前回のQ＆A

Q1) Fisher の正確な確率の説明で 0 がでてくる場合がどうして 1 通りしかないのかわからなかった。

A1) すみません。0 が出てくるときは全部で 1 通りというわけではなくて，0 個を選ぶ組み合わせ，というのが，どれも選ばれないという 1 通りしかないということを言ったつもりでした。わかりにくかったと思うので，データ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きます。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します。いま仮に，下のようなクロス集計表が得られたとします。

	A あり	A なし	合計
B あり	4	3	7
B なし	1	7	8
合計	5	10	15

15 人のうち，5 人が要因 A をもっていて，7 人が要因 B をもっているとき[1] に，この表が得られる確率は，15 人のうち要因 A をもっている 5 人の内訳が，要因 B をもっている 7 人から 4 人と，要因 B をもっていない 8 人から 1 人になる確率となります。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，${}_7\mathrm{C}_4 \cdot {}_8\mathrm{C}_1 / {}_{15}\mathrm{C}_5 \simeq 0.0932$ です。

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 ということです。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，A の有無と B の有無は関係がないと判断していいことになります。

しかし，有意確率，つまり第一種の過誤を起こす確率は，A の有無と B 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りません。周辺度数が上の表と同じ表は，

(1)

	A あり	A なし
B あり	0	7
B なし	5	3
合計	5	10

(2)

	A あり	A なし
B あり	1	6
B なし	4	4
合計	5	10

(3)

	A あり	A なし	合計
B あり	2	5	7
B なし	3	5	8
合計	5	10	15

(4)

	A あり	A なし
B あり	3	4
B なし	2	6
合計	5	10

(5)

	A あり	A なし
B あり	4	3
B なし	1	7
合計	5	10

(6)

	A あり	A なし	合計
B あり	5	2	7
B なし	0	8	8
合計	5	10	15

の計６種類しかありません。(1) や (6) の表よりもさらに稀な場合を考えると，(1) の先は A も B もある人の数がマイナスになってしまいますし，(6) の先は A があって B がない人の数がマイナスになってしまいます。

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，(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 ${}_7\mathrm{C}_5 \cdot {}_8\mathrm{C}_0 / {}_{15}\mathrm{C}_5 \simeq 0.0070$ となります。[2]

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，$0.0932 + 0.0186 + 0.0070 = 0.1188$ となります。以上，フィッシャーの正確な確率を数値例で説明してみました。

Q2) 関連性の指標のオッズ比の話がわからなかった。

A2) 今回やります。

Q3) カイ二乗って一体何なんですか？

A3) 単なる記号の名前なので，講義で説明したような値をそう呼ぶのだと思ってください。統計学ではギリシャ文字と人名がよく使われるのですが，たいていの場合，それらの名前と，その名前が表す統計量や関数の間には，歴史的な意味しかありません。

Q4) （改善の要望）専門用語が突然出てくるとわからないことがある，プリントのどこを喋っているかわからないことがある，プリントを１回前に配って欲しい。

A4) 努力します。

---

[1] 「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」ともいいます。

[2] これらの確率をすべて足すと 1 になります。上の計算値として書いた値を使うと 0.9999 となりますが丸め誤差のせいで，厳密に計算すれば 1 になります。

統計学第７回　カテゴリ変数２つの分析（２）

（１）研究デザインとリスク，オッズ

- 前回も触れたが，研究デザインによって得られる関連性の指標は異なる。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の，発症しなかった人に対する比である。
- 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[3] ）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[4] 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），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）[5] とか断面研究（Cross Sectional Study）[6] 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，曝露群の病気の人の病気でない人に対する比が，対照群に比べてどれくらい大きいかを示す値だからである[7] 。
- 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。[8]
- 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。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年間の白血病の発症率が調査されたことがある。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

[3] 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。

[4] 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。

[5] 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（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。

[6] 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。

[7] 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。実は，病気群で曝露した人の曝露していない人に対する比が，病気なし群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。

[8] 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を表示することが多い。

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った。

- 仮に[9] 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。[10]
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[11] 。

	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，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。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。[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。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，ピアソンの相関係数，ファイ係数といったものがある（後述）。[13]
- 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 係数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

[9] 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に環境研と癌センターの研究チームが調べていたはずだが，その結果がどうなったのかは知らない。

[10] 上述のように最尤推定された条件付きオッズ比は，R の `fisher.test(matrix(c(4,2,9996,9998), nc=2))` で計算すると，2.000322 である。

[11] くどいようだが，あくまで架空のデータである。

[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。

[13] 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。

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。

- 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。
- 対数変換の場合は，95%信頼区間の下限は $RR \cdot \exp(-\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2})$，上限が $RR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2})$ 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，対数変換か Cornfield の方法（４次方程式の解を求めねばならないので手計算は不可能）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 $OR \cdot \exp(-qnorm(0.975)\sqrt{1/a + 1/b + 1/c + 1/d})$，上限は $OR \cdot \exp(qnorm(0.975)\sqrt{1/a + 1/b + 1/c + 1/d})$ 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。[14] なお，Cornfield の方法は煩雑なので省略する。

（２）その他の関連性の指標

- 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。$RD = \pi_1 - \pi_2$
- 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。$RelD = (\pi_1 - \pi_2)/(1 - \pi_2)$
- 曝露寄与率：真に要因の影響によって発症した者の割合。$AFe = (\pi_1 - \pi_2)/\pi_1$
- 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，$AFp = (\pi - \pi_2)/\pi$
- 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 で表した場合の相関係数[15] 。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，$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$

（３）一致度の指標

- 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。

	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 $1-\mathtt{pnorm}(\kappa/\sqrt{V(\kappa)})$ 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，$\kappa \pm \mathtt{qnorm(0.975)} \cdot \sqrt{P_o \cdot (1 - P_o)/(N \cdot (1 - P_e)^2)}$ として計算できる。
- 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。

（４）付録

- table 型の object である X を与えることで今回示した指標をすべて計算する関数，crosstab(X) を定義してみた。l7-1.R としてダウンロードできる。R の組み込みデータの infert（スイスの女性で不妊がらみのデータ）を使って，既往出生児数が２人以上かどうかと自然流産の経験の有無の関連をみた使用例もつけてみたので，参考にされたい。

[14] R 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，単純な計算よりも幅が広くなる。

[15] 相関係数とは，第 11 回で触れる予定だが，-1 から 1 までの値をとる量で，まったく関連がない場合に 0 となり，完全に一致するとき 1 となる。