医学情報処理演習第 10 回「クロス集計」

中澤　港 (nminato@med.gunma-u.ac.jp)

2004 年 12 月 13 日

前回の課題の回答例

下枠内のように，村別のマラリア原虫陽性者人数を変数 malaria に，検査総数を変数 pop に付値して，まず，「マラリア原虫陽性割合には村落間に差がない」という帰無仮説を検定する。

```
malaria <- c(6,10,18)
pop <- c(180,220,80)
names(malaria) <- c("内陸","川沿い","海沿い")
prop.test(malaria,pop)
```

有意確率は 10^{-8} のオーダーなので帰無仮説は有意水準 5%で棄却される。そこで，次に多重比較により，どの村落とどの村落で差があるのかを検討する。そのため，`pairwise.prop.test(malaria,pop)` と入力する。

```
        内陸     川沿い
川沿い  0.72     -
海沿い  8.0e-06  1.3e-05
```

ホルムの方法で第 1 種の過誤を調整した有意確率が上枠内のように得られるので，有意水準 5%で検定すると，内陸と川沿いにはマラリア原虫陽性割合に有意な差がなく，海沿いと内陸，海沿いと川沿いにはそれぞれ有意差があると判断される。

ハマダラカの相対的な密度のスコアを `mosquito <- c(1,2,4)` と与え，この順に原虫陽性割合が大きくなっていく傾向があるかどうかをコクラン＝アーミテイジ検定するには，`prop.trend.test(malaria,pop,mosquito)` と打てばよい。結果は，$\chi^2 = 30.043$ で，有意確率は 10^{-8} のオーダーなので，対数オッズがスコアと比例して変化する傾向があるという対立仮説が採択される。

２つのカテゴリ変数の独立性の検定

今回は，２つのカテゴリ変数の関係を扱う。とくに関連性についての分析は，`vcd` ライブラリや `epitools` ライブラリを導入しておくと楽なのだが，一般ユーザ権限では，演習室のコンピュータに新しくライブラリをインストールすることができないので，この演習では，定義式を入力して計算を行う。

まずは２つのカテゴリ変数が独立である（つまり，関係がない）という帰無仮説を検定する場合を考える。

例えば，肺がんと判明した男性患者 100 人と，年齢が同じくらいの健康な男性 100 人を標本としてもってきて，それまで 10 年間にどれくらい喫煙をしたかという聞き取りを行うという「患者対照研究 (case control

study)」を実施した場合に，喫煙の程度を「一度も吸ったことがない」から「ずっとヘビースモーカーだった」まで何段階かのスコアを振れば，喫煙状況という変数と肺がんの有無という変数の組み合わせが得られる。そこで，それらが独立であるかどうか（関連がないかどうか）を検討することになるわけである[*1]。

クロス集計とは？

前回みたとおり，カテゴリ変数のもつ統計的な情報は，カテゴリごとの度数だけである。そこで，２つのカテゴリ変数の間に関係について検討したいときには，まず**それらの組み合わせの度数**を調べた表を作成する。これをクロス集計表と呼ぶ。クロス集計表を図示するには，`mosaicplot()` という関数が使える。

とくに，２つのカテゴリ変数が，ともに２値変数のとき，そのクロス集計は２×２クロス集計表（２×２分割表）と呼ばれ，その統計的性質が良く調べられている。

独立性のカイ二乗検定の原理

独立性の検定としては，２つのカテゴリ変数の間に**関連がない**と仮定した場合に推定される期待度数を求めて，それに観測度数が適合するかを検定するカイ二乗検定が最も有名である（だから，実はカイ二乗適合度検定と同じ原理である）。もちろん，ある種の関連が仮定できれば，その仮定の元に推定される期待度数と観測度数との適合を調べてもいいが，一般に，２つのカテゴリ変数の間にどれくらいの関連がありそうかという仮定はできないことが多い。そこで，関連がない場合の期待度数を推定し，それが観測値に適合しなければ関連がないとはいえない，と推論するのである。

	A	$\bar{A}$
B	a 人	b 人
$\bar{B}$	c 人	d 人

２つのカテゴリ変数 A と B が，それぞれ「あり」「なし」の２つのカテゴリ値しかとらないとき，これら２つのカテゴリ変数の組み合わせは「A も B もあり（$A \cap B$）」「A なし B あり（$\bar{A} \cap B$）」「A あり B なし（$A \cap \bar{B}$）」「A も B もなし（$\bar{A} \cap \bar{B}$）」の４通りしかない。それぞれの度数を数えあげた結果が，上記の表として得られたときに，母集団の確率構造が，

	A	$\bar{A}$
B	π_{11}	π_{12}
$\bar{B}$	π_{21}	π_{22}

であるとわかっていれば，$N = a + b + c + d$ として，期待される度数は，

	A	$\bar{A}$
B	$N\pi_{11}$	$N\pi_{12}$
$\bar{B}$	$N\pi_{21}$	$N\pi_{22}$

であるから，

$$\chi^2 = \frac{(a - N\pi_{11})^2}{N\pi_{11}} + \frac{(b - N\pi_{12})^2}{N\pi_{12}} + \frac{(c - N\pi_{21})^2}{N\pi_{21}} + \frac{(d - N\pi_{22})^2}{N\pi_{22}}$$

として，自由度３のカイ二乗検定をすればよいことになる。しかし，普通，π は未知である。そこで，$Pr(\bar{A}) = 1 - Pr(A)$ かつ，この２つのカテゴリ変数が独立ならば $Pr(A \cap B) = Pr(A)Pr(B)$ と考えれば良

[*1] ただし，むしろこのデザインは，肺がん患者は健康な人に比べて，どれくらい喫煙していた割合が高いか，を評価するためのデザインである（既に亡くなっている人が除かれてしまっているので，発生リスクは過小評価されるかもしれない）。逆に，喫煙者と非喫煙者を 100 人ずつ集めて，その後の肺がん発生率を追跡調査する前向き研究（フォローアップ研究）では，非喫煙群に比べて，喫煙者ではどれくらい肺がんの発生率が高いかを評価できる。「……に比べてどれくらい高いか」を示すためには，リスク比とかオッズ比のような「比」を用いるのが普通である。これらの「比」については後半で扱う。

いことを使って，$Pr(A)$ と $Pr(B)$ を母数として推定する[*2]。$Pr(A)$ の点推定量は，B を無視して A の割合と考えれば $(a+c)/N$ であることは自明である。同様に，$Pr(B)$ の点推定量は，$(a+b)/N$ となる。したがって，$\pi_{11} = Pr(A \cap B) = Pr(A)Pr(B) = (a+c)(a+b)/(N^2)$ となる。

同様に考えれば，母集団の各セルの確率は下式で得られる。

$$\pi_{12} = (b+d)(a+b)/(N^2)$$

$$\pi_{21} = (a+c)(c+d)/(N^2)$$

$$\pi_{22} = (b+d)(c+d)/(N^2)$$

これらの値を使えば，

$$\chi^2 = \frac{\{a-(a+c)(a+b)/N\}^2}{\{(a+c)(a+b)/N\}} + \frac{\{b-(b+d)(a+b)/N\}^2}{\{(b+d)(a+b)/N\}} + \frac{\{c-(a+c)(c+d)/N\}^2}{\{(a+c)(c+d)/N\}} + \frac{\{d-(b+d)(c+d)/N\}^2}{\{(b+d)(c+d)/N\}}$$

$$= \frac{(ad-bc)^2\{(b+d)(c+d)+(a+c)(c+d)+(b+d)(a+b)+(a+c)(a+b)\}}{(a+c)(b+d)(a+b)(c+d)N}$$

分子の中括弧の中は N^2 なので，結局，

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

ただし通常は，イェーツの連続性の補正を行う。カイ二乗分布は連続分布なので，各度数に 0.5 を足したり引いたりしてやると，より近似が良くなるという発想である。この場合，

$$\chi_c{}^2 = \frac{N(|ad-bc|-N/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

が自由度 1 のカイ二乗分布に従うと考えて検定する。ただし，$|ad-bc|$ が $N/2$ より小さいときは補正の意味がないので，$\chi^2 = 0$ とする。

実際の検定は R を使えば，クロス集計表が既に得られているとき，例えば a=12, b=8, c=9, d=10 などとわかっていれば，`x <- matrix(c(12,9,8,10),nr=2)` として表を与え（あまり 2×2 の場合は意味がないが，必要なら `mosaicplot(x)` として図示してから），`chisq.test(x)` とするだけでいい[*3]。各度数が未知で，各個人についてカテゴリ変数 A と B の生の値が名義尺度として得られているときは，`table(A,B)` とすればクロス集計表が作成できる。そこで，`chisq.test(table(A,B))` とすれば，独立性のカイ二乗検定ができる[*4]。

R では，`chisq.test()` 関数の中で，`simulate.p.value=TRUE` というオプションを使えば，シミュレーションによってそのカイ二乗値より大きなカイ二乗値が得られる確率を計算させることもできる。この方がたんなるカイ二乗検定よりも正確な p 値が得られるが，遅いコンピュータだと計算時間がかかる欠点がある。

例題

肺ガンの患者 100 人に対して，1 人ずつ性・年齢が同じ健康な人を対照として 100 人選び（この操作をペアマッチサンプリングという），それぞれについて過去の喫煙の有無を尋ねた結果，患者群では過去に喫煙を経験した人が 80 人，対照群では過去に喫煙を経験した人が 55 人だった。肺ガンと喫煙は無関係といえるか？　独立性のカイ二乗検定をせよ。

帰無仮説は，肺ガンと喫煙が無関係（独立）ということである。クロス集計表を作ってみると，

[*2] $Pr(X)$ はカテゴリ X の出現確率を示す記号である。また，2 つの母数をデータから推定するので，得られるカイ二乗統計量が従う分布の自由度は 3 より 2 少なくなり，自由度 1 のカイ二乗分布となる。

[*3] 連続性の補正を行わないときは `chisq.test(x,correct=F)` とするが，通常その必要はない。

[*4] 実は `chisq.test(A,B)` でもカイ二乗検定は可能だが，表を与える形にしておく方がよい。

	肺ガン患者群	健康な対照群	合計
過去の喫煙経験あり	80	55	135
過去の喫煙経験なし	20	45	65
合計	100	100	200

となる。この表は，`matrix(c(80,20,55,45),nr=2)` で得られる。肺ガンと喫煙が無関係だという帰無仮説の下で期待される各カテゴリの人数は，

	肺ガンあり	肺ガンなし
喫煙あり	$135 \times 100/200 = 67.5$	$135 \times 100/200 = 67.5$
喫煙なし	$65 \times 100/200 = 32.5$	$65 \times 100/200 = 32.5$

となる。従って，連続性の補正を行なったカイ二乗統計量は，

$$\chi_c^2 = (80-68)^2/67.5 + (55-67)^2/67.5 + (20-32)^2/32.5 + (45-33)^2/32.5 = 13.128...$$

となり，自由度１のカイ二乗分布で検定すると `1-pchisq(13.128,1)` より有意確率は 0.00029... となり，有意水準 5%で帰無仮説は棄却される。つまり，肺ガンの有無と過去の喫煙の有無は独立とはいえない。R では

```
X <- matrix(c(80,20,55,45),nr=2)
chisq.test(X)
```

と入力すれば，下枠内の結果が得られる。

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  X
X-squared = 13.1282, df = 1, p-value = 0.0002909
```

この検定は，肺ガン群と対照群の間で，過去の喫煙者の割合に差があるかどうかを検定することと数学的に同値である。下枠内を実行すれば，まったく同じ検定結果が得られる。

```
smoker <- c(80,55)
pop <- c(100,100)
prop.test(smoker,pop)
```

ただし，カイ二乗検定はあくまで正規近似なので，ある程度各カテゴリの組み合わせごとの期待頻度が大きくないと近似が悪くなってしまう。一般に，期待度数が５以下の組み合わせが検討すべき組み合わせ数の 20%以上あるときは（例えば２×２クロス集計表なら，１つでも期待度数５以下の組み合わせがあれば）カイ二乗検定は適当でないとされる。

フィッシャーの直接確率（正確な確率）

期待度数が低い組み合わせがあるときには，前回の資料に書いたようにカテゴリを併合して変数を作り直す方法もあるが，もっといい方法が考案されている。

ここで調べたいのは組み合わせの数なので，周辺度数を固定して（各々の変数については母比率が決まっていると仮定して）すべての組み合わせを考え，それらが起こる確率（超幾何分布に従う）を１つずつ計算し，得られている集計表が得られる確率よりも低い確率になるような表が得られる確率をすべて足し合わせてしまえば，２つのカテゴリ変数の間に関連がないという帰無仮説の下でそういう表が偶然得られる確率がどれほど

低いのかを，直接計算することができる。こうして計算される確率を，フィッシャーの直接確率，あるいは，フィッシャーの正確な確率（検定）という。これなら，近似ではないので，期待度数が低い組み合わせがあっても問題ない。

もう少し丁寧に言うと，サイズ N の有限母集団があって，そのうち変数 A の値が 1 である個体数が m_1，1 でない個体数が m_2 あるときに，変数 B の値が 1 である個体数が n_1 個（1 でない個体数が $n_2 = N - n_1$ 個）あるという状況を考え，この n_1 個のうち変数 A の値が 1 である個体数がちょうど a である確率を求めることになる。これは，m_1 個から a 個を取り出す組み合わせの数と m_2 個から $n_1 - a$ 個を取り出す組み合わせの数を掛けて，N 個から n_1 個を取り出す組み合わせの数で割った値になる。これと同じ周辺度数をもつ 2×2 分割表のうち，確率がこれと同じかこれよりも小さい表の確率をすべて足し合わせたものが，「変数 A と変数 B が独立」という帰無仮説が成り立つ確率になる[*5]。

フィッシャーの正確な確率は，R では，`fisher.test(table(A,B))` で実行できる。この方がカイ二乗検定よりも正確である。クロス集計表を使って 2 つのカテゴリ変数間の独立性の検定をするときは，コンピュータが使えるならば，サンプルサイズがよほど大きくない限り，常にカイ二乗検定ではなく，フィッシャーの正確な確率を求めるべきである。

例題

上記の肺ガンの有無と過去の喫煙の有無のデータでフィッシャーの直接確率を計算せよ。

既に `X` にクロス集計表が付値されているので，`fisher.test(X)` を実行すると，有意確率は 0.0002590 と得られ，有意水準 5%で「肺ガンの有無と過去の喫煙の有無は独立」という帰無仮説は棄却される。なお，このように 2×2 クロス集計表を分析する場合は，`fisher.test()` 関数は，後で説明するオッズ比とその 95%信頼区間も同時に計算してくれる。

サンプル数が少ない場合について，実際に数値を使って説明しておく。Fisher の正確な確率は仮定が少ない分析法で，とくにデータ数が少なくてカイ二乗検定が使えない場合にも使えるので，動物実験などでは重宝する。

例題

15 人の健康なボランティアに数値計算をしてもらったところ，得点が高得点群と低得点群の 2 群にきれいに分かれたとする。この人たちに，その日に朝食を食べてきたかどうかを尋ねた結果，食べてきた人とこなかった人がいたとする。個人別のデータは下表の通りであったとする。朝食を食べたかどうかと数値計算の得点が独立かどうか検定せよ。

ID 番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
得点（高得点:H，低得点:L）	H	H	L	H	L	L	H	H	L	L	L	L	L	L	L
朝食（食べた:B，食べない:N）	B	B	B	N	N	N	B	B	N	N	N	B	N	B	N

R で答えを得るには，下枠内を入力すればよい（入力が面倒なら，もちろん，H を 1，L を 2，B を 1，N を 2 と数値で置き換えても問題ない）。

[*5] 有限母集団からの非復元抽出になるので，平均 $E(a)$ と分散 $V(a)$ は，

$$E(a) = n_1 m_1 / N$$

$$V(a) = \{(N - n_1)/(N - 1)\} n_1 (m_1/N)(m_2/N) = (m_1 m_2 n_1 n_2)/\{N^2(N - 1)\}$$

となる。実際には組み合わせ計算が多いので，手計算で実行することはまずありえず，統計ソフトにやらせることになる。また，個々の 2×2 分割表の確率は離散値をとるので，同じ確率の表がありうる場合に，それを足し算に含めるのかどうかは難しい点である。これを乱数によって決める「ランダム検定」という手法もあるが，あまり一般的ではない。

```
calc <- as.factor(c('H','H','L','H','L','L','H','H','L','L','L','L','L','L','L'))
bf <- as.factor(c('B','B','B','N','N','N','B','B','N','N','N','B','N','B','N'))
print(X <- table(bf,calc))
fisher.test(X)
```

3 行目で得られる結果からクロス集計表を書いてみると，

	高得点	低得点	合計
朝食あり	4	3	7
朝食なし	1	7	8
合計	5	10	15

である。15 人のうち 5 人が高得点で，7 人が朝食あり，という条件が決まっているとき[*6]，偶然この表が得られる確率は，15 人のうち高得点の 5 人の内訳が，朝食を食べた 7 人から 4 人と，食べていない 8 人から 1 人になる確率となる。つまり，15 から 5 を取り出す組み合わせのうち，7 から 4 を取り出し，かつ残りの 8 から 1 を取り出す組み合わせをすべて合わせたものが占める割合になるので，${}_7\mathrm{C}_4 \cdot {}_8\mathrm{C}_1 / {}_{15}\mathrm{C}_5 \simeq 0.0932$ である。

つまり，上のクロス集計表が，偶然（2 つの変数に何も関係がないとき）得られる確率は 0.0932 ということである。これだけでも既に 5%より大きいので，「2 つの変数が独立」という帰無仮説は棄却されず，得点の高低と朝食の有無は関係がないと判断していいことになる。

しかし，有意確率，つまり第一種の過誤を起こす確率は，得点の高低と朝食の有無には関係がないと判断した場合にそれが間違っている確率なので，この表だけではなく，この表よりも偶然得られる確率が低い表が得られる確率をすべて足さねばならない。周辺度数が上の表と同じ表は，

(1)	高得点	低得点	(2)	高得点	低得点	(3)	高得点	低得点	合計
朝食あり	0	7		1	6		2	5	7
朝食なし	5	3		4	4		3	5	8
合計	5	10		5	10		5	10	15

(4)	高得点	低得点	(5)	高得点	低得点	(6)	高得点	低得点	合計
朝食あり	3	4		4	3		5	2	7
朝食なし	2	6		1	7		0	8	8
合計	5	10		5	10		5	10	15

の計 6 種類しかない。(1) や (6) の表よりもさらに稀な場合を考えると，(1) の先は高得点かつ朝食ありの人の数がマイナスになってしまうし，(6) の先は高得点かつ朝食なしの人の数がマイナスになってしまう。

そこで，すべての表について，それが偶然得られる確率を計算すると，[*7](1) は ${}_7\mathrm{C}_0 \cdot {}_8\mathrm{C}_5 / {}_{15}\mathrm{C}_5 \simeq 0.0186$，(2) は ${}_7\mathrm{C}_1 \cdot {}_8\mathrm{C}_4 / {}_{15}\mathrm{C}_5 \simeq 0.1632$，(3) は ${}_7\mathrm{C}_2 \cdot {}_8\mathrm{C}_3 / {}_{15}\mathrm{C}_5 \simeq 0.3916$，(4) は ${}_7\mathrm{C}_3 \cdot {}_8\mathrm{C}_2 / {}_{15}\mathrm{C}_5 \simeq 0.3263$，(5) は上で計算した通り ${}_7\mathrm{C}_4 \cdot {}_8\mathrm{C}_1 / {}_{15}\mathrm{C}_5 \simeq 0.0932$，(6) は ${}_7\mathrm{C}_5 \cdot {}_8\mathrm{C}_0 / {}_{15}\mathrm{C}_5 \simeq 0.0070$ となる[*8]。

以上の計算より，元の表（= (5)）より得られる確率が低い（つまりより偶然では得られにくい）表は (1) と (6) なので，それらを足して，元の表の両側検定（どちらに歪んでいるかわからない場合）での有意確率は，$0.0932 + 0.0186 + 0.0070 = 0.1188$ となる（`fisher.test()` の結果と小数第 4 位で 1 違うのは丸め誤差のせいである）。

[*6] 「各変数については母比率が決まっているとき」ということで，このことを「全ての周辺度数が固定されているとき」ともいうのである。

[*7] 既出の通り，R で組み合わせ計算を行う関数は `choose()` である。例えば ${}_7\mathrm{C}_3$ は，`choose(7,3)` で計算できる。

[*8] これらの確率をすべて足すと 1 になる。上の計算値として書いた値を使うと 0.9999 となるが，これは丸め誤差のせいであり，厳密に計算すれば 1 になる。

研究デザインと疫学指標

独立とはいえないなら，次に調べることは，どの程度の関連性があるのかということである。研究デザインによって得られる関連性の指標は異なる。具体的な解析方法に入る前に，疫学の基礎知識が必要なのでまとめておく。

疫学指標の基礎知識 (1) 頻度の指標

集団内の疾病の状況を表すためには，たんに患者数だけでは不十分である。まず，どのくらいの規模のどういう集団をどのくらいの期間観察したのか，という意味で，分母を定義することが必要である。分子を定義する上では，どういう診断基準で判定したのかということと，その診断基準の信頼性 (reliability)・妥当性 (validity)・正確さ (accuracy)・精度 (precision) を把握し高めることが重要である。

具体的な指標としては，まず，以下の 3 つを区別する必要がある[a]。

有病率 (prevalence)　有病割合と呼ぶ方が紛れがない。一時点での人口に対する患者の割合で無次元である。一時点でのということを明示するには，point prevalence という。急性感染症で prevalence が高いなら患者が次々に発生していることを意味するが，慢性疾患の場合はそうとは限らない。行政施策として必要な医療資源や社会福祉資源の算定に役立つ。例：高血圧や高コレステロール血症は prevalence が高いので，対策がとられている。

累積罹患率 (cumulative incidence)　通常，たんにリスク (risk) といえば，この累積罹患率を指す。期首人口のうち観察期間中に病気になった人数の割合であり，無次元である。当然，観察期間が短ければ小さい値になるので，「20 年間のがんの発症リスク」のような表現になる。脱落者は分母から除外する（脱落分を正しく扱うためには生存時間解析という手法を用いる）。無作為割付けの介入研究でよく使われる指標である。

罹患率 (incidence rate)　発生率ともいう。個々の観察人年の総和で発生数を割った値。そのため，観察期間によらない値になる。次元は 1 ／年。International Epidemiological Association の Last JM [Ed.]“A Dictionary of Epidemiology, 4th Ed.”(Oxford Univ. Press, 2001) に明記されているように，incidence は発生数である。感受性の人の中で新たに罹患する人が分子。再発を含む場合はそう明記する必要がある。意味としては，瞬時における病気へのかかりやすさ。つまり疾病罹患の危険度（ハザード）を示す。疾病発生状況と有病期間が安定していれば，平均有病期間＝有病割合／罹患率という関係が成り立つ。ランダム化臨床試験でよく使われる指標である。

さらに，オッズ (odds) という概念を押さえておく必要がある。オッズとは，**ある事象が起きる確率の起きない確率に対する比**である。一時点での非患者数に対する患者数の比を疾病オッズ (disease-odds) と呼ぶ。また，患者対照研究などで，過去に何らかの危険因子に曝露した人数の，曝露していない人数に対する比を曝露オッズ (exposure-odds) と呼ぶ。

---

[a] 参考までにまとめておくと，以下の指標も疫学的には重要である。

死亡率 (mortality rate)　人口のうち，ある一定期間に死亡した人数の割合。分母分子ともカテゴリ分けしてカテゴリごとに計算した死亡率はカテゴリ別死亡率 (category-specific mortality rate) となる。死因別死亡率 (disease-specific mortality rate) など。一般に期間は 1 年間とするので，分母は 1 年間の半ばの人口を使い，それを年央人口と呼ぶ（日本の人口統計では 10 月 1 日人口を用いる）。意味としては，疾病がもたらす結果の 1 つを示す指標といえる。年齢によって大きく異なるので，年齢で標準化することが多い。

致命率 (case-fatality rate)　ある疾病に罹患した人のうち，その疾病で死亡した人の割合（%で表す）。意味としては，疾病の重篤度を示すものである。ただし慢性疾患では有病期間が長いので，観察期間の設定が重要である。一般に，致命率＝死亡率／罹患率という関係が成り立つ。

死因別死亡割合 (proportional mortality rate; PMR)　ある特定の死因による死亡が全死亡に占める割合。増減はその疾患の増減だけでなく，他の疾患の増減とも連動する。

PMI (proportional mortality indicator)　50 歳以上死亡割合と訳す。全死亡数に対する 50 歳以上死亡数の占める割合を%表示した値である。計算に必要なのは年齢 2 区分の死亡数のみなので，統計資料が整備されていない途上国でも信頼性が高い値を得ることができる利点がある。

疫学指標の基礎知識 (2) 効果・関連性の指標

その上で，何らかの危険因子への曝露があると，なかった場合に比べて何倍くらい病気に罹りやすさが上昇する**効果**をもつかといったことを推論することになる。効果の指標としては，以下の４つを区別しておこう。

相対危険（Relative Risk） 以下の３つの総称。リスク比を指している場合が多い。

リスク比 (risk ratio) 累積罹患率比（cumulative incidence rate ratio）ともいう。曝露群のリスクの非曝露群のリスクに対する比である。

罹患率比 (incidence rate ratio) 曝露群の罹患率の非曝露群の罹患率に対する比をいう。

死亡率比 (mortality rate ratio) 曝露群の死亡率の非曝露群の死亡率に対する比をいう。罹患率比と死亡率比を合わせて率比（rate ratio）という。

オッズ比 (odds ratio) オッズの比。2 種類のオッズ比（コホート研究における累積罹患率のオッズ比と患者対照研究における曝露率のオッズ比）は数値としては一致する。オッズ比は比較的簡単に得られる値なので，率比の近似値として価値がある。

寄与危険 (attributable risk) 危険因子への曝露による発症増加を累積罹患率（リスク）または罹患率の差で表した値。つまり，累積罹患率差＝リスク差 (risk difference)，または罹患率差 (incidence rate difference) である。超過危険 (excess risk) ともいう。

寄与割合 (attributable proportion) 曝露群の罹患率のうちその曝露が原因となっている割合。つまり罹患率差を曝露群の罹患率で割った値になる。罹患率比から 1 を引いて罹患率比で割った値とも等しい。

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，例えば 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）[*9]とか断面研究（cross-sectional study）[*10]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ことが利点である。断面研究や患者対照研究における曝露オッズの比を曝露オッズ比 (exposure-odds ratio)，断面研究やコホート研究における疾病オッズの比を疾病オッズ比 (disease-odds ratio) と呼ぶ。

では，クロス集計表から，これらの値を計算してみよう。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

[*9] 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。

[*10] 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。

リスク比とオッズ比の点推定量

点推定量の計算は簡単である。この表でいえば，リスク比は $\frac{a/m_1}{c/m_2}$ となり，疾病オッズ比は $\frac{a/b}{c/d} = \frac{ad}{bc}$ である。曝露オッズ比は $\frac{a/c}{b/d} = \frac{ad}{bc}$ となり，数値としては疾病オッズ比と一致する。ただし，R の `fisher.test()` で計算されるオッズ比は，この単純な計算式から得られる値と異なっている。`fisher.test()` では周辺分布をすべて固定したクロス集計表の最初の要素に対して，非心度パラメータがオッズ比で与えられるような非心超幾何分布を仮定して最尤推定がなされる。`vcd` ライブラリの `oddsratio()` 関数でも連続性の修正がなされるため，`log=F` として対数オッズにしない指定をしても，$\frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)}$ が計算される。

オッズ比が重要なのは，稀な現象をみるときに，リスク比のよい近似になるからであると言われている。例えば，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年間の白血病の累積罹患率（リスク）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	99996	100000
送電線から離れて居住	2	99998	100000
合計	6	199994	200000

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（$(4/100000)/(2/100000) = 2$，つまりリスク比が 2 なので）ということができる。ここで疾病オッズ比をみると，$(4 * 99998)/(2 * 99996) \simeq 2.00004$ と，ほぼリスク比と一致していることがわかる。[*11]

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。

	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，白血病患者からのサンプルだから）が，白血病の人の送電線の近くに居住した経験の曝露オッズは 0.25 となり，白血病でない人ではそのオッズが 0.111... となるので，これらの曝露オッズの比は 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，患者対照研究で曝露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。例えば，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

[*11] 上述のように R のプログラムを使って `fisher.test(matrix(c(4,2,99996,99998), nr=2))` として最尤推定しようとすると，Windows マシンではメモリ不足でエラーを起こしてしまって計算できない（サンプルサイズが大きすぎるため）。`vcd` ライブラリの `oddsratio()` 関数を使うと，`oddsratio(matrix(c(4,2,99996,99998), nr=2),log=F)` より，1.800036 が得られる。

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[*12]。

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。

目的によっては，リスク比やオッズ比の他に，寄与危険（＝リスク差），寄与割合（＝曝露寄与率），相対差，母集団寄与率，Yule の Q，ファイ係数といった指標も用いられるけれども，これらは点推定量だけが求められることが多い。

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，κ 係数などの一致度の指標を計算するべきである（後述）。

リスク比とオッズ比の 95%信頼区間

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する。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点推定量は，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。リスク比の点推定量は $RR = \pi_1/\pi_2 = (Xm_2)/(Ym_1)$ となる。

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，

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

$$RR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad \text{（上限）} \tag{2}$$

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（下枠内参照）。

```
X <- 4
m1 <- 100000
Y <- 2
m2 <- 100000
print(RR <- (X/m1)/(Y/m2))
RR*exp(-qnorm(0.975)*sqrt(1/X-1/m1+1/Y-1/m2))
RR*exp(qnorm(0.975)*sqrt(1/X-1/m1+1/Y-1/m2))
```

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。

[*12] 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。

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[*13]。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，ここでは扱わない。

その他の関連性の指標

寄与危険（リスク差） 曝露群のリスクと対照群のリスクの差である。リスク比の計算で用いた記号で表せば，$\pi_1 - \pi_2$

寄与割合（曝露寄与率） 真に要因の影響によって発症した者の割合。同じ記号で表せば，$(\pi_1 - \pi_2)/\pi_1$

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。同じ記号で表せば，$(\pi_1 - \pi_2)/(1 - \pi_2)$

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，$(\pi - \pi_2)/\pi$

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。$Q = (OR - 1)/(OR + 1)$

ファイ係数 (ρ) 要因の有無，発症の有無を 1,0 で表した場合のピアソンの積率相関係数である。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，$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$。なお，ファイ係数は `vcd` ライブラリの `assoc.stats()` 関数で計算できる。

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。test-retest reliability（検査再検査信頼性）の指標といえる。ちなみに繰り返し調査に限らないが，カテゴリ変数間の一致度をみるための作図には，`vcd` ライブラリに含まれている `agreementplot()` という関数が有用である。

	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 の分散 $V(\kappa)$ は，$V(\kappa) = P_e/(N \cdot (1 - P_e))$ 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，帰無仮説「$\kappa = 0$」を検定したり，κ の 95%信頼区間を求めたりすることができる。ここは `vcd` ライブラリの `Kappa()` 関数に倣って[*14]，2×2 のクロス集計表を与えたときに κ の点推定量と 95%信頼区間を計算し，有意確率を計算する R の関数を書いてみよう。

[*13] 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。

[*14] 注：この `Kappa()` 関数はm×mのクロス集計表について一致度と信頼区間を計算し検定もしてくれる。

```
kappa.test <- function(x) {
x <- as.matrix(x)
a <- x[1,1]; b <- x[1,2]; c <- x[2,1]; d <- x[2,2]
m1 <- a+b; m2 <- c+d; n1 <- a+c; n2 <- b+d; N <- sum(x)
Pe <- (n1*m1/N+n2*m2/N)/N
Po <- (a+d)/N
kappa <- (Po-Pe)/(1-Pe)
seK0 <- sqrt(Pe/(N*(1-Pe)))
seK <- sqrt(Po*(1-Po)/(N*(1-Pe)^2))
p.value <- 1-pnorm(kappa/seK0)
kappaL<-kappa-qnorm(0.975)*seK
kappaU<-kappa+qnorm(0.975)*seK
list(kappa=kappa,conf.int=c(kappaL,kappaU),p.value=p.value)}
```

上枠内のように関数定義した後で，例えば，○×で回答する項目について 2 回の繰り返し調査をしたときに，1 度目も 2 度目も○であった人数が 10 人，1 度目は○で 2 度目は×であった人数が 2 人，1 度目は×で 2 度目は○であった人数が 3 人，1 度目も 2 度目も×であった人数が 19 人であったときは下枠内のように入力する。

```
kappa.test(matrix(c(10,3,2,19),nr=2))
```

Cronbach の α

一致度つながりで Cronbach の α にも一言触れておく。質問紙調査においては，多くの概念は直接聞き取ることができないので，複数の質問を組み合わせることによって対象者の差異をより細かく把握しようと試みることがある。回答が同じように変動していれば，それらの質問によって同じ上位概念を聞き取れている信頼性が高いと考え，それを指標化したものが Cronbach の α 係数である。
例えば，自然への親近感を聞き取りたい場合に，

（1）あなたは自然が好きですか？　嫌いですか？　（好き，どちらかといえば好き，どちらかといえば嫌い，嫌い）

だけでは対象者は4群にしか分かれない（順序尺度として数値化すると，好きを4点，嫌いを1点として1点から4点の4段階)。しかし，

（2）休日に海や山で過ごすのと映画館や遊園地で遊ぶのとどちらが好きですか？　（海や山，どちらかといえば海や山，どちらかといえば映画館や遊園地，映画館や遊園地）

を加えて，これも「海や山」を4点，「映画館や遊園地」を1点とする順序尺度として扱うことにすれば，（1）と（2）の回答の合計点を計算すると，2点から8点までの7群に回答者が類別される可能性があり，より細かい把握が可能になる。さらに，

（3）無人のジャングルで野生生物の観察をする仕事に魅力を感じますか？　それとも感じませんか？　（感じる，どちらかといえば感じる，どちらかといえば感じない，感じない）

の4点を加えると，3点から 12 点までの 10 段階になる。この合計得点を「自然への親近感」を表す尺度として考えてみると，3つの項目は同じ概念を構成する項目（下位概念）として聞き取られているので，互いに回答が同じ傾向になることが期待される。つまり（1）で好きと答えた人なら，（2）では海や山と答える人が多いだろうし，（3）では感じないと答えるよりも感じると答える人が多いと考える。同じ概念を構成する質問に対して同じ傾向の回答が得られれば，その合計得点によって示される尺度は，信頼性が高いと考えられる。
上記3つの質問に対して一貫した答えが得られたかどうかを調べる方法の1つに折半法がある。例えば質問（1）と（3）の合計点の変数 x_{13} と質問（2）の点の変数 x_2 という具合に，同じ概念を構成する全質問を2つにわけて，x_{13} と x_2 の相関係数を $r_{x_{13}x_2}$ とすれば，これらの質問の信頼性係数 $\alpha_{x_{13}x_2}$ は，$\alpha_{x_{13}x_2} = \frac{2r_{x_{13}x_2}}{1+r_{x_{13}x_2}}$ となるというのがスピアマン・ブラウンの公式である。
折半法では通常，奇数番目の項目と偶数番目の項目に二分するが，（1）の点と（2）と（3）の合計点という分け方もあるわけで，下位概念が3つ以上ある質問だったら，これらの回答に一貫して同じ傾向があるかどうかをスピアマン・ブラウンの公式で出そうと思うと，α の値はいくつも（n 個の下位概念からなるなら，n 項目を2つに分ける組み合わせの数だけ）できる。この例では，$\alpha_{x_1x_{23}}$, $\alpha_{x_{12}x_3}$ も計算する必要がある。
それをまとめてしまおうというのが **Cronbach の** α で，仮に（1）（2）（3）の合計得点が「自然への親近感」を表す変数 x_t だとして，（1）（2）（3）の得点をそれぞれ変数 x_1，x_2，x_3 とすれば，Cronbach の α は，

$$\alpha = \frac{3}{3-1}(1 - \frac{s_{x_1}^2 + s_{x_2}^2 + s_{x_3}^2}{s_{x_t}^2})$$

となる（s_{x_1} は x_1 の不偏標準偏差である。以下同様)。Cronbach の α が 0.8 以上なら十分な，0.7 でもまあまあの，内的一貫性（信頼性）がその項目群にはあるとみなされる。

Cronbach の α の計算例

学生 5 人から上の 3 問に対する回答が下表のように得られたとする。

(1)	(2)	(3)	合計
3	2	1	6
3	3	2	8
3	2	3	8
4	3	3	10
3	3	2	8

まず，スピアマン・ブラウンの公式で信頼性係数 $\alpha_{x_{12}x_3}$ は，x_{12} と x_3 の相関係数 $r_{x_{12}x_3}$ は，$\bar{x_{12}} = 5.8$，$\bar{x_3} = 2.2$ なので，

$$s_{x_{12}} = \sqrt{\frac{(5-5.8)^2 \times 2 + (6-5.8)^2 \times 2 + (7-5.8)^2}{(5-1)}}$$

$$s_{x_3} = \sqrt{\frac{(1-2.2)^2 + (2-2.2)^2 \times 2 + (3-2.2)^2 \times 2}{(5-1)}}$$

$$s^2_{x_{12}x_3} = \frac{(5-5.8)(1-2.2) + (6-5.8)(2-2.2) + (5-5.8)(3-2.2) + (7-5.8)(3-2.2) + (6-5.8)(2-2.2)}{(5-1)}$$

から計算して，

$$r_{x_{12}x_3} = \frac{s^2_{x_{12}x_3}}{(s_{x_{12}} s_{x_3})} = 0.52$$

となる。これを使って，

$$\alpha_{x_{12}x_3} = \frac{2 \times 0.52}{1 + 0.52} = 0.68$$

が得られる。

一方，Cronbach の α は，$s^2_{x_1} = 0.2, s^2_{x_2} = 0.3, s^2_{x_3} = 0.7, s^2_{x_t} = 2$ より，$\alpha = 3/2 * (1 - (0.2 + 0.3 + 0.7)/2) = 0.6$ となる。

ちなみに，`X`，`Y`，`Z` が同じ概念の下位尺度となるスコアの変数だとして，Cronbach の α 係数を計算するための R のプログラムは以下の通り。

```
T <- X+Y+Z; VX <- var(X); VY <- var(Y); VZ <- var(Z); VT <- var(T)
alpha <- (3/2)*(1-(VX+VY+VZ)/VT)
print(alpha)
```

課題

ある国で炉心溶融事故を起こした原子力発電所を含む郡の人口 40000 人のうち，事故後 5 年間で 400 人が白血病を発症したとする。事故の 10 年前には人口 35000 人で，事故の 5 年前までの 5 年間では，100 人が白血病を発症したとする。この炉心溶融事故は白血病リスクをどれくらい高めたか？　リスク比とその 95%信頼区間を計算して答えよ（注：この数値例はまったく架空の話である）。

結果は配布する紙に学籍番号，氏名と共に自筆して提出すること。結果の提出をもって出席確認とする。