SOME TESTS OF SIGNIFICANCE, TREATED BY THE THEORY OF PROBABILITY

BY HAROLD JEFFREYS, M.A., St John's College

[*Received* 1 January, *read* 11 March 1935]

It often happens that when two sets of data obtained by observation give slightly different estimates of the true value we wish to know whether the difference is significant. The usual procedure is to say that it is significant if it exceeds a certain rather arbitrary multiple of the standard error; but this is not very satisfactory, and it seems worth while to see whether any precise criterion can be obtained by a thorough application of the theory of probability.

I. CONTINGENCY

Suppose that two different large, but not infinite, populations have been sampled in respect of a certain property. One gives x specimens with the property, y without; the other gives x' and y' respectively. The question is, whether the difference between x/y and x'/y' gives any ground for inferring a difference between the corresponding ratios in the complete populations. Let us suppose that in the first population the fraction of the whole possessing the property is p, in the second p'. Then we are really being asked whether $p=p'$; and further, if $p=p'$, what is the posterior probability distribution among values of p; but, if $p \neq p'$, what is the distribution among values of p and p'.

Let h denote our previous knowledge and q the proposition $p=p'$. Then in the absence of relevant information we take the prior probabilities

$$P(q \mid h) = P(\sim q \mid h) = \tfrac{1}{2}. \tag{1}$$

This is the natural application of a principle from an earlier paper*. If we are asked a question and given no further information, the prior probabilities of the alternatives stated in the question are equal. It may be necessary to say again that this expresses no opinion about the frequency of the truth of q among any real or imaginary populations; it is simply the formal way of saying that we do not know whether it is true or not of the actual populations under consideration at the moment.

If q is true, then the prior probability of p is uniformly distributed from 0 to 1,

* Jeffreys, *Proc. Camb. Phil. Soc.* 29 (1933), 83–7.

according to the usual theory of sampling. If q is not true, then the prior probabilities of p and p' are uniformly and independently distributed. Hence*

$$P(dp \mid qh) = dp; \quad P(dp\,dp' \mid \sim q.h) = dp\,dp' \tag{2}$$

and

$$P(q, dp \mid h) = \tfrac{1}{2}dp; \quad P(\sim q.dp\,dp' \mid h) = \tfrac{1}{2}dp\,dp' \tag{3}$$

by the fundamental rule

$$P(pq \mid r) = P(p \mid r)\,P(q \mid pr). \tag{4}$$

We have now to apply the theorem of inverse probability that among a set of hypotheses p_n, on data h and θ, $P(p_n \mid h\theta)/P(p_n \mid h)\,P(\theta \mid p_n h)$ is independent of p_n. The data θ are in this case the compositions of the samples. Then

$$P(\theta \mid dp.qh) = \frac{(x+y)!}{x!\,y!}\,\frac{(x'+y')!}{x'!\,y'!}\,p^x(1-p)^y p^{x'}(1-p)^{y'}, \tag{5}$$

$$P(\theta \mid dp\,dp'.\sim q.h) = \frac{(x+y)!}{x!\,y!}\,\frac{(x'+y')!}{x'!\,y'!}\,p^x(1-p)^y p'^{x'}(1-p')^{y'}, \tag{6}$$

whence

$$P(q, dp \mid \theta, h) \propto p^{x+x'}(1-p)^{y+y'}\,dp, \tag{7}$$

$$P(\sim q, dp, dp' \mid \theta, h) \propto p^x(1-p)^y p'^{x'}(1-p')^{y'}\,dp\,dp', \tag{8}$$

the factor of proportionality being the same in both cases†. The $\tfrac{1}{2}$ and the factorial functions, being the same for all alternatives, have been cancelled.

Now each of p and p' can range from 0 to 1; hence we can estimate the posterior probabilities of q and $\sim q$ by integrating (7) with regard to p, thus adding up the contributions to the probability of q from all the admissible values of p; similarly we integrate (8) with regard to both p and p'. Then

$$P(q \mid \theta, h) \propto \frac{(x+x')!\,(y+y')!}{(x+x'+y+y'+1)!}, \tag{9}$$

$$P(\sim q \mid \theta, h) \propto \frac{x!\,y!}{(x+y+1)!}\,\frac{x'!\,y'!}{(x'+y'+1)!}, \tag{10}$$

and

$$\frac{P(q \mid \theta, h)}{P(\sim q \mid \theta, h)} = \frac{(x+x')!\,(y+y')!\,(x+y+1)!\,(x'+y'+1)!}{x!\,y!\,x'!\,y'!\,(x+x'+y+y'+1)!}. \tag{11}$$

With the further datum that the sum of the probabilities of q and $\sim q$ on any data is equal to 1, this gives the required solution.

When the numbers in the samples are large an approximation to this ratio may be useful. It can be written

$$\frac{(x+y+1)(x'+y'+1)}{(x+x'+y+y'+1)}\binom{x+y}{x}\binom{x'+y'}{x}\Big/\binom{x+x'+y+y'}{x+x'}, \tag{12}$$

where the quantities in brackets denote the numbers of combinations of $x+y$

* dp in this expression for a probability means the proposition that p lies in a *particular* range of length dp.

† This is true throughout the paper in analogous pairs of equations.

things taken x at a time, and so on. The last three factors have already been evaluated*; they give

$$\left\{\frac{(x+x'+y+y')^3}{2\pi(x+x')(y+y')(x+y)(x'+y')}\right\}^{\frac{1}{2}}$$
$$\exp\left\{-\tfrac{1}{2}\frac{(x+x'+y+y')^3}{(x+x')(y+y')(x+y)(x'+y')}p^2\right\}\left\{1+O\left(\frac{1}{x},\frac{1}{y},\frac{1}{x'},\frac{1}{y'}\right)\right\}, \quad (13)$$

where
$$p=x-\frac{(x+y)(x+x')}{x+x'+y+y'}=\frac{xy'-x'y}{x+x'+y+y'}. \quad (14)$$

Restoring the first factor in (12) we have

$$\frac{P(q\mid\theta,h)}{P(\sim q\mid\theta,h)}\sim\left\{\frac{(x+x'+y+y')(x+y)(x'+y')}{2\pi(x+x')(y+y')}\right\}^{\frac{1}{2}}$$
$$\exp\left\{-\tfrac{1}{2}\frac{(x+x'+y+y')(xy'-x'y)^2}{(x+x')(y+y')(x+y)(x'+y')}\right\}. \quad (15)$$

When $xy'-x'y$ is small this is large of the order of the square root of the numbers in the samples, and q approaches certainty; but when it is large the ratio is very small and q approaches impossibility. The theory therefore shows that a small difference between the sampling ratios may establish a high probability that the ratios in the main populations are equal, while a large one may show that they are different. This is in accordance with ordinary practice, but has not, so far as I know, been related to the general theory before. In one respect, however, there is a departure from ordinary practice. It would be natural to define a standard error of $xy'-x'y$ in terms of the coefficient of its square in the exponential; but the range of values of the exponent that make the ratio of the posterior probabilities greater than 1 is not a constant, since it depends on the outside factor, which increases with the sizes of the samples. This variability is of course connected directly with the fact that agreement between the two populations becomes more probable if the samples are large and the difference of the sampling ratios small; when the ratio is large at $xy'-x'y=0$, a larger value of the exponent is obviously needed to reduce the product to unity.

Some numerical values are given by way of illustration. In each case $x=y$, $x'+y'=x+y$, but in general $x'\neq y'$. The table gives $x+y$, the maximum value of the ratio of the posterior probabilities, and that of $x'-y'$ needed to make the ratio equal to unity.

$x+y$	$P(q)/P(\sim q)$	$x'-y'$	$(x'-y')/(x+y)^{\frac{1}{2}}$
40	3·57	14·3	2·26
100	5·65	26·4	2·64
200	7·97	40·8	2·89
400	11·3	61·5	3·07
1,000	17·8	107·3	3·39
10,000	56·4	401	4·01
100,000	178	1440	4·57

* Jeffreys, *Scientific Inference*, 1931, Lemma II, equation (8).

The ratio of the critical value of $x'-y'$ to $(x+y)^{\frac{1}{2}}$ is given in a further column to show how little it varies when the sizes of the samples change by a factor of 2500.

II. MEASUREMENT WHEN THE STANDARD ERRORS ARE KNOWN

Suppose that we have two different ways of measuring a quantity and want to know whether there is any systematic difference between the results. As before, let h denote our previous knowledge. If the true values (including the systematic error) are x and x', we put the systematic difference equal to y. The prior probability of x is uniformly distributed over a long range l; that of x', given x, is not so distributed, because the expected values of the difference are presumably much smaller than l. We shall therefore suppose that the prior probability of y is uniformly distributed over a range $-m$ to $+m$ (m much less than l), and that it is independent of that of x. We are supposing the standard errors σ, σ' known already. We denote the proposition $y=0$ by q. Then

$$P(q, dx \mid \sigma, \sigma', h) = \tfrac{1}{2}dx/l; \quad P(\sim q.dxdy \mid \sigma, \sigma', h) = \tfrac{1}{4}dxdy/lm. \tag{1}$$

We denote the measures $x_1 \ldots x_n$, $x'_1 \ldots x'_n$ briefly by θ; then

$$\begin{aligned} P(\theta \mid dx, q, \sigma, \sigma', h) &\\ = n!\,n'!\,(2\pi\sigma^2)^{-\frac{1}{2}n}(2\pi\sigma'^2)^{-\frac{1}{2}n'} &\exp\left\{-\frac{1}{2\sigma^2}\Sigma(x_1-x)^2 - \frac{1}{2\sigma'^2}\Sigma(x'_1-x)^2\right\} \\ = n!\,n'!\,(2\pi\sigma^2)^{-\frac{1}{2}n}(2\pi\sigma'^2)^{-\frac{1}{2}n'} &\exp\left\{-\frac{n\tau^2}{2\sigma^2} - \frac{n'\tau'^2}{2\sigma'^2} - \frac{n}{2\sigma^2}(\bar{x}-x)^2 - \frac{n'}{2\sigma'^2}(\bar{x}'-x)^2\right\}, \end{aligned} \tag{2}$$

while $P(\theta \mid dx, dy, \sim q, \sigma, \sigma', h)$ is got from this by putting $\bar{x}'-x-y$ for $\bar{x}'-x$. Here $\bar{x}$ and $\bar{x}'$ are the means and τ and τ' the standard deviations. Many factors do not involve the unknowns. Then

$$P(q, dx \mid \theta, \sigma, \sigma', h) \propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-x)^2 - \frac{n'}{2\sigma'^2}(\bar{x}'-x)^2\right\}dx, \tag{3}$$

$$P(\sim q, dxdy \mid \theta, \sigma, \sigma', h) \propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-x)^2 - \frac{n'}{2\sigma'^2}(\bar{x}'-x-y)^2\right\}\frac{dxdy}{2m}, \tag{4}$$

and by integration and cancelling another factor,

$$P(q \mid \theta, \sigma, \sigma', h) \propto \pi^{\frac{1}{2}}\exp\left\{-\left(\frac{2\sigma^2}{n}+\frac{2\sigma'^2}{n'}\right)^{-1}(\bar{x}-\bar{x}')^2\right\}, \tag{5}$$

$$P(\sim q \mid \theta, \sigma, \sigma', h) \propto \frac{\pi}{2m}\left(\frac{2\sigma^2}{n}+\frac{2\sigma'^2}{n'}\right)^{\frac{1}{2}}\operatorname{erf}\left[\frac{\bar{x}-\bar{x}'+y}{(2\sigma^2/n+2\sigma'^2/n')^{\frac{1}{2}}}\right]_{y=-m}^{m}, \tag{6}$$

and finally

$$\frac{P(q \mid \theta, \sigma, \sigma', h)}{P(\sim q \mid \theta, \sigma, \sigma', h)} = \frac{4\mu}{\pi^{\frac{1}{2}}}\frac{\exp(-\lambda^2)}{\operatorname{erf}(\lambda+\mu)-\operatorname{erf}(\lambda-\mu)}, \tag{7}$$

where

$$\left(\frac{2\sigma^2}{n}+\frac{2\sigma'^2}{n'}\right)^{\frac{1}{2}}\lambda = \bar{x}-\bar{x}'; \quad \left(\frac{2\sigma^2}{n}+\frac{2\sigma'^2}{n'}\right)^{\frac{1}{2}}\mu = m. \tag{8}$$

The ratio of the posterior probabilities is a function of λ and μ only. If μ is small, we can approximate to the difference of the error functions, and the ratio reduces to unity. The observations therefore tell us nothing with respect to the truth of q that we did not know already. This was to be expected, for μ is the ratio of m to $\sqrt{2}$ times the standard error of y when x and x' are estimated independently. If μ is small it means that we already know that the systematic difference, if any, is much less than can be estimated from the number of observations available, and we are no further forward.

If μ is large, on the other hand, and λ is not very large, the error functions approach 1 and -1; then

$$\frac{P(q \mid \theta, \sigma, \sigma', h)}{P(\sim q \mid \theta, \sigma, \sigma', h)} = \frac{2\mu}{\pi^{\frac{1}{2}}} \exp(-\lambda^2) \tag{9}$$

nearly. Its maximum is $2\mu/\pi^{\frac{1}{2}}$, which is large, and it is unity when

$$\lambda^2 = \log(2\mu/\pi^{\frac{1}{2}}). \tag{10}$$

When λ^2 exceeds this substantially we can infer that q is untrue and that there is a systematic difference.

This result agrees qualitatively with what we should expect; its peculiarity is that whenever significant results are obtained our previous estimate of m appears explicitly in the answer. Given m, the larger the number of observations the larger is μ, and the larger λ has to be to make the estimate of y significant (though at the same time the critical value of $\bar{x} - \bar{x}'$ of course diminishes). It is therefore not correct to say that a systematic difference becomes significant when it reaches any constant multiple of its standard error; the multiple varies with the ratio of m to this standard error. To apply this theory it is therefore necessary that we should have previous knowledge of the range of possible values of y. Such knowledge may exist; thus if we are comparing the times of arrival of an earthquake wave as recorded on two different instruments, we may know from the theory of the instruments what the delays in giving an observable response to the motion of the ground are likely to be. Since m enters only through its logarithm its effect is in any case not great in practical cases, and it does not need to be determined very accurately.

III. Possible failure of the normal law of errors

It may happen, however, that we have no previous information about the range of admissible values of y; then m is effectively infinite, and it appears that no matter how many observations we have we shall never be able to infer a systematic difference. Two possible ways of avoiding this difficulty are easily seen to fail. We might suppose that the prior probability of y in such cases is not uniformly distributed, $P(dy \mid \sim q.h)$ being proportional to $dy/|y|$ instead of dy; but finite limits must still be inserted to keep the total probability of $\sim q$ finite, and these will appear explicitly in the answer. Again, the above analysis supposes σ and σ' known already. In practice they are often unknown; then we must introduce their values into the hypothesis to be tested, and proceed on previous lines*. But then if there are at least two observations in each set we effectively determine σ and σ' in terms of τ and τ'; we do not get rid of m. These attempts therefore lead nowhere.

It seems that the true source of the difficulty is that the normal law is not exact. Its only justification is as an approximation in the case where the actual error is the resultant of a large number of independent components of comparable size, and it is exact only in the limiting case when the component errors are infinitely small and infinite in number. In practice most of the contribution to the standard error comes from components that are finite in size and in number. As an illustration of the effect of allowing for this, let us suppose that the component errors are all $\pm\epsilon$, and that their number is k in both cases, so that

$$\sigma = \sigma' = \epsilon\sqrt{k}.$$

Then a value $\bar{x}-x$ means that in the set of observations $x_1, \ldots, x_n$ there are $n(\bar{x}-x)/\epsilon$ more positive than negative errors, arising from a sampling process that gives nk such errors in all. Thus our problem resembles that discussed in the first section of this paper. It is asking whether, given the sampling ratios from two different large classes, the difference between them gives any reason for inferring a real difference. But when n becomes large the outside factor in (I, 15) is large of order $\sqrt{nk}$, and the ratio of the observed difference to its standard error must be large of order $(\frac{1}{2}\log nk)^{\frac{1}{2}}$ before it becomes significant. It could therefore have been expected from (I, 15) that the conditions for the normal law to be exact are inconsistent with the possibility of finding a definite criterion for significance unless supplementary information is available†. Conversely, the finding of such a criterion depends on the estimation of the number of component errors.

The failure of the normal law for large errors has often been suggested previously, as a result of the occurrence of more large deviations, in comparison with the standard deviation, than the normal law would suggest. This argument has usually been answered by pointing out that the normal law does in fact imply large deviations, and that any set of deviations would in fact be consistent with it. But this answer does not deal with the real question, whether the normal law is in fact the most probable, or indeed possesses any appreciable probability, in comparison with others. We have seen that it is at best an approximation, and we must ask whether the observed deviations are due just to the imperfection of this approximation.

It appears that it is at least possible that they may be. We may think of the normal curve as divided into three regions, according as the error is small, moderate, or large. Let us compare it with an error curve where the component

* Jeffreys, *Scientific Inference*, 66–70.

† The analogy is illustrative and not complete. $y=0$ here would correspond in Section I to $(p-\frac{1}{2})k\epsilon = (p'-\frac{1}{2})k'\epsilon'$, not to $p=p'$.

errors are all finite. The area of this curve and its second moment must by construction have the same values as for the normal curve. But it extends only a finite distance from the origin, and greater errors make no contribution to the area or the second moment. Consequently the contributions from the small and moderate errors must be increased to restore the total values. But the large errors contribute appreciably to the ratio of the second moment to the area, and much less in proportion to the area itself; when the large errors are cut off the ratio can be restored only by increasing the ordinates corresponding to errors greater than σ. Hence the moderate errors must have greater ordinates than in the normal curve. This can be illustrated by considering the expectation of error when the individuals are 10 in number and all ± 1, and the number of trials is 1024. For comparison we give the numbers in ranges centred on the same values computed from the normal distribution with the same area and second moment:

	0	2	4	6	8	10	> 11
Discontinuous distribution	252	210	120	45	10	1	0
Normal distribution	254·2	209·4	117·2	44·5	11·5	2·0	0·3
Difference	− 2·2	+ 0·6	+ 2·8	+ 0·5	− 1·5	− 1·0	− 0·3

The standard error is 3·162. In the discontinuous distribution the reduction in the groups centred on 8 and larger errors is compensated by an increase in the groups centred on 4 and 6. The difference is however very slight. For various reasons this is still unsatisfactory. The increased number of errors is confined to those not much over the standard error; in fact the difficulty is that there are too many errors three and four times the standard error. To explain this we must apparently have recourse again to the proof of the normal law. It is proved only for errors that are a moderate multiple of σ and small compared with the maximum error possible; greater errors are shown to have a very small probability in comparison with small ones, but the ratio of the actual probability of error to that given by the normal law is incapable of being estimated by the usual approximation. The complete form is

$$P(\xi)=\left(1+\sum_{r=3}^{\infty}(-1)^r\frac{P_r}{r!}\frac{\partial^r}{\partial\xi^r}\right)\frac{1}{\sqrt{(2\pi)}\,\sigma}\exp\left(-\frac{\xi^2}{2\sigma^2}\right). \qquad (1)$$

The P_r are of order $k\epsilon^r$, where k is the number of component errors and ϵ^r the mean of their rth moments, and the derivative contains a term in $(\xi/\sigma^2)^r$. Hence a typical term is of order $k\epsilon^r\xi^r/\sigma^{2r}$, or, since $\sigma^2=k\epsilon^2$, of order $\xi^r/k^{\frac{1}{2}r-1}\sigma^r$. So long as ξ is of order σ and k is large, therefore, the convergence is reasonably rapid. But when ξ approaches $\sigma\sqrt{k}$ the convergence is bad, and the first term is not a good approximation to the series, as indeed is obvious from inspection of the binomial distribution. But suppose that one component error is usually small compared with the others, but on one occasion out of p is of order $\epsilon\sqrt{p}$; then its standard value is still ϵ. But its contribution to P_r is of order $\epsilon^r p^{\frac{1}{2}r-1}$, and the typical term is of order

$$(k-1+p^{\frac{1}{2}r-1})\frac{\epsilon^r \xi^r}{\sigma^{2r}} \doteqdot \left\{k^{1-\frac{1}{2}r}+\left(\frac{p}{k}\right)^{\frac{1}{2}r}\frac{1}{p}\right\}\frac{\xi^r}{\sigma^r}. \tag{2}$$

For $r=2$, p cancels, and the first term of $P(\xi)$ is unchanged. But the second term in (2) does not decrease with r unless $\xi/\sigma<(k/p)^{\frac{1}{2}}$. All these extra terms may be small on account of the factor $1/p$, but rapid convergence is not to be expected. Accordingly it should be expected that the presence of a component error of this type would make the normal law a still worse approximation for the larger errors.

The difficulty is evidently a physical one. The normal law, the application of which brought it into sight, rests on physical assumptions that are certainly only approximate, and the application of the theory requires some supplementary information. Unfortunately the nature of the component errors themselves makes it difficult to present this information in a very definite form. If they all followed identical laws we could apply the theory of the binomial distribution to find their number, but in fact they clearly follow different laws and the binomial distribution is not a good approximation. For the normal law the ratio of the fourth moment to the square of the second is 3; for the binomial one it is $3-2/k$. In fact the ratio is usually greater than 3, so that the binomial law is a worse approximation than the normal law.

It appears, however, that the departure of the true law from the normal law may resemble that of the binomial law in one respect, that all the component errors have finite ranges and the range of the total error possible is finite. This is in accordance with the language of many observers, who will describe an error of more than a certain amount as "impossible". Most writers on the normal law would apparently interpret this in a Pickwickian sense and expect that with a sufficient number of observations an error of any amount would occur sooner or later. I suggest, however, that what the observer says is literally true; that if an astronomer is observing time to a standard error of 0·1 sec. and says that an error of 1 sec. is impossible, he does not mean that an error of 1 sec. is so rare that it will not be expected to occur until about e^{50} observations have been made, but that it will not occur at all so long as the same conditions of observation are maintained; and I am prepared to believe him.

IV. THE REJECTION OF OBSERVATIONS

The failure of the normal law for large errors is easily seen to be related to the question of the criterion for rejecting observations. If the true law of error is

$$P(d\xi \mid \sigma h)=\frac{1}{\sigma}f\left(\frac{\xi}{\sigma}\right)d\xi,$$

the likelihood of the observations is a maximum when

$$\Sigma_r \frac{f'(x_r-x)}{f(x_r-x)}=0.$$

For the smaller errors each term is practically $(x_r - x)/\sigma^2$, and if we are limited to these the mean is the most probable value. But it seems that in actual cases the large errors fall off less rapidly in number than the normal law would suggest, and if we redetermine f to allow for this the terms arising from them are much less than $(x_r - x)/\sigma^2$. Consequently they should correctly have much less weight in estimating x than the central observations have. The large errors also arise from the component errors that are liable to be specially large, and we have no reason to expect these to be symmetrical about zero. We can in fact say nothing at all about the form of the error law for large errors except from the distribution of these errors in the actual observations. Nearly all the information provided by the large errors is therefore used up in determining the behaviour of $f(\xi)$ in the corresponding regions, and they have little to say that is relevant to the value of x. It is therefore wrong to retain these observations at full weight, and dubious whether they have enough weight to be worth retaining at all.

On the other hand there will be a range where the variation of $f'(\xi/\sigma)/f(\xi/\sigma)$ with x, though less than what it is on the normal law, is comparable with it. In this range the observations have some weight in determining x, but less than those nearer the centre. The weight is therefore a continuous function of the deviation. I have already discussed one problem of this type*; it was found there that the weight was about $\frac{1}{2}$ at deviation $2{\cdot}2\sigma$, falling to $0{\cdot}1$ at deviation $3{\cdot}3\sigma$. This was an unusually intractable case, where the number of large residuals was so large that the determination of a mean and a standard error in the usual way would have been practically impossible. Where the distribution is more nearly normal it may be better to fit one of Pearson's curves; but in any case the weight of the larger errors in determining the distribution of the probability of the true value will be small.

V. Measurement when the standard errors are previously unknown

The reason for the explicit appearance of m in the answer in Section II is now easily seen. The prior probabilities of a real difference and no real difference were taken equal. On the former hypothesis, with the possible real differences ranging from $-m$ to m, the chance of the observed difference being a moderate multiple of σ, σ' is small, and therefore the posterior probability of a real difference contains a small factor of order $\sqrt{\left(\frac{\sigma^2}{n}+\frac{\sigma'^2}{n'}\right)}\Big/m$. But if the possible errors really have finite upper limits we can present the argument in a different way. If we have made two long series of observations and they do not overlap we should hardly consider it worth while to examine the possibility that there is no difference between the quantities measured; the question arises only when the ranges overlap, and then it is acute. But then we can assign a definite value to m; if

* Jeffreys, *Proc. Roy. Soc.* A, 137 (1932), 78–87.

the ranges overlap, the maximum difference possible is the difference between the extreme observations possible. On the normal law this alternative does not arise because there is no extreme limit to the errors possible. Here we can take the difference of the extreme observations as our estimate of m. It will probably be a slight underestimate, but not a serious one.

In Section II we regarded the standard errors as known already; here m can be found only from the observations, and we may as well determine the standard errors also. The alternatives have then to be stated as $y=0$, prior probability $\frac{1}{2}$; $|y| < \frac{1}{2}m$, prior probability $\frac{1}{2}$. Denoting these by q and $\sim q$, we have

$$\left.\begin{aligned} &P(q, dx, d\sigma d\sigma' \mid h) \propto \tfrac{1}{2} dx\, d\sigma\, d\sigma'/l\sigma\sigma'; \\ &P(\sim q, dx, dy, d\sigma, d\sigma' \mid h) \propto \tfrac{1}{2} dx\, dy\, d\sigma\, d\sigma'/lm\sigma\sigma' \end{aligned}\right\}. \tag{1}$$

If the two laws of error are

$$d\phi = \frac{1}{\sigma} f\left(\frac{\xi}{\sigma}\right), \quad d\phi' = \frac{1}{\sigma'} g\left(\frac{\xi'}{\sigma'}\right), \tag{2}$$

subject to
$$\int_{-\infty}^{\infty} f(u)\, du = 1, \quad \int_{-\infty}^{\infty} g(u)\, du = 1, \tag{3}$$

and the numbers of observations are n, n', we have

$$P(\theta \mid q, dx, d\sigma, d\sigma', h) \propto \sigma^{-n}\sigma'^{-n'} \Pi f\left(\frac{x_r - x}{\sigma}\right) \Pi g\left(\frac{x_s' - x}{\sigma'}\right). \tag{4}$$

On our hypothesis $f(u) = \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}u^2}$ so long as u is not more than 2 or so; beyond that the observations are in any case few and contribute little to the variation of $f(u)$ with x. If p, p' are the total weights of the observations in determining x, $x+y$, σ, σ', $\frac{1}{\sigma} \Pi f\left(\frac{\xi}{\sigma}\right)$ is practically proportional to

$$(2\pi)^{-\frac{1}{2}p} \sigma^{-p} \exp - \frac{p}{2\sigma^2}((x - \bar{x})^2 + \tau^2), \tag{5}$$

where $\bar{x}$ and τ are the mean and standard deviation of the observations. Other factors in the product are useful only in determining the form of f for the larger values of u. Then

$$P(\theta \mid q, dx, d\sigma, d\sigma', h) \propto (2\pi)^{-\frac{1}{2}(p+p')} \sigma^{-p} \sigma'^{-p'}$$
$$\exp\left\{-\frac{p}{2\sigma^2}(\bar{x} - x)^2 - \frac{p'}{2\sigma'^2}(\bar{x}' - x)^2 - \frac{p\tau^2}{2\sigma^2} - \frac{p'\tau'^2}{2\sigma'^2}\right\}, \tag{6}$$

while $P(\theta \mid \sim q, dx, dy, d\sigma, d\sigma', h)$ is got from this by putting $\bar{x}' - y$ for $\bar{x}'$. Then

$$P(q, dx, d\sigma, d\sigma' \mid \theta h) \propto \sigma^{-(p+1)} \sigma'^{-(p'+1)}$$
$$\exp\left\{-\frac{p}{2\sigma^2}(\bar{x} - x)^2 - \frac{p'}{2\sigma'^2}(\bar{x}' - x)^2 - \frac{p\tau^2}{2\sigma^2} - \frac{p'\tau'^2}{2\sigma'^2}\right\} dx\, d\sigma\, d\sigma', \tag{7}$$

$$P(\sim q, dx, dy, d\sigma, d\sigma' \mid \theta h) \propto \sigma^{-(p+1)} \sigma'^{-(p'+1)}$$
$$\exp\left\{-\frac{p}{2\sigma^2}(\bar{x} - x)^2 - \frac{p'}{2\sigma'^2}(\bar{x} - x - y)^2 - \frac{p\tau^2}{2\sigma^2} - \frac{p'\tau'^2}{2\sigma'^2}\right\} dx\, dy\, d\sigma\, d\sigma'/m. \tag{8}$$

To compare the posterior probabilities of q and $\sim q$ we must integrate with regard to x from $-\frac{1}{2}l$ to $+\frac{1}{2}l$, y from about $-\frac{1}{2}m$ to $+\frac{1}{2}m$, σ and σ' from 0 to infinity. In actual cases however l and m will be large enough for the limits for x and y to be replaced by $\pm\infty$. Then making certain approximations based on p and p' being large we find

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \frac{\mu}{\sqrt{\pi}} e^{-\lambda^2}, \tag{9}$$

with the previous definitions of λ and μ; the present m however replaces the previous $2m$, so that the results are formally identical. The difference is in the changed value of m corresponding to the abandonment of the normal law of error for large errors. The following table expresses the results in terms of the parameters $\mu\surd 2$ and $\lambda\surd 2$, these being the easiest to calculate. $\lambda\surd 2$ is the ratio of the observed difference to its standard error $\left(\frac{\tau^2}{p}+\frac{\tau'^2}{p'}\right)^{\frac{1}{2}}$ as usually computed; $\mu\surd 2$ is the ratio of the whole range of the observations of both quantities together to this standard error. The maximum value of the ratio and the critical value of $\lambda\surd 2$ are given.

$\mu\surd 2$	$\mu/\surd\pi$	$\lambda\surd 2$
5	1·99	1·17
10	3·99	1·66
20	7·98	2·04
50	19·9	2·44
100	39·9	2·72
200	79·8	2·96
500	199	3·25
1000	399	3·46

In a fairly ordinary case, suppose that the numbers and standard errors of individual observations of the two sets are equal and that observations with deviations beyond 3τ make no contribution; then $\mu\surd 2 = \surd(18p)$. As p ranges from 5 to 1000, $\lambda\surd 2$ ranges from 1·6 to 2·8. The rule that a difference becomes significant at about two or three times its standard error is therefore about right for ordinary numbers of observations.

VI. Correlation

The usual theory of correlation seems to rest on the postulate that the conditions are such that the probability density for two or more variables is proportional to the exponential of a quadratic function in them*. This is an extension of the normal law of error for one variable, and must rest on similar postulates. We must suppose that the variations of two quantities x and y are composed of a large number of comparable variations, with a tendency for those with positive signs in x to be associated with positive signs in y, to give a positive correlation. Without some such postulate we cannot expect the supposed form of the prob-

* Cf. Brunt, *Combination of observations* (1931), 165.

ability density to hold. It is open to the same objections as the normal law for one variable, but like it may be a useful approximation if due care is used. Such a law involves three parameters, and to estimate the prior probability of these requires some analysis of the foundations of the law. We suppose that the variations of x are all $\pm\alpha$, each being as likely to be positive as negative, and $m+n$ in number; those in y are all $\pm\beta$. For m of the components the variations of x have the same sign and for the other n they have opposite signs. Suppose then that in an actual case the variation of x is $p\alpha$, and of y is $q\beta$. Then in x, $\frac{1}{2}(m+n-p)$ components are positive and the rest negative; for y the corresponding number is $\frac{1}{2}(m+n+q)$. Then $\frac{1}{2}(p-q)$ more components in the n have positive signs in x than in y, and the distribution of signs in x is: from m, $\frac{1}{2}m+\frac{1}{4}(p+q)$ positive, $\frac{1}{2}m-\frac{1}{4}(p+q)$ negative; from n, $\frac{1}{2}n+\frac{1}{4}(p-q)$ positive, $\frac{1}{2}n-\frac{1}{4}(p-q)$ negative. The probability of this arrangement is proportional to

$$\frac{m!}{\{\frac{1}{2}m+\frac{1}{4}(p+q)\}!\{\frac{1}{2}m-\frac{1}{4}(p-q)\}!}\,\frac{n!}{\{\frac{1}{2}m+\frac{1}{4}(p-q)\}!\{\frac{1}{2}m-\frac{1}{4}(p-q)\}!}. \tag{1}$$

Approximating by Stirling's theorem we get

$$P(p\alpha, q\beta \mid \alpha, \beta, m, n) \propto \exp-\left\{\frac{(p+q)^2}{8m}+\frac{(p-q)^2}{8n}\right\}, \tag{2}$$

and to make the total probability of all values of p, q unity the factor must be $1/4\pi\surd(mn)$. The expectations of x^2, y^2, xy are $(m+n)\alpha^2$, $(m+n)\beta^2$, $(m-n)\alpha\beta$. Calling these s^2, t^2, rst, we can write

$$P(dx, dy \mid \alpha, \beta, m, n) = \frac{1}{2\pi\surd(1-r^2)st}\exp-\frac{1}{2(1-r^2)}\left\{\frac{x^2}{s^2}+\frac{y^2}{t^2}-\frac{2rxy}{st}\right\}dxdy. \tag{3}$$

We see that it is expressed wholly in terms of three parameters s, t, r instead of the original four. This of course is a standard result. But subject to the fundamental hypothesis we have also to find the distribution of the prior probability of s, t, r on the supposition that we have no previous information. Now s and t are wholly determined by $m+n$, α and β, and provided that α gives no information about β we can take their prior probability to be proportional to $ds\,dt/st$, the extreme values permissible (supposed very large and very small) being the same for all values of r. Also $r=(m-n)/(m+n)$; thus it depends on the fraction of the component errors that have opposite signs in x and y, and as in sampling we can take its prior probability to be uniformly distributed among the various values of n, given $m+n$. Hence the prior probability of r is uniformly distributed from $r=-1$ to $r=1$. On the hypothesis, then, that x can give relevant information about y, we have finally

$$P(ds\,dt\,dr \mid \sim q.h) \propto \frac{1}{2}\frac{ds}{s}\frac{dt}{t}dr. \tag{4}$$

We are, however, considering as a serious alternative that there is no relation between x and y, and as before we assign equal prior probabilities to the proposition q, that $r=0$, and the proposition $\sim q$, that $r \neq 0$. Then

$$P(dx\,dy \mid q, s, t, h) = \frac{1}{2\pi st} \exp -\frac{1}{2}\left(\frac{x^2}{s^2}+\frac{y^2}{t^2}\right) dx\,dy, \tag{5}$$

$$P(q, ds, dt \mid h) \propto \frac{1}{2}\frac{ds}{s}\frac{dt}{t}, \tag{6}$$

$$P(dx\,dy \mid \sim q, s, t, r, h) = \frac{1}{2\pi st\sqrt{(1-r^2)}} \exp -\frac{1}{2(1-r^2)}\left\{\frac{x^2}{s^2}+\frac{y^2}{t^2}-\frac{2rxy}{st}\right\} dx\,dy, \tag{7}$$

$$P(\sim q, ds, dt, dr \mid h) \propto \frac{1}{4}\frac{ds}{s}\frac{dt}{t}\,dr. \tag{8}$$

Now we have n pairs of observed values of x and y. If we put

$$\Sigma x^2 = n\sigma^2; \quad \Sigma y^2 = n\tau^2; \quad \Sigma xy = n\rho\sigma\tau, \tag{9}$$

and denote the observations collectively by θ, we have

$$P(\theta \mid q, s, t, h) \propto (st)^{-n} \exp -\frac{n}{2}\left(\frac{\sigma^2}{s^2}+\frac{\tau^2}{t^2}\right), \tag{10}$$

$$P(\theta \mid \sim q, s, t, r) \propto (st)^{-n}(1-r^2)^{-\frac{1}{2}n} \exp -\frac{n}{2(1-r^2)}\left\{\frac{\sigma^2}{s^2}+\frac{\tau^2}{t^2}-\frac{2r\rho\sigma\tau}{st}\right\}, \tag{11}$$

and

$$P(q, ds, dt \mid \theta h) \propto (st)^{-(n+1)} \exp -\frac{n}{2}\left(\frac{\sigma^2}{s^2}+\frac{\tau^2}{t^2}\right) ds\,dt, \tag{12}$$

$$P(\sim q, ds, dt, dr \mid \theta h) \propto \tfrac{1}{2}(st)^{-(n+1)}(1-r^2)^{-\frac{1}{2}n} \exp -\frac{n}{2(1-r^2)}\left\{\frac{\sigma^2}{s^2}+\frac{\tau^2}{t^2}-\frac{2r\rho\sigma\tau}{st}\right\} ds\,dt\,dr. \tag{13}$$

Finally to compare the posterior probabilities of q and $\sim q$ we must integrate with respect to all values of s, t, and r.

First consider the case $n=1$. Here if x and y vary at all, ρ, the computed correlation coefficient, is ± 1, being simply $xy/|x||y|$. Then

$$P(q \mid \theta h) \propto \int_0^\infty \int_0^\infty \exp -\frac{1}{2}\left(\frac{\sigma^2}{s^2}+\frac{\tau^2}{t^2}\right)\frac{ds\,dt}{s^2t^2} = \frac{1}{2}\frac{\pi}{\sigma\tau}, \tag{14}$$

$$P(\sim q \mid \theta h) \propto \frac{1}{2}\int_0^\infty \int_0^\infty \int_{-1}^{1} \frac{(1-r^2)^{-\frac{1}{2}}}{s^2t^2} \exp -\frac{1}{2(1-r^2)}(\lambda\xi^2+\mu\eta^2)\,ds\,dt\,dr, \tag{15}$$

where ξ and η are obtained from $1/s$ and $1/t$ by a rotation of axes and λ, μ are the roots of

$$(\lambda-\sigma^2)(\lambda-\tau^2) = r^2\sigma^2\tau^2.$$

On integration then (15) gives

$$\tfrac{1}{4}\pi \int_{-1}^{1} (1-r^2)^{-\frac{1}{2}} \frac{(1-r^2)}{\sqrt{(\lambda\mu)}}\,dr = \frac{1}{2}\frac{\pi}{\sigma\tau}. \tag{16}$$

This is equal to (14), and indicates that if q and $\sim q$ were equally probable before one observation is made, they are equally probable after it; a single observation tells us nothing with respect to this alternative. This was to be expected, since one observation will give unit correlation whether there is any relation between the variables or not. It is necessary, however, to verify it because it checks the whole of the work up to this point.

We proceed to an approximation for n large. It is convenient to transform the variables in (12) by the relations

$$2h^2\sigma^2 = 1, \quad 2k^2\tau^2 = 1, \tag{17}$$

and in (13) by
$$2h^2(1-r^2)s^2 = 1, \quad 2k^2(1-r^2)s^2 = 1. \tag{18}$$

Then
$$P(q \mid \theta h) \propto \int_0^\infty \int_0^\infty (hk)^{n-1} \exp - n(h^2\sigma^2 + k^2\tau^2)\, dh\, dk, \tag{19}$$

$$P(\sim q \mid \theta h) \propto \frac{1}{2}\int_0^\infty \int_0^\infty \int_{-1}^1 (hk)^{n-1}(1-r)^{\frac{1}{2}n} \exp - n(h^2\sigma^2 + k^2\tau^2 - 2hk\sigma\tau r\rho)\, dh\, dk\, dr. \tag{20}$$

When n is large (19) is nearly
$$\frac{\pi}{n}(2e\sigma\tau)^{-n}. \tag{21}$$

Apart from the factor $(hk)^{-1}$ the integrand in (20) has a maximum at

$$r = \rho, \quad 2h^2\sigma^2 = 2k^2\tau^2 = 1/(1-\rho^2); \quad \text{that is, at } s = \sigma,\ t = \tau. \tag{22}$$

Near this it is proportional to

$$\exp - n\left\{(2-\rho^2)\sigma^2 h'^2 - 2\rho^2\sigma\tau h'k' + (2-\rho^2)\tau^2 k'^2 - \frac{2k\rho\sigma\tau}{\sqrt{\{2(1-\rho^2)\}}} h'r' - \frac{2h\rho\sigma\tau}{\sqrt{\{2(1-\rho^2)\}}} k'r' + \frac{1}{2}\frac{1+\rho^2}{(1-\rho^2)^2} r^2\right\}, \tag{23}$$

accents denoting departures from the values at the maximum. The discriminant of the quadratic is $-2n^3\sigma^2\tau^2/(1-\rho^2)$. Hence

$$P(\sim q \mid \theta h) \propto \frac{\pi^{\frac{3}{2}}}{2^{n+\frac{1}{2}}} \frac{e^{-n}}{n^{\frac{3}{2}}(1-\rho^2)^{\frac{1}{2}n-\frac{3}{2}}} (\sigma\tau)^{-n}, \tag{24}$$

and
$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \left(\frac{2n}{\pi}\right)^{\frac{1}{2}} (1-\rho^2)^{\frac{1}{2}(n-3)}. \tag{25}$$

The following table gives the maximum value of this ratio and the critical value of ρ that makes it equal to unity; the value of $(1-\rho^2)/(n-1)^{\frac{1}{2}}$ is given in an extra column for comparison:

n	$(2n/\pi)^{\frac{1}{2}}$	ρ	$(1-\rho^2)/\sqrt{(n-1)}$
5	1·78	0·66	0·281
10	2·52	0·48	0·255
20	3·57	0·37	0·198
50	5·64	0·266	0·133
100	7·97	0·205	0·0963
200	11·3	0·156	0·0691
500	17·8	0·107	0·0442
1000	25·2	0·080	0·0314

The usual formula for the standard error of a correlation coefficient is $(1-\rho^2)/\sqrt{n}$; here n has been replaced by $n-1$ to express the indeterminacy for $n=1$. It appears that a correlation coefficient becomes significant at about twice the standard error computed from the usual formula. This is because when n is large and ρ small $(1-\rho^2)^{\frac{1}{2}(n-3)}$ approaches $e^{-\frac{1}{2}n\rho^2}$ and the variation of the other factor with n is slow in comparison.

It has been supposed in the above analysis that the undisturbed values of x and y are zero or known. If they have to be found from the observations we must denote them by a and b, and replace x, y by $x-a$, $y-b$ in (5) and (7). (6) and (8) become

$$P(q, da, db, ds, dt \mid h) \propto \tfrac{1}{2} da\, db \frac{ds}{s} \frac{dt}{t}, \tag{26}$$

$$P(\sim q, da, db, ds, dt, dr \mid h) \propto \tfrac{1}{4} da\, db \frac{ds}{s} \frac{dt}{t} dr. \tag{27}$$

If now $\Sigma x = n\bar{x}$; $\Sigma y = n\bar{y}$; $\Sigma(x-\bar{x})^2 = (n-1)\sigma^2$; $\Sigma(y-\bar{y})^2 = (n-1)\tau^2$;

$$\Sigma(x-\bar{x})(y-\bar{y}) = (n-1)\rho\sigma\tau, \tag{28}$$

(19) and (20) become

$$P(q \mid \theta h) \propto \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} (hk)^{n-1} \exp - n\{h^2(\bar{x}-a)^2 + k^2(\bar{y}-b)^2 + h^2\sigma^2 + k^2\tau^2\}\, da\, db\, dh\, dk, \tag{29}$$

$$P(\sim q \mid \theta h) \propto \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty}\int_{-1}^{1} (hk)^{n-1}(1-r^2)^{\frac{1}{2}n} \exp\left[-n\{h^2(\bar{x}-a)^2 + k^2(\bar{y}-b)^2 - 2hkr(\bar{x}-a)(\bar{y}-b)\} - (n-1)(h^2\sigma^2 + k^2\tau^2 - 2hk\sigma\tau r\rho)\right] da\, db\, dh\, dk\, dr. \tag{30}$$

The integration with regard to a and b removes a factor hk from (29) and $hk\sqrt{(1-r^2)}$ from (30), and the rest of the work is as before with $n-1$ replacing n. Thus no information is given about the significance of ρ until $n > 2$, again as we should expect. The table is still correct except that all values of n have to be increased by 1, the other columns remaining unchanged.

VII. Periodicity

Suppose that we have $2n+1$ values of a variable at equal intervals of time and that we compute from these a mean value and a set of Fourier coefficients a_r and b_r for sines and cosines. This can be done even if there is no systematic law of variation; we want to test the results for the possibility that the period corresponding to the largest amplitude is real. If the true mean square departure from the mean is s, and there is no systematic law, each Fourier coefficient arises in the way contemplated in the proof of the normal law, and its probability is distributed according to the law

$$P(da_r \mid qsh) = \frac{1}{s}\sqrt{\frac{n}{\pi}} \exp\left(-\frac{na_r^2}{s^2}\right) da_r. \tag{1}$$

If there are real terms corresponding to $r=p$, let their true coefficients be α_p and β_p, and put

$$c_r^2 = a_r^2 + b_r^2; \quad \gamma^2 = \alpha_p^2 + \beta_p^2; \quad c_p'^2 = (a_r - \alpha_r)^2 + (b_r - \beta_r)^2. \tag{2}$$

The mean square of the rest of the variation is only $\sqrt{(s^2-\gamma^2)}$, and

$$P(da_r \mid \sim q, s, p, \alpha, \beta, h) = \frac{1}{\sqrt{(s^2-\gamma^2)}}\sqrt{\frac{n}{\pi}} \exp\left(-\frac{na_r^2}{s^2-\gamma^2}\right) da_r. \tag{3}$$

We do not exclude the possibility that the irregular variation may make contributions to a_p and b_p as calculated, but (3) must be modified for $r=p$ by putting $a_p - \alpha_p$ for a_p. Suppose that in fact we calculate coefficients for m periods, then according to (1) the probability of the observed values is

$$P(\theta \mid q, s, h) = \frac{1}{s^{2m}}\left(\frac{n}{\pi}\right)^m \exp\left(-\frac{n}{s^2}\Sigma c_r^2\right) \Pi da_r db_r, \tag{4}$$

and according to (3) is

$$P(\theta \mid \sim q, s, p, \alpha, \beta, h) = \left(\frac{n}{\pi(s^2-\gamma^2)}\right)^m \exp\left(-\frac{n\,\Sigma' c_r^2}{s^2-\gamma^2}\right) \Pi da_r db_r, \tag{5}$$

where the accent indicates that c_p is replaced by c_p'.

We assess the prior probability of q as $\frac{1}{2}$; this means that m has been chosen so that the m periods considered are as likely as not to include a real one. That of s is distributed according to the law ds/s. If there is no previous opinion as to which value of p is likely to arise we take that of p equal to $1/m$. For α_p and β_p we know that γ must be less than s, and express no opinion about the expected phase. We take

$$P(d\alpha_p d\beta_p \mid \sim q, s, h) = \frac{1}{\pi s^2} d\alpha_p d\beta_p = \frac{1}{\pi s^2}\gamma d\gamma d\theta. \tag{6}$$

Since $d\gamma^2 = -d(s^2-\gamma^2)$ this amounts to saying that all proportions of s^2 that may arise from γ^2 are equally likely. Then

$$P(q, ds \mid h) \propto \tfrac{1}{2} ds \mid s, \tag{7}$$

$$P(\sim q, ds, p, d\alpha_p, d\beta_p \mid h) \propto \frac{1}{2}\frac{ds}{s}\frac{1}{m}\frac{1}{\pi s^2} d\alpha_p d\beta_p, \tag{8}$$

$$P(q\,ds \mid \theta h) \propto \frac{1}{s^{2m+1}} \exp\left(-\frac{n}{s^2}\Sigma c_r^2\right) ds, \tag{9}$$

$$P(\sim q, ds, p, d\alpha_p, d\beta_p \mid \theta h) \propto \frac{1}{m\pi s^3 (s^2-\gamma^2)^m} \exp\left(-\frac{n}{s^2-\gamma^2}\Sigma' c_r^2\right) ds\, d\alpha_p d\beta_p. \tag{10}$$

Here s ranges from 0 to ∞, γ from 0 to s, and θ from 0 to 2π. To compare posterior probabilities of q and $\sim q$ we must integrate with regard to all three. We consider only large values of m and n. Then (9) gives

$$P(q \mid \theta h) \propto \frac{1}{2}\sqrt{\frac{\pi}{m}}\, e^{-m}\left(\frac{m}{n\Sigma c_r^2}\right)^m \tag{11}$$

To integrate (10) it is convenient to change the order of integration, s going from γ to ∞, and γ from 0 to ∞. Integration with regard to s gives

$$P(\sim q, p, d\alpha_p, d\beta_p \mid \theta h) \propto \frac{1}{2\sqrt{(\pi m^3)}} e^{-m} \left(\frac{m}{n\Sigma' c_r^2}\right)^{m-1} \left(\frac{1}{\gamma^2 + (n/m)\Sigma' c_r^2}\right)^2 d\alpha_p d\beta_p. \tag{12}$$

The chief factor is stationary for $\alpha_p = a_p$, $\beta_p = b_p$, as we should expect. So long as these do not exceed the rest of the coefficients too much we can now write

$$\Sigma'' c_r^2 = \Sigma' c_r^2 - c_p'^2 = \Sigma c_r^2 - c_p^2, \tag{13}$$

$$P(\sim q, p \mid \theta h) \propto \frac{1}{2\sqrt{(\pi m^3)}} e^{-m} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{m}{n\Sigma''}\right)^{m+1} \exp\left(-\frac{mc_p'^2}{\Sigma''}\right) d\alpha_p d\beta_p$$

$$\propto \frac{\pi^{\frac{1}{2}}}{2} \frac{e^{-m}}{m^{\frac{1}{2}}} \frac{m^{m-1}}{n^{m+1}(\Sigma'')^m}, \tag{14}$$

and finally
$$\frac{P(\sim q \mid \theta h)}{P(q \mid \theta h)} = \frac{1}{mn}\left(\frac{\Sigma c_r^2}{\Sigma'' c_r^2}\right)^m \tag{15}$$

nearly. This answer is expressed in terms of the available material, but is awkward for computation. But we may write

$$\Sigma c_r^2 = m\sigma^2/n, \tag{16}$$

and then σ is the most probable value of s and nearly the mean square value of the original deviations; it would be exactly the latter if the whole n harmonics had been computed. If c_p^2 is not small compared with $\Sigma'' c_r^2$ the right side of (15) is obviously large; if it is small we can use the approximation

$$\frac{P(\sim q \mid \theta h)}{P(q \mid \theta h)} = \frac{1}{mn} \exp \frac{nc_p^2}{\sigma^2}. \tag{17}$$

Thus the ratio may be small, though it will not often approach $1/mn$.

(*a*) This result resembles Sir Gilbert Walker's criterion* in form, but differs in the appearance of n in the outside factor. Walker considers the probability, given s and q, that the largest of m computed amplitudes will exceed a definite value, so that his discussion is entirely in terms of likelihood. The n comes originally from the factor $1/\pi s^2$ in (6), and this or something very like it seems to be necessary if we are to express the condition that $\gamma < s$. If, however, we have other knowledge that fixes a limit to γ much less than s, the n will be replaced by something still larger, and the test becomes more stringent still.

(*b*) If our problem is to decide on the reality of a particular period, as when an annual period in earthquake frequency is suspected, most of the argument is unchanged, except that p is no longer to be found from the observations, being already assigned. Then q is the proposition that this particular period is not real, $\sim q$ the proposition that it is; if they have equal prior probabilities, the factor $1/m$

* Walker, *Indian Meteor. Mem.* 21 (1914); *Q.J.R. Met. Soc.* 51 (1925), 337–346, and later papers.

in (8) becomes 1, and the m disappears from (17). This confirms the argument generally, since the reality of a preassigned period should not depend greatly on the calculation of the amplitudes corresponding to a number of irrelevant other periods. This form replaces Schuster's criterion. We see from (9) and (10) that if we determine σ from the whole of the material, m is replaced by n, and the dependence on m disappears entirely.

mn or n	$c_p/(\sigma/\sqrt{n})$
10	1·52
30	1·84
100	2·15
300	2·39
1,000	2·63
3,000	2·83
10,000	3·03

In this table the first column is to be taken as mn when the question is about the reality of the largest amplitude that may be obtained, and as n when it is about the reality of a period previously specified by other considerations.

(*c*) When the observations to be tested for periodicity represent a continuous function this argument needs some modification. It might seem that since for a continuous function n, the whole number of calculable amplitudes, is infinite the whole argument breaks down, but this is not so. In any case s is finite, and if the argument leading to (1) still held all the standard values of the expected amplitudes would be equal and zero. But in fact for a continuous function they are not all equal. If a period is so short that two complete periods elapse between any maximum of the function and the next minimum, the contributions to the computed amplitude will nearly cancel out; it is only for longer periods than this that the computed amplitudes have any chance of being considerable. The continuity of the function, in other words, ensures a strong correlation between values at less than this interval, and we cannot treat them as independent in assessing the probabilities of the amplitudes, even if there is no real periodicity. But for the longer periods the assumption of independence is probably a good enough approximation. The number of periods that do not complete themselves twice between consecutive stationary values of the function can be estimated roughly from the appearance of the function itself. If T is the whole range of the time and τ the average interval between consecutive stationary values, the least period with appreciable amplitude will be about $\frac{1}{2}\tau$, and the argument will apply to periods greater than this. The number of these is $2T/\tau$ or $4k$, where k is the whole number of maxima in the range. This is taken for n and the rest of the argument is as before.

(*d*) The reality of a period is equivalent to the proposition that of $2n$ calculated coefficients, the probabilities of most of which are distributed according to the normal law, one or two are so large that they must be considered as due to a departure from the law. This is formally the same as the proposition that of a number of observations, mostly with errors derived from the normal law, some show deviations too large to be attributed to that law. The difference is that in the periodicity problem $2n$ corresponds to k, the number of component errors, and unless k in the error problem can be found in some way we are no further forward.

VIII. General remarks

All the results of this paper are capable of two extensions. We have in each case considered the existence and the non-existence of a real difference between the two quantities estimated as two equivalent alternatives, each with prior probability $\frac{1}{2}$. This is a common case, but not general. If however the prior probabilities are unequal the only difference is that the expression obtained for $P(q \mid \theta h)/P(\sim q \mid \theta h)$ now represents $\dfrac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} \Big/ \dfrac{P(q \mid h)}{P(\sim q \mid h)}$. Thus if the estimated ratio exceeds 1, the proposition q is rendered more probable by the observations, and if it is less than 1, q is less probable than before. It still remains true that there is a critical value of the observed difference, such that smaller values reduce the probability of a real difference. The usual practice is to say that a difference becomes significant at some rather arbitrary multiple of the standard error; the present method enables us to say what that value should be. If however the difference examined is one that previous considerations make unlikely to exist, then we are entitled to ask for a greater increase of the probability before we accept it, and therefore for a larger ratio of the difference to its standard error. To raise the probability of a proposition from 0·01 to 0·1 does not make it the most likely alternative. The increase in such cases, however, depends wholly on the prior probability, and this investigation therefore separates into two parts the ratio of the observed difference to its standard error needed to make the existence of a real difference more likely than not; the first can be definitely evaluated from the observational material, while the second depends wholly on the prior probability.

We have always used q to denote the proposition that the difference under investigation is exactly zero. It might appear that when the difference found is small it will establish that it is zero, in spite of other considerations that may suggest that there should be a real difference, but smaller than what we have called the critical difference. This however is not so. In such a case, where the other considerations suggest a possible variation between $\pm\epsilon$, say, where ϵ is less than the standard error of the difference, we may denote by q the proposition that the difference is within this range. Then the observations have practically the same probability for all values of the difference in this range, and the total posterior probability of all these values is practically what we got for $P(q \mid \theta h)$ before.

The ratios of their probabilities among themselves, however, are almost unaltered. Thus if we expect for some reason a variation up to $0''\cdot1$ in the longitude of a place, but for some other, but dubious, reason there is a possibility of a larger variation, we may make a series of observations and determine a variation of $1''\cdot5$ with a standard error of $1''$, the critical value being $2''$. Then the reality of the larger variation is less probable than before; our confidence in the belief that the real variation is under $0''\cdot1$ is increased correspondingly, but as between different values less than $0''\cdot1$ our choice remains exactly as previous considerations indicated.

Our results are in terms of probabilities. Strictly, when we have evaluated the posterior probability of q, and we want the distribution of the probability of any other parameter x, this is made up of two terms

$$P(x \mid \theta h) = P(x \mid q\theta h)\, P(q \mid \theta h) + P(x \mid \sim q\theta h)\, P(\sim q \mid \theta h).$$

Even if q has a high probability on the data, the second term is not zero unless $P(q \mid \theta h)$ is exactly 1, which will not often happen. This equation is formally correct, expressing the fact that the probability of x is the sum of the parts arising from all the alternative hypotheses together; but nobody is likely to use it. In practice, for sheer convenience, we shall work with a single hypothesis and choose the most probable.

This remark needs a little qualification, since the change from a simple hypothesis to a more complicated one is not what F. P. Ramsey called "ethically neutral". The hypothesis that the unknowns tested by two methods are the same may be preferred even if it has a somewhat smaller probability than the proposition that they are different, simply because it is easier to work with. Further, a journal may be unwilling to publish a new hypothesis if its probability is only slightly more than that of an old one, though the time has not been reached when an improvement of the probability in any specified ratio can be given as the standard for publication. These considerations lie outside the theory of probability, but will affect the application of its results, and the limits of significance indicated may be slightly widened for practical purposes.