REVIEWS OF
MODERN PHYSICS

VOLUME 6 JULY, 1934 NUMBER 3

On the Statistical Theory of Errors

W. EDWARDS DEMING, *Bureau of Chemistry and Soils, U. S. Department of Agriculture*
AND
RAYMOND T. BIRGE, *University of California, Berkeley*

TABLE OF CONTENTS

A. LIST OF FIGURES

B. LIST OF TABLES

C. NOMENCLATURE

Small Greek letters will be used for the parent population, small English letters for a sample, and capital English letters for a collection of samples. The left hand column lists the page on which the symbol first occurs.

Page	*Symbol*	*Explanation*
123	ν,n	the number of items in the parent population, and the number of items in the sample.
123	$\mu,\bar{x}$	the means of the parent population and of the sample.
125, 126	σ,s	the square roots of the second moments about the means, or the standard deviations (S.D.), of the parent population and of the sample.
125	$x_1, x_2, \cdots, x_n$	the observations constituting a sample.
125	$\epsilon_i \equiv x_i - \mu$	true errors.
125	$v_i \equiv x_i - \bar{x}$	residuals.
126	$u \equiv \bar{x} - \mu$	true error of a sample.
125	ρ	probable error (p.e.) of a single observation.
127	r	p.e. of the mean of n observations.
126	r.m.s.	root mean square.
127	N	an indefinitely large number of samples of n drawn from the parent population.
128	$\check{s}$	the mode of the sampling distribution of s, Eq. (14), Helmert's equation.
128	$\bar{s}$	the mean of the sampling distribution of s.
128	$\dot{s} = \sigma/f$	the median of the sampling distribution of s. This defines f. σ would be the median on the sampling distribution of fs.
128	$B(m,n)$	the beta function $\int_0^1 x^{m-1}(1-x)^{n-1}dx$. The arguments m and n are interchangeable.
129	$\Gamma_x(n)$	the incomplete gamma function $\int_0^x x^{n-1}e^{-x}dx$.
129	$\Gamma(n)$	the complete gamma function (the same integral with limits 0 to ∞).
129	z	$\equiv u/s$, the true error of the mean in units of the S.D. s of the sample (abscissa of Student's distribution, Eq. (21)).
132	δ	defines a contour of arbitrary constant altitude on the u,s frequency surface.

Page	*Symbol*	*Explanation*
132	λ	defines a contour on the u,s frequency surface along which the probability of a given set of errors bears the constant ratio λ to the maximum value that this probability can attain.
132	P_u	probability of drawing a sample with true error greater than $\|u\|$.
133	γ	denotes $0.674\cdots/\sqrt{2}=0.476936276\cdots$.
133	P_s	probability of drawing a sample with S.D. greater than s.
133	χ^2	argument in the chi-test.
135	P_z	probability of drawing a sample with $\|u/s\|=\|z\|$ greater than a specified value.
137	P_λ	probability of drawing a sample lying outside a specified λ contour of the u,s frequency surface.
137	P_δ	probability of drawing a sample lying outside a specified δ contour of the u,s frequency surface.
140	ζ	value of $\|z\|$ for which $P_z=\frac{1}{2}$, i.e., the quartile deviation in Student's distribution of z in samples of n.
141	ϕ	denotes the factor $0.674\cdots f/\sqrt{n}$. r would be the median on the sampling distribution of ϕs.
142	$\sigma(s,5)$	the 5 percent fiducial value of σ corresponding to a given value of s. There is 1 chance in 20 that the S. D. of the parent population is greater than $\sigma(s,5)$ for the given value of s.
142	$s(\sigma,95)$	the 95 percent fiducial value of s corresponding to a given value of σ. There are 19 chances in 20 that the S. D. of the sample is greater than $s(\sigma,95)$ for the given value of σ.
143	$r(s,5)$	the 5 percent fiducial value of r corresponding to a given value of s. There is 1 chance in 20 that the probable error of the mean of n observations is greater than $r(s,5)=\phi_{95}s$.
142	f_{95}	that particular value of σ/s designated by $\sigma(s,5)/s$ or by $\sigma/s(\sigma,95)$.
143	ϕ_{95}	denotes the factor $0.674\cdots f_{95}/\sqrt{n}$; $\phi_{95}s=r(s,5)$.
143	f_{50}	the same as f; the subscript 50 is used for emphasis, especially in discussing 50 percent fiducial values of σ and s.
143	ϕ_{50}	the same as ϕ; the subscript 50 is used for emphasis, especially in discussing 50 percent fiducial values of σ and s.
145	σ_s, r_s	*estimates* of σ and of r derived from the sample alone.
147	ωs	some multiple of s, denoting an estimate, σ_s, made from a sample.
148	F	the r.m.s. error in an estimate of σ, in units of σ, or the r.m.s. error in an estimate of r, in units of r, both of which are equal to the *estimated proportional* r.m.s. error in an estimate of σ or of r.
150, 151	$\phi(\sigma)$	the ordinate at the abscissa σ on a *prior existence curve* for the S. D. of the parent population.
150, 151	p or $p(\sigma)$	the ordinate at the abscissa σ on a *posterior curve* for the S. D. of the parent population.
154	s_0	an observed S. D. in a sample of n.
154	r_q	the "posterior quartile deviation" of u, the quartile deviation at the section $s=$const. on the posterior probability surface for u and s.
155	$\theta(\mu)$	the ordinate at the abscissa μ on a *prior existence curve* for the mean of the parent population.
155	a, b, c	adjustable parameters in Molina and Wilkinson's forms of prior existence curves for μ and σ.
156	T	denotes $n+2+c+b$ in Molina and Wilkinson's curves.
156	$q(u)$	the ordinate at the abscissa u on the posterior surface for u and s, taken at the section $s=$const.
156	t	the quartile deviation on Student's curve when n is replaced by $T=n+2+c+b$.
156	$r_q(50)$, $r_q(80)$, $r_q(90)$, $r_q(99.73)$	the 50, 80, 90 and 99.73 percentile deviations at the section $s=$const. on the posterior probability surface for u and s. r_q is generally used in place of $r_q(50)$.
158	n_i	the number of observations on the mean μ_i ($i=1, 2, \cdots, m$).

Page	*Symbol*	*Explanation*
158	μ_i	the mean of the parent population from which the n_i observations constitute a sample ($i=1, 2, \cdots, m$). The m means $\mu_1, \mu_2, \cdots, \mu_m$ may or may not all be distinct, but the m parent populations all have the same S.D. σ.
158	m	a finite number of samples or series of observations on means that may all be distinct and in which n may vary from one sample to another, but for which σ is constant.
158	u_i	the error in the mean of the n_i observations on the mean μ_i.
158	s_i	the S.D. of these n_i observations.

§1. Introduction

SOME of the recent advances in probability and mathematical statistics throw considerable light on the theory of errors. Problems that arise in drawing conclusions from observations are essentially statistical and should be handled as such. Unfortunately the literature on statistics has received but scant notice from writers of treatises on errors. In the present paper we shall attempt to put the pertinent results of statistics in such a form that they will be useful for the interpretation of physical data.

Pursuit of the theory of errors is often considered to be futile for the reason that systematic errors, suspected or unsuspected, may be so large as to eclipse any accidental error likely to occur. It is true that a statistical treatment of the data obtained from a single experiment performed under controlled conditions can never disclose the systematic errors in that one experiment. It is only by comparing the results of several observers that it is possible to form some idea as to whether all observers were really measuring the same thing or if, on the contrary, the systematic errors present in one experiment were different from those in the others. Such comparisons are possible only when the data of each observer have been correctly treated, statistically, on the assumption that all systematic corrections have been eliminated. For this reason a working knowledge of the theory of errors is indispensable to the interpretation of experimental data. The detection of systematic errors by statistical analysis has been discussed and applied by one of the writers.[1]

The branch of statistics that concerns the theory of errors is called "sampling" or "the theory of small samples." The object of sampling is to make possible an estimation of the magnitude and variability of some measureable property of a very large number of items by testing only a portion of them. From the measurements of the individuals in a random sample of 5, 10, 20, 30 or more items, and from previous experience with similar items, some estimate of the mean of the measureable magnitude and of its variability in the entire lot can be made by statistical methods of induction. The confidence that one may place in such an estimate depends on the size of the sample and on previous experience with similar items, when such experience is available. Complete confidence or certainty can only be approached as a limit by indefinitely increasing the size of the sample. No guarantee can be made beforehand as to how large the sample must be in order that an estimate shall lie within a specified amount from the true value[2]; however, it may be possible to *lay odds* beforehand that an estimate will fall within the specified range. The theory of sampling furnishes both the methods of estimation and the odds.

A "frequency curve" is a curve so constructed that the area included between two abscissas is equal to the number of items having a measured quality lying within the range defined by these abscissas. Since the area of any strip must be integral and therefore finite, even though the

[1] Raymond T. Birge, Phys. Rev. **40**, 207–227 (1932); **40**, 228–261 (1932).

[2] The reader may consult J. M. Keynes, *A Treatise on Probability*, Ch. 29 (Macmillan, 1921); W. A. Shewhart, *The Economic Control of Quality*, pp. 362, 438 (Van Nostrand, 1931); Thornton C. Fry, *Probability*, Ch. 3 (Van Nostrand, 1928); M. S. Bartlett, Proc. Roy. Soc. **A141**, 518–534 (1933), especially pages 520 and 521.

abscissas differ only infinitesimally, it is clear that the total area under any frequency curve must be infinite and that its actual construction would require an unattainable number of measurements. A frequency *curve* therefore is an attribute of a hypothetical and indefinitely large aggregate, known by the term "parent population." An actual sample, no matter how large, is finite, and therefore will have not a frequency curve but a frequency *polygon*.

As the size of the sample is indefinitely increased and the "class interval" along the abscissa indefinitely decreased, the frequency polygon of the sample approaches the frequency curve of the parent population from which it is drawn. The parent population and its frequency curve have the same objective existence as any statistical limit; hence they can be approached to any desired degree by the two expedients (a) taking a large enough sample, and (b) refining the measurements so that enough figures are recorded for each item to allow a sufficiently small class interval.

In the theory of errors a set of n equally reliable observations may be considered as a sample of n drawn at random from an indefinitely large number ν of observations that *might* be made if time and opportunity would permit and if the apparatus would not wear out. This hypothetical aggregate will be the parent population in the problem.

If there were no systematic errors present, the mean of the parent population would be the true value of the quantity being measured. The effect of a systematic error is to displace the mean of the parent population of observations above or below the true value. This correction, if ever isolated and evaluated, can be added to or subtracted from the mean of the parent population to give the true value.

The object of making the n observations is to *estimate* what would be obtained for the mean of an indefinitely large number of observations; in other words, the object is to estimate the position of the mean of the parent population. Its exact value remains unknown because n is finite. As our hopes vanish of ever knowing exactly the mean of the parent population, we become increasingly interested in the number of significant figures in the estimate. That is, if $\bar{x}$ is an estimate of the mean μ of the parent population, we should like to know what is the chance that $\bar{x}$ differs from μ by a stated amount. On the basis of certain assumptions regarding the form of the parent population, the study of statistics furnishes the answers to this question and to several others that arise.

The true value of the quantity being measured is approached by correcting for systematic errors, one after another. The effect of accidental errors can be reduced as far as desired by taking enough observations. The measurement of each systematic correction presents a problem in statistics, for a correction cannot be intelligently applied unless its precision is stated.

§2. The Specification of the Parent Population

The frequency curve for the parent population will be assumed "normal." There are several reasons for this choice. In the first place, for error theory the normal curve is nearly always an excellent approximation. Furthermore, several investigations on non-normal populations have shown that even considerable departures from normality do not produce appreciable changes in many important deductions based on the normal curve. It has also been established that the frequency curve formed by the means of samples drawn from a non-normal parent population is often much more nearly normal than the population itself. While there exist several types of measurement that by nature do not have normal parent populations, rarely will deductions based on the normal law fail to be valid.

It is therefore idle to investigate whether a parent population is *exactly* normal. However, it may be worth while to discuss some arguments that are commonly advanced as proof that the normal law cannot possibly ever be obeyed. The most incisive arguments run as follows: (a) Since only certain discrete values can be recorded, the probability for all intermediate values is zero. Therefore the law of error cannot be continuous, hence cannot be the normal curve, because of the inherent discontinuous nature of measurement. (b) The frequency polygon of a set of measurements is nearly always skew and irregular, whereas a symmetrical and regular figure should be obtained if the normal law holds. (c) Ex-

tremely large residuals apparently do not occur, whereas according to the normal law they should occur once in a while. When the statistical view is taken and the normal curve becomes a frequency curve for the parent population of observations, the fallacies in these objections become evident, as will now be explained.

The discontinuous nature of measurement has nothing to do with the law of error, which is the specification of the parent population. The step or least count of the instrument, being finite, simply has the effect of grouping the observations into class intervals. Such grouping must always be accomplished before a frequency polygon can be constructed: if the instrument did not attend to this, the computer would have to do it.

It might be expected that the moments of a set of n measurements would vary somewhat as the least count and the zero of the measuring scale are changed, and such is in fact the case. This effect has been carefully investigated by Sheppard,[3] Fisher,[4] and Wilson;[5] and the corrections to be applied to the various moments on account of the finite width of the class interval have properly come to be known as "Sheppard's corrections." These serve to bridge the gap between a continuous law of error and the discontinuous nature of measurement. Such investigations have served to show that the least count of the instrument should be small enough so that when a large number of readings (perhaps a hundred or more) are taken, there will be a variation in the recorded terminal digits of around 20 units, for otherwise a considerable portion of a set of observations is, in effect, scrapped. An astonishingly large number of observations may be required to overcome the damage done by unnecessarily coarse reading or graduation of the scale.

The appearance of a frequency polygon can be very misleading. Even when there are many hundred observations in a set, the appearance of the polygon may be of little value for inferring the law of error. Fortunately the adequacy of a chosen parent population, whatever it may be and however arrived at, can be tested quantitatively and objectively by Karl Pearson's chi-test or criterion for goodness of fit.[6] This test determines the probability that a given set of observations follows the normal law or some other proposed form. The chi-test provides the only decisive criterion, yet it is almost never used by physicists. One good reason is that at least 500 observations are required in order that confidence may be placed in the result.[7] Even when the test shows a small probability that the set of observations came from a normal parent population, conclusions based on the normal law will usually be safe.

If the least count of the instrument were infinitesimal, the normal law would admit the occurrence of a certain small proportion of very large residuals. But in practice the least count is always finite, and this serves to divide the area under the frequency curve into rectangular strips every one having width equal to the least count, and the one of maximum height being centered at the mean of the curve. The readings that can be made on the instrument are the abscissas of the centers of these strips, and if an infinite number of readings were taken, the number recorded of a particular magnitude would be the area of the corresponding strip. Now where the curve approaches the horizontal axis, the areas of the successive strips decrease very rapidly because of the infinitely high order of contact made by the curve. This will especially be true if the graduations on the scale are coarse, for unless the least count is extremely fine there will always be some outlying strip whose area is much greater than all the area lying beyond. The abscissa of the center of this strip will then, in the long run, be

[3] W. F. Sheppard, Proc. London Math. Soc. **29**, 353–380 (1897); J. Roy. Stat. Soc. **60**, 698–703 (1897).

[4] R. A. Fisher, Phil. Trans. Roy. Soc. **A222**, 309–368 (1921–22).

[5] E. B. Wilson, Proc. Nat. Acad. Sci. **13**, 151–156 (1927).

[6] Karl Pearson, Phil. Mag. **50**, 157–175 (1900). This was Pearson's first paper on the chi-test. Tables for using the criterion were computed by W. Palin Elderton, and appeared first in Biometrika **1**, 155–163 (1901–02). These, with additions and examples, are found in *Tables for Statisticians and Biometricians*, Part I, edited by Karl Pearson and published in 1914 by the Biometric Laboratory, University College, London, W. C. 1. Some important discussions of the chi-test are summarized by R. A. Fisher in his *Statistical Methods for Research Workers* (published by Oliver and Boyd, 1925, 4th edition, 1932).

[7] It is interesting to notice the frequency polygon for 500 measurements of a spectral line made by one of us (reference 1, p. 210). The chi-test gives $P=0.22$, which means that in about 1 out of 5 trials we should expect in random sampling a larger χ^2 than that here obtained *if* the real distribution is normal. This probability is not only high, but is a result that could never have been deduced from the mere appearance of the polygon.

recorded more frequently than all the further outlying readings combined, which means that in practice the residuals apparently have an upper limit. Extremely large residuals will occur once in a while, but their frequency is much diminished by the discontinuity of measurement and the shape of the normal curve. The fact that extremely large residuals are seldom found *supports* the normal law and does not subvert it, As was pointed out by Pearson[6] in his original paper on the chi-test, and as has been clearly explained by all later writers on the same subject, it is necessary to lump the tail of a frequency curve into a single "cell"; consequently slight disagreements between calculated and observed frequencies in the tails of the curve are of no concern whatever, either in making an objective test (such as the chi-test) of the fit of the curve or in speculations on the extent to which departures from normality may invalidate deductions that are based on a normal parent population. Thus the last argument is found to be irrelevant.

§3. The Distribution of Certain Properties of Samples Drawn from a Normal Parent Population, and Some Deductions

(3a). The distribution of u, s, and z

The normal curve[8] is fully specified by its mean μ and S.D. σ. If $x_1, x_2, \cdots, x_n$ are n observations of equal reliability and $\bar{x}$ is their arithmetic mean, the *n true errors* are defined as $x_i-\mu\equiv\epsilon_i$ and the *n residuals* as $x_i-\bar{x}\equiv v_i$. By definition, the S.D. of the parent population is σ, where

$$\sigma^2=\sum_\nu(x_i-\mu)^2/\nu=\sum_\nu\epsilon_i^2/\nu. \tag{1}$$

The algebraic form of the normal curve is[9]

$$y\,d\epsilon=[\nu/\sigma\sqrt{(2\pi)}]e^{-\epsilon^2/2\sigma^2}\,d\epsilon. \tag{2}$$

The total area under the curve is ν, the number of observations (and hence errors) in the parent population.

The "probable error" of a single one—any one —of the observations is that *constant* quantity ρ that divides the area of the curve into quarters. It is therefore defined by the equation

$$\int_{-\rho}^{\rho}y\,d\epsilon=\tfrac{1}{2}\int_{-\infty}^{\infty}y\,d\epsilon=\tfrac{1}{2}\nu, \tag{3}$$

wherein y has the value assigned by Eq. (2). The value of ρ is found to be an irrational fractional multiple of σ, namely,

$$\rho=0.6744897502\cdots\sigma. \tag{4}$$

It is an even bet that any one of the ν observations taken at random lies within $\mu\pm\rho$, for half of them lie inside $\mu\pm\rho$ and the other half outside. Curve (a) in Fig. 1 shows a normal frequency curve and the abscissas that divide it symmetrically into quarters.

The division of a symmetrical curve into quarters is called a "quartile" division, and the distance from the center to the dividing lines on either side is known as the "quartile distance." In the normal curve (a) of Fig. 1, the probable

---

[8] The normal curve is sometimes called a Gaussian error curve. It has been attributed to Gauss rather than to Laplace solely because Gauss' *Theoria Motus Corporum Coelestium* appeared in 1809, three years prior to the appearance of Laplace's *Théorie Analitique des Probabilités*. But this was not Laplace's first treatment of the normal curve; in 1774 (Mémoires . . . présentés a l'Académie T. vi, p. 628) he arrived at the normal curve as an approximation to the hypergeometric series, and in 1778 (Mémoire sur les Probabilités) he dealt further with it and emphasized the need of tabulating the normal probability integral. Accordingly Laplace should be credited with the normal curve and its integrals rather than Gauss. However, both men were considerably antedated by Abraham De Moivre, according to evidence presented in a historical note by Karl Pearson, Biometrika **16**, 402–404 (1924). De Moivre arrived at the normal curve and its integrals as approximations to binomial series in about 1721, and printed his findings under the title *Approximatio ad Summam Terminorum Binomii* $(a+b)^n$ *in Seriem expansi*, dated Nov. 12, 1733. This seven page pamphlet was bound into the unsold copies of his *Miscellanea Analytica* as a second supplement. Only two copies of this book complete with the second supplement have been reported extant, but these rare pages have been made generally accessible by a photographic reproduction in a commentary by R. C. Archibald, Isis **8**, 671–683 (1926). De Moivre himself translated the *Approximatio* . . . into English and amplified it for portions of the second and third editions of his *Doctrine of Chances*, published in 1738 and 1756, respectively. This English translation is quoted in full on pages 567–575 of David Eugene Smith's *A Source Book in Mathematics* (McGraw-Hill, 1929). The essential parts of this translation are found on pages 14–17 of Helen M. Walker's *History of Statistical Method* (Williams and Wilkins, Baltimore, 1929).

[9] In this paper, frequency curves will be written in differential form. y will be used indiscriminately for the ordinates of all of them. The differential specifies what sort of frequency curve y is the ordinate of, and the whole expression gives the frequency in the elementary cell. Thus in Eq. (2), $y\,d\epsilon$ is the number of errors in the interval $\epsilon\pm\frac{1}{2}d\epsilon$.

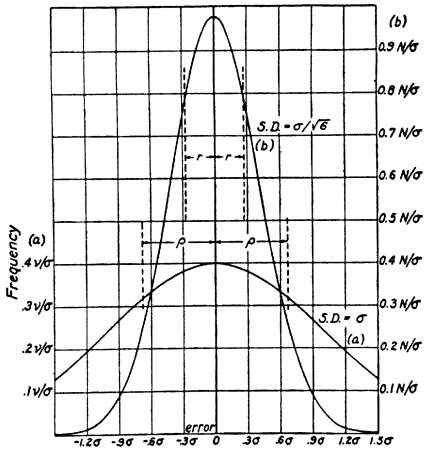


Fig. 1. (a) The normal frequency curve of errors in the parent population; its S.D., or square root of the second moment about the mean, is σ. The area under the curve is the total number of errors, ν. The abscissas $\pm\rho = 0.674\cdots\sigma$ divide the curve symmetrically into quarters. (b) The frequency curve of the errors of the means of N samples of 6 each, drawn at random from the preceding parent population of errors. This curve is also normal, but its S.D. is $\sigma/\sqrt{6}$, hence the abscissas that divide it into quarters are $\pm r = 0.674\cdots\sigma/\sqrt{6}$. The area under the curve is N, the number of samples.

error ρ is therefore the quartile distance of the ν observations from the mean μ.

The S.D. s of the sample of n observations is by definition the r.m.s. residual, so

$$s^2 = \sum_1^n (x_i - \bar{x})^2/n = \sum_1^n v_i^2/n. \tag{5}$$

The true error of the mean of the sample will be

$$u = \bar{x} - \mu. \tag{6}$$

s and $\bar{x}$ can always be computed, but u is unknown as long as μ remains unknown.

Our study of the theory of errors depends mainly on the distribution of u and s in samples of n drawn from the parent population. This was first found by Helmert in three neglected papers that appeared in 1875 and 1876. He found first an expression for the distribution of $\sum_1^n \epsilon_i^2$ in a set of n measurements.[10] The following year, 1876, in discussing the precision of Bessel's correction, he had occasion to find the distribution of s in samples of n.[11]

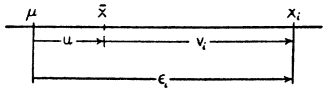


FIG. 2. An observation falls at x_i and the average of a sample of n falls at $\bar{x}$. The figure shows the relations between the error ϵ_i, the residual v_i, and the error u of the sample. Here $\mu < \bar{x} < x_i$, hence ϵ_i, v_i, and u are positive, as the arrows indicate.

Eq. (2) gives the number of errors in the parent population lying in $\epsilon \pm \frac{1}{2}d\epsilon$; whence the probability of the coexistence of n errors lying in the ranges $\epsilon_i \pm \frac{1}{2}d\epsilon_i$ $(i = 1, 2, \cdots, n)$ is

$$[1/\sigma\sqrt{(2\pi)}]^n \exp\left(-\sum_1^n \epsilon_i^2/2\sigma^2\right) d\epsilon_1\, d\epsilon_2 \cdots d\epsilon_n. \tag{7}$$

This can be expressed in terms of u and s by noting the relations between errors and residuals that are exhibited in Fig. 2 and expressed algebraically by

$$\left.\begin{aligned} \epsilon_1 &= v_1 + u \\ \epsilon_2 &= v_2 + u \\ &\cdots \\ \epsilon_{n-1} &= v_{n-1} + u \\ \epsilon_n &= v_n + u = -v_1 - v_2 - \cdots - v_{n-1} + u \end{aligned}\right\}. \tag{8}$$

These follow directly from the definitions. Since the algebraic sum of the residuals is zero, it is evident that

$$\sum_1^n \epsilon_i^2 = \sum_1^n v_i^2 + nu^2 = ns^2 + nu^2. \tag{9}$$

This resembles the formula for the moment of inertia of n points of equal mass about μ. s is the radius of gyration about $\bar{x}$, and u is the distance from μ to $\bar{x}$.

The Jacobian of the transformation (8) is n, so that $d\epsilon_1\, d\epsilon_2 \cdots d\epsilon_n$ becomes $n\, du\, dv_1\, dv_2 \cdots dv_{n-1}$; whence the probability of the coexistence of the n residuals $v_1, v_2, \cdots, v_n$ is

[10] F. R. Helmert, Schlömlich's Zeits. f. Math. und Phys. **20**, 300–303 (1875); ibid. **21**, 192–218 (1876). Helmert's derivation is reproduced in Emanuel Czuber's *Beobachtungsfehler* (Teubner (1891)) on pages 147–150.

[11] F. R. Helmert, Astronomische Nachrichten **88**, No. 2096, 122 (1876). This is given in Czuber's book on pages 159–163. References to Helmert's work are often inaccurately given.

$$y\,du\,dv_1\,dv_2\cdots dv_{n-1}=n[1/\sigma\sqrt{(2\pi)}]^n \exp(-ns^2/2\sigma^2-nu^2/2\sigma^2)du\,dv_1\,dv_2\cdots dv_{n-1}. \quad (10)$$

By a clever transformation, Helmert changed the element of volume from $du\,dv_1\,dv_2\cdots dv_{n-1}$ in the residual space to the element $du\,ds$ in the u,s space. A shorter method than Helmert's is the geometrical one introduced by Karl Pearson,[6] which for brevity we shall follow. Since the integral of the right-hand side of Eq. (10) over all values of $u, v_1, v_2, \cdots, v_{n-1}$ is convergent, integration with respect to $v_1, v_2, \cdots, v_{n-1}$ can be accomplished by using an ellipsoidal shell in the orthogonal $v_1, v_2, \cdots, v_{n-1}$ space in place of the rectangular element $dv_1\,dv_2\cdots dv_{n-1}$. The volume of the thin ellipsoidal shell defined by the two surfaces over which $(v_1^2+v_2^2+\cdots+v_n^2)^{\frac{1}{2}}$ has the pair of constant values $n^{\frac{1}{2}}(s\pm\frac{1}{2}ds)$ is $\left[2\pi^{\frac{1}{2}(n-1)}/\Gamma\left(\frac{n-1}{2}\right)\right]n^{\frac{1}{2}(n-2)}s^{n-2}\,ds$—a result that is known from studies in hyper-space. Now since the right-hand side of Eq. (10) up to the differentials is constant over either surface of the shell that has just been described, it can be integrated throughout this shell simply by replacing $dv_1\,dv_2\cdots dv_{n-1}$ by $\left[2\pi^{\frac{1}{2}(n-1)}/\Gamma\left(\frac{n-1}{2}\right)\right]n^{\frac{1}{2}(n-2)}s^{n-2}\,ds$. Multiplication by N, which denotes an indefinitely large number of samples of n drawn at random from the same parent population, will then give

$$y\,du\,ds=Nn[1/\sigma\sqrt{(2\pi)}]^n\{2\pi^{\frac{1}{2}(n-1)}/\Gamma[\tfrac{1}{2}(n-1)]\}n^{\frac{1}{2}(n-2)}s^{n-2}\exp(-ns^2/2\sigma^2-nu^2/2\sigma^2)du\,ds$$
$$=\left[\frac{N\sqrt{n}}{\sigma\sqrt{(2\pi)}}e^{-nu^2/2\sigma^2}\,du\right]\left[\frac{n^{\frac{1}{2}(n-1)}}{\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma}(s/\sigma)^{n-2}e^{-ns^2/2\sigma^2}\,ds\right] \quad (11)$$

for the frequency distribution of u and s. $y\,du\,ds$ is the number of samples that have S.D. in the interval $s\pm\frac{1}{2}ds$ and means in the interval $u\pm\frac{1}{2}du$ measured from the mean μ of the parent population. n is the number of observations in each sample, and N is the number of samples.[12]

Eq. (11) is a very important one. In the first place, by integrating it with respect to s from 0 to ∞ there results

$$y\,du=[N\sqrt{n}/\sigma\sqrt{(2\pi)}]e^{-nu^2/2\sigma^2}\,du \quad (12)$$

for the number of means having errors in the interval $u\pm\frac{1}{2}du$. Eq. (12) is another normal curve, and its S.D. is $\sigma/\sqrt{n}$. This is an important property of samples from a normal parent population. The probable error r (or the quartile distance) of the mean of n observations is accordingly $\rho/\sqrt{n}$; that is[13]

$$r=0.674\cdots\sigma/\sqrt{n}. \quad (13)$$

A frequency curve for the means of samples of 6 is given as Curve (b) of Fig. 1. The vertical lines with abscissas $\pm r$ divide its area symmetrically into quarters. It is an even bet that the mean of n observations does not differ from the mean of the parent population by more than r.

In the second place, integration of u from $-\infty$ to $+\infty$ in Eq. (11) gives

$$y\,ds=\frac{Nn^{\frac{1}{2}(n-1)}}{\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2}e^{-ns^2/2\sigma^2}ds \quad (14)$$

for the number of samples having S.D. lying in the interval $s\pm\frac{1}{2}ds$ and with $\bar{x}$ lying anywhere. This is equivalent to a result obtained by Helmert[11] in 1876, and for this reason it will be called "Helmert's equation." A graph for $n=6$ is shown as Fig. 3. Karl Pearson[14] has discussed the

[12] Instead of using the actual volume of the ellipsoidal shell, it is perhaps more convenient simply to say that the volume contained between the two ellipsoidal surfaces must be some constant times $s^{n-2}ds$, since it is in a space of $n-1$ dimensions. Then from Eq. (10)

$$y\,du\,ds=\text{const.}\ s^{n-2}\exp(-ns^2/2\sigma^2-nu^2/2\sigma^2)\,du\,ds$$

will be the frequency distribution of u and s if the factor of proportionality is properly chosen. This factor can be found by equating $(1/N)\int_{s=0}^{\infty}\int_{u=-\infty}^{\infty}y\,du\,ds$ to unity; its value so determined and inserted back into the expression for $y\,du\,ds$ gives Eq. (11) immediately.

[13] A table showing the factor $0.674\cdots/\sqrt{n}$ to five figures, for n running from 1 to 1000, was published by Winifred Gibson, Biometrika **4**, 385–393 (1906). This is reproduced as Table V in the *Tables for Statisticians and Biometricians*, Part I. Table 26 in the *Smithsonian Physical Tables* shows $0.674\cdots/\sqrt{(n-1)}$ to four figures up to $n=99$, whence the factor $0.674\cdots/\sqrt{n}$ can be read if one takes care to increase the argument by unity.

[14] Karl Pearson, Biometrika **10**, 522–529 (1915).

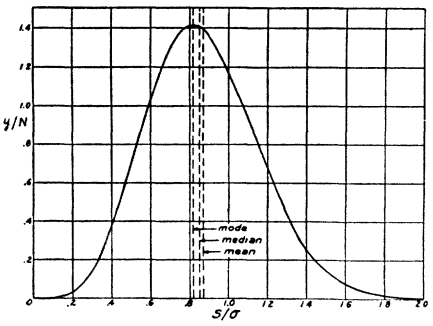


FIG. 3. Frequency distribution of the standard deviation s in samples of 6 from a parent population whose standard deviation is σ.

$$y\,ds=\frac{Nn^{\frac{1}{2}(n-1)}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2}e^{-(n/2)(s/\sigma)^2}\,ds,\qquad n=6$$

N is the number of samples; n is the number in each sample and is here equal to 6. The mode comes at $s/\sigma=0.8165$. The median comes at $s/\sigma=0.8516$. The mean comes at $s/\sigma=0.8686$.

geometry of these curves. They are decidedly skew when n is small, but as n increases they become normal about the point $s=\sigma$ with S.D. $\sigma/\sqrt{(2n)}$, as Pearson showed analytically and as is exhibited graphically in Fig. 4. Each full line curve is the true graph of Helmert's equation, while the corresponding broken line is a normal curve of S.D. $\sigma/\sqrt{(2n)}$ so placed that its center (peak) comes at the mean $\bar{s}$ of the full line curve. The approaching coincidence of the full and broken curves with increasing n shows how Helmert's curves lose their skewness and become normal with S.D. $\sigma/\sqrt{(2n)}$.

The mode (maximum) is at

$$\check{s}=\sigma[(n-2)/n]^{\frac{1}{2}}\rightarrow\sigma(1-1/n-1/2n^2-\cdots). \quad (15)$$

The mean (first moment of area) is at

$$\bar{s}=\int_0^\infty s\,y\,ds\Big/\int_0^\infty y\,ds$$
$$=\sigma(2\pi/n)^{\frac{1}{2}}/B[\tfrac{1}{2}(n-1),\tfrac{1}{2}]\rightarrow\sigma(1-3/4n-7/32n^2-\cdots). \quad (16)$$

The last parenthesis comes from applying the De Moivre—Stirling approximation

$$n!=(2\pi n)^{\frac{1}{2}}(n/e)^n(1+1/12n+1/288n^2-139/51840n^3+\cdots) \quad (17)$$

TABLE I. *The median σ/f of the standard deviation frequency curves.* σ/f is defined by

$$\int_0^{\sigma/f}\frac{n^{\frac{1}{2}(n-1)}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2}e^{-ns^2/2\sigma^2}ds=\tfrac{1}{2} \quad (19)$$

Comparison with the mean and mode.

n	Median σ/f	Mode $\sigma\sqrt{[(n-2)/n]}$	Mean $\sigma(2\pi/n)^{\frac{1}{2}}/B\left(\frac{n-1}{2},\frac{1}{2}\right)$
2	0.476 9363 σ	0	0.564 1896 σ
3	.679 7782	0.577 3503 σ	.723 6012
4	.769 0862	.707 1068	.797 8846
5	.819 3527	.774 5967	.840 7487
6	.851 6120	.816 4966	.868 6267
7	.874 0808	.845 1543	.888 2029
8	.890 6326	.866 0254	.902 7033
9	.903 3347	.881 9171	.913 8749
10	.913 3911	.894 4272	.922 7456
11	.921 5509	.904 5340	.929 9598
12	.928 3048	.912 8709	.935 9418
13	.933 9874	.919 8662	.940 9825
14	.938 8347	.925 8201	.945 2877
15	.943 0191	.930 9493	.949 0076
16	.946 6671	.935 4143	.952 2538
17	.949 8761	.939 3364	.955 1115
18	.952 7207	.942 8090	.957 6464
19	.955 2598	.945 9053	.959 9103
20	.957 5399	.948 6833	.961 9445
21	.959 5989	.951 1897	.963 7823
22	.961 4675	.953 4626	.965 4507
23	.963 1706	.955 5331	.966 9721
24	.964 7297	.957 4271	.968 3652
25	.966 1620	.959 1663	.969 6456
49	.982 8634	.979 3792	.984 6022
75	.988 8337	.986 5766	.989 9609

to the factorials that arise from the beta function.[15]

The median $\hat{s}$ of one of these curves is the abscissa that divides its area into halves. This abscissa will be some multiple of σ, say σ/f, which by definition will satisfy

$$\int_0^{\sigma/f}y\,ds=\tfrac{1}{2}\int_0^\infty y\,ds=\tfrac{1}{2}N, \quad (18)$$

wherein the integrand is given by Eq. (14). The

[15] $$\frac{1}{B\left(\frac{n-1}{2},\frac{1}{2}\right)}=\frac{\Gamma(\frac{1}{2}n)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma(\frac{1}{2})}$$
$$=\frac{1}{\pi}\frac{n-2}{n-3}\frac{n-4}{n-5}\frac{n-6}{n-7}\cdots\frac{6\cdot4\cdot2}{5\cdot3\cdot1}=\frac{2^{n-2}\left(\frac{n-2}{2}!\right)^2}{\pi(n-2)!} \quad n \text{ even}$$
$$=\frac{1}{2}\frac{n-2}{n-3}\frac{n-4}{n-5}\frac{n-6}{n-7}\cdots\frac{5\cdot3\cdot1}{4\cdot2}=\frac{(n-2)!}{2^{n-2}\left(\frac{n-3}{2}!\right)^2} \quad n \text{ odd}$$
These products can be derived from the recursion formula $\Gamma(n+1)=n\,\Gamma(n)$, which leads to
$$\Gamma[\tfrac{1}{2}(n-1)]=\frac{(n-2)!\sqrt{\pi}}{2^{n-2}[\frac{1}{2}(n-2)]!} \quad n \text{ even}$$
$$=[\tfrac{1}{2}(n-3)]! \quad n \text{ odd}$$
since $\Gamma(\frac{1}{2})=\sqrt{\pi}$.

upper limit can be found by inverse interpolation in the *Tables of the Incomplete Gamma Function*.[16] These calculations have been made for us by Lola S. Deming, and are given in Table I, together with the abscissas of the mean and mode. The positions of the mean, median, and mode are shown graphically for $n=6$ in Fig. 3. Clearly, as n increases, the mean, median, and mode all approach the value σ, as is already evident from the discussion of Fig. 4.

In the third place, it is evident that the distributions of u and s in Eq. (11) are completely independent; in any sample u may be large and s small, and conversely. This is the fundamental reason for the difficulties that are encountered in attempting to find the true mean μ and the probable error of $\bar{x}$ when the only information at hand is that provided by the sample itself. These difficulties disappear as n increases, as will be clear from a later section. The independence of u and s is a property peculiar to samples drawn from a normal parent population. This property does not hold for a non-normal distribution.

The fourth result to be derived from the simultaneous distribution of u and s is the distribution of u/s, with s lying anywhere between 0 and ∞. u/s can be thought of as the distance from the mean of the sample to the mean of the parent population measured in terms of the S.D. of the sample. The distribution of u/s was first found by Student[17] in 1908. To accomplish this he needed the distribution of s. Unaware of Helmert's work, Student established the distribution of s beyond reasonable doubt by an ingenious empirical process. Then after proving that there is no correlation between u and s, nor between u^2 and s^2, he assumed that u and s are independent, and proceeded by the following method to find the distribution of u/s.

[16] *Tables of the Incomplete Gamma Function*, edited by Karl Pearson, published by His Majesty's Stationery Office, Imperial House, Kingsway, London W. C. 2. (1922). The incomplete gamma function is defined by the integral

$$\Gamma_x(n)=\int_0^x x^{n-1}e^{-x}dx.$$

In the same symbolism the complete gamma function would be $\Gamma_\infty(n)$, but for brevity and by convention we drop the subscript ∞ and write simply $\Gamma(n)$. The left-hand side of Eq. (18) is $N\,\Gamma_v\left(\frac{n-1}{2}\right)\Big/\Gamma\left(\frac{n-1}{2}\right)$, where $v=n/2f^2$.

[17] Student, Biometrika **6**, 1-25 (1908); **11**, 414-417 (1915-17).

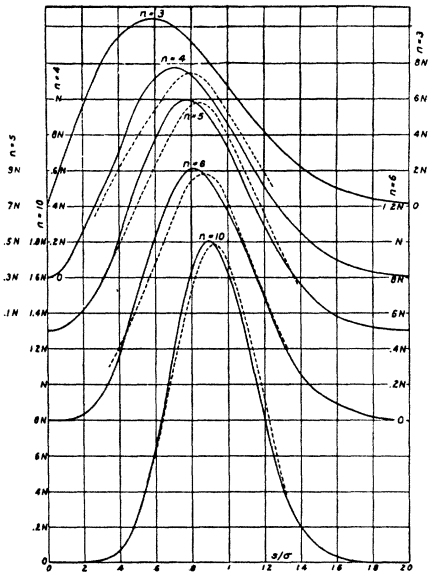


FIG. 4. Frequency distribution of the standard deviation s in samples of n from a parent population whose standard deviation is σ.

$$\text{———}\,y\,ds=\frac{Nn^{\frac{1}{2}(n-1)}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2}e^{-ns^2/2\sigma^2}\,ds$$

$$\text{- - - -}\,y\,ds=\frac{N\sqrt{n}}{\sigma\sqrt{\pi}}e^{-n(s-\bar{s})^2/\sigma^2}\,ds$$

N is the number of samples. $\bar{s}/\sigma$ is the abscissa of the center of area for a particular full line curve. These curves illustrate the mode approaching the mean and the frequency distribution of s becoming normal with standard deviation $\sigma/(2n)^{\frac{1}{2}}$, as n increases.

In Eq. (11) let u/s be replaced by z. Then if s and z be used as orthogonal axes in place of u and s, the elementary volume $y\,du\,ds$ becomes $y\,s\,ds\,dz$, so that the simultaneous distribution of s and z is

$$y\,ds\,dz=\frac{N\,n^{\frac{1}{2}n}}{\sqrt{(2\pi)}\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma^2}\cdot(s/\sigma)^{n-2}\,e^{-(ns^2/2\sigma^2)(1+z^2)}\,s\,ds\,dz. \quad (20)$$

Integration of this with respect to s from 0 to ∞ gives

$$y\,dz=\frac{N}{B[\frac{1}{2}(n-1),\,\frac{1}{2}]}(1+z^2)^{-\frac{1}{2}n}dz \quad (21)$$

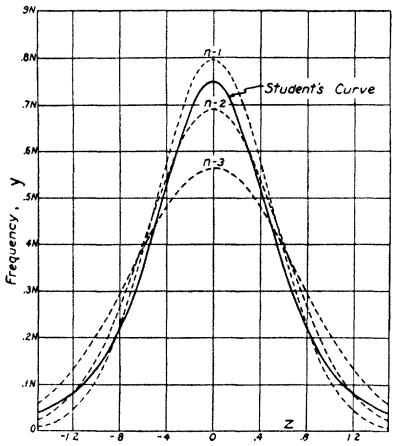


FIG. 5. ————Student's distribution of $z \equiv u/s$,

$$y\,dz = \frac{N}{B\left(\frac{n-1}{2}, \frac{1}{2}\right)} (1+z^2)^{-\frac{1}{2}n}\,dz.$$

– – – – –Normal distribution of S.D. $1/\sqrt{m}$,

$$y\,dz = N\sqrt{(m/2\pi)}e^{-\frac{1}{2}mz^2}\,dz, \quad m = n-1,\ n-2,\ n-3.$$

The curves are plotted for $n=5$. The normal curve of S.D. $1/(n-3/2)^{\frac{1}{2}}$ is not shown because it lies so close to Student's distribution that it would cause confusion. The area under all curves is N, the number of samples. n is the number in each sample. z is the distance from the mean of the sample to the mean of the parent population (the true value), measured in terms of the S.D. s of the sample.

for the number of samples having z in the range $z \pm \frac{1}{2}dz$ and any S.D. s whatever. This is called "Student's distribution." The most important property of this equation is the absence of σ. Student's 1908 paper was a powerful stimulus to the theory of sampling, not alone for the distribution of u/s but for the distribution of s itself, since not until long afterward was Helmert's prior work discovered by statisticians.[18]

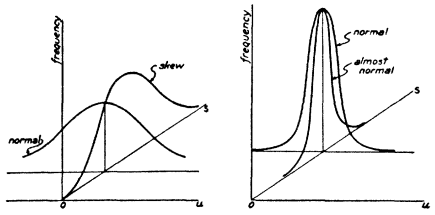
Student's curves are symmetrical in z, as would be expected, since for any value of s, u is as likely to be positive as negative. As n increases they become normal near the center, with S.D. $1/(n-3/2)^{\frac{1}{2}}$. The full line curve in Fig. 5 is Student's distribution for $n=5$. The dashed ones are the normal curves of S.D. $1/(n-1)^{\frac{1}{2}}$, $1/(n-2)^{\frac{1}{2}}$, and $1/(n-3)^{\frac{1}{2}}$, for comparison. The figure shows that a normal curve of S.D. $1/(n-3/2)^{\frac{1}{2}}$ will fall very close to Student's distribution, especially near the center; in fact such a curve could not be shown on the same figure without confusion and so has been omitted. The agreement in the quartile distances of the two curves is shown in Table II, which will be needed later.

[18] Karl Pearson, Biometrika **23**, 416–418 (1931–32).

(3b). The u,s frequency surface

The simultaneous distribution of u and s is important not only for the four conclusions that have already been deduced from it, but also because it is the equation of the "u,s frequency surface"—a surface whose altitude y on the orthogonal axes u and s is given by Eq. (11). The elementary volume $y\,du\,ds$ is the number of samples whose errors fall in the range $u \pm \frac{1}{2}du$ while their standard deviations fall in the range $s \pm \frac{1}{2}ds$; consequently, by integration, the volume erected on any closed figure in the u,s plane is the number of samples whose errors and standard deviations fall simultaneously within the ranges defined by the boundary of the given figure. The total volume under the surface is N, the number of samples. The authors have found this surface to be extremely valuable in describing certain properties of small samples.

Because of the complete independence of u and s, all plane sections $u = \text{const.}$ on this surface will be skew curves similar to the curve defined



n small, about 10 n larger, about 30

FIG. 6. The frequency surface

$$y\,du\,ds = \left[\frac{N\sqrt{n}}{\sigma\sqrt{(2\pi)}} e^{-nu^2/2\sigma^2}\,du\right] \times \left[\frac{n^{\frac{1}{2}(n-1)}}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{1}{2}(n-3)}\sigma} \left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{1}{2}n(s/\sigma)^2} ds\right]$$

illustrated by sections. As n increases, the volume becomes more and more concentrated about the point $u=0$, $s=\sigma$. The total volume is always N, the number of samples. The $s = \text{const.}$ curves are always normal with S.D. $= \sigma/\sqrt{n}$. The $u = \text{const.}$ curves approximate normal curves with S.D. $= \sigma/\sqrt{(2n)}$ as n increases sufficiently.

by Helmert's equation, which has already been discussed. Fig. 3 is then typical of any of these curves. They will all have the same mode, mean, and median that have been found in Eqs. (15), (16), (18), for Helmert's equation. As n increases, the mode, mean, and median approach coincidence with the value σ while the curves lose their skewness and become normal with center at $s=\sigma$ and with S.D. $\sigma/\sqrt{(2n)}$.

The $s=$ const. curves will be normal, all with center at $u=0$ and with S.D. $\sigma/\sqrt{n}$. Clearly, as n increases, the u,s frequency surface becomes more and more concentrated about the point $u=0$, $s=\sigma$. Two u,s frequency surfaces are represented by sections in Fig. 6. The one on the left is for a small value of n and the one on the right is for a comparatively large value of n.

(3c). Tests for hypotheses concerning the parent population

Since the frequency surface for a normal parent population is completely determined when its mean μ and S.D. σ are given, it is sufficient in our problem to state that the object of making n observations is to enable something to be conjectured regarding the mean or the S.D., or both, of the hypothetical indefinitely large number of observations that *might* be taken and from which the n observations constitute a sample. By keeping in mind the u,s frequency surface it is possible to make certain objective statements regarding the parent population from which a sample is drawn.

As long as the parent population remains unknown, the position of a sample in the u,s plane remains unknown so far as its u coordinate is concerned. The S.D. s and the mean $\bar{x}$ can be computed for the sample, but the error $u=\bar{x}-\mu$ obviously cannot be computed, for μ is unknown. Moreover, on account of the independence of u and s, the known value of s gives us no clue regarding the value of u; however, it may help us to lay odds on any specified range within which u might be found.

Since the same sample can come from many sources, the exact parent population cannot be determined from the sample. On the other hand, considerations of the u,s frequency surface are often very helpful in deciding whether a suggested hypothesis regarding the parent population is improbable. To be more specific, there are certain tests which determine the probability that the given sample could have been drawn from a suggested parent population—that is, a parent population having a *proposed* mean and S.D. These various tests will not all give the same answer to the problem, in fact at times they may differ so widely that a suggested hypothesis will be accepted on the basis of one test but rejected by another. Such a situation is, of course, a difficult one, but it is apt to arise when dealing with small samples. The larger n is, the finer will be the distinctions that can be drawn between one hypothesis and another, and the closer will all tests agree. In the limit, as n becomes infinite,

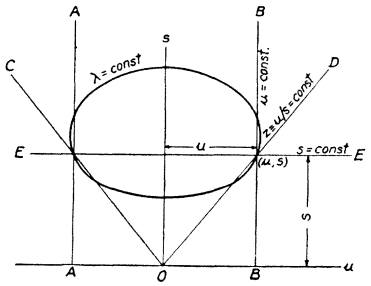


FIG. 7. Contours in the u,s plane. A sample of S.D. s and of error u can be plotted in the u,s plane. The sample point (u,s) lies on the four contours shown:

$|u|=$ const. $\quad |z|\equiv|u/s|=$ const.
$s=$ const. $\quad \lambda=$ const.

the sample becomes identical with the parent population and any proposed hypothesis can be decided with certainty. However, n is for various reasons usually limited to a small integer, and the problem is to learn how much can be safely inferred from such a sample.

By proposing values of μ and σ, a u coordinate for the sample is provided for testing purposes, and the sample may be placed at the point (u,s) in Fig. 7, and certain conclusions drawn. The volume of the u,s frequency surface lying outside any one of certain contours that pass through the point furnishes a test of the hypothesis.

Through the given point in the u,s plane there can be drawn five contours that divide the

volume symmetrically each side of the s axis. They are

$$\pm u = \text{const.}, \tag{22}$$

$$s = \text{const.}, \tag{23}$$

$$\pm z = u/s = \text{const.}, \tag{24}$$

$$\delta = (s/\sigma)^{n-2} \exp\left[-\tfrac{1}{2}n(u^2+s^2)/\sigma^2\right] = \text{const.}, \tag{25}$$

$$\lambda = (s/\sigma)^{n} \exp\left[-\tfrac{1}{2}n(u^2+s^2-\sigma^2)/\sigma^2\right] = \text{const.} \tag{26}$$

The first three are straight lines extending to infinity; the last two are oval closed curves surrounding the highest point ($u=0$, $s=\sigma[(n-2)/n]^{\frac{1}{2}}$ of the volume defined by Eq. (11). Only the z contours are independent of σ.

A certain fraction of the volume lies outside the symmetrically placed u contours AA and BB; this fraction is the probability of drawing a sample of n items having an absolute error in their mean greater than the proposed value of u. This fractional part of the volume can be computed easily from a table of the normal probability integral. Its value is

$$P_u = 2\frac{\sqrt{n}}{\sigma\sqrt{(2\pi)}}\int_u^\infty e^{-nu^2/2\sigma^2}\,du \cdot \frac{n^{\frac{1}{2}(n-1)}}{\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma}\int_0^\infty \left(\frac{s}{\sigma}\right)^{n-2} e^{-ns^2/2\sigma^2}\,ds$$

$$= 2\frac{\sqrt{n}}{\sigma\sqrt{(2\pi)}}\int_u^\infty e^{-nu^2/2\sigma^2}\,du = 1-\sqrt{(2/\pi)}\int_0^{u/(\sigma/\sqrt{n})} e^{-\frac{1}{2}t^2}dt. \tag{27}$$

If P_u turns out to be small, say 0.01, then only once in 100 trials could the mean of the sample be expected to differ so widely from the mean of the proposed parent population; in such a case the hypothesis would immediately be placed under suspicion, but it cannot be definitely rejected until other tests have been made and the circumstances carefully reviewed. On the other hand, if P_u turns out to be fairly large, say 0.2 or higher, then in at least 1 trial out of 5 a greater error would, in the long run, be obtained, and there would be no grounds for rejecting the hypothesis on this criterion. The test just described will be called the "u test."

The upper limit in the last integral of Eq. (27) is the ratio of u to $\sigma/\sqrt{n}$, i.e., the ratio of u to the S.D. of the means of samples of n. In this form the value of P_u is easily found from *Sheppard's Table*.[19] If the form of the integral in Eq. (27) is changed so that

$$P_u = 1-(2/\sqrt{\pi})\int_0^{u/\sqrt{(2\sigma^2/n)}} e^{-t^2}\,dt, \tag{28}$$

the upper limit becomes the argument in various other tables of the normal probability integral.[20] The upper limit could also be made to depend on the ratio u/r with an attending increase in convenience for some problems; thus,

[19] W. F. Sheppard, Biometrika **2**, 174–190 (1902). This table is reproduced as Table II in *Tables for Statisticians and Biometricians*, Part I. The upper limit in the integral of Eq. (27) is Sheppard's x, and our P_u is his $1-\alpha$ or $2[1-\frac{1}{2}(1+\alpha)]$.

[20] The first table of the normal probability integral was computed by M. Kramp and published in his *Analyse des Réfractions*, pp. 195–206 (Strasbourg, 1789). This formed the basis for all tables down to 1898, when James F. Burgess in the Trans. Roy. Soc. Edinburgh **39**, Part II, pp. 257–322 (1898) tabulated the integral in Eq. (28) to 15 decimals, together with first and second differences, the argument being the upper limit of this integral and proceeding in steps of 0.001 from 0 to 1.499 and then in steps of 0.002 from 1.500 to 3. Shorter tables, based on Burgess', are given in B. O. Peirce's *A Short Table of Integrals* (Ginn and Company), in the *Smithsonian Physical Tables* (pp. 56 and 57 of the 7*th* and 8*th* editions), and in many texts on the theory of errors, least squares, and statistics. Notable also is the Kelley-Wood table, Appendix C of Truman L. Kelley's *Statistical Method* (Macmillan, 1924), where the upper limit of the integral in Eq. (28) is tabulated with $\frac{1}{2}(1-P_u)$ as argument in steps of 0.001 from 0 to 0.499. One of the handiest tables for P_u are Tables I and II in R. A. Fisher's *Statistical Methods for Research Workers* (page 79 in the fourth edition), where $u/(\sigma/\sqrt{n})$ is listed to six decimals for values of P_u proceeding in steps of 0.01 from $P_u=0.01$ to $P_u=1.00$, and also for $P_u=10^{-3}$, 10^{-4}, $\cdots$, 10^{-7}. It is interesting to note that the "Diffusion Integral" of Table 31 in the 7*th* and 8*th* editions of the *Smithsonian Physical Tables* is just our P_u.

$$P_u = 1 - 0.674\cdots(2/\pi)^{\frac{1}{2}}\int_0^{u/r} e^{-\frac{1}{2}(0.674\cdots)^2 t^2}\,dt = 1 - (2\gamma/\sqrt{\pi})\int_0^{u/r} e^{-\gamma^2 t^2}\,dt$$

$$= 1 - (2/\sqrt{\pi})\int_0^{\gamma u/r} e^{-t^2}\,dt, \tag{29}$$

wherein $\gamma = 0.674\cdots/\sqrt{2} = 0.476936276\cdots$. The probability integral was first tabled with u/r as argument by Encke,[21] and this form has been adopted by several later writers.

The necessity of using tables of the normal probability integral is to a large measure obviated by Fig. 8, which shows closely enough for most purposes the chance P_u of the occurrence of an error (in the absolute sense) as great as or greater than given multiples both of the S.D. (or r.m.s. error) $\sigma/\sqrt{n}$ and of the probable error $r = 0.674\cdots\sigma/\sqrt{n}$.

Tables of the normal probability integral generally tabulate the *in*ternal portion of the area under the normal curve, that is, the *un*shaded portion of the area in the upper right-hand corner of Fig. 8. This internal area is to be subtracted from the whole area (unity) to obtain the external portion, which we designate by P_u. The reader should note carefully that in the headings of some tables, the letter P is used for the internal portion of the area, and is then just the complement of our P_u.

The other contours in Fig. 7 provide other tests. Thus, the fraction P_s of the volume that lies above the s contour EE is the chance of drawing a sample of n having a S.D. greater than s. This leads to another type of probability integral, the incomplete gamma function, which has been tabled by Karl Pearson and his staff.[16] From Eq. (11) the fraction of the volume above EE in Fig. 7 is

$$P_s = \left[\frac{\sqrt{n}}{\sigma\sqrt{(2\pi)}}\int_{-\infty}^{\infty} e^{-nu^2/2\sigma^2}\,du\right]\left[\frac{n^{\frac{1}{2}(n-1)}}{\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma}\int_s^{\infty}\left(\frac{s}{\sigma}\right)^{n-2} e^{-ns^2/2\sigma^2}\,ds\right]$$

$$= 1 - \frac{n^{\frac{1}{2}(n-1)}}{\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma}\int_0^{s}\left(\frac{s}{\sigma}\right)^{n-2} e^{-ns^2/2\sigma^2}\,ds$$

$$= 1 - \Gamma_v\left(\frac{n-1}{2}\right)\Big/\Gamma\left(\frac{n-1}{2}\right), \tag{30}$$

wherein $v = ns^2/2\sigma^2$, and Γ_v and Γ represent the incomplete and the complete gamma functions.[16] Here it should be noted that the ratio of s to σ is required in order that this integral can be found, but no value of μ is needed. If P_s is small, there is an equally small chance that a sample of S.D. as large as the known s could have been drawn from a parent population having the suggested S.D. σ, and the interpretation is that the hypothesis, as far as σ is concerned, is unlikely. If P_s turns out to be nearly unity, it is practically certain that *if* the suggested σ were the true value, the S.D. of the sample would have been larger than that observed. Hence the suggested value of σ would again appear unlikely. When P_s is anywhere near $\frac{1}{2}$, there is no ground for rejecting the hypothesis on the basis of this criterion. This test will be called the "s test."

Instead of using tables of the incomplete gamma function for calculating P_s, it is usually easier in this work to use tables for the chi-test.[6] In the chi-tables, $P(\chi^2)$ depends on two arguments, χ^2 and the number of "degrees of freedom." P_s will be identical with $P(\chi^2)$ if ns^2/σ^2 replaces χ^2 and if $n-1$ be taken for the number of degrees of freedom. In Elderton's table the number of

[21] Encke, Berliner Astronomisches Jahrbuch für 1834, pp. 249–312 (1832). The tables on these pages are reproduced in Encke's *Astronomische Abhandlungen* Vol. 1, No. 7 (Berlin, 1866). Kramp's tables (see preceding footnote) formed the basis for Encke's computations. Abbreviations of Encke's tables are given in several more recent books, among which are T. W. Wright's *Adjustment of Observations* (Van Nostrand, 1884; revised by J. F. Hayford in 1906), David Brunt's *Combination of Observations* (Cambridge University Press, 1917), W. W. Johnson's *Theory of Errors and Method of Least Squares* (John Wiley, 1912), A. de Forest Palmer's *Theory of Measurements* (McGraw-Hill, 1912), *The Smithsonian Physical Tables*, page 57.

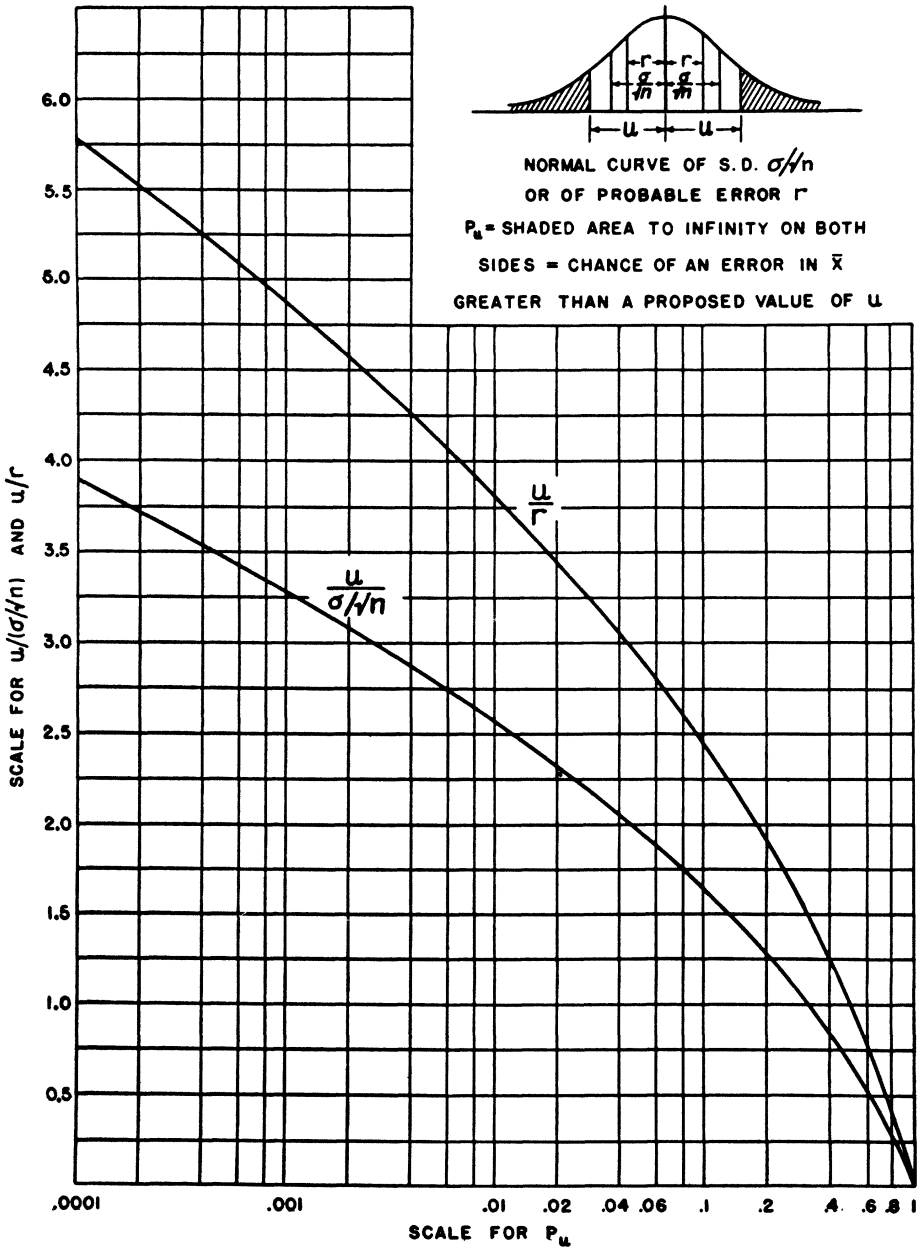


FIG. 8. Chart for making the u-test.

degrees of freedom is denoted by $n'-1$, and in Fisher's table it is denoted by n. The identity of P_s and $P(\chi^2)$ is more than a mathematical coincidence, for it turns out in studying the chi-test that ns^2/σ^2 actually *is* χ^2 for n observations made on a single magnitude—but we cannot pursue the matter further here. Besides the chi-tables, another short cut to calculating P_s is possible when n is around 20 or more, for then the normal curve of S.D. $\sigma/\sqrt{(2n)}$ is a close enough approximation to Helmert's equation near the mode, as was learned from Fig. 4, and a table of the normal probability integral can be used to ascertain whether the sample is unusual. For smaller values of n this approximation may not be close enough.

A third criterion comes from the z contours. CO and OD are drawn making the angles $\pm$ arc tan u/s with the s axis. The fraction P_z of the volume lying outside these contours is the probability of drawing a sample of n having a ratio of u to s greater than the ratio arising from the proposed u and observed s. The calculation of P_z leads to a third type of probability integral, the incomplete beta function; however, the special type here encountered is generally known as "Student's integral," since it is simply an integral of Student's distribution and Student himself prepared rather extensive tables.[17] From Eq. (21)

$$P_z = 1 - \frac{2}{B(\frac{1}{2}(n-1), \frac{1}{2})} \int_0^z (1+z^2)^{-\frac{1}{2}n}\, dz. \quad (31)$$

P_z is the fractional part of the area lying beyond $\pm z$ under Student's distribution of z (Fig. 5), just as P_u is the fractional part of the area lying beyond $\pm u$ under the normal distribution of u, and shown shaded in the upper right-hand corner of Fig. 8. For the calculation of Student's integral it is not necessary to postulate a value of σ, since P_z is simply the probability that a sample of n will fall outside a proposed pair of z contours, and these are independent of σ. Probably the handiest scheme for looking up the value of Student's integral is with the nomograph devised by V. A. Nekrassoff[22] and reproduced as our Fig. 9 with the kind permission of the Bell Telephone Laboratories. The curved portion of the $z\sqrt{(n-1)}$ scale will give better results than the straight portion, which it supersedes over a short range, but both the curved and straight portions will give practically the same results.

The reader familiar with Fisher's methods will realize that the z test here described is equivalent to his t test for the significance of the mean of a single sample.

A very small value of P_z signifies that the sample has an exceptionally small value of z; thus, on the average, only once in 1000 trials will u/s ($\equiv z$) be so large that $P_z = 0.001$. In such cases either the proposed error u is unusually large or else the S.D. s of the sample is accidently very low. Evidently, then, if we reject the idea that the error in $\bar{x}$ is as great as the proposed value of u every time P_z turns out to be small, we shall occasionally reject a perfectly good hypothesis, for not only will the error in the sample actually be large sometimes but also the S.D. s will occasionally be unusually small. When, however, P_z is closer to unity, say 0.2 or greater, the sample is not unusual, and the interpretation is either that u is not exceptionally large or that if it is, then s is also. In such a case it would evidently be unwise to conclude that the error in $\bar{x}$ can easily be as great as the postulated value of u unless there is good reason to believe that the S.D. of the sample is not unusual.

If the S.D. of the sample happens to be exceptional, the u and z tests will give different results regarding the proposed value of u, and it is the latter test that will be misleading. Without even a guess as to where σ lies there is no way of surmising whether s is or is not extraordinary and the z test will accordingly be hazardous when considering the error of the sample. On the other hand, if there is some fairly definite knowledge concerning σ, the u test can be applied; the z test is in this case irrelevant except that it serves as an indication of whether the S.D. of the sample is or is not extraordinary. If the sample is not exceptional, the u and z tests will indicate substantially the same conclusions; and conversely, if the sample is exceptional they will disagree.

This has an important bearing in those problems in physics wherein, having given the mean $\bar{x}$ and the S.D. s of n observations, we seek merely the probability that the error in $\bar{x}$ could

[22] V. A. Nekrassoff, Metron **8**, No. 3, 95–101 (1930).

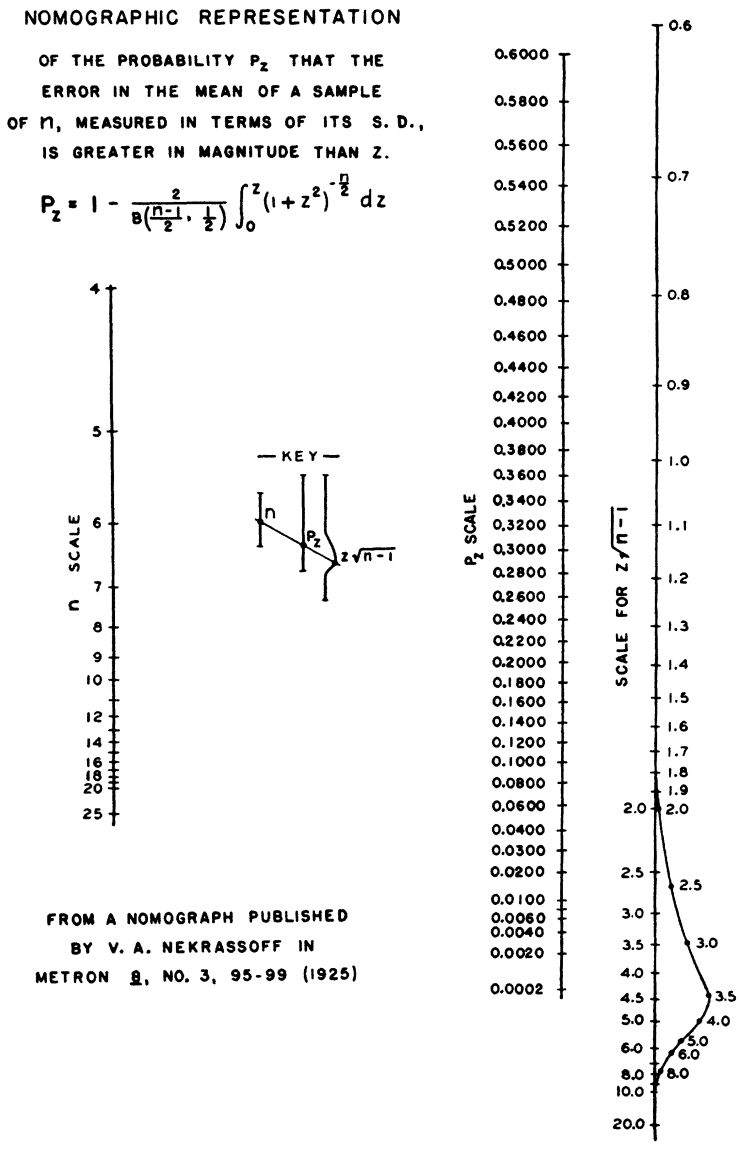


FIG. 9. Chart for making the s-test. Published by permission of the Bell Telephone Laboratories, Inc. (The reference to Metron should be 1930 instead of 1925.)

be as great or greater than a proposed error u; in other words, where we seek the probability that the given sample could have been drawn from some normal parent population having its mean at $\mu(=\bar{x}-u)$ *and having any S.D. whatever*. This is a natural question to ask, especially when there is no information at hand concerning σ. Now since σ is not needed for the z test, it might seem that here is a criterion peculiarly adapted to the problem. Unfortunately, however, this is not so. For although its value may be unknown, nevertheless the parent population does have a certain S.D., and as has just been learned, it is necessary to have some notion what this S.D. is before the proposed error u can be judged with confidence on the basis of the z test. Evidently, then, it is impossible to make any progress without postulating some value of σ, and all conclusions respecting the error, whether drawn from the u or the z test, will depend on this postulate.

The z test simply tells whether the value of u/s obtained in a given sample is extraordinarily large or small, and for *this*, it is, of course, perfectly valid. Usually, however, we are more interested in knowing whether $u \equiv \bar{x}-\mu$ is exceptionally large or small, and the trouble with testing $z=u/s$ is that z expresses u in units of s, which is itself a variable, being subject to the fluctuations of sampling according to Eq. (14).

Careful considerations of the u, s and z tests will generally disclose about all the information concerning the parent population that the sample alone is capable of giving. Any one of the three tests by itself may be misleading, because they all possess an inherent weakness owing to the fact that the contours on which they depend extend to infinity.

An important contribution was made by J. Neyman and Egon S. Pearson[23] when they developed a single test depending on a unique family of *closed* contours for the probability associated with a proposed parent population. They devised for this purpose the λ contours, and the test depending on them will be called the "λ test." Along a λ curve the ratio of the altitude at any point of the u,s frequency surface to the maximum value that it can be made to take (by putting $u=0$ and $\sigma=s$) remains constant. The fraction of the volume under the u,s frequency surface lying outside the λ contour drawn through the point (u,s) is

$$P_\lambda = \frac{n^{\frac{1}{2}n}}{\sqrt{(2\pi)}\Gamma[\frac{1}{2}(n-1)]2^{\frac{1}{2}(n-3)}\sigma^2} \cdot \int\int (s/\sigma)^{n-2} e^{-(n/2\sigma^2)(u^2+s^2)}\, du\, ds, \quad (32)$$

the integral being taken outside the λ curve. Neyman and Pearson published values of P_λ as a function of n, u/σ, and s/σ. By means of their diagram and table the λ test is as easy to apply as any of the others. When P_λ turns out to be small, the hypothesis respecting μ or σ, or both, appears questionable. The diagram published by Neyman and Pearson enables the computer to ascertain at a glance just where the trouble lies when P_λ turns out to be small.

A fifth test is provided by the δ contours of Eq. (25), but the difference between P_δ and P_λ is insignificant, and there is a theoretical reason why the λ contours are better suited to the purpose. The δ contours are curves of equal altitude on the u,s frequency surface, but for small values of n they would not be curves of equal altitude on a u,s^2 or on a u,s^3 frequency surface. But the significance of the λ contours is always the same, regardless of the coordinate system. As n increases, the δ and λ contours approach coincidence; in fact at $n=10$ they are already very close together.

The significance of each test depends not only on the value of P (P_u, P_s, $\cdots$) that is found, but also on how much is known *a priori* regarding the parent population. A hypothesis regarding μ and σ cannot be accepted merely because the tests give high values of P, for it may seem wise to abandon this hypothesis in favor of one that leads to smaller values of P but which is *a priori* more logical or has a more rational basis. For this reason considerable caution must be exercised before *accepting* a hypothesis purely on the basis of any one or all of these tests. High values of P simply show that there are no grounds for rejecting the proposed values of μ and σ on the basis of these criteria alone.

[23] J. Neyman and Egon S. Pearson, Biometrika **20a**, 175–241 (1928). The diagrams and tables published by Neyman and Pearson, together with remarks on their use, will be found in *Tables for Statisticians and Biometricians*, Part II.

On the other hand, a very low value of P does not present such difficulties, for it forces us, regardless of *a priori* considerations, either to admit that the sample is exceptional or to regard the hypothesis with suspicion. Which one of these alternatives is to be chosen will depend for one thing on how compelling were the reasons for selecting the particular hypothesis in the first place. It is thus clear that statistical tests are more readily useful for rejecting a hypothesis than for accepting one.

In rejecting a hypothesis we may reject one that is true: in accepting one we may accept one that is false. The frequency of the former mistake can be controlled to a large extent by lowering the limit for rejection. Thus, if we decide to reject a hypothesis when any $P<0.01$, we shall commit the mistake of rejecting a perfectly good one on the average of once in 100 such tests, but by lowering the rejection limit to 0.001 we lower this average to once in 1000. It is, however, impossible to control so easily the mistake of accepting a hypothesis on the basis of a high value of P when actually it is false, for there will always be false hypotheses that give higher values of P than the true one gives, so that it is impossible to distinguish between the true and the false by objective tests alone. Methods for making quantitative use of other information concerning the hypotheses under test have been devised by J. Neyman and Egon S. Pearson,[23, 24] who have given an excellent discussion of this whole subject.

In some cases it is more important to avoid rejecting a hypothesis that is true than it is to avoid accepting one though it be false; and in other cases just the reverse is true. The seriousness of either mistake depends on the action that is to follow the decision and on the interests involved. A clear illustration of this statement is found in the conflicting interests of producer and consumer in the results of sampling tests on a consignment of goods. For the proposed hypothesis we might say that the consignment which is sampled complies with certain specifications; then a low rejection limit works to the advantage of the producer but to the disadvantage of the consumer, whereas a high rejection limit does just the opposite.

As an example for illustrating the application of the different tests let us consider the following 10 readings made on a micrometer: 1.078, 1.080, 1.071, 1.076, 1.081, 1.077, 1.075, 1.073, 1.079, 1.070. There is reason to suppose that these are of equal reliability, so they will be given equal weight. Their mean is $\bar{x}=1.0760$ and their S.D. $s=0.00355$.[25]

Let us first consider the hypothesis that the sample was drawn from a parent population with true mean 1.0740. If this is the case, the true error of the mean of our sample is $+0.0020$, and we may now ask the question, what is the chance that the true error could be as large as or larger than 0.0020? Without some knowledge concerning σ the only thing we can do is to postulate that the sample was not extraordinary, and apply the z test. If $u=+0.0020$ or greater, then $u/s=+0.0020/0.00355=+0.563$ or greater. Now with $n=10$ and $z=0.563$, Fig. 9 shows that $P_z=0.13$. So in about 1 out of 8 samples of 10, $|u/s|$ will be as large as or larger than 0.563, or in 1 out of 16 samples, u/s will be as large as or larger than $+0.563$. Hence on the assumption that the S.D. of the 10 readings is not unusual, there is no compelling reason to reject the proposal that if the number of measurements were to be indefinitely increased, their mean would finally settle down to the value 1.0740.

Suppose now that there has been some previous work done by the same observer with the same instrument, and there is good reason to believe that σ lies very close to 0.0040. It is clear, without actually calculating P_s, that 0.00355 was in fact not an extraordinary S.D., for the average S.D. in samples of 10 drawn from a normal parent population having $\sigma=0.0040$ is, by

[24] J. Neyman and Egon S. Pearson, Phil. Trans. Roy. Soc. **A231**, 289–337 (1933); Proc. Camb. Phil. Soc. **29**, 492–510 (1933). See also Thornton C. Fry, *Probability and Its Engineering Uses*, pp. 269–270 (Van Nostrand, 1928).

[25] One of the slowest ways to compute $\bar{x}$ and s is to follow their definitions, i.e., take the sum and divide by n, and then find the square root of the average squared residual. Considerable time can be saved by computing $\bar{x}$ and s simultaneously by using the departures from some selected point (instead of from $\bar{x}$), and then applying a correction. In this example 1.075 might be selected as a datum. The departures from this point are 3, 5, -4, 1, 6, 2, 0, -2, 4, -5, all times 10^{-3}. The average of these numbers is $+1.0$, whence $\bar{x}=1.075+0.0010=1.0760$. The sum of their squares is 136; hence, by a well-known formula in mechanics, $s^2=(136/10-1.0^2)\cdot 10^{-6}=12.6\cdot 10^{-6}$ and $s=0.00355$. See Whittaker and Robinson, *The Calculus of Observations*, Art. 96 (Blackie and Sons, 1924 and 1926).

Table I, $0.0040 \cdot 0.9227 = 0.0037$, which is close to 0.0040. So in this case the conclusion indicated by the z test can be accepted with confidence. But if σ is known pretty definitely, the u test is possible. The probable error of the ten measurements is $r = 0.674\cdots\sigma/\sqrt{10} = 0.00085$, so the ratio $u : r = 0.0020 : 0.00085 = 2.34$. The ratio $u : \sigma/\sqrt{n}$ is $0.0020 : 0.0040/\sqrt{10} = \sqrt{(5/2)} = 1.58$. Either of these ratios enables P_u to be read quickly from Fig. 8. The result is $P_u = 0.114$, which means that there is about 1 chance in 9 that $|u| \geqq 0.0020$, or that there is about 1 chance in 18 that $u \geqq +0.0020$. The u and z tests therefore concur, as they will when the S.D. of the sample is not extraordinary.

In the preceding paragraphs we have made the hypothesis that the mean of the parent population is 1.0740 and on the basis of the z and u tests have calculated the chance that a sample of 10 with $\bar{x} = 1.0760$ or greater (i.e., with $u \geqq +0.0020$) could have been drawn from such a parent population. In making the z test it was necessary to assume that the S.D. of the sample was not extraordinary, and in the u test to assume and use some definite value of σ. The reliance that can be placed on the results depends entirely on the validity of these assumptions. When σ is not known very definitely, but reasons exist for thinking that it may be in the neighborhood of (e.g.) 0.0040, we might be interested in the question of what fraction of the samples drawn from a normal parent population with $\mu = 1.0740$ and $\sigma = 0.0040$ would, on the average, lie outside the oval shaped λ contour drawn through the point in the u,s plane corresponding to the 10 observations. With $u = 0.0020$ and $s = 0.00355$ it is found that $P_\lambda = 0.27$, which means that about 3 out of 11 samples will fall outside this λ contour. On the basis of the λ test, then, there is no reason to reject the proposal that $\mu = 1.0740$ and $\sigma = 0.0040$.

For the sake of illustration, it is interesting to assume that $\sigma = 0.0025$ instead of 0.0040. This will reverse some of the previous conclusions. In the first place, the S.D. of the sample now appears to be exceptionally high, for with $v = n s^2/2\sigma^2 = 10(0.00355)^2/2(0.0025)^2 = 10.1$, Eq. (30) gives

$$P_s = 1 - \Gamma_v(9/2)/\Gamma(9/2) = 0.0168,$$

which means that in only about 17 samples out of 1000 could the S.D. be as high as or higher than that found. We may now expect the z and u tests to disagree. The probable error of $\bar{x}$ is now only $(0.674\cdots)(0.0025)/\sqrt{10} = 0.000533$, and the proposed error 0.0020 is accordingly 3.75 times the probable error, for which P_u is 0.0114—just about 1/10 of what it was before. So if $\sigma = 0.0025$, an error as large in magnitude as 0.0020 could occur in only 11 or 12 samples out of 1000, and the proposal that 1.0740 could be the true value should be looked upon with suspicion. Certainly in the face of such odds the proposal could hardly be other than rejected without some very forceful arguments to support it. This conclusion contrasts with that which would be drawn from the z test, for P_z retains its former value, 0.13. The disagreement between the u and z tests shows how misleading the latter would be if used alone. The trouble comes, of course, from the fact that the S.D. of the sample is now exceptionally high.

Finally, we may examine the λ contour on the double assumption that $\mu = 1.0740$ and $\sigma = 0.0025$. In this case P_λ is found to be 0.013, which is so low that the assumption appears improbable. From the position of the sample in the u,s diagram it is evident that the low value of P_λ arises almost solely from the high value of the ratio s/σ.

(3d). Three important relations when $P = \frac{1}{2}$

The u,s, and z tests lead to three important statistical relations. If the straight line contours of Fig. 7 take positions such that the volume under the u,s frequency surface is divided symmetrically into quarters by each of them, it will be an even bet that a random sample will fall inside or outside the u and z contours, and above or below the s contour. Fig. 10 illustrates this situation.

In Fig. 10a, $P_u = \frac{1}{2}$. The lines AA and BB effect quartile divisions of every one of the normal curves obtained by taking sections $s = \text{const.}$ through the u,s frequency surface. The particular constant value of $|u|$ along these lines is therefore $r = 0.674\cdots\sigma/\sqrt{n}$, the probable error of the mean of n observations.

In Fig. 10b, $P_s = \frac{1}{2}$. The line EE divides into halves the area under each of the Helmert's

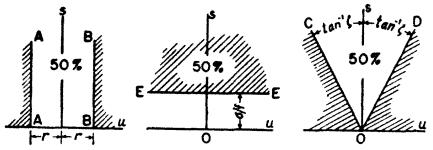


FIG. 10. The volume under the u,s frequency surface

$$y\,du\,ds=\left[\frac{N\sqrt{n}}{\sqrt{(2\pi)}\sigma}\,e^{-nu^2/2\sigma^2}\,du\right]\times\left[\frac{n^{\frac{1}{2}(n-1)}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2}e^{-ns^2/2\sigma^2}\,ds\right] \quad (11)$$

can be divided into quarters in several different ways. Here the division is effected with u, s, and z contours by setting the shaded areas (to infinity) each equal to $\frac{1}{4}$. In (a) the lines AA and BB are a distance r from the s axis, r being the "probable error." r is determined from the normal probability integral by setting $P_u=\frac{1}{2}$. In (b) the line EE divides all the $u=$const. curves into halves and therefore lies the median distance $\dot{s}=\sigma/f$ above the u axis. $1/f$ is determined from the *Tables of the Incomplete Gamma Function* by setting $P_s=\frac{1}{2}$, and its values are given in Table I, column 2. In (c) the lines CO and DO make equal angles with the s axis, this angle being $\tan^{-1}\zeta$. ζ is determined from Student's integral by setting $P_z=\frac{1}{2}$, and its values are given in Table II, column 2. r and $\dot{s}$ depend on σ and n both, whereas ζ depends only on n. $z=u/s$ is constant and equal to ζ along the lines CO and DO.

curves that are obtained by taking a section $s=$const. through the u,s frequency surface. The particular constant value of s along EE is $\dot{s}=\sigma/f$, the median of Helmert's distribution, given by Eq. (18) and Table I, column 2.

In Fig. 10c, $P_z=\frac{1}{2}$. The constant value of $z=u/s$ along the z contours is always the tangent of the angle that these contours make with the s axis. In Fig. 10c, the lines OD and OC effect a quartile division of Student's distribution of z, and the particular constant value of $|z|$ along them is denoted by ζ. Values of ζ for n running from 3 to 25 have been calculated for us by Lola S. Deming and are listed in Table II.[26]

[26] The values of ζ in Table II were calculated by putting $z=\tan\theta$ and then making successive approximations to find the limits of the integral written in the heading of Table II. The same purpose could be accomplished with less precision by inverse interpolation in Student's original tables (see footnote 17) or in later tables by Student and R. A. Fisher, Metron 5, No. 3, pp. 90–120 (1925). Another possibility is inverse interpolation in the *Tables of the Incomplete Beta Function*, recently prepared by Karl Pearson and his staff (issued by the Biometric Laboratory, University College, London, W. C. 1, 1934), but our Table II was calculated and used several years before the appearance of the *Tables of the Incomplete Beta Function*.

TABLE II. *The quartile deviation ζ in Student's distribution.* ζ is defined by

$$\frac{1}{B[\frac{1}{2},\frac{1}{2}(n-1)]}\int_{-\zeta}^{\zeta}(1+z^2)^{-\frac{1}{2}n}dz=\tfrac{1}{2}. \quad (33)$$

Comparison with the normal curve of S.D. $1/(n-3/2)^{\frac{1}{2}}$.

n	ζ	$\frac{0.674\cdots}{\sqrt{(n-3/2)}}$	Discrepancy, percent low
3	0.577 349	0.550 719	4.612
4	.441 614	.426 585	3.403
5	.370 348	.360 530	2.651
6	.324 981	.317 957	2.161
7	.292 942	.287 603	1.822
8	.268 786	.264 557	1.573
9	.249 745	.246 289	1.384
10	.234 241	.231 348	1.235
11	.221 300	.218 833	1.115
12	.210 288	.208 152	1.016
13	.200 768	.198 896	0.9324
14	.192 434	.190 774	0.8621
15	.185 056	.183 573	0.8014
16	.178 467	.177 130	0.7492
17	.172 533	.171 321	0.7025
18	.167 154	.166 048	0.6617
19	.162 249	.161 234	0.6256
20	.157 752	.156 816	0.5930
21	.153 607	.152 742	0.5631
22	.149 774	.148 970	0.5368
23	.146 214	.145 464	0.5129
24	.142 896	.142 195	0.4906
25	.139 794	.139 137	0.4707

As has already been pointed out, σ does not enter Student's distribution of z, hence ζ is independent of σ and depends only on n. Further, since the normal curve of S.D. $1/(n-3/2)^{\frac{1}{2}}$ or of probable error $0.674\cdots/(n-3/2)^{\frac{1}{2}}$ was found to be an excellent approximation to Student's distribution of z near the center, we should expect this last expression to be a good approximation to ζ, provided n is not too small. The actual discrepancy is given in Table II, column 4. In practice, the approximation $\zeta=0.674\cdots/(n-2)^{\frac{1}{2}}$ will be found entirely satisfactory when $n>20$, though of course, $0.674\cdots/(n-3/2)^{\frac{1}{2}}$ is always a better one.

If s be computed for each of an indefinitely large number N of samples, half the values of s will be less than $\dot{s}=\sigma/f$, and the other half will be greater, by definition of the median $\dot{s}=\sigma/f$ in Eq. (18). Clearly, then, if fs be computed for each sample, half the values of fs will be less than σ and half will be greater. Finally, if $0.674\cdots fs/\sqrt{n}$ be computed for each sample, half will be less than r and half will be greater.

It is convenient to denote $0.674\cdots f/\sqrt{n}$ by the symbol ϕ, so that ϕ and ϕs bear the same relation to the probable error r that f and fs do to the S.D. σ. Values of ϕ are given in the second column of Table III. The heading of this column is ϕ_{50}, for reasons that will become clear later.

The preceding discussion shows that the contours in Figs. 10a and 10c correspond to quartile distances r and ζ on the distributions of u and z, respectively, and that the contour in Fig. 10b corresponds to the medians σ/f, σ, and r on the distributions of s, fs, and ϕs, respectively. Hence in the case of a large number of samples of n observations each, it will be found that

(a) in $(\frac{1}{2}\pm\epsilon_1)$ the cases, $|u| \gtrless r$;

(b) in $(\frac{1}{2}\pm\epsilon_2)$ the cases, $\left.\begin{matrix}\phi s \gtrless r \\ fs \gtrless \sigma\end{matrix}\right\}$;

(c) in $(\frac{1}{2}\pm\epsilon_3)$ the cases, $|u/s| \gtrless \zeta$;

wherein ϵ_1, ϵ_2, ϵ_3 approach zero as a statistical limit[2] as the number of samples is indefinitely increased; that is, the odds that ϵ_1, ϵ_2 or ϵ_3 shall differ from zero by less than a stated amount can be made as great as desired by taking enough samples. No one can say in advance just how many samples must be taken in order that ϵ_1 may be less than (e.g.) 0.01, but it is possible to find the *probability* that $\epsilon_1<0.01$ for a given number of samples.

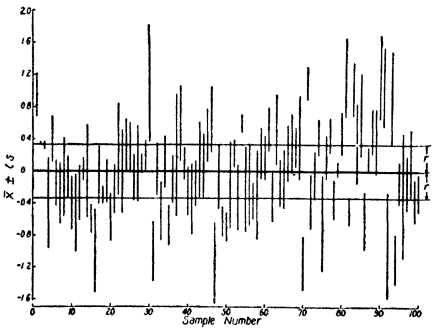


FIG. 11. For each of 100 samples of 4, $\pm\zeta s$ is laid off in the vertical from the point that represents $\bar{x}$. The mean of the parent population is $\mu=0$, and its S.D. is unity. $r=0.674\cdots\sigma/\sqrt{4}=0.337$. The horizontal lines at distances $\pm r$ from the true value show the range covered by the probable error. In 51 out of 100 samples $|u|<\zeta s$. In 52 out of 100 samples $|u|<r$. In 53 out of 100 samples $\phi s<r$. As the number of samples is indefinitely increased, the fractions of them satisfying these three inequalities each approach $\frac{1}{2}$ as a statistical limit.

The relations (a), (b), and (c) just given can be stated still more simply as follows. It is an even bet that for a random sample

(a) $|u|>r$ or $|u|<r$;

(b) $\left.\begin{matrix}fs>\sigma \\ \phi s>r\end{matrix}\right\}$ or $\left.\begin{matrix}fs<\sigma \\ \phi s<r\end{matrix}\right\}$;

(c) $\left.\begin{matrix}|u/s|>\zeta \\ |u|>\zeta s\end{matrix}\right\}$ or $\left.\begin{matrix}|u/s|<\zeta \\ |u|<\zeta s\end{matrix}\right\}$.

The character of each of these quantities, for any given value of n, is worthy of notice. In (a) r is a *constant* while u varies from sample to sample. In (b) σ and r are constants, while fs and ϕs vary from sample to sample. In (c) ζ is a constant while u/s varies, and in the second form, *both* ζs and u vary from sample to sample. These facts and relations are illustrated in Fig. 11, where the value of $\bar{x}$ for each of a number of samples is measured along the vertical and marked by a heavy dot, then the distance ζs for the sample is laid off in the vertical above and below the dot. Thus a vertical line of length $2\zeta s$ with center at $\bar{x}$ marks each sample. In Fig. 11 these lines represent the first 100 samples of 4 drawn from a normal parent population of S.D. $\sigma=1$ and mean $\mu=0$.[27] From Table II, $\zeta=0.4416$ when $n=4$.

It will be noticed that in 51 out of the 100 samples, the range $\pm\zeta s$ measured from $\bar{x}$ overlaps the true value $\mu=0$. For a random sample, there is by relation (c) above an even chance that $|u|<\zeta s$, so we should expect to find approximately half these ranges to overlap $\mu=0$.

A pair of horizontal lines equally spaced at a distance $r=0.674\cdots\sigma/\sqrt{4}=0.337\cdots$ above and below the true value $\mu=0$ show the range covered by the probable error. Before a sample is drawn, it is an even bet by relation (a) above that

[27] These are listed in W. A. Shewhart's book, *The Economic Control of Quality*, Table D, page 454 (Van Nostrand, 1931). The authors are indebted to Dr. Shewhart for the idea of this figure. It was first exhibited by him at a joint meeting of the American Mathematical Society and Section K of the A. A. A. S. in Atlantic City, December 27, 1932.

$|u| < r$, and it is interesting to note that 52 dots fall inside the range $\pm r$ and 48 fall outside. If for each sample in the figure the range $\pm \phi s$ were laid off from the horizontal line $\mu = 0$, it would be found that $\phi s < r$ in 53 samples, $\phi s > r$ in 46 samples, and $\phi s = r$ to three digits in one sample. (To avoid confusion, the ranges $\pm \phi s$ are not shown on the figure.)

Thus the 100 values of $|u|$ and the 100 values of ϕs are separately about equally divided each side of the probable error r. At the same time the 100 values of $|u/s|$ are about equally divided each side of ζ, so that about half the 100 values of $|u|$ are greater than the corresponding values of ζs. If the number of samples were indefinitely increased, the ratio of the number for which $|u| > r$ to the number for which $|u| < r$ would approach unity as a statistical limit, and the same can be said of the other inequalities written in (b) and (c).

(3e). Fiducially related values of σ and s

Closely related to the tests that have previously been described is the notion of *fiducially* related values of σ and s. The adjective *fiducial* was introduced in 1930 by Fisher[28] for the description of a certain objective relation that exists between a parameter of the parent population and the corresponding parameter of a sample when the sampling distribution of the latter depends only on the former. Such is the case with σ and s. Thus, if a set of n observations has been taken and the S.D. is found to be s, we can arbitrarily put $P_s = 0.95$, using the observed value of s for the limit of integration in Eq. (30), and then make the perfectly objective statement that there is only 1 chance in 20 that the S.D. of the parent population can be greater than the value of σ required to be used in the integral. This is the same thing as drawing the s contour of Fig. 7 at a distance from the u axis equal to the observed S.D. s, and then arbitrarily selecting for σ that value which will put 95 percent of the volume of the u,s frequency surface above the contour and the remaining 5 percent below it. These particular values of σ and s are accordingly so related to each other that if σ were actually the S.D. of the parent population then there would be 19 chances in 20 that a sample drawn therefrom would have a S.D. as large as or larger than s; and conversely, since s has actually been observed, there is only 1 chance in 20 that the S.D. of the parent population is as large as or larger than σ.

The value of σ required to be used in the integrals of Eq. (30) will for a given value of n be a function both of P_s and of the limit of integration s, so it seems desirable that the nomenclature for fiducial values should express this fact. If P_s has been placed equal to 0.95, we designate the required value of σ by the symbol $\sigma(s,5)$ and call it "the 5 percent fiducial value of σ corresponding to the given value of s," because there are 5 chances in 100 that the S.D. of the parent population is greater than $\sigma(s,5)$ for the given value of s. Likewise the value of s required to be used as a limit of integration in the same equation will be a function of σ and P_s, so when $P_s = 0.95$ we denote the required value of s by the symbol $s(\sigma,95)$ and call it "the 95 percent fiducial value of s corresponding to the given value of σ," because there are 95 chances in 100 that the S.D. of the sample is greater than $s(\sigma,95)$ for the given value of σ.

Now it so happens that in the incomplete gamma function to which Eq. (30) reduces, s and σ occur only in the ratio $s : \sigma$. This ratio will of course be a function of P_s for a given value of n. If, then, for $P_s = 0.95$ this ratio be denoted by $1/f_{95}$, Eq. (30) gives

$$\frac{1}{\Gamma[\frac{1}{2}(n-1)]}\int_0^{n/2f^2_{95}} x^{\frac{1}{2}(n-3)}\, e^{-x}dx$$

$$= \Gamma_{n/2f^2_{95}}[\tfrac{1}{2}(n-1)]/\Gamma[\tfrac{1}{2}(n-1)] = 0.05, \quad (34)$$

from which the numerical evaluation of f_{95} for different values of n can be accomplished. When $n < 9$, the most satisfactory method seems to be to integrate in series, retaining enough terms to give the accuracy desired, and then to solve for $n/2f^2_{95}$ by any scheme that happens to be suitable for finding the numerical roots of the resulting algebraic equation. When $n \gtrless 9$, interpolation in the *Tables of the Incomplete Gamma Function*[16] by means of a central difference formula will give 7 place accuracy. Values of f_{95} obtained by a combination of these methods are shown in Table III for n running from 2 to 25.

[28] R. A. Fisher, Proc. Camb. Phil. Soc. **26**, 528–535 (1930).

These values of f_{95} provide the reciprocal relation between the S.D. of the parent population and the S.D. of the sample that has been described above: when a sample of n shows a S.D. s, there is only 1 chance in 20 that the S.D. of the parent population whence it came can be greater then $f_{95}s$, and conversely, if the S.D. of a parent population is σ, there are 19 chances out of 20 that the S.D. of a sample of n drawn therefrom will be greater than σ/f_{95}.

The notion of fiducial values can easily be extended to the probable error of the mean of n observations, for if there is only 1 chance in 20 that the S.D. of the parent population is greater than $f_{95}s$, there is the same chance that the probable error of the mean of n observations is greater than

$$r(s,5) = 0.674\cdots f_{95}s/\sqrt{n} \equiv \phi_{95}s, \qquad (35)$$

which may accordingly be termed "the 5 percent fiducial value of r corresponding to the given S.D. s." The factor $0.674\cdots f_{95}/\sqrt{n}$ is denoted by ϕ_{95}, as just indicated, and its values for n between 2 and 25 are listed alongside the values of f_{95} in Table III. The factor ϕ_{95} gives a very useful relation, because although the value of σ, and hence that of r, may be unknown, we can be "19/20 sure" that r is not greater than $r(s,5)$ as calculated in the last equation. Thus, to go back to the 10 observations previously under consideration, since their S.D. is 0.00355 there is only 1 chance in 20 that the probable error of their mean $\bar{x}$ is greater than $0.3699\times 0.00355 = 0.00131$. The values of ϕ_{95} in Table III make the calculation of 5 percent fiducial values of r a very simple matter.

With the notation here introduced, extension to other fiducial points can be conveniently accomplished. Thus with some value of P_s other than 0.95, the subscripts for f and ϕ can be changed to the new percentage; likewise $\sigma(s,5)$, $s(\sigma,95)$, $r(s,5)$ can be rewritten to correspond with the new value of P_s. In particular, the 50 percent point is of special interest, for it corresponds to the *medium* of Helmert's distribution of s, as is evident from a comparison of Eqs. (19) and (30). The values of $1/f_{50}$ are accordingly just those ratios of s/σ that were labeled $1/f$ in Eqs. (18), (19), and Table I, and the corresponding factors $\phi_{50} = 0.674\cdots f_{50}/\sqrt{n}$ are just those that were denoted by ϕ in the preceding section. The median of Helmert's curves is so frequently used that for brevity and convenience the subscript 50 will ordinarily be omitted, so that except when emphasis is desired, f_{50} and ϕ_{50} will appear simply as f and ϕ.

Values of f_{50} for n between 2 and 25 are shown in the second column of Table III; these are, of course, simply the reciprocals of $1/f$ in Table I. Alongside these are shown the factors $\phi_{50} = 0.674\cdots f/\sqrt{n}$. (Later on, ϕs will have still another significance, and it will be convenient to have ϕ_{50} retabulated in Table IV for comparison with two other functions yet to be introduced.)

When the S.D. of a sample of n turns out to be s, there is an even chance that $\sigma \gtrless fs$, and

TABLE III. *Fiducial values of σ and s.* Multiplying factors for getting the 5 and 50 percent fiducial values of σ, and the 5 and 50 percent fiducial values of the probable error r, corresponding to a given S.D. s in a sample of n. f_{95} is defined as the ratio of the 5 percent fiducial value of σ to the observed value of s. f_{95} is obtained by setting $P_s = 0.95$, whereupon Eq. (30) gives

$$\frac{1}{\Gamma[\frac{1}{2}(n-1)]}\int_0^{n/2f_{95}^2} x^{\frac{1}{2}(n-3)}e^{-x}dx = 0.05. \qquad (34)$$

The 5 percent fiducial value of σ is $f_{95}s$, and the 5 percent fiducial value of r is accordingly

$$r(s,5) = 0.674\cdots f_{95}s/\sqrt{n} = \phi_{95}s. \qquad (35)$$

The odds are 19 : 1 that r is not greater than $\phi_{95}s$. f_{50} and ϕ_{50} (or simply f and ϕ) are defined in a similar manner by setting $P_s = 0.50$. $1/f_{50}$ is then just the median value of s/σ, and has already been given in Table I. The odds are even that r is not greater than $r(s,50) = \phi_{50}s$, which is the 50 percent fiducial value corresponding to the given S.D. s.

n	f_{50}	$\phi_{50} = 0.674\ldots f_{50}/\sqrt{n}$	f_{95}	$\phi_{95} = 0.674\ldots f_{95}/\sqrt{n}$
2	2.096 716	1	22.552 803	10.756 2497
3	1.471 068	0.572 8587	5.353 057	2.084 5706
4	1.300 244	.438 5007	3.371 735	1.137 1005
5	1.220 476	.368 1455	2.652 372	0.800 0640
6	1.174 244	.323 3389	2.288 667	0.630 2057
7	1.144 059	.291 6586	2.068 899	.527 4310
8	1.122 797	.267 7514	1.921 235	.458 1533
9	1.107 009	.248 8888	1.815 807	.408 0230
10	1.094 821	.233 5170	1.734 191	.369 8896
11	1.085 127	.220 6783	1.670 828	.339 7901
12	1.077 232	.209 7462	1.619 586	.315 3470
13	1.070 678	.200 2916	1.577 196	.295 0457
14	1.065 150	.192 0093	1.541 478	.277 8745
15	1.060 424	.184 6755	1.510 922	.263 1308
16	1.056 338	.178 1222	1.484 443	.250 3104
17	1.052 769	.172 2201	1.461 245	.239 0418
18	1.049 628	.166 8682	1.440 730	.229 0455
19	1.046 836	.161 9858	1.422 439	.220 1062
20	1.044 313	.157 5083	1.406 011	.212 0553
21	1.042 102	.153 3825	1.391 165	.204 7596
22	1.040 077	.149 5648	1.377 670	.198 1113
23	1.038 238	.146 0186	1.365 341	.192 0227
24	1.036 560	.142 7132	1.354 027	.186 4220
25	1.035 023	.139 6225	1.343 599	.181 2487

that $r \gtrless \phi s$. This statement is only a repetition of relation (b) in the previous section, but it is now seen that the even odds that were obtained by placing the line *EE* in Fig. 10 at the *median* is only one of an infinite set of odds that can be laid on pairs of values of σ and s through the fiducial relation. In practice, it has been found that the odds 1 : 1 and 19 : 1, given by f_{50} and f_{95}, will yield sufficient information. Thus, although we may know nothing beforehand concerning σ, we can by a glance at Table III say that the probable error of 20 observations is as likely as not to be greater than $\phi s = 0.158s$, but that there is only 1 chance in 20 that it is greater than $\phi_{95} s = 0.212\ s$. Since these two multiples of s are so close together, we can be fairly confident that the probable error of the 20 observations is in the neighborhood of ϕs. On the other hand, while the probable error of 3 observations is as likely as not greater than $0.573\ s$, it has 1 chance in 20 of being greater than $5.40\ s$. On account of the disparity between these last two multiples of s (they differ by almost ten-fold), we should be extremely cautious about assigning any value to the probable error of 3 observations, on the basis of their S.D. alone.

It is interesting to note from Table III that the values of f_{50} and f_{95} are widely different when n is small, but that they both approach unity monotonically and are not so greatly different toward the end of the table. The approaching coincidence of f_{50} and f_{95} is, of course, brought about by the tendency of the Helmert curves to become more and more concentrated about the abscissa $s/\sigma = 1$ as n increases, as is illustrated by the curves in Fig. 4. This shows that as n increases, the fluctuations in s are confined more and more to a narrow band about σ.

§4. The Estimation of the Probable Error

(4a). Introduction

R. A. Fisher[4] has divided the problems of statistics into three classes: (a) the *specification* of the form of the frequency curve of the parent population, and of the necessary parameters; (b) the *distribution* of various properties (means, errors, standard deviations, etc.) of samples drawn from a given parent population; (c) the *estimation* of the parameters of the parent population from information provided, at least in part, by the sample. The first and second class can be handled independently of the third, but the third is intimately related to the others. In this treatment of the theory of errors, the problem of *specification* was disposed of by making the *assumption* that the parent population of observations is normal. The simultaneous *distribution* of errors and standard deviations in samples was then found, and certain deductions were drawn from it.

These deductions are most conveniently expressed in terms of the u, s, z, and λ tests, and by means of the fiducial relation between σ and s, which have been described in the preceding sections. These tests lead to statements such as the following, "*If* the S.D. of the parent population is σ, then there is not more than one chance in 100 that the error in $\bar{x}$ could be as large as the proposed value of u," or "It is an even bet that the error in $\bar{x}$ is not more than ζs." Such statements are entirely objective, and involve none of the risks of estimation. These tests make no pretense of estimating σ; the u test, for example, though it depends on σ, simply finds the odds against the occurrence of an error as large as or larger than the proposed error, and the odds so found will of course vary as σ varies.

The parent population of observations is, by assumption, normal, and is therefore completely specified by the three parameters ν, μ and σ. When a set of n observations is taken, their mean $\bar{x}$ differs from μ by an unknown error u. Odds against the occurrence of an error as large as or larger than a given magnitude can be found by the u test, but, as has been noted, the results of this test depend on the value of σ chosen for the purpose. Clearly, then, it is desirable to use a value of σ that is as close as possible to the actual S.D. of the parent population. It is the purpose of any process of estimation to provide a value of σ that will make the u test valid, or, what is the same thing, to provide an estimate of the probable error of $\bar{x}$.

The problem of *estimation* has necessarily been deferred to the last, since it is a process of attempting to reckon from the sample back to the parent population, and therefore depends on the distribution of u and s. It is a problem that involves all the entanglements of induction.

There are three methods of attempting to say something about the parent population—maximum likelihood, empirical or arbitrary schemes, and the posterior method. The first two disregard all prior knowledge and base the estimate purely on the sample. The last one utilizes the methods of Bayes and Laplace to combine previous experience or knowledge with the information contained in the sample. As the size of the sample increases, the results provided by all these methods become indistinguishable. The three methods will be treated here in the order named.

(4b). Maximum likelihood

It is evident from Helmert's Eq. (14) and the curve for $n=6$ in Fig. 3 that when a sample of n is drawn from a parent population having a certain S.D. σ, the S.D. s of the sample may lie anywhere between 0 and ∞, whether σ be large or small.[28] It is further evident that a sample of S.D. s *may* have come from any one of an infinite number of parent populations. Out of this infinity of parent populations there is a particular one that is *most favorable* to the given sample; that is, there is a particular one for which the probability of drawing a sample of S.D. $s\pm\frac{1}{2}ds$ is greater than for any other parent population. To arrive at this particular one, Helmert[11] simply found the value of σ that makes y in Eq. (14) a maximum for the given value of s by setting $dy/d\sigma=0$. The necessary relation between s and σ is easily found to be

$$\sigma=s[n/(n-1)]^{\frac{1}{2}}. \tag{36}$$

This value of σ, which will be called σ_s, may be adopted as an *estimate* of the unknown S.D. σ of the parent population. When it is substituted into Eq. (13) and used with the definition of s in Eq. (5), it gives

$$r_s=0.674\cdots s/(n-1)^{\frac{1}{2}}$$
$$=0.674\cdots[\sum v^2/n(n-1)]^{\frac{1}{2}} \tag{37}$$

for an estimate of the probable error r of n equally reliable observations. The subscript s, attached to any quantity such as σ or r, signifies that the quantity is an *estimate* derived from the *sample alone*. The factor $0.674\cdots/(n-1)^{\frac{1}{2}}$ is tabulated in the second column of Table IV for n between 2 and 25.[13]

Eq. (37) is a familiar formula. In textbooks it is usually called the "formula for the probable error," but it should be carefully noted that this is a misnomer; r_s is *not* the probable error r of $\bar{x}$, it is an *estimate* of r, and only one of many possible estimates. Failure to realize this is very likely responsible for the disrepute of "probable error" in some quarters. Just as $\bar{x}$ is an estimate of μ, and is subject to statistical fluctuations for which r is a convenient measure, so r_s is an estimate of r, and is similarly subject to statistical fluctuations the measure of which will be described presently. When n is small these fluctuations are serious. As n increases, they become less and less bothersome, for we have seen from the curves of Fig. 4 that as n increases s becomes more and more restricted to the neighborhood of σ, so that the estimate r_s becomes more and more restricted to the true probable error r.

The introduction of the factor $[n/(n-1)]^{\frac{1}{2}}$ in Eq. (37) is called "Bessel's correction," since it seems to have been first used by Bessel. The history of just how and when he derived it is at present obscure. The process that Helmert used in deriving Bessel's correction has been named by R. A. Fisher[4, 30, 31] the "method of maximum likelihood," and the estimate so obtained the "optimum value"; Eq. (37) then gives the "*optimum* estimate of r." Another interpretation of the relation between s and σ in Eq. (36) will be given in the derivation of Eq. (42).

(4c). Empirical estimates

There are other methods of attempting to reckon from the sample alone what the S.D. of the parent population actually was. One might arbitrarily assume that the observed S.D. s of the sample is the average of all those that would be observed if a very large number of samples were to be drawn. Geometrically this is equivalent to placing the observed value s at the *mean* $\bar{s}$ of the S.D. frequency curve (Eq. (14) and Fig. 3). If this is done, the estimate of σ is, by Eq. (16),

[28] Since the least count of any measuring instrument must be finite, the S.D. of a sample will in practice have an upper limit.

[30] R. A. Fisher, Messenger of Mathematics **41**, 155–160 (1912).

[31] R. A. Fisher, Proc. Camb. Phil. Soc. **22**, 700–725 (1925); **26**, 528–535 (1930); **28**, 257–261 (1932).

$$\sigma_s = s\sqrt{(n/2\pi)}\, B(\tfrac{1}{2}(n-1), \tfrac{1}{2}) \rightarrow s/(1-3/4n-7/32n^2-\cdots) \quad (38)$$

and by introducing this into Eq. (13), the corresponding estimate of the probable error r is

$$r_s = 0.674\cdots s\sqrt{(1/2\pi)}\, B(\tfrac{1}{2}(n-1), \tfrac{1}{2}). \quad (39)$$

We shall call this the "*mean* estimate of r." The factors multiplying s have been worked out by Lola S. Deming for n running from 2 to 25 and are shown in the third column of Table IV.

Another possibility is to assume that if more samples were to be drawn, as many would be found with S.D. $> s$ as have S.D. $< s$. Geometrically this is the same thing as arbitrarily placing the observed S.D. at the *median* $\dot{s} = \sigma/f$ of the S.D. frequency curve; hence this estimate of σ is fs, and by Eq. (13) it leads to

$$r_s = 0.674\cdots fs/\sqrt{n} = \phi s \quad (40)$$

for the corresponding estimate of r. We call this the "*median* estimate of r." It is identical with the 50 percent fiducial value of r. It will be recalled that in the discussion of the median, f was defined by Eq. (18), and that values of $1/f$ and f (or f_{50}) have been shown in Tables I and III respectively. The factors $\phi = 0.674\cdots f/\sqrt{n}$ have also been given in Table III, in the column headed ϕ_{50}. For ready comparison between *median*, *optimum*, and *mean* estimates, ϕ is again listed in Table IV.

There are other possibilities without number. Only two more will be mentioned. One is to place the observed S.D. at the mode (maximum) of the S.D. frequency curve; this leads to

$$r_s = 0.674\cdots s/(n-2)^{\frac{1}{2}}, \quad (41)$$

which may be called the "*modal* estimate of r."

Another is to assume that the observed s^2 is the mean square of all the standard deviations that would be obtained from a very large number of samples. It is a simple matter to prove by Helmert's equation that the mean square of the standard deviation in a very large number of samples is $\sigma^2(n-1)/n$. Thus, using Helmert's Eq. (14),

$$\bar{s^2} = \int_0^\infty s^2(s/\sigma)^{n-2}\exp(-ns^2/2\sigma^2)ds \Big/ \int_0^\infty (s/\sigma)^{n-2}\exp(-ns^2/2\sigma^2)ds$$

$$= \sigma^2(2/n)\Gamma\left(\frac{n+1}{2}\right)\Big/\Gamma\left(\frac{n-1}{2}\right) = \sigma^2(n-1)/n. \quad \cdots(42)$$

This scheme of estimating σ brings in the factor $(n-1)/n$ and therefore leads to none other than the *optimum* value, and the corresponding estimate of r is identical with Eq. (37).

It is not necessary to know the distribution of standard deviations in samples in order to find the mean square standard deviation; it can be found by writing Eq. (9) for each of a large number N of samples of n items each, and adding the N equations so obtained. This procedure gives

$$\sum_1^N\sum_1^n \epsilon_i^2 = n\sum_1^N s^2 + n\sum_1^N u^2;$$

$$(1/Nn)\sum_1^N\sum_1^n \epsilon_i^2 = (1/N)\sum_1^N s^2 + (1/N)\sum_1^N u^2.$$

The left-hand side of the last equation is the mean square (true) error in N samples, and is therefore

TABLE IV. *Factors that multiply s to get various estimates of the probable error r of n observations.*

The "optimum estimate,"

$$r_s = 0.674\cdots s/\sqrt{(n-1)}. \quad (37)$$

The "mean estimate,"

$$r_s = 0.674\cdots s/\sqrt{(1/2\pi)}\, B(\tfrac{1}{2}(n-1), \tfrac{1}{2}). \quad (39)$$

The "median estimate,"

$$r_s = 0.674\cdots fs/\sqrt{n} = \phi s. \quad (40)$$

n	$0.674.../\sqrt{(n-1)}$	$0.674...\sqrt{(1/2\pi)} \times B(\frac{1}{2}(n-1), \frac{1}{2})$	$\phi = 0.674.../\sqrt{n}$
2	0.674 4898	0.845 3475	1
3	.476 9363	.538 1650	0.572 8587
4	.389 4168	.422 6738	.438 5007
5	.337 2449	.358 7766	.368 1455
6	.301 6410	.317 0053	.323 3389
7	.275 3593	.287 0213	.291 6586
8	.254 9332	.264 1711	.267 7514
9	.238 4681	.246 0183	.248 8888
10	.224 8299	.231 1497	.233 5170
11	.213 2924	.218 6829	.220 6783
12	.203 3663	.208 0348	.209 7462
13	.194 7084	.198 8026	.200 2916
14	.187 0698	.190 6985	.192 0093
15	.180 2650	.183 5101	.184 6755
16	.174 1525	.177 0772	.178 1222
17	.168 6224	.171 2761	.172 2201
18	.163 5878	.166 0099	.166 8682
19	.158 9788	.161 2011	.161 9858
20	.154 7386	.156 7871	.157 5083
21	.150 8205	.152 7168	.153 3825
22	.147 1857	.148 9477	.149 5648
23	.143 8017	.145 4446	.146 0186
24	.140 6408	.142 1774	.142 7132
25	.137 6796	.139 1209	.139 6225

σ^2. The first term on the right is $\overline{s^2}$, the mean square value of s, and the last term is the mean square error of the means of the N samples, which by Eq. (12) is simply σ^2/n. So we have

$$\sigma^2 = \overline{s^2} + \sigma^2/n$$

or

$$\overline{s^2} = \sigma^2(n-1)/n,$$

as before. This is the method adopted in some textbooks for the derivation of Eq. (37). Usually, however, the texts forget to warn the reader that it is only the *average* value of s^2 that is equal to $\sigma^2(n-1)/n$; the S.D. s of any one sample may give an estimate that differs considerably from the true value. Further, the texts usually do not mention the fact that this is only one of many possible methods of estimating r.

It is evident from a comparison of the columns of Tables I and IV that the *optimum*, *mean*, *median* and *modal* estimates are approaching coincidence as n increases. Table I and Fig. 4 have already shown that the mean, median, and mode on Helmert's curve approach σ as n increases, and that the values of s become restricted practically to a very small range near the abscissa $s=\sigma$. So when n is very large, σ may be equated to $s[n/(n-1)]^{\frac{1}{2}}$, or, closely enough, simply to s, with considerable confidence.

(4d). Fluctuations in estimates. The r.m.s. error in an estimate of r. Significant figures

Estimates of r made by maximum likelihood or any empirical method are subject to the statistical fluctuations of sampling. Just as there is no way of judging how much significance dare be attached to the mean $\bar{x}$ of a sample without knowing the r.m.s. fluctuation $\sigma/\sqrt{n}$ (i.e., the S.D.) of the means of such samples—or what amounts to the same thing, their probable error r—so there is no way of judging the significance of an estimate of σ or of r without having some idea of the r.m.s. fluctuation of such estimates. It is therefore desirable to study the precision of estimates of the probable error.

Every method for estimating σ from the sample alone places

$$\sigma_s = \omega s \tag{43}$$

where ω is some function of n that approaches unity as n increases. This proposed relation between s and σ gives, from Eq. (13),

$$r_s = 0.674\cdots\omega s/\sqrt{n} = r\omega s/\sigma \tag{44}$$

for the corresponding estimate of r. The error in writing r_s for r is

$$r_s - r = (0.674\cdots/\sqrt{n})(\omega s - \sigma). \tag{45}$$

If this were written for an indefinitely large number of samples, the mean square error committed would be the average of $(r_s-r)^2$ taken over all samples. Now $y\,ds$ in Eq. (14) is the number of samples having S.D. $s\pm\frac{1}{2}\,ds$, so with this it is easy to write down the contribution from each interval ds between $s=0$ and $s=\infty$ toward the sum of $(r_s-r)^2$. The sum of all these contributions divided by N is the desired average of $(r_s-r)^2$; wherefore

$$\overline{(r_s-r)^2} = (1/N)(0.674\cdots/\sqrt{n})^2 \int_0^\infty (\omega s-\sigma)^2 y\,ds$$

$$= (1/N)(0.674\cdots\sigma/\sqrt{n})^2 \int_0^\infty (1-2\omega s/\sigma+\omega^2 s^2/\sigma^2) y\,ds.$$

The three integrations that arise from the three terms in the parenthesis correspond, save for constant factors, to the integrations that would be used for computing the zero, first, and second moments of the area under Helmert's curve, Eq. (14), all of which have been found. The zero moment is of course unity; the first moment or mean is $\bar{s}$ and is given by Eq. (16); and the second moment is $\overline{s^2} = \sigma^2(n-1)/n$, as was found in Eq. (42). Whereupon it follows that

$$\overline{(r_s-r)^2} = r^2\{1-2\omega\bar{s}/\sigma+\omega^2(n-1)/n\}, \tag{46}$$

and that the

$$\frac{\text{r.m.s. error in writing } r_s \text{ for } r}{r} = \{1-2\omega\bar{s}/\sigma+\omega^2(n-1)/n\}^{\frac{1}{2}}. \tag{47}$$

For convenience, the right-hand member of this equation will be designated by the letter F.

For the *optimum* estimate of r, $\omega=[n/(n-1)]^{\frac{1}{2}}$, so the r.m.s. error in the classical formula Eq. (37) is, in units of r,

$$\{2-2(\bar{s}/\sigma)\sqrt{[n/(n-1)]}\}^{\frac{1}{2}}=\{2-2\sqrt{[2\pi/(n-1)]}/B[\tfrac{1}{2}(n-1),\tfrac{1}{2}]\}^{\frac{1}{2}}$$
$$\rightarrow\{2(n-1)\}^{-\frac{1}{2}}\{1-1/16(n-1)-\cdots\}. \quad (48)$$

The gamma functions come from the value of s given in Eq. (16). Helmert[11] published this result in 1876. It was to this end that he derived Eq. (14).

Eq. (47) gives the r.m.s. error of any estimate of r in fractional parts of r. But when r is unknown, we have only r_s, and this increases and decreases with ω. Hence it would be interesting to express the r.m.s. error of an estimate in units of r_s. To accomplish this it is only necessary to multiply Eq. (47) through by $r/r_s=\sigma/\omega s$. Accordingly the

$$\frac{\text{r.m.s. error in writing } r_s \text{ for } r}{r_s}=(\sigma/\omega s)F, \quad (49)$$

F being, as already noted, the expression in ω on the right-hand side of Eq. (47). We shall call the expression $(\sigma/\omega s)F$, just derived, the *proportional* r.m.s. error in r_s, and shall abbreviate it "p.r.m.s." error. It has its minimum value when $\omega=\sigma/\bar{s}$, as is easily found by equating to zero the derivative of $(\sigma/\omega s)F$ with respect to ω. This result shows that the *mean* estimate of r, given in Eq. (39), has the smallest possible p.r.m.s. error.

Like σ and r, the p.r.m.s. error $(\sigma/\omega s)F$ can only be *estimated* from a sample; its true value, as far as can be learned from the sample, remains unknown. Now it so happens that the estimate of $(\sigma/\omega s)F$ is simply F, since the estimate of σ is σ_s, and $\sigma_s/\omega s$ is unity by Eq. (43).

We therefore have shown that

$$F\equiv\{1-2\omega\bar{s}/\sigma+\omega^2(n-1)/n\}^{\frac{1}{2}} \quad (50)$$

is not only by Eq. (47) the r.m.s. error in r_s, in units of r, but that it is also the *estimated p.r.m.s. error in* r_s.

To get the estimated p.r.m.s. error of the *optimum* (classical) estimate of r, we put $\omega=\{n/(n-1)\}^{\frac{1}{2}}$ in the expression just written for F, and for the *mean* estimate we put $\omega=\sigma/\bar{s}$. The numerical values of F for these two estimates are given in Table V for n running from 2 to 10. The estimated p.r.m.s. error F in either estimate of r is seen to be roughly 25 percent when $n=9$, and it increases rapidly as n decreases. Evidently, then, an estimate of σ is subject to rather violent fluctuations when n is very small.

In the last column of Table V are shown values of $1/[2(n-1)]^{\frac{1}{2}}$ for comparison with the second and third columns. Evidently $1/[2(n-1)]^{\frac{1}{2}}$ comes about midway between the *optimum* and *mean* values of F; it is a little larger than the former and a little smaller than the latter. It is perhaps a good enough approximation for either estimate even down to $n=2$ and 3, since little significance can be attached to such small samples anyway. The values of F in the second and third columns of Table V clearly approach those of $1/[2(n-1)]^{\frac{1}{2}}$ in the last column. It should be mentioned that Helmert in his 1876 paper[11] gave a three-place table of F for the *optimum* estimate running from $n=2$ to $n=8$, and compared it with $1/[2(n-1)]^{\frac{1}{2}}$.

On account of certain considerations arising from the notion of maximum likelihood, it is probably safe to say that when an estimate of r is to be made from the sample alone, there is no better procedure than the classical one of using the *optimum* estimate, Eq. (37). We have here discussed other ways of estimating the probable

TABLE V. Values of

$$F=\{1-2\omega\bar{s}/\sigma+\omega^2(n-1)/n\}^{\frac{1}{2}} \quad (50)$$

for the *optimum* (classical) and the *mean* estimates of the probable error. Comparison with $1/[2(n-1)]^{\frac{1}{2}}$. For the *optimum* estimate, $\omega=[n/(n-1)]^{\frac{1}{2}}$. For the *mean* estimate, $\omega=\sigma/\bar{s}=\sqrt{(n/2\pi)}\,B[\frac{1}{2}(n-1),\frac{1}{2}]$.

n	F optimum	F mean	$1/[2(n-1)]^{\frac{1}{2}}$
2	0.635 7915	0.755 5106	0.707 1068
3	.477 0180	.522 7231	.500 0000
4	.396 6920	.422 0157	.408 2483
5	.346 4517	.362 9993	.353 5534
6	.311 3427	.323 2123	.316 2278
7	.285 0656	.294 1050	.288 6751
8	.264 4600	.271 6367	.267 2612
9	.247 7471	.253 6224	.250 0000
10	.233 8406	.238 7648	.235 7023

error mainly to emphasize the fact that all of them are subject to fluctuations arising from the sampling distribution of s as given by Helmert in Eq. (14).

If n is so large that the sampling distribution of s (Helmert's Eq. (14)) can be considered normal, its area can be divided into quarters that for practical purposes are symmetrically situated about the mean. An estimate of r then has a probable error, and since the curve is normal, this probable error will be $0.674\cdots$ times the S.D. or the r.m.s. fluctuation. But we have already observed from Table V that the r.m.s. errors in the *optimum* and *mean* estimates approach $1/[2(n-1)]^{\frac{1}{2}}$, and when n is large enough for one of these estimates to have a probable error, any of the other possible estimates that have been considered would have practically the same r.m.s. error; hence we can say that when a probable error of an estimate of the probable error r exists, its estimated value is $0.674\cdots/[2(n-1)]^{\frac{1}{2}}r_s$, that is, the

$$\text{estimated probable error in } r_s = 0.674\cdots r_s/[2(n-1)]^{\frac{1}{2}} = \gamma r_s/\sqrt{(n-1)}, \quad (51)$$

γ having the value $0.674\cdots/\sqrt{2}=0.4769\cdots$ as in Eq. (29). This is often loosely called "the probable error of the probable error." Strictly, the probable error r has no probable error, since it is a definite, though perhaps unknown, magnitude for any set of n observations. The *estimate* r_s made from the sample alone does, however, always have a r.m.s. error, but cannot have a *probable* error, as just explained, unless n is so large that the distribution of s is practically normal. This condition is perhaps approached closely enough when $n=20$, but of course no definite line can be drawn there. Now either from choice or circumstances, 20 is about as large a number of observations as physicists are in the habit of taking, so that only rarely does an estimate of r actually have a probable error. It therefore seems best to deal exclusively with the estimated p.r.m.s. error of r_s, which has been designated by the letter F in Eq. (47), calculated in Table V, and which is well enough approximated by the simple expression $1/[2(n-1)]^{\frac{1}{2}}$. Accordingly, the mean of n observations, together with either the *optimum* or the *mean* estimate r_s of the probable error, should then be written

$$\bar{x}\pm r_s(1\pm 1/[2(n-1)]^{\frac{1}{2}}).$$

In so doing, it is important to remember that although r_s is the estimated *probable* error in $\bar{x}$, the quantity $1/[2(n-1)]^{\frac{1}{2}}$ is the estimated proportional *r.m.s.* error in r_s.

Only when n is large can any reasonable degree of belief be placed in an estimate of r. For this reason a statement of the estimated probable error r_s is by itself of little use; we require also the source of this estimate and whether it be from 5 observations or from 25. If it is from 5 observations we know immediately that it is subject to an estimated p.r.m.s. error of over one-third and it must therefore not be taken too seriously. One way of overcoming this difficulty is to bring in prior knowledge by the methods to be outlined later, but this is not always feasible nor possible. On the other hand, if the estimate is made from 25 observations, some significance can be attached to it. In publishing an estimate made from a sample alone, either n or the estimated p.r.m.s. error should be stated. Thus, the result of the 10 observations made on a micrometer, previously considered, should be written

$$1.0760\pm 0.0008(1\pm 0.24)$$

or

$$1.0760\pm 0.0008, \text{ (10 observations)}.$$

Either line conveys the information that the estimated probable error is subject to considerable doubt. The estimated p.r.m.s. error tells how many figures are significant in r_s, and in turn r_s tells how many are significant in $\bar{x}$. A proper appreciation of these principles is essential when correcting data for systematic errors, or when drawing any conclusion from experimental results.

(4e). The posterior method. The prior and posterior curves for σ

Estimates of σ obtained by maximum likelihood or by any empirical method are based on the sample alone and hence are subject to statistical fluctuations. They take no account of knowledge concerning σ that may exist in varying amounts before the sample is taken. The confidence that any one places in an estimate made by one of the foregoing devices will depend in

some manner on his previously formed ideas concerning the range in which σ lies and on how large the sample is. As n is indefinitely increased, previous experience and ideas are gradually and unconsciously relegated into insignificance.

The posterior method of reasoning combines prior knowledge with the information contained in the sample. It is applied in a qualitative way quite generally. Everyone who thinks to himself, "This result seems higher (or lower) than I had for good reasons expected to find it; I wonder therefore if by chance it is not too high (or too low)," is combining prior knowledge with new information provided by the sample and is therefore employing, qualitatively, the posterior method.

Prior[32] knowledge concerning σ may range from none at all to the ability to place it within very narrow limits. As an example of the latter situation we may cite cases where it is possible to make a long series of measurements (perhaps a hundred) on a single magnitude. The S.D. of this long series multiplied by $(100/99)^{\frac{1}{2}}$ may confidently be adopted as the correct value of σ for computing the probable error of subsequent shorter series of observations made with the same instrument and under the same conditions. In such a situation, the value of σ is established so definitely that the S.D. of the subsequent small samples need not be computed at all, and the uncertainties of trying to estimate σ from each one of them alone are eliminated.

At the other extreme stands the less fortunate situation where nothing at all is known regarding σ and where there is no hope of taking a longer series of measurements under comparable conditions in order to establish it. Between the two extremes come more or less hazy notions, often no more than enough to state wide limits between which σ must lie. At other times the limits may be narrower.

These notions might be expressed graphically in a probability curve, to be called a *prior existence curve*, so drawn that the area between any two abscissas is the probability of finding σ between them. The total area under the curve would then be unity, since σ must lie somewhere within the range of the curve. A simple curve is shown in Fig. 12. Here it is supposed that σ is known to lie somewhere between 1 and 2, and the probability that it lies in any intermediate interval is proportional to the width of that interval; hence the curve is flat. Such a prior curve, having finite discontinuities at $\sigma = 1$ and $\sigma = 2$, would, of course, never be used in practice, but it is a convenient one mathematically and so will serve well for the first example.

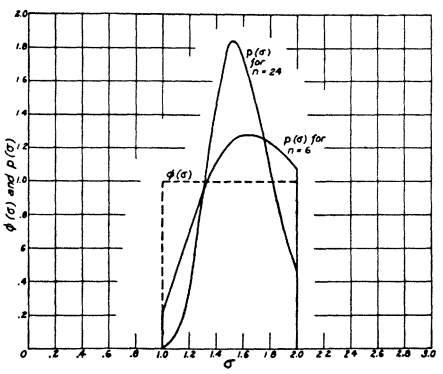


FIG. 12. Prior and posterior estimates of σ. – – – – – $\phi(\sigma)$ *vs.* σ. The *prior* existence curve for σ shows the state of knowledge concerning the S.D. of the parent population *before* a sample is drawn from it. In this example, σ is known to lie with constant probability between 1 and 2. ——— $p(\sigma)$ *vs.* σ. The *posterior* curve for σ shows the state of knowledge concerning the S.D. of the parent population *after* a sample has been drawn and its S.D. computed. Here, the sample was found to have a S.D. of 3/2. The probability is no longer constant between 1 and 2, but becomes more and more concentrated about the point $\sigma = s$ as n increases. The area under all curves is unity.

In the situation where a long series of observations has provided rather definite information, the prior existence curve would have nearly all of its area enclosed in a narrow strip centered at the S.D. of the long series. The exact shape of the curve over this short interval would be unimportant.

A horizontal line extending from very small to very large values of σ and including unit area with the σ axis implies that the S.D. of the parent population has equal probability in equal ranges. Such a graph might seem to be the appropriate prior existence curve in the absence

[32] Prior knowledge can sometimes be obtained *after* the n observations are taken as well as *before*. Our adjectives relating to time are chosen for convenience to fit the usual descriptions of the law of causality, but they may be changed if desired.

of any previous knowledge whatever concerning the precision of a set of observations; but if there is no knowledge concerning σ, then there is none concerning σ^2, σ^3, $\ln \sigma$, $\cdots$; and if the horizontal line expresses ignorance of σ, it must also express ignorance of these functions of σ. But if σ has equal probability in equal ranges, then σ^2, σ^3, $\ln \sigma$, $\cdots$ do *not* have equal probabilities in equal ranges of σ^2, σ^3, $\ln \sigma$, $\cdots$. So it appears hazardous to attempt to express mathematically a state of complete ignorance concerning σ. Nevertheless, Harold Jeffreys[33] has argued that the correct procedure in such cases is to make the ordinates on the prior existence curve proportional to σ^{-1}, i.e., to assume that $\ln \sigma$ is uniformly distributed. Be that as it may, it will be clear later that when the prior information is so hazy that there is difficulty in expressing it, the posterior method is affected by the statistical fluctuations of small samples nearly as much as the estimates made by maximum likelihood or any empirical method, and so is hardly worth the effort. Jeffreys' curve is a special case of one introduced by Molina and Wilkinson in 1929, which will be studied later.

The quantitative application of the posterior method of approaching the parent population is always possible by Laplace's generalization of Bayes' theorem[34–38] *provided* the state of prior knowledge is expressed graphically or analytically in a prior existence curve. The process involves only simple principles in the theory of probability.

If $\phi(\sigma)$ is the ordinate on the prior existence curve at the abscissa σ, then $\phi(\sigma)\,d\sigma$ is the *prior existence probability*—the probability that the S.D. of the parent population lies in the interval $\sigma \pm \frac{1}{2}d\sigma$ according to the state of knowledge existing *before* the sample was drawn. Now *if* the S.D. of the parent population is σ, the probability of drawing a sample having the S.D. $s \pm \frac{1}{2}ds$ is, by Helmert's Eq. (14), *const.* $\times \sigma^{-1}(s/\sigma)^{n-2} \exp(-ns^2/2\sigma^2)\,ds$. This is called the *prior productive probability* of σ. The probability that the S.D. of the parent population lies in the interval $\sigma \pm \frac{1}{2}d\sigma$ *and* that the S.D. of a sample of n drawn therefrom will lie in $s \pm \frac{1}{2}ds$ is the probability of a compound event, and will therefore be proportional to the product of the *prior existence* and the *prior productive* probabilities, namely,

$$p\,d\sigma\,ds = const. \times \phi(\sigma)\sigma^{-1}(s/\sigma)^{n-2} \exp(-ns^2/2\sigma^2)\,d\sigma\,ds. \quad (52)$$

We can imagine a surface of ordinate p plotted on the orthogonal axes σ and s. Let us take a slab of thickness ds at s, parallel to the p,σ plane. The equation of the curve made by this section is

$$p\,d\sigma = const.\ \phi(\sigma)\sigma^{-1}(s/\sigma)^{n-2} \exp(-ns^2/2\sigma^2)\,d\sigma.$$

$p\,d\sigma$ will be proportional to the *posterior probability* of σ, which is the name given to the probability that the S.D. of the parent population lies within the interval $\sigma \pm \frac{1}{2}d\sigma$ *after* the sample is drawn and found to have S.D. s. The factor of proportionality will be unity if the area under the curve is unity, as it will be if the constant is properly chosen. This is insured if the last equation is written

$$p\,d\sigma = \frac{\phi(\sigma)\ \sigma^{-1}\ (s/\sigma)^{n-2} \exp(-ns^2/2\sigma^2)}{\int_0^\infty \phi(\sigma)\sigma^{-1}(s/\sigma)^{n-2} \exp(-ns^2/2\sigma^2)d\sigma}\,d\sigma. \quad (53)$$

When the constant factor in the equation of any probability curve is so chosen that the total area under the curve is unity, the equation is said to be "normalized," and the required constant factor is called the "normalizing factor." It simply serves to identify *unity* with *certainty*. As in the equation just written, the process of normalization is nearly always most conveniently accomplished by writing a denominator identical with the numerator, and then integrating in the denominator over all values of the variable whose probability is being written.

The following example will illustrate the use of the method and will exhibit some of its

[33] Harold Jeffreys, *Scientific Inference*, Ch. 5 (Cambridge University Press 1931); Proc. Roy. Soc. **A138**, 48–55 (1932); Proc. Camb. Phil. Soc. **29**, 83–87 (1933); Proc. Roy. Soc. **A140**, 523–534 (1933). Jeffreys' arguments are disputed by R. A. Fisher, Proc. Roy. Soc. **A139**, 343–348 (1933).

[34] Thomas Bayes, Phil. Trans. Roy. Soc. **53**, 370–418 (1763).

[35] Laplace, *Théorie Analytique des Probabilités* (1812).

[36] Poisson, *Recherchés sur la Probabilité des Jugements* (1837).

[37] See also Edward C. Molina, Bull. Am. Math. Soc. **36**, 369–392 (1930); Ann. Math. Stat. **2**, No. 1, 23–37 (1931).

[38] An excellent treatment of Laplace's generalization of Bayes' theorem is in Ch. 5 of Thornton C. Fry's, *Probability and Its Engineering Uses* (Van Nostrand, 1928). See also Ch. 6 in Arne Fisher's *Mathematical Theory of Probabilities* (Macmillan, second edition 1922).

features. We shall suppose that before any sample is drawn, σ is known to lie between 1 and 2, and that equal intervals are equally probable in this range. Then the ordinates of the prior existence curve will be

$$\left.\begin{aligned} &\phi(\sigma)=0 \quad 0<\sigma<1 \\ &\phi(\sigma)=1 \quad 1<\sigma<2 \\ &\phi(\sigma)=0 \quad 2<\sigma \end{aligned}\right\}. \tag{54}$$

The graph is shown dashed in Fig. 12. Now let us suppose that a sample of 6 is drawn and that its S.D. is computed and found to be 1.5. Are all values of σ between 1 and 2 equally probable now? The posterior curve furnishes the answer. Its ordinates are found by substituting the proper values of $\phi(\sigma)$, n, and s into Eq. (53). The result is

$$\left.\begin{aligned} &p(\sigma)=0 \quad 0<\sigma<1 \\ &p(\sigma)\,d\sigma=\frac{\sigma^{-1}(s/\sigma)^{n-2}\exp\,(-ns^2/2\sigma^2)}{\int_1^2\sigma^{-1}(s/\sigma)^{n-2}\exp\,(-ns^2/2\sigma^2)d\sigma}d\sigma \\ &\quad=187.13\sigma^{-5}\exp\,(-27/4\sigma^2)d\sigma \\ &\qquad\qquad 1<\sigma<2 \\ &p(\sigma)=0 \quad 2<\sigma \end{aligned}\right\}. \tag{55}$$

The normalizing factor 187.13 was obtained by using the *Tables of the Incomplete Gamma Function* to evaluate the denominator in the preceding line.

Eq. (55) is plotted in the same figure. Instead of being flat, the posterior curve has a maximum. Approximately half the area is included between the abscissas 1.46 and 1.86, so the location of σ is now a little more definite than it was. The area of the posterior curve would be more concentrated, and σ more definitely located, if the prior curve had had a maximum near the middle instead of being flat.

If n had been 24 instead of 6, the equation of the posterior curve would have been

$$p(\sigma)\,d\sigma=33.371\times10^8\sigma^{-23}\exp\,(-27/\sigma^2)\,d\sigma, \qquad 1<\sigma<2. \tag{56}$$

This is also shown in the figure. The area is now much more concentrated in the neighborhood of the maximum, so that with $n=24$ we should have a much better idea of where σ actually lies.

The posterior method furnishes a probability curve for σ by changing the prior existence curve in accordance with the new information contained in the sample. *Before* the sample was drawn the probability was given by $\phi(\sigma)$; *afterward*, by $p(\sigma)$.

The shape of the posterior curve changes more or less as s changes; it is therefore not entirely free from the statistical fluctuations of sampling. Just how sensitive it is to variations in s will depend on how large n is and on how definite the prior information was; as one would expect, when the prior curve confines σ to fairly narrow limits and n is not large, variations in s have little effect; in fact if the prior information is extremely definite, a very large value of n will be required to affect noticeably the posterior curve through changes in s. This is why the value of σ that has once been established by means of a long series of measurements can be used for subsequent shorter series; the standard deviations of these shorter series need not be computed at all because their influence on the posterior curve would be negligible. However, if the prior information fixes σ only loosely, the sample may influence the posterior curve considerably, even when n is small. When n is large, the posterior curve rises to a sharp peak at the abscissa provided by maximum likelihood, irrespective of the shape of the prior curve. Furthermore, as n increases, the fluctuations in s become inappreciable. It is therefore correct to say that a value of σ can be established by taking a long series of measurements.

The form of the prior existence curve shown in Fig. 12 is useful for illustration, but on account of its discontinuities it lacks some of the practical features of the curve proposed by Molina and Wilkinson to be considered in a later section.

(4f). Further remarks on the method of maximum likelihood

Before leaving the prior existence curve of Fig. 12 it may be worth while to examine further the position of the maximum of the resulting posterior curve. Starting with a flat prior existence curve like that in Fig. 12, the maximum

will always come at the abscissa $\sigma = s[n/(n-1)]^{\frac{1}{2}}$. This result arises, of course, from differentiating with respect to σ the expression for $p(\sigma)$ in Eq. (55), holding s constant, and setting the derivative equal to zero. The resulting relation between σ and s is independent of the denominator, which is merely a constant; hence this relation is independent of the range over which the prior existence curve extends, provided only that it is flat. If the prior existence curve for σ were other than flat, the maximum on the posterior curves would in general lie elsewhere, because $p(\sigma)\, d\sigma$ would no longer be given by the right-hand side of Eq. (55) nor anything proportional to it, but would be given by Eq. (53) wherein $\phi(\sigma)$ would not be a constant but some function of σ.

Now the position of the maximum (or the mode) on the posterior curve that comes from using a flat prior existence curve for σ turns out to be identically the same relation between σ and s as was obtained in Eq. (36), which was arrived at in our search for the parent population that is most favorable (or *most likely* in Fisher's sense) to the S.D. that was actually observed in the sample. It will be recalled that we arrived at this most favorable parent population by differentiating Helmert's Eq. (14) with respect to σ and setting the derivative equal to zero; also that we called this process the *method of maximum likelihood*, after Fisher. That the two results—the position of the maximum on the posterior curve and the application of the method of maximum likelihood—must be identical is evident from the fact that when the prior existence curve for σ is flat, $\phi(\sigma)$ is simply a constant and the right-hand side of Eq. (53) then expresses, save for a constant factor, the same relation between σ and s as occurs in Helmert's equation, so that we are really differentiating the same function in both cases.

Because of this coincidence, the method of maximum likelihood has often been described as the process of finding the mode of the posterior curve that arises from a flat prior existence curve. This explanation, although it masks the true nature of the notion of maximum likelihood, would in itself do no harm were it not that by implications it leads to misinterpretations. Thus, as has been pointed out, the abscissa of the mode of the posterior curve changes as the prior existence curve changes, and the particular abscissa $\sigma = s[n/(n-1)]^{\frac{1}{2}}$ is the mode of the posterior curve in general only when the prior existence curve for σ is flat; whence such an explanation as proposed above leads innocently to the statement that the method of maximum likelihood is a posterior method and depends on a uniform (flat) prior existence curve for the parameter sought—in our case, σ. But if we had used some function of σ such as σ^2, σ^3, $\ln \sigma$, $\cdots$ in place of σ as the equally spaced abscissas along the axis of the prior and posterior curves, we should likewise have found that, starting with a flat prior curve for σ^2, σ^3, $\ln \sigma$, $\cdots$ as the case may be, the relation $\sigma = s[n/(n-1)]^{\frac{1}{2}}$ is *not* that existing at the mode of the new posterior curve; whereupon any uniqueness that the method of maximum likelihood might have seemed to possess now appears to have been an illusion.

The resolution of the difficulties that we are led to by such an explanation lies in the realization that the method of maximum likelihood is not a posterior method at all. It is simply a process for finding the parent population that is most favorable to the event that was observed to happen—in our case a sample having S.D. s. Obviously the answer to such a problem as finding the most favorable parent population should not, in fact *must* not, depend on the choice of coordinates nor on any state of prior knowledge, and it is interesting to note that if Helmert's Eq. (14) be expressed in terms of any function of σ, rather than in terms of σ itself, the result of setting the derivative with respect to σ or any function of σ equal to zero is always the same as that already found in Eq. (36), namely, $\sigma = s[n/(n-1)]^{\frac{1}{2}}$. This invariance is a general property of the method of maximum likelihood, and the proof is very simple: if the function $f(x)$, continuous in any interval, be expressed in terms of v so that $f(x) = F(v)$ and $v = g(x)$ over that interval, we shall find that the values of x that maximize or minimize $f(x)$ correspond through the relation $v = g(x)$ precisely to the values of v that maximize or minimize $F(v)$, provided dv/dx is neither 0 nor ∞.

A graphical illustration of the meaning of maximum likelihood is provided by Fig. 13, which shows three Helmert curves for $n = 6$. One

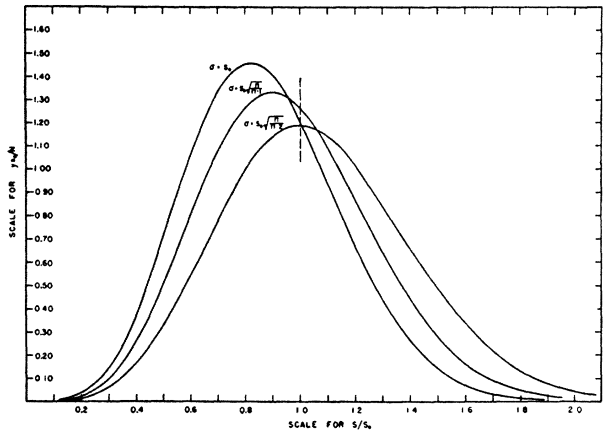


FIG. 13. Curves illustrating the meaning of maximum likelihood. A sample of n is drawn, and its S.D. proves to be s_0. The S.D. σ of the parent population can be anything between 0 and ∞, but the value $\sigma = s_0\sqrt{(n/(n-1))}$ is "most likely," for this gives the greatest possible ordinate at $s = s_0$ on the frequency distribution curve for the S.D. of samples of n (Helmert's equation). The curves illustrate that the ordinate at $s = s_0$ is higher when $\sigma = s_0\sqrt{(n/(n-1))}$ than when $\sigma = s_0$ or $\sigma = s_0\sqrt{(n/(n-2))}$. With the latter value, the mode of the curve comes at $s = s_0$. The equation of the curves is

$$y\,ds = \frac{Nn^{\frac{1}{2}(n-1)}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2}e^{-ns^2/2\sigma^2}\,ds \qquad \text{(Helmert's equation)}$$

in which n has been placed equal to 6.

curve is plotted with the maximum likelihood value of σ, namely, $\sigma = s_0[n/(n-1)]^{\frac{1}{2}}$, s_0 being the S.D. observed in a sample of n; and the other two are plotted with slightly less and slightly greater values of σ. At $s = s_0$, or at $s/s_0 = 1$, the ordinate along the curve having S.D. $\sigma = s_0[n/(n-1)]^{\frac{1}{2}}$ is clearly greater than the ordinates of the two other curves. This fact illustrates that out of the infinity of parent populations that the sample could have come from, that having the maximum likelihood value of σ is most favorable, since it gives the greatest possible ordinate at $s = s_0$ and therefore maximizes the probability of drawing a sample of S.D. $s_0 \pm \frac{1}{2}ds$.

(4g). The posterior method, continued. The probability curve of the unknown mean, and the calculation of the posterior quartile deviation

One particular value of σ gives the u,s frequency surface that was studied in previous sections. A u,s frequency surface having its total volume equal to unity but made up of contributions from several values of σ would be a composite surface. Its sections would no longer be the u and s curves that were studied, since all values of σ under the prior existence curve for σ make their contributions to the volume according to their relative probabilities, which are designated by the ordinates $\phi(\sigma)$.

To make the posterior method complete, it is necessary to consider also the prior existence curve for the mean μ of the parent population. The prior curve for μ, as well as that for σ, will have its effect on the composite surface.

We may take sections $s = \text{const.}$ on this composite u,s surface, just as before, but such sections will not now be normal curves as they were with the simple surface. We shall assume that they are symmetrical, however; and we shall define the "posterior quartile deviation" r_q to be the absolute magnitude of the u abscissas that divide an s section symmetrically into quarters. Sometimes. if not always, these abscissas r_q will vary as the s coordinate of the

section varies, whereas with the simple u,s surface the abscissas $\pm r$ cut *all* $s=$ const. sections symmetrically into quarters.

Mathematically manageable forms, allowing sufficient freedom for any degree of prior knowledge likely to be encountered, have been introduced by Molina and Wilkinson[39] for the prior existence probabilities of the mean μ and the S.D. σ of the parent population. They are

$$\phi(\sigma)\,d\sigma=\frac{1}{2^{\frac{1}{2}c}\,\Gamma(\frac{1}{2}c+1)\,\sigma}(a/\sigma)^{c+2}e^{-a^2/2\sigma^2}\,d\sigma, \qquad (57)$$

$$\theta(\mu)\,d\mu=A\left[1+\frac{1}{1+a^2/ns^2}\left(\frac{\bar{x}-\mu}{s}\right)^2\right]^{-\frac{1}{2}b}d\mu. \qquad (58)$$

a, b and c are adjustable constants. A can easily be found by setting $\int_{-\infty}^{\infty}\theta(\mu)\,d\mu=1$, but its value will not be needed.

Graphs of Eq. (57) with $c=3$ and $c=10$ are shown in Fig. 14. They are skew curves; the mode comes at $a/(c+3)^{\frac{1}{2}}$ and the mean at $[a/\sqrt{(2\pi)}]B(\frac{1}{2}(c+1),\frac{1}{2})$. The σ axis is tangent to the curves at 0 and ∞, where it makes high order of contact, so extremely small and extremely large values of σ are always excluded. The larger c is, the narrower is the range in which the greater part of the area is confined. The two constants a and c permit whatever concentration of area happens to fit the state of prior knowledge and also permit the mean or mode of the curve to be placed at will. It will be noticed that if $a=0$ and $c=-2$, Molina and Wilkinson's prior curve for σ reduces to the one proposed by Jeffreys,[38] namely, $\phi(\sigma)=$ const. σ^{-1}, and that if $a=0$ and $c=-3$, we obtain the flat prior existence curve $\phi(\sigma)=$ const.

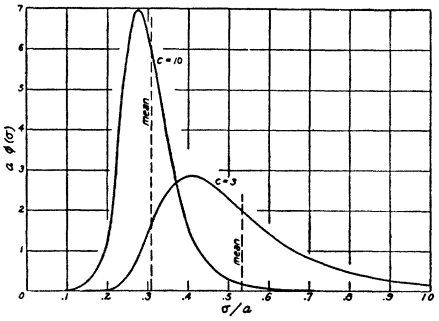


FIG. 14. Molina and Wilkinson's prior existence curve for σ.

$$\phi(\sigma)\,d\sigma=\frac{1}{2^{\frac{1}{2}c}\Gamma(\frac{1}{2}c+1)\sigma}(a/\sigma)^{c+2}e^{-a^2/2\sigma^2}\,d\sigma \qquad (57)$$

a and c are arbitrary constants. The area included between any two abscissas is the prior probability that σ lies within that interval. The total area under each curve is unity. The curves here drawn with $c=3$ and $c=10$ show that increasing values of c correspond to increasingly definite prior knowledge concerning the S.D. of the parent population.

The prior curve for the mean is of the Student type (see Fig. 5). It is symmetrical about the mean $\bar{x}$ of the sample, so when $b>0$ this curve implies that the mean of the sample is *a priori* to be preferred as the mean of the parent population. When $b=0$, the curve is flat from 0 to ∞, meaning that equal ranges from $-\infty$ to $+\infty$ are, *a priori*, equally probable. This is the most conservative value of b.

If μ and σ were *known*, the probability of drawing a sample with S.D. $s\pm\frac{1}{2}ds$ and with mean at $\bar{x}\pm\frac{1}{2}d\bar{x}$ would be given immediately by Eq. (11), and can be written

$$y(\bar{x},s)\,d\bar{x}\,ds=C\sigma^{-2}(s/\sigma)^{n-2}\exp\left[-ns^2/2\sigma^2-n(\bar{x}-\mu)^2/2\sigma^2\right]d\bar{x}\,ds. \qquad (59)$$

This is the *prior productive probability* of μ and σ.

The *posterior* probability of μ and σ, i.e., the probability that the mean and S.D. of the parent population lie in the ranges $\mu\pm\frac{1}{2}d\mu$ and $\sigma\pm\frac{1}{2}d\sigma$ while the mean and S.D. of the sample lie in the ranges $\bar{x}\pm\frac{1}{2}d\bar{x}$ and $s\pm\frac{1}{2}ds$, is given, except for the normalizing factor, by the product of the *prior existence* and *prior productive* probabilities as expressed in Eqs. (57), (58), (59). Finally, integration of this product over all possible values of σ gives the posterior probability of μ, namely

$$y\,d\mu=\frac{\int_0^{\infty}\theta(\mu)\,\phi(\sigma)\,y(\bar{x},s)\,d\sigma}{\int_{-\infty}^{\infty}d\mu\int_0^{\infty}\theta(\mu)\,\phi(\sigma)\,y(\bar{x},s)\,d\sigma}d\mu. \qquad (60)$$

[39] E. C. Molina and R. I. Wilkinson, Bell Syst. Tech. J. **8**, 632–645 (1929).

This is the probability, *after* the sample of mean $\bar{x}$ and S.D. s has been drawn, that the mean of the parent population lies in the interval $\mu \pm \frac{1}{2}d\mu$. As usual, the denominator is simply the normalizing factor. The constant C in Eq. (59) cancels, so its value need not be determined.

The integrations with respect to σ in this fraction are easily performed when the prior existence probability functions $\theta(\mu)$ and $\phi(\sigma)$ have the forms suggested by Molina and Wilkinson. The result is

$$y\,d\mu = \frac{1}{s\sqrt{(1+a^2/ns^2)}\,B\left(\frac{n+1+c+b}{2}, \frac{1}{2}\right)}\left[1+\frac{1}{1+a^2/ns^2}\left(\frac{\bar{x}-\mu}{s}\right)^2\right]^{-\frac{1}{2}(n+2+c+b)} d\mu \tag{61}$$

for the posterior probability of μ.

It is here convenient to replace the error $\bar{x}-\mu$ by its usual symbol u, $d\mu$ by $-du$, and to denote $n+2+c+b$ by T and the entire resulting expression by $-q(u)\,du$. Then

$$q(u)du = \frac{1}{s\sqrt{(1+a^2/ns^2)}\,B(\frac{1}{2}(T-1), \frac{1}{2})}\left[1+\frac{1}{1+a^2/ns^2}\frac{u^2}{s^2}\right]^{-\frac{1}{2}T} du \tag{62}$$

is the *posterior probability* curve for the error u when the S.D. of the sample is s.

This is the equation for a section $s=$ const. on the composite u,s frequency surface formed by the contributions of *all* values of σ in the assumed $\phi(\sigma)$ distribution. The posterior quartile deviation, r_q, previously defined is then given by the integral

$$\int_{-r_q}^{r_q} q(u)\,du = \tfrac{1}{2}\int_{-\infty}^{\infty} q(u)\,du = \tfrac{1}{2}, \tag{63}$$

since it must divide the $s=$ const. curve symmetrically into quarters. The value of r_q will then be expressed by

$$r_q = st(1+a^2/ns^2)^{\frac{1}{2}} \tag{64}$$

where t is a function of T only, and satisfies

$$\left.\begin{aligned} &\frac{1}{B\left(\frac{T-1}{2}, \frac{1}{2}\right)}\int_{-t}^{t}(1+t^2)^{-\frac{1}{2}T}\,dt = \tfrac{1}{2}, \\ &T = n+2+c+b. \end{aligned}\right\} \tag{65}$$

The integral by which t is determined is of the Student type; in fact t is just the value of ζ given by Table II when the n in that table is replaced by T.[40] If the integral were equated to 0.80, 0.90, and 0.9973, the corresponding limits would determine the posterior 80, 90, and 99.73 percentile deviations. These can be denoted by $r_q(80)$, $r_q(90)$, $r_q(99.73)$. The posterior probable error, or 50 percent error, could be denoted by $r_q(50)$, but unless emphasis is desired it will usually be written simply as r_q.

Curves showing t as a function of T for the four values of the integral of Eq. (65) are shown in Fig. 15. The ordinates for the 50 percent curve come from Table II; the others were kindly furnished by Molina and Wilkinson. They show a similar chart in their paper. The procedure is very simple after the constants a, b, and c are settled upon. It is only necessary to find t for the abscissa $T=n+2+c+b$ by means of Fig. 15; then to compute r_q by Eq. (64).

It is interesting now to notice certain features in the results that have been obtained. In Fig. 15 the ordinates for large values of T drop off more and more slowly with increase in T, so when n is large, t is not very sensitive to changes in n, b, and c. Hence as n increases indefinitely, t approaches coincidence with ζ regardless of b and c. Further, as $n\to\infty$, $a^2/ns^2\to 0$ and $1+a^2/ns^2 \to 1$; therefore $r_q \to st \to s\zeta$, which in turn ap-

[40] The value of T to be used in Table II must not be confused with the actual number of items n in the sample. T and n are numerically the same only when $2+c+b=0$, as Eq. (65) shows. In the prior existence function assumed by Jeffreys (footnote 33), $c=-2$ and $b=0$, and this relation is satisfied. Since Jeffreys also assumed $a=0$ we have from Eq. (64) the further interesting relation that $r_q = ts = \zeta s$. Thus when $\phi(\sigma) = \text{const.}/\sigma$, the posterior quartile deviation is numerically equal to what may be called "Student's 50 percent error" (see Table II and Fig. 10c and the accompanying discussion). It should be emphasized, however, that this is a mere numerical coincidence and that the two quantities r_q and ζs have very different theoretical meanings.

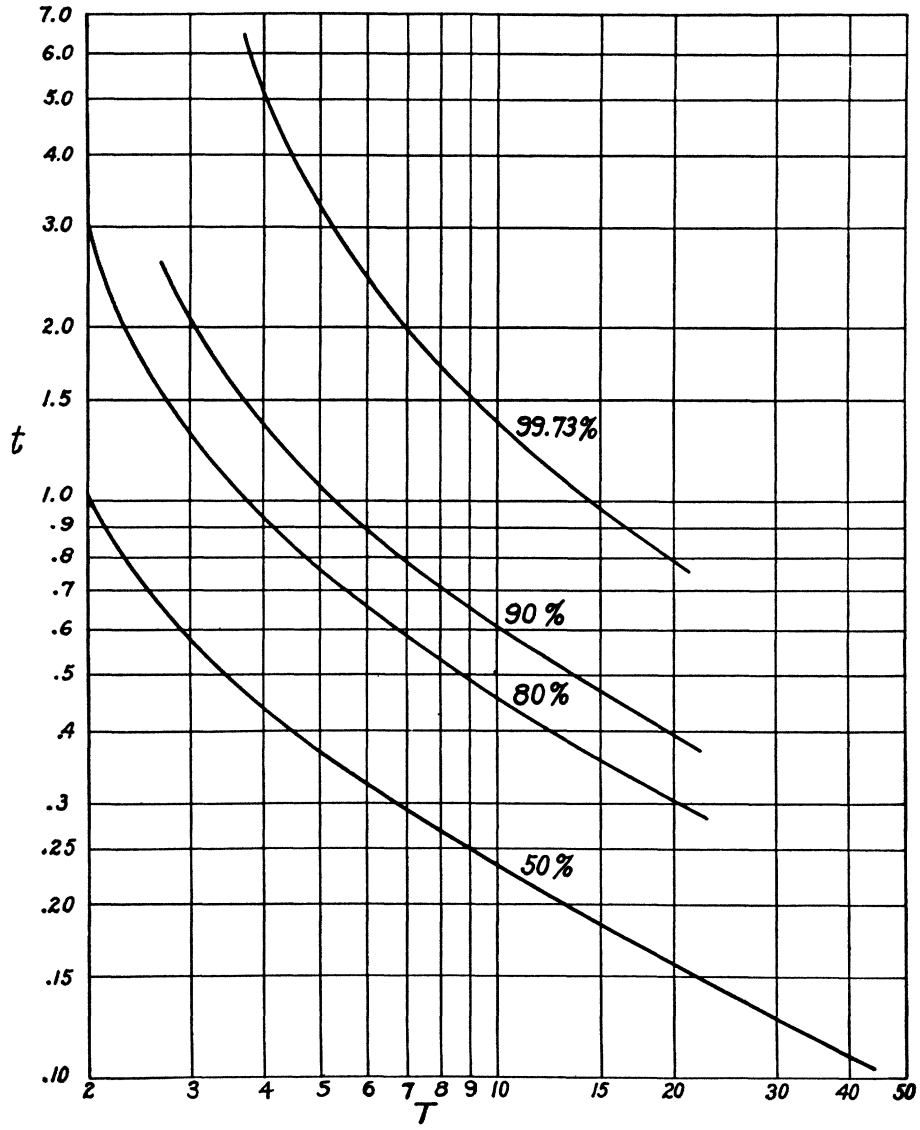


FIG. 15. Chart for using Molina and Wilkinson's prior existence curves. The ordinate t on any curve multiplied by $s\sqrt{(1+a^2/ns^2)}$ gives the indicated posterior percentile deviation of u. The abscissa $T=n+2+b+c$. n = number in sample; a, b, c are constants used in fitting Molina and Wilkinson's curves to the prior knowledge.

proaches r as a statistical limit. Thus the true probable error will be attained as the sample is indefinitely increased, irrespective of the prior information, for the constants a, b, c then have negligible influence.

When n is small, the situation is different, for the value of t, and hence that of r_q, will depend considerably on b and c. Also if $a \neq 0$, the term a^2/ns^2 in Eq. (64) will be important on account of its stabilizing action for it will prevent r_q from fluctuating as widely as s does. But if $a = 0$, the term a^2/ns^2 will be absent, and r_q will be proportional to s, and will therefore fluctuate with s. This is the situation in Jeffreys' assumption.[33, 40]

The significance of $r_q(50)$ is that according to our knowledge and beliefs concerning μ and σ, derived from all sources including the sample, we are willing to lay even odds that $|u| \lesseqgtr r_q$. The significances of $r_q(80)$, $r_q(90)$, and $r_q(99.73)$ are similar except that the odds are 80 : 20, 90 : 10, and 99.73 : 0.27 that $|u| < r_q(80)$, $r_q(90)$, and $r_q(99.73)$, respectively.

r_q is not the probable error of the mean of n observations, nor is it an estimate of the probable error, any more than ζs is. r_q simply provides another statistical relation; it differs from ζs in that by taking account of prior information it is not subject to fluctuations to the same degree as s and ζs. It is interesting to note that Molina and Wilkinson[39] made 21 different assumptions regarding the prior existence curves for μ and σ and thereby obtained 21 different values for the posterior quartile deviation r_q. For a sample of $n = 5$ the highest and lowest of these values of r_q are closely in the ratio 2 : 1, which shows that prior information may have considerable influence on r_q when n is small.

(4h). The estimation of σ from several samples

We have seen that a value of σ can be established by taking a long series of measurements on a particular magnitude; if s is the S.D. of this long series, we may with considerable confidence estimate σ to be $s[n/(n-1)]^{\frac{1}{2}}(1 \pm 1/[2(n-1)]^{\frac{1}{2}})$. If n is large the estimated p.r.m.s. error $1/[2(n-1)]^{\frac{1}{2}}$ will be small and the effect of prior knowledge will be negligible. We may then use this value for σ in calculating the probable error of subsequent shorter series of observations made under similar conditions.

Unfortunately it is not always practicable nor possible to take a long series of measurements in order to establish a value of σ. Oftentimes, however, there do exist records of many short series of observations, all presumably made under approximately the same conditions and therefore all with practically the same precision. In such cases it is desirable to have a method for estimating σ from these several sets of observations.

Let there be n_1 observations on the mean μ_1, n_2 on the mean μ_2, $\cdots$, n_m on the mean μ_m. Let the means of these m series of observations be in error by the amounts u_1, u_2, $\cdots$, u_m, and let their S.D. be s_1, s_2, $\cdots$, s_m. By writing down the probabilities of the occurrence of errors and residuals after the manner of the development of Helmert's equation it is not difficult to see that

$$y\,ds_1 ds_2 \cdots ds_m = \text{const.}\,\frac{s_1^{n_1-2} s_2^{n_2-2} \cdots s_m^{n_m-2}}{\sigma^{n_1+n_2+\cdots+n_m-m}} \exp\left(-\frac{n_1 s_1^2 + n_2 s_2^2 + \cdots + n_m s_m^2}{2\sigma^2}\right) ds_1 ds_2 \cdots ds_m \qquad (66)$$

is the probability that the S.D. of the m series will lie in the m ranges $s_1 \pm \frac{1}{2}ds_1$, $s_2 \pm \frac{1}{2}ds_2$, $\cdots$, $s_m \pm \frac{1}{2}ds_m$ while their means lie anywhere between $-\infty$ and $+\infty$. σ is the same for all sets since we are assuming that all the observations are made under the same conditions as far as precision is concerned. It is σ that is to be estimated. To accomplish this we can apply the method of maximum likelihood—that is, differentiate the above expression with respect to σ, set this derivative equal to zero, and solve for σ. The result is

$$\sigma_s^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + \cdots + n_m s_m^2}{n_1 + n_2 + \cdots + n_m - m} = s^2 \frac{n_1 + n_2 + \cdots + n_m}{n_1 + n_2 + \cdots + n_m - m}, \qquad (67)$$

where

$$s^2=\frac{n_1s_1^2+n_2s_2^2+\cdots+n_ms_m^2}{n_1+n_2+\cdots+n_m}. \qquad (68)$$

s as here defined is just the S.D. that would be calculated for the entire lot of $n_1+n_2+\cdots+n_m$ observations if each series of observations were held rigid with respect to its own mean and the m sample means moved into coincidence.

Eq. (67) gives the *optimum* estimate of σ, found from the m series of observations. Its estimated p.r.m.s. error is very closely $1/[2(n_1+n_2+\cdots+n_m-m)]^{\frac{1}{2}}$, which of course reduces to $1/[2(n-1)]^{\frac{1}{2}}$ for a single set, as has already been found in Table V. This optimum estimate, together with its estimated p.r.m.s. error is then statistically more reliable than an estimate made from any one of the individual series of observations that make up the entire lot; it is also statistically more reliable than an estimate from a subsequent short series of measurements yet to be made under the same conditions. We should therefore not bother to compute the S.D. of subsequent short series, but should rather calculate their probable errors immediately by Eq. (13) using therein the more reliable estimate of σ that comes from Eq. (67). There is, of course, no reason why the S.D. of any short series should not be combined with previous ones to get a still more reliable estimate if such a course seems advisable, and it should be noted that the form of the middle member of Eq. (67) is such that this is very easy to accomplish. The point that we wish to emphasize is that the S.D. of short series should not be used by themselves if there is any way to avoid doing so.

An interesting special case is where measurements are made in duplicate. Here $n_1=n_2=n_3=\cdots=n_m=2$, and m, the number of items measured, is equal to $\frac{1}{2}(n_1+n_2+\cdots+n_m)$. Eq. (67) then reduces to

$$\sigma_s^2=(s_1^2+s_2^2+\cdots+s_m^2)/\tfrac{1}{2}m \qquad (69)$$

for the optimum estimate of σ. The S.D. of any pair of measurements is obviously just half the difference between the pair. Now any single pair of measurements constitutes a sample of 2 and is by Table V almost useless for estimating σ, but if several hundred items have been measured in duplicate, the pairs of observations can be combined and used in Eq. (69) to get a fairly reliable estimate, since the r.m.s. error of this estimate will be $1/(2m)^{\frac{1}{2}}$.

As an example in the use of Eq. (67) we take 20 samples of 5 each from the 500 readings on a spectral line that were made by one of us.[1, 7] The fact that all these sets of 5 readings were observations on a single magnitude rather than on distinct means $\mu_1, \mu_2, \cdots, \mu_{20}$ is of no consequence in the application of Eq. (67); there is in fact an advantage for purposes of illustration in having the 500 readings all on the same magnitude, because after we estimate σ by means of Eq. (67) from the 20 samples of 5 each we shall have for comparison the still more reliable estimate obtained from the entire 500. The 20 samples of 5 each were made up from the 500 observations in the following way: Readings No. 1, 11, 21, 31, 41 constitute the first sample, readings No. 51, 61, 71, 81, 91 constitute the second sample, $\cdots$, readings No. 451, 461, 471, 481, 491 constitute the tenth, readings No. 2, 12, 22, 32, 42 constitute the eleventh, readings No. 52, 62, 72, 82, 92 the twelfth, etc. The S.D. and individual estimates of σ made by both the *optimum* and *mean* formulas (Eqs. (37) and (39)) are shown in Table VI. Here $n_1=n_2=n_3=\cdots=5$ and $m=20$. With the squares of the S.D. in the second column Eq. (67) then gives

$$\sigma_s^2=\frac{5\times1336+5\times1976+\cdots+5\times1464}{5+5+\cdots+5-20}\times10^{-8}$$

$$=\frac{1336+1976+\cdots+1464}{16}\times10^{-8}=1517\times10^{-8},$$

$$\sigma_s=0.00389.$$

Here the estimated p.r.m.s. error is $1/[2(n_1+n_2+\cdots+n_m-m)]^{\frac{1}{2}}=1/\sqrt{160}=0.079$, so we write

$$\sigma_s=0.0039(1\pm0.08). \qquad (70)$$

The averages (r.m.s. and arithmetic) of the optimum and mean estimates in the fourth and fifth columns of Table VI compare very favorably with this result, but it is interesting to see how the individual estimates in these same columns fluctuate. Until the estimate of σ written in Eq. (70) has been displaced by a still better one, the probable error of the mean $\bar{x}$ of *any* one of the 20 series of 5 observations each, or indeed of

TABLE VI. *An estimate of σ made from 20 samples of 5 each. Comparison with the optimum and mean estimates of σ made from the individual samples.*

By Eq. (37) the *optimum* estimate of σ is $s[n/(n-1)]^{\frac{1}{2}} = 1.1180s$ when $n=5$.
By Eq. (39) the *mean* estimate of σ is $s(n/2\pi)^{\frac{1}{2}}B(\frac{1}{2}(n-1), \frac{1}{2}) = 1.1894s$ when $n=5$.

Sample No.	$(\text{S.D.})^2 = s^2$ mm²	s mm	Estimates of σ, in mm Optimum	Mean
	$\times 10^{-8}$	$\times 10^{-4}$	$\times 10^{-}$	$\times 10^{-4}$
1	1336	36.55	40.87	43.47
2	1976	44.45	49.70	52.87
3	0936	30.59	34.21	36.39
4	0256	16.00	17.89	19.03
5	0896	29.93	33.47	35.60
6	1064	32.62	36.47	38.80
7	0704	26.53	29.66	31.56
8	0200	14.14	15.81	16.82
9	0544	23.32	26.08	27.74
10	1056	32.50	36.33	38.65
11	3944	62.80	70.21	74.70
12	0256	16.00	17.89	19.03
13	3384	58.17	65.04	69.19
14	2296	47.92	53.57	56.99
15	0800	28.28	31.62	33.64
16	0704	26.53	29.66	31.56
17	0400	20.00	22.36	23.79
18	0776	27.86	31.14	33.13
19	1280	35.78	40.00	42.55
20	1464	38.26	42.78	45.51
	Average	32.41*	38.95**	38.55+*

The *optimum* estimate of σ made from the 20 samples of 5 each is found from Eq. (67):

$$\sigma_s^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + \cdots + n_m s_m^2}{n_1 + n_2 + \cdots + n_m - m}$$

$$= \frac{5\times 1336 + 5\times 1976 + \cdots + 5\times 1464}{5+5+\cdots+5-20} \times 10^{-8}$$

$$= \frac{1336 + 1976 + \cdots + 1464}{16} \times 10^{-8} = 1517 \times 10^{-8},$$

$\sigma_s = 0.00389$ mm.

* arithmetic mean.
** root mean square.

any subsequent 5 observations taken under the same conditions, should be written as

$$[0.674\cdots \times 0.0039/\sqrt{5}](1 \pm 0.08)$$
$$= 0.0013(1 \pm 0.08), \quad (71)$$

which makes use of the estimate of σ furnished by the 20 samples rather than by any individual sample of 5.

In this particular example we have at hand 400 more readings, since we have used only 100 ($=20\times 5$) so far. When σ is estimated from the entire 500 the result is

$$\sigma_s = 0.003583(500/499)^{\frac{1}{2}}[1 \pm 1/\sqrt{2(500-1)}]$$
$$= 0.00359(1 \pm 0.032). \quad (72)$$

The figure 0.003583 is the S.D. of the 500 readings. The factor $(500/499)^{\frac{1}{2}}$ is hardly necessary, since n is so large. The previous estimate of σ furnished by Eq. (70) and used in Eq. (71) should now be replaced by the estimate in Eq. (72). In practice we are generally not so fortunate as to have a series of 500 observations from which to estimate σ but must instead be content to combine several small samples by the method of Eq. (67); indeed, more often the estimate of σ must be made from a single small sample. In such a case, Eq. (67) reduces to Eq. (37), the use of which has been discussed earlier.

§5. Conclusion

So far, we have dealt with methods for laying odds on the error of the mean of a single sample. The error of the mean has referred throughout the paper to the difference between the mean of the n observations in the sample and what the mean would be if n were indefinitely increased. We have therefore considered only accidental errors. As was stated at the beginning of the paper, no amount of analysis of a single sample, regardless of how large it is, can of itself lead one to suspect the presence of constant errors.

The parent population of errors, and any sample therefrom, is one of accidental errors only. The mean of the parent population is not necessarily the true value of the thing being measured; it is displaced by an amount equal to the sum of all the constant errors that happen to be operating. Only by considering several sets of observations (samples) from different arrangements of apparatus or from different laboratories, but supposedly made on the same unknown magnitude or on the same function, can statistical tests indicate the presence of constant errors.

A large portion of the work that has been done in mathematical statistics during the last few years has been directed toward the problem of several samples, or toward the more general

problem presented by observations on points in the plane or in space when the true coordinates would supposedly satisfy a given functional relation. Statistical methods, together with the necessary tables and charts for facilitating computation, have been devised from the results of recent advances in theory for getting a quantitative answer to the important question of how well or how poorly a proposed law of physics is substantiated by experiment. This question, as far as statistics goes, is closely related to the detection of constant errors.

The theory and the method for handling several samples is a more general problem, but not necessarily a more difficult one, than the treatment of a single sample. In order that safe conclusions may be drawn from several series of observations, it is essential that each series receive correct statistical treatment, or none at all. It follows that although a single sample cannot by itself lead to the detection of constant errors either with correct or incorrect treatment, the statistics of a single sample must form the background for the interpretation of several samples. The present paper is the result of an attempt to gather the elements of the statistics of a single sample into one place for ready reference, in order to promote the study of general methods for the interpretation of observational data.

The authors acknowledge with pleasure their indebtedness to numerous friends for advice and encouragement. In particular there should be mentioned Dr. Walter Shewhart of the Bell Telephone Laboratories, Messrs. E. C. Molina and R. I. Wilkinson of the American Telephone and Telegraph Company (now of the Bell Telephone Laboratories), and Professor J. Robert Oppenheimer of the University of California. It is, of course, to be understood that the authors assume full responsibility for all statements made in the paper.