TECHNICAL REPORT
4-73
業 績 報 告 書

A REDUCTION METHOD OF BINOMIAL OR MULTINOMIAL DATA WITH MULTIPLE CLASSIFICATIONS

多次元分割表に伴う二項型または多項型データの一削減法

MASANORI OTAKE, B.A. 大竹正徳

ABCC JNIH

ATOMIC BOMB CASUALTY COMMISSION

国立予防衛生研究所－原爆傷害調査委員会

JAPANESE NATIONAL INSTITUTE OF HEALTH OF THE MINISTRY OF HEALTH AND WELFARE

TECHNICAL REPORT SERIES
業 績 報 告 書 集

The ABCC Technical Reports provide the official bilingual statements required to meet the needs of Japanese and American staff members, consultants, advisory groups, and affiliated government and private organizations. The Technical Report Series is in no way intended to supplant regular journal publication.

ABCC業績報告書は，ABCCの日米専門職員，顧問，諮問機関ならびに政府および民間の関係諸団体の要求に応ずるための日英両語による公式報告記録であって，業績報告書集は決して通例の誌上発表論文に代わるものではない.

TECHNICAL REPORT
4-73
業 績 報 告 書

A REDUCTION METHOD OF BINOMIAL OR MULTINOMIAL DATA WITH MULTIPLE CLASSIFICATIONS

多次元分割表に伴う二項型または多項型データの一削減法

MASANORI OTAKE, B.A.　大竹正徳

ABCC JNIH

ATOMIC BOMB CASUALTY COMMISSION
HIROSHIMA AND NAGASAKI, JAPAN

A Cooperative Research Agency of
U.S.A. NATIONAL ACADEMY OF SCIENCES · NATIONAL RESEARCH COUNCIL
and
JAPANESE NATIONAL INSTITUTE OF HEALTH OF THE MINISTRY OF HEALTH AND WELFARE

with funds provided by
U.S.A. ATOMIC ENERGY COMMISSION
JAPANESE NATIONAL INSTITUTE OF HEALTH
U.S.A. PUBLIC HEALTH SERVICE

原 爆 傷 害 調 査 委 員 会
広島および長崎

米国学士院一学術会議と厚生省国立予防衛生研究所
との日米共同調査研究機関

米国原子力委員会，厚生省国立予防衛生研究所および米国公衆衛生局の研究費による

ACKNOWLEDGMENT

謝 辞

The author wishes to express his appreciation to Mr. Seymour Jablon, former chief of the ABCC Department of Statistics, Professor Sumiyasu Yamamoto and Assistant Professor Nariaki Sugiura, Faculty of Sciences, Hiroshima University for their valuable suggestions.

貴重な提案をいただいた Seymour Jablon 元 ABCC 統計部長ならびに広島大学理学部山本純恭教授および杉浦成昭助教授に謝意を表する.

CONTENTS

目 次

Approved 承認 *22 February 1973*

A REDUCTION METHOD OF BINOMIAL OR MULTINOMIAL DATA WITH MULTIPLE CLASSIFICATIONS

多次元分割表に伴う二項型または多項型データの一削減法

MASANORI OTAKE, B.A. (大竹正徳)

Department of Statistics
統計部

SUMMARY

Medical data such as the presence or absence of a specific disease or grades of clinically abnormal findings are frequently classified into groups of one or more factors. An important step in the classification of data is the determination of how the events or characteristics should be grouped. This study proposes a data reduction method which identifies the point of maximum χ^2 value where the total range of data may be divided into two parts. The procedure involves the calculation of χ^2 values for possible 2×2 or $s \times 2$ tables that can be formed by a step-by-step combination of marginal totals. If the maximum χ^2 value is less than a predetermined value, it is presumed that these groups are homogeneous and may therefore be combined. If the maximum χ^2 value is greater than or equal to a specific value, this is taken as an indication that each group is meaningful, statistically speaking.

Leukemia data preclassified by city, sex, age and dose are used to illustrate the method. In this example, the data suggest that, in males, the risk of leukemia is the highest among those exposed to radiation doses of 200 rad and over who were aged under 15 at the time of the atomic bomb.

要　約

特定疾患の有無あるいは臨床的異常所見の評価などに関する医学データは1要因またはそれ以上の要因によって種々の群にしばしば分割される．データ分類のうえで重要な段階は，事象または特性をいかに群別するかを決定することである．本研究はデータの全領域が2分割されたときのχ^2統計量の最大値点によってデータを削減する方法を提案する．この方法は各事象の周辺合計を漸次結合することによって作成される可能な2×2表または$s \times 2$表に対するχ^2統計量を計算することである．もしχ^2統計量の最大値が前もって定めたある判定値より小さければこれらの群は同じ程度の変動であるとみなし，一つの群に統合できると考える．もしχ^2統計量の最大値がある判定値より大きいか等しい場合，統計的にいって，各群は意味のある徴候を示唆すると考える．

データの削減法を示す1例として，事前に都市別，性別，年齢群別および被曝線量群別に分類された白血病データが利用された．この例題については，白血病発生の危険率は，男で，しかも原爆時15歳未満で200 rad以上の被曝線量を受けた被爆者間に最も高いということを示唆する．

I. INTRODUCTION

Data in a medical survey often involve the presence or absence of an attribute such as a specific disease, or several grades of the morphological abnormalities which can be ordered in terms of severity. The former constitutes a binomial sample and the latter a multinomial sample. The characteristics of such a binomial or multinomial sample cannot be explained simply in terms of any single factor but may instead be related simultaneously to many other factors. There are also interaction effects among those factors. Even when the best measure is examined, such a factor may actually have nonlinear effects on the characteristics of such a binomial or multinomial sample.

The purpose of this paper is to reasonably evaluate the essential features of a binomial or multinomial sample which has been preclassified in some order by one or more factors. The multiple classifications of such a binomial or multinomial sample give a large number of groups composed of x_1 numbers with specific disease and x_2 without disease, or composed of x_1 (marked), x_2 (moderate), x_3 (light), x_4 (normal), etc. In many instances, multiple classifications will result in zero or low frequencies in some cells among many preclassified groups. The reduction of such data has been attempted by regrouping of homogeneous groups through the process of a series of binary splits.

Belson[1] has suggested the empirical matching of samples which can be classified into two different groups by a sign of difference between expected and observed cases in low score groups. Such a division always allows a 50 to 50 split of expected cases. Morgan and Sonquist[2] and Sonquist and Morgan[3] have proposed an interesting approach such that the partitioning of the group into two groups on an arbitrary decision rule provides the largest reduction in the unexplained sum of squares within the group, using mean squares criteria.

The author attempted a different approach to data reduction which identifies the point of maximum χ^2 value where the marginal binomial or multinomial data for multiple classifications may be divided into two parts, i.e., binary splits. The definition and criterion of binary splits are given in Section II. Two sets of leukemia data with one and four factors are used as numerical examples 1 and 2 in Section III, and a discussion is presented in Section IV.

I．序

医学調査におけるデータは特定疾患のような属性の有無あるいはその程度によって順位づけられる種々の形態学的異常度をしばしば包含する．前者は二項型標本を構成し，後者は多項型標本を形成する．かかる二項型標本または多項型標本の特性はある単一の因子によって説明されるものではなく，他の多くの要因と密接な関係をもっている．またこれら要因間の交互作用も存在する．たとえ最良の測定値が得られる場合にあっても，このような要因は二項型標本または多項型標本の特性に対して，実際的には非線型的効果をもつこともある．

本論文の目的は1要因またはそれ以上の要因によって順位づけて事前に分類した二項型標本または多項型標本に対する本質的特徴を合理的に評価することにある．かかる二項型標本または多項型標本の多次元分割表は特殊疾患者数である x_1 および非特殊疾患者数である x_2 または形態学的分類による対象者数である x_1（高程度），x_2（中程度），x_3（小程度），x_4（正常）などから成る多くの群を構成する．多くの場合，多次元分割表では事前に分類した多くの群内のある区分の幾つかにおいて，ゼロまたは低頻度数を結果づけるであろう．一連の2分割の手順を通じて同じ程度の変動をもつ群結合によるデータの削減が試みられた．

Belson[1] は低頻度群の期待値と観測値との間の差の符号によって二つの異なる群に分類される経験的標本結合を示唆した．このような分割は常に期待数の50対50分割を許容する．Morgan と Sonquist[2] および Sonquist と Morgan[3] は，平均平方和による判定法を用いて，任意の判定規準に基づき群の分割を2群に分割することが，群内の説明のつかない平方和における最も大きな削減を提供するという興味ある方法を提案した．

著者は二項型標本または多項型標本の多次元分割表の周辺合計データから2区分，すなわち，2分割したときの χ^2 統計量の最大値点によってデータを削減する異なる方法を示した．2分割の定義および判定基準は第II章で述べる．第III章では数値例1および数値例2として1要因および4要因を伴う2例の白血病データを用いる．考察は第IV章で述べる．

II. DEFINITION AND CRITERION OF BINARY SPLITS

1. Binomial or Multinomial Sample Preclassified by One Factor

Consider a sample drawn at random from a binomial or multinomial population, and assume that such a sample has been preclassified into p groups in some order by one factor.

The binomial sample preclassified in terms of one factor gives a 2 × p table as

Ⅱ．２分割の定義および判定基準

１．１要因の項によって事前に分類された二項型標本または多項型標本

二項型母集団または多項型母集団から無作為抽出された標本を考える．そして，この標本は１要因によって順位づけられたｐ群に事前に分類されているとする．

１要因の項によって事前に分類される二項型標本は

$$\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}, \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}, \ldots, \begin{pmatrix} x_{p1} \\ x_{p2} \end{pmatrix}$$

where x_{i1} $(i = 1, 2, \ldots, p)$ is the number of cases with the presence of an attribute in the i-th group and x_{i2} is the number of cases without that attribute.

The multinomial sample preclassified in terms of one factor gives an s × p table as

としての２×ｐ分割表を提供する．ただし，x_{i1}（$i = 1, 2, \cdots, p$）はｉ番目の群におけるある属性をもつ対象者数を示し，一方 x_{i2} はその属性をもたない対象者数を示す．

１要因の項で事前に分類される多項型標本は

$$\begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1s} \end{pmatrix}, \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2s} \end{pmatrix}, \ldots, \begin{pmatrix} x_{p1} \\ x_{p2} \\ \vdots \\ x_{ps} \end{pmatrix}$$

where x_{i1} is the number of cases which can be specified in terms of grade 1, x_{i2} is that of grade 2, ..., and x_{is} is that of grade s.

Let the data in each group of a binomial or multinomial sample be expressed as $X_1, X_2, \ldots, X_p$ respectively. The binary splits (B.S.) of $X_i' \equiv (x_{i1}, x_{i2})$ for $i = 1, 2, \ldots, p$ of binomial sample or $X_i' \equiv (x_{i1}, x_{i2}, \ldots, x_{is})$ of a multinomial sample are defined by

としてのｓ×ｐ分割表を提供する．ただし x_{i1} は評価１の項によって定められた例数であり，x_{i2} は評価２の例数，…，および x_{is} は評価ｓの例数をさす．

二項型標本または多項型標本の各群におけるデータをそれぞれ $X_1, X_2, \cdots, X_p$ として表わす．二項型標本 $X_i' \equiv (x_{i1}, x_{i2})$ ただし $i = 1, 2, \cdots, p$ または多項型標本 $X_i' \equiv (x_{i1}, x_{i2}, \cdots, x_{is})$ の２分割は

$$\begin{array}{l} (X_1, X_2 + \ldots + X_p) \\ (X_1 + X_2, X_3 + \ldots + X_p) \\ \quad\vdots \\ (X_1 + \ldots + X_{p-1}, X_p) \qquad (1) \end{array}$$

Hence, the number of B.S. at the beginning is equal to (p−1). Each B.S. of (1) is composed of a 2 × 2 table for a binomial sample or an s × 2 contingency table for a multinomial sample. Criterion for $\omega = (X_1, ..., X_p)$ is executed:

によって定義される．このように，最初の段階での２分割数は（p−１）個に等しい．(1)の各２分割は二項型標本に対しては２×２表を構成し，また多項型標本に対してはs×２表を構成する．$\omega = (X_1, \cdots, X_p)$に対する判定基準は次のように作成される．

(i) If p = 1, the process of B.S. will be terminated.

（i） もしp＝１であれば，２分割の手順は終了する．

(ii) If $p \geqslant 2$, each of $\chi^2_{(1)}, \chi^2_{(2)}, ..., \chi^2_{(p-1)}$ is calculated for the corresponding B.S. defined in (1) and

（ii） もしp≧２であれば，$\chi^2_{(1)}, \chi^2_{(2)}, \cdots, \chi^2_{(p-1)}$個の$\chi^2$統計量が(1)で定義した対応する２分割によって計算される．そして

$$\chi^2_{(\gamma)} = \max [\, \chi^2_{(1)}, \chi^2_{(2)}, ..., \chi^2_{(p-1)} \,]$$

is determined where γ is a cutting point which gives the maximum of possible χ^2 values obtained through the process of a series of B.S. defined by (1).

が決定される．ただし，γは(1)で定義した一連の２分割の手順を通じて計算された可能なχ^2統計量の最大値を与える分割点である．

(iii) If $\chi^2_{(\gamma)} < c$, the process of B.S. is terminated. We conclude that p groups preclassified in some order for one factor are homogeneous. Therefore, all groups may be combined as $X_1 + ... + X_p$. A constant c is available for the criterion. See IV Discussion about a constant c.

（iii） もし$\chi^2_{(\gamma)} < c$であれば，２分割の手順は終了する．１要因を順位づけて事前に分類したp群は同じ程度の変動であると結論づける．したがって，すべての群は$X_1 + \cdots + X_p$として統合される．定数cは判定基準として用いられる．定数cについては第Ⅳ章の考察を参照されたい．

(iv) If $\chi^2_{(\gamma)} \geqslant c$, $\omega = (X_1, ..., X_p)$ is divided into two new groups,

（iv） もし$\chi^2_{(\gamma)} \geqq c$であれば，$\omega = (X_1, \cdots, X_p)$は新しい二つの群，すなわち，

$$\omega \begin{cases} \omega_1 = (X_1, ..., X_\gamma) \\ \omega_2 = (X_{\gamma+1}, ..., X_p). \end{cases}$$

The same criterion as described in (i) to (iv) will be applied to ω_1 and ω_2, respectively, until the process of B.S. is terminated.

に分割される．上記の（i）から（iv）までの２分割の同様の判定基準がそれぞれω_1およびω_2に適用される．

2. Binomial or Multinomial Sample Preclassified by Two Factors

２．２要因の項で事前に分類された二項型標本および多項型標本

We can consider a sample drawn at random from a binomial or multinomial population, and assume that such a sample has been preclassified in some order by two factors.

二項型母集団および多項型母集団から無作為抽出された標本を考える．そして，この標本は２要因によって順位づけられて事前に分類されていると仮定する．

The binomial sample preclassified in terms of two factors has the following tabular form:

２要因の項によって事前に分類された二項型標本は次の分割形式

$$\begin{pmatrix} x_{111} \\ x_{112} \end{pmatrix}, \begin{pmatrix} x_{121} \\ x_{122} \end{pmatrix}, \ldots, \begin{pmatrix} x_{1q1} \\ x_{1q2} \end{pmatrix}$$

$$\begin{pmatrix} x_{211} \\ x_{212} \end{pmatrix}, \begin{pmatrix} x_{221} \\ x_{222} \end{pmatrix}, \ldots, \begin{pmatrix} x_{2q1} \\ x_{2q2} \end{pmatrix}$$

$$\vdots \qquad \vdots$$

$$\begin{pmatrix} x_{p11} \\ x_{p12} \end{pmatrix}, \begin{pmatrix} x_{p21} \\ x_{p22} \end{pmatrix}, \ldots, \begin{pmatrix} x_{pq1} \\ x_{pq2} \end{pmatrix}$$

where each component of x_{ij1} ($i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, q$) in the (i, j)th group is the number of cases which can be specified in terms of grade 1, x_{ij2} is that in terms of grade 2. We can represent each group in (2) as $X'_{ij} \equiv (x_{ij1}, x_{ij2})$ for the binomial sample. When we have the multinomial sample preclassified in some order for two factors, each component of the group is expressed by $X'_{ij} \equiv (x_{ij1}, x_{ij2}, \cdots x_{ijs})$. The same procedure (B.S.) as for one factor is performed for two marginal totals which are represented by $X_{i\cdot} = (X_{1\cdot}, X_{2\cdot}, \ldots, X_{p\cdot})$ and $X_{\cdot j} = (X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot q})$. Each marginal total is composed of a $2 \times p$ or an $s \times p$ contingency table for $X_{i\cdot}$ and a $2 \times q$ or an $s \times q$ contingency table for $X_{\cdot j}$.

を与える．ただし，（i，j）番目の群における x_{ij1}（$i=1, 2, \cdots, p$；$j=1, 2, \cdots, q$）の各成分は評価1の項の例数を示し，x_{ij2} は評価2の例数を示す．二項型標本に対する(2)の各群は $X'_{ij} \equiv (x_{ij1}, x_{ij2})$ として表わされる．2要因により順位づけられて事前に分類された多項型標本の場合，群の各成分は $X'_{ij} \equiv (x_{ij1}, x_{ij2}, \cdots, x_{ijs})$ として示される．1要因に対すると同様の2分割の手順が $X_{i\cdot} = (X_{1\cdot}, X_{2\cdot}, \cdots, X_{p\cdot})$ および $X_{\cdot j} = (X_{\cdot 1}, X_{\cdot 2}, \cdots, X_{\cdot q})$ によって代表される二つの周辺合計に適用される．各周辺合計は $X_{i\cdot}$ に対して $2 \times p$ または $s \times p$ 分割表から成り，また，$X_{\cdot j}$ に対しては $2 \times q$ または $s \times q$ 分割表から成っている．

The definition of the B.S. for a $2 \times p$ or an $s \times p$ table and a $2 \times q$ or an $s \times q$ table is given as

$2 \times p$ または $s \times p$ 分割表および $2 \times q$ または $s \times q$ 分割表における2分割の定義は

$$\begin{gathered}
(X_{1\cdot}, X_{2\cdot} + \ldots + X_{p\cdot}) \\
(X_{1\cdot} + X_{2\cdot}, X_{3\cdot} + \ldots + X_{p\cdot}) \\
\vdots \\
(X_{1\cdot} + \ldots + X_{p-1\cdot}, X_{p\cdot}) \\
\text{and} \quad \text{および} \qquad\qquad (3) \\
(X_{\cdot 1}, X_{\cdot 2} + \ldots + X_{\cdot q}) \\
(X_{\cdot 1} + X_{\cdot 2}, X_{\cdot 3} + \ldots + X_{\cdot q}) \\
\vdots \\
(X_{\cdot 1} + \ldots + X_{\cdot q-1}, X_{\cdot q}).
\end{gathered}$$

The number of possible B.S. is equal to $(p-1) + (q-1)$. Criterion for $\omega = (X_{i1}, \ldots, X_{iq}; i=1, \ldots, p)$ is executed:

として表わされる．可能な2分割数は（$p-1$）+（$q-1$）に等しい．$\omega = (X_{i1}, \cdots, X_{iq}$；$i=1, \cdots, p$）に対して次の判定基準が作成される．

(i) If p=1 and q=1, all the processes of B.S. will be terminated.

（ｉ） もしｐ＝１かつｑ＝１であれば，２分割のすべての手順は終了する．

(ii) If $p \geqslant 2$ and q=1 or p=1 and $q \geqslant 2$, the process of B.S. will be reduced to the B.S. of one factor.

（ii） もしｐ≧２かつｑ＝１またはｐ＝１かつｑ≧２であれば，２分割の手順は１要因の２分割の手順と同じになる．

(iii) If $p \geqslant 2$ and $q \geqslant 2$, the B.S. for the p $\times$ q groups are executed for two marginal totals for $X_{i\cdot} = (X_{1\cdot}, \ldots, X_{p\cdot})$ and $X_{\cdot j} = (X_{\cdot 1}, \ldots, X_{\cdot q})$. Each of $\chi^2_{(1\cdot)}, \ldots, \chi^2_{(p-1\cdot)}$ and $\chi^2_{(\cdot 1)}, \ldots, \chi^2_{(\cdot q-1)}$ is calculated for each B.S. defined in (3), and

（iii） もしｐ≧２かつｑ≧２であれば，ｐ×ｑ群の２分割は$X_{i\cdot} = (X_{1\cdot}, \cdots, X_{p\cdot})$および$X_{\cdot j} = (X_{\cdot 1}, \cdots, X_{\cdot q})$に対する二つの周辺合計を用いて実行される．$\chi^2_{(1\cdot)}, \cdots, \chi^2_{(p-1\cdot)}$および$\chi^2_{(\cdot 1)}, \cdots, \chi^2_{(\cdot q-1)}$の各$\chi^2$統計量は(3)において定義された２分割表から計算される．そして，

$$\chi^2_{(\gamma^d)} = \max \left[\max_{1 \leqslant i \leqslant p-1} \chi^2_{(i\cdot)}, \max_{1 \leqslant j \leqslant q-1} \chi^2_{(\cdot j)} \right]$$

is determined, where d = 1 for p groups of 2 for q groups indicates the factor giving the maximum of $(p+q-2)\ \chi^2$ values.

が決定される．ただし，ｐ群の要因に対してはｄ＝１またはｑ群の要因に対してはｄ＝２によって（ｐ＋ｑ－２）個のχ^2統計量の最大値を与える要因を表わす．

(iv) If $\chi^2_{(\gamma^d)} < c$, the process of B.S. is terminated. We shall conclude that the p $\times$ q groups preclassified in some order by two factors are homogeneous. Therefore, all groups may be combined.

（iv） もし$\chi^2_{(\gamma^d)} < c$であれば，２分割の手順は終了する．２要因によって順位づけて事前に分類したｐ×ｑ群は同じ程度の変動であるとみなし，１つの群として扱ってよいと結論づける．

A constant c is available for the criterion. See IV Discussion about a constant c.

定数ｃが判定基準として用いられる．定数ｃについては第Ⅳ章の考察を参照されたい．

(v) If $\chi^2_{(\gamma^d)} \geqslant c$, ω is divided into two new groups,

（ｖ） もし$\chi^2_{(\gamma^d)} \geqq c$であれば，ωが二つの新しいグループに分割される．

$$\omega \left\langle \begin{array}{l} \omega_1^d \\ \\ \omega_2^d \end{array} \right.$$

If d = 1,
もし

$$\omega \left\langle \begin{array}{l} \omega_1^1 = (X_{1j}, \ldots, X_{\gamma^1 j}) \\ \omega_2^1 = (X_{\gamma^1+1j}, \ldots, X_{pj}) \end{array} \right.$$

where j = 1, ..., q.
ただし

If d = 2,
もし

$$\omega \left\langle \begin{array}{l} \omega_1^2 = (X_{i1}, \ldots, X_{i\gamma^2}) \\ \omega_2^2 = (X_{i\gamma^2+1}, \ldots, X_{iq}) \end{array} \right.$$

where i = 1, ..., p.
ただし

The same criterion as described in (i) to (v) will be applied to ω_1^1 and ω_2^1 for d = 1 or to ω_1^2 and ω_2^2 for d = 2 until the process of B.S. is terminated.

上記の（ｉ）から（ｖ）までの同様の判定基準が２分割の条件を満足しなくなるまでｄ＝１に対してはω_1^1およびω_2^1に，また，ｄ＝２に対してはω_1^2およびω_2^2に適用される．

3. Binomial or Multinomial Sample Preclassified by m Factors

In general, we can consider a sample drawn at random from a binomial or multinomial population, and assume that such a sample has been preclassified in some order by m factors.

The definition and criterion of B.S. are considered as an extension of two factors. The process of a series of B.S. is performed by the same procedure as that for two factors.

3．m要因によって事前に分類された二項型標本または多項型標本

一般に，二項型母集団または多項型母集団から無作為抽出された標本を考える．そして，標本はm要因によって順位づけて事前に分類されていると仮定する．

2分割の定義および判定基準については2要因の一般化として考えられる．一連の2分割の手順は2要因に対する手順と同様の方法で行なわれる．

III. NUMERICAL APPROACH

Ⅲ．数値解法

Example 1: One factor

A portion of the leukemia data prepared by Ishimaru, et al[4] was used. Table 1 shows the number of leukemia cases observed for the period 1 October 1950-30 September 1966 among male atomic bomb survivors in Hiroshima aged 15-39 at the time of the A-bomb (ATB) belonging to the Life Span Study Extended sample (LSSE)[5]

例題1：1要因の場合

石丸ら[4]が行なった白血病データの一部が利用された．表1は広島の寿命調査拡大集団[5]に属する男の原爆時年齢15－39歳の被爆生存者について，1950年10月1日から1966年9月30日までの期間中に観察された白血病症例を示す．

TABLE 1 NUMBER OF LEUKEMIA CASES BY RADIATION DOSE FOR HIROSHIMA MALES AGED 15-39 ATB IN LSSE, 1950-1966

表1　広島の寿命調査拡大集団のうち，原爆時年齢15－39歳の男の被曝線量群別白血病症例，1950－1966年

	Radiation Dose (rad)							
	<5 X_1	5-19 X_2	20-49 X_3	50-99 X_4	100-199 X_5	200-299 X_6	300+ X_7	Total $X.$
Leukemia (x_{i1})	2	0	3	2	2	2	5	16
Not leukemia (x_{i2})	4601	1161	477	271	243	98	149	7000
Total ($x_{i\cdot}$)	4603	1161	480	273	245	100	154	7016

where i = 1, 2, ..., 7.

From the definition of B.S. in (1), $\chi^2_{(1)}, \chi^2_{(2)}, \ldots, \chi^2_{(6)}$ are calculated. The $\chi^2_{(1)}$ value is obtained by the $(X_1, X_2 + \ldots + X_7)$ table, i.e.,

(1)の2分割の定義から，$\chi^2_{(1)}$，$\chi^2_{(2)}$，…，$\chi^2_{(6)}$が計算される．$\chi^2_{(1)}$の統計量は$(X_1, X_2+\cdots+X_7)$表，すなわち，

	X_1	$X_2 + \ldots + X_7$	Total
Leukemia	2	14	16
Not leukemia	4601	2399	7000
Total	4603	2413	7016

In order to evaluate the maximum of χ^2 values obtained through a series of B.S., we used a simple procedure involving less calculation:

から求められる．一連の２分割の手順を通じて得られた種々のχ^2統計量の最大値を評価するために，計算上の簡略法，すなわち，

$$\chi^2_{(1)} = C_1 * (A_{(1)} - C_2)$$

$$\chi^2_{(2)} = C_1 * (A_{(2)} - C_2)$$

$$\vdots$$

$$\chi^2_{(6)} = C_1 * (A_{(6)} - C_2)$$

Hence, the maximum of these χ^2 values is obtained by

を用いた．したがって，これらのχ^2統計量の最大値は次の式で表わされる．

$$\chi^2_{(\gamma)} = C_1 * [\max_{1 \leqslant \gamma \leqslant 6} A_{(\gamma)} - C_2]$$

$$\text{where } C_1 = (x._1 + x._2)^2 / (x._1 * x._2) = 7016^2 / (16*7000) = 439.50$$

$$C_2 = x^2._1 / (x._1 + x._2) = 16^2/7016 = 0.036488, \text{ and}$$

$$A_{(\gamma)} = (\sum_{i=1}^{\gamma} x_{i1})^2 / \sum_{i=1}^{\gamma} x_i. + (\sum_{i=\gamma+1}^{7} x_{i1})^2 / \sum_{i=\gamma+1}^{7} x_i., \text{ i.e.,}$$

$$\max [A_{(1)} = 0.082096, A_{(2)} = 0.15724, A_{(3)} = 0.16074, A_{(4)} = 0.16984,$$

$$A_{(5)} = 0.20489, A_{(6)} = 0.17997] = A_{(5)} = 0.20489$$

Thus, the maximum χ^2 value is given by

かくして，χ^2の最大値は

$$\chi^2_{(\gamma=5)} = 439.50 * [0.20489 - 0.036488] = 74.013.$$

A constant c = 3.841 was evaluated as the critical value at the 5% point of a χ^2 distribution with 1 degree of freedom.

として得られる．判定値は自由度１のχ^2分布の５％点の定数 c ＝3.841として評価された．

As $\chi^2_{(\gamma=5)} \geqslant 3.841$, $\omega = (X_1, \ldots, X_7)$ is divided into two new groups,

$\chi^2_{(\gamma=5)} \geqq 3.841$の不等式が成立するので$\omega = (X_1, \ldots, X_7)$は新しい二つの群，

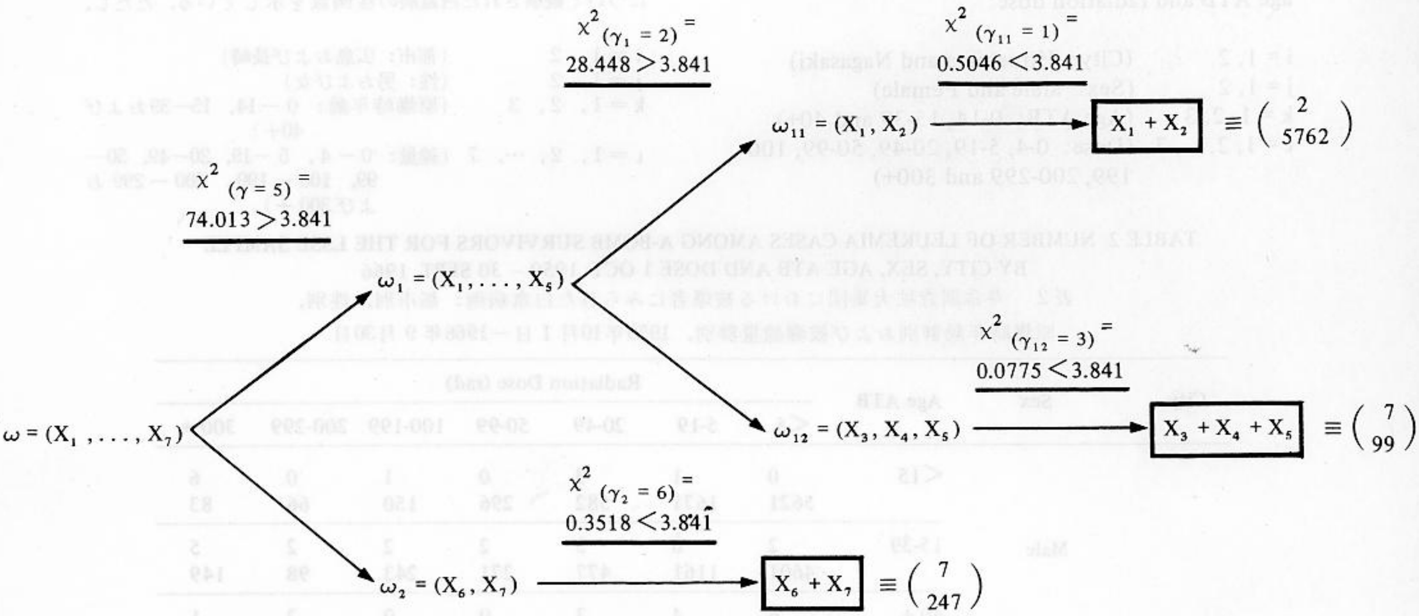
$$\omega \begin{cases} \omega_1 = (X_1, \ldots, X_5) \\ \omega_2 = (X_6, X_7) \end{cases}$$

and the process of possible B.S. is continued for ω_1 and ω_2 as shown in Figure 1.

に分割される．そして，可能な２分割の手順が図１に示しているようにω1およびω2に対して続行される．

FIGURE 1 REDUCTION OF BINOMIAL DATA PRECLASSIFIED BY RADIATION DOSE

図１ 被曝線量によって事前に分類された二項型データの削減



Note: ☐ shows that the process of B.S. was terminated.

The final results through the process of possible B.S. are given by

可能な２分割の手順によって得られた最終結果は次のようである．

	$(X_1 + X_2)$	$(X_3 + X_4 + X_5)$	$(X_6 + X_7)$	Total
Leukemia	2	7	7	16
Not leukemia	5762	991	247	7000
Total	5764	998	254	7016

Through the use of these three meaningful new groups, we can easily determine the relative region of high or low incidence. There may be almost no problem for one factor, but it will be more complex to handle or to detect the degree of association existing among two or more classifications.

意味ある三つの新しい群から，容易に白血病の相対的発生率の高・低領域を決定することができる．１要因の場合にはほとんど問題はないが，多次元分割表間に存在する関連の程度を評価し，検出することはきわめて複雑になる．

Example 2: Four factors

The leukemia data which have been preclassified in some order by four factors were used. Table 2 shows the number of leukemia cases for the period 1 October 1950 – 30 September 1966 among the A-bomb survivors for the LSSE sample by city, sex, age ATB and radiation dose:

i = 1, 2 (City: Hiroshima and Nagasaki)
j = 1, 2 (Sex: Male and Female)
k = 1, 2, 3 (Age ATB: 0-14, 15-39 and 40+)
t = 1, 2, ..., 7 (Dose: 0-4, 5-19, 20-49, 50-99, 100-199, 200-299 and 300+)

例題2：4要因の場合

4要因によって順位づけて事前に分類された白血病データが利用された．表2は寿命調査拡大集団において，都市別，性別，原爆時年齢群別および被曝線量群別に，1950年10月1日から1966年9月30日までの期間中の被爆生存者について観察された白血病の症例数を示している．ただし，

i＝1，2 （都市：広島および長崎）
j＝1，2 （性：男および女）
k＝1，2，3 （原爆時年齢：0－14，15—39および40＋）
ι＝1，2，…，7 （線量：0－4，5－19，20—49，50—99，100－199，200－299および300＋）．

TABLE 2 NUMBER OF LEUKEMIA CASES AMONG A-BOMB SURVIVORS FOR THE LSSE SAMPLE BY CITY, SEX, AGE ATB AND DOSE 1 OCT. 1950 – 30 SEPT. 1966

表2 寿命調査拡大集団における被爆者にみられた白血病例：都市別，性別，原爆時年齢群別および被曝線量群別，1950年10月1日—1966年9月30日

City	Sex	Age ATB	Radiation Dose (rad)						
			<5	5-19	20-49	50-99	100-199	200-299	300 +
Hiroshima	Male	<15	0 5621	1 1671	3 582	0 296	1 150	0 66	6 83
		15-39	2 4601	0 1161	3 477	2 271	2 243	2 98	5 149
		40 +	6 5944	4 1625	3 681	0 407	0 317	2 88	1 131
	Female	<15	2 5455	1 1342	1 618	1 324	4 211	0 77	0 112
		15-39	1 10396	2 2816	2 1233	3 737	2 451	1 169	4 217
		40 +	8 7278	1 1911	3 946	2 567	2 269	0 103	3 114
Nagasaki	Male	<15	1 1483	0 903	0 238	0 191	1 172	3 84	2 86
		15-39	1 1318	0 387	0 174	0 169	1 231	1 173	1 99
		40 +	1 1088	1 502	0 207	0 149	1 121	0 79	2 70
	Female	<15	1 1638	1 951	0 261	0 202	0 195	0 109	2 85
		15-39	1 2315	0 1024	0 308	0 315	0 317	2 197	0 172
		40 +	0 1147	0 732	0 166	0 147	0 135	0 45	0 47

Note: The number of cases with leukemia and without leukemia is shown for each subset, say, $X_{1111} = (0, 5621)$

注：白血病例と非白血病例の数を各区分群ごとに示した．たとえば，X_{1111}＝(0，5621)．

Source data of Table 2 were obtained from ABCC IBM Tab # 01822-7.

表2の原資料はABCC IBM Tab．01822－7から求めた．

Let $X'_{ijkt} = (x_{ijkt1}, x_{ijkt2})$ be each component of groups preclassified by four factors.

$X'_{ijkl} = (x_{ijkt1}, x_{ijkt2})$ を4要因によって事前に分類された群の各成分であるとしよう．

Each marginal total for X_{ijkt} groups is composed of

X_{ijkt} 群の各周辺合計は都市に対して

$$X_{i\cdots} = \begin{pmatrix} 86, & 23 \\ 60008, & 18432 \end{pmatrix}$$

for city,

性に対して

$$X_{\cdot j\cdot\cdot} = \begin{pmatrix} 59, & 50 \\ 32586, & 45854 \end{pmatrix}$$

for sex,

年齢群に対して

$$X_{\cdot\cdot k\cdot} = \begin{pmatrix} 31, & 38, & 40 \\ 23206, & 30218, & 25016 \end{pmatrix}$$

and for age,

および　線量群に対して

$$X_{\cdots t} = \begin{pmatrix} 24, & 11, & 15, & 8, & 14, & 11, & 26 \\ 48284, & 15025, & 5891, & 3775, & 2812, & 1288, & 1365 \end{pmatrix}$$

for dose.

から構成される．四つの周辺合計から各2分割によって得られた χ^2 統計量の最大値は次の式によって決定される．

The maximum of χ^2 values obtained for each B.S. of four marginal totals is determined by

$$\chi^2_{(\gamma^d)} = C_1 * \left\{ \max_{\gamma^1=1} \left[A_{(\gamma^1)}, A_{(\gamma^2=1)}, \max_{1\leqslant\gamma^3\leqslant 2} A_{(\gamma^3)}, \max_{1\leqslant\gamma^4\leqslant 6} A_{(\gamma^4)} \right] - C_2 \right\}$$

where d = 1 (city), 2 (sex), 3 (age) or 4 (dose), and

ただし，d = 1（都市），2（性），3（年齢群）または4（線量群）を示す．そして

$$C_1 = \frac{(x_{\cdots\cdot 1} + x_{\cdots\cdot 2})^2}{(x_{\cdots\cdot 1} * x_{\cdots\cdot 2})} = \frac{(109 + 78440)^2}{(109 * 78440)} = 721.63$$

$$C_2 = \frac{x_{\cdots\cdot 1}^{\ 2}}{(x_{\cdots\cdot 1} + x_{\cdots\cdot 2})} = \frac{109^2}{78549} = 0.15126$$

$$A_{(\gamma^1=1)} = \frac{86^2}{60094} + \frac{23^2}{18455} = 0.15173$$

$$A_{(\gamma^2=1)} = \frac{59^2}{32645} + \frac{50^2}{45904} = 0.16109$$

$$\max_{1 \leqslant \gamma^3 \leqslant 2} A_{(\gamma^3)} = \max \left[A_{(\gamma^3=1)} = \frac{31^2}{23237} + \frac{78^2}{55312} = 0.15135 \right.$$

$$\left. A_{(\gamma^3=2)} = \frac{69^2}{53493} + \frac{40^2}{25056} = 0.15286 \right] = A_{(\gamma^3=2)} = 0.15286$$

$$\max_{1 \leqslant \gamma^4 \leqslant 6} A_{(\gamma^4)} = \left[\max A_{(\gamma^4=1)} = \frac{24^2}{48308} + \frac{85^2}{30241} = 0.25084, \right.$$

$$A_{(\gamma^4=2)} = \frac{35^2}{63344} + \frac{74^2}{15205} = 0.37948,$$

$$A_{(\gamma^4=3)} = \frac{50^2}{69250} + \frac{59^2}{9299} = 0.41044,$$

$$A_{(\gamma^4=4)} = \frac{58^2}{73033} + \frac{51^2}{5516} = 0.51760$$

$$A_{(\gamma^4=5)} = \frac{72^2}{75859} + \frac{37^2}{2690} = 0.57726$$

$$\left. A_{(\gamma^4=6)} = \frac{83^2}{77158} + \frac{26^2}{1391} = 0.57527 \right] = A_{(\gamma^4=5)} = 0.57726.$$

Thus, we have

このようにして,

$$\max \left[A_{(\gamma^1=1)} = 0.15173, A_{(\gamma^2=1)} = 0.16109, A_{(\gamma^3=2)} = 0.15286, A_{(\gamma^4=5)} = 0.57726 \right] = A_{(\gamma^4=5)} = 0.57726.$$

The maximum of χ^2 values is given by

が求められる. χ^2 統計量の最大値は

$$\chi^2_{(\gamma^4=5)} = 721.63 * [0.57726 - 0.15126] = 307.42.$$

As $\chi^2_{(\gamma^d)} = 307.42 > 3.841$, $\omega = (X_{ijk1}, \ldots, X_{ijk7})$ is divided into two new groups,

である. $\chi^2_{(\gamma^d)} = 307.42 > 3.841$ の不等式が成立するので, $\omega = (X_{ijk1}, \cdots, X_{ijk7})$ は二つの新しいグループに分割される.

$$\omega \begin{cases} \omega_1^4 = (X_{ijk1}, \ldots, X_{ijk5}) \\ \omega_2^4 = (X_{ijk6}, X_{ijk7}) \end{cases}$$

where i = 1, 2; j = 1, 2; k = 1, 2, 3.

ただし i＝1，2；j＝1，2；k＝1，2，3である．

The same process as described for ω will be applied by ω_1^4 and ω_2^4.

ωに対してと同様の手順がω_1^4およびω_2^4のそれぞれに適用される．

Figure 2 shows each step of groups obtained by means of a series of possible B.S. The final results are shown in Figure 3. The appropriate incidence of leukemia is evaluated through nine new groups as shown in Figure 3. It was noted that the highest region of leukemia incidence observed for the period 1 October 1950 – 30 September 1966 was 3.3% (11/330) of a new group composed of males who had been exposed to 200 rad or over at age under 15 ATB. The second highest region was 1.6% (14/901) of a new group composed of males who had been exposed to 200 rad or over at age 15 or over ATB. Thus, we can conclude that the risk of leukemia is the highest among males who were exposed to high radiation dose of 200 rad or over at ages under 15 ATB.

図2は一連の可能な2分割の手段による群削減の各段階を示している．最終結果は図3に示している．図3によれば，白血病の特定発生率は九つの新しい群を通じて評価される．1950年10月1日から1966年9月30日の期間中に観察された白血病発生率の最も高い領域は，原爆時年齢15歳未満の男で200 rad以上の被曝線量を受けた被爆者で構成される群である．この新しい群の白血病発生率は3.3％(11／330)であった．次に，高い発生率を示す領域はやはり男で，原爆時年齢15歳以上の群で200 rad以上の被曝線量を受けた被爆者から成る新しい群であり，その白血病発生率は1.6％(14／901)であった．このように，白血病の最も高い危険率は原爆時年齢15歳未満の男で200 rad以上の高線量を受けた被爆者であると結論づけることができよう．

IV. DISCUSSION

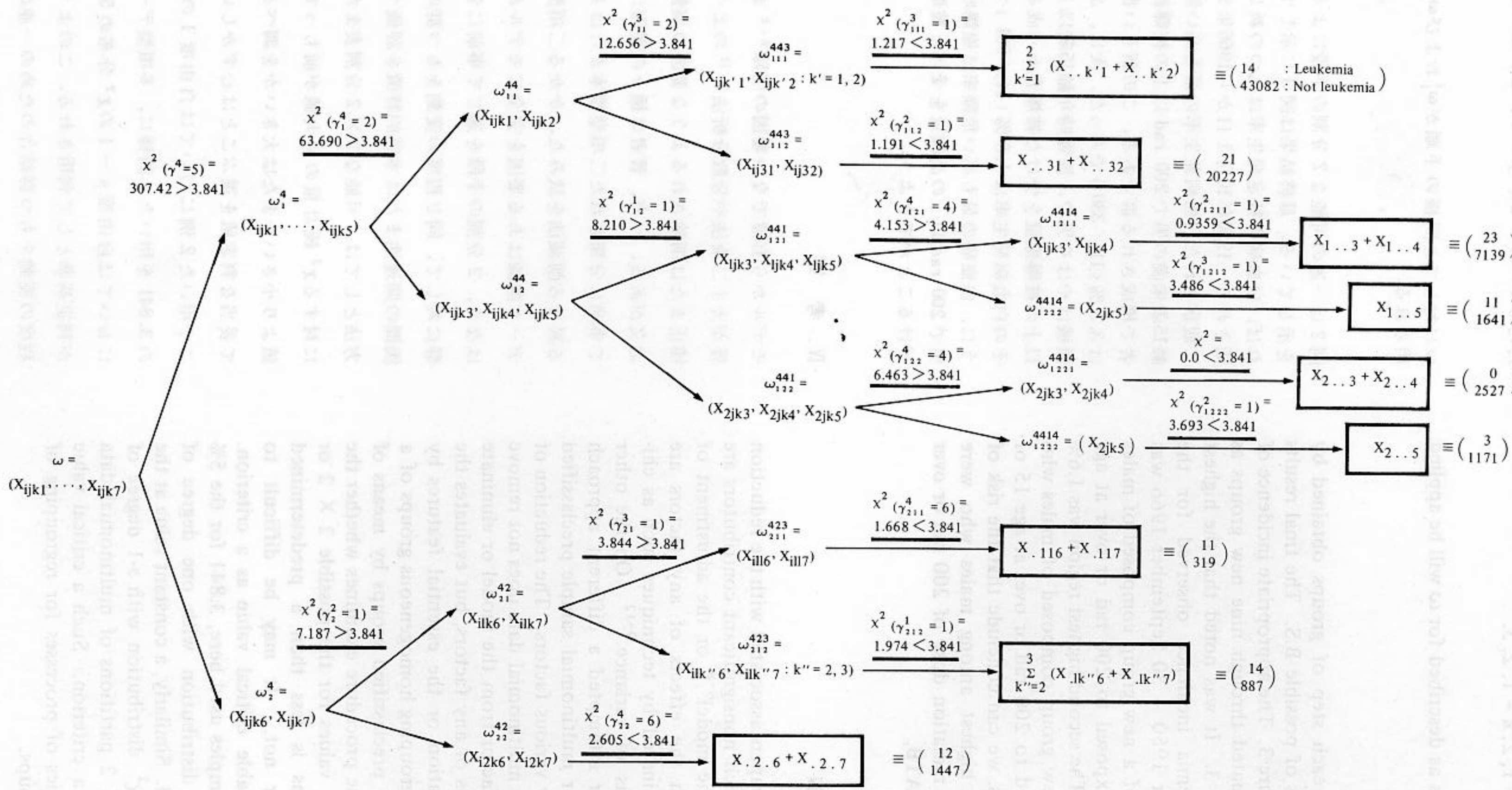
There are many papers associated with the reduction of factors in which nonsignificant contributors are removed from the model[6-8] or the adjustment of factors in which the effects of any factors are adjusted or eliminated by techniques such as chi-square or analysis of variance.[9-12] On the other hand, the author attempted a different approach for a binomial or multinomial sample preclassified in some order by various factors. The reduction of such binomial or multinomial data does not remove completely any factor from the model or eliminate simply the effects of any factors, but evaluates the degree of association or the essential features by combining or regrouping homogeneous groups of a large number of preclassified groups by means of binary splits. The procedure examines whether the maximum of χ^2 values for the possible 2 × 2 or s × 2 partitions is less than a predetermined critical value or not. It may be difficult to choose a reasonable critical value as a criterion. For the two examples used here, 3.841 for the 5% point of a χ^2 distribution with one degree of freedom was used. Similarly, a constant value at the 5% point of a χ^2 distribution with s-1 degree of freedom for s × 2 partitions of multinomial data will be used as a criterion. Such a critical value determines a series of processes for regrouping of homogeneous groups.

IV. 考　察

モデルから有意でない要因の削除[6-8]またある要因の影響がカイ二乗法や分散分析法[9-12]のような技法によって補正または削除されるような要因の調整に関する多くの論文がある．一方，著者は種々の要因によって順位づけて事前に分類された二項型標本または多項型標本に対する異なる削減法を試みた．かかる二項型または多項型のデータ削減はある要因を完全にモデルから削除するのではなく，2分割の手段を通じて事前に分類された多くの群に対して，同じ程度の変動をもつ群の結合によって要因間の関連性または本質的特徴を評価することにある．方法としては，可能な2×2分割表またはs×2分割表に対するχ^2統計量の最大値が前もって定められた判定値より小さいかまたは大きいかを調べる．判定規準として妥当な判定値を選ぶことはむずかしい問題である．ここで用いた2例については自由度1のχ^2分布の5％点の3.841を用いた．同様に，多項型データのs×2分割においては自由度s－1のχ^2分布の5％点での定数値が判定基準として利用される．このような判定値が同じ程度の変動をもつ群結合のための一連の手順を決定する．

FIGURE 2 REDUCTION OF BINOMIAL DATA PRECLASSIFIED BY CITY, SEX, AGE AND RADIATION DOSE

図２　都市，性，年齢，および被曝線量によって事前に分類された二項型データの削減

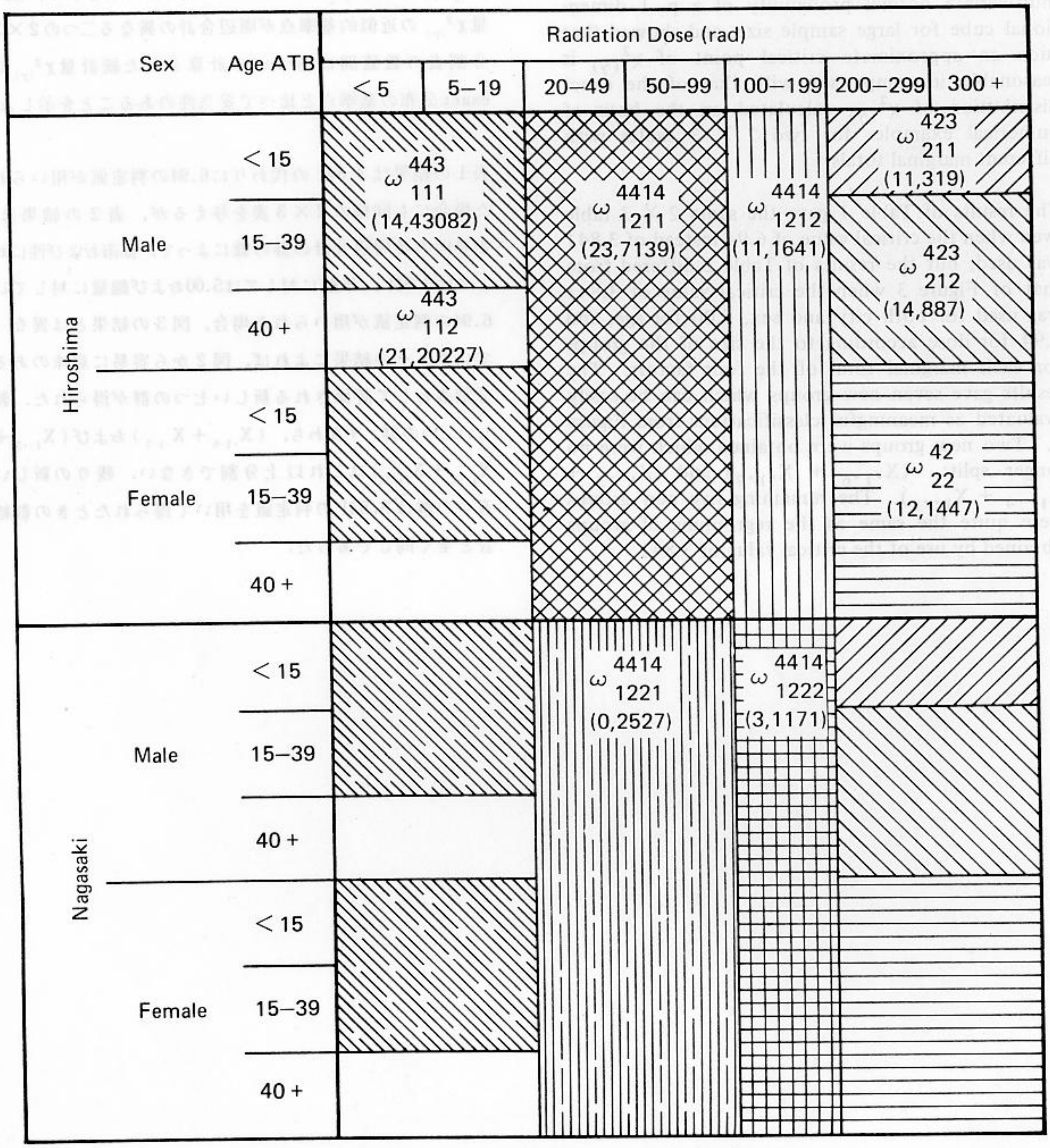


where i = 1, 2 (city); j = 1, 2 (sex); k = 1, 2, 3 (age), and $\boxed{}$ shows that the process of B.S. was terminated.

ただし i＝1，２（都市）；j＝1，２（性）；k＝1，２，３（年齢），および□は２分割の手順の終了を示す．

FIGURE 3 FINAL RESULT OBTAINED BY MEANS OF B.S.

図3　2分割法によって得られた最終結果



Note: Numbers in parentheses indicate number of leukemia cases and not leukemia cases.

注：かっこの数は白血病例数，非白血病例数を示す.

Sugiura and Otake[13] proposed the use of the approximate critical point of 6.94 instead of 3.841. They gave the approximate critical point for the distribution of $\chi^2_{(\gamma)}$ = max $[\chi^2_{(1)}, \chi^2_{(2)}, \ldots, \chi^2_{(p-1)}]$ which is expressed in terms of the multivariate normal probability of a p−1 dimensional cube for large sample size, and showed that such an approximate critical point of $\chi^2_{(\gamma)}$ is reasonable in comparison with that of the exact distribution of $\chi^2_{(\gamma)}$ calculated on the basis of numerical examples for two 2 X 3 tables with different marginal totals.

The results of Table 1 gave the same 2 X 3 table even when the critical value of 6.94 instead of 3.841 was used, but the results of Table 2 differed from that of Figure 3 when the critical value of 3.841 was used for both city and sex, 5.00 for age, and 6.94 for dose according to the size of the groups for each marginal total of the four factors. The results gave seven new groups which can be easily evaluated as meaningful classifications from Figure 2. Two new groups were obtained which were no longer split: $(X_{\cdot 1 \cdot 6} + X_{\cdot 1 \cdot 7})$ and $(X_{1 \cdot\cdot 3} + X_{1 \cdot\cdot 4} + X_{1 \cdot\cdot 5})$. The remaining five new groups were quite the same as the regrouping of groups obtained by use of the critical value of 3.841.

杉浦と大竹[13]は3.841の代わりに6.94の近似的基準点の利用を提案した．かれらはp－1次元正規分布の正方形領域で表わせる統計量$\chi^2_{(\gamma)}$ = max $\{\chi^2_{(1)}, \chi^2_{(2)}, \cdots, \chi^2_{(p-1)}\}$の極限分布における近似的基準点を示し，統計量$\chi^2_{(\gamma)}$の近似的基準点が周辺合計の異なる二つの2×3分割表の数値例に基づいて計算された統計量$\chi^2_{(\gamma)}$のexact分布の基準点と比べて妥当性のあることを示した．

表1の結果は3.841の代わりに6.94の判定値が用いられた場合にも同様の2×3表を与えるが，表2の結果は，4要因の各周辺合計の群の数によって，都市および性に対しては3.841，年齢に対しては5.00および線量に対しては6.94の判定値が用いられた場合，図3の結果とは異なっていた．その結果によれば，図2から容易に意味のある分割表として評価される新しい七つの群が得られた．新しい二つの群，すなわち，($X_{\cdot 1 \cdot 6}+X_{\cdot 1 \cdot 7}$)および($X_{1 \cdot\cdot 3}+X_{1 \cdot\cdot 4}+X_{1 \cdot\cdot 5}$)はこれ以上分割できない．残りの新しい五つの群は3.841の判定値を用いて得られたときの群結合と全く同じであった．

REFERENCES

参考文献

1. BELSON WA: Matching and prediction on the principle of biological classification. Appl Stat 8:65-75, 1959

2. MORGAN JN, SONQUIST JA: Problems in the analysis of survey data, and a proposal. J Am Stat Assoc 58:415-34, 1963

3. SONQUIST JA, MORGAN JN: The detection of interaction effects. Monograph No. 35, Survey Research Center Instit. for Social Research, University of Michigan, 6th ed, 1970 Lib of Cong. Catalog Card No. 64-63935

4. ISHIMARU T, HOSHINO T, ICHIMARU M, OKADA H, TOMIYASU T, TSUCHIMOTO T, YAMAMOTO T: Leukemia in atomic bomb survivors, Hiroshima and Nagasaki 1 October 1950-30 September 1966. Radiat Res 45, 216-33, 1971

5. BEEBE GW, USAGAWA M: The major ABCC samples, ABCC TR 12-68

6. PALSTON A, WILF HS: Mathematical methods for digital computers. John Wiley & Sons, Inc, 1960. Multiple regression analysis by Efroymson, 191-203

7. DRAPER NR, SMITH H: Applied regression analysis, John Wiley & Sons. Inc, 163-216, 1966

8. LAMOTTE LR, HOCKING RR: Computational efficiency in the selection of regression variables. Technometrics 12:83-93, 1970

9. MANTEL N: Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure. J Am Stat Assoc 58, 690-700, 1963

10. ARMITAGE P: The chi-square test for heterogeneity of proportions, after adjustment for stratification. J R Statist Soc., C, 21, 113-120, 1972

11. BIRTH MW: The detection of partial association, II: the general case. J R Statist Soc., B, 27, 111-124, 1966

12. STERLING TD, BINKS RG, HABERMAN S, POLLACK SV: Robot data screening – a ubiquitous automatic search technique. In Statistical Computation ed by Milton RC, Nelder JA, New York, Academic Press, 1969. pp. 319-33

13. SUGIURA N, OTAKE M: Approximate distribution of the maximum of c-1 χ^2-statistics (2X2) derived from 2Xc contingency table. Communications in Statistics, 1(1), 9-16, 1973