SOME MATRIX OCCUPANCY PROBLEMS
WITH DICHOTOMOUS ENTRIES

Lars Holst

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

August 1976
(Received June 4, 1976)

18p.

MRC-TSR-1663

DAAG29-75-C-0024

Approved for public release
Distribution unlimited

Sponsored by
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709
An $R \times N$ matrix is generated in the following way. In each row a predetermined number of positions are randomly assigned the value 1; the remaining positions are assigned the value 0. For each column a real valued function of the elements is given. In this paper the sum of the values of these functions is studied when $N \to \infty$. The results can be applied to e.g. "committee" problems and contingency tables of 0-1-variables.
1. Introduction

In Feller (1968), p. 112, exercise 16, the following problem is posed:

"A cell contains N chromosomes, between any two of which an interchange of parts may occur. If R interchanges occur (which can happen in \( \binom{N}{2}^R \) distinct ways), find the probability that exactly m chromosomes will be involved".

Various generalizations of this problem have been considered as e.g. committee problems. A typical generalization is: "R committees are formed from N individuals, the j:th committee has size \( n_j \), the committees are formed by independent simple random sampling, find the distribution of the number of individuals belonging to all committees".

This type of problem has also been applied to a certain health problem, see Mantel (1974) and the references given there.

Let us consider the following situation. Consider a finite population of N units. Take R independent simple random samples without replacement of sizes \( n_1, \ldots, n_R \). Define \( Y_{jk} = 1 \), if the \( k \)th unit occur in the \( j \)th sample, otherwise let \( Y_{jk} = 0 \) for \( j = 1, \ldots, R \) and \( k = 1, \ldots, N \). Consider the \( R \times N \) matrix or special contingency table with the \( Y \)'s as entries. This matrix has fixed row totals \( n_1, \ldots, n_R \).
and the different rows are independent random vectors. Let $f_1, \ldots, f_N$ be given real functions defined on $\{0, 1\}^R$ and consider the random variable:

$$Z_N = \sum_{k=1}^{N} f_k(Y_{1k}, \ldots, Y_{Rk}).$$

How is $Z_N$ distributed?

In this paper limit distributions of $Z_N$ are derived when $N \to \infty$. Similar results have been proved by Eicker et al. (1972) using different methods. Some limit results can also be obtained from Theorem 2 in Holst (1973). As an application the limiting distribution of a chi-square statistic proposed in Mantel (1974) is discussed.

Some words about notation. We set, for all $j$, $p_j = n_j/N$. To state the limit theorems properly we should use an extra index $\nu$, but to facilitate notation we suppress $\nu$. In the following it is always assumed that $N \to \infty$ and $n_j \to \infty$ in such a way that

$$Np_j(1 - p_j) \to +\infty, \quad j = 1, \ldots, R.$$ 

A random variable $X$ with $P(X = 1) = 1 - P(X = 0) = p$ is called $\text{Be}(p)$.

The binomial distribution with parameters $n$ and $p$ is abbreviated $\text{Bin}(n, p)$, $N(m, \Sigma)$ is the normal with mean $m$ and variance (matrix) $\Sigma$, $\text{Po}(m)$ the Poisson with mean $m$ and $\chi^2(f)$ the chi-square with $f$ degrees of freedom. The integer part of a real number $A$ is $[A]$. We write converges in distribution as $\mathcal{F} \to \cdots$. In the following $X_{jk}$
for $j = 1, \ldots, R$ and $k = 1, \ldots, N$ will always be independent random variables where $X_{jk}$ is $\text{Be}(p_j)$. We set $X_j = \sum_{k=1}^{N} X_{jk}$. Note that $X_j$ is $\text{Bin}(N, p_j)$. 
2. The characteristic function

Let \( X' \)s and \( Y' \)s be defined as above and consider for a given function \( f \) the random variables

\[
Z = f(Y_{11}, \ldots, Y_{RN})
\]

and

\[
U = f(X_{11}, \ldots, X_{RN}).
\]

Theorem 1. The following relation holds

\[
E(e^{itZ}) = a_N \cdot \prod_{j=1}^{R} (\frac{Np_j(1-p_j)}{2\pi})^{1/2} \cdot \\
\cdot \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \exp(itU + \sum_{j=1}^{R} \theta_j(X_j - Np_j))d\theta_1 \cdots d\theta_R
\]

where

\[
a_N = \frac{\prod_{j=1}^{R} n_j!}{(N-p_j)^{N-n_j} \cdot (2\pi Np_j(1-p_j))^{1/2} \cdot (2\pi n_j)!}.
\]

Proof. The conditional distribution of \((X_{j1}, \ldots, X_{jN})\) given \(X_j = n_j\) is the same as the distribution of \((Y_{j1}, \ldots, Y_{jN})\). Therefore we have

\[
E(e^{itZ}) = E(e^{itU} | X_j = n_j, j = 1, \ldots, R).
\]

The distribution is given by

\[
P(Y_{j1} = v_{j1}, \ldots, Y_{jN} = v_{jN}) = \frac{1}{n_j!}
\]

for \(v_{jk} = 0, 1\) and \(v_j = \sum_{k=1}^{N} v_{jk} = n_j\). Furthermore, we have
\[
\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} E(\exp(itU + i \sum_{j}^{R} \theta_j(X_j, N))d\theta_1 \cdots d\theta_R \\
= \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \sum_{j}^{R} \exp(itf(\nu_j) + i \sum_{j}^{R} \theta_j(\nu_j - n_j)) \cdot \\
\cdot \prod_{j,k} p_j^k (1-p_j)^{1-v_j} d\theta_1 \cdots d\theta_R = \\
= \sum_{\nu} \exp(itf(\nu)) \prod_{j=1}^{R} (\int_{-\pi}^{\pi} e^{i\theta_j(\nu - n_j)} d\theta_j \cdot p_j^{\nu_j} (1-p_j)^{N-\nu_j}) = \\
= \sum_{\nu_1=1}^{\infty} \cdots \sum_{\nu_R=1}^{\infty} \exp(itf(\nu)) \cdot (2\pi)^R \prod_{j=1}^{R} P(X_j = n_j) = \\
= E(e^{itZ}) \cdot (2\pi)^R \prod_{j=1}^{R} P(X_j = n_j),
\]

from which the assertion follows.

Using Stirling's formula we get:

**Lemma.** If \( Np_j(1-p_j) \to \infty \), all \( j \), then \( \alpha_N \to 1 \). If \( p_j \to \gamma_j \), \( 0 < \gamma_j < 1 \), all \( j \), then \( \alpha_N = 1 + O(1/N) \).
3. Limit theorems

In this section we consider random variables of the form

\[ Z_N = \sum_{k=1}^{N} f_k(Y_{1k}, \ldots, Y_{Rk}). \]

As \((X_{1k}, \ldots, X_{Rk})\) and \((Y_{1k}, \ldots, Y_{Rk})\) have the same distribution, they also have the same moments so

\[ E(f_k(X_{1k}, \ldots, X_{Rk})) = E(f_k(Y_{1k}, \ldots, Y_{Rk})) = \mu_k. \]

**Theorem 2.** Suppose that there exists a \( q_0 < 1 \) such that for \( q > q_0 \),

\[ M = \lfloor Nq \rfloor \]

and

\[ U_M = \sum_{k=1}^{M} f_k(X_{1k}, \ldots, X_{Rk}) \]

we have

\[
\mathcal{L}(U_M - \sum_{k=1}^{M} \mu_k)
\]

\[
\begin{pmatrix}
\sum_{k=1}^{M} \frac{(X_{1k} - p_1)/(Np_1(1 - p_1))^{1/2}}{Np_1(1 - p_1)} \\
\vdots \\
\sum_{k=1}^{M} \frac{(X_{Rk} - p_R)/(Np_R(1 - p_R))^{1/2}}{Np_R(1 - p_R)}
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 \\
B_{1q} \\
\vdots \\
B_{Rq}
\end{pmatrix}
\]

\[
\begin{pmatrix}
A_q \\
B_{1q} \\
\vdots \\
B_{Rq}
\end{pmatrix}
\]

\[
\begin{pmatrix}
B_{1q} \\
\vdots \\
B_{Rq}
\end{pmatrix}
\]

where \( A_q = A_1, B_{jq} = B_j, \) all \( j, \) when \( q \rightarrow 1. \) Then when \( N \rightarrow \infty \)

\[ \mathcal{L}(Z_N - \sum_{k=1}^{N} \mu_k) \rightarrow N(0, A_1 - \sum_{j=1}^{R} B_j^2). \]
Proof. Without loss of generality we can suppose that \( \mu_k = 0 \). Define \( Z_M \)

analogously to \( U_M \). Set \( X_j \cdot M = \sum_{k=1}^{M} X_{jk} \) and \( X'_j \cdot M = \sum_{k=M+1}^{N} X_{jk} = X_j \cdot M' - X_j \cdot M' \).

Let us also introduce \( \sigma_j = (Np_j (1 - p_j))^{1/2} \). From Theorem 1 it follows that

\[
E(e^{itZ_M}) = \alpha_N \cdot (2\pi)^{-R/2} \cdot \sigma_1 \cdots \sigma_R \cdot \\
\cdot \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} E(\exp(itU_M + i \sum_{j=1}^{R} \theta_j (X_j \cdot M' - Mp_j))) \cdot \\
\cdot E(\exp(i \sum_{j=1}^{R} \psi_j (X'_j \cdot M - (N - M)p_j)))d\theta_1 \cdots d\theta_R = \\
= \alpha_N \cdot (2\pi)^{-R/2} \int_{-\pi\sigma_1}^{\pi\sigma_1} \cdots \int_{-\pi\sigma_R}^{\pi\sigma_R} E(\exp(itU_M + i \sum_{j=1}^{R} \psi_j (X_j \cdot M' - Mp_j)))d\theta_1 \cdots d\theta_R = \\
\cdot E(\exp(i \sum_{j=1}^{R} \psi_j (X'_j \cdot M - (N - M)p_j)))d\psi_1 \cdots d\psi_R = \\
= \alpha_N (2\pi)^{-R/2} \int \cdots \int g_M(t, \psi_1, \ldots, \psi_R) h_M(\psi_1, \ldots, \psi_R) d\psi_1 \cdots d\psi_R.
\]

From DeMoivre's theorem we have

\[
\ln h_M(\psi_1, \ldots, \psi_R) = \exp(-\frac{1}{2} \sum_{j=1}^{R} \psi_j^2 (1 - q)) = h(\psi_1, \ldots, \psi_R),
\]

and it is not difficult to prove that

\[
\int |h_M| = \int h.
\]

From the assumptions we get

\[
g_M(t, \psi_1, \ldots, \psi_R) = \exp(-\frac{1}{2} (t^2 A_q + 2t \sum_{j=1}^{R} \psi_j B_{jq} + \sum_{j=1}^{R} \psi_j^2 q)) .
\]
As \[ |g_M| \leq 1 \text{ and } a_N \to 1 , \] it follows from the extended Lebesgue Convergence Theorem, see Rao (1973), p. 136, that
\[
\mathbb{E}(e^{-tZ_M}) \to (2\pi)^{-R/2} .
\]
\[
\int \cdots \int \exp\left( -\frac{1}{2} \left( t^2 a_q + 2t \sum_{j=1}^{R} \psi_j b_{jq} + \sum_{j=1}^{R} \psi_j^2 \right) \right) d\psi_1 \cdots d\psi_R
\]
\[
= \exp\left( -\frac{1}{2} t^2 (a_q - \sum_{j=1}^{R} b_{jq}^2) \right) .
\]
Thus we have proved that
\[ \mathcal{L}(Z_M) \to N(0, A_q - \sum_{j=1}^{R} b_{jq}^2) . \]
Analogously we prove
\[ \mathcal{L}(Z_N - Z_M) \to N(0, A_1 - A_q - \sum_{j=1}^{R} (b_j - b_{jq})^2) . \]
The assumptions imply that
\[ A_q - \sum_{j=1}^{R} b_{jq}^2 \to A_1 - \sum_{j=1}^{R} b_j^2 \]
and
\[ A_1 - A_q - \sum_{j=1}^{R} (b_j - b_{jq})^2 \to 0 . \]
Using the argument by LeCam (1958), p. 13-14, we obtain
\[ \mathcal{L}(Z_N) \to N(0, A_1 - \sum_{j=1}^{R} b_j^2) . \]
When the same function \( f \) is used for each column then we get a simple and useful theorem.

**Theorem 3.** Suppose that for some random vector \((U, V_1, \ldots, V_R)\) and
\[
s_j = (Np_j(1 - p_j))^{1/2}
\]
we have
\[
\begin{pmatrix}
U - Np_j / \sigma_j \\
X_1 - Np_1 / \sigma_1 \\
\vdots \\
X_R - Np_R / \sigma_R
\end{pmatrix}
\rightarrow
\begin{pmatrix}
U \\
V_1 \\
\vdots \\
V_R
\end{pmatrix}
\]
Then the infinitely divisible random vector has the characteristic function
\[
E(e^{itU}) = \phi(t) = e^{\frac{1}{2} t^2 A + 2 \sum_{j=1}^{R} t s_j B_j + \sum_{j=1}^{R} s_j^2}
\]
where
\[
E(e^{itU}) = \phi(t) = e^{-\frac{1}{2} t^2 A}
\]
and \( \phi \) has no normal component. Furthermore
\[
\mathcal{L}(Z_N - N \mu) \rightarrow \mathcal{L}(Z)
\]
where the random variable \( Z \) has the characteristic function
\[
E(e^{itZ}) = \phi(t) = \exp(-\frac{1}{2} t^2 (A - \sum_{j=1}^{R} B_j^2))
\]

**Proof.** The first part of the assertion follows from classical limit theorems, cf. LeCam (1958), p. 8.

Without loss of generality let us assume that \( \mu = 0 \). We observe that with \( M \) and \( Nq \), \( 0 < q < 1 \),
\begin{equation*}
E(\exp(it U_M + \frac{1}{M} \sum_{j=1}^{R} s_j (X_j - M\mu_j/\sigma_j))) \to (H(t, s))^{\mathbb{Q}} ,
\end{equation*}

using the notation of the previous proof. The assertion then follows as in Theorem 2. 

\textbf{Remark 1.} If \( p_j \to \gamma_j, 0 < \gamma_j < 1, \) all \( j, \) then the limit distribution has no non-normal component, because \( f_N(X_{1_k}, \ldots, X_{R_k}) \) can only take at most \( 2^R \) different values. Hence non-normal limits can only occur when some \( p_j \to 0 \) or 1.

\textbf{Remark 2.} Many of the theorems in Eicker et al. (1972) are special cases of Theorem 3.

The case with no normal component is particularly simple.

\textbf{Theorem 4.} Suppose that

\begin{equation*}
\mathcal{L}(U_N - N\mu) \to \mathcal{L}(U),
\end{equation*}

where \( U \) has no normal component. Then

\begin{equation*}
\mathcal{L}(Z_N - N\mu) \to \mathcal{L}(U).
\end{equation*}

\textbf{Proof.} As \( U_N - N\mu \) and \( (X_j - N\mu_j)/\sigma_j, j = 1, \ldots, R, \) converges in distribution it follows that we can from every subsequence of the vectors in Theorem 3 select a convergent subsequence. For such a convergent subsequence we can apply Theorem 3. As \( U \) had no normal component we have \( A = 0 \) and so \( B_1 = \cdots = B_R = 0. \) Thus the limiting characteristic function is just \( \phi(t). \) As this limit does not depend on the particular subsequence it follows that

\begin{equation*}
E(\exp(it(U_N - N\mu))) \to \phi(t),
\end{equation*}

which proves the theorem. 

-10-
If we consider sequences such that \( f_1 = \cdots = f_N = f \) and \( p_j = \gamma_j \), \( 0 < \gamma_j < 1 \), independent of \( N \), then the following local limit theorem hold.

**Theorem 5.** Suppose that the random variable \( f(Y_1, \ldots, Y_R) \) is integer one-lattice. Then uniformly in \( \nu \) when \( N \to \infty \)

\[
P\left( \sum_{k=1}^{N} f(Y_{1k}, \ldots, Y_{Rk}) = \nu \right) \approx (N\sigma^2(1 - p^2))^{1/2} \cdot (2\pi)^{1/2} \cdot \exp\left(-\frac{(\nu - Np)^2}{2N\sigma^2(1 - p^2)}\right) \to 0,
\]

where

\[
\mu = \mathbb{E} f(Y_1, \ldots, Y_R),
\]

\[
\sigma^2 = \text{Var} f(Y_1, \ldots, Y_R)
\]

and

\[
\rho^2 = \sum_{j=1}^{R} \frac{(\text{Cov}(f(Y_1, \ldots, Y_R), Y_{R_j}))^2}{\sigma^2 Y_j (1 - Y_j)}.
\]

**Proof.** Using Theorem 1 and the inversion formula for characteristic functions we obtain

\[
P\left( \sum_{k=1}^{N} f(Y_{1k}, \ldots, Y_{Rk}) = \nu \right) = \left(2\pi\right)^{(R+2)/2} \cdot \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \mathbb{E}(\exp(itU_N + i \sum_{j=1}^{R} \theta(X_j, - Np_j))) e^{-it\nu} dt \cdots d\theta_R
\]

\[
= \left(2\pi\right)^{R/2} \cdot P(U_N = \nu, X_j = n_j, j = 1, \ldots, R)
\]

\[
= \left(2\pi\right)^{R/2} \cdot P(\nu, N, n_1, \ldots, n_R).
\]

Using a multidimensional local limit theorem by Rvàčeja (1954), p. 202, we have uniformly in \( \nu \)
\[ P(v, N, n_1, \ldots, n_R) \cdot (2\pi N\sigma^2 (1 - \rho^2))^{1/2} \cdot \exp\left(-\frac{(v - N\mu)^2}{2N\sigma^2 (1 - \rho^2)} + 0\right) \to 0. \]

(Note that \( (Y_{11}, \ldots, Y_{R1}) \) and \( (X_{11}, \ldots, X_{R1}) \) have the same distribution and therefore the same moments.) Combining the last two expressions and using the lemma of Section 2 proves the assertion.)
4. Applications

**Example 1.** Suppose that the number of columns with no zeros are of interest. In the committee problem this corresponds to the number of persons which are members of all committees. The appropriate function for this case is

$$f(Y_1, \ldots, Y_R) = Y_1 \cdot \ldots \cdot Y_R.$$  

Suppose now that $N, n_1, \ldots, n_R \to \infty$ so that $Np_1 \cdot \ldots \cdot p_R \to \lambda$, $0 < \lambda < \infty$. We have

$$P(f(X_{IL}, \ldots, X_{RI}) = 1) = p_1 \cdot \ldots \cdot p_R$$

so from the usual Poisson approximation of the binomial it follows that

$$\sum_{k=1}^{N} f(X_{1k}, \ldots, X_{Rk}) \to \text{Po}(\lambda).$$

Hence by Theorem 4 the number of columns with no zero is in the limit $\text{Po}(\lambda)$ - distributed.

**Example 2.** In connection with a health research problem Mantel (1974) proposes tests for differences between columns. Essentially using the sample variance of the column totals is suggested as a test-statistic. In our notation Mantel's statistic could be written

$$Z_N = (N - 1)\left( \sum_{k=1}^{N} \sum_{j=1}^{R} (Y_{jk} - p_j)^2 / N \sum_{l=1}^{R} p_j(1 - p_j) \right).$$

Using normal random variables Mantel approximates $Z_N$'s distribution by $\chi^2(N - 1)$. A limit distribution of $Z_N$ can be obtained from
Theorem 3 if \( n_j/N \rightarrow \gamma_j, \ 0 < \gamma_j < 1 \). The limit law has no nonnormal component and after some calculations one finds

\[
A - \sum_{j=1}^{R} B_j^2 = 2(\sum_{j=1}^{R} \gamma_j(l - \gamma_j))^2 - \sum_{j=1}^{R} (\gamma_j(l - \gamma_j))^2.
\]

We may note that in fact the exact mean and variance are

\[
EZ_N = N - 1
\]

\[
\text{Var} Z_N = 2(N - 1) \cdot (1 - \sum_{j=1}^{R} (p_j(l - p_j))^2 / (\sum_{j=1}^{R} p_j(l - p_j))^2).
\]

By Theorem 3 we can state

\[
\chi^2((Z_N - (N - 1))/(2(N - 1))^{1/2}) \rightarrow N(0, 1 - K)
\]

where

\[
K = \sum_{j=1}^{R} (\gamma_j(l - \gamma_j))^2 / (\sum_{j=1}^{R} \gamma_j(l - \gamma_j))^2.
\]

So the limit distribution has smaller variance than the chi-square approximation indicates. By the Cauchy-Schwarz inequality we get \( K \geq 1/R \) with equality if and only if \( \gamma_1 = \ldots = \gamma_R \). Unless \( R \) is big and the \( \gamma \)'s are not too unequal this approximation is likely to give a conservative test.

Mantel also discusses a \( \chi^2(1) \) approach. This is actually the same as using the normal approximation suggested by the limit law above.

As Mantel points out, the distribution of \( Z_N \) is right skew in typical cases. Thus the normal approximation may be inappropriate. A right skew distribution having the right mean, variance and limit law is

\[
(1 - K) \cdot \chi^2((N - 1)/(1 - K)).
\]

One may expect that this distribution better approximates that of \( Z_N \) than any of the other approximations.
REFERENCES


An $R \times N$ matrix is generated in the following way. In each row a pre-determined number of positions are randomly assigned the value 1; the remaining positions are assigned the value 0. For each column a real valued function of the elements is given. In this paper the sum of the values of these functions is studied when $N \leq R$. The results can be applied to e.g. "committee" problems and contingency tables of 0-1-variables that approaches $1_{\text{inf}}$ of infinity.