MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Report S 192 (VP 6)

Analysis of factorial design for Poisson-variates

by

H.J. Prins

February
1956

1. Summary

In this report a method is proposed for the analysis of factorial design, if the variates follow the Poisson law. BARTLETT [1] gives a treatment of the problem which uses a normalizing transformation. In the present paper this is not done. Some known tests are used to derive an analysis procedure which is often simpler to carry out than the one proposed by BARTLETT.

2. Introduction

As the analysis of factorial design for normal variates (analysis of variance) the terms one way and two way classification will be used.

There is one main difference between analysis of variance and the proposed analysis. The hypotheses tested are here multiplicated and not additive as in the analysis of variance. Moreover a difference is caused by the fact that the normal distribution contains two parameters and the Poisson distribution only one.

The aim of the present investigation is to avoid normalizing transformations by adapting the analysis to the Poisson distribution.

References to previous results of other authors are very incomplete in this report. They may be found however in [3]. Most of the results in this report are not claimed to be new. The generalisation to more classifications however has to the authors knowledge not been published before.

3. The one way classification (i.e. the k -sample test)

If $x_1, \ldots, x_k$ are independent Poisson variates with means $\mu_1, \ldots, \mu_k$ the following tests are considered (H_0 denoting the hypothesis tested, H the alternatives)

$$(1)\begin{cases} H_0: & \dfrac{\mu_i}{\sum \mu_i} = p_i; \qquad p_i \text{ is given}; \; i = 1, \ldots, k \\ H: & \text{at least for one } i: \\ & \mu_i \neq p_i \sum \mu_i . \end{cases}$$

A special case is $p_i = 1/k$;

$$(2)\begin{cases} H_0: & \mu_1 = \ldots = \mu_k \\ H: & \text{at least two of the } \mu_i \text{ differ} \end{cases}$$

To derive a test the following well known property of Poisson-variates is used:

If $\underline{x}_1, \ldots, \underline{x}_k$ are independent Poisson variates with means $\mu_1, \ldots, \mu_k$, than the simultaneous distribution of $\underline{x}_1, \ldots, \underline{x}_k$ under the condition that their sum $\underline{X} = \Sigma \underline{x}_i$ has a given value X is given by a multinomial distribution with probabilities $p_i = \frac{\mu_i}{\Sigma \mu_i}$ and n trials:

$$(3) \quad P\{\underline{x}_1 = x_1; \ldots; \underline{x}_k = x_k / \underline{X} = X\} = \frac{X!}{\prod x_i!} \prod_i p_i^{x_i}.$$

and the simultaneous unconditional distribution of the x_i may be written as the product of (3) and the distribution of $\underline{X}$ which is again a Poisson distribution with mean $\Sigma \mu_i$.

If (3) is approximated in the usual way by a multinormal distribution (see for instance MOOD [2] p. 270 seq) it follows that

$$(4) \qquad \sum_{i=1}^{k} \frac{(\underline{x}_i - p_i X)^2}{p_i X}$$

has asymptotically for $X \to \infty$ a χ^2-distribution with $(k-1)$ degrees of freedom. In (4) X is a given quantity. The unconditional distribution of (4) is asymptotically for $\Sigma \mu_i \to \infty$ a χ^2 distribution as well (for a proof see e.g. J. VAN KLINKEN and H.J. PRINS [3] p. 16).

Thus if $\underline{x}_1, \ldots, \underline{x}_k$ are Poisson variates and the p_i are given, then

$$(5) \qquad \underline{z}_{1,k} \stackrel{def}{=} \sum_{i=1}^{k} \frac{(x_i - p_i \underline{X})^2}{p_i \underline{X}}$$

has asymptotically for $\Sigma \mu_i \to \infty$ a χ^2-distribution with $(k-1)$ degrees of freedom.

Now (5) is proposed as statistic for the test mentioned above with a critical region consisting of large values of $\underline{z}_{1,k}$ Asymptotically this test procedure is aequivalent to the appropriate likelihood ratio test (see VAN KLINKEN and PRINS 3 p. 24 seq.). Moreover the approximation is already for small values very close. If $p_i = 1/k$ $(i=1,\ldots,k)$ then for $X/k > 1$ and $k > 6$ the approximation is good enough for testing purposes on the 5% level. If the p_i do not differ widely from $1/k$ it seems that for the same values the approximation is also satisfying. This last result was found by investigating individual cases. A more detailed account of the closeness of approximation is given in [3] p. 33 seq. and in the papers mentioned there.

For small values of k and X it is feasible to construct an exact test, which is a simple summing of the probabilities (3) of all possible sets of observation (given X) arranged according to increasing $\underline{z}_{1,k}$. For $p_i = 1/k$ $(i=1,\dots,k)$ the order of magnitude of $\underline{z}_{1,k}$ is given by the order of magnitude of $\sum_{i=1}^{k} x_i^2$ which simplifies the procedure. Moreover in this last case a permutation of a given set of observations has the same probability (an example of this test is given in [3] p. 27 seq.)

As mentioned above the distribution (3) is a conditional distribution and so the exact test is a conditional test.

4. The two-way-classification

The variates of the classification scheme

$$(6)\qquad \begin{matrix} \underline{x}_{1,1}, & \dots, & \underline{x}_{1,n} \\ \vdots & & \vdots \\ \underline{x}_{m,1}, & \dots, & \underline{x}_{m,n} \end{matrix}$$

are supposed to be independent Poisson variates with means

$$(7)\qquad \begin{matrix} \mu_{1,1}, & \dots, & \mu_{1,n} \\ \vdots & & \vdots \\ \mu_{m,1}, & \dots, & \mu_{m,n} \end{matrix}$$

In order to derive tests for this case a similar property is used as was used for the one-way-classification. The simultaneous distribution of the $\underline{x}_{ij}$ is decomposed into the product of several distribution functions

$$(8)\qquad \prod_{i,j} \frac{e^{-\mu_{ij}}\mu_{ij}^{x_{ij}}}{x_{ij}!} = \left[\frac{e^{-M} M^{X}}{X!}\right]\cdot\left[\frac{X!}{\prod_i X_{i.}!}\prod_i p_{i.}^{X_{i.}}\cdot C\cdot\frac{X!}{\prod_j X_{.j}!}\prod_j p_{.j}^{X_{.j}}\right]\cdot\left[\frac{\prod_i X_{i.}!\,\prod_j X_{.j}!}{X!\,\prod_{ij} x_{ij}!}\;\frac{\prod_{ij} p_{ij}^{x_{ij}}}{C}\right]$$

with abbreviations:

$X_{i.} = \sum_j x_{ij}$; $X_{.j} = \sum_i x_{ij}$; $M = \sum_{ij}\mu_{ij}$; $M_{i.} = \sum_j \mu_{ij}$; $p_{i.} = \frac{M_{i.}}{M}$;

$M_{.j} = \sum_i \mu_{ij}$; $p_{.j} = \frac{M_{.j}}{M}$; $p_{ij} = \frac{\mu_{ij}\,M}{M_{i.}\,M_{.j}}$ and $C = \sum_{x_{ij}} \frac{\prod X_{i.}!\,\prod X_{.j}!\,\prod_{ij} p_{ij}^{x_{ij}}}{X!\,\prod_{ij} x_{ij}!}$

and C is a standardizing factor and depends on the p_{ij} and the $X_{i.}$ and $X_{.j}$. Thus $\mu_{ij} = M\,p_{i.}\,p_{.j}\,p_{ij}$. The summation $\sum_{x_{ij}}$ indicates the sum over all values of the x_{ij} for given totals $X_{i.}$, $X_{.j}$ and X.

In (8) the right side is the product of the distribution of the sum $\underline{X}$ of all observations, the simultaneous distribution of the marginal totals $\underline{X}_{i.}$ and $\underline{X}_{.j}$ given X and the distribution of the $\underline{x}_{ij}$ given the $X_{i.}$, the $X_{.j}$ and X. So except for the first one all distributions are conditional distributions. A special case arises if $p_{ij} = 1$ $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ or $\mu_{ij} = \frac{M_{i.} M_{.j}}{M}$ then $C = 1$, as the last distribution is a generalization of the well known hypergeometric distribution, the distribution of the observations in a $m \times n$ contingency table, and

(9)
$$\prod_{ij} \frac{e^{-\mu_{ij}} \mu_{ij}^{x_{ij}}}{x_{ij}} = \left[\frac{e^{-M} M^{X}}{X!}\right] \cdot \left[\frac{X!}{\prod_i X_{i.}!} \prod_i p_{i.}^{X_{i.}}\right] \cdot \left[\frac{X!}{\prod_j X_{.j}!} \prod_j p_{.j}^{X_{.j}}\right] \left[\frac{\prod_i X_{i.}! \prod_j X_{.j}!}{X! \prod_{ij} X_{ij}!}\right]$$

In this case the conditional distributions of the marginal totals given X are independent.

For tests relative to the means of the marginal totals the criteria of the preceding section are used. In the special case (9) these tests are, given X, independent asymptotically as well as exactly. The hypothesis tested are, for row totals

$$(10) \begin{cases} H_0: & M_{i.} = p_{i.} \cdot M \quad \text{with given } p_{i.} \quad (i = 1, \ldots, m) \\ H: & \text{at least for one } i \quad M_{i.} \neq p_{i.} M \end{cases}$$

and for column totals

$$(11) \begin{cases} H_0: & M_{.j} = p_{.j} M \quad \text{with given } p_{.j} \quad (j = 1, \ldots, n) \\ H: & \text{at least for one } j \quad M_{.j} \neq p_{.j} M \end{cases}$$

In the way stated in the preceding section exact test may be contructed.

We will now consider another test. If in (8) all $p_{ij} = 1$ we will say that no interaction is present, whereas if at least one $p_{ij} \neq 1$ we will state that there is interaction.

Consider the following test

$$(12) \begin{cases} H_0: & \text{all } p_{ij} = 1 \\ H: & \text{at least one } p_{ij} \neq 1 \end{cases}$$

Now under H_0 it follows from (9) that the conditional distribution of the $\underline{x}_{ij}$ given all marginal totals is

$$(13) \quad P\left[\underline{x}_{ij} = x_{ij} \mid X_{i.},\ X_{.j}\right] = \frac{\prod_i X_{i.}! \prod_j X_{.j}!}{X! \prod x_{ij}!}$$

This is the probability distribution of the contingency table with $m \times n$ cells. As for the contingency table tests the expression (13) may be approximated by a multinormal density and from this it follows that

$$\underline{Z}_{n.m,\, n+m-1} = \sum_{i,j} \frac{\left(\underline{x}_{ij} - \frac{X_{i.}\, X_{.j}}{X}\right)^2}{\frac{X_{i.}\, X_{.j}}{X}} = X \left(\sum_{i,j} \frac{\underline{x}_{ij}^2}{X_{i.} X_{.j}} - 1 \right)$$

has asymptotically a χ^2-distribution with $(n-1)(m-1)$ degrees of freedom. For the same reason as mentioned in section 2, the unconditional distribution of $\underline{Z}_{n.m,\, n+m-1}$ is asymptotically the same χ^2 distribution.

It may be proved along the same lines as in the one way classification, that the test is asymptotically aequivalent to the likelihood ratio test. As the hypotheses tested are aequivalent to those for the contingency table, all further information concerning the closeness of approximation, exact tests etc. may be found in any description of the contingency table tests (see for instance MOOD [2] p. 274 seq. and KENDALL [4] p. 3A seq. and literature quoted in the index of the latter).

This test allows a slight generalization. Suppose under H_0 the p_{ij} are not equal to one but have given values, then a test procedure for the following test may be derived:

$$(14) \quad \begin{cases} H_0 : & \text{all } p_{ij} \text{ have given values} \\ H : & \text{at least one } p_{ij} \text{ has not the given value} \end{cases}$$

Now from (8) it follows that the distribution of the $\underline{x}_{ij}$ is given by

$$(15) \quad P = P\left[\underline{x}_{ij} = x_{ij} \mid X_{i.}, X_{j.}, \Pi_{ij}\right] = \frac{1}{C} \frac{\prod_i X_{i.}!\, \prod_j X_{.j}!}{X!\, \prod_{ij} x_{ij}!} \prod_{ij} p_{ij}^{x_{ij}}$$

(15) may be approximated if $x_{ij} \neq 0$ for all i and j by a normal distribution. Using Stirling's formula

$$(16) \qquad x! = e^{-x} x^{x+1/2} \sqrt{2\pi}$$

(15) becomes:

$$(17) \quad (2\pi)^{-\frac{(m-1)(n-1)}{2}} \frac{1}{C} \frac{e^{-2X}}{e^{-2X}} \frac{\prod_i X_{i.}^{X_{i.}+1/2} \prod_j X_{.j}^{X_{.j}+1/2}}{X^{X+1/2} \prod_{ij} x_{ij}^{x_{ij}+1/2}} \prod_{ij} p_{ij}^{x_{ij}}$$

Now (17) equals

$$(2\pi)^{-\frac{(m-1)(n-1)}{2}} \cdot \frac{1}{c} \cdot \left(\prod_{i,j} \frac{X}{p_{ij} X_{i.} X_{.j}}\right)^{1/2} \left(\frac{\prod_i X_{i.} \prod_j X_{.j}}{X}\right)^{1/2} e^{-\sum_{i,j} (x_{ij}+1/2) \cdot \ln \frac{x_{ij}}{p_{ij}} \frac{X}{X_{i.} X_{j.}}}$$

Developing the logarithm of the power of e, and using the symbols $P_{ij} = \frac{p_{ij} X_{i.} X_{.j}}{X}$ and $\tilde{x}_{ij} = x_{ij} - P_{ij}$, the exponent of e becomes

$$(19) \quad -\sum_{i,j} \left(\tilde{x}_{ij} + P_{ij} + 1/2\right)\left(\frac{\tilde{x}_{ij}}{P_{ij}} - \frac{\tilde{x}_{ij}^2}{2P_{ij}^2} + \frac{\tilde{x}_{ij}^3}{3P_{ij}^3} - \dots\right)$$

The expansion of the logarithm is only valid for values of x_{ij} such that $\frac{\tilde{x}_{ij}}{P_{ij}} < 1$. We will require that $\frac{\tilde{x}_{ij}}{P_{ij}} = \alpha << 1$ then (18) equals

$$(19) \quad -\sum_{i,j} \left\{ 1/2 \, \frac{\tilde{x}_{ij}^2}{P_{ij}} \left(1 + O(\alpha)\right) + \tilde{x}_{ij}\left(1 + O(\alpha)\right) + O(\alpha) \right\}$$

As

$$\sum_{i,j} p_{ij} \, p_{i.} \, p_{.j} = 1$$

for arbitrary $p_{i.}$ and $p_{.j}$ (from the definition of p_{ij}),

$$\sum_{i,j} p_{ij} \frac{X_{i.}}{X} \cdot \frac{X_{.j}}{X} = 1$$

and so

$$\sum \tilde{x}_{ij} = 0$$

Thus (15) may be approximated by

$$(20) \quad P = (2\pi)^{-\frac{(m-1)(n-1)}{2}} \cdot \frac{1}{c} \left(\prod_{i,j} \frac{X}{p_{ij} X_{i.} X_{.j}}\right)^{1/2} \left(\frac{\prod_i X_{i.} \prod_j X_{.j}}{X}\right)^{1/2} e^{-\sum_{i,j} 1/2 \frac{\tilde{x}_{ij}}{P_{ij}}}$$

a multinormal density, if α is small, i.e. for values of x_{ij} such that $\tilde{x}_{ij} < \alpha . P_{ij}$

or for values of $\frac{\tilde{x}_{ij}^2}{P_{ij}}$ such that $\frac{\tilde{x}_{ij}^2}{P_{ij}} < \alpha^2 . P_{ij}$

The approximation is good if the approximation is good for an area of the normal density (20) which has a probability close to one. This is so if $\alpha^2 . P_{ij} > k$ being a positive number. For $\alpha = O(P_{ij}^{-1/2})$ and P_{ij} large these conditions are statisfied.

So (20) is valid if P_{ij} is large. From (20) it follows that

$$(21) \qquad Z_{m.n,m+n-1} = \sum_{i,j} \frac{\tilde{x}_{ij}^2}{P_{ij}}$$

has approximately a χ^2 distribution with $(m-1)(n-1)$ degrees of freedom. (21) is proposed as a test criterion for the test mentioned above. To the author's knowledge nothing is known about the closeness of the approximation. Exact tests may be carried out by summing the probabilities (15) in order of magnitude of the test criterion, and again the critical region consists of large values.

It must be remarked, that if at least one $p_{ij} \neq 1$ the tests for row and column means are dependent and may even be highly dependent as is easily seen by a simple example. In a $2 \times m$ scheme the means in the first column increase monotoneously and in the second column decrease monotoneously, then if the totals in the first row are large there is a large probability that the column total for the first column is high as well.

5. Other tests in a two way classification

It is possible to construct other tests for a two way classification. One will be given here. All of them are asymptotically aequivalent to likelihood ratio tests (see [3]).

In order to give a first example we will consider another decomposition of the simultaneous distribution.

$$(22) \quad \prod_{i,j} \frac{e^{-\mu_{ij}} \mu_{ij}^{x_{ij}}}{x_{ij}} = \left[\frac{e^{-M} M^X}{X!}\right] \left[\frac{X!}{\prod_i X_{i.}!} \prod_i p_{i.}^{X_{i.}}\right] \left[\prod_{i,j} \frac{X_{i.}!}{x_{ij}!} c_{ij}^{x_{ij}}\right]$$

when $\quad c_{ij} = \dfrac{\mu_{ij}}{M_{i.}} = p_{ij} \cdot p_{.j}$

In the last part of the right side of (22) in multinomial distributions are given and will be used for the following test

$$(23) \begin{cases} H_0 : \text{all } c_{ij} \text{ have given values} \\ H : \text{at least one } c_{ij} \text{ is not equal to the given value.} \end{cases}$$

A special case is

$$(24) \begin{cases} H_0 : \quad \mu_{i1} = \mu_{i2} = \cdots = \mu_{in} \qquad (i = 1, \ldots, k) \\ H : \text{one } \mu_{ij} \text{ does not follow the restriction under } H_0. \end{cases}$$

The test criterion is derived by summing all exponents of the individual normal approximations to the multinomial distributions. Thus a composite test is contructed from the one way classification test for every row.

The test criterion is

$$Z_{n,m,m} = \sum_{i,j} \frac{(x_{ij} - X_{i.}\, c_{ij})^2}{X_{i.}\, c_{ij}}$$

and has asymptotically a χ^2 distribution with $m(n-1)$ degrees of freedom. From the composition (22) it is clear that the test is independent of a test for the means of the row totals. It will be dependent of all the other tests discussed so far.

The above test may be carried out for columns in the same manner. An exact test would be very cumbersome as an extra summation occurs.

6. Multi-way classification

In order to derive tests for these cases the same procedure is used, i.e. developing the simultaneous distribution of the variates. The case of the 3-way classification will be given here as an example of the method. The variates are $\underline{x}_{ijk}$ with means μ_{ijk} $(i=1,\ldots,n;\ j=1,\ldots,n_2;\ k=1,\ldots,n_3)$

The decomposition of the simultaneous distribution of the $\underline{x}_{ijk}$ is arrived at in equating the simultaneous distribution of the $\underline{x}_{ijk}$ to the product of 4 factors i.e. the distribution of

a. the same of all observations

b. the joint distribution of the totals $\underline{X}_{i..} = \sum_{jk} \underline{x}_{ijk}$, $\underline{X}_{.j.} = \sum_{ik} \underline{x}_{ijk}$ and $\underline{X}_{..k} = \sum_{ij} \underline{x}_{ijk}$, given X

c. the joint distributions of the totals $\underline{X}_{ij.} = \sum_k \underline{x}_{ijk}$, $\underline{X}_{i.k} = \sum_j \underline{x}_{ijk}$ and $\underline{X}_{.jk} = \sum_i \underline{x}_{ijk}$ given the preceding totals

d. the distribution of the $\underline{x}_{ijk}$ given all totals

From these 4 types the marginal distribution of 3 are known except for interaction factors and standardizing factors. With due regard to the totals on which the standardizing factors depend, the following decomposition is obtained.

$$\prod_{ijk} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{x_{ijk}}}{x_{ijk}!} = \left[\frac{e^{-M} M^X}{X!}\right] \cdot \left[\left(X! \prod_i \frac{p_{i..}^{X_{i..}}}{X_{i..}}\right) \left(X! \prod_j \frac{p_{.j.}^{X_{.j.}}}{X_{.j.}!}\right) \cdot \right. \tag{25}$$

$$\left. \cdot \left(X! \prod_k \frac{p_{..k}^{X_{..k}}}{X_{..k}!}\right) \cdot C_1 \left(\underline{X}_{i..}, \underline{X}_{.j.}, \underline{X}_{..k}, p_{ij.}, p_{i.k}, p_{.jk}, p_{ijk}\right) \right] \cdot$$

$$\cdot\left[\left(\frac{\prod_i X_{i..}!\prod_j X_{.j.}!}{X!\prod_{i,j} \underline{X}_{ij.}}\prod_{i,j} p_{ij.}^{x_{ij.}}\right)\left(\ldots\right)\left(\ldots\right)\frac{1}{C_1}\cdot C_2\left(\underline{X}_{ij.},\underline{X}_{i.k},\underline{X}_{.jk}, p_{ijk}\right)\right]\cdot$$

(25)

$$\cdot\left[\frac{1}{C_2}\frac{X!\prod_{i,j} X_{ij.}!\prod_{i,k} X_{i.k}!\prod_{j,k} X_{.jk}!}{\prod_i X_{i..}!\prod_j X_{.j.}!\prod_k X_{..k}!\prod_{i,j,k} x_{ijk}!}\prod_{i,j,k} p_{ijk}^{x_{ijk}}\right]$$

with $p_{i..}=\frac{M_{i..}}{M}$ etc., $p_{ij.}=\frac{M_{ij.}M}{M_{i..}M_{.j.}}$ etc., $p_{ijk}=\frac{M_{i..}M_{.j.}M_{..k}\,\mu_{ijk}}{M\,M_{ij.}M_{i.k}M_{.jk}}$ etc., with the M's the same sums over the μ's as the X's are over the x's. Thus $\mu_{ijk}=M.\,p_{i..}\,p_{.j.}\,p_{..k}\,p_{ij.}\,p_{i.k}\,p_{.jk}\,p_{ijk}$ with restrictions on the p's:

$\sum_i p_{i..}=1$ etc. $\sum_i p_{i..}p_{.j.}\,p_{ij.}=1$ etc. and $\sum_i p_{i..}p_{ij.}\,p_{i.k}\,p_{ijk}=1$ etc.

he constants C_1 and C_2 serve two purposes; in the last distribution of (25) C_2 is a standardizing factor and depends on the $\underline{X}_{ij.}$, $\underline{X}_{i.k}$ and $\underline{X}_{.jk}$ and the p_{ijk}. In the last distribution but one C_2 is the interaction term demonstrating the independence of the three marginal distributions of the hypergeometric type, whereas C_1 is in that case the standardizing factor. In the second distribution from the top C_1 is the interdependence factor as it depends on $\underline{X}_{i..}$, $\underline{X}_{.j.}$ and $\underline{X}_{..k}$ and the p's.

The eight distributions between brackets will be used to derive tests for the various hypotheses, whereas the factors C_1 and C_2 demonstrate, that the tests thus obtained are dependent.

For the sake of simplicity we will only consider tests of a special type:

$$(26)\begin{cases} H_0: & p_{i..}=\frac{1}{m} \quad i=1,\ldots,m \\ H: & \text{at least one } p_{i..}\neq\frac{1}{m}\end{cases}$$

and so for the $p_{.j.}$ and $p_{..k}$;

$$(27)\begin{cases} H_0: & p_{ij.}=1 \quad i=1,\ldots,m \quad j=1,\ldots,n \\ H: & \text{at least one } p_{ij.}\neq 1\end{cases}$$

and so for the $p_{i.k}$ and $p_{.jk}$;

$$(28)\begin{cases} H_0: & p_{ijk}=1 \quad i=1,\ldots,m \quad j=1,\ldots,n \quad k=1,\ldots,r \\ H: & \text{at least one } p_{ijk}\neq 1.\end{cases}$$

For (26) the test criterion was derived in discussing the one way classification, for (27) the test criterion was derived in discussing the two way classification. For (28) the test criterion is

$$(29)\quad \Xi_{n.m.r,\; nm+nr+mr-n-m-r+1} = \sum_{i,j,k} \frac{\left(x_{ijk} - \dfrac{X_{ij.}X_{i.k}X_{.jk}X}{X_{i..}X_{.j.}X_{..k}}\right)^2}{\dfrac{X_{ij.}X_{i.k}X_{.jk}X}{X_{i..}X_{.j.}X_{..k}}}$$

The asymptotic distribution is a χ^2 with $(n-1)(m-1)(r-1)$ degrees of freedom. The derivation is obtained in the same way as that given in section 4, using Stirling and the first terms of the logarithmic expansion. The demonstration is rather extensive and won't be given here.

Test criteria for test where under H_0 the values of the p's are given, but not equal were given already in sections 3 and 4 except for the p_{ijk}.

Now that the decomposition is known for the 3-way classification it may by the same method be derived for a 4-way classification from these results and by iteration for the r way classification the numbers of tests of the various types are easily seen to be

1 test for the general mean

r tests for row- and column effects

$\binom{r}{2}$ tests for first order interaction

$\binom{r}{3}$ tests for second order interaction etc.

and for the number of degrees of freedom of the asymptotic distribution of χ^2, which equals the number of free variates in the test minus the number of linear restrictions as marginal totals have given values, is found

1	for the test for the general means
$(n_1-1), \dots, (n_r-1)$	for the tests for row- and column effects
$(n_1-1)(n_2-1), \dots, (n_{r-1}-1)(n_r-1)$	for the tests for first order interaction

etc.

The sum of these degrees of freedom must equal the number of independent variates as is easily seen to be the case

$$(30)\qquad \begin{aligned} &1+\sum_{i=1}^{r}(n_i-1)+\sum_{\substack{i,j\\ i\neq j}}(n_i-1)(n_j-1)+\dots+(n_1-1)\cdots(n_r-1)= \\ &\prod_{i=1}^{r}\left\{(n_i-1)+1\right\}=n_1\dots n_r \end{aligned}$$

It is seen that the notion degrees of freedom, as defined by: the number of variates minus the number of restrictions on these variates under a specified hypothesis is of some importance in this type of analysis.

In the general case the tests may be carried out by computing the exact probabilities in the way as was indicated above, but already for a second order interaction test this is very cumbersome.

7. Tests for goodness of fit for Poisson variates and their dependence on the tests discussed above

If $x_1, \dots, x_k$ are independent discrete variates known to have the same distribution there exists a test, to test

$$(31)\qquad \begin{cases} H_0: & \text{the } x_i \text{ follow the Poisson law} \quad (\mu=\sigma^2) \\ H: & \text{the } x_i \text{ follows a law with either } \mu<\sigma^2 \text{ or } \mu>\sigma^2 \end{cases}$$

This test is well known and the test criterion is a special case of the test discribed in the one way classification:

$$(32)\qquad \underline{a}=\frac{\sum(x_i-\bar{x})^2}{\bar{x}}=\frac{\sum x_i^2}{\bar{x}}-\bar{x}$$

and has asymptotically the same distribution as χ^2 with $(k-1)$ degrees of freedom.

The test is described in further detail in [3]. It is usefull if the following kind of analysis occurs.

It is wanted to investigate the influence of various factors on the mean of a discrete variate, but it is not known whether the discrete distribution is a Poisson distribution. For instance in a cotton mill one wants to study the influence of various factors on the mean number of threadbreaks, the number of threadbreaks being the discrete variate.

Some classification scheme is developped to enable an efficient analysis, but instead of one observation per cell of the classification scheme, more observations are done. The sum of these observations per cell is used for the analysis of the factorial design , whereas the values of the observations constituting the sum are used to test for every cell whether the observations come from a Poisson population. It may be that one wants an over all test and then the best procedure seems to sum the a's as the asymptotic distribution of the sum is again a χ^2 distribution.

The tests for goodness of fit in the scheme are asymptotically independent of the tests of the analysis of the factorial design, as the subdivision of the sum in a number of variates is independent of the value of this sum.

The test may be incorporated in the decomposition scheme used in the previous sections. In the left side of the decompositions (9) and (25) the various x are now supposed to be the sums per cell. For one cell by the property (9), if $x_1, \ldots, x_k$ are the observations within the cell and x is their sum and μ their mean

$$(33) \quad \prod_i \frac{e^{-\mu}\mu^{x_i}}{x_i!} = \left[\frac{e^{-k\mu}(k\mu)^x}{x!}\right] \cdot \left[\frac{x!}{\prod x_i!}\left(\frac{1}{k}\right)^x\right]$$

In (33) the first distribution in the right hand member is the distribution of the cell sum and so if right and left side of the decomposition for the factorial design are multiplied by the second factor of (33), the simultaneous distribution of all vari ates is decomposed and the new decomposition contains the product of the simultaneous distribution of the variates within cells, given the cell sums.

Remarks

BARTLETT [1] and other authors discussed an analysis of factorial designs procedure for Poisson variates using a transform of the variates $\sqrt{x}$ or $\sqrt{x+a}$ which is approximately normal, with known variances.

The method proposed in this report seems to have some advantages over the one stated above.

Some of the differences between the methods will be given here briefly.

1. Using the $\sqrt{x}$ transform method, it is difficult to state in terms of the means μ what the null hypothesis for interaction is. For large μ, we have $\mathcal{E}\sqrt{x} \approx \sqrt{\mu}$ and the hypothesis is less complicated. The method used above states the hypotheses tested in a concise form in terms of the means.
2. For small means an exact test for the $\sqrt{x}$ transform method is difficult and the normal approximation seems rather poor, whereas in the preceding sections it was stated that such tests are not too cumbersome for the method described there.
3. The exact distributions of the test criteria for the tests in this report are not simple. An asymptotic distribution is derived, the closeness of approximation of which seems not so difficult to judge as in the $\sqrt{x}$ analysis.
4. Computations in the case of the $\sqrt{x}$ transform necessitate the computation of the roots of all observations.

We may consider the situation from anotherpoint of view. It is desired to test certain specified hypotheses, which are specified b the nature of an experiment. If these hypotheses are of the type discussed in this report, it seems best to use the tests proposed here. In terms of FISHER's concepts of amount of information and sufficient statistics.

The decomposition method gains importance, as the total amount f information concerning the μ is decomposed in the total amount of information concerning the transforms p of the means. Thus the total amount of information on the $p_{i.}$ in a two way classification is given by

$$\frac{X!}{\prod_i X_{i.}!} \prod_i p_{i.}^{X_{i.}}$$

and in no other part of the decomposition the $p_{i.}$ occur.

It is seen that the only reason for using the type of analysis proposed in this report is, that it is most suitable for a certain set of hypotheses. Whether these hypotheses are usefull in a practical situation is undecided. As the type of analysis proposed here has a certain mathematical straight forwardness, it would be satisfactory if the hypotheses considered have general practical importance.

References

[1] M.S. BARTLETT, The square-root transformation in the analysis of variance. Supp. Jrn. Royal Stat.Soc., 3, 1936, pp. 68-78.

[2] A.M. MOOD, Introduction to the theory of statistics, Mac Graw-Hill, New York, 1950.

[3] J. VAN KLINKEN and H.J. PRINS, Overzicht van toetsings- en schattingsmethoden met betrekking tot Poisson-verdelingen (Survey of tests for Poisson variates). Report S 133 (Ov3) of the statistical department of the Mathematical Centre, Amsterdam, 1954.

[4] M.G. KENDALL, The advanced theory of Statistics, London, 1947.