ON A DISTRIBUTION YIELDING THE ERROR FUNCTIONS OF SEVERAL WELL KNOWN STATISTICS

By Mr. R. A. Fisher,
Statistical Department, Rothamsted Experimental Station, Harpenden, England.

1. Theoretical Distributions

The idea of an error function is usually introduced to students in connection with experimental errors; the normal curve itself is often introduced almost as if it had been obtained experimentally, as an error function. The student is not usually told that little or nothing is known about experimental errors, that it is not improbable that every instrument, and every observer, and every possible combination of the two has a different error curve, or that the error functions of experimental errors are only remotely related to the error functions which are in practical use, because these are applied in practice not to single observations but to the *means* and other *statistics* derived from a number of observations.

Many statistics tend to be normally distributed as the data from which they are calculated are increased indefinitely; and this I suggest is the genuine reason for the importance which is universally attached to the normal curve. On the other hand some of the most important statistics do not tend to a normal distribution, and in many other cases, with small samples, of the size usually available, the distribution is far from normal. In these cases tests of *Significance* based upon the calculation of a "probable error" or "standard error" are inadequate, and may be very misleading. In addition to tests of significance, tests of goodness of fit also require accurate error functions; both types of test are constantly required in practical research work; the test of goodness of fit may be regarded as a kind of generalized test of significance, and affords an *a posteriori* justification of the error curves employed in other tests.

Historically, three distributions of importance had been evolved by students of the theory of probability before the rise of modern statistics; they are

Distribution	Due to	Date
Binomial expansion,	Bernoulli	c. 1700,
Normal curve,	Laplace, Gauss	1783,
Exponential expansion,	Poisson	1837;

of these the series of Bernoulli and Poisson, although of great importance, especially the latter, are in a different class from the group of distributions with which we are concerned, for they give the distribution of *frequencies*, and are consequently discontinuous distributions.

2. Pearson's χ^2 Distribution

In 1900 Pearson devised the χ^2 test of goodness of fit. If $x_1, x_2, \ldots, x_{n'}$, are the observed frequencies in a series of n' classes, and $m_1, m_2, \ldots, m_{n'}$ the corresponding expectations, then the discrepancy between expectation and observation may be measured by calculating

$$\chi^2 = S\frac{(x-m)^2}{m} .$$

The discrepancy is *significant* if χ^2 has a value much greater than usually occurs, when the discrepancy measured is that between a random sample, and the population from which it is drawn. To judge of this we need to know the random sampling distribution of χ^2. This distribution Pearson gave in his paper of 1900. The distribution for large samples is not normal; it is independent of the actual values of $m_1, \ldots, m_{n'}$; but it includes a parameter which, according to Pearson's original exposition, was to be identified with n', the number of frequency classes. Consequently in Pearson's original table, and in the fuller table given soon after by Elderton, the table is entered with the parameter n', which can take all integer values from 2 upwards.

More recently, it has been shown that Pearson neglected, as small quantities of the second order, certain corrections, which in fact do not tend to zero, but to a finite value, as the sample is increased. These corrections are, in fact, not small at all. In consequence of this, most of the tests of goodness of fit made in the last 25 years require revision. The important point is, however, that the distributions found by Pearson still hold, if we interpret n', not as the number of frequency classes, but as one more than the number of *degrees of freedom*, in which observation may differ from expectation. For example, in a contingency table of r rows and c columns, we ought not to take

$$n' = cr$$

but

$$n'-1 \ = (c-1)\ (r-1)$$

in recognition of the fact that the marginal totals of the table of expectation have been arrived at by copying down the marginal totals of the table of observations. For instance in a 3×5 table we should put $n'=9$, and not $n'=15$. In a 2×2 table $n'=2$, not $n'=4$.

One consequence of this is that it is more convenient to take $n=n'-1$, representing the number of degrees of freedom, as the parameter of the tables; in fact, to number the tables from (say) 1 to 50 instead of from 2 to 51. The real importance of n is shown by the fact that if we have a number of quantities $x_1, \ldots, x_n$, distributed independently in the normal distribution with unit standard deviation, and if

$$\chi^2 = S(x^2),$$

then χ^2, so defined, will be distributed as is the Pearsonian measure of goodness of fit; n is, in fact, the number of independent squares contributing to χ^2. The mean value of χ^2 is equal to n.

The χ^2 distribution is the first of the family of distributions of which I will speak, and like the others it turns up more or less unexpectedly in the distributions of a variety of statistics. In a noteworthy paper in 1908, "Student" investigated the error curve of the Standard Deviation of a small sample from a normal distribution, and with remarkable penetration he suggested a form for this error curve which has proved to be exact. The relation of this curve with that of χ^2 is close; if x stand for any value of a normal sample, $\bar{x}$ for the mean, and σ for the standard deviation of the population, then

$$\chi^2 = \frac{S(x - {}^{\bar{x}})^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$$

where n, the number of degrees of freedom, is one less than the number in the sample, and s^2 is the best estimate from the sample of the true *variance*, σ^2.

Another example of the occurrence of the χ^2 distribution crops up in the study of small samples from the Poisson Series. In studying the accuracy of methods of estimating bacterial populations by the dilution method I was led to the fact that the statistic

$$\chi^2 = \frac{S(x-\bar{x})^2}{\bar{x}}$$

when x is a single observation, and $\bar{x}$ the mean of the sample, is distributed wholly independently of the true density of the population sampled; for ordinary values of $\bar{x}$, but not for very small values, it is also distributed independently of $\bar{x}$, and its distribution is that of χ^2 with n one less than the number in the sample.

A similarly distributed index of dispersion may be used for testing the variation of samples of the binomial and multinomial series. The case of the binomial is interesting to economists, in that it leads at once to a test of the significance of the Divergence-Coefficient of Lexis. In fact, the method of Lexis was completed, and made capable of exact application, from the time of the first publication of the table of χ^2. I do not think, however, that this has been observed, either by exponents of the method of Lexis and his successors, or by exponents of the test of goodness of fit.

3. The General z Distribution

The most direct way of passing from the χ^2 distribution to the more general distribution to which I wish to call attention is to consider two samples of normal distributions, and how the two estimates of the variance may be compared. We have two estimates $s_1{}^2$ and $s_2{}^2$ derived from the two small samples, and we wish to know, for example, if the variances are significantly different. If we introduce hypothetical true values $\sigma_1{}^2$ and $\sigma_2{}^2$ we could theoretically calculate in terms of σ_1 and σ_2, how often $s_1{}^2 - s_2{}^2$ (or $s_1 - s_2$) would exceed its observed value. The probability would of course involve the hypothetical σ_1 and σ_2, and our formulae could not be applied unless we were willing to substitute the observed values $s_1{}^2$ and $s_2{}^2$ for $\sigma_1{}^2$ and $\sigma_2{}^?$; but such a substitution, though quite legitimate with large samples, for which the errors are small, becomes extremely misleading for small samples; the probability derived from such a substitution would be far from exact. The only exact treatment is to eliminate the unknown quantities σ_1 and σ_2 from the distribution by replacing the distribution of s by that of log s, and so deriving the distribution of $\log s_1/s_2$. Whereas the sampling errors in s_1 are proportional to σ_1, the sampling errors of log s_1 depend only upon the size of the sample from which s_1 was calculated.

We may now write

$$n_1 s_1^2 = \sigma_1^2 \chi_1^2 = \sigma_1^2 S_1(x^2)$$

$$n_2 s_2^2 = \sigma_2^2 \chi_2^2 = \sigma_2^2 S_2(x^2)$$

$$e^{2z} = \frac{s_1^2}{s_2^2} = \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{n_2 S_1(x^2)}{n_1 S_2(x^2)}$$

where S_1 and S_2 are the sums of squares of (respectively) n_1 and n_2 independent quantities; then z will be distributed about $\log \frac{\sigma_1}{\sigma_2}$ as mode, in a distribution which depends wholly on the integers n_1 and n_2. Knowing this distribution we can tell at once if an observed value of z is or is not consistent with any hypothetical value of the ratio σ_1/σ_2.

The distribution of z involves the two integers n_1 and n_2 symmetrically, in the sense that if we interchange n_1 and n_2 we change the sign of z. In more detail, if P is the probability of exceeding any value, z, then interchanging n_1 and n_2, $1-P$ will be the probability of exceeding $-z$.

Values of special interest for n_1 and n_2 are ∞ and unity. If n_2 is infinite, S_2/n_2 tends to unity and consequently we have the χ^2 distribution, subject to the transformation

$$e^{2z} = \frac{\chi^2}{n}, \; n_1 = n$$

or

$$z = \tfrac{1}{2} \log \frac{\chi^2}{n};$$

similarly if n_1 is infinite, we have the χ^2 distribution again with

$$z_p = -\tfrac{1}{2} \log \frac{1}{n} \chi^2_{1-p}$$

the curves being now reversed so that the P of one curve corresponds to $1-P$ of the other.

In the second special case, when $n_1 = 1$, we find a second important series of distributions, first found by "Student" in 1908. In discussing the accuracy to be ascribed to the mean of a small sample, "Student" took the revolutionary step of allowing for the random sampling variation of his estimate of the standard error. If the standard error were known with accuracy the deviation of an observed value from expectation (say zero), divided by the standard error, would be distributed normally with unit standard deviation; but if for the accurate standard deviation we substitute an estimate based on n degrees of freedom we have

$$t=\frac{x\sqrt{n}}{\sqrt{S(x^2)}}$$

$$t^2=\frac{nx^2}{S(x^2)}=e^{2z}\text{ if }\begin{cases}n_1=1\\ n_2=n\end{cases}$$

consequently the distribution of t is given by putting $n_1=1$, and substituting $z=\frac{1}{2}\log t^2$.

The third special case occurs, when both $n_1=1$, and $n_2=\infty$, and as is obvious from the above formulae, it reduces to the normal distribution with

$$z=\tfrac{1}{2}\log x^2.$$

In fact, one series of modifications of the normal distribution gives the χ^2 distributions, a second series of modifications gives the curves found by "Student", while if both modifications are applied simultaneously we have the double series of distributions appropriate to z.

Like the χ^2 distribution, the distribution found by "Student" has many more applications beyond that for which it was first introduced. It was introduced to test the significance of a mean of a unique sample; but as its relation to the z distribution shows, it occurs wherever we have to do with a normally distributed variate, the standard deviation of which is not known exactly, but is estimated independently from deviations amounting to n degrees of freedom. For example, in a more complicated form it gives a solution for testing the significance of the difference between two means, a test constantly needed in experimental work.

An enormously wide extension of the applications of "Student's" curves flows from the fact that not only means, but the immense class of statistics known as regression coefficients may all be treated in the same way; and indeed must be treated in the same way if tests of significance are to be made in cases where the number of observations is not large. And in many practical cases the number is not large; if a meteorologist with 20 years records of a place wishes to ask if the observed increase or decrease in rainfall is significant, or if in an agricultural experiment carried out for 20 years, one plot has seemed to gain in yield compared to a second plot differently treated, Student's curves provide an accurate test, where the ordinary use of standard errors or probable errors are altogether misleading.

The more general distribution of z like its special cases, crops up very frequently. I found it first in studying the error functions of the correlation coefficient. If the correlation, let us say, between pairs of n brothers, is obtained by forming a symmetrical table, we obtain what is called an *intraclass* correlation. If r is such a correlation, let

$$r=\tanh z, n_1=n-1,\ n_2=n,$$

then this transformation expresses the random sampling distribution of r in terms of that of z, when n is the number in the sample.

It was the practical advantages of the transformation that appealed to me at the time. The distribution of r is very far from normal for all values of the correlation, (ρ); and even for large samples when the correlation is high; its accuracy depends greatly upon the true value of ρ, which is of course unknown, and the form of the distribution changes rapidly as ρ is changed. On the other hand the distribution of z is nearly normal even for small values of n; it is absolutely constant in accuracy and form for all values of ρ, so that if we are not satisfied with its normality we can make more accurate calculations.

The distribution shows a small but constant bias in the value of z, when r is derived from the symmetrical table; this bias disappears if instead of starting from the correlation as given by the symmetrical table we approach the matter from a more fundamental standpoint. Essentially we estimate the value of an intraclass correlation by estimating the ratio of two variances, the intraclass variance found by comparing numbers of the same class, and the variance between the observed means of the different classes. From n classes of s observations each, we have $n(s-1)$ degrees of freedom for the intraclass variance, and $n-1$ degrees of freedom for the variance of the means. From the definition of the z distribution it will obviously be reproduced by the errors in the ratio of two independent estimates of the variance, and if we estimate the variances accurately the bias in the estimate of the correlation will be found to have disappeared; it was, in fact, introduced by the procedure of forming the symmetrical table.

The practical working of cases involving the z distribution can usually be shown most simply in the form of an analysis of variance. If x is any value, $\bar{x}_p$ the mean of any class, and $\bar{x}$ the general mean, n the number of classes of s observations each, the following table shows the form of such an analysis:

Analysis of Variance

Variance	Degrees of Freedom	Sum of Squares	Mean Square
Between classes.......	$n_1=n-1$	$sS_1^n(\bar{x}_p-\bar{x})^2$	s_1^2
Within classes.........	$n_2=n(s-1)$	$S_1\ (x-\bar{x}_p)^2$	s_2^2
Total.............	$ns-1$	$S_1^{ns}(x-\bar{x})^2$	

The two columns headed Degrees of Freedom and Sum of Squares must add up to the totals shown; the mean squares are obtained by dividing the sums of squares by the corresponding degrees of freedom, then $z=\log s_1/s_2$ may be used to test the significance of the intraclass correlation, or the significance of its deviations from any assigned value.

4. Multiple Correlations

The same method may be used to solve the long outstanding problem of the significance of the multiple correlation. If y is the dependent variate, and $x_1, x_2, \ldots, x_p$ are p independent variates, all measured from their means, and if the regression of y on $x_1, \ldots, x_p$ is expressed by the equation

$$Y=b_1x_1+\ldots+b_px_p$$

such that the correlation of y with Y is R, then R is termed the multiple correlation of y with $x_1, \ldots, x_p$, and the ingredients of the analysis of variance are as follows:

Variance	Degrees of Freedom	Sum of Squares	
Of regression formula.......	p	$S(Y^2)$	$=nR^2\sigma^2$
Deviations from regression formula.................	$n-p-1$	$S(y-Y)^2$	$=n(1-R^2)\sigma^2$
Total.................	$n-1$	$S(y^2)$	$=n\sigma^2$

For samples from uncorrelated material the distribution of R may therefore be inferred from that of z, the actual curve for n observations and p independent variates being

$$df = \frac{\frac{n-3}{2}!}{\frac{n-p-3}{2}!\,\frac{p-2}{2}!}\,(R^2)^{\frac{1}{2}(p-2)}\,(1-R^2)^{\frac{1}{2}(n-p-3)}\,d(R^2)$$

which degenerates when $p=1$, into the better known distribution for a single independent variate

$$df = \frac{2\,\frac{n-3}{2}!}{\frac{n-4}{2}!\,\sqrt{\pi}}\,(1-R^2)^{\frac{n-4}{2}}\,dR,$$

a distribution first suggested by "Student"[7], which has been established since 1915.

5. The Correlation Ratio

The distribution, for uncorrelated material of the correlation ratio η, is clearly similar to that of the multiple correlation, R, and resembles the case of the intraclass correlation when the number of observations varies from class to class.

A number of values of the variate y are observed for each of a series of values of the variate x; n_p is the number observed in each array, $\bar{y}$ the mean of the observed values and $\bar{y}_p$ the mean of any array; the variance of the variate y may be analysed as follows, there being a arrays:

Variance	Degrees of freedom	Sum of Squares	
Between arrays............	$a-1$	$S\{n_p(\bar{y}_p-\bar{y})^2\}$	$=N\eta^2\sigma^2$
Within arrays.............	$S(n_p-1)$	$S(y-\bar{y}_p)^2$	$=N(1-\eta^2)\sigma^2$
Total................	$S(n_p)-1$	$S(y-\bar{y})^2$	$=N\sigma^2$

η^2 is thus distributed just like R^2. The transformation is

$$\frac{\eta^2}{1-\eta^2} = \frac{n_1}{n_2}e^{2z} = \frac{S_1(x^2)}{S_2(x^2)}\,.$$

If the observations are increased so that $n_2 \rightarrow \infty$, then

$$n_2 \frac{\eta^2}{1-\eta^2}$$

tends to be distributed in the χ^2 distribution corresponding to $(a-1)$ degrees of freedom, while for the multiple correlation with p independent variates

$$n_2 \frac{R^2}{1-R^2}$$

tends to be distributed in the χ^2 distribution with p degrees of freedom. These are two examples of statistics not tending to the normal distribution for large samples.

More important than either of these is the use of the z distribution in testing the goodness of fit of regression lines, whether straight or curved. If Y stand for the function of x to be tested representing the variation of the mean value of y for different values of x, let there be a arrays and let q constants of the regression formulae have been adjusted to fit the data by least squares, then the deviations of the observations from the regression line may be analysed as follows

Variance due to	Degrees of freedom	Sum of Squares	
Deviation of array mean from formula...........	$a-q$	$S\{n_p(\bar{y}_p - Y_p)^2\}$	$=N\sigma^2(\eta^2-R^2)$
Deviation within array.....	$N-a$	$S(y-\bar{y}_p)^2$	$=N\sigma^2(1-\eta^2)$
Total...............	$N-q$	$S(y-Y)^2$	$=N\sigma^2(1-R^2)$

where R is the correlation between y and Y. The transformation this time is

$$\frac{\eta^2-R^2}{1-\eta^2} = \frac{S_1}{S_2} = \frac{n_1}{n_2} e^{2z},$$

and if the sample is increased indefinitely

$$n_2 \frac{\eta^2-R^2}{1-\eta^2}$$

tends to χ^2 distribution for $(a-q)$ degrees of freedom.

This result is in striking contrast to a test which is in common use under the name of Blakeman's criterion, which is designed to test the linearity of regression lines, and which also uses the quantity η^2-r^2. Our formula shows that, for large samples, with linear regression,

$$(N-a)\frac{\eta^2-r^2}{1-\eta^2}$$

has a mean value $a-2$. The failure of Blakeman's criterion to give even a first approximation, lies in the fact that, following some of Pearson's work, the number of the arrays is ignored, whereas, in fact, the number of arrays governs the whole distribution.

Summary

The four chief cases of the z distribution have the following applications

I. Normal Curve	II. χ^2	III. Student's	IV. z
Many statistics from large samples	Goodness of fit of frequencies Index of dispersion for Poisson and Binomial samples Variance of normal samples	Mean Regression coefficient Comparison of means and regressions	Intraclass correlations Multiple correlation Comparison of variances Correlation ratio Goodness of fit of regressions

COMPARISON OF DISTRIBUTION FORMULAE

	Constant factor	Frequency
χ^2	$\dfrac{1}{2^{\frac{n-2}{2}}\,\dfrac{n-2}{2}!}$	$\chi^{n-1}e^{-\frac{1}{2}\chi^2}d\chi \;=\; n^{\frac{1}{2}n}e^{nz-\frac{1}{2}ne^{2z}}\,dz$
Normal	$\dfrac{2}{\sqrt{2\pi}}$	$e^{-\frac{1}{2}x^2}dx \;=e^{z-\frac{1}{2}e^{2z}}dz$
Student's	$\dfrac{\dfrac{n-1}{2}!\;\;2n^{\frac{1}{2}n}}{\dfrac{n-2}{2}!\;\;\sqrt{\pi}}$	$(n+t^2)^{-\frac{n+1}{2}}dt = \dfrac{e^z\,dz}{(e^{2z}+n)^{\frac{1}{2}(n+1)}}$
z	$\dfrac{2\,.\,\dfrac{n_1+n_2-2}{2}!}{\dfrac{n_1-2}{2}!\;\dfrac{n_2-2}{2}!}\,n_1^{\frac{1}{2}n^1}n_2^{\frac{1}{2}n_2}$	$\dfrac{e^{n_1z}dz}{(n_1e^{2z}+n_2)^{\frac{1}{2}(n_1+n_2)}}$

Addendum Aug., 1927.

Since the International Mathematical Congress (Toronto, 1924) the practical applications of the developments summarized in this paper have been more fully illustrated in the author's book *Statisiical Methods for Research Workers* (Oliver and Boyd, Edinburgh, 1925). The Toronto paper supplies in outline the mathematical framework around which the book has been built, for a formal statement of which some reviewers would seem to have appreciated the need.