2. 自由度

概要: 統計では自由度が議論されることが多い。しかし、自由度とは何かを詳しく議論されることはない。ここでは、自由度とは何かを議論する。

キーワード:自由度

2.1. 序

これから多変量解析をしていくが、その解析の中で自由度が大事になってくる。

自由度とは、自由に定めることができる値の数のことである。もう少し詳しくいうと、ある代表値や合計値があるとき、自由に値を取れる数のことである。これは、標本分散のところで少しふれた。標本分散においては、それはそれほど重要ではなかった。その自由度が非常に大事になる場面がこれから登場するので、ここで自由度について触れておく。

2.2. 自由度の評価

例えば、サンプルサイズが 3 のデータ$\left(a,b,c\right)$から得られた平均が 4 であるとき、一つ目のa、二つ目のbは自由に決めることができる。しかし、a,bの値が決まれば、cの値は、平均が 4 であるから

$$c = 4\times 3 - \left(a+b\right) \tag{1}$$

と決まってしまう。つまり、自由に決めることのできる値の数は 1 つ減ることになる。上の例では、

$$\mu = \frac{1}{3}\sum_{i=1}^{3} x_i \tag{2}$$

と表現される。データは 3 個あるが、平均が 4 であることをあらかじめ分かっているから、この場合は自由度は$3-1=2$となる。つまり、自由度はデータの数から、制約条件の数を引いたものになる。

一般に、n個のデータがある場合

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

などが定義される。その$\bar{x}$を使い、その他の量を定義する場合がある。

n個のデータを用いているので、標本平均はnで割る。統計においては、

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{4}$$

などの量が評価される。この中の

$$\left(x_1-\bar{x}\right)^2+\left(x_2-\bar{x}\right)^2+\cdots+\left(x_n-\bar{x}\right)^2=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \tag{5}$$

をデータ数nで割って、平均の分散としている。しかし、n で割っているが、実際は$\bar{x}$を使っているので、n 個目のデータは

$$x_n=n\bar{x}-\sum_{i=1}^{n-1}x_i \tag{6}$$

と決まってしまう。標本分散S^2は同じn個のデータを用いているが、$\bar{x}$が含まれているので、自由に決めることのできる数が 1 個減って、自由度は$n-1$になる。つまり、

$$s^2=\frac{1}{n-1}\sum_{i=1}^{3}\left(x_i-\bar{x}\right)^2 \tag{7}$$

として、分散に利用する。つまり、nで割るのではなく、$n-1$で割る。

以下で示す期待値$E\left[X\right]$は無限回データXを取った場合に相当する。この期待値が求めるものと一致するようにする。つまり、

$$\sigma^2=E\left[s^2\right]=E\left[\frac{1}{n-1}\sum_{i=1}^{3}\left(x_i-\bar{x}\right)^2\right] \tag{8}$$

となるようにする。

これは以下のように証明される。

$$
\begin{aligned}
&E\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\\
&=E\left[\sum_{i=1}^{n}\left(x_i-\bar{x}-\mu+\mu\right)^2\right]\\
&=E\left[\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right]+E\left[\sum_{i=1}^{n}\left(\mu-\bar{x}\right)^2\right]+2E\left[\left(\mu-\bar{x}\right)\sum_{i=1}^{n}\left(x_i-\mu\right)\right]\\
&=n\sigma^2+E\left[\sum_{i=1}^{n}\left(\mu-\bar{x}\right)^2\right]-2E\left[\sum_{i=1}^{n}\left(\mu-\bar{x}\right)^2\right]\\
&=n\sigma^2-E\left[\sum_{i=1}^{n}\left(\bar{x}-\mu\right)^2\right]\\
&=n\sigma^2-E\left[\sum_{i=1}^{n}\left(\frac{1}{n}\left(x_1+x_2+\cdots+x_n\right)-\mu\right)^2\right]\\
&=n\sigma^2-E\left[\sum_{i=1}^{n}\left(\frac{1}{n}\left(x_1-\mu\right)+\frac{1}{n}\left(x_2-\mu\right)+\cdots+\frac{1}{n}\left(x_n-\mu\right)\right)^2\right]\\
&=n\sigma^2-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_1-\mu\right)^2\right]-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_2-\mu\right)^2\right]-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_2-\mu\right)^2\right]-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_n-\mu\right)^2\right]\\
&-2E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_1-\mu\right)\left(x_2-\mu\right)\right]-2E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_1-\mu\right)\left(x_3-\mu\right)\right]-\cdots-2E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_{n-1}-\mu\right)\left(x_n-\mu\right)\right]\\
&=n\sigma^2-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_1-\mu\right)^2\right]-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_2-\mu\right)^2\right]-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_2-\mu\right)^2\right]-E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(x_n-\mu\right)^2\right]\\
&-2E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_1-\mu\right)\right]E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_2-\mu\right)\right]-2E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_1-\mu\right)\right]E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_3-\mu\right)\right]-\\
&\cdots-2E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_{n-1}-\mu\right)\right]E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_n-\mu\right)\right]\\
&=n\sigma^2-E\left[\frac{n}{n^2}\frac{1}{n}\sum_{i=1}^{n}\left(x_1-\mu\right)^2\right]-E\left[\frac{n}{n^2}\frac{1}{n}\sum_{i=1}^{n}\left(x_2-\mu\right)^2\right]-E\left[\frac{n}{n^2}\frac{1}{n}\sum_{i=1}^{n}\left(x_2-\mu\right)^2\right]-E\left[\frac{n}{n^2}\frac{1}{n}\sum_{i=1}^{n}\left(x_n-\mu\right)^2\right]\\
&=n\sigma^2-n\times\frac{n}{n^2}\sigma^2\\
&=\left(n-1\right)\sigma^2
\end{aligned}
\tag{9}
$$

となる。ここで、x_i は独立としている。

したがって、

$$
\begin{aligned}
\sigma^2&=\frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\\
&\approx\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\\
&=s^2
\end{aligned}
\tag{10}
$$

となる。

2.3. 自由度の方程式

自由度が計算できない場合がある。例えば方程式

$$A = B + C \tag{11}$$

があった場合、これらの自由度は

$$\text{自由度}_A = \text{自由度}_B + \text{自由度}_C \tag{12}$$

の関係にある。この中の自由度 A,B が分かっており、自由度 C のみが分かっていない場合、この方程式を利用できる。

2.4. 自由度の具体例

ここでは、自由度が登場するのはそれほど多くはない。その場合の自由度はどう評価されるのか具体例を示す。

2.4.1. 不偏分散

不偏分散 s^2 については、すでに詳しくのべた。不偏分散の自由度はデータの数を n とすると、その自由度は $n-1$ となる。すなわち、不偏分散は

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{13}$$

となる。

2.4.2. 回帰直線

回帰を行う場合、1 種類の目的変数と p 種類の説明変数があるとすると

$$S_y{}^2 = S_Y{}^2 + S_e{}^2 \tag{14}$$

の関係がある。この詳しい解説は重回帰分析のところで行う。

ここで、$S_y{}^2$ は目的変数の標本分散で、$\bar{y}$ を使っているから、n 個のデータであれば、その自由度は $n-1$ である。$S_e{}^2$ は誤差の標本分散で、

$$\begin{aligned} e_k &= y_k - Y_k \\ &= y_k - \left(\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_p x_p\right) \end{aligned} \tag{15}$$

であり、係数$\hat{a}_0,\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_p$を使っているから、その自由度は$n-(p+1)$である。

回帰の自由度n_Yは

$$n-1=n_Y+n-(p+1) \tag{16}$$

より、

$$n_Y=p \tag{17}$$

となる。したがって、この場合の自由度は

$$\begin{cases} n_y=n-1 \\ n_e=n-(p+1) \\ n_Y=p \end{cases} \tag{18}$$

となる。

2.4.3. クロス集計表

k水準、l水準のクロス集計を行う場合がある。この場合、独立値を求める。独立値はそれぞれの確率値を持ち、各確率の和は1である。つまり、その自由度は$(k-1)(l-1)$である。この議論をする。

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、

$$\chi^2=\sum_{i=1}^{k}\frac{(x_i-a_i)^2}{a_i} \tag{19}$$

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

$$x_k=N-\sum_{i=1}^{k-1}x_i \tag{20}$$

と定まる。したがって、その自由度ϕは

$$\phi=k-1 \tag{21}$$

となる。

データがクロス集計表である場合、

$$\chi^2=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(x_{ij}-a_{ij})^2}{a_{ij}} \tag{22}$$

となる。この場合の全データ数は

$$kl \tag{23}$$

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

$$kl - k \tag{24}$$

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、

$$\begin{aligned} \phi &= kl - k - (l-1) \\ &= (k-1)(l-1) \end{aligned} \tag{25}$$

となる。

2.5. まとめ

ここではこの章の結果をまとめる。

自由度とは変動できる変数の数のことである。

典型的な例に対してその自由度を求めた。