第 6 章

カテゴリ変数 2 つの分析（1）

6.1　2 つのカテゴリ変数を分析する 2 つのアプローチ

前章では，1 つのカテゴリ変数のもつ情報から母比率を推定したり，期待される母比率と一致するかどうかを検定する方法を示した。本章では，2 つのカテゴリ変数を分析する方法を示す。

2 つのカテゴリ変数を分析するには，2 つのアプローチがある。1 つは，2 つの変数についての母比率に差があるかどうかを調べるアプローチであり，もう 1 つは，2 つの変数の関係を調べるアプローチである。後者を調べる際には，クロス集計表を作るのが普通である。その上で，2 つの変数の独立性を検定したり，関連の程度を調べたりする。[*1]

6.2　2 つのカテゴリ変数の母比率の差の検定と信頼区間

前章で説明したように，個々のカテゴリ変数がもつ情報はデータ数（標本サイズ）と，各カテゴリの割合である。そこから，各カテゴリの母集団における割合（母比率）を推定することができる。

2 つのカテゴリ変数の母比率 p_1, p_2 が，各々の標本比率 $\hat{p_1} = r_1/n_1, \hat{p_2} = r_2/n_2$ として推定されるとき，それらの差を考える。差 $(\hat{p_1} - \hat{p_2})$ の平均と分散は，$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2, V(\hat{p_1} - \hat{p_2}) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ となる。2 つの母比率に差が無いならば，$p_1 = p_2 = p$ とおけるはずなので，$V(\hat{p_1} - \hat{p_2}) = p(1-p)(1/n_1 + 1/n_2)$ となる。この p の推定値として，$\hat{p} = (r_1 + r_2)/(n_1 + n_2)$ を使い，$\hat{q} = 1 - \hat{p}$ とおけ

[*1] ただし母比率の差の検定は，後で述べるように，2 × 2 のクロス集計表とみなして独立性の検定をすることと数学的に等価である。

ば，n_1p_1 と n_2p_2 がともに5より大きければ，標準化して正規近似を使い，

$$Z = \frac{\hat{p_1} - \hat{p_2} - E(\hat{p_1} - \hat{p_2})}{\sqrt{V(\hat{p_1} - \hat{p_2})}} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} \sim N(0,1)$$

によって[*2]検定できる。

例をあげよう。2002年6月1日に山口県立大学の2つのキャンパスを隔てるバイパスで，交通の様子を観察したデータを考える[*3]。1人で観察する場合，観察対象は車，歩行者などが考えられるが，ここでは車とする。車1台から得られるいくつかの特性を1組のデータとして扱う（こういう1組を1つの「オブザーベーション(observation)」と呼ぶ)。簡単に捉えられる特性としては，進行方向，車の種類（普通乗用車かそれ以外か)，車の色，といったものが考えられる。データ解析を考える上では，これらの特性が「変数」となる。つまり，生データを表の形にまとめると，

	変数		
オブザーベーション番号	進行方向	車の種類	車の色
1	津和野方面	乗用車	白
2	山口市街地	乗用車	白
3	山口市街地	乗用車	銀
⋮			

これを数値としてコーディングするときは，典型的なカテゴリを1にするとわかりやすい。進行方向という変数の変数名をDest（値は，津和野方面を1，山口市街地方面を2）とし，車の種類の変数名をType（乗用車が1，トラックなどそれ以外のものを2)，車の色の変数名をColor（1が白，2が黒，3はそれ以外）とすれば，次の表のようにコード化される。

[*2] この Z は離散値しかとれないため，連続分布である正規分布による近似の精度を上げるために，連続性の補正と呼ばれる操作を加え，かつ $p_1 > p_2$ の場合（つまり $Z > 0$ の場合）と $p_1 < p_2$ の場合（つまり $Z < 0$ の場合）と両方考える（両側検定という）のだが，正規分布は原点について対称なので，絶対値をとって $Z > 0$ の場合だけ考え，有意確率を2倍すればよい（逆に5%水準で検定したいなら，97.5%点より Z が大きいかどうかを見ればよい)。すなわち，

$$Z = \frac{|\hat{p_1} - \hat{p_2}| - (1/n_1 + 1/n_2)/2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

として，この Z の値が標準正規分布の97.5%点（Rならば `qnorm(0.975,0,1)`）より大きければ帰無仮説を棄却するのが普通である。

[*3] `http://phi.ypu.jp/statlib/tf.mpg` としてMPEG1形式のムービーファイルを公開している。

	変数		
Obs	Dest	Type	Color
1	1	1	1
2	2	1	1
3	2	1	3
⋮			

これを表計算ソフト（Excel や StarSuite/OpenOffice.org の scalc など）やテキストエディタで入力し，CSV（コンマ区切り値）形式のファイル `L6-1.csv` として R の作業ディレクトリに保存すれば，R のコンソールで `x <- read.csv("L6-1.csv")` として，`x` というデータフレームに読み込めるし，タブ区切りテキスト形式のファイル `L6-1.dat` として保存すれば，`x <- read.delim("L6-1.dat")` として読み込める[*4]。R では，各変数はデータフレーム名$変数名として参照できるので，たとえば進行方向別の頻度を出したいときは，`table(x$Dest)` とすれば良い。総観察数 89 台のうち，津和野方面が 60 台，山口市街地方面が 29 台であったことがわかる。

ここで，進行方向によって乗用車割合が異なるかという仮説を考えてみる。帰無仮説は，「進行方向が反対でも乗用車割合には差が無い」ということになる。

`table(x$Type[x$Dest==1])` とすれば，津和野方面の乗用車が 60 台中 57 台であることがわかり，`table(x$Type[x$Dest==2])` とすれば，山口市街地方面の乗用車が 29 台中 25 台であったことがわかる。

上で説明した式にあてはめて計算すると，

$$\hat{p} = (57+25)/(60+29) = 0.92\ldots$$
$$\hat{q} = 1-\hat{p} = 0.079\ldots$$
$$Z = \frac{|0.95-0.86| - (1/60+1/29)/2}{\sqrt{0.92\cdot 0.079\cdot(1/60+1/29)}} = 1.024$$

となるので，標準正規分布の 97.5% 点である 1.96 よりずっと小さく，5% 水準で有意ではない。つまり帰無仮説は棄却されず，差はないと考えてよい（囲み解説『過誤』を見よ）。

[*4] これらのファイルは `http://phi.ypu.jp/statlib/L6-1.csv` などとしてダウンロードできる。

過誤

厳密に言えば，差がないとしたときに偶然この値以上の値が得られる確率が5% よりずっと多い，ということである。この確率がいくらかといえば，R で，`2*(1-pnorm(1.024,0,1))` とすれば，0.305... という値が得られるので，約 31% である。ついでに書いておくと，有意確率は，それに従って帰無仮説を棄却した場合にその判断が誤りであった（=実は差がなかった）確率なので，第1種の過誤（α-Error）とも呼ばれる。反対に，検定の検出力が足りなくて本当は差があるのに差がないと判断してしまう確率を第2種の過誤（β-Error）と呼ぶ。第2種の過誤は標本サイズに依存する。

差の 95% 信頼区間を出すことも簡単である。信頼区間を出すには，標本サイズが大きければ正規分布を仮定できるので，原則どおりに差から分散の平方根の 1.96 倍を引いた値を下限，足した値を上限とすればよい。上の例では，$\hat{p_1}-\hat{p_2}=0.0879\ldots$，$V(\hat{p_1}-\hat{p_2})=\hat{p_1}(1-\hat{p_1})/n_1+\hat{p_2}(1-\hat{p_2})/n_2=57/60(1-57/60)/60+(25/29(1-25/29)/29=0.00489\ldots$ となるので，信頼区間の下限は $0.0879-1.96*\sqrt{0.00489}=-0.049$，上限は $0.0879+1.96*\sqrt{0.00489}=0.225$ となる。しかし，通常は連続性の補正を行うので，下限からはさらに $(1/n_1+1/n_2)/2=(1/60+1/29)/2=0.0255\ldots$ を引き，上限には同じ値を加えて，95% 信頼区間は $(-0.0747, 0.251)$ となる。実は R では，

```
type1 <- c(57,25)
total <- c(60,29)
prop.test(type1,total)
```

とすれば各々の母比率の推定と，その差があるかどうかの検定（連続性の補正済み，ただし正規近似そのままではなく，カイ二乗分布で検定したものだが，数学的にはまったく同値である），差の 95% 信頼区間を一気に出してくれる。$p=0.3057$ より有意な差は無く，95% 信頼区間は $(-0.0747, 0.251)$ であることがわかる。

6.3 2つのカテゴリ変数の関係を調べることと研究のデザイン

こんどは，2つの変数の関係を調べるアプローチについて説明する。関係を調べるといっても，研究デザインによって，検討すべき関係の種類はさまざまである。たとえば，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＝ケースコントロール研究」[*5]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。もちろん，それらが独立であるかどうか（関連がないかどうか）を検討することもできる。

しかし，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[*6]，断面研究で得られた2つの変数には時間的な前後関係がないので，独立性の検定を行ったり，リスク比やオッズ比以外の関連性の指標を計算することが多い[*7]。関連性の指標については次章で詳しく説明することにして，本章の後半では，独立性の検定について説明する。

6.4 クロス集計とは？

2つのカテゴリ変数の間に関係があるかどうかを検討したいとき，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。

とくに，2つのカテゴリ変数が，ともに2分変数のとき，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。

6.5 独立性の検定の原理

独立性の検定は，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。

---

*5 詳しくは疫学の教科書を参照されたい。

*6 それらの値は次章で説明するリスク比やオッズ比という指標で表され，疫学研究上非常に重要である。

*7 ただし，オッズ比は断面研究でも計算できる。

	特性 A あり	特性 A なし
特性 B あり	a 人	b 人
特性 B なし	c 人	d 人

標本が，上記の表のような度数をもっているとき，母集団の確率構造が，

	特性 A あり	特性 A なし
特性 B あり	π_{11}	π_{12}
特性 B なし	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，

	特性 A あり	特性 A なし
特性 B あり	$N\pi_{11}$	$N\pi_{12}$
特性 B なし	$N\pi_{21}$	$N\pi_{22}$

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が，普通は π が未知なので，$p(A\cap B)=p(A)p(B)$ と考えて，各々の変数については特性のある人とない人の人数が決まっている（周辺度数が固定している）と考え，$p(A)$ の推定値 $(a+c)/N$ と $p(B)$ の推定値 $(a+b)/N$ の積として π_{11} を，$p(\bar{A})$ の推定値 $(b+d)/N$ と $p(B)$ の推定値 $(a+b)/N$ の積として π_{12} を，$p(A)$ の推定値 $(a+c)/N$ と $p(\bar{B})$ の推定値 $(c+d)/N$ の積として π_{21} を，$p(\bar{A})$ の推定値 $(b+d)/N$ と $p(\bar{B})$ の推定値 $(c+d)/N$ の積として π_{22} を推定すれば，

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

となる。この場合は，母数を2つ推定したので，自由度1のカイ二乗分布に従うと考えて検定できる。

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，$a=12, b=8, c=9, d=10$ などとわかっているときは，`x <- matrix(c(12,8,9,10),nc=2)` として表を与え，`chisq.test(x)` とするだけでもできる（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）。各度数が未知で，各個人についてのカテゴリ変数AとBの生の値が与えられているときも，Rを使うと，`chisq.test(A,B)` で計算できる。クロス集計表を作るには，`table(A,B)` とする。もちろん，`chisq.test(table(A,B))` でもカイ二乗検定ができる。

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1

上の交通量調査データで独立性のカイ二乗検定をせよ。

帰無仮説は，進行方向と車の種類が独立（無関係）ということである。クロス集計表を作ってみると，

	乗用車	それ以外	合計
津和野方面	57	3	60
山口市街地方面	25	4	29
合計	82	7	89

となる。進行方向と車の種類が無関係であった場合に期待される度数は，

	乗用車	それ以外	合計
津和野方面	60*82/89	60*7/89	60
山口市街地方面	29*82/89	29*7/89	29
合計	82	7	89

となる。これから定義の通りに計算してもいいが，連続性の補正も考えると公式に代入するのが現実的である[*8]。実際に代入してみると，連続性の補正済みのカイ二乗統計量は $\chi_c^2 = 89*(|57*4-3*25|-89/2)^2/(60*29*82*7) = 1.049$ となる。自由度1のカイ二乗分布で分布関数の値を1から引くと，$p=0.3057\ldots$ となり，有意確率が約31%である（つまり帰無仮説は棄却されず，独立である可能性が十分にある）ことがわかる[*9]。ただし，Rで実行した結果に警告メッセージが出ていることか

[*8] もちろん，Rで `chisq.test(matrix(c(57,25,3,4),nc=2))` とするのが手軽である。

[*9] この値が母比率の差の検定の有意確率と一致していることに注意されたい。

らもわかるが，この例ではサイズの小さなセルがあるので，カイ二乗検定における正規近似は適当でない可能性があり（一般に期待度数が5以下のセルが全体の20%以上あるときはカイ二乗検定は適当でないとされる），次に説明するフィッシャーの直接確率を使った方がよい。

6.6 フィッシャーの直接確率（正確な確率）

周辺度数を固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）を直接計算し，与えられた表が得られる確率よりも低い確率になる場合をすべて足し合わせたもの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。

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，そのうち変数 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[*10]。

フィッシャーの正確な確率検定は，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。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，標本サイズがよほど大きくない限り，常にFisherの正確な確率を求めるべきである。

例題2

上の交通量調査データでフィッシャーの直接確率を計算せよ。

`fisher.test(matrix(c(57,25,3,4),nc=2))` をRで計算させると，0.2089となる。カイ二乗検定の場合よりも小さな有意確率が得られたことに注意されたい（一般に第1種の過誤をしにくい）。

[*10] 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。

わかりにくいと思うので，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。いま仮に，下のようなクロス集計表が得られたとする。

	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 をもっているときに[*11]，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，${}_7C_4 \cdot {}_8C_1 / {}_{15}C_5 \simeq 0.0932$ 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，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る。

しかし，有意確率，つまり第 1 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らない。周辺度数が上の表と同じ表は，

(1)	A あり	A なし	(2)	A あり	A なし	(3)	A あり	A なし	合計
B あり	0	7		1	6		2	5	7
B なし	5	3		4	4		3	5	8
合計	5	10		5	10		5	10	15

(4)	A あり	A なし	(5)	A あり	A なし	(6)	A あり	A なし	合計
B あり	3	4		4	3		5	2	7
B 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

[*11] 「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。

A も B もある人の数がマイナスになってしまうし，(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[*12]，(1) は ${}_7C_0 \cdot {}_8C_5 / {}_{15}C_5 \simeq 0.0186$，(2) は ${}_7C_1 \cdot {}_8C_4 / {}_{15}C_5 \simeq 0.1632$，(3) は ${}_7C_2 \cdot {}_8C_3 / {}_{15}C_5 \simeq 0.3916$，(4) は ${}_7C_3 \cdot {}_8C_2 / {}_{15}C_5 \simeq 0.3263$，(5) は上で計算した通り ${}_7C_4 \cdot {}_8C_1 / {}_{15}C_5 \simeq 0.0932$，(6) は ${}_7C_5 \cdot {}_8C_0 / {}_{15}C_5 \simeq 0.0070$ となる[*13]。

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。

[*12] R で組み合わせ計算を行う関数は choose() である。たとえば ${}_7C_3$ は，choose(7,3) で計算できる。

[*13] 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。