

# はじめに

## 統計遺伝学、遺伝統計学とは——遺伝学、ゲノム学と統計学

本書は統計遺伝学・遺伝統計学の教科書として使うことを目的として書きました。統計遺伝学・遺伝統計学は、生命現象の根幹の1つである遺伝という現象を軸に数理的に生命現象を解釈する学問です。親と子は似ています。それは、遺伝子が親から子へと伝達されるからです。よく似てはいますが異なる部分も持っていて、集団で見るとばらつきがあります。このように、似ているけれどもばらついていることを、遺伝子の伝達とその機能の発揮とに照らして理解しようとするのが遺伝学です。

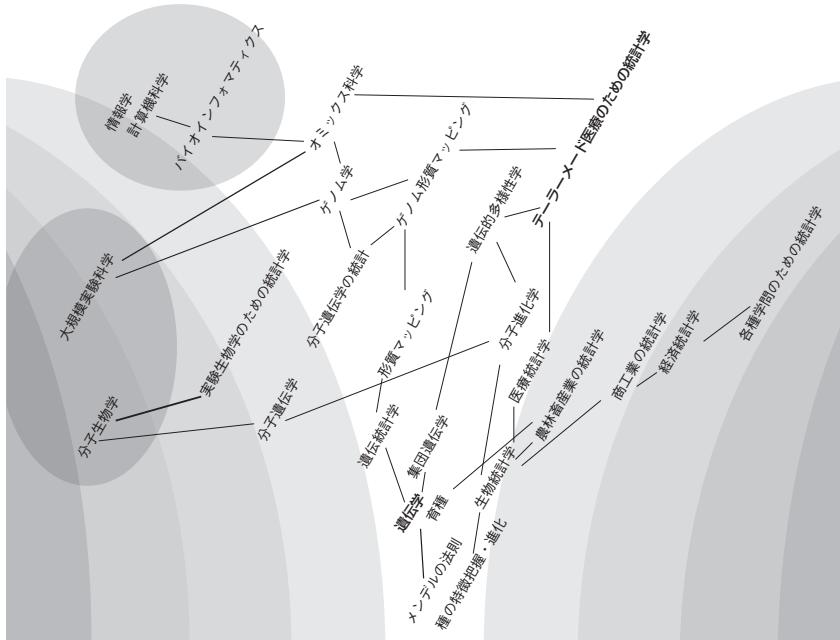
子の遺伝子のセットは、母親と父親から遺伝子のセットの半分ずつを受け継いでいます。逆に言うと、親の遺伝子セットの半分だけが子へと伝わります。この半分の選び方がランダムに起きるので、遺伝現象は確率的です。このため、遺伝学は確率的に現象を理解することを基本とします。

このように、「ばらつき」と「確率」は遺伝学の基本ですが、同じく「ばらつき」と「確率」を基本とする学間に統計学があります。このように遺伝学と統計学は非常に似た部分があり、実際、両者は一体となって発展してきたという関係があります。

現在、遺伝学は、現象としての遺伝から、実体としての遺伝子とその分子機構の解明、さらには、遺伝子の役割を遺伝子セット全体として理解しようとするゲノム学の展開により、大規模データ科学としての側面を強くしています。一方、統計学はありとあらゆる学間に影響を与え、非常に幅の広い学問となっています（図0.1）。

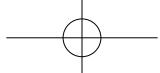
遺伝学・ゲノム学における数理統計学的アプローチは非常に多岐に渡ります

が、本書では、遺伝現象・ゲノム解析に特徴的な事柄として、離散的のこと・確率と尤度のこと・大規模なこと・不均一なことを取り上げ、いろいろな角度から焦点を当てました。



統計遺伝学・ゲノムのための統計学に関与する、さまざまな関連分野を例挙し、その中で関係の強いものを線で結んでいます。ただし、各用語の配置は筆者の主觀を反映しているものであり、参考程度に眺めるのが望ましいでしょう。遺伝学と統計学はほとんど区別のない時期があり、そこから、生物現象・形質の多様性を測定する手法と、生物現象のはらつきのあるデータを用いて検定をする方法論が発展してきました。測定・検定は、一般的な統計学として、生物学分野にとどまらず、すべての科学分野の基盤として发展を遂げてきました。一方、遺伝学・生物多様性のための数理科学は形質マッピング学と集団遺伝学について展開を続けました。他方、生物学・遺伝学は分子生物学の勃興とともに、実験生物学のデータ解析のための統計解析手法の発展をもたらしました。そして分子遺伝学が発展し、それは巨大情報分子であるDNAのための統計学の展開をもたらしました。分子生物学・分子遺伝学は大規模データ科学技術の発展とともに、大規模データ統計学として大きく展開したのです。その中には、長大ゲノムを用いたゲノム形質マッピングとその他のオミックス科学データ解析が含まれます。また、分子遺伝学は進化学と集団遺伝学にも取り込まれ、分子進化学・分子集団遺伝学として、分子生物学・分子遺伝学の諸分野との連携が進みました。この発展には、計算機科学・情報学の発展、その生物分野への応用であるバイオインフォマティクスと密接な関係があります。

図 0.1 縦軸は時間軸、下が過去、上が現在



## 本書の構成

本書は遺伝現象と遺伝子の機能に関する研究について幅広く取り扱い、それらの統計学的側面を理解するときに共通している事柄に焦点を当てて構成しました。

第 I 部では、遺伝・遺伝子に関連する生物学的事項を取り扱います。生物学の基礎知識がある人にとっては内容は平易なので、その内容を例に本書の第 II 部以降で活用するフリーの統計ソフト R の扱いを覚えて欲しいと思います。本書では R の利用・R ソースを理解していることを前提として話を進めます。ごく簡単に R の導入・利用に関して記載しましたが、R に関して全くの初学者が R を覚えられるように記載していません。R の活用のためには後述の参考書・ウェブサイトなどを利用してください。逆に数理統計系の基礎がある人にとっては、遺伝学・生物学の基礎事項の導入の役割を持つと思います。

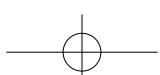
第 II 部は、データの取り扱いの基礎に関する事項を取り扱います。遺伝子の離散性からカテゴリカルデータ型に関して重点的に説明しています。また、データを相対的に評価することについて説明しています。また、個々のサンプルを区別して、その個々のサンプルの間の関係を扱うのか、集団として扱うのか、という視点から、その手法を利用例に即して説明しています。

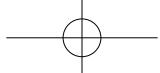
第 III 部は、記述統計に関する事項、推定や検定の基礎となる概念である次元・自由度、確率・尤度、分布、指標について説明しています。この中で、尤度の考え方を活用して特に発達した解析手法である、連鎖解析に関して、少し詳しく説明しています。

データを統計学的に取り扱う道具立ての説明である第 IV 部を受けて、第 V 部では、データからどのようにして意味を引き出すかについて説明しています。具体的には、推定・棄却・検定・関連に関する事項を取り扱っています。

第 VI 部は、データが大規模であるときや、複雑であるときに発生する諸事情に関する事柄を扱っています。場合の数を数え上げること、場合の数が大きすぎると、部分を使って全体に関する情報を取得すること、数多くの検定をするときの解釈方法について順に説明しています。

そして最後の付録で、R を使って勉強するときに便利だと思われる事項と、本書で使用した数式表現をまとめました。





## 本書の使い方

前述した通り、本書は遺伝現象と遺伝子の機能に関する研究の統計学的側面を理解するときに共通している事柄に焦点を当てて構成しました。逆に言えば、遺伝子マッピングや遺伝子機能解析などに使われる諸手法の解説や、各種指標の定義などはあえて扱わないとしました。

それら個々の手法を活用したり、指標を適切に用いなくてはならないことは当然ですが、それらについての理解を目指したのでは、取り扱わなかった事項はわからなくなりますし、今後、開発・提案されるであろう、新規の方法や指標などに対応する力はつかないからです。それよりは、いろいろな手法や指標に共通する考え方は何なのかという基礎に重みをおくことで、応用力をつけて欲しいと願っているからです。語学の机上學習において、基礎を押さえすることで、初めての文章も読みこなせるようになることを目指すようなイメージです。

ですから、本書では、形質マッピングに興味があるならこの章を読むとよい、クラスタリングを知りたいならこの章を、というように構成されています。1つの研究テーマに関する事項があちこちに出てきて、それらを相互参照しながら読むのが最善、という作りになっています。

また、複数回、登場する用語や考え方には、必ずしも初出のところで丁寧に説明することはせず、丁寧な説明が最も有用と思われるところに記載しました。これは、初出時に定義や説明を書くことで、前半が重くなることを避けるためです。したがって、多少、曖昧なままでも、意味が通る限りは読み進み、再び読み返したときに納得がいけばよし、と考えてもらえばと思います。読み進むにあたり、**太字1**でハイライトされた用語がわかりにくいときには、索引でその用語を調べてください。その用語が登場しているページがどこかわかります。複数箇所に登場しているときには、最も丁寧に説明しているページが太字にしてありますので、そのページの説明を読んでください。最も丁寧に説明しているページでは、その用語は**太字2**のように別タイプのフォントで表現しています。ウェブの記事で言えば、相互リンクがあちこちに張ってあるような具合です。索引を介してジャンプするようなつもりで使ってください。

また、図を多用しました。さらに、図を描くためのフリーの統計ソフトRのソースも掲載しました。自然言語の文章で内容を押さえ、図でイメージを湧かせた上で、数式で正確な表現を確認しつつ、Rのソースも理解の助けとし

て欲しいと思います。また、Rのソースはシミュレーションであるものが多いので、是非、使用して、異なる条件での結果を出力してみることをお勧めします。これらの図はカラーで見る方がわかりやすいかもしれません。<http://www.genome.med.kyoto-u.ac.jp/func-gen-photo/index.php?album=StatGenetTextbook>で閲覧できます。

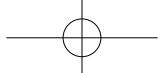
本書で使用したRのソース、は<http://www.genome.med.kyoto-u.ac.jp/StatGenet/lectures/2010/StatGenetTextbook/Rsrc.zip>からダウンロードできます。

### 本書の執筆でお世話になったみなさん

本書を執筆するにあたり、多くの方にお世話になりました。

特に、日々の研究のよろず相談に乗っていただいている私的会合のみなさん（早稲田大学先進理工学部電気・情報生命工学科・井上真郷先生、東京大学医科学研究所ヒトゲノム解析センター・井元清哉先生、第一三共株式会社・遠藤聖子さん、筑波大学大学院人間総合科学研究科・大橋順先生、理化学研究所ゲノム医科学研究センター統計解析研究チーム・熊坂夏彦さん（五十音順））には非常に多くのことを教えていただきました。この場を借りて御礼を申し上げます。また、このメンバーの中でも、特に大橋順先生には、本書の執筆にあたって多岐に渡る多くのご助言をいただきました。本当にありがとうございました。また、京都大学医学部附属病院免疫・膠原病内科の寺尾知可史さん、同学大学院医学研究科附属ゲノム医学センターの川口喬久さんには、原稿・Rソースの確認をしていただくなど、平素の共同研究活動に加えてお世話になるとともに、川口さんには、本書で使用したRソースのデモンストレーションプログラムを書いていただきました。この他、遺伝統計学とは縁遠い立場から、原稿全体に目を通し本の構成についての意見をくれた妻の純子にも感謝しています。そして何よりも、遺伝統計学・統計遺伝学の分野での書籍発行の機会を提供してくださったオーム社様、また、筆を進めながら内容・構成をすいぶんと変えたにも関わらず我慢してお付き合いくださったオーム社開発部の皆様に御礼申し上げます。

このように、多くの方のお蔭で出来上がった本書ですが、少しでも多くの読者の方が遺伝統計学・統計遺伝学に興味を持ち、多様な学問背景の方が本分野に取り組むきっかけとなることを祈っています。



## 参考書

本書を読むにあたっての遺伝学、統計学、Rに関する関連書を挙げます。本書は、ここに挙げる本を理解するための入門書でもなく、ここに挙げた書を読んでから、読むべき専門書でもありません。どちらを先に読んでもよいですが、両方を読むと、理解が深まると思われます。

### ①統計遺伝学

- [1] "Handbook of Statistical Genetics (3rd ed.)", David J. Balding et al. John Wiley & Sons

本書が扱った内容のほぼすべてを網羅的に扱った大部な書。

- [2] "Statistics in Human Genetics", Pak Sham Arnold

コンパクトにまとめた良書。

- [3] 『遺伝統計学入門』 鎌谷直之著、岩波書店

形質マッピングを中心に扱った書。

### ②遺伝学・集団遺伝学

- [1] 『初步からの集団遺伝学』 安田徳一著、裳華房

集団遺伝学の概念を体系的に扱った書。

### ③離散数学

- [1] 『離散数学「数え上げ理論」』 野崎昭弘著、講談社

数え上げについての入門書。

- [2] 『[マクグロウヒル大学演習] 离散数学—コンピュータサイエンスの基礎

数学』 Seymour Lipschutz 著、成嶋弘監訳、オーム社

離散数学を計算機活用を念頭において扱った良書。

### ④確率・統計学

- [1] 『確率概論』 河野敬雄著、京都大学学術出版会

確率を正面から扱った書。

- [2] 『統計学を拓いた異才たち』 ディヴィッド・サルツブルグ著、竹内惠行・

熊谷悦生訳、日本経済新聞社

統計学の発展に関する読み物。

- [3] 『入門ベイズ統計』 松原望著、東京図書

ベイズ統計に関する入門書。

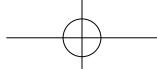
- [4] 『統計分布ハンドブック』 萩谷千鳳彦著、朝倉書店

## ⑤数学

- [1] 『統計学のための数学入門 30 講』永田靖著、朝倉書店  
統計学での利用の面から数学の諸項目を解説した良書。
- [2] 『[数理科学ライブラリ (2)] グラフ理論入門』N. ハーツフィールド・G. リンゲル著、鈴木晋一訳、サイエンス社  
グラフ理論の平易な入門書。

## ⑥ R

- [1] 『Rによる統計解析』青木繁伸著、オーム社  
統計解析一般を R で実施することを念頭においた良書。
- [2] 『Rで学ぶクラスタ解析』新納浩幸著、オーム社  
クラスタ解析を網羅的に概説した入門書。

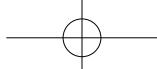


x

# 目 次

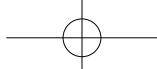
はじめに .....	iii
<b>第Ⅰ部 遺伝子型から表現型まで</b>	<b>1</b>
<b>第1章 遺伝——似ていることと似ていないこと</b> .....	<b>3</b>
1.1 形質が遺伝する .....	3
1.1.1 遺伝 .....	3
1.1.2 生物の特徴——形質とフェノタイプ（表現型）.....	4
1.1.3 同一性と多様性 .....	4
1.2 遺伝子 .....	5
1.2.1 遺伝子とは.....	5
1.2.2 染色体 .....	6
1.2.3 遺伝子座、アレル、ハプロタイプ、ディプロタイプ、フェノタイプ .....	8
1.2.4 2倍体、ホモ接合、ヘテロ接合、ジェノタイプ、フェノタイプ、遺伝形式 .....	10
<b>第2章 DNA、RNA、タンパク質、形質 .....</b>	<b>15</b>
2.1 DNA二重鎖 .....	15
2.1.1 複製、変異、組み換え .....	17
2.1.2 起源が同じ——IBD .....	18
2.1.3 1つの数値で表して扱いやすくする——IBDの期待値.....	19
2.1.4 同胞のアレルの一一致率 .....	22
2.1.5 変異の消長——遺伝的浮動 .....	32
2.2 DNAからRNA、タンパク質へ .....	40
2.2.1 DNAからRNAへ——転写 .....	40
2.2.2 RNAからタンパク質へ——翻訳 .....	40
<b>第3章 多様性の諸相 .....</b>	<b>43</b>
3.1 核酸、タンパク質の多様性 .....	43
3.1.1 DNA配列の多様性、種の違い、遺伝子多型 .....	43
3.1.2 RNAとタンパク質の多様性 .....	46

<b>3.2 多様性と分散.....</b>	<b>49</b>
3.2.1 分散の分解——分散、共分散 .....	49
3.2.2 遺伝率と分散 .....	53
3.2.3 ハーディ・ワインバーグ平衡（HWE）と分散 .....	53
3.2.4 アレル関連、連鎖不平衡と分散 .....	55
<b>3.3 データの取り扱い方と分散、共分散 .....</b>	<b>57</b>
3.3.1 HWE とアレル関連、連鎖平衡を 2 列のデータで考える .....	57
3.3.2 遺伝形式（優性、劣性）は第 3 の列.....	59
<b>3.4 たくさんの要因——多因子遺伝.....</b>	<b>61</b>
<b>第Ⅱ部 データ、サンプル、サンプルの集まり</b>	<b>63</b>
<b>第 4 章 観察して評価すること .....</b>	<b>65</b>
<b>4.1 データの種類と構成 .....</b>	<b>65</b>
4.1.1 遺伝子から見たデータの種類——遺伝子型と表現型、最終形質と中間形質 .....	65
4.1.2 解析対象としてのデータの種類——データ型 .....	67
4.1.3 一部に順序があること .....	70
4.1.4 カテゴリの組み合わせ .....	71
4.1.5 唯一選択、重複選択 .....	71
4.1.6 2 倍体という特殊性——ハーディ・ワインバーグ平衡（HWE）の正確検定 ...	73
4.1.7 親項目と子項目 .....	76
4.1.8 カテゴリの配置、カテゴリ間の非独立性、正単体 .....	77
<b>4.2 データを比較する .....</b>	<b>79</b>
4.2.1 2 つのデータの関係——対称的な関係と非対称的な関係 .....	79
4.2.2 非対称な関係を対称にする——距離 .....	80
4.2.3 ヨークリッド距離とそれ以外の距離 .....	82
4.2.4 配列の違いとマンハッタン距離 .....	83
4.2.5 距離の代わりに角度で表す——相関係数 .....	84
<b>4.3 複数のサンプル、たくさんの比較 .....</b>	<b>84</b>
4.3.1 1 対 $N-1$ と $N$ 対 $N$ .....	84
4.3.2 一部に関係がないとき——半順序 .....	85
4.3.3 距離行列と木 .....	86
<b>第 5 章 サンプルを個別に捉える .....</b>	<b>89</b>
<b>5.1 グラフとは.....</b>	<b>89</b>
5.1.1 グラフの定義 .....	89
<b>5.2 サンプルを並べる——数直線というグラフ .....</b>	<b>90</b>
<b>5.3 木というグラフ .....</b>	<b>91</b>

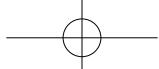


## xii 目 次

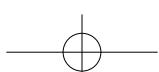
5.3.1 木とは .....	91
5.3.2 木の形状——トポロジー .....	92
<b>5.4 木構造でのデータの理解——階層的クラスタリング .....</b>	<b>93</b>
5.4.1 進化系統樹 .....	93
5.4.2 階層的クラスタリング .....	94
<b>5.5 行列データを眺める .....</b>	<b>98</b>
5.5.1 要素を並び替えて眺める——ヒートマップ .....	98
5.5.2 要素を並び替えずに眺める——連鎖不平衡係数プロット .....	99
5.5.3 片方の軸に着目、両方の軸に着目 .....	100
<b>5.6 個体の家系図、アレルの系図——同一種の中のグラフ .....</b>	<b>101</b>
5.6.1 個人の関係のグラフと染色体の関係のグラフ .....	102
5.6.2 染色体の伝達グラフと組み換え .....	104
5.6.3 祖先にさかのぼる——コアレセント .....	106
<b>5.7 ネットワーク .....</b>	<b>108</b>
<b>第 6 章 サンプルを集団として捉える .....</b>	<b>111</b>
6.1 分布として捉える .....	111
6.1.1 1 次元 .....	111
6.1.2 2 次元 .....	113
6.2 非階層的クラスタリング .....	116
6.3 集団遺伝学 .....	117
6.3.1 不均一と不平衡 .....	117
6.3.2 均一な集団とハーディ・ワインバーグ平衡 (HWE) ——均一な集団の混合 ..	117
6.3.3 時間的な変化 .....	118
6.3.4 空間的移動 .....	122
6.4 热力学、統計力学、流体力学 .....	123
6.4.1 時空間、有限と無限 .....	123
6.4.2 均一、平衡、定常 .....	124
<b>第 III 部 サンプルの集まりの特徴づけ</b>	<b>127</b>
<b>第 7 章 尺度、変数、自由度、次元 .....</b>	<b>129</b>
7.1 データをかいつまんで伝える .....	129
7.1.1 分割表の情報をかいつまんで伝える .....	129
7.1.2 量的データをかいつまむ .....	133
7.2 次元と独立と直交 .....	138
7.2.1 自由度と次元 .....	138
7.2.2 分割表の自由度と線形独立——行列 .....	139

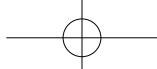


7.2.3 確率的独立と直交 .....	140
7.2.4 線形独立と直交基底 .....	141
7.2.5 正規直交基底を取り出す——固有値分解.....	142
<b>7.3 変数の構造と意味 .....</b>	<b>146</b>
7.3.1 変数の構造.....	146
7.3.2 意味から選ぶ変数——データ構造で決める変数 .....	147
<b>第 8 章 統計量、指數、確率、尤度.....</b>	<b>149</b>
8.1 確率分布 .....	149
8.1.1 分布とは.....	149
8.1.2 離散的な確率分布 .....	150
8.1.3 連続的な場合——指數分布 .....	151
8.1.4 指數分布と正規分布との違い.....	153
8.1.5 一様分布、指數分布、正規分布、矩形分布——一般正規分布 .....	154
8.1.6 正規分布、カイ分布と次元.....	156
8.1.7 カイ分布からカイ自乗分布.....	160
8.1.8 最も観察されそうなカイ自乗値 .....	161
<b>第 9 章 確率と尤度.....</b>	<b>163</b>
9.1 確率、尤度.....	163
9.1.1 確率 .....	163
9.1.2 尤度 .....	166
9.1.3 確率の和は 1、尤度の和は 1 ではない .....	168
9.1.4 尤度の指數化——尤度比と事前確率、事後確率.....	170
9.2 条件付確率、確率、尤度、非独立 .....	171
9.2.1 カテゴリ型の条件付確率.....	171
9.2.2 量的な軸での条件付確率 .....	174
9.2.3 事前確率「当初の予想」と陽性的中率 (PPV) と陰性的中率 (NPV) .....	175
<b>第 10 章 連鎖解析に見る尤度と変数 .....</b>	<b>179</b>
10.1 尤度を使った形質マッピング——連鎖解析.....	179
10.2 パラメトリック連鎖解析と尤度 .....	180
10.2.1 マーカーの伝達木と原因座位の伝達木 .....	180
10.2.2 マーカーと原因座位の間の組み換え .....	185
10.2.3 隠れマルコフモデルと連鎖解析の尤度計算 .....	189
10.3 ノンパラメトリック連鎖解析——罹患同胞対解析.....	192
10.3.1 相対危険度を変数とする .....	192
10.3.2 IBD 数ごとに場合分けする .....	193

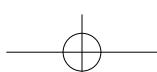


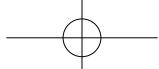
<b>第 11 章 指数（インデックス）とは .....</b>	<b>199</b>
11.1 指数は相対的な値 .....	199
11.2 不平衡の指数.....	200
11.2.1 ハーディ・ワインバーグ平衡 (HWE) .....	200
11.2.2 連鎖不平衡.....	200
11.2.3 $p$ 値——確率変数を指數化する .....	201
<b>第IV部 推定、仮説、棄却、関連、因果</b>	<b>203</b>
<b>第 12 章 推 定 .....</b>	<b>205</b>
12.1 最尤推定.....	205
12.2 信頼区間.....	206
12.3 いろいろな信頼区間 .....	210
12.3.1 ベイズ推定——観察していないとき.....	212
12.3.2 ベイズ推定、事前確率、共役事前分布 .....	213
12.3.3 多項分布とその共役事前分布——ディリクレ分布 .....	216
12.3.4 最尤推定とハプロタイプ頻度推定——連鎖不平衡係数推定.....	220
12.4 EM アルゴリズム .....	225
<b>第 13 章 棄却と検定.....</b>	<b>229</b>
13.1 信じるのが難しい仮説を棄却する——3 カテゴリの観察.....	229
13.2 分割表検定.....	232
13.2.1 ピアソンの独立性検定——カイ自乗検定.....	236
13.2.2 帰無仮説と最尤仮説を比較して統計量にする——尤度比検定.....	237
13.3 3 つの検定方法の比較——正確確率検定、ピアソンの独立性検定、尤度比検定 ..	239
13.3.1 サンプル数が小さいときと大きいとき .....	239
13.3.2 検定の対称性 .....	243
13.3.3 有限な範囲と無限の広がりの違い .....	244
13.3.4 計算量の違い .....	244
13.3.5 計算量の違いのまとめ .....	244
13.4 仮説に制約を定めて検定する.....	245
13.4.1 1 つの分割表にいろいろな検定を適用してみる .....	247
13.4.2 離散的な仮説空間での尤度比の比較 .....	251
13.5 検定同士の非独立な関係.....	251
13.6 表のサイズを変える .....	254
13.6.1 表形式のデータ .....	254
13.6.2 順序のあり・なしと検定手法.....	254





13.6.3 複数の手法の挙動の比較 .....	257
<b>第 14 章 関係と因果.....</b>	<b>261</b>
14.1 原因と結果と時間 .....	261
14.2 原因としてのジェノタイプ .....	262
14.3 有向グラフ、ベイジアンネットワーク .....	263
<b>第V部 大規模なこと</b>	<b>269</b>
<b>第 15 章 数え上げる .....</b>	<b>271</b>
15.1 順列、重複順列、分割表の正確生起確率 .....	271
15.1.2 組み合わせ、重複組み合わせ、2倍体ジェノタイプの種類数 .....	274
15.2 分割の数——スターリング数とベル数 .....	276
15.3 分割とカテゴリの統合 .....	279
15.3.1 順序のないカテゴリの場合 .....	279
15.3.2 順序のあるカテゴリの場合 .....	279
15.4 木の形の数、グラフの数——木、クラスタリング、ベイジアンネットワーク ...	280
15.4.1 木のパターンの数 .....	280
15.4.2 クラスタリングのパターンの数 .....	281
15.4.3 無向グラフの数、有向グラフの数、非循環有向グラフの数 .....	283
<b>第 16 章 省略する .....</b>	<b>287</b>
16.1 ランダムに抽出する、ランダムに巡回する .....	287
16.1.1 既知の分布からのランダムサンプリング .....	287
16.1.2 サンプルを使ってランダムサンプリング、リサンプリングとパーミュテーション ..	288
16.1.3 ランダムウォーク .....	292
16.2 主要な部分のみを使う .....	294
16.2.1 近似する .....	294
16.3 意義の大きい方から選ぶ、小さい方から捨てる .....	299
<b>第 17 章 たくさんの検定 .....</b>	<b>301</b>
17.1 多重検定 .....	301
17.1.1 独立な検定の繰り返し .....	301
17.1.2 多重検定時の $p$ 値の期待値 .....	301
17.1.3 一番小さい $p$ 値の補正 .....	302
17.1.4 非独立な検定の繰り返し .....	306
17.1.5 モンテカルロ・パーミュテーションによる多重検定補正 .....	306
17.1.6 非独立な検定を繰り返したときの最も小さい $p$ 値 .....	309





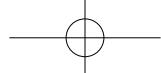
## xvi 目 次

17.2 <i>p</i> 値が均一に分布しないとき .....	309
17.2.1 <i>p</i> 値が小さめに出るとき——ジェノミックコントロール法 .....	309
17.2.2 対立仮説が成り立つとき——非心カイ自乗分布 .....	313
17.2.3 検定のパワー .....	316
17.3 たくさんの結果の分布を活用する .....	318
17.3.1 主成分分析を使って補正する .....	318
17.3.2 帰無仮説が必ずしも棄却されるべきではないとき .....	326
17.4 複数の結果を合わせる——メタアナリシス .....	330
17.4.1 相互に独立な検定を合わせる .....	330
17.4.2 2 表を単純に足し合わせる .....	334
17.4.3 メタアナリシス .....	337

## 付 錄 345

---

付録 A R .....	347
A.1 R のインストールと起動と終了 .....	347
A.2 R のパッケージを使う .....	348
A.3 本書の R のソースを利用する .....	349
A.3.1 本書で用いた R の関数 .....	350
A.4 R のヘルプを出す関数を使ってみる .....	352
A.4.1 R のソースを確認する .....	353
A.5 確率分布関数、疑似乱数列の発生 .....	356
付録 B 数式記号 .....	357



# 第 I 部

## 遺伝子型から表現型まで



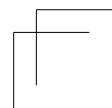
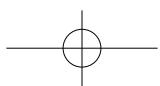
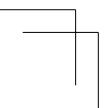
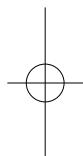
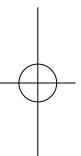
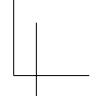
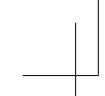
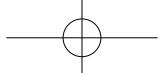
第1章 遺伝——似ていることと似ていないこと

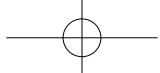
第2章 DNA、RNA、タンパク質、形質

第3章 多様性の諸相

第I部は遺伝現象の生物学的な基礎をまとめます。生物学・遺伝学を背景にしている人は数理解析的処理とフリー統計解析ソフト「R」の利用法についてを、数理解析を背景にしている人は生物学・遺伝学に関して整理してください。







# 第 1 章

## 遺伝——似ていることと似ていないこと

### 1.1 形質が遺伝する

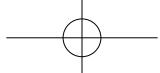
#### ◎ 1.1.1 遺伝

遺伝とは何でしょうか。血のつながりがあるときに、ある特徴が共有されることです。人間の親から人間の子が生まれたとき、人間であるという特徴は親子で共有されています。このことがあまりにも当たり前であったとき、人間は遺伝という現象に言葉を与えることはなかったと思います。遺伝という現象に単語を与えたということは、それが当たり前とは言えなかつたからです。

人間の親から人間の子が生まれるその傍らで、「カエルの子はカエル」であり、「瓜の蔓に茄子はならな」かったために、種という特徴が遺伝していることに気づいたのでしょうか。しかしながら、種に気づいただけでは、遺伝現象に気づいたとは言い難いです。ここで挙げたカエルと瓜のことわざも、「鳶が鷹を生む」という逆の意味のことわざも、「親子は似るのが普通であって、ごく稀に親子に際だった違いが見られる」という人間の親子の観察の結果です。これらのことわざが示すように、

「血のつながりがあるときに、ある特徴について似たり似なかつたりするものだけれども、似ることの方が似ないことよりも多い」

ということに気づいたことをもって、遺伝現象に気づいたと言えるでしょう。



さらには、

「遺伝する特徴もある一方で、遺伝しない特徴もある」

ことに気づいたときに、遺伝現象には原因があるだろうと気づいたと思われます。これが遺伝学の第一歩です。

### ◎ 1.1.2 生物の特徴——形質とフェノタイプ（表現型）

生物が持つ特徴は形質<sup>※1</sup>と呼ばれます。特徴はなんでも形質です。どんな形質に興味を持つかが、研究の内容を決めますし、新たな解析の視点を持つことは、新しく形質を定義することとも言い換えられます。そういう意味で、形質を考える視点を整理することは有用です。

形質を分類するときの1つの方法は、五感（視覚・聴覚・味覚・嗅覚・触覚）のどれによる評価であるかが挙げられます。別の分類の仕方としては、解剖学・構造的、生理・機能的、分子生物学・薬理学的という分け方があります。また、個体そのものに備わった特徴であるのか、個体が外敵や環境との関係の持ち方（行動・免疫反応など）なのか、という分類もできます。その他には、数学的な概念なのか、物理的に測定するものなのか、変化する（化ける）様子なのか、という視点も形質の分類に役立ちます。これらの形質を観察することでデータが得られます。

この形質の観察データがフェノタイプ（表現型）です。

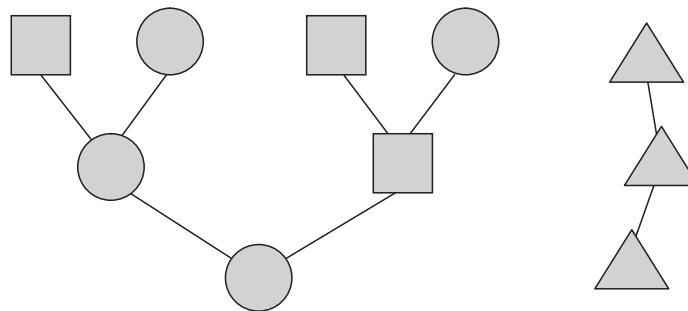
### ◎ 1.1.3 同一性と多様性

生物は自分と同じものを次世代に残すことを1つの特徴とします。無性生殖<sup>※2</sup>では子は親の遺伝子をそのまま引き継ぎますが、ヒトをはじめとする有性生殖<sup>※3</sup>の生物では全く同一の個体を残すわけではなく、両親の遺伝子の半分ずつを引き継ぎます（図1.1）。

※1 太字の単語は索引掲載語です。索引には、その単語を丁寧に説明したページ（索引の太字のページ）が示されているので、ウェブのリンクをたどるように索引を介して、適宜、説明をたどってください。

※2 無性生殖：1つの個体が単独で新しい個体を生み出すような生殖形式。

※3 有性生殖：男と女という異なる性別の個体から構成される種がとる生殖戦略で、子の遺伝子は両親の遺伝子の半分ずつからなる。



□はオス、○はメス、△は性別がない個体を表します。親子の関係はグラフで表されます。有性生殖では、オスとメスの対から子が生まれます。無性生殖では、性別のない親から子が生まれます。図では、線で結ばれた個体は、上にあるのが親、下にあるのが子です。

図 1.1 有性生殖と無性生殖

**適応**という面から言うと、親がうまく生きているわけだから、子も親と同じならうまくいくんだろうと考えた戦略と言えるかもしれません。他方、個体ではなく、個体が属する種に着目してみます。種の個体がどれもみな同じだと、環境が変化したときに、全個体がそろって生きにくくなることもあるでしょう。それは得策ではないと考えることができます。環境の変化があっても、いろいろな個体がいれば、環境の変化が起きても種の全滅する可能性は減りますから、いろいろな個体で構成しておくことも得策と考えられます。遺伝現象は、このように、同一性の確保と多様性の確保の二面性を持っています。

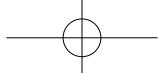


## 1.2 遺伝子

### ● 1.2.1 遺伝子とは

遺伝現象は、親子関係と形質との間の関係です。なぜ、親子関係があると形質が似るのか、それが何によってもたらされるのかわかりませんでした。その遺伝現象をもたらす何かの基本単位が**遺伝子**です。

わからないなりに、遺伝現象の原因となる実体の存在が信じられ、それはいくつもあると予測されたので、遺伝現象の基本単位を遺伝子としました。現在の生物学で言えば、遺伝現象の原因となる実体が**ゲノム**であり、ゲノムの構成要素が遺伝子と言えるでしょう。遺伝子は親子の間で共有されるものであるべ

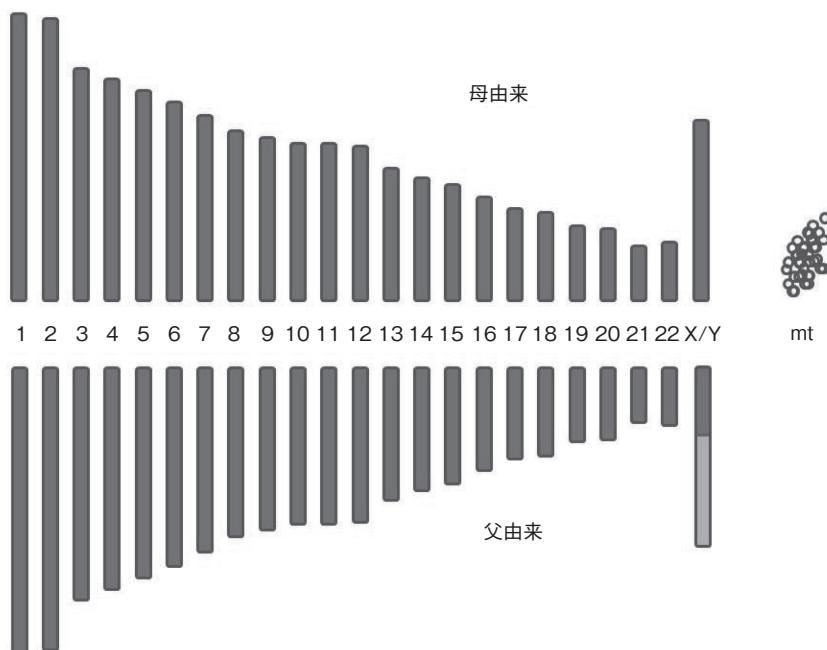


きです。習慣や環境も親と子の間で共有されますが、個体の始まりは受精卵であり、受精卵は母親の体内で母親からの卵子と父親からの精子とが合わさってできます。卵子と精子は、それぞれ母と父からの遺伝情報を染色体として持ち込むので、ここに遺伝子があるはずです。

実際、この染色体が遺伝子のすべてではないかもしないものの、大部分を担うと考えられています。染色体は長いDNA分子とそれをとりまくタンパク質からできています。DNA分子は第2章で説明する通り「情報を担う物質」として優れた特長を持っています。

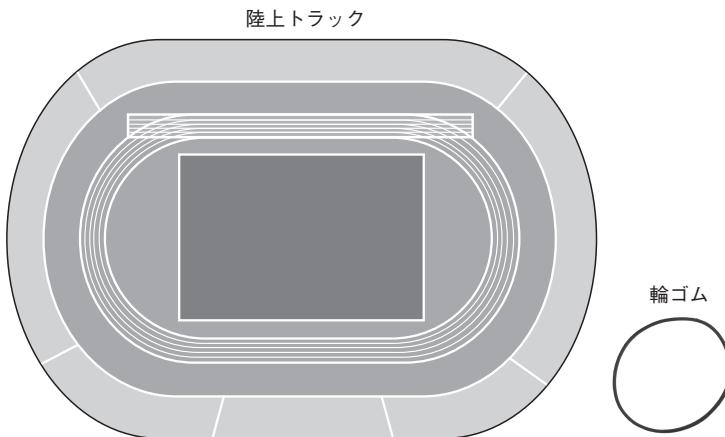
### ◎ 1.2.2 染色体

**染色体**は非常に長い紐のような構造になっています。受精に際して、卵子と精子から持ち込まれた染色体のセットは、新しい個体の遺伝情報となります。ヒトの身体は多数の細胞でできていますが、すべての細胞は原則として、同じ染色体のセットを持ちます（図1.2）。



(a) 上段・下段はそれぞれ母由来・父由来のセットです。1番から22番までの常染色体と性染色体のX・Y、mt（ミトコンドリア染色体）とが示されています。

図1.2 染色体

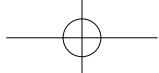


(b) 核染色体とミトコンドリア染色体の大きさのイメージ図です。ミトコンドリア染色体（約 15,000 塩基対）を輪ゴムとすると、核染色体の短いものだと、1 本（約 0.5 億塩基対）は陸上トラックを 1 周（400 メートル）するくらいの 1 本の紐に相当します。長い核染色体だと、トラック 5 周分くらいになります。ミトコンドリア染色体は 1 細胞あたり 100 から 10,000 個あるので、箱売りの輪ゴム数箱から数十箱分の輪ゴムがあることになります。

図 1.2 染色体（つづき）

染色体は大きく 2 つに分けられます。**核染色体**と**ミトコンドリア染色体**です。核染色体は 22 対の**常染色体**と 1 対の**性染色体**とからなります。性染色体は X, Y の 2 種類で女は X を 2 本、男は X と Y を 1 本ずつ持ります。受精に際して、卵子からは 1 セットの核染色体（22 本の常染色体と 1 本の性染色体（X 染色体））と多数のミトコンドリア染色体が、精子からは 1 セットの核染色体（22 本の常染色体と 1 本の性染色体（X もしくは Y 染色体））が持ち込まれます。父親からはミトコンドリア染色体は受け継ぎません。対となる常染色体はよく似ていますが、2 本の性染色体はかなり違います。X 染色体と Y 染色体は共通の部分も持いますが、それぞれに特異的な部分があり、長さも大きく異なります。

ミトコンドリア染色体は、紐の両端がくっついた輪の形（環状）をしていますが、この環状染色体はバクテリアの染色体に見られる特徴です。はるか昔に真核細胞が今のミトコンドリアの祖先であるバクテリアを細胞内に取り込んだことに由来しています。ミトコンドリア染色体は核染色体に較べて非常に小さく、長さにして 1 万分の 1 程度ですが、個々の細胞は多くのミトコンドリア染色体を持ちます。その数は細胞によってばらつきがあり、100 から 10,000 個くらいです。



## ■ R を使ってみる

本書ではフリーの統計ソフト R のソースを使って内容の理解を助けることにしています。R のソースが読めなくても、内容の枠を理解することは可能ですが、R のソースを読み取ることを前提として本書は書いてあります。18.1 節を参考に、是非、R のインストールをして、R を使えるようになってください。R を覚えるためには、「はじめに」に挙げた R 関連書や、"R 統計" という検索語でインターネット検索をすると必要な情報は得られると思います。

本書に掲載した R のソースは、<http://www.genome.med.kyoto-u.ac.jp/StatGenet/lectures/2010/StatGenetTextbook/Rsrc.zip> からダウンロード可能です。R をインストールして起動したのち、ダウンロードしたコマンドをコピー & ペーストして下さい。`StatGenetDemo.R` というファイルを読み込ませると、すべてのソースファイルが順次実行されます。R を覚えたい場合には、面倒くさくとも、コマンドを自分で打つことも非常に有用です。ヒト染色体の大きさを表す棒グラフを描いてみます（R1-1.R）。

### R ソース 1.1 R1-1.R : ベクトルに値を代入し、棒グラフを描く（関数 barplot()）

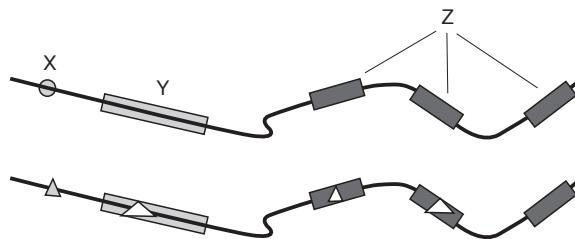
※左端の数値は行数、それ以外が R のコマンド

```
1 # 染色体の長さ ( 単位 : 塩基対数 )
2 #chromLen に 24 個の数値のベクトルを代入する
3 chromLen<-c(247249719,242951149,199501827,191273063,
180857866,
4 170899992,158821424,146274826,140273252,135374737,
134452384,
5 132349534,114142980,106368585,100338915,88827254,78774742,
6 76117153,63811651,62435964,46944323,49691432,154913754,
57772954)
7 #barplot(棒グラフ) を描く。データは chromLen、棒の名前は、
1:22(1,2,...,22) と "X", "Y" のベクトル。色は黒
8 barplot(chromLen,names=c(1:22,"X","Y"),col="black")
```

### ◎ 1.2.3 遺伝子座、アレル、ハプロタイプ、ディプロタイプ、フェノタイプ

遺伝情報は染色体が担っています。遺伝子は遺伝情報の単位ですが、この単位はひも状の染色体の特定の部分に存在します。この位置のことを**遺伝子座（ローカス）**と言います。染色体とその主要構成要素である DNA 分子は線状

の構造をしているので、遺伝子座は、線状の地図の位置として表現されます。遺伝子は、遺伝情報を担うものであって、染色体・DNA 上の位置で表すことができるものである、とも言い換えられるかもしれません。



Xは長さが1塩基、Yは複数塩基の長さのある線分、Zは複数の線分の集合であるような遺伝子座です。△は上下で異なる塩基の位置を示しています。

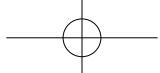
図1.3 X, Y, Zが遺伝子座

ここで言う遺伝情報は、あるタンパク質のアミノ酸配列に関する情報の書かれた範囲全体を指してもよいですし、その発現制御領域のことを指してもよく、また、それらを合わせたものを指しても、反対にそれらのごく一部だけを指してもよいです。

後述するように染色体の主要構成成分であるDNA分子は、4種類の塩基(A, T, G, C)と呼ばれる部品が線状に並んだ構造をしていますが、その塩基1つ1つにも遺伝情報はありますし、それが100万個連なった塊にも情報があります。長さが1塩基の部分を遺伝子座とみなし、染色体によって異なる塩基が対応しているとき、そこの遺伝子座を**1塩基多型 (SNP)**と呼び、その座を占める塩基のバリエーションを**アレル**と言います(図1.3のX)。

ある長さにわたって遺伝子座とみなし、染色体によってその遺伝子座の塩基の並びが異なるとき、その配列の1つ1つがアレルで、**ハプロタイプ**とも呼びます(図1.3のY)。

また、タンパク質はアミノ酸がつながった構造をしていて、その情報がDNA上にコードされていますが、そのDNA上の情報は連続してコードされていることもあります。いくつかの線分に分かれて並んでいることもあります。そのコードしている線分の1つ1つを**エクソン**と呼び、エクソンとエクソンの間を**イントロン**と呼びます。このエクソンとイントロンとで構成されるDNA配列全体を遺伝子座とすることもできます(図1.3のZ)。



このように、遺伝子座に配列のバリエーションがあるとき、そのバリエーションの1つ1つがアレル（**対立遺伝子**）です。

DNA配列のタイプがアレルですが、遺伝子座の取り方によって別名があります、名前の付け方が違ったりします。

遺伝子座の単位	1塩基	塩基の組み合わせ・長さのある配列	遺伝子
アレルの別称		ハプロタイプ	遺伝子タイプ
例	"A", "G"	"AGCT", "CCAT", "0011", "1101"	"野生型", "変異型X", "0401", "タイプ2"

染色体が遺伝子座に持つタイプがアレルです。個体は、染色体のセットを持ちます。常染色体の場合には、個体は両親に由来する2つのアレルを持ちます。個人が遺伝子座に持つタイプ（アレルの組み合わせ）が**ジェノタイプ（遺伝子型）**です。常染色体の場合には2つのアレルの組み合わせのタイプなので、**ディプロタイプ**とも呼びます。

単位	1塩基	塩基の組み合わせ・長さのある配列	遺伝子
例	"A, A", "A, G"	"AGCT, AGCT", "0011, 1101"	"野生, 野生", "0401, *0901", "2, 2"

個体はジェノタイプを持つ一方で、フェノタイプも持ります。

これらは、単位をDNA分子とするか、そのペアとするか、個体とするかの違いがありますが、それぞれの単位でのタイプのことです。その名称がアレル（ハプロタイプを含む）、ジェノタイプ（ディプロタイプ）、フェノタイプです。

単位	染色体	染色体のペア	個体
名称	アレル（ハプロタイプ）	ジェノタイプ（ディプロタイプ）	フェノタイプ

### ● 1.2.4 2倍体、ホモ接合、ヘテロ接合、ジェノタイプ、フェノタイプ、遺伝形式

常染色体は父母由来の1本ずつの対で存在します。遺伝情報を重複して2組持っていることを意味します。このように2組持っている生物を**2倍体**と言います。ある遺伝子座について同じアレルを2つ持っているとき、その個体はその遺伝子座について**ホモ接合体**であると言い、異なるアレルを持っていると

きには、**ヘテロ接合体**であると言います。

ここから先では、2倍体生物のみを考えます。ヒトも2倍体です。2倍体生物は、常染色体上の遺伝子座にはアレルが2つあります。この2つのアレルの持ち方を個体のジェノタイプと言います。これに対して、個体の形質の様子がフェノタイプです。ジェノタイプがフェノタイプに影響を与える関係の中で最も明快な関係に、**優性遺伝形式**と**劣性遺伝形式**という関係があります。これについて考えます。

今、2つのアレル M と m があり、ある特徴を持つか持たないかの区別として、「あり」と「なし」というフェノタイプがあるとします。個体は MM, Mm, mm のいずれかのアレルのパターン（ジェノタイプ）でアレルを持ちます。アレル m があるフェノタイプをもたらす因子であるとします。

表1.1 ジェノタイプとフェノタイプの関係：2 × 3表

	MM	Mm	mm	計
あり	$n_{10}$	$n_{11}$	$n_{12}$	$n_{1\cdot}$
なし	$n_{20}$	$n_{21}$	$n_{22}$	$n_{2\cdot}$
計	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot \cdot}$

$n_{ij}$  の  $i, j$  はそれぞれ行と列の番号を表していて、特に列番号は m の本数を表しています。また、最右の列は行に関する和、最下行は列に関する和を表しています。

$$\sum_{i=1}^2 n_{ij} = n_{\cdot j}; \sum_{j=0}^2 n_{ij} = n_j; \sum_{i=1}^2 \sum_{j=0}^2 n_{ij} = n_{\cdot \cdot};$$

記号  $\Sigma$  については、付録 B を参照。

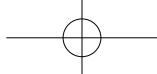
フェノタイプが「あり」である率  $f_j = \frac{n_{1j}}{n_{\cdot j}}$ ,  $j = 0, 1, 2$  はジェノタイプごとに異なります。これはジェノタイプ別の**浸透率**と呼ばれます。

MM を基準として、各ジェノタイプの浸透率が何倍高いか ( $\lambda_j = \frac{f_j}{f_0}$ ) は、

**ジェノタイプ相対リスク (genotype relative risk: GRR)** と呼ばれます。 $\lambda_0 = 1$  です。

	MM	Mm	mm
GRR	$\lambda_0 = 1$	$\lambda_1$	$\lambda_2$

ヘテロ型 (Mm) の GRR がホモ型 (mm) のそれと等しいとき ( $\lambda_1 = \lambda_2$ )、



アレル m は 1 本でも 2 本でも同じ強さでフェノタイプに影響を与えていることがあります、このとき、m のこの表現型への影響の仕方を優性遺伝形式と呼びます。逆に、ヘテロ型の GRR が基準のホモ型 (MM) のそれと等しい ( $\lambda_1 = \lambda_0$ ) とき、アレル m は 2 本揃って初めてフェノタイプに影響を与えていくことになります。この形式を劣性遺伝形式と呼びます<sup>※4</sup>。

優性とも劣性ともつかない中間的な形式には大きく分けて 2 つの定義がありますが、

$$\lambda_1 = x\lambda_2 + (1-x)\lambda_0$$

と表せば、 $x=0$  が劣性、 $x=0.5$  が**相加的 (additive)**、 $x=1$  が優性です。

$$\lambda_1 = \lambda_2^y \times \lambda_0^{1-y}$$

と表せば、 $y=0$  が劣性、 $y=0.5$  が**相乗的 (multiplicative)**、 $y=1$  が優性です。

$x=0.5$  のときは、

$$\lambda_1 = \frac{\lambda_0 + \lambda_2}{2}$$

( $\lambda_0$  と  $\lambda_2$  の相加**平均**<sup>※5</sup>) となるので、additive (相加的) モデル、

※4 今、着目しているフェノタイプを持つことに関して M が優性の形式を持つときには、「このフェノタイプを持たないこと」に関して、m が劣性の形式を持ち、M が劣性の形式を持つときには、m が優性の形式を持ちます。

※5 平均と言えば、

$$\text{相加平均 (算術平均)} : A = \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

$$\text{乗平均 (幾何平均)} : G = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}} = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}}$$

$$\text{の他に、調和平均} : H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

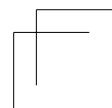
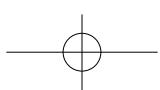
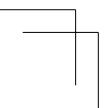
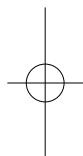
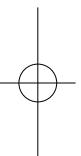
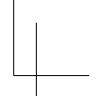
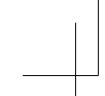
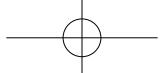
がありますが、調和平均に基づいた遺伝形式の議論はありません。2 要素の場合には、 $G = \sqrt{AH}$  であることや、一般化した形式  $M_p = \left( \frac{1}{n} \sum_{i=1}^n x_i^p \right)^{\frac{1}{p}}$  の平均の定義（一般化平均）を使えば相加平均、相乗平均、調和平均は、 $M_1, M_\infty, M_{-1}$  であるとして捉えておくことは悪くないです。

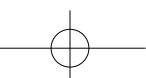
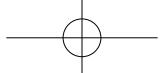
$y=0.5$  のときは、

$$\lambda_1 = \sqrt{\lambda_0 \times \lambda_2}$$

( $\lambda_0$  と  $\lambda_2$  の相乗平均) となるので、multiplicative (相乗的) モデルと呼びます。

また、アレル m を 1 本持つ効果が 2 本持つ効果以下であるときは、 $0 \leq x, y \leq 1$  ですが、アレル m を 1 本持つ効果が 2 本持つよりも強いことあることもあります、それは**超優性 (overdominance)** と呼ばれます。 $x, y > 1$  の場合に相当します。 $x, y < 0$  の場合はヘテロ型が着目している表現型を持たないことについて超優性である場合になり、 $x, y \in (-\infty, \infty)$  により、**GRR**に基づき、すべての遺伝形式が網羅できることになります。





## 第 2 章

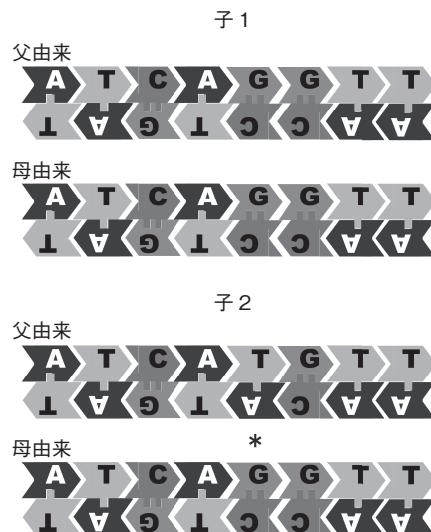
# DNA、RNA、タンパク質、形質

### 2.1 DNA 二重鎖

染色体は非常に長いDNA分子と多くのタンパク質からなっています。DNA分子は塩基と呼ばれる4種類の化学構造体(ATGC)を一列に並べた構造をしており、この4種類の塩基が作る文字列を塩基配列と言います。遺伝情報はこの塩基配列と、DNA分子の化学修飾状態によって伝えられる場合がほとんどです。特に塩基配列は明快に情報を保持し伝えることができ、遺伝子の本体とも言えるので、ここから先は、遺伝情報は塩基配列が伝えるものとして話を進めます。

DNAの配列はATGCの4アルファベットの連なりであり、その配列には向きがあります。順方向と逆方向の2本の塩基配列が、対をなしてよじれることで、二重鎖という構造をとります。対をなすためには、4文字に対をなす規則が必要です。AとTがペア、GとCがペアになります。対になる2つの文字列は異なっているので、DNA二重鎖にある塩基配列には、2通りの読み方があることになります(図2.1)。

ある配列に対して、それと対をなす配列を相補鎖配列と言います。



塩基の部品は向きと対になるための凹凸を持たせています。最上段の父由来のDNA分子は "ATCAGGTT" という配列と、その逆の "AACCTGAT" という配列の2通りの読み方を持っています。子1の父由来・母由来のセットは、全く同一なので、子1は、この配列範囲について、すべてホモ接合体です。子2の母由来のDNA分子は、左から右への読み方向での5番目の塩基(\*)がTではなくGに変わっています。DNA二重鎖では、対をなす塩基は決まっているので、この分子ではもう1つの配列がAからCに変わっています。子2は、この箇所の塩基の対が父由来と母由来とで異なっていますからヘテロ接合体です。

図 2.1 ある両親から生まれた2人の子のDNA配列

DNAが二重鎖として、常に両方向の文字列を持っていることには少なくとも、2つの意味がありそうです。第一に、2本の文字列は相互にもう片方の文字列に関する情報のすべてを持っていることから、情報のバックアップとしての役割を果たします。これは、DNA分子の複製をしたり、壊れかけたときに修復をしたりするときに役立ちます。第二に、いつでも読み取れる状態の情報の量を2倍にしておくことができます。1本鎖として配列を保存していた場合には、その鎖の配列の情報はすぐに読み取れますし、逆方向の配列の読み方は、相補配列がどうなっているかを作るひと手間がかかります。二重鎖として持つていれば、どちらの向きの鎖が持つ情報もすぐに取り出すことができます。

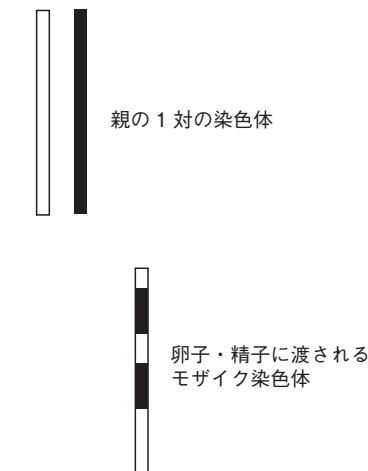
### ● 2.1.1 複製、変異、組み換え

1 個の受精卵から多数の細胞が作られ、個体が形成されるとき、DNA 二重鎖は複製されてすべての細胞に同じ染色体のセットが渡されます。

個体の一生の間には DNA 配列に変化することがあります。その変化は修復されることが普通ですが、修復されずに残ることもあり、そのような変化を**体細胞変異**と呼びます。体細胞変異はその細胞から分裂してきた細胞に受け継がれます。癌というのは、この体細胞変異が原因で同じ DNA 変異を共有する細胞の数が増え、不都合が生じた状態のことです。この体細胞変異は子に伝えられることはできません。他方、卵子と精子を作る過程でも DNA が複製されますが、その途中に**変異（生殖細胞変異）**と呼ばれて体細胞変異と区別されます）が起きます。生殖細胞変異は卵子・精子が作られるときに確率的に一定数は必ず起きると考えてよいです。この変異は受精卵に伝えられ、子の細胞のすべてに引き継がれます。

また、親から子への DNA 配列の伝達にあたっては、親の DNA 配列の半分が伝えられます。第 1 章で説明したように、親は対になった 2 本ずつの染色体を持っていますが、そのうちの 1 本分が卵子・精子に引き継がれ、その合体が受精卵の新しい遺伝子のセットを作ります。ここで、2 本のうちの「1 本」が伝えられるのではなく、「1 本分」が伝えられると言いました。これは、卵子・精子に渡す DNA 配列のセットを作るときに、2 本の対になる染色体の一部分ずつを継ぎ合わせて、1 本分の染色体にしてそれを引き渡すことを意味しています。

図 2.2 に示したように、親は 1 対の染色体を持ちます。その 2 本を白と黒で表しています。卵子・精子へは 1 本分の染色体が渡されます。渡される 1 本は白と黒とが交代したモザイクになります。1 本分として過不足ないことができます。白黒の交代する部分では、染色体が**交叉**したと言い、白黒が交代することを**組み換え**と言います。2 本の染色体が作ったモザイク状の染色体を**組み換え体**と言います。交差する箇所は同じではないので、モザイクの様子が異なります。



上段の2本はある個体の染色体の対です。黒が母親由来、白が父親由来とします。下段は、この個体からできる卵子または精子の染色体です。黑白のモザイク1本分が受け継がれています。

図 2.2 染色体の対

### ◎ 2.1.2 起源が同じ——IBD

最初に常染色体上有る1つの遺伝子座について考えます。これは染色体が1対しかないような生物での遺伝について考えることと同じです。一卵性双生児は、同一の受精卵を起源にしているので、2アレルの起源が確率1で一致しています。親子の関係の場合、子は1アレルを片親から引き継いでいますが、もう片方のアレルは別の親から引き継いでおり、その起源は異なっていると考えることにすれば（近親婚でないと仮定すれば）、1アレルの遺伝子起源が一致する確率が1で、2アレルの起源が一致する確率は0、いずれのアレルも一致しない確率も0です。兄弟姉妹（同胞）の場合は、片親から引き継いだアレルが一致する確率が0.5ですので、両親から引き継いだアレルがともに一致する確率は、 $0.5 \times 0.5 = 0.25$ 。両親から引き継いだアレルがそろって起源を異なる確率も同様に0.25。2アレルのうち片方が起源を同じにする確率がその残りで0.5です。

このようにアレルの起源が同じであることを、IBD（identity by descent）と言い、IBDである数（IBD数）は0か1か2かのいずれかになります。その確率は血縁関係によって決まります。ですから遺伝的な近さを、IBD

数の確率によって表現することができます（表 2.1）。

表 2.1 血縁関係と IBD 数の確率の関係

血縁関係	IBD 数別確率			IBD 数の期待値	一致率	
	2	1	0		期待値	分散
自身・一卵性 双生児	1	0	0	2	1	0
親子	0	1	0	1	$\frac{1}{2}$	0
同胞	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1	$\frac{1}{2}$	$\frac{1}{8}$
祖父母 - 孫	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{64}$

### ● 2.1.3 1 つの数値で表して扱いやすくする —— IBD の期待値

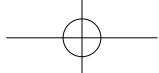
IBD 数が 0, 1, 2 になる確率は、確かに、遺伝的な近さの情報として有用ですが、血縁関係ごとに 3 つの数値を持つベクトルを情報として扱うのは、不便なこともあります。血縁関係の近さに興味があるときには、1 つの数値で表されているのが便利です<sup>※1</sup>。ここでは、3 つの値が確率変数（3 つの場合ですべて尽くされ、それぞれを観測する確率を足し合わせると 1 になる）であることから、確率変数の代表値として有用な期待値を代表的な値としてみます。

期待値とは、いろいろな値をとりうるときに、それぞれの値をとる確率で重み付けをした相加平均のことです。 $X = \{x_1, x_2, \dots\}$  という値を  $P = \{p_1, p_2, \dots\}$  ( $\sum_i p_i = 1$ ) という確率でとるときに、期待値  $E(X)$  は

$$E(X) = \sum_i p_i x_i$$

で表されます。ですから、自身・一卵性双生児の場合に  $E(X) = 2$  となります。自身の血縁関係の強さは、最強の血縁関係だと考えてこれを 1 とすることにして、 $\frac{E(X)}{2}$  としてみるのも得策でしょう。これをアレルの一致率と呼ぶこ

※1 血縁関係を 1 つの数値で表す方法には、近縁係数（kinship coefficient）や相関係数など、遺伝学・集団遺伝学ではいくつかの異なる定義と名称とがあります。それぞれの定義や意味がありますが、一番大切なことは、「複数の値のセットで表されている情報から 1 つの代表的な値に変えて、理解を助けていくこと」です。



とにします。表 2.1 の一致率の期待値がこの値です。

親子と同胞とを較べてみることにしましょう。これらは IBD 数が 2, 1, 0 である確率が、 $(0, 1, 0)$  と  $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$  と異なっています。しかしながら、アレルの一致率の期待値はどちらも  $\frac{1}{2}$  で同じになっています。3つの値で表されていた血縁関係の強さを、1つの数値で表したことにより、ある側面で便利になりましたが、情報が失われたことがわかります。

### (1) 平均、分散、モーメント、期待値

値  $X$  が値  $\{x_1, x_2, \dots\}$  を確率  $P = \{p_1, p_2, \dots\}$  ( $\sum_i p_i = 1$ ) でとるとき、**期待値**  $E(X)$  は、

$$E(X) = \sum_i p_i x_i$$

で表されます。このような分布の  $X$  の期待値は、分布の**平均**とも言います。

**分散**  $V(X)$  は値のばらつきを表す指標で、

$$V(X) = \sum_i p_i (x_i - E(X))^2$$

と算出されます。

平均も分散とともに分布の様子を説明する指標です。これらを統一的に説明するものに**モーメント**があります。モーメントはより一般的な分布の指標で、平均と分散はとともにモーメントのうちの1つです。モーメントは次数と中心のとり方で決まります。中心のとり方としては原点 (O) か平均かの2通りを考えるのが普通です。

$c$  を中心とした  $k$  次モーメントは

$$\mu_k(c) = \sum_i p_i (x_i - c)^k$$

で表され、原点を中心としたそれは

$$\mu_k(0) = \sum_i p_i x_i^k$$

で表され、平均を中心としたそれは

$$\mu_k(E(X)) = \sum_i p_i(x_i - E(X))^k$$

です。平均とは、原点を中心とした1次のモーメントになります。分散は、平均を中心とした2次のモーメントです。

モーメントをRで計算することで、**Rの関数の作り方**を覚えておくことにします。

値のベクトル(x)とその確率のベクトル(p)、次数(order)、中心を原点にするか平均にするか(center=FALSEは平均にしない、TRUEはする)を引数としてモーメントを計算する関数はR2-1.R、R2-2.R、R2-3.Rの通りです。作成した関数の中で、その関数自体を呼び出していることに注意してください。

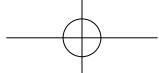
#### R ソース 2.1 R2-1.R : モーメントを計算する関数

```

1  # 関数を作るときには、function() 関数を使います
2  #'momentProb' という名前の関数を作ります
3  #function() 関数は、引数を()の中に、処理を{}の中に書きます。
4  #xが値ベクトル、pがその確率ベクトル
5  #引数のうちx,pは"="で値を指定していませんが、
6  #orderとcenterは値が指定されています
7  #x,pは引数を与えないといけませんが、orderとcenterは
8  #値を与えなければ、デフォルト値(1, FALSE)が用いられます
9  momentProb<-function (x, p, order = 1, center = FALSE)
10 {
11   if(center) # 平均を中心とするならxの値から平均を引く
12   x <- x - momentProb(x,p,order=1,center=FALSE)
13   sum(x^order*p)
14 }
15 # 使ってみる
16 datax<-c(1,2,3)
17 datap<-c(0.2,0.3,0.5)
18 momentProb(x=datax,p=datap,order=2,center=TRUE)

```

学校でのテストの点数などから平均や分散を計算するときには、点数のデータが「標本」であることから、標本平均、標本分散と呼ばれます。その場合には、すべての標本の確率を等しいものとして、計算します。

**R ソース 2.2 R2-2.R : 標本モーメントを計算する関数**

```
1 momentX<-function (x, order = 1, center = FALSE)
2 {
3   # すべての標本に等確率 rep(1,length(x))/length(x) を与えた
4   # momentProb() と同じ
5   momentProb(x,p=rep(1,length(x))/length(x),order=order,
6               center=center)
6   # length(x) は x の要素数、rep(v,L) は v が L 個並んだベクトル
}
```

ここでは、全員の点数を平等に評価しているとも言えますし、全員の確率が等しいと考えて、全員の確率の和が 1 になるように、各人の確率を定め、各人の値の次数乗にその確率を掛けている、とも言えます。

同胞のアレルの一致率の平均と分散を求めるコマンドは R2-3.R のようになります。

**R ソース 2.3 R2-3.R : 同胞の IBD 数の期待値、分散を求める**

```
1 x<-c(1,0.5,0) # IBD 数別の一致の値
2 pDoho<-c(1/4,1/2,1/4) # 同胞の IBD 数別確率
3 momentProb(x,pDoho,order=1,center=FALSE) # 期待値
4 momentProb(x,pDoho,order=2,center=TRUE) # 分散
```

この結果が表 2.1 に記されています。一卵性双生児、親子、同胞、祖父母 - 孫のそれぞれで分散は  $0, 0, \frac{1}{8}, \frac{3}{64}$  です。親子と同胞の一致率の期待値はともに 0.5 で同じですから、両者の IBD 数分布の違いは、期待値には現れず、分散に現れているといえます。

### ● 2.1.4 同胞のアレルの一致率

#### (1) 染色体は 23 対

染色体が 1 対の場合を考えました。1 つの遺伝子座について考えたとも言えます。実際には、ヒトのゲノム DNA は 23 対の染色体・DNA 分子に分かれています。今、単純化して考えるために、 $k$  対の常染色体のみが存在し、それらは同じ長さを持つものとします。 $k$  対の常染色体はそれぞれ父母起源の 1 本ずつです。この起源の一致の程度を考えることにします。一卵性双生児の場合に

は、 $k$ 対、 $2k$ 本の常染色体のすべてで一致率が1です。親子の場合は、 $k$ 対、 $2k$ 本のうち、 $k$ 本分<sup>※2</sup>が必ず一致していて一致率が0.5です。したがって、一致率は、一卵性双生児では平均1、分散0、親子では平均0.5、分散0です。

さて、同胞についてです。 $k$ 対、 $2k$ 本分のうち、どれくらいが一致するでしょうか。IBD 数を0, 1, 2として表してきましたが、表し方を次のように変えてみます。同胞が両親から引き継ぐアレルについて、それぞれ、同胞間で起源が一致した場合に1、一致しなかった場合を0で表すと、第1番目の染色体の1対、2本が一致する場合は"11"、父由来は一致して母由来は一致しない場合を"10"、逆に父由来は一致せず母由来が一致する場合を"01"、両方とも一致しない場合を"00"と表すことにします。それぞれの場合の確率は等確率で0.25ずつです。0, 1の数字の並べ方（重複順列）は、 $2^2$ の4通りで、各々の確率は $\frac{1}{2^2}$ でした。一致率が1, 0.5, 0の確率がそれぞれ0.25, 0.5, 0.25です。

$k$ 対、 $2k$ 本の染色体について考えると $2k$ 個の0, 1の並べ方について考えることになります。並べ方は $2^{2k}$ 通りあり、各々の確率は $\frac{1}{2^{2k}}$ です。 $2k$ 本の染色体のすべてで起源が同一になる場合というのは、"11...11"という並びのみですから、その確率は $\frac{1}{2^{2k}}$ です。これが、一致率1の確率です。起源の同じ $2k$ 本のうち $i$ 本の起源が一致する場合というのは、一致率が $\frac{i}{2k}$ の場合ですが、これは、 $2k$ から $i$ をとる組み合わせになるので、その確率は、

$$\binom{2k}{i} = \frac{(2k)!}{i!(2k-i)!}$$

と $\frac{i}{2^{2k}}$ との積

$$\frac{i}{2^{2k}} \binom{2k}{i}$$

---

※2 親から子に伝わる染色体はモザイクですから、1本が一致するのではなく、「1本分」が一致します。

になります（R2-4.R の `sibIdProb()` が算出関数）※3。

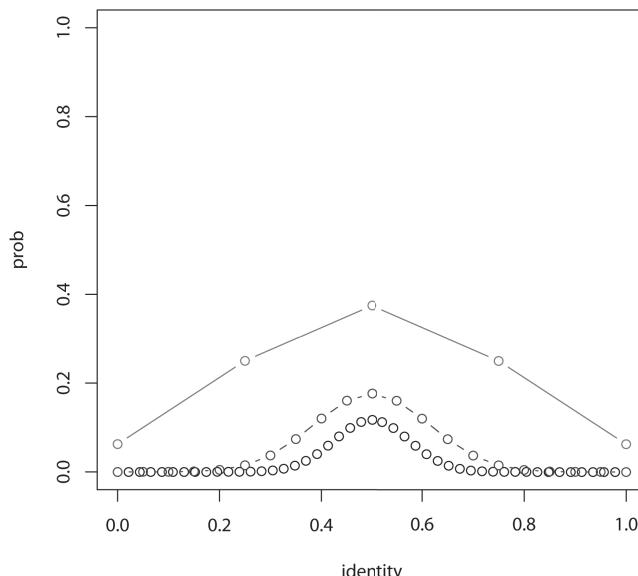


図 2.3 ゲノムを構成する染色体の長さがすべて同じで、本数が 1, 5, 23 本であると仮定したときの、同胞のアレルの一致率の分布

染色体数が 1, 5, 23 の場合を図 2.3 に示しました。山が高い方から、染色体数が 1, 5, 23 です。染色体数が増えるに連れて、山が細かくなると同時に、一致率が 0.5 付近である確率が高くなっています（R2-4.R）。

染色体数が多いということは、同胞間の違いの程度のばらつきを少なくする効果があることがわかります。どの同胞同士を較べても、だいたい半分くらいが一致していて、極端に一致度が高かったり、低かったりする間柄は稀になっています。

---

※3 二項分布： $1 = 1^k = (p + (1-p))^k = \sum_{i=0}^k A_i p^i (1-p)^{k-i}$  としたときに、 $A_i$  を二項係数と言います。

$$A_i = \frac{k!}{i!(k-i)!} = \frac{k(k-1)\dots 1}{i(i-1)\dots 1 \times (k-i)(k-i-1)\dots 1}$$

### R ソース 2.4 R2-4.R : 同胞一致率分布の計算

```

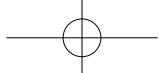
1  # 等長染色体 k 本のときに i=0,1,...,2k 本が一致した場合の一一致率
2  sibIdValue<-function(k=1){
3    (0:(2*k))/(2*k) # i/(2k)
4  }
5  # 2k 本のうち i=0,1,...,2k 本が一致する確率
6  # dbinom() 関数については、18.4 確率分布関数、疑似乱数列の発生を参照
7  sibIdProb<-function(k=1){
8    dbinom(0:(2*k),2*k,0.5)
9  }
10 numch<-1:23 # 染色体数の例として 1 から 23
11 means<-vars<-rep(0,length(numch))
12 for(i in 1:length(numch)){ # ループ処理
13   identity<-sibIdValue(numch[i])
14   prob<-sibIdProb(numch[i])
15   if(i>1){par(new=TRUE)} # 図を重ねて描く
16   plot(identity,prob,type="b",ylim=c(0,1))
17   means[i]<-momentProb(identity,prob,order=1,center=FALSE)
18   # 平均
19   vars[i]<-momentProb(identity,prob,order=2,center=TRUE)
19   # 分散
20 }
21 plot(means) # 期待値をプロット
21 plot(vars) # 分散をプロット

```

## (2) 染色体の長さはばらばら

前項では、染色体の長さが同じものとして計算しましたが、実際には、図 1.2 のように長さは異なります。長さを実際の染色体のそれにして考えてみます。染色体の長さが違うと、長い染色体が一致した場合と、短い染色体が一致した場合とでは、一致率への貢献の程度が異なるので、 $2k$  本のうち  $i$  本が一致した場合でも、長い染色体で  $i$  本なのか、短い染色体ばかりで  $i$  本なのかでは一致率が違ってきます。これをシミュレーションで確かめてみます。乱数を発生させてシミュレーションしてみる練習として、以下のことを行ってみることとします。

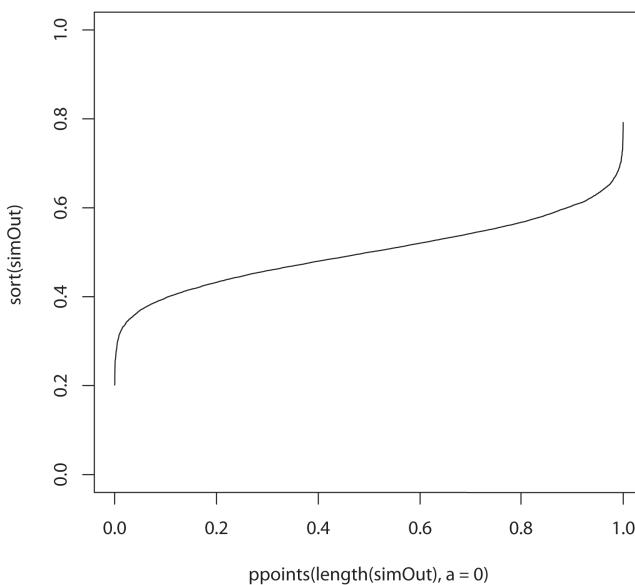
今、常染色体と X 染色体とを対で計 46 本持っているとします。それぞれの一致不一致を二項分布からの乱数 (`rbinom()`) でランダムに割り当てて、`Niter=10,000` 回数 (Niter 同胞ペア) のシミュレーションを行います。比較



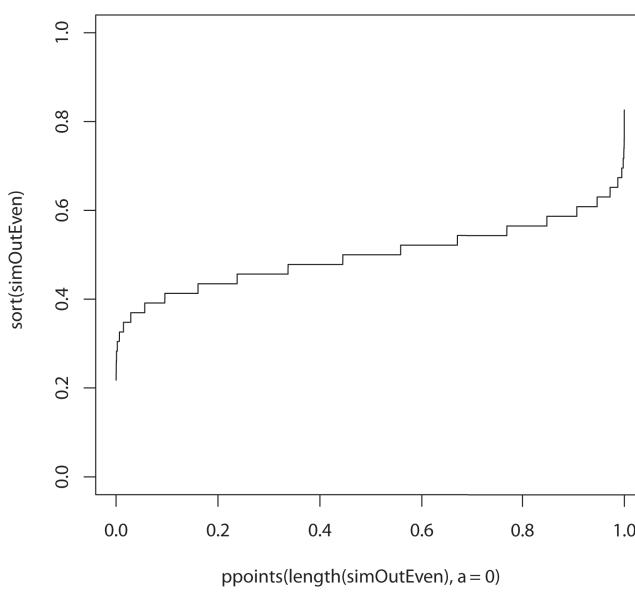
のために、同じく 46 本の等しい長さの染色体の場合も同様に実験します。この結果、どちらの場合も一致率の平均は 0.5 に非常に近く、同程度であることがわかります。また、分散は、等長の染色体よりも、長さが本当の染色体の方が大きいようです。10,000 回の一一致率の分布を見るために、一致率をソートしてプロットしてみます (R2-5.R)。

R ソース 2.5 R2-5.R : 23 対の染色体の長さが実際の長さである場合と等しい長さである場合とで、同胞一致率の分布を比較する

```
1  # f: 個々の染色体が占める割合 ; Niter: シミュレーション回数を与えて、  
# 一致率がいくつになるかをシミュレーションする  
2  SibSim<-function(f=f,Niter=10000){  
3  # 0(非共有)か1(共有)かの乱数の行列を作る  
4  # (行数: 試行回数行、列数: 染色体本数)  
5  rs<-matrix(rbinom(Niter*length(f),1,0.5),nrow=Niter)      # 行列作成  
6  # 行列の外積。共有染色体の割合を全染色体について合算  
7  # 亂数が1のときに一致したとして、その染色体の分の割合をかける  
8  rs%*%f # 外積  
9  }  
10 # 1,2,...,X染色体を2本ずつで46本  
11 chs<-rep(chromLen[1:23],2) # chromLen は R1-1.R で与えたもの  
12 f<-chs/sum(chs) # 個々の染色体の占める割合  
13 Niter<-10000 # 試行回数  
14 simOut<-SibSim(f,Niter=Niter)  
15 # 染色体が等長だと...  
16 cheven<-rep(1,length(f))  
17 feven<-cheven/sum(cheven)  
18 simOutEven<-SibSim(feiven,Niter=Niter)  
19 # 横軸を0-1に納めた上で、Niter回試行の一一致率を昇順ソートしてプロット  
20 plot(ppoints(Niter,a=0),sort(simOut),type="l",ylim=c(0,1))  
# ppoints() 確率点生成 sort() ソート  
21 plot(ppoints(Niter,a=0),sort(simOutEven),type="l",  
ylim=c(0,1))
```

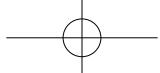


(a) 染色体の長さがヒト染色体の実際の長さでのシミュレーション結果



(b) 等長でのシミュレーション結果

図 2.4 23 対の染色体を持つときの同胞間の一致率の分布



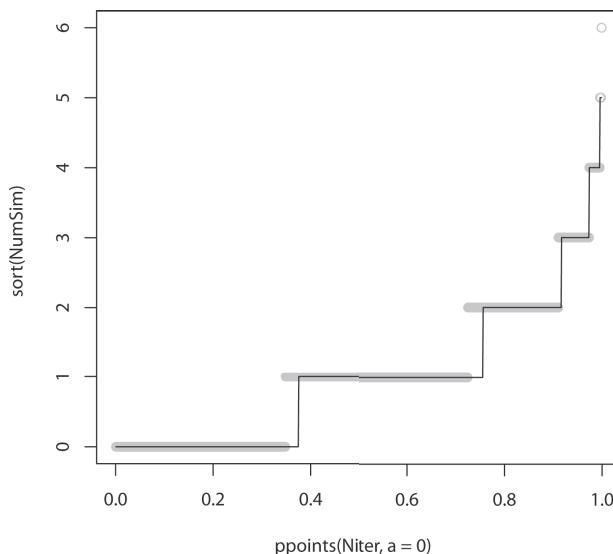
平均と分散など、分布の様子を要約した情報では両者に大きな違いはありませんが、グラフにしてみて初めてわかる情報があります（図2.4）。染色体の長さに忠実に行なったシミュレーション（a）では、曲線がかなり滑らかなのに対して、等長の場合（b）が階段状になっていることです。「等長」の場合には、一致率としてとりうる値が限定されるために、離散的なパターンをとるのであります。ここで確認しておきたいのは、次の3点です。

- データはプロットしてみると見逃す情報があること
- 特殊な条件（今回は「等長」）には注意が必要であること
- 离散的な性質は分布に影響を与えること

### （3）染色体は23対よりももっと細かくなる——交叉・組み換え：ポアッソン過程と指數分布

これまで、2本ある常染色体のどちらが子に伝えられるかは、0.5の確率で決まるものとし、さらに、染色体2本のうちの片方が伝えられると決まれば、その染色体全体が伝えられるものとしてシミュレーションしてきました。実際には、ペアとなる染色体は交叉・組み換えという現象を通じて、父由来・母由来の染色体の一部が置き換わったモザイク状の染色体が作られて伝達されます。したがって、この交叉・組み換えの結果、23本の2倍（46ピース）以上に細切れになったものとして、一致率を考える必要があります。

**交叉**は、対となる染色体同士が細胞内で並んで起こります。卵子・精子に染色体のセットを渡すにあたり、各染色体の対の間で1箇所以上、何箇所かで交叉が起こるとされています。XとY染色体も、大きさがずいぶん違いますが、一部がよく似た配列（相同）であり、その部分で寄り添い、交叉が起きることが知られています。このように、指で数えられる程度に稀に起きる事象がランダムに起きるとき、そんな事象の起き方を**ポアッソン過程**と言い、その起きる回数が従う分布として、**ポアッソン分布**が知られています。



灰色丸は交叉をランダムに発生させたときの分布、黒線はポアッソング分布から乱数発生の結果。

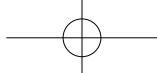
図 2.5 ポアッソング分布による交叉箇所数分布の近似

ある染色体の長さが  $L$  塩基とします。長さ  $L$  の DNA は、 $L - 1$  節所の塩基と塩基のつなぎ目があります。このつなぎ目のどこでも等確率で交叉が起きるものとし、その確率を  $r$  とします。平均交叉回数は

$$meanN = (L - 1) \times r$$

です。

交叉回数の分布をシミュレーションしてみます (R2-6.R)。すべての交叉可能箇所で交叉が起きるかどうかをシミュレーション（全箇所シミュレーション）して、交叉箇所数を数え上げる方法と、平均回数が  $meanN$  のポアッソング分布から発生させた交叉箇所数とする方法の 2 方法を実施してみます。その結果が図 2.5 です。大きめの灰色丸が作るグラフが全箇所シミュレーションでの分布です。その上に、細線で描いたのが、ポアッソング分布からの交叉箇所数の分布です。よく似ています (図 2.5)。



## R ソース 2.6 R2-6.R : 交又回数のシミュレーション

```
1 # 可能箇所すべてで交叉が起きるかどうかを試す方法
2 RecombSim<-function(L=10000,r=0.001,Niter=1000) {
3 # Lは配列長 ,rは箇所あたりの交叉確率 ,Niterはシミュレーション試行回数
4 # 行数 Niter、列数 L-1 箇所の行列にする
5 m<-matrix(rbinom((L-1)*Niter,1,r),nrow=Niter)
6 apply(m,1,sum)
7 }
8 # ポアソン分布を使う方法
9 RecombPois<-function(L=10000,r=0.001,Niter=1000) {
10 rpois(Niter,(L-1)*r)
    # rpois() 関数については、18.4 確率分布関数、疑似乱数列の発生を参照
11 }
12 L<-10000;r<-0.0001;Niter<-1000
13 NumSim<-RecombSim(L=L,r=r,Niter=Niter)
14 NumPois<-RecombPois(L=L,r=r,Niter=Niter)
15 ylim<-c(0,max(NumSim,NumPois))
16 plot(ppoints(Niter,a=0),sort(NumSim),ylim=ylim,
       col=gray(6/8))
17 par(new=T)
18 plot(ppoints(Niter,a=0),sort(NumPois),type="l",ylim=ylim)
```

交叉・組み換えを起こしてモザイク状の染色体を作つて渡すということは、図 2.6 に示すように、2 本のモザイクができるときに起きた交叉場所ごとに交互に IBD かそうでないかが決まります。

図 2.6 の上の 2 本は、染色体のモザイク状態を白黒パターンで示したもので



染色体 1 と染色体 2 はそれぞれ白黒のモザイクです。2 本の染色体が IBD であるかどうかは白黒が一致するかどうかで決まり、それを最下段「同じ？」に示しています。染色体 1, 2 のそれぞれで交叉が起きるたびに、IBD か否かの交代する様子がわかります。

図 2.6 2 本の染色体が IBD であるかどうか

す。モザイク染色体ができるにあたって起きた交叉箇所は、染色体 1 が 2 箇所、染色体 2 が 3 箇所です。下は、2 本のモザイクの IBD である範囲を灰色、そうでない範囲を白としたものです。モザイク染色体 2 本の IBD のパターンは、2 本の染色体の交叉箇所を合わせたすべてで、交代している様子がわかります。

今、1 本の染色体に平均して  $k$  回の交叉が起きてモザイク化するとすると、2 本のモザイク染色体の IBD の割合を計算するためには、平均して  $2k$  回の交叉箇所を考えればよいです。交叉・組み換えが起きないものとして染色体数を増やしたところ、一致率の分散が小さくなることを観察しました。交叉・組み換えがあると、染色体数を増やしたのと同じ効果が出ますから、同胞間の一致率の分散は、さらに小さくなることになります。

さて、交叉の箇所数を考えましたが、今度は、交叉と交叉との間隔について考えてみます。交叉の箇所数をシミュレートするときに、ポアッソン分布からの乱数を使いました。以下のように、交叉間距離（染色体の端から交叉までの距離）をシミュレーションで発生させてみた上で、その分布が指數分布になっていることを確認してみます（R2-7.R、図 2.7）。

#### R ソース 2.7 R2-7.R : 交叉間距離のシミュレーション

```

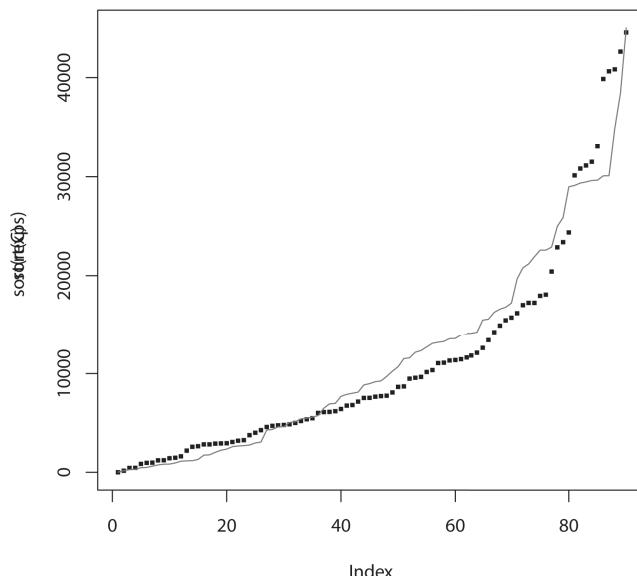
1 Niter<-1000 # シミュレーション回数
2 L<-1000000 # 染色体の長さ
3 r<-0.0001 # 塩基間あたりの交叉確率
4 # 交叉箇所数をポアッソン分布からの乱数で指定し、
   # 交叉箇所を sample 関数で指定する
5 crosses<-sort(sample(1:(L-1),rpois(1,(L-1)*r),
   replace=FALSE))
6 # 交叉間距離のベクトルを作る
7 A<-c(0,crosses) # 染色体の始点と交叉箇所のベクトル
8 B<-c(crosses,L) # 交叉箇所と染色体の終点のベクトル
9 C<-B-A # 交叉間距離のベクトル
10 # 平均が mean(C) の指數分布からの乱数を length(C) の数だけ
    # 発生させてプロット
11 rexps<-rexp(length(C),1/mean(C))
   # rexp() 関数については、18.4 確率分布関数、疑似乱数列の発生を参照
12 # 交叉間距離をソートしてプロット
13 ylim<-c(0,max(C,rexps))
14 plot(sort(C),ylim=ylim,cex=0.5,pch=15) # 交叉間距離の昇順プロット

```

```

15 par(new=T)
16 plot(sort(rexps), col="red", ylim=ylim, type="l")
  # 指数分布からの乱数の昇順プロット

```



ドットは交差を塩基間にランダムに発生させて観測した交叉間距離の昇順プロットです。線は指數分布からの乱数の昇順プロットです。

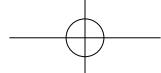
図 2.7 染色体にランダムに交叉が起きたとき、交叉間距離の分布は指數分布で近似される

### ◎ 2.1.5 変異の消長——遺伝的浮動

生殖細胞が染色体のセットを引き継ぐとき、必ず、ある数の変異が起き、塩基配列が変わります。変異率は非常に小さいので、この変異を受け継いだ子は、この変異箇所についてヘテロ接合体であるでしょう。すると、この子が次世代へと引き継ぐ卵子または精子が、この変異を持つ確率は 0.5 です。この子が 2 人の子を持つとします。子 2 人がともにこの変異を引き継ぐ確率は 0.25、2 人のうち 1 人が引き継ぐ確率は 0.5、2 人とも引き継がない確率は 0.25 です。

このように、発生した変異は、1 世代すら引き継がれずに消滅することもあります。集団の中で頻度が高くなることもあります。このようにアレルの頻度は、ランダムに上下します。このような変動を **遺伝的浮動** と言います。

では、実際に発生した変異が消滅するかどうかというのはどのような動きを



するのでしょうか？ 今、人口が  $N$  人（染色体数が  $2N$  本）で固定しているとします。 $2N$  本の染色体があり、それぞれの染色体は、0 か 1 か……何本か不明ながら、何本かを次世代に残し、その総本数が  $2N$  だということです。

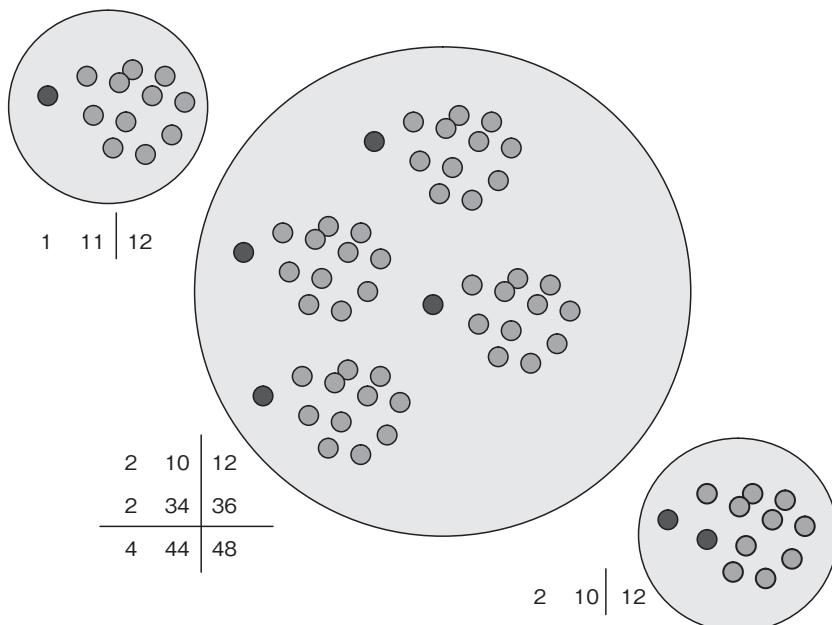
すべての個体が最大  $k$  人くらいの子を次世代に残せる（すべての染色体が最大  $k$  本のコピーを残せる）という制限を設けて、その上で、どの染色体が多く残るか少なく残るかは、ランダムに決まるというモデルで考えてみることにします。

### (1) 分割表

図 2.8 を見てください。左上が現世代の染色体のプールです。1 本の変異染色体（濃い灰色）と、11 本の非変異染色体（薄い灰色）の計 12 本の染色体があります。最大の子の数が 4 人として、それぞれ 4 本、44 本の計、48 本の染色体プールを作り、そこから 12 本の染色体を抜き取るとします。ある抜き取り操作により変異染色体が 2 本、非変異染色体が 10 本になったことを右下の状態で示しています。プールからの抜き取りは、 $2 \times 2$  の分割表で考えることができます。全部で 48 本の染色体は 4 本が変異染色体で、44 本が非変異染色体です。それが列の和として表されています。12 本を抜き取るので、抜き取られる 12 本と抜き取られずに残る 36 本が行の和として表されています。

今、抜き取った 12 本の内訳が変異染色体 2 本、非変異染色体 10 本ですから、図に示したような分割表ができます。このような表が観察される確率（生起確率）は  $\frac{4!44!12!36!}{48!2!10!2!34!}$  で計算されます<sup>※4</sup>。これが、変異染色体が 1 本の状態から変異染色体が 2 本の状態へ遷移する確率です。

※4 このような分割表の生起確率の計算は 15.1 節を参照。



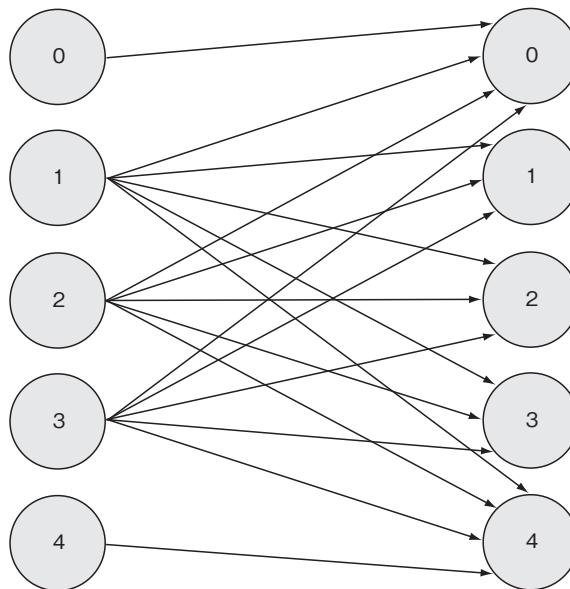
左上にある世代の染色体のプールを表し、染色体総本数が 12 本（6 人）。中央のプールは、左上の染色体がすべて 4 倍になった状態を表し、右下のプールは、中央のプールからランダムに 12 本の染色体を抜き取ってできた新世代の染色体プールを表しています。薄い灰色の染色体と濃い灰色の染色体の本数を区別して記載しています。中央の図では、下段の (4, 44) は左上のプールの (1, 11) を 4 倍した数値を表し、右端の列は、総本数 48 本のうち、新世代に伝えられる 12 本と伝えられない 36 本とを分けて示したものです。上段の (2, 10) は新世代に伝えられた染色体の色別内訳数を示しています。

図 2.8 染色体の伝達モデル

## (2) 状態推移

状態遷移という考え方をしてみます。たった 2 人（染色体 4 本）の小集団で考えてみます。変異染色体の本数としてありえる場合は 0, 1, 2, 3, 4 の 5 つです。図 2.9 を見てください。0, 1, 2, 3, 4 はそれぞれ、変異染色体の本数がその数である状態を表します。左側から右側へ世代が 1 つ進むとします。左側・右側それぞれの状態には確率がありますが、その確率は、世代ごとにすべての状態について足し合わせると 1 になります。

左側の 0, 1, 2, 3, 4 の 5 つの状態から右側の 0, 1, 2, 3, 4 の 5 つの状態へと、 $\rightarrow$ が状態の推移を表します。0 または 4 の状態は変異染色体がないか、すべてが変異染色体であるかであり、この場合には、変異染色体の本数が 0, 4 から変



左側のある状態から右側のある状態へ推移が可能であれば $\rightarrow$ があり、不可能なときは $\rightarrow$ はない。

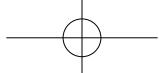
図 2.9 左側の 5 状態から右側の 5 状態へと推移する

化しようがないので、 $\rightarrow$ は 1 本しか出でていません。1, 2, 3 の状態からは、0, 1, 2, 3, 4 のすべての状態に $\rightarrow$ が出でています。すべての $\rightarrow$ が同じ太さで描かれていますが、2 の状態からは 2 の状態に遷移する確率が他の状態に遷移する確率よりも高い、というように、 $\rightarrow$ には確率の重みがあります。左側のそれぞれの状態からは 1 本以上何本かの $\rightarrow$ が出ます。左側の状態の 1 つから出た $\rightarrow$ すべてについて、その $\rightarrow$ が表す遷移の確率を足し合わせると 1 になります。

一般に、 $2N$  本の染色体のうち、変異のある染色体の本数は  $i = 0, 1, 2, \dots, 2N$  のいずれかです。 $2N + 1$  個の場合（状態）があり、それ以外はありません。変異が起きたときから始めて、一斉に新たな  $2N$  本の染色体のセットに切り替わることとし、その切り替わりの単位を世代と呼ぶことにします。第  $i$  世代で変異染色体本数が  $j$  本である確率を  $p_{i,j}$  とし、 $2N + 1$  個の状態を

$$P_i = \{p_{i,0}, p_{i,1}, \dots, p_{i,2N}\}; \sum_{j=0}^{2N} p_{i,j} = 1$$

と表すこととします。集団において、同じ座位に同じ変異が起きることはあ



りえないとすると、第0世代では、変異染色体の本数は1本ですから、 $p_{0,1}=1$ で、それ以外の状態の確率は0です。

集団のサイズを固定し、初代の変異染色体本数を指定してシミュレーションを開始します。すべての染色体は平等に最大  $k$  本の染色体を次世代に残せるものとします。したがって、ある世代から次の世代への変化においては、変異染色体と非変異染色体の本数を  $k$  倍した後、そのプールから次世代集団サイズだけ染色体をランダムに抜き取ることとして、 $P_i$  の変化を追跡します。

初期本数の確率が1の状態から徐々に初期本数を中心に増減した本数の状態の確率が増え、変異染色体の本数が0の状態の確率が上昇してきます。これは、一度、変異染色体が集団からなくなると、決して、その変異が集団に復活してこないことを意味しています。逆に、変異染色体のみが集団を占める形で、多型性が失われることもあります。時間が経つに連れ、多型性のない状態（変異染色体がない状態もしくは変異染色体のみの状態）の確率が高くなります。また、多型性のある状態は、変異染色体の本数によらずほぼ同じ確率で観察されるようになります。

関数に異なる変数を与えて、その様子を確認してください（R2-8.R 図 2.10）。

#### R ソース 2.8 R2-8.R : 遺伝的浮動シミュレーション

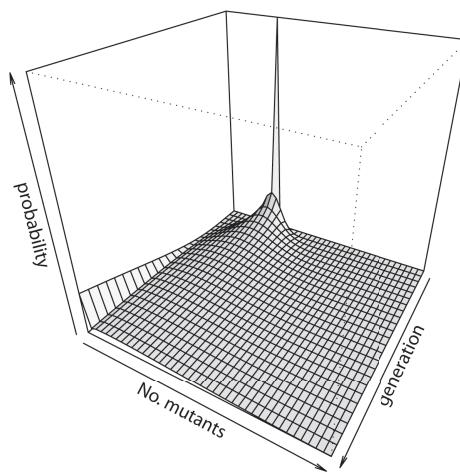
```
1  # 集団サイズ・変異アレル本数などを与え、  
# 何世代後に変異アレル本数が何本であるかの確率を計算する  
2  # N 染色体集団サイズ,p 変異アレル本数,k 個々の染色体が次世代に残す  
# 染色体最大本数,infty 制限なく残せるなら infty=TRUE  
3  probDrift02NInf<-function(N,p,k,infty=FALSE){  
4    if(infty){ # 染色体を抜き取るプールのサイズを無限大にする  
5      ret<-rep(0,(N+1)) # 変異染色体本数 0,1,2,...,N の状態  
6      pr1<-p/N # 変異染色体頻度  
7      pr2<-1-pr1 # 非変異染色体頻度  
8      if(pr1==0){ # 変異染色体がなければ、次世代も変異染色体はない  
9        ret[N+1]<-1  
10     }else if(pr2==0){ # 非変異染色体がなければ、次世代は変異染色体ばかり  
11       ret[1]<-1  
12     }else{  
13       for(i in 0:N){ # N本中 i 本が変異染色体である確率  
14         ret[i+1]<-exp(lgamma(N+1)-lgamma(i+1)-lgamma(N-i+1)+  
15           i*log(pr1)+(N-i)*log(pr2))  
16     }  
17   }
```

```
16    }
17    ret
18 }else{ # 次世代染色体を抜き出すプールは有限本数とする
19   ret<-rep(0, (N+1))
20   kN<-k*N
21   kp<-k*p
22   kNp<-kN-kp
23   k1N<-(k-1)*N
24   commonLN<-lgamma(kp+1)+lgamma(kNp+1)+lgamma(k1N+1) +
      lgamma(N+1)-lgamma(kN+1)
25   for(i in 0:N){
26     if(kp>=i){ # プールにある変異染色体本数より多い本数を抜き出すことはできない
27       n11<-i
28       n12<-kp-n11
29       n21<-N-n11
30       n22<-k1N-n12
31       if(n11>=0 && n12>=0 && n21>=0 && n22>=0) {
32         ret[i+1]<-exp(commonLN-lgamma(n11+1)-lgamma(n12+1) -
            lgamma(n21+1)-lgamma(n22+1))
33       }
34     }
35   }
36 }
37 ret
38 }
39 }
40 }
41 # ある状態において、次世代の変異染色体本数別の確率を計算する
42 nextGenExProb2<-function(x,k,infty){
43   N<-length(x)-1 # 染色体集団サイズ
44   ret<-rep(0,length(x))
45   for(i in 1:length(x)){ # 状態 0,1,...,N ごとに、次世代の状態への
        移り変わる確率を計算する
46     p<-i-1
47     tmpret<-rep(0,length(x))
48     tmpret<-probDrift02NIInf(N,p,k,infty=infty)
49     ret<-ret+tmpret*x[i]
      # 今の世代における変異本数が i の確率が x[i] なので、
      # その比率に応じて、次世代の状態 0,1,2,...,N になる
      # 確率を加算
50   }
51 return(ret)
```

```

52 }
53
54 DriftSim3<-function(k=2,initial=1,np=20,ng=100,infty=FALSE)
{ # k: 最大子孫染色体本数, initial: 初期変異染色体本数, np: 集団サイズ,
  ng: シミュレーション世代数, infty: 無限大ブールにするかどうか
55 m<-matrix(rep(0,(np*2+1)*ng),nrow=ng)
      # 全世代の全状態の確率を納める
56 m[1,initial+1]<-1 # 第1世代は、初期本数の状態の確率が1
57 for(i in 2:ng){ # 第2世代以降を順次シミュレート
58   m[i,]<-nextGenExProb2(m[i-1,],k=k,infty=infty)
59 }
60 }
61 return(m)
62 }
63 # 人数 20人、染色体数 40本、最大子染色体数 2、初期変異染色体数 10本、世代数 25
64 out<-DriftSim3(k=2,np=20,initial=10,ng=25,infty=FALSE)
      # 実行
65 # 結果を鳥瞰図表示
66 # theta,phi は鳥瞰の視点を決める変数、shade は陰影をつけて見やすくするための変数
67 persp(out,theta=120,phi=30,shade=0.2,xlab="generation",
      ylab="No. mutants",zlab="probability")

```



底面のうち、横幅の広い軸（No.mutants）が変異染色体本数です。縦に広い軸（generation）が奥から手前に向かって、世代が進んでいます。高さは、世代ごと・変異染色体本数ごとの確率です。奥のピークが世代 0 における状態（確率が 1）。世代が進むにつれ、変異染色体が 0 である確率が高くなることが、左端の高さの増加から読み取れます。

図 2.10 遺伝的浮動シミュレーション

### (3) 推移行列

R を使って、状態の変化を描きましたが、この状態の推移は行列（**推移行列**）を使って表すこともできます。行が、推移前の変異染色体本数を表し、列が推移後の変異染色体本数を表します。 $s_i$  本の推移前状態のときに  $s_j$  本の推移後状態に変わる確率が  $(i, j)$  セルで  $pr_{i,j} = s_i, j = s_j$  と表されています。第  $i$  行の確率は、推移可能な推移後状態全部について足し合わせると 1 になります。

$$\sum_{j=1}^{2N} pr_{i=s_i, j=s_j} = 1$$

before  after	$s_0$	...	$s_j$	...	$s_{2N}$
$s_0$	$pr_{i=0, j=0} = 1$	...	$pr_{i=0, j=j} = 0$	...	$pr_{i=0, j=2N} = 0$
...	...	...	...	...	...
$s_i$	$pr_{i=i, j=0}$	...	$pr_{i=i, j=j}$	...	$pr_{i=i, j=2N}$
...	...	...	...	...	...
$s_{2N}$	$pr_{i=2N, j=0} = 0$	...	$pr_{i=2N, j=j} = 0$	...	$pr_{i=2N, j=2N} = 1$

ただし、

$$pr_{i=i, j=j} = \frac{(2N)!(2N(k-1))!(ki)!(k(2N-i)!}{(2kN)! j!(2N-j)!(ki-j)!(2N(k-1)-(ki-j))!}$$

であり、これは、

	変異本数	非変異本数	計
次世代	$j$	$2N-j$	$2N$
非伝達分	$ki-j$	$2N(k-1)-(ki-j)$	$2N(k-1)$
$k$ 倍プール全体	$ki$	$k(2N-i)$	$2kN$

という分割表の生起確率です。

ここでは、変異アレルの頻度が集団内で変化する様子を扱いました。世代の変化による頻度の推移の計算にあたっては、次世代の確率が現世代の確率のみを用いて計算できました。このように次の状態の確率を現世代の確率だけを用いて計算する処理をマルコフ連鎖と言います。マルコフ連鎖は第 8 章の連鎖解析のところでもう一度扱いますが、その特徴は、処理が逐次処理であること

と、組み合わせを使わぬことです。この2つの特徴は、手続処理を簡単にしてくれます。



## 2.2 DNA から RNA、タンパク質へ

### ◎ 2.2.1 DNA から RNA へ——転写

DNAに記録された遺伝情報はRNAに写し取られて機能を発揮します。図2.11のように、DNAの上側の情報部分"CAGGTT"は、"CAGGUU"のRNAとして写しとられます。これを**転写**と言います。DNAは"A", "C", "G", "T"を使いますが、RNAは"T"の代わりに"U"を使います。RNAはさらにタンパク質に翻訳されることも多く、その場合のRNAはメッセンジャーRNA(mRNA)と呼ばれます。翻訳されずにRNAとして機能を発揮することもあり、そのときには、機能性RNA遺伝子と呼ばれます。

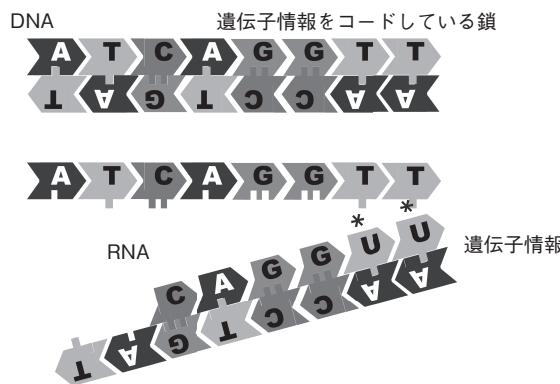


図 2.11 転写：DNA から RNA が転写される

### ◎ 2.2.2 RNA からタンパク質へ——翻訳

mRNAの情報を読み取ってアミノ酸の連なり（タンパク質）を作ることを**翻訳**と言います。翻訳に用いられるアミノ酸は20種類あります。塩基4種類の順列で20種類のアミノ酸を識別するための最短の塩基列の長さは $4^2 \leq 20 \leq 4^3$ ですが、実際、塩基3個の順列のそれぞれがアミノ酸に対応して翻訳さ

れます。3 塩基の並びをコドンと言い、コドンとアミノ酸の対応関係を図 2.12 に示しました。タンパク質への翻訳にあたっては、翻訳処理の開始を意味する特別なコドンと、翻訳の終了を意味する特別なコドンとを決めることで、翻訳処理の開始と終了が制御されます。開始コドンである "AUG" は特定のアミノ酸である "Met" に対応しているので、タンパク質が作られるときの最初のアミノ酸はいつも "Met" です。終了のコドンは複数ありますが、それにはなんのアミノ酸も対応させないことで、翻訳作業が終了します。

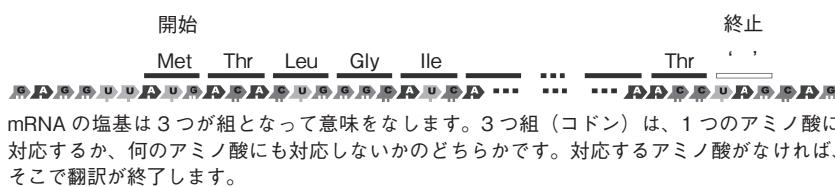
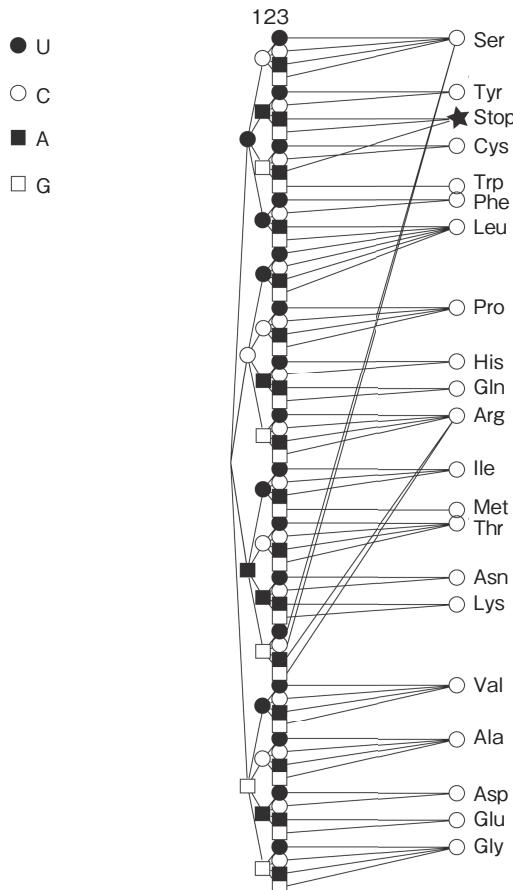
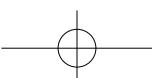
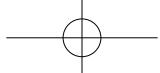


図 2.12 コドンとアミノ酸の対応関係



4つの塩基（U,C,A,G）をマーク（●、○、■、□）で表しています。分岐木の一番上はUCUに相当し、それに対応しているアミノ酸はSerであることがわかります。 $4^3=64$ 通りの組み合わせがあり、それが、20種類のアミノ酸と「対応アミノ酸なし（終止コドン）」の21通りのどれかに相当します。

図 2.13 mRNA の塩基の 3 つ組を分岐木で表示



# 第3章

## 多様性の諸相

### 3.1 核酸、タンパク質の多様性

#### ● 3.1.1 DNA配列の多様性、種の違い、遺伝子多型

DNA配列の多様性は変異によって生まれ、それが集団中で増えるか減るか、なくなるかは遺伝的浮動という現象で決まります。また、複数の変異が同じ染色体上に生まれた後に、組み換えによって異なる染色体上に分かれたり、逆に、異なる染色体上に生まれた変異が組み換えによって同じ染色体上に乗ることによって変異の存在状態が変わります。種の遺伝的な違いはDNA配列の違いです。同じ種の個体の集団で、配列が同一でないときに遺伝子多型（以下、多型）であると言います。

多型はDNA配列の違方ににより、置換・挿入欠失・リピート（繰り返し）・逆位・転座などに分類され、それぞれの種類に短いものから、長いものまであります。遺伝子座もそのサイズが1塩基だったり、ある長さを持つ範囲だったり、とびとびの範囲だったりしましたが、多型でも同様にDNA配列の任意の位置・範囲について定めることができます。

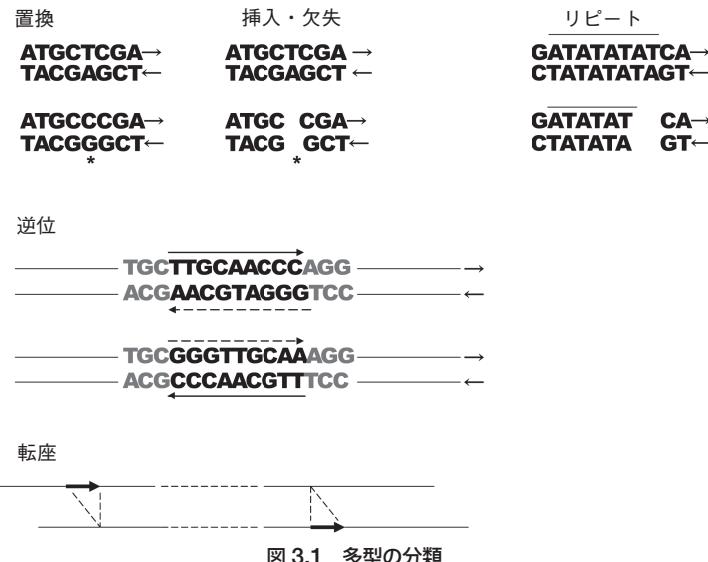


図 3.1 多型の分類

図 3.1 を見てください。転座以外は 1 対の DNA 二重鎖配列が示してあり、それぞれの鎖の右端の→はその方向です。置換は 1 塩基対の例です。\* の部分が上の二重鎖では T と A の対であるのに対し、下の二重鎖では C と G の対になっています。挿入・欠失の例も 1 塩基です。上の二重鎖に較べると下の二重鎖は欠失で、下の二重鎖に較べると上の二重鎖は挿入です。リピートの例は、"AT" が 4 回繰り返されている上の二重鎖に較べて、下の二重鎖は 3 回の繰り返します。逆位では、灰色の塩基に挟まれた黒い塩基対部分が逆位になっています。上の二重鎖の上側の配列 "TTGCAACCC" が、下の二重鎖の下側の配列になっています。転座は→で示された配列が上と下とでは異なる位置に入っていることを示しています。

図 3.2 には遺伝子多型の大きさが示されています。スケールが最上段に対数スケールで示されています。多型は顕微鏡で観察可能なもの (microscopic variants) と観察できないもの (sub-microscopic variants) に大別され、特に大きめながら視覚観察できないものを構造多型 (structural variants) と呼びます。多型のタイプのそれぞれに、さまざまなサイズの多型があります。SNP は 1 塩基の置換型多型、コピーナンバー多型 (CNV) はリピート型の構造多型です。X 染色体と Y 染色体の違いを多型とみなせば、染色体サイズ

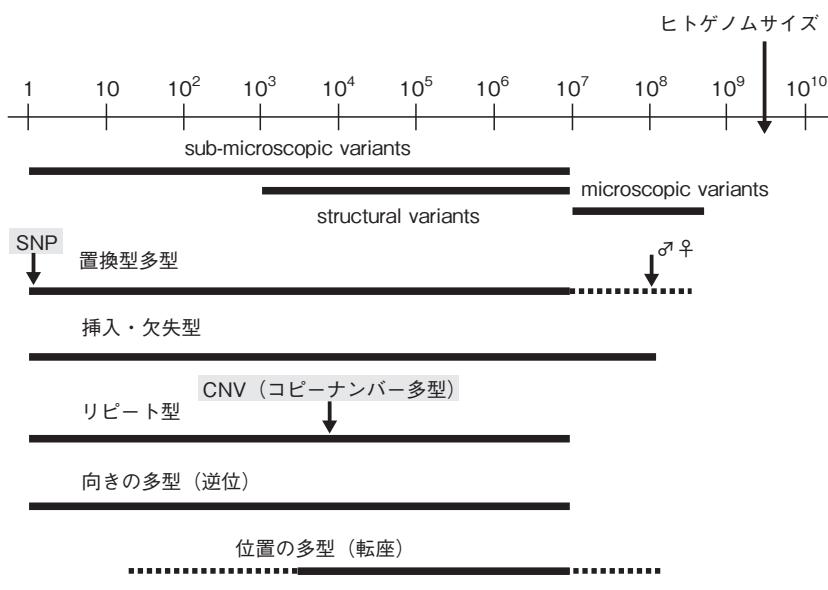


図 3.2 多型のいろいろ

の挿入・欠失とも言えますし、XとYとをそっくり入れ替えているとみなせば置換とも言えるでしょう。

多型を遺伝子座のそれとすれば、多型の異なる配列の1つ1つはアレルです。1塩基多型は4種類の塩基をとることが可能ですが、多くの場合、2種類の塩基のみで存在します。このような場合、2アレル型の多型であると言います。個々の多型は塩基配列上に並んでいますので、ある長さを持つ範囲に複数の小さな多型が存在することができます。この複数の小多型を含む範囲を1つの単位として取り扱うときには、そのアレルをハプロタイプと呼びます。図3.3はある領域に3つの2アレル型の1塩基多型があり、その組み合わせとしてのハプロタイプは $2^3=8$ 種類あることを示しています。2アレル型の場合には"A" "T" "G" "C" の4文字を使わずに"0" "1"で表すと便利なことが多いです。

**ATGCTCGACATGGTCCCATGTTA→  
ATGCACGAGATGGTCCCATTTA→**

TCG			
AGA			
000	001	010	100
111	011	101	110
		111	

2つの配列は3箇所の1塩基多型で配列が異なります。多型箇所の塩基は黒く、その他の塩基は灰色で示してあります。多型箇所だけを抜き出すと "TCG", "AGA" という配列になります。この違う部分だけを用いてハプロタイプとしてもよいです。2アレル型なので、第1, 2, 3の1塩基多型の "T" "C" "G" を "0"、"A" "G" "A" を "1" で表すことにすると、この2ハプロタイプは "000", "111" と表せます。1塩基多型のアレルの組み合わせが8通りあるので、8種類のハプロタイプがありえることを、右下の "0" "1" の配列が示しています。

図 3.3 2つのハプロタイプ

### ● 3.1.2 RNA とタンパク質の多様性

RNA の配列は DNA からそっくり転写されてできますから、塩基配列の多様性は DNA 配列のそれと同じです。ある DNA 配列から RNA が転写されてできるときに、鋳型としての DNA は変わらなくても、DNA 配列のどの部分が転写されるかが変わることによって、多様な RNA 配列が得られます。ある DNA 配列上にあるいくつかのエクソンが mRNA に転写され、さらにタンパク質に翻訳されますが、このエクソン領域のとり方が、場合によって変わってくることがあります。あるエクソンがそっくり抜け落ちたり、あるエクソンの長さが伸びたりします。DNA から mRNA を転写する際に、エクソンだけを切り出す過程を **スプライシング** と言うので、このように異なる mRNA 配列のそれぞれを **スプライシングバリエント** と呼びます。この結果、出来上がるタンパク質のアミノ酸配列も当然、違ってきます。

R を使って、スプライシングバリエントの配列を作ってみます (R3-1.R)。

#### R ソース 3.1 R3-1.R : スプライシングバリエントの配列

```

1 seq1<-sample(c( "A" , " T" , " G" , " C" ),100,replace=TRUE)
      # 長さ 100 の DNA 配列をランダムに作る
2 exonpattern1<-c(11:20,41:60,81:90)
      # mRNA1 は 3 エクソン。そのパターン
3 exonpattern2<-c(11:20,41:60,66:77,81:90)
      # mRNA2 は mRNA1 の第 2 エクソンと第 3 エクソンの間にエクソンが挿入されています

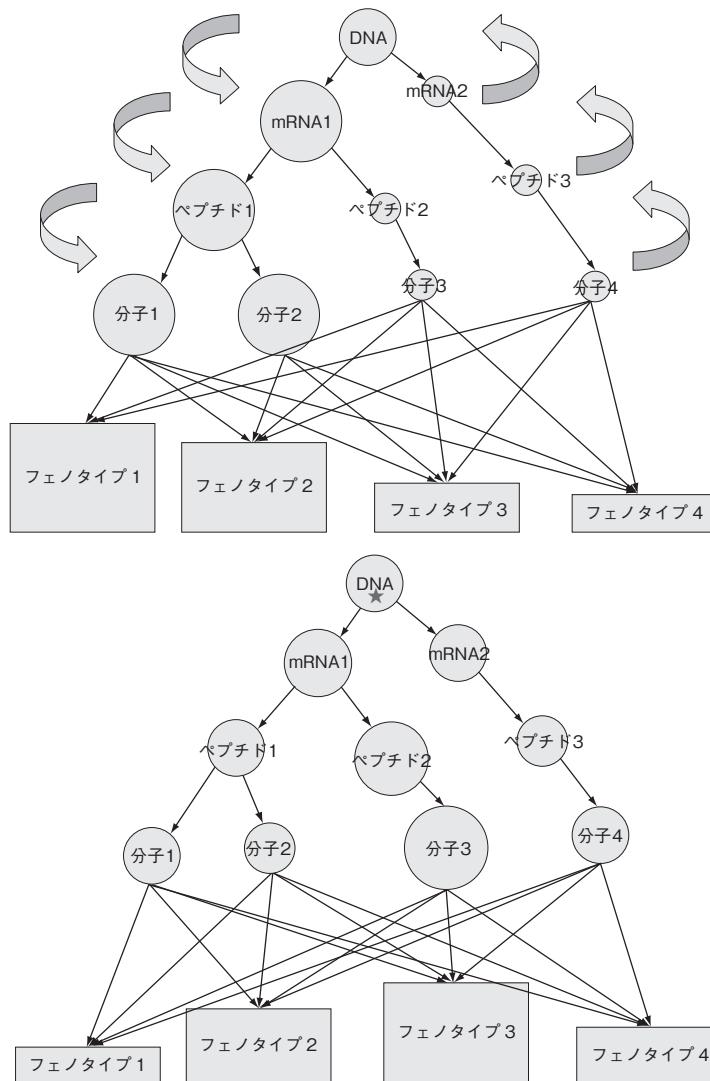
```

```
4 seq1 [seq1=="T"] <- "U" # mRNAではDNAの "T" が "U" になります。
5 mRNA1<-seq1 [exonpattern1] # mRNA1 の配列を抜き出します
6 mRNA2<-seq1 [exonpattern2]
```

mRNA では DNA で "T" の塩基が "U" になります。1つの DNA 配列からエクソンの位置が定まることによって、1つの mRNA が作られました。実際には、異なるエクソンの位置を用いることによって、1つの DNA 配列から複数の mRNA が転写されます（スプライシングバリアント）。多くの場合は、余分のエクソンが転写されたり、一部のエクソンが長かったり、短かったりという違います。

このように RNA は DNA のどの部分を取り出すか、という「位置的」な変化をつけることで多様性を持っています。さらに、DNA 二重鎖は各細胞に 1 セットずつあり、それより多くも少なくもなく（量が一定）、しかも、その細胞が新たに生まれてから死ぬまで変わらない（時間的に一定）のですが、RNA の方は、どのくらいたくさん転写をするかによって、量の増減があり、しかも、その量は「時間的」に変わります。RNA の多様性は DNA の多様性よりも格段に大きいことがわかります。実際には、DNA の方も量と塩基配列こそ変わらないものの、その分子状態は化学修飾を受けることによって、状態が「位置」によって多様ですし、その状態が「時間的」にも変化しています。

タンパク質は mRNA が翻訳されることによってできます。DNA から mRNA に引き継がれた配列の多様性が、コドン表でのアミノ酸への対応関係を変えるようなものであれば、アミノ酸配列を変えます。スプライシングバリアントはタンパク質のアミノ酸配列を長さも含めて変化させます。それ以外の多様性は、アミノ酸配列以外での多様性に影響します。タンパク質への翻訳も「いつ」「どこで」「どれくらい」という点で多様度を持ち、それは、生物が生きていく上で微妙かつ精密に調節されているので、遺伝子産物としてのタンパク質のアミノ酸配列レベルでの多様性、その量・時間的変化の多様性がより大きくなります。生物個体の特徴である形質は、このタンパク質の多様性を含めた、多くの構成分子の多様性の結果として現れてくるものなので、DNA 配列にその根を持つつつも、大変大きな多様性を持っていることが理解できます（図 3.4）。



DNA から mRNA、アミノ酸配列が転写・翻訳され、機能を持つ分子となり、さまざまな形質に影響を与えます。DNA 配列の多様性（★）は mRNA、アミノ酸配列、分子を介して形質に影響を与えます。DNA 上の 1 つの遺伝子から複数の mRNA が転写され、その量（○の大きさ）に多様性が生まれます。さらにアミノ酸配列のレベルで量を含めた多様性がもたらされ、さまざまな形質の出現しやすさへ影響する様子が、形質の四角の高低で表現してあります。さらに、DNA → RNA → タンパク質 → 形質という方向だけではなく、RNA は DNA に調節的に働き、タンパク質は RNA や DNA に働きかけるという逆方向の反応もあり、DNA・RNA・タンパク質・形質の間の関係は大変複雑になります。

図 3.4 DNA から転写・翻訳され、機能を持つ分子になるまで

## 3.2 多様性と分散

### ● 3.2.1 分散の分解——分散、共分散

多様性はばらつき加減ですので、ばらつきの指標の1つである**分散**として値にすることができます。

分散は足し合わせることができます、逆に、分解することもできます。ここでは分解を見てみます。あるものが複数のものから構成されているとき、その分散は、構成要素の分散と、構成要素間の関係とに分解できます。

まず、この話を進める前に、構成要素の分散と構成要素間の関係という2種類の要素を考えていますが、これを統一して取り扱うことにします。構成要素の分散はその要素と要素自身との関係であるということにすれば、構成要素  $s$  の  $s$  に対する関係を、 $\sigma_{s,s}$  と書くことができます。他方、 $s$  自身ではない  $t$  との関係を  $\sigma_{s,t}$  とすれば、 $s = t$ ,  $s \neq t$  のいずれでも同じように取り扱うことができます。さて、ある変数  $X_s$  について、 $N$  個のサンプルに  $x_{s,i}$  があったときに、 $X_1$  分散は、

$$V(X_s) = \frac{1}{N} \sum_{i=1}^N (x_{s,i} - \mu_s)^2$$

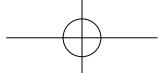
で計算できるのでした。ただし、 $\mu_s = \frac{1}{N} \sum_{i=1}^N x_{s,i}$  は  $X_s$  の**平均**です。分散の計算式の  $(x_{s,i} - \mu_s)^2 = (x_{s,i} - \mu_s) \times (x_{s,i} - \mu_s)$  に着目します。ここに  $s$  が2回登場しているので、

$$\sigma_{s,s} = V(X_s) = \frac{1}{N} \sum_{i=1}^N (x_{s,i} - \mu_s)(x_{s,i} - \mu_s)$$

としてやれば、

$$\sigma_{s,t} = \frac{1}{N} \sum_{i=1}^N (x_{s,i} - \mu_s)(x_{t,i} - \mu_t)$$

と書き換えてやることで、分散の定義の拡張として、2変数  $X_s$ ,  $X_t$  の間の関係



$\sigma_{s,t}$  が定義できます。これが**共分散**です。

今、2つの変数  $X_1, X_2$  があって、 $N$  個のサンプル  $i = 1, 2, \dots, N$  についてその値  $x_{1,i}, x_{2,i}$  があり、その和  $P = p_i = x_{1,i} + x_{2,i}$  を考えます。物理の試験の点  $X_1$  と生物の試験の点  $X_2$  の和で、理科の総合点とするようなものです。

$$\sigma_{P,P} = \sigma_{X_1,X_1} + \sigma_{X_2,X_2} + \sigma_{X_1,X_2} + \sigma_{X_2,X_1} = \sum_{s=1}^2 \sum_{t=1}^2 \sigma_{X_s,X_t}$$

となります。また、共分散は、2因子が独立のときにゼロで、関係があるときに大きくなります。計算式での証明は、統計学の本にはよく出ていますので、ここでは省略して、それよりは、「本当にそうなることの確認」をすることにしてみます。

このことを R で見てみます。 $X_1, X_2$  が乱数であるとします。 $P$  をその和とします (R3-2.R)。

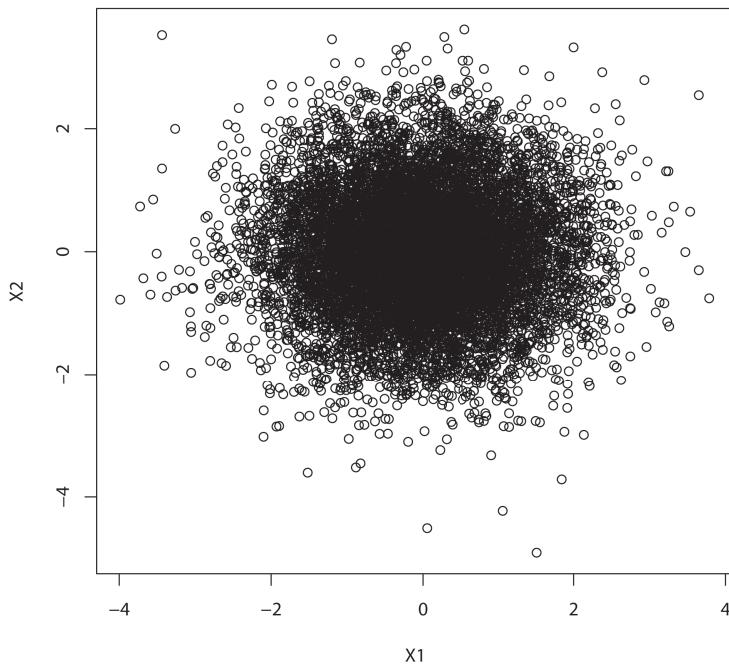
図 3.5 のように  $X_1, X_2$  が相互に独立な場合と、相互に関連のある場合とを試してみます。その上で、分散、共分散の関係を確認します。

#### R ソース 3.2 R3-2.R : 分散・共分散

```
1 nsample<-10000;X1<-rnorm(nsample);X2<-rnorm(nsample)
# rnorm() 関数については、18.4 確率分布関数、疑似乱数列の発生を参照
2 p12<-X1+X2
3 cov12<-cov(cbind(X1,X2))
# X1, X2 の分散・共分散行列を計算します
4 cov12 # 分散共分散行列を表示します 共分散は小さいです（右上、左下のセル）
5 vpp_2<-cov(p12,p12) # p12 の分散
6 vpp_2-sum(cov12) # p12 の分散が、分散・共分散の和である。
7 plot(X1,X2) # X1, X2 のばらつきを示します
8 # x1 と関係のある x2 を作ります
9 X3<-X1+0.1*rnorm(nsample)
10 p13<-X1+X3
11 cov13<-cov(cbind(X1,X3)) # X1, X2 の分散・共分散行列を計算します
12 cov13 # 分散共分散行列を表示します 共分散は大きいです（右上、左下のセル）
13 vpp_3<-cov(p13,p13) # p12 の分散
14 vpp_3-sum(cov13) # p13 の分散が、分散・共分散の和である。
15 plot(X1,X3) # プロットは強い相関を示します
```

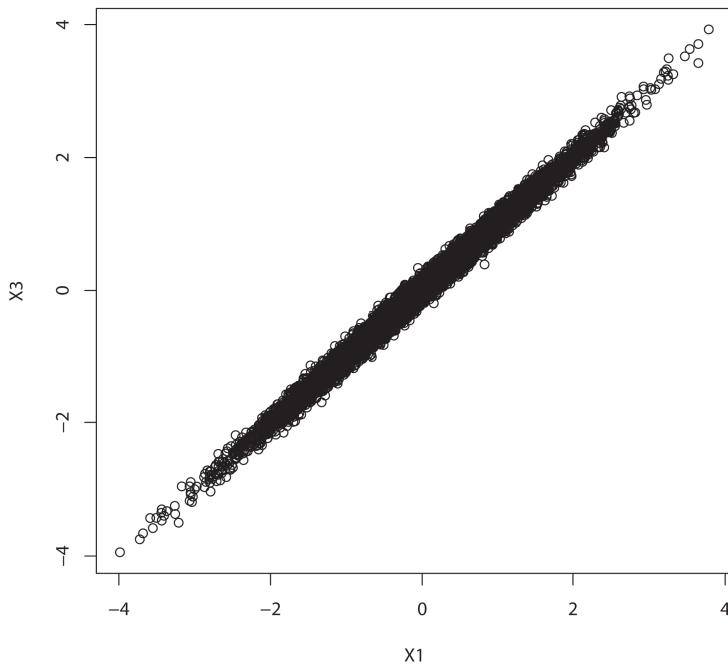
**■ R の出力結果**

```
> cov12
      X1          X2
X1  1.009561663 -0.006727835
X2 -0.006727835  1.018274923
> vpp_2 - sum(cov12)
[1] 0
> cov13
      X1          X3
X1  1.009562  1.010659
X3  1.010659  1.021820
> vpp_3 - sum(cov13)
[1] 0
```



(a) 相関のない例

図 3.5 相関のない例と強い例



(b) 相関の強い例

図 3.5 相関のない例と強い例（つづき）

2つの例では  $X_1, X_2$  の相関の強さが、一方はなしで、一方は非常に強いという違いがありますが、どちらも  $P$  の分散  $V(P) = V(X_1) + V(X_2) + 2*Cov(X_1, X_2)$  になっていることが確かめられます。 $X_1, X_2$  の相関の強さの違いは、 $V(P)$  に占める  $Cov(X_1, X_2)$  の割合が、ほとんどないか、非常に大きいかに現れています。ここでは正規分布を使いましたが、これは分布によりませんし、サンプルサイズが小さくても成り立ちます。また、ここまで、2つの因子とその因子間の関係で考えてきましたが、因子の数は2つに限らずいくつでもよいので、

$$\sigma_{P,P} = \sum_{s=1}^k \sum_{t=1}^k \sigma_{X_s, X_t}$$

となります。

### ● 3.2.2 遺伝率と分散

これが、遺伝因子に関してどのように使われているかを見てみます。ある形質 ( $P$ ) のばらつき加減が、遺伝因子とそれとは無関係の（独立な）他の因子で説明できるとしたとします。「遺伝因子 ( $G$ )」と「他の因子 ( $E$ )」の2つが登場し、

$$\sigma_{P,P} = \sigma_{G,G} + \sigma_{E,E}$$

と分解されます。 $G$  と  $E$  とは独立なので、 $\sigma_{G,E} = 0$  であって、上式では省略しています。  $\frac{\sigma_{G,G}}{\sigma_{P,P}}$  は、形質を決める遺伝因子の強さとして、**遺伝率** (heritability) と呼ばれます。

### ● 3.2.3 ハーディ・ワインバーグ平衡 (HWE) と分散

2アレル型の多型を考えます。アレルが  $A$  と  $a$ 、その頻度が  $p$  と  $(1-p)$  とします。2倍体のジェノタイプは "AA", "Aa", "aa" の3種類です。 $a$  の本数を用いて3ジェノタイプを  $G0$ ,  $G1$ ,  $G2$  と書き表すことにします。この3ジェノタイプの頻度を考えます。 $a$ ,  $A$  の選ばれ方が独立であるという仮定のもとでの期待値は

$$G0 = p^2$$

$$G1 = 2p(1-p)$$

$$G2 = (1-p)^2$$

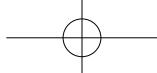
となります。この値からのずれがあるとして、それを変数で表してみます。

$$G0 = p^2 + \Delta$$

$$G1 = 2p(1-p) - 2 \times \Delta$$

$$G2 = (1-p)^2 + \Delta$$

と1変数  $\Delta$  を使って表すことができます。 $\Delta = 0$  のとき、 $a$ ,  $A$  が独立な仮定のもとでの期待頻度となり、逆に  $\Delta = p(1-p)$  のときに、ヘテロ型がいなくなります。 $\Delta = 0$  であるとき ( $A$  と  $a$  の組み合わせ方が独立であるとき) に、遺伝学ではその集団が「**ハーディ・ワインバーグ平衡**」(Hardy-Weinberg 平衡)



(HWE) にある」と言います。そして、そこからずれていることを、HW 不平衡と言います。また、そのずれを、

$$\begin{aligned}G0 &= p^2 + p(1-p)F \\G1 &= 2p(1-p)(1-F) \\G2 &= (1-p)^2 + p(1-p)F\end{aligned}$$

と表すこともあり、この  $F$  を**近交係数**と呼びます。

アレル  $A$  と  $a$  を因子のありなしとして、 $A$  に値 1 を、 $a$  を値 0 とカウントすることとします。このようなアレルがもたらす値の分散の大小について考えてみます。今、人数  $N$  の集団を考えます。この集団を染色体単位で考えると、 $2N$  本の集団とします。

その平均と分散  $\mu(h)$ ,  $v(h)$  は、

$$\begin{aligned}\mu(h) &= (p \times 1 + (1-p) \times 0) = p \\v(h) &= p \times (1-p)^2 + (1-p) \times (0-p)^2 = p(1-p)\end{aligned}$$

となります。

さて、2倍体ではどうでしょうか。2倍体なので、ジェノタイプは " $AA$ ", " $Aa$ ", " $aa$ " の3種類あります。3ジェノタイプの値はそれぞれ 2, 1, 0 となります。さて、3つのジェノタイプの頻度が定まれば、この集団における、この影響力の分散が計算できます。2倍体での平均 ( $\mu(d)$ ) と分散 ( $v(d)$ ) とします。

$$\begin{aligned}\mu(d) &= G0 \times 0 + G1 \times 1 + G2 \times 2 = 2p = 2\mu(h) \\v(d) &= G0 \times (0 - \mu(d))^2 + G1 \times (1 - \mu(d))^2 + G2 \times (2 - \mu(d))^2 \\&= 2p(1-p)(1+F) = 2v(h) + 2 \times F \times v(h)\end{aligned}$$

2倍体の分散は、1倍体の分散の和 ( $2v(h)$  と  $2 \times F \times v(h)$  の和) となっています。この  $F$  に依存する項が共分散成分になっており、ここから共分散は  $Cov(d) = F \times v(h)$  であることが見てとれます。実際、

$$F = \frac{Cov(d)}{v(h)} = \frac{Cov(d)}{p(1-p)}$$



ですが、これは相関係数です。 $F$ と相関係数は、共分散に比例する値で、ヘテロ型がいないう状態のときに1をとるような指標であることがわかります\*1。

### ● 3.2.4 アレル関連、連鎖不均衡と分散

今、2つの2アレル型多型が同じ染色体の比較的近いところにあるとします。多型を  $Ma, Mb$  とします。それらのアレルを  $A/a, B/b$  とし、そのアレル頻度を  $p, (1-p), q, (1-q)$  とします。 $Ma, Mb$  とともに、 $A, B$  が値1を持ち、 $a, b$  は値0であるとします。前項で見たように、 $Ma, Mb$  の値の平均と分散はそれぞれ、

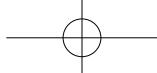
$$\begin{aligned}\mu(Ma) &= p, \mu(Mb) = q \\ v(Ma) &= p(1-p), v(Mb) = q(1-q)\end{aligned}$$

です。2多型の組み合わせとして4ハプロタイプ  $AB, Ab, aB, ab$  ができます。そのハプロタイプ頻度を  $h1, h2, h3, h4$  とします。先ほど、HWEのときに、独立の仮定での頻度を算出し、そこからの逸脱で、ジェノタイプの頻度を表すことにしました。今回もそれと同じ手順を踏むことにします。2多型が独立であると仮定すると、

$$\begin{aligned}h1 &= pq \\ h2 &= p(1-q) \\ h3 &= (1-p)q \\ h4 &= (1-p)(1-q)\end{aligned}$$

異なる座位のアレルの組み合わせの頻度が、独立を仮定した頻度となっているとき、その座位はお互いにアレルが独立である・連鎖平衡にあると言います。逆に、独立ではないときにアレルに関連がある・連鎖不均衡にあると言います。連鎖不均衡のときには、ハプロタイプ頻度がこの値からずれるので、それの項を持ち込んで表せば、次のようにになります。今、 $p=h1+h2, q=h1+h3$  なので、

\*1 指数とは何かについては第11章でまとめます。



$$h1 = pq + \delta$$

$$h2 = p(1-q) - \delta$$

$$h3 = (1-p)q - \delta$$

$$h4 = (1-p)(1-q) + \delta$$

と表せます。

ここで、

$$\delta = r\sqrt{p(1-p)q(1-q)}$$

とおくと、 $h1 = p = q$ ,  $h2 = h3 = 0$ ,  $h4 = (1-p) = (1-q)$  のときに  $r = 1$  となります。これは、アレル  $A$  と  $B$  とが強く関係し、 $a$  と  $b$  とが強く関係し、4つの可能なハプロタイプのうち、2つだけが存在するという特別な状態に対応しています。先ほどの HWE の例で、 $F = 1$  がホモ型だけの状態を表していたのに対応します。

この  $r$  は 2 つの 2 アレル型多型のアレル同士が関係する程度の指数（アレル関連指数、連鎖不平衡指数）と呼ばれるものの 1 つです。逆に言うと、多型のアレル同士が同じ染色体の上に乗るかどうかは独立なこともあるけれども、特定のアレルの組み合わせが高頻度で同じ染色体上にあるという場合が存在し、それを遺伝学では、アレル関連と呼んだり、連鎖不平衡と呼んだりする、というように言い換えることもできます。

さて、

$$h1 = pq + \delta$$

$$h2 = p(1-q) - \delta$$

$$h3 = (1-p)q - \delta$$

$$h4 = (1-p)(1-q) + \delta$$

と表すこととし、この 4 ハプロタイプが独立して 2 倍体のジェノタイプを構成しているとします。 $A$ ,  $B$  ともに得点 1 を持つものとし、 $AB$ ,  $Ab$   $aB$ ,  $ab$  の得点はそれぞれ 2, 1, 1, 0 であるとします。このときの、染色体の集団の得点の平均と分散は、

$$\begin{aligned}\mu(Ma + Mb) &= p + q \\ v(Ma + Mb) &= p(1-p) + q(1-q) + 2\delta \\ &= v(Ma) + v(Mb) + 2\delta\end{aligned}$$

これから、2座位の得点の和の分散は、個々の座位の分散の和と、それ以外に分解されました。 $\delta$ は2多型のアレル間の共分散に関する項なわけです。実は、連鎖不平衡の指数である  $r$  ( $r^2$ が頻用される) は、

$$r = \frac{\delta}{\sqrt{p(1-p)q(1-q)}}$$

と表され、これは相関係数です。これは共分散と比例し、存在しうる4ハプロタイプのうち、2ハプロタイプしか存在しない状態で1となるように調整された指数であることがわかります。

## 3.3 データの取り扱い方と分散、共分散

### ● 3.3.1 HWE とアレル関連、連鎖平衡を2列のデータで考える

HWEとアレル関連・連鎖平衡について、それぞれの指標  $F, r$  と共分散の関係を見てきました。両者を2カラムのデータとして考えてみます。図3.6を見てください。

まず、このデータをHWEのデータとして考えます。左の2列は、個人が持つ2つのアレルを表しています。右のカラムは2倍体としての点数となっています。2倍体の点数の分散は、2つのアレルの分散と、2つのアレルの間の共分散の和であって、それは  $F$  と比例します。 $11, (10, 01), 00$  は3ジェノタイプに分かれますので、それに相当する頻度が列の左側に示してあります。

次にこのデータを2多型のハプロタイプのものとして考えます。左の2列は、それぞれ  $Ma, Mb$  のアレルを表しています。4つのハプロタイプが、2列にまたがった数値ペア  $11, 10, 01, 00$  で表されます。右の列はハプロタイプご

最左はアレル頻度  $p$  と近交係数  $F$  で表したもの。最右は 2 多型のアレル頻度と連鎖不平衡指数  $r$  で表したもの。 $V1, V2, Cov12$  は因子 1, 2 の分散と共に共分散。

図 3.6 HWE と連鎖不平衡のデータを 2 カラムデータとして表す

との得点に相当します。このとき、ハプロタイプの点数の分散が、個々の多型の分散の和と、多型のアレル間の共分散の和であることが示されています。この場合の共分散は、アレル間の関連指標  $r$  に比例しています。4ハプロタイプの頻度が右側に示してあります。

HWE として扱う場合と連鎖平衡として扱う場合との違いは、 $(1, 0), (0, 1)$  の行数が HWE のデータの場合には常に等しいのに対して、アレル関連・連鎖平衡のデータの場合には等しくなくてもよいという点です。

### ● 3.3.2 遺伝形式（優性、劣性）は第3の列

HWE、アレル関連・連鎖平衡の場合には、独立しているかもしれないもの（2倍体の個々のアレル、2つの多型）を図3.6のように2列で扱い、それぞれの分散と2列間の共分散とに分解しました。では、優性・劣性・相加的といった遺伝形式は、それと同じように扱うことができるでしょうか。あるアレルを2本持つジェノタイプの影響度を2としたときに、相加的形式では、アレル1本のジェノタイプの影響度は1です。これは、HWE・連鎖平衡を考えたときと同じ枠組みです。したがって、2列で扱うことは適當です。

では、優性形式の場合はどう考えればよいのでしょうか。ヘテロ型の場合だけ、影響度を1だけ増やすことが必要です。純粹に優性の場合はこの効果が1、純粹に劣性の場合には-1、それ以外の場合も、この値を適当に定めることで表現することができます。相加的形式は、この値が0である特別な場合とも言えます。そのために図3.7のように、この効果専用のカラム  $Vd$  を付け加えることにします。

2列ではうまく表せませんでしたが、3列目を加えることで表すことができます。遺伝形式を考えるときには、ヘテロ型に特有の効果の要素と、それと2つのアレルとの関係の要素の3要素を付け加えて考えなくてはいけないことがわかります。

新たに加えた第3列の分散  $Vd$  と第3列と第1列、第2列との間の共分散  $Cov1d$ ,  $Cov2d$  は次のように表されます（細かい式の導出は省略しました）。

$$Vd = G1(1 - G1)d^2$$

$$Cov1d = Cov2d = \frac{1}{2}G1(G2 - G0)d$$

$G0$ ,  $G1$ ,  $G2$  は、点数が0, 1, 2の人数です。このように、ジェノタイプの頻度の関数で表され、 $d=0$ のとき（相加的形式）に両方とも0となり、 $G1=0$ （ヘテロ型）が存在しないときにもこの効果はなく、また、 $G0=G2$ （2つのホモ型の頻度が同じとき=アレル頻度が0.5のとき）には共分散の項が0になります。このことなどがわかります。 $d=\frac{G0-G2}{G0+G2}$ のときには、 $Vd + Cov1d + Cov2d = 0$ となって、遺伝形式の寄与が全体として相殺されることなどがわかりま

$Vd$ は第3のカラムの分散です。第3のカラムと第1、第2のカラムとの間の共分散  $Cov1d$ ,  $Cov2d$ が生じています。全体の分散は、分散と共分散の和で表せます。

図 3.7 図 3.6 に第 3 のカラムを付け加えた図

す。逆に言えば、アレル頻度が 0.5 でないときには、遺伝形式が相加的以外の場合には、形式の寄与の列がアレルの列と独立であることはありえないことを意味しています。遺伝形式の効果は、アレルの持ち方に依存して決まる要素であることを思い出せば、当然のことと言えます。

## 3.4 たくさんの要因——多因子遺伝

形質の多くは、複数のそして多数の遺伝子と環境因子の影響を受けて決まっていると考えられています。そのような形質を**多因子遺伝形質**とか**複合遺伝形質**と言います。

たとえば、ある形質に 100 個の 2 アレル型多型が関係しているとします。これらの多型は相互に連鎖平衡にあるとし、また HWE を仮定した集団を考えます。このような集団で、100 個の多型がさまざまな強さでさまざまな遺伝形式で寄与しているとします。

このような場合には、たくさんの要素が影響するために、遺伝因子全体の形質への影響は、個人ごとにばらついて、正規分布様の分布になることが示せます。そしてその遺伝因子全体の寄与の分散は、寄与因子の分散・共分散の和であることも確かめられます。多型間に連鎖不平衡だったり、HWE からずれている要素を加味するためには、それらのために変数を追加してやるだけですから、ここでは、単純な設定としました。

R のソース (R3-3.R) を確認してください。

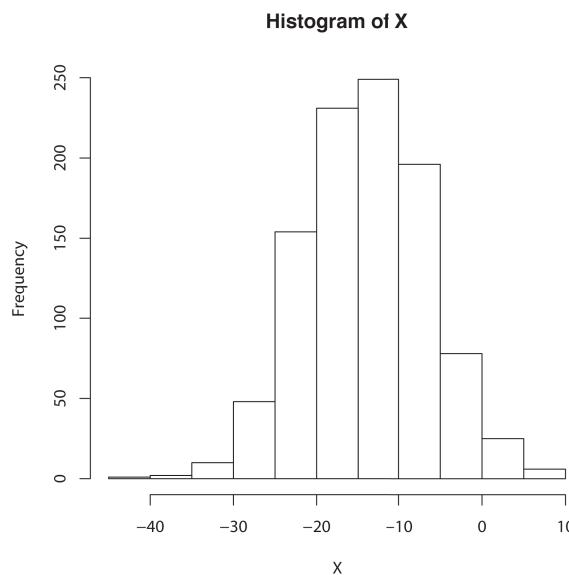
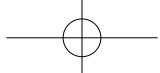


図 3.8 100 座位がいろいろな強さといろいろな遺伝形式で寄与するような形質のリスク分布

## R ソース 3.3 R3-3.R : ●●●

```
1 covXY<-function(x,y) { # 2列の値ベクトルから共分散を算出
2   mx<-momentX(x,order=1,center=FALSE) (●? R2-Rの関数であること触れなくてOK ? ●)
3   my<-momentX(y,order=1,center=FALSE)
4   sum((x-mx)*(y-my))/length(x)
5 }
6 covMat<-function(m) { # すべての列ベクトルのペアについて共分散を計算
7   ret<-matrix(rep(0,length(m[1,])^2),nrow=length(m[1,]))
8   for(i in 1:length(m[1,])){
9     for(j in 1:length(m[1,])){
10       ret[i,j]=covXY(m[,i],m[,j])
11     }
12   }
13   ret
14 }
15 N<-1000;M<-100 # N サンプル数 M 2アレル型多型数
16 af<-runif(M) # アレル頻度 runif() 関数については、18.4 確率分布関数、
    疑似乱数列の発生を参照
17 r<-rnorm(M) # アレル 1本分の得点（座位の強さ）
18 d<-rnorm(M) # 座位ごとの相加モデルからのずれ
19 Xmat<-matrix(rep(0,N*M),nrow=N)
20 for(i in 1:M){ # 座位ごとに人数分の得点を計算
21   x<-rep(0,N)
22   x<-x+rbinom(N,1,af[i])
23   x<-x+rbinom(N,1,af[i])
24   x[x==1]<-1+d[i] # ヘテロ接合体特有の値（相加モデルからのずれ）を付与
25   x<-x*r[i] # 座位の強さを掛ける
26   Xmat[,i]<-x # この座位由来の得点を各人に加算
27 }
28 X<-apply(Xmat,1,sum) # 個人の全多型分の得点
29 momentX(X,order=2,center=TRUE) # 全多型分の得点の分散と
30 sum(covMat(Xmat)) # 分散共分散行列の値の和は一致する
31 hist(X) # 個人の全多型分得点の分布
```



## 第 II 部

### データ、サンプル、 サンプルの集まり



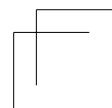
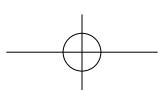
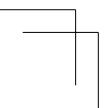
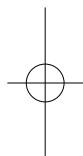
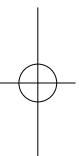
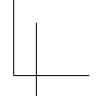
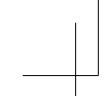
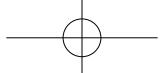
第4章 観察して評価すること

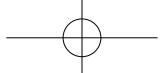
第5章 サンプルを個別に捉える

第6章 サンプルを集団として捉える

生物のジェノタイプ・フェノタイプとのその関係を理解するためには、ジェノタイプ・フェノタイプを観察し、それをデータとして取り扱うことが必要です。第II部では、データを取り扱うための根本であるデータの型についてを考えた後、データが帰属するサンプルを取り扱うためのアプローチを大きく2分して考えます。アプローチの1つは個々のサンプルを区別する方法であり、もう1つはサンプルを集団として捉える方法です。







# 第4章

## 観察して評価すること

### 4.1 データの種類と構成

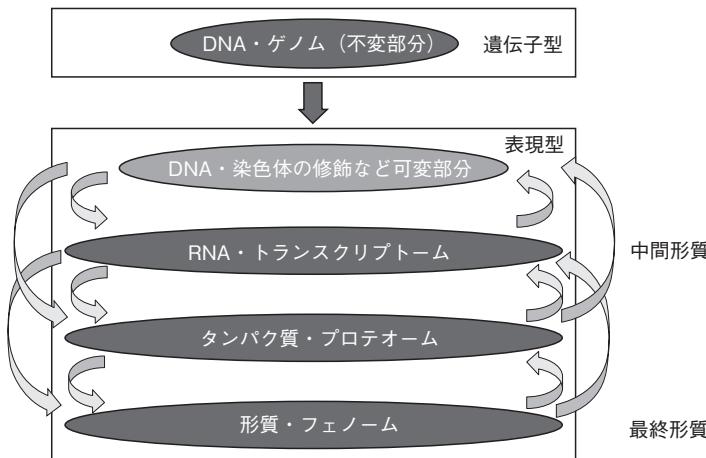
#### ● 4.1.1 遺伝子から見たデータの種類——遺伝子型と表現型、最終形質と中間形質

遺伝統計学では生命現象を遺伝子という観点で整理・解釈するので、データの種類と構成を考えるときにも、その視点から種類分けをします。2つに分かれます。

- ジエノタイプ（遺伝子型）
- フエノタイプ（表現型）

ジエノタイプというのは、親から子に伝えられる情報である「遺伝情報」に関するデータである、ということです。これは個体が生まれるとき（受精したとき）から死ぬまで、変わらずに保持する「何か」です。生まれるときからあるので、形質に影響を与えます。変わらずに存在するので、形質から影響を受けることはありません。ゲノムの型と言い換えてもよいかもしれません。DNA分子とそれを含む染色体は化学修飾を受けたりして、変化することもあるので、DNA分子・染色体の状態のうち、変化しないものの型をジエノタイプと考えましょう。塩基配列がそれに相当します。

そして、ジエノタイプでないものをすべてひっくるめて、フエノタイプとし



データはジェノタイプ（遺伝子型）とフェノタイプ（表現型）に分けられます。ジェノタイプは不变。フェノタイプは可変。フェノタイプは形質のものですが、それらは、ジェノタイプからの遠近関係により、最終形質とその中間形質という関係で見ることもできます。

図4.1 ジェノタイプとフェノタイプ

ます。表現型は相互に影響を及ぼしたり、及ぼされたりします。RNA の発現の状態（トランスクリプトーム）は DNA の配列に影響を受けますが、発現調節をするタンパク質の影響を受けますし、そのタンパク質の発現は、さまざまなタンパク質と RNA の影響を受けます<sup>※1</sup>。

同じフェノタイプと言っても、それらは、段階的に構成されています。DNA から転写されて mRNA ができます。さらに翻訳されてタンパク質ができます。そしてさらに、タンパク質をはじめとするさまざまな分子がさまざまな現象を起こします。ときには、その分子の量や機能を測定する検査などがあります。その検査は病気などの判断の有力な根拠になっているかもしれません。

このように、遺伝子から段階的に離れていきますが、mRNA もタンパク質も、検査も病気もすべて形質です。今、病気に特に興味がある場合、それは最終的に興味がある形質であるとして、**最終形質**と呼ばれます。このようなと

※1　塩基配列以外で親から子に伝達する情報もあります。その要素の影響を**エピジェネティクス**と言います。そのことを考えると、DNA の塩基配列と DNA・染色体の塩基配列以外の状態とを分けてしまうのは、乱暴に過ぎますが、本書では「割り切り方」の一方方法として、このようにしておきます。

き、最終形質と関係する、最終形質以外の形質は**中間形質**とも呼ばれます。

「遺伝子から最終形質までの道のり」の中間にあるからです。

遺伝子発現を研究標的にするときには、mRNA 自体が最終形質になります。

### ● 4.1.2 解析対象としてのデータの種類——データ型

#### (1) データ型

データを遺伝子の視点で分類しましたが、解釈・解析する視点から分類することにします。その分類が**データ型**です。分類の視点は次の 2 つです。

- とりうる値が限られている（離散的）かそうでない（連続的）か
- 順序があるかないか

離散的なデータ型は 2 個以上のカテゴリからなります。

表 4.1 データ型の分類

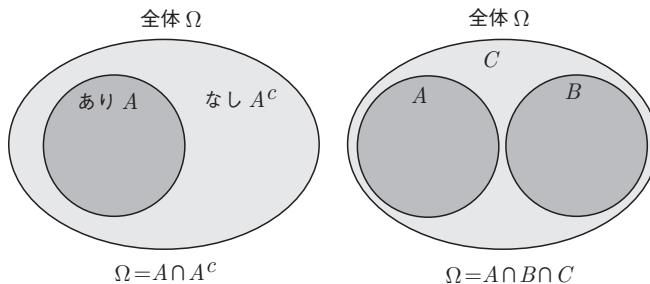
	離散		連続
順序の有無	2 カテゴリ	3 以上カテゴリ	量的
順序あり	△	○	○
一部順序あり	×	○	×
順序なし	△	○	×

この表では、○印はそのようなデータ型があることを示し、×印はないことを、△印は事情によって、あるともないとも言えることを示しています。

#### (2) 2 カテゴリ型は順序があるともないとも言える

- A さんは病気 X と診断されています（されていません）
- A さんの遺伝子 G の mRNA は検出されています（されていません）

このデータは、数え上げができる場合（カテゴリ）のどちらかに該当するので離散的です。場合の数が 2 つしかないので、カテゴリ数は 2 です。必ずどちらかのカテゴリに属します。どちらにも属さなかったり、両方に属することはできません。2 カテゴリ型は相互に排他的で、2 つのカテゴリを合わせると全体になります。全体集合に対して、部分集合とその補集合になっている、とも言います（図 4.2）。



全体を3カテゴリに分けると、相互に重なりのない（相互に排他的な）3つの集合に分かれます。 $A^c$ はAの補集合。

図4.2 全体を2つのカテゴリに分けると、集合とその補集合に分かれる

表4.1で、2カテゴリ型は、順序がある、ないのどちらにも△印がついています。これは、「診断・検出されている」ことを「されていない」ことに対して、「何かがある」と考えると、「ある」ことに1点、「ない」ことに0点をつければ、順序があることになりますし、「ある」と「ない」ことは並列な2つのこと、と考えれば順序がないことになります。以降の話では、順序があるデータ型と一緒に取り扱われることもあり、順序がないデータ型と一緒に取り扱われることもあります。

### (3) 3以上カテゴリ型

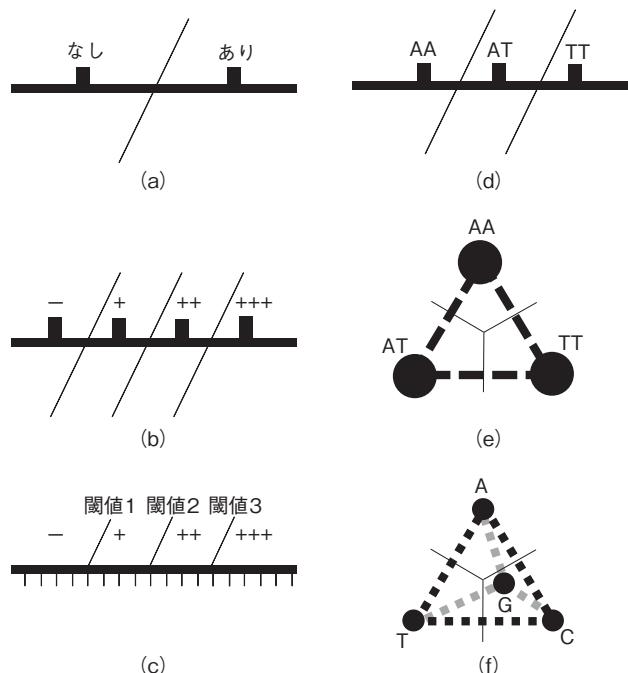
次の例を見てみます。

- Aさんの遺伝子Gの遺伝子多型Pの父由来の塩基は"A"です

塩基は"A" "T" "G" "C"の4つの可能性があります。そしてそれ以外の可能性はなく、必ずどれか1つをとります。ですから、4つのカテゴリを持つデータ型です。ここで、"A" "T" "G" "C"には特に順番の定めようもないで、順序なしです。図4.3 (f) のように3次元空間を4つの部分に分けることに相当します。

- Aさんの遺伝子多型Pのジェノタイプは"AT"です

この例では、3つのジェノタイプ"AA", "AT", "TT"がありえます。離散的な3つのカテゴリで、順序がないと見ることができます。この場合は図4.3 (e) のように平面を3つの部分に分けることに相当します。別の見方では、この遺



直線を区切る線を引くことによっていくつかの線分に分けると、順序のあるカテゴリに分けられます (a) (b) (d)。区切りの線を引かずに数直線として扱えば連続型です (c)。3 以上のカテゴリは、全体が直線では表せず、平面・立体・高次元空間になります。この場合も全体を複数の相互に排他的なカテゴリに仕切る点は同じです。(e) (f) の 2 つの図は 2 次元平面を 3 カテゴリに分ける様子と、3 次元空間を 4 カテゴリに分ける様子です。

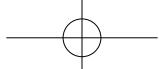
図 4.3 全体が直線で表せるときは、順序ありのデータ型とみなせる

伝子多型の遺伝子型としてはアレル "T" の本数に着目すると、0, 1, 2 本というように、数値を割り当てることができます。この場合は、順序のあるカテゴリです。この場合は、直線を 3 つに仕切ることに相当します (図 4.3 (d))。どちらで扱うかは、どのような点に着目したいかによって変わります。

一方、次の例は

- A さんの検査 T (定性) の結果は "++" (強い) です

カテゴリ数は 3 個以上ですが、順序があります。"++" (強く陽性 (20 ~ 30)) は、"−" (陰性 (0-10)) , "+" (陽性 (10-20)) , "++" (強く陽性 (20 ~ 30)) , "+++" (非常に強く陽性 (30 以上)) という 4 個のカテゴリのうち、



2番目に強いカテゴリです。順序があるのは、直線で仕切ったものであるとして定義されているからです。量的データに、仕切り線の位置を決める値（閾値（10, 20, 30））が存在する場合もこのような形になります。

#### (4) 連続型

- Aさんの検査T（タンパク質T：遺伝子Gの翻訳産物）（定量）の結果は256ユニットです

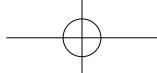
量で量ると数直線という1本の直線で表すことができ、データはこの直線上の点として表すことができます。では、AさんとBさんの2人のデータが、ともに256ユニットであったとき、2人のタンパク質量は同じなのでしょうか。たとえ256ユニットと同じ値だとしても、十分に精密に較べれば、全く同じということはないでしょう。そういう意味で、量的データの値は、数直線上の点で表される値を適当な精度で丸めた値であると考えておくのがよいでしょう。精度という細かさで多くの順序ありカテゴリに分けられたカテゴリ型のデータであるが、そのカテゴリ数が十分多いので、カテゴリであることを忘れてもよいというように考えてもよいかもしれません。

##### ● 4.1.3 一部に順序があること

では、順序が一部にある場合というはどういう場合でしょうか。2つの2アレル型の多型のジェノタイプのデータを見てみます。

	CC	CG	GG		0	1	2
AA	AA, CC	AA, CG	AA, GG	0	(0, 0)	(0, 1)	(0, 2)
AT	AT, CC	AT, CG	AT, GG	1	(1, 0)	(1, 1)	(1, 2)
TT	TT, CC	TT, CG	TT, GG	2	(2, 0)	(2, 1)	(2, 2)

この表では、縦に多型P(A/T)、横に多型Q(C/G)のディプロタイプが表されています。これは、3個のカテゴリを持つ2つの因子の組み合わせのカテゴリ型データで、そのカテゴリは $3^2 = 9$ 個ある場合です。では、Pのアレル "T" の本数と Q のアレル "G" の本数に着目して P も Q も順序ありと考えます。9通りのカテゴリはそれぞれ P と Q とのアレルの本数のパターンで表すこともできます。特に特殊事情を考慮しなければ、P の "T" の本数に順序があつ



て、Q の "G" の本数に順序を考えてよく、P の "T" も Q の "G" も病気 X のリスクを上げているとき、

$$(0, 0) < (0, 1) < (0, 2)$$

$$(0, 0) < (1, 0) < (2, 0)$$

$$(0, 0) < (1, 1) < (2, 2)$$

の順序は問題がないでしょう。

はっきりしないのは、(0, 1) と (1, 0) の順序、(0, 2) と (1, 1) と (2, 0) の順序、(1, 2) と (2, 1) の順序です。これが、一部に順序があるけれども、一部に順序がない場合です。このような場合を**半順序**と言います。全体に順序をつけるためには、何かしらのルールを入れる必要があります。すべてに順序がある場合は**全順序**と言います<sup>※2</sup>。

#### ● 4.1.4 カテゴリの組み合わせ

多型 P, Q の例では、2 つのカテゴリ型を組み合わせていました。カテゴリ数が  $k_i$  であるようなカテゴリ変数を  $N$  個組み合わせれば、

$$\prod_{i=1}^N k_i = k_1 \times k_2 \times \dots \times k_N$$

通りのカテゴリが生じます。

#### ● 4.1.5 唯一選択、重複選択

複数の項目からなる診断基準の例を見てみます。

- Aさんは 5 項目の診断基準のうち、3 項目を満足しています

これは 5 項目のうちの該当する項目をいくつでも選ぶというタイプのデータ型です。それらはいくつかのパターンに分けられます。5 カテゴリ ("a", "b", "c", "d", "e") から 1 つ選ぶときに得られたデータの記録方法としては次の表のような方法が考えられます。

※2 順序、全順序、半順序は数学で定義された概念です。ここでは、データの型を考えるためにあたって、順序という既存の概念を使えることを理解することが大事であって、その詳細には立ち入りません。詳細は成書を参照してください。

サンプル	カテゴリ
サンプル 1	b
サンプル 2	a
サンプル 3	a
サンプル 4	c

「 $n$ 個の中から選んでください。選ぶ数はいくつでも構いません」というような選択肢の場合もよく見かけます。選ぶ数に上限を設けて、「 $n$ 個の中から選んでください。ただし、最大3個までとします」ということも比較的多いかもしれません。こんな場合は、次のように記録することになるでしょう。

サンプル	選択カテゴリ 1	選択カテゴリ 2	選択カテゴリ 3
サンプル 1	b	—	—
サンプル 2	a	d	c
サンプル 3	a	c	d
サンプル 4	c	e	—

サンプルによって、選択カテゴリ数が異なると表の形で記録するのが難しくなります。サンプル2とサンプル3はともにa, c, dを選んでいますが、表の上では異なっています。次のようにしてみます。

サンプル	a	b	c	d	e
サンプル 1	0	1	0	0	0
サンプル 2	1	0	1	1	0
サンプル 3	1	0	1	1	0
サンプル 4	0	0	1	0	1

列の数が増えましたが、表の中はシンプルになりましたし、選択された項目の順番について気をつかわなくても同じ選択パターンは同じレコードになっています。

この方式だと、「最大3個を選び、それに順序をつけなさい」という選択方法にも簡単に対応できます。「選ばないカテゴリを0点、1番目に選んだ項目に3点、2番目に選んだ項目に2点、3番目に選んだ項目に1点」と重みを付けることにして、

サンプル	a	b	c	d	e
サンプル 1	0	3	0	0	0
サンプル 2	3	0	1	2	0
サンプル 3	3	0	1	2	0
サンプル 4	0	0	3	0	2

のように、表の中の値に重みを反映するだけでうまく記録ができます。

言い方を変えると、順序のない 3 以上カテゴリ型の変数は、カテゴリごとに 2 カテゴリ型変数を割り当てて、それを組み合わせたものとして捉えることができる事が、データの記録方式からわかります。

ただし、個々のカテゴリに割り当てた変数同士が独立かというと、そうではありません。5 カテゴリから 1 個だけを選択する条件の場合には、5 カテゴリに入れる値は 0 か 1 で、その和は 1 でなくてはならない、という制約が 5 つの変数の間にありますし、最大 3 個を選ぶという条件の場合には、5 カテゴリに入れる値は 0 か 1 で、その和は 3 以下でなくてはならない、という制約があります。

$N$  個のカテゴリから、 $k$  個を選ぶとき、その選び方は  $\binom{N}{k} = \frac{N!}{k!(N-k)!}$  通り

あります（組み合わせ。選んだ項目が何番目かも問題にするなら、 $\frac{N!}{(N-k)!}$  通りあります（順列））。

A, T, G, C を並べて長さ  $L$  の塩基配列を作るとき、その配列の場合の数は、重複順列で  $4^L$  通りでした。これに対して、3 以上カテゴリから選択するときの場合の数はこれより小さくなっています。この違いは、塩基配列のときに組み合わせた各塩基同士は独立であったのに対して、複数項目からの選択の場合には、選択項目間は独立でないことを反映しているとも言えます。

### ● 4.1.6 2 倍体という特殊性——ハーディ・ワインバーグ 平衡 (HWE) の正確検定

カテゴリに注意して表を作ることが便利であるとの例として、HWE の正確確率検定を挙げることにします。

ジェノタイプは父母由来のアレルの組み合わせです。今、ある 2 アレル型多型があり、その 3 ディプロタイプ "AA", "AT", "TT" の人数がそれぞれ、 $n_0$ ,  $n_1$ ,  $n_2$ ;  $n_0+n_1+n_2=N$  人だとします。アレル A と T の染色体本数が  $n_A$ ,  $n_T$  とし

ます。 $N$ 行2列の表があり、各行の和がすべて2で、第1、2列の和が $n_A$ ,  $n_T$ で、総和が $2N$ である表ができました。

サンプル	A の本数	T の本数	計
AA1	2	0	2
AA2	2	0	2
...	...	...	...
AA $n_0$	2	0	2
AT1	1	1	2
AT2	1	1	2
...	...	...	...
AT $n_1$	1	1	2
TT1	0	2	2
TT2	0	2	2
...	...	...	...
TT $n_2$	0	2	2
計	$n_A$	$n_T$	$2N$

これは「2カテゴリから、重複を許して必ず2個選べ」というタイプのデータ型です。

第1列の値が0, 1, 2のときには、第2列の値は必ず2, 1, 0という関係になっていて、片方の列はなくともわかります。2列分の情報は冗長です。では、この2列は全く不要なのかというと、そうではありません。2アレル型多型のハーディ・ワインバーグ平衡を考えるときには、この2列を持つ表が活用されます。この表を用いて、HWEの正確確率検定をしてみます。

一般的な分割表の正確確率検定については、第13章を参照してください。

このような分割表の生起確率は、2のセルが $n_0 + n_2$ 個、0のセルが $n_0 + n_2$ 個、1のセルが $2 \times n_1$ 個、それぞれあるので、

$$\frac{n_A! n_T! \times \prod_{i=1}^N (2!)^{n_i}}{(2N)! (2!)^{n_0+n_2} (0!)^{n_0+n_2} (1!)^{2 \times n_1}} = \frac{n_A! n_T! \times 2^{n_1}}{(2N)!}$$

となります。

$0! = 1$ ,  $1! = 1$ ,  $2! = 2$ ,  $N = n_0 + n_1 + n_2$ を使って式変形しました。

ここまででは、 $N$ 行を区別していました。これは $N$ 人のそれぞれを区別していたことになるのですが、今は $N$ 人のうち3ジエノタイプの内訳が $n_0, n_1, n_2$ である場合を区別しないので、 $n_0, n_1, n_2$ 人で観測する場合は  $\frac{N!}{n_0!n_1!n_2!}$  通りあり、これらの区別がつきません。したがって、そのような確率は結局、

$$\frac{N!}{n_0!n_1!n_2!} \times \frac{2^{n_1} n_A! n_T!}{(2N)!} = \frac{2^{n_1} n_A! n_T! N!}{(2N)! n_0! n_1! n_2!}$$

となります。

観測した表の生起確率がわかりましたので、正確検定をするためには、次のようにします。

まず、与えられた条件 ( $n_0, n_1, n_2$  が与えられた条件。必然的に  $n_A, n_T$  も決まります) のもとで、起こりえるすべての表を数え上げ、その生起確率を求めます。その上で、観測表の生起確率以下の生起確率をとる表について足し合わせます。

今、 $n_A + n_T = 2N$  なので、 $n_A$  と  $n_T$  はともに偶数か、ともに奇数のどちらかです。 $n_A, n_T$  が奇数のときには、ヘテロ型の人数  $n_1$  は奇数しかとらず、 $n_A, n_T$  が偶数のときには、 $n_1$  は偶数しかありません。

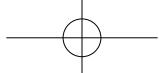
そのことを用いて HWE の正確確率検定を R で行ってみるソースが R4-1.R です。とりうる場合の確率のすべてを足し合わせると確かに 1 になっていることも確認できます。

#### R ソース 4.1 R4-1.R : HWE 正確確率検定

```

1 hweExact<-function(g=c(813,182,5)) { # 3ジエノタイプの人数
2 n<-sum(g) # 総人数
3 nA<-2*g[1]+g[2] # Aアレル本数
4 na<-2*g[3]+g[2] # aアレル本数
5 evod<-g[2]%%2 # ヘテロ人数の偶数奇数判断
6 maxAa<-min(nA,na)-evod
7 Aa<-seq(from=evod,to=maxAa,by=2) # 観測しうるヘテロ人数のベクトル
8 AA<-(nA-Aa)/2 # 観測しうるAA人数
9 aa<-(na-Aa)/2 # 観測しうるaa人数
10 obs<-(g[2]-evod)/2+1 # 観察データのヘテロ人数はAa[obs]
11 prob<-rep(0,length(Aa)) # 観測しうる表の生起確率
12 prob<-exp(n*lgamma(2+1)+lgamma(nA+1)+lgamma(na+1)-

```



```
lgamma(2*n+1) - (AA*lgamma(2+1)+Aa*lgamma(1+1)
+aa*lgamma(2+1))+lgamma(n+1)-(lgamma(AA+1) +
lgamma(Aa+1)+lgamma(aa+1)))
13 p.value<-sum(prob [prob<=prob [obs]])
# 観測表の生起確率以下の生起確率を持つ表の生起確率の和
14 # Aa 観測しうるヘテロ人数リスト
15 # prob ヘテロ人数別の生起確率
16 # obsprob 観察テーブルの生起確率
17 # p.value 正確検定p値
18 list(Aa=Aa,prob=prob,obsprob=prob [obs],p.value=p.value)
19 }
20 xx<-hweExact(c(813,182,5))
21 xx$p.value # 検定p値
22 sum(xx$prob) # 全表の生起確率の和は1になります
```

### ● 4.1.7 親項目と子項目

複数の項目を組み合わせたときに、それらが独立である場合もあれば、相互に関係がある場合もあることはすでに述べてきました。

- Aさんは病気Xのタイプbと診断されています

このカテゴリ型項目はどうでしょうか。この項目のデータは病気Xという項目が「あり」の場合にのみ存在します。病気Xという項目を親項目とすれば、その子項目と言えます。アンケート調査などにもこのような項目間の帰属関係があります。「前の設問で（1）と答えた場合に、以下の設間に答えよ」という形式の場合です。データの構造が複雑になってきたときには、このような項目間の相互帰属関係を明確にして取り扱うことも適切です。

このような項目の帰属関係はフェノタイプにのみあるわけではありません。ジェノタイプの場合には、次のような例があります。Y染色体上にあって、Y染色体固有の遺伝子に関するジェノタイプは、Yを持っていて初めて意味をなしますから、Y染色体の保有に関する項目と、Y染色体固有のジェノタイプに関する項目は親子関係です。また、挿入型の遺伝子多型があって、その挿入配列がさらに多型性を持つ場合にも、遺伝子型の間に親子関係があります。親子関係はグラフ理論で言うところの木で表すことができますので、第5章で扱うグラフに関する説明が参考になります。

### ● 4.1.8 カテゴリの配置、カテゴリ間の非独立性、正単体

5 カテゴリから 3 カテゴリを選択するときに、5 列の表でデータを表すときには、表に 0 か 1 の値を入れ、行についての和が 3 になるようにするという制約がありました。ある列に入れる値によって、別の列に入れる値が影響を受けています。これが、カテゴリ間の非独立性です。このことを異なる視点から眺めてみます。

相互に関係のない  $n$  カテゴリは、**正単体**の頂点に配置できることが知られています。正単体とは、正三角形を多次元一般化したものと考えればよく、次のように説明できます。

- $k$  次元空間にある  $k$ -正単体は  $k+1$  個の頂点を持ち、すべての頂点は中心から等距離にあって、中心から各頂点への方向はどの 2 頂点を取り上げて

$$\text{も等しい角度 } \theta \text{ をなす } (\cos\theta = -\frac{1}{k})^{\text{※3}}$$

---

※3  $k+1$  次元空間の  $k+1$  個のベクトルを、 $(e_i = 0, 0, \dots, 0, 1, 0, \dots, 0); i = 1, 2, \dots, k+1$

という単位ベクトルで考えます。この  $k+1$  個のベクトルは  $k+1$  次元空間にあって、すべてのベクトルは直交しています。今、このすべての中心を

$$o = \frac{1}{k+1} \sum_{i=1}^{k+1} e_i = \frac{1}{k+1} (1, 1, 1, \dots, 1)$$

とします。ここで、

$$f_i = e_i - o \\ = \left( -\frac{1}{k+1}, -\frac{1}{k+1}, \dots, -\frac{1}{k+1}, 1 - \frac{1}{k+1}, -\frac{1}{k+1}, \dots, -\frac{1}{k+1} \right)$$

とします。 $\sum_{i=1}^k f_i = -f_{k+1}$  なので、これらは、 $k$  次元空間に納まっており、すべての  $f_i$  は同じ長さ

$$|f_i| = \sqrt{\frac{1}{(k+1)^2} (k * 1 + k^2)} = \sqrt{\frac{k}{k+1}}$$

を持ち、任意の  $i, j, i \neq j$  について、

$$f_i f_j = (k-1) * \left( \frac{1}{k+1} \right)^2 + 2 \times \left( -\frac{1}{k+1} \right) \left( 1 - \frac{1}{k+1} \right) = -\frac{1}{k+1}$$

であるから、そのなす角  $\theta_{ij}$  は  $\cos(\theta_{ij}) = \frac{f_i f_j}{|f_i| |f_j|} = -\frac{1}{k+1} \frac{k+1}{k} = -\frac{1}{k}$  です。

低次元の場合には次のようにになります。

- 1次元空間（直線）：1- 正単体 線分
- 2次元空間（平面）：2- 正単体 正三角形
- 3次元空間（空間）：3- 正単体 正四面体

$k+1$  個のカテゴリデータ型ではすべてのカテゴリが相互に対等な関係にあって、 $k$  次元空間中の  $k$ - 正単体の頂点とみなすことができます。

$k=1$  のとき（カテゴリ数が 2 のとき）は、2- 正単体の頂点に配置できるわけですが、2- 正単体とは、ただの線分で、その両端に 2 つのカテゴリが配置されています。直線配置できる点が連続データ型と共通であって、直線配置できることが「順序あり」扱いできることですので、2 カテゴリカルデータ型は必ず順序あります。また、2 カテゴリは相互に逆向きのベクトルなので、片方のデータを知れば、もう片方のデータは自ずから確定するのも、こういう事情です。

さて、カテゴリ数  $k$  を大きくしていくと、 $\cos(\theta_{ij}) = -\frac{1}{k}$  は 0 に収束しています。 $\theta_{ij}$  は直角に近づいていくということです。直交関係は独立な関係なので、カテゴリ数が大きくなると、カテゴリ同士は相互に独立な関係に近づいていくことがわかります。第5章で扱うグラフで正単体を描けば、頂点数  $k+1$  の完全グラフとなります。すべての頂点の間に辺を引いたグラフで、**完全グラフ**という名前で呼ばれます。また、正単体という高次元表現にも、グラフという平面表現にも共通するのは、すべての頂点を相互に平等に扱う方法であるという点です。

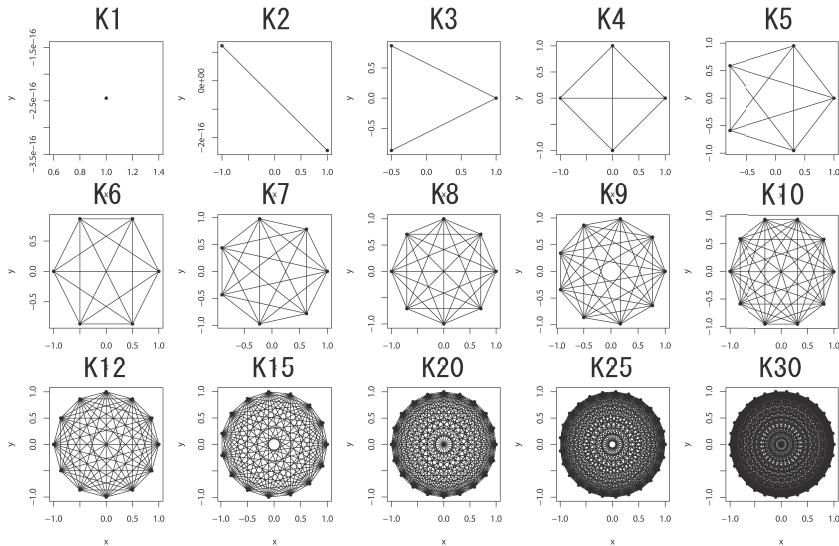
頂点数 : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25, 30。 $K_n$  の  $n$  が頂点数。

図 4.4 カテゴリを表す正単体は完全グラフ

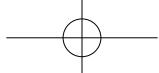
## 4.2 データを比較する

### ● 4.2.1 2つのデータの関係——対称的な関係と非対称的な関係

前節ではデータの分類についてまとめました。今度はデータを引き較べることにします。引き較べることを、**関係**<sup>※4</sup>を見る作業と呼ぶことにします。まず、関係には方向性があって、A のデータの B のデータに対する関係を見るのか、B のデータの A のデータに対する関係を見るのかの 2通りのやり方があります。

- 病気 X と診断されているかどうかという項目について、A さんは "1" (はい、そうである)、B さんも "1" です

※4 ここで言う「関係」は数学的には**二項関係**と定義されています。



A と B との関係は、 $A=B$  か  $A \neq B$  のいずれかです。どちらの場合も A のデータの B のデータに対する関係と B のデータの A のデータに対する関係と同じです。これを、「**対称的**」な関係と言います。次は「**非対称的**」な関係の例です。

- AさんはXを1つ持ち、Bさんは2つ持ちます

このデータは順序があります。そして、AさんのデータはBさんのそれに対して「小さく」、BさんのデータはAさんのそれに対して「大きい」です。2方向の関係が異なりますから、これは、「**非対称的**」な関係です。

### ◎ 4.2.2 非対称な関係を対称にする——距離

順序のあるデータ型は非対称な関係をもたらしました。次の例を見てみます。

- Aさんの検査T（定性）の結果は "++" で、Bさんのそれは "+++" です

順序のあるカテゴリ型データで、非対称な関係です。これを

- Aさんの検査T（定性）の結果は Bさんのそれと "1" 段階違います
- Bさんの検査T（定性）の結果は Aさんのそれと "1" 段階違います

と評価するとします。こうすると、「対称的」な関係になります。順序があるデータについて「差」をとることによって、「対称的」にしています。その結果、負の値をとらなくなりました。絶対値を採用したとも言えます。これは、**距離**と呼ばれるものです。

距離を扱いやすいように定義します。

#### (1) 距離の定義

- 2つの何か (A, B) の間に定められ
- 非負 (0以上) の値をとり
- その値は、A の B に対する値も B の A に対する値も同じであり
- A と B とが同じときには 0 をとり
- 3つの何か (A, B, C) があるときには、A-B, B-C, C-A の間の距離は、A, B, C の3つを平面に配置して三角形（つぶれてしまって直線になってしまっても

よい) が描けるような値をとるもの (三角不等式を満足する)

## (2) 三角不等式

ここで、距離を定義するのに三角不等式という用語が出てきました。これを説明します。

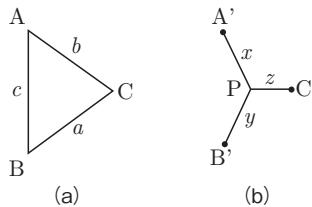


図 4.5 距離

図 4.5 (a) を見てください。平面上に 3 点 A, B, C があります。A から C へ行くのに、まっすぐ行く方が B を経由して行くよりも距離が短いです。この関係がどの点からどの点へ行く場合にも当てはまります。△ ABC の辺 AB, BC, CA の長さがそれぞれ  $c, a, b$  とすると、

$$b + c > a, c + a > b, a + b > c$$

が成り立ちます。このとき、

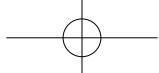
$$x = \frac{a+b+c}{2} - a = \frac{1}{2}(b+c-a) \geq 0$$

$$y = \frac{a+b+c}{2} - b = \frac{1}{2}(c+a-b) \geq 0$$

$$z = \frac{a+b+c}{2} - c = \frac{1}{2}(a+b-c) \geq 0$$

$$x + y = c, y + z = a, z + x = b$$

が成り立ちますから、 $A', B', C'$  の 3 点を点 P からそれぞれ  $x, y, z$  の距離に配置すれば (図 4.5 (b))、 $A'$  から  $B'$  へ P を経由して行く道のりの長さは、AB の道のりと同じです。 $B'C', C'A'$  の場合も同様です。したがって、**三角不等式** が成り立つような 3 点の距離関係は、3 点間の道のりを変えることなく、枝分かれで描けることがわかります。本章の後半では、複数の点の間の距離の情報



から、この枝分かれ道（木）を描く手法のことを述べますが、距離の定義はそのような木を作れることを保証しています。

### ◎ 4.2.3 ユークリッド距離とそれ以外の距離

データが数直線上にあるとき、関係を距離として表すことができました。私たちが日常生活で使っているいわゆる距離は、平面上（2次元）にも空間中（3次元）にも定義されています。それは**ユークリッド距離**と言われるもので、 $k$ 次元空間においては点  $A = (a_1, a_2, \dots, a_k)$  と点  $B = (b_1, b_2, \dots, b_k)$  の間のユークリッド距離は

$$d_E(A, B) = \sqrt{\sum_{i=1}^k (a_i - b_i)^2} = \sqrt{\sum_{i=1}^k \delta_i^2}$$

ただし、 $\delta_i = |a_i - b_i|$  と定義されます。

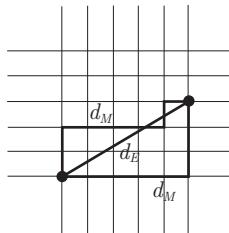
しかしながら、距離の定義はすでに述べた通りで、この式を必ずしも要求しません。ですから、遺伝子型や表現型のデータを使って距離を定めるときも、ユークリッド距離に限る必要はありません。たとえば、

$$d_M(A, B) = \sum_{i=1}^k |a_i - b_i| = \sum_{i=1}^k \delta_i$$

で定義される**マンハッタン距離**というものがあります。マンハッタン距離は、図 4.6 のように格子の辺をたどっていくときの距離です。

ユークリッド距離とマンハッタン距離以外にも距離はいくつも定義できます。脚注を参考に R を活用していろいろな距離の名称と定義を確認してみてください※5。

※5 R で距離を扱う関数である `dist()` 関数では、"euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski" の 6 種類から選択することができます。`dist()` 関数の説明文書（R のヘルプ記事）を表示したり、`dist()` 関数のソースコード（R のソース）を表示することで、これらの内容がわかります。



2つの黒丸の間に直線で結んだ距離がユークリッド距離 ( $d_E$ ) で、格子（マンハッタン街区のような）を最短でたどるのがマンハッタン距離 ( $d_M$ ) です。ユークリッド距離の道は1通りですが、マンハッタン距離の道のとり方は1通りではありません。今、 $k$  次元で  $d_M = \sum_{i=1}^k \delta_i$  であるとすると、通り道のとり方の場合の数は、 $\binom{d_M}{\delta_1 \delta_2 \dots \delta_k} = \frac{d_M!}{\prod_{i=1}^k \delta_i!}$  です（第15章参照）。

図 4.6 マンハッタン距離

#### ◎ 4.2.4 配列の違いとマンハッタン距離

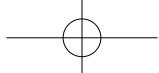
たとえば、DNA 塩基配列同士の距離は、塩基の異なる箇所数で表すことができます。

- ヒト : cccggaCAcCgActtcccGgggctcatt
- マウス : cccggaTGcAgGcttcccAgggctcatt

この部分の配列の長さは 29 塩基で、5 塩基が違います。これは、29 次元空間で、マンハッタン距離が 5 と考えることができます。変異が 5 箇所に起きたことをマンハッタン距離 5 とみなしているわけです。配列の比較には、塩基の置換の他に、挿入・欠失も考慮が必要です。

- ヒト : ...cccggaCAcCgActtcccGgggctcattACcctCAc...
- マウス : ...cccggaTGcAgGcttcccAgggctcatt=Tcct=Tc...

この場合には、置換と挿入・欠失という異質な出来事を総合的に考えて距離を決める必要があります。それぞれの出来事に、生物学的な背景を考慮して重み付けをします。また、塩基配列の比較では、2 配列がランダムに作られたという仮説に立つと、ありえないほど似ており、そのありえなさを数値にして、配列の似ている程度を数値化しますが、その珍しさを極値分布に照らして  $p$



値として数値化することができます。よく使われる公共サイトのblastなどがその例です。

### ◎ 4.2.5 距離の代わりに角度で表す——相関係数

距離は対称な関係を0以上の値で表現しました。負の数も含めて関係を量的に評価することはできないでしょうか。カテゴリ型を空間に配置したときに、

すべてのカテゴリが相互に  $\theta = \cos^{-1} \frac{1}{k}$  なる関係<sup>※6</sup>にあって、この角度が均等

であることが、カテゴリの平等な関係を意味していると述べました。このように角度も2つのものの間の関係を表すことができます。 $\theta$ をその関係を表す量とすることができますし、 $\cos \theta$ をその量としてもよいでしょう。今、2つのデータがベクトル形式であれば、ベクトル同士の内積と長さで表されます。

$$\cos \theta = \frac{\sum_{i=1}^k x_i y_i}{|x||y|}$$

これは相関係数と呼ばれる値です。



## 4.3 複数のサンプル、たくさんの比較



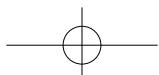
### ◎ 4.3.1 1対N-1とN対N

2個のサンプルを比較してきました。ここからは、N個のサンプルを較べます。

ある1個に特に興味があって、それ以外の(N-1)個との関係が調べたいときもあります。N個の全部に興味があるときには、すべての関係を調べることになります。関係が対称でないときには、 $N \times N$ (重複順列)の比較をしますし、関係が対称であるときには自身との関係の情報もとりたければ、

$\frac{N(N+1)}{2}$ ペア(重複組み合わせ)の関係を調べますし、自身との関係は"0"

※6  $\cos \theta = x$  のとき  $\theta = \cos^{-1} x$



であって調べるには及ばないということであれば、 $\frac{N(N-1)}{2}$  ペア（組み合わせ）の関係を調べます。 $N \times N$  の関係を図 4.7 のように行列で表せば、 $N \times N$  行列全体、対称的な関係のときは**下三角行列**部分のみ、1 対その他ならば 1 行のみを使います。（●?本文中は N、図中は n、OKor どちらかに統一?●）

$v(1, 1)$	$v(1, 2)$	...	$v(1, i)$	...	$v(1, n)$
$v(2, 1)$	$v(2, 2)$	...	$v(2, i)$	...	$v(2, n)$
...	...	...	...	...	...
$v(i, 1)$	$v(i, 2)$	...	$v(i, i)$	...	$v(i, n)$
...	...	...	...	...	...
$v(n, 1)$	$v(n, 2)$	...	$v(n, i)$	...	$v(n, n)$

すべての組み合わせは正方形行列全体 ( $N \times N$ )、対称な関係を見るときは、三角行列部分のみが対象。

図 4.7 比較の仕方

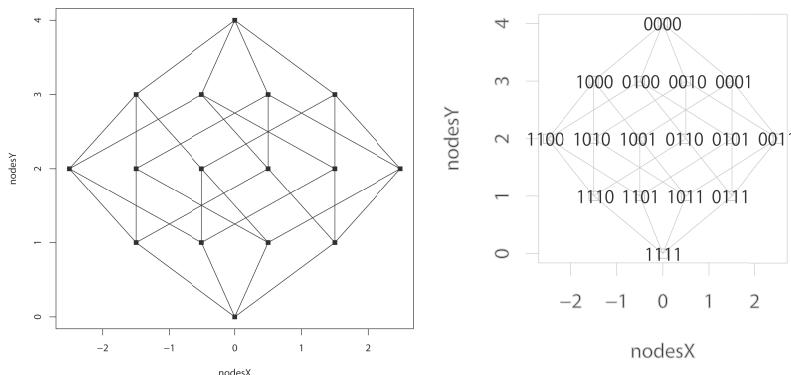
### ● 4.3.2 一部に関係がないとき——半順序

$N \times N$  のすべてに関係が定義できれば行列状に関係を見ればよいですが、関係が定義できない場合が含まれることもあります。

例を挙げます。4 個の選択肢からの選択を考えます。1 個も選ばない場合を  $\{0, 0, 0, 0\}$ 、1 番と 3 番を選ぶ場合を  $\{1, 0, 1, 0\}$  と表すことにします。こうすると、選び方は全部で  $2^4 = 16$  通りあります。1, 2, 3 番を選んだ場合と 1, 3 番を選んだ場合は、前者の選択は後者の選択を含みます（包含関係にあります  $\{1, 2, 3, 0\} \subset \{1, 0, 1, 0\}$ ）。しかしながら、 $\{1, 1, 1, 0\}$  と  $\{1, 0, 0, 1\}$  との関係は、包含関係が成り立ちません。このように包含関係が定まる場合と定まらない場合があります。 $2^4$  通りを 4 次元空間の点みなして、 $k = 4$  次元の長さ 1 の**立方格子**<sup>※7</sup>の点として表すとします。

要素数が 1 個だけ異なっていて、包含関係にある点同士が辺で結ばれます。 $2^k$  個の点を持つ多次元の格子は平面に描くのが難しいので、図 4.8 のように平面に押しつぶして描くことができます。図の上部にある点から下向きの辺だけをたどって下部にある点たどりつける場合には、上の点は下の点に含まれる関

※7 1 次元の立方格子点  $(0), (1)$  を結んだ線分、2 次元立方格子は点  $(0, 0), (0, 1), (1, 0), (1, 1)$  を結んだ正方形、3 次元立方格子は点  $(0, 0, 0), (0, 0, 1), (0, 1, 0), (1, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0), (1, 1, 1)$  を結んだ立方体。 $k$  次元立方格子は  $2^k$  個の点からなり、その点の座標の成分は 0 か 1 でできている。



ハッセ図とも言います。要素数 4 の集合の各要素を含むか否かを 0, 1 で表して、ハッセ図の各頂点に部分集合を割り当てれば（右図）、ハッセ図は部分集合間の包含関係を表します。0 と 1 の数値の並びをハプロタイプとみなせば、1 回の変異で変化しうる間柄のハプロタイプ同士が結ばれていることになります（R4-sup1.R（この R ソースはダウンロードすることができますが、本書には掲載していません）でハッセ図を描けます）。

図 4.8 4 次元立方格子を平面に描く

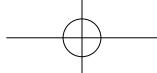
係になっていきます。

このように一部に順序があり、一部に順序がないときに、このような図で表されることがわかります。ハッセ図と言います。「一部に順序がある」という呼び方をしていましたが、それを半順序と言います。「全部に順序がある」のが全順序です。

"0, 1" を 2 つのアレルとすれば、4箇所の 2 アレル型多型が作る 16 種類のハプロタイプを変異回数で結んだ場合も多次元立方格子で表されます。

### ● 4.3.3 距離行列と木

今、 $N$  個の要素同士のすべてに対称的な関係があつて距離として表されるとします。 $N$  個のサンプルの  $N$  個のサンプルに対する関係を調べるとします。距離で表す関係は対称なので、（下）三角行列と呼ばれる行列に納めることができます。さらに、距離であるときには、自身と自身との関係は 0 であつて不要なので、以下に示すように、対角成分のない三角行列成分として表すことができます。これが距離行列です。行番号（2, 3, 4, 5）と列番号（1, 2, 3, 4）とともに距離が表示されています。



## ■ R の出力結果

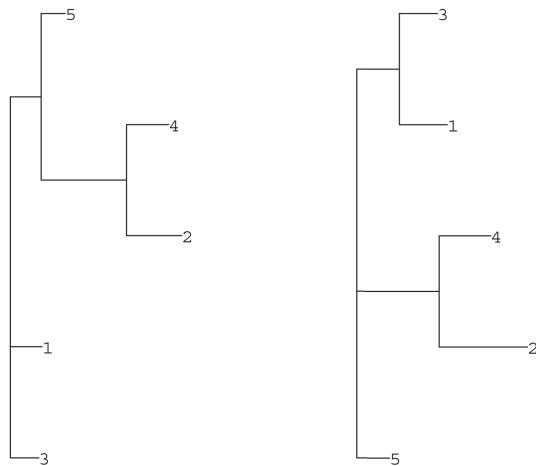
1	2	3	4
2 4.725108			
3 1.372943 4.600435			
4 4.135837 2.220284 4.311534			
5 2.020896 3.608861 1.844573 3.591590			

距離の定義としてユークリッド距離とマンハッタン距離を用いて、木を作ります。R のパッケージ<sup>※8</sup>ape を取り込み、ape が持つ、近隣結合法 (Neighbor-Joining) による木作成関数 nj() を使います。距離の定義が違うので、距離行列も異なりますから、そこから作られる木も少し違ってきます。木の作成はサンプルの階層的クラスタリングの一種です。階層的クラスタリングは引き続き第 5 章で扱います。

### R ソース 4.2 R4-2.R : 距離行列と近隣結合法による木の作成

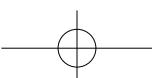
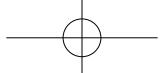
```
1 # サンプル数 Ns=5 が 5 次元のデータを持っているようなデータセットを作成
2 Ns<-5;k<-5;x <- matrix(rnorm(Ns*k), nrow=Ns)
3 dist(x,method="euclidean") # ユークリッド距離の距離行列
4 library(ape) # 木を作る関数 nj() を持つパッケージ ape の読み込み
5 treu<-nj(dist(x,method="euclidean"))
   # ape の近隣結合法関数 nj により木を作る
6 trman<-nj(dist(x,method="manhattan"))
   # マンハッタン距離で木を作る
7 par(mfcol=c(1,2)) # 画面を 1 行 2 列に分割
8 plot(treu);plot(trman) # 2 つの距離法で木の表示
9 par(mfcol=c(1,1)) # 画面分割を 1 行 1 列に戻す
```

※8 R のパッケージの取り込みは付録 A の A.2 節を参照。



同じデータを使っても距離の定義が異なると、作成される木の形状が異なることがわかります。

図 4.9 ユークリッド距離（左）とマンハッタン距離（右）による距離行列に基づく近隣結合法による木



# 第 5 章

## サンプルを個別に捉える

サンプルからデータをとって、データから意味を取り出すのが遺伝統計学の役割ですが、サンプルの取り扱いには大きく2つのやり方があります。

1つは、個々のサンプルを個別に扱う方法で、もう1つは、サンプルを集団として扱う方法です。第5章では個別に扱う方法についてを、第6章では集団として扱う方法についてを取り上げます。

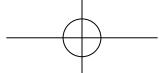
### 5.1 グラフとは

#### ● 5.1.1 グラフの定義

この章では、個々のサンプルを区別して扱います。区別して扱うために、個々のサンプルはそれぞれを点として扱います。サンプルを点で表して、それらの関係を表す方法の1つが**グラフ**です。すでに、この本でもいくつかのグラフの例が登場しています。

用語を整理する意味も込めて、グラフ理論の基礎事項をまとめます。

- 「頂点」と「辺」とで構成されたものをグラフと言います
- 頂点は点であり、辺は頂点と頂点を結んだものです
- 辺は2つの異なる頂点を結ぶ場合と、1つの頂点から出て同じ頂点に戻る場合とがあります



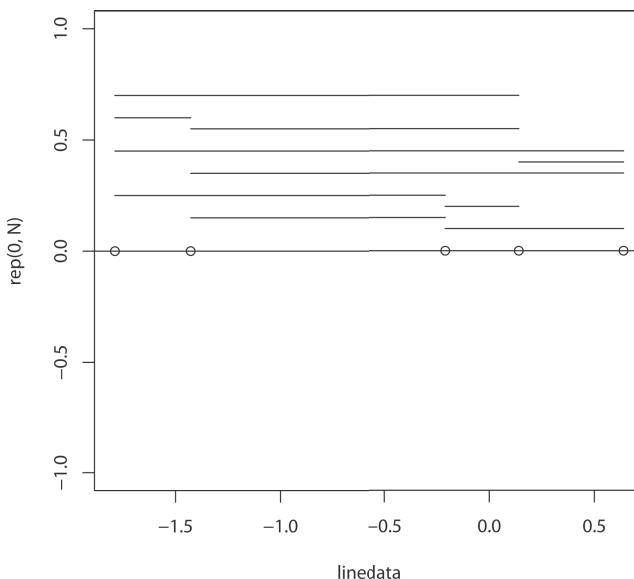
- 異なる2頂点を結ぶ辺のみを許せば、辺は頂点ペアの間の関係を表すことになります
- 辺に向きがあるとする場合と、向きがないとする場合があります
- 向きのあるグラフを有向グラフ、ないグラフを無向グラフと言います
- 辺は頂点に「接続」し、ある頂点に「接続」している辺の数をその頂点の「次数」と言います
- 辺で結ばれた頂点は互いに「隣」であると言います
- ある頂点からある頂点まで辺をたどるとき、両端の頂点と途中の頂点と辺とを合わせて「道」と言います
- 「道」がある点から元の点に戻ってくるとき、それは「サイクル」と言います
- グラフ上での2頂点間の距離は、2頂点を結ぶ最短の道の長さです
- 特徴的なグラフには名前がついています
- すべての頂点の間に辺があるグラフは、「完全グラフ」<sup>※1</sup>と呼ばれます
- すべての頂点同士に道があり、サイクルがないグラフは「木」と呼ばれます
- 関係をグラフにすることの利点の1つは、グラフを扱うアルゴリズムやグラフに関する証明済みの事項を利用できることです

## 5.2 サンプルを並べる——数直線というグラフ

順序のある連続データ型の値を持つサンプルセットがあるとします。サンプルを値に応じて数直線上に並べます（図5.1の中央の水平線とその上の点）。この並べ方はどのように読み取ればよいでしょうか。

数直線上の点のすべての点のペアを取り出して、その長さが見えるように数直線の上部に並べたのが、図5.1の中央の数直線の上部に描かれた線分です。このようにしますと、点を乗せた数直線には、 $N \times N$ の関係が距離として埋め込まれていることがわかります。このことから、量的データを用いて定めたサンプル間距離を表したグラフは、直線状のグラフとして表せることがわかります。

※1 図4.4は完全グラフの例です。



中央の数直線上に 5 点あります。それらが作る  $5 \times 4 / 2 = 10$  個のペアの距離を表す線分が数直線上に示されています。この 10 ペアの情報を埋め込んだのが、この数直線状のグラフです。

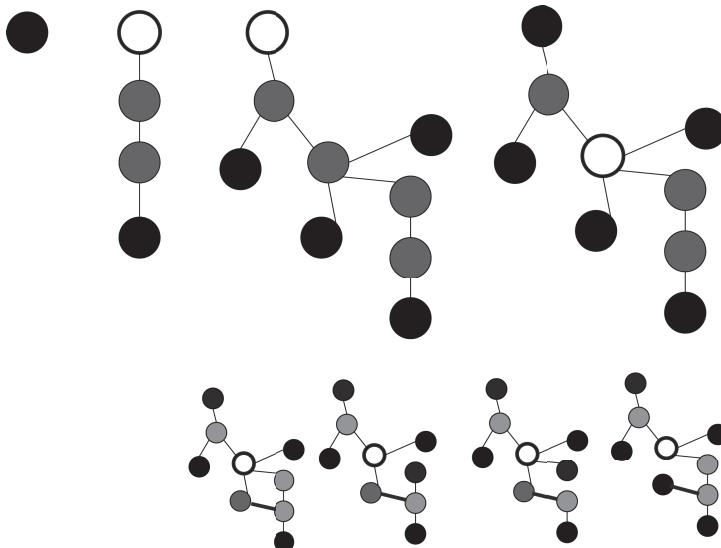
図 5.1 数直線

## 5.3 木というグラフ

### ● 5.3.1 木とは

前節の数直線の例もグラフには違いありませんが、ずいぶんと単純な形をしていました。今度は、グラフの中で**木**と呼ばれるタイプのものを扱います。

数直線も木の特別な場合です。図 5.2 を見てください。グラフで言うところの木は、植物の木と同様に、ある点から成長してできる形をとります。その成長にあたっては、先端を伸ばすか、枝分かれするかの 2 通りの方法があります。成長するときにできる辺の先は必ず新しい頂点であって、既存の頂点には接続しないので、木にはサイクルはできません。木の上の 2 点間に辺を渡すと、必ずサイクルができるので、木ではなくなります。そのサイクルを構成する 1 辺を取り去ると、再び木に戻ります。木の頂点の数と辺の数は



頂点 1 のグラフも木です。白丸が「根」、黒丸が「葉」、灰色丸がそれ以外の点です。上段の右 2 つのグラフは無向グラフとしては同じですが、「根」を取り替えて有向グラフとすると、グラフとして異なったものになります。下段の最左は、1 本の辺を 2 頂点間に差し渡して、木でなくしたグラフです。生じたサイクルの辺を 1 本取り去ると、必ず木に戻ります。

図 5.2 木

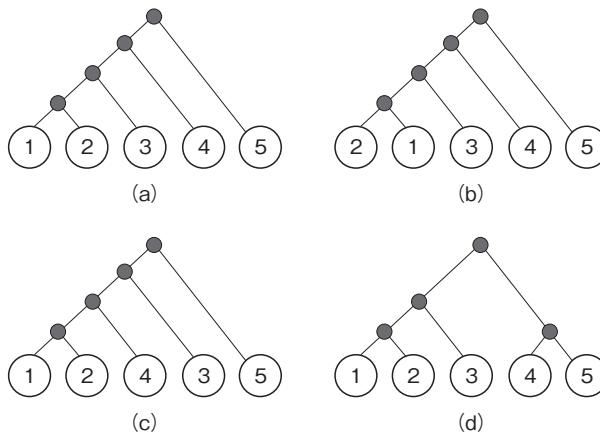
$$(T \text{ の辺の個数}) = (T \text{ の節点の個数}) - 1$$

の関係にあります。

木には「根 (root)」がある場合（「根付き木」）とない場合があります。根付き木は、木であって、その 1 つの頂点を特別扱いして、それを「根」と呼びます。根付き木は辺に向きがあり（有向）、辺の向きは「根」から離れる方向です。「根つき木」の先端には、次数 1（接続している辺の数が 1）の頂点（端末点）があり、それは「葉」と呼ばれます。

### ◎ 5.3.2 木の形状——トポロジー

木でデータを理解するためには、木の違いがわかる必要があります。木の違いは、枝分かれの具合と、分岐間の長さ・末梢の枝の長さで決まります。長さには意味を持たせずに枝分かれの具合だけに着目したとき、その形状を木のトポロジーと言います（図 5.3）。



木としてのトポロジーと分岐木としてのトポロジーは違います。分岐木は階層的分類を表しています。その場合には、トーナメント方式の試合をしている場合を考えるとわかりやすいです。トーナメント戦のやり方が変わらなければ、分岐木のトポロジーは同じです。それが変われば、分岐木のトポロジーは異なります。(b) を (a) と較べると、1, 2 のラベルの付け方は変わりましたが、トーナメント戦の対戦順は同じです。(a) と (c) は、3, 4 のラベルを付け替えました。この場合は、トーナメント戦の対戦順が変わります。(d) は、見るからに形が変わっていて、トーナメント戦の対戦順は変わります。

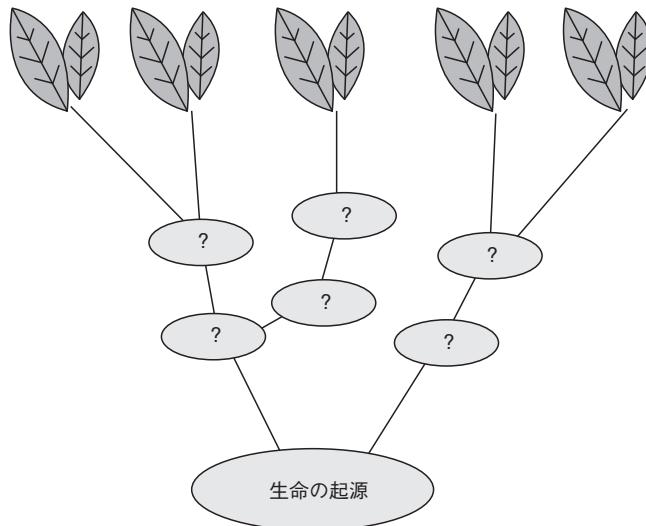
図 5.3 分岐木としてのトポロジー

## 5.4 木構造でのデータの理解——階層的クラスタリング

### ● 5.4.1 進化系統樹

種の進化の過程を樹で表したもの **系統樹** と言います。系統樹は「根付き」の木グラフで、現在の種は「葉」にあたります。

複数サンプル（今の場合は種）の外見的特長・塩基配列など、何かしらで遠近関係があるときに、その関係を木として捉えようとした例です。木の構造を、階層的クラスタ構造とも言います。枝分かれがサンプルの分類（クラスタ化）をしており、その枝分かれが、根元から末梢に向かって段階的（階層的）に繰り返されているからです。**分岐図** (cladogram) とも呼ばれます。サンプルを分類することで理解しようとする手法全体を**クラスタ解析・クラスタリング**と言いますが、階層的クラスタリングは木構造を作ってサンプルを分類する方法の1つです。



進化系統樹は生命の共通祖先から、現在の種まで木状の関係があると考えて作られるものです。現在の種は系統樹の葉に相当します。過去には、共通の祖先がいるとみなしています。

図 5.4 進化系統樹

### ● 5.4.2 階層的クラスタリング

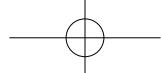
クラスタリング手法には、階層的クラスタリング以外に非階層的クラスタリングもあります。非階層的クラスタリングはサンプルを集団として扱うので、第6章で改めて取り上げます。

階層的クラスタリングは、サンプルペア間の遠近関係・類似関係から、木のトポロジーと辺の長さを答えとして出します。初めは、すべての要素を独立したクラスタとし、クラスタ同士をだんだんに合併することを繰り返すことで、最終的に1つにまとめ上げます。

この処理を実行するにあたって、原則として次の3つの決まりを定める必要があります。

- 遠近関係・類似関係の測り方
- 合併してできたクラスタと、その他のクラスタとの距離の決め方
- 合併の順序の規則

です。サンプル同士の遠近関係・類似関係の測り方については、対称的で量的



な関係であれば、距離でもよいですし、データベクトルのなす角に基づく値をとってもよいです。4.2 節（距離、相関係数）の議論に基づいて選べばよいです。

個々のサンプル同士の間の遠近関係・類似関係は選びましたが、複数サンプルが作るクラスタと、他のクラスタとの間の関係を量として表すためには、クラスタ内のどの点の情報をどのように用いるのかを決める必要があります。その方法を説明したのが図 5.5 です。

図 5.5 に示す通り、生じたクラスタとその他のサンプルや他のクラスタとの間の遠近関係・類似関係を測るにあたり、クラスタの代表点を新たに定める場合と、定めない場合があります。定める場合には、その定め方の定義が必要です。本書の目的は、個々の手法の理解ではありませんから、細かい点は不要です。手法ごとにどういう点が違うのかをつかんでください。

(b) は**近隣結合法** (NJ 法) を示しています。あくまでも、分岐図であることを意識した方法ですので、まとめる（結合する）2 点の根元側に共通な点を定め、そこから 2 点が分岐したものとして新しい点を定めます。

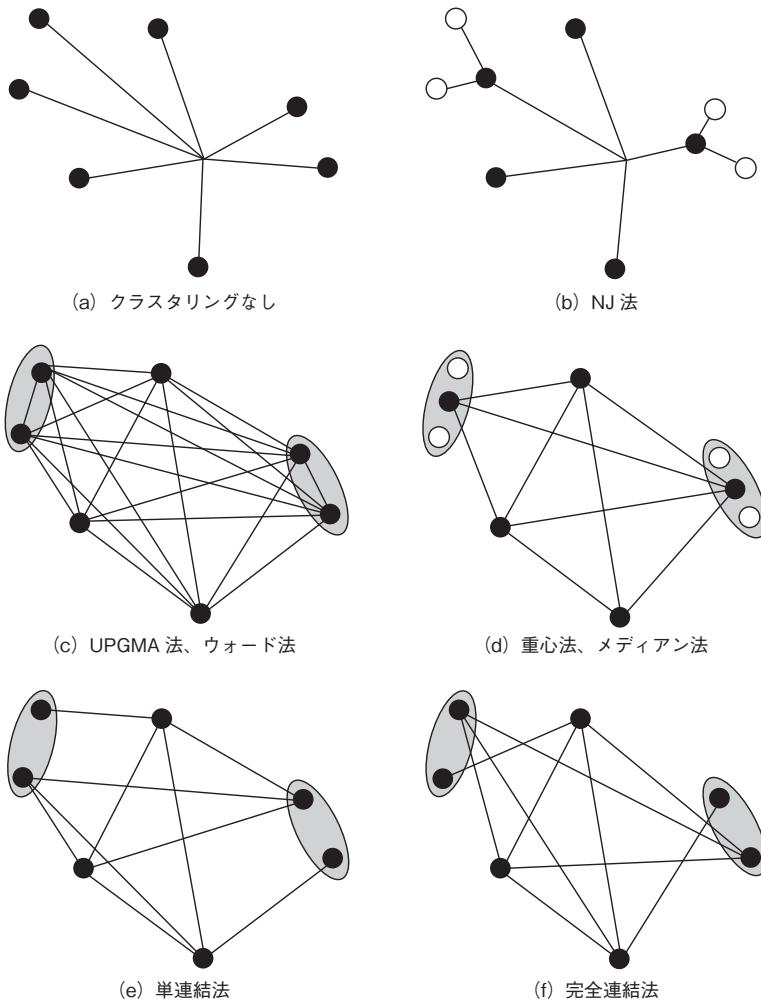
(d) は、根元に近いところに新点を置くのではなく、新たに生じたクラスタの「真ん中」にクラスタを代表する点を置きます (**重心法・メディアン法**)。この「真ん中」の定義もいくつかの定め方があります。

他方、新たな代表点を定めない場合には、クラスタが範囲を持った領域であるとみなします。このやり方では、クラスタを表す領域とクラスタ外部の点ないし領域との関係を測る方法を定める必要があります。(c) は、クラスタ化した後も、クラスタの構成要素のすべてが測定に意味を持つ方法です (**群平均法** (**UPGMA** (Unweighted Pair-Group Method using Arithmetic averages) 法)、**ウォード法**)。 (e) はクラスタとクラスタ外部の関係は最小関係を、(f) は最大関係をとるように定めます (**単連結法**、**完全連結法**)。

合併の順序のルールは、最も近いペア同士を合併していくのが素直な考え方なので、そのようにします。

このように、距離・関係の定義がいろいろあり、クラスタ間の距離・関係を決める方法にもいろいろなやり方があるので、出来上がるクラスタもいろいろです。

第 4 章で、距離行列の扱いのときには、進化・集団遺伝学系のクラスタリング法である NJ 法を用いました。ここでは、データマイニング系の R の



(a) は、7個のサンプルを空間に配置しただけで、クラスタリングしていない状態です。他の5つの図は、それぞれの方法でクラスタを進めている途中経過を示しています。

(b) (NJ法) では、クラスタの生成に伴って、クラスタを代表する点を中心寄りに定めています。クラスタとその他の点やクラスタとの関係は、この代表点によって決められます。クラスタリングされたサンプルが白丸で表され、新たに作られた代表点が黒丸で表されています。

(d) (重心法、メディアン法) でも、クラスタを代表する点が生成されます。それは、クラスタを構成する要素の分布の「真ん中」に選ばれます。中心寄りに選ぶNJ法とはここが違います。クラスタとその他の点やクラスタの関係は、この代表点によって決められます。

その他の3つの図は、クラスタリングした後に、クラスタに新たな代表点を生成しない点で共通しています。クラスタとその他の点やクラスタの関係を定める方法の違いで名称が異なります。その説明は本文を参照してください。

図 5.5 クラスタリングの方法

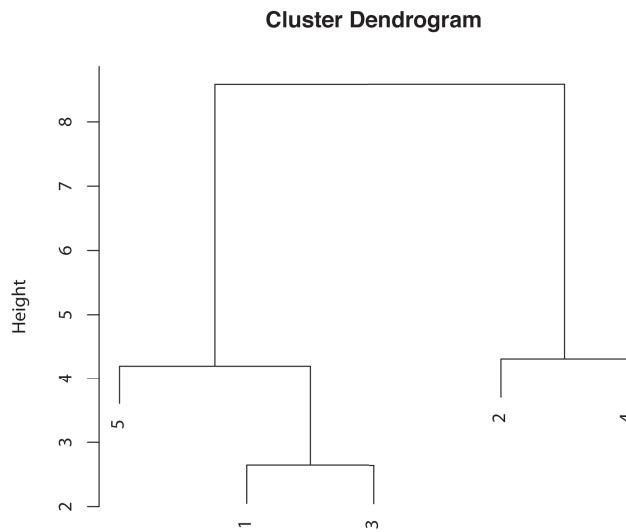
`hclust()` 関数を使ってみます<sup>※2</sup>(R5-1.R、図 5.6)。

#### R ソース 5.1 R5-1.R : クラスタリング

```

1 distMatrix<-dist(x,method="manhattan")
   # 距離にはマンハッタン距離を使用
2 trclust<-hclust(distMatrix,method="ward")
   # クラスタ間距離の定義にはウォード法を使用
3 plot(trclust)

```



```

dman
hclust(*, "complete")

```

図 5.6 クラスタリング結果

※2 `hclust()` の R のヘルプを見ることにより、複数の手法 "ward", "single", "complete", "average", "mcquitty", "median", "centroid" が利用可能であることがわかります。R のソースを確認すれば、その定義が確認できます。

## 5.5 行列データを眺める

### ◎ 5.5.1 要素を並び替えて眺める——ヒートマップ

$N \times N$  の関係を階層的クラスタリングにて木にまとめ上げましたが、 $N$  個のサンプルの間の関係が複数 ( $M$  個) の項目によって決まるときに、項目の方の関係にも興味があったらどうしたらよいでしょうか？ 両方で階層的クラスタリングをして、 $N \times M$  個のデータを表示させてみればよいでしょう。

図 5.7 は、**ヒートマップ** という図示方法です。データは  $6 \times 20$  の長方形型の行列です（2 アレル型多型が 6 箇所あり、20 本の染色体のハプロタイプを想定しています）。R5-2.R のようにしてデータを作り、ヒートマップを作る `heatmap()` 関数に処理させています。

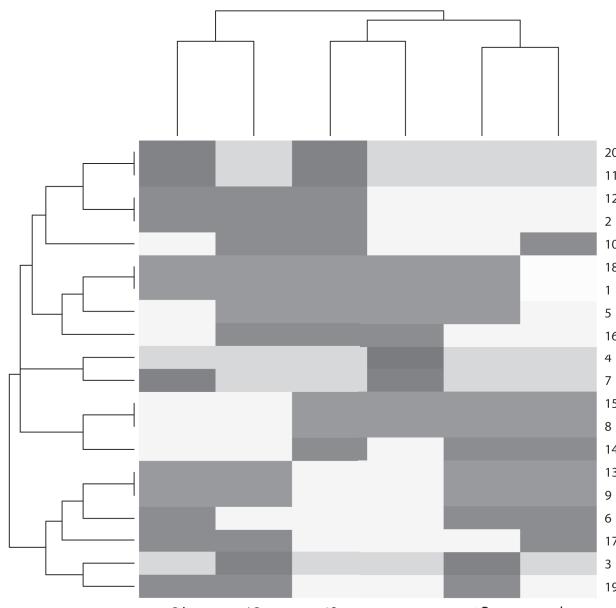
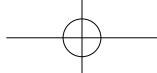


図 5.7 6 多型・20 染色体の多型データのヒートマップ



### R ソース 5.2 R5-2.R : ヒートマップ

```
1 m<-matrix(rbinom(120,1,0.5),20,6)
2 heatmap(m)
```

水平軸では 6 個の多型がクラスタリングされ、垂直軸では 20 本の染色体がクラスタリングされています。 $6 \times 20$  個のデータはその値に応じて濃淡がついています<sup>※3</sup>。

### ● 5.5.2 要素を並び替えずに眺める——連鎖不平衡係数プロット

前項のヒートマップでは  $N \times M$  行列の情報を、 $N$  側と  $M$  側の両方でクラスタリングし、 $N \times M$  をカラースケールで視覚的に表現しました。クラスタリングの結果、要素の並び順が変更されました。

要素の並び順に意味があるときは、並び順を変えずに表示します。多型マークーが染色体上に並んでいるときに、マークー間の  $N \times N$  の関係を表示してみます。マークー間の関係として連鎖不平衡係数を使って、図示することにします。連鎖不平衡の係数のうち  $r$  と呼ばれるものは相関係数でしたので、R5-3.R の処理により図 5.8 のように表示することができます。

実際には、相関係数が 1 のときと -1 のときは、片方の多型のアレルの "0", "1" のラベル付けを取り替えれば同じことです。 $r^2$  を連鎖不平衡係数としてよく使いますので、図 5.8 もそのようになっています。これは、連鎖不平衡の視覚的表現で最もよく使われるプロットです。

### R ソース 5.3 R5-3.R : 連鎖不平衡マップ

```
1 cormatrix<-cor(m);rsqmatrix<-cormatrix^2
2 image(1:nrow(rsqmatrix),1:ncol(rsqmatrix),rsqmatrix,
       col=gray((100:0)/100))
```

※3 `heatmap()` 関数は、内部で、距離行列を作る部分に `dist()` 関数を、クラスタリングに `hclust()` 関数を用いています。R のソースを確認するとわかります。

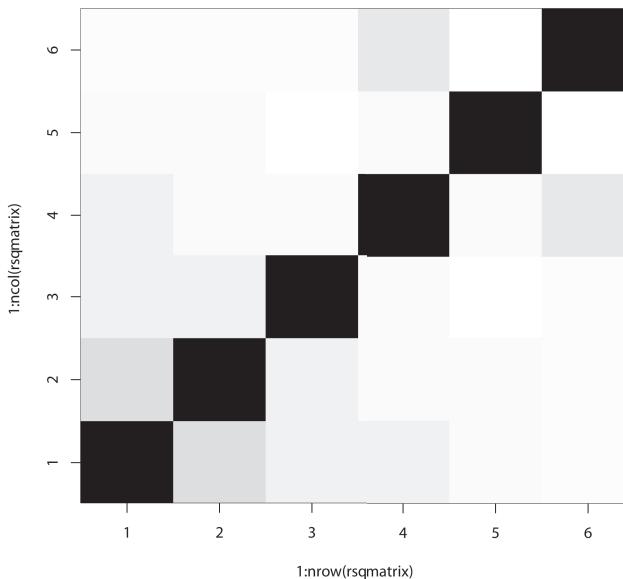
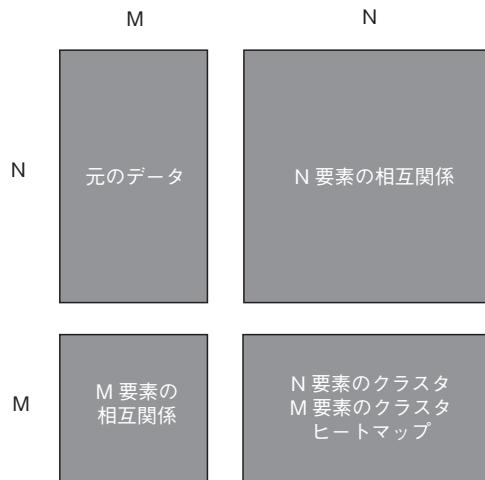


図 5.8 6 多型の連鎖不平衡マップ

### ● 5.5.3 片方の軸に着目、両方の軸に着目

もともとのデータが  $N \times M$  の形をしていたときにクラスタリングをすれば、要素の関係の強弱により、要素の順序が変わります。行についてクラスタリングを行ったり、列についてクラスタリングを行ったり、行・列の両方についてクラスタリングを行ったり（ヒートマップ）することができました。

要素順を変更せずに、 $N \times N$  の関係表示をすることもできました（連鎖不平衡係数プロット）。 $N \times N$  とは逆に  $M \times M$  で相関をとって図示することもできます。これらの関係は図 5.9 のようになっています。



2次元データそのものは左上。行の  $N$  個の要素について処理すれば、右上になります。要素の順序を変えればクラスタリングですし、変えなければ、相関パターンの図示になります。列の  $M$  個の要素についても同様です（左下）。両方でクラスタリングをするとヒートマップ表示になります。

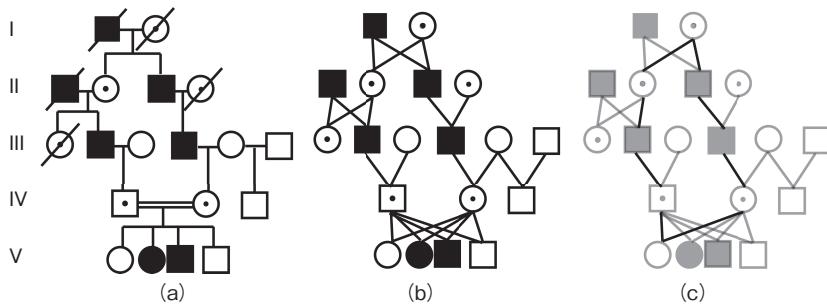
図 5.9  $N \times M$  の形のデータ

## 5.6 個体の家系図、アレルの系図——同一種の中のグラフ

親子・血縁関係を図で表した**家系図**というものがあります（図 5.10）。遺伝学では血縁関係は非常に重要ですから、家系図も大変重要です。

図 5.10 は点（丸と四角）が線で結ばれていますが、すべての線が点と点とで結ばれているわけではないのでグラフではありません。また点にフェノタイプ（発病・非発病、死亡・生存）の情報が付随し、夫婦間の線に近親婚か否かの情報があるなど、複雑な情報を搭載している点も、データ解析的には取り扱いを難しくしています。家系図をグラフとして扱うためにはどのようにすればよいでしょう。

図 5.10 (b) が血縁関係をグラフ化した図です。遺伝因子の伝達関係を表すためのグラフとしたので、夫婦の関係を表す水平線は消え、子を表す頂点を介して間接的に夫婦の関係が示されるようになりました。近親婚を表す2重線も消えました。近親婚であることは、近親婚で生まれた子から出発すると、グラ



左のⅠ、Ⅱ、Ⅲ、Ⅳ、Ⅴは世代番号です。四角は男性、丸は女性を表します。黒塗りは発病者、白抜きは非発病者です。白抜きのうち、中央に点がある場合は、遺伝因子を有する人（保因者）です。男女が水平線で結ばれているときには、生物学的な父母であることを示しています。男女の水平線から垂直に下りた線の先に子が続きます。子が複数あるときには、その線に水平線が接続し、水平線から垂直線が枝分かれしてその先に子が接続します。両親を同じくする子は、このように水平線からの枝分かれで示されます。

第Ⅲ世代の右から2番目の女性は2人の男性と水平線で接続し、それぞれの関係から子を持っています。この場合、第Ⅳ世代の右端と中央の2人は半同胞です。第Ⅳ世代の夫婦は二重線です。これは、近親婚関係を表しています。斜めの線は死亡を表しています。

個人を頂点に、親子関係を辺に変えたのが（b）です。見やすくするために死亡の斜線は省略しています。（c）は、サイクルの1つを強調した図です。

図 5.10 常染色体劣性遺伝病の家系図の例

フをぐるりと回って、本人に戻ってこられることからわかります。なお、辺に向きをつけていませんが、遺伝子の伝達は親から子へ伝わりますから、伝達関係を考慮すれば有向グラフです。木ではありません。

### ● 5.6.1 個人の関係のグラフと染色体の関係のグラフ

#### （1）個人の関係のグラフ——家系図

図 5.11（a）は常染色体劣性遺伝病の家系図です。それをグラフにしたのが図 5.11（b）です。

図 5.12 は個人が持つ染色体を三角で表して、その伝達関係をグラフにしたもので、子に伝わる染色体は、必ず親の染色体セットの両方から一部分ずつを受け継ぎます。したがって、すべての染色体は2本の辺を親の染色体から受け継ぎます。染色体数は個体数の2倍があるので、その分だけグラフが混雑していますが、ある染色体に着目して祖先の方向に向かってたどると、世代ごとに祖先染色体が2倍になる2分岐木であることがわかります。

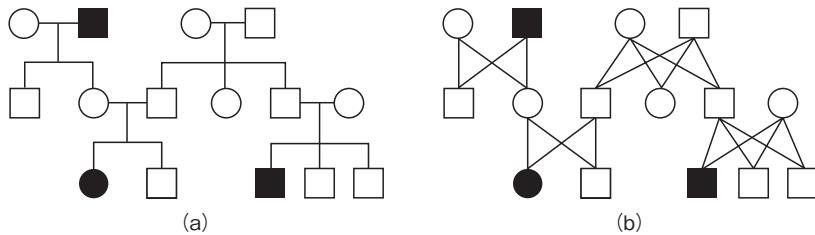


図 5.11 常染色体性劣性遺伝形式の形質の家系図とそのグラフ化表現

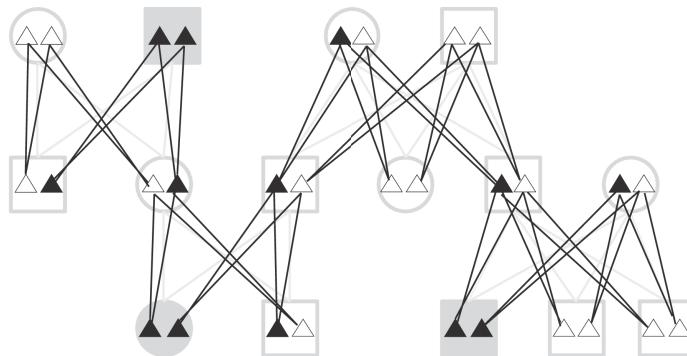


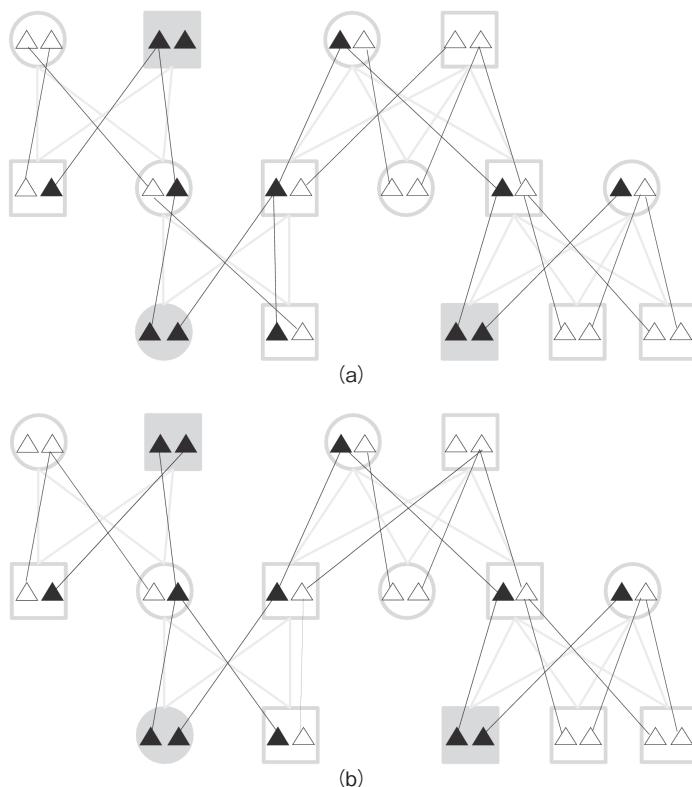
図 5.12 染色体の伝達グラフ

## (2) 個人の関係のグラフは複数の染色体の関係のグラフを含む

ある親の 2 倍体染色体から、精子もしくは卵子の 1 倍体染色体が取り出される過程を観測することはできません。したがって、その取り出しパターンは、推測するしかありません。

ある家系において、ジェノタイプを調べたとします。そのジェノタイプを満足する、染色体の伝達グラフのパターンは 1 つとは限りません。

2 アレル型多型の例を示します。図 5.13 を見てください。2 アレル型の多型のジェノタイプを観察したときに、(a) と (b) と、どちらも白黒の三角パターン（ジェノタイプパターン）は同じですが、辺の結び方（伝達パターン）が違います。遺伝因子解析では、ジェノタイプデータに基づいて、アレルがどのように伝達してきたかを推定しますが、なぜ推定しなければいけないのかをこの図は示しています。



第Ⅱ世代の左から3番目の個体が受け取る白△の染色体が、(b)では同胞と共有されているのに対して、(a)では共有されていません。また、第Ⅲ世代の左から2番目の個体の黒▲の染色体は、(a)では父親から受け取っていますが、(b)では母親から受け取っています。

図 5.13 2 アレル型多型の伝達グラフ

### ◎ 5.6.2 染色体の伝達グラフと組み換え

染色体の伝達グラフを、染色体の場所ごとに分解してみます。

図 5.14 (a) は、2組の染色体対（黒と白）から交叉が1箇所で起きて、組み換え染色体ができる様子を示しています。染色体を、3つの部分 A, B, C に分けて考えます。それぞれの部分での染色体の伝達の様子を (b) に示します。A, B, C に分けてグラフを描くと、子染色体と親染色体との関係は1対1で、枝分かれがありません。逆に、交叉・組み換えが起きた点をまたいで伝達グラフを描くと、(c) にあるように、枝分かれが生じます。交叉・組み換えが

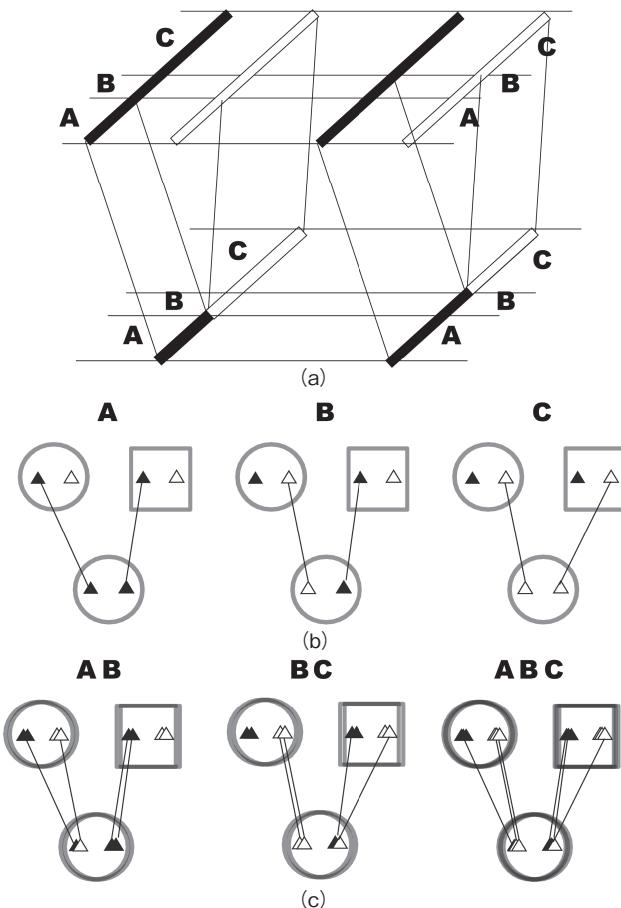
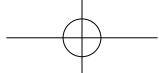


図 5.14 染色体の伝達グラフ

起きていらない範囲だけを取り出すと、何世代さかのぼっても、枝分かれのない一本道の伝達関係になります。

(c) ではAとB、BとC、AとBとCのすべてを重ね合わせました。このように重ね合わせると、染色体のグラフが分岐木になります。なお、長さ1塩



基の線分の内部で交叉・組み換えは起きませんので、塩基に関する伝達グラフは必ず枝分かれのない一本道になります。

### ◎ 5.6.3 祖先にさかのぼる——コアレセント

染色体の伝達グラフをより多くの世代にわたって考えてみることにします。染色体の集団があって、交叉・組み換えをしつつ、次世代の染色体を作る過程をシミュレーションしてみます。

図 5.15 は、8 本の染色体で 10 世代の経過を見ています。縦軸が世代経過を表しています。各世代にある 8 個の点が染色体を表しています。(a) は、ある特定の塩基についての伝達グラフです。1 塩基の伝達グラフなので一本道です。最も若い世代（最下端の世代）の 8 本の染色体の親をたどると、一番上の世代では、2 本の染色体に行き着いていることがわかります。現在の染色体の祖先染色体をさかのぼるとだんだんに同じ染色体に行き着くわけですが、これを **コアレセント**（合体・統合）と呼びます。

(b) のグラフは、(a) のグラフの塩基の近傍の塩基での伝達の様子です。木の形はよく似ていますが、(a) の方が (b) よりも濃い木が大きいことがわかります。この 2 つのグラフを重ね合わせたのが、(c) のグラフです。第 6 世代の左から 2 番目の染色体だけが、上から 2 本の辺を受けています。祖先に向かって、一本道ではなくなり、分岐が生じています。これは、2 つの塩基の間で交叉・組み換えが起きたことを示しています。

(d) のグラフでは、薄い色の木の途中に変異を表す★印が描き込まれ、その変異を引き継ぐ部分が太い木として描いてあります。この木はこの★印が「根」になります。

このように、特定の塩基に着目すると、複数の木が独立に存在して、根から枝を介する部分は過去にあり、現在は、その末端である葉として現れています。この葉を 2 つ持っているのが個人です。すべての塩基について、伝達の木を描くことができて、隣合う塩基の木々のパターンはほぼ同じで、交叉・組み換えが起きたときにのみ、パターンが変わります。塩基ごとの伝達グラフはゲノム全体にわたって存在しており、そのすべての木を重ね合わせたものが、塩基配列情報の現在と過去とその伝達に関する情報のすべてです。

個人はフェノタイプを持っています。したがって、ある変異がフェノタイプ

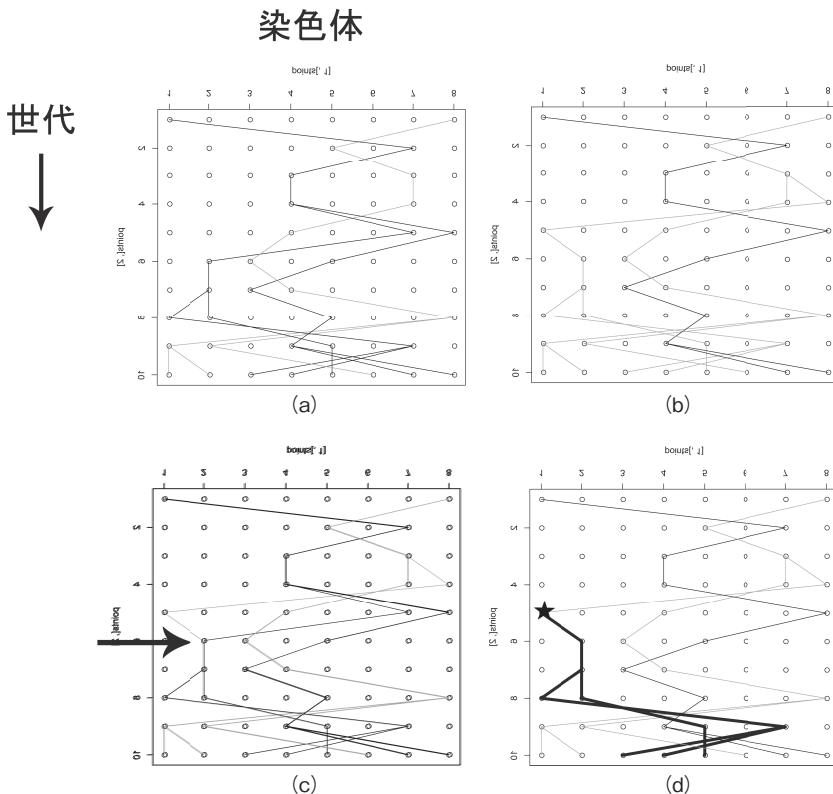
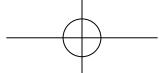


図 5.15 コアレント

と関係していることを調べる作業は、この過去にさかのぼるたくさんの木々のパターンのうちのどれが、葉に現れているフェノタイプのパターンとよく合致するかを判断することと同じです。この点は、家系サンプルを用いた連鎖解析でも、集団からのサンプルを用いた関連解析でも同じです。

この図を描くRのソースは掲載しませんが、ダウンロードできます (R5-sup1.R)。



## 5.7 ネットワーク

個々のサンプルを個別に取り扱う方法としてのグラフに関する最後のトピックとして、**ネットワーク**を取り上げます。木はサイクルを持たないグラフですが、サイクルを持ったグラフを活用することもできます。サイクルを持った有向グラフでそこに何かしらの流れを想定するとき、それはネットワークと呼ばれます。

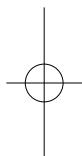
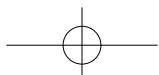
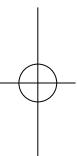
分子としてのDNA、RNA、タンパク質、その他の因子は、同時に存在して相互に影響を及ぼし合うことができます。これらの関係を表現しようとするとき、**サイクル**が生じます。AがBを促し、BがAを促すという関係は**ポジティブフィードバック**と呼ばれ、AがBを促し、BがAを抑制するという関係は**ネガティブフィードバック**と呼ばれ、どちらも単純に表せば、図5.16(a) (b) のようになります。要素を増やしてA、B、Cの3要素にしてみます。CがAに影響を与える様子を「作用している」印象を強めて図にしたのが図5.16の(c)～(f)です。

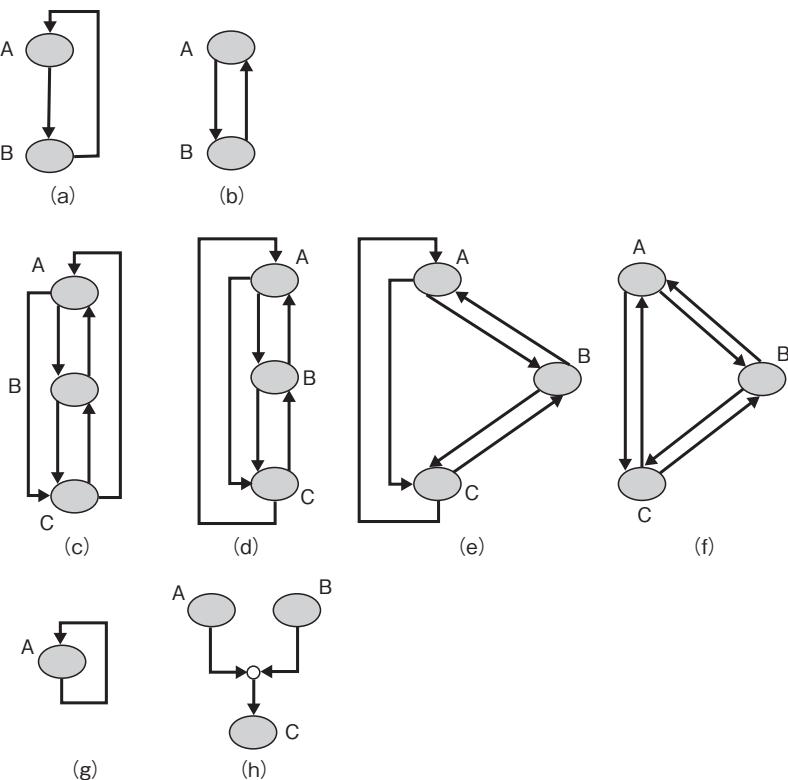
少しずつ配置を変えていくと、3頂点の間に方向の異なる2つの辺を引いた状態という意味では、変わりがないことがわかります。

要素間のネットワークの解析では、図5.16(h)のような関係も登場します。2つの要素を組み合わせたときに初めて、第3の要素への効果が生じる、という関係です。このような場合には、「2要素の組み合わせ」に対応した点をグラフに付け加えることになります。遺伝形式を取り扱うときに列を追加する必要があったことと同じことです(図3.7参照)。

要素の組み合わせを考えるときには、組み合わせの数が要素数に応じて猛烈な勢いで増えしていくためにグラフが巨大となり、全部を数え上げることは現実的でなくなることが多いです。

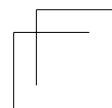
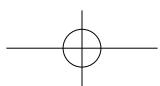
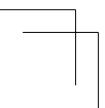
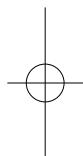
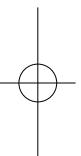
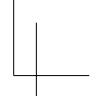
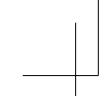
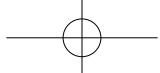
また、サイクルのあるグラフは無向にしろ有向にしろ、読み取りが難しくなります。どこから見始めて、どのように見終えたらよいかがわかりにくいいからです。ですから、サイクルのあるグラフを理解するときには、そのグラフを特徴づけたり、説明したりする指標を取り出すという作業が発生します。

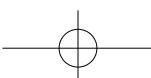
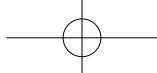




(a) がポジティブフィードバック、(b) がネガティブフィードバック、(c) ~ (f) は 3 点間の相互関係を点の配置を換えて描いたもの、(g) は自身へのフィードバック、(h) は 2 つの要素の作用に相互の影響がある場合を示しています。

図 5.16 フィードバックとグラフ





# 第 6 章

## サンプルを集団として捉える

第5章では個々のサンプルを区別しました。今度は、サンプルの集まりに着目します。

### 6.1 分布として捉える

#### ◎ 6.1.1 1次元

複数のサンプルについて1つの量的データ型のデータがあるものとします。そのデータの集まりの特徴を知るために、**箱ひげ図**と**密度分布**と**累積密度分布**を描いてみます（R6-1.R）。

R ソース 6.1 R6-1.R : 記述統計

```
1 n1<-1000;n2<-500 # サンプル数
2 # 1峰性サンプル作成
3 popdata1<-rnorm(n1,0,0.5) # 正規分布からの乱数発生
4 par(mfcol=c(1,3)) # 画面を1x3に分割
5 boxplot(popdata1) # 箱ひげ図
6 plot(ecdf(popdata1)) # 標本の累積分布
7 plot(density(popdata1)) # 密度分布
8 par(mfcol=c(1,1))
9 summary(popdata1) # データの基本統計
10 #2峰性サンプル作成
```

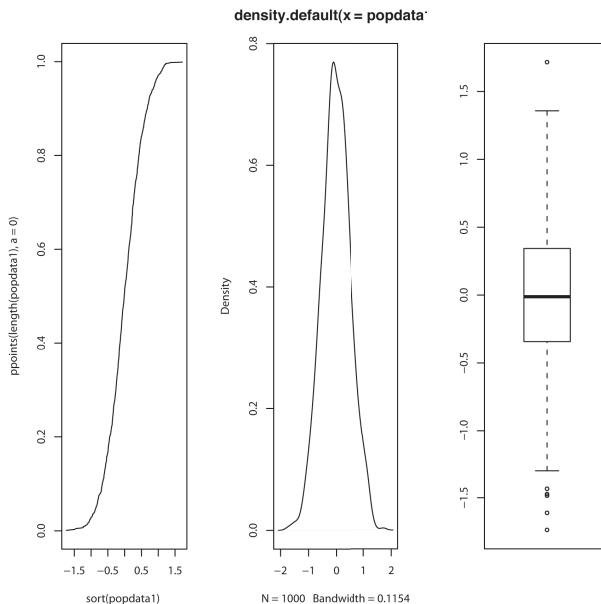
```

11 popdata2<-c(rnorm(n1,0,0.5),rnorm(n2,5,1))
12 par(mfcol=c(1,3)) # 画面を1x3に分割
13 boxplot(popdata2) # 箱ひげ図
14 plot(ecdf(popdata2)) # 標本の累積分布
15 plot(density(popdata2)) # 密度分布
16 par(mfcol=c(1,1))

```

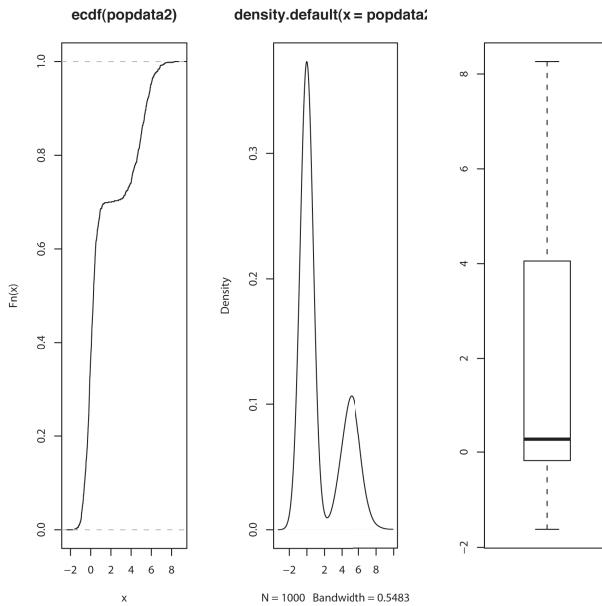
図6.1 (a) の例では、密度分布プロットから、1峰性の分布であることが読み取れます。累積分布プロットの勾配の具合からも同様のことを読み取れます。一峰性であれば、箱ひげ図やそれが表している**基本統計量**である最小値・最大値・**四分位点**・**中央値**・**平均**などの情報が、かいつまんだ**統計量**として有用です。

(b) の例は、2つの峰が並んでいて、右の峰は左の峰より背が低く、幅が広いです。累積プロットの方でも、この同じ情報を読み取れます。2峰性であることは、急峻な部分が2箇所あることからわかります。また、急峻な部分の長さが、左側で長く、右側で短いことから密度分布の峰の高さが値の小さい峰で



(a) 1峰性分布の累積分布、密度分布、箱ひげ図

図6.1 1次元データの記述統計



(b) 2峰性分布の累積分布、密度分布、箱ひげ図

図 6.1 1次元データの記述統計（つづき）

高いことがわかります。急峻な部分の左右の幅が左側で小さいことから、峰の幅は狭いことが読み取れます。このような場合には2峰性であることが最も重要なですから、1峰性を想定したときにかいづまんだ情報を提供する箱ひげ図や基本統計量の情報が役に立ちません。

### ◎ 6.1.2 2次元

データが2次元になっても事情は同じです。3峰性のデータを作って、散布図、濃淡での散図、鳥瞰図、クラスタリング図を描きます（R6-2.R、図 6.2）。

#### R ソース 6.2 R6-2.R : 2次元分布表現

```

1  # 正規乱数を用いてデータを作る
2  n1<-500;n2<-300;n3<-200;x<-c(rnorm(n1,0,0.5),rnorm(n2,5,1),
   rnorm(n3,8,2));y<-c(rnorm(n1,0,2),rnorm(n2,3,2),
   rnorm(n3,-3,1))
3  library(gregmisc) # hist2d() を持つパッケージ
4  h2d <- hist2d(x,y, show=FALSE,same.scale=TRUE,

```

```
nbins=c(10,10)) # 2次元ヒストグラム情報を取る  
5 plot(x,y) # 散布図  
6 filled.contour( h2d$x, h2d$y, h2d$counts, nlevels=9,  
    col=gray((8:0)/8) ) # 2次元ヒストグラムを濃淡で  
7 persp( h2d$x, h2d$y, h2d$counts,ticktype="detailed",  
    theta=60, phi=30,shade=0.5, col="cyan")  
# 2次元ヒストグラムを鳥瞰図で
```

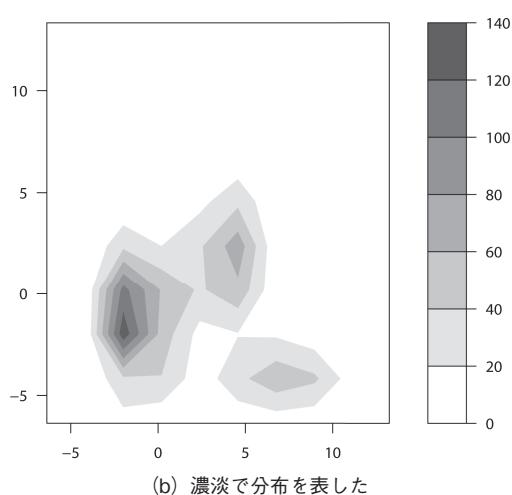
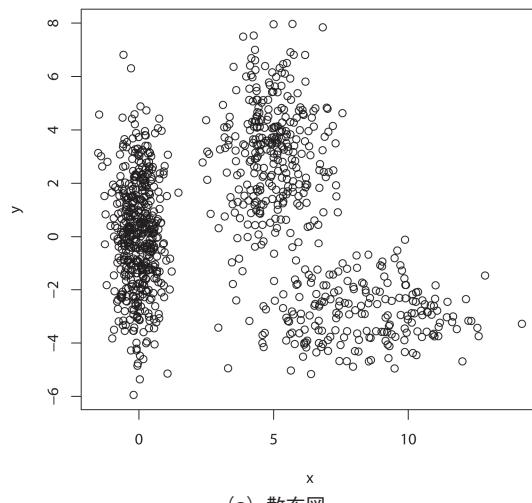
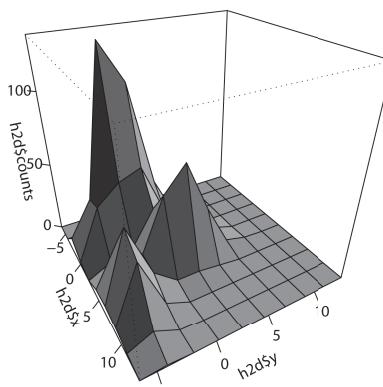
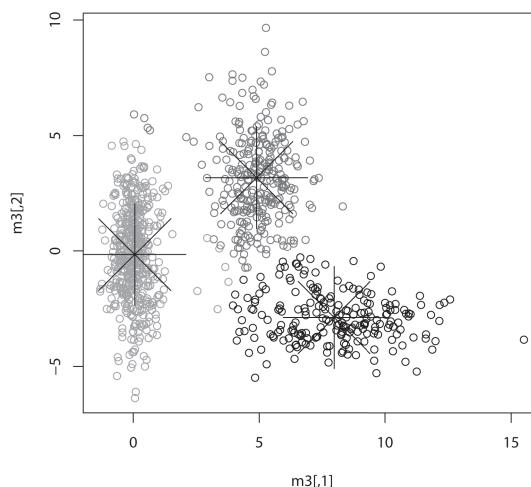


図 6.2 2次元分布表現



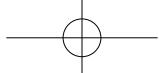
(c) 密度を高さにとった鳥瞰図



(d) クラスタリングをして色分けをした散布図 (R6-3.R によるグラフ)

図 6.2 2 次元分布表現（つづき）

分布を2次元平面にプロットしたり、それを濃淡や高低鳥瞰図にしたりして、分布の特徴を捉えます。1峰性であれば、その峰の特徴を捉える値でかいづまることになります。多峰性であれば、いくつの峰があるか、峰同士の相互位置関係はどうなっているか、個々の峰の特徴はどうか、ということをデータから読み取る必要があります。この例では、3峰性です。次元が3以上になると、図にするのが難しくなりますが、考え方は同じです。



## 6.2 非階層的クラスタリング

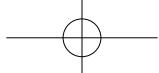
6.1.2 項で示した2次元に広がっているサンプルが3つの峰を作っているようなので、**非階層的クラスタリング**手法でサンプルを3つの峰に帰属させてみることにします。

代表的な手法である**k-means 法**では、まずクラスタ数  $k$  を与え、サンプルを  $k$  個のクラスタのうちのどれかに暫定的に帰属させます。次にクラスタの暫定的構成サンプルを用いて、クラスタの中心を定めます。クラスタの中心が定まつたら、個々のサンプルの位置と、クラスタとの位置関係から、サンプルが帰属するべきクラスタを変更します。こうすることで、クラスタの構成サンプルが変わります。構成サンプルが変わるとクラスタの中心が動くので、中心を更新します。このように適当にクラスタ（峰）の中心を定め、サンプルを  $k$  のクラスタに帰属させつつ、帰属の変更を繰り返すことでその中心を移動させていきます。最終的にサンプルの帰属クラスタが変わらなくなるまで続け、その状態を答えとする方法です。

R6-3.R によるグラフが図 6.2 (d) です。

R ソース 6.3 R6-3.R : 非階層的クラスタリング

```
1 # 非階層的クラスタリング
2 m3<-matrix(c(x,y),ncol=2)
3 cl <- kmeans(m3, 3) # kmeans 法で3群にクラスタリング
4 plot(m3, col = cl$cluster)
5 points(cl$centers,pch = 8, cex=10) # クラスタの中央に印をつける
```



## 6.3 集団遺伝学

生物個体が集まって集団を構成します。この個体の集団の様子を取り扱う遺伝学の一分野として**集団遺伝学**という独立した分野があります。いくつかのトピックスを取り上げて、その内容について触れていきます。

### ● 6.3.1 不均一と不平衡

ある一定以上の個人が構成する集団で出会いの機会が完全にランダムであることは考えにくく、実際には、偏りが生じます。偏る要因としては、物理的に遠いところに存在する場合や、往来を制約する地理的な要因（山・川・海）などの自然の要因もあれば、言語・宗教・民族・国家などの人的要因もあります。隔たりがあっても、完全に隔絶されていなければ、徐々に混ざり合って、隔たりがなかった場合と同じような状態に向かいます。また、人口が増加していく過程では、構成メンバーが十分に混じり合う暇がないので、人口の増加も不均一化の要因になります。ですから、不均一な集団を、均一になるまでの時間的途中経過の状態と捉えることもできます。

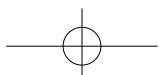
### ● 6.3.2 均一な集団とハーディ・ワインバーグ平衡 (HWE) ——均一な集団の混合

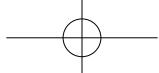
集団内のメイティング<sup>※1</sup>がランダムであるときに生じる平衡が HWE です。現実には、平衡に達していないことが多いです。しかしながら、不均一な状態をそのまま取り扱うことは面倒なことが多く、また理解もしにくいので、よく行われるのは、HWE を満足する集団が複数存在し、それが混ざったものとして、不均一な状態を表そうとする方法です。

最も基本的な場合として、2 集団の混合を考えます。

2 つの集団があり、2 アレル型多型  $A, a$  について、それぞれ、 $A$  のアレル頻度が  $p, q$  であり、集団内部では HWE であるときに、この 2 集団が  $c, (1-c)$  の比率で混合しているときを考えます。

※1 メイティングとは、雌雄がペアを作ること、または、ペアを作つて子を成すこと。





第1集団ではジェノタイプ頻度が、

$$p^2, 2p(1-p), (1-p)^2$$

で、第2集団では

$$q^2, 2q(1-q), (1-q)^2$$

です。混合集団では、

$$(gm_1, gm_2, gm_3) = (cp^2 + (1-c)q^2, 2cp(1-p) + 2(1-c)q(1-q), c(1-p)^2 + (1-c)(1-q)^2)$$

となります。 $A$ に値1、 $a$ に値0を与えると、それぞれの集団での平均と分散は、

$$m_1 = 2p, v_1 = 2p(1-p)$$

$$m_2 = 2q, v_2 = 2q(1-q)$$

です。混合集団でのアレル頻度は  $cp + (1-c)q$  と構成比率に比例した値になり、ジェノタイプの値の平均もその2倍で、

$$mm = 2(cp + (1-c)q) = 2(cm_1 + (1-c)m_2)$$

ですが、分散の方はそれほど簡単にはいきません。第3章で見たように、**共分散**の項が入ってくるからです。

このように複数の均一な集団が混合した集団は、構造のある集団、**構造化した集団**と呼びます。

### ● 6.3.3 時間的な変化

#### (1) 拡散方程式

今度は時間経過を考えます。

今、2つの島  $A, B$  があって、2島はそれぞれ隔離していたとします。ある多型について、 $A$  はアレル M ばかり、 $B$  はアレル m ばかりだったところにその島が陸続きになった（地殻変動でも橋がかけられたでもなんでもよいのですが）とします。 $A, B$  の人口は  $P_A, P_B$  で変わらず、両島の間では、単位時間あ

たり  $d$  の人が  $A$  から  $B$  へ移住し、逆に  $B$  から  $A$  へも同じ人数が移住するとします。 $A, B$  のアレル M の頻度を時間  $t$  の関数で表し、 $Fa(t), Fb(t)$  とし、陸続きになったときを  $t=0$  とします。 $t \leq 0$  のとき  $Fa(t)=1, Fb(t)=0$  です。ここで、 $t=T$  から  $t=T+\delta$  への変化を考えます。 $A, B$  それぞれについて、 $t=T$  のときのアレル M の数のうち、移住しない分から、移住して出て行く分を引き、逆に移住して入ってくる分を足しますから、

$$\begin{aligned} 2P_A Fa(T+\delta) &= 2P_A Fa(T) - 2d\delta Fa(T) + 2d\delta Fb(T) \\ 2P_B Fb(T+\delta) &= 2P_B Fb(T) - 2d\delta Fb(T) + 2d\delta Fa(T) \end{aligned}$$

という関係にあります。2つの式にそれぞれ  $P_B, P_A$  を掛けて引くと、

$$\begin{aligned} 2P_A P_B (Fa(T+\delta) - Fb(T+\delta)) \\ = 2P_A P_B (Fa(T) - Fb(T)) - 4d\delta(P_A + P_B)(Fa(T) - Fb(T)) \end{aligned}$$

となり、

$$\begin{aligned} (Fa(T+\delta) - Fb(T+\delta)) - (Fa(T) - Fb(T)) = \\ - \frac{2(P_A + P_B)}{P_A P_B} (Fa(T) - Fb(T)) \end{aligned}$$

となります。ここで、 $G(T) = Fa(T) - Fb(T)$  とおけば、

$$G(T+\delta) - G(T) = -\frac{2(P_A + P_B)}{P_A P_B} G(T)$$

です。ここから、

$$\frac{d}{dt} G(T) = -\frac{2(P_A + P_B)}{P_A P_B} G(T)$$

なので、

$$G(T) = K e^{-\frac{2(P_A + P_B)}{P_A P_B} t}$$

今  $G(0) = 1$  ですから  $K = 1$  です。 $G(t)$  の微分方程式は、アレル頻度の差の変化量が、アレル頻度の差に比例していることを示した式です。

$$P_A Fa(t) + P_B Fb(t) = P_A Fa(0) + p_B Fb(0) = P_A$$

なので、

$$Fa(t) = \frac{P_A + P_B e^{-\frac{2(P_A+P_B)t}{P_A P_B}}}{P_A + P_B}$$

$$Fb(t) = \frac{P_A (1 - e^{-\frac{2(P_A+P_B)t}{P_A P_B}})}{P_A + P_B}$$

となります。

このアレル頻度の変化の様子を R で描いてみます (R6-4.R、図 6.3)。

#### R ソース 6.4 R6-4.R : 移住

```

1 pa<-9000;pb<-1000;d<-100;t<-0:100
    # pa,pb:2集団の人口,d:単位時間あたりの移住人数,t:世代
2 fa<- (pa+pb*exp(-2*d*(pa+pb)/(pa*pb)*t))/(pa+pb)
3 fb<- (pa*(1-exp(-2*d*(pa+pb)/(pa*pb)*t)))/(pa+pb)
4 plot(t,fa,ylim=c(0,1),type="l")
5 par(new=T)
6 plot(t,fb,ylim=c(0,1),type="l")

```

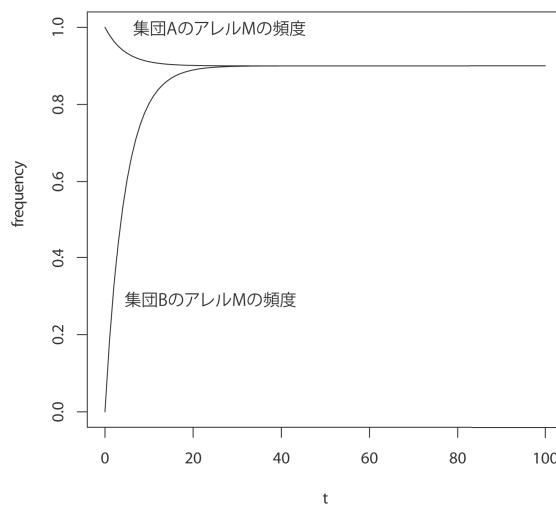


図 6.3 時間とともに 2 地域のアレル頻度の差が縮まって、最後には同じ値に収束

時間（横軸）とともに、2島のアレル頻度は同じ値に収束します。収束する頻度は、移住開始前に2島を合わせた全員に占めるアレルの頻度です。また、収束までの時間は、移住の人数が島人数に占める割合によります。2島が隔絶していた状態から、交流のある状態に変わったことで、アレル頻度が変化しました。時間が十分にあれば、移住のスピードによらず、あるアレル頻度に落ち着いて、変化しなくなります。この状態が平衡状態です。

このときのアレル頻度の変化の量が、アレル頻度に比例する形の微分の式で出てきました。これは、熱力学で**拡散方程式**と呼ばれる形式の式の最も単純なものです。

## (2) 推移行列でも表す

島の数を増やします。島が5つあり、それぞれの島から他の島へと移住をするとします。移住する人数は、島の人口の一定の割合で、移住先を選ぶ割合も変化しないとすると、A島から、B, C, D, E島へ移住する人数は、A島の人口  $P_A$  を用いて、

$$P_A * m_{A \rightarrow X}; X = B, C, D, E$$

と表されます。移住しない割合を

$$m_{A \rightarrow A} = 1 - \sum_{X \in \{B, C, D, E\}} m_{A \rightarrow X}$$

と表せば、

$$\begin{aligned} P_A(T+1) &= P_A(T)m_{A \rightarrow A} + \sum_{X \in \{B, C, D, E\}} P_X m_{X \rightarrow A} \\ &= \sum_{X \in \{A, B, C, D, E\}} P_X m_{X \rightarrow A} \end{aligned}$$

となります。これは、推移を表す行列を用いて、

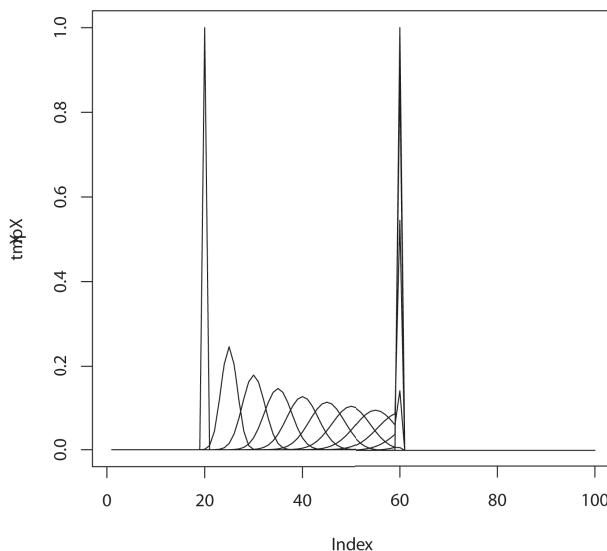
$$\begin{aligned} &\begin{pmatrix} m_{A \rightarrow A} & m_{B \rightarrow A} & m_{C \rightarrow A} & m_{D \rightarrow A} & m_{E \rightarrow A} \\ m_{A \rightarrow B} & m_{B \rightarrow B} & m_{C \rightarrow B} & m_{D \rightarrow B} & m_{E \rightarrow B} \\ m_{A \rightarrow C} & m_{B \rightarrow C} & m_{C \rightarrow C} & m_{D \rightarrow C} & m_{E \rightarrow C} \\ m_{A \rightarrow D} & m_{B \rightarrow D} & m_{C \rightarrow D} & m_{D \rightarrow D} & m_{E \rightarrow D} \\ m_{A \rightarrow E} & m_{B \rightarrow E} & m_{C \rightarrow E} & m_{D \rightarrow E} & m_{E \rightarrow E} \end{pmatrix} \begin{pmatrix} f_A(T) \\ f_B(T) \\ f_C(T) \\ f_D(T) \\ f_E(T) \end{pmatrix} \\ &= \begin{pmatrix} f_A(T+1) & f_B(T+1) & f_C(T+1) & f_D(T+1) & f_E(T+1) \end{pmatrix} \end{aligned}$$

のように表されます。これは、第2章の遺伝的浮動のときのやり方と同じです。

### ◎ 6.3.4 空間の移動

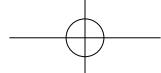
6.3.3項の例では、個体が存在する位置は帰属する島以外には、特に気にしませんでした。全員が島の1箇所に存在しているように扱ったとも言えます。今度は、ある個体が存在する位置にも着目します。簡単のために、1次元空間（直線上）に存在するとします。そして、ある因子を持っていると、その空間において生活しやすさが異なるために、より生活しやすい場所に移動するというような場合を考えます。そして、その因子を持っているときには、より快適な方へ移動したがる傾向があるとします。今、このような因子を持つ集団がある位置に集中的に存在していたとすると、少しづつ快適な方向へ移動し、最終的には快適なところに集結します。

このような空間中での動きも拡散方程式の枠組みです。図6.4に示すように、集団はある場所に集中していましたが、移動を開始します。移動中は存在範囲が広めになり、ゴールに到着した個体はそこから動かないで、最終的にすべての個体がゴールに終結して、動きが止まります。



左のピークに集中していた集団は、時間とともに条件のよい右へと動きます。分布は左から右に向かって、時刻の遅い場合を表しています。

図6.4 空間の移動



この図を描く R のソースは掲載しませんが、ダウンロードできます (R6-sup1.R)。



## 6.4 热力学、统计力学、流体力学

### ● 6.4.1 時空間、有限と無限

集団内の因子の頻度が時間の経過について**平衡**状態に達することや、因子を持つ個体が空間内を時間とともに移動する様子を見てきました。**拡散方程式**も登場しました。時空間を定める変数の関数として扱ったわけです。ここで、空間というものが出てきましたが、これはわれわれの住んでいる 3 次元空間に限るものではありません。地球の表面を空間にするならば、それは、球面が空間です。球面の特徴は、2 次元の面としての特徴を持つつ、有限な面積を持っているのに、果てに行き着くということがないことです。

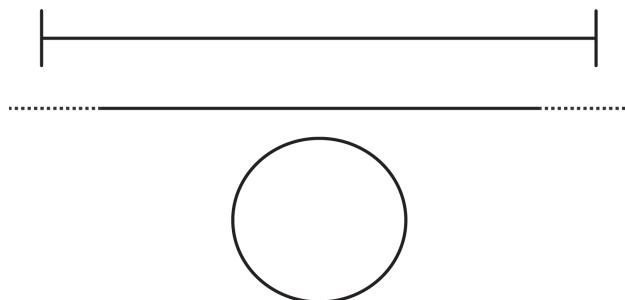
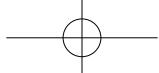
生物個体が存在する時空間だけについて考えることはありません。多変量のデータは項目の数だけ次元を持たせることができます。非常に高次元な空間に存在します。データが作る時空間も、生物個体が生きる時空間も、時間の軸だけは一方向性の唯一の軸とするのが適当かもしれません。データの時間以外の空間の次元は自在です。空間の定め方はさまざまですが、定め方によらず共通して注意する点があるとすれば、次のような点が挙げられます。

- 空間は**有限**か、**無限**か
- 空間は閉じているか、閉じていないか

この 2 つは似ているようですが、少し違います。先ほど挙げた地表面という空間は、面積を計算できることからも「有限」です。しかしながら、どこまでも進んでいくことができて、果てはありません。ぐるっと回って元に戻ってくるだけです（図 6.5）。

何かをモデル化するときには、無限を仮定することが多いです。数理的取り扱いが簡単になるからです。

たとえば、人口の増加について、いくらでも増えられるように仮定することがよくあります。それは、比較的少数の場合には現実とよく当てはまります



上段は果てがあって有限の長さの線。中段は果てがなく、無限の長さの線。下段は果てがなく、有限の長さの線。

図 6.5 有限な線と無限の線、果てがある線と果てのない線

が、ある程度以上になると増加に制約が出てその影響が無視できないことはよくあります。化学反応の開始初期には無限の仮定が当てはまるものの、反応産物の量が増えてくると頭打ちになる場合も同様で、**飽和現象**と呼ばれます。

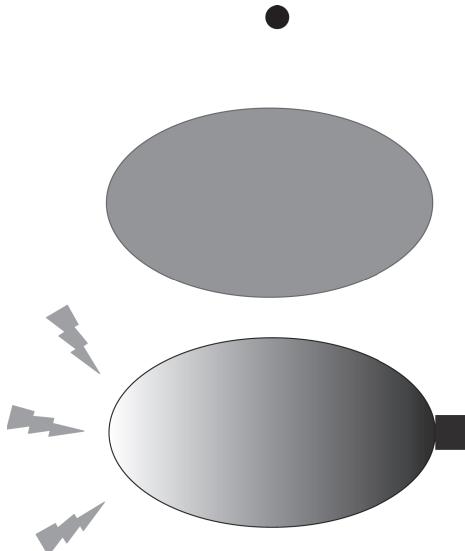
変異による遺伝子多型箇所の増大についても、同じことが言えます。変異の発生確率は非常に低く、DNAは非常に長いので、どこも平等に無限に変異が起きると仮定することができます。この仮定が有効な場合も多いですが、実際には、飽和の影響を考えなくてはならない場合も出てきます。

### ◎ 6.4.2 均一、平衡、定常

時空間の次元の広さについて述べましたが、もう一度、時間も空間も広がりがない状態に話を戻します。**物理学**や**熱力学**的な視点で考えます。

時間も空間も広がりを持たない世界では、すべてのものは同じ場所に存在し、すべてのことは一瞬で起きるのと同じことです。HWEを考える際に、すべての染色体はお互いに同じ確率で出会って、ペアを作ると仮定しましたが、これは、すべてのものが同じ場所に存在しているとみなしたとも言えます。また、平衡状態に達するには時間がかかりましたが、すべてのことが一瞬で起きるとすれば、一瞬で平衡状態に達しますから、HWEを考えたときは、時間もないものとしたと考えることもできます。

時間には広がりを持たせない今まで、空間に広がりを持たせるとします。そうすると、ものが空間に分布します。分布すると濃い場所と薄い場所ができます。濃度の分布や、密度の分布です。非常に多くの構成要素があって、それに



点には大きさがなく、内部構造もありません。連続体には広がりがあります。内部は平衡に達することもありますが、外部との交渉により内部にばらつきを持った状態で安定することもあります。

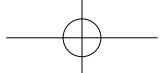
図 6.6 点と連続体

切れ目はないけれども、空間的に偏りがあるようなものとして、**連続体**という考え方があります（図 6.6）。物理・化学では、**流体**という捉え方があります。流体には気体、液体の両方が含まれますが、液体は気体よりも個々の分子の動きが制限されているので、液体をイメージする方がよいです。

このように、空間に広がりを持たせて解釈するときには、連続体とか流体とかに関する概念やモデルを用いることができる事がわかります。この視点から、先ほどの空間の広がりのない状態を見直すと、これは**質点系**で考えていたことに相当することがわかります。

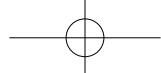
しかしながら、時間の広がりがないとすると、すべてのことは一瞬で起きてしまいますから、安定した状態に達するのも一瞬です。ですから、このような時空間では、ある一定の状態をとります。特に条件がなければ、内部は均一になるでしょう。構成要素自体は絶えず変化しているかもしれません、全体で見たとき変化がないように見える状態のことで、**平衡状態**と呼びます。

もし、ある塊のある場所を熱し、それ以外のところで放熱すると、いつしか熱している近くは熱く、そこから遠いところはそれほどでもない状態に落ち着



くでしょう。これも、落ち着いた一定状態です。外と何かしらのやりとりをした常態での安定状態で、平衡状態とは区別して**定常**状態と呼ぶこともあります。これは、時間に関して不变である場合ですが、もう少し時間について大きな目で見ることもできます。個々の構成要素は変化しながらも、全体として時間的に無変化な状態は、生命現象の一部しか説明していません。個体に関して言えば、日周期で起きる現象もあれば、年単位での変化もあります。その他にも、長短さまざまな周期で変化する現象も多いです。これらについては、周期を持って一定の状態変化を繰り返しているわけですが、周期的変化という安定状態にありますから、時間を周期単位まで大きくしてやれば、定常状態です。

これらの考え方は、熱力学の枠組みですが、熱力学の考え方は個々の区別をせずに全体を捉えるときの道具立てです。熱力学に個々の要素への考慮をして拡大したのが**統計力学**です。また、安定した状態に焦点を当ててきましたが、生物現象は、逆に、思いもよらない状態変化を対象にすることもあります。熱力学・統計力学では、**非線形熱力学カオス**理論などが、不安定な変化を理論立てており、それらも、生命現象の取り扱いにあっては道具立てとして有効です。



第7章 尺度、変数、自由度、次元

第8章 統計量、指数、確率、尤度

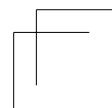
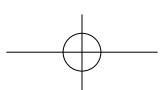
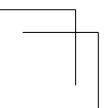
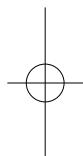
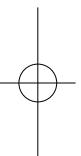
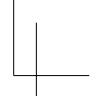
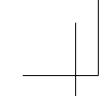
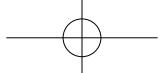
第9章 確率と尤度

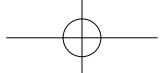
第10章 連鎖解析に見る尤度と変数

第11章 指数（インデックス）とは

第Ⅱ部において、データとは何か、データの型とは何か、データを持つサンプルをどのように取り扱うために何を考慮するべきかについての概念を論じてきました。第Ⅲ部では、実際にデータを取り扱うために必要となる概念を導入します。







# 第 7 章

## 尺度、変数、自由度、次元

### 7.1 データをかいづまんで伝える

#### ● 7.1.1 分割表の情報をかいづまんで伝える

##### (1) 情報を伝えるために必要な数値の数

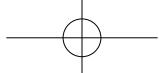
今、ある医療機関で治療薬の選択に関する集計をとることになったとします。ある病気に対して、3種類の薬 ( $x_1, x_2, x_3$ ) を処方することが可能であり、この医療機関では2人の医師 ( $d_1, d_2$ ) が診療をしています。集計の結果、以下の表のような結果だったそうです。

	$x_1$	$x_2$	$x_3$	計
$d_1$	25	23	12	60
$d_2$	15	17	8	40
計	40	40	20	100

この結果を報告する方法について考えてみます。2人の医師が処方している人数を薬別に6個の数値 (25, 23, 12, 15, 17, 8) を用いて

「 $x_1, x_2, x_3$  を  $d_1$  医師は 25, 23, 12、 $d_2$  医師は 15, 17, 8 人ずつ処方していました」

と報告することができます。



別のやり方でも報告してみます。医師別の人數と薬別の人數と総人數だけを報告するのです。表で書けば以下のような具合です。■は「伏せ字」と読みます。あえて伝えない数値です。

	$x_1$	$x_2$	$x_3$	計
$d_1$	■	■	■	60
$d_2$	■	■	■	40
計	40	40	20	100

医師別・薬別の細かい内訳はともかく、医師別の人數、薬ごとの人數についての報告になっています。これは分割表の周辺度数を報告したことになります。

今、これだけを報告されたときには、医師と薬とには特に関係がないと考えて、

	$x_1$	$x_2$	$x_3$	計
$d_1$	$60 \times 40/100 = 24$	$60 \times 40/100 = 24$	$60 \times 20/100 = 12$	60
$d_2$	$40 \times 40/100 = 16$	$40 \times 40/100 = 16$	$40 \times 20/100 = 8$	40
計	40	40	20	100

だと思うこととします。これを（無関係・独立の仮定のもとで）期待する、と言います。これが期待値の表です。医師別・薬別の内訳を詳しく報告することにすれば、この期待値からのずれを使って、次のように6個のセルのうち2個のセルについて、期待値との差を報告すると、伏せてあるセルの値も計算によって知ることができます。この方法を使えば、すべてのセルの値を知らせることができます。実際には、次の表で伏せていない数値だけを伝えればよいので、6個の値がすべての情報を伝えていることになります。

	$x_1$	$x_2$	$x_3$	計
$d_1$	+1	-1	■	60
$d_2$	■	■	■	■
計	40	40	■	100

## (2) 変数セットの間の関係

6個の数値を用いた2つの報告方法の数値の間の関係は次のようになります。

	$x_1$	$x_2$	$x_3$	計
$d_1$	$\delta_{11}$	$\delta_{12}$	■	$n_{1..}$
$d_2$	■	■	■	■
計	$n_{..1}$	$n_{..2}$	■	$n_{..}$

	$x_1$	$x_2$	$x_3$
$d_1$	$n_{11}$	$n_{12}$	$n_{13}$
$d_2$	$n_{21}$	$n_{22}$	$n_{23}$

$$n_{..} = \sum_{i=1}^2 \sum_{j=1}^3 n_{ij}$$

$$n_{1..} = \sum_{j=1}^3 n_{1j}, n_{..1} = \sum_{i=1}^2 n_{i1}, n_{..2} = \sum_{i=1}^2 n_{i2}$$

$$\delta_{11} = n_{11} - \frac{n_{1..} \times n_{..1}}{n_{..}}, \delta_{12} = n_{12} - \frac{n_{1..} \times n_{..2}}{n_{..}}$$

2つの報告の仕方のそれぞれが6変数を持ち、それらが6個の連立方程式で関係づけられていることが見てとれます。つまり変数の付け替え作業をしていくだけなわけです。

### (3) 大まかに伝えるか、細かく伝えるか

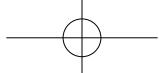
分割表の情報をすべて伝えるとすれば、変数セットのとり方によらず、必要な変数の数は同じでした。さて、本当に「すべて」を伝えることがよいことでしょうか。今、観察表は期待値表とほとんど一緒です。こんな場合には、

「全部で100人処方され、 $x_1, x_2, x_3$ の処方内訳は、40, 40, ■で、 $d_1, d_2$ の担当内訳は60, ■で、期待値表と大差ありませんでした」

と報告してもかなり正確な報告になります。6個の変数のすべてを報告する代わりに、意味の大きな4個の変数だけで報告をまとめることの方が適切です。変数セットを変えることは、少ない変数で情報の大変な部分を取り出すための工夫であるということがわかります。

### (4) 大まかな報告の定量化と自由度

さて、「大差ありませんでした」という報告の言葉に着目します。「大差(ない)」とか「大いに違(う)」という言葉は人によって感じ方が違うかもしれないのに、この部分を誤解のないような数値に変えることを考えます。期待値表と観測表とがどれくらい違うかを数値で表しましょう。それぞれの表は6個の数値でできていますので、それを使った「距離」を使うとよさそうです。その



違いは  $\delta_{11}$ ,  $\delta_{12}$  の 2 変数のとり方で決まるので、距離は 2 変数の関数です。変数の数が多ければ多いほど距離は長くなる機会が増えるので、距離が長いか短いかの判断はいくつの変数に由来するかを勘案して決めるのが適当です。

今、距離としてカイ自乗統計量と呼ばれる値を用いることにはすれば、その値は変数の数に応じて評価してやることが適当だということになります。これが、カイ自乗統計量の大小を評価するときに、変数の数（自由度）を考慮して  $p$  値という変数の数と無関係な指標に変換する理由です。

R でこのプロセスを見てみましょう (R7-1.R)。

#### R ソース 7.1 R7-1.R : カイ自乗検定

```
1  obtable2 <- matrix(c(25, 23, 12, 15, 17, 8),  
2                           nrow = 2, byrow = TRUE) # 2x3 表を作る  
2  chisq.test(obtable2)  
   # カイ自乗統計量で「距離」を測り、それに基づいて p 値を計算する
```

#### ■ R の出力結果

```
Pearson's Chi-squared test  
  
data: obtable2  
X-squared = 0.2083, df = 2, p-value = 0.901
```

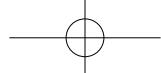
出力は上記のようになります。これらをまとめて言うと、

「全部で 100 人処方され、 $x_1$ ,  $x_2$ ,  $x_3$  の処方内訳は、40, 40, ■で、 $d_1$ ,  $d_2$  の担当内訳は 60, ■で、期待値表と大差ありませんでした ( $\chi^2 = 0.2083$ , 自由度 = 2,  $p = 0.901$ )」

のように言えることになります。

ここまで話は次のようにまとめられます。

- $2 \times 3$  分割表では 6 個の変数すべての情報を伝えることができること
- 総サンプル数と周辺度数のために、4 個の変数が必要なこと
- この 4 個の変数からは期待値表が作れること
- 実際の観察表についての完全な情報を伝えるためには、残りの 2 個の変数を使えばよいこと



- 観察表が期待値表から離れているかどうかは、それぞれの表を報告するときには使う変数の個数の差に応じて解釈すること
- この変数の個数の差が、検定における自由度に相当すること

### (5) 分割表の自由度

自由に 6 個の値を使って、 $2 \times 3$  表を作るとき、**自由度**は 6 です。では、「 $2 \times 3$  表の検定は自由度が 2」であるというのは、どういうことなのでしょうか。「 $2 \times 3$  表で行と列に関して独立性の検定をするときに、その統計量が自由度 2 である」とことを、短縮して「 $2 \times 3$  表は自由度 2」と言っています。「行と列に関して」とのただし書きがあります。これは、「行と列は与えられたものとする」ということでもあります。「独立性の検定」とのただし書きがあります。これは、「独立の仮定を基準とする」ということでもあります。したがって、 $n_{..}$ ,  $f_{..}$ ,  $n_{1.}$ ,  $n_{2.}$  は与えられたものとして、 $\delta_{ij}$ について考えよ、ということになります。そのような状況では、変数は 2 個ですよ、というのが「自由度 2」の意味です。確かに、 $\delta_{11}$ ,  $\delta_{12}$  しか自由に変えられませんから、自由度は 2 です。

一般的に、 $N \times M$  分割表について独立性の検定をするときの自由度は  $(N-1) \times (M-1)$  です。

## ◎ 7.1.2 量的データをかいづまむ

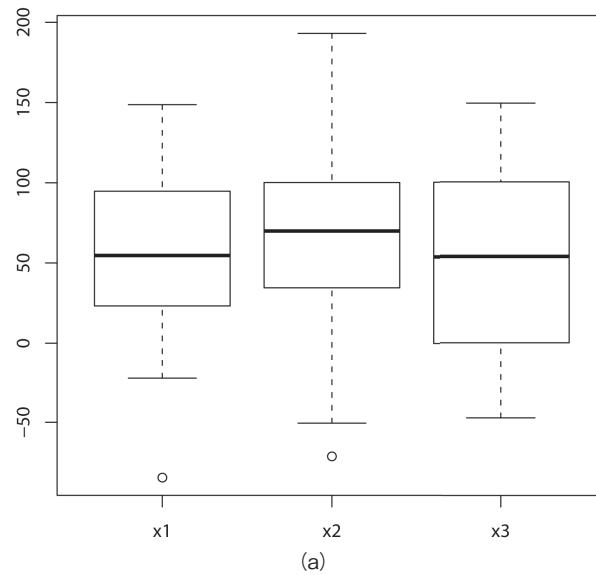
### (1) ばらつきを使って情報を伝える

薬  $x_1$ ,  $x_2$ ,  $x_3$  の服用者が 40, 40, 20 人ずついました。

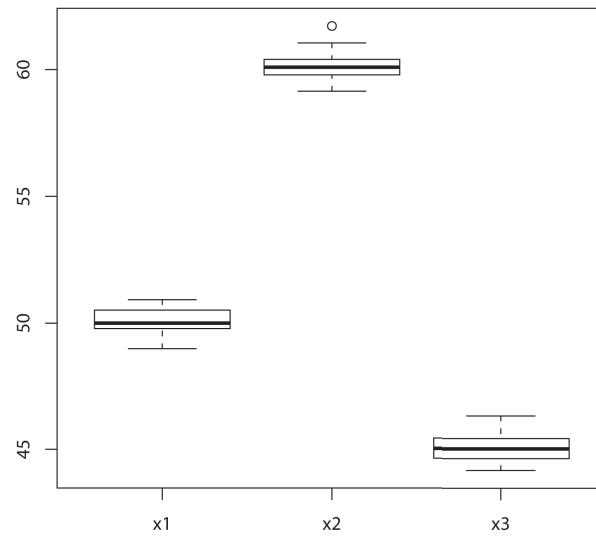
それぞれが病気の重さを表す検査を受けたところ、図 7.1 (a) に示すような分布をとったそうです。この結果を報告することを考えます。

全 100 人の検査結果を順番に伝えることもできるでしょうし、少しかいつまんで、全員の平均値と 99 人の検査結果を伝えてよいでしょう。100 人目の検査結果は平均値と 99 人の結果から逆算できますから、伝えるには及びません。どちらの方法でも 100 個の値を用いてすべての人の検査結果を伝えることができます。

図 7.1 (b) のように、3 群に差があるときには、3 種類の薬ごとに平均を出して、その差を伝える方が、優れたかいづまみ方です。全員の平均 ( $m_w$ ) がわかっているとき、3 種類の薬の服用人数 ( $n_1$ ,  $n_2$ ,  $n_3$ ) と 3 種類の薬のうち 2 種



(a)



(b)

図 7.1 3種類の薬  $x_1$ ,  $x_2$ ,  $x_3$  と検査結果

類の平均 ( $m_{x1}$ ,  $m_{x2}$ ) がわかっていれば、3番目の平均  $m_{x3}$  は逆算可能です。

また、それぞれの薬の服用者のうち、1人の値は残りの服用者の値を用いて逆算できるので、

$$\begin{aligned} 1 \text{ (全平均)} + 2 \text{ (2種類の薬の平均)} + (\text{第1薬の人数} - 1) + \\ (\text{第2薬の人数} - 1) + (\text{第3薬の人数} - 1) = \text{全人数} \end{aligned}$$

となり、やはり、全人数分の値を使って報告することができます。全体の自由度が 100 と大きくなりましたが、考え方は、分割表のときと同じです。

## (2) 群別のまとまりのよさ

さて、分割表のときにも問題にした通り、100 個の変数を使って完全な報告をするのか、たいした違いがないときには、その分を省略して報告するのかについて考えてみます。3種類の薬で検査値に違いがないと考えるか、違いがあると考えるかは、それを判断するための数値（検定統計量）とそれを評価するための自由度とそれらを使って出す  $p$  値に基づくのがよいです。

2人の医師の処方の違いのあり、なしの判断にあたって、観測表と期待値表との距離（カイ自乗統計量）とそれを評価するための自由度と、 $p$  値で判断したのと同じことです。今回は、ばらつきに着目します。

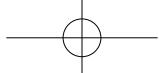
図 7.1 (b) のような分布だと、薬別に報告することが適當だろうと思われます。この場合に、群ごとの平均を報告するのが適當だと感じられるのは、各群のまとまりがよいかからです。各群のまとまりがどれくらいよいときに群別の平均値を報告することに意味があるかを決めるために、まず、まとまりをどうやって数値にするかを考えます。

すべてのサンプルの平均  $m_w$  (全平均) を中心にした 2 次モーメントは**全平方和** ( $SSw$ ) と呼ばれ、

$$SSw = \frac{1}{n} \sum_{i,j} (v_{ij} - m_w)^2$$

です。

群ごとのまとまりのよさは**群内平方和** ( $SSI$ ) と呼ばれ、群平均  $m_{xi}$  を中心とした 2 次モーメントを全群について足し合わせたもので、



$$SSi = \sum_{i=1}^3 \sum_{j=1}^{n_i} (v_{ij} - m_{x_i})^2$$

で表されます。

群の平均がばらつく具合は、**群間平方和** ( $SSb$ ) と呼ばれ、

$$SSb = \sum_{i=1}^3 n_i \times (m_{x_i} - m_w)^2$$

と表されます。

実は、

$$SSw = SSi + SSb$$

となっています。

データセットが与えられると、全平方和が決まります。このときに、群内平方和が小さめで群間平方和が大きめなときに、群ごとのまとまりがよいと感じますから、

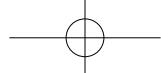
$$\frac{SSi}{SSw}, \frac{SSb}{SSw}, \frac{SSb}{SSi}$$

などの大小を用いて評価することができるはずです。

実際には、サンプルの数や群の数によらずに検定ができるようにさらに工夫がされています。群に関する自由度  $df_b =$  群数 - 1、群ごとのサンプルの自由度の和  $df_i = \sum_i^3 (n_i - 1)$  で補正をして

$$F = \frac{\frac{SSb}{df_b}}{\frac{SSi}{df_i}}$$

という値が使われます。この  $F$  は、全平方和の群間平方和と群内平方和への分配割合と、全体の自由度の大きさとその分配具合で決まる値です。この  $F$  を算出して、群別の平均値を報告するかしないかに目安を与える検定を（1次



元配置) **分散分析**と呼びますが、それは  $\frac{SS_b}{df_b}$ ,  $\frac{SS_i}{df_i}$  が、それぞれ、群間・群内分散の**不偏推定量**\*1となっているからです。

この  $F$  は群間に差がないとする仮説のもとで、 **$F$ 分布**に従うことが知られているので、この  $F$  を算出してそれを  $F$  分布に照らして検定することができます。R では以下の通りです (R7-2.R)。

### R ソース 7.2 R7-2.R : ANOVA

```
1 # 3 群の群別サンプル数、平均、SD
2 nx <- c(40, 40, 20); mb<- c(50, 60, 45);
  sdb <- c(50, 50, 50)*0.01
3 # 生起乱数でデータ作成
4 t1 <- rnorm(nx[1], mb[1], sdb[1]); t2 <- rnorm(nx[2],
  mb[2], sdb[2]); t3 <- rnorm(nx[3], mb[3], sdb[3])
5 tb <- c(t1, t2, t3)
6 # サンプルに群ラベルをつける
7 x <- c(rep("x1", nx[1]), rep("x2", nx[2]),
  rep("x3", nx[3]))
8 boxplot(tb ~ x)
9 # ANOVA
10 summary(aov(tb ~ x))
```

### ■ R の出力結果

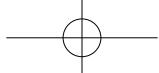
	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
x	2	3587.6	1793.80	7237.6	< 2.2e-16 ***						
Residuals	97	24.0	0.25								
	---										
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	.	0.1	' '	1

この出力は 2 行 5 列の表になっており、1 行目は 5 列、2 行目は 3 列のみに

\*1 分散の不偏推定量。今、ある分布があって、平均が  $m$ 、分散が  $v$  だったとします。そこから、 $N$  個をサンプリングして標本データが得られたとします。標本平均と標本分散がモーメントとして計算できます。標本分散を計算するときに、との分布の平均  $m$ を中心とした 2 次モーメントが  $v$  なのですが、 $m$  の代わりに標本平均を中心とした 2 次モーメントを計算すると、 $v$  よりも小さい値になります。そんなときに、標本分散

を  $\frac{N-1}{N}$  倍すると、元の分布の分散のよい推定値になります。これが不偏分散です。

$\frac{SS_b}{df_b}$  と  $\frac{SS_i}{df_i}$  は群間・群内分散の不偏分散になっています。



値が入ります。行は  $x$  が群間、`Residuals` が群内を意味し、列は `Df` (自由度)、`Sum Sq` (平方和)、`Mean Sq` (平均平方) がそれぞれの行に出ています。`F value` は  $F$  値、`Pr(>F)` は  $p$  値です。

## 7.2 次元と独立と直交

### 7.2.1 自由度と次元

自由度は変数の数の差であることがわかりました。今、 $k$  個の変数でデータを考えているとします。自由度を構成する  $k$  個の変数に 1 つずつ軸を割り振ることで、データは  $k$  次元座標上の 1 点に対応づけることができます。たとえば、 $2 \times 2$  分割表は自由度が 1 ですので、変数  $x$  を用いて、次のように表すことができます。

	A	a	計
B	$x$	$y = 9 - x$	9
b	$z = 12 - x$	$w = 18 + x$	21
計	12	18	30

$x$  は 0 から 9 までの値をとることができますので、それぞれの値について、複数の独立性検定を実施し、 $x$  を横軸に  $p$  値を  $y$  軸にプロットできます。その例が図 13.3 です。

同様に  $2 \times 3$  分割表を作ってみると、自由度が 2 なので、期待値  $e_{ij}$  と  $\delta_{11}$ ,  $\delta_{12}$  を使って、観察表を次のように表すことができます。

	AA	Aa	aa	
ケース	$e_{11} + \delta_{11}$	$e_{12} + \delta_{12}$	$e_{13} - (\delta_{11} + \delta_{12})$	60
コントロール	$e_{21} - \delta_{11}$	$e_{22} - \delta_{12}$	$e_{23} + (\delta_{11} + \delta_{12})$	40
計	9	42	49	100

2 つの変数  $\delta_1$ ,  $\delta_2$  を縦軸と横軸にとれば、観察表は 2 次元平面の点として表現できます。このような表に関する検定が 2 次元平面に表現されている例が、第 13 章(図 13.1) に出てています。このように、自由度を次元として扱うことができます。

## ● 7.2.2 分割表の自由度と線形独立——行列

自由度を行列演算で考えてみます。 $2 \times 3$  分割表に戻ります。周辺度数 ( $R, S, t_1, t_2, t_3, T$ ) がわかっているとき、 $2 \times 3$  個のセルの値を 6 変数で表します。

	$x_1$	$x_2$	$x_3$	計
$d_1$	$r_1$	$r_2$	$r_3$	$R$
$d_2$	$s_1$	$s_2$	$s_3$	$S$
計	$t_1$	$t_2$	$t_3$	$T$

分割表の定義から、

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{pmatrix} = \begin{pmatrix} R \\ S \\ t_1 \\ t_2 \\ t_3 \\ T \end{pmatrix}$$

と書き表せる行列の積に関する連立方程式が成り立ちます。

連立方程式を解くにあたり、きれいに解けて変数の値が全部決まるか、値が定められないかは、行列の用語で言えば一次独立であるかないか、ということになります。

一次独立かどうかの判定では、行列のランク（階数）<sup>※2</sup>というものを使って表します。ランクと変数の数が等しければ、すべての変数の値が決まり、ランクよりも変数の数が多ければ、「値が定まらない=自由」な変数であることになります。これを使って自由度を求めれば、「自由度 = 変数の数 - ランク」です。

R では次のようにして、算出します（周辺度数だけを条件にすれば、自由度は 2、(1, 1) と (1, 2) のセルの値を与えれば、自由度は 0 になります）(R7-3R)。

※2 行列のランクとは、列ベクトルの一次独立なものの中の最大個数。



## R ソース 7.3 R7-3.R : 行列の一次独立と自由度

```

1 m<-matrix(c(1,1,1,0,0,0,
2           0,0,0,1,1,1,
3           1,0,0,1,0,0,
4           0,1,0,0,1,0,
5           0,0,1,0,0,1,
6           1,1,1,1,1,1, # 以上 6 行が周辺度数の制約条件に対応する
7           1,0,0,0,0,0, # (1,1) セルに値を与える式に対応する行
8           0,1,0,0,0,0), # (1,2) セルに値を与える式に対応する行
9           ncol=6,byrow=TRUE)
10  # qr() はランクを求める関数
11  6-qr(m[1:6,])$rank # 第1-6 行のみを用いると自由度は 2 になる
12  6-qr(m[1:8,])$rank # 第1-8 行のすべてを用いると自由度は 0 になる

```

## ● 7.2.3 確率的独立と直交

7.2.2 項では自由度と変数の数と線形独立の話をしました。**独立**は確率の世界にもあります。確率の世界での独立は、ある事象とある事象とが相互に独立に起こるとき、その 2 つの事象が独立であるというの、両方が同時に生起する確率が、それぞれの確率の積で表されることを言います。

今、相互に独立な事象  $A$  と事象  $B$  の起きる確率  $P(A)$ ,  $P(B)$  と、両方の事象が同時に起きる確率  $P(A) \times P(B)$  との関係は、図 7.2 で表すように、2 つの大きな長方形の重なり部分になります。2 つの事象が独立であるときには、それぞれの事象に対応する軸を直交させることに相当します。

別の見方をします。今、2 つの相互に独立な連続型のデータを観測します。2 セットのデータの内積を使って、2 ベクトルのなす角度が約  $\frac{\pi}{2}$  であること（直交している）ことが確かめられます（R7-4.R）。

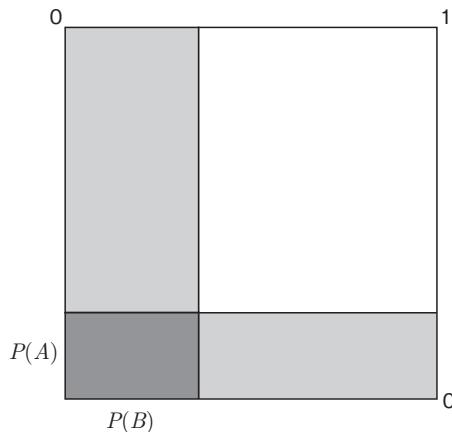
## R ソース 7.4 R7-4.R : 独立な 2 事象の作る余弦

```

1 Ns <- 10000; A <- rnorm(Ns); B <- rnorm(Ns) # 独立な値の Ns ペア
2 ip<-sum(A*B) # 内積
3 cos<- sum(A * B)/sqrt(sum(A * A) * sum(B * B)) # 余弦
4 acos(cos)/pi # 角度 (単位は $\\pi)。

```

このようにして算出した値が約 0.5 となり、2 つの独立な事象のデータベクトルが直交していることがわかります。4.2.5 項でデータの似ている程度を角



横軸に事象  $A$  の生起確率、縦軸に事象  $B$  の生起確率をとります。生起確率は帯で表した面積と考えることができます。 $A$  と  $B$  が独立なとき縦軸と横軸は直交し、 $A$  と  $B$  とが同時に起きる確率は、2つの帯の重なった部分の面積となっています。

図 7.2 同時に起きる確率

度で紹介したことと同じです。

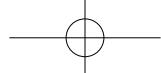
#### ● 7.2.4 線形独立と直交基底

確率で考えると独立とは直交と関係していましたが、行列では独立と言えば線形独立のこと、直交しているというわけではありません。 $k$  次元のデータを説明するために必要な最少のベクトルの数は  $k$  本であり、そのような  $k$  本のベクトルが満足するべき条件が、「 $k$  本のベクトルが線形独立である」ことです。

このような  $k$  本のベクトルは  $k$  次元空間を張る**基底**と呼ばれます、この基底の各ベクトルの長さや、ベクトル間の角度に制約はありません。しかしながら、基底の構成ベクトル同士がなす角を揃えると対称となり、対称であることは価値があることがあるので、そのようにしてとる基底を**直交基底**と呼びます。また、基底の構成ベクトルの長さが均一なことも、それだけで意味があるので、そのような基底には名前があって、**正規直行基底**と呼ばれます。

多次元のデータがあるときに、データのばらつきがよく見えるような正規直行基底を取り出すことがあります。特異値分解や固有値分解（主成分分析）と呼ばれる手法です。

これらを R で試してみることにします (R7-5.R)。



### ● 7.2.5 正規直交基底を取り出す——固有値分解

複数のグループに分けられる標本が多数あって、それについて、比較的多くの量的項目についてデータをとる状況を考えます。

データの作成方法とその処理は以下の通りです。

#### R ソース 7.5 R7-5.R : 固有値分解

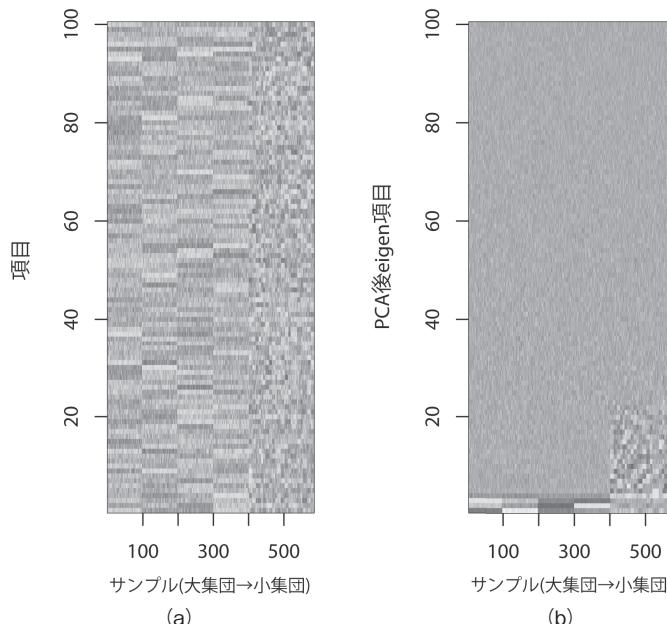
```
1 # 傾った集団構成 (100 人規模の亜集団 4 つと 10 人規模の亜集団を 20 個) で
2 # 100 項目のデータを作成
3 Nm<-100 # 項目数
4 # 亜集団別の人数発生 (100 人くらいの 4 亜集団と 20 人くらいの 10 亜集団)
5 Ns<-c(rpois(4,100),rpois(20,10))
6 Npop<-length(Ns) # 亜集団数
7 M<-NULL # 全ジエノタイプデータを納める行列
8 # 亜集団別に平均を振ってシミュレーション
9 for(j in 1:Npop){
10   tmpM<-matrix(rep(0,Nm*Ns[j]),ncol=Nm)
11   for(i in 1:Nm){ # 項目ごとのループ
12     af<-rnorm(1) # 項目の頻度のおよその値
13     tmpM[,i]<-rnorm(Ns[j],af) # 亜集団別の頻度
14   }
15   # 全データ行列に格納
16   M<-rbind(M,tmpM)
17 }
18 # データを標準化
19 wholemean<-mean(M)
20 M<-M-wholemean # 全平均が 0 になるように
21 mu<-apply(M,2,mean) # 列平均
22 M<-t(t(M)-mu) # 列平均が 0 になるように
23 # 固有値分解
24 svdout<-svd(M)
25 M2<-svdout$u%*%diag(svdout$d) # 分解後データ行列
26 par(mfcol=c(1,2))
27 # 固有値分解前後を image() プロット
28 image(1:sum(Ns),1:Nm,M,xlab="サンプル (大集団→小集団)",ylab="項目")
29 image(1:sum(Ns),1:Nm,M2,xlab="サンプル (大集団→小集団)",
30 ylab="PCA 後 eigen 項目")
31 df1<-as.data.frame(M);df2<-as.data.frame(M2) # データフレーム化
32 L<-1:5;par(mfcol=c(1,1))
33 plot(df1[,L]) # 5 軸がなす軸ペアでサンプルをプロット。分離しない
```

```

34 plot(df2[,L]) # 固有値分解後に分離力のあるトップ5軸でのプロットは分離する
35 VM1<-apply(M,2,var) # 項目ごとの分散(分解前)
36 VM2<-apply(M2,2,var) # 項目ごとの分散(分解後)
37 ylim<-c(min(VM1,VM2),max(VM1,VM2))
38 # 固有値分解前の各項目の分散はどれも同じ程度だが
39 # 固有値分解には、分散の大きいものと小さいものとのコントラストが大きくなっている
40 plot(VM1,ylim=ylim,type="b")
41 par(new=T)
42 plot(VM2,ylim=ylim,type="b",col="red")

```

項目数の次元空間に標本がばらついて存在していることになります。まず、標本には構成標本の多いグループが4つと、構成標本の少ないグループが20個で、観察項目100個でデータをシミュレーションしています。観察データそのものをRのimage()関数で表示すると、横軸に集団ごとのある程度のまとまりは見えますが、まとまりがよいとは言えません(図7.3(a))。これに固有値分解を実行します。固有値分解を施すと、正規直交基底が取り出されま



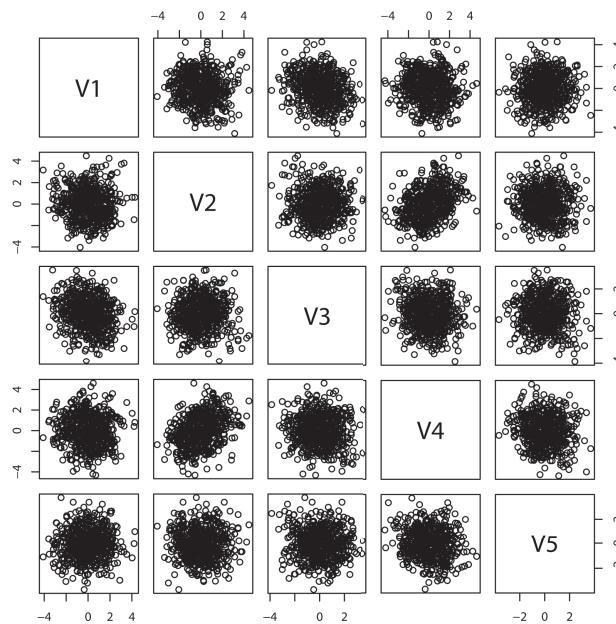
(a)では、左3分の2くらいが粗い濃淡となり、右3分の1くらいが細かい濃淡となっています。  
(b)では、それを固有値分解することで、各列について、濃淡のばらつきが下辺に集中してきています。

図7.3 image()関数による表示

す。直交基底なので、データ全体の分散は、軸ごとの分散の和になるわけですが、このとき、できるだけ軸ごとの分散にコントラストが出るように基底がとられます。

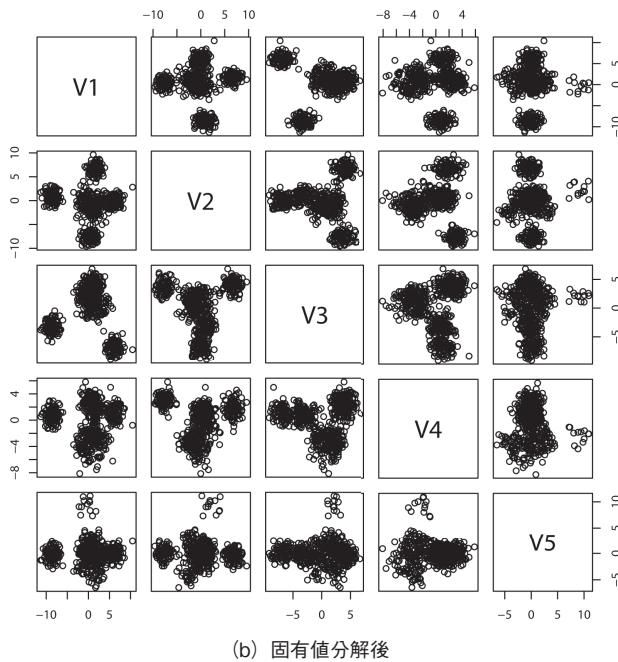
実際、図7.3 (b) では固有値分解によって取り出された少数の軸が下辺に集中していますが、この軸によって大半のばらつきが説明されています。それは `image()` 関数による表示の少数の下部では濃淡コントラストが大きいのに較べ、それ以外は濃淡がほぼ均一であることからわかります。

図7.4を見てください。固有値分解の前 (a) に、適当な5つの項目を選んで、それらを縦横軸にとってサンプルをプロットしてもサンプルが群に分かれているように見えませんが、固有値分解実施後 (b) に、分離力のあるトップ5軸を取り出して同様にプロットした方では、サンプルが分離されているのがわかります。ここでは、4つの大きなグループと多くの小さなグループを想定してシミュレーションしたデータを使っているので、第3軸までが効果的であることもわかります。



(a) 固有値分解前

図7.4 5項目のサンプル



(b) 固有値分解後

サンプルがクラスタリングしていることがわかります。

図 7.4 5 項目のサンプル (つづき)

特異値分解前の項目の分散と特異値分解後の軸ごとの分散をプロットすると、図 7.5 のようになります。分解前は、どの項目も似たり寄ったりの分散を持っていますが、分解後の項目の分散は、大きいものからだんだんに小さくなっています。これは、ばらつきが大きくなる軸を第 1 軸とし、残ったばらつきをなるべく大きくとれるような軸を第 2 軸としてとるように正規直交系を取り出しているからです。

固有値分解では、サンプルのばらつきを大きく説明する軸から選び出していき、最終的に項目数と同じだけの軸が得られますが、その新たな軸が説明するばらつきは、大きい方から降順に並びます。その様子を見てみます。分解前の場合は、すべての軸が似たり寄ったりの分散しか持っていないのに対して、分解によって選ばれた軸は、大から小へと変化しています。なお、それぞれの軸のとり方で、軸別の分散の振り分け状態は違いますが、全部の軸について分散を足し合わせれば、一致します。

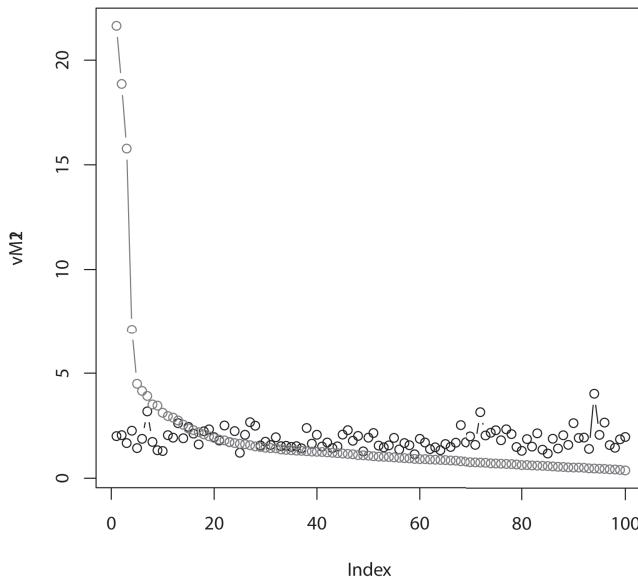


図 7.5 固有値分解前の軸ごとの分散（高さが低いところでばらばらとしているプロット）と分解後の分散（左肩上がりのプロット）

## 7.3 変数の構造と意味

### ● 7.3.1 変数の構造

これまで、変数を自由度の数だけ使ってデータのすべてを表現すること、また、かいづまんでは少數の変数だけで表現することについて考えてきました。また、変数のセットは、線形変換を施すことによって使いやすくできることも見てきました。

変数のとり方はどのようにとってもよいのですが、よくやるやり方というものがあります。

- すべてを平等に扱う方法
- 順序をつけて並べる方法
- 階層化する方法

がよく使われます

$N$  カテゴリの比率を例にとって見てみます。

$N$  カテゴリの占める比率を  $N$  個の数値で表現するときには、すべてのカテゴリが平等に扱われています。自由度が  $N-1$  です。この方法のよいところは、興味の対象に直属する性質そのものを対象としていることです。一方、 $N$  個の変数のすべてが平等なので、少ない変数でかいつまんで全体を捉えることが難しいです。

階層化する方法というのは、全体を徐々に小分けにしていく方法です。分岐木的な構造です。大づかみにしたいときには、木の根元寄りの変数だけでかいづまめば、データ全体の様子を大づかみにすることが可能です。分布を平均、分散で捉えるのは、1 次からだんだんと高次になる階層のうちの、1 次と 2 次のモーメントで捉えることです。

全体の分散を分散共分散行列の成分に分けるのも、同じ発想です。分散分析も群内分散と群間分散への分解であって、階層化処理をしています。また、集団の分集団構造を  $F_{ST}$  を含む、複数の  $F^{**3}$  に分解するのも、階層的な処理です。

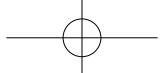
### ◎ 7.3.2 意味から選ぶ変数——データ構造で決める変数

情報をかいつまむ変数を定めることによってデータを理解するときに、変数が何を表しているのかは重要です。ですから、生物学的な意味や現象モデル上のパラメタなど、何かしら意味の定まった変数を選ぶというやり方があります。 $N$  カテゴリのそれぞれの頻度に変数を割り当てる、というのも、「あるカテゴリの頻度」という意味があるので、その値の解釈に迷うことはありません。

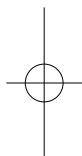
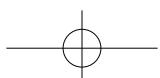
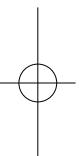
また、遺伝・生命現象を説明する数理モデルを立てて、データとの当てはまりを検討することも可能です。この場合の変数もモデルの中で意味を持っていますから、値の解釈は容易です。そのモデルを構成する変数を選ぶ方法です。変異率や組み換え率を変数化してモデルを立てたり、因子のリスクの大きさを変数化するなどがこの方法にあたります。

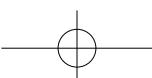
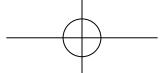
他方、データの塊に対して、その塊を説明しやすい角度から眺めるという方

※3 集団遺伝学で集団の偏り具合を定量的に表す指標です。



法もあります。行列型データを特異値分解して、説明力の大きな軸を取り出すのがこれにあたります。この方法の利点は、相互に直交する複数の軸が見いだされ、しかも、説明力の大きいものが得られることです。それらは、直交する因子となっています。しかしながら、得られた軸（因子）自体には自明な意味がないので、その軸の生物現象における意味を見いだす必要があります。





# 第 8 章

## 統計量、指標、確率、尤度

データの特徴を取り出したら、その取り出した値に意味を持たせて解釈したくなります。この章では値に意味を持たせるための仕組みについて考えます。

### 8.1 確率分布

#### ● 8.1.1 分布とは

本書では、ところどころで分布という用語を使ってきました。ここで分布について整理しましょう。

**分布**は、離散的にしろ連続的にしろ、順序ありにしろなしにしろ、とりうる値の範囲が決まっているときに、その範囲の中のどの値をどのくらいとるかを説明したものです。いわゆる距離（ユークリッドの距離）だったら、0以上の実数が範囲です。さいころの目だったら、1から6までの自然数が範囲です。DNAを構成する塩基だったら、A, T, G, Cの4カテゴリが範囲です。この範囲を**確率空間**とも呼びます。

**確率変数**が確率空間内の値をとるときに、めったにとらない値があったり、しばしばとる値があったりしますが、その様子を表したのが**確率分布**です。いろんな値をとる可能性がありますが、必ず確率空間内のどれかの値をとります。確率空間内の値について、とりやすさの程度を数値で表すことにします。とる可能性がないときの数値を0とします。少しでもとる可能性のある値には

正の値を与えます。とりやすさの程度に比例した値を与えることにします。その上で、範囲内のすべてについてとりやすさの値を足し合わせたら、それが1になるように補正します。

このように定まった範囲に、とりやすさの分布を持ったものを確率変数と呼んで、そのとりやすさの分布をその確率変数の確率分布と呼びます。

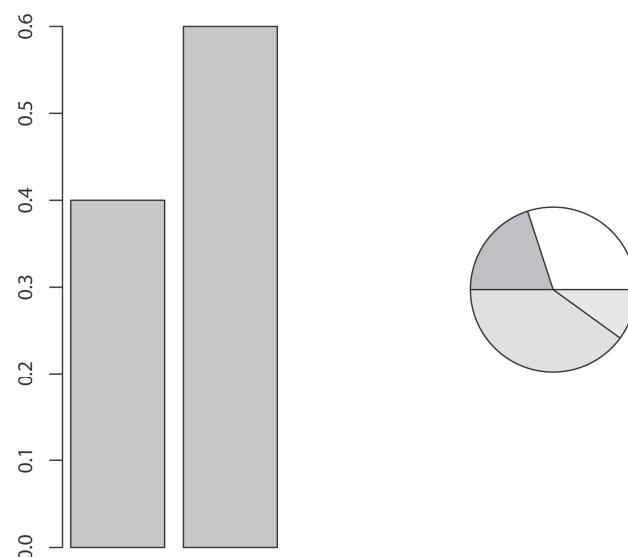
### ◎ 8.1.2 離散的な確率分布

赤玉4個、白玉6個の入った袋から、1個取り出すときに、赤玉である確率が0.4、白玉のそれは0.6です。確率は合わせて1です。4色になれば、赤0.3、白0.2、青0.4、黒0.1のようになります。

これをRで図示したグラフが図8.1です（R8-1.R）。

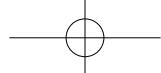
Rソース 8.1 R8-1.R : 離散的な確率分布

```
1 barplot(c(0.4,0.6),names.arg=c("赤","白"),ylab="確率")
2 pie(c(0.3,0.2,0.4,0.1),labels=c("赤","白","青","黒"))
```



2カテゴリの確率分布が2本の棒で表されています（左）。4カテゴリの確率分布が円グラフで表されています（右）。

図 8.1 離散的な確率分布



### ● 8.1.3 連続的な場合——指数分布

8.1.2 項では確率空間が離散的な場合を示しました。次に連続的な値をとる確率変数の場合を考えます。

連続的な場合には、数え上げて、足し合わせるわけにいきません。確率空間が連続的な例として、0 以上の実数を確率空間とする統計量の確率分布を作つてみることにします。実数直線という 1 次元空間において、確率空間は  $0 \leq x \leq \infty^{*1}$  で定義されているとします。この確率変数は  $x=0$  のときの確率を基準として、単位距離  $L$  増えるごとに生起確率が半分になるとします。この場合は、次の式が成り立ちます。

$$P(x+L) = \frac{1}{2} \times P(x)$$

$P(x)$  は定義された範囲で正なので、両辺の対数をとって、それを  $Q(x)$  とすれば、

$$\begin{aligned} Q(x+L) &= \log(P(x+L)) = \log(P(x)) - \log(2) \\ &= Q(x) - \log(2) \end{aligned}$$

ここで、 $\lambda = \frac{\log(2)}{L}$  とすれば

$$P(x) = P(0) \times e^{-\lambda x}$$

となります<sup>\*2</sup>。ここで、確率空間全体について、この  $P(x)$  を足し合わせると 1 になることから、

\*1 無限大。

\*2 少し変形して、 $Q(x+L) - Q(L) = -\log(2)$ 。ここから  $x$  の増加に比例して  $Q(x)$  が変化するので、

$$Q(x) = Q(0) - \frac{x}{L} \log(2)$$

となり、これを  $P(x)$  に戻すと

$$\log(P(x)) = \log(P(0)) - \frac{x}{L} \log(2)$$

となって、

$$P(x) = P(0) \times \exp\left(-\frac{x}{L} \log(2)\right)$$

であることがわかります。

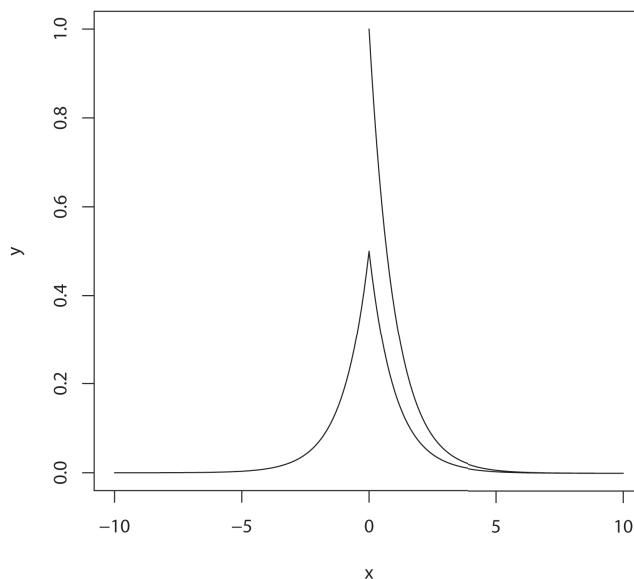
$$\lim_{X \rightarrow \infty} \int_0^X P(x) dx = \lim_{X \rightarrow \infty} P(0) \left[ -\frac{1}{\lambda} e^{-\lambda x} \right]_0^X = P(0) \times \frac{1}{\lambda} = 1$$

これを満足する条件から、 $P(0) = \lambda$  が得られて、結局、

$$P(x) = \lambda e^{-\frac{x}{\lambda}}; \lambda = \frac{\log(2)}{L}$$

が得られます。これは、**指數分布**と呼ばれる分布の**確率密度関数**です。このように、連続な確率空間では、確率密度が関数で表せることがあり、確率密度関数と、その足し合わせである**累積分布関数**との間は、**微分**と**積分**とでつながっています。

$P(x) = \lambda e^{-\lambda x}$  の式の由来を確認します。原点から遠ざかる方向にある距離を進むと、生起確率が  $\frac{1}{e}$  になるような分布でした。時間のように 1 方向性の数直線ならばこうなりますが、両方向性の 1 次元空間の場合には、原点から遠ざかる方向が 2 つありますから、確率空間を  $-\infty \leq x \leq \infty$  としてやって、



ラプラス分布は正負に値を持ち、指數分布は正にのみ値を持ちます。指數分布の高さはラプラス分布の高さの 2 倍になっています。

図 8.2 指數分布とラプラス分布



$$P(x) = \frac{1}{2} \lambda e^{-\lambda|x|}$$

としてやることにします。 $\frac{1}{2}$ は1方向性から2方向性への変化、 $|x|$ と絶対値をとっているのは、 $x$ が負の場合に距離にするためです。これは2方向性の指数分布ですが、**ラプラス分布**とも呼ばれます。指数分布とその折り返しを含むラプラス分布の両方を重ねて描いた図8.2から、高さが半分であること、対称性がでていることが読み取れます。

#### ● 8.1.4 指数分布と正規分布との違い

さて、8.1.3項で原点からの距離がある単位だけ変化すると、生起確率が $\frac{1}{e}$ になる場合を考えました。では、原点からの距離の2乗がある単位だけ変化すると、生起確率が $\frac{1}{e}$ になるような確率変数だったら、どうなるのでしょうか？

指数分布・ラプラス分布を少し変形して、

$$P(x) = Ce^{-\lambda|x|^2}$$

という形をしていることが予想され、実際、

$$\int_0^\infty P(x) = \frac{1}{2}$$

を満足するように、 $C$ を求めてやると、

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

であることが知られています。実は、この式は、分布の紹介で最初に出てくることが多い、平均0の**正規分布**の式と同じです。正規分布のよくある表記法で書き直せば、

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

です。

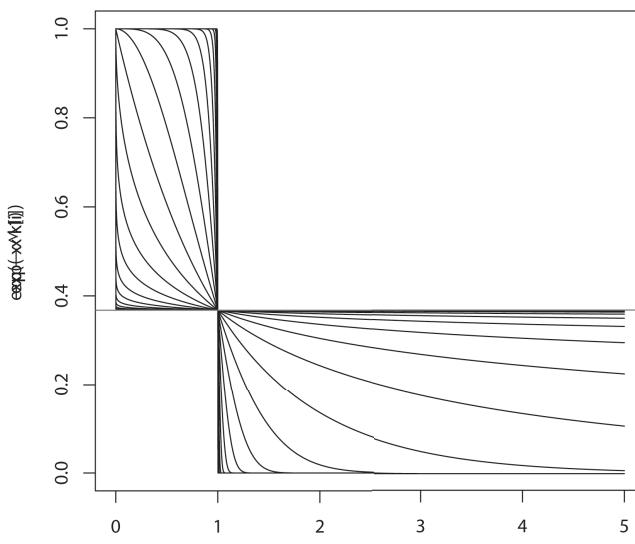
$$k=2, \lambda = \frac{1}{2\sigma^2}$$

の関係があります。

### ● 8.1.5 一様分布、指數分布、正規分布、矩形分布——一般正規分布

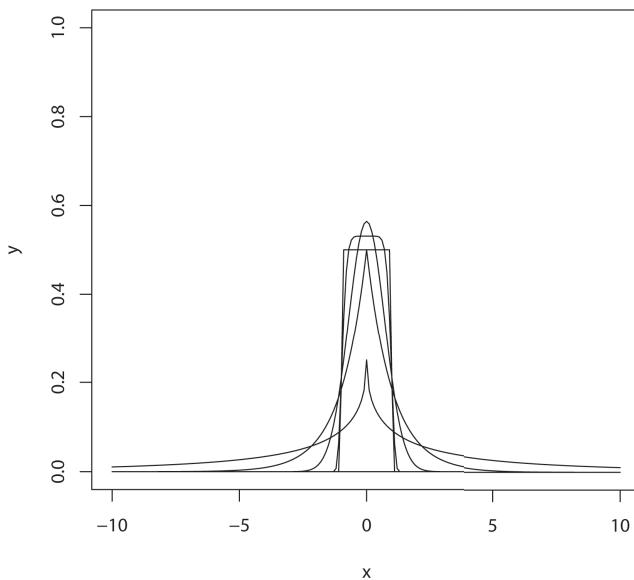
実際、 $e^{-\lambda|x|^k}$  という関数は、 $k=1$  のときに指數分布（ラプラス分布）、 $k=2$  のときに正規分布、 $k$  が 0 に近くなると  $x=0$  のときだけ 1 で、他は  $e^{-1}$  で一様（**一様分布**）です。 $k$  が無限に大きくなると、 $x < 1$  で 1、 $x = 1$  では  $e^{-\lambda}$ 、 $x > 1$  では  $f(x) = 0$  であるような、**矩形分布**です。

これら一様分布、指數分布（ラプラス分布）、正規分布、矩形分布の確率密度分布は、確率空間での積分が 1 になるように補正を施すことで確率密度分布にできます。ここで、ラプラス分布  $\frac{1}{2}\lambda e^{-\lambda|x|}$ 、正規分布  $\frac{1}{\sqrt{\pi}}\lambda^2 e^{-\lambda x^2}$  をよく眺めると、 $C\lambda^k e^{-\lambda|x|^k}$  というように表せそうです。実際、次のようにすると、確



(a)  $e^{-\lambda|x|^k}; k = -10, -9, \dots, 0, 1, \dots, 10$  のグラフ

図 8.3 一般正規分布



(b) (a) を確率密度分布にしたもの

図 8.3 一般正規分布（つづき）

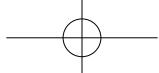
率空間全体での積分が 1 になることが示されています。

$$P(x; k) = \frac{1}{2} \frac{1}{\Gamma\left(\frac{1}{k} + 1\right)} \lambda^{\frac{1}{k}} e^{-\lambda|x|^k}$$

ただし、 $\Gamma$  はガンマ関数<sup>※3</sup>と呼ばれる関数です。

これを**一般正規分布**と呼んだり、**一般誤差分布**と呼んだりします。さまざまな対称な分布がこの表記で表現できることがわかりました。図 8.3 にあるように、矩形の場合は幅が 2、高さが 0.5 です。一様分布は高さが 0 です。指数分布の折り返しであるラプラス分布は、原点で最高点 0.5 を持ち、そこが尖っています。正規分布は、指数分布よりも最高点が高く、頂上が滑らかです。

※3 ガンマ関数は、ここで示した分布の関数にも登場しますし、 $\Gamma(n+1) = n!$  であること、 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$  であるなど、場合の数の計算や、円・球などとも関係する関数です。



### ● 8.1.6 正規分布、カイ分布と次元

平均 0 の正規分布の確率密度分布の表記法として、

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

がよく用いられるのは、この正規分布の分散が  $\sigma^2$  であり、分布を分散から特定できるというメリットがあるからです。では、平均 0、分散 1 の正規分布（標準正規分布）の式を示しましょう。 $\sigma^2=1$  ですから、

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

です。

さて、指數分布を考えるときに、原点からの距離について考え始めました。その上で、1 次元の数直線上で負の領域にも話を広げてラプラス分布の式を作りました。原点を中心に対称にするために、確率密度を半分にしたのでした。

正規分布について、この指數分布からラプラス分布への変化の逆を行います。今、正規分布は、正負の実数が確率空間ですが、原点からの距離のみに興味があるとしますと、0 以上の範囲だけが確率空間となります。確率密度は、2 倍にすればよいです。

$$2 \times \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{x^2}{2}}$$

です。

これは、自由度 1 のカイ分布と呼ばれる分布の確率密度関数です。「カイ分布」であって、「カイ自乗分布」ではありません。「カイ自乗分布」の方が統計的検定でよく使われる方の分布です。

この分布は「自由度 1」のカイ分布でした。自由度 1 のカイ分布は、1 つの変数が自由に動かせるときの分布であって、自由度を 1 以外にも変えることができるところが、その名称から示唆されています。自由度が 1 なのは、数直線という 1 次元空間を扱っているからです。標準正規分布から、自由度 1 のカイ分布を作ったときに行つたことは、数直線上の座標を、「原点からの距離」に変

更するための処理でした。1次元数直線上の座標を距離にするには、同じ距離の点が2つあるので、2倍したのです。

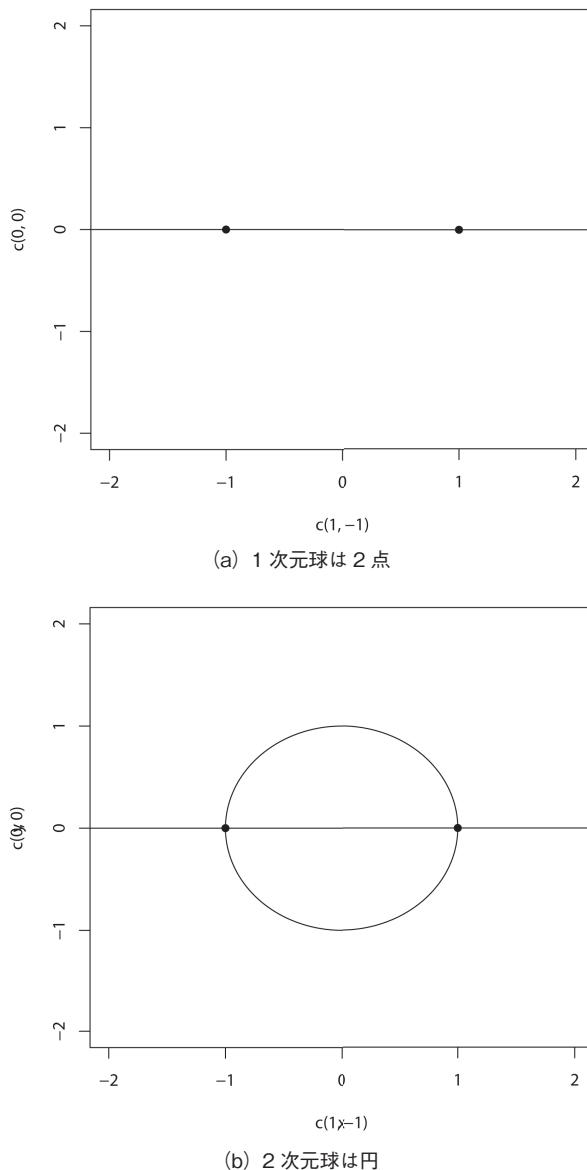
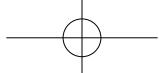


図 8.4 球面：ある点からのユークリッド距離が等しい点の集まり



では、自由度が2のときはどうしたらよいのでしょうか？ 自由度2ということは、2次元空間です。2次元空間で同じ距離の点を集めると、それは円になります。円を構成する点の個数がいくつであるかは、数えられません。その代わりに長さは計算して求めることができます。

$$2\pi x$$

です。 $x$ は原点からの距離で**半径**です。では、この一周を構成する点がすべて  $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$  の確率で発生しますから、自由度2のカイ分布の確率密度関数は、

$$2\pi x \times \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = \sqrt{2\pi x}e^{-\frac{x^2}{2}}$$

でしょうか。幸い、この関数の積分は  $\frac{de^{ax^2}}{dx} = 2axe^{ax^2}$  であることを利用すれば簡単ですから、実際に確率空間  $0 \leq x \leq \infty$ について積分して求めてみて、それが1になるかどうかで確認ができます。

$$\int_0^\infty \sqrt{2\pi x}e^{-\frac{x^2}{2}} dx = \sqrt{2\pi} \left[ (-e^{-\frac{x^2}{2}}) \right]_0^\infty = \sqrt{2\pi}$$

1次元のときに2つ折にした場合と異なり、距離  $x$ に応じて対応する点の数が異なるために、すべて足し合わせて1にならっていません。しかしながら、距離  $x_1$  と  $x_2$  の点を集めた確率の比は、

$$\frac{\sqrt{2\pi x}e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi x}e^{-\frac{x_2^2}{2}}}$$

でよいのですから、全体の積分が1になるように、 $\sqrt{2\pi}$ で補正してやればよいでしょう。したがって、2次元平面において、原点から遠ざかるにつれて生起確率が正規分布的に小さくなっていくようなときに、距離が等しい点（生起確率が等しい点）の確率を足し合わせたような確率変数の確率密度関数は、

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

となるはずです。そして、これは、確かに、自由度  $k$  のカイ分布の一般式に  $k = 2$  を当てはめた式

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

と一致しています。

では、正規分布から、任意の自由度  $k$  のカイ分布を作つてみることにしましょう。1 次元のときは 2 倍、2 次元のときは  $2\pi x$  倍しました。これは、1 次元と 2 次元空間の半径  $x$  の「円周」です。一般次元の「円周」(**多次元球**の大きさ) は、

$$S(x; k) = 2 \frac{1}{\Gamma\left(\frac{k}{2}\right)} \pi^{\frac{k}{2}} x^{k-1}$$

であることが知られています。

したがつて、自由度  $k$  のカイ分布の確率密度分布は、

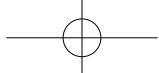
$$S(x; k) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{2\pi}} \frac{2\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{k-1} e^{-\frac{x^2}{2}}$$

に比例して、空間全体の積分が 1 になるような補正を加えられた式になっていくことでしょう。

実際、自由度  $k$  のカイ分布は、

$$C \times S(x; k) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \left( \frac{1}{\sqrt{2\pi}} \right)^k S(k) e^{-\frac{x^2}{2}}$$

$$C = \left( \frac{1}{2\pi} \right)^{\frac{k-1}{2}}$$



と書き表せることがわかります。これは、 $e^{-\frac{x^2}{2}}$  が表すように、原点から遠ざかると確率が小さくなる分布であって、その小さくなり方が、正規分布と同じように、 $x^2$  が一定量増えると  $\frac{1}{e}$  になるような小さくなり方である分布です。そして、次元が  $k$  であるときには、 $k$  次元球の「表面積」 $S(x; k)$  に関する要素を除けば、 $\left(\frac{1}{\sqrt{2\pi}}\right)^k$  で補正することで、空間全体の積分を 1 にできることを示しています。

1 次元の正規分布のときに  $\frac{1}{\sqrt{2\pi}}$  が同様に補正項であったことと、それを  $k$  次元に一般化したこととを考え合わせれば、納得がいきます。実際、**多変量正規分布**の確率密度関数でも、補正の項として、 $\left(\frac{1}{\sqrt{2\pi}}\right)^k$  が現れるのですが、同じことです※4。

### ◎ 8.1.7 カイ分布からカイ自乗分布

8.1.6 項で、正規分布を多次元に展開して、距離ごとに確率を足し合わせることで、多次元を 1 変数で取り扱っている分布としてカイ分布が出てきました。どちらの分布も原点からの距離  $x$  を変数として扱いました。しかしながら、これらは距離の 2 乗の多寡を気にする分布です。それが  $e^{-\lambda x^2}$  を式に含む理由です。ですが、気になるものが  $x^2$  のなだったら、いっそのこと、 $X = x^2$  の  $X$ 、つまり原点からの距離の 2 乗そのものを変数として確率分布にしてしまうのも、すっきりしそうです。そのようにして得られた確率密度分布関数が、

$$Pr(\chi^2 = X) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} X^{\frac{k}{2}-1} e^{-\frac{X}{2}}$$

です。検定でしばしば登場する**カイ自乗分布**です。

※4 多次元正規分布の確率密度関数は、ベクトル  $(x - m)$  と行列  $\Sigma$  を用いて次式で表されます。

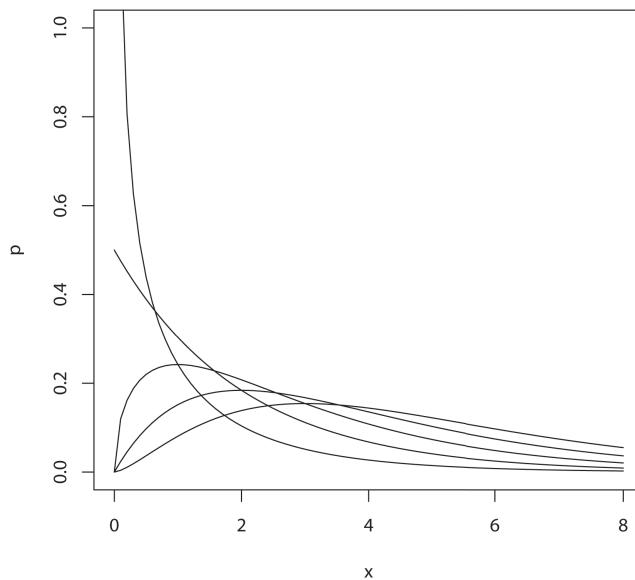
$$f(x) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - m)^T \Sigma^{-1}(x - m)\right)$$

### ● 8.1.8 最も観察されそうなカイ自乗値

カイ自乗分布の自由度別の確率密度分布を見てみます。自由度が 1, 2, 3, 4, 5 の場合を図示します（図 8.5）。

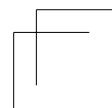
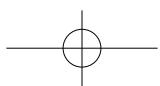
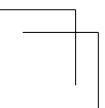
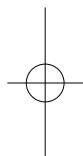
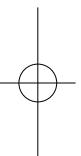
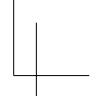
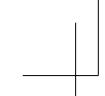
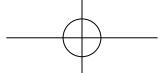
自由度が 1, 2 のカイ自乗分布では確率密度が最大なのは、距離が 0 のときですが、自由度が 3 以上になると、ある程度距離のあるときに確率密度が最大になっている様子が見てとれます。

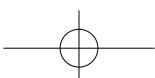
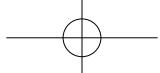
原点からの距離（距離の 2 乗）が大きくなると、生起確率は小さくなります。が、自由度が大きくなると、距離が小さい場合よりも距離がある程度大きい場合の方が、空間が広い（多次元球の表面積として広い）ので、最もとりやすいカイ自乗値が大きい方にシフトしているのです。



最大の値を与えるカイ自乗値（横軸）が、カイ自乗値として最もとりやすい値です。自由度が上がると、最もとりやすい値が次第に大きくなります。

図 8.5 カイ自乗分布の確率密度分布（自由度 1, 2, 3, 4, 5）





# 第 9 章

## 確率と尤度

### 9.1 確率、尤度

ここまで、確率というものがわかったものとして話を進めてきました。この点を少し詳しく見るにあたり、**確率**と**尤度**というよく似た概念に焦点を当てて整理することとします。

#### ◎ 9.1.1 確率

A と B との 2 種類にラベル付けされた玉が全部で 40 個（A 12 個、B 28 個）入った袋から、全部で 10 個の玉を取り出す場合を考えます。2 種類の玉の取り出し方の場合の数は、A の個数に着目して 0 から 10 個の 11 通りあります。1 個取り出してラベルを確認しては、取り出した玉を袋に戻して、再度、取り出すという作業を 10 回繰り返す場合と、10 個を一度に取り出す場合とを考えます。どちらのやり方でも、一番確率が高いのは、3 個が A で 7 個が B の場合です。もちろん、0 個から 10 個までの場合の確率を足し合わせると 1 です（R9-1.R、図 9.1）。

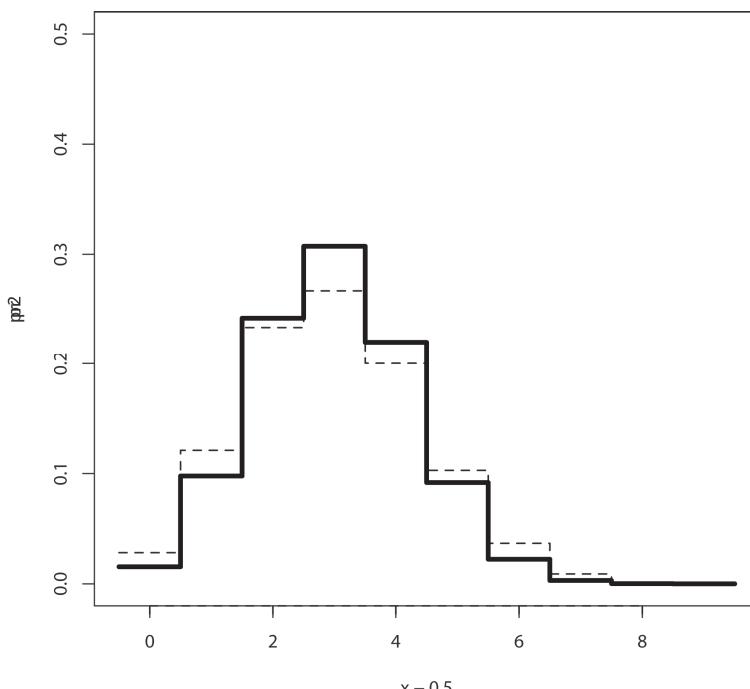
#### R ソース 9.1 R9-1.R : ランダムサンプリング

```
1 # 1 個ずつ取り出しては戻すのを 10 回繰り返しす場合
2 p <- 0.3; x <- 0:10 # A玉の割合は 0.3
3 pr <- choose(10, x) * p^x * (1 - p)^(10 - x)
   # choose(a,b) は a 個から b 個を取り出す場合の数
```

```

4 A <- 12; B <- 28; S <- A + B
5 x <- 0:10; y <- 10 - x
6 # 10個を取り出す場合
7 pr2 <- exp(lgamma(10 + 1) + lgamma(30 + 1) +
   lgamma(A + 1) + lgamma(B + 1) - (lgamma(S + 1) +
   lgamma(x + 1) + lgamma(y + 1) + lgamma(A - x + 1) +
   lgamma(B - y + 1)))
8 ylim <- c(0, 0.5)
9 plot(x-0.5, pr, ylim = ylim, type = "s")
10 par(new = T);
plot(x - 0.5, pr2, ylim = ylim, type = "s", col = "red")

```



40個中12個がA玉であるときに、10個の玉を取り出してそのうちA玉の個数がいくつになるかの確率。1個、取り出して戻すのを10回繰り返した場合が破線、10個を一度に取り出した場合が太い実線です。Rのソースは2種類の線を色で区別しています。

図9.1 ランダムサンプリング

40個の玉がA, Bの2種類である点は変えずに、その内訳がいろいろあることにします。Aは0個から40個までのいずれかです。Bは逆に40個から0

個です。全体の A, B の内訳が変わると、取り出される 10 個の A, B の内訳ももちろん変わります。袋全体の中味に A が多いときには、取り出される 10 個も A が多くなりがちです。

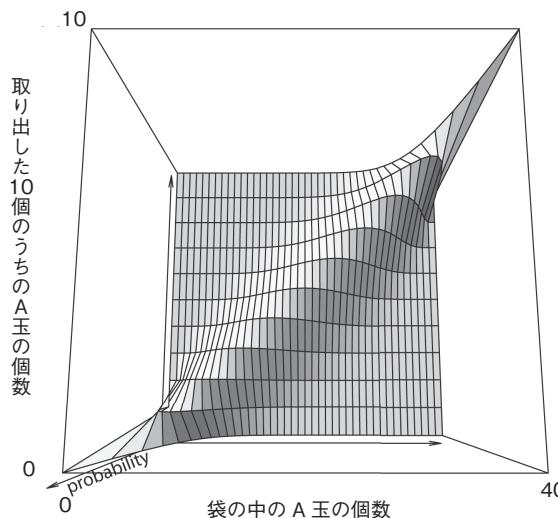
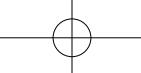
その確率を図にしたのが図 9.2 です (R9-2.R)。

### R ソース 9.2 R9-2.R : 確率と尤度

```

1 exProb<-function(m) { # 正確確率を算出
2   m1<-apply(m,1,sum) # 行の和
3   m2<-apply(m,2,sum) # 列の和
4   s<-sum(m)
5   exp(sum(lgamma(m1+1))+sum(lgamma(m2+1))-
     (lgamma(s+1)+sum(lgamma(m+1))))
6 }
7 # 総数 S そのうちの A 玉数、B 玉数。抜き取り総数 N そのうちの A 玉数 x
8 S<-40;A<-0:S;B<-S-A;N<-10;x<-0:N
9
10 probmat<-matrix(0,length(A),length(x))
11 for(i in A){
12   for(j in x){
13     y<-N-j; z<-i-j; w<-S-(j+y+z);
14     if(j>=0 & y>=0 & z>=0 & w>=0){
15       data<-c(j,y,z,w)
16       probmat[i+1,j+1]<-exProb(matrix(c(j,y,z,w),nrow=2))
17     }
18   }
19 }
20 phi<-80;theta<-0;shade<-0.3 # 描図パラメタ
21 persp(A,x,probmat,xlab="No. A in 10", ylab="No. A in 40",
22         zlab="probability",phi=phi,theta=theta,shade=shade)
22 plot(A - 0.5, probmat[, 4], type = "s")
23           # 取り出した玉のうち 3 個が A 玉だった確率
23 abline(h = 0, col = "red")

```



全部で 40 個の 2 種類の玉のセットから 10 個を取り出したときの、玉の内訳別確率。横軸が袋の玉 40 個のうち A 玉の個数、縦軸が 10 個を一度に抜き出したときの A 玉の個数。

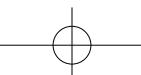
図 9.2 玉の内訳別確率

### ◎ 9.1.2 尤度

引き続き A, B、2 種類の玉が合計 40 個入っている袋から、10 個を取り出すことを考えます。今、10 個を一度に取り出したところ、A, B が 3 個と 7 個だったとします。袋全体の 40 個の A, B の内訳について、どんなことが言えるでしょうか？

A が 3 個、取り出されたことから、元の袋には A が 3 個以上はあったことが「確実」に言えます。また、B が 7 個、取り出されたことから、元の袋には B が 7 個以上あったことも「確実」です。このことから、(A, B) の内訳は、(3, 37), (4, 36), ..., (33, 7) の全部で 31 通りの可能性があることがわかります。9.1.1 項では、元の袋の A, B の内訳ごとに、10 個を取り出したときの A, B の内訳の確率を計算してありましたから、それを見てみることにします。

図 9.3 の 3 次元グラフでは、横軸に袋の中の A 玉の数（0 から 40）が、縦軸に取り出された 10 個の中の A 玉の個数が対応しています。その縦横軸で 2 次元の格子が作られ、高さは、袋の A 玉の数を条件として与えたときに、取り出される A 玉の数に関する確率です。袋の中の A 玉の数を「仮説」、取り出

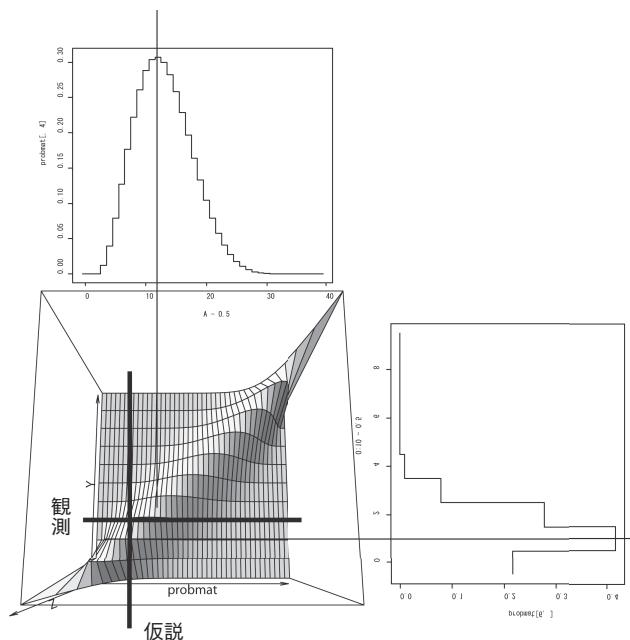


される A 玉の数を「観測」とすれば、観測の仮説ごとの確率を、すべての条件と、可能なすべての観察についてプロットしたものと言えます。

ある仮説に着目するとき、3 次元グラフ（図 9.3）の太い縦線の部分を見ればよいです。袋の中に A 玉が 5 個ある場合について抜き出したのが、図 9.3 の右側のグラフです。縦軸が抜き出される A 玉の数、横軸が確率となっています。これは、「袋の中に A 玉が 5 個ある場合」の「取り出される A 玉の数」に関する確率密度分布のグラフです。

別の方に切り出してみます。「A 玉が 3 個取り出された」という観察をする確率を、「袋の中の A 玉の数」について抜き出したのが、図 9.3 の上部のグラフで、それに対応するのが、3 次元グラフでは太い横線になります。

A, B の取り出し個数が 3, 7 個のときの部分だけを抜き出して描いた、図 9.3 の上側のグラフを読み取ってみましょう。



3 次元グラフは横軸に袋の中の A 玉の個数、縦軸に取り出した玉のうちの A 玉の個数。袋の中の A 玉の個数が 5 個のときのグラフの断面が右側に、取り出した玉のうち 3 個が A 玉だったときの場合についての断面が上部に表示してあります。

図 9.3 観測の仮説ごとの確率

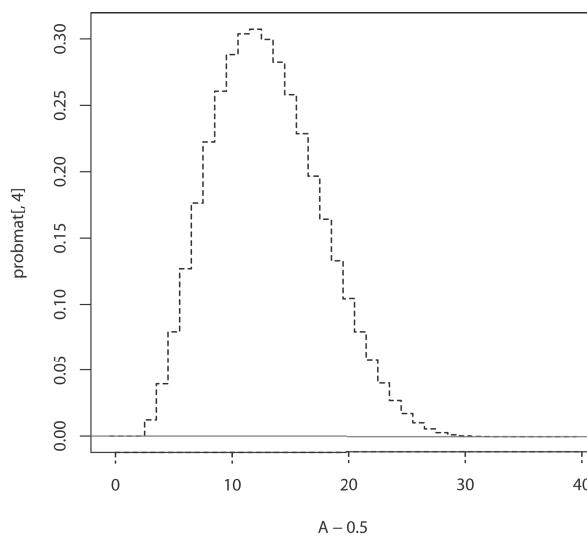
$A$  が 0, 1, 2 の場合と 34, 35, 36, 37, 38, 39, 40 の場合は確率が 0 です。ある情報（10 個を抜き出したらその内訳が 3, 7 だった）が与えられたときに、知りたいこと（40 個入りの袋の内訳）に関する知識が増えて、それがこのグラフに表されているということです。40 個の内訳のうち、ある場合は「ありえない」で、ある場合はある場合にくらべて「ありえる程度が大きい」という知識です。この「ありえる程度」を「尤度 (likelihood)」と言います。尤度を元の袋の A 玉の数  $a$  の関数として表すと、

$$L(A = a | (3, 7)) = \frac{a!(40-a)!10!30!}{40!3!7!(a-3!)(33-a)!}$$

となり、これを**尤度関数**と言います。

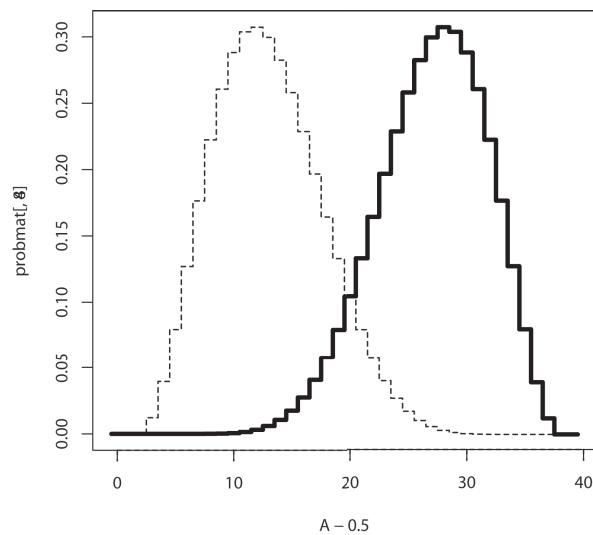
### ◎ 9.1.3 確率の和は 1、尤度の和は 1 ではない

図 9.4 の尤度関数のグラフは両端で値が 0 になった、山形のグラフです。形が確率密度関数のグラフに似ていますが、このグラフの線の下部分の面積（積分）をしてみると、和が 1 にはなりません。R9-3.R で確認することができます。

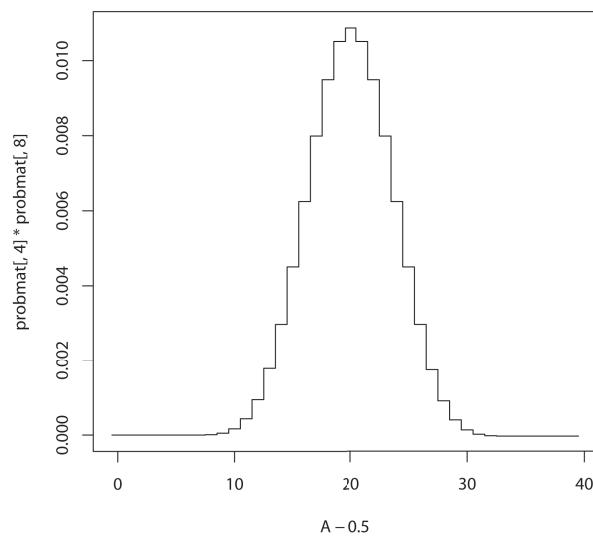


(a) (3, 7) 個の観測のときの尤度

図 9.4 尤度

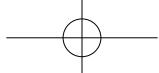


(b) (3, 7) 個の観察（破線）と (7, 3) 個の観察（実線）のときの尤度



(c) 2 度の観察情報の両方を使った尤度

図 9.4 尤度（つづき）



## R ソース 9.3 R9-3.R : ●●

```
1 sum(probm[, 4])
# 確率・尤度の行列のうち、10個中3個がA玉だったという観測に対応する列
```

尤度の分布は確率密度分布ではないことがわかります。足して1にはなりませんが、袋全体の中のAの個数に関する尤度は、次のように相対的に意味があります。Aが12個の尤度は $L(A=12 | (3, 7)) = 3.073032e - 01$ 、それに対してAが3個の尤度は $L(A=3 | (3, 7)) = 1.214575e - 02$ です。

$$L(A=12 | (3, 7))/L(A=3 | (3, 7)) = 25.3013$$

12個の尤度は、3個の尤度の25.3倍です。これが、2つの可能性（Aが12個か3個か）の尤度の比（尤度比）です。

さて、ここで、抜き出した10個を元に戻して、再度、10個を抜き出し直したら、今度は、A, Bが7個と3個だったとします。この情報をもとにした尤度を描いてみます（図9.4（b））。

では、元のAが0から40までのそれぞれの尤度を見てみましょう。2度の抜き取りは相互に独立しているので、1度目の尤度と2度目の尤度を掛け合わせます（R9-4.R）。

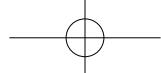
## R ソース 9.4 R9-4.R : 抜き取りを2回したとき

```
1 # probmat[, 4] : 確率・尤度の行列のうち、
# 10個中3個がA玉だったという観測に対応する列と
2 # probmat[, 8] : 確率・尤度の行列のうち、
# 10個中7個がA玉だったという観測に対応する列との積
3 plot(A - 0.5, probmat[, 4] * probmat[, 8], type = "s")
```

2度の結果を合わせると、尤度が最も大きいのは、40個のうちA, Bが半分ずつの場合でした（図9.4（c））。

### ◎ 9.1.4 尤度の指数化——尤度比と事前確率、事後確率

抜き取りを繰り返したとき、抜き取り1回分が教える尤度を掛け合わせて、統合した尤度を求めました。すべての場合について尤度を足し合わせたときに1になるように調整することにします。なぜなら、1は何回掛け合わせても1のままなので、尤度の掛け算をするときに便利だからです。



このように調整してやった上で、観察するたびに、尤度を掛け算で計算してやると、ある仮説の尤度が  $L_1$  から  $L_2$  に変わったとき、 $L_1$  と  $L_2$  の大小を比較することができます。

今、観測の前後で、比較可能にした観測前後の尤度を観測という「事」の「前」と「後」の尤度であって、「確率と同じように、足して 1 になっている」という意味で、仮説の**事前確率**と**事後確率**と呼びます。この事前確率と「事」の観察と事後確率の関係のことを、**ベイズの定理**と言います。

9.1.3 項の例では、1 回目の抜き取りの前は、40 個の玉のうちの A 玉の数は、0, 1, ..., 40 のいずれも等しいと考えていました。1 回目の抜き取りの結果、A が少なめであると考えました。この考え方の変化が事前確率から事後確率への変化です。2 回目の抜き取りに際しては、1 回目の抜き取りの後の事後確率が、2 回目の抜き取りの前の事前確率となり、抜き取り結果を見た後で、A の数は半分くらいという事後確率に変化したわけです。

この例では、最初には、すべての仮説が同等にもっともらしいしていましたが、そうする必要はありません。たとえば、「風のうわさ」で A 玉が少なめだと知っていれば、A 玉の数が 0, 1, 2, ..., 19 の尤度を 20, 21, ..., 40 の尤度の 2 倍としてスタートすることも可能だからです。

## 9.2 条件付確率、確率、尤度、非独立

### ● 9.2.1 カテゴリ型の条件付確率

2 つの因子  $A/a, B/b$  で 4 グループに分けられた集団を考えます。全体を 1 とすると、4 グループの比率を下のような表にすることができます。2 通りの表を作ります。片方は、 $A$  と  $B$  とが**独立**な場合、もう片方は独立ではない場合です。

独立な場合

	$A$	$a$	計
$B$	0.3	0.2	0.5
$b$	0.3	0.2	0.5
計	0.6	0.4	1

独立でない場合

	$A$	$a$	計
$B$	0.4	0.1	0.5
$b$	0.2	0.3	0.5
計	0.6	0.4	1

比率の様子を図9.5で表してみます。面積が比率です。独立な場合には格子状になり、独立でない場合にはでこぼこすることがわかります。

#### R ソース 9.5 R9-5.R : 独立か否かの格子表示

```

1 t1 <- as.table(matrix(c(0.3, 0.2, 0.3, 0.2), nrow = 2,
2   byrow = TRUE))
3 t2 <- as.table(matrix(c(0.4, 0.1, 0.2, 0.3), nrow = 2,
4   byrow = TRUE))
5 plot(t(t1))
6 plot(t(t2))

```

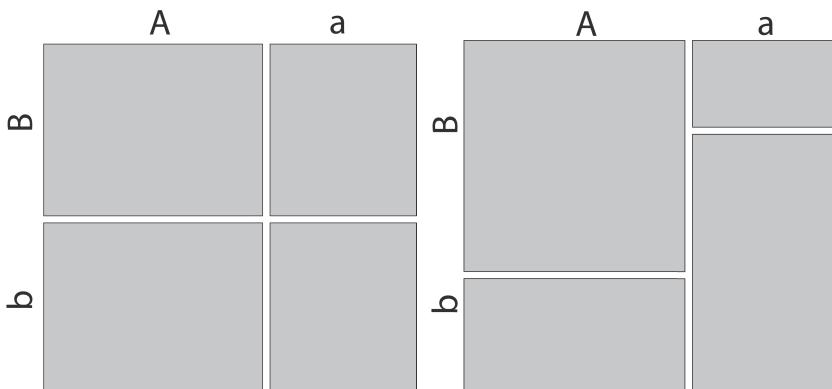


図9.5 比率：左は独立な場合、右は独立でない場合

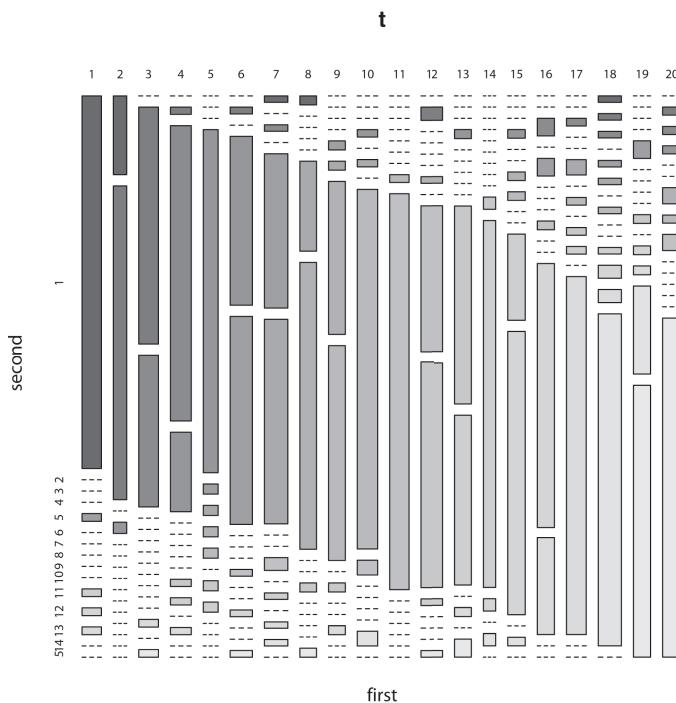
この表と図から、情報を引き出してみます。 $A$  の比率は 0.6、 $a$  は 0.4、 $B$  は 0.5、 $b$  は 0.5 です。どちらの表でも同じです。では、このような 2 種類のラベルを持つ集団から、ある 1 つのサンプルを抜き出したとします。 $"A, B"$ ,  $"A, b"$ ,  $"a, B"$ ,  $"a, b"$  の 4 通りのラベルパターンの取り出される確率は、独立な表の場合には、 $(0.3, 0.3, 0.2, 0.2)$  ですし、独立でない表の場合には  $(0.4, 0.2, 0.1, 0.3)$  です。

今、ある 1 つのサンプルを抜き出して、それが  $B$  か  $b$  かを当てたいとします。何の情報もなければ、 $B:b = 0.5:0.5 = 1:1$  で予想するのがよいでしょう。今、このサンプルは  $A$  であることがわかったとします。このとき、 $B$  なのか  $b$  なのかを予想し直すこととします。表で言えば、 $A$  の列、図9.5で言えば、左側の列に注目すればよいです。独立な場合には、 $B:b = 0.3:0.3 = 1:1$  と予想し

ます。

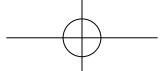
残念ながら  $A$  であることがわかったものの、 $B$ についての予想に変化がありませんでした。独立でない場合には、 $B:b=0.4:0.2=2:1$ と予想します。 $A$ に関する情報がなかったときとずいぶんと予想が変わりました。 $\frac{0.4}{0.4+0.2}=\frac{2}{3}$ です。これは、 $A$ であるという条件のもとでの、 $B/b$ の**条件付確率**です。

今、横軸も縦軸も 2 カテゴリ型で考えました。カテゴリ数を増やしても同じことです。横軸に 20 カテゴリ ( $A=\{A_1, A_2, \dots, A_{20}\}$ )、縦軸に 15 カテゴリ ( $B=\{B_1, B_2, \dots, B_{15}\}$ ) として、 $20 \times 15$  通りのラベルパターンのすべての頻度がわかっているときには、 $20 \times 15$  表がわかっていて、その図（図 9.6）も描けます。そして、そこからサンプリングしたときに、 $A$  の 20 カテゴリのうち



横軸で左側のカテゴリの場合には、縦軸で上側のカテゴリの比率が高く、横軸で右側のカテゴリの場合には、縦軸で下側のカテゴリの比率が高いことが示されています。長方形ブロックの面積は頻度の高さを表しています。

図 9.6 条件付確率：20 カテゴリ×15 カテゴリの場合

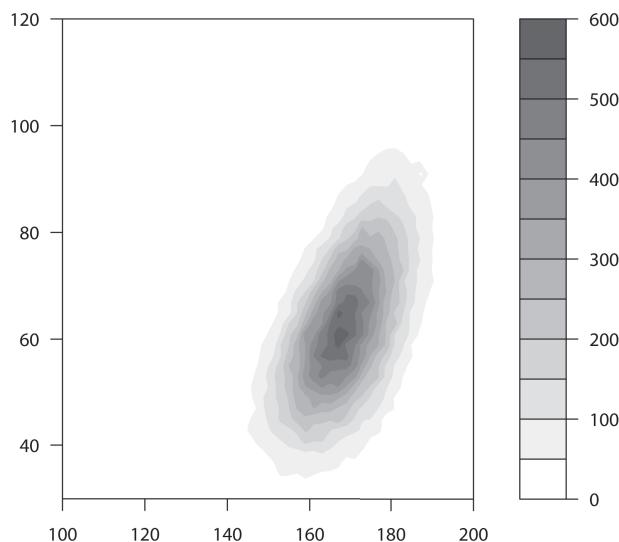


のどれであるか ( $A_i$ ) が判明したときに、 $B$  の 15 カテゴリのうちのどれがどれくらいの確率かは、 $A_i$  という条件付確率としてわかります。縦軸と横軸が独立であるときには、 $A$  に関する情報が判明しても、 $B$  について得られる情報はありませんでしたが、縦軸と横軸が独立でないときには、 $B$  についての情報を増やしてくれる、という点も、 $2 \times 2$  の表の場合と同様です。

### ◎ 9.2.2 量的な軸での条件付確率

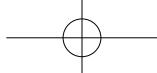
縦軸と横軸について、たまたま、カテゴリ型で話を進めてきました。そうする必要はないかもしれません。縦軸と横軸とに入るべきものは、ある尺度の可能性のあるものすべて（確率変数の空間全体）であればよいです。

たとえば、横軸に身長、縦軸に体重として、図 9.7 に示すような分布があるとき、身長の情報が得られれば、身長の情報がないときよりも、体重がどれくらいかの予想は正確になるでしょう。それは、身長と体重が独立ではないからです。カテゴリ型の場合と同じ理由です。



横軸が身長、縦軸が体重で正の相関があります。

図 9.7 条件付確率：量的変数×量的変数の場合



### ● 9.2.3 事前確率「当初の予想」と陽性的中率 (PPV) と陰性的中率 (NPV)

ここまでの一例では、ある集団が2つの確率変数を持っているときに、2つの確率変数の組み合わせの確率が、縦と横の軸にとってできる長方形部分の面積であるようにグラフ化できました。このようなグラフがあって、2つの確率変数同士が非独立であるなら、片方の変数を観察すると、もう片方の変数についての情報が増えるという話でした。

今度は、少し事情を変えます。2つの確率変数を使って、2軸に対応させる点は同じです。ただし、わかっているのは、片方の変数を条件としたときのもう片方の変数の条件付確率だけです。

2つの変数が  $A/a, B/b$  の2カテゴリ型とします。 $B$  であるときに  $A$  である確率（条件付確率）を  $Pr(A|B)$  とし、それ以外の場合も、同様に  $Pr(a|B), Pr(A|b), Pr(a|b)$  とします。これらがわかっています。たとえば、 $Pr(A|B) = 0.8, Pr(a|B) = 0.2, Pr(A|b) = 0.4, Pr(a|b) = 0.6$  とします。 $A/a$  が男女、 $B/b$  が1年生と2年生とします。1年は男が8割、2年は男が4割です。今、1年生と2年生の比率が0.5対0.5とすると、

	$A$	$a$	計
$B$	0.4	0.1	0.5
$b$	0.2	0.3	0.5
計	0.6	0.4	1

という表ができます。このときに、ある人が男か女かという情報が得られれば、その人の学年に関する予測の精度が上がるるのは前項までと同じです。男( $A$ )と観察したら、1年生の確率は、 $Pr(B|A) = \frac{0.4}{0.6} = \frac{2}{3}$  です。

もしも、1年生と2年生の比率が0.8対0.2のときには、表が変わって、

	$A$	$a$	計
$B$	0.64	0.16	0.8
$b$	0.08	0.12	0.2
計	0.72	0.28	1

となります。表は変わりましたが、1年生 ( $B$ ) の男女比、2年生 ( $b$ ) の男女

比は変わっていません。この表から、男女を観察したときに学年を予測することもできます。この表から、男である ( $A$ ) と観察したとき、1年生である ( $B$ ) 確率は、 $Pr(B|A) = \frac{0.64}{0.72} = \frac{8}{9}$  です（図9.8）。

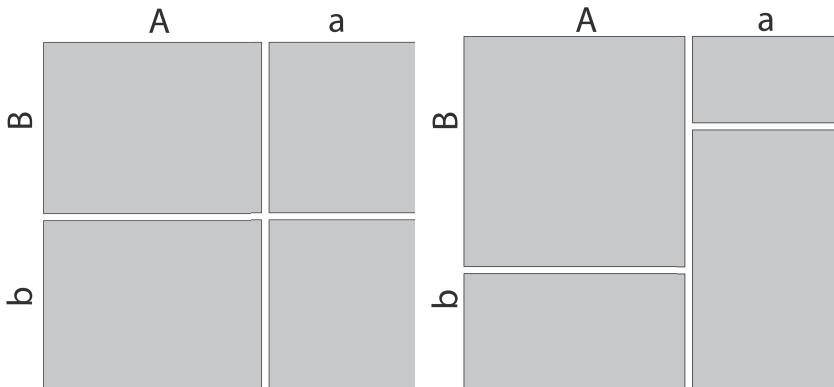


図9.8 比率：左は  $B:b=0.5:0.5$ 、右は  $B:b=0.8:0.2$

2つの表は、 $A/a$  の比率を変えたことで、長方形の中身が変わり、横軸に関する条件付確率は変わっていないけれども、縦軸に関する条件付確率が変わったわけです。

このように、観察前の  $B/b$  の比率（初めの例での  $0.5/0.5$ 、2つ目の例での  $0.8/0.2$ ）を、**事前確率**と言い、それを用いて作った表からわかる  $Pr(B|A)$ ,  $Pr(b|A)$ ,  $Pr(B|a)$ ,  $Pr(b|a)$  を**事後確率**と言います。

医療の現場では、異なった用語を使います。 $A/a$  を検査が「陽性／陰性」として、 $B/b$  を病気で「ある／ない」とします。表を作るために必要だった  $Pr(A|B)$ （疾患ありの場合の検査陽性率）,  $Pr(a|B)$ ,  $Pr(A|b)$ ,  $Pr(a|b)$ （疾患なしの場合の検査陰性率）は、 $Pr(B|A)$  が検査の**感度**、 $Pr(b|a)$  が検査の**特異度**と呼ばれます。また、 $B/b$  の事前確率は、検査前に集めた症状などの情報から予測した病気である予想確率です。そして、 $B/b$  の事後確率は、 $Pr(B|A)$  が「検査が陽性だったときに病気であると考える確率」で、これが**陽性的中率 (PPV)** です。また、 $Pr(b|a)$  が「検査が陰性だったときに病気でないと考える確率」で、これが**陰性的中率 (NPV)** です。

ここでは、縦横2軸に2カテゴリ型変数を用いましたが、カテゴリ数が増え

たり、量的変数にしたりしても、考え方は同じです。

### (1) 仮説空間と観測空間の条件付確率

さて、検査と病気の $2 \times 2$ 表では、列についての処理は検査結果がわかったときの病気のあり・なしの尤度を表し、行についての処理は病気のあり・なしがわかっているときの、検査の陽性・陰性の確率を表していました。袋からの玉の抜き取りのときに作った、確率と尤度のグラフを見直してみましょう（図9.3）。列に確率が、行に尤度が出ていたのと同じ構図です。

ここまでの一例で、縦軸と横軸にとったものがどういうものだったかを確認します。分割表の一例では、あるカテゴリ変数のとりうるカテゴリすべてを縦軸に、もう1つの確率変数のカテゴリすべてを横軸にとりました。

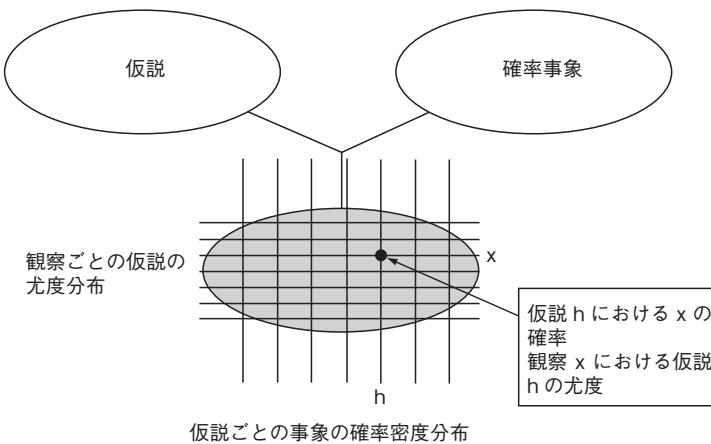
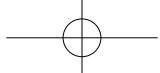


図 9.9 確率と尤度と仮説と観察との関係

確率と尤度とは、図9.9に示すように、同じグラフを縦軸を基準にするか横軸を基準にするかの違いを反映したものです。身長・体重の一例では、縦軸に体重としてとりうる値の範囲のすべてをとり、横軸に身長としてとりうる値の範囲のすべてをとりました。そして、縦軸と横軸の間に非独立な関係があることを利用して、片方の観察によって、もう片方に関する情報を改善しました。

2つの確率変数を取り上げて、それぞれの確率空間を軸においてやり、片方の確率変数を観察することによって、もう片方の確率変数に関する情報を改善すること、とまとめられます。



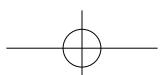
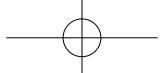
2種類の玉が入った袋から抜き出して、袋の中の玉の内訳を当てる、という作業は、この考え方でいくと、次のように言えます。

- 横軸に、袋の中の玉の比率のとりうる範囲をとる
- 縦軸には、取り出すという試行で起こりうる事象の場合をとる

袋の中の玉の比率と取り出す事象の起こる確率は、もちろん関連しているので、縦軸の観察によって横軸の情報が改善するわけです。袋から玉を取り出すことを繰り返して、袋の中の玉の比率を予測する作業は、次のような作業であることがわかります。

まず、玉の比率の事前確率を考えます。次に観察をして、事後確率を得ます。そして、2回目の取り出しにあたっては、事後確率を事前確率と考え直します。1回目の観察によって事前確率が変わったため、縦軸と横軸で作られた長方形部分の分布を変えます。2回目の観察をします。事後確率が得られます。もう一度取り出すなら、事前確率を事後確率に置き換えてから取り出しを……

という繰り返しです。



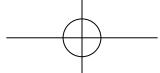
# 第 10 章

## 連鎖解析に見る尤度と変数

### 10.1 尤度を使った形質マッピング——連鎖解析

尤度の計算は、とりうる仮説のすべてについて、観察データをもたらす確率を計算することから始まります。連鎖解析とは、DNA配列上にある、形質の原因変異の位置を探索する手法の1つで、数多くの遺伝因子の同定に成功してきた手法ですが、家系情報とフェノタイプ情報とジェノタイプ情報とから、尤度を計算する解析手法です。連鎖解析では非常に多くの仮説について尤度を計算する必要があります。この章では、尤度を利用することの一環として、その手法の概要を見てみることにします。

連鎖解析は大きく2つに分けられ、パラメトリック手法とノンパラメトリック手法があります。対象とする形質としては、前者が比較的少数の大家系に認められる強い遺伝因子に向いているのに対して、後者は比較的弱い遺伝因子を小規模家系を多数集めて解析するに向いています。また、前者は尤度を計算すべき仮説空間が広く、込み入っていますので、尤度についての理解を深める題材として適当であり、後者は遺伝因子を変数化することについて考えるのに好適と思われますので、それぞれ、順を追って説明することとします。



## 10.2 パラメトリック連鎖解析と尤度

### ◎ 10.2.1 マーカーの伝達木と原因座位の伝達木

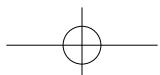
**パラメトリック連鎖解析**では、染色体の家系内での受け渡しと組み換えとのパターンをすべて数え上げます。そして、フェノタイプの情報とジェノタイプの情報を利用して、どの染色体のどの位置に原因因子が存在していると考えることがもっともらしいかを調べます。そして、その原因因子の存在場所としてもっともらしい箇所が、どれくらい原因因子のありかとして信憑性があるかを数値で示します。

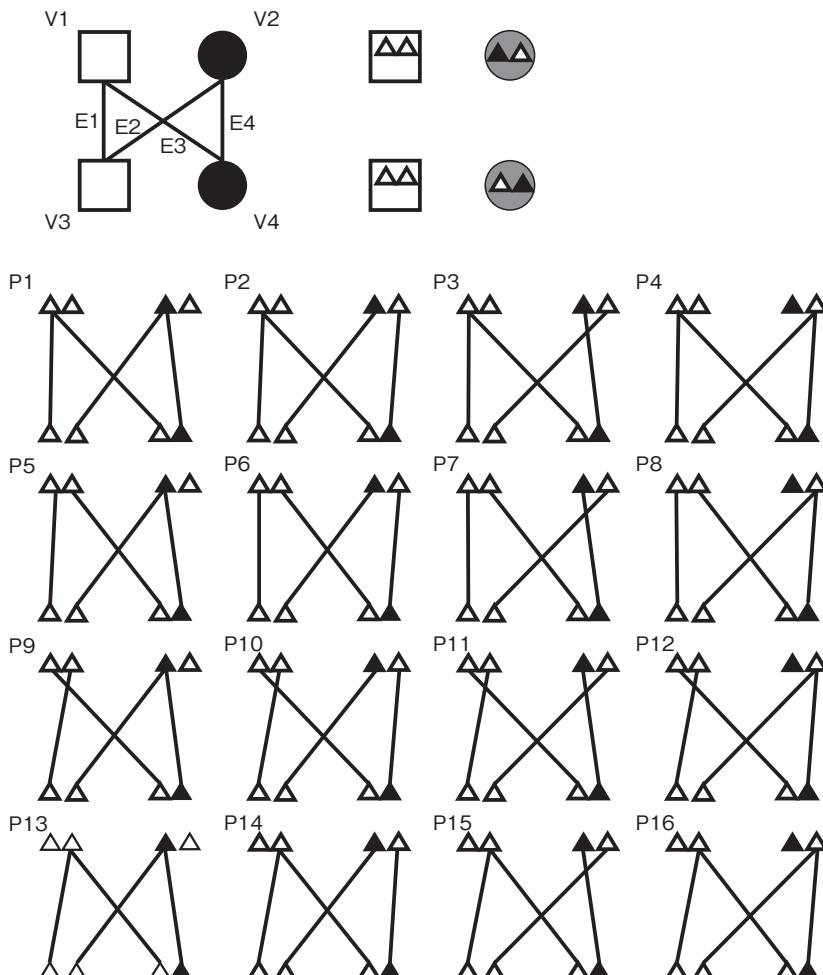
まず、染色体の家系内での受け渡しと組み換えのパターンとはどういうことかを見てみましょう。家系図は個人のつながりでしたが、それは、染色体の伝達のグラフを中に隠し持っているものだ、ということは第5章で述べました。そして、染色体上の1点について限れば、必ず木が描けることも確認しました。

今、家系図が与えられたとき、可能性のある木のパターンは、家系図での辺の数（伝達の数） $n$ に対して $2^n$ あります。図10.1の例では、4人はV1, V2, V3, V4の4点と、それを結ぶ4本の辺 E1, E2, E3, E4とでできているグラフを構成します。染色体・アレルに着目すれば、各個人には2つの要素があります。染色体・アレルの伝達関係でグラフにすると、辺1本につき、2通りの引き方があるので、 $2^{\text{辺の数}} = 16$ パターンが作れます。これが、ゲノム上の1箇所に関する、木の形の数です。図10.1では16パターンのすべてを示しています。

ゲノム上のすべての箇所は、同様に $2^n$ パターンの可能性があります。この木のパターンを行列で表すことにします。染色体・アレルのうち、父親から受け取ったものを0、母親から受け取ったものを1として、図に表すときには、父由来のそれを左に母由来のそれを右に置くことにしましょう。図の染色体の伝達パターンの左上隅のパターン（P1）は、(E1, E2, E3, E4)=(0, 0, 0, 0)と表せて、その隣のパターン（P2）は(0, 0, 0, 1)と表せます。

Rでパターンを作ってみます (R10-1.R)。





優性遺伝形式で父親と娘が表現型陽性。左上は個人のつながりを表すグラフで、右上は染色体のアレルのパターンをします。下段には P1, P2, ..., P16 の染色体伝達パターンが示されています。すべての伝達パターンを示していますので、色の違う染色体が辺で結ばれている場合も含んでいます。

図 10.1 両親と 2 人の子の例

R ソース 10.1 R10-1.R : 順列

```

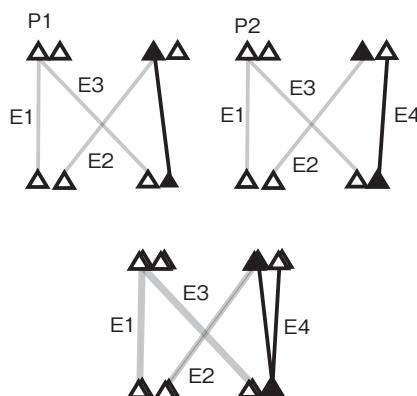
1 library(gtools)
2 permutations(2,4,c(0,1),repeats=TRUE)
# 長さ 4 のベクトルを作る。抜き出し元は要素数 2 個で、それは c(0,1)。
# 抜いては戻し (repeats) を TRUE で実行する

```

### ■ R の出力結果

	[,1]	[,2]	[,3]	[,4]
[1, ]	0	0	0	0
[2, ]	1	0	0	0
[3, ]	0	1	0	0
[4, ]	1	1	0	0
[5, ]	0	0	1	0
[6, ]	1	0	1	0
[7, ]	0	1	1	0
[8, ]	1	1	1	0
[9, ]	0	0	0	1
[10, ]	1	0	0	1
[11, ]	0	1	0	1
[12, ]	1	1	0	1
[13, ]	0	0	1	1
[14, ]	1	0	1	1
[15, ]	0	1	1	1
[16, ]	1	1	1	1

全部で、 $L$ 箇所のことを考えれば、 $16^L$ と、膨大な数になります。2箇所(M1, M2) ( $L=2$ )で考えます。親子4人という小さい単位で考えます。M1は、この16パターンのどれかをとります。M2も同様です。したがって、すべての可能性を考えると  $16^2$  のパターンを考慮すればよいです。M1がP1パターンで、M2がP2パターンだったとします。



下段のE4(右側の親から右側の子への伝達)が、M1の木では右親の▲からの辺であるのに對して、M2の木では右親の△からの辺となっており、V字を作っています。

図 10.2 上段は M1, M2 の木、下段はそれを重ね合わせた図

辺 E4 に相当する部分で、V の字が生じています。これは、組み換えが起きたことを表しています。(E1, E2, E3, E4) の表し方でいうと、P1 は (0, 0, 0, 0)、P2 は (0, 0, 0, 1) です。P1 と P2 の間のマンハッタン距離が 1 なので、それがこの 4 人家族での組み換えの伝達数であることを確認してください。全部の伝達（全部の辺）で組み換えが起きたときが、組み換え回数の最大値で、この図の例ではそれは 4 です。組み換えの起きていない伝達数はこの図の場合  $4 - 1 = 3$  です。

R で M1 の 16 パターン対 M2 の 16 パターン ( $16^2$  パターン) のすべてについて、組み換えありの伝達の数と組み換えなしの伝達の数を表す行列を作つてみます。

#### R ソース 10.2 R10-2.R : 組み換えの数

```

1 n<-4
2 m<-permutations(2,4,c(0,1),repeats=TRUE)
3 RecNumberMat<-as.matrix(dist(m,method="manhattan",diag=TRUE,
      upper=TRUE)) # 隣接ツリー間の組み換え回数
4 NonRecNumberMat<-n-RecNumberMat # 隣接ツリー間の非組み換え回数
5 RecNumberMat

```

組み換えあり伝達数行列は次のようになります。1 行目と 1 列目は行番号・列番号を表しています。

#### ■ R の出力結果

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	1	1	2	1	2	2	3	1	2	2	3	2	3	3	4
2	1	0	2	1	2	1	3	2	2	1	3	2	3	2	4	3
3	1	2	0	1	2	3	1	2	2	3	1	2	3	4	2	3
4	2	1	1	0	3	2	2	1	3	2	2	1	4	3	3	2
5	1	2	2	3	0	1	1	2	2	3	3	4	1	2	2	3
6	2	1	3	2	1	0	2	1	3	2	4	3	2	1	3	2
7	2	3	1	2	1	2	0	1	3	4	2	3	2	3	1	2
8	3	2	2	1	2	1	1	0	4	3	3	2	3	2	2	1
9	1	2	2	3	2	3	3	4	0	1	1	2	1	2	2	3
10	2	1	3	2	3	2	4	3	1	0	2	1	2	1	3	2
11	2	3	1	2	3	4	2	3	1	2	0	1	2	3	1	2
12	3	2	2	1	4	3	3	2	2	1	1	0	3	2	2	1
13	2	3	3	4	1	2	2	3	1	2	2	3	0	1	1	2
14	3	2	4	3	2	1	3	2	2	1	3	2	1	0	2	1
15	3	4	2	3	2	3	1	2	2	3	1	2	1	2	0	1
16	4	3	3	2	3	2	2	1	3	2	2	1	2	1	1	0

さて、M1, M2 のジェノタイプデータがあると、M1, M2 のそれぞれについて、16パターンの条件付確率（ジェノタイプデータという条件）がわかります。今、M1 では、図 10.1 のように V1, V2, V3, V4 のジェノタイプが、(0, 0), (1, 0), (0, 0), (0, 1) だったとします。図 10.1 の伝達パターンでは、P1 と P3 はありますですが、P2 と P4 はありません。このように、ジェノタイプの情報から、伝達の木にはある木とありえない木があります。ありえないパターンは確率が 0 で、あるいはパターンはみな同じ条件付確率を持ちます。

一方、病気の原因座位 (G) については、別の理由から木の条件付確率が決まります。こちらの理由は、家系の表現型です。今、病気の遺伝形式を決めるに、フェノタイプのパターンを起こしうる伝達パターンと起こしえないパターン、より起こしやすいパターンの条件付確率が決まります。優性遺伝形式であるとか、劣性遺伝形式であるとかを仮定します。このとき、ジェノタイプによってフェノタイプが確実に決まると言います<sup>\*1</sup>。ジェノタイプが決まるとフェノタイプが決まるとしても、その逆は必ずしも正しくはありません。優性遺伝形式のときは、フェノタイプが出ているときに、ホモ接合体かもしれないし、ヘテロ接合体かもしれないからです。ですから、フェノタイプがわかつていて、遺伝形式がわかつているときには、個人の「原因遺伝子のジェノタイプ」を確率的に割り当てます<sup>\*2</sup>。確率的にではありますが、原因遺伝子のジェノタイプがわかつたので、原因遺伝子の伝達パターンとして、どれがどれくらいありそうかも確率的に決まります。

さて、ここまでで、M1, M2 におけるあるいは伝達パターンと、原因座位 (G) の伝達パターンとがわかりました。G は M1, M2 の間にあるかもしれないを考えます。「Ma, Mb, G の伝達パターンがそれぞれ、Pi, Pj, Pk である場合」があるいはかどうかもわかります。「Pi, Pj, Pk」の 3 つ組が、起きやすい組み合わせかそうでないかは、Pi と Pk の間と Pj と Pk の間で組み換え数と非組み換え数がそれぞれ何回ずつかに応じて決まります。

Ma と G の間で組み換えが起きる確率が  $\theta_{aG}$ 、同様に G と Mb の間のそれが  $\theta_{Gb}$  とします。組み換えが起きる確率は、0 から 0.5 の値をとります。0.5 の場合は相互に独立な場合で、異なる染色体に乗っているような場合とも言えま

<sup>\*1</sup> このようにジェノタイプが決まるとフェノタイプが完全に決まることを、**浸透率**が 1 である、と言います。

<sup>\*2</sup> 浸透率が 1 ではないときには、それに応じて、個々人の「原因遺伝子のジェノタイプ」を確率的に割り当てる必要です。

す。 $P_i$  と  $P_j$  の間で、組み換えが起きた伝達数と組み換えの起きなかった伝達数を  $N_{rec_{i,j}}$ ,  $N_{non_{i,j}}$  とすれば、

$$L(P_i, P_j, P_k) = \theta_{aG}^{N_{rec_{i,k}}} \times (1 - \theta_{aG})^{N_{non_{i,k}}} \times \theta_{Gb}^{N_{rec_{k,j}}} \times (1 - \theta_{Gb})^{N_{non_{k,j}}}$$

が尤度です。

これにより、「ありそうな  $P_i$ ,  $P_j$ ,  $P_k$  の組み合わせ」と「ありそうでない  $P_i$ ,  $P_j$ ,  $P_k$  の組み合わせ」に高低がつきました。組み換えあり伝達数が多い区間は、距離が長い方が尤度が上がりますし、組み換えなし伝達数が多い区間は、距離が短い方が尤度が上がるからです。これをすべての  $P_i$ ,  $P_j$ ,  $P_k$  の組み合わせについて足し合わせます。考慮するべきのは  $M1$  と  $G$  の間と  $M2$  と  $G$  の間です。足し合わせるときには、 $P_i$ ,  $P_j$ ,  $P_k$  の条件付確率 ( $M1$ ,  $M2$  のジェノタイプ条件と  $G$  のフェノタイプ条件) を  $p_i$ ,  $p_j$ ,  $p_k$  としてこれらを考慮します。

$$L = \sum_{\text{すべての } (P_i, P_j, P_k) \text{ の組み合わせ}} l(P_i, P_j, P_k) p_i p_j p_k$$

これが、すべての可能性のあるパターンに関して得られる尤度です。家族4人の例では、全部で  $16^3$  通りの  $(P_i, P_j, P_k)$  の組み合わせがあります。

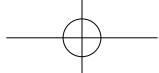
### ◎ 10.2.2 マーカーと原因座位の間の組み換え

今、 $M1$  と  $M2$  の間の組み換え確率  $\theta_{12}$  は、マーカーと原因座位の間の組み換え確率  $\theta_{1G}$ ,  $\theta_{2G}$  と次のような関係にあります。

$$\theta_{12} = \theta_{1G}(1 - \theta_{2G}) + \theta_{2G}(1 - \theta_{1G})$$

$M1-G$  間で組み換えが起きて、 $M2-G$  間で起きなかった場合と、 $M1-G$  間では起きずに  $M2-G$  間で起きた場合との和が、 $M1-M2$  間で起きた場合だからです。少し工夫をして

$$\begin{aligned} \theta_{1G} &= 0.5 - \sqrt{\frac{0.5 - \theta_{12}}{2} \frac{\cos(t)}{\sin(t)}} \\ \theta_{2G} &= 0.5 - \sqrt{\frac{0.5 - \theta_{12}}{2} \frac{\sin(t)}{\cos(t)}} \end{aligned}$$



のように変数  $t$  の関数で表すことができます。ここで、 $t$  は  $\theta_{1G}, \theta_{2G}$  が 0 以上になるような範囲にとります。 $t$  を変化させてやって、

$$L = \sum_{\text{all } Pi, Pj, Pk} l(Pi, Pj, Pk) pi pj pk$$

を最大になるような  $t$  を探すことが連鎖解析です。

考えなくてはいけないパターンはたくさんありました。場合の数が多いだけです。唯一、知りたくて、変化させたいのは  $G$  の位置です。そして観察されたジェノタイプとフェノタイプをもたらす尤度を最大にする  $G$  の位置を知ることが目的です。 $G$  の位置は  $t$  で変数化しましたから、この値をたくさん調べて、最大の尤度が出る値を見つけてやりましょう。 $G$  の位置を動かすと、 $Ma$  と  $G$  の間と  $Mb$  と  $G$  の間で組み換えあり・なし伝達が多い方がよいのか、少ない方がよいのかの値が変わるために、尤度が変化します。

実際の連鎖解析では、ゲノム上のマーカーを用います。それは、マーカー間の組み換え割合 ( $\theta_{12}$ ) がわかっているからです。その値と、 $G$  の位置 ( $t$ ) から、 $M1-G, M2-G$  の間の組み換え割合が決まりますから、 $G$  がその場所だったときの尤度をすべてのパターンに関して計算してやりましょう。

残念ながら 4 人家族では、伝達に関する情報が大変少ないので、試しに計算するにしてもつまらないで、次のようにして、勝手にありえる伝達パターンを作ってみます。作り方の考え方としては、 $M1, M2$  のマーカーの伝達パターンとして「ありえる」か「ありえないか」の二者択一を適当に割り振ります。 $G$  のそれについては、 $M1, M2$  の伝達パターンと似ていることが、 $G$  がこのあたりに存在するということなので、両者の伝達パターンに似させることとし、浸透率などの影響で、どのパターンも可能性があるようになります。

まずは、尤度を計算して、 $G$  の位置に関してプロットしてみましょう。ソースは少し長くなりますが、コメントを頼りに処理を追いかけてみてください。

#### R ソース 10.3 R10-3.R : 連鎖解析の尤度計算

```
1 # 以下の引数を用いて対立仮説と帰無仮説の尤度を計算
2 # n: 木の枝の数
3 # m1,m2,G: 2 マーカーの木の確率分布とジェノタイプが決める木の確率分布
4 # theta: マーカー間の組み換え率
5 # k: マーカー間の尤度計算地点数
```

```

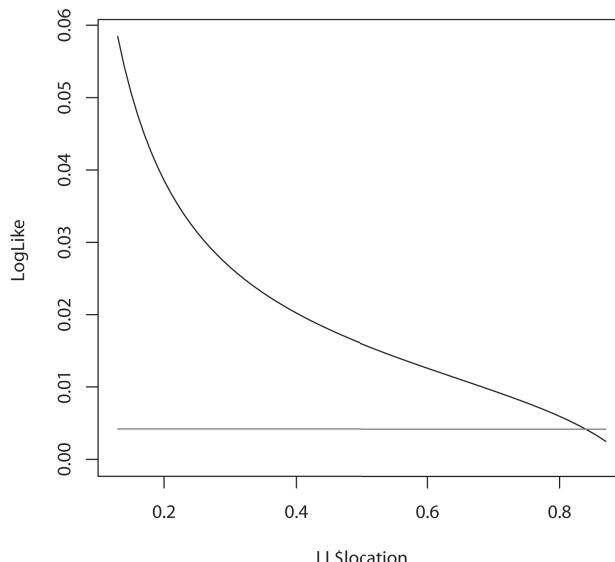
6 CalcLike<-function(n,m1,m2,G,theta,k) {
7   x<-seq(from=0,to=1,by=1/k)
8   t<-x*pi/2
9   theta1<-sqrt((0.5-theta)/(2*sin(t))*cos(t))+0.5
10  theta2<-sqrt((0.5-theta)/(2*cos(t))*sin(t))+0.5
11  range<-which(theta1>=0 & theta2>=0)
12  x<-x[range]
13  theta1<-theta1[range]
14  theta2<-theta2[range]
15  library(gtools)
16  trvec<-permutations(2,n,c(0,1),repeats=TRUE)
17  #隣接ツリー間の組み換え回数と非組み換え回数
18  RecNumberMat<-as.matrix(dist(trvec,method="manhattan",
19    diag=TRUE,upper=TRUE))
20  NonRecNumberMat<-n-RecNumberMat
21  Lalt<-rep(0,length(x)) # 対立仮説の場合の尤度を格納
22  for(i in 1:length(Lalt)){
23    m1G<-m1%*%t(G) #Ma と Gとの確率の積
24    #組み換えあり伝達数を考慮
25    m1Gx<-m1G*theta1[i]^RecNumberMat*
26      (1-theta1[i])^NonRecNumberMat
27    m1Gx<-apply(m1Gx,2,sum)
28    # $(2^n)^2$  の行列から、Gのパターン  $(2^n)$  パターンに集約
29    m2G<-m1Gx%*%t(m2) #Mb の確率との積
30    m2G<-m2G*theta2[i]^RecNumberMat*
31      (1-theta2[i])^NonRecNumberMat
32    # G と Mb との組み換えあり伝達数を考慮
33    Lalt[i]<-sum(m2G) # 全パターンについて足し合わせ
34  }
35  # 帰無仮説の場合
36  m12<-m1%*%t(m2)
37  m12x<-m12*theta1^RecNumberMat*
38  (1-theta1)^NonRecNumberMat/(2^n)
39  Lnull<-rep(sum(m12x),length(Lalt))
40  list(logLikeAlt=Lalt,logLikeNull=Lnull,location=x)
41  n<-4 # 木の枝の数
42  set.seed(65432) # 適当なデータを作るために疑似乱数列の種を与える
43  m1<-sample(c(0,1),n^2,replace=TRUE,prob=c(0.8,0.2))
44  m1<-m1/sum(m1) # マーカー 1 での木パターンの確率分布
45  m2<-sample(c(0,1),n^2,replace=TRUE,prob=c(0.8,0.2))
46  m2<-m2/sum(m2) # マーカー 2 での木パターンの確率分布

```

```

42 G<-0.9*m1+0.1*m2
43 G<-G/sum(G) # ジエノタイプが決める木パターンの確率分布
44 theta<-0.4 # 2マーカー間の組み換え率
45 k<-100 # マーカー間を100等分した場所で尤度を計算
46 LL<-CalcLike(n,m1,m2,G,theta,k)
47 ylim<-c(min(LL$logLikeAlt,LL$logLikeNull),
        max(LL$logLikeAlt,LL$logLikeNull))
48 ylim<-c(0,max(LL$logLikeAlt,LL$logLikeNull))
49 plot(LL$location,LL$logLikeAlt,type="l",ylim=ylim)
        # 対立仮説の尤度のグラフ
50 par(new=T)
51 plot(LL$location,LL$logLikeNull,type="l",col="red",
      ylim=ylim) # 帰無仮説の尤度のグラフ

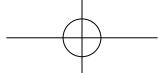
```



横軸が G の位置 (変数 t)、縦軸が尤度、水平線は帰無仮説における尤度

図 10.3 連鎖解析の尤度計算

このようにして計算した尤度が最も大きくなる位置に、G があると予想します。次に、この G の予想位置が、信じるに足るものなのかの判断をしないといけません。G がどのマーカーとも同じ染色体上に乗っていないとした場合（帰無仮説）の尤度と較べて、十分に高い尤度が得られていたら、信じる



に足るとみなします。図 10.3 では、水平線が帰無仮説での尤度です。尤度は比で判断すればよいので、帰無仮説の場合の尤度に較べて、10 の何乗倍尤度が高いかを数値で表すことが多く、それをロッドスコア (LOD, logarithm of odds) と呼びます。

実際の計算では、木のパターンの組み合わせの網羅をどういう手順で計算するか、とか、木のパターンの組み合わせとはいっても、全マーカーでの組み合わせはせず、近いところのマーカーについてのみパターンの組み合わせを考慮するなどの実際上の工夫がなされます。

### ◎ 10.2.3 隠れマルコフモデルと連鎖解析の尤度計算

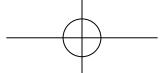
2つのマーカーに挟まれた座位の位置決めをするために座位数 3 個（マーカー 2 個、原因座位 1 個）に関する伝達パターンと、位置をずらす変数との 4 変数が作る場合をすべて網羅するのは気が利かないでの、気を利かせる方法を考えてみます。**隠れマルコフモデル**と呼ばれる方法です。

M1, G, M2 と並んだ座位について考えます。M1 の伝達パターンから、G のパターンへ変化すると考えます。M1 のパターンは複数あって、G のパターンも複数あります。M1 のどのパターンからも、G のすべてのパターンへと変化する可能性はあります。問題は、M1 と G との間の組み換え割合の多寡によって、移りやすいパターンと移りにくいパターンがあることです。M1 の  $2^n = 16$  パターンと G の  $2^n = 16$  パターンの組み合わせで、 $16^2$  のパターンを考える必要があります。

次に、G から M2 への変化を考えます。このとき、 $M1 \rightarrow G$  のパターン  $16^2$  通りのすべてについて、 $M2$  の  $2^n = 16$  パターンへの推移を考えるわけで、このようにすると、 $16^3$  通りを考えなくてはなりません。しかしながら、G と M2 との関係（組み換え割合と伝達パターンの変化）は、G のパターンにはよりますが、M1 のパターンにはよりません<sup>※3</sup>。

したがって、M1 と G の  $16^2$  パターンのうち、G のパターンが共通するパターン（それは 16 パターンあります）は M2 との関係において、同じに扱って

※3 実際の染色体の組み換えでは、Ma と G の間で組み換えが起きていると、G と Mb の間で組み換えは起きにくくなりますが（紐をねじるときに、続けざまにねじるには力がいります）、 $16^3$  回のループを回して計算しているときにも、そこまで考慮していないので、今回も考慮するには及びません。



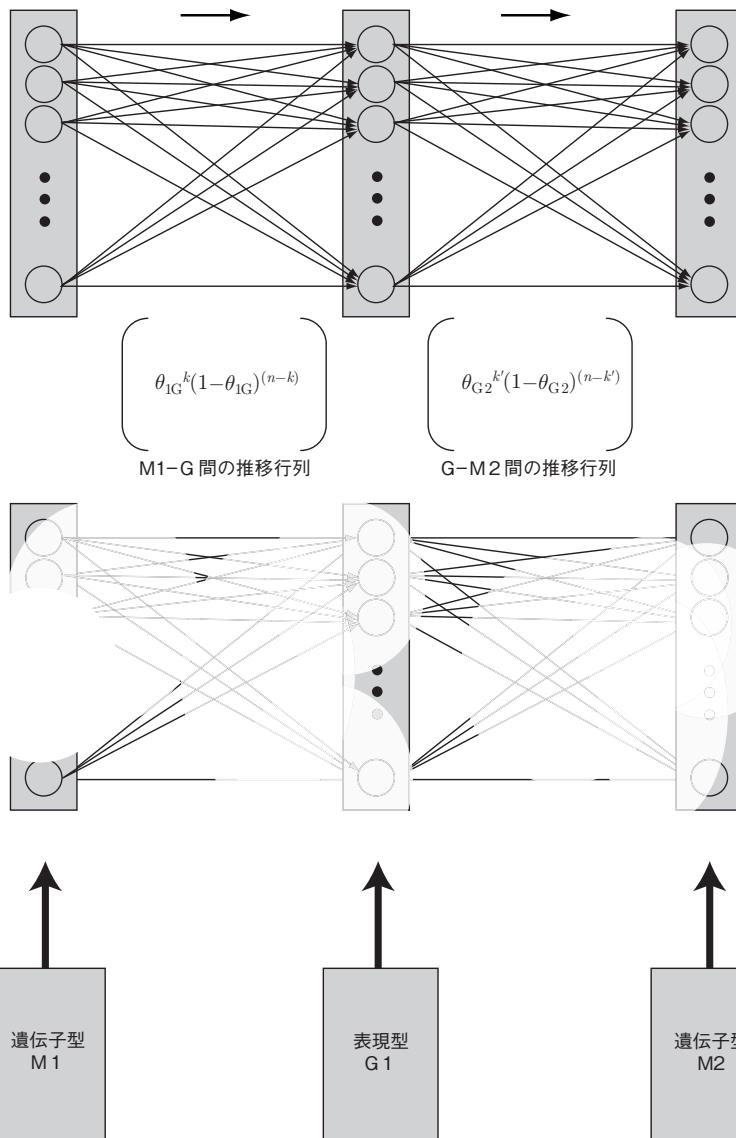
しまうことができます。同じに扱うということは、まとめてしまうわけです。 $16 \times 16$  行列から、G のパターンが同じものを合算して、長さ 16 のベクトルにしてやります。その上で、Mb の 16 パターンとの関係を組み換えに考慮して計算すればよいです。この方法のよいところは、増えたパターンを元のパターン数に戻してから、次のステップに進めることで、これならば、次から次へと処理を続けていくときに、同じことを繰り返すだけで済みます。

このように、順番に移り変わっていくときに、次のステップへの変化は、現在の状態のみから（確率的に）決められるような移り変わりを**単純マルコフ連鎖**と言います<sup>※4</sup>。

座位のアレルの伝達パターンの塩基配列上での移り变わりは、座位間の組み換え割合によって決まりますので、マルコフ連鎖です。マルコフ連鎖の定義にあるように、「次のステップに関係しない前段階」はあってもなくても、「次のステップ」について決められます。それを利用して、パターンが増えた後で、少ないパターンに戻してやることができました。

今、観察しているのは、マーカーのジェノタイプや個人のフェノタイプです。これらは、「隠れて」いないので観察できます。一方、最も興味のある原因座位 G の位置はもちろん、観察することはできませんし、その位置のありそうな場所を知るための情報である、各座位の  $2^n$  の伝達パターンと、座位間のパターン同士をつなぐ組み換えのあり・なしも、観察することができません。「隠されて」いるのです。「隠れていない」観察データから、興味の対象である「隠された」状態の推移を探索する作業なので、「隠れ（た）マルコフ連鎖」と呼ばれます。

※4 単純でないマルコフ連鎖として次のステップへの変化が、現在とそれより前の何段階の状態から決まるような移り変わりが定義されますが、いずれにしろ、次のステップが限定された前段階のみから決められるようなものがマルコフ連鎖です。



3座位のそれぞれに  $2^n$  の伝達パターンがあり、左から右へと伝達パターンが推移します。パターン間の推移のしやすさは、組み換えあり伝達数 ( $k$ (M1-FG 間の組み換えあり伝達数)、 $k'$ (G-M2 間の組み換えあり伝達数)) と組み換えなし伝達数 ( $n - k$ ,  $n - k'$ ) と座位間の組み換え割合 ( $\theta$ ) で決まります。推移のしやすさは、パターン数×パターン数の行列です。2番目の座位のパターンと3番目の座位のパターンとの推移には、1番目の座位のパターンは関係していません。伝達パターンはすべて推定されたもので、観察できないので「隠れ」ています。

図 10.4 伝達パターン

## 10.3 ノンパラメトリック連鎖解析——罹患同胞対解析

### ◎ 10.3.1 相対危険度を変数とする

連鎖解析では、伝達パターンについて推定することを基本にします。10.2節では、遺伝形式や浸透率などをパラメタとして仮定して尤度を計算しました。そしてその尤度に基づいて、原因座位の位置を探しました。本節では、**ノンパラメトリック**という名前が示す通り、遺伝形式や浸透率を仮定せずに解析を進める方法を扱います。こちらの方法では、ジェノタイプがフェノタイプに影響しているか否か、影響しているとしたら、その強さはどれくらいかを変数とします。そして、影響がないとみなすよりも、あるとみなす方がもっともらしい位置をゲノム上から探します。

このノンパラメトリック連鎖解析手法は、ある病気を同胞（兄弟姉妹）が揃って発病しているペア（**罹患同胞対**）を収集して解析する研究で用いられる例が多いので、それに沿って話を進めます。では、ジェノタイプがフェノタイプに影響している強さを変数にする方法と、それが、伝達パターンとどのような関係にあるかを見ていくことにします。

ある座位のあるアレルが、あるフェノタイプを起こしやすくしているとします。そのアレルを A、それ以外のアレルを a とすると、3つの遺伝子型 AA, Aa, aa が考えられます。それぞれのジェノタイプのときに、 $R_2$ ,  $R_1$ ,  $R_0$  の確率でフェノタイプを起こすとします。この座位がフェノタイプと関係がなければ  $R_2 = R_1 = R_0$  です。

罹患同胞対のジェノタイプの組み合わせは、 $3^2 = 9$  通りあります。それぞれの場合で、同胞対が揃って病気である確率は、1人目のジェノタイプから決まる病気の確率と2人目のジェノタイプから決まる病気の確率の積です。表にすると以下の通りです。

	AA	Aa	aa
Risk	$R_2$	$R_1$	$R_0$

	AA	Aa	aa
AA	$R_2^2$	$R_2 \times R_1$	$R_2 \times R_0$
Aa	$R_1 \times R_2$	$R_1^2$	$R_1 \times R_0$
aa	$R_2 \times R_0$	$R_1 \times R_0$	$R_0^2$

### ● 10.3.2 IBD 数ごとに場合分けする

ここで、同胞対について伝達パターンを考えます。1人につき親子2人からの伝達があるので、伝達数は4つあります。4つの伝達が、それぞれ親の持つ2つの染色体のどちらからの伝達かの区別をするので、 $2^4=16$ の伝達パターンがあるのでした。それを $4 \times 4$ の行列で表してみます。父親の2つの染色体をF1, F2、母親のそれをM1, M2と書くことにします。同胞の1人にF1M1, F1M2, F2M1, F2M2の4パターンがあります。同胞2人の1人目を縦軸に、2人目を横軸にして、各軸に4パターンを対応させます。

ここで、同胞対が揃ってフェノタイプを持っている場合について考えるためには、同胞対のジェノタイプを考慮する必要があります。同胞対のジェノタイプを考慮するためには、親が持つ2アレルがそれぞれ、リスク型かそうでないかの区別が必要ですから、そのことを考えます。

両親の4染色体の中に、リスク型（A）がいくつあるかで分類します。0, 1, 2, 3, 4の5通りあります。0の場合、子どもは必ず非リスク型のホモ接合体（aa）です。1の場合、子どもは非リスク型のホモ接合体（aa）か、ヘテロ接合体（Aa）です。2の場合、その2本を片親が持っている場合には、子どもは必ずヘテロ接合体（Aa）です。2の場合で、両親が1本ずつ持っている場合には、子どもは非リスク型のホモ接合体（aa）も、リスク型のホモ接合体（AA）も、ヘテロ接合体（Aa）もあります。3の場合、子どもはリスク型のホモ接合体（AA）か、ヘテロ接合体（Aa）です。4の場合、子どもは必ずリスク型のホモ接合体（AA）です。

両親のリスク型保有数別に考えていくこととして、代表パターンを決めます。両親リスクアレル数が0の場合は、F1 = M1 = F2 = M2 = aです。両親リスクアレル数が1の場合は、F1 = A, M1 = F2 = M2 = aを代表にします。両親リスクアレル数が2の場合は、2通り考える必要があります。リスクアレルが片親に偏っていて、子どもが必ずヘテロ接合体の場合はF1 = F2 = A, M1 = M2 = aを代表にします。他方、両親がリスクアレルを1本ずつ持つ場合は、F1 = M1 = A, F2 = M2 = aを代表にします。両親リスクアレル数が3の場合は、F1 = a, M1 = F2 = M2 = Aを代表にします。両親リスクアレル数が4の場合は、F1 = M1 = F2 = M2 = Aです。

16伝達パターンごとに、2人の子どものジェノタイプと、2人が揃ってフェ

ノタイプを持つ確率を表10.1に示します。

表10.1 ジェノタイプと確率

$F_1 = M_1 = F_2 = M_2 = a$  の場合 (2人揃って発病する確率が、伝達パターンによりません)

	F1M1	F1M2	F2M1	F2M2
F1M1	aa aa	aa aa	aa aa	aa aa
F1M2	aa aa	aa aa	aa aa	aa aa
F2M1	aa aa	aa aa	aa aa	aa aa
F2M2	aa aa	aa aa	aa aa	aa aa

	F1M1	F1M2	F2M1	F2M2
F1M1	$R_0^2$	$R_0^2$	$R_0^2$	$R_0^2$
F1M2	$R_0^2$	$R_0^2$	$R_0^2$	$R_0^2$
F2M1	$R_0^2$	$R_0^2$	$R_0^2$	$R_0^2$
F2M2	$R_0^2$	$R_0^2$	$R_0^2$	$R_0^2$

$F_1 = A, M_1 = F_2 = M_2 = a$  の場合 (2人揃って発病する確率が、伝達パターンによりません)

	F1M1	F1M2	F2M1	F2M2
F1M1	Aa Aa	Aa Aa	Aa AA	Aa AA
F1M2	Aa Aa	Aa Aa	Aa AA	Aa AA
F2M1	AA Aa	AA Aa	AA AA	AA AA
F2M2	AA Aa	AA Aa	AA AA	AA AA

	F1M1	F1M2	F2M1	F2M2
F1M1	$R_1^2$	$R_1^2$	$R_0 \times R_1$	$R_0 \times R_1$
F1M2	$R_1^2$	$R_1^2$	$R_0 \times R_1$	$R_0 \times R_1$
F2M1	$R_0 \times R_1$	$R_0 \times R_1$	$R_0^2$	$R_0^2$
F2M2	$R_0 \times R_1$	$R_0 \times R_1$	$R_0^2$	$R_0^2$

$F_1 = F_2 = A, M_1 = M_2 = a$  の場合 (2人揃って発病する確率が、伝達パターンによりません)

	F1M1	F1M2	F2M1	F2M2
F1M1	Aa Aa	Aa Aa	Aa Aa	Aa Aa
F1M2	Aa Aa	Aa Aa	Aa Aa	Aa Aa
F2M1	Aa Aa	Aa Aa	Aa Aa	Aa Aa
F2M2	Aa Aa	Aa Aa	Aa Aa	Aa Aa

	F1M1	F1M2	F2M1	F2M2
F1M1	$R_1^2$	$R_1^2$	$R_1^2$	$R_1^2$
F1M2	$R_1^2$	$R_1^2$	$R_1^2$	$R_1^2$
F2M1	$R_1^2$	$R_1^2$	$R_1^2$	$R_1^2$
F2M2	$R_1^2$	$R_1^2$	$R_1^2$	$R_1^2$

$F_1 = M_1 = A, F_2 = M_2 = a$  の場合 (2人揃って発病する確率が、伝達パターンによって変わります)

	F1M1	F1M2	F2M1	F2M2
F1M1	AA AA	AA Aa	AA Aa	AA aa
F1M2	Aa AA	Aa Aa	Aa Aa	Aa aa
F2M1	Aa AA	Aa Aa	Aa Aa	Aa aa
F2M2	aa AA	aa Aa	aa Aa	aa aa

	F1M1	F1M2	F2M1	F2M2
F1M1	$R_2^2$	$R_1 \times R_2$	$R_1 \times R_2$	$R_0 \times R_2$
F1M2	$R_1 \times R_2$	$R_1^2$	$R_1^2$	$R_0 \times R_1$
F2M1	$R_1 \times R_2$	$R_1^2$	$R_1^2$	$R_0 \times R_1$
F2M2	$R_0 \times R_2$	$R_0 \times R_1$	$R_0 \times R_1$	$R_0^2$

$F_1 = a, M_1 = A, F_2 = M_2 = A$  の場合 (2人揃って発病する確率が、伝達パターンによって変わります)

	F1M1	F1M2	F2M1	F2M2
F1M1	Aa Aa	Aa Aa	Aa AA	Aa AA
F1M2	Aa Aa	Aa Aa	Aa AA	Aa AA
F2M1	AA Aa	AA Aa	AA AA	AA AA
F2M2	AA Aa	AA Aa	AA AA	AA AA

	F1M1	F1M2	F2M1	F2M2
F1M1	$R_1^2$	$R_1^2$	$R_1 \times R_2$	$R_1 \times R_2$
F1M2	$R_1^2$	$R_1^2$	$R_2 \times R_1$	$R_1 \times R_2$
F2M1	$R_1 \times R_2$	$R_1 \times R_2$	$R_2^2$	$R_2^2$
F2M2	$R_1 \times R_2$	$R_1 \times R_2$	$R_2^2$	$R_2^2$

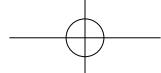


表 10.1 ジェノタイプと確率（つづき）

$F1 = M1 = F2 = M2 = A$  の場合（2人揃って発病する確率が、伝達パターンによりません）

	F1M1	F1M2	F2M1	F2M2
F1M1	AA AA	AA AA	AA AA	AA AA
F1M2	AA AA	AA AA	AA AA	AA AA
F2M1	AA AA	AA AA	AA AA	AA AA
F2M2	AA AA	AA AA	AA AA	AA AA

	F1M1	F1M2	F2M1	F2M2
F1M1	$R_2^2$	$R_2^2$	$R_2^2$	$R_2^2$
F1M2	$R_2^2$	$R_2^2$	$R_2^2$	$R_2^2$
F2M1	$R_2^2$	$R_2^2$	$R_2^2$	$R_2^2$
F2M2	$R_2^2$	$R_2^2$	$R_2^2$	$R_2^2$

両親のアレルの持ち方を場合分けした上で、子どものアレルの持ち方 16通りごとに、同胞が揃って病気になる確率がわかりましたから、伝達パターン 16通りのそれぞれの尤度を計算することができます。もう少し工夫してみることにします。

16 伝達を場合分けします。IBD という考え方がありました（2.1.3 項参照）。

アレルの由来染色体が同じであることでした。この IBD の数で 16 パターンを分類してみます。IBD 数は 2, 1, 0 の 3 通りあります。16 パターンよりは 3 パターンの方が、考えるのが簡単そうです。

IBD 数を同胞対のパターンの組み合わせごとに表に記します。IBD 数が 2 の場合が 4 通り、1 の場合が 8 通り、0 の場合が 4 通りあります。16 パターンはどれも等確率でおきますから、偏りがなければ、IBD 数が 2, 1, 0 になる確率は、0.25, 0.5, 0.25 であることがわかります。

表 10.2 IBD 数

	F1M1	F1M2	F2M1	F2M2
F1M1	2	1	1	0
F1M2	1	2	0	1
F2M1	1	0	2	1
F2M2	0	1	1	2

同胞対がフェノタイプを持っているときに、IBD 数が 0, 1, 2 のどれにどれくらいなりやすいかの計算は、16 伝達パターンの IBD 数と、16 伝達パターンごとの 2 人揃ってフェノタイプを持つ確率とから、計算が可能です。実際に計算すると以下の表のようになります。

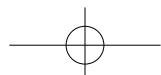


表10.3 IBD数のなりやすいかの計算①

F1	M1	F2	M2	IBD数2	IBD数1	IBD数0
a	a	a	a	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
A	a	a	a	$\frac{1}{4} \left( 1 + \left( \frac{R_1 - R_0}{R_1 + R_0} \right)^2 \right)$	$\frac{1}{2}$	$\frac{1}{4} \left( 1 - \left( \frac{R_1 - R_0}{R_1 + R_0} \right)^2 \right)$
A	a	A	a	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
A	A	a	a	$\frac{1}{2} \left( (1-S) \left( \frac{R_2 - R_0}{R_2 + 2R_1 + R_0} \right)^2 \right)$	S	$\frac{1}{2} \left( (1-S) - \left( \frac{R_2 - R_0}{R_2 + 2R_1 + R_0} \right)^2 \right)$
a	A	A	A	$\frac{1}{4} \left( 1 + \left( \frac{R_2 - R_1}{R_2 + R_1} \right)^2 \right)$	$\frac{1}{2}$	$\frac{1}{4} \left( 1 - \left( \frac{R_2 - R_1}{R_2 + R_1} \right)^2 \right)$
A	A	A	A	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$S = \frac{1}{2} \left( 1 - \left( \frac{R_2 - 2R_1 + R_0}{R_2 + 2R_1 + R_0} \right)^2 \right)$$

表10.4 IBD数のなりやすいかの計算②

F1	M1	F2	M2	IBD数2	IBD数1	IBD数0
a	a	a	a	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$
A	a	a	a	$\frac{1}{4} (1 + \Delta_{1>0}^2)$	$\frac{1}{2}$	$\frac{1}{4} (1 - \Delta_{1>0}^2)$
A	a	A	a	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
A	A	a	a	$\frac{1}{4} (1 + \Delta_{2>1,0>1}^2 + 2\Delta_{2>1>0}^2)$	$\frac{1}{2} (1 - \Delta_{2>1,0>1}^2)$	$\frac{1}{4} (1 + \Delta_{2>1,1>0}^2 - 2\Delta_{2>1>0}^2)$
a	A	A	A	$\frac{1}{4} (1 + \Delta_{2>1}^2)$	$\frac{1}{2}$	$\frac{1}{4} (1 - \Delta_{2>1}^2)$
A	A	A	A	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

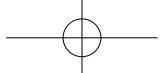
$$\begin{aligned}\Delta_{1>0} &= \frac{R_1 - R_0}{R_1 + R_0} \\ \Delta_{2>1} &= \frac{R_2 - R_1}{R_2 + R_1} \\ \Delta_{2>1,0>1} &= \frac{R_2 - 2R_1 + R_0}{R_2 + 2R_1 + R_0} = \frac{(R_2 - R_1) + (R_0 - R_1)}{(R_2 + R_1) + (R_0 + R_1)} \\ &= \frac{(R_2 + R_0) - 2R_1}{(R_2 + R_1) + (R_0 + R_1)} \\ \Delta_{2>1>0} &= \frac{R_2 - R_0}{R_2 + 2R_1 + R_0} = \frac{(R_2 - R_1) + (R_1 - R_0)}{(R_2 + R_1) + (R_1 + R_0)}\end{aligned}$$

16パターンのすべてで2人揃ってフェノタイプを持つ確率が等しい場合があります。両親のリスクアレル保有数が0の場合と4の場合（表の1行目と6行目）、そして2であってその2つのリスクアレルが片親に偏っている場合（表の3行目）です。この場合には、IBD数が0, 1, 2の尤度は0.25, 0.5, 0.25です。

両親の4アレルのうち、1アレルだけがリスクアレルの場合（表の2行目）と、1アレルだけが非リスクアレルの場合（表の5行目）とは、IBD数1の尤度は0.5ですがIBD数2の尤度が0.25より大きくなり、それと引き換えにIBD数0の尤度が小さくなります。この出入りの大きさは、表では $\Delta^2$ で表しているように、IBD数が2のときに0.25より大きくなり、IBD数が0のときには0.25より小さくなります。その逆はありません。

次に両親が揃ってAaの場合（表の4行目）を見てみます。この行だけ、IBD数1の列の値が0.5ではありません。そして、そのずれは $-\Delta^2$ と表している通り、0.5より小さくなる方向にれます。このとき、IBD数2の場合とIBD数0の場合はIBD数1の場合のずれの分を均等に引き受けて大きくなり（ $\Delta_{2>1,0>1}^2$ の項）、その上で、IBD数2が大きくなり（ $\Delta_{2>1>0}^2$ の項）、その分だけIBD数0が小さくなります。

ずれの項の大きさがどのように決まるかを見てみましょう。両親がリスクアレルを1つだけ持つ場合（表の2行目）のずれの項は、 $\Delta_{1>0}^2 = \left( \frac{R_1 - R_0}{R_1 + R_0} \right)^2$ です。これは、Aaとaaとのリスクの差が0のときに0となり、それ以外は正の値を持つ項です。これは優性遺伝形式の場合に最もよく当てはまります。優性遺伝形式では、両親のうちの片方だけが有病であることが多く、また、AaのリスクとAAのリスクは同じです。実際に優性遺伝形式の場合の有病者は



たいてい、リスクアレルを1本だけ持つことで、Aaとaaのリスクの差が問題となることと、符合します。

両親が3本のリスクアレルを持つ場合（表の5行目）のずれの項は、

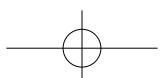
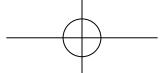
$$\Delta_{1>0}^2 = \left( \frac{R_1 - R_0}{R_1 + R_0} \right)^2$$
 です。これは、AAのリスクとAaのリスクの差で**尤度**の

ずれが決まります。劣性遺伝形式の場合にAAとAaの違いが問題となります。実際には、劣性遺伝形式の場合には、両親が揃ってリスクアレルを1本保有していることが多いので、このパターンを調べることは多くありませんが、IBD数の尤度に関しては、このパターンが劣性形式に対応していることがわかります。

最後に両親が揃ってAaの場合（表の3行目）の尤度のずれの項についてです。まず、IBD数1の場合のずれの項  $\Delta_{2>1,0>1}^2 = \left( \frac{(R_2 + R_0) - 2R_1}{(R_2 + R_1) + (R_0 + R_1)} \right)^2$  は、 $R_2 + R_0 = 2R_1$  の場合に0になります。AaのリスクがAAとaaの中間のときのことです。この項は、IBD数0のときも2のときも尤度のずれをもたらします。リスクアレルを1本持つ場合のリスクが、リスクアレルを2本持つ場合のちょうど半分であるか、それよりずれるかが、IBD数のすべての場合に影響することになります。リスクアレルが相加的にリスクを持つ状態を基準にして、優性遺伝形式寄りか劣性遺伝形式寄りにずれていることを表す項です。

IBD数0と2のときに、さらにずれる項  $\Delta_{2>1>0}^2 = \left( \frac{R_2 - R_0}{R_2 + 2R_1 + R_0} \right)^2$  を見てください。これは、AAとaaのリスクの差の項です。つまり、AA, Aa, aaのリスクはAAとaaのリスクの差を問題にし、その上で、AaのリスクについてはAAとaaの中間を基準にして、そこからのずれについて考慮している、という構図になっています。

$R_2, R_1, R_0$  が相互に異なっていれば、両親から罹患同胞対への伝達パターン16通りの尤度が影響を受けることを見てきました。その16パターンをIBD数によって3グループに分けたときには、3グループの尤度が影響を受けることも見ました。罹患同胞対を用いたノンパラメトリック連鎖解析では、このことを使って、原因座位の位置を探します。 $R_0, R_1, R_2$  の値を変えれば、IBD数0, 1, 2の尤度は増減しますから、それを変化させた上で、原因座位がどこにありそうかを尤度で表し、さらにそれが帰無仮説と比較して有意なかどうかを判定します。



# 第 11 章

## 指数（インデックス）とは

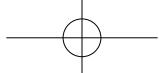
### 11.1 指数は相対的な値

データをかいづまむ話の最後として、**指数（インデックス）**というものについて考えます。ここでの指数というのは、相対評価して数値化したものです。基準に照らしてそれとの相対評価をするときを使います。

ある分子の機能が十全な場合と、全く失われた場合と、ある程度失われた場合とを評価するようなものを想定することができます。2つの集団の混ざり具合として、全く混ざっていない状態と均一な1集団にまで混ざった状態の2状態を両極として、その間としてどのくらいの混ざり具合か、というような評価も、そのような例です。

必ずしも、上限と下限とがあって、その範囲で評価する場合とは限りません。原点に対してある基準値を定め、そこからの増減を割合で表すような場合があります。経済成長率などはよくこうした値を用います。これらに共通するのは、原点とそれ以外の基準点があることです。下限と上限があれば、それを0と1にして指数は0から1の値をとります。原点と基準点があれば、原点を0、基準点を1として、基準点より原点に近ければ0から1の値、基準点より原点から遠ければ1より大きい値を与えます。また、基準点1からの増減を増減率とすることもあります。1から0.9への変化を「0.9倍」と評価したり、「0.1減少」と評価することです。

指数のよいところは、数値を見るだけでその大きさが理解しやすいことで



す。これは、確率分布が無限の値に対応しているときに、変数の値の大きさがイメージしにくいことと対照的です。



## 11.2 不平衡の指数

### ◎ 11.2.1 ハーディ・ワインバーグ平衡 (HWE)

ハーディ・ワインバーグ平衡の評価で用いた  $F$  (3.3.3 項) は指数です。

平衡なときに  $F=0$

	A	a	計
A	$f_A^2$	$f_A \times f_a$	$f_A$
a	$f_A \times f_a$	$f_a^2$	$f_a$
計	$f_A$	$f_a$	1

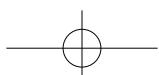
最も偏ったときに  $F=1$  をとる

	A	a	計
A	$f_{AA}=f_A$	0	$f_A$
a	0	$f_{aa}=f_a$	$f_a$
計	$f_A$	$f_a$	1

### ◎ 11.2.2 連鎖不平衡

連鎖不平衡でも指数を使います。平衡なときが 0 で、最も偏ったときが 1 です。第 3 章で扱った  $r$ ,  $r^2$  もこれを満足する指数です。連鎖不平衡の場合には、「最も偏ったとき」の定義を変えることで、別の指標が使われることもあります。2 多型のアレル頻度が等しくなければ、 $r=1$  は実現しないので、アレル頻度が異なる場合にも実現する範囲での平衡からのずれの最大状態のときに 1 となるように調整することができます。連鎖不平衡係数として知られる  $D'$  は、そのような指標で、次の表に示すようなハプロタイプ頻度状態 (4 ハプロタイプのうち、1 ハプロタイプが頻度 0) を最も極端な状態としています。

	A	a	計
B	$f_A$	$f_B - f_A = f_a - f_b$	$f_B$
b	0	$f_b$	$f_b$
計	$f_A$	$f_a$	1



### ● 11.2.3 $p$ 値——確率変数を指数化する

確率分布は、変数の値が無限大まで広がっていることがあり、指数と較べて値の大きさがわかりにくいと述べました。それを解消するのが、 $p$  値です。確率密度関数は確率空間について積分すると 1 になることを利用します。確率密度関数を変数の小さい方から累積していった関数を、累積分布関数と呼びます。こうすることで、確率変数の値はその確率分布のクオントイル（分位数）に対応づけることができます。

クオントイルは、確率変数が最小のときに 0 であり、最大のときに 1 となります。また、確率分布を統計学的検定に用いるときには、値が大きいほど、仮説を信じない方がよい（仮説を棄却する）ことを表しますので、クオントイルが 1 に近いほど、仮説を棄却すべきということになります。 $p$  値は  $1 - \text{クオントイル}$  と大小の方向を反転することで、 $p$  値が小さいほど仮説が棄却されるべきであるというようにしています。図 11.1 は自由度 4 のカイ自乗分布の確率密度関数と累積分布関数（クオントイル値）と  $p$  値のプロットです。確率密度関数の積分が累積分布関数となり、 $p$  値は累積分布関数と上下対称なグラフです (R11-1.R)。

R ソース 11.1 R11-1.R : カイ自乗分布の確率密度関数、累積分布関数と  $p$  値

```

1 df<-4;x <- seq(from = 0, to = 15, by = 0.1)
2 d <- dchisq(x, df);q <- pchisq(x, df);
   p <- pchisq(x, df, lower.tail = FALSE)
3 ylim<-c(0,1)
4 plot(x, d, ylim = ylim, type = "l")
5 par(new = T);plot(x, q, ylim = ylim, type = "l")
6 par(new = T);plot(x, p, ylim = ylim, type = "l")

```

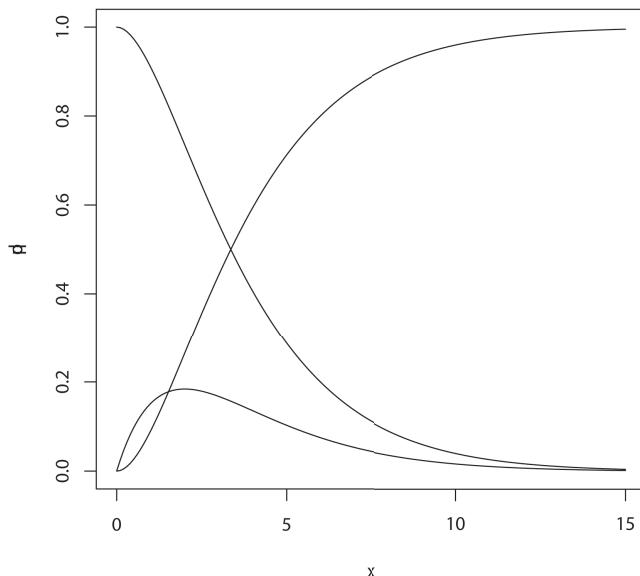
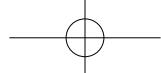


図 11.1 自由度 4 のカイ自乗分布の確率密度（最も背の低いグラフ）、累積密度（右上がりのグラフ）、 $p$  値（左上がりのグラフ）



# 第 IV 部

## 推定、仮説、棄却、関連、因果



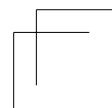
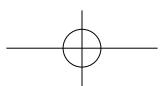
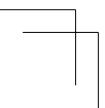
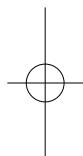
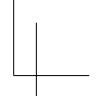
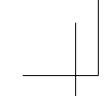
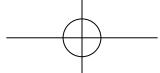
第12章 推 定

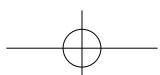
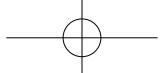
第13章 检定と棄却

第14章 関係と因果

第IV部はデータに潜む知りたい真実を推定し、データに基づいて仮説の真偽を判断（棄却と検定）し、現象の原因なのか結果なのか（因果）を検討することを取り上げます。







# 第 12 章

## 推 定

### 12.1 最尤推定

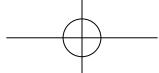
ある大規模な集団について、2 カテゴリ型（因子ありとなし）の調べものをしているとします。何人かをサンプルとして調べることにします。20 人を調べたところ、因子ありが 6 人、なしが 14 人だったとします。さて、大規模集団では、因子ありの割合はどれくらいと考えればよいでしょうか。 $\frac{6}{6+14} = 0.3$  ですから、0.3 (30%) が因子ありと**推定**します。それでよいです。因子ありの割合が 0.3 であるとき、抜き出しを 20 回繰り返して、そのうち 6 回が因子ありである確率は、

$$\frac{20!}{6!14!} 0.3^6 \times (1-0.7)^{14}$$

です。今、集団での因子ありの割合を  $p$  とすると、20 人中 6 人が因子ありの確率は、

$$f(p) = \frac{20!}{6!14!} p^6 \times (1-p)^{14}$$

です。これはサンプル 20 人中 6 人が因子を持っていたときの尤度関数です。これを、 $p$  で微分します。微分するのは、最大値をとるときには傾きが 0 であるからです。



$$\begin{aligned}f'(p) &= \frac{20!}{6!14!} \times (6 \times p^5 \times (1-p)^{14} - 14 \times p^6(1-p)^{13}) \\&= \frac{20!}{6!14!} \times p^5 \times (1-p)^{13}(6(1-p) - 14p) \\&= \frac{20!}{6!14!} \times p^5 \times (1-p)^{13}(6 - 20p)\end{aligned}$$

です。

$$f(p=0.3)=0$$

と、 $f(p)$  は  $p = 0.3$  で唯一の**極値**（**極大値**もしくは**極小値**）をとることがわかります。実際、この極値が最大値であることは、 $f(p=0.3)>0$  で、 $f(p=0)=f(p=1)=0$  であることから、容易にわかります。ですから、 $p=0.3$  というのには、20人中6人が因子ありだったという情報があるときに、元の集団の因子あり割合のうちで、尤度を最大にするものであることがわかります。尤度を最大にする  $p$  の値はわかりました。これは、元の集団の因子あり割合について推定した値です。この場合は尤度を最大にする推定値なので**最尤推定値**と言います。

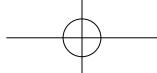
## 12.2 信頼区間

元の集団の因子あり割合は 0.3 かもしれません、0.3 以外がありえないわけではありません。推定範囲に幅を持たせて考えることにします。「幅があるのはよいとして、おおよそ、これより高く、これより低い」という目安の値が知りたいものとします。今、推定したい  $p$  のとりうる値は 0 から 1 まであって、尤度のピークは 0.3 です。尤度関数  $f(p)$  は、 $p$  の値ごとに、何倍ありそうかを表した関数ですから、定数  $C$  を使って、

$$f_c(p) = Cp^6 \times (1-p)^{14}$$

と書けます。ここで、 $C$  を、

$$\int_0^1 f_c(p) dp = \int_0^1 Cp^6 \times (1-p)^{14} dp = 1$$



を満足するようにとることとして、

$$\int_L^U f_c(p)dp : \left( \int_0^L f_c(p)dp + \int_U^1 f_c(p)dp \right) = \alpha : (1 - \alpha)$$

となるような値  $L$  と  $U$ があれば、 $p$  は  $L \leq p \leq U$  の範囲に  $\alpha$  の確率で入るだろう、と言えるでしょう。実際、

$$C = \frac{1}{\int_0^1 p^6 \times (1-p)^{14} dp}$$

であることが知られています。この値は、割合が  $p$  のときに、20人中因子ありが6人である確率

$$\frac{20!}{6!14!} p^6 \times (1-p)^{14}$$

と関係のある係数で、

$$C = \frac{(20+1)!}{6!14!} = \frac{21!}{6!14!}$$

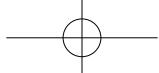
であることも知られています。

さらに一般化しておきます。ここで用いている階乗 ( $n!$ ) を、整数以外でも計算できるようにした関数である**ガンマ関数**と**ベータ関数**を使って、次のように書き換えることができます<sup>\*1</sup>。 $N$ 人中  $k$ 人の場合として、一般に、

$$\begin{aligned} C &= \frac{(N+1)!}{k!(N-k)!} = \frac{\Gamma((N+1)+1)}{\Gamma(k+1)\Gamma((N-k)+1)} \\ &= \frac{\Gamma((k+1)+(N-k+1))}{\Gamma(k+1)\Gamma((N-k)+1)} = \frac{1}{B(k+1, (N-k+1))} \end{aligned}$$

となります。したがって、

<sup>\*1</sup> ガンマ関数、ベータ関数、ベータ分布は付録 B を参照。



$$\begin{aligned}f_c(p) &= \frac{1}{B(k+1, N-k+1)} p^k (1-p)^{N-k} \\&= \beta(p; k+1, N-k+1)\end{aligned}$$

が求める分布です。そしてこれには  $\beta$  分布という名前がついています。ここで、

$$\int_0^L f_c(p) dp = \int_U^1 f_c(p) dp$$

であること、という条件を付け加えれば、

$$\int_L^U f_c(p) dp : \left( \int_0^L f_c(p) dp + \int_U^1 f_c(p) dp \right) = \alpha : (1 - \alpha)$$

を満足するような  $\alpha$  の値に対して、 $L, U$  の値が確定します。これが、元の集団の因子あり割合の推定値を、尤度関数に基づいて決めた  $\alpha$  信頼区間です。

では、実際に R を使ってこの信頼区間を求めてみることにします (R12-1.R)。

#### R ソース 12.1 R12-1.R : 信頼区間

```
1 set.seed(.Random.seed[1]) # 疑似乱数を揃えるため
2 N <- 20; k <- 6 # 観測情報
3 p <- seq(from = 0, to = 1, by = 0.01) # 数値を計算する p のリスト
4 v <- dbeta(p, k + 1, N - k + 1) # β 関数の確率密度
5 plot(p, v, type = "l")
6 abline(v = k/N) # 最尤推定値
7 cirange <- 0.95 # 信頼区間を与える上下 0.025 の範囲を指定します
8 # β 関数のクォンタイル
9 ci <- qbeta(c((1 - cirange)/2, 1 - (1 - cirange)/2), k + 1,
   N - k + 1)
10 abline(v = ci) # 95% 信頼区間
```

$p$  の値についてグラフを描けば、最大の尤度を与える  $p=0.3$  にピークがくることも見てとれます。95% 信頼区間が  $(0.1458769 - 0.5217511)$  となります<sup>※2</sup>。

※2  $p=0.3$  にピークがあり、最尤推定値であることを示しています。95% 上限と下限を垂直線で示しています。上限よりも  $p$  が大きい場合と下限よりも  $p$  が小さい場合のグラフ下の面積は等しいです。

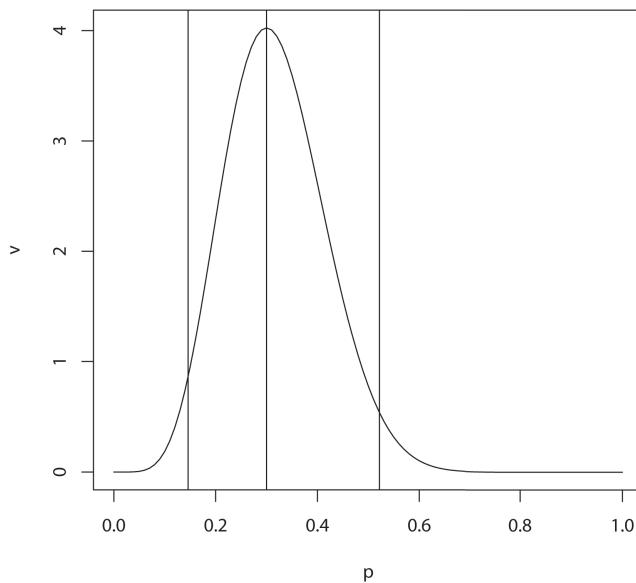
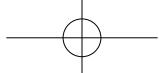


図 12.1 信頼区間

20 人中 6 人が因子あり、という観測に基づいて、 $p$  の確率密度分布が、

$$\frac{1}{B(7,15)} p^6 \times (1-p)^{14} = \\ \beta(p; k+1 = \alpha = 7, N-k+1 = \beta = 15)$$

というベータ分布であると推定しています。ここで、再度、1 人をサンプリングすることとして、そのサンプルが因子である確率を予想してみることとします。ベータ分布では、**最頻値**が  $\frac{\alpha-1}{\alpha-1+\beta-1} = 0.3$ 、平均が  $\frac{\alpha}{\alpha+\beta} = 0.318$  となることが知られていますから、最も可能性が高いのは 0.3 のときです。また、期待値は 0.318 となります。これは、図 12.1 のピークが  $p = 0.3$  のところにあることと、分布が右に裾を引いていることと符合します。



## 12.3 いろいろな信頼区間

元の集団の因子あり割合の推定値、尤度関数に基づいた信頼区間を求めました。ここで、「尤度関数に基づいた」信頼区間と書きました。なぜなら、信頼区間にはいろいろな信頼区間の定義があるからです。

Rで見てみましょう。「あり・なし」のような2つの値をとるかどうかの分布を**二項分布**と呼びますが、そのパッケージ "binom" を使います (R12-2.R)。

### R ソース 12.2 R12-2.R : いろいろな信頼区間

```
1 library(binom)
2 binom.confint(6, 20 ,conf.int=0.95, prior.shape1 = 1,
    prior.shape2 = 1)
```

### ■ R の出力結果

	method	x	n	mean	lower	upper
1	agresti-coull	6	20	0.3000000	0.14315926	0.5212908
2	asymptotic	6	20	0.3000000	0.09916346	0.5008365
3	bayes	6	20	0.3181818	0.14587694	0.5217511
4	cloglog	6	20	0.3000000	0.12252643	0.5013504
5	exact	6	20	0.3000000	0.11893159	0.5427892
6	logit	6	20	0.3000000	0.14140583	0.5272397
7	probit	6	20	0.3000000	0.13522492	0.5212242
8	profile	6	20	0.3000000	0.13184834	0.5165045
9	lrt	6	20	0.3000000	0.13178688	0.5165046
10	prop.test	6	20	0.3000000	0.12839086	0.5433071
11	wilson	6	20	0.3000000	0.14547724	0.5189728

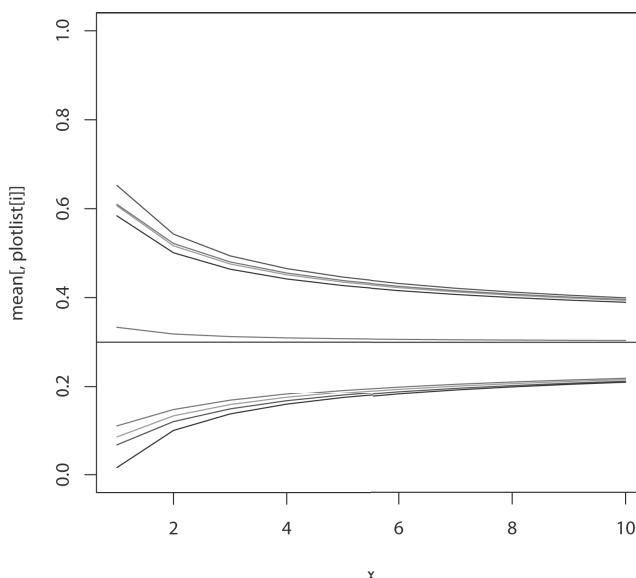
11種類の方法があることがわかります。3番目の "bayes" が12.2節の方法で求めた値と一致しています。それについて推定値を1つの値で答えるなら、この値という出力 "mean" と、信頼区間の上下値 ("upper", "lower") がお出力されます。

これらの手法の意味と求め方は、ここでは触れないこととしますが、異なる手法はどういう点に着目するとよいかを指摘しておきます。これらの手法は、既知の分布を仮定して算出するか、そうでないかで分けることができます。ま

た、上下限値が最尤推定値から同じだけ離れている信頼区間（対称性）とそうでないものとに分けることも有用です。たとえば、上下限値の平均と最尤推定値が等しいことから、2番目の "asymptotic" という方法が対称性であることがわかります。

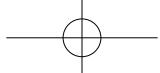
いくつも方法がありますが、いずれの方法でもサンプル数が増えると、平均が0.3に収束し、信頼区間も同じ値に収束します。その様子をプロットしたのが図10.2です。11種類の方法のうち、4方法のみについてプロットしています。横軸は、10, 20, ..., 100人を調べて、それぞれ、3, 6, ..., 30人が因子ありだったときの、割合の推定値（平均）と95%信頼区間です。

サンプル数が多くなると手法による差がなくなるということは、逆に言うと、サンプル数が少ないとには方法によって推定に差があるということです。そして、サンプル数が少ないとには、推定値としてどのような挙動をするのが望ましいかによって、使い分けをすることが適当である、ということでもあります。



横軸が総人数／10、縦軸は平均と信頼区間。Rのソースは掲載しませんが、ダウンロードできるR12-sup1.Rで描画できます。

図12.2 総人数のうち3割が因子ありだったときの推定結果



### ● 12.3.1 ベイズ推定——観察していないとき

因子のあり・なしについて、母集団の割合の推定について見てきました。二項分布の割合  $p$  についての尤度関数としてベータ分布が登場しました。また、最尤推定値と期待値・平均値がずれていることが示されました。信頼区間の計算もを行い、そのときに "bayes" (ベイズ) という名前のついた区間推定方法が出てきました。ベイズという名のつく推定は、**ベイズの定理**（事前分布と観察から**事後分布**を計算すること）が基本です。その意味で、先ほどの最尤推定・信頼区間の算出 (R12-2.R) を見直します。

```
binom.confint(6, 20, conf.level = 0.95 prior.shape1 = 1,  
               prior.shape2 = 1)
```

と実行しました、引数として与えている "6, 20" は 20 人中 6 人という意味ですし、"conf.level=0.95" というのは信頼区間の指定です。"prior.shape1=1,prior.shape2=1" というのはなんでしょうか。これは、ベータ分布

$$\frac{1}{B(a,b)} p^{a-1} (1-p)^{b-1}$$

の "a=prior.shape1, b=prior.shape2" に相当する変数です。20 人中 6 人が因子ありという観察をしたときの、母集団の因子保有者率の尤度の分布を描くときに、図 12.1 を描く R のコマンド (R12-1.R) ではベータ分布の関数を用いて、"v<-dbeta(p, k+1, N-k+1)" としました。ここで、"k+1=a, N-k+1=b" です。k=6, N-k=14 ( $a=k+1=7, b=N-k+1=15$ ) の場合と、 $a=1, b=1$  ( $k=a-1=0, N-k=b-1=0$ ) の場合をプロットしてみると図 12.3 になります。

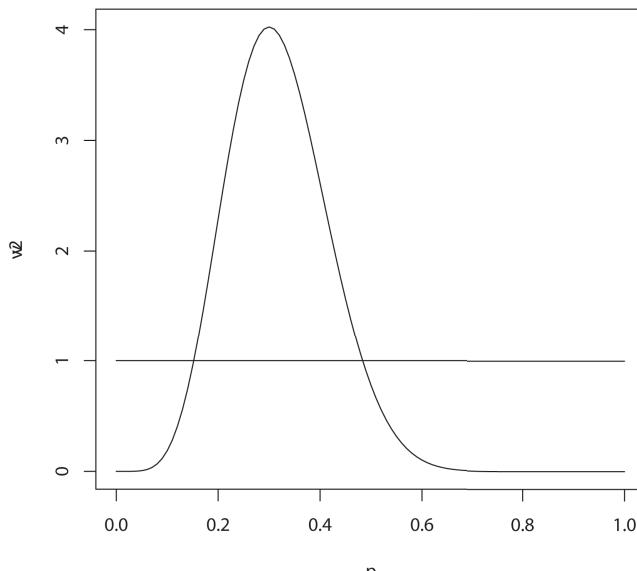


図 12.3  $a=7, b=15$  の場合と  $a=1, b=1$  の場合のベータ関数の確率密度分布(前者が山型、後者が水平線)

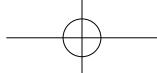
" $a=1, b=1$ " の方は均一分布であることが見てとれます。そのことは以下に示す定義式からもわかります。

$$\frac{1}{B(1,1)} p^{1-1} (1-p)^{1-1} = \frac{\Gamma(1+1)}{\Gamma(1+1)\Gamma(1+1)} \times 1 = 1$$

また、別の見方をすれば、" $a=1, b=1$ " は、" $k=0, N-k=0$ " のことですが、これは1人も調べていないことを意味しています。何の情報が何もない状態という意味です。そんなときは  $p$  の値に優劣をつけられないために均一分布を推定しているわけです。

### ● 12.3.2 ベイズ推定、事前確率、共役事前分布

今、何がしかの情報があって、因子ありの方が少なめだという事前確率があるとします。因子ありが因子なしよりも少なめな事前分布としてどんな分布を仮定しても構わないので、ここでベータ分布を事前分布にしてみることにします。`dbeta()` 関数の引数として  $x < y$  のような 2 数を与えれば、因子ありが少なめになりますから、適当に  $x, y$  を変えてその分布の形を見てみま



す。この分布が思っている分布だというのがあれば、それを採用すればよいです。

### R ソース 12.3 R12-3.R : 事前分布

```
1 x <- 1.2; y <- 1.5; p<-seq(from=0,to=1,by=0.01)
2 v3 <- dbeta(p, x, y)
3 plot(p, v3, type = "l", ylim = c(0, 4))
4 v4 <- dbeta(p,x+6,y+14)
5 par(new=TRUE)
6 plot(p, v4, type = "l", ylim = c(0, 4))
```

たとえば、次のような分布（図 12.4 (a)）が思い描いている分布だとします（R12-3.R）。

$$\frac{1}{B(x=1.2, y=1.5)} p^{1.2-1} (1-p)^{1.5-1}$$

この事前分布を思い描いているときに、20 人中 6 人が因子あり、という情報が得られたとすると、事後分布は、

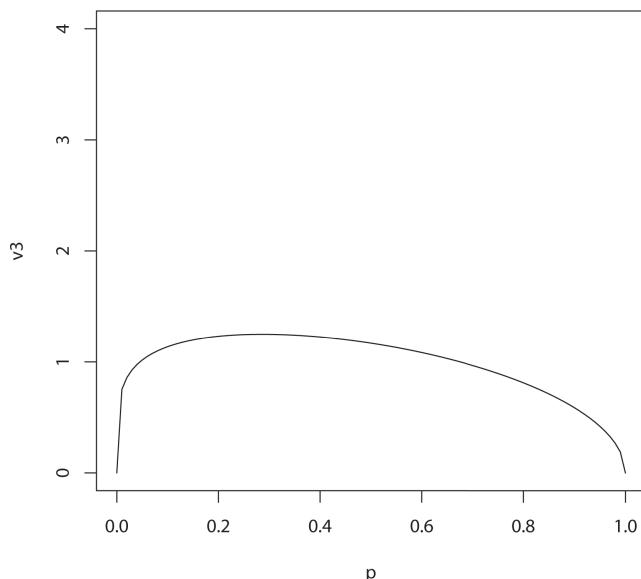
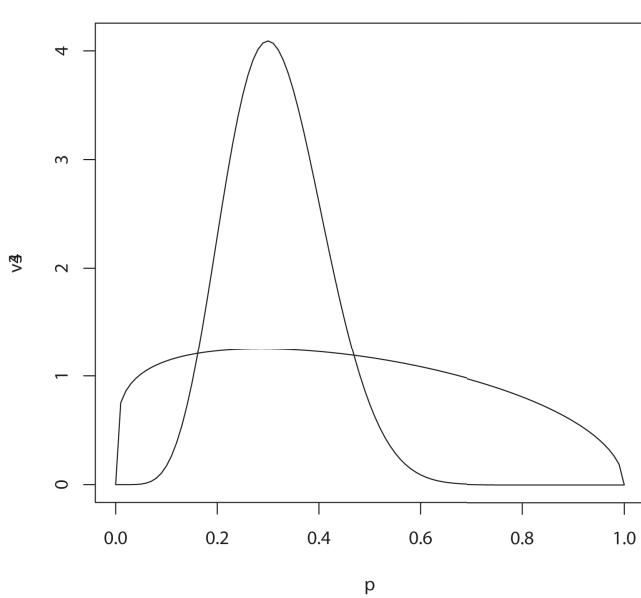
$$\begin{aligned} \frac{1}{B(x=1.2, y=1.5)} p^{x-1} (1-p)^{y-1} \times p^6 (1-p)^{14} = \\ \frac{1}{B(x=1.2, y=1.5)} p^{1.2+6-1} (1-p)^{1.5+14-1} \end{aligned}$$

に比例することになります。

このような  $p$  の関数で、 $p$  がとりうる範囲 ( $0 \leq p \leq 1$ ) について積分して 1 になるように補正してやれば

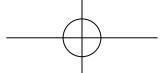
$$\frac{1}{B(x+6, y+14)} p^{x+6-1} (1-p)^{y+14-1}$$

となるので、結局、事前分布にベータ分布を仮定すると、観察結果（ $N$  人中  $k$  人が因子あり）の情報によって、

(a)  $x = 1.2, y = 1.5$  の事前分布

(b) 事前分布を受けて、6:14 の観察をした後の事後分布を描き加えている

図 12.4 事前分布と事後分布



$$\frac{1}{B(x,y)} p^{x-1} (1-p)^{y-1} \rightarrow \\ \frac{1}{B(x+k, y+(N-k))} p^{x+k-1} (1-p)^{x+(N-k)-1}$$

という変化になります。関数を決めているパラメタのみに着目すれば、

$$(x, y) \rightarrow (x+k, y+(N-k))$$

という変化です。もし、この後さらに情報が追加されて、 $N'$ 人中  $k'$ 人が因子ありだったら、

$$(x, y) \rightarrow (x+k, y+(N-k)) \rightarrow (x+k+k', y+(N-k)+(N'-k'))$$

というようにパラメタの変化だけを気にすればよいです。

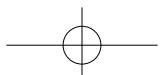
このように、事前分布と事後分布が同じ分布であると、分布を変化させていくときに、分布のパラメタの簡単な計算のみを考えればよくなって、好都合です。確率的に起きる現象が二項分布であるときには、ベータ分布を事前分布とおくと便利なのです。この二項分布とベータ分布との便利な関係を、「二項分布の**共役事前分布**はベータ分布である」と言います。12.3.3 項で説明する**多項分布**の共役事前分布はディリクレ分布ですし、**ポアッソン分布**のそれはガンマ分布です。

これを用いて、R12-3.R で "x=1.2, y=1.5" の事前分布からスタートして、20 人中 6 人という観察に基づいて求めた事後分布は図 12.4 (b) のようになります。

### ◎ 12.3.3 多項分布とその共役事前分布——ディリクレ分布

3 以上カテゴリに関して、**多項分布**があります。3 以上カテゴリがある比率であるときに、そこからいくつかを取り出したときにある内訳となる確率に関する分布です。二項分布をカテゴリ数が 3 以上でも使えるように拡張した分布です。

$$\frac{N!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k (p_i)^{n_i}$$



ただし、 $\sum_{i=1}^k n_i = N, n_i \geq 0$ と表されます。

二項分布の共役事前分布はベータ分布でしたが、多項分布の共役事前分布は**ディリクレ分布**と呼ばれます。

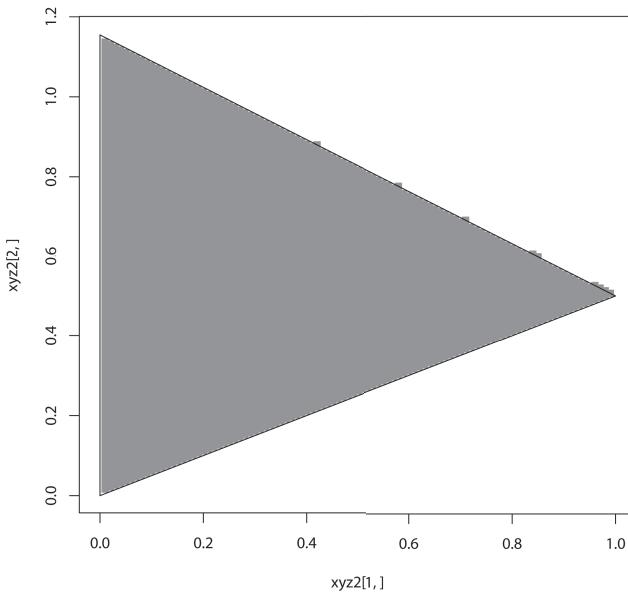
今、母集団の3カテゴリの比率を推定しようとしているとしています。20人をサンプリングして観察したら、内訳が10, 7, 3人ずつだったとします。事前分布として全く無情報のときには、全部で0人観察して各カテゴリに0人ずつの情報がある状態です。ベータ分布のときには、このような場合に"(1, 1)"を関数に渡しました。ディリクレ分布のときにも同様に、"(1, 1, 1)"を渡します。そして、観測したならば、ベータ分布のときには"(1+6, 1+14)"のように、観察人数をそれぞれ足してやりました。ディリクレ分布も同様で、"(1+10, 1+7, 1+3)"としてやります。

図12.5の図を描くRのコマンドは長くなるので掲載しませんが、ダウンロードすることができるR12-sup2.R(三項の分布)がそのソースです。簡単に説明すると、ベータ分布のときのdbeta()関数に相当するのがddirichlet()関数です。

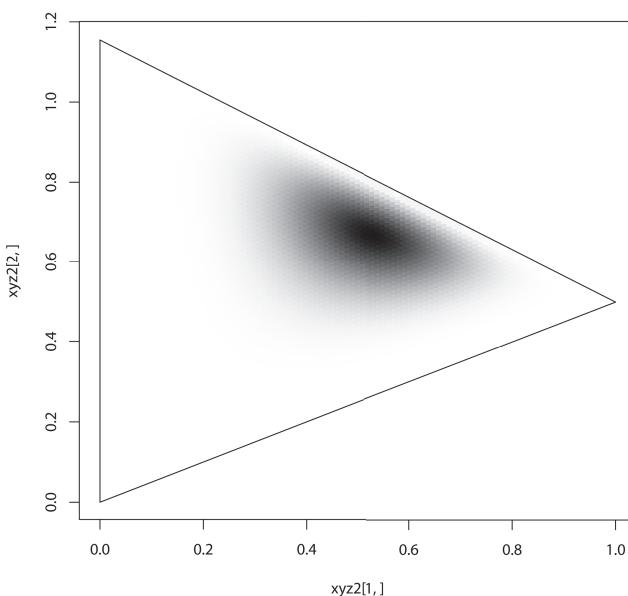
3カテゴリは2次元正単体である正三角形の頂点に配置することができます(第4章)<sup>※3</sup>。3カテゴリの内訳分布はこの正三角形の内部に描けますから、それを描いて、濃淡で分布を示せば、事前分布は一様(図12.5(a))で、ある観察に基づいて推定した事後分布は図12.5(b)のように濃淡ができます。

---

※3  $k$  カテゴリに  $v_i$  ( $i=1, 2, \dots, k$ ) を対応させたとき、これらの線形結合  $x = \sum_{i=1}^k a_i v_i$  で表される  $x$  は  $\sum_{i=1}^k a_i = 1$  という制約をつけると、この  $k$  個のベクトルの頂点を通る面の上の点です。さらに制約  $a_i \geq 0$ を入れると、 $k$  個のベクトルの頂点同士を結んだ直線に囲まれた領域(凸包)となります。

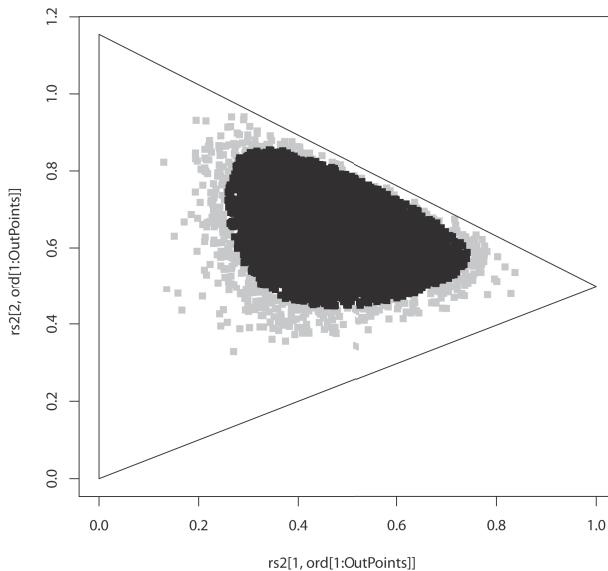


(a) 未観測状態の事前分布（均一分布）



(b) 観測 (10, 7, 3) 後の事後分布

図 12.5 三項の分布



(c) 亂数による 95% CI 区間の表示（中央の黒が 95% CI 内、周辺の淡い領域が CI の外側）

図 12.5 三項の分布（つづき）

では、この場合の信頼区間について、考えてみることにします。

ベータ分布を用いて 95% 信頼区間を考えたときには、両端にそれぞれ 5% の半分をとることで、信頼区間を決めました。今回は 2 次元平面に広がっていて「両端」がありません。その代わりに、「周辺地帯」があります。「周辺地帯」が合算して 5% になるようにして、中央付近に 95% がくるように線引きをしたいのです。

ある分布からランダムにサンプリングができるときには、たくさんのサンプルを発生させて、その分布から信頼区間を決めることもできます。今の例では、 $(10+1, 7+1, 3+1)$  をパラメタとしたディリクレ分布が対象ですが、これから乱数を発生させることもできますし (`rdirichlet()` 関数)、そのようにして発生させた 3 カテゴリの比率のデータについて、そのディリクレ分布に基づく生起確率も求められます (`ddirichlet()` 関数)。発生させたサンプルの確率が小さい方から 5% と、大きい方から 95% を塗り分ければ、境界線が 95% 信頼区間の辺縁になります。

### ● 12.3.4 最尤推定とハプロタイプ頻度推定——連鎖不平衡 係数推定

#### (1) 疑似乱数を用いた最尤推定

2個の2アレル型多型 (A/a と B/b) があるとします。2多型が作るハプロタイプとしては、AB, Ab, aB, ab の4種類があります。今、2多型のそれぞれについて2倍体ジェノタイプを観察するとします。9タイプの人数 ( $n_{ij}$ ) が観察されたとしましょう。 $n_{ij}$  の値は総人数を 100 とし、AA, Aa, aa は A のアレル頻度  $p_A = 0.4$  で HWE を満足するように定め、BB, Bb, bb は、B のアレル頻度  $p_B = 0.3$  で HWE を満足するように定め、その上で  $n_{ij}$  の内訳を、(AA, BB), (aa, bb) が多めになり、(AA, bb), (aa, BB) が少なめになるように適当に足し引きして作った表が表 12.1 です。

表 12.1 ジェノタイプ観察表

	AA	Aa	aa	計
BB	$n_{11} = 2$	$n_{12} = 5$	$n_{13} = 2$	9
Bb	$n_{21} = 10$	$n_{22} = 18$	$n_{23} = 14$	42
bb	$n_{31} = 4$	$n_{32} = 25$	$n_{33} = 20$	49
計	16	48	36	100

さて、この集団の4ハプロタイプ AB, aB, Ab, ab の頻度を推定してみることとします。4ハプロタイプの頻度を  $f_1, f_2, f_3, f_4$ ,  $\sum_i^4 f_i = 1$  として HWE を仮定すれば、ディプロタイプの頻度  $F_{ij}$  はジェノタイプ頻度表のように与えられます。これを用いて、9ディプロタイプの観測の尤度  $L(F)$  の対数（**対数尤度**） $\log(L(F))$  を計算すると、

$$\log(L(F)) = \sum_{i=1}^3 \sum_{j=1}^3 \log(F_{ij}^{n_{ij}}) = \sum_{i=1}^3 \sum_{j=1}^3 n_{ij} \log F_{ij}$$

です。

表 12.2 HWE を仮定したときのジェノタイプ頻度表

	AA	Aa	aa
BB	$F_{11} = f_1^2$	$F_{12} = 2f_1f_2$	$F_{13} = f_2^2$
Bb	$F_{21} = 2f_1f_3$	$F_{22} = 2(f_1f_4 + f_2f_3)$	$F_{23} = 2f_2f_4$
bb	$F_{31} = f_3^2$	$F_{32} = 2f_3f_4$	$F_{33} = f_4^2$

4 ハプロタイプの頻度  $f_k$ ,  $k = 1, 2, 3, 4$  をディリクレ分布を使って、ランダムに発生させて、「 $\log(L(F))$ 」を一番大きくするような  $f_k$ 」を見つけてやることにしましょう。

#### R ソース 12.4 R12-4.R : ハプロタイプ頻度推定

```

1 library(MCMCpack)
2 # 4 ハプロタイプの頻度 $f$ から、9 ジェノタイプの HWE 下頻度を作成
3 Make3x3 <- function(f) {
4   tmp <- f %*% t(f)
5   matrix(c(tmp[1, 1], 2 * tmp[1, 2], tmp[2, 2], 2 * tmp[1,
6     3], 2 * (tmp[1, 4] + tmp[2, 3]), 2 * tmp[2, 4], tmp[3,
7     3], 2 * tmp[3, 4], tmp[4, 4]), nrow = 3, byrow = TRUE)
8 }
9 # 4 ハプロタイプ頻度 $f$ のもとでの
10 # 9 ジェノタイプカウント $g$ を観察する対数尤度を算出
11 CalcLike3x3 <- function(g = matrix(1, 3, 3),
12   f = c(0.25, 0.25, 0.25, 0.25)) {
13   F <- Make3x3(f)
14   sum(g * log(F))
15 }
16 CalcR <- function(h) { # 4 ハプロタイプ頻度から連鎖不平衡係数 r を計算
17   p_A <- h[1] + h[3]
18   p_B <- h[1] + h[2]
19   (h[1] -
20     p_A * p_B)/sqrt(p_A * (1 - p_A) * p_B * (1 - p_B))
21 }
22 ns <- matrix(c(2, 5, 2, 10, 18, 14, 4, 25, 20), nrow = 3,
23   byrow = TRUE) # 初期 9 ジェノタイプカウント
24 N <- 10000 # 試行回数
25 p <- c(1, 1, 1, 1) # ディリクレ事前分布パラメタ (一様分布)
26 sampled <- rdirichlet(N, p) # ディリクレ乱数
27 loglikes <- apply(sampled, 1, CalcLike3x3, g=ns)
28 maxset <- sampled[which(loglikes == max(loglikes)), ]

```

```

# 最大尤度を返した 4 ハプロタイプ頻度
26 maxp_A <- maxset[1] + maxset[3]
# 最大尤度を返したハプロタイプ頻度から、SNPA のアレル頻度を計算
27 maxp_B <- maxset[1] + maxset[2] # 同じくSNPB のアレル頻度を計算
28 maxr <- CalcR(maxset) # 同じく連鎖不平衡係数 r を計算
29 maxset

```

この R の処理によって 4 ハプロタイプの頻度は、次のように推定されました。

### ■ R の出力結果

```
[1] 0.1520910 0.1296524 0.2513286 0.4669280
```

4 ハプロタイプ AB, aB, Ab, ab の頻度として 0.16, 0.14, 0.24, 0.46 くらいの値が返り、A, B の頻度としては 0.4 前後、0.3 前後の値（提示例では  $0.1520910 + 0.2513286 = 0.4034196$ ,  $0.1520910 + 0.1296524 = 0.2817434$ ）となります。

4 ハプロタイプの頻度も、 $N$  回の試行をした限りでは、尤度が最大な値なわけですから、まずはの **最尤推定値**と言えます。

### (2) 制約を入れて最尤推定

別のやり方を考えてみます。今、AA, Aa, aa の人数 16, 48, 36 が与えられています。このとき、A アレル本数は  $16 \times 2 + 48 = 80$  と数え上げることが可能で、この数値から A アレルの標本頻度が 0.4 であることがわかります。同様に、B アレルの標本頻度は 0.3 です。これらはそれぞれのアレル頻度の最尤推定値です<sup>※4</sup>。では、 $p_A$ ,  $p_B$  はこの多型ごとの最尤推定値を用いるとして、9 ジェノタイプの人数がわかっているときに、4 ハプロタイプの頻度の推定をして

※4 HWE 条件のもと、

$$\begin{aligned}\log(L(p)) &= n_{AA} \log(p_A^2) + n_{Aa} \log(2p_A p_a) + n_{aa} \log(p_a^2) \\ &= (2n_{AA} + n_{Aa}) \log(A) + (2n_{aa} + n_{Aa}) \log(a) + n_{Aa} \log(2)\end{aligned}$$

で微分すると、

$$(\log(L(p)))' = \frac{2n_{AA} + n_{Aa}}{p} - \frac{2n_{aa} + n_{Aa}}{1-p} (\log(L(p)))' = 0$$

を解いて、

$$p = \frac{2n_{AA} + n_{Aa}}{2(n_{AA} + n_{Aa} + n_{aa})}$$

となります。

みましょう。

4ハプロタイプの頻度は、2多型のアレル頻度  $p_A, p_a, p_B, p_b$  と、**連鎖不平衝係数  $r$** を用いて、

$$\begin{aligned} h_1 &= p_A p_B + r \sqrt{p_A p_a p_B p_b} \\ h_2 &= p_A p_b - r \sqrt{p_A p_a p_B p_b} \\ h_3 &= p_a p_B - r \sqrt{p_A p_a p_B p_b} \\ h_4 &= p_a p_b + r \sqrt{p_A p_a p_B p_b} \end{aligned}$$

と表せたことを思い出せば、 $p_A, p_B$  が与えられているときには、 $r$ のみで4ハプロタイプ頻度が決まります。したがって、尤度関数も  $r$ のみの関数になります。今度は、 $r$ を  $-1 \leq r \leq 1$  の範囲で、適当な刻み幅でしらみつぶしに調べてみるとこにしましょう (R12-5.R)。

#### R ソース 12.5 R12-5.R : ハプロタイプ頻度推定②

```

1  # p_A, p_B を与えて、r を変数として 4 ハプロタイプの頻度を作る
2  p_A <- 0.4; p_B <- 0.3
3  # 連鎖平衡のときのハプロタイプの頻度
4  f1 <- p_A * p_B; f2 <- (1 - p_A) * p_B; f3 <- p_A *
   (1 - p_B); f4 <- (1 - p_A) * (1 - p_B)
5  rs <- seq(from = -1, to = 1, by = 0.001) # r の値を振る
6  # ハプロタイプ頻度
7  ds <- rs * sqrt(p_A * (1 - p_A) * p_B * (1 - p_B))
8  f1 <- ds + f1; f2 <- -ds + f2; f3 <- -ds + f3;
   f4 <- ds + f4
9  F <- matrix(c(f1, f2, f3, f4), nrow = length(f1))
10 minF <- apply(F, 1, min)
    # 4 ハプロタイプとも正の頻度を持つ場合を取り出す。
11 rs <- rs[minF > 0]
12 F <- F[minF > 0, ]
13 loglikes2 <- apply(F, 1, CalcLike3x3, g=ns)
14 maxset2 <- F[which(loglikes == max(loglikes2)), ]
15 maxp_A2 <- maxset2[1] + maxset2[3];
   maxp_B2 <- maxset2[1] + maxset2[2]
16 maxr2 <- CalcR(maxset2)
17 plot(rs, loglikes2, type="l")

```

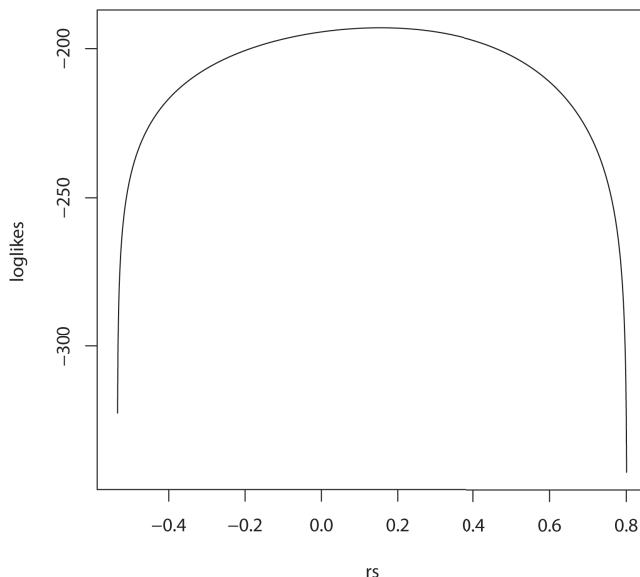


図 12.6 ハプロタイプ頻度推定 2：横軸に  $r$ 、縦軸に対数尤度

この結果は図 12.6 のようになります。

尤度がピークを作ることが示されました。このピークをもたらす値が  $r$  の最尤推定値です。

実際、求めた尤度を先に実施したディリクレ分布乱数によって得られた尤度と較べると、今回の尤度の方が高くなります。

## 12.4 EM アルゴリズム

2通りの方法で最尤推定値に近づいてみました。1つ目の方法では4ハプロタイプの頻度をやみくもに探索（ディリクレ分布で4変数の3自由度空間を探索）し、2つ目の方法では、規則正しく（1変数の空間を、細かく定まった間隔で探索）調べ上げました。変数の数が多いとき空間は広く、空間が広くなると空間全体を調べることが大変になります。そのようなときにうまく探索する方法が必要になります。その1つが**EM アルゴリズム**という方法です<sup>※5</sup>。

EM アルゴリズムを用いて、観測データ（ディプロタイプ頻度データ）をもとにハプロタイプ頻度を推定するときには、ハプロタイプ頻度に暫定的に値を割り当てるところからスタートします。そして、与えた暫定頻度のもとでのハプロタイプ頻度の期待値を計算し直して、暫定的な値を更新します。このような手順を踏むと、だんだんに尤度が高くなることが知られています。では、実際の処理の動きを見ていくことにします<sup>※6</sup>。

ここでは、4ハプロタイプに均一の頻度を与えて、それを初期値とすることとします。頻度の更新回数は1,000回とします。4ハプロタイプの頻度の初期値が与えられると、観察ジェノタイプデータから、4ハプロタイプの本数の期待値を算出することになります。今、9ジェノタイプのうち、2多型ともヘテロ型（Aa, Bb）である場合以外は、2本のハプロタイプの内訳が確定します。

2座位がホモならば、同一のハプロタイプを2本持りますし、1座位がホモでもう1座位がヘテロのときには、異なる2つのハプロタイプの対であることが確定します。ですから、その9ジェノタイプのうち8つについては、ハプロタイプ頻度の推定値とは無関係にハプロタイプ本数が数えられます。最後の

※5 ちなみに、「アルゴリズム」というのは、何かしらの目的を達成するための「方法に関する工夫」とでも言うものです。たいていは完璧な方法がなかったり、あったとしても非常に時間や労力がかかる場合に、少ない労力で満足できる結果を得るような工夫であって、その工夫が確かにそうなっていることが示されている（論文などで発表されている）もののこと、と言ってもよいでしょう。

※6 最初に与える暫定的なハプロタイプ頻度（初期値）によらずに、得られるべき結果（最尤推定値）に行き着くといいのですが、場合によっては最尤推定値に到達しないこともあります。なぜなら、このアルゴリズムは、だんだんにより大きな尤度を与える値に移動していくので、**極大値**が複数あるときには、最大ではない極大値に行き着くこともあります。しかしながら、多くの場合には、最尤推定値に行き着くものと思っておくことにしましょう。

1つのディプロタイプについては、(AB, ab) というハプロタイプの組み合わせか、(Ab, aB) という組み合わせかの、いずれかです。それぞれの組み合わせの起きる確率は、HWE を仮定すれば  $f_1 \times f_4$ ,  $f_2 \times f_3$  ですので、 $n_{22}$  人は次のように振り分けられます。

$$\begin{cases} (AB, ab) & \frac{f_1 \times f_4}{f_1 f_4 + f_2 f_3} \\ (Ab, Ab) & \frac{f_2 \times f_3}{f_1 f_4 + f_2 f_3} \end{cases}$$

これを用いて、観察しているディプロタイプのデータに占める、4ハプロタイプの標本頻度を計算します。このようにして計算したハプロタイプの頻度を、新たなハプロタイプ頻度として暫定的に認めます。後は、同じ処理を繰り返します。

この更新の手続きと、その経過を保管する処理をする R のコマンドを示します (R12-6.R)。

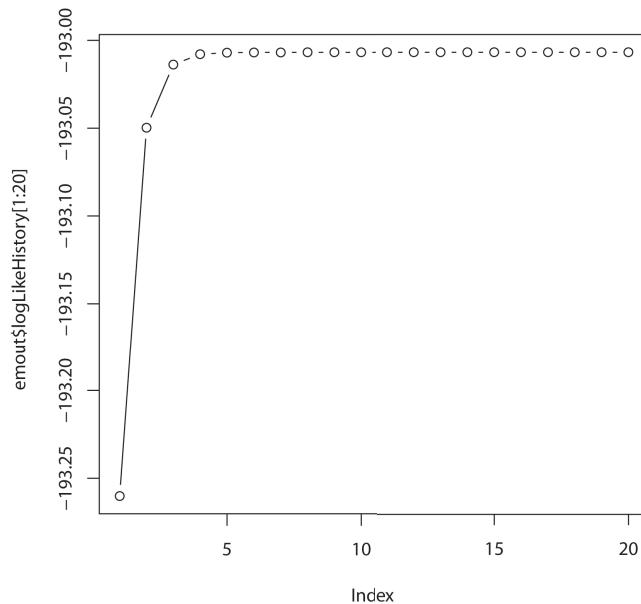
#### R ソース 12.6 R12-6.R : EM アルゴリズム

```

1  EM2loci <- function(n, p = NULL, Niter = 1000) {
2      if (is.null(p))          p <- rep(0.25, 4)
          # 初期頻度を与えないければ、均等に割り当て
3      f <- p
4      rs <- rep(0, Niter) # 連鎖不平衡係数の収束ログ
5      logLikes <- rep(0, Niter) # 対数尤度の収束ログ
6      fs <- matrix(0, Niter, 4) # ハプロタイプ頻度の収束ログ
7      # ハプロタイプ頻度によらず、確定するハプロタイプ本数
8      fixed<- c(n[1, 1] * 2 + n[1, 2] + n[2, 1],
         n[1, 3] * 2 + n[1, 2] + n[2, 3],
         n[3, 1] * 2 + n[2, 1] + n[3, 2], n[3, 3] * 2 +
9         n[2, 3] + n[3, 2])
10     for (i in 1:Niter) {
11     # 暫定ハプロタイプ頻度における2重ヘテロジエノタイプを
12     # f[1]-f[4] と f[2]-f[3] との2組に分配する比率
13     tmpratio<-f[1]*f[4]/(f[1]*f[4]+f[2]*f[3])
14     tmp <- rep(0,4)
15     tmp[1] <- fixed[1] + n[2, 2] * tmpratio
           tmp[2] <- fixed[2] + n[2, 2] * (1 - tmpratio)
```

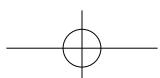
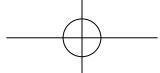
```
16      tmp[3] <- fixed[3] + n[2, 2] * (1 - tmpratio)
17      tmp[4] <- fixed[4] + n[2, 2] * tmpratio
18  # この回の更新結果を格納
19      fs[i, ] <- tmp/sum(tmp);
20      logLikes[i] <- CalcLike3x3(n, tmp);
21      rs[i] <- CalcR(tmp)
22      f <- tmp # 暫定ハプロタイプ頻度の更新
23  }
24  list(f = f, r = rs[Niter], logLike = logLikes[Niter],
25       fHistory = fs, rHistory = rs,
26       logLikeHistory = logLikes)
27 }
28 emout <- EM2loci(ns)
29 maxp_A <- emout$f[1] + emout$f[3];
30 maxp_B <- emout$f[1] + emout$f[2]; maxr <- emout$r
31 plot(emout$logLikeHistory[1:20], type = "b")
32                                # 20回目までの対数尤度の更新履歴をプロット
```

尤度の増え方を、繰り返し処理の初めの 20 回について描いたものです。確かに尤度がだんだんに大きくなり、収束していく様子が見てとれます。3, 4 回の繰り返し計算ではほぼ頂上にたどり着いています（図 12.7）。



繰り返し処理が進むにつれ、対数尤度が大きくなり、収束していく様子がわかります。横軸が繰り返し回数、縦軸が対数尤度。

図 12.7 EM アルゴリズム



# 第 13 章

## 棄却と検定

### 13.1 信じるのが難しい仮説を棄却する—— 3 カテゴリの観察

これまで、データから変数の期待値や最尤推定量、信頼区間などを推定しました。今度は、ある変数がある値であるという仮説を信じるべきか信じないべきか、その程度を数値で表す、という話です。仮説を信じる程度が低いときに、その仮説をある基準で棄却する、と言うので、この話は「仮説の棄却」についてです。

3 カテゴリについて、(10, 7, 3) と観測した場合を考えます。3 カテゴリが同一頻度だと仮定します。N 人が 3 カテゴリに  $n_1, n_2, n_3$  人に分かれて観察する確率は

$$\frac{N!}{n_1!n_2!n_3!} \left(\frac{1}{3}\right)^N$$

です。20 人を 3 カテゴリに分ける分け方のすべてについて、その確率を計算してみます。

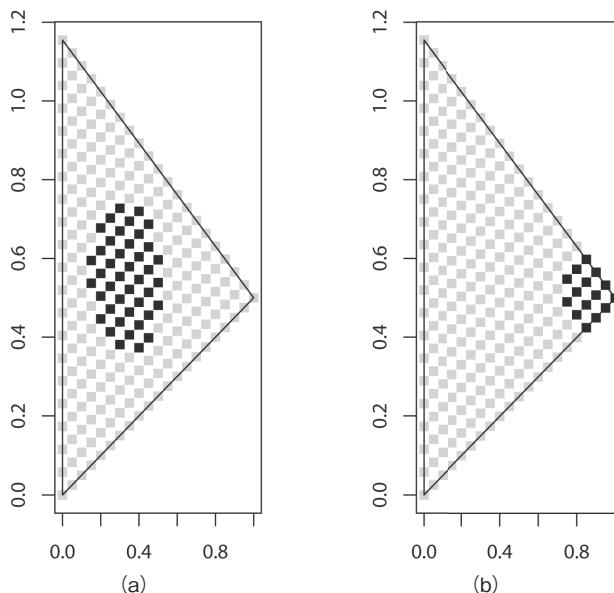
R の処理は以下の通りです (R13-1.R)。

## R ソース 13.1 R13-1.R : 3 カテゴリの正確確率

```

1  dpoly<-function(n=c(1,2,3),p=NULL) {
2      # apply() 関数で使うための関数を作る（確率計算）
3      N<-sum(n)
4      if(is.null(p)){
5          p<-rep(1/length(n),length(n))
6      }
7      exp(lgamma(N+1)-sum(lgamma(n+1))+sum(n*log(p)))
8
9  DrawHigherLower<-function(g=c(10,7,3),p=NULL) {
10     N<-sum(g)
11     x<-0:N;y<-0:N
12     xy<-expand.grid(x,y) # x,yの全組み合わせを作る
13     z<--(xy[,1]+xy[,2])+N
14     xyz<-matrix(c(xy[,1],xy[,2],z),nrow=length(z))
15     xyz<-xyz[apply(xyz,1,min)>=0,] # xyz のすべての要素は非負
16
17     probs<-apply(xyz,1,dpoly,p=p)
18     # xyz の各行 (3 数値) について、上で定義した dpoly() 関数を適用する
19     probObs<-dpoly(c(10,7,3)) # 観測データの確率
20     higher<-which(probs>probObs)
21         # 観測データの生起確率より大きい確率の場合
22     lower<-which(probs<=probObs)
23         # 観察データの生起確率以下の確率の場合
24     xyz2<-apply(xyz,1,CoordTriangle)
25     x2<-xyz2[1,];y2<-xyz2[2,]
26     xlim<-c(0,1);ylim<-c(0,2/sqrt(3))
27     # 高確率は黒で、それ以外は灰色でプロット
28     plot(xyz2[1,which(probs>probObs)]/N,xyz2[2,
29             which(probs>probObs)]/N,xlim=xlim,ylim=ylim,col =
30             "black",pch=15)
31     par(new=T)
32     plot(xyz2[1,which(probs<=probObs)]/N,xyz2[2,
33             which(probs<=probObs)]/N,xlim=xlim,ylim=ylim,
34             col = gray(4/5),pch=15)
35     DrawTriangleFrame()
36
37 }
38 par(mfcol=c(1,2))
39 DrawHigherLower(c(10,7,3))
40 DrawHigherLower(c(10,7,3),c(0.9,0.05,0.05))
41 par(mfcol=c(1,1))

```



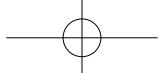
20人の内訳が $(20, 0, 0)$ ,  $(0, 20, 0)$ ,  $(0, 0, 20)$ のときが三角形の頂点。(a)は、3要素の確率が等しいとしたときに $(10, 7, 3)$ と観測するより高確率の領域が黒で、低確率が灰色。(b)は、3要素の確率が $(0.9, 0.05, 0.05)$ との仮説のもとで $(10, 7, 3)$ と観測するより高確率か低確率かで色分けしています。

図 13.1 3要素の三角プロット

$(10, 7, 3)$ と観測した場合の確率と比較して色分けしたのが、図 13.1 の (a) です。

すべての点の確率を足し合わせると 1 になっています。灰色は $(10, 7, 3)$ と観測する確率以下の確率で起きるペアを表しています。この部分の確率を足し合わせると 0.177 になるので、 $(10, 7, 3)$ と同程度か、それよりも珍しいことが起きる確率は 0.177 です。これは、3 カテゴリが同頻度だという仮説のもとで、 $(10, 7, 3)$ を観察する珍しさであり、この値が小さいほど珍しく、大きくなるとその最大値は 1 で、最もありふれていますことになります。

この数値が正確確率をもとにした  $p$  値です。3 カテゴリが同一頻度であるという仮説検定の  $p$  値です。今、観察される珍しさが低い方から 0.05 に含まれれば、仮説を信じることをやめることにする、とするとき、0.05 を **棄却水準** と言います。そしてこの  $p$  値が 0.05 未満であれば、仮説が棄却された、と言います。今回は、 $p = 0.177$  ですから、棄却水準を 0.05 とすれば、棄却されない



ことがわかります。

図13.1 (a) では、3 カテゴリの頻度が同じであるという仮説に基づいて生起確率を計算しましたが、3 カテゴリの頻度が異なるという仮説に基づいても同様に計算して図を描けます。また、灰色の部分を足し合わせることもできます。もしも、3 カテゴリの比率が (0.9, 0.05, 0.05) だという仮説に立っていたら、プロットは図13.1 (b) のようになります。通常は、このような特別な比率の仮説が棄却されるかどうかは問題にしませんが、考え方としては同じことです。



## 13.2 分割表検定

次に、2 カテゴリについて、集団からケース (G1) とコントロール (G2) とをサンプリングした場合を考えます。観察データは  $2 \times 2$  分割表として得られます。

表 13.1 観察データ

	A	a	計
G1	$n_{11} = 15$	$n_{12} = 25$	$n_{1\cdot} = 40$
G2	$n_{21} = 15$	$n_{22} = 45$	$n_{2\cdot} = 60$
計	$n_{\cdot 1} = 30$	$n_{\cdot 2} = 70$	$n_{\cdot \cdot} = 100$

G1 と G2 が属する集団には A, a の比率がありますが、それが今はわからぬいとします。G1 と G2 のサンプルはどちらも、この集団からランダムに抜き取られたものであるという仮説を立てます。そして、その仮説に立ったときに、観察データがどれくらいありふれたものなのか、珍しいものなのかを評価して、仮説を棄却するかどうかを考えてみます。

この集団の 2 カテゴリの比率は特定しませんが、変数を使って、 $p, 1 - p$  とします。G1, G2 ともに、この比率からのサンプリングですから、表のようなサンプリングがなされる確率は、

$$\frac{n_{1\cdot}!}{n_{11}! n_{12}!} \frac{n_{2\cdot}!}{n_{21}! n_{22}!} p^{n_{11}} (1-p)^{n_{12}}$$

です。今、 $p$  の値がなんであれ、ある値に確定しているとすれば、与えられた周辺度数の条件のもとで  $n_{11}, n_{12}, n_{21}, n_{22}$  と観察される確率は、

$$\frac{n_1!}{n_{11}!n_{12}!} \frac{n_2!}{n_{21}!n_{22}!}$$

に比例したものになります。実際、

$$\frac{n_1!}{n_{11}!n_{12}!} \frac{n_2!}{n_{21}!n_{22}!} \times \frac{n_1!n_2!}{n_{..}!}$$

としてやると、与えられた周辺度数のもとで、観察されうるすべての  $n_{ij}$  の場合について足し合わせたときに 1 になるので、 $p$  の値がなんであれ、G1, G2 が同じ集団からのサンプルであるという仮定のもとでの、表の観察確率は上の式になります。これに基づいて、他の観察されうる表と観察の珍しさを比較して  $p$  値化すれば、それがこの  $2 \times 2$  表において因子の独立に関するフィッシャーの**正確確率検定**となります。

さて、今、表 13.1 の周辺度数を満足する表を変数  $\delta$  を用いて、表 13.2 のように表します。

表 13.2 変数  $\delta$  を用いた表

	A	a	計
G1	$n_{11} = \frac{n_1 n_{..}}{n_{..}} + \delta$	$n_{12} = \frac{n_1 n_{..}}{n_{..}} - \delta$	$n_{1..} = 40$
G2	$n_{21} = \frac{n_2 n_{..}}{n_{..}} - \delta$	$n_{22} = \frac{n_2 n_{..}}{n_{..}} + \delta$	$n_{2..} = 60$
計	$n_{..1} = 30$	$n_{..2} = 70$	$n_{..} = 100$

この  $\delta$  を 0 から少しづつ増やしていきます。

R を用いて次のように計算します (R13-2.R)。

#### R ソース 13.2 R13-2.R : 正確確率とカイ自乗検定

```

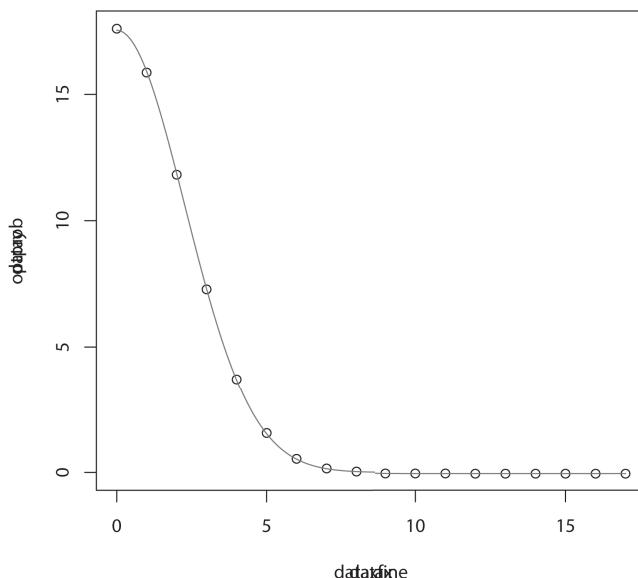
1 # 期待値表を作る
2 makeExptable<-function(m = matrix(c(10, 20, 30, 40),
nrow = 2)) {
3   m1 <- apply(m, 1, sum); m2 <- apply(m, 2, sum);
N <- sum(m); etable <- m1 %*% t(m2)/N

```

```

4 }
5 # 2x2 表の第1セルを期待値から1ずつ動かして生起確率を返す
6 ProbDistance <- function(m = matrix(c(10, 20, 30, 40),
7 nrow = 2)) {
8 # 周辺度数・期待値表を作る
9 m1 <- apply(m, 1, sum)
10 m2 <- apply(m, 2, sum)
11 etable <- makeExptable(m)
12 N<-sum(etable)
13 d <- seq(from = 0, to = 20, by = 1) # 期待値からのずれ
14 # 4セルとも正であるような2x2表の4値を作る
15 x <- d + etable[1, 1]; y <- -d + etable[1, 2];
16 z <- -d + etable[2, 1]; w <- d + etable[2, 2]
17 mat <- matrix(c(x, y, z, w), ncol = 4);
18 mins <- apply(mat, 1, min); matOK <- mat[mins > 0, ]
19 # 生起確率の対数の周辺度数部分
20 tmp <- sum(lgamma(m1 + 1)) + sum(lgamma(m2 + 1)) -
21 lgamma(N)
22 # 表ごとに正確生起確率を算出
23 prob <- exp(-apply(lgamma(matOK+1), 1, sum)+tmp)
24 list(x = d[mins > 0], prob = prob)
25 # xに期待値からのずれ、probに生起確率
26 }
27 d <- ProbDistance(m = matrix(c(10, 20, 30, 40), nrow = 2))
28 #c(10,20,30,40) の2x2表について実施
29 ylim <- c(0, max(d$prob))
30 plot(d$x, d$prob, ylim = ylim)
31 # optim() 関数を使って推定する関数を作る
32 fForOptim <- function(x) {
33 a <- x[1]; b <- x[2] ;c<-x[3]# 2変数関数
34 # optim() 関数が目指すのは、a * exp(-b*x^2) と観察値との距離の
35 # 自乗を全観察点について合算したもの
36 sum((d$prob - (a * exp(-b * d$x^2)))^2)
37 }
38 optout <- optim(c(1, 1), fForOptim)
39 # 推定する関数の係数の初期値は1,1から開始させる
40 xfine <- seq(from = min(optout$par), to = max(optout$par),
41 by = 0.01) # 横軸の値
42 optprob <- optout$par[1] * exp(-optout$par[2] * xfine^2)
43 # 推定関数の値
44 par(new=T) # 表の正確生起確率に推定関数を重ねて描く
45 plot(xfine, optprob, ylim = ylim, col = "red", type = "l")

```



2×2分割表を期待値表から少しずつずらして得られる表の正確生起確率を○で表示し、そのプロットに近似曲線を重ね合わせています。

図 13.2 正確確率とカイ自乗検定

すると G1, G2 が同じ集団からのサンプルだとした場合の生起確率は、図 13.2 の○のようにだんだん小さくなります。

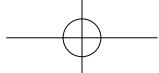
この○の作るカーブが何かしらの規則に則っているかもしれないと予想してみます。 $\delta$  の 2 乗に応じて小さくなっていくのではないかと、当たりをつけてみます。

$$a \times e^{-b\delta^2}$$

に近似できるかもしれないということです。

R の `optim()` 関数<sup>※1</sup>を用いて  $a$ ,  $b$  の近似値を求めてみます。そのようにして推定した関数の曲線が図 13.2 の曲線です。非常によく一致します。つまり、2×2 表の表の生起確率は、期待値からのずれ  $\delta$  の 2 乗に応じて減じていくことがわかりました。2 乗で減じるのは、正規分布・カイ分布・カイ自乗分布

※1 `optim()` 関数は係数とともに指定した関数への最適化関数です。`help(optim)` で確認できる通り、係数の値の求め方がいくつか実装されています。第 13 章では、線形近似に使用しています。



の性質でした（第8章）。この性質から、 $2 \times 2$ 分割表では期待値からのずれをカイ自乗分布で評価できるのです。また、 $2 \times 2$ 表は自由度が1の場合ですが、同様に任意の自由度の表でも、こうした対応関係があるので、カイ自乗値を用いた検定（ピアソンの独立性検定）が有用となります。

### ◎ 13.2.1 ピアソンの独立性検定——カイ自乗検定

$N \times M$  サイズの任意の表についてピアソンの独立性検定（カイ自乗検定）が行えます。

観察表とその周辺度数から独立を仮定して作る期待値表との違いを表13.3 のように数値化します。

表 13.3 観察表と期待地表

観察表							期待地表						
	B1	...	Bj	...	BN	計		B1	...	Bj	...	BN	計
A1	$n_{11}$	...	$n_{1j}$	...	$n_{1N}$	$n_{1\cdot}$	A1	$e_{11}$	...	$e_{1j}$	...	$e_{1N}$	$n_{1\cdot}$
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Ai	$n_{i1}$	...	$n_{ij}$	...	$n_{iN}$	$n_{i\cdot}$	Ai	$e_{i1}$	...	$e_{ij}$	...	$e_{iN}$	$n_{i\cdot}$
...	...	...	...	...	...	...	...	...	...	...	...	...	...
AM	$n_{M1}$	...	$n_{Mj}$	...	$n_{MN}$	$n_{M\cdot}$	AM	$e_{M1}$	...	$e_{Mj}$	...	$e_{MN}$	$n_{M\cdot}$
計	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot N}$	$n_{\cdot \cdot}$	計	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot N}$	$n_{\cdot \cdot}$

$$e_{ij} = \frac{n_i \cdot n_{\cdot j}}{n_{\cdot \cdot}}$$

独立性の検定のピアソンのカイ自乗値は次の式で与えられます。

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

縦軸と横軸が独立であるときに、この値を観察する確率が自由度  $(M-1) \times (N-1)$  のカイ自乗分布の確率密度分布と似ていることを利用して  $p$  値化します。

R13-3.R から、検定の計算法を読み取ってください。

### R ソース 13.3 R13-3.R : カイ自乗検定

```

1 chisqCalc<-function(m=matrix(c(10,20,30,40),nrow=2)) {
  # 行列を引数とする
2   etable<-makeExptable(m)
3   m1<-length(m[,1]);m2<-length(m[1,]) # 行・列の数
4   chi2<-sum((m-etable)^2/etable)
      # 全セルについて、(観測値 - 期待値)^2 / 期待値
5   df<-(length(m1)-1)*(length(m2)-1) # 自由度は (N-1)x(M-1)
6   p<-pchisq(chi2,df,lower.tail=FALSE)
      # p 値は、カイ自乗分布から得る
7   list(chi2=chi2,p=p,df=df)
8 }
9 chisqCalc(m=matrix(c(10,20,30,40),nrow=2))

```

なお、この関数は、R の関数 "chisq.test(correct=FALSE)" と同じことです<sup>※2</sup>。

## ● 13.2.2 帰無仮説と最尤仮説を比較して統計量にする——尤度比検定

13.2 節の冒頭（13.2.1 項の前）の正確確率検定では、G1, G2 が同一集団からのサンプルであるという仮説のもとでの表の観察確率を使って珍しさを算出しました。今度は、その裏側、尤度を使います。

ある表が観察されたときに、G1, G2 が、因子保有率が異なる 2 集団からのサンプルであるという仮説に立ちます。2 集団の因子保有率には観察データから得られる**最尤推定値**  $\frac{n_{11}}{n_1}, \frac{n_{21}}{n_2}$  を仮定して、その仮定のもとで、観察データが観察される確率をまず考えます。

---

※2 “correct=FALSE”とは*イエーツの連続性補正*をしないというオプションです。イエーツの連続性補正とは、セルの値が小さめのときに行なうことが推奨される補正で、カイ自乗値を小さめ（p 値を大きめ）にします。本書では、この補正については取り扱いません。特定の条件の場合のみに補正を入れると、本書で行っている方法間の比較に際して、数値の整合性を悪くするなどの影響があるためです。なお、補正が必要な場合には、正確確率検定によって、その代用をすることを勧めます。

$$\begin{aligned} L(G1 \neq G2) &= \frac{n_{..}!}{n_{11}! n_{21}!} \frac{n_{..}!}{n_{12}! n_{22}!} \left( \frac{n_{11}}{n_{..}} \right)^{n_{11}} \left( \frac{n_{12}}{n_{..}} \right)^{n_{12}} \left( \frac{n_{21}}{n_{..}} \right)^{n_{21}} \left( \frac{n_{22}}{n_{..}} \right)^{n_{22}} \\ &= \frac{n_{..}!}{n_{11}! n_{21}!} \frac{n_{..}!}{n_{12}! n_{22}!} \frac{n_{11}^{n_{11}} n_{12}^{n_{12}} n_{21}^{n_{21}} n_{22}^{n_{22}}}{n_{1.}^{n_{1.}} n_{2.}^{n_{2.}}} \end{aligned}$$

次に、G1, G2 とが同一集団からのサンプルであるとしたときの、その保有率の最尤推定値  $\frac{n_{..}}{n_{..}}$  を仮定して、その仮定のもとで、観察データが観察される確率を考えます。

$$\begin{aligned} L(G1 = G2) &= \frac{n_{..}!}{n_{11}! n_{21}!} \frac{n_{..}!}{n_{12}! n_{22}!} \left( \frac{n_{..}}{n_{..}} \right)^{n_{11}+n_{21}} \left( \frac{n_{..}}{n_{..}} \right)^{n_{12}+n_{22}} \\ &= \frac{n_{..}!}{n_{11}! n_{21}!} \frac{n_{..}!}{n_{12}! n_{22}!} \frac{n_{..}^{n_{..}}}{n_{..}^{n_{..}}} \end{aligned}$$

G1, G2 が同一集団からのサンプルであるとしたときの尤度の方が小さくなりますが、その尤度の違いを数値にします。小ささを測るのに、比をとることにします。この比を**尤度比**と言います。これは、G1, G2 が異なる集団からのサンプルだという仮説を立てて、最も尤度が高くなるようにした場合と、G1, G2 が同じ集団からのサンプルだという仮説とを比較したものです。

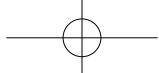
比をとると階乗の部分はきれいになくなって、

$$\frac{L(G1 \neq G2)}{L(G1 = G2)} = \frac{n_{..}^{n_{..}} n_{11}^{n_{11}} n_{12}^{n_{12}} n_{21}^{n_{21}} n_{22}^{n_{22}}}{n_{1.}^{n_{1.}} n_{2.}^{n_{2.}} n_{1.}^{n_{1.}} n_{2.}^{n_{2.}}}$$

実際、この式は、 $n_{..} = \sum_j n_{ij}$ ,  $n_{..j} = \sum_i n_{ij}$  に注意して  $N \times M$  に一般化すると、

$$\frac{L(G1 \neq G2)}{L(G1 = G2)} = \prod_{i,j} \left( \frac{n_{ij}}{e_{ij}} \right)^{n_{ij}}$$

となることが示せます。そして、この値の対数の 2 倍が、この表のピアソンのカイ自乗値とほぼ一致します。したがって、この値（尤度比の対数の 2 倍）を



用いても、帰無仮説の棄却検定を行うことができます。**尤度比検定**と呼ばれる手法です。

R13-4.R でその処理を確認してください。

#### R ソース 13.4 R13-4.R : 尤度比検定

```
1 likelihoodRatioTest<-function(m=matrix(c(10,20,30,40),  
2 nrow=2)){  
2   etable<-makeExptable(m)  
3   chi2<-2*sum(log(m/etable)*m)  
4   df<-(length(m[,1])-1)*(length(m[1,])-1)  
5   p<-pchisq(chi2,df,lower.tail=FALSE)  
6   return(list(statistic=chi2,p.value=p,df=df))  
7 }  
8 likelihoodRatioTest(m=matrix(c(10,20,30,40),nrow=2))
```

## 13.3 3つの検定方法の比較——正確確率検定、ピアソンの独立性検定、尤度比検定

さて、同じ分割表に 3 種類の相互に似た結果を返す検定手法が 3 つ出てきました。それぞれの特徴を知るために、比較してみることにします。 $p$  値で較べます。

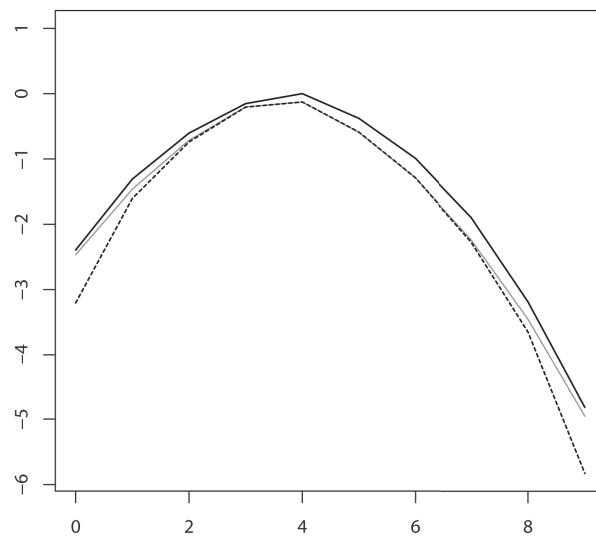
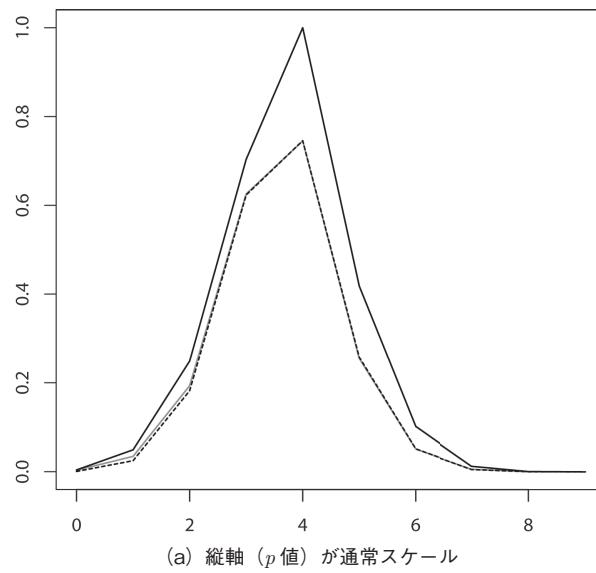
### ● 13.3.1 サンプル数が小さいときと大きいとき

#### (1) サンプル数が小さいとき

サンプル数が小さいとき (3, 6, 9, 12) と、大きいとき (100, 200, 300, 400) の 2 通りで、ピアソンの独立性検定、尤度比検定、フィッシャーの正確確率検定の結果を較べてみます。

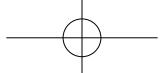
図 13.3、図 13.4、図 13.5 を描く R のソースは掲載しませんが、ダウンロードできる R13-sup1.R です。

まず、サンプル数が小さいときです。図 13.3 を見てください。横軸に第 1 セルの値を、縦軸に  $p$  値をとりました。縦軸を通常スケールにした (a) では、全般的に大きめの値が出ているのが正確検定で、小さめな値が出ていてほぼ一致しているのがピアソンの方法と尤度比検定です。一方、縦軸を対数スケ



黒の実線が正確検定。灰色の実線がピアソンの独立性検定、破線が尤度比検定。

図 13.3 サンプル数が小さいとき



ールにすると (b)、横軸の両端 ( $p$  値が小さい範囲) で、尤度比検定が他の 2 法とずれている傾向があります。この範囲ではピアソン法と正確法が近くなっています。

このように、サンプル数が小さいときは正確確率検定に較べて、カイ自乗分布になぞらえた検定（ピアソンの方法と尤度比検定）は  $p$  値が小さめに出ます。 $p$  値が小さく出るということは、「統計的に有意」な判定を下しやすいということから、好ましくありません（検定では、 $p$  値を大きめに算出することを「保守的」と呼んで、尊重することが多いです）。カイ自乗分布は整数にも実数にも対応した連続な分布であるのに、観察は整数であるために、離散的なデータとずれが生じます。

サンプル数が大きくなると離散的とは言っても、かなり滑らかな動きになりますが、サンプル数が小さいと滑らかでなくなるために、連続なカイ自乗分布による近似がうまくいかなくなっているわけです。このことから、サンプル数が小さいときの、カイ自乗分布になぞらえた検定については、解釈を慎重にすることが必要で、正確検定を使う方が適当である場合が生じます。

また、正確検定の値に近づくように補正する方法などが提唱されたりしています。これは、信頼区間の推定方法として提案されていくつかの方法が、やはり、分布の連続性がもたらす不都合を補正するものとの同じことです<sup>※3</sup>。

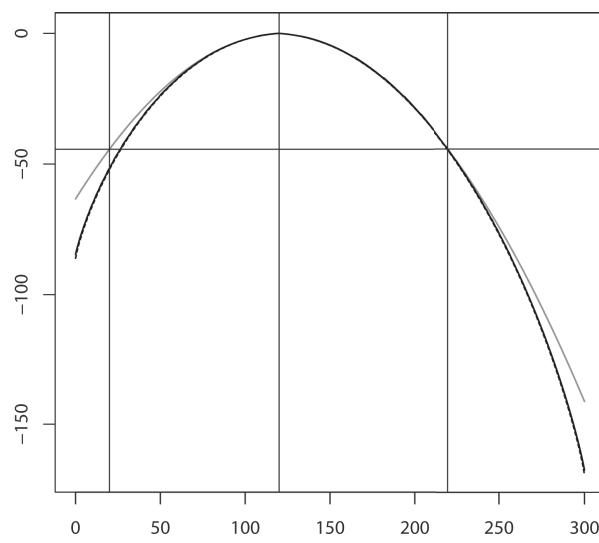
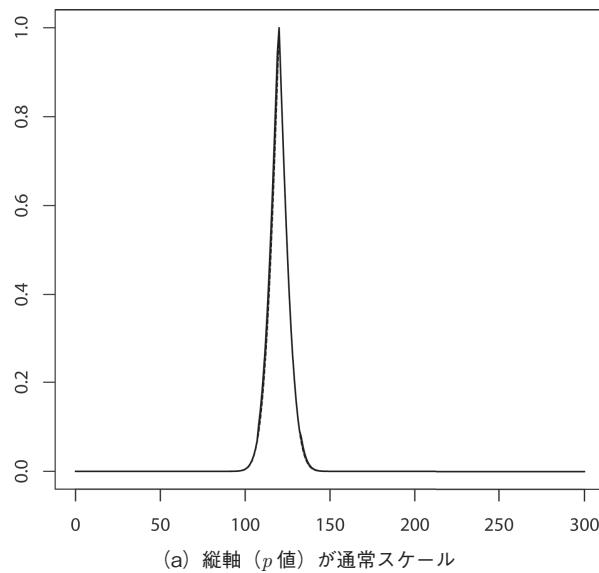
## (2) サンプル数が大きいとき

次に、サンプル数を大きくした場合を見ます。図 13.4 を見てください。

サンプル数が小さいときとの違いは、グラフが滑らかになっている点です。それ以外の違いとしては、縦軸が通常スケールのときには、3 法でほとんど差がないことです。このように、3 種類の  $p$  値は、サンプルの数を大きくするとほとんど一致するほどに、似てきます。この、「完全には一致しないけれども、ほぼ同じとみなせるくらいに似ていく」ということを**漸近近似**と言うので、カイ自乗検定は漸近近似検定の 1 つです。

一方、縦軸を対数スケールにすると、 $p$  値が非常に小さい範囲でずれている様子がわかります。ずれているのがピアソンの方法です。正確検定と尤度比検定はほぼ一致しています。

※3 “correct=FALSE” でコメントしたイエーツの連続性補正是、サンプル数が小さいときに起きるこの影響の補正法です。



黒の実線が正確検定、灰色の実線がピアソンの独立性検定、破線が尤度比検定。

図 13.4 サンプル数が大きいとき

### ● 13.3.2 検定の対称性

図 13.4 (b) をもう一度見ることにしましょう。

サンプル数が大きい場合で縦軸を対数スケールにしたものです。この図で、 $p$  値が最大となるのは、期待値表 ( $x = 120$ ) のときです。そこから、 $x$  が 100 減ったときと 100 増えたときとを、縦線で示しております。 $x = 120 + 100 = 220$  のときの 3 つの  $p$  値はほぼ同じですが、その点を通る水平線を引くと、ピアソンの方法の  $p$  値は、 $x = 120 - 100 = 20$  のときに同じ値をとっていることがわかります。このことから、ピアソンの方法の  $p$  値は左右対称であることがわかります。

他方、尤度比検定と正確検定の  $p$  値は、それよりも小さい値であることがわかります。この 2 つの手法は左右が非対称になっています。サンプル数が大きく、期待値表からさして遠くない範囲では、この非対称性は問題になりませんが、手法の性質として対称性のあり・なしという違いがあることがわかります。

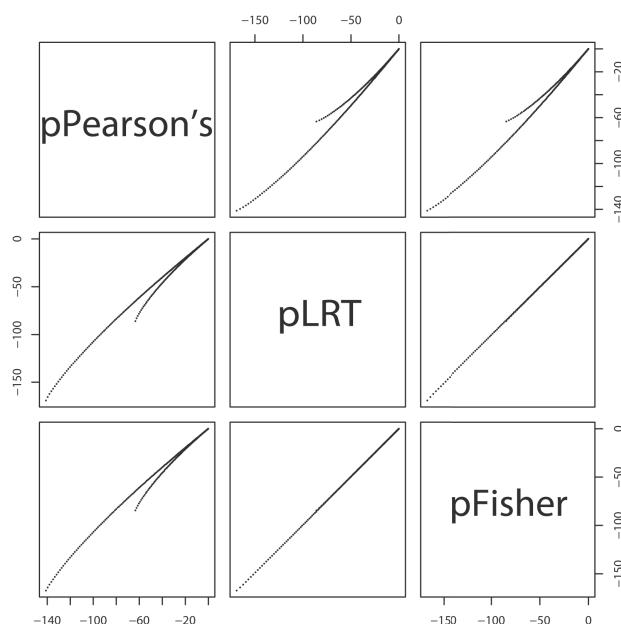
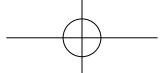


図 13.5 pPearson's : ピアソンの方法の  $p$  値、pLRT : 尤度比検定の  $p$  値、pFisher : フィッシャーの正確確率検定の  $p$  値



サンプル数が大きい場合の3方法の $p$ 値の一致の様子を対数スケールで散布図にしてみます。尤度比検定と正確検定はほぼ完全一致している様子が見えますが、ピアソンの方法は、その他2法との関係で、2本の線が現れます。

この2つの線のうちの片方は、期待表から(1, 1)セルが増える方向の表に由来し、もう片方の線は、逆に減る方向の表に由来しています。

### ◎ 13.3.3 有限な範囲と無限の広がりの違い

対称性を確認した図13.4 (b)を見ると、別のことにも気づきます。 $x$ 軸の両端寄りの部分では、ピアソンの方法の $p$ 値に較べて、正確確率検定と尤度比検定の $p$ 値が小さい傾向があります。この傾向は、サンプル数が小さいときに現れませんが、サンプル数が大きくなると見えてきます。

これは、ピアソンの方法が、 $x$ の範囲を $-\infty \leq x \leq \infty$ としているのに対して、正確確率検定と尤度比検定では、図に表示した $x$ の範囲の外側は、ありえない範囲（観察数が負になる、比率が負になる）としている、という違いに由来します。ピアソンの方法では、ありえない領域にも小さいながらもある程度の確率を割り振り、その分、期待値に近い領域の確率をほんのわずかずつですが小さめにしています。そして、ありえない領域の確率は観察データよりも珍しいとして、 $p$ 値の計算に足し合わせています。ありえるけれども期待値から遠いデータの場合には、このありえない確率の分が無視できない大きさとなるために、 $p$ 値が大きめに出てくるわけです。

正確確率検定と尤度比検定ではありえない領域に確率を割り振らないので、この影響を受けていません。この影響は、 $p$ 値が非常に小さい場合にのみ現れます、検定方法の特性の1つです。

### ◎ 13.3.4 計算量の違い

この他の大きな違いに、計算量の多寡があります。

正確確率検定はすべての場合を数え上げるので、サンプル数の自由度乗のオーダーで計算量が増えます。したがって、自由度が大きい場合には非常に計算負荷が大きくなり、実際的ではなくなります。

### ◎ 13.3.5 計算量の違いのまとめ

ここまで見てきた3方法の特徴を表にします。

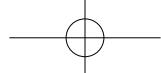


表 13.4 計算量の違い

方法	計算量	少サンプルのとき	対称性	漸近近似	連続・離散
正確確率検定	多	保守的・正確	非対称	正確	離散
ピアソンの方法	少	正確検定との乖離が大きい	対称	漸近近似	連続
尤度比検定	少	正確検定との乖離が大きい	非対称	漸近近似	連続

## 13.4 仮説に制約を定めて検定する

分割表について、行と列とが独立であるという仮説の棄却をする方法が複数あることを述べてきました。

本節では、同一の分割表に対して、対立仮説の立て方にいろいろな制約を定めた上で、検定することを考えます。対立仮説に定める制約をモデルとも言います。モデルと変数の関係、自由度について少し話を進めます。

ある 2 アレル型座位で、次のような  $2 \times 3$  表が得られたとします。

	AA	Aa	aa	計
ケース	$n_{12} = 30$	$n_{11} = 55$	$n_{10} = 15$	$n_{1\cdot} = 100$
コントロール	$n_{22} = 16$	$n_{21} = 48$	$n_{20} = 36$	$n_{2\cdot} = 100$
計	$n_{\cdot 2} = 46$	$n_{\cdot 1} = 103$	$n_{\cdot 0} = 51$	$n_{\cdot \cdot} = 200$

尤度比検定では、独立仮説とデータから最尤推定した変数を仮説として比較の相手としました。尤度はどんな仮説にも計算できますから、尤度比もさまざまな 2 つの仮説について計算することができます。しばしば立てる仮説にどんなものがあるかを見てみることにします。

- 仮説 1. ジエノタイプとフェノタイプに関係がないという仮説（帰無仮説）
- 仮説 2. A は優性遺伝形式でフェノタイプに影響しているというモデルに基づく仮説
- 仮説 3. A は劣性遺伝形式でフェノタイプに影響しているというモデルに基づく仮説
- 仮説 4. A は相加的遺伝形式でフェノタイプに影響しているというモデルに基づく仮説
- 仮説 5. A は優性遺伝形式・相加的遺伝形式・劣性遺伝形式のいずれか 1

つで影響しているというモデルに基づく仮説<sup>※4</sup>

仮説6. ジエノタイプとフェノタイプには、(どういう形式にしろ、なんでもよいので) 独立ではない関係があるという仮説

3 ジエノタイプのリスクが同一であるという仮説（仮説1）は、ジエノタイプ頻度を決めるのに2 (=3-1) 変数、有病率を定めるのに1 変数の合計3 変数で定めるモデル（自由度3）です。

仮説6は、ケースとコントロールの相対危険度を自由に決めるために2 変数増やして、全部で変数5 個を必要とするモデルです（自由度5）。自由度が2 つ増えました。優性遺伝形式・劣性遺伝形式では、ヘテロ接合体の相対危険度 ( $RR$ ) がリスクホモ接合体と同じか、非リスクホモ接合体と同じなので、自由度は1 増えるだけで4になります。

ヘテロ接合体のリスクが2つのホモ接合体のリスクの算術平均の場合（仮説5（相加的モデル））は、ヘテロ接合体のリスクを2つのホモ接合体の「中間」に定めている仮説ですが、この場合も、ヘテロ接合体のリスクが中間に固定されているので自由度は4です。仮説5では、3つの遺伝形式（優性・劣性・相加的）がありえるとするモデルですが、以下の表のように、 $k$ という変数を用いて、ヘテロ接合体のリスクを表すことができて、これによりある程度の自由があります。しかしながら、どんな値でもとれる、というのに較べ、大幅に限定的ですから、この場合は自由な変数4個と制約された変数1個となり、全体の自由度は4より大きいですが、5よりは小さいです。

表 13.5 6つの仮説の比較

仮説	AA リスク	Aa リスク	自由な 変数の数	限定的な 変数の数
1	0	0	3	0
2	$1 < RR_{AA} \leq \infty$	$RR_{Aa} = RR_{AA}$	4	0
3	$1 < RR_{AA} \leq \infty$	1	4	0
4	$1 < RR_{AA} \leq \infty$	$RR_{Aa} = \frac{1}{2}(RR_{AA} + 1)$	4	0
5	$1 < RR_{AA} \leq \infty$	$RR_{Aa} = k(RR_{AA} - 1) + 1; k = 0, 0.5, 1$	4	1
6	$0 \leq RR_{AA} \leq \infty$	$0 \leq RR_{Aa} \leq \infty$	5	0

※4 3 遺伝形式の検定統計量の最大値を採用する方法 MAX3 テストで検定します。Rなら Rassoc パッケージの MAX3() 関数。

今、この6つの仮説から2個を取り出して、それぞれの尤度を計算し、それを比較することが可能です。6仮説同士の比較を表にすると表13.6のようになります。表の左上から右下の対角線の上側には、仮説間の自由度の差が書き込まれ、対角線の下側には、比較・検定手法の例が記されています。仮説5との比較の場合には、不自由な変数の分の自由度が整数で表せないので、その分を $\alpha$ としてあります。

表13.6 6つの仮説同士の比較

仮説	1	2	3	4	5	6
1		1	1	1	$1 + \alpha$	2
2	$2 \times 2$ 表ピアソン ( $df = 1$ )		0	0	$\alpha$	1
3	$2 \times 2$ 表ピアソン ( $df = 1$ )	LR比較		0	$\alpha$	1
4	$2 \times 3$ 表傾向性検定 ( $df = 1$ )	LR比較	LR比較		$\alpha$	1
5	MAX3 テスト	—	—	—		$1 - \alpha$
6	$2 \times 3$ 表ピアソン ( $df = 2$ )	LRT ( $df = 1$ )	LRT ( $df = 1$ )	LRT ( $df = 1$ )	—	

※ LR：尤度比、LRT：尤度比検定

### ● 13.4.1 1つの分割表にいろいろな検定を適用してみる

仮説1との比較においてはピアソンの独立性検定が適用できます。 $2 \times 2$ 表を自由度1のカイ自乗検定する場合と、 $2 \times 3$ 表を自由度2のカイ自乗検定する場合とがあります。尤度比を用いて、自由度 $k$ のカイ自乗分布に従う統計量を算出して検定することもできます。自由な変数の少ない方の仮説の棄却検定になります。

また、仮説5（相加的遺伝形式）の場合には、 $2 \times 3$ 表について、3ジエノタイプに線形のリスクを想定して、**傾向性の検定**を実施することと同じです。特に、**Cockran-Armitage の傾向性検定**と呼ぶこともあります。優性・劣性・相加的形式は、傾向性の検定において、 $k = 1, 0, 0.5$ という係数を与えた傾向性検定と考えることもできます。

R13.5.Rでは、RassocパッケージのCATT()関数にその係数を引数として渡して算出させています。

R ソース 13.5 R13-5.R :  $2 \times 3$  分割表にいろいろな検定を行う

```

1 library(Rassoc)
2 # 2次元平面の座標から、観察テーブルを作成する
3 casecontRTheta<-function(R,t,af,f,N){
4   popW<-c(af^2+af*(1-af)*f,2*af*(1-af)*(1-f),
5   (1-af)^2+af*(1-af)*f)
6   table<-matrix(c(popW*N[1],popW*N[2]),nrow=2,byrow=TRUE)
7   table[1,1]<-table[1,1]-R/2*cos(t)+sqrt(3)/2*R*sin(t)
8   table[2,1]<-table[2,1]+R/2*cos(t)-sqrt(3)/2*R*sin(t)
9   table[1,2]<-table[1,2]+R*cos(t)
10  table[2,2]<-table[2,2]-R*cos(t)
11  table[1,3]<-table[1,3]-R/2*cos(t)-sqrt(3)/2*R*sin(t)
12  table[2,3]<-table[2,3]+R/2*cos(t)+sqrt(3)/2*R*sin(t)
13  return(table)
14 }
15 af<-0.4;f<-0;N<-c(100,100)
16 # allele frequency, HWE-f, sample size
17 x<-y<-seq(from=-10,to=10,by=1)
18 #2次元格子点のテーブルを作成し各種検定
19 cattp<-max3p<-df2p<-domp<-recp
20   <-matrix(rep(0,length(x)*length(y)),nrow=length(x))
21 for(i in 1:length(x)){
22   for(j in 1:length(y)){
23     R<-sqrt(x[i]^2+y[j]^2)
24     if(x[i]==0){
25       t<-pi/2
26     }else{
27       t<-atan(y[j]/x[i])
28     }
29   }
30   popdata<-casecontRTheta(R=R,t=t,af=af,f=f,N=N)
31   if(min(popdata)>0){
32     catout<-CATT(popdata,x=0.5)
33       # Cochran-Armitage の傾向性検定を係数 0.5 で実施
34     domout<-CATT(popdata,x=1)
35       # Cochran-Armitage の傾向性検定を係数 1 で実施
36     recout<-CATT(popdata,x=0)
37       # Cochran-Armitage の傾向性検定を係数 0 で実施
38     max3out<-MAX3(popdata)
39     df2chi2out<-chisq.test(popdata,2)
40     cattp[i,j]<-catout$p

```

```

35   max3p[i,j]<-max3out$p
36   df2p[i,j]<-df2chi2out$p.value
37   domp[i,j]<-domout$p
38   recp[i,j]<-recout$p
39   }
40   }
41   }
42   filled.contour(x,y,-log(cattp,10),color=terrain.colors)
43   filled.contour(x,y,-log(max3p,10),color=terrain.colors)
44   filled.contour(x,y,-log(df2p,10),color=terrain.colors)
45   filled.contour(x,y,-log(domp,10),color=terrain.colors)
46   filled.contour(x,y,-log(recp,10),color=terrain.colors)

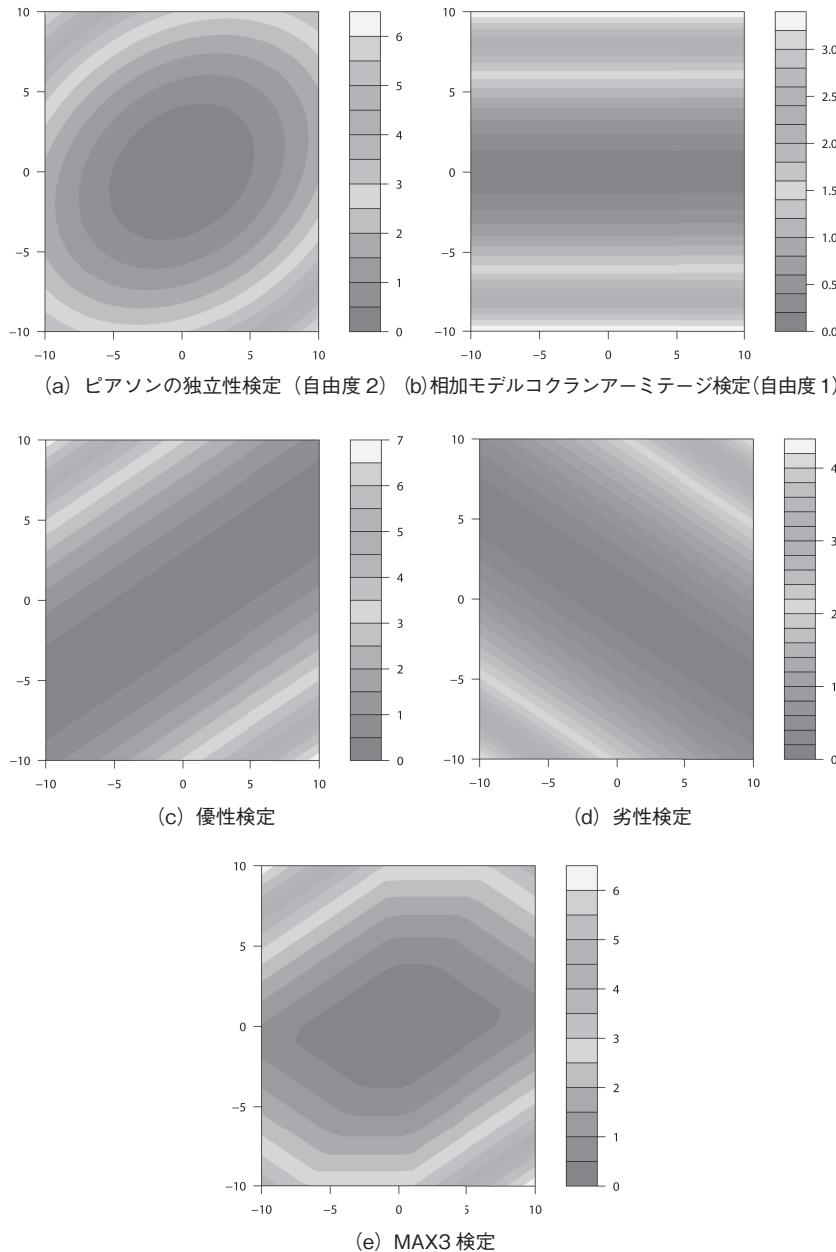
```

いずれにしろ、モデルのおき方に合わせて手法を選んでいるということになります。表で示した分割表の周辺度数が与えられているとします。このとき、G1 の 3 ジエノタイプのカウントを決めると、G2 のカウントも決まります。また、G1 の 3 ジエノタイプカウントは 3 カテゴリで自由度が 2 なので、正三角形上の点として表せることになります（第 4 章の正単体）。そのような座標をとり、それぞれの表について、自由度 2 のピアソンの独立性検定を実施すると、その  $p$  値が等しい表は楕円を描きます（図 13.6 (a)）。

同様に、相加的モデル (b)、優性モデル (c)、劣性モデル (d) の  $p$  値の等高線は直線を作ります。MAX3 検定 (e) は、3 つのモデルを合わせたものなので、等高線もこの 3 つのモデルを合わせたものになっています。

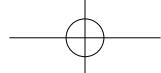
このことからわかるのは、以下のことです。

- 周辺度数を共有する  $2 \times 3$  表は、自由度 2 の平面に分布させることができて、自由度 2 の検定は、そこに楕円を描くこと
- その表について自由度 1 の検定をすることは、直線状の等高線を引いてデータを評価すること
- それ以外にも等高線の引き方が作れること
- 楕円や直線の等高線を引く検定には、検定手法が用意されていること



それぞれ  $p$  値の常用対数  $x$  ( $-1$ ) の等高線を描いています。

図 13.6 G1 のジェノタイプカウントを正三角形座標上に表したもの



### ● 13.4.2 離散的な仮説空間での尤度比の比較

仮説の比較方法の表には、2つの仮説しかありえなくて、そのどちらかの判断をするような場合もあります。親子の判定（親子なのか、そうでないのか）はこの部類に入ります。これらの場合には、次のような検定になるでしょう。

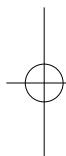
2つの仮説だけが「仮説の空間」です。したがって、2つの仮説から算出される尤度 ( $L(h_1) + L(h_2)$ ) を足し合わせたものが、仮説空間の全体にわたる尤度の和です。したがって、2つの尤度の和に占める、それぞれの仮説の尤度の割合

$$\frac{L(h_i)}{L(h_1) + L(h_2)}; i = 1, 2$$

が仮説  $h_i$  を信じるべき程度になります。特に、確率分布などを用いる必要のない判断になります。



## 13.5 検定同士の非独立な関係



さて、前節では1つの $2 \times 3$ 表にいろいろな仮説を立てて、いろいろな較べ方をしてみました。自由度1の検定が3種類（優性、劣性、相加的）、自由度2の検定（ピアソンの独立性検定）が1種類ありました。これらの  $p$  値は、相互に独立ではありません。

その様子を見てみることにします。適当に表を作り、その検定結果を比較してみます（R13-6.R）。

#### R ソース 13.6 R13-6.R : 検定同士の非独立な関係

```
1 library(Rassoc)
2 Niter<-1000
3 Nca<-1000
4 Nco<-1000
5 chiP<-cattP<-domP<-recP<-rep(0,Niter)
6 for(i in 1:Niter){
7   af<-runif(1)*0.6+0.2
8   df<-c(af^2,2*af*(1-af),(1-af)^2)
```

```

9  case<-sample(c(0,1,2),Nca,df,replace=TRUE)
10 cont<-sample(c(0,1,2),Nco,df,replace=TRUE)
11 m<-matrix(c(length(which(case==0)),length(which(case==1)),
12   length(which(case==2)),length(which(cont==0)),
13   length(which(cont==1)),length(which(cont==2))),
14   nrow=2,byrow=TRUE)
15 chiP[i]<-chisq.test(m,correct=FALSE)$p.value
16 cattP[i]<-CATT(m,0.5)$p
17 domP[i]<-CATT(m,1)$p
18 recP[i]<-CATT(m,0)$p
19 }
20 plot(as.data.frame(cbind(chiP, domP, cattP, recP)),
21       cex = 0.1)
22 cormat <- cor(cbind(chiP, domP, cattP, recP))
23 cormat

```

まず、独立な2つの検定があったとき、その2つの検定の  $p$  値には全く相関がありません。そのようなときにペアとなる  $p$  値の散布図は図 13.7 のように、ばらばらになります。

$p$  値間の相関を調べて数値で表せば次のようになり、関連があることがわかります。

### ■ R の出力結果

	chiP	domP	cattP	recP
chiP	1.0000000	0.63872128	0.6711991	0.63932199
domP	0.6387213	1.00000000	0.5568540	0.04097478
cattP	0.6711991	0.55685396	1.0000000	0.53840145
recP	0.6393220	0.04097478	0.5384015	1.00000000

優性・相加的・劣性の3検定の間では、優性と劣性とが比較的独立性が強いことがわかります。これは、優性と劣性のモデル同士が、優性と相加的のそれよりも「遠い」関係だからです。ピアソンの方法と3つの自由度1の検定との相関は同程度であることもわかります。

このように独立でない検定を複数、行ったときには、個々の検定の結果の解釈には注意が必要です。このことは、第17章の多重検定のところで扱います。

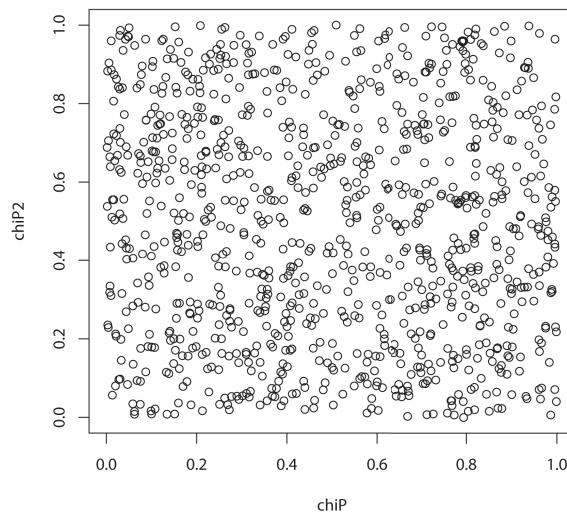
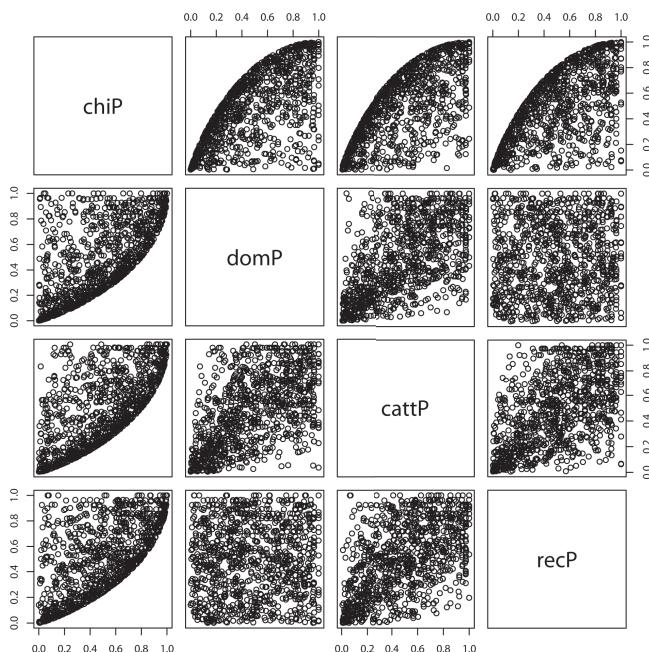
(a) 独立な 2 検定の  $p$  値の散布図(b)  $2 \times 3$  表に自由度 2 の独立性検定と 3 種類の遺伝形式モデルに関する自由度 1 の検定を実施したときの、4 検定 (chiP : ピアソンの方法、domP : 優性モデル、cattP : 相加的モデル、recP : 劣性モデル) の  $p$  値の散布図

図 13.7 検定同士の非独立な関係

## 13.6 表のサイズを変える

### ◎ 13.6.1 表形式のデータ

表というのは、行と列に2つの変数を当てはめたデータの表現方法の1つです。あるサンプルセットがあり、そのサンプルが2つの変数に関するデータを持っているときに作れるものです。そして、サンプルを通じた2変数の関係について検討することができます。本節では、この2つの尺度に関するデータについて、データ型ごとにどのような対応をするのかについて見ていきます<sup>※5</sup>。

### ◎ 13.6.2 順序のあり・なしと検定手法

#### (1) 検定の詳細はRのソースを調べよう

以降の説明では、いくつかの検定手法名が出てきます。それらの定義式などは掲載しません。その代わりにRのソースを提示し、各検定手法で用いられるRの関数がわかるようにしています。Rの関数の内容をどのように表示するかは付録Aに記載しましたので、検定の定義の確認はそれを活用してください。

#### Rソース 13.7 R13-7.R : 複数の分割表検定の結果を比較

```

1  # テーブルサイズ (NxM) と与えて各種検定結果の比較をプロットする
2  CompareTests<-function(N=2,M=2,Niter=1000,Ns=100,k1=10,
   k2=3){
3  library(MCMCpack)
4  library(clinfun)
5  pearsonp<-trendp<-kwp<-jtp<-lmp<-rep(0,Niter)
6  for(i in 1:Niter){
7  # $N\times M$ テーブルをランダムに作る
8  fn<-rdirichlet(1,rep(k1,N))
9  fm<-rdirichlet(1,rep(k2,M))
10 first<-sample(1:N,Ns,prob=fn,replace=TRUE)

```

<sup>※5</sup> 量的変数に関しては、その分布の様子などにより適切な解析手法が分かれますが、本書では、カテゴリカルなデータ型からの延長として量的尺度を眺めることに重点をおき、その点についてはあえて触れません。分布を仮定する手法を**パラメトリック**、仮定しない手法を**ノンパラメトリック**と呼びますので、基本的にはノンパラメトリックな手法を中心とした話になっています。

```

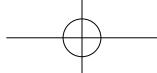
11 second<-sample(1:M,Ns,prob=fm,replace=TRUE)
12 t<-table(first,second) # 分割表を作る
13 pearsonp[i]<-chisq.test(t,correct=FALSE)$p.value
14                                     # ピアソンの独立性検定
15 if(N==2){
16   trendp[i]<-prop.trend.test(t[1,],t[1,]+t[2,],
17                               score=1:M)$p.value # レンド検定
18 }
19 kwp[i]<-kruskal.test(second~first)$p.value
20                                     # Kruskal-Wallis
21 jtp[i]<-jonckheere.test(second,first,
22   alternative="two.sided")$p.value # Jonckheere-Terpstra
23 lmp[i]<-summary(lm(second~first))$coefficients[2,4]
24                                     # 線形回帰
25 }
26 databind<-cbind(pearsonp,kwp,jtp,lmp,trendp)
27 plot(as.data.frame(databind))
28 return(databind)
29 }
30 N<-2;M<-2;Niter<-100;Ns<-1000 # 2x2 テーブル
31 cout22<-CompareTests(N,M,Niter,Ns)
32 N<-2;M<-3;Niter<-100;Ns<-1000 # 2x3 テーブル
33 cout22<-CompareTests(N,M,Niter,Ns)
34 N<-2;M<-10;Niter<-100;Ns<-1000 # 2x10 テーブル
35 cout22<-CompareTests(N,M,Niter,Ns)
36 N<-3;M<-3;Niter<-100;Ns<-1000 # 3x3 テーブル
37 cout22<-CompareTests(N,M,Niter,Ns)

```

## (2) 片方の軸が 2 カテゴリの場合

ピアソンの独立性検定は、「順序なしのカテゴリ × 順序なしのカテゴリ」の手法です。傾向性の検定は、片方の軸が 2 カテゴリであるときに、もう片方の軸が順序のあるカテゴリであるときの検定です。順序のあるカテゴリに 0, 1, 2, ... という重みを付けるのが最も単純なやり方ですが、重み付けにカーブをかけたりすることもできます。遺伝形式の場合で言えば、0, 1, 1 が優性モデル、0, 0.5, 1 が相加的モデルで「最も単純なやり方」、0, 0, 1 が劣性モデルです。

このように重みを付けるのは「線形な関係」に持ち込むためですが、重みを忘れてただ単に、値の大小関係のみを使うこともできます。値に応じて順位を付けるのです。片方の軸が 2 カテゴリで、もう 1 つの軸に大小を考慮して順位



につけるのは、順位和検定です。順序の方のカテゴリ数がどんどん増えて、同じカテゴリに含まれるサンプルが1つもないような状態にまでなると、サンプルの数だけカテゴリ数があるような状態になります。このときに、サンプルの値の大小順に順序を与えて検定するのが、順位和検定です。

量的データ型のときも、すべてのサンプルが異なる値を持っていて、その値に順序がつけられますから、同様に順位和検定ができます。片方の軸が順序のない3以上カテゴリで、もう片方の軸に順序がある場合には、クラスカル・ウォリス検定が使えます。2カテゴリは順位ありとみなせますから、「2カテゴリ×順序のない3以上カテゴリ」の場合にも、これを適用することが可能です。線形回帰を適用してしまうことも可能です。この場合は「線形」の回帰ですから、同じく「線形」を仮定している傾向性検定と似た挙動をとります。

### (3) 片方の軸が順序なしの3以上カテゴリの場合

片方の軸が3軸以上のカテゴリで順序なしになった場合には、もう片方の軸が順序なしであれば、ピアソンの独立性検定になります。もう片方の軸が順序ありの場合には、クラスカル・ウォリスの検定があります。片方が3以上カテゴリで順序がないときには、もう片方の軸が量的形質になっても、同じくクラスカル・ウォリスの検定が実施可能です。

### (4) 両軸に順序がある場合

片方の軸が順序ありのカテゴリになり、もう片方の軸も順序がある場合には、Jonckheere-Terpstra検定が使えます。この検定も、値の大小関係を用いた処理をしているので、3以上の順序ありカテゴリも量的データの場合も同様に使えます。両方の軸が連続な場合も順序ありという意味では同じです。したがって、この場合もJonckheere-Terpstra検定が使えます。また、線形回帰も使えます。

表13.7 適用できる検定

	2	3順序なし	3順序あり	連続
2	ピアソン $\chi^2$	ピアソン $\chi^2$	順位和検定・傾向性の検定	KW
3順序なし	—	ピアソン $\chi^2$	KW	KW
3順序あり	—	—	JT	JT
連続	—	—	—	JT/（線形）回帰

※ 2:2カテゴリ、3順序なし:3以上順序なしカテゴリ、3順序あり:3以上順序ありカテゴリ、連続:連続（順序あり）、KW:クラスカル・ウォリス検定、JT:Jonckheere-Terpstra検定

### ● 13.6.3 複数の手法の挙動の比較

あるデータ型のときにある手法が「使えます」と述べてきましたが、それは、データの処理がそのデータ型を受け付けるという意味です。複数の方法が使えるデータ型のときにはそれぞれの結果が異なりますから、その違いの由来が何なのかを意識する必要があります。

実際に複数の手法の結果を比較してみましょう。「2 カテゴリ × 2 カテゴリ」の場合は、順序ありとも順序なしともみなせる軸の組み合わせなので、上述したすべての検定を実行することが可能で、それらはほぼ同じ結果をもたらします。他との一致が多少なりとも悪いのは Jonckheere-Terpstra 検定ですが、これは軸が高次になったときなどにうまくいくように工夫された計算式を使っている分、単純な場合にしわ寄せがきていると思えばよいでしょう（図 13.8 (a)）。

次に  $2 \times 3$  表にします。その結果が図 13.8 (b) です。ピアソンの独立性検定は自由度 2 ですから、その他の検定（自由度 1）とは異質です。それ以外の検定は、クラスカル - ウオリス検定も Yonckheere-Terpstra 検定も傾向性検定も線形回帰もだいたい同じです。特に、傾向性検定と線形回帰はどちらも線形を仮定した検定ですから、計算誤差を除いて完全に一致します。

次に、片方の軸を 2 カテゴリにしたまま、もう片方のカテゴリ数を 10 まで増やすと、ピアソンとそれ以外の検定はほとんど独立になります。しかしながら、それ以外はおまかに同じです。傾向性の検定と線形回帰の一一致は相変わらずよいです（図 13.8 (c)）

今度は  $3 \times 3$  表にします。両方に順序のある Yonckheere-Tersptra 検定と線形回帰とは似ています。それ以外は挙動が違います（図 13.8 (d)）。両軸に傾向がない、片方に傾向がある、両軸に傾向がある、の 3 グループに分かれますので、それらは「ピアソン > クラスカル - ウオリス検定 > 線形 2 法 (Yonckheere-Terpstra 検定、線形回帰)」という順番で関係がある様子も見てとれます。

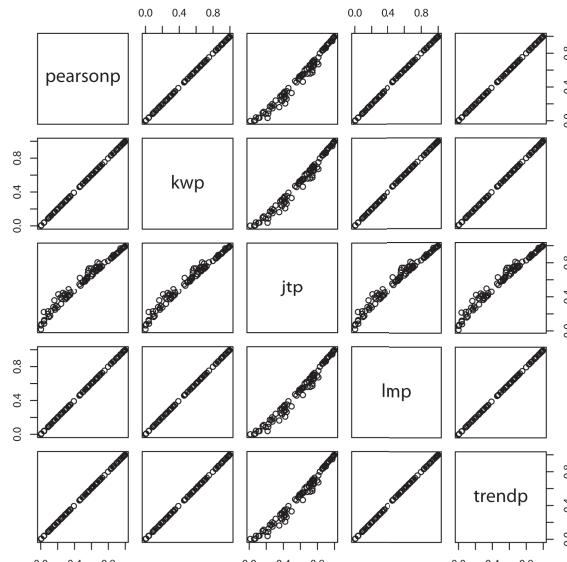
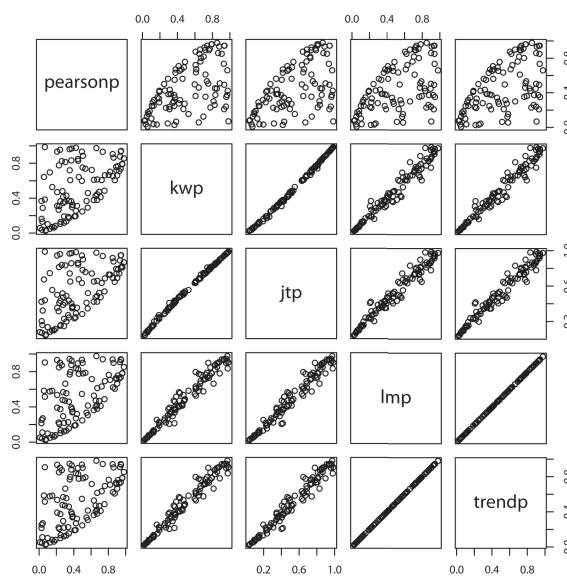
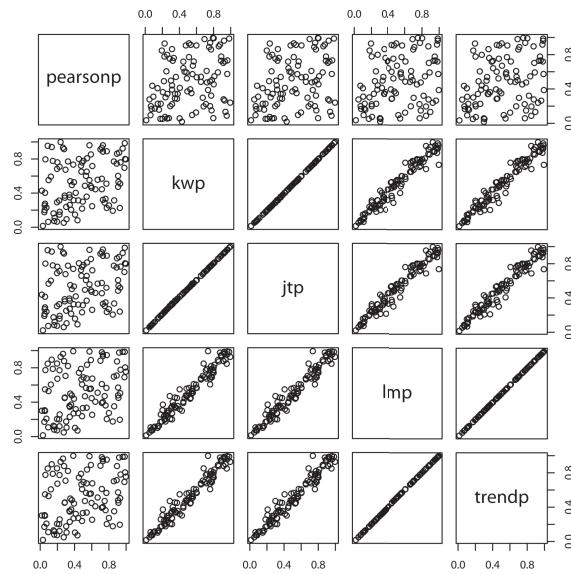
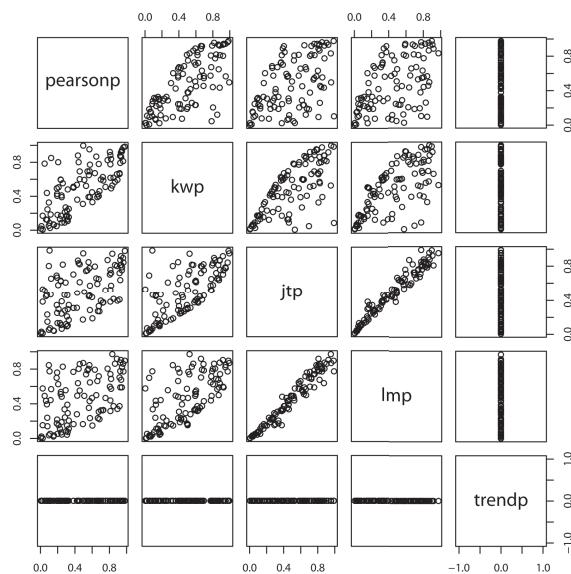
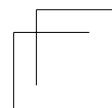
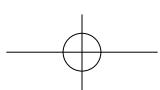
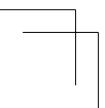
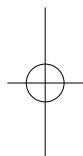
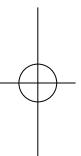
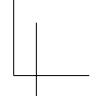
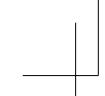
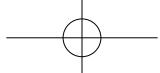
(a)  $2 \times 2$  表(b)  $2 \times 3$  表

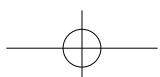
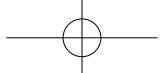
図 13.8 複数の分割表検定の結果を比較

(c)  $2 \times 10$  表(d)  $3 \times 3$  表

上から、次の諸検定の  $p$  値、ピアソンの検定 (pearsonp)、クラスカル - ウオリス検定 (kwp)、Jonckheere-Terpstra 検定 (jtp)、線形回帰 (Imp)、傾向性検定 (trendp)

図 13.8 複数の分割表検定の結果を比較 (つづき)





# 第 14 章

## 関係と因果

### 14.1 原因と結果と時間

何かを調べるとき、何かの原因が知りたい、ということが多いです。「A ならば B である」です。しかしながら、データから得られる情報は、「A と B とは関係がある」であることが多いです。データは「関連」について教えてくれるが、その「**因果関係**」について教えてくれるわけではないです。「因果関係」を知ることは、「関係」を知ることとは異なることなので、そのための工夫が必要です。

**疫学**という学問分野は、まさに因果関係のあり・なしについて研究しているので、因果関係の妥当性に関する体系的理解はそちらを参照していただくこととします。しかしその中で、遺伝子の解析に関係の深い点である、「時間」の要素を取り上げます。

「A と B とに関係がある」ときに

- A が B の原因である場合
- B が A の原因である場合
- A と B とがともに、C を原因とする結果である場合
- A と B とがともに、C の原因である場合

の 4 パターンが考えられるでしょう。図で表すと図 14.1 のようになります。

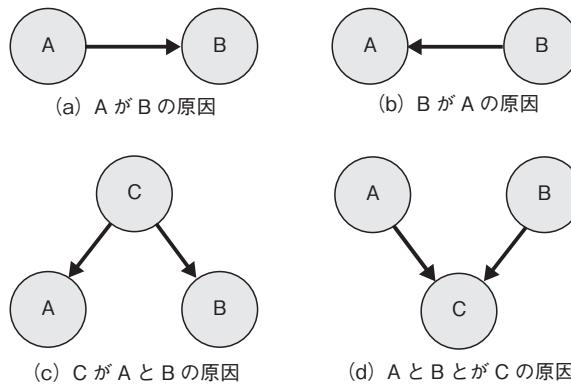


図14.1 AとBとに関連があるときの因果関係のいろいろ

今、「A」が「B」より「時間的」に先に存在しているときを考えます。その時間的な前後関係がわかれば、4つの場合のうち「BがAの原因である場合」の可能性がなくなります。残りの3つの場合は相変わらず残ります。

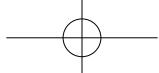
研究のスタイルを後ろ向き研究と前向き研究とに2分するのも、「因果関係」の観点から言うと、「時間」の要素をどのように取り扱うという点で大きく異なるからです。「結果と注目している」因子が研究開始時点では確かに存在しなかったのに、研究を進めていくうちに確かに存在するようになったことを、前向き研究は「確認」することができます。

多くの実験研究は前向き研究です。「原因と注目している」因子のあり・なしを操作して、「結果と注目している」因子がどのよう影響を受けるかを観察するように実験を組み立てるのが普通だからです。

## 14.2 原因としてのジェノタイプ

遺伝子の場合はどうでしょうか？ ジェノタイプとフェノタイプの区別をするときに、「染色体のうち、変化しない部分（塩基配列）」をジェノタイプとし、それ以外の化学修飾などはRNA、タンパク質、形質一般とともに、フェノタイプとすることを第4章で述べました（図4.1）。

このようにして定義したジェノタイプは、細胞の誕生時点・個体の誕生時点すでに存在しており、細胞・個体に起きる各種の現象は、ジェノタイプの確



定より後に起きますから、ジェノタイプとその他すべてのフェノタイプについての解析に限ると、時間的関係を限定した解析として扱えることになります。この点が、統計遺伝学の特化した点の1つと言えるでしょう。



## 14.3 有向グラフ、ベイジアンネットワーク

因果関係には時間的順序が大事な要素であることを説明したときに、図14.1にある通り、向きのある辺を持つグラフ（有向グラフ）を用いました。時間という軸が持つ、1方向性の情報をグラフでは矢印の向きとして取り扱えます。

また、第10章でマルコフ連鎖のこと触れました。マルコフ連鎖では、ある時点の様子が、その直前（もしくは数段階前まで）の状態によって確率的に決まっていました。マルコフ連鎖の状態と状態の間には、データが支持する「関係」があります。その「関係」がどちらを向いているかは、両者の「関係」のみに着目していれば決まりません。なんらかの理由で向きを与えると、その関係はある状態から次の状態への推移確率になります。前の状態が確率的に表され、それに基づいて次の状態が確率的に決められていました。

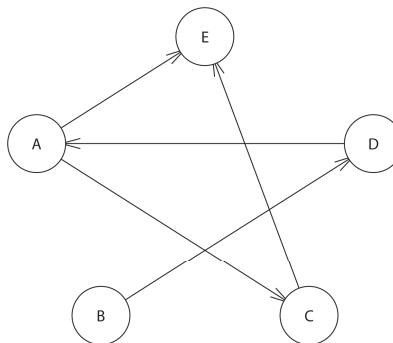
今、有向グラフに着目して、ある点の様子がその点に向かう辺の出所の点の状態によって確率的に決まると考えれば、直線上ではないけれども、似たような取り扱いが可能です。有向グラフであって、どの点から出発しても自身の点に戻ってこられないようなグラフを**非循環有向グラフ**と言います。

要因とその因果関係をこの非循環有向グラフを用いて表し、観察データを最もよく説明するグラフを推定する手法に**ベイジアンネットワーク**があります。

以下では、その概略をRのbnlearnパッケージを用いて試してみます（R14-1.R）。まず、非循環有向グラフを描いてみます。

図14.2のグラフを見てください。これは点が5個のグラフで、辺に向きがありますから、有向グラフです。ACEに三角形ができていますが、矢印の向きに注意するとぐるぐると回ることができないので、これは「非循環」です。

次に、適当に依存関係のある5要素のデータを作ってみます。5個の頂点のある有向グラフは、 $5 \times 4/2 = 10$ 本の辺を引くことが可能で、10本の辺のそれぞれに線を引くか引かないか、引くとして向きをどうするかの3つの場合がありますから、 $3^{10} = 59,049$ 通りの辺の引き方があります。頂点の数を1つ



相互に結ばれて三角形を作る3点がありますが、辺の向きに注意すると非循環であることがわかります。

図14.2 頂点数5のグラフ

増やして6になると、この数は10,000,000を越えます。非循環という制約をつけるのでこれより減りますが、非常に多いことは明らかです<sup>※1</sup>。したがって、たくさんのグラフパターンの中から、特定のグラフパターンを選び出す手順が必要です。

bnlearnパッケージにも複数のアルゴリズムが実装されていますが、その1つでグラフパターンを選択してみます。

この選ばれたグラフパターンが図14.3です。このパターンが他のパターンよりもデータに即して優れていることを確認したいのです。しかしながら、すべてのグラフパターンと比較することはできませんから、いくつかを抜き出

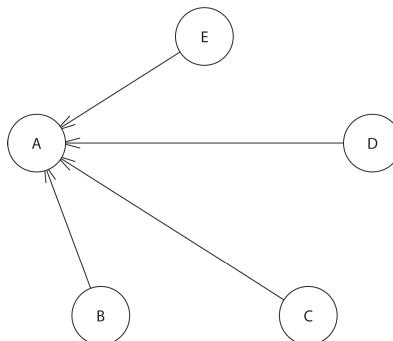


図14.3 推定されたネットワークグラフ

※1 非循環有効グラフのパターン数は15.3節で扱います。

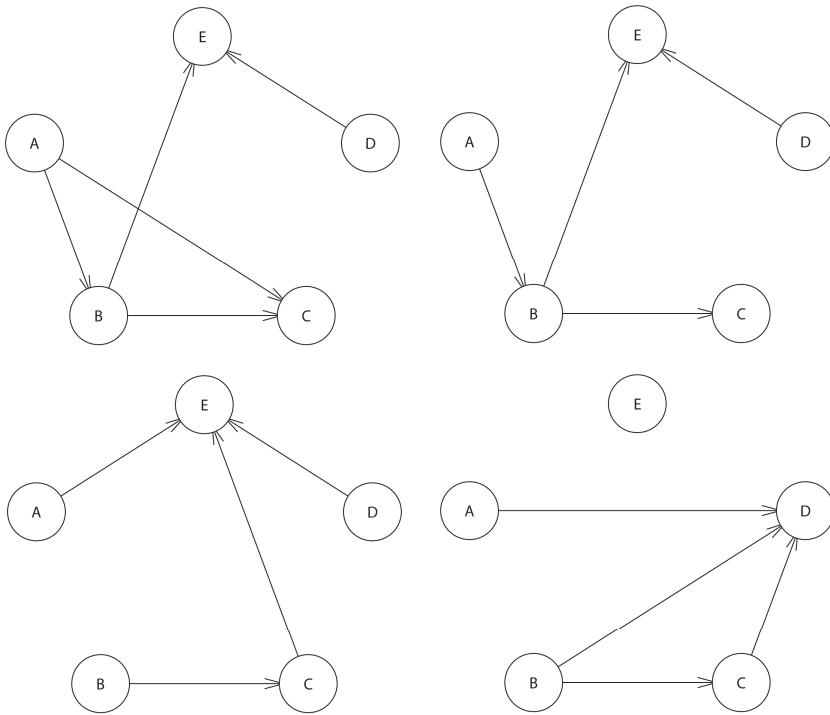


図 14.4 ランダムに作成した 4 個のネットワークグラフ

して比較してみましょう。まず、ランダムにグラフパターンを作ってみて、それらを図示してみます（図 14.4）。

いろいろなパターンが作れることがわかりましたので、100 個ほど作成して、それをあるアルゴリズム<sup>※2</sup>が選択したグラフと比較してみます。比較するためには、なんらかの優劣の判定方法が必要ですが、そのための関数（score()）もあります。推定グラフのスコアを水平線で、ランダムに作ったグラフのスコアを点で示したのが図 14.5 です。100 個のランダムなグラフパターンの得点が、推定パターンの得点（水平線）より小さいことが見てとれます。

※2 Grow-Shrink アルゴリズムを使いました。

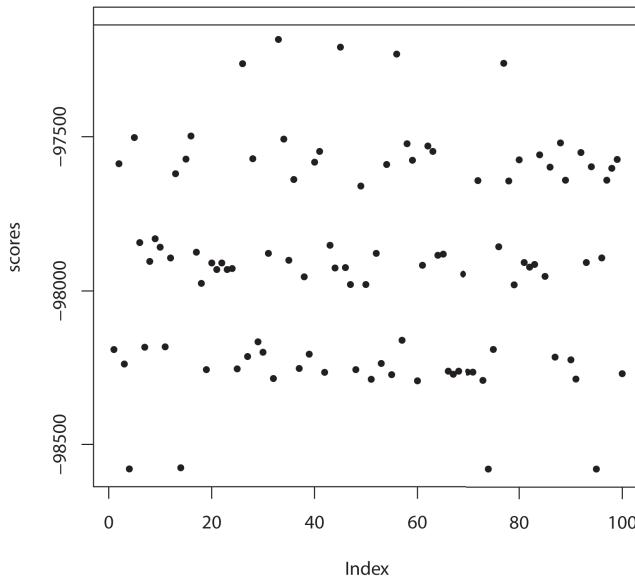


図 14.5 推定ネットワークグラフの評価点（上部の水平線）は、ランダムに作ったグラフの評価点（ドット）よりも高い

このようにして因果関係を表す非循環有向グラフを選んだ後に、有向辺に要素間の関係の値を与えることで、選んだグラフにさらにモデルとしての意味づけをすることができます。

処理内容については、R14-1.R に示すソースで確認してください。

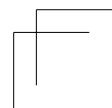
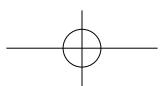
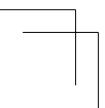
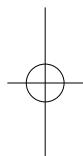
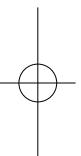
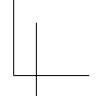
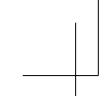
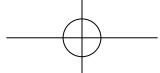
#### R ソース 14.1 R14-1.R : ベイジアンネットワーク

```

1 library(bnlearn)
2 e <- empty.graph(LETTERS[1:5]) # 頂点数 5 の辺のないグラフ
3 arc.set <- matrix(c("A", "C", "B", "D", "C", "E", "D", "A",
                      "A", "E"), ncol = 2, byrow = TRUE)
4 arcs(e) <- arc.set # 辺を与える
5 plot(e)
6 # 適当に関連のある 5 項目のデータを作る
7 set.seed(123456)
8 Ns<-10000
9 A<-rpois(Ns,3);B<-rpois(Ns,3);C<-rpois(Ns,3);D<-rpois(Ns,3);
  E<-rpois(Ns,3)
10
11 B[1:(Ns/4)] <- A[1:(Ns/4)];

```

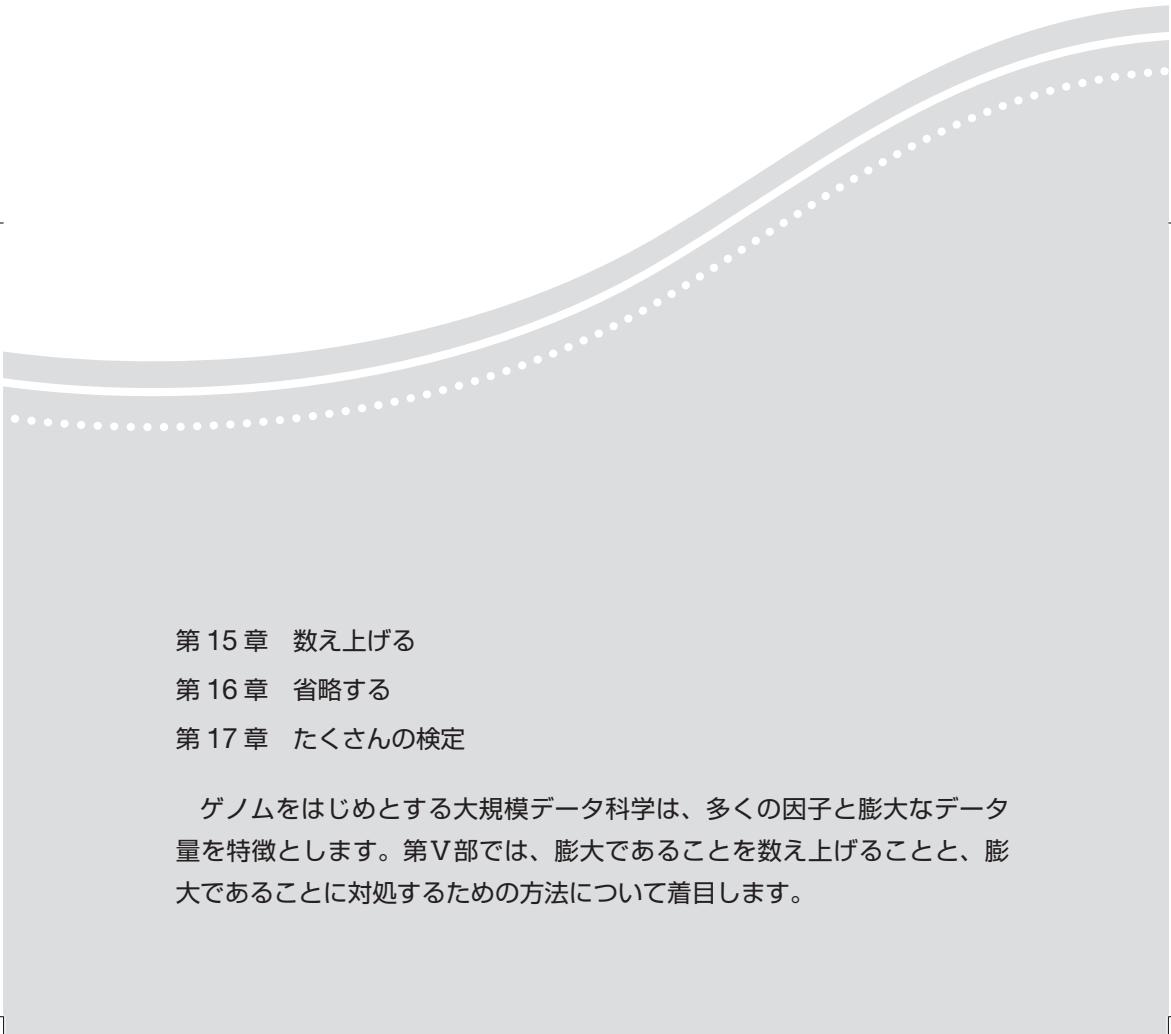
```
C[(1+Ns/4):(Ns/2)]<-A[(1+Ns/4):(Ns/2)]
12 D[(1+Ns/2):(3*Ns/4)]<-A[(1+Ns/2):(3*Ns/4)];
   E[(1+3*Ns/4):Ns]<-A[(1+3*Ns/4):Ns]
13 d<-as.data.frame(cbind(A,B,C,D,E))
14 res<-gs(d)
      # 手法の1つであるGrow-Shrink法を用いてネットワークを推定する
15 plot(res)
16 # 頂点数5のグラフをランダムに作ってプロットしてみる
17 plot(random.graph(LETTERS[1:5], num = 1))
18 # 推定されたグラフとランダムに作成したグラフとで、
19 # データの説明の良さを評価して比較してみる
20 GSscore<-score(res, d) # 推定ネットワークグラフの評価
21 Niter<-100 # ランダムに作るグラフの数
22 scores<-rep(0,Niter) # ランダムに作ったネットワークの評価
23 for(i in 1:Niter){
24   randomGraph<-random.graph(LETTERS[1:5], num = 1)
25   scores[i]<-score(randomGraph,d)
26 }
27 ylim<-c(min(scores,GSscore),max(scores,GSscore))
28 plot(scores,ylim=ylim,pch=20)
29 abline(h=GSscore)
30 bn.fit(res,d) # 推定結果の表示
```





# 第 V 部

## 大規模なこと



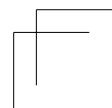
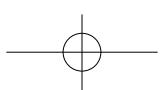
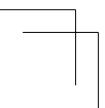
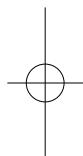
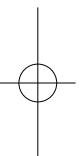
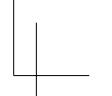
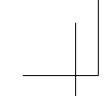
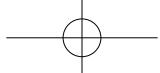
第 15 章 数え上げる

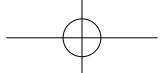
第 16 章 省略する

第 17 章 たくさんの検定

ゲノムをはじめとする大規模データ科学は、多くの因子と膨大なデータ量を特徴とします。第V部では、膨大であることを数え上げることと、膨大であることに対処するための方法について着目します。







# 第 15 章

## 数え上げる

### 15.1 順列、重複順列、分割表の正確生起確率

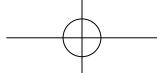
ゲノムをはじめとする「オミックス」研究<sup>※1</sup>は、そもそも多くの要素を全体として捉えることに主眼をおいた研究なので、取り扱い要素が非常に多いです。ヒトゲノムは30億塩基対、ヒトのコーディング遺伝子数は2～3万です。さらに、要素間の関係を調べたり、組み合わせについて調べることも多いです。また、木グラフやネットワークを使いますから、俄然、場合の数が大きくなります。まずは、場合の数に関してまとめましょう。

#### (1) 順列と組み合わせ

$N$ 個の要素があるときに、 $k$ 個を抜き出して並べる並べ方を**順列**と言います。順列の場合の数  $P(N, k)$  は次の式で表されます。ガンマ関数も使います(図 15.1 (a))。

$$\begin{aligned} P(N, k) &= N(N-1)(N-2)\dots(N-(k-1)) \\ &= \frac{N!}{(N-k)!} = \frac{\Gamma(N+1)}{\Gamma(N-k+1)} \end{aligned}$$

※1 ゲノム、ranscripTom、Proteomeなど、「オーム」という接尾語を持つ研究。それぞれ、遺伝子、遺伝子発現、タンパク質のすべてを対象に総合的に解析するための学問のこと。非常に多くの要素(全遺伝子、全タンパク質)に関する実験を行うことを特徴とします。



$k$  個のうち 1 個目は  $N$  個の要素から選ぶことができて、2 個目は残っている  $N-1$  個から、3 個目は  $N-2$  個から、……というような選び方なので、上記のような式となります。

#### R ソース 15.1 R15-1.R : 順列の数

```
1 permN<-function(N=10,k=3){exp(lgamma(N+1)-lgamma((N-k)+1))}  
2 permN(N=10,k=3)
```

(a) 順列  $P(3,3)=3!=6$

{1 2 3} {1 3 2} {2 1 3} {2 3 1} {3 1 2} {3 2 1} 1 と 2 と 3

(b) 重複順列  $\Pi(3,3)=3^3=27$

{1 2 3} {1 3 2} {2 1 3} {2 3 1} {3 1 2} {3 2 1} 1 と 2 と 3

{1 1 2} {1 2 1} {2 1 1}  
{1 2 2} {2 1 2} {2 2 1} 1 と 2

{1 1 3} {1 3 1} {3 3 1}  
{1 3 3} {3 1 3} {3 3 1} 1 と 3

{2 2 3} {2 3 2} {3 3 2}  
{2 3 3} {3 2 3} {3 3 2} 2 と 3

{1 1 1} {2 2 2} {3 3 3} 単独

図 15.1 順列と重複順列

$N$  個の異なる要素から、重複を許して  $k$  個を並べる並べ方を**重複順列**と呼びます。こちらは単純で、 $k$  回とも  $N$  個のどれかを当てはめられるので、

$$\Pi(N, k)=N^k$$

です（図 15.1 (b)）。

#### R ソース 15.2 R15-2.R : 重複順列の数

```
1 repPermN<-function(N=10,k=3){ N^k }  
2 repPermN(N=10,k=3)
```

選び出す要素ごとに個数を指定して、全部で  $k$  個にして並べ方は、

$$\frac{k!}{\prod_{i=1}^N n_i!}; \sum_{i=1}^N n_i = k$$

です。これは、 $k$  個の要素に  $N$  種類のラベルをラベルごとの枚数を指定してラベル付けする場合の数です。 $N \times M$  の分割表で考えます。

今、総数  $n_{..}$  に、2通りのラベルを貼ります。 $N$  種類である 1 番目のラベルが  $n_{i..}$ ;  $i = 1, 2, \dots, N$  枚ずつあり、その貼り方の場合の数は、

$$\frac{n_{..}!}{\prod_{i=1}^N n_{i..}!}$$

で、 $M$  種類である 2 番目のラベルが  $n_{..j}$ ;  $j = 1, 2, \dots, M$  枚ずつあり、その貼り方の場合の数は、

$$\frac{n_{..}!}{\prod_{j=1}^M n_{..j}!}$$

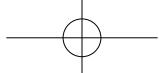
です。2つのラベルは、それぞれお互いに関係なく貼ることができるとき、その貼り方は 2つの貼り方の積なので、

$$X = \frac{n_{..}!}{\prod_{i=1}^N n_{i..}!} \times \frac{n_{..}!}{\prod_{j=1}^M n_{..j}!}$$

です。一方、総数  $n_{..}$  に  $N \times M$  種類のラベルを 1つだけ貼る場合を考えます。各ラベルの枚数が  $n_{ij}$ ;  $i = 1, 2, \dots, N$ ;  $j = 1, 2, \dots, M$  とすると、

$$Y = \frac{n_{..}!}{\prod_{i=1}^N \prod_{j=1}^M n_{ij}!}$$

です。



$$\begin{aligned} \frac{Y}{X} &= \frac{\frac{n_{..}!}{\prod_{i=1}^N \prod_{j=1}^M n_{ij}!} \cdot \prod_{i=1}^N n_{i..}! \cdot \prod_{j=1}^M n_{.j}!}{n_{..}!} \\ &= \frac{\prod_{i=1}^N n_{i..}! \prod_{j=1}^M n_{.j}!}{n_{..}! \prod_{i=1}^N \prod_{j=1}^M n_{ij}!} \end{aligned}$$

が、2軸が独立と仮定したときに表を観察する**正確生起確率**になります。この計算は13.2節での計算式と同じです。

正確生起確率というのは、すべてのラベル付けの場合の数（順列）のすべてを数え上げているのですが、本当に数え上げると面倒なので、同じラベルの枚数に関する考慮をしていることがわかります。

### ● 15.1.2 組み合わせ、重複組み合わせ、2倍体ジェノタイプの種類数

順列と似たものに**組み合わせ**というのがあります。

$N$ 個から  $k$  個を選ぶ組み合わせは、

$$C(N, k) = \frac{N!}{k!(N-k)!} = \frac{P(N, k)}{k!}$$

です。 $k$  個の順列を作った上で、 $k$  個の要素の並び方はどうでもよいので、その場合の数  $k!$  で割っています。これは、 $N$  個に 2 種類のラベルを貼る貼り方で、 $k$  枚と  $N-k$  枚に貼る貼り方のことです。

R ならば R15-3.R のようになります。

R ソース 15.3 R15-3.R : 組み合わせ

```
1 choose(n=20, k=3)
```

$N$  種類から重複を許して  $k$  個を選んで作る組み合わせ（**重複組み合わせ**）は次のようになります。

$$\begin{aligned}
 H(N, k) &= C(N + k - 1, k) \\
 &= \sum_{i=0}^k (C(N, k-i) \times H(k-i, i)) \\
 &= \sum_{i=0}^k (C(N, k-i) \times C(k-1, i))
 \end{aligned}$$

これは、 $N$ 種類を1列に並べて、それを分ける $N-1$ 本の仕切り線を考えます。 $k$ 個を選ぶということを、 $N-1$ 本の仕切り線で区分けされた場所に $k$ 個のものを置くという作業と考えます。すると、 $N-1$ 本の仕切り線と $k$ 個のものとを1列に並べる方法（順列）が、数え上げる場合の数になります。ただし、 $N-1$ 本の線と $k$ 個のものは、互いに区別がつかないことを考慮して、

$$H(N, k) = \frac{P(N-1+k)}{P(N-1)P(k)} = C(N+k-1, k)$$

とします。

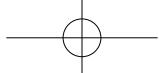
たとえば、 $N$ 種類のアレルを持つ多型が作る2倍体のジェノタイプは、 $H(N, 2)$ です。

#### R ソース 15.4 R15-4.R : 重複組み合わせ (2通りの計算方法)

```

1 repCombN<-function(N=10, k=3) {
2   choose(N+k-1, k)
3 }
4 repchoose<-function(n=10, k=3) {
5   ret<-0
6   for(i in 0:k) {
7     ret<-ret+choose(n, k-i)*repCombN(k-i, i)
8   }
9   ret
10 }
11 repCombN(N=10, k=3)
12 repchoose(n=10, k=3)

```



## 15.2 分割の数——スターリング数とベル数

$N$  個の要素を **分割**することを考えます。まず  $k$  群に分割するとします。どの要素も必ずどれかの群に属するものとして、複数の群に属する要素はないものとします。また、 $k$  群のすべてが、必ず 1 個以上の要素を持つように分割します。これを**第2スターリング数**  $St2(N, k)$  と言います。 $k$  は  $1, 2, \dots, N$  のいずれかです。 $k=1, N$  のときには  $St2(N, k)=1$  です。

今、 $St2(N, k)$  がわかっているとします。すると、 $St2(N+1, k)$  は、次のような式で表されます。

$$St2(N+1, k) = St2(N, k) \times k + St2(N, k-1)$$

これは、要素数が  $N$  から  $N+1$  に増えたとき、 $k$  群に分ける分け方は、

- $N$  個の要素がすでに  $k$  群に分かれているときに、追加された 1 要素を  $k$  群のどれか 1 つに入れる場合が  $St2(N, k) \times k$  通り
- 追加する 1 要素が単独で 1 つの群に属するようにするとすると、残りの  $N$  個の要素は  $k-1$  群に分けるので、その場合は  $St2(N, k-1)$  通り

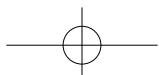
の 2 パターンであることを利用しています。2 分割の場合の数は、 $N$  個から  $k=1, 2, \dots, N-1$  個を取り出す場合の数と関係しています。 $k$  個取り出す場合をすべて数え上げると、 $N-k$  個を取り出す場合も数えてしまうので、2 倍になることを考慮して、

$$\begin{aligned} St2(N, 2) &= \frac{1}{2} \sum_{i=1}^{N-1} C(N, i) \\ &= \frac{1}{2} \left( \sum_{i=0}^N C(N, i) - (C(N, 0) + C(N, N)) \right) \\ &= \frac{1}{2} ((1+1)^N - 2) = 2^{N-1} - 1 \end{aligned}$$

とします<sup>※2</sup>。

---

※2  $N$  個の要素を 2 群に分けることが、 $N$  個の要素の集合の部分集合を選ぶことに関係していることとも関係しています。なぜなら、要素数  $N$  の集合の部分集合の数は、集合そのものと空集合とを合わせたときに  $2^N$  だからです。



要素数を指定した重複順列  $5!/(3!2!)$

2種類の3+2の重複順列	5個から3個を選ぶ	5個から2個を選ぶ	5+1個を2+1人に分ける	5-1種類からを2個を重複を許して選ぶ
{A A A B B}	{1 2 3}	{4 5}	{OOOO P Q}	{OO o q r}
{A A B A B}	{1 2 4}	{3 5}	{OOO PP Q}	{O P o r}
{A A B B A}	{1 2 5}	{3 4}	{OOO P QQ}	{O o Q r}
{A B A A B}	{1 3 4}	{2 5}	{OO PPP Q}	{O o q R}
{A B A B A}	{1 3 5}	{2 4}	{OO PP QQ}	{o PP q r}
{A B B A A}	{1 4 5}	{2 3}	{OO P QQQ}	{o P Q r}
{B A A A B}	{2 3 4}	{1 5}	{O PPP Q}	{o P q R}
{B A A B A}	{2 3 5}	{1 4}	{O PPP QQ}	{o p QQ r}
{B A B A A}	{2 4 5}	{1 3}	{O PP QQQ}	{o p Q R}
{B B A A A}	{3 4 5}	{1 2}	{O P QQQQ}	{o p q RR}

図 15.2 左から重複順列、組み合わせ、組み合わせ、分割、重複組み合わせ（同じ計算式になる）

#### R ソース 15.5 R15-5.R : 第2スターリング数

```

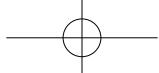
1 Stirling2N<-function (N=10, k=3) {
2   ret<-0
3   if(k<=N && k>=1) {
4     if(k==1) {
5       ret<-1
6     }else{
7       ret<-Stirling2N(N-1,k)*k+Stirling2N(N-1,k-1)
8     }
9   }
10  ret
11 }
12 Stirling2N(N=10, k=3)

```

重複順列と組み合わせ、分割の場合の数には図 15.2 のような相互関係があります。

$N$  個の要素を、群の数は問わずに分ける分け方は**ベル数** ( $B(N)$ ) と呼ばれます。これは、 $N$  個を  $k = 1, 2, \dots, N$  群に分ける場合の数 ( $St2(N, k)$ ) を足し合わせたものなので、

$$B(N) = \sum_{i=1}^N St2(N, i)$$



です。次のようにも書けます。

$$B(N+1) = \sum_{i=0}^N C(N,i)B(i)$$

これは、次のように考えます。 $N+1$  番目の要素が属する群が、 $i+1$  個の要素を持つ群だとします。そうすると、その群に属する要素を除いた、 $N-i$  個の要素を分割するパターンが  $B(N-i)$  通り存在します。 $N$  個の要素から、 $N+1$  番目の要素と一緒にする  $i$  個の要素の選び方は  $C(N, i)$  なので、このような関係が生じます。ただし、0 個の要素の分割という定義しにくい値がありますから、 $B(0)=1$  と定めて、漸化式がうまくいくようにしてあります。

#### R ソース 15.6 R15-6.R : ベル数

```
1 bellN<-function(N=10) {  
2   bs<-rep(0,N+1)  
3   bs[1]<-1  
4   if(N>=1) {  
5     for(i in 2:(N+1)) {  
6       for(j in 0:(i-2)) {  
7         bs[i]<-bs[i]+choose(i-2,j)*bs[j+1]  
8       }  
9     }  
10   }  
11   bs[2:length(bs)]  
12 }  
13 bellN(N=10)
```

## 15.3 分割とカテゴリの統合

### ● 15.3.1 順序のないカテゴリの場合

第2スターリング数もベル数も集合を分割する場合の数でした。順序のない  $N$  種類のカテゴリを持つ尺度があったときには、 $N$  カテゴリをそれぞれ独立に扱うこともあります、いくつかのカテゴリを分割する（共通項でまとめる）ことを考慮することがあるかもしれません。そのようなときには、ここで示した場合の数のすべてを対象とするか、一部の組み合わせだけをなんらかの根拠とともにプールすることになります。ここで見てきた集合の分割は、データ解析にあたっては、順序のないカテゴリをグループ分けすることに関連することだったわけです。

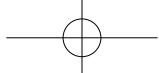
### ● 15.3.2 順序のあるカテゴリの場合

では、順序のあるカテゴリ型尺度のときにはどうなるでしょうか（図 15.3）。たとえば、重症度が 5 段階で評価されている場合に、重症度と 2 カテゴリ型の要素（治療反応性など）との関係を調べたいとします。5 段階に (1, 2, 3, 4, 5) と線形の重みを付ければ、それに応じた傾向性の検定になります。 $((1), (2, 3, 4, 5))$  と分けたり、 $((1), (2, 3), (4, 5))$  と分けたりすることが適当なことがあるかもしれません。

では、どれくらいの分け方があるかを考えます。 $N$  個の要素を 1 列に並べると、 $N - 1$ 箇所の仕切り線があります。この  $N - 1$  本の仕切り線のうち 1 本だけを有効にすれば 2 分割、2 本を有効にすれば 3 分割、 $k = 1, 2, \dots, N - 1$  本を有効にすれば  $k + 1$  分割なので、結局、順序のあるカテゴリの分割の場合の数は、

$$B_{\text{order}}(N) = \sum_{i=1}^{N-1} C(N-1, i) = 2^{N-1} - 1 = St2(N, 2)$$

R では R15-7.R のようになります。



## R ソース 15.7 R15-7.R : 順序のあるカテゴリの分割

```

1 Border<-function(N) {
2   2^(N-1)-1
3 }
4 Border(10)

```

## 3要素の分割

$\{1\ 2\ 3\}$	$St2(3,1)=1$
$\{\{1\ 2\}\{3\}\}$	$St2(3,2)=3$
$\{\{1\}\{2\ 3\}\}$	$B(3)=1+3+1=5$
$\{\{1\ 3\}\{2\}\}$	$St2(3,3)=1$
$\{\{1\}\{2\}\{3\}\}$	

## 秩序のある3要素の分割

$\{1\ 2\ 3\}$	$C(2,0)=1$
$\{\{1\ 2\}\{3\}\}$	$C(2,1)=2$
$\{\{1\}\{2\ 3\}\}$	$2^2=1+3+1=4$
$\{\{1\ 3\}\{2\}\}$	$=St2(3,2)+1$
$\{\{1\}\{2\}\{3\}\}$	$C(2,2)=1$

$\{\{1\ 3\}\{2\}\}$  は 123 の順序を守っていない

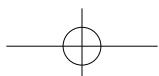
図 15.3 順序のないカテゴリの分割と順序のあるカテゴリの分割

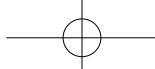


## 15.4 木の形の数、グラフの数——木、クラスタリング、ベイジアンネットワーク

### ◎ 15.4.1 木のパターンの数

場合分けに**カタラン数**というものがあります。根のある木で表す現象が遺伝学には多いですが、根のある木で  $N$  本の辺を持つグラフの数がカタラン数です。また、 $N$  個の頂点を内部頂点とするグラフの形の数もカタラン数です。根のある 2 分岐木の場合、 $N$  個の内部頂点を持つものは  $N+1$  個の葉を持つ ( $N+1$  チームのトーナメント大会は  $N$  試合) ので、 $N+1$  個の要素のまとめ上げ 2 分岐木の形の数が、カタラン数であるとも言えます。





## R ソース 15.8 R15-8.R : カタラン数

```
1 CatalanN <- function(N = 10) {  
2   exp(lgamma(2 * N + 1) - lgamma(N + 2) - lgamma(N + 1))  
3 }  
4 CatalanN(N=10)
```

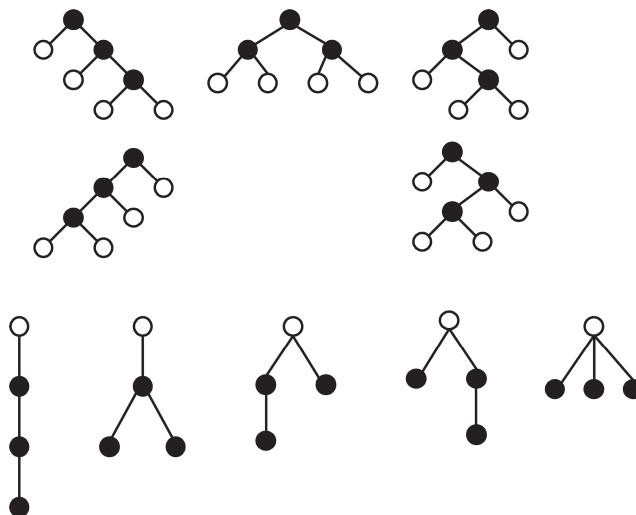


図 15.4 上段の 5 つの木：内部頂点が 3 の木、下段：根のある木で辺の数が 3 の木

## ◎ 15.4.2 クラスタリングのパターンの数

分岐木と言えば、クラスタリングでも分岐木としてまとめ上げる作業をしました。 $N$  個の要素を 2 分岐木にまとめ上げるとき、ある分岐点の先を入れ替えるてもクラスタリングのパターンとしては変わらないので、そのことを考慮することにします。 $N$  個の要素を 2 分岐木パターンでクラスタリングする方法の数はダブルファクトリアルナンバーと呼ばれる数列

$$(2Z-1)!! = 1 \times 3 \times 5 \times \dots \times (2Z-1)$$

を使って、

$$2((N-2)-1)!!$$

で計算されます。

Rには2分岐木を数え上げて描くツールがあって、少数であれば、それを使っても数え上げることができ、図を描くこともできます(R15-9.R、図15.5)※3。

#### Rソース 15.9 R15-9.R : 2分岐木のパターン数

```

1 library(phangorn)
2 allTrees(5) # 頂点数5でクラスタリングパターンを列挙する
3 trees<-allTrees(5)
4 plot(trees)
5 doubleFactorialN<-function(N=5) {
6   if(N<3) {
7     print("no answer")
8   }else{
9     ret<-1
10    for(i in 3:N){
11      ret<-ret*(2*(i-2)-1)
12    }
13    ret
14  }
15 }
16 doubleFactorialN(N=5)

```

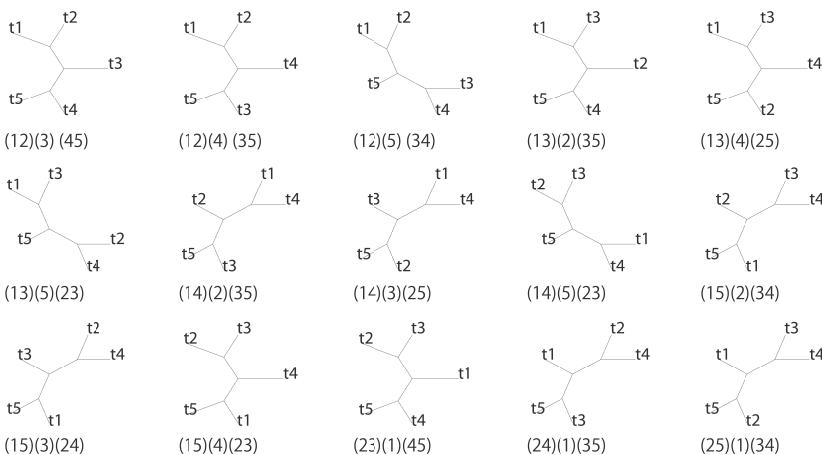
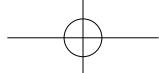


図15.5 頂点数5の2分岐木パターンのすべて(15通り)

※3 R15-9.Rを実行すると、グラフの表示がなされるたびに、次の処理に移ってよいかどうかを尋ねられます。マウスをクリックするなどで、次の処理に進めます。



### ● 15.4.3 無向グラフの数、有向グラフの数、非循環有向グラフの数

頂点数  $N$  個の無向グラフを作るときには、頂点数  $N$  に対して、頂点のペア  $\frac{N(N-1)}{2}$  に、辺を引くか引かないかの選択ができるので、

$$\frac{N(N-1)}{2^2}$$

です。

有向グラフの場合は、辺を引かないか、引くとしたら 2 方向のどちらに引くか、の 3 つの選択肢があるので、

$$\frac{N(N-1)}{3^2}$$

です。

ベイジアンネットワークで出てきた非循環有向グラフの種類数 ( $N_{\text{acg}}$ ) は、

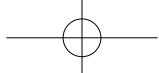
$$N_{\text{acg}}(0) = 1$$

$$N_{\text{acg}}(N) = \sum_{i=1}^N (-1)^{k+1} C(N, k) \times 2^{k(N-k)} \times N_{\text{acg}}(N - k)$$

で与えられます。

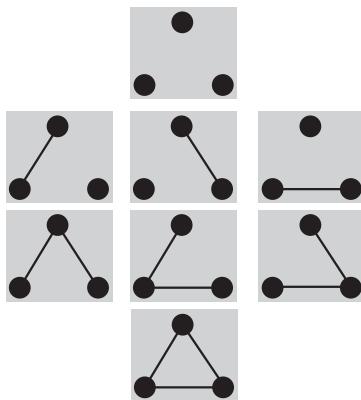
この数列の場合は、計算はソースを示す代わりに、数列の検索サイトの紹介をすることにします。

頂点数 3 の場合の無向グラフと非循環有向グラフを図 15.6 に示します。

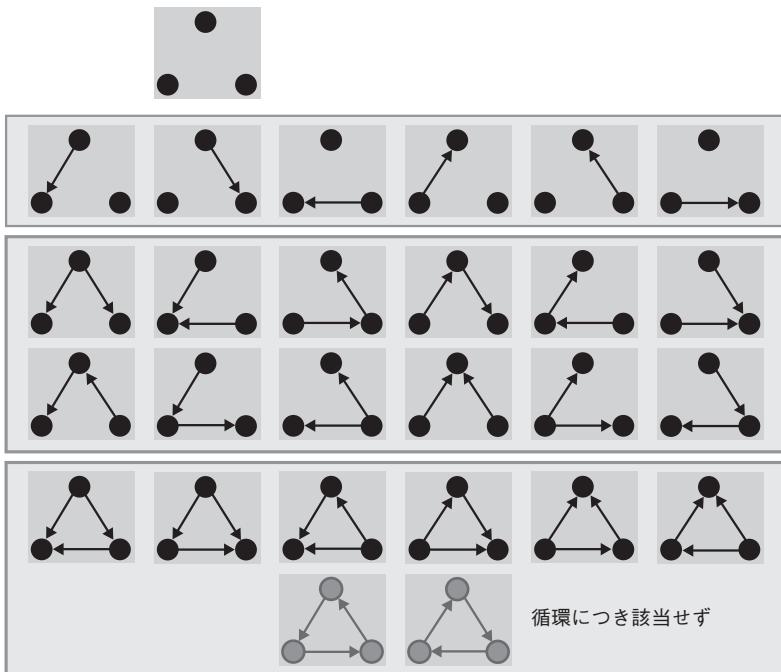


ここで出た数列は "AT&T Labs Research" (<http://www.research.att.com/~njas/sequences/>) が提供している整数列サイト（整数列大事典）で確認できます。

- 順列 (Factorial numbers) : id A000142
- 第2スターリング数 (Stirling numbers of second kind) : id A008277
- ベル数 (Bell or exponential numbers) : id A000110
- カタラン数 (Catalan numbers) : id A000108
- ダブルファクトリアルナンバー (Double factorial numbers) : id A001147
- 非循環有向グラフ数 (Number of acyclic digraphs (or DAGs) with n labeled nodes) : id A003024

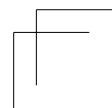
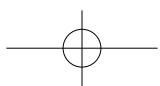
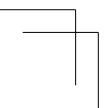
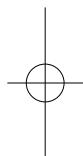
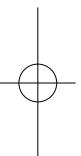
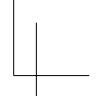
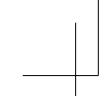
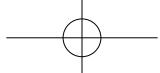


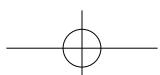
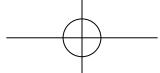
(a) 頂点数 3 の無向グラフのパターン (8 種類)



(b) 非循環有向グラフのパターン (25 種類)、最下段の 2 パターンは循環しているもので数えていない

図 15.6 項目数 3 のグラフ





# 第 16 章

省略する

## 16.1 ランダムに抽出する、ランダムに巡回する

遺伝統計学領域の解析では、場合の数が多いことが頻繁にあるため、工夫が必要になります。何かしらの処理を手順として定め、その手順が効率的であるように工夫したものをアルゴリズムと呼びます。アルゴリズムのもう1つの特徴は、いろいろな場面で使い回せることです。分野を超えて利用されているものも多いです。また、1つの目的に複数のアルゴリズムがあり、それぞれが長所・短所を持ちます。また、多数の亜流が派生的に生じることも多いです。

まず、すべての場合を調べる代わりに、一部を調べることで、全体を調べたことと似た結果を得ることを目指した工夫について見てきます。

### ● 16.1.1 既知の分布からのランダムサンプリング

既知の分布からのランダムサンプリングとは、値がとるべき分布に沿って乱数を発生させて、有限個の値が目指す分布と似ているようにすることです。

乱数を発生させるときに、計算機では疑似乱数列というものを使います（付録 A の A）。疑似乱数列はアルゴリズムに基づいて発生させます。しばしば用いる既知の分布からの乱数発生は、そのための関数が用意されています。3 カテゴリの頻度分布は正三角形領域に表現できますが、そこにディリクレ分布からの疑似乱数列発生により、均等に頻度分布が発生している様子を示したのが R16-1.R、図 16.1 です。

## R ソース 16.1 R16-1.R : 均等に発生する頻度分布

```

1 library(MCMCpack)
2 d<-rdirichlet(1000,c(1,1,1))
  # ディリクレ分布からの均一乱数の発生
3 plot(d[,1],sqrt(3)/2*(d[,2]-d[,3]),xlim=c(0,1),
      ylim=c(-sqrt(3)/2,sqrt(3)/2))
4 # 三角形の周を描く
5 segments(c(0,1,0),c(-sqrt(3)/2,0,sqrt(3)/2),c(1,0,0),
           c(0,sqrt(3)/2,-sqrt(3)/2))

```

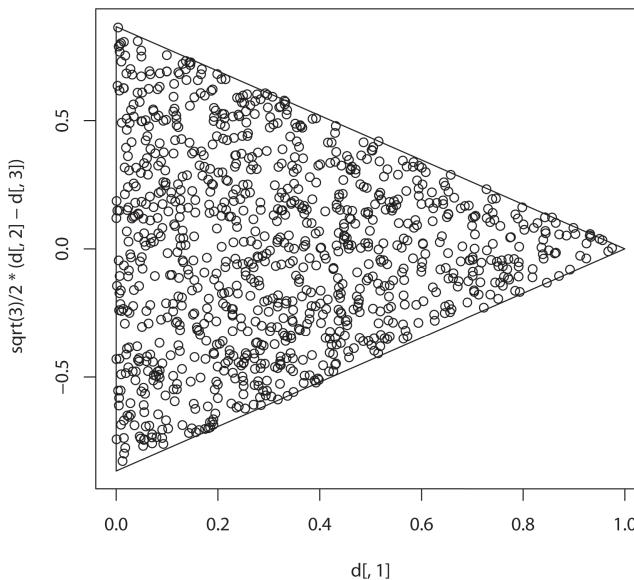
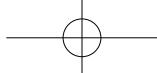


図 16.1 3 カテゴリの頻度の一様分布は正三角形領域に均一に広がる

## ◎ 16.1.2 サンプルを使ってランダムサンプリング、リサンプリングとパーミュテーション

### (1) リサンプリング

得られたデータを用いて、ランダムサンプリングをすることもあります。データがサンプルものなので、「再度」サンプリングするという意味で**リサンプリング**と言います。母集団の分布はわからないながら、サンプルの分布が母集



団の分布に準じたものになっていると考えてこのようにします。データの一部を用いて繰り返しサンプリングをすることで、標本統計量の信頼区間などを推定することに利用します。この方法の1つであるブートストラップ法では、サンプルセットから重複を許してリサンプリングします。

#### R ソース 16.2 R16-2.R : ブートストラップ法

```
1 x<-1:10
2 #bootstrap
3 sample(x,replace=TRUE)
4 #permutation
5 sample(x,replace=FALSE)
```

重複を許したサンプリングの例を示します。

#### ■ R の出力結果

```
[1] 7 4 8 6 7 8 8 3 5 4
```

サンプルの一部を使ってモデル推定をして、モデル推定に使わなかったサンプルを用いてそのモデルを検証することを、**クロス - バリデーション**と呼びます。ここでも、どのサンプルを検証用に選ぶかを決めるために、ランダムサンプリングを行います。また、重複を許さないでリサンプリングすることがあります。重複を許さないすべてのサンプルをリサンプルして、サンプルの順序を入れ替えると、それは、**置換（パーミュテーション）**と呼ばれます。重複を許さないサンプリングは次のようにになります。

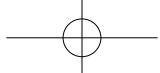
#### ■ R の出力結果

```
[1] 6 2 5 8 7 3 1 9 4 10
```

#### (2) パーミュテーション（順列、置換）

$N$  個の要素のすべてを用いて順列を作ったものは特に**置換（パーミュテーション）**と呼びますが、この置換を用いて、独立性の検定をすることができます。

8 サンプルあって、2 項目 A, B のデータが次に示す表のような場合を考えます。



	A	B
S1	1	1
S2	2	3
S3	3	2
S4	4	6
S5	5	4
S6	6	7
S7	7	8
S8	8	5

ここで、帰無仮説の棄却検定をするために、A と B とが無関係であると仮定します。無関係ならば、8 個のサンプルがとった A の値はどのサンプルがどの値をとるのも自由のはずなので、A の値をシャッフルしてやることにします。今、このシャッフルする場合の数は順列  $P(8, 8)$  (15.1 節) ですから、すべての置換（順列）を観察データと比較してやります。

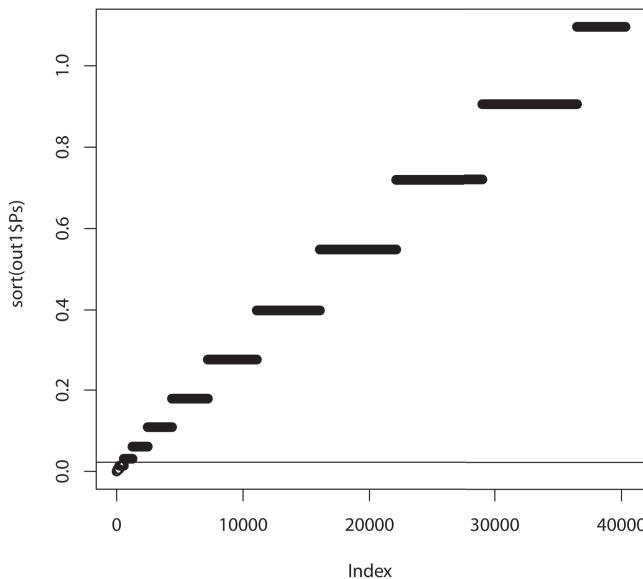
比較するときには、何かしら基準がいります。この場合は、A の値と B の値について、Jonckheere-Terpstra 検定の  $p$  値 (jtp) の大小で比較することにしましょう。jtp が小さいほど A と B とに関係がないという仮説のもとでは「珍しい」わけですから、シャッフルして出た jtp の値（置換（順列）の数だけあります）のうち、観察データの jtp の値以下の置換（順列）数を数え上げます。その数が置換（順列）の総数に占める割合が、パーミュテーション法の  $p$  値です。置換（順列）の総数がさほど多くなければ、すべての置換（順列）を調べ上げればよいですが、たいていの場合は数え切れないで、そのときには適当な数の置換（順列）を作成してやり、それに対応する jtp の値と観察 jtp の値とを比較してパーミュテーション法の  $p$  値とします。

一部の置換（順列）で済ます方法は、計算機がランダムに作った置換（順列）に基づくのでモンテカルロ<sup>※1</sup>なパーミュテーション法と言いますが、現実的には、パーミュテーション法はモンテカルロにて行うことが普通なので、パーミュテーション法と言えばモンテカルロのことを指すことが多いです。

R16-3.R で実行できます<sup>※2</sup>。パーミュテーション法の  $p$  値は離散的な値をとっており、観測表の  $p$  値を水平線で示したので、水平線以下の点の比率がパ

※1 モンテカルロ法とは計算機を使って乱数を発生させて行う方法の総称です。

※2 この処理は時間がかかります。

図 16.2 モンテカルロ・パーミュテーション法による  $p$  値の累積分布、水平線は観測表の  $p$  値

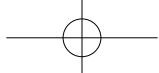
一ミュテーション法の  $p$  値となります。

### R ソース 16.3 R16-3.R : パーミュテーション法

```

1 # データ作成
2 A<-1:8
3 B<-c(1,3,2,6,4,7,8,5)
4 # 全置換でのパーミュテーションかモンテカルロ・パーミュテーションかを
# MonteCarlo で選択
5 PermutationTestJT<-function(A,B,MonteCarlo=TRUE,n=10000) {
  # n モンテカルロのときの試行回数
  library(clinfun) # jonckheere.test() のパッケージ
  ret<-0
  Ps<-NULL
  jtp<-jonckheere.test(A,B)$p.value # 観測表 p 値
  if(MonteCarlo || length(A)>10) {
    # 要素数が 10 より大なら、必ずモンテカルロ
    Ps<-rep(0,n)
    for(i in 1:n){
      tmp<-sample(1:length(A),length(A))
      Ps[i]<-jonckheere.test(A[tmp],B)$p.value
    }
  }
}

```



```
16 }else{ # 全部の置換を実行
17   library(gtools) # permutations() のパッケージ
18   perms<-permutations(length(A),length(A)) # 全置換を返す
19   Ps<-rep(0,length(perms[,1]))
20   for(i in 1:length(perms[,1])){
21     Ps[i]<-jonckheere.test(A[perms[i,]],B)$p.value
22   }
23 }
24 ret<-length(Ps[Ps<=jtp])/length(Ps)
25 list(originalp.value=jtp,permpp.value=ret,Ps=Ps)
26 }
27 out1<-PermutationTestJT(A,B,MonteCarlo=FALSE)
28                                     # 全置換に基づく：正確検定
29 out2<-PermutationTestJT(A,B,MonteCarlo=TRUE,n=100)
30                                     # モンテカルロパーミュテーション法
31 out1$originalp.value # 観察表のjpt
32 out1$permpp.value # 全置換に基づくp
33 out2$permpp.value # モンテカルロパーミュテーションに基づくp。
34                                     シミュレーション回数が値の最小値を決める
35 plot(sort(out1$Ps))
36 abline(h=out1$permpp)
```

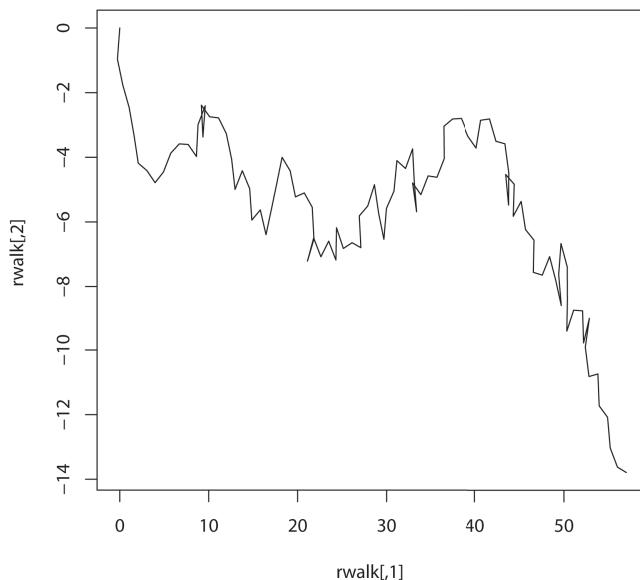
### ◎ 16.1.3 ランダムウォーク

いろいろな状態（変数の値）を確率的にとるようなときに、その状態を網羅的にランダムに作成することが難しいこともあります。そのようなとき、ある状態からスタートして、そこから変化しうる状態を選んで次の状態とすることで、いろいろな状態を作成する方法が**ランダムウォーク**です。

図 16.3 では、2 次元平面のランダムウォークを描きました。

ランダムウォークは、変数の**推定**をしている（探し物をしている）ときに、変数がとりうる空間の全体を歩き回れるなら、適切なところを見つけることができるだろう、という考えに基づく方法です。推定途中の変数の値（位置）を用いて、次のステップでの変数の値（位置）を決めるので、これは**マルコフ連鎖**です。

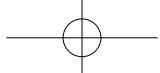
探し物ですから、最終的には求めている点（変数の推定値）へ向かって移動して行って欲しいのです。ですから、ランダムウォークでは空間をやみくもに



ソースは掲載していませんが、ダウンロードできる R16-sup1.R で実行できます。

図 16.3 2 次元平面上のランダムウォーク

歩くのではなく、適切な方向へ進む確率が高いように工夫（アルゴリズム）がなされます。メトロポリス・ヘイスティング法や、共役事前分布を用いるサンプリングなどは、このランダムウォークを効率よくする工夫です。このタイプの処理では、どこから歩き始めるかが結果に影響を与えますから、その影響がなくなるまで（なくなったと判断できるまで）ランダムウォークを続けることや、開始点を複数採用して複数の結果を求め、その一致性を確認することなどもなされます。これらも「工夫」の一部です。



## 16.2 主要な部分のみを使う

### ◎ 16.2.1 近似する

#### (1) モデルに当てはめる

分割表の正確確率がカイ分布の確率密度関数と似ていることを第13章（図13.2）で確認しましたが、そこでは、ある特定の関数の式を想定しその係数を推定しました。Rの`optim()`関数を用いました。この関数は、観測データをある関数に当てはめるにあたり、そこからのずれが最小になるような係数を推定します。繰り返し処理をして係数を改善し、ずれが十分に小さくなるまで続けます。そして、この係数の改善方法についても複数のアルゴリズムが実装されていて、選択することができます。図13.2の例は、当てはめる関数がデータをほとんど完全に説明できる場合でした。

別の例で見てみることにします。**極値分布**というものに触れてみることにします。配列比較の評価（4.2.4項）で極値分布を使うと説明しました。配列の異同を調べるとき、よく似た配列を較べるため2つの配列に関連がないという仮説の棄却という観点で評価すると、非常に極端な値が得られます。このように非常に極端な値は、その由来に関わらず、ある関数で表される分布に従うことが知られており、それを極値分布と言います<sup>※3</sup>。この極値分布の係数を推定してみることにします。

たくさんの仮説をカイ自乗検定して、多くのカイ自乗値を得たときに、そのカイ自乗値の中で最も大きな値を考えます。このようなたくさんの仮説の検定を何セットも行うと、セットの数だけ「最も大きなカイ自乗値」が得られます。この「最も大きなカイ自乗値」が極値分布で近似できます。たくさんの検定を行ったときに、どのくらいの統計量を有意であると考えるかを**多重検定問題**と言いますが、この「最も大きなカイ自乗値」の分布に照らして検定結果を評価することが、多重検定問題を解く鍵です。多重検定問題は第17章で扱い

※3 大水害や、飛行機事故など、稀な事象が起きる確率は、実際に起きることが稀なので、同レベルの事象の発生確率を観察することからは、わかりません。比較的稀な事象の観察を重ね、その分布をとった上で、極値の分布の形を推定することを介して、極端な事象の確率を推定します。このとき用いる分布が極値分布です。

ます。

さて、極値分布は、

$$G(x) = e^{-\left(1+\zeta\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\zeta}}}$$

という式で表されます。

複数の「最も大きなカイ自乗値」を得たら、その分布に適合するような係数  $\zeta, \mu, \sigma$  を推定してやります。

実際に R で行ってみます。R16-4.R を追いかけてみてください。

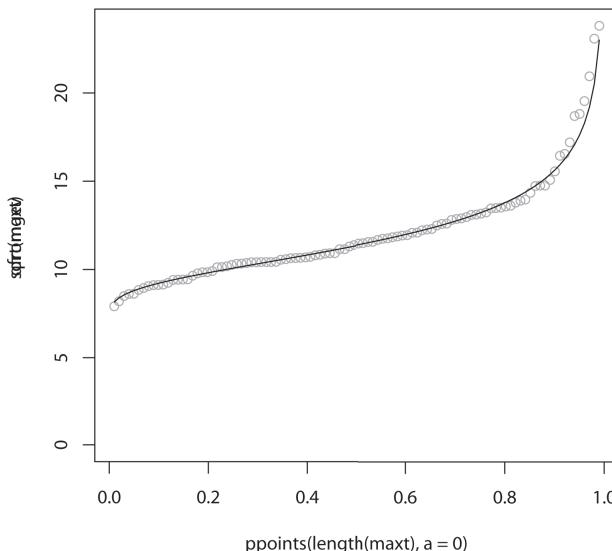
#### R ソース 16.4 R16-4.R : 多重検定と極値分布

```

1 # 自由度 1 のカイ自乗分布乱数を 100000 個発生
2 N<-100000;M<-100;t<-rchisq(N,1)
3 #100 セットに分割
4 matt<-matrix(t,nrow=M)
5 # それぞれのセットの最大値を取り出し
6 maxt<-apply(matt,1,max)
7 #evd パッケージを読み込んでから、その関数を使用して、
8 #Generalized Extreme Value Distribution の 3 パラメタを推定した上で
9 # その累積確率を算出してプロットし、観察値の累積分布と比較
10 library(evd)
11 gevest<-fgev(maxt)
12 qfromgev<-qgev(ppoints(length(maxt),a=0),
13   loc=gevest$estimate[1],scale=gevest$estimate[2],
14   shape=gevest$estimate[3])
15 plot(ppoints(length(maxt),a=0),sort(maxt),
16   ylim=c(0,max(maxt)),type="p",cex=1,col=gray(5/8))
17 par(new=T)
18 plot(ppoints(length(maxt),a=0),qfromgev,
19   ylim=c(0,max(maxt)),type="l")

```

ランダムに発生した 1,000 個のカイ自乗値の中の最大値を 100 セット分作り、それをもとに極値分布を推定しています。100 個の観察された最大カイ自乗値と、推定された極値分布との累積分布をプロットしています。



カイ自乗分布からの乱数 1,000 個のうちの最大値を 100 回観察したときの累積分布が、でこぼこした灰色のプロット。それに合致する極値分布が滑らかな細線。

図 16.4 多重検定と極値分布

## (2) 形式に当てはめる

前項の例は、関数とそれを定める係数とがあって、データを用いて係数を推定するという手続でした。

限られた数の変数で作ったモデルを観察データに当てはめることと同じです。

データをかいづまることを取り上げた第 7 章では、このように、意味のある変数によってデータを捉える方法と、データそのものから変数を引き出す方法があることを述べました（7.3 節）。ここでもその視点で眺めてみます。

データをうまく説明するために、大きく説明する変数から徐々に細かく説明する変数を加えていくことを考えます。固有値分解の手続きがこれに相当しました。

同じく、大づかみから徐々に細かくつかんでいく方法に**多項式近似**というのがあります。

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_k^k$$

という式での近似です。 $k$ 次の多項式近似と言います。線形関数への回帰式の推定と同じことです。平均をとるというのは、0次多項式近似とも言えます。ランダムなカイ自乗値1,000個の最大値の分布について、`optim()`関数を用いて次数を上げるに従って、線形近似がよくなる様子を見てみましょう（R16-5.R）。

#### R ソース 16.5 R16-5.R : 線形近似

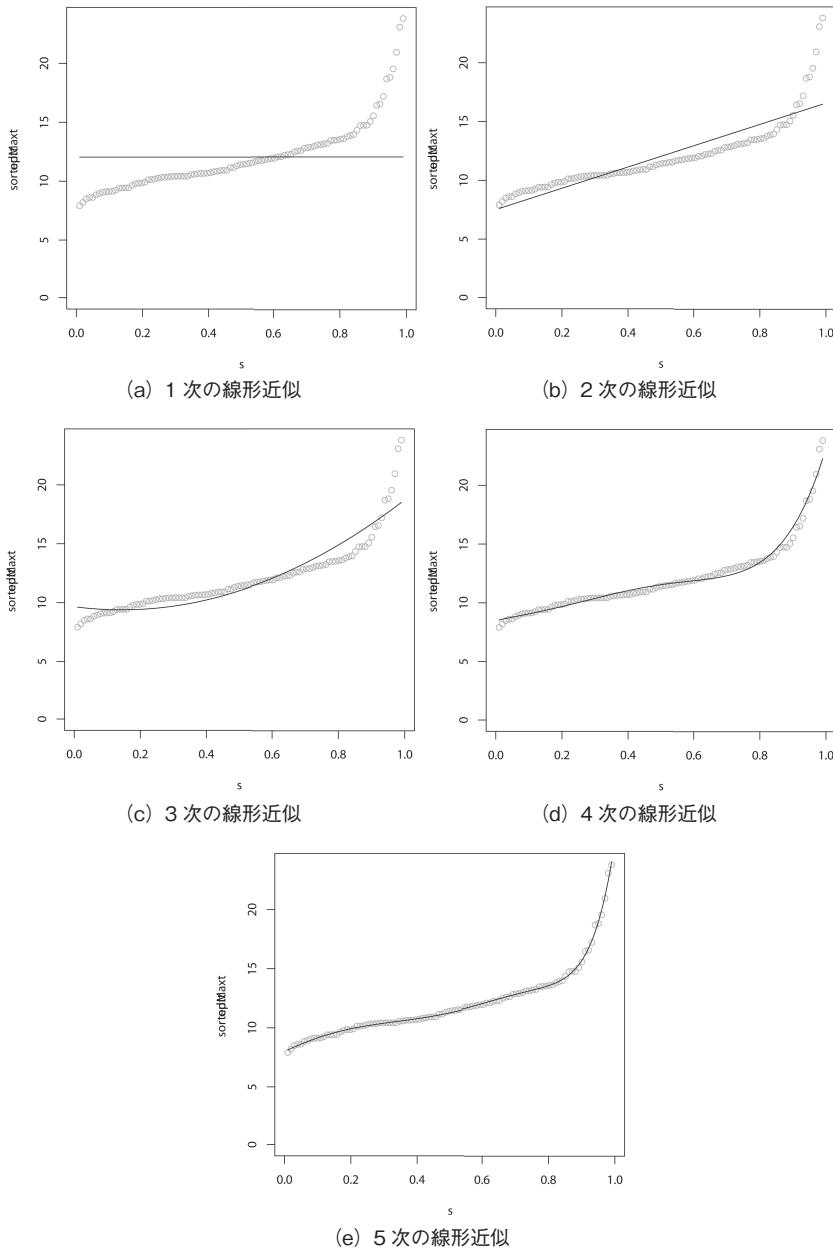
```

1 sortedMaxt<-sort(maxt) # 最大力自乗値の昇順ソートリスト
2 s<-ppoints(length(sortedMaxt), a=0)
3 #たとえば2次での近似
4 f2 <- function(x) {
5   sum((sortedMaxt-(x[1]*s^0+x[2]*s^1+x[3]*s^2))^2)
6 }
7 optout2<-optim(rep(1,3),f2,method="BFGS")
8 ylim<-c(0,max(sortedMaxt))
9 plot(s,sortedMaxt,ylim=ylim,type="l")
10 optt<-rep(0,length(s))
11 for(i in 1:length(optt)){
12   optt[i]<-sum(optout2$par * s[i]^(0:(length(optout2$par)-1)))
13 }
14 par(new=T)
15 plot(s,optt,ylim=ylim,col="red",type="l")

```

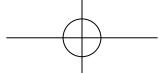
データの説明のときに、自由度の数だけの変数を使って、データのすべてを説明することができる、という話をしました。今回の多項式近似では、 $k$ はどこまでも大きくできます。本来のデータが多項式で表すことができなくても、自由度の数を超えて用いることで、近似をどんどんよくすることができます。

同様に、複雑な関数を取り扱いが容易な成分で近似する方法に、**テイラー展開**などの**無限級数**展開があります。これも微分を繰り返して、無限に細かくする近似です。



カイ自乗分布からの乱数 1,000 個のうちの最大値の累積分布の線形近似。でこぼこの線が観測データ、滑らかな線が近似の結果。

図 16.5 線形近似



## 16.3 意義の大きい方から選ぶ、小さい方から捨てる

分散の分解（固有値分解）と多項式近似・無限級数展開とは前者が有限に分け、後者が無限に分けるという点に違いがありました。次の2点が共通です。

- 大きな特徴を捉えたのち、徐々に細部を説明していく
- 大づかみから詳細説明へと進める手順が決まっている

似たような方法として、複数の要素から、意味のある要素の組み合わせを絞り込むという作業があります。要素の組み合わせは、全部を調べ上げようすると膨大な場合の数となるために、すべての組み合わせを試す代わりに、次のような手順で一部の組み合わせのみを調べることができます。

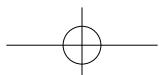
1つ目を選ぶときに一番強い要素を選択し、次に初めに選択した要素は必ず取り入れることとして、残った要素について最大のものを取り出します。常に、その時点での対象から一番のものを選び出すことの繰り返しです。

この逆もあります。

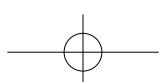
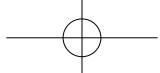
すべての要素をリストに入れた状態からスタートします。ここで、もしだけ1つを外すとしたら、どれが一番外されるべきかを検討します。次に外すときは、すでに外したものは除外することを前提に、外すべき要素を一番不要のものとして外します。多数の要素が順番に並びます。

有限個の変数への分解にあたっても、説明力の強さの順番がありました。無限個への展開の場合も順番がありました。

これらはみな、要素が順番に並んでいて、徐々に説明する内容が小さくなっています。基本的には、大きく捉えている部分に意味が大きく、小さく捉えている部分は意味が小さめです。この順序のあるリストの初めの方からいくつかを取り上げることで、全体の説明とすることが近似なわけですが、では何個目までを取り上げることにして、何個目以降は取り上げないようにするのがよいかの判断が必要なことがあります。



段階を1段進めたときに、それが重要な1段なのか、意味のない1段なのかの判定をするのも、統計的判断です。その判断にあたっては、採用される変数と観察されたデータとの間の尤度、変数の数と自由度の関係に基づいて、**情報量基準**と呼ばれる基準とそれを用いた判断方法が複数あります。詳細には触れませんが、これらの判断方法がこの処理を支えています。



# 第 17 章

## たくさんの検定

### 17.1 多重検定

#### ● 17.1.1 独立な検定の繰り返し

「下手な鉄砲も数撃てば当たる」ということわざにもあるように、珍しい出来事も繰り返して何度も行えば一度くらいは起こるもので。これと同じことで、小さい  $p$  値に相当する珍しい統計量も、繰り返して測れば珍しいことはなくなります。これが、たくさんの検定を行ったときに  $p$  値の解釈を変える理由です。**多重検定**と呼ばれる問題です。

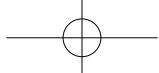
本章では、 $p$  値の解釈の変更方法（補正方法）について述べていきます。

#### ● 17.1.2 多重検定時の $p$ 値の期待値

今、独立な  $k$  個の検定を行ったときに、一番小さい  $p$  値はどのくらい小さいかを考えます。 $k=1$  のときは  $0 \sim 1$  の均一分布で、その期待値は 0.5 です。この期待値が  $k$  とどういう関係にあるかというと、

$$E(\min(p) | k) = \frac{1}{k+1}$$

です。さらに、 $k$  個の検定を行ったときに  $i$  番目に小さい  $p$  値の期待値は、



$$E(i_{th} p \mid k) = \frac{i}{k+1}$$

です<sup>※1</sup>。

### ◎ 17.1.3 一番小さい $p$ 値の補正

#### (1) Sidak の方法

たくさんの要素について多くの検定を繰り返すと、最も小さい  $p$  値の期待値が  $p$  値の名目上の値より小さくなることがわかりました。では、名目上の  $p$  値を実質上の珍しさを表すように補正する方法について考えてみます。

最も基本となるのは、相互に独立な検定を繰り返す場合です。個々の検定の  $p$  値が  $0 \sim 1$  の均一分布となるとします。 $k$  個の検定のそれぞれで  $p < a$  となる確率が  $a$  なので、 $k$  個の検定のすべてで  $p < a$  となる確率は  $a^k$  です。逆に、すべてで  $p \geq a$  となる確率は  $(1 - a)^k$  です。今、 $k$  個の検定の中で 1 番小さい

---

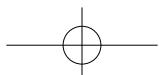
※1 すべての  $p$  値が  $a$  以上である確率は  $Pr(\text{all } p \geq a \mid k) = (1 - a)^k$  です。少なくとも 1 個以上の  $p$  値が  $a$  未満である確率は  $1 - Pr(\text{all } p \geq a \mid k) = 1 - (1 - a)^k$  です。1 番小さい  $p$  値 ( $\min p$ ) が  $a$  から  $a + \delta a$  ( $\delta a > 0$ ) の間である確率は、すべての  $p$  値が  $a + \delta a$  以上ではなくて、すべての  $p$  値が  $a$  以上である場合です。したがって、その確率は  $-(Pr(\text{all } p \geq a + \delta a \mid k) - Pr(\text{all } p \geq a \mid k))$  です。 $\min p = a$  である確率は、 $\delta a$  を極限まで小さくした値なので、

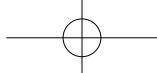
$$Pr(\min p = a \mid k) = -d Pr(\text{all } p \geq a \mid k) / da = k(1 - a)^{k-1}$$

となります。さて、期待値は、とりうる値について起きる確率の重みを付けて積分すればよいので、

$$\begin{aligned} \int_0^1 a Pr(\min p = a \mid k) da &= \int_0^1 ka(1 - a)^{k-1} da \\ &= k \left( \int_0^1 (1 - a)^{k-1} da - \int_0^1 (1 - a)^k da \right) \\ &= \left[ k \left( -\frac{1}{k}(1 - a)^k + \frac{1}{k+1}(1 - a)^{k+1} \right) \right]_0^1 \\ &= k \left( \frac{1}{k} - \frac{1}{k+1} \right) = \frac{1}{k+1} \end{aligned}$$

さらに一般に、 $k$  個のうち  $i$  番目に小さい  $p$  値の期待値は、 $i$  個が  $a$  未満で  $k - i$  個が  $a$  以上である確率が  $\frac{k!}{i!(k-i)!} a^i (1 - a)^{k-i}$  であることから、最小  $p$  値の場合より多少面倒ですが、 $\frac{i}{k+1}$  であることを示すことができます。





$p$  値が  $a$  であったとします。これよりも珍しいのは、

「 $k$  個のうちどれか 1 つでも  $p$  値が  $a$  未満である」

= 「すべての検定で  $p \geq a$  であるわけではない」

ような珍しさと同じなので、「すべての検定で  $p \geq a$  でない」確率となります。それは、 $1 - (1 - a)^k$  です。逆に言えば、 $1 - (1 - a)^k = 0.01$  のような  $a$  を観測したら、それが  $k$  個の独立な検定を行ったときに、100 回に 1 回しか起きないほど珍しいこと、という意味で、「多重検定をしたときの補正後  $p$  値  $p_c = 0.01$ 」です。

$$1 - (1 - a)^k = p_c$$

今、ある検定で  $p = 0.00001$  を観察したとします。ただし、一緒に  $k$  個の検定をしていて、その  $k$  個の検定の中で最も小さい  $p$  値だったとします。 $k$  を 1 個から 10 万個まで増やして、 $p_c$  の値がどのように変化するかをプロットしてみます。

$k = 1$  のときには  $p_c = 0.00001$  と、 $p$  値をそのままの値で評価すればよいですが、 $k = 100000$  のときは  $p_c = 0.6321224$  と、全く帰無仮説を棄却しないことがわかります。

逆に、 $k$  個の検定をしたときに、どれくらい小さい名目上の  $p$  値をとると、それはたくさんの検定をしたにも関わらず十分珍しく、0.01 の棄却水準でも仮説を棄却できるでしょうか？

$$a = 1 - (1 - p_c)^{\frac{1}{k}}$$

ですから、

- $k = 1$  の場合には、 $p = 0.01$  ならば補正後も  $p_c = 0.01$  です
- $k = 100000$  のときは、名目上の  $p$  値が  $1.005034 \times 10^{-4}$  くらい小さいときに、補正後が  $p_c = 0.01$  となります

この方法で個々の検定の  $p$  值を多数の同時検定の数に応じて補正する方法を Sidak の方法と呼びます。

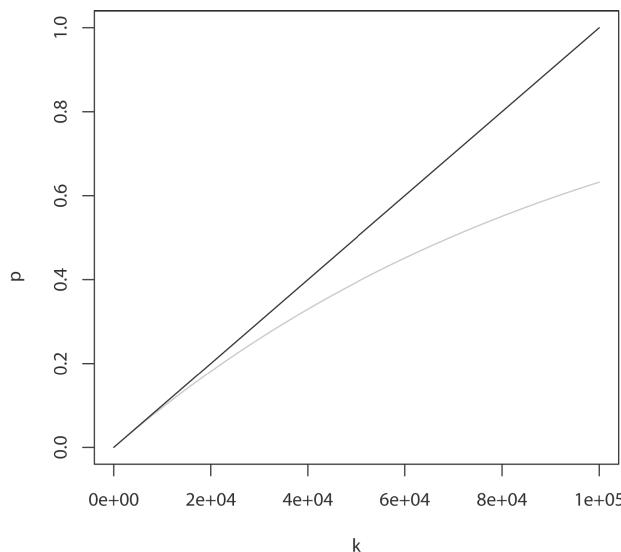


## (2) Bonferroni の方法

これよりももっと単純に、

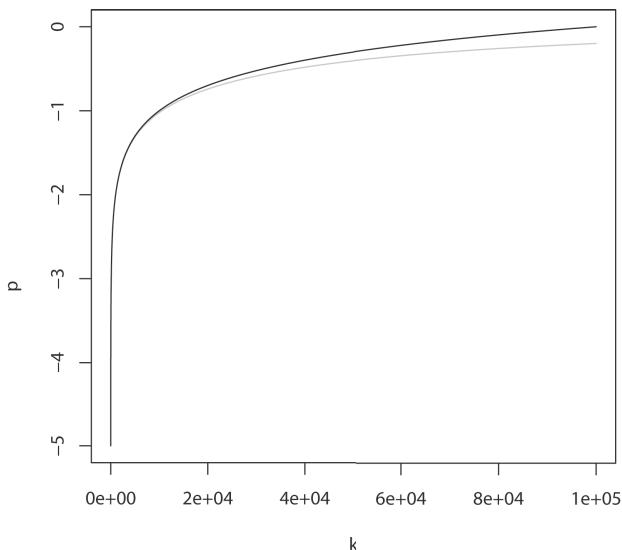
$$p_c = a \times k$$

とする方法もあります。1,000 個の検定を行ったら、名目上の  $p$  値を 1,000 倍して補正後の  $p$  値とする、という単純なものです。Sidak の方法と Bonferroni の方法を比較したのが図 17.1 (a) (b) です。Bonferroni の方法では、 $k$  が大きくなるに連れて  $p_c$  が Sidak の方法よりずいぶんと大きめになることがわかります。Bonferroni の方法では  $p$  値が大きめに出ますから、保守的な補正法であると言われます。逆に言えば、Bonferroni の方法を適用しても  $p_c$  が棄却水準を越えている場合には、十二分の自信を持って、仮説を棄却できるとも言えます。

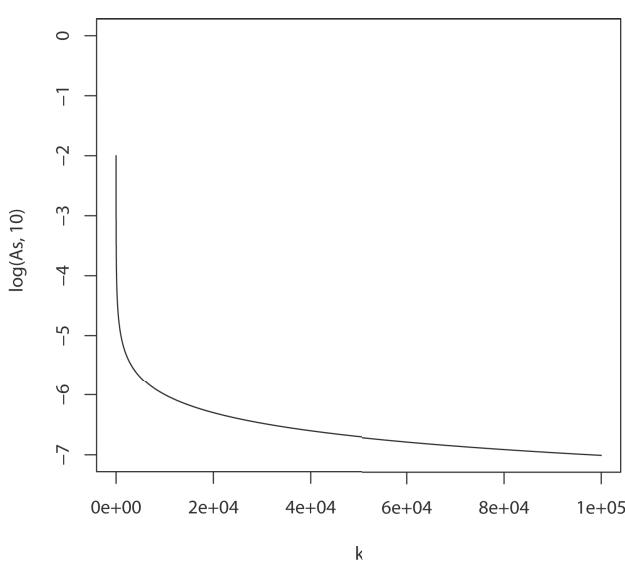


(a)  $k$  個の独立な検定を繰り返したときに、 $p = 0.00001$  をいくつに補正するか（下の灰色線が Sidak の方法、上の黒線が Bonferroni の方法）：通常スケール

図 17.1 Sidak の方法と Bonferroni の方法



(b)  $k$  個の独立な検定を繰り返したときに、 $p = 0.00001$  をいくつに補正するか（下の灰色線が Sidak の方法、上の黒線が Bonferroni の方法）：常用対数スケール



(c)  $k$  個の検定を繰り返したときに、Sidak の補正後  $p$  値が 0.01 となるような名目上の  $p$  値をプロットしたもの

ソースは掲載しませんが、ダウンロードできる R17-sup1.R で描画できます。

図 17.1 Sidak の方法と Bonferroni の方法（つづき）

### ● 17.1.4 非独立な検定の繰り返し

第11章で $2 \times 3$ 表に複数の検定を実施したときに、その検定の  $p$  値は相互に独立でないことを見ました。形質同士に共通点があったり、遺伝マーカー同士に共通点があったりする（連鎖不平衡・集団構造化）ときには、個々の検定は独立ではありません。そのようなときに、Bonferroni の補正法や Sidak の補正法を用いると保守的に過ぎて、偽陰性が増えてしまいます。

しかしながら、数多くの検定の  $p$  値の中の最小の  $p$  値がどのような分布をとるのかが不明なので、補正の仕方がわかりません。逆に言えば、数多くの検定の  $p$  値の中の最小の  $p$  値の分布がわかれば、その分布に照らして補正すればよいのです。たくさんの検定をしているときには、手元にたくさんのデータがありますから、それを用いてリサンプリングの手法を利用して分布を推定してやり、その推定した分布におけるクオンタイルから多重検定補正を行うことができます。

リサンプリングとしてパーミュテーション（置換・順列）を用いる方法を利用することができます。それは、パーミュテーション法と呼ばれました。多重検定のパーミュテーションによる補正をしてみます。

### ● 17.1.5 モンテカルロ・パーミュテーションによる多重検定補正

今、ある因子  $P$  と複数の因子  $X_i, i = 1, 2, \dots$  を検定するとして、 $X_i, X_j$  の独立が言えないような場合を考えます。まず、 $P$  と  $X_i$  との検定を実行します。次に、 $P$  の値は  $X_i$  の値と独立なのか、そうでないかを検定したいので、 $P$  の値をリサンプリングによって変更します。パーミュテーション法は繰り返し抽出を許さないタイプのリサンプリング方法ですから、置換（順列）( $P_p$ )を作ることになります。置換（順列）ごとに、 $P_p$  と  $X_i$  との検定を実施します。興味があるのは、 $P$  と  $X_i$  との検定の最小  $p$  値が、 $P_p$  と  $X_i$  との検定の最小  $p$  値の分布におけるクオンタイルになることです。このクオンタイルがパーミュテーション法による補正後の  $p$  値です。

以下では、相互に依存関係のある複数項目のデータ  $X_i$  を作り、 $P$  との  $t$  検定を行っています。

リサンプリングの繰り返しとともに、補正後  $p$  値がどのように変化するかをプロットしたのが図17.2です。徐々にある値に収束していく様子がわかります。十分に収束すれば、それがパーミュテーション法での補正  $p$  値と言えます。

なお、パーミュテーション法による補正  $p$  値は、繰り返しリサンプリングの回数が  $N$  回のときには、 $\frac{1}{N}$  より小さくなりません。逆に、 $N$  回の繰り返しによって、観察最小  $p$  値以下の最小  $p$  値が出なくとも、せいぜい  $\frac{1}{N}$  未満であるとしか言えません。

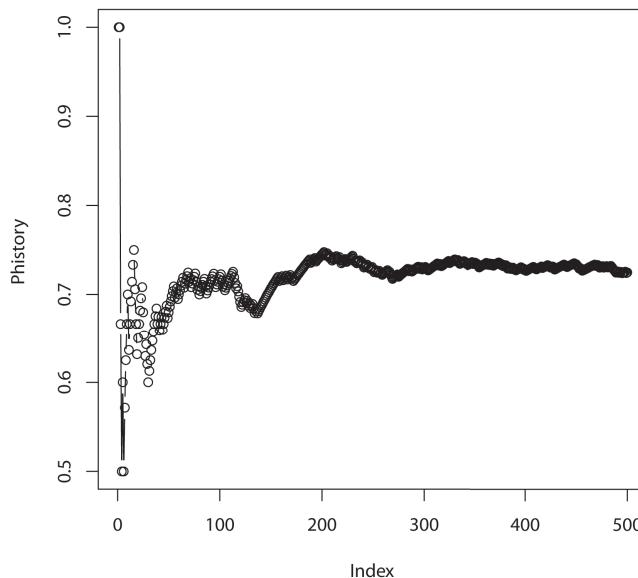
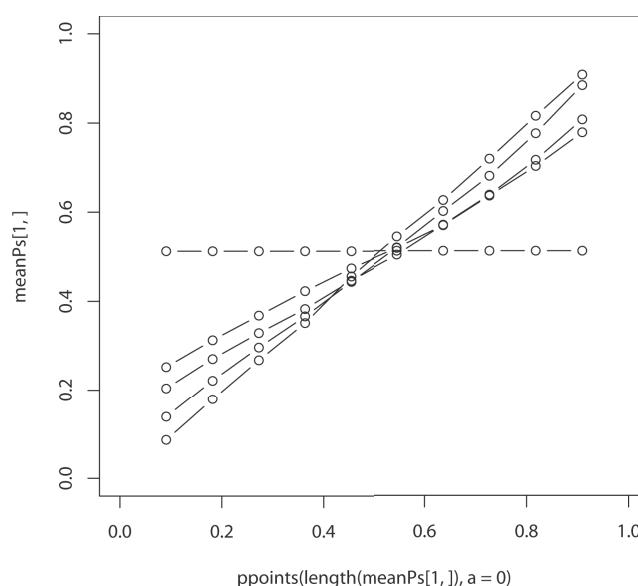
図 17.2 は R17-1.R のソースで実行した結果です。

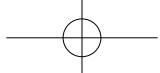
#### R ソース 17.1 R17-1.R : パーミュテーション法による $p$ 値の収束

```

1  # パーミュテーション法による最小 p 値のクオンタイル算出
2  Ns<-100;Nm<-10;Niter<-500
   # サンプル数、マーカー数、パーミュテーション試行回数
3  # P の置換 (順列) と d との t 検定をする関数
4  Fx<-function(d){
5    t.test(d-P[shuffle])$p.value
6  }
7  P<-rbinom(Ns,1,0.5) # ケース・コントロールの 2 値型フェノタイプ
8  X<-matrix(rnorm(Ns*Nm),nrow=Ns)
9  meanPs<-matrix(0,Ns,Nm);Pminhistories<-matrix(0,Ns,Niter)
10
11 X<-matrix(rnorm(Ns*Nm),nrow=Ns)
12 r<-0.3
13 for(i in 2:Nm){
14   R<-sample(c(0,1),Ns,replace=TRUE,prob=c(r,1-r))
15   X[,i]<-X[,i-1]*(1-R)+X[,i]*R
16 }
17 shuffle<-1:Ns
18 obsPs<-apply(X,2,Fx)
19 obsMinP<-min(obsPs)
20
21 Plist<-matrix(0,Niter,Nm)
22 Phistory<-rep(0,Niter)
23 counter<-0
24 for(i in 1:Niter){
25   shuffle<-sample(1:Ns)
26   Plist[i,]<-sort(apply(X,2,Fx))
27   if(min(Plist[i,])<=obsMinP){
28     counter<-counter+1
29   }
30   Phistory[i]<-counter/i
31 }
32 plot(Phistory, type = "b")

```

(a) リサンプリング回数が増えるとパーミュテーション補正後  $p$  値は収束していく(b) 1, 2, ..., 10 番目に小さい  $p$  値の期待値は検定の関連の強さによって変化する図 17.2 パーミュテーション法による  $p$  値の収束



### ● 17.1.6 非独立な検定を繰り返したときの最も小さい $p$ 値

さて、 $X$  の各因子が相互に独立なときには、 $i$  番目に小さい  $p$  値の期待値は  $\frac{i}{N+1}$  でした。 $p$  値の  $X$  のデータの依存の程度を強くすると、 $i$  番目に小さい  $p$  値の期待値はどのように変わるのでしょうか？

10 個の変数について検定をします。この 10 変数のデータは 5 段階の関連を持たせます。全く関連がない場合には、17.1.2 項で述べたように、 $i$  番目に小さい  $p$  値の期待値は  $\frac{i}{10+1}$  となって、図 17.2 (b) では最も傾きの大きなグラフとして描かれています。徐々に変数間の関連を強くするとグラフの傾きが小さくなり、最も関連が強い場合（10 変数のすべてが常に同じ値をとるようにしてあります）では、水平になっています。10 個の変数 ( $X_1, \dots, X_{10}$ ) がすべて同じときにはすべての変数が同じ  $p$  値をとるので、大小の差が出ません。その結果、最も小さい  $p$  値の期待値も最も大きい  $p$  値の期待値も同じ（水平グラフ）になります。



## 17.2 $p$ 値が均一に分布しないとき



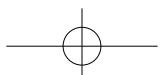
17.1 節では、帰無仮説が成り立つ場合、 $p$  値の期待値が 0.5 となるような場合に関して取り上げました。本節では、そうでない場合について、いくつかのパターンに分けて考えてみます。

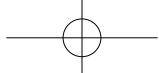
### ● 17.2.1 $p$ 値が小さめに出るとき——ジェノミックコントロール法

$p$  値の期待値が 0.5 ではなく、小さめに出る場合というのを考えてみます。

今、ケースとコントロールが、少々異なる集団からサンプリングされたとします。**集団構造化**<sup>※2</sup>のある集団からケースとコントロールが不均一にサンプリ

※2 サンプルは遺伝的に多様です。その多様性は、完全には混ざりきっていないために、偏りがあります。HWE からずれているのもそのような例です。また、多民族が構成している集団などでは、社会的要因などがあり、ランダムメイティングが実現しないことなどもその要因です。このような集団から、着目形質の解析のためにサンプリングをすると、ケースサンプルとコントロールサンプルの遺伝的背景が完全に一致しないことがあります。このように、サンプリングする母集団が不均一な集団であることを、集団が遺伝的に構造化していると言います。





ングされたような場合です。

ゲノム上の多数のマーカーについて、ケース・コントロール間で関連検定をすることにします。たくさんのマーカーの大多数は、ケースとコントロールの形質の違いとは無関係だとします。その代わり、ケースとコントロールは遺伝的に少々違うので、マーカーごとの関連検定の  $p$  値は  $0 \sim 1$  の均一分布よりも小さめに出ると思われます。どの程度、どのように小さめに出るかを調べてみます。

ケースとコントロールは、HWE にある集団からのランダムなサンプルであるとします。ケースはすべて第1群からサンプリングし、コントロールはすべて第2群からサンプリングすることにします。そして、それぞれの母集団は、マーカーのアレル頻度が同じこともあるし、異なることもあるものの、小さく異なることの方が多く、大きく異なることは少ないと思います。マーカーはたくさんありますし、2集団間で混ざりのよいマーカーとそうでもないマーカーがあると考えることは悪くないでしょうから、アレル頻度の違いが平均0の正規分布に従うとしてシミュレーションしてみます。

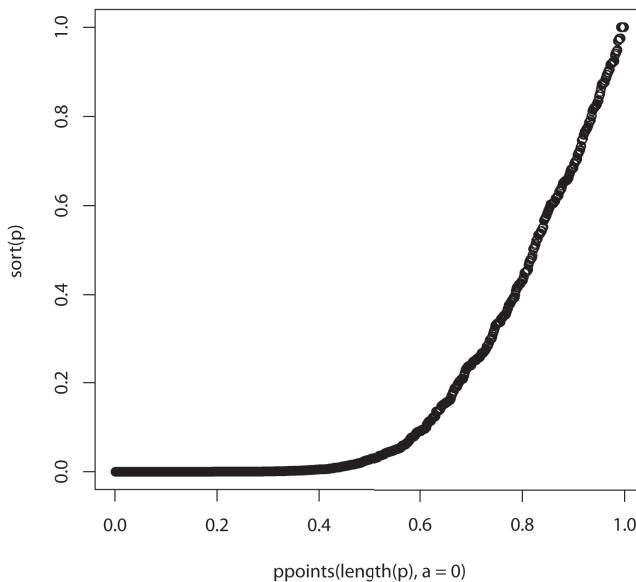
### R ソース 17.2 R17-2.R : ジェノミックコントロール

```
1 set.seed(995599);Niter<-1000 # 試行回数
2 library(Rassoc)
3 st<-rep(0,Niter);p<-rep(0,Niter) # 結果格納ベクトル
4 for(i in 1:Niter){
5   af<-runif(1)*0.6+0.2 # およそのアレル頻度
6   delta<-rnorm(1) # 2群のアレル頻度の違いは正規分布で決める
7   af1<-af+af*0.05*delta; af2<-af-af*0.05*delta
                                # 群別のアレル頻度
8   # ケース・コントロールをそれぞれ第1群・第2群から
     ランダムサンプリング (HWEを仮定)
9   case<-sample(c(0,1,2),1000,c(af1^2,2*af1*(1-af1),
    (1-af1)^2),replace=TRUE)
10  cont<-sample(c(0,1,2),1000,c(af2^2,2*af2*(1-af2),
    (1-af2)^2),replace=TRUE)
11  # 分割表作成
12  t<-matrix(c(length(case[case==0]),length(case[case==1]),
    length(case[case==2]),
    length(cont[cont==0]),length(cont[cont==1]),
    length(cont[cont==2])),nrow=2,byrow=TRUE)
```

```

14 cattout<-CATT(t) # 相加的モデルで検定
15 st[i]<-(cattout$statistic)^2; p[i]<-cattout$p
16 }
17 plot(ppoints(length(p),a=0),sort(p)) # p 値の昇順プロット
18 # 色々なλで補正してみる
19 ylim=c(0,1)
20 plot(ppoints(length(p),a=0),sort(p),ylim=ylim,type="l") # 観測 p のプロット
21 for(i in 2:20){ # 補正 p のプロット
22 par(new=T)
23 plot(ppoints(length(p),a=0),sort(pchisq(st/i,1,
24 lower.tail=FALSE)),ylim=ylim,col="red",type="l")
25 } # 観測 p 値の中央値が 0.5 となるように λ の値を算出
26 lambda<-quantile(st,0.5)/qchisq(0.5,1)
27 stc<-st/lambda
28 pc<-pchisq(stc,1,lower.tail=FALSE)
29 par(new=T)
30 plot(ppoints(length(pc),a=0),sort(pc),type="l") # 中央値を利用した補正法による補正後 p のプロット

```

図 17.3 構造化集団でのケース・コントロール関連検定の  $p$  値

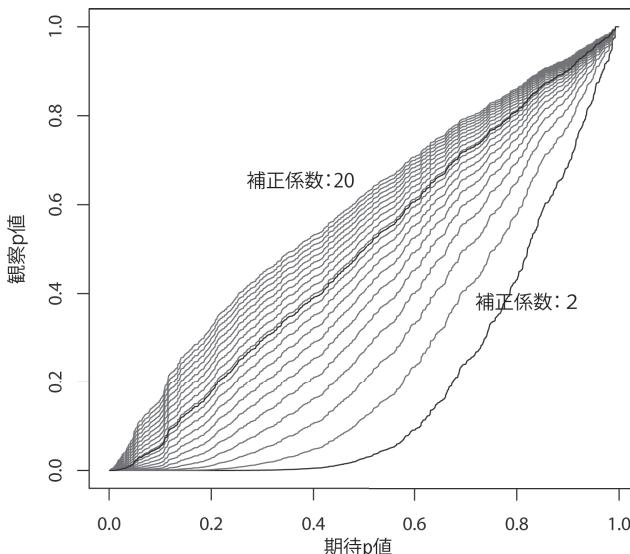
帰無仮説が成り立っているときには、昇順に並べた  $p$  値はグラフで対角線上に並ぶはずですが、図 17.3 では対角線より下に凸になっています。これは、期待するより小さい  $p$  値が多いことを意味します。

この  $p$  値を次のように補正してみることにします。

今回の検定は相加的遺伝形式に関する Cockran-Armitage の傾向性検定（自由度 1）を用いました。その検定統計量であるカイ自乗値を正のある値  $\lambda$  で割り、それを補正後カイ自乗統計量とし、この補正後値を自由度 1 のカイ自乗分布で評価して  $p$  値にするのです。

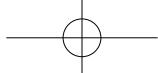
その様子を見てみます。図 17.4 では  $\lambda = 2, 3, \dots, 20$  としてあります。下に凸だったプロットが徐々に、対角線（均一分布）に近づいて、さらに行き過ぎて、上に凸になっていく様子が見てとれます。

この性質を利用して、観察された統計量から補正項  $\lambda$  を計算して、均一分布になるように補正することを考えます。観察カイ自乗値の中央値が、理論的



下に大きく凸の黒線が構造化集団でのケース・コントロール関連検定の  $p$  値。ほぼ対角線の黒太線が中央値を利用した補正後の  $p$  値。その他の細いたくさんの線は、補正係数を 2 から 20 まで変えたときの補正後の  $p$  値のプロット。

図 17.4 ジェノミックコントロール法



カイ自乗値の中央値の何倍であるかを補正項  $\lambda$  にするのが 1 つの方法です<sup>※3</sup>。

この方法を **ゲノムワイド形質マッピング** では **ジェノミックコントロール法** と呼んで、ゲノム上のマーカーを用いて大規模に関連検定を行うときの集団構造化補正に用います。

### ◎ 17.2.2 対立仮説が成り立つとき——非心カイ自乗分布

17.2.1 項ではケースとコントロールとが異なる集団からのサンプルである場合でした。そして、検定する多数の因子の頻度の集団間の違いが 0 を中心とした正規分布になるような例でした。

今度は、帰無仮説が成り立たず**対立仮説**が成り立つとき（関連が真であるとき）に、帰無仮説の棄却検定を実施する場合を考えてみます。当然のことながら、**帰無仮説**を棄却する検定の  $p$  値は均一ではなくなります。ジェノミックコントロールの例では、ケースとコントロールとでアレル頻度の差をゼロを中心とする正規分布でとりましたが、今回は、あるマーカーで 2 群間にアレル頻度の差があるという対立仮説が成り立つときにランダムサンプリングをし、その検定  $p$  値の累積分布を示します（図 17.5 (a)）。

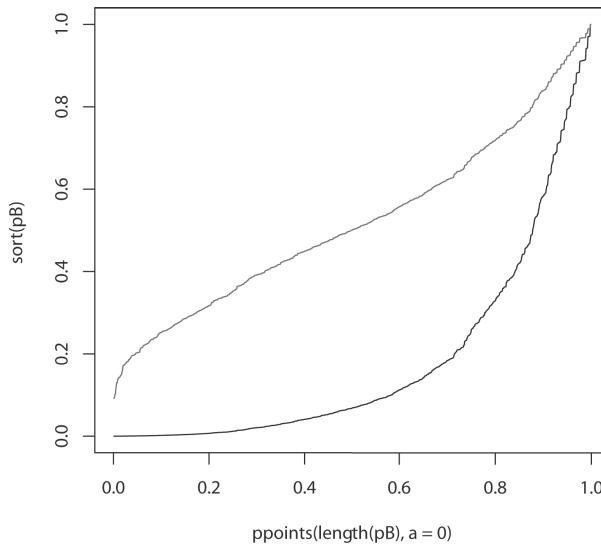
大きく下に凸の線が、関連が真であるときの  $p$  値をソートしてプロットしたもので、図 17.3 の図と同様に下に凸ですが、17.2.1 項で紹介したジェノミックコントロール法で補正してみると対角線になりません（図 17.5 (a) の立ち上がりの早い線）。補正後  $p$  値が対角線にならないことから、この補正が無効であることがわかります。

対立仮説が真の場合のカイ自乗値は、行と列との独立を仮定した期待値ではない値（対立仮説での期待値）が最も観察されやすく、そこを中心として距離の 2 乗に応じて生起確率が減少するからです。このような分布を**非心カイ自乗分布**と言います。

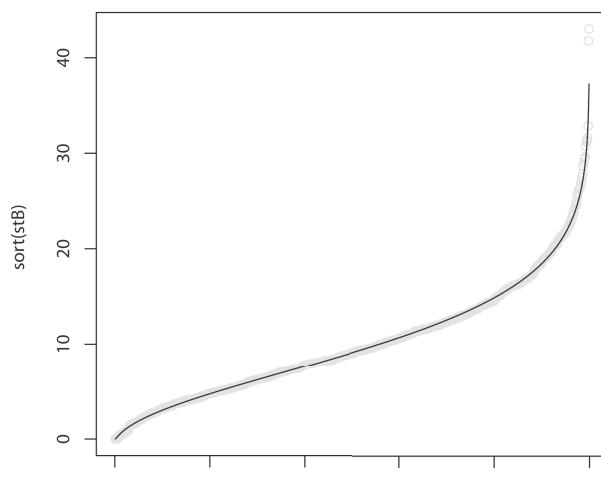
非心カイ自乗分布は自由度 ( $k$ ) と、ピークの原点からのずれを表す変数 ( $r$ ) とで形が決まります。非心カイ自乗分布の平均が  $k+r$  であることを利用して、発生させた検定統計量の平均から  $r$  を算出し、それに基づいて非心カイ自乗分布をプロットしてみると、確かに、シミュレーションして得られた統

※3 中央値は、値のセットの中にはずれ値があってもその影響が及びにくい値ですから、たくさんのマーカーの中に、少しくらいの本当に形質と関連しているマーカーが混入していても、影響はほとんどないという利点があります。

計量の分布とよく重なります（図17.5（b））。



(a) 対立仮説が成立立つきの  $p$  値の昇順ソートプロット（下側の線）。また、それをジェノミックコントロール法補正した  $p$  値を合わせて示しています。ジェノミックコントロールによって補正が適切になされないことがわかります。



(b) 対立仮説下での相加的モデル検定  $p$  値と非心カイ自乗分布との一致を示しています。灰色の太い線が観測値。黒い細線が非心カイ自乗分布。

図 17.5 非心カイ自乗分布

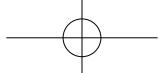
このシミュレーションは R17-3.R で実施しました。

### R ソース 17.3 R17-3.R : 非心カイ自乗分布

```

1 library(Rassoc)
2 stB<-rep(0,Niter);pB<-rep(0,Niter) # 統計量と p 値の格納ベクトル
3 # 2 群に異なるアレル頻度を与える
4 af<-runif(1)*0.6+0.2 ;delta<-0.05;af1<-af+af*delta;
5 af2<-af-af*delta
6 exptable<-matrix(c(af1^2,2*af1*(1-af1),(1-af1)^2,af2^2,
7 2*af2*(1-af2),(1-af2)^2),nrow=2,byrow=TRUE)*1000
8 # 相加的モデル検定
9 cattout<-CATT(exptable)
10 expSt<- (cattout$statistic)^2 # 対立仮説下でのカイ自乗値期待値
11 for(i in 1:Niter){ # 対立仮説下でのランダムサンプリング
12 case<-sample(c(0,1,2),1000,c(af1^2,2*af1*(1-af1),
13 (1-af1)^2),replace=TRUE)
14 cont<-sample(c(0,1,2),1000,c(af2^2,2*af2*(1-af2),
15 (1-af2)^2),replace=TRUE)
16 t<-matrix(c(length(case[case==0]),length(case[case==1]),
17 length(case[case==2]),
18 length(cont[cont==0]),length(cont[cont==1]),
19 length(cont[cont==2])),nrow=2,byrow=TRUE)
20 cattout<-CATT(t); stB[i]<- (cattout$statistic)^2;
21 pB[i]<-cattout$p
22 }
23 # ジェノミックコントロール法を適用してみる
24 lambdaB<-quantile(stB,0.5)/qchisq(0.5,1) ;
25 stcB<-stB/lambdaB;pcB<-pchisq(stcB,1,lower.tail=FALSE)
26 plot(ppoints(length(pB),a=0),sort(pB),type="l",ylim=ylim)
27 par(new=T)
28 plot(ppoints(length(pcB),a=0),sort(pcB),type="l",col="red",
29 ylim=ylim)
30 # 対立仮説下でのカイ自乗統計量の平均と分散
31 meanstB<-mean(stB);varstB<-var(stB)
32 # 非心カイ自乗分布の算出
33 # 非心パラメタは平均 - 自由度であることを利用して
34 ncp<-meanstB-1
35 # 非心カイ自乗分布と観察カイ自乗とをプロット
36 ncpvalue<-qchisq(ppoints(length(pcB),a=0),1,ncp)
37 ylim<-c(0,max(stB,ncpvalue))

```



```
30 plot(ppoints(length(stB),a=0),sort(stB),ylim=ylim,  
       col=gray(7/8))  
31 par(new=T)  
32 plot(ppoints(length(stB),a=0),ncpvalue,type="l",ylim=ylim)
```

### ● 17.2.3 検定のパワー

対立仮説が成り立っているという条件では、統計量の分布が帰無仮説の場合と異なりました。今、2つの仮説があるとき、1つ目の仮説のもとで得られる確率密度分布と2つ目の仮説のもとで得られる確率密度分布とが決まります。17.2.2項の場合は、自由度1のカイ自乗分布と自由度1の非心カイ自乗分布でした。図17.6にこの2つの分布を描きました。右下がりの曲線が自由度1のカイ自乗分布、低い山型の分布が非心カイ自乗分布です。ここで、統計量がある値（図17.6の垂直線）よりも大きければ、1つ目の仮説を棄却することになります。1つ目の仮説が真のときに統計量がこの値より大きい確率（P1）は、この垂直線よりも右側の部分について、この分布を積分した値になります。

他方、2つ目の仮説が真のときにこの垂直線よりも右側のカイ自乗値を観察する確率（P2）は、低い山型曲線に関して積分した値になります。この値P2が相当程度大きいとき、仮説1は棄却され、仮説2は棄却されません。P1が検定の棄却水準で、P2が検定のパワーです。

R17-4.Rで実行できます。

#### R ソース 17.4 R17-4.R : 棄却水準とパワー

```
1 x<-seq(from=0,to=30,by=0.01)  
2 df<-1;lambda<-ncp  
3 chi<-dchisq(x,df,0) # カイ自乗分布  
4 ncChi<-dchisq(x,df,lambda) # 非心カイ自乗分布  
5 ylim=c(0,0.2)  
6 plot(x,chi,ylim=ylim,type="l")  
7 par(new=T)  
8 plot(x,ncChi,ylim=ylim,type="l",col="red")  
9 abline(v=qchisq(0.05,1,lower.tail=FALSE))  
10 pchisq(qchisq(0.05,1,lower.tail=FALSE),1,ncp,  
           lower.tail=FALSE) # パワーの計算
```

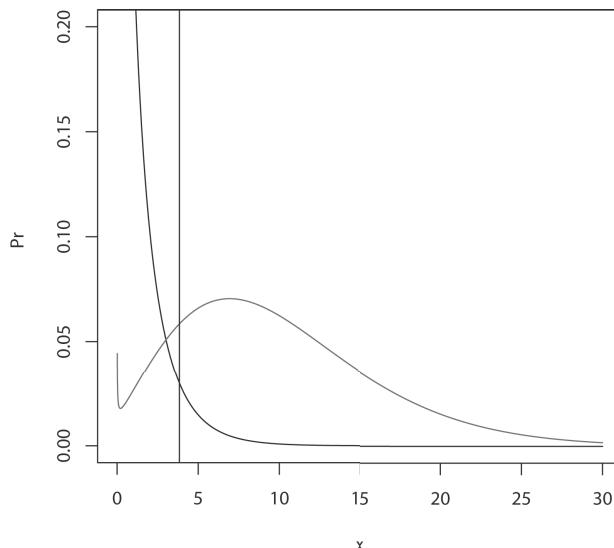


図 17.6 自由度 1 のカイ自乗分布と非心カイ自乗分布の確率密度分布、垂直線は  $p=0.05$  の棄却閾値

帰無仮説（1つ目の分布）と対立仮説（2つ目の分布）のもとでの統計量の分布がわかっていて、帰無仮説を棄却する基準（垂直線）が決まればパワーが計算できるわけです。

逆に言えば、帰無仮説と対立仮説のもとでの統計量の分布を定めるだけの情報（変数）と棄却水準とのすべてとが決まらなければパワーは決まりません。帰無仮説と対立仮説のもとでの統計量は、それぞれの仮説の構成要素（因子ごとの頻度など）とサンプルサイズとで決まりますから、結局、以下の要素がパワーに関わる諸因子です。

- 仮説の構成要素
- サンプルサイズ
- 棄却水準
- パワー

この諸条件のうち1つ以外を既知としてやることで、残りの1つの条件の値が決まります。その様子をRで見てみます。Rの`power.prop.test()`関数では、サンプル数、2群の因子陽性比率、パワーのいずれかを未知として、

他を指定することで、未知変数の値が返ります。

#### R ソース 17.5 R17-5.R : パワー

```
1 help(power.prop.test)
2 help(power.t.test)
3 help(power.anova.test)
4 power.prop.test(p1 = 0.5, p2 = 0.4, sig.level = 0.01,
                  power = 0.9)
```

R17-5.R の出力は以下のようになります。指定しなかったサンプル数  $n$  が算出されて、指定した変数の値とともに示されています。

#### ■ R の出力結果

```
Two-sample comparison of proportions power calculation

n = 734.0538
p1 = 0.5
p2 = 0.4
sig.level = 0.01
power = 0.9
alternative = two.sided

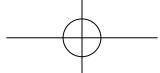
NOTE: n is number in *each* group
```



## 17.3 たくさんの結果の分布を活用する

### ◎ 17.3.1 主成分分析を使って補正する

サンプル間に非独立性があるときに、 $p$  値が小さめに出ました。そしてそれをジェノミックコントロール法で補正をしました。この方法では、帰無仮説が成立立つとみなせる多くの検定結果の分布を、あるべき分布（ $p$  値として均一分布）になるように1個の変数で補正をしていました。この補正法では、算出した補正係数（1より大）によりカイ自乗値を割って、それを補正統計量として用いましたから、すべての検定結果が、補正後には帰無仮説を棄却しにくい方向へと補正します。



今、あるマーカーがフェノタイプに関係しており、ケース群で多いはずとします。しかしながら、ケースとコントロールとが異なる2集団からサンプリングされてしまったとしましょう。そしてこのマーカーはケース群をサンプリングした母集団での頻度が低めで、コントロール群をサンプリングした母集団での頻度が高めであったとすると、マーカーとフェノタイプとの関係が希釈されて、検出しにくくなります。逆の場合には、バイアスのかかったサンプリングの影響は、マーカーとフェノタイプの関係を実際より大きく見せる働きをするでしょう。

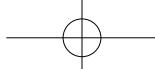
このように、サンプリングバイアスの影響はマーカーごとに異なる向き、異なる強さで出てきます。したがって、すべてのマーカーについて同じように補正をしてしまうのは乱暴です。もしも、マーカーごとに「適切に」補正することができれば、より的確な評価ができるでしょう。

多数のサンプルについて、多数のマーカーの値があるとします。サンプルには集団の構造化があるので、マーカーに関する個々のサンプルの値を利用して、サンプル同士の遠近関係を求めるすることができます。情報は、マーカーの数の次元だけありますが、その中から、軸の情報量に大小がつくように、直交座標を選び直すことができます。このようにして、新たに選んだ座標系では、主要な軸にサンプルのばらつき具合が集約されていますから、主要な軸のみを使ってサンプルの位置を定めることにします。

少数の軸で近似していると考えればよいです。このようにすると、サンプルが空間に配置され、マーカーはこの空間に濃淡のパターンを描きます。空間上である特定の領域に配置されたサンプルが、あるマーカーの特定のアレルばかりを持てば、そのマーカーは空間に濃淡を作ります。逆に、空間に均等に分布するようなマーカーもあるでしょう。また、マーカーの濃淡は、特定の軸に沿っていることもあるでしょうし、そうでないこともあるでしょう。いずれにしろ、サンプルの位置に応じて、個々のマーカーの値として、とりやすい値とりにくい値ができます。

フェノタイプの方も、空間内に値の分布ができますから、個人の位置によって、とりやすい値とりにくい値ができます。ジェノタイプもフェノタイプも位置による値のとりやすさ・とりにくさで補正をした後に、ジェノタイプとフェノタイプの検定をします。

このプロセスをRで見てみることにします。R17-6.Rのソースを追ってくだ



さい。

まず構造化のある集団を作ります。

#### R ソース 17.6 R17-6.R : 構造化集団からのデータ作成

```
1 # 構造化集団をシミュレート
2 Nm<-1000 ;Npop<-4; # マーカー数 亜集団数
3 Ns<-c(100,200,200,200) # 集団別人数
4 M<-NULL # 全ジエノタイプデータを納める行列
5 # 亜集団別にアレル頻度・HWE-f を振ってジエノタイプのシミュレーション
6 for(j in 1:Npop){
7   tmpM<-matrix(rep(0,Nm*Ns[j]),nrow=Nm)
8   for(i in 1:Nm){
9     af<-runif(1)*0.8+0.1; f<-rnorm(1, sd=0.01);
10    if(abs(f)>1) f=0
11    df<-c(af^2,2*af*(1-af),(1-af)^2)
12    df[1]<-df[1]+f/2*df[2]; df[3]<-df[3]+f/2*df[2];
13    df[2]<-1-df[1]-df[3]
14    tmpM[i,]<-sample(c(0,1,2),Ns[j],replace=TRUE,prob=df)
15  }
16  # 全データ行列に格納
17  M<-cbind(M,tmpM)
18 }
```

次にこのデータから、**主成分分析**によって、主要な軸を取り出します。処理を追ってください (R17-7.R)。

#### R ソース 17.7 R17-7.R : 主成分分析

```
1 ##PCA##
2 mu<-apply(M,1,mean) # マーカー別平均
3 M<-M-mu # マーカー平均で標準化
4 M<-M/sqrt(mu/2*(1-mu/2)) # 分散で標準化
5 X<-1/Nm*t(M) %*%M # 個人間の分散・共分散行列
6 eiout<-eigen(X) # 固有値分解
7 plot(eiout$values) # 固有値をプロット
```

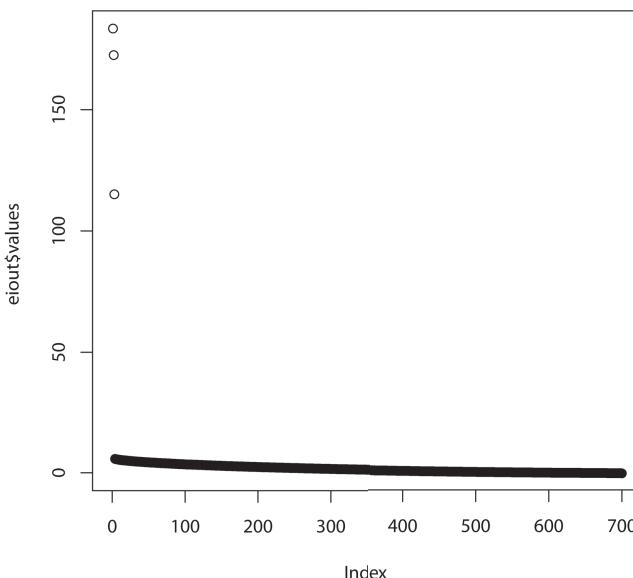


図 17.7 固有値をプロット：上位の 3 軸が重要であることが示されている

大きな固有値を持つのは 3 個で、残りは小さく、また大差がありません（図 17.7）。したがって、取り出された意味のある固有値の数は 3 個とわかります。この数は、シミュレーションで作った亜集団の数 - 1 です。

第 1 軸から第 5 軸までの軸を縦横の軸にとって、サンプルの分離の様子をプロットしてみます。確かに、第 1 軸から第 3 軸までは分離していますが、第 4 軸、第 5 軸は意味を持っていない様子が見てとれます。したがって、この 3 つの主要な軸を用いて補正をするのが適切であることもわかります（図 17.8）。

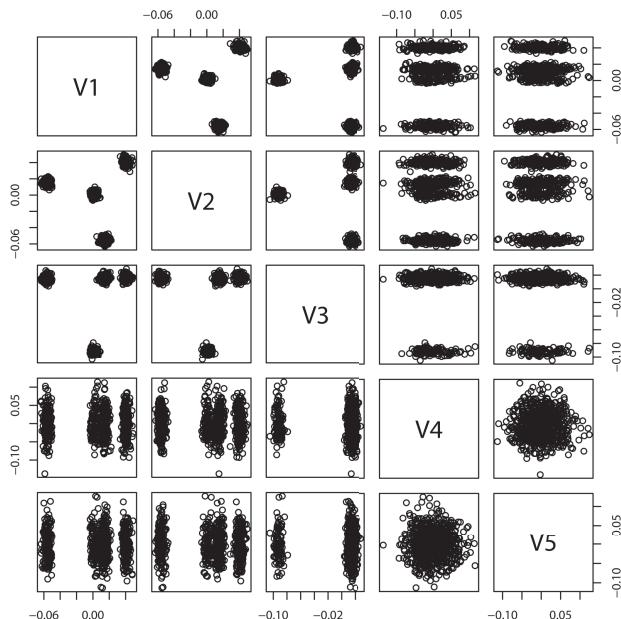
処理は R17-8.R の通りです。

#### R ソース 17.8 R17-8.R : 主成分軸でのプロット

```

1 # 意味のある固有値の数は、亜集団の数 - 1
2 # plot してみる
3 eivect<-as.data.frame(eiout$eigenvectors)
4 eilist<-1:(Npop+1)
5 plot(eivect[,eilist])

```



第1軸から第3軸はサンプルをクラスタリングするが、第4軸、第5軸はクラスタリングしていません。

図 17.8 第1軸から第5軸での散布図

次に、ケースとコントロールを構造化亜集団から異なる比率でサンプリングします (R17-9.R)。

#### R ソース 17.9 R17-9.R : 構造化サンプリング

```

1  # 偏らせて形質を与える
2  phenotype<-c(sample(c(0,1),sum(Ns)/2,replace=TRUE,
prob=c(0.45,0.55)),sample(c(0,1),sum(Ns)/2,
replace=TRUE,prob=c(0.55,0.45)))

```

構造化した集団のジェノタイプと、そこからアンバランスにサンプリングしたケースとコントロールのサンプルについて、フェノタイプとジェノタイプの関連を検定します。また、主成分分析を実行して、それによってフェノタイプとジェノタイプの補正をした上で、フェノタイプとジェノタイプの関連を検定します。

R17-10.R の 8 行目、12 行目で補正しています。ソースを追ってください。

## R ソース 17.10 R17-10.R : ジェノタイプ・フェノタイプの補正

```

1  # 検定統計量と p 値の格納用
2  Chisq<-rep(0,Nm) ; CorrChisq<-rep(0,Nm) ; Ps<-rep(0,Nm) ;
   CorrPs<-rep(0,Nm)
3  L<-3 # 考慮する軸数
4  Emat<-eiout$vectors[,1:L]
   # 考慮する軸数分を eigenvector 行列から抜き出し
5  Esqs<-apply(Emat*Emat,2,sum)
   # 抜けデータがなければ、これはすべて値が 1
6  phenotype<-phenotype-mean(phenotype) # phenotype の平均値補正
7  Gamma<-apply(Emat*phenotype,2,sum)/Esqs # 補正項
8  corrphenotype<-phenotype-Emat%*%Gamma # 補正形質
9  for(i in 1:Nm){ # マーカーごとにループ
10  genotype<-M[i,]
11  Gamma<-apply(Emat*genotype,2,sum)/Esqs # マーカーごとの補正項
12  corrgenotype<-genotype-Emat%*%Gamma
   # マーカー別の補正ジェノタイプ
13  # 補正なしのトレンド検定統計量は、ジェノタイプ - フェノタイプの
   # cor^2 のサンプル数倍
14  Chisq[i]<-(sum(Ns))*cor(genotype,phenotype)^2
15  # 補正ありのそれは、補正後ジェノタイプ - 補正後フェノタイプの
   # cor^2 の (サンプル-亜集団数) 倍
16  CorrChisq[i]<-(sum(Ns)-(L+1))*cor(corrgenotype,
   corrphenotype)^2
17  }
18  # 自由度 1 で p 値算出
19  Ps<-pchisq(Chisq,1,lower.tail=FALSE)
20  CorrPs<-pchisq(CorrChisq,1,lower.tail=FALSE)
21  # 補正前・後 p 値昇順 plot
22  ylim<-c(0,1)
23  plot(ppoints(length(Ps),a=0),sort(Ps),ylim=ylim)
24  par(new=T)
25  plot(ppoints(length(Ps),a=0),sort(CorrPs),ylim=ylim,
   col="red")
26  # 補正前・後 p 値散布図
27  plot(Ps, CorrPs)
28  # ジェノミックコントロール法との比較
29  chivalue<- qchisq(Ps,1,lower.tail=FALSE)
30  lambda<-quantile(chivalue,0.5)/qchisq(0.5,1)
31  chivalueGC<-chivalue/lambda
32  pGC<-pchisq(chivalueGC,1,lower.tail=FALSE)
33  plot(Ps,pGC)

```

補正前の  $p$  値と補正後の  $p$  値の昇順プロット（図 17.9）を見ると、補正によって  $p$  値が均一分布（対角線）に近づいていることがわかります。

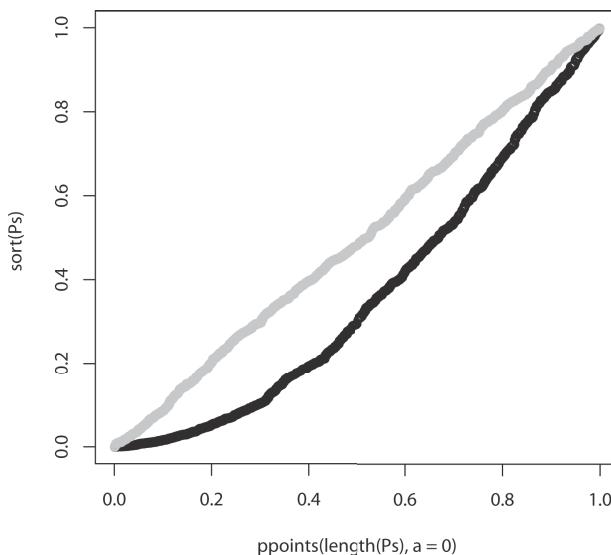


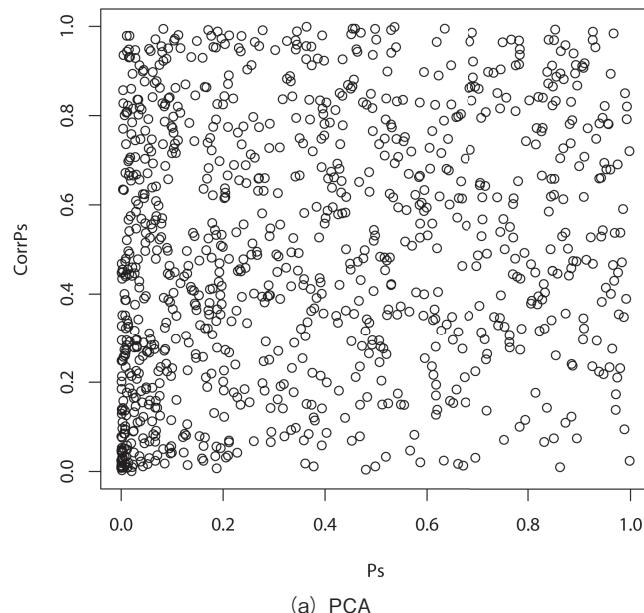
図 17.9 補正前の  $p$  値（黒色）と主成分分析による補正後の  $p$  値（灰色）の昇順プロット

また、補正前の  $p$  値と補正後の  $p$  値との関係（図 17.10 (a)）を見ると、補正によって  $p$  値が大きくなるもの ( $y=x$  の対角線の上側) と、逆に小さくなるもの ( $y=x$  の対角線の下側) とがあることがわかります。

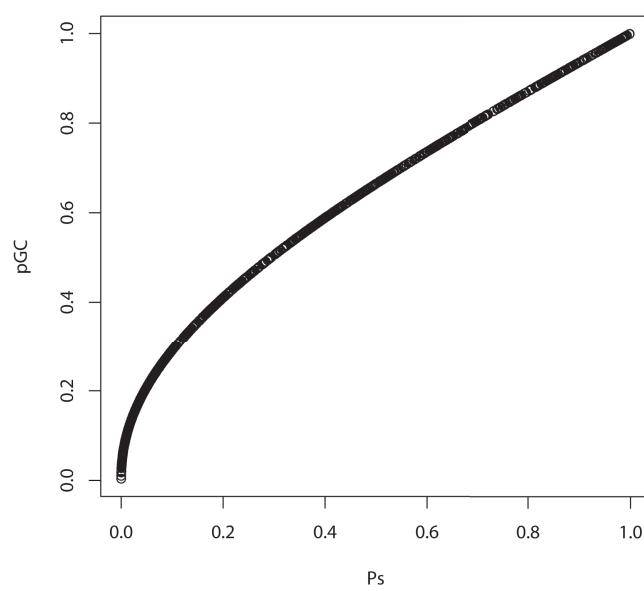
ジェノミックコントロールによって補正した場合（図 17.10 (b)）は、すべて対角線の上側にきているのと対照的です。

主成分分析による方法ではマーカーごとに補正をするのに対し、ジェノミックコントロール法ではすべてのマーカーを一律に補正することが、この散布図の違いを作っています。

このように、多くのデータがあるときには、サンプル間の依存性を抽出して、補正に利用できることがわかりました。

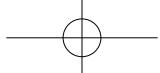


(a) PCA



(b) ジェノミックコントロール法

図 17.10 補正前後の  $p$  値の散布図



## ● 17.3.2 帰無仮説が必ずしも棄却されるべきではないとき

### (1) False Discovery Rate (FDR)

多くの帰無仮説棄却検定を行うにあたり、 $p$  値が一様分布とならないときの補正について述べましたが、検定仮説の少なからぬ部分が帰無仮説を棄却するような場合はどのように考えればよいでしょうか？ 大規模な遺伝子発現データの場合などは、関連があって当然な因子を数多く取り扱います。そのようなときの対処法として、

「そもそも、検定のうちのある割合は帰無仮説を棄却するものとする」

という発想をします。雑な考え方をすれば、全部で  $N$  個の仮説を検定するときに、陽性となるべき仮設の割合が  $a$  としたら、 $p$  値が小さい方から  $N \times p$  個の仮説が帰無仮説を棄却したとみなす、というのはどうでしょうか。

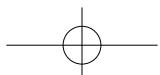
これでは、 $p$  値が小さい方の仮説の  $p$  値の大小に関わらず、採択されてしまうので、さすがに乱暴に過ぎるでしょう。では、 $N$  個の  $p$  値が  $\frac{i}{N+1}$  ;  $i = 1, 2, \dots, N$  というように、多重検定における、 $p$  値の期待値通りに出でいたらどうでしょう（図 17.11）。この場合も、すべてが帰無仮説を満足しているときに、「普通に出るべき  $p$  値」なわけですから、このようなときに上位  $a$  の仮説を棄却するのは、問題があるでしょう。

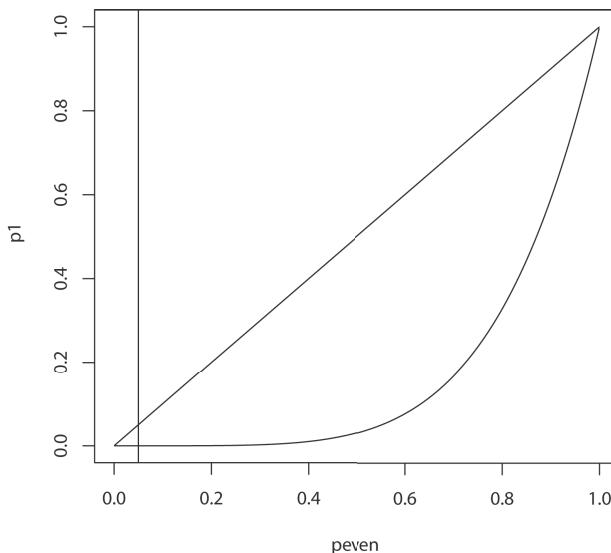
R17-11.R を用いて、人工的な  $p$  値分布を作ります（図 17.11）。

#### R ソース 17.11 R17-11.R : $p$ 値リスト

```
1 # 人工的な p 値分布を作ります
2 N<-1000
3 peven<-p1<-ppoints(N,a=0) # N 個の p 値の期待値のベクトル
4 p1<-peven^5 # 期待値から人工的に逸脱させる
5 plot(peven, peven, xlim=c(0,1), ylim=c(0,1), type="l")
6 par(new=T)
7 plot(peven,p1,xlim=c(0,1),ylim=c(0,1),type="l")
8 abline(v=0.05) # 棄却水準
```

すべての仮説が帰無仮説を満足するときに期待される  $p$  値の線よりも、小さめの方に  $p$  値のプロットがあるときには、そのうちの  $p$  値の小さい方から採択することは妥当のように思えます。問題は、どこまでを採択するかの線引





対角線を描いているのが、すべて帰無仮説が真である場合です。下に凸の場合は、名目上の  $p$  値が小さめに出ています（対立仮説が真である検定が多い）。垂直線が棄却水準 0.05。垂直線より左側を有意とすれば、全体の 0.05 の検定を有意とすることになります。

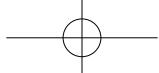
図 17.11 1,000 個の  $p$  値の昇順プロット

きのルールをどうやって作るかです。今、考えたいのは「 $a$  の割合の仮説が採択される」ことが、標準的であるように工夫したルールです。

**False Discovery Rate (FDR)** と呼ばれる手法がこれに相当します。工夫にはいろいろやり方がありますが、その中で最も単純な **Benjamini and Hochberg 法** と呼ばれるものを見てみることにします。

まず初めに、次のように考えてみます。「可能性は極めて低いですが、もしかしたら、すべての仮説が帰無仮説を棄却するかもしれない」です。そのようなときには、最も大きい  $p$  値を返した仮説も帰無仮説を棄却するはずです。今、棄却水準を  $b$  としますと、最も大きい  $p$  値が  $b$  よりも小さかったら、その仮説も含めて、すべての仮説が帰無仮説を棄却したと考えることにしましょう。最も大きい  $p$  値だけではなく、それ以外の  $p$  値にも採択するべきかどうかの基準を定めることにします。第  $i$  番目に大きい  $p$  値に関しては、

$$b \times \frac{N - i + 1}{N}$$



の値よりも小さかったら採択することにすると、最も大きい  $p$  値の場合には、

$$b \times \frac{N-1+1}{N} = b$$

となって、この方法でうまくいきます。

図にしてみます。図 17.11 で下に凸になっている  $p$  値を大きい方から並べます。それをプロットしたのが図 17.12 です。下に凸です。図 17.12 には、下に凸な曲線が 2 本引かれています。下側の線が補正前の  $p$  値のプロットです。そこに、

$$y = b \frac{N-i+1}{N} \quad (\text{たとえば, } b=0.05)$$

の斜めの線を引きます。この線より下回った検定が採択されます。

R では、`p.adjust()` という関数があります。この関数は、検定結果の  $p$  値の列を補正して、それぞれの  $p$  値が直線  $y = b \frac{N-i+1}{N}$  と交わるとしたら、 $b$  の値がいくつのときかを算出して、それを補正後の  $p$  値として返します。したがって、 $b=0.05$  のときには、補正後の  $p$  値が 0.05 より小さければ、FDR 法で採択されたことになります。図 17.12 では、`p.adjust()` 関数の補正後の  $p$  値もプロットしてあります。そのラインが、下に凸な 2 本の線のうちの上側の線です。観察  $p$  値のラインが直線  $y = 0.05 \times \frac{N-i+1}{N}$  と交わった点を通る垂直線が引いてあり、その垂直線と補正後の  $p$  値のプロットの交点の  $y$  軸値が 0.05 となっていることを示しています。

説明が少しややこしいですが、R17-12.R を頼りに実行してみてください。

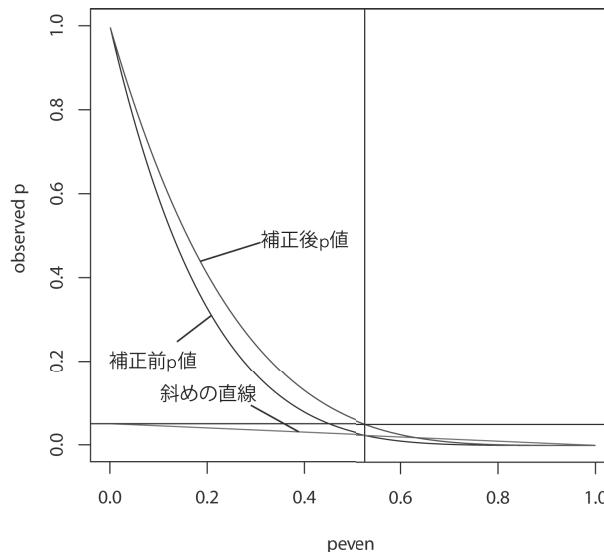
#### R ソース 17.12 R17-12.R : FDR

```
1 N<-1000
2 peven<-p1<-ppoints(N,a=0) # N 個の p 値の期待値のベクトル
3 peven<-ppoints(length(p),a=0)
4 p1<-peven^5 # 期待値から人工的に逸脱させる
5 pfdr<-p.adjust(p1,"fdr")# FDR
6 b<-0.05 # 5%で帰無仮説が成立すると仮定する
7 # 名目 p 値
8 plot(peven,sort(p1,decreasing=TRUE),xlim=c(0,1),
```

```

9  ylim=c(0,1),type="l")
10 par(new=T)
11 # BH 法の補助直線
12 plot(peven,b*((length(p1)-(1:length(p1))+1)/length(p1)),
13       xlim=c(0,1),ylim=c(0,1),type="l",col="red")
14 # 補正後 p 値
15 par(new=T)
16 plot(peven,sort(pfdr,decreasing=TRUE),xlim=c(0,1),
17       ylim=c(0,1),type="l",col="blue")
18 abline(h=b) # y=b の水平線
19 abline(v=peven[length(which(pfdr>b))])
20 # 名目 p 値の線と補助直線の交点を通る垂線

```

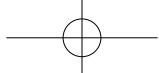


下に凸な曲線が 2 本あります。下側の曲線は観測された  $p$  値を降順にプロットしたものです。  
上側の曲線は FDR 補正後の  $p$  値。 $y = 0.05 \times \frac{N-i+1}{N}$  の斜めの直線 ( $y$  切片が 0.05) と観測  $p$  値曲線との交点を通る垂直線が引かれています。その垂直線と  $y = 0.05$  との交点を補正後  $p$  値曲線が通ります。

図 17.12 FDR

## (2) ベイズ因子

FDR では、仮説のうちのある割合が採択されるべきであるというところから出発して、採択基準を定めました。



別の方法もあります。今、ゲノム上の多数のマーカーとある形質との関係を調べているものとします。ジェノミックコントロールのときに仮定した通り、ほとんどすべてのマーカーは形質と無関係かもしれません。しかしながら、ゲノム上には非常に多くの要素があり、それらは、大きな力、小さな力、いろいろな力を持っていて、ごく微細な力も含めれば、ゲノム上のほぼすべての場所が形質となんらかの関連があると考えることも、悪くないかもしれません。

実際、ジェノミックコントロールでは、数多くのマーカーがその程度はまちまちだけれども、亜集団ごとにアレル頻度が違っていることを仮定しました。そのときの仮定として、アレル頻度の違いは平均を0とする正規分布に従うものとしました。配列多様性も大部分は微弱な関連を持ち、ごくわずかに目立った関連があるとすることもあります。このように、帰無ではない関連の分布を仮定することは、**事前分布**をある意味で勝手に仮定しているわけで、ベイズの考え方です。

このように大多数のマーカーに帰無ではない仮説を仮定してデータを解釈する手法として、**ベイズ因子**があります。本書では用語を挙げるに留めます。



## 17.4 複数の結果を合わせる——メタアナリシス

### ◎ 17.4.1 相互に独立な検定を合わせる

$2 \times 2$ 分割表による独立性の検定が2つあるときのことを考えます。この2つの検定は、ある1つのこと（因子A/aと因子B/bの関係）を調べるもので、全く独立にサンプルを集めて検定するものとします。このように、あることに2つのサンプルセットを使って調べるときに、その2セットの結果をどうやって統合するかの話です。

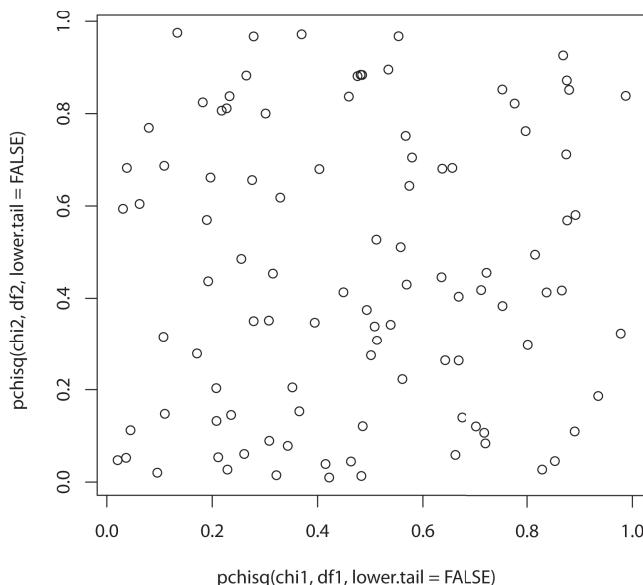
この2つの検定をサンプルを取り直して何度も繰り返すことを考えます。それぞれの帰無仮説のもとでのピアソンの独立性検定のカイ自乗値は、自由度1のカイ自乗分布に従いますので、2つの検定のピアソン検定のカイ自乗検定の  $p$  値の散布図を描くと、図 17.13 のようになります (R17-13.R)。

## R ソース 17.13 R17-13.R : 独立な 2 検定

```

1 Nt<-10000
2 df1<-1
3 df2<-2
4 chi1<-rchisq(Nt,df1)
5 chi2<-rchisq(Nt,df2)
6 plot(pchisq(chi1,df1,lower.tail=FALSE),
      pchisq(chi2,df2,lower.tail=FALSE))
7 chisum<-chi1+chi2 # カイ自乗値の和
8 plot(ppoints(Nt,a=0),sort(pchisq(chisum,df=df1+df2,
      lower.tail=FALSE)),type="l") # 自由度の和を自由度として p 値化

```

図 17.13 互いに独立な 2 つの検定を行ったときの  $p$  値の散布図

2 つの検定からのカイ自乗値の和は、2 つの検定の自由度の和である 2 を自由度としたカイ自乗分布に従うことは、シミュレーションデータの  $p$  値が均一分布となっていることで示せます（R17-13.R、図 17.14）。

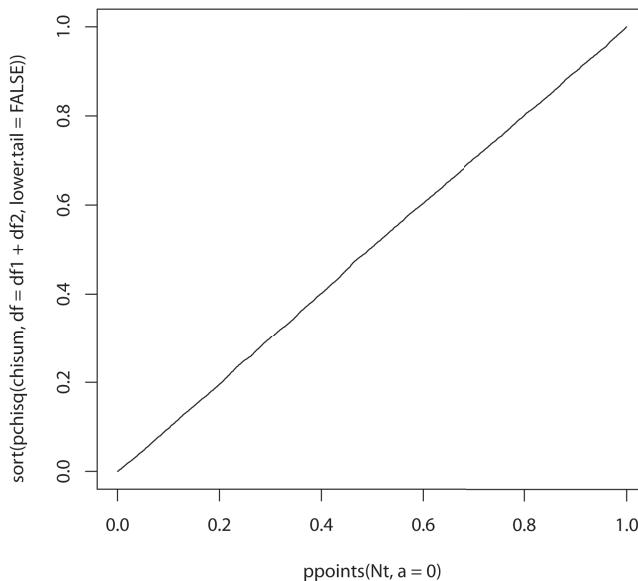


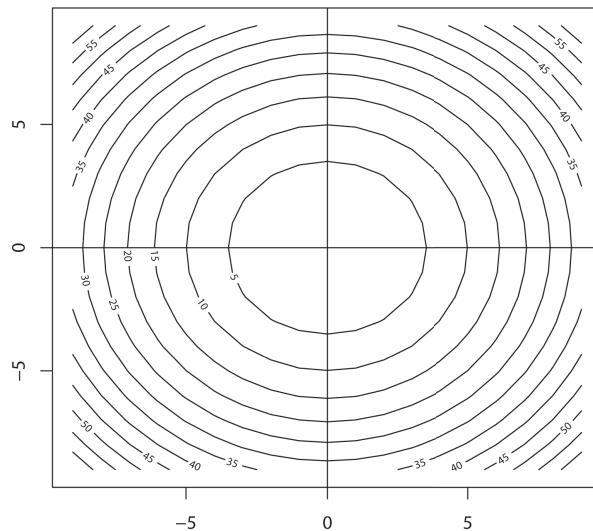
図 17.14 2つの相互に無関係なピアソンの方法のカイ自乗値を足した値を、もとの2つのピアソンの方法の自由度の和を自由度とするカイ自乗分布に照らして  $p$  値としたときの  $p$  値の昇順プロット：均一分布になっている

実際、検定の数は2に限らず、個々の検定の自由度が任意であってもこの関係は成り立ちます。相互に独立な検定を複数行うこととは、個々の検定の自由度の和の検定を1回行うことであることも、同様に示せます。このことを、カイ自乗分布が再帰性を持つと言います。確率変数  $X$  と  $Y$  がそれぞれ自由度  $dx$ ,  $dy$  のカイ自乗分布に従い、相互に独立であるとき、 $X+Y$  は自由度  $dx+dy$  のカイ自乗分布に従うという性質のことです。

このことを示すソースは、ダウンロードできるソースの R17-sup3.R です。

2つの  $2 \times 2$  表検定に話を戻します。2つの表があったときに、この表のそれぞれの周辺分布のもとでとりうる表を全部作成して、2次元にプロットします（図 17.15）。これまで、多くの2次元プロットが正三角形でしたが、今度は2つの表の軸を直交させます。2つの表が独立からです。横軸は1つ目の表の第1セルの値の期待値からのずれを、縦軸は2つ目の表のそれをとて描いてあります。橢円の等高線が描けていることがわかります。

2次元に橢円の等高線が描けるところは、自由度2の表（ $2 \times 3$  表）の自由



2つの $2 \times 2$ 表があるときに、第1の表の期待値表からのずれを横軸に、第2の表のそれを縦軸にとります。2表のピアソンの独立性検定カイ自乗値の和が等しいような表のペアに相当する点をつなぐと楕円になります。

図 17.15 メタアナリシス①

度2のピアソンのカイ自乗検定と同じです。 $2 \times 3$ 表で自由度2の検定をしたときには、周辺度数が許す範囲で、3つの列が完全に自由に値をとれることを前提とした検定でした。今、2つの $2 \times 2$ 表について、それぞれが自由に値をとれることを対立仮説として、2つの表の両方が帰無であるという仮説を帰無仮説とすると、自由度2の検定となるわけです。

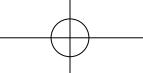
ソースは R17-14.R の通りです。

#### R ソース 17.14 R17-14.R : メタアナリシス①

```

1  # 2表とその周辺度数・期待値テーブルを作る
2  t1<-t2<-matrix(c(10,10,10,10),nrow=2,byrow=TRUE)
3  m11<-m21<-apply(t1,1,sum);m12<-m22<-apply(t1,2,sum);
   M1<-M2<-sum(t1)
4  e1<-m11%*%t(m12)/M1;e2<-m21%*%t(m22)/M2
5  # とりうるテーブルをx11,y11を変数として網羅する
6  x11<-seq(from=-M1,to=M1,by=1);x12<-x11+m11[1];
   x21<-x11+m12[1];x22<-x12+m12[2]
7  xbind<-cbind(x11,x12,x21,x22);

```



```
okx<-which(apply(xbind,1,min)>0)
8 x11<-x11[okx];x12<-x12[okx];x21<-x21[okx];x22<-x22[okx]
9 y11<-seq(from=-M2,to=M2,by=1);y12<-y11+m21[1];
y21<-y11+m22[1];y22<-y12+m22[2];
ybind<-cbind(y11,y12,y21,y22);
oky<-which(apply(ybind,1,min)>0)
10 y11<-oky;y12<-oky;y21<-oky;y22<-oky
11 # カイ自乗を計算する
12 chi1<-(x11-e1[1,1])^2/e1[1,1]+(x12-e1[1,2])^2/e1[1,2] +
(x21-e1[2,1])^2/e1[2,1]+(x22-e1[2,2])^2/e1[2,2]
13 chi2<-(y11-e2[1,1])^2/e2[1,1]+(y12-e2[1,2])^2/e2[1,2] +
(y21-e2[2,1])^2/e2[2,1]+(y22-e2[2,2])^2/e2[2,2]
14 # 2つのカイ自乗値を足し合わせる
15 z<-outer(chi1,chi2,FUN="+")
16 xlim<-ylim<-c(min(x11-e1[1,1],y11-e2[1,1]),max(x11-e1[1,1],
y11-e2[1,1]))
17 contour(x11-e1[1,1],y11-e2[1,1],z,xlim=xlim,ylim=ylim)
18 abline(h=0);abline(v=0) # 補助線
```

## ● 17.4.2 2表を単純に足し合わせる

では、本節の本題であるメタアナリシスに話を移します。2つの $2 \times 2$ 表分割表検定を統合しようとしているとします。どちらの表も因子A/aと因子B/bとの間の関連について調べているものとします。

このメタアナリシスでは、帰無仮説は、2つの検定のそれぞれで帰無仮説が成り立つ、というものです。そして、対立仮説は、2つの検定に共通した何かしらの仮説です。

図17.15を見てみます。この図は、2つの表のどちらもが期待値をとる場合が原点になっています。そして、第1象限（右上の4半分）は2つの検定の両方で、(1, 1)のセルが期待値よりも大きくなっている領域です。第3象限（左下の4半分）は逆に2つの検定の両方で、(1, 1)のセルが期待値よりも小さくなっている領域です。残りの2つの象限は、(1, 1)のセルが片方の検定では大きくなり、もう片方の検定では小さくなるような場合に対応した領域です。

4つの象限において、2つの検定がどの因子の結びつき（「A=B」は「AとBとが正の関連を持つ」を表すものとします）を指示しているかを表にすれば、次のようになります。

	第1検定	第2検定
第1象限	A=B,a=b	A=B,a=b
第2象限	A=b,a=B	A=B,a=b
第3象限	A=b,a=B	A=b,a=B
第4象限	A=B,a=b	A=b,a=B

したがって、2つの検定が、A/a, B/b の関係において同様の傾向を示しているかどうかで言えば、第1、第3象限はそう言えますが、第2、第4象限は反対の傾向を示していることになります。

さて、次のように考えてみましょう。

実は、これは1つのスタディであったとします。たまたま、だれかがデータの一部を集計し、その残りを別に集計したものだったのです。この場合は、もともと1つのスタディであったわけですから、単純に合わせて1つのスタディと考えるのがよいです。

そうしてやると、2つのスタディの和に対して、カイ自乗値が1つ得られます。それを描いてみます（R17-16.R、図17.16）。

このようにしてやると、対立仮説は、2つの表を足し合せた表から得られる、群別・因子別の頻度がこの合算表の最尤推定値になりますから、それを仮説とした検定になります。合算する前には、2つの検定のそれぞれに自由な比率を仮説として検討できたのに、自由な程度が減少してしまったわけです。そうすると、 $2 \times 3$ 表に対して、何かしらの遺伝形式をモデルとして適用したときと同様に、2次元空間に直線状の等高線が現れました。そして、その直線状の等高線は、自由度2の楕円の接線となります。その接線は傾きが-1です。これは、第1の表の(1, 1)のセルの期待値からのずれと、第2の表の(1, 1)のセルの期待値からのずれを足した値が同じであるときには、メタアナリシスの結果も同じになることを意味します。

そして、先ほど、2つの表が共通の対立仮説を持つならば、第1、第3象限がそれを支持し、第2、第4象限はそれを否定することを確認しました。確かに、この合算表による等高線は、共通仮説を支持する部分で大きなカイ自乗値を持ち、共通仮説を否定する領域では中心から遠くなても（個々の検定でカイ自乗値が大きくなっても）、統合した上での判定は、カイ自乗値が増えません（図17.16）。

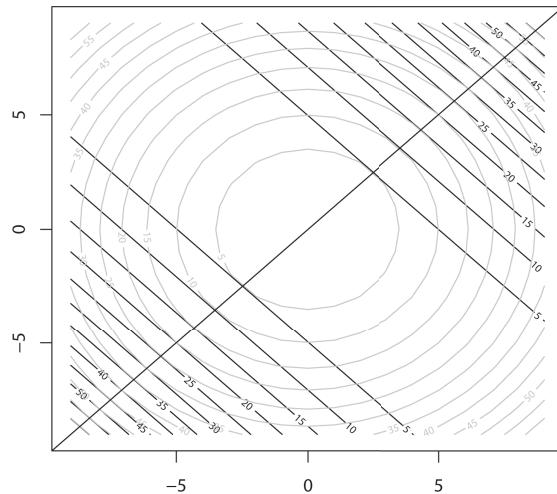
R の処理を示します (R17-15.R)。

#### R ソース 17.15 R17-15.R : 2表を足し合わせる

```

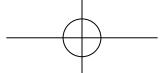
1  # 格子状に作った2表のセルの値を足し合わせる
2  sum11<-outer(x11,y11,FUN="+");sum12<-outer(x12,y12,FUN="+")
3  sum21<-outer(x21,y21,FUN="+");sum22<-outer(x22,y22,FUN="+")
4  sume11<-e1[1,1]+e2[1,1];sume12<-e1[1,2]+e2[1,2];
   sume21<-e1[2,1]+e2[2,1];sume22<-e1[2,2]+e2[2,2]
5  # 足し合せた表のカイ自乗値
6  sumz<-(sum11-sume11)^2/sume11+(sum12-sume12)^2/sume12+
   (sum21-sume21)^2/sume21+(sum22-sume22)^2/sume22
7  xlim<-ylim<-c(min(x11-e1[1,1],y11-e2[1,1]),max(x11-e1[1,1],
   y11-e2[1,1]))
8  contour(x11-e1[1,1],y11-e2[1,1],sumz,xlim=xlim,ylim=ylim)
   # 表の足し合せによる等高線
9  par(new=T)
10 contour(x11-e1[1,1],y11-e2[1,1],z,xlim=xlim,ylim=ylim,
   col=gray(6/8)) # 橋円等高線
11 abline(a=0,b=1)

```



2つの $2 \times 2$ 表のカイ自乗値の和が等しいような表のペアに相当する点を結ぶと、橋円になります（灰色）。1つ目の表の期待値表からのずれと2つ目の表の期待値表からのずれとの和が等しいような表のペアに相当する点を結ぶと、傾きが-1の直線になります（黒色）。

図 17.16 2表を足し合わせる

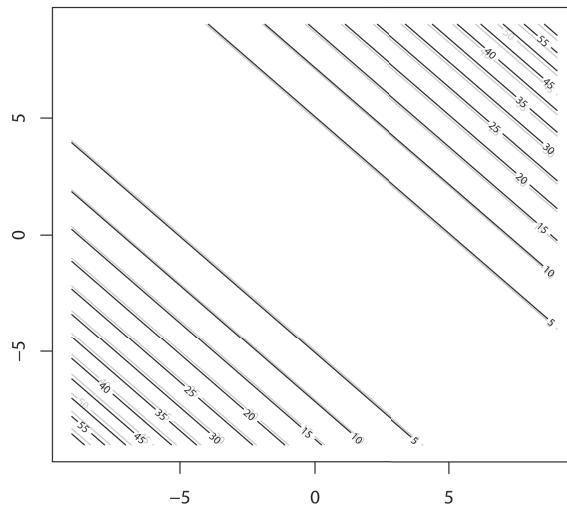


### ● 17.4.3 メタアナリシス

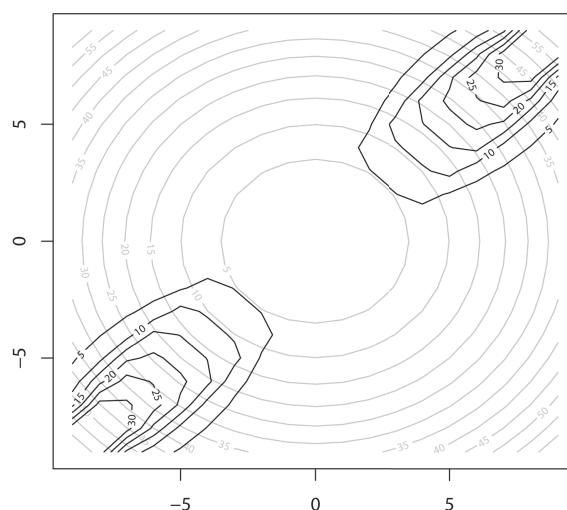
複数の $2 \times 2$ 表を統合する**メタアナリシス**の手法の等高線がどうなるかを見てみることにします。

メタアナリシスの方法には大きく2つあります。**固定効果モデル**と呼ばれるものと、**変量効果モデル**と呼ばれるものです。簡単に言うと、固定効果モデルは、統合するスタディに同じモデルが当てはまる（同じオッズ比が当てはまるなど）ことを想定した方法で、変量効果モデルは、統合するスタディのそれぞれにある程度違いのあるモデルが当てはまることを許容した上で、スタディ間に共通する傾向があるかどうかを検定する手法です。

それらについてのイメージを知るには等高線を眺める方が早いので、描いてみます。固定効果モデルの代表的な手法である**マンテル-ヘンツェル法**では、表を足し合わせたものと同じく、傾きが負の直線状の等高線になります。一方、変量効果モデルの代表的な手法である**DerSimonian-Laird法**では、統計量が大きくなる特定の尾根があって、そこから離れると急速に統計量が減少します。逆に言うと、スタディ間で同じ傾向（オッズ比が等しいなど）のとき（図17.17のような描き方のときには、「座標が「傾きが正の直線（ $y=x$ の対角線）に沿っているとき」」には、固定効果モデルも変量効果モデルも違いはないが、スタディ間の違いが大きいときには、固定効果モデルよりも変量効果モデルの方が保守的であることがわかります。



(a) 固定効果モデルで同じ検定結果となる表のペアを結んだ線（黒色）。2表を単純に足し合わせて検定したときの等高線を補助線（灰色）として引いてあります、それとほぼ同じになります。



(b) 変量効果モデルで同じ検定結果となる表のペアを結んだ線（黒色）。2表の統計量を足し合わせたときの等高線（楕円）を補助線（灰色）として引いてあります。

図 17.17 メタアナリシス②

## R ソース 17.16 R17-16.R : メタアナリシス②

```
1 library(rmeta)
2 n.case<-c(m11[1],m21[1])
3 n.ctrl<-c(m11[2],m21[2])
4 zMH<-z
5 zDSL<-z
6 for(i in 1:length(z[,1])){
7   for(j in 1:length(z[1])){
8     mhout<-meta.MH(n.case,n.ctrl,c(x11[i],y11[j]),
9                      c(m12[1]-x11[i],m22[1]-y11[j]))
10    zMH[i,j]<-mhout$MHtest[1]
11    dslout<-meta.DSL(n.case,n.ctrl,c(x11[i],y11[j]),
12                      c(m12[1]-x11[i],m22[1]-y11[j]))
13    zDSL[i,j]<-dslout$test[1]^2
14  }
15  zlim<-c(0,max(sumz))
16  contour(x11-e1[1,1],y11-e2[1,1],sumz,xlim=xlim,ylim=ylim,
17            zlim=zlim,nlevels=10,col=gray(6/8))
18  par(new=T)
19  contour(x11-e1[1,1],y11-e2[1,1],zMH,xlim=xlim,ylim=ylim,
20            zlim=zlim,nlevels=10)
21  #####
22  contour(x11-e1[1,1],y11-e2[1,1],z,xlim=xlim,ylim=ylim,
23            zlim=zlim,,nlevels=10,col=gray(6/8))
24  par(new=T)
25  contour(x11-e1[1,1],y11-e2[1,1],zDSL,xlim=xlim,ylim=ylim,
26            zlim=zlim,nlevels=10)
27  #####
28 library(rmeta)
29
30 t1<-matrix(c(10,20,30,40),nrow=2,byrow=TRUE)
31 t2<-matrix(c(200,300,50,40),nrow=2,byrow=TRUE)
32 m11<-apply(t1,1,sum)
33 m12<-apply(t1,2,sum)
34 M1<-sum(t1)
```

```
35 m21<-apply(t2,1,sum)
36 m22<-apply(t2,2,sum)
37 M2<-sum(t2)
38
39
40 m11<-apply(t1,1,sum)
41 m12<-apply(t1,2,sum)
42 M1<-sum(t1)
43
44 m21<-apply(t2,1,sum)
45 m22<-apply(t2,2,sum)
46 M2<-sum(t2)
47
48 e1<-m11%*%t(m12)/M1
49 e2<-m21%*%t(m22)/M2
50
51 x11<-seq(from=-M1,to=M1,by=1)
52
53 y11<-seq(from=-M2,to=M2,by=1)
54
55 x12<-x11+m11[1]
56 x21<-x11+m12[1]
57 x22<-x12+m12[2]
58
59 xbind<-cbind(x11,x12,x21,x22)
60 okx<-which(apply(xbind,1,min)>0)
61
62 x11<-x11[okx]
63 x12<-x12[okx]
64 x21<-x21[okx]
65 x22<-x22[okx]
66
67 y12<-y11+m21[1]
68 y21<-y11+m22[1]
69 y22<-y12+m22[2]
70
71 ybind<-cbind(y11,y12,y21,y22)
72 oky<-which(apply(ybind,1,min)>0)
73
74 y11<-y11[oky]
75 y12<-y12[oky]
76 y21<-y21[oky]
```

```

77 y22<-y22[oky]
78
79
80 # 格子状に作った2表のセルの値を足し併せる
81 sum11<-outer(x11,y11,FUN="+");
82 sum12<-outer(x12,y12,FUN="+")
83 sum21<-outer(x21,y21,FUN="+");
84 sum22<-outer(x22,y22,FUN="+")
85 sume11<-e1[1,1]+e2[1,1];sume12<-e1[1,2]+e2[1,2];
86 sume21<-e1[2,1]+e2[2,1];sume22<-e1[2,2]+e2[2,2]
87 # 足し併せた表のカイ自乗値
88 sumz<-(sum11-sume11)^2/sume11+(sum12-sume12)^2/sume12+
89   (sum21-sume21)^2/sume21+(sum22-sume22)^2/sume22
90 xlim<-ylim<-c(min(x11-e1[1,1],y11-e2[1,1]),
91   max(x11-e1[1,1],y11-e2[1,1]))
92
93 # chi1+chi2
94 # カイ自乗を計算する
95 chi1<-(x11-e1[1,1])^2/e1[1,1]+(x12-e1[1,2])^2/e1[1,2]+
96   (x21-e1[2,1])^2/e1[2,1]+(x22-e1[2,2])^2/e1[2,2]
97 chi2<-(y11-e2[1,1])^2/e2[1,1]+(y12-e2[1,2])^2/e2[1,2]+
98   (y21-e2[2,1])^2/e2[2,1]+(y22-e2[2,2])^2/e2[2,2]
99 # 2つのカイ自乗値を足し合わせる
100 z<-outer(chi1,chi2,FUN="+")
101
102 n.case<-c(m11[1],m21[1])
103 n.ctrl<-c(m11[2],m21[2])
104 zMH<-zDSL<-z
105 for(i in 1:length(z[,1])){
106   for(j in 1:length(z[1,])){
107     mhout<-meta.MH(n.case,n.ctrl,c(x11[i],y11[j]),
108       c(m12[1]-x11[i],m22[1]-y11[j]))
109     zMH[i,j]<-mhout$MHtest[1]
110     dslout<-meta.DSL(n.case,n.ctrl,c(x11[i],y11[j]),
111       c(m12[1]-x11[i],m22[1]-y11[j]))
112     zDSL[i,j]<-dslout$test[1]^2
113   }
114 }
115 zlim<-c(0,max(zMH,zDSL))
116 contour(x11-e1[1,1],y11-e2[1,1],sumz,xlim=xlim,ylim=ylim,
117   zlim=zlim,nlevels=10,col=gray(6/8))
118 par(new=T)

```

```

109 contour(x11-e1[1,1],y11-e2[1,1],zMH,xlim=xlim,ylim=ylim,
           zlim=zlim,nlevels=10)
110
111 contour(x11-e1[1,1],y11-e2[1,1],z,xlim=xlim,ylim=ylim,
           zlim=zlim,,nlevels=10,col=gray(6/8))
112 par(new=T)
113 contour(x11-e1[1,1],y11-e2[1,1],zDSL,xlim=xlim,ylim=ylim,
           zlim=zlim,nlevels=10)

```

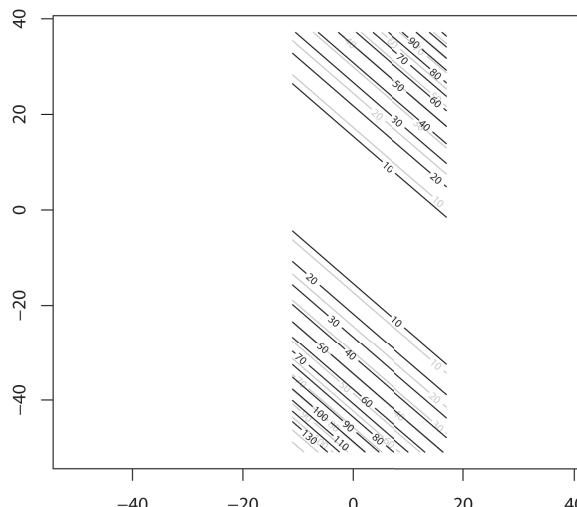
ここまででは、対称的な表でしかも2個の表が同一という、極めて均整のとれた場合で図を描きました。今度は個々の表の対称性を次のRソース（R17-16.Rから抜粋）が示すように崩してみます。

```

30 t1<-matrix(c(10,20,30,40),nrow=2,byrow=TRUE)
31 t2<-matrix(c(200,300,50,40),nrow=2,byrow=TRUE)

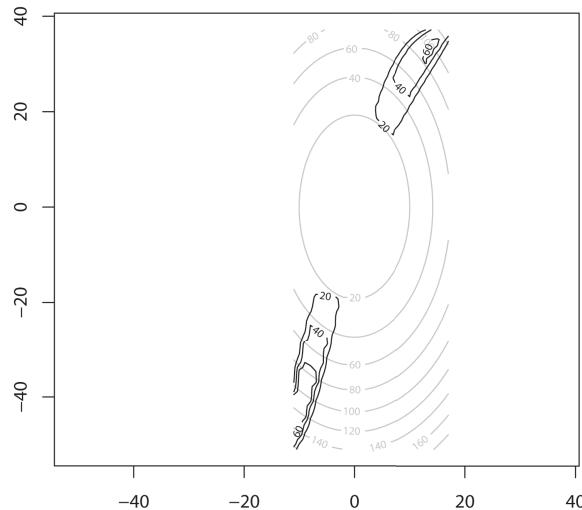
```

このようにすると楕円がゆがんだり、楕円と直線との相対的位置関係が均整のとれていた場合と均整が崩れた場合とで異なる点などがありますが、基本的には、2次元平面にどのような等高線を引くかによる違いであることがわかります。



(a) 固定効果モデルでの等高線：表を足し合わせて検定したときの補助線（黒）とそれが生じていますが、等高線の傾きは同じです。

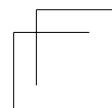
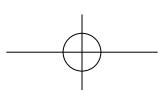
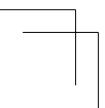
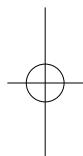
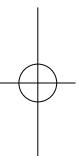
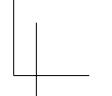
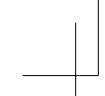
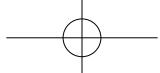
図 17.18 メタアナリシス③

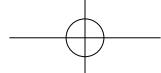


(b) 変量効果モデルでの等高線：等高線が作る尾根の方向や幅が図 18.17 (b) と異なっています

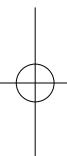
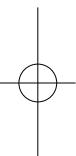
図 17.18 メタアナリシス③ (つづき)

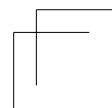
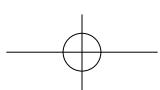
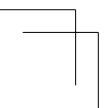
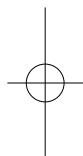
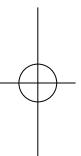
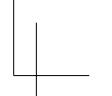
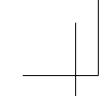
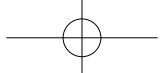
メタアナリシスは、帰無仮説を棄却するかどうかの判断と同時に、共通する仮説における「効果」の大きさの値を統合して、信頼区間を与えることも重要です。また、メタアナリシスは、そもそも統合するべきスタディをどうやってみつけるか、といった事柄も重要です。しかし、それらを含めると話が大きくなりますので、本書ではメタアナリシスを、他の章の話題との関連でイメージすることを目的として、ここで終了とします。

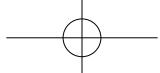




## 付 錄







# 付録 A

---

## R

### A.1 R のインストールと起動と終了

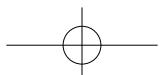
R は統計関係の処理がしやすいように作られたフリーソフトです。CRAN (The Comprehensive R Archive Network、<http://cran.r-project.org/>) がその管理サイトです。このサイトから利用環境 (Linux, Mac, Windows) に合わせて配布用バイナリをダウンロードしてインストールすることで利用が可能です。普通のフリーソフトのインストールと同じです。デフォルトのインストール後には、デスクトップに起動用のアイコンが作成されていますから、ダブルクリックして R を起動すると、初期画面が出ます。カーソルが ">" (プロンプト) の後にあります。ここに、コマンドを打ちます。

R のバージョンアップ、利用コンピュータのバージョンアップなどの影響でうまくいかないこともあるかもしれません、ウェブ上で発達しているソフトウェアなので、ウェブ検索などにより解決の糸口が見つかると思います。

#### ■ R 起動後の初期画面

```
R version ...
...
'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます。
'q()' と入力すれば R を終了します。

>
```



## A.2 R のパッケージを使う

R をインストールすると、基本的な統計解析とプログラミングが可能となります。利用者が限定されているような処理については、パッケージと呼ばれる関数群が CRAN およびそのミラーからダウンロードすることによって可能になります。

R のインターフェースのメニューバーから「パッケージ」→「CRAN ミラーサイトの設定」を選択し、CRAN ミラーサイトを選択（身近なところがよいです）して「OK」ボタンをクリックします。続けてメニューバーから「パッケージ」→「パッケージのインストール」を選択すると、パッケージ名のリストが表示されるので、目的のパッケージを選び、「OK」ボタンをクリックします。パッケージは Ctrl キーを押しながら選択すると、複数のパッケージを選択できます。そうすると、選択したパッケージがダウンロードされます。ダウンロードしたパッケージは、R をインストールしたディレクトリの "...¥R¥R-2.10.0¥library" に置かれます（2.10.0 はバージョンによって異なります）。なお、Windows Vista/7 で R を標準のディレクトリ (C:¥Program Files) にインストールすると、個人的なライブラリの場所を作りたいですか、と聞かれますので「はい」ボタンをクリックします。R をインストールしたディレクトリの "...¥R¥R-2.10.0¥library" は、R のインストール時についてきた基本的なパッケージの置き場です。

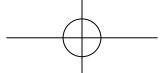
無事に済んだら、

```
> library(ape)
```

とパッケージを読み込みます。これで、"ape" パッケージの関数が使えるようになります。

本書で使用したパッケージは次の通りです。これらを上述の手順で入手してください。

```
"ape", "binom", "bnlearn", "clinfun", "evd", "gregmisc",
"gttools", "MCMCpack", "phangorn", "Rassoc", "rmeta"
```



## A.3 本書のRのソースを利用する

本書でリスト番号付きで掲載してあるRのソースをそのまま利用することができます。ソースは <http://www.genome.med.kyoto-u.ac.jp/StatGenet/lectures/2010/StatGenetTextbook/Rsrc.zip> からダウンロードすることができます。

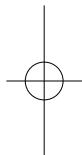
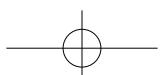
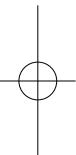
ダウンロードしたファイルはすべてのRのソースを一括して収録したファイル "StatGenetDemo.R" と、個々のソースファイル ("R1-1.R" など) に分かれています。Rを起動して、zipファイルを解凍したディレクトリを作業ディレクトリとした上で、メニューバーから「ファイル」→「Rコードのソースを読み込み」を選択すると、ファイルを選択することができるので、まず "StatGenetDemo.R" を読み込んでください。本書で利用したRソースファイルが順に実行されます。Rのコンソールに "Type <Return> to start : " という表示がなされるので、「リターン」してください。次のソースファイルが実行されます。

また、グラフの表示がなされた後に、「ページ変更を確認するまで待機」という表示で止まることがあります。この場合は、グラフをクリックしたり、「リターン」キーを押すことで次に進めます。また、次の処理に移ってよいかどうかを尋ねられることがあります。マウスをクリックするなどで、次の処理に進めます。

なお、"StatGenetDemo.R" を実行すると、インターネットに接続した環境であれば、必要なパッケージのダウンロードが行われます。

また、個々のソースはRの起動ごとに順番に実行しないと、不具合が生じる場合があることにも留意してください。

その他、1つのソースで複数の図を描いている場合には、古い図が新しい図に上書きされて、見ることができない場合もあります。そのような場合には、ソースファイルのコマンドを1行ずつ実行することが必要です。(●?以下、追加OKor不要?●) 1つの描画ウインドウに複数のグラフを重ねて表示している場合もあります。そのような場合、複数の軸ラベルが重ねて描画されているものもあります。本書に掲載のグラフは、モノクロの紙面のため色を変えた線では判別しにくい場合があることを考慮し、Rのソースでは色を変えた線を点線にして



掲載している場合などがあります。

### ◎ A.3.1 本書で用いた R の関数

以下に、本書で用いた R の関数の初出ソースを記載します。

関数	意味	ソース ファイル
%%	剰余	R4-1.R
%%*	外積	R2-5.R
:	数列	R2-4.R
abline()	図に線を加える	R9-2.R
apply()	行列の列単位・行単位の操作	R3-3.R
as.data.frame()	「データフレーム」型のオブジェクトにする	R7-5.R
as.matrix()	「行列」型のオブジェクトにする	R10-2.R
atan()	アークタンジェント	R13-5.R
barplot()	棒グラフ	R1-1.R
binom.confint()	信頼区間のいろいろ	R12-2.R
bn.fit()	bnlearn パッケージで推定結果を表示する	R14-1.R
boxplot()	箱ひげ図	R6-1.R
CATT()	傾向性検定コクラン・アーミテージ	R13-5.R
cbind()	列結合	R3-2.R
chisq.test()	カイ自乗検定	R7-1.R
choose()	組み合わせ・抜き取り	R9-1.R
cor()	相関係数行列	R5-2.R
cos()	コサイン	R10-3.R
cov()	分散共分散行列	R3-2.R
dbeta()	ベータ分布	R12-1.R
dbinom()	二項分布	R2-4.R
dchisq()	カイ自乗分布	R11-1.R
density()	密度関数作成	R6-1.R
dist()	距離	R4-2.R
dpoly()	多項分布	R13-1.R
ecdf()	経験分布	R6-1.R
eigen()	主成分分析	R17-7.R
elseif()	条件判断 2	R2-8.R
empty.graph()	空のグラフオブジェクトを作る	R14-1.R
exp()	指数関数	R2-8.R
expand.grid()	2 次元格子座標	R13-1.R

関数	意味	ソース ファイル
fgev()	極値分布	R16-4.R
filled.contour()	等高線塗りつぶし	R6-2.R
for()	ループ	R2-4.R
function()	関数作成	R2-1.R
gray()	モノクロ濃淡	R5-2.R
gs()	Grow-Shrink 法でのネットワーク推定	R14-1.R
hclust()	クラスタリング	R4-3.R
heatmap()	ヒートマップ	R5-1.R
hist()	ヒストグラム	R3-3.R
hist2d()	2 次元ヒストグラム	R6-2.R
if()	条件判断	R2-8.R
image()	格子スケール図	R5-2.R
jonckheere.test()	Jonckheere-Terpstra 検定	R13-7.R
kmeans()	クラスタリング k-means 法	R6-3.R
kruskal.test()	クラスカル - ウォリス検定	R13-7.R
length()	長さ・要素数	R2-7.R
lgamma()	対数ガンマ	R2-8.R
library()	パッケージを読み込む	R4-2.R
list()	リスト作成	R4-1.R
lm()	線形回帰	R13-7.R
log()	対数	R2-8.R
matrix()	行列作成	R2-5.R
max()	最大値	R2-6.R
MAX3()	MAX3 検定	R13-5.R
mean()	平均	R2-7.R
min()	最小値	R4-1.R
nj()	近隣結合法	R4-2.R
optim()	近似	R13-2.R
outer()	外積	R17-15.R
p.adjust()	FDR	R17-12.R
par()	図のパラメタ定義	R2-4.R
permutations()	順列	R10-1.R
persp()	鳥瞰図	R2-8.R
plot()	プロット	R2-4.R
points()	中心点の追加	R6-3.R
ppoints()	確率点の生成	R2-5.R
prop.trend.test()	傾向性検定	R13-7.R

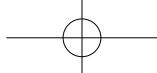
関数	意味	ソース ファイル
<code>qr()</code>	行列の QR 分解、行列のランク	R7-3.R
<code>quantile()</code>	クオンタイル	R17-2.R
<code>random.graph()</code>	グラフをランダムに作る	R14-1.R
<code>rbind()</code>	行結合	R7-5.R
<code>rdirichlet()</code>	ディリクレ分布	R12-4.R
<code>rep()</code>	同じ値の繰り返し	R2-8.R
<code>rexp()</code>	指数分布	R2-7.7
<code>rnorm()</code>	正規分布	R3-2.R
<code>rpois()</code>	ポアソン分布	R2-6.6
<code>runif()</code>	一様分布	R3-3.R
<code>sample()</code>	ランダムサンプリング	R2-7.R
<code>score()</code>	bnlearn パッケージでグラフの点数を計算する	R14-1.R
<code>segments()</code>	線分を引く	R16-1.R
<code>seq()</code>	数列作成	R10-3.R
<code>set.seed()</code>	疑似乱数列の種設定	R10-3.R
<code>sin()</code>	サイン	R10-3.R
<code>sort()</code>	ソート	R2-5.R
<code>sqrt()</code>	平方根	R10-3.R
<code>sum()</code>	和	R2-1.R
<code>svd()</code>	固有値分解	R7-5.R
<code>var()</code>	分散	R7-5.R
<code>which()</code>	該当するものを選ぶ	R10-3.R
関数の再帰定義	関数の再帰定義	R2-1.R

## A.4 R のヘルプを出す関数を使ってみる

R には常に新しい関数が付け加わっています。少しでも利用できるかもしれない関数を見つけたら、その内容を確認することが大事です。そのために解説記事がついています。たとえば、`dist()` 関数についての説明を読むなら、

```
> help(dist)
```

とします。ヘルプ画面の終わりの方に使い方 "Examples" があります。コピー & ペーストすれば R で実行できるようになっていますから、それをそのまま



ま実行してみてください。一通りの使い方がわかります。

### ● A.4.1 R のソースを確認する

#### (1) 言語と構成

R の関数は、R という言語で書かれたものと、C 言語、Fortran で書かれたものでできています。R のアプリケーションとしての骨格、計算関数の基本部分は C 言語、Fortran の関数を用いて作られていて、統計的処理の関数に R の関数が用いられている場合が多いという構成です。また、R の関数は、C 言語の関数、Fortran の関数を呼び出して使っている場合も多いです。

#### (2) R の構成——バイナリー配布とソース配布

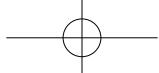
R はアプリケーションとして動く状態で配布されます。そのときに元のソースコードなしに動く状態のファイルのみで入手することもできますし（バイナリー配布）、ソースコードとともに入手することもできます（ソースでの配布）。いずれの方法にしろ、通常配布されたものは、統計ソフトとしての基本的なことが一通りできるアプリケーションを提供します。

場合によっては、この基本セットではできない解析などがあります。R はフリーであり、R で作成した解析ツールを利用者相互で共有する仕組みを持っています。パッケージという後付けできる関数セットがあり、これもインターネットからダウンロードすることができます。このパッケージは、一定のルールで作られている場合には、R 本体から簡単に入手することができます。その他にも、ルールに合ってはいないけれども、入手して取り込むことができるパッケージが個人的に公開されていることがあります。また、自分や仲間内で関数を作って共有することもあります。この場合には、パッケージとしての体裁をとる必要もありません。

#### (3) R のソースの確認

ソースを読むと、統計処理の内容がよくわかりますから、ソースを表示してみることは有用です。簡単に表示できる場合とそうでない場合があります。簡単に表示できる場合は関数名を入力するだけです。

```
> dist  
function (x, method = "euclidean", diag = FALSE, upper = FALSE,  
         p = 2)
```



```
{  
  if (!is.na(pmatch(method, "euclidian"))))  
  ...  
  return(d)  
}  
<environment: namespace:stats>  
>
```

このように関数のソースが表示されることもあります。

一方、ソースが表示されない場合もあります。その場合はいくつかのパターンに分かれます。

```
> median  
function (x, na.rm = FALSE)  
UseMethod("median")  
<environment: namespace:stats>  
> methods(median)  
[1] median.default  
> median.default  
function (x, na.rm = FALSE)  
{  
  if (is.factor(x))  
  ...  
}  
<environment: namespace:stats>  
>
```

median() 関数の場合は、UseMethod() 関数を使っていて、その詳細が隠されていることがわかります。UseMethod() 関数を使っている関数の場合には、methods() 関数を使うと隠れている中味に関する情報が増えます。median() 関数は median.default() 関数という中味を持っていることがわかるので、それをコマンドで与えるとソースが表示されます。

```
> t.test  
function (x, ...)  
UseMethod("t.test")  
<environment: namespace:stats>  
> methods(t.test)  
[1] t.test.default* t.test.formula*
```

```

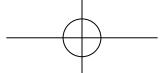
Non-visible functions are asterisked
> t.test.default
エラー: オブジェクト 't.test.default' がありません
> getS3method("t.test","default")
function (x, y = NULL, alternative = c("two.sided", "less",
  "greater"),
...
  return(rval)
}
<environment: namespace:stats>
>
```

`t.test()` 関数の場合に `methods()` 関数を使うと、`t.test.default()` 関数と `t.test.formula()` 関数という中味が隠れていることがわかるのですが、さらに隠されていることが関数の末尾の "\*" で示されています。この場合には `t.test.default()` 関数というコマンドは無効で、`getS3method("t.test","default")` とか `getS3method("t.test","foumula")` というコマンドによって、初めてソースが表示されます。

```

> runif
function (n, min = 0, max = 1)
.Internal(runif(n, min, max))
<environment: namespace:stats>
> sum
function (... , na.rm = FALSE) .Primitive("sum")
> chol
function (x, ...)
UseMethod("chol")
<environment: namespace:base>
> chol.default
function (x, pivot = FALSE, LINPACK = pivot, ...)
{
...
}
<environment: namespace:base>
>
```

`runif()` 関数の場合は `UseMethod()` ではなく、`.Internal()` とな



っています。これは、R のアプリケーション本体として登録された関数を呼び出していることを表しています。また、R 言語ではなく C 言語で書かれていることも意味します。C 言語のソースはバイナリで入手した場合は、ソースを読むことができません。ソースとして入手した場合には、解凍先の `src/main` 以下にある C 言語のソースファイルのどこかにあります。C 言語の関数としての呼び名は、解凍先の `src/main/names.c` というファイルに書かれているので、そこからたどることになりますが、関数名のファイルがあるわけではないので、見つけるのが簡単でないことが多いです。

`sum()` 関数の場合には、`.Internal()` の代わりに `.Primitive()` となっています。この場合も `.Internal()` と同様です。

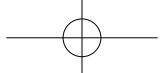
`chol()` 関数の場合には、再び `UseMethod()` が使われていますので、`chol.default` でソースを表示できます。`dist()` 関数はソースをすぐに表示できました。この 2 つの関数では処理の主要な部分で、`.Call()`, `.Fortran()`, `.C()` 関数が使われています（上記のリストでは省略されています）。この 3 つともう 1 つ `.External()` を合わせた 4 つの関数は、すでに述べた `.Internal()`, `.Primitive()` と同様に、R 言語以外で書かれた関数を呼び出すものです。それらの関数は、ソースコードファイルの中にあります。探し方の原則は、パッケージの名前などの付加情報を活用することでしょう。必ずしも、容易にたどり着けるわけではないようです。



## A.5 確率分布関数、疑似乱数列の発生

本書では図を作成するのに R の分布関数と疑似乱数列とを多用しました。R にはさまざまな確率分布があります。それらについて、R では確率密度、累積確率、クオントイル、疑似乱数列を与える関数が用意されています。

たとえば、二項分布であれば分布名を表す文字列 `"binom"` と、確率密度 (`"d"` : density)、累積確率 (`"p"` : 検定  $p$  値を求めるときに使う)、クオントイル (`"q"` : quantile)、疑似乱数 (`"r"` : random) とを組み合わせて、`dbinom()`, `pbinom()`, `qbinom()`, `rbinom()` という関数があります。ボアソン分布なら `dpois()`, `ppois()`, `qpois()`, `rpois()` 関数です。



# 付録 B

.....

## 数式記号

- $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$
- $\prod_{i=1}^n x_i = x_1 \times x_2 \times \dots \times x_n$
- $N! = N \times (N-1) \times \dots \times 2 \times 1$
- $\binom{N}{k} = \frac{N!}{k!(N-k)!}$
- **集合の交わり** :  $A \cap B$ 、**集合の和** :  $A \cup B$ 、**補集合** :  $A^c$ ,  $A \cap A^c = \emptyset$  (**空集合**)
- **階乗、ガンマ関数** ( $\Gamma(z)$ )、**ベータ関数** ( $B(x, y)$ )、**ベータ分布**
  - $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$
  - $\Gamma(z+1) = z!$
  - $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
  - $B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt = \frac{\Gamma(x) + \Gamma(y)}{\Gamma(x+y)}$
  - $\beta(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$

