１０章 質的データの検定

この章では質的データについて、その検定手法を見て行きます。基本的に検定はχ^2統計量を用いたχ^2検定を用いますが、特別な場合として直感的に行える比率を用いた検定方法も付け加えておこうと思います。

前章の終わりに述べた通り、ここからは辞書的な使い方をすることも想定しています。そのため一貫して、最初に検定の「例」、次にそれを解くための「理論」、その理論を用いた「解答」という順番で書いて行きます。「理論」の部分については、後で見たとき一目で分かるように、要点だけ列挙します。その後に、「解説」として検定についての一般的な説明をします。数学の苦手な人はここまで読めば十分です。

節の最後に「数学的解説」として、簡単に示せる範囲で理論の成り立ちを紹介する場合があります。数学に興味のある人が読んで下さい。数式を追っていくと、文系大学生としては十分な内容になっていると思います。

10.1 母集団の比率と指定比率との検定

ここでは、9.1節で例として与えた超能力の検定を一般化して話をします。適合度検定とはいくつかの事象のそれぞれの出現比率が、想定した比率と異なっているかどうかを見分ける検定方法です。コインの表裏やサイコロの目の出現比率のように分かり易いものから、出現比率がある確率分布に従っているかどうかなどの少し難しいものまで、利用範囲の広い検定方法です。以下の例から始めましょう。

例

ある町で1年間に発生した交通事故の件数を平日の曜日ごとに調べたところ、以下の表が得られた。事故には曜日によるばらつきがある（一様でない）といえるか？有意水準5%で判定せよ。

曜日	月	火	水	木	金	計
事故件数	16	14	16	11	23	80

理論

n回の観測の中で、事象1はn_1回、事象2はn_2回、・・・、事象kはn_k回起こるとする。出現比率は想定比率 $p_1, p_2, \cdots, p_k$ に比べて差があるといえるか。出現の想定値を $m_1, m_2, \cdots, m_k$ （$m_i = np_i$）として、$\alpha \times 100\%$の有意水準で判定する。

帰無仮説 H_0：事象 i の出現比率は p_i　　（想定比率と比べて差がない）

対立仮説 H_1：H_0 でない　　（想定比率と比べて差がある）

H_0のもとで　　$\chi^2 = \dfrac{(n_1 - m_1)^2}{m_1} + \dfrac{(n_2 - m_2)^2}{m_2} + \cdots + \dfrac{(n_k - m_k)^2}{m_k} \underset{n\to\infty}{\sim} \chi^2_{k-1}$分布　(10.1)

$p = chidist(\chi^2, k-1)$として、 $p < \alpha$ のとき、H_0を棄却してH_1を採択する。

解答

この場合事象は「月曜に発生」「火曜に発生」・・・と考えられますので、事象の数は5となり、$k=5$ です。帰無仮説を仮定すると一様な出現比率ですから、想定比率は $p_1=p_2=\cdots=p_5=1/5$ となります。これより出現の想定値は以下のようになります。

$$m_1=m_2=\cdots=m_5=80/5=16$$

この数値を用いると、χ^2 値は以下となります。

$$\chi^2=\frac{(23-16)^2}{16}+\frac{(14-16)^2}{16}+\frac{(16-16)^2}{16}+\frac{(11-16)^2}{16}+\frac{(16-16)^2}{16}$$

$$=\frac{78}{16}=4.875$$

自由度は 5-1=4 となり、検定確率値は *chidist*() 関数を使うと以下となります。

$$p=chidist(4.875,4)=0.300365\cong 0.300$$

これより $p>0.05$ ですから、指定比率と比べて差がある（一様でない）といえないという結論になります。

解説

事象の出現比率を指定比率と比較する検定を適合度検定といいます。検定統計量の式で「$\underset{n\to\infty}{\sim}\chi^2_{k-1}$ 分布」としましたが、これはデータ数が十分大きくなると自由度 $k-1$ の χ^2 分布に従う、と解釈します。そのためこの検定を利用するには、ある程度のデータ数が必要になります。具体的には、各事象の出現数が大体 10 以上と考えておけばよいでしょう。

実現値と想定値を用いて χ^2 値を計算してみると、$\chi^2=4.875$ という値が求まります。自由度を $5-1=4$ として、Excel の χ^2 分布の確率を求める関数 *chidist*() を用いて、上側確率 $p=0.300365\cong 0.300$ を得ます。近似式としての誤差がありますので、確率 p の小数点以下の桁数は 3 桁程度にしましたが、読者の方が計算されるとき、値がはっきり分る方が安心感を与えると考え、Excel で計算した値も標準的な桁数で表示することにしました。報告書などに書かれるときは、データの有効桁数も考慮して、小数点以下 3~4 桁として四捨五入するのがよいと思います。また、χ^2 値などの検定統計量を関数に代入する場合、四捨五入した値を用いず、セルを参照する形で代入することにしました。即ち、小数の計算は Excel の最大桁数で行っています。これも結果の値に微妙な差が出ないようにするためです。以後も同様に表示させてもらいますので、ご了承下さい。

検定結果の判定の部分について、この教科書では Excel の関数をそのまま用いた

$p=chidist(\chi^2,k-1)$ として、 $p<\alpha$ のとき

という形式にしています。一般の教科書では、「$\chi^2_{k-1}(p)=\chi^2$ として、 $p<\alpha$ のとき、H_0 を棄却し、H_1 を採択する。」とか「$\chi^2>\chi^2_{k-1}(\alpha)$ のとき、H_0 を棄却し、H_1 を採択する。」のような形式に書かれていると思います。検定統計量が正規分布の場合は、「$Z(p/2)=|Z|$ として、 $p<\alpha$ のとき」または「$|Z|>Z(\alpha/2)$ のとき」となり、F 分布の場合は、「$F_{n_1,n_2}(p)=F$ として、 $p<\alpha$ のとき」または「$F>F_{n_1,n_2}(\alpha)$ のとき」となります。また t

分布の場合は、「$t_n(p/2)=|t|$として、 $p<\alpha$ のとき」または「$|t|>t_n(\alpha/2)$ のとき」になります。今後は理論の部分に Excel を利用した形式しか書きませんので、一般の教科書と比較する際はここの記述を参考にして下さい。

数学的解説　[Skip OK]

さて、この理論の中心的な統計量 χ^2 が自由度 $k-1$ の χ^2 分布に従う理由を考えてみましょう。ここでは簡単のため事象をはい・いいえと答える 2 つの場合（$k=2$）と考えてみます。はいと答える確率を p_0 とすると、いいえと答える確率は $1-p_0$ となります。ここでは検定確率値 p と区別するために、指定比率（確率）を p_0 としています。そのとき統計量 χ^2 はどうなるでしょうか。 X をはいと答える度数とすると以下のようになります。

$$\chi^2=\frac{(X-np_0)^2}{np_0}+\frac{[(n-X)-n(1-p_0)]^2}{n(1-p_0)}$$

$$=\frac{(X-np_0)^2}{np_0}+\frac{(X-np_0)^2}{n(1-p_0)}=\frac{(X-np_0)^2}{np_0(1-p_0)}$$

さて、この中で用いた確率変数 X が従う分布は何でしょうか。これは 6.4 節で紹介した 2 項分布と呼ばれる分布です。2 項分布に従う確率変数 X は、十分大きな度数で、近似的に平均 np_0、分散 $np_0(1-p_0)$ の正規分布に従うことが知られています。即ち、

$$Z=\frac{X-np_0}{\sqrt{np_0(1-p_0)}} \underset{n\to\infty}{\sim} N(0,1) \text{ 分布}$$

です。また、8.1 節で述べたように、$Z\sim N(0,1)$ 分布の場合、$Z^2\sim\chi_1^2$ 分布であることも知られています。ここで上に求めた統計量 χ^2 の式を見ると、丁度 Z^2 に一致しており、χ^2 は χ_1^2 分布に従うことが分ります。

さて、上の式を少し書き換えてみます。

$$Z=\frac{X-np_0}{\sqrt{np_0(1-p_0)}}=\frac{n(X/n-p_0)}{\sqrt{np_0(1-p_0)}}=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \underset{n\to\infty}{\sim} N(0,1) \text{ 分布}$$

これは、実測値 X から実測比率 $\hat{p}=X/n$ を用いた式に変わっています。この式を使うと、比率を意識して検定を行うことができます。ただし、この方法だと事象が 3 つ以上のときには使えません。

有意水準を $\alpha\times100\%$ として両側検定の場合、Z の値から以下のように確率 p を求めます。

$$p/2=1-normsdist(|Z|)$$

もし $p<\alpha$ であれば、帰無仮説を棄却して、対立仮説を採用します。

2 項分布は確率 p_0 を有限にして、度数を十分大きくすると正規分布に近づきますが、余り大きくない度数では正規分布からずれています。そのため上の Z の式では不十分で、以下のような補正項を含んだ式を使うのが良いとされています。

$$Z = \frac{|\hat{p} - p_0| - 1/2n}{\sqrt{\frac{p_0(1-p_0)}{n}}} \underset{n\to\infty}{\sim} N(0,1) \text{ 分布}$$

これはイェーツ（Yates）の補正と呼ばれています。この補正項を含む式を度数で表わした式に書き換えると以下のようになります。

$$Z^2 = \frac{(|n_1 - m_1| - 1/2)^2}{m_1} + \frac{(|n_2 - m_2| - 1/2)^2}{m_2} \underset{n\to\infty}{\sim} \chi_1^2 \text{ 分布}$$

但し、$n_1 = n\hat{p}$，$n_2 = n(1-\hat{p})$，$m_1 = np_0$，$m_2 = n(1-p_0)$ としています。

一般に事象の数が k の場合、理論のところで与えた式に補正項を加えると以下のようになります。

$$\chi^2 = \frac{(|n_1 - m_1| - 1/2)^2}{m_1} + \frac{(|n_2 - m_2| - 1/2)^2}{m_2} + \cdots + \frac{(|n_k - m_k| - 1/2)^2}{m_k} \underset{n\to\infty}{\sim} \chi_{k-1}^2 \text{ 分布} \quad (10.2)$$

度数がそれほど多くない場合はこの式を用いる方が無難なようです。

注）イェーツ補正を用いた式は、検定確率が有意水準に近くなると正確な値に近づくようになっています。有意水準からかなり離れた場合は、正確な確率からずれており、むしろ補正しない方が良い値となります。そのため、有意でない場合に確率の値を書くことはお勧めできません。報告書などでは、n.s.（有意差なし）としておくべきでしょう。

問題

ある大学の学生 50 人を任意抽出し、大学改革のアンケート調査を行ったところ、賛成 35、反対 15 であった。学生の過半数が賛成している（賛成の比率が 1/2 と異なる）といえるか、有意水準 5%で判定せよ。

解答

帰無仮説 H_0：賛成と反対は比率 1/2 である。

対立仮説 H_1：H_0 でない。

$$\chi^2 = \frac{(35-25)^2}{25} + \frac{(15-25)^2}{25} = \frac{200}{25} = 8$$

$$p = chidist(8,1) = 0.004678 \cong 0.005$$

$p < 0.05$ より、賛成は過半数であるといえる。

（正確には、賛成と反対は確率 1/2 でないといえる。）

問題

上の例題で、他の曜日を 1 つにまとめた場合、金曜日は特に事故が起こっているといえるか。有意水準 5%で判定せよ。

解答

曜日	金曜	その他
事故件数	23	57
予想確率	1/5	4/5
予想値	16	64

帰無仮説 H_0：事故は金曜に 1/5 の比率で起きている。

対立仮説 H_1：H_0 でない。

$$\chi^2 = \frac{(23-16)^2}{16} + \frac{(57-64)^2}{64} = 3.828125$$

$$p = chidist(3.828125, 1) = 0.050399 \cong 0.0504$$

$p > 0.05$ より、金曜日に多いとは言えない。しかし、結果がぎりぎりなので考察の余地は残る。一様性の検定と検定結果が異なるが、データをまとめることによりこのようなこともあり得る。

10.2 対応のない多群間の比率の検定

10.2.1 2×2 表の検定

例

ある商品の購入意欲に男女差があるかどうか調べるために、男女によって購入意思の有無を分けたところ以下の結果を得た。男女差はあるといえるか。有意水準 5%で判定せよ。

	意欲あり	意欲なし	計
男	18	10	28
女	12	14	26
計	30	24	54

理論（χ^2 検定）

ある 2 つの事象 1 と事象 2 の実現度数を 2 つの要因 1 と要因 2 により分けると以下のようになった。事象 1 と事象 2 の出現比率の間に 2 つの要因による差が認められるか。有意水準 $\alpha \times 100\%$ で判定する。

	事象 1	事象 2	計
要因 1	*a*	*b*	*a*+*b*
要因 2	*c*	*d*	*c*+*d*
計	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*=*n*

帰無仮説 H_0：要因間に差がない。（事象の出現比率に差がない）

対立仮説 H_1：要因間に差がある。（事象の出現比率に差がある）

H_0 のもとで $\chi^2 = \dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \underset{n\to\infty}{\sim} \chi_1^2$ 分布 (10.3)

$p = chidist(\chi^2, 1)$ として、$p < \alpha$ ならば、H_0 を棄却して H_1 を採択する。

解答

2 次元分割表から、$a = 18, b = 10, c = 12, d = 14, n = 54$ となり、これを用いて χ^2 統計値

を求めると以下のようになります。

$$\chi^2 = \frac{54\times(18\times14-10\times12)^2}{28\times26\times30\times24} = 1.795055$$

自由度は 1 ですから、検定確率値は *chidist*() 関数を用いて以下のようになります。

$$p = chidist(1.795055, 1) = 0.180312 \cong 0.180$$

結局 $p > 0.05$ ですから、要因による差があるとはいえないという結論になります。

解説

この χ^2 統計量は一般の χ^2 統計量の 2×2 分割表についての特別な形式です。一般的な書き方は次の $m\times n$ 表の検定で示しますが、少々厄介なので特によく利用される 2×2 分割表の場合だけ別にしておきました。解答の計算は分割表の度数を使うだけですから、特に問題はないと思います。

数学的解説　[Skip OK]

さて、ここで述べた統計量 χ^2 が自由度 1 の χ^2 分布に従うことは直感では分りませんので、少し理論の背景を探ってみることにします。

今、要因に関わらず事象 1 の出現比率を p_0 とします。要因 i における事象 1 の出現度数を確率変数として X_i とし、その他の度数を以下の表のように与えます。

	事象 1	事象 2	合計
要因 1	X_1	n_1-X_1	n_1
要因 2	X_2	n_2-X_2	n_2

確率変数 X_i が 2 項分布に従うことを利用すると以下のようになります。

$$\frac{X_i - n_i p_0}{\sqrt{n_i p_0(1-p_0)}} \underset{n_i\to\infty}{\sim} N(0,1) \text{ 分布}$$

左辺の統計量に $1/\sqrt{n_i}$ を掛けると、正規分布の性質から、

$$\frac{X_i/n_i - p_0}{\sqrt{p_0(1-p_0)}} \underset{n_i\to\infty}{\sim} N\left(0, \frac{1}{n_i}\right) \text{分布}$$

となります。さらに、$X_i/n_i = \hat{p}_i$ として、$i=1$ の場合の統計量から $i=2$ の場合の統計量を引くと、以下の関係が成り立ちます。

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)}} \underset{n_i\to\infty}{\sim} N\left(0, \frac{1}{n_1}+\frac{1}{n_2}\right) \text{分布}$$

左辺の統計量を $\sqrt{1/n_1 + 1/n_2}$ で割って Z とすると、データ数が多い場合、Z は標準正規分布に近づくことが示されます。

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \underset{n_i\to\infty}{\sim} N(0,1) \text{ 分布}$$

ここで、問題は p_0 が何かということです。我々は p_0 について知りませんので、予想するしかありません。そこで、p_0 の代わりに以下の $\bar{p}$ を用います。

$$\bar{p}=\frac{X_1+X_2}{n_1+n_2} \qquad Z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \underset{n_i\to\infty}{\sim} N(0,1)\text{ 分布}$$

ここに $\bar{p}$ は標本全体から p_0 を見積もったものですので、あくまで近似です。検定手順は、両側検定の場合 $p/2=1-normsdist(Z)$ として、$p<\alpha$ ならば帰無仮説を棄却し、対立仮説を採択することになります。これは比率を元にした検定手法で、直感的に分り易いのでよく利用されます。

さて、ここで理論のところで述べた実際の観測値を入れて Z^2 を計算してみましょう。

$$n_1=a+b,\ \ n_2=c+d,\ \ \hat{p}_1=\frac{a}{a+b},\ \ \hat{p}_2=\frac{c}{c+d},\ \ \bar{p}=\frac{a+c}{a+b+c+d}=\frac{a+c}{n}$$

として、少々計算すると

$$Z^2=\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \underset{n\to\infty}{\sim} \chi_1^2\text{ 分布}$$

となり、理論式が導かれます。

データ数があまり多くない場合、以下のような形で補正項が入ります。

$$\chi^2=\frac{n(|ad-bc|-n/2)^2}{(a+b)(c+d)(a+c)(b+d)} \underset{n\to\infty}{\sim} \chi_1^2\text{ 分布} \qquad (10.4)$$

また、比率を元にすると補正項は以下のようになります。

$$Z=\frac{|\hat{p}_1-\hat{p}_2|-\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \underset{n_i\to\infty}{\sim} N(0,1)\text{ 分布}$$

実際の検定では、補正項を含む式を利用するのが安全なようです。

10.2.2 m×n 表の検定　[Skip OK]

例

ある地域の女性について、ある商品の所有の有無を職業別に分類してみると、以下の結果が得られた。職業間で商品所有の割合に差が認められるか。有意水準 5%で判定せよ。

	所有有り	所有無し	計
主婦	90	199	289
事務	40	39	79
販売・生産	53	71	124
計	183	309	492

理論

要因（r 種）により事象（s 種）の出現状況を分けると以下のようになった。出現比率に要因による差が認められるか。有意水準 $\alpha\times100\%$ で判定する。

	事象 1	事象 2	…	事象 s	計
要因 1	x_{11}	x_{12}	⋯	x_{1s}	$x_{1\cdot}$
要因 2	x_{21}	x_{22}	⋯	x_{2s}	$x_{2\cdot}$
:	:	:		:	:
要因 r	x_{r1}	x_{r2}	⋯	x_{rs}	$x_{r\cdot}$
計	$x_{\cdot 1}$	$x_{\cdot 2}$	⋯	$x_{\cdot s}$	n

H_0：出現比率に要因による差はない（要因と独立である）

H_1：出現比率に要因による差がある（要因と独立でない）

H_0 のもとで $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(x_{ij} - x_{i\bullet}x_{\bullet j}/n)^2}{x_{i\bullet}x_{\bullet j}/n} \underset{n\to\infty}{\sim} \chi^2_{(r-1)(s-1)} \text{ 分布} \qquad (10.5)$$

$p = chidist(\chi^2, (r-1)(s-1))$ とし、$p < \alpha$ ならば、H_0 を棄却して H_1 を採択する。

解答

これは 3×2 の分割表なのですが、計算が相当複雑です。

$$x_{11} = 90,\ x_{12} = 199,\ x_{21} = 40,\ x_{22} = 39,\ x_{31} = 53,\ x_{32} = 71$$
$$x_{1\bullet} = 289,\ x_{2\bullet} = 79,\ x_{3\bullet} = 124,\ x_{\bullet 1} = 183,\ x_{\bullet 2} = 309,\ n = 492$$

として、χ^2 統計値は以下のようになります。

$$\chi^2 = \frac{(90 - 289\times 183/492)^2}{289\times 183/492} + \frac{(199 - 289\times 309/492)^2}{289\times 309/492}$$
$$+ \frac{(40 - 79\times 183/492)^2}{79\times 183/492} + \frac{(39 - 79\times 309/492)^2}{79\times 309/492}$$
$$+ \frac{(53 - 124\times 183/492)^2}{124\times 183/492} + \frac{(71 - 124\times 309/492)^2}{124\times 309/492}$$
$$= 12.27293$$

自由度は $(3-1)(2-1) = 2$ ですから、検定確率値は $chidist()$ 関数を用いて以下のようになります。

$$p = chidist(12.27293, 2) = 0.002163 \cong 0.002$$

これより $p < 0.05$ ですから、職業（要因）間に差があるといえるという結論になります。

解説

この一般の $m\times n$ 表の検定については、Excel を用いて簡単に計算するというには厄介ですので、ここでは簡単な例と理論をあげておくに留めます。計算には統計分析の専用ソフトウェアを利用することをお勧めします。著者のホームページ上からダウンロードできる分析ツールを利用するのもよいでしょう。

この検定では複数の事象の出現比率の比較をしていますが、どこに差があるのか明らかにすることはできません。これは比率に関する多重比較の問題として、別の本に譲ることにします。

数学的解説　[Skip OK]

さて、ここで与えた表式はかなり複雑な形をしています。この形と 2×2 表で与えた形が

とても同じだとは思えないほどです。そこで、少し面倒ですがこれらが実際に一致することを示しておきましょう。但し、$x_{11}=a$，$x_{12}=b$，$x_{21}=c$，$x_{22}=d$，$x_{1\bullet}=a+b$，$x_{2\bullet}=c+d$，$x_{\bullet 1}=a+c$，$x_{\bullet 2}=b+d$ となります。

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(x_{ij}-x_{i\bullet}x_{\bullet j}/n)^2}{x_{i\bullet}x_{\bullet j}/n}=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(x_{ij}n-x_{i\bullet}x_{\bullet j})^2}{nx_{i\bullet}x_{\bullet j}}$$

$$=\frac{(x_{11}n-x_{1\bullet}x_{\bullet 1})^2}{nx_{1\bullet}x_{\bullet 1}}+\frac{(x_{12}n-x_{1\bullet}x_{\bullet 2})^2}{nx_{1\bullet}x_{\bullet 2}}+\frac{(x_{21}n-x_{2\bullet}x_{\bullet 1})^2}{nx_{2\bullet}x_{\bullet 1}}+\frac{(x_{22}n-x_{2\bullet}x_{\bullet 2})^2}{nx_{2\bullet}x_{\bullet 2}}$$

$$=\frac{(an-(a+b)(a+c))^2}{n(a+b)(a+c)}+\frac{(bn-(a+b)(b+d))^2}{n(a+b)(b+d)}$$

$$+\frac{(cn-(c+d)(a+c))^2}{n(c+d)(a+c)}+\frac{(dn-(c+d)(b+d))^2}{n(c+d)(b+d)}$$

ここで、分子を計算すると、すべて $(ad-bc)^2$ となりますので、以下のように計算が続きます。

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

$$\times[(c+d)(b+d)+(c+d)(a+c)+(a+b)(b+d)+(a+b)(a+c)]$$

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

一般の $m\times n$ 表の統計量 χ^2 について、その表式を適合度検定から見直してみましょう。要因 i の出現確率を p_i、事象 j の出現確率を q_j とし、それらが独立であるとすると、要因 i，事象 j の出現確率は、p_iq_j となります。この場合、理論的な出現度数は、np_iq_j となりますので、適合度検定の統計量は以下のように表現できます。

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(x_{ij}-np_iq_j)^2}{np_iq_j}\underset{n\to\infty}{\sim}\chi^2_{rs-1}\text{分布}$$

ここに自由度が 1 減っているのは、$\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-np_iq_j)=0$ の制約が 1 つあるからです。

しかし、我々にはこの理論確率が分りませんので各行、各列の合計から推測する他はありません。そこで、$p_i=x_{i\bullet}/n$，$q_j=x_{\bullet j}/n$ とおくことにします。これを利用して、統計量を書き直すと、以下のようになります。

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(x_{ij}-x_{i\bullet}x_{\bullet j}/n)^2}{x_{i\bullet}x_{\bullet j}/n}\underset{n\to\infty}{\sim}\chi^2_{(r-1)(s-1)}\text{分布}$$

ここで自由度について直感的に考えてみます。理論確率を上のようにおくと、分子の確率変数に、

$$\sum_{j=1}^{s}(x_{ij}-x_{i\bullet}x_{\bullet j}/n)=0,\quad\sum_{i=1}^{r}(x_{ij}-x_{i\bullet}x_{\bullet j}/n)=0$$

の制約が付くことになります。制約式の数は左が r 個、右が s 個です。但し、これらの制約

より導かれる

$$\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-x_{i\bullet}x_{\bullet j}/n)=0$$

の制約はどちらの式からも導かれますので、制約式の数は全部で$r+s-1$個になります。それゆえ、自由度は$rs-(r+s-1)=(r-1)(s-1)$で与えられます。

データ数がそれほど多くない場合、理論で与えた検定量には補正項が入り以下のようになります。

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(|x_{ij}-x_{i\bullet}x_{\bullet j}/n|-\frac{1}{2})^2}{x_{i\bullet}x_{\bullet j}/n}\underset{n\to\infty}{\sim}\chi^2_{(r-1)(s-1)}\text{ 分布} \qquad (10.6)$$

一般にこちらの統計量を使うことをお勧めします。

10.3 対応のある２群間の比率の検定

例

経営状態の良い支店と悪い支店とを規模でマッチングさせて、ある要因の有無で分類させたところ以下の表を得た。経営状態にこの要因による差があると考えられるか。有意水準5%で判定せよ。

良＼悪	要因有	要因無
要因有	10	24
要因無	11	40

理論　マクネマー（McNemar）検定

データと対照データとをある条件でマッチさせて、要因の有無で分類したところ以下の表を得た。データと対照データに要因による差があると考えられるか。有意水準$\alpha\times100\%$で判定する。

群 1＼群 2	要因有	要因無
要因有	a	b
要因無	c	d

帰無仮説 H_0：要因による差がない

対立仮説 H_1：要因による差がある

H_0のもとで　　$\chi^2=\dfrac{(b-c)^2}{b+c}\underset{b,c\to\infty}{\sim}\chi^2_1$ 分布　　(10.7)

$p=chidist(\chi^2,1)$として、$p<\alpha$ならば、H_0を棄却してH_1を採択する。

解答

これは計算が極めて簡単です。$b=24, c=11$ですから、χ^2統計値は以下となります。

$$\chi^2=\frac{(24-11)^2}{24+11}=4.828571$$

これから自由度を 1 として、$chidist()$関数を用い、検定確率値を求めると、以下のように

なります。

$$p = chidist(4.828571, 1) = 0.027992 \cong 0.028$$

これより $p < 0.05$ ですから、要因による差があるといえるという結論になります。

解説

この例の場合、経営状態の良い支店と悪い支店を対応させた以下のような形式のデータを元にしています。

番号	経営良	経営悪
1	要因有り	要因有り
2	要因有り	要因無し
3	要因無し	要因有り
:	:	:
85	要因無し	要因有り

このデータをまとめて、例で述べた集計表を作ります。

これまではデータを２つの群に分けるとき、以下のような分け方をしていました。

	要因有	要因無
群 1	a'	b'
群 2	c'	d'

これは、群1と群2について要因の有無の割合の比較になります。これに対して上のように、2 群のデータ間に 1 対 1 の対応が付けられる場合、より有効な検定方法があります。それがここで述べる McNemar 検定です。この検定は、要因について 2 群のデータを対応の組ごとに有－有，有－無，無－有，無－無の 4 つの場合に分け、それぞれの組の数を表に記入します。

群 1＼群 2	要因有	要因無
要因有	a	b
要因無	c	d

この対応を考えない表と対応を考えた表とでは、後者が 1 組を 1 つと数えることから、要素数の合計に 2 倍の差がでます。

$$a' + b' + c' + d' = 2(a + b + c + d)$$

後者の集計では要因の有無について差がないとすると、2 群で有－無，無－有となる確率は等しくなるでしょう。そこで２つの場合について、それぞれの出現確率が 1/2 であるかどうか検定します。この検定は適合度検定ですので、全データ数を $b + c$ として検定統計量は以下のようになります。

$$\chi^2 = \frac{[b - (b+c)/2]^2}{(b+c)/2} + \frac{[c - (b+c)/2]^2}{(b+c)/2} = \frac{(b-c)^2}{b+c} \sim \chi_1^2 \text{ 分布}$$

データ数が少ない場合、適合度検定では以下のように補正項が入り、

$$\chi^2 = \frac{[|b - (b+c)/2| - ½]^2}{(b+c)/2} + \frac{[|c - (b+c)/2| - ½]^2}{(b+c)/2}$$

まとめて次のようになります。

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \underset{b,c\to\infty}{\sim} \chi_1^2 \text{ 分布} \qquad (10.8)$$

10.4 比率の検定のためのデータ数の決定

ここでは検定自体の話から離れて、ある種の調査を行なう場合の、有効な調査対象数の選び方の話をします。一般に調査対象数が多ければ多いほど検定精度が上がり、有意差が出易くなることが知られていますが、人手や予算の関係で調査の規模が制限されることも事実です。このことから、調べたいことの有意性を出す最低限の対象数を知っておくことは重要です。これには予め小さな規模の予備調査を行ない、例えばある案に賛成する割合がどの程度あるのかということを知った上で、これから述べる方法を適用して本番の対象数の決定を行ないます。ここでは、10.1 節で学んだ適合度検定で 2 分割の簡単な場合を用いて、データ数決定の考え方を学びます。もう少し詳しいことは解説のところでも再度話をします。

例

アンケート調査で、「はい」と答えた回答が 60%と予想されるとき、有意水準 5%で過半数である（「はい」が 1/2 でない）と判定するために必要なデータ数はいくらか。

理論

２つの事象の想定比率がそれぞれ、p_0，$1-p_0$ であるとき、有意水準$\alpha\times100\%$ で実現比率 $\hat{p}$ を想定比率と異なると判定するために必要なデータ数を求める。

適合度検定の検定統計量の性質を利用して、データ数は以下で与えられる。

$$n > chiinv(\alpha,1)\times\frac{p_0(1-p_0)}{(\hat{p}-p_0)^2} \qquad (10.9)$$

解答

有意水準の確率値 0.05 の χ^2 統計値は、$\chi_1^2(0.05) = chiinv(0.05,1) = 3.841455$ で与えられます。また、$p_0 = 0.5$，$\hat{p} = 0.6$ ですから、上の式を用いて以下のようになります。

$$n > chiinv(0.05,1)\times\frac{0.5\times0.5}{(0.6-0.5)^2} = 96.03638$$

これから、データ数は 97 以上必要であることが分かります。

解説

この節は検定にまつわる話題にページを割いてみましょう。アンケート調査をする際、我々は何を求めているのでしょうか。例えば、支持率が過半数かどうかを知りたいというのは、過半数であることをはっきりさせたい場合に調査することが多いと思います。対立仮説が採択されて過半数であると示されない限り、帰無仮説が採択された検定結果は「この段階では過半数とはいえない」となるだけで、はっきりと過半数か否かを判定できているわけではないからです。

検定の結果は標本のデータ数に依存します。より多くのデータを集めるほど、母集団を推

測し易くなることは直感的に理解できると思います。ではどれほどのデータを集めればよいのでしょうか。予め予備調査を行っており、集まったデータから大体何割の人が支持するかということが分っていたなら、過半数を示すのに必要なデータ数を割り出すことができるというのがこの節の話です。これは調査の規模を決める問題として非常に重要です。もちろん調査対象は多いに越したことはありません。しかし費用や人的な制限から、調査規模は制約を受けます。その際、大体何人位調査するとはっきりとした結果が得られるのかが分かれば、調査計画も立て易くなります。

この節では想定比率と実測比率を比較する場合の調査対象数を決定する手法を示しています。特に事象が 2 つの場合、取り扱いが簡単なので、ここではこの場合に限って説明しています。事象が 3 つ以上の場合は、比率の設定が厄介ですが、それさえ分れば同様な考えでデータ数を決めることができます。

理論のところで、データ数の決定には適合度検定の検定統計量の性質を利用すると書きましたが、もう少し詳しく説明しておきます。適合度検定の検定統計量 χ^2 は以下のように与えられますが、

$$\chi^2 = \frac{(n_1 - np_0)^2}{np_0} + \frac{[n_2 - n(1-p_0)]^2}{n(1-p_0)} \underset{n\to\infty}{\sim} \chi_1^2 \text{分布}$$

これを次のように変形します。

$$\begin{aligned}\chi^2 &= \frac{(n_1 - np_0)^2}{np_0} + \frac{[n_2 - n(1-p_0)]^2}{n(1-p_0)} \\ &= \frac{n^2(\hat{p} - p_0)^2}{np_0} + \frac{n^2[(1-\hat{p}) - (1-p_0)]^2}{n(1-p_0)} \\ &= \frac{n(\hat{p} - p_0)^2}{p_0(1-p_0)}\end{aligned}$$

変形には事前調査などによる実現比率を $\hat{p}$ として、$n_1 = n\hat{p}$, $n_2 = n(1-\hat{p})$ の関係を用いています。ここで χ^2 の値がどの程度になれば有意水準 $\alpha \times 100\%$ で有意性を示すことができるか、ということは Excel 関数 *chiinv*(確率, 自由度)を使って容易に求めることができます。即ち、

$$\frac{n(\hat{p} - p_0)^2}{p_0(1-p_0)} > chiinv(\alpha, 1)$$

この式から理論で与えた(10.9)式が出ます。

問題

以下の場合、想定比率 0.5 と有意差を出すためのデータ数はいくら必要か？

1) 実測比率 0.7 で、有意水準 5%として有意
2) 実測比率 0.55 で、有意水準 5%として有意
3) 実測比率 0.6 で、有意水準 1%として有意

解答

1) 25 以上　　2) 385 以上　　3) 166 以上