7. 適合度検定

概要:ある理論を考え、その理論の成否を確かめるために実験データを得たとする。適合度検定では、考えた理論がデータを説明できるのかを判定する。つまり、各グループに属するデータ数の比率が理論通りかを検定する。これは、各グループに入るデータ数から規格化値を構成し、その規格化値の二乗が χ^2 分布に従うとして導出する。

キーワード: 適合度検定;自由度;標準正規分布;二項分布; χ^2 分布

7.1. 序

ある現象に対してその理論値を提案したとする。そしてその理論が正しいのか対応する実験をする。その理論が実際のデータと一致しているかどうかを確かめたい。その検定を可能にするのが適合度検定である。

7.2. 二つのグループの場合

ここでは、まず、グループが二つの場合を考える。二つのグループのデータ数が x_1, x_2 であり、二つのグループに対応する理論比率が p_1, p_2 であったとする。全体のデータ数を N とする。すると、グループ 1,2 に対応する理論データ数 a_1, a_2 は以下のようになる。

$$a_1 = Np_1 \tag{1}$$

$$a_2 = Np_2 \tag{2}$$

ただし、

$$p_1 + p_2 = 1 \tag{3}$$

である。

ここで、以下の量を考える。

$$T = \frac{\left(x_1 - a_1\right)^2}{a_1} + \frac{\left(x_2 - a_2\right)^2}{a_2} \tag{4}$$

これは、変形すると以下のようになる。

$$
\begin{aligned}
T &= \frac{\left(x_1 - a_1\right)^2}{a_1} + \frac{\left(x_2 - a_2\right)^2}{a_2} \\
&= \frac{\left(x_1 - Np_1\right)^2}{Np_1} + \frac{\left(x_2 - Np_2\right)^2}{Np_2} \\
&= \frac{\left(x_1 - Np_1\right)^2}{Np_1} + \frac{\left[\left(N - x_1\right) - N\left(1 - p_1\right)\right]^2}{N\left(1 - p_1\right)} \\
&= \frac{\left(x_1 - Np_1\right)^2}{Np_1} + \frac{\left(x_1 - Np_1\right)^2}{N\left(1 - p_1\right)} \\
&= \left(x_1 - Np_1\right)^2 \left[\frac{1}{Np_1} + \frac{1}{N\left(1 - p_1\right)}\right] \\
&= \left[\frac{x_1 - Np_1}{Np_1\left(1 - p_1\right)}\right]^2
\end{aligned} \tag{5}
$$

ここで、p_1 はデータがグループ 1 に属する確率であるから、x_1 は二項分布に従う。二項分布の平均は Np_1 で分散は $Np_1\left(1-p_1\right)$ である。N が大きいとき、二項分布は正規分布近似できる。したがって、

$$
\frac{x_1 - Np_1}{Np_1\left(1 - p_1\right)} \tag{6}
$$

は標準正規分布に従う。T はその 2 次の量であるから、自由度 1 の χ^2 分布に従う。

7.3. 適合度検定の一般化

ここでは、2 つのグループを扱ったが、これは k 個のグループに拡張される。

k 個のグループのデータ数が $x_1, x_2, \cdots, x_k$ であり、k 個のグループに対応する理論比率が $p_1, p_2, \cdots, p_k$ であったとする。全体のデータ数を N とする。すると、グループ $1, 2, \cdots, k$ に対応する理論データ数 $a_1, a_2, \cdots, a_k$ は、$a_1 = Np_1, a_2 = Np_2, \cdots, a_k = Np_k$ となる。ただし、

$$
p_1 + p_2 + \cdots + p_k = 1 \tag{7}
$$

である。

ここで、以下の量を考える。

$$
\begin{aligned}
T &= \frac{(x_1-a_1)^2}{a_1} + \frac{(x_2-a_2)^2}{a_2} + \cdots + \frac{(x_k-a_k)^2}{a_k} \\
&= \sum_{i=1}^{k} \frac{(x_i-a_i)^2}{a_i}
\end{aligned}
\tag{8}
$$

Nが大きい場合、Tはその2次の量であるから、自由度$k-1$ のχ^2分布に従う。

ここで自由度とは、変化できる量の数のことで、データ数から制約条件の数を引いたものである。例えば、上の式においては、a_1、a_2、・・・、a_{k-1}と任意の値をとるとすると、全確率は1であるから、a_kは

$$
\begin{aligned}
a_k &= p_k N \\
&= \left[1-(p_1+p_2+\cdots+p_{k-1})\right]N
\end{aligned}
\tag{9}
$$

と定まる。したがって、変化できる数は$k-1$となる。

7.4.　最尤法による適合度検定

前節ではあまり詳細に立ち入らずに一般化の議論をしていた。ここでは最尤法を利用して詳細に議論し、理論を再現する。

k 種類の事象$E_1, E_2, \cdots, E_k$が互いに排他的に独立に起こるとする。我々は、N回の試行をし、事象$E_1, E_2, \cdots, E_k$ がそれぞれ $n_1, n_2, \cdots, n_k$ 回起こったとする。事象E_i が起こる確率をθ_iとする。 データ$(n_1, n_2, \cdots, n_k)$ を得る確率は幾何関数で

$$
f(n_1, n_2, \cdots, n_k; \theta) = \frac{N!}{n_1! n_2! \cdots n_k!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k} \tag{10}
$$

と与えられる。ここで、

$$
N = n_1 + n_2 + \cdots + n_k \tag{11}
$$

である。ここで、事象が起こった回数は$(n_1, n_2, \cdots, n_k)$であるから、対応する確率は以下となる。

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \left(\hat{\theta}_{1N}, \hat{\theta}_{2N}, \cdots, \hat{\theta}_{kN}\right) \\
&= \left(\frac{n_1}{N}, \frac{n_2}{N}, \cdots, \frac{n_2}{N}\right)
\end{aligned}
\tag{12}
$$

ただし、

$$\hat{\theta}_{iN} = \frac{n_i}{N} \tag{13}$$

である。これは、実際に起こったデータをもとにした確率である。

事象が独立であった場合、これらの起こる確率は理論的に評価できる。この場合は、各事象が起こる確率は分かっているとする。それを

$$\boldsymbol{\theta}_0 = \left(\theta_{10}, \theta_{20}, \cdots, \theta_{k0}\right) \tag{14}$$

とする。

この場合のデータ$\left(n_1, n_2, \cdots, n_k\right)$を得る確率は

$$f_0\left(n_1, n_2, \cdots, n_k; \boldsymbol{\theta}_0\right) = \frac{N!}{n_1! n_2! \cdots n_k!} \theta_{10}{}^{n_1} \theta_{20}{}^{n_2} \cdots \theta_{k0}{}^{n_k} \tag{15}$$

となる。

この実際に得られた確率と独立値を得る確率の比をλとすると、そえは以下になる。

$$\begin{aligned}
\lambda\left(n_1, n_2, \cdots, n_k\right) &= \frac{f_0\left(n_1, n_2, \cdots, n_k; \boldsymbol{\theta}_0\right)}{f\left(n_1, n_2, \cdots, n_k; \hat{\boldsymbol{\theta}}\right)} \\
&= \frac{\dfrac{N!}{n_1! n_2! \cdots n_k!} \theta_{10}{}^{n_1} \theta_{20}{}^{n_2} \cdots \theta_{k0}{}^{n_k}}{\dfrac{N!}{n_1! n_2! \cdots n_k!} \hat{\theta}_{1N}{}^{n_1} \hat{\theta}_{2N}{}^{n_2} \cdots \hat{\theta}_{kN}{}^{n_k}} \\
&= \left(\frac{\theta_{10}{}^{n_1}}{\hat{\theta}_{1N}{}^{n_1}}\right)\left(\frac{\theta_{20}{}^{n_2}}{\hat{\theta}_{2N}{}^{n_2}}\right) \cdots \left(\frac{\theta_{k0}{}^{n_k}}{\hat{\theta}_{kN}{}^{n_k}}\right)
\end{aligned} \tag{16}$$

実際に得られるデータが理論から予想されるものと一致していれば、それは 1 になるはずである。しかし、データは標本であるから、理想値からずれる。このずれが、許容できるかどうかを判断するのが検定である。

この比の対数をとって、-2 倍する量を考える。もしも、両者が同じであれば、0 になるが、標本であるから、同じにはならない。その同じにならない量が有意かを判断する。この量は以下になる

$$\begin{aligned}
-2\ln\left[\lambda\left(n_1, n_2, \cdots, n_k\right)\right] &= 2\sum_{i=1}^{k} n_i \left[\ln \hat{\theta}_{iN} - \ln \theta_{i0}\right] \\
&= 2\sum_{i=1}^{k} n_i \left[\ln\left(\frac{n_i}{N}\right) - \ln \theta_{i0}\right]
\end{aligned} \tag{17}$$

これをテーラー展開して以下を得る。

$$
\begin{aligned}
2\sum_{i=1}^{k} n_i \left[\ln\hat{\theta}_{iN} - \ln\theta_{i0}\right] &= 2\sum_{i=1}^{k} n_i \ln\left(\frac{\hat{\theta}_{iN}}{\theta_{i0}}\right) \\
&= 2\sum_{i=1}^{k} n_i \ln\left[\left(\frac{\hat{\theta}_{iN} - \theta_{i0} + \theta_{i0}}{\theta_{i0}}\right)\right] \\
&= 2\sum_{i=1}^{k} n_i \ln\left[1 + \left(\frac{\hat{\theta}_{iN} - \theta_{i0}}{\theta_{i0}}\right)\right] \\
&\approx 2\sum_{i=1}^{k} n_i \left[\frac{\hat{\theta}_{iN} - \theta_{i0}}{\theta_{i0}} - \frac{1}{2}\left(\frac{\hat{\theta}_{iN} - \theta_{i0}}{\theta_{i0}}\right)^2\right] \\
&= 2\sum_{i=1}^{k} n_i \left[\frac{\frac{n_i}{N} - \theta_{i0}}{\theta_{i0}} - \frac{1}{2}\left(\frac{\frac{n_i}{N} - \theta_{i0}}{\theta_{i0}}\right)^2\right] \\
&= 2\sum_{i=1}^{k} n_i \left[\frac{n_i - N\theta_{i0}}{N\theta_{i0}} - \frac{1}{2}\left(\frac{n_i - N\theta_{i0}}{N\theta_{i0}}\right)^2\right]
\end{aligned}
\tag{18}
$$

これは、さらに以下のように変形される。

$$
\begin{aligned}
&2\sum_{i=1}^{k} n_i \left[\frac{n_i - N\theta_{i0}}{N\theta_{i0}} - \frac{1}{2}\left(\frac{n_i - N\theta_{i0}}{N\theta_{i0}}\right)^2\right] \\
&= 2\sum_{i=1}^{k} \left[\left(n_i - N\theta_{i0}\right) + N\theta_{i0}\right]\left[\frac{n_i - N\theta_{i0}}{N\theta_{i0}} - \frac{1}{2}\left(\frac{n_i - N\theta_{i0}}{N\theta_{i0}}\right)^2\right] \\
&= 2\sum_{i=1}^{k} \left[\frac{\left(n_i - N\theta_{i0}\right)^2}{N\theta_{i0}} + \left(n_i - N\theta_{i0}\right) - \frac{1}{2}\frac{\left(n_i - N\theta_{i0}\right)^3 + N\theta_{i0}\left(n_i - N\theta_{i0}\right)^2}{N^2{\theta_{i0}}^2}\right] \\
&= 2\sum_{i=1}^{k} \left[\frac{\left(n_i - N\theta_{i0}\right)^2}{2N\theta_{i0}} + \left(n_i - N\theta_{i0}\right) - \frac{1}{2}\frac{\left(n_i - N\theta_{i0}\right)^3}{N^2{\theta_{i0}}^2}\right] \\
&= \sum_{i=1}^{k} \left[\frac{\left(n_i - N\theta_{i0}\right)^2}{N\theta_{i0}} + 2\left(n_i - N\theta_{i0}\right) - \frac{\left(n_i - N\theta_{i0}\right)^3}{N^2{\theta_{i0}}^2}\right]
\end{aligned}
\tag{19}
$$

Eq. (19)以下の項は下記に示すように 0 になる。

$$
\begin{aligned}
\sum_{i=1}^{k}\left(n_i - N\theta_{i0}\right) &= \sum_{i=1}^{k} n_i + N\sum_{i=1}^{k}\theta_{i0} \\
&= N - N \\
&= 0
\end{aligned}
\tag{20}
$$

また、3 次の項も正負の値をとるため、無視できる。(N が大きいときこれは対称分布に

近づくから 0 と近似できる。）したがって、この値は以下のように変形できる。

$$\begin{aligned}\chi^2 &\approx \sum_{i=1}^{k}\left[\frac{\left(n_i - N\theta_{i0}\right)^2}{N\theta_{i0}} + 2\left(n_i - N\theta_{i0}\right) - \frac{\left(n_i - N\theta_{i0}\right)^3}{N^2{\theta_{i0}}^2}\right] \\ &\approx \sum_{i=1}^{k}\frac{\left(n_i - N\theta_{i0}\right)^2}{N\theta_{i0}}\end{aligned} \tag{21}$$

ここで、$k=2$ の場合はこれは自由度 $k-1=1$ の χ^2 分布従うことが示された。任意の値 k で、成り立つとすると、あきらかに $k+1$ でも成り立つ。したがって、この確率変数は χ^2 分布に従う。

量 $N\theta_{i0}$ は理論値 a_i で置き換えることができる。また、実際に得られたデータ n_i を x_i で表す。すると、 χ^2 は以下となる。

$$\chi^2 = \sum_i \frac{\left(x_i - a_i\right)^2}{a_i} \tag{22}$$

この各項は正であるから、その和も正になる。もしも、x_i が理論値 a_i と同じであれば、この量は 0 になる。すなわち、この量が大きければ大きいほど、理論値からはずれており、小さければ理論値に近い、と判断できる。

k 個のデータがあったとすると、全ての確率の和を 1 であるという制約条件が一個あるから、この自由度は $k-1$ となる。

つまり、推定確率を設定し、自由度 $k-1$ の χ^2 分布の P 点と上の χ^2 の値を比較すればいい、ということになる。

7.5. クロス集計表への適合度検定の一般化

ここでは $k\times l$ クロス集計表の適合度検定を行う。この場合は、先の場合と異なるのは、行と列に対する比率が独立に定義されていることである。行に対する比率を p_i 、列に対する比率を q_j とする。また、全データ数を N とする。この場合は

$$p_1 + p_2 + \cdots + p_k = 1 \tag{23}$$

$$q_1 + q_2 + \cdots + q_l = 1 \tag{24}$$

である。この場合のセル(i,j)に対する理論値a_{ij}は

$$a_{ij} = Np_i q_j \tag{25}$$

である。セル(i,j)に対する実測値をx_{ij}とする。以下の量を考える。

ここで、以下の量と考える。

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(x_{ij}-a_{ij}\right)^2}{a_{ij}} \tag{26}$$

Nが大きい場合、Tはその 2 次の量であるから、これはは自由度$(k-1)(l-1)$のχ^2分布に従う。

つまり、推定確率を設定し、自由度$(k-1)(l-1)$のχ^2分布の P 点と上のχ^2の値を比較すればいい、ということになる。

7.6. 自由度に関する議論

ここでは、自由度に関する議論を改めてする。

データがクロス集計でない場合、データ数がk個ある場合、

$$\chi^2 = \sum_{i=1}^{k}\frac{\left(x_i-a_i\right)^2}{a_i} \tag{27}$$

と表される。この場合、トータルのデータ数はわかっているから、最後のデータ数は

$$x_k = N - \sum_{i=1}^{k-1} x_i \tag{28}$$

と定まる。したがって、その自由度ϕは

$$\phi = k-1 \tag{29}$$

となる。

データがクロス集計表である場合、

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(x_{ij}-a_{ij}\right)^2}{a_{ij}} \tag{30}$$

となる。この場合の全データ数は

$$kl \tag{31}$$

となる。

この場合の行を考える。周辺合計は分かっているから、各行の最後の数は決まってしまう。したがって、自由に変動できる数は

$$kl-k \tag{32}$$

となる。

次に列を考える。この場合も周辺合計は分かっているから、各列の最後の数は決まってしまう。この場合、最後の列に関してはすでに行で考えているから$l-1$列で考えればいい。つまり、変動できるデータ数、つまり自由度ϕは、

$$\begin{aligned}\phi &= kl-k-(l-1)\\ &= (k-1)(l-1)\end{aligned} \tag{33}$$

となる。

7.7. まとめ

この章のまとめを行う。

k水準の適合度検定は、

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - a_i)^2}{a_i}$$

を評価し、これが自由度$k-1$のχ^2分布に従うとして行う。ただし、a_iは水準iに対する理論値、x_iは水準iに対する実際のデータである。

$k \times l$クロス集計表の適合度検定は、

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(x_{ij} - a_{ij})^2}{a_{ij}}$$

を評価し、これが自由度$(k-1)(l-1)$のχ^2分布に従うとして行う。ただし、a_{ij}は水準(i,j)に対する理論値、x_{ij}は水準(i,j)に対する実際のデータである。