K217
WAS

3 15

BELL TELEPHONE LABORATORIES
INCORPORATED

JULY 1928

REPRINT B-327

SMALL SAMPLES--NEW EXPERIMENTAL RESULTS

BY

W. A. SHEWHART and F. W. WINTERS

EXPERIMENTAL RESULTS
ILLUSTRATING SOME ADVANTAGES AND LIMITATIONS
OF THE LATEST ERROR THEORY OF SMALL SAMPLES

SMALL SAMPLES—NEW EXPERIMENTAL RESULTS

By W. A. Shewhart and F. W. Winters, *Bell Telephone Laboratories, Inc.*

The problem of determining the error of an average of a small sample is merely one of a whole class of sampling problems which belong to the very important and rapidly developing branch of investigation commonly classified as mathematical statistics. The researches in this field, by no means a new one, have received great impetus from the recent work of Pearson, R. A. Fisher, "Student" and others who, within the past decade, have provided the scientific tools which make possible the solution of certain practical problems previously unsolved.

Needless to say, however, all of the important questions in this field have not been answered; and even those that have, very frequently suffer in a practical way from the limitations underlying such answers. Also in some of the recent developments, already referred to, it is not possible to meet in a practical way all of the conditions upon which the theoretical solution rests and hence we must rely to a certain extent upon experimental investigation of the nature offered herein to justify the application of the theory to the practical problems which arise. In fact it is just the effects of such limitations that this paper is to emphasize in such a way, it is hoped, as will indicate that the fruits of future research may be as great or even greater than those so recently obtained.

Specifically we shall do three things:

1. Show the practical necessity for using averages of small samples.
2. Show empirically that "Student's" theory is a marked improvement over customary error theory, and indicate theoretically why even this latest theory fails in most *practical* cases to give the error of the average obtained from a small sample.
3. Show by experimental results that the discrepancy between present theory and practice warrants further theoretical studies.

WIDESPREAD USE OF SMALL SAMPLES

The necessity for knowing how to analyze the results of small numbers of measurements is becoming well established. For example, most physicists and chemists deal with comparatively few observations and many large industrial laboratories carry on routine analyses of raw materials, inspections of product and research investigations where, for one reason or another, often not more than five observations are made. Many of the properties of materials have been established on comparatively small numbers of widely dispersed observations as, for example, the modulus of rupture of saturated woods of different species where the standard deviation of a set of tests is usually at least 25 per cent of the average of this same set. This is true even though extensive use is to be made of the results.

In the above instances it would be possible to obtain a greater number of observations although sometimes the cost per observation may run into hundreds of dollars. Occasionally, however, it is humanly impossible to obtain further measurements even though we desire to do so. An example would be the design of a levee of sufficient height to hold back the flood waters of the Mississippi. An engineer's estimate would of necessity be based upon the records of the run-off of that area whereas such records have only been maintained for a comparatively few years.[1]

"STUDENT'S" THEORY AN IMPROVEMENT OVER CUSTOMARY ERROR THEORY

Let us review briefly the modern error theory for a sample drawn from a normal universe and then consider the nature of the limitations imposed by this theory. Specifically, customary error theory assumes a set of n observed data constituting a sample to be drawn from a normal universe characterized[2] by equation 1,

$$dy=\frac{1}{\sigma'\sqrt{2\Pi}}e^{-\frac{(X-\bar{X}')^2}{2\sigma'^2}}\quad dX, \qquad (1)$$

where $\bar{X}'$ and σ' are the mean and standard deviation respectively of the universe, and dy represents to within infinitesimals of higher order the probability of an observed value of X falling within the range X to $X + dX$. It follows, as is well known, that the distribution of means of samples of size n will be given by the same expression where $\frac{\sigma'}{\sqrt{n}}$ is substituted for σ' and the variable X now becomes the average $\bar{X}$ of the sample.

Now, customary error theory uses as an estimate of σ' the following expression calculated from the sample

$$\sigma_1=\sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}, \qquad (2)$$

X being measured from an arbitrary origin. Obviously σ_1 may or may not be equal to σ' and in general will be distributed asymmetrically

[1] Still other cases, to be more amply discussed in a future paper, arise where more information can be gained by breaking up a large sample into sub-samples and analyzing the sub-samples upon the basis of modern theory than can be obtained by treating the large sample as a unit.

[2] Throughout this article the *primed* notation is used to represent parameters of the theoretical frequency function, as distinct from the unprimed estimates of these calculated from the sample.

about σ', the lower limit of the distribution being, of course, $\sigma_1=0$. In 1908 "Student" pointed out that, if we use σ_1 as an estimate of true standard deviation σ' of the universe, we can no longer assume that $t=\frac{\sqrt{n}(\overline{X}-\overline{X}')}{\sigma_1}$ will be distributed normally. In the investigation of this problem, however, "Student" actually found the distribution of a slightly different variable $z=\frac{\overline{X}-\overline{X}'}{\sigma}$ where $\sigma=\sqrt{\frac{\Sigma(X-\overline{X})^2}{n}}$, and in this case X is measured from the true mean $\overline{X}'$ of the universe, $\overline{X}$ being the mean of the sample.

The distribution of z is derived upon the following assumptions:

1. That the distribution of means of samples of size n is normal.
2. That the distribution of σ is given by the expression[1]

$$f(\sigma)d\sigma=\frac{n^{\frac{n-1}{2}}}{2^{\frac{n-3}{2}}\left|\underline{\frac{n-3}{2}}\right.}\frac{\sigma^{n-2}}{\sigma'^{n-1}}e^{-\frac{n\sigma^2}{2\sigma'^2}}d\sigma, \qquad (3)$$

   where $f(\sigma)d\sigma$ is to infinitesimals of higher order the probability of occurrence of a value of σ within the range σ to $\sigma+d\sigma$.

3. That $\overline{X}$ and σ of a sample of size n are not correlated.

Papers by "Student,"[2] Pearson,[3] Fisher[4] and others have established the justification of these three assumptions for the case of the normal law and have provided a comprehensive set of tables for estimating the error of the average in terms of the observed standard deviation. Now, we shall consider the application of this theory to some experimental results.

A study of the data obtained from 1000 samples of four drawn from each of three universes, normal, rectangular and right-triangular gives some interesting information as shown in Chart I. The observed frequency distributions of z from the three universes are shown in this figure, but at present we shall confine our attention to the normal case.

The solid curves give the probability (in per cent) of the occurrence of a value of z between $-\infty$ and any assigned value, say z_0. Of course these curves also may be used to give the probability of observing a

[1] Provided we define $\left|\underline{\frac{n-3}{2}}\right.=\Gamma\left(\frac{n-1}{2}\right)$. ($n$ even.)

[2] "Student," *Biometrika*, Vol. VI, pp. 1–15, 1908; Vol. XI, pp. 416–417, 1917; *Metron*, Vol. V, No. 3, pp. 18–21, 1925.

[3] Karl Pearson, *Biometrika*, Vol. X, pp. 522–529, 1915.

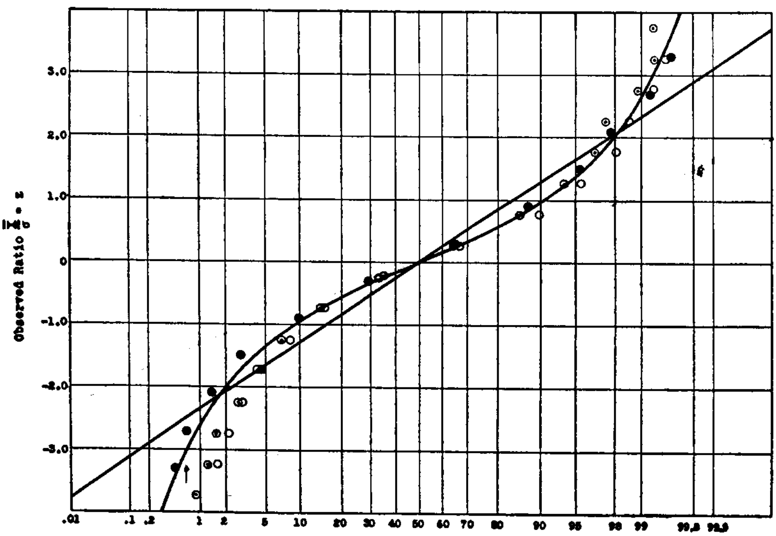
[4] R. A. Fisher, *Biometrika*, Vol. X, pp. 507–521, 1915; *Proceedings of the Cambridge Philosophical Society*, Vol. XXI, pp. 655–658; *Metron*, Vol. V, pp. 3–17 and pp. 22–32, 1925.

value of z between $-z_0$ and $+z_0$. The straight line represents customary theory which assumes that the probability of an observed mean differing from the true mean of the universe by any assigned multiple of the observed standard deviation of the mean, is given by the normal law frequency function. It is clear that the twisted curve representing "Student's" integral gives quite a different result from that obtained by the older theory.

Doubtless anyone acquainted with "Student's" work will recognize

CHART I

PROBABILITY ASSOCIATED WITH RANGE ($-\infty$ TO z)



—— Customary Theory

Student's Theory

● Observed ratio Normal Universe

⊙ Observed ratio Rectangular Universe

O Observed ratio Triangular Universe

at once the essential points of difference between it and normal law theory. The important fact to be noted here is that for small samples like four or five such as often occur in practical problems, "Student's" theory gives a much better representation of the facts than does customary theory.

The black dots represent the observed distribution of z for 1,000 samples of four drawn from a normal universe. We see how closely these dots follow the theoretical curve of "Student" even though the range of the sampled universe was of necessity limited and not infinite as theory assumes. However, does "Student's" theory still apply when the sampled universe is not normal?

Inasmuch as the distribution of z is based upon the three stated assumptions, we naturally would be led to inquire whether these assumptions were still justified when the universe is not normal. The rigorous answer is no, not a single one of them is justified. What then is the effect of the limitations underlying "Student's" distribution?

We give now the theoretical basis for these effects and show later how some of them contribute to the failure of "Student's" theory when, for want of a better method, it is applied to other more general cases than that for which it was intended.

From the work of Tchouproff, Pearson and others as cited in a recent article by Church,[1] we have

$$\sigma'_{\overline{X}} = \frac{\sigma'}{\sqrt{n}},$$

$$\beta'_{1_{\overline{X}}} = \frac{\beta'_1}{n},$$

$$\beta'_{2_{\overline{X}}} = \frac{\beta'_2 - 3}{n} + 3$$

where $\sigma'_{\overline{X}}$, $\beta'_{1_{\overline{X}}}$, $\beta'_{2_{\overline{X}}}$ represent the standard deviation, skewness, and kurtosis of the distribution of means of samples of size n drawn from a universe characterized by σ', β'_1 and β'_2. We see that, as n becomes large, $\beta'_{1_{\overline{X}}} \to 0$ and $\beta'_{2_{\overline{X}}} \to 3$ or the values of skewness and kurtosis corresponding to the normal law. Irrespective, therefore, of the shape of the universe the distribution of means of samples of size n, so far as characterized by the parameters $\beta'_{1_{\overline{X}}}$ and $\beta'_{2_{\overline{X}}}$, approaches normality with a standard deviation $\frac{\sigma'}{\sqrt{n}}$. In the same article by Church, he gives the first four moments of the distribution of the variance σ^2 for samples of size n drawn from any universe characterized by the parameters β'_1, β'_2, β'_3, β'_4 and β'_6. From a study of these it seems that the distribution of σ for universes other than normal may differ materially from that given by equation 3.

Hence, in general, though the first assumption may be approximately fulfilled even for comparatively small values of n, we may not expect the second assumption to be even approximately satisfied except in rare cases. Another important limitation is introduced through the correlation between the mean $\overline{X}$ and variance σ^2 of a sample. Neyman[2] has shown that the average $\overline{X}$ and variance σ^2 are correlated for

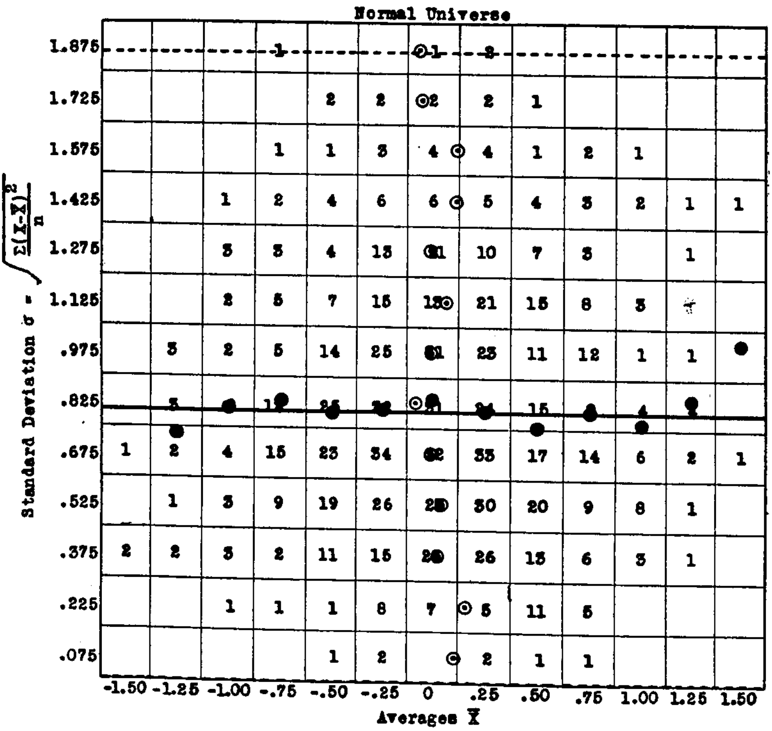
[1] A. E. R. Church, "On the Means and Squared Standard Deviations of Small Samples from Any Population," *Biometrika*, Vol. XVIII, pp. 320–394, November, 1926.

[2] J. Neyman, "On the Correlation of the Mean and Variance in Samples Drawn from an Infinite Population," *Biometrika*, Vol. XVIII, pp. 401–413, 1926.

all samples drawn from other than a normal universe. Moreover, in practice even though we know the universe to be normal in appearance it nevertheless embraces only a finite range. Hence we must always expect correlation between the mean $\overline{X}$ and variance σ^2, *i. e.* even "Student's" theory probably never rigorously applies to a practical problem.

CHART II

CONTROL CHART TAKING INTO ACCOUNT THE CORRELATION BETWEEN STANDARD DEVIATION AND AVERAGES FOR SAMPLES OF 4

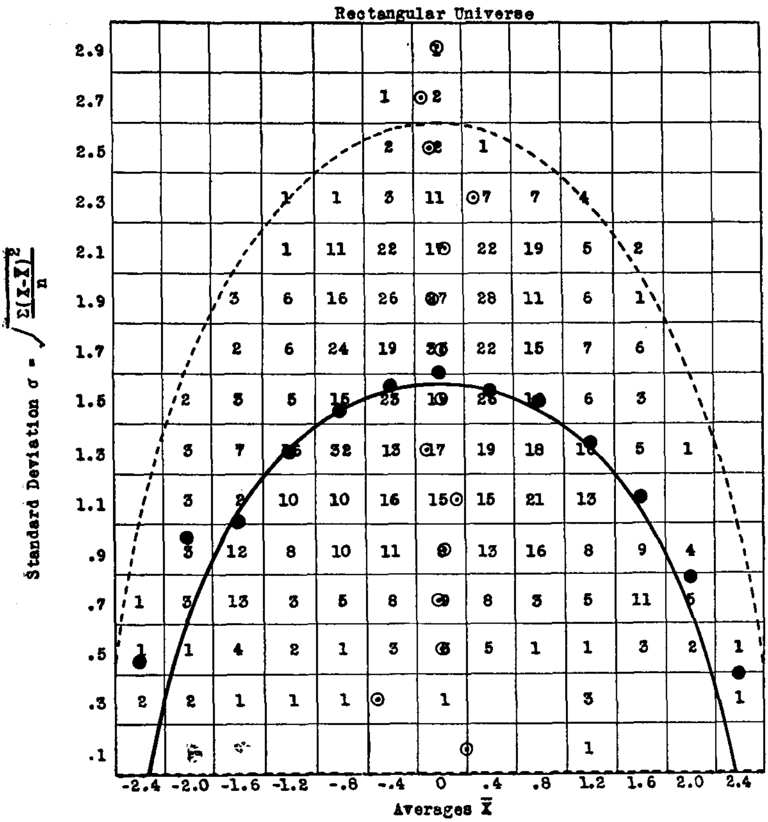


The truth of this statement becomes evident when we study the regression[1] of standard deviation on average for the three universes as shown in Chart II. Even in the case of the normal universe there seems to be some correlation between σ and $\overline{X}$, at the ends of the range of averages. This is to be expected, however, since the range of the experimental normal universe, as also of the other universes was of necessity limited. The dotted limit lines theoretically should include 99.7 per cent of the observations.

[1] The curves of regression were constructed following the methods described by Neyman. We are indebted to Miss Marion B. Cater and Miss Miriam S. Harold for carrying out the experimental work and making all calculations.

For samples from universes characterized by values of β'_1 and β'_2 satisfying the relation $\beta'_2-\beta'_1-3=0$, the regression of variance on the mean is linear. For samples from universes with β''s lying above this line, the regression to the first order of approximation is parabolic with

CHART II (*Continued*)



——— Theoretical Regression Curve of Standard Deviation on Average

● Observed Points of Regression of Standard Deviation on Average

– – – – Theoretical Limits

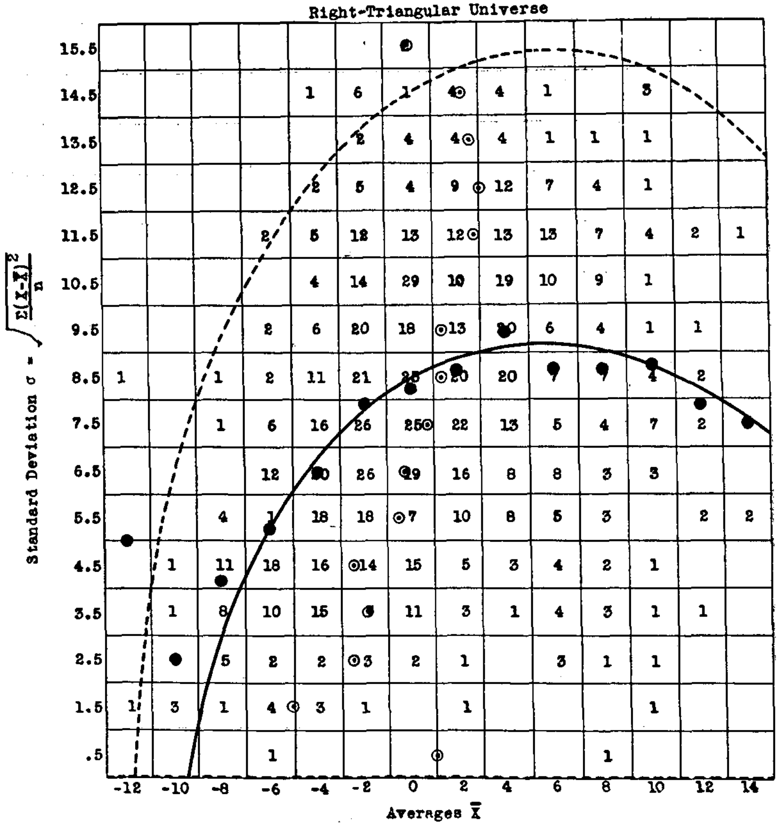
branches directed in the negative sense (downward), and for samples from universes with β'_1's lying below this line, the branches of the parabola are directed in the positive sense (upward). The universes chosen for study are examples of the first two types of regression.

The effect of this correlation upon the distribution of $z=\frac{x}{\sigma}$ where $x=\bar{X}-\bar{X}'$ becomes evident. For linear regression we would expect the distribution to be skew and therefore not the same as that given by

"Student." For the first of the two above-mentioned types of non-linear regression, "Student's" integral will give probabilities corresponding to a given range of z which are too high and, for the second type of non-linear regression, "Student's" integral will give probabili-

CHART II (*Continued*)

Right-Triangular Universe



ties which are too low, provided in each case that both assumptions (1) and (2) are met.

We present now some empirical evidence to serve as a background for the above statements as well as to show in picturesque form the failure of even the best available theory.

EMPIRICAL RESULTS SHOWING IMPORTANCE OF FURTHER RESEARCH

Since in practice the three stated assumptions are, strictly speaking never met, we have no general way of determining how closely "Student's" integral will give the distribution of z. We are obliged, therefore, at the present time to analyze the situation empirically.

The dots ⊙ and ○ in Chart I represent the observed distribution of z for 1,000 averages of four drawn from the rectangular and right-triangular universes respectively. Apparently the effect of the correlation is the controlling factor in these two cases. At least the results are as would be expected, since β'_1 and β'_2 for each of the two non-normal universes lie above the line $\beta'_2-\beta'_1-3=0$. The observed distributions of means and of standard deviations obtained from samples of four drawn from the rectangular universe are very much like those of the means and standard deviations of samples drawn from the normal universe. Assumptions (1) and (2) are, therefore, approximately met in this case whereas a glance at the regression between x and σ for this universe will suffice to show that assumption (3) is not even approximately fulfilled. We note that as x increases in absolute value the corresponding average value of σ decreases, which means that $z=\frac{x}{\sigma}$ increases. The significance of this fact is that a greater number of z's will lie outside a certain $|z|$ (the curve of regression of σ on x being approximately symmetrical) than is the case for the z's from a normal universe where the average value of σ for a given x is practically constant.

In terms of "Student's" integral we may, therefore, state, with some confidence, that the value obtained from the integral for the range $-z$ to $+z$ is too large, at least for most values of z between 0 and 3. The physical significance of this fact is seen in the arrangement of the dots ⊙ relative to "Student's" curve, that is they are on the concave side of the latter and therefore subtend between the range $-z$ to $+z$, a smaller probability than that indicated by "Student's" theory, *i. e.* the range[1] corresponding to a given probability as given by "Student's" theory is too small.

How is it for the case of the right-triangular universe? The observed distributions of means and standard deviations of samples drawn from this universe are quite similar in appearance to the corresponding distributions from the rectangular universe. The curve of regression of σ on x for this case is in general character like that for the rectangular universe only not so symmetrical. What was said about z, as the value of x increases in the rectangular, may be repeated in general for this case, particularly for those z's corresponding to negative values of

[1] A still more important problem than that described in this paper has to do with the expected probability associated with the range given by $\bar{X}\pm t\sigma_{\bar{X}}$ where

$$\sigma_{\bar{X}}=\sqrt{\frac{\Sigma(X-\bar{X})^2}{n^2}}.$$

Empirical information obtained from 150 observations shows that this probability is approximately 92 per cent when $t=3$.

x. On the positive side the average σ for a given x does not decrease so rapidly as x increases and we would, therefore, expect a closer agreement between "Student's" curve and that of the observed z's when the latter are positive. The position of the dots O relative to "Student's" curve over the observed range seem to agree very well with what we might expect from such a study of the regression curve. Here then as in the rectangular case we feel justified in concluding that "Student's" integral for the probability of an observed value of z falling within the range $-z_0$ to $+z_0$ is too large, at least for most values of $z_0 \leqq 3$, *i. e.* the range associated with a given probability as indicated by "Student's" theory is too small.

In the light of the theoretical and empirical results here presented it seems likely that the probability associated with a given range $-z$ to $+z$ as given by "Student's" integral must be considered as an upper bound. In other words, the range associated with a given probability as given by "Student's" integral must be considered as a lower bound when sampling from a universe whose β''s lie above the line $\beta'_2 - \beta'_1 - 3 = 0$.

SUMMARY

In closing let us describe briefly the ideas we have attempted to convey. We have the practical problem of estimating the error of the average obtained from a small sample. For the case of sampling from a normal universe the empirical results here presented show that "Student's" theory applies even though assumptions (1), (2) and (3) are, strictly speaking, not met because of the physical limitations of any universe such as must be used to check theory.

For the case of sampling from a non-normal universe, we have pointed out the theoretical reasons for the failure of "Student's" theory. The empirical results show that the correction that may be necessary, if we attempt to apply "Student's" theory, when sampling from other than a normal universe is of sufficient magnitude to warrant further extension of theory to cover the cases which arise in practice.