9

カイ2乗検定

9-1 カイ2乗検定

一般に，帰無仮説の下である検定統計量が χ^2 分布に従う場合，この検定を**カイ2乗検定**という名でよぶことが多い．カイ2乗検定には，適合度検定や分割表での独立性の検定などがある．

9-2 適合度検定

母集団は互いに排反な k 個の級 $C_1, C_2, \cdots, C_k$ に分れているとし，この母集団から1個の標本をとるとき，それが $C_1, C_2, \cdots, C_k$ に入る確率を $p_1, p_2, \cdots, p_k$ $(\sum_{i=1}^{k} p_i = 1)$ とする．いま，母集団から n 個の標本をとるとき，それらが $C_1, \cdots, C_k$ に入る**観測度数**を $n_1, n_2, \cdots, n_k$ $(\sum_{i=1}^{k} n_i = n)$ とすれば，これら級に入る**期待度数**は $m_1 = np_1$, $m_2 = np_2$, $\cdots$, $m_k = np_k$ となる．

級	C_1	C_2	$\cdots$	C_k	計
観測度数	n_1	n_2	$\cdots$	n_k	n
期待度数	m_1	m_2	$\cdots$	m_k	n

このとき，

帰無仮説 H_0：母集団の各級の確率（または確率分布）は

$$p_1 = p_{10},\ p_2 = p_{20},\ \cdots,\ p_k = p_{k0}$$

である

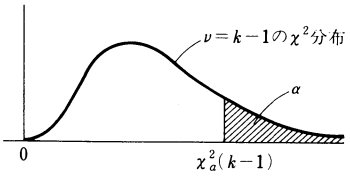
の検定に，検定統計量として

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-m_i)^2}{m_i}=\sum_{i=1}^{k}\frac{(n_i-np_{i0})^2}{np_{i0}}$$

が使われる．この統計量は n が大きく，各 m_i が5以上であれば H_0 の下で近似的に自由度 $\nu=k-1$ の χ^2 分布に従う．この検定を**適合度検定**という．有意水準が α %のこの検定の棄却域は

$$\chi^2>\chi_\alpha^2(k-1)$$

で与えられる．



期待度数の計算で，母集団の未知母数を推定することが必要な場合には，推定される母数が c 個ならば，このときのカイ2乗検定の棄却域は

$$\chi^2>\chi_\alpha^2(k-c-1)$$

となる．

9-3　独立性の検定

n 個の標本を2つの属性 A, B によって，次のような2元表に分割する．ここで，n_{ij} は級 (A_i, B_j) に入る標本の個数で，この表は **$r\times s$ 分割表**とよばれる．分類に用いる属性は定性的なものでも定量的なものでもよい．

$r\times s$ 分割表

A \ B	B_1	B_2	$\cdots$	B_s	計
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	n

$$n_{i\cdot}=\sum_{j=1}^{s}n_{ij},\qquad n_{\cdot j}=\sum_{i=1}^{r}n_{ij}$$

$$n=\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}=\sum_{i=1}^{r}n_{i\cdot}=\sum_{j=1}^{s}n_{\cdot j}$$

1個の標本が級 (A_i, B_j) に入る確率を p_{ij}，級 A_i に入る確率を $p_{i\cdot}$，級 B_j に入る確率を $p_{\cdot j}$ とする．ここで検定すべき仮説は，2つの属性 A, B が独立であるという仮説，すなわち $P(A_i\cap B_j)=P(A_i)P(B_j)$ で，これは，

$$H_0: p_{ij}=p_{i\cdot}\,p_{\cdot j}\qquad (i=1,2,\cdots,r;\ j=1,2,\cdots,s)$$

で表される．H_0 が真の下で級 (A_i, B_j) に入る標本の期待度数 m_{ij} は，

$$m_{ij}=\frac{n_{i\cdot}\,n_{\cdot j}}{n}$$

で与えられる．よって，H_0 を検定する統計量は

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\dfrac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\dfrac{n_{i\cdot}n_{\cdot j}}{n}}$$

で，nが十分大きいならば，H_0が真のとき，これは近似的に自由度$\nu=(r-1)(s-1)$のχ^2分布に従う．よって，このカイ2乗検定の有意水準α%の棄却域は

$$\chi^2>\chi_\alpha^2((r-1)(s-1))$$

2×2分割表　$r=s=2$の場合を**2×2分割表**という．このときχ^2は，次の式になる．

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

2×2分割表

A \ B	B_1	B_2	計
A_1	a	b	$a+b$
A_2	c	d	$c+d$
計	$a+c$	$b+d$	n

イエーツの補正　2×2の分割表で期待度数が小さい場合にはχ^2分布への近似をよくするために，$\frac{1}{2}$だけずらして

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(|n_{ij}-m_{ij}|-\dfrac{1}{2}\right)^2}{m_{ij}}$$

を使うのがよいとされている．これを**イエーツの補正**という．

例　題

例題 1　**（一様分布の適合度検定）**

観測者が測定器具の目盛りを読むとき，最後の桁の数字は目測で判読される．その際，特定の数字を好む傾向のあることが指適されている．いま，200個の数字について，次の結果が得られた．この観測者は特別な数字を好む傾向があるだろうか．5%有意水準で検定せよ．

最後の桁の数字	0	1	2	3	4	5	6	7	8	9	計
観測度数	32	16	18	19	17	25	11	16	30	16	200

解 この場合の帰無仮説としては「観測者は特別な数字を好む傾向をもたない」をとるのが適当で，観測者が最後の桁の数字を i と読む確率を p_i とすると，この仮説は

$$H_0 : p_0=p_1=\cdots=p_9=\frac{1}{10}$$

で表される．H_0 が真であれば，期待度数は明らかに

$$m_i=np_{i0}=200\times\frac{1}{10}=20 \qquad (i=0, 1, \cdots, 9)$$

であるから，観測度数 n_i と期待度数 m_i を表にして示すと

	0	1	2	3	4	5	6	7	8	9	計
n_i	32	16	18	19	17	25	11	16	30	16	200
m_i	20	20	20	20	20	20	20	20	20	20	200

よって，

$$\chi^2=\sum_i\frac{(n_i-m_i)^2}{m_i}=\frac{(32-20)^2}{20}+\frac{(16-20)^2}{20}+\cdots+\frac{(16-20)^2}{20}=20.6$$

$\nu=10-1=9$, $\alpha=0.05$. χ^2 分布表より $\chi^2_{0.05}(9)=16.92$. ゆえに，棄却域は

$$\chi^2>16.92$$

データから求めた χ^2 の値 20.6 は棄却域に入るから，H_0 は棄却される．

したがって，この観測者には特別な数字を好む傾向があるといえる．実際，この観測者は 0 と 8 を好むようである．

例題 2 （適合度検定）

人間の血液型は 4 種類で，その構成比率は $q^2 : p^2+2pq : r^2+2qr : 2pr$ であるという．ただし $p+q+r=1$. いま，ある職業についている 770 人の血液型を調べて，

180, 360, 132, 98

なる観測度数を得た．これより，仮説「この職業人の血液型の分布は，$p=0.4$, $q=0.4$, $r=0.2$ で定まる構成比率をもつ」を検定せよ．

解 仮説が正しいとして 4 種類の血液型の期待度数を求めれば

$$n_1=nq^2=770\times0.4^2=123.2$$

$$n_2=n(p^2+2pq)=770\times(0.4^2+2\times0.4\times0.4)=369.6$$

$$n_3=n(r^2+2qr)=770\times(0.2^2+2\times0.4\times0.2)=154.0$$
$$n_4=n(2pr)=770\times(2\times0.4\times0.2)=123.2$$

観測度数と比較すると，

					計
n_i	180	360	132	98	770
m_i	123.2	369.6	154.0	123.2	770.0

$$\chi^2=\frac{(180-123.2)^2}{123.2}+\frac{(360-369.6)^2}{369.6}+\frac{(132-154.0)^2}{154.0}+\frac{(98-123.2)^2}{123.2}$$
$$=34.73$$

$\alpha=0.05$, $\nu=4-1=3$. χ^2 分布表より $\chi^2_{0.05}(3)=7.81$. よって棄却域は

$$\chi^2>7.81$$

χ^2 の値は棄却域に入っているから，仮説は棄却される．したがってこの職業人の血液型の分布は仮説が与える構成比率とは**異なる**ものである．

例題 3　（ポアソン分布の適合度検定）

大気中に浮遊するある微小な物質の量を推定するため，空間内にいくつかの点を選び，その点のまわりの単位体積内の粒子数を計測する．いま300点を選んで観測した結果，つぎのデータが得られた．

（a）　粒子の数の平均と分散を求めよ．

（b）　このデータにポアソン分布をあてはめ，その適合性を調べよ．

粒子の数	0	1	2	3	4	5	6以上	計
観測度数	38	75	89	54	20	19	5	300

解　（a）　粒子数を x，度数を f とすると，平均は

$$\bar{x}=\frac{\sum x_if_i}{n}=\frac{0\times38+1\times75+\cdots+6\times5}{300}=\frac{31}{15}\fallingdotseq\mathbf{2.07}$$

分散は

$$s^2=\frac{\sum x_i^2f_i}{n}-\bar{x}^2=\frac{0^2\times38+1^2\times75+\cdots+6^2\times5}{300}-\left(\frac{31}{15}\right)^2\fallingdotseq\mathbf{2.04}$$

平均と分散がほぼ等しいので，粒子の数の分布としてポアソン分布が予想される．

（b） 粒子の数は $\lambda=2.07$ のポアソン分布に従うという帰無仮説の下で，期待度数を次のように求める．

値 x	0	1	2	3	4	5	6以上	計
確率 $e^{-2.07}\frac{2.07^x}{x!}$	0.13	0.26	0.27	0.19	0.10	0.04	0.01	1.0
期待度数＝300×確率	39	78	81	57	30	12	3	300
観測度数	38	75	89	54	20	19	5	300

これから

$$\chi^2=\frac{(38-40.6)^2}{40.6}+\frac{(75-81.2)^2}{81.2}+\cdots+\frac{(5-5.0)^2}{5.0}=9.39$$

$\alpha=0.05$, $\nu=7-1=6$ より $\chi^2_{0.05}(6)=12.59$. $\chi^2=9.39$ はこの値を超えないから，仮説は採択である．よってこのデータはポアソン分布に**適合している**．

例題 4 （2×2 分割表）

ある会社の社員60名に，タバコをすうかすわないかと，パチンコをするかしないかを調査した．その結果は次のようであった．タバコをすうこととパチンコをすることは独立かどうかを5%有意水準で検定せよ．

	パチンコをする	しない	
タバコをすう	9	3	12
すわない	18	30	48
	27	33	60

解 この問題の帰無仮説は，H_0：タバコとパチンコとは独立，である．H_0 の下での期待度数は

	パチンコ	しない
タバコ	27×12/60＝ 5.4	33×12/60＝ 6.6
すわない	27×48/60＝21.6	33×48/60＝26.4

イエーツの補正を施してカイ2乗検定を行う．

観測度数 n_i	期待度数 m_i	n_i-m_i	$\lvert n_i-m_i\rvert-0.5$	$\frac{(\lvert n_i-m_i\rvert-0.5)^2}{m_i}$
9	5.4	3.6	3.1	1,780
3	6.6	−3.6	3.1	1,456
18	21.6	−3.6	3.1	0.445
30	26.4	3.6	3.1	0.364
				計　4.045

$\alpha=0.05$, $\nu=1$ より $\chi^2_{0.05}(1)=3.84$ であるから，棄却域は $\chi^2>3.84$. よって帰無仮説は棄却される．したがってタバコとパチンコは**独立ではない**.

例題 5　**(2×4 分割表)**

アメリカでの調査によると，息子が父親の職業と同じ職業を選ぶかどうかを調べて，次の結果を得た.

息子の職業	父親の職業 医者	銀行員	教員	弁護士	計
父親と同じ職業	34	27	28	19	108
父親と異なる職業	166	123	152	81	522
計	200	150	180	100	630

「息子の職業選択と親の職業とは独立である」という仮説を 5%有意水準で検定せよ.

解　これは 2×4 分割表での独立性の検定の問題である.

各枡の期待度数を求めると，

$$m_{11}=\frac{200\times108}{630}=34.3,\qquad m_{12}=\frac{150\times108}{630}=25.7,$$

$$m_{13}=\frac{180\times108}{630}=30.9,\qquad m_{14}=\frac{100\times108}{630}=17.1,$$

$$m_{21}=\frac{200\times522}{630}=165.7,\qquad m_{22}=\frac{150\times522}{630}=124.3,$$

$$m_{23}=\frac{180\times522}{630}=149.1,\qquad m_{24}=\frac{100\times522}{630}=82.9$$

よって期待度数の表は

34.3	25.7	30.9	17.1
165.7	124.3	149.1	82.9

これから

$$\chi^2=\sum_{i,j}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}=\frac{(34-34.3)^2}{34.3}+\frac{(27-25.7)^2}{25.7}+\cdots+\frac{(81-82.9)^2}{82.9}$$
$$=0.64$$

自由度 $\nu=(2-1)(4-1)=3$, $\chi^2_{0.05}(3)=7.81$ であるから，仮説は棄却されない．したがって息子の職業選択に親の職業は**関係しない**．

9章の問題

9.1 乱数サイを200回実際に振って，次の結果を得た．この乱数サイは正しいと認められるか．

目の数	0	1	2	3	4	5	6	7	8	9	計
度数	26	27	20	13	19	19	15	19	27	15	200

9.2 次の表は100個の乱数の標本であるといわれている．0から9までの各数字の度数はその期待度数と有意に異なるか．5%有意水準で検定せよ．

35230	66852	50395	59228	28896	48780	00845	39797	86339	57380
92264	95450	41210	66273	91350	52137	02829	62316	46155	16031

9.3 次の表は10騎馬兵団の20年間の記録で，馬にけられて死んだ兵士の数である．これはBortkewitchによって集められ，Fisherが引用した有名な例である．これに適当なポアソン分布をあてはめ，その適合性を論ぜよ．

死者の数	0	1	2	3	4	計
兵団数	109	65	22	3	1	200

9.4 流行性感冒の予防注射の効果を調べるため，流行性感冒にかかった人とかからなかった人について，それぞれ予防注射をしたか，しなかったかをきき，次の結果を得た．この予防注射は流行性感冒の予防に効果があるといえるか．1%有意水準で検定せよ．