NOTE ON THE PROBABILITY ASSOCIATED WITH THE ERROR OF A SINGLE OBSERVATION

By Dr. W. A. Shewhart
Bell Telephone Laboratories

The Problem: The object of the present note is to give an empirical indication of the magnitude of the correction which must be applied to the customary Theory of Error estimate of the probability associated with the error of a single observation or, in other words, to the estimate of the probability associated with a given rang.

We can make the practical significance of this problem clear by considering a very simple type of example. The determination of the strength of wood entering into any kind of construction involves a knowledge of the modulus of rupture of the particular kind of timber being used where, of course, modulus of rupture is a certain particularly located stress per unit area under breaking load. Obviously, it is of considerable importance in this as in many similar problems to know the probability associated with any range so that we may determine the expected number of pieces of wood of the particular species under consideration to be found in the future within any chosen limits. The most comprehensive and valuable source of information on the strengths of timber is perhaps the series of bulletins published by our own government laboratories. For example, Table 1 of Bulletin 556 of the United States Department of Agriculture gives results of modulus of rupture tests on 126 species of wood. The number of trees tested per species, however, varies from 2 to 60, the modal number being 5. Tests of this nature are very expensive because they involve the selection, the preparation and shipment of trees to the testing laboratories, not to mention the cost of making the physical tests in the laboratory. This problem is merely typical of that of setting standards for physical properties of materials whether they be of engineering or other interest.

Now an engineer making use of a given kind of wood naturally wants to know the probable number of trees having a particular quality within a given range so that he may compare one species of trees with that of another to determine which meets his needs better. In other words he wants to know the probability associated with a given range.

The results herein presented show that the customary estimate of this probability *is too large by several per cent* when the sample size "n" is small as so often is the case in many important engineering problems such as the one indicated.

In the particular case in hand more extensive data are not available primarily because of the cost of accumulating the same. Many cases arise, however, where it is impossible to secure large numbers of observations. This would be true if we wished, as the Civil Engineers often do, to determine the probability that the annual flood run-off in future years will fall within a certain range. In many such cases records have been kept for only a few years and hence only a few observations are available upon which to base an estimate of this probability.

Customary Solution: Engineering practice is, of course, to use error theory for estimating the probability associated with any range. Thus, if $X_1, X_2, \cdots, X_n$ represent n observed values of a chance variable X, the probability P_2 of a future observation falling within a range $X=L_1$ to $X=L_2$ is assumed to be given by the integral

$$P_1=\int_{L_1}^{L_2}\frac{1}{\sigma\sqrt{2\pi}}e^{-(X-\overline{X})^2/2\sigma^2}dX, \qquad (1)$$

where $\overline{X}$ is the arithmetic mean of the sample and σ is the estimate, determined from the sample of size n, of the standard deviation σ' of the universe. Naturally the use of this integral involves the assumption that the universe is normal, and the following discussion will be limited by this same assumption except where otherwise noted. It is apparent that this integral would give the true probability provided we knew the average $\overline{X}'$, and the standard deviation σ' of the universe. In that case P_2 would be equal to P_1. However, in practice we seldom, if ever, have this information, hence P_2 is not in general equal to P_1.

If we let $l_1=L_1-\overline{X}$ and $l_2=L_2-\overline{X}$, then what we really need to know is the theoretical distribution function for the probability associated with the range $\overline{X}+l_1$ to $\overline{X}+l_2$. Such functions for the ranges of interest are not available and so recourse was made to an empirical determination of some of these.

Experimental Results: In most engineering work, as in practically every field of science, two ranges are of great interest, viz., $\overline{X}\pm.6745\sigma$ and $\overline{X}\pm 3\sigma$ where as before σ is the estimate of the true standard deviation σ' of the universe. It is generally assumed that

the probability associated with either one of these ranges is that given by the normal law integral, equation (1). For the first range, this gives .500 and for the second .997. As already stated, experimental results presented below show that these two probabilities are higher than they should be, particularly for small samples. In our discussion we shall also consider the range $\overline{X} \pm t\sigma$ where t is any real number. In general we shall find that the probability associated with a given range as found from equation (1) is always greater than the expected probability P_2 that should be associated with this range.

Now, there are two obvious reasons why the expected probability P_2 associated with the range $\overline{X} \pm t\sigma$ should be different from the expected probability P_1 associated with the corresponding range $\overline{X}' \pm t\sigma'$ although the use of equation (1) assumes the equality of these two probabilities. One reason is that a range of given value subtends a greater area of the universe when spaced symmetrically in respect to the average $\overline{X}'$ of the universe than the same range subtends when spaced symmetrically about $\overline{X}$ unless, of course, $\overline{X} = \overline{X}'$. The other reason is that the distribution of σ is unsymmetrical and the difference $P_1 - P_2$ depends upon the method of calculating σ or, in other words, of estimating σ'.

If we make use of the method of maximum likelihood[1], we have

$$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}} \tag{2}$$

and the difference $P_1 - P_2$ for $n = 4$, as will be seen below, is approximately .14 and .10 respectively for $t = .6745$ and $t = 3$. Specifically the use of the integral of equation (1) together with the estimate of σ' given by equation (2) indicates probabilities .500 and .997. For ranges corresponding to the two values of t just noted, we should correct these results to read .36 and .90, respectively. For example in the pole problem mentioned above we should say that about 40% and 92% of future samples of poles should be expected to have moduli of rupture lying within the respective ranges $\overline{X} \pm .6745\sigma$ and $\overline{X} \pm 3\sigma$, where σ is calculated by equation (2) and $\overline{X}$ is the average of the moduli of rupture for the 5 poles, instead of 50.0% and 99.7% given by the method of maximum likelihood.

[1] Whittaker and Robinson, Calculus of Observations, First Edition, pp. 186-187.

In a similar way we find that, if the customary Theory of Error estimate of σ' is used, viz.,

$$\sigma = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}, \tag{3}$$

then the difference $P_1 - P_2$ for $n=4$ becomes approximately only .07. Even this difference, however, is quite too large an effect to be overlooked unnecessarily in many practical problems.

Another method of estimating σ' is to make use of the most likely estimate σ of σ' determined from the relation[2]

$$\sigma = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-2}} \tag{4}$$

The present study shows that the use of this value of σ reduces the difference $P_1 - P_2$ for $n=4$ to approximately .05 and .01, respectively for the ranges $\overline{X} \pm 3\sigma$ and $\overline{X} \pm .6745\sigma$.

This difference is not large enough to be of great engineering importance in most cases. Hence we seem to be justified, upon the basis of these empirical results, in using the integral of equation (1), together with σ given by equation (4), it being apparent that the difference $P_1 - P_2$ for other than a symmetrical range in respect to $\overline{X}'$ should be less than that for a symmetrical range, and that, as the sample size increases, the difference $P_1 - P_2$ rapidly decreases.

The first part of the experimental study consisted of drawing 1000 samples of four from a normal universe and calculating the ranges $\overline{X} \pm .6745\sigma$, $\overline{X} \pm 3\sigma$, $\overline{X}' \pm .6745\sigma$, $\overline{X}' \pm 3\sigma$, for σ as given by equations (2) and (4). For each range for each of the samples of four, the fraction of the universe standing on this range as base was determined. This fraction represents the probability of an observed value X falling within this range. The averages of the observed probabilities (1000 for each kind of range for each σ) associated with the different ranges are found in column 2 of Table 1. The average probability corresponding to σ determined from equation (3) was computed by interpolation.

---

[2] Pearson, Karl, "On the Distribution of Standard Deviations of Small Samples," *Biometrika*, Vol. 10, 1915, pp. 522-529. The engineering use of this relationship is discussed by the present author in an article "Correction of Data for Errors of Averages," *Bell System Technical Journal*, Vol. V, April 1926, pp. 308-319. In the writer's opinion the present paper empirically justifies some of the applications there made.

The ranges $\overline{X}' \pm t\sigma$ are included to indicate the magnitude of the difference between the expected probability associated with the range $\overline{X} \pm t\sigma$ and the probability associated with the range $\overline{X}' \pm t\sigma$. From these results it becomes evident that the variation in σ from sample to sample apparently has more effect in producing a difference $P_1 - P_2$ than does the variation in $\overline{X}$ about $\overline{X}'$.

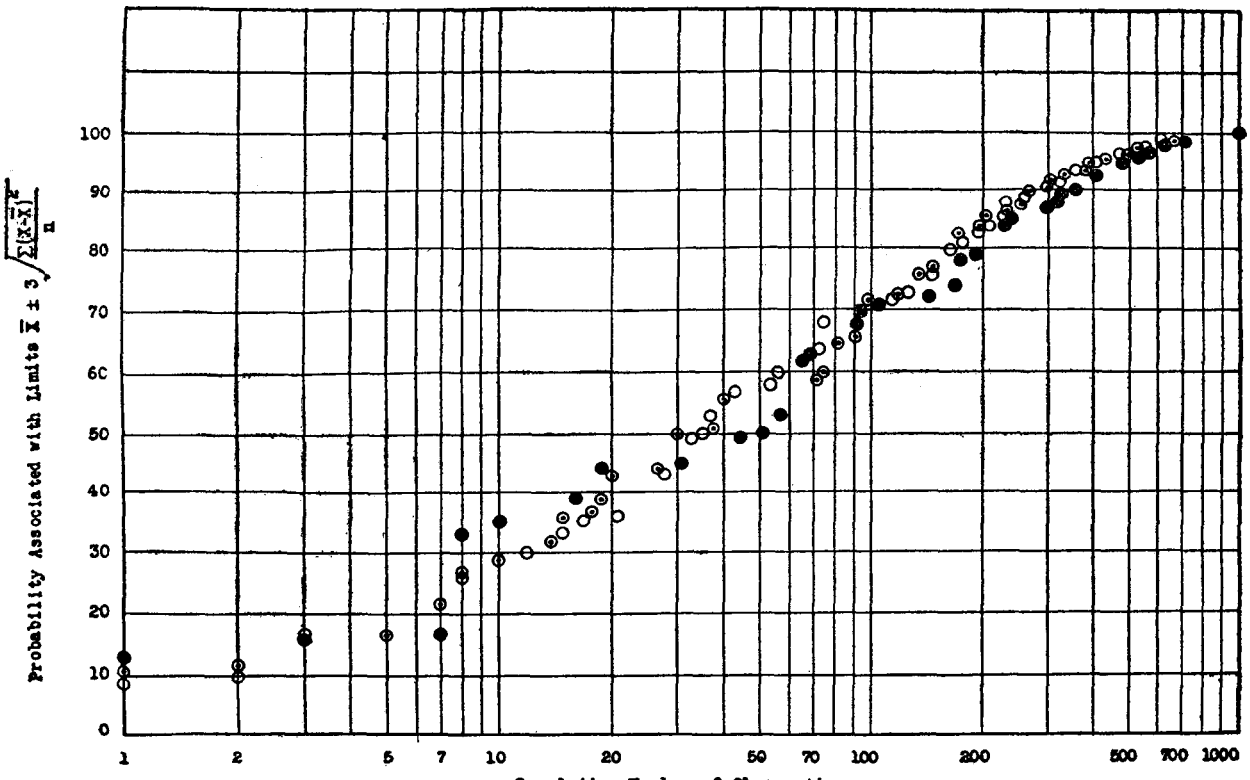
TABLE 1. PROBABILITY ASSOCIATED WITH CERTAIN RANGES*

Range	Normal Universe	Rectangular Universe	Triangular Universe
$\overline{X}' \pm 3\sigma$	.92	—	—
$\overline{X} \pm 3\sigma$	.90	.91	.91
$\overline{X}' \pm 3\sqrt{\frac{n}{n-2}}\sigma$	.97	—	—
$\overline{X} \pm 3\sqrt{\frac{n}{n-2}}\sigma$	.95	.96	.96
$\overline{X}' \pm .6745\sigma$	.40	—	—
$\overline{X} \pm .6745\sigma$	.36	—	—
$\overline{X}' \pm .6745\sqrt{\frac{n}{n-2}}\sigma$	.53	—	—
$\overline{X} \pm .6745\sqrt{\frac{n}{n-2}}\sigma$	.49	—	—

* In this table $\sigma = \sqrt{\frac{\Sigma(X-)\overline{X}^2}{n}}$

It is also of interest to note the wide dispersion of the distribution of probabilities associated with a given range. The black dots in Fig. 1 give as a typical case the observed distribution for the range $\overline{X} \pm 3\sigma (n = 4)$. We see that one observed probability was as low as .13, that seven did not exceed .18; that 50 did not exceed .50 and that 500 or one-half of the probabilities did not exceed .95. This means that about 500 of the observed probabilities fell between .95 and 1.00, and that once in this series of 1000 trials we obtained a range which corresponded to a probability of .13 when we would have assumed that it corresponded to a probability of .90.

It need scarcely be mentioned of course that the experimental results given above do not prove that the magnitude of the difference $P_1 - P_2$ is of the order indicated above. In fact, it would not be possible to prove that the true difference had been observed even though the numbers of observations were increased at will. The probability is approximately .99, however, that the observed dif-



Empirical results showing that customary error theory does not give expected probability associated with a given range.

M8645F

● Normal Universe
⊙ Rectangular Universe
○ Triangular Universe

ference $P_1 - P_2$ in any case does not differ by more than .01 from the true difference.

So far, our discussion has been limited to the consideration of samples drawn from a normal universe. What is the significance of this limitation? No complete answer to this question is forthcoming but sufficient work has been done to indicate that some type of correction must be considered in the interpretation of some of the important generalized formulas of sampling, such as those given by Tchebycheff, Camp, and others.

In connection with some other work, 1000 samples of four were drawn from each of two universes,—one rectangular, $\beta_1' = 0$ and $\beta_2' = 1.8$ and the other a right triangular universe $\beta_1' = 1.32$ and $\beta_2' = 2.4$. The average probability associated with the ranges $\overline{X} \pm 3\sqrt{\frac{\Sigma(X-\overline{X})^2}{n}}$ and $\overline{X} \pm 3\sqrt{\frac{n}{n-2}}\sqrt{\frac{\Sigma(X-\overline{X})^2}{n}}$ were determined for each of these universes and the results are included in Table 1. A very striking coincidence is observed, namely, that the average probabilities are practically the same for the three kinds of universe. That this must be a mere coincidence is obvious upon a little consideration of the reasons why these probabilities should not be the same in general for all ranges. For both of these universes the probability associated with the range $\overline{X}' \pm 3\sigma'$ is 1.00 which corresponds to certainty. Hence we see that there is a really important difference $P_1 - P_2$ for each of these universes.

It appears, therefore, that in practice we must take into consideration the effect of applying the general sampling formulae a posteriori that are supposed to be applied a priori.

Conclusion: If we feel that we are justified in assuming that we are sampling from a normal universe, it appears that we may use the integral of equation (1) as giving the probability associated with a given range provided we use the estimate of σ given by equation (3) without introducing errors of greater magnitude than those indicated in this paper *and which are always less than those given by customray theory.* If we know nothing about the universe and still use this method of finding the probability associated with a given range, enough has been done to show that sometimes at least we shall not be led very far astray, although much more work remains to be done to cover this general case.