III-2-6. Student の t 分布

t 分布は student の t 検定に使われる確率分布です。Student の t 検定は、データの差の有意性の検定です。ですから、t 分布とは正規分布すると仮定されるあるデータと、あるデータの平均値の差の分布のことです。この確率分布は正規分布に似ていて、実際自由度が十分大きければ正規分布に近似できます。左右対称というところも正規分布に似ています。 正規分布と大きく違うところは、自由度によって分布が変わるところです。この点は χ^2 分布です。というのも、この確率分布が正規分布とχ^2分布の合成によってできているからです。

これを発見したのはビール会社ギネスの技師だった W.Gosset ですが、Student の t 分布という名前がついているのは、会社から論文投稿を禁じられていたために、彼が Student という筆名でその発見を論文化したからです。おそらく、彼の発見の動機は、今、私たちが抱いている疑問と同じだったろうと思います。 つまり、「正規分布することが想定されるデーターについて、データーから得られた平均値 がどのくらい正規分布の期待値（平均値・中央値）に近いのかは、母集団の標準偏差(σ)を尺度にして、 標準化して、標準正規分布 N(0,1)の分布の中でデーターの平均値がどの位置にあるのかを考 えれば良いというのはわかるにしても、そもそも母集団の標準偏差(σ)を知らないのだから、標準化す ることができません。データーから得られるのは標本分散から推定した母分散の推定値だ から確率的に変動する。その変動をどう読みこむのか。」、という疑問です。

我々の疑問を数式的に表し、それをどのように解決すればよいかを考えます。 まず、我々は正規分布というものの存在を認めています。u が標準正規分布 N(0,1)に従うの ならば、その確率は

$$N_{(0,1)} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-u^2}{2}}$$

次はuをどう求めるかですが、その作業が標準化と言われる作業で、実際のデーター、期待 値、分散（偏差）が使われます。 これは習った通り。

$$u = \frac{x - \mu}{\sigma}$$

なのですが、我々が疑問としているのは、「この式で使っているのは正規分 布している本来の理想的なデータの期待値=平均値（μ）と分散（$\sigma 2$）なのだから、実際のデ ータしか知らない私たちが、そんなものは知るわけがないだろう。」ということです。仮に、母集団の平均を知っていることにして、実際に得られた値と真の平均値との差を考えられる偏差を基準値としてその平方和を求めると、

$SS = v$として、

$$v = \sum_{i=1}^{m} \left(\frac{y_i - \mu}{\sigma_{se}} \right)^2$$

です。σ_{se}と添え字を付けたのは、まだ、y_iが何だかわからないからです。とりあえず、

something estimated だと覚えておいて下さい。そこれをよく見ると、σ_{se}が外から与えた期待値ならば、第一の式の総和記号の中はχ^2の定義式で、vはχ^2分布します。このこの場合、平均値という形で合計の値が縛られている のだから、χ^2の値を- 1 個目まで決めた後の最後の値は自動的に決まります。だから全体 の自由度は m-1 です。つまりvは自由度 m-1 のχ^2分布します。このuとvは互いに独立なのです （関係ない。uが変化してもvは変化しない。反対にvが変化してuの値に変化はない。とい う意味です。これが W.Gosset のやった最大の発見なのです。だから、２つの変数を合成し た変数を作り、その変数の確率を２つの確率の積として計算できるのです。)。私は、この２ つが独立かと問われると、一瞬、言葉に詰まってわからなくなります。でも落ちつて考え ると、確かに独立ですね。
そこで、２つの関数を関係づけるために、この２つの変数からできる合成関数を考えます。母集団の分散 σ は平方和を(n-1)で割ったものが推定になるというのをやりましたが。この場合は、μとσ_{se}を既知のものとして、外から与えているので、自由度は m です。

$$v = \sum_{i=1}^{m} \left(\frac{x_i - \mu}{\sigma_{se}}\right)^2 = \frac{1}{{\sigma_{se}}^2} \sum_{i=1}^{m} \left(x_{i-\mu}\right)^2$$

$${s_{se}}^2 = \frac{1}{n} \sum_{i=1}^{m} \left(x_{i-\mu}\right)^2$$

だから

$$s^2 = \frac{1}{n} v {\sigma_{se}}^2$$

$$s = \sqrt{\frac{v}{n}} \sigma_{se}$$

$$\frac{s}{\sigma_{se}} = \sqrt{\frac{v}{n}}$$

u = $x - \mu\,\sigma$ は正規分布しますが、問題は、分母つまり単位となる物差しの長さが、データーからもと めた偏差と真の偏差との間で違っているということです。その解決として「実際のデータ ーから求めた値に、実際の値から得た偏差と真の偏差の比を掛けたものを、合 成変数として、その値になる確率を考える。」というのが W.Gosset の提案内容です。
vはχ^2分布し、uは正規分布して、互いに独立ですから、２つを合成した関数変数の確率は両者の確率の積です。
具体的には、合成変数（ｔ:ｔ 値）は次の式です。

$$t = \frac{y - \mu}{\sigma_{se}} \cdot \frac{\sigma}{s}$$

$$t = u\sqrt{\frac{n}{v}}$$

確率変数uとvが同時にそれぞれ独立にある値をとって、ｔ の値が決まるので、uになる確率

W(u)とvになる確率 P(v)の積が t になる確率 S(t)です。

$$S(t) = W(u)P(v)$$

確率の総和は次の重積分で計算 できて、その値は 1 です（確率の総和は 1 ）。

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} W(u)P(v)\,dudv = 1$$

これを t と s の式に変換します。その際、$t = u\sqrt{\frac{n}{v}}$からヤコビアンをもとめます。ヤコビアンとは、正規分変数の変換によって生ずる。係数のようなものです。プログに解説を載せておきました。わからなければ読んでください。

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} W(u)P(v)\,dudv$$

$$= \int_{-\infty}^{\infty}\int_{0}^{\infty} J(u,v/t,s)W(u)P(v)\,dtds$$

vからsの変換ですが、sはvそのものでも良さそうだから、$s = v$として、

$$\frac{du}{dt} = \sqrt{\frac{v}{n}}, \qquad \frac{du}{ds} = 0, \frac{dv}{dt} = \frac{-2nu^2}{t^{-3}}, \frac{dv}{ds} = 1$$

$$J(u,v/t,s) = \begin{vmatrix} \sqrt{\frac{v}{n}} & 0 \\ \frac{-2nu^2}{t^{-3}} & 1 \end{vmatrix} = \sqrt{\frac{v}{n}}$$

$$W(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-u^2}{2}}$$

$$P(v) = P\left(\chi^2{}_{\varphi}\right) = \frac{\chi^2{}_{\varphi}^{\left(\frac{\varphi}{2}-1\right)} e^{\frac{-\chi^2{}_n}{2}}}{2^{\frac{\varphi}{2}}\Gamma\left(\frac{\varphi}{2}\right)}$$

なのですが、今、考えているのは、自由度は外から与えた平均値と分散を与えた時の確率分布ですから、自由度は$m-1$ ではなくてmなのです。W.Gosset が、考えた t 分布は、 2 つの群の平均値の差が、 0 という帰無仮説を立てて、実測された二つの群の差から t 値を計算し、帰無仮説から導かれる t 値=0 という期待値に対して、実測された t 値がどのような確率で存在するかという確率分布です。だから、自由度は 1 です。

$$P(v) = \frac{v^{\frac{-1}{2}} e^{\frac{-v^2}{2}}}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}$$

Γ ( )はガンマ関数（ガンマ関数については カイ二乗分布を作るときに説明しました。これらを使って、以下の重積分を変形します。

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} W(u)P(v)\,dudv$$

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-u^2}{2}}\frac{v^{\frac{-1}{2}}e^{\frac{-v^2}{2}}}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}dudv$$

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\sqrt{\frac{v}{n}}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-u^2}{2}}\frac{v^{\frac{-1}{2}}e^{\frac{-v^2}{2}}}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}dsdt$$

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sqrt{2n\pi}\sigma}e^{\frac{-u^2}{2}}\frac{e^{\frac{-v^2}{2}}}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}dsdt$$

かかなり長い計算になるので、省略します。興味があれば、ブログを読んでください。少し誤りがあるので、近々、修正しておきます。
最終的に、

$$\int_{-\infty}^{\infty}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)\left(\frac{t^2}{n^2}+1\right)^{\frac{n}{2}+1}}dt=1$$

となりますが
個々のｔの値をとる確率は、−∞からｔの値までの定積分を微分すれば良いので、積分記号の中の関数になって、

$$S(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)\left(\frac{t^2}{n^2}+1\right)^{\frac{n}{2}+1}}$$

です。この複雑な式は、β関数を使えば、簡略化して表現できます。β関数はガンマ関数の積で以下の式で表せます。

$$\beta(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \qquad \textit{式 22}$$

$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$ですから

$$\beta\left(\frac{n}{2},\frac{1}{2}\right)=\frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}=\frac{\pi\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}$$

で

$$S(t)=\frac{1}{\sqrt{n\pi}\beta\left(\frac{t^2}{n^2}+1\right)^{\frac{n}{2}+1}} \qquad \text{式 23}$$

となります。一般に student の t 値の確率密度関数として紹介されているのは、この式です。
話を元に戻します。

$$v = \sum_{i=1}^{m} \left(\frac{y_i - \mu}{\sigma_{se}}\right)^2$$

のところで、y_iは何だかわからないとしましたが、今はわかります。y_iは t 値です。μはその期待値で 0 です。つまり、グループ A とグループ B の平均値の差です。μは 0 なんだから、$\left(\frac{y_i-\mu}{\sigma_{se}}\right) = \left(\frac{y_A-y_B}{\sigma_{se}}\right)$です。

A の平均値y_Aが期待値 0 だとすれば、y_Bは期待値 0 の周辺で確率的に変動するでしょう。

また、B の平均値y_Bが期待値 0 だとすれば、y_Aは期待値 0 の周辺で確率的に変動するでしょう。これは、真の平均値がデータから得られた平均値の周りで変動するときの積率でもあります。データから推定された平均値mが真の平均値の周り変動するときの積率を平均誤差と言うということは、すでに説明しました。平均誤差は$\frac{\sigma}{\sqrt{n}}$です。つまり、２つの群の平均値から推定した、 t 値の分散は$\frac{\sigma^2}{n}$で偏差は$\frac{\sigma}{\sqrt{n}}$です。$\sigma_{se} = \frac{\sigma}{\sqrt{n}}$ つまり、σ_{se}(Something estimated)は S tandard error だという話です。t値の定義式を書き換えておきます。

$$t = \frac{y-\mu}{\sigma_{se}} \cdot \frac{\sigma}{s} = \frac{y-\mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{\sigma}{s} = \frac{y-\mu}{\frac{s}{\sqrt{n}}} \qquad \text{式 24}$$

sはサンプル集団の偏差

ポイントを要約すると、
t 分布は、正規分布する分散の等しい二つのデータ群の平均値の差から、 t 値を作って、その値がカイ二乗分布するとして、求めた確率分布で、データ群の差を標準化するために、データ群から求めた標準誤差で差を割ったもので。