JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 70B, No. 2, April–June 1966

A Note on Contaminated Samples of Size Three

T. A. Willke*

(February 25, 1966)

Estimation of the mean and standard deviation using the closest two of three observations in a sample from a normal population with contamination by slippage of the mean is investigated by a sampling study. Lieblein's results, which indicated that the use of these statistics is not advisable for noncontaminated samples, are borne out by this study for contaminated samples as well.

Key Words: Outliers, contaminated samples, robust estimates, best two of three.

1. Introduction

In the physical sciences samples of only three measurements of a quantity are not uncommon, and estimation of the actual value of the quantity from these few measurements poses some difficult problems. If it is known that all of the measurements are good, that is, that they are measurements of the same quantity and that they contain no blunders or gross errors, then the sample mean has no serious competitor as an estimate of the true mean for measurement data which are approximately normally distributed.

Quite often, though, there is a definite possibility that one or more of the measurements contain an error which is not just due to the uncertainties of the measurement process, but which results from a slip in the procedure, a failure of some component of the measurement apparatus, a misread dial, etc. Such a measurement is called a contaminant and sometimes, depending on the purpose of the experiment, should be discarded from the sample. Unfortunately, unless the error is very large it is usually difficult to determine whether a measurement is a contaminant or not. The chemistry lab teacher who advises his students to take three measurements and use only the closest two of them in their calculations has recognized this problem and uses this device in an attempt to get robust estimates which are not likely to be as affected by a contaminant as the ordinary estimates are. The main purpose of this paper is to examine how sound this procedure is.

Lieblein [1955] derived distributions of some statistics, especially the mean and the range of the closest two of three independent observations from the same normal distribution, and he discussed their properties as estimators of the mean and the standard deviation of the population. He found them to be inefficient and generally unreliable compared to the mean and the range of all three observations. However, an experimenter would use the closest two of three observations to compute his estimates only if he thought that the sample might be contaminated. Lieblein considered the null case where no contamination exists, hence he has evaluated the penalty one must pay by using these protected estimates when they are not really needed. To complete the picture we must find out how much, if any, the experimenter stands to gain if there is contamination, and thus see how robust these estimates really are. This, of course, depends on how much and what kind of contamination is present, and this note describes the results of a sampling experiment in the important case where the contamination is by slippage of the mean. It is assumed throughout that the standard deviation is not known. If some prior knowledge of the standard deviation is available the contaminants are easier to detect and the treatment of the problem is changed.

2. Estimation of the Mean With Exactly One Contaminant

Let x_1, x_2, x_3, be an independent sample of size three with x_1 and x_2 from a normal distribution with the mean μ and the variance σ^2. Let x_3 be from a normal distribution with the mean $\mu+\delta\sigma$ and variance σ^2. The order statistics for the sample will be denoted $x_{(1)}<x_{(2)}<x_{(3)}$. Let x' and x'' be the closest two of the three observations with $x'<x''$. Then consider the following statistics as estimates of the mean;

$$\bar{x}=(1/3)\sum x_i$$

$$m=x_{(2)}$$

$$y_3=(1/2)(x'+x'').$$

In particular we are interested in (1) the bias or the difference between the expected value of the statistic and μ, and (2) the root mean square error of the estimate from μ. In the null case, $\delta=0$, all three esti-

*University of Maryland, College Park, Md.; part time at the National Bureau of Standards.

mates are unbiased and the root mean square errors of $\bar{x}$ and m are known to be

$$\{E[(\bar{x}-\mu)^2]\}^{1/2}=\frac{\sigma}{\sqrt{3}}=0.577\sigma, \text{ and}$$

$$\{E[(m-\mu)^2]\}^{1/2}=0.670\sigma;$$

and of y_3 was found by Lieblein to be

$$\{E[(y_3-\mu)^2]\}^{1/2}=\left(\frac{1}{2}+\frac{\sqrt{3}}{4\pi}\right)^{1/2}\sigma=0.799\sigma.$$

Thus, in the null case, the sample mean is about twice as efficient as y_3. It should be noted that y_3 has a larger standard error than the mean of an uncontaminated sample of size two has (0.707σ), so that when there is no contamination the chemistry student is better off taking just two measurements and averaging them than he is taking three and using the best two of the three.

In the sampling experiment 1,000 samples of size three (containing two uncontaminated values, x_1 and x_2, and one contaminated value, x_3) were taken for each of the values $\delta=0$, 1, 2, 3, 4, and 6, to see how the above estimators performed in nonnull situations. The bias and the root mean square error of these estimates as determined by the sampling experiment are graphed in figure 1. (The results of the sampling experiment agree with the exact values, which are known for $\delta=0$ and for $\bar{x}$, almost to within the accuracy of the graphs. The graphs for the median, m, are exact values calculated from its probability distribution function.)

The three lines in the graphs labeled $\eta=1/3$, $\eta=1/6$ and $\eta=1/11$ correspond to estimates of the mean where either $\bar{x}$ or y_3 is used depending on the spacing of the three measurements as follows. If the spacing between the three measurements is about the same, then there is little evidence of a contaminant and one would want to use $\bar{x}$. However, if one of the three is relatively far removed from the other two, then the natural tendency is to discard it and use the average of the other two, namely y_3. A decision rule for this policy can be formulated as follows. Let

$$u=x''-x'$$

$$\omega=x_{(3)}-x_{(1)}.$$

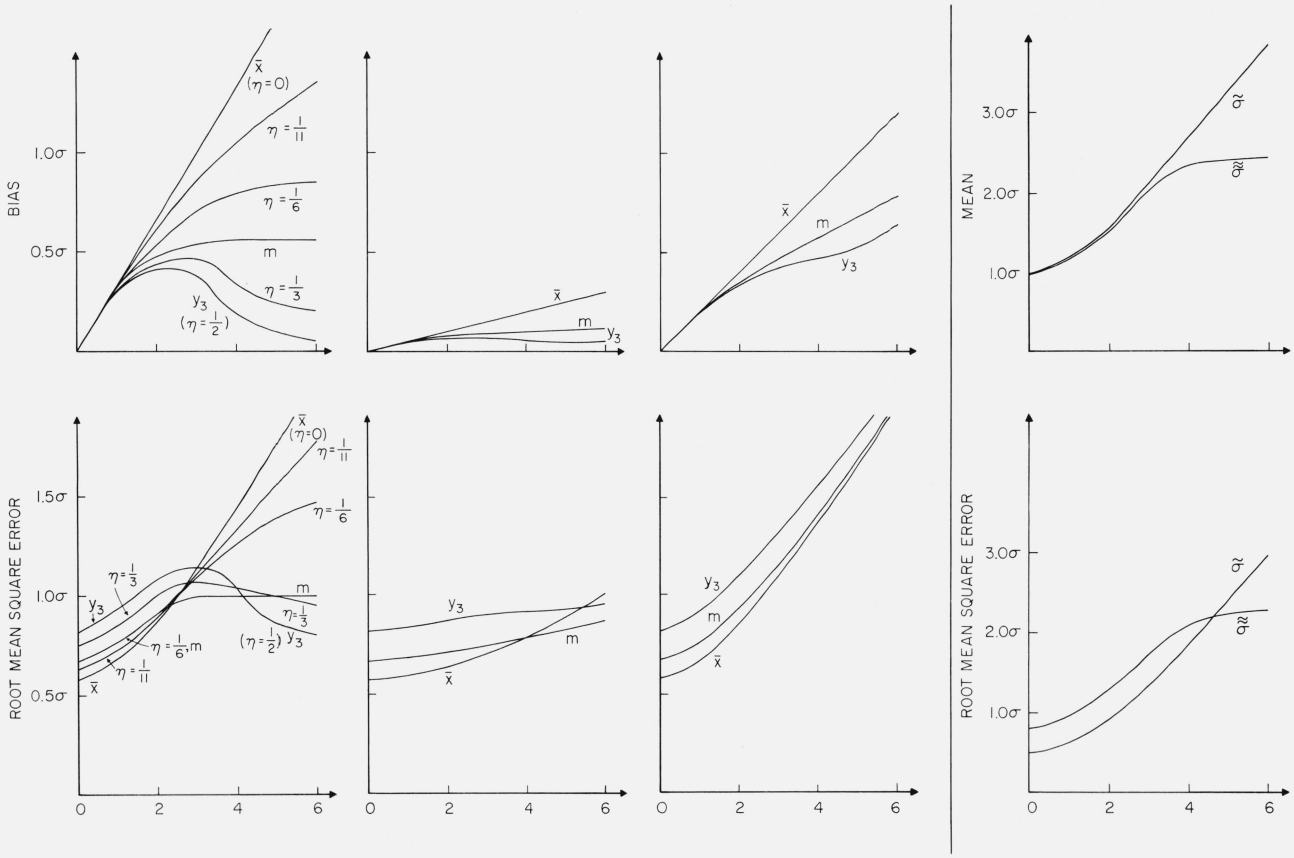

FIGURE 1
Exactly one contaminant

FIGURE 2
5% contamination

FIGURE 3
20% contamination

FIGURE 4
Exactly one contaminant

Biases and root mean square errors for various estimates of the mean.

Means and root mean square errors for two estimates of the standard deviation.

Then, if $\frac{u}{\omega} < \eta$, where η is a preassigned constant, use y_3 as the estimate, and if $\frac{u}{\omega} \geq \eta$ use $\bar{x}$. For example, suppose η is chosen to be $\frac{1}{11}$, then if the measurement on one end is more than ten times as far from the middle measurement as the one on the other end, y_3 is used; otherwise $\bar{x}$ is used. Notice that $\eta = 0$ corresponds to using $\bar{x}$ always and $\eta = \frac{1}{2}$ corresponds to using y_3 always.

The bias in $\bar{x}$ increases linearly with δ whereas that of y_3 goes to 0 since the two good measurements are almost always used for y_3 when δ is large. For the same reason the bias in the median levels out to 0.564σ, the expected value of the second order statistic in a sample of size two.

The graph of the root mean square error shows that although in the null case y_3 is quite inefficient, it results in a real saving for large values of δ and this fact would seem to support the opinion of the chemistry teacher.

3. Estimation of the Mean With a Random Number of Contaminants

The graphs discussed above can be misleading since they are for a model in which there is known to be exactly one contaminant. In practice the difficult task usually is to decide whether there is a contaminant present. Even in noncontaminated samples the ratio $\frac{u}{\omega}$ can be deceptively small (the probability that $\frac{u}{\omega} \leq \frac{1}{11}$ is about 0.157). Someone who does not fully appreciate the vagaries of small samples can easily be led to believe there is a contaminant present when there are none.

Perhaps a more realistic model would be to assume that any one of the measurements has a certain chance of being contaminated, hence there could be 0, 1, 2, or even 3 contaminants in the sample. If there are two or three contaminants in a sample of three then estimation of the mean is hopeless anyway, but it is possible that the use of y_3 leads to a false sense of security, or does even more damage than $\bar{x}$ or m, particularly when all contaminants are from the same source.

In figures 2 and 3 are graphed the bias and the root mean square error for these estimates when there is a 5 percent and a 20 percent chance respectively that each particular measurement is a contaminant and when this probability is independent of the probability for the other two measurements. The graphs were calculated from the results of the sampling experiment above by the use of binomial probabilities. For this model y_3 loses much of its advantage, especially for the high contamination of 20 percent because there is a reasonable chance that two of the sample values are contaminants and then y_3 exhibits even a larger bias than $\bar{x}$. Moreover, x has a uniformly smaller root mean square error than either y_3 or m has up to $\delta = 6$. This, of course, is true only because in this model all contaminants have their mean displaced in the same direction. If contamination comes from both sides the bias may be eliminated.

The graphs for the optional estimates with $\eta = 1/3$, 1/6, and 1/11 are not given, but they lie midway between the graphs of y_3 and $\bar{x}$ in that respective order.

4. Estimation of the Standard Deviation

The same sampling experiment was used to evaluate properties of different estimators of the standard deviation, σ. Two estimators were considered,

$$\tilde{\sigma} = 0.591\omega = 0.591(x_{(3)} - x_{(1)})$$

$$\tilde{\tilde{\sigma}} = 2.205y_1 = 2.205(x'' - x').$$

The estimate based on the range, ω, is the usual estimate for very small sample sizes, but again, if contamination is feared, one might want to use $\tilde{\tilde{\sigma}}$, the estimate based on the range of the closest two. The factors, 0.591 and 2.205, make the estimates unbiased in the null case. Lieblein [1955] has shown that $\tilde{\tilde{\sigma}}$ is quite inefficient compared to $\tilde{\sigma}$ in the null case, in fact

$$\{E(\tilde{\sigma} - \sigma)^2\}^{1/2} = 0.524\sigma$$

$$\{E(\tilde{\tilde{\sigma}} - \sigma)^2\}^{1/2} = 0.826\sigma.$$

Here, just as for the mean, the range of just two true duplicates provides a better estimate than $\tilde{\tilde{\sigma}}$.

The bias and the root mean square error of $\tilde{\sigma}$ and $\tilde{\tilde{\sigma}}$ as determined from the sampling experiment are graphed in units of σ in figure 4. These graphs are for the model in which there is exactly one contaminant in the sample. Graphs for the 100γ percent contamination model (corresponding to figures 2 and 3 for the mean) are not included in this note, but they also indicate the superiority of $\tilde{\sigma}$ over $\tilde{\tilde{\sigma}}$ when all contaminants come from the same source and $\delta \leq 3$.

5. References

Anscombe, F. J., Rejection of outliers, Technometrics **2**, 123–147 (1960).

Dixon, W. J., Rejection of outliers, ch.IOH in Contributions to Order Statistics (edited by A. E. Sarhan and B. G. Greenberg) (John Wiley & Sons, Inc., New York, N.Y., 1962).

Dixon, W. J. and Massey, F. J., Jr., Introduction to Statistical Analysis, pp 275–278 (McGraw-Hill Book Co., New York, N.Y., 1957).

Lieblein, Julius, Properties of certain statistics involving the closest pair in a sample of three observations, J. Res. NBS **48**, 255–268, (1955) RP2311.

Proschan, Frank, Testing suspected observations, Industrial Quality Control **13**, No. 7, 14–19, (1957).

(Paper 70B2–175)