2021.11.10(水 13:10 ～)

兵庫高校

13　比率データの分析と分割表

（この章のポイント）

1）比率のデータは二項分布を前提とする。
2）二項分布は、標準化して正規分布で近似する。
3）分割表や適合度検定では標準化して二乗和をとり、カイ二乗分布で近似する。

13.1　比率データの分析

まず、ドラマを見る可能性がある調査の対象から、独立に n 人の調査を実施し、見た人を 1, 見てない人を 0 とすれば、その結果は **ベルヌーイ試行**（5.2 節）と考えられます。この場合、見ている人の総数 X は **二項分布** $B(n,P)$（5.3 節）に従います。このとき、母比率 P は真の視聴率を意味します。これに対して、標本比率 $p=X/n$ が母比率 P の点推定量であり、一般的に **視聴率** といわれるのはこの値です。

点推定量：$\widehat{P}=p=\dfrac{X}{n}$

データ数 n が比較的小さい場合には、確率は二項分布の確率関数を用いれば直接計算できますが、比較的大きい場合にはこれが難しいため正規分布で近似します。

13.1.1　正規分布による近似法

二項分布を正規分布で近似できる条件として、以下の条件があります。

$$\underline{\text{正規近似条件}}：np \geqq 5,\ n(1-p) \geqq 5 \tag{13.1}$$

この条件が成り立つ時に、標本比率を正規分布で近似する方法を紹介します。二項分布に従う確率変数 X の母平均と母分散は、5 章の (5.5) 式より

$$E(X)=nP,\quad Var(X)=nP(1-P)$$

です。したがって、標本比率 p の母平均と母分散は、

$$E(p)=P,\quad Var(p)=P(1-P)/n \tag{13.2}$$

とわかります。正規分布で近似するとき、標本比率 p を標準化した値 u が、標準正規分布 $N(0,1)$ に（近似的に）従うことを仮定します。

$$\text{仮定１}：u=\frac{p-P}{\sqrt{P(1-P)/n}} \sim N(0,1) \tag{13.3}$$

まず、母比率 P の区間推定を行うには、分母の P に p を代入します。

$$\text{仮定２}：u=\frac{p-P}{\sqrt{p(1-p)/n}} \sim N(0,1) \tag{13.4}$$

この (13.4) 式を、$P(-1.960 \leqq u \leqq 1.960)=0.95$ に代入すると、

$$P\left(-1.960 \leqq \frac{p-P}{\sqrt{p(1-p)/n}} \leqq 1.960\right)=0.95$$

となります。この式を分子の P に関して解くことで、次の P の **信頼率 95% の信頼区間** が得られます。

$$P\left(p-1.960\sqrt{p(1-p)/n} \leqq P \leqq p+1.960\sqrt{p(1-p)/n}\right)=0.95 \tag{13.5}$$

次に、母比率 P が従来の比率 P_0 より大きいかどうか調べたいとき、帰無仮説は等号、対立仮説は立証したいことなので、

帰無仮説 $H_0 : P = P_0$

対立仮説 $H_1 : P > P_0$

の仮説を立てます。このとき、(13.3) 式中の P に P_0 を代入した u_0 を求めます。

$$\textbf{検定統計量} : u_0 = \frac{p - P_0}{\sqrt{P_0(1-P_0)/n}} \tag{13.6}$$

この u_0 は、帰無仮説 $P = P_0$ の下では、標準正規分布 $N(0,1)$ に従います。そこで、次の棄却域を設定すると、有意水準は 5% で、検出力が高い手法となります。

棄却域 $R : u_0 \geqq u(0.05) = 1.645$

また、立証したいことが、従来 P_0 より母比率 P が小さくなっているかどうかや、異なっているかどうかを調べたいときは、表 13.1 のように行います。

表 13.1　興味の対象（対立仮説）による検定方法の違い

検定名称	興味の対象	帰無仮説	対立仮説	棄　却　域
両側検定	異なるか	$P = P_0$	$P \neq P_0$	$\lvert u_0 \rvert \geqq u(0.025) = 1.960$
右片側検定	大きいか	$P = P_0$	$P > P_0$	$u_0 \geqq u(0.05) = 1.645$
左片側検定	小さいか	$P = P_0$	$P < P_0$	$u_0 \leqq -u(0.05) = -1.645$

13.1.2　直接計算法による母比率の分析

標本の大きさ n が小さくて、$np < 5$ のときは、正規分布への近似条件 (13,1) が成り立ちません。このとき次の定理より、区間推定が行えます（章末問題 13.1,2)。

定理 13.1　X が二項分布 $B(n,p)$ に従うとき、以下の関係式が成り立つ。

$$P(X \geqq x) = \frac{1}{B(x, n-x+1)} \int_0^p t^{x-1}(1-t)^{n-x} dt, \quad x = 1, 2, \cdots, n \tag{13.7}$$

ここで、(n,x) のときの信頼限界 (P_L, P_U) は、$P(X \geqq x) = 0.025$ となる p と、$P(X \leqq x) = 0.025$ となる p と考えれば、次の 正確な 95% 区間推定 が得られます。

$$\frac{\phi_2}{\phi_1 F(\phi_1, \phi_2; 0.025) + \phi_2} \leqq P \leqq \frac{\phi_2'}{\phi_1'/F(\phi_2', \phi_1'; 0.025) + \phi_2'} \tag{13.8}$$

ただし、$\phi_1 = 2(n-x+1),\ \phi_2 = 2x,\ \phi_1' = 2(n-x),\ \phi_2' = 2(x+1).$

13.2　2 群の比率の違いの分析

2 つの母集団における母比率 $P_A,\ P_B$ に関して、次の例題の状況を考えて下さい。

13.2.1　2 群の標本比率の差の分布

母集団 A では m 人のうち X 人、母集団 B では n 人のうち Y 人である場合、各々の母集団ごとに、標本比率に関して母平均と母分散を求めて、以下のように近似できます。

$$p_A = \frac{X}{m} \sim N\left(P_A, \frac{P_A(1-P_A)}{m}\right), \quad p_B = \frac{Y}{n} \sim N\left(P_B, \frac{P_B(1-P_B)}{n}\right)$$

これらの差を考えると、12.2 節と同様に、以下の式が得られます。

$$p_A - p_B \sim N\left(P_A - P_B, \frac{P_A(1-P_A)}{m} + \frac{P_B(1-P_B)}{n}\right) \tag{13.9}$$

これを標準化すると、次の式が導かれます。

$$\text{仮定 1}：u=\frac{(p_A-p_B)-(P_A-P_B)}{\sqrt{\dfrac{P_A(1-P_A)}{m}+\dfrac{P_B(1-P_B)}{n}}}\sim N(0,1) \tag{13.10}$$

この性質を用いて、比率の差に関する、検定や区間推定が行われます。

13.2.2　2 群の比率の差の推定

推定の際には、上記の (13.10) 式の分母の P_A, P_B を p_A, p_B で置き換えます。

$$\text{仮定 2}：u=\frac{(p_A-p_B)-(P_A-P_B)}{\sqrt{p_A(1-p_A)/m+p_B(1-p_B)/n}}\sim N(0,1) \tag{13.11}$$

このとき、母比率 P_A-P_B の区間推定は、以下のように導きます。まず、

$$P(-1.960 \leqq u \leqq 1.960)=0.95$$

の式の u に (13.11) 式を代入して、次の式を導きます。

$$P\left(-1.960 \leqq \frac{(p_A-p_B)-(P_A-P_B)}{\sqrt{p_A(1-p_A)/m+p_B(1-p_B)/n}} \leqq 1.960\right)=0.95$$

この (　) 内を P_A-P_B に関して解き直すことで、以下の信頼限界を得ます。

$$\text{点推定値}：\widehat{P_A-P_B}=p_A-p_B$$

$$\text{信頼限界}：p_A-p_B \pm 1.960\sqrt{\frac{p_A(1-p_A)}{m}+\frac{p_B(1-p_B)}{n}}$$

13.2.3　2 群の比率の差の検定

P_A より P_B が大きいかどうかを調べる場合は、以下の仮説となります。

帰無仮説 H_0: $P_A=P_B$
対立仮説 H_1: $P_A<P_B$

検定に用いる統計量（検定統計量）は、帰無仮説 $P_A=P_B$ の場合を考えます。そこで、(13.10) 式の P_A, P_B に共通の $\overline{p}$ を代入すると、以下の式となります。

$$\text{検定統計量}：u_0=\frac{p_A-p_B}{\sqrt{\left(\dfrac{1}{m}+\dfrac{1}{n}\right)\overline{p}(1-\overline{p})}} \tag{13.12}$$

ただし、$\overline{p}$ とは、次の式で計算される併合比率です。

$$\text{併合比率}：\overline{p}=\frac{x+y}{m+n}$$

対立仮説ごとに、どのように棄却域を設定するかは、次の表 13.4 のように行います。

表 13.4　興味の対象（対立仮説）による検定方法の違い

検定名称	興味の対象	帰無仮説	対立仮説	棄　却　域
両側検定	異なるか	$P_A=P_B$	$P_A\neq P_B$	$\lvert u_0\rvert \geqq u(0.025)=1.960$
右片側検定	大きいか	$P_A=P_B$	$P_A>P_B$	$u_0 \geqq u(0.05)=1.645$
左片側検定	小さいか	$P_A=P_B$	$P_A<P_B$	$u_0 \leqq -u(0.05)=-1.645$

13.3　分割表

例題 13.6　4 台の機械で作った製品の品質を 3 つの級に分類した部品数データを、表 13.5 の形式で集計した。有意水準 5% で品質に差があるかどうかを検定せよ。

表 13.5　機械ごとの部品数

	1 級品	2 級品	3 級品	計
A_1	13	31	6	50
A_2	27	15	8	50
A_3	29	16	5	50
A_4	20	19	11	50
計	89	81	30	200

表 13.6　$a \times b$ 分割表

	B_1	B_2	$\cdots$	B_b	$T_{i\bullet}$
A_1	x_{11}	x_{12}	$\cdots$	x_{1b}	$T_{1\bullet}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2b}	$T_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_a	x_{a1}	x_{a2}	$\cdots$	x_{ab}	$T_{a\bullet}$
$T_{\bullet j}$	$T_{\bullet 1}$	$T_{\bullet 2}$	$\cdots$	$T_{\bullet b}$	T

13.3.1　分割表におけるデータ

一般的に、2 つの分類項目 A, B で A は $A_1 \sim A_a$ の a 種類、B は $B_1 \sim B_b$ の b 種類に分類して、その度数を x_{ij} とおくとき、上記の表 13.6 の形式のデータが得られます。これを **分割表データ** といい、x_{ij} は <u>観測度数</u> と呼ばれます。また、A_iB_j のことを <u>セル</u> といいます。

13.3.2　分割表における仮説検定

1) 帰無仮説と対立仮説

帰無仮説 H_0：A と B に関係はない　(すべての i, j で、$p_{ij} = p_{i\bullet}p_{\bullet j}$)
対立仮説 H_1：A と B に関係がある　(ある i, j で、$p_{ij} \neq p_{i\bullet}p_{\bullet j}$)

ここで、$p_{ij} = P(A_iB_j),\ p_{i\bullet} = P(A_i),\ p_{\bullet j} = P(B_j)$ とします。

2) 期待度数

A_i の確率 $p_{i\bullet}$, B_j の確率 $p_{\bullet j}$ は、$\widehat{p}_{i\bullet} = \dfrac{T_{i\bullet}}{T},\ \widehat{p}_{\bullet j} = \dfrac{T_{\bullet j}}{T}$ の式で推定します。

帰無仮説の下では $\widehat{p}_{ij} = \widehat{p}_{i\bullet}\widehat{p}_{\bullet j}$ と考えると、起こると期待される度数（<u>期待度数</u>）が計算できます。

期待度数：$t_{ij} = T\widehat{p}_{ij} = \dfrac{T_{i\bullet} \times T_{\bullet j}}{T}$

3) 検定統計量と棄却域

この観測度数と期待度数の違いを、次の式で評価すると、帰無仮説の下で、自由度 $\phi = (a-1)(b-1)$ に近似的に従うことがわかっています。（章末問題 13.4 参照）

検定統計量：$\chi_0^2 = \displaystyle\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(x_{ij} - t_{ij})^2}{t_{ij}}$

そこで、次の棄却域に入れば帰無仮説を棄却して、A と B に関係があると判定します。

棄却域 R：$\chi_0^2 \geqq \chi^2(\phi, \alpha),\ \ \phi = (a-1)(b-1)$

13.4　適合度検定

適合度検定では、次の検定統計量を用います。

検定統計量：$\chi_0^2 = \displaystyle\sum_{i=1}^{k}\frac{(x_i - t_i)^2}{t_i}$

また、棄却域は次のように設定します。

棄却域 R：$\chi_0^2 \geqq \chi^2(\phi, \alpha),\ \ \phi = k - t - 1$

ここで、期待度数を求めるために、構造を考えて母数を推定する場合、その個数を t とします。