

厚生省「がんの疫学」班分担研究報告

がんの疫学統計の方法に関する研究

分担研究者 山 本 俊 一

(東京大学医学部)

研究協力者 稲 葉 裕

柳 井 晴 夫

高 木 廣 文

1980年 3 月

はじめに

がんの疫学の分野における研究は、最近ますます盛んになってきて、毎年発表される情報量は急速に蓄積しつつあり、その進歩には目をみはるものがある。しかしながら、一方ではそのような努力に対して高い評価を払いつつも、疫学が生産する大量のアウトプットデータに対しては必ずしも全面的な信頼がよせられていないのが、残念ながら事実である。

このように、がんに限らず一般に非感染症の疫学的アプローチに対する批判については、疫学者の側においても反省しなければならない面があるように思われる。もちろん、かりに疫学に対して不信感があるとすれば、それは単に疫学者だけに責任があるのではなく、疫学者に基礎データを提供する側にもやはり問題があることは言うまでもない。しかしながら、疫学者に対して良質のデータが提供されない風潮が存在するとするならば、それはあるいは疫学に対する不信感に由来しているのかも知れない。もしそうであるならば、疫学が振わないということは、そこに疫学者とその周囲にある専門家との間に相互不信の悪循環があるからだと言えよう。

デカルトが全てを懐疑してみて、最終的にどうしても懐疑できないものとして、「われ思う。故にわれあり。」という命題に到達し、これを出発点としてその哲学大系を樹立したのは有名である。

そこでわれわれは、これにならって、がんの疫学の原点を求めてみたいと思う。すなわち、たとえ疫学が提示したデータが全く疑わしいものであったとして、その中にどうしても捨てきれずに残るものがあるだろうか。もしありとすれば、それは何であろうかを求めることである。

先ず、がんの各種の統計の中でも比較的信頼性のうすいとみられている全国死亡統計における部位別がん死亡の年次的変動を取り上げてみよう。全死亡、中高年令層における年齢別死亡、胃がん、乳がん、子宮がんによる死亡などにおいては実数、比率とも、少なくとも最近10～20年間はそれぞれ一定値を維持し、その変動幅の小さいことが認められている。これに対して発生頻度の低い部位別がん死亡あるいは、若年層における年齢別死亡にはかなり年次的な変動がみられる。このことについては、われわれは当然のこととして受けとめていて別に不審をいだくことはない。それは統計学の法則により、大数の場合には変動は少なく、小数の場合には変動が大きいものだからである。しかし、われわれは、「当り前のことである。」として何物かを受け入れているとき、その中に含まれている大きな真理を見逃し勝ちであることは、ニュートンの林檎の逸話がよくそれを物語っている。そして、ここで見落されている大きな真理とは、「人間集団の中のがんの出現には、それぞれ与えられた前提に応じた安定した事前確率がある。」ということではないかと思うのである。

死因別死亡統計に示されている数値がそのまま正確に国民における死亡パターンを示しているとは誰も思っていないが、しかし、正確ではないにしてもそれが近似的には正しいはずだという一歩も二歩も譲歩した主張に対しても異議を唱える人が決して少なくない。衛生統計値は、い

目 次

はじめに	1
1. 資料に関する検討	4
1.1 「Cancer Mortality in 24 countries」について	
1.2 「Cancer Incidence in Five Continents」について	
1.3 国際比較の環境要因の資料について	
1.4 「悪性新生物死亡統計」について	
1.5 日本国内の県別環境要因について	
1.6 山梨県内市町村別比較について	
1.7 Case-Control Study 資料について	
2. 標準化の方法	6
2.1 直接法による年齢訂正死亡率	
2.2 間接法による年齢訂正死亡率	
2.3 40才以上特殊死亡率など	
2.4 その他	
3. 相関係数について	10
3.1 相関係数の各種の指標	
3.2 相関係数についての性質	
4. 多変量解析について	17
4.1 多変量解析法の分類	
4.2 因子分析によるがんの部位別・地域別死亡率の分析	
4.3 重回帰分析の疫学データへの適用	
4.4 数量化二類を用いた胃がんの risk factor に関する解析	
5. 地域相関と個人相関	28
5.1 データの質について	
5.2 統計方法の適用について	
5.3 地域単位と個人単位で得られた結果が矛盾する場合の考察	
5.4 地域相関と個人相関の数学的関連	
6. 因果関係の推論をめぐって	31
6.1 偏相関と因果関係	
6.2 パス解析	
6.3 相関関係と因果関係	
6.4 リスクファクターの推定をめぐって	
おわりに	37

わば実数の影をみているのであって、それは実数の2倍であるかもしれないし、あるいはその半分であるかもしれない。しかも、それを知る人はいないのである。しかし、それにもかかわらず、われわれは特定の人口集団の中で特定のがんで死亡する確率を定常化させようとする何らかの潜在力が働いていることはどうしても否定できない。確かに何年にもわたって近似的に同数あるいは同率が続くのは、同じような方法で同じような水準の調査を続けているからであって、もし、調査精度を上げるならば、異なった数値の得られることは明らかであろう。しかし、だからといって、それによって本質が変わるのではなく、その実態は相変わらず定常性を維持し続けているのが真実であろう。

われわれは、がん死亡に関するこの安定化された事前確率をがんの疫学の原点とするよう提唱したいのである。

この原点に立ってわが国におけるがん死亡のすう勢をもう一度眺めてみると、統計学的変動を示す発生頻度の低いがんを除いて、全般的に安定した永年変動をみるのであるが、しかし、その一部は明らかに増加ないし減少傾向を示している。このような例外のあることは、しかしながら、決して前記の安定化作用を否定するものではなく、むしろ、この作用に拮抗してもう一つこれを変化させようとする別の力の働いていることの証拠であろう。そして、そのような第二の力は、時間とともに変化していくもので疫学者が環境要因と名付けているもの、あるいは少なくとも、それを主力とするものであらうと考えてよいであろう。

するとここで問題となるのは、この二つの力、特により基本的と考えられるこのような安定化作用の本態は何かということである。

さて、ここで生物学的な立場に立ってこのような疫学の原点のさらにその基礎を探るならば、われわれは地球上に存在する人類のそれぞれの個人が所有する複雑な遺伝子機構を全人口にわたって総合した巨大な集合体、あるいはもっと具体的に表現すれば、全人類にまたがる遺伝子の集合と配分が、がんばかりでなく、すべての人間の生物学的現象を規定する基盤として存在していると思うのである。しかし、この本態は決して固いわくにはめられた決定論的な静止体ではなく、極めて活発に変動する確率論的な存在なのである。何故ならば、先ず生物としての人間集団が新しい個体の子孫として生み出していくとき、そこには先ず交配という極めて偶発的な、そして人間の場合には社会的な動機が必要であり、これに引き続いて受精の段階では、生物学の法則に従った、精子と卵子との間の高度に確率的な結合の実現がその本態となっているからである。どの個人もその実現の確率がほとんど無に等しい中から生まれてくるのである。

このようにして出生した新しい個体には、その将来の発育から老化への生物学的経過の中のあらゆる事象の発現に関して、genotypeとしての確率分布が与えられてあるはずであるが、これに対して多くの後天的要因が作用してそれを大きく変動させ、その結果としてphenotypeとしての生物学的事象が発現してくると思うのであるが、その間の分布の修飾については、上に想定したgenotypeが不可知的な想定である以上、その変動を推定することはできないのである。

がんについても、受精の瞬間において、その新しい生命の将来の発がんの確率分布が定まっ

てしまうという決定論があり得ると思うのであるが、そのような思弁的な論議はともかくとして、われわれには知ることのできない確率法則に従って、現実の中で、ある発育時間のある個体の、ある組織の中に、誰にも気付かれないある種のがん化の最初の段階が実現するのである。

この一つの小さな変化が、がん組織へと自己増殖する pathogenesis についてもわれわれの知識は極めて乏しいが、しかし、個体側とがん組織側とのからみ合いとして発展していくその過程は、われわれにとっては近寄り難い複雑な確率過程をふんでいるはずである。

それは最終的にはがん発病までの潜伏期間の長さの分布としてわれわれには理念的に理解できるが、しかし個体内でのがん細胞の誕生時点が不明である以上、それも観測することのできないものであって、われわれの知り得るものは、偶発的な「早期発見」によって観察された極度の偏りをもつがんの発育段階の分布でしかない。

次に、そこで治療を受けたとき、その予後を決めるのも極めて偶発的なものばかりである。われわれの医師との遭遇、その医師が決定した方針、外科手術における切除の実態などが大きな決定因子となるばかりでなく、がん組織の悪性度、転移の規模と分布などすべてが確率事象である。これらの各因子の影響の強さに関しては最終的には治療を受けてからがんが死亡するまでの余命を指標として観測し評価することしかできない。

しかし、ここでも、がんで死ぬ前に他の病気で死ぬという確率が待ちうけているのである。

このように考えてくると、がん死亡はどうしようもなく複雑にからみ合った確率事象の終点をみているのであって、これについて疫学的・統計学的な解析を加えようという意欲は、考えれば考えるほど消えていくように思える。

しかしながら、それはどんなに困難であろうと、われわれががんの疫学者は、ないものを追っているのではないという確信はゆるがない。しかも、その内容はどのように複雑であろうと、現象としては定常性という取り扱いを容易にする条件をがんの側が提示してくれていて、あたかもわれわれの解明を待っているかのようである。

このような立場に立ってわれわれは、無理に先きに進もうとするのではなく、むしろ独断と偏見による、誤認と誤解を極力避けながら、以下にがんの疫学について統計学的な観点をふまえた内省的な考察を展開した積りである。しかし、われわれの能力の不足から、そこには多くの誤りや欠陥があるものと思われ、これに対する御助言や御批判をいただけるよう、ここにお願いする次第である。

1. 資料に関する検討

最近の疫学的データ解析の1つの特徴として、既存の資料を用いた統計的解析がある。いかなる統計的手法を選択するについても、もとなる数値がどこから得られたものかを十分に吟味する必要があることはいうまでもない。

今回の研究に我々が用いた資料を表1-1に示す。「がんの疫学」によく利用される資料は、既存資料としては限られており、その内容については周知の研究者が多いと思われるが、確認の意味も含めて、以下に簡単な説明を試みる。

表1-1 がんの疫学の統計的方法に関する研究に使用した資料

I 国際比較

死亡率：「Cancer Mortality in 24 Countries」(Segi et al., 1969)

罹患率：「Cancer Incidence in Five Continents」(UICC, 1976)

環境要因：理科年表(東京天文台編, 1975)

FAO Production Year-book (FAO, 1971)

Monthly Bulletin of Agricultural Economics and Statistics

(FAO 1976, 1977)

II 日本国内の県別比較

死亡率：「悪性新生物死亡統計」(厚生省, 1973)

環境要因：理科年表(東京天文台編, 1975)

III 山梨県内市町村別比較

死亡率：県庁資料等より作成(稲葉ら, 1977)

環境要因：山梨県総合地域指標(1975)

IV 山梨県の肝がん・肝硬変 case control study (山本ら, 1980)

1.1 「Cancer Mortality in 24 countries」について

これは全6冊よりなり、1950年～1967年の18年間、男13, 女14の部位のがんについて、信頼できる資料が存在する24ヶ国の死亡数, 死亡率, 年令訂正死亡率(世界人口を基準とする直接法: 後述)が掲載されている。そのもとなる数値は、各国からWHO(世界保健機構)に届けられたもので、いわゆる死亡診断書に基いたものと考えられる。国により死亡診断書の信頼度も多少は異なるが、いわゆる先進国中心のこの24ヶ国の資料は、国際比較をする上ではしばしば用いられるもので、有用な資料といえよう。ただし欧米が中心となるため、アジアでは日本, 中東でイスラエル, 南米でチリという具合に、限られた国の資料しか入手できないのが残念である。また、1968年以後の最近のデータは1975年までのものが、国の数を増やして出版されている。

1.2 「Cancer Incidence in Five Continents」について

これは1966年に第1巻が出て、現在第3巻まで刊行されている。がん登録を行なっている国または都市の資料を集めたもので、男42, 女44部位のがん罹患率が25ヶ国67人口集団について解析が可能である。年令訂正罹患率も、第3巻には世界人口を基準にしたものの他に Truncated Population を基準にしたものも掲載されている。罹患率という感度の高いものさしによって、世界各国の比較が可能であるという点で利用価値は高いが、各地域のがん登録のシステムがかなり異なっており、診断技術の差や、人口の年令構成の差、等も十分に考慮した上で使用しなくてはならない。

1.3 国際比較の環境要因の資料について

がんの死亡、罹患に影響を与える環境要因については、非常に多くのものが考えられる。今回は入手しやすく、しかも国際比較がある程度可能なものということで、人口密度、気温、降雨量（理科年表, 1975）と食物消費（FAO 1971）を使用した。国単位または都市単位での比較が可能なものとしては、国民所得などの経済学の指標、医療関係者の数、上下水道の普及率など国連やWHOで出している多くの数値がある。それをすべてがんと関連でみることも可能ではあるが、今後はやはりある程度標的を絞って検討することの方が望ましい。国際的に環境要因とがん死亡率または罹患率とを比較し、疫学的な因果関係を推論するにはまだ数多くの考察すべき問題があるが、何らかの形で手をつけ始めることは必要なことではないだろうか。

1.4 「悪性新生物死亡統計」について

日本のがん死亡率に関して、厚生省がまとめたものが、1960年および1973年の2回刊行されている。また Segi は「Cancer Mortality in Japan」(Segi, 1965)として、1899～1962年の死亡統計をまとめたものを1965年に発表している。今回我々の使用したものは1973年の「悪性新生物死亡統計」のみである。これには1958年～1971年の14年間のがん死亡数および死亡率が、男18, 女19の部位について掲載されている。もともになる資料は保健所を通して厚生省に届出られる死亡診断書による人口動態統計の形である。したがって、死因はほとんどが臨床診断と考えられる。部位によっては、病理剖検検報によって補正係数が算出されているものもあり（丸地, 1974, 松下 1975）、診断の難しい部位では、診落とし、診過しの率のかなり高いことも考慮しておく必要はある。

1.5 日本国内の県別環境要因について

今回は理科年表から気温、降雨量、人口密度、第一次産業従事者数のみを取りあげた。他の研究者らの報告でも食物消費（豊川, 1976, Maruchi et al., 1977）がとりあげられており、それなりの成果を挙げている。ここでも国際比較の所で述べたと同様な標的を絞ること、均質な資料を得ることなどの注意が必要とされることはいうまでもない。

1.6 山梨県内市町村別比較について

がんの部位別死亡率は、重松らの全国市町村別死亡率の資料から、山梨県のを借用し、特に肝がん・肝硬変死亡率については県庁厚生部より資料を提供してもらって計算を実施した。これらはすべて人口動態統計の資料であり、1968年以後のものしか入手できなかった。診断の信頼度については、前述したものと同様であるが、山梨県の場合には剖検施設も少なく、全国水準より信頼度はやや低いと考えておいた方がよい。（稲葉ら，1977 a, b）

環境要因としては、山梨県総合地域指標（1975）にとりあげられた指標のうちから、いくつかを選択した。市町村単位で入手しやすいもの（所得、上水道普及率、各種死亡率など）が選ばれ、がん発生との関連が考えられるが資料としては市町村単位では得にくいもの（アルコール摂取、喫煙、食物消費など）が除外されるのはやむを得ない。

1.7 Case-Control Study 資料について

我々は1976年より、山梨県内の主要病院の協力を得て、肝がん・肝硬変患者の症例対照研究を行ってきた。その資料を若干利用することにした。診断は臨床医の判断に一任したが、組織所見まで調べているものもかなり多く、少くとも人口動態統計のものより信頼度は高い。しかし、調査という性質上、症例数は3年間で肝がん64例、肝硬変153例とかなり少なくなっている。環境要因としては、問診による居住歴、職歴、既往歴、飲酒、喫煙の程度、食生活の他に、日本住血吸虫症皮内反応、HB_s抗原およびその抗体の検査を実施した。詳細については別に公表の予定である。（山本ら，1980）

2. 標準化の方法

死亡数にしても死亡率にしても、異なる人口集団の比較をするためには、年齢による補正がまず必要とされる。この年齢補正の方法はこれまでに多くの提案がされており、いずれも一長一短のあることが知られている。

我々がこれまで使用してきた方法についてここで概観しておきたい。

2.1 直接法による年齢訂正死亡率

直接法は最もよく使用される標準化の方法で、一つの標準人口を選び、取扱うべき人口集団の年齢階級別死亡率(a_i)をこの標準人口の対応する年齢階級別人口(p_i)にかけ、各群の和を求めたもの($\sum a_i p_i$)を標準人口の総数で割ることによって訂正死亡率を得る方法である。（表2-1）最近はその信頼限界を求めるために標準誤差も計算されることが多くなっている。（Ching Long Chiang, 1961）ここでよく問題とされるのは、標準人口としてどのような集団を設定するかということである。国際的な比較をする場合には、表2-2に示される

表2-1 直接法による訂正死亡率の算出

年令階級	観察人口集団死亡率	標準人口	期待死亡数
0 - 4	a_1	p_1	$a_1 p_1$
5 - 9	a_2	p_2	$a_2 p_2$
10 - 14	a_3	p_3	$a_3 p_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	a_i	p_i	$a_i p_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
80 才以上	a_n	p_n	$a_n p_n$
合 計		Σp_i	$\Sigma a_i p_i$

$$\text{訂正死亡率} = \frac{\Sigma a_i p_i}{\Sigma p_i}$$

$$\text{標準誤差} = \sqrt{\frac{\Sigma a_i (1 - a_i)}{\Sigma q_i}}$$

(q_i : 観察人口集団の人口)

表2-2 標準人口の例

(UICC, 1976)

Age (in years)	African	World	European	Truncated
0-	2,000	2,400	1,600	-
1- 4	8,000	9,600	6,400	-
5- 9	10,000	10,000	7,000	-
10-14	10,000	9,000	7,000	-
15-19	10,000	9,000	7,000	-
20-24	10,000	8,000	7,000	-
25-29	10,000	8,000	7,000	-
30-34	10,000	6,000	7,000	-
35-39	10,000	6,000	7,000	6,000
40-44	5,000	6,000	7,000	6,000
45-49	5,000	6,000	7,000	6,000
50-54	3,000	5,000	7,000	5,000
55-59	2,000	4,000	6,000	4,000
60-64	2,000	4,000	5,000	4,000
65-69	1,000	3,000	4,000	-
70-74	1,000	2,000	3,000	-
75-79	500	1,000	2,000	-
80-84	300	500	1,000	-
85 and over	200	500	1,000	-
Total	100,000	100,000	100,000	31,000

ような人口が標準集団としてよく使用される。世界人口は Segi らによって提唱された全世界の人口をまとめて、総数 100,000 人となるように端数を切りすてたものである。アフリカ、ヨーロッパの人口はそれぞれの地域内での比較に用いられることが多い。Truncated Population は Doll によって提唱されたもので、若年層、高年層のバラツキが大きいことを考慮して、一定の年齢層のみで切って比較しようとするものである。

日本の国内での比較には、年次変化をみる場合には昭和 10 年（1935 年）を標準人口とすることが多いが、最近の地域別の比較を目的とする時には、昭和 35 年（1960 年）が使用されている。（厚生省、1975）

2.2 間接法による年齢訂正死亡率

間接法を使用するのは、直接法の計算に使用すべきデータが入手できない時が主であるが、計算すべき死亡率が非常に低い時にもよく用いられる。期待死亡数と実際の観察死亡数が比較できるので、この比を標準化死亡率比（Standard Mortality Ratio S. M. R.）として表現することもある。 χ^2 検定や t 検定も可能であり、標準誤差、信頼限界も比較的容易に求められる利点がある。算出方法は表 2-3 に示すように、標準人口集団の年齢階級別死亡率（ m_i ）に観

表 2-2 間接法による訂正死亡率の算出

年齢階級	標準人口 死亡率	取り扱うべき 人口集団の人口	期待死亡数	$S.M.R. = \frac{T}{\sum m_i q_i}$ (T は観察死亡総数) $\text{訂正死亡率} = M \times S.M.R.$ $\text{標準誤差} = \frac{1}{\sum m_i q_i}$
0 - 4	m_1	q_1	$m_1 q_1$	
5 - 9	m_2	q_2	$m_2 q_2$	
10 - 14	m_3	q_3	$m_3 q_3$	
-----	m_i	q_i	$m_i q_i$	
80 才以上	m_n	q_n	$m_n q_n$	
合 計	M (粗死亡率)	$\sum q_i$	$\sum m_i q_i$	

察人口人口集団の年齢階級別人口（ q_i ）をかけたものの総和（期待死亡数）を観察死亡総数で割り（SMR）これに標準人口の粗死亡率をかけたものが訂正死亡率となる。

ここでも標準人口集団は前述の直接法の場合と同様に任意に選択できるが、年齢階級別の死亡率を仮定することは難しいので、実際に存在する人口集団を用いるため、世界人口のような形の標準集団は考えない。

実際には日本の県別比較や市町村別比較の際には、5 年間（例えば昭和 33 年から 37 年）の死亡データを一括して、昭和 35 年（国勢調査の年）の全国人口を標準集団として、間接法による訂正死亡率を計算することが多い。

2.3 40 才以上特殊死亡率など

Segi らが日本の県別比較をする際に簡便であり、比較的信頼できるものとして提唱したも

のである。(Segi, 1965) 確かに通常のがんの発現は 40 才から急速に増加していくので、部位別の比較には利用できる。ただし、部位によっては白血病のように若年者に多発するがんもあるため、部位相互の関連を調べようとする時には使用できない。

これを乗り越えるために、我々は〔平均死亡年齢 - $2 \times$ (標準偏差)〕以上の年齢の特殊死亡率を計算してみた。(Yanai et al., 1978) 各部位別の平均死亡年齢の計算はやゝ煩雑であり、コンピューターによる他はないが、その結果は表 2-4 に示すようになった。平均死亡年

表 2-4 Means and standard deviations of the ages at death among the patients who died of cancers with various sites 1958-1962)

sites	male	female
1) buccal	60.6 14.2	61.5 16.6
2) esophagus	65.4 9.9	66.1 12.3
3) stomach	62.5 11.3	62.0 13.9
4) intestine	61.8 14.2	63.9 14.0
5) rectum	63.3 13.1	63.1 13.9
6) liver	62.3 11.9	64.3 12.8
7) bile duct	63.4 10.6	64.1 11.2
8) pancreas	60.1 11.0	60.5 11.9
9) larynx	65.1 10.1	65.6 11.7
10) lung	63.1 10.9	61.3 12.8
11) bone	52.8 19.6	53.6 20.0
12) skin	63.4 17.1	64.4 18.2
13) breast	64.3 13.2	55.6 14.1
14) uterus	—	56.9 13.0
15) ovary	—	48.4 16.7
16) prostate	70.7 10.1	—
17) bladder	67.8 11.2	66.7 11.3
18) kidney	53.8 20.3	51.0 23.2
19) lymphoma	49.1 20.2	51.3 20.2
20) leukemia	29.6 21.3	30.7 21.0

齢はほぼ 60 才前後、標準偏差が約 10 才のところにある部位が多く、Segi らの 40 才以上の死亡率がほぼあてはまることになるが、女性性器のがんや、リンパ腫、白血病などは上記の方法で計算したものと比較するとかなりずれてくることがわかる。

ただしこれは簡便法と呼ぶには手間がかかり、訂正死亡率のような信頼限界の計算もできないので、あまり一般的な利用にはすすめられない。

2.4 その他

死亡率や罹患率など人口 10 万対で示されるようなものと、気温、人口密度、食物消費などのまるで単位の異なるものとの関連をみようとする時、もう一つ別の標準化が必要になってくる。通常よく使用されるのは、平均値との差を標準偏差で割って無名数としたものである。

$$Z = \frac{x - \bar{x}}{s} \quad \left(\begin{array}{l} x : \text{実数}, \bar{x} : x \text{ の平均} \\ s : x \text{ の標準偏差} \end{array} \right)$$

ただし、これはマイナス符号をもつことになって取扱いにくいので、これに一定数 (K) を加えて正数になるようにして使用することもできる。例えば、以下のように定義される指標

$$T = \frac{(x - \bar{x})}{s} \times 10 + 50$$

は偏差値とよばれ、学力試験の解析によく用いられている。似たような考え方で数値の小さいものから順位をつけてしまうという方法もある。

いずれも問題となるのは、かけ離れた数値の意味がなくなってくるという点である。しかしこれはまた利点とも考えられるので、目的に応じて実数をそのまま使用したり、この種の標準化や順序尺度に変換して使用することが望まれる。

3. 相関係数について

疫学では通常、各資料から得られたデータの標準化を行なったのち、いろいろな形での関連、関連を計算する。ここではその際に使用する相関係数について、やや詳しく考察を試みる。

3.1 相関係数の各種の指標

相関係数とは、2 つの変数に関する多くの測定値の組 (x_i, y_i) ($i = 1 \sim N$) が与えられている場合に、それらの測定値間の関連の程度を量的に測定する指標である。相関係数は 2 組の変数が量的データ (間隔尺度, 比尺度) で与えられているか、質的データ (順位尺度, 名義尺度) で与えられているかによって、用いる公式が異ってくるが、2 組の変数が量的データである場合、ピアソンの積率相関 (Pearson's product moment correlation) と呼ばれる次式が用いられる。

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (1)$$

図3-1(a)のように、一方の変数(X)が大きくなるにつれて他方の変数(Y)も大きくなる
とき、上式は正の値をとる。このときXとYは正の相関を持つという。また図3-1(b)のよう

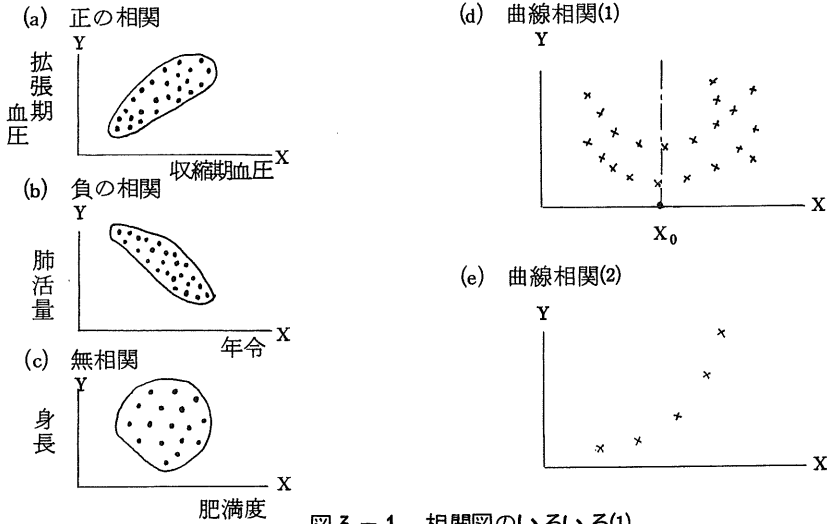


図3-1 相関図のいろいろ(1)

な場合には、負の相関、(c)の場合には、無相関(XとYは独立)となる。

ところで、図3-1(a), (b)のように、相関図上の各点が1本の直線上のまわりに集まるとき、
(1)式で示される相関係数の絶対値は1に近づく。しかし図3-1(d)のように、 $X \geq X_0$ のとき
XとYは正の相関、 $X < X_0$ のときXとYは負の相関を持つ場合、全体としてのXとYの相関
は消失する。このような場合、XとYの関係を、より明確にとらえるには、直線のかわりに、
2次曲線を用いると相関図上の各点がよりよく近似される。このとき、XとYは曲線相関を持
つという。曲線相関を表現する1つの方法としては、測定値を順位値におきかえて、次式で定
義されるスピアマンの“順位相関係数”(Spearman's rank correlation coefficient)

$$R_{xy} = 1 - \frac{1}{6} \frac{\sum_{i=1}^N (R_{xi} - R_{yi})^2}{N(N^2 - 1)} \quad (2)$$

(R_{xi} , R_{yi} は X_i , Y_i の順位値, Nは総データ数)

を用いることが考えられる。すると図3-1(e)のような場合、ピアソンの相関は1とはならな
いが、スピアマンの順位相関の(2)式は1.0となる。

表 3-1 相関比を求めるデータ

カテゴリー区分	1	2	...	m	計
	x_{11}	x_{21}	...	x_{m1}	
	x_{12}	x_{22}	...	x_{m2}	
	x_{1N}	x_{2N}	...	x_{mN}	
データ数	N_1	N_2	...	N_m	N
平均	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_m$	$\bar{x}$

表 3-2 関連係数を求めるデータ (分割表)

Y \ X	1	2	3	...	m	計
1	N_{11}	N_{12}	N_{13}	...	N_{1m}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	...	N_{2m}	$N_{2.}$
...	...	...	...	...	...	...
n	N_{n1}	N_{n2}	N_{n3}	...	N_{nm}	$N_{n.}$
計	$N_{.1}$	$N_{.2}$	$N_{.3}$	...	$N_{.m}$	N

表 3-3
ファイ相関係数を求めるデータ (四分表)

Y \ X	1	2	計
1	a	b	a + b
2	c	d	c + d
計	a + c	b + d	N

$N = a + b + c + d$

もう 1 つの曲線相関を表現する方法としては、X と Y の一方の変数 (例えば Y) がいくつかのカテゴリーに分かれる (年齢区分を 20 代、30 代、40 代、50 代に分けられるように) 名義尺度の変数と仮定ができる場合、次式で定義される相関比 (correlation ratio) が有用である。

$$\eta_x^2 = \frac{\sum_{i=1}^m N_i (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^m \sum_{j=1}^m (x_{ij} - \bar{x})^2} \quad (3)$$

(なお、 $\bar{x}_i$ 、 N_i 、 x_{ij} については表 3-1 を参照のこと) つまり、相関比とは、変数 X の全分散 s_x^2 に対する級間平方和の比で、X と Y の間に、直線的な対応がみられない場合でも、X と Y に何らかの対応がみられる、たとえば X と Y に曲線的な対応がみられる場合には η_x^2 は 1 に近づく。なお、X の区分の数 m が 2 の場合、 η_x^2 の平方根は

$$\eta_x = \frac{(\bar{x}_1 - \bar{x}_2)}{s_x} \sqrt{\frac{N_1}{N} \left(1 - \frac{N_1}{N}\right)} \quad (4)$$

となり、点双列相関係数 (point biserial correlation) となる。

二つの変数の関連が症状 (X) と経過 (Y) といったように、いくつかのカテゴリーの組み合わせで与えられている場合 (表 3-2 参照) の頻度表はクロス集計表 (または分割表) とよばれる。このとき X と Y のカテゴリーの組み合わせに、何らかの対応が見られるか否かを検定するには

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{\left(N_{ij} - \frac{1}{N} N_{i.} N_{.j}\right)^2}{\frac{1}{N} N_{i.} N_{.j}} \quad (5)$$

が自由度 $(n-1)(m-1)$ のカイ二乗分布にしたがうことを用いればよい。これに対して、データ総数 N に無関係な X と Y の関連の強さを示す指標としては、クレマーの関連係数を用いればよい。

$$C_1 = \sqrt{\chi^2 / N(r-1)} \quad (r \text{ は } m, n \text{ の小さい方})$$

この値は、2 つの名義尺度からなる変数のカテゴリー区分にどの程度の対応関係があるかを示すもので、ふつう、X と Y の関連 (association) を示す指標といわれる。

さらには、ピアソンの関連係数

$$C_2 = \sqrt{\chi^2 / (N + \chi^2)}$$

が使われる。 C_1 、 C_2 の上限はともに 1 になる。

なお、X, Y のカテゴリー数がともに2の場合（表3-3 参照） ϕ_1 は、ファイ相関係数

$$\phi = \frac{(ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad (6)$$

に一致する。しかし、上記の中の値は小さくなりやすいので、血圧値(X), コレステロール値(Y)を高い、低いといった二つのカテゴリーで区分して相関係数を求める場合は、X, Y のもともとの分布がほぼ正規分布にしたがうものと仮定できるから、近似的に

$$r_t = \cos \left(\pi \frac{\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right) \quad (7)$$

によって定義される4分相関係数（tetracholic correlation）を用いた方がよい。

（例1） 右のような2×2の分割表で
ファイ相関と4分相関を計算すると、前者が0.189 後者が0.310
と明らかに後者の方が大きくなる。

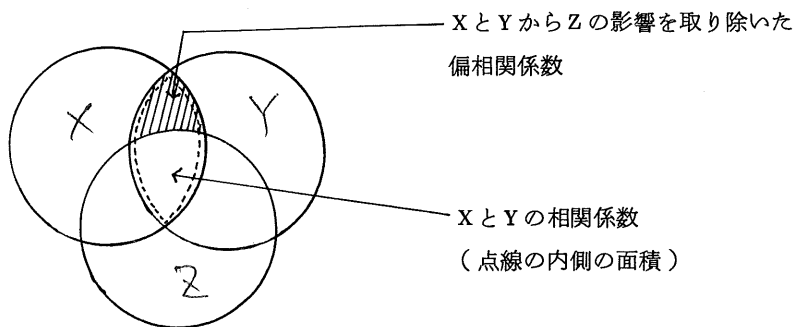
タバコ	肺癌	コントロール	計
20本以上	60	80	140
20本以下	40	120	160
計	100	200	300

この他に癌の疫学データの解析に有用と思われる相関係数としては、次式で定義される偏相関係数（partial correlation coefficient）がある。

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2} \sqrt{1 - r_{YZ}^2}} \quad (8)$$

上式は、図3-2のベン図形で示されているように、XとYの相関から第3の変数Zによって説明される相関を取り除いたものである。

図3-2 偏相関係数のベン図形による表現



または $Z = (Z_1, Z_2, \dots, Z_p)$ と影響を取り除く変数が多数ある場合には、

$$r_{XY \cdot Z} = \frac{r_{XY} - R_{XY \cdot Z}}{\sqrt{1 - R_{X \cdot Z}^2} \sqrt{1 - R_{Y \cdot Z}^2}} \dots \quad (9)$$

を用いればよい。（ただし、 $R_{X \cdot Z}^2, R_{Y \cdot Z}^2$ はX, Y に対する $Z = (Z_1, \dots, Z_p)$ に基づく重相関係数の平方。また、

$$R_{XY, Z}^2 = \sum_{i=1}^p \sum_{j=1}^q r_{XZ_i} r_{YZ_j} r_{Z_i Z_j}^{(-1)} \quad (10)$$

(ただし, $r_{Z_i Z_j}^{(-1)}$ は, Z の相関行列の逆行列の (i, j) 成分)

表3-4 相 関 係 数 の 種 類

名 称	変 域	適用できる変数の種類
2 変 数 の 相 関 係	ピアソンの 積率相関係数 r	$-1 \leq r \leq +1$ 距離尺度・比例尺度
	点相関係数 ϕ (ファイ)	$-1 \leq \phi \leq +1$ 名義尺度・順序尺度
	周辺分布によって 最大値・最小値は 制限される	
	スピアマンの 順位相関係数 r	$-1 \leq r \leq +1$ 順序尺度
	四分相関係数 r	$-1 \leq r \leq +1$ 距離尺度・比例尺度
	点双列相関係数 r	$-1 \leq r \leq +1$ 1変数(X)は名義尺度, 順序尺度 1変数(Y)は距離尺度, 比例尺度
3に相 関係 数 以 上 の	相関比 η (イータ)	$0 \leq \eta \leq +1$ 1変数(X)はどの尺度でもよい 1変数(Y)は距離尺度, 比例尺度
	重相関係数 r	$0 \leq r \leq +1$ 距離尺度, 比例尺度
	偏相関係数 r	$-1 \leq r \leq +1$ 距離尺度, 比例尺度

3.2 相関係数についての性質

相関係数の大きさは, 分析に用いる標本データの選び方によって著しく変わってくことに注意しなければならない。

例えば, 一方の変数 Y のデータが何らかの理由である値 (Y_0) 以上のデータしか集められなかった場合, X と Y の相関係数を求めると, 図3-2(a)のように, Y の値に制限のない場合の標本データによって計算される相関係数より小さくなりやすい。この理由から, 一般に, 症状が比較的類似している患者のデータによって計算される各種の検査値間の相関係数は, そうでない場合に比べて小さくなる。次に, 図3-3 b (16頁参照) のように, 母集団から選択された標本が, 著しく偏った標本がある場合, 母集団では, プラスの相関がみられるのに, 標本ではマイナスの相関がみられることがある。

(例2) 表3-5に世界の24ヶ国のデータによって計算される6つの部位の訂正死亡率間の相関係数と、我が国の都道府県別に計算される間接法にもとづく訂正死亡率間の相関係数を示した。(ともに男子のデータで、外国は1950～1966, 日本は1958～1971年の平均を用いた)

表3-5 悪性新生物の7つの部位に関する相関係数

悪性新生物の部位		(上段は24ヶ国, 下段括弧内は日本46都道府県のデータを用いて計算したもの)				
1	食道	1.000 (1.000)				
2	胃	0.344 (0.385)				
3	腸	-0.221 (0.278)	-0.729 (0.548)			
4	直腸	-0.175 (0.405)	-0.339 (0.589)	0.754 (0.574)		
5	肺	-0.001 (0.203)	-0.107 (0.244)	0.516 (0.540)	0.589 (0.377)	
6	皮膚	-0.188 (-0.266)	-0.429 (-0.607)	0.462 (-0.513)	0.097 (-0.502)	0.114 (-0.118)
7	白血病	-0.387 (-0.155)	-0.603 (-0.347)	0.423 (-0.073)	0.329 (-0.255)	0.138 (0.229)
		1)	2)	3)	4)	5)
						6)

表3-5から明らかに、同一の部位間で、外国と日本における相関係数の大きさに最も大きな相異が見られるのが、胃と腸で、外国-0.728, 日本0.548である。この他、皮膚と腸、胃と直腸の間にも相関係数に大きな差がみられる。

外国のデータは24ヶ国、しかもその殆んどが西欧諸国に偏っているという欠陥もあり、外国のデータを用いた相関係数の方が、より真実を反映しているとは、必ずしも断定できないが、外国のデータにもとづく相関の方が世界各国を母集団とした場合の母集団係数の推定値としては好ましいことは明らかである。(外国と日本の相関の相違を説明するには、第3の変数に関する偏相関を求めることが有効であるかもしれない。)

この他、標本データのとり方によって、相関係数の大きさはさまざまに変化する。図3-3(c)のように、5つの部分集団の内部では、負の相関がみられているが、全体では、正の相関がみられることがある。このような傾向は、部分集団を20才代、30才代、40才代といったような年齢区分にとった場合、加齢によってその値が増加する2つの(臨床)検査値間の相関にみられることが多い。(収縮期血圧と拡張期血圧の相関係数は各年齢区分別に計算すると、低い値となるが、各年齢区分をまとめて計算すると高い値が得られるのはその一例である。)さらに図3-3(d)に示したケースは注目に値する。この図に示された2つの検査値X, Yにおいて、Xでは女子、Yでは男子の平均値が高く、しかも男女いずれもXとYに強い相関がみられる。しかしこの場合、図から明らかに、男女あわせて相関を求めると、相関はマイナスになる。こ

のように、標本データがいくつかの部分集団に分別され、それぞれの部分集団において、測定値の平均値に著しい差がみられる場合には、それぞれの部分集団別に相関係数を求める方が全体をまとめて相関係数を求める場合に比して事象の本質をより正確に反映することになる。

なお、相関係数の求め方には、個人を単位とした個人相関と地域を単位とした地域相関（または生態学的相関）があるが、地域相関には次のような問題がある。

〔例3〕 4で述べる方法にしたがって計算された20（男子18，女子19）の部位のがんの都道府県別死亡率の異なる年次間の相関係数を求めたところ表3-5のような値が得られていることがわかった。

このことは、1958年から1971年にわたる期間中において、がんの各種の部位の死亡率は地域的にみて、極めて安定していることを示すものであるが、ここで、それぞれの年次間における相関係数と各部位の10万人当たりの粗死亡率（昭和46年）の相関係数を求めてみると、男子の場合、 $0.572 (T_1 \sim T_2)$ 、 $0.423 (T_2 \sim T_3)$ 、 $0.516 (T_1 \sim T_3)$ 、女子の場合、 $0.455 (T_1 \sim T_2)$ 、 $0.340 (T_2 \sim T_3)$ 、 $0.493 (T_1 \sim T_3)$ ときわめて高い相関が得られた。

すなわち、年次間の相関の強いほど粗死亡率が高いことが示された。

この事実は、地域相関（または間接相関）の大きさが、個々の地域あたりのデータ総数の増加にともなって増加していくことを示すもので、興味ある現象といえよう。

つまり、地域相関の大きさの解釈にあたっては、地域単位あたりのデータ数についての考慮が必要であることを示唆するものである。

図3-3 相関のとり方

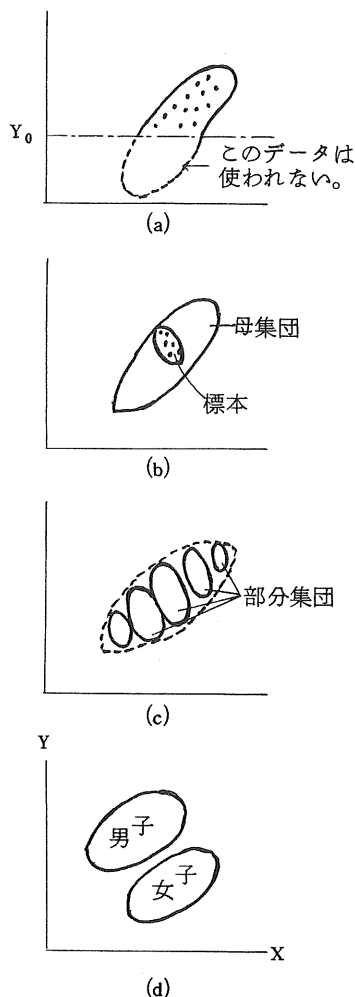


表 3-6 癌の部位別地域別死亡率の異なる年次間の相関係数
(死亡率は間接法の訂正死亡率)

Sex Periods Sites	Male			Female		
	T1-T2	T2-T3	T1-T3	T1-T2	T2-T3	T1-T3
1) Buccal	0.556	0.673	0.476	0.551	0.679	0.519
2) Esophagus	0.932	0.923	0.862	0.937	0.914	0.848
3) Stomach	0.971	0.932	0.934	0.969	0.935	0.928
4) Intestine	0.609	0.539	0.523	0.481	0.562	0.419
5) Rectum	0.606	0.753	0.551	0.545	0.401	0.448
6) Liver	0.906	0.926	0.844	0.856	0.912	0.843
7) Bile duct	0.820	0.814	0.770	0.777	0.795	0.694
8) Pancreas	0.685	0.814	0.694	0.672	0.826	0.693
9) Larynx	0.470	0.575	0.497	0.511	0.634	0.521
10) Lung	0.701	0.687	0.655	0.792	0.801	0.616
11) Bone	0.507	0.271	0.232	0.427	0.156	0.163
12) Skin	0.571	0.680	0.499	0.400	0.619	0.331
13) Breast	0.225	-0.200	-0.212	0.405	0.561	0.329
14) Uterus	—	—	—	0.403	0.310	0.424
15) Ovary	—	—	—	0.450	0.432	0.391
16) Prostate	0.406	0.313	0.454	—	—	—
17) Bladder	0.486	0.412	0.356	0.411	0.629	0.339
18) Kidney	0.286	0.561	0.311	0.645	0.258	0.127
19) Lymphatic tissue	0.846	0.880	0.847	0.812	0.836	0.878
20) Leukemia	0.666	0.540	0.483	0.290	0.402	0.336

(T₁ : 1958 - 1962 , T₂ : 1963 - 1971 , T₃ : 1968 - 1971)

4. 多変量解析について

多変量解析法とは、多くの個体について測定された多種の観測値からなる変数が与えられている場合、それらの変数を個々に独立させずに、変数間の相互の関連を分析する手法の総称で、電子計算機の発展とともに、ここ10数年来、経済学、心理学、農学、生物学、医学の分野で幅広く使用されている。さらに、医学の分野においては計量診断の方法、さらに prospective study にもとづくさまざまな risk factor の解析に用いられるようになってきた。

4.1 多変量解析法の分類

多変量解析の手法は表 4-1 に示したように、多数あるが、その分類の基準は次の 3 つに大別される。

- (1) 外的基準のありなし
- (2) 説明変数または基準変数の個数
- (3) 説明変数（または基準変数）の各尺度が名義尺度または間隔尺度のいずれからなるか。

表4-1 多変量解析法の分類

		説明変数		基準変数	
		名義尺度	間隔尺度	名義尺度	間隔尺度
外的基準のある場合	重回帰分析	—	多数	—	1
	正準相関分析	—	多数	—	多数
	重判別分析（正準分析）	—	多数	多数	—
	（線型）判別分析	—	多数	2	—
	多重ロジスティック関数	—	多数	2	—
	数量化第一類	多数	—	—	1
外的基準のない場合	数量化第二類	多数	—	多数	—
	主成分分析	—	多数	—	（多数）
	変形主成分分析	—	多数	—	（多数）
	数量化第三類	多数	—	—	—
	クラスター分析	—	（多数）	—	多数
類似性指数を用いる	因子分析	—	（多数）	—	多数
	数量化四類，多次元尺度構成法，最小次元解析，潜在構造分析				

まず，外的基準とは，測定目標のようなもので，医療検査値を説明する場合，患者の病名，またはある疾患の重症度，さらには，疫学調査における prospective study を基盤にすると，被調査者が追跡期間中にある疾患になるかならないかの情報が外的基準となる。外的基準がある場合の分析法は，外的基準が病名のように名義尺度（分類）の場合，判別分析，疾病の重症度のように間隔尺度の場合，重回帰分析になる。さらに，分析に用いる説明変数が名義尺度からなるカテゴリー変数である場合，重回帰分析は数量化一類，判別分析は数量化二類に対応する。また，prospective study において，調査の開始時に $X = (x_1, x_2, \dots, x_p)$ という P 個の変数の観測値で特徴づけられる患者が，ある一定年度を区切った期間中に死亡などの事故をおこす確率を

$$P(x_1, x_2, \dots, x_p) = \frac{1}{1 + \exp\left(-\sum_{j=1}^P b_j x_j\right)} \quad (1)$$

で表わす多重ロジスティック曲線による解析は，最近，多くの疾病に適用され，成果をあげている。上記の多重ロジスティック曲線における係数 b_j は，データに正規性が仮定されない場合でも，いわゆる判別関数によって得られた判別係数に，きわめて近い解が得られる利点があるが，時間の変数が考慮されていないので，この点を補う方法としては，多重ワイブル分布による解析，検査時点から一定時間 t 経過した後の死亡率曲線を，

$$\lambda(t|X) = \exp\left(-\sum_{i=1}^p b_i x_i\right) \lambda_0(t) \quad (12)$$

として定義する Cox (1971) の方法等が有効である。

外的基準がない場合の分析法には、与えられた変数間の相関係数を分析して、少数の合成された変数に還元する主成分分析、さらには、変数を、いくつかの潜在因子の和（1次結合）に分解する因子分析法があるが、最近では、変数（または固体）間の類似の指標にもとづいて、それらをいくつかの群に分類するクラスター分析の方法が脚光をあびている。

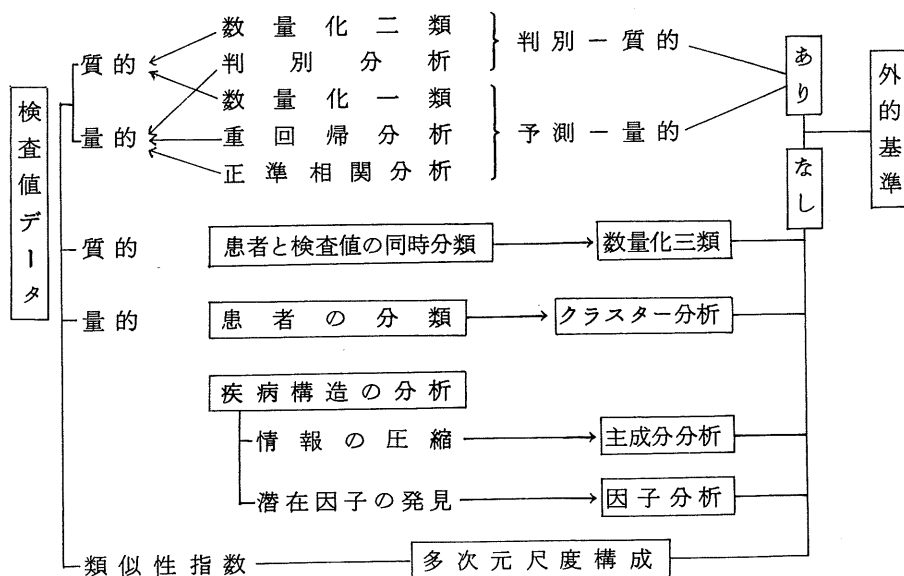
これまで、計量診断の方法としては、病名を最初に定めて、検査値に判別関数を適用するのが通常であったが、ベーチェット病や SLE のように診断基準が明確に定義しにくい疾患が多数現われている今日、外的基準のない多変量解析の手法の適用によって、既存の診断基準に束縛されない新たな疾患の定義が生まれることが期待される。（病名とは多變量的に言及すれば、多次元空間の領域であり、疾病の診断とは、多次元空間上に設けられたある座標軸に症候ベクトルを射影することに他ならない。）

クラスター分析を用いた SLE の患者の分類、因子分析を用いたベーチェット病の症候群の多次元分類については、宮原（1975）、山本他（1973）を参照されたい。

ここで、図 4-1 に患者の疾病の予測、判別、疾病構造分析という観点から、多変量解析の各種の技法の相互関連を示した。（多変量解析の各種の技法についての詳細は、竹内・柳井（1972）、柳井・高根（1977）を参照されたい。）

つづいて、次節以降に、癌の疫学データを用いた因子分析、クラスター分析、重回帰分析、判別分析、数量化2類による分析例を示す。

図 4-1 多変量解析法の分類



なお、がんの部位別死亡率の地域的類似性を検討するもう一つの方法としては、クラスター分析がある。Burbank (1972) は、がんの各部位間の階層的構造 (Hierarchical structure) を容易に探ることができるという点で、因子分析に比べて、クラスター分析の有用性を主張している。しかし、クラスター分析によって得られた階層的構造が、病因学的構造 (Pathogenic structure) とは必ずしも対応しない点、さらにクラスター間の類似性が必ずしも明確にされない点で、われわれはこの主張には賛成できない。

4.2 因子分析によるがんの部位別地域別死亡率の分析

〔資料および分析方法〕

【分析1】厚生省より刊行された「悪性新生物統計」を用いて、20部位（男子18部位、女子19部位）についての性別、都道府県別の死亡率（昭和33年～46年）を(a)特殊死亡率（文献1）、(b)間接法による訂正死亡率（文献2）に変換し、因子分析のバリマックス法を用い、部位別の因子負荷量と県別の因子得点を算出した。なお、(a)を用いた分析では全年度を、第1期（昭和33年～37年）、第2期（昭和38年～42年）、第3期（昭和43年～46年）に分けて、死亡率を算出し男子の場合 $54 (= 18 \times 3) \times 54$ 、女子の場合 57×57 の相関係数行列に因子分析を適用した。さらに、この場合、5歳区分の年齢区別データにもとづく因子分析も行った。(b)による方法では、クラスター分析も実施した。）

【分析2】分析に用いた資料は「Cancer Mortality in 24 Countries」(Segi and Kurihara)で、男子13、女子14の部位を選び、1950年から67年間の各部位についての訂正死亡率の平均値を用い、【分析1】と同様の手法で因子分析を行った。さらに、得られた国別の因子得点とF.A.O. Production Year Bookより得られる20の食品群と4種の環境要因との相関を求めて、各因子の持つ意味を探った。

特殊死亡率にもとづく因子分析により、次の結果が得られた。

男子	負荷の高い部位	因子得点の大きい県
第1因子	リンパ組織、皮ふ、肝臓、膀胱、前立腺	九州全域、高知、徳島、岡山
第2因子	肺、喉頭、口腔	東京、大阪、兵庫、京都、神奈川
第3因子	胃、食道、腸（特に第1、2期）	奈良、和歌山、佐賀、秋田、山形
第4因子	すい臓、胆道	宮城、青森、秋田、新潟

女子	負荷の高い部位	因子得点の大きい県
第1因子	リンパ組織、皮ふ、肝臓、子宮	九州全域、高知、愛媛、鳥取
第2因子	肺、乳房、すい臓	東京、京都、鳥取、宮城、福岡
第3因子	胃、食道、直腸、肝臓	山形、新潟、鳥取、奈良、京都
第4因子	すい臓、胆道、白血病（特に第3期）	宮城、福島、新潟、埼玉

上記の結果から、男女にほぼ共通した4つの因子が得られ、これらの因子は、調査期間（昭

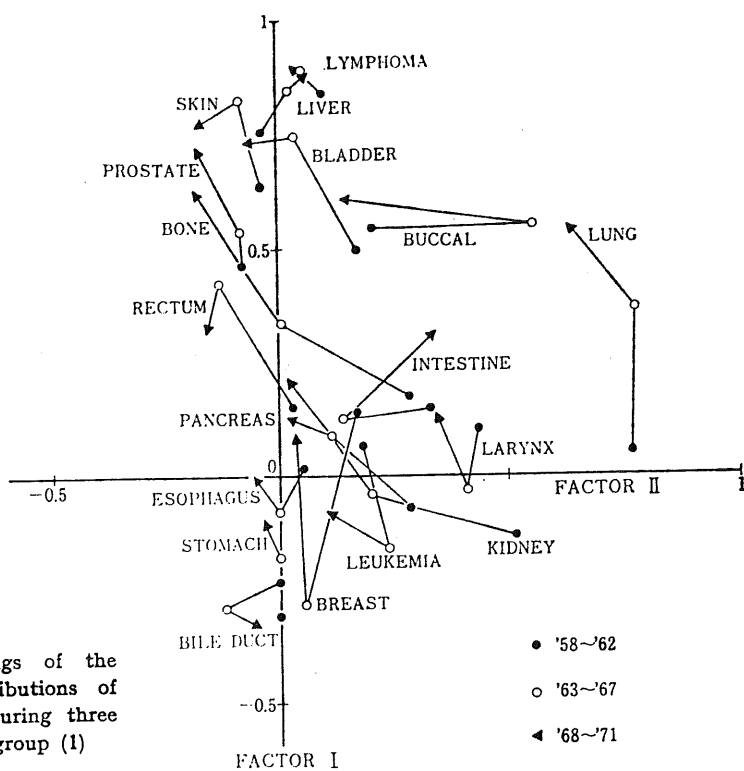


图 4-2(a) 18 种癌种的地理分布因素负荷量 (1) 男性组

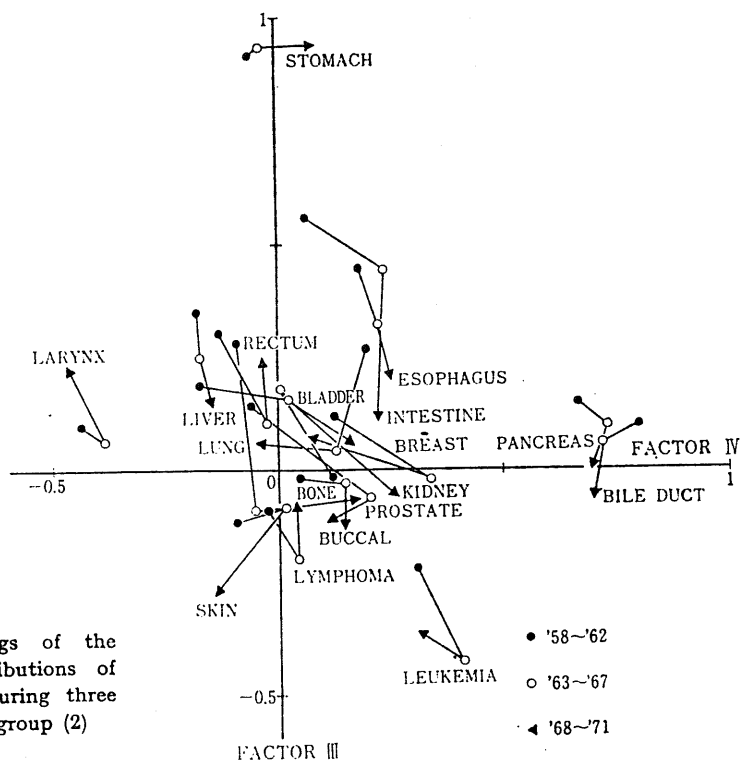


图 4-2(b) 18 种癌种的地理分布因素负荷量 (2) 男性组

和 33 年～46 年) 中、ほぼ安定した構造を有していることが確められた。

なお、ここで図 4-2 (a)(b) に、男子の場合についてのみ、第 1 因子と第 3 因子を縦軸に、第 2 因子と第 4 因子を横軸にして、それぞれの部位の時系列変化をふまえた因子負荷量を示した。第 1, 2, 3 期の間の点の位置が近い部位は、これらの期間に亘って死亡率の地域分布が類似していることを意味する。

なお、第 1, 2, 3 期の年次的にみた因子構造の変化で注目されることは、男女とも肺癌の場合、年次が進むにつれて第 1 因子の負荷が高くなること、さらに男子の食道、肺癌は第 3 因子の負荷が小さくなることである。同一のデータに間接法による死亡率にもとづいて因子分析を行った場合、先の分析で得られた第 2, 第 3 の因子が消失し、それにかわって、肺、直腸、口腔 (女子の場合、膀胱も) が 1 つの因子を構成し、胃、食道はそれぞれ独自因子を構成することが示唆された。つづいて、5 歳間隔の年齢別死亡率の因子分析の結果から、2 つの因子 (寄与率: 男子 81.2 %, 9.3 %, 女子 80.9 %, 10.3 %) が抽出された。第 1 因子は加齢にしたがって指数的に死亡率が増加する因子、第 2 因子は 60 歳時に死亡率がピークとなる因子で、それぞれ後天的 (postnatal), 先天的 (prenatal) な発癌因子と関連があるものと推定される。(図 4-3 参照)

第 1 因子は、白血病を除くすべての部位に高い因子負荷量がみられるが、第 2 因子は、正の方向に白血病、すい臓、腎臓、マイナスの方向に前立腺、皮膚、膀胱がんが高く負荷する。

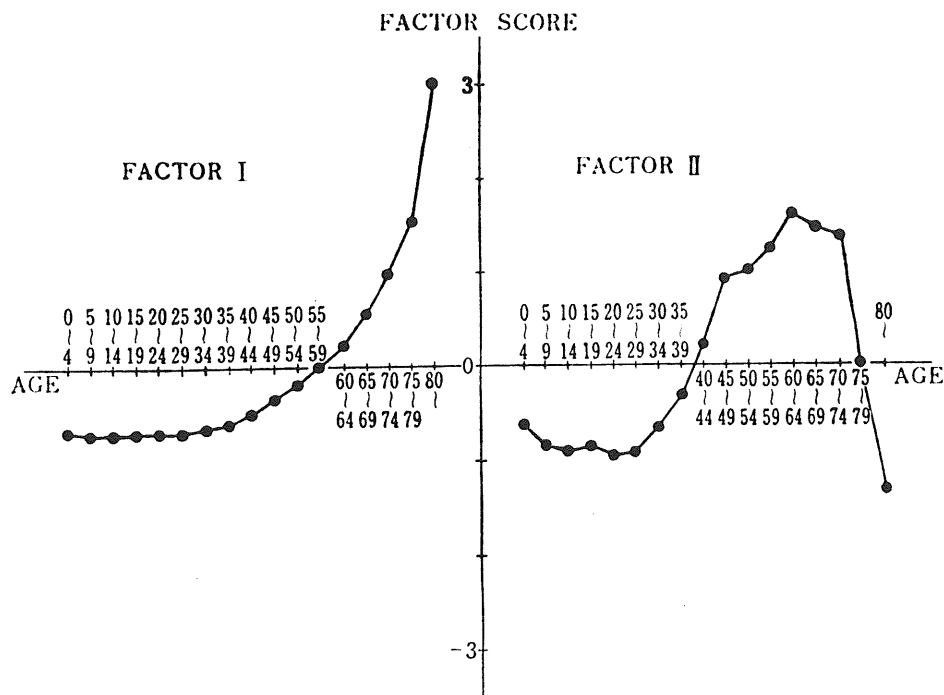


図 4-3 Curve of factor scores of the 5 years age group distributions of 18 cancer sites with three periods for males

すなわち、多くの部位の癌の死亡率の分布 x_j は、図 4-3 の第 1 因子 f_1 と第 2 因子 f_2 の重みづけの和、すなわち

$$x_j \doteq \alpha_{j1} f_1 + \alpha_{j2} f_2$$

によって表わされることを意味している。この重みを, Armitage & Doll (1954) の Multi-stage 理論と合わせて考察すると興味深い。

なお、上記の詳細な内容については Yanai et al. (1978) を参照してほしい。

このような年齢分布の解析に関しては、すでに稲葉他 (1975) が、8 つの特定疾患の有病率曲線に関しても行っている。

なお、間接法による訂正死亡率を用いた因子分析に関しても、上記の結果とほぼ同様の傾向がみられる。

世界 24 カ国のデータを用いて因子分析を行い、次の結果が得られた。

男子	高い負荷をもつ部位	高い因子得点をもつ国	関連する指標
第 1 因子	すい臓、前立腺、皮ふ、白血病 (以上 正) 胃、肝臓(以上 負)	米国、オーストラリア (正) 日本、イタリア、 チリ(負)	砂糖、動物性たん白(正) 穀類(負)、人口密度(負)
第 2 因子	腸、直腸、肺	デンマーク、英国	動物性油脂、熱量、肉
第 3 因子	喉頭、口腔、食道	フランス、イタリア	アルコール、野菜
第 4 因子	胃、食道、甲状腺	フィンランド、チリ	果物摂取(負)
女子	高い負荷をもつ部位	高い因子得点をもつ国	関連する指標
第 1 因子	白血病、皮ふ(以上 正) 胃、 子宮、肝臓(以上 負)	米国(白人)、ノルウェイ、 ニュージーランド(正) チリ、日本(負)	牛乳、動物性食品、砂糖 人口密度(負)
第 2 因子	腸、直腸、肺	カナダ、ベルギー、英国	油脂、熱量、肉
第 3 因子	喉頭、口腔、食道	アイルランド、スコットランド	植物性油脂(負)
第 4 因子	甲状腺、すい臓、肺	イスラエル、オーストラリア	野菜摂取(負)、雨量(負)

上記の表から、男女とも、第 1、第 2、第 3 の 3 つの因子に高く負荷する部位はほぼ同一であるが、第 3 因子のみ因子得点の分布、食物摂取との関連が男女間で全く異なる点が興味深い。

なお、重回帰分析の結果から、男子の場合、喉頭、口腔、食道癌の発症にアルコール多飲が何らかの形で関与していることが推定される。上記の内容の詳細については Yanai et al. (1980) を参照されたい。

4.3 重回帰分析の疫学データへの適用

前節の後半で示したところの世界 24 ケ国のデータを用いて、男子 13 の部位の癌の死亡率のそれぞれに対し、20 の食品別摂取状況と 4 つの環境要因を説明変数とした重回帰分析を行った。

重回帰分析の方法は、最初に基準変数と最も高い相関をもつ変数を選び、つづいて、最初に選ばれた変数の影響を取り除いた偏相関係数が最大となる変数を選ぶといった手順で行われる前進選択法を適用し、各々の部位について第5位までに選ばれた説明変数と、それらの変数を用いた場合に得られる重相関係数、個々の変数に対する偏回帰係数の大きさを表4-2に示した。例として、(1)の口腔がんのケースについて説明してみよう。まず最初にアルコールが選ばれ、つづいて、人口密度が二番目に選ばれる。この二つの変数を用いて、重相関係数を計算すると0.596となる。このような手順を繰り返していくことによって、アルコール、人口密度、果物、砂糖、穀類の五つの変数が選択され、それぞれの標準化偏回帰係数は、0.810, -0.263, -0.207, 0.688, 0.489となる。(したがって、口腔癌の死亡率(標準得点)の予測値、 $\hat{Y}$ は次式で予測されることになる。

$$\hat{Y} = 0.81 \times \text{アルコール} - 0.263 \times \text{人口密度} - 0.207 \times \text{果物} \\ + 0.688 \times \text{砂糖} + 0.488 \times \text{穀類}$$

(それぞれの変数も標準得点で表わされると仮定)選ばれた五つの変数によって最も高い値の重相関係数が得られた部位は、前立腺で0.897つづいて直腸で0.986、喉頭癌(Larynx)で0.878と続く。

表4-2には示されていないが、女子の場合は乳癌で0.843である。つづいて、偏回帰係数の値からみて興味ある点は、口腔癌、食道癌、喉頭癌において、いずれもアルコールの偏回帰係数がプラスで大きく、果物の偏回帰係数がマイナスであることである。

アルコールの過剰摂取と果物摂取の不足は、ビタミン不足を推測させることから、ビタミン不足が、口腔癌、食道癌、喉頭癌の一つのリスクファクターと考えることもできる。この他、肉類(meat)は直腸がんに対して、きわめて高い偏回帰係数をもつこと、砂糖がすい臓がんには高い係数をもつ反面、肝臓がんにはマイナスの係数をもつ事等、砂糖の消費量が文明水準の高さと関係する事を考慮すると、興味深い結果といえよう。しかし、これらの結果は、すべて国を単位とした地域相関にもとづく多変量解析であるので、次章5で述べられているような、地域相関と個人相関に関するさまざまな問題点を考慮しなければならない。

4.4 数量化二類を用いた胃がんの risk factor に関する解析

分析に用いた資料は、1967年1月から1972年5月までの5年5カ月間に、愛知県がんセンターと名古屋市内のある専門病院に入院した胃癌患者総計2,051名(男子1,366名、女子685名)、対照群として上記期間中に愛知県がんセンターを訪ずれ、精密検査の結果、異常所見を認めなかった者523名と萎縮性胃炎1,619名計2,142名(男子1,193名、女子949名)で、栗田英男氏(名古屋市立大医助教授)が上記患者に直接面接し個々の患者についての種々の疫学的事項や医療経過をretrospectiveに聴取したものである。

表4-2 The Results of Stepwise Multiple Regression Analysis (Male)

Site	Item	Beta	Cor.	M.Cor.	Site	Item	Beta	Cor.	M.Cor.
1.	Oral				8.	Larynx			
1.	Alcohol	0.910	(0.456)	0.456	1.	Alcohol	0.878 (0.760)	0.760	
2.	D.Population*	-0.263	(-0.300)	0.596	2.	Fruits	-0.363 (0.112)	0.797	
3.	Fruits	-0.207	(-0.013)	0.648	3.	Vegetable	0.181 (0.474)	0.823	
4.	Sugar	0.688	(0.150)	0.702	4.	P.Industry	0.476 (0.271)	0.845	
5.	Cereals	0.489	(0.065)	0.765	5.	Eggs	0.402 (-0.165)	0.878	
2.	Esophagus				9.	Lung			
1.	Alcohol	0.624	(0.398)	0.398	1.	Energy	0.224 (0.465)	0.465	
2.	Fruits	-0.348	(-0.086)	0.490	2.	Vegetable	-0.315 (-0.363)	0.535	
3.	Tree nuts	0.197	(0.096)	0.523	3.	Fish	-0.425 (-0.416)	0.606	
4.	D.Population	-0.218	(-0.042)	0.555	4.	D.Population	0.308 (0.209)	0.674	
5.	Meat	-0.173	(-0.160)	0.579	5.	Fruits	-0.143 (-0.078)	0.686	
3.	Stomach				10.	Skin			
1.	Meat	-0.706	(-0.659)	0.659	1.	Fish	-0.079 (-0.404)	0.404	
2.	Rainfall	0.465	(-0.041)	0.698	2.	D.Population	-0.195 (-0.386)	0.515	
3.	Tree nuts	-0.306	(0.149)	0.744	3.	Spices	0.192 (0.400)	0.573	
4.	Temperature	-0.205	(-0.085)	0.753	4.	Cereals	0.829 (0.106)	0.649	
5.	Energy	-0.297	(-0.585)	0.767	5.	Meat	0.826 (0.381)	0.766	
4.	Intestine				11.	Thyroid			
1.	Meat	1.092	(0.731)	0.731	1.	Fruits	0.712 (0.479)	0.479	
2.	Rainfall	-0.292	(0.120)	0.752	2.	Vegetable	-0.163 (-0.240)	0.636	
3.	P.Industry**	-0.452	(-0.605)	0.777	3.	Cereals	0.595 (0.039)	0.711	
4.	Cereals	0.680	(-0.501)	0.811	4.	Spices	0.462 (0.389)	0.769	
5.	Temperature	-0.321	(-0.243)	0.854	5.	Temperature	-0.409 (-0.219)	0.821	
5.	Rectum				12.	Leukemia			
1.	Energy	0.857	(0.707)	0.707	1.	Sugar	-0.078 (0.570)	0.570	
2.	D.Population	0.130	(0.313)	0.765	2.	Spices	0.590 (0.555)	0.633	
3.	Temperature	-0.643	(-0.509)	0.805	3.	Rainfall	-0.462 (-0.183)	0.732	
4.	Milk	-0.977	(0.403)	0.844	4.	Cereals	-0.239 (-0.493)	0.766	
5.	Sugar	0.521	(0.428)	0.896	5.	Milk	0.211 (0.543)	0.775	
6.	Liver				13.	Prostate			
1.	Sugar	-0.809	(-0.759)	0.759	1.	Spices	0.622 (0.680)	0.680	
2.	D.Population	0.256	(0.446)	0.799	2.	Tree nuts	-0.394 (-0.437)	0.812	
3.	Spices	0.203	(-0.279)	0.830	3.	Temperature	0.390 (-0.088)	0.835	
4.	Tree nuts	0.227	(0.524)	0.841	4.	Cereals	-0.560 (-0.540)	0.880	
5.	Fish	-0.191	(0.319)	0.854	5.	Eggs	-0.288 (0.207)	0.897	
7.	Pancreas								
1.	Sugar	0.502	(0.664)	0.664					
2.	D.Population	-0.498	(-0.634)	0.814					
3.	Milk	-0.241	(0.444)	0.828					
4.	Meat	0.309	(0.560)	0.839					
5.	Rainfall	-0.161	(-0.047)	0.850					

Beta : 標準化偏回帰係数

Cor. : 死亡率と各変数との相関係数

M.Cor.: 重相関係数

* Density of Population (人口密度)

** Primary Industry (第一次産業者の割合)

表4-3 分析に用いた対象者数

性別・ 年齢層 対象		男 子				女 子
		30 歳 代	40 歳 代	50 歳 代	60 歳 代	30～60歳代
胃 癌 群		129	200	200	200	640
対 照 群		200	200	179	92	750
内 訳	異常所見なし	100	66	18	7	280
	萎縮性胃炎	100	134	161	85	470

上記の結果のデータに数量化二類の適用により算出された重みを表4-4に示した。重みが正でその値が大きいほど、そのカテゴリーが胃がんに対して高危険因子であることを示す。しかし、この値は重回帰分析での偏回帰係数に相当するもので、各カテゴリー単独の胃がんに対する危険度を示すものでない。

ここで得られた表4-4の重みを巨視的な立場からみると職業（とくに30代を除く技能職）、職場環境（塵埃、肉体、疲労）、高塩食品摂取、多量米飯摂取、食事の速さ、多量飲酒、高喫煙などが胃癌の高危険因子群として浮び上がってくることになる。ここで年齢階層別に特徴的なことを二～三拾いあげてみると、多量飲酒は30歳代、40歳代、多量喫煙はレンジ（各カテゴリーの重みの最大値と最小値の差）のもっとも大きい30歳代に高リスクをもたらしやすい。これに対して加齢が進むにつれて、米飯多量摂取者は胃癌になりやすく、その反対にパン食者は胃癌になりにくくなること、さらに50歳代、60歳代と進むにつれて多量コーヒー摂取が胃癌の危険因子となってくることは、50歳代の胃癌の多くが慢性的な胃粘膜の刺激作用と加齢現象（胃腺の萎縮）とによって、幽門前庭部に生じやすいことを考え合えると興味深い結果である。なお同様の項目を用いて胃潰瘍群と対照群にもとづく判別を行うと各年齢層を通じてコーヒー摂取が強い危険因子となることが示されている。

なお判別値による胃がん群と対照群の2群の重心を分割点（cutting point）としたとき、各年齢階層を通じての正診率は70%台であった。図4-4が、判別値による分布で、2群がそれぞれ正規分布にしたがうものと仮定すると、50歳代では判別値が胃がん群の平均値より1標準偏差大きい7.85以上ならば、胃がんになる確率（尤度）は92%になる。

（なお、本研究は、柳井、高木と栗田英男氏（名古屋市立大医助教授）吉本泰彦（放射線影響研究所）との共同研究で、その詳細は柳井他（1977, 1979）に掲載されている。

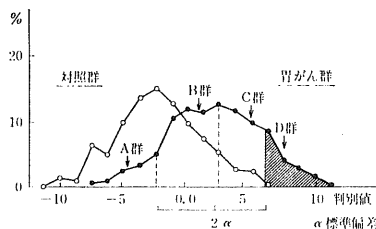


図4-4 胃がん群の集団分割

表 4-5 数量化二類による各項目別カテゴリーの重み

項目	カテゴリー	年齢階層			
		30歳代	40歳代	50歳代	60歳代
学歴	義務教育	0.58 -0.45	-0.25 0.25	-0.42 0.90	0.49 -1.47
	専門職	0.37 0.25	-0.01 0.65	0.09 0.08	-3.06 -0.16
職種	専任販売	0.07	-3.24	-3.79	-2.17
	運業・漁	-3.54	2.34	2.51	1.01
	技能作	0.48	-1.40	-0.97	0.31
	単純作	-1.73	-0.72	-0.27	2.57
職場	埃肉・体・疲・熱・埃・神・特記	2.00 1.22 -2.21 -4.54	2.93 -0.11 -2.30 -1.86	2.21 0.91 -3.02 -0.90	0.29 0.79 -2.69 -0.07
	埃肉・体・疲・熱・埃・神・特記	2.00 1.22 -2.21 -4.54	2.93 -0.11 -2.30 -1.86	2.21 0.91 -3.02 -0.90	0.29 0.79 -2.69 -0.07
収入 (月収)	5万円以下	-0.33	3.48	-1.06	-0.51
	6万円以下	3.78	2.28	0.88	-1.93
	7万円以下	1.11	2.80	1.94	0.85
	8万円以下	-3.70	-2.15	-1.53	-0.30
	9万円以下	-3.69	-0.08	-1.91	3.31
	10万円以上	1.41	-1.03	-2.45	-0.20
労働時間	8時間未満	2.05	-0.90	0.90	1.47
	8時間以上	-0.60	0.86	-0.22	-0.32
米飯量	パ・ン	2.91	0.68	-1.02	-3.88
	6.5杯以下	-1.74	-0.54	-0.29	-0.89
食事のさ	遅普通	1.11	-0.09	-1.75	1.84
	速普通	0.32	0.41	0.54	-0.29
食事のさ	冷普通	-0.85	-4.52	0.77	0.11
	熱普通	0.84	1.40	-0.05	0.30
コーヒー摂取	なと毎	0.57	1.67	-0.12	-0.88
	きどしき日	-0.90	-1.40	-0.06	0.71
牛乳摂取	なと毎	0.33	-1.19	0.22	-0.40
	きどしき日	-0.99	-0.13	-0.38	0.49
漬物摂取	なと毎	0.97	1.57	0.15	0.17
	きどしき日	3.58	0.59	2.60	2.87
高塩分食品摂取	なと毎	-0.50	0.49	1.35	0.32
	きどしき日	-0.59	-0.26	-1.20	-1.11
飲酒量	なと毎	-3.04	-1.22	-1.38	-2.24
	きどしき日	-0.77	-0.78	-2.79	-2.73
	180ml/日	0.76	0.26	1.46	1.99
	360ml/日	0.27	0.47	0.37	0.81
	540ml/日	-1.22	-1.47	-1.81	0.21
	540ml/日以上	0.31	-2.28	1.54	-1.78
喫煙量 (本/日) × (喫煙年数)	100	-1.52	0.30	-1.68	-0.88
	200	2.10	-0.91	0.42	0.76
	400	-0.44	3.40	3.01	1.02
	600	5.71	4.76	2.74	-6.68
	800	-2.03	-2.36	-4.64	-2.45
	1000以上	0.84	-1.68	-1.40	-2.29
薬服の用	なと毎	-1.96	0.62	0.16	-3.69
	きどしき日	1.10	-1.80	-1.28	-0.37
親族のん	2人	-0.12	0.69	-1.98	0.26
	2人	4.45	2.62	2.92	-1.33
	1人	7.40	3.26	3.83	2.44
	1人	0.34	-0.28	1.60	0.53
	1人	-0.17	-0.24	-2.74	-0.83
	1人	-3.30	0.94	-1.88	-3.28

* 1人が胃がん, 1人が他のがん

5. 地域相関と個人相関

統計的な手法を適用していく場合に、地域を単位とするものと、個人を単位にするものを同じように扱ってよいであろうか。またその結果が矛盾するようなものである時にはどう考えていったらよいのであろうか。この項ではこれらの問題について検討してみたい。

5.1 データの質について

まず地域単位のデータと個人単位のデータの内容の相異が考えられる。例えば気温、人口密度などは地域単位でしか考えられないデータである。一方、既往歴、家族歴などは個人単位でしか得られない。両者に共通に得られるデータとしては、食物消費、アルコール消費、職業、定期的実施されている調査や検査の成績（結核罹患率¹⁾、ツベルクリン陽性率²⁾など）がある。

地域単位のデータの利点は、きちんとした手続きさえ踏めば、比較的簡単に入手できること、最近やかましい個人のプライバシーをあまり侵害せずに解析ができること、結果の還元もそれほど厳密に考慮する必要がないことなどであろう。逆に個人データの利点としては、精度が高いこと、結果の解釈が直接的にできること、興味をより強くひくことができることなどであろう。

5.2 統計的方法の適用について

地域を単位とする場合には相関分析の手法が使用され、個人を単位とする場合には case control study や cohort study などの分析疫学的手法が用いられることが多い。

地域を単位とした場合は比較的連続量としての数値が得られることが多く、個人が単位の場合は逆にカテゴリカルデータ（あり、なしなど）しか得られないことが多いためであろう。ただし、カテゴリカルデータを用いても相関係数の算出は可能であるし、連続量をカテゴリカルデータに直すことも、情報量を失うという損失はあるが可能である。

そこで今回、すでに得られている山梨県のデータを利用して、地域単位と個人単位のそれぞれに全く同じ統計手法を試みた場合、どのような結果を生ずるかを検討した。結果は表 5-1 に示す。この表の相関係数は(1)式から求めたものであり、地域の場合は x_i は各市町村の訂正死亡率（ここでは、昭和 47～52 年の合計を、昭和 50 年の国勢調査人口を基準とする間接法で計算したもの）であり、 y_i が各市町村の H_B 抗原陽性率（山梨県血液センター献血者の資料を用いた）、日本住血吸虫症受診率（昭和 33～37 年の平均）である。個人の場合、 x_i は肝がんまたは肝硬変患者を 1、対照群を 0 とし、 H_B 抗原は RPHA 法または RI 法で陽性のもの 1、陰性のもの 0；日本住血吸虫症は既往歴ありのものを 1、なしのものを 0 として計算したものである。

相対危険度は McNemar (1961) の Matched Pair Analysis を用いて、算出したものである。地域の場合には、昭和 50 年の全国死亡率と比較して、 χ^2 検定を行ない、5% 以下の危険率で有

表 5 - 1 データ解析法による地域（市町村）と個人の比較

		HB _s 抗原	日本住血吸虫症	
肝がん	地 域	0.0 6 0	0.5 7 7 **	(N=64)
	個 人	0.4 2 7 **	0.2 9 4 **	(N=101)
相対危険度	地 域	1.0	8.0 *	(14 pair)
	個 人	1 0.0 * (59 pair)	9.5 * (56 pair)	
肝硬変	地 域	0.0 6 5	0.6 2 3 **	(N=64)
	個 人	0.2 1 0 **	0.3 4 7 **	(N=236)
相対危険度	地 域	0.6	1 2.0 **	(16 pair)
	個 人	7.7 ** (146 pair)	5.9 ** (150 pair)	

(** P < 0.0 1)
 * P < 0.0 5)

意に高い地域を肝がん、肝硬変（＋）として、ほぼ同じ人口規模の他の市町村を任意に選び、ペアを作った。HB_s 抗原陽性率に関しては、山梨県の平均値（1.7 %）より高率の地域を（＋）とした。日本住血吸虫症に関しては、昭和 33 ～ 37 年に、1 例でも患者の報告された市町村は（＋）とした。

個人の場合には、相関係数の時と全く同様にして、ペア分析を行なった。

結果をみると、日本住血吸虫症に関しては地域でも個人でも同様の結果が示されており、この因子が肝がんにも肝硬変にも関連していることを疑わせる。

一方、HB_s 抗原に関しては、地域を単位とした場合には、あまり明瞭な差を認めないが、個人単位では、はっきりとした関連を認めることができた。

以上、きわめて単純な形で統計的解析の比較を試みたわけであるが、実際には、県単位、国単位の変量解析を行なってみると、これらの関係がきわめて複雑な構造をもっているものと考えられる。

5.3 地域単位の個人単位で得られた結果が矛盾する場合の考察

上述の例では、HB_s 抗原と肝がん・肝硬変の間に、個人単位では非常に明瞭な関連が認められたにもかかわらず地域単位ではそれが出てこなかった。この場合には、HB_s 抗原陽性率の資料が、血液センターのものであって、データそのものに偏りがあることである程度の説明はつけられる。すなわち、供血者は、自発的に供血を申し出るような健康に関心の高い人であり、肝疾患等の疾患にかかっておらず、年齢も 20 ～ 59 才と限られているという点である。

一般に個人に単位とする調査で関連の見出された要因に対しては、地域単位で関係がないとされても、これを無視することはできない。データの精度は個人単位の方がずっとよいからで

ある。ただし、実際の予防活動へ結びつけていこうとする時、地域単位であまり関連を見出せないということが示されれば地域単位の前データを取り扱うよりは、個人を中心に分析が展開されていく方がより望ましいという示唆は得られるかもしれない。

一方、地域単位の相関が明瞭にみられるのに、個人単位での調査でははっきりしないという例もある。これは食物消費との関連でよく議論される。我々の解析でも砂糖の消費量はかなり多くの部位のがんと高い相関を示す。しかし、実際に調査は実施していないが、症例対照研究などではあまりはっきりした関連は出て来ないのではないかと考えている。これは、「相関のとり方」の項でも述べられているように、症例対照研究のような場合は、ある程度同質の集団から選ばれたものの関係になるため、差が小さくなって係数としては高くならないことが考えられるからである。砂糖の消費量などは、ある意味では文明発達のインディケーターとも考えられ、文明の発達と共に出てくる各種の発がん物質への暴露の総和としてみることも必要なのではないだろうか。豊川らが食物消費を類型化して、いくつかのパターンに分類し、それぞれの類型の中での疾患の傾向をとらえようとしているが、がんの疫学にもそのような姿勢が今後望まれよう。地域を単位として、ある種のがんと関連を認められた要因に対しては、個人単位の因果関係に結びつけていくためには、さらに多くの考察が必要とされるということである。

最後に前述したように地域を単位としたものでなければ得られないデータがある。これに関しては、個人のレベルにまでもってくるには、一層の手間がかかる。しかし、だからといってその解析をするのは無意味であるとは言いきれないように思う。地域単位の解析によって、その地域の特徴を把握することは、広い意味での公衆衛生に資するデータとなると考える。

5.4 地域相関と個人相関の数学的関連

一般に、個人相関の大きさと比較すると、地域相関の値は大きくなりやすい。このことをもう少し厳密に考察してみよう。

いま、全データが m 組の集団（各組には $N_1, N_2, \dots, N_m$ 個のデータが含まれていると仮定）からなる二つの測定値を x, y とすると、分散分析の考え方から、 x と y の分散 S_x^2, S_y^2 と共分散 S_{xy} は

$$S_x^2 = S_{xA}^2 + S_{xe}^2$$

$$S_y^2 = S_{yA}^2 + S_{ye}^2$$

$$S_{xy} = S_{xyA} + S_{xye}$$

のように分解される。

したがって、個人相関の指標となる通常のピアソンの積率相関係数 r_{xy} は、

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{S_{xyA} + S_{xye}}{\sqrt{S_{xA}^2 + S_{xe}^2} \sqrt{S_{yA}^2 + S_{ye}^2}}$$

$$\begin{aligned}
&= \frac{r_{XYA}}{\sqrt{1 + \left(\frac{S_{Xe}}{S_{XA}}\right)^2} \sqrt{1 + \left(\frac{S_{Ye}}{S_{YA}}\right)^2}} + \frac{r_{XYe}}{\sqrt{1 + \left(\frac{S_{XA}}{S_{Xe}}\right)^2} \sqrt{1 + \left(\frac{S_{YA}}{S_{Ye}}\right)^2}} \\
&= h_{XA} h_{YA} r_{XYA} + \sqrt{1 - h_{XA}^2} \sqrt{1 - h_{YA}^2} r_{XYe} \quad (13)
\end{aligned}$$

となる。ただし、 r_{XYA} は地域の大きさに重みづけした、各地域データ（平均値）間の地域相関係数、 r_{XYe} は、個人のデータから、それぞれの個人が所属する地域の平均を減らした値にもとづく、地域内個人相関係数である。（さらに、 h_{XA}^2 、 h_{YA}^2 は x 、 y の相関比）

ここで(13)式より次式が導びかれる。

$$\begin{aligned}
r_{XYA} &= \frac{1}{h_{XA} h_{YA}} r_{XY} - \sqrt{\frac{1}{h_{XA}^2} - 1} \sqrt{\frac{1}{h_{YA}^2} - 1} r_{XYe} \\
&= \frac{1}{h_{XA} h_{YA}} (r_{XY} - \sqrt{1 - h_{XA}^2} \sqrt{1 - h_{YA}^2} r_{XYe}) \quad (14)
\end{aligned}$$

ところで、地域単位の増大（地域数の減少）は、個人相関 r_{XYe} を^低高めるが、地域内部の異質性が高まり、相関比を小さくするので、全体として地域相関 r_{XYA} は^低低くなる。上記の事実^低は、Robinson (1950) によって導びかれ、安田 (1969) によって我が国に紹介されているが、疫学的データを用いた詳細な研究は行われていない。今後の研究課題というべきであろう。

6. 因果関係の推論をめぐって

疫学の大きな目的の一つは、因果関係の解明である。「がん」のような高令者に発症することの多い非感染性の疾患の原因は、単純な急性感染症とちがって「因果の緩」と呼ばれるような複雑な道筋を想定して解明していかなければならない。そのためには4で述べられような各種の多変量解析の技法が有効であろうと考えられる。また、「疫学」という立場からは、実験医学とちがって、厳密な実験条件を設定した因果関係の解明方法とはとることができにくい。どのような新しい統計的分析技法が発達したとしても、その基本となるのは、疫学の場合各変量間の相関関係でしかありえないともいえる。したがって、多数の変数間の相関関係のもつ情報を最大限に生かした統計的解析を行っていくことが必要とされるわけである。

6.1 偏相関と因果関係

X と Y の間に強い相関がみられる場合でも、 X が Y の原因である（または Y が X の原因）といったように、 X と Y の因果関係が推定できるとは限らない。

場合によっては、 X と Y の間に第三の変数 Z が介在して、その影響で、 X と Y の相関が大き

くなっていることも考えられる。例えば、ある薬の服用(X)がある疾患の発症(Y)と強い相関がみられたと仮定しよう。この場合、その薬(X)は高年齢(Z)になるほど多く服用し、その疾患(Y)も高年齢(Z)になるにつれて発症率が高くなると仮定しよう。すなわち、 r_{XY} , r_{XZ} , r_{YZ} に、いずれも高い相関が見られる場合、XとYの相関は、Zによって惹き起こされた疑似相関となり、薬の服用(X)が疾患(Y)の発症の原因とは推定できなくなる。

以上の例から明らかなように、相関の大きさのみから因果関係の有無を推定することは、きわめて危険である。ところで、Blalock(1964)は、3つの変数X, Y, Zの間に、(a) $X \rightarrow Z \rightarrow Y$, (b) $Y \rightarrow Z \rightarrow X$, (c) $X \leftarrow Z \rightarrow Y$ のような因果連鎖が存在すると予想される場合、XとYからZの影響を取り除いた(7)式で定義される偏相関係数がほぼ0になることを示した。例えば、妊娠前の一定期間中における妊婦の体重(X)、妊婦の食事量(Y)、出産児の体重(Z)の間に予想される因果連鎖モデルとして、 $X \rightarrow Y \rightarrow Z$, $Y \rightarrow X \rightarrow Z$ の2つが考えられるが、偏相関係数として、 $r_{X/Y \cdot Z/Y} = 0.5$, $r_{Y/X \cdot Z/X} = 0.1$ が得られたとすれば、後者のモデル、すなわち、妊婦の食事量の増加によって体重が増え、その結果として出産児の体重が大きくなったものと解釈する方が望ましいことになる。

また、Zが、XとYの間に介在する第三の変数群が $Z = (Z_1, Z_2, \dots, Z_p)$ と多変量で与えられる場合、(9)式で定義される偏相関係数がほぼゼロとなれば、

$$\begin{aligned} (a) \quad & Z = (Z_1, Z_2, \dots, Z_p) \begin{cases} \nearrow X \\ \searrow Y \end{cases} \\ (b) \quad & X \rightarrow Z \rightarrow Y, \text{ または } Y \rightarrow Z \rightarrow X \end{aligned}$$

のいずれかの因果連鎖が想定される。

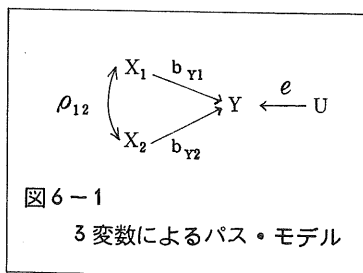
6.2 パス解析

この他、多くの変数が存在する場合の因果モデルの解明には、重回帰分析の考え方をその理論的支柱とするパス解析の方法が有効である。

パス解析は生物学者 Sewall Wright (1921) によって考案され、その後、主に社会学の分野で用いられてきた。図6-1をもとに、パス解析の規則を簡単に述べる。 X_1 , X_2 は外生変数 exogenous variable であり、内生変数 endogenous variable Yの直接的な原因であるとする。

この関係は矢印のついた直線によって表わし、関連の強さは直線上の記号によって示す。 b_{Y1} , b_{Y2} はパス係数 path coefficient, もしくは単にパラメータと呼ぶ。Yは、その他にUで示した誤差要因 disturbance の影響を受けている。2変数 X_1 , X_2 の間には相関があり、この関係は矢印のついた曲線で表わす。またパス解析では全ての変数は平均0, 分散1に標準化されているものとする。

図6-1の関係は、



$$Y = b_{Y1} X_1 + b_{Y2} X_2 + eU$$

とかけ、このような式を重回帰式と区別するために、構造方程式 structure equation と呼ぶ。上式は重回帰分析の手法で解くことが可能であるが、ここではパス解析の法則を用いる。

次に、パス解析の代表的な三つの法則を解説しよう。

1) 第一法則 (the first law)

2 変数 Y , Z の相関係数 ρ_{YZ} は、

$$\rho_{YZ} = \sum b_{YX_i} \rho_{X_i Z}$$

と表わせる。ここで、 b_{YX_1} は X_1 から Y へのパス係数、 $\rho_{X_1 Z}$ は X_1 と Z の相関係数であり、 $\sum$ 記号は Y の原因となる全ての変数 X_i について加え合わすことを示す。

例 1. 図 1 から ρ_{Y1} , ρ_{Y2} を求めると、

$$\rho_{Y1} = b_{Y1} \rho_{11} + b_{Y2} \rho_{21} = b_{Y1} + b_{Y2} \rho_{21}$$

$$\rho_{Y2} = b_{Y1} \rho_{12} + b_{Y2} \rho_{22} = b_{Y1} \rho_{12} + b_{Y2}$$

となり、

$$b_{Y1} = \frac{\rho_{Y1} - \rho_{12} \rho_{Y2}}{1 - \rho_{12}^2}, \quad b_{Y2} = \frac{\rho_{Y2} - \rho_{12} \rho_{Y1}}{1 - \rho_{12}^2}$$

と求まる。また $e^2 = V(eU) = V(Y - b_{Y1} X_1 - b_{Y2} X_2)$

$$= 1 + b_{Y1}^2 + b_{Y2}^2 + 2b_{Y1}b_{Y2}\rho_{12} - 2b_{Y1}\rho_{Y1} - 2b_{Y2}\rho_{Y2}$$

となる。

2) トレーシング法則 (the tracing rule)

2 変数 Y , Z の相関係数 ρ_{YZ} は、 Z と Y との間の全ての可能な道のり tracing を考えたとき、各道のりでのパス係数の積を全て加えたものに等しい。ただし、同一変数は 1 つの道のりで 2 度以上は通れず、中間の変数で矢頭が両方向から入っている変数は通れない。

例 2. 図 6-2 から各相関係数とパス係数の関係

を求める。

$$\rho_{Y1} = b_{Y2} b_{21}, \quad \rho_{Y2} = b_{Y2}, \quad \rho_{12} = b_{21}$$

これにより、 b_{Y1} , b_{Y2} のパラメータには 2 つの解があることになる。このような状態を overiden-

tified であるといい、逆にパラメータ数が式の数より多い場合を underidentified であるという。後者では全パラメータを決めることができない。

図 6-2 のモデルでは解を一意に決めるために、

$$\rho_{Y1} = \rho_{Y2} \rho_{12}$$

が成立する必要がある。これを制約条件 restriction と呼ぶ。制約条件が成立しない時には、モデルは妥当性を失う。制約条件がほぼ成立する時には、各パラメータの推定値の平均値をとればよい。

最後に、多くの変数の間に共通する原因がある場合について述べる。

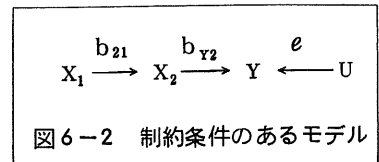


図6-3は、2つの共通因子 f_1, f_2 が変数 $X_1, X_2, \dots, X_6$ に関与しているとした場合のモデルであり、パス係数が0のものがあったとしてもよい。各変数 X_i ($i = 1, 2, \dots, 6$)は、 $X_i = a_{i1}f_1 + a_{i2}f_2 + e_i U_i$ と表わせる。トレーシング法則を用いると、変数 X_i と X_j との相関係数 ρ_{ij} は、 $\rho_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2}$ のように表わせる。このモデルは一般に因子分析として知られているもので、パス解析的な考えが、因子分析の応用にも使えることを示すものと考えられる。

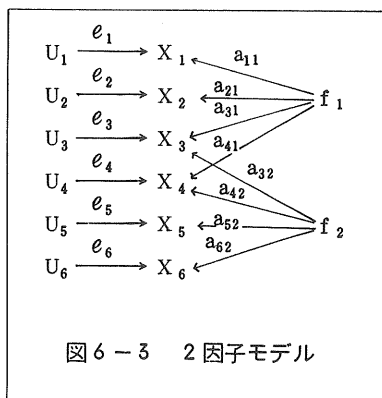


図6-3 2因子モデル

なお、パス解析に関する最近の文献としてはKenny (1979)とHeise (1975)がある。

以上述べたように、多くの変数の間にある連鎖的な構造が仮定される場合に、各変数間の結びつきの強さを分析していくものが、パス解析であるといつてよい。このようにして、パス解析により、一つの因果関係のモデルの統計学的検討が可能となる。

〔例1〕前述の山梨県の市町村の資料により、重回帰分析とパス解析を実施した結果を表6-1、および図6-4に示す。地域を単位とした解析では日本住血吸虫症の影響が強く出てくるのは前述したとおりである。ただこの解析では、各変数の相互間の関係がある程度明らかにされているといえよう。

表6-1 山梨県肝がんに関する重回帰分析(市町村を単位として)

説明変数	単相関係数	標準化偏回帰係数	重相関係数
X_1 HB _s 抗原陽性率	0.060	0.005	0.060
X_2 肝硬変死亡率	0.453	0.152	0.454
X_3 日本住血吸虫症受診率	0.577	0.482	0.589

(重相関係数の値は、 X_1 を説明変数に用いた場合が0.060、 X_1, X_2 を用いた場合が0.453、 X_1, X_2, X_3 を用いた場合が0.589となることを意味する。)

4

同様の解析を質的データに実施しようとする時は、数量化の考え方が必要となってくる。

重回帰分析に対応するのは数量化第一類である。これにしたがって実施した症例対照研究の結果を図6-5に示す。ここではHB_s抗原の役割が強く示されている。

パス解析の利点は、一つのモデル

図6-4 山梨県肝がんに関するパス解析

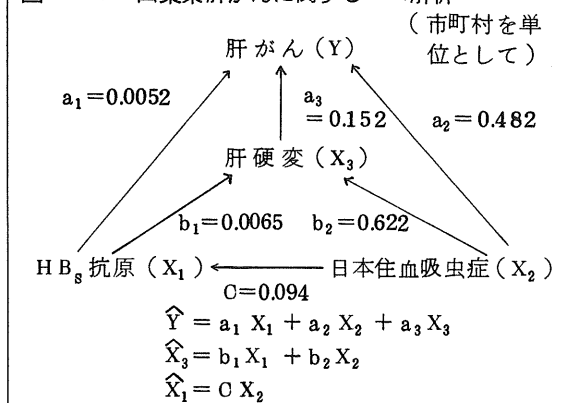
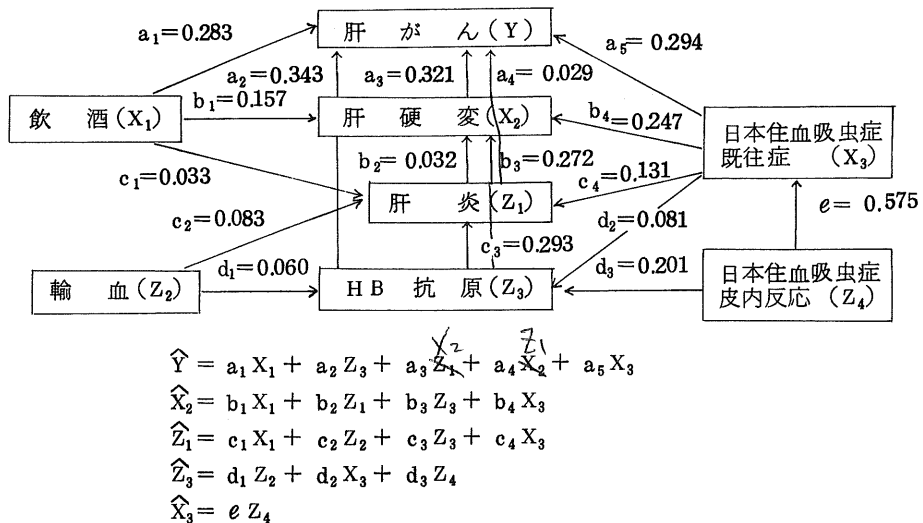


図6-5 山梨県肝がんに関するパス解析

(症例対照研究を資料として)



を設定することにより、無意味な変数を減らすことができ、これまでの知識を総合した形でモデルの適否を数量的に判断できることにある。ただし欠点としては、既存の知識のわくを越えることができないこと、取扱う変数が多くなるほど作業が複雑となり、時間がかかること、各変数の単独の影響を前提としているため、相加作用、相乗作用等の想定がしにくいことなどである。

6.3 相関関係と因果関係

ここでは相関関係を示す疫学データがどのようにして因果関係の解明に役立つかを中心に考察を加えてみたい。

a 使用するデータの意味づけ

コンピューターの発達と共に、多量データの一括処理が容易になり、ともすると何でもコンピューターに投入してみれば何か結果が出てくるであろうと、安易にデータを処理する傾向がある。

我々の今回の研究過程においても、いく分かはこのような傾向があったことを反省している。因果関係の解明を目指すためのデータであるならば、ある程度その方向に沿って解析するデータを整理しておくことが望ましい。例えば世界24ヶ国のがんの部位別死亡率を説明する要因としては自然現象(気温、温度、降雨量)、社会現象(第一次産業従事者の比率、人口密度)と食物消費量を取りあげたが、これらは並列的に解析するデータではなかったのではないかと考えられる。すなわち、自然現象がある程度社会現象を規定しており、食物

消費はそれらの結果としての要素が強い。がん死亡の原因を疫学的に探策し、自然現象との相関を認めたとしても、これは生物学的、社会学的要因を媒介として考慮しなければ、そのままでは説明したことにならない。

一つの方角としては、因果関係の究明をはかる時、気温とか降雨量のようなコントロール不可能な要因ははずして、食物摂取等ある程度予防対策も可能な要因にしぼっていく必要があるのではないか。

b 相関のとり方

すでに述べたように、相関係数の大きさ、対象とする集団のとり方でかなり変わってくる。その意味では、世界24ヶ国のようなとり方は果して妥当性があるかどうか疑問である。実際問題としてデータのない国はとり扱えないため、止むを得ないともいえるが、国単位の比較はもう少し慎重を期した方がよい。

その意味では日本の中の県別、市町村別資料はある程度平均したものが集められているという点で比較する意味がでてくる。

症例対照研究のような場合にも、母集団が必ずしも明確でないことをいつも自覚しておく必要はある。

6.4 リスクファクターの推定をめぐって

最後に、リスクファクター推定の問題に言及してみよう。

多くのリスクファクターが関与するものと予想される場合には、判別分析、重回帰分析、数量化一類、二類などの多変量解析の手法を用いることによって、多変数のリスクを求めるうえでのマッチングの手間を省くことができる。しかし、多変量解析の手法にも問題がないわけではない。周知のようにP個の連続量を説明変数とする判別関数の推定値は、

$$\hat{Y} = b_1x_1 + b_2x_2 + \cdots + b_px_p$$

と表される。ここで判別係数 b_j は変数 X_j が1単位増加したときの $\hat{Y}$ の増分を示すものであるが、日常の医療場面において最高血圧を下げるような治療をすれば、最低血圧も下がり、米食をパン食に切り替える栄養指導をすれば牛乳を飲む量や果物を摂る回数も多くなりやすい。したがって判別分析（または数量化二類）に含めた変数が異なる項目のカテゴリー間で強い相関が見られる場合には、個々の変数単独でリスクファクターを考慮することなく、相関の強い変数群を1つのセットとしてそれらの変数全体としてのリスクを評価すべきであろう。たとえば、牛乳摂取と米飯摂取量の間には逆相関があり、牛乳摂取と果物摂取には正相関がみられることから（豊川，1976），牛乳，パン食，果物といった西欧化された食生活のパターンが胃癌のリスクファクターの1つのセットとして抑制的に作用するものと考えるのは、その1例であろう。（富永，1979）

つまり、セットとしてのリスクファクターを考察するという意味は、判別関数（または数量化二類）の術語でいえば、判別値を高めるすべての要因が、全体としてがんのリスクファクタ

一になっており、個々の要因はどれもそれ自身では決定的な発病要因とはなっていないことを意味するものである。

すなわち予防医学的な観点からは、たとえば家系にがん患者が多発するような遺伝的な負荷が高い人は、食習慣を変えとか、禁煙をするといったように、判別値を低めるように生活習慣を変えていくことが望まれるわけである。

しかし上記の多変量解析におけるリスクファクターの解析は、あくまで各要因の相加作用（一次式による重みづけ）のみを前提としたものであるが、もちろん、二つの要因の積を一つの新しい変数として考察するいわゆる相乗作用の効果についても今後検討していかなければならない。

おわりに

3年間の成果を小冊子にまとめようと考え出したのは、それほど以前のことでない。多分、はっきり決めたのは11月の班会議のあとだったように記憶している。

いつも討論し合っていることも、いざ文章にしてみると、思ったほど簡単ではない。あれもこれも入れておきたいと思いながら、今年度中という制約のもとで、相当省略してしまった。でき上がってみると、ますます検討すべきことが増えてきたような気がしている。

とりあえずこの3年間の成果ということでまとめあげたもので、今後この問題に関する研究にはいろいろな形で関わっていくことになると思う。我々の浅学がさらけ出されて、多少恥ずかしい気もしているが、一つの叩き台として、出版した次第である。御一読頂ければ幸いである。

《文 献》（A B C順）

- Armitage, P. and Doll, R. (1954) "The age distribution of cancer and a multi-stage theory of carcinogenesis. Brit. J. Cancer, 8.
- Blalock, H.M. (1964): Causal Inferences in Nonexperimental Research, University of North Carolina, Chapel Hill
- Burbank, F. (1972). A sequential space-time cluster analysis of cancer mortality in the United States: Etiologic implications, American Journal of Epidemiology, 95, 393-416.
- Ohian, C.L. (1961): Standard Error of the Age-Adjusted Death Rate; Vital Statistics Special Reports, Selected Studies 47 (9)
- Cox, P.R. (1972): Regression models and life tables. J. Roy Statist. Soc. B, 34, 187 ~ 220
- FAO (1971) Production Year book, 435 ~ 441

- FAO (1976, 1977) : Monthly Bulletin of Agricultural Economics and Statistics, Vol 25 (4, 7-8), Vol 26 (1, 5)
- Heise, D.R. (1975) : Causal Analysis, John Wiley & Sons, New York.
- 稲葉裕, 高木広文, 柳井晴夫他 (1975) : 疫学調査における地域および年齢因子に関する研究, 日本公衛誌, 22 (2), 83 - 93
- 稲葉裕, 丸地信弘, 山本俊一, 高橋月容, 中村清純, 小野哲, 長橋捷 (1977 a) : 山梨県における肝がん, 肝硬変死亡と環境要因との関連, 日本公衛誌 24 (8) 465-473
- 稲葉裕, 高橋月容, 丸地信弘 (1977 b) : 山梨県の肝がん・肝硬変死亡に関する人口動態統計の解析, 日本公衛誌 24 (12) 811 - 815
- 稲葉裕, 高木広文, 柳井晴夫 (1979) : わが国における部位別がん死亡率の地理的分布に関する因子分析, 日本公衛誌 26 (2), 67 - 75
- Kenny, D.A. (1979) : Correlation and Causality, John Wiley & Sons, New York.
- 厚生省統計調査部編 (1960, 1973) : 悪性新生物死亡統計
- 厚生省統計調査部編 (1978) : 昭和 45 年 主要死因別訂正死亡率
- 宮原英夫 (1975) : 全身性エリテマトーデスのクラスタ分析, 医用電子と生体工学, 13 (3): 133 ~ 140
- 松下寛 (1975) : 日本の肝硬変と肝癌, 臨床科学 11 (3) 275 - 286
- 丸地信弘 (1974) : 病理剖検よりみた脾がんの臨床疫学的検討, 癌の臨床 20 (1) 136 - 139
- Maruchi, N. et al. (1977) : Relation of food consumption to cancer mortality in Japan with special reference to international figures. Gann, 68 : 1 - 13
- McNemar, Q. (1947) : Note on the sampling error of the difference between corrected proportions or percentages, Psychometrika 12, 153 - 157
- Robinson, S. (1950) : Ecological Correlations and the Behavior of Individuals, Sociological Rev., 15 : 351 - 357
- Segi, M. & Kurihara, M. (1960, 1962, 1964, 1966, 1969, 1972) : Cancer mortality for selected sites in 24 countries No. 1 ~ 6. Japan Cancer Society
- Segi, M. (1965) : Cancer Mortality in Japan, Tohoku University, Sendai.
- 高木広文, 豊川裕之, 柳井晴夫, 栗田英男 (1976) : 胃がん患者家族におけるがん発症頻度に関する研究, 日本公衛誌, 23 (12), 723 - 730
- 竹内啓, 柳井晴夫 (1972) : 多変量解析の基礎, 東洋経済新報社
- 富永祐民 (1979) : 発がん性の疫学的評価: 重松逸造編: 新しい疫学の方法論, ソフトサイエンス社, 224 - 275
- 豊川裕之 (1976) : 食生活と健康, 大修館
- 東京天文台編 (1975) : 理科年表, (丸善)

- UIOC(1966,1970,1976) : Cancer Incidence in Five Continents, Vol. I - III
- Wright, S. (1960) : Path coefficients and path regression : Alternative or complementary concepts, Biometrics, 16, 189 - 202
- 安田三郎, 海野道郎 (1976) : 「社会統計学, 改訂版」丸善
- Yanai, Inaba, Takagi, Yamamoto, S.(1980) : Factor analysis of cancer mortality rates for selected sites in the world, to appear in J.of National Institute of Environmental Health Perspectives.
- Yanai, H., Inaba, U., Takagi, H., et al. (1978) : Factor analytic study of Cancer mortalities in Japan from 1950 to 1967. Behaviormetrika, Vol 5, 55 - 74
- 柳井晴夫, 吉本泰彦, 高木廣文, 豊川裕之, 前田和甫, 栗田英男 (1977) : 胃がんのリスクファクターに関する統計的解析, 日本公衛誌, 24(9) : 547 - 556
- 柳井晴夫, 栗田英男 (1979) : 胃がん, 医学の歩み, 110 (13), 896 - 903
- 柳井晴夫, 高根芳雄 (1977) : 多変量解析法, 朝倉書店
- 山梨県企画管理室 (1962) : 山梨県総合地域指標
- 山本俊一他 (1974) : ベーチェット病患者全国疫学調査成績, 厚生省特定疾患疫学調査協議会他
- 山本俊一他 (1980) : 山梨県における肝がん・肝硬変発生要因に関する研究, 文部省科学研究費報告書 (印刷中)