Correction of Data for Errors of Averages Obtained from Small Samples

By W. A. SHEWHART

Bell Telephone Laboratories, Incorporated

SYNOPSIS: Recent contributions to the theory of statistics make possible the calculation of the error of the average of a small sample—something that cannot be done accurately with customary error theory. Obviously, these contributions are of very general importance, because experimental and engineering sciences alike rest upon averages which in a majority of cases are determined from small samples, and because an average cannot be used to advantage without its probable error being known.

The present paper attempts to show in a simple way why we cannot use customary error theory to calculate the error of the average of a small sample and to show what we should use instead. The points of interest are illustrated with actual data taken for this purpose. The paper closes with applications of the theory to four types of problems involving samples of small size for each of which numerous examples arise in practice. These types are:

1. Determination of error of average.
2. Determination of error of average difference.
3. Determination of most probable value of the root mean square deviation of the universe when only one sample of n pieces has been examined.
4. Determination of most probable value of the root mean square deviation of the universe when several samples of n pieces each have been examined.

USEFUL THEORY OVERLOOKED: WHY?

PRACTICALLY everyone uses averages—research workers and engineers in particular. Moreover, all of us have long appreciated the fact that an average is often only of value when we know its probable error. Naturally, we turn to the theory of errors to guide us in calculating the probable error. Naturally, because from 1733 to 1908 there was nothing else that we could turn to. Since 1908 the recognition has been gradually making headway that to use customary error theory for determining the probable errors of averages of small samples is a mistake.

The story of how to calculate the probable error of a small sample was originally told in Biometrika, a journal for the statistical study of biological problems—a veritable mine of useful information. The truth was given in equations involving terms familiar only to statisticians and hence was concealed from many. The story, however, with the aid of such experimental results as are used in this paper can be told in a simple manner: it is of interest to all of us who, for one reason or another, cannot make large numbers of observations on every quantity that we measure, but must nevertheless estimate the probable errors of our results. In this discussion, diagrams will be

used instead of equations, and, because of this rather popular presentation, many readers may want to consult, as the original sources, the intensely interesting mathematical contributions of "Student",[1] Professor Karl Pearson,[2] and R. A. Fisher.[3]

Case Where Customary Theory Applies

We start, as in customary error theory, with the assumption that the probability distribution of errors is normal. This simply means that the probability of the occurrence of an error within any range is assumed to be equal to the area under the so-called normal curve[4] (such a curve is shown in Fig. 1) between the limits of the same range.

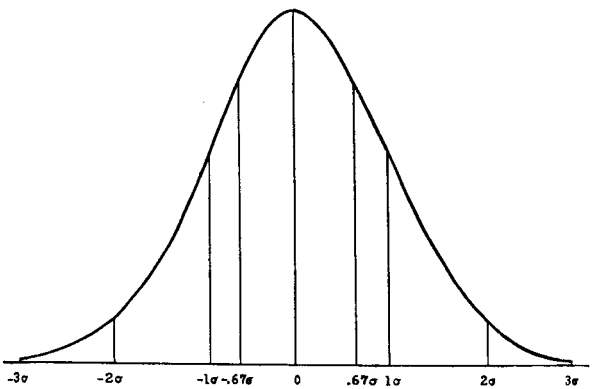

Fig. 1—Customarily assumed law of error curve—normal law

50.00000% of area within $0 \pm .67449\sigma$
68.26894% of area within $0 \pm 1\sigma$
95.44998% of area within $0 \pm 2\sigma$
99.73002% of area within $0 \pm 3\sigma$

The total area under the curve is, of course, unity. This curve is plotted with the origin at the true value and with the errors measured in units of the root mean square error σ. The fractions of the area bounded by certain multiples of the root mean square error are shown for reference.

Let us make an experiment and see how far customary error theory

[1] Biometrika, Vol. VI, 1908, pp. 1–15. Vol. XI, 1917, pp. 416–417.

[2] Biometrika, Vol. X, 1915, pp. 522–529.

[3] Biometrika, Vol. X, 1915, pp. 507–521. Proc. Camb. Phil. Soc., Vol. XXI, 1923, pp. 655–658.

[4] The equation for this has recently been traced back to Abraham De Moivre (1733) by Professor Pearson. See Biometrika, Vol. XVI, 1924, pp. 402–404.

carries us, see where it breaks down, see why it breaks down, and then avail ourselves of the new theory—a powerful tool of great value, because it makes possible for the first time the solution of many practical problems. Here is the experiment. Take 998 small circular chips, 499 green and 499 white. Mark 20 white ones with 0, 40 white ones with 0.1, 39 white ones with 0.2, etc., in accordance with the normal law. Do the same for the green chips except that all numbers on the chips are minus. Put the 998 chips in a bowl, mix thoroughly, draw out one and record it. Replace the chip, again mix thoroughly, and repeat the process until 4000 values are observed. A little reflection shows that this experiment is equivalent to making 4000 measurements of a quantity by a method subject to a normal law of error with a root mean square error of approximately unity.

Let us group these 4000 values into 1000 groups of 4, and determine the average for each group, taking the first four observations as the first group, the second four as the second group and so on. This gives

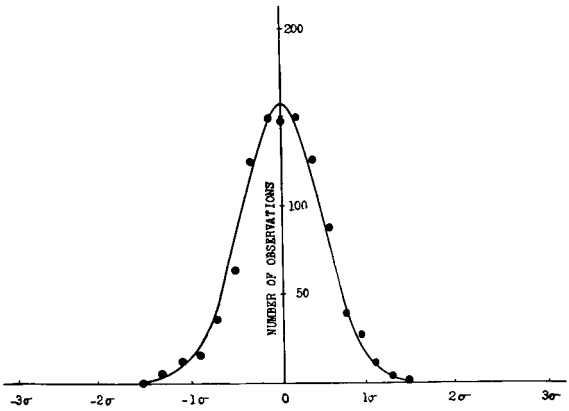

Fig. 2—Curve showing customary error theory to be satisfactory on one condition not often met in practice; *i.e.*, σ is known

Distribution of 1000 averages of 4

– Normal law with root mean square error $\frac{\sigma}{\sqrt{4}}$

us 1000 averages. Suppose we subtract the true value m (in this case zero) from each average and divide this result by the root mean square error of the frequency distribution of values within the bowl. This gives us 1000 observations of the error of the average of 4 observations measured in terms of σ. Customary error theory shows that these averages should be distributed normally as indicated by the smooth

curve in Fig. 2 with a root mean square error of $\frac{\sigma}{\sqrt{4}}$ or one half that in Fig. 1. The dots show the experimental results.[5]

So far the customary error theory is satisfactory. But we do not often have this case in practice; that is, we do not know the root mean square error σ, and instead know only the observed root mean square error s of the sample.[6]

Case Where Customary Theory Does not Apply

Let us next recall just the way we use the customary theory in practice and then see what mistake we usually make. Take the results of drawing the first sample of 4 in the experiment previously cited. The four observed values are .6, −.2, 1.1, −2.0, the average $\overline{X}$ of these

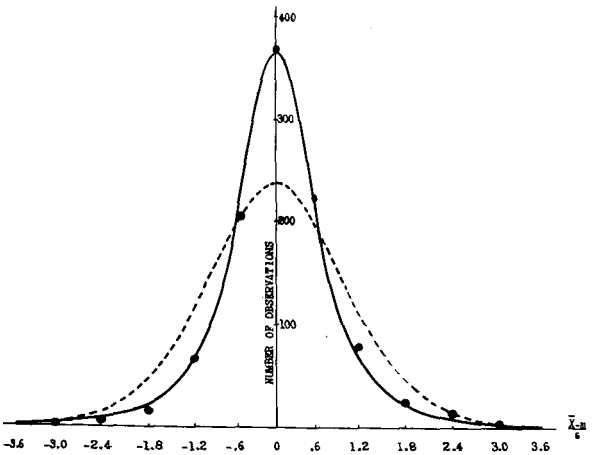

Fig. 3—Curves showing inaccuracy of customary error theory in finding error of average in terms of the observed standard deviation s

– – – – – Customary theory
——— New theory
• Distribution of 1000 z's

is −.125, and the observed root mean square deviation s is 1.177. Assuming no knowledge of the root mean square error σ of the distribution from which the sample of 4 was taken and using customary theory, we should assume the probable or 50% error to be $.6745\frac{1.177}{\sqrt{4}}$

[5] I am indebted to Miss Victoria Mial and Miss Marion Cater for securing the experimental results, making all necessary calculations, and drawing the curves given in this paper.

[6] Customarily we do not know the true value m, hence instead of knowing the root mean square errors we know the root mean square or standard deviations.

This follows from the fact that the observed values of the ratio $z = \frac{\bar{X} - m}{s}$ where m is the true value, are customarily assumed to be distributed normally. *Here we come to the crux of the discussion: these observed values of the ratio are not distributed normally.* "Student"[7], in 1908, was the first to show how they are distributed.

Let us look at the observed frequency distribution of the 1000 z's given by the above experiment (dots Fig. 3). To be normally distributed, as customarily assumed, these dots would have to lie on the dotted normal curve. Obviously they do not. Instead they lie on a much more peaked curve (solid line) than the normal. This was calculated with the aid of "Student's" theory. We must therefore conclude: the probability that the mean of a sample of n, drawn at random from a normal distribution, will not exceed (in the algebraic sense) the mean of that distribution by more than z times the root mean square deviation of the sample cannot be found from the normal law when n is small. We must use the tables provided by "Student" in the two papers referred to above.

Why the Customary Theory Fails to Give the Error of the Average in Case of Small Samples

Let us look a little further into the reason why the z's are not distributed normally, before we consider the question as to the magnitude

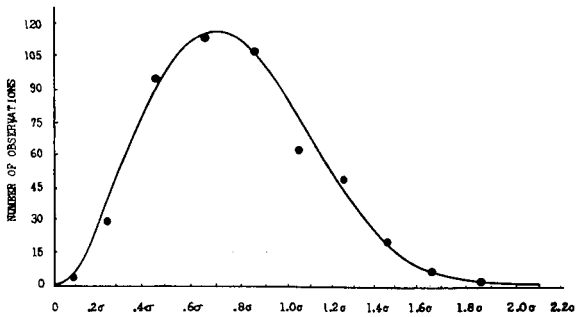

Fig. 4—Data furnishing a clue to reason for inadequacy of customary error theory
• Observed distribution of standard deviations of 1000 samples of four
– Theoretical curve of asymmetrical type

of the difference between the probable error determined from one theory and that determined from the other.

Let us look at the distribution of the 1000 standard deviations, the s's, Fig. 4, for here we shall find the secret revealed: The distri-

[7] Loc. cit.

bution of s's, as we might expect, is asymmetrical; the most probable standard deviation s, to be observed is not the average s. Of course, the customary theory assumes that the average s is the most probable s, and that the distribution of s is normal. We should therefore expect to find the z's distributed normally for values of n such that the dis-

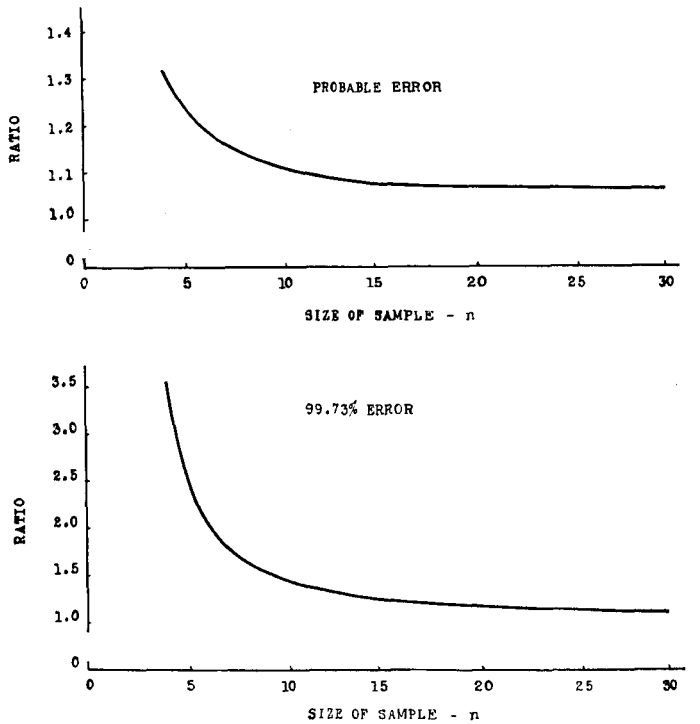

Fig. 5—Chart showing magnitude of correction for size of sample—ratio of the errors to their customarily accepted values

tribution of observed standard deviations is approximately normal. Now, Professor Pearson[8] has developed the theory underlying the distribution of s. He finds that as n increases, the distribution of s rapidly approaches normality. Even for n greater than 25 the distribution has approached normality to such an extent that we should expect the z's to be distributed approximately in normal fashion. The study of the distribution of z shows this to be true, as we shall see below.

In passing, we should note how closely the theoretical curve, Fig. 4, fits the observed points and also note two other checks between theory

[8] Loc. cit.

and observation furnished by the new data given herein. According to theory, the modal and mean values of s for samples of size 4 expressed in units of σ should be .707 and .798 respectively. The experimental results are .717 and .801.

How Much Larger Are the Probable and 99.73% Errors of an Average Than the Customarily Accepted Values?

The difference between the error of an average and its customarily accepted value increases as the number of observations n (or size of

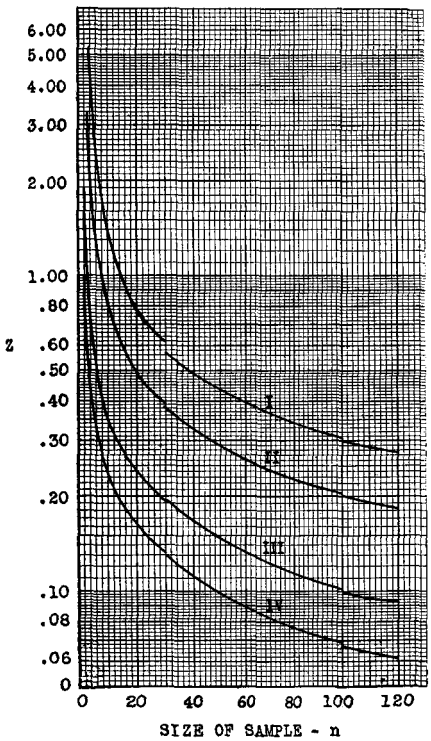

Fig. 6—Errors of averages of samples of size n
I -99.73002% error
II -95.44998% error
III -68.26894% error
IV -50.00000% error
z = the ratio of the error of the average to the observed standard deviation

sample) decreases. This fact is illustrated in Fig. 5. This figure shows the ratios of the errors to their customarily accepted values plotted for values of n from 4 to 30.

Curves showing the most frequently used errors of averages measured in terms of z (*i.e.* in terms of the ratio of the error to the observed standard deviation) are given in Fig. 6. The error curves for n less than 30 have been obtained with the aid of "Student's" original tables, those for n between 30 and 100 have been obtained from the normal law integral tables using the standard deviation of z; *i.e.* $\frac{s}{\sqrt{n-3}}$ as given by "Student." For n greater than 100, customary error theory has been used.[9]

Typical Practical Applications

But few, if any, recent developments of statistical theory are of more general application in most fields of scientific research and engineering than the one herein described.[10] This follows because the theory herein discussed must be used in calculating the required probable error (or other measure of dispersion) of the averages obtained from small numbers of observations. The number of applications of this character is legion.

Problem Type 1, Determination of Error of Average

Example 1:

Five samples of granular carbon taken from a crucible show resistances of 47.5, 49.4, 43.2, 48.0 and 46.2 ohms respectively. What are the probable and 99.73% errors of the average of these resistances?

Solution:

The observed values of average resistance $\overline{X}$, and standard deviation $s=\sqrt{\frac{\Sigma(\overline{X}-X)^2}{n}}$ are 46.9 ohms and 2.097 ohms respectively. Hence from Fig. 6 we see that the probable and 99.73% errors are respectively $.372s=.780$ ohms and $3.33s=6.99$ ohms respectively whereas from customary theory they would be $.302s=.633$ ohms and $1.34s=2.81$

[9] For the curves in this figure as in the preceding one, I have assumed the customary theory for the case where the true value of X is known so that the root mean square error of the average $\overline{X}$ of sample of size n is the ratio $\frac{s}{\sqrt{n}}$. Of course, as we know from customary error theory, if we assume no knowledge of the true value of X, we should use $\frac{s}{\sqrt{n-1}}$.

[10] Since this paper was written, a very interesting article, "Statistics in Administration," has appeared in *Nature* (V. 117, pp. 37–38, Jan. 9, 1926), calling attention to the importance of the theory of small samples.

ohms respectively. The true probable and 99.73% errors are 23% and 148% higher respectively than those calculated by customary theory, as is evident from Fig. 5.

Discussion of Type 1:

Examples of this type of problem are obviously so numerous that further illustrations need not be given. They occur every day in practically every science. We see that in such cases it is certainly necessary to allow for the effect of the small size of sample.

PROBLEM TYPE 2, DETERMINATION OF ERROR OF AVERAGE DIFFERENCE

Example 1:

Five instruments are measured for some characteristic X, first on one machine and then on another, giving two sets of values $X_{11}, X_{12}, \ldots X_{15}$, and $X_{21}, X_{22}, \ldots x_{25}$ respectively. Calculate the 5 differences $X_{11}-X_{21}=x_1$, $X_{12}-X_{22}=x_2$, $\ldots X_{15}-X_{25}=x_5$. Assume that the average difference is $\bar{x}$ and the standard deviation of the differences is s. Assuming the two machines give the same results except for random variations, what is the probability that the observed difference would occur? Are we justified in the assumption that the machines give the same results?

Solution:

The true difference is zero on this assumption. The observed difference is $z=\frac{\bar{X}-0}{s}$, and "Student's" tables may be used to evaluate this probability.[11] If this probability is very small, let us say .001 or less, it may be taken as indicating that the machines do not give the same results.

Example 2:

We wish to compare the depth of penetration obtained from two different methods of preserving chestnut telephone poles. We choose n poles for test. A sample from each pole is treated by one process, and a sample from each pole is treated by another process. The depths of penetration are measured. Are we justified in assuming the two methods to give significantly different results?

[11] Approximate values can be obtained from the curves in Fig. 6.

Solution:

If n is small, we proceed as in the previous case, to find the probability of occurrence of the observed difference. If this probability is small, we conclude that the difference is significant; *i.e.*, the two methods of preservation give different results.

Example 3:

Three-bolt guy clamps are used for clamping the guy wires on telephone poles. These are supplied from different sources. Those from one source fail to hold the wire as well as those from another and inspection shows that these same clamps fail to meet a certain specified dimension. The force required to slip the wire in each of 10 clamps from this source is measured. These clamps are then modified to meet the specified dimension and the force required to slip the wire in each clamp is again measured. Are we justified in attributing the failure to hold the wire to the fact that these clamps did not meet the specification?

Solution:

The solution follows the same line as in the first case.

Discussion of Problems of Type 2:

Problems of this type are very numerous. It is obvious that significant differences calculated as above indicated are always larger than those calculated by customary theory.

Problem Type 3, Determination of Most Probable Value of the Root Mean Square Deviation of the Universe When Only One Sample of n Pieces Has Been Examined

Example 1:

Five tool-made models are tested for their efficiency, giving values $X_1, X_2, \ldots X_5$. What is the most probable value of the range within which the efficiencies of product instruments may be expected to lie approximately 99.7% of the time, assuming that a manufacturing process can be developed which is the same as that used in producing the tool-made models?

Solution:

Customary practice would answer: the average of the five values plus or minus 3 times their standard deviation. The better answer is: the average plus or minus $\frac{3}{.7746}$ times the standard deviation.

This follows from Professor Pearson's work previously quoted. He has shown that the most probable observed standard deviation $\breve{s}$ of

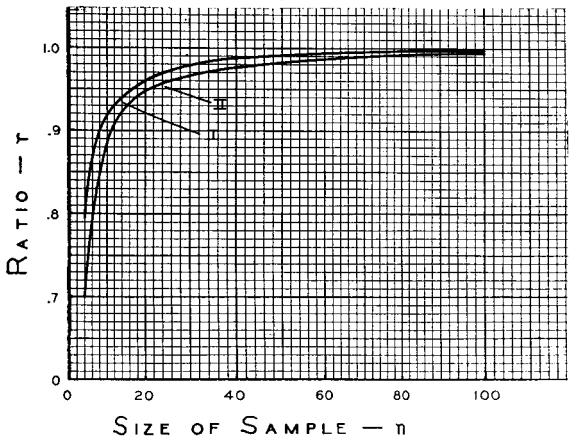

Fig. 7—Curves giving the most probable value of the true standard deviation σ
I When the average $\bar{s}$ of standard deviations of many samples is known. $r\sigma = \bar{s}$
II When the standard deviations of one sample is known. $r\sigma = \breve{s}$

a sample of n from a normal distribution with standard deviation σ is $\breve{s} = \sqrt{\frac{n-2}{n}}\,\sigma$. Substituting the value $n=5$ in this equation we get $\breve{s} = .7746\sigma$.

A curve of the values of $\frac{\breve{s}}{\sigma}$ vs. n is presented in Fig. 7 for reference in solving problems of this character.

Problem Type 4, Determination of Most Probable Root Mean Square Deviation of the Universe When Several Samples of n Pieces Each Have Been Examined

Example 1:

One thousand transmitters, known to have different efficiencies, have been tested five times each for efficiency. Find the standard deviation of the machine method of measurement.

Solution:

Calculate the standard deviation of the five tests for each transmitter. Find the average value of these 1000 values and divide it by .8407. This follows from the fact that the average $\bar{s}$ of the observed standard deviation for a series of samples of size n drawn from a normal distribution with standard deviation σ is

$$\bar{s} = \sqrt{\frac{2}{n}} \frac{\left|\frac{n-2}{2}\right.}{\left|\frac{n-3}{2}\right.} \sigma.$$

where the symbol $\left|\underline{X}\right.$ is equivalent to $\Gamma(X+1)$

Thus for $n=5$ we get [12] $\bar{s}=.8407\sigma$.

Fig. 7 also presents the values of the ratio $\frac{\bar{s}}{\sigma}$ for reference and with sufficient accuracy for solving problems similar to the example cited. Greater accuracy than that afforded by the curves can be secured by direct substitution in the equations for $\bar{s}$ and $\bar{s}$ or by referring to the original tables.

[12] We will recall with interest how closely the observed average, $\bar{s}=.798\sigma$, of the 1000 values of s corresponding to the 1000 samples of four herein presented checked the theoretical average of $.801\sigma$.