THE PROBLEM OF k SAMPLES FOR POISSON POPULATION.

By P. V. SUKHATME, *Imperial Institute of Sugar Technology, Cawnpore.*

(Communicated by Prof. P. C. Mahalanobis.)

(*Read January 5, 1937.*)

INTRODUCTION.

A good many contributions have been added in recent years to the 'Problem of k Samples' as formulated by Neyman and Pearson. The discussion is confined to Normal Law Variation and little or no attention is paid to non-normal populations. The first attempt in this direction was probably that of the author (1) who considered the case of Exponential Law Variation and developed a technique of analysis completely analogous to that of Neyman and Pearson. The object of this paper is to consider the problem in relation to yet another two populations—the Poisson and the Binomial Series—the technique for which is fairly known in the works of Fisher. The main interest of the paper, however, is that the technique of this paper forms an alternative approach to the 'Interval' technique of analysis discussed by the author (1).

THE HYPOTHESES CONSIDERED.

It is known that if an event occurs randomly in space or time and the variable considered is the number of occurrences counted in a fixed space or time interval, then the chances that this variable takes values 0, 1, ... x ... are given by the terms of the Poisson series

$$e^{-m}\left(1, m, m^2, \ldots\ldots \frac{m^x}{x!} \ldots\right).$$

Suppose we have a type of data arranged as follows :—

$$\begin{array}{cccccc} x_{11} & x_{21} & \ldots\ldots & x_{t1} & \ldots. & x_{k1} \\ x_{12} & x_{22} & \ldots\ldots & x_{t2} & \ldots. & x_{k2} \\ \vdots & \vdots & & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1n} & \ldots. x_{2n} & \ldots\ldots & x_{tn} & \ldots. & x_{kn} \end{array} \quad .. \quad .. \quad .. \quad (1)$$

where x_{ti} is the number of occurrences counted in a fixed time or space interval and the chance that the variable in the t^{th} column and the i^{th} row takes the value x_{ti} given by

$$e^{-m_{ti}} \frac{(m_{ti})^{x_{ti}}}{x_{ti}\,!} \quad .. \quad .. \quad .. \quad .. \quad (2)$$

Thus in the case of the telephone problem considered in my paper (1), if n represents the number of time units during the day and k the number of

stations, then x_{ti} will be the number of calls arrived during the i^{th} unit at the t^{th} exchange. The problems for consideration would then be whether

(*a*) the intensity of traffic during the day (or part of the day) varies more than might be expected through chance causes ;

(*b*) the traffic is significantly different at different exchanges, assuming that during the day its intensity is the same at every station ;

(*c*) the whole set of record (or part of the set within a rectangle) could be combined together without loss of homogeneity.

Analytically these problems would take the following form :—

(*a*) The hypothesis H_1 that the variation within columns is no more than might be expected through chance causes. That is to say

$$m_{t1} = m_{t2} = \ldots\ldots = m_{tn} \quad (t = 1, \ldots . k) \ldots \quad \ldots \quad \ldots \quad (3)$$

(*b*) The hypothesis H_2. Here it is assumed that

$$m_{t1} = m_{t2} = \ldots\ldots = m_{tn} = m_t \text{ say}$$

the hypothesis to be tested is then that

$$m_1 = m_2 = \ldots\ldots = m_t = \ldots = m_k \quad \ldots \quad \ldots \quad \ldots \quad (4)$$

(*c*) The hypothesis H that the whole set of observations have come from some common Poisson population. That is to say whether (3) and (4) are true.

Our purpose is to determine from the observed data suitable criteria to test the hypotheses H_1, H_2 and H, and to obtain their sampling distributions. The likelihood ratio as defined by Neyman and Pearson (2, 3) has proved to be a very powerful method in determining suitable tests of the statistical hypotheses. In the case of Normal Law Variation the method leads to the usual analysis of variance tests and to certain new tests of practical value. In the case of the χ^2 law variation with two degrees of freedom the method has again proved to be of immense use. We shall use the same in the present instance and shall find that the method leads to the well-known χ^2 tests first given by R. A. Fisher (4).

The derivation of criteria.

The probability function for the joint occurrence of the nk values is obtained by multiplying expressions of the type (2) and may therefore be written

$$p = e^{-S(m_{ti})} \prod_{t=1}^{k} \prod_{i=1}^{n} \frac{(m_{ti})^{x_{ti}}}{x_{ti}\,!} \quad \ldots \quad \ldots \quad \ldots \quad (5)$$

The method of likelihood ratio consists in defining two sets of conditions : (*a*) the conditions which are assumed to be satisfied ; (*b*) the conditions which define the hypothesis to be tested. Thus considering the hypothesis H_1, the conditions (*a*) are that the observations x_{ti} have been drawn from the same Poisson population m_{ti} and the conditions (*b*) are that same

$$m_{t1} = m_{t2} = \ldots . = m_{tn} = m_t \text{ say ;}$$

so that the observations in the t^{th} group are drawn from a common Poisson population $m_t (t = 1, \ldots k)$. The conditions (a) define a class Ω of admissible set of populations m_{ti} and the conditions (b) define a sub-class ω, of Ω to which m_{ti} must belong if the hypothesis tested be true.

The maximum value of p in equation (5) associated with Ω is called $p(\Omega \text{ max.})$ and associated with ω is called $p(\omega \text{ max.})$. The likelihood ratio is then defined as

$$\lambda = \frac{p(\omega \text{ max.})}{p(\Omega \text{ max.})} \quad .. \quad .. \quad .. \quad .. \quad (6)$$

which is supposed to give the suitable criterion of the hypothesis. For it is clear that λ must vary between 0 and 1 and that the smaller the value of λ the less likely is it that the populations belong to ω, that is that the hypothesis tested is true.

Consider the hypothesis H_1.

The conditions (a) and (b) are already defined above. The set Ω consists of populations $m_{ti} (t = 1, \ldots, k; i = 1, \ldots n)$ and the subset ω consists of populations $m_t (t = 1, \ldots k)$ to which the t^{th} group $(t = 1, \ldots k)$ belongs.

We have from the relation (5)

$$\log p = -S(m_{ti}) + S\{x_{ti} \log m_{ti}\} - S\{\log x_{ti}!\}$$

$$\therefore \quad \frac{\partial \log p}{\partial m_{ti}} = -1 + \frac{x_{ti}}{m_{ti}}$$

whence for the maximum value of p, we have

$$m_{ti} = x_{ti}$$

giving us

$$p(\Omega \text{ max.}) = e^{-S(x_{ti})} \prod_{t=1}^{k} \prod_{i=1}^{n} \frac{(x_{ti})^{x_{ti}}}{x_{ti}!} \quad .. \quad .. \quad .. \quad (7)$$

In exactly the same way we have for the subset ω

$$\log p = -S(m_t) + S(x_{ti} \log m_t) - S\{\log x_{ti}!\}$$

giving us

$$\frac{\partial \log p}{\partial m_t} = -n + \frac{\underset{i}{S}(x_{ti})}{m_t}$$

where the summation $\underset{i}{S}$ extends over the values in the t^{th} group.

For the maximum value of p we have

$$m_t = \bar{x}_t \, . = \frac{\underset{i}{S}(x_{ti})}{n}$$

and

$$p(\omega \text{ max.}) = e^{-S(\bar{x}_t)} \prod_{t=1}^{k} \prod_{i=1}^{n} \frac{(\bar{x}_t \, .)^{x_{ti}}}{x_{ti}!} \quad .. \quad .. \quad (8)$$

Hence

$$\lambda_{H_1} = \frac{p(\omega \text{ max.})}{p(\Omega \text{ max.})} = \frac{\prod_{t=1}^{k} (\bar{x}_{t.})^{n\bar{x}_{t.}}}{\prod_{t=1}^{k} \prod_{i=1}^{n} (x_{ti})^{x_{ti}}} \quad .. \quad .. \quad (9)$$

Following the same procedure we shall have

$$\lambda_{H_2} = \prod_{t=1}^{k} \left(\frac{\bar{x}_{..}}{\bar{x}_{t.}}\right)^{n\bar{x}_{t.}} \quad .. \quad .. \quad .. \quad .. \quad (10)$$

where $\bar{x}_{..}$ denotes the mean of the whole set of N observations and

$$\lambda_H = \frac{(\bar{x}_{..})^{N\bar{x}_{..}}}{\prod_{t=1}^{k} \prod_{i=1}^{n} (x_{ti})^{x_{ti}}} \quad .. \quad .. \quad .. \quad .. \quad (11)$$

where $N = nk$.

It will be noticed that

$$\lambda_H = \lambda_{H_1} \times \lambda_{H_2} \quad .. \quad .. \quad .. \quad .. \quad (12)$$

If λ is to be regarded as the fundamental criterion of the hypothesis, we shall have to obtain the probability $P\{\lambda \leqslant \lambda_0\}$ that λ is less than or equal to a given value λ_0 if the hypothesis tested be true. It will be found, however, that the approach to this problem necessarily involves limiting approximations, one of these being the approximate expression for λ itself. These approximate expressions for λ's are found identical with Fisher's indices of dispersion. Thus for λ_H we have

$$-\log \lambda_H = S\{x_{ti}(\log x_{ti} - \log \bar{x})_{..}\} \quad .. \quad .. \quad (13)$$

Substituting

$$x_{ti} = \bar{x}_{..} + z_{ti}\sqrt{\bar{x}_{..}} \quad .. \quad .. \quad .. \quad (14)$$

$$-\log \lambda_H = S(\bar{x}_{..} + z_{ti}\sqrt{\bar{x}_{..}}) \left\{\log\left(1 + \frac{z_{ti}}{\sqrt{\bar{x}_{..}}}\right)\right\}^{*}$$

$$= \frac{1}{2} S(z^2_{ti}) - \frac{1}{2}\frac{1}{\sqrt{\bar{x}_{..}}} S(z^3_{ti})$$

$$+ \frac{1}{3}\frac{1}{\sqrt{\bar{x}_{..}}} S \frac{z^3_{ti}}{\left(1 + \theta\frac{z_{ti}}{\sqrt{\bar{x}_{..}}}\right)^3} + \frac{1}{3}\frac{1}{\bar{x}_{..}} S \left\{\frac{z^4_{ti}}{\left(1 + \theta\frac{z_{ti}}{\sqrt{\bar{x}_{..}}}\right)^3}\right\}_{0<\theta<1}$$

$$= \frac{1}{2} S(z^2_{ti}) + \eta \quad .. \quad .. \quad .. \quad .. \quad .. \quad (15)$$

---

* The expansion of $\log\left(1+\frac{z}{x}\right)$ is valid if z/x is less than one. This may not, however, always happen. But it may be shown that within the boundary of the Best Critical Region its value is less than one and that the expansion is valid.

where η can be made as small as possible by making $\bar{x}_{..}$ sufficiently large, which is a direct consequence of the fact that m is large.

We thus have the result that the distribution of $-2 \log \lambda_H$ approximates to that of $\frac{S(x_{ti}-\bar{x}_{..})^2}{\bar{x}_{..}}$ as m becomes large.

In the same way we shall have

$$-2 \log \lambda_{H_1} = S\left\{\frac{(x_{ti}-\bar{x}_{t.})^2}{\bar{x}_{t.}}\right\} \quad .. \quad .. \quad .. \quad (16)$$

and

$$-2 \log \lambda_{H_2} = S\left\{\frac{(\bar{x}_{t} -\bar{x}_{..})^2}{\bar{x}_{..}}\right\} \quad .. \quad .. \quad .. \quad (17)$$

THE DISTRIBUTIONS.

The expressions on the right hand side of equations (15), (16) and (17) resemble the ordinary χ^2 and are known to be distributed in a Pearsonian χ^2 distribution with $kn-1$, $k(n-1)$ and $k-1$ degrees of freedom. A rigorous proof giving the distribution of χ^2 may be made in more than one way. It is proposed in this paper to follow the method of Kolodziejczyk (5) who gave the proof for the distribution of χ^2 for a single sample of the Binomial series, and show that the χ^2 for samples of the Poisson series follows with good approximation the Pearsonian distribution for large value of m.

It is clear that we need consider the distribution of only one of the three forms of χ^2 given above—say the first. Denote by $P\{\chi^2 \leqslant \chi^2_0\}$ the probability that χ^2 is less than say χ^2_0. It follows :—

$$P\{\chi^2 \leqslant \chi^2_0\} = S\left\{\prod_{t=1}^{k}\prod_{i=1}^{n} e^{-m}\frac{m^{x_{ti}}}{x_{ti}!}\right\} \quad .. \quad .. \quad (18)$$

where the summation S extends over the system of values given by

$$\chi^2 = \frac{S(x_{ti}-\bar{x}_{..})^2}{\bar{x}_{..}} \leqslant \chi^2_0 \quad .. \quad .. \quad .. \quad (19)$$

Since it may be shown that the sum of a given number of terms of the Poisson series can be represented with good approximation by the integral of a normal curve with the same mean and standard deviation when m is large, we may for large values of m write

$$P\{\chi^2 \leqslant \chi^2_0\} = \frac{1}{(\sqrt{2\pi m})^N}\int_{W_0}\cdots\int e^{-\frac{S(x_{ti}-m)^2}{2m}} \prod_{t=1}^{k}\prod_{i=1}^{n} dx_{ti} + \eta \quad .. \quad (20)$$

where x_{ti} are assumed continuous, W_0 denotes the region defined in (19) and η can be made as small as possible by making m sufficiently large. That is to say, given ϵ, any $+ve$ number, however small, we can always find a number m_0 such that for $m \geqslant m_0$, $|\eta| < \epsilon$.

Let us put

$$\frac{1}{(\sqrt{2\pi m})^N}\int_{W_0}\cdots\int e^{-\frac{S(x_{ti}-m)^2}{2m}}\prod_{t=1}^{k}\prod_{i=1}^{n}dx_{ti}=I(W_0) \quad .. \quad .. \quad (21)$$

and choose a positive number Q such that

$$I(W_1)=\left(\frac{1}{\sqrt{2\pi}}\right)^N\int_{W_1}\cdots\int e^{-\frac{S(y^2_{ti})}{2}}\prod_{t=1}^{k}\prod_{i=1}^{n}dy_{ti}>1-\epsilon \quad .. \quad (22)$$

where W_1 is defined by

$$\frac{S(x_{ti}-m)^2}{m}\leqslant Q^2 \quad .. \quad .. \quad .. \quad .. \quad (23)$$

and

$$y_{ti}=\frac{x_{ti}-m}{\sqrt{m}} \quad .. \quad .. \quad .. \quad .. \quad (24)$$

Consider the region W_2—the part common to W_0 and W_1. It follows that

$$P\{\chi^2\leqslant\chi^2{}_0\}=I(W_2)+\eta_2 \quad .. \quad .. \quad .. \quad (25)$$

where

$$|\,\eta_2\,|<2\epsilon \quad \text{for } m\geqslant m_0.$$

It should be noted that for all points in the region W_2 we have

$$|\,x_{ti}-m\,|\leqslant Q\sqrt{m} \quad .. \quad .. \quad .. \quad .. \quad (26)$$

$$\therefore\ |\,\bar{x}_{..}-m\,|\leqslant Q\sqrt{m} \quad .. \quad .. \quad .. \quad .. \quad (27)$$

and hence

$$|\,x_{ti}-\bar{x}_{..}\,|\leqslant 2Q\sqrt{m} \quad .. \quad .. \quad .. \quad (28)$$

Consider the regions W_3 and W_4 defined by

$$\left.\begin{aligned}&\frac{S(x_{ti}-\bar{x}_{..})^2}{\bar{x}_{..}}\leqslant\chi^2{}_0-\epsilon \quad .. \quad .. \quad .. \quad (29)\\ &\frac{S(x_{ti}-m)^2}{m}\leqslant Q^2 \quad .. \quad .. \quad .. \quad (30)\end{aligned}\right\}$$

and

$$\left.\begin{aligned}&\frac{S(x_{ti}-\bar{x}_{..})^2}{\bar{x}_{..}}\leqslant\chi^2{}_0+\epsilon \quad .. \quad .. \quad .. \quad (31)\\ &\frac{S(x_{ti}-m)^2}{m}\leqslant Q^2 \quad .. \quad .. \quad .. \quad (32)\end{aligned}\right\}$$

It follows that

$$W_3<W_2<W_4 \quad .. \quad .. \quad .. \quad .. \quad (33)$$

where the symbol $<$ denotes ' contained in ' and

$$I(W_3)\leqslant I(W_2)\leqslant I(W_4) \quad .. \quad .. \quad .. \quad (34)$$

Now the inequalities (29) and (31) may be written as

$$\frac{S(x_{ti}-\bar{x}_{..})^2}{m} \leqslant (\chi^2{}_0-\epsilon)\left(1+\frac{\bar{x}_{..}-m}{m}\right) \quad .. \quad .. \quad (35)$$

and

$$\frac{S(x_{ti}-\bar{x}_{..})^2}{m} \leqslant (\chi^2{}_0+\epsilon)\left(1+\frac{\bar{x}_{..}-m}{m}\right) \quad .. \quad .. \quad (36)$$

whence it is clear that by increasing m if necessary we shall have

$$\chi^2{}_0 > (\chi^2{}_0-\epsilon)\left(1+\frac{\bar{x}_{..}-m}{m}\right) > \chi^2{}_0-2\epsilon \quad .. \quad .. \quad (37)$$

and

$$\chi^2{}_0 < (\chi^2{}_0+\epsilon)\left(1+\frac{\bar{x}_{..}-m}{m}\right) < \chi^2{}_0+2\epsilon. \quad .. \quad .. \quad (38)$$

Denote by W_5 and W_6 the regions defined by

$$\left.\begin{aligned} &\frac{S(x_{ti}-\bar{x}_{..})^2}{m} \leqslant \chi^2{}_0-2\epsilon \\ &\frac{S(x_{ti}-m)^2}{m} \leqslant Q^2 \end{aligned}\right\} \quad .. \quad .. \quad .. \quad (39), \quad .. \quad .. \quad .. \quad (40)$$

and

$$\left.\begin{aligned} &\frac{S(x_{ti}-\bar{x}_{..})^2}{m} \leqslant \chi^2{}_0+2\epsilon \\ &\frac{S(x_{ti}-m)^2}{m} \leqslant Q^2 \end{aligned}\right\} \quad .. \quad .. \quad .. \quad (41), \quad .. \quad .. \quad .. \quad (42)$$

We then have for m sufficiently large

$$W_5 < W_3 < W_2 < W_4 < W_6 \quad .. \quad .. \quad .. \quad (43)$$

giving us the result

$$I(W_5) \leqslant I(W_2) \leqslant I(W_6) \quad .. \quad .. \quad .. \quad (44)$$

Consider now the regions W_7 and W_8 corresponding to the inequalities (39) and (41) respectively. Then owing to the property of the number Q, we have for $m > m_0$

$$I(W_7) = I(W_5)+\eta_4 \quad .. \quad .. \quad .. \quad .. \quad (45)$$

and

$$I(W_8) = I(W_6)+\eta_5 \quad .. \quad .. \quad .. \quad .. \quad (46)$$

where

$$|\ \eta_4\ | \text{ and } |\ \eta_5\ | \text{ are each less than } \epsilon.$$

But the probability integral of $\frac{S(x_{ti}-\bar{x}_{..})^2}{m}$ is the well-known ordinary χ^2 integral, whence we have

$$\lim_{m\to\infty} P\{\chi^2 \geqslant \chi^2{}_0\} = (2)^{\frac{1}{2}f}\Gamma(\tfrac{1}{2}f)\int_{\chi^2_0}^{\infty}(\chi^2)^{\frac{1}{2}f-1}e^{-\frac{1}{2}\chi^2}d(\chi^2) \quad .. \quad .. \quad (47)$$

where f = number of degrees of freedom
$= N-1$.

In exactly the same way it may be shown that the two forms of χ^2 for H_1 and H_2 hypotheses follow χ^2-distribution with degrees of freedom equal to $k(n-1)$ and $k-1$ respectively.

The binomial series.

Analogous procedure may be followed in the case of Binomial series. The data such as those on page 297 may now arise in a variety of ways. Thus for example, the n values of x in the t^{th} column may represent the number of infested barley ears in n samples of, say, 20 each, drawn from the i^{th} plot and k will represent the number of plots and the problems for consideration would then be whether the material is homogeneous within a plot, whether the proportion of infestation varies considerably from plot to plot and so on.

The complete form of χ^2-technique is given in Table I where in the case of the Binomial series

$p_{t.}$ represents the proportion of infested ears on the t^{th} plot;

$p_{..}$ is the proportion of infested ears on all the plots;

$q_{t.} = 1-p_{t.}$ and $q_{..} = 1-p_{..}$

and the notation used for the Poisson series is one that has been already explained.

Table I.

Variation.	λ-criteria for Poisson Series.	χ^2 Poisson Series.	χ^2 Binomial Series.	Degrees of freedom.
Within Columns ..	$\dfrac{\prod_{t=1}^{k} (\bar{x}_{t.})^{n\bar{x}_{t.}}}{\prod_{t=1}^{k}\prod_{i=1}^{n} (x_{ti})^{x_{ti}}}$	$S\left\{\dfrac{(x_{ti}-\bar{x}_{t.})^2}{\bar{x}_{t.}}\right\}$	$S\left\{\dfrac{(x_{ti}-\bar{x}_{t.})^2}{\bar{x}_{t.}\,q_{t.}}\right\}$	$k(n-1)$
Between Columns ..	$\prod_{t=1}^{k}\left(\dfrac{\bar{x}_{..}}{\bar{x}_{t.}}\right)^{n\bar{x}_{t}}$	$S\left\{\dfrac{(\bar{x}_{t.}-\bar{x}_{..})^2}{\bar{x}_{..}}\right\}$	$S\left\{\dfrac{(\bar{x}_{t.}-\bar{x}_{..})^2}{\bar{x}_{..}q_{..}}\right\}$	$k-1$
Total	$\dfrac{(\bar{x}_{..})^{N\bar{x}_{..}}}{\prod_{t=1}^{k}\prod_{i=1}^{n} (x_{ti})^{x_{ti}}}$	$\dfrac{S(x_{ti}-\bar{x}_{..})^2}{\bar{x}_{..}}$	$\dfrac{S(x_{ti}-\bar{x}_{..})^2}{\bar{x}_{..}\,q_{..}}$	$kn-1$

In conclusion it must be emphasized that the expressions obtained for the probability $P\{\chi^2 \leqslant \chi^2_0\}$ are only approximate and hold true in the limit when m is sufficiently large. The effect of this stipulation in practice and the illustration of the uses of these tests are discussed elsewhere (6). It is found

that when m is as large or larger than 3, the χ^2-forms for the Poisson series follow the ordinary χ^2-distribution with good approximation and that the χ^2-tables may be used with confidence (7).

Summary.

A statistical technique called the technique of 'Count' analysis for samples drawn at random from a Poisson population has been developed. In particular three statistical hypotheses corresponding to Neyman and Pearson's H_1, H_2 and H hypotheses for Normal Law Variation are considered and the principle of likelihood ratio is applied to get suitable criteria based on observations. It is shown that the criteria follow the well-known Pearsonian χ^2-distribution with degrees of freedom appropriate to respective hypotheses. It is emphasized that the tests are approximate and that the approximation becomes satisfactory when m, the population parameter, is as large or larger than 3. A similar set of hypotheses with their criteria are given for samples of the Binomial series.

References.

(1) P. V. Sukhatme, *Statistical Research Memoirs*, Vol. I, 1936.
(2) J. Neyman and E. S. Pearson, *Biometrika*, Vol. XX, 1928.
(3) ,, ,, *Bull. Academie Polonaise*, Science series A, 1931.
(4) R. A. Fisher, Statistical Methods for Research Workers, 1934.
(5) S. Kolodziejczyk, *Annals Soc. Polonaise Math.*, Vol. IX, 1930.
(6) P. V. Sukhatme, *Jour. Agricultural Science*, Vol. VI, Part VI, 1936.
(7) P. V. Sukhatme, *Supp. Jour. Royal Stat. Soc.*, 1937. (In Press.)