20th May, 1935

Professor E.B.Wilson,
School of Public Health,
Harvard University,
Van Dyke St.,
BOSTON, Mass. U.S.A.

Dear Dr. Wilson.

I have now looked through W.E.Deming's paper and my reaction to it is not unlike yours. He seems to be much concerned to put others (unnamed) right on points on which it is very doubtful whether anyone has been mistaken, or has been under any misapprehension. His impression as shown on page 5, that what he calls the u test is somehow more fundamental than Student's t test seems to rest on a confusion between problems ems of estimation & tests of significance – confusion which has been, I am afraid, accentuated bythe discussions of Neyman and Pearson on the subject.

It is a little unfortunate that he should persist in calling the t test "the z test," since Student changed his notation now ten years ago, and the z test is normally used for a more inclusive test, of which t and χ^2 are special cases.

I believe the root of his misapprehension, which is a confusing one, may be traced in the opening paragraph of his discussion on page 11, for he seems to think that the u test and the t test are applicable in the same problem, whereas the u test can only be applied when the population variance is known, and the t test when it is unknown, but capable of estimation from the data. He proceeds solemnly to show that, if applied to the same data, and if each is used to select the same fraction, 1% or 5% of samples from the hypothetical population on testing which the hypothesis shall be rejected, somewhat different sets of samples would be chosen by the two tests. This point, for what it may be worth, has been sufficiently emphasised in a rather silly paper by Treloar and Wilder which he cites (No. 13 of his references). What first Treloar and Wilder, and later, it seems, Deming failed to see is that, in setting our level of significance at any value such as 1%, we are choosing voluntarily to make the mistake of rejecting the hypothesis when it is true in this proportion of cases; and as, on the hypothesis discussed, we are always equally wrong in such rejection, it is a matter of complete indifference, provided the proportion is kept right, in which particular samples this mistake is made.

Obviously, of course, when the population variance is known, the test which utilises this knowledge is preferable to one using only an estimate, and the reason for this preference is not that we should prefer to be deceived by one sort of sample rather than by another.

It is curious that the simple history of the problem should not have kept the auther straight. From the beginning of the Theory of Errors, it must have been clear that, knowing the population variance, one could use the normal curve to test the significance of a mean. The test can easily be extended to a weighted mean, or to the regression coefficient, as soon as this is recognised as merely a weighted mean.

In the majority of practical problems, it was easily perceived that the population variance was not provided in the data, but, if the sample were large, could be estimated with some confidence, so that the problem was discussed by Gauss, what estimate it was best to take. Gauss satisfied himself that the mean square error, with allowance for degrees of freedom, was most proper, though, under the influence of Peters, astronomers were largely persuaded to use the mean error instead. As soon, however, as any estimate is used rather than a known value, the question of its errors of estimation arises; and throughout the nineteenth century it was usual to say that we needed some large number, e.g. 50 observations, before relying on the test. 'Student's' work showed that, using the mean square error, the frequency of rejection of the hypothesis, when true, can be calculated exactly, and for small samples was larger than that derived by substituting the estimate in the formula for the normal curve. 'Student's' tables, in fact, show how to make the classical test exact.

I therefore disagree entirely with the paragraph commencing at the foot of page 13, and should agree with Sterne, that it is a refinement of the classical procedure, but not of P_α as defined by Deming, as this only refers to the relatively infrequent case in which the sampling variance is known before the data are collected.

The author says that the main purpose of his paper is interpretation, and, if that is so, much of his exposition is redundant; it should also require the author to be more careful to understand the history and present ideas on the subject.

Yours sincerely,