バイオインフォマティクスシリーズ 3

生　物　統　計

浜田 道昭 監修
木立 尚孝 著

コロナ社

シリーズ刊行のことば

現在の生命科学においては，シークエンサーや質量分析器に代表される計測機器の急速な進歩により，ゲノム，トランスクリプトーム，エピゲノム，プロテオーム，インタラクトーム，メタボロームなどの多種多様・大規模な分子レベルの「情報」が蓄積しています。これらの情報は生物ビッグデータ（あるいはオミクスデータ）と呼ばれ，このようなデータからいかにして新しい生命科学の発見をしていくかが非常に重要となっています。

このような状況の中でその重要性を増しているのが，生命科学と情報科学を融合した学際分野である「バイオインフォマティクス」（生命情報科学，生物情報科学）です。バイオインフォマティクスは，DNA やタンパク質の配列などの，生物の配列情報をディジタル情報として捉え，コンピュータにより解析を行うことを目的として誕生しました。このような，生物の配列情報を解析するバイオインフォマティクスの一分野は「配列解析」と呼ばれます（これは本シリーズでも主要なテーマとなっています）。上述の計測機器の進歩とともに，バイオインフォマティクスはここ数十年で飛躍的に発展し，いまや配列解析にとどまらずに，トランスクリプトーム解析，メタボローム解析，プロテオーム解析，生物ネットワーク解析など多岐にわたってきています。また，必要な知識も，統計学，機械学習，物理学，化学，数学などの多くの分野にまたがっています。しかしながら，これらのバイオインフォマティクスの多岐にわたる分野を，教科書的・体系的に学ぶことができる成書シリーズは，国内外を見てもほとんどありません。

そこで，大学生，大学院生，技術者，研究者などに，バイオインフォマティクスの各分野を体系的に学習することを可能とするための教科書を提供することを目的として本シリーズを企画しました。これを実現するために，バイオイン

フォマティクス分野の最前線で活躍をしている，若手・中堅の研究者に執筆を依頼しております。執筆者の方々には，バイオインフォマティクス研究の基盤となる理論やアルゴリズムを中心に，可能な限り厳密かつ自己完結的に解説を行うようにお願いしています。そのため，本シリーズは，大学などにおけるバイオインフォマティクスの講義の教科書として活用可能であるのみならず，読者が独学する場合にも最適な書籍になっていると確信しています。

最後になりますが，本シリーズの企画の段階から辛抱強くサポートしてくださったコロナ社の皆様に御礼を申し上げます。本シリーズが，今後のバイオインフォマティクス研究さらには生命科学研究の一助となることを切に願います。

2021 年 9 月

「バイオインフォマティクスシリーズ」監修者　浜田道昭

ま　え　が　き

DNA の塩基の並びを決定する DNA シークエンサーの高速化など，測定技術の飛躍的な進歩により，生命現象に関する情報は爆発的に増大している。例えば典型的な DNA シークエンシング実験では，一度の実験で 100 文字程度の A，C，G，T からなる文字列が数千万本出力される。人間の目では，一生をかけてもこのようなデータを見尽くすことはできず，そこにどのような生物学的情報が含まれているかを推測することも困難である。

したがってこのような膨大なデータの解析には，計算機を効率的に活用してデータが持つ特徴やパターンを抽出するデータサイエンスの手法が必要となる。データサイエンスを用いた解析をするためには，計算機の仕組み，プログラミング言語，統計科学，機械学習，人工知能など，さまざまな分野の知識が必要とされる。中でも統計科学は，測定データのランダム性の特徴を把握し，おのおののデータに対し適切な解析手法を選択するための強力な概念と方法を提供する。本書では，データサイエンスを活用した生命研究をするために必要な統計科学の基礎を解説する。

本書の特徴の一つは，紹介する解析手法の多くに数式を用いた導出をつけたことである。すでにある統計解析の書籍は，手法の使い方は説明しても，その数学的導出については省略することが多い。例えば t 検定の使い方を説明する書籍は多くあるが，t 検定統計量が帰無仮説のもとでどうしてスチューデントの t 分布に従うのかを解説したものはあまりない。本書は，そのような指南本に物足りなさを感じる読者が各手法の由来を理解し，自信を持って使えるようになることを目標とした。

本書のもう一つの特徴は，仮説検定など 20 世紀前半までに成立した古典的統計解析手法と，ベイズ統計やマルコフ連鎖モンテカルロ法など 20 世紀後半から

発展してきた比較的新しい解析手法を同程度の分量で記述したことである。生命データ解析ではどちらの手法も頻繁に使われることが理由だが，これにより統計科学の多様な背景思想を統一的に理解する助けになるのではないかと期待している。さらに本書では，生命データの解析で頻繁に現れる超高次元データに仮説検定を適用するときに重要となる多重検定補正の概念や，機械学習や人工知能技術における重要概念である過適合現象についても具体例を交えて詳しく解説した。

本書は，大学教養課程程度の線形代数と微積分の基礎知識がある読者を想定している。具体例の説明では生命科学の用語を用いることもあるが，解説する手法は生命データ解析に限らず一般のデータサイエンスで使えるものであり，生命科学の知識がなくとも理解できると考えている。

本書は多くの方の支援と助力により完成することができました。本書の執筆を勧めていただいた早稲田大学の浜田道昭先生とコロナ社に感謝します。また，早稲田大学の福永津嵩先生は査読の段階で有益なコメントをしてくれました。また，筆者の主宰する研究室の大学院生と東京大学理学部生物情報科学科の学生からは，本書の元となった生物統計論の講義に際し多くの有益なコメントをもらいました。最後に，いつも支えてくれた両親と，励ましと癒やしを与えてくれた妻と息子に感謝します。

2022 年 3 月

木立尚孝

本書で使われるおもな記号

記号	説明
$\mathbb{R}$	実数全体の集合
$\mathbb{C}$	複素数全体の集合
$a \bmod b$	整数 a を b で割った余り
$\gcd(A)$	整数集合 A の最大公約数
$\max_{\boldsymbol{x}} g$ ($\min_{\boldsymbol{x}} g$)	$\boldsymbol{x}$ を動かしたときの関数 $g(\boldsymbol{x})$ の最大値（最小値）
$\operatorname{argmax}_{\boldsymbol{x}} g$ ($\operatorname{argmin}_{\boldsymbol{x}} g$)	関数 $g(\boldsymbol{x})$ の値を最大（最小）にする変数 $\boldsymbol{x}$ の値
$\lim_{\boldsymbol{x}\to\boldsymbol{a}} g$	$\boldsymbol{x}$ を $\boldsymbol{a}$ に近づけたときの関数 $g(\boldsymbol{x})$ の極限値
$n!$	非負整数 n の階乗 $1 \cdot 2 \cdots n$。特に $0! = 1$
$\binom{n}{i}$	二項係数 $n!/(i!(n-i)!)$（2.2 節）
$\Gamma(z)$	ガンマ関数（2.6 節）
$B(\alpha, \beta)$	ベータ関数（6.5 節）
$\lvert a \rvert$	実数または複素数 a の絶対値
$\lvert \boldsymbol{A} \rvert$	行列 $\boldsymbol{A}$ の行列式（付録 A.3）
$\lvert A \rvert$	有限集合 A の要素数（付録 A.1）
$\lVert \boldsymbol{a} \rVert$	ベクトル $\boldsymbol{a}$ の長さ（付録 A.2）
$\boldsymbol{A}^T$	行列（ベクトル）$\boldsymbol{A}$ の転置行列（ベクトル）（付録 A.3）
a^{-1}	実数または複素数の逆数 $1/a$
$g^{-1}(\boldsymbol{x})$	関数（写像）g の逆関数（逆写像）（付録 A.1）
$\boldsymbol{A}^{-1}$	行列 $\boldsymbol{A}$ の逆行列（付録 A.3）
$[\boldsymbol{A}]_{ij}$ ($[\boldsymbol{x}]_i$)	行列 $\boldsymbol{A}$（ベクトル $\boldsymbol{x}$）の第 (i,j) 成分（第 i 成分）（付録 A.3）
$\boldsymbol{0}$	0 ベクトル（付録 A.2）
$\boldsymbol{1}$	すべての成分が 1 のベクトル（付録 A.2）
$\boldsymbol{I}$	単位行列（付録 A.3）
$\mathbf{diag}(a_1, \cdots, a_n)$	対角行列（付録 A.3）
π	円周率 $3.14\cdots$
π_0，$\hat{\pi}_0$	仮説検定集合のうち真の帰無仮説が占める割合（6.8 節）
$\boldsymbol{\pi}$，π_i，$\pi(\boldsymbol{x})$，$\pi_{\boldsymbol{x}}$	マルコフ過程の平衡分布（10.7 節，12.2 節）
δ_{ij}	クロネッカーのデルタ（式 (A.1)）
$\delta(x)$	ディラックのデルタ関数（2.10 節）
$\mathcal{O}(g(n))$	ランダウの $\mathcal{O}$ 記法（付録 A.5）
$\boldsymbol{\nabla}_{\boldsymbol{x}} g$	関数 $g(\boldsymbol{x})$ の変数 $\boldsymbol{x}$ に関する勾配（付録 A.4）
Ω	標本空間（1.3 節）
$\mathcal{E}$	事象空間（1.3 節）

記号	説明
$\mathbb{P}$	確率測度（1.3 節）
$\mathcal{B}(\mathbb{R})$	実直線上のボレル代数（1.4 節）
$X \sim \mathcal{P}$	X は確率分布 $\mathcal{P}$ に従う確率変数（1.5 節）
$x \sim \mathcal{P}$	$\mathcal{P}$ から値をランダムサンプリングし x が得られた（1.7 節）
$\mathbb{P}_T$，p_T	データがランダムサンプリングされる真の確率分布（1.2 節）
$\mathbb{P}_X$	X の確率分布（1.5 節）
$p_X(x)$，$p_{\mathcal{P}}(x)$，$p(x)$	x を出力する確率分布（1.9 節）
$p(x, y)$	同時確率分布（1.10 節）
$p(x\|y)$	条件付き確率分布（1.10 節）
P-value(s_{obs})	P 値（4.3 節）
$p(\boldsymbol{x}\|\boldsymbol{\theta})$	モデル分布（7.1 節）
$p(D\|\boldsymbol{\theta})$，$L(\boldsymbol{\theta}\|D)$	尤度（7.2, 9.5 節）
$l(\boldsymbol{\theta}\|D)$	対数尤度（7.2 節）
$p(\boldsymbol{\theta}\|D)$	事後分布（9.5 節）
$p(\boldsymbol{\theta})$	事前分布（9.5 節）
$p(D)$	エビデンス（9.5 節）
$\boldsymbol{P}$	マルコフ過程の遷移確率行列（10.2 節）
$\boldsymbol{P}^n$	マルコフ過程の n ステップ遷移確率行列（10.3 節）
$\mathbb{F}_X(x)$	X の確率分布関数（1.8 節）
$\boldsymbol{F}(\boldsymbol{\theta})$	フィッシャー情報量（7.5 節）
$f_X(x)$	X の確率密度関数（1.8 節）
$KL(p_1\|\|p_2)$	カルバック・ライブラー情報量（1.17 節）
$\mathbb{E}_{X\sim\mathcal{P}}(g(X))$	$X \sim \mathcal{P}$ のもとでの関数 $g(X)$ の期待値（1.11 節）
$\mathbb{E}(g(X,Y)\|Y=y)$	$Y = y$ のもとでの関数 $g(X, Y)$ の条件付き期待値（1.11 節）
$\mathbb{I}_b(x)$，$\mathbb{I}(a < X < b)$	指示関数（1.12 節）
$\mathrm{var}(X)$	X の分散（1.13 節）
$\mathrm{cov}(X, Y)$	X と Y の共分散（1.13 節）
$\mathbb{V}(\boldsymbol{X})$	$\boldsymbol{X}$ の分散共分散行列（1.13 節）
ρ_{XY}	X と Y の相関係数（1.14 節）
$\varphi_X(t)$	X の特性関数（1.16 節）
$\bar{x}$	標本平均（1.15 節）
s^2	標本分散（1.15 節）
c_{XY}	X と Y の標本共分散（1.14 節）
r_{XY}	X と Y の標本相関係数（式 (1.30)）
$\mathrm{Ber}(q)$	ベルヌーイ分布（2.1 節）
$\mathrm{Binom}(n, q)$	二項分布（2.2 節）
$\mathrm{Cat}(\boldsymbol{q})$	カテゴリカル分布（2.3 節）
$\mathrm{Mult}(n, \boldsymbol{q})$	多項分布（2.4 節）
$\mathcal{N}(\mu, \sigma^2)$	1 変数正規分布（2.5 節）
$\mathrm{Gamma}(\alpha, \beta)$	ガンマ分布（2.6 節）
$\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	多変数正規分布（2.7 節）
$\mathrm{Unif}(a, b)$	一様分布（2.8 節）

記号	説明
$\mathrm{Deg}(c)$	退化分布（2.9 節）
$\mathrm{Emp}(D)$	経験分布（2.11 節）
$\mathrm{HyperGeom}(n, m, k)$	超幾何分布（5.2 節）
$\chi^2(m)$	χ^2 分布（5.3 節）
$t(\nu)$	スチューデントの t 分布（5.7 節）
$\mathrm{Beta}(\alpha, \beta)$	ベータ分布（6.5 節）
$\mathrm{Geom}(q)$	幾何分布（13.3 節）
$\mathrm{Exp}(\lambda)$	指数分布（13.4 節）
$\mathrm{Pois}(\mu)$	ポアソン分布（13.7 節）
$\mathrm{NegBinom}(r, p)$	負の二項分布（13.7 節）
$\mathrm{Erlang}(r, \lambda)$	アーラン分布（13.7 節）
H_0	帰無仮説（4.2 節）
H_1	対立仮説（4.2 節）
$\boldsymbol{H}(\boldsymbol{\theta})$	ヘッセ行列（付録 A.5）
$H(\boldsymbol{\theta})$	エントロピー（8.2 節）
$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$	最尤推定量（7.3 節）
$\hat{\boldsymbol{w}}_{\mathrm{LS}}$	最小 2 乗推定量（9.1 節）
$\hat{\boldsymbol{w}}_{\mathrm{RLS}}$	正則化最小 2 乗推定量（9.4 節）
$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$	最大事後確率推定量（9.5 節）
$\hat{\boldsymbol{\theta}}_{\mathrm{PME}}$	事後平均推定量（9.5 節）
$Q_{\mathrm{EM}}(\boldsymbol{\theta}\|\boldsymbol{\theta}')$	期待値最大化法の Q 関数（8.2 節）
$RSS(\boldsymbol{w}\|D)$	予測値と測定データとの 2 乗誤差（9.1 節）
$RSS_\lambda(\boldsymbol{w}\|D)$	正則化項付き 2 乗誤差（9.4 節）
$n_{\mathrm{eff},i}$	有効サンプルサイズ（12.4 節）
$\tau_A(\omega)$	到達時刻（13.1 節）

目　　次

1. 統計解析の目的と確率空間

2. 確率分布の具体例

3. 大数の法則と中心極限定理

4. 仮説検定とP値

5. 仮説検定の具体例

6. 多重検定補正と false discovery rate

7. 確率モデル解析と最尤推定法

8. 混合正規分布と期待値最大化法

9. 回帰モデルの正則化とベイズ推定

10. マルコフ過程と平衡分布

11.　ランダムサンプリングと数値積分

12.　事後分布とマルコフ連鎖モンテカルロ法

13.　到達時刻とポアソン過程

広まることで進化する。

- 測定ノイズ：　生命現象の測定は多くの場合，測定方法固有のノイズの影響を受ける。特に，DNA シークエンシング実験は，サンプル中の DNA 分子の部分配列をランダムに選択して読み取る手法のため，同じ実験をしても，得られる測定データは DNA の塩基配列の集合としては毎回まったく異なるものとなる。
- 複雑な決定論的機構：　環境条件や実験者の違いなど，ある程度は決定論的に決まる過程であっても，法則性の推定や制御・予測が難しい場合には，確率的現象と同様の扱いがされる。

このように，調べたい研究対象自身に由来する確率的現象（**システムノイズ**（system noise））と，測定過程に由来する確率的現象（**測定ノイズ**（measurement noise））により同じ実験を繰り返しても，得られるデータは実験ごとに違ったものとなる。このようなばらつきのあるデータから意味のある生物学的情報を抽出するために統計解析が必要となる。

1.2　統計解析の目的と限界

統計解析では，測定の背後に，システムノイズと測定ノイズの両方に由来する**「真の」確率分布**（true distribution）$\mathbb{P}_T$ が存在しており，観測される測定データは，真の確率分布から**ランダムサンプリング**（random sampling）されたものであると仮定する。この仮定により確率論を用いたデータの分析が可能となる。解析者は，確率論を活用してデータのモデリングと分析を行い，真の確率分布が持つ性質を推定する。すなわち，統計解析は，確率論を用いて測定データから真の確率分布の性質を明らかにすることを目的とする（**図 1.1**）。

実験で得られた測定データから，背後にある真の確率分布の性質をどの程度詳細に推測できるかは，測定データの質と量に大きく依存する。データにノイズが少なくデータ量も十分にあれば，真の確率分布の詳細な性質がわかるだろ

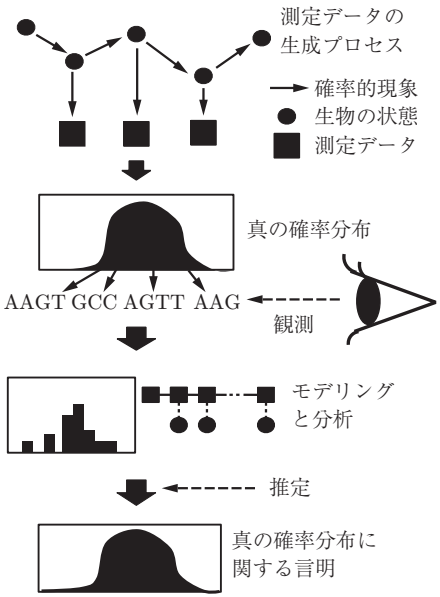

図 1.1　統計解析の目的

うし，データにばらつきが多かったり，サンプル数が少なかったりする場合は，大ざっぱな性質しかわからないだろう。データを解析する際には，対象となる測定データがどの程度の情報を含んでいるのか，あるいは引き出しうるかを見極め，適切な解析手法の選択や，妥当な到達目標を設定することが大事である。

たとえ無限に多いデータがあったとしても，統計解析が答えられない問がある。例えば，コインを振って表と裏のどちらが出たかを記録したデータがあるとしよう。コインをたくさん振って，表と裏が出た割合がどちらも 0.5 に限りなく近づけば，そのコインは表と裏の出方に偏りがないとほとんど確実にいえる。しかしこのときでも，つぎにコインを振ったときに表が出るか裏が出るかは予測できない。いえるのは，つぎに表と裏のどちらが出るかは五分五分であるということだけである。統計解析は，決定論的に決まる部分とランダム性によりばらつく部分とに分離し，ランダムに見える部分についてはそれ以上の原因の追究を諦める手法であるともいえる。

$$D_k = \frac{S_k^2}{n_k}$$
$$\Rightarrow T \overset{\text{近似的に}}{\sim} t(\nu)$$
$$\nu = \frac{(D_A + D_B)^2}{D_A^2/(n_A - 1) + D_B^2/(n_B - 1)}$$

5.8　スチューデントの t 分布の導出

ここでは 1 サンプル t 検定の検定統計量がスチューデントの t 分布に従うことを示す。まず，n 個のデータは正規分布 $\mathcal{N}(\mu_0, \sigma^2)$ に従うので，その確率密度は

$$\begin{aligned} f_X(x)d^n x &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{h=1}^{n}(x^{(h)} - \mu_0)^2\right) d^n x \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{h=1}^{n} z_h^2\right) d^n z \end{aligned} \tag{5.12}$$

である。ただし，2 行目では $z_h = (x^{(h)} - \mu_0)/\sigma$ と変数変換した。

ここで，n 番目の要素が長さ 1 のベクトル $\boldsymbol{e}^{(n)} = (1, \cdots, 1)/\sqrt{n}$ となるような正規直交基底 $\{\boldsymbol{e}^{(h)} \in \mathbb{R}^n | h = 1, \cdots, n,\ \boldsymbol{e}^{(h)T}\boldsymbol{e}^{(h')} = \delta_{hh'}\}$ を一つとる（付録 A.2 を参照）。このとき，式 (A.2) で説明したように，任意のベクトル $\boldsymbol{z}$ は正規直交基底の線形結合で表され（$\boldsymbol{z} = \sum_h \boldsymbol{e}^{(h)} y_h$），各項の係数は $\boldsymbol{z}$ と基底の内積をとり求められる（$y_h = \boldsymbol{e}^{(h)T}\boldsymbol{z}$）。また，各列に正規直交基底を並べた $(n \times n)$ 次元の行列 $\boldsymbol{O} = (\boldsymbol{e}^{(1)}, \cdots, \boldsymbol{e}^{(n)})$ は直交行列となり（付録 A.3 を参照），式 (A.11)，(A.12) より $\boldsymbol{O}$ は関係式 $\boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}$，$|\boldsymbol{O}| = \pm 1$ を満たす。ただし，$\boldsymbol{I}$ は $(n \times n)$ 次元の単位行列である。この直交行列の性質から次式のように $\boldsymbol{z}$ と $\boldsymbol{y}$ の内積が等しいことがわかる。

$$\begin{aligned} \boldsymbol{z}^T\boldsymbol{z} &= \boldsymbol{y}^T\boldsymbol{O}^T\boldsymbol{O}\boldsymbol{y} \\ &= \boldsymbol{y}^T\boldsymbol{y} \end{aligned}$$

この正規直交基底を用いると，以下のような関係式が成り立つ。

$$
\begin{aligned}
y_n &= \frac{1}{\sqrt{n}}(1,\cdots,1)z \\
&= \sqrt{n}\bar{z} \\
&= \sqrt{n}\frac{\bar{x}-\mu_0}{\sigma} \\
\sum_{h=1}^{n-1} y_h^2 &= \boldsymbol{y}^T\boldsymbol{y} - y_n^2 \\
&= \boldsymbol{z}^T\boldsymbol{z} - n\bar{z}^2 \\
&= \sum_{h=1}^{n}(z_h-\bar{z})^2 \\
&= \frac{1}{\sigma^2}\sum_{h=1}^{n}(x^{(h)}-\bar{x})^2
\end{aligned}
$$

さらに式 (5.12) で $\boldsymbol{z}=\boldsymbol{O}\boldsymbol{y}$ への積分の変数変換をすると（式 (A.22)），ヤコビ行列式 J の絶対値は $|J|=||\boldsymbol{O}||=|\pm 1|=1$ なので，$d^nz=d^ny$ が成り立つ。さらに，$(n-1)$ 次元ベクトル $\boldsymbol{y}'=(y_1,\cdots,y_{n-1})$ について極座標系への変数変換を行い（式 (A.23)），5.3 節のように角度方向について周辺化すると

$$
\begin{aligned}
&\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right)d^nz = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}'^T\boldsymbol{y}'+y_n^2)\right)d^{n-1}y'dy_n \\
&\quad\xrightarrow{\text{周辺化}} f_{\chi^2(\nu)}(a)f_{\mathcal{N}(0,1)}(b)dadb \\
&\quad= \frac{1}{2^{\nu/2}\Gamma(\nu/2)}a^{\nu/2-1}e^{-a/2}\frac{1}{\sqrt{2\pi}}e^{-b^2/2}dadb \\
&a = \boldsymbol{y}'^T\boldsymbol{y}' = \sum_{h=1}^{n-1} y_h^2 \\
&b = y_n \\
&\nu = (n-1)
\end{aligned}
$$

が得られる。

ここで (a,b) から $(s,t)=(a,\sqrt{\nu}b/\sqrt{a})$ へ変数変換し，t 検定統計量に関係

のない変数 s について周辺化すると

$$
\begin{aligned}
(s,t) &= (a, \sqrt{\nu}a^{-1/2}b) \\
&\Rightarrow (a,b) = (s, \nu^{-1/2}s^{1/2}t) \\
\left|\begin{matrix} \dfrac{\partial a}{\partial s} & \dfrac{\partial a}{\partial t} \\ \dfrac{\partial b}{\partial s} & \dfrac{\partial b}{\partial t} \end{matrix}\right| &= \left|\begin{matrix} 1 & 0 \\ \nu^{-1/2}s^{-1/2}/2 & \nu^{-1/2}s^{1/2} \end{matrix}\right| \\
&= \frac{1}{\sqrt{\nu}}s^{1/2} \\
f_{\chi^2(\nu)}(a)f_{\mathcal{N}(0,1)}(b)dadb &= \frac{1}{2^{(\nu+1)/2}\sqrt{\nu\pi}\Gamma(\nu/2)}s^{(\nu+1)/2-1}e^{-s(1+t^2/\nu)/2}dsdt \\
&\xrightarrow{\text{周辺化}} \frac{1}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}\left(\int_0^\infty u^{(\nu+1)/2-1}e^{-u}du\right)dt \\
&= \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \\
&= f_{t(\nu)}(t)dt
\end{aligned}
$$

となりスチューデントの t 分布が現れる。ただし s に関する積分では，$u = s(1+t^2/\nu)/2$ へ変数変換した後に式 (2.15) のガンマ関数の定義を使った。

一方，この変数 t の元をたどると

$$
\begin{aligned}
t &= \sqrt{\nu}a^{-1/2}b \\
&= \sqrt{\nu}\left(\sum_{h=1}^{n-1}y_h^2\right)^{-1/2}y_n \\
&= \sqrt{\nu}\left(\frac{1}{\sigma^2}\sum_{h=1}^{n}(x^{(h)}-\bar{x})^2\right)^{-1/2}\left(\sqrt{n}\frac{\bar{x}-\mu_0}{\sigma}\right) \\
&= \sqrt{n}\frac{\bar{x}-\mu_0}{\sqrt{\sum_{h=1}^{n}(x^{(h)}-\bar{x})^2/(n-1)}}
\end{aligned}
$$

となり t 検定統計量に等しい。これより t 検定の検定統計量 T が自由度 $\nu = n-1$ のスチューデントの t 分布に従うことが示された。

索　　引

—— 監修者・著者略歴 ——

浜田　道昭（はまだ　みちあき）
2000年　東北大学理学部数学科卒業
2002年　東北大学大学院理学研究科修士課程修了（数学専攻）
2002年　株式会社富士総合研究所研究員
2009年　東京工業大学大学院総合理工学研究科博士後期課程（社会人博士）修了（知能システム科学専攻）
博士（理学）
2010年　東京大学特任准教授
2014年　早稲田大学准教授
2018年　早稲田大学教授
現在に至る

木立　尚孝（きりゅう　ひさのり）
1997年　京都大学理学部卒業
1999年　東京大学大学院総合文化研究科修士課程修了（広域科学専攻）
2003年　奈良先端科学技術大学院大学情報科学研究科修士課程修了
2004年　博士（学術）（東京大学）
2005年　産業技術総合研究所特別研究員
2007年　奈良先端科学技術大学院大学情報科学研究科博士後期課程修了
博士（理学）
2009年　東京大学准教授
現在に至る

生　物　統　計

Biostatistics

2022 年 5 月 27 日　初版第 1 刷発行

検印省略

監 修 者　浜　田　道　昭
著　　者　木　立　尚　孝
発 行 者　株式会社　コ ロ ナ 社
代 表 者　牛 来 真 也
印 刷 所　三 美 印 刷 株 式 会 社
製 本 所　株式会社　グ リ ー ン

112–0011　東京都文京区千石 4–46–10
発 行 所　株式会社　**コ ロ ナ 社**
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話(03)3941–3131(代)
ホームページ　https://www.coronasha.co.jp

ISBN 978–4–339–02733–4　C3355　Printed in Japan　（新井）