第7章

カテゴリ変数2つの分析（2）

7.1 研究デザインとリスク，オッズ

前章でも触れたように，関連性の指標は研究デザインによって異なる。まず，関連性の指標の中でもやや毛色が異なる（統計学よりも疫学でよく使われる）リスクとオッズという考え方を説明する。疫学分野で主に発達した理論なので，病気を例にとって説明するが，因果関係が想定できる変数間であれば，別に病気の話に限らず成立する考え方である。

病気のリスクといえば，全体のうちでその病気を発症する人の割合である。これとは別に，オッズという考え方もある。病気のオッズといえば，その病気を発症した人数の，発症しなかった人数に対する比である。

さてしかし，リスクとかオッズそのものでは，病気の発症と要因の有無の関係はわからない。要因があった場合のリスクやオッズを，要因がなかった場合のリスクやオッズと比べることによって，初めて要因の有無と病気の発症がどれくらい関係していたかがわかる。すなわち，ある要因をもつ人たち（曝露群）の病気のオッズが，その要因がない人たち（対照群とかコントロール群という[*1]）の病気のオッズに対して何倍になっているか，というのがオッズ比（英語では Odds Ratio）である。同じように，曝露群のリスクの，対照群のリスクに対する比がリスク比[*2]である。

要因の有無と病気の有無がまったく関係がなければ，リスク比もオッズ比も1になることが期待される。それぞれ信頼区間を計算して（計算方法は難しいので後述），

[*1] 理想的な対照群は，その要因がない点だけが曝露群と違っていて，それ以外の条件はすべて同じであることが望ましい。

[*2] 英語では Risk Ratio だが，Rate Ratio とか Relative Risk という言い方もある。Relative Risk の訳から相対危険ということもあるが，同じ意味。

たとえば95%信頼区間が1を含まなければ，5%水準で有意な関連が見られるといえる。

ところで，病気のリスクは，全体のうちで病気を発症する人の割合であったから，まず全体を把握していないと定義できない。つまり，まず観察対象全体で曝露群と対照群を把握しておいて，経時的に追跡調査して，それぞれの群で何人ずつ発症するかを調べるという，「前向き研究」（コホート研究とかフォローアップ研究ということもある）でないと，リスク比は計算できないことになる。

これに対して，患者対照研究（Case Control Study）*3とか断面研究（Cross Sectional Study）*4では，曝露時点での全体が未知なので，原理的にリスクを計算できないことになる。激しい曝露を受けた人は調査時点よりずっと前に病気を発症して死んでしまった可能性があるので，患者対照研究や断面研究から無理にリスクを見積もろうとするとリスクを過小評価してしまうことになるからである。

一方，オッズ比はどんなデザインの研究でも計算できる。たんに，曝露群の病気の人数の病気でない人数に対する比が，対照群のそれに比べてどれくらい大きいかを示す値だからである*5。

ここで，クロス集計表ではどう計算するのかということを示す。以下の表を考えてみる。

	疾病あり	疾病なし	合計
曝露あり	a	b	m_1
曝露なし	c	d	m_2
合計	n_1	n_2	N

この表でいえば，リスク比は $(a/m_1)/(c/m_2)$ となり，疾病オッズ比は $(a/b)/(c/d) = ad/bc$ である。曝露オッズ比は $(a/c)/(b/d) = ad/bc$ となるので，疾病オッズ比と一致することがわかる*6。

*3 調査時点で，患者を何人サンプリングすると決め，それと同じ人数の対照（その病気でないことだけが患者と違って，それ以外の条件はすべて患者と同じことが望ましい）を選んで，それぞれが過去に受けた曝露要因や，現在の生活習慣，態度などを調べることによって，その病気の原因を探る方法論。

*4 調べてみないと患者かどうかさえわからないような場合や，因果の向きがはっきりしない変数間の関係を見たいときは，全体で何人サンプリングすると決めて一時点で調査する。こういう方法論を断面研究という。

*5 この場合のオッズ比は，「曝露なし群での疾病ありのオッズ」に対する「曝露あり群での疾病ありのオッズ」の比なので，**疾病オッズ比**という。逆に，疾病あり群で曝露した人数の曝露していない人数に対する比が，疾病なし群のそれに比べてどれくらい大きいかを示す値として**曝露オッズ比**というものも考えられるが，数学的には同じ値になる。

*6 ただし，統計パッケージでは，単純なこの値でなく，最尤推定をして得られる条件付きオッズ比が表示されることが多い。

オッズ比が重要なのは，稀な現象をみるときには，リスク比のよい近似になるからであると言われている。たとえば，送電線からの高周波が白血病の原因になるという仮説を検証するために，送電線からの距離が近い場所に住んでいる人（曝露群）と，遠いところに住んでいる人（対照群）をサンプリングして，5年間の追跡調査をして，5年間の白血病の罹患率を調査することを考えよう。白血病は稀な疾患だし，高周波に曝露しなくても発症することはあるので，このデザインでリスク比を計算するためには，莫大な数のサンプルをフォローアップする必要があり，大規模な予算とマンパワーが投入される必要があるだろう。

仮に[*7]調査結果が，下表のようであったとすると，

	白血病発症	発症せず	合計
送電線近くに居住	4	9996	10000
送電線から離れて居住	2	9998	10000
合計	6	19994	20000

送電線の近くに住むことで白血病を発症するリスクは，送電線から離れて住む場合の2倍になった（$(4/10000)/(2/10000) = 2$，つまりリスク比が2なので）ということができる。ここでオッズ比をみると，$(4*9998)/(2*9996) \simeq 2.0004$ と，ほぼリスク比と一致していることがわかる[*8]。

こうして得られるリスク比は，確かに原理的に正しくリスクを評価するのだが，稀なリスクの評価のためには大規模な調査が必要になるので，効率が良いとはいえない。そこで，通常は，前向き研究ではなく，患者対照研究を行って，過去の曝露との関係をみることが行われる。この場合だったら，白血病患者100人と対照100人に対して，過去に送電線の近くに居住していたかどうかを聞くわけである。それで得られた結果が，仮に下表のようになったとしよう[*9]。

	白血病	白血病でない	合計
送電線近くに居住した経験あり	20	10	30
送電線から離れて居住	80	90	170
合計	100	100	200

この場合，リスク比は計算しても意味がない（白血病かつ送電線の近くに居住した経験がある20人は，送電線の近くに住んだ経験がある人からのサンプルではなく，

[*7] これはあくまで架空のデータである。本当の送電線と白血病の関係は，数年前から，WHOのプロジェクトの一環として，国立環境研究所と国立癌センターの研究チームが調べたらしいが，その結果がどうなったのかは知らない。

[*8] 上述のように最尤推定された条件付きオッズ比は，Rのプログラムを使って `fisher.test(matrix(c(4,2,9996,9998), nc=2))` として計算すると，2.000322である。

[*9] くどいようだが，あくまで架空のデータである。

白血病患者からのサンプルだから）が，送電線の近くに居住した経験がある人のうち，白血病の人の，白血病でない人に対するオッズは2となり，送電線から離れて居住した人ではそのオッズが0.888...となるので，これらのオッズの比は2.25となる。この値は母集団におけるリスク比のよい近似になることが知られている。このように稀な疾患の場合は，患者対照研究でオッズ比を求める方が効率が良い。

原理的に前向き調査ができない場合もある。とくに，薬害と呼ばれる現象は，妙な病気が見つかったときに，後付けで原因を探ることになるので，患者対照研究にならざるを得ない。たとえば，スモンとかサリドマイドは，そうやって原因がわかった問題である。腕が短く生まれた子どもの母親と，そうでない子どもの母親に，妊娠中に飲んだ薬の有無を尋ねて，特定の時期にサリドマイドを飲んだという曝露による疾病オッズ比が有意に大きい結果が得られたのだ*10。

また，問題があるかどうかが事前に明らかでない場合は，断面研究をせざるを得ない。聞き取りや質問紙などで調べる，心理学的，あるいは社会学的な調査項目間の関係を見る場合は，断面研究をする場合が多い。なお，断面研究の場合は，リスク比やオッズ比の他に，リスク差，相対差，曝露寄与率，母集団寄与率，YuleのQ，ファイ係数といったものがある（後述）*11。

なお，同じ質問を2回した場合に同じ変数がどれくらい一致するかについては，普通にクロス集計表を作って独立性の検定ができそうな気がするかもしれないが，してはいけない。この場合はtest-retest-reliabilityを測ることになるので，クロンバックのα係数やκ統計量などの一致度の指標を計算するべきである（後述）。

では，リスク比とオッズ比の95%信頼区間を考えよう。まずリスク比の場合から考えると，前向き研究でないとリスク比は計算できないので，曝露あり群となし群をそれぞれm_1人，m_2人フォローアップして，曝露あり群でX人，なし群でY人が病気を発症したとしよう。得られる表は，

	発症	発症なし	合計
曝露あり	X	$m_1 - X$	m_1
曝露なし	Y	$m_2 - Y$	m_2
合計	$X+Y$	$N-X-Y$	N

となる。このとき，母集団でのリスクの推定値は，曝露があったとき$\pi_1 = X/m_1$，曝露がなかったとき$\pi_2 = Y/m_2$である。リスク比は，$RR = \pi_1/\pi_2$なので，その推定量は，$(Xm_2)/(Ym_1)$となる。

*10 ここで有意と書いたが，統計的に有意かどうかをいうためには検定するか，95%信頼区間を出さねばならない。その方法は後述する。

*11 2×2でないクロス集計表で，たとえば5×5以上ならば，順位相関係数を使うことも可能。

リスク比の分布は N が大きくなれば正規分布に近づくので，正規分布を当てはめて信頼区間を求めることができるが，普通は右裾を引いているので対数変換か立方根変換（Bailey の方法）をしなくてはならない。対数変換の場合，95% 信頼区間の下限と上限はそれぞれ，

$$RR \cdot \exp(-\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad \text{（下限）} \tag{7.1}$$

$$RR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad \text{（上限）} \tag{7.2}$$

となる。RR が大きい場合は立方根変換しなくてはいけないが，煩雑なので省略する。前述の白血病の例で計算してみると，95% 信頼区間は，$(0.37, 10.9)$ となる。

次にオッズ比の信頼区間を考える。前述の表の a, b, c, d という記号を使うと，オッズ比の点推定値 OR は，$OR = (ad)/(bc)$ である。オッズ比の分布も右裾を引いているので，対数変換または Cornfield (1956) の方法によって正規分布に近づけ，正規近似を使って 95% 信頼区間を求めることになる。対数変換の場合，95% 信頼区間の下限は $OR \cdot \exp(-\mathrm{qnorm}(0.975)\sqrt{1/a + 1/b + 1/c + 1/d})$，上限は $OR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/a + 1/b + 1/c + 1/d})$ となる。前述の白血病の例で計算してみると，オッズ比の 95% 信頼区間も $(0.37, 10.9)$ となる[*12]。Cornfield の方法はやや複雑であり，高次方程式の解を Newton 法などで数値的に求める必要があるので，本書では扱わない。

7.2 その他の関連性の指標

7.2.1 リスク差

曝露によるリスクの増減を絶対的な変化の大きさで表した値。過剰危険 (Excess Risk) ともいう。

$$RD = \pi_1 - \pi_2$$

7.2.2 相対差

要因ももたず発症もしていない者のうち，要因をもった場合にのみ発症する割合。

$$RelD = (\pi_1 - \pi_2)/(1 - \pi_2)$$

[*12] R の `fisher.test()` 関数で計算した結果では，オッズ比の 95% 信頼区間は $(0.29, 22.1)$ となり，対数変換を使った単純な計算よりも幅が広くなる。

7.2.3 曝露寄与率

真に要因の影響によって発症した者の割合。

$$AFe = (\pi_1 - \pi_2)/\pi_1$$

7.2.4 母集団寄与率

母集団において真に要因の影響によって発症した者の割合。$\pi = (X + Y)/(m_1 + m_2)$ として，

$$AFp = (\pi - \pi_2)/\pi$$

7.2.5 Yule の Q

オッズ比を -1 から 1 の値を取るようにスケーリングしたもの。

$$Q = (OR - 1)/(OR + 1)$$

7.2.6 ファイ係数 (ρ)

要因の有無，発症の有無を 1,0 で表した場合の相関係数[*13]。θ_1, θ_2 を発症者中の要因あり割合，非発症者中の要因あり割合として，

$$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$$

7.3 一致度の指標

7.3.1 κ 統計量

2 回の繰り返し調査をしたときに，あるカテゴリ変数がどれくらい一致するかを示す指標である。

[*13] 相関係数については，第 11 章で詳しく説明するが，-1 から 1 までの値をとる量で，2 つの変数間にまったく関連がない場合に 0 となり，片方が大きくなればもう片方の変数も常に同じ割合で大きくなる関係があるとき 1 となる。

	2回目○	2回目×	合計
1回目○	a	b	m_1
1回目×	c	d	m_2
合計	n_1	n_2	N

という表から，偶然でもこれくらいは一致するだろうと思われる値は，1回目と2回目の間に関連がない場合の各セルの期待値を足して全数で割った値になるので $P_e = (n_1 \cdot m_1/N + n_2 \cdot m_2/N)/N$，実際の一致割合（1回目も2回目も○か，1回目も2回目も×であった割合）は $P_o = (a+d)/N$ とわかる。ここで，$\kappa = (P_o - P_e)/(1-P_e)$ と定義すると，κ は，完全一致のとき1，偶然と同じとき0，それ以下で負となる統計量となる。

κ 統計量は，有意性の検定ができる。κ の分散 $V(\kappa) = P_e/(N \cdot (1-P_e))$ となるので，$\kappa/\sqrt{V(\kappa)}$ が標準正規分布に従うことを利用して検定できる。つまり，帰無仮説「κ が偶然一致する程度と差がない」が正しい確率が $1 - \mathrm{pnorm}(\kappa/\sqrt{V(\kappa)})$ となる。

ここで pnorm() は正規分布の分布関数を表す R の関数である。上の表の記号を使って R のプログラムを書けば，

```
Pe<-(n1*m1/N+n2/N*m2)/N
Po<-(a+d)/N
kappa<-(Po-Pe)/(1-Pe)
SEkappazero<-sqrt(Pe/(N*(1-Pe)))
pkappa<-1-pnorm(kappa/SEkappazero)
cat("Kappa=",kappa," (p=",pkappa,")\n")
```

となる。

この確率が5%未満ならば，得られた一致度は有意水準5%で信頼できる（偶然の一致より大きい）といえる。

κ 統計量の95%信頼区間は，

$$\kappa \pm \mathrm{qnorm}(0.975) \cdot \sqrt{P_o \cdot (1-P_o)/(N \cdot (1-P_e)^2)}$$

として計算できる。

有意性の検定の場合と同様に，上の表の記号を使って R のプログラムを書けば，

```
Pe<-(n1*m1/N+n2/N*m2)/N
Po<-(a+d)/N
kappa<-(Po-Pe)/(1-Pe)
SEkappa<-sqrt(Po*(1-Po)/(N*(1-Pe)^2))
kappaL<-kappa-qnorm(0.975)*SEkappa
kappaU<-kappa+qnorm(0.975)*SEkappa
cat(95%CI=[",kappaL,",",kappaU,"]\n")
```

となる。

なおκ統計量は，2×2だけでなく，m×mのクロス集計表に適用できる概念である。

7.4 利用例

本章で紹介した全ての指標を計算する関数 crosstab() を定義し，R の組み込みデータであるスイス女性の出生データに適用する例を挙げておくので，参考にされたい。

```
#
# Defining a function to combine several calculation for the indices of
# relationship.
# developed by Minato Nakazawa on 16th November 2001.
crosstab <- function(X) {
if (length(X)>4) stop("Given data cannot constitute 2x2 cross table")
cat(rep("=",35),"\n The results may include inappropriate statistics
for given table\n
(e.g. Risk Ratio can stand only for cohort study).
Take care.\n",rep("=",35),"\n")
a<-X[1,1]; b<-X[1,2]; c<-X[2,1]; d<-X[2,2]
m1<-a+b; m2<-c+d; N<-m1+m2; n1<-a+c; n2<-b+d
# risk ratio
RR<-(a*m2)/(c*m1)
RRL<-RR*exp(-qnorm(0.975)*sqrt(1/a-1/m1+1/c-1/m2))
RRU<-RR*exp(qnorm(0.975)*sqrt(1/a-1/m1+1/c-1/m2))
cat("Risk Ratio=",RR,"\t 95%CI=[",RRL,",",RRU,"]\n")
# odds ratio
OR<-(a*d)/(b*c)
ORL<-OR*exp(-qnorm(0.975)*sqrt(1/a+1/b+1/c+1/d))
ORU<-OR*exp(qnorm(0.975)*sqrt(1/a+1/b+1/c+1/d))
cat("Odds Ratio=",OR,"\t 95%CI=[",ORL,",",ORU,"]\n")
```

```
# risk difference
cat("Risk Difference=",RD<-a/m1-c/m2,"\n")
# relative difference
cat("Relative Difference=",RelD<-(a/m1-c/m2)/(1-c/m2),"\n")
# "曝露寄与率"
cat("AFe=",AFe<-(a/m1-c/m2)/(a/m1),"\n")
# "母集団寄与率"
cat("AFp=",AFp<-(n1/N-c/m2)/(n1/N),"\n")
# Yule's Q
cat("Yule's Q=",Q<-(OR-1)/(OR+1),"\n")
# "ファイ係数"
cat("phi coefficient=",rho<-sqrt((a/m1-c/m2)*(a/n1-b/m2)),"\n")
# kappa
Pe<-(n1*m1/N+n2/N*m2)/N
Po<-(a+d)/N
kappa<-(Po-Pe)/(1-Pe)
SEkappa<-sqrt(Po*(1-Po)/(N*(1-Pe)^2))
kappaL<-kappa-qnorm(0.975)*SEkappa
kappaU<-kappa+qnorm(0.975)*SEkappa
SEkappazero<-sqrt(Pe/(N*(1-Pe)))
pkappa<-1-pnorm(kappa/SEkappazero)
cat("Kappa=",kappa," (p=",pkappa,")\t 95%CI=[",kappaL,",",kappaU,"]\n")
}

data(infert)
fewchild<-(infert$parity<=2)
noabort<-(infert$spontaneous==0)
Y<-table(fewchild,noabort)
print(Y)

# output of table(fewchild,noabort)
#
#         noabort
# fewchild FALSE TRUE
#    FALSE    43   25
#    TRUE     64  116
#
# i.e.
#           abort
# morechild  TRUE FALSE
#     TRUE     43    25
#    FALSE     64   116
```

```
# Chi-square test
print(chisq.test(Y))
# Fisher's exact test (where odds ratio is conditional MLE)
print(fisher.test(Y))

crosstab(Y)
# same as crosstab(matrix(c(43,64,25,116),nc=2))
```