5日間で理解するt分布

浜田　宏

東北大学文学部

このノートの目的は統計学でよく使う（その割に，それが実際になんであるか多くの人にとってよくわからない）《t分布》の理解を深めることです．t分布の導出をフォローする過程で，微積分と線形代数と確率論の理解も深めたいと思います[1]．

みなさんt分布はご存じでしょうか？ 統計学のテキストに必ず登場する分布の一つで，仮説検定でよく使われます．t分布のイメージとして多くの人は「正規分布みたいなもの」「サンプルサイズが小さい時に正規分布の代わりに使うもの」といった特徴を思い描くでしょう．あるいはもう少し統計の応用に慣れている人は，「母分散が未知な場合に使う推定量がしたがう分布」というイメージを持っているかもしれません．

t分布とは

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad (n \geq 1)$$

という確率密度関数を持つ連続確率変数のことです．独立な標準正規分布Xと自由度nのχ^2分布Yの比でつくった確率変数

$$T = \frac{X}{\sqrt{Y/n}}$$

は自由度nのt分布に従います．

と，言われても……．

…… ちょっとなに言ってるか，わからない．

と思いませんか？

[1] このノートは浜田宏．2020『その問題，やっぱり数理モデルが解決します』第11章の補足資料です．私自身の理解不足で間違ったことを書いているかもしれません．その際はご指摘いただけると助かります．ver.1.0.; 2020年10月24日公開．ver.1.0.1.; 誤字修正．ver.1.1.; 5節に独立性の確認を追加．ver.1.2.; 3.2節例5の計算を簡略化．ver.1.3.; 5節の独立性の確認を修正． ver.1.4.; 5節に平均の差の検定を追加．

私は t 分布の定義をはじめて見たとき，頭に無数の？ が浮かびました．

「あのややこしい確率密度関数はどこからでてきたのか？」「標準正規分布 X と自由度 n の χ^2 分布 Y の比で確率変数をつくるとは，どういうことか？」このような疑問を持つのは当然でしょう．

こうした疑問を解消するために，本ノートは次のような流れで構成しています．

1 日目 数学的準備（合成関数の微分と置換積分法）

2 日目 確率変数の変換と合成

3 日目 χ^2 分布の導出[2]

4 日目 t 分布の導出

5 日目 統計量と t 分布の関係

まず，1 日目は計算に必要な数学の確認です．高校で習う微積分 $+\alpha$ なので，知っている人はとばしてください．

2 日目は，確率変数の変換と合成（確率変数同士の演算）です．たとえば X の分布が既知であるとき，$Y = aX + b$ がどんな分布にしたがうのか？ $Z = X + Y$ とおくとき，Z はどんな分布に従うのか？ といった問題を考えます．

3 日目は，χ^2 分布です．χ^2 分布も t 分布同様に，統計学のテキストに必ずと言っていいほど登場するにもかかわらず，知名度ほどには読者の理解が追いつかない分布の 1 つです．そこで，t 分布の導出に必要な χ^2 分布をどうやってつくるのかを確認します．なお，ここで直交行列による確率変数ベクトルの変換というおもしろい操作を使います．

4 日目は，2 日目で確認した確率変数同士の合成と，3 日目に確認した χ^2 分布を使って，t 分布を導出します．

5 日目は，最後に統計の応用場面で実際によく使う t 分布の利用法を確認します．具体的には平均にかんする仮説検定と，回帰分析の OLS 推定量の仮説検定で，t 分布が使える理由を確認します．

なお，このノートは「5 日間で理解する」と題していますが，私自身が学生だった頃，ここに書かれている内容を理解するのに半年から 1 年はかかっていたと思います．ですので自分が納得できるまで，時間をかけて読んでいただければ幸いです．

[2] χ はギリシア文字でカイと読みます．χ^2 の読み方は《カイ 2 乗》あるいは《カイ自乗》です．

1　数学的準備

1.1　対数関数の微分

対数関数の微分は，いろいろな場面で使います．

命題 1 (対数関数の微分)**.**

$$(\log x)' = \frac{1}{x}$$

証明.

$$\begin{aligned}\frac{\log(x+h)-\log(x)}{h} &= \frac{1}{h}\log\frac{x+h}{x} \\ &= \frac{1}{x}\times\frac{x}{h}\times\log\left(1+\frac{h}{x}\right) \\ &= \frac{1}{x}\log\left(1+\frac{h}{x}\right)^{\frac{x}{h}}\end{aligned}$$

ここで極限をとると

$$\begin{aligned}\lim_{h\to 0}\frac{1}{x}\log\left(1+\frac{h}{x}\right)^{\frac{x}{h}} &= \lim_{t\to\infty}\frac{1}{x}\log\left(1+\frac{1}{t}\right)^{t} \\ &= \frac{1}{x}\log e = \frac{1}{x}\end{aligned}$$

ここで e は《自然対数の底》と呼ばれる無理数 $2.7182818285\cdots\cdots$ です (矢野・田代 1993: 83-84)．　□

1.2　合成関数の微分

合成関数の微分もよく使います．まずは例から確認しましょう．

例 1. 2 つの関数 $z = 2y + 5, y = x^2$ があると仮定します．このとき z を x で微分するには，一旦

$$z = 2y + 5 = 2(x^2) + 5$$

と代入してから dz/dx を計算すれば OK です．つまり

$$\frac{dz}{dx} = 4x.$$

です．いま，ためしに

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

とおいて，計算してみましょう．すると

$$\frac{dz}{dy}\frac{dy}{dx} = 2 \cdot 2x = 4x$$

となり，計算結果が一致します．このことは偶然ではありません．これは，一般的な命題（合成関数の微分）として常に

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

が成立するからです．

合成関数の微分にはどんな便利な点があるのでしょうか．例えば

$$z = \log_e y, \quad y = \sqrt{3x}$$

のような関数だと，代入して計算することが簡単ではありません．一方，合成関数の微分を使えば，それぞれの導関数を掛けあわせるだけなので，計算が比較的容易です．

では命題の証明を確認しておきましょう．

命題 2 (合成関数の微分)**.** 関数 $y = f(x), z = g(y)$ がそれぞれ x, y について微分可能ならば

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = g'(y)f'(x)$$

が成立する

証明. 導関数の定義から

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = f'(x)$$

なので，極限をとらない場合は

$$\frac{f(x+h)-f(x)}{h} = f'(x) + \varepsilon.$$

分母を払うと

$$f(x+h) - f(x) = (f'(x) + \varepsilon)h$$

です．いま x の増分 Δx に対する y の増分を Δy, そして y の増分 Δy に対する z の増分を Δz とおきます．

$$\Delta y = f(x+\Delta x) - f(x) = (f'(x) + \varepsilon_1)\Delta x$$
$$\Delta z = g(y+\Delta y) - g(y) = (g'(y) + \varepsilon_2)\Delta y$$

ここで，微分可能の仮定から, $\varepsilon_1 \to 0(\Delta x \to 0), \varepsilon_2 \to 0(\Delta y \to 0)$ です.

2 つめの式 Δy の部分に 1 つめの式を代入すると，

$$\Delta z = (g'(y) + \varepsilon_2)(f'(x) + \varepsilon_1)\Delta x$$

両辺を Δx でわると

$$\frac{\Delta z}{\Delta x} = (g'(y) + \varepsilon_2)(f'(x) + \varepsilon_1)$$

ここで $\Delta x \to 0$ と仮定すると，$f(x)$ が連続なので $\Delta y \to 0$ である．よって $\Delta x \to 0$ のとき $\varepsilon_1 \to 0, \varepsilon_2 \to 0$ なので

$$\frac{\Delta z}{\Delta x} = (g'(y) + \varepsilon_2)(f'(x) + \varepsilon_1) \to g'(y)f'(x)$$

です (矢野・田代 1993: 82). □

1.3 置換積分法

確率の計算には積分を使います．そして積分の計算でとても便利な命題の 1 つが《置換積分法》です．この計算法はさまざまな場面で役立つので，この機会に是非覚えてください.

まず置換積分法の具体的な例を示します.

例えば

$$\int_0^1 4x(2x^2 - 3)^4 dx$$

という定積分を考えます．これは関数 $4x(2x^2 - 3)^4$ グラフを x 軸の 0 から 1 までの範囲で積分するという意味で，図中のグレーで着色した部分の面積を求める操作と一致します.

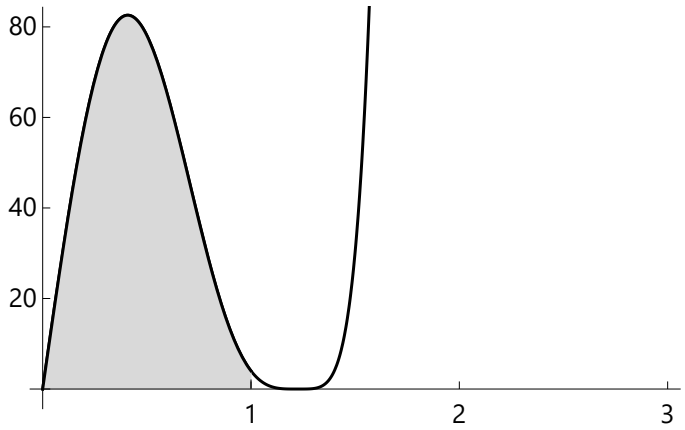

図 1: $4x(2x^2 - 3)^4$ のグラフ

次に，別の関数 t^4 のグラフの面積を -3 から -1 の範囲で求める計算を考えます．

図 2: t^4 のグラフ

この図形の面積は次の定積分と一致します．

$$\int_{-3}^{-1} t^4 dt$$

さて図 1 と図 2 では関数が異なるので，グラフの形状は違います．またグレーで色づけした部分の面積も，一見したところ異なるように感じます．

《置換積分法》とは，上に並べた二つのグラフ (図 (1) と図 (2)) の面積が，「ぴたりと一致する」事実を示した命題です．式で書けば

$$\int_0^1 4x(2x^2-3)^4 dx = \int_{-3}^{-1} t^4 dt = \frac{242}{5}$$

です．

このように「ある関数を別の関数に変換して計算しても面積が一致すること」を保証してくれる便利な命題が，置換積分法です．

以下に，一般的な命題としての《置換積分法》を示します．

命題 3 (置換積分法)**.** 関数 $f(x)$ が区間 $[a,b]$ で連続であり $x = g(t)$ が連続な導関数 $dx/dt = g'(t)$ を持ち，t の値 α, β に対して $\alpha = g^{-1}(a), \beta = g^{-1}(b)$ ならば

$$\int_a^b f(x)dx = \int_\alpha^\beta f(g(x))g'(t)dt$$

が成り立つ．

証明. この命題の証明には初日で復習した《合成関数の微分》を使います．まず $f(x)$ の不定積分を $F(x)$ とおけば, $F'(x) = f(x)$ だから，合成関数の微分により

$$\frac{d}{dt}F(g(t)) = \frac{dF(x)}{dx}\frac{dx}{dt} = f(x)g'(t) = f(g(t))g'(t)$$

です．これを区間 $[\alpha, \beta]$ で積分すると

$$\begin{aligned}\int_\alpha^\beta f(g(x))g'(t)dt &= [F(g(t))]_\alpha^\beta \\ &= F(g(\beta)) - F(g(\alpha)) = F(b) - F(a) \\ &= \int_a^b f(x)dx\end{aligned}$$

です (矢野・田代 1993: 101). □

一般的な命題とその証明では計算過程を少しイメージしにくいかもしれません．そのような場合は，具体的な関数を使って命題が成立するかどうかを確かめてみましょう．先ほど例に挙げた

$$\int_0^1 4x(2x^2-3)^4dx$$

の計算に置換積分法を使います.

$$x = \sqrt{\frac{t+3}{2}}$$

とおけば,

$$\frac{dx}{dt} = \frac{1}{2\sqrt{2}\sqrt{t+3}}$$

です (ここでまた合成関数の微分を使いました)．また x が 0 から 1 に動くとき, t は $t = -3$ から $t = -1$ まで動きますので置換積分法によって積分の範囲は -3 から -1 に変化します.

これで準備が整ったので，$x = \sqrt{\frac{t+3}{2}}$ を代入して，置換積分法を適用します.

$$\begin{aligned}\int_0^1 4x(2x^2-3)^4dx &= \int_0^1 3\sqrt{\frac{t+3}{2}}\left(2\left(\sqrt{\frac{t+3}{2}}\right)^2-3\right)^4\frac{dx}{dt}dt \\ &= \int_0^1 4\sqrt{\frac{t+3}{2}}\left(2\left(\frac{t+3}{2}\right)-3\right)^4\frac{1}{2\sqrt{2}\sqrt{t+3}}dt \\ &= \int_0^1 \frac{4}{\sqrt{2}\cdot 2\sqrt{2}}\left(t+3-3\right)^4 dt \\ &= \int_{-3}^{-1} t^4 dt\end{aligned}$$

です．置換積分法のおかげで被積分関数が計算しやすい簡単な形になりました．これを計算すると

$$\int_{-3}^{-1} t^4 dt = \frac{242}{5}$$

です．

以上で初日の内容は終了です．お疲れ様でした．

2　確率変数の変換と合成

2.1　確率変数の変換

ここからは，確率変数の便利な性質をいくつか確認しておきます．よく使う正規分布を例に考えてみましょう．

まず確率変数 X が平均 μ, 分散 σ^2 の正規分布に従うと仮定します．記号ではこれを，

$$X \sim N(\mu, \sigma^2)$$

のように書きます．これは，確率変数 X の確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

であることを意味します．次に記号 $P(X < c)$ を

$$P(X < c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

と定義します．$P(X < c)$ は，確率変数 X の実現値が c より小さい確率を表しています．いま

$$Y = aX + b$$

という変換を考えると，Y の分布はどうなるでしょうか？

$Y = aX + b$ を変形すれば $X = (Y - b)/a$ なので，X の確率密度関数に $x = (y - b)/a$ をためしに代入してみましょう．すると

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\frac{y-b}{a}-\mu)^2}{2\sigma^2}\right\}$$

となりますが，実はこれだけでは確率変数の変換はうまくいきません．右辺の関数は確率密度関数の定義（定義域の範囲で積分すると 1 になる）を満たしていないからです．

X は確率変数ですから，$P(-\infty < X < \infty) = 1$ です．変換後の Y も $P(-\infty < Y < \infty) = 1$ が成り立ってくれないと困ります．

そこで変換後の Y も $P(-\infty < Y < \infty) = 1$ となるように，Y の確率密度関数を計算しなければなりません．そのために《置換積分法》を使います．

置換積分法は，面積を変えずに被積分関数を別の表現へと変換する方法でした．したがってある確率密度関数 $f(x)$ の面積が 1 であるならば，総面積を変えずに別の確率密度関数へと変換することができるはずです．

さっそく，正規分布に従う確率変数の変換を置換積分を利用して計算してみましょう．

確率変数 X の実現値が c より小さい確率を積分

$$P(X<c)=\int_{-\infty}^{c}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx$$

で表します（再掲）．

$$y=ax+b \iff x=\frac{y-b}{a}$$

という変数変換を考えるので，

$$\frac{dx}{dy}=\frac{1}{a}$$

です．また x が $-\infty$ から c まで動くとき，$y=ax+b$ は $-\infty$ から $y=ac+b$ まで動きます．ゆえに置換積分法により，

$$\begin{aligned}
P(X<c)&=\int_{-\infty}^{c}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx\\
&=\int_{-\infty}^{ac+b}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\frac{y-b}{a}-\mu)^2}{2\sigma^2}\right\}\frac{dx}{dy}dy\\
&=\int_{-\infty}^{ac+b}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\frac{1}{a^2}(y-b-a\mu)^2}{2\sigma^2}\right\}\frac{1}{a}dy\\
&=\int_{-\infty}^{ac+b}\frac{1}{\sqrt{2\pi a^2\sigma^2}}\exp\left\{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}\right\}dy\\
&=P(Y<ac+b)
\end{aligned}$$

です．最後の Y の確率密度関数をみれば，

$$Y\sim N(a\mu+b,a^2\sigma^2)$$

であることが分かります．

このことを一般的な命題として表現しておきましょう．

命題 4 (正規分布の一次変換)**.** 正規分布 $X\sim N(\mu,\sigma^2)$ を $Y=aX+b$ と変換すると，Y の分布は正規分布 $N(a\mu+b,a^2\sigma^2)$ にしたがう

例 2. $X\sim N(0,1)$ のとき $Y=2X+1$ とおいて変換すると

$$Y\sim N(1,4)$$

である．このとき，たとえば確率 $P(1<X<2)$ は

$$P(1<X<2)=P(3<Y<5)$$

と等しい．

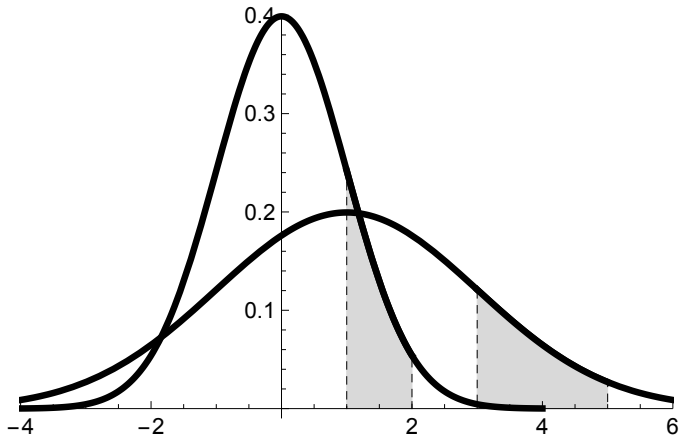

図 3: $X \sim N(0,1)$ のとき $P(1 < X < 2)$ を $Y = 2X + 1$ とおいて変換した $P(3 < Y < 5)$ のグラフ．グレーの部分の面積は一致する．

なお確率変数の変換で, より重要な演算が

$$Z = X + Y$$

という《確率変数同士を足す》タイプの合成です．この和は確率論と統計学のいろんな場面で登場します．その計算には 1 変数の置換積分法を 2 変数以上に一般化した次の《変数変換定理》を用います．

2.2 確率変数の合成

命題 5 (変数変換定理)**.** 確率変数の組 (X, Y) と (U, V) に

$$\begin{cases} X &= g(U,V) \\ Y &= h(U,V) \end{cases}$$

という関係があるとき, (x, y) 上の確率分布 $p(x, y)$ は，(u, v) 上の確率分布 $q(u, v)$

$$\begin{aligned} q(u,v) &= p(x,y) \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \\ &= p(g(u,v), h(u,v)) \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \end{aligned}$$

に変換される (小針 1973)**.**

この命題に登場する

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

という記号は

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

という行列の行列式（determinant）です．この行列式をヤコビアンと呼びます．ヤコビアンの記号として

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right|$$

という省略記号も使うので覚えておいてください．

例 3. 確率変数の組 (X,Y) から $U=X+Y$ という和を作り，U の分布を知りたいとします．変数変換定理を使って，確率分布 $p(x,y)$ を $q(u,v)$ に変換して，$q(u,v)$ を周辺化して U の分布を計算してみましょう．まず，

$$\begin{cases} u &= x+y \\ v &= y \end{cases}$$

という二組の関係を仮定します（v はあとで $\frac{\partial y}{\partial u}, \frac{\partial y}{\partial v}$ を計算しやすいように適当に定義しました）．これを x,y についての式と見なせば

$$\begin{cases} x &= u-v \\ y &= v \end{cases}$$

と書けます．変数変換定理における $X=g(U,V)$ を $x=u-v$，$Y=h(U,V)$ を $y=v$ と見なす，と解釈してください．確率分布 $p(x,y)$ は，変換定理によって確率分布 $q(u,v)$ に移るので

$$q(u,v)=p(g(u,v),h(u,v))\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = p(u-v,v)\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

です．

ここで行列の要素である偏導関数

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

をそれぞれ計算します．

$$\frac{\partial x}{\partial u} = 1 \qquad \frac{\partial x}{\partial v} = -1$$
$$\frac{\partial y}{\partial u} = 0 \qquad \frac{\partial y}{\partial v} = 1$$

したがって行列式は

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1 \cdot 1 - (-1) \cdot 0 = 1$$

なので

$$q(u, v) = p(u - v, v) \cdot 1$$

です．最後に U の確率密度関数 $\varphi(u)$ 取り出すために周辺化します．

$$\varphi(u) = \int_{-\infty}^{\infty} p(u - v, v) dv$$

結局，$U = X + Y$ の分布は同時分布 $p(x, y)$ を使って

$$\int_{-\infty}^{\infty} p(u - v, v) dv$$

と計算できる，というわけです．

この例を一般的な形で書くと，次のように表現できます．

命題 6. 同時確率分布 $p(x, y)$ があるとき，$U = X + Y$ の分布 $\varphi(u)$ は

$$\varphi(u) = \int_{-\infty}^{\infty} p(u - v, v) dv$$

によって定まる (小針 1973)．

(X, Y) が独立な場合の例も確認しておきましょう．

例 4. つぎの独立な 2 つの連続一様分布の合成を考え，その確率密度関数を示します．すなわち

$$X \sim \text{Uniform}(0, 1), Y \sim \text{Uniform}(0, 1)$$

であるとき $U = X + Y$ を作り U の確率密度関数を特定します．

X, Y の確率密度関数をそれぞれ $f_X(x), f_Y(y)$ とおきます．

$$f_X(x) = \begin{cases} 1 & , 0 \leq x \leq 1 \\ 0 & , \text{その他} \end{cases}, \qquad f_Y(y) = \begin{cases} 1 & , 0 \leq y \leq 1 \\ 0 & , \text{その他} \end{cases}$$

つぎに

$$\begin{cases} u &= x+y \\ v &= y \end{cases} \qquad \begin{cases} x &= u-v \\ y &= v \end{cases}$$

とおいて，変数変換定理を適用します．ヤコビアンは 1 です．求める u の確率密度関数を $g(u)$ とおいて

$$g(u) = \int_{-\infty}^{\infty} f_X(u-v) f_Y(v) dv$$

を計算します．ここで確率密度関数はそれぞれ

$$f_X(u-v) = 1, f_Y(v) = 1$$

だから

$$\int_{-\infty}^{\infty} 1 \cdot 1 dv = [v]_{-\infty}^{\infty} = \cdots\cdots \text{ナンジャコレ?}$$

と考えると，うまくいきません．

一様分布の確率密度関数は変数の範囲によって形が変わるので，そのことを考慮する必要があります．$0 \leq v \leq 1$ のとき $f_Y(v) = 1$ だから

$$g(u) = \int_{-\infty}^{0} f_X(u-v) \cdot 0 \, dv + \int_{0}^{1} f_X(u-v) \cdot 1 \, dv + \int_{1}^{\infty} f_X(u-v) \cdot 0 \, dv$$

です．ここで

$$\int_{0}^{1} f_X(u-v) \, dv$$

の $f_X(u-v)$ は $f_X(x)$ の定義より，$0 \leq u-v \leq 1$ ならば 1 で，それ以外は 0 です．いま，

$$0 \leq u \leq 1 \text{ かつ } 0 \leq v \leq u$$

ならば

$$\inf(u-v) = 0, \sup(u-v) = 1$$

また

$$1 < u \leq 2 \text{ かつ } u-1 \leq v \leq 1$$

ならば

$$\inf(u-v) = 0, \sup(u-v) = 1$$

ゆえに $0 \leq u \leq 1$ のとき

$$g(u) = \int_{0}^{u} f_X(u-v) \, dv = u$$

です．また $1 < u \leq 2$ のとき

$$\int_{u-1}^{1} f_X(u-v)\,dv = [v]_{u-1}^{1} = 1-(u-1) = 2-u$$

です．以上をまとめると

$$g(u) = \begin{cases} u & , \quad 0 \leq u \leq 1 \\ 2-u & , \quad 1 < u \leq 2 \\ 0 & , \quad \text{その他} \end{cases}$$

です．この確率密度関数のグラフは次のような形をしています．

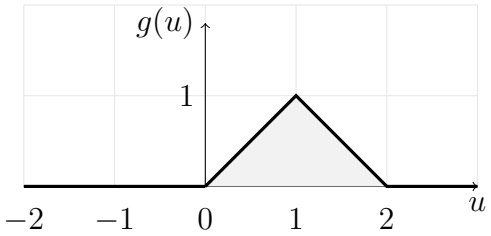

図 4: 2 つの一様分布を合成して作った分布

一様分布を足して作った確率変数の密度関数は《三角形》になるのでした．

確率変数同士の合成は，統計モデリングの基本です．経済学や心理学や社会学で自分が研究している対象をモデル化すると，既存の分布では表現できないことがしばしばあります．そのような場合は，上記の合成を利用して，必要な分布をつくってみましょう．

以上で 2 日目は終了です．

3　χ^2 分布の導出

数学的準備が整ったので，さっそく χ^2 分布の導出過程をフォローしてみましょう．本節の流れは次の通りです．

1. 逆関数が多価の場合（$Y = X^2$ のタイプ）の変数変換
2. 標準正規分布の 2 乗から自由度 1 の χ^2 分布をつくる
3. 数学的帰納法で自由度 n の χ^2 分布をつくる

3.1　逆関数が多価の場合の変数変換

みなさんは既に

$$Z = X + Y$$

$$Z = XY$$

といったタイプの合成はできると思います (1〜2 日目の内容を参照)．

統計学でよく使う自由度 n の χ^2 分布は，$X_i \sim N(0,1)$ を使って

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

という分布を作った場合の，Y がしたがう分布です．このような合成は X^2 という関数を使っているために，少し工夫が必要です．

3.2　自由度 1 の χ^2 分布

例 5 (標準正規分布の 2 乗の分布)．$X \sim N(0,1)$ とします．$Y = X^2$ と変換した場合の Y の確率密度関数 $q(y)$ を変数変換定理を使って計算してみましょう． X の確率密度関数を $f(x)$ とおけば

$$\begin{aligned}
P(X^2 < a) &= P(-\sqrt{a} < X < \sqrt{a}) \\
&= \int_{-\sqrt{a}}^{\sqrt{a}} f(x)dx \\
&= 2\int_{0}^{\sqrt{a}} f(x)dx && f(x) \text{ が偶関数のため} \\
&= 2\int_{0}^{a} f(\sqrt{y})\frac{dx}{dy}dy && x = \sqrt{y} \text{ とおいて変数変換} \\
&= \int_{0}^{a} \frac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2} dy
\end{aligned}$$

です．ところで

$$P(Y = X^2 < a)$$

と考えれば，

$$P(Y < a) = \int_0^a \frac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2} dy$$

は Y の分布関数なので，被積分関数は Y の確率密度関数です．ゆえに $X \sim N(0,1)$ であるとき，$Y = X^2$ と変換した場合の Y の確率密度関数 $q(y)$ は

$$q(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2}$$

です．これを自由度 1 の χ^2 分布と呼びます．

3.3 自由度 n の χ^2 分布

前節の例で，$X \sim N(0,1)$ であるとき $Y = X^2$ と変換した場合の Y の分布が，自由度 1 の χ^2 分布にしたがうことが分かりました．これを n 個足したときの分布，すなわち

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

は自由度 n の χ^2 分布にしたがいます[3]．このことを数学的帰納法によって示します．

命題 7 (自由度 n の χ^2 分布)**.** $X_i \sim N(0,1)$ と仮定する．

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

とおくと，Y の確率密度関数は

$$f(y) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} y^{\frac{n-2}{2}} e^{-y/2}, & y > 0 \text{ のとき} \\ 0, & y \leq 0 \text{ のとき} \end{cases}$$

となる。ここで $\Gamma(a)$ はガンマ関数と呼ばれ，定義は

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx \quad (a > 0)$$

である．このとき Y は自由度 n の χ^2 分布にしたがう，という．

[3] 自由度という謎の言葉が出てきましたが，ここでは単に $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ という合成に使う独立な確率変数 X_i の数，という意味で使います

証明. 数学的帰納法で証明します．まず $n=1$ の場合は，Y の確率密度関数は

$$f_1(y)=\frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}y^{-\frac{1}{2}}e^{-y/2}=\frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-y/2}, y>0 \text{ のとき}$$

です．これは $Y=X^2$ と変換した場合の Y の確率密度関数と一致します（ここで $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$ という命題を使いました ）.

次に n のときに成立すると仮定して，$n+1$ のときに成立することを示します.

帰納法の仮定から

$$\begin{aligned}\int_0^\infty f_n(x-t)f_1(t)dt&=\int_0^\infty \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}(x-t)^{\frac{n-2}{2}}e^{-(x-t)/2}\frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}t^{-\frac{1}{2}}e^{-t/2}dt\\&=\int_0^x \frac{1}{2^{\frac{n+1}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}(x-t)^{\frac{n-2}{2}}e^{-x/2}t^{-\frac{1}{2}}dt\\&\qquad (t>x \text{ の範囲で確率密度は } 0)\\&=\frac{e^{-x/2}}{2^{\frac{n+1}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}\int_0^x (x-t)^{\frac{n-2}{2}}t^{-\frac{1}{2}}dt\end{aligned}$$

ここで $t=xy$ とおいて変数変換します．積分の部分だけを考えると

$$\begin{aligned}\int_0^x (x-t)^{\frac{n-2}{2}}t^{-\frac{1}{2}}dt&=\int_0^1 (x-xy)^{\frac{n-2}{2}}(xy)^{-\frac{1}{2}}\frac{dt}{dy}dy\\&=\int_0^1 (1-y)^{\frac{n-2}{2}}x^{\frac{n-2}{2}}x^{-\frac{1}{2}}y^{-\frac{1}{2}}xdy\\&=x^{\frac{n-2-1+2}{2}}\int_0^1 (1-y)^{\frac{n-2}{2}}y^{-\frac{1}{2}}xdy\\&=x^{\frac{n-1}{2}}B\left(\frac{n}{2},\frac{1}{2}\right)\\&=x^{\frac{n-1}{2}}\frac{\Gamma(n/2)\Gamma(1/2)}{\Gamma(n/2+1/2)}\end{aligned}$$

です．もとの式にもどせば

$$\frac{e^{-x/2}}{2^{\frac{n+1}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}x^{\frac{n-1}{2}}\frac{\Gamma(n/2)\Gamma(1/2)}{\Gamma(n/2+1/2)}=\frac{1}{2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2})}e^{-x/2}x^{\frac{n-1}{2}}$$

です．これは $n+1$ の場合に成立することを示しています．数学的帰納法により，X_i^2 を n 個足した場合の確率密度関数は，命題の示すとおりです. □

3.4 PC による計算例

証明だけだと，分布を合成するイメージがつかめない場合は，コンピュータを使って計算してみるとよいでしょう（以下の計算では Mathematica を使いました）.

```
data = RandomVariate[NormalDistribution[0, 1], 10^4];
data2 = data*data;
Show[
    Histogram[data2, 20, "ProbabilityDensity"],
    Plot[PDF[ChiSquareDistribution[1], x], {x, 0, 12},
    PlotStyle -> Thick, PlotRange -> {0, 1}]]
```

1行目は標準正規分布の乱数を 10^4 個作っています．2行目で各値を2乗して，3行目以降でヒストグラムを作成しています．曲線は自由度1の χ^2 分布の確率密度関数です．

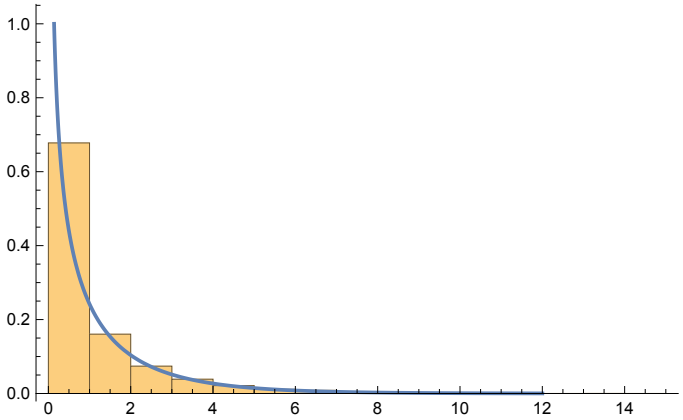

図 5: $X \sim N(0,1)$ のとき $Y = X^2$ と変換した Y の分布．

同様の操作で

$$Y = X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2$$

と変換すると Y が自由度5の χ^2 分布にしたがうと，命題は主張しています．

```
data1 = RandomVariate[NormalDistribution[0, 1], 10^4];
data2 = RandomVariate[NormalDistribution[0, 1], 10^4];
data3 = RandomVariate[NormalDistribution[0, 1], 10^4];
data4 = RandomVariate[NormalDistribution[0, 1], 10^4];
data5 = RandomVariate[NormalDistribution[0, 1], 10^4];
data = data1^2 + data2^2 + data3^2 + data4^2 + data5^2;
Show[
    Histogram[data, 20, "Probability"],
    Plot[PDF[ChiSquareDistribution[5], x], {x, 0, Max[data]},
    PlotStyle -> Thick, PlotRange -> {0, 0.3}]]
```

1行目から5行目で標準正規分布の乱数を 10^4 個 $\times 5$ セット作っています．6行目で各値を2乗しながら足します．7行目以降でヒストグラムを作成しています．曲線は自由度5の χ^2 分布の確率密度関数です．

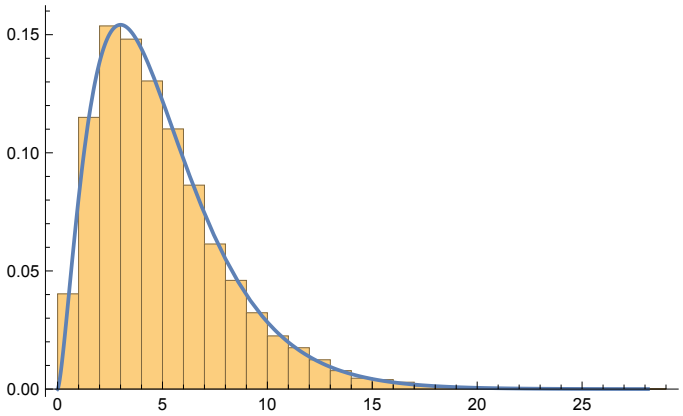

図 6: $X_1, X_2 \sim N(0,1)$ のとき $Y = X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2$ と変換した Y の分布.

標準正規分布の 2 乗和の分布と，χ^2 分布の確率密度関数が一致する様子が分かります.
以上で 3 日目は終了です.

4　t 分布の導出

さて，ようやく本題である t 分布を導出する準備ができました．

まず t 分布をつくるまでの流れを図で確認します浜田 (2020: 241)．

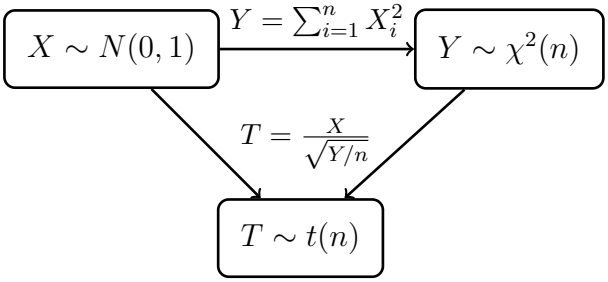

図 7: 正規分布、χ^2 分布、t 分布の関係

はじめに n 個の独立な標準正規分布を 2 乗して足して Y をつくります．すると Y は、パラメータ n の χ^2 分布にしたがいます（3 日目参照)。つまり $X \sim N(0,1)$ であるとき、

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

という確率変数 Y をつくると、Y の確率密度関数は

$$f(y) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n-2}{2}} e^{-y/2}, & y > 0 \text{ のとき} \\ 0, & y \leq 0 \text{ のとき} \end{cases}$$

です。この確率密度関数を持つ Y の分布を《**自由度 n の χ^2 分布**》と呼び，記号で

$$Y \sim \chi^2(n)$$

と書きます。

次に、標準正規分布 X と自由度 n の χ^2 分布 Y の比をとって、新しい確率変数 T をつくります。

$$T = \frac{X}{\sqrt{Y/n}}$$

X と Y が独立なら、この確率変数 T も特定の分布にしたがいます。

計算結果だけを書くと、T の確率密度関数は

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad (n \geq 1)$$

です。この確率密度関数を持つ T の分布を《**自由度 n の t 分布**》と呼び、記号で

$$T \sim t(n)$$

と書きます。

命題のかたちで一般的に表現します.

命題 8 (t 分布)**.** 標準正規分布 X と自由度 n の χ^2 分布 Y の比によって、確率変数 T を定義する。

$$T = \frac{X}{\sqrt{Y/n}}$$

X と Y が独立なら、T の確率密度関数は

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad (n \geq 1)$$

となる。この確率密度関数を持つ T の分布を《**自由度 n の t 分布**》と呼び、記号で

$$T \sim t(n)$$

と書く。

ここまではイントロで書いたことの繰り返しです.

証明. 変数変換定理を使います．確率変数の組 (X, Y) を次の関数を満たす (T, U) に変換します.

$$T = \frac{X}{\sqrt{Y/n}}$$
$$U = \frac{Y}{n}$$

n は定数です．この関係を (X, Y) について表現し直すと

$$X = T\sqrt{U}$$
$$Y = nU$$

です.

変数変換定理により

$$q(t,u) = p(x,y)\begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{vmatrix}$$

であり，これを周辺化した分布が求める確率密度関数です．

$$f(t) = \int_{-\infty}^{\infty} q(t,u)du = \int_{-\infty}^{\infty} p(t\sqrt{u}, nu) \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{vmatrix} du$$

まず $p(x,y)$ は X, Y が独立なので積の形に分解できます．X が標準正規分布で Y が χ^2 分布なので

$$p(x,y) = p_x(x)p_y(y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n-2}{2}} e^{-y/2}$$

です．さらに t, u に変換すると $x = t\sqrt{u}, y = nu$ なので

$$\frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u})^2/2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} (nu)^{\frac{n-2}{2}} e^{-nu/2}$$

です．次にヤコビアンを計算すると

$$\begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{vmatrix} = \begin{vmatrix} \sqrt{u} & \frac{t}{2\sqrt{u}} \\ 0 & n \end{vmatrix} = n\sqrt{u}$$

です．以上を組み合わせて積分を計算して周辺分布を求めます．

$$\begin{aligned}
f(t) &= \int_{-\infty}^{\infty} p(t\sqrt{u}, nu) \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{vmatrix} du \\
&= \int_{-\infty}^{\infty} p(t\sqrt{u}, nu) n\sqrt{u} \ du \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u})^2/2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} (nu)^{\frac{n-2}{2}} e^{-nu/2} n\sqrt{u} \ du \\
&= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u})^2/2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} (nu)^{\frac{n-2}{2}} e^{-nu/2} n\sqrt{u} \ du \\
&\qquad (u \leq 0 \text{ の範囲で確率密度関数は } 0) \\
&= \frac{n^{\frac{n-2}{2}} n}{\sqrt{2\pi} 2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_{0}^{\infty} e^{-(t\sqrt{u})^2/2} u^{\frac{n-2}{2}} e^{-nu/2} \sqrt{u} \ du \\
&\qquad (u \text{ の積分に関係しない項を外にだす}) \\
&= \frac{n^{\frac{n}{2}}}{\sqrt{2\pi} 2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_{0}^{\infty} e^{-(t\sqrt{u})^2/2} u^{\frac{n-2}{2}} e^{-nu/2} \sqrt{u} \ du
\end{aligned}$$

積分の外に出した項を K とおきます．

$$K = \frac{n^{\frac{n}{2}}}{\sqrt{2\pi} 2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$$

積分の中身を整理すると

$$\begin{aligned} f(t) &= K\int_0^\infty e^{-(t\sqrt{u})^2/2}u^{\frac{n-2}{2}}e^{-nu/2}\sqrt{u}\ du \\ &= K\int_0^\infty e^{-(t\sqrt{u})^2/2}e^{-nu/2}u^{\frac{n-2}{2}}u^{1/2}\ du \\ &= K\int_0^\infty e^{-\frac{u(n+t^2)}{2}}u^{\frac{n-1}{2}}\ du \end{aligned}$$

です．ここで $w = u(n+t^2)$ とおいて変数変換します．

$$\begin{aligned} f(t) &= K\int_0^\infty e^{-\frac{u(n+t^2)}{2}}u^{\frac{n-1}{2}}\ du \\ &= K\int_0^\infty e^{-\frac{w}{2}}\left(\frac{w}{n+t^2}\right)^{\frac{n-1}{2}}\frac{du}{dw}\ dw \\ &= K\int_0^\infty e^{-\frac{1}{2}w}w^{\frac{n-1}{2}}\left(\frac{1}{n+t^2}\right)^{\frac{n-1}{2}}\frac{1}{n+t^2}\ dw \\ &= K\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}}\int_0^\infty e^{-\frac{w}{2}}w^{\frac{n-1}{2}}\ dw \\ &= K\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}}2^{\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right) \qquad \text{ガンマ関数の性質を使います} \end{aligned}$$

最後に K をもとに戻して整理します．

$$\begin{aligned} f(t) &= K\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}}2^{\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right) \\ &= \frac{n^{\frac{n}{2}}}{\sqrt{2\pi}2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}}2^{\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right) \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\frac{n^{\frac{n}{2}}}{\sqrt{\pi}2^{\frac{1}{2}}2^{\frac{n}{2}}}2^{\frac{n+1}{2}}\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}} \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\frac{n^{\frac{n+1}{2}}n^{-\frac{1}{2}}}{\sqrt{\pi}}\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}} \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(\frac{n}{n+t^2}\right)^{\frac{n+1}{2}} \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \end{aligned}$$

最後の関数は t 分布の確率密度関数です．　□

以上で 4 日目は終わりです．

5　t 分布の意味

今日で最後（5 日目）です．4 日目でようやく t 分布の導出に成功しました．しかしこのややこしい確率密度関数を持った分布は，どんな場面で役立つのでしょうか？ 最後にその利用法を確認しましょう．

5.1　母平均の仮説検定

母平均にかんする仮説検定では，$X_1, X_2, \ldots, X_n$ が独立で各 X_i が平均 μ と標準偏差 σ をもつとき

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$$

という確率変数を使います．この統計量の実現値を計算する際に，$\bar{X}$ には標本平均値を，母平均 μ にはその候補 μ^* を代入します．しかし σ^2 は未知母数なので，標本からは分かりません．

統計学のテキストには，しばしば推定量の代用として

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \tag{1}$$

が使えると書いてあります．ここで S^2 は標本分散（確率変数）で

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

です．

われわれはすでに標準正規分布 X と自由度 n の χ^2 分布 Y の比によって、確率変数 T をつくると，この

$$T = \frac{X}{\sqrt{Y/n}}$$

が t 分布に従うことを知っています．

この命題を利用して，推定量の代用 (1) が t 分布に従うことを示せるかどうか考えてみましょう．まず (1) の分母分子に母標準偏差 σ をかけると

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \frac{\sigma}{\sigma}\frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \frac{\sigma}{S}\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{S/\sigma}$$

です．ここで分母 S/σ に注目すると

$$
\begin{aligned}
S/\sigma &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}\cdot\frac{1}{\sigma} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\frac{1}{\sigma^2}} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}
\end{aligned}
$$

です．よくみると

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$$

の部分は（X_i が正規分布なら）標準正規分布の 2 乗和

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

に形が似ています．標準正規分布の 2 乗和なら自由度 n の χ^2 分布に従いますが（3 日目を参照），母平均 μ が標本平均 $\bar{X}$ になっているため，どんな分布にしたがうのかまだよく分かりません．

そこで次の補題を使います．

命題 9 (自由度 $n-1$ の χ^2 分布)**.** $X_i \sim N(\mu,\sigma^2)$ とおく．$X_1, X_2, \ldots, X_n$ が独立であるとき

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$$

は自由度 $n-1$ の χ^2 分布に従う．

証明. 証明の方針は以下の通りです (小針 1973).

$$Z_i = \frac{X_i-\mu}{\sigma}$$

とおけば $Z_i \sim N(0,1)$ です．

$$
\begin{aligned}
\bar{Z} &= \frac{1}{n}\sum_{i=1}^{n}Z_i = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} \\
&= \frac{1}{n\sigma}(\sum_{i=1}^{n}X_i - n\mu) = \frac{1}{n\sigma}(n\bar{X}-n\mu) \\
&= \frac{\bar{X}-\mu}{\sigma}
\end{aligned}
$$

であり,

$$\begin{aligned}\frac{X_i-\bar{X}}{\sigma} &= \frac{X_i-\mu}{\sigma}+\frac{\mu-\bar{X}}{\sigma}\\ &= Z_i-\bar{Z}\end{aligned}$$

なので

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{Z_i-\bar{Z}}{1}\right)^2$$

が成立します．右辺が自由度 $n-1$ の χ^2 分布に従うことを示せば命題の証明は完了します．

そこで，標準正規分布にしたがう Z_i を，別の標準正規分布 Y_i に変換して

$$\sum_{i=1}^{n}\left(\frac{Z_i-\bar{Z}}{1}\right)^2=\sum_{i=1}^{n-1}Y_i^2 \tag{2}$$

という関係が成立することを示します．以上が証明の方針です．

まず $(Z_1,Z_2,\ldots,Z_n)$ から $(Y_1,Y_2,\ldots,Y_n)$ の変換として

$$\begin{aligned}y_1 &= a_{11}z_1+a_{12}z_2+\cdots+a_{1n}z_n\\ y_2 &= a_{21}z_1+a_{22}z_2+\cdots+a_{2n}z_n\\ &\vdots\\ y_n &= a_{n1}z_1+a_{n2}z_2+\cdots+a_{nn}z_n\end{aligned}$$

を考え，A を a_{ij} を成分とする n 次の直交行列であると仮定します．$\boldsymbol{z}=(Z_1,Z_2,\ldots,Z_n)^T$, $\boldsymbol{y}=(Y_1,Y_2,\ldots,Y_n)^T$ とおけば,

$$A\boldsymbol{z}=\boldsymbol{y}$$

です．直交行列の性質より

$$||\boldsymbol{y}||=||A\boldsymbol{z}||=||\boldsymbol{z}||$$

なのでベクトル $\boldsymbol{y}$ のノルムと $\boldsymbol{z}$ のノルムは一致します．これを成分で書けば

$$\sum_{i=1}^{n}Z_i^2=\sum_{i=1}^{n}Y_i^2 \tag{3}$$

が成立しています.

さらに直交行列の n 行目が

$$\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}$$

であると仮定すれば,

$$Y_n = \frac{1}{\sqrt{n}}(Z_1 + Z_2 + \cdots + Z_n)$$
$$Y_n^2 = \frac{1}{n}(Z_1 + Z_2 + \cdots + Z_n)^2$$
$$Y_n^2 = \frac{1}{n}(n\bar{Z})^2 = \frac{1}{n}n^2\bar{Z}^2 = n\bar{Z}^2$$

より

$$Y_n^2 = n\bar{Z}^2 \tag{4}$$

が成立します.

ところで変数変換定理(2日目参照)により確率密度関数$q(z_1, z_2, \ldots, z_n)$と$p(y_1, y_2, \ldots, y_n)$の間には

$$q(z_1, z_2, \ldots, z_n) = p(y_1, y_2, \ldots, y_n)\left|\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(z_1, z_2, \ldots, z_n)}\right|$$

が成立するのでした[4]. この変数変換は直交行列による1次変換のため, ヤコビアンは直交行列Aの行列式(determinant)に一致します. そして直交行列の性質により, 行列式は1です. よって

$$\begin{aligned} q(z_1, z_2, \ldots, z_n) &= p(y_1, y_2, \ldots, y_n)\left|\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(z_1, z_2, \ldots, z_n)}\right| \\ &= p(y_1, y_2, \ldots, y_n) \cdot 1 \end{aligned}$$

です. 仮定より$Z_1, Z_2, \ldots, Z_n$は独立な標準正規分布なので,

$$q(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-z_i^2/2} = \frac{1}{(\sqrt{2\pi})^n}e^{-\frac{1}{2}\sum z_i^2}$$

です. $q(z_1, z_2, \ldots, z_n) = p(y_1, y_2, \ldots, y_n)$ より

$$\begin{aligned} p(y_1, y_2, \ldots, y_n) &= q(z_1, z_2, \ldots, z_n) \\ &= \frac{1}{(\sqrt{2\pi})^n}e^{-\frac{1}{2}\sum z_i^2} \\ &= \frac{1}{(\sqrt{2\pi})^n}e^{-\frac{1}{2}\sum y_i^2} \qquad (3)\text{ より} \\ &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-y_i^2/2} \end{aligned}$$

です. つまり直交変換された$(Y_1, Y_2, \ldots, Y_n)$の各Y_iは独立に標準正規分布にしたがいます.

[4]この条件下ではヤコビアンはn次正方行列の行列式です. ここでは省略記号を使って表現していますのでご注意ください

そして

$$\begin{aligned}\sum_{i=1}^{n}(Z_i-\bar{Z})^2 &= \sum_{i=1}^{n}Z_i^2-n\bar{Z}^2\\ &= \sum_{i=1}^{n}Y_i^2-Y_n^2 \qquad (3)\text{ および }(4)\text{ より}\\ &= \sum_{i=1}^{n-1}Y_i^2\end{aligned}$$

なので右辺の和は《$n-1$ 個の独立な標準正規分布の 2 乗和》となり，自由度 $n-1$ の χ^2 分布にしたがいます．

したがって当然，左辺の

$$\sum_{i=1}^{n}(Z_i-\bar{Z})^2$$

も自由度 $n-1$ の χ^2 分布にしたがいます．

ところで，最初に確認したように

$$\sum_{i=1}^{n}(Z_i-\bar{Z})^2=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$$

だったことを思い出します．ここまでに左辺が自由度 $n-1$ の χ^2 分布にしたがうことを示したので，当然右辺も自由度 $n-1$ の χ^2 分布にしたがいます．

これで命題が証明されました．

□

さて話を戻すと，われわれは母平均の推定量として母分散の代わりに標本分散を使った

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$$

という推定量 (1) の分布が知りたいのでした．

簡単な変形によって推定量 (1) は

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}}=\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{S/\sigma}=\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}{n-1}}}$$

と書けることをさきほど確認しました．ここで分子 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ と分母 $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$ の独立性を確認する必要があります．

5.1.1 独立性の確認

$Y_i \sim N(0,1), Z_i = (X_i - \mu)/\sigma$ とおけば，命題 9 の証明より

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n-1} Y_i^2 + Y_n^2$$

かつ

$$\underbrace{\sum_{i=1}^{n-1} Y_i^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2}_{\chi^2(n-1)}, \underbrace{Y_n^2 = n\bar{Z}^2}_{\chi^2(1)}$$

です．これは 2 次形式の和なので，コクランの定理 (野田・宮岡 1992) により，$\sum_{i=1}^{n-1} Y_i^2$ と Y_n^2 は独立です[5].

それゆえ

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \quad と \quad n\bar{Z}^2$$

は独立です．また確率変数 X, Y が独立であるとき $f(X), Y$ も関数 f に逆変換が存在する場合には独立である，という命題により

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \qquad と \quad \sqrt{n\bar{Z}^2}$$

も独立です．

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

かつ

$$\sqrt{n\bar{Z}^2} = \sqrt{n}\bar{Z} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

より，推定量

$$\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}{n-1}}}$$

の分母と分子が独立であることがわかりました[6]．ゆえに 4 日目に確認した t 分布の命題により，推定量

$$\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}{n-1}}} = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

[5]直感的に言うと，前節で確認したように $\sum_{i=1}^{n-1} Y_i^2 + Y_n^2$ は自由度 n の χ^2 分布に従いますので，独立でないと χ^2 分布の再生性が成立しません．

[6]5.2 の例でも同様の独立性の確認が必要です．本質的なロジックは同じです

は自由度 $n-1$ の t 分布にしたがいます[7].

母平均の仮説検定を行う場合に，母分散の代わりに標本分散を代用できることの根拠は，『標本分散が母分散に似ているから』では不十分です．厳密な根拠は，推定量が《t 分布にしたがうこと》を証明できることです[8].

例 6 (2 次元の直交行列による変数変換)．《自由度 $n-1$ の χ^2 分布》命題の証明では，n 行目が $\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}$ であるような直交行列を使いました．そのような直交行列が実際に作れることを確認しましょう．$n=2$ の場合に

$$\begin{pmatrix} a & b \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{pmatrix} = \begin{pmatrix} a & b \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

という行列を考えます．これが直交行列であるためには，自身の転置行列との積が単位行列と等しくなければなりません．

$$\begin{pmatrix} a & b \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} a & \frac{1}{\sqrt{2}} \\ b & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

この条件を成分で表すと

$$a^2 + b^2 = 1$$
$$\frac{a}{\sqrt{2}} + \frac{b}{\sqrt{2}} = 0$$

です．2 つめの式は $a+b=0$ と書けます．この条件を満たす a, b は，たとえば

$$a = \frac{1}{\sqrt{2}}, b = -\frac{1}{\sqrt{2}}$$

があります．この直交行列をつかって標準正規分布 Z_1, Z_2 の次のような変換を考えます．

$$Y_1 = \frac{1}{\sqrt{2}} Z_1 - \frac{1}{\sqrt{2}} Z_2$$
$$Y_2 = \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2$$

[7]なお，この結果が成立するためには，サンプル X_i が正規分布にしたがう，という仮定が必要です．X_i の分布がなんであれ，標本平均 $\bar{X}$ が中心極限定理によって正規分布で近似できるなら，母平均の仮説検定ができると私は信じていたのですが，この証明では X_i 自体の分布形に制約があるため，サンプルによってはいわゆる《t 検定》が適用できない場合があることに気づきました．私の勘違いでしょうか？

[8]『母平均の仮説検定を行う場合に，母分散が未知なとき，その代用として標本分散（母分散の推定量）を使えば，正規分布にしたがう検定統計量の変わりとして，t 分布にしたがう推定量が使える』ことは多くの入門的テキストに書かれています．しかし，その理屈を説明している本は少ないようです．私が授業で使うテキストでも，この説明は省略されていることが多いので，このノートを書いて勉強することにしました．

すると，2 行目の条件から

$$Y_2^2 = \frac{Z_1^2}{2} + Z_1 Z_2 + \frac{Z_2^2}{2} = 2\bar{Z}^2$$

です．また Y_1 と $Y_2 2$ 乗して合計すると

$$Y_1^2 + Y_2^2 = \frac{Z_1^2}{2} - Z_1 Z_2 + \frac{Z_2^2}{2} + \frac{Z_1^2}{2} + Z_1 Z_2 + \frac{Z_2^2}{2} = Z_1^2 + Z_2^2$$

が成立します．変数変換定理より

$$\begin{aligned} q(z_1, z_2) &= p(y_1, y_2) \left| \frac{\partial(y_1, y_2)}{\partial(z_1, z_2)} \right| \\ &= p(y_1, y_2) \begin{vmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix} \\ &= p(y_1, y_2) \cdot 1 \end{aligned}$$

ヤコビアンを計算すると確かに 1 です．

$$\begin{aligned} p(y_1, y_2) &= q(z_1, z_2) \\ &= \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-z_2^2/2} = \frac{1}{(\sqrt{2\pi})^2} e^{-\frac{1}{2}(z_1^2 + z_2^2)} \\ &= \frac{1}{(\sqrt{2\pi})^2} e^{-\frac{1}{2}(y_1^2 + y_2^2)} = \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \end{aligned}$$

Y_1, Y_2 は確かに独立に標準正規分布に従います．

そして変数変換の条件から導出した

$$Y_1^2 + Y_2^2 = Z_1^2 + Z_2^2, \quad Y_2^2 = 2\bar{Z}^2$$

という関係より

$$\begin{aligned} \sum_{i=1}^{2} (Z_i - \bar{Z})^2 &= \sum_{i=1}^{2} Z_i^2 - 2\bar{Z}^2 \\ &= \sum_{i=1}^{2} Y_i^2 - Y_2^2 \qquad \text{上の関係より} \\ &= Y_1^2 + Y_2^2 - Y_2^2 = Y_1^2 \end{aligned}$$

なので，左辺と右辺が自由度 1 の χ^2 分布にしたがうことが分かります．

5.2 母平均の差の検定

2 集団の平均の差の検定で，母分散が分からない場合に t 分布を使う理由を説明します．
集団 1 のサンプル（確率変数）を

$$X_1, X_2, ..., X_n$$

とおき，集団 2 のサンプル（確率変数）を

$$Y_1, Y_2, ..., Y_m$$

とおきます．X_i の平均と分散は μ_1, σ_1^2，Y_i の平均と分散は μ_2, σ_2^2 とおきます．
このとき，次が成立します．

命題 10 (自由度 $n+m-2$ の t 分布と平均の差の検定)**.**

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) U^2}}$$

ただし

$$U^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2}$$

であるとき t は自由度 $n+m-2$ の t 分布に従う．

証明の補題として次を使います．

- $X_i \sim N(\mu, \sigma^2)$ とおく．$X_1, X_2, \ldots, X_n$ が独立であるとき

  $$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  は自由度 $n-1$ の χ^2 分布に従う（証明済み）．
- X, Y が独立でそれぞれ 自由度 n, m の χ^2 分布に従うとき，$X+Y$ は自由度 $n+m$ の χ^2 分布に従う（χ^2 分布の再生性）
- X が標準正規分布，Y が自由度 $n+m-2$ の χ^2 分布に従うとき，

  $$T = \frac{X}{\sqrt{Y/(n+m-2)}}$$

  は自由度 $n+m-2$ の t 分布に従う．

証明. 補題より X が標準正規分布，Y が自由度 $n+m-2$ の χ^2 分布に従うとき，

$$T = X\frac{1}{\sqrt{Y/(n+m-2)}}$$

は自由度 $n+m-2$ の t 分布に従う．

X は標準正規分布なので，これを

$$X = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

とおく．すると正規分布の再生性を使うと，右辺は標準正規分布に従う（細かなことを言うと，$\bar{X}, \bar{Y}$ に中心極限定理を適用して正規分布で近似する）．また分母の Y を

$$Y = \frac{1}{\sigma_1^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{1}{\sigma_2^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

と置き換える．第 1 項は自由度 $n-1$ の χ^2 分布に従い，第 2 項は自由度 $m-1$ の χ^2 分布に従うから，その和は自由度 $n+m-2$ の χ^2 分布に従う

さらに σ_1^2 と σ_2^2 が等しいと仮定する．以上の仮定の下で，次のように変形する．

$$\begin{aligned}
T &= X\frac{1}{\sqrt{Y/(n+m-2)}} \\
&= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \frac{1}{\sqrt{\frac{\frac{1}{\sigma_1^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 + \frac{1}{\sigma_2^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n+m-2}}} \\
&= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n} + \frac{1}{m})\sigma^2}} \frac{1}{\sqrt{\frac{\frac{1}{\sigma^2}(\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(Y_i-\bar{Y})^2)}{n+m-2}}} \\
&= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \frac{1}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n+m-2}}} \\
&= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n} + \frac{1}{m})U^2}}
\end{aligned}$$

□

仮定より，左辺が自由度 $n+m-2$ の t 分布に従うので，右辺の最下段も自由度 $n+m-2$ の t 分布に従う[9]．

[9] ここで χ^2 分布を使う際に，X_i, Y_i が正規分布に従うことを仮定している．よって母集団が正規分布に従っていない場合に，t 分布を仮定した差の検定はできないように思える（標本平均自体は中心極限定理を適用すれば正規分布で近似できる）．しかし現実には母集団の分布がなんであれ，皆平気で t 検定を使っているから，他に使える χ^2 分布の仮定が存在するのかもしれない

5.3 回帰係数の仮説検定

データ y_i $(i = 1, 2, \ldots, n)$ を生成する未知の分布を推測するための確率モデルとして

$$Y_i = a + bx_i + U_i$$

を仮定します。誤差 U_i が確率変数ならば、OLS 推定量 $\hat{B}$ も確率変数で

$$\hat{B} = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}$$

と表すことができます (浜田 2020). さらに

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \frac{x_i - \bar{x}}{S_{xx}} = w_i$$

とおけば, 推定量 $\hat{B}$ は

$$\hat{B} = b + \sum_{i=1}^{n} w_i U_i = b + w_1 U_1 + w_2 U_2 + \cdots + w_n U_n$$

とシンプルに誤差項の和として書けます。$U_i \sim N(0, \sigma^2)$ すなわち

誤差項 $U_1, U_2, \ldots, U_n$ が独立に平均 0 で分散 σ^2 の正規分布にしたがう

と仮定すれば

$$b + w_1 U_1 + w_2 U_2 + \cdots + w_n U_n$$

も正規分布にしたがいます. よって OLS 推定量 $\hat{B}$ の分布はより簡潔に

$$\hat{B} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$$

と書けます。つまり OLS 推定量 $\hat{B}$ は平均が b で分散が $\frac{\sigma^2}{S_{xx}}$ の正規分布にしたがいます。正規分布の性質によって、OLS 推定量 $\hat{B}$ を、その平均 b と標準偏差 $\sqrt{\frac{\sigma^2}{S_{xx}}}$ で標準化した確率変数

$$\frac{\hat{B} - \text{平均}}{\text{標準偏差}} = \frac{\hat{B} - b}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

は標準正規分布にしたがいます。

$$\frac{\hat{B} - b}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0, 1)$$

したがって誤差項の分散 σ^2 が分かれば, 係数の検定のために必要な統計量を計算できます.

しかし一般には誤差項の分散 σ^2 は分からないので，未知の σ^2 の代用品として、

$$\hat{S}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{U}_i^2$$

という確率変数 $\hat{S}^2$ を使います[10]。ここで $\hat{U}_i$ は

$$\hat{U}_i = Y_i - (\hat{A} + \hat{B}x_i) \quad \hat{A}, \hat{B}\text{は}\, a, b\, \text{の OLS 推定量}$$

という確率変数です.

この $\hat{S}^2$ を誤差項の分散 σ^2 の代用品として使い、

$$T = \frac{\hat{B} - b}{\sqrt{\frac{\hat{S}^2}{S_{xx}}}}$$

という確率変数を新たにつくります．そして分母分子に σ をかけると

$$T = \frac{\hat{B} - b}{\sqrt{\frac{\hat{S}^2}{S_{xx}}}} = \frac{\hat{B} - b}{\sqrt{\frac{\hat{S}^2}{S_{xx}}}}\frac{\sigma}{\sigma} = \frac{\hat{B} - b}{\sqrt{\frac{\sigma^2}{S_{xx}}}}\frac{\sigma}{\sqrt{\hat{S}^2}}$$
$$= \frac{Z}{\sqrt{\frac{\sum_{i=1}^{n}\hat{U_i}^2/\sigma^2}{n-2}}}$$

だから

$$\sum_{i=1}^{n}\hat{U_i}^2/\sigma^2$$

の分布を考えます.

命題 (自由度 $n-2$ の χ^2 分布). OLS 残差の二乗和

$$\sum_{i=1}^{n}\hat{U_i}^2 = \sum_{i=1}^{n}(Y_i - (\hat{A} + \hat{B}x_i))^2$$

は自由度 $n-2$ の χ^2 分布にしたがう

証明. [11]回帰モデルの誤差項を

$$U_i = Y_i - a - bx_i$$

とおけば，代数的な変形により，OLS 残差の 2 乗和は

$$\sum_{i=1}^{n}\hat{U_i}^2 = \sum_{i=1}^{n}U_i^2 - n\bar{U}^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2(\hat{B} - b)^2$$

[10]推定量 $\hat{S}^2$ は誤差項の分散の不偏推定量（期待値が分散 σ^2 に一致）です。

[11]この証明は Larsen & Marx (2012) を参照しました．証明の方針は基本的には《自由度 $n-1$ の χ^2 分布》命題の証明と同じです.

と書けます.

ここで 1 行目が

$$\frac{x_1-\bar{x}}{\sqrt{S_{xx}}}, \frac{x_2-\bar{x}}{\sqrt{S_{xx}}}, \ldots, \frac{x_n-\bar{x}}{\sqrt{S_{xx}}}$$

で 2 行目が

$$\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}$$

であるような直交行列 A を使い，確率変数ベクトル $\boldsymbol{u} = (U_1, U_2, \ldots, U_n)^T$ から $\boldsymbol{z} = (Z_1, Z_2, \ldots, Z_n)^T$ への変換

$$\boldsymbol{z} = A\boldsymbol{u}$$

を考えます．すると A の 1 行目から

$$z_1 = \frac{1}{\sqrt{S_{xx}}}\sum_{i=1}^{n}(x_i-\bar{x})U_i = \frac{S_{xx}}{\sqrt{S_{xx}}}(\hat{B}-b)$$
$$z_1^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2(\hat{B}-b)^2$$

であり，A の 2 行目から

$$z_2^2 = \left(\frac{1}{\sqrt{n}}(U_1+U_2+\cdots+U_n)\right)^2 = n\bar{U}^2$$

です．ゆえに

$$\begin{aligned}\sum_{i=1}^{n}\hat{U_i}^2 &= \sum_{i=1}^{n}U_i^2 - n\bar{U}^2 - \sum_{i=1}^{n}(x_i-\bar{x})^2(\hat{B}-b)^2 \\ &= \sum_{i=1}^{n}U_i^2 - z_2^2 - z_1^2 \\ &= \sum_{i=1}^{n}Z_i^2 - z_2^2 - z_1^2 = \sum_{i=3}^{n}Z_i^2\end{aligned}$$

です.

よって

$$\sum_{i=1}^{n}\hat{U_i}^2/\sigma^2 = \sum_{i=3}^{n}Z_i^2/\sigma^2 = \sum_{i=3}^{n}\left(\frac{Z_i-0}{\sigma}\right)^2$$

です．変数変換定理と直交行列 A の仮定より Z_i は平均 0，標準偏差 σ の正規分布に従います．$\frac{Z_i-0}{\sigma}$ は標準化によって標準正規分布にしたがうので，右辺は，標準正規分布の 2 乗を $n-2$ 個足したつくった確率変数です.

したがって，左辺の $\sum_{i=1}^{n}\hat{U_i}^2/\sigma^2$ は，自由度 $n-2$ の χ^2 分布にしたがいます.

□

以上の命題により，推定量

$$T = \frac{Z}{\sqrt{\frac{\sum_{i=1}^{n} \hat{U_i}^2/\sigma^2}{n-2}}}$$

の分母にある $\sum_{i=1}^{n} \hat{U_i}^2/\sigma^2$ が自由度 $n-2$ の χ^2 分布にしたがうことが分かりました．分子は標準正規分布なので，分子と分母が独立であれば T は自由度 $n-2$ の t 分布にしたがいます．確率変数間の関係は次の図をイメージしてください．

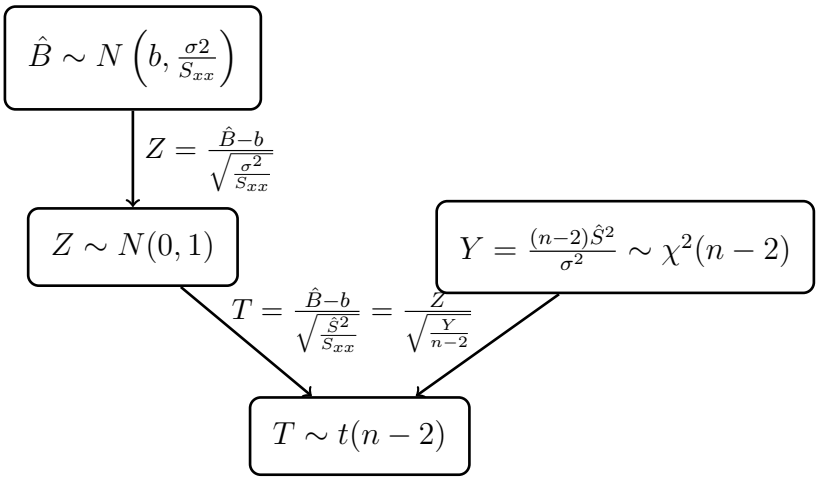

図 8: 正規分布（OLS 推定量）、χ^2 分布、t 分布の関係

最後にもう一度，内容をまとめます．

回帰係数の検定で，誤差項の分散の代用として，

$$\hat{S}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{U}_i^2$$

を使った場合の，検定統計量は

$$T = \frac{Z}{\sqrt{\frac{\sum_{i=1}^{n} \hat{U_i}^2/\sigma^2}{n-2}}}$$

です．この推定量は正規分布ではなく，自由度 $n-2$ の t 分布にしたがいます（その理由が，いま証明した命題です）．

以上でノートを終わります．

5.4　エンディング

青葉は、花京院の描いた図を見直した。

「よくわかんないけど、矢印の上や横に書いた式が、確率変数の合成の方法を表しているんだね」

「そうだよ。合成積や変数変換定理という方法を使うと確率変数同士の合成ができる。そうやって導出した分布が χ^2 分布や t 分布なんだよ」

「つまり裏でややこしい計算が必要なわけね」

「おもしろいから、再現してみようか？ ちょっと時間かかるけど」

「いや、遠慮しとく」

「一度自分で計算してみるといいよ、楽しいから」

「そんな計算が楽しいのは、花京院くんだけだよ……」

『その問題，やっぱり数理モデルが解決します』11 章:241-242

文献一覧

浜田宏，2020，『その問題，やっぱり数理モデルが解決します』ベレ出版.

小針晛宏，1973，『確率・統計入門』岩波書店.

Larsen, Richard J. & Morris L. Marx, 2012, *An Introduction to Mathematical Statistics and Its Applications, Fifth Edition,* Pearson.

野田一雄・宮岡悦良，1992，『数理統計学の基礎』共立出版.

矢野健太郎・田代嘉宏，1993，『社会科学者のための基礎数学改訂版』裳華房.