

厚生労働科学研究費補助金

健康科学総合研究事業

健康度の測定法及び計算式の開発に関する研究

平成15年度 総括・分担報告書

主任研究者 川村則行

平成16(2004)年 3月

目 次

I 総括研究報告		
健康度の測定法及び計算式の開発に関する研究	川村則行	1
I-A 研究協力者報告書		
本報告書において用いられた統計手法の解説	宮崎隆穂	9
管理検査の研究に使用する統計手法について	吉野 陸	12
うつデータを用いたマハラノビス・田口の方法における変数選択に関する検討	吉野陸 川村則行	22
ストレス脆弱性モデルについての解説	澁谷美穂子	31
職域に於けるストレスと健康 ～ストレス因子の解明～	杉本日出子 川村則行	37
職域に於けるストレスと健康 ～CAID分析とストレス因子の解明～	杉本日出子	42
職域に於けるストレスと健康 ～ストレス対策～	杉本日出子	44
遺伝疫学の研究手法	酒見正太郎 小牧元 川村則行	50
II 分担研究報告		
労働者の喫煙習慣と免疫系	中田光紀	54
III 研究成果の刊行に関する一覧表		63
IV 研究成果の刊行物・別刷り		65

厚生労働科学研究費補助金(健康科学総合研究事業)
総括研究報告書
健康度の測定法及び計算式の開発に関する研究

主任研究者 川村則行

国立精神・神経センター 精神保健研究所
心身医学研究部

研究要旨

過去の研究において, Exposure(曝露)として800項目の質問紙の情報(生育歴, スレッサー, 気質, 性格, 服薬歴, 緩衝要因), Surrogate(エンドポイントの代用:細胞性および液性免疫, 健康診断データ)と Outcome(エンドポイント:疾病休業日数, 各種生活習慣病の発症)に関する情報を蓄積した. これらを用い, 個人の健康度の指標を表現する目的で, Exposure, Surrogate と Outcome の subcategory の関係を繋ぐ数式・アルゴリズムを, discriminant analysis と general linear model を用いて, 非公開で作成した. 今年度はこの統計手法を改め, Maharanobis-田口メソッドや生存分析を用いた. 数式・アルゴリズムが, 同一集団および異なる集団にて, Surrogate と Outcome の subcategory の予測に役立つか否かを検証し, 適宜, 修正を加え, 簡略化し, 実用に耐えうる, 健康度の指標として確立することを目指す. そのために, (1)Exposure, Surrogate と Outcome の精度を高めること (2)Exposure に遺伝子多型の情報を含めること, (3)Exposure のうち主要なものに, 簡便な介入を加え, Surrogate と Outcome が, 介入後も同一の数式・アルゴリズムで表現できるか否か検討すること(4)非線形モデルを数式・アルゴリズムに活用すべきか否かの検討を行うこと, を4本柱として研究を進める. 本研究の主なる成果である健康度:数式・アルゴリズムは, 厚生労働省の指導の元で開発し特許化する. 副次的なる成果として, Exposure, Surrogate と Outcome の相互関係や因果関係に関する新しい知見は, 国際学術雑誌に投稿し, 英文・日本語の書籍として刊行する. 数式については, 平成 14年度の報告書にて開示する方針であったが, 実用化に向けての社会的手続きを考案中で, 一般には開示しない事に決定した. 数式を記載することは, 特許化する目的に反するので記載しない点につき了解を得たい. 総括報告書中の図の緩衝要因(Buffering System)に該当する項目の多くが, スレッサーのレギュレーションを変え, 健康度数式の分岐点を形成することが明らかとなった. 今後, この点に注意して, 開発を進めたい.

研究協力者

宮崎隆穂

長寿科学振興財団 リサーチレジデント
学術振興会特別研究員
精神保健研究所心身医学研究部

酒見正太郎

東京大学公衆衛生大学院

井筒節

東京大学健康科学保健学科
精神衛生

富岡光直

九州大学心療内科

石川俊男

国立精神・神経センター国府台病院
心療内科

長谷川明弘

東京都立大学都市科学大学院

志村翠

国立精神・神経センター
精神保健研究所 心身医学研究部

中田光紀

産業医学総合研究所

分担研究者

小牧元

国立精神・神経センター
精神保健研究所 心身医学研究部

中田光紀

産業医学総合研究所

研究協力者

前田基成

女子美術大学芸術学部

有村達之

九州大学大学院心身医学

篠田晴男 立正大学

緒方一子

営団地下鉄総合生活相談室

久保千春

九州大学大学院医学院心身医学

A 研究目的

過去の研究において、Exposure (曝露)として 800 項目の質問紙の情報(生育歴, スレッサー, 気質, 性格, 服薬歴, 緩衝要因), Surrogate (エンドポイントの代用: 細胞性および液性免疫, 健康診断データ)と Outcome (エンドポイント: 疾病休業日数, 各種生活習慣病の発症)に関する情報を蓄積した。

これらを用い、個人の健康度の指標を表現する目的で、Exposure, Surrogate と Outcome の subcategory の関係を繋ぐ数式・アルゴリズムを、discriminant analysis と general linear model を用いて、非公開で作成した。この数式・アルゴリズムが、同一集団および異なる集団にて、Surrogate と Outcome の subcategory の予測に役立つか否かを検証し、適宜、修正を加え、簡略化し、実用に耐えうる、健康度の指標として確立することを目指す。

主任研究者は、The rise in medical care expenditures in Japan: 1977-81. Am J Public Health. 1985 において、国民の2/3が罹患する、がんおよび循環器の死に至る病が日本の医療費の多くを消費することを示した。

昨今では、これらを含む、生活習慣病の予防法の開発が急務となった。

この事業に関わる、本研究の最終目的は、21世紀に望まれる order-made preventive medicine の構築である。

その第一段階として、いかなる個人からも、簡便に得ることの出来る必要最小限の情報によって、疾患の発症を何処まで予測できるか、何を変えれば、発症しないかという帰納的知見を、日本人集団で構築する必要がある。

本研究では、健康度を幾つかのサブカテゴリーにわけ、サブカテゴリー毎の健康度を、ある一時点の個人の情報から、一定の計算式および、アルゴリズムで数値化し、それらサブカテゴリー毎の数値に基づき、個別の個人用の

order-made preventive strategy を作成し、個人においては、QOL が高く、active で vigorous な長寿命の達成を目指す。

地域や国家のレベルにおいては、新しい予防医学ビジネスや雇用の創出、医療費の削減を目指す。

そのために、(1) Exposure, Surrogate と Outcome の精度を高めること (2) Exposure に遺伝子多型の情報を含めること、(3) Exposure のうち主要なものに、簡便な介入を加え、Surrogate と Outcome が、介入後も同一の数式・アルゴリズムで表現できるか否か検討すること、(4) 非線形モデルを数式・アルゴリズムに活用すべきか否かの検討を行うこと、以上を4本柱として、本研究の主なる成果である、健康度：数式・アルゴリズムは、厚生労働省の指導の元で開発し特許化する。副次的なる成果として、Exposure, Surrogate と Outcome の相互関係や因果関係に関する新しい知見は、国際学術雑誌に投稿し、英文・日本語の書籍として刊行する。

B. 研究方法

discriminant analysis と general linear model の独立要因として採用された項目(尺度等の制約をはずして選択されている)によって、Surrogate として発がんやウイルスへの抵抗力の指標として有用な NK活性の素点と順位数を説明する数式、同じくCD4数を説明する数式、アレルギーの程度をあらわすIgEを説明する数式、老化の指標としても使用できる可能性を持つTh1:Th2サイトカイン比やアポトーシスを説明する数式、Outcome として風邪引き日数、疾病休業日数、高脂血症、他を説明する数式をこれまでに開発している。

検証は以下2つの方法によって行う。

①これまでに数式を開発するためにデータを収集したのと同じ集団の内の他の一部と、そ

れとは地域も職種も異なる別の集団において、cross sectional study を行い、cross validation にて数式の再現性を見る。

②その集団で、介入可能な独立要因に、何らかの介入を行い、2時点目での予測値に変化をおこし、Prospective に数式と実測値との差異を見る。

改良は以下3つの方法により行う。

①discriminant analysis と general linear model に従った線形式に、非線形要素を取り入れるとしていたが、今年度から、MTS法を取り入れることとした。

②項目の文言の変更や未測定概念の追加による独立変数の変更と、従属変数の増加。(肥満、血圧、自殺企図他)

③遺伝子多型の情報を取り入れる。遺伝子多型は、全血からゲノム DNA を抽出し、それをPCR法によって増幅後、制限酵素で切断して、断片長から遺伝子変異を特定したり、PCR産物の直接シーケンス法を用いて調べる。更に、DNA アレイ法によって調べる。

統合は以下のように行う

Outcome としての疾患の発症と Surrogate の免疫系の関係をあらわす数式が、コホートの年数を追うごとに作成される。Exposure → Surrogate → Outcome の流れや Exposure → Outcome の流れなどの幾つかのモデルを比較し、最も個人の値を示す式を作成し、最終的に、それを、高免疫、中免疫、低免疫のように単純化して表現する。

倫理面への配慮

インフォームドコンセントおよび倫理関係書類は、昨年度の報告書参照。

平成13年度に文部科学省・厚生労働省・経済産業省からヒトゲノム・遺伝子研究に関する倫理指針が作成され、それに準拠し、倫理

委員会を経て、十分なインフォームドコンセントのもとに研究を遂行する。

本研究課題に関する倫理委員会書類は、平成13年10月に国立精神・神経センター国府台地区倫理委員会に提出し、平成14年1月に条件付き承認を得て、平成14年4月14日に最終的に、承認を受けた。

その後、被験者からの研究協力の合意を得た。

対象

企業: 国内の製造業、小売業、鉄鋼業に働く
労働者 約4000人
地域: 静岡県、兵庫県の地域住民
約800人

C. 研究成果

統計手法の再検討

今年度は主に数式の根幹をなす統計手法に関して再検討を行った。

「個人の健康度の指標を表現する目的で、Exposure, Surrogate と Outcome の subcategory の関係を繋ぐ数式・アルゴリズムを、discriminant analysis と general linear model を用いて、非公開で作成した。この数式・アルゴリズムが、同一集団および異なる集団にて、Surrogate と Outcome の subcategory の予測に役立つか否かを検証し、適宜、修正を加え、簡略化し、実用に耐えうる、健康度の指標として確立することを目指す。」と概要のところに記載したが、判別分析は全く使用できないことがわかった。

品質管理学会等で活躍する、企業の統計手法に関する専門家等と、何回か話し合いを行った。

品質管理学では、生産された各製品やその部品の特質のうち、不良品や欠陥、事故発生などの、販売後のリスクの推定を行う際に、製

品や部品の特性のうち、Endpointとしての欠陥や事故発生を予測するために最も有用な特性(代用特性)の発見のために、様々な統計手法が用いられている。最近の傾向では、Maharanobis-Taguchi Method を使用する機会が増えたという。

この統計手法に関しては今回の報告書の中で、吉野によって記載されている。医学領域でもすでに、糖尿病の発症予測に関して、この統計手法は使用されている。

実際に、がんとうつ病の発症予測に関して、この統計手法にて、予測を試みた。

うつ病の発症や発症時を目的変数として、生存分析やMTS法(マハラノビス田ロメソッド)を行った。MTSでは、平成12年度にうつ病と診断された5人を比較データとして、

Sheehanの不安尺度、SDS、NEO、JCQ、睡眠時間、食生活、運動、健康診断データからは、中性脂肪、コレステロール、血圧、BMIなどを選んで解析した。MTS法で解析した結果、睡眠時間、神経質傾向、誠実性傾向、伏在型攻撃性などの変数においてSN比が高く、これらの変数で、2001年度以降のうつ病の発症を予測した。距離の閾値を1.5として、うつ病予備軍と、正常者群の2群に分けたとき、うつ病予備軍のグループは、790人中15人がうつ病を発症し、正常群2938人中22人がうつ病を発症した。(オッズ比2.57、95%信頼区間1.33-4.97)

このほか胃癌などの解析も行っているが、特許取得を目指すため秘匿する。

Datasets

今年度は、分担研究者と共に、シンシナティのNIOSH(米国国立職業性安全衛生研究所)に訪れた。

健康度の数式に関する議論を、Nomi Swanson, Steven Saughter 博士らと協議

した。その結果、下記のデータセットにて
そうした数式の検討が可能であることを知
らされた。

1. CDC Behavioral Risk Factor Survey
2. CDC National Center for Health Statistics
3. National Longitudinal Surveys
4. Health and Retirement Study
5. Wisconsin Longitudinal Study
6. Framingham Heart Study

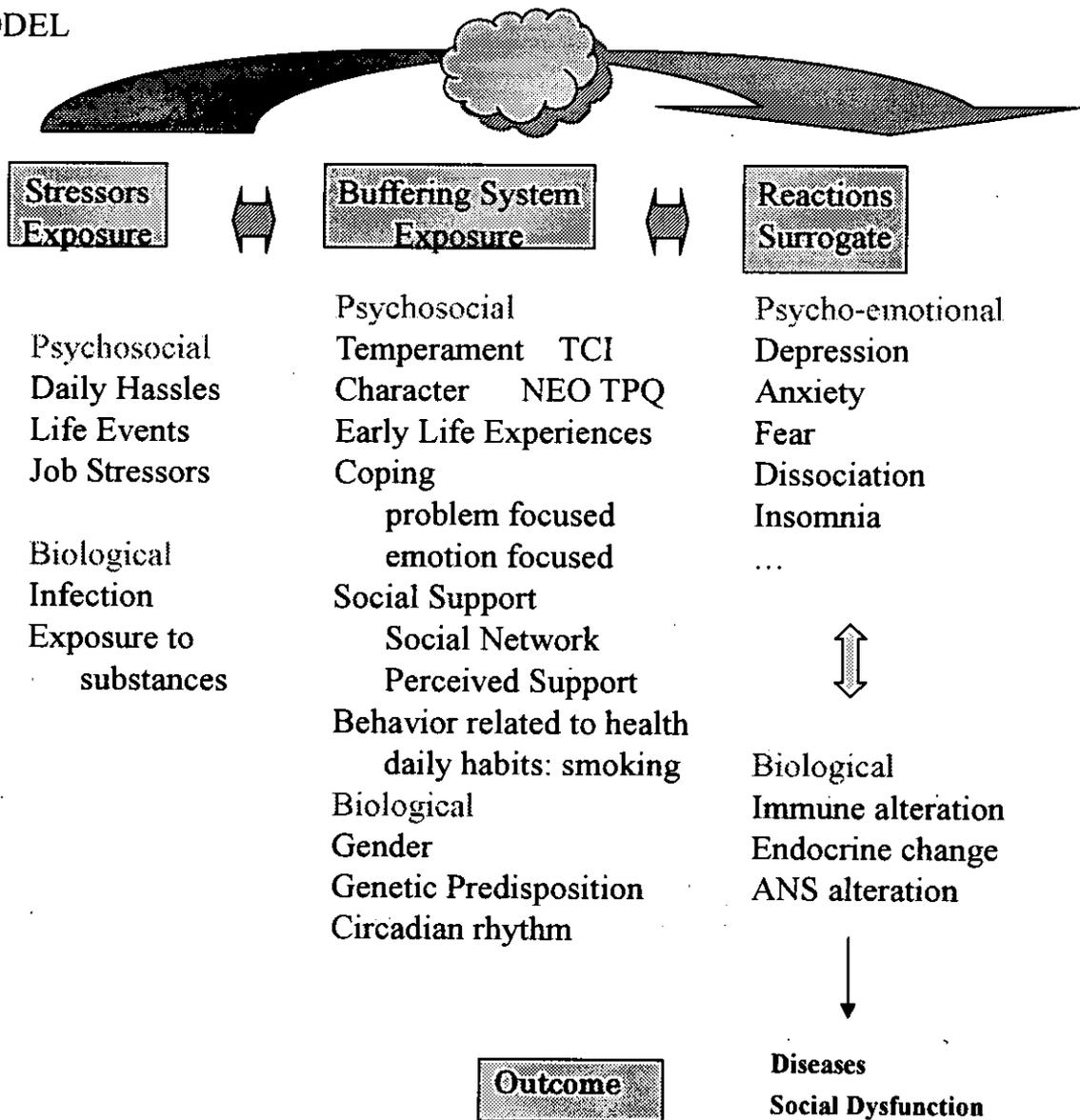
7. National Institute for Mental Health
8. NIH Office of Women's Health Report on
women's health research in NIH

9. National Cancer Institute

10. National Institutes of Health

今後これらのデータから得られる情報との比
較を行う。

MODEL



D.E. 考察と結論

欧米を中心に、過去、疾病罹患と健康関連行動、性格、ストレスなどの健康影響が調べられてきた。acute and chronic stress は60年代から、hostility は70年代から、depression は50年代から、social support は70年代から、socioeconomic status は、50年代から、報告が始まった。これらの成果を日本で検証する仕事も近年盛んであるが、日本人と、欧米人では影響度がことなることが明らかになりつつある。

前記のモデルの中で、緩衝要因(Buffereing System)に該当する項目が、多くの場合、ストレスラーのレギュレーションを変え、健康度数式の分岐点を形成することが、明らかとなっている。今後、この点に注意して、開発を進めたい。

F. 健康危険情報 なし

G. 研究発表

1 論文発表

1 小牧元, 志村翠, 川村則行, 前田基成, 有村達之, 久保千春, 中田光紀, 篠田晴男, 緒方一子 日本語版 The20-item Toronto Alexithymia Scale(TAS-20)の信頼性, 因子的妥当性の検討 心身医学 Vol.43, No.12, Page839-846 2003

2 古賀章子(久留米大学 大学院 比較文化研究科), 前田正治, 進藤啓子, 丸岡隆之, 川村則行 消防業務とトラウマティック・ストレス 福岡市消防隊員に対する疫学調査の結果から 九州神経精神医学(0023-6144)49 巻 1 号 Page44-50 2003.

我々も、欧米の尺度を日本語化したり、日本人向けに多数の尺度を開発して、日本人向けの成果を得ようと努力の途上にある。

常に使用している思考モデルは下記の図の通りである。

遺伝及び環境によって形成される個人のストレスに対する強さが各人によって異なり、諸事象の経験を契機に、メンタルヘルス不全を生じたり、生じなかつたりする場合があるというものであるという、精神科領域に於けるストレス脆弱性モデルと規を一にしている。

3 Bedoui S, Miyake S, Lin Y, Miyamoto K, Oki S, Kawamura N, Beck-Sickinger A, von Hoersten S, Yamamura T: Neuropeptide Y(NPY)suppresses Experimental Autoimmune Encephalomyelitis :NPY1 Receptor-specific induction of autoreactive Th1 response in vivo. J.Immunology 171(7):3451-3458, 2003.10.1

4 Sakami S, Ishikawa T, Kawakami N, Haratani T, Fukui A, Kobayashi F, Fujita O, Araki S, Kawamura N.:Coemergence of Insomnia and a Shift in the Th1/Th2 Balance toward Th2 Dominance. Neuroimmunomodulation 10 : 337-343, 2002-2003.

5 Miyazaki T, Shimizu T, Komaki G, Fujita O, Tsuboi H, Kobayashi F, Kawamura N: Development of Overt-Covert Aggression Inventory. Psychological Report 93: 23-34, 2003.

6 Miyazaki T, Ishikawa T, Iimori H, Miki A, Wenner M, Fukunishi I, Kawamura N :Relationship between perceived social support and immuniefunction. Stress and Health 19:3-7, 2003.

7 Tanigawa T, Iso H, Yamagishi K, Muraki I,

Kawamura N, Nakata A, Sakurai S, Ohira T, Shimamoto T : Association of lymphocyte subpopulations with clustered features of metabolic syndrome in middle-aged Japanese men. *Atherosclerosis in press*

8 Sakami S, Maeda M, Maruoka T, Nakata A, Komaki G, Kawamura N. Positive coping up- and down-regulates in vitro cytokine productions from T cells dependent on stress levels. *Psychotherapy and Psychosomatics*. in press

総説

1 川村則行, 宮崎隆穂, 酒見正太郎, 中田光紀: 特集 脳と免疫 3 精神的ストレスと免疫. *BRAIN MEDICAL* 15(4)12, 2003.

著書

2 川村則行: 心が病気をつくる—自己治癒力の高め方. 代替療法と免疫力・自然治癒力 自然治癒力を高める連載講座 1, ほんの木, 70-85, 東京, 2003.

3 川村則行: 心と体を守る—職業現場における心身のサリュートジェネシス①. 働く人の健康づくり 平成 15 年 4 月号, 2003.4.20.

4 川村則行: 心と体を守る—職業現場における心身のサリュートジェネシス②. 働く人の健康づくり 平成 15 年 6 月号, 2003.6.20.

5 川村則行: 心と体を守る—職業現場における心身のサリュートジェネシス③. 働く人の健康づくり 平成 15 年 7 月号, 2003.7.20.

6 川村則行: 心と体を守る—職業現場における心身のサリュートジェネシス④. 働く人の健康づくり 平成 15 年 8 月号, 2003.8.20.

7 川村則行: 第 44 回日本心身医学会 心身医学におけるサリュートジェネシスの実践を. *Medical Tribune* 36(23): 8, 2003.6.5.

8 川村則行: 神経・内分泌・免疫系の情報ネットワーク. *武田薬報* 433, 22-27, 2003.

9 川村則行: 心と体を守る—職業現場における心身のサリュートジェネシス⑤. 働く人の健康づくり 平成 15 年 9 月号, 2003.9.20.

10 川村則行: 心と体を守る—職業現場における心身のサリュートジェネシス⑥. 働く人の健康づくり 平成 15 年 10 月号, 2003.10.20.

2. 学会発表

一般口演

1: 宮崎隆穂, 川村則行, 小牧元, 石川俊男: 知覚されたソーシャルサポートと楽観的信念が NK 細胞数に与える影響—仮説的因果モデルによる検討—第 44 回日本心身医学会総会ならびに学術講演会シンポジウム VI「ヘルスプロモーションの心身医学」. 那覇市, 2003.5.8-9

2: 川村則行, 宮崎隆穂, 志村翠, 小牧元, 飯森洋史:

高齢者のソーシャルサポート・健康度の精神神経免疫学的研究. 第 44 回日本心身医学会総会ならびに学術講演会シンポジウム VI「ヘルスプロモーションの心身医学」. 那覇市, 2003.5.8-9.

3 安藤哲也, 石川俊男, 成尾哲朗, 岡部憲二郎, 滝井正人, 立川直子, 竹内香織, 増田彰典, 苅部正巳, 山口利昌, 川村則行, 小牧元: 摂食障害でのグレリン遺伝子の多型解析. 第 44 回日本心身医学会総会ならびに学術講演会パネルディスカッション IV「摂食障害の最前

線」, 那覇市, 2003.5.8-9.

4 川合嘉子, 飯森洋史, 野口有紀子, 宮田敬一, 川村則行:

心理的要因や自律神経機能と尿酸値.

第44回日本心身医学会総会ならびに学術講演会, 那覇市, 2003.5.8-9.

5 酒見正太郎, 石川俊男, 川上憲人, 原谷隆史, 福井明, 小林章雄, 藤田定, 荒記俊一, 川村則行: 不眠症と免疫機能の関係. 第62回日本公衆衛生学会総会, 国立京都国際会館, 京都, 2003.10.24.

6 宮崎隆穂, 小牧元, 藤原定, 坪井宏仁, 小林章雄, 清水貴裕, 川村則行:

Overt-Covert Aggression Inventory の信頼性・妥当性の検討.

第19回日本ストレス学会総会, 東京, 2003.11.27-28.

7 酒見正太郎, 石川俊男, 飛鳥井, 望原谷隆史, 小林章雄, 藤田定, 川上憲, 荒記俊一, 福井明, 杉本日出子, 飯森洋史, 川村則行: PTSDの既往のある男性の再適応上の問題. 第8回日本心療内科学会学術総会, 大分, 2004.1.9-10.

8 Miyazaki T, Sakami S, Hasegawa A, Tsuboi H, Oguni I, Hoshi T, Shimura M, Ago Y, Komaki G, Kobayashi F, Kawamura N: The psycho-educational intervention for the promotion of perceived social support. :A randomized control trial for the elderly people at the rural city in Japan. 17th World Congress on Psychosomatic Medicine, Waikoloa, Hawaii, August 25, 2003.

9 Iimori H, Kawai Y, Miyata K, Kawamura N: Relationship between depressive and anxious state, the function of autonomic nervous

system, and the level of uric acid. 17th World Congress on Psychosomatic Medicine, Waikoloa, Hawaii, August 25, 2003.

10 Komaki G, Maeda M, Arimura T, Nakata A, Shinoda H, Ogata I, Shimura M, Kawamura N, Kubo C: The reliability and factorial validity of the Japanese version of the 20-Item Toronto Alexithymia Scale (TAS-20). 17th World Congress on Psychosomatic Medicine, Waikoloa, Hawaii, August 25, 2003.

3 委員等

1 心療内科学会 編集委員
心身医学会 評議員

2 武蔵野女子大学 人間関係学科
非常勤講師

3 大阪大学大学院衛生学
非常勤講師

H. 知的所有権の出願・登録状況

1 特許取得
なし

2 実用新案登録
なし

3 その他
なし

本報告書において用いられた統計手法の解説

宮崎隆徳^{1,2)}

¹⁾学術振興会 特別研究員

²⁾精神・神経センター 精神保健研究所

概要

本章では、当報告書において用いられた統計手法についてその背景と思想、さらには特徴や用いる際の注意点について簡単に概説する。具体的には大きく分けると一般化線形モデルとデータマイニングの手法について解説している。一般化線形モデルの項では、t 検定・分散分析から始まりロジスティック回帰モデル、生存時間解析まで含み、それら分析の背景と特徴を記述した。データマイニングの項ではオペレーションズリサーチの分野から CHAID、AHP の2つの手法を取り上げそれぞれ解説を試みた。一般化線形モデルにおけるいわゆるけちの (parsimony) 論理とデータマイニングの手法の思想上の違いをあきらかにし、それぞれの目的について理解を深めることを目的とした。

一般化線形モデルについて

統計手法を用いた研究とは、単純化すれば入力と出力との関係の研究であるといえる。入力である説明変数 (独立変数・independent variable とともに記述される) は出力である被説明変数 (従属変数・dependent variable とともに記述される) となんらかの関係 (効果・effect) があると研究初期に想定されている。また、研究では直接関心がないが入力と出力双方に関連があると思われる別の因子、つまり交絡因子 (confounding factor) があるとその関係は複雑なものになる。いかに交絡因子の影響を統制しながら目的とする入力と出力との関係を記述するのかが、実質科学的研究ではその根幹をなすほど重要であるといえよう。

さて、あるサンプル (データ) をとってくれば、こうした入出力の変数間の関係をモデル化することができ、それぞれのパラメータを推定し、そのデータの当てはめ式が得られる。このデータとモデルとの関係を最も簡単に表すと⁽¹⁾、

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

になる。ここで、FIT はそのモデルからの予測値となり、RESIDUAL は DATA と FIT との差である。従来これらの線形重回帰モデルは t 検定や分散分析といった従属変数

が連続変数でかつ正規分布を仮定した検定法で用いられることが多かった。分散分析に関しては従属変数の正規性が保証されない場合でも、比較的結果が頑健であることが知られていたが黙認される程度にとどまっていた。もともとこうした変数に関する制約条件は分散分析や t 検定などの古典的統計手法が、Fisher に始まる農産物の品種改良などの従属変数が正規分布にしたがうような実験的手法から一般的な知見を導き出す手法として発達してきたことに由来するものと思われる。しかし、統計的手法を用いた研究が医学、経済学、心理学など様々な分野に応用されるに至って必ずしも従属変数が正規分布に従わなかったり、連続変数ではない場合が多くなり理論的な拡張が要請された。その結果提案されたのが一般化線形モデル (Generalized Linear Model)⁽²⁾であり、今日の医学分野における統計学的推測の根幹をなす基本モデルである⁽³⁾とされている。この手法は、従来の正規分布の制約をはずし、正規分布になじまない確率変数に対しても統一的な線形推測が可能となるようにしたものである。扱われる線形モデルとしては従来の線形重回帰モデルに加えて以下のような分析が内包さ

(2) Nelder, J.A. and Wedderburn R.W.M. Generalized Linear Models. J.R. Statist. Soc., A. 135, 370-384(1972)

(3) McCullagh, P. and J.A. Nelder. Generalized Linear Models, Second edition, London: Chapman and Hall, 1989

(1) Chatfield C. Problem Solving. A statistician's guide. London: Chapman and Hall, 1995

れる。

- * 対数線形モデル
- * ロジスティック回帰モデル
- * 生存時間解析
- * ポワソン回帰モデル

一般化線形モデルでは、従属変数の正規性の制約が外れているが、制約が全くないわけではなく正規分布、二項分布、ポアソン分布など指数型分布族である事が条件になるが、この扱える分布の拡張により扱える問題の範囲が飛躍的に広がっている。端的にまとめると、従属変数が連続変量の時には重回帰モデルをつかい、2 値データの時にはロジスティック回帰モデルや生存時間解析、頻度データであればポワソン回帰モデルを用いるということになる。

一般化線形モデルにおいては、①独立変数の効果の分散を誤差分散との対比によって明らかにする ②線形結合モデル全体から導かれた予測値とデータとの一致率 を有効なモデルが構築されたかどうかの判断基準とすることでは共通している。本来であれば想定されるサンプルの全数調査を行えばこれらの統計手法は必要ではない。しかし、現実的にも理論的にもある問題に対する全数調査を行うことは非常に困難であることが多く、より少ない標本集団から効率的に母集団を推測することを目的としているのである。しかし注意しなければならないのは「より少なく(効率的な)、より正確な(誤差の少ない)」モデルは強力な情報を所有するが、前提として入力と出力の変数の設定(因果関係の設定)は統計手法の選択ではなく、いくつかの条件を満たした実験計画(研究計画)によらなければ誤った結論を支持する可能性があるということである。

データマイニングについて (CHAID)

ボルヘスは「バベルの図書館」という短編の中で無限の図書館について記述している。本棚のある部屋が終わりなくつながっている構造物がそれこそ無尽蔵の本を収納している。しかしながら大部分の本は無意味なタイトルがついていたり、内容のないものである。すなわちその図書館には無

限の量のデータが存在するが、情報は存在しない。こうした皮肉な風刺が現代人が置かれている状況と重なるのは明白であろう。データマイニングはこうした大量のデータの中から有意義な情報を取り出すために考案された一連の方法である。つまり扱われる変数の数が多く、サンプル数も数千数万単位のデータに適用されることが多い。しかも、分析にかかるコストが少なく(計算量が少ない)、結果が直感的にすぐ理解でき、効率的に情報を発掘するという思想の下にデータマイニングは行われる。前述した推測統計学の範疇である一般化線形モデルとは、推測の正確さにどれくらい力点を置くかが決定的に異なる。データマイニングでは扱えるデータ量が膨大なため、データ内での交差妥当化やさらには再分析すら可能なのである。

今回はこのデータマイニングの手法から決定木による分析について解説する。そもそも決定木による分析は 1960 年代にさかのぼる人工知能研究の副産物として生まれた。パーセプトロンなどの単純な学習モデル機械が、既存のデータから決定規則を見つけ分類を行う際のその分類の容態を示すのが決定木分析である。扱う変数の内容と用途に応じて以下のようないくつかのアルゴリズムが提案されている。

今回は主に CHAID について解説する。CHAID は Chi-squared Automatic Interaction Detector であり χ^2 乗統計量による相互作用の自動検出を行う分析手法である。従属変数が量的な物の場合にはいくつかの区間に分割しカテゴリー変数化を行う。まず有意確率水準を利用して、独立変数を評価を行う。評価の基準は従属変数に対して等質であるか異質であるかということであり、従属変数に対して独立変数が等質ならば値を結合し、異質ならば結合しないという操作を行う。結果として最良の独立変数群を選択して、等質なグループになるように分岐を構成していく。検定の方法は従属変数が量的変数の時には F 検定を行い、名義尺度・順序尺度の時には、2 方向クロス表、Pearson の χ^2 乗検定、尤度

比検定、などを行う。データが十分にある場合には、決定木が算出された後に、テストサンプル法などで交差妥当化の検証を行いある程度算出された決定木モデルが一般化可能かどうか検証を行う。

にあたることができる。

- ② 比率をペア比較で答える際、同じくらい、やや、かなり、非常に、きわめて、といったファジィな表現を用いることによって意思決定者の負担を軽くしている。

アルゴリズム	従属変数のモード	分岐
CHAID	名義、順序、量的	2進木以上
Exhaustive CHAID	名義、順序、量的	2進木以上
C&RT(CART)	名義、順序、量的	2進木
QUEST	名義のみ	2進木

AHP について

Analytic Hierarchy Process(AHP:階層分析法)は主に、オペレーションズリサーチ(OR)の領域で、人間なら誰もが持っている経験や勘という感覚的な情報を意思決定のプロセスに取り入れた問題解決型意思決定手法として T.L.Saaty が提唱した方法である。AHP 手法は次の3段階から成り立つ。

- ③ 首尾一貫性のないデータを扱え、しかも首尾一貫性の度合いが同時にわかるので修正が用意である。
- ④ データがない、または取りにくい環境下で意思決定しなければならないような問題の解決にあたる事が出来る。
- ⑤ 決定に先立って、様々な場合を想定して意思決定の影響を予測したいような問題の解決にあたる事が出来る。

(1) 第1段階

複雑な状況下にある問題を階層構造に分解する。ただし、階層の最上層は1つの総合目的であり、下位目的が意思決定者によりいくつか決定される。

(2) 第2段階

各レベルの要素間の重み付けを行う。それぞれのレベルに存在する要素を一对比較法により全ての順列組み合わせで行う。この結果から線形代数の固有値の考え方をを用いてそれぞれの重みを計算する。

(3) 第3段階

各レベルの要素間の重み付けが決定されると、階層全体の重み付けを行い、総合目的に対する各代替案の優先順位が決定する。

AHP の特徴としては以下に整理される⁽⁴⁾。

- ① 評価基準がたくさんあり、しかもお互いに共通の尺度がないような問題の解決

(4) 木下栄蔵 マネジメントサイエンス入門 経営・政策科学の戦略モデル 啓学出版株式会社 東京 1993

健康診断のような多変量データに適用する統計手法について品質管理の立場から留意点をまとめた。

医療分野では、観測データが極めて小数でかつクラス分けされた副作用の大きさなどが用いられるため、用いられる統計手法はノンパラメトリック検定法が中心となっている。従って、多くの観測変数を処理する多変量解析法が適用されることは比較的少ない。

しかし、健康診断データでは、観測変数が多いうえに、それに属する標本の異常状態を指標的に表して正常群と区別する必要があるため、記述的な多変量解析法が有効である。

そこで、本稿では管理や検査で用いる多変量解析法について、それらを推奨する理由と適用する上での留意点について論ずる。特に、データのスペクトル分解に相当し、異常状態の大きさを視覚的に捉える主成分分析法とその延長上にあるMTS法、および異常状態の大きさではなくパターンを検出する対応分析について実例を用いて比較する。

1. はじめに

品質管理という学問領域がある。管理とは、良い状態を保つように処置することを意味する。健康管理も広義の品質管理である。品質管理においては、管理や検査という目的を効率的に行なうために統計的手法が活用されている。本稿では、管理や検査で用いる多変量解析法について論じる。そして、その統計的手法を推奨する理由や、研究デザインに対する留意点について述べたい。

さて、品質管理の分野では、工程より抜取ったサンプルの「できばえ」から工程の安定度を判定するというように現場的側面が強い。そこで直感的なグラフなど記述的方法が活用される。それに対して医療分野では管理よりもむしろ薬効のような介入

効果を測定することが主眼となる。そのため用いられる統計的手法は検定が主流であるといっても過言ではない⁽¹⁾。

また、医療・薬事の分野における観測データはプラセボのような対照群(統制群)との比較データや、薬効の程度のような分割表データであり、多くがノンパラメトリックである。工学分野の観測データのようにきれいな正規分布が得られることは少なく、検定手法が煩雑になる。このことが医療・薬事の分野におけるパラメトリックな多変量解析手法の適用を困難にしている。

以上のような背景から、医療・薬事の分野において記述的な多変量解析法が活用されることは比較的少ない。

そもそも、統計学は社会統計学から発生し計量経済学に適用され数量化理論へと進歩したものと、遺伝学から発生し確率

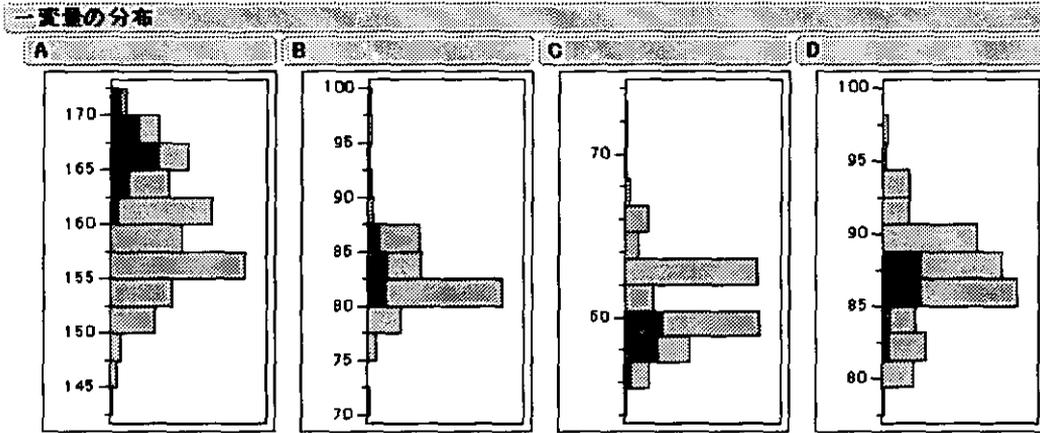


図-1. 各変数のヒストグラム (濃い色がⅡ群)

変数名	A	B	C	D
A	1	0.189	-0.095	0.173
B	0.189	1	0.338	0.428
C	-0.095	0.338	1	0.534
D	0.173	0.428	0.534	1

表-1. 各変数間の相関係数

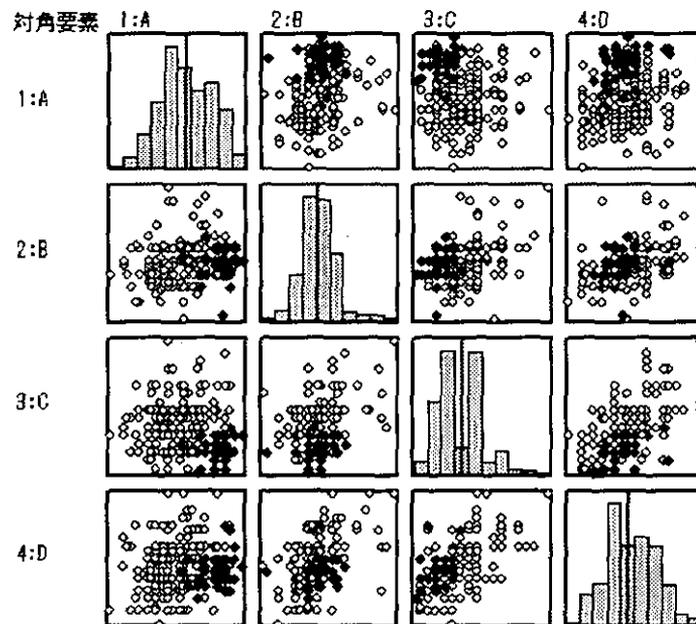


図-2. 各変数間の散布図 (濃い色がⅡ群)

論を持ち込み検定・推定のような帰納的推論へ発展した数理統計学とがある⁽²⁾。医療・薬事に適用される統計学は後者を基盤としており、その中で自己完結しており、記述的方法を用いなくても不自由は無い。

しかし、最近では両者が融合し探索的データ解析へと進歩し多様化している。いわゆる多変量解析法の発展形であるが、今後、医療・薬事分野のデータにもそれら新しい統計的手法が広く適用されていくのではないかと思われる。

そこで本稿では、健診データを取り上げ、品質管理に用いられる記述的な多変量解析手法を適用したらどうなるかということ和管理・検査の立場から検討する。

2. 推測統計と記述統計

多くの教科書では、母集団と標本の説明から始まって、正規分布に代表される標本分布や検定・推定のような統計的推測手法の解説へと進む。これは数理統計学を基盤にした「推測統計」であるが、それと対極をなす統計学がある。それが統計的記述を目的とした「記述統計」である。記述統計は確率的分布を先験的に仮定しない点で推測統計と異なっている。記述統計ではむしろデータの中から規則性を発見し、それを管理や予測に使用しようとする。測定対象が全数調査ということもあり得るほどデータ主体であり、国勢調査データなどが対象になる。記述統計の究極の姿がデータマイニング手法といえる。

以上のように、記述統計は目の前にある現象を見やすく整理要約して数量的秩序

を発見しようという方法である。この方法は、統計データの大型化とコンピュータの利用によって大きな進歩を遂げてきた。現在、主な解析手法として、分類法として因子分析、クラスター分析、縮約方法として主成分分析や多次元尺度構成法、因果関係の解明手段として共分散構造分析などが一般的に用いられている。

ここでは縮約方法に的を絞って、その手法の概要を述べる。

観測者がまずすべきことは、データに潜む情報を顕在化させることである。そこでヒストグラムや散布図を描き、それから平均値、中央値、最頻値を求めることは記述統計の第一歩である。しかし、多変量のデータでは、変数間に相関があったりして、それだけで全体のデータ構造を把握することは困難である。相関係数を見ればよいわけだが、無味乾燥な数値列を眺めていてもなかなか情報は顕在化してこない。そこで、この相関係数行列をスペクトル分解して眺めてみようというのが縮約である。たとえば、音の波だけを眺めていてもどんな高さの音が出ているのか全く分からないが、スペクトル分析して周波数成分に分けて見ればド・ミ・ソの音が大きいことが分かるというようなものである。ここでは、元のデータに対して確率的分布のような仮定は一切持ち込まない。縮約されたデータは、元のデータに対して完全に対応している。

利点は次に述べるとおりである。多変量のデータを少ない座標軸上にプロットしても元の姿を残しているということであるから、大方の特徴がその座標軸に要約されているということである。そこで特徴のある座標軸が元のデータのどんな変数から成り立

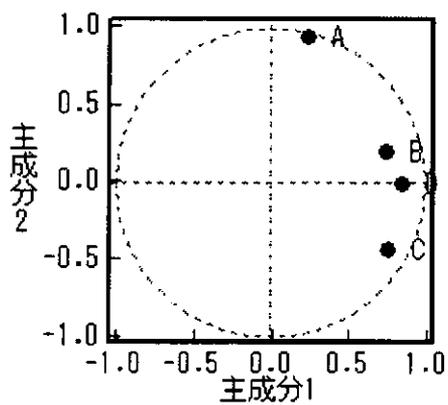


図-3.主成分分析結果—因子負荷量

分散 = 1.0 表示物*数: 250/250 出力基準値: 0.00

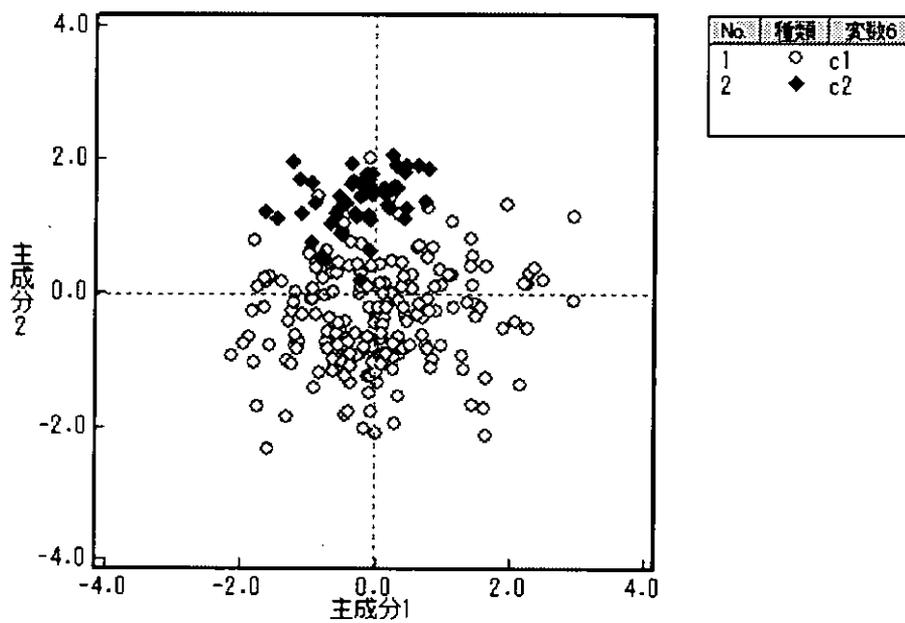


図-4.主成分分析結果—主成分得点 (濃い色がⅡ群)

っているかを調べることにより、その特徴を生んでいる要因が把握できる。すなわち管理すべき要因が判明するのである。さらに各々のサンプルが要約された座標軸上などの部位に来るかにより、そのサンプルの分類も可能となる。すなわち検査への適用が可能になるのである。次節では具体的事例を用いてその利点を述べる。

3. 記述統計の適用例

①主成分分析

健診データに記述統計を適用してみた。このデータは4変数で2群に分かれており、I群サンプル数200、II群サンプル数50である。II群は異常群と考える。詳細データは付録に載せる。

まず、ヒストグラムを図1に示す。ヒストグラムから各変数の分布は分かるが、I群とII群は重なり合い、分離は鮮明ではない。そこでさらに相関構造を検討することとする。相関係数行列を表1に、多変量散布図を図2に示す。これより、変数間に相関があることが分かる。多変量散布図を眺めると何となくII群は分離できるようである。

データの構造が不鮮明なのは、4変数全てを同時に眺めているからである。変数間に相関があるのなら、その構造を反映させて眺めてみたい。そこでI群・II群あわせて主成分分析を行なってみる。その結果を図3,4に示す。この健診データのように変数間に相関があるサンプルを空間上にプロットしてみると図5のように分布していると考えられる。主成分分析では図5に示す平面のように、この分布のばらつきが最大になるように第1座標軸 p_1 を決め、さ

らにその座標軸に直交するばらつきが最大になるように第2座標軸 p_2 を決める。以下同様にして座標軸を決めていく。

図3は因子負荷量と呼ばれる散布図である。解析に用いた4つの変数が、縮約された2本の座標軸にどのように影響しているかを表している。第1座標軸に相当する成分を第1主成分というが、これには変数A,B,C,Dの全てが正の効果を与えている。そこで、このような主成分を各サンプルの大小を表す成分と考え「サイズファクタ」と呼ぶことがある。それに対して第2主成分は、A,Bが正の効果、Cが負の効果を与えている。そこで、このように正負の効果が入り混じっている主成分を各変数の凹凸度合いを表す成分と考え「シェイプファクタ」と呼ぶことがある。

図4は主成分得点と呼ばれる散布図である。解析に使用した各々のサンプルが、縮約された2本の座標軸上などの位置にプロットされるかを示している。特定のサンプルが全体の中でサイズが大きいのかどうか、凹凸が大きいのかどうか分かる。さらに、I群とII群の違いについて第2主成分軸(シェイプファクタ)の方向で分離可能であることが分かる。

このように、健診データは4変数であったが、主成分分析により2軸への縮約が可能となった。また、2つの群が分類できることが明らかとなった。群の差を生んでいるのは変数A,BとCの凹凸であるということも判明した。しかし、検査によってII群を分類するために、わざわざ座標軸上に布置しなければならないというのでは非効率である。そこで、特定の軸上でII群が分類できることに着目して、さらに効率的な検

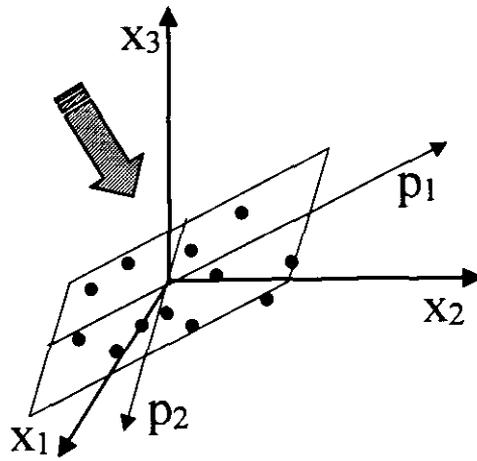


図-5.主成分分析の概念

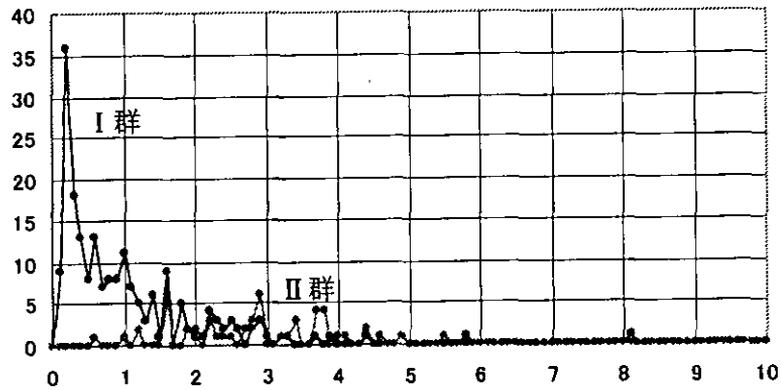


図-6.マハラノビス・タグチの距離の分布

総合順位	出身国	選手名	J1	J2	J3	J4	J5	J6	J7	J8	J9
			SUI	USA	GBR	JPN	GDR	FRG	URS	CAN	TCH
			スイス	米	英	日	東独	西独	ソ連	カナダ	チェコ
P1	1 東独	カタリナ・ビット	58	56	56	57	58	57	57	57	58
P2	2 加	エリザベス・マンリー	59	57	58	59	59	59	57	58	59
P3	3 米	デブラ・トーマス	57	55	55	57	57	58	56	56	56
P4	4 米	ジル・トレナー	56	55	54	57	57	56	56	56	57
P5	5 日	伊藤みどり	59	58	59	59	59	59	58	59	59
P6	6 西独	クラウディア・ライストナー	56	56	54	54	56	57	54	54	56
P7	7 ソ	キラ・イワノワ	52	45	47	54	54	52	54	50	51
P8	8 ソ	アンナ・コンドラシヨワ	52	52	48	55	55	56	55	51	50
P9	9 東独	シモーネ・コッホ	50	51	49	53	53	52	53	53	51
P10	10 西独	マリナ・キエルマン	51	50	53	53	53	53	52	52	53
P11	11 伊	ベアトリス・ジェルミニ	51	50	50	46	51	47	51	51	49

表-2.カルガリ五輪の女子フィギアスケート技術点の採点

出方法を検討する。具体的にはマハラノビス・タグチの方法(以下 MTS 法)を使用する。

②MTS 法

MTS 法とは、多変量正規分布データで表現される集団の裾にある異常群を検出する検査手法で、田口が提案した。多変数の観測データに対して適用し、変数間の相関関係を考慮して、個々のサンプルの重心からのマハラノビスの汎距離を算出し、重心から遠いサンプルを異常値として検出する。主成分分析が縮約後も全ての変数を使用しているのに対して、MTS 法では変数を適切に選択して分離を明瞭化する。また、各サンプルの性質はたった一つのスカラー量に縮約される。

本データの主成分分析の結果では、I 群が原点まわりに布置されるのに対して、II 群は原点から離れた位置に布置されている。この構造を利用するのである。

MTS 法は主成分分析と関連が強い。主成分分析において、第 k 主成分の固有値 λ_k で規準化した i 番目のサンプルの主成分得点を、

$$f_{ik}^*$$

とすると、 i 番目サンプルの(マハラノビスの距離)² は、規準化した主成分得点を 2 乗して加えることにより求められる⁽³⁾。すなわち、

$$D^2 = f_{i1}^{*2} + f_{i2}^{*2} + f_{i3}^{*2} + f_{i4}^{*2} + f_{i5}^{*2} + \dots$$

よって、主成分分析のプロットはマハラノビスの汎距離を反映していることが分かる。このことから MTS 法は主成分分析の延長上に捉えることができる。

MTS 法を適用して検出を行なった結果、選択された変数は、A、C となった。距離を計

算した結果を図 6 に示す。図 6 は横軸がマハラノビス・タグチの距離、縦軸が頻度である。この図から、閾値を 2 に設ければ、ほぼ I 群と II 群を分類できると思われる。分類した結果、散布図上ではどうなっているかを図 7 に示す。これより、過剰判定と検査漏れが生じていることが分かるが、検査の性格上この程度は止むを得ないであろう。

③対応分析

対応分析はコレスポンディング分析ともいう。主成分分析と同じくスペクトル分解であるが、距離の捉え方が異なる。

ここでは、カルガリ五輪におけるフィギアスケートの採点を取り上げ、主成分分析との違いを明らかにしたい。データは表 2 に示すとおり、11 人の選手に対して 9 人の審判が採点を行なっている。なお、表 2 は実際の得点に対し 10 倍して整数化してある。採点には技術点と芸術点があるが、ここでは技術点について詳細に検討し、後ほど芸術点についても検討してみる。

まず、このデータに対し代表的な記述統計手法である主成分分析を行なってみる。選手をサンプルとしたものと、表 2 を転置し審判をサンプルとしたもの 2 つのケースを実施してみる。予想される結果として、選手をサンプルにすれば、サイズファクタ軸には選手の実力が現れ、審判をサンプルにすれば、サイズファクタ軸には採点の甘辛が現れることは容易に類推できる。それらに直交して現れる第 2 主成分は、選手がサンプルのときは選手の実力以外の変動分が現れる。いったいそれは何なのだろうか。

図 8 に選手をサンプルとした主成分分析の主成分得点を示す。図 9 に審判をサン