

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

フォー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、バイオインフォー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ー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下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ー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

「バイオインフォー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
$\max_{\mathbf{x}} g$ ( $\min_{\mathbf{x}} g$ )	$\mathbf{x}$ を動かしたときの関数 $g(\mathbf{x})$ の最大値 (最小値)
$\operatorname{argmax}_{\mathbf{x}} g$ ( $\operatorname{argmin}_{\mathbf{x}} g$ )	関数 $g(\mathbf{x})$ の値を最大 (最小) にする変数 $\mathbf{x}$ の値
$\lim_{\mathbf{x} \rightarrow \mathbf{a}} g$	$\mathbf{x}$ を $\mathbf{a}$ に近づけたときの関数 $g(\mathbf{x})$ の極限值
$n!$	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)
$ a $	実数または複素数 $a$ の絶対値
$ \mathbf{A} $	行列 $\mathbf{A}$ の行列式 (付録 A.3)
$ A $	有限集合 $A$ の要素数 (付録 A.1)
$\ \mathbf{a}\ $	ベクトル $\mathbf{a}$ の長さ (付録 A.2)
$\mathbf{A}^T$	行列 (ベクトル) $\mathbf{A}$ 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 $1/a$
$g^{-1}(\mathbf{x})$	関数 (写像) $g$ の逆関数 (逆写像) (付録 A.1)
$\mathbf{A}^{-1}$	行列 $\mathbf{A}$ の逆行列 (付録 A.3)
$[\mathbf{A}]_{ij}$ ( $[\mathbf{x}]_i$ )	行列 $\mathbf{A}$ (ベクトル $\mathbf{x}$ ) の第 $(i, j)$ 成分 (第 $i$ 成分) (付録 A.3)
$\mathbf{0}$	$0$ ベクトル (付録 A.2)
$\mathbf{1}$	すべての成分が $1$ のベクトル (付録 A.2)
$\mathbf{I}$	単位行列 (付録 A.3)
$\operatorname{diag}(a_1, \dots, a_n)$	対角行列 (付録 A.3)
$\pi$	円周率 $3.14 \dots$
$\pi_0, \hat{\pi}_0$	仮説検定集合のうち真の帰無仮説が占める割合 (6.8 節)
$\boldsymbol{\pi}, \pi_i, \boldsymbol{\pi}(\mathbf{x}), \pi_{\mathbf{x}}$	マルコフ過程の平衡分布 (10.7 節, 12.2 節)
$\delta_{ij}$	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)
$\nabla_{\mathbf{x}} g$	関数 $g(\mathbf{x})$ の変数 $\mathbf{x}$ に関する勾配 (付録 A.4)
$\Omega$	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)
$P\text{-value}(s_{\text{obs}})$	P 値 (4.3 節)
$p(\mathbf{x} \boldsymbol{\theta})$	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)
$\mathbf{P}$	マルコフ過程の遷移確率行列 (10.2 節)
$\mathbf{P}^n$	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)
$\mathbf{F}(\boldsymbol{\theta})$	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)
$\text{var}(X)$	$X$ の分散 (1.13 節)
$\text{cov}(X, Y)$	$X$ と $Y$ の共分散 (1.13 節)
$\mathbb{V}(\mathbf{X})$	$\mathbf{X}$ の分散共分散行列 (1.13 節)
$\rho_{XY}$	$X$ 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))
$\text{Ber}(q)$	ベルヌーイ分布 (2.1 節)
$\text{Binom}(n, q)$	二項分布 (2.2 節)
$\text{Cat}(\mathbf{q})$	カテゴリカル分布 (2.3 節)
$\text{Mult}(n, \mathbf{q})$	多項分布 (2.4 節)
$\mathcal{N}(\mu, \sigma^2)$	1 変数正規分布 (2.5 節)
$\text{Gamma}(\alpha, \beta)$	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)
$\text{Unif}(a, b)$	一様分布 (2.8 節)

記号	説明
$\text{Deg}(c)$	退化分布 (2.9 節)
$\text{Emp}(D)$	経験分布 (2.11 節)
$\text{HyperGeom}(n, m, k)$	超幾何分布 (5.2 節)
$\chi^2(m)$	$\chi^2$ 分布 (5.3 節)
$t(\nu)$	スチューデントの $t$ 分布 (5.7 節)
$\text{Beta}(\alpha, \beta)$	ベータ分布 (6.5 節)
$\text{Geom}(q)$	幾何分布 (13.3 節)
$\text{Exp}(\lambda)$	指数分布 (13.4 節)
$\text{Pois}(\mu)$	ポアソン分布 (13.7 節)
$\text{NegBinom}(r, p)$	負の二項分布 (13.7 節)
$\text{Erlang}(r, \lambda)$	アールン分布 (13.7 節)
$H_0$	帰無仮説 (4.2 節)
$H_1$	対立仮説 (4.2 節)
$\mathbf{H}(\boldsymbol{\theta})$	ヘッセ行列 (付録 A.5)
$H(\boldsymbol{\theta})$	エントロピー (8.2 節)
$\hat{\boldsymbol{\theta}}_{\text{ML}}$	最尤推定量 (7.3 節)
$\hat{\boldsymbol{w}}_{\text{LS}}$	最小 2 乗推定量 (9.1 節)
$\hat{\boldsymbol{w}}_{\text{RLS}}$	正則化最小 2 乗推定量 (9.4 節)
$\hat{\boldsymbol{\theta}}_{\text{MAP}}$	最大事後確率推定量 (9.5 節)
$\hat{\boldsymbol{\theta}}_{\text{PME}}$	事後平均推定量 (9.5 節)
$Q_{\text{EM}}(\boldsymbol{\theta} \boldsymbol{\theta}')$	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)
$n_{\text{eff}, i}$	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

1.1	確率的現象	1
1.2	統計解析の目的と限界	2
1.3	確率空間の定義	4
1.4	確率空間の例	6
1.5	確率変数	7
1.6	確率変数の例	9
1.7	確率変数とランダムサンプリングの解釈	10
1.8	確率分布関数と確率密度関数	11
1.9	確率分布の表記	12
1.10	複数の確率変数の同時確率分布	13
1.10.1	同時確率分布	13
1.10.2	周辺分布	14
1.10.3	条件付き確率分布	16
1.10.4	統計的に独立な確率変数	17
1.11	期待値	18
1.12	指示関数	20
1.13	分散と共分散	21
1.14	相関係数	25
1.15	サンプル値からの推定	27
1.16	特性関数	30
1.17	カルバック・ライブラー情報量	31

## 2. 確率分布の具体例

2.1	ベルヌーイ分布	34
2.2	二項分布	35
2.3	カテゴリカル分布	37
2.4	多項分布	39
2.5	1変数正規分布	40
2.6	ガンマ分布	44
2.7	多変数正規分布	47
2.8	一様分布	49
2.9	退化分布	51
2.10	ディラックのデルタ関数	53
2.11	経験分布	53

## 3. 大数の法則と中心極限定理

3.1	観測データの頻度分布	55
3.2	標本平均が従う確率分布	56
3.3	大数の法則	57
3.4	大数の法則の例	59
3.5	大数の法則の極限へ近づく速さ	60
3.6	中心極限定理	61
3.7	中心極限定理の例	66

## 4. 仮説検定とP値

4.1	仮説検定の概念	68
4.2	仮説検定の手順	69
4.3	P値	72

4.4	経験分布を用いた仮説検定	74
4.5	統計的有意性の解釈	75

## 5. 仮説検定の具体例

5.1	二項検定	77
5.2	フィッシャーの正確確率検定	79
5.3	$\chi^2$ 検定と $\chi^2$ 分布	80
5.4	$\chi^2$ 適合度検定	83
5.5	$\chi^2$ 独立性検定	84
5.6	$\chi^2$ 適合度検定の導出	86
5.7	$t$ 検定	89
5.8	スチューデントの $t$ 分布の導出	92
5.9	マン・ホイットニーの $U$ 検定	95
5.10	コルモゴロフ・スミルノフ検定	98

## 6. 多重検定補正と false discovery rate

6.1	多重検定補正の必要性	100
6.2	ボンフェローニ補正	102
6.3	false discovery rate	104
6.4	Benjamini-Hochberg 法	106
6.5	quantile-quantile プロットと順序統計量	107
6.6	Benjamini-Hochberg 法の導出	110
6.7	Benjamini-Yekutieli 法	116
6.8	Storey 法	118

## 7. 確率モデル解析と最尤推定法

7.1	仮説検定の問題と確率モデル解析	124
7.2	尤度	127

7.3	最尤推定法	128
7.4	最尤推定法の例	130
7.5	最尤推定量の漸近的性質	131
7.6	モデル分布の同一性とヘッセ行列	136

## 8. 混合正規分布と期待値最大化法

8.1	混合正規分布	138
8.2	期待値最大化法の原理	140
8.3	期待値最大化法の例	143
8.4	交差検証による成分数の決定	147

## 9. 回帰モデルの正則化とベイズ推定

9.1	多項式回帰と最小2乗法	151
9.2	多項式回帰の確率モデル	157
9.3	過適合	159
9.4	正則化最小2乗法	162
9.5	ベイズ推定	165
9.6	正則化最小2乗法の確率モデルによる解釈	168

## 10. マルコフ過程と平衡分布

10.1	確率過程の定義	171
10.2	マルコフ過程	173
10.3	遷移確率行列の性質	174
10.4	生成消滅過程	176
10.5	マルコフ鎖のランダムサンプリング	177
10.6	$P^n$ の漸近的振る舞いの例	178
10.7	平衡分布	181
10.8	平衡分布からのランダムサンプリング	183

10.9	連続状態マルコフ過程の平衡分布	184
10.10	連続状態マルコフ過程の例	186

## 11. ランダムサンプリングと数値積分

11.1	ランダムサンプリングと乱数生成法	192
11.2	線形合同法	193
11.3	確率分布関数からのランダムサンプリング	195
11.4	棄却法によるランダムサンプリング	197
11.5	確率変数の変数変換を用いる方法	198
11.6	期待値計算と数値積分計算	200

## 12. 事後分布とマルコフ連鎖モンテカルロ法

12.1	事後分布からのランダムサンプリング	202
12.2	メトロポリス・ヘイスティングス法	203
12.3	マルコフ連鎖モンテカルロ法の例	206
12.4	期待値計算と有効サンプルサイズ	209
12.5	提案分布のパラメータ調節の例	212
12.6	ギブスサンプリング	213
12.7	ギブスサンプリングの例	215

## 13. 到達時刻とポアソン過程

13.1	到達時刻の定義	219
13.2	ベルヌーイ過程の例	220
13.3	幾何分布	220
13.4	指数分布	222
13.5	指数分布の無記憶性	224
13.6	無記憶性の証明	225
13.7	ポアソン過程	226

---

付	録	230
A.1	集合と写像	230
A.2	ベクトル空間	232
A.3	行列	233
A.4	微分と積分	238
A.5	関数論	243
引用・参考文献		246
索引		248

## 2 1. 統計解析の目的と確率空間

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

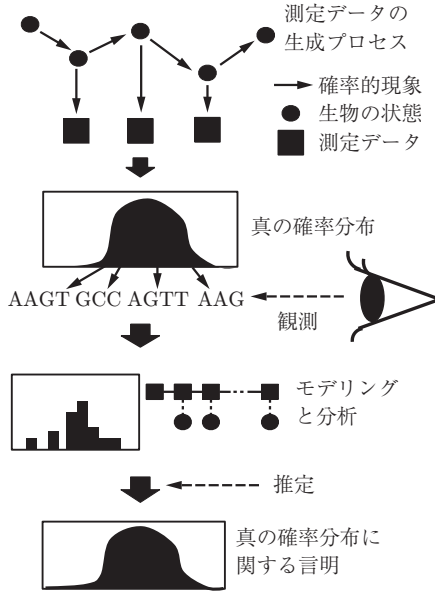


図 1.1 統計解析の目的

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無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。



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

## 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  $\mathbf{e}^{(n)} = (1, \dots, 1)/\sqrt{n}$  となるような正規直交基底  $\{\mathbf{e}^{(h)} \in \mathbb{R}^n | h = 1, \dots, n, \mathbf{e}^{(h)T} \mathbf{e}^{(h')} = \delta_{hh'}\}$  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  $\mathbf{z}$  は正規直交基底の線形結合で表され ( $\mathbf{z} = \sum_h \mathbf{e}^{(h)} y_h$ )、各項の係数は  $\mathbf{z}$  と基底の内積をとり求められる ( $y_h = \mathbf{e}^{(h)T} \mathbf{z}$ )。また、各列に正規直交基底を並べた ( $n \times n$ ) 次元の行列  $\mathbf{O} = (\mathbf{e}^{(1)}, \dots, \mathbf{e}^{(n)})$  は直交行列となり (付録 A.3 を参照)、式 (A.11), (A.12) より  $\mathbf{O}$  は関係式  $\mathbf{O}^T \mathbf{O} = \mathbf{I}$ ,  $|\mathbf{O}| = \pm 1$  を満たす。ただし、 $\mathbf{I}$  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  $\mathbf{z}$  と  $\mathbf{y}$  の内積が等しいことがわかる。

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

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。

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

さらに式 (5.12) で  $\mathbf{z} = \mathbf{O}\mathbf{y}$  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  $|J| = \|\mathbf{O}\| = |\pm 1| = 1$  なので,  $d^n z = d^n y$  が成り立つ。さらに,  $(n-1)$  次元ベクトル  $\mathbf{y}' = (y_1, \dots, y_{n-1})$  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

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

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

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

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

<b>【あ】</b>		確率モデル解析	125	グラフ表現	176
アーラン分布	229	仮説検定	69	クロネッカーのデルタ	233
<b>【い】</b>		過適合	161	訓練データセット	148
一様分布	49	カテゴリカル分布	37	<b>【け】</b>	
<b>【う】</b>		カルバック・ライブラー		経験分布	53
ウェルチの $t$ 検定	91	情報量	31	決定論的手法	193
<b>【え】</b>		ガンマ関数	45	元	230
エビデンス	166	ガンマ分布	44	原始的	237
エントロピー	142	<b>【き】</b>		検証データセット	149
<b>【お】</b>		偽陰性	70	検定統計量	70
オルンシュタイン・		幾何分布	220	<b>【こ】</b>	
ウーレンバック過程	189	棄却する	69	交差検証	149
<b>【か】</b>		棄却法	197	合成写像	231
回帰モデル	152	擬似乱数列	193	勾配	239
解析関数	243	期待値	18	勾配降下法	129
解析的	243	期待値最大化法	140	互換	231
可逆	183	ギブスサンプリング	213	コーシーの積分定理	245
拡散方程式	190	帰無仮説	69	固有値	235
確率	5	帰無分布	70	固有値分解	235
確率過程	171	既約	181, 237	コルモゴロフ・	
確率行列	238	逆行列	235	スミルノフ検定	98
確率空間	4	逆写像	231	根元事象	4
確率測度	5	逆像	231	混合係数	139
確率的現象	1	逆フーリエ変換	242	混合正規分布	139
確率分布	8, 12	偽陽性	70	<b>【さ】</b>	
確率分布関数	11	共分散	22	最小 2 乗推定量	153
確率ベクトル	232	共役事前分布	170	最小 2 乗法	152
確率変数	7	行列	233	最大事後確率推定量	167
確率密度関数	12	行列式	235	最尤推定法	128
		行列積	233	最尤推定量	128
		極座標系	242	差集合	230
		<b>【く】</b>			
		空集合	230		

**【し】**

識別モデル 158  
 次元 233  
 事後分布 166  
 事後平均推定量 168  
 指示関数 20  
 事象 5  
 事象空間 5  
 指数分布 223  
 システムノイズ 2  
 事前分布 165  
 シード 194  
 写像 231  
 周期 237  
 周期的 180  
 集合 230  
 周辺化 14  
 周辺分布 15  
 受容する 69  
 受容率 204  
 順序統計量 108  
 条件付き確率 16  
 条件付き確率分布 16  
 条件付き期待値 19  
 詳細釣り合いの式 183  
 初期状態確率 174  
 真の確率分布 2  
 真部分集合 230

**【す】**

スチューデントの  $t$  分布 89

**【せ】**

正規直交基底 233  
 正規分布 40  
 正行列 237  
 生成消滅過程 176  
 生成モデル 158  
 正則 244  
 正則化項 162  
 正則化最小 2 乗法 162  
 正則化パラメータ 162

正則関数 244  
 正定値 236  
 精度パラメータ 42  
 成分 139  
 積集合 230  
 積分 239  
 説明変数 152  
 遷移確率行列 174  
 遷移確率密度 185  
 線形回帰 152  
 線形合同法 193  
 線形独立 232  
 全射 231  
 線積分 245  
 全単射 231  
 全変動距離 185

**【そ】**

像 231  
 相関係数 25  
 測定ノイズ 2

**【た】**

第 1 種の過誤 70  
 対角行列 234  
 退化分布 51  
 対称行列 236  
 大数の法則 57, 120, 131, 200, 209  
 対数尤度 128  
 第 2 種の過誤 70  
 対立仮説 69  
 多項式回帰 152  
 多項分布 39  
 多重検定補正 101  
 単位行列 234  
 単位ベクトル 232  
 単射 231

**【ち】**

置換 231  
 中心極限定理 62, 81, 88, 131, 200, 209

超幾何分布 80  
 直交行列 236

**【て】**

提案分布 204  
 提案分布比 204  
 定常分布 182, 185  
 定常マルコフ過程 174  
 定数確率変数 51  
 ディラックのデルタ関数 53  
 テイラー展開 243  
 テイラーの定理 243  
 停留方程式の方法 129  
 転置 233

**【と】**

同一性 136  
 統計解析 1  
 統計的に独立 17  
 統計的有意性 75  
 同時確率 13  
 同時確率分布 13  
 到達時刻 219  
 特性関数 30

**【な】**

内積 232

**【に】**

二項係数 36  
 二項検定 78  
 二項定理 36  
 二項分布 36

**【の】**

ノンパラメトリック検定 95

**【は】**

パーミュテーション検定 75  
 半正定値 236  
 半負定値 236

		ベイズ推定	165				
		ベイズの定理	165			<b>【も】</b>	
		ベクトル空間	232		モーメント	31	
		ベータ関数	108		目的変数	152	
		ベータ分布	108		モデル分布	125	
		ヘッセ行列	244				<b>【や】</b>
		ベルヌーイ過程	220		焼入れ	208	
		ベルヌーイ分布	34		ヤコビ行列	241	
		ペロン・フロベニウスの			ヤコビ行列式	241	
		定理	237				<b>【ゆ】</b>
		偏微分	239				
							<b>【ほ】</b>
		<b>【ほ】</b>			有意水準	70	
		ポアソン過程	226		有効サンプルサイズ	210	
		ポアソン分布	227		尤度	127	
		ボックス・ミュラー					<b>【ら】</b>
		変換	199		ラグ	210	
		ボレル代数	6		ランク	234	
		ボンフェローニ補正	102		乱数生成器	192	
					乱数生成法	193	
		<b>【ま】</b>			ランダウの $O$ 記法	243	
		マルコフ過程	173		ランダム経路	171	
		マルコフ鎖	177		ランダムサンプリング	2, 10	
		マルコフ連鎖					<b>【り】</b>
		モンテカルロ法	203		リーマン積分	239	
		マン・ホイットニーの					<b>【る】</b>
		$U$ 検定	95				
							<b>【れ】</b>
		<b>【み】</b>			ルベーグ積分	240	
		右固有ベクトル	235				<b>【わ】</b>
		<b>【む】</b>			レヴィの連続性定理	30	
		無記憶性	224				<b>【わ】</b>
		<b>【め】</b>			和集合	230	
		メトロポリス・					
		ヘイスティングス法	203				
<b>【ひ】</b>							
非決定論的手法	193						
非周期的	181, 237						
微積分学の基本定理	240						
左固有ベクトル	235						
非負行列	237						
微分	238						
微分同相写像	241						
標準正規分布	42						
標準偏差	22						
標本共分散	27						
標本空間	4						
標本自己相関係数	210						
標本相関係数	27						
標本点	4						
標本分散	27						
標本平均	27						
<b>【ふ】</b>							
フィッシャー情報量	135						
フィッシャーの							
正確確率検定	80						
負定値	236						
負の二項分布	227						
部分集合	230						
部分積分	240						
不偏推定量	28						
ブラウン運動	190						
フーリエ変換	242						
分割表	79						
分散	21						
分散共分散行列	24						
分布収束	30						
<b>【へ】</b>							
平均対数尤度	132						
平衡分布	182						

<p><b>【B】</b></p> <p>Benjamini-Hochberg 法 106</p> <p>Benjamini-Yekutieli 法 118</p> <p>BH 法 106</p> <p><b>【E】</b></p> <p>EM アルゴリズム 140</p> <p>E ステップ 141</p> <p><b>【F】</b></p> <p>false discovery rate 104</p> <p>family-wise error rate 102</p> <p>FDR 104</p> <p>FWER 102</p> <p><b>【M】</b></p> <p>M ステップ 141</p>	<p><b>【N】</b></p> <p><math>n</math> ステップ遷移確率行列 175</p> <p><b>【P】</b></p> <p>P 値 72</p> <p><b>【Q】</b></p> <p>quantile-quantile プロット 107</p> <p>Q 関数 140</p> <p>Q 値 122</p> <p><b>【S】</b></p> <p>Storey 法 118</p> <p><b>【T】</b></p> <p><math>t</math> 検定 89</p>	<p><b>【ギリシャ文字】</b></p> <p><math>\sigma</math>-代数 5</p> <p><math>\chi^2</math> 検定 80</p> <p><math>\chi^2</math> 適合度検定 83</p> <p><math>\chi^2</math> 独立性検定 84</p> <p><math>\chi^2</math> 分布 81</p> <p>~~~~~</p> <p><b>【数字】</b></p> <p>0 ベクトル 232</p> <p>1 サンプル <math>t</math> 検定 90</p> <p>1 対 1 の写像 231</p> <p>2 サンプル <math>t</math> 検定 90</p>
--	---	---

—— 監修者・著者略歴 ——

浜田 道昭 (はまだ みちあき)	木立 尚孝 (きりゅう ひさのり)
2000年 東北大学理学部数学科卒業	1997年 京都大学理学部卒業
2002年 東北大学大学院理学研究科修士課程修了 (数学専攻)	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)
2002年 株式会社富士総合研究所研究員	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
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)	2004年 博士(学術)(東京大学)
2010年 東京大学特任准教授	2005年 産業技術総合研究所特別研究員
2014年 早稲田大学准教授	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)
2018年 早稲田大学教授 現在に至る	2009年 東京大学准教授 現在に至る

## 生物統計

Biostatistics

© Hisanori Kiryu 2022

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)



＜出版者著作権管理機構 委託出版物＞

本書の無断複製は著作権法上での例外を除き禁じられています。複製される場合は、そのつど事前に、出版者著作権管理機構（電話 03-5244-5088, FAX 03-5244-5089, e-mail: info@jcopy.or.jp）の許諾を得てください。

本書のコピー、スキャン、デジタル化等の無断複製・転載は著作権法上での例外を除き禁じられています。購入者以外の第三者による本書の電子データ化及び電子書籍化は、いかなる場合も認めていません。落丁・乱丁はお取替えいたします。