On the Value of a Mean as Calculated from a Sample.

By L. Isserlis, D.Sc., *West Ham Municipal Technical Institute.*

§ 1. We are frequently compelled to determine the mean value of a certain character from an examination of a sample only of the whole aggregate of individuals possessing the character. The sample must of course be a random one, and the result obtained will vary from random sample to random sample. It is therefore of great practical importance to determine as precisely as we can the nature of the variation in the results, or in the language of mathematical statistics—the *frequency distribution* of the values of the average in various samples. Now it is well known that when the total population is infinite the frequency distribution of the mean approaches the normal or Gaussian distribution as the size of the sample increases, and this independently of the normality or otherwise of the sampled population.* When the sampled population is itself normal and the size of the sample is negligible in comparison with that of the whole population, the frequency distribution of values of the mean in the sample is a Gaussian one with a standard deviation approximately equal to that of the sampled population divided by the square root of the number of individuals composing the sample.

The problem to be discussed may be stated in the following form, which with some slight modifications is due to Professor A. L. Bowley.†

Let $\bar{x}$ be the mean value of the character in the whole population N. *Consider a sample of size n and suppose that n/*N *is finite, and* N *large. The values of the mean in samples of size n have a certain (unknown) curve of frequency. If n/*N *were negligible, the chance that the mean as deduced from a sample, would differ from $\bar{x}$ by as much as d would be*

$$2\int_d^\infty \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-x^2/2\sigma^2}\, dx,$$

*where $\sigma\sqrt{n}$ is the standard deviation of the universe. It is required to investigate how this chance is altered if n/*N *be not neglected.*

§ 2. We shall not, in what follows, assume that the sampled population is itself normal, but it will appear that the special case of

* *Cf.* Henderson, R., *Journal Inst. of Actuaries*, Vol. 41.

† I must express my obligation to Professor Bowley both for drawing my attention to the importance of the problem in the theory of random sampling and for his very pertinent suggestions with regard to the nature of the results to be expected.

normality yields a result of the greatest practical importance, as the simple value of the chance stated above, then differs inappreciably from the correct value, for any possible size of n/N, even when N is only moderately large.

Adopting Prof. Pearson's well-known notation, let , β_2 be the constants specifying the type of the frequency distribution of the whole population and σ its standard deviation. Let B_1, B_2 and Σ denote the same constants for the frequency curve of values of the mean in samples of size n extracted out of the total population N. We make no restrictions, such as that n/N should be small, or N large, nor that the latter population is normal. Let z be the amount by which the mean of the sample differs from the mean of the whole population; z will vary from one random sample to another. In a large number of such samples the mean values of z, z^2, z^3 and z^4 will have values calculated by the present writer in a previous paper.*

Mean value of $z = 0$

Mean value of $z^2 = \chi\mu_2/n$.

Mean value of $z^3 = \chi\chi'\mu_3/n^2$.

Mean value of $z^4 = \dfrac{3\chi}{n^2}\left(\phi + 2\dfrac{\phi'}{n}\right) + \dfrac{\chi\phi'}{n^3}\,(\mu_4 - 3\mu_2^2)$.

Here μ_2, μ_3, μ_4 are the moment coefficients of the distribution of the total population taken about the mean, μ_2 for instance being what we denoted above by σ^2 and μ_3^2/μ_2^3 being equal to β_1 while μ_4/μ_2^2 is equal to β_2. The other constants are defined in terms of n and N by the equations—

$$\chi = \frac{N-n}{N-1},\ \chi' = \frac{N-2n}{N-2},\ \chi'' = \frac{N-3n}{N-3}$$

$$\phi = 2 - 4\chi' + 3\chi'' + 4\chi'/n - 6\chi''/n$$

$$\frac{\phi'}{n} = -1 + 3\chi' - 2\chi'' - 3\chi'/n + 4\chi''/n.$$

It follows that the constants of the required frequency distribution of the mean are given by

$$\Sigma^2 = \chi\mu_2/n = \chi\sigma^2/n$$

or

$$\Sigma = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}} \qquad (1)$$

This is the correct value of this important constant, the value $\sigma/\sqrt{n}$ frequently employed instead is an approximation which assumes both that n/N is very small and that N is very great †

$$B_1 = \frac{\chi^2\chi'^2\mu_3^2/n^4}{\chi^3\mu_2^3/n^3}$$

* *Roy. Soc. Proc. A.* Vol. 92. 1915, p. 32.

† This result is not new, and is given by Pearson. *Biometrika.* Vol. V, p. 183.

or
$$B_1 = \frac{\chi'^2}{\chi}\beta_1/n \tag{2}$$

The corresponding approximation gives in this case
$$B_1 = \beta_1/n.$$

Finally,
$$B_2 = \frac{\frac{3\chi}{n^2}\left(\phi+2\frac{\phi'}{n}\right)\mu_2{}^2+\frac{\chi\phi'}{n^3}(\mu_4-3\mu_2{}^2)}{\chi^2\mu_2{}^2/n^2}$$

or
$$B_2 = \frac{3\left(\phi+2\frac{\phi'}{n}\right)}{\chi}+\frac{\phi'}{n\chi}(\beta_2-3) \tag{3}$$

When N is very great and n/N is small, this becomes
$$B_2-3=\frac{\beta_2-3}{n}.$$

Since $\beta_1=0$, $\beta_2=3$ are the conditions for normality the approximate forms show that the required frequency curve is approximately normal when the distribution of the original population is so, in the case when n/N is negligible; and further that it tends to normality for *large* samples even when the original distribution is not Gaussian.

§ 3. To deal with the general case we begin by expressing the constants of the distribution of the mean directly in terms of n, N and the constants σ, β_1, β_2 of the sampled population. Using the values of χ, χ', χ'' given above we find—
$$\chi'^2/n\chi = \frac{N-1}{(N-2)^2}\left\{\frac{N^2}{n(N-n)}-4\right\}$$

so that
$$B_1 = \frac{N-1}{(N-2)^2}\left\{\frac{N^2}{n(N-n)}-4\right\}\beta_1 \tag{4}$$

The previously quoted values of ϕ and ϕ' lead to
$$\frac{3}{\chi}(\phi+2\phi'n) = \frac{3(N-1)(N-6)}{(N-2)(N-3)}+\frac{6N(N-1)}{(N-2)(N-3)}\frac{1}{n(N-n)}$$

and
$$\phi'/n\chi = \frac{N-1}{(N-2)(N-3)}\left[\frac{N(N+1)}{n(N-n)}-6\right]$$

so that
$$B_2 = \frac{3(N-1)(N-6)}{(N-2)(N-3)}+\frac{6N(N-1)}{(N-2)(N-3)}\frac{1}{n(N-n)} + \frac{N-1}{(N-2)(N-3)}\left[\frac{N(N+1)}{n(N-n)}-6\right](\beta_2-3) \tag{5}$$

We have already shown that
$$\Sigma = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{(N-n)\,n}{N-1}}\frac{\sigma}{n} \tag{6}$$

Equations (4), (5) and (6) which determine the fundamental constants of the sought for frequency distribution deserve closer examination. The manner in which the frequency curve differs from the normal in shape—what Pearson calls the *skewness* and the *curtosis*—depends only on the values of B_1 and B_2. Now equations (4) and (5) show that these are symmetrical with respect to n and $N-n$ or with regard to the size of the *sample* and that of the *remainder*. Whenever we select a sample of size n we are at the same time rejecting a sample of size $N-n$. It is therefore not surprising to find that the frequency curves for the distribution of the mean (or for that matter of any moment coefficient) of a sample and its complement should be of the same *type*. The standard deviation Σ is in another category, its magnitude is a measure of accuracy, and we see from equation (6) that the standard deviations of the sample and of its complement are inversely proportional to their populations; if we break up a population of 1,000 many times into two portions, one of size 100 and the other 900, the frequency curve deduced from the latter will have a standard deviation one-ninth of that deduced from the former, and may be said roughly, to give determinations nine times more accurate. This must be carefully distinguished from what is practically Bernoulli's theorem, that samples of 900 extracted from an infinite population will yield results *three* times more accurate than samples of 100. The form $\Sigma = \sqrt{\frac{(N-n)\,n}{N-1}}\,\frac{\sigma}{n}$ illustrates the theorem of "complementary samples," but the form $\Sigma = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}}$ gives more information when n/N is small, and illustrates Bernoulli's theorem.

§ 4. *The frequency distribution of the mean, when the sampled population is normal.*

When the sampled population is normal, we know that β_1 is zero and β_2 is three. For this case the β coefficients and the standard deviation of the frequency distribution of the mean become

$$\left.\begin{aligned} B_1 &= 0 \\ B_2 &= \frac{3(N-1)(N-6)}{(N-2)(N-3)} + \frac{6N(N-1)}{(N-2)(N-3)}\,\frac{1}{n(N-n)} \\ \Sigma &= \sqrt{\frac{(N-n)\,n}{N-1}}\,\frac{\sigma}{n} \end{aligned}\right\} \qquad (7)$$

The first term in the value of B_2 is independent of the population n of the sample. The second diminishes as n increases until $n=\frac{1}{2}N$ and then increases again. When $n=1$ the value of B_2 is found to reduce to 3 and this is also its value when $n=N-1$. When $n=\frac{1}{2}N$ we find $B_2 = \frac{3(N-1)(N-4)}{N(N-3)}$.

Thus as the size of the sample increases from 1 to $N-1$, the value of B_2 diminishes from the normal value 3 to the minimum value 3 (N—1) (N—4)/N (N—3) corresponding to half-size samples and then increases until the normal value is again reached at $n = N-1$. Now the conditions for Pearson's symmetrical Frequency Curve Type II_L.

$$y = y_o \left(1 - \frac{x^2}{a^2}\right)^m \tag{8}$$

m, positive, are $\beta_1 = 0$, $1.8 < \beta_2 < 3$. These conditions will be satisfied by our frequency distribution for all possible sizes of the sample, if the minimum value found above for β_2 is greater than 1·8. As a matter of fact

$$3\,(N-1)\,(N-4)/N\,(N-3) > 1{\cdot}8$$

so long as $N^2 - 8N + 10$ is positive, or if $N > 6$.

Our first conclusion then is, that (excepting the cases in which the sampled population contains fewer than six individuals) *the frequency distributions of the values of the mean in samples of any size* (n) *extracted out of a* NORMAL *population of any size N is a Pearsonian frequency curve of type* II_L.*

The constants of this type are—

$$m = \tfrac{1}{2}\,(5B_2 - 9)/(3 - B_2)$$
$$a^2 = \Sigma^2\, 2B_2/(3 - B_2)$$
$$y_o = \frac{n}{\sqrt{2\pi}\Sigma}\,\frac{\Gamma(\frac{3}{2}+m)}{\Gamma(m+1)\sqrt{(\frac{3}{2}+m)}}$$

In an important class of practical applications the sampled population is comparatively large and the minimum value of β_2 is then considerably greater than 1·8. For an original population of one million we find

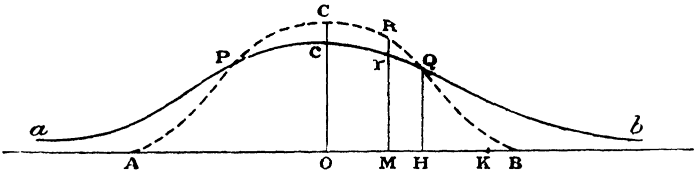
$$\text{Minimum } B_2 = 3\,(N-1)\,(N-2)/N\,(N-3)$$
$$= 2.999{,}994$$

while for N = 1000, minimum $B_2 = 2.993{,}994$. Thus for samples of size n extracted out of a normal population of a million the frequency distribution of the mean is exactly normal for samples of one and differs inappreciably from normality even in the extreme case of samples of half a million; also there are very few practical applications in which the frequency distribution for samples extracted out of a population of a thousand or more need be treated as otherwise than normal. Thus the answer to Prof. Bowley's problem in this case is that the chance of a deviation d in the value of the mean, when n/N is not neglected is for practical purposes the same as if n/N were negligible.

The more exact answer (of theoretical interest only) is to be obtained from a comparison of the normal curve corresponding to n/N negligible and the limited curve II_L obtained above.

* For the constants of this curve *cf. Phil. Trans. A.* Vol. 216, p. 449.

In the figure $a\ c\ b$ represents the normal curve, and A C B the limited range type II_L curve. For samples out of a population of a million the two curves would be indistinguishable on any reasonable scale. Nevertheless since they are of equal areas the modal ordinate of the limited curve must exceed that of the normal curve. Hence the curves must intersect in two points P, Q. Now if RrM



be any ordinate between CcO and QH the area under the limited curve between these ordinates must be greater than the area under the normal curve. This will hold for all ordinates up to one through a certain point K (to the right of H). If $\mathrm{OK} = k$, we can say that the chance of any deviation d between the value of the mean in the sample and its value in the general population is *increased* by treating n/N as negligible if $d < k$, *i.e.*, for *moderate* errors, and is diminished if $\mathrm{d} > k$.

§ 5. *The frequency distribution of the mean when the sampled population is not normal.*

In the general case the values of B_1, B_2 and Σ for the frequency distribution of the mean are given by formulæ (4), (5) and (6) of § 3. Let us consider the case in which the total population N is considerable and suppose very small samples to be excluded. Thus N is large and n not small and n/N is not to be treated as negligible. If we neglect squares and higher powers of $1/\mathrm{N}$, but make no other approximation, our formulæ become

$$\mathrm{B}_1 = \frac{1}{\mathrm{N}}\left[\frac{1}{\frac{n}{\mathrm{N}}\left(1-\frac{n}{\mathrm{N}}\right)}-4\right]\beta_1 \tag{9}$$

$$\mathrm{B}_2-3 = \frac{1}{\mathrm{N}}\left[\frac{1}{\frac{n}{\mathrm{N}}\left(1-\frac{n}{\mathrm{N}}\right)}-6\right](\beta_2-3) \tag{10}$$

$$\Sigma^2 = \left(1-\frac{n-1}{\mathrm{N}}\right)\frac{\sigma^2}{n} \tag{11}$$

It follows that for large values of N the distribution of the mean is approximately normal even when the distribution of the sampled population is decidedly skew. To get a rough idea of how nearly normal this distribution remains in the most extreme cases, let us consider a universe whose abnormality of distribution is almost out of the range of experience. We will suppose that for the sampled population $\beta_1=10$ and $\beta_2-3=20$. A glance at Rhind's diagram,* or at Pearson's recent extension of it,† will show that such abnormality is altogether outside the ordinary run of frequency distributions.

Let $N=10^6$ and $n/N>1/1000$ then with $\beta_1=10$ and $\beta_2=23$

$$B_1 = \frac{1}{N}\left[\frac{1}{\frac{n}{N}\left(1-\frac{n}{N}\right)}-4\right]\beta_1$$
$$< .01$$

while
$$B_2-3 = \frac{1}{N}\left[\frac{1}{\frac{n}{N}\left(1-\frac{n}{N}\right)}-6\right](\beta_2-3)$$
$$< \cdot 02$$

Or if samples of over a thousand be extracted out of a population of a million, then even if the sampled population be extremely abnormal, the frequency distribution of the mean as obtained from random samples will not differ from a normal distribution more than is the case for a frequency curve where β_1 and β_2 are $\cdot 01$ and $3\cdot 02$ respectively. If the sample is bigger than 10,000 in the above case, the corresponding values of β_1 and β_2 are $\cdot 001$ and $3\cdot 002$.

The frequency curves are of the same type II as in the case already considered, in which the sampled population was normal, and we are again entitled to state the solution of our problem as follows :—

If samples of size n are taken from a population of size N, whose distribution is not normal, then the distribution of the values of the mean has a known frequency curve (Pearson's Type II_L) provided N is large and n/N not very small. If we treat this curve as a normal one, and n/N as negligible, we obtain for the probability of a certain deviation in the value of the mean, a result which is very nearly correct ; it errs in excess for moderate deviations and in defect for large ones.

* "*Tables for Statisticians,*" p. 69.
† *Phil. Trans. A.* Vol. 216. p. 456.