離散変量と連続変量が混在する場合の
統計的異常検出法

2011 年 8 月

飯田 孝久

主　論　文　要　旨

報告番号	甲 ㊁ 第　　号	氏 名	飯田 孝久

主論文題目：

離散変量と連続変量が混在する場合の統計的異常検出法

（内容の要旨）

本論文では、連続変量と離散変量が混在する異常検出問題において、離散変量の値を与えたとき、連続変量が分散共分散行列が共通の正規分布にしたがうとするロケーションモデルを仮定し、分布の母数が既知の場合と未知の場合について、異常検出法を構成した。誤報率が設定値に一致する、あるいはなるべく近い値になるように棄却限界値を定める方法を与え、誤報率および検出力の性質を明らかにし、手法間の比較を行った。なお、確率が小さい離散変量の水準での異常を確実に検出することは本研究での一つの目標である。

分布の母数が既知の場合は、離散変量の値を与えて連続変量のみに基づく検定を用いる条件付法(C 法)、離散変量を連続変量と併せて求めたマハラノビス平方距離を用いるマハラノビス距離法(M 法)と、全変量を用いた尤度比検定に基づく尤度比法(L 法)を構成した。どの方法においても、異常検出統計量は連続変量のみによるマハラノビス平方距離と補正項の和として表現されることから、誤報率が正確に設定値と一致する異常検出法を構成した。誤報率ならびに検出力に関する性質を明らかにした。補正項の性質から、M 法と L 法では正常状態で確率が小さい離散変量の水準ほど異常と判定しやすくなることがわかった。2 値変量の場合についての数値計算を基に手法の比較を行った結果、L 法とM法は確率が小さい水準で高い検出力を与える方法であることが確認できた。母数の状況により最適な方法は変化するが、総合的に判断してL法が優れていると結論づけられた。

分布の母数が未知の場合は、母数が既知の場合の 3 手法の異常検出統計量に初期データによる分布の母数の推定量を代入する推定方式と、初期データに判定標本を併せた全データに対する尤度比検定に基づく検定法(T 法)を構成した。棄却限界値は、初期データについて期待値をとった期待誤報率が設定値に近くなるよう、連続変量のみに基づくマハラノビス平方距離の分布として F 分布を用いて決定し、４つの手法について、誤報率ならびに検出力の基本的性質を明らかにした。期待誤報率が設定値に一致しないため、検出力の期待誤報率に対するオッズ比を用いて手法の比較を行った。C 法は離散変量を積極的に異常検出に用いていないことから、また、M 法は期待誤報率が設定値から大きく乖離する可能性があることから、詳細な比較はL 法と T 法について行った。既知の場合と同様に、母数の状況によりその優劣は変化するが、総合的にみて期待誤報率が安定し広い範囲でオッズ比の高いT法が優れていると結論づけられた。

SUMMARY OF Ph.D. DISSERTATION

School	Student Identification Number	SURNAME, First name
		IIDA Takahisa

Title

Statistical Methods for Detecting Abnormal Items When There Exist Both Categorical and Continuous Variables

Abstract

The problem of detecting abnormal items is discussed as a hypothesis testing problem for the case when both continuous and categorical variables are observed. Assuming the location model where continuous variables are multivariate normally distributed with common covariance matrix when categorical variables are observed, detection methods are derived with the false alarm probability as near the nominal value as possible for both cases when parameter values are known and unknown.

For the case when all parameter values of the distribution for the group of normal items are known, three detection methods are constructed. Conditional (C) method is based on the conditional distribution of continuous variables when categorical variables are observed. Mahalanobis distance (M) method uses the squared Mahalanobis distance by replacing the categorical variables by their dummy variables. Likelihood ratio (L) method is based on the likelihood ratio test. For these three methods, it is shown that the test statistics are expressed as the sums of Mahalanobis distance based on continuous variables and the correction terms which are determined by the probabilities of categorical variables. So, the distribution of these test statistics for normal items is a mixture of shifted χ^2 distributions. Based on this result, critical values are determined so that the nominal value of the false alarm probability is attained.

For the case when the parameter values are unknown, two types of detection methods are considered. One is estimative method, where estimates are substituted for the unknown parameters in the test statistics of C, M, and L methods. The other is the testing method (T method), which is derived as the likelihood ratio test using all data including initial data for normal items and testing sample.

Some basic properties for these methods are shown concerning their false alarm probabilities and conditional powers given the frequencies of categorical variables in the initial data and/or the values of categorical variables of testing sample. For comparing these methods, their expected error rates for normal items and the detecting powers are numerically evaluated when one dichotomous variable is included. For the case when parameter values are known, L method has higher power for a wide range of parameter values compared to other methods. For the case when parameter values are unknown, T method performs better in the sense that it has stable false alarm probability and higher power for a wide range of parameter values.

目 次

図 目 次

表 目 次

第1章　序論

1.1　異常検出と判別の問題

本論文は、離散変量と連続変量が混在する異常検出の問題について、統計学、特に仮説検定の観点から議論するものであり、著者の四編の論文（飯田, 他 (2008)、飯田, 他 (2009)、飯田・篠崎 (2010)、飯田・篠崎 (2011)) に基づいている。異常検出とは、個体についての観測値を基に、その個体が正常状態から乖離しているか否かを判断するものである。異常検出の問題の特徴をより明らかにするために、類似した問題として、測定されたデータを用いた、統計的手法による判別の問題を考えよう。判別の問題では、あらかじめいくつかの群を想定し、各群において変量が多変量正規分布にしたがうと考えられ、かつ、分散共分散行列が等しいと想定できる場合には線形判別関数を用いて、各個体がどの群に属するかを決定する。分散共分散行列が異なると考える場合には、マハラノビス平方距離による 2 次判別関数を用いるのが通常である。このような判別方法が使えるのは、観察される変量が各群内で一定の確率分布にしたがっていると考えてよい場合である。つまり、各群において、観測値の全体が何らかの分布にしたがっていると考えるのが自然であり、新しい観測値も同様な挙動をすると考えられる。

一方、異常検出の問題の例として健康診断を取り上げよう。健康診断では、身長・体重のような身体計測データや、コレステロールや GTP などの血液生化学検査などから健康かどうかを判定する。このとき、健康な人の集団について各種の計測値が一定の分布にしたがっていると想定することは自然であるが、さまざまな病気に罹っている人々のデータが、一定の分布にしたがっているとは考えにくい。健康診断（血液生化学検査など）の結果は、罹っている病気によって異なってくるし、罹っている病気の種類・頻度について想定するのは難しく、何らかの分布として記述することが困難であるからである。したがって、健康すなわち正常状態とそうでないかを判断する問題は、2 群の判別とは異なる問題であり、健康かどうか、すなわち、ある群に属すかどうかを判定する問題であると考えられる。また、火災報知器システムを考えてみても、通常の状況においては、温度などの状態は一定の分布に従っていると考えてもよいが、火災には煙が多く発生するものや火の回りが速いものなど様々な状況があり、火災発生時における観測値が一定の分布に従っているとは考えにくい。火災報知器は、正常か火災かを判定するというよりは、正常かそうでないかを判定するものであると考えるのが自然である。このほかにも異常検出問題の例としは、偽貨の判定、ネット上での攻撃の検出、倒産しそうな会社の特定など様々なものがある。このように、正常か異常かを判定する異常検出問題は、正常状態では観測値はある一定の分布にしたがっていると考えられるが、異常状態は様々であり、そのときの観測値について分布が想定できないという特徴がある。そのため、通常の判別問題との違いを考慮して、異常検出を非対称判別ということがある。

1.2 管理図法とMTS

異常検出の問題については、品質管理分野で用いられている統計手法として管理図法が知られている。連続的に製品を製造している工程において、一定の時間間隔ごとに標本を抽出し、管理特性の値を計測し平均と範囲をプロットし、その時系列的変化から工程に異常が発生しているかどうかを判定する $\bar{x} - R$ 管理図がその代表例として挙げられる。異常検出や判別分析では、手法の評価に正しく判定する確率 (あるいは逆に誤判別率) を用いるが、管理図法では、異常と判定するまでのサンプル数の期待値である平均連長 (ARL:Average Run Length) を用いることが多い。正常状態ではこの値が大きいほど、異常状態では小さいほど優れた手法と考えられる。管理図法は長い歴史をもち、数多くの研究がなされている。その基本的議論については、例えば、「JIS 品質管理」(1993) や仁科 (2009) などを参照されたい。なお、管理図法では、用いる標本の大きさは通常 1 よりも大きく、また、時系列データとして扱うことにも意味があり、データに時系列的構造を想定して解析されることが多いという特徴がある。

観測変数が複数個ある場合には、多変量管理図も用いられている。Hotelling の提唱した T^2 統計量を用いた管理図がよく知られているが、T^2 統計量は、変量間の分散共分散行列を用いたマハラノビス平方距離に他ならない。その研究も歴史をもっており文献も膨大な数に上るが、基本的文献として、Alt(1985), Tracy, Young and Mason (1992) および Mason and Young (2002) を挙げておく。T^2 統計量による平均値の検定問題における検出力について論じた文献として、Das Gupta and Perlman (1974) がある。管理図の文脈で、Wierda and Steerneman (1995) は、検出力の挙動について、また、Champ et al.(2005) は時系列的観点からの検討も含めて ARL の挙動について調べている。さらに、Jiang and Tsui (2008) は、広く仮説検定理論の立場から、T^2 管理図と他の多変量管理図との性能の比較を行っている。

一方、田口玄一氏は、健康診断や火災報知器システムのような異常検出を目的とする問題に対して、マハラノビス・タグチ (MT) システム、略して、MTS を提唱した（田口 (1993), 田口 (2002), 田口・兼高 (2002), 宮川 (2000) 等を参照)。正常群（単位空間とよばれている）における平均ベクトルと分散共分散行列を基にしたマハラノビス平方距離を用いて、新たな個体について異常か否かの判断を行うという手法である。MTS の一つの特徴は、前もって異常な個体についての観測値が得られていることを前提として、それらのデータに対するマハラノビス平方距離がなるべく大きくなるように変数選択を行うことである。(Taguchi and Rajesh(2000))。さまざまな解説書（立林 (2004)、長谷川 (2004)、立林, 他 (2008) など）や研究成果をまとめた書物（田口・兼高 (2002)、椿・河村 (2008)）が出版されており、工業分野などで豊富な応用例が示されている。しかし、Woodall et al. (2003) でも議論されているように、特に、統計学の観点から研究すべき課題は多く、さまざまな側面について研究が進められている。例えば、宮川・永田 (2003), 永田・久富 (2008), 永田・土居 (2009) などを参照されたい。

工程管理のために管理図法を用いる場合には、工程のさまざまな異常を検出したい、すなわち、既に考慮すべきことが判っている異常だけではなく、現状では想定できない異常も検出したいので、工程に関する情報を持つ変数はなるべく用いるのが自然である。そのように、想定できない異常も検出することが期待される管理図法に比べ、異常な個体について特定のデータが得られており、そのような異常な個体の検出を主な目的とした MTS の場合の方が、変数選択を行うことがより重要になる。

いくつかの連続変量が観測される場合には、判別分析・多変量管理図・MTS のどれをとってもマハラノビス平方距離を用いている。観測変量が多変量正規分布にしたがっていると想定できる場合は、誤判別率などを用いて判定の正確さを統計学的に正確に評価することが可能である。しかし、

観測変量が多変量正規分布にしたがっているとは考えられない場合には、マハラノビス平方距離に相当する量は計算できるが、それを用いたときの判定精度を正確に評価することは困難である。

1.3 離散変量の活用

これまでの異常検出の議論では、多くの場合、連続変量のみを用いている。離散変量を用いる場合でも、その変量は標本の欠点数（傷の数）などのように順序尺度構造をもっており、必要ならば変数変換を用いることで連続変量とみなして議論することがほとんどであった。

しかし、変量の中には、例えば健康診断における性別や喫煙・飲酒・定期的運動習慣の有無のような名義尺度構造しかもたない 2 値変量が含まれることも多い。火災報知器の問題における、設置場所が厨房か否かを表す変量も一つの例である。このような場合でも、MTS では 2 値変量をダミー変数化し、これを含めてマハラノビス平方距離を求め、変量の数を自由度とする χ^2 分布を用いて異常を判定するということが行われている（例えば兼高 (1987) など)。これは、多変量正規分布をもつ母集団からの標本について計算されるマハラノビス平方距離が χ^2 分布をすることを根拠としている。しかし、２値変量をあたかも正規変量であるかのごとく扱い、マハラノビス平方距離を用いるのは、統計学的厳密さを欠いている。これに関しては、Woodall et al.(2003) に対するコメントの中で Abraham and Variyath(2003) が言及している。なお、この他にも、分布の仮定をしないことやサンプリングに関することなどいくつかの問題点が MTS について指摘されている (Woodall et al. (2003))。

本論文では、離散変量と連続変量が混在しており、それらが同時に観測される場合の異常検出問題を取り扱う。例えば、健康診断において、身長・体重のような身体計測データや、コレステロールや γ–GTP などの血液生化学検査などの結果に加えて、性別や運動習慣の有無などを表す 2 値変量データを用いて、身体に異常があるかどうかを判定する場合がそれにあたる。異変を知らせる火災報知機の置かれているのが火気を使用する場所か否かを表す 2 値変量を用いるのもその例である。

あるいは、製品が最終検査で正常（OK) か異常（NG) と決定されるが、中間工程での検査あるいは簡便な検査で NG 品を検出することが望ましい状況を考える。中間あるいは簡便な検査の際に、計量データ以外に、ライン・機械・作業者の熟練度などの識別番号、原材料の種類などの層別因子が観測されているとすれば、これも連続変量とともに離散変量が観測されている場合の異常検出問題の一つの例となる。また、新型インフルエンザ感染に伴う海外渡航歴の有無などもその例である。

取上げる離散変量は、異常発生を起こしやすい状況であるか否かを表わすものであったり、あるいは、異常発生を示唆する現象の生起を表すものなど、何らかの意味で異常発生に関連している変量である。このような変量を適切に用いることが、正確に異常を検出することにつながると考えられる。

1.4 ロケーションモデル

本論文では、離散変量と連続変量が混在して観測される状況についての基本的統計モデルとして、Olkin and Tate(1961) が導入したロケーションモデルを用いる。それに基づき、統計学的に正確な議論を行う。ロケーションモデルでは、離散変量を与えたときの連続変量の分布が、平均ベクトルは異なるが分散共分散行列が共通の多変量正規分布にしたがうと仮定する。

いま、離散変量が 1 つの場合について具体的に述べよう。水準数 I の離散変量を X とし、その取りうる値を $1,\ldots,I$ とする。また、それぞれの値をとる確率を

$$p_x = P\{X = x\}, \quad x = 1,\ldots,I$$

とする。ただし、本論文では、２値変量の場合には、ダミー変数との対応関係から、水準値として 0 と 1 を用いることとする。m 次元の連続変量を $\boldsymbol{Y}$ とし、$X = x$ が与えられたとき $\boldsymbol{Y}$ は、平均 $\boldsymbol{\mu_x}$、分散共分散行列 Σ の m 次元正規分布 $N(\boldsymbol{\mu_x}, \Sigma)$ に従うものとする。つまり、$(\boldsymbol{Y}', X)'$ の実現値 $(\boldsymbol{y}', x)'$ が観測される確率・確率密度関数は

$$p_x(2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\{-(\boldsymbol{y}-\boldsymbol{\mu}_x)'\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_x)/2\} \tag{1.1}$$

と表現される。したがって、分布の母数は、離散変量 X の確率分布を記述する $p_x, x = 1,\ldots,I$ と、正規分布にしたがう連続変量 Y の平均 $\boldsymbol{\mu_x}, x = 1,\ldots,I$ および分散共分散行列 Σ である。

なお、離散変量が複数個存在する場合は、それらの水準組合せを新たな水準とみなすことで、離散変量が一つの場合に帰着させることができることに注意する。例えば、それぞれ 2 水準の値をとる変量を２つ考える場合、水準組み合わせとしては 4 通りとなり、水準数が４の１つの離散変量の場合に帰着できる。しかし、すべての水準組み合わせを考えると、変量の数が大きくなるにつれ水準数は大きくなり、それに伴って母数の合計数も増えてしまうことになる。考える水準組み合わせの数を減らしたり、あるいは、水準組み合わせの数を増やしても母数の数を増やさぬようにモデル化することが考えられている。（Krzanowski(1982) などを参照。）

ロケーションモデルは、離散変量と連続変量が混在する判別の問題ではしばしば適用されている。その研究を先導的に推進したのは Krzanowski(1975,1980,1982,1983,1986) である。特に、Krzanowski (1983) では、ロケーションモデルにおいて、松下距離に基づく判別分析を提案し、Bar-Hen and Daudin(1995) は Kullback-Leibler 情報量の離散変量部分と連続変量部分への分解ならびにその漸近的性質を論じた。中西 (1999) は、母集団間の距離と誤判別率の関係を論じ、それを用いた変数選択法を提案した。これらはいずれも 2 群の判別問題についての議論であった。Nakanishi(2003) はロケーションモデルにおける判別問題を検定問題として定式化し、その検定統計量である Kullback-Leibler 情報量の漸近正規性を示し、検出力をシミュレーションにより求めた。この中では、2 群の判別に加えて、新たなデータが与えられた群に属すかどうかという判定の問題（異常検出問題と解釈することができる）も検定の枠組みで取りあげている。

1.5 統計的仮説検定問題としての異常検出

異常検出の問題を統計的仮説検定の問題として定式化することについて一通り述べておこう。いま、連続変量と離散変量を併せた $(\boldsymbol{Y}', X)'$ の確率分布がロケーションモデルによって記述され、確率・確率密度関数が (1.1) 式で与えられるものとする。正常群の確率分布を π で表わし、母数がすべて既知であるとし、それを $p_x^{(0)}, \boldsymbol{\mu}_x^{(0)}, x = 1,\ldots,I, \Sigma$ とする。異常であるか否かを判定したい個体についても判定標本 $(\boldsymbol{Y}', X)'$ が観測され、その分布が π でないと判断すべきかどうかが問題なのだと考えられる。したがって、異常検出の問題は、仮説：判定標本 $(\boldsymbol{Y}', X)'$ の確率分布は π である、の検定問題に定式化される。より具体的な仮説は、$H_0 : p_x = p_x^{(0)}, \boldsymbol{\mu} = \boldsymbol{\mu}_x^{(0)}, x = 1,\cdots,I$ と記述されることになる。

正常群での確率分布の母数は未知である場合が通常であろうが、その場合には正常群からの無作為標本が得られるとし、判定標本と併せて、判定標本の確率分布が π であるとの仮説を検定するこ

とになる。素朴な方法は、既知の場合の検定方式に現れる母数をその推定量で代用することであるが、その性質について議論するためには厳密な取り扱いが必要となる。

仮説検定における基本的概念について、異常検出の問題に即して述べておこう。第 1 種の誤りは、正常であるのに異常であると判断してしまう誤りであり、第 2 種の誤りは、異常であるのにそれを見過ごして正常と判断してしまう誤りである。2 種類の誤りを犯す確率を問題にすることになるが、第 1 種の誤りを犯す確率については、誤報率と表現することにする。誤報率をこの値にしたいという設定値が仮説検定での有意水準である。誤報率をその設定値に一致させることは、母数が既知の場合には可能であることが示されるが、母数が未知の場合には一般には一致させることができない。そこで、異常検出法を評価するためには、期待誤報率が設定値にどの程度合致しているかも議論しなければならない。そのためには、離散変量の値を与えたときの条件付誤報率の概念も必要になる。第 2 種の誤りを犯す確率を 1 から引いた値が検出力であり、仮説検定の議論では検出力が用いられる。異常検出の問題でも検出力、つまり、異常であるときに正しく異常と判定する確率を用いて議論する。

離散変量の混在する異常検出の問題を仮説検定問題としてとらえるとき、一つの素朴な方法として、離散変量の値が与えられたという条件の下での連続変量の分布を用いて検出を行うことが考えられる。母数を既知とする場合について述べれば以下の通りである。離散変量 $X = x$ という条件の下での連続変量 $\boldsymbol{Y}$ の条件付分布が $N(\boldsymbol{\mu}_{\boldsymbol{x}}, \Sigma)$ であることから、異常であるかどうかを判断したい個体について $(\boldsymbol{Y}', X)'$ を観測したとき、$X = x$ であったならば $\boldsymbol{Y}$ に基づき $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_x^{(0)}$ の検定を行うことにより、異常か否かを判断するのである。この検定は χ^2 分布を用いて行うことができ、誤報率をその設定値に一致させることができる。本論文では、この方法を条件付法とよんでいる。健康診断等において、男女で標準値が異なるような場合には、このような判断方法も妥当性をもつかもしれない。しかし、この方法は、離散変量の値を言わば層別のためにしか活かしておらず、異常状態で離散変量の確率分布自身が変化し、離散変量の値そのものが異常であるか否かの判断のために重要な情報をもつ場合には、有効な方法ではないことに注意する。

1.6 本論文の概要

本論文では、離散変量と連続変量が混在する場合について、ロケーションモデルを仮定して、統計的仮説検定の観点から異常検出問題を議論する。より具体的には

1. 異常検出法の構成
2. 誤報率の設定値を実現するための棄却限界値の決定法
3. 検出力についての検討

という問題を取り扱う。議論は、大きく

. 正常群での確率分布の母数が既知の場合

. 正常群での確率分布の母数が未知の場合

の 2 つに分かれており、それぞれ 2,3 章および 4,5 章の議論に対応する。母数が既知の場合には 3 つの異常検出法を取り扱い、未知の場合には 4 つの異常検出法について議論することになる。以下、章ごとにその概要を述べておこう。

第 2 章では、正常群での確率分布の母数が既知の場合の異常検出法を構成し、誤報率の設定値を実現するための棄却限界値の設定、さらに、誤報率などについて手法間の比較を行う。はじめに、判定標本の離散変量 X の値が与えられたという条件の下での異常検出法について述べる。ロケーションモデルの下で、正常群については、$X = x$ のとき連続変量 $\boldsymbol{Y}$ が正規分布 $N(\boldsymbol{\mu}_x^{(0)}, \Sigma)$ にしたがっているので、$X = x$ が与えられたとき異常検出の問題は平均値の検定に帰着し、χ^2 分布を用いて誤報率の設定値を各 x ごとに実現することができる。これを条件付法とよび、C 法と表記する。

つぎに、MTS で通常用いられる、離散変量をダミー変数化し全変量を基にして求めるマハラノビス平方距離を用いる異常検出法について議論している。この全変量に基づくマハラノビス平方距離が、χ^2 分布にしたがう連続変量に基づくマハラノビス平方距離と、離散変量の確率によって定まる定数（これを補正項とよぶ）の和として表現されることを示す。つまり、$X = x$ のとき

$$(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})' \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)}) + (1 - p_x^{(0)})/p_x^{(0)}$$

と表されることがわかる。したがって、正常状態での全変量に基づくマハラノビス平方距離が、位置をずらした χ^2 分布の混合分布にしたがうので、誤報率の設定値を正確に実現するように棄却限界値を決定することができる。この方法を、マハラノビス距離法とよび M 法と表記する。また、自由度が [連続変量の数+用いたダミー変数の数] の χ^2 分布を用いて棄却限界値を定める MTS では、誤報率は設定値に一致せず大きく異なることがあることが示される。さらに、複数の 2 値変量が存在する場合について、2 値変量の連続変量の平均への効果が加法的であれば、全変量に基づくマハラノビス平方距離がやはり χ^2 分布にしたがう連続変量に基づくマハラノビス平方距離と補正項の和として表現され、正確な誤報率を実現する異常検出が可能であることが示される。

また、判定標本が、(1.1) 式で p_x を $p_x^{(0)}$, $\boldsymbol{\mu}_x$ を $\boldsymbol{\mu}_x^{(0)}$ で置き換えた正常群の確率分布にしたがっているという仮説の尤度比検定に基づいて異常検出法を導出し提案する。このとき、導かれる統計量は、M 法と同様に χ^2 分布にしたがう連続変量に基づくマハラノビス平方距離と離散変量の確率によって決まる補正項の和として表現されることが確認される。これに基づき、正確に誤報率の設定値を実現するように棄却限界値を定めた異常検出法が構成できる。これを尤度比法 (L 法) とよぶ。M 法と L 法では補正項の関数形が異なり、M 法の方が離散変量の水準による差が大きく、補正項の影響が大きい。この差が、離散変量の値を与えたときの条件付誤報率の挙動の違いとして現れることを示している。

第 3 章では、2 章で構成した 2 つの異常検出法である M 法、L 法に C 法を含め、3 つの異常検出法の性能について比較検討を行っている。分布の母数が既知の場合は、誤報率を正しく設定値に一致させることができるので、基本的には検出力の挙動について議論している。そのための準備としての意味もあって、離散変量の値が与えられたときの条件付誤報率について、3 つの手法による差異を明らかにしている。C 法では条件付誤報率は離散変量の水準によらず一定であるが、M 法の場合に水準による差が最も大きくなることが示される。これが 3 手法の条件付検出力の挙動の差異につながることになる。

検出力は、異常状態における離散変量の値を与えたときの条件付検出力の、離散変量の分布についての平均として表現される。また、条件付検出力は、離散変量の各水準での連続変量の平均の正常状態での値からのずれを表す非心度により定まることがわかる。さらに、異常状態での離散変量の分布の変化のあり方により、どの方法が最も検出力の高い優れた方法であるかは変わり、一概に優劣をつけることができないことがわかる。

3 つの方法を、離散変量の水準数が 3 以上の場合について比較することは困難なので、2 値変量が 1 つの場合について比較する。2 値変量の確率のみが変化する場合は、正常状態で確率が小さい

水準の確率が異常状態では増大するとき、M 法の検出力が最も高く、つぎに L 法、そして C 法の順となる。確率の変化の方向が逆の場合は検出力も逆の順となる。連続変量の平均値も変化する場合には、M 法、L 法は、C 法に比べると、正常状態で確率が小さい 2 値変量の値における条件付検出力を大幅に大きくし、正常状態で確率が大きい場合は条件付検出力を小さくすることがわかる。M 法の場合に特に顕著である。2 値変量の各水準における非心度を固定するとき、$p_0^{(0)} \geq 0.5$ とすると、$X = 0$ である確率が $p_0^{(0)}$ から小さくなるにつれて最適な手法は C 法、L 法、M 法と変化することがわかる。総合的には、L 法が、母数の通常考えるべき領域の広い範囲で、つまり、正常状態で確率の小さい離散変量の水準の確率が大きくなり、非心度もある程度大きくなる状況ならば、3 つの手法の中で最も高い検出力をもつことが確認される。

第 4 章では、分布の母数が未知の場合について、4 つの異常検出法を構成し、誤報率の設定値を実現するための棄却限界値の設定法を議論し、さらに、誤報率の観点から 4 つの方法について比較する。

母数が未知の場合は、正常群からの無作為標本（これを初期データとよぶ）が得られるものとし、これと判定標本を併せて、異常検出法を構成している。異常検出法は初期データに依存するため、実際の誤報率は変動してしまう。そのため、初期データについて期待値をとった期待誤報率を問題にすることになる。

一つの構成法は、正常群の確率分布の母数が既知の場合の異常検出法に現れる母数に、初期データに基づく推定量を代入するという、推定方式であり、C 法、M 法、L 法の 3 通りが考えられる。離散変量の確率分布は相対頻度を用いて推定し、離散変量を与えたときの連続変量の平均は初期データにおいて離散変量の水準ごとの平均を用いる。連続変量の分散共分散行列については、全体でプールした偏差積和行列を定数で割った量を推定量として用いる。最尤推定量、Σ の不偏推定量、逆行列が Σ^{-1} の不偏推定量となる推定量の 3 通りの推定量の選び方を取り上げ、どの選択が期待誤報率を設定値に近づけるという意味で適切であるかを検討している。

さらに、正常群からの標本である初期データと判定標本とが同じ分布をもつとの仮説の尤度比検定に基づく異常検出法を導いている。これを検定方式（T 法）とよぶが、やはり、連続変量に基づくマハラノビス平方距離を用いることになる。

以上の 4 方法について、期待誤報率がなるべく設定値に一致するように棄却限界値を定める方法について議論している。その方法には 2 つあり、1 つは χ^2 分布法であり、もう 1 つは F 分布法である。3 つの推定方式については、χ^2 分布法とは、母数が既知のとき離散変量の値を与えたとき連続変量だけに基づくマハラノビス平方距離が χ^2 分布にしたがうので、これを近似的分布として用いて棄却限界値を定める方法である。T 法については、χ^2 分布法とは一般の尤度比検定統計量の漸近分布としての χ^2 分布を用いる方法である。F 分布法は、4 つの方法に共通であり、初期データの離散変量の頻度分布および判定標本の離散変量の値を与えたとき、連続変量に基づくマハラノビス平方距離の定数倍が F 分布にしたがうことを用いて棄却限界値を定める方法である。いずれも期待誤報率を設定値に正確に一致させることができないので、その正確さについて評価することが必要になる。

4 つの異常検出法と 2 つの棄却限界値を定める方法による期待誤報率の正確さについて、1 つの 2 値変量の場合について数値的に評価している。4 つの異常検出法のいずれについても、χ^2 分布法よりも F 分布法の方が期待誤報率が設定値に近い値で安定していることが確認される。F 分布法を用いる場合、分散共分散行列の推定量の選択はあまり影響せず、総合的には T 法が L 法、M 法、C 法に比べ優れていることが示される。2 値変量の両水準とも初期データで観測されるという条件を課すとき、特に T 法では、設定値に非常に近い期待誤報率を与えることも確認される。

第 5 章では、４章で与えられた４つの母数が未知のときの異常検出法について、主に検出力の観点から比較を行っている。

そのために、まず、判定標本の離散変量の値が与えられたときの条件付誤報率について、相対的に確率の小さい水準の方が条件付誤報率は大きくなることを示している。それを用いて、離散変量の分布のみが変化した場合の検出力について、正常状態で相対的に確率の小さい水準の確率が大きくなるほど、検出力は大きくなることを示している。逆に言えば、このような離散分布の変化を検出することを主眼とすべきであるということになる。連続変量の平均も変化するときの検出力については、非心度が大きくなるほど検出力が大きくなることは示される。それ以上の一般的性質を厳密に議論することは困難であるが、非心度が水準によらずほぼ等しい場合には、正常状態で相対的に確率の小さい離散変量の水準の確率が大きくなるほど、検出力が大きくなることが示唆される。

さらに、１つの 2 値変量が混在する場合について、数値計算を基に４つの異常検出法についての比較を行っている。４つの手法とも、期待誤報率を設定値に一致させることができないので、単に検出力の大きさだけでは公平な比較が難しい。そこで、各手法について検出力の期待誤報率に対するオッズ比を用いて比較を行っている。その結果、推定方式の中の L 法と検定方式の T 法が、異常状態の広い範囲で安定して優れた異常検出法であることが確認される。特に、正常状態で確率の値が相対的に小さい離散変量の水準の確率が、異常状態で格段に大きくなるということがなければ、T 法が優れていることが示される。また、C 法に比べ、他の 3 法は、正常状態で確率の小さい離散変量の水準での条件付検出力を大きくし、逆に、確率の大きい水準での条件付検出力を下げることになっており、その度合いは M 法、L 法、T 法の順に強いことがわかる。

M 法は、期待誤報率が設定値を大幅に超えることがあり、検出力を議論する以前に、異常検出法として問題があると言うべきである。さらに、C 法は、離散変量の分布が異常状態で変化する可能性を無視しており、離散変量の分布の情報が全く生かされない方法である。総合的に判断して、T 法が推奨される。

最後に第 6 章では、本論文の成果をまとめ結論を述べている。

第2章　母数が既知の場合の異常検出法

この章では、ロケーションモデルにおいて、分布の母数が既知の場合に正確な誤報率を実現する異常検出法を構成する。まずはじめに、離散変量の水準が与えられたという条件の下で連続変量のみに基づき判断する条件付法を紹介する。その後、離散変量をダミー変数化して求めたマハラノビス平方距離の表現を求め、それに基づいて異常検出法を構成する。特に複数の 2 値変量が混在する場合のマハラノビス平方距離の表現とそれに基づく異常検出法について議論する。最後に、判定標本が正常群の分布からのものであるとする仮説の検定に対する尤度比検定統計量を用いる方法を構成する。いずれも正確な誤報率を実現できることに注意する。なお、各手法における誤報率ならびに条件付誤報率の性質については、手法を構成した後に示すが、手法間での条件付誤報率の比較は、検出力を比較するときの基礎となるので、第 3 章でまとめて議論する。

2.1　条件付異常検出

この方法は、離散変量 X の値が与えられたときの m 次元連続変量 $\boldsymbol{Y}$ の条件付分布に対して、それぞれ有意水準 α の検定を行い異常を検出する方法である。ここで、正常群の確率分布の母数は、母数が既知であることから右上に (0) を付けて表すことにする。正常状態では $X = x$ のとき連続変量 $\boldsymbol{Y}$ は平均ベクトル $\boldsymbol{\mu}_x^{(0)}$、分散共分散行列 Σ の正規分布に従うので、この検定は帰無仮説

$$H_0 : \boldsymbol{\mu}_x = \boldsymbol{\mu}_x^{(0)}$$

を対立仮説

$$H_1 : \boldsymbol{\mu}_x \neq \boldsymbol{\mu}_x^{(0)}$$

に対して有意水準 α で検定する問題と表現される。

これは通常の平均値の検定であり、尤度比検定統計量は連続変量のみを基にして求めたマハラノビス平方距離 $(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})$ になる。帰無仮説が真のとき、このマハラノビス平方距離が自由度 m の χ^2 分布にしたがうことを用いて検定する。すなわち、離散変量 X の値が与えられたときの条件付誤報率が、それぞれ水準 α に一致するように棄却域を定める方法である。したがって、離散変量の水準が与えられたときの条件付誤報率は全て α なので、全体としての誤報率も α となることがわかる。

この方法では、離散変量の分布が正常状態のそれから変化する可能性を検出することを目標としてはいないことに注意する。なお、この方法は離散変量の値が与えられたという条件の下での異常検出なので、条件付法とよび C 法と表記する。

2.2　離散変量も含めたマハラノビス平方距離に基づく異常検出

この節では、離散変量が一つの場合に、それをダミー変数化して求めたマハラノビス平方距離の表現を求め、それに基づいて、正確な異常判定ルールを構成する。MTS の議論では、離散変量を連

続変量であるかのようにしてマハラノビス平方距離を求めて異常検出が行われているが、それは根拠のないことではなく、正確に誤報率を実現するように修正することが可能であることを示すことになる。

2.2.1 2 値変量が一つの場合のマハラノビス平方距離

まず、2 値変量が一つの場合を取り上げる。健康診断データにおける性別や、火災報知システムが設置される場所における火気の有無などを用いるのがこの場合に相当する。

2 値変量 X と、m 次元連続変量 $\boldsymbol{Y}$ が観測されるとし、それらを合わせて $\boldsymbol{U} = (\boldsymbol{Y}', X)'$ と表すことにする。2 値変量 X は、一方の値を 0、他方を 1 とする。$X = x$ となる確率を $p_x, x = 0, 1,$ とする。

連続変量 $\boldsymbol{Y}$ はロケーションモデルの仮定から、$X = x$ のとき平均 $\boldsymbol{\mu}_x$、分散共分散行列 Σ の正規分布にしたがうとする。

このとき $\boldsymbol{U}$ の期待値は、$E(X) = p_1, \quad E(\boldsymbol{Y}) = p_0\boldsymbol{\mu}_0 + p_1\boldsymbol{\mu}_1 = \boldsymbol{\mu}_*$ で与えられ、分散共分散行列 Ω は、積の期待値から期待値の積を引くことで、

$$\Omega = \begin{bmatrix} V(\boldsymbol{Y}) & Cov(\boldsymbol{Y}, X) \\ Cov(X, \boldsymbol{Y}) & V(X) \end{bmatrix} = \begin{bmatrix} \Sigma + p_0p_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' & p_0p_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ p_0p_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' & p_0p_1 \end{bmatrix} \tag{2.1}$$

と表現できることがわかる。ここで、分割された正定符号行列の逆行列についての以下の表現を用いる。

$$\begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} H^{-1} & -H^{-1}G' \\ -GH^{-1} & C^{-1} + GH^{-1}G' \end{bmatrix}, \tag{2.2}$$

ここで、

$$G = C^{-1}B', \quad H = A - BC^{-1}B'$$

である。分散共分散行列 Ω に適用すると、$G = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)', H = \Sigma$ となるので、分散共分散行列 Ω の逆行列は

$$\Omega^{-1} = \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ -(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1} & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' + 1/(p_0p_1) \end{bmatrix} \tag{2.3}$$

となる。

マハラノビス平方距離 Δ^2 は、上で求めた分散共分散行列 Ω の逆行列を用いて、

$$\begin{pmatrix} \boldsymbol{Y} - \boldsymbol{\mu}_* \\ X - p_1 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} \boldsymbol{Y} - \boldsymbol{\mu}_* \\ X - p_1 \end{pmatrix} \tag{2.4}$$

で与えられる。これを、Σ^{-1} を含む部分とそれ以外の部分に分けて整理すると、

$$\Delta^2 = \{\boldsymbol{Y} - (1 - X)\boldsymbol{\mu}_0 - X\boldsymbol{\mu}_1\}'\Sigma^{-1}\{\boldsymbol{Y} - (1 - X)\boldsymbol{\mu}_0 - X\boldsymbol{\mu}_1\} + (X - p_1)^2/(p_0p_1) \tag{2.5}$$

となるので、$X = x$ のとき Δ^2 は

$$\Delta^2 = (\boldsymbol{Y} - \boldsymbol{\mu}_x)'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_x) + (1 - p_x)/p_x \tag{2.6}$$

と表現できる。$X = x$ のとき $\boldsymbol{Y} \sim N(\boldsymbol{\mu}_x, \Sigma)$ であり、(2.6) 式の第 1 項は自由度 m の χ^2 分布にしたがう。第 2 項は、2 値変量の確率分布によって決まり、確率 p_x が小さいときほど大きな値をとる。これを以降、補正項と呼ぶことにする。これより、マハラノビス平方距離は、X が観測されたとき、χ^2 分布に従う部分と、2 値変量の確率によって決まる定数の和として表現できる。さらに、Δ^2 は、それらを確率 p_0, p_1 で混合した確率分布を持つことがわかる。マハラノビス平方距離は (2.4) 式で計算すれば、$\boldsymbol{U}$ の平均と分散共分散行列を用いて求めることができる。2 値変量の値で条件付けした連続変量 $\boldsymbol{Y}$ だけを用いたマハラノビス平方距離を求める必要はない。

2.2.2　2 値変量の場合におけるマハラノビス平方距離を用いた異常検出

マハラノビス平方距離の分布がわかったので、それを利用して異常検出ルールを仮説検定の観点から構成することができる。正常な場合には

$$Pr(X = x) = p_x^{(0)}, \quad E(\boldsymbol{Y}|X = x) = \boldsymbol{\mu}_x^{(0)}$$

とし、$X = x$ が観測されたときの $\boldsymbol{Y}$ の分散共分散行列については、x によらず一定とし、

$$V(\boldsymbol{Y}|X = x) = \Sigma$$

とする。新たな異常であることが疑われる個体についての観測値 $(\boldsymbol{Y}', X)'$ について、

$$Pr(X = x) = p_x, \quad E(\boldsymbol{Y}|X = x) = \boldsymbol{\mu}_x$$

としたとき、帰無仮説

$$H_0 : p_x = p_x^{(0)},\ \boldsymbol{\mu}_x = \boldsymbol{\mu}_x^{(0)}, \qquad x = 0, 1 \tag{2.7}$$

を検定する。$\boldsymbol{\mu}_*^{(0)} = p_0^{(0)}\boldsymbol{\mu}_0^{(0)} + p_1^{(0)}\boldsymbol{\mu}_1^{(0)}$ とおくとき、検定に用いるマハラノビス平方距離は

$$\Delta^2 = \begin{pmatrix} \boldsymbol{Y} - \boldsymbol{\mu}_*^{(0)} \\ X - p_1^{(0)} \end{pmatrix}' \Omega^{-1} \begin{pmatrix} \boldsymbol{Y} - \boldsymbol{\mu}_*^{(0)} \\ X - p_1^{(0)} \end{pmatrix} \tag{2.8}$$

であり、(2.6) 式から $X = x$ のとき

$$\Delta^2 = (\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)}) + (1 - p_x^{(0)})/p_x^{(0)} \tag{2.9}$$

と表現される。

有意水準 α の異常検出ルールを定めるには、前節で求めたマハラノビス平方距離の仮説が真のときの分布の上側 $100\alpha\%$ 点 K を求めることが必要になる。すなわち、

$$Pr(\Delta^2 \geq K) = p_0^{(0)}Pr(\Delta^2 \geq K|X = 0) + p_1^{(0)}Pr(\Delta^2 \geq K|X = 1) = \alpha$$

となる K を求めることになる。仮説が正しいとき、$X = x$ という条件の下で (2.9) 式の第 1 項が自由度 m の χ^2 分布に従うことに注意して、自由度 m の χ^2 分布の分布関数を $F(\cdot)$ で表すとき、K は方程式

$$p_0^{(0)}F(K - p_1^{(0)}/p_0^{(0)}) + p_1^{(0)}F(K - p_0^{(0)}/p_1^{(0)}) = 1 - \alpha \tag{2.10}$$

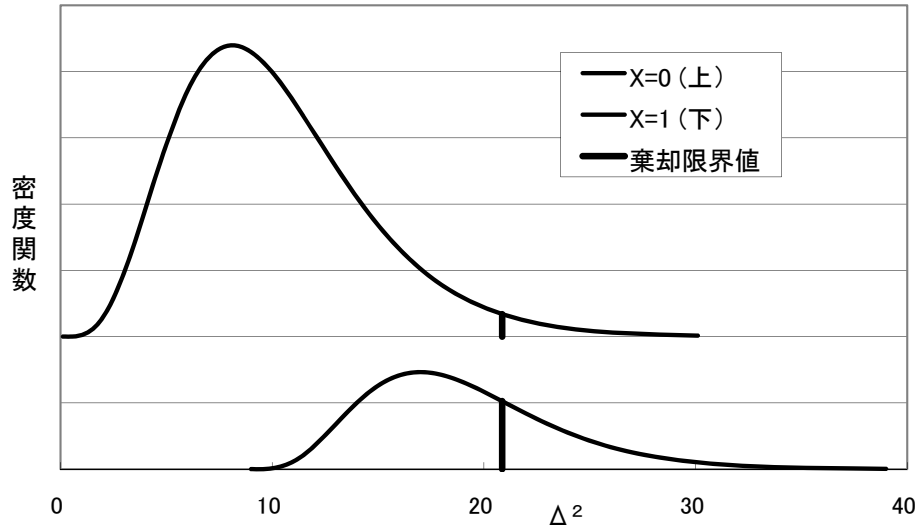

図 2.1: 棄却限界値の決定 ($p_0^{(0)} = 0.9, m = 10$)

の解として与えられる。この様子を図示したのが図 2.1 であり上側が $X = 0$、下側が $X = 1$ のときの分布である。それぞれ、平行移動した χ^2 分布で、面積はそれぞれ $p_0^{(0)}, p_1^{(0)}$ で、左右の位置の違いは補正項の値の差に対応している。図の中に縦に入れた線が K の位置を示し、この線の右側の面積が両者合わせて α になる。(この図は見やすくするため、$X = 1$ の密度関数の値を 3 倍している。)

(2.10) 式の左辺は K の単調関数なので、方程式 (2.10) は反復法で容易に解を求めることができる。その解を $K(p_0^{(0)})$ と表す。このようにして定まる異常検出ルールをマハラノビス距離法と呼び M 法と表記する。また、仮説が正しいときの $Pr(\Delta^2 \geq K|X = x)$ を、$X = x$ のときの条件付誤報率と呼ぶことにする。$p_0^{(0)} > 0.5$ ならば、条件付誤報率は、$X = 1$ のときの方が $X = 0$ のときより大きい。

2.2.3 条件付誤報率の挙動

マハラノビス距離法において、$X = 0$ および 1 のときの条件付誤報率を、それぞれ α_0, α_1 と表すことにする。この α_0 と α_1 は $p_0^{(0)}$ の関数である。ここでは、$p_0^{(0)}$ が 1/2 から 1 まで変化したときの条件付誤報率の変化を調べてみよう。便宜上、$p_0^{(0)} = p$ とおいて α_x を p の関数として $\alpha_x(p), x = 0, 1$ と表すことにすると、以下の性質が成立する。

性質 2.1

(i) $\alpha_0(1/2) = \alpha_1(1/2) = \alpha$. $\quad \lim_{p\to 1} \alpha_0(p) = \alpha, \quad \lim_{p\to 1} \alpha_1(p) = 1$.

(ii) $\alpha_0(p) < \alpha < \alpha_1(p), \qquad 1/2 < p < 1$.

(iii) $\alpha_1(p)$ は $1/2 < p < 1$ の範囲で p の単調増加関数であり、$1 - \alpha < p^* < 1$ なる p^*が存在し、
$p^* \leq p \leq 1$ で$\alpha_1(p) = 1$ である。

(iv) $\alpha_0(1 - \alpha) < 1 - F((1 - 2\alpha)/(\alpha(1 - \alpha)))$.

(証明) $p_0^{(0)}$ を p と表すことにし、全体での棄却限界値を $K(p)$ で表したが、$X = x$ のときの統計量 $(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})$ の条件付棄却限界値を $K_x(p)$ で表す。すなわち、

$$K_0(p) = K(p) - (1 - p)/p, \qquad K_1(p) = K(p) - p/(1 - p) \tag{2.11}$$

とする。また、Δ^2 の棄却限界値 $K(p)$ を与える (2.10) 式は

$$pF(K(p)-(1-p)/p)+(1-p)F(K(p)-p/(1-p))=1-\alpha \tag{2.12}$$

となる。

性質 (i) (2.12) 式において、$p=1/2$ または $p\to 1$ とすればよい。$p=1/2$ のときは、補正項の値はどちらも 1 になるので、条件付誤報率はどちらも α になる。$p=1$ のときは、確率 1 で $X=0$ となり、$X=0$ における条件付誤報率は α になる。また、p が 1 に近づくとき、(2.12) 式で定まる $K(p)$ について $\lim_{p\to 1}K(p)<\infty$ であり、$X=1$ の場合の補正項は ∞ に発散するので、$K_1(p)\to-\infty$ である。つまり、p がある一定値より大きいとき、$X=1$ のときの条件付誤報率は 1 となる。

性質 (ii) 自由度 m の χ^2 分布の上側 $100\alpha\%$ 点を $\chi^2_m(\alpha)$ とする。(2.12) 式の左辺の $K(p)$ に $\chi^2_m(\alpha)+(1-p)/p$ を代入すると、$p/(1-p)>(1-p)/p$ より左辺の値は $1-\alpha$ より小さくなる。また、$\chi^2_m(\alpha)+p/(1-p)$ を代入すると $1-\alpha$ より大きくなる。したがって

$$\chi^2_m(\alpha)+(1-p)/p<K(p)<\chi^2_m(\alpha)+p/(1-p) \tag{2.13}$$

となり、性質 (ii) が示される。

性質 (iii) 2 つの補正項の差を

$$g(p)=p/(1-p)-(1-p)/p$$

で表す。$g(p)$ は $1/2<p<1$ の範囲で正の値をとる増加関数である。(2.12) 式を $K_1(p)$ を用いて表現すると、

$$pF(K_1(p)+g(p))+(1-p)F(K_1(p))=1-\alpha \tag{2.14}$$

となる。(2.14) 式の両辺を p で微分すると、

$$F(K_1(p)+g(p))+p(K_1'(p)+g'(p))f(K_1(p)+g(p))-F(K_1(p))+(1-p)K_1'(p)f(K_1(p))=0$$

となる。ここで、$f(\cdot)$ は、自由度 m の χ^2 分布の密度関数である。$K_1'(p)$ について整理すると、

$$K_1'(p)\{pf(K_1(p)+g(p))+(1-p)f(K_1(p))\}=-F(K_1(p)+g(p))+F(K_1(p))-pf(K_1(p)+g(p))g'(p) \tag{2.15}$$

となる。左辺の $K_1'(p)$ にかかる因子は正で、右辺は負なので、$K_1'(p)<0$ である。よって、$X=1$ における条件付棄却限界値 $K_1(p)$ は、単調に減少する。(i) の証明で述べたように、$p\to 1$ のとき、$K_1(p)\to-\infty$ なので、条件付誤報率はその値が 1 に達するまで単調に増加する。

性質 (iv) (2.12) 式左辺の p に $1-\alpha$ を代入すると

$$(1-\alpha)F(K(1-\alpha)-\alpha/(1-\alpha))+\alpha F(K(1-\alpha)-(1-\alpha)/\alpha)=1-\alpha$$

であり、

$$F(K(1-\alpha)-(1-\alpha)/\alpha)=\{(1-\alpha)/\alpha\}\{1-F(K(1-\alpha)-\alpha/(1-\alpha))\}$$

となる。したがって $K(1-\alpha)-(1-\alpha)/\alpha>0$ なので、$X=0$ のときの条件付棄却限界値 $K_0(1-\alpha)$ は

$$K_0(1-\alpha)=K(1-\alpha)-\alpha/(1-\alpha)>(1-\alpha)/\alpha-\alpha/(1-\alpha)=(1-2\alpha)/\alpha(1-\alpha) \tag{2.16}$$

となり、$X=0$ のときの条件付誤報率 $1-F(K_0(1-\alpha))$ は $1-F((1-2\alpha)/\{\alpha(1-\alpha)\})$ より小さいことがわかる。

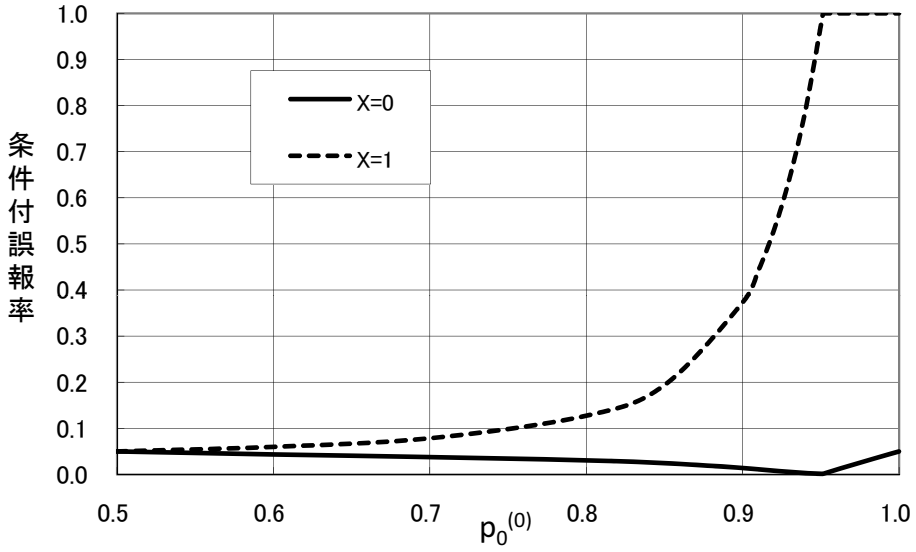

図 2.2: M 法の条件付誤報率 ($m = 10, \alpha = 0.05$)

□

$m = 10$ のときの条件付誤報率を p の関数として与えたのが図 2.2 である。(ii) からわかるように、条件付誤報率は、$X = 0$ のとき全体での水準 α より小さく、$X = 1$ のとき α より大きくなる。また、(i) より $\alpha_0(p)$ は、$p = 1/2$ のときの値も、p が 1 に近づいたときの極限値もともに α であることから、一度減少した後に増加に転じるので単調関数ではないことがわかる。(iii) から、$\alpha_1(p)$ は p が 1 に近いとき 1 になってしまう、つまり、$X = 1$ ならば必ず帰無仮説を棄却することになる。

(iv) から、α が 0 に近づくと、$(1 - 2\alpha)/(\alpha(1 - \alpha))$ の値が発散するため、α_0 が 0 に近づくことがわかる。α_0 の挙動を、自由度が 5，10，20 の場合について示したのが図 2.3 である。

自由度が小さい場合、すなわち連続変量の数が少ないときほど α_0 の小さくなり方は顕著である。このとき、正常な場合に異常と判定することはほとんどないが、後で述べるように、異常が発生した場合の $X = 0$ のときの条件付検出力が低下する可能性がある。

2.2.4 2 値変量を正規変量とみなす異常検出法における誤報率

2 値変量を正規変量とみなし、自由度が［連続変量の数＋ 2 値変量の数］の χ^2 分布の上側 $100\alpha\%$ 点を超えたときに異常と判定することがある。この方法を、「簡便法」と呼ぶ。この場合、帰無仮説のもとで実際に仮説を棄却する確率は設定した有意水準 α に必ずしも一致しないことが数値計算から確認できる。2 値変量の数を 1 とし有意水準を 5% として、2 値変量が 0 をとる確率 $p_0^{(0)}$ の値を 0.5 から 1 まで変化させたときに実際に仮説を棄却する確率を連続変量の数 m が $5, 10, 20$ の場合について求めた。連続変量の数が 5 の場合、$X = 0$ のとき（図 2.4) の条件付誤報率は 4.08% から 2.72% へ単調に減少し、$X = 1$ のとき（図 2.5）の条件付誤報率は 4.08% から単調に増加し、$p_0^{(0)} = 0.923$ 付近で 100% になった後は 100% のままである。全体での誤報率は、これら 2 つの値を $p_0^{(0)}, 1 - p_0^{(0)}$ で重み付けした平均値であり、図 2.6 に示すように、$p_0^{(0)} = 0.5$ から 0.8 付近までは誤報率は 5% より小さいが、0.8 付近で 5% を超え、0.9 付近で最大値をとった後 0.97 付近で再度 5% を下回ることがみてとれる。

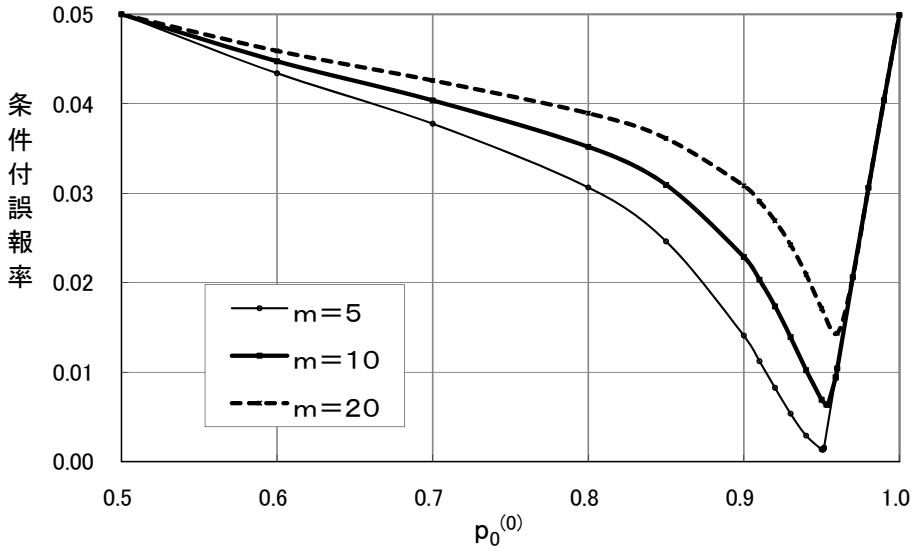

図 2.3: M 法の条件付誤報率 $[X = 0](m = 5, 10, 20, \alpha = 0.05)$

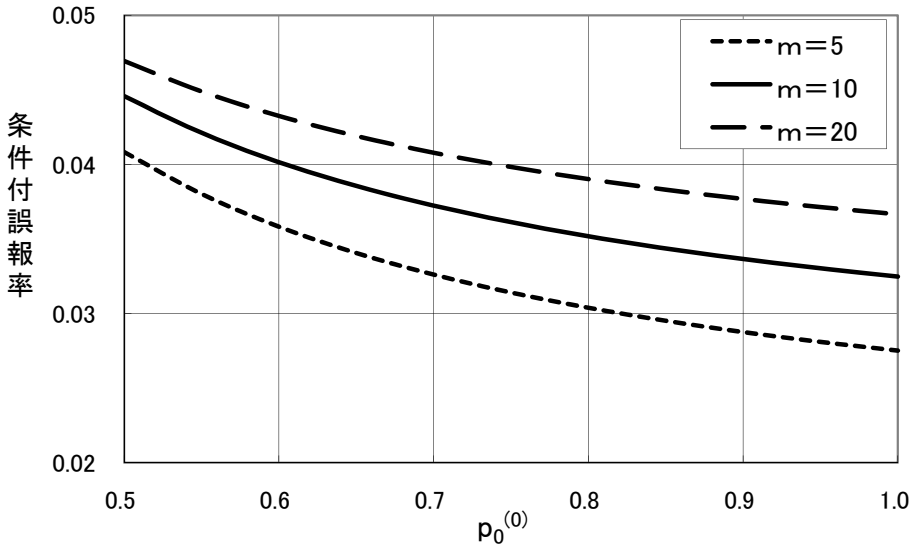

図 2.4: 簡便法の条件付誤報率 $[X = 0](m = 5, 10, 20, \alpha = 0.05)$

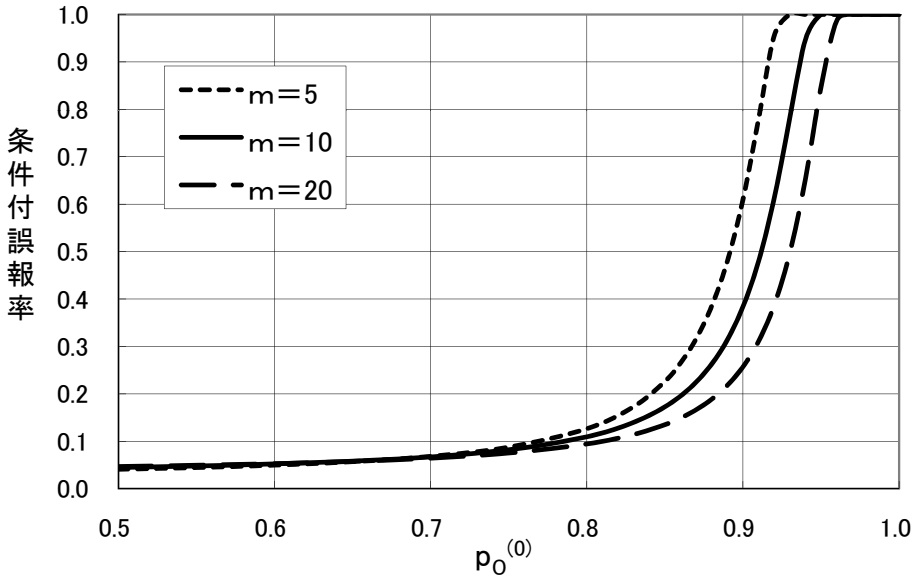

図 2.5: 簡便法の条件付誤報率 $[X = 1](m = 5, 10, 20, \alpha = 0.05)$

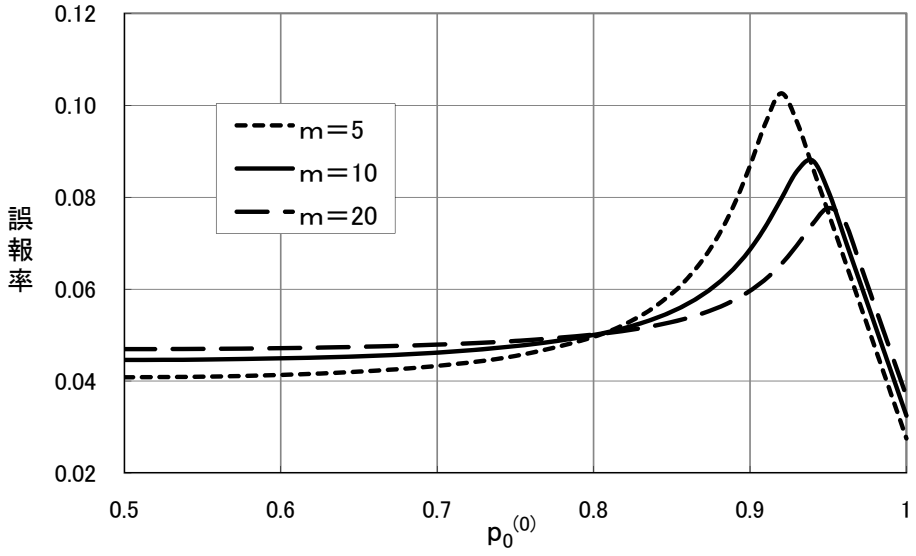

図 2.6: 簡便法の誤報率 $(\mathrm{m}= 5, 10, 20, \alpha = 0.05)$

$m = 10, 20$ の場合も、$m = 5$ の場合と同様の挙動をするが、変動幅は縮小し全体として有意水準 0.05 に近づき、誤報率の最大値は減少する。したがって、連続変量の数が多いほど、与えられた有意水準からの乖離は小さくなることがわかる。

以上より、簡便法では、$p_0^{(0)}$ の値により誤報率が有意水準より小さい場合や大きい場合があり、$p_0^{(0)}$ の値によっては大きく異なることもあることが示された。数値計算は 2 値変量の場合で行ったが、多値変量の場合にも同様の問題が発生すると考えられる。

したがって、離散変量をダミー変数化し、正規変量のごとく扱った場合、誤報率を設定値に一致させることが困難であることが確認できた。以降、離散変量をダミー変数化した場合の簡便法による棄却限界値の決定は採用しないこととする。

2.2.5 多水準離散変量の場合

ここでも、ロケーションモデルを仮定する。すなわち、$X = x, x = 1, \ldots, I,$ のとき連続変量 $\boldsymbol{Y}$ の分布は $N(\boldsymbol{\mu}_x, \Sigma)$ とする。離散変量の水準数が 3 以上の場合、各水準に対して値 1 をとるダミー変数列を構成して水準による違いを表現することができる。しかし、すべてのダミー変数を用いると、分散共分散行列がランク落ちするため逆行列が存在せず、マハラノビス平方距離を求めることはできない。そこで、ダミー変数を 1 つ取り除いてマハラノビス平方距離を求める。正常状態で $X = x$ となる確率を $p_x^{(0)}$ とし $p_x^{(0)} > 0$ とすると、$X = x$ のときのマハラノビス平方距離は、取り除くダミー変数によらず (2.9) 式の形に表されることが確認できる。したがって、2 値変量の場合と同様に、仮説が正しいとき、χ^2 分布に従う変量と確率 $p_x^{(0)}$ によって定まる定数の和になる。この結果から、有意水準 α の異常検出ルールは、

$$\sum_{x=1}^{I} p_x^{(0)} F(K - (1 - p_x^{(0)})/p_x^{(0)}) = 1 - \alpha \tag{2.17}$$

を満足する K を用いて、$\Delta^2 > K$ と定められる。2 値の場合と同様、確率 $p_x^{(0)}$ が小さい水準 x ほど補正項が大きいので、条件付誤報率が大きくなることがわかる。

MTS 関連の文献では、脳疾患患者の排尿自立達成の予測 (田口、兼高 (2002)pp295-304) で取り上げている 6 種類の疾患、衝突防止センシングシステム (田口、兼高 (2002)pp189-195) におけるシーン分けなどが多水準離散変量として用いられている。これらの場合、ロケーションモデルの仮定が妥当であれば正確な有意水準を持つように棄却限界値を決めることができる。

2.3 複数の 2 値変量が混在する場合のマハラノビス平方距離

2.3.1 マハラノビス平方距離の表現

ここでは、複数の 2 値変量が存在する場合におけるマハラノビス平方距離の分布について述べる。健康診断において、飲酒・喫煙・運動などの習慣の有無を尋ねる場合がこれにあたる。火災報知システムの例においても、データとしてガスコンロを用いた調理を行っているかどうかや喫煙者の有無など複数の要因が用いられている。一般に 2 値変量の数を k とし、とる値は 0 または 1 とする。これらを並べた k 次元ベクトル確率変数を $\boldsymbol{X}$ で、その取りうる値を$\boldsymbol{t}$ で表す。2^k 個の$\boldsymbol{t}$ を辞書式順序

で並べた $k \times 2^k$ 行列を A とする。$k = 3$ のとき、A はつぎのように、各列が左から順に 0 から 7 の 2 進展開に対応している行列になる。

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

前節と同様にロケーションモデルを仮定する。すなわち、$\boldsymbol{X} = \boldsymbol{t}$ が与えられたときの連続変量 $\boldsymbol{Y}$ の分布は、平均が $\boldsymbol{\mu_t}$ で分散共分散行列が Σ の正規分布とし、連続変量の数を m とする。平均ベクトル $\boldsymbol{\mu_t}$ を $\boldsymbol{t}$ の値の辞書式順序で並べた $m \times 2^k$ 行列を

$$M = [\boldsymbol{\mu_t}; \boldsymbol{t} = \boldsymbol{0}, \ldots, \boldsymbol{1}]$$

とする。$\boldsymbol{X} = \boldsymbol{t}$ の確率 $Pr(\boldsymbol{X} = \boldsymbol{t})$ を $p_{\boldsymbol{t}}$ とし、これらを$\boldsymbol{t}$ の値の辞書式順序で並べた列ベクトルを $\boldsymbol{p}$ とし、これらを対角要素にもつ対角行列を D とする。ここでも、前節と同様に平均ベクトルと分散共分散行列は既知とする。

マハラノビス平方距離の表現を求めるのに必要な分散共分散行列ならびにその逆行列の表現を求める。$\boldsymbol{X}$ の期待値は、対応する$\boldsymbol{t}$ の要素が 1 である水準組合せの確率 $p_{\boldsymbol{t}}$ を加えたものであり、$\boldsymbol{Y}$ の期待値は、$\boldsymbol{\mu_t}$ の重み付平均である。変量の積についても同様にして、

$$E(\boldsymbol{X}) = A\boldsymbol{p}, \quad E(\boldsymbol{Y}) = M\boldsymbol{p} = \boldsymbol{\mu}_*,$$

$$E(\boldsymbol{X}\boldsymbol{X}') = ADA', \quad E(\boldsymbol{Y}\boldsymbol{Y}') = \Sigma + MDM', \quad E(\boldsymbol{Y}\boldsymbol{X}') = MDA'$$

を得る。分散および共分散は、積の期待値から期待値の積を引いて求められるので、$E = D - \boldsymbol{p}\boldsymbol{p}'$ とおくと、以下を得る。

$$V(\boldsymbol{X}) = ADA' - A\boldsymbol{p}\boldsymbol{p}'A' = AEA', \qquad V(\boldsymbol{Y}) = \Sigma + MDM' - M\boldsymbol{p}\boldsymbol{p}'M' = \Sigma + MEM',$$

$$Cov(\boldsymbol{Y}, \boldsymbol{X}) = MDA' - M\boldsymbol{p}\boldsymbol{p}'A' = MEA'.$$

ここで、離散変量の分散共分散行列 $V(\boldsymbol{X}) = AEA'$ が正則であることを仮定する。すべての$\boldsymbol{t}$に対し $p_{\boldsymbol{t}} > 0$、すなわちどの水準組合せも起こりうる場合には成立することが確認できる。

分散共分散行列 Ω は

$$\Omega = \begin{bmatrix} V(\boldsymbol{Y}) & Cov(\boldsymbol{Y}, \boldsymbol{X}) \\ Cov(\boldsymbol{X}, \boldsymbol{Y}) & V(\boldsymbol{X}) \end{bmatrix} = \begin{bmatrix} \Sigma + MEM' & MEA' \\ AEM' & AEA' \end{bmatrix}$$

と表現できる。前節と同様にして分散共分散行列 Ω の逆行列を求めると

$$\Omega^{-1} = \begin{bmatrix} H^{-1} & -H^{-1}G' \\ -GH^{-1} & (AEA')^{-1} + GH^{-1}G' \end{bmatrix} \tag{2.18}$$

となる、ここで

$$G = (AEA')^{-1}AEM', \quad H = \Sigma + MEM' - MEA'(AEA')^{-1}AEM'$$

である。マハラノビス平方距離 Δ^2 は、この Ω^{-1} を用いて、

$$\begin{bmatrix} \boldsymbol{Y} - M\boldsymbol{p} \\ \boldsymbol{X} - A\boldsymbol{p} \end{bmatrix}' \Omega^{-1} \begin{bmatrix} \boldsymbol{Y} - M\boldsymbol{p} \\ \boldsymbol{X} - A\boldsymbol{p} \end{bmatrix} \tag{2.19}$$

と表すことができる。

2.3.2　加法性の仮定が成立している場合のマハラノビス平方距離

行列 H は、2 値変量が 1 つの場合と異なり、必ずしも Σ に一致するわけではない。しかし、次に定義する平均の加法性が成立している場合には、$H = \Sigma$ となり、$\boldsymbol{X} = \boldsymbol{x}$ のとき Δ^2 を (2.6) 式の形で表現できる。

定義 2.1（平均の加法性）
ある m 次ベクトル $\boldsymbol{\mu}_0$ および $m \times k$ 行列 Λ が存在して、

$$\boldsymbol{\mu_t} = \boldsymbol{\mu}_0 + \Lambda \boldsymbol{t} \tag{2.20}$$

と表すことができるとき、平均は 2 値変量に対して加法的であるという。

(2.20) 式中の Λ の各列は、対応する 2 値変量が 0 から 1 へ変化したときの連続変量の平均の変化量を表すので、(2.20) 式は、2 値変量の水準による連続変量の平均の変化が、主効果のみで表され交互作用が存在しないことを表現している。Krzanowski(1980,1982) は、加法モデルを含め、2 値変量の高次の交互作用を無視するモデルを 2 群の判別問題について提案している。実験計画法の要因実験において高次の交互作用を無視するのと同じ考え方であり、加法性が成立していると考えてよい状況も珍しくはないと考えられる。

$\boldsymbol{\mu_t}$ を列ベクトルに並べた行列が M なので、

$$M = \boldsymbol{\mu}_0 1' + \Lambda A = \begin{bmatrix} \boldsymbol{\mu}_0 & \Lambda \end{bmatrix} B \tag{2.21}$$

と表せる。ここで、$B' = \begin{bmatrix} \mathbf{1} & A' \end{bmatrix}$ であり、$\mathbf{1}$ はすべての要素が 1 の 2^k 次ベクトルである。(2.21) 式は M' の値空間が B' の値空間に含まれていると言い換えることができる。このとき、次の定理を示すことができる。

定理 2.2　平均の加法性が成立しているならば $H = \Sigma$ である。

（証明）いま、

$$E\mathbf{1} = (D - \boldsymbol{p}\boldsymbol{p}')\mathbf{1} = D\mathbf{1} - \boldsymbol{p}\boldsymbol{p}'\mathbf{1} = \boldsymbol{p} - \boldsymbol{p} = \mathbf{0}$$

であることに注意する。仮定より M' の値空間が B' の値空間に含まれており、m 次元ベクトル $\boldsymbol{\mu}_0$ と $m \times k$ 行列 Λ を用いて、$M' = \mathbf{1}\boldsymbol{\mu}_0' + A'\Lambda'$ と表すことができるので、

$$\begin{aligned} MEM' - MEA'(AEA')^{-1}AEM' &= ME(\mathbf{1}\boldsymbol{\mu}_0' + A'\Lambda') - MEA'(AEA')^{-1}AE(\mathbf{1}\boldsymbol{\mu}_0' + A'\Lambda') \\ &= MEA'\Lambda' - MEA'\Lambda' \\ &= O \end{aligned}$$

となり、$H = \Sigma$ を得る。

□

注意 1.　$p_{\boldsymbol{t}} = 0$ のとき、$\boldsymbol{\mu_t}$ の値はマハラノビス平方距離に影響を与えないので、定理での加法性の条件は、さらに『$p_{\boldsymbol{t}} > 0$ なる任意の$\boldsymbol{t}$ に対し、平均の加法性が成立している』とすることができる。

注意 2. すべての$\boldsymbol{t}$ に対して $p_{\boldsymbol{t}} > 0$ とすると、$H = \Sigma$ ならば平均の加法性が成立することが示せるので、平均の加法性が成り立たなければ $H \neq \Sigma$ となる。

注意 2 の証明

$S = AEA'(= V(\boldsymbol{X}))$ とおくと、S は正則であることが確認できる。さらに、E は非負定符号であることも示されるので、

$$E - EA'S^{-1}AE = E^{1/2}(I - E^{1/2}A'S^{-1}AE^{1/2})E^{1/2} \tag{2.22}$$

と表され、$(I - E^{1/2}A'S^{-1}AE^{1/2})$ は正射影行列なので、(2.22) 式は非負定符号である。したがって、

$$\begin{aligned} M(E - EA'S^{-1}AE)M' = O &\Leftrightarrow (E - EA'S^{-1}AE)M' = O \\ &\Leftrightarrow \mathcal{R}(M') \subset \mathcal{N}(E - EA'S^{-1}AE). \end{aligned} \tag{2.23}$$

これより $\mathcal{N}(E - EA'S^{-1}AE) = \mathcal{R}(B')$ を示せばよいことがわかる。

$\mathcal{N}(E - EA'S^{-1}AE) \supset \mathcal{R}(B')$　であることは、

$$(E - EA'S^{-1}AE)\mathbf{1} = \mathbf{0}, \qquad (E - EA'S^{-1}AE)A' = EA' - EA' = O$$

から確認される。したがって、$rank(E - EA'S^{-1}AE) \geq 2^m - rank(B')$　を示せばよい。いま、

$$E = (E - EA'S^{-1}AE) + EA'S^{-1}AE$$

と分解する。一般に $rank(A + B) \leq rankA + rankB$ より

$$rank(E) \leq rank(E - EA'S^{-1}AE) + rank(EA'S^{-1}AE),$$

であり、$rank(E) = 2^m - 1, \qquad rank(EA'S^{-1}AE) = rank(A)$ であることが示されるので、

$$rank(E - EA'S^{-1}AE) \geq 2^m - 1 - rankA = 2^m - rankB'$$

が得られる。

□

平均の加法性が成立するとき、分散共分散行列 Ω の逆行列は、(2.18) 式において $H = \Sigma$ とおくことにより

$$\Omega^{-1} = \begin{bmatrix} I \\ -(AEA')^{-1}AEM' \end{bmatrix} \Sigma^{-1} \begin{bmatrix} I & -MEA'(AEA')^{-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & (AEA')^{-1} \end{bmatrix} \tag{2.24}$$

となる。したがって、(2.19) 式で表されるマハラノビス平方距離は、

$$\boldsymbol{Z} = (\boldsymbol{Y} - M\boldsymbol{p}) - MEA'(AEA')^{-1}(\boldsymbol{X} - A\boldsymbol{p}) \tag{2.25}$$

とおくと、

$$\Delta^2 = \boldsymbol{Z}'\Sigma^{-1}\boldsymbol{Z} + (\boldsymbol{X} - A\boldsymbol{p})'(AEA')^{-1}(\boldsymbol{X} - A\boldsymbol{p}) \tag{2.26}$$

となる。

ここで、(2.25) 式に (2.21) 式を代入し、 $\mathbf{1}'\boldsymbol{p} = 1, \mathbf{1}'E = \mathbf{0}'$ を用いると

$$\begin{aligned}\boldsymbol{Z} &= \boldsymbol{Y} - (\boldsymbol{\mu}_0\mathbf{1}' + \Lambda A)\boldsymbol{p} - (\boldsymbol{\mu}_0\mathbf{1}' + \Lambda A)EA'(AEA')^{-1}(\boldsymbol{X} - A\boldsymbol{p}) \\ &= \boldsymbol{Y} - \boldsymbol{\mu}_0 - \Lambda A\boldsymbol{p} - \Lambda(\boldsymbol{X} - A\boldsymbol{p}) \\ &= \boldsymbol{Y} - (\boldsymbol{\mu}_0 + \Lambda\boldsymbol{X})\end{aligned}$$

となる。したがって $\boldsymbol{X} = \boldsymbol{t}$ のとき、$\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{\mu_t}$ であることがわかり、(2.26) 式の右辺の第 1 項は

$$\boldsymbol{Z}'\Sigma^{-1}\boldsymbol{Z} = (\boldsymbol{Y} - \boldsymbol{\mu_t})'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_t}) \tag{2.27}$$

となり、自由度 m の χ^2 分布に従うことがわかる。

(2.26) 式の右辺の第 2 項は、k 個の 2 値変量が独立、すなわち $p_{\boldsymbol{t}} = \prod_{i=1}^{k} Pr(X_i = t_i)$ であれば、

$$AEA' = diag(Pr(X_j = 0)Pr(X_j = 1))$$

となることから、$\boldsymbol{t}$ の第 i 成分を t_i としたとき、

$$\sum_{i=1}^{k} \frac{1 - Pr(X_i = t_i)}{Pr(X_i = t_i)} \tag{2.28}$$

と表現できる。2 値変量が独立でない場合 (2.26) 式右辺の第 2 項の表現は一般には簡単にならない。

2.3.3 加法性が成立している場合の異常検出

加法性が成立している場合のマハラノビス平方距離の分布が示されたので、検定の棄却域を定めることができる。自由度 m の χ^2 分布の分布関数を $F(\cdot)$、$\boldsymbol{X} = \boldsymbol{t}$ のときの補正項（(2.26) 式右辺の第 2 項）を $b_{\boldsymbol{t}}$ とおき、

$$\sum_{\boldsymbol{t}} p_{\boldsymbol{t}}^{(0)} F(K - b_{\boldsymbol{t}}) = 1 - \alpha \tag{2.29}$$

を満足する値 K を用いて、棄却域を $\Delta^2 > K$ と定めればよい。(2.29) 式からわかるように、補正項 $b_{\boldsymbol{t}}$ が大きいほど $F(K - b_{\boldsymbol{t}})$ が小さくなるので、$\boldsymbol{X} = \boldsymbol{t}$ のときの条件付誤報率 $1 - F(K - b_{\boldsymbol{t}})$ は大きくなる。

2.3.4 加法性の仮定が成立しない場合のマハラノビス平方距離の分布

ここでは、加法性の仮定が成立しない場合には、マハラノビス平方距離の分布がより複雑になることを示す。(2.25) 式で与えられる $\boldsymbol{Z}$ は、$\boldsymbol{X} = \boldsymbol{t}$ のとき $\boldsymbol{Z} = \boldsymbol{Y} - \mu_{\boldsymbol{t}}$ となるとは限らないことに注意する。(2.19) 式のマハラノビス平方距離は

$$\Delta^2 = \boldsymbol{Z}'H^{-1}\boldsymbol{Z} + (\boldsymbol{X} - A\boldsymbol{p})'(AEA')^{-1}(\boldsymbol{X} - A\boldsymbol{p}) \tag{2.30}$$

と表される。ここで、加法性の仮定が成立していないので、$H = \Sigma$ とは限らないが、$\boldsymbol{X}$ が与えられたとき、$\boldsymbol{Z}$ の分散共分散行列は Σ なので、$\Sigma^{-1/2}\boldsymbol{Z}$ の要素は独立に分散が 1 の正規分布に従って

いる。また、$H-\Sigma = MEM' - MEA'(AEA')^{-1}AEM'$ は非負定値対称行列なので、適当な行列 C を用いて

$$H = \Sigma + CC'$$

と表現できる。したがって、

$$\begin{aligned} H^{-1} &= \Sigma^{-1} - \Sigma^{-1}C(I + C'\Sigma^{-1}C)^{-1}C'\Sigma^{-1} \\ &= \Sigma^{-1/2}(I - \Sigma^{-1/2}C(I + C'\Sigma^{-1}C)^{-1}C'\Sigma^{-1/2})\Sigma^{-1/2} \end{aligned}$$

と表すと、2 次形式

$$\boldsymbol{Z}'H^{-1}\boldsymbol{Z} = (\Sigma^{-1/2}\boldsymbol{Z})'(I - \Sigma^{-1/2}C(I + C'\Sigma^{-1}C)^{-1}C'\Sigma^{-1/2})\Sigma^{-1/2}\boldsymbol{Z} \tag{2.31}$$

は、$\boldsymbol{X}$ が与えられたとき、平均が 0 とは限らない分散 1 の独立な正規変量の 2 乗の重み付き和として表現されることがわかる。重みは、行列 $(I - \Sigma^{-1/2}C(I + C'\Sigma^{-1}C)^{-1}\Sigma^{-1/2})$ の固有値で与えられる。つまり、非心 χ^2 変量の重み付き和として表現されるので、正確な有意水準を持つ異常検出ルールを構成するために棄却点を計算するには、より複雑な計算が必要になる。

2.4 尤度比検定による異常検出

ここでは、尤度比検定の考え方に基づき、設定された誤報率を正確に実現する異常検出方法を導出する。さらに、その方法の条件付誤報率についての基本的性質についても述べる。

2.4.1 尤度比検定による異常検出

データ $(\boldsymbol{y}', x)'$ が得られるとき、(2.7) を一般化した帰無仮説

$$H_0 : p_x = p_x^{(0)},\ \boldsymbol{\mu}_x = \boldsymbol{\mu}_x^{(0)}, \qquad x = 1, \ldots, I \tag{2.32}$$

を検定するための尤度比検定統計量を求める。基本的には、$(\boldsymbol{y}', x)'$ が確率・確率密度関数

$$p_x^{(0)}(2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\{-(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})/2\} \tag{2.33}$$

で与えられる母集団分布からの観測値であるという帰無仮説 H_0 の真偽を判断する問題である。帰無仮説 H_0 の下での尤度関数の最大値は、(2.33) 式で与えられる。一方、制約条件のないとき尤度関数が最大になるのは、$p_x = 1, \boldsymbol{\mu}_x = \boldsymbol{y}$ のときである。定数部分を省略すると、尤度比検定統計量 (対数尤度比の -2 倍) は、

$$(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)}) - 2\log(p_x^{(0)}) \tag{2.34}$$

となる。この式の第 1 項は、連続変量のみに基づくマハラノビス平方距離であり、第 2 項は離散変量の確率による補正項である。M法と全く類似の形の検定統計量が得られたが、M法では、補正項が $(1 - p_x^{(0)})/p_x^{(0)}$ であったことに注意する。(補正項の違いとその影響については第 3 章で議論する) この結果、尤度比統計量もM法の場合と同様に位置をずらした χ^2 分布の混合分布にしたがうことがわかる。

ここで、ロケーションモデルが成立しない場合すなわち、離散変量の値を与えたときの連続変量の分布が正規分布ではあるが分散共分散行列が異なる場合を考えてみる。$X = x$ のときの分散共分散行列を Σ_x で表すとき、(2.34) 式を修正して

$$(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})'\Sigma_x^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)}) - 2\log(p_x^{(0)}) \tag{2.35}$$

を考えれば、正常状態における確率分布は位置をずらした χ^2 分布の混合分布である。したがって、(2.35) 式を用いることから出発するならば、ロケーションモデルの仮定は不必要であり、本節の議論はそのまま成立することがわかる。

離散変量が独立な 2 つの変量 X_1 と X_2 を組み合わせたものである場合を考える。$X_1 = x$ の確率と $X_2 = y$ の確率をそれぞれ p_{x+}、p_{+y} で表せば $X_1 = x, X_2 = y$ となる確率は $p_{x+}p_{+y}$ なので、補正項は、両者の周辺確率による補正項の和として表現できることがわかる。

(2.34) 式で与えられる検定統計量は、分布間の距離からも導かれることを述べておく。確率・確率密度関数 $f(t)$ と $g(t)$ で表される 2 つの分布間の距離の表現のひとつである松下距離の 2 乗 $\int(\sqrt{f(t)} - \sqrt{g(t)})^2 dt$ は

$$\rho = \int \sqrt{f(t)g(t)}dt = \int \sqrt{g(t)/f(t)}f(t)dt$$

を用いて $2(1-\rho)$ と表される。いま、$(\boldsymbol{y}', x)'$ が (2.33) 式で与えられる確率分布からの観測値とみなせるかどうかを判断するために、$f(t)$ として (2.33) 式で与えられる帰無仮説の下での確率・確率密度関数、$g(t)$ として、得られたデータ $(\boldsymbol{y}', x)'$ に確率 1 を与える退化した分布を考える。そのとき、

$$-4\log\rho = (\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)}) - 2\log(p_x^{(0)})$$

となり、尤度比法と同じ異常検出法が得られる。松下距離に関する議論は、判別分析の枠組みで Krzanowski(1986) に与えられている。

なお、中西、加藤 (2008) は、分布間の距離としての λ ダイバージェンスについて議論している。分散共分散行列が異なるものとして、観測データ $X = x, \boldsymbol{Y} = \boldsymbol{y}$ が得られるとき、$-1/2 \leq \lambda < 0$ となる λ に対して

$$MTD^{\lambda} = \frac{1}{\lambda}(p_x^{(0)})^{(1+\lambda)}\exp\Big(\frac{\lambda(1+\lambda)}{2}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})'\Sigma_x^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)}) - 1\Big)$$

で距離を定義しているので、

$$\frac{2}{\lambda(1+\lambda)}\log(\lambda MTD^{\lambda}) = (\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)})'\Sigma_x^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x^{(0)}) + \frac{2}{\lambda}\log(p_x^{(0)})$$

となり、同種の統計量が得られる。

(2.34) 式で与えられた尤度比検定統計量を用いた有意水準 α の検定を構成するためには、棄却限界値を、正常状態でその値以下になる確率が α となるように定めればよい。(2.34) 式で与えられる量は、離散変量が与えられたときには平行移動した χ^2 分布にしたがうことから、I 個の平行移動した χ^2 分布の混合分布にしたがうことがわかる。したがって、棄却限界値は、自由度が m の χ^2 分布の累積分布関数を $F(\cdot)$ で表すとき、

$$\sum_{x=1}^{I} p_x^{(0)} F(K + 2\log(p_x^{(0)})) = 1 - \alpha \tag{2.36}$$

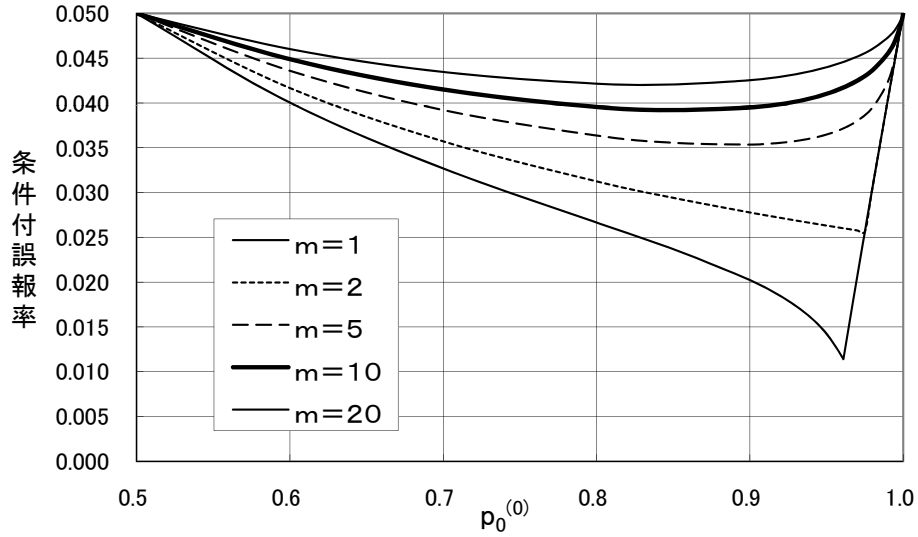

図 2.7: L 法の条件付誤報率 $[X = 0](\alpha = 0.05)$

を満足する K で与えられる。(2.36) 式の左辺は、K の単調な関数なので、(2.36) 式を満足する K の値は、χ^2 分布の分布関数を計算することができれば、反復法により容易に求められる。以上により定まった、尤度比検定統計量を用いる異常検出法を尤度比法とよび、 L法と表記する。また、

$$1 - F(K + 2\log(p_x^{(0)}))$$

は、$X = x$ のときに仮説を棄却する確率である。仮説が正しいときすなわち正常状態において、離散変量 X の値を与えたときの条件付誤報率である。

L 法における条件付誤報率に関しても、補正項の性質から、離散変量の確率が小さいほど条件付誤報率が大きくなることがわかる。

2.4.2 2 値変量の場合

離散変量 X の取る値を $0, 1$ の 2 値とし、$\frac{1}{2} \leq p_0^{(0)}$ とする。$\alpha_x(p_0^{(0)})$ を $X = x$ のときの条件付誤報率とすると、M法における性質 $2.1(i), (ii), (iii)$ と同様に次の性質が成立することも容易に確認できる。記号簡略化のため、$p_0^{(0)}$ を p と表わす。

性質 2.2

(i) $\alpha_0(1/2) = \alpha_1(1/2) = \alpha.$ $\quad \lim_{p\to 1} \alpha_0(p) = \alpha, \quad \lim_{p\to 1} \alpha_1(p) = 1$

(ii) $\alpha_0(p) < \alpha < \alpha_1(p), \qquad 1/2 < p < 1$

(iii) $\alpha_1(p)$ は $1/2 < p < 1$ の範囲で p の単調増加関数であり、$1 - \alpha < p^* < 1$ なる p^*が存在し、
$p^* \leq p \leq 1$ で$\alpha_1(p) = 1$ である。

ここで、2 値変量の場合の条件付誤報率のグラフを図 $2.7(X = 0)$ と図 $2.8(X = 1)$ に与える。補正項が異なるためM法と値は異なっているが、全体としてみると同じような挙動をしている。

なお、この章で紹介した各手法に関する条件付誤報率の比較は、第 3 章の検出力の比較と合わせて行う。

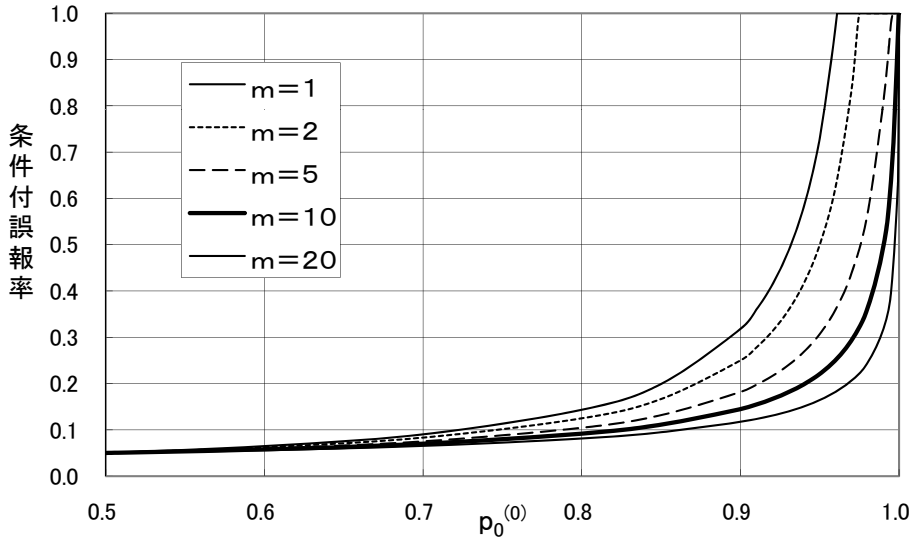

図 2.8: L 法の条件付誤報率 $[X = 1](\alpha = 0.05)$

2.5 まとめ

離散変量が混在する場合の異常検出問題をロケーションモデルを用いて定式化し、3 つの異常検出法として条件付法 (C 法)、マハラノビス距離法 (M 法)、尤度比法 (L 法) を、分布の母数が既知である場合について構成した。どの方法においても、異常検出統計量は連続変量によるマハラノビス平方距離と離散変量の分布で決まる補正項の和になることが確認された。この結果から、設定した誤報率を達成する異常検出法を与えた。

L 法と M 法では、離散変量の確率の小さい水準で条件付誤報率が大きくなることを示した。また、2 値変量の場合の数値計算から、その度合いは L 法に比べて M 法で顕著であることが確認できた。これらの結果を基に次章で検出力に関する評価・検討を行う。

第3章　母数が既知の場合の異常検出法の検出力

この章では、離散変量が一つでロケーションモデルが想定される状況において、分布の母数がすべて既知の場合における異常検出法の検出力の性質を考察し、3 つの手法を比較する。まず、多値変量における条件付誤報率の性質を整理する。その後、2 値変量における条件付誤報率の手法間での比較をおこなう。これらの性質は、検出力の性質ならびにその手法間での比較を行うのに必要である。

この議論を踏まえて、検出力の性質を明らかにし、手法の優劣を比較検討する。検出力については、初めに離散変量の分布のみが変化した場合について考察し、その後、連続変量の平均ベクトルも変化した場合について比較検討する。2 値変量の場合における誤報率および検出力については、数値計算により詳細な検討を行う。特に、平均ベクトルの変化量のマハラノビス平方距離が離散変量の各水準で等しい状況について 3 手法の優劣の比較を行う。

3.1　離散変量を与えたときの条件付誤報率

初めに、多値変量の場合について述べ、続いて 2 値変量の場合について説明する。なお、誤報率の設定値を α で表す。

3.1.1　多値変量の場合

離散変量の水準を $x = 1, \ldots, I$ とし、正常状態でのその確率 $p_x^{(0)}$ は $p_1^{(0)} \geq \cdots \geq p_I^{(0)}$ を満たすものとする。また、3 つの手法、条件付法 (C 法)、尤度比法 (L 法)、マハラノビス距離法 (M 法) を表す記号を A とし、各手法に対し C、L、M を対応させる。手法Aによる異常検出のための統計量において、離散変量の水準 x での補正項を $h_A(x)$ と表わすことにする。各手法に対し、$h_C(x) = 0$、$h_M(x) = (1 - p_x^{(0)})/p_x^{(0)}$、$h_L(x) = -2 \log p_x^{(0)}$ である。さらに、手法 A の $X = x$ のときの条件付誤報率を $G_A(x)$ で表わすことにする。

C 法は、離散変量の観測値を与えたときの条件付検定による方法なので、2.2 節で示したように離散変量の値によらず条件付誤報率は設定値 α に等しくなる。従って、全体での誤報率も α となる。

M 法と L 法では、$A = M, L$ として

$$h_A(1) \leq \cdots \leq h_A(I)$$

が成り立つので、条件付誤報率については

$$G_A(1) \leq \cdots \leq G_A(I) \tag{3.1}$$

が成り立つ。離散変量の両水準での確率が等しい場合にのみ等号が成立する。

このように、L 法や M 法では、確率の小さい水準が観測された場合は、誤って異常であると判定する確率が大きくなる性質を持っている。

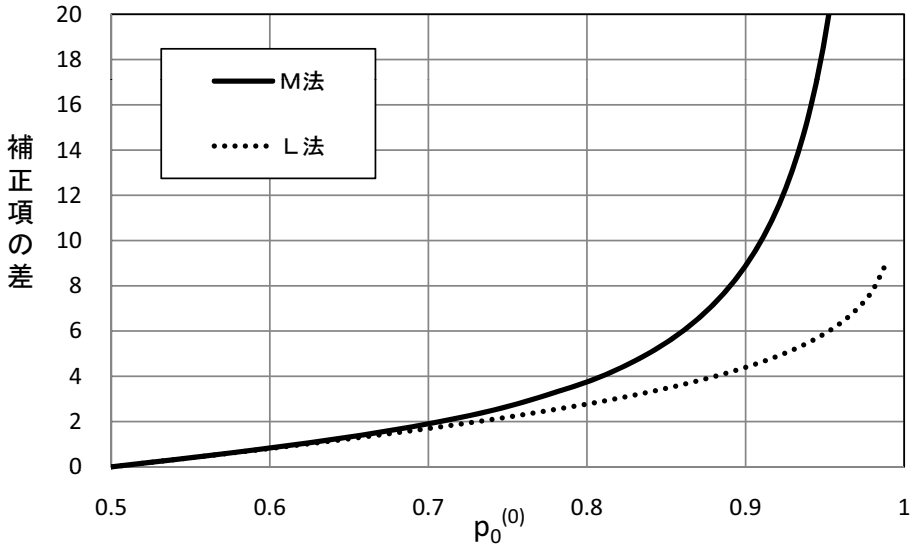

図 3.1: 補正項の差 (L 法と M 法)

3.1.2 2 値変量の場合

2 値変量なので、離散変量の水準は 0 と 1 とする。また、$p_0^{(0)} \geq p_1^{(0)}$ とする。

2 章で示した様に、$X = 0$ のときの条件付誤報率は設定値 α 以下で、$p_0^{(0)}$ が 0.5 から 1 に変化するとき、初めは減少し途中から上昇に転じ両端 (0.5 と 1) では α になること、$X = 1$ では設定値 α 以上で、$p_1^{(0)}$ が 0.5 から 0 に変化するにつれ単調に増加し最後は 1 になることがわかっている。ここでは、手法間での条件付誤報率を比較する。

実際に条件付誤報率の挙動に影響を与えているのは、補正項の値ではなく水準間での補正項の差である。2 値変量の場合の補正項の差は、M 法では $p_0^{(0)}/(1-p_0^{(0)}) - (1-p_0^{(0)})/p_0^{(0)}$、L 法では $2\log\{p_0^{(0)}/(1-p_0^{(0)})\}$ である。これらの補正項の差については、$p_0^{(0)} > 0.5$ のとき、

$$p_0^{(0)}/(1-p_0^{(0)}) - (1-p_0^{(0)})/p_0^{(0)} > 2\log\{p_0^{(0)}/(1-p_0^{(0)})\}$$

であり、M 法の補正項の差の方が L 法のそれより大きいことがわかる。$0.5 \leq p_0^{(0)} \leq 1$ での補正項の差を示したのが図 3.1 である。

この図から、$p_0^{(0)} = 0.5$ ではどちらも 0 であるが、$p_0^{(0)} > 0.5$ では M 法の方が大きい値をとり、その差は $p_0^{(0)}$ が大きくなるにつれ拡大していることがわかる。補正項の差の大小から、$X = 0$ では L 法が、$X = 1$ では M 法が条件付誤報率が大きいことがわかる。これに C 法を加えると、

$$G_M(0) \leq G_L(0) \leq G_C(0) = \alpha = G_C(1) \leq G_L(1) \leq G_M(1) \tag{3.2}$$

が成り立っていることがわかる。ただし、どの手法においても誤報率は設定値 α に一致していること、すなわち

$$p_0^{(0)} G_A(0) + p_1^{(0)} G_A(1) = \alpha, \quad A = C, L, M$$

であることに注意する。2 値変量の場合における L 法と M 法の $p_0^{(0)}$ を 0.5 から 1 まで変化させたときの $X = 0$ のときの条件付誤報率のグラフが図 3.2 である。

この図から、$p_0^{(0)}$ の値が $1-\alpha$ を越えたところでの $X = 0$ における条件付誤報率の下がり方は、L 法は M 法に比べてそれほど顕著でないことがわかる。また、図は省略したが、$X = 1$ では、M 法

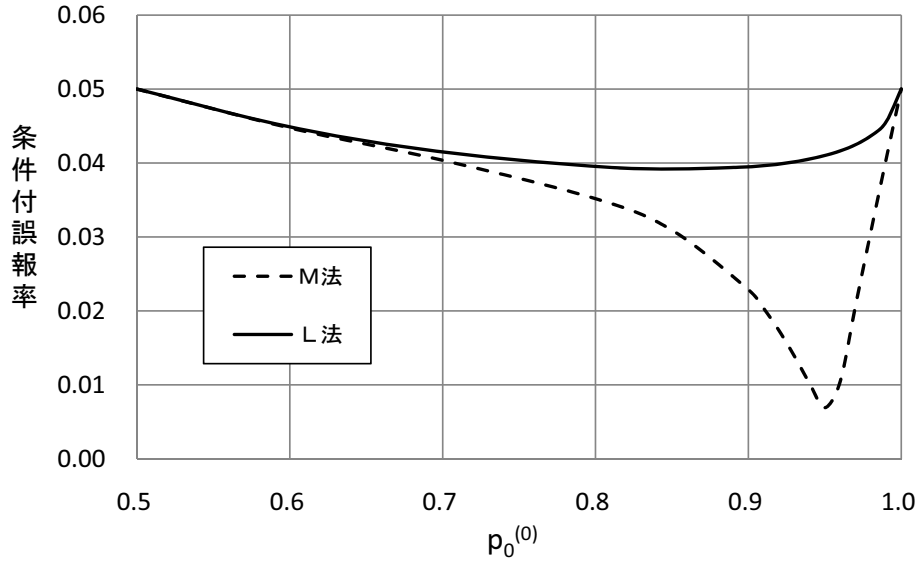

図 3.2: 条件付誤報率 $[X=0]$(L 法と M 法)

の方が 1 への近付き方が早いことが確認できた。これらの性質は、これから述べる各手法の検出力の挙動と密接にかかわっている。

3.2 離散変量の分布だけが変化したときの検出力

一般に異常状態をつぎのように表わすことにする。X の確率分布を定める母数ベクトル $\boldsymbol{p} = (p_1, \ldots, p_I)$ が $\boldsymbol{p}^{(0)} = (p_1^{(0)}, \ldots, p_I^{(0)})$ から $\boldsymbol{q} = (q_1, \ldots, q_I)$ に変化することと、$X = x$ のときの $\boldsymbol{Y}$ の平均ベクトル $\boldsymbol{\mu}_x$ が $\boldsymbol{\mu}_x^{(0)}$ から $\boldsymbol{\eta}_x, x = 1, \ldots, I$ に変化することとが考えられる。$\boldsymbol{\mu}_x^{(0)}$ から $\boldsymbol{\eta}_x$ への変化は、$(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})$ の分布に $\psi_x = (\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)})$ を通して影響するので、$\boldsymbol{\psi} = (\psi_1, \ldots, \psi_I)$ とおくと、一般に異常状態を $(\boldsymbol{q}, \boldsymbol{\psi})$ で表わすことができる。そのときの方法 A の検出力を $P_A(\boldsymbol{q}, \boldsymbol{\psi})$ と表わすことにする。

3.2.1 多値変量の場合

まず、M 法と L 法について $\boldsymbol{\psi} = \boldsymbol{0}$ の場合の検出力の挙動について考える。誤報率は $\sum_{x=1}^{I} p_x^{(0)} G_A(x)$ と表され、離散変量 X の分布の母数が $\boldsymbol{q} = (q_1, \ldots, q_I)$ へと変化したときの検出力は $\sum_{x=1}^{I} q_x G_A(x)$ で与えられる。不等式 (3.1) より $p_x^{(0)}$ の値が小さいほど $G_A(x)$ は大きいので、$p_x^{(0)}$ の値が相対的に小さい x に対し q_x の値が大きくなるとき検出力が誤報率よりも大きくなることが期待される。

より一般的な性質として議論するため、つぎの概念・記号を用意する。いま X の確率分布を定める 2 通りの母数ベクトル $\boldsymbol{q}_1 = (q_{11}, \ldots, q_{1I})$ と $\boldsymbol{q}_2 = (q_{21}, \ldots, q_{2I})$ に対し

$$\sum_{i=1}^{k} q_{1i} \leq \sum_{i=1}^{k} q_{2i}, \qquad k = 1, \ldots, I$$

が成立ち、かつ、ある k に対して厳密な不等号が成立つとき、$\boldsymbol{q}_1$ は $\boldsymbol{q}_2$ より確率的に大きいといい、$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$ と表わす。これは、集合 $\{1, 2, \ldots, I\}$ 上の離散分布として、$\boldsymbol{q}_1$ の方が $\boldsymbol{q}_2$ よりも大きい値が

出やすいことを述べている。$p_1^{(0)} = \cdots = p_I^{(0)} = 1/I$ のときは $G_A(x)$ が x の値と無関係な定数なので、この場合を除外するため $\boldsymbol{p}^{(0)}$ について $\mathbf{1}_I/I \succ \boldsymbol{p}^{(0)}$ という条件をおく。

性質 3.1 $\mathbf{1}_I/I \succ \boldsymbol{p}^{(0)}$ とする。$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$ ならば $P_A(\boldsymbol{q}_1, \mathbf{0}) > P_A(\boldsymbol{q}_2, \mathbf{0})$ である。

（証明）3.1.1 で述べたように、$G_A(x)$ は x の非減少関数であるので、 $G_A(X)$ は X の確率分布が $\boldsymbol{q}_1$ で与えられる場合の方が $\boldsymbol{q}_2$ の場合よりも確率的に大きくなる。したがって、その期待値である検出力についても $\boldsymbol{q}_1$ の場合の方が大きくなる。

□

性質 3.1 から、$\boldsymbol{q} \succ \boldsymbol{p}^{(0)}$ ならば、検出力は誤報率よりも大きくなることがわかる。逆に $\boldsymbol{p}^{(0)} \succ \boldsymbol{q}$ ならば、検出力は誤報率よりも小さくなってしまうこともわかる。

3.2.2 2 値変量の場合

まず、2 値変量の分布つまり $X = 0$ となる確率 p_0 が $p_0^{(0)}$ から q_0 に変化した場合の検出力について考える。各方法の条件付誤報率は $G_A(x)$ で与えられるが、連続変量の分布は変化しないので、2 値変量 X の値を与えたときの条件付検出力は条件付誤報率に一致する。$0.5 < p_0^{(0)} < 1$ であるので、

$$G_M(0) \le G_L(0) \le G_C(0) = \alpha = G_C(1) \le G_L(1) \le G_M(1)$$

であり、

$$p_0^{(0)} G_A(0) + (1 - p_0^{(0)}) G_A(1) = \alpha, \quad A = L, M, C$$

が成立つ。したがって、検出力　$q_0 G_A(0) + (1 - q_0) G_A(1)$　は、$q_0 > p_0^{(0)}$ のときは大きい順に C 法＞ L 法＞M法であり、$q_0 < p_0^{(0)}$ のときは逆に、M法＞ L 法＞ C 法の順となる。

3.3 連続変量の平均も変化した場合の検出力

ここでは、連続変量の平均ベクトルも変化した場合の検出力について考察する。多値の場合の一般論に続いて、2 値の場合について手法間の比較・検討を数値計算の結果を基に行う。

3.3.1 多値変量の場合

$X = x$ における連続変量 $\boldsymbol{Y}$ の平均が $\boldsymbol{\mu}_x^{(0)}$ から $\boldsymbol{\eta}_x$ に変化したとき、$X = x$ を与えたときの $\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)}$ の分布は、平均 $\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)}$、分散共分散行列が Σ の正規分布であるので、$(\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})' \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_x^{(0)})$ は、自由度 m、非心度 $(\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)})' \Sigma^{-1} (\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)})$ の非心 χ^2 分布にしたがう。

ここで、手法 A について $X = x$ を与えたときの非心度 ψ のときの条件付検出力を、前節の関数 G と類似の記号を用いて、$G_A(x; \psi)$ と表現する。前節の $G_A(x)$ は非心度を 0 とした $G_A(x; 0)$ に相当する。

自由度 m、非心度 ψ の非心 χ^2 分布の分布関数を $F(\cdot;\psi)$ で表わせば、$X=x$ を与えたときの条件付検出力 $G_A(x;\psi_x)$ は、

$$G_A(x;\psi_x) = 1 - F\left[K_A - h_A(x);\psi_x\right], \quad A = M, L, C \tag{3.3}$$

で与えられる。ここで、K_A は、第 2 章で求めた手法 A での棄却限界値であり、$\boldsymbol{p}^{(0)}$ に依存するが、省略している。

さらに異常状態で X の分布の母数を $\boldsymbol{q}$ とすれば、検出力は

$$P_A(\boldsymbol{q},\boldsymbol{\psi}) = \sum_{x=1}^{I} q_x G_A(x;\psi_x), \quad A = M, L, C \tag{3.4}$$

と表現される。

$\boldsymbol{\eta}_x, x = 1,\ldots,I$ が変化したときの検出力の挙動について述べることができる一つの性質は、ψ_x についての単調増加性である。一般に、非心度が大きくなるとき、非心 χ^2 分布は確率的に大きくなる。したがって、$\boldsymbol{q}_1 = \boldsymbol{q}_2 = \boldsymbol{q}$ であり、$\boldsymbol{\psi}_1 = (\psi_{11},\ldots,\psi_{1I})$, $\boldsymbol{\psi}_2 = (\psi_{21},\ldots,\psi_{2I})$ について $\psi_{1x} \geq \psi_{2x}, x = 1,\ldots,I,$ が成立つとき、$P_A(\boldsymbol{q},\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q},\boldsymbol{\psi}_2)$ となることがわかる。

M 法と L 法においては、各水準の確率の大小から補正項の大小が決まるので、各水準における非心度が同じ場合は、条件付検出力についても条件付誤報率と同様

$$G_A(1;\psi) \leq \cdots \leq G_A(I;\psi), \quad A = L, M \tag{3.5}$$

が成立する。C 法の場合は、この式の不等号が全て等号に置き換わるので、どの方法でも条件付検出力は x の単調非減少関数である。これらの性質から、以下の性質が成り立つ。

性質 3.2 2 つの異常状態 $(\boldsymbol{q}_1,\boldsymbol{\psi}_1)$ と $(\boldsymbol{q}_2,\boldsymbol{\psi}_2)$ について、$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$, $\boldsymbol{\psi}_1 \geq \boldsymbol{\psi}_2$ $(\psi_{1x} \geq \psi_{2x}, x = 1,\ldots,I)$, かつ $\psi_{21} \leq \cdots \leq \psi_{2I}$ であれば、$P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q}_2,\boldsymbol{\psi}_2)$ である。

(証明) X の 2 水準 $x < y$ について、$\psi_{2x} \leq \psi_{2y}$ より

$$G_A(x;\psi_{2x}) \leq G_A(x;\psi_{2y}) \leq G_A(y;\psi_{2y})$$

なので、$G_A(x;\psi_{2x})$ は x の単調非減少関数であり、$\boldsymbol{\psi}_2$ に関して

$$G_A(1;\psi_{21}) \leq \cdots \leq G_A(I;\psi_{2I}) \tag{3.6}$$

が成立する。したがって、$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$ より

$$P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_2) \geq P_A(\boldsymbol{q}_2,\boldsymbol{\psi}_2) \tag{3.7}$$

が成り立つ。さらに、$\boldsymbol{\psi}_1 \geq \boldsymbol{\psi}_2$ より

$$P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_2)$$

が成り立つので、

$$P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q}_2,\boldsymbol{\psi}_2) \tag{3.8}$$

が成立する。

□

一つの典型的で重要と考えられるのは、ψ_x が x の値によらず同程度の大きさの場合である。例えば、海外発の新型インフルエンザ感染において、離散変量として海外渡航歴の有無を取り上げた場合、海外渡航歴の有無により感染確率が変動するが、感染したときの病状は渡航歴の有無にはよらないと考えられる。これについては、この章の終りで 2 値の場合について数値計算で検出力の比較を行う。

3.3.2 2 値変量の場合

2 値変量の水準は 0 と 1 とし、その確率について $p_0^{(0)} > 1/2$ とする。誤報率と検出力は、正常状態における平均ベクトル $\boldsymbol{\mu}_0^{(0)}$ および $\boldsymbol{\mu}_1^{(0)}$ の値にはよらないことから、一般性を失うことなく $\boldsymbol{\mu}_0^{(0)} = \boldsymbol{\mu}_1^{(0)} = \mathbf{0}$ とする。また、異常状態において、$X = x$ のときの Y の平均ベクトルを $\boldsymbol{\eta}_x$ とおく。

3.3.2.1 条件付検出力の性質

ここでは、各手法の条件付検出力の大小についての性質を明らかにする。どの手法でも、異常検出統計量は連続変量のみに基づくマハラノビス平方距離と補正項の和として与えられる。方法Aにおいて有意水準 α を実現する棄却限界値を K_A とする。K_A は $p_0^{(0)}$ の関数であるが、簡単のため確率 $p_0^{(0)}$ は省略した。したがって、連続変量のみに基づくマハラノビス平方距離の棄却限界値は、$K_A - h_A(x)$ である。条件付誤報率の大小関係を示す (3.2) 式から、$X = 0$ では

$$K_C - h_C(0) \le K_L - h_L(0) \le K_M - h_M(0)$$

$X = 1$ では

$$K_M - h_M(1) \le K_L - h_L(1) \le K_C - h_C(1)$$

である。

この結果、異常状態で平均が変わったときの条件付検出力についても、条件付誤報率と同様の大小関係が成り立つ。つまり、$X = 0$ では、条件付検出力の大きい順に、C 法、L 法、M 法であり、$X = 1$ では逆に M 法、L 法、C 法の順となる。

3.3.2.2 M 法の検出力

異常検出に用いるマハラノビス平方距離 Δ^2 の表現 (2.6) 式の第 1 項は自由度 m、非心度 $\psi_x = (\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)})'\Sigma^{-1}(\boldsymbol{\eta}_x - \boldsymbol{\mu}_x^{(0)})$ の非心 χ^2 分布に従う。したがって、正常状態で $X = 0$ となる確率 $p_0^{(0)}$ と異常状態で $X = x$ における非心度 ψ_x を指定すれば、2 値変量の値を与えたときの条件付検出力が定まる。自由度 m、非心度 ψ の非心 χ^2 分布の分布関数を $F(\bullet; m, \psi)$ とすると、(2.10) 式の解 K として定まる棄却限界値 K_M を用いるとき、検出力は以下の式で与えられる。

$$1 - \{q_0 F(K_M - (1 - p_0^{(0)})/p_0^{(0)}; m, \psi_0) + (1 - q_0) F(K_M - p_0^{(0)}/(1 - p_0^{(0)}); m, \psi_1)\}, \tag{3.9}$$

このとき

$$1 - F(K_M - (1 - p_x^{(0)})/p_x^{(0)}; m, \psi_x), \quad x = 0, 1 \tag{3.10}$$

表 3.1: 非心度による条件付検出力の変化 (M 法) $[X=0](m=10,\alpha=0.05)$

	非心度							
$p_0^{(0)}$	0	2	5	10	15	20	25	30
0.50	0.050	0.121	0.268	0.542	0.760	0.891	0.955	0.984
0.60	0.045	0.111	0.252	0.523	0.745	0.882	0.951	0.982
0.70	0.040	0.102	0.238	0.506	0.731	0.873	0.947	0.980
0.80	0.035	0.092	0.220	0.483	0.712	0.861	0.940	0.977
0.90	0.023	0.066	0.172	0.416	0.653	0.821	0.918	0.966
0.94	0.010	0.035	0.107	0.308	0.543	0.738	0.867	0.939
0.95	0.007	0.026	0.084	0.264	0.491	0.695	0.838	0.922
0.9538	0.006	0.024	0.080	0.255	0.481	0.685	0.831	0.919
0.96	0.010	0.035	0.108	0.310	0.545	0.740	0.868	0.940
1.00	0.050	0.121	0.268	0.542	0.760	0.891	0.955	0.984

は $X=x$ のときの条件付検出力である。(3.9) 式からわかるように、検出力は条件付検出力を確率 q_x で重み付けした和で表されている。また、(3.10) 式より、2 値変量が与えられたときの条件付検出力は、非心 χ^2 分布に従う変量が $K_M-h_M(x)$ を超える確率である。非心度が 0 のときが条件付誤報率なので、条件付誤報率が小さい (大きい) ということは $K_M-h_M(x)$ の値が大きい (小さい) ことに対応し、したがって条件付検出力も小さく (大きく) なることに注意する。$p_0^{(0)}$ と ψ_x の値による条件付検出力の変化の様子を、$m=10,\alpha=0.05$ の場合について示したのが表 3.1 と表 3.2 で、この中の一部を図示したのが図 3.3 と図 3.4 である。

まず、$p_0^{(0)}=0.9$ のときの X の値による条件付検出力の違いを調べる。$\psi_0=\psi_1$ の場合は、$X=0$ の場合に比べて $X=1$ の場合の方が検出力が大きいことがわかる。また、$X=1$ の場合非心度が小さいうちに検出力が大きく増加しているが、$X=0$ の場合非心度がある程度大きいところで急激に増加している。この傾向は、すべての $p_0^{(0)}$ の値についてあてはまる。

つぎに、$p_0^{(0)}$ の値による条件付検出力の変化を調べる。$X=0$ のときの条件付検出力は、非心度 ψ_0 の値が一定のとき、2 節で計算した条件付誤報率と同様、$p_0^{(0)}$ が 0.5 から 1 へと変化するのに伴い、はじめは減少し $p_0^{(0)}$ の値が 1 の近くで増加に転じる。また、$p_0^{(0)}$ が 0.95 の付近で条件付検出力が最も低くなっていることが確認できる。なお、この表の中にある $p_0^{(0)}=0.9538$ は、$m=10$ のとき $X=0$ における条件付誤報率が最小、すなわち $K_M-h_M(0)$ が最大となり、条件付検出力が最小となる点である。一方、$X=1$ における条件付検出力は、性質 $2.1(iii)$ より条件付誤報率が $p_0^{(0)}$ に関して単調に増加することからわかるように、$p_0^{(0)}$ の増加に伴い増加していることがわかる。

表 3.1,3.2 から M 法における検出力は以下のようにして求められる。まず、正常状態における 2 値変量の分布を表す $p_0^{(0)}$ の値を定める。指定する非心度における条件付検出力の値を、$X=0$ と $X=1$ の双方について表から求める。求めた 2 つの値を、異常状態における 2 値変量の確率で重み付平均をとった値が求める検出力である。

m が 10 の場合、$X=0$ のときの条件付誤報率が最も小さくなるのは、$p_0^{(0)}=0.9538$ のときであっ

表 3.2: 非心度による条件付検出力の変化 (M 法) $[X=1](m=10,\alpha=0.05)$

	非心度							
$p_0^{(0)}$	0	2	5	10	15	20	25	30
0.50	0.050	0.121	0.268	0.542	0.760	0.891	0.955	0.984
0.60	0.058	0.135	0.290	0.568	0.779	0.902	0.961	0.986
0.70	0.072	0.160	0.328	0.608	0.808	0.918	0.969	0.989
0.80	0.109	0.219	0.406	0.684	0.858	0.944	0.980	0.994
0.90	0.294	0.452	0.653	0.862	0.953	0.985	0.996	0.999
0.94	0.673	0.795	0.903	0.975	0.994	0.999	1.000	1.000
0.95	0.869	0.928	0.972	0.994	0.999	1.000	1.000	1.000
0.9538	0.951	0.976	0.992	0.999	1.000	1.000	1.000	1.000
0.96	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

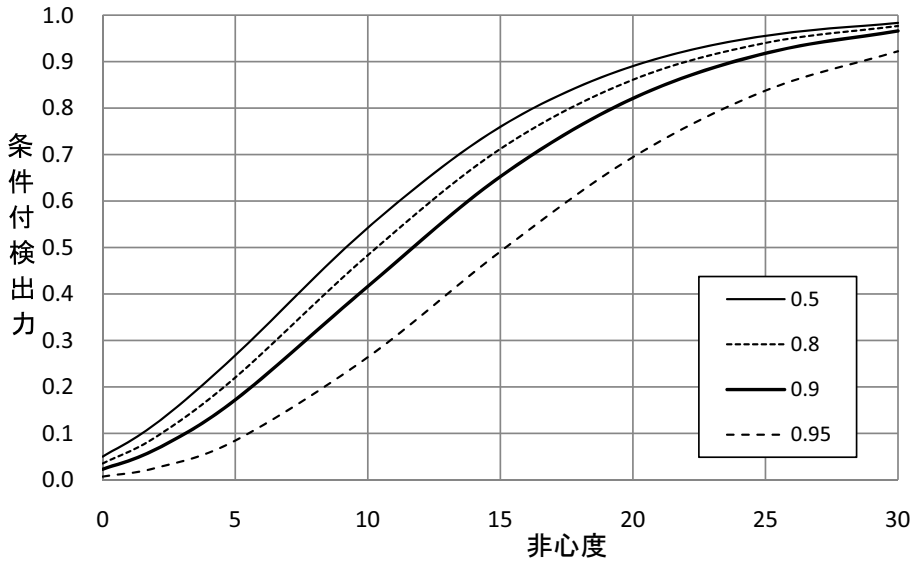

図 3.3: M 法の条件付検出力 $[X=0](m=10,\alpha=0.05)$ (凡例は $p_0^{(0)}$ の値)

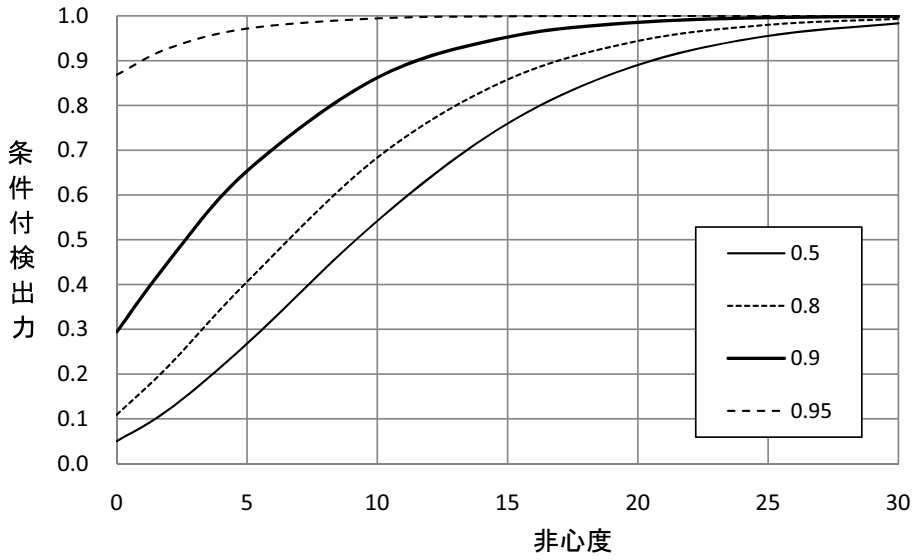

図 3.4: M 法の条件付検出力 $[X=1](m=10, \alpha=0.05)$ (凡例は $p_0^{(0)}$ の値)

た。補正項による位置のズレが大きく、$X=1$ の場合だけで棄却する確率が $\alpha=0.05$ に近くなり、$X=0$ で棄却されにくくなっているためである。この場合の条件付検出力ならびに $q_0=0.6$ での検出力を示したのが図 3.5 である。M 法の検出力は、C 法の検出力（中央の破線）と比べて、$X=1$（一番上の実線）では高く、$X=0$（一番下の実線）では低いことがわかる。しかし、$q_0=0.6$（中央の実線）では、非心度が 20 付近で多少検出力が低くなってはいるが、非心度が 10 以下では検出力が高く、全体として遜色ない結果を示している。したがって、この場合でも M 法を用いることに大きな問題はないと考える。

3.3.2.3 L 法の検出力

L 法においても、平均が変化した場合の異常検出統計量は、2 値変量 X の値が与えられたとき、位置をずらした非心 χ^2 分布にしたがう。したがって、条件付検出力は非心 χ^2 分布の分布関数を計算することで求めることができる。そのようにして求めた結果が表 3.3 と 3.4 である。連続変量の数 $m=10$、誤報率の設定値 $\alpha=0.05$ とした。

$X=1$ の場合の方が $X=0$ の場合に比べて $p_0^{(0)}$ が 1 に近づいたときの変化が大きいので、その部分をより詳細に示した。$X=1$ の場合について、検出力関数の様子をグラフにしたのが図 3.6 である。

$X=0$ の場合、条件付検出力を $p_0^{(0)}$ の関数として見ると条件付誤報率と同様、一度減少したあと増加に転じる ($m=10$ では $p_0^{(0)}=0.85$ で増加に転じる) が、表 3.3 からわかるように、その変化は非常に小さい。$X=1$ の場合も、条件付誤報率と同様、$p_0^{(0)}$ が増加するにつれて条件付検出力が増加 (グラフが上の方に移動）する。とくに $p_0^{(0)}$ の値が 1 の近くでは急激に増加し、最終的にはすべてが 1 に近づくことがわかる。

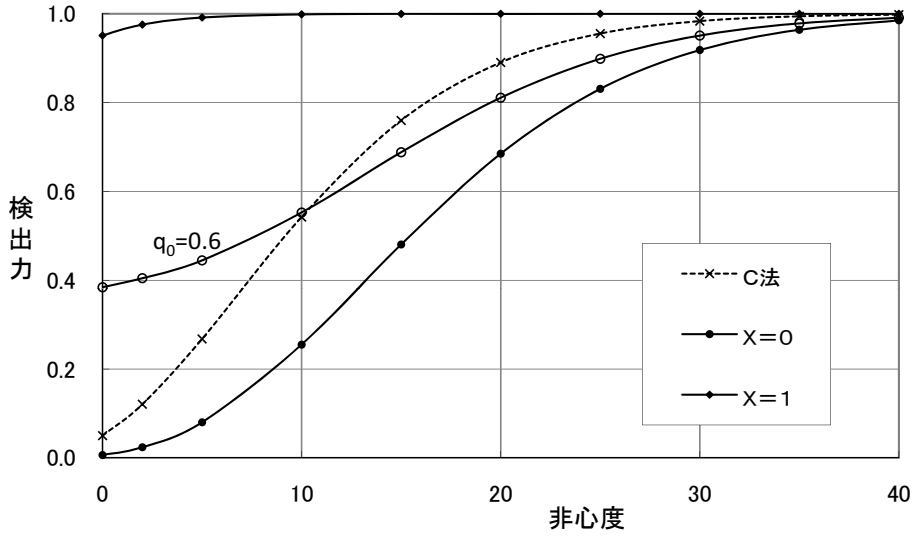

図 3.5: $p_0^{(0)} = 0.9538$ のときの検出力 [M 法、$q_0 = 0.6$] $(m = 10, \alpha = 0.05)$

表 3.3: 非心度による条件付検出力の変化 (L 法) $[X = 0](m = 10, \alpha = 0.05)$

	非心度							
$p_0^{(0)}$	0	2	5	10	15	20	25	30
0.5	0.050	0.121	0.268	0.542	0.760	0.891	0.955	0.984
0.6	0.045	0.111	0.252	0.524	0.745	0.882	0.951	0.982
0.7	0.041	0.105	0.242	0.511	0.735	0.875	0.948	0.980
0.8	0.040	0.101	0.235	0.503	0.728	0.871	0.946	0.979
0.85	0.039	0.100	0.234	0.501	0.727	0.870	0.945	0.979
0.9	0.039	0.101	0.235	0.502	0.728	0.871	0.946	0.979
0.95	0.041	0.104	0.240	0.509	0.733	0.874	0.947	0.980
0.99	0.046	0.112	0.254	0.526	0.747	0.883	0.952	0.982

表 3.4: 非心度による条件付検出力の変化 (L 法) $[X = 1](m = 10, \alpha = 0.05)$

	非心度							
$p_0^{(0)}$	0	2	5	10	15	20	25	30
0.5	0.050	0.121	0.268	0.542	0.760	0.891	0.955	0.984
0.6	0.058	0.135	0.290	0.567	0.779	0.902	0.961	0.986
0.7	0.070	0.156	0.321	0.601	0.803	0.916	0.968	0.989
0.8	0.092	0.192	0.371	0.651	0.837	0.934	0.976	0.992
0.85	0.111	0.222	0.410	0.687	0.860	0.945	0.981	0.994
0.9	0.145	0.270	0.468	0.736	0.889	0.959	0.986	0.996
0.91	0.155	0.284	0.484	0.748	0.896	0.962	0.988	0.996
0.92	0.166	0.299	0.501	0.761	0.903	0.965	0.989	0.997
0.93	0.181	0.318	0.522	0.777	0.911	0.969	0.990	0.997
0.94	0.198	0.340	0.545	0.793	0.920	0.973	0.991	0.998
0.95	0.221	0.368	0.574	0.813	0.930	0.977	0.993	0.998
0.96	0.251	0.403	0.608	0.835	0.940	0.981	0.994	0.998
0.97	0.293	0.451	0.653	0.862	0.952	0.985	0.996	0.999
0.98	0.361	0.522	0.713	0.895	0.966	0.990	0.997	0.999
0.99	0.493	0.647	0.809	0.939	0.983	0.996	0.999	1.000
0.995	0.637	0.767	0.886	0.969	0.993	0.998	1.000	1.000
0.999	0.919	0.958	0.985	0.997	1.000	1.000	1.000	1.000

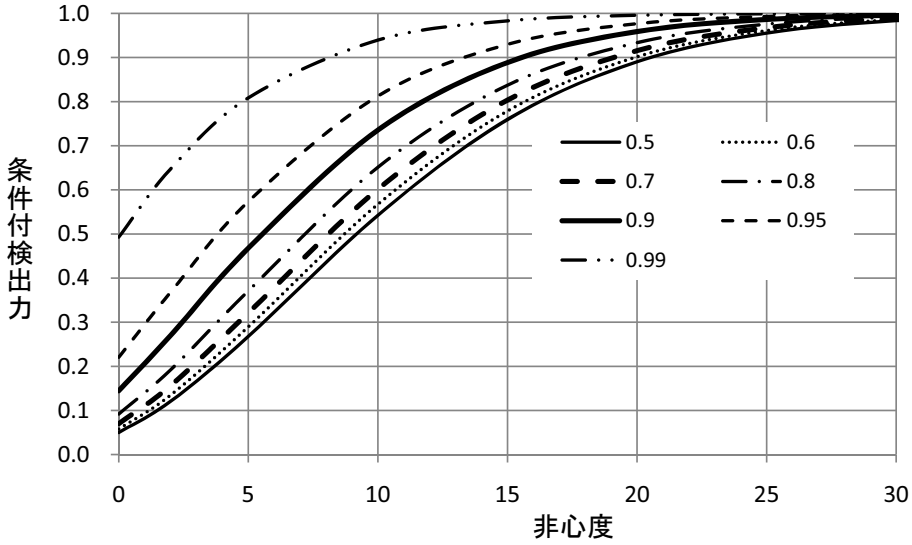

図 3.6: L 法の条件付検出力 $[X = 1](m = 10, \alpha = 0.05)$(凡例は $p_0^{(0)}$ の値)

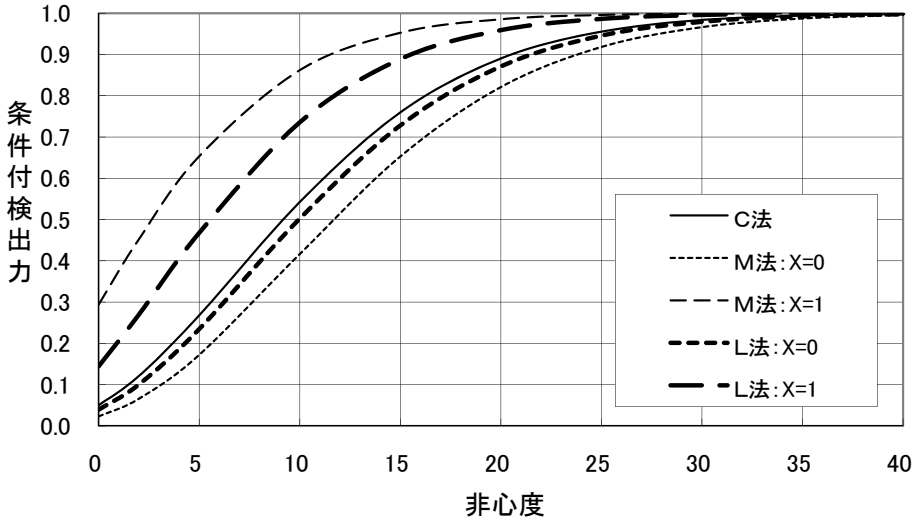

図 3.7: 条件付検出力の比較 ($m = 10, \alpha = 0.05, p_0^{(0)} = 0.9$)

3.3.2.4 検出力による 3 手法の比較

3 つの方法の条件付検出力を、$p_0^{(0)} = 0.9$ の場合について比較したのが図 3.7 である。条件付検出力は、上の方から M 法 ($X = 1$)、L法 ($X = 1$)、C法 ($X = 0, 1$ とも同じ)、L法 ($X = 0$)、M 法 ($X = 0$) である。L 法、M法どちらにおいても、C 法に比較して条件付検出力の $X = 0$ における減少度に比べて、$X = 1$ における上昇度が大きいことがわかる。この図から、2 値変量が混在した場合の L 法および M 法による異常検出は、$X = 0$ つまり正常状態で頻度が大きい値においては C 法（M 法と L 法における $p_0^{(0)} = 0.5$ の場合と同じ挙動をする）とほぼ同等の条件付検出力を維持しながら、$X = 1$ つまり正常状態で頻度が小さい 2 値変量の値における条件付検出力を大きくできる手法であることが確認できる。

全体での検出力は、異常状態における 2 値変量の分布および、$X = 0$ のときおよび $X = 1$ のときの非心度により変化するので、どの手法が良いかについて簡単に結論づけられない。また、正常状態における 2 値変量の分布 $p_0^{(0)}$ もこれに影響することに注意する。

ここでは 1 つの目安を得るために、$X = 0$ と $X = 1$ における非心度が等しい場合について検出力の比較を行う。この場合、q_0 の値が 1 に近ければ C 法がよく、0 に近ければ M 法がよい。したがって、異常状態での q_0 の値によりその優劣が入れ替わる。その境界の q_0 の値を境界確率と呼ぶことにする。L 法と C 法および L 法とM法との境界確率を表 3.5 および表 3.6 に与えた。L 法とM法の境界確率を示したのが図 3.8 である。

非心度が 0 の時は、すべての境界確率は $p_0^{(0)}$ に一致する。非心度が増加するにつれ、境界確率は減少する。L 法と C 法の境界確率の場合、その下がり方は緩やかで、異常状態において q_1 がある程度増加する場合は、L 法が優位であると考えられる。L 法とM法との境界確率は、$p_0^{(0)}$ が 0.9 までは非心度に関する緩やかな減少関数であるが、0.9 を越えてからは、非心度 10 あたりで急激に減少するカーブを描く。とくに $p_0^{(0)} = 0.95$ では、その下がり方が極端である。これは、$p_0^{(0)} = 0.95$ 付近でM法の $X = 0$ における条件付誤報率が 0 に近い値をとることに起因している。

一般に、q_0 の値が 1 に近いときは C 法、0 に近いときはM法が良く、その中間に L 法が優位である領域が存在する。$p_0^{(0)} = 0.9$ のとき、それぞれの手法が検出力の意味で最も優位となる領域を示したのが図 3.9 である。

表 3.5: L 法と C 法の境界確率 $(m = 10, \alpha = 0.05)$

	非心度									
$p_0^{(0)}$	0	2	5	10	15	20	25	30	35	40
0.6	0.600	0.592	0.584	0.574	0.566	0.559	0.553	0.548	0.543	0.538
0.7	0.700	0.685	0.670	0.650	0.635	0.621	0.609	0.599	0.588	0.579
0.8	0.800	0.781	0.760	0.733	0.710	0.691	0.673	0.657	0.642	0.628
0.85	0.850	0.830	0.808	0.778	0.753	0.731	0.711	0.692	0.675	0.658
0.9	0.900	0.881	0.859	0.828	0.802	0.778	0.756	0.736	0.717	0.699
0.91	0.910	0.892	0.870	0.839	0.813	0.789	0.767	0.746	0.727	0.709
0.92	0.920	0.902	0.881	0.851	0.825	0.801	0.778	0.758	0.738	0.719
0.93	0.930	0.913	0.892	0.863	0.837	0.813	0.791	0.770	0.750	0.732
0.94	0.940	0.924	0.904	0.876	0.850	0.826	0.804	0.784	0.764	0.745
0.95	0.950	0.935	0.917	0.889	0.864	0.841	0.819	0.799	0.780	0.761
0.96	0.960	0.947	0.930	0.904	0.880	0.858	0.837	0.817	0.798	0.781
0.97	0.970	0.959	0.943	0.920	0.898	0.877	0.857	0.839	0.821	0.804
0.98	0.980	0.971	0.958	0.938	0.919	0.901	0.883	0.867	0.851	0.836
0.99	0.990	0.984	0.976	0.961	0.947	0.933	0.920	0.907	0.895	0.884

表 3.6: L 法とM法の境界確率 $(m = 10, \alpha = 0.05)$

	非心度									
$p_0^{(0)}$	0	2	5	10	15	20	25	30	35	40
0.6	0.600	0.584	0.568	0.547	0.531	0.517	0.504	0.493	0.483	0.473
0.7	0.700	0.669	0.635	0.592	0.556	0.526	0.498	0.473	0.451	0.430
0.8	0.800	0.755	0.701	0.626	0.562	0.506	0.456	0.411	0.371	0.336
0.85	0.850	0.798	0.730	0.629	0.540	0.462	0.394	0.336	0.286	0.244
0.9	0.900	0.840	0.747	0.595	0.459	0.347	0.261	0.196	0.147	0.112
0.91	0.910	0.848	0.747	0.578	0.427	0.309	0.222	0.159	0.115	0.084
0.92	0.920	0.855	0.745	0.554	0.388	0.264	0.179	0.122	0.084	0.059
0.93	0.930	0.863	0.740	0.521	0.339	0.214	0.135	0.087	0.057	0.038
0.94	0.940	0.870	0.731	0.477	0.282	0.162	0.094	0.056	0.034	0.022
0.95	0.950	0.878	0.719	0.426	0.223	0.115	0.061	0.033	0.019	0.011
0.96	0.960	0.896	0.745	0.451	0.239	0.124	0.066	0.037	0.021	0.013
0.97	0.970	0.923	0.808	0.550	0.323	0.180	0.101	0.058	0.035	0.021
0.98	0.980	0.948	0.864	0.644	0.409	0.237	0.135	0.078	0.046	0.028
0.99	0.990	0.973	0.921	0.751	0.512	0.302	0.167	0.091	0.051	0.029

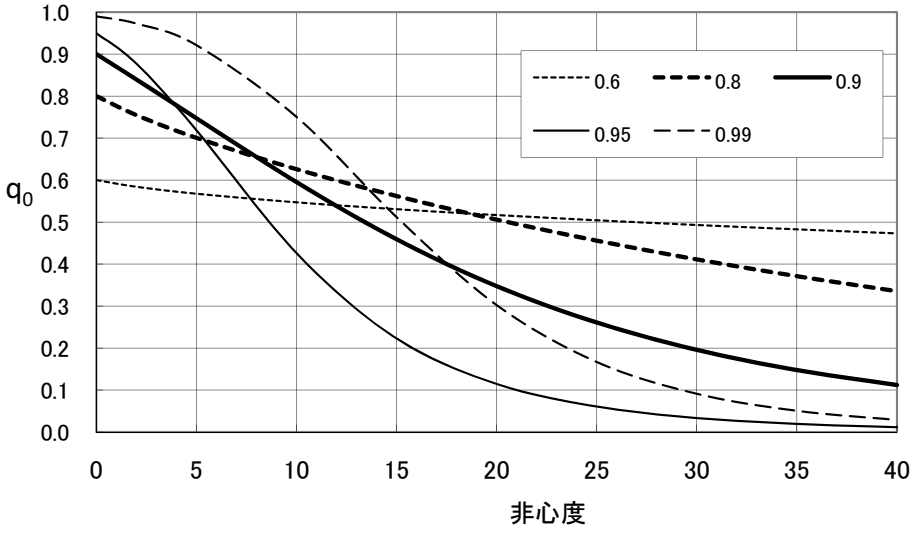

図 3.8: L 法とM法の境界確率 $(m = 10, \alpha = 0.05)$ (凡例は $p_0^{(0)}$ の値)

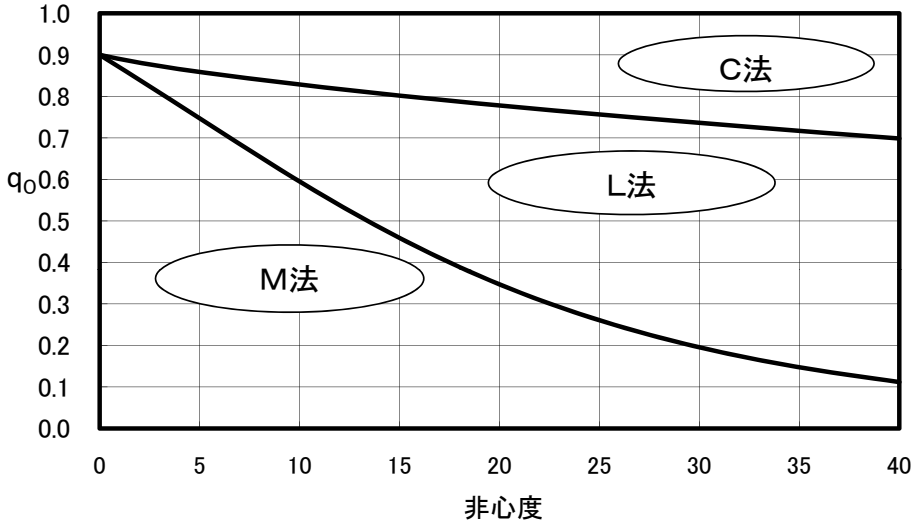

図 3.9: 各方法が最適な領域 $(m = 10, \alpha = 0.05, p_0^{(0)} = 0.9)$

図 3.9 において、C 法が優位な領域は上側に、M法が優位な領域は下側に、L 法が優位な領域はその中間に存在する。非心度が 0 から増加するにつれて、L 法が優位である領域が増加しているのが確認できる。非心度が 20 以上では、条件付検出力が 3 方法とも 0.87 以上なので、大きな差はないと考えられる。非心度 10 では、L 法とM法の境界確率は 0.6 程度である。この値と正常状態での値 $p_0^{(0)} = 0.9$ から、オッズ比を計算すると $(0.4/0.6)/(0.1/0.9) = 6$ であり、これ以内の変化ならば L 法が優位となる。

このように、異常な状態では $X = 1$ の確率がある程度上昇（p_0 の値が減少）し、非心度がある程度大きくなるような状況では、L 法が 3 つの手法の中で検出力の高い異常検出法であることがわかった。

したがって、2 値変量 X として、そのような性質を持つものを取り上げることができれば、非心度もある程度大きくなる状況では、L 法は優れた異常検出法であると考えられる。

3.3.3 計算例

兼高 (1987) による肝臓診断のデータの一部を用いて、3 つの方法による異常検出について例示・検証する。このデータは、正常人 200 名、肝疾患 17 名の、年齢・性別を含む 17 変量からなるデータで、Woodall, et al.(2003) の論文でも用いられている。ただし、肝疾患の 17 名の一部に重複があるので、異常データは 16 名として計算した。

性別 (X) が唯一の離散変量であり、男性 70 名、女性 130 名である。ロケーションモデルの仮定と矛盾しないように、男女間で分散があまり変わらない変量の中から、平均が適度に異なる連続変量を候補とし、さらに、異常判定にある程度寄与するものとして、Y_1(総タンパク)、Y_2(ロイシンアミノペプチダーゼ)、Y_3(総コレステロール) を取り上げた。この 3 変量間の相関はそれほど大きくなく、男女間での違いもあまりないと判断された。実際、分散共分散行列の同等性の尤度比検定 (例えば、竹村 (1991)p94) を行ったところ、検定統計量の値は 8.43 になり、自由度は 6 なので、p 値は約 0.21 であった。分散共分散行列が同じであるとしても特に大きな問題は無いと判断した。

男女それぞれのデータから推定した平均ベクトルは、$(7.42, 67.4, 179.9), (7.49, 77.6, 170.0)$ であり、これを母平均ベクトルとして用いた。さらに、プールした偏差積和行列から求めた分散共分散行列

$$\begin{pmatrix} 0.0826 & 0.5044 & 0.5446 \\ 0.5044 & 128.4 & 62.18 \\ 0.5446 & 62.18 & 375.8 \end{pmatrix}$$

を共通の分散共分散行列として用いた。

$X = 0$ を男性に対応させ、$p_0^{(0)}$ としてはデータでの相対頻度とは異なる値であるが、0.8 および 0.9 の 2 つの値に設定した。例えば、健康診断の受診者の男女比が偏る場合には、このような $p_0^{(0)}$ を用いることも考えられる。

各 $p_0^{(0)}$ の値に対する C 法、L 法および M 法における理論上の条件付誤報率、および実際のデータで正常人を肝疾患があると判定した数、肝疾患がある人を正しく判定した数は表 3.7 の通りである。

C 法、L 法、M法の順に、$X = 0$ における補正項が小さくなる（$X = 1$ における補正項が大きくなる）ので、正常な場合は、$X = 0$ では C 法が、$X = 1$ では M 法が最も肝疾患があると判定されやすくなると予想される。$p_0^{(0)} = 0.8$ の場合、正常な個体については、$X = 0$ の場合はM法が、$X = 1$ の場合では C 法が最も誤りが少ないことが見てとれる。一方、肝疾患がある個体については、$X = 0$ の場合では誤りの程度には差がないが、$X = 1$ の場合では L 法およびM法で C 法よりも多く肝疾患を正しく検出しているのがわかる。このような事情は、$p_0^{(0)} = 0.9$ の場合により顕著になることが表

表 3.7: 各手法による異常判定数の比較

	条件付誤報率		正常を異常と判定		異常を異常と判定	
$p_0^{(0)} = 0.8$	$X = 0$	$X = 1$	$X = 0(70)$	$X = 1(130)$	$X = 0(4)$	$X = 1(12)$
C 法	0.0500	0.0500	3	3	3	10
L 法	0.0336	0.1155	1	11	3	11
M 法	0.0269	0.1424	1	14	3	11

	条件付誤報率		正常を異常と判定		異常を異常と判定	
$p_0^{(0)} = 0.9$	$X = 0$	$X = 1$	$X = 0(70)$	$X = 1(130)$	$X = 0(4)$	$X = 1(12)$
C 法	0.0500	0.0500	3	3	3	10
L 法	0.0315	0.2169	1	31	3	11
M 法	0.0085	0.4233	0	64	3	12

から確認できる。このデータは、異常の度合いが強い観測値が多いので、これだけの結果から明確な判断をすることはできないが、q_0 が $p_0^{(0)}$ よりある程度減少する場合にはL 法が優位であることが示唆されている。

3.4 結論

離散変量を含む異常検出問題でロケーションモデルが仮定できるときの 3 つの異常検出法、C 法、M法、L 法の誤報率と検出力についてその性質を明らかにした。さらに、2 値変量の場合については、数値計算で検出力の観点から 3 つの方法をを比較した。その結果、連続変量の変化量である非心度が各水準で等しい場合は、離散変量の確率分布の変化が小さい場合は C 法が、正常状態で確率が小さい水準の確率が異常状態で急激に大きくなる場合は M 法が最適となるが、その中間領域ではL 法が最適となることが確認でき、その中間領域は非心度の増加とともに広がることが確認できた。正常状態で確率が小さい事象の確率が大きくなり、非心度がある程度増加するときは、 L法が優れた異常検出法であると考えられる。

この章では、すべてのパラメータを既知としたが、実際にはこれらの値を推定しなければならない。それについては、第 4 章と第 5 章で議論する。

第4章　母数が未知の場合の異常検出法

この章では、分布の母数が未知の場合における、離散変量と連続変量が混在する異常検出問題を取り扱う。正常状態での個体についての観測値 (初期データ) を想定し、それを基にして異常検出法を構成する。まずはじめに、母数が既知のときの異常検出のための統計量の母数に、正常状態での観測値に基づく推定量を代入するという、推定方式による異常検出法について述べる。さらに、判定標本も併せた全データに対する仮説検定問題における尤度比検定に基づく異常検出法（検定方式）を構成する。また、初期データについて期待値をとった期待誤報率を設定値になるべく一致させるような棄却限界値の決定法についても議論する。

4.1　推定方式による異常検出

4.1.1　推定方式

水準数 I の離散変量 X と m 次元連続変量 $\boldsymbol{Y}$ が観測されるものとし、これらをまとめて $\boldsymbol{U} = (\boldsymbol{Y}', X)'$ と表記する。ここでもロケーションモデルを仮定する。すなわち、正常な個体からなる母集団（正常群とよぶ）π において、離散変量 X について $X = x,\ x = 1, \ldots, I,$ となる確率を p_x とし、$X = x$ が与えられた条件のもとで連続変量 $\boldsymbol{Y}$ は、平均 $\boldsymbol{\mu}_x$、分散共分散行列 Σ の正規分布 $N(\boldsymbol{\mu}_x, \Sigma)$ にしたがうとする。新たに、正常群に属さないかどうかを判定する標本（判定標本とよぶ）について、$\boldsymbol{U} = (\boldsymbol{Y}', X)'$ が観測されるとする。

第 2 章、第 3 章では、正常群での分布の母数 $p_x,\ \boldsymbol{\mu}_x, x = 1, \ldots, I,$ およびΣ を既知として、マハラノビス距離法（M 法）と尤度比法（L 法）、さらに、離散変量の値を与えたときの条件付分布に基づいて異常検出を行う条件付法（C 法）について議論した。これらの方法では、異常検出のための統計量はすべて、正常状態において χ^2 分布にしたがう連続変量に基づくマハラノビス平方距離 $(\boldsymbol{Y} - \boldsymbol{\mu}_x)'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_x)$ と離散変量のしたがう分布によって定まる補正項の和として表現された。分布の母数が未知の場合における素朴な方法は、これらの異常検出のための統計量に分布の母数の推定量を代入する方法である。この方法を推定方式とよぶ。

4.1.2　初期データに基づく未知母数の推定

推定のためには、正常群についての観測値が得られていることが前提となる。正常群からの大きさ n の無作為標本である初期データを $\boldsymbol{u}_i = (\boldsymbol{y}_i', x_i)',\ i = 1, \ldots, n,$ とすると、その同時確率・確率密度関数は

$$\prod_{i=1}^{n} \left[p_{x_i} (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_{x_i})' \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{x_i}) \right\} \right]$$

である。

初期データに基づく母数 $p_x,\ \boldsymbol{\mu}_x, x = 1, \ldots, I,\ \Sigma$ の推定量はつぎのように与えることができる。$I(\cdot)$ を、条件を満足するときに 1、それ以外で 0 をとる定義関数とするとき

$$n_x = \sum_{i=1}^{n} I(x_i = x), \qquad x = 1, \ldots, I$$

とおき、$\boldsymbol{n} = (n_1, n_2, \ldots, n_I)'$ とする。そのとき、$p_x,\ \boldsymbol{\mu}_x$ の最尤推定量は

$$\hat{p}_x = \frac{n_x}{n}, \qquad \hat{\boldsymbol{\mu}}_x = \frac{1}{n_x} \sum_{i=1}^{n} I(x_i = x)\boldsymbol{y}_i, \qquad x = 1, \ldots, I$$

で与えられる。$p_x,\ \boldsymbol{\mu}_x$ の推定量としてはこれらを用いる。なお、$n_x = 0$ の場合 $\boldsymbol{\mu}_x$ は推定できない。

$n_x = 0$ となる x の個数を $z(\boldsymbol{n})$ とおくと、$x_1, \ldots, x_n$ を与えたとき

$$S = \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_{x_i})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_{x_i})'$$

はウィッシャート分布 $W(n - I + z(\boldsymbol{n}), \Sigma)$ にしたがう。自由度に $z(\boldsymbol{n})$ があるが、これは、頻度が 0 の水準の数だけ平均を推定しないので、偏差平方和の自由度が増加するためである。

Σ の最尤推定量は S/n であり、不偏推定量は $S/(n - I + z(\boldsymbol{n}))$ である。さらに、その逆行列が Σ^{-1} の不偏推定量となる $S/(n - I - m - 1 + z(\boldsymbol{n}))$ も候補となる。棄却限界値の決定法によっては、Σ の推定量の選択が大きな影響を与えることが考えられるので、本章では一般に、c を正の定数として S/c という形の推定量について議論する。2 値変量の場合については、4.3.1 節,4.3.2 節で数値計算に基づき期待誤報率について議論し、c の選択についても論じる。

4.1.3 推定方式による異常検出

L 法を例にとり議論を始める。正常群の分布の母数が既知の場合、2 章で示したように新たな個体について観測値 $(\boldsymbol{y}', x)$ を得たとき、仮説 $H_0 : (\boldsymbol{y}', x) \in \pi$ を検定する尤度比検定は

$$(\boldsymbol{y} - \boldsymbol{\mu}_x)'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_x) - 2\log p_x > K \tag{4.1}$$

のときに仮説を棄却する。K は、誤報率が指定される値に一致するように定められた。分布の母数が未知の場合、推定方式による L 法の異常検出統計量は、(4.1) 式の $p_x,\ \boldsymbol{\mu}_x, \Sigma$ に推定量を代入することで得られる。なお、M 法は、(4.1) 式左辺の第 2 項を $(1 - p_x)/p_x$ で置き換えた場合に相当し、C 法は、第 2 項が 0 の場合に相当する。

いま、

$$\Delta^2(c) = c(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_x)'S^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_x), \quad h_L(n_x) = -2\log\left(\frac{n_x}{n}\right) \tag{4.2}$$

とおくとき、推定方式による L 法の異常検出は、定数 K_L を適当に定めて

$$\Delta^2(c) + h_L(n_x) > K_L$$

のとき、棄却することになる。M 法、C 法については、$h_L(n_x)$ を $h_M(n_x) = (n - n_x)/n_x$、$h_C(n_x) = 0$ で、K_L を K_M、K_C でそれぞれ置き換えればよい。なお、$n_x = 0$ のときは、M 法と L 法では補正

項が ∞ になるため、結果として連続変量の値によらず棄却することになるが、C 法においても連続変量の値によらず棄却することとする。

推定方式による異常検出の議論は、補正項が異なるだけで本質的には手法にはよらない。そこで、一般に記号 A で方法を表すことにし、$A = M, L, C$ であり、それぞれ M 法、L 法、C 法を表すことにする。

連続変量のみ存在する場合、母数が未知のとき、マハラノビス平方距離の計算に初期データを基にした母数の最尤推定値を用い、母数が既知の場合の分布である χ^2 分布により棄却限界値を定めると、正常な個体を異常と判定する誤報率の期待値（期待誤報率とよぶ）が指定される設定値より大きくなることが宮川, 他 (2007) により報告されている。さらに、初期データから計算される統計量の分布を考慮し F 分布を用いて棄却限界値を設定することで、期待誤報率を設定値に一致させることが可能であることも示している。ここでも、4.1.4 節で各分布を用いる棄却限界値の決定法について議論する。

誤報率としては、初期データの観測値を固定して、判定標本についての判定を繰り返し行ったときの誤報率も考えられる。これを実際の誤報率とよぶ。ここでは実際の誤報率の初期データの分布についての期待値である期待誤報率を主に問題にする。実際の誤報率の分布に関しては、2 値変量の場合についてシミュレーションに基づいて 4.4 節で考察する。

4.1.4 棄却限界値の決定と期待誤報率

4.1.4.1 χ^2 分布法

(4.2) 式で与えられる $\Delta^2(c)$ は、cS^{-1} を Σ^{-1} とすれば、自由度 m の χ^2 分布にしたがうので、そのようにみなして期待誤報率の推定値が設定値 α に一致するように棄却限界値を定める方法が考えられる。すなわち、自由度 m の χ^2 分布の分布関数を $F(\cdot)$ として

$$\sum_{x=1}^{I} \frac{n_x}{n} F(K_A - h_A(n_x)) = 1 - \alpha$$

を満たす K_A を棄却限界値とする方法である。この方法を χ^2 分布法とよぶ。棄却限界値 K_A を決定する式に $n_x, x = 1, \ldots, I$ が入っていることからわかるように、棄却限界値は離散変量の頻度 $n_1, \ldots, n_I$ の関数になるので、$K_A(\boldsymbol{n})$ と書くことにする。このように棄却限界値を定めたときの期待誤報率は、設定値 α に完全には一致しないが、その表現については、つぎの F 分布法の項であわせて述べることにする。

4.1.4.2 F 分布法

(4.2) 式で定義される $\Delta^2(c)$ について、$n_1, \ldots, n_I$ および x の値を固定し $n_x > 0$ とすれば、$\boldsymbol{y}$ と $\hat{\boldsymbol{\mu}}_x$ および S の標本変動を考慮したとき、

$$\frac{n_x}{n_x+1} \frac{n-m-I+1+z(\boldsymbol{n})}{m} \frac{1}{c} \Delta^2(c) = \frac{n_x}{n_x+1} \frac{n-m-I+1+z(\boldsymbol{n})}{m} (\boldsymbol{y}-\hat{\boldsymbol{\mu}}_x)' S^{-1} (\boldsymbol{y}-\hat{\boldsymbol{\mu}}_x) \quad (4.3)$$

は自由度 $(m,\ n-m-I+1+z(\boldsymbol{n}))$ の F 分布にしたがう。これは、初期データの分布も考慮した場合、$X = x$ のとき $(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_x)$ が平均 $\boldsymbol{0}$、分散共分散行列 $(1 + \frac{1}{n_x})\Sigma$ の正規分布に、S がウィッシャート分布 $W(n - I + z(\boldsymbol{n}), \Sigma)$ にしたがうことからわかる (例えば Seber(2008) の p468 を参照)。

自由度 $(m,\ n-m-I+1+z(\boldsymbol{n}))$ の F 分布の分布関数を $F(\cdot)$ で表せば、$X=x$ のときの条件付誤報率を

$$1-F\left[\frac{n_x}{n_x+1}\frac{n-m-I+1+z(\boldsymbol{n})}{m}\frac{1}{c}\left\{K_A-h_A(n_x)\right\}\right] \tag{4.4}$$

で推定するのが自然である。なお $n_x=0$ のときには棄却するのであるが、その場合 (4.4) 式は $1-F(0)=1$ となるので、この表現をそのまま用いてよい。したがってこの条件付誤報率の推定値に、仮説が正しいときの $X=x$ となる確率の推定値である n_x/n をかけて平均することで期待誤報率の推定値が得られることになる。誤報率の設定値を α とするとき、この推定値を α に等しくなるように K_A を定めることにする。つまり

$$\sum_{x=1}^{I}\frac{n_x}{n}F\left[\frac{n_x}{n_x+1}\frac{n-m-I+1+z(\boldsymbol{n})}{m}\frac{1}{c}\left\{K_A-h_A(n_x)\right\}\right]=1-\alpha \tag{4.5}$$

を満たすように A 法での棄却限界値 K_A を決定するのである。K_A は χ^2 分布法と同様、$n_1,\ldots,n_I$ の関数として定まるので $K_A(\boldsymbol{n})$ と表すことにする。このとき、期待誤報率を設定値 α に完全に一致させることはできないが、初期データの大きさがある程度大きければ、α に近い値になることが期待される。このように、判定標本の分布だけでなく初期データの連続変量の分布を考慮して棄却限界値を定める方法では F 分布を用いるので、これを F 分布法と呼ぶことにする。

以上のように $K_A(\boldsymbol{n})$ の値を定めたとき、初期データの離散変量の頻度と判定標本の離散変量の値が与えられたときの条件付誤報率は (4.4) 式の K_A を $K_A(\boldsymbol{n})$ で置き換えた

$$1-F\left[\frac{n_x}{n_x+1}\frac{n-m-I+1+z(\boldsymbol{n})}{m}\frac{1}{c}\left\{K_A(\boldsymbol{n})-h_A(n_x)\right\}\right] \tag{4.6}$$

で与えられるので、期待誤報率は、(4.6) 式を X および $n_1,\ldots,n_I$ の確率分布で期待値をとることで得られる。つまり、I 項分布 $M_I(n,p_1,\ldots,p_I)$ の確率関数を $f(n_1,\ldots,n_I)$ で表せば、期待誤報率は

$$\sum_{x=1}^{I}p_x\sum_{n_1,\ldots,n_I}f(n_1,\ldots,n_I)\left(1-F\left[\frac{n_x}{n_x+1}\frac{n-m-I+1+z(\boldsymbol{n})}{m}\frac{1}{c}\left\{K_A(\boldsymbol{n})-h_A(n_x)\right\}\right]\right) \tag{4.7}$$

で与えられる。4.1.4.1 で述べた χ^2 分布法における期待誤報率は、$K_A(\boldsymbol{n})$ の意味が異なるだけでまったく同じ表現となる。

注意 1. 初期データにおいて観測されない離散変量の水準があるとき、これを異常判定に用いるということは考えにくい状況も多いだろう。4.3.3 節では、2 値変量の場合で初期データにおいて両水準とも観測されているという条件 (これを正条件と呼ぶ) の下での期待誤報率を数値的に求め、考察する。

注意 2. 連続変量のみの場合に Σ の推定量として S の定数倍を用いるとき、定数 c として何を用いても F 分布法により棄却域を定めれば同じ異常検出法に帰着する。しかし、離散変量が存在する場合には補正項 $h_A(n_x)$ が存在するので、定数 c の選び方により、得られる異常検出方式が異なることに注意する。

4.2 検定方式による異常検出

ここでは、まず母数がすべて未知のときの異常検出問題に対し、尤度比検定の考え方により検定統計量を導出し、それに基づく検定方式の異常検出法における棄却限界値の決定方法と期待誤報率について議論する。

4.2.1 尤度比検定統計量

この方法では、初期データも検定の枠組みに入れ、初期データと判定標本が同じ母集団分布にしたがうという帰無仮説を、判定標本は初期データとは異なる母集団分布にしたがうという対立仮説に対して検定を行う。つまり、仮説および対立仮説をそれぞれ

$$H_0 : \boldsymbol{u}_i = (\boldsymbol{y}_i', x_i)' \in \pi, \quad i = 1, \ldots, n, \quad (\boldsymbol{y}', x)' \in \pi,$$

$$H_1 : \boldsymbol{u}_i = (\boldsymbol{y}_i', x_i)' \in \pi, \quad i = 1, \ldots, n, \quad (\boldsymbol{y}', x)' \notin \pi$$

としたときの尤度比検定統計量を求める。判定標本について $X = x, x = 1, \ldots, I,$ となる確率を q_x とし、$X = x$ が与えられたとき連続変量$\boldsymbol{Y}$は、平均 $\boldsymbol{\eta}_x$, 分散共分散行列 Σ の正規分布 $N(\boldsymbol{\eta}_x, \Sigma)$ にしたがうとする。一般に、初期データおよび判定標本の同時確率・確率密度関数は

$$\prod_{i=1}^{n} \left[p_{x_i} (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_{x_i})' \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{x_i}) \right\} \right]$$

$$\times q_x (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\eta}_x)' \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\eta}_x) \right\}$$

で与えられる。以下では、記号の混乱を避けるため、判定標本の離散変量 X の値が 1 すなわち $x = 1$ の場合について記述する。また、初期データだけに基づく最尤推定量と区別するため、帰無仮説の下で尤度を最大にする推定量は、$\check{\Sigma}$ のように表記する。

$n_1 = 0$ の場合については後で議論することにして、$n_1 > 0$ として議論を始める。H_0 が正しいとき、$q_1 = p_1, \boldsymbol{\eta}_1 = \boldsymbol{\mu}_1$ なので、尤度関数

$$p_1^{n_1+1} \left(\prod_{x=2}^{I} p_x^{n_x} \right) (2\pi)^{-\frac{m(n+1)}{2}} |\Sigma|^{-\frac{n+1}{2}}$$
$$\times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{x_i})' \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{x_i}) - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}_1)' \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_1) \right\}$$

を最大にするのは

$$\check{p}_1 = (n_1 + 1)/(n + 1), \quad \check{p}_x = n_x/(n + 1), \quad x = 2, \ldots, I,$$

$$\check{\boldsymbol{\mu}}_1 = \frac{1}{n_1 + 1} (n_1 \hat{\boldsymbol{\mu}}_1 + \boldsymbol{y}), \quad \check{\boldsymbol{\mu}}_x = \hat{\boldsymbol{\mu}}_x, \ x = 2, \ldots, I,$$

$$\check{\Sigma} = \frac{1}{n + 1} \left\{ \sum_{i=1}^{n} (\boldsymbol{y}_i - \check{\boldsymbol{\mu}}_{x_i})(\boldsymbol{y}_i - \check{\boldsymbol{\mu}}_{x_i})' + (\boldsymbol{y} - \check{\boldsymbol{\mu}}_1)(\boldsymbol{y} - \check{\boldsymbol{\mu}}_1)' \right\} \tag{4.8}$$

のときである。なお、$n_x = 0,\ x = 2, \ldots, I$ のとき $\boldsymbol{\mu}_x$ は推定できない。したがって、H_0 が正しいときの最大尤度は

$$\left(\frac{n_1+1}{n+1}\right)^{n_1+1} \prod_{x=2}^{I} \left(\frac{n_x}{n+1}\right)^{n_x} (2\pi)^{-\frac{m(n+1)}{2}} |\check{\Sigma}|^{-\frac{n+1}{2}} \exp\left(-\frac{m(n+1)}{2}\right) \tag{4.9}$$

で与えられる。なお、$n_x = 0,\ x = 2, \ldots, I$ のとき $(n_x)^{n_x} = 1$ とする。

H_1 が正しいときの最尤推定量は、$p_x, \boldsymbol{\mu}_x$ については $\hat{p}_x,\ \hat{\boldsymbol{\mu}}_x$ で与えられ、Σ については $n/(n+1)\hat{\Sigma}$ で与えられる。ここで $\hat{\Sigma} = \frac{1}{n}S$ である。判定標本の分布の母数 $q_1, \boldsymbol{\eta}_1$ についてはそれぞれ $1,\ \boldsymbol{y}$ で与えられる。そのときの最大尤度は、

$$\prod_{x=1}^{I} \left(\frac{n_x}{n}\right)^{n_x} (2\pi)^{-\frac{m(n+1)}{2}} \left|\frac{n}{n+1}\hat{\Sigma}\right|^{-(n+1)/2} \exp\left(-\frac{m(n+1)}{2}\right) \tag{4.10}$$

となる。

式 (4.9) と (4.10) の比を取れば、尤度比検定統計量は

$$\frac{(n+1)^{n+1}}{n^n} \frac{n_1^{n_1}}{(n_1+1)^{n_1+1}} \left(\frac{|\check{\Sigma}|}{|\frac{n}{n+1}\hat{\Sigma}|}\right)^{(n+1)/2}$$

となる。

ここで、$\hat{\boldsymbol{\mu}}_1 - \check{\boldsymbol{\mu}}_1 = -(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)/(n_1+1),\ \ (\boldsymbol{y} - \check{\boldsymbol{\mu}}_1) = n_1/(n_1+1)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)$ を用いて、(4.8) 式から

$$(n+1)\check{\Sigma} = S + \frac{n_1}{n_1+1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)'$$

を得ることができる。したがって $\check{\Sigma} = n/(n+1)\hat{\Sigma} + n_1/\{(n+1)(n_1+1)\}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)'$ であるので、これに $|A + \boldsymbol{d}\boldsymbol{d}'| = |A|(1 + \boldsymbol{d}'A^{-1}\boldsymbol{d})$ なる等式を用いれば、対数尤度比検定統計量の 2 倍は

$$(n+1)\log\left\{1 + \frac{n_1}{n(n_1+1)}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)'\hat{\Sigma}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)\right\} + h_T(n_1)$$

となる、ここで

$$h_T(n_1) = 2\left\{(n+1)\log(n+1) - n\log(n) + n_1\log(n_1) - (n_1+1)\log(n_1+1)\right\}, \qquad n_1 > 0$$

である。したがって、対数尤度比検定統計量がマハラノビス平方距離 $(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)'\hat{\Sigma}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)$ の単調増加関数と離散変量の観測度数による補正項との和になることが確認できた。

つぎに $n_1 = 0$ の場合について考えよう。その場合、H_0 が正しいときの $\boldsymbol{\mu}_1$ の最尤推定量は、$\check{\boldsymbol{\mu}}_1 = \boldsymbol{y}$ となる。その結果、$\check{\Sigma} = S/(n+1)$ となり、最大尤度の連続変量の関わる部分は両仮説で等しくなる。したがって、$n_1 = 0$ のときの対数尤度比検定統計量の 2 倍は、$2\{(n+1)\log(n+1) - n\log(n)\}$ となる。これは定数であり、棄却限界値との大小関係により棄却するか否かが定まることになる。

設定値 α の値にも依存するが、正確な尤度比検定の考え方によるならば、$n_1 = 0$ のときにも棄却しない可能性もありうる。しかし、そこまで考慮すると議論を複雑にしてしまうし、また、推定方式との比較のためにも、$n_1 = 0$ のときには棄却することにする。$n_1 = 0$ の場合にすべて棄却してしまうことで、尤度比検定と多少のずれが生じうるが、n がある程度大きければ、観測度数が 0 になる確率が無視できるため、実質上ほとんど差異はないと考えられる。以上により定まる検定方式による異常検出法を T 法とよぶ。

4.2.2 棄却限界値の決定と期待誤報率

ここでは、棄却限界値の決定と期待誤報率について議論する。尤度比統計量の分布を χ^2 分布法で近似する方法と F 分布法を取り上げる。

4.2.2.1 χ^2 分布法

$X = x$ に対し $n_x > 0$ のとき、対立仮説の下では異常状態での離散変量の確率と、その観測値における連続変量の平均を推定するので、2つの仮説の下で推定する母数の数の差は $m + 1$ である。そこで、尤度比検定統計量の分布を自由度 $m + 1$ の χ^2 分布で近似して、

$$(n+1)\log\left\{1 + \frac{n_x}{n(n_x+1)}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_x)'\hat{\Sigma}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_x)\right\} + h_T(n_x) > \chi^2(m+1, \alpha)$$

のとき異常と判定することが考えられる。4.1.4.1 節の場合とその意味は異なるが、この方法も χ^2 分布法とよぶことにする。

このとき、$X = x$ および $n_1, \ldots, n_I$ が与えられたときの条件付誤報率は、(4.3) 式が自由度 $(m, n - m - I + 1 + z(\boldsymbol{n}))$ の F 分布にしたがうことから、

$$1 - F\left(\frac{n - m - I + 1 + z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n+1}\left(\chi^2(m+1,\alpha) - h_T(n_x)\right)\right\} - 1\right]\right)$$

となる。$n_x = 0$ の場合は棄却することにしたので、$h_T(0) = \infty$ とすればよい。

このようにして求められた条件付誤報率を X および $n_1, \ldots, n_I$ の分布について期待値をとれば期待誤報率が求められる。

4.2.2.2 F 分布法

本来の尤度比検定法の考え方からすれば、棄却限界値は $n_1, \ldots, n_I$ の値と無関係に定めるべきものであるが、設定値を実現するよう定めるのが困難であるのでつぎのように棄却限界値 K_T の値を $n_1, \ldots, n_I$ の値に依存して定めることにする。まず、$n_1, \ldots, n_I$ の値を固定し $X = x$ のときの誤報率の推定量として

$$1 - F\left(\frac{n - m - I + 1 + z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n+1}\left(K_T - h_T(n_x)\right)\right\} - 1\right]\right)$$

を用いるのが一つの単純な方法である。期待誤報率の推定値は、この $X = x$ のときの期待誤報率の推定値を、$p_x, x = 1, \ldots, I$ の代わりに標本相対度数 $n_x/n,\ x = 1, \ldots, I$ を用いて平均することで得られる。したがって

$$\sum_{x=1}^{I} \frac{n_x}{n} F\left(\frac{n - m - I + 1 + z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n+1}\left(K_T - h_T(n_x)\right)\right\} - 1\right]\right) = 1 - \alpha \tag{4.11}$$

を満たす K_T を棄却限界値とする。K_T は、他の K_A と同様、$n_1, \ldots, n_I$ の関数として定まるので $K_T(\boldsymbol{n})$ と表す。また、$n_x = 0$ となる水準での判定方法をどのように決めても、$X = x$ のときの推定誤報率に相対度数 0 を掛けるため、棄却限界値を決める (4.11) 式には影響がないことにも注意す

る。このように定めた棄却限界値 $K_T(\boldsymbol{n})$ を用いるとき、初期データの離散変量の頻度 $\boldsymbol{n}$ と $X = x$ を与えたときの条件付誤報率は

$$1 - F\left(\frac{n-m-I+1+z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n+1}\left(K_T(\boldsymbol{n}) - h_T(n_x)\right)\right\} - 1\right]\right) \tag{4.12}$$

である。期待誤報率は、(4.12) 式を X と $n_1, \ldots, n_I$ の分布についての期待値として (4.7) 式と同様に、

$$1 - \sum_{x=1}^{I} p_x \sum_{n_1,\ldots,n_I} f(n_1, \ldots, n_I) F\left(\frac{n-m-I+1+z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n+1}\left(K_T(\boldsymbol{n}) - h_T(n_x)\right)\right\} - 1\right]\right) \tag{4.13}$$

で与えられる。

注意 3：離散変量の確率分布が既知の場合は、式 (4.11) において、n_x/n の代わりに p_x を用いて棄却限界値 K_T を決定すればよい。したがって、期待誤報率が設定値 α に一致することがわかる。推定方式においても同様に、棄却限界値を F 分布法を用いて決定すれば期待誤報率は設定値 α に一致する。このように、離散変量の分布が既知の場合は、正確な期待誤報率をもつ異常検出法の設計が可能となる。たとえば、健康診断の受診者の男女の割合が他の情報源から正確にわかる場合や、製品検査においてライン別製造割合がわかっている場合などでは、離散変量の分布が正確に把握できると考えられる。

4.3 2 値変量のときの期待誤報率の挙動

ここでは、離散変量が 2 値の場合における期待誤報率について、数値計算の結果を基に考察する。検定方式の T 法と推定方法の L 法、M 法、C 法を取り上げる。m は 10、α は 0.05 とし、n は 20 から 100 について、離散変量が値 0 をとる確率 p_0 を 0.5 から 1 まで変化させたときの期待誤報率を数値的に求めた。さらに、正条件のもとでの期待誤報率および、m と α を変化させた場合についても議論する。また、特に断らない限り $m = 10, \alpha = 0.05$ とする。

4.3.1 χ^2 分布法を用いるときの期待誤報率

連続変量のみの場合、分散共分散行列の推定量として最尤推定量を用いて χ^2 分布法を適用する場合は、期待誤報率が非常に大きな値になることが報告されている（宮川, 他 (2007)）。しかし、分散共分散行列の推定量を修正することで、χ^2 分布法による棄却限界値を用いる場合でもそれほど期待誤報率を大きくしないことが可能であることが数値計算から確認できた。そこで、2 値変量が含まれる場合についても、用いる分散共分散行列の推定量による期待誤報率の違いを、L 法について考察する。M 法、C 法でも状況はほとんど同じであった。

L 法で n が 50 の場合について期待誤報率を計算した結果が図 4.1 である。(4.2) 式で与えられる $\Delta^2(c)$ における定数 c としては、$c_1 = n$(最尤推定量)、$c_2 = n - m - 3 + z(\boldsymbol{n})$($\Sigma^{-1}$ の不偏推定量) に加えて、$c_3 = n - 2 + z(\boldsymbol{n})$($\Sigma$ の不偏推定量) を取り上げた。図からわかるように、離散変量が混在した場合でも、c_1 や c_3 を用いた場合は、期待誤報率が非常に大きな値になることが確認できる。一方、c_2 を用いた場合は、それほど大きな値にはならず、設定した値 α にかなり近い値になること

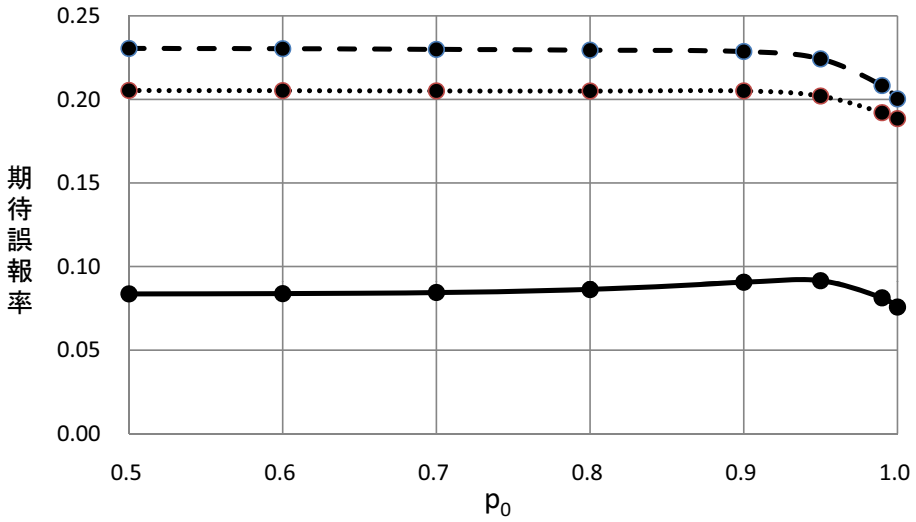

図 4.1: χ^2 分布法において用いる Σ の推定量による期待誤報率の比較 ($n = 50$) (破線:最尤推定量 (c_1)、点線:不偏推定量 (c_3)、実線:Σ^{-1} の不偏推定量 (c_2))

がわかる。この傾向は、除数の大きさからの当然の帰結であるが、Σ^{-1} の不偏推定量を用いた場合でも後で示すように F 分布法を用いた場合と比べると設定値に十分近いとは言い難い。

T 法および、c_2 を用いた各推定方式において、χ^2 分布法を用いたときの期待誤報率を $n = 50$ の場合について示したのが図 4.2 である。$n = 50$ でも T 法の期待誤報率は推定方式と比べてかなり大きいので、T 法で χ^2 分布法を用いるのは適当ではないと判断できる。

4.3.2 F 分布法を用いるときの期待誤報率

$m = 5, 10, 20$、$n = 20, 50, 100$、$\alpha = 0.05, 0.01$ について c_2 を用いた χ^2 分布法と、F 分布法を比較した。F 分布法では、c_1 と c_2 の双方を用いた場合について比較した。この範囲において期待誤報率はすべて設定値 α 以上であることが確認できた。また、期待誤報率と設定値 α の差は、$0.5 \leq p_0 \leq 0.9$ の範囲で χ^2 分布法が F 分布法の 2.5 倍以上あることが確認され、最大は 100 倍以上であった。この様子を $n = 50$、L 法について示したのが図 4.3 である。F 分布法の期待誤報率が設定値 α に近いのに対して、χ^2 分布法が大きな値をとっていることがわかる。F 分布法は、初期データの連続変量の分布を考慮したときの分布を用いて棄却限界値を定めているので、近似分布を用いる χ^2 分布法よりも優れているのは当然と考えられるが、この結果は、離散変量の存在しない場合 (宮川, 他 (2007) 参照) と同様である。この結果を踏まえ、以降では χ^2 分布法については考察の対象としないこととする。

また、この図からわかるように、F 分布法では用いる Σ の推定量による期待誤報率の差がほとんどない。二本の曲線をよく見ると、最尤推定量を用いたほうが少し下にあり、設定値に近いことが確認できる。この傾向は他の方法でもほぼ同様であり、$0.5 \leq p_0 \leq 0.9$ の範囲では c_1 を用いた方が概して期待誤報率が設定値に近くなっていた。この結果から、推定方式において F 分布法で棄却限界値を決定する場合には、Σ の推定量として最尤推定量を用いることにする。

F 分布法を用いたときの各手法の比較を行うが、推定方式では上の結果から $c_1 = n$ を用いた最尤推定量の場合のみを示すこととする。$n = 50$ のときの検定方式と各推定方式において F 分布法を用いたときの期待誤報率を示したのが図 4.4 である。

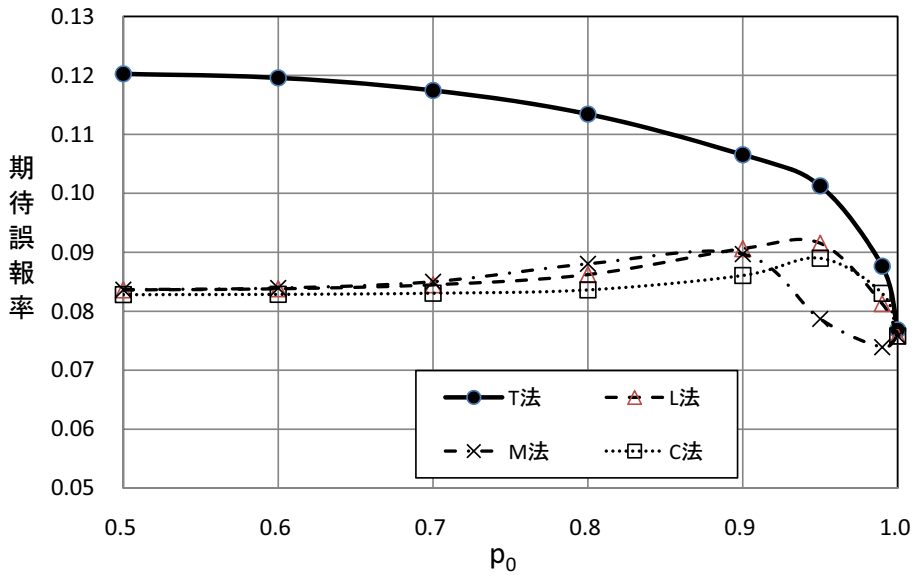

図 4.2: χ^2 分布法の比較 (T 法と推定方式)($n = 50$)

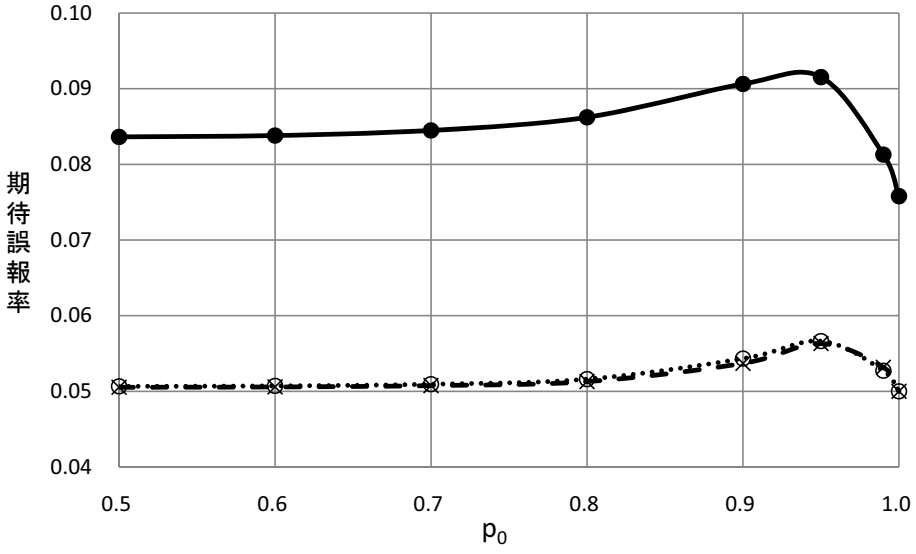

図 4.3: F 分布法と χ^2 分布法の比較 (L 法,n=50) (実線:χ^2 分布法 (Σ^{-1} の不偏推定量)、点線:F 分布法 (Σ^{-1} の不偏推定量)、破線:F 分布法 (最尤推定量))

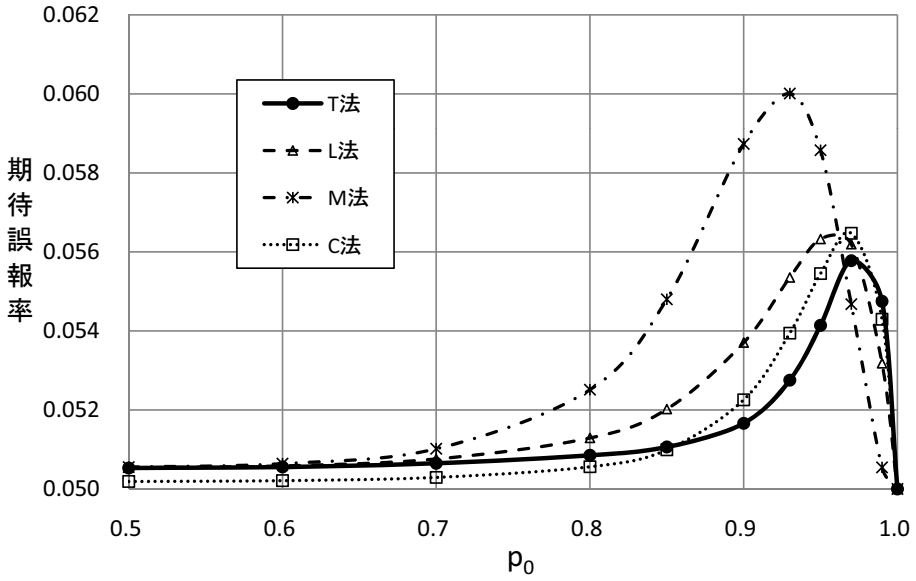

図 4.4: F 分布法を用いたときの期待誤報率 $(n = 50)$

L 法、M 法とも、T 法と比べると $p_0 = 0.8$ から 0.95 にかけて概して大きい値をとっていることがわかる。特に M 法の場合、p_0 が 0.8 を超えたあたりから急に増加し、0.9 の近くでは他の手法に比べてかなり大きな値をとることが確認される。C 法と T 法は全般的に設定値に近い値をとる。0.9 を過ぎたあたりで増加するが、最大値は T 法のほうが小さく、設定値に近いことが分かる。初期データの標本変動を直接的に考慮する T 法の方が推定方式よりも優れていると考えられる。

初期データ数 n の増加に伴う期待誤報率の変化を T 法の場合について示したのが図 4.5 である。どの手法でも同じであるが、n の増加に伴い期待誤報率が設定値に近づいていくことがわかる。また、$0.5 < p_0 < 0.9$ の範囲で期待誤報率と設定値 0.05 との差を 0.005 以下にしたいのであれば、$n = 30$ 程度で達成できることがわかる。

4.3.3 正条件のもとでの期待誤報率

これまでの結果は、2 値変量の片方の水準が初期データで観測されない場合も含めて誤判別率の期待値を求めている。しかし、実際には観測されない水準を取り上げるのは不自然な場合もある。ここでは、4.1 の注意 1 で述べた正条件の下で期待誤報率を求める。$n = 50$ における正条件の下での期待誤報率を示したのが図 4.6 である。

正条件をつけない場合、p_0 が 0.9 を越えたあたりで期待誤報率が大きな値になったが、正条件をつけることで $p_0 = 0.9$ 付近での期待誤報率の増加はある程度抑えられることがわかる。逆に p_0 が極端に 1 に近いとき設定値 α より小さくなることが確認される。これは、$n_x = 0 (x = 1, 2)$ の場合に必ず異常と判定すること、さらに、その場合を除外すると誤報率が減少することで説明できる。$0.5 \leq p_0 \leq 0.95$ の範囲での期待誤報率の設定値からの乖離の大きさは、M 法が最も大きく、次いで L 法、C 法で、T 法が最も小さい。また、p が 1 に近いとき設定値より小さくなることに関しても、その大きさは T 法が最も小さいことがわかる。

ここでは示さないが、T 法は C 法と比べて n が小さい場合においても、p_0 の値によらず設定値に近い値を維持していることが確認できた。また、C 法で期待誤報率が設定値に近いのは、補正項の値が離散変数の値によらず 0 であるため、マハラノビス平方距離での棄却限界値が離散変量の値に

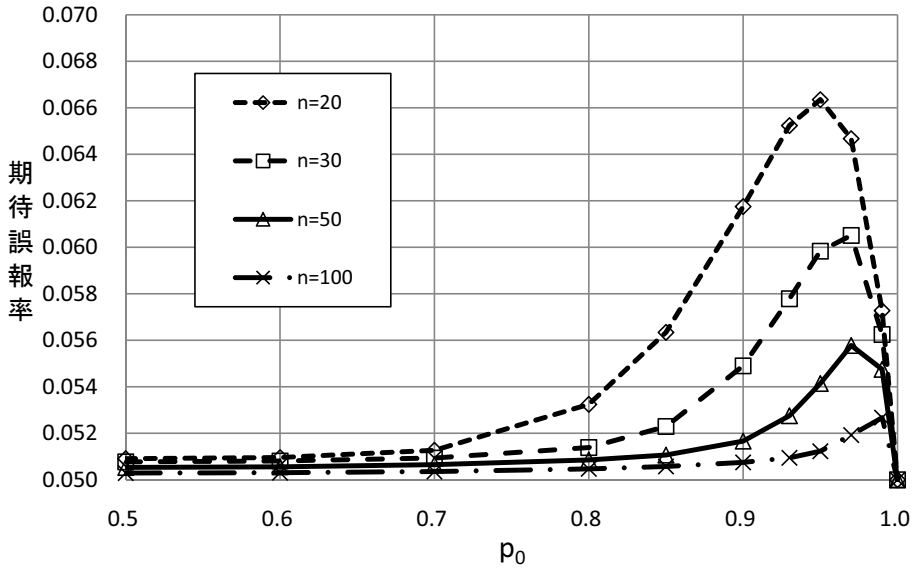

図 4.5: n による期待誤報率の変化 (T 法)

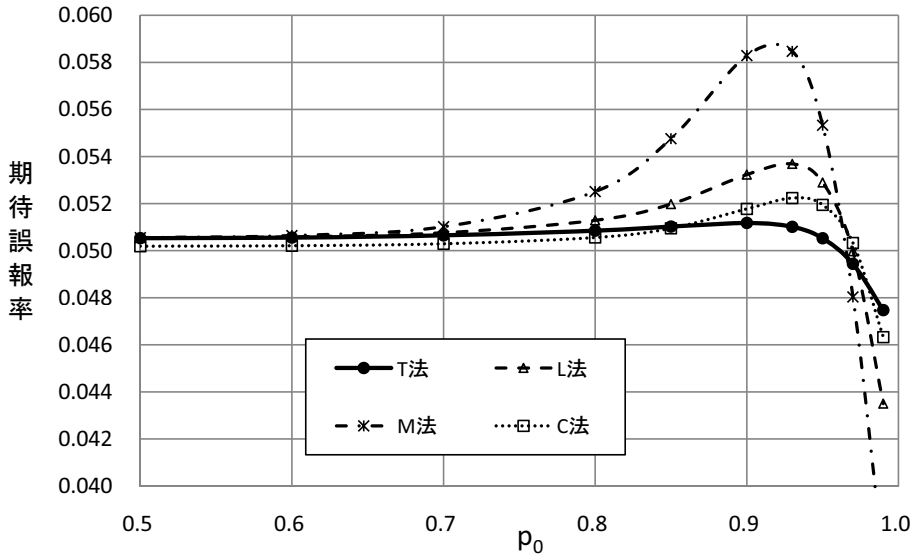

図 4.6: 正条件の下での期待誤報率 ($n = 50$)

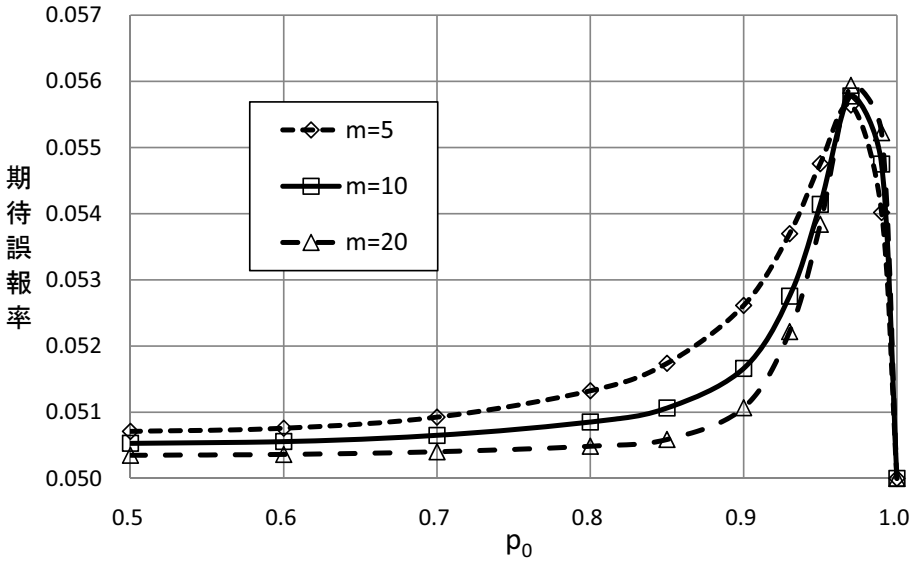

図 4.7: m による期待誤報率の変化 (T 法,$n = 50$)

より変わらないことが大きな理由であると考えられる。これは、離散変量を積極的に異常検出に利用する立場からは必ずしも好ましいとは言えない。その意味で、離散変量の値によりマハラノビス平方距離での棄却限界値を変えながら安定した期待誤報率を実現できている T 法は優れているといえる。

4.3.4 m および α の値が変化するときの期待誤報率の挙動

$n = 50$ の場合について、m を $5, 10, 20$ と変えた時の期待誤報率の変化を示したのが図 4.7 である。この図から、n が一定の場合、取り上げる連続変量の数を増やすと期待誤報率が設定値に近づくことが分かる。この結果は自明ではないし、この計算結果だけから結論づけられるわけではないが、連続変量はある程度多く取り上げた方が期待誤報率を設定値に近づける効果があることが示唆される。

つぎに、設定値 α が 0.05 より小さい場合を考える。管理図で通常用いられている 3σ *ルール*は、$\alpha = 0.0027$ に相当するように、異常検出は設定値 α が小さい状況で用いられる可能性が高い。ここでは、$\alpha = 0.01, 0.001$ の場合について期待誤報率を求めた。図 4.8 は、$\alpha = 0.01, n = 50$ の場合の期待誤報率である。

$\alpha = 0.01$ の場合、0.05 の場合に比べて n が小さい時には手法間の違いは小さい。しかし、n が 50 程度になると $p = 0.9$ 近辺で M 法、L 法、C 法の順に推定方式は、T 法に比べて大きな値をとることがわかる。また、α の減少に伴い、設定値との相対誤差は増加する傾向があることも確認できた。

正条件のもとでの期待誤報率は、T 法が他の手法に比べて設定値に近くなることはすでに述べたが、この傾向は α の値が小さいときでも成立していることが確認できた。これらの傾向は、$\alpha = 0.001$ でも確認できた。実際の異常検出では期待誤報率の設定値 α を小さく設定することが多い。そのような場合でも、T 法は安定した期待誤報率を維持できる方法であると考える。

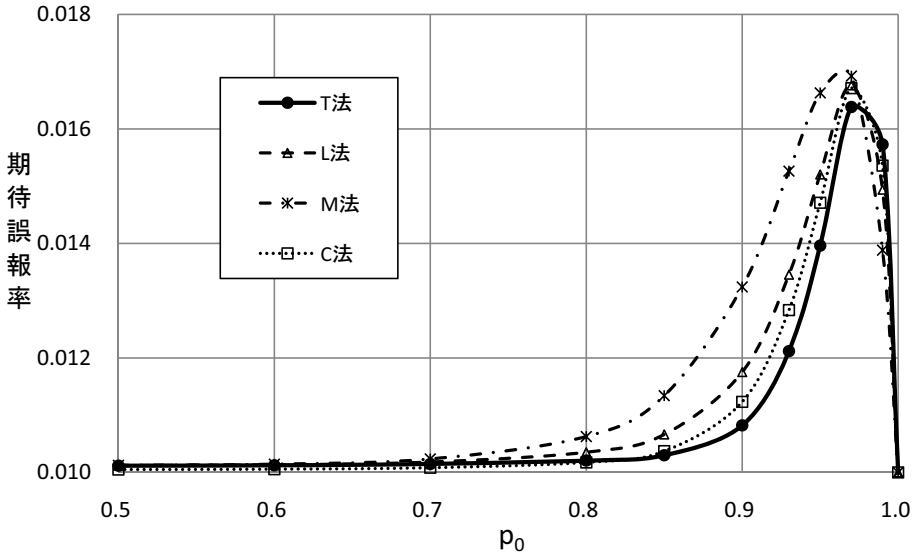

図 4.8: $\alpha = 0.01$ での期待誤報率 ($n = 50$)

4.4 実際の誤報率の分布

初期データにより定まる異常検出法を固定したとき、実際の誤報率がどの程度変動するかを知ることは、異常検出を行う際に考慮すべき重要な問題である。ここでは、2 値変量の場合の実際の誤報率の分布状況を、$m = 5, 10$ についてシミュレーションで確認する。

n を $20, 30, 40, 50, 100$ として、手法について共通の初期データを用い、実際の誤報率を推定することを 200 回繰り返した。誤報率の推定用のデータは 2 値変量の両水準について 100,000 組ずつ用意し、これを共通に用いて 2 値変量の値を与えたときの条件付誤報率を計算し、実際の誤報率の推定値を求めた。

$m = 10$ で F 分布法を用いた場合の、$n = 30, 50, 100, p = 0.7, 0.8, 0.9$ のときの実際の誤報率の分布状況を箱ひげ図で示したのが図 4.9 である。なお、箱は 25% 点と 75% 点を表し、ひげの先はそれぞれ 5% 点と 95% 点を示している。シミュレーションの繰り返しでの変動の影響が多少残っており、明確な差ではないものの、T 法においては、箱の長さとひげを含めた全体の長さが L 法および M 法に比べて短く、実際の誤報率の変動が小さい傾向があることが確認された。この差は、$p_0 = 0.7$ のときは小さいが、p が 1 に近づくにつれ大きくなっていくことも確認できる。

また、図 4.9 からは、実際の誤報率の分布が大きい側に裾を引いていることや、n が大きくなるにつれ、変動が減少することも確認できる。$m = 5$ の場合についてもほぼ同様の結果が得られている。

なお、$m = 5$ の場合には、χ^2 分布法についても実際の誤報率を求めた。その結果から、推定方式で χ^2 分布法で棄却限界値を決定した場合、c_2 を用いても F 分布法と比べて四分位範囲が広いことがわかった。F 分布法の χ^2 分布法に対する優位性は、その期待誤報率だけでなく実際の誤報率の分布でも確認された。

以上シミュレーションの結果からも、T 法が設定値 α に近い実際の誤報率を実現するという意味で、L 法や M 法に比べて優れていることが確認された。

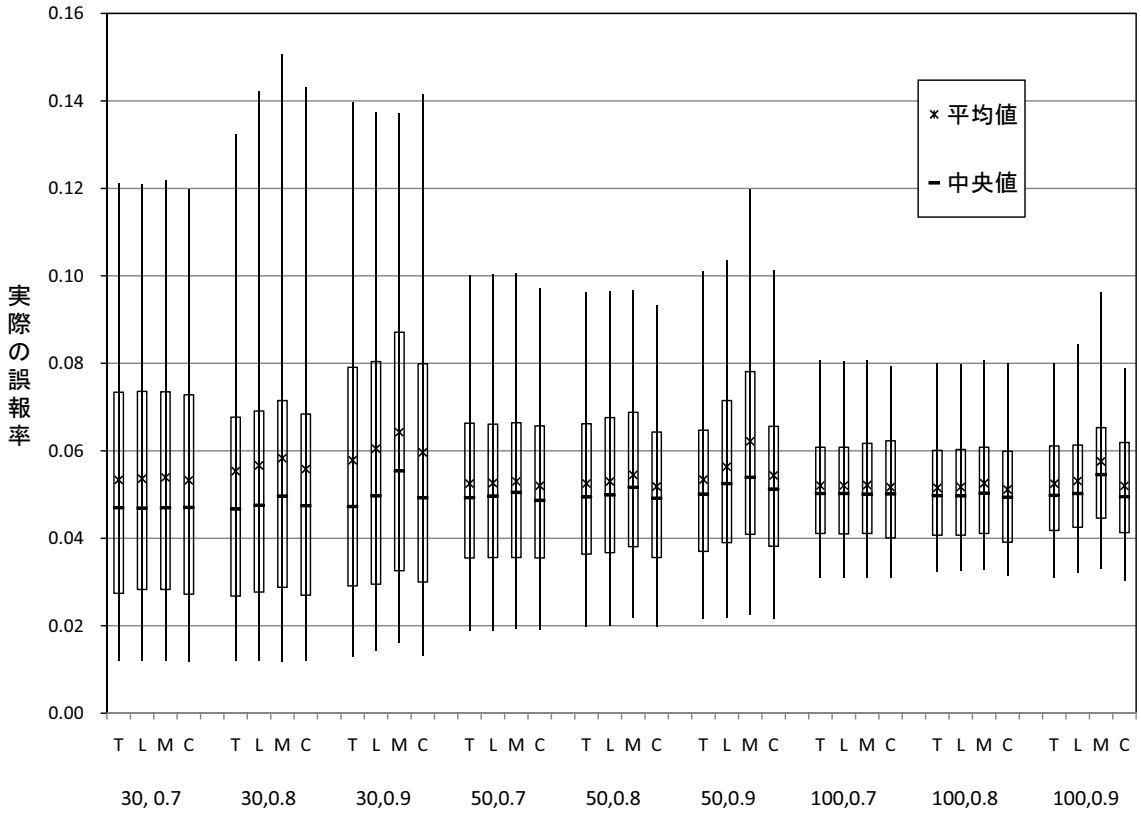

図 4.9: 実際の誤報率の箱ひげ図 $(m = 10, n = 30, 50, 100, p = 0.7, 0.8, 0.9, \alpha = 0.05)$ (T,L,M,C は各手法を表し、その下の数字は左が n で右が p である)

第5章　母数が未知の場合の異常検出法の検出力

この章では、前章で構成した母数が未知の場合の異常検出法について、それらの条件付誤報率ならびに期待誤報率の性質を明らかにし、その結果を基に各手法の検出力の一般的性質を明らかにする。さらに、2 値変量の場合については、数値計算の結果から手法の優劣について検討する。

5.1　条件付期待誤報率の性質

ここでは、正常状態における分布の母数が未知の場合について第 4 章で構成した各方法の条件付誤報率ならびに期待誤報率に関する性質を明らかにする。これらの性質は、検出力についての議論の基礎を与えるものである。なお、以降では一般性を失わず、$p_1 \geq \cdots \geq p_I$ とする。さらに、推定方式においては、棄却限界値の決定法として F 分布法を用い Σ の推定量として最尤推定量 S/n を用いることにする。

条件付誤報率および条件付検出力を表す関数として、第 3 章で用いた記号 G_A を母数が未知の場合にも用いることにする。条件として、離散変量の水準値のほかに、初期データにおける離散変量の度数分布を与える場合についても考える。これは、異常検出統計量で用いた補正項および棄却限界値の決定において、離散変量の相対頻度を用いていることに起因している。また、手法を表す記号 A のとる値として、検定法を表す T が加わることに注意する。

初期データの離散変量の頻度 $\boldsymbol{n}$ と判定標本の離散変量の値 x が与えられたときの条件付誤報率は、推定方式については (4.6) 式より

$$G_A(x,\boldsymbol{n}) = 1 - F\left[\frac{n_x}{n_x+1}\frac{n-m-I+1+z(\boldsymbol{n})}{m}\frac{1}{n}\left\{K_A(\boldsymbol{n}) - h_A(n_x)\right\}\right], \quad A = C, L, M \tag{5.1}$$

検定方式については (4.12) 式より

$$G_T(x,\boldsymbol{n}) = 1 - F\left(\frac{n-m-I+1+z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n+1}\left(K_T(\boldsymbol{n}) - h_T(n_x)\right)\right\} - 1\right]\right) \tag{5.2}$$

となる。さらに、$\boldsymbol{n}$ の確率関数 $f(n_1,\ldots,n_I)$ で期待値をとることで、判定標本の離散変量の値 x が与えられたときの条件付誤報率は、

$$G_A(x) = \sum_{n_1,\ldots,n_I} f(n_1,\ldots,n_I) G_A(x,\boldsymbol{n}), \qquad A = C, L, M, T \tag{5.3}$$

と表現される。

まず、条件付期待誤報率 $G_A(x,\boldsymbol{n})$ について以下の性質が成り立つ。

性質 5.1　$n_i > n_j$ であれば、$G_A(i,\boldsymbol{n}) \leq G_A(j,\boldsymbol{n})$ が成り立つ。等号は、$G_A(i,\boldsymbol{n}) = 1$ のときにのみ成立する。

(証明)　$n_j = 0$ の場合は $G_A(j,\boldsymbol{n}) = 1$ なので、手法によらず不等式が成り立っている。

推定方式の場合、(5.1) 式の右辺において x の値により異なるのは補正項 $h_A(n_x)$ と $n_x/(n_x+1)$ の部分である。補正項はどの手法でも n_x の単調非増加関数であり、$n_x/(n_x+1)$ は n_x の単調増加関数なので、分布関数 $F(\cdot)$ の引数は n_x の単調非減少関数となり、(5.1) 式で与えられる条件付期待誤報率は $n_x > 0$ に対し単調非増加関数となる。検定方式の場合も $h_T(n_x)$ が n_x の減少関数であることから、(5.2) 式より n_x についての単調非増加性を確認できる。

等号が成り立つのは、$x = i$ のときの分布関数 $F(\cdot)$ の引数が 0 以下の場合のみである。このとき、$G_A(i, \boldsymbol{n}) = G_A(j, \boldsymbol{n}) = 1$ となる。

□

$p_i > p_j$ であれば、n_i の方が n_j よりも大きな値が出やすいので、性質 5.1 から $G_A(i) \leq G_A(j)$ が成立することが予想される。これが正しいことを示すのが次の性質である。

性質 5.2　$p_i > p_j$ ならば $G_A(i) < G_A(j)$ が成り立つ。

(証明)　一般性を失うことなく、$p_1 > p_2$ ならば $G_A(1) < G_A(2)$ であることを示せばよい。$(n_3, \ldots, n_I) = \boldsymbol{n}_{(12)},\ n_1 + n_2 = n_{12}$ とおけば、$\boldsymbol{n}_{(12)}$ を与えたときの n_1 の確率分布は二項分布 $B(n_{12},\ p_1/(p_1+p_2))$ である。(4.5) 式あるいは (4.11) 式により棄却限界値 $K_A(\boldsymbol{n})$ が定義されているので、$K_A(\boldsymbol{n})$ は集合 $\{n_1, \ldots, n_I\}$ を通してのみ $\boldsymbol{n}$ の関数であることがわかる。したがって

$$G_A(1, (n_1, n_2, \boldsymbol{n}_{(12)})) = G_A(2, (n_2, n_1, \boldsymbol{n}_{(12)})) \tag{5.4}$$

が成立つ。$G_A(i)$ は、$G_A(i, (n_1, n_2, \boldsymbol{n}_{(12)}))$ の $\boldsymbol{n}_{(12)}$ が与えられたときの n_1 の確率分布での期待値を、さらに $\boldsymbol{n}_{(12)}$ の確率分布で期待値をとった値である。そこで、二項分布 $B(n_{12},\ p_1/(p_1+p_2))$ の確率関数を $f_1(n_1)$ と表わすとき、

$$\sum_{n_1=0}^{n_{12}} f_1(n_1) \left\{ G_A(1, (n_1, n_2, \boldsymbol{n}_{(12)})) - G_A(2, (n_1, n_2, \boldsymbol{n}_{(12)})) \right\} \leq 0$$

を示すことにする。左辺は、(5.4) 式を用いることで、$(n_1, n_2, \boldsymbol{n}_{(12)}) = \boldsymbol{n}$ とすれば

$$\sum_{n_1 > n_2} \{f_1(n_1) - f_1(n_2)\}\{G_A(1, \boldsymbol{n}) - G_A(2, \boldsymbol{n})\}$$

と書き直される。$n_1 > n_2$ のとき、$p_1 > p_2$ より $f_1(n_1) > f_1(n_2)$ であり、性質 5.1 より $G_A(1, \boldsymbol{n}) \leq G_A(2, \boldsymbol{n})$ である。さらに、$K_A(\boldsymbol{n}) > 0$ であり、少なくとも $n_1 = n$ のとき $G_A(1, \boldsymbol{n}) < 1$ となり、$G_A(1, \boldsymbol{n}) < G_A(2, \boldsymbol{n})$ が成立つので、$G_A(1) < G_A(2)$ が確認された。

□

いま $p_1 \geq \cdots \geq p_I$ としているので、性質 5.2 から

$$G_A(1) \leq \cdots \leq G_A(I) \tag{5.5}$$

が成立することがわかる。p_x の値が相対的に小さい x が X の値として観測される場合に、条件付期待誤報率が大きいという母数が既知の場合と同様の結果が成り立っていることに注意する。この性質が条件付検出力、ひいては検出力の基本的挙動を決定づけると考えられる。

5.2 離散変量の分布のみが変化したときの検出力

第 3 章と同様に異常状態をつぎのように表わすことにする。X の確率分布を定める母数ベクトルが正常状態での値$\boldsymbol{p} = (p_1, \ldots, p_I)$から$\boldsymbol{q} = (q_1, \ldots, q_I)$に変化することと、$X = x$のときの$\boldsymbol{Y}$ の平均ベクトルが正常状態での値$\boldsymbol{\mu}_x$から$\boldsymbol{\eta}_x, x = 1, \ldots, I$に変化することとが考えられる。5.3 節でも述べるように、$\boldsymbol{\mu}_x$から$\boldsymbol{\eta}_x$への変化は、$\Delta^2(n) = n(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_x)'S^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_x)$の分布に$\psi_x = (\boldsymbol{\eta}_x-\boldsymbol{\mu}_x)'\Sigma^{-1}(\boldsymbol{\eta}_x-\boldsymbol{\mu}_x)$を通して影響するので、$\boldsymbol{\psi} = (\psi_1, \ldots, \psi_I)$とおくと、一般に異常状態を$(\boldsymbol{q}, \boldsymbol{\psi})$で表わすことができる。そのときの方法 A の検出力を $P_A(\boldsymbol{q}, \boldsymbol{\psi})$ と表わすことにする。まず、$\boldsymbol{\psi} = \boldsymbol{0}$ の場合の検出力の挙動について考える。

期待誤報率が、判定標本の離散変量の値xが与えられたときの条件付誤報率を用いて$\sum_{x=1}^{I} p_x G_A(x)$と表されるのと同様に、離散変量 X の分布の母数が $\boldsymbol{q} = (q_1, \ldots, q_I)$ へと変化したときの検出力は$\sum_{x=1}^{I} q_x G_A(x)$ で与えられる。性質 5.2 より p_x の値が小さいほど $G_A(x)$ は大きいので、p_x の値が相対的に小さい x に対し q_x の値が大きくなるとき検出力が期待誤報率よりも大きくなることが期待される。

この状況は、母数が既知の場合について 3.2.1 節で述べたのと同様である。したがって、確率的大小の概念を用いて性質 3.2 と同様、次の性質が母数が未知の場合にも成立する。

性質 5.3　$\boldsymbol{1}_I/I \succ \boldsymbol{p}$ とする。$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$ ならば $P_A(\boldsymbol{q}_1, \boldsymbol{0}) > P_A(\boldsymbol{q}_2, \boldsymbol{0})$ である。

証明は性質 3.2 の場合と全く同様であるので省略する。

性質 5.3 から、母数が未知の場合においても、$\boldsymbol{q} \succ \boldsymbol{p}$ ならば、検出力は期待誤報率よりも大きくなることがわかる。逆に $\boldsymbol{p} \succ \boldsymbol{q}$ ならば、検出力は期待誤報率よりも小さくなってしまうこともわかる。

5.3 連続変量の平均も変化したときの検出力

$X = x$ における連続変量 $\boldsymbol{Y}$ の平均が $\boldsymbol{\mu}_x$ から $\boldsymbol{\eta}_x$ に変化したとき、$\boldsymbol{n}$ および $X = x$ を与えたときの $\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_x$ の分布は、平均 $\boldsymbol{\eta}_x - \boldsymbol{\mu}_x$、分散共分散行列が $((n_x + 1)/n_x)\Sigma$ の正規分布であるので、(4.3) 式のマハラノビス平方距離を定数倍した量は、自由度が $(m,\ n - m - I + 1 + z(\boldsymbol{n}))$、非心度が $\lambda_x = (n_x/(n_x + 1))\psi_x$ の非心 F 変量の定数倍として分布することがわかる。λ_x が n_x に依存する量であることに注意する。ここで、$X = x$、$\boldsymbol{n}$ および非心度 ψ_x を与えたときの条件付検出力を $G_A(x, \boldsymbol{n}; \psi_x)$ と表わす。5.1 節の条件付誤報率の表現は平均の変化量が 0 すなわち非心度が 0 の場合なので、$G_A(x, \boldsymbol{n}) = G_A(x, \boldsymbol{n}; 0)$ であることに注意する。自由度 $(m,\ n - m - I + 1 + z(\boldsymbol{n}))$、非心度 λ の非心 F 分布の分布関数を $F(\cdot; \lambda)$ で表わせば、$X = x, \boldsymbol{n}$ および非心度 ψ_x を与えたときの条件付検出力 $G_A(x, \boldsymbol{n}; \psi_x)$ は、推定方式の場合

$$G_A(x, \boldsymbol{n}; \psi_x) = 1 - F\left[\frac{n_x}{n_x + 1}\frac{n - m - I + 1 + z(\boldsymbol{n})}{m}\frac{1}{n}\left\{K_A(\boldsymbol{n}) - h_A(n_x)\right\}; \lambda_x\right],\quad A = M, L, C \tag{5.6}$$

検定方式の場合

$$G_T(x, \boldsymbol{n}; \psi_x) = 1 - F\left(\frac{n - m - I + 1 + z(\boldsymbol{n})}{m}\left[\exp\left\{\frac{1}{n + 1}\left(K_T(\boldsymbol{n}) - h_T(n_x)\right)\right\} - 1\right]; \lambda_x\right) \tag{5.7}$$

で与えられる。これらを $\boldsymbol{n}$ の確率分布で期待値をとれば、$X = x$ が与えられたときの条件付検出力は

$$G_A(x;\psi_x) = \sum_{n_1,\ldots,n_I} f(n_1,\ldots,n_I) G_A(x,\boldsymbol{n};\psi_x), \quad A = M,L,C,T \tag{5.8}$$

となり、さらに X の分布の母数を $\boldsymbol{q}$ とすれば、検出力は

$$P_A(\boldsymbol{q},\boldsymbol{\psi}) = \sum_{x=1}^{I} q_i G_A(x;\psi_x), \quad A = M,L,C,T \tag{5.9}$$

と表現される。

$\boldsymbol{\mu}_x, x = 1,\ldots,I$ が変化したときの検出力の挙動について述べることができる一つの性質は、ψ_x についての単調増加性である。一般に、非心度 λ が大きくなるとき、非心 F 分布は確率的に大きくなる。したがって、$\boldsymbol{q}_1 = \boldsymbol{q}_2 = \boldsymbol{q}$ であり、$\boldsymbol{\psi}_1 = (\psi_{11},\ldots,\psi_{1I})$, $\boldsymbol{\psi}_2 = (\psi_{21},\ldots,\psi_{2I})$ について $\psi_{1i} \geq \psi_{2i}, i = 1,\ldots,I,$ が成立つとき、$P_A(\boldsymbol{q},\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q},\boldsymbol{\psi}_2)$ となることがわかる。

性質 5.1 で述べた $G_A(x,\boldsymbol{n})$ についての性質と同様に、$\psi_1 \leq \cdots \leq \psi_I$ とするとき、$i < j$ に対し $n_i > n_j$ のときに $G_A(i,\boldsymbol{n};\psi_i) \leq G_A(j,\boldsymbol{n};\psi_j)$ が成立すれば、つぎの性質を示すことができる：

2 つの異常状態 $(\boldsymbol{q}_1,\boldsymbol{\psi}_1)$ と $(\boldsymbol{q}_2,\boldsymbol{\psi}_2)$ について、$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$, $\psi_{1i} \geq \psi_{2i}, i = 1,\ldots,I$ かつ $\psi_{21} \leq \cdots \leq \psi_{2I}$ であれば、$P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q}_2,\boldsymbol{\psi}_2)$ である。

これに対応する性質は、正常群での母数 $\boldsymbol{p}, \boldsymbol{\mu}_x, x = 1,\ldots,I,$ および Σ が既知の場合には成立することが性質 3.2 で確認されている。残念ながら、母数が未知の場合には、F 分布の非心度 $\lambda_x = n_x\psi_x/(n_x+1)$ が n_x に依存するので、$\psi_1 \leq \cdots \leq \psi_I$ が成り立っても $G_A(x,\boldsymbol{n};\psi_x)$ については性質 5.1 に相当するものが一般には成立たないことが数値的に確認される。このことが議論を困難にしてしまっているが、n がある程度大きければ上に述べたものに近い性質が成立していると考えられる。

一つの典型的で重要と考えられるのは、ψ_x が x の値によらず同程度の大きさである場合である。ここまでの議論から、$\psi_{2i} = \psi,\ i = 1,\ldots,I,$ の場合、$\boldsymbol{q}_1 \succ \boldsymbol{q}_2$, $\psi_{1i} \geq \psi, i = 1,\ldots,I,$ ならば $P_A(\boldsymbol{q}_1,\boldsymbol{\psi}_1) \geq P_A(\boldsymbol{q}_2,\boldsymbol{\psi}_2)$ がほぼ成立していると期待される。次節に与えるものを含め数値計算の結果からも、少数の例を除き大半の場合に成立していることがわかる。

5.4 2 値変量の場合

ここでは、離散変量が 2 値の場合について検出力の数値計算を行い、各手法の持つ性質を明らかにし、どの手法が有効であるかを比較・検討する。正常状態において 2 値変量が 0,1 という値をとる確率 p_0, p_1 について、$p_0 \geq p_1$ とする。まず、異常状態において 2 値変量の分布だけが変化する場合について調べ、その後、連続変量の平均ベクトルも変化する場合について検討する。

すべての方法で期待誤報率が設定値に一致していれば、検出力の大小だけで比較することが可能である。しかし、4 章の数値計算の結果からわかるように、p_0, p_1 が未知の場合、F 分布法により棄却限界値を定めたときの期待誤報率は設定値に一致しないし、手法により期待誤報率の設定値からの乖離度が大きく異なる場合もある。そこで、期待誤報率の大きさと検出力の大きさを総合的に判断して、方法の比較を行わなければならない。ここでは、期待誤報率が設定値から大幅には乖離していない場合について、各手法の性能をオッズ比

$$\frac{\text{検出力}}{1-\text{検出力}} \times \frac{1-\text{期待誤報率}}{\text{期待誤報率}} \tag{5.10}$$

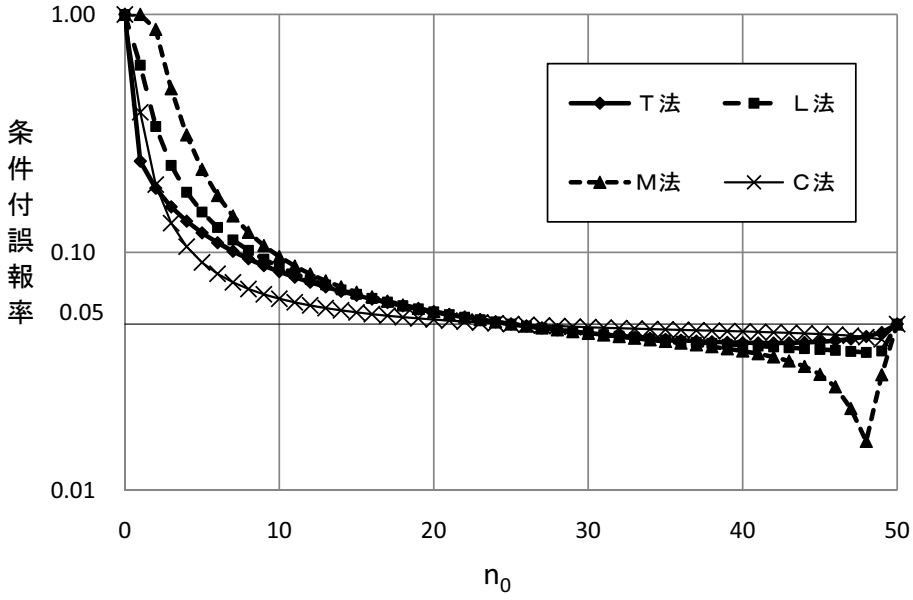

図 5.1: 条件付期待誤報率 $G_A(0,(n_0,n_1))(m=10,n=50,\alpha=0.05)$

を用いて測ることにする。これは、検出力の大きさを期待誤報率の大きさとの相対的関係で測ることになり、手法を比較するためには一つの合理的評価基準であると考える。

離散変量を取り上げるのは、その水準による連続変量の平均ベクトルの違いを考慮に入れるためのみならず、異常状態では離散変量の分布自身に変化が生じることを想定しているからである。異常検出法は、異常状態のため離散変量の分布が変化した場合にも、それを的確に検出できることが望ましい。その意味で、C 法は補正項が 0 であり、離散変量の分布の情報が生かされないので、ここでは単に比較の対象として補助的に扱うものとする。さらに、5.2、5.3 の議論から、考えている手法の検出力が大きくなるのは、異常状態における 2 値変量の確率 q_0 について $q_0 < p_0$ が成立つ場合なので、主にこの場合の検出力を基に比較を行う。

ここでは、$n=50$ と固定して $m=3,5,10$、$\alpha=0.05,0.01$ の場合について調べた。さらに、n と α による変化をみるため、$m=10,n=50,\alpha=0.001$ および $m=5,n=20,\alpha=0.05,0.01$ の場合についても調べた。

5.4.1 条件付期待誤報率および期待誤報率

$m=10$、$n=50$、$\alpha=0.05$ の場合の条件付期待誤報率 $G_A(0,(n_0,n_1))$ を、横軸に n_0 をとり示したのが図 5.1 である。n_0 の全範囲での様子と n_0 が $n=50$ に近いときの様子を一枚の図に示すため、条件付誤報率 (縦軸) は対数目盛りとした。横軸に n_1 をとったときの $G_A(1,(n_0,n_1))$ の図は、図 5.1 を $n_0=25$ について折り返した図になることに注意する。異常検出法の定式化より条件付誤報率 $G_A(0,(n_0,n_1))$ は、$n_0=0$ で 1 であり、n_0 が増加するとともに減少し $n_0=25=n_1$ で $\alpha=0.05$ となる。その後も n_0 とともに減少し、50 の手前で最小値をとり、50 では α となる。

推定方式の中では、2 水準間の補正項の差 $h_A(n_0)-h_A(n_1)$ の値の大きさの違いを反映して、$G_A(0,(n_0,n_1))$ の値は $1 \le n_0 < n/2$ の範囲では M 法、L 法、C 法の順に大きく、$n_0 > n/2$ ではその逆の大小関係になる。T 法では $G_T(0,(n_0,n_1))$ の値は、$n_0=0$ および $n_0=50$ に近い場合を除いて L 法と C 法の間にあり、どちらかというと L 法に近い挙動をしている。n_0 の値が n に近い

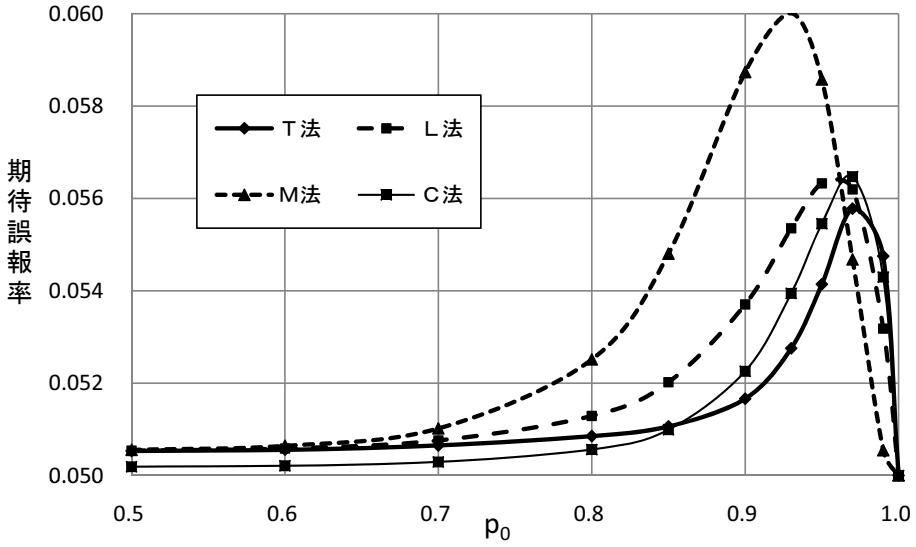

図 5.2: 期待誤報率 $(m = 10, n = 50, \alpha = 0.05)$

とき、M 法の条件付期待誤報率が極端に小さい値をとっていることに注意する。これは、補正項の差が他の方法に比べて極端に大きくなることによると考えられるが、後に述べる検出力の大きさに影響を与えることになる。

p_0 の関数としての期待誤報率 $p_0 G_A(0) + p_1 G_A(1)$ は 4.3.2 に与えられているが、ここで必要な部分を再度図 5.2 に与える。いずれの手法でも、全般に期待誤報率は設定値よりも多少大きく、p_0 の値が 0.9 を超えたところにピークがあるが、M 法の場合は特に大きな乖離が見られる。期待誤報率が設定値を大幅に超える場合があるならば、そのような手法を異常検出法として用いることには問題があるというべきである。p_0 の値が 0.9 前後の広い範囲で M 法の期待誤報率は設定値よりも 1 割以上大きくなるので、異常検出法としては適当とは考えにくい。

ここまでの議論を踏まえて、以降は T 法と L 法の比較を中心に議論することにする。

5.4.2 離散変量の分布のみが変化したときの検出力

数値計算の結果の一例を図 5.3 に示すが、一般につぎのことがわかった。推定方式に限れば、$q_0 < p_0$ の場合は M 法が最もオッズ比が大きく、次いで L 法、C 法の順となる。逆に、$q_0 > p_0$ の場合はオッズ比も逆の大小関係になる。$q_0 = p_0$ の場合、検出力と期待誤報率は等しくなるので、全ての手法でオッズ比は当然 1 となる。方法ごとの 2 水準間での補正項の差の大小を考えると、このような結果になることは納得できるものである。なお、母数が既知の場合、期待誤報率は設定値と一致するので検出力だけを問題にすればよいが、3.2.2 に与えられた母数が既知の場合の検出力の挙動についての議論と、これらの結果は同様の結論を与えている。T 法の場合オッズ比は、補正項 $h_T(x)$ の検定方法へのかかわり方が異なるので一概には言えないが、多くの場合 C 法と L 法の中間的挙動をする。

5.4.3 連続変量の平均も変化したときの検出力

ここでは、2 値変量の値が x であるときの平均の変化の度合いを測る量である ψ_x が、x の値によらず一定の値 ψ である場合について議論する。前にも述べたが、新型インフルエンザ感染での海外

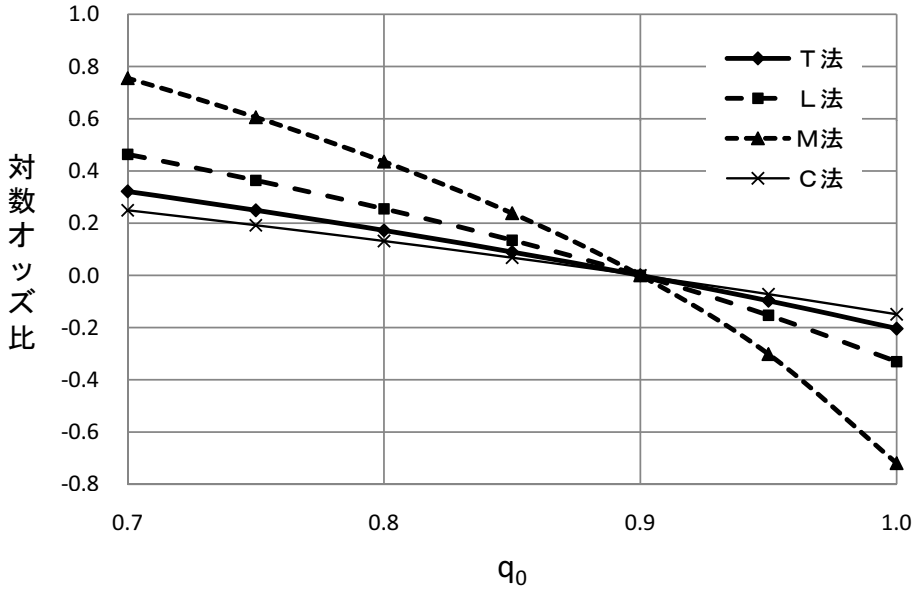

図 5.3: 離散変量の分布が変化したときの対数オッズ比 $[p_0 = 0.9](m = 10, n = 50, \alpha = 0.05)$

渡航歴の有無は感染のリスク要因の一つであるが、感染したときの症状には海外渡航歴の有無によりそれほど大きな違いはないと考えられる。このような状況では、異常状態での連続変量の平均の変化量は、離散変量の値によらないと考えてよい。

$m = 10, n = 50, \alpha = 0.05$ の場合において、条件付検出力 $G_A(0, (n_0, n_1); \psi)$ を示したのが図 5.4、5.5、5.6 である。ψ の値は、10,20,30 と選んでいる。これらの図から、M 法では n_0 が $n = 50$ に近いところで条件付検出力が減少しているのが見て取れる。これは、5.4.1 で示したように条件付期待誤報率 $G_M(0, (n_0, n_1))$ が同じように極端に小さくなることに起因している。また、n_0 が 1 に近いところで $G_T(0, (n_0, n_1); \psi)$ の値が大きく減少し、ψ の値が大きいときには 5.3 で述べたように、$n_0 < n_1$ でも $G_T(0, (n_0, n_1); \psi) \geq G_T(1, (n_0, n_1); \psi)$ が成立たないことがわかる。これは、(5.7) 式において、n_1 が小さいとき、非心 F 分布の非心度 $n_1\psi/(n_1 + 1)$ が小さくなることが原因と考えられる。

$p_0 = 0.9$ の場合に x の値を与えたときの条件付検出力 $G_A(x; \psi)$ を、横軸に ψ をとり描いたのが図 5.7 である。各方法で $x = 1$ の場合の曲線が上に、$X = 0$ の場合が下にある。 8 本の曲線の中で真ん中に位置する 2 本の実線が C 法の場合であり、それから上下に隔たっていく形で T 法、L 法、M 法が位置している。少し見難いが、C 法の $X = 0$ の検出力は T 法のそれとほぼ一致している。

この図から、L 法,M 法,T 法は C 法に比べて、$x = 1$ の場合における検出力の増加が、$x = 0$ の場合における検出力の減少に比べて大きいことがわかる。L 法、M 法、T 法は、C 法と比べて $x = 0$ での条件付き検出力を多少犠牲にして、$x = 1$ での条件付き検出力を大きくすることで、全体の検出力を大きくしようとする方法であることがわかる。これは p_0 の値が大きいときにより顕著になる。

各手法での検出力は、図 5.7 の各手法の上下の条件付検出力の曲線の値を異常時の離散変量の確率 q_0, q_1 で重みをつけた平均となる。したがって、この図から一般に q_0 が大きいときは C,T,L,M の順に、q_1 が大きいときは M,L,T,C の順に検出力が大きくなると予想され、離散変量の確率の値により検出力の意味で優れた手法が入れ代ることになる。しかし、手法の良さを比較するオッズ比には誤報率の値も影響を与えているため、このことだけから手法の優劣を判定することはできないことに注意する。

先に示した各 m, n, α の組合せについて、ψ の値を 0,2,5,10,15,20,25,30,40 と変化させ、q_0 を 0.00

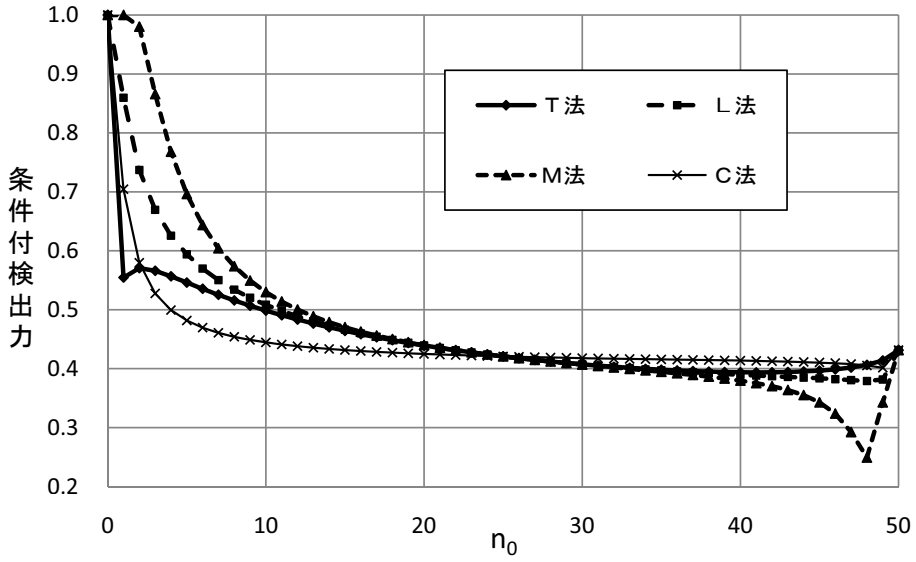

図 5.4: 条件付検出力 $G_A(0,(n_0,n_1);10)(m=10,n=50,\alpha=0.05)$

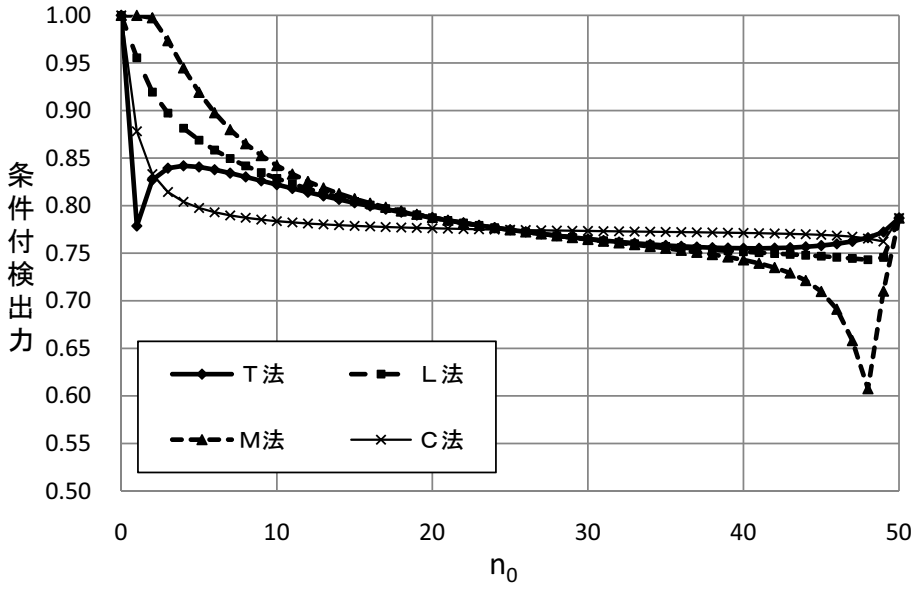

図 5.5: 条件付検出力 $G_A(0,(n_0,n_1);20)(m=10,n=50,\alpha=0.05)$

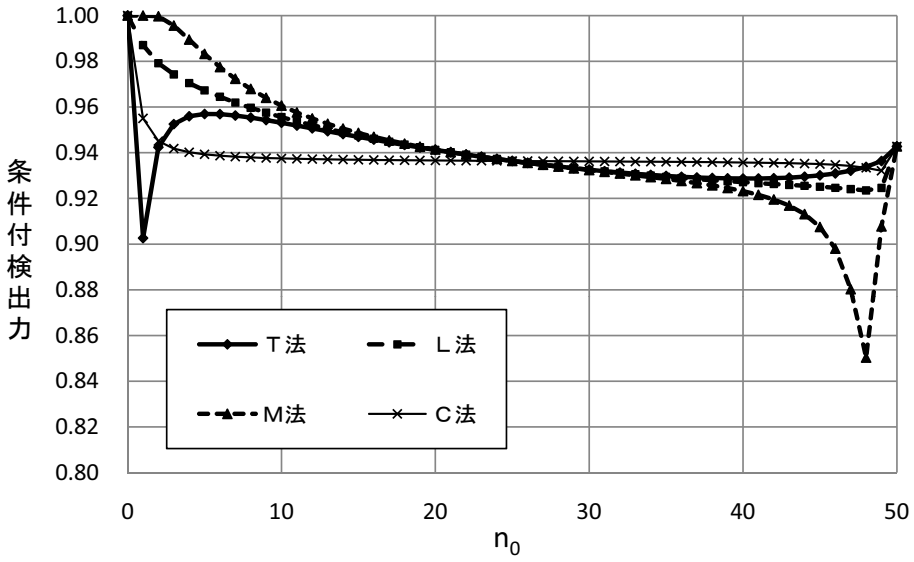

図 5.6: 条件付検出力 $G_A(0,(n_0,n_1);30)(m=10,n=50,\alpha=0.05)$

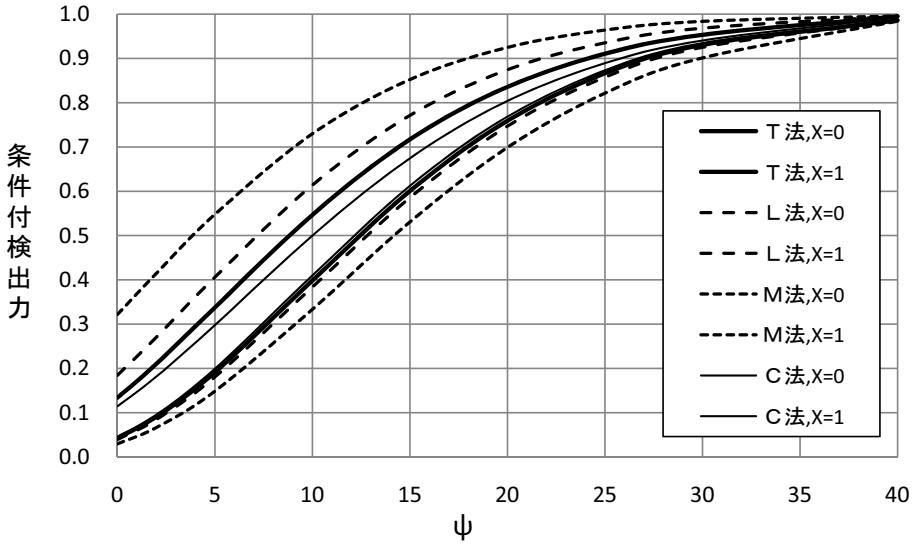

図 5.7: 条件付検出力 $G_A(x;\psi)(m=10,n=50,\alpha=0.05,p_0=0.9)$

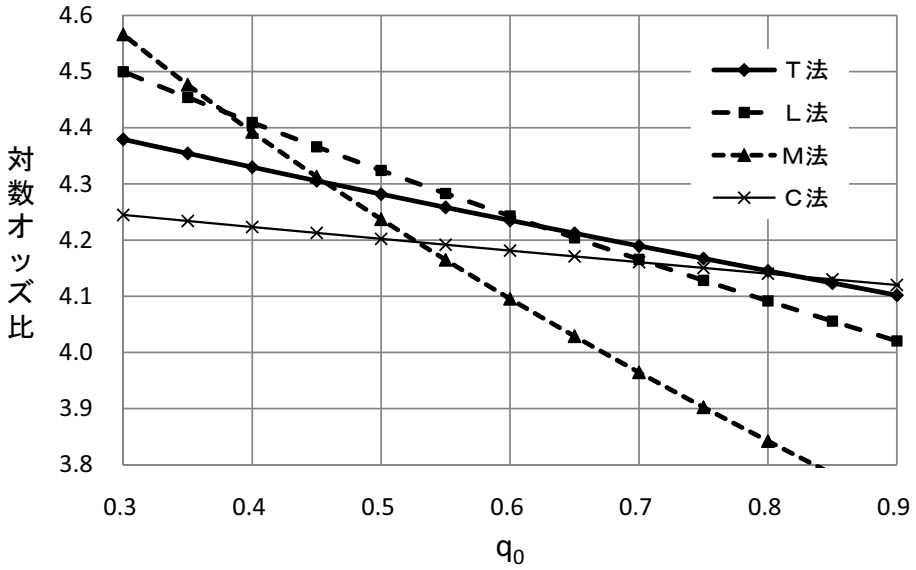

図 5.8: 離散変量の分布が変化したときの対数オッズ比 $[p_0 = 0.9, \psi = 20](m = 10, n = 50, \alpha = 0.05)$

から 1.00 まで 0.05 刻みで動かして、各手法について検出力を計算し、オッズ比を用いて手法を比較した。q_0 を横軸にとり、各手法についてオッズ比の挙動を示す一例が図 5.8 である。$0.3 < q_0 < 0.9$ の範囲に着目すると、一般にオッズ比を最大にするという意味で最適な手法は、p_0, ψ の値を固定するとき q_0 の大きさにより変化し、q_0 の値が小さくなるにつれ C 法、T 法、L 法、M 法と入れ替わっていく。M 法は、ψ の値が 0 に近いところでは優れているが、ψ の値の増加とともに M 法が最適であるような q_0 の値は、p_0 よりかなり小さい値の狭い範囲に限定されてしまうことがわかった。また、C 法は、$q_0 > p_0$ のとき優れているが、これは本研究で目標としている領域ではないことに注意する。$q_0 < p_0$ という主眼とする領域の広い範囲で、T 法と L 法が最適でない場合でも、最適な手法に近いオッズ比を与えるという意味で、優れた手法であることがわかった。

T 法および L 法の最適性が入れ替わる q_0 の値を、T 法と L 法の境界確率とよぶ。$m = 5, 10, n = 50, \alpha = 0.05$ の場合に各 ψ の値における境界確率を p_0 に対し示したのが図 5.9 である。各曲線の上部で T 法、下部では L 法のオッズ比が大きい。境界確率の値は、計算した範囲では n の値が大きく、M の値が小さく、α の値が小さくなるときに小さくなり、T 法が最適である領域が広がることが確認できた。例えば、$m = 10, \alpha = 0.05$ のとき、$0.7 \leq p_0 \leq 0.9$ の範囲で、境界確率の値は、$\psi = 10$ のとき $0.8p_0$ 程度、$\psi = 20$ のとき $0.7p_0$ 程度である。$m = 3$ のときには $\psi = 5$ でも $0.65p_0$ 程度にまで小さくなり、T 法が最適である範囲が広がる。したがって、変数が少なく期待誤報率の設定値が小さいときは T 法を用いるのが特に有利となる。

5.4.4 まとめ

以上、本節では、2 値変量が混在するロケーションモデルで、分布の母数が未知の場合の異常検出問題について議論してきた。異常検出のための手法の優劣を比較する場合に留意すべき点がいくつかある。

まず、期待誤報率に関しては、その安定性が問題となる。期待誤報率が設定された値から大きく変化しないことは、手法を用いる際の重要な要素である。M 法では、p_0 が 0.9 を越えたときに期待

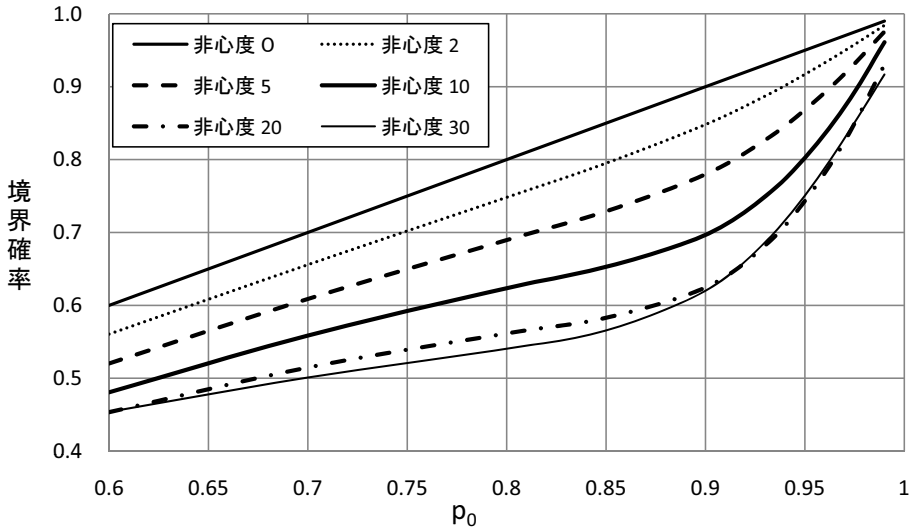

図 5.9: 境界確率 [T 法と L 法]$(m = 10, n = 50, \alpha = 0.05)$

誤報率が他の方法に比べて設定値 α から大きく離れることが数値計算で確認されている。M 法は、期待誤報率の安定性の面で問題があると考えられる。

つぎに、検出力に関しては、まず第一に 2 値変量の分布が正常状態から変化したとき、それをどれほど正しく検出できるかが問題になる。2 値変量で確率が小さい水準は全体として少数部分である。この少数部分の割合が異常状態で増加するとき、これを確実に発見するのも大事なことであり、本論文が目指しているところでもある。C 法はこの点からみて不十分であると考えられる。

最後に、連続変量の平均も変化した場合については、平均が大きく変化する場合はどの方法を用いてもほぼ確実に異常を検出できるので、ある程度の平均の変化に対する検出力が問題になる。この観点からみて、T 法は L 法に比べて 2 値変量の分布の変化がそれほど大きくない場合には優れていると考えられる。

5.5 結論

本章では、分布の母数が未知の場合における 4 つの異常検出法の条件付誤報率および検出力の性質を明らかにした。その結果を用いて、離散変量の分布のみが変化した場合および連続変量の平均も変化した場合の検出力の挙動に関する基本的性質を一般的に明らかにした。

4 つの異常検出法について誤報率や検出力の挙動を一般的に比較するのは困難なので、2 値変量の場合についてその挙動を数値計算により調べた。4 章の結果も併せて総合的に結論を述べることにする。棄却限界値の決定については、χ^2 分布法よりも F 分布法の方が期待誤報率が設定値に近いことが確認された。また、用いる Σ の推定量による違いはそれほど大きくなかったが、最尤推定量を用いた場合が期待誤報率が設定値に近いことがわかった。さらに、正条件における期待誤報率については、T 法が優れて安定していることも確認された。離散変量の値を与えたときの条件付検出力については、離散変量の情報を積極的に用いない C 法と比較すると、他の手法は $X = 0$ における条件付検出力を少し犠牲にして $X = 1$ での条件付検出力を大きくする手法であることがわかった。

検出力の挙動について手法間の比較を一般的に行うのは困難なので、各水準での非心度が等しい場合を取り上げ、数値計算に基づいて比較した。また、分布の母数が未知の場合は期待誤報率を設

定値に一致させることができず、検出力だけを用いての比較は公平でないので、期待誤報率に対するオッズ比を用いて行った。ここで、C 法は離散変量の情報を積極的に利用しないこと、M 法は期待誤報率の変動が大きいことを考慮し、T 法と L 法を中心に比較した。その結果、非心度がある程度大きいときは T 法が L 法に比べて優位である 2 値変量の確率の範囲が広く、T 法が優れていることが確認できた。

第6章　結論

6.1　離散変量が混在するときの異常検出法

本論文では、連続変量と離散変量が混在する異常検出問題について議論した。ここで取り上げる離散変量は、異常を起こしやすい状況や異常発生を示唆するなど、何らかの意味で異常検出に寄与しうるものであることを想定している。このような変量を適切に活用することで、異常を素早く確実に検出することを目指している。離散変量の値を与えたとき、連続変量が分散共分散行列が共通の正規分布にしたがうとするロケーションモデルを仮定し、分布の母数が既知の場合と未知の場合について、異常検出法を構成した。正常であるのに異常と判定する誤報率あるいは期待誤報率が、設定値に一致する、あるいは、なるべく近い値になるように棄却限界値を定める方法が与えられた。さらに、誤報率あるいは期待誤報率および正しく異常を検出する検出力の性質を明らかにし、手法間の比較を行った。

6.2　分布の母数が既知の場合

分布の母数が既知の場合は、離散変量をダミー変数として連続変量と併せて求めるマハラノビス平方距離を用いるマハラノビス距離法 (M 法) と、全変量を用いた尤度比検定に基づく尤度比法 (L 法) を構成した。この 2 つの方法に、離散変量の値を与えた条件の下での連続変量の平均ベクトルについての検定による条件付法 (C 法) を加えた 3 方法について議論した。

離散変量が 1 つの場合、全変量に基づくマハラノビス平方距離が離散変量 $X = x$ のとき連続変量のみによるマハラノビス平方距離と $(1 - p_x^{(0)})/p_x^{(0)}$ という補正項の和として表現されることが示された。ここで、$p_x^{(0)}$ は正常状態で $X = x$ となる確率である。また、複数の 2 値変量が存在する場合、離散変量の値の連続変量の平均への効果に加法性があるときは、連続変量のみによるマハラノビス平方距離と離散変量の確率に基づく補正項の和として表現されることが示された。この結果から、誤報率が正確に設定値に一致する異常検出法を構成したのが M 法である。

全変量を用いた尤度比検定に基づく L 法の異常検出統計量は、M 法と同様に連続変量のみによるマハラノビス平方距離と補正項 $-2\log(p_x^{(0)})$ の和になることが示された。なお、C 法は補正項を 0 とした場合と考えられる。

それぞれの手法について誤報率が正確に設定値に一致するような棄却限界値の定め方を与えた。M 法と L 法では離散変量の水準により補正項の値が異なるため、判定標本の離散変量 X の値 x を与えたときの条件付誤報率は水準により異なる。すなわち確率が小さい水準での補正項が大きいことから、どちらの方法においても、正常状態で確率が小さい水準ほど異常と判定しやすくなることがわかる。なお、C 法の場合の条件付誤報率は、どの水準でも設定した誤報率に一致している。

条件付誤報率および検出力の挙動について一般的に手法間で比較することは困難なので、2 値変量の場合について数値計算を基に比較した。2 つの水準間での補正項の差は、M 法が最も大きく、次いで L 法、C 法 $(= 0)$ である。これより、2 つの水準を $0, 1$ とし、$p_0^{(0)} > \frac{1}{2}$ とすると $X = 1$ での条

件付誤報率は大きい順に M 法、L 法、C 法となり、$X = 0$ での条件付誤報率の大きさは、逆の順になる。また、M 法では、$p_1^{(0)}$ が誤報率の設定値に近いとき、$X = 0$ における条件付誤報率が極端に小さい値になりうることが確認された。

異常状態において離散変量の分布のみが q_0, q_1 に変化したときの検出力は、条件付誤報率をこの分布で平均をとることで得られる。したがって、$q_0 < p_0^{(0)}$ の場合の検出力は大きい順に、M 法、L 法、C 法であり、$q_0 > p_0^{(0)}$ の場合は C 法、L 法、M 法の順となる。

連続変量の分布も変化した場合の条件付検出力は、平均の変化量を表す非心度の関数として与えられる。また、非心度を固定したときの条件付検出力の値は、条件付誤報率と同様 $X = 1$ では大きい順に M 法、L 法、C 法であり、$X = 0$ ではその逆の順になる。3 つの手法における検出力の比較には、X の値による非心度の違いや、X の分布の変化状況により優劣が変わるため、一般的に比較するのは困難である。本論文では、X の各水準で非心度が等しい場合について、手法による検出力の挙動の違いについて考察した。非心度を固定したとき、q_0 が 1 に近いときは C 法が、q_0 が 0 に近いときは M 法が最も検出力が大きく、その中間に L 法が最適となる領域が存在する。ある程度非心度が大きい場合、q_0 が $p_0^{(0)}$ よりある程度減少する範囲で L 法が最適であることが確認できた。また、C 法は離散変量を異常検出に積極的に用いていないこと、M 法では条件付誤報率が非常に低くなる場合があるなどの問題点がある。以上より、L 法が母数の広い範囲で安定して検出力の高い方法であると考えられる。

6.3 分布の母数が未知の場合

分布の母数が未知の場合には、正常であることがわかっている個体についての観測値 (初期データ) が得られるものとして異常検出法を構成した。一つの構成法は、母数が既知の場合の異常検出統計量に、初期データによる分布の母数の推定量を代入するという推定方式である。推定方式による 3 手法 (C 法、L 法、M 法）に加えて、初期データに判定標本を合わせた全データに対する検定を行う検定方式により検定法 (T 法) を構成した。棄却限界値は、初期データについて期待値をとった期待誤報率が設定値に近くなるよう決定する。そのとき、連続変量のみに基づくマハラノビス平方距離の分布として χ^2 分布ではなくて、より正確な F 分布に基づき棄却限界値を定めることが、全ての手法において設定値に近い期待誤報率を与えることが確認された。

判定標本の離散変量の水準 $X = x$ および初期データにおける離散変量の観測度数を与えたときの条件付期待誤報率の性質を示し、さらに、判定標本の離散変量の値を与えたときの条件付期待誤報率の性質を明らかにした。2 値変量の場合の数値計算の結果から、推定方式における方法間で、母数が既知の場合と同様に確率が小さい水準では条件付期待誤報率は大きい順に M 法、L 法、C 法であり、確率が大きい水準では逆の順になることが確認された。T 法は状況により多少変化するが、おおむね L 法に近い挙動をすることがわかった。また、2 値変量の場合に双方の水準で観測値が得られるという正条件の下では T 法の期待誤報率が他と比べて安定して設定値に近いことが確認できた。

母数が未知の場合は期待誤報率が設定値に一致しないので、検出力だけでは公平な比較ができない。ここでは、期待誤報率に対する検出力のオッズ比を基準として用いて比較を行った。ただし、期待誤報率が設定値から大きく離れることはそれ自体大きな問題であることに注意する。

離散変量の分布のみが変化した場合については、どの手法においても、正常状態において確率が小さい水準の確率が異常状態で増加するとき、オッズ比が増加することがわかった。2 値変量の場合についての数値計算の結果を基に手法間の比較を行ったが、推定方式に限れば分布が既知の場合と同様に、$p_0 > 1/2$ に対し、$q_0 < p_0$ の場合のオッズ比は、大きい順に M 法、L 法、C 法であり、

$q_0 > p_0$ の場合は逆の順になった。T 法については、検定統計量への補正項のかかわり方が推定方式と異なるので一概には言えないが、多くの場合に L 法と C 法の中間的挙動をしていた。

連続変量の平均も変化した場合、観測度数が極端に小さい水準については、観測度数および離散変量の値を与えたときの条件付検出力が、小さくなる場合があることがわかった。しかし、初期データがある程度大きければ、このようなことが起こる可能性は低くなると考えるので、これはそれほど大きな問題ではないと考えられる。L 法、M 法、T 法は、正常状態における各水準の確率について $p_0 > p_1$ とするとき、$X = 0$ での条件付検出力を多少犠牲にして、$X = 1$ での条件付検出力を大きくする方法であることが確認できた。

C 法は離散変量を積極的に異常検出に用いていないことから、また、M 法は期待誤報率が設定値から大きく乖離する可能性があることから、手法の総合的な比較は L 法と T 法について行った。連続変量の平均の非心度が等しいときに 2 つの手法の優劣が入れ替わる離散変量の確率の境界値を調べた結果、異常状態である程度平均が変化する場合は、広い範囲で T 法が優れていることが確認できた。

2 値変量の数値計算の結果からではあるが、期待誤報率が特に正条件の下で設定値に近いという安定性と、期待誤報率に対する検出力のオッズ比に基づく比較での広い範囲での優越性から、T 法が異常検出において安心して使える優れた手法であると結論付けられる。

参考文献

[1] Abraham, B. and Variyath, A. M. (2003): Discussion on “ A review and analysis of the Mahalanobis-Taguchi system, ” *Technometrics*, **45** , 1, 22-24.

[2] Alt, F. B.(1985): “ Multivariate quality control, ” *Encyclopedia of Statistical Science*, 6 (Kotz, S. and Johnson, N., eds.), John Wiley & Sons, Inc., 110-122.

[3] Bar-Hen, A. and Daudin, J. J. (1995): “ Generalization of the Mahalanobis distance in the mixed case, ” *Jour. of Multivariate Analysis*, **53**, 332-342.

[4] Champ, C. W., Jones, L. A. and Rigdon, S. E. (2005): “ Properties of the T^2 control chart when parameters are estimated, ” *Technometrics*, **47**, 437-445.

[5] Das Gupta, S. and Perlman, M. D. (1974): “ Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2-test, ” *J. Amer. Statist. Assoc.*, **69**, 345, 174-180.

[6] 長谷川良子（2004）:『マハラノビス・タグチ（MT）システムのはなし』, 日科技連出版社.

[7] 飯田孝久, 福島崇博, 篠崎信雄 (2008):“ 2 値変量が混在する場合のマハラノビス距離による異常検出, ”「応用統計学」,**37**,2,55-76.

[8] 飯田孝久, 福島崇博, 篠崎信雄 (2009):“ 離散変量と連続変量が混在する場合の尤度比検定法による異常検出, ”「品質」,**39**,3,102-111.

[9] 飯田孝久, 篠崎信雄 (2010) :“ 離散変量と連続変量が混在する異常検出問題における母数が未知のときの期待誤報率, ”「品質」,**41**,1, 121-130

[10] 飯田孝久, 篠崎信雄 (2011) :“ 離散変量と連続変量が混在する異常検出問題における母数が未知のときの検出力, ”「品質」, **41**,3, 採択済み.

[11] Jiang, W. and Tsui, K. L.(2008) : “ A theoretical framework and efficiency study of multivariate statistical process control charts, ” *IIE Transactions*, 40, 650-663.

[12] 兼高達貳 (1987):“ マハラノビスの汎距離の応用例-特殊健康診断の事例, ”「標準化と品質管理」, **40**, 10, 57-64.

[13] Krzanowski, W. J. (1975): “ Discrimination and classification using both binary and continuous variables, ” *J. Amer. Statist. Assoc.*, **70**, 782-790.

[14] Krzanowski, W. J. (1980): “ Mixture of continuous and categorical variables in discriminant analysis, ” *Biometrics*, **36**, 493-499.

[15] Krzanowski, W. J. (1982): “ Mixture of continuous and categorical variables in discriminant analysis :A hypothesis-testing approach, ” *Biometrics*, **38**, 991-1002.

[16] Krzanowski, W. J. (1983): “ Distance between populations using mixed continuous and categorical variables, ” *Biometrika*, **70**, 235-243.

[17] Krzanowski, W. J. (1986): “ Multiple discriminant analysis in the presence of mixed continuous and categorical data, ” *Comp. Math. Appl.*, **12A**, 179-185.

[18] Mason, R. L. and Young, J. C. (2002): Multivariate Statistical Process Control with Industrial Applications, Philadelphia : ASA-SIAM.

[19] 宮川雅巳 (2000):『品質を獲得する技術』, 日科技連出版社.

[20] 宮川雅巳 (2003):“ SQC から見たタグチメソッド, ”「品質」 **33**,1,27-35.

[21] 宮川雅巳, 田中研太郎, 岩澤智之, 中西寛子 (2007):“ マハラノビス・タグチ・システムにおける実際の誤報率, ”「品質」,**37**,1,101-106.

[22] 宮川雅巳・永田靖 (2003):“ マハラノビス・タグチ・システムにおける多重共線性対策について, ”「品質」, **33**, 4, 77-85.

[23] 永田靖, 土居大地（2009）:“ タグチの RT 法で用いる距離の性質とその改良, ”「品質」 , **39**,3, 90 － 101.

[24] 永田靖, 久富剛（2008）:“ 項目数が $n-1$ 以上の場合の MT システムの第 1 種の距離, ”「品質」, **38**,1, 142 － 146.

[25] 中西寛子 (1999):“ 離散と連続変量が混在する場合の距離と誤報率, ”「応用統計学」, **28**, 2, 79-89.

[26] 中西寛子, 加藤貴一 (2008):“ 離散と連続変量が混在する場合のマハラノビス・タグチ・システム, ”「*JSQC* 第 86 回研究発表会要旨集」, 171-174.

[27] Nakanishi, H. (2003): “ Tests of hypotheses for the distance between populations on the mixture of categorical and continuous variables, ” *J. Jpn. Soc. Comp. Statist.*,**16**, 53-62.

[28] 日本規格協会編 (1993):『JIS ハンドブック 14　品質管理』, 日本規格協会.

[29] 仁科健 (2009):『統計的工程管理』, シリーズ現代の品質管理３, 日本規格協会.

[30] Olkin, I. and Tate, R. F. (1961): “ Multivariate correlation models with mixed discrete and continuous variables, ” *Ann. Math. Statist.*, **32** , 448-465.

[31] Seber(2008):*A Matrix Handbook for Statisticians*, John Wiely & Sons, Inc.

[32] 田口玄一 (1993):“ 品質工学入門　連載 24　医学部門への応用, ”「標準化と品質管理」, **46**,5,88-94.

[33] 田口玄一 (2002):“ 20 世紀の MTS 法と 21 世紀の MT 法, ”「標準化と品質管理」, **55**, 2, 61-70.

[34] 田口玄一, 兼高達貳 (2002): 『MT システムにおける技術開発』, 日本規格協会.

[35] Taguchi, G ., and Rajesh, J. (2000): “ New trends in multivariate diagnosis, ” *Sankhya*, Series B, **62**, 233-248.

[36] 竹村彰通 (1991):『多変量推測統計の基礎』, 共立出版.

[37] 立林和夫 (2004):『入門タグチメソッド』, 日科技連出版社.

[38] 立林和夫 編著, 手島昌一, 長谷川良子 著（2008）:『入門 MT システム』, 日科技連出版社.

[39] 椿広計, 河村敏彦（2008）:『設計科学におけるタグチメソッド』, 日科技連出版社.

[40] Tracy, N. D., Young, J. C. and Mason, R. L. (1992): “ Multivariate control charts for individual observations, ” *Jour. Quality Technology*, **24** , 2 , 88-95.

[41] Wierda, S. J. and Steerneman, T. (1995): “ Power properties of the T^2 control chart, ” *International J. of Reliability, Quality and Safety Engineering*, **2**, 1, 1-14.

[42] Woodall, W. H., Koudelic, R., Tsui, K., Kim, B., Stoumbos, Z. G. and Carvounis, C. P. (2003): “ A review and analysis of the Mahalanobis-Taguchi system, ” *Technometrics*, **45**, 1, 1-30 (with discussions).

謝辞

本論文の執筆にあたり、主査である篠崎信雄教授には、長期間にわたり懇切丁寧に御指導していただきました。深く感謝の意を表します。

また、本論文のテーマの発端になる修士論文を著し、共著者としても協力いただいた福島崇博君に感謝します。

副査をお願いした清水邦夫教授、櫻井彰人教授、鈴木秀男教授には、有益な助言をいただき、論文をさらに充実させることができました。ありがとうございました。

最後に、私を研究者の道に導いてくれた、鷲尾泰俊名誉教授ならびに故坂元平八名誉教授に感謝します。