Rによる統計解析の基礎

中澤　港

目次

第 1 章　統計学とは何だろうか？ 9
1.1　統計学の歴史 9
1.2　不確実性とランダム（乱雑さ） 10
1.3　統計解析の手順 11
1.4　統計解析の２大方針 13
1.5　統計解析の道具 13

第 2 章　統計的な考え方の基礎―確率と確率分布 15
2.1　本章のテーマ 15
2.2　確率的な現象を統計的事象と呼ぶ 15
2.3　準備その１：「標本空間」 16
2.4　準備その２：「事象」 16
2.5　準備その３：余事象・和事象・積事象・排反事象 17
2.6　準備その４：相互排反性と加法定理 17
2.7　準備その５：事象の独立性と乗法定理 18
2.8　確率を定義するための４種類のアプローチ 19
2.9　大数の法則（操作的接近の根拠） 21
2.10　確率変数・期待値・分散の感覚的把握 21
2.11　確率変数・期待値・分散を数式で書く 22
2.12　ベルヌーイ試行と２項分布 22
2.13　２項分布のシミュレーション 23
2.14　２項分布の理論分布 24
2.15　正規分布 25

第 3 章　データの尺度・データの図示 27
3.1　尺度と変数 27

3.2 名義尺度 (nominal scale) 28
3.3 順序尺度 (ordinal scale) 29
3.4 間隔尺度 (interval scale) 30
3.5 比尺度 (ratio scale) 30
3.6 データの図示 30
3.7 離散変数の図示の例 33
3.8 連続変数の図示の例 34

第 4 章 データを１つの値にまとめる（代表値） 37
4.1 ２つの戦略 37
4.2 中心傾向（Central Tendency） 38
4.3 ばらつき（Variability） 47
4.4 まとめ 52

第 5 章 比率に関する推定と検定 55
5.1 母比率を推定する方法 55
5.2 推定値の確からしさ 56
5.3 信頼区間 57
5.4 正規近似による信頼区間の推定 58
5.5 母比率の検定 59

第 6 章 カテゴリ変数２つの分析（１） 65
6.1 ２つのカテゴリ変数を分析する２つのアプローチ 65
6.2 ２つのカテゴリ変数の母比率の差の検定と信頼区間 65
6.3 ２つのカテゴリ変数の関係を調べることと研究のデザイン 68
6.4 クロス集計とは？ 69
6.5 独立性の検定の原理 69
6.6 フィッシャーの直接確率（正確な確率） 72

第 7 章 カテゴリ変数２つの分析（２） 75
7.1 研究デザインとリスク，オッズ 75
7.2 その他の関連性の指標 79
7.3 一致度の指標 81
7.4 利用例 82

第 8 章　平均値に関する推定と検定 85
8.1　母平均値と標本平均の差の検定 85
8.2　独立 2 標本の平均値の差の検定 87
8.3　両側検定と片側検定 90
8.4　対応のある 2 標本の平均値の差の検定 91

第 9 章　2 群の差に関するノンパラメトリックな検定 93
9.1　ノンパラメトリックな検定とは？ 93
9.2　Wilcoxon の順位和検定 94
9.3　正規スコア検定 100
9.4　メディアン検定 100
9.5　符号付き順位和検定 101
9.6　符号検定 102
9.7　並べ換え検定 103

第 10 章　多群間の差を調べる〜一元配置分散分析と多重比較 105
10.1　多群間の比較を考える 105
10.2　一元配置分散分析 106
10.3　クラスカル＝ウォリス (Kruskal-Wallis) の検定 109
10.4　多重比較 111

第 11 章　相関と回帰 119
11.1　量的変数の関連を調べる 119
11.2　相関関係の具体的な捉え方 122
11.3　回帰の考え方 124

第 12 章　時系列データと間隔データの扱い方 131
12.1　時間を扱うとはどういうことか？ 131
12.2　時系列解析の基礎 132
12.3　生存時間解析の基礎 141

第 13 章　一般化線型モデル入門 147
13.1　一般化線型モデルとは？ 147
13.2　変数の種類と数の違いによる線型モデルの分類 147
13.3　重回帰分析 150

13.4 共分散分析 151
13.5 補足：一般線型混合モデル 156

第 14 章 高度な解析法についての概説 157
14.1 主成分分析 157
14.2 因子分析 158
14.3 クラスター分析 160

第 15 章 参考文献 163

付録：R について 165
A.1 なぜ R を使うべきなのか？ 165
A.2 R を使うための最初の 1 歩 167
A.3 R の参考書・web サイトなど 169

はじめに〜本書の狙い

およそ世の中のすべての現象は不確実性，予測不可能性を含んでいる。とくに人間が絡む現象はそうである。しかし，予測能力がきわめて大きいことはヒトという動物の特徴であり（そう，リチャード・ドーキンスも語っている通り），我々は不確実なながらも先の見通しを立てて，何とかうまく生きていこうと考えざるを得ない。

では，どうやって見通しを立てればいいのだろうか？　不確実で不安定な現象でも，数多く集めれば，何らかの法則性が見えてくることがある。この，法則性を見出す（検証を含む）方法論が統計学である。本書の目的は，統計学の考え方の基礎を説明しながら，実際に多くのデータを集めて分析する技術の初歩を解説することにある。

実際に統計手法を身につけるには，データを使って分析してみることが近道だが，従来は，高価な統計ソフトに依存したり，市販ソフトの機能制約版を使ったり，あるいは独自開発のソフトを使って解説するしかないのが難点であった。しかし現在では，オープンソースで国際共同開発されている **R** (`http://www.r-project.org/`) という素晴らしいソフトウェアが存在し，機能的にも市販ソフトに引けを取らないし，大勢の専門家の目によって吟味されており信頼性も高いので，本書では **R** を使った分析法を説明する。日本において R が普及していない原因は日本語による解説書がほとんどなかった[*1]ためだと思うので，本書が **R** の普及の一助になれば幸いである。

なお，本書は，高崎経済大学地域政策学部における 2001 年度の社会統計学の講義，山口県立大学における 2002 年度の統計学の講義資料，及び講義の際に受けた質問を元にして，大幅に加筆修正を行ったものである。講義に出席し，さまざまな質問をし

[*1] R-jp メーリングリスト [`http://epidemiology.md.tsukuba.ac.jp/~mokada/ml/R-jp.html`] 有志により翻訳されたマニュアルの pdf 版が `http://phi.ypu.jp/swtips/R-jp-docs/` から入手できる。なお，R-jp メーリングリストとは，筑波大学の岡田昌史さんによって運営されている，R について日本語で情報交換をするためのメーリングリストである。

てくれた学生諸氏に感謝申し上げる。また，本書の草稿を読んで有益なご指摘をくださった R-jp メーリングリストの方々に感謝申し上げる。なかでも，数多くの丁寧なコメントをくださった群馬大学社会情報学部の青木繁伸先生には深く感謝申し上げる。もちろん，本書の内容に間違いがあれば，それは著者個人の責任である。

2003 年 6 月 16 日中澤　港

第 1 章

統計学とは何だろうか？

1.1 統計学の歴史

統計学の歴史については，さまざまなことが言われている。ラオ（1993）によれば，統計学の歴史をずっと過去に遡っていくと，本来は，「家畜や他の財産の帳簿をつけるために原始人が木につけた刻み目」だという[*1]。その後，国家を経営するための基礎資料的な意味あいが強くなった。ラオ（1993）には，「ある国およびそこに生きている生命の状態や発展についての，最も完全で，最も根拠のある知識」という，マルシャスの言葉が引用されている。

英語の statistics は，ラテン語で国家を意味する status を語源として，18 世紀半ばにドイツの哲学者アッヘンウォールが作った言葉が元になっている（ラオ，1993 年）。国家としての人口規模が大きくなるとともに，雑然とした大量の生データを，解釈をやさしくしたり種々の方策決定に用いるためにまとめ上げる手法が必要になり，グラントの生命表やケトレーの度数分布図など質的にも量的にも統計手法が開発されてきた。産業革命が進行する中，1834 年，英国王立統計協会設立により，学問としての「統計学」が成立し，「人間に関係することがらで，数量で表現することが可能で，一般的な法則を導き出すのに十分なだけ積み重ねられたもの」と定義された。

20 世紀末，コンピュータの進歩とともに統計学の理論や技術も大きく進展し，同時にパッケージソフトウェアの開発によって，統計学の専門家でなくても統計解析を行うことが可能になった。そのため，統計学は，いまやすべての自然科学や社会科学で適用される科学的分析の技術となっている。もっとも広い意味で定義するならば，

[*1] 「原始人」という呼び方には問題があるが，大事なことは，文字が無かった時代であっても統計的な概念は必要だったし，ありえたということである。

「不確実性を考慮した論理的推論」ということになるだろう。

1.2　不確実性とランダム（乱雑さ）

世の中のほぼすべての事象は不確実性を含んでいる。素粒子レベルでは物理法則も不確実性を含むし（ある原子核に含まれる電子が存在する確率がゼロでない場所という意味で電子雲は決まるけれども，電子がある瞬間にどこに存在するかということは確率的にしかいえない），遺伝子の発現や社会における個人の行動なども，決して決定されてはいない。

不確実性を数学的に扱うには，確率的に起こるできごと（確率事象）を扱わねばならない。確率事象は，一般に，何回中何回くらい起こりそうかはわかっているが，いつ起こるかがわからない。そのために，ふつうは既知の分布関数が使われる。ただし，コンピュータ上で扱うときは，ランダムな数字の列，即ち乱数列[*2]を利用することができる。例えば区間 (0,1) の実数値をとる一様乱数列[*3]を確率事象に割り当てれば，その値が確率 p より小さいときに事象が起こり，確率 p より大きいときに事象が起こらないと解釈することによって，確率を事象が起こるか起こらないかに置き換えることができる。このやり方で確率分布をシミュレートすることは，コンピュータ集約型統計学と呼ばれる分野で近年盛んに行われている。

乱数列については，線型合同法など，数式を使って生成される擬似乱数列というものがあり，数式がわかれば次の数字は予想できるのだが，見かけ上はでたらめな数の並びに見え，実用上十分なでたらめさをもっているので，コンピュータ上で乱数列が必要な場合は良く使われている。現在ではコンピュータ上でも熱拡散の状態を測って真の乱数を得る拡張ボードが市販されていて（例えば東芝のランダムマスターなど）利用可能だが，メルセンヌツイスター[*4]のような優れたアルゴリズムで生成された擬似乱数を使う方が普通である。

[*2] 次の数字が予想できない，意味のないでたらめな数の集まり。例えば，20 桁の対数表の 15〜19 桁目を並べたものとか，袋に入れた 500 個ずつの白ビーズと黒ビーズから，よく混ぜて 1 個ビーズを取り出して色を記録し，元の袋の中に戻して，取り出す前と同じ状態に戻してからまたビーズを取り出して色を記録し，と繰り返した（復元抽出した）ときの色の列など。

[*3] 乱数列のうち，各数字の出現頻度が等しいと期待されるものを一様乱数列と呼ぶ。

[*4] 松本眞，西村拓士両氏によって，1996 年から 1997 年にわたって開発された擬似乱数生成アルゴリズムで，生成速度が速く，きわめて周期が長く，C で書かれたプログラムソースコードが自由に利用できるという利点をもつ。詳細は `http://www.math.keio.ac.jp/home2/matumoto/public_html/mt.html` を参照されたい。

1.3 統計解析の手順

データの分析技術としての統計解析は，一定の手順を踏んで行われる。箇条書きすると，以下のような手順が典型的と思われる。

- 目的を明確にする
- 生データをとる
- データ化（エディティング，コーディング，データ入力）
- データの図示（幹葉表示やヒストグラムなど）
- 代表値（分布の位置やばらつきを示す値）の計算
- 作業仮説の明確化（ここで因果関係についての仮説を立てることが多い）[*5]
- 仮説検定や区間推定を行う（攪乱要因に配慮し，その影響を制御する必要がある）
- 因果関係についての推論を行う（先行研究の知見なども総合する必要がある）

データ化以前の段階についての細かい説明は調査法についての本やデザインの本を参照されたいが，大事なことは，データを取る前の段階で，統計解析をどうするか考えておくべきだということである。実際にはなかなかそうはできず，データを取った後で解析法が考えられる場合が多いのだが，後付けの分析はバイアスの元になるし，言いたいことを検討するための解析に必要なデータが取れていないことが解析段階で判明しても，後の祭りなのである。だから，統計解析は，データを取ったあとで始まるものではなく，データを取る前の段階で始まっていることを肝に銘じておくべきである。フィールドワーカーや実験科学者や政策担当者は，データを取ってから解析法に困って統計学者に相談するのではなく，デザインの段階から相談すべきである。

因果関係については，数値間に常に関連があるだけではなく，時間的前後関係など，いくつかの条件を満たさないと因果関係があるとはいえないし，その条件についてもいろいろ議論がある（Rothman, 2002）。一般には，図 1.1 に示すように，その分析で着目している結果（英語では outcome と呼ばれる。この図の場合なら高血圧）を評価軸としたとき，「A：この分析で結果との関係を評価したい因子（この図では体脂肪割合や食塩摂取量）」「B：この分析で結果との関係をみたいわけではないが結

[*5] 因果関係については，佐藤・松山 (2002) の議論がすばらしく良くまとまっているし，Rothman (2002) の第 2 章の議論も参考になる。後者は `http://www.oup-usa.org/sc/0195135547/media/0195135547_ch2.pdf` として web で全文が公開されている。

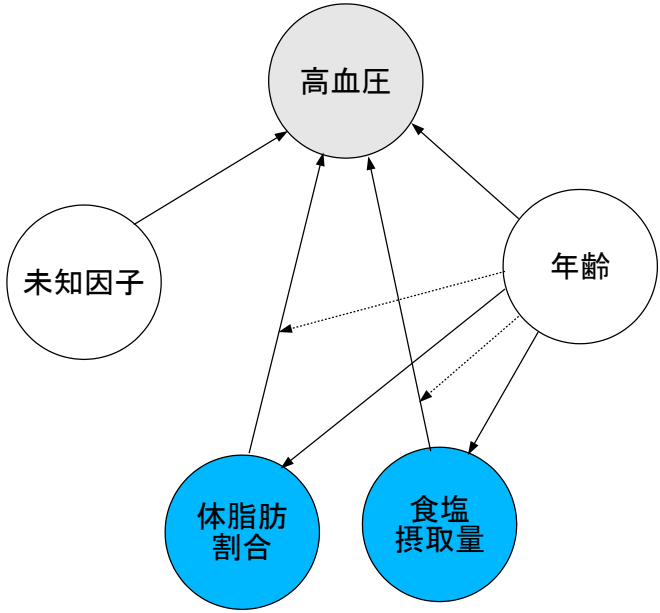


図 1.1: 因果関係と攪乱要因の例

果やAやそれらの関係に影響することがわかっている因子（この図では年齢）」「C：この分析では調べていないが，結果に影響する因子（この図では未知因子）」に分けて捉えることができる。Bは攪乱要因とか交絡要因と呼ばれ，Bの影響を調整した上で，Aと結果の関係を調べないと，真の関係はわからない。結果のうち，Aによって説明できない部分が偶然変動及びCとして残るので，Cの影響は小さいほどよい。その意味で，統計解析では結果のどれくらいがAによって説明できるかを調べたり，Cの影響がどれくらいあるかを調べる手法があり，因果関係の確からしさを判断する上で重要である。

1.4 統計解析の２大方針

上に述べた意味での手順は同じだが，推論の根拠について考えたとき，統計解析には，大きく分けて２つの異なる方針がある[*6]。

１つは，デザインに基づく解析である。仮説検定の例としては，並べ替え検定やログランク検定がこれに当たる。デザインに基づく解析では，データが，ランダムにデータを取った場合に得られるパタンの１つであると考え，その確率を直接計算する。攪乱要因はデータを層別することで制御する。一般に計算量は多くなるが，分布を仮定する必要がないのが利点である。

もう１つは，モデルに基づいた解析である。ｔ検定や重回帰分析や比例ハザードモデルなど，有名な多くの統計解析は，この考え方に基づいている。デザインに基づく解析に比べると，一般に計算量は少なくて済む。結果の分布を記述するために確率分布を仮定し，その未知パラメータをデータから推定する。データがモデルにもっとも良く当てはまるようなパラメータを最小二乗法や最尤法などで推定し，当てはまりの悪さを AIC などの指標を使って評価する。攪乱要因の影響は，説明変数としてモデルに入れることで調整することになる。

1.5 統計解析の道具

実際の統計解析は，コンピュータのソフトを使って行われるのが普通である。SAS と SPSS が最も有名でよく使われているソフトだし高機能だが高価である。会社などで十分な予算があれば，SAS を使ってもよいであろう。しかし個人が使うには非現実的な値段である。JMP や Statview などはそれほど高価ではなく，マウスでデータを見ながら操作できるので取っ付きやすいが，複雑な操作を一度に実行させたり，大量の計算をさせるにはあまり向いていないように思われる（JMP はバージョン４からスクリプト言語をサポートしたので，複雑な操作を予めスクリプトファイルに書いておくという使い方も可能になったが)。

また，世間では，MS-Excel を使って初歩的な統計解析をすることも多いようである。覚えておくといろいろ便利だが，解析の中身がブラックボックスだし，それがないと何も出来ないという状況では困る。そもそも統計ソフトではないので，少量

[*6] この考え方について詳しくは，松山裕「統計解析の２つの原理」，帝京大学研究用コンピュータ室ニュース No.25，pp.54-64，1992 年 7 月 31 日を参照されたい。

のデータ入力や変換には便利（大量のデータ入力にはデータベースソフトを使うか，html のフォームと cgi を組み合わせて入力環境を作ると良い）だが，MS-Excel のマクロ言語などで統計解析をすると，後で何をしたのかわからなくなりやすいし（その点は Statview なども同じ危険を孕んでいる），ちょっとだけ修正するといったことがやりにくいので，本格的に統計データ解析をするには薦められない。

本書では R を利用した解析方法を説明する。R の最大の利点は，オープンソースであり，かつ拡張性が高い点だと思うが，慣れれば使いやすさもかなり高い水準にある。少なくとも表計算ソフトや汎用言語を使うよりは，ずっと簡単に統計処理ができる。R についての詳細は，付録を参照されたい。

第 2 章

統計的な考え方の基礎—確率と確率分布

2.1 本章のテーマ

前章で，統計学とは「不確実性を考慮した論理的推論である」と述べた。不確実性とは，0%でも 100%でもないファジーな確率をもっているということである。

しかし，ここで簡単に「確率」と言ってしまったが，さて，改めて確率とは何かと訊かれたら，答えに詰まってしまうのではないだろうか。

そこで，本章では，あらゆる統計的な考え方の基礎となる「確率」というものを徹底的に考えてみることにする。高校数学の復習になってしまうかもしれないが，頭を整理しておくという意味で，役に立つのではないだろうか。このテーマについて，もっと詳しく知りたい方は，伏見 (1987) など，確率論の本を参照されたい。

2.2 確率的な現象を統計的事象と呼ぶ

2.2.1 どういう現象が確率的か？

- サイコロを振ったときの目：振ってみるまでは 1 から 6 のどれが出るかはわからない。どの目がでる可能性も等しいから。
- 天気予報：「明日の天気予報は晴れ」といっても「必ず晴れる」とは限らない。「曇ったり雨が降ったりする可能性も少しはあるが，晴れる可能性が高い」ことを意味する。
- 喫煙と肺がんの関係：「タバコを吸うと肺がんになる」という命題は，「タバコ

を吸った人と吸わなかった人を比べて，肺がんになった人の割合が吸った人の方で高い」という関係を示す。タバコを吸っても肺がんにならない人もいるし，吸わなくても肺がんになる人もいる。

こういう「不確かさ」に潜む法則性（長期間繰り返し観察したり，大集団で観察すると見られる）を考える学問を確率論と呼ぶ。大雑把に言えば，この種の法則性をもつ現象を，「統計的事象」と呼び，統計的事象の確かさの度合いを示すのに便利なモノサシが「確率」である。そこで，「確率」をきちんと定義してみることにする。その前に，いくつかの準備が必要である。

2.3　準備その１：「標本空間」

統計的事象を捉えるには，「どんなことが起こりうるか」という範囲を定めることが必要である。

現象は一般に多面的で，様々な観察方法がある。以下３点によって統計的現象を捉えた，記号化された結果の集合のことを「標本空間」と呼ぶ。

- 観察を行う側面を特定する
- 起こりうる結果の範囲を規定する
- その範囲内の各結果に記号を対応させる

個々の結果の起こりうる可能性を示す数値（これを「確率」と呼ぶ）を考える。もっとも単純には「どの結果も同程度に起こる」と考える。各結果に対応付けられた確率は０から１までの数値であり，各確率の値の総和は１にならねばならない。例えば，サイコロの目では，標本空間は｛1，2，3，4，5，6｝である。

2.4　準備その２：「事象」

問題は，個々の結果の可能性よりも，いくつかの結果が複合された集合（これを「事象」と呼ぶ）の起こる可能性がどのくらいか，ということである。つまり，「事象」＝「標本空間の部分集合」である。

サイコロの例では，「目が偶数（丁）」とか「目が５以上」とか「目が１」とかいうことが事象といえる。

ある事象の確率は，その事象に含まれる各結果の生起確率の和である。従って，各結果の生起確率が等しい場合は，その事象に含まれる結果の場合の数をすべての場合

の数で割ると，その事象の確率になる。サイコロの例では，「目が５以上」という事象の確率は，2/6=0.333... である。

2.5 準備その３：余事象・和事象・積事象・排反事象

起こりうるすべての結果の集合を「全事象」という。つまり，全事象は標本空間に等しい。

決して起こらない事象を「空事象」といい，空集合 ϕ で表す。

事象 E に対して，E が起こらないという事象を E の「余事象」という。E の余事象を $\bar{E}$ と書く。サイコロの例では，「目が偶数」という事象の余事象は「目が奇数」である。$Pr(E) + Pr(\bar{E}) = 1$ が常に成立する。

事象 E と F の少なくとも一方が起こるという事象を，E と F の「和事象」といい，$E \cup F$ で表す。

事象 E と F の両方が起こるという事象を，E と F の「積事象」といい，$E \cap F$ で表す。

事象 E が起これば F は決して起こらないとき，E と F は「排反事象」であるという。E と F が排反事象なら，$E \cap F = \phi$ である。

2.6 準備その４：相互排反性と加法定理

サイコロで考えると，１回振ったとき「偶数の目が出る」という事象 E が起こる確率（これを $Pr(E)$ という記号で書くことにする）は，2,4,6 の場合の数３を，1,2,3,4,5,6 の場合の数６で割った値なので $Pr(E) = 0.5$ である。

では，２回振って「少なくとも１回は偶数の目」の確率はどうなるだろうか？　まず，0.5+0.5=1.0 ではないのは自明である（1.0 ということは，必ずそうなるということだから）。ここで大切なのは，『１回目に「偶数の目が出る」事象 E_1 と２回目に「偶数の目が出る」事象 E_2 とは排反ではない』ことに注意することである。集合のベン図（図 2.1）から考えると，$Pr(E_1 \cup E_2) = Pr(E_1) + Pr(E_2) - Pr(E_1 \cap E_2)$ であることが直感的にわかる。この式を「加法法則」と呼ぶ。ベン図をよく見ると，「２回とも奇数」の余事象なので，$1 - Pr(\bar{E_1} \cap \bar{E_2}) = 1 - 9/36 = 0.75$ と考えてよいこともわかるだろう。因みに，事象 E と事象 F が排反なら，$Pr(E_1 \cap E_2) = 0$ なので，$Pr(E \cup F) = Pr(E) + Pr(F)$ という「加法定理」が成立する。

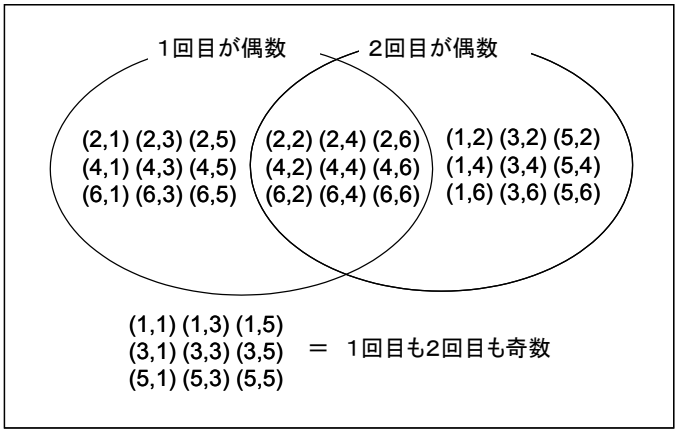


図 2.1: サイコロを 2 回振って「少なくとも 1 回は偶数の目」の確率を考えるためのベン図

2.7 準備その５：事象の独立性と乗法定理

事象 E が起こったときに事象 F が起こる確率を，「E が起こったときの F の条件付き確率」といい，$Pr(F|E)$ と書く。

「E が起こった」うちの「E も F も起こった」場合なので，$Pr(F|E) = Pr(F \cap E)/Pr(E)$ である。

E と F が互いに無関係（＝独立）なら，$Pr(F|E) = Pr(F)$。逆にいえば，$Pr(F) = Pr(F|E)$ のときに事象 E と事象 F は互いに独立であるという。独立でないとき「従属である」という。

上記 2 つの式から，E と F が独立なら，$Pr(F \cap E) = Pr(F) \times Pr(E)$ という「乗法定理」が成立する。

2.8 確率を定義するための４種類のアプローチ

確率を定義するには，以下の４種類のアプローチがある。

- 操作的アプローチ（統計的定義）：数多く試したときの相対度数の極限。例えば，事象 E が起こる確率 $Pr(E)$ は，N 回試したときに N_1 回事象 E が起こるとして，N_1/N という相対度数が，N を無限大にしたときに漸近する値である。
- 対称的確率：サイコロの場合，６通りの目の出る確率はどれも等しくなければならず，その和は１でなくてはならないので，例えば１の目が出る確率は $1/6$ となる。限定的かつ循環論法。
- 公理的客観確率：標本空間の各要素を e_i として，$Pr(e_i) >= 0$ かつ $Pr(e_1) + Pr(e_2) + ... + Pr(e_N) = 1$ かつ事象 E が起こる確率 $Pr(E) = \sum Pr(e_i)$ を公理とする。[*1]。

[*1] 要素は互いに排反であるということ。当然である。より厳密に定義するならば，次のようになる（伏見，1987 より）。

確率の議論をする際に考える事象群 $\mathfrak{A}$ は，次の条件を満たしていなければならない。

【B1】標本空間 Ω が $\mathfrak{A}$ に含まれている。

【B2】事象 A が $\mathfrak{A}$ に含まれているならば，A の余事象 $\bar{A}$ も $\mathfrak{A}$ に含まれている。

【B3】$A_1, A_2, ...$ が $\mathfrak{A}$ に含まれているならば，それらの和事象 $A_1 \cup A_2 \cup ... \left(= \bigcup_{i=1}^{\infty} A_i \right)$ も $\mathfrak{A}$ に含まれている。

これら３つの条件をすべて満たす $\mathfrak{A}$ のことをボレル集合体という。このとき，以下の条件も自然に成立する。

【B4】空集合 ϕ が $\mathfrak{A}$ に含まれている。（ϕ は標本空間の余事象なので，【B1】と【B2】より自明である）

【B5】$A_1, A_2, ...$ が $\mathfrak{A}$ に含まれているならば，それらの積事象 $A_1 \cap A_2 \cap ... \left(= \bigcap_{i=1}^{\infty} A_i \right)$ も $\mathfrak{A}$ に含まれている。（集合論におけるド・モルガンの法則〔＝積事象の余事象は余事象の和事象に等しい〕と【B2】と【B3】より自明である）

こうしてボレル集合体 $\mathfrak{A}$ を定めた上で，$\mathfrak{A}$ に含まれる個々の事象が起こる確率（確率測度ということもある）を定義することができる。事象 A の起こる確率を $Pr(A)$ という記号で書くと，確率 $Pr(\cdot)$ は，次の性質をもつものとして定義できる（これが公理的客観確率の厳密な定義である）。

【P1】任意の事象 A の確率は 0 と 1 の間の実数である。

【P2】標本空間全体 Ω の確率は 1 である。

【P3】$A_1, A_2, ...$ が互いに排反な事象であるならば，それらの和事象の確率は，それらの確率の和に等しい（「完全加法性」と呼ばれる）。

まず標本空間を考え，その部分集合の集まりの一種としてボレル集合体を考え，最後にボレル集合体

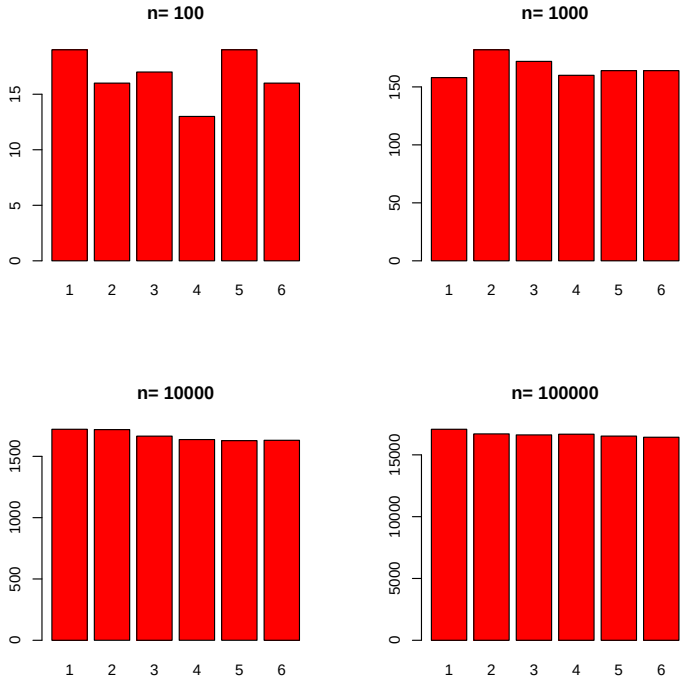


図 2.2: 大数の法則のシミュレーション

- 主観確率：観念的にも二度と繰り返すことのできない事象についての「見込み」を扱う。決定理論において重要。

2.9 大数の法則（操作的接近の根拠）

操作的アプローチが成り立つ様子を，コンピュータを使って調べてみよう。R のプログラムでは（runif() は一様乱数を発生させる関数），

```
a <- c(100,1000,10000,100000)
op <- par(mfrow=c(2,2))
for (i in 1:4) {
y <- as.integer(runif(a[i],1,7))
s <- paste("n=",as.integer(a[i]))
barplot(table(y),main=s) }
par(op)
```

とすれば，図 2.2 のように，試行回数を増やすと，サイコロの特定の目が出る割合が，ある一定値に近づくことがわかる。「1 の目が出る」事象 E_1 が起こる確率 $Pr(E_1) = p$ とおけば，N 回サイコロを振って N_1 回 1 の目が出たとして，任意の正の小さな数 ε に対して，$\lim_{N\to\infty} Pr(|N_1/N - p| < \varepsilon) = 1$ ということで，これをベルヌーイ (Bernoulli) の大数の法則という。

2.10 確率変数・期待値・分散の感覚的把握

確率変数と期待値について，まず感覚的に把握することは重要である。そこで，鈴木義一郎「情報量基準による統計解析入門」（講談社サイエンティフィク）に掲載されている，スロットマシンの例を紹介しよう。

スロットマシンでは，ごくたまに，投入金額の何十倍ものコインが出てくることがある。マシン利用者全員に返ってくる賞金の合計を利用回数で割った値が，1 回に期待される賞金額で，これを賭け金で割った値を「賞金還元率」と呼ぶ。言い換えると，1 から賞金還元率を引いた値が，賭け事の胴元が儲けると期待される値である。一般に，賞金額が $x_1, x_2, x_3, ...$ で，その賞金が得られる確率が $p_1, p_2, p_3, ...$ のように設定されたスロットマシンの期待賞金額 M は，$M = x_1p_1 + x_2p_2 + x_3p_3 + ...$ で与えら

の要素に実数値を対応させる関数で性質【P1】〜【P3】を満たすものとして確率を定めたので，これら 3 つを組にして $(\Omega, \mathfrak{A}, P)$ と書き，これを確率空間と呼ぶ。

れる。このスロットマシンのようなものを確率変数といい，期待賞金を期待値と呼ぶ（厳密には後述）。

期待賞金が同じでも，値動きの幅が小さいと一喜一憂の程度が小さく，逆に幅が大きいと滅多に当たらないが当たったときの喜びは大きくなる。つまり，ギャンブル性は，値動きの幅と，チャンスの大きさに依存している。

各賞金がどれくらい期待賞金から隔たりがあり，それを獲得できる可能性がどれくらいあるのかを見積もれば，ギャンブル性が表せる。

マシンのギャンブル性を V とおけば，$V = \sum$（期待値からの隔たり）$\times$（可能性）という値が定義できる。この V を「分散」と呼ぶ。このとき，各賞金額 x と期待値 M の隔たりは，差を二乗した値 $D = (x - M)^2$ で表す。

2.11　確率変数・期待値・分散を数式で書く

一般に，とりうる値の集合 $\mathbf{x} = (x_1, x_2, x_3, ...)$ と，それぞれの値が実現する確率 $\mathbf{p} = (p_1, p_2, p_3, ...)$ が与えられていて，事象として $\mathbf{x}$ のうちどれか 1 つの値のみ実現するとき，$(\mathbf{x}, \mathbf{p})$ という 1 セットを，「確率変数」と呼んで，X で表す。このとき，期待値は $E(X) = \mu = \sum x_i p_i$ であり，分散は $V(X) = \sigma^2 = \sum (x_i - \mu)^2 p_i$ となる。また，分散の平方根 σ を標準偏差と呼ぶ。

このとき，横軸に $\mathbf{x}$ の各々の値を示す位置をとり，その各々に $\mathbf{p}$ の各々の可能性を示す高さの棒を立ててみれば，これが確率変数の「確率分布」ということになる。

2.12　ベルヌーイ試行と 2 項分布

1 回の実験で事象 S か事象 F のどちらかが起こり，しかもそれらが起こる可能性が，$Pr(S) = p, Pr(F) = 1 - p = q$ で何回実験しても変わらないとき，これを「ベルヌーイ試行」という。ベルヌーイ試行では，事象 F は事象 S の余事象になっている。

例えば，不透明な袋に黒い玉と白い玉が 500 個ずつ入っていて，そこから中を見ないで 1 つの玉を取り出して色を記録して（事象 S は「玉の色が黒」，事象 F は「玉の色が白」）袋に戻す実験はベルヌーイ試行である（注：袋に戻さないと 1 回実験するごとに事象の生起確率が変わっていくのでベルヌーイ試行にならない）。

ベルヌーイ試行を n 回行って，S がちょうど k 回起こる確率は，$Pr(X = k) =_n C_k p^k q^{n-k}$ である。${}_nC_k$ は言うまでもなく n 個のものから k 個を取り出す組み合わせの数である。2 項係数と呼ばれる。このような確率変数 X は，「2 項分布に従う」といい，$X \sim B(n, p)$ と表す。$E(X) = np$, $V(X) = npq$ である。

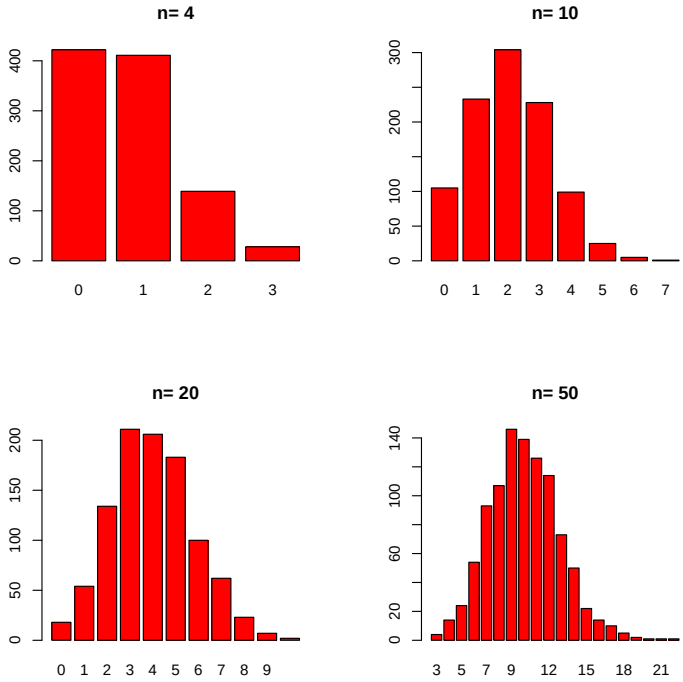


図 2.3:　2項分布のシミュレーション

2.13　2項分布のシミュレーション

正二十面体（各面には 1 から 20 までの数字が割り振られている）サイコロを n 回 $(n = 4, 10, 20, 50)$ 投げたときの，1 から 4 までの目が出る回数を 1 試行と考えれば，これはベルヌーイ試行である。1 回投げたときに 1 から 4 までの目が出る確率は 0.2 であるとして（＝母比率を 0.2 とする），試行 1000 セットの度数分布を図 2.3 に示す。R のプログラムは下記の通り。

```
times <- function(n) {
```

```
hit <- 0
dice <- as.integer(runif(n,1,21))
for (j in 1:n) { if (dice[j]<5) { hit <- hit+1 } }
return(hit)}

a <- c(4,10,20,50)
op <- par(mfrow=c(2,2))
for (i in 1:4) {
nx <- a[i]
y <- c(1:1000)
for (k in 1:1000) { y[k] <- times(nx) }
barplot(table(y),main=paste("n=",nx))
}
par(op)
```

2.14 2 項分布の理論分布

この例で，各 n についての理論的な確率分布は，$Pr(X = k) =_n C_k 0.2^k 0.8^{n-k}$ より図 2.4 のようになる。R のプログラムは下記の通り。

```
a <- c(4,10,20,50)
op <- par(mfrow=c(2,2))
for (i in 1:4) {
n <- a[i]
k <- 0
chk <- c(1:n+1)
while (k <= n) { chk[k+1] <- choose(n,k)*(0.2^k)*(0.8^(n-k)); k <- k+1 }
barplot(chk,col='red',main=paste("n=",n))
}
par(op)
```

ただし，R には様々な確率分布についての関数があり，`choose(n,k)*(0.2^k)*(0.8^(n-k))` は `dbinom(k,n,0.2)` と同値である。このように，確率変数が取りうる各値に対して，その値をとる確率を与える関数を確率密度関数という。値が小さいほうからそれ

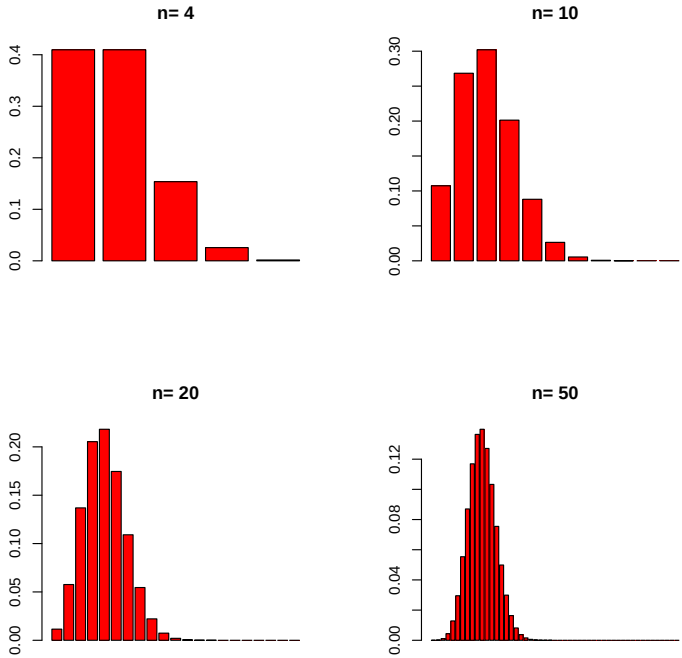


図 2.4:　2 項分布の理論分布

を全部足した値を与える関数（つまり，その確率変数の標本空間の下限から各値までの確率密度関数の定積分）を分布関数（あるいは確率母関数，累積確率密度関数）と呼ぶ。

2.15　正規分布

n が非常に大きい場合は，2 項分布 $B(n,p)$ の確率 $Pr(X = np + d)$ という値が，

$$\frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{d^2}{2npq}\right)$$

で近似できる。一般にこの極限（n を無限大に限りなく近づけた場合）である，

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

という形をもつ確率分布を正規分布と呼び，$N(\mu, \sigma^2)$ と書く。

$z = (x - \mu)/\sigma$ と置けば，

$$Pr(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

となる。これを標準正規分布と呼び，$N(0, 1)$ と書く。統計学でよく使われる確率分布であるカイ二乗分布とか t 分布とか F 分布は，正規分布から導かれる分布である。

参考

よく使われる確率密度関数，分布関数，分位点関数について R での表現の一覧を下表にまとめておくので，参考にされたい。

分布の種類	確率密度関数 (probability density function)	分布関数=確率母関数=累積確率密度関数 (distribution function = probability generating function cumulative probability density function)(4)	分位点関数 (quartile function)
カイ二乗分布	dchisq(カイ二乗値，自由度)	pchisq(カイ二乗値，自由度)	qchisq(%，自由度)
2 項分布	dbinom(生起回数，試行回数，母比率)	pbinom(生起回数，試行回数，母比率)	qbinom(%，試行回数，母比率)
ポアソン分布	dpois(生起回数，期待値)	ppois(生起回数，期待値)	qpois(%，期待値)
正規分布 (1)	dnorm(Z スコア，平均値，標準偏差)	pnorm(Z スコア，平均値，標準偏差)	qnorm(%，平均値，標準偏差)
対数正規分布 (2)	dlnorm(Z スコア，対数平均値，対数標準偏差)	plnorm(Z スコア，対数平均値，対数標準偏差)	qlnorm(%，対数平均値，対数標準偏差)
一様分布 (3)	dunif(値，最小値，最大値)	punif(値，最小値，最大値)	qunif(%，最小値，最大値)
t 分布	dt(t 値，自由度)	pt(t 値，自由度)	qt(%，自由度)
F 分布	df(F 値，第 1 自由度，第 2 自由度)	pf(F 値，第 1 自由度，第 2 自由度)	qf(%，第 1 自由度，第 2 自由度)

(1) 平均値と標準偏差は省略可能。省略時は標準正規分布 (平均 0，標準偏差 1) になる。

(2) 対数平均値と対数標準偏差は省略可能。省略時は対数平均 0，対数標準偏差 1 になる。なお，対数平均とは自然対数をとった値の平均，対数標準偏差とは自然対数をとった値の標準偏差をいう。dlnorm(1) は dnorm(0) と等しい。

(3) 閉区間である。省略時は 0 と 1 になる。

(4) R には，これらの分布関数に従う乱数を生成する関数もある。例えば，0 から 1 までの一様乱数を 1000 個生成する関数は runif(1000,0,1) である。試行回数 100 回，母比率 0.2 の 2 項分布に従う乱数を 1000 個発生させるには，rbinom(1000,100,0.2) とすれば良い。

第 3 章

データの尺度・データの図示

3.1 尺度と変数

尺度とは，研究対象として取り上げる操作的概念を数値として扱うときのモノサシの目盛り（の種類），言い換えると，「データに何らかの値を対応させる基準」である[*1]。尺度は，名義尺度，順序尺度，間隔尺度，比尺度（比例尺度ともいう）の４つに分類される[*2]。

研究対象として取り上げる操作的概念は，変数という形で具体化される。変数は，それが表す尺度の水準によって分類されるが，一般には，名義尺度は定性的変数（カテゴリ変数），順序尺度，間隔尺度，比尺度は定量的変数に相当する。定量的変数には，整数値しかとらない離散変数と，実数値をとりうると考えられる連続変数がある。順序尺度は離散変数，間隔尺度は離散の場合も連続の場合もあるが連続変数であること

[*1] より厳密には次の通り。非空の集合 A の要素間にいくつかの関係 $R_1, R_2, ..., R_n$ が成り立っているときに，これを $\alpha =< A, R_1, R_2, ..., R_n >$ と書くことにし，数量的な要素からなる非空の集合 B の要素間に関係 $S_1, S_2, ..., S_n$ が成り立っているときに，これを $\beta =< B, S_1, S_2, ..., S_n >$ と書くとき，もし B の中の要素が A の中のすべての要素 $x, y (x, y \in A)$ の写像 $f(x), f(y)$ からなり $(f(x), f(y) \in B)$，x と y の間に関係 $R_1, R_2, ..., R_n$ が成り立っているときに B の中の $f(x)$ と $f(y)$ の間に関係 $S_1, S_2, ..., S_n$ が成り立っている（これを準同型という）ならば，関係系 α は関係系 β によって “表現される”，という。測定とは，経験的世界の関係系 α が数量的な形式関係系 β によって表現されることをいう。尺度とは，このような $< A, B, f >$ の組である。B の各要素に変換 ϕ を施して得られる集合の要素を考えるとき，それがやはりもとの経験的関係系 α を表現しているなら，変換 ϕ はもとの表現 f に対して許容的であるといい，尺度は，許容的な変換の型が，それ自身のみであるか（絶対尺度），正の実定数との積であるか（比例尺度），正の実定数との積に定数を加えた一次変換であるか（間隔尺度），単調関数なら何でも良いのか（順序尺度），１対１対応の写像なら何でも良いのか（名義尺度），によって５種類に分けられる。

[*2] これは Stevens が提唱した分類だが，上述のごとく，絶対尺度を加えて５つとする分類もある。

が多く，比尺度は連続変数である。定性的変数と離散変数の中には，1か0，あるいは1か2，のように，2種類の値しかとらない「2分変数 (dichotomous variable)」や，1か2か3，のように3種類の値しかとらない「3分変数 (trichotomous variable)」がある。変数がとり得る値の範囲を，その変数の**定義域**と呼ぶ。変数は，被験者や研究対象のちがいによって，複数の異なったカテゴリあるいは数値に分かれるのでなければ意味がない。例えば，その研究のすべての対象者が男性であれば，性別という変数を作ることは無意味である。

対応する尺度の種類によって，変数は，図示の仕方も違うし，代表値も違うし，適用できる統計解析手法も違ってくる。尺度についてより詳しく知りたい方には，池田央『調査と測定』（新曜社）をお薦めする。

3.2 名義尺度 (nominal scale)

- 値の差も値の順序も意味をもたず，たんに質的データの分類基準を与える。
- 例えば，性別とか職業とか居住地といったものは，名義尺度である。
- 性別という名義尺度をあらわす変数は，例えば，男性なら M，女性なら F という具合に文字列値をとることもできるが，一般には男性なら 1，女性なら 2 というように，数値を対応させる。これは，第1回の講義で触れたとおり，コーディング（coding）と呼ばれる手続きである。関心のある事象が，例えば血液中のヘモグロビン濃度のように，性別ばかりでなく，授乳や妊娠によって影響を受ける場合は，調査対象者を，男性なら 1，授乳も妊娠もしていない女性は 2，授乳中の女性は 3，妊娠中の女性は 4，という具合に，生殖状態（性別及び授乳，妊娠）という名義尺度をあらわす変数にコード化する場合もある[*3]。
- 名義尺度を表す値にはそれを他の値と識別する意味しかない。統計解析では，クロス集計表を作って解析する他には，グループ分けや層別化に用いられるのが普通である[*4]。

[*3] このようにコーディングのやり方は一通りに限ったものではなく，分析の目的によって多様である。場合によっては再コーディングが必要となることもある。ここで注意すべきは，性別という名義尺度と，生殖状態という名義尺度は，別の尺度だということである。しかし，男性を M，女性を F と表しても，男性を 1，女性を 2 と表しても，1対1変換である限り，それは同じ性別という名義尺度である。

[*4] より複雑な統計解析に使う場合は，ダミー変数として値ごとの2分変数に変えることもある。例えば，居住地という変数の定義域が {東京，長野，山口} であれば，この変数の尺度は名義尺度である。東京を1，長野を2，山口を3と数値を割り振っても，名義尺度であるには違いない。しかし，居住地という変数を無くして，代わりに，東京に住んでいるか（1）いないか（0），長野に住んで

3.3 **順序尺度** (ordinal scale)

- 値の差には意味がないが，値の順序には意味があるような尺度。
- 例えば，鉱物の強度，地震の震度，尿検査でのタンパクの検出の程度について+++，++，+，±，− で表される尺度[*5]，「好き」「普通」「嫌い」に 3, 2, 1 点の得点を割り付けた尺度などは，順序尺度である。
- 順序尺度を表す値は，順序の情報だけに意味があるので，変数の定義域が 3, 2, 1 であろうと，15, 3.14159265358979, 1 であろうと同じ意味をもつ。しかし，順序の情報としては，1 から連続した整数値を割り当てるのが通例であり（同順位がある場合の扱いも何通りか提案されている），その場合に使えるノンパラメトリックな統計手法が数多く開発されている（順位相関や順位和検定など）。順序尺度を表す変数の平均値[*6]を求めることには意味がないが，中央値[*7]には意味がある。
- ただし，もっともらしい仮定を導入して間隔尺度であるとみなし，平均や相関を計算することも多い。例えば，「好き」「普通」「嫌い」の 3, 2, 1 とか，「まったくその通り」「まあそう思う」「どちらともいえない」「たぶん違うと思う」「絶対に違う」の 5, 4, 3, 2, 1 などは本来は順序尺度なのだが，等間隔であるという仮定をおいて間隔尺度として分析される場合が多い。質問紙調査などで，いくつかの質問から得られるこのような得点の合計によって何らかの傾向を表す合成得点を得ることが頻繁に行われるが，得点を合計する，という操作は各質問への回答がすべて等間隔であるという仮定を置いているわけである。合成得点が示す尺度の信頼性を調べるためにクロンバックの α 係数という統計量がよく使われるが，α 係数の計算には平均や分散が使われていることからも，それが間隔尺度扱いされていることがわかる。

いるか（1）いないか（0），という 2 つのダミー変数を導入することによって，同じ情報を表現することができる。ダミー変数を平均すると，1 に当てはまるケースの割合になる性質をもつために，ダミー変数は多くの統計手法の対象になりうる。

[*5] +の数を数値として，例えば 3, 2, 1, 0.5, 0 とコーディングしても，3 と 2 の差と 2 と 1 の差が等しいわけではなく，3 は 2 よりも尿タンパクが高濃度に検出され，2 は 1 よりも高濃度だという順序にしか意味がないから，順序尺度である

[*6] 次章で説明するが，ここでは，全部の値を足し合わせて値の数で割ったもの，と普通に考えておけば良い。

[*7] これも次章で説明するが，ここでは，小さいほうから順番に値を並べて，ちょうど中央にくるものと考えれば良い

3.4 **間隔尺度** (interval scale)

- 値の差に意味があるが，ゼロに意味がない尺度。[*8]
- 例えば，摂氏温度や西暦年は，間隔尺度である。気温が摂氏 30 度であることは，摂氏 10 度より摂氏 20 度分，温度が高いことを意味するが，3 倍高いことは意味しない。しかし，平年なら最高気温が摂氏 20 度であるようなときに摂氏 30 度であれば，摂氏 25 度であるのに比べて，平年との差が 2 倍あるとは言って良い。
- 間隔尺度をもつ変数に対しては，平均や相関など，かなり多くの統計手法が適用できるが，意味をもたない統計量もある。[*9]

3.5 **比尺度** (ratio scale)

- 値の差に意味があり，かつゼロに意味がある尺度。[*10]
- 例えば，cm 単位で表した身長とか，kg 単位で表した体重といったものは，比尺度である。予算額といったものも，0 円に意味がある以上，比尺度である[*11]。

3.6 **データの図示**

データの大局的性質を把握するには，図示をするのが便利である。人間の視覚的認識能力は，パターン認識に関してはコンピュータより遥かに優れていると言われているから，それを生かさない手はない。

[*8] より正確に言えば，値の比に意味がない尺度ということになる。ただし，値の差の比には意味がある。

[*9] 例えば，標準偏差を平均値で割った値を%表示したものを変動係数というが，身長という変数でも，普通に cm 単位や m 単位やフィート単位で表した比尺度なら変動係数に意味があるが，100cm を基準とした cm 単位や，170cm を基準とした 2cm 単位のように間隔尺度にしてしまった場合の変動係数には意味がない。変動係数は，分布の位置に対する分布のばらつきの相対的な大きさを意味するので，分布の位置がゼロに対して固定されていないと意味がなくなってしまうのである。

[*10] より正確に言えば，値の比にも意味がある尺度ということになる。

[*11] 予算額には 0 円やマイナスが普通にありえるし，何%成長とか何%削減という扱いより絶対値の増減が問題にされる場合が多いので間隔尺度とすべきという見方もありうる。

変数が表す尺度の種類によって，さまざまな図示の方法があるので，それをざっと示すことにする。

3.6.1 離散変数の場合

- 度数分布図：値ごとの頻度を縦棒として，異なる値ごとに，この縦棒を横に並べた図である。離散変数の名前を X とすれば，R では `barplot(table(X))` で描画される。
- 積み上げ棒グラフ：値ごとの頻度の縦棒を積み上げた図である。R では

  ```
  fx <- table(X)
  barplot(matrix(fx,NROW(fx)),beside=F)
  ```

  で描画される。
- 帯グラフ：横棒を全体を 100 %として各値の割合にしたがって区切って塗り分けた図である。R では

  ```
  px <- table(X)/NROW(X)
  barplot(matrix(pc,NROW(pc)),horiz=T,beside=F)
  ```

  で描画される。
- 円グラフ（ドーナツグラフ・パイチャート）：円全体を 100 %として，各値の割合にしたがって中心から区切り線を引き，塗り分けた図である。ドーナツグラフでは 2 つの同心円にして，内側の円内を空白にする。R では `pie()` 関数を用いる[*12]。

3.6.2 連続変数の場合

- ヒストグラム：変数値を適当に区切って度数分布を求め，分布の様子を見るものである。R では `hist()` 関数を用いる。
- 正規確率プロット：連続変数が正規分布しているかどうかを見るものである（正規分布に当てはまっていれば点が直線上に並ぶ）。R では `qqnorm()` 関数を用いる。

[*12] R-1.5 以前は `piechart()` 関数だったが置き換えられた

- 幹葉表示 (stem and leaf plot)：大体の概数（整数区切りとか 5 の倍数とか 10 の倍数にすることが多い）を縦に並べて幹とし，それぞれの概数に相当する値の細かい部分を葉として横に並べて作成する図。R では `stem()` 関数を用いる。
- 箱ヒゲ図 (box and whisker plot)：データを小さい方から順番に並べて，ちょうど真中にくる値を中央値 (median) といい，小さい方から 1/4 の位置の値を第 1 四分位 (first quartile)，大きいほうから 1/4 の位置の値を第 3 四分位 (third quartile) という。縦軸に変数値をとって，第 1 四分位を下に，第 3 四分位を上にした箱を書き，中央値の位置にも線を引いて，さらに第 1 四分位と第 3 四分位の差（四分位範囲）を 1.5 倍した線分をヒゲとして第 1 四分位の下と第 3 四分位の上に伸ばし，ヒゲの先より外れた値を外れ値として○をプロットした図である。カテゴリによって層別した箱ヒゲ図を横に並べて描くと，大体の分布の様子と外れ値の様子が同時に比較できるので便利である。R では `boxplot()` 関数を用いる。
- レーダーチャート：複数の連続変数を中心点から放射状に数直線としてとり，データ点をつないで表される図である。それら複数の変数によって特徴付けられる性質のバランスをみるのに役立つ。1 つのケースについて 1 つのレーダーチャートができるので，他のケースと比較するには，並べて描画するか，重ね描きする。R では `stars()` 関数を用いる。
- 散布図 (scatter plot)：2 つの連続変数の関係を 2 次元の平面上の点として示した図である。R では `plot()` 関数を用いる。異なる群ごとに別々のプロットをしたい場合は `plot()` の `pch` オプションで塗り分けたり，`points()` 関数を使って重ね打ちしたりできる。点ごとに異なる情報を示したい場合は `symbols()` 関数を用いることができるし，複数の連続変数間の関係を調べるために，重ね描きしたい場合は `matplot()` 関数と `matpoints()` 関数を，別々のグラフとして並べて同時に示したい場合は `pairs()` 関数を用いることができる。データ点に文字列を付記したい場合は `text()` 関数が使えるし，マウスで選んだデータ点にだけ文字列を付記したい場合は `identify()` 関数が使える。

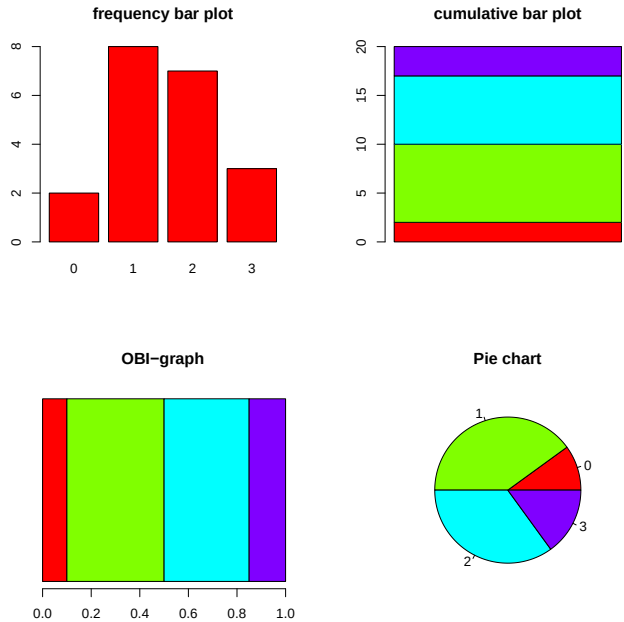


図 3.1: 離散変数の図示の例

3.7 離散変数の図示の例

20 組の夫婦について，その子ども数が，2, 3, 1, 0, 3, 2, 2, 1, 1, 1, 2, 2, 1, 3, 2, 1, 0, 2, 1, 1 だった場合，たとえば図 3.1 のような図示ができる[*13]

[*13] R のプログラムは下記の通り。

```
child <- c(2, 3, 1, 0, 3, 2, 2, 1, 1, 1, 2, 2, 1, 3, 2, 1, 0, 2, 1, 1)
fc <- table(child)
pc <- fc/sum(fc)
op <- par(mfrow=c(2,2))
barplot(fc, main="frequency bar plot")
barplot(matrix(fc,NROW(fc)), beside=F, main="cumulative bar plot", col=rainbow(NROW(fc)))
```

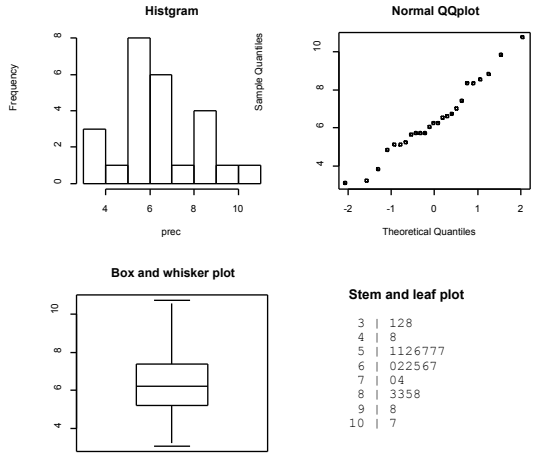


図 3.2: 離散変数の図示の例

3.8　連続変数の図示の例

平成元年 3 月 9 日から 4 月 2 日の東京地区の最低気温（℃）が次のようであったとする。

3.2, 3.1, 5.1, 4.8, 8.3, 9.8, 8.3, 6.6, 5.1, 3.8, 5.2, 5.6, 6.5, 5.7, 5.7, 7.4, 6.2, 7.0, 6.7, 5.7, 6.2, 6.0, 8.8, 10.7, 8.5

このとき，次の R のプログラムを使えば，図 3.2 のような図示ができる。

```
prec <- c(3.2, 3.1, 5.1, 4.8, 8.3, 9.8, 8.3, 6.6, 5.1, 3.8,
```

```
barplot(matrix(pc,NROW(pc)), horiz=T, beside=F, main="OBI-graph", col=rainbow(NROW(pc)))
pie(pc, main="Pie chart", col=rainbow(NROW(pc)))
par(op)
```

```
5.2, 5.6, 6.5, 5.7, 5.7, 7.4, 6.2, 7.0, 6.7, 5.7, 6.2,
6.0, 8.8, 10.7, 8.5)
op <- par(mfrow=c(2,2))
hist(prec, main="Histgram")
qqnorm(prec, main="Normal QQplot")
stem(prec,2)
boxplot(prec,main="Box and whisker plot")
par(op)
```

第 4 章

データを１つの値にまとめる（代表値）

4.1 ２つの戦略

データを１つの値にまとめるとは，分布の特徴を１つの値で代表させる，ということである。このような値を，代表値と呼ぶ。代表値は，記述統計量 (descriptive statistics) の１つである。

代表値を求める単純な戦略としては，誰でも思いつくだろうが，２つが考えられる。分布の位置と，分布の広がりである。例えば，正規分布だったら，$N(\mu, \sigma^2)$ という形で表されるように，平均 μ，分散 σ^2 という２つの値によって分布が決まるわけだが，この場合，μ が分布の位置を決める情報で，σ^2 が分布の広がりを決める情報である。

一般に，調査データは，仮想的な母集団からの標本（サンプル）[*1]と考えられ，データから計算される代表値は，母集団での分布の形を推定するために使われる。その意味で，これらの代表値は母数（パラメータ）と呼ばれる。

分布の位置を示す代表値は central tendency（中心傾向）と呼ばれ，分布の広がりを示す代表値は variability（ばらつき）と呼ばれる。

本章で以下用いる例題は，Grimm (1993) の第３章と第４章から引用してアレンジしたものが多い。代表値のような基礎的なことについてきちんと説明された教科書は意外に少ない中で，Grimm の本は丁寧に書かれていて，名著といってよい。英文も平易なのでお薦めする。

[*1] サンプリング理論については，統計学というよりは調査法や実験計画法の範疇になるので，それらの成書を参照されたい。

4.2 中心傾向（Central Tendency）

4.2.1 平均値（mean）

平均値は，分布の位置を示す指標として，もっとも頻繁に用いられる。実験的仮説検証のためにデザインされた式の中でも，頻繁に用いられる。記述的な指標の 1 つとして，平均値は，いくつかの利点と欠点をもっている。日常生活の中でも平均をとるという操作は普通に行われるから説明不要かもしれないが，数式で書くと以下の通りである。

母集団の平均値 μ（ミューと発音する）は，

$$\mu = \frac{\sum X}{N}$$

である。X はその分布における個々の値であり，N は値の総数である。$\sum$（シグマと発音する）は，一群の値の和を求める記号である。すなわち，$\sum X = X_1 + X_2 + X_3 + ... + X_N$ である。

標本についての平均値を求める式も，母集団についての式と同一である。ただし，数式で使う記号が若干異なっている。標本平均 $\bar{X}$（エックスバーと発音する）は，

$$\bar{X} = \frac{\sum X}{n}$$

である。n は，もちろん標本サイズである[*2]。

例題 1

値が {5, 8, 10, 11, 12} である母集団の平均値はいくらか？

値が 5 つしかない母集団というものは想像しにくいかもしれないが，$\mu = (5 + 8 + 10 + 11 + 12)/5 = 9.2$ であることは，小学生でもわかるだろう。R で平均値を計算するには，`mean()` という関数を使う。たとえば，例題 1 の解を得るには，`mean(c(5,8,10,11,12))` とすればよい。

さて，本章で取り上げる Central Tendency には，平均値の他に，あと 2 つ，中央

[*2] 記号について注記しておくと，集合論では $\bar{X}$ は集合 X の補集合の意味で使われるが，代数では確率変数 X の標本平均が $\bar{X}$ で表されるということである。同じような記号が別の意味で使われるので混乱しないように注意されたい。補集合は X^C という表記がなされる場合も多いようである。標本平均は $\bar{X}$ と表すのが普通である。

値 (median) と最頻値 (mode) がある。どれも分布の中心の位置がどの辺りかを説明するものだが，中心性 (centrality) へのアプローチが異なっている。

平均値は，中心性を示すために，どんなやり方をとっているのだろうか？　たまたまその値が平均値と同じであったという希な値を除けば，各々の値は，平均値からある距離をもって存在する。言い換えると，各々の値は，平均値からある程度の量，ばらついている。ある値が平均値から離れている程度は，単純に $X - \bar{X}$ である。この，平均からの距離を，偏差（あるいは誤差）といい，x という記号で書く。つまり，$x = X - \bar{X}$ である。次の例を見ればわかるように，偏差は正の値も負の値もとるが，その合計は 0 になるという特徴をもつ。どんな形をしたどんな平均値のどんなに標本数が多い分布だろうと，偏差の和は常に 0 である。式で書くと，$\sum x = \sum(X - \bar{X}) = 0$ ということである。言い方を変えると，偏差の和が 0 になるように，平均値によって調整が行われたと見ることもできる。平均値は，この意味で，分布の中心であるといえる。

例題 2

分布 A が {2,4,6,8,10} という 5 つの値をもち，分布 B が {2,4,6,8,30} という 5 つの値をもっているとき，分布 A の標本平均は 6 であるから，それぞれの値の偏差は $\{-4, -2, 0, 2, 4\}$ となり，その合計は 0 である。分布 B についても確かめよ。

標本平均は (2+4+6+8+30)/5=10 で，それぞれの値の偏差は $\{-8, -6, -4, -2, 20\}$ となるので，確かにその合計は 0 となる。分布 B は分布 A よりも平均値が大きい。

4.2.2 **重み付き平均** (weighted mean)

重み付き平均は，各々の値にある重みをかけて合計したものを，重みの合計で割った値である。式で書くと，

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + ... + n_n(\bar{X}_n)}{n_1 + n_2 + ... + n_n}$$

ここでは標本サイズが異なる複数の平均値の総平均 (grand mean) を計算する場合について説明する。

Y大学の 3 つの学部の学生が TOEIC を受験したところ，学部ごとの得点の平均値がA学部 440 点，B学部 470 点，C学部 410 点だったとしよう。これらの値からY大学全体の TOEIC の平均点を求めたいときは，どうしたらよいだろうか？　単純にこれらを足して 3 で割った 440 点としていいのだろうか？

大学全体としての TOEIC の平均値は，3 つの学部のどこに所属する学生であるかにかかわりなく，全員の得点を足して，その人数で割って得るべきものである。そうだとすれば，単純に 3 つの値を足して 3 で割るのでは具合が悪い。各学部の人数は異なるので，人数の多い学部の得点の方が，総平均には余計に影響するだろうからだ。こういう場合は，各学部の人数をそれぞれの平均点に掛けて（つまり各学部の得点総和に戻して）足し合わせ，それを人数の和（つまり大学全体の人数）で割れば良いことが直感的にわかるだろう。これが重み付き平均の発想である。

例題 3

TOEIC の平均点が {440, 470, 410} であった 3 つの学部それぞれの人数が {200 人，100 人，300 人} であったなら，この大学の TOEIC の総平均は何点か？

$(200 \times 440 + 100 \times 470 + 300 \times 410)/(200 + 100 + 300) = 430$ より，430 点となる。R で実行するときは，

```
p <- c(440,470,410)
n <- c(200,100,300)
gm <- sum(p*n)/sum(n)
print(gm)
```

とすると見やすい。

練習問題

3 つの年齢群ごとの平均血圧が下の表のように記録されているとき，すべての年齢群をプールした，血圧の総平均値を求めよ[a]。

	年齢		
	20-39	40-59	60+
収縮期血圧 (mmHg)	118	128	145
拡張期血圧 (mmHg)	70	78	82
人数	13	12	16

[a] 但し，血圧の意味合いは年齢によって変わってくるからこそ，ふつう敢えて年齢群別に平均値を出すわけだから，年齢群をプールした血圧の平均値を出すことには，あまり意味はない。ここは単なる計算練習だと思って欲しい。また，ここで述べたような意味での重み付き平均を計算する必要があるのは，集計済みの二次資料から指標値を再計算するような場合なので，生データがあれば生データから計算すれば済むことである。

収縮期血圧の総平均値は，$(118 \times 13 + 128 \times 12 + 145 \times 16)/(13 + 12 + 16) = 131$ よ

り，131 mmHg となり，拡張期血圧の平均値は，$(70 \times 13 + 78 \times 12 + 82 \times 16)/(13 + 12 + 16) = 77$ より，77 mmHg となる。これも R で実行するときは，例題 3 と同様に，

```
SBP <- c(118,128,145)
DBP <- c(70,78,82)
N <- c(13,12,16)
gSBP <- sum(SBP*N)/sum(N)
gDBP <- sum(DBP*N)/sum(N)
cat("SBP 総平均=",gSBP,", DBP 総平均=",gDBP,"\n")
```

とするとわかりやすい。

4.2.3 度数分布の平均

度数分布の平均も，重み付き平均に似た概念である。離散変数の平均の場合に，度数分布を出して，各値にその度数を掛けたものの和を度数の総和で割ることで得られる。これは，言い換えると，度数で重み付けした平均値である。

$$\mu = \frac{\sum Xf}{\sum f}$$

という式になる。

平均値は，例題 2 を見ればわかるように，少数の極端な値の影響を受けやすいという欠点をもつ。1 つだけ極端な値があったからといって，あまりに値がそちらに引っ張られてしまっては，分布の位置を代表する値としては具合が良くない[*3]。

[*3] その 1 つが，実は測定ミスであったり，異質な対象だったりして，外れ値である場合もあり，その場合は平均値の計算に入れないこともある。あまり機械的にやるのは良くないが，ネイマンの外れ値の検定を使うのも一案である。

例題 4

A 大学の学長選挙で，B 氏が，A 大学の研究水準を上げるという公約を掲げて当選したとしよう。4 年後の次の選挙のときに，B 氏は自分が公約を果たしたと宣伝したいわけだが，彼の定義によると，大学の研究水準が上がるとは，教員の論文数の平均値が増えるということである。ところで，B 氏が当選した当時の教員数は 100 人いて，そのうち発表論文数が 5 本の人が 80 人，10 本の人が 15 人，30 本の人が 5 人いたとしよう。この時点での平均論文数は $(5 \times 80 + 10 \times 15 + 30 \times 5)/100 = 7$ なので 7 本である。その後 4 年間誰も 1 本も論文を書かなかったとしても，2 年目にたまたま 2330 本の論文をもつ教員が 1 人着任したら，何が起こるかを考えてみよう。

平均論文数は $(5 \times 80 + 10 \times 15 + 30 \times 5 + 2330)/101 = 30$ から，30 本となってしまう。そこで，B 氏は，大威張りで，任期中に平均論文数は 4 倍以上に増えたと報告することができる。元々 A 大学にいた教員の論文数はまったく変わらず，従って大した研究環境を提供できていないと思われるにもかかわらず，である。B 氏が公約を果たしたと宣伝しても嘘ではないことになるが，何か妙である。

例題 4 は，極端に高い値が，平均値を高く押し上げてしまったという例である。分布の位置の指標としては，極端な外れ値に対してこんなに敏感であっては具合が良くない。こういう極端な値が含まれている歪んだ分布の場合には，平均値という指標は誤解を生んでしまうので，相応しくないことになる。

4.2.4 中央値（median）

そこで登場するのが**中央値**である。中央値は，全体の半分がその値より小さく，半分がその値より大きい，という意味で，分布の中央である。言い換えると，中央値は，頻度あるいは値の数に基づいて分布を 2 つに等分割する値である。中央値を求めるには式は使わない（決まった手続き＝アルゴリズムとして，並べ替え (sorting) は必要）。極端な外れ値の影響を受けにくい（言い換えると，外れ値に対して頑健である）。歪んだ分布に対する最も重要な central tendency の指標が中央値である。

例題 5

次の分布の中央値は何か？　{1, 4, 6, 8, 40, 50, 58, 60, 62}

この場合，小さい方から数えても大きいほうから数えても 5 番目の値である 40 が中央値であることは自明である。次に小さい値である 50 との距離や次に大きい値で

ある 8 との距離は中央値を考える際には無関係である。中央値を求めるには，値を小さい順に並べ替えて[*4]，ちょうど真中に位置する値を探せばよい。この意味で，中央値は値の順序だけに感受性をもつ（= rank sensitive である）といえる[*5]。

R で中央値を計算するには，`median()` という関数を使う。たとえば，例題 5 の解を得るには，`median(c(1, 4, 6, 8, 40, 50, 58, 60, 62))` とすればよい。

例題 6

次の標本分布の平均値と中央値は何か？　{2, 4, 7, 9, 12, 15, 17}

R で

```
x <- c(2,4,7,9,12,15,17)
mean(x)
median(x)
```

とすると，平均値は約 9.43，中央値は 9 であるとわかる。

例題 7

次の標本分布の平均値と中央値は何か？　{2, 4, 7, 9, 12, 15, 17, 46, 54}

例題 6 と同様に計算すると，平均値は 18.4，中央値は 12 となる。例題 6 に比べると，右側に 2 つの極端な値を加えただけだが，平均値はほぼ倍増してしまう。それに対して，中央値は 1 つ右側の値に移るだけであり，中央値の方が極端な値が入ることに対して頑健といえる。

ところで，値の数が奇数だったら，このように順番が真中というのは簡単に決められるが，値が偶数個だったらどうするのだろうか？

例題 8

次の分布の中央値は何か？　{4, 6, 9, 10, 11, 12}

中央値が 9 と 10 の間にくることは明らかである。そこで，普通は 9 と 10 を平均した 9.5 を中央値として使うことになっている。もっとも，本来整数値しかとらないような値について，中央値や平均値として小数値を提示することに意味があるかどうかは問題である。例えば，例題 8 の分布が，ある地方の水泳プールで 6 日間観察した

[*4] 値の数が少ない場合には，手作業で並べ替えを行えばよいが，大量のデータを手作業で並べ替えるのは大変である。コンピュータのプログラムに値を並べ替えさせるアルゴリズムには，単純ソート，バブルソート，シェルソート，クイックソートなどがある。

[*5] 平均値は値の大きさによって変わるので，value sensitive であるといえる。

ときの，1 日当たりの飛び込みの回数を示すものだとしよう。中央値が 9.5 ということになると，9.5 回の飛び込みというのは何を表すのか？　半分だけ飛び込むということはありえない。つまり実体はない，単なる指標値だということになる。同様に平均値についても，世帯当たりの平均子ども数が 2.4 人とかいうとき，0.4 人の子どもは実体としてはありえない。しかし，分布の位置を示す指標としては有用なので，便宜的に使っているのである。

例題 9

次の分布の中央値は何か？　{7, 7, 7, 8, 8, 8, 9, 9, 10, 10}

例題 10

次の分布の中央値は何か？　{7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10}

このように同順位の値（tie という）がある場合は，事態はやや複雑である。順番で言えば，例題 9 でも例題 10 でも中央値は 8 と 8 の間に来るはずだから，8 と思うであろう。実際，SAS，SPSS などの有名ソフトを初めとして，MS-Excel や R に至るまで，ほぼすべての統計ソフトは，8 という答えを出してくるし，一般にはそれで問題ない。[*6]

[*6] ただし，厳密に考えると，簡単に 8 と言えない。Grimm (1993) が指摘するように，分布の値を示す数値は，間隔の中点と考えるべきだからである。普通はそこまで厳密に考える必要はないが，参考までに説明しておこう。

要点は，『それぞれの値を，表示単位によって規定される区間の中点と考え，同順位の値があるときは，それが区間内に均等に散らばると考える』ということである。これは直感的に考えても合理的であろう。

たとえば，1 1 1 2 2 2 3 3 3 という，表示単位 1 のデータがあるとき，真の値がそれぞれ等間隔に散らばっているならば，0.67 1.00 1.33 1.67 2.00 2.33 2.67 3.00 3.33 と考えるのが自然である。これなら，それぞれの値が 1/3 間隔になっているし，中点 1 で示される値 0.67 1.00 1.33 の平均は 1 となるので，どこにも矛盾がない。

この例から帰納的に考えて，その区間の下限の値を L とし，階級幅を h とし，同順位の個数を fm 個とし，1 つ下の区間までに F 個のサンプルがあるとすれば，F+1 番目, F+2 番目, ..., F+fm 番目 の値はそれぞれ，L+1/(2fm)*h, L+3/(2fm)*h, ..., L+(2fm-1)/(2fm)*h となる。つまり，**F+x 番目の値は，L+(2x-1)/(2fm)*h** となる。

この式から例題 9 の 3 つの 8 の真の値がいくつになるか計算すると，

4 番	5 番	6 番
7.67	8.00	8.33

となって，5 番と 6 番の間は 8.17 となる。

同じく例題 10 で真の値を計算すると，{6.67 7.00 7.33 7.60 7.80 8.00 8.20 8.40 8.75 9.25 9.75 10.25} となるので，中央値は 8.00 と 8.20 の間で 8.10 となる。{1 1 2 2 3 3} という表示単位 1 のデータでは，真の値は {0.75 1.25 1.75 2.25 2.75 3.25} と推定されるので，中央値は 1.75

さて，もう 1 歩進めて，度数分布表から中央値を計算する場合を考えてみよう。ちょっと複雑だが，理解するのは難しくない。下表は，年齢階級ごとの人数の分布であり，これから年齢の中央値を求める方法を考えることにする。

年齢階級	度数	累積度数
45-49	1	76
40-44	2	75
35-39	3	73
30-34	6	70
25-29	8	64
20-24	17	56
15-19	**26**	**39**
10-14	11	13
5-9	2	2
0-4	0	0

まず，累積度数の最大の数をみる（つまり総数をみる）。この例では 76 である。中央値の順位は $(76+1)/2 = 38.5$ 位となる[*7]。38.5 番目の値を含む年齢階級を探すと，15-19 である。そこで，単純に統計ソフトが出してくる中央値は 15-19 歳となる[*8]。

5 歳の階級幅の中のどこに中央値があるのかというところまで推定しようとなると，もう少し厳密に考えねばならなくなる。つまり，Grimm 流に 15-19 歳の 26 人の値が均等に散らばっていると考えると，{14.5+5/52, 14.5+15/52, 14.5+25/52, ..., 14.5+245/52, 14.5+255/52} となるから，38.5 位の値は，最後の 2 つの平均をとって，$14.5 + (245 + 255)/104 \approx 19.3$ から約 19.3 歳となる。

このやり方は，中央値が正確な分布の中央（少なくともその近似）になっているという特性を強めるものである。式で書けば，中央値は，

$$L + \left[\frac{N/2 - F}{f_m} \cdot h \right]$$

となる。ここで，L は中央順位を含む階級の正確な下限，F は中央順位を含む階級より下の値の総度数，f_m は中央順位を含む階級の度数，h は階級幅である。

この式は以下のように導かれる。

と 2.25 の平均で 2 となる。

[*7] Grimm (1993) には 76 を 2 で割って 38 番目の値が中央値であると書かれているが，論理的整合性を欠く。もし総数を 2 で割った順位の値が中央値だとすると，例題 8 の答えが下から 3 番目で 9 ということになってしまう。総数に 1 を加えて 2 で割らなくてはいけない。

[*8] 繰り返すが，普通はこの解で問題ない。

1. サンプル数 N が奇数のとき, $(N+1)/2$ 番目が中央値なので, $F+x=(N+1)/2$ を x について解いて $L+(2x-1)/(2f_m)*h$ に代入すれば,

$$L+(N+1-2F-1)/(2f_m)*h = L+(N/2-F)/f_m*h$$

となる。

2. N が偶数のとき, 中央値は $N/2$ 番目と $N/2+1$ 番目の間なので, $F+x=N/2$ と $F+x=N/2+1$ を x について解いて $L+(2x-1)/(2f_m)*h$ に代入した

$$L+(2(N/2-F)-1)h/(2f_m)$$

と

$$L+(2(N/2+1-F)-1)h/(2f_m)$$

の平均となって，やはり

$$L+(N/2-F)/f_m*h$$

で良いことになる。

4.2.5 最頻値（Mode）

残る最頻値は，きわめて単純である。もっとも度数が多い値を探すだけである。もっとも数が多い値が，もっとも典型的だと考えるわけである。データを見ると，最頻値が 2 つある場合があり，この場合は分布が二峰性 (bimodal) だという[*9]。すべての値の出現頻度が等しい場合は，最頻値は存在しない。

分布の形によって，平均値，中央値，最頻値の関係は変わってくる。歪んでいない分布ならば，ばらつきの程度によらず，これら 3 つの値は一致する。二峰性だと最頻値は 2 つに分かれるが，平均値と中央値はその間に入るのが普通である。左すそを引いた分布では，平均値が最も小さく，中央値が次で，最頻値が最も大きくなる。右すそを引いた分布では逆になる。

平均値は，(1) 分布のすべての値を考慮した値である，(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる，(3) 多くの統計量や検定で使われている，という特長をもつ。標本調査値から母集団の因果関係を推論したい場合に，もっとも

[*9] しかし隣り合う 2 つの値がともに最頻値である場合は二峰性だとはいわず，離れた 2 つの値が最頻値あるいはそれに近い場合，つまり度数分布やヒストグラムの山が 2 つある場合に，分布が二峰性だといい，2 つの異なる分布が混ざっていると考えるのが普通である。

普通に使われる。しかし，(1) 極端な外れ値の影響を受けやすい，(2) 打ち切りのある分布では代表性を失う場合がある[*10]，という欠点があり，外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。最頻値は，標本をとったときの偶然性の影響を受けやすいし，もっとも頻度が高い値以外の情報はまったく使われない。しかし，試験の点で何点の人が多かったかを見たい場合は最頻値が役に立つし，名義尺度については最頻値しか使えない。

ここで上げた 3 つの他に，幾何平均 (geometric mean) や調和平均 (harmonic mean) も，分布の位置の指標として使われることがある。幾何平均はデータの積の累乗根（対数をとって平均値を出して元に戻したもの），調和平均はデータの逆数の平均値の逆数であり，どちらもゼロを含むデータには使えない。大きな外れ値の影響を受けにくいという利点があり，幾何平均は，とくにデータの分布が対数正規分布に近い場合によく用いられる。

4.3 ばらつき（Variability）

分布を特徴付けるには，分布の位置だけではなく，分布の広がり具合の情報も必要である。例えば，図 4.1 の 2 つの分布は[*11]，どちらも平均 0 の正規分布なので中央値も最頻値も共通だが，実線で書かれた幅が狭い方が標準偏差 1，破線で書かれた幅が広い方が標準偏差 4 と，標準偏差が大きく異なるために，まったく違った外見になっている。標準偏差は，もっとも良く使われる分布の広がり具合の指標である。

広がり具合を示す指標は，ばらつき (variability) と総称される。ばらつきの指標には，範囲，四分位範囲，四分位偏差，平均偏差，分散（及び不偏分散），標準偏差（及び不偏標準偏差）がある。

[*10] 氷水で痛みがとれるまでにかかる時間とか，年収とか。無限に観察を続けるわけにはいかないし，年収は下限がゼロで上限はビル・ゲイツのそれのように極端に高い値があるから右すそを長く引いた分布になる。平均年収を出している統計表を見るときは注意が必要である。年収の平均的な水準は中央値で表示されるべきである。

[*11] この図を書くための R のプログラムは次の通り。

```
x <- c(1:1000)/100-5
z1 <- dnorm(x,0,1)
z2 <- dnorm(x,0,4)
plot(x,z1,type='l',lty=1,ylab='probability density',xlab='')
points(x,z2,type='l',lty=2)
```

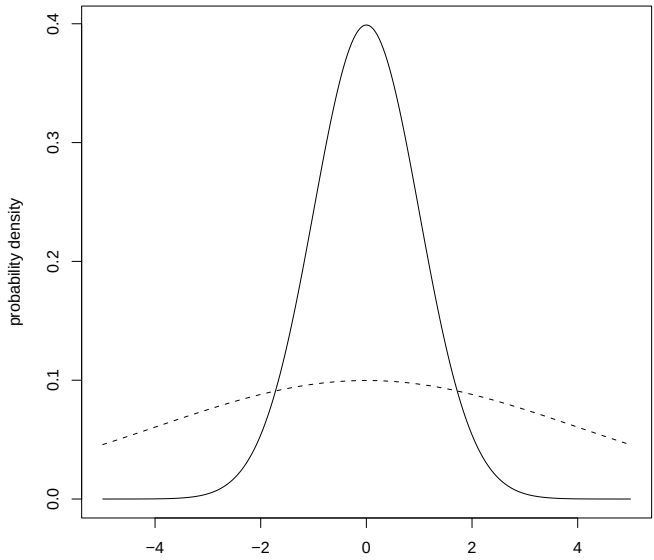


図 4.1: 平均が同じで標準偏差が異なる正規分布

4.3.1 **範囲** (range)

範囲は，最も単純なばらつきの尺度である。値のとる全範囲そのものである。つまり，最大値から最小値を引いた値になる。

例題 11

次の分布の範囲はいくらか？　$\{17, 23, 42, 44, 50\}$

いうまでもなく，$50 - 17 = 33$ である。ばらつきの尺度として範囲を使うには，若干の問題が生じる場合がある。極端な外れ値の影響をダイレクトに受けてしまうのである。次の例を考えてみよう。

例題 12

次の分布の範囲はいくらか？　{2, 4, 5, 7, 34}

答えは $34 - 2 = 32$ なのだが，2, 4, 5, 7 というきわめて近い値４つと，かけ離れて大きい 34 という値からなるのに，32 という範囲は，全体のばらつきが大きいかのような誤った印象を与えてしまう。ばらつきの指標としては，分布の端の極端な値の影響を受けにくい値の方がよい。

4.3.2 **四分位範囲** (Inter-Quartile Range; IQR)

そこで登場するのが四分位範囲である。その前に，分位数について説明しよう。値を小さい方から順番に並べ替えて，４つの等しい数の群に分けたときの 1/4, 2/4, 3/4 にあたる値を，四分位数 (quartile) という。1/4 の点が第１四分位，3/4 の点が第３四分位である（つまり全体の 25 %の値が第１四分位より小さく，全体の 75 %の値が第３四分位より小さい）。2/4 の点というのは，ちょうど順番が真中ということだから，第２四分位は中央値に等しい。

ちょっと考えればわかるように，ちょうど４等分などできない場合がもちろんあって，上から数えた場合と下から数えた場合で四分位数がずれる可能性があるが，その場合はそれらを平均するのが普通である。また，最小値，最大値に，第１四分位，第３四分位と中央値を加えた５つの値を五数要約値と呼ぶことがある。第１四分位，第２四分位，第３四分位は，それぞれ Q1, Q2, Q3 と略記することがある。

これを一般化して，値を小さい方から順番に並べ替えて，同数の群に区切る点を分位数 (quantile) という。百等分した場合を，とくにパーセンタイル (percentile) という。言い換えると，第１四分位は 25 パーセンタイル，第３四分位は 75 パーセンタイルである。

四分位範囲とは，第３四分位と第１四分位の間隔である。パーセンタイルでいえば，75 パーセンタイルと 25 パーセンタイルの間隔である。上と下の極端な値を排除して，全体の中央付近の 50 %（つまり代表性が高いと考えられる半数）が含まれる範囲を示すことができる。

4.3.3 **四分位偏差** (Semi Inter-Quartile Range; SIQR)

四分位範囲を２で割った値を四分位偏差と呼ぶ。もし分布が左右対称型の正規分布であれば，中央値マイナス四分位偏差から中央値プラス四分位偏差までの幅に全データの半分が含まれるという意味で，四分位偏差は重要な指標である。IQR も SIQR も

少数の極端な外れ値の影響を受けにくいし，分布が歪んでいても使える指標である。

例題 13

パプアニューギニアのある村で成人男性 28 人の体重を量ったところ，{50.5, 58.0, 47.5, 53.0, 54.5, 61.0, 56.5, 65.5, 56.0, 53.0, 54.0, 56.0, 51.0, 59.0, 44.0, 53.0, 62.5, 55.0, 64.5, 55.0, 67.0, 70.5, 46.5, 63.0, 51.0, 44.5, 57.5, 64.0}（単位は kg）という結果が得られた。このデータから，四分位範囲と四分位偏差を求めよ。

R の `fivenum()` 関数を使うと，Q1=52.00，Q2=55.50，Q3=61.75 とわかる。これより，四分位範囲は $Q3 - Q1 = 9.75$，四分位偏差はそれを 2 で割って 4.975 である。

4.3.4 平均偏差 (mean deviation)

偏差の絶対値の平均値を平均偏差と呼ぶ。四分位範囲や四分位偏差は，全データのうちの限られた情報しか使わないので，分布のばらつきを正しく反映しない可能性がある。そこで，すべてのデータを使ってばらつきを表す方法を考えよう。すべての生の値は，平均値からある距離をもって分布している。この距離は既に述べたように偏差あるいは誤差と呼ばれる。偏差の大きさは，分布のばらつきを反映している。

例題 14

分布 A が {11, 12, 13, 14, 15, 16, 17}，分布 B が {5, 8, 11, 14, 17, 20, 23} だとする。どちらも平均値は 14 である。しかし，分布 B は分布 A よりもばらつきが大きい。言い換えると，分布 B の方が分布 A よりも平均値からの距離が大きい。しかし，それをどうやって 1 つの値として表すことができるだろうか？

ただ合計しただけでは，平均値のところで述べたように，偏差の総和は必ずゼロになってしまう。これはマイナス側の偏差がプラス側の偏差と打ち消しあってしまうためなので，偏差の**絶対値の**総和を出してやればいいというのが最も単純な発想である。それだけだとサンプル数が多いほど大きくなってしまうので，値 1 つあたりの偏差の絶対値を出してやるためにサンプル数で割ることが考えられる。これが平均偏差の考え方である。

すなわち，平均偏差 MD は，

$$MD = \frac{\sum |X - \mu|}{N}$$

で定義される。μ は平均値，N はサンプル数である。例題 14 の場合，分布 A の平均偏差は約 1.71，分布 B の平均偏差は約 5.14 である。これらの値は，次の R プログラムによって計算される。

```
A <- c(11, 12, 13, 14, 15, 16, 17)
B <- c(5, 8, 11, 14, 17, 20, 23)
mA <- mean(A)
mB <- mean(B)
sum(abs(A-mA))/NROW(A)
sum(abs(B-mB))/NROW(B)
```

平均偏差はすべてのデータを使い，かつ少数の外れ値の影響は受けにくいという利点があるが，絶対値を使うために他の統計量との数学的な関係がなく，標本データから母集団統計量を推定するのに使えないという欠点がある。

4.3.5 **分散** (variance)

マイナス側の偏差とプラス側の偏差を同等に扱うためには，絶対値にするかわりに二乗しても良い。つまり，偏差の二乗和の平均をとるわけである。これが分散という値になる。分散 V は，

$$V = \frac{\sum (X - \mu)^2}{N}$$

で定義される[*12]。標本数 n で割る代わりに自由度 $n-1$ で割って，不偏分散 (unbiased variance) という値にすると，標本データから母集団の分散を推定するのに使える。即ち，不偏分散 V_{ub} は，

$$V_{ub} = \frac{\sum (X - \bar{X})^2}{n - 1}$$

である。

4.3.6 **標準偏差** (standard deviation)

分散の平方根をとったものが標準偏差である。平均値と次元を揃えるという意味をもつ。不偏分散の平方根をとったものは，不偏標準偏差となる。もし分布が正規分布

[*12] 実際に計算するときは 2 乗の平均から平均の 2 乗を引くと簡単である。

ならば，Mean±2SD[*13]の範囲にデータの 95%が含まれるという意味で，標準偏差は便利な指標である。

練習問題

例題 14 の 2 つの分布について，不偏分散と不偏標準偏差を計算せよ。

以下の R プログラムにより，A の不偏分散が 4.67，A の不偏標準偏差が 2.16，B の不偏分散が 42，B の不偏標準偏差が 6.48 だとわかる。

```
A <- c(11, 12, 13, 14, 15, 16, 17)
B <- c(5, 8, 11, 14, 17, 20, 23)
cat("A の不偏分散=",var(A)," / A の不偏標準偏差=",sd(A),"\n")
cat("B の不偏分散=",var(B)," / B の不偏標準偏差=",sd(B),"\n")
```

4.3.7　標準誤差 (standard error) と変動係数 (coefficient of variation)

生データの分布のばらつきの指標ではないが，関連するのでここで示しておく。不偏標準偏差を $\sqrt{N}$ で割った値は，平均値の推定幅を示す値となり[*14]，標準誤差 (standard error) として知られている。SD と SE は論文などでは良く混用されているが，意味がまったく違う。また，標準偏差（不偏標準偏差ではない）を平均値で割って 100 を掛けた値を変動係数という。即ち，平均値に対して，全測定値が何%ばらついているかを示す，相対的なばらつきの指標である。これは測定誤差を示すときなどに使われる値であり，母集団統計量である。

4.4　まとめ

データの分布は，位置とばらつきを示す 2 つの代表値に集約して示すのが普通である。分布に外れ値が多い・歪みが大きい・尺度水準が低いなどの理由で，分布を仮定できない場合は，中央値と四分位偏差を用い，そうでない場合は平均値と（不偏）標準偏差を用いて，位置±ばらつき，という形で示すのが普通である。

参考までに，MS-Excel と R による代表値の求め方を一覧形式でまとめておくの

[*13] 普通このように 2SD と書かれるが，正規分布の 97.5 パーセント点は 1.959964... なので，この 2 は，だいたい 2 くらいという意味である。

[*14] 平均値の分散は生データの分散の 1/N になることと，N が大きいとき，元の分布によらず平均値は正規分布に近づく（中心極限定理）ため。

で，必要に応じて参照されたい。

求める代表値など	EXCEL の関数または手順 （範囲 A1:Y1 にデータがあるとして）	R の関数または手順 （x <- c(...) などのやり方で変数 x にデータを入れたとして）
最頻値	離散データなら=MODE(A1:Y1) で良いが，連続量なら，ツール>分析ツール>ヒストグラムでヒストグラムを書いて最大度数のデータ区間を探し，その区間の中点を最頻値とする。	hist(x) でヒストグラムを書いて最大度数のデータ区間を探し，その区間の中点を最頻値とする。hist(x,c(min(x),5,8,max(x))) などとすれば，x の最小値から 5 まで，5 から 8 まで，8 から x の最大値までという 3 つの区間で度数を計算させることができる。本来は hist(x,5) とすれば 5 つの区間という形の指定ができるはずなのだが，区間の数によってうまくいったりいかなかったりした。 なお，hist(x,plot=F) とすれば，グラフを書く代わりに数値を表示させられる。
中央値	=MEDIAN(A1:Y1)	median(x) 但し，x の中に NA（欠損値）を含む場合は， median(x,na.rm=T) または median(x[!is.na(x)]) とする。以下同様。
平均値	=AVERAGE(A1:Y1) 調和平均は=HARMEAN(A1:Y1)，幾何平均は=GEOMEAN(A1:Y1) で求められる。	mean(x) 調和平均は 1/mean(1/x)，幾何平均は exp(mean(log(x))) で求められる。
範囲	=MAX(A1:Y1)−MIN(A1:Y1)	max(x)-min(x)
四分位範囲	=QUARTILE(A1:Y1,3) − QUARTILE(A1:Y1,1)	IQR(x) または，y <- quantile(x); y[4] - y[2] または，fivenum(x)[4] - fivenum(x)[2] でも良い。
四分位偏差	=(QUARTILE(A1:Y1,3) − QUARTILE(A1:Y1,1))/2	IQR(x)/2 または，y <- quantile(x); (y[4] - y[2])/2 または，(fivenum(x)[4] - fivenum(x)[2])/2 でも良い。
平均偏差	=AVEDEV(A1:Y1)	組み込み関数にはないが，sum(abs(x-mean(x)))/NROW(x) で得られる。
不偏分散	=VAR(A1:Y1) （不偏でない分散は=VARP(A1:Y1) で得られる）	var(x) 不偏でない分散は組み込み関数にはないが，sum((x-mean(x))^2)/NROW(x) で得られる。
不偏標準偏差	=STDEV(A1:Y1) （不偏でない標準偏差は=STDEVP(A1:Y1) で得られる）	sd(x) 不偏でない標準偏差は，sqrt(sum((x-mean(x))^2)/NROW(x)) で得られる。(*)
タブ区切りデータファイルの読み込み	そのままドラッグ&ドロップ	1 行目に変数名が入っているなら，x <- read.delim("d:/sample.dat" ,header=T) などとする (**)。 それぞれの変数は，例えば x$age のようにして参照できる。1 行目が変数名でなくすぐにデータである場合は，x <- read.delim("d:/sample.dat" ,header=F) とする。この場合，変数名は x$V1，x$V2，... として参照できる。いちいち x$とつけるのが面倒なら，attach(x) とすれば V1 とか V2 だけで参照できる。
カンマ区切りデータファイルの読み込み	そのままドラッグ&ドロップ	1 行目に変数名が入っているなら，x <- read.csv("d:/sample.dat" ,header=T) とする (**)。1 行目が変数名でなくすぐにデータである場合は，x <- read.csv("d:/sample.dat" ,header=F) とする。
データの編集	表にそのまま打ち込む	de(x$V1,x$V5) などとすれば表形式で指定した変数の値を編集できる。表の上でマウスを右クリックすると操作メニューがでる。
データの書き出し	ファイルから保存を選ぶ	コンマ区切りでデータフレーム x をマイドキュメントの sample.dat に書き出すには，write.table(x,"d:/sample.dat", sep=",") とする。タブ区切りなら sep="\t"とすればよい。

(*) もちろん，不偏でない分散を出すときに，Vx<-sum((x-mean(x))^2)/NROW(x) などとして値を保存しておいて，sqrt(Vx) とするのがエレガントである。

(**) \を/に置き換えたファイル名をフルパスで書く。ただし，2 バイトコードが入ったディレクトリ名やファイル名は，文字化けするので使いにくい（半角英数字のファイル名に書き換えておくべきである）。また，Windows2000 の場合，マイドキュメントフォルダのフルパスは，普通，C:/Documents and Settings/nakazawa/My Documents/のようになるが，長いパスを打つのは面倒なので，D:ドライブのルートディレクトリにデータを置くと，指定が容易である。

第 5 章

比率に関する推定と検定

5.1 母比率を推定する方法

今回は，名義尺度や順序尺度をもつカテゴリ変数を分析する方法に入る。まずは変数が１つの場合を考える。カテゴリ変数１つがもっている情報は，データ数と，個々のカテゴリが占める割合（標本比率）である。したがって，このデータから求める統計的な指標は，母比率，即ち個々のカテゴリが母集団で占めるであろう割合である。通常，標本比率とほぼ一致する。

例えば，手元の容器の中に，数百個の白い碁石があるとする。この概数を手っ取り早く当てるために，数十個の黒い碁石を混ぜる。よくかき混ぜてから 20 個程度の石を取り出してみて（標本），その中で黒い石が占めていた割合（標本比率）を求め，それが母比率と等しいと仮定して加えた黒い碁石の数を割って総数を求め，黒い碁石の数を引けば，元々の白い碁石の数が得られる。生態学で，野原のバッタの数を調べたいときに全数を調べるわけにはいかないので，捕まえてペンキでマークして放して暫く経ってからまた捕まえてマークされているバッタの割合を求めて，マークした数をそれで割って総数を推定する，というリンカーン法（Capture-Mark-Recapture ともいう）のやり方と同じである。

例題１．

最初に混入した黒い石の数が 40 個，かき混ぜてから 20 個の石を取り出してみたら黒石 2 個，白石 18 個だった場合，元の白石の数はいくつと推定されるか？

元の白石の数を x とすると，40/(40+x)=2/(2+18) となるので，これを x について解けば，x=360 が得られる。したがって 360 個と推定される。

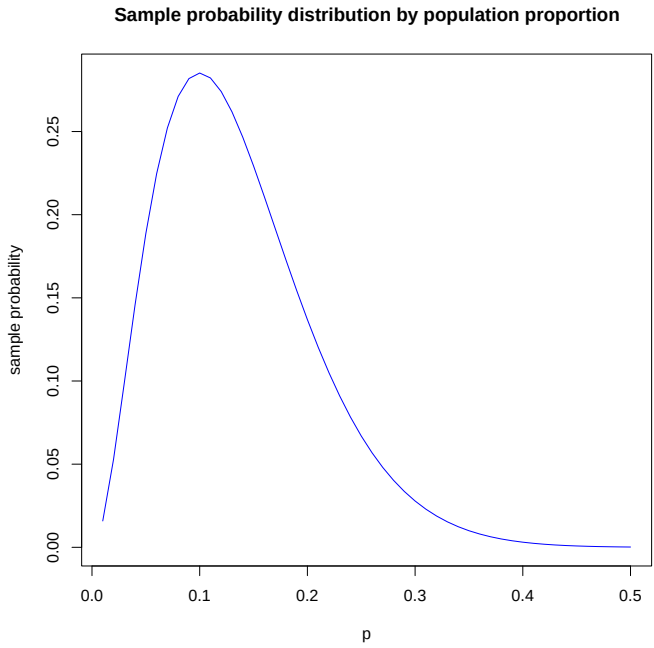


図 5.1: さまざまな p に対して，20 個取り出したときに黒石がちょうど 2 個である確率

5.2　推定値の確からしさ

ここで，このようにして求めた推定値がどれほど確からしいか？　を考えよう。例えば，黒石の割合（母比率）が p である容器から 20 個の石を取り出したときに，黒石がちょうど 2 個である確率を考えると，これは二項分布に従う[*1]。

つまり，確率 p の現象が 20 回中 2 回起こり，残りの 18 回は確率 (1-p) の現象が起こったわけだから，その確率をすべて掛け合わせ，20 回中どの 2 回で起こるのかと

[*1] 厳密に考えると，この確率が二項分布に従うためには復元抽出でなければならないので，1 個取り出しては戻す必要がある。CMR ではそうしないが，それは標本に比べて母集団が十分に大きく，抽出操作が母比率に影響を与えないと仮定できるからである。

いう組み合わせの数だけパタンがありうるので ${}_{20}C_2$ 回だけそれを足し合わせた確率になる[*2]。この値が最大となるのは，図 5.1 のように $p = 0.1$ の時である[*3]。

40 個入れて全体の 0.1 を占めるのだから，40/0.1=400 が全体の数で，400-40=360 が元の白石の数だと推定できる。ただし，図を見ればわかるように，$p = 0.09$ だろうが $p = 0.11$ だろうが，黒石がちょうど 2 個である確率には大した差はない。だから，360 個という点推定値は，404 個とか 324 個に比べて，それほど信頼性は高くない。

5.3 信頼区間

では，ある程度の信頼性が見込める範囲を示すことは可能だろうか？　という考え方で示されるのが信頼区間である。例をあげよう。ビデオリサーチ (http://www.videor.co.jp/index.html) によれば，NHK の朝のテレビ小説「ほんまもん」の 2001 年 10 月 8 日の関東地区の視聴率は 22.9%であった。関東地区の調査対象世帯は 600 だから，137 世帯が見ていたことになる。このとき，関東地区全体の真の視聴率（母比率）は，どのくらいの範囲をとれば，95%の確率でその中に収まるのか？　というのが問題である。

「ほんまもん」を見る／見ないという事象が各世帯独立に起こるとすれば，二項分布で考えることができる。母比率が 137/600 の時にちょうど 137 世帯が見た（裏返して言えば 463 世帯が見なかった）確率は，`choose(600,137)*(137/600)^137*(463/600)^463` で，たかだか 3.9%に過ぎない。

しかし，例えば，母比率が 10%だったのに 137 世帯が見たという確率は，`2.5*10^(-20)` であり，まったくありそうにない。137/600 の前後適当な幅をとれば，かなり高い確率で，ちょうど 137 世帯が見た，という事象が起こることになる。この幅を**「信頼区間」**という。95%の確率でちょうど 137 世帯が見たという事象が起こるための母比率の推定幅を，「95%信頼区間」という。

95%信頼区間を求めるには，下側 2.5%の点と上側 2.5%の点を求めればよいので，

[*2] R では `choose(20,2)*p^2*(1-p)^18` あるいは `dbinom(2,20,p)` で得られる。

[*3] この図を描かせる R のプログラムは次の通り。

```
p <- c(1:50)/100
prob <- dbinom(2,20,p)
plot(p,prob,col="blue",type="l",main="Sample probability distribution by population proportion",
ylab="sample probability")
```

R なら,

```
z<-0; k<-0; while (z<0.025) {k <- k+1; z <- z+zz[k]}; print(k)
```

として下側 2.5%の点を求め,

```
z<-0; k<-600; while (z<0.025) {k <- k-1; z <- z+zz[k]}; print(k)
```

として上側 2.5%の点を求めればよい。

結果として，600 世帯の調査で 22.9%の視聴率だったら，無限母集団の視聴率（真の視聴率）の 95%信頼区間は，19.8%から 26.5%の間と言える。

5.4 正規近似による信頼区間の推定

二項分布は，n が大きいときは正規分布で近似できる。このことを利用すれば，母比率 p，標本数（調査世帯数）n で，その標本の中で注目している属性をもつ標本数（「ほんまもん」を見た世帯数）を X，観測比率を $p' = X/n$ とすれば，X が近似的に正規分布 $N(np, np(1-p))$ に従うことになる。正規分布の 95%のサンプルは，平均 ± 標準偏差× 1.96 に含まれるので,

$$\mathrm{Prob}[-1.96 \leq (X - np)/\sqrt{np(1-p)} \leq 1.96] = 0.95$$

これから式変形すると, $\mathrm{Prob}[p' - 1.96\sqrt{p'(1-p')/n} \leq p \leq p' + 1.96\sqrt{p'(1-p')/n}] = 0.95$ となるので，母比率 p は 95%の確率で $(p' - 1.96\sqrt{p'(1-p')/n}, p' + 1.96\sqrt{p'(1-p')/n})$ の範囲にあるといえる。即ちこれが，母比率 p の 95%信頼区間となる。

練習問題

ある大学の正門の前で，ある朝登校して来る学生の男女比を調べてみたところ，300 人中，女子学生が 75 人であった。この大学の女子学生の割合の点推定値と 95%信頼区間を求めよ。

点推定値は言うまでもなく 75/300=0.25，つまり 25%である。95%信頼区間の下限を求める R の式は，`75/300-2*sqrt(75/300*225/300/300)`，上限は `75/300+2*sqrt(75/300*225/300/300)` であるから，95%信頼区間は [20%, 30%] となる。なお，この推定には，朝登校して来る学生に男女の偏りがないという仮定があるので，実は真の値を過大評価することになっている。どうすれば正しい推定ができるような標本がとれるか，考えてみるのも一興であろう。

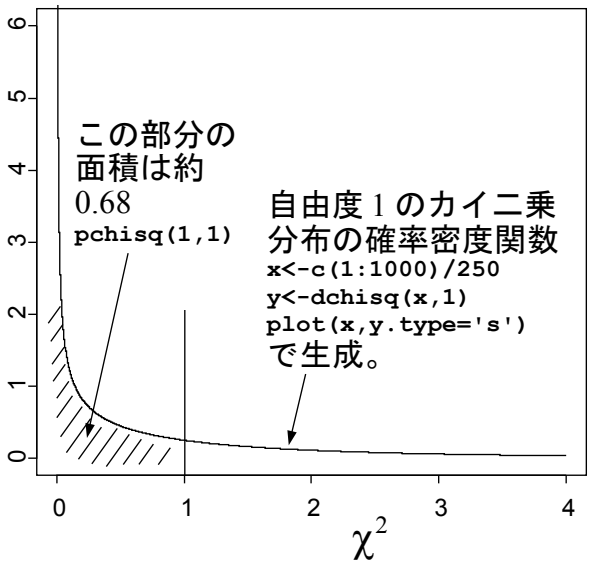


図 5.2: 自由度 1 のカイ二乗分布

5.5 母比率の検定

予め母比率について何らかの期待があるとき（50%であるとか），標本から推定された母比率が，それと違っていないかどうかを調べたい，ということが起こる。こういう場合の基本的な考え方としては，標本データの度数分布が，母集団について期待される分布と一致するという仮説（帰無仮説）が成り立っている確率を調べて，それが普通では考えられないほど小さい場合（通常は 5%未満）に，滅多にないことだから偶然ではない（これを「有意である」という），と考えて帰無仮説を棄却する。

カテゴリ数が全部で n 個あるとき，i 番目のカテゴリの観測度数が O_i，期待度数が

E_i であるとき，$\chi^2 = \sum (O_i - E_i)^2/E_i$ が[*4]，自由度 $n-1$ のカイ二乗分布に従うことを利用して検定する（但し，不明な母数があるときは，その数も自由度から引く。E_i が 1 未満のときはカテゴリ分けをやり直す)。このような χ^2 が大きな値になることは，観測された度数分布が期待される分布と一致している可能性が極めて低いことを意味する。一般に，χ^2 が自由度 $n-1$ のカイ二乗分布の 95%点よりも大きいときは，統計的に有意であるとみなして，帰無仮説を棄却する（適合しないと判断する)。

ちなみに，自由度 1 のカイ二乗分布は，図 5.2 のような形になる。χ^2 値が 1 より大きくなる確率は，約 0.32 ということである。 参考までに，自由度 n のカイ二乗分布の確率密度関数は，$x > 0$ について， $f_n(x) = 1/(2^{(n/2)}\Gamma(n/2))x^{(n/2-1)}\exp(-x/2)$ であり，平均 n，分散 $2n$ である。なお，**自由度 (degree of freedom; d.f.)** とは，標本の数から，前もって推定する母数の数を引いた値である[*5]。この例なら $\sum E_i$ だけを $\sum O_i$ として推定すれば，E_1 から E_{n-1} まで定めて E_n が決まることになるので，自由度は 1 を引く。

分布関数（確率母関数）は，確率密度関数を積分したものであり，図で見れば面積に当たる。その逆関数，つまり面積に対応する値を与える関数を分位点関数という。自由度 1 のカイ二乗分布の場合，R では，カイ二乗値 x について，確率密度関数が `dchisq(x,1)`，分布関数が `pchisq(x,1)` で与えられ，95%点を与える分位点関数が `qchisq(0.95,1)` で与えられる。これは，χ^2 値が `qchisq(0.95,1)` 以下である確率が 95%であることを意味する。逆にいえば，χ^2 値が `qchisq(0.95,1)` より大きくなることは，確率 5%もない，滅多にないことである。言い換えると，「観測された分布が期待される分布と一致している可能性は 5%もない」ということである。このようなとき，「観測された分布が期待される分布と違いがない」という仮説は有意水準 5%で棄却されたといい，観測された分布は期待される分布と一致するとはいえない，と解釈する。

例題 2 .

ある病院で生まれた子ども 900 人中，男児は 480 人であった。このデータから，（1）男女の生まれる比率は半々であるという仮説，（2）男児 1.06 に対して女児 1 という割合で生まれるという仮説，は支持されるか？（出典：豊川・柳井，1982）

[*4] χ は「カイ」と発音する。英語では chi-square と書かれるので，英文を読むときに間違って「チ」と読んでしまうと大変恥ずかしい。

[*5] 合計を母数と考えなければ，標本数から 1 を引いて母数の数を引く，と捉えてよい。前章で不偏分散を求めるときにも標本数から 1 を引いて自由度を求めた。

（1）の場合，χ^2 は，`X <- (480-450)^2/450+(420-450)^2/450` として計算される。この値が自由度 1 のカイ二乗分布に従うので，R で `1-pchisq(X,1)` とすれば，男女の生まれる比率が半々である場合に 900 人中男児 480 人という観察値が得られる確率が計算できる。その確率がきわめて小さければ（通常 5%未満），統計的に意味があるほど有り得なさそうな（「統計的に有意な」という）現象であると考えて，仮説を棄却する。

実は，この場合は母比率が 0.5 であるとして 2 項分布で計算してもよい。480 人以上になる確率と 420 人以下になる確率の合計がきわめて小さければ，「男女の生まれる比率は半々である」という仮説はありそうもないと考えてよいことになる。母比率 0.5 で起こる現象が，900 回中ちょうど 480 回起こる確率は，`choose(900,480)*0.5^480*0.5^420` で与えられるが，R には二項分布についてもカイ二乗分布と同じように確率密度関数を与える関数があり，この確率は `dbinom(480,900,0.5)` で与えられる。

480 人以上になる確率は，`dbinom(480,900,0.5) + dbinom(481,900,0.5) + ... + dbinom(900,900,0.5)` となるが，これは分布関数を使えば，`1-pbinom(480,900,0.5)` で計算できる。420 人以下になる確率は，`dbinom(0,900,0.5) + dbinom(1,900,0.5) + ... + dbinom(420,900,0.5)` であり，分布関数を使って書けば，`pbinom(420,900,0.5)` である。従って，求める確率はこれらの和，即ち，

```
1-pbinom(480,900,0.5)+pbinom(420,900,0.5)
```

である。計算してみると 0.045... となるので，有意水準 5%で仮説は棄却されることがわかる。

（2）の場合，χ^2 は，

```
EM <- 900*1.06/2.06; EF <- 900*1/2.06
X <- (480-EM)^2/EM+(420-EF)^2/EF
1- pchisq(X,1)
```

を計算すると，約 0.26 となるので，仮説の下で偶然，男児が 900 人中 480 人以上になる確率は約 26%あると解釈され，この仮説は棄却されないことがわかる。

応用：

1 日の交通事故件数を 155 日間について調べたところ，0 件の日が 79 日，1 件の日が 61 日，2 件の日が 13 日，3 件の日が 1 日，4 件以上の日が 1 日だったとする。このとき，1 日あたりの交通事故件数はポアソン分布に従うと言えるか？（出典：豊川・柳井，1982）[a]

[a] 一般に，稀な事象についてベルヌーイ試行を行うときの事象生起数がポアソン分布に従うことが知られている。交通事故は稀な事象であり，ある日に交通事故が起こる件数と翌日に交通事故が起こる件数は独立と考えられるので，交通事故件数はポアソン分布に従うための条件を満たしている。

R では，ポアソン分布の確率関数（離散分布の場合は，確率密度関数と言わずに確率関数というのが普通）は，dpois(**件数，期待値**) で与えられる。

ポアソン分布の期待値（これは母数である）がわからないので，データから推定すれば，(0 × 79 + 1 × 61 + 2 × 13 + 3 × 1 + 4 × 1)/155 で得られる。R で書けば，この値を Ehh に保存するとして，

```
cc <- c(0:4); hh <- c(79,61,13,1,1); Ehh <- sum(cc*hh)/sum(hh)
```

となる。

従って，1 日の事故件数が期待値 Ehh のポアソン分布に従うとしたときの，事故件数 0〜4 の期待日数 epp は，epp <- dpois(cc,Ehh)*sum(hh) で得られる。

こうなれば，X <- sum((hh-epp)^2/epp) としてカイ二乗値を求め，これが自由度 3（件数の種類が 5 種類あって，ポアソン分布の期待値が母数として推定されたので，$5-1-1=3$ となる）のカイ二乗分布に従うとして 1-pchisq(X,3) が 0.05 より小さいかどうかで判定すれば良さそうなものだが，そうはいかない。

epp の値を見ればわかるが，epp[cc==4] が 1 より小さいのである。期待度数が 1 より小さいときはカテゴリを併合しなくてはならないので，epp[cc==4] を epp[cc==3] と併合する。

即ち，

```
ep <- epp[cc<3]
ep[cc==3] <- epp[cc==3]+epp[cc==4]
ep <- ep[!is.na(ep)]
```

として期待度数の分布 ep を得，

```
h <- hh[cc<3]
```

```
h[cc==3] <- hh[cc==3]+hh[cc==4]
h<-h[!is.na(h)]
```

として観測度数の分布 h を得る。

後は，XX<-sum((h-ep)^2/ep) としてカイ二乗値を求め，1-pchisq(XX,2) を計算すると（カテゴリが 1 つ減ったので自由度も 1 減って 2 となる），約 0.187 となることがわかる。即ち，1 日の交通事故件数がポアソン分布にしたがっていると仮定したとき所与のデータよりも偏ったデータが得られる確率は約 19%あり，珍しいこととはいえない。

第 6 章

カテゴリ変数２つの分析（１）

6.1　２つのカテゴリ変数を分析する２つのアプローチ

前章では，１つのカテゴリ変数のもつ情報から母比率を推定したり，期待される母比率と一致するかどうかを検定する方法を示した。本章では，２つのカテゴリ変数を分析する方法を示す。

２つのカテゴリ変数を分析するには，２つのアプローチがある。１つは，２つの変数についての母比率に差があるかどうかを調べるアプローチであり，もう１つは，２つの変数の関係を調べるアプローチである。後者を調べる際には，クロス集計表を作るのが普通である。その上で，２つの変数の独立性を検定したり，関連の程度を調べたりする。[*1]

6.2　２つのカテゴリ変数の母比率の差の検定と信頼区間

前章で説明したように，個々のカテゴリ変数がもつ情報はデータ数（標本数）と，各カテゴリの割合である。そこから，各カテゴリの母集団における割合（母比率）を推定することができる。

２つのカテゴリ変数の母比率 p_1, p_2 が，各々の標本比率 $\hat{p_1} = r_1/n_1, \hat{p_2} = r_2/n_2$ として推定されるとき，それらの差を考える。差 $(\hat{p_1} - \hat{p_2})$ の平均値と分散は，$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2, V(\hat{p_1} - \hat{p_2}) = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$ となる。２つの母比率に差が無いならば，$p_1 = p_2 = p$ とおけるはずなので，$V(\hat{p_1} - \hat{p_2}) =$

[*1] ただし母比率の差の検定は，後で述べるように，２×２のクロス集計表とみなして独立性の検定をすることと数学的に等価である。

$p(1-p)(1/n_1+1/n_2)$ となる。この p の推定値として，$\hat{p}=(r_1+r_2)/(n_1+n_2)$ を使い，$\hat{q}=1-\hat{p}$ とおけば，n_1p_1 と n_2p_2 がともに 5 より大きければ，標準化して正規近似を使い，

$$Z=\frac{\hat{p_1}-\hat{p_2}-E(\hat{p_1}-\hat{p_2})}{\sqrt{V(\hat{p_1}-\hat{p_2})}}=\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1+1/n_2)}}\sim N(0,1)$$

によって[*2]検定できる。

例をあげよう。2002 年 6 月 1 日に山口県立大学の 2 つのキャンパスを隔てるバイパスで，交通の様子を観察したデータを考える[*3]。1 人で観察する場合，観察対象は車，歩行者などが考えられるが，ここでは車とする。車 1 台から得られるいくつかの特性を 1 組のデータとして扱う（こういう 1 組を 1 つの「オブザーベーション (observation)」と呼ぶ)。簡単に捉えられる特性としては，進行方向，車の種類（普通乗用車かそれ以外か），車の色，といったものが考えられる。データ解析を考える上では，これらの特性が「変数」となる。つまり，生データを表の形にまとめると，

	変数		
オブザーベーション番号	進行方向	車の種類	車の色
1	津和野方面	乗用車	白
2	山口市街地	乗用車	白
3	山口市街地	乗用車	銀
⋮			

これを数値としてコーディングするときは，典型的なカテゴリを 1 にするとわかりやすい。進行方向という変数の変数名を Dest（値は，津和野方面を 1，山口市街地方面を 2）とし，車の種類の変数名を Type（乗用車が 1，トラックなどそれ以外のものを 2)，車の色の変数名を Color（1 が白，2 が黒，3 はそれ以外）とすれば，次の表のようにコード化される。

[*2] この Z は離散値しかとれないため，連続分布である正規分布による近似の精度を上げるために，連続性の補正と呼ばれる操作を加え，かつ $p_1>p_2$ の場合（つまり $Z>0$ の場合）と $p_1<p_2$ の場合 (つまり $Z<0$ の場合）と両方考える（両側検定という）のだが，正規分布は原点について対称なので，絶対値をとって $Z>0$ の場合だけ考え，有意確率を 2 倍すればよい（逆に 5%水準で検定したいなら，97.5%点より Z が大きいかどうかを見ればよい)。即ち，

$$Z=\frac{|\hat{p_1}-\hat{p_2}|-(1/n_1+1/n_2)/2}{\sqrt{\hat{p}\hat{q}(1/n_1+1/n_2)}}$$

として，この Z の値が標準正規分布の 97.5%点（R ならば `qnorm(0.975,0,1)`）より大きければ帰無仮説を棄却するのが普通である。

[*3] http://phi.ypu.jp/statlib/tf.mpg として MPEG1 形式のムービーファイルを公開している。

	変数		
Obs	Dest	Type	Color
1	1	1	1
2	2	1	1
3	2	1	3
⋮			

これを表計算ソフト（Excel や StarSuite/OpenOffice.org の scalc など）やテキストエディタで入力し，CSV（コンマ区切り値）形式のファイル `L6-1.csv` として R の作業ディレクトリに保存すれば，R のコンソールで `x <- read.csv("L6-1.csv")` として，`x` というデータフレームに読み込めるし，タブ区切りテキスト形式のファイル `L6-1.dat` として保存すれば，`x <- read.delim("L6-1.dat")` として読み込める[*4]。R では，各変数はデータフレーム名$変数名として参照できるので，例えば進行方向別の頻度を出したいときは，`table(x$Dest)` とすれば良い。総観察数 89 台のうち，津和野方面が 60 台，山口市街地方面が 29 台であったことがわかる。

ここで，進行方向によって乗用車割合が異なるかという仮説を考えてみる。帰無仮説は，「進行方向が反対でも乗用車割合には差が無い」ということになる。

`table(x$Type[x$Dest==1])` とすれば，津和野方面の乗用車が 60 台中 57 台であることがわかり，`table(x$Type[x$Dest==2])` とすれば，山口市街地方面の乗用車が 29 台中 25 台であったことがわかる。

上で説明した式にあてはめて計算すると，

$$\hat{p} = (57 + 25)/(60 + 29) = 0.92...$$

$$\hat{q} = 1 - \hat{p} = 0.079...$$

$$Z = (|0.95 - 0.86| - (1/60 + 1/29)/2)/\sqrt{0.92 \cdot 0.079 \cdot (1/60 + 1/29)} = 1.024$$

となるので，標準正規分布の 97.5%点である 1.96 よりずっと小さく，5%水準で有意ではない。つまり帰無仮説は棄却されず，差はないと考えてよい[*5]。

[*4] これらのファイルは `http://phi.ypu.jp/statlib/L6-1.csv` などとしてダウンロードできる。

[*5] 厳密に言えば，差がないとしたときに偶然この値以上の値が得られる確率が 5%よりずっと多い，ということである。この確率がいくらかといえば，R で，`2*(1-pnorm(1.024,0,1))` とすれば，0.305... という値が得られるので，約 31%である。ついでに書いておくと，有意確率は，それに従って帰無仮説を棄却した場合にその判断が誤りであった（＝実は差がなかった）確率なので，第一種の過誤（α-Error）とも呼ばれる。反対に，検定の検出力が足りなくて本当は差があるのに差がないと判断してしまう確率を第二種の過誤（β-Error）と呼ぶ。第二種の過誤は標本数に依存する。

差の 95%信頼区間を出すことも簡単である。信頼区間を出すには，サンプルサイズが大きければ正規分布を仮定できるので，原則どおりに差から分散の平方根の 1.96 倍を引いた値を下限，足した値を上限とすればよい。上の例では，$\hat{p_1} - \hat{p_2} = 0.0879...$，$V(\hat{p_1} - \hat{p_2}) = \hat{p_1}(1 - \hat{p_1})/n_1 + \hat{p_2}(1 - \hat{p_2})/n_2 = 57/60(1 - 57/60)/60 + (25/29(1 - 25/29)/29 = 0.00489...$ となるので，信頼区間の下限は $0.0879 - 1.96 * \sqrt{0.00489} = -0.049$，上限は $0.0879 + 1.96 * \sqrt{0.00489} = 0.225$ となる。しかし，通常は連続性の補正を行うので，下限からはさらに $(1/n_1 + 1/n_2)/2 = (1/60 + 1/29)/2 = 0.0255...$ を引き，上限には同じ値を加えて，95%信頼区間は $(-0.0747, 0.251)$ となる。

実は R では，

```
type1 <- c(57,25)
total <- c(60,29)
prop.test(type1,total)
```

とすれば各々の母比率の推定と，その差があるかどうかの検定（連続性の補正済み，ただし正規近似そのままではなく，カイ二乗分布で検定したものだが，数学的にはまったく同値である），差の 95%信頼区間を一気に出してくれる。$p = 0.3057$ より有意な差は無く，95%信頼区間は $(-0.0747, 0.251)$ であることがわかる。

6.3　2 つのカテゴリ変数の関係を調べることと研究のデザイン

こんどは，2 つの変数の関係を調べるアプローチについて説明する。関係を調べるといっても，研究デザインによって，検討すべき関係の種類はさまざまである。例えば，肺がんと判明した男性患者 100 人と，年齢が同じくらいの健康な男性 100 人を標本としてもってきて，それまで 10 年間にどれくらい喫煙をしたかという聞き取りを行うという「患者対照研究＝ケースコントロール研究」[*6]を実施した場合に，喫煙の程度を「一度も吸ったことがない」から「ずっとヘビースモーカーだった」まで何段階かのスコアを振れば，喫煙状況という変数と肺がんの有無という変数の組み合わせが得られる。もちろん，それらが独立であるかどうか（関連がないかどうか）を検討することもできる。

しかし，むしろこのデザインは，肺がん患者は健康な人に比べて，どれくらい喫煙していた割合が高いか，を評価するためのデザインである（既に亡くなっている人が

[*6] 詳しくは疫学の教科書を参照されたい。

除かれてしまっているので，発生リスクは過小評価されるかもしれない)。逆に，喫煙者と非喫煙者を 100 人ずつ集めて，その後の肺がん発生率を追跡調査する前向き研究（フォローアップ研究）では，非喫煙群に比べて，喫煙者ではどれくらい肺がんの発生率が高いかを評価でき[*7]，断面研究で得られた２つの変数には時間的な前後関係がないので，独立性の検定を行ったり，リスク比やオッズ比以外の関連性の指標を計算することが多い[*8]。関連性の指標については次章で詳しく説明することにして，本章の後半では，独立性の検定について説明する。

6.4　クロス集計とは？

２つのカテゴリカル変数の間に関係があるかどうかを検討したいとき，それらの組み合わせの度数を調べた表を作成する。これをクロス集計表と呼ぶ。

とくに，２つのカテゴリカル変数が，ともに２値変数のとき，そのクロス集計は２×２クロス集計表（２×２分割表）と呼ばれ，その統計的性質が良く調べられている。

6.5　独立性の検定の原理

独立性の検定は，２つのカテゴリカル変数の間に関連がないと仮定した場合に推定される期待度数を求めて，それに観測度数が適合するかを検定するカイ二乗検定である。もちろん，ある種の関連が仮定できれば，その仮定の元に推定される期待度数と観測度数との適合を調べてもいいが，一般に，２つのカテゴリカル変数の間にどれくらいの関連がありそうかという仮定はできないことが多い。そこで，関連がない場合の期待度数を推定し，それが観測値に適合しなければ関連がないとはいえない，と推論するのである。

	特性 A あり	特性 A なし
特性 B あり	a 人	b 人
特性 B なし	c 人	d 人

標本が，上記の表のような度数をもっているとき，母集団の確率構造が，

[*7] それらの値は次章で説明するリスク比やオッズ比という指標で表され，疫学研究上非常に重要である。

[*8] ただし，オッズ比は断面研究でも計算できる。

	特性 A あり	特性 A なし
特性 B あり	π_{11}	π_{12}
特性 B なし	π_{21}	π_{22}

であるとわかっていれば，$N = a + b + c + d$ として，期待される度数は，

	特性 A あり	特性 A なし
特性 B あり	$N\pi_{11}$	$N\pi_{12}$
特性 B なし	$N\pi_{21}$	$N\pi_{22}$

であるから，

$$\chi^2 = \frac{(a - N\pi_{11})^2}{N\pi_{11}} + \frac{(b - N\pi_{12})^2}{N\pi_{12}} + \frac{(c - N\pi_{21})^2}{N\pi_{21}} + \frac{(d - N\pi_{22})^2}{N\pi_{22}}$$

として，自由度 3 のカイ二乗検定をすればよいが，普通は π が未知なので，$p(A \cap B) = p(A)p(B)$ と考えて，各々の変数については特性のある人とない人の人数が決まっている（周辺度数が固定している）と考え，$p(A)$ の推定値 $(a+c)/N$ と $p(B)$ の推定値 $(a+b)/N$ の積として π_{11} を，$p(\bar{A})$ の推定値 $(b+d)/N$ と $p(B)$ の推定値 $(a+b)/N$ の積として π_{12} を，$p(A)$ の推定値 $(a+c)/N$ と $p(\bar{B})$ の推定値 $(c+d)/N$ の積として π_{21} を，$p(\bar{A})$ の推定値 $(b+d)/N$ と $p(\bar{B})$ の推定値 $(c+d)/N$ の積として π_{22} を推定すれば，

$$\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

となる。この場合は，母数を 2 つ推定したので，自由度 1 のカイ二乗分布に従うと考えて検定できる。

ただし通常は，イェーツの連続性の補正を行う。カイ二乗分布は連続分布なので，各度数に 0.5 を足したり引いたりしてやると，より近似が良くなるという発想である。この場合，

$${\chi_c}^2 = \frac{N(|ad - bc| - N/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

が自由度 1 のカイ二乗分布に従うと考えて検定する。ただし，$|ad - bc|$ が $N/2$ より小さいときは補正の意味がないので，$\chi^2 = 0$ とする。

実際の検定は R を使えば，a=12, b=8, c=9, d=10 などとわかっているときは，`x <- matrix(c(12,8,9,10),nc=2)` として表を与え，`chisq.test(x)` とするだけでもできる（連続性の補正を行わないときは `chisq.test(x,correct=F)` とするが，通常その必要はない）。各度数が未知で，各個人についてのカテゴリカル変数 A と B の生の値が与えられているときも，R を使うと，`chisq.test(A,B)` で計算できる。ク

ロス集計表を作るには，`table(A,B)` とする。もちろん，`chisq.test(table(A,B))` としてもよい。

R では，`chisq.test()` 関数の中で，`simulate.p.value=TRUE` というオプションを使えば，シミュレーションによってそのカイ二乗値より大きなカイ二乗値が得られる確率を計算させることもできる。この方がたんなるカイ二乗検定よりも正確な p 値が得られるが，遅いコンピュータだと計算時間がかかる欠点がある。

例題 1

上の交通量調査データで独立性のカイ二乗検定をせよ。

帰無仮説は，進行方向と車の種類が独立（無関係）ということである。クロス集計表を作ってみると，

	乗用車	それ以外	合計
津和野方面	57	3	60
山口市街地方面	25	4	29
合計	82	7	89

となる。進行方向と車の種類が無関係であった場合に期待される度数は，

	乗用車	それ以外	合計
津和野方面	60*82/89	60*7/89	60
山口市街地方面	29*82/89	29*7/89	29
合計	82	7	89

となる。これから定義の通りに計算してもいいが，連続性の修正も考えると公式に代入するのが現実的である[*9]。実際に代入してみると，連続修正済みのカイ二乗統計量は $\chi_c^2 = 89 * (|57 * 4 - 3 * 25| - 89/2)^2/(60 * 29 * 82 * 7) = 1.049$ となる。自由度 1 のカイ二乗分布で分布関数の値を 1 から引くと，p=0.3057... となり，有意確率が約 31%である（つまり帰無仮説は棄却されず，独立である可能性が十分にある）ことがわかる[*10]。ただし，R で実行した結果に警告メッセージが出ていることからもわかるが，この例ではサイズの小さなセルがあるので，カイ二乗検定における正規近似は適当でない可能性があり（一般に期待度数が 5 以下のセルが全体の 20%以上あるときはカイ二乗検定は適当でないとされる），次に説明するフィッシャーの直接確率を使った方がよい。

[*9] もちろん，R で `chisq.test(matrix(c(57,25,3,4),nc=2))` とするのが手軽である。

[*10] この値が**母比率の差の検定の有意確率と一致していること**に注意されたい。

6.6 フィッシャーの直接確率（正確な確率）

周辺度数を固定してすべての組み合わせを考え，それらが起こる確率（超幾何分布に従う）を直接計算し，与えられた表が得られる確率よりも低い確率になる場合をすべて足し合わせたものをフィッシャーの直接確率，あるいは，フィッシャーの正確な確率（検定）という。

もう少し丁寧に言うと，サイズ N の有限母集団があって，そのうち変数 A の値が 1 である個体数が m_1，1 でない個体数が m_2 あるときに，変数 B の値が 1 である個体数が n_1 個（1 でない個体数が $n_2 = N - n_1$ 個）あるという状況を考え，そのうち変数 A の値が 1 である個体数がちょうど a である確率を求めることになる。これは，m_1 個から a 個を取り出す組み合わせの数と m_2 個から $n_1 - a$ 個を取り出す組み合わせの数を掛けて，N 個から n_1 個を取り出す組み合わせの数で割った値になる。これと同じ周辺度数をもつ 2 × 2 分割表のうち，確率がこれと同じかこれよりも小さい表の確率をすべて足し合わせたものが，「変数 A と変数 B が独立」という帰無仮説が成り立つ確率になる[*11]。

フィッシャーの正確な確率は，R では，`fisher.test(table(A,B))` で実行できる。この方がカイ二乗検定よりも正確である。独立性の検定をするときは，コンピュータが使えるならば，サンプルサイズがよほど大きくない限り，常に Fisher の正確な確率を求めるべきである。

例題 2

上の交通量調査データでフィッシャーの直接確率を計算せよ。

`fisher.test(matrix(c(57,25,3,4),nc=2))` を R で計算させると，0.2089 となる。カイ二乗検定の場合よりも小さな有意確率が得られたことに注意されたい（一般に第一種の過誤をしにくい）。

わかりにくいと思うので，サンプル数が少ない場合について，実際に数値を使って説明しておく。Fisher の正確な確率は仮定が少ない分析法で，とくにデータ数が少な

[*11] 有限母集団からの非復元抽出になるので，平均 $E(a)$ と分散 $V(a)$ は，$E(a) = n_1 m_1/N$，$V(a) = \{(N-n_1)/(N-1)\} n_1 (m_1/N)(m_2/N) = (m_1 m_2 n_1 n_2)/\{N^2(N-1)\}$ となる。実際には組み合わせ計算が多いので，手計算で実行することはまずありえず，統計ソフトにやらせることになる。また，個々の 2 × 2 分割表の確率は離散値をとるので，同じ確率の表がありうる場合に，それを足し算に含めるのかどうかは難しい点である。これを乱数によって決める「ランダム検定」という手法もあるが，あまり一般的ではない。

くてカイ二乗検定が使えない場合にも使えるので，動物実験などでは重宝する。いま仮に，下のようなクロス集計表が得られたとする。

	A あり	A なし	合計
B あり	4	3	7
B なし	1	7	8
合計	5	10	15

15 人のうち，5 人が要因 A をもっていて，7 人が要因 B をもっているときに[*12]，この表が得られる確率は，15 人のうち要因 A をもっている 5 人の内訳が，要因 B をもっている 7 人から 4 人と，要因 B をもっていない 8 人から 1 人になる確率となる。つまり，15 から 5 を取り出す組み合わせのうち，7 から 4 を取り出し，かつ残りの 8 から 1 を取り出す組み合わせをすべて合わせたものが占める割合になるので，${}_7C_4 \cdot {}_8C_1 / {}_{15}C_5 \simeq 0.0932$ である。

つまり，上のクロス集計表が，偶然（2つの変数に何も関係がないとき）得られる確率は 0.0932 ということである。これだけでも既に 5%より大きいので，「2つの変数が独立」という帰無仮説は棄却されず，A の有無と B の有無は関係がないと判断していいことになる。

しかし，有意確率，つまり第一種の過誤を起こす確率は，A の有無と B の有無には関係がないと判断した場合にそれが間違っている確率なので，この表だけではなく，この表よりも偶然得られる確率が低い表が得られる確率をすべて足さねばならない。周辺度数が上の表と同じ表は，

(1)	A あり	A なし	(2)	A あり	A なし	(3)	A あり	A なし	合計
B あり	0	7		1	6		2	5	7
B なし	5	3		4	4		3	5	8
合計	5	10		5	10		5	10	15

(4)	A あり	A なし	(5)	A あり	A なし	(6)	A あり	A なし	合計
B あり	3	4		4	3		5	2	7
B なし	2	6		1	7		0	8	8
合計	5	10		5	10		5	10	15

の計 6 種類しかない。(1) や (6) の表よりもさらに稀な場合を考えると，(1) の先は

[*12] 「各変数については母比率が決まっているとき」ということで，このことを「全ての周辺度数が固定されているとき」ともいうのである。

A も B もある人の数がマイナスになってしまうし，(6) の先は A があって B がない人の数がマイナスになってしまう。

そこで，すべての表について，それが偶然得られる確率を計算すると，[*13](1) は ${}_7\mathrm{C}_0 \cdot {}_8\mathrm{C}_5 / {}_{15}\mathrm{C}_5 \simeq 0.0186$，(2) は ${}_7\mathrm{C}_1 \cdot {}_8\mathrm{C}_4 / {}_{15}\mathrm{C}_5 \simeq 0.1632$，(3) は ${}_7\mathrm{C}_2 \cdot {}_8\mathrm{C}_3 / {}_{15}\mathrm{C}_5 \simeq 0.3916$，(4) は ${}_7\mathrm{C}_3 \cdot {}_8\mathrm{C}_2 / {}_{15}\mathrm{C}_5 \simeq 0.3263$，(5) は上で計算した通り ${}_7\mathrm{C}_4 \cdot {}_8\mathrm{C}_1 / {}_{15}\mathrm{C}_5 \simeq 0.0932$，(6) は ${}_7\mathrm{C}_5 \cdot {}_8\mathrm{C}_0 / {}_{15}\mathrm{C}_5 \simeq 0.0070$ となる[*14]。

以上の計算より，元の表（= (5)）より得られる確率が低い（つまりより偶然では得られにくい）表は (1) と (6) なので，それらを足して，元の表の両側検定（どちらに歪んでいるかわからない場合）での有意確率は，$0.0932 + 0.0186 + 0.0070 = 0.1188$ となる。

[*13] R で組み合わせ計算を行う関数は `choose()` である。例えば ${}_7\mathrm{C}_3$ は，`choose(7,3)` で計算できる。

[*14] これらの確率をすべて足すと 1 になる。上の計算値として書いた値を使うと 0.9999 となるが，これは丸め誤差のせいであり，厳密に計算すれば 1 になる。

第 7 章

カテゴリ変数２つの分析（２）

7.1 研究デザインとリスク，オッズ

前章でも触れたように，研究デザインによって得られる関連性の指標は異なる。まず，関連性の指標の中でもやや毛色が異なる（統計学よりも疫学でよく使われる）リスクとオッズという考え方を説明する。疫学分野で主に発達した理論なので，病気を例にとって説明するが，因果関係が想定できる変数間であれば，別に病気の話に限らず成立する考え方である。

病気のリスクといえば，全体のうちでその病気を発症する人の割合である。これとは別に，オッズという考え方もある。病気のオッズといえば，その病気を発症した人数の，発症しなかった人数に対する比である。

さてしかし，リスクとかオッズそのものでは，病気の発症と要因の有無の関係はわからない。要因があった場合のリスクやオッズを，要因がなかった場合のリスクやオッズと比べることによって，初めて要因の有無と病気の発症がどれくらい関係していたかがわかる。すなわち，ある要因をもつ人たち（曝露群）の病気のオッズが，その要因がない人たち（対照群とかコントロール群という[*1]）の病気のオッズに対して何倍になっているか，というのがオッズ比（英語では Odds Ratio）である。同じように，曝露群のリスクの，対照群のリスクに対する比がリスク比[*2]である。

要因の有無と病気の有無がまったく関係がなければ，リスク比もオッズ比も 1 になることが期待される。それぞれ信頼区間を計算して（計算方法は難しいので後述），

[*1] 理想的な対照群は，その要因がない点だけが曝露群と違っていて，それ以外の条件はすべて同じであることが望ましい。

[*2] 英語では Risk Ratio だが，Rate Ratio とか Relative Risk という言い方もある。Relative Risk の訳から相対危険ということもあるが，同じ意味。

例えば 95%信頼区間が 1 を含まなければ，5%水準で有意な関連が見られるといえる。

ところで，病気のリスクは，全体のうちで病気を発症する人の割合であったから，まず全体を把握していないと定義できない。つまり，まず観察対象全体で曝露群と対照群を把握しておいて，経時的に追跡調査して，それぞれの群で何人ずつ発症するかを調べるという，「前向き研究」（コホート研究とかフォローアップ研究ということもある）でないと，リスク比は計算できないことになる。

これに対して，患者対照研究（Case Control Study）[*3]とか断面研究（Cross Sectional Study）[*4]では，曝露時点での全体が未知なので，原理的にリスクを計算できないことになる。激しい曝露を受けた人は調査時点よりずっと前に病気を発症して死んでしまった可能性があるので，患者対照研究や断面研究から無理にリスクを見積もろうとするとリスクを過小評価してしまうことになるからである。

一方，オッズ比はどんなデザインの研究でも計算できる。たんに，曝露群の病気の人数の病気でない人数に対する比が，対照群のそれに比べてどれくらい大きいかを示す値だからである[*5]。

ここで，クロス集計表ではどう計算するのかということを示す。以下の表を考えてみる。

	疾病あり	疾病なし	合計
曝露あり	a	b	m_1
曝露なし	c	d	m_2
合計	n_1	n_2	N

この表でいえば，リスク比は $(a/m_1)/(c/m_2)$ となり，疾病オッズ比は $(a/b)/(c/d) = ad/bc$ である。曝露オッズ比は $(a/c)/(b/d) = ad/bc$ となるので，疾病オッズ比と一致することがわかる[*6]。

[*3] 調査時点で，患者を何人サンプリングすると決め，それと同じ人数の対照（その病気でないことだけが患者と違って，それ以外の条件はすべて患者と同じことが望ましい）を選んで，それぞれが過去に受けた曝露要因や，現在の生活習慣，態度などを調べることによって，その病気の原因を探る方法論。

[*4] 調べてみないと患者かどうかさえわからないような場合や，因果の向きがはっきりしない変数間の関係を見たいときは，全体で何人サンプリングすると決めて一時点で調査する。こういう方法論を断面研究という。

[*5] この場合のオッズ比は，「曝露なし群での疾病ありのオッズ」に対する「曝露あり群での疾病ありのオッズ」の比なので，**疾病オッズ比**という。逆に，疾病あり群で曝露した人数の曝露していない人数に対する比が，疾病なし群のそれに比べてどれくらい大きいかを示す値として**曝露オッズ比**というものも考えられるが，数学的には同じ値になる。

[*6] ただし，統計パッケージでは，単純なこの値でなく，最尤推定をして得られる条件付きオッズ比が表示されることが多い。

オッズ比が重要なのは，稀な現象をみるときには，リスク比のよい近似になるからであると言われている。例えば，送電線からの高周波が白血病の原因になるという仮説を検証するために，送電線からの距離が近い場所に住んでいる人（曝露群）と，遠いところに住んでいる人（対照群）をサンプリングして，５年間の追跡調査をして，５年間の白血病の罹患率を調査することを考えよう。白血病は稀な疾患だし，高周波に曝露しなくても発症することはあるので，このデザインでリスク比を計算するためには，莫大な数のサンプルをフォローアップする必要があり，大規模な予算とマンパワーが投入される必要があるだろう。

仮に[*7]調査結果が，下表のようであったとすると，

	白血病発症	発症せず	合計
送電線近くに居住	4	9996	10000
送電線から離れて居住	2	9998	10000
合計	6	19994	20000

送電線の近くに住むことで白血病を発症するリスクは，送電線から離れて住む場合の２倍になった（$(4/10000)/(2/10000) = 2$，つまりリスク比が２なので）ということができる。ここでオッズ比をみると，$(4 * 9998)/(2 * 9996) \simeq 2.0004$ と，ほぼリスク比と一致していることがわかる。[*8]

こうして得られるリスク比は，確かに原理的に正しくリスクを評価するのだが，稀なリスクの評価のためには大規模な調査が必要になるので，効率が良いとはいえない。そこで，通常は，前向き研究ではなく，患者対照研究を行って，過去の曝露との関係をみることが行われる。この場合だったら，白血病患者 100 人と対照 100 人に対して，過去に送電線の近くに居住していたかどうかを聞くわけである。それで得られた結果が，仮に下表のようになったとしよう[*9]。

	白血病	白血病でない	合計
送電線近くに居住した経験あり	20	10	30
送電線から離れて居住	80	90	170
合計	100	100	200

この場合，リスク比は計算しても意味がない（白血病かつ送電線の近くに居住した

[*7] これはあくまで架空のデータである。本当の送電線と白血病の関係は，数年前から，WHO のプロジェクトの一環として，国立環境研究所と国立癌センターの研究チームが調べたらしいが，その結果がどうなったのかは知らない。

[*8] 上述のように最尤推定された条件付きオッズ比は，R のプログラムを使って `fisher.test(matrix(c(4,2,9996,9998), nc=2))` として計算すると，2.000322 である。

[*9] くどいようだが，あくまで架空のデータである。

経験がある 20 人は，送電線の近くに住んだ経験がある人からのサンプルではなく，白血病患者からのサンプルだから）が，送電線の近くに居住した経験がある人のうち，白血病の人の，白血病でない人に対するオッズは 2 となり，送電線から離れて居住した人ではそのオッズが 0.888... となるので，これらのオッズの比は 2.25 となる。この値は母集団におけるリスク比のよい近似になることが知られている。このように稀な疾患の場合は，患者対照研究でオッズ比を求める方が効率が良い。

原理的に前向き調査ができない場合もある。とくに，薬害と呼ばれる現象は，妙な病気が見つかったときに，後付けで原因を探ることになるので，患者対照研究にならざるを得ない。例えば，スモンとかサリドマイドは，そうやって原因がわかった問題である。腕が短く生まれた子どもの母親と，そうでない子どもの母親に，妊娠中に飲んだ薬の有無を尋ねて，特定の時期にサリドマイドを飲んだという曝露による疾病オッズ比が有意に大きい結果が得られたのだ[*10]。

また，問題があるかどうかが事前に明らかでない場合は，断面研究をせざるを得ない。聞き取りや質問紙などで調べる，心理学的，あるいは社会学的な調査項目間の関係を見る場合は，断面研究をする場合が多い。なお，断面研究の場合は，リスク比やオッズ比の他に，リスク差，相対差，曝露寄与率，母集団寄与率，Yule の Q，ピアソンの相関係数，ファイ係数といったものがある（後述）[*11]。

なお，同じ質問を 2 回した場合に同じ変数がどれくらい一致するかについては，普通にクロス集計表を作って独立性の検定ができそうな気がするかもしれないが，してはいけない。この場合は test-retest-reliability を測ることになるので，クロンバックの α 係数や κ 係数などの一致度の指標を計算するべきである（後述）。

では，リスク比とオッズ比の 95%信頼区間を考えよう。まずリスク比の場合から考えると，前向き研究でないとリスク比は計算できないので，曝露あり群となし群をそれぞれ m_1 人，m_2 人フォローアップして，曝露あり群で X 人，なし群で Y 人が病気を発症したとしよう。得られる表は，

	発症	発症なし	合計
曝露あり	X	$m_1 - X$	m_1
曝露なし	Y	$m_2 - Y$	m_2
合計	$X + Y$	$N - X - Y$	N

となる。このとき，母集団でのリスクの推定値は，曝露があったとき $\pi_1 = X/m_1$，

[*10] ここで有意と書いたが，統計的に有意かどうかをいうためには検定するか，95%信頼区間を出さねばならない。その方法は後述する。

[*11] 2 × 2でないクロス集計表で，たとえば 5 × 5 以上ならば，順位相関係数を使うことも可能。

曝露がなかったとき $\pi_2 = Y/m_2$ である。リスク比は，$RR = \pi_1/\pi_2$ なので，その推定量は，$(Xm_2)/(Ym_1)$ となる。

リスク比の分布は N が大きくなれば正規分布に近づくので，正規分布を当てはめて信頼区間を求めることができるが，普通は右裾を引いているので対数変換か立方根変換（Bailey の方法）をしなくてはならない。対数変換の場合，95%信頼区間の下限と上限はそれぞれ，

$$RR \cdot \exp(-\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad \text{（下限）} \tag{7.1}$$

$$RR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/X - 1/m_1 + 1/Y - 1/m_2}) \quad \text{（上限）} \tag{7.2}$$

となる。RR が大きい場合は立方根変換しなくてはいけないが，煩雑なので省略する。前述の白血病の例で計算してみると，95%信頼区間は，$(0.37, 10.9)$ となる。

次にオッズ比の信頼区間を考える。前述の表の a, b, c, d という記号を使うと，オッズ比の点推定値 OR は，$OR = (ad)/(bc)$ である。オッズ比の分布も右裾を引いているので，対数変換または Cornfield (1956) の方法によって正規分布に近づけ，正規近似を使って 95%信頼区間を求めることになる。対数変換の場合，95%信頼区間の下限は $OR \cdot \exp(-\mathrm{qnorm}(0.975)\sqrt{1/a + 1/b + 1/c + 1/d})$，上限は $OR \cdot \exp(\mathrm{qnorm}(0.975)\sqrt{1/a + 1/b + 1/c + 1/d})$ となる。前述の白血病の例で計算してみると，オッズ比の 95%信頼区間も $(0.37, 10.9)$ となる[*12]。Cornfield の方法はやや複雑であり，高次方程式の解を Newton 法などで数値的に求める必要があるので，本書では扱わない。

7.2 その他の関連性の指標

7.2.1 リスク差

曝露によるリスクの増減を絶対的な変化の大きさで表した値。過剰危険 (Excess Risk) ともいう。

$$RD = \pi_1 - \pi_2$$

[*12] R の `fisher.test()` 関数で計算した結果では，オッズ比の 95%信頼区間は $(0.29, 22.1)$ となり，対数変換を使った単純な計算よりも幅が広くなる。

7.2.2 相対差

要因ももたず発症もしていない者のうち，要因をもった場合にのみ発症する割合。

$$RelD = (\pi_1 - \pi_2)/(1 - \pi_2)$$

7.2.3 曝露寄与率

真に要因の影響によって発症した者の割合。

$$AFe = (\pi_1 - \pi_2)/\pi_1$$

7.2.4 母集団寄与率

母集団において真に要因の影響によって発症した者の割合。$\pi = (X + Y)/(m_1 + m_2)$ として，

$$AFp = (\pi - \pi_2)/\pi$$

7.2.5 Yule の Q

オッズ比を -1 から 1 の値を取るようにスケーリングしたもの。

$$Q = (OR - 1)/(OR + 1)$$

7.2.6 ファイ係数 (ρ)

要因の有無，発症の有無を 1,0 で表した場合の相関係数[*13]。θ_1, θ_2 を発症者中の要因あり割合，非発症者中の要因あり割合として，

$$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$$

[*13] 相関係数については，第 11 章で詳しく説明するが，-1 から 1 までの値をとる量で，2 つの変数間にまったく関連がない場合に 0 となり，片方が大きくなればもう片方の変数も常に同じ割合で大きくなる関係があるとき 1 となる。

7.3 一致度の指標

7.3.1 κ 統計量

2回の繰り返し調査をしたときに，あるカテゴリ変数がどれくらい一致するかを示す指標である。

	2回目○	2回目×	合計
1回目○	a	b	m_1
1回目×	c	d	m_2
合計	n_1	n_2	N

という表から，偶然でもこれくらいは一致するだろうと思われる値は，1回目と2回目の間に関連がない場合の各セルの期待値を足して全数で割った値になるので $P_e = (n_1 \cdot m_1/N + n_2 \cdot m_2/N)/N$, 実際の一致割合（1回目も2回目も○か，1回目も2回目も×であった割合）は $P_o = (a+d)/N$ とわかる。ここで, $\kappa = (P_o - P_e)/(1-P_e)$ と定義すると，κ は，完全一致のとき1，偶然と同じとき0，それ以下で負となる統計量となる。

κ 統計量は，有意性の検定ができる。κ の分散 $V(\kappa) = P_e/(N \cdot (1-P_e))$ となるので，$\kappa/\sqrt{V(\kappa)}$ が標準正規分布に従うことを利用して検定できる。つまり，帰無仮説「κ が偶然一致する程度と差がない」が正しい確率が $1 - \mathtt{pnorm}(\kappa/\sqrt{V(\kappa)})$ となる[*14] 。この確率が5%未満ならば，得られた一致度は有意水準5%で信頼できる（偶然の一致より大きい）といえる。

κ 統計量の95%信頼区間は，

$$\kappa \pm \mathtt{qnorm}(0.975) \cdot \sqrt{P_o \cdot (1-P_o)/(N \cdot (1-P_e)^2)}$$

[*14] pnorm は正規分布の分布関数を表す R の関数である。上の表の記号を使って R のプログラムを書けば，

```
Pe<-(n1*m1/N+n2/N*m2)/N
Po<-(a+d)/N
kappa<-(Po-Pe)/(1-Pe)
SEkappazero<-sqrt(Pe/(N*(1-Pe)))
pkappa<-1-pnorm(kappa/SEkappazero)
cat("Kappa=",kappa," (p=",pkappa,")\n")
```

として計算できる[*15] 。なお κ 統計量は，2×2 だけでなく，m×mのクロス集計表に適用できる概念である。

7.4 利用例

本章で紹介した全ての指標を計算する関数 crosstab() を定義し，R の組み込みデータであるスイス女性の出生データに適用する例を挙げておくので，参考にされたい。

```
#
# Defining a function to combine several calculation for the indices of relationship.
# developed by Minato Nakazawa on 16th November 2001.
crosstab <- function(X) {
if (length(X)>4) stop("Given data cannot constitute 2x2 cross table")
cat(rep("=",35),"\n The results may include inappropriate statistics for given table\n
(e.g. Risk Ratio can stand only for cohort study). Take care.\n",rep("=",35),"\n")
a<-X[1,1]; b<-X[1,2]; c<-X[2,1]; d<-X[2,2]
m1<-a+b; m2<-c+d; N<-m1+m2; n1<-a+c; n2<-b+d
# risk ratio
RR<-(a*m2)/(c*m1)
RRL<-RR*exp(-qnorm(0.975)*sqrt(1/a-1/m1+1/c-1/m2))
RRU<-RR*exp(qnorm(0.975)*sqrt(1/a-1/m1+1/c-1/m2))
cat("Risk Ratio=",RR,"\t 95%CI=[",RRL,",",RRU,"]\n")
# odds ratio
OR<-(a*d)/(b*c)
ORL<-OR*exp(-qnorm(0.975)*sqrt(1/a+1/b+1/c+1/d))
ORU<-OR*exp(qnorm(0.975)*sqrt(1/a+1/b+1/c+1/d))
cat("Odds Ratio=",OR,"\t 95%CI=[",ORL,",",ORU,"]\n")
# risk difference
cat("Risk Difference=",RD<-a/m1-c/m2,"\n")
# relative difference
cat("Relative Difference=",RelD<-(a/m1-c/m2)/(1-c/m2),"\n")
# "曝露寄与率"
cat("AFe=",AFe<-(a/m1-c/m2)/(a/m1),"\n")
# "母集団寄与率"
cat("AFp=",AFp<-(n1/N-c/m2)/(n1/N),"\n")
# Yule's Q
cat("Yule's Q=",Q<-(OR-1)/(OR+1),"\n")
# "ファイ係数"
cat("phi coefficient=",rho<-sqrt((a/m1-c/m2)*(a/n1-b/m2)),"\n")
# kappa
Pe<-(n1*m1/N+n2/N*m2)/N
Po<-(a+d)/N
kappa<-(Po-Pe)/(1-Pe)
SEkappa<-sqrt(Po*(1-Po)/(N*(1-Pe)^2))
kappaL<-kappa-qnorm(0.975)*SEkappa
kappaU<-kappa+qnorm(0.975)*SEkappa
SEkappazero<-sqrt(Pe/(N*(1-Pe)))
pkappa<-1-pnorm(kappa/SEkappazero)
cat("Kappa=",kappa," (p=",pkappa,")\t 95%CI=[",kappaL,",",kappaU,"]\n")
}
```

[*15] pnorm は正規分布の分布関数を表す R の関数である。上の表の記号を使って R のプログラムを書けば，

```
Pe<-(n1*m1/N+n2/N*m2)/N
Po<-(a+d)/N
kappa<-(Po-Pe)/(1-Pe)
SEkappa<-sqrt(Po*(1-Po)/(N*(1-Pe)^2))
kappaL<-kappa-qnorm(0.975)*SEkappa
kappaU<-kappa+qnorm(0.975)*SEkappa
cat(95%CI=[",kappaL,",",kappaU,"]\n")
```

```
data(infert)
fewchild<-(infert$parity<=2)
noabort<-(infert$spontaneous==0)
Y<-table(fewchild,noabort)
print(Y)
# output of table(fewchild,noabort)
#
#         noabort
# fewchild FALSE TRUE
#    FALSE    43   25
#    TRUE     64  116
#
# i.e.
#           abort
# morechild  TRUE FALSE
#     TRUE     43    25
#    FALSE     64   116

# Chi-square test
print(chisq.test(Y))
# Fisher's exact test (where odds ratio is conditional MLE)
print(fisher.test(Y))

crosstab(Y)
# same as crosstab(matrix(c(43,64,25,116),nc=2))
```

第 8 章

平均値に関する推定と検定

本章では，量的な変数の代表値である平均値についての分析法を説明する。まず，何らかの標本データの平均値が母集団の平均値に一致するかどうかを探る場合を考えよう。例えば，山口市のある保健所で 100 人の 3 歳児の体重を測ったときに，その平均が全国平均として報告されている値と一致するかどうかを検討するような場合がこれに当たる。

8.1 母平均値と標本平均の差の検定

サイズ n の標本 X について，標本平均 $E(X) = \sum X/n$ と既知の母平均値 μ_X の差の検定は，母分散 V_X が既知のとき，$z_0 = |E(X) - \mu_X|/\sqrt{V_X/n}$ が標準正規分布に従うことを使って検定できる。

V_X が未知のときは，標本の不偏分散 $S_X = \sum (X_i - E(X))^2/(n-1) = \mathrm{var}(X)$ を使って，$t_0 = |E(X) - \mu_X|/\sqrt{S_X/n}$ が自由度 $n-1$ の t 分布に従うことを使って検定できる（暗黙の仮定として，ランダムサンプルで，母集団の分布が正規分布であることが必要)。

未知の母平均値の信頼区間の推定はこの裏返しである。 つまり, 母平均値の 95%信頼区間の下限は，不偏分散を標本数 n で割ったものの平方根に自由度 $n-1$ の t 分布の 97.5%点を掛けた値を標本平均から引いた値になり，上限は，同じ値を標本平均に足した値になる。なお，R では，変数 X について，`t.test(`X`,mu=`**既知の母平均値**`)` とすれば，上記の検定と推定を両方やってくれる。

例題 1

2001 年に厚生科学研究で行われた「少子化の見通しに関する専門家調査」の結果の一部を見てみる。この調査は，「人口学，経済学，家族社会学，公衆衛生学を中心とした専門家を対象として少子化研究会のメンバーが対象候補者を抽出し，回答者の偏りや不足等について検討を加えた上で，748 名を対象として調査を実施した」もので，回収率は 44 %であった。この調査では，2025 年の合計出生率がいくつになるかという推定値が尋ねられていて，生データを見ると，

```
1.38 1.50 1.30 1.40 1.40 1.15 1.31 1.37 1.50 1.55 1.55 1.56
1.50 1.56 1.50 1.38 1.50 1.20 1.20 1.50 1.25 1.25 1.22 1.40
1.80 1.37 1.35 1.70 1.35 1.50 ...（後略）
```

のようになっていた（回答数は 311，平均値は 1.385，不偏分散は 0.0252 であった）。調査用紙には，厚生労働省『人口動態統計』から，国立社会保障・人口問題研究所による低位推計 1.38，中位推計 1.61，高位推計 1.85 という情報が掲載されていた。仮にこれらの値を母平均とすると，専門家たちが出した推定値は，それに一致しているといえるだろうか？

仮に中位推計を母平均としたとき，得られたデータがそれに一致するかどうかを見てみよう。母分散は不明なので，S_X を使って，$t_0 = |1.385 - 1.61|/\sqrt{0.0252/311} = 25.0$ より，自由度 310 の t 分布で 25.0 の上側確率はほぼ 0 なので両側検定のために 2 倍しても[*1]ほぼ 0 であり，有意に異なるといえる。なお，元のデータを `x<-c(1.38,1.50,...)` のように `x` に付値しておけば，`t.test(x,mu=1.61,alternative="two.sided")` で母平均との差の検定結果が出力される。

低位推計を母平均としたときは，$t_0 = |1.385 - 1.38|/\sqrt{0.0252/311} = 0.555$ となり，自由度 310 の t 分布で上側確率をみると 0.289 となる。つまり，得られたデータの母平均が 1.38 と等しいという帰無仮説が成立する確率は $0.289 \times 2 = 0.579$ となり，棄却されない。

95%信頼区間を計算すると，下限が $1.385 - 1.968 \cdot \sqrt{0.0252/311} = 1.367$，上限が $1.385 + 1.968 \cdot \sqrt{0.0252/311} = 1.403$ となる。

[*1] 両側検定と片側検定については第 6 章脚注 2 でも触れたが，後で詳述する。

例題2

平成 10 年の国民栄養調査によれば，50-59 歳男性の平均 BMI は 23.6 であった。同じ年にA社の職員健診を受診した 50-59 歳男性 248 人の平均 BMI が 24.6 で，その不偏分散が 8.6 であったとき，A社の 50-59 歳男性は全国平均に比べて BMI に差があると言えるかどうか検定せよ。

母分散が未知なので, 標本の不偏分散で代用すれば, $t_0 = |24.6-23.6|/\sqrt{8.6/248} = 5.37$ より，自由度 247 の t 分布で 5.37 の上側確率はほぼ 0[*2]なので，両側検定のために2倍しても有意に異なるといえる。

8.2 独立2標本の平均値の差の検定

標本調査によって得られた独立した2つの量的変数 X と Y（サンプル数が各々 n_X と n_Y とする）について，母分散が既知で等しい V である場合は，$z_0 = |E(X) - E(Y)|/\sqrt{V/n_X + V/n_Y}$ が標準正規分布に従うことを使って検定する[*3]。

8.2.1 母分散が未知の場合

調査データを分析する場合は母分散が既知であることはほとんどなく，これが普通である。手順としては以下の通りである。

1. F 検定（分散が等しいかどうか）：2つの量的変数 X と Y の不偏分散の大きい方を小さい方で割った $F_0 = S_X/S_Y$ が第 1 自由度 $n_X - 1$，第 2 自由度 $n_Y - 1$ の F 分布に従うことを使って検定する。[*4]
2. 分散に差があるか差がないかによって，平均値が等しいかどうかの検定法は異なる。[*5]

[*2] R で `1-pt(5.37,247)` を計算すると結果が `9.081154e-08` と表示される（9.081154×10^{-8} という意味）。

[*3] 分布がひどく歪んでいる場合には，Mann-Whitney の U 検定（Wilcoxon の順位和検定ともいう）を行う。詳細は次章で説明するが，その場合は，代表値としても平均値と標準偏差でなく，中央値と四分位偏差を表示するのが相応しい。

[*4] R では `1-pf(`F_0`,`$n_X - 1$`,`$n_Y - 1$`)` が有意確率になる。しかし，F_0 を手計算しなくても，`var.test(`X`,`Y`)` で等分散かどうかの検定が実行できる。この場合は，R が勝手に入れ替えてくれるので，X の不偏分散の方が Y の不偏分散より大きいかどうか気にしなくてもよい。

[*5] 分散に差があるだけでも，別の母集団からとられた標本であると判断して，平均値が等しいかどうかを検定する意味はないとする考え方もありうるが，Welch の方法を使うか，ノンパラメトリックな方法を使って検定するのが普通である。

8.2.2 分散に差がない場合

母分散 S を $S = [(n_X - 1)S_X + (n_Y - 1)S_Y]/(n_X + n_Y - 2)$ として推定し，$t_0 = |E(X) - E(Y)|/\sqrt{S/n_X + S/n_Y}$ が自由度 $n_X + n_Y - 2$ の t 分布に従うことを利用して検定する。[*6]

8.2.3 分散が差がある場合（Welch の方法）

$t_0 = |E(X) - E(Y)|/\sqrt{S_X/n_X + S_Y/n_Y}$ が自由度 ϕ の t 分布に従うことを使って検定する[*7]。但し

$$\phi = \frac{(S_X/n_X + S_Y/n_Y)^2}{\{(S_X/n_X)^2/(n_X - 1) + (S_Y/n_Y)^2/(n_Y - 1)\}}$$

例題 3

先の調査では，少子化に対するイメージを問う質問項目もあった。「明るいイメージ」「どちらかといえば明るいイメージ」「どちらかといえば暗いイメージ」「暗いイメージ」の 4 つから選ぶのだが，仮に「明るいイメージ」または「どちらかといえば明るいイメージ」と答えた人（楽観主義と呼ぶことにする）と，「どちらかといえば暗いイメージ」または「暗いイメージ」と答えた人（悲観主義と呼ぶことにする）に分けると，楽観主義と悲観主義で比べたら，2025 年の合計出生率の推定値に有意な差はあるだろうか？

楽観主義のデータは，

```
1.40 1.15 1.55 1.56 1.50 1.50 1.20 1.80 1.30 1.54 1.40 1.60 1.50
1.25 1.38 1.50 1.30 1.35 1.50 1.20 1.70 1.60 1.40 1.30 1.05 1.62
1.25 1.40 1.50 1.10 ...  （後略）
```

となっていて（サンプル数 $n_X = 68$，平均 $E(X) = 1.384$，不偏分散 $S_X = 0.0337$），悲観主義のデータは，

[*6] R では，`t.test(X,Y,var.equal=T)` とすれば検定してくれる。

[*7] R では，`t.test(X,Y)`（または `t.test(X,Y,var.equal=F)` だが，`var.equal` の指定を省略した時は等分散でないと仮定して Welch の検定がなされるので省略していい）とすれば検定してくれる。

```
1.38 1.50 1.30 1.40 1.31 1.37 1.50 1.55 1.56 1.50 1.38 1.20 1.50
1.25 1.25 1.22 1.40 1.37 1.35 1.70 1.35 1.55 1.60 1.70 1.20 1.31
1.40 1.40 1.60 1.10 ...  （後略）
```

となっていて（サンプル数 $n_Y = 235$, 平均 $E(Y) = 1.383$, 不偏分散 $S_Y = 0.0234$），生のデータを見ただけでは差があるかないかさっぱりわからない。そこでまず思いつくのが，平均値を比べてやろうということである。楽観主義でも悲観主義でも同じ母集団に属していて，合計出生率の推定そのものには差がないと仮定すると，これらのデータは平均も分散も一致するはずである。実際はどうなっているだろうか？

母分散は不明なので，まず F 検定を行う。$F_0 = 0.0337/0.0234 = 1.443$ なので，第 1 自由度 67，第 2 自由度 234 の F 分布で上側確率を計算すると，0.0246 となり，分散が等しいという帰無仮説は棄却される。そこで，Welch の方法によって検定を行うと，$t_0 = 0.031, \phi = 95.465$ となるので，有意確率は 0.9753 であり，2 群の平均値に差がないという帰無仮説は採択される。よって，楽観主義でも悲観主義でも 2025 年の合計出生率の推測値には差がないといえる。

例題 4

件の専門家調査には，出生率がそのうち回復するとみるか，低下し続けるとみるかという質問項目もあり，この答えの違いによって，2025 年の予測値には違いがありそうである。

回復するとみる人たちの 2025 年の合計出生率の予測値は，

```
1.40 1.40 1.56 1.50 1.40 1.80 1.37 1.40 1.40 1.60 1.60 1.25 1.50
1.50 1.70 ...  （後略）
```

となっており（サンプル数 58，平均 1.487，不偏分散 0.0275），低下し続けるとみる人たちの予測値は，

```
1.38 1.30 1.15 1.31 1.37 1.50 1.55 1.55 1.56 1.50 1.50 1.38
1.20 1.20 1.25 ...  （後略）
```

となっている（サンプル数 221，平均 1.356，不偏分散 0.0211）。2 群間に違いがあると言っていいか？

R で計算すると，

```
F0 <- 0.0275/0.0211
1-pf(F0,57,220)
```

とすれば，0.0928 という結果が得られるので分散に有意差はないといえる。従って，Welch にしなくていい。続けて R で計算すると，

```
S <- ((58-1)*1.487+(221-1)*1.356)/(58+221-2)
t0 <- abs(1.487-1.356)/sqrt(S/58+S/221)
2*(1-pt(t0,58+221-2))
```

の結果として 7.966E-09 が得られ（7.966×10^{-9} という意味），ほぼゼロに近いので，平均には有意差があるといえる。もっとも，R では，X と Y に各群の生データを付値して，t.test(X,Y,var.equal=T) とすれば，同じ結果が得られる。通常はそれで十分である。

8.3　両側検定と片側検定

これまでも何度か両側検定をしてきたが，ここで両側検定と片側検定の意味をもう一度きちんと押さえておこう。

2 つの量的変数 X と Y の平均値の差の検定をする場合，それぞれの母平均を μ_X，μ_Y と書けば，その推定量は $\mu_X = \mathrm{mean}(X) = \sum X/n$ と $\mu_Y = \mathrm{mean}(Y) = \sum Y/n$ となる。

両側検定では，帰無仮説 H0 : $\mu_X = \mu_Y$ に対して対立仮説（帰無仮説が棄却された場合に採択される仮説）H1 : $\mu_X \neq \mu_Y$ である。H1 を書き直すと，「$\mu_X > \mu_Y$ または $\mu_X < \mu_Y$」ということである。つまり，t_0 を「平均値の差を標準誤差で割った値」として求めると，t_0 が負になる場合も正になる場合もあるので，有意水準 5%で検定して有意になる場合というのは，t_0 が負で t 分布の下側 2.5%点より小さい場合と，t_0 が正で t 分布の上側 2.5%点（つまり 97.5%点）より大きい場合の両方を含む。t 分布は原点について対称なので，結局両側検定の場合は，上述のように差の絶対値を分子にして，t_0 の t 分布の上側確率[*8]を 2 倍すれば有意確率が得られることになる。

片側検定は，先験的に X と Y の間に大小関係が仮定できる場合に行い，例えば，X の方が Y より小さくなっているかどうかを検定したい場合なら，帰無仮説 H0 : $\mu_X \geq \mu_Y$ に対して対立仮説 H1 : $\mu_X < \mu_Y$ となる。この場合は，t_0 が正になる場合だけ考えればよい。有意水準 5%で検定して有意になるのは，t_0 が t 分布の上側 5%点（つまり 95%点）より大きい場合である。R で片側検定をしたい

[*8] t 分布の確率密度関数を t_0 から無限大まで積分した値，即ち，t 分布の分布関数の t_0 のところの値を 1 から引いた値。R では 1-pt(t0, **自由度**)。

場合は，alternative という指定を追加する。例えば，$X > Y$ が対立仮説なら，t.test(X,Y,alternative="greater") とする。指定しなければ両側検定である。alternative に指定できる文字列は，greater の他には less と two.sided がある（指定しない場合は two.sided を指定したのと同じ意味，つまり両側検定になる）。

8.4 対応のある 2 標本の平均値の差の検定

例えば，先に説明した専門家調査の結果で，2005 年の予測値と 2025 年の予測値に差があるかないかという問題を考えよう。この場合は同じ人について両方の値があるので，全体の平均に差があるかないかだけをみるのではなく，個人ごとの違いを見るほうが情報量が失われない。このような場合は，独立 2 標本の平均値の差の検定をするよりも，対応のある 2 標本として分析する方が切れ味がよい（差の検出力が高い）[*9]。対応のある 2 標本の差の検定は，paired-t 検定と呼ばれ，意味合いとしてはペア間の値の差を計算して値の差の母平均が 0 であるかどうかを調べることになる。R で対応のある変数 X と Y の paired-t 検定をするには，t.test(X,Y,paired=T) で実行できるし，それは t.test(X-Y,mu=0) と等価である。2025 年の予測値は，

```
1.38 1.50 1.30 1.40 1.40 1.15 1.31 1.37 1.50 1.55 1.55 1.56 1.50
1.56 1.50 1.38 1.50 1.20 1.20 1.50 1.25 1.25 1.22 1.40 1.80 1.37
1.35 1.70 1.35 1.50 ...（後略）
```

のようになっていた（回答数は 311，平均値は 1.385，不偏分散は 0.0252）。2005 年の予測値は，

```
1.30 1.35 1.34 1.35 1.32 1.25 1.34 1.34 1.40 1.40 1.35 1.30 1.30
1.32 1.35 1.39 1.30 1.30 1.30 1.20 1.33 1.35 1.30 1.37 1.40 1.33
1.39 1.35 1.35 1.30 ...（後略）
```

であった（回答数は 311，平均値は 1.334，不偏分散は 0.00259）。これを普通に t 検定するなら，明らかに分散が異なるので，Welch の検定によって $t_0 = 5.37$，自由度が 373.1 より両側検定の有意確率は 1.37×10^{-7} となるが，対応のある t 検定をすると，2025 年と 2005 年の予測値の差が，

[*9] 分布が歪んでいる場合や，分布が仮定できない場合の対応のある 2 標本の分布の位置の差があるかどうか検定するには，ウィルコクソンの符号順位検定を用いる。詳しくは次章で説明するが，R では wilcox.test(**変数 1**, **変数 2**,paired=T) で実行できる。この場合も U 検定のときと同じく，代表値は中央値と四分位偏差で表示するべきである。

```
-0.08 -0.15 0.04 -0.05 -0.08 0.10 0.03 -0.03 -0.10 -0.15 -0.20
-0.26 -0.20 -0.24 -0.15 0.01 -0.20 0.10 0.10 -0.30 0.08 0.10
0.08 -0.03 ...（後略）
```

となりサンプル数 311，平均 -0.0508，不偏分散 0.0192 より，$t_0 = 6.46$ となり自由度 310 の t 分布で上側確率を求めて 2 倍すれば，$p = 3.942 \times 10^{-10}$ となり，こちらの方が有意確率は小さくなる。いずれにせよ 5%よりずっと小さいので，2025 年の予測値と 2005 年の予測値は 5%水準で有意に異なるといえる。

第 9 章

２群の差に関するノンパラメトリックな検定

9.1 ノンパラメトリックな検定とは？

パラメータ（parameter）とは母数という意味である。これまで説明してきた検定法の多くは，母数，つまり母集団の分布に関する何らかの仮定をおいていた。その意味で，t 検定も F 検定もパラメトリックな分析法といえる。一方，フィッシャーの正確な確率は母数を仮定しないのでパラメトリックでない。ノンパラメトリックな分析とは，パラメトリックでない分析，つまり母数を仮定しない分析をさす[*1]。

問題を定式化すると，次のようになる。

1. 標本データ $X_1, X_2, ..., X_n$ が互いに独立に分布 $F(x)$ に従い，別の標本データ $Y_1, Y_2, ..., Y_n$ が互いに独立かつ X とも独立で分布 $G(y)$ に従う。
2. F と G には連続分布であるという以外には制約をおかない。
3. このとき，「２つの分布に差はない」という帰無仮説（$H : F(x) \equiv G(x)$）を検定する。

[*1] ただし，厳密に考えると区分はそれほど明確ではない。例えば，カイ二乗検定では母集団の分布には特定の仮定は置いていないので，定義からすると，実はノンパラメトリックな分析になる。ただし，カイ二乗統計量がカイ二乗分布に従うためにはデータ数が十分に多いことが必要である。もっとも，そう言ってしまえば正規近似する場合の順位和検定もデータ数が多いことが必要なので，問題は何を検定の本質と見なすかという話になってくる。一般には，量的な変数を分析するのに，量の情報を使わずに大小関係，即ち順位の情報だけを使う分析をノンパラメトリックな解析と呼ぶことが多い。

つまり，ノンパラメトリックな検定では，「母数を仮定しない」とは言っても，**連続分布である**ことだけは仮定する。もっとも理想的には分布の形が同じで位置だけがずれているという，「ズレのモデル」が仮定できると話は簡単である。

2 群の差に関するノンパラメトリックな検定の具体的な方法としては，前章でも軽く触れたように，Wilcoxon の順位和検定（Mann-Whitney の U 検定），符号付順位和検定，符号検定などがある。得られたデータがある種の経験分布関数に一致するかどうかを調べるために良く使われる検定法としてはコルモゴロフ＝スミルノフ検定（KS 検定）がある[*2]。

パラメトリックな分析法が使える前提としては，理想的には母集団の分布は正規分布にしたがっていなくてはならない。しかし，実際には正規分布にしたがっていない場合もある。この場合の戦略としては，(1) 対数正規分布とかガンマ分布のような別な分布を考える，(2) 正規分布に近づくような変換を施す，といったことが考えられるが，真の分布がわかっていないためにうまく行くとは限らない。そこで，ヒストグラムを描いてみて，どうも正規分布ではなさそうだと思ったら，分布によらない方法を試してみるというのも一案である。2 群の分布の位置の差に関する検定の場合，Wilcoxon の順位和検定の検出力は，最良の場合の t 検定の 95%程度だが，分布が歪んでいる場合には t 検定よりも検出力が良くなる場合もある。

本章のテーマについて詳しく知りたい場合は，竹内，大橋 (1981) の第 4 章や，伊藤ら (1984) の第 2 章を参照されたい。

9.2　Wilcoxon の順位和検定

Wilcoxon の順位和検定は，Mann-Whitney の U 検定と（見かけはちょっと違うが）同じ内容の検定である（詳しくは後述する）。

データがもつ情報の中で，単調変換に対して頑健なのは順位なので，これを使って検定しようという発想である。以下，手順を箇条書きする。

1. 変数 X のデータを $x_1, x_2, ..., x_m$ とし，変数 Y のデータを $y_1, y_2, ..., y_n$ とする。
2. まず，これらをまぜこぜにして小さい方から順に番号をつける[*3]。例えば，$x_8[1], y_2[2], y_{17}[3], ..., x_4[N]$ のようになる（但し $N = m + n$）。
3. ここで問題にしたいのは，それぞれの変数の順位の合計がいくつになるかとい

[*2] 説明は省略するが，R では `ks.test(変数 1, 変数 2)` で実行可能である。

[*3] 同順位がある場合の扱いは後述する。

うことである。ただし，順位の総合計は $(N+1)N/2$ に決まっているので，片方の変数だけ考えれば残りは引き算でわかる。そこで，変数 X だけ考えることにする。

4. X に属する x_i（$i = 1, 2, ..., m$）の順位を R_i と書くと，X の順位の合計は

$$R_X = \sum_{i=1}^{m} R_i$$

となる。R_X があまり大きすぎたり小さすぎたりすると，X の分布と Y の分布に差がないという帰無仮説が疑わしいと判断されるわけである。では，帰無仮説が成り立つ場合に，R_X はどのくらいの値になるのだろうか？[*4]

5. もし X と Y に差がなければ，X は N 個のサンプルから偶然によって m 個取り出したものであり，Y がその残りである，と考えることができる。順位についてみると，$1, 2, 3, ..., N$ の順位から m 個の数値を取り出すことになる。ありうる組み合わせは，${}_NC_m$ 通りある[*5]。

6. $X > Y$ の場合には，${}_NC_m$ 通りのうち，合計順位が R_X と等しいかより大きい場合の数を k とする（$X < Y$ の場合は，合計順位が R_X と等しいかより小さい場合の数を k とする）。

7. $k/{}_NC_m$ が有意水準 α より小さいときに H_0 を疑う。N が小さいときは有意になりにくいが，N が大きすぎると計算が大変面倒である[*6]。そこで，正規近似を行う（つまり，期待値と分散を求めて，統計量から期待値を引いて分散の平方根で割った値が標準正規分布に近似的に従うという関係を用いて検定する）。

8. 帰無仮説 H_0 のもとでは，期待値は

$$E(R) = \sum_{i=1}^{m} E(R_i) = m(1 + 2 + ... + N)/N = m(N+1)/2$$

[*4] 以下説明するように，順位和 R をそのまま検定統計量として用いるのが Wilcoxon の順位和検定であり，R_X, R_Y の代わりに，$U_X = mn + n(n+1)/2 - R_Y$，$U_Y = mn + m(m+1)/2 - R_X$ として，U_X と U_Y の小さいほうを U として検定統計量として用いるのが，Mann-Whitney の U 検定である。有意確率を求めるために参照する表は違うが，数学的には同じ意味をもつ。R では，Wilcoxon の順位和統計量の分布関数が提供されているので，例えばここで得られた順位和を RS と書くことにすると，`2*(1-pwilcox(RS,m,n))` で両側検定の正確な有意確率が得られる。

[*5] R では `choose(N,m)`。

[*6] もっとも，今ではコンピュータにやらせればよい。例えば R であれば，`wilcox.test(X,Y,exact=T)` とすれば，サンプル数の合計が 50 未満で同順位の値がなければ，総当りして正確な確率を計算してくれる。が，つい 15 年くらいまではコンピュータは誰もが使える道具ではなかったし，総当りをするには計算時間がかかりすぎた。今のコンピュータでもサンプルサイズが大きいと，総当りでは計算時間がかかりすぎて実用的でない。

(1 から N までの値を等確率 $1/N$ でとるから)。分散はちょっと面倒で,

$$var(R) = E(R^2) - (E(R)^2)$$

から,

$$E(R^2) = E((\sum_{i=1}^{m} R_i)^2) = \sum_{i=1}^{m} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

となるので[*7],

$$E(R_i^2) = (1^2 + 2^2 + ... + N^2)/N = (N+1)(2N+1)/6$$

と

$$E(R_i R_j) = \frac{1}{N(N-1)}\{(\sum_{k=1}^{N} k)^2 - \sum_{k=1}^{N} k\}$$

$$= \frac{1}{N(N-1)}(\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6})$$

$$= \frac{(N+1)(3N+2)}{12}$$

を代入して整理すると,結局, $var(R_X) = m(N+1)(N-m)/12 = mn(N+1)/12$ となる。

9. 標準化[*8]して連続修正[*9]し, $z_0 = \{|R_X - E(R_X)| - 1/2\}/\sqrt{var(R_X)}$ を求める。m と n が共に大きければこの値が標準正規分布に従うので,例えば $z_0 > 1.96$ ならば,両側検定で有意水準 5%で有意である。R で有意確率を求めるには, z_0 を `z0` と書けば, `2*(1-pnorm(z0,0,1)` とすればよい。
10. ただし,同順位があった場合は,ステップ 2) の「小さい方から順に番号をつける」ところで困ってしまう。例えば,変数 X が $\{2,6,3,5\}$, 変数 Y が $\{4,7,3,1\}$ であるような場合には, X にも Y にも 3 という値が含まれる。こ

[*7] 第 1 項が対角成分,第 2 項がそれ以外に相当する。$m = 2$ の場合を考えてやればわかるが,

$$E((\sum_{i=1}^{2} R_i)^2) = E((R_1 + R_2)^2) = E(R_1^2 + R_2^2 + 2R_1R_2) = \sum_{i=1}^{2} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

となる。

[*8] 何度も出てくるが,平均(期待値)を引いて分散の平方根で割る操作である。

[*9] これも何度も出てくるが,連続分布に近づけるために 1/2 を引く操作である。

ういう場合は，下表のように平均順位を両方に与えることで，とりあえず解決できる。

属する変数	Y	X	X	Y	Y	X	X	Y
値	1	2	3	3	4	5	6	7
順位	1	2	3.5	3.5	5	6	7	8

11. ただし，このやり方では，正規近似をする場合に分散が変わる[*10]。帰無仮説の下で，$E(R_X) = m(N+1)/2$ はステップ 8) と同じだが，分散が

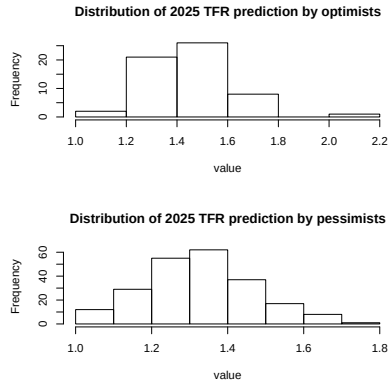
$$var(R_X) = mn(N+1)/12 - mn/\{12N(N-1)\} \cdot \sum_{t=1}^{T}(d_t^3 - d_t)$$

となる。ここで T は同順位が存在する値の総数であり，d_t は t 番目の同順位のところにいくつのデータが重なっているかを示す。上の例では，$T = 1$, $d_1 = 2$ となる。なお，あまりに同順位のものが多い場合は，この程度の補正では追いつかないので，値の大小があるクロス集計表として分析するべきである（例えば Cochran-Armitage 検定などが考えられる）。

[*10] 正確な確率を求めるならば問題ない。

例題 1

前章で紹介した，少子化の見通しに関する専門家調査の結果を再び取り上げることにする。低出生率は今後回復すると見る 58 人と，このまま出生率は低下し続けると見る 221 人の間で，2025 年の合計出生率の予測値には t 検定で差があったわけだが，予測値の分布を調べると，下図のようにかなり右裾を引いた歪んだ形になっていることがわかる。ノンパラメトリックな検定である Wilcoxon の順位和検定で検定をやり直してみよう。



まず，それぞれの順位を計算する。R で 2 群を合わせた順位を計算するには，データフレームに戻すと便利である。今後回復すると見る人の予測値を示す変数を opt，低下し続けると見る人の予測値を示す変数を pes とすると，

```
d <- data.frame(gr=c(rep(1,NROW(opt)),rep(2,NROW(pes)),val=c(opt,pes))
```

とすれば，d というデータフレームができる。次に，順位を得る関数 rank() を使って，rnk <- rank(d$val) とし，これを元のデータフレームに結合して，dd <- data.frame(d,rk=rnk) とすれば，opt の順位の合計が sum(dd$rk[dd$gr==1]) によって得られ，pes の順位の合計が sum(dd$rk[dd$gr==2]) で得られる。opt と pes の順位はそれぞれ，

```
(opt) 165.5 165.5 245.5 218.0 165.5 277.0 131.0 ...（後略）
(pes) 141.0  83.0  17.5 103.0 131.0 218.0 241.0 ...（後略）
```

となり，それぞれの合計は，11157，27903 となる[*11]。つまり，$R_{opt} = 11157$ である。ここでは，opt の方が pes より確率的に大きいと考えられるので，opt≤pes を帰無仮説として片側検定をする。本当は同順位がいくつもあるので分散の計算が面倒なのだが，とりあえず同順位の存在を無視すると，R_{opt} の期待値 $\mathrm{E}(R_{opt})$ は，$\mathrm{E}(R_{opt})=58*(58+221+1)/2 = 8120$ となり，分散 $\mathrm{V}(R_{opt})$ は，$\mathrm{V}(R_{opt})=58*221*(58+221+1)/12 = 299086.7$ となるので，

$$\frac{R_{opt} - E(R_{opt}) - 1/2}{\sqrt{V(R_{opt})}} = 5.552323$$

より，`1-pnorm(5.552323,0,1)`$=1.4 \times 10^{-8}$ である。この値はほとんどゼロなので帰無仮説は棄却され，opt は pes より大きいといえる。なお，R で同順位を考慮して計算すると `wilcox.test(opt,pes,alternative="greater",exact=F)` で良く，$p = 1.248 \times 10^{-8}$ となり，大差ないことがわかる。これくらいデータ数が大きくて同順位が少なければ，同順位を考慮せずに計算しても，有意確率のオーダーは変わらない。

練習問題

ある大学の学生実習で水道水の水質検査をしたところ，遊離残留塩素濃度 (mg/L) が，集合住宅群 (A) では {0.3346, 0.6230, 0.8580, 1.1031, 0.4580, 0.6210, 0.9071, 0.4760, 0.5020, 0.9670, 0.7100, 0.1350, 1.1390, 0.5741, 0.9090, 1.0400, 0.4190, 0.6296, 1.1080, 0.5793, 1.0420, 1.2826, 1.8280, 0.1630} で，一戸建て群 (B) では {0.8583, 0.9320, 0.4220, 0.3570, 0.0641, 0.5338, 0.8280, 1.1400, 0.7229, 0.0000} であったとしよう（下表のように，この 2 群間には同順位はない)。この 2 群の間で遊離残留塩素濃度に差があるかどうか，Wilcoxon の順位和検定で調べてみよ。

(A)										
値	0.3346	0.6230	0.8580	1.1031	0.4580	0.6210	0.9071	0.4760	0.5020	0.9670
順位	5	16	21	29	9	15	23	10	11	26
値	0.7100	0.1350	1.1390	0.5741	0.9090	1.0400	0.4190	0.6296	1.1080	0.5793
順位	18	3	31	13	24	27	7	17	30	14
値	1.0420	1.2826	1.8280	0.1630						
順位	28	33	34	4						
(B)										
値	0.8583	0.9320	0.4220	0.3570	0.0641	0.5338	0.8280	1.1400	0.7229	0.0000
順位	22	25	8	6	2	12	20	32	19	1

詳しい計算手順は省くが，実際に試されたい。なお，R の `wilcox.test()` 関数を

[*11] もっとも，`d<-cbind(c(rep(1,NROW(opt)),rep(2,NROW(pes))),c(opt,pes))` かつ `x<-rank(d[,2])` としてから，`x[d[,1]==1]` として opt 群の順位を，`x[d[,1]==2]` として pes 群の順位を得ることもできる。

使って検定すると，正規近似では $p = 0.2986$，正確な確率では $p = 0.3040$ となる。すなわち，(A) と (B) の 2 群間に 5%水準で有意差はないことがわかる[*12]。

9.3　正規スコア検定

Wilcoxon の順位和検定では R_i として順位そのものを用いたが，これは大小関係が保存されるならば順位の代わりに適当なスコア $s(R_i)$ を使って構わない。スコアとして正規スコアを用いるのが，正規スコア検定である。

正規スコアとしては，$Z_{(1|N)} \leq Z_{(2|N)} \leq ... \leq Z_{(N|N)}$ を標準正規分布からの大きさ N の順序統計量としたとき，

$$s(R_i) = E(Z_{(i|N)}) \simeq \Phi^{-1}\left(\frac{i}{N+1}\right)$$

を用いる。正規スコア検定は，ズレのモデルと分布の正規性の仮定の下では，t 検定と漸近的に同等である。

9.4　メディアン検定

$s(R_i)$ として $i \geq [(N+1)/2]$ のとき 1，$i < [(N+1)/2]$ のとき 0 を用いるのがメディアン検定である。次のように言い換えることもできる。

m 個のデータからなる X と n 個のデータからなる Y を合わせた $N = m + n$ 個のデータを，全体のメディアン以上かメディアンより小さいかによって分類すると，以下の 2 × 2 クロス集計表が得られる。

[*12] 「有意」という考え方は重要なので，念のため復習しておく。この例題では，どちらが高いとか低いとかいった事前情報はないので，「集合住宅群と一戸建て群の間で水道水の遊離残留塩素濃度に差はない」を帰無仮説として両側検定をする。「有意水準を 5%にする」とは，「帰無仮説が偶然に成り立つ確率が 5%未満であれば，統計的に意味があるほど稀な現象なので帰無仮説は成り立たないとみなす」ということなので，「5%水準で有意でない」といえば，「帰無仮説が偶然に成り立つ確率が 5 %未満であれば，統計的に意味があるほど稀な現象なので帰無仮説は成り立たないとみなす，としたのに，データから計算するとその確率が 5%より大きくなってしまったので，統計的に意味があるほど稀ではなく，帰無仮説が成り立たないとはみなせない」ということになる。この例でいえば，有意水準を 5%にしたのに，「集合住宅群と一戸建て群の間で水道水の遊離残留塩素濃度に差がない」条件下で，実際に得られているデータが偶然得られる確率は 5%より大きいので，「差がない」という帰無仮説が棄却されなかったということを意味するわけである。

	X	Y	合計
メディアン以上	H	$(m+n/2)-H$	$(m+n)/2$
メディアンより小さい	$m-H$	$H+(n-m)/2$	$(m+n)/2$
合計	m	n	$m+n$

帰無仮説の下では H は $m/2$ の周りに分布する（超幾何分布）ので，$Pr(H = h') =_n C_{h'} \cdot_n C_{(m+n)/2-h'} / {}_{m+n}C_{(m+n)/2}$ より，$Pr(H \geq h')$ をすべて合計して 2 倍すれば，両側検定での有意確率が得られる。

9.5 符号付き順位和検定

2 群間の各サンプルに対応がある場合には，単純な順位和検定よりも切れ味がよい方法がある。符号化順位検定とも呼ばれるこの方法は，対応のある t 検定の場合と同じような考え方に基づく。

変数 X の任意の i 番目（i は 1 から n までの整数値）のデータが $x_i = e_i + \theta_i$ のように，誤差変動 e_i と真の効果 θ_i の和であると捉えれば，もし X と Y が同じ母集団からのサンプルであるならば $X-Y$ により X と Y に共通する真の効果を打ち消すことができて，$U_i = x_i - y_i = e_i - e\prime_i$ が得られる。このとき帰無仮説は，e_i と $e\prime_1$ の分布が同じということなので，U_i は原点に対して対称になるはずである。そこで，U_i の絶対値が小さい方から順に順位 R_i をつける。さらに，$\varepsilon_i = 1(U_i > 0), \varepsilon_i = -1(U_i < 0)$ とすれば，帰無仮説の下で $Pr(\varepsilon_i = 1) = Pr(\varepsilon_i = -1) = 1/2$ となる。いま，

$$R^* = \sum_{i=1}^{n} \varepsilon_i R_i$$

とおけば，R^* の大きさによって検定ができる。

すべての場合（ε_i の値が各 i について 2 通りあるので，2^n 通り）を計算してやれば正確な確率が計算できるが[*13]，n が大きくなると計算が大変なので，$n \geq 15$ ならば近似を行ってよいことになっている。R^* の期待値は

$$E(R^*) = \sum_{i=1}^{n} R_i E(\varepsilon_i) = \sum_{i=1}^{n} R_i (1 \times 1/2 + (-1) \times 1/2) = 0$$

[*13] この正確な確率の計算法は，R. A. Fisher が考案した「並べかえ検定」(permutation test) と呼ばれている。後述する。

分散は

$$\mathrm{var}(R^*) = \sum_{i=1}^{n} {R_i}^2 var(\varepsilon_i)$$

$$= \sum_{i=1}^{n} {R_i}^2 (1^2 \times 1/2 + (-1)^2 \times 1/2)$$

$$= \sum_{i=1}^{n} {R_i}^2 = n(n+1)(2n+1)/6$$

となるので，標準化と連続性の補正をして，

$$\frac{|R^*| - 1/2}{\sqrt{var(R^*)}}$$

が標準正規分布に従うことを利用して検定する。なお，R では，対応のある 2 群の生のデータを `X` と `Y` に付値しておき，`wilcox.test(X,Y,paired=TRUE)` とすればこの検定ができる。

9.6　符号検定

対応がない場合と同様，対応がある場合でも，R_i という順位そのものを用いる代わりに，スコア $s(R_i)$ を使うことが可能である。ただし，すべての $1 \leq i, j \leq n$ について，$R_i \leq R_j$ ならば $s(R_i) \leq s(R_j)$ となっている必要がある。

もっとも単純なスコアとして，すべての i について $s(R_i) = 1$ とすることを考えると，

$$R^* = \sum_{i=1}^{n} \varepsilon_i$$

となるので，これは $U_i = x_i - y_i$ が正となるオブザーベーション数 K から負のオブザーベーション数を引いた値となり，$2K - n$ に等しくなるので，K をそのまま検定統計量としてもよい。

帰無仮説の下での K の分布を考える。X と Y に差がなければ U_i が正となるか負となるかは確率 $1/2$ で起こるので，n 個のオブザーベーション中で正のオブザーベーション数が x 個になる確率 $p(x)$ は 2 項分布で表され，$p(x) =_n C_x \cdot (1/2)^x \cdot (1/2)^{n-x} =_n C_x \cdot (1/2)^n$ となる。

したがって，K よりも稀な値が偶然得られる確率は，$K > n/2$ の場合には，$2 \times \sum_{x=K}^{n} p(x)$ であり，この値が両側検定の有意水準と考えられる。なお，対立仮説が

$X > Y$ である片側検定の場合は，有意確率は $\sum_{x=K}^{n} p(x)$ となる。

もちろん，n が大きければ，2項分布は正規近似することができるが，n が小さいときに近似を用いずに確率を簡単に計算できるのが符号検定の利点である。

9.7 並べ換え検定

前述の通り R. A. Fisher が考案したが，コンピュータが発達するまでは計算量が多すぎて，データが少ない場合にしか使えなかった。コンピュータ集約型統計学の代表的な手法の1つで，分布に依存しない正確な確率が出せる。

例えば，符号付き順位和検定で正確な確率を計算するには，すべてのありうる R^* を計算して，その絶対値が R^* の実現値の絶対値以上の値になる場合の数 M を数えれば，$M/2^n$ が有意確率となる。R^* が正の場合だけ考えて2倍しても，ほぼ同じ値になる（片側検定の場合は2倍しなくてよい）。

R では，CRAN から `install.packages("exactRankTests")` として追加パッケージをインストールし，`library(exactRankTests)` とすれば，`perm.test()` という関数で並べ換え検定が可能になる。

第 10 章

多群間の差を調べる〜一元配置分散分析と多重比較

10.1　多群間の比較を考える

t 検定や順位和検定では 2 群間の差を比べた。では，3 群以上の場合はどうしたらいいだろうか？

単純に 2 群間の差の検定を繰り返してはいけない。なぜなら，n 群から 2 群を抽出するやりかたは ${}_nC_2$ 通りあって，1 回あたりの第 1 種の過誤を 5%未満にしたとしても，3 群以上の比較全体として「少なくとも 1 組の差のある群がある」というと，全体としての第 1 種の過誤が 5%よりずっと大きくなってしまうからである。

この問題を解消するには，大別して 2 つのアプローチがある。1 つは，多群間の比較という捉え方をやめて，群分け変数が注目している量の変数に与える効果があるかどうかという捉え方にする，というアプローチである。具体例でいえば，東京と長野と山口で年降水量の平均に差があるかどうかを見たいときに，東京と長野，長野と山口，という具合に比べるのではなくて，年降水量という変数に対して，地域という変数が有意な効果をもっているかどうか？　と立論するのである。このやり方に当たるのが一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定（ノンパラメトリックな一元配置分散分析）である。

もう 1 つのアプローチは，有意水準 5%の 2 群間の検定を繰り返すことによって全体としては大きくなってしまう第 1 種の過誤を調整することによって，全体としての検定の有意水準を 5%に抑えることである。このやり方は「多重比較法」と呼ばれる。さまざまな方法が提案されているが，中には数学的に不適切なものが歴史的に古くか

ら使われているからというだけの理由で使われ続けている場合もあり，注意が必要である。

これら 2 つのアプローチは別々に行うというよりも，段階を踏んで行うものと考えるのが一般的である[*1]。一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定によって群間に何らかの差があると結論されてから，初めてどの群とどの群の差があるのかを調べるために多重比較法を使うというわけである。その意味で，多重比較法は *post hoc* な解析と呼ばれることがある。仮に多重比較法で有意な結果が出たとしても，一元配置分散分析の結果が有意でなければ，偶然のばらつきの効果が群間の差よりも大きいということなので，特定群間の差に意味があると考えることは解釈のし過ぎである（少なくともそのことに配慮した解釈を加えなくてはいけない）。

10.2　一元配置分散分析

一元配置分散分析は，データのばらつき（変動）を，群間の違いという意味のはっきりしているばらつき（群間変動）と，各データが群ごとの平均からどれくらいばらついているか（誤差）をすべての群について合計したもの（誤差変動）に分解して，前者が後者よりもどれくらい大きいかを検討することによって，群分け変数がデータの変数に与える効果があるかどうかを調べるものである。

例えば，南太平洋の 3 つの村 X, Y, Z で健診をやって，成人男性の身長や体重を測ったとしよう。このとき，データは例えば次のようになる（架空のものである）。

ID 番号	村落 (vg)	身長 (cm)(height)
1	X	161.5
2	X	167.0
（中略）		
22	Z	166.0
（中略）		
37	Y	155.5

身長と体重の関係を図示すると図 10.1 のようになる。

[*1] ただし，永田，吉田 (1997) が指摘するように，段階を踏んで実行すると，ここにまた検定の多重性の問題が生じるので，両方はやるべきではない，という考え方にも一理ある。つまり，厳密に考えれば，群分け変数が量的変数に与える効果があるかどうかを調べたいのか，群間で量的変数に差があるかどうかを調べたいのかによって，これら 2 つのアプローチを使い分けるべきだというのである。この点に関しては，多くの学術雑誌が現在でも「段階を踏め」式の指摘をしてくるので，思想の違いと考えるしかないし，どこかの群間にはっきりした違いがあれば，どちらの考え方をしても結果に違いは出てこないはずだから，当面は「段階を踏む」式の考え方をしておく方が無難であろう。

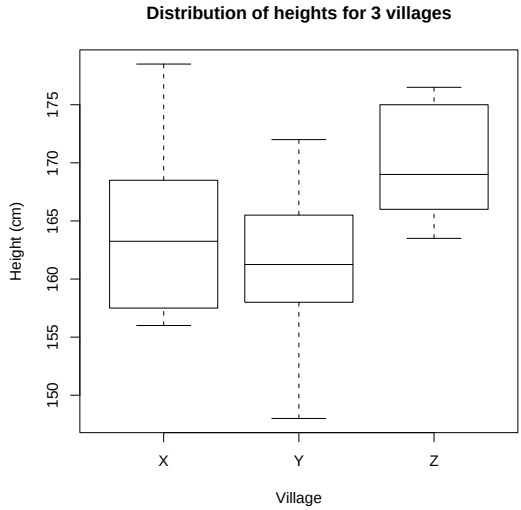


図 10.1: 南太平洋 3 村落の男性における身長の分布

村落によって身長に差があるかどうかを検定したいならば，height という量的変数に対して，vg という群分け変数の効果があるかどうかを一元配置分散分析することになる。R でデータを読み込んでから，summary(aov(height ~ vg)) とすれば（実は anova(lm(height ~ vg)) でも同等），例えば次のような結果が得られる。

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
vg           2  422.72  211.36  5.7777 0.006918 **
Residuals   34 1243.80   36.58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

このような結果の表を分散分析表という。右端の*の数は有意性を示す目安だが，確率そのものに注目してみるほうがよい。Sum Sq のカラムは偏差平方和を意味する。vg の Sum Sq の値 422.72 は，村ごとの平均値から総平均を引いて二乗した値を村ごとの人数で重み付けした和であり，群間変動または級間変動と呼ばれ，vg 間でのばらつきの程度を意味する。Residuals の Sum Sq の値 1243.80 は各個人の身長

からその個人が属する村の平均身長を引いて二乗したものの総和であり，誤差変動と呼ばれ，村によらない（それ以外の要因がないとすれば偶然の）ばらつきの程度を意味する。`Mean Sq` は平均平方和と呼ばれ，偏差平方和を自由度 (`Df`) で割ったものである。平均平方和は分散なので，`vg` の `Mean Sq` の値 211.36 は群間分散または級間分散と呼ばれることがあり，`Residuals` の `Mean Sq` の値 36.58 は誤差分散と呼ばれることがある。`F value` は分散比と呼ばれ，群間分散の誤差分散に対する比である。この場合の分散比は第 1 自由度 2，第 2 自由度 34 の F 分布に従うことがわかっているので，それを使った検定の結果，分散比がこの実現値よりも偶然大きくなる確率（`Pr(>F)` に得られる）が得られる。この例では 0.006918 なので，`vg` の効果は 5%水準で有意であり，帰無仮説は棄却される。つまり，身長は村落によって有意に異なることになる。

きちんと数式で説明すると，次のようになる。X 村の N_1 人の身長が X_{11}, X_{12}, ..., X_{1N_1}，Y 村の N_2 人の身長が X_{21}, X_{22}, ..., X_{2N_2}，Z 村の N_3 人の身長が X_{31}, X_{32}, ..., X_{3N_3} だとする（総人口 $N_1 + N_2 + N_3 = N$ 人とする）。村毎の平均身長を $\bar{X_1}$, $\bar{X_2}$, $\bar{X_3}$ と書き，全体の平均を $\bar{X_T}$ と書くことにする。このとき，総変動（総平方和）S_T は，

$$S_T = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(X_{ij} - \bar{X_T})^2$$

級間変動（群間平方和）S_A は，

$$S_A = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(\bar{X_i} - \bar{X_T})^2$$

誤差変動（級内変動，群内平方和，または誤差平方和ともいう）S_E は，

$$S_E = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(X_{ij} - \bar{X_i})^2$$

となる[*2]。自由度は，群の効果に関して $P_A = 3 - 1 = 2$ で，残差の効果に関して $P_E = N - 3 = 34$ である。よって，級間分散 $V_A = S_A/P_A$，誤差分散 $V_E = S_E/P_E$ と推定でき，F 統計量 $F_0 = V_A/V_E$ が，第 1 自由度 P_A，第 2 自由度 P_E の F 分布に従うことを使って検定できる[*3]。つまり，繰り返しになるが，分散分析とは，全体のばらつき S_T を，群間の違いという意味のはっきりしているばらつき S_A と，それ

[*2] ちなみに簡単な式変形で，このとき $S_T = S_A + S_E$ であることがわかる。確かめられよ。

[*3] R では，p=`1-pf(`F_0`,`P_A`,`P_E`)` として有意確率が得られる。

では説明できないばらつき，つまり誤差である S_E に分けて比べることを意味するのである。

念のため上の数値例の値が数式のどれに当たるかをまとめておくと，P_A が 2，P_E が 34（N が 37），S_A が 422.72，S_E が 1243.80，V_A が 211.36，V_E が 36.58，F_0 が 5.7777，p が 0.006918 である。

なお，この例のように，群分けをするカテゴリ変数が 1 つの場合を，一元配置分散分析 (ONE-WAY ANOVA)，2 つの場合を二元配置分散分析 (TWO-WAY ANOVA)，3 つなら三元配置分散分析 (THREE-WAY ANOVA) などと呼ぶ。二元配置以上の場合は，カテゴリ変数間での交互作用による影響を調べるための交互作用項がモデルに入ってくるし，その従属変数への効果を見るために母数モデルと変量モデルの違いを区別しなくてはならない。また，量的変数による交絡がある場合は共分散分析 (ANACOVA) をすることになる。[*4]

10.3 クラスカル＝ウォリス (Kruskal-Wallis) の検定

一元配置の分散分析は，各群が等しい母分散をもつ正規分布に従うことを仮定して行っているパラメトリックな分析法なので，データの分布がひどく歪んでいる場合は結果がおかしくなる場合がある[*5]。そこで，多群間の差を調べるためにもノンパラメトリックな方法がある。クラスカル＝ウォリス (Kruskal-Wallis) の検定と呼ばれる方法である。R では，`kruskal.test(`**量的変数 ~ 群分け変数**`)` で実行できる。以下，仕組みを箇条書きで説明する。

- 「少なくともどれか 1 組の群間で大小の差がある」という対立仮説に対する「すべての群の間で大小の差がない」という帰無仮説を検定する。
- まず 2 群の比較の場合の順位和検定と同じく，すべてのデータを込みにして小さい方から順に順位をつける（同順位がある場合は平均順位を与える）。

[*4] これらの一部は第 12 章と第 13 章で説明する。

[*5] 各群の母分散が等しいかどうかを調べる検定法として，バートレット (Bartlett) の検定と呼ばれる方法がある。R では `bartlett.test(`**量的変数 ~ 群分け変数**`)` で実行できる。同じ目的のノンパラメトリックな方法として，Fligner-Killeen の検定という方法もある。R では，`fligner.test(`**量的変数 ~ 群分け変数**`)` で実行できる。また，母集団が正規分布しているかどうかを調べる方法としては，既に説明したヒストグラムや正規確率プロットなどのグラフ表示による方法の他に，シャピロ＝ウィルク (Shapiro-Wilk) の検定と呼ばれる方法もある。詳しくは説明しないが，R では `shapiro.test(`**量的変数**`)` で実行できる。厳密に言えば，これらの検定で等分散性と分布の正規性が確認されない限り，一元配置分散分析の結果を解釈するには注意が必要なのだが，論文や本でもそこまで考慮されずに使われていることが多い。

- 次に，各群ごとに順位を足し合わせて，順位和 $R_i (i = 1, 2, ..., k;\ k$ は群の数) を求める。
- 各群のオブザーベーションの数をそれぞれ n_i とし，全オブザーベーション数を N としたとき，各群について統計量 B_i を $B_i = n_i\{R_i/n_i - (N+1)/2\}^2$ として計算し，

$$B = \sum_{i=1}^{k} B_i$$

として B を求め，$H = 12 \cdot B/\{N(N+1)\}$ として H を求める。同順位を含むときは，すべての同順位の値について，その個数に個数の 2 乗から 1 を引いた値を掛けたものを計算し，その総和を A として，

$$H' = \frac{H}{1 - \frac{A}{N(N^2-1)}}$$

により H を補正した値 H' を求める。

- H または H' から表を使って（データ数が少なければ並べかえ検定によって）有意確率を求めるのが普通だが，$k \geq 4$ で各群のオブザーベーション数が最低でも 4 以上か，または $k = 3$ で各群のオブザーベーション数が最低でも 5 以上なら，H や H' が自由度 $k - 1$ のカイ二乗分布に従うものとして検定できる。

なお，対応のある多群間の差をノンパラメトリックな方法で調べるには，フリードマン (Friedman) の検定と呼ばれる手法を用いる。R では，`friedman.test(`**量的変数 ~ 群分け変数**`)` で実行できる。簡単に説明すると，まず同じ個体について群間で順位をつける（群といっても，対応がある場合だから，例えば 2005 年の予測値と 2010 年の予測値と 2025 年の予測値というように，個々の個体について順位をつけることが可能である)。次に，群ごとにこの順位の合計（順位和）を計算する。順位和の二乗和から順位和の平均の二乗を引いた値を統計量 S として，サンプル数が少ない場合は表によって（コンピュータシミュレーションによってもよい）有意確率を計算し，サンプル数が多い場合は自由度が群数より 1 少ないカイ二乗分布に従う統計量 Q を S の 12 倍を個体数と群数と「群数＋1」の積で割った値として計算して有意確率を計算する。ただし同順位がある場合は調整が必要であり，煩雑なので，通常はコンピュータソフトウェアに計算させる。

10.4 多重比較

仮に，上述の南太平洋の島の３つの村での健診の例で，一元配置分散分析か Kruskal-Wallis の検定で有意差があったときに，具体的にどの村の間に有意差があるのかを調べるには，単純に考えると，t 検定[*6]や順位和検定[*7]を繰り返せば良さそうである。この方法が使われている本や論文もないわけではない。しかし，３つの村でこれをやると３つから２つを取り出す全ての組み合わせについて検定するので，３回の比較をすることになり，個々の検定について有意水準を 5%にすると，全体としての第１種の過誤は明らかに 5%より大きくなる。もし村が７つあったら，７つから２つを取り出す組み合わせは 21 通りあるので，１つくらいは偶然によって有意差が出てしまう比較があっても全然おかしくない。したがって，先に述べた通り，t 検定の繰り返しは第１種の過誤が大きくなってしまって不都合である。これに似た方法として無制約 LSD（最小有意差）法や Fisher の制約つき LSD 法（一元配置分散分析を行って有意だった場合にのみ LSD 法を行うという方法）があるが，これらも第１種の過誤を適切に調整できない（ただし制約つきの場合は３群なら大丈夫）ことがわかっているので，使ってはいけない。現在では，この問題は広く知られているので，t 検定の繰り返しや LSD 法で分析しても論文は accept されない。

多重比較の方法にはいろいろあるが，良く使われているものとして，ボンフェローニ (Bonferroni) の方法，シェフェ (Scheffé) の方法，ダンカン (Duncan) の方法，テューキー (Tukey) の HSD，ダネット (Dunnett) の方法，ウィリアムズ (Williams) の方法がある。しかしこの中で，ダンカンの方法は，新多範囲検定などと呼ばれた時期もあったが，数学的に間違っていることがわかっているので，使ってはいけない。ボンフェローニの方法とシェフェの方法も検出力が悪いので，特別な場合を除いては使わない方が良い。せめてテューキーの HSD を使うべきである。ダネットの方法は対照群が存在する場合に対照群と他の群との比較に使われるので，適用場面が限定されている[*8]。ウィリアムズの方法は対照群があって他の群にも一定の傾向が仮定される場合には最高の検出力を発揮するが，ダネットの方法よりもさらに限られた場合にしか使えない。

[*6] 第８章を参照。R では `t.test(height[vg=="X"],height[vg=="Y"])` など。

[*7] 第９章を参照。R では `wilcox.test(height[vg=="X"],height[vg=="Y"])` など。

[*8] ただし，対照群が他の群との比較のすべての場合において差があるといいたい場合は，多重比較をするのではなくて，t 検定を繰り返して使うのが正しいので，注意が必要である。もちろんそういう場合は多くはないが。

上記いくつかの方法が良く使われている原因は，用途が限定されているダネットとウィリアムズを除けば，たんにそれらが歴史的に古く考案され，昔の統計学の教科書にも説明されているからに過ぎない。現在では，かなり広い用途をもち，ノンパラメトリックな分析にも適応可能なホルム (Holm) の方法（ボンフェローニの方法を改良して開発された方法）が第一に考慮されるべきである。その上で，全ての群間の比較をしたい場合はペリ (Peritz) の方法，対照群との比較をしたいならダネットの逐次棄却型検定（これはステップダウン法と呼ばれる方法の 1 つであり，既に触れたダネットの方法とは別）も考慮すればよい。とはいえ，ソフトウェアによってはこれらの方法をサポートしていない場合もあると思われ，その場合はテューキーの HSD を使うべきである（もちろん場合によっては，ダネットかウィリアムズを使い分けねばならない)[*9]。

多重比較においては，帰無仮説が単純ではない。例えば，4 群間の差を調べるとしよう。一元配置分散分析での帰無仮説は，$\mu_1 = \mu_2 = \mu_3 = \mu_4$ である。これを包括的帰無仮説と呼び，$H_{\{1,2,3,4\}}$ と書くことにする。さて第 1 群から第 4 群までの母平均 $\mu_1 \sim \mu_4$ の間で等号関係が成り立つ場合をすべて書き上げてみると，$H_{\{1,2,3,4\}} : \mu_1 = \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2,3\}} : \mu_1 = \mu_2 = \mu_3$，$H_{\{1,2,4\}} : \mu_1 = \mu_2 = \mu_4$，$H_{\{1,3,4\}} : \mu_1 = \mu_3 = \mu_4$，$H_{\{2,3,4\}} : \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2\},\{3,4\}} : \mu_1 = \mu_2$ かつ $\mu_3 = \mu_4$，$H_{\{1,3\},\{2,4\}} : \mu_1 = \mu_3$ かつ $\mu_2 = \mu_4$，$H_{\{1,4\},\{2,3\}} : \mu_1 = \mu_4$ かつ $\mu_2 = \mu_3$，$H_{\{1,2\}} : \mu_1 = \mu_2$，$H_{\{1,3\}} : \mu_1 = \mu_3$，$H_{\{1,4\}} : \mu_1 = \mu_4$，$H_{\{2,3\}} : \mu_2 = \mu_3$，$H_{\{2,4\}} : \mu_2 = \mu_4$，$H_{\{3,4\}} : \mu_3 = \mu_4$ の 14 通りである。このうち，$H_{\{1,2,3,4\}}$ 以外のものを部分帰無仮説と呼ぶ。すべての 2 つの群の組み合わせについて差を調べるということは，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}, H_{\{2,3\}}, H_{\{2,4\}}, H_{\{3,4\}}\}$ が，考慮すべき部分帰無仮説の集合となる。一方，例えば第 1 群が対照群であって，他の群のそれぞれが第 1 群と差があるかどうかを調べたい場合は，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}$ が考慮すべき帰無仮説の集合となる。これらの集合をその多重比較における「帰無仮説族」と呼ぶ。

ここで多重比較の目的を「帰無仮説族」というコトバを使って言い換えてみる。個々の帰無仮説で有意水準を 5%にしてしまうと，帰無仮説族に含まれる帰無仮説のどれか 1 つが誤って棄却されてしまう確率が 5%より大きくなってしまう。それではまずいので，その確率が 5%以下になるようにするために，何らかの調整を必要とするわけで，この調整をする方法が多重比較なのである。つまり，帰無仮説族の有意水準を定める（例えば 5%にする）ことが，多重比較の目的である[*10]。

[*9] もっとも，オープンソースで多くのコンピュータで無料で使える R がホルムの方法をデフォルトとしている現実を考えれば，そういう言い訳はもはや通用しないと思う。

[*10] このことからわかるように，差のなさそうな群をわざと入れておいて帰無仮説族を棄却されにくく

R では，`pairwise.t.test(height,vg,p.adjust.method="bonferroni")` とすれば，ボンフェローニの方法で有意水準を調整した，すべての村落間での身長の差を t 検定した結果を出してくれる[*11]。

また，`pairwise.wilcox.test(height,vg,p.adjust.method="bonferroni")` とすれば，ボンフェローニの方法で有意水準を調整した，すべての村落間での身長の差を順位和検定した結果を出してくれる。これらの関数で，`p.adjust.method` を指定しなければホルムの方法になるが，明示したければ，`p.adjust.method="holm"` とすればよい。R でもボンフェローニが可能なのは，一番単純な方法であるという理由と，ホルムの方法に必要な計算がボンフェローニの計算を含むからだと思われる。なお，R を使って分析するのだけれども，データがきれいな正規分布をしていて，かつ古典的な方法の論文しか accept しない雑誌に対してどうしても投稿したい，という場合は，`TukeyHSD(aov(height ~ vg))` などとして，テューキーの HSD を行うことも可能である。また，CRAN (http://cran.r-project.org/) から `multcomp` パッケージをインストールすることによって，`simtest(height ~ vg, type="Dunnett")` あるいは `simtest(height ~ vg, type="Williams")` としてダネットやウィリアムズの方法を使うことも可能である。

これらの方法の中身に立ち入って説明しつくすことは不可能なので，ここではボンフェローニとホルム，テューキーの HSD だけを簡単に説明する。より詳しく知りたい場合には，永田，吉田 (1997) を参照されたい。

10.4.1 ボンフェローニの方法とホルムの方法

ボンフェローニの方法とは，ボンフェローニの不等式に基づく多重比較法である。きわめて単純な考え方に基づいているために，適用可能な範囲が広い。しかし，検出力が落ちてしまいがちなので，ベストな方法ではない。

ボンフェローニの不等式とは，k 個の事象 E_i $(i = 1, 2, ..., k)$ に対して成り立つ，

$$Pr(\cup_{i=1}^{k} E_i) \leq \sum_{i=1}^{k} Pr(E_i)$$

をいう。左辺は k 個の事象 E_i のうち少なくとも 1 つが成り立つ確率を示し，右辺は

したり，事後的に帰無仮説を追加したりすることは，統計を悪用していることになり，やってはいけない。

[*11] ただし，t 検定とは言っても，`pool.sd=F` というオプションをつけない限りは，t_0 を計算するときに全体の誤差分散を使うので，ただの t 検定の繰り返しとは違う。

各事象 E_i が成り立つ確率を加え合わせたものなので，この式が成り立つことは自明であろう（個々の事象がすべて独立な場合にのみ等号が成立する）。

次に，この不等式を多重比較にどうやって応用するかを示す。まず，帰無仮説族を $\{H_{01}, H_{02}, ..., H_{0k}\}$ とする。E_i を「正しい帰無仮説 H_{0i} が誤って棄却される事象」と考える。この表現をボンフェローニの不等式にあてはめれば，

$$Pr(\text{正しい帰無仮説のうちの少なくとも 1 つの } H_{0i} \text{ が誤って棄却される})$$

$$\leq \sum_{i=1}^{k} Pr(\text{正しい帰無仮説 } H_{0i} \text{ が誤って棄却される})$$

右辺が α 以下になるためには，もっとも単純に考えれば，足しあわされる各項が α/k に等しいかより小さければよい。つまり，ボンフェローニの方法とは，有意水準 α で帰無仮説族を検定するために，個々の帰無仮説の有意水準を α/k にするものである[*12]。手順としてまとめると，以下の通りである。

1. 帰無仮説族を明示し，そこに含まれる帰無仮説の個数 k を求める。
2. 帰無仮説族についての有意水準 α を定める。$\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。
3. 帰無仮説族に含まれているそれぞれの帰無仮説に対して検定統計量 T_i $(i = 1, 2, ..., k)$ を選定する。
4. データを取り，検定統計量 T_i を計算する。
5. 各検定統計量 T_i について有意水準 α/k に対応する棄却限界値（通常は分布関数の $(1 - \alpha/k) \times 100\%$点）を c_i とするとき，$T_i \geq c_i$ ならば H_{0i} を棄却し，$T_i < c_i$ なら H_{0i} を保留する（採択ではない）。

なお，R では，各々の帰無仮説の有意水準を α/k とする代わりに，各々の帰無仮説に対して得られる有意確率が k 倍されて（ただし 1 を超えるときは 1 として）表示されるので，各々の比較に対して表示される有意確率と帰無仮説族について設定したい有意水準との大小によって仮説の棄却／保留を判断してよい。

ボンフェローニの方法では，すべての H_{0i} について有意水準を α/k としたのが良くなかったので，ホルムの方法は，そこを改良したものである。以下，ホルムの方法の手順をまとめる。

[*12] ここで注意しなければいけないことは，検定すべき帰無仮説族に含まれる個々の帰無仮説は，データをとるまえに定められていなければいけないことである。データをとった後で有意になりそうな帰無仮説を k 個とってきて帰無仮説族を構成するのでは，帰無仮説族に対しての第 1 種の過誤をコントロールできないのでダメである。

1. 帰無仮説族を明示し，そこに含まれる帰無仮説の個数 k を求める。
2. 帰無仮説族についての有意水準 α を定める。$\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。ここまではボンフェローニと同じ。
3. $\alpha_1 = \alpha/k$，$\alpha_2 = \alpha/(k-1)$，...，$\alpha_k = \alpha$ を計算する。
4. 帰無仮説族に含まれているそれぞれの帰無仮説に対して検定統計量 T_i $(i = 1, 2, ..., k)$ を選定する。
5. データを取り，検定統計量 T_i を計算する。
6. 各検定統計量 T_i について有意確率 P_i を求め，小さい順に並べ換える。
7. P_i の小さいほうから順に α_i と P_i の大小を比べる。
8. $P_i > \alpha_i$ ならばそれよりも有意確率が大きい場合の帰無仮説をすべて保留して終了する。$P_i \leq \alpha_i$ なら H_{0i} を棄却して，次に小さい P_i について比較する。$i = k$ となるまで繰り返す。

ホルムの方法についても，R では，7) で P_i と α_i の大小を比べる代わりに $P_i' = P_i \times (k - i + 1)$ が表示されるので，値そのものを有意水準と比較すればよい。ただし，上記手順 8 からすると，P_i' が有意でなかったら，P_{i+1}' が有意水準より小さくてもその仮説は保留されるべきなのだが，その点がどう表示されるのかは未確認である。

計算例を示すと，南太平洋の 3 つの村の問題で，ボンフェローニの方法とホルムの方法で検定した有意確率は，次のようになる。

検定と調整の方法	X-Y	X-Z	Y-Z
誤差分散を使った t 検定，Bonferroni で調整	0.8841	0.1283	0.0052
t 検定の繰り返し，Bonferroni で調整	1.0000	0.1422	0.0026
誤差分散を使った t 検定，Holm で調整	0.2947	0.0855	0.0052
t 検定の繰り返し，Holm で調整	0.3475	0.0948	0.0026
順位和検定の繰り返し，Bonferroni で調整	1.0000	0.2162	0.0078
順位和検定の繰り返し，Holm で調整	0.4865	0.1441	0.0078

10.4.2 テューキーの HSD

テューキーの HSD では，母集団の分布は正規分布とし，すべての群を通して母分散は等しいと仮定する。

データが第 1 群から第 a 群まであって，各々が n_i 個（$i = 1, 2, ..., a$）のデータからなるものとする。第 i 群の j 番目のデータを x_{ij} と書くことにすると，第 i 群の平

均 $\bar{x_i}$ と分散 V_i は,

$$\bar{x_i} = \sum_{j=1}^{n_i} x_{ij}/n_i$$

$$V_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2/(n_i - 1)$$

となり，誤差自由度 P_E と誤差分散 V_E は,

$$P_E = N - a = n_1 + n_2 + ... + n_a - a$$

$$V_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2/P_E = \sum_{i=1}^{a} (n_i - 1)V_i/P_E$$

で得られる。

簡単にいえば，テューキーの HSD は，すべての群間の比較について誤差分散を使った t_0 統計量を計算し，t 分布ではなくて，**ステューデント化された範囲の分布** (Studentized range distribution) と呼ばれる分布の $(1-\alpha)\times 100\%$点を $\sqrt{2}$ で割った値との大小で有意水準 α の検定をする方法である。以下手順としてまとめる。

1. 帰無仮説族を明示する。テューキーの HSD の場合は，通常,

$$\{H_{\{1,2\}}, H_{\{1,3\}}, ..., H_{\{1,a\}}, H_{\{2,3\}}, ..., H_{\{a-1,a\}}\}$$

2. 有意水準 α を定める。$\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。
3. データを取り，すべての群について $\bar{x_i}, V_i$ を計算し，P_E, V_E を計算する。
4. すべての 2 群間の組み合わせについて，検定統計量 t_{ij} を

$$t_{ij} = (\bar{x_i} - \bar{x_j})/\sqrt{V_E(1/n_i + 1/n_j)}$$

   により計算する（$i, j = 1, 2, ..., a; i < j$)。
5. $|t_{ij}| \geq q(a, P_E; \alpha)/\sqrt{2}$ なら $H_{\{i,j\}}$ を棄却し，i 群と j 群の平均値には差があると判断する（比較の形からわかるように，これは両側検定である)。$|t_{ij}| < q(a, P_E; \alpha)/\sqrt{2}$ なら $H_{\{i,j\}}$ を保留する。ここで $q(a, P_E; \alpha)$ は，群数 a，自由度 P_E のステューデント化された範囲の分布の $(1-\alpha)\times 100\%$点である。つまり，$\alpha = 0.05$ ならば，$q(a, P_E, 0.05)$ は，群数 a，自由度 P_E のステューデント化された範囲の分布の 95%点である。R では，この値を与える関数は，`qtukey(0.95,`a`,`P_E`)` である。が，すべての群間比較を手計算するよりも，パッケージに計算させるのが普通である。

上述の例題に対する R の `TukeyHSD(aov(height ~ vg))` の出力は以下の通り。95%同時信頼区間が 0 を含まない Y 村と Z 村の身長だけが，5%水準で有意に異なる（Z−Y が正なので，Z 村の平均身長の方が Y 村の平均身長より有意に高い）と読める。

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = height ~ vg)

$vg
         diff        lwr       upr
Y-X -2.538889 -8.3843982  3.306620
Z-X  5.850000 -0.9598123 12.659812
Z-Y  8.388889  2.3382119 14.439566
```

第 11 章

相関と回帰

11.1 量的変数の関連を調べる

相関と回帰は混同されやすいが，思想はまったく違う。相関は，変数間の関連の強さを表すものである。回帰は，ある変数の値のばらつきが，どの程度他の変数の値のばらつきによって説明されるかを示すものである。回帰の際に，説明される変数を従属変数または目的変数，説明するための変数を独立変数または説明変数と呼ぶ。2 つの変数間の関係を予測に使うためには，回帰を用いる。

まず相関について，前章であげた南太平洋の 3 村落 X, Y, Z の成人男性の例を使って説明しよう。前章では示さなかった体重のデータも加えると，データは例えば次のようになる（架空のものである）。

ID 番号	村落 (VG)	身長 (cm)(HEIGHT)	体重 (kg)(WEIGHT)
1	X	161.5	49.2
2	X	167.0	72.8
（中略）			
22	Z	166.0	58.0
（中略）			
37	Y	155.5	53.6

身長と体重の関係を，身長を横軸にとって，体重を縦軸にとって二次元平面にプロットすると，図 11.1 のようになる[*1]。第 3 章で触れたが，このような図を散布図

[*1] R を使って，村ごとにプロットするマークを変えてプロットした。プログラムは下記の通り。

```
x <- read.delim("l11-1.dat")
attach(x)
```

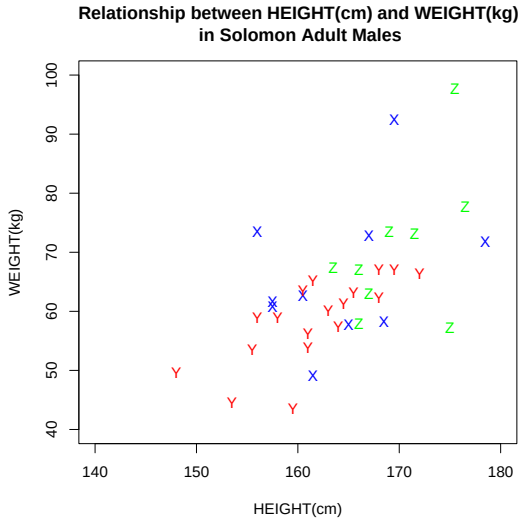


図 11.1: 南太平洋 3 村落の成人男性における身長と体重の関係

(scatter plot または scattergram）と呼ぶ。2 つの量的変数間の関係をみるときには，基本として絶対に作成しなければならない。

関係とか関連とかいっても，その中身は多様である。例えば，$pV = nRT$ のような物理法則は，測定誤差を別にすれば 100%成り立つ関係である。身長と体重の間の関係はそうではないが，無関係ではないことは直感的にも理解できるし，上の図を見ても「身長の高い人は体重も概して重い傾向がある」ことは間違いない。

一般に，2 個以上の変量が「かなりの程度の規則正しさをもって，増減をともにする関係」のことを相関関係 (correlation) という。相関には正の相関 (positive correlation) と負の相関 (negative correlation) があり，一方が増えれば他方も増え

```
plot(HEIGHT[VG==’X’], WEIGHT[VG==’X’], main="Relationship between HEIGHT(cm)
and WEIGHT(kg)\n in Adult Males in South Pacific 3 villages", pch=’X’,
col="blue", xlim=c(140,180), ylim=c(40,100), xlab="HEIGHT(cm)", ylab="WEIGHT(kg)")
points(HEIGHT[VG==’Y’], WEIGHT[VG==’Y’], pch=’Y’,col="red")
points(HEIGHT[VG==’Z’], WEIGHT[VG==’Z’], pch=’Z’,col="green")
detach(x)
```

る場合を正の相関，一方が増えると他方は減る場合を負の相関と呼ぶ。例えば，上の図に示されている身長と体重の関係は正の相関である。

相関関係があることは，因果関係 (causal relationship) が成り立つための重要な要件ではあるが，それだけで因果関係があるとは結論付けるのは勇み足である。では，因果関係があるというための基準はあるのだろうか？　古来いろいろな説があった中でも有力とされていて，Rothman(2002) にも載っている Hill(1965) の基準によれば，因果関係を因果関係のない関連と区別するためには，次の９条件が満たされる必要がある。[*2]

1. 相関関係が強い。
2. 相関関係が常に成り立つ。
3. 相関関係に特異性がある。
4. 時間的前後関係がはっきりしている。
5. 生物学的なメカニズムが想定できる（これは疫学の教科書での説明だから「生物学的な」なので，社会学的であっても物理化学的であってもよい)。
6. もっともらしい。
7. 首尾一貫している（他の知見と矛盾がない)。
8. 実験的な証拠がある。
9. アナロジーがなりたつ。

相関関係があっても，それが見かけ上のものである（それらの変量がともに，別の変量と真の相関関係をもっている）場合がある。具体例としては，血圧と所得の間に正の相関があるという命題は，データをとってみれば，多くの場合に成り立つであろう。しかしこれは，おそらくどちらも年齢や摂取エネルギー量との間に真の相関関係が存在するのであって，それらの影響を制御したら（例えば同年齢で同じような食生活をしている人だけについて見る，という層別化をしたら)，血圧と所得の間の正の相関は消えてしまうだろう。この場合，見かけ上の相関があることは，たまたまそのデータで成り立っているだけであって，科学的仮説としての意味に乏しい。因果関係

[*2] もっとも，Hill 自身が，必ずしも９条件すべてが成り立たない場合もあるし，ヒトについては実験的な証拠が得られない場合が普通であるなど，この条件が決定的とすることは因果推論の現実性を失わせてしまうことも認めているので，あくまでこうした基準は目安程度に考えるべきである。なかでも，相関関係の特異性という考え方は，因果推論の適用範囲を著しく狭めてしまう。現実の因果関係の大部分は，複数の要因と複数の結果が網の目のように絡んでいるし，環境によって同じ遺伝的要因がまったく逆の結果をもたらす場合もある。だからこそ Rothman(2002) は，簡単な基準は存在しないことを強調し，因果構成要因群 (Component causes) と充分要因 (Sufficient cause) という考え方（第１章を参照）を提起したのである。

に迫ることが大事なのであって，相関関係はその入り口に過ぎないことを再び強調しておく[*3]。

時系列データや地域相関のデータでは，擬似相関 (spurious correalation) が見られる場合もある。例えば，日本の砂糖輸入量と溺死・溺水者数の年次別データをプロットしてみると，負の相関関係があるように見えるのだが，両者の間には真の関係はない。ある年に日本で植えた木の幹の太さと，同じ年にイギリスで生れた少年の身長を 15 年分，毎年 1 回測ったデータをプロットすると，おそらくは正の相関関係があるように見えるのだが，両者の間には関係がないのは明らかである（どちらも年次と真の相関があるとはいえるだろうが)。

複数の種類の異なるデータをまとめて見ることで見かけの相関が生じてしまう場合もある。上に示した南太平洋の 3 つの村の身長と体重の関係を良く見ると，相関関係は村によって随分違っていることがわかる。それをまとめたことで身長の分散と体重の分散が広がって，見かけ上強い正の相関がでたと解される。こういう場合は，村で層別して村ごとに相関を検討する必要がある[*4]。

11.2 相関関係の具体的な捉え方

上で定義したように，相関関係は増減をともにする関係であればいいので，その関係が線形（一次式で表される，散布図で直線として表される）であろうと非線形（二次式以上または階段関数などで表される）であろうと問題ない。しかし，一般には，線形の関係があるという限定的な意味で使われる場合が多い。なぜなら，相関を表すための代表的な指標である相関係数[*5]r が，線形の関係を示すための指標だからである。もっといえば，r が意味をもつためには，2 つの変量が二次元正規分布に従っていなければならない。

非線形の相関関係を捉えるには，2 つのアプローチがある。1 つは線形になるように対数変換などの変換をほどこすことで，もう 1 つはノンパラメトリックな相関係数（分布の形によらない，例えば順位の情報だけを使った相関係数）を使うことである。

[*3] 変数 X と変数 Y の間に因果関係があるとは，変数 X が変数 Y を引き起こすために必要であるような関係があるということである。Hill の条件では，因果関係があるといえるためには，X と Y の間の相関は，常に強くなくてはならない。しかし，見かけの相関でない相関があっても，直接の因果関係がない場合はありえる。つまり，変数 X が変数 Y を引き起こすのに関与することもあるが，それは必要ではない，というような場合がそれに該当する。

[*4] または，目的次第では，ダミー変数を使った重回帰分析をするべきである（第 13 章で触れる予定)。

[*5] 普通，ただ相関係数といえば，ピアソンの積率相関係数 (Pearson's Product Moment Correlation Coefficient) を指し，通常，r という記号で表す。

ノンパラメトリックな相関係数にはスピアマン (Spearman) の順位相関係数 ρ や，ケンドール (Kendall) の順位相関係数 τ がある。

ピアソンの積率相関係数とは，X と Y の共分散を X の分散と Y の分散の積の平方根で割った値である。式で書けば，相関係数の推定値 r は，X の平均を $\bar{X}$，Y の平均を $\bar{Y}$ と書けば，

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

となる。母相関係数がゼロかどうかという両側検定のためには，それがゼロであるという帰無仮説の下で，検定統計量

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

が，自由度 $n-2$ の t 分布に従うことを利用して検定すればよい。

なお R では，

```
r <- cov(X,Y)/sqrt(var(X)*var(Y))
n <- NROW(X)
t0 <- r*sqrt(n-2)/sqrt(1-r^2)
```

として，`2*(1-pt(t0,n-2))` で有意確率が得られるが，`cor.test()` 関数（下記）を使う方が簡単である[*6]。`cor.test()` 関数を使った場合は，信頼区間も計算される。なお，信頼区間は，サンプルサイズがある程度大きければ（通常は 20 以上），正規近似を使って計算できる。

$$a = \frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{\sqrt{n-3}}Z(\alpha/2), \quad b = \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{1}{\sqrt{n-3}}Z(\alpha/2)$$

と書くことにすると[*7]，母相関係数の $100 \times (1-\alpha)\%$信頼区間の下限は $(\exp(2a)-1)/(\exp(2a)+1)$，上限は $(\exp(2b)-1)/(\exp(2b)+1)$ である[*8]。

[*6] また，`cov(X,Y)/sqrt(var(X)*var(Y))` と同値な関数として `cor(X,Y)` がある。

[*7] $Z(\alpha/2)$ は，標準正規分布の $100 \times (1-\alpha/2)$ パーセント点，つまり R では（α を `alpha` と書くことにして），`qnorm(1-alpha/2,0,1)` である。例えば $\alpha = 0.05$ なら，`qnorm(0.975,0,1)` である。

[*8] なお，ln は自然対数，exp は指数関数を表す。

順位相関係数は，非線形の相関関係を捉えたい場合以外にも，分布が歪んでいたり，外れ値がある場合に使うと有効である。スピアマンの順位相関係数 ρ は[*9]，値を順位で置き換えた（同順位には平均順位を与えた）ピアソンの積率相関係数になる。X_i の順位を R_i，Y_i の順位を Q_i とかけば，

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

となる。スピアマンの順位相関係数がゼロかどうかという両側検定は，サンプル数が 10 以上ならばピアソンの場合と同様に，$T = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$ が自由度 $n-2$ の t 分布に従うことを利用して行うことができる。

ケンドールの順位相関係数 τ は，

$$\tau = \frac{(A-B)}{n(n-1)/2}$$

によって得られる。ここで A は順位の大小関係が一致する組の数，B は不一致数である。

いずれにせよ，R では `cor.test(X, Y, method="pearson")` とすればピアソンの相関係数が，`cor.test(X, Y, method="spearman")` でスピアマンの順位相関係数が，`cor.test(X, Y, method="kendall")` でケンドールの順位相関係数が得られる。同時に，`alternative` を指定しないときは，「相関係数がゼロである」を帰無仮説として両側検定した有意確率と 95%信頼区間が表示される。なお，例えば `cor.test(X, Y, alternative="g")` とすれば，ピアソンの相関係数が計算され，対立仮説を「正の相関がある」とした片側検定の結果が得られる。なお，ケンドールに関しては並べ換えによる正確な確率も求めることができ，その場合は `exact=T` というオプションを指定する。

11.3　回帰の考え方

1984 年から 1993 年までのプロ野球の 1 試合平均入場者数（単位：千人）の推移は下表のようになっている（注：この表の出典は，鈴木 (1995) である）。

年次	84	85	86	87	88	89	90	91	92	93
セリーグ	28	29	29	31	31	31	31	32	35	34
パリーグ	13	12	16	18	21	23	22	24	24	24

[*9] ピアソンの相関係数の母相関係数を ρ と書き，スピアマンの順位相関係数を r_s と書く流儀もある。

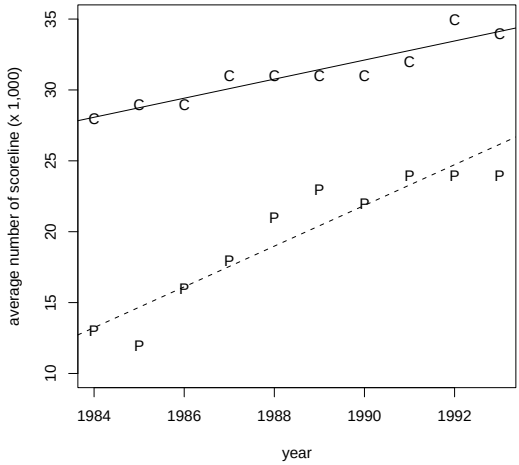


図 11.2: プロ野球 1 試合平均入場者数の推移

図 11.2 を見ると，両リーグとも途中から伸び悩んでいるが，ある程度直線的に増加しているように見える。このように，ほぼ比例関係にある量的なデータ群をうまく代表する直線を求めるには，「各データの点から直線までのずれの大きさの合計」を最小にすればよい。この場合，「年次」が予め決まっている値であるのに対して，入場者数は測定値であり，誤差を含む可能性があるので，「データ点から直線までのずれ」を評価するには，データ点と直線の最短距離よりも，データ点から直線に垂直に下ろした線分の長さの二乗和を使う方がよい。この，ずれを最小にする直線を，「年次を独立変数，入場者数を従属変数とする回帰直線」と呼ぶ。計算方法は，数学的には次に述べる検量線の求め方と同じである。R では，

```
year <- c(1984:1993)
central.league <- c(28,29,29,31,31,31,31,32,35,34)
pacific.league <- c(13,12,16,18,21,23,22,24,24,24)
plot(central.league~year,pch='C',xlim=c(1984,1993),ylim=c(10,35),
ylab="average number of scoreline (x 1,000)")
abline(lm(central.league~year),lty=1)
```

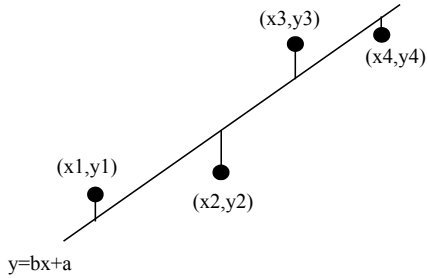


図 11.3: 標準物質の希釈系列濃度に対して測定された吸光度の関係

```
points(pacific.league~year,pch='P')
abline(lm(pacific.league~year),lty=2)
```

とすれば図 11.2 が描かれる。

11.3.1 検量線

実験によって，あるサンプルの濃度を求めるやり方の 1 つに，検量線の利用がある。検量線とは，予め濃度がわかっている標準物質を測ったときの吸光度のばらつきが，その濃度によってほぼ完全に（通常 98%以上）説明されるときに，その回帰を利用して，サンプルを測ったときの吸光度からサンプルの濃度を逆算するための回帰直線である（曲線の場合もあるが，通常は何らかの変換をほどこし，線形回帰にして利用する）。検量線の計算には，(A) 試薬ブランクでゼロ点調整をした場合の原点を通る回帰直線を用いる場合と，(B) 純水でゼロ点調整をした場合の切片のある回帰直線を用いる場合がある。いずれも，量がわかっているもの（この場合は濃度）を x，誤差を含んでいる可能性がある測定値（この場合は吸光度）を y として $y = bx + a$ という形の回帰式の係数 a と b を最小二乗法で推定し，サンプルを測定した値 y から $x = (y - a)/b$ によってサンプルの濃度 x を求める。回帰直線の適合度の目安としては，学生実習でも相関係数の 2 乗が 0.98 以上あることが望ましい。また，後で述べるように，データ点の最小，最大より外で直線関係が成立する保証はない。従って，

サンプル測定値が標準物質の測定値の最小より低いか，最大より高いときは，限界を超えていることになって値は使えない[*10]。

図 11.3 のような測定点 (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) が得られたときに，検量線 $y = bx + a$ を推定するには，図に示した線分の二乗和が最小になるように a と b を設定すればよい，というのが最小二乗法の考え方である。つまり，

$$f(a,b) = \sum_{i=1}^{n} \{y_i - (bx_i + a)\}^2$$

$$= b^2 \sum_{i=1}^{n} x_i^2 - 2b \sum_{i=1}^{n} x_i y_i + 2ab \sum_{i=1}^{n} x_i - 2a \sum_{i=1}^{n} y_i + na^2 + \sum_{i=1}^{n} y_i^2$$

が最小になるような a と b を推定すればよい。通常，a と b で偏微分した値がそれぞれ 0 となることを利用して計算すると簡単である。つまり，

$$\frac{\partial f(a,b)}{\partial a} = 2na + 2(b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i) = 0$$

$$i.e. \quad na = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i$$

$$i.e. \quad a = (y \text{ の平均}) - (x \text{ の平均}) * b$$

$$\frac{\partial f(a,b)}{\partial b} = 2b \sum_{i=1}^{n} x_i^2 + 2(a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i) = 0$$

$$i.e. \quad b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i$$

を連立方程式として a と b について解けばよい。これを解くと，

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

が得られる[*11]。b の値を上の式に代入すれば a も得られる。検量線に限らず，一般の回帰直線でも，計算方法は原則として同じである。名称の説明をしておくと，一般に，$y = bx + a$ という回帰直線について，b を回帰係数 (regression coefficient)，a を切片 (intercept) と呼ぶ。

[*10] 余談だが，このような場合はサンプルを希釈するか濃縮して測定するのが普通である。

[*11] 分母分子を n^2 で割れば，b は $x_i y_i$ の平均から x_i の平均と y_i の平均の積を引いて，x_i の二乗の平均から x_i の平均の二乗を引いた値で割った形になる。

データから得た回帰直線は，$pV = nRT$ のような物理法則と違って，完璧にデータに乗ることはない。そこで，回帰直線の当てはまりのよさを評価する必要が出てくる。a と b が決まったとして，$z_i = a + bx_i$ とおいたとき，$e_i = y_i - z_i$ を残差 (residual) と呼ぶ。残差は，y_i のばらつきのうち，回帰直線では説明できなかった残りに該当する。つまり，残差が大きいほど，回帰直線の当てはまりは悪いと考えられる。残差にはプラスもマイナスもあるので，例によって二乗和をとって，

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - z_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n - \frac{(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}/n$$

として, 値は使えない。回帰直線の当てはまりの悪さを示す尺度となる。この Q を「残差平方和」と呼び，それを n で割った Q/n を残差分散という。この残差分散 $var(e)$ と Y の分散 $var(Y)$ とピアソンの相関係数 r の間には，$var(e) = var(Y)(1 - r^2)$ という関係が常に成り立つので，$r^2 = 1 - var(e)/var(Y)$ となる。このことから r^2 が 1 に近いほど回帰直線の当てはまりがよいことになる。その意味で，r^2 を「決定係数」と呼ぶ。また，決定係数は，Y のばらつきがどの程度 X のばらつきによって説明されるかを意味するので，X の「寄与率」と呼ぶこともある。

回帰直線は最小二乗法でもっとも残差平方和が小さくなるように選ぶわけだが，データの配置によっては，何通りもの回帰直線の残差平方和が大差ないという状況がありうる。例えば，独立変数と従属変数（として選んだ変数）が実はまったく無関係であった場合は，データの重心を通るどのような傾きの線を引いても残差平方和はほとんど同じになってしまう。その意味で，回帰直線のパラメータ（回帰係数 b と切片 a）の推定値の安定性を評価することが大事である。そのためには，t 値というものが使われている。いま，Y と X の関係が $Y = a_0 + b_0 X + e$ というモデルで表されるとして，誤差項 e が平均 0，分散 σ^2 の正規分布に従うものとすれば，回帰係数の推定値 a も，平均 a_0，分散 $\sigma^2/n)(1 + M^2/V)$（ただし M と V は x の平均と分散）の正規分布に従い，残差平方和 Q を誤差分散 σ^2 で割った Q/σ^2 が自由度 $(n-2)$ のカイ二乗分布に従うことから，

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

が自由度 $(n-2)$ の t 分布に従うことになる。しかしこの値は a_0 がわからないと計算できない。a_0 が 0 に近ければこの式で $a_0 = 0$ と置いた値（つまり $t_0(0)$。これを

切片に関する t 値と呼ぶ）を観測データから計算した値が $t_0(a_0)$ とほぼ一致し，自由度 $(n-2)$ の t 分布に従うはずなので，その絶対値は 95%の確率で t 分布の 97.5%点（サンプルサイズが大きければ約 2 である）よりも小さくなる。つまり，データから計算された t 値がそれより大きければ，切片は 0 でない可能性が高いことになる。t 分布の分布関数を使えば，「切片が 0 である」という帰無仮説に対する有意確率が計算できることになる。回帰係数についても同様に，

$$t_0(b) = \frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

が自由度 $(n-2)$ の t 分布に従うことを利用して，「回帰係数が 0」であるという帰無仮説に対する有意確率が計算できる。

以上の説明からすると，身長と体重のように，どちらも誤差を含んでいる可能性がある測定値である場合には，どちらかを独立変数，どちらかを従属変数とみなしてよいのかということが問題になってくる。一般には，身長によって体重が決まってくるというように方向性が仮定できれば，身長を独立変数と見なしてもよいことになっているが，回帰分析をしてしまうと，独立変数に測定誤差がある可能性が排除されてしまうことには注意しておくべきである。つまり，測定誤差が大きい可能性がある変数を独立変数とした回帰分析は，できれば避けたほうが良い（が，そうもいかないのが実情である）。また，最小二乗推定の説明から自明なように，独立変数と従属変数を入れ替えた回帰直線は一致しないので，どちらを従属変数とみなし，どちらを独立変数とみなすか，ということは，因果関係の方向性に基づいてきちんと決めるべきである。

ところで，回帰 (regression) とは本章第 1 節で説明した通り，被説明変数（従属変数）のばらつきが，説明変数（独立変数）のばらつきで説明されるという考え方だが，もともとは，生物統計学者 Francis Galton が，父親と息子の身長をペアとして測定し，背の高い父親の息子の平均身長が父親ほど高くなく，背の低い父親の息子の平均身長が父親ほど低くないこと，つまり第二世代の身長が平均の方向に「回帰」するという意味で用いた言葉である。この現象は，父親群でも息子群でも身長の平均と分散が等しいと仮定し，父親の身長と息子の身長の分布が二次元正規分布に従うとすると以下のようにクリアに説明できる。

父親の身長が x の息子の身長 Y の期待値 $\mu_{Y\cdot x}$ は，父親の身長と息子の身長の母相関係数を R と書くことにすると，$\mu_{Y\cdot x} = \mu_y + R\frac{\sigma_y}{\sigma_x}(x-\mu_x)$ となるので，これを式変形すれば，$\mu_{Y\cdot x} - x = -(1-R)(x-\mu_x)$ となるので，$x > \mu_x$ ならば $\mu_{Y\cdot x} < x$ となり，$x < \mu_x$ ならば $\mu_{Y\cdot x} > x$ となる。この式は，Galton が観察した現象と符合する。

回帰を使って予測をするとき，外挿には注意が必要である。前述の通り，検量線は，原則として外挿してはいけない。実際に測った濃度より濃かったり薄かったりするサンプルに対して，同じ関係が成り立つという保証はどこにもないからである（吸光度を y とする場合は，濃度が高くなると分子の重なりが増えるので飽和 (saturate) してしまい，吸光度の相対的な上がり方が小さくなっていき，直線から外れていく）。しかし，外挿による予測は，実際にはかなり行われている。例えば世界人口の将来予測とか，河川工学における基本高水計算式とか，感染症の発症数の将来予測は，回帰の外挿による場合が多い。このやり方が妥当性をもつためには，その回帰関係が（1）かなり説明力が大きく，（2）因果関係がある程度認められ，（3）それぞれの変数の分布が端の切れた分布でない（truncated distribution でない）という条件を満たす必要がある。そうでない場合は，その予測結果が正しい保証はどこにもない[*12]。

R では，今回説明したような線形回帰を行うための関数は `lm()` である。例えば，`lm(Y~X)` のように用いれば，回帰直線の推定値が得られる。決定係数や回帰係数と切片の検定結果は，`summary(lm(Y~X))` とすれば出力される（他の統計ソフトでも簡単に得られるはずである）。なお，普通の量的変数の間の線型回帰を一般化すれば，t 検定，分散分析，共分散分析，回帰分析，判別分析，正準相関分析などをすべて共通の数学モデルで扱うことができる。このモデルを一般化線型モデルと呼ぶ。英語では Generalized Linear Model といい，R での関数名も `glm()` である。一般化線形モデルは，基本的には，$Y = \beta_0 + \beta X + \varepsilon$ という形で表される（Y が従属変数群，X が独立変数群，β_0 が切片群，β が係数群，ε が誤差項である）。詳しくは第 13 章で説明する。

[*12] それでも簡便さのために回帰の外挿による予測はかなり行われてしまっているのが現状だが，本来そういう場合は，単純な回帰でなく確率的な因果モデルを立て，シミュレーションを行うべきである

第 12 章

時系列データと間隔データの扱い方

12.1　時間を扱うとはどういうことか？

時間の入ったデータとしては，大きく分けて２種類を考えるべきである。１つは，データ間が独立でない場合である。これまで説明してきた，時間が入っていないデータは，個々のオブザーベーションが独立である。例えば，身長と体重の関係を取り上げるときは，Ａさんの身長とＢさんの身長には関係がないし，Ａさんの体重とＢさんの体重には関係がない。２次元平面にプロットされる点と点が互いに独立だからこそ，２次元正規分布に従うという仮定もできるわけである。しかし，例えば，ある人が生まれてから，毎年誕生日に身長を測って 18 歳くらいまで記録したとしたとき，年齢と身長の関係をみると，点と点の間は独立ではない。８歳のときの身長は，７歳のときの身長がどこまで伸びていたかということに，ある程度依存する。

横軸に西暦年をとり，縦軸に世界人口をとってプロットした場合も同様で，1950 年の人口は，1949 年にどこまで人口が増えていたかということと無関係ではありえない。この種のデータを時系列データと呼び，時系列データを扱う解析法を総称して時系列解析という。時系列データにおける点と点の関係は微分方程式や差分方程式で表されるが，微分方程式や差分方程式を解いて非線形回帰をするよりも，自己相関をみたり，複数の波の重ね合わせとしてパタンを解析することが多い。

第２のパタンは，期間をデータとして扱う場合に生じる。何かのイベントが発生するまでの時間をデータとして使う場合を考えよう。例えば，結婚から第１子受胎までの時間とか，チェルノブイリの事故で流出した放射性物質に曝露した子どもたちが白

血病を発症するまでの時間とかいったデータである。この種のデータを間隔データと総称する。時間の情報を間隔データとしてうまく使えると，少ないサンプル数でも，ある瞬間にイベントが発生する確率（ハザード）を効果的に推定することができる。出生力を推定するときに，閉経後の女性にインタビューをすることが行われるが，たんに一生のうちに子どもを何人産んだかを聞くよりも，出産暦としてすべての出産間隔を聞くほうが情報量が多いのは自明であろう。間隔データでは，観察期間中にそのイベントが起こらなかったケースは，打ち切りとなる（多くは右側打ち切り）。打ち切りデータは，イベントが起こるまでの時間がそれよりも長いケースなので，解析から取り除くと全体の推定値が過小評価されてしまう。それゆえ，「少なくとも打ち切りまでの期間より長い」という情報をうまく生かす分析法が要求される。生存時間解析とかハザード解析と呼ばれる分野で，この種の研究は多く行われてきた。

本章では，この 2 つ，即ち，時系列データと間隔データの扱い方の基礎を説明する。

12.2　時系列解析の基礎

まず，単純な例として，さきほども取り上げた，世界人口の推移を見てみよう。コーエン (1998) に載っている Kremer (1993) の推定値[*1]を使うと，世界人口の推移は図 12.1 のようになっている[*2]。普通の軸（左上）でみると近年の急増が激しすぎて

[*1] 米国センサス局のウェブサイト (http://www.census.gov/) からダウンロードすることもできる。

[*2] この図を描くための R のプログラムは以下の通り。

```
# world population (x 1 million) estimated by Kremer 1993
YEAR <- c(-1000000, -300000, -25000, -10000, -5000, -4000,
          -3000, -2000, -1000, -500, -200, 1, 200, 400, 600,
          800, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1650,
          1700, 1750, 1800, 1850, 1900, 1920, 1930, 1940, 1950,
          1960, 1970, 1980, 1990, 2000)
POP <- c(0.125, 1, 3.34, 4, 5, 7, 14, 27, 50, 100, 150, 170, 190,
         190, 200, 220, 265, 320, 360, 360, 350, 425, 545, 545, 610,
         720, 900, 1200, 1625, 1813, 1987, 2213, 2516, 3019, 3693,
         4450, 5333, 6000)
#
POP <- POP*1000000
BP <- 2001 - YEAR
#
op <- par(mfrow=c(2,2))
#
plot(-BP,POP,type="b",xlab="- \"Years before present (=X)\"",
ylab="World population (=Y)",axes=F)
```

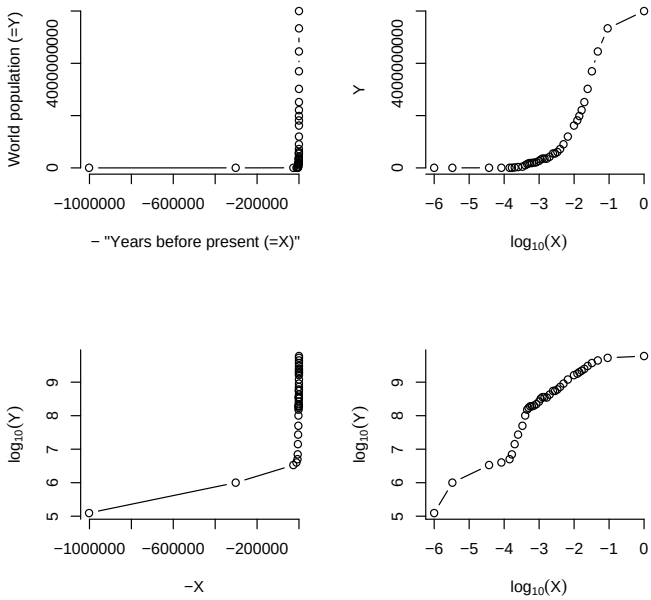


図 12.1: 世界人口の経時変化

```
axis(1,at = my.at <- c(-1000000,-800000,-600000,-400000,-200000,0),
labels=formatC(my.at,format="fg"))
axis(2,at = my.at <- c(0,2*10^9,4*10^9,6*10^9), labels=formatC(my.at,format="fg"))
#
plot(-log10(BP),POP,type="b",xlab=expression(log[10](X)),ylab="Y",axes=F)
axis(1,at = my.at <- c(-7,-6,-5,-4,-3,-2,-1,0), labels=formatC(my.at,format="fg"))
axis(2,at = my.at <- c(0,2*10^9,4*10^9,6*10^9), labels=formatC(my.at,format="fg"))
#
plot(-BP,log10(POP),type="b",xlab="-X",ylab=expression(log[10](Y)),axes=F)
axis(1,at = my.at <- c(-1000000,-800000,-600000,-400000,-200000,0),
labels=formatC(my.at,format="fg"))
axis(2,at = my.at <- c(2:10), labels=formatC(my.at,format="fg"))
#
plot(-log10(BP),log10(POP),type="b",xlab=expression(log[10](X)),
```

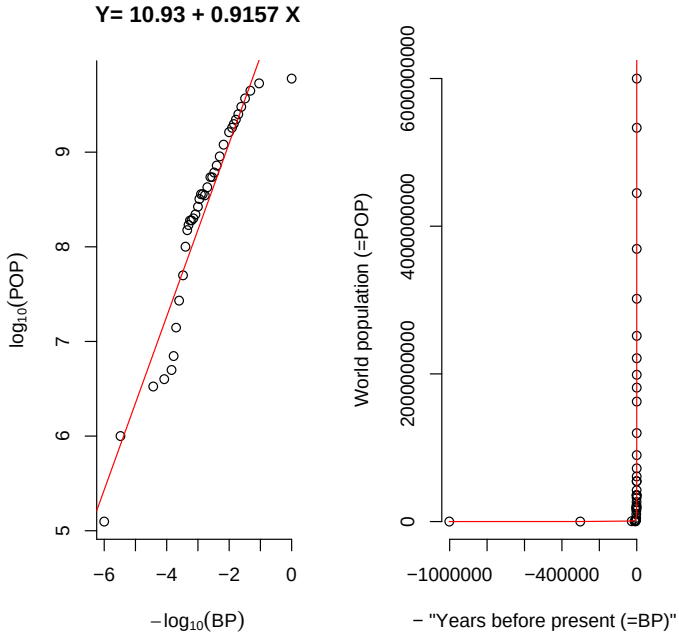


図 12.2: 世界人口の経時変化への数式の当てはめ

初期の変化がわからないが，両対数でプロットすると（右下）3 つの階段状（Deevey の階段，と呼ばれる）に見える[*3]。

これらのデータから世界人口の将来予測をするために，変化のパタンを数学的に表

```
ylab=expression(log[10](Y)),axes=F)
axis(1,at = my.at <- c(-7,-6,-5,-4,-3,-2,-1,0), labels=formatC(my.at,format="fg"))
axis(2,at = my.at <- c(2:10), labels=formatC(my.at,format="fg"))
#
par(op)
```

[*3] 10000 年ほど前までは，ヒトも狩猟採集に頼って生活しており，自然の生態系の一員だったと考えられるが，農耕が始まると人口支持力（環境収容力）が上がり（農耕革命），200 年ほど前の産業革命で非生物的資源を大規模に使うようになってさらに環境収容力が上がったという，不連続な人口増加率に対する説明が，一応与えられている。

そうという試みがいくつもなされてきた。最も単純なアプローチが，見かけの変化に数式を当てはめることである。前章でも説明したように，直線的でない関連に対して回帰を行うには，変換して見かけ上の関連を直線に近づける方法と，非線形の数式を当てはめて最小二乗法でもっとも良く当てはまるパラメータを得る方法（非線形回帰）がある。それらの変化のパタンを使って予測することは，既に説明した回帰の外挿に当たるから，正しい保証はないし，実際，どれも十分な説明にならないことが既知である[*4]。

見かけの当てはめをしたものを図 12.2 に示す[*5]。まず説明変数であるそのときから現在までの経過年数と，目的変数である世界人口を，両方とも対数変換する。上で示した Deevey の階段ができあがるが，これは比較的直線に近いので，とりあえず直線回帰をしてみる（左図）。対数変換していたのを元に戻したのが右の図である。かなり良く当てはまっているように見えないこともないが，左の図を見ると明らかに最

[*4] ただし，もし微分方程式がメカニズム（因果関係）を正しく説明しているならば，外挿してもいいことになる（もちろん，メカニズムが変わらなければ，という限定条件の下での話である）。だが，農耕革命や産業革命に匹敵する変化が起こったら，当然メカニズムが変わるだろうから，外挿による予測が現実に合わなくなってくるのは当然である。20 世紀後半からの少子化が，そうしたメカニズムの変化なのかどうかは，まだ誰にもわからない。

[*5] この図を描くための R のプログラムは次の通り。ただしデータ定義部分は前の脚注に示したものと同じなので省略する。

```
POP <- POP*1000000
BP <- 2001 - YEAR
LPOP <- log10(POP)
LBP <- -log10(BP)
reg <- lm(LPOP~LBP)
op<-par(mfrow=c(1,2))
plot(LBP,LPOP,type="p",axes=F,xlab=expression(-log[10](BP)),ylab=expression(log[10](POP)),
main=paste("Y=",formatC(reg$coefficient[[1]],width=5),"+",
formatC(reg$coefficient[[2]],width=5),"X"))
axis(1,at = my.at <- c(-7,-6,-5,-4,-3,-2,-1,0), labels=formatC(my.at,format="fg"))
axis(2,at = my.at <- c(2:10), labels=formatC(my.at,format="fg"))
abline(reg,col="red")
plot(-BP,POP,type="p",xlab="- \"Years before present (=BP)\"",
ylab="World population (=POP)",axes=F)
axis(1,at = my.at <- c(-1000000,-800000,-600000,-400000,-200000,0),
labels=formatC(my.at,format="fg"))
axis(2,at = my.at <- c(0,2*10^9,4*10^9,6*10^9), labels=formatC(my.at,format="fg"))
points(-BP,10^(reg$coefficient[[1]]+reg$coefficient[[2]]*LBP),col="red",type="l")
par(op)
```

近の値が過大評価になっているので，予測の信頼性はない。もっといえば，明らかに回帰の外挿だし，そもそも，点と点が独立でないのに回帰を使うのは思想的にもおかしいから，三重の間違いである。

12.2.1 非線形回帰

次に，微分方程式や差分方程式で隣り合う点の間の関係を説明するアプローチを説明する。これは，時系列データが互いに独立ではないことは正しく反映している。未知の係数を求めるには，非線形の最小二乗法を用いる。仮に代数的に解けない場合は，関数の最小化を数値的に行う方法がいくつか提案されているので，コンピュータを使って数値解を得る。関数の最小化を行う方法の中で最も単純なのは，Nelder-Mead の滑降シンプレックス法と呼ばれる方法である。もっと効率が良い方法としては，Powell の方法などがあるが，難解である。R でも optim() という関数で Nelder-Mead の他にも BFGS や SANN など計 5 つの方法が提供されているが，Powell の方法はない。

世界人口の変化についての微分方程式モデルには，時刻を t，人口を N，増加率を r，初期人口を N_0，人口収容力を K として，

- 指数的増加モデル: $dN/dt = rN$，即ち $N = N_0 e^{rt}$
- ロジスティック増加モデル: $dN/dt = rN(K - N)$，即ち

$$N = \frac{K}{1 + (K/N_0 - 1)e^{-rKt}}$$

- 最後の審判日モデル: 相対増加速度が現在人口に比例する，つまり $dN/dt = rN^2$ とするモデル。1958 年までのデータに当てはめると，2026 年には無限大に発散してしまうが，1980 年頃までは良く当てはまる（相対増加率が初めは実際より低く，後半では高すぎ）。現実の相対増加速度が減少に転じたので否定された。
- 指数的増加の和のモデル: 2 つの部分集団（先進国と途上国）に分けて，それぞれが指数的増加をすると考えたもの。先進国の割合を p として，$dN/dt = dN_1/dt + dp(N - N_1)/dt = r_1 N_1 + pr_2(N - N_1)$ とする。現在までのところ，あてはまりは悪くない。

これらのように微分方程式で考え，それを解いた方程式によって非線形回帰する方が[*6]，変数変換によって直線に近づけ，線形回帰するよりは本質的である。人口の場

[*6] R では，非線形回帰は nls というライブラリを使って実行する。

合は必ず整数なので，微分方程式というよりも本質的には差分方程式であり，その意味ではカオスが生じる可能性もあるので，そもそも予測が安定しない可能性があるが，局所的には微分で考えても悪くないと仮定されている。しかし，ここに上げたモデルはどれも実際の世界人口の変化を十分に説明しきれないことがわかっている。考えてみれば，人口を構成する中身の人々は常に出生と死亡によって入れ替わり，文化も自然環境も変わっていくので，微分方程式自体がが途中で変化するかもしれないから，当てはまらないのは当然である。時系列解析の本質的な難しさは，ここにある。システムの定常性を仮定できないのである。そうなると，シナリオを仮定したシミュレーション以外には手口はない。世界人口の予測には，シミュレーション（多くは，地域を分けて出生と死亡に要因分解して各々のトレンドを予測するコウホート要因法と呼ばれる手法である）も行われている。しかし決定的な予測に成功した研究はない。

12.2.2 自己回帰モデル

時系列データの予測に関しては，微分方程式や差分方程式とはまったく違うアプローチもあり，広く使われている。アプリオリに，そのデータの変化のパタンがいくつかの成分によって構成されると決めてしまい，その中身を探るアプローチである。時系列データには，繰り返し起こる（周期性がある）変化を含むように見えるものがある。例えば，日本のような中緯度に位置する場所では，一日の平均気温は，季節ごとに周期的に変化する。しかし，繰り返しは完全ではなく，地球温暖化が起これば長期的には上昇傾向をもつし，天気などによって毎日微妙に変化する。このような時系列データは，季節成分，傾向成分，不規則成分という 3 つの成分に分解して考えることができる。この考え方の応用としては，株価変動のような経済データから季節的な変動を除去する方法として，季節調整法と呼ばれる方法が広く用いられている。

周期的な変動を表す考え方の，もっとも基礎的なものは自己回帰 (Auto Regression の略で AR と呼ばれる）である。適当なタイムラグ s を置いて周期的に同じ成

```
dat <- data.frame(N=POP,t=YEAR)
library(nls)
x <- getInitial(N~SSlogis(t,Asym,xmid,scal),data=dat)
xx <- nls(N~SSlogis(t,Asym,xmid,scal),data=dat,x)
summary(xx)
```

とすれば，N=Asym/(1+exp((xmid−t)/scal)) としたロジスティック増加モデルの係数が得られる筈だが，世界人口データではロジスティック増加モデルの当てはまりが悪いので収束せず，解が得られない。

分によって決まる値が出現するならば，任意の時点 t における値 $x(t)$ と，時点 $t+s$ における値 $x(t+s)$ が相関をもっていることになり，$x(t+s)$ を目的変数，$x(t)$ を説明変数として回帰式を出したときに十分に良い当てはまりが得られれば，s だけ後の値が予測できることになる[*7]。このとき，過去の値を十分多く使えば，予測誤差が過去の値と関係をもたなくなると期待される。式で書けば，現在の時刻を n として，時刻 n における値を示す確率変数を $x(n)$ とするとき，$x(n)=a_1x(n-1)+a_2x(n-2)+...+a_Mx(n-M)+w(n)$ と書くと，誤差項 $w(n)$ が $x(n)$ の過去の値と独立になるということである。適切な次数 M を選ぶ方法は，期待される予測誤差の二乗が最小になるようにするのが実用的な方法の 1 つである。一般には，AIC などの情報量基準を用いて，モデルのあてはまりの悪さが最小になるようにする。$x(t)$ が 2 次定常であれば，$x(t)$ と $x(t+s)$ の共分散 $R_{xx}(s)$ は，s だけの関数となり，$R_{xx}(s)=E((x(t)-\mu)(x(t+s)-\mu))$ を自己共分散関数と呼ぶ。明らかに，$R_{xx}(0)$ は $x(t)$ の分散に等しく，$R_{xx}(-s)=R_{xx}(s)$，つまり自己共分散関数は原点について対称である。共分散の代わりに $x(t)$ と $x(t+s)$ の相関係数を考えると，$\rho_{xx}(s)=R_{xx}(s)/R_{xx}(0)$ となり，やはり s だけの関数となる。明らかに，$\rho_{xx}(0)=1$ である。隣同士の相関が小さい確率過程の場合は，s を大きくすると急速に $\rho_{xx}(s)$ は 0 に近づく。

過去の時系列データから予測をすることを定式化してみよう。確率過程 $x(t)$ の $s-1$ までの観測値に基づいて，次の時刻 s における値を予測することを考えるということである。簡単のため，$x(t)$ の期待値は t によらず常に 0 とする。何らかの手段で係数の列 $\{a(m);m=1,2,...,M\}$ を構成し，

$$\hat{x}(s)=\sum_{m=1}^{M}a(m)x(s-m)$$

によって $x(s)$ の予測値とすることが考えられる。このやり方で得られる予測を M 次線形予測と呼ぶ。このとき，予測誤差 $\varepsilon(s)=x(s)-\hat{x}(s)$ の 2 乗平均

$$\lim_{N\to\infty}\sum_{s=1}^{N}\varepsilon^2(s)/N$$

が最小になるように係数列 $\{a(m)\}$ を選んだときの M 次線形予測を M 次最良線形予測という。このとき，$\varepsilon(s)$ と $x(s-1),...,x(s-M)$ は無相関となる。M を大きく

[*7] このためには，$x(t)$ が 2 次定常であると都合がよい。すなわち，平均と分散が時間に関して不変であり，任意の t，s に関して $E(x(t)x(s))$ が $t-s$ だけの関数となっていると都合がよい。そうでない場合は，差分を取ったり傾向成分を除去したりして，2 次定常なデータに変換することもある。

していくと，それ以上予測誤差の最小二乗平均が小さくならなくなる点が存在し，そのときの係数列を最良線形予測子と呼ぶ。このときの予測誤差の系列は，互いに相関をもたない，平均 0，分散一定の確率変数の列（すなわちホワイトノイズ）になっている。

R で，非線形回帰以外の時系列解析をするには，まずデータを時系列データであると認識させる必要がある。例えば，ソロモン諸島ガダルカナル島で 1995 年 11 月 23 日から 12 月 29 日まで毎日朝 6 時と昼 2 時の 2 回ずつ気温を測定したデータがある。これを順番に `temp` という変数に代入するならば，

```
temp <- c(23.5, 28.7, 24.4, 28.5, 24.5, 30.5, 25.0, 30.9, 25.0,
26.7, 24.1, 30.3, 25.4, 28.3, 24.5, 32.9, 26.0, 29.4, 25.7,
31.2, 24.9, 29.3, 24.6, 29.9, 24.8, 32.0, 26.3, 31.8, 24.2,
31.2, 24.7, 29.6, 24.8, 30.7, 25.8, 31.7, 25.4, 30.1, 24.8,
29.1, 25.4, 30.9, 23.6, 31.2, 26.1, 30.8, 24.9, 32.2, 26.2,
31.2, 25.1, 31.9, 25.8, 32.5, 24.8, 32.4, 25.3, 31.7, 25.8,
33.6, 25.5, 31.6, 26.0, 30.5, 25.0, 33.0, 25.5, 30.0, 23.5,
31.6, 25.9, 33.0, 24.8, 33.3)
```

となる（本当のデータはいくつか欠損を含んでいるのだが，それだと解析が面倒なので適当に補った）。これを `ttemp` という時系列データとして認識させるには，`library(ts)` として時系列解析パッケージが使えるようにしてから，`ttemp<-ts(temp, freq=2, start=c(23, 1))` とすれば，毎日 2 回の測定値があるデータを 23 日の 1 回目から，`temp` という配列から読んで，`ttemp` という時系列データに代入する操作をしたことになる。その後の時系列解析は，この `ttemp` に対して行う。このデータに対して AR モデルを当てはめるには，`ar(ttemp)` とすればよい。この例では，4 次の自己回帰係数が計算される。

AR 過程 $x(n)$ で表される確率システムの時刻 n における状態を $z(n)$ で表すことにすると，$z(n) = (x(n), x(n-1), ..., x(n-M+1))$ で与えられることがわかる。これと将来の入力 $w(n+1), w(n+2), ...$ がわかれば，将来の動き $x(n+1), x(n+2), ...$ が確定される。この場合，$w(n)$ は誤差項ではなく，システムを動かす入力と考えられる。すべての周波数成分を一様に含むホワイトノイズ $w(n)$ から当面の観測値の系列を生み出す確率過程の形で，多くの時系列モデルが与えられる。たとえば，$x(n)$ を 1 変量または多変量の確率過程として，状態を表すベクトル $z(n)$ を使って，$z(n) = Fz(n-1) + Gw(n)$, $x(n) = Hz(n)$（ここで F, G, H は，すべて適当な行列である）のように与えられる。ここで $w(n)$ を $w(n) + b_1 w(n-1) + ... + b_L w(n-L)$

で置き換えると，自己回帰移動平均過程 (ARMA モデル）が得られるが，詳細はここでは触れない。関心がある方は放送大学のテキストである尾崎統「時系列論」日本放送出版協会などを参照されたい。

12.2.3 フーリエ解析

次に周期的な変動をする関数としては正弦関数 (sin) と余弦関数 (cos) があるので，観察された周期性がいくつかの正弦関数と余弦関数の定数倍の和として表されると決めてしまい，この定数のセットを求めるアプローチがある（実は，すべての周期的な関数は，正弦関数と余弦関数の有限個または無限個の和で表現できることがわかっている)。物理学などで，ある光がどのような周波数の波がそれぞれどのくらいの強さで足しあわされたものかを調べる方法は，スペクトル解析 (spectral analysis) と呼ばれているが，実は周期的な波だけでなく，非周期的な波の表現方法としても有効なことがわかっており，一般にフーリエ解析 (Fourier analysis) と呼ばれる。フーリエ解析の計算方法としてよく行われるのが，高速フーリエ変換 (Fast Fourier Transformation; FFT) である。フーリエ解析を定式化すると，以下のようになる。いま，区間 $-T/2$ から $T/2$ までの間に，等間隔にとられた $2n$ 個の時系列データが与えられているとする。$x(t)$ は，$t = 0, \pm T/2n, \pm 2T/2n, ... \pm (n-1)T/2n, -T/2$ で与えられている[*8]。このとき，

$$x(t) = a_0 + 2\sum_{m=1}^{n-1}(a_m \cos\frac{2\pi mt}{T} + b_m \sin\frac{2\pi mt}{T}) + a_n \cos\frac{2\pi nt}{T}$$

と書ける。周波数 $1/T$ の整数倍の波の重ね合わせと考えるので，$1/T$ は基本周波数と呼ばれる。R では，例えば `fft(ttemp)` とすれば時系列データ `ttemp` に対して FFT が行われる。フーリエ解析については，ヒッポファミリークラブによって作られた「フーリエの冒険」という素晴らしい入門書があり，お薦めである。

フーリエ解析では，正弦関数や余弦関数は時間と独立である。しかし，世の中には，時間とともにばらつきが大きくなるような繰り返しもある。その場合は，観察された関数を，時刻を説明変数に含む周期的な関数であるウェーブレット (Wavelet) 関数の足し合わせに分解する方法がある。このやり方はウェーブレット解析と呼ばれ，非常に強力だが，まだ実際に使われた事例が少ない（2001 年 12 月第 1 週の Nature に，麻疹の流行パタンの分析に適用した論文が掲載されていた)。

[*8] データの個数が $2n$ 個なので，区間が $-T/2$ 以上 $T/2$ 未満であると考えて，$T/2$ は入れない。

12.3 生存時間解析の基礎

期間データを扱う方法としては，一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれるものがある。なかでもよく知られているものが Kaplan-Meier の積・極限推定量である（現在では，普通，カプラン・マイヤ推定量と呼ばれている）。カプラン・マイヤ推定量は，イベントが起こった各時点での，イベントが起こる可能性がある人口（リスク集合）あたりのイベント発生数を 1 から引いたものを掛け合わせて得られる，ノンパラメトリックな最尤推定量である。複数の期間データ列の差の比較には，ログランク検定や一般化ウィルコクソン検定が使われる。が，ログランク検定でも Mantel-Haenzel 流のログランク検定と Peto and Peto 流のログランク検定があったり，一般化ウィルコクソン検定でも Gehan-Breslow 流と Peto-Prentice 流があったりして，非常に面倒な話になってくるので，本書では説明しない。それらのノンパラメトリックな方法とは別に，イベントが起こるまでの時間が何らかのパラメトリックな分布に当てはまるかどうかを調べる方法もある。当てはめる分布としては指数分布やワイブル分布がある。イベントが起こるまでの期間に何らかの別の要因が与える効果を調べたいときはコックス回帰（それらが基準となる個人のハザードに対して $\exp(\sum \beta_i z_i)$ という比例定数の形で掛かるとする比例ハザード性を仮定する方法）と，パラメトリックなモデルに対数線形モデルの独立変数項として入れてしまう加速モデルがある[*9]。生存時間解析も，時系列解析と同じく，それだけで一冊の本ができるほど奥が深いので，今回はカプラン・マイヤ推定量の求め方だけ説明する。より詳しくは，大橋，浜田 (1995) などを参照されたい。

なお，データ数が多い場合は，個々の間隔データを集計して，生命表解析を行うこともある。生命表解析の代表的なものは，ヒトの平均寿命を計算するときに行われている（官庁統計としても，まさしく生命表という形で発表されている）。平均寿命とは 0 歳平均余命のことだが，これは，0 歳児 10 万人が，ある時点での年齢別死亡率に従って死んでいったとすると，生まれてから平均してどれくらいの期間生存するのかという値である[*10]。一般に x 歳平均余命は，x 歳以降の延べ生存期間の総和 (T_x) を x 歳時点の個体数 (l_x) で割れば得られる。延べ生存期間の総和は，年齢別死亡率 q_x が変化しないとして，$l_x(1-q_x/2)$ によって x 歳から $x+1$ 歳まで生きた人口 L_x（開

[*9] R では生存時間解析をするための関数は `survival` パッケージで提供されており，`library(survival)` とすれば使えるようになる。カプラン・マイヤ法は `survfit()` 関数，コックス回帰は `coxph()` 関数，加速モデルは `survreg()` 関数で実行できる。

[*10] 誤解されることが多いが，死亡年齢の平均値ではないので注意されたい。

MO_ID	MO_BD	C1_BD	C2_BD	C3_BD	C4_BD	C5_BD	C6_BD	C7_BD	C8_BD	C9_BD	C10_BD	C11_BD
20102	390000	0	640600	680000	711014	760000						
60202	250000	480415	560921	630000								
50102	400000	550000	590000	630000	660810	681011	710319	741018	760611	0		
30602	450000	580000	601004	630000	650000	670000	670000	720000	740000	750000	780714	
10502	400000	600716	630000	650807	670000	690609						
10102	400000	651103	681225	0	720200	0	790517	0	820000	840503	860527	890302
30102	490000	680000	700000	720000	750000	770000	820927					
10202	490000	680000	720826	760000	830000							
40302	580000	700000	780606	820906	901012	910606						
40102	570000	710114	730000	750000	770000	810621	840101	870802	920813			
20502	580000	720906	740704	761106	800407	811126	860516	910406				
50302	520000	730000	780000	800000	830000	870000	0					
10402	441101	730324	760723	770801	880119							
60302	460000	740000	770000	790000	800000	820000						
70202	550000	740000	780000	800000	840000	870000	890000	920000	941100			
70302	600000	750000	780000	800000	820000	850500	860000	880000	920000	940000		
20302	610000	760709	771020	790309	811002	850415	890803					
30702	600000	810500	820000	830000	840000	850000	900924	930430	950604			
30502	501205	820921	840803	881228								
60402	530000	830212	850216	900916	950921							
10802	650521	840623	861009	890727	920329	940416						
50402	670000	861114	880430	900130	910000	930325	950108					
20602	651114	870904	881111	900519	911104							
60102	570000	880000	950905									
10902	670000	900000	910000	950319								
30202	710000	900408	920210	940305								
40202	640000	901007	931109									
60204	680000	910000	920000									
50202	640000	911001	921020									
20202	711014	920801	931127									
10302	720826	920823	940308									
11002	700917	930303	950513									
10702	670304	930701										
80102	720229	940125										
11102	670809	940406										
30302	720000	940611										
50303	730000	950300										
10602	740700	950317										
20504	740704	950905										
60303	740000	951024										
70102	0	420000	450000	470000	520000	531225	550000	630000	670000			

図 12.3: ソロモン諸島のある村の女性全員の再生産史

始時点の人口が決まっていて死亡率も変化しないので x 歳の静止人口と呼ばれる）を求め，それを x 歳以降の全年齢について計算して和をとることで得られる。ヒトの人口学では年齢別死亡率から q_x を求めて生命表を計算するのが普通だが，生物一般について考えるときは，同時に生まれた複数個体（コホート）を追跡して年齢別生存数として l_x を直接求めてしまう方法（コホート生命表）とか，たんに年齢別個体数を l_x と見なしてしまう方法（静態生命表，偶然変動で高齢の個体数の方が多い場合があるので平滑化するのが普通）がよく行われる。

では，簡単な例を使って，カプラン・マイヤ推定量を説明しよう。表 12.3 は，ソロモン諸島のある村で，既婚女性全員に，自分の誕生日，第 1 子誕生日，第 2 子誕生日，……，末子誕生日（まだ出産を完了していない年齢の女性も含めて，ともかくそれまでに産んだ子どもの誕生日を全部）聞き取った結果である。間隔データを使わなければ，このデータから出生力について何かいうためには，出産を完了した女性についての平均出産数（平均完結パリティという）くらいしか指標がないが，間隔データを使えば，時間当たりの出生力を考えることができるので，出産を完了していない女

性のデータも使うことができる。

この種のデータには，以下の利点と欠点がある。

- 母親に対して，全ての子どもの出生年月日を聞き取ることは，統計がしっかりしていない社会でも比較的信頼性の高い方法である。
- 人口規模が小さくても使える上，過去の推計もできるという利点がある。
- 古くなるほど誤差が大きくなるバイアスや，他に影響を受ける要因が多いのは欠点。
- 結婚から第１子誕生までの期間や，第１子と第２子の出生間隔がよく使われるが，上にあげたソロモン諸島の社会では，結婚記念日はあまり正確に記憶されていなかったために，第１子と第２子の出産間隔を使うことにした。第１子と第２子の出産間隔には，第２子の在胎期間が含まれるために，その期間のハザードは原理的にゼロであることに注意する必要がある[*11]。

まず，カプラン・マイヤ推定量についての一般論を示す。イベントが起こる可能性がある状態になってから，イベントが起こった時点を $t_1, t_2, ...$ とし，t_1 時点でのイベント発生数を d_1，t_2 時点でのイベント発生数を d_2，以下同様であるとする。また，時点 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは，その直前でまだイベントが起きていない（この例では第１子出産後で第２子出産前の）個体数である。観察途中で死亡や転居などによって打ち切りが生じるために，リスク集合の大きさはイベント発生によってだけではなく，打ち切りによっても減少する。従って n_i は，時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる。なお，イベント発生と打ち切りが同時点で起きている場合は，打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。このとき，カプラン・マイヤ推定量 $\hat{S}(t)$ は，

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

として得られる。その標準誤差はグリーンウッドの公式により，説明は省略するが，

$$var(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

で得られる。なお，カプラン・マイヤ推定量を計算するときは，階段状のプロットを同時に行うのが普通である。

[*11] 例えば，在胎期間の推定値として 9 ヶ月を引いた値をデータにしたり，または在胎期間を切片として含んだハザード関数を推定することも考慮するべきである。

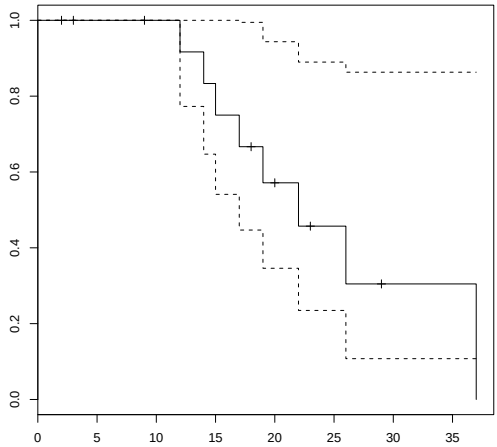


図 12.4: ソロモン諸島女性の第 1 出産間隔についてのカプラン・マイヤプロット

R では，library(survival) としてパッケージを呼び出し，dat <- Surv(**生存時間，打ち切りフラグ**) 関数で生存時間データを作り（打ち切りフラグは 1 でイベント発生，0 が打ち切り。ただし区間打ち切りの場合は 2 とか 3 も使う），res <- survfit(dat) でカプラン・マイヤ法によるメディアン生存時間が得られ，plot(res) とすれば階段関数が描かれる。イベント発生時点ごとの値を見るには，summary(res) とすればよい[*12]。

[*12] 参考までに書いておくと，生データがイベント発生の日付を示している場合，間隔を計算するには difftime() 関数や ISOdate() 関数を使うと便利である。例えば，

```
dob
1964-8-21
```

という形のテキストファイル L12-1.dat で，日付が与えられているとしよう。これを x <- read.delim("L12-1.dat") として読み込み，2003 年 6 月 11 日までの間隔を計算したければ，difftime(ISOdate(2003,6,11),x$dob) とすれば，その間の経過日数が DateTimeClasses のオブジェクトとして得られる。日数から年に変換したければ，例えば as.integer(difftime(ISOdate(2003,6,11),x$dob))/365.24 とすればいいし，さらに 12 を掛ければ月単位になる。

例えば，区間打ち切り（イベント発生までの時間がある幅をもってしかわからないデータ）を無視して，上で示したソロモン諸島のデータのうち，第 1 子出生が 1986 年以降のものの出産間隔データを R で分析すると[*13]，右側打ち切りを考慮した出産間隔のメディアンが 22 ヶ月であることがわかる（プロットを図 12.4 に示す）。

[*13] プログラムは下記の通り。

```
library(survival)
time <- c(17,14,22,37,12,15,19,26,29,23,20,18,9,9,3,2)
event <- c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
dat <- Surv(time,event)
res <- survfit(dat)
print(res)
summary(res)
plot(res)
```

第 13 章

一般化線型モデル入門

13.1 一般化線型モデルとは？

第 11 章で説明したように，普通の量的変数の間の線型回帰を一般化すれば，t 検定，分散分析，共分散分析，回帰分析，重回帰分析，ロジスティック回帰分析，正準相関分析など多くの分析方法を共通の数学モデルで扱うことができる。このモデルは一般化線型モデルと呼ばれる。英語では Generalized Linear Model といい，R での関数名も `glm()` である[*1]。

一般化線型モデルは，基本的には，$Y = \beta_0 + \beta X + \varepsilon$ という形で表される（Y が従属変数群，X が独立変数群（及びそれらの交互作用項），β_0 が切片群，β が係数群，ε が誤差項である）。係数は最小二乗法または最尤法で数値的に求める。以下，先にあげたいくつかの分析が，どのように一般化線型モデルを特殊化したものなのかを説明し，その中で重回帰分析と共分散分析について若干の補足説明を加える。

13.2 変数の種類と数の違いによる線型モデルの分類

以下のように整理すると，t 検定，分散分析，回帰分析といった分析法が，すべて一般化線型モデルの枠組みで扱えることがわかる。

例えば，建物の型の変数 (BD) を集合住宅 1，一戸建て 2 とした場合の，東京のと

[*1] 線型は linear の訳で，一次結合という意味なのだが，漢字としては線形と書かれることもある。厳密な区分はないように思われるが，`glm()` の場合は「型」の字を使う方が普通のようである。なお，一般化線型モデルのうち，ある条件を満たすものを一般線型モデル (General Linear Models) と呼び，SAS の PROC GLM はこれに当たる。

分析名	従属変数 (Y)	独立変数 (X)
t 検定（注 1）	量的変数 1 つ	2 値変数 1 つ
一元配置分散分析	量的変数 1 つ	カテゴリ変数 1 つ
多元配置分散分析	量的変数 1 つ	カテゴリ変数複数
回帰分析	量的変数 1 つ	量的変数 1 つ
重回帰分析	量的変数 1 つ	量的変数複数（注 2）
共分散分析	量的変数 1 つ	（注 3）
ロジスティック回帰分析	2 値変数 1 つ	2 値変数，カテゴリ変数，量的変数複数
正準相関分析	量的変数複数	量的変数複数

（注 1）Welch の方法でない場合。
（注 2）カテゴリ変数はダミー変数化
（注 3）2 値変数 1 つと量的変数 1 つの場合が多いが，「2 値変数またはカテゴリ変数 1 つまたは複数」と「量的変数 1 つまたは複数」を両方含めば使える。

ある大学の学生実習で測定した水道水質の総硬度 (HARD) の平均値に，建物の型によって差があるかどうかを検定したいとする。
　等分散性を仮定すれば，R では，

```
BD <- c(1,1,1,1,1,1,2,2,1,1,2,1,1,2,1,1,2,2,1,1,1,1,1,2,1,1,2,1,1,2,1,2,1,1)
HARD <- c(88.280, 103.500, 119.600, 96.210, 109.340, 100.500, 81.390, 75.715,
112.880, 101.150, 84.400, 102.900, 65.000, 97.445, 101.850, 79.100, 103.620,
69.270, 97.090, 101.150, 89.820, 108.560, 98.810, 103.620, 85.940, 89.230,
69.300, 101.150, 101.150, 73.070, 62.695, 148.590, 93.080, 103.500)
```

　としてデータを定義した後，`t.test(HARD ~ BD, var.equal=T)` とすることによって，以下の結果が得られる。

```
Two Sample t-test
data:  HARD by BD
t = 0.8843, df = 32, p-value = 0.3831
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.444719 18.867802
sample estimates:
mean in group 1 mean in group 2
```

```
96.35354        90.64200
```

一般化線型モデルを使って，建物の型を独立変数として総硬度を従属変数としたモデルの当てはめをしてみるには，R では，データ定義後に，`summary(glm(HARD ~ BD))` とすればよい。以下の結果が得られる。

```
Call:
glm(formula = HARD ~ BD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-33.659   -8.957    3.301    7.061   57.948

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  102.065      8.861  11.518 6.41e-13 ***
BD            -5.712      6.459  -0.884    0.383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 294.4717)

    Null deviance: 9653.4  on 33  degrees of freedom
Residual deviance: 9423.1  on 32  degrees of freedom
AIC: 293.72

Number of Fisher Scoring iterations: 2
```

が得られる。`Coefficients:` の `BD` のところを見ると，`t value` が -0.884，その有意確率が 0.383 となっていて，t 検定の結果と一致している（t 値の符号が違うが，t 分布は左右対称なので両側検定では符号が違っても同じ意味）ことがわかる。

この場合は，当然のことながら，普通の線型モデルでも同じ結果が得られるし，分散分析でも同じ結果となる。つまり，t 検定は分散分析の特殊な場合ということができるし，分散分析は線型モデルの特殊な場合ということができるし，線型モデルは一般化線型モデルの特殊な場合（当然だが）ということができる。

13.3　重回帰分析

複数の独立変数を同時にモデルに投入することにより，従属変数に対する，他の影響を調整した個々の変数の影響をみることができる。

重回帰分析は，何よりもモデル全体で評価することが大切である。例えば，独立変数が年齢と体重と一日当たりエネルギー摂取量，従属変数が血圧というモデルを立てれば，年齢の偏回帰係数（または偏相関係数または標準化偏回帰係数）は，体重と一日当たりエネルギー摂取量の影響を調整した（取り除いた）後の年齢と血圧の関係を示す値だし，体重の偏回帰係数は年齢と一日当たりエネルギー摂取量の影響を調整した後の体重と血圧の関係を示す値だし，一日当たりエネルギー摂取量の偏回帰係数は，年齢と体重の影響を調整した後の一日当たりエネルギー摂取量と血圧の関係を示す値である。独立変数が年齢と体重で従属変数が血圧である場合の年齢の偏回帰係数は，独立変数に一日当たりエネルギー摂取量も入っている場合の年齢の偏回帰係数とは**異なる**。

モデル全体としてのデータへの当てはまりは，重相関係数の 2 乗（決定係数）や，AIC で評価する。

偏回帰係数の有意性検定は，偏相関係数がゼロである確率を t 検定によって求める。１つの重回帰式の中で，相対的にどの独立変数が従属変数（の分散）に対して大きな影響を与えているかは，偏相関係数の二乗の大小によって評価するか，または標準化偏回帰係数によって比較することができる。しかし，原則としては，別の重回帰モデルとの間では比較不可能である。

たくさんの独立変数の候補からステップワイズ法によって比較的少数の独立変数を選択することが良く行われる。しかし，モデル全体で評価するという観点からは，あまり薦められない。数値以外の根拠により投入する変数を決めて，各々の偏回帰係数（または偏相関係数）が有意であるかないかを見る方が筋がよい。十分な理由があれば，有意でない変数も含めた重回帰式を作っても良い。

しかし，数値以外の根拠が薄い場合もあるし，偏回帰係数が有意でない（偏相関係数がゼロであるという帰無仮説が成り立つ確率が 5%より大きい）変数を重回帰モデルに含めることを嫌う立場もある。従って，数値から最適なモデルを求める必要もありうる。そのためには，独立変数が１個の場合，２個の場合，３個の場合，……，のそれぞれについてすべての組み合わせの重回帰モデルを試して，最も重相関係数の二乗が大きなモデルを求めて，独立変数がｎ個の場合が，ｎ－１個の場合のすべての変数を含むならば尤度比検定を行って，尤度が有意に大きくならないところまでのｎ－

1個を独立変数として採用するのが良い。SAS では PROC REG の MAXR というオプションで可能である。

13.4 共分散分析

典型的には，$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ というモデルになる。2値変数 X_1 によって示される2群間で，量的変数 Y の平均値に差があるかどうかを比べるのだが，Y が量的変数 X_2 と相関がある場合に（このとき X_2 を共変量と呼ぶ），X_2 と Y の回帰直線の傾き (slope) が X_1 の示す2群間で差がないときに，X_2 による影響を調整した Y の修正平均（adjusted mean; 調整平均ともいう）に，X_1 の2群間で差があるかどうかを検定する。

R では，X_1 を示す変数名を `C`（注：`C` は `factor` である必要がある），X_2 を示す変数名を `X` とし，Y を示す変数名を `Y` とすると，`summary(glm(Y~C+X))` とすれば，`X` の影響を調整した上で，`C` 間で `Y` の修正平均（調整平均）が等しいという帰無仮説についての検定結果が得られる（`C2` と表示される行の右端に出ているのがその有意確率である）。ただし，この検定をする前に，2本の回帰直線がともに有意にデータに適合していて，かつ2本の回帰直線の間で傾き (slope) が等しいかどうかを検定して，傾きが等しいことを確かめておかないと，修正平均の比較には意味がない。そこで，まず例えば，`summary(lm(Y[C==1]~X[C==1]); summary(lm(Y[C==2]~X[C==2])` として2つの回帰直線それぞれの適合を確かめ，`summary(glm(Y~C+X+C*X))` として傾きが等しいかどうかを確かめなければならない。傾きが有意に違っていることは，`C` と `X` の交互作用項が有意に `Y` に効いていることと同値なので，`Coefficients` の `C2:X` と書かれている行の右端を見れば，「傾きが等しい」を帰無仮説とした場合の有意確率が得られる。そもそも回帰直線の適合が悪ければその独立変数は共変量として考慮する必要がないし，傾きが違っていれば群分け変数と独立変数の交互作用が従属変数に関して有意に影響しているということなので，2群を層別して別々に解釈する方が良い。

参考までに数式でも説明しておく。いま，`C` で群分けされる2つの母集団における，(X, Y) の間の母回帰直線を，$y = \alpha_1 + \beta_1 x$，$y = \alpha_2 + \beta_2 x$ とすれば，次の2つの仮説が考えられる。まず傾きに差があるかどうか？　を考える。つまり，$H_0 : \beta_1 = \beta_2$，$H_1 : \beta_1 \neq \beta_2$ である。次に，もし傾きが等しかったら，y 切片も等しいかどうかを考える。つまり，$\beta_1 = \beta_2$ のもとで，$H_0' : \alpha_1 = \alpha_2$，$H_1' : \alpha_1 \neq \alpha_2$ を検定する。各群

について，X と Y の平均と変動と共変動を出しておけば[*2]，仮説 H_1 のもとでの残差平方和

$$d_1 = SS_{Y1} - (SS_{XY1})^2/SS_{X1} + SS_{Y2} - (SS_{XY2})^2/SS_{X2}$$

と仮説 H_0 のもとでの残差平方和

$$d_2 = SS_{Y1} + SS_{Y2} - (SS_{XY1} + SS_{XY2})^2/(SS_{X1} + SS_{X2})$$

を計算して $F = (d_2 - d_1)/(d_1/(N-4))$ が H_0 のもとで第 1 自由度 1，第 2 自由度 $N-4$ の F 分布に従うことを使って傾きが等しいかどうかの検定ができる。H_0 が棄却されたときは，$\beta_1 = SS_{XY1}/SS_{X1}$，$\beta_2 = SS_{XY2}/SS_{X2}$ として別々に傾きを推定し，y 切片 α もそれぞれの式に各群の平均値を入れて計算できる。H_0 が採択されたときは，共通の傾き β を，$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$ として推定する。この場合はさらに y 切片が等しいという帰無仮説 H_0' のもとで全部のデータを使った残差平方和 $d_3 = SS_Y - (SS_{XY})^2/SS_X$ を計算して，$F = (d_3 - d_2)/(d_2/(N-3))$ が第 1 自由度 1，第 2 自由度 $N-3$ の F 分布に従うことを使って検定できる。H_0' が棄却された場合は各群の平均を共通の傾きに代入すれば各群の切片が求められるし，採択されたら，要するに 2 群間に差がないということになるので，2 群を一緒にして普通の単回帰分析をしていいことになる。

[*2] サンプルサイズ $N1$ の第 1 群に属する x_i, y_i について，$E_{X1} = \sum x_i/N1$，$SS_{X1} = \sum(x_i - E_{X1})^2$，$E_{Y1} = \sum y_i/N1$，$SS_{Y1} = \sum(y_i - E_{Y1})^2$，$E_{XY1} = \sum x_i y_i/N1$，$SS_{XY1} = \sum(x_i y_i - E_{XY1})^2$。第 2 群も同様。

例題

下表は，都道府県別のデータで，1990 年の 100 世帯あたり乗用車台数（CAR1990）と，1989 年の人口 10 万人当たり交通事故死者数（TA1989）と，1985 年の国勢調査による人口集中地区居住割合（DIDP1985）である。REGION の 1 は東日本，2 は西日本を意味する。東日本は西日本よりも，人口集中地区居住割合を調整しても乗用車保有台数が多いと言えるか？

PREF	REGION	CAR1990	TA1989	DIDP1985
Hokkaido	1	86	11.6	66.7
Aomori	1	78.9	9.5	42.2
Iwate	1	86.6	9.7	27.5
Miyagi	1	92.7	7.9	50.7
Akita	1	90.3	8.1	31.2
Yamagata	1	104.7	7.1	36.7
Fukushima	1	102.7	12.1	33.6
Ibaraki	1	120.7	16.4	29.2
Tochigi	1	122.2	16.5	35.1
Gunma	1	123.9	11.5	38.2
Saitama	1	88.7	7.3	71.7
Chiba	1	86.4	8.8	65
Tokyo	1	58.2	4.1	97.1
Kanagawa	1	75.5	7.2	89.1
Niigata	1	93.2	11.1	42.6
Toyama	1	113	11.1	37.9
Ishikawa	1	99.1	9.5	46.4
Fukui	1	109.4	14.7	35.9
Yamanashi	1	112.8	13.8	31.2
Nagano	1	110.9	9.6	31.1
Gifu	1	119.7	12	36.8
Shizuoka	1	107.5	10.5	51.5
Aichi	1	107.2	8.2	67.2
Mie	1	106.7	13.7	38
Shiga	2	104.4	14.5	29.1
Kyoto	2	75.5	8.9	79.5
Osaka	2	62.8	5.9	93.8
Hyogo	2	75.6	8.9	71.7
Nara	2	86	9.3	52.7
Wakayama	2	83	11.6	42.3
Tottori	2	92.1	11.8	26.2
Shimane	2	86.9	9.9	23.4
Okayama	2	95.7	11.3	33.9
Hiroshima	2	79.6	9.7	58.5
Yamaguchi	2	84.4	11.5	44
Tokushima	2	90.7	10.9	27.4
Kagawa	2	89.8	14.3	32.3
Ehime	2	72.3	10.9	43.1
Kochi	2	74.9	11.3	38.4
Fukuoka	2	82.3	8	63.3
Saga	2	97.4	12.8	27.6
Nagasaki	2	69.3	5.9	41.6
Kumamoto	2	87.3	8.5	36.6
Oita	2	82.5	8.7	40.4
Miyazaki	2	85.7	7.4	39
Kagoshima	2	70.5	7.3	36.3
Okinawa	2	100.3	7.6	56.5

人口集中地区[*3]人口割合が高い都道府県ほど人がまとまって住んでいるわけだから，先験的に，そういう都道府県ほどマイカー保有率は低くて済みそうだと思われる。したがって，人口集中地区人口割合によってマイカー保有率を調整しなくては，それ以外の要因（例えば，公共交通機関の整備の割合や，自動車産業の発達の度合い，ディーラーの営業活動の熱心さ，平均世帯規模，郊外型大型店舗の展開の度合い，道路政策，等々）による東日本と西日本のマイカー保有率への影響を評価できないことになる。

東日本を○で，西日本を△でプロットし，東日本の回帰直線を実線，西日本の回帰

[*3] 1 km^2 当たりの人口密度が 4,000 人以上の集合地区で，かつ合計人口が 5,000 人以上の地区をいう。

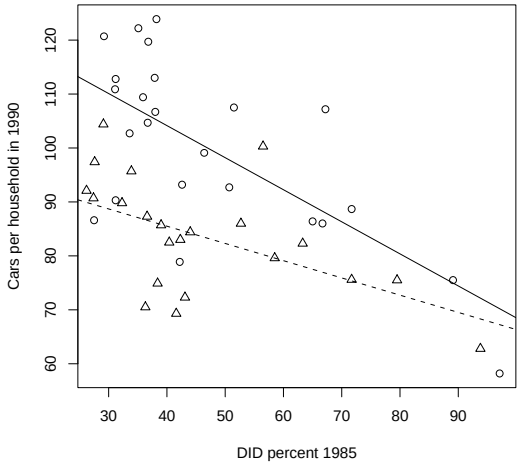


図 13.1: 交通事故件数と世帯当たりの自家用車保有台数の関係の東日本と西日本の比較

直線を点線で追加すると，図 13.1 のようになる[*4]。

R での共分散分析の手順は，まず

```
summary(lm(CAR1990[REGION==1] ~ DIDP1985[REGION==1]))
summary(lm(CAR1990[REGION==2] ~ DIDP1985[REGION==2]))
```

[*4] この図を描く R のプログラムは，例えば

```
x <- read.table("anacova.dat")
attach(x)
```

としてデータを読み込んでから，

```
plot(CAR1990[REGION==1]~DIDP1985[REGION==1],pch=1,
xlab='DID percent 1985',ylab='Cars per household in 1990')
points(CAR1990[REGION==2]~DIDP1985[REGION==2],pch=2)
abline(lm(CAR1990[REGION==2]~DIDP1985[REGION==2]),lty=2)
abline(lm(CAR1990[REGION==1]~DIDP1985[REGION==1]),lty=1)
```

とすればよい

とする。得られる結果

```
Call:
lm(formula = CAR1990[REGION == 1] ~ DIDP1985[REGION == 1])

Residuals:
     Min       1Q   Median       3Q      Max
-24.9808  -5.1307   0.9493   8.2336  19.2190

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          127.9283     6.7699  18.897 4.35e-15 ***
DIDP1985[REGION == 1]  -0.5945     0.1333  -4.459 0.000197 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.24 on 22 degrees of freedom
Multiple R-Squared: 0.4747,Adjusted R-squared: 0.4508
F-statistic: 19.88 on 1 and 22 DF,  p-value: 0.0001967

Call:
lm(formula = CAR1990[REGION == 2] ~ DIDP1985[REGION == 2])

Residuals:
     Min       1Q   Median       3Q      Max
-16.1869  -3.3935   0.2297   3.4338  20.0706

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           98.2912     5.0750  19.368 7.12e-15 ***
DIDP1985[REGION == 2]  -0.3197     0.1047  -3.053  0.00604 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 8.904 on 21 degrees of freedom
Multiple R-Squared: 0.3075,Adjusted R-squared: 0.2745
F-statistic: 9.323 on 1 and 21 DF,  p-value: 0.006037
```

から，これらの回帰式が両方とも有意にデータに適合していることがわかる。次に，

```
summary(glm(CAR1990 ~ as.factor(REGION)+DIDP1985
+as.factor(REGION)*DIDP1985))
```

とすれば交互作用項の係数の有意性をみることができて，有意確率が 0.118 という結果が得られるので傾きには差がないとわかる。最後に

```
summary(glm(CAR1990 ~ as.factor(REGION)+DIDP1985))
```

として as.factor(REGION)2 の有意確率をみると 0.05 より遥かに小さいので，修正平均にも差があるとわかる。つまり，東日本と西日本では，人口集中地区への居住割合の影響を調整しても，世帯当たりの自動車保有台数には有意に差があるといえる。

13.5　補足：一般線型混合モデル

複数の対象についての経時的観察データが複数あるときに，個体間の経時的な変化のパタンの違いをモデルに取り込むことによって一般化線型モデルをさらに一般化したのが一般線型混合モデル (General Linear Mixed Model) である。高度な分析なのでここでは説明しないが，非常に強力である。R では，nlme というライブラリが提供されている。8 歳から 14 歳まで 2 年おきに歯列矯正の指標として，頭蓋の X 線写真により下垂体から翼上顎裂までの距離 (mm) を，男児 16 人，女児 11 人について測定したデータ（Orthodont という組み込みデータ）による実行例は，library(nlme) としてから，example(lme) とすれば見ることができる。年齢によるモデル，性と年齢と個体差によるモデルについて出力される。

第 14 章

高度な解析法についての概説

14.1 主成分分析

n 個体のサンプルがあって，それぞれについて，p 個の変数 $x_1, x_2, ..., x_p$ の観測値が得られているとする。一般に，p 個の変数の情報を全部一度に考えて n 個体の情報を把握することは難しい。そこで考えられるのが，p 個の変数を，もっと少ない数の，互いに独立な主成分 (principal component) で表せないかということである。

いま，主成分 $\xi_1, \xi_2, ..., \xi_p$ を考え，これらを x の一次関数で表すことにする。つまり，

$$\xi_i = \sum_{j=1}^{p} l_{ij} x_j$$

として，p^2 個の適当な係数 l_{ij} を見つけることを考える。各 x_j をそれぞれの平均からの偏差として測れば，どの x_j も n 個体についての和はゼロになり，従って ξ_i の和もゼロになる。ここで p 個の ξ は互いに無相関であるとする。すなわち

$$E(\xi_i \xi_j) = E(\{\sum_{k=1}^{p} l_{ik} x_k \sum_{m=1}^{p} l_{jm} x_m\} = 0 \qquad (i \neq j)$$

とする。これだけではまだ $p(p+1)/2$ 個の自由度が残っているので，この変換を直交変換であると条件付ける，すなわち

$$\sum_{k=1}^{p} l_{ik} l_{jk} = 0 (i \neq j), = 1 (i = j)$$

とすれば，符号の付け替えの自由度を加味しても有限組の解が得られることになる (数学的な解は行列の固有値と固有ベクトルを求めることによって得られるが，普通

はコンピュータソフトにやらせるので説明は省略する)。より詳しくは，ケンドール (1981) を参照されたい。

この新しい変数 ξ は主成分と呼ばれる。ξ は，もとの変数 x が正規分布に従うなら互いに独立である。行列の固有値の大きさの順に $\xi_1, \xi_2, ..., \xi_p$ と番号をつけると，これらは順に第 1 主成分，第 2 主成分，...，第 p 主成分と呼ばれる。第 1 主成分は，あらゆる一次関数の中で可能な最大の分散をもつ。第 2 主成分は第 1 主成分と無相関な一次関数の中で可能な最大の分散をもつ。このようにして主成分を決めると，それぞれの固有値の，固有値の和に対する割合を使って，それぞれの主成分が全変動の何パーセントを説明するかを表すことができる。それを主成分の寄与率と呼ぶ。普通は，たくさんの変数から少数（例えば 2 つとか 3 つ）の主成分だけを使って全変動の 80%が説明できる，のように使う。

本当はこんなに少数のデータに使うような分析法ではないのだが，前章の例題で使ったデータについて，R を使って主成分分析をしてみる。まず，`library(mva)` として多変量解析ライブラリを呼び出しておく必要がある。ついで，`mat<-matrix(c(CAR1990,TA1989,DIDP1985),nrow=47)` として `res<-princomp(mat); summary(res)` とすれば，下表が得られる。

	Comp.1	Comp.2	Comp.3
Standard deviation	21.1842224	11.7510982	1.897637799
Proportion of Variance	0.7600359	0.2338654	0.006098678
Cumulative Proportion	0.7600359	0.9939013	1.000000000

この結果から, 第 1 主成分の寄与率が 76%, 第 2 主成分までの累積寄与率が 99%で, 取り上げた 3 つの変数のばらつきは，ほぼ完全に 2 つの直交する主成分に分解できることがわかった。そこで，各都道府県の第 1 主成分（得点）と第 2 主成分（得点）を図示するには，`biplot(res,xlabs=PREF)` とすれば，図 14.1 が得られる。

14.2　因子分析

思想は逆だが，数学的には因子分析は，主成分分析に良く似ている。つまり，p 個の観測された変数 x があるときに，個々の x が m 個（$m < p$）の潜在因子の線型結合と誤差によって表されると考える。たくさんの変数を，別の少数の変数の線型結合によって表すことによって情報を集約する方法論である。R では `factanal` という最尤法で因子分析を行う関数があるが，3 つの変数を 2 つの因子で説明することはできず，1 つの因子しか想定できない（上の例題のデータで `factanal(mat,2)` とする

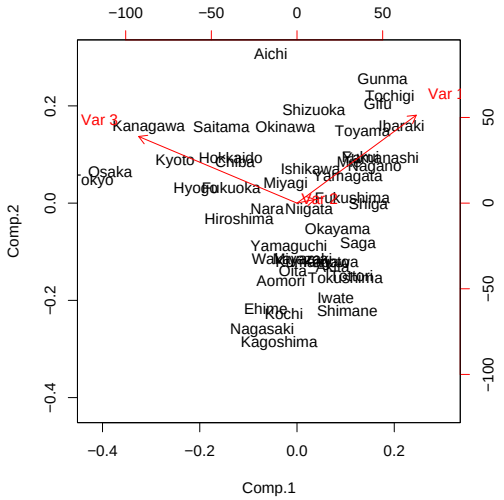


図 14.1: 都道府県別の交通事故件数，人口集中地区割合，世帯当たりの自家用車保有台数についての主成分分析の結果

とエラーが出る。)。`factanal(mat,1)` とすると，第 1 因子の因子負荷量は 1.764 であり，寄与率は 0.588 である。このことは，取り上げた 3 つの変数は，共通の潜在因子によって約 59%説明されるということを意味する。少数の主成分また因子による累積寄与率を最大にするために varimax 回転や promax 回転を行うことがあるが，R ではこれらの関数も用意されている。

因子分析は，観測された変数 (observed variables; 観測変数) 間の関係が，実は測定不可能な構成概念 (construct)，即ち因子 (factor) との関係によって説明されると捉えるモデルであるということもできる。しかし因子分析には，観測変数間の関係は，因子との関係においてしか説明できないし，因子間の因果関係を論じることができないし，仮説検証ができないという欠点がある。そこで，測定不可能な因子間の関係もあるだろうけれど，すべてをそれで説明しようとするのではなくて，観測変数間の直接的な関係をまず考えて，それで説明しきれない部分を測定不可能な潜在変数 (latent variables) と変数間の因果関係を不完全にする偶然変動としての誤差変数 (error variables) によって補う，というアプローチが考えられる。これが共分散構造

分析 (covariance structure analysis)[*1]である。その前段階として，因子分析に仮説検証機能を追加した確認的因子分析がある。共分散構造分析は，潜在変数間の関係を表す構造方程式モデルと，観測変数間の関係を表す測定方程式モデルを，誤差変数を入れて結合したものであるということができる。統計パッケージでは，SAS では PROC CALIS，SPSS では AMOS という追加パッケージを使う。R でも sem というライブラリで実行できる。

14.3　クラスター分析

変数間のでなく，データ間の関係を表したいときに使うのがクラスター分析である。クラスター分析には，距離行列に基づいて個体を結合しながらクラスターを積み上げていく（出力は樹状図またはネットワーク図になる）階層的手法と，予めいくつくらいの塊（クラスター）に分かれるかを決めて，データを適当に振り分ける非階層的手法がある。

距離行列の計算法にも多々あり，結合法にも多々ある。いくつかの方法でやってみて，樹状図に差がなければ，そのクラスター分析の結果は安定していて，信頼できるといえる。樹状図が大きく変わるようなら信頼できない。解釈としては，変数が足りないために，個体間の関係が十分にわからないと考える。例としては，R で，先ほどのデータを読んで mva ライブラリを呼び出した後で，

```
mat <- matrix(c(CAR1990,TA1989,DIDP1985),nrow=47)
dis <- dist(mat,method="euclidean")
clus <- hclust(dis)
op<-par(ps=8)
plot(clus,PREF,xlab="",ylab="",sub="")
par(op)
```

とすれば，樹状図 (dendrogram とか tree とかいう）が図 14.2 のように描ける。R ではデフォルトの距離の計算法はユークリッド距離（要するに差の二乗和），クラスター結合法は，完全連結法 (complete linkage) である。クラスター分析の結果は見やすいが，解釈には主観が入りがちである。ちなみに山口は和歌山と最も近いようである。

非階層的手法の k-means 法の R での実行例も示しておく。 5 つのクラスタを仮定

[*1] 共分散構造解析と訳すこともある。

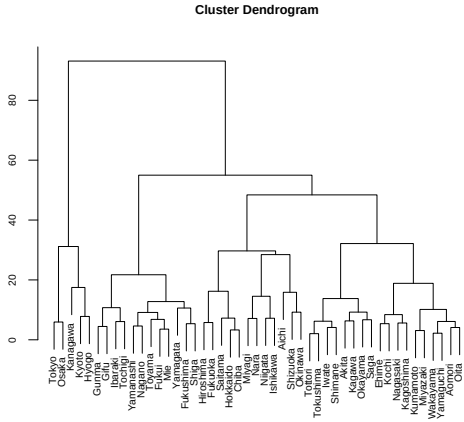


図 14.2: 都道府県別の交通事故件数，人口集中地区割合，世帯当たりの自家用車保有台数についての階層的クラスター分析結果

すると，データを読んで mva ライブラリ呼出し後，

```
clus5 <- kmeans(mat,5)
op<-par(ps=7)
plot(CAR1990,TA1989,pch=clus5$cluster,xlim=c(50,130),ylim=c(2,18))
text(CAR1990,TA1989,paste(PREF),pos=1)
par(op)
```

によって図 14.3 が得られる。

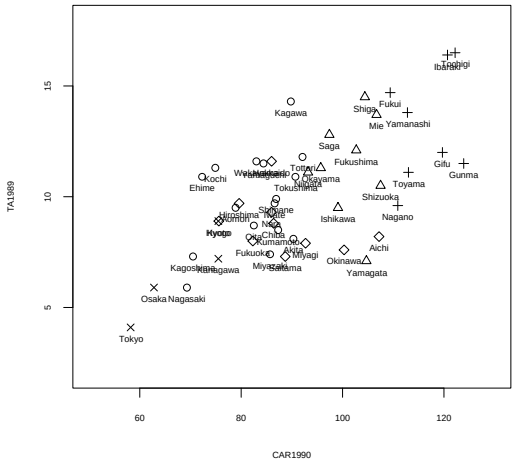


図 14.3: 都道府県別の交通事故件数，人口集中地区割合，世帯当たりの自家用車保有台数についての非階層的クラスター分析（kmeans 法）の結果

第 15 章

参考文献

参考文献というよりも，英語で言えば Further Readings の推薦なのだが，本文中でも触れた以下の書籍はいずれも良書だと思う。

- ラオ，C．R．（藤越康祝，柳井晴夫，田栗正章訳）『統計学とは何か　■偶然を生かす■』，丸善，1993 年：統計的な「ものの考え方」について，古典的な例や身近な例から，かなり高度な話題まで幅広く取り上げ，切れ味のよい解説を加えた名著である。
- 鈴木義一郎『情報量規準による統計解析入門』，講談社，1995 年：モデルベースの解析をするために，統計モデルがどういう意味をもつのか，その当てはめはどのように評価すべきか，ということを基礎から丁寧に解説している本であり，初学者にも薦められると思う。
- 浜田知久馬『学会・論文発表のための統計学　統計パッケージを誤用しないために』，真興交易（株）医書出版部，1999 年：自分でパッケージを使って統計解析をするときに気をつけなくてはいけないポイントを要領よくまとめた本。論文を読むときに，そこで使われている手法のどういう点に注意して結果を読み取らなくてはいけないか，ということもわかる。
- 粕谷英一『生物学を学ぶ人のための統計のはなし―きみにも出せる有意差』，文一総合出版，1998 年：統計解析をやり始めた大学院生などが陥りやすい罠，統計結果を読むときに間違いやすい点などを対話形式の軽妙な調子で書いた本であり，とっつきやすいと思う。
- ケンドール，M．G．（奥野忠一，大橋靖雄訳）『多変量解析』，培風館，1981 年：方法の羅列やパッケージの出力の見方に終始する多変量解析の解説書が多い中で，この本は多変量解析の意味を丁寧に，しかも数式は必要最小限しか使

わずに解説した良書である。絶版らしいのは残念なことである。
- 竹村彰通『現代数理統計学』，創文社，年：統計学を本気で学びたい人は，この本を理解することから入るとよいと思う。腰を据えてかからないと制覇できない高い山であるが，統計学に対する理解の次元が変わる。
- Rothman KJ ”Epidemiology: An Introduction”, Oxford University Press, 2002.
- 伏見正則『理工学者が書いた数学の本　確率と確率過程』，講談社，1987 年：確率の捉え方について明快に書かれている。
- 池田央『調査と測定』，新曜社，1980 年：尺度について厳密な説明が与えられている。
- Grimm LG ”Statistical Applications for the Behavioral Sciences”, John Wiley & Sons, New York, 1993.
- 豊川裕之，柳井晴夫（編著）『医学・保健学の例題による統計学』，現代数学社，1982 年.
- 永田靖，吉田道弘『統計的多重比較法の基礎』，サイエンティスト社, 1997 年：多重比較についてきわめて丁寧に論じ尽くした教科書である。ただし，「基礎」と銘打たれてはいるが，経験を積んだ研究者を対象として書かれており，学部学生が読むにはかなり難しい。
- 大橋靖雄，浜田知久馬『生存時間解析　SAS による生物統計』，東京大学出版会，1995 年.
- 竹内啓，大橋靖雄『数学セミナー増刊　入門｜現代の数学 [11] 統計的推測－２標本問題』，日本評論社，1981 年
- 伊藤嘉昭監修・粕谷英一，藤田和幸『動物行動学のための統計学』，東海大学出版会，1984 年
- ジョエル・E・コーエン著，重定南奈子・瀬野裕美・高須夫悟訳『新「人口論」：生態学的アプローチ』，農山漁村文化協会，1998 年

付録：R について

A.1 なぜ R を使うべきなのか？

R は R-project (`http://www.r-project.org/`) という国際共同プロジェクトで開発されている，ソースが公開されていて，誰でも自由に利用できる統計ソフトである。ベル研で開発され市販されている S や S-Plus と 8 割くらいは共通の言語仕様をもつとされ，Poorman's S と呼ばれることもある。

R の操作性を市販統計ソフトと比べると，STATISTICA とか JMP とか SPSS のようなメニューから統計手法を選ぶものとはまったく異なり，関数を打たねばならないので，初めて使うときはとっつきにくいかもしれない。が，市販ソフトの中でも，プログラムをタイプするという意味では，SAS と似ている[*1]。記述統計から多変量解析まで，およそ必要な統計解析はすべてできるといっていいし，計算に使われている手法も新しい。大手市販ソフトでは新しい方法のフォローが遅いので，時としてベストではない統計手法が使われたりするのだが，R は現時点でベストとされるかあるいは標準的な手法がデフォルトになっている。例えば擬似乱数列生成のアルゴリズムはバージョン 1.7.0 からメルセンヌツイスター（第 1 章参照）がデフォルトになったし，多重比較のデフォルトは Holm の方法になっている（第 10 章参照）。

プログラムをタイプするのが面倒だと思う人もいるかもしれないが，よく考えてみれば，それが誤解に過ぎないことがわかると思う。少なくとも MS-Excel でマクロを使うよりもずっと簡単な場合が多い。

例えば，MS-Excel で独立 2 標本の平均値の差の検定をするには，ツールの分析ツール（アドインなので，フルインストールするか，インストール時に指定しないと入らない）を使うわけだが，まず等分散性の検定を選び，標本の範囲をそれぞれ指定して実行し，その結果等分散という帰無仮説が棄却されなければ，等分散を仮定した

[*1] もっとも，SAS のプログラムを FORTRAN とすれば，C++や APL くらいに洗練されている言語体系だと思う。

２群の平均値の差の検定を選んで再び２つの標本の範囲を選んで実行するし，棄却されたときは等分散でないときの２群の平均値の差の検定を選んで２つの標本の範囲を選んで実行する，という手順を踏む必要がある。結果は別々のシートに出力され，それは表として提示できるような形にはなっていない。少なくとも 10 ステップくらいのマウスの操作が必要であり，わずらわしい。

R ならば，サンプルサイズが小さければ，変数 x と y（変数名は何でもよい）に直接２つの標本データを付値（代入）してから，var.test(x,y) をして，p が有意水準未満ならば t.test(x,y) でいいし，そうでなければ t.test(x,y,var.equal=T) とすればよい。データ範囲を何度も選ぶよりも，x とか y とかタイプする方が一般的にいえばずっと楽だと思う。もちろん，表形式のデータを読み込んで分析する関数だけ指定することもできる。

しかも，R では，結果を変数に代入して保存したり加工したりできる。xtable というライブラリ[*2]をインストールして読み込めば，結果を xtable( ) の括弧内に入れるだけで，HTML 形式や LaTeX 形式に変換できたりする。

美しい図を作るのも実に簡単で，しかもその図を PDF とか postscript とか png とか jpeg とか Windows 拡張メタファイル (emf) の形式で保存でき，他のソフトに容易に取り込める。例えば emf 形式で保存すれば，Microsoft PowerPoint や OpenOffice.org の Draw などの中で，ベクトルグラフィックスとして再編集できる。PDF 形式で出力してから，pTeX に入っている pdftops プログラムで-eps オプションをつけて変換すれば，Encapsulated Postscript 形式（EPS 形式）のファイルを作るのも容易である。

おそらく多くの日本人にとって最大の難点は，日本語が使えない（グラフィック表示は面倒な指定をすればできないこともないし，フォントを変更すればコンソール表示もできるし，データに日本語が入っていてもだいたい扱えるが，変数名としては使えないので，例えば１行目に日本語を使って変数名を打ってある表データは読み込めない）ことだろう。日本語変数名がなければ，Excel のデータならタブ区切りテキスト形式で保存すれば読みこめるし，foreign という標準ライブラリを使って SAS（Transport 形式）や SPSS（.sav 形式）や S-PLUS や Stata や EPIINFO や Octave のデータを読み込むことができる。

日本語による解説があまり出回っていない（英語が読めれば無料でもたくさん出回っている）のも，多くの日本人にとっては難点かもしれない。統計手法がわかって

[*2] R は拡張が楽なので世界中の研究者が追加のライブラリを作って，R 本体と同じような配布条件で公開しているものが山ほどある。

いも，関数名がわからないと実行できないから，統計手法から関数名を探せるようなサービスが欲しいところだろう。本書がその一助になれば幸いである。

A.2 R を使うための最初の１歩

A.2.1 インストール

- R は，現在のところ，Windows，MacOS，Linux，FreeBSD などの OS の上で利用可能である。
- Windows 版バージョン 1.7.0 のインストールは，CRAN（または会津大学にあるミラー）から rw1070.exe をダウンロードして実行するだけなので簡単である。rw1070.exe に該当するものは，R-1.6.0 では rw1060.exe だったし，それ以前は SetupR.EXE というファイル名だったが，いずれにせよインストールの仕方は同じで，ただ実行して，ダイアログに答えていくだけでいい。Windows 版では，デフォルトのインストール先ディレクトリは，`C:\Program Files\R\rw1070` などとバージョン番号が付き，アップグレードしても旧バージョンは自動的には消去されない。なお，手動で追加したライブラリはバージョンアップの際には継承されないので（バージョン依存性があるかもしれないから当然だが），それらのライブラリが zip 形式で公開されているならばそれもどこかに保存しておき，R 本体をバージョンアップした後で，Packages メニューの Install package from local zip file を選んで，保存しておいた追加ライブラリを１つずつ選択するという手順を踏む必要がある[*3]。
- FreeBSD 4.5R でも，ソースの tar ボールをダウンロードして展開し，そのディレクトリで`./configure` をやってから `make` するだけでコンパイルできるので，`su` して `make install` すればインストールが完了した。Vine Linux 2.6rc1 では g77 などのフォートランコンパイラを追加インストールする必要があったが，それさえしておけば FreeBSD 4.5R と同様にコンパイルやインストールができた。起動コマンド `R` で起動する Rconsole はテキストベースのシェルでも利用できるので，最尤推定などの時間がかかる計算はサーバでやらせると良い。おそらく他の UNIX 系 OS でも似たようなものであろう。

[*3] ネットワークにつながれた環境であれば，ライブラリは CRAN からダウンロードする方が安全かもしれない。MacOS では `install.packages()` 関数が使えないという話を聞いたことがあり，できるかどうか未確認だが，Windows2000，FreeBSD 4.5R，Vine Linux 2.6rc1 の環境では，`install.packages("xtable")` などとするだけで済んで簡単であった。

Debian など，いくつかの Linux ディストリビューションではバイナリパッケージが公開されているので，それをインストールする方が簡単かもしれない。

A.2.2 最も基本的な操作

- **起動**は，Windows ではデスクトップにできるアイコン（またはスタートメニューのプログラムの R にできるアイコン）をクリックするだけでいい。Windows2000 なら，コマンドラインでも `Rterm --no-save` として起動できる。Linux や FreeBSD のシェルを telnet や ssh で使う場合は，`R` と打てば良い。いずれの場合でも > というプロンプトが表示されて入力待ちになる。
- **終了**は，プロンプトに対して `q()` と打てば良い。コマンドラインパラメータとして`--no-save` などとつけて起動した場合以外はワークスペースを保存するかどうかの問い合わせがあるので，その回のセッションを記録しておきたいならば y を，そうでなければ n と打つ。ワークスペースを保存しすぎると.RData というファイルが大きくなって起動が遅くなるが，作業中は便利な機能である。
- 基本的に，関数にデータを与えて得られる結果を表示したり，変数に付値したりして使う。アルファベットとドットからなる文字列は変数になりうる。付値とは，ほぼ代入を意味する。例えば，`x` という変数に `3, 5, 7` という 3 つの値からなるベクトルを付値するには，`x <- c(3,5,7)` とする。これら 3 つの値の平均値を得るには，`mean()` という関数を使って，`mean(x)` とすれば良い。付値せずに関数だけを打てば結果を表示するが，もちろん関数の値を別の変数に付値することもできる。例えば `y <- mean(x)` として，`y` をまた別の計算に使うこともできる。変数の情報を見るには `str()` という関数が便利である。
- GUI 環境では，Help メニューから R Manual(html) を選べば，階層構造で説明を参照できる。関数へのインデックスもある。
- 関数の使い方を忘れたときは，`help(`**関数**`)` とか `help.search("`**キーワード**`")` で説明が得られる。
- `example(`**関数**`)` で関数の利用例が得られる。

A.2.3 R Commander を使ってみる

sem を初めとして様々なパッケージを開発し，”An R and S-Plus companion to applied regression”という優れた教科書も書いている McMaster 大学の John Fox 教授が，最近になって Rcmdr という，R をメニュー形式で操作するためのパッケージを発表した。

R-1.7.0 以上でないと使えないが，メニュー項目はテキストファイルで定義されているので書き換え可能である。Rcmdr のテキスト表示は tcl/tk で行われており，tcl/tk はバージョン 8.1 から国際化対応しているので，文字コードとして UTF-8 を使えば，メニューを日本語化することもできる[*4]。

インストールのためには，同じ John Fox が開発した”car”というパッケージを必要とするので，そちらを予め入れておかなくてはならない。手順としては，

```
> install.packages("car")
> install.packages("Rcmdr")
```

だけでインストールは完了する。この後で，ハーバード大学の林啓一さんが http://plaza.umin.ac.jp/~epi/Rcmdr-menus.txt として公開されている，ある程度日本語化したメニューで，R の library ディレクトリ内の Rcmdr/menus/Rcmdr-menus.txt を上書きし，library(Rcmdr) とすれば，日本語メニューで R を操作できる環境になる。あまり複雑な操作はメニューではできないが，入門としては便利だと思う。

A.3 R の参考書・web サイトなど

- R Project (http://www.r-project.org/)：プロジェクトのサイト
- R-announce Info Page (https://www.stat.math.ethz.ch/mailman/listinfo/r-announce/)：重要なお知らせが英語で流れる ML の情報ページ。
- CRAN (http://cran.r-project.org/)：プログラムやライブラリのダウンロード用サイト
- 会津大学のミラー

[*4] R-1.7.0 に含まれている tcl/tk はバージョン 8.4 である。

(ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN/index.html)：CRAN のミラーサイト。日本国内のミラーサイトは，公式にはここだけである。

- 公式入門書やマニュアルの日本語訳は，東京工業大学・間瀬教授のサイト (http://www.is.titech.ac.jp/~mase/R.html) で公開されている。その pdf 版は学芸大の森厚さんのサイト (http://buran.u-gakugei.ac.jp/~mori/LEARN/R/) からダウンロードできる。なお，2002 年末での最新版及び 2003 年 5 月に公開されたバージョン 1.7.0 の R-intro の暫定和訳については，間瀬教授のサイトからソースファイルをダウンロードし，日本語コードを SJIS に変えて Windows 版 pTeX と dvipdfm を使って個人的に pdf 化したものを，http://phi.ypu.jp/swtips/R-jp-docs/にも置いてある（フォントを内蔵していないのでファイルサイズは小さい)。
- 群馬大学社会情報学部・青木繁伸教授が R による統計処理 (http://aoki2.si.gunma-u.ac.jp/R/) という凄いページを作られている。R を使おうと思う方は必見である。
- 多摩大学・山本義郎助教授による R–統計解析とグラフィックスの環境 (http://datamining.tama.ac.jp/~yama/R/) と R 入門 (http://datamining.tama.ac.jp/~yama/R/Rintro.html) は，ちょっと古いバージョンでの解説だが丁寧に書かれていて役に立つと思う。
- R についての日本語メーリングリスト（R-jp）が，筑波大学のサーバで，岡田昌史さんによって運営されている。元々はドキュメント翻訳用のメーリングリストとしてスタートしたらしいが，R について日本語で議論されている ML としては唯一のものである。登録の仕方などの説明は http://epidemiology.md.tsukuba.ac.jp/~mokada/ml/R-jp.html にあり，過去に投稿されたメールのアーカイヴも公開されている。
- 山口県立大学の中澤のサイト内でも http://phi.ypu.jp/swtips/R.html として R についての使い方の tips や最新情報を公開している。本書のサポートもここで行う予定である。