研究員の眼

統計分析を理解しよう：自由度の概念と活用について

生活研究部 主任研究員　金　明中
(03)3512-1825　kim@nli-research.co.jp

自由度とは？

統計学を勉強する時、よく登場するのが自由度（degree of freedom）である。統計学の教科書では、自由度を、「自由に決めることができる値の数」、「観察値の数から推計値を除いた数」などと定義している。しかし、自由度の意味がよく理解されていないのは「何に対する（あるいは、何の）」自由度かを明確に示していないからではないかと考えられる。

例えば、平均の場合は身長に対する平均（身長の平均）、成績に対する平均（成績の平均）、収入に対する平均（収入の平均）といったように、平均の具体的な内容が示されていることに比べ、自由度は修飾語を付けないまま自由度だけに呼ばれているケースが多い。従って、「計算に対する（計算の）自由度」、「標本分散に対する（標本分散の）自由度」のように何に対する自由度なのかを明確に示すと、自由度に対する理解がより深まると考えられる。

自由度の定義から考える自由度

ここでは、自由度の定義「自由に決めることができる値の数」と「観察値の数から推計値を除いた数」に基づいて自由度の説明を試みる。

例えば、サンプルサイズが3のデータ（a、b、c）から得られた標本平均が4であるとき、1つ目のaの値と2つ目のbの値は自由に決めることができる。そこで、ここではaの値が3、bの値が5だと仮定しよう。すると、すでに標本平均が4であることが分かっているので、3つ目のcの値は「4」しか入れることができない。つまり、「自由に決めることができる値の数」が1つ減ることになるので、「計算の自由度」はサンプルサイズから1を差し引いた「3-1=2」となる。

$$\text{平均}=\frac{a+b+c}{n} \quad \Rightarrow \quad 4=\frac{3+5+c(?)}{3} \quad \Rightarrow \quad c=4$$

次は「観察値の数から推計値を除いた数」の定義から自由度を見てみよう。上記の例から説明すると、観察値の数は、a、b、c の 3 つであり、a、b、c から算出された「平均」は推計値である。そこで、観察値の数「3」から、推計値である平均「1」を除くと自由度は 2 になる。

標本分散、t 分布に利用される自由度

自由度は、標本から計算した標本分散等を求めるとき、推定・検定のために t 分布表や F 分布表を引くとき等に使われる。標本分散は、母分散を推定するためのものであり、式（2)のように書くことができる。標本分散の平方根は標準偏差になる。

式（1）$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

式（2）$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \rightarrow \quad s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

標本平均を求める時には式（1）のように全体の合計を n で割るのに、標本分散を求める時には式（2）のように n-1 で割るのが一般的である。なぜ、標本分散は n ではなく n－1 を用いるのだろうか？

上述した通り、独立した n 個のデータを用いて、推計値である「標本平均」を求めているので、「標本平均」は n で割るのが適切である。それに対して、標本分散の場合には、式（2）のように平方和（個々のデータと平均値の差を二乗した値の和）を求める式（式（３））に、推計値である標本平均（$\bar{x}$）が含まれているので、「自由に決めることができる値の数」が一つ減ることになる。言い換えると、データの各値（x_i）と標本平均（$\bar{x}$） との差である「偏差」の合計は 0（式（4））になるので、自由に入れることができる値の数が一つ制約されてしまう。そこで、標本分散を求める際には一般的に n ではなく n-1 が用いられている。

式（3）$\sum_{i=1}^{n} \dfrac{(x_i - \bar{x})^2}{n}$

式（4）
$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_{n-1} - \bar{x})(x_n - \bar{x}) \\ &= (x_1 + x_2 + \cdots + x_{n-1} + x_n) - n(\bar{x}) \\ &= \sum_{i=1}^{n} x_i - n(\bar{x}) = 0 \end{aligned}$$

また、t 分布表を引くときにも自由度が使われる。母集団が正規分布だと分かっている場合におけ

る母平均の区間推定には、「正規分布を用いた推定」と「ｔ 分布を用いた推定」がある。一般的に母分散の値が分かっていれば標準正規分布を利用して推定をするものの、現実には母平均の値が不明なら母分散の値も分からないので、母集団から標本を抽出して推定を行うのが普通である。そこで、母分散の値が分からず、サンプルサイズの小さい場合に、母平均の区間を推定する確率分布がｔ分布である。

つまり、標本平均$\bar{x}$の分布が、平均がμで、分散が$\frac{\sigma^2}{n}$である正規分布に従う場合（式（5））の母平均の統計的推定は、母分散の値を分かっている時には式（6）のように平均が0、分散が1である標準正規分布を、母分散の値を分かっていない時には、式（7）のように自由度がn−1のｔ分布を利用すれば良い。

式（5）$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$　標本平均$\bar{x}$の分布は、平均がμで、分散が$\frac{\sigma^2}{n}$である正規分布に従う。

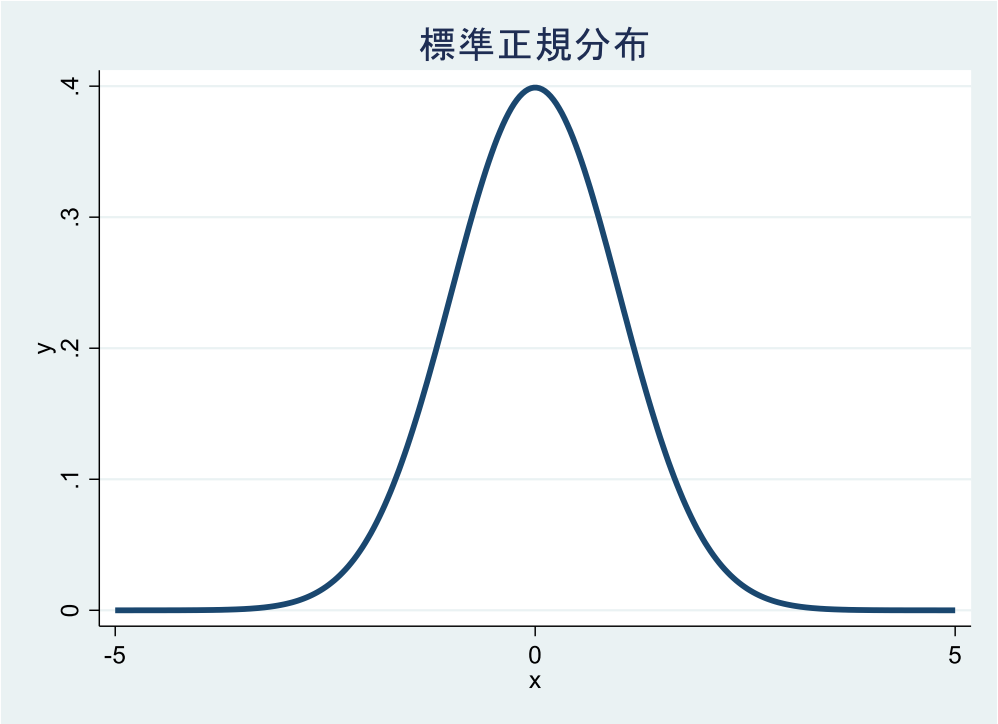
式（6）母分散の値を分かっている時　⇒　標準正規分布を利用：$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

式（7）母分散の値を分かっていない時　⇒　ｔ分布を利用：$\frac{\bar{x}-\mu}{s/\sqrt{n}}$　～　ｔ(n−1)

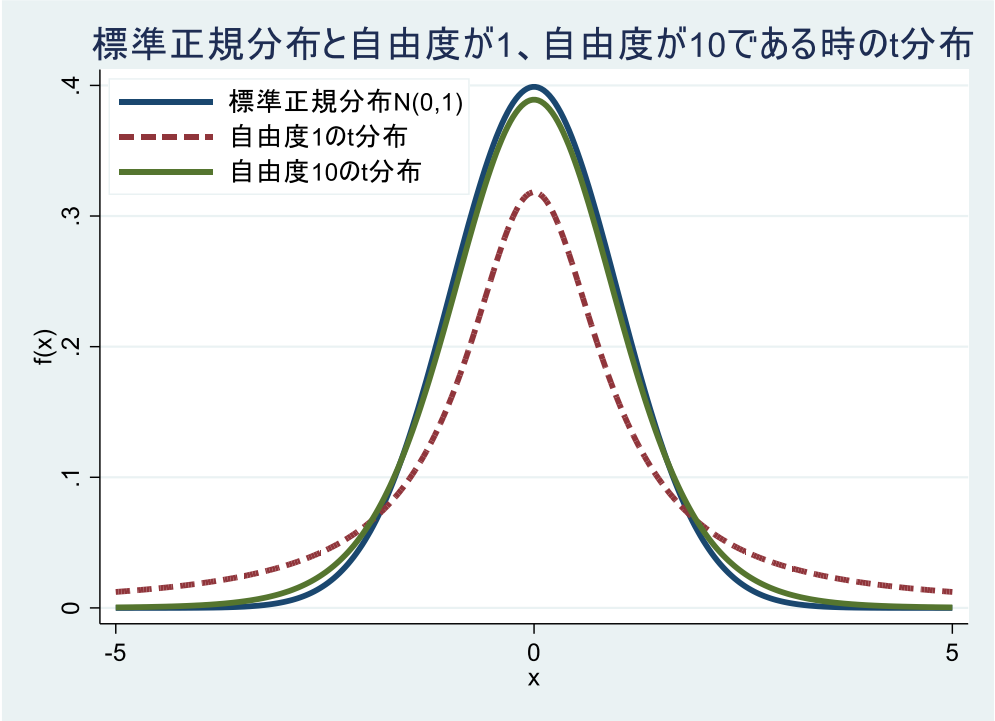
標準正規分布曲線は図表1のような形をしており、ｔ分布の曲線も図表1と似たような曲線になることが予想できる。なぜならば、式（6）と式（7）を見ると、異なるのはσとsだけだからである。式（6）のσは母集団の標準偏差であり、その値は分からない「常数」である。一方、標本の標準偏差であるsは、標本を抽出するたびにその値が変わるので「変数」である。つまり、式（6）と式（7）を比較すると、式（6）に比べ、式（7）の方が標本を抽出するたびに変わるので変動が大きいと言える。従って、ｔ 分布の曲線は標準正規分布曲線より分布が大きく、横軸に広がっている可能性が高い（図表2）。また、ｔ分布の曲線はサンプルサイズの影響を受ける。つまり、サンプルサイズｎが大きくなれば大きくなるほど、ｔ分布は標準正規分布に近くなる。図表2を見ると、自由度が1のｔ分布曲線より、サンプルサイズ（自由度）が大きい自由度10のｔ分布の方が分布が小さく、より標準正規分布に近い曲線になっていることが分かる。

※　サンプルサイズｎが大きくなれば大きくなるほど、sはσに近くなり、ｔ分布は標準正規分布とほぼ一致する。

図表１標準正規分布曲線



図表２標準正規分布曲線とｔ分布の曲線



本文で説明した自由度の基本概念を理解し、統計的推定等に有効に活用されることを願うところである。

