

調査データのための分析方法

大学入試センター 研究開発部
大津 起夫

1 はじめに

ここでは、調査などによって得られるデータの分析法と、それに基づく因果性の推論方法について、特に統計的な分析を行う際の問題に重点を置いて紹介する。

入試研究に関わるデータの分析には、1) 試験成績(テストデータ)に関わるものと、2) 試験に関連する質問紙調査などによって得られるデータがある。前者に関わる分析手法は、一般的に「テスト理論」という名称で呼ばれており、洗練された手法が計量心理学の研究者によって開発されてきた。「因子分析」や「項目反応理論(項目応答理論)」は、各種の心理テストを分析する目的で開発されてきたものであり、分析のためのソフトウェアも良く整備されている。

ここでは、後者の一般的な調査データのために利用される分析方法について簡単に紹介し、特に統計データに基づいて因果性の推定を行う場合の問題を指摘する。

2 2×2表の関連性

離散的な値(合格・不合格、性別、現役・既卒など)を持つ変数間の関係を調べる基本的な方法は、これらの変数間についての分割表を作成することである。ただし、適切な層別を行わないと、不適切な見かけ上の関係が得られる場合がある。

表1は1973年におけるカリフォルニア大学バークレイ校の5つの学科での入学志願者数と合否の集計結果である。このデータは一見すると不可解な情報を与える例としてよく知られている。Bickel et al. (1975)によれば、1973年のバークレイ校での大学院の学科またはコースは、全体で101あり、志願者の総数は12,763名である。Bickelらは、入学許可に男女差別があるのではないかと疑問に答える形での分析を行い、単純な分析が誤った結論を導く可能性を指摘している。

ここに示したのは、この入学にかかわるデータのうち、Freedman et al.(1978)に示されている6つの学科についてのものである¹。表1に示した性別の集計結果を見ると、男子の合格率は44.5%であるのに対し、女子の合格率は30.4%であり、明らかに女子の合格率が低い。女子が大学院に合格しづらい傾向があるのだろうか？

¹このデータはフリーの統計分析ソフトウェアであるRシステム(Ihaka & Gentleman, 1996)の例題にも含まれている。

表 1: 1973 年におけるカリフォルニア大学バークレイ校大学院 6 学科への入学志願と可否の結果 (性別)

	実数			率 (%)		
	合格	不合格	合計	合格	不合格	合計
男子	1198	1493	2691	44.5	55.5	100.0
女子	557	1278	1835	30.4	69.6	100.0
合計	1755	2771	4526	38.8	61.2	100.0

Source: Freedman et al. (1978) による。これらは人数の多い 6 つの学科についてのもの。大学院全体の受験者数は 12,763 である (Bickel et al., 1975)。

以下では、まず 2×2 の分割表の関係を調べるための統計的な指標とそれらの性質について検討し、ついで学科別に層別された分割表の検討を行う。

2.1 2×2 分割表の関連性指標

2 つの変数 X_1 と X_2 とがそれぞれ 2 値 (0 と 1, または Yes と No など) の値をとるものとする。これらの関係を示すためにいくつかの方法が用いられている。

ふたつの 2 値変数 X_1 と X_2 のクロス集計されたデータの件数を 2×2 の行列 n_{ij} , ($i = 1, 2; j = 1, 2$) で表すものとする。行は X_1 に対応し、列が X_2 に対応する。また、 n_{i+} によって第 i 行の和を、 n_{+j} によって第 j 列の和を表す。標本の総件数は N とする。また、クロス表の相対頻度を p_{ij} , ($i = 1, 2; j = 1, 2$) で表し、その周辺相対頻度を p_{i+} および p_{+j} で表す。

より具体的には、 p_{1+} は $p_{11} + p_{12}$ であり、 $p_{2+} = p_{21} + p_{22}$ である。ここでの例では、それぞれ男子の志願者と女子の志願者の相対頻度である。また $p_{+1} = p_{11} + p_{21}$ であり、 $p_{+2} = p_{12} + p_{22}$ である。ここではそれぞれ合格者と不合格者の相対頻度となる。

	実数			相対頻度 (率)		
	$X_2 = 1$	$X_2 = 2$	行計	合格	不合格	行計
$X_1 = 1$	n_{11}	n_{12}	n_{1+}	p_{11}	p_{12}	p_{1+}
$X_1 = 2$	n_{21}	n_{22}	n_{2+}	p_{21}	p_{22}	p_{2+}
列計	n_{+1}	n_{+2}	$N = n_{++}$	p_{+1}	p_{+2}	1.0

ここで、行 X_1 と列 X_2 の間に統計的な関係がないことを、 X_1 と X_2 とが統計的に独立であるという。これが成立するのは

$$\frac{p_{11}}{p_{12}} = \frac{p_{21}}{p_{22}}$$

である場合である。表 1 の例においては、男子と女子で合格率が同じであることを意味する。

2×2 表の行と列との関係を示すための指標としては、次のものが用いられている。

1. オッズ比 (odds ratio)

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}},$$

合格率を「賭け率」(オッズ)とみなすと、上の値 α は二つの条件における賭け率の比となる。このためこの値 α はオッズ比と呼ばれる。オッズ比は行と列とが独立なら1である。最小はゼロであり、上限はない。

2. 対数オッズ比.

$$\log \alpha = \log \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

オッズ比の対数をとったもの。オッズ比自体よりも扱いやすい性質があるためしばしば利用される。行と列が独立なら(オッズ比1であり)ゼロとなる。

対数オッズ比の分散は各 n_{ij} が大きい場合に次の式で近似される。

$$\text{Var} \left[\log \frac{n_{11}n_{22}}{n_{12}n_{21}} \right] \simeq \frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}$$

3. 4 分点相関係数 (ファイ ϕ 係数). 2 値データが0または1値をとるものとみなして Spearman の積率相関係数 (一般的な相関係数) を求めると次の式が得られる。

$$\phi = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1+}p_{2+}p_{+1}p_{+2}}}$$

この値は通常相関係数と同じ手順で求められるので、便利ではあるが、後述するように周辺度数の偏りの影響を受けやすい。この値は-1から1の範囲をとる。

独立性の検定に用いられる X^2 統計量

$$X^2 = \sum_{i,j} \frac{(n_{ij} - Np_{i+}p_{+j})^2}{Np_{i+}p_{+j}}$$

は ϕ^2 の N 倍になる。

4. 4 分相関係数 (tetrachoric correlation).

分析対象となる 2×2 表の背後に、相関のある2変量正規分布を仮定し、2つの変数がある閾値によって分割されることにより、4つに区分されたデータが得られていると解釈する。図1は2変量正規分布に基づくデータの例であるが、ここで縦線と横線の位置および分布の相関を調節することにより、分布から期待される各領域の確率と得られたデータとが適合するようにする。この仮定にもとづいて、データにもっとも良く当てはまる2変量正規分布を推定し、その相関係数を求める。このようにして得られる値を4分相関係数と呼ぶ。この値は簡単な式で求めることができないので、近似式による推定か反復を伴う数値計算が必要になる。

これらの指標にはそれぞれの特徴がある。4分点相関係数は周辺度数の偏りによって大きく影響を受けるため、相対頻度が小さいカテゴリーがある場合には適切ではない。

ここで、表1について、これらの指標の値を求めてみると次のようになる。

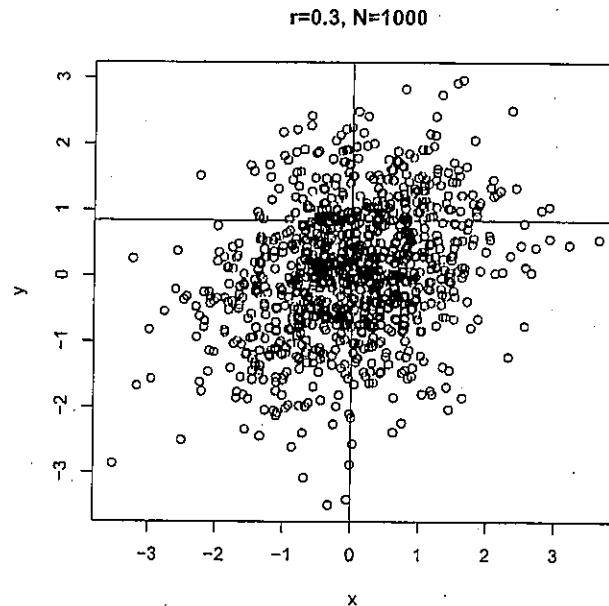


図 1: 2 変量正規分布に基づくデータ例 (母相関は 0.3)

1. オッズ比

$$\frac{1198 \times 1278}{557 \times 1493} = 1.84$$

値の範囲はゼロから無限大まで。性別と合格率が独立なら 1。

2. 対数オッズ比

$$\log \frac{1198 \times 1278}{557 \times 1493} = 0.610$$

求められた値の標準偏差は約 0.064 と推定される。値の範囲は、負の無限大から正の無限大までどの値も取りうる。性別と合格率が独立ならゼロ。

3. 4 分点相関係数 (ファイ係数 ϕ)

$$\phi = \frac{1198 \times 1278 - 557 \times 1493}{\sqrt{(1198 + 557)(1493 + 1278)(1198 + 1493)(557 + 1278)}} = 0.1427$$

-1 から +1 の範囲の値をとる。性別と合格率が独立ならゼロ。この例では、より正確には $\phi = 0.1427318$ 。

ファイ係数から導かれるカイ 2 乗統計量 X^2 は必ずゼロ以上の値をとる。この場合

$$X^2 = \phi^2 \times N = 0.1427^2 \times 4526 = 92.21$$

表 2: 合格者を4分の1にした仮想例

	実数			率 (%)		
	合格	不合格	合計	合格	不合格	合計
男子	299	1493	1792	16.7	83.3	100.0
女子	139	1278	1417	9.8	90.2	100.0
合計	438	2771	3209	0.136	0.864	100.0

である。ゼロであるのは、性別と合格率とが独立である場合に限られる。このデータについて独立性のカイ2乗検定を行うと、上側確率 (p 値) は極めてゼロに近い値となり、性別と合格率が独立ではないことが確認される。

4. 4分相関係数

専用のソフトウェアを用いた反復計算により推定すると、次の値が得られる。

$$r_{tetra} = 0.230$$

2.2 関連性指標の特徴

これらの指標にはそれぞれ個性があり、利用にあたって注意を払う必要がある。特に利用される機会の多い指標は (対数) オッズ比とファイ係数であると思われるが、これらは相反する性格を持っている。

ここで表1のデータに変更を加え合格者数を (ほぼ) 4分の1に変更してみる。

表2について、指標を求めると次のようになる。

表 3: 合格者が4分の1の場合の関連性の指標 (仮想例)

指標	変更後	変更前
1 オッズ比	1.84	1.84
2 対数オッズ比	0.610	0.610
3 4分点相関係数 (ファイ係数)	0.099	0.1427
4 4分相関係数	0.200	0.230

このように、オッズ比は、ある行 (または列) が一定の比率で減少 (増加) しても変化しないが、他の指標は変化する。特定の列 (または行) の比率が小さい場合 (例えば合格率が極端に小さい場合) には、ファイ係数 ϕ は小さい値になる傾向がある。このため、周辺度数の異なる 2×2 表についてファイ係数を求めると、それらの大小が関連性の強さではなく、周辺度数の偏りの程度を表す場合がある。

表4は、2変量正規分布を各変数の閾値で区分し、 $(p_{ij}, i = 1, 2; j = 1, 2)$ を求め、4分点相関係数 ϕ とオッズ比 α を計算した結果である。同一の p について、 ϕ と α が逆の傾向を示している。 ϕ は、周辺度数が偏ると小さくなる傾向があるが、逆に α は大きな値を示している。

表 4: 相関係数とオッズ比

母相関 ρ	周辺確率 (p_{1+}, p_{+1})	同時確率 (p_{11})	4 分点相関係数 ϕ	オッズ比 α	対数オッズ比 $\log \alpha$
0.3	(0.05, 0.95)	0.0495	0.043	5.94	1.78
	(0.20, 0.80)	0.1809	0.131	2.78	1.02
	(0.20, 0.95)	0.1968	0.078	3.81	1.34
	(0.50, 0.50)	0.2985	0.194	2.19	0.79
	(0.50, 0.80)	0.4336	0.168	2.38	0.87
	(0.50, 0.95)	0.4870	0.110	3.00	1.10
	(0.80, 0.80)	0.6661	0.163	2.46	0.90
	(0.80, 0.95)	0.7704	0.119	2.99	1.08
	(0.95, 0.95)	0.9071	0.098	3.52	1.26
0.9	(0.05, 0.95)	0.0500	0.053	極めて大	—
	(0.20, 0.80)	0.2000	0.250	18398.0	9.82
	(0.20, 0.95)	0.2000	0.115	極めて大	—
	(0.50, 0.50)	0.4282	0.713	35.59	3.57
	(0.50, 0.80)	0.4981	0.490	169.97	5.14
	(0.50, 0.95)	0.5000	0.229	極めて大	—
	(0.80, 0.80)	0.7499	0.687	44.85	3.80
	(0.80, 0.95)	0.7988	0.445	219.70	5.39
	(0.95, 0.95)	0.9319	0.618	90.32	4.50

2.3 シンプソンのパラドックスと層別の分析

ここで表 1 のデータをより詳細に学科別に検討してみる。表 5 は、学科別の内訳を示したものである。各学科別に、オッズ比を求めると 6 学科のうち、女子の合格率が低いのは、C と E の 2 つのみであることがわかる。また、独立性のカイ 2 乗検定を行うと、顕著な関係がみられるのは A 学科のみであり、しかも傾向は女子の合格率の方が高い。これは全体を集計した傾向とは、逆向きのものである。

では、なぜこのような相反する傾向が生じたのだろうか。ここで、学科ごとに女子の志願者の比率と男女合わせての合格率の散布図を図 2 に示す。これを検討すると、女子の志願者の多い学科は、合格率の低いつまり入学の難しい学科であることがわかる。このため、個別の学科においては、女子の合格率が低い傾向がなくとも、全体では女子の合格率が男子よりも低くなってしまう。

多重分割表の分析において、全体での傾向と層別での傾向とが相反するこのような逆説的な現象は、シンプソンのパラドックス (Simpson's paradox) と呼ばれている。

ここでデータの構造を直観的に理解するために、表 5 の内容をグラフに表示したものが、図 3 である。この表示方法は Bertin(1981) によるもので、重み付きマトリックスと呼ばれている。この場合は、各列の横方向の幅が各学科別・性別の志願者数に比例するようにとってある。長方形の高さは、合格率と不合格率を表している。横方向の幅がそれぞれのケースの層別の人数に比例しているので、長方形の面積が該当するセルの人数を表す。また、破線は全体での合格率と不合格率を表しており、これを超える部分は表示色を変えて強調してある。カイ 2 乗検定の結果からもわかるように、A 学科を除いては男女で大きな合格率の違いはない。また、A 学科では女子の方が際立って合格率の高いことがわかる。

表 5: 学科別・性別の可否の結果

性別	可否	学科						合計
		A	B	C	D	E	F	
男子	合格	512	353	120	138	53	22	1198
	不合格	313	207	205	279	138	351	1493
女子	合格	89	17	202	131	94	24	557
	不合格	19	8	391	244	299	317	1278
合計		933	585	918	792	584	714	4526
オッズ比		0.35	0.80	1.13	0.92	1.22	0.83	1.84
ϕ		-0.14	-0.02	0.03	-0.02	0.04	-0.02	0.61
X^2		17.25	0.25	0.75	0.30	1.00	0.38	92.20
(p 値)		0.00	0.61	0.39	0.59	0.32	0.54	0.00

Source: Freedman et al. (1978). オッズ比が1より大きい (ϕ がゼロより大きい) ことは, 男子の合格率が女子より大きいことを示す.

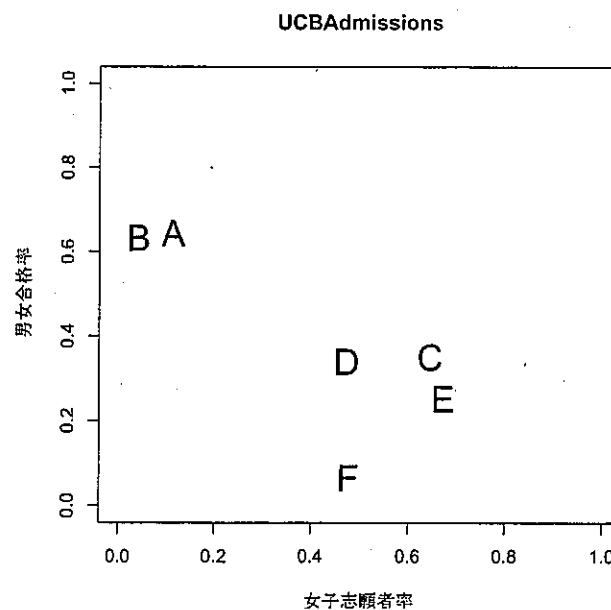


図 2: 学科別の女子志願者率と合格率 (男女込み)

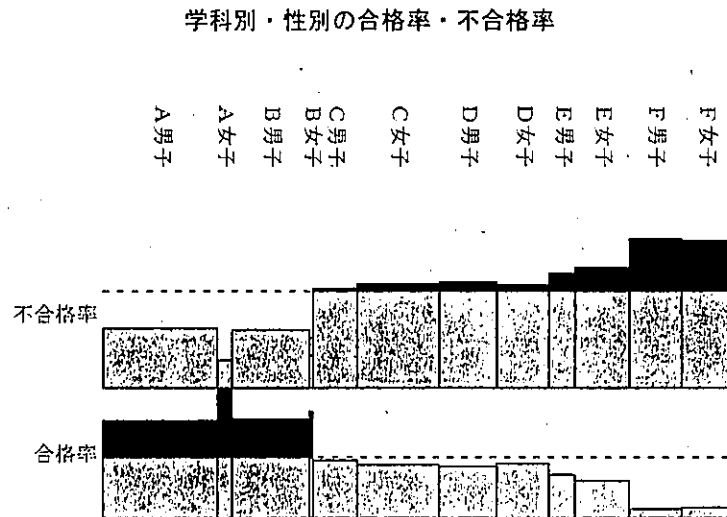


図 3: 学科別性別合格率不合格率の重み付きマトリックス表示

3 調査データにおける因果推定の難しさ

多重分割表の分析を行って因果関係の推論を行う場合には注意すべき点がいくつかある。

3.1 多くの層に対応する推定法

ひとつは、推定上の（数値的な）性質に関わるものである。

ここでは説明しなかったが、多重分割表の構造を分析する手法として最も一般的なものはロジスティック回帰および対数線形モデルと呼ばれる手法である。これらのモデルの推定にあたっては、通常、最尤法という手法を用いている。最尤法はデータの件数が十分あるなら、ある意味において最善の推定方法であること、つまり最も推定精度の良いことが理論的に保証されている。しかしながら、各層に含まれるデータの件数が多くない場合（それぞれ数件などの場合）には、必ずしも適切な推定値を与えるとは限らない。特に、ケースコントロールスタディとよばれる研究デザイン（各層には実験群に属する被験者が1名、対照群に属する被験者が1名の計2名しか含まれない）では、最尤法による推定では著しい偏りが生じる。このようなケースについては、条件付最尤法とよばれる推定手法を用いるか、またはコ克蘭・マンテル・ヘンツェル検定などの手法を用いる必要がある。

3.2 影響関係の方向

もうひとつの問題は、より概念的な枠組みに関わるものである。上に述べた例では、性別と合格率の双方に学科が関係しているため、これについて層別することにより、妥当な検討を行うことができた。では、一般的に、変数 X と変数 Y の関係を調べる場合、それら両者に関係している全ての変数 $Z_1, \dots, Z_p$ を列挙し、これらについて同時に層別を行えばよいだろうか。

これについては、影響関係の方向までも含めてより慎重に検討する必要がある。図4は、影響関係についてのいくつかのケースを示したものである。ここで取り上げた例は(i)に示したものである。多重分割表による分析は影響関係の強さを記述することを目的としており、影響関係の向きについての情報を用いてはいない。しかし、現実の解釈においては、影響関係の向きが重要になる。上に示した合格率の例(i)では、「学科」について層別することにより、「性別」(X)の「合否」(Y)に及ぼす直接の影響を検討することが可能になる。また(ii)の例のように X と Y とに Z が共通に影響を与えている場合も、 Z について層別を行うことにより、 X から Y への直接の影響を推定することができる。

しかし、常に層別を行うことが妥当であるわけではない。(iii)の例のように影響関係の下流にある変数について層別を行うと、知ろうとしている関係とは異なる見かけ上の関係が X と Y に生じるため、分析方法として適切ではない。

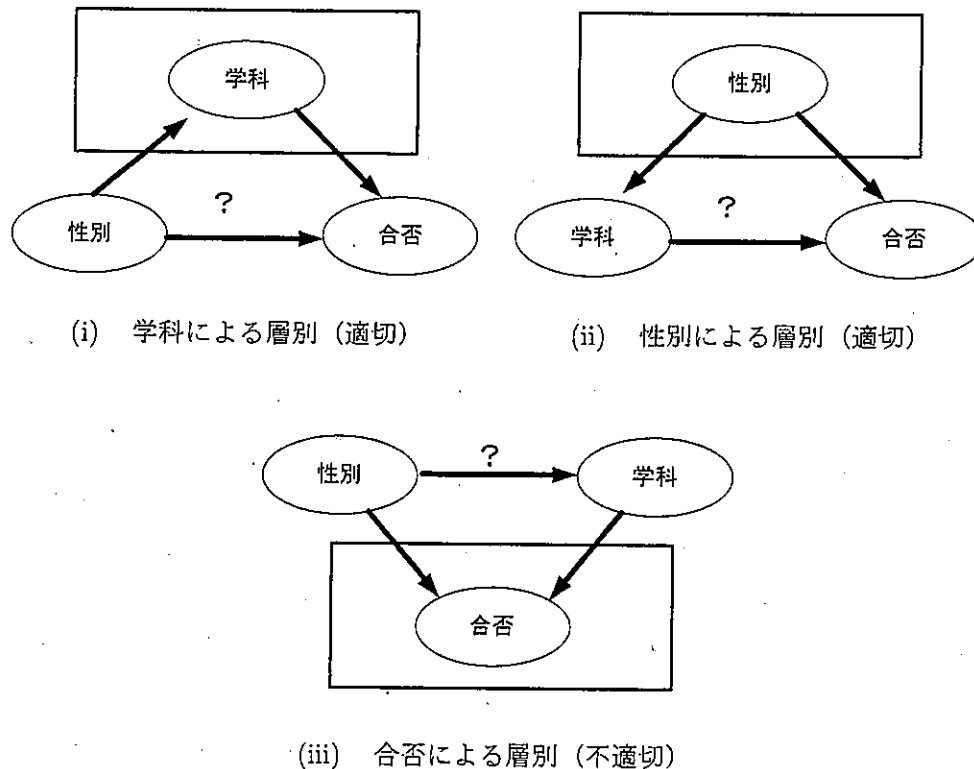


図 4: 影響構造と層別の適切さ

3.3 未知の背景要因の影響

また外的な介入の有無は、影響関係の推論の確さと深い関係を持つ。一般的には、外的介入なしでの推論から決定的な結論を得るのは難しい。以下は、介入を伴わない（大量の）調査データから得られた結論が、他の証拠なしには、明快な解釈が困難であることを示している。

Smith et al. (1992) は、入学試験とは異なる分野の例ではあるが、調査データの結果の解釈に、慎重な注意が必要であることを指摘した論文である。ここで著者らは、大規模な危険要因についての調査 (MRFIT) を分析することによって、まず喫煙と自殺の間に関係が見られることを指摘している。MRFIT (Multiple Risk Factor Intervention Trial) は米国の 35 歳から 57 歳までの 361,662 人の男性を 12 年間にわたって追跡調査した研究である。年齢、人種、1 日の喫煙本数、糖尿病治療中であるか、心筋梗塞の有無、社会経済地位（居住地域から推定）がベースラインデータとしてとられており、これらの効果を考慮してもなお喫煙と自殺との間に関係は見られる。また、MRFIT 以外の他の大規模な複数の追跡調査でも同様の結果が得られている。

しかし、Smith et al. (1992) は、MRFIT について喫煙と他殺による死の間にも同様の関係が見られることを報告している。喫煙がその薬理作用によって自殺を増加させるという説明は、一見もっともらしく思えるが、他殺を増加させる原因となるとは考えづらい。喫煙は本当に自殺に影響を及ぼしているのだろうか？

これについての決定的な答えは現状では分かっていない。しかし、タバコが直接自殺の危険を高めるとこれらの調査結果から解釈するのは、困難であるように思える。研究者によっては捉えられていない何らかの心理的または社会的要因があり、これが喫煙の習慣と自殺傾向の両者に影響を及ぼしているとみなすことも可能と思われる。

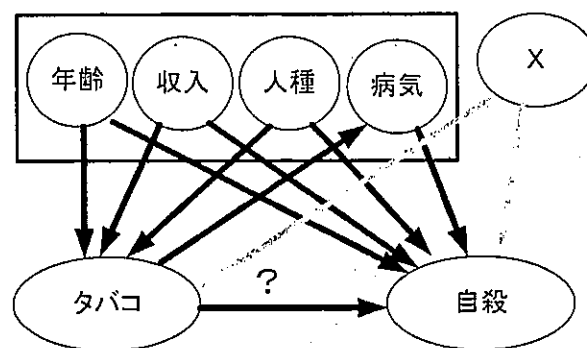


図 5: 喫煙の影響構造

表 6: 喫煙行動と自殺率 (Smith et al., 1992)

タバコ/日	人数	自殺件数	自殺件数 1万人×年
0	228,545	291	1.09
1-19	29,333	50	1.47
20-39	72,200	166	2.00
40-59	27,844	78	2.46
60+	3,740	16	3.78

表 7: リスク要因と自殺の相対リスク (95% 信頼区間) (Smith et al. 1992)

リスク要因	年齢補正済み 相対リスク	完全補正済み 相対リスク
喫煙		
0	1.00	1.00
1-19	1.36(1.00-1.83)	1.36(1.00-1.84)
20-39	1.88 (1.55-2.27)	1.86(1.54-2.26)
40-59	2.31 (1.79-2.96)	2.27(1.77-2.92)
60+	3.44 (2.08-5.69)	3.33(2.01-5.52)
(χ^2 for trend)	(78.66, $p < 0.0001$)	(75.98, $p < 0.0001$)
収入		
低	1.35(1.13-1.60)	1.33(1.13-1.59)
他	1.00	1.00
人種		
黒人	0.85(0.60-1.21)	0.76(0.53-1.08)
他	1.00	1.00
心筋梗塞		
病歴あり	1.79(1.07-2.99)	1.73(1.03-2.90)
なし	1.00	1.00
糖尿病		
有病	1.66(0.99-2.77)	1.61(0.96-2.70)
なし	1.00	1.00

表 8: タバコと他殺による死 (Smith et al. 1992 に基づく)
 総数 222, 人種と収入 (居住地域から推定) について補正された
 10 万人あたり死亡率の比

リスク要因 (一日あたり喫煙本数)	他殺による死亡の相対比率 (95% 信頼区間)
0	1.00
1-39	1.71 (1.29-2.28)
40-	2.04 (1.32-3.15)

参考文献

- [1] Agresti, A. (2003) カテゴリカルデータ解析入門, サイエンス社 (原著 1996, *An introduction to categorical data analysis*, Wiley).
- [2] Agresti, A. (2002). *Categorical data analysis*, Wiley.
- [3] Bertin, J. (1981). *Graphics and graphic information-processing*, (translated by W.J.Berg and P.Scott) New York: Walter de Gruyter.
- [4] Bickel, P.J., Hammel, E.A., & O'Connell, J.W. (1975). Sex bias in graduate admissions: Data from Berkeley, *Science*, 187, 398-404.
- [5] Freedman, D., Pisani, R. & Purves, R. (1978). *Statistics*, New York: W.W. Norton.
- [6] Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5, 299-314. (<http://www.r-project.org/>)
- [7] Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.
- [8] Smith, G.D., Phillips, A.N. & Neaton, J.D. (1992). Smoking as "independent" risk factor for suicide: illustration of an artifact from observational epidemiology? *Lancet*, 340, 709-12.