JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 284 DECEMBER 1958 *Volume 53*

'STUDENT' AND SMALL SAMPLE THEORY

B. L. WELCH
University of Leeds, England

This year marks the Fiftieth Anniversary of the publication of "Student's" distribution. It is an appropriate time to reconsider the impact of this part of "Student's" work on the development of statistics.

1. INTRODUCTION

WHEN he died in 1937 W. S. Gosset occupied an enviable position among statisticians. He was universally respected for the originality of his statistical work and for the attractive way in which he presented it. The contributions which are now best remembered when we allude to 'Student'—using the pseudonym under which he wrote—are the early papers on theoretical topics. He was, however, equally admired for the nice sense of proportion which governed all his statistical reasoning. This sense was evident in the valuable suggestions which he made concerning the conduct of experiments and surveys, looking ahead always to an eventual statistical analysis which would be both simple and informative. The many-sided nature of the man is apparent to anyone who glances, however casually, through the volume of his collected papers [14] and was brought out clearly by a number of writers when he died (c.f. particularly Pearson [10]). It is difficult to add much to this general picture but I intend to refer to certain aspects which are made topical by the fact that just fifty years have elapsed since the publication of the two well-known papers (i) "The probable error of a mean" and (ii) "The probable error of a correlation coefficient." I propose to discuss these two 1908 papers but, in doing so, I shall take for granted some familiarity with their main contents and try to see them against the background of the contributions made by other authors to related problems. This will necessitate some brief description of inverse probability arguments although, as we shall see, Gosset's work was ultimately to strengthen the reaction against this approach to statistical inference.

2. THE THEORY OF ERRORS

A large number of books on the reduction of observations were written in the later decades of the nineteenth century, most of them aiming to illustrate and thus make more accessible the very general computational methods published by Gauss in 1821–6 [7]. Their object was to show how to obtain from scientific observations estimates of physical quantities together with indications of their reliability. It had become usual to express precision in terms of "proba-

ble errors" and most authors made at least some brief attempt to say how a "probable error" was to be interpreted in terms of probability theory. A typical statement, for instance, is the following from an American text by W. W. Johnson [9, p. 50].

> "The probable error of a final result is frequently written after it with the sign $\pm$. Thus, if the final determination of an angle is given as 36° 42′ .3 $\pm$ 1′ .22, the meaning is that the true value of the angle is exactly as likely to lie between the limits thus assigned (that is, between 36° 41′ .08 and 36° 43′ .52) as it is to lie outside of these limits."

In thus asserting the equal likelihood that a "true" value will be contained within or excluded from the assigned range, writers on the theory of errors almost invariably had in mind a hypothetical long run of repetitions, consisting not only of the results one might obtain by repeating the measurements on the same true quantity but also by measuring other true quantities of a similar nature. Out of this global set of hypothetical repetitions one might then, in theory, construct the sub-set for which the measurements are identical with those actually realized in the investigation under review. The assertion then made was that, *in this sub-set*, on 50 per cent of occasions the true quantity being measured would lie in the range calculated with the aid of the probable error formula.

That the accuracy of such probability statements depends among other things on prior assumptions about the distribution of the true values being presented for measurement was of course well known; it was also realized that, if the number of measurements made is small, as it is very apt to be in practice, a change in the prior assumption can seriously alter the probability which should be associated with the calculated limits. Since, however, this does not seem to have been regarded as a matter of great importance, one must conclude that contemporary scientific users of the method of least squares were as a rule content with it simply as a very convenient method of estimation. They were happy to have standard errors (or probable errors) as indicating, in a broad comparative way, the merits of the estimates obtained, but the exact expression of probability, derived from an application of the Bayes-Laplace method, must to the majority of them have been of secondary importance.

The position of the application of inverse probability theory has not, in the opinion of the present writer, changed much down to the present day, despite the illuminating attempts by Jeffreys [8] to put the choice of prior distribution functions on a rational basis and to remove the whole theory from the context of a frequency interpretation of probability.

3. SMALL SAMPLE THEORY OF THE MEAN

In the presence of this overriding uncertainty about prior distribution functions of true values, writers approaching the subject from this viewpoint could comfortably overlook several other difficulties. In particular it was customary to derive a standard error of a mean by using the observed minimized sum of squares, but in calculating probable error therefrom the random fluctuation of this quantity was ignored. For in fact if the data were sufficiently sparse for this fluctuation to matter, the assumed prior distribution for the true values would

at the same time become of critical importance. This was realized by no one better than F. Y. Edgeworth, who nevertheless did in 1883 still consider it worth while to develop the small sample theory of a mean value to some extent [2]. In describing what he had to say in this context I shall translate his remarks from the language of the theory of errors to that of present-day statistics, but trust that otherwise I shall not alter the sense.

Suppose $x_1, x_2, \cdots x_n$ are independent Gaussian variables with expectation μ and standard deviation σ. Denoting them collectively by S, we may write their distribution, for given μ and σ, as

$$f(S \mid \mu, \sigma)dS = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-2^{-1}\sigma^{-2}\sum (x-\mu)^2\right\}dx_1 \cdots dx_n. \quad (1)$$

Suppose also that we are given a prior distribution $g(\mu, \sigma)\, d\mu\, d\sigma$, for μ and σ. Then the *joint* distribution of sample and parameters is

$$h(S, \mu, \sigma)dSd\mu d\sigma = f(S \mid \mu, \sigma)g(\mu, \sigma)dSd\mu d\sigma. \quad (2)$$

The posterior distribution of μ and σ, *given* a realized S, is therefore

$$p(\mu, \sigma \mid S)d\mu d\sigma \propto h(S, \mu, \sigma)d\mu d\sigma \quad (3)$$

(the constant of proportionality being obtained by integrating out over μ and σ). The posterior distribution of μ alone is then

$$p(\mu \mid S)d\mu = \int_{\sigma=0}^{\infty} p(\mu, \sigma \mid S)d\mu d\sigma. \quad (4)$$

For the prior distribution, $g(\mu, \sigma)\, d\mu\, d\sigma$, Edgeworth assumed the form $C\sigma^{-2}d\mu d\sigma$ which follows by taking μ and σ to be independent and the precision constant $h = (2^{-1/2}\sigma^{-1})$ to have a uniform distribution. Making the necessary substitutions [2, p. 367] he arrived at the equation

$$p(\mu \mid S)d\mu = K\left\{1 + n(\bar{x}-\mu)^2/\sum (x-\bar{x})^2\right\}^{-(n+1)/2}d\mu. \quad (5)$$

On writing $t = \sqrt{n(n-1)}(\bar{x}-\mu)/[\sum(x-\bar{x})^2]^{1/2}$, this yields

$$p(t \mid S)dt \propto \left\{1 + t^2/(n-1)\right\}^{-(n+1)/2}dt. \quad (6)$$

If n is large we may expand (6) in powers of $(n-1)^{-1}$ to give

$$p(t)dt \propto \exp\left\{-t^2/2 + (t^4 - 4t^2)/4(n-1) + \text{etc.}\right\}dt. \quad (7)$$

Edgeworth termed (5) a sub-exponential distribution and noted that the factor needed to give the "probable error" now differs from the standard Gaussian multiple, although, as equation (7) shows, with large enough n there is no difference. Since one has in practice to deal with small groups of observations it might appear that one should attach great importance to equation (5). Edgeworth never did so, however, because he realized that a change in the assumed form of $g(\mu, \sigma)$ would have decisive influence. In equation (7) the corrective term $(t^4-4t^2)/4(n-1)$ to the large sample result would have to be replaced by something else if $g(\mu, \sigma)$ were altered. Unless, therefore, we possess, as Edgeworth did not, some unequivocal method of deciding upon $g(\mu, \sigma)$, we are not much further forward and the use of the Gaussian multiple, as generally practised, could scarcely be subjected to severe criticism.

When we come to consider Gosset's contribution to the problem of the mean we shall find him arriving by a different route, not at the same expression as (5), but at an expression of similar form, but, even so, with a very different interpretation. The present paper of Edgeworth was not, however, known to Gosset—it has indeed been largely overlooked by statisticians—and we can only speculate what his reaction might have been had he seen it.

4. THE CORRELATION COEFFICIENT

At this point it will be convenient also to describe in a very formal way the general inverse probability approach to the problem of the correlation coefficient.

Suppose (x_i, y_i) are now n pairs of random variables, independent as between pairs, but each pair following the normal bivariate distribution with means μ_x, μ_y, variances σ_x^2, σ_y^2 and covariance $\rho\sigma_x\sigma_y$. Again denoting the whole sample by S, we have therefore as its distribution for given values of the population parameters:

$$f(S \mid \mu_x, \mu_y, \sigma_x, \sigma_y, \rho)dS = \exp\left[-\frac{1}{2(1-\rho^2)}\sum\left\{\frac{(x_i-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\sigma_y} + \frac{(y_i-\mu_y)^2}{\sigma_y^2}\right\}\right] \times \left\{\sigma_x\sigma_y\sqrt{1-\rho^2}\right\}^{-n}(2\pi)^{-n}dx_1dy_1 \cdots dx_ndy_n. \tag{8}$$

Then, if the prior distribution function of the parameters is $g(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)d\mu_x\,d\mu_y d\sigma_x\,d\sigma_y\,d\rho$, the joint distribution of sample and parameters is given by

$$h(S, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = f(S \mid \mu_x, \mu_y, \sigma_x, \sigma_y, \rho)g(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho) \tag{9}$$

where the differential elements $dS\,d\mu_x\,d\mu_y\,d\sigma_x\,d\sigma_y\,d\rho$ must be appended to each side of the equation. The posterior distribution of the parameters *given* S is proportional to $h(S,\mu_x,\mu_y,\sigma_x,\sigma_y,\rho)$ and hence the posterior distribution of ρ alone is

$$p(\rho \mid S)d\rho \propto \int_{\mu_x}\int_{\mu_y}\int_{\sigma_x}\int_{\sigma_y} h(S, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho)d\mu_x d\mu_y d\sigma_x d\sigma_y d\rho. \tag{10}$$

The result of this analysis for the particular case where $g(\mu_x,\mu_y,\sigma_x,\sigma_y,\rho)$ is taken to be uniform and where n is large was given in 1898 by K. Pearson and L. N. G. Filon [11]. If r is the sample correlation coefficient the posterior distribution of ρ is then normal with mean r and standard deviation $(1-r^2)n^{-1/2}$. (Pearson and Filon indeed gave the joint posterior distribution of all the parameters but this need not concern us for the moment.) Within limits we can alter $g(\mu_x,\mu_y,\sigma_x,\sigma_y,\rho)$ considerably and still obtain the same large sample result. If n is not large enough, however, the choice of the form of $g(\mu_x,\mu_y,\sigma_x,\sigma_y,\rho)$ will be critical and all the familiar objections to the method will begin to carry weight.

5. GOSSET'S DISCUSSION OF THE CORRELATION COEFFICIENT

In his 1908 paper on the correlation coefficient [12], Gosset mentions two typical questions. (i) He introduces the subject by referring to the problem of

judging whether an observed r is consistent with an assumed ρ (in his case $\rho=0$), but also (ii) he states that "we require the probability that ρ for the population from which the sample is drawn shall lie between any given limits." (Gosset actually uses R for the population correlation coefficient but I have taken the liberty of changing his R to ρ in the present quotations to conform with modern usage. The important point is that Gosset did use different symbols for population and sample statistics). He continues [12, p. 302],

> "It is clear that in order to solve this problem we must know two things: (1) the distribution of values of r derived from samples of a population which has a given ρ, and (2) the *a priori* probability that ρ for the population lies between any given limits. Now (2) can hardly ever be known, so that some arbitrary assumption must in general be made; when we know (1) it will be time enough to discuss what will be the best assumption to make, but meanwhile I may suggest two more or less obvious distributions. The first is that any value is equally likely between $+1$ and -1, and the second that the probability that x is the value is proportional to $1-x^2$: this I think is more in accordance with ordinary experience: the distribution of *a priori* probability would then be expressed by the equation $y=\frac{3}{4}(1-x^2)$.
>
> But whatever assumption be made, it will be necessary to know (1), so that the solution really turns on the distribution of r for samples drawn from the same population."

Although he does not produce the sought solution in final mathematical form Gosset, by a mixture of empirical and theoretical reasoning which has often been admired, succeeds in telling us almost as much about the distribution of r as any symbolic expression could convey. However, since he could not write down in a short convenient way the expression for $f(r \mid \rho)dr$, he was unable to take the further step envisaged in the above quotation. For to complete his solution, given a prior probability distribution $g(\rho)d\rho$, he would have had to write down the joint distribution

$$h(r, \rho)drd\rho = f(r \mid \rho)g(\rho)drd\rho \tag{11}$$

and then the posterior distribution of ρ given r

$$p(\rho \mid r)d\rho \propto h(r, \rho)d\rho \tag{12}$$

(the factor of proportionality being obtained by integrating out with respect to ρ). From (12), for given observed r, he could then have calculated the chance that ρ lies between any limits that might have been prescribed beforehand and thus have solved his second problem.

The solution of equation (12) is not, however, necessarily the same as that of equation (10) of our previous section. In (12) we are assuming, in calculating the posterior distribution of ρ, that r is the only feature of the sample that need be considered whereas formally (10) implies that we consider all the sample values, although in virtue of (8) the quantities needed reduce immediately to ($\bar{x}$, $\bar{y}$, s_x, s_y and r). To investigate this further let us write

$$f(S \mid \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = f(r \mid \rho)f(S \mid r, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) \tag{13}$$

and

$$g(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = g(\rho)g(\mu_x, \mu_y, \sigma_x, \sigma_y \mid \rho). \tag{14}$$

We shall then have from (10) and (11)

$$p(\rho \mid S) \propto h(r, \rho) \cdot \int_{\mu_x} \int_{\mu_y} \int_{\sigma_x} \int_{\sigma_y} f(S \mid r, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) g(\mu_x, \mu_y, \sigma_x, \sigma_y \mid \rho) d\mu_x d\mu_y d\sigma_x d\sigma_y \tag{15}$$

If the result of the integrations in (15) does not depend on ρ then $p(\rho \mid S)$ is the same for all S leading to a given r and therefore (10) and (12) will give the same posterior distribution of ρ. This clearly will not always be the case irrespective of the form of $g(\mu_x, \mu_y, \sigma_x, \sigma_y \mid \rho)$, but it is not difficult to make a choice of $g(\mu_x \mu_y \sigma_x \sigma_y \mid \rho)$ for which it will be the case (e.g. Jeffreys, [8], p. 152).

I am not concerned to pursue the matter beyond this point at the moment for all these assignations of prior probability have an uncomfortable air of contrivance about them and in the present situation there are other obvious grounds why r, alone, should enter into the picture. For, starting with the set of quantities ($\bar{x}$, $\bar{y}$, s_x, s_y and r), there is no other function of them which has a direct distribution depending upon ρ but not depending upon the nuisance parameters μ_x, μ_y, σ_x and σ_y in addition to ρ. Being, as we are in the situation which Gosset has in mind, completely ignorant about the values the nuisance parameters are likely to possess, any rule for making inferences about ρ which can be expressed in *exact* probability terms must therefore be based on the distribution of r alone among the sample quantities available. If there did not exist a quantity like r depending in its distribution only on the parameter ρ at issue or, if we were concerned rather with probability statements which were to be expressed in terms of *inequalities* the position might conceivably be changed, but, as it is, we are fortunate that we can here, as Gosset does, simplify at the outset and consider a single statistic r alone.

Even in Gosset's treatment there still remains the question of the prior distribution of ρ. He would have been forced to give more consideration to this if he had actually solved his main problem and found $f(r \mid \rho)$. However, as the above quotation shows, his assumed forms for $g(\rho)$ are put forward only very tentatively and he might easily have decided to dispense with them on further consideration and have tried what he could do without any assumed prior knowledge at all.

6. GOSSET'S DISCUSSION OF THE MEAN

Although there are parts of Gosset's paper on the mean [13] which, as with his treatment of the correlation coefficient, suggest an outlook based ultimately on inverse probability, there is nowhere explicit reference to the prior functions which are an indispensable item in the practical working out of such an approach and in places we see indeed a very different outlook taking shape. The major part of the paper is concerned with an investigation of the direct probability distributions of the quantities

$$s^2 = \sum (x - \bar{x})^2/n \quad \text{and} \quad z = (\bar{x} - \mu)/s \tag{16}$$

(We shall maintain Gosset's definition of s although most of us would use a divisor $(n-1)$ and prefer to discuss, instead of z, the quantity

$$t = \sqrt{n(n-1)}(\bar{x} - \mu)/[\sum (x - \bar{x})^2]^{1/2} = \sqrt{(n-1)}z \tag{17}$$

which tends to have unit standard deviation as n becomes large).

On this occasion Gosset succeeded in writing down the direct distributions sought and, in particular, he found

$$p(z)dz \propto (1 + z^2)^{-n/2}dz \tag{18}$$

As R. A. Fisher ([4], p. 81) has noted, Gosset's discussion of the particular case $n=2$ is specially interesting. In this case we have the simple expressions

$$s = |x_1 - x_2|/2 \quad \text{and} \quad z = (x_1 + x_2 - 2\mu)/|x_1 - x_2| \tag{19}$$

and

$$p(z)dz \propto (1 + z^2)^{-1}dz. \tag{20}$$

The chance that z lies between any two values z_1 and z_2 is

$$|\tan^{-1} z_2 - \tan^{-1} z_1|/\pi$$

and in particular the chance is $\frac{1}{2}$ that z lies between -1 and $+1$. Symbolically

$$\Pr\{-1 < (x_1 + x_2 - 2\mu)/|x_1 - x_2| < 1\} = \tfrac{1}{2}. \tag{21}$$

Thence we may deduce that

$$\Pr\{(x_1 + x_2) - |x_1 - x_2| < 2\mu < (x_1 + x_2) + |x_1 - x_2|\} = \tfrac{1}{2} \tag{22}$$

$$\text{i.e. } \Pr\cdot\{\mu \text{ lies between } x_1 \text{ and } x_2\} = \tfrac{1}{2}. \tag{23}$$

In Gosset's own words ([13], p. 13), where he is first broaching the question of tabulating (18), this deduction is expressed thus:

> "The table for $n=2$ can be readily constructed by looking out $\theta = \tan^{-1} z$ in Chambers's tables and then $0.5 + \theta/\pi$ gives the corresponding value.
>
> Similarly $\frac{1}{2}\sin\theta + 0.5$ gives the values when $n=3$.
>
> There are two points of interest in the $n=2$ curve. Here s is equal to half the distance between the two observations. $\tan^{-1} s/s = \pi/4$, so that between $+s$ and $-s$ lies $2\times\pi/4\times 1/\pi$ or half the probability, i.e. if two observations have been made and we have no other information, it is an even chance that the mean of the (normal) population will lie between them. On the other hand the second moment coefficient is
>
> $$\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \tan^2\theta d\theta = \frac{1}{\pi}\Big[\tan\theta - \theta\Big]_{-\pi/2}^{\pi/2} = \infty,$$
>
> or the standard deviation is infinite while the probable error is finite."

Later on, following the short table of the probability integral of z which he provides for $n=4(1)10$, Gosset again gives expression to a similar interpretation ([13], p. 20):

> "The tables give the probability that the value of the mean, measured from the mean of the population, in terms of the standard deviation of the sample, will lie between $-\infty$ and z. Thus, to take the tables for samples of 6, the probability of the mean of the population lying between $-\infty$ and once the standard deviation of the sample is 0.9622, or the odds are about 24 to 1 that the mean of the population lies between these limits.
>
> The probability is therefore 0.0378 that it is greater than once the standard deviation and 0.0756 that it lies outside ± 1.0 times the standard deviation."

In other words the table provides first the information that, for $n=6$,

$$\Pr\{z = (\bar{x} - \mu)/s < 1\} = 0.9622 \tag{24}$$

and then again we may make the transition to an equivalent statement of which μ is the subject, viz.

$$\Pr(\mu > \bar{x} - s) = 0.9622 \tag{25}$$

and by symmetry this implies

$$\Pr(\mu < \bar{x} + s) = 0.9622 \doteq 24/25. \tag{26}$$

Furthermore, although μ has become the subject of statements (23) and (26) the probability is still a direct one related to hypothetical repeated sampling from a population with fixed mean μ and standard deviation σ. (The two above quotations are indeed separated in the text by the description of a sampling experiment where, among other things, the theoretical expression (18) is tested out on empirical material consisting of 750 samples of 4 from a given population). The status of the concept of probability is not changed by the mere alterations of emphasis which Gosset makes as he proceeds in these passages from one sentence to the next, for what he is saying at this point is deduced all the time from the direct distribution of z without the intervention of any further principles.

7. APPLICATION TO SPECIFIC EXAMPLES

Although Gosset was concerned with direct probabilities in the part of his paper to which allusion has just been made, as soon as he began to apply his results to specific examples he used language that to readers at that time might easily have suggested that a posterior probability interpretation in the Bayes-Laplace sense was intended. For instance, one of his sets of data relates to the additional hours of sleep obtained by 10 patients when given a certain hypnotic drug (Treatment 1) compared with the sleep obtained without hypnotic. The individual gains were 0.7, −1.6, −0.2, −1.2, −0.1, 3.4, 3.7, 0.8, 0.0, and 2.0. These have mean $\bar{x}=0.75$ and standard deviation s (using his definition) $=1.70$. Of these figures Gosset writes (p. 20),

> "First let us see what is the probability that 1 will on the average give increase of sleep; i.e. what is the chance that the mean of the population of which these experiments are a sample is positive. $+0.75/1.70=0.44$, and looking out $z=0.44$ in the table for ten experiments we find by interpolating between 0.8697 and 0.9161 that 0.44 corresponds to 0.8873, or the odds are 0.887 to 0.113 that the mean is positive."

This is elliptic. All that Gosset's developed theory, supported by his tabulation, had shown was that

$$\Pr\{z = (\bar{x} - \mu)/s < 0.44\} = 0.887. \tag{27}$$

On making the kind of transition described in the previous section this becomes

$$\Pr(\mu > \bar{x} - 0.44s) = 0.887. \tag{28}$$

Now Gosset's statement that the chance is 0.887 that μ is positive can be obtained from (28) by substituting for $\bar{x}$ the realized value 0.75 and for s the

realized value 1.70. Thus $\mu > 0$ can be regarded as a realized value of the random inequality $\mu > (\bar{x} - 0.44s)$, which, in repeated samples, has a chance 0.887 of being satisfied. One may object that the random quantity $(\bar{x} - 0.44s)$ has been defined with the assistance of the figure 0.44 which is itself derived from the realized sample values, but this is a point which I shall not enter into, for it is in any case not certain that an interpretation on the present lines was what Gosset actually intended, despite some pointers in this direction from earlier sections. However, whatever he had in mind, there is no doubt that, by many readers, a stated chance that $\mu > 0$ would automatically be regarded as a posterior probability such as might be deduced if some prior distribution of μ and σ were available. It is perfectly true that Gosset mentions no such prior function in the present context and therefore strictly should not be suspected here of using the classical inverse probability argument. It is also true, however, that particularly at the time he was writing, users of the Bayes-Laplace method often introduced prior distributions almost by sleight of hand and he would not have been out of fashion if he had been doing something similar. This practice was relatively innocuous when large samples were available, but in Gosset's work, which was avowedly designed to deal with very small samples, a tacit and unreasoned adoption of a particular prior distribution function could have been fatal to his purpose. Possibly to prevent any chance of misunderstanding, in later papers he abandoned the present kind of statement, as far as I am aware altogether, and gave his conclusions in the form of a direct summary of the type:—if $\mu = 0$ then a value $z = \bar{x}/s$ less than that observed will occur with probability 0.887 and a greater value with probability 0.113, i.e. not sufficiently rarely to throw doubt upon the hypothesis that $\mu = 0$. Whether such statements, impeccable as they are as deductions from the initial assumptions, are in fact ever in themselves sufficient for action is arguable. But at least they have the merit of being easily understood.

8. THE NATURE OF GOSSET'S ACHIEVEMENT

I have thus far chosen to emphasize the position of these papers of Gosset in the context of the views of statistical inference current at his time. I am, however, far from wishing to imply that he himself was much concerned with any theory of inference, suggestive as some of his remarks may have been on this score. He was primarily interested to find the distribution of r and z in direct sampling from normal populations and the occasional references he makes to the problem of inference are, perhaps, no more than an acknowledgment of its existence. It is then as an extension of our knowledge of direct sampling distributions that he would have wished the 1908 papers to be assessed. If, as it happens, I have said very little above about the actual derivations, it is only because the facts are so well known. There may be readers who are not greatly impressed by the papers on account of the incompleteness of the mathematical proofs which are given, but the final verdict of mathematical statisticians will, I believe, be that they have lasting value. They have the rare quality of showing us how an exceptional man was able to make mathematical progress without paying too much regard to the rules. He fortified what he knew with some tentative guessing, but this was backed by subsequent careful testing of his

results. In this he exemplifies an attitude more common, perhaps, to mathematical innovators than they care sometimes to admit. We have become accustomed to-day to a standard of published mathematical proof which can hide rather than reveal the actual process by which discoveries are made. With Gosset on the other hand, we can almost observe his initial thinking, whilst the nature of the final proof is secondary provided only it is sufficient to convince us that the results are right.

9. SUBSEQUENT DEVELOPMENTS

The successful generalization of the 'Student' distribution which forms the basis of so much statistical work in the modern period, was, as is well known, provided by R. A. Fisher. If I may venture to express a particular preference among his papers, it is for one which was published in 1925 [3] in which the whole theory is very succinctly developed. Fisher showed that it applied to the most general situation in "least squares" where observations are interpreted as being equal to linear functions of parameters plus random normal errors whose sampling variances are proportional to known numbers (but where the actual scale of residual variance has to be estimated from the minimized sum of squares). Also about this time Fisher published the book [5] which was to become a classic and which exploits the "general linear hypothesis" in a variety of experimental situations very different from the typical ones encountered in physics and astronomy. In this field of application Gosset also had made a great deal of the running but the conduct of the general advance now lay in Fisher's hands, and the impetus which he then gave to the subject is far from being exhausted.

Also should be mentioned further work on the correlation coefficient. As we noted above, Gosset did not succeed in discovering an explicit form for the distribution of r in normal samples. Fisher in 1915 [6] was, however, successful, although the complexity of the result was such that it was not surprising that Gosset's unorthodox approach had failed to reveal it. Later Fisher was to provide also simplifying approximations to the distribution and to make the generalization to partial correlations and, in effect, to write yet another chapter in the history of the development of statistical methods.

10. EFFECTS OF NON-NORMALITY

Anyone who works in this field of normal small sample theory must reflect at some stage on the importance or otherwise of the "assumption" of normality in the populations sampled. The reasons which have been put forward from time to time for making this assumption are not wholly convincing but are worthy of some notice:

(i) It is said that many empirical populations are in fact Gaussian. We can, I think, accept that to a good approximation this is so, or becomes so, by some simple transformation of the variables. Nevertheless the onus would always seem to be on the experimenter to produce positive evidence that the Gaussian assumption is in a general way applicable in the particular field in which he is operating.

(ii) Gosset expressed the opinion that there might be fields of inquiry where

skewed populations are expected but where the direction of skewness would not be known beforehand. He seemed to envisage the possibility of a distribution of skewness (equally likely positive and negative)—whether in an actual superpopulation or just in some logical sense is not clear—and by this means the normal theory distribution of z might be maintained (c.f. Letter quoted by E. S. Pearson, [10], p. 245). This, as it stands, is too vague to be of much assistance.

(iii) We often wish to draw deductions of a symmetrical kind (e.g. confidence limits equally spaced about the mean) and moderate skewness in the population does not affect such statements seriously. This is true but almost as often we wish to make "one-tailed" statements which are affected by skewness.

(iv) There is as a rule no other simpler mathematical assumption than the Gaussian in better accordance with the empirical facts for which at the same time the sampling theory has been worked out in such complete form. Here, perhaps, we are coming close to the real reason why normal theory holds the position it does, but it is not a reason which is convincing to a person who questions the necessity of using any small sample theory at all.

In trying to give weight to these pros and cons it may be helpful to recall just exactly how the presence of skewness in a population does influence the distribution of the quantity $t=\sqrt{n}(\bar{x}-\mu)/s$, using now our standard definition. We may note firstly that, if n is very large, t tends to be normally distributed irrespective of the form of the population sampled (excepting some extreme cases). If n is only moderately large, however, M. S. Bartlett [1] has shown that the distribution of t differs from the standard normal theory t distribution by an amount of order $n^{-1/2}$. But the normal theory t itself differs from the unit Gaussian distribution only by an amount of order n^{-1}. If, therefore, we decide to ignore the influence of skewness on the t distribution we might well go further and act as if t were unit Gaussian. If this position were accepted we would, of course, be returning to the use of the ubiquitous figure 0.67449 for determining probable error from an estimated standard error, despite the fact that the latter may be based on only a moderate number of degrees of freedom. The normal theory t-multiple will undoubtedly constitute a refinement if we are *actually* sampling from a Gaussian population, but otherwise, it is difficult to see how we can press its use upon recalcitrant statisticians who say that they have no confidence that their data are Gaussian and that therefore, for simplicity, they are content to use with small samples the multiples which they know are at least valid with large ones. We may reply that there is little chance of making things much worse by using a normal theory t-multiple rather than the unit Gaussian multiple, but we can give no positive assurance that there will be gain.

11. CONCLUSION

The expression of inferences from sample to population means in terms of probability has never been free from an admixture of arbitrary elements, e.g. (i) the nature of the law of "facility of error" (or the form to be assumed for the population distribution), and (ii) the nature of prior functions in the Bayes-Laplace method. It was early shown that the effects of this arbitrariness dis-

appeared if samples were large enough, but the development of a specific small sample theory was still inhibited. In the modern period, often dated from 1908, we have seen a gradual abandonment of inverse probability arguments and attempts to confine conclusions to those which may be deduced from direct distributional facts. Whether we believe, with some, that inverse probability has finally been scotched or, with others, that a stroke of inverse probability will always be required at some point, we must note that a large part of the development of the normal small sample theory, at least in the twenty years following 1908 when the immediate influence of Gosset was being felt, was rendered possible by the removal from the argument of the arbitrariness associated with the postulation of particular prior distribution functions of parameters. Without prejudging the success or otherwise of these developments as providing a sufficient basis for probability inference, and without attempting to evaluate what has been written on inference since 1928, we can still unreservedly commemorate in Gosset a man who played an outstanding part in contributing to our understanding of these questions.

The source of arbitrariness associated with the assumption of normality in the population remains, however, whatever our general views on inference may be. The standard 'Student' theory is an unqualified improvement on large sample theory only if the populations sampled are close to the Gaussian form.

REFERENCES

[1] Bartlett, M. S., "The effect of non-normality on the t distribution," *Proc. Camb. Phil. Soc.*, 31 (1935), 223–31.
[2] Edgeworth, F. Y., "The method of least squares," *Phil. Mag.*, 16 (1883), 360–75.
[3] Fisher, R. A., "Applications of "Student's" distribution," *Metron*, 5 (1925), 3–17.
[4] Fisher, R. A., *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd, 1956.
[5] Fisher, R. A., *Statistical Methods for Research Workers*. Edinburgh, Oliver and Boyd, 1925.
[6] Fisher, R. A., "The frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, 10 (1915), 507–21.
[7] Gauss, C. F., "Theoria combinationis observationum erroribus minimis obnoxiae, *Werke*, IV (1821–6), 1–93.
[8] Jeffreys, H., *Theory of Probability*, 2nd ed. Oxford: Clarendon Press, 1948.
[9] Johnson, W. W., *The Theory of Errors and Method of Least Squares*. New York: John Wiley and Sons, 1892.
[10] Pearson, E. S., " 'Student' as Statistician," *Biometrika*, 30 (1939), 210–50.
[11] Pearson, K. and Filon, L. N. G., "On the probable errors of frequency constants and on the influence of random selection on variation and correlation," *Phil. Trans. Roy. Soc.*, A 191 (1898), 229–311.
[12] 'Student,' "Probable error of a correlation coefficient," *Biometrika*, 6 (1908), 302–10.
[13] 'Student,' "The probable error of a mean," *Biometrika*, 6 (1908), 1–25.
[14] *"Student's" Collected Papers*. (Edited by E. S. Pearson and John Wishart with a foreword by L. McMullen. Issued by Cambridge University Press for *Biometrika*.)