頻度データの分析

適合度
カイ2乗検定
R×C表

そら豆の遺伝の実験

- 仮説
  - メンデルの法則は正しい
  - 実験も正しく行なわれた
- 観察頻度と期待頻度の一致の程度を量的に示す

項目(i)	種	観察頻度(O_i)	理論比率
1	丸，黄	315	9/16
2	しわ，黄	101	3/16
3	丸，緑	108	3/16
4	しわ，緑	32	1/16
計		556	1

適合度の検定

項目(i)	種	観察頻度(O_i)	理論比率	期待頻度(E_i)	残差
1	丸，黄	315	9/16	312.75	2.25
2	しわ，黄	101	3/16	104.25	−3.25
3	丸，緑	108	3/16	104.25	3.75
4	しわ，緑	32	1/16	34.75	−2.75
計		556	1	556	0

- 項目ごとに　残差^2／期待頻度　を計算し加える

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

- 理論比率が正しいときにはX^2は自由度k-1のカイ二乗分布に従う
  - 期待頻度の和は観察度数の和と一致するよう定められるため，自由度は項目数-1

帰無仮説：観察頻度は期待頻度に適合

$$\chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(101-104.25)^2}{104.25} + \frac{(108-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75}$$
$$= 0.47$$
$$\chi^2_{0.05}(3) = 7.81$$

- 観測頻度が期待頻度からはずれるほどX^2の値は大きくなる
- カイ二乗分布の表(P.208表3)より
  - 自由度3の上側5%点：7.81
  - 0.47はこれより小さい
    ⇒帰無仮説は棄却されない
    　観察頻度は期待頻度の誤差の範囲内

R×C表

観測値

	列 1	j	C	計
1				N_1
行 i		O_{ij}		N_i
R				N_R
計	M_1	M_j	M_C	T

期待値

	列 1	j	C	計
1				N_1
行 i		$\frac{N_i M_j}{T}$		N_i
R				N_R
計	M_1	M_j	M_C	T

- i 行 j 列の期待頻度は　$E_{ij} = T \times \frac{N_i}{T} \times \frac{M_j}{T} = \frac{N_i M_j}{T}$

R×C表

- 行と列の要因が独立のとき

  $\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$　は，自由度(R-1)×(C-1)のカイ二乗分布に従う
  - ∑はR×C個の和
- 期待頻度が5より小さいセルは全体の20%以下
  - 検定精度が落ちる
  - 必要に応じて行または列の合併を行なう

R×C表（サンプル）

観測値

	医師	看護師	その他	計
賛成	60	50	120	230
反対	40	150	180	370
計	100	200	300	600

期待値

	医師	看護師	その他	計
賛成	38.33333	76.66667	115	230
反対	61.66667	123.3333	185	370
計	100	200	300	600

R×C表（サンプル）

残差

	医師	看護師	その他	計
賛成	21.66667	−26.6667	5	0
反対	−21.6667	26.66667	−5	0
計	0	0	0	0

残差平方/期待値

	医師	看護師	その他	計
賛成	12.24638	9.275362	0.217391	21.73913
反対	7.612613	5.765766	0.135135	13.51351
計	19.85899	15.04113	0.352526	35.25264

R×C表（サンプル）

- $X^2=35.25$　自由度=(2-1)×(3-1)=2
- $X^2_{0.01}(2)=9.21$　カイ二乗表より（P.208 表3）
- $X^2 > X^2_{0.01}(2)$
  - 帰無仮説：行と列の要因は独立
    ⇒病院全体での3つの職種での賛否は同率
  - 仮説は危険率1%で棄却
- 病院全体での3つの職種での賛否の比率は異なることが示唆される

2×2表

$$\chi^2 = \frac{T(O_{11}O_{22} - O_{12}O_{21})^2}{N_1 N_2 M_1 M_2}$$

- 2×2表はR×C表で求めるカイ二乗と同じ