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分布的です。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、母分散σ^2を尺度にして、標準化して、標準正規分布$\mathrm{N}(0,1)$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、そもそも母分散σ^2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。

さて、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$\mathrm{N}(0,1)$に従うのならば、その確率は

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

となるということに、今更、疑問をはさむことはないでしょう。

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。

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

確かのその通りなのですが、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（σ^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。

データを標準化するために、データから母集団の分散を推定するならば、

$$v=\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$\sigma^2 = \frac{v}{n-1}$$

となりますが、これをよく見ると、第一の式の総和記号の中はχ^2の定義式で、χ^2の和はχ^2分布するのだから、vはχ^2分布します。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n-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n-1です。つまりvは自由度n-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。標本集団の分散s^2を求めるときは、平方和（平均値からの距離の２乗の和）をnで割りますが、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ね。

$$v = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

ということです。ということは

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

$$s = \sigma\sqrt{\frac{v}{n}}$$

$$\frac{\sigma}{s} = \sqrt{\frac{n}{v}}$$

ここで合成関数tについて考えます。

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

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。これがｔで、独立したuとvの２つの変数を一つの変数にまとめているのです。具体的には以下の式です。

$$\mathrm{t} = \frac{\bar{x} - \mu}{\sigma} \cdot \frac{\sigma}{s}$$

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

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。

$$\mathrm{S(t)} = \mathrm{W}(u)P(v)$$

右辺のそれぞれの確率はたがいに独立して直交していますから、体積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）。

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

という重積分です。この立体をtと直交する平面（ｔに直交するs軸上の平面）で切り取った時の面積が、tとなる確率ということになります。

数学に詳しい人風にやるとすると、ヤコビアンと座標変換の知識が必要です。ヤコビアンについてはIII-3-3に書きましたが、座標変換するときの拡大倍率のようなものだと思ってください。

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

$\mathrm{t} = \mathrm{u}\sqrt{\frac{n}{v}}$からヤコビアンをもとめて

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

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

J(u, v/t, s)はヤコビアン、内側の積分範囲はよくわからないから適当です。sはvそのもので良いから、v=s、とします。

ヤコビアンを計算します。

$$\frac{du}{dt} = \frac{\sqrt{v}}{\sqrt{n}}, \quad \frac{du}{ds} = o, \ \frac{dv}{dt} =, \ \frac{dv}{ds} = 1$$

$$\mathrm{J} = \begin{bmatrix} \frac{du}{dt} & \frac{du}{ds} \\ \frac{dv}{dt} & \frac{dv}{ds} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{v}}{\sqrt{n}} & 0 \\ \frac{-2nu^2}{t^{-3}} & 1 \end{bmatrix} = \frac{\sqrt{v}}{\sqrt{n}}$$

$$\mathrm{J(u,v/t,s)} = \frac{\sqrt{v}}{\sqrt{n}}$$

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

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

Γ(　)はガンマ関数（ガンマ関数については III-2-5 カイ二乗分布の式 18 を参照してください。

これらを代入すると

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

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

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

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

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

から

$$= \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot s^{\frac{n-1}{2}} e^{\frac{-(\frac{t^2}{n^2}v+v)}{2}} dsdt$$

$$= \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot s^{\frac{n-1}{2}} e^{\frac{-(\frac{t^2}{n^2}+1)s}{2}} dsdt$$

$$= \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot s^{\frac{n-1}{2}} e^{-(\frac{t^2}{n^2}+1)\frac{s}{2}} dsdt$$

ここで、積分記号内の指数関数をwと置いて、計算を簡略化してまとめます。

$$(\frac{t^2}{n^2}+1)\frac{s}{2} = w$$

$$\frac{dw}{ds} = (\frac{t^2}{n^2}+1)\frac{1}{2}$$

$$s = \frac{2w}{(\frac{t^2}{n^2}+1)}$$

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\cdot s^{\frac{n-1}{2}}e^{-(\frac{t^2}{n^2}+1)\frac{s}{2}}dsdt$$

$$=\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\cdot\left(\frac{2w}{(\frac{t^2}{n^2}+1)}\right)^{\frac{n-1}{2}}e^{-w}\frac{2}{(\frac{t^2}{n^2}+1)}dwdt$$

$$=\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{1}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\cdot\left(\frac{2}{(\frac{t^2}{n^2}+1)}\right)^{\left(\frac{n+1}{2}\right)}w^{\frac{n-1}{2}}e^{-w}dwdt$$

$(\frac{t^2}{n^2}+1)\frac{s}{2}=w$としたときに、sはtと互いに独立した変数で、sを$w$の関数とすれば、tしか含まれない$(\frac{t^2}{n^2}+1)$は$w$による内側の積分記号の外に出せます。

$$=\int_{-\infty}^{\infty}\frac{2^{\frac{n}{2}+1}}{\sqrt{2n\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)(\frac{t^2}{n^2}+1)^{\frac{n}{2}+1}}\int_{0}^{\infty}\cdot w^{\left(\frac{n+1}{2}-1\right)}e^{-w}dwdt$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)(\frac{t^2}{n^2}+1)^{\frac{n}{2}+1}}\int_{0}^{\infty}\cdot w^{\left(\frac{n+1}{2}-1\right)}e^{-w}dwdt$$

ところで、内側の積分記号は　Γ関数の形をしていて、$\Gamma\left(\frac{n+1}{2}\right)$です。

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

この積分の意味は、tに沿って$-\infty$から∞まで積分すれば、確率の総和１になるということですから、tの個々の値の確率は、この積分関数（積分の中の関数）ということになって

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

これが求める答えですが、この答えはβ(ベータ)関数という関数を使って、さらに簡略化して書くことができます。β関数はΓ関数を組み合わせたもので。

$$\beta(p,q))=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

式22

です。

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

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

$$S(t) = \frac{1}{\sqrt{n}\mathrm{B}\left(\frac{n}{2},\frac{1}{2}\right)\left(\frac{t^2}{n^2}+1\right)^{\frac{n}{2}+1}}$$

式23

この説明では、

$$\mathrm{t} = \frac{\bar{x}-\mu}{\sigma}\cdot\frac{\sigma}{s}$$

としています。ふつうの教科書には

$$\mathrm{t} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\cdot\frac{\sigma}{s}$$

式24

と書いてあります。その通りです。悩まないでください。教科書通りにやればOKです。これは、平均値の推定値の分散（標準誤差）が$\sqrt{n}$に反比例するからです。このことは、すでに２項分布のところで確かめてありますね。ｔ検定はデータから推測された平均値の差の検定ですから、全体データーのばらつきを表す偏差ではなくて、標準誤差$\frac{\sigma}{\sqrt{n}}$を使うのです。こうして求められたt^2も真の平均値（期待値）の周りにχ２乗分布しているのです。