統計学教育における数式処理システムの利用

Examples of Using a Computer Algebra System in Statistics Education

ネットワーク情報学部　石鎚英也
School of Network and Information　Hideya ISHIZUCHI

Keywords: statistics, education, CAS, Maxima

Abstract

This essay suggests the possibility of statistics education using a computer algebra system (CAS) by attempting to derive typical probability distributions from normal distributions using CAS.

はじめに

統計学は難しい．例えば，最も重要な分布である正規分布からしてその由来がよく分からないという学生さんが多いのではないだろうか．正規分布から派生する確率分布については言わずもがなである．しかしながら，初等的な統計学の教育でも，オーソドックスな推定･検定の授業では，そうした分布の話を避けては通れない．

ある統計学のテキスト[5]に以下のような記述がある：「伝統的統計学の初等的学習では，t分布・F分布・カイ 2 乗分布などの標本分布を導出しません.『～という統計量は～分布することが知られている』と天下りに教えられました.（中略）それどころか統計学の授業を担当している教員自身が導いたことがないという場合も少なくないのである.」このように，統計量の分布について曖昧さなしに教えることは非常に難しいが，それを使わずに済ませることはできないので，天下り的に認めてもらうというのが，多くの授業で採られているやり方ではないだろうか．

ビッグデータとかデータサイエンスとかエビデンス・ベースト・アプローチといった新語が人口に膾炙し，大学教育でも実務的・実践的なデータ解析の教育が重視されてきているようである．しかしながら，昨今の数式離れの風潮からすると，基礎的な内容であっても，統計学の数理的な側面については，今後さらに手薄になっていくのではないかと危惧される．例えば，ほとんど数式がでてこない統計学の入門書もある[9]（誤解のないよう付言すると，文章と図版だけで統計学の概念や手法を説明しようとするそうした試みの価値を否定するものではない）．

特に文系の学部では，理論は必ずしも必要ないとする考え方もあり得るが，統計学の数理的な背景がある程度は分かる，あるいは，時間をかければ理解できるはずだといった感触を得ることは，よく分からないまま手順をただ暗記するのと比して，少なくとも，精神的には健全であろう．

本小論は，数式処理システム（CAS）を使って，正規分布から派生する典型的な確率分布の導出を試み，CAS を利用した統計教育の可能性を示唆するものである．手で行うには煩雑な積分も，数式処理システムを利用することで，腕力に自信のない人でも計算できるし，些末な計算より導出の大きな流れに意識を集中できるというメリットもある．ただし，当然ながら，その代償として，ソフトの利用方法を知り，書式に則してスクリプトを作る必要がある．従って，このようなアプローチによる教育は，他学部と比して，ネットワーク情報学部のような文理・情報系の学部により適していると思われる．

数式処理システムとして，専修大学の端末室では Mathematica が利用可能であるが，以下では，フリーウェアである Maxima（wxMaxima 17.10.0）を用いることにする．なお，Maxima のスクリプトについての詳しい解説は省略する（ヘルプや[10]などを参照されたい）．

準備

2.1. 逆関数

確率密度関数（PDF，probability density function）の導出の過程では，しばしば変数変換を行うことになるが，その際に変換の逆関数を扱う場合がある．そこで，まず関数gの逆関数を Maxima で求めてみよう（図 1 参照）．

関数$g: X \to Y$の逆関数g^{-1}を求めるには，$y = g(x)$をxについて解けばよいわけだが，gが 1 対 1 対応でない場合（例えば$y = x^2$）には，一般に複数の解（$x = \pm\sqrt{y}$）が存在する．その場合，このコードでは逆関数を求めず，警告を表示させている．

「使用例」の最後の 2 行は$y = \sqrt{x}$の逆関数を求める例である（本稿では，変数は全て実数とする）．コードの「assume」は，変数に対する仮定を示す．$y = \sqrt{x}$なので，実数の変数

しか扱わない場合「$x > 0$」は必要だが，「$y > 0$」は不要なはずである．しかし，これがないと，実行時に「Is y positive, negative or zero?」というメッセージが表れて処理が中断されてしまう．「p;」（positive の意）のようにyの符号をそこで入力すれば処理が継続されるが，煩雑なので「$y > 0$」を仮定に加えている．

```
(%i12)  /· 定義域sの関数gの逆関数 ·/
        inv(g, s):=block(_g:solve(g, s),
          if length(_g)=1 then return(_g[1])
          else "multivalued! use 'solve' instead.")$
        /· 使用例 ·/
        s:[x, y]$ g:[u=(x+y)/2, v=(x−y)/2];
          inv(g,s);
        s:[x]$ g:[y=x^2]; inv(g,s);
        solve(g,s);
        s:[x]$ assume(x>0, y>0)$
          g:[y=sqrt(x)]; inv(g,s);
(g)     [u=(y+x)/2, v=(x−y)/2]
(%o4)   [x=v+u, y=u−v]
(g)     [y=x^2]
(%o7)   multivalued! use 'solve' instead.
(%o8)   [x=−sqrt(y), x=sqrt(y)]
(g)     [y=sqrt(x)]
(%o12)  x=y^2
```

図 1　逆関数

2.2. 変数変換

確率変数の変数変換による確率密度関数の変化について，簡単に復習しておきたい(Maxima での実装は図 2参照)．なお，以下では，確率変数もその値も英小文字で表す．

まず，1 変数同士の変数変換については，変換$g: y \mapsto x$によって，確率密度関数f_xに従う確率変数xを新しい確率変数yに写したとき，yの確率密度関数f_yは，f_xとgにより次式のように表現することができる（図の tfrm 関数）．

$$f_y(y) = f_x(g(y)) \cdot \left|\frac{dx}{dy}\right|$$

これは，置換積分に相当する．

もし，定義域，値域が上記と逆の変数変換$g: x \mapsto y$の場合だと，yの確率密度関数f_yは，次式のように表現できる（図の tfrm1 関数）．

$$f_y(y) = f_x(g^{-1}(y)) \cdot \left|\frac{dx}{dy}\right|$$

また，2 変数以上の変数変換$g: x \mapsto y$の場合は，重積分の変数変換と同じように以下の関係が成り立つ（図の tfrm2 関数）．

$$f_y(y) = f_x(g^{-1}(y)) \cdot |J|$$

ここで，Jはg^{-1}のヤコビアンを表し，$|J|$はJの絶対値である．

```
/· 確率密度関数の変数変換 ·/
tfrm(f, si, gi):=block(_g2:map('rhs, gi),
  J:jacobian(_g2, si), j:determinant(J),
  trigreduce, subst(gi, f)·abs(j))$
tfrm1(f, s, g):=block(_g:[inv(g, s)], _g2:map('rhs, _g),
  _s:map('lhs, g), J:jacobian(_g2, _s),
  j:determinant(J), trigreduce, subst(_g, f)·abs(j))$
tfrm2(f, s, g):=block(_g:inv(g, s), _g2:map('rhs, _g),
  _s:map('lhs, g), J:jacobian(_g2, _s),
  j:determinant(J), trigreduce, subst(_g, f)·abs(j))$
```

図 2　変数変換

次に，変数変換の簡単な例を図 3 に示す．xを$[0,1]$上の一様分布（$f(x) = 1$）に従う確率変数とする．関係$x = y^2$を満たす確率変数yの確率密度関数は$f_y(y) = 2y$であることが tfrm の出力から分る（台は$0 \le y \le 1$だが，こうした計算はこのコードでは行っていない）．関係$y = \sqrt{x}$で変数変換しても同じ結果である（tfrm1 の出力）．

また，(x, y)を$[0,1] \times [0,1]$上の一様分布（$f(x, y) = 1$）に従う確率変数とする．変数変換$u = x + y, v = x - y$による確率変数(u, v)の分布も一様分布（$f_{uv}(u, v) = 1/2$）であることが tfrm2 の出力から分る（台は 4 点$(0,0)$，$(1,1)$，$(2,0)$，$(1,-1)$を順に結ぶ正方形の辺と内部である）．

```
        /· 使用例(1) ·/
        f:1;
        assume(y>0)$
        si:[y]$ gi:[x=y^2]; tfrm(f, si, gi);
        s:[x]$ g:[y=sqrt(x)]; tfrm1(f, s, g);
        s:[x, y]$ g:[u=x+y, v=x−y]; tfrm2(f, s, g);

        /· 使用例(2)　畳み込みの導出 ·/
        f:f1(x)·f2(y); s:[x, y]$ g:[u=x+y, v=y];
        h:tfrm2(f, s, g); integrate(h, v, minf, inf);
(f)     1
(gi)    [x=y^2]
(%o20)  2 y
(g)     [y=sqrt(x)]
(%o23)  2 y
(g)     [u=y+x, v=x−y]
(%o26)  1/2
(f)     f1(x) f2(y)
(g)     [u=y+x, v=y]
(h)     f1(u−v) f2(v)
(%o31)  ∫ f1(u−v) f2(v) dv   (−∞ to ∞)
```

図 3　変数変換（例）

「使用例(2)」では，畳み込み積分（convolution）の公式を導いている．すなわち，独立な 2 つの（連続型）確率変数(x,y)の和$u=x+y$の分布を求めるには，(x,y)の同時分布f_{xy}を各変数の密度関数の積$f_{xy}(x,y)=f_x(x)\cdot f_y(y)$として，$u=x+y, v=y$により変数変換したときの$(u,v)$の同時分布$f_{uv}(u,v)$を求め，それを$v$について積分，周辺化すればよい．結果は，畳み込み積分 $\int_{-\infty}^{\infty} f_x(u-v)\cdot f_y(v)dv$となる．

χ2 分布

3.1. 分布の定義と視覚化

n個の独立な確率変数$z_i(i=1,\cdots,n)$がいずれも標準正規分布$N(0,1^2)$に従うとき，それらの平方和$x=z_1^2+\cdots+z_n^2$の従う確率分布が（自由度nの）χ^2分布である．

図 4 では，標準正規分布$N(0,1^2)$に従う（疑似）乱数 10,000 個のヒストグラムと標準正規分布の確率密度関数$f(x)$を重ねて描いている．確率密度関数$f(x)$は，

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

で定義される．この乱数を 2 乗した 10,000 個の値は，図 5 のヒストグラムのように分布した．図には自由度 1 のχ^2分布の確率密度関数$f_1(x)$を重ねて描いている．確率密度関数$f_1(x)$は，以下のように定義される．

$$f_1(x)=\begin{cases}\frac{1}{\sqrt{2\pi}}x^{-\frac{1}{2}}\cdot e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$$

また，標準正規分布$N(0,1^2)$に従う 10,000 個の乱数を 3 組（計 30,000 個）用意し，各組から 1 つずつ順に取り出したデータの平方和 10,000 個を求めると，図 6 のヒストグラムのように分布した．図には自由度 3 のχ^2分布の確率密度関数$f_3(x)$を重ねて描いている．関数$f_3(x)$は，以下の通り：

$$f_3(x)=\begin{cases}\frac{1}{\sqrt{2\pi}}x^{\frac{1}{2}}\cdot e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$$

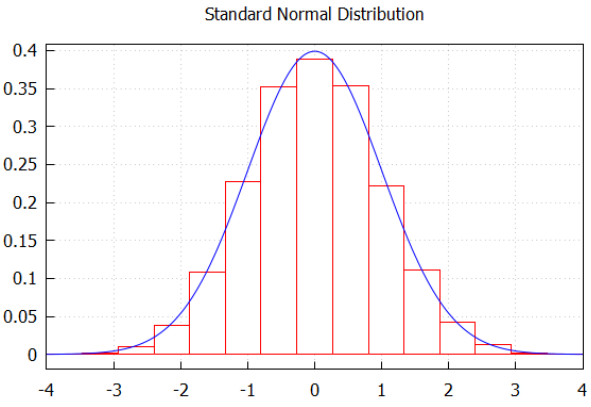


図 4　標準正規分布

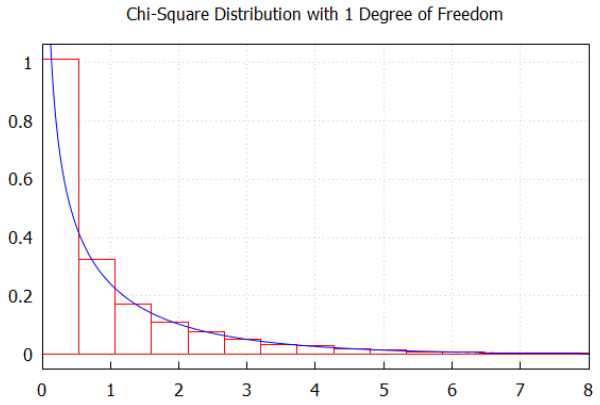


図 5　自由度 1 のχ^2分布

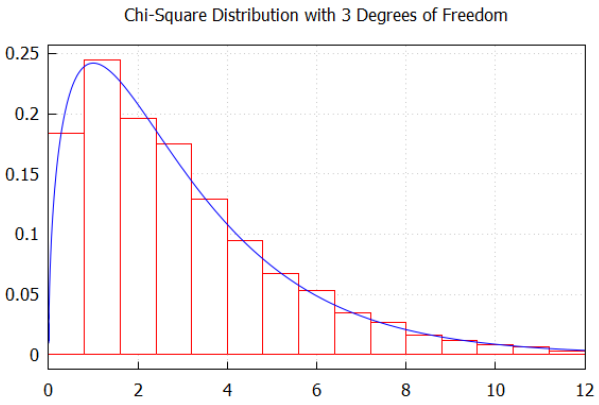


図 6　自由度 3 のχ^2分布

図 4，図 5，図 6 を表示する Maxima のコードを付録に示す)．

より一般に，(自由度nの) χ^2分布の確率密度関数$f_n(x)$は，

$$f_n(x)=\begin{cases}\frac{1}{2^{n/2}\cdot\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\cdot e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$$

で与えられる．自由度nは自然数であり，また自由度が 1 以外では，確率密度関数の台は 0 を含めて良いようである．また，Γはガンマ関数を示し，

$$\Gamma(n)=\int_0^{\infty}t^{n-1}\cdot e^{-t}dt$$

と定義される．ガンマ関数の引数nは，一般的には，（実部が正の）複素数である．また，

$$\Gamma(1)=1$$
$$\Gamma(n+1)=n\Gamma(n)$$

が成り立つ（図 7）．従って，特に，nが自然数のときには，

$$\Gamma(n+1)=n!$$

であることが分る．つまり，ガンマ関数は階乗を一般化した関数である．また，$\Gamma(1/2)=\sqrt{\pi}$が成立する．

これらの関係と$f_n(x)$から，自由度 1, 3 のχ^2分布の確率密度関数が前出の式$f_1(x), f_3(x)$であることが容易に確かめら

れる．なお，図 7 の「%e」はネイピア数（自然対数の底）を意味している．

```
(%i8)   /· 「関数の定義・性質 ·/
        gamma_expand:true$
        assume(n>0)$
        f(n):=integrate(t^(n-1)·%e^(-t), t, 0, inf);
        gamma(n); f(n)-gamma(n);
        f(1); f(n+1); f(1/2);
(%o3)   f(n) := ∫_0^∞ t^(n-1) %e^(-t) dt
(%o4)   Γ(n)
(%o5)   0
(%o6)   1
(%o7)   n Γ(n)
(%o8)   √π
```

図 7　Γ関数の定義・性質

ガンマ関数のグラフは，図 8 のようになる．また，自由度$n = 1, \cdots, 5$について，χ^2分布の確率密度関数のグラフを描くと図 9 のようになる．

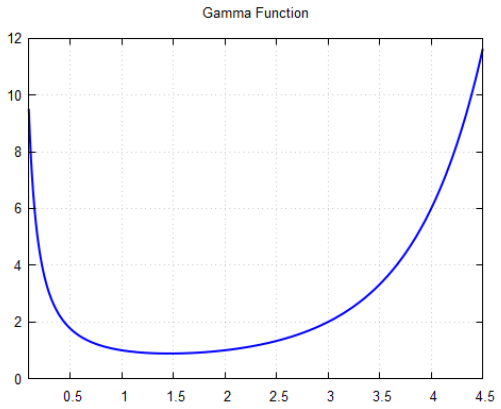


図 8　Γ関数

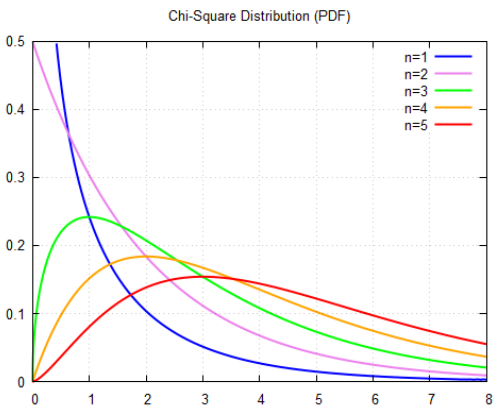


図 9　χ^2分布の確率密度関数

3.2. 分布の導出と性質

まず，n個の独立な確率変数$z_i (i = 1, \cdots, n)$がいずれも標準正規分布$N(0, 1^2)$に従うとき，それらの平方和$x = z_1^2 + \cdots + z_n^2$の確率分布が（自由度$n$の）$\chi^2$分布に従うことを Maxima で示そう．これは，以下のように，自由度に関する帰納法で証明できる．

(1)　標準正規分布に従う確率変数の平方は，自由度 1 のχ^2分布$\frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \cdot e^{-\frac{x}{2}}$に従う．

(2)　標準正規分布に従うn個の独立な確率変数の平方和が自由度nのχ^2分布に従うならば，標準正規分布に従う$n+1$個の独立な確率変数の平方和は自由度$n+1$のχ^2分布に従う．

ステップ(1)については，$x > 0$として図 10 のように確かめられる．ここで，「tfrm」は「準備」のところで定義した変数変換を行う関数である．「%」は，直前の結果を示す．

```
(%i13)  /· χ^2分布の導出(自由度1) ·/
        f:pdf_normal(z, 0, 1)$
        gi1:[z=sqrt(x)]$ gi2:[z=-sqrt(x)]$
        si:[x]$
        assume(x>0)$
        tfrm(f, si, gi1)+tfrm(f, si, gi2); rootscontract(%);
        rootscontract(pdf_chi2(x,1));
(%o11)  %e^(-x/2) / (√2 √π √x)
(%o12)  √(1/(2πx)) %e^(-x/2)
(%o13)  √(1/(2πx)) %e^(-x/2)
```

図 10　χ^2分布の導出（自由度 1）

ステップ(2)については，ステップ(1)の結果と帰納法の仮定により，自由度nのχ^2分布に従う確率変数と，（それと独立な）自由度 1 のχ^2分布に従う確率変数の和が自由度$n+1$のχ^2分布に従うことが示されればよいが，より一般に，χ^2分布の再生性（reproductive property）を確かめることができる．すなわち，自由度mとnのχ^2分布に従う独立な 2 つの確率変数の和は，（自由度$m+n$の）χ^2分布に従うことが分る（図 11）．同図では，以下の順で計算の経過を出力している．

(1)　自由度mのχ^2分布と自由度nのχ^2分布に従う独立な確率変数の和の分布を畳み込みで求める（図中の「res」変数）．

(2)　res に現れる積分の非積分関数を部分的に簡約化する．

(3)　簡約化した結果を使って，積分計算を改めて行う．

(4) その結果を使って，res の表現を書き換える．
(5) その結果と自由度$m+n$のχ^2分布の確率密度関数の差を計算する．

最後の計算結果が 0 となり，確率変数の和が自由度$m+n$のχ^2分布に従うことが確かめられた．

```
(%i16) /· χ^2分布の再生性 ·/
assume(m>0)$ assume(n>0)$
fy(y):=pdf_chi2(y, m)$ fz(z):=pdf_chi2(z, n)$
assume(x>v)$ assume(v>0)$
res:integrate(fy(x-v)·fz(v), v, 0, x);
part(res,1,2,1,[1,2])·ratsimp(part(res,1,2,1,3));
integrate(%,v,0,x);
makegamma(part(res,1,1)·%/part(res,2));
ratsimp(%-pdf_chi2(x, m+n));
```

$$(\%o13)\quad v^{n/2-1}(x-v)^{m/2-1}\,\%e^{-\frac{x}{2}}$$

$$(\%o14)\quad \beta\left(\frac{m}{2},\frac{n}{2}\right)x^{n/2+\frac{m}{2}-1}\,\%e^{-\frac{x}{2}}$$

$$(\%o15)\quad \frac{2^{-\frac{n}{2}-\frac{m}{2}}\,x^{n/2+\frac{m}{2}-1}\,\%e^{-\frac{x}{2}}}{\Gamma\left(\frac{n}{2}+\frac{m}{2}\right)}$$

$$(\%o16)\quad 0$$

図 11 χ^2分布の導出（再生性）

また，（母平均を未知とする）母分散の区間推定や検定では，標本の偏差平方和の分布に関する次のような性質が用いられる：n個の独立な確率変数$z_i(i=1,\cdots,n)$がいずれも同一の正規分布$N(\mu,\sigma^2)$に従うとき，（母標準偏差で標準化した）偏差平方和$x=((z_1-\bar{z})/\sigma)^2+\cdots+((z_n-\bar{z})/\sigma)^2$の確率分布は（自由度$n-1$の）$\chi^2$分布に従う．ただし，$\bar{z}$は標本平均である．

母平均μが既知なら，確率変数$(z_i-\mu)/\sigma$は標準正規分布に従う．よって，その平方$((z_i-\mu)/\sigma)^2$は自由度 1 のχ^2分布に従い，それらの和$((z_1-\mu)/\sigma)^2+\cdots+((z_n-\mu)/\sigma)^2$は自由度$n$の$\chi^2$分布に従う．しかし，母平均が未知で，母平均μを標本平均$\bar{z}$で置き換えると，自由度が 1 つ減ったχ^2分布になるということである．これは，「$(z_i-\bar{z})/\sigma\ (i=1,\cdots,n)$が独立でないため自由度が 1 つ下がる」という風に説明されることがあるが，そうした説明では，まじめな学生は納得できないのではないだろうか．

この性質については，色々な証明方法があるようである（例えば，コクランの定理によるもの[4]，モーメント母関数によるもの[7]，他の変数変換によるもの[3][8]など）．付録に証明の一例を示す．

t 分布

4.1. 分布の定義と視覚化

2 個の独立な確率変数y,zについて，yが標準正規分布$N(0,1^2)$に従い，zが自由度nのχ^2分布$\chi^2(n)$に従うとき，それらから定義される確率変数

$$x=\frac{y}{\sqrt{z/n}}$$

の従う確率分布が（自由度nの）t分布である．

標準正規分布$N(0,1^2)$に従う乱数yと自由度 1 のχ^2分布$\chi^2(1)$に従う乱数zをそれぞれ 10,000 個ずつ用意し，1 つずつ順に取り出したy,zに上式を適用してxを 10,000 個求めると，図 12 のヒストグラムのように分布した．図には自由度 1 のt分布の確率密度関数$f_1(x)$を重ねて描いている．確率密度関数$f_1(x)$は，

$$f_1(x)=\frac{1}{\pi(x^2+1)}$$

で定義され，（標準）コーシー分布とも呼ばれる．

また，zを自由度 3 のχ^2分布$\chi^2(3)$に従う乱数に変え，上と同様にxを 10,000 個求めると，図 13 のヒストグラムのように分布した．図には自由度 3 のt分布の確率密度関数$f_3(x)$を重ねて描いている．確率密度関数$f_3(x)$は，

$$f_3(x)=\frac{1}{\frac{\sqrt{3}\pi}{2}\left(\frac{x^2}{3}+1\right)^2}$$

で定義される．

より一般に，（自由度nの）t分布の確率密度関数$f_n(x)$は，

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

で与えられる．ただし，Bはベータ関数を示す．

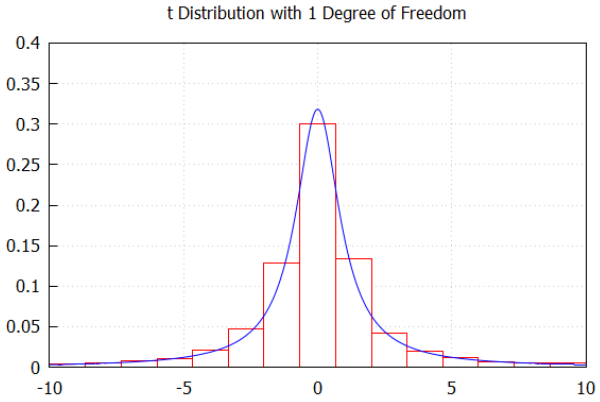


図 12 自由度 1 のt分布

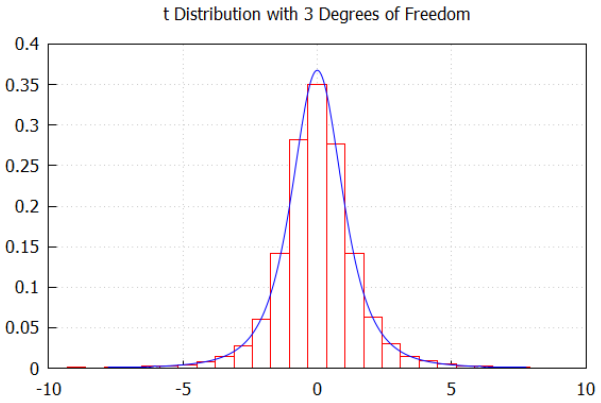


図 13　自由度 3 のt分布

ベータ関数Bは，

$$B(m,n)=\int_0^1 t^{m-1}\cdot(1-t)^{n-1}dt$$

と定義される．ベータ関数の引数m,nは，一般的には，（実部が正の）複素数である．また，ガンマ関数を使って

$$B(m,n)=\frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)}$$

と表現できる（図 14）．関数のグラフを図 15 に示す．

```
(%i7)   /· B関数の定義・性質 ·/
        assume(m>0, n>0)$
        f(m, n):=integrate(t^(m-1)·(1-t)^(n-1), t, 0, 1);
        beta(m, n); f(m, n)-beta(m, n);
        makegamma(f(m,n));
        f(1/2, 1/2); f(3/2, 1/2);
(%o2)   f(m,n) := ∫_0^1 t^(m-1) (1-t)^(n-1) dt
(%o3)   β(m,n)
(%o4)   0
(%o5)   Γ(m) Γ(n) / Γ(n+m)
(%o6)   π
(%o7)   π/2
```

図 14　B関数の定義・性質

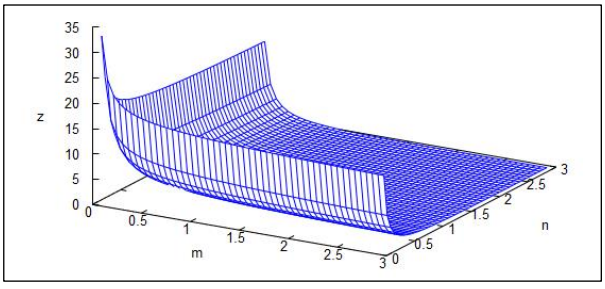


図 15　B関数

また，自由度$n=1,2,5,\infty$について，t分布の確率密度関数のグラフを描くと図 16 のようになる．ただし，$n=\infty$(inf)については標準正規分布の確率密度関数にしている．これは，t分布の自由度nを大きくすると，標準正規分布に近づくことによる（図 17）．

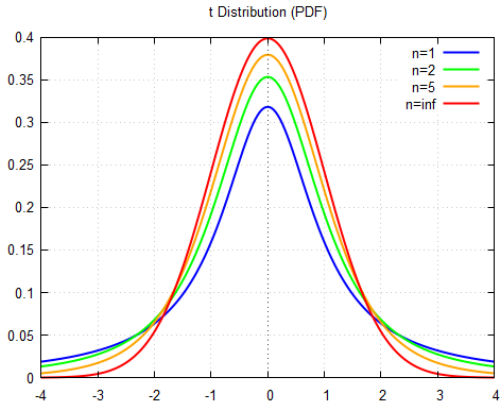


図 16　t分布の確率密度関数

```
(%i5)   /· t分布の極限 ·/
        load("distrib")$
        f(x, n):=pdf_student_t(x, n)$
        g(x):=pdf_normal(x, 0, 1)$
        limit(f(x, n), n, inf);
        %-g(x);
(%o4)   %e^(-x^2/2) / (√2 √π)
(%o5)   0
```

図 17　t分布の極限

4.2. 分布の導出と性質

2 個の独立な確率変数y,zについて，yが標準正規分布$N(0,1^2)$に従い，zが自由度nのχ^2分布$\chi^2(n)$に従うとき，それらから定義される確率変数

$$x=\frac{y}{\sqrt{z/n}}$$

が(自由度nの)t分布に従うことを図 18 のように示せる．同図では，以下の順で計算の経過を出力している．

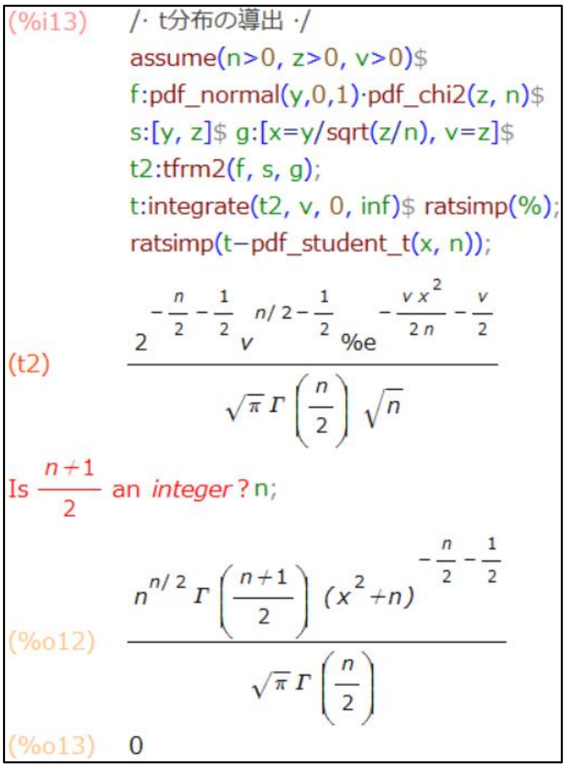
(1) 標準正規分布に従う確率変数yと自由度nのχ^2分布に従う確率変数zの確率密度関数の積をfとする（これは，y,zが独立なときの同時分布）．

(2) 変数変換$g:(y,z)\mapsto(x,v)$を$x=\frac{y}{\sqrt{z/n}}, v=z$とし，「準備」で定義した「tfrm2」を使って，$(x,v)$の同時分布を求める（図中の「t2」変数）．

(3)　t2 をvについて積分してxの（周辺）分布を求める．なお，この時点で，システムから「$(n+1)/2$は整数か？」との問い合わせが現れるので，「y」(yes の意）または「n」（no の意）を入力する．図では「n」と入力しているが，「y」としても結果は変わらない．

(4)　得られたxの（周辺）分布とt分布の確率密度関数の差を計算する．

最後の計算結果が 0 となり，確率変数xが自由度nのt分布に従うことが確かめられた．


```
(%i13)  /* t分布の導出 */
        assume(n>0, z>0, v>0)$
        f:pdf_normal(y,0,1)*pdf_chi2(z, n)$
        s:[y, z]$ g:[x=y/sqrt(z/n), v=z]$
        t2:tfrm2(f, s, g);
        t:integrate(t2, v, 0, inf)$ ratsimp(%);
        ratsimp(t-pdf_student_t(x, n));
```

$$\text{(t2)}\quad \frac{2^{-\frac{n}{2}-\frac{1}{2}}\, v^{n/2-\frac{1}{2}}\, \%e^{-\frac{v\,x^2}{2n}-\frac{v}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)\sqrt{n}}$$

Is $\frac{n+1}{2}$ an *integer*? n;

$$\text{(\%o12)}\quad \frac{n^{n/2}\,\Gamma\left(\frac{n+1}{2}\right)(x^2+n)^{-\frac{n}{2}-\frac{1}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}$$

(%o13)　0

図 18　t分布の導出

F 分布

5.1. 分布の定義と視覚化

2 個の独立な確率変数y,zについて，それぞれ自由度m,nのχ^2分布に従うとき，それらの変数の比$x=\frac{y/m}{z/n}$の従う確率分布が（自由度(m,n)の）F分布である．

自由度 1 のχ^2分布に従う乱数 10,000 個を 2 組用意し，1 つずつ順に取り出した乱数y,zの比$x=y/z$を 10,000 個求めると，図 19 のヒストグラムのように分布した．図には自由度$(1,1)$のF分布の確率密度関数$f_{11}(x)$を重ねて描いている．確率密度関数$f_{11}(x)$は，

$$f_{11}(x)=\frac{1}{\pi\sqrt{x}(x+1)}$$

で定義される．

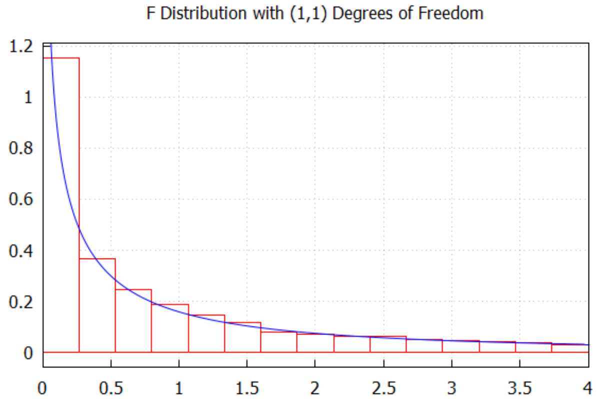


図 19　自由度$(1,1)$のF分布

また，2 組の乱数y,zをそれぞれ自由度 5, 2 のχ^2分布に変えた時の比$x=\frac{y/5}{z/2}$を 10,000 個求めると，図 20 のヒストグラムのように分布した．図には自由度$(5,2)$のF分布の確率密度関数$f_{52}(x)$を重ねて描いている．確率密度関数$f_{52}(x)$は，

$$f_{52}(x)=\frac{x^{3/2}}{(x+2/5)^{7/2}}$$

で定義される．

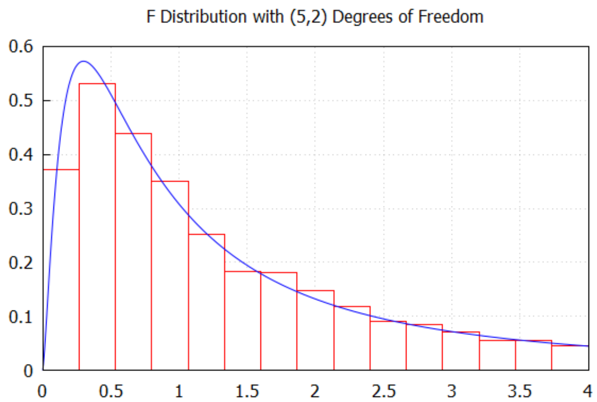


図 20　自由度$(5,2)$のF分布

さらに，自由度 100 の 2 組のχ^2分布について，同様の比を求めると，図 21 のヒストグラムのように分布した．図には自由度$(100,100)$のF分布の確率密度関数$f_{100\,100}(x)$を重ねて描いている．確率密度関数$f_{100\,100}(x)$は，

$$f_{100\,100}(x)=\frac{\alpha\cdot x^{49}}{(x+1)^{100}}$$

で定義される．ただし，αは

$$\alpha=25222836136391048333703124314 00$$

なる定数である（この数値については後述）．

より一般に，（自由度(m,n)の）F分布の確率密度関数$f_{mn}(x)$は，

$$f_{mn}(x)=\begin{cases}\dfrac{m^{\frac{m}{2}}\cdot n^{\frac{n}{2}}}{B\left(\frac{m}{2},\frac{n}{2}\right)}\cdot\dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, & x\geq 0\\ 0, & x<0\end{cases}$$

で与えられる．$f_{11}(x)$，$f_{52}(x)$，$f_{100\,100}(x)$の関数形は図 22 のように確かめられる．

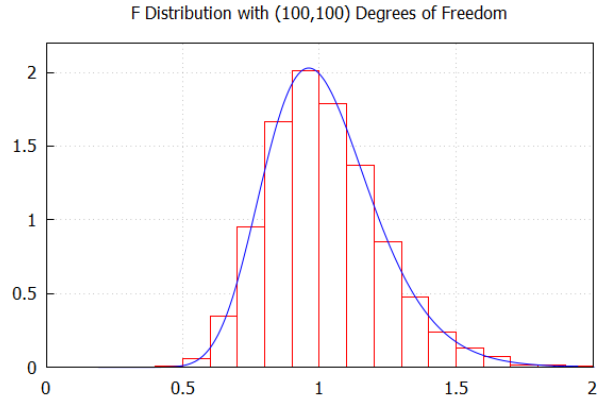


図 21　自由度(100, 100)のF分布

```
(%i8)   /· F分布の関数形確認 ·/
        load("distrib")$
        assume(x>0)$
        f:(m^(m/2)·n^(n/2))/beta(m/2,n/2) ·
          x^(m/2−1)/(m·x+n)^((m+n)/2) ;
        radcan(makegamma(f)−pdf_f(x, m, n));
        pdf_f(x, 1, 1);
        pdf_f(x, 5, 2);
        ratsimp(%−x^(3/2)/(x+2/5)^(7/2));
        pdf_f(x, 100, 100);
```

$$(f)\quad \frac{m^{m/2}\,n^{n/2}\,x^{m/2-1}}{\beta\left(\frac{m}{2},\frac{n}{2}\right)\,(mx+n)^{\frac{n+m}{2}}}$$

$$(\%o4)\quad 0$$

$$(\%o5)\quad \frac{1}{\pi\sqrt{x}\,(x+1)}$$

$$(\%o6)\quad \frac{5^{7/2}\,x^{3/2}}{2^{7/2}\left(\frac{5x}{2}+1\right)^{7/2}}$$

$$(\%o7)\quad 0$$

$$(\%o8)\quad \frac{252228361363910483337031243140 0\,x^{49}}{(x+1)^{100}}$$

図 22　F分布の関数形確認

また，自由度$(m,n)=(1,1),(2,1),(5,2),(10,1),(100,100)$について，$F$分布の確率密度関数のグラフを描くと図 23 のようになる．

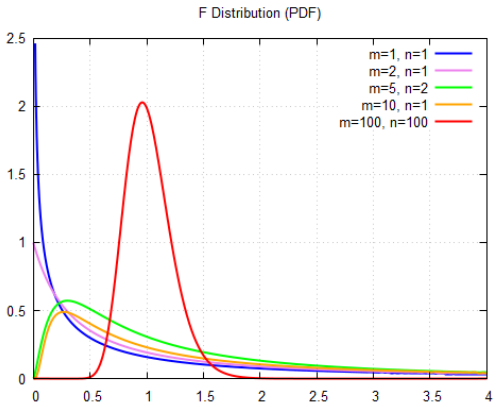


図 23　F分布の確率密度関数

5.2. 分布の導出と性質

2 個の独立な確率変数y,zについて，それぞれ自由度m,nのχ^2分布に従うとき，それらから定義される確率変数$x=\frac{y/m}{z/n}$が（自由度(m,n)の）F分布に従うことを図 24 のように示せる．

```
(%i17)  /· F分布の導出 ·/
        declare(m, integer)$ declare(n, integer)$
        assume(m>0, n>0, x>0, y>0)$
        f:pdf_chi2(y, m)·pdf_chi2(z, n)$
        s:[y, z]$
        g:[x=(y/m)/(z/n), v=z]$
        t2:tfrm2(f, s, g)$
        t:integrate(t2, v, 0, inf)$ radcan(t);
        pdf_f(x, m, n)$ radcan(%);
        ratsimp(t−pdf_f(x, m, n));
```

Is $\frac{n+m}{2}$ an *integer*?n;

$$(\%o14)\quad \frac{m^{m/2}\,n^{n/2}\,\Gamma\left(\frac{n+m}{2}\right)\,x^{\frac{m-2}{2}}}{\Gamma\left(\frac{m}{2}\right)\,\Gamma\left(\frac{n}{2}\right)\,(mx+n)^{\frac{n+m}{2}}}$$

$$(\%o16)\quad \frac{m^{m/2}\,n^{n/2}\,\Gamma\left(\frac{n+m}{2}\right)\,x^{\frac{m-2}{2}}}{\Gamma\left(\frac{m}{2}\right)\,\Gamma\left(\frac{n}{2}\right)\,(mx+n)^{\frac{n+m}{2}}}$$

$$(\%o17)\quad 0$$

図 24　F分布の導出

同図では，以下の順で計算の経過を出力している．これはt分布の導出と同じような手順である．

(1) 自由度mのχ^2分布に従う確率変数yと自由度nのχ^2分布に従う確率変数zの確率密度関数の積をfとする（これは，y,zが独立なときの同時分布）．

(2) 変数変換$g:(y,z)\mapsto(x,v)$を$x=\frac{y/m}{z/n}, v=z$とし，「準備」で定義した「tfrm2」を使って，(x,v)の同時分布を求める（図中の「t2」変数）．

(3) t2 をvについて積分してxの（周辺）分布を求める．なお，この時点で，システムから「$(n+m)/2$は整数か？」との問い合わせが現れるので，「y」(yes の意）または「n」（no の意）を入力する．図では「n」と入力しているが，「y」としても結果は変わらない．

(4) 得られたxの（周辺）分布とt分布の確率密度関数の差を計算する．

最後の計算結果が 0 となり，確率変数xが自由度m,nのF分布に従うことが確かめられた．

おわりに

統計学教育において，数式処理システム（CAS）を利用してできることをいくつか例示した．

- 乱数を使ったシミュレーションによって，理論分布の適合性を視覚的に確認した．
- 確率変数の変換によって定義される確率密度関数の理論式をいくつか導出した．

このうち特に後者においては，CAS の利用は，煩雑な積分計算から解放される点は大きなメリットであるが，しばしば，数式の色々な簡約化（ratsimp など）を組み合わせて試みる必要性が生じた．ただし，逆説的だが，「全自動」ではなく，こうした何らかの試行錯誤を伴う（＝手を動かし，時間を要す）方が，より理解が深まるかも知れない．

退職記念号ということで，最後にひとこと．

綿貫先生とは研究室が近く，プライベートでは，廊下ですれ違う時によく声をかけて下さいました．3 年次科目の「プロジェクト」では，環境問題をメインテーマに，熱心に学生を指導してこられました：大学の消費電力をリアルタイムに表示するシステムや，自転車型のトレーニングマシンで創エネ（人力発電）できる器具は特に印象に残っています．いつの頃からか，夏の暑い日も冬の寒い日も，在室されているときは，ドアが少し開いています．エアコン（や照明）を極力使わない省エネをストイックに実践されているものとお見受けしました．

益々のご健勝を

参考文献

[1] 総務省統計局 「統計学習の指導のために（先生向け）」 http://www.stat.go.jp/teacher/

[2] 総務省統計局 「高等学校学習指導要領解説 数学 統計関係部分抜粋」 http://www.stat.go.jp/teacher/dl/pdf/c3index/guideline/high/math.pdf （2017/11/20 閲覧）

[3] 高杉豊, 馬場敬之 2014 『演習 統計学キャンパス・ゼミ』 マセマ出版社

[4] 田坂誠男 1977 『品質管理の基礎』 朝倉書店

[5] 豊田秀樹 2016 『はじめての統計データ分析』 朝倉書店

[6] 馬場敬之，久池井茂 2003 『確率統計キャンパス・ゼミ』 マセマ出版社

[7] ホーエル P., G. 1978 『入門数理統計学』 培風館

[8] マスオ 「不偏分散と自由度 n-1 のカイ二乗分布」 https://mathtrain.jp/chinijoproof

[9] ロウントリー D. 2001 『新・涙なしの統計学』 新世社

[10] Maxima 5.41.0 Manual, http://maxima.osdn.jp/maxima.html

付録

a.1. 視覚化（Maximaスクリプトの例）

```
/* 正規分布と自由度 1, 3 の χ 2 分布 */
load ("descriptive")$
load("distrib")$
r_s:make_random_state(123)$
set_random_state(r_s)$

n:10000$
z1:random_normal(0,1,n)$
z2:random_normal(0,1,n)$
z3:random_normal(0,1,n)$

f(x):=pdf_normal(x,0,1)$
wxdraw2d(grid=true,title="Standard Normal Distribution",
  histogram_description(z1, nclasses=[-4,4,15], fill_color=red,
  frequency=density), explicit(f(x),x,-4,4), xrange=[-4,4])$
  rootscontract(f(x));

assume(x>0)$
f1(x):=pdf_chi2(x,1)$
wxdraw2d(grid=true,title="Chi-Square Distribution with 1
Degree of Freedom", histogram_description(z1^2, nclasses=
[0,8,15], fill_color=red, frequency=density), explicit
(f1(x),x,0,8), xrange=[0,8])$
rootscontract(f1(x));

f3(x):=pdf_chi2(x,3)$
wxdraw2d(grid=true,title="Chi-Square Distribution with 3
Degrees of Freedom", histogram_description(z1^2+z2^2+
z3^2, nclasses=[0,12,15], fill_color=red, frequency =density),
explicit(f3(x),x,0,12), xrange=[0,12])$
  rootscontract(f3(x));
```

a.2. 偏差平方和の分布

ここでは，下記の命題について，[8]に示されている証明を紹介し，また，特に，$n=2$と$n=3$の場合について，より具体的に Maxima で証明の流れを確かめる．なお，この証明では，直交行列による変数変換を用いており，基礎的な線形代数の知識を仮定する．

> n個の独立な確率変数$z_i\ (i=1,\cdots,n)$がいずれも同一の正規分布$N(\mu,\sigma^2)$に従うとき，（標準偏差で基準化された）偏差平方和$x=((z_1-\bar{z})/\sigma)^2+\cdots+((z_n-\bar{z})/\sigma)^2$の確率分布は（自由度$n-1$の）$\chi^2$分布に従う．ただし，$\bar{z}$は標本平均である．

まず，z_iを標準化して$w_i=(z_i-\mu)/\sigma$と置くと，w_iは標準正規分布に従うが，$w_i-\bar{w}=(z_i-\mu)/\sigma-(\bar{z}-\mu)/\sigma=(z_i-\bar{z})/\sigma$なので，下記の命題を示せばよいことが分る（ただし，$\bar{w}$は$w_i\ (i=1,\cdots,n)$の標本平均）．以下こちらを示す．

> n個の独立な確率変数$w_i\ (i=1,\cdots,n)$がいずれも標準正規分布に従うとき，偏差平方和$x=(w_1-\bar{w})^2+\cdots+(w_n-\bar{w})^2$の確率分布は（自由度$n-1$の）$\chi^2$分布に従う．

一行目の要素が全て$1/\sqrt{n}$であるようなn次元正規直交行列の一つをQとする．このような行列は実際に存在する．例えば，q_1を要素が全て$1/\sqrt{n}$であるようなn次元行ベクトルとし，$q_i\ (i=2,\cdots,n)$を第i要素のみが 1 で他は 0 であるようなn次元行ベクトルとすると，これらは線形独立なので，グラム・シュミットの正規直交化法で正規直交系を得て，それらを並べた行列を作ればよい．

その直交行列Qを使って，$w=(w_1\ \cdots\ w_n)^t$から$y=(y_1\ \cdots\ y_n)^t$への変数変換

$$\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}=Q\begin{pmatrix}w_1\\ \vdots\\ w_n\end{pmatrix}$$

を考える．

最初に，$y_i\ (i=1,\cdots,n)$が独立に標準正規分布に従うことを示す：w_iの同時分布は，独立性の仮定から標準正規分布の積

$$\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w_i^2}{2}\right)=(2\pi)^{-\frac{n}{2}}\cdot\exp\left(-\sum_{i=1}^{n}\frac{w_i^2}{2}\right)$$

$$=(2\pi)^{-\frac{n}{2}}\cdot\exp\left(-\frac{1}{2}w^t\,w\right)$$

で与えられる．また，Qの直交性から$Q^{-1}=Q^t$なので

$$w^t\,w=(Q^{-1}y)^tQ^{-1}y=(Q^ty)^tQ^{-1}y=y^t\,y$$

である．そして，$w=Q^ty$のヤコビアンはQの直交性から

$$|Q^t|=|Q|=\pm1$$

で，絶対値が 1 なので，y_iの同時分布も$(2\pi)^{-\frac{n}{2}}\cdot\exp\left(-\frac{1}{2}y^t\,y\right)$となる．すなわち，$y_i$は独立に標準正規分布に従う．

次に，

$$\sum_{i=1}^{n}(w_i-\bar{w})^2=\sum_{i=2}^{n}y_i^2$$

を示す：

$$\sum_{i=1}^{n}(w_i-\bar{w})^2=\sum_{i=1}^{n}w_i^2-2\bar{w}\sum_{i=1}^{n}w_i+\sum_{i=1}^{n}\bar{w}^2=\sum_{i=1}^{n}w_i^2-n\bar{w}^2$$

$$=w^t\,w-n\bar{w}^2=y^t\,y-n\left(\frac{1}{n}\sum_{i=1}^{n}w_i\right)^2$$

$$=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}\frac{w_i}{\sqrt{n}}\right)^2=\sum_{i=1}^{n}y_i^2-y_1^2$$

$$=\sum_{i=2}^{n}y_i^2$$

以上から，$x=\sum_{i=1}^{n}(w_i-\bar{w})^2$は，$n-1$個の独立な標準正規分布の平方和なので，自由度$n-1$の$\chi^2$分布に従うことが示された．□

さて，特に，$n=2$と$n=3$の場合について，より具体的にMaxima で証明の流れを確かめてみる（図 25 と図 26）．

```
(%i23)  /· (基準化された)偏差平方和の分布 (n=2) ·/
        load("eigen")$ load("distrib")$ n:2$
        Q:matrix([1/sqrt(n), 1/sqrt(n)], [0, 1])$
        q:gramschmidt(Q)$
        q1:unitvector(q[1])$ q2:unitvector(q[2])$
        Q:matrix(q1, q2)$
        f:pdf_normal(w1, 0, 1)·pdf_normal(w2, 0, 1)$
        radcan(f); s:[w1, w2]$ g:[y1=q1.s, y2=q2.s];
        t2:tfrm2(f, s, g)$ radcan(t2); m:(w1+w2)/2$
        ratsimp((w1−m)^2+(w2−m)^2);
        %−ratsimp(rhs(g[2])^2);

                                   − (w2^2 + w1^2)/2
                                %e
(%o16)                          ------------------
                                      2 π

(g)     [y1 = w2/sqrt(2) + w1/sqrt(2), y2 = w2/sqrt(2) − w1/sqrt(2)]

                                   − (y2^2 + y1^2)/2
                                %e
(%o20)                          ------------------
                                      2 π

                                 w2^2 − 2 w1 w2 + w1^2
(%o22)                           ---------------------
                                           2

(%o23)  0
```

図 25　偏差平方和の分布（$n=2$）

```
(%i22)  /· (基準化された)偏差平方和の分布 (n=3) ·/
        n:3$ Q:matrix([1/sqrt(n), 1/sqrt(n), 1/sqrt(n)],
          [0, 1, 0], [0, 0, 1])$ q:gramschmidt(Q)$
        q1:unitvector(q[1])$ q2:unitvector(q[2])$
        q3:unitvector(q[3])$ Q:matrix(q1, q2, q3)$
        f:pdf_normal(w1, 0, 1)·pdf_normal(w2, 0, 1)
          ·pdf_normal(w3, 0, 1)$ radcan(f);
        s:[w1, w2, w3]$ g:[y1=q1.s, y2=q2.s, y3=q3.s];
        t2:tfrm2(f, s, g)$ radcan(t2); m:(w1+w2+w3)/3$
        ratsimp((w1−m)^2+(w2−m)^2+(w3−m)^2);
        %−ratsimp(rhs(g[2])^2+rhs(g[3])^2);

                                   − (w3^2 + w2^2 + w1^2)/2
                                %e
(%o15)                          ------------------------
                                        3/2   3/2
                                   2    π

(g)     [y1 = w3/sqrt(3) + w2/sqrt(3) + w1/sqrt(3), y2 = − w3/(sqrt(2) sqrt(3)) + sqrt(2) w2/sqrt(3)
        − w1/(sqrt(2) sqrt(3)), y3 = w3/sqrt(2) − w1/sqrt(2)]

                                   − (y3^2 + y2^2 + y1^2)/2
                                %e
(%o19)                          ------------------------
                                        3/2   3/2
                                   2    π

(%o21)

                                2 w3^2 + (−2 w2 − 2 w1) w3 + 2 w2^2 − 2 w1 w2 + 2 w1^2
                                -------------------------------------------------------
                                                        3

(%o22)  0
```

図 26　偏差平方和の分布（$n=3$）

$n=2$の場合の処理の概要は，以下の通りである：

- 行列$\begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ 0 & 1\end{pmatrix}$にグラム・シュミットの直交化を行い（eigen パッケージの gramschmidt 関数），単位ベクトルに直して（unitvector 関数）並べ，正規直交行列Qを得る．図では出力指定されていないが，$\mathrm{Q}=\begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ -1/\sqrt{2} & 1/\sqrt{2}\end{pmatrix}$となる．
- 独立に標準正規分布に従う確率変数w_1, w_2の同時分布fを標準正規分布の PDF の積$\frac{1}{2\pi}\cdot\exp(-(w_1^2+w_2^2)/2)$として定義する．また，Qによる変数変換$g:(w_1,w_2)\mapsto(y_1,y_2)$を$y_1=(w_1+w_2)/\sqrt{2}, y_2=(-w_1+w_2)/\sqrt{2}$とし，「準備」で定義した「tfrm2」を使って，$(y_1,y_2)$の同時分布を求める（図中の「t2」変数）．結果は，$\frac{1}{2\pi}\cdot\exp(-(y_1^2+y_2^2)/2)$であることが分る．
- w_1, w_2の平均$m=(w_1+w_2)/2$からの偏差平方和$(w_1-m)^2+(w_2-m)^2$を計算し，これが$y_2^2=(-w_1+w_2)^2/2$と等しいことを示している．

また，$n=3$の場合の処理も同様である．正規直交行列は

$$\mathrm{Q}=\begin{pmatrix}1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3}\\ -1/\sqrt{6} & \sqrt{2}/\sqrt{3} & -1/\sqrt{6}\\ -1/\sqrt{2} & 0 & 1/\sqrt{2}\end{pmatrix}$$

であり，w_1, w_2, w_3の平均$m=(w_1+w_2+w_3)/3$からの偏差平方和$(w_1-m)^2+(w_2-m)^2+(w_3-m)^2$を計算し，これが$y_2^2+y_3^2$と等しいことを確かめている．