AD-A083 525 CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES F/G 12/1
POLARIZATION TEST FOR THE MULTINOMIAL DISTRIBUTION.(U)
FEB 79 K ALAM, A MITRA N00014-75-C-0451
UNCLASSIFIED N105 NL

1 OF 1
AD A
083525

END
DATE
FILMED
DTIC

12 LEVEL II

B.S.

DEPARTMENT OF MATHEMATICAL SCIENCES

CLEMSON UNIVERSITY
Clemson, South Carolina

CLEMSON UNIVERSITY
SOUTH CAROLINA
1889

POLARIZATION TEST FOR THE MULTINOMIAL DISTRIBUTION,

BY

Khursheed Alam & Amitava Mitra

Clemson University

Report N105

Technical Report #299

February, 1979

Research Supported in part by

THE OFFICE OF NAVAL RESEARCH

Task NR ~~042-271~~ Contract N00014-75-C-0451
047 302

POLARIZATION TEST FOR THE MULTINOMIAL DISTRIBUTION

Khursheed Alam* & Amitava Mitra

Clemson University & University of Southern California

ABSTRACT

The vectors $\underset{\sim}{p} = (p_1,\ldots,p_k)'$ representing the cell probabilities of a multinomial distribution are partially ordered according to the majorization relation: $\underset{\sim}{p}$ majorizes $\underset{\sim}{p}'$ ($\underset{\sim}{p} \succ \underset{\sim}{p}'$) if $\Sigma_{i=1}^{j} p_{(i)} \geq \Sigma_{i=1}^{j} p'_{(1)}$, $j = 1,\ldots,k$, where $p_{(i)}$ denotes the ith largest value among $p_1,\ldots,p_k$. If the reverse inequality holds we say that $\underset{\sim}{p}$ minorizes $\underset{\sim}{p}'$ ($\underset{\sim}{p} \prec \underset{\sim}{p}'$). In this paper we consider a test of the hypothesis H: $\underset{\sim}{p} \prec \underset{\sim}{p}^o$ against the alternative hypothesis H': $p \succ \underset{\sim}{p}^o$, where $\underset{\sim}{p}^o$ is a given vector. The test discriminates between the situations where the total multinomial probability is distributed more or less evenly among the k cells. It is therefore called a polarization test.

Key words: Multinomial Distribution; Schur-Convex Function

AMS Classification: 62F05

*The author's work was supported by the Office of Naval Research under Contract No. N00014-75-C-0451

1. Introduction. Let $M(\underset{\sim}{x};\ \underset{\sim}{p}, n)$ denote the multinomial distribution, where $\underset{\sim}{x} = (x_1,\dots,x_k)'$ denotes the vector of cell probabilities, $\underset{\sim}{p} = (p_1,\dots,p_k)'$ denotes the vector of cell frequencies, $\Sigma_{i=1}^{k} x_i = n$ and $\Sigma_{i=1}^{k} p_i = 1$. Consider a partial ordering of the probability vectors $\underset{\sim}{p}$, given by the majorization relation: $\underset{\sim}{p}$ majorizes $\underset{\sim}{p}'$ ($\underset{\sim}{p} \succ \underset{\sim}{p}'$) if $\Sigma_{i=1}^{j} p_{(i)} \geq \Sigma_{i=1}^{j} p'_{(i)}$, $j = 1,\dots,k$, where $p_{(i)}$ denotes the ith largest value among $p_1,\dots,p_k$. If the reverse inequality holds we say that $\underset{\sim}{p}$ minorizes $\underset{\sim}{p}'$ ($\underset{\sim}{p} \prec \underset{\sim}{p}'$). A symmetric function f is said to be Schur-convex if $f(\underset{\sim}{p}) \geq f(\underset{\sim}{p}')$ for all $\underset{\sim}{p} \succ \underset{\sim}{p}'$. For example, $Q = \Sigma_{i=1}^{k} p_i^2$ is a Schur-convex function. Clearly, $\frac{1}{k} \leq Q \leq 1$. If the k multinomial events are nearly equally probable then the value of Q is close to its lower bound. On the other hand if the total probability is almost concentrated into a single cell then the value of Q is close to 1. Thus, the value of Q measures, so to speak, "polarization" of the multinomial distribution.

More generally, the multinomial distribution associated with $\underset{\sim}{p}$ is said to be more polarized than the multinomial distribution associated with $\underset{\sim}{p}'$ if $\underset{\sim}{p} \succ \underset{\sim}{p}'$. Note that the vector $(\frac{1}{k},\dots,\frac{1}{k})$ is majorized by every vector $\underset{\sim}{p}$. In this paper we consider a test of the hypothesis, $H : \underset{\sim}{p} \prec \underset{\sim}{p}^o$ against the alternative hypothesis $H' : \underset{\sim}{p} \succ \underset{\sim}{p}^o$, where $\underset{\sim}{p}^o$ is a given value of $\underset{\sim}{p}$. The test is based on the statistic $T = (\Sigma_{i=1}^{k} x_i^2)/n$, rejecting H for large values of T. We call it a polarization test. We note that the polarization test is a one-sided test, whereas, Pearson's Chi-square test for goodness of fit is two-sided, designed to test the hypothesis that $\underset{\sim}{p} = \underset{\sim}{p}^o$ against the alternative hypothesis that $\underset{\sim}{p} = \underset{\sim}{p}^o$. For

$k = 2$ the hypothesis H states that $|p_1 - p_2| \leq |p_1^o - p_2^o|$, while the reverse inequality holds for H'.

The problem of testing H against H' arises in various situations. Suppose, for example, that k political parties are contesting in an election. Let p_i denote the proportion of voters in favor of the ith party $(i = 1,\dots,k)$ at a certain period of time before the election. It might be of interest to know at a subsequent period of time before the election whether, due to the emergence of certain issue or the occurrence of certain event, the voting preference had polarized in the sense that a single party or, at most, a few parties out of the k parties would share together almost all the votes.

For another example, suppose that the population of a variety of fish is spread out in certain parts of a lake. It might be of interest to know whether the fish population had concentrated into fewer parts of the lake at a certain time, that is, the fish population had polarized due to a change in weather condition or some other factor.

In the following section it is shown that the given test is unbiased. For the application of the test we need to know the distribution of T. Formulas for exact as well as the asymptotic distribution for large n are given. Numerical results are given showing asymptotic convergence of the distribution.

2. Polarization test. Consider the hypothesis H. We reject H for large values of T. By Theorem 3.7 of Hollander, Proschan and Sethuraman (1977)

$$P\ T > t\ \ ^{P} .$$

is Schur-convex function of $\underset{\sim}{p}$ for any positive number t. Therefore, the polarization test is unbiased. For the application of the theorem note that T is a Schur-convex function of $\underset{\sim}{x}$ and that the multinomial probability function satisfies the condition required for $\phi(\underset{\sim}{\lambda},\underset{\sim}{x})$ in the theorem.

For the application of the test we need to find the distribution of T. First we consider the exact distribution of T. Let $[x]^+$ denote the smallest non-negative integer $\geq x$, and $[x]^-$ denote the smallest integer $\leq x$. Let $D_k(t;p_1,\dots,p_k,n)$ $= P\{T \leq t\}$ denote the cumulative distribution function (cdf) of T for $k \geq 2$. The cdf is recursively given by

$$(2.1)\qquad D_2\,(t;p_1,1-p_1,n)=\begin{cases} 0, & t<n/2 \\ \displaystyle\sum_{r=[\frac{n}{2}-(\frac{nt}{2}-\frac{n^2}{4})^{\frac{1}{2}}]^+}^{[\frac{n}{2}+(\frac{nt}{2}-\frac{n^2}{4})^{\frac{1}{2}}]^-} \binom{n}{r}\, p_1^r(1-p_1)^{n-r} & \frac{n}{2}\leq t\leq n \\ 1, & t>n \end{cases}$$

$$(2.2)\qquad D_k(t;p_1,\dots,p_k,n)=\begin{cases} 0, & t<n/k \\ \displaystyle\sum_{r=[\frac{n}{k}-(n(t-\frac{n}{k})(1-\frac{1}{k}))^{\frac{1}{2}}]^+}^{[\frac{n}{k}+(n(t-\frac{n}{k})(1-\frac{1}{k}))^{\frac{1}{2}}]^-} \binom{n}{r}p_k^r(1-p_k)^{n-r} \\ \qquad D_k\left(\frac{nt-r^2}{n-r};\frac{p_1}{1-p_k},\dots,\frac{p_{k-1}}{1-p_k},\ n-r\right), & \frac{n}{k}\leq t\leq n \\ 1, & t>n \end{cases}$$

The recursive relation is based on the fact that the conditional distribution of $\underset{\sim}{u} = (x_1,\dots,x_{k-1})$, given x_k, is $M(\underset{\sim}{u};\frac{p_1}{1-p_k},\dots,\frac{p_{k-1}}{1-p_k},\ n-x_k)$.

Next we consider the asymptotic distribution of T for large n. The covariance matrix of $\underset{\sim}{x}$ is $n\Sigma = n(D - \underset{\sim}{p}\,\underset{\sim}{p}')$, where prime denotes the transpose and D denotes the diagonal matrix with ith diagonal element equal to p_1. Let $\lambda_1, \ldots, \lambda_{k-1}$ denote the non-zero eigen value of Σ. An eigen vector corresponding to the zero eigen value is

$\xi = (1, 1, \ldots, 1)'$. Let P denote an orthogonal matrix diagonalizing Σ whose first row is equal to $(\frac{1}{\sqrt{k}}, \ldots, \frac{1}{\sqrt{k}})$ and let $y = P\,x$. The first component of y is equal to $\frac{n}{\sqrt{k}}$. Let the mean of y be written as

$$(2.3) \qquad E\,y = (\tfrac{n}{\sqrt{k}}, n\delta_1\sqrt{\lambda_1}, \ldots, n\delta_{k-1}\sqrt{\lambda_{k-1}})'$$

$$= n\,P\,p.$$

From (2.3) we have

$$(2.4) \qquad \frac{1}{k} + \sum_{i=1}^{k-1} \lambda_i\,\delta_i^2 = p'P'P\,p = p'p = Q$$

$$(2.5) \qquad \sum_{i=1}^{k-1} \lambda_i^2\,\delta_i^2 = p'\,P'(P\Sigma P')P\,p$$

$$= p'\,(D - p\,p')p$$

$$= Q_1 - Q^2$$

where $Q_1 = \sum_{i=1}^{k} p_i^3$. It is easy to see that $Q_1 \geq Q^2$.

We have $n\,T = x'\,x = y'\,y'$. Since x is asymptotically distributed according to the multivariate normal distribution with mean np and covariance $n\Sigma$, we have for large n

$$(2.6) \qquad T \overset{d}{\sim} \frac{n}{k} + \sum_{i=1}^{k} \lambda_i\,Z_i^2$$

where $\overset{d}{\sim}$ means "asymptotically distributed as" and Z_i is normally distributed with variance 1 and mean equal to $\sqrt{n}\,\delta_i$. Moreover $Z_1, \ldots, Z_{k-1}$ are independent.

Let $\theta_i = \delta_i^2$. Since Z_i^2 is distributed as $\chi^2_{1,n\theta_i}$ (non-central chi-square with 1 degree of freedom and non-centrality parameter equal to $n\theta_i$). The moment generating function of $R = \sum_{i=1}^{k-1} \lambda_i Z_i^2$ is given by

(2.7) $$M(t) = E\, e^{tR}$$

$$= \text{enp}\left(n \sum_{i=1}^{k-1} \frac{\lambda_i \delta_i^2 t}{1-2\lambda_i t}\right) \prod_{i=1}^{k-1} (1-2\lambda_i t)^{-\frac{1}{2}} \quad t < 0.$$

From (2.7) it is seen that

(2.8) $$\sqrt{n}\left(\frac{R}{n} - \sum_{i=1}^{k-1} \lambda_i \delta_i^2\right)\left(4 \sum_{i=1}^{k-1} \lambda_i^2 \delta_i^2\right)^{-\frac{1}{2}} \overset{d}{\to} V.$$

where V denotes a standard normal random variable. From (2.4), (2.5), (2.6) and (2.8) we have that

(2.9) $$\frac{T - nQ}{2\sqrt{n}} (Q_1 - Q^2)^{-\frac{1}{2}} \overset{d}{\to} V.$$

Suppose that $Q_1 - Q^2 = 0$. Two cases arise: (i) $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$ and (ii) $p_i p_j = 0$ for $i \neq j$. In case (i) we have $\lambda_1 = \lambda_2 = \ldots = \lambda_{k-1} = \frac{1}{k}$ and from (2.3) $\delta_1 = \delta_2 = \ldots = \delta_{k-1} = 0$. Therefore, Z_i^2 is distributed as χ_1^2 (central chi-square with 1 degree of freedom) and (2.6) we have

(2.10) $$k(T - nQ) \overset{d}{\to} \chi_{k-1}^2.$$

In case (ii) we have $T = n$ with probability 1.

Let Q^o and Q_1^o denote the values of Q and Q_1, respectively, for $p = p^o$, and let V_α denote the upper α - quantile of the standard normal distribution. Let T_α denote the critical value of the polarization test for a level of significance equal to α, as derived from (2.1) and (2.2). From (2.9) the value of T_α for large n is approximately given by

(2.11) $$T_\alpha = n\, Q^o + 2\sqrt{n}\, (Q_1^o - (Q^o)^2)^{\frac{1}{2}}\, V_\alpha.$$

Also, the asymptotic power of the polarization test is equal to

$$\Phi\left(\frac{\sqrt{n}(Q - Q^o)}{2(Q_1 - Q^2)} - \left(\frac{Q_1^o - (Q^o)^2}{Q_1 - Q^2}\right)^{\frac{1}{2}} V_\alpha\right)$$

where Φ denotes the standard normal cdf. For $Q - Q^o = \frac{c}{\sqrt{n}}$, where c is given positive number, the asymptotic power, that is, the Pitman efficiency is equal to $\Phi(\frac{c}{2}(Q_1^o - (Q^o)^2)^{-\frac{1}{2}} - V_\alpha)$.

In order to compare the asymptotic formula with the exact formula we show in the table below values of $P\{T \leq T_\alpha\}$, derived from (2.1) and (2.2), where T_α is given by (2.11) for $\alpha = .95$, k = 2, 3, 4 and certain values of n and $\underset{\sim}{p}^o$. It is seen from the table that n = 100 is not sufficiently large for the asymptotic probability to match with the exact probability. It is interesting to observe that the figures in columns 2, 3 and 5 agree except for one entry. We have checked the figures given in the table with the result obtained from a simulation study.

Values of $P\{T \leq T_{.95}\}$

k	2	3		4	
$\underset{\sim}{p}^o$	(.10,.90)	(.05,.05,.90)	(.3,.3,.4)	(.03,.03,.04,.90)	(.2,.2,.2,.4)
n=10	1.000	1.000	.596	1.000	.773
20	.873	1.000	.643	1.000	.812
40	.920	.920	.757	.920	.844
60	.947	.947	.802	.947	.861
80	.965	.965	.802	.965	.877
100	.942	.942	.822	.942	.886

Reference

[1] Hollander, M., Proschan, F. and Sethuraman, J. (1977) Functions decreasing in transposition and their applications in ranking problems. Ann. Statist. (5) 722-733.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N-105 2. GOVT ACCESSION NO. AD-A083525	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Polarization test for the multinomial distribution	5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER 299
7. AUTHOR(s) Khursheed Alam Amitava Mitra	8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 047-002 NR ~~042-271~~
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code ~~436~~ 434 Arlington, Va. 22217	12. REPORT DATE 12-2-1979 13. NUMBER OF PAGES 7
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multinomial Distribution, Schur-Convex Function.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper gives a test of the hypothesis that the total probability associated with a multinomial distribution with k cells is evenly distributed among the k cells against the alternative hypothesis that it is less evenly distributed.