t 分布 (t-distribution)

オックスフォード大学で化学と数学の学位をとったウィリアム・ゴセットは、1899 年に老舗のビール製造会社ギネス社に採用された。そこで、彼は、麦芽汁の発酵させる酵母を精密に測定する仕事をしていた。ビールの味は、麦芽汁に投入する酵母の量で決まるのである。培養液の中で増殖を続ける酵母の数を測るためには、培養液からサンプルを取り出し、酵母細胞の数を数える必要がある。

今、n コのサンプルを取り出し、酵母細胞の数を数えたところ、x_1, x_2, $\cdots$, x_n の値が得られたとしよう。これらは、正規分布 $N(\mu, \sigma^2)$ から取り出された n コの標本とみなすことができる。知りたいことは、n コのサンプルから得られた値の平均 $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$ と真の平均 μ とのずれの大きさである。標準偏差 σ が分っていれば、

$$u = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

が標準正規分布 $N(0,1)$ に従うことから、ずれの大きさの確率が求まる。しかし、当然のことながら平均 μ も分らないのに分散 σ^2 の分かるはずもない。そこで、母分散 σ^2 の代わりに、標本不偏分散

$$s^2 = \frac{1}{n-1}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}$$

を用いることになる。すると、s^2 も誤差を含むので、分布

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

の含む誤差は、正規分布の場合よりも大きくなるであろう。ゴセットは、変数 t の従う確率分布を求め、スチューデントのペンネームでバイオメトリカ誌に発表した。そこで、この分布を**スチューデントの t 分布**と言う。本名ではなくペンネームを用いたのは、ギネス社は、社員が研究成果を発表することを禁じていたためである。ギネス社がこのような方針を採っていたのは、会社の重要な機密の漏洩を防ぐためであった。

ところで、

$$v = \left(\frac{x_1 - \bar{x}}{\sigma}\right)^2 + \left(\frac{x_2 - \bar{x}}{\sigma}\right)^2 \cdots + \left(\frac{x_n - \bar{x}}{\sigma}\right)^2$$

は、自由度 $n-1$ の χ^2 分布にしたがう。これを用いると、$s^2 = \sigma^2 v/(n-1)$ と書くことができる。そこで、t を u と v で表すと、

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{s} = \frac{u}{\sqrt{\dfrac{v}{n-1}}}$$

となる。ここで、u と v は、互いに独立な自由度であることに注意する必要がある。互いに独立であるとは、たとえ一方の値を決めたとしても、もう一方の値は決まらないことを意味する。

t 分布曲線

平均も分散も分からない正規分布から n コの標本を無作為抽出して作った変数 $t = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$ の確率分布は、標準正規分に従う変数 u と自由度 $n-1$ の χ^2 分布に従う変数 v をもちいて、$t = \dfrac{u}{\sqrt{v/(n-1)}}$ と書けるので、これら 2 つの分布を用いて表すことができる。

ここでは、簡単のために、変数 $t = \dfrac{u}{\sqrt{v/(n-1)}}$ の分布の代わりに、$t = \dfrac{u}{\sqrt{v/n}}$ の分布を求めることにしよう。

標準正規分布の確率密度関数を $g(u)$、自由度 n の χ^2 分布の確率密度関数を $T_n(v)$ と書くことにすると、

$$
\begin{aligned}
g(u) &= \frac{1}{\sqrt{2\pi}} e^{-u^2/2} && (1) \\
T_n(v) &= \frac{1}{2^{n/2}\Gamma(n/2)} v^{(n-2)/2} e^{-v/2} && (2)
\end{aligned}
$$

である。このとき、変数 $t = u/\sqrt{v/n}$ の従う確率分布関数 $f_n(t)$ は、

$$
f_n(t) = \int \delta\left(t - \frac{u}{\sqrt{v/n}}\right) g(u) T_n(v)\, du dv
$$

で与えられる。ここで u についての積分をおこなうのだが、その前に、u の代わりに $y = u/\sqrt{v/n}$ を用いて書き直してから積分をおこなうと、

$$
\begin{aligned}
f_n(t) &= \int \sqrt{\frac{v}{n}} \delta\left(t - y\right) g(\sqrt{v/n}\, y) T_n(v)\, dy dv \\
&= \int \sqrt{\frac{v}{n}} g(\sqrt{v/n}\, t) T_n(v)\, dv
\end{aligned}
$$

となる。ここで、(1) 式と (2) 式を代入すると、

$$
f_n(t) = \frac{1}{\sqrt{2\pi n}\, 2^{n/2}\Gamma(n/2)} \int_0^\infty v^{(n-1)/2} e^{-(1+t^2/n)v/2}\, dv
$$

さらに、$x = \left(1 + \dfrac{t^2}{n}\right)\dfrac{v}{2}$ とおいて x についての積分に書き直すと、

$$
f_n(t) = \frac{(1+t^2/n)^{-(n+1)/2}}{\sqrt{\pi n}\, \Gamma(n/2)} \int_0^\infty x^{(n+1)/2-1} e^{-x}\, dx
$$

となる。この右辺の積分は、ガンマ関数

$$
\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx
$$

をもちいて表すことができるので、

$$
f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\, \Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}
$$

となる。

さらに、ベータ関数

$$B(1/2,\, n/2) = \frac{\Gamma(1/2)\,\Gamma(n/2)}{\Gamma((n+1)/2)} = \frac{\sqrt{\pi}\,\Gamma(n/2)}{\Gamma((n+1)/2)}$$

を用いて表すと,

$$f_n(t) = \frac{1}{\sqrt{\pi}\,B(1/2,\, n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

となる。これを、自由度 n の t **分布**という。

正規母集団からサンプル数 n の標本を抜き出して作った変数 $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ は、自由度 $n-1$ の t 分布に従う。変数 t の中の標本平均 $\bar{x}$ と標本不偏分散 s^2 は、サンプルから計算できる。唯一不明なのは、母平均 μ である。そこで、t 分布は、母平均の検定や推定に使われる。

t 分布曲線を図 1 に示す。参考のために、標準正規分布 ($N(0,1)$ - 実線) も描いてある。サンプルの数が小さい場合の t 分布曲線は、標本不偏分散 s^2 の持つ誤差のために、幅の広い曲線となる。サンプル数が多くなると不偏標本分散 s^2 は、標準偏差 σ^2 に近づくので、曲線は、正規分布曲線と区別ができなくなる。別の言い方をすると、サンプル数が大きい極限では、中心極限定理により t 分布曲線は、正規分布曲線に一致する。サンプル数が 30 未満の場合を小標本と言い t 分布を用いる必要がある。サンプル数が 30 以上の場合を大標本と言う。この場合には、正規分布を用いることができる。

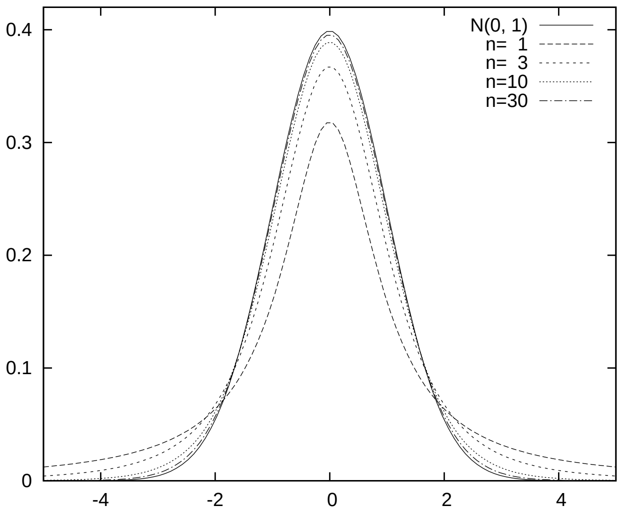

図 1: t 分布曲線 $y = f_n(t)$ と正規分布 $N(0,\,1)$