医学情報処理演習第 10 回「クロス集計」[*1]

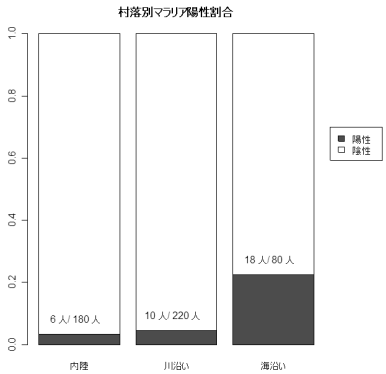
2006 年 12 月 11 日 中澤　港 (nminato@med.gunma-u.ac.jp)
演習サポート web ページ：http://phi.med.gunma-u.ac.jp/medstat/

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。

it09-ans-2006.R

```
malaria <- c(6,10,18)
pop <- c(180,220,80)
names(malaria) <- c("内陸","川沿い","海沿い")
positive <- malaria/pop
negative <- 1-positive
tab <- rbind(positive,negative)
rownames(tab) <- c("陽性","陰性")
print(tab)
op <- par(mar=c(5,4,4,6.5)+0.1,xpd=NA)
ip <- barplot(tab,main="村落別マラリア陽性割合",col=c("red","white"))
legend(ip[3]+0.7,0.7,legend=rownames(tab),fill=c("red","white"))
text(ip-0.05,positive+0.05,paste(malaria,"人/",pop,"人"))
par(op)
prop.test(malaria,pop)
pairwise.prop.test(malaria,pop)
mosquito <- c(1,2,4)
prop.trend.test(malaria,pop,mosquito)
```



prop.test() の結果，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，pairwise.prop.test() を実行すると，2 村落のペアごとに「原虫陽性割合に差が無い」を帰無仮説とする検定の有意確率（Holm の方法で検定の多重性を調整済み）は，内陸と川沿いの間で 0.72，内陸と海沿いの間で 8.0×10^{-6}，川沿いと海沿いの間で 1.3×10^{-5} となる。つまり，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。

最後に，ハマダラカの相対的な密度のスコアを mosquito という変数に与え，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と，$\chi^2 = 30.043$ 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。つまり，ハマダラカの相対的な密度が高いほどマラリア原虫陽性割合が高い傾向が有意にあるといえる。

[*1] 本資料は http://phi.med.gunma-u.ac.jp/medstat/it10-2006.pdf としてダウンロード可能である。

複数のカテゴリ変数を分析するために

今回は，複数のカテゴリ変数の関係を分析する方法を扱う[*2]。カテゴリデータの分析，とくに関連性についての分析には，vcd ライブラリや epitools ライブラリを導入しておくと非常に便利である[*3]。これらの追加ライブラリ内の関数を使いたいときは，library(vcd) のようにライブラリをメモリに読み込むことで，そこに含まれる関数やデータが使える状態になる。なお，ややマニアックな余談だが，vcd ライブラリには goodfit() という適合度検定を行う関数が含まれていて，前回の「複雑な例題」も以下のように簡単に実行することができる[*4]。

```
library(vcd)
hh <- as.table(c(79,61,13,2))
names(hh) <- 0:3
print(res <- goodfit(hh,"poisson","ML"))
summary(res)
plot(res)
```

２つのカテゴリ変数の独立性の検定

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方法について説明する。

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「症例対照研究 (case control study)[*5]」を実施したとする。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，200 人の一人一人について喫煙状況という変数に値が割り振られることになる。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を考え，**クロス集計**することによって，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[*6]。

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。例えば，ある村で健診に訪れた 13 人の性別と病気の有無の調査結果が下表の通りであったとする[*7]。

人	1	2	3	4	5	6	7	8	9	10	11	12	13
性別	男	男	男	男	男	男	女	女	女	女	女	女	女
病気	有	有	有	無	無	無	有	有	有	有	無	無	無

[*2] 実は前回の課題も村落とマラリア原虫陽性という２つのカテゴリ変数の関係の分析とみなせるが，敢えてそういう見方をしなかった。

[*3] 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，昭和分室にお願いして，予め CRAN にあるすべてのライブラリをインストールしてもらった。ついでに R のバージョンも最新の 2.4.0 にしてもらった。もちろん，自分のコンピュータに管理者権限でインストールした R に vcd ライブラリを導入したい場合は，`install.packages("vcd")` として，出てくるウィンドウで適当なミラーサーバ（国内では Japan(Tsukuba) または Japan(Tokyo) を推奨）を選ぶだけでよい。ブロードバンドなら 1 分もかからないだろう。

[*4] ただし分布の推定法に ML 法と MinChisq 法があり，前回説明した期待値を出すのは ML 法であるが，その場合，適合度の検定法も尤度比検定になってしまうので，前回とまったく同じ結果にはならない。なお，期待値推定の中身は goodfit と打てばわかるし，適合度検定の中身は getS3method("summary","goodfit") とすれば見える。UseMethod を使って定義された関数（generic.class という形式）の中身は，一般にただ generic.class と打つのでは見えず，getS3method("generic","class") とする必要がある。

[*5] 患者対照研究ともいう。

[*6] 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。独立性の検定は，一時点の断面研究（英語では cross-sectional study で，「横断的研究」ともいう）で調べた属性変数間でなされるのが普通である。症例対照研究では，既に亡くなっている人が除かれてしまっているので，注目している要因によってその疾患が起こりやすくなる程度が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のコホート研究 (cohort study) 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くなるかを評価できる。「……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。

[*7] このデザインは断面研究である。

この生データをもっとも簡単に R に入力し，性別と病気のクロス集計表を作るには次の枠内を打つ。下から 2 行目の `table()` という関数が，生のカテゴリ変数値の組み合わせをカウントしてクロス集計表を作ってくれ，最下行の `mosaicplot()` という関数が，それをグラフ表示してくれる（この場合はグラフにするまでもないので，この資料には掲載しない）。

it10-1-2006.R

```
pid <- 1:13
sex <- as.factor(c(rep(1,6),rep(2,7)))
levels(sex) <- c("男","女")
disease <- as.factor(c(1,1,1,2,2,2,1,1,1,1,2,2,2))
levels(disease) <- c("有","無")
print(ctab <- table(sex,disease))
mosaicplot(ctab,main="２×２クロス集計表のモザイクプロット例")
```

クロス集計表としては，次の枠内の結果が得られる。

```
    disease
sex  有 無
  男  3  3
  女  4  3
```

とくに，２つのカテゴリ変数が，この例のようにともに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（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い場合に「統計的に有意な関連があった」と判断する。

	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た結果が上表として得られたとき，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，期待される度数は[*8]，

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

であるから，

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

[*8] ただし $N = a + b + c + d$ である。

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，一般に 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 $Pr(A \cap B) = Pr(A)Pr(B)$ と考えられる**ことを使って，$Pr(A)$ と $Pr(B)$ を母数として推定する[*9]。2 つの母数をデータから推定したため，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。$Pr(A)$ 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。

もちろん，R にはこの検定を簡単に行う関数が実装されている。例えば a=12, b=8, c=9, d=10 なら次の通り[*10]。

it10-2-2006.R

```
x <- matrix(c(12,9,8,10),nr=2)
chisq.test(x)
```

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**，`chisq.test(table(A,B))` とすれば，独立性のカイ二乗検定ができる[*11]。

例題

肺がんの患者 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[a]，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。肺がん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。

[a] この操作をペアマッチサンプリングという。ただし，このような症例対照研究でマッチングをすると，却ってバイアスが生じる場合があるので注意されたい。

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

[*9] $Pr(X)$ はカテゴリ X の出現確率を示す記号である。

[*10] 連続性の補正なしなら 2 行目が `chisq.test(x,correct=F)` となるが，通常その必要はない。

[*11] 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。なお，R の `chisq.test()` 関数では，`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長くなってしまうのが欠点である。

	肺がん患者群	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。肺がんと喫煙が無関係という帰無仮説の下で期待される各カテゴリの人数は,

	肺がんあり	肺がんなし
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，肺がん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，肺がん群と対照群の間で，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[*12]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

[*12] 例えば 2 × 2 クロス集計表なら 1 つでも期待度数 5 以下のセルがあれば該当する。

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[*13]。

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。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

上記の肺がん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「肺がん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。サンプルサイズが小さい場合について，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なので，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

この検定を実行するための R のコードは下枠内の通り。

it10-3-2006.R

```
calc <- as.factor(c(1,1,2,1,2,2,1,1,2,2,2,2,2,2,2))
levels(calc) <- c("高得点","低得点")
bf <- as.factor(c(1,1,1,2,2,2,1,1,2,2,2,1,2,1,2))
levels(bf) <- c("朝食あり","朝食なし")
print(X <- table(bf,calc))
fisher.test(X)
```

最下行の `fisher.test()` の結果，`p-value` が 0.1189 なので，5%水準で帰無仮説は棄却されない。つまり，このデータは，数値計算の得点と朝食を食べたかどうかが独立であることを示唆する（もっとも，データが少ないので，検出力が足りずに第 2 種の過誤が起きている可能性はあるが）。この計算結果は，以下のように考えて導かれる。まず，下から 2 行目の結果からクロス集計表を書いてみると，次の通りである[*14]。

	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

[*13] 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。

[*14] `library(vcd)` として `vcd` ライブラリが使えるようにしておいて，`mar_table(X)` とすれば，このような周辺度数を含めた集計表の作成も可能。`epitools` ライブラリの `table.margins()` 関数も同じ機能をもつ。

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[*15]，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，[*16](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[*17]。

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）。

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。カテゴリ変数間の関連については，従来より疫学分野で多くの研究が蓄積されてきた。疫学研究では，研究デザインによって，得られる関連性の指標は異なることに注意しなければならない。その意味で，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。なお，疫学データについて関連性の指標を説明する都合上，以下では，何らかのリスクファクターへの曝露の有無と疾病の有無の関連性の分析について説明するが，数学的には，そうでなくても，2 つのカテゴリ変数間の関連性の程度について一般に通用する部分もある。

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。

大雑把に言えば，信頼性は再現性が確保されているかを意味する。妥当性は測りたいものがきちんと測れているかどうかを意味する。正確さは系統的なズレがないかどうかを示す。精度は偶然誤差の小ささを示す（例えば小数点以下何桁目まで測れているか）。

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。

[*15] 「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。

[*16] 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。

[*17] 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。

有病割合 (prevalence)

有病率と呼ばれることもあるが，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，症例対照研究などで，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。ここでは深入りしないが，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。

頻度の指標を押さえた上で，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。とくにリスク比やオッズ比はよく使われる指標である。

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[*18]。

ところで，病気のリスクは，全体（期首人口）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 (prospective study)」（この意味ではコホート研究 (cohort study) とかフォローアップ研究 (follow-up study) と言ってもいい）でないと，リスク比（に限らず相対危険すべて）は計算できない。

つまり，症例対照研究（case-control study）[*19]とか断面研究（cross-sectional study）[*20]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。これらの研究デザインでは，オッズ比を計算するのが普通である。断面研究や症例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) と呼ぶが，これらは数学的には一致する。

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（表☆としよう）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

[*18] ただし，重要なのは 95%信頼区間が 1 を含むかどうかという意思決定だけではなく，むしろリスク比やオッズ比の点推定量と信頼区間の値そのものである。知りたいのは，非曝露群に比べて曝露群のオッズやリスクが何倍になっているかということであって，それが 1 と差が無いかという判定ではない。Rothman や Greenland に代表される現代の疫学者は，有意性をみる仮説検定は，せっかく関連性の程度が得られているのに，それを有無という 2 値に還元してしまうので情報量の損失が大きく，あまり意味がないと言っている。それゆえ，疫学研究では検定結果よりも 95%信頼区間そのものの方が重要である。Rothman は関連の程度に応じた有意確率の変化を示すという意味で，p-value 関数（リスク比やオッズ比を横軸にとって，「真の値が横軸の値と差が無い」帰無仮説の検定の有意確率=p-value を縦軸にとって，とりうるすべてのリスク比やオッズ比について線で結んだグラフ）を求めるべきだと主張している（Rothman KJ 著，矢野栄二・橋本英樹監訳『ロスマンの疫学』篠原出版新社，pp.150-156）。

[*19] 調査時点で，患者を何人サンプリングすると決め，同数でもいいが通常はその何倍かの人数の対照（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。ただし原則としてマッチングに使った変数で層別解析しなくてはいけない）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。

[*20] 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

$$\frac{a/m_1}{c/m_2} = \frac{am_2}{cm_1}$$

となる。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，結局，疾病オッズ比と一致することがわかる。

ただし，R の `fisher.test()` で計算されるオッズ比は，ad/bc という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。また，`vcd` ライブラリの `oddsratio()` 関数で `log=F` オプションを付けると[*21]定義通りの計算をしてくれる。

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

$(4/100000)/(2/100000) = 2$ から，リスク比が 2 なので，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 倍になったということができる[*22]。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，前向きコホート研究ではなく，症例対照研究を行って，過去の曝露との関係をみる。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，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なので，リスク（累積罹患率）が定義できず，リスク比も計算できない。形の上から無理やり計算しても意味はない。しかし，曝露オッズは計算できる。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

[*21] ただし，どこかのセルが 0 のときは，

$$\frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)}$$

が計算されるので一致しないことに注意。この関数では対数オッズにしないと `summary(oddsratio())` による有意性の検定はできないが，`confint(oddsratio())` による信頼区間の推定は，対数オッズでなくてもできる。また，`epitools` ライブラリにある同名の `oddsratio` 関数は，もっと複雑なことをしているので，注意が必要である。

[*22] 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。

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ので，このように稀な疾患の場合は，大規模コホート研究をするよりも，症例対照研究で曝露オッズ比を求める方が効率が良い（もちろん，コストさえかければ大規模コホート研究の方が強い証拠となるデータが得られるが）。

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，症例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[*23]。

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，２つのカテゴリ変数の関連性を表す指標として，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。また，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。独立でないことは自明だからである。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（これらについては後述する）。

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**から考えるために，曝露あり群と曝露なし群をそれぞれ m_1 人，m_2 人フォローアップして，曝露あり群で X 人，曝露なし群で 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，

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

となる。なお，RR が大きくなると対数変換ではうまく近似できないので，立方根変換しなくてはいけないが，複雑なのでここでは説明しない。R のコードは以下の通り。p 値は帰無仮説 $RR = 1$ の検定の有意確率である。

it10-4-2006.R

```
riskratio2 <- function(X,Y,m1,m2) {
 data <- matrix(c(X,Y,m1-X,m2-Y,m1,m2),nr=2)
 colnames(data) <- c("疾病あり","疾病なし","合計")
 rownames(data) <- c("曝露群","対照群")
 print(data)
 RR <- (X/m1)/(Y/m2)
 n1 <- X+Y; T <- m1+m2; n2 <- T-n1
 p.v <- 2*(1-pnorm(abs((X-n1*m1/T)/sqrt(n1*n2*m1*m2/T/T/(T-1)))))
 RRL <- RR*exp(-qnorm(0.975)*sqrt(1/X-1/m1+1/Y-1/m2))
 RRU <- RR*exp(qnorm(0.975)*sqrt(1/X-1/m1+1/Y-1/m2))
 cat("リスク比の点推定量:",RR,"(p=",p.v,") 95%信頼区間=[",RRL,",",RRU,"]\n")
}
riskratio2(4,2,100000,100000)
```

[*23] 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。

結果は以下の通り。

```
        疾病あり 疾病なし   合計
曝露群         4    99996 1e+05
対照群         2    99998 1e+05
リスク比の点推定量: 2 (p= 0.4142103) 95%信頼区間=[ 0.3663344 , 10.91899 ]
```

ちなみに epitools ライブラリには riskratio() という関数があり，**先に非曝露，発症なしのデータを与える仕様なので注意が必要だが**，

```
library(epitools)
riskratio(c(99998,2,99996,4))
```

によって，曝露 2 (Exposed2) の行に点推定量 2 と 95%信頼区間 $(0.37, 10.9)$ が得られる。

また，率比については別に rateratio() という関数があって，分母を観察人年とした率比とその信頼区間を計算してくれる。信頼区間の計算は"midp"または"wald"または"boot"の 3 種類が指定できる。**非曝露群のデータを曝露群のデータより先に指定することに注意しなければならないが**，比較的使い方は簡単である。なお，この関数は，method="wald"オプションをつけないと，点推定量についても median-unbiased な推定値を計算するので，率比といっても単純な率の比とはやや異なる。簡単のため曝露群でも対照群でも白血病発症時点は観察終了直前だったとすれば，

```
library(epitools)
rateratio(c(2,4,5*100000,5*100000),method="wald")
```

により, 率比の点推定量は 2, 95%信頼区間は $(0.37, 10.9)$ が得られ, リスク比の値と一致する（ただし, median-unbiased な推定結果だと，これよりかなり幅が広くなる)。

次に**オッズ比の信頼区間**を考える。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%信頼区間は，

$$OR \cdot \exp(-\mathrm{qnorm}(0.975)\sqrt{1/a + 1/b + 1/c + 1/d}) \quad \text{（下限）}$$
$$OR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/a + 1/b + 1/c + 1/d}) \quad \text{（上限）}$$

となる。Cornfield の方法の方が大きなオッズ比については近似がよいが，手順がやや複雑であるため，ここでは扱わない。現在では Exact 法を用いることが推奨されているので，基本的に fisher.test() の結果を採用すればよい。R のコードは以下の通り。p 値は帰無仮説 $OR = 1$ の検定の有意確率である。

it10-5-2006.R

```
oddsratio2 <- function(a,b,c,d) {
 data <- matrix(c(a,b,a+b,c,d,c+d,a+c,b+d,a+b+c+d),nr=3)
 colnames(data) <- c("疾病あり","疾病なし","合計")
 rownames(data) <- c("曝露群","対照群","合計")
 print(data)
 OR <- (a*d)/(b*c)
 N1 <- a+c; M1 <- a+b; N0 <- b+d; M0 <- c+d; T <- a+b+c+d
 p.v <- 2*(1-pnorm(abs((a-N1*M1/T)/sqrt(N1*N0*M1*M0/T/T/(T-1)))))
 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("オッズ比の点推定量:",OR," (p=",p.v,") 95%信頼区間 = [",ORL,",",ORU,"]\n")
}
oddsratio2(4,2,99996,99998)
```

なお，サンプルサイズが大きいときは，R の fisher.test() 関数や，それを内部的に利用している epitools ライブラリの oddsratio.fisher() 関数[*24]では Out of workspace というエラーを起こして計算できないことがあるが，

[*24] oddsratio(...,method="fisher") でも呼び出される。epitools ライブラリの oddsratio() 関数には midp, fisher, wald, small と

vcd ライブラリの oddsratio() 関数は計算方法が異なるため実行できる。ただし，fisher.test() 関数の Out of workspace エラーはデフォルトで 20 万バイト確保されている計算用メモリでは不足したというエラーなので，呼び出すときに大きめの workspace を確保すれば回避可能である。これらを使って計算するためのコードを次の枠内に示す。

it10-6-2006.R

```
X <- matrix(c(4,2,99996,99998),nr=2)
fisher.test(X, workspace=10000000)
require(epitools)
oddsratio(c(4,99996,2,99998),method="fisher")
detach(package:epitools)
require(vcd)
OR <- oddsratio(X,log=F)
ORL <- summary(oddsratio(X))
ORCI <- confint(OR)
M <- c("オッズ比の点推定量","  (p=",")  95%信頼区間 = [ "," , "," ]\n")
cat(M[1],OR,M[2],ORL[1,4],M[3],ORCI[1],M[4],ORCI[2],M[5])
```

結果を次の表にまとめて示す。

			95%信頼区間	
方法	点推定量	有意確率	下限	上限
上で定義した oddsratio2()	2.00004	0.4142	0.366	10.9
fisher.test()	2.000022	0.6875	0.2866	22.11
epitools の oddsratio.fisher()	2.000022	0.6875(midp 0.453)	0.2866	22.11
vcd の oddsratio()	2.00004	0.1898	0.4262	9.386

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，$(\pi - \pi_2)/\pi$

ユールの 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。$Q = (OR - 1)/(OR + 1)$。独立な場合は 0 となる。

ファイ係数 (ϕ)　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，$\phi = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$。この値は 2×2 に限らず，一般の $k \times m$ の分割表について計算でき，ピアソンのカイ二乗統計量 χ_0^2 と総人数 n を用いて，$\sqrt{\chi_0^2/n}$ と定義される。k と m のどちらか小さな方の値が t だとすると，ファイ係数は 0 から $\sqrt{t-1}$ の範囲をとる。

ピアソンのコンティンジェンシー係数 C　ファイ係数はカテゴリ数の影響を受けるので，それを除去したものである。ファイ係数を用いて，$C = \sqrt{\phi^2/(1 + \phi^2)}$ として計算される。取りうる値の範囲は 0 から $\sqrt{(t-1)/t}$ である。

クラメールの V　ファイ係数を用いて，$V = \phi/\sqrt{t-1}$ と表せる。取りうる値の範囲は 0 から 1 となり，変数のカテゴリ数によらないのが利点である。

なお，ファイ係数，ピアソンのコンティンジェンシー係数，クラメールの V（これらは総称して属性相関係数と呼ばれることがある）は vcd ライブラリの assocstats() 関数で計算できる。この関数は，これらの係数の他，関連がないという仮説検定を実行してピアソンのカイ二乗統計量と尤度比カイ二乗統計量（ここでは説明しないが，多くの場合にピアソンのカイ二乗統計量を使った通常のカイ二乗検定よりもよいとされる），さらに，それらの有意確率を計算してくれる。属性相関係数はすべてピアソンのカイ二乗統計量に基づいて計算されるので，その有意性検定はカイ二乗検定の結果と等価と考えてよい。上記白血病のコホート研究の例でこれらを計算するには下枠内を打てばよいが，これら

いう 4 種類の method があり，それぞれ別々の関数を内部的に呼び出している（これは S3method の継承ではない）。riskratio() 関数とも rateratio() 関数とも引数を与える順序が異なるので注意されたい。この辺り，epitools ライブラリは若干思想が良くないと思う。

の係数の値はすべて 0.002 となるので，ほぼ関連はないと判定される。

```
require(vcd)
assocstats(matrix(c(4,2,99996,99998),nr=2))
```

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。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。下枠内は，２×２のクロス集計表を与えたときに，κ の点推定量と 95%信頼区間と有意確率を計算する R の関数を定義してから，○×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 人であったときにその計算を実行させる命令である。

it10-7-2006.R

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

vcd ライブラリの Kappa() 関数は $m \times m$ のクロス集計表について，重みなしと重みつきで κ 係数を計算してくれる[*25]。結果を confint() 関数に渡せば信頼区間も推定できる。同じデータに適用するには，下枠内のように打つ。上枠内の結果と同じ結果が得られる[*26]。

[*25] 重みは, Po や Pe を計算する際に weights=オプションを指定しないとき, あるいは weights="Equal-Spacing"にマッチしない任意の文字を指定した場合は, weights="Fleiss-Cohen"と指定したのと同じで, カテゴリ数が nc として 1-(abs(outer(1:nc,1:nc,"-"))/(nc-1))^2 となり，weights="Equal-Spacing"を指定したときは 1-abs(outer(1:nc,1:nc,"-"))/(nc-1) が重みとなる。つまり，○×の一致をみるならカテゴリ数は 2 なので，重みはどちらの方法でも matrix(c(1,0,0,1),nc=2) となる。

[*26] 有意確率は出ないが，κ 係数は，有意性の検定をするよりも，95 %信頼区間を示すことと，目安としての一致度の判定基準（負だと poor な一致，0-0.2 で slight な一致，0.21-0.4 で fair な一致，0.41-0.6 で moderate な一致，0.61-0.8 で substantial な一致，0.81-0.99 で almost perfect な一致，1 で perfect な一致とする，Landis and Koch, 1977, Biometrics, 33: 159-174 など）を参照して一致度を判定するとい

```
require(vcd)
print(myKappa <- Kappa(matrix(c(10,3,2,19),nr=2)))
confint(myKappa)
```

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。
`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 のコードを下枠内に示す。

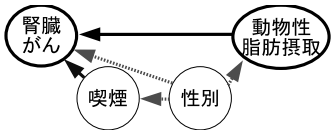
```
T <- X+Y+Z; VX <- var(X); VY <- var(Y); VZ <- var(Z); VT <- var(T)
alpha <- (3/2)*(1-(VX+VY+VZ)/VT)
print(alpha)
```

う使い方が普通らしく，`vcd` ライブラリでもそのような実装がされているのだと思われる。考えてみれば，一致度を評価する上で $\kappa = 0$ という帰無仮説の検定には意味が乏しいのは当然かもしれない。

交絡を考える

ここまでは（Cronbach の α を除けば），２つのカテゴリ変数の関係の分析であった。しかし，相関係数にも見かけの相関や擬似相関があったことを思い出してほしいが，クロス集計表でも見かけの関連が出てしまうことがある。大きな原因は，交絡因子があることである。

２つのカテゴリ変数の関係を調べているとき，それらの両方に影響を与えている変数があれば，それは交絡因子（交絡変数）である。交絡はカテゴリ変数に限らず，量的な変数にも起こるが，その制御については第 12 回に扱う。次の図に示す例では，動物性脂肪摂取と腎臓がん罹患の関係を調べようとしているのだが，性別が動物性脂肪摂取と腎臓がん罹患の両方に影響しているため，交絡を起こす可能性がある。喫煙は腎臓がん罹患には影響するが，動物性脂肪摂取と関係するとは思われないため，このフレームでは交絡ではない。



Rothman(2002)[*27]によると，交絡の条件は３つあって，

1. 交絡因子は疾病と関連していなくてはならない
2. 交絡因子は曝露と関連していなくてはならない
3. 交絡因子は曝露の効果であってはならない

が満たされていなくてはならない，とまとめられている。

シンプソンのパラドックス

第３の変数による交絡があるために，真の関係と見かけの関係が異なる事例として最も有名なものの一つが，シンプソンのパラドックスである。

it10-8.R

```
# 出典：Simpson EH (1951) The interpretation of interaction in
# contingency tables. J. Royal Stat. Soc. Ser. B, 13: 238-241.
males <- matrix(c(4,3,8,5),nr=2)
dimnames(males) <- list(c("生存","死亡"),c("処置なし","処置あり"))
females <- matrix(c(2,3,12,15),nr=2)
dimnames(females) <- list(c("生存","死亡"),c("処置なし","処置あり"))
total <- males+females
require(vcd)
prop.table(males,2)
summary(oddsratio(males))
prop.table(females,2)
summary(oddsratio(females))
prop.table(total,2)
summary(oddsratio(total))
```

上枠に示した Simpson の論文に載っている例では，男女別にみれば処置ありの方が生存割合が高いのに，男女をプールしてみると処置による生存割合の差が消失している。もっとも，消失してしまうといっても，サンプルサイズが小さいこともあって，統計的に有意とはいえない。しかし，現実にもこういうことは頻繁にあって，雑な解析では真の関連を見誤ってしまう危険がある。

例えば，下枠内は，スティーヴン・セン（松浦俊輔訳）『確率と統計のパラドックス』青土社の p.39 に掲載されている例を分析するプログラムだが[*28]，年齢をプールすると糖尿病の型と死亡率は独立でない（$p = 0.001$）のに，40 歳

[*27] Rothman, KJ (2002) Epidemiology: An Introduction. Oxford Univ. Press.（矢野栄二，橋本英樹監訳 (2004)「ロスマンの疫学：科学的思考への誘い」，篠原出版新社，として邦訳がでている）

[*28] 「生存」は，censored なので観察終了時までにイベントが起こっていないことを意味し，通常「打ち切り」と訳されるが，この本の訳文は「調査中」となっていて，大胆に意味を酌んでいいなら「生存」と訳してしまってもいいだろうと判断した。

以上と 40 歳未満で区切って層別に解析するとどちらの層でも独立性の帰無仮説は棄却されない（それぞれ $p = 0.27$, $p = 1$）。しかも，40 歳以上でも 40 歳未満でも IDDM 群の死亡率の方が NIDDM 群の死亡率より高いのに（40 歳以上では IDDM 群 0.46 に対して NIDDM 群 0.41，40 歳未満では IDDM 群 0.008 に対して NIDDM 群 0），年齢をプールすると IDDM 群の死亡率（0.29）よりも NIDDM 群の死亡率（0.40）の方が高くなる。これは年齢が交絡しているために，本来はない見かけ上の関連が見えてしまったことを意味する。40 歳未満群の大半が IDDM であって，かつ 40 歳未満群の死亡率が 40 歳以上群の死亡率より遥かに低いために，こうなったのである。

it10-9-2006.R(1)

```
over40 <- matrix(c(311,218,124,104),nc=2)
under40 <- matrix(c(15,0,129,1),nc=2)
dimnames(over40) <- list(c("生存","死亡"),c("NIDDM","IDDM"))
dimnames(under40) <- list(c("生存","死亡"),c("NIDDM","IDDM"))
print(over40)
fisher.test(over40)
print(under40)
fisher.test(under40)
total <- under40+over40
print(total)
fisher.test(total)
```

交絡を制御するには

このような交絡が起こらないようにするには，もちろん，臨床試験を実施する場合のように研究をデザインできる状況であれば，ランダム割付けを行ってフォローアップし，率比を分析するなど，**デザイン上で交絡を防ぐ工夫をすればよいし，それが王道である。**

しかし，実際問題として，研究は実験ばかりではないし，調査において完全に交絡を防ぐデザインをすることは，ほぼ不可能である。交絡があるかもしれないデータについて，それを制御して真の関連を検討するには，大まかにいって２つのアプローチがある。

１つは，交絡変数も原因となる変数とともに独立変数として投入し，それらの交互作用も考えながら，結果となる変数（多くは疾病発生）を従属変数として説明するようなモデルの当てはめを行う方法である。ロジスティック回帰分析を含むこの方法は，一般化線型モデルというフレームの中で扱え，第 12 回に説明する。

もう１つは，交絡因子によって層別解析を行うか，または限定を行うことである。層別解析とは，交絡因子のカテゴリによってデータを分割し，それぞれ別々の層として分析を進めることをさす。層別解析をした上で，どの層でも同じ向きに関連がありそうなら，たんにプールするのではなくて，「どの層でも同じ向きに関連がある」を対立仮説として，クロス集計表を併合した分析を行う。具体的な方法としては，マンテルヘンツェルの要約カイ二乗検定とか，共通オッズ比といったものが有名である[*29]。R では `mantelhaen.test()` 関数により，マンテルヘンツェルの要約カイ二乗統計量とその検定，さらに各層が２×２分割表のときは共通オッズ比とその 95%信頼区間を計算することができる。ここで得られる共通オッズ比は，層の違いを調整した関連の強さを示す指標となる。

ただし，マンテルヘンツェルの要約カイ二乗検定は３次の交互作用が存在しないこと（言い換えると，クロス表を作っている変数間の関連がどの層でも同じということ）を前提として行うものなので，それに先立って Woolf の検定（経験的ロジスティック変換を用いて，帰無仮説「どの層でも変数間の関連が共通」を検定する）によってそれを確認しておくべきとされる[*30]。Woolf の検定は `vcd` ライブラリの `woolf_test()` 関数で可能である。グラフ表示は `vcd` ライブラリの `fourfold()` 関数を用いるとよい。拡張モザイクプロットも `vcd` ライブラリの `mosaic()` 関数でできるが，引数を与える順序を変えねばならず，かつあまり見やすくないので，個人的にはお勧めしない。

先に作った２つの２×２クロス集計表 `under40` と `over40` は，次の枠内のコードのようにすれば３次元のクロス集計表 `x` にすることができる。

[*29] マンテルヘンツェルの方法による複数の層の関連の指標の併合は，オッズ比だけでなく，リスク差やユールの Q やファイ係数についても可能である。文献：佐藤俊哉，前田和甫 (1987) 疫学研究から得られる層別データの要約. 日本公衆衛生学雑誌, 34(5): 255-260.

[*30] なお，向きは異なるかもしれないがともかく何らかの関連があるかどうかを調べたい場合は，自由度 1 のカイ二乗分布する変数 k 個の和が自由度 k のカイ二乗分布に従うことを使って，各層で得られたカイ二乗統計量の総和を出してやれば検定できる。

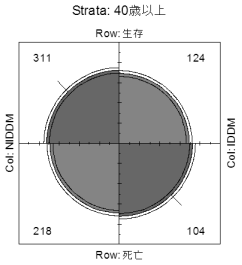
it10-9-2006.R(2)

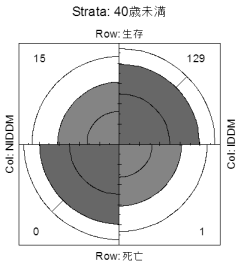
```
x <- array(c(over40,under40),dim=c(2,2,2))
dimnames(x) <- list(c("生存","死亡"),c("NIDDM","IDDM"),c("40 歳以上","40 歳未満"))
print(x)
```

または，いきなり数値を array() 関数に渡して次の枠内のように定義してもよい。Woolf の検定で有意でないので３次の交互作用はなく，マンテルヘンツェルの要約カイ二乗検定でも有意でないので，どの層でも同じ向きに関連があるとはいえないことがわかる。なお，fourfold() は grid グラフィックスを使うので layout() することができないようだ。

it10-10-2006.R

```
x <- array(c(311,218,124,104,15,0,129,1),dim=c(2,2,2))
dimnames(x) <- list(c("生存","死亡"),c("NIDDM","IDDM"),c("40 歳以上","40 歳未満"))
require(vcd)
woolf_test(x)
mantelhaen.test(x)
y <- array(c(x[1,1,],x[2,1,],x[1,2,],x[2,2,]),dim=c(2,2,2))
dimnames(y) <- list(c("40 歳以上","40 歳未満"),c("生存","死亡"),c("NIDDM","IDDM"))
structable(y,split_vertical=T)
fourfold(x)
mosaic(y)
```





課題

アルコール摂取と食道がんの症例対照研究を実施したとする。当初，食道がん患者 180 人と対照 575 人のサンプリングを行ったが，年齢が交絡している可能性を考え，年齢群別に集計したところ，55 歳未満では，患者群 46 人のうちアルコール多量摂取者が 30 人，対照群 372 人のうちアルコール多量摂取者が 64 人で，55 歳以上では，患者群 134 人のうちアルコール多量摂取者が 66 人，対照群 203 人のうちアルコール多量摂取者が 45 人だったとする。

このデータから，アルコール摂取と食道がんには関連があるか，あるとしたらどの程度の関連か評価せよ。まず年齢層別に２つのクロス集計表を作り，別々に分析してから，その結果から判断して，必要な場合には，どの層でも共通した関連があるかどうか，あるとすればどの程度の関連か評価せよ。

適切な作図も行って PowerPoint に貼り付け，統計処理の結果と学籍番号，氏名を記入して印刷し，氏名を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。