Note on Dr Burnside's recent paper on errors of observation. By Mr R. A. FISHER, Fellow of Gonville and Caius College.

[Received 16 July 1923.]

That branch of applied mathematics which is now known as Statistics has been gradually built up to meet very different needs among different classes of workers. Widely different notations have been employed to represent the same relations, and still more widely different methods of treatment have been designed for essentially the same statistical problem. It is therefore not surprising that Dr Burnside* writing on errors of observation in 1923 should have overlooked the brilliant work of "Student" in 1908† largely anticipates his conclusion.

Student's work is so fundamental from the theoretical stand- point, and has so direct a bearing on the practical conclusions to be drawn from small samples, that it deserves to be far more widely known than it is at present.

A set of n observations is regarded as a random sample from an indefinitely large population of possible observations, which population obeys the normal, or Gaussian, law of error, and is therefore characterised by two parameters, m, the mean, and σ, the standard deviation. The latter is related to the "precision constant," h, by the equation

$$h = \frac{1}{2\sigma^2}$$

and it is a matter of indifference, provided we steer clear of all assumptions as to a priori probability, which parameter is used. The frequency of observations in the range dx is given by

$$df = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx.$$

It is essential to remember that both m and σ are necessarily unknown; all that is known is the set of observations $x_1, x_2, \ldots x_n$. From these certain statistics may be calculated, which may be regarded as estimates of the unknowns, but are not to be confused with, or substituted for, them. For the normal distribution we have the two familiar statistics

$$\bar{x} = \frac{1}{n} S(x)$$
$$s^2 = \frac{1}{n} S(x - \bar{x})^2$$

For each sample of n observations we shall obtain generally a differnt pair of values of x and s. In order to draw correct con- clusions from any observed pair of values, it is necessary to know how these values are distributed in different samples from a single population.

*W. Burnside (1923), "On errors of observation," *Proceedings of the Cambridge Philosophical Society* 21, pp. 482–7.

†Student (1908), "The probable error of a mean," *Biometrika*, 6, pp. 1–25.

If we regard the observations x_1, x_2, $\dots x_n$ as coordinates in n-dimensional space, any set of observations will be represented by a single point, and the frequency element, in any volume element $dx_1\, dx_2\, \dots\, dx_n$, will be

$$\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{1}{2\sigma^2}S(x-m)^2} dx_1\, dx_2\, \dots\, dx_n.$$

This may be expressed in terms of the statistics $\bar{x}$ and s by recognising the geometrical meaning of these two quantities, for if P be the point $(x_1, x_2, \dots x_n)$, and PM be drawn perpendicular to the line

$$x_1 = x_2 = \dots = x_n$$

then PM will lie in the "plane" space, determined by $\bar{x}$,

$$S(x) = n\bar{x},$$

and M will be the point $(\bar{x}, \bar{x}, \dots \bar{x})$.

Hence we see that $\bar{x}$ is constant in plane regions perpendicular to a fixed straight line, and the distance of M from the origin is $\bar{x}\sqrt{n}$; also that the distance PM is $s\sqrt{n}$, so that, for given values of $\bar{x}$ and s, P lies on a sphere in $n-1$ dimensions, of radius proportional to s; therefore the volume corresponding to $d\bar{x}ds$ will be proportional to

$$s^{n-2} ds\, d\bar{x}$$

and will be a region of constant density, proportional to

$$e^{-\frac{1}{2\sigma^2}S(x-m)^2}$$
$$= e^{-\frac{n}{2\sigma^2}(\bar{x}-m)^2} \,.\, e^{-\frac{ns^2}{2\sigma^2}}.$$

The frequency with which $\bar{x}$ and s fall into assigned elementary ranges $d\bar{x}$, ds is therefore proportional to

$$e^{-\frac{n}{2\sigma^2}(\bar{x}-m)^2} d\bar{x} \,.\, s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds.$$

from which it appears that the distribution of the two quantities is wholly independent, that of x being

$$df = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{n}{2\sigma^2}(\bar{x}-m)^2} d\bar{x} \qquad \text{(I)}$$

and that of s

$$df = \frac{n^{\frac{1}{2}(n-1)}}{2^{\frac{1}{2}(n-3)} \,.\, \frac{n-3}{2}!} \frac{s^{n-2}}{\sigma^{n-1}} e^{-\frac{ns^2}{2\sigma^2}} ds \qquad \text{(II)}$$

It will be observed that the distributions both of $\bar{x} - m$ and of s depend upon σ, and, if σ is unknown, are not of direct service; but in statistical practice,

including the practices ordinarily applied to errors of observation, it is the ratio of these two quantities which is of importance. If now

$$z = \frac{\bar{x} - m}{s}$$

we may substitute sz for $\bar{x} - m$, and $s\,dz$ for $d\bar{x}$, so that the simultaneous distribution of s and z is

$$df = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}(n-2)}.\frac{n-3}{2}!\sqrt{\pi}} \frac{s^{n-1}}{\sigma^n} e^{-\frac{ns^2}{2\sigma^2}(1+z^2)} ds$$

and integrating with respect to s from 0 to ∞, we have for the distribution of z

$$df = \frac{\frac{n-2}{2}!}{\frac{n-3}{2}!\sqrt{\pi}} . \frac{dz}{(1+z^2}^{\frac{1}{2}n} \qquad \text{(III)}$$

The distributions of s, (II), and of z, (III), were given by Student in 1908.

The traditional treatment of the probable error of the mean depends upon the distribution of $\bar{x}$, (I). The mean varies about its population value, m, in a normal distribution, with standard deviation $\sigma/\sqrt{n}$. If, therefore, σ were known, we could accurately assign to $\bar{x}$x the probable error, $\cdot 6745\sigma/\sqrt{n}$, and test whether the observed value, $\bar{x}$, were in accord with any hypothetical value, m, by means of the probability integral of the normal curve

$$P = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt, \qquad x = \frac{(\bar{x} - m)\sqrt{n}}{\sigma}.$$

But if, in fact, σ is not known, and we only have an estimate of σ, such as s, then the above reasoning collapses, for the distribution of

$$\frac{\bar{x} - m}{s} = z$$

is not a normal distribution; the "probable error," whether calculated as the quartile distance, or as a conventional multiple of the standard deviation, ceases to supply a test of the significance of the departure of x from its hypothetical value, m. Such a test is supplied by the probability integral of the Type VII curve, which gives the actual distribution of z, that is by

$$P = \int_z^\infty \frac{\frac{n-2}{2}!}{\frac{n-3}{2}!\sqrt{\pi}} . \frac{dt}{(1+t^2}^{\frac{1}{2}n}$$

Tables of this integral, for different values of z and n, have been given by Student[‡] in 1917. Fuller tables are now in course of preparation. The slight difference between the above formula and that given by Dr Burnside is traceable

[‡]Student (1917), "Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn," *Biometrika*, 11, pp. 414–17.

to Dr Burnside's assumption of an *a priori* probability for the precision constant, whereas Student's formula gives the actual distribution of z in random samples.

[From *Proceedings of the Cambridge Philosophical Society* **21** (1923), 655–658, reprinted in *Collected Papers* **1**, 455–458.]