FURTHER SIGNIFICANCE TESTS

BY MR HAROLD JEFFREYS, St John's College

[*Received* 13 December 1935, *read* 26 October 1936]

I. GENERAL REMARKS ON SIGNIFICANCE TESTS

In a previous paper* (afterwards referred to as Paper I) tests have been given for the significance of some quantities found statistically. The results are given in the form $P(q \mid \theta h)/P(\sim q \mid \theta h)$; here h denotes the previous knowledge and θ the experimental evidence used, while q is the hypothesis that all the variations outstanding can be attributed to accidental error or random variation, and $\sim q$ the hypothesis that at least part of them is systematic. It has been supposed in the analysis that q and $\sim q$ are equally probable on the information h; but if they are not, the only alteration is that the ratios evaluated now represent

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} \bigg/ \frac{P(q \mid h)}{P(\sim q \mid h)}.$$

If successive batches of relevant information are available the total effect on the probability of q can therefore be got by multiplying the values of

$$P(q \mid \theta h)/P(\sim q \mid \theta h)$$

given by the investigations separately. In each case the assumption that q has prior probability $\frac{1}{2}$ is really a practical working rule rather than a statement of fact. It means strictly that we have no previous information to tell us whether q contains the whole of the facts or whether some modification would give better inferences in the future. It is reasonably clear that q has a moderate prior probability in practical cases, for if it had one near 0 we should not consider its truth worth investigating, and if the ratio found is very different from unity any ordinary error in the prior probability will still leave the posterior probability of q near 0 or 1 as the case may be. If on the other hand the ratio is moderate, the observations have told us little that we did not know already, and the question remains undecided, awaiting more observations or more crucial ones. As a result of a recent discussion of the possible applications of James Bernoulli's theorem† it appears that the prior probability of a general law is not an *a priori* statement in the present state of our knowledge, though it may have been one at a much earlier stage; it is really an inductive inference based on the frequency of success of generalizations suggested in the past, but the available data are selected, because many of the failures have not been published. The practical rule

* Jeffreys, *Proc. Camb. Phil. Soc.* 31 (1935), 203–22.

† *Mind*, 45 (1936), 324–33; *Phil. Mag.* 22 (1936) (in the press).

suggested is that all alternatives seriously advanced should be given the same prior probability.

The results are usually of the form $\alpha n^{\frac{1}{2}} \exp(-\frac{1}{2}x^2/\sigma^2)$, where n is the number of observations and x is the difference found statistically, which may be a difference of two sampling ratios or measurements, a correlation or a harmonic coefficient. σ is the standard error of x as found from the usual statistical theories. α is usually a moderate coefficient. The form of the results can be explained simply in general terms. Suppose that the difference which we are trying to find might have had any value within a range m. It is actually found to be within a certain small range of length τ about x. Then, on the hypothesis that there is a real difference, the probability that the result would be in this range is τ/m. But on the hypothesis that there is no real difference the corresponding probability is $\tau(2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2/\sigma^2)$. Hence by the theorem of inverse probability the probabilities of no real difference and of a real difference are in the ratio $(m/\sigma)(2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2/\sigma^2)$. But if the accuracy of the observations remains constant the standard error of the mean decreases like $n^{-\frac{1}{2}}$; hence the outside factor is of order $n^{\frac{1}{2}}$. If r possible differences are considered at the same time, each contributes its factor m/σ to the result, and the outside factor is proportional to $n^{\frac{1}{2}r}$.

To put the matter in other words, if an observed difference is found to be of order σ, then on the hypothesis that there is no real difference this is what would be expected; but if there was a real difference that might have been anywhere within a range m it is a remarkable coincidence that it should have happened to be in just this particular stretch near zero. On the other hand if the observed difference is several times its standard error it is very unlikely to have occurred if there was no real difference, but it is as likely as ever to have occurred if there was a real difference. In this case beyond a certain value of x the more remarkable coincidence is for the hypothesis of no real difference, and as we have to decide from the facts as presented we shall accept the difference. The theory merely develops these elementary considerations quantitatively and evaluates the factor α. If $P(q \mid \theta h) > \frac{1}{2}$, we shall expect the difference found to persist with more and more accurate observations; if it is less than $\frac{1}{2}$ we shall expect the estimated difference to diminish.

The usual statistical method is to evaluate the observed difference and its standard error, and to say that it is not significant if it is less than a certain constant multiple of this error. No explanation of this rule is given, the probability of the observations being found only on the hypothesis that there is no difference, and not compared with that on the alternative hypothesis. The present method provides an explanation; but the multiple found is not constant, depending on the number of observations and on the ratio of the standard error of one observation to the whole difference possible, but since it involves these numbers only through the square roots of their logarithms the variation in actual cases is not very large.

The method is based on the simplicity postulate first stated in general terms by Dr Wrinch and myself, and further developed in my *Scientific Inference*. This rests on the fact that the widespread acceptance of general laws based on experience is not explained by Laplace's assessment of the prior probability for sampling or by the natural extension of it for measured quantities, but it is explained if the prior probability of a simple law is taken to be comparable with the total prior probability of all modifications of it. The aim of the theory is not to alter common sense, but to state it in formal terms and assist its application to be more systematic. In the cases considered, Laplace's assessment for sampling is rejected for the extreme cases where the class sampled may be pure in respect of the property under examination*, because it leads to results that are not in accordance with general belief. There is no logical justification for the simplicity postulate (or, for that matter, for logic); its justification is that its results explain the inductive inferences actually made, correspond to general belief, and help us to say more definitely in any particular case whether the observations support the hypothesis under consideration or not.

I have already pointed out† that the quantitative law established by experience is not an approximation to the simple law, but the exact simple law itself. When it is found that any natural modification of the simple law leads to a reduction of the probability, we assert the simple law, on the ground that it is more likely to lead to results agreeing with further observations than any of the changes suggested. The question then arises, what is our attitude to possible modifications that may be real but are smaller than the amounts that would be significant on the criterion? When we assert q in this way do we deny any results of later discovery? Thus in the seismological problem that is mentioned later, when it is found that the cubic formula gives a good agreement with the observations and that the formula with a fourth power is less probable, do we mean that in an accurate formula the coefficient of the fourth power would be rigorously zero? In this case the assertion would be fairly certainly untrue, because the cubic formula must break down at some distance from the origin, and an exact one would contain some higher terms. But these would be small within the range considered, and their inclusion with the cubic formula would hardly change the probability of the observations. Thus the observations have nothing to say directly about their inclusion, except that their coefficients are probably considerably smaller than their estimated standard errors. Thus our criterion does not decide on the exactness of the cubic formula. It only asserts that the true formula is unlikely to differ from this except by quantities much less than their standard errors. All hypotheses within this range are then effectively included in q. The essential point is that a formula with three adjustable constants has been found to be the most

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

† *Scientific Inference*, 51.

reliable, and that we should not introduce a fourth without new positive evidence. The cubic formula is for instance an approximation to the result of assuming that the velocity of the wave in question varies uniformly with depth. If we took this as our q it would still give a formula with three adjustable constants, and our data would be unable to decide between this and the exact cubic formula; the decision between them would rest on the theoretical fact that the new formula can be exact, while the cubic one can never be more than an approximation. The success of the cubic formula in fitting the facts therefore provides the observational side of the justification of the other, showing that departures from it must be small. It appears then that our q may not represent a single hypothesis, but what would ordinarily be regarded as a bundle of hypotheses differing in their consequences only by quantities too small to be found from the observations. This leads to an important methodological point: so long as two hypotheses lead to the same observable inferences, they must be regarded, not as different hypotheses, but as subdivisions of the same hypothesis. (If they give exactly the same inferences they are not even subdivisions, but different ways of stating the same hypothesis.) In this sense I am not certain how far the various apparently conflicting views current regarding the "expanding universe" are really distinct or merely different ways of saying the same thing; they will become distinct only when observable differences are predicted between inferences from them, and observational test becomes possible. Similarly when we increase the accuracy of the individual observations or their number or make new crucial tests, we may now find terms of the types previously rejected; but this expresses not a rejection of our original q but a choice between its subdivisions.

In approximating to a series of observed values by means of assigned functions it may turn out that a large number of terms have significant coefficients, and the question will arise as to whether we have chosen the method of approximation in the best way. Thus if we had a four-figure table of $\sin x$ for values of x in circular measure at intervals of 0·01 up to a radian and did not know that it represented $\sin x$, we might analyse it in terms of powers of x and find that the odd ones up to x^7 had significant coefficients. But the ratios of the coefficients would then turn out to be nearly those in the expansion of $\sin x$, and a further analysis in terms of trigonometrical functions would be suggested. Then the second term would be found to have a non-significant coefficient and we should have an approximation in one term; and this would have a higher probability for extrapolation because it contains no adjustable coefficient instead of four.

In the present work successive elements in a contingency table, or successive coefficients in an approximation, are tested in turn; all those that have been found to be supported by the observations in previous tests are treated as definitely established. This is an approximation stated in the last paragraph of Paper I. If θ denotes the observations yet made, and x is some new proposition to be considered, then if the observations give a high probability to q the second term in the equation

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

is small and the probability of x will be nearly the same as if q was definitely proved. In further work q is therefore associated with h and is effectively added to our previous knowledge. This is the fundamental justification of the scientific use of "reality". When a law has received a high probability from experiment we can estimate the probabilities of inferences involving it with considerable accuracy without needing to estimate the contributions from various alternatives not formally excluded but with low probabilities. At this stage, then, any inferences will have almost the same probabilities as if the law was certain. Any parameters in such a law will be regarded as "real" because they affect our inferences. The only philosophical question that arises is whether we take reality as a primitive idea or not. If we do, we shall take this statement as a criterion of what is real; if we do not, we shall take it as a definition of reality. The choice makes no difference to the problem of inference. But on the former attitude we have to be prepared to modify our notion of what is real as more laws are discovered. An external object was the most complete instance of reality before the development of chemistry; a chemist would say that it is a collection of various kinds of atoms; a modern physicist would say that it is composed of protons and electrons or of ψ-waves. Reality may be a primitive idea as a general notion that *something* is real, but it is quite certain that reality of any particular kind cannot be asserted *a priori*, but must emerge from the treatment of observed data. Since the choice of what is real depends on the establishment of laws that will permit further inferences, and the method of establishing them involves the hypothesis of the possibility of inference, the theory of inference must come before the notion of reality in any practical sense. But it seems astonishing that anything so changeable as the scientific use of reality can be a primitive notion. On the other hand if we define reality in the above way it ceases to be a notion capable of truth or falsehood; these lie in the laws that coordinate sensations, and the concepts involved in them, which we choose to call "real", arise of their own accord as new laws are established, and the possibility of change in what we consider real is provided automatically. Reality then becomes a convention based on a mathematical approximation. On a strict phenomenalist philosophy we should have to evaluate all the small remainders of the type rejected above, and the notion of reality would remain somewhat vague; but in practice we do not, not for any *a priori* reason, but because the work would become intolerably cumbersome and the results would hardly differ. If workers in different branches of science call different things "real" there is then no contradiction; they are at liberty to choose their conventions to suit the laws of their respective subjects.

It is not asserted that the notion of, for instance, the external object rests in all cases on the deliberate application of the contingency theory given later in this paper; but it does seem to be true that without elaborate mathematical analysis common-sense is able to recognize contingencies when they have sufficiently high probabilities on the data, and this rough practical way of making inductive inferences and devising concepts to coordinate them is enough in many cases where the data are abundant.

An objection that has been made to my theory is that it is not "objective", because probability is defined as degree of knowledge and not in terms of observations. It might be held to be a sufficient answer that the fact that we make inferences and expect them to hold beyond the original data is as much a fact of observation as any other. Further, any "objective" theory has to take objectivity as a primitive idea; whereas my theory makes it possible to assign a definite meaning to "real" in terms of observation. If we begin by deciding *a priori* what is objective we impose our preconceived ideas of what the world is like upon our theory from the start, and make it impossible ever to remove the subjective element thus introduced; unless indeed we say that nothing but sensations can be objective, but as far as I know nobody says that. It would in fact be possible to define things in terms of sensations in an infinite number of ways, and one of our problems is to explain why we choose any particular way rather than another; and this is explained by the theory of inference. Behind the objection there is, I believe, the idea that it is impossible to make a consistent theory of common-sense reasoning; but that has been sufficiently answered, and it is indeed rather surprising that it survived the publication of Karl Pearson's *Grammar of Science*.

Since the function $\dfrac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} \Big/ \dfrac{P(q \mid h)}{P(\sim q \mid h)}$, which arises in the present work, is a function of the observations alone, it seems worth while to give it a distinctive name. It is greater than unity when the observations support q, and less when they oppose q. If then we define "support" as meaning the logarithm of this function, the support vanishes at the critical value, is negative beyond it, and positive within the critical range. If the support is evaluated for another series of observations, that from the two together can be found by addition.

II. Assessments of the prior probability

1. *Sampling for a class containing several types.* When there are only two alternatives the principle that the most probable value of the ratio of their numbers in the class sampled is the same as in the sample leads to Laplace's assessment of the prior probability, except for an ambiguity when the possibility that the class is all of one type has not yet been excluded. When there are several alternatives we can extend Laplace's assessment. Suppose that the whole number of members is n, divided among r types, the numbers of the respective types being

$m_1, m_2, \ldots, m_r$. Then we say that all compositions are equally probable. The number of ways of dividing n things into r classes is $(n+r-1)!/n!\,(r-1)!$; but m_r is determined when the rest are known, and hence

$$P(m_1, m_2, \ldots, m_{r-1} \mid nh) = n!\,(r-1)!/(n+r-1)!. \tag{1}$$

Of these possibilities, if m_1 is considered fixed, the number of partitions among the others is the number of ways of dividing $n-m_1$ things into $r-1$ classes, which is $(n-m_1+r-2)!/(n-m_1)!(r-2)!$; hence for m_1 by itself

$$P(m_1 \mid nh) = \frac{n!\,(r-1)\,(n-m_1+r-2)!}{(n+r-1)!\,(n-m_1)!}. \tag{2}$$

2. *Chance capable of a continuous series of values.* In the last problem suppose n very large and put $m_1 = np_1$ and so on; the proposition that m_1 has a particular value is the proposition that p_1 is in a particular range dp_1 of length $1/n$. Then

$$P(dp_1 dp_2 \ldots dp_{r-1} \mid nh) = n^{r-1}\,dp_1 \ldots dp_{r-1} \frac{n!\,(r-1)!}{(n+r-1)!}$$
$$\to (r-1)!\,dp_1 \ldots dp_{r-1}. \tag{3}$$

Here n has disappeared and need not be considered further. This gives the distribution of the joint prior probability that the chances of the various types lie in particular ranges. For p_1 separately we can approximate to 1 (2) by Stirling's theorem ($n-m_1$ large) or integrate (3); then

$$P(dp_1 \mid h) = (r-1)\,(1-p_1)^{r-2}\,dp_1. \tag{4}$$

If Σp is restricted to be μ instead of 1, a factor $\mu^{-(r-1)}$ must be included in (3).

In (4) the probability of p_1 is no longer uniformly distributed as on Laplace's assessment. This expresses the restriction that the average value of all the p's is now $1/r$ instead of $\frac{1}{2}$ as for the case of two alternatives; it would now be impossible for more than two of them to exceed $\frac{1}{3}$. But if all but two of them are fixed the prior probability is uniformly distributed as between those two.

Suppose that we have made a sample and that the numbers of various types are $x_1, x_2, \ldots, x_r$. The probability of the sample, given the p's and the actual order of occurrence, is $p_1^{x_1} \ldots p_r^{x_r}$; whence by (3)

$$P(dp_1 \ldots dp_{r-1} \mid \theta h) \propto p_1^{x_1} \ldots p_r^{x_r}\,dp_1 \ldots dp_{r-1}, \tag{5}$$

factors independent of the p's having been dropped. Integrating with respect to all the p's except p_1, the sum of the others being restricted to be less than $1-p_1$, we have

$$P(dp_1 \mid \theta h) \propto \frac{x_1! \ldots x_r!}{(x_1 + \ldots + x_r + r - 2)!} p_1^{x_1} (1-p_1)^{x_2 + \ldots + x_r + r - 2} dp_1$$
$$\propto p_1^{x_1} (1-p_1)^{\Sigma x - x_1 + r - 2}\,dp_1. \tag{6}$$

But if we are given only p_1, the probability of getting x_1 of the first type and $\Sigma x - x_1$ of all the others together is $p_1^{x_1}(1-p_1)^{\Sigma x - x_1}$; and combining this with (4) we get (6) again, the factor $r-1$ being the same for all values of p_1. Hence if we want only the fraction of the class that is of one particular type we need consider only the numbers of that type and the total of the other types in the sample; the distribution among the other types is irrelevant.

W. E. Johnson*, assuming the last result, which is plausible by itself, and working entirely with the posterior probability, has shown by an ingenious method that the probability that the next specimen will be of the first type is linear in x_1; his formula in the present notation is $(1+wx_1)/(r+w\,\Sigma x)$. Applying (6) I find that this probability is $(1+x_1)/(r+\Sigma x)$, whence $w=1$. The undetermined constant in Johnson's formula is thus identified.

3. *Set of quantities the sum of whose squares is known not to exceed s^2, where s is given.* In finding an approximation by Fourier series or other expansion in orthogonal functions up to some definite order we take the least squares solution as giving the most probable values of the coefficients, when the standard error is known, even when the number of observations is small. As for sampling, this is equivalent to taking the prior probability of each coefficient uniformly distributed. The condition that the sum of the squares is not more than s^2 is additional, but the value s^2 is permitted, and any coefficient can reach $\pm s$. Within this range we can still take the distribution uniform. Then

$$\int\int\dots\int da_1\,da_2\dots da_m=\frac{\{\Gamma(\tfrac{1}{2})\}^m}{\Gamma(\tfrac{1}{2}m+1)}s^m=\frac{\pi^{\frac{1}{2}m}s^m}{\Pi(\tfrac{1}{2}m)},\tag{1}$$

whence

$$P(da_1\dots da_m\mid sh)=\frac{\Pi(\tfrac{1}{2}m)}{\pi^{\frac{1}{2}m}s^m}\,da_1\dots da_m.\tag{2}$$

If we did not take the prior probability as uniformly distributed it would express an additional opinion on the relative probabilities of possible values of the coefficients beyond the restriction already stated; but this must not be done without positive evidence.

4. *The distribution of the prior probability of the precision constant.* In the case where we have no previous information about the value of the precision constant, I originally took the prior probability that it should be in a range dh to be proportional to dh/h, on the ground that there seemed to be no reason for preferring either h or the standard error as the criterion of accuracy, and dh/h and $d\sigma/\sigma$ are equal and opposite. Later I noticed that in the same conditions the only factor that can occur without referring to a previous estimate is a power of h, and if this power was not $1/h$ it would express certainty that h was either 0 or infinity. This argument will evidently apply to a large number of other essentially positive quantities. Further, this hypothesis has been shown to be equivalent to the proposition that if we make two observations x_1 and x_2 of a quantity, then whatever their difference may be the probability that a third observation will lie

* *Mind*, 41 (1932), 421–3.

between them is $\frac{1}{3}$. The last result was obtained on the assumption that the normal law of errors holds; but the first two arguments would be equally valid for any law involving only a true value and a precision constant. It is therefore desirable to see whether the third argument holds for such a general law. We suppose that the law of error is expressed by $hf\{h(x-a)\}\,dx$, where

$$\int_{-\infty}^{z} f(z)\,dz = F(z), \quad F(\infty) = 1. \tag{1}$$

The previous knowledge being denoted by k, we have

$$P(da\,dh \mid k) \propto da\,dh/h, \tag{2}$$

$$P(da\,dh \mid x_1, x_2, k) \propto hf\{h(x_1-a)\}f\{h(x_2-a)\}\,da\,dh, \tag{3}$$

$$P(dx_3 \mid x_1, x_2, k) \propto \iint h^2 f\{h(x_1-a)\}f\{h(x_2-a)\}f\{h(x_3-a)\}\,da\,dh\,dx_3. \tag{4}$$

The probability that x_3 lies between x_1 and x_2 is got by integrating with regard to x_3 between these limits and comparing with the result of integrating from $-\infty$ to $+\infty$. Thus it is I_1/I_2, where

$$I_1 = \int_{-\infty}^{\infty}\int_{0}^{\infty} hf\{h(x_1-a)\}f\{h(x_2-a)\}\left[F\{h(x_2-a)\} - F\{h(x_1-a)\}\right] da\,dh, \tag{5}$$

$$I_2 = \int_{-\infty}^{\infty}\int_{0}^{\infty} hf\{h(x_1-a)\}f\{h(x_2-a)\}\,da\,dh. \tag{6}$$

Let us transform the variables to

$$\theta = h(x_1-a), \quad \phi = h(x_2-a). \tag{7}$$

Since $x_2 > x_1$, ϕ is always greater than θ, and the transformation gives

$$\begin{aligned}(x_2-x_1)\,I_1 &= \int_{-\infty}^{\infty}\int_{\theta}^{\infty} f(\theta)f(\phi)\{F(\phi)-F(\theta)\}\,d\theta\,d\phi \\ &= \int_{-\infty}^{\infty} \left[\tfrac{1}{2}f(\theta)\{1-F^2(\theta)\} - f(\theta)\,F(\theta)\{1-F(\theta)\}\right] d\theta \\ &= \tfrac{1}{2} - \tfrac{1}{6} - \tfrac{1}{2} + \tfrac{1}{3} = \tfrac{1}{6}, \end{aligned} \tag{8}$$

$$\begin{aligned}(x_2-x_1)\,I_2 &= \int_{-\infty}^{\infty}\int_{\theta}^{\infty} f(\theta)f(\phi)\,d\theta\,d\phi \\ &= \int_{-\infty}^{\infty} f(\theta)\{1-F(\theta)\}\,d\theta = \tfrac{1}{2}. \end{aligned} \tag{9}$$

Thus $I_1 = \frac{1}{3}I_2$, which proves the proposition. The assumption of the normal law of error was therefore not necessary to the previous result.

Let us now return to (3) and consider the probability that the parameter a may lie between x_1 and x_2. This is I_3/I_4, where

$$(x_2 - x_1) I_3 = \int_{-\infty}^{0} \int_{0}^{\infty} f(\theta) f(\phi) \, d\theta \, d\phi$$

$$= F(0) \{1 - F(0)\}, \tag{10}$$

$$(x_2 - x_1) I_4 = \int_{-\infty}^{\infty} \int_{\theta}^{\infty} f(\theta) f(\phi) \, d\theta \, d\phi = \tfrac{1}{2}. \tag{11}$$

Thus the probability that a will lie between x_1 and x_2 is $\frac{1}{2}$ if, and only if, $F(0) = \frac{1}{2}$. That is, a is such that when it is known an observation is as likely to exceed it as not; it is the median of the distribution. Hence with any law of error, if the accuracy is unknown, the median of the law is as likely as not to lie between the first two observed values. This was proved previously for the normal law. If $F(0)$ is not $\frac{1}{2}$, the probability that a will lie between the first two observations is smaller. As for the proposition for the third observation, if we have previous information about h the result will not hold except for a particular interval between the first two observations.

One use of this discussion is that in comparing different hypotheses to obtain significance tests a factor of proportionality has to be associated with dh/h; in practice h is always limited above by the step of the scale, and below by some consideration such as the total possible range of the observations. It is only when these limits are well separated that the law can be expected to hold, and near them it will not express the correct distribution of the prior probability, because they give additional information about the accuracy when it is very large or very small; but in practice the assumption of complete ignorance of the accuracy is an approximation that should be very accurate over a long range. It is therefore desirable to verify that to assume it in the practical case leads to the same results as in the ideal one, and this can be done by seeing whether it leads to the result that, whatever the separation of the first two observations may have been, the probability that the third will lie between them is $\frac{1}{3}$. The extreme limits are found to give no trouble on account of the convergence of the integrals, but in the practical case they will make a small, though negligible, alteration in the probability in question. Thus, to consider extreme cases, if the first two observations agree exactly we shall suspect that the total variation does not exceed the step of the scale and that the third will repeat them; if they differ by the step of the scale it will be impossible for the third to lie between them, though it will probably agree with one or the other. If the first two differ by nearly the whole range of possible variation, on the other hand, it will be highly probable that the third will be between them. The rule and its verification are good approximations only when the separation of the first two observations is sufficiently different from both the upper and lower limits of possible error; but this is the usual case.

III. CONTINGENCY

1. Problems of several different types may be combined in this term. In the most elementary one a single large class is sampled in respect of several properties at once, and the results are arranged in a table showing the numbers with the various possible conjunctions of pairs of properties. The number with any one of the properties in the original class is then one of our unknowns, and all as far as we know before the investigation are independent. The properties may be quantitative, but as a rule are not. If the chance of an observation falling in the sth column is P_s, and that of one in the tth row is Q_t, then if the row and column are mutually irrelevant the chance of one in the (s, t) element is $P_s Q_t$. If they are related, however, the chance will be p_{st}, which will have to be found from the sample, and we shall want to know whether the numbers in the sample support the hypothesis that p_{st} is different from $P_s Q_t$. It appears that this type of contingency is the most fundamental one in our recognition of the associations of properties that are sufficiently regular to enable us to give names to things*. It is not usually worth while to invent a concept and a name for it on a single property, but only on a conjunction of properties that is too common to be attributed to accidental association.

The next case is that considered in Paper I, with some possible extensions. Classes already separated are sampled in respect of various properties, and we want to know whether the numbers in the samples provide evidence of any difference in the composition of the classes. If the results for one class are given in a column, the total chance in the column is not one of the things we wish to know, and P_s does not arise. The chances for the elements in the column are p_{st}, where $\Sigma_t p_{st} = 1$; the random selection that may influence the total number in the column in the first problem has here been deliberately excluded, and this fact is part of the data. The question is now whether the p_{st} are the same in all columns or not. The results in consequence are not symmetrical with regard to the rows and columns, contrary to what is found in the first case.

The difference can perhaps be seen best from a particular example. Suppose that we wish to compare the statures of Englishmen and Scotsmen. To apply the first method we might use data for the Englishmen and Scotsmen (born in Scotland) that we meet in London; to apply the second we should use samples taken in England and Scotland separately. In the first case the numbers that we meet will depend in part on the actual numbers of Englishmen and Scotsmen in London, and so far as the ratio in the sample differs from that in the total it will introduce some uncertainty. In the second case the sizes of the samples are separately at our disposal; they are not influenced by the ratio of the populations of England and Scotland, and would not be used to estimate it.

* Karl Pearson, *Grammar of Science*, Chapter v.

The third case arises in such problems as dice throwing. The question here is whether the result of one trial affects the probability of the possible results at the next trial. In the case of dice we can say at once that it does not, because the laws of dynamics are already established, and in proper conditions where the die turns over a large number of times after being thrown a small random variation in the conditions of throwing will be enough to spread the probability nearly uniformly over the possible results. However, as N. R. Campbell has suggested, there are cases where the existence of chance is not known initially and has to be tested by experiment. This problem is separate from that of the existence of bias. If the probability of one possible alternative is P_s, and p_{st} is the probability of the tth alternative following the sth, then if previous results do not affect the probability at any one trial $p_{st} = P_s P_t$. We can then test whether the data of succession support this hypothesis.

In all these problems the exponential factor has the same form, and is Pearson's $\exp(-\tfrac{1}{2}\chi^2)$. The other factor, however, is different, because in each case we are asking a different question, or using different data to answer it.

2. *General contingency.* With the definitions stated above, q is the hypothesis that $p_{st} = P_s Q_t$ for all values of s and t, $\sim q$ its denial. Suppose that the actual number of such observations in the (s, t) element is x_{st}; there are m columns and n rows. Put

$$\sum_{s=1}^{m} x_{st} = x_t, \quad \sum_{t=1}^{n} x_{st} = y_s, \quad \sum_{t=1}^{n} x_t = \sum_{s=1}^{m} y_s = N. \tag{1}$$

On hypothesis q, we have $m-1$ independent P_s and $n-1$ independent Q_t. Then

$$P(dP_1 \ldots dP_{m-1} dQ_1 \ldots dQ_{n-1} \mid qh) = (m-1)!\,(n-1)!\,dP_1 \ldots dP_{m-1} dQ_1 \ldots dQ_{n-1}. \tag{2}$$

On hypothesis $\sim q$, the p_{st} have to be assessed separately, so that $mn-1$ of them are independent. Then

$$P(dp_{11} dp_{12} \ldots dp_{st} \ldots dp_{m,n-1} \mid \sim qh) = (mn-1)!\,dp_{11} \ldots dp_{m,n-1}. \tag{3}$$

If the observed x_{st} are collectively denoted by θ, the probability of the observations on hypothesis q is

$$P(\theta \mid P_s Q_t qh) = \prod_{s,t} (P_s Q_t)^{x_{st}} = \Pi_s (P_s^{y_s})\, \Pi_t (Q_t^{x_t}), \tag{4}$$

and on hypothesis $\sim q$ is $\quad P(\theta \mid p_{st}, \sim q, h) = \prod_{s,t} p_{st}^{x_{st}}. \quad (5)$

We treat q and $\sim q$ as equally probable. Then

$$P(dP_1 \ldots dP_{m-1} dQ_1 \ldots dQ_{n-1} q \mid \theta h) \propto (m-1)!\,(n-1)!\, \Pi_s (P_s^{y_s})\, \Pi_t (Q_t^{x_t})\, dP_1 \ldots dP_{m-1}\, dQ_1 \ldots dQ_{n-1}, \tag{6}$$

$$P(dp_{11} \ldots dp_{m,n-1}, \sim q \mid \theta h) \propto (mn-1)! \prod_{s,t} p_{st}^{x_{st}}\, dp_{11} \ldots dp_{m,n-1}. \tag{7}$$

We have now to integrate with regard to all the P_s, Q_t, and p_{st}, subject to the conditions

$$\left.\begin{aligned} 0 &\leqslant P_m = 1 - P_1 - \dots - P_{m-1}; \\ 0 &\leqslant Q_n = 1 - Q_1 - \dots - Q_{n-1}; \\ 0 &\leqslant p_{m,n} = 1 - p_{11} - \dots - p_{m,n-1}. \end{aligned}\right\} \tag{8}$$

The integrals are of Dirichlet's type, and we find

$$P(q \mid \theta h) \propto (m-1)!\,(n-1)!\,\frac{y_1! \dots y_m!}{(N+m-1)!}\,\frac{x_1! \dots x_n!}{(N+n-1)!}, \tag{9}$$

$$P(\sim q \mid \theta h) \propto (mn-1)!\,\frac{x_{11}!\,x_{12}! \dots x_{st}! \dots x_{mn}!}{(N+mn-1)!}. \tag{10}$$

The proportionality factor is the same in both, and our answer is the ratio of these expressions. As a check we may notice that they are equal when m is 1 and n has any value, when the data are all in one column and can give no information about the contingency. To put this into a practical form when the samples are large we proceed to consider the function

$$f = \frac{(x+r-1)!}{x_1! \dots x_r!}\,a_1^{x_1} \dots a_r^{x_r}, \tag{11}$$

where the a's are fractions subject to their sum being 1, and $x_1 + \dots + x_r = x$; then applying Stirling's theorem we have

$$\begin{aligned} f &= \left\{\frac{x+r-1}{(2\pi)^{r-1}\,x_1 \dots x_r}\right\}^{\frac{1}{2}} \frac{(x+r-1)^{x+r-1}}{x_1^{x_1} \dots x_r^{x_r}}\, e^{-(r-1)}\, a_1^{x_1} \dots a_r^{x_r} \\ &= \left\{\frac{x}{(2\pi)^{r-1}\,x_1 \dots x_r}\right\}^{\frac{1}{2}} x^{r-1} \left(\frac{a_1 x}{x_1}\right)^{x_1} \dots \left(\frac{a_r x}{x_r}\right)^{x_r}, \end{aligned} \tag{12}$$

where the terms of the order of $1/x$ of this have been dropped. Now denote the last set of factors by $1/g$, and put

$$x_1 = a_1 x + \alpha_1 x^{\frac{1}{2}} + \eta_1, \text{ etc.}, \tag{13}$$

where α_1 and η_1 are moderate numbers; the a's can be chosen to make this possible. Then

$$\begin{aligned} \log g &= \Sigma x_s \log\left(\frac{x_s}{a_s x}\right) = \Sigma\,(a_s x + \alpha_s x^{\frac{1}{2}} + \eta_s) \log\left(1 + \frac{\alpha_s}{a_s} x^{-\frac{1}{2}} + \frac{\eta_s}{x}\right) \\ &= \Sigma\left(a_s x^{\frac{1}{2}} + \eta_s + \frac{\alpha_s^2}{2a_s}\right) + O(x^{-\frac{1}{2}}) \\ &= \Sigma\,\frac{\alpha_s^2}{2a_s}, \end{aligned} \tag{14}$$

the first two terms cancelling because

$$\Sigma x_s = x = \Sigma a_s x. \tag{15}$$

Then

$$f = \frac{1}{(a_1 \dots a_r)^{\frac{1}{2}}} \left(\frac{x}{2\pi}\right)^{\frac{1}{2}(r-1)} \exp\left(-\Sigma\,\frac{\alpha_s^2}{2a_s}\right). \tag{16}$$

We now apply this to (9) and (10); we can define a_s as y_s/N, and b_t as x_t/N. Then

$$P(q \mid \theta h) \propto (m-1)!\,(n-1)!\,(a_1 \ldots a_m)^{\frac{1}{2}}\,(b_1 \ldots b_n)^{\frac{1}{2}} \left(\frac{2\pi}{N}\right)^{\frac{1}{2}(m+n-2)} a_1^{y_1} \ldots a_m^{y_m} b_1^{x_1} \ldots b_n^{y_n}. \tag{17}$$

So long as x_{st} does not differ too much from $Na_s b_t$ we can write

$$x_{st} = Na_s b_t + N^{\frac{1}{2}}\alpha_{st} + \eta_s, \tag{18}$$

and $$P(\sim q \mid \theta h) \propto (mn-1)!\,\prod_{s,t}(a_s b_t)^{\frac{1}{2}} \left(\frac{2\pi}{N}\right)^{\frac{1}{2}(mn-1)} \prod_{s,t}(a_s b_t)^{x_{st}} \exp\left(-\Sigma \frac{\alpha_{st}}{2a_s b_t}\right). \tag{19}$$

Finally combining these we have

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \frac{m!\,n!}{(mn)!} \frac{1}{(a_1 \ldots a_m)^{\frac{1}{2}(n-1)}(b_1 \ldots b_n)^{\frac{1}{2}(m-1)}} \left(\frac{N}{2\pi}\right)^{\frac{1}{2}(m-1)(n-1)} \exp\left(-\Sigma \frac{\alpha_{st}^2}{2a_s b_t}\right). \tag{20}$$

The expression in the exponent is seen to be Karl Pearson's $\frac{1}{2}\chi^2$. It can be written

$$\tfrac{1}{2}\chi^2 = \Sigma\Sigma \frac{N(x_{st} - Na_s b_t)^2}{2x_s y_t} = \Sigma\Sigma \frac{(x_{st} - Na_s b_t)^2}{2Na_s b_t}. \tag{21}$$

If N is large this ratio will be very large for moderate values of χ^2, indicating that hypothesis q is supported by the observations; but if χ^2 is large the exponential factor will outweigh the outside one and a real variation will be confirmed.

2·1. The above result compares only the two extreme cases. In one there is supposed to be strict proportionality between the chances for the corresponding elements of any two rows or columns; in the other there is no relation at all. In many ordinary cases the conditions are intermediate, most of the chances being proportional but others departing from the rule. In such a case the formula would tell us only that some are proportional or that others are not, whereas we shall usually want to identify the exceptional ones. We must then consider the significance of the observed departures from proportionality in turn; then q becomes the hypothesis that no others are real beyond a certain stage, $\sim q$ the hypothesis that another particular one is real. Some progress has been made with this problem, but the integrations become intractable, and only an approximation will be attempted here. Suppose that we have reached a stage where the doubtful member is a particular p_{st}; as our data we take only the actual number of observations in the element, those in the corresponding row and column, and the total number of observations in the elements not yet found to depart significantly. Suppose that the total chance for all elements in the column, not yet found significant, is P_s, and that for the row Q_t. These have some uncertainty, but we shall ignore it and include the values of P_s and Q_t in h. Then on hypothesis q the chance of one observation falling in the element is $P_s Q_t$. On hypothesis $\sim q$ we know only that p_{st} is less than both P_s and Q_t; let P_s be the smaller. It is possible that p_{st} might be found to absorb the whole of P_s, and neither condition expresses any preference between different values of p_{st}. Hence

$$P(dp_{st} \mid \sim q, h) = dp_{st}/P_s. \tag{1}$$

The probability of an observation in the rest of the column is $P'_s = P_s - p_{st}$, and in the rest of the row $Q'_t = Q_t - p_{st}$. If x_s and y_t are the whole number of observations in the row and column, and N is the whole number of observations,

$$P(\theta \mid \sim q, p_{st}, h) = p_{st}^{x_{st}} (P'_s)^{x_s - x_{st}} (Q'_t)^{y_t - x_{st}} (1 - P'_s - Q'_t - p_{st})^{N - x_s - y_t + x_{st}}. \tag{2}$$

Also

$$P(\theta \mid q, h) = (P_s^{x_s} Q_t^{y_t}) (1 - Q_t)^{x_s - x_{st}} (1 - P_s)^{y_t - x_{st}} \{(1 - P_s)(1 - Q_t)\}^{N - x_s - y_t + x_{st}}. \tag{3}$$

Then

$$P(q \mid \theta h) \propto P_s^{x_s} Q_t^{y_t} (1 - P_s)^{N - x_s} (1 - Q_t)^{N - y_t}, \tag{4}$$

$$P(\sim q, dp \mid \theta h) \propto p_{st}^{x_{st}} (P_s - p_{st})^{x_s - x_{st}} (Q_t - p_{st})^{y_t - x_{st}} (1 - P_s - Q_t + p_{st})^{N - x_s - y_t + x_{st}} dp_{st}/P_s. \tag{5}$$

Put

$$p_{st} = P_s Q_t (1 + \alpha), \quad x_s = NP_s, \quad y_t = NQ_t, \quad x_{st} = NP_s Q_t (1 + x'). \tag{6}$$

Then approximately

$$\frac{P(\sim q, dp_{st} \mid \theta h)}{P(q \mid \theta h)} = \exp\left[-\frac{1}{2}\frac{NP_s Q_t}{(1 - P_s)(1 - Q_t)}\{(\alpha - x')^2 - x'^2\}\right]\frac{dp_{st}}{P_s}, \tag{7}$$

$$\frac{P(\sim q \mid \theta h)}{P(q \mid \theta h)} = Q_t \left\{\frac{2\pi(1 - P_s)(1 - Q_t)}{NP_s Q_t}\right\}^{\frac{1}{2}} \exp\left\{\frac{1}{2}\frac{NP_s Q_t}{(1 - P_s)(1 - Q_t)}\left(\frac{x_{st}}{NP_s Q_t} - 1\right)^2\right\}, \tag{8}$$

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \left\{\frac{NP_s}{2\pi Q_t (1 - P_s)(1 - Q_t)}\right\}^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{(x_{st} - NP_s Q_t)^2}{NP_s Q_t (1 - P_s)(1 - Q_t)}\right\}, \tag{9}$$

which gives the required criterion. For significance the element has to make a contribution to χ^2 of about $\log(N/2\pi)$ or more.

Equation 2 (20) is equivalent to testing $(m-1)(n-1)$ elements in succession, and the expression on the right should be the product of $(m-1)(n-1)$ factors of the form (9). The power of N agrees, but comparison of the other factors is difficult because P_s and Q_t change at every trial.

The hypothesis $P(q \mid h) = \frac{1}{2}$ in this problem says that the particular element discussed is initially as likely to be abnormal as not; the solution therefore applies when previous knowledge directs special attention to this element, or when we have reason to suppose half the elements abnormal. If the problem is whether there is any abnormal element among those not yet discussed, and there is nothing to indicate that it is likely to be any one rather than any other, the prior probability is considerably smaller. The total prior probability that there is an abnormal element is then to be taken as $\frac{1}{2}$, and the probability that any particular element is normal is $1 - k$; if there are r untested elements the probability that they are all normal is $(1 - k)^r = \frac{1}{2}$, whence $k = r^{-1}\log 2$ nearly. If then this describes the conditions, a factor $r/\log 2$ must be included in (9).

3. *Multiple sampling.* This is the extension of the problem of Paper I to the case where m classes are sampled and each sample considered in respect to n alternatives. Let p_{st} be now the chance of a member of the sth class being in the tth subclass. Then the p_{st}, with respect to t, follow the law (2); proposition q is that they do not depend on s. Then on proposition q we can write simply p_t for p_{st}, and

$$P(dp_1 \ldots dp_{n-1} \mid qh) = (n-1)!\, dp_1 \ldots dp_{n-1}, \tag{1}$$

$$P(dp_{11} \ldots dp_{1,n-1} \ldots dp_{m,n-1} \mid \sim q, h) = \{(n-1)\,!\}^m dp_{11} \ldots dp_{m,n-1}, \tag{2}$$

$$P(\theta \mid p_1 \ldots p_{n-1}, qh) = p_1^{x_1} \ldots p_n^{x_n}, \tag{3}$$

$$P(\theta \mid p_{11} \ldots p_{m,n-1}, \sim q, h) = \Pi\,(p_{st}^{x_{st}}), \tag{4}$$

$$P(q, dp_1 \ldots dp_{n-1} \mid \theta h) \propto p_1^{x_1} \ldots p_n^{x_n} dp_1 \ldots dp_{n-1}, \tag{5}$$

$$P(\sim q, dp_{11} \ldots dp_{m,n-1} \mid \theta h)\ \{(n-1)\,!\}^{m-1}\,\Pi\,(p_{st}^{x_{st}})\, dp_{11} \ldots dp_{1,n-1} \ldots dp_{m,n-1}. \tag{6}$$

Integrating (5), we get $$P(q \mid \theta h) \propto \frac{x_1!\, x_2! \ldots x_n!}{(\Sigma x_t + n - 1)!}. \tag{7}$$

In (6) each batch of dp_{st} is distributed according to (1) without reference to the others. Then integrating we have

$$P(\sim q \mid \theta h) \propto \{(n-1)!\}^{m-1} \prod_{s=1}^{m} \frac{x_{s1}!\, x_{s2}! \ldots x_{sn}!}{(y_s + n - 1)!}, \tag{8}$$

and the required ratio is that of (7) and (8). Whem $m = n = 2$ this reduces to the solution of Paper I. To approximate when the ratios in the various classes are not very different, we proceed as before. Put

$$y_s = \Sigma_t x_{st} = N a_s, \quad \Sigma y_s = N, \quad \Sigma_s x_{st} = x_t = N b_t, \tag{9}$$

$$x_{st} = N a_s b_t + N^{\frac{1}{2}} \alpha_{st} + \eta_{st}. \tag{10}$$

Then

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \frac{1}{\{(n-1)!\}^{m-1}} \frac{(a_1 \ldots a_m)^{\frac{1}{2}(n-1)}}{(b_1 \ldots b_n)^{\frac{1}{2}(m-1)}} \left(\frac{N}{2\pi}\right)^{\frac{1}{2}(n-1)(m-1)} \exp\left(-\sum_{s,t} \frac{\alpha_{st}^2}{2a_s b_t}\right). \tag{11}$$

The result is somewhat different from that of III (2), but the last two factors, which are the most important, are the same in both. There seems to be no fixed rule about which is the greater, except that the first is the greater if n is greater than m, the dominant factor in their ratio being $(n/m)^{mn}$.

For $m = 2$, $n = 2$, (11) becomes

$$\left\{\frac{x_1 x_2}{2\pi N b_1 (1 - b_1)}\right\}^{\frac{1}{2}} \exp\left\{-\frac{1}{2} \frac{x_1 x_2 (\beta_1 - \beta_2)^2}{N b_1 (1 - b_1)}\right\}, \tag{12}$$

where β_1 and β_2 are x_{11}/x_1 and x_{12}/x_2, $(x_1 + x_2)\, b_1 = x_{11} + x_{12}$, and $N = x_1 + x_2$.

3·1. As in the problem of general contingency we may need to test one element at a time. This can be reduced to a case of that of Paper I. Suppose that the doubtful element is the tth in the sth class. Classes where the tth element is abnormal, and abnormal elements in the sth class, are irrelevant; consequently we may omit all such classes and rows, and estimate p_t for the whole of the tth row from what are left. Then we can apply (7) or (12) for the case of $n=2$.

4. In the test for randomness the p_{st} are restricted by the relation $\Sigma_\tau p_{s\tau} = \Sigma_\tau p_{\tau s}$ with $P_s = Q_s$. This complicates the analysis, but by the general principle stated in the first section it appears that the condition will be given approximately by II, 2·1 (9), with the larger of the P_s and P_t corresponding to any element replacing Q_t, and the smaller replacing P_s.

In both these cases a correction will be needed analogous to that stated at the end of 2·1 when the element tested is indicated only by the observations and not by any previous considerations specially directed to it.

IV. REPRESENTATION OF A SET OF OBSERVATIONS BY ASSIGNED FUNCTIONS

1. We often have a large number of observations of a quantity y for different values of an argument x, and want to find a function of x that will represent y at other values of x, possibly outside the range covered by the observations. We proceed by choosing a number of linearly independent functions of x and fit a linear combination of them, with adjustable coefficients, to the observed values by least squares. If the number of terms used was equal to n, the number of observations, we could fit the observed values exactly. But if y is affected by errors of observation the resulting representation would be incorrect by the whole error of observation at the values observed, by comparable amounts at intermediate points, and by considerably more if we tried to use it for extrapolation. The constituent functions that vary most rapidly with x then actually diminish the accuracy of the representation. It is therefore better to try to represent y by means of a number of functions smaller than n, and to attribute the residuals to accidental error. Our problem is to decide how many terms are worth retaining in any particular case.

The functions chosen will be chosen with reference to their appropriateness to the particular problem. They arise usually in a definite order of complexity. Thus for a polynomial representation we should take the first constant, the second linear in x, the third quadratic, and so on; for a representation by harmonic functions we should take harmonics of increasing integral multiples of x. In many cases the need for the first two or three will be obvious from inspection of the observed values and their differences; but beyond this we need some criterion.

The observed values may be at unequal intervals of the argument, or some values may be repeated. Without loss of generality we can make the functions mutually orthogonal, and we can normalize them; thus if we are seeking a representation in the form

$$y = Sa_r f_r(x), \tag{1}$$

where S denotes summation over all the functions f_r, we choose the functions in such a way that

$$\Sigma f_l^2(x) = n, \quad \Sigma f_l(x) f_m(x) = 0, \tag{2}$$

each summation being over the observed values of the argument. The first function will usually be unity. We test the reality of the terms in a definite order, starting with the ones that vary least rapidly. If the first r are real, the standard error will be found from the standard residual after we have allowed for them. For a known standard error, the most probable values of the coefficients are found by least squares; this is equivalent to saying that the prior probabilities of all coefficients already established are uniformly distributed. But at this stage we shall want to know whether a further term is worth retaining. If its true coefficient is a_{r+1}, it makes a contribution nearly equal to a_{r+1}^2 to the mean square residual after allowing for r terms, so that the previous estimate of the standard error would be too high. However, let us define s to be the true mean square variation of y from the sum of the true values of the sum of the first r terms. This is unknown apart from the observations, and we take

$$P(ds \mid qh) = P(ds \mid \sim qh) \propto ds/s, \tag{3}$$

where q denotes the proposition that the first r terms alone are significant and $\sim q$ the proposition that we need also to retain the $(r+1)$th term. We treat these as equivalent, so that

$$P(q \mid h) = P(\sim q \mid h) = \tfrac{1}{2}. \tag{4}$$

Now a_{r+1} is restricted to lie between $\pm s$, but within this range its prior probability is uniformly distributed; its best value will still be got by the least squares solution, and this term might account for the whole of the outstanding variation. Thus

$$P(da_{r+1} \mid \sim q, s, h) = da_{r+1}/2s. \tag{5}$$

In some problems it happens that at any stage in the work several terms enter on an equal footing. Thus if we are using trigonometrical functions and know nothing about their phases we must consider a corresponding sine and cosine together; in expressing gravity over the surface of the earth in spherical harmonics there is no ground for expecting any harmonic of the same degree (beyond degree 2) rather than any other. Instead of taking one term at a time we therefore consider the proposition that after the first r terms m new ones arise simultaneously. This does not affect (3) and (4); but now we must take the prior probabilities of the m coefficients all uniformly distributed, subject to the restriction that the sum of their squares is not more than s^2. Within this range we can apply II (3), whence (5) is generalized into

$$P(da_{r+1} \dots da_{r+m} \mid \sim q, s, h) = \frac{\Pi(\frac{1}{2}m)}{\pi^{\frac{1}{2}m} s^m} da_{r+1} \dots da_{r+m}. \tag{6}$$

Thus in all

$$P(q, ds, da_1 \dots da_r \mid h) \propto \frac{ds}{s} da_1 \dots da_r, \tag{7}$$

$$P(\sim q, ds\, da_1 \dots da_{r+m} \mid h) \propto \frac{ds}{s} \frac{\Pi(\frac{1}{2}m)}{\pi^{\frac{1}{2}m} s^m} da_1 \dots da_{r+m}. \tag{8}$$

Since the reality of the first r terms is already established we need express no restrictions on their values, except that they must cover the same ranges in both (8) and (9); for their original ranges of possible values were much greater than those of order $sn^{-\frac{1}{2}}$ that they are now restricted to. The last m are restricted to the same range as in (6).

Now if we denote the values of y collectively by θ, we have

$$P(\theta \mid q, s, a_1 \dots a_r, h) = \frac{1}{(2\pi s^2)^{\frac{1}{2}n}} \exp\left\{-\sum_{l=1}^{r} \frac{(y - Sa_l f_l)^2}{2s^2}\right\} \Pi(dy). \tag{9}$$

But

$$\Sigma(y - Sa_l f_l)^2 = \Sigma\{(y - S\alpha_l f_l) - S(a_l - \alpha_l) f_l\}^2$$

$$= n\sigma^2 + nS(a_l - \alpha_l)^2, \tag{10}$$

where α_l is the coefficient found by the method of least squares, given by

$$n\alpha_l = \Sigma y f_l, \tag{11}$$

and σ^2 is the mean square residual after allowing for r terms.

Similarly

$$P(\theta \mid \sim q, s, a_1 \dots a_{r+m}, h)$$
$$= \frac{1}{\{2\pi(s^2 - S'a^2)\}^{\frac{1}{2}n}} \exp\left\{-\frac{n}{2} \frac{\sigma^2 - S'\alpha^2 + S(a_l - \alpha_l)^2 + S'(a_l - \alpha_l)^2}{s^2 - S'a^2}\right\} \Pi(dy), \tag{12}$$

where S now means summation from $l = 1$ to r, and S' summation from $r+1$ to $r+m$. Hence

$$P(q, ds, da_1 \dots da_r \mid \theta h) \propto \frac{ds}{s} da_1 \dots da_r \frac{1}{(2\pi s^2)^{\frac{1}{2}n}} \exp -\frac{n}{2} \frac{\sigma^2 + S(a_l - \alpha_l)^2}{s^2}, \tag{13}$$

$$P(\sim q, ds, da_1 \dots da_{r+m} \mid \theta h) \propto \frac{ds}{s} \frac{\Pi(\frac{1}{2}m)}{\pi^{\frac{1}{2}m} s^m} da_1 \dots da_{r+m} \frac{1}{\{2\pi(s^2 - S'a^2)\}^{\frac{1}{2}n}}$$
$$\times \exp\left\{-\frac{n}{2} \frac{\sigma^2 - S'\alpha^2 + S(a_l - \alpha_l)^2 + S'(a_l - \alpha_l)^2}{s^2 - S'a^2}\right\}. \tag{14}$$

To get our test we must integrate these with respect to s and all the a's. We need consider only the case when n is large compared with $r+m$, and $S'\alpha^2$ appreciably less than σ^2; if the later terms absorbed nearly all the outstanding residuals their significance would be clear on inspection. Subject to our conditions the last two factors in (13) and (14) vary much more rapidly than the others and have a sharp maximum near $s = \sigma$, $a = \alpha$, and the limits may be replaced by $\pm\infty$ without important loss of accuracy. These two factors in (13) take the form $(2\pi)^{-\frac{1}{2}n} s^{-n} e^{-\frac{1}{2}n-\phi}$, where ϕ is a quadratic in the variations of s and the a's from their optimum values. The discriminant of ϕ is found to be $2^{-r}(n/\sigma^2)^{r+1}$; whence

$$P(q \mid \theta h) \propto \frac{1}{\sigma} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}n}} e^{-\frac{1}{2}n} (2\pi)^{\frac{1}{2}r} \pi^{\frac{1}{2}} \frac{\sigma^{r+1}}{n^{\frac{1}{2}(r+1)}}. \tag{15}$$

The corresponding factors in (15) give a ϕ with the discriminant

$$D=\left\{\frac{n}{(\sigma^2-S'\alpha^2)^2}\right\}^{r+m+1}\begin{vmatrix} \sigma^2 & -\sigma\alpha_1 & -\sigma\alpha_2 & \cdots \\ -\sigma\alpha_1 & \tfrac{1}{2}(\sigma^2-S'\alpha^2+2\alpha_1^2) & \alpha_1\alpha_2 & \cdots \\ -\sigma\alpha_2 & \alpha_1\alpha_2 & \tfrac{1}{2}(\sigma^2-S'\alpha^2+2\alpha_2^2) & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix}$$

$$=\frac{n^{r+m+1}\sigma^2}{2^{r+m}(\sigma^2-S'\alpha^2)^{r+m+2}} \tag{16}$$

and

$$P(\sim q\mid\theta h)\propto\frac{1}{\sigma}\frac{\Pi(\frac{1}{2}m)}{\pi^{\frac{1}{2}m}\sigma^m}\frac{1}{\{2\pi(\sigma^2-S'\alpha^2)\}^{\frac{1}{2}n}}e^{-\frac{1}{2}n}(2\pi)^{\frac{1}{2}(r+m)}\frac{\pi^{\frac{1}{2}}(\sigma^2-S'\alpha^2)^{\frac{1}{2}(r+m+2)}}{n^{\frac{1}{2}(r+m+1)}\sigma}. \tag{17}$$

Finally
$$\frac{P(q\mid\theta h)}{P(\sim q\mid\theta h)}=\frac{(\frac{1}{2}n)^{\frac{1}{2}m}}{\Pi(\frac{1}{2}m)}\left(\frac{\sigma^2-S'\alpha^2}{\sigma^2}\right)^{\frac{1}{2}(n-r-m-2)} \tag{18}$$

To put this into a convenient form we notice that $(\sigma^2-S'\alpha^2)^{\frac{1}{2}}$ is the mean square residual after allowing for the m new terms, and that each of them has a standard error
$$\tau=\{(\sigma^2-S'\alpha^2)/(n-r-m)\}^{\frac{1}{2}}, \tag{19}$$
according to the usual treatment. Then if $S'\alpha^2$ is small compared with σ^2, we have nearly
$$\frac{P(q\mid\theta h)}{P(\sim q\mid\theta h)}=\frac{(\frac{1}{2}n)^{\frac{1}{2}m}}{\Pi(\frac{1}{2}m)}e^{-\frac{1}{2}S'\alpha^2/\tau^2}. \tag{20}$$

The critical value is therefore given by
$$\frac{S'\alpha^2}{\tau^2}=m\log\tfrac{1}{2}n-2\log\Pi(\tfrac{1}{2}m). \tag{21}$$

With regard to the assumption that the same factor of proportionality is to be used in the two equations (3) and the further equations (7) and (8) some comment is needed. The method used has involved the "reality" approximation stated in the first section of this paper. When we say that the first r terms in the foregoing problem are real we mean that their inclusion leads to more reliable inferences than their omission; but actually it may not be quite certain that all these terms are necessary, and a fuller treatment might allow for the small probabilities that some of them should be omitted. In consequence of the practical convenience of the notion of reality, however, we have treated these terms as established with certainty. But when this is done we can take the reality of the terms as part of our fundamental data h, leaving their actual values to be assessed in the course of the work.

Now the law that if an essentially positive quantity s is quite unknown the prior probability that it is in a range ds is proportional to ds/s requires for definiteness a factor of proportionality to make the total probability of all values of s equal to unity. This factor depends on the range of values that are admissible, and is in practice determined by the fact that s is limited below by the step of the measuring instrument and above by any vague knowledge that we may have about the extent that the quantity observed is likely to vary. Near these extremes the law may break down; if s as found by the use of the law is found to be comparable with the step of the scale it may be thought worth while to examine the possibility that the step is in fact the only source of error. Now in the present problem s^2 is the mean square variation of the difference between y and the sum of the true values of the first r terms. It might appear that if the next batch of terms was real it would contribute to s^2, and therefore that the values of s are limited below at a larger value than if these terms were absent. This would alter the factor of proportionality, and (3) would not be true. This, however, is to ask the wrong question. At the stage of the work that we are considering we have a certain set of residuals, and our question is, whether these agree better with the hypothesis that they are entirely due to accidental error or with that that they are at least in part due to real terms in the next batch. If there are real terms, for any value of s the standard accidental error is correspondingly reduced, and the range of possible values of s remains as before.

In (9), again, the first r coefficients appear to be treated differently from the next m, for they are allowed to range from $-\infty$ to $+\infty$, while the latter are confined between $\pm s$. This is because by h they are already established. We take h to assert only their reality and not their values, which might have been anywhere within wide ranges, so that their prior probabilities can be taken as uniformly distributed within any range such that the factor expressing the chance is appreciable. We must however take the permissible ranges to be the same for hypotheses q and $\sim q$, otherwise we shall be asking for the probability of q on one set of data and that of $\sim q$ on another, and to compare such estimates would tell us nothing. But making the ranges the same makes the proportionality factor the same. If the m new terms pass the test for significance we can then consider a further batch, and stop when no increase of probability is obtained. Then as a final step we can re-estimate all the coefficients and the standard error, allowing all now to have large ranges, but since in the conditions of our approximation to the integrals we have effectively taken all the permissible ranges large compared with the standard errors there will be no appreciable change in the distribution of probability from that found already.

The value of r has disappeared from (20), so that the critical ratio of the coefficients to their standard errors does not depend on how many terms have already been included.

1·1. If $m=1$, (20) becomes

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \left(\frac{2n}{\pi}\right)^{\frac{1}{2}} e^{-\alpha^2_{r+1}/2\tau^2}. \tag{22}$$

This is identical with the result for a correlation coefficient between two variables

(VI (25) of previous paper) when the number of observations is large. It is more general, because the former treatment supposed the probability of deviations normally distributed for both variables, which is not assumed here.

As a specimen we may consider the times of transmission of the longitudinal elastic wave in the earth up to a distance of 19°, where they show a discontinuity. The term in the square of the distance should theoretically be absent. The linear and cubic terms are easily seen to be needed. The coefficient of the fourth power was found to be 1·3 times its standard error. The observations were not all of equal weight, but the total weight was that of about 400 good observations. In these conditions the critical value is 2·2 times the standard error; the term was therefore rejected and the solution retaining nothing beyond the cubic term considered established. Between 20° and 30° the square term had a coefficient differing from the value extrapolated from greater distances by 2·2 times the standard error; the effective number of observations was about 230. The critical value is 2·1 times the standard error, so that this difference is somewhat supported by the observations, though not well determined and subject to further test.

1·2. If we are using trigonometrical series, $m=2$, and (20) becomes

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \tfrac{1}{2}n \exp\left\{-\frac{n(\alpha_1^2+\alpha_2^2)}{2\sigma^2}\right\}. \tag{23}$$

Equation VII (17) of the previous paper may be written

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = mn \exp\left\{-\frac{n(a^2+b^2)}{\sigma^2}\right\}, \tag{24}$$

for the case where we are testing the reality of one of m suggested periods. If only one period is suggested m is unity; n of (23) is $2n+1$ of (24). Also in (24) a and b are actual Fourier coefficients; here the terms are supposed normalized, so that $\alpha_1 = a/\sqrt{2}$, $\alpha_2 = b/\sqrt{2}$. The exponents therefore differ by a factor of 2. This arises from an error in equation (1) of the previous discussion; na_r^2/s^2 should be $na_r^2/2s^2$. Correcting this we have agreement. The previous table of critical values should be multiplied by $\sqrt{2}$.

1·3. In the problem of representing gravity by spherical harmonics, let us suppose those up to some degree to have been established. We know that the three first degree harmonics and two of the five second degree ones are theoretically absent. After the constant term and the main ellipticity term are recognized there are therefore 2 second degree ones to consider, expressing the ellipticity of the equator; then 7 third degree ones together; then 9 fourth degree ones, and so on. Thus if we are testing terms of degree 2, (23) holds; for terms of order $k \geqslant 3$, m in (21) must be replaced by $2k+1$.

1·4. Since there is nothing in this work to restrict the number of independent variables, the extension to the significance of partial correlation coefficients is immediate.

2. When the observed values represent a continuous function the conditions depart from those of the last discussion in two ways. So long as the chief source of irregularity in the observed values is error of observation there is some reason to believe that errors at different values of the argument are independent and follow the normal law approximately. But a real continuous variation is often superposed on the normal sources of error, keeping the same sign over finite, though varying, ranges of the argument. A complete Fourier analysis would determine the high harmonics needed to represent this variation within the range of observation, but when the period is obviously varying rapidly a large number of terms will be needed, and there may be little reason to believe in their permanency even if they are evaluated. It will then be better to regard the total variation as composed of a few slowly varying terms with an irregular continuous variation added. This would apply, for instance, to the variation of temperature through the year when the obvious diurnal term has been removed. Thus the chance of the variation at any moment is not distributed according to the normal law, and those at neighbouring moments are not independent. The first difficulty is not serious so long as the number of observations is large, because what matters is the joint chance of all the observations, and when the variation has a finite range this is nearly the same as if the observations followed the normal law with the same standard error. The correlation between neighbouring values, however, implies a reduction in the effective number of observations. This can be dealt with by estimating the average interval through which the residuals keep the same sign. As suggested before, we can simply count the maxima and take the effective number of observations to be four times the number of maxima; alternatively we can work out the correlation between values at different distances apart and find the shortest interval such that this correlation vanishes. This should be nearly $\frac{1}{4}$ of the average interval between consecutive maxima. But then the irregular variation will ordinarily be reversed in sign at somewhat greater intervals, and we shall eliminate it more effectively by taking means over four times this interval. If the short period variation was exactly harmonic this process would eliminate it exactly; if we use only an average period we shall still eliminate most of it. Thus we can adapt our observations to the present theory in the following way. First find the correlation between the residuals at various small distances apart, and hence the shortest interval such that the correlation disappears. Multiply this interval by 4 and divide the observations into ranges (not overlapping) of the latter length, and take a mean over each range. Then the resulting means can be analysed as for discrete observations and the above criterion for significance will apply. The extent of our success in eliminating the irregular variation will depend on how much its quasi-period varies, and the standard error of one mean, while presumably much less than the mean square variation of the original values, will have to be determined directly from the

means and not from the original values. The averaging process will somewhat reduce the coefficients of the genuine harmonic terms found, but this is easily allowed for.

It may be noticed that the real long-period variations will tend to produce positive correlations between neighbouring observations, which will tend to mask the change of sign of the correlation coefficient when the distance is varied; but this can be avoided by first finding approximate coefficients for the early terms and removing them. An accurate determination will not be needed at this stage, and a new one can be made later if it appears necessary.

3. In analysing gravity we have not a continuous record, but in several regions there are enough observations to give a determination of the correlations at different distances and hence of the best size of mesh to use to remove the local irregularities*; this can then be applied to other regions. The true mean over any mesh has however to be replaced by the mean at the finite number of stations within it, and allowance has to be made for the probable value of the difference, which will depend on the number of stations. Thus we have two independent sources of accidental error, apart from the error of observation, which is much smaller. There is this failure of the observations within any mesh to represent the mesh as a whole, which can be estimated from the differences of the observed values within any mesh; but there are many possible terms that together represent an irregular variation that keeps the same sign through any cell. Our problem is to find out up to what degree it is worth while estimating the various harmonics and leaving the rest to be regarded as irregular variation of these two types. A difficulty that arises in this problem and not in our previous one is that the weights are unequal, some meshes being represented by many stations and others by one or two, or none at all. This would give no trouble in our previous conditions; but here if we have to revise any estimate of the scatter of the means in the meshes, the error of representation remaining the same, the ratios of the weights will be altered. We cannot therefore choose our orthogonal functions once for all, as we have supposed done so far. An estimate using the spherical harmonics themselves and not any combinations of them chosen to be orthogonal at the actual weights will therefore be less troublesome. But inclusion of, say, the fourth harmonics will now alter the most probable values found for the coefficients of the third, and the adaptation of our analysis becomes very formidable. The determination of the accidental variation, given the number of terms worth retaining, is not too difficult. If s is the standard variation due to the intermediate inequalities, s' that in a single mesh, and n_p the number of observations per mesh, the total standard error of the mean taken for the mesh is $(s^2+s'^2/n_p)^{\frac{1}{2}}$. This can be substituted in the expression for the chance of the observations, and s and the coefficients found

* J. de Graaff Hunter, *Phil. Trans.* A, 234 (1935), 407–28.

by making this a maximum. The significance test needed to say how many harmonics should be kept, however, involves too difficult integrations.

The following investigation for a simpler case may be of some use toward determining how many terms to retain. The observations are now supposed to have the same known standard error s', and the standard real variation after the first r terms have been removed is s, part of which also may be random error. On hypothesis q the standard error of one observation is $(s^2+s'^2)^{\frac{1}{2}}$. This is the same for all observations, so that the functions stay orthogonal if s is changed. Then (13) and (14) are altered by the substitution of $s^2+s'^2$ for s^2 in the last two factors the others remaining the same. The best value of s^2 is now $\sigma^2-s'^2$. We find as before

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \frac{(\frac{1}{2}n)^{\frac{1}{2}m}}{\Pi(\frac{1}{2}m)} \frac{(\sigma^2-S'\alpha^2)^{\frac{1}{2}(n-r-m-2)}(\sigma^2-s'^2)^{\frac{1}{2}m}}{\sigma^{n-r-2}}..$$

If then s' is much less than σ, so that most of the error is due to the variation that was previously unknown, the criterion is hardly altered. But if s' contributes a notable part of σ, the ratio is much reduced, and the new terms may be retained at a smaller multiple of their apparent standard error than was previously possible. The reason is that in this case much of the residuals is already accounted for by s' which contributes little to the calculated amplitudes. The standard error as found is of course greater than it would be if s' was absent.

To adapt this to the discussion of gravity, we should take n to be the actual number of meshes used and give the meshes weights inversely proportional to $s^2+s'^2/n_p$ with the values found, making the total weight equal to n. Then weighted means of the squares of the residuals and of s'^2/n_p can be substituted for σ^2 and s'^2. $S'\alpha^2$ will then be replaced by the weighted mean of the squares of the contributions to the observed values made by the approximation under test. Such a method can be only approximate, but is likely to be practically applicable without a prohibitive amount of labour.

V. EXAMPLES

1. *Simple sampling.* K. Pearson* quotes from W. F. R. Weldon a series of experiments where 12 dice were repeatedly thrown up simultaneously, a 5 or a 6 being counted as a "success", and the numbers of occasions when 0, 1, ..., 12 successes occurred were recorded. We wish to know whether the results give evidence of bias. The total number of throws is 26,306, so that 315,672 dice were thrown in all, and the number of successes was 106,602; the observed frequency of success is then 0·337699, instead of $\frac{1}{3}$. In this case we have to test the hypothesis of no bias; the latter can be adapted to the present treatment by regarding it as determined by a number of throws of different dice so large that its uncertainty may be neglected, say 10^9, though this is only a mathematical device and the

* *Phil. Mag.* 50 (1900), 167.

theory for such a case could equally well be constructed from the start. Then we apply I (15) of Paper I, or III 3 (11) of this paper; x_2 is so large that $x_1 + x_2$ can be put equal to x_2; $a = \frac{1}{3} = \beta$; $\beta' - \beta = 0{\cdot}004366$; $x_1 = 315672$;

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \left(\frac{x_1}{2\pi \cdot \frac{1}{3} \cdot \frac{2}{3}}\right)^{\frac{1}{2}} \exp\left(-\frac{x_1(\beta-\beta')^2}{2 \cdot \frac{1}{3} \cdot \frac{2}{3}}\right)$$

$$= 476 \exp(-13{\cdot}539) = 6{\cdot}27 \times 10^{-4}.$$

There is therefore strong evidence of bias, as Pearson also infers. For this purpose we need consider only the totals; this is equivalent to assuming that in any throw the probability of a success with any one die is independent of the success or failure of the others. Thus if p is the chance of a success, the chance of a successes in a throw of 12 is $\frac{12!}{a!(12-a)!} p^a (1-p)^{12-a}$. Then the chance in m throws of 12 that 0 will occur n_0 times, 1 n_1 times, and so on, is

$$\frac{m!}{n_0!\, n_1!\, n_2! \ldots n_{12}!} \left(\frac{12!}{0!\,12!}\right)^{n_0} \left(\frac{12!}{1!\,11}\right)^{n_1} \cdots \left(\frac{12!}{12!\,0!}\right)^{n_{12}} p^{\Sigma na} (1-p)^{12m - \Sigma na}.$$

Here only the last two factors involve p and the rest cancel in any comparison of probabilities. But Σna is the whole number of successes and $12m$ the whole number of throws, counting all dice. The distribution of the numbers of successes, apart from their total, could give additional information only if there was a question whether the dice were unequally biased or whether different conditions of throwing encouraged several dice to give similar results in any one throw.

2. *Test of randomness.* There has been discussion recently on the possibility that an earthquake in one region of the earth may stimulate one in another region, and S. Yamaguti* has dealt with the problem by dividing the earthquake regions into eight, and considering how often one in any region is followed by one in each of the eight. F. J. W. Whipple† has objected that Yamaguti's results as stated are consistent with the hypothesis that they are due to ordinary errors of sampling. But the list of earthquakes used is very incomplete, containing only 420 for 32 years. During this time the real number of earthquakes that occurred must have been about 50 times this. Thus even if the phenomenon was real the list might be a random sample, and Whipple's criticism may apply only to the list actually used and might give different results when applied to the material as a whole. A fresh classification has therefore been made, using the International Seismological Summary from July 1926 to December 1930. Two new regions, the Indian Ocean and the Atlantic, were added to Yamaguti's eight; there were eight earthquakes in Africa, but these were ignored because they were too few to be of any use. A few cases where several widely different epicentres would fit the data

* *Bull. Earthquake Research Inst. Tokyo*, 11 (1933), 46–68.

† *Monthly Notices Roy. Astr. Soc.*, Geoph. Suppl. 3 (1934), 233–8.

equally well were also omitted. The North Pacific in west longitude was included with North America; the eastern North Pacific was divided between Japan (with the Loo-Choo Islands and Formosa) and the Philippines; the East Indies were included with the South Pacific, and the West Indies with Central America. The Mediterranean region and the northern coast of Africa are included with Europe. The results are as follows:

First \ Second	Europe	Asia	Indian Ocean	Japan	Philippines	South Pacific	North America	Central America	South America	Atlantic	Total
Europe	97	58	11	73	12	60	22	22	23	19	397
Asia	69	119	13	93	21	56	16	20	22	15	444
Indian Ocean	10	17	8	23	4	10	5	3	6	2	88
Japan	84	90	21	179	22	82	24	36	26	26	590
Philippines	8	18	4	31	33	22	5	6	8	4	139
South Pacific	57	62	14	81	17	115	22	16	22	19	425
North America	17	18	3	32	6	18	21	6	6	5	132
Central America	16	28	4	26	5	22	2	16	10	2	131
South America	29	19	4	33	9	27	7	4	24	1	157
Atlantic	10	15	6	19	10	13	8	2	10	8	101
											2604

In all the four regions with the most earthquakes the successor was found to be oftenest in the same region. Since aftershocks from the same epicentre are a known phenomenon, excess chance in the diagonal elements is a serious possibility. Whether it is enough to show in the present data, however, requires investigation.

We therefore consider first the diagonal elements in the table to see whether they are abnormal. The formula for testing any one of them separately is

$$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \left(\frac{NP_s}{2\pi P_t(1-P_s)(1-P_t)}\right)^{\frac{1}{2}} \exp\left\{-\frac{(x_{st}-NP_sP_t)^2}{2NP_sP_t(1-P_s)(1-P_t)}\right\},$$

where for this purpose $P_t = P_s$. Then for European shocks $P_s = \frac{397}{2604} = 0{\cdot}153$; expected number on the hypothesis of randomness is $NP_s^2 = 61$; observed excess is $+36$. Denoting the denominator of the exponential by $2\sigma^2$ we find $\sigma = 6{\cdot}6$, so that the difference found is $5{\cdot}6\sigma$, and the exponential factor is $1{\cdot}5 \times 10^{-7}$. The other factor is 24, so that the ratio is $3{\cdot}6 \times 10^{-6}$. The difference can therefore be taken as real. For the other diagonal elements we have:

	NP_s^2	Excess	σ	Excess/σ	$P(q\mid\theta h)/P(\sim q\mid\theta h)$
Europe	61	36	6·6	+5·6	$3{\cdot}6\times10^{-6}$
Asia	76	43	7·2	+6·0	$3{\cdot}7\times10^{-7}$
Indian Ocean	3	5	1·7	+2·9	0·3
Japan	134	45	9·0	+5·0	$9{\cdot}0\times10^{-5}$
Philippines	7	26	2·7	+9·7	$2{\cdot}0\times10^{-10}$
South Pacific	70	45	8·4	+5·4	$1{\cdot}1\times10^{-6}$
North America	7	14	2·6	+5·4	$1{\cdot}0\times10^{-6}$
Central America	7	9	2·6	+3·5	$4{\cdot}8\times10^{-2}$
South America	9	15	3·1	+4·8	$2{\cdot}2\times10^{-4}$
Atlantic	4	4	2·0	+2·0	2·8

All except the Atlantic region support the existence of a real difference, the numbers in the last column being less than 1. Even for the Atlantic there is an excess: and at this stage of the work, when the other regions have been found to give excess chances of a repetition of about 0·1, we should be entitled to consider the permissible range of p_{ss} as from P_s^2 to $P_s^2+0{\cdot}1$ instead of from 0 to 1; this would divide the outside factor by $\sqrt{10}$ and leave the result indecisive. The results show that there is a definite excess chance that the successor of a given earthquake will be in the same region.

We have now to test the other elements; the diagonal ones are known to be exceptional and are not considered further. There are 1984 others. P_s for a European earthquake is now the chance that the second earthquake will be in Europe, given that the first was not in Europe; our first estimate of it is therefore $300/1684 = 0{\cdot}177$. But this requires a correction, because the sum of all the $P_s P_t$

	Europe	Asia	Indian Ocean	Japan	Philippines	South Pacific	North America	Central America	South America	Atlantic
Europe	—	60−2	13−2	80−7	17−5*	56+4	18+4	19+3	22+1	15+4
Asia	60+9*	—	14−1	88+5	19+2	62−6	20−4	21−1	24−2	17−2
Indian Ocean	13−3	14+3	—	19+4	4+0	13−3	4+1	5−2	5+1	4−2
Japan	80+4	88+2	19+2	—	25−3	83−1	27−3	29+7*	32−6*	22+4
Philippines	17−9*	19−1	4+0	25+6*	—	18+4	6−1	6+0	7+1	5−1
South Pacific	56+1	62+0	13+1	83−2	18−1	—	19+3	20−4	23−1	16+3
North America	18−1	20−2	4−1	27+5	6+0	19−1	—	6+0	7−1	5+0
Central America	19−3	21+7*	5−1	29−3	6−1	20+2	6−4*	—	8+2	5−3
South America	22+7*	24−5*	5−1	32+1	7+2	23+4	7+0	8−4	—	9−8*
Atlantic	15−5*	17−2	4+2	22−3	5+5*	16−3	5+3	5−3	9+1	—
P_s	0·165	0·181	0·039	0·244	0·052	0·172	0·055	0·059	0·067	0·046

for $s \neq t$ has to be made equal to 1; the ratios have to be kept the same as for our first estimates. If
$$P_s = \lambda x'_s/(N' - x'_s),$$
$$\lambda^2 \Sigma' \frac{x'_s}{N' - x'_s} \frac{x'_t}{N - x'_t} = 1 = \lambda^2 \left\{ \left(\Sigma \frac{x'_s}{N - x'_s} \right)^2 - \Sigma \left(\frac{x'_s}{N - x'_s} \right)^2 \right\},$$
accents meaning that diagonal elements are excluded. Hence λ is found to be 0·931, and we can compute the P_s and the expected numbers. For each succession the. expected number and the difference between the observed and expected numbers are given. The standard error in each case is about equal to 0·8 or 0·9 of the square root of the expected number; the cases where the residual exceeds this root are indicated by asterisks. There are 13 such entries out of 90. We need consider only these. The residual is most marked for the successions Philippines-Europe, Central America-Asia, South America-Europe, South America-Atlantic, and Atlantic-Philippines. Denoting these by a, b, c, d, e we have by applying the formula:

	NP_sP_t	Excess	σ	Excess/σ	$P(q\mid\theta h)/P(\sim q\mid\theta h)$
a	17	-9	3·7	$-2·5$	$11·6 \times 0·044 = 0·51$
b	21	$+7$	4·0	$+1·7$	$11·6 \times 0·236 = 2·7$
c	22	$+7$	4·1	$+1·7$	$12·9 \times 0·236 = 2·9$
d	9	-8	2·8	$-2·9$	$15·6 \times 0·015 = 0·24$
e	5	$+5$	2·1	$+2·4$	$17·6 \times 0·056 = 0·99$

The corresponding ratios for the other elements would be greater and mostly between 10 and 17. The only successions where the data support the existence of any anomaly are for Philippines-Europe and South America-Atlantic, both of which are deficiencies, and neither of which is at all decisive.

At this stage it is necessary to specify more clearly what question we are asking. The ratio that we have called $\dfrac{P(q \mid \theta h)}{P(\sim q \mid \theta h)}$ is really $\dfrac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} \Big/ \dfrac{P(q \mid h)}{P(\sim q \mid h)}$, and therefore gives the true posterior probability only if q was as likely as not initially. In the present case this is roughly true for the diagonal elements, but for the others modification is needed. Our question for them is whether there is any relation between the locality of one earthquake and that of its successor, apart from repetitions. If we take this as a general proposition for all elements, with prior probability $\frac{1}{2}$, 87 successions out of 90 speak definitely against it, and we must regard it as disproved, its posterior probability being 10^{-80} or 10^{-90}. But if we put it on one side we may consider the alternative question, whether any pair of regions has such a connection, granting that the majority of pairs have not. In this case we have nothing to tell us which regions are likely to be concerned; the prior probability that there is at least one such pair is $\frac{1}{2}$, and the prior probability that it will be any one particular pair is $\frac{1}{90} \log_e 2 = \frac{1}{130}$. Thus in testing this suggestion we should take $P(q \mid h)/P(\sim q \mid h) = 130$; then all the posterior probabilities of no connection are large. The distinction is that in the first case we are considering a general proposition to be tested against the whole of the data, while in the second we are deliberately seeking for data that may support the proposition; and in the latter case we must apply a more severe standard of criticism.

It appears therefore that an earthquake in one region has an extra chance of being followed by one in the same region, but that those in different regions occur at random.

That Yamaguti's list is a random selection is confirmed by the fact that in three of his eight regions the excess of successors in the same region is replaced by a deficiency.

3. *Approximation by assigned functions.* Two applications of the criterion have already been mentioned. Another is the problem of the motion of the node of Venus, which was a difficulty in the immediate acceptance of Einstein's law of gravitation, because the assumption of enough extra-planetary matter to explain it would spoil the agreement of the motion of the perihelion of Mercury with the theory; it was left aside only when it was found that the quantity of matter needed to explain it would be too great to have escaped being seen. In finding the motion of the node two unknowns are determined together, effectively from the latitude of Venus, the other being the rate of change of orbital inclination; hence $m = 2$. The observed excess motions with their standard errors are

$$\sin i\delta\Omega = 0{\cdot}60 \pm 0{\cdot}17, \quad \delta i = 0{\cdot}38 \pm 0{\cdot}33$$

in seconds of arc per century. The exponential factor is then

$$\exp\left\{-\frac{1}{2}\left(\frac{60}{17}\right)^2 - \frac{1}{2}\left(\frac{38}{33}\right)^2\right\} = \exp(-6{\cdot}95) = \frac{1}{1030}.$$

There were 12,319 observations. Thus the outside factor is 6160, and the ratio $P(q \mid \theta h)/P(\sim q \mid \theta h)$ is about 6. These two results therefore support the law of gravitation when this is taken as the only serious alternative to accepting them at their face value. On the usual theory the probability of an accidental variation exceeding 3·5 times its standard error is 4×10^{-4}, and the anomaly would have to be taken as real. Such a value will in any case be exceptional, but with the actual number and accuracy of the observations it is more exceptional on the hypothesis that it is real than on the hypothesis that it is due to accidental error.