前回のQ＆A＋成績評価について

Q1) (480-463)^ 2/463... の^ って何ですか？

A1) これは累乗するという意味の記号です。記号の意味を説明し忘れることがときどきあるので，そういうときは途中で遮って構わないので指摘してください。

Q2) 計算を省略しないでください。

A2) あまり計算手順の説明に時間をかけるわけにもいかないので，細かいところは省略せざるを得ません。悪しからず。

成績評価 方法を決定しました。筆記試験（概ね穴埋めか選択式にする予定）を 8 月 5 日（月）に実施します。統計学は覚えているかどうかよりも理解しているかどうか，使えるかどうかが勝負なので，ノートや配布資料を持ち込み可とします。計算能力をみたいわけではないので，できれば平方根の計算ができる電卓を持参して欲しいのですが（大内のナフコで 300 円くらいで売っています），電卓を持っていない人が多ければ何か方法を考えます。試験は 100 点満点とし，出席 1 回を 2 点または 3 点（どちらにするかは検討中）として加算します。ただし，100 点を超える人は 100 点になります。追試は行いません。

統計学第６回　カテゴリ変数２つの分析（１）

（１）２つのカテゴリ変数を分析する２つのアプローチ

- 前回は，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。今回は，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]

（２）２つのカテゴリ変数の母比率の差の検定と信頼区間

- 前回説明したように，個々のカテゴリ変数がもつ情報はデータ数（標本数）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 $(\hat{p_1} - \hat{p_2})$ の平均値と分散は，$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2, V(\hat{p_1} - \hat{p_2}) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ 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，$\hat{q} = 1 - \hat{p}$ とおけば，$n_1 p_1$ と $n_2 p_2$ 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（映像を参照）。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，

---

[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。

[2] このままの Z では正規分布から若干ずれるので，それを修正する（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）。即ち，

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

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。

	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，下表のようになる。

	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 など）で入力し，CSV 形式か TAB 区切りテキストで保存すれば，R のコンソールで `x _ read.csv("d:/work/LECTURE/ypu/statistics/L6-1.dat")` などとして読み込める[3] 。R では，各変数はデータフレーム名$変数名として参照できるので，例えば進行方向別の頻度を出したいときは，`table(x$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])` などとすれば，津和野方面の乗用車割合は $57/60 = 0.95$，山口市街地方面の乗用車割合は $25/29 = 0.86...$ と計算できる。
- 上で説明した式にあてはめて計算すると，$\hat{p} = (57+25)/(60+29) = 0.92...$，$\hat{q} = 1-\hat{p} = 0.079...$，$Z = (|0.95-0.86| - (1/60+1/29)/2)/\sqrt{0.92 \cdot 0.079 \cdot (1/60+1/29)} = 1.024$ 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[4] 。
- 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，標本数が大きければ，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，$\hat{p_1}-\hat{p_2} = 0.0879...$，$V(\hat{p_1}-\hat{p_2}) = \hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2 = 57/60(1-57/60)/60 + (25/29(1-25/29)/29 = 0.00489...$ 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 $(1/n_1 + 1/n_2)/2 = (1/60+1/29)/2 = 0.0255...$ 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。

（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。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＝

[3] 前回まで代入記号を < − で示していたが，− と紛らわしいことと，_（アンダースコアと呼ぶ）でもまったく同じように代入を意味する記号として動作するので，今回から記号を変える。

[4] 厳密に言えば，差がないとしたときに偶然こ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，有意確率とは，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，第一種の過誤（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二種の過誤（β-Error）と呼ぶ。第二種の過誤は標本数に依存する。

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「全然吸わない」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（それらの値は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），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。関連性の指標については次回詳しく説明することにして，今回の講義の後半では，独立性の検定について説明する。

（4）クロス集計とは？

- ２つのカテゴリカル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つのカテゴリカル変数が，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。

（5）独立性の検定の原理

- 独立性の検定は，２つのカテゴリカル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つのカテゴリカル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。

	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。

[5] この場合もそうだが，患者対照研究は多くの場合ある時点での１回の調査によって行われる。ある時点での１回の調査によって行われる研究を横断的研究とか断面研究という。この辺りは疫学の講義でもう少し詳しく触れられるであろう。

- 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カル変数 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)` で計算させる。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))` としてもよい。
- R では，`chisq.test` 関数の中で，`simulate.p.value` 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 上の交通量調査データで独立性のカイ二乗検定をすることを考えてみる。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。[6] 連続修正済みのカイ二乗統計量は $\chi_c^2 = 89*(|57*4-3*25|-89/2)^2/(60*29*82*7) = 1.049$ となり，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... 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。[7]

- 周辺度数を固定して，すべての組み合わせを考えて，それらが起こる確率（超幾何分布に従う）を直接計算し，与えられた表よりも偏った表になる確率（偏っているかどうかは，それぞれの表のカイ二乗値を連続修正なしで計算し，大きければ偏っていると判断する）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，全標本数 N の有限母集団があって，そのうち変数 A のカテゴリが 1 である標本が m_1，1 でない標本が m_2 あるときに，変数 B のカテゴリが 1 である標本が n_1 個（1 でない標本が $n_2=N-n_1$ 個）あるとき，そのうち変数 A のカテゴリが 1 である標本がちょうど a 個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。[8]
- 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，標本数がよほど多くない限り，常に Fisher の正確な確率を求めるべきである。

例題 上の交通量調査データで計算すると，`fisher.test(matrix(c(57,25,3,4),nc=2))=0.2089` となる。カイ二乗検定の場合よりも小さな[9] 有意確率が得られたことに注意（一般に第一種の過誤をしにくい）。

---

[6] 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，連続性の修正も考えると公式に代入するのが現実的である。

[7] 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。

[8] 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，$E(a)=n_1m_1/N$, $V(a)=\{(N-n_1)/(N-1)\}n_1(m_1/N)(m_2/N)=(m_1m_2n_1n_2)/\{N^2(N-1)\}$ 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，一般的ではない。

[9] 講義時配布資料では，ここが「大きな」となっていたが，言うまでも無く typo である