目　次

第1章　統計学とは

自然科学，社会科学を問わず現実に起こる現象を解析する場合，あるいは実験や医療の現場においてデータに向き合う場合，そこには観測誤差や背景因子の多様性に伴う不確実性が存在する．例えば，病気の患者にある薬剤を投与したときの効果は，その薬剤の効能だけでなく，患者の体調や遺伝的背景，生活習慣など様々な因子の影響を受け，事前にその結果を知ることはできない．

しかしこれら不確実な事象には，個々の現象を取り上げれば不確実でも，全体として何らかの傾向，法則性が存在し，データに蓄積された過去の経験をもとに合理的な推論を行うことが可能な場合もある．そのための，データを収集し解析し解釈する方法論が「統計学」である．データに含まれる不確実性は確率的現象としてモデル化されるが，その確率的現象を扱う数学理論が「確率論」と呼ばれるものである．すなわち，不確実性や多様性を伴った事象に対して，観察されたデータを基に合理的な推論を行うための方法を提供するのが統計学であり，その理論的枠組みを数学的に支えるのが確率論，ということになる．

もし，自然現象あるいは社会現象において関連するすべての情報を得られれば，不確実性は除かれ，現在の状況と将来の予測を完全に理解できるはずである．しかし，現実には不確実な現象についてすべての情報を得ることは不可能であり，100%誤りのない判断をすることは困難である．ではどうするか．すべての情報を得ることは無理でも，部分的な情報を集め，それを基に全体を推論することであれば可能である．「不確実性」のないところに，統計学は必要ない．

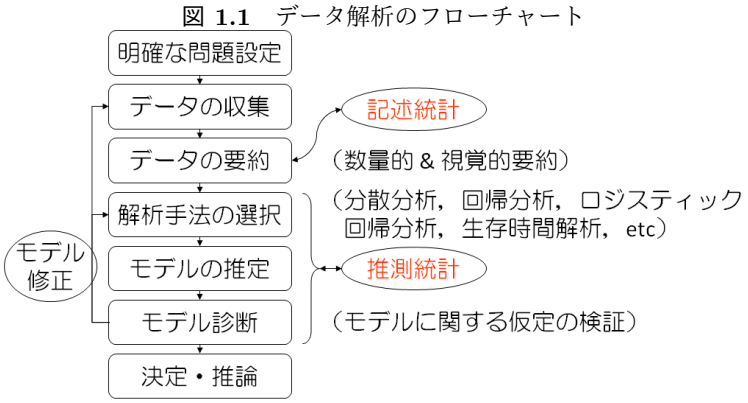
データの解析には，大規模な計算が必要になる場合もある．また，現代の統

計学ではデータを可視化 (visualization) し，視覚的にデータの特徴を捉えることが重要である．いずれの場合にもコンピューターを利用し統計解析ソフトによって解析を行うことが必要になる．本書では，インターネット上で公開されている統計解析ソフトウエア R を用いる．R は無料で利用できるフリーウエアであり，標準的な解析から最先端の統計手法まで兼ね備えた優れたソフトウエアである．R の入手とインストールの方法は，付録にまとめておいた．本書では今後 R を利用できることを前提に議論を進める．R の具体的な使用方法は，本書の中で順次触れていく．

1.1　データ解析の手順

実際のデータ解析において，興味の対象となる事象に関する情報をすべて集めることは不可能である．(例えば，血圧を下げる薬の効果を知るために，全ての高血圧患者に薬を投与して降圧効果を確かめるのは現実的ではない)　したがって可能な戦略は，興味の対象について部分的な（収集可能な）情報を集め，それを基に全体を推論することになる．データ解析の手順については様々な考え方がありうるが，図 1.1 のフローチャートのようにまとめることができる．

図 1.1　データ解析のフローチャート



明確な問題設定 データの解析を始めるためには，まずいかなる対象に対して，何を知りたいのか，そのためにはどのような情報をいかなる方法で集めればよいか明らかにする必要がある．解析の目的とデータの性質によって，適切な解析の手法も変わってくるため，問題を明確に設定することは解析の枠組みを決める大切なステップであり，本章で詳しく説明する．

データの収集 解析の目的が定まったら，次は目的に合わせてデータを収集する段階になる．このステップで大切なのは，解析対象から偏りなくデータを集めることである．一概に「偏りなく」データを集める，といっても簡単でないが，これも本章の中で解説する．

データの要約 データが収集されても，いきなり解析に移るわけではない．まず，データの特徴や傾向を大づかみに把握するためにデータの要約を行う．データを解析するための様々な手法では数学の理論が用いられるわけだが，数学的モデルには必ず何らかの前提条件が存在する．データを要約することで，解析しようとするデータがモデルの前提条件を満たしているか吟味することが出来る．

データの要約は，1) データの位置や散らばりの大きさなどデータを特徴付ける代表値を求める**数量的要約 (Numerical Summary)** と，2) 図による視覚的情報によって要約をする**視覚的要約 (Graphical Summary)** の二つに分けられる．数量的要約と視覚的要約の二つをあわせて**記述統計学 (Descriptive Statistics)** と呼ぶ．これについては，第 2 章で取り上げる．

解析手法の選択 データの要約によってデータの大まかな傾向をつかんだあと，解析目的に合わせた手法が選択される．統計学には，解析目的に従って**分散分析 (analysis of variance)**, **線型回帰分析 (linear regression analysis)**, **ロジスティック回帰分析 (logistic regression analysis)**, **生存時間解析 (survival time analysis)** など様々な解析手法（統計的モデル）が存在する．これら統計的モデルについては，後の章で詳述する．前述したとおり，解析手法にはその前提となる数学的条件があり，データがそれを満たさないようなモデルは選択できない．

モデルの推定 このステップで，いよいよデータから統計的モデルを推定する．統計的モデルとは解析対象となる要因の間の関係を数学的に記述するもの

であるが，その推定のための計算は，R などの統計解析ソフトによって行われる．

モデル診断 データにモデルを当てはめるといっても，推定したモデルに必要な前提条件が今手元にあるデータで実際に満たされているかは別問題である．データにモデルを当てはめた後，モデルの前提条件が満たされているか否かを事後的に確認することを**モデル診断 (Model diagnostics)** という．モデルの仮定が満たされていれないときは，解析結果を受け入れることはできず，前のステップに戻ってモデルを修正する必要がある．例えば，使用する解析モデルを変更する，データをほかの形に変換する，場合によっては，最初からデータを取り直すなどの対応をとることになる．

決定・推論 モデル診断によって，すべての仮定が満たされたことが確認されたら，最終的なモデルの結果を評価し，当初の解析目的にしたがって推測を行う．また解析結果は，わかりやすい形で報告されなければならない．統計解析パッケージ R は，図の表示に関しても柔軟で有用な機能を持っており，結果のプレゼンテーションにも有効である．

以上で示した，問題の設定から結果の表示までの一連のプロセスがデータ解析の手順であり，統計学が解決すべき課題といえる．ここでは一般的，抽象的に説明したが，以降の章では問題ごとに具体的にデータ解析の流れを説明する．

1.2 データ解析の問題設定

データを解析するとき最初にやるべきことは，図 1.1 の第一段階にある通り，そのデータを解析することで何を知りたいのか明確に問題を設定することである．一般に，次の 6 つの概念を定義することで，データ解析の目的を厳密に設定できるといわれている．

母集団 (Population) 解析対象となる個体の集合．もし世論調査で日本の政党の支持率を知ることが目的であれば，母集団は例えば「日本人全体の集合」が考えられる．病気の患者に対する薬剤の効果を知ることが解析目的なら，その病気に罹患したヒトの集合が母集団になる．

パラメター (Parameter) 母集団を特徴付ける未知の定数．世論調査を例にすれば政党支持率がパラメター，薬剤効果であれば薬の奏功率がパラメターになる．

標本 (Sample) 母集団から抽出された部分．標本が持つ部分的な情報を基に，母集団のパラメターを推測するのがデータ解析の目的である．

Sampling frame 標本として抽出されうる個体の集合．すなわち，標本となる可能性のある個体の集合．当然，データ解析に用いる情報は Sampling frame の中からしか得られない．

変数 (Variable) 母集団において，個体間で確率的に変わりうる特性，量．上の世論調査の例であれば，各有権者の各政党への態度（支持，不支持），薬剤効果の例であれば，薬を投与された各患者の応答性（効果あり，なし）が変数になる．

統計量 (Statistic) 変数の値に基づき，標本から計算される量．政党支持率を調べる世論調査であれば標本のデータから計算した支持率が，薬の奏効率であれば薬を投与された人のうち薬効のあった人の割合が統計量の値となる．統計量の値によって，パラメターを推測する．

以下，上で述べた 6 つの概念について，詳しく説明する．

1.2.1 母集団 (Population)

上で述べたとおり，**母集団**とはデータ解析の対象全体からなる集合である．例として日本における各政党の支持率を知ることを目的とした世論調査の場合，母集団は「日本人全体の集合」が考えられる，と述べた．しかし，よく考えてみると世論調査の母集団として他の定義を考えることも可能である．「日本人全体の集合」を世論調査の母集団とすると，そこには生まれたばかりの乳児も含まれる．果たして 0 歳児の政治的意向を知ることが，世論調査の目的なのだろうか．それでは，「日本における有権者の集合」を母集団としたらどうであろうか．その場合，日本で生活する投票権を持たない外国籍の方は母集団に含まれない．同じ社会で暮らす外国人の政治意識は，本当に世論調査の興味の対象外なのだろうか．あるいは，「日本に住む 18 歳以上の人」を母集団とすることも

可能である．しかしその場合，外国籍の成人は母集団に含まれるが投票権を持たない 17 歳以下の高校生は含まれない．次の世代を担う高校生全体の政治意識を知りたくはないだろうか．

このように考えると，母集団を定義することは，解析者が「知りたいことの範囲」を明らかにすることと同義である．世論調査で政党支持率を調べる場合，「日本人全体」「日本の有権者全体」「日本に住む 18 歳以上の人全体」など，どの集団における政党支持率を知りたいのか，解析者自身が選択しなければならない．また，母集団を定義するときは，「日本の有権者全体」に対する外国人のように，母集団に含まれない存在は何であるのか考えることも重要である．

1.2.2 パラメター (Parameter)

パラメターの定義は，母集団を特徴付ける未知の定数，とした．世論調査における政党の支持率は，母集団である (例えば) 日本の有権者の集合を特徴づける量である．パラメターの定義のポイントは二つ，「未知である」ことと「定数である」[1] ことである．もし政党の支持率が正確に知られていれば，世論調査を行う必要はない．支持率が未知であるからこそ，調査をするのである．またパラメターは定数であり，必ず実数値で表現できるものである．政党の支持率，薬の奏効率，すべて数字で表される概念である．

データ解析の目的をもっとも抽象的に定義すれば，それは「**母集団のパラメターについて何かを知ること**」すなわち，解析対象となる母集団の関心のあるパラメターについて推測をすることになる．世論調査を例にとれば，「母集団である日本の有権者集団における，政党支持率というパラメターについて推測すること」が世論調査の目的となる．

1.2.3 標本 (Sample)

母集団に不確実性が存在する場合，母集団全体について 100 %完全な情報を手に入れることは不可能である．しかし母集団全体の情報を収集することは困

[1] パラメターを定数ととらえるアプローチを統計学では頻度論 (frequentism) と呼び，本書ではこの立場をとる．ベイズ統計学 (Bayesian statistics) と呼ばれる別のアプローチでは，パラメターを確率的に変動するものとするが本書では扱わない．

難であっても，その母集団の一部を「標本（サンプル）」として抽出し，標本中の個体の情報を調べることは可能である．政党支持率を調べるための世論調査であれば，母集団に含まれる個人全体にインタビューすることは不可能でも，その中から 1,000 人なり 10,000 人なりを標本として抽出しインタビューすることは可能である．その結果，ある人は政党 A を「支持する」「支持しない」「わからない」などの回答を得ることができる．すなわち，標本として抽出された部分については，100%完全な情報が手に入るわけである．

1.2.4 Sampling frame

この Sampling frame（サンプリングフレーム）という概念については適切な日本語訳がないようであるが，標本として抽出されうる個体の集合，あるいは標本となる可能性のある個体の集合を指している．Sampling frame は，標本の抽出方法にも深い関係がある．再び世論調査を例にとると，世論調査の方法として思いつくのはどのような方法であろうか．

RDD (Random Digit Dialing) 報道機関の世論調査などでよく用いられる方法で，ランダムに発生させた電話番号に電話をかけ，出た相手にインタビューする方法である．（実際にはもう少し複雑な手順がある）

インターネット投票 インターネット上に投票サイトを立ち上げ，「支持する」「支持しない」といったラジオボタンを置く．サイトを訪れた人に，いずれかのボタンを選択してもらい一定期間後に集計する方法である．

街角インタビュー テレビ番組などでよく登場する方法で，街に出たレポーターが通行人に直接インタビューする方法で行われる世論調査である．

さて，このように世論調査といっても色々な方法があるが，それぞれの方法の Sampling frame（この場合，世論調査の対象として抽出される可能性のある人の集合）はどのような集団であろうか．

RDD (Random Digit Dialing) ランダムな番号に電話を掛ける方法で，標本として抽出される可能性があるのは「電話を持っている人」だけである．したがって，電話を持っていれば小学生でも RDD の Sampling frame に含ま

れるし，電話を持っていなかったり通信圏外にいる人は RDD の Sampling frame には含まれない．

インターネット投票 インターネット上のボタンを押すことで投票する方法の場合，サンプルとして抽出される最低限の条件は，インターネットにアクセスできることである．しかし，それだけで十分であろうか．インターネットにアクセスできる人の中で，実際にネット上の投票をしたことがある人は，むしろ少数派ではないだろうか．インターネット投票の Sampling frame は，「インターネットにアクセスでき，かつネット上で投票する意思のある人」とするべきである．

街角インタビュー 例えば，「月曜日の 15:00 に，東京の銀座で 100 人に聞きました」という街角インタビューを考えよう．この場合，普通の事務職の会社員やまじめな学生は Sampling frame に含まれない．なぜなら，営業で外出しているわけでもない会社員や学校で勉強している学生が，平日の昼過ぎに銀座を歩いているはずがないからである．また，東京近郊以外の地域，例えば農村部に暮らす人も含まれない．「月曜の午後，銀ブラをしている人」というのが，この場合の Sampling frame である．読者はどのような人たちを思い浮かべるであろうか．

Sampling frame を考える際に重要なのは，データ解析の目的が母集団の推測にあるのに対して，そのための情報は Sampling frame からしか得られない，という点である．従って，もし母集団の個体全てに標本となる可能性があるとは限らない場合，すなわち Sampling frame と母集団が一致しない場合，標本は母集団全体を代表せず解析に**偏り (bias)** が生じることになる．Sampling frame を考えるときは，Sampling frame に含まれない個体は何か，Sampling frame には含まれるが母集団には含まれない個体は何かを考えることが重要である．

以上の**標本**と **Sampling frame** を定義することで，データ解析において何処からどのようにして情報を集めてくるか，が明確になる．

1.2.5 変数 (Variable)

標本として抽出された個体は，その情報を詳しく調べられる．政党支持率を

調べるための世論調査であれば抽出された人がどの政党を支持しているのか，薬の奏効率を調べるための治験であれば投与された患者に薬効があるかを調べることになる．標本として抽出され観察される前は，母集団に含まれる人がどの政党を支持するか，その態度は未確定であり政党 A を支持する可能性も支持しない可能性もある．薬の奏効率を推測する場合，投与された薬が有効であるか無効であるかは診察前には未確定である．また，血圧，心拍数などの検査数値も測定前には未確定である．このように母集団の各個体において観察前には確率的に値が変わりうる量を**変数 (variable)** と呼んでいる．変数の値は，標本抽出前は未確定であるが，抽出後は観察され観測値が確定する．（世論調査であれば，インタビューが行われ各個人の支持政党が明らかになる）このように，実際に観察され確定した変数の観測値を，**データ (Data)** と呼ぶ．

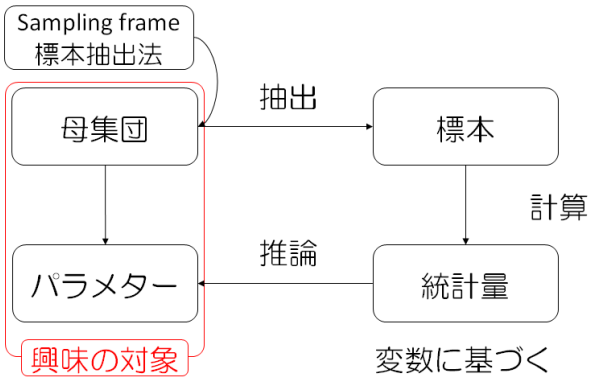
1.2.6 統計量 (Statistic)

標本から観察されたデータは，単なる数字や記号の羅列である．これを計算によって一つの値にまとめ，パラメターの推測に用いるのが統計量である．例えば，データから計算された標本平均は母集団平均を推定するための統計量の一つである．変数が観察されその値が確定する前は統計量の値も未確定であるが，データの値が確定した後は，例えば “56.9%” のように具体的な実数値となることに注意する．この具体的な統計量の観測値を用いて未知のパラメターに関する推論を行うことが，データ解析の最終目標である．

以上述べた，データ解析における問題設定の流れをまとめたのが図 1.2 である．

解析対象となる母集団のパラメターについて，推論を行うことがデータ解析の目的である．実際には母集団について 100%完全な情報は手に入らないので，その母集団の一部を標本として抽出し，標本中の個体の変数の値を調べたうえで，データから統計量を計算し，統計量の値からパラメターに関する推測を行う，というのがデータ解析の流れになる．

図 **1.2** 問題設定のダイアグラム



1.3 標本抽出法

前節で，図 1.1 のデータ解析のフローチャートの第一段階「明確な問題の設定」について説明した．本節では図 1.1 の第二段階「データの収集」，つまり母集団から標本を抽出する際の**標本抽出法 (Sampling scheme)** について説明する．よく使われる標本抽出法には，以下のようなものがある．

全数調査 (Census) 母集団中のすべての個体が，標本として抽出される．全数調査が実行可能であれば，母集団についての完全な情報を得ることが出来る．例えば国勢調査は，統計法に基づき日本に居住している全ての人及び世帯を対象として実施されるもので，全数調査を目指したものと言える．ただし，実際には経済的物理的理由で実行は困難な場合がほとんどであり，中には実行が不可能な場合もある．（例えば，缶詰工場の品質管理のため製品の缶を開封して検査する場合を考える．もし全数調査をしようとすれば，すべての缶を開封することになり，売り物がなくなってしまう．）

単純無作為抽出 (Simple Random Sampling, SRS) 標本は，母集団中の個体から等しい確率で抽出されるように設計された抽出法．例えば母集団が 10,000 の個体からなれば，それぞれの個体が標本として抽出される確率は（標本数）$\times \frac{1}{10000}$ になる．SRS の場合，母集団に含まれる個体はすべて標本として抽出される可能性を持つので，母集団と Sampling frame が一致する抽出法であると言える．無作為抽出には**多段抽出法**，**層別抽出法**など様々なヴァリエーションの発展型が存在するが，本書ではこれ以上触れない．

Voluntary Response Sampling 標本は，任意で調査に参加した個体のみから抽出されるような抽出法．第 1.2.4 節で取り上げた「インターネット投票」などが，これに当たる．Voluntary Response Sampling では調査に参加する意思を持つものだけが Sampling frame に入る，という点で，Sampling frame と母集団は一致しない．

Convenience Sampling「街頭調査」と訳されることもあるが，街頭に限らず母集団から恣意的に選ばれた部分からのみ標本を抽出する抽出法．第 1.2.4 節で触れた「街角インタビュー」などが典型的であるが，この場合 Sampling

frame は標本抽出が行われた時間や場所あるいは抽出の基準に依存し，やはり母集団には一致しない．

標本を抽出するときは，母集団全体の傾向を反映し，構造的な偏り（バイアス, bias）が起こらないように抽出方法を定める必要がある．そのためには，母集団中のすべての個体は標本として等しい確率で抽出されることが必要である．この観点から，Convenience sampling と Voluntary response sampling は，簡便であるが母集団と Sampling frame が一致せずバイアスを含む可能性がある．Census は母集団のすべての情報を集められるが，母集団が十分小さい場合を除いては実行は困難である．これに対し単純無作為抽出 (SRS) は，その定義からバイアスがなく望ましいものといえる．

実際の解析の場面では，最初から実験，調査をすべて計画することができず，すでに集められたデータを解析せざるを得ないこともある．そのときにも，解析の目的を明確にするために，前節の 6 つの概念を確認し，また標本抽出が適切に行われたかを確認するため，とくに母集団と Sampling frame が一致するか，標本の抽出は十分ランダムに行われたかを確認する必要がある．もし母集団と Sampling frame が一致しない場合，Sampling frame に含まれないものに対しては，いかなる推測も出来ない．解析の結果は，あくまで Sampling frame に含まれた部分に制約されることになる．

第2章　記述統計

第 1 章では，P. 7, 図 1.1「データ解析のフローチャート」のうち，「明確な問題の設定」と「データの収集」まで説明した．データが得られたあと，データ解析の第三段階は「データの要約」である．**データの要約 (summary)** の目的はデータの分布の形状と特徴を理解することであるが，その方法は 1) **数量的要約 (numerical summary)** と，2) **視覚的要約 (graphical summary)** に分けられる．これらを総称して，**記述統計学 (descriptive statistics)** と呼ぶ．

記述統計の内容について説明する前に，なぜ記述統計によってデータの特徴を理解することが重要なのか，今一度考えておく．

2.1　記述統計の重要性

記述統計はデータを要約し，データの持つ全体的な特徴，傾向を表現する．特にデータの分布の**位置 (location)** と，分布の**散らばり (分散，variance)** ，およびその**形状 (shape)** の要約を重視する．では，なぜこのようなデータの要約が必要なのか，その理由として以下のものが考えられる．

母集団と解析データの異同を示すため　第 1 章で見た通り，データ解析の目的は母集団のパラメターについて推論を行うことにある．その際，推論に用いるデータの分布が母集団とどの程度同じでどの程度異なるのか，事前に明らかにする必要がある．このため臨床医学・生物学系の論文では，多く

の場合論文の最初でデータの記述統計が報告される．

適切な解析手段を選択するため 統計学では，同じ解析目的に対して複数の解析手法が存在する場合がある．例えば分布の中心の位置を推定する場合でも，分布の形状が左右対称なのか，左右いずれかに歪んでいるのか，あるいはデータの中に質の異なるサブグループが存在するのか，状況によって異なる手法を用いる必要がある．適切な解析方法を選択するためには，データの特徴を理解することが必要である．

データが誤りなく公正に収集されていることを示すため 例えば二つの群を比較する比較対照実験の場合，サンプルが二群にランダムに割り付けられ，対照のための条件以外の背景因子には極端な違いがないことが理想である．比較群と対照群のデータの分布に大きな違いがなければ，偏りなくデータが収集されたことの傍証になる．また，本来正の値をとるはずのデータに負の値が混じっている，あるいは異常に欠測値が多いなど，データ収集の際の誤りを思わせる要素がないことを示すのも，記述統計の役割である．

2.2 数量的なデータの要約

数量的なデータの要約の目的は，データから分布の形状を特徴づける**統計量** (p. 14) を計算し，データの大まかな傾向を理解することである．分布を特徴づける統計量には，データの**位置 (中心, Location)** を表す量や，データの**変動や散らばり (Variability, Dispersion)** を表す量などがある．

2.2.1 データの位置

2.2.1.1 平均 (Mean)

データの位置（中心）を表す統計量として，もっともよく使われるのが**平均 (Mean)** である．n 個の観測値 $x_1, x_2, \ldots, x_n$ が与えられたとき，平均は以下の式で定義される．

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

ただし，$\sum_{i=1}^{n}$ は**総和記号**と呼ばれ，$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$ を意味する．（総和記号と似た記号で，**総積記号** $\prod_{i=1}^{n} x_i = \prod_{i=1}^{n} x_i = x_1 \times \cdots \times x_n$ というものもある）ここでは，実際に平均を計算するために統計解析ソフトウエア R を使ってみよう．付録「R のインストールと使い方」に従って，コンピューターに R をインストールする．R を起動したのち，平均を計算するため以下のようにサンプルデータを入力する．ここで使われる c() は “combine” を意味する R コマンドで，カッコの中をまとめて一つのベクトルを作る．

```
> x <- c(0.684, 1.406, -0.663, -0.124, 0.849, -0.888, -0.492,
+ -0.044, -0.188, 0.536, 2.063, -1.379, 0.920, 0.453, 1.239)
```

前記のプログラム例の最初に現れる > は R のコマンドプロンプトであり，二行目先頭の+は入力が完了していない場合のプロンプトである．(一行目で c() コマンドの入力途中で改行したため，継続行の先頭に+が現れている)

まず定義に従って，ベクトル x の和を求め，それをデータの個数すなわちベクトルの長さで割る．数値型ベクトルの和を求めるコマンドは sum()，ベクトルの長さを求めるコマンドは length() である．

```
> sum(x)
[1] 4.372
> length(x)
[1] 15
> sum(x)/length(x)
[1] 0.2914667
```

平均を一度に計算するための R コマンドは，mean() である．

```
> mean(x)
[1] 0.2914667
```

確かに mean() が計算した結果は，定義に従った計算結果に一致している．

2.2.1.2 中央値 (Median)

平均は標本一つ一つの値に注目し，等しく $1/n$ の重みをかけて足し合わせたものである．これに対して，標本の大小の順序に注目し，ちょうど真ん中に来た値でデータの中心を表す代表値を**中央値 (Median)** と言う．いま，大きさ n の標本 $(x_1, x_2, \ldots, x_n)$ が与えられたとする．これら n 個の値を大きさの順に並べなおして $x_{(1)} \leq x_{(2)} \leq, \ldots, \leq x_{(n)}$ としたものを**順序統計量 (order statistic)** という．$x_{(1)}$ は標本 $(x_1, x_2, \ldots, x_n)$ の最小値，$x_{(n)}$ は最大値になる．順序統計量の概念を用いて，中央値 (Median) は以下のように定義される．

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & n \text{ が奇数} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & n \text{ が偶数} \end{cases} \tag{2.2}$$

$n = 3$ ならば，n が奇数であるから $\tilde{x} = x_{((3+1)/2)} = x_{(2)}$，$n = 4$ ならば，n が偶数であるから $\tilde{x} = (x_{(n/2)} + x_{(n/2+1)})/2 = (x_{(4/2)} + x_{(4/2+1)})/2 = (x_{(2)} + x_{(3)})/2$. つまり中央値とは，標本を大きさ順に並べたとき「真ん中」にくる値（n：奇数），もしくはもっとも真ん中に近い二つの値の平均（n：偶数）である．R で中央値を計算するためのコマンドは，`median()` になる．

```
> median(x)
[1] 0.453
```

ここで考えたのは，標本から求めた平均と中央値である．しかし標本が抽出されてくる元となった母集団についても平均と中央値を考えることができる．例えば，母集団が N 個の実数からなる集合であった場合，

$$\text{母集団平均} = \mu = (\text{母集団の } N \text{ 個の値の和})/N,$$
$$\text{母集団中央値} = \tilde{\mu} = \text{母集団の } N \text{ 個の値の，50\%順位の値.}$$

上に記したとおり，母集団平均，母集団中央値は，標本平均 $\bar{x}$，標本中央値 $\tilde{x}$ と区別して，それぞれ $\mu, \tilde{\mu}$ と記す．$\mu, \tilde{\mu}$ が母集団のパラメターであるとき，$\bar{x}, \tilde{x}$ はそれぞれ $\mu, \tilde{\mu}$ を推定するための統計量になる．このように，何らかのパラメターを推定するための統計量を**推定量 (Estimator)**，データから計算された推

定量の観測値を**推定値 (Estimate)** と呼ぶ．推定量，推定値については，第 4 章 4.1 節で，詳しく述べる．

平均値と中央値の関係

平均値は，分布の中心を推定する統計量として最もよく用いられる．一方で，標本の中に極端に大きい若しくは小さい外れ値があるとき，平均値はそれら外れ値に強く影響されることも知られている．

例題 2.1. いま *1000* 軒の家計の年収を調査するとする．簡単にするため，すべての家計の年収は *500* 万円であるとする．このとき家計の年収の平均と中央値を *R* で求めると，以下のように求められる．

```
> (5000000*1000)/1000   # 標本平均
[1] 5e+06
> (5000000+5000000)/2   # 標本中央値
[1] 5e+06
```

計算結果の *"5e+06"* は，$5 \times 10^6 = 5,000,000$ の意である．平均も中央値もどちらも *500* 万円で等しいが，定義により計算の仕方が異なることに注意する．さてここで，集団に年収 *100* 億円のお金持ちが一人加わったとする．このときの平均は

```
> (5000000*1000+10000000000)/1001
[1] 14985015
```

であるから，*1001* 軒の家計の平均年収は *1498* 万 *5015* 円になる．他方，$n = 1001$ が奇数であるから，$\tilde{x} = x_{((1001+1)/2)} = x_{(501)} = 5000000$ となり年収の中央値は *500* 万円のままである．さて，いま *1001* 軒の家計のうち *1000* 軒の家計の年収が *500* 万円であるとき，平均 *1498* 万 *5015* 円と中央値 *500* 万円では，どちらが分布の中心としてふさわしいだろうか？

上の例で，標本に極端に大きな値が加わったとき平均が大きく変動したのは，明らかに平均の欠点である．これに対して，中央値は外れ値の影響を受けなかっ

た．これは中央値の利点であり，データに多くの**外れ値（Outlier, 極端に大きい若しくは小さい値）**が含まれるときは中央値を用いるべきである．

一方で，平均は数学的に取り扱いやすく理論的に多くの利点がある．例えば，$x_2, \ldots, x_n$ を任意の値に固定したとき，平均 $\bar{x}$ を x_1 で微分した場合の導関数の値は以下のように与えられる．

$$\frac{\partial}{\partial x_1}\bar{x} = \frac{\partial}{\partial x_1}\frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\left(\frac{d}{dx_1}x_1\right) = \frac{1}{n}.$$

しかし，中央値 $\tilde{x}$ に対して x_1 で微分した場合の導関数は明示的に定義することはできない．つまり，中央値の定義 (2.2) そのものは極めて単純であるにもかかわらず，中央値に対しては微分，積分といった最も基本的な数学的方法すら定義できない，数学的に極めて扱いづらい概念であるということができる．

また，中央値の定義 (2.2) から標本数 n が奇数の場合 $\tilde{x} = x_{(n+1)/2}$，n が偶数の場合 $(x_{(n/2)} + x_{(n/2)+1})/2$ であり高々二つのサンプルの値しか用いられない．サンプル数 n がどれだけ大きかろうとも，$(n-2)$ 個のサンプルの値は使われずに捨てられる，という意味で中央値は情報の損失が極めて大きい概念であるともいえる．結局，平均と中央値の使い分けは外れ値の有無によって判断することになるが，この点については後でもう一度触れる（p. 37）．

2.2.1.3 パーセント点 (Percentile)

中央値は，その定義から標本を小さいほうから $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ と並べなおしたとき，50%の順位にある値である．この考え方を拡張し，データの小さいほうから $100 \times k$%の順位にある値を k **パーセント点 (百分位点，k-th percentile)** あるいは**標本パーセント点 (sample percentile)** という．特に 25 パーセント点 (25-th percentile) を**第 1 四分位点 (first quartile)**，75 パーセント点 (75-th percentile) を**第 3 四分位点 (third quartile)** という．50 パーセント点=**第 2 四分位点 (second quartile)** は中央値そのものになる．これらパーセント点，四分位点も順序統計量を基に定義されているので，中央値と同様外れ値に対して影響されにくい性質を持っている．四分位点，パーセント点を計算するための R コマンドは，`quantile()` になる．x の四分位点を求めるときは，ただ x を `quantile()` に代入する．k パーセント点を求めるには，`quantile()`

に probs=0.01*k オプションを与える．

```
> quantile(x)
     0%     25%     50%     75%    100%
-1.3790 -0.3400  0.4530  0.8845  2.0630
> quantile(x, prob=0.2) # x の 20%パーセント点
    20%
-0.5262
```

(最小値, 第1四分位点, 第2四分位点, 第3四分位点, 最大値) の組を **Five numbers summary** と呼び，quantile() コマンドで得られる．さらに summary() コマンドを用いることで，Five numbers summary に平均を加えた組を得ることが出来る．

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.3790 -0.3400  0.4530  0.2915  0.8845  2.0630
```

通常，Five numbers summary, あるいはそれに平均を付け加えたものまで求めれば，分布の中心（位置）に関する数量的な要約としては十分である．

2.2.2 データの散らばり

平均などデータの中心を表す代表値は，データが分布する位置を示す．分布の形状を特徴付けるもう一つの重要な概念に，データの**変動 (variability)** や**散らばり (dispersion)** がある．例えばデータが二つの群に分けられるとする．それぞれの群の平均に意味のある差があるか検討するとき，各群のデータの散らばりが相対的に大きければ平均のわずかな差はノイズに埋もれてしまう．一方，平均の差に比べてデータの散らばりが小さければ，よりたやすく平均の差を見いだすことが出来る．データの散らばりの大きさを測る尺度には，以下のような概念がある．

2.2.2.1 分散 (variance) と標準偏差 (standard deviation)

データの散らばりを測る尺度として最もよく用いられるのは，**分散 (vari-**

ance) とその平方根である**標準偏差 (standard deviation)** である．n 個の観測値 $x_1, x_2, \ldots, x_n$ が与えられたとき，分散 s^2 と標準偏差 s は以下のように定義される．

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.3}$$

$$s = \sqrt{s^2}. \tag{2.4}$$

分散は，データ全体の散らばりの大きさを測るため，まず個々の観測値 x_i とデータの分布の中心である標本平均 $\bar{x}$ の間の距離を偏差の二乗 $(x_i - \bar{x})^2$ で測る．この，個々の x_i とデータの中心である $\bar{x}$ との二乗距離をデータ全体で平均したものが分散 s^2 である．（実際には，分散は偏差の二乗和を $(n-1)$ で割ったもの．なぜ $(n-1)$ で割るかは，命題 4.1，(p. 97) で触れる）x_i が $\bar{x}$ の周りに集中して分布していれば，$(x_i - \bar{x})^2$ の値は小さく分散 s^2 も小さい．逆にデータの散らばりが大きければ，分散 s^2 も大きな値をとる．標準偏差 s は分散 s^2 の平方根である．分散を計算する R のコマンドは `var()`，標準偏差を求めるコマンドは `sd()` である．

```
> var(x)
[1] 0.8889764
> sd(x)
[1] 0.9428555
```

R で平方根を求めるコマンドは `sqrt()` である．これを用いて，標準偏差の定義を確認する．

```
> sqrt(var(x))
[1] 0.9428555
```

確かに，分散 `var(x)` の平方根は標準偏差 `sd(x)` の値に一致している．

2.2.2.2 四分位点間距離 (Inter Quartile Range, IQR)

データの散らばりの尺度として，分散と標準偏差は代表的なものである．し

かし分散の定義には平均 $\bar{x}$ が用いられ，各 x_i の $\bar{x}$ からの散らばりの大きさは両者の二乗距離 $(x_i - \bar{x})^2$ で測られる．「平均値と中央値の関係」で述べたとおり，$\bar{x}$ は極端に大きいあるいは小さい外れ値に対して敏感 (sensitive) に反応し，$(x_i - \bar{x})^2$ もまた外れ値に対して極端に大きな値をとりがちである．したがって分散 s^2（及びその平方根である標準偏差 s）もまた，外れ値に対して影響されやすいという欠点を持っている．外れ値 (outlier) に対して影響されにくい（「頑健な」あるいは「ロバスト (robust)」な）散らばりの尺度として，以下に定義する**四分位点間距離 (Inter Quartile Range, IQR)** がある．

$$IQR = \text{第 3 四分位点} - \text{第 1 四分位点}. \tag{2.5}$$

第 1 四分位点，第 3 四分位点ともに順序統計量を基に定義されるので，IQR は外れ値に対して影響されにくい尺度になっている．なお，第 1 四分位点から第 3 四分位点までの範囲を**四分位範囲**と呼ぶが，これも Inter Quartile Range, IQR と称する．IQR を求める R のコマンドは，`IQR()` である．

```
> IQR(x)
[1] 1.2245
> quantile(x, 0.75) - quantile(x, 0.25)
   75%
1.2245
```

上の例の 3 行目では，第 3 四分位点の値 `quantile(x, 0.75)` から第 1 四分位点の値 `quantile(x, 0.25)` を引いた値が，確かに `IQR(x)` の値 1.2245 に一致することを確認している．なお，その下の出力に “75%” とあるが，これは第 3 四分位点が 75%点であることを受け継いだもので，気にする必要はない．四分位範囲を表記する際は，

$$\text{IQR} : (-0.3400, 0.8845)$$

のように第 1 四分位点，第 3 四分位点を明示して範囲を記す方法もよく用いられる．R では，大文字と小文字は異なるオブジェクトとして扱われる．`iqr()` という表現は，四分位点間距離を求める R のコマンドとしてまったく意味がない．

2.2.3 標準偏差と標準誤差と IQR

データ解析においては，以上述べてきた平均，分散などを用いてデータの数量的要約を行うが，論文などで実際に要約を行う際は，いくつか決まったやり方で要約されることが多い．論文の中では，しばしば次のような表現を見かける．

> Continuous variables were expressed as mean ± SD, mean ± SE or median (interquartile range), as appropriate.

これは，「連続変数（実数値であらわされる変数）は，平均 ± 標準偏差，平均 ± 標準誤差，あるいは中央値（四分位範囲）の，いずれか適当なもので表現される」という意味である．まず，新しい概念である**標準誤差 (standard error, SE)** を定義する．いま母集団平均にならい，母集団が N 個の実数からなる集合であった場合の母集団分散と母集団標準偏差を定める．

$$\text{母集団分散} = \sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2/N,$$
$$\text{母集団標準偏差} = \sigma = \sqrt{\sigma^2}.$$

ただし，μ は母集団平均．このとき，標本平均 $\bar{X}$ の標準偏差は $\sigma/\sqrt{n}$ で与えられる．（命題 3.17, (p. 86) 参照）ただし，n は標本数である． すなわち，同じ母集団から何度もサンプル収集を行いその都度標本平均を計算したとき，標本平均の標準偏差 $\sigma/\sqrt{n}$ はデータ全体の標準偏差 σ よりずっと小さくなる．この「標本平均の標準偏差」の推定値 $s/\sqrt{n}$ が標準誤差である．（より正確には，何らかの統計量の標準偏差を標準誤差と言う．特に言及なしに標準誤差というときは，通常上に示したように標本平均の標準誤差 (Standard Error of Mean, SEM) を意味する．）このとき，平均 ± 標準偏差，平均 ± 標準誤差，あるいは中央値（四分位範囲）の使い分けは以下のようになる．

平均 ± 標準偏差 標準偏差は観測データ全体の散らばりの大きさを表す．データが後述する正規分布（第 3.3.4 節 (p. 60) 参照）という確率分布に従う場合は，平均 ± 標準偏差の範囲にデータの 60〜70%が分布していると想定できる．「平均 ± 標準偏差」という表現は，観測データ全体の位置と散らば

りを表現するのに適している．

平均 ± 標準誤差 この場合データが正規分布に従うのであれば，平均 ± 標準誤差の範囲に繰り返し求めた標本平均の 60〜70%が分布している．標本平均は母集団平均を推定するための推定量であるから，SE は標本平均による母集団平均の推定の精確さ (precision) を測っていることになる．

中央値（四分位範囲） 中央値 (Median) を中心に，IQR の範囲にデータ全体の 50%が分布している．観測データ全体の散らばりを記述している点で，平均 ± 標準偏差に対応する概念である．四分位範囲（第一四分位点から第三四分位点までの範囲）は，データの散らばる範囲を逸脱しない点に注意する．

薬剤を投与した群と投与していない群のように二群以上を比較する場合は，平均の推定と比較を問題にしているので「平均 ± 標準誤差」が適切である．データ全体の散らばりの範囲に興味があれば，平均 ± 標準偏差も可能である．平均 ± 標準偏差を用いるときの注意点として，データの分布が歪んでいるとき平均 ± 標準偏差は以下の例題のように不合理な値をとる可能性があることが挙げられる．

例題 2.2. 脳性ナトリウム利尿ペプチド *(brain natriuretic peptide, BNP)* は心臓から分泌されるホルモンであり，心不全のマーカーとしてよく用いられる．*BNP* は必ず正の値をとり，その分布は右に長い裾を持つ（右に歪んでいる）ことが知られている．例えばある心不全患者集団の *BNP* の平均*=195.9*，中央値*=104.0*，標準偏差*=292.4*，四分位範囲*=(41.3, 238.0)* だとする．このとき *BNP* の値の分布を平均 ± 標準偏差 $\Longleftrightarrow$ 195.9 ± 292.4 と表してしまうと，*BNP* の値の下限が $195.9 - 292.4 = -96.5$ 以下であると主張していることになってしまい，*BNP* が正の値を持つことに矛盾する．このようなときは，中央値（四分位範囲）$\Longleftrightarrow$ *104.0 (41.3, 238.0)* のように表記すれば，四分位範囲は必ずデータの取り得る範囲に値をとるので矛盾がない．

2.3 視覚的なデータの要約

数量的なデータの要約によって分布を特徴づける様々な数値情報を得られる

が，それによって分布の形状がすべて理解できるとは限らない．分布の形状を把握するには，グラフを用いたデータの要約によって視覚的に分布をとらえることが有用である．本節では，最も基本的な視覚的要約として**ヒストグラム**と**ボックスプロット**を取り上げる．

2.3.1 ヒストグラム (Histogram)

観測値が得られたとき，標本の範囲 (Range) をいくつかの隣接する区間 (subinterval) に分割する．この区間を**階級 (Class or Bin)** といい，各階級の上限と下限の中間値を**階級値**という．各階級の中に値をとる観測値の個数を**度数 (Frequency)**，標本の総数を 1 としたときの各階級の度数の割合 (度数/標本数) を**相対度数 (Relative Frequency)** という．このとき横軸に観測値をとり，縦軸に度数もしくは相対度数をとった棒グラフを**ヒストグラム (Histogram)** という．もし階級の幅がそれぞれ異なるときは，各階級の上の棒の面積が度数あるいは相対度数と比例するように，(度数あるいは相対度数) = (階級の幅) × (棒の高さ) となるように棒グラフの高さを決める．

例題 2.3 (Low Infant Birth Weight データ)．本章では，例として *1986* 年にアメリカ，マサチューセッツ州スプリングフィールドの *Baystate Medical Center* で生まれた *189* 人の幼児に関する，*Low Infant Birth Weight* データ[6)12)] を用いる．*Low Infant Birth Weight* データは *R* の *MASS* パッケージに *birthwt* オブジェクトとして保存されている．

```
> library(MASS)
> data(birthwt)
> head(birthwt, n=2)
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
```

`library()` コマンドは，R の拡張パッケージをロードするコマンド（付録 .6.5 も参照）`data()` コマンドはデータオブジェクトをロードする．`head()` は，ベクトル，行列，配列，データフレームの最初の数行を表示する（デフォルトは

6 行．n オプションで行数を指定する）birthwt オブジェクトに含まれる変数の説明は，以下の通りである．

low	出生体重が 2.5kg 未満であるか否かのダミー変数．(0/1)
age	母親の年齢
lwt	最終月経期間における母親の体重 (ポンド)
race	母親の人種 (1=白人, 2=黒人, 3=その他)
smoke	妊娠期間中の喫煙の有無 (0/1)
ptl	過去の早産の数
ht	高血圧の有無 (0/1)
ui	子宮炎症の有無 (0/1)
ftv	妊娠後最初の 3 ヶ月に医師の診断を受けた回数
bwt	出生体重 (グラム)

ここで，birthwt オブジェクトの変数の扱いに注意する．race は人種の別を表すが，数値型データであり，これを因子型に型変換する必要がある．（R の変数の型については，付録 .3.2 を参照）

```
> birthwt$race <- factor(birthwt$race, labels=c("white", "black", "other"))
# 因子型に変換し，水準のラベルを 1，2，3 から，white，black，other に変更する．
```

なお，low のように 0 あるいは 1 の二通りの値しかとらない変数を，**ダミー変数 (dummy variable)** と呼ぶ．また，上のプログラムに “# 因子型に変換し，...” という表現があるが，R では “#” 以降に書かれた文はコメントとして無視される．factor() コマンドは，ベクトルを因子型に変換する．label オプションにより，変数 race の水準のラベルを 1, 2, 3 から，white, black, other に変更する．（付録 p. 150 も参照）

ヒストグラムを描くための R のコマンドは，hist() である．R のデフォルトでは，縦軸に度数をとったヒストグラムが描かれる（図 2.1 左）．縦軸に相対頻度をとったヒストグラムを描くときは，hist() コマンドに freq=FALSE オプションか prob=TRUE オプションのいずれかを与える（図 2.1 右：縦軸のラベルが “Density” となっているが，「相対頻度」と読み替えてほしい．）．

```
> attach(birthwt) # データフレーム（この場合は birthwt）の要素を，直接参照できるようにする
> hist(lwt)
> hist(lwt, freq=FALSE)
```

図 2.1

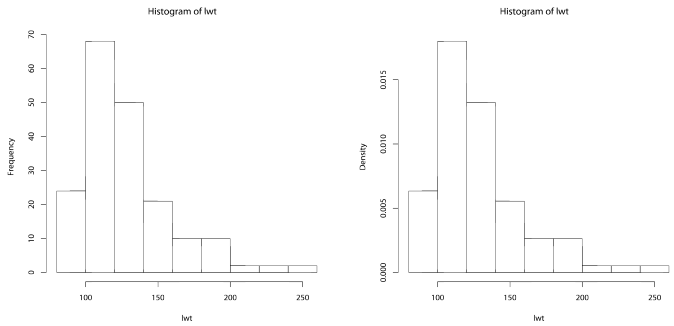


図 2.1 から，母親の体重 (lwt) の分布はサンプルの多くが比較的小さい値をとる一方，数は少ないが大きな値を持つものもある右裾の長い形状をしていることが分かる．また，縦軸に頻度をとった場合（図 2.1 左）と，縦軸に相対頻度をとった場合（図 2.1 左）では，ヒストグラムの形状は全く同じだが，縦軸の値が異なっていることが分かる．

ヒストグラムの階級の数 k を決めるための方法はいくつか提案されているが，まだ決定的なものはないようである．階級の数 k を決めるための古典的な方法として，以下の「Sturges の公式」[1] が知られている．

$$k \approx 1 + \log_2 n$$

ただし，k は階級の数，n は標本数である．

[1] Scott, David W., *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons, New York; Chichester, 2 edition 1992, pp. 48-49

ヒストグラムの形状:

ヒストグラムはデータの分布の形状について，わかりやすい要約を与える．図 2.2 に様々な形のヒストグラムを示した．図 2.2 の左上は**単峰型 (unimodal)** かつ**対称 (symmetric)** な分布で，ただひとつのピークを持つ．これに対して右上は，**二峰型 (bimodal)** の分布で二つのピークを持つ．さらに多くのピークを持つ分布は**多峰型 (multimodal)** と呼ばれるが，複数のピークを持つ分布は，データの中にそれぞれのピークに対応して質の異なるサブグループが含まれる場合があり注意が必要である．

分布の対称性に注目すると，左上のように左右対称な分布がある一方で，左右非対称な分布も存在する．分布の右裾が長い左下のような分布は**右にゆがんだ分布，right skewed or positively skewed** と呼ばれる．逆に分布の左裾が長い右下のような分布は**左にゆがんだ分布，left skewed or negatively skewed** と呼ばれる．

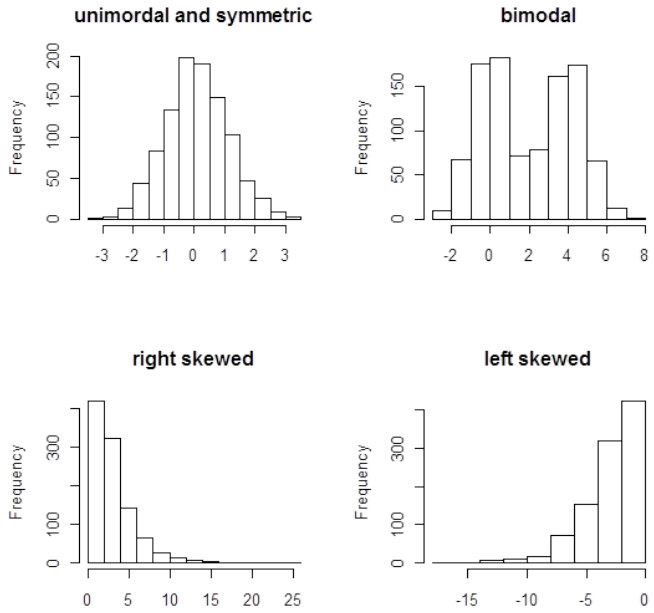
2.3.2 ボックスプロット (Box-plot)

ヒストグラムは，分布の全般的な形状を図示するのには適しているが，データの位置や広がりを示す統計量を明示することは出来ない．また，平均や分散の値に大きな影響を与える「外れ値 (Outlier)」を示すことも出来ない．これらの点に対応する方法として，以下に定義する**ボックスプロット (Box-plot)** がある．

定義: いま，IQR をデータの四分位点間距離 (Inter Quartile Range) とする．（第 1 四分位点 - $1.5 \times IQR$）より小さい観測値，もしくは（第 3 四分位点 + $1.5 \times IQR$）より大きい観測値を**外れ値 (Outlier)** という．外れ値は四分位点から $3 \times IQR$ 以上離れているとき **extreme outlier**，そうでなければ **mild outlier** であるという．

定義: ボックスプロットは以下の手順で描かれる．1)　縦軸に変数値をとり，下限が第 1 四分位点，上限が第 3 四分位点となる長方形を描く．2)　長方形の中の中央値に当たる位置に線を描く．3)　長方形の上下辺から観測値の最大値，最小値まで線を描く．この線のことを「ひげ (whisker)」と呼ぶ．ただし，データ

図 2.2 ヒストグラムの形状



の中に外れ値があるときは，長方形の上下辺から（第 1 四分位点 - 1.5 × IQR）および（第 3 四分位点 + 1.5 × IQR）まで「ひげ」を描き，外れ値は点で表す（図 2.3 左）．ボックスプロットを描くための R のコマンドは `boxplot()` である．ここでは，母親の年齢 age のボックスプロットを描いている（図 2.3 右）．`boxplot()` コマンドには，タイトルを表すためのオプション `main` と，y 軸のラベルを表すオプション `ylab` を付けてある．

```
> boxplot(age, main="age", ylab="years")
```

図 2.3

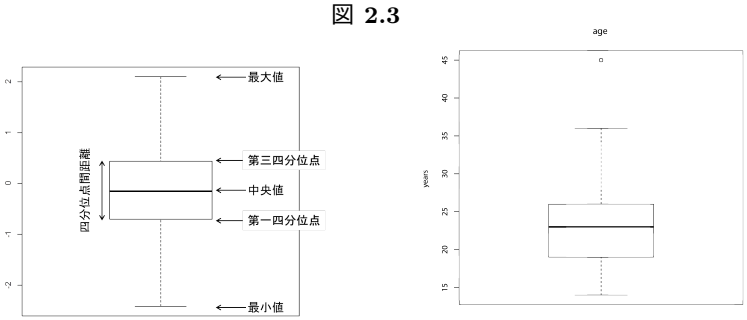


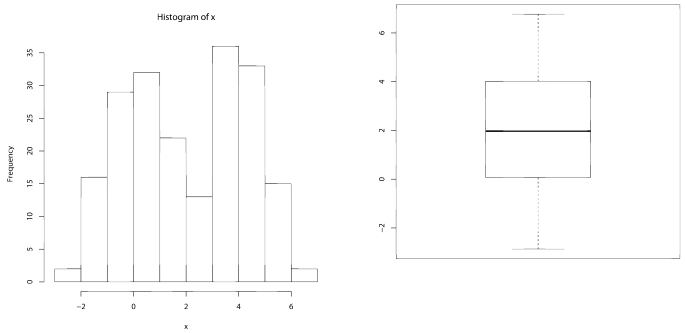
図 2.3 右の age のボックスプロットを見ると，age=45 のあたりに特に大きい外れ値があることがわかる．また，ボックスプロットを見るときのチェックポイントとして，最小値（髭の下端）から中央値までの距離と，中央値から最大値（髭の上端）までの距離を比較することがある．図 2.3 右の場合，最小値から中央値までの距離に比べ，中央値から最大値までの距離のほうが大きいことから，分布が右にゆがんでいる（右に長い裾を引いている）ことがわかる．

2.3.3 ヒストグラムとボックスプロット

本節では，ヒストグラムとボックスプロットという 2 種類の図を紹介した．この二つがどのような特徴を持つかを示すため，以下の例を考える．

二峰型データ 図 2.4 は同一の二峰型データ（ピークを二つ持つデータ）のヒストグラムとボックスプロットである．ヒストグラムは明らかに二峰型の特徴を示しているが，ボックスプロットからは二つのピークを持つという分布の特徴を捉えられていない．このことから，ヒストグラムはデータの分布の**全体的な特徴**をとらえるのに適している，といえる．

図 2.4

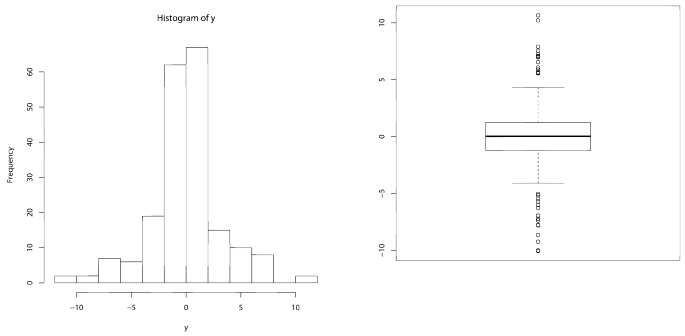


裾の重いデータ 図 2.5 は，極端に大きいあるいは小さい外れ値を多数含んだ所謂裾の重いデータのヒストグラムとボックスプロットである．

ボックスプロットは，その定義から（第 1 四分位点 - $1.5 \times IQR$）より小さい，もしくは（第 3 四分位点 + $1.5 \times IQR$）より大きいデータを外れ値として表示するため，分布の**裾の部分にある外れ値**をとらえるのに適している．一方ヒストグラムの形状は単峰型で左右対称な分布のヒストグラムに似ており，分布の裾が重いという特徴を十分に捉えていない．

最後に，本節で検討した「ピークが二つある」とか「データの裾が重い」といったデータの形状に関する情報は，平均や分散といった数量的なデータの要約ではとらえることが出来ない，という点を強調しておく．例えば，データの中心を推定するのに「平均」と「中央値」のどちらを使うのか，という判断には，データに極端に大きいもしくは小さい外れ値が含まれているか，データの

図 2.5



分布は左右いずれかの方向に歪んでいるかどうか，と言った分布の形状に関する情報が必要であるが，それはグラフを用いた視覚的なデータの要約によってしか得られないものである．他方，視覚的なデータの解釈は多分に主観的であるから，数値による客観的な要約で補完する必要がある．

結局，データの特徴を十分に捉えるには，数量的な要約と視覚的な要約が共に必要であるということになる．

2.4 二変量データの要約

第 2.2, 2.3 節では，一変量のデータに対して数量的，視覚的要約を行った．本節では，二つの確率変数 X, Y の n 個の観測値のペア $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられた場合の二変量データの要約を行う．

2.4.1 二変量データの数量的要約：標本共分散と標本相関係数

定義 2.1. x の標本平均 $\bar{x}$ と y の標本平均 $\bar{y}$ に対して以下の量を**標本共分散** ***(sample covariance)*** という．

$$\frac{S_{xy}}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

一般に，大きい x に対して大きい y が対応し，小さい x に対して小さい y が対応する右肩上がりの関係があるとき，x と y の間に**正の相関 (positive correlation)** があるという．逆に大きい x に小さい y が対応する右肩下がりの関係があるとき，x と y の間には**負の相関 (negative correlation)** が存在するという．

もし，x と y の間に強い正の相関があるならば，$\bar{x}$ より大きい x は $\bar{y}$ より大きい y に対応付けられ，$\bar{x}$ より小さい x は $\bar{y}$ より小さい y に対応付けられ，$(x_i - \bar{x})(y_i - \bar{y}) > 0$ となる傾向があるはずである．したがって正の相関は $S_{xy} > 0$ すなわち正の共分散を意味する．逆に，x と y の間に負の相関があるならば，多くの $(x_i - \bar{x})(y_i - \bar{y})$ は負の値をとり，$S_{xy} < 0$ すなわち標本共分散は負の値をとる．

標本共分散は n 個の観測値のペアにおける x と y の関係の強さを測る尺度の

一つではあるが，その値は x と y の単位に依存するため，値そのものはあまり使われず符号によって正負の相関を判断するのに用いられることが多い．x と y の関係を測る尺度としては，次に定義する標本相関係数の方が有用である．

定義 2.2. n 個の観測値のペア $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ に対して，**標本相関係数** ***(sample correlation coefficient)*** を以下で定義する．

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{2.6}$$

命題 2.1. *[*標本相関係数の性質*]*

1. 任意の定数 a, b, c, d，ただし a, c は同符号，に対して，$(ax_i + b, cy_i + d), i = 1, \ldots, n$ の標本相関係数は，$(x_i, y_i), i = 1, \ldots, n$ の標本相関係数 r に等しい．
2. $-1 \leq r \leq 1$
3. $r = 1$ のとき，$(x_i, y_i), i = 1, \ldots, n$ は正の傾きを持つ直線上にある．$r = -1$ のとき，$(x_i, y_i), i = 1, \ldots, n$ は負の傾きを持つ直線上にある．

命題 2.1, 1 は，r が x と y の単位の変換に関して不変であることを意味している．すなわち，モノを測る異なる単位は 1m = 3.2808 フィート, 1kg = 2.2046 ポンド, 温度のセ氏 (C) とカ氏 (F) であれば $C = (5/9) \times (F - 32)$ のように一般に $z = ax + b$ の関係にあるが，r はそのような変換によって変化しない．

命題 2.1, 2, 3 は，絶対値で 1 の正（負）の相関は正（負）の傾きを持つ直線で達成され，それより弱い相関は r の絶対値で 0 に近いことに対応する，すなわち r は x と y の「線形関係の強さ」を測る尺度であることを示している．

標本共分散，標本相関係数を求める R コマンドは，それぞれ `cov()` コマンドと `cor()` コマンドである．ここでは Low Infant Birth Weight データ（例題 2.3, (p. 29)）を用いて，母親の体重 `lwt` と子の出生体重 `bwt` の標本共分散と標本相関係数を求めてみよう．

```
> cov(lwt, bwt)# lwt と bwt の標本共分散 #
[1] 4141.652
> cor(lwt, bwt)# lwt と bwt の標本相関係数 #
[1] 0.1857333
```

標本共分散，標本相関係数共に正の値であることから，lwt と bwt の間に正の相関があることがわかる．ただし，標本相関係数が $r = 0.186$ と 0 に近いことから，lwt と bwt の間の線形関係は弱いと思われる．

2.4.2 二変量データの視覚的要約：散布図

定義 2.3. n 個の観測値のペア $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられた時，視覚的要約のためそれを二次元の座標系にプロットしたものを**散布図** ***(scatter plot)*** という．

R で散布図を描くためのコマンドは plot() コマンドである．以下に，母親の体重 lwt と子の出生体重 bwt の散布図を描いてみよう．図 2.6 に，lwt と bwt の散布図を示す．

```
plot(lwt, bwt)
```

図 2.6 Low Infant Birth Weight データ：lwt と bwt の散布図

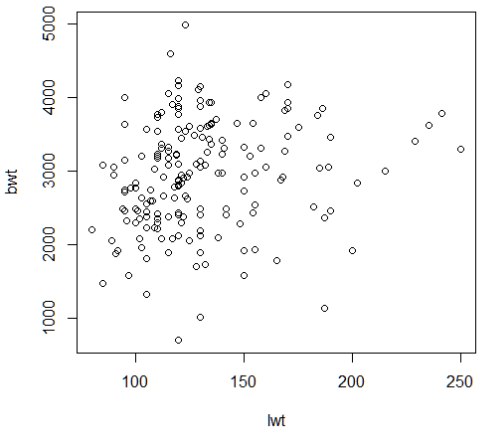


図 2.6 より，lwt が増えれば bwt も増える正の相関があるように見える．た

だし，直線関係は強くはなく相関は弱いものであり，それは標本相関係数が 0 に近い値であったことに矛盾しない.

表 2.1 第 2 章の R コマンドリスト

コマンド名	目的	使い方
c()	オブジェクトを結合する	c(3, 4, 10, 6, 5)
sum()	ベクトルの要素の和を求める	sum(x)
length()	ベクトルの長さを求める	length(x)
mean()	算術平均を求める	mean(x)
median()	中央値を求める	median(x)
quantile()	四分位点を求める パーセント点を求める	quantile(x) quantile(x, prob=k)
summary()	five numbers summary + mean	summary(x)
var()	分散を求める	var(x)
sd()	標準偏差を求める	sd(x)
sqrt()	平方根を求める	sqrt(x)
IQR()	四分位点間距離を求める	IQR(x)
hist()	ヒストグラムを描く（度数） （相対度数）	hist(x) hist(x, freq=F)
boxplot()	ボックスプロットを描く	boxplot(x)
cov()	標本共分散の計算	cov(x, y)
cor()	標本相関係数の計算	cor(x, y)
plot()	散布図の描画	plot(x, y)

第3章　確率論

ここまで第1章，第2章を通じて，統計学の目的とデータ解析の手順，データの収集の仕方およびデータの要約の方法を議論してきた．図 1.1 (P. 7) 「データ解析のフローチャート」に従えば，初めの三段階を済ませたことになる．当然次は具体的なデータ解析手法を含む推測統計に進むところである．

しかし本章では，データ解析の手法に議論を進める前に，まずデータに含まれる**不確実性 (randomness)**，すなわち偶然発生する現象の起こりやすさを測る**確率 (probability)** という概念を数学的にモデル化することを考える．だが「確率とは何ぞや」を考える前に，まずそもそも確率とはどのようなものに対して定義されるのか，確率を語る「舞台」を設定することから始めよう．

3.1　確率

定義 3.1.

標本空間 結果が偶然によって左右される実験，観察などの**試行** ***(trial or experiment)*** において，起こりうる可能な結果全てからなる集合を**標本空間** ***(sample space)*** と呼ぶ．標本空間は S, Ω などと記される．

事象 標本空間の部分集合を，**事象** ***(event)*** と呼ぶ．

ここでは，確率的な現象を表現する概念として標本空間と事象を定義した．確率とは標本空間上の事象に対して定義されるものであり，「標本空間と事象の

組」こそが，確率を考える「舞台」である．

例題 3.1.

コイン投げ コインを投げて，表裏どちらが出るか実験する．起こりうる結果は表か裏の二つに一つであるから，コイン投げの標本空間 S は

$$S = \{H, T\}$$

となる．ただし H:表 *(Head)*，T:裏 *(Tail)* とする．コイン投げにおける事象としてまず考えられるのは，「表が出る $\{H\}$」「裏が出る $\{T\}$」である．事象に対して確率が定義されるから，「表が出る確率」「裏が出る確率」を考えられるわけである．さらに標本空間 S そのものも標本空間の部分集合であるから $\{H, T\}$ も事象の一つである．$S = \{H, T\}$ の確率は「表もしくは裏が出る」確率に他ならないから，もちろん標本空間 S の確率は $1 = 100\%$ と定義されることになる．

最後に，「コインを投げた時，表と裏が同時に出る確率は *0* である」というように，起こる可能性がなく確率が *0* となる事象も定義しておく必要がある．このような要素を持たず起こりえない事象を**空事象** ***(empty event)*** と言い，$\emptyset$ と記す．結局，コイン投げにおけるすべての事象の集合は，以下の *4* つの要素からなる集合になる．

$$\text{コイン投げの事象の集合} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$$

サイコロ投げ サイコロを投げ，その目を読む試行を考える．起こりうる結果，すなわち標本空間 S は *1* から *6* までの整数からなる集合である．

$$S = \{1, 2, 3, 4, 5, 6\}$$

サイコロ投げにおける事象は，$\{1\}, \{2\}, \ldots, \{6\}$ のような事象から，標本空間 $\{S\}$ 自体まで様々なものが考えられる（標本空間 S が有限の n 個の要素からなるとき，その部分集合の全体からなる集合（冪（べき）集合, *power set*）は 2^n 個の要素をを持つことが知られている．サイコロ投げの場合，空

事象 $\emptyset$ を含め，全部で $2^6 = 64$ 通りある）．このうち，$\{1\}, \ldots, \{6\}$ のようにひとつの要素からなる事象を**根元事象** ***(elementary event)*** という．一方「偶数の目が出る」$= \{2, 4, 6\}$，「3 以下の目が出る」 $= \{1, 2, 3\}$ など複数の要素からなる事象を，**複合事象** ***(compound event)*** と呼ぶ．これらも全て標本空間の部分集合であり，サイコロ投げにおける事象である．

世論調査 いま n 人のサンプルに対して「現在の内閣を支持するか否か」の世論調査をしたとする．棄権や保留を考えないとすると，各個人の可能な態度は「支持する」か「支持しない」かの択一であるから，i 番目の個人の標本空間は $S_i = \{Y, N\}$ となる．ただし，Y：支持する，N：支持しない，とする．n 人のサンプル全体の標本空間は

$$S = \prod_{i=1}^{n} S_i = \{Y, N\} \times \{Y, N\} \times \cdots \times \{Y, N\}$$

となる．S は 2^n 通りの要素を持つ巨大な集合である．さらに S における事象の総数は，$\emptyset$ から S まで合わせて 2^{2^n} 通りという天文学的な数になる．（$n = 3$ であれば $2^n = 2^3 = 8$ であるから，$n = 3$ 人でも事象の総数は $2^{2^n} = 2^8 = 256$ 通りとなる．）

このように標本空間と事象を定めれば，標本空間上の事象に対して確率を考えることができる．確率という概念の定義に移る前に，事象についてもう少し考えよう．上では一つの事象について考えたが，二つ以上の事象が存在するとき，複数の事象の関係について，以下の概念を定義する．

定義 3.2.

和事象 A, B 二つの事象があるとき，A あるいは B に含まれる結果を $A \cup B$ と記して**和事象** ***(union of events)*** と呼ぶ．事象 A, B の何れか，あるいは両方が起こる事象である．$A \cup B$ を *“A **cup** B”* と称することがある．

積事象 A, B 二つの事象があるとき，A と B に共通に含まれる結果を $A \cap B$ と記して**積事象** ***(intersection of events)*** と呼ぶ．事象 A, B の両方が同時に起こる事象である．$A \cap B$ を *“A **cap** B”* と称する．

排反事象 二つの事象 A, B が共通部分を持たないとき，A と B は互いに**排反**

(disjoint) であるという．A と B が互いに排反であるならば，$A \cap B = \emptyset$，つまり A, B の共通部分は空事象となり A と B は同時には起こらない．

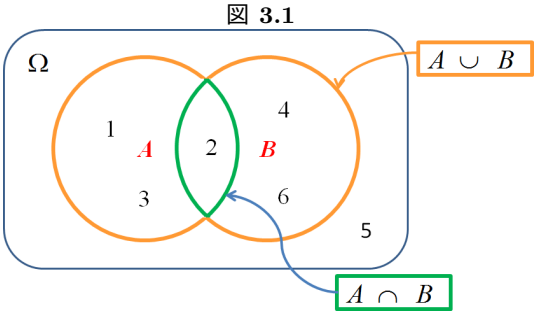
補事象（余事象） 事象 A があるとき，標本空間 S の中で A に含まれない結果を A^C と記して，**補事象** ***(complementary event)*** と呼ぶ．A と A^C は共通部分を持たないから，A とその補事象 A^C は必ず排反である．

例題 3.2. ここでは，サイコロ投げを例に考える．標本空間は $S = \{1, 2, \ldots, 6\}$．二つの事象として $A =$「偶数の目が出る」$= \{2, 4, 6\}$，$B =$「3 以下の目が出る」$= \{1, 2, 3\}$ を考える．このとき，

$$A \cup B = \{1, 2, 3, 4, 6\},\ A \cap B = \{2\},\ A^C =「奇数の目が出る」= \{1, 3, 5\}$$

であり，例えば A と $\{5, 6\}$ は共通部分を持たないから互いに排反である．

図 3.1



以上の定義の下で，「確率」という概念を以下のように定義する．

定義 3.3 (確率の公理 (Axiom of Probability))．標本空間 S 上の事象に対して定義された実数値関数 P は，以下の条件を満たすとき**確率** ***(Probability)*** と呼ぶ．

1. 任意の事象 $A \in S$ に対して，$0 \leq P(A) \leq 1$.
2. $P(S) = 1$.
3. 事象の列 $A_1, A_2, \ldots$ が互いに排反である時，すなわち $A_i \cap A_j = \emptyset,\ i \neq j$

である時,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right) \tag{3.1}$$

ただし，$\sum$ は和を表す記号で $\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$, $\sum_{i=1}^{\infty} x_i = x_1 + x_2 + \ldots$ を意味する．同様に積を表す記号 $\prod$ として，$\prod_{i=1}^{n} x_i = x_1 \times \cdots \times x_n$, $\prod_{i=1}^{\infty} = x_1 \times x_2 \times \ldots$ も存在する．$\bigcup$ は集合の和を表す記号で，$\bigcup_{i=1}^{\infty} A_i$ は無限個の事象 $A_1, A_2, \ldots$ の和事象を表す．同様に $\bigcap$ は集合の積を表す記号で，$\bigcap_{i=1}^{\infty} A_i$ は事象 $A_1, A_2, \ldots$ の積事象を示す．

上で述べた「確率の公理」のうち，公理 1 は，任意の事象に対して確率は 0 以上 1 以下であることを述べていて，確率と呼ばれるものがすべからく満たしている性質である．（ときどき文学的表現で「絶対確実」を意味するために「確率 200%だぁ」などと述べているのを見かけるが，数学的には全くナンセンスである．）

公理 2 は標本空間全体の確率は 1 であることを述べている．定義から標本空間は起こりうる可能な結果全てからなる集合であるから，逆に言えば標本空間の要素のいずれかは確率 1 で必ず実現するはずである．その意味で，公理 2 は確率の性質として当然のものである．

公理 3 は，互いに排反な事象の確率に関するものである．互いに排反である事象は共通部分を持たないから，同時には起こりえない．公理 3 の式 (3.1) の左辺は，互いに排反な事象の和集合の確率，すなわち同時に起こりえない事象のうちどれかが起こる確率を示しているが，それが個々の事象の確率の和である右辺に等しくなることを主張している．公正なサイコロを投げた時,「偶数の目もしくは奇数の目が出る確率」=1 は,「偶数の目が出る確率」=1/2 と「奇数の目が出る確率」=1/2 の和に等しい，という事実の一般化ともいえる．公理 3 の特徴は，この「同時に起こりえない事象全体の確率は，個々の事象の確率の和である」という命題が，無限個の排反な事象に対しても成り立つとすることであるが，確率と呼ばれるものの性質として納得できるものである．

このように「確率の公理」は確率と呼ばれるものが当然備えるべき性質を述べたものであるが，この公理が素晴らしいのは,「確率論の公理」を満たすものはすべからく「確率」と呼ぼうという発想の逆転にある．この「確率の公理」か

ら，応用上有用なあらゆる確率の性質が導かれるのである．

例題 3.3. 改めて，コイン投げを考える．例題 *3.1* にあるように $S = \{H, T\}$ である．表が出る確率を $P(\{H\}) = p$ とすると，公理 *1* により，$0 \leq p \leq 1$．$S = \{H\} \cup \{T\}$ であるから，公理 *2* より $P(S) = P(\{H\} \cup \{T\}) = 1$．一方，$\{H\} \cap \{T\} = \emptyset$ で $\{H\}$ と $\{T\}$ は互いに排反であるから，公理 *3* により $P(\{H\} \cup \{T\}) = P(\{H\}) + P(\{T\}) = p + P(\{T\}) = 1$ より $P(\{T\}) = 1 - p$.（公理 *3* において，$A_1 = \{H\}, A_2 = \{T\}, A_3 = A_4 = \cdots = \emptyset$ とする）もしコインが公正なコインであれば「表 *(H)*」と「裏 *(T)*」は**同様に確からしい *(equally likely)*** ので表と裏の出る確率は等しく $P(\{H\}) = P(\{T\})$ と置くことができる．このとき，

$$P(\{H\}) = P(\{T\}) \ \Rightarrow \ p = 1 - p \ \Rightarrow \ 2p = 1 \ \Rightarrow \ p = 1/2.$$

このように，サイコロの表と裏の *2* 通りが同様に確からしいならば，それぞれの確率が *1/2* となることは確率の公理から導かれる．

同様に，サイコロの *1* から *6* の目が同様に確からしいならば（言葉を換えれば，いずれかの目が特に出やすいと考える理由がないならば），それぞれの目が出る確率が *1/6* となることも確率の公理から導かれる．

例題 3.4 (ボンフェローニの不等式 (Bonferroni inequality))．$E_1, E_2, \ldots, E_K$ を，任意の標本空間 S 上の K 個の事象とする．このとき S 上の確率 P に対して，以下が成り立つ．

$$P\left(\bigcup_{k=1}^{K} E_k\right) \leq \sum_{k=1}^{K} P(E_k) \tag{3.2}$$

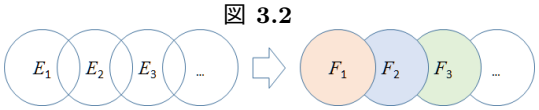
Proof. $E_1, E_2, \ldots, E_K$ に対して，

$$F_1 = E_1, F_2 = E_2 \cap E_1^C, F_3 = E_3 \cap (E_1 \cup E_2)^C, \ldots, F_k = E_k \cap \left(\bigcup_{j=1}^{k-1} E_j\right)^C, k = 2, \ldots, K$$

とする．（図 3.2 参照）このとき，F_k は互いに背反 $(F_i \cap F_j = \emptyset, i \neq j)$ で，$F_k \subset E_k$ より $E_k = F_k \cup (E_k \cap F_k^C)$ かつ $F_k \cap (E_k \cap F_k^C) = \emptyset$．よって，

$$P(E_k) = P(F_k \cup (E_k \cap F_k^C)) = P(F_k) + P(E_k \cap F_k^C) \quad \text{確率論の公理 3}$$

図 3.2



$$\geq P(F_k) \quad \text{確率論の公理 1 より } 0 \leq P(E_k \cap F_k^C) \tag{3.3}$$

$\bigcup_{k=1}^{K} E_k = \bigcup_{k=1}^{K} F_k$ であるから，

$$P\left(\bigcup_{k=1}^{K} E_k\right) = P\left(\bigcup_{k=1}^{K} F_k\right) = \sum_{k=1}^{K} P(F_k) \quad \text{確率論の公理 3}$$
$$\leq \sum_{k=1^K} P(E_k) \quad (3.3) \text{ より}$$

□

3.2 条件付き確率と独立性

前節で，標本空間上の事象に対して確率という概念を定義した．それでは，互いに影響を与え合う複数の事象が存在したとき，双方の事象にかかわる確率をどのように考えればよいであろうか．再びサイコロ投げの例を考えることにして，事象 A を $A =$「偶数の目が出る」$= \{2, 4, 6\}$，事象 B を $B =$「3 以下の目が出る」 $= \{1, 2, 3\}$ とする（p. 44, 図 3.1 参照）．いま B が起こったとすれば可能な結果は $\{1\}$, $\{2\}$, $\{3\}$ の三通りであるから，そのうち A が起こりうるのは $\{2\}$ の一通りだけである．一方 B が起こらなかったとすれば可能な結果は $\{4\}$, $\{5\}$, $\{6\}$ の三通りであり，その中で A が起こりうるのは $\{4\}$, $\{6\}$ の二通りである．つまり A, B 二つの事象を同時に考えた時，B が起こるか起こらないかで A の起こりうる場合の数が異なってくる訳である．

定義 3.4. ある事象 B が起こったという条件の下で事象 A の起こる確率を

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \ P(B) > 0 \tag{3.4}$$

と定義し，B を条件とする A の**条件付き確率** ***(conditional probability of*** A ***given*** B***)*** という．$P(B) = 0$ なる B に対しては，$P(A|B)$ は定義しない．

条件付き確率を定義する (3.4) 式は，以下のように解釈できる．すなわち，事象 B が起こったとするとその条件の下で事象 A が起こる可能な場合は $(A \cap B)$

だけである．その時，$P(A|B)$ は $P(B)$ を 1 とした時の $P(A \cap B)$ の割合であるから，$P(A|B)$ は以下のように求められる．

$$\begin{aligned} P(A|B) : 1 = P(A \cap B) : P(B) &\iff P(A \cap B) = P(A|B)P(B) \\ &\iff P(A|B) = \frac{P(A \cap B)}{P(B)} \end{aligned} \tag{3.5}$$

なお $P(B) = 0$ となる場合は，条件付けの前提となる事象 B の起こる確率が 0 であるということである．前提が成り立たないわけだから，$P(B) = 0$ なるときに A の条件付き確率 $P(A|B)$ を考えることは，そもそも無意味である．

条件付き確率の概念は，互いに影響を与え合う複数の確率的現象をモデル化するうえで本質的に重要である．確率の概念を導入することで，偶然に影響される事象が起こる確からしさをモデル化した．しかし，現実の世界ではある事象がそれ単独で存在することはまれである．例えば，肺がんの患者にある抗がん剤を投与するとする．事象 A を「抗がん剤が効果を上げる」こととすると，$P(A)$ は抗がん剤の奏効率を表している．しかし同じ肺がんの患者でも，がんのステージ，喫煙習慣の有無，ある種の遺伝子における突然変異の有無など様々な条件によって，抗がん剤の効果は異なってくるはずである．であるならば，患者の状態を無視して単に抗がん剤が効く確率 $P(A)$ を考えることはほとんど無意味であり，がんのステージ，喫煙経験，遺伝的変異など薬剤の効果に影響を与える要因 B で条件づけた条件付き確率 $P(A|B)$ を考えて初めて意味が出てくることになる．

その意味で，条件付き確率によって現象に関与する複数の要因の相互関係をモデル化することは，データ解析の本質と言っても過言ではない．

定義 3.5. 二つの事象 A，B に対して

$$P(A \cap B) = P(A)P(B) \tag{3.6}$$

が成り立つとき，A と B は互いに**独立** ***(independent)*** であるという．

A と B が互いに独立である時，(3.5), (3.6) から以下が成り立つ．

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A),$$
$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B).$$

すなわち A と B が独立ならば，B を条件とする A の条件付き確率 $P(A|B)$ は A 単独の確率 $P(A)$ に等しく，$P(B|A)$ は $P(B)$ に等しい．言葉を換えれば，A と B が独立ならば A と B は互いの発生確率に影響を与え合わない．

3.3 確率変数と確率分布

前節までで，「確率」という概念の定義を導入した．しかし確率を定義する対象である事象には，サイコロの目のように「数字で表せるもの」もあれば，コインの表裏，薬剤の効果の有無のように「数字で表せないもの」もある．これではあまりに一般的であり，また数学の世界で発展した微分，積分のような便利な概念を適用することもできない．そこで，標本空間上で考えられた確率的な現象を，数の世界と結びつけることを考える．

定義 3.6. ある試行に関する標本空間 S に対して，S 上で定義された実数値関数を**確率変数** ***(random variable)*** という．S 上の確率変数を X とし，X が値を取り得る集合（これを X の**値域**という）を D とすれば，

$$X : S \to D \subset \mathcal{R}$$

ただし，$\mathcal{R}$ は実数の集合である．

（数学的には，確率変数は S 上で定義された「可測関数」と呼ばれるものであり，確率変数の値域として実数の集合以外のものを考えることもできるが，本書では扱わない）サイコロの目のように元々標本空間の要素が数字で表されるときは，確率変数として $1 \mapsto 1, 2 \mapsto 2, \ldots$ のように標本空間の要素を自分自身に対応付ける変換（「恒等変換」という）を考えればよい．

確率変数は，その取り得る値により以下の二通りに分けられる

定義 3.7. 確率変数の値域が有限もしくは可算無限集合であるとき，**離散確率**

変数 ***(discrete random variable)*** という．確率変数の値域が実数上の区間 $[a,b], a<b$ である時（$(-\infty,\infty)$ を含む）もしくは互いに排反な区間の和集合であるとき，**連続確率変数** ***(continuous random variable)*** という．

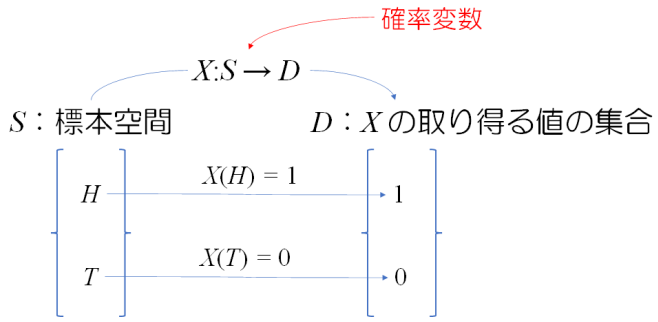
離散確率変数の定義にある「可算無限集合」とは，その要素が自然数と一対一に対応付けられる集合を意味する．つまり離散確率変数とは，その取り得る値に「一番目，二番目,...」というように順番を振ることができる確率変数である．整数値を値としてとる確率変数は，典型的な離散確率変数である．一方，実数全体を値としてとる，あるいは正の実数全体を値としてとる確率変数は連続確率変数である．

例題 3.5. ここではコイン投げを例に，確率変数を考えてみる．例題 *3.1* から，コイン投げの標本空間は $S=\{H,T\}$ である．このとき，表 *(H)* には *1*，裏 *(T)* には *0* を対応付ける関数 X を考えると，X は S 上で定義され $D=\{0,1\}$ を値域とする離散確率変数である．

$$X: S=\{H,T\} \to D=\{0,1\} \in \mathcal{R}$$
$$X(H)=1,\ X(T)=0.$$

このように二通りの結果しか持たない試行に対して $D=\{0,1\}$ を値域とする確

図 3.3



率変数を，**ベルヌーイ確率変数** ***(Bernoulli random variable)*** とよぶ．すな

わち，X の値域 $D = \{0, 1\}$ において $x = 1$ に対して確率 $p(1) = p$ を対応付け，$x = 0$ に対して確率 $p(0) = 1 - p$ を対応付けるルールそのものが，コイン投げに対するベルヌーイ確率変数 X の確率分布となる．

さてここまでで確率を考えたい試行に対して，1）確率を定義する舞台として「標本空間」と「事象」の概念を導入し，2）標本空間上の事象に対して「確率」を定義し，3)「確率変数」により元の標本空間を数の世界に結び付けた．ところで例題 3.5 のコイン投げに対するベルヌーイ確率変数 X についていえば，コインの表 (H) が出る事象と確率変数 X が $x = 1$ という値を持つことは同値であるから，$X = 1$ となる確率 $P(X = 1)$ はコインの表が出る確率 $P(\{H\}) = p$ と等しいはずである．このように，もともとは標本空間上で考えた確率を，確率変数を通じて数の世界の上に「写す」ことを考えよう．まずは，離散確率変数の場合から始める．

3.3.1 離散確率変数の確率分布

定義 3.8. X を離散確率変数とする．X の値域 D の任意の点 x に対して，$(X = x)$ となる確率 $p(x) = P(X = x)$ を x の関数とみて，X **の確率分布** ***(probability distribution)*** あるいは**確率関数** ***(probability mass function (pmf))*** という．

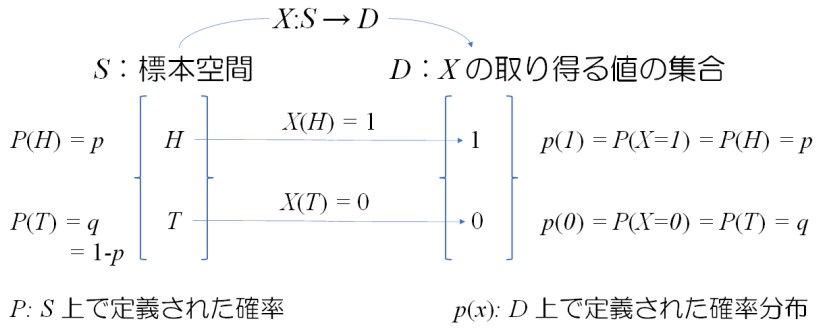
上の定義にある通り，確率変数はアルファベットの大文字 X，確率変数が実際にとった実現値は小文字の x というように表記を区別する．

例題 3.6. 例題 *3.5* に挙げた，コイン投げに対するベルヌーイ確率変数 X を例に考える．X の値域は $D = \{0, 1\}$ であり，X の確率分布は D 上で定義されて D の各点（*0* か *1*）に対して，その値を取り得る確率を対応付ける関数であったから，X の *pmf* $p(x)$ は以下のように与えられる．

$$p(1) = P(X = 1) = P(H) = p,$$
$$p(0) = P(X = 0) = P(T) = 1 - p.$$

例題 3.7. 次に，サイコロ投げの出た目を Y として，Y の確率分布を考えよう．

図 3.4


X:S → D
S：標本空間
D：X の取り得る値の集合
P(H) = p
H
X(H) = 1
1
p(1) = P(X=1) = P(H) = p
P(T) = q
= 1-p
T
X(T) = 0
0
p(0) = P(X=0) = P(T) = q
P: S 上で定義された確率
p(x): D 上で定義された確率分布

サイコロ投げの標本空間は $S = \{1, 2, 3, 4, 5, 6\}$ である．サイコロの目は「数字で表せるもの」であるから，確率変数 Y は S の各要素に自分自身を対応付ける恒等変換であり，Y の値域も $D = \{1, 2, 3, 4, 5, 6\}$ であって標本空間 S に一致する．このように標本空間が数の集合であるときは，標本空間とそのうえで定義された確率変数の値域は一致し，両者は通常区別されない．

さて，例題 *3.3* で見たとおりサイコロの目が同様に確からしいならば，それぞれの目が出る確率は *1/6* であるから，$D = \{1, 2, 3, 4, 5, 6\}$ 上で定義された Y の確率分布 $p(y)$ は以下のように与えられる．

$$p(y) = P(Y = y) = P(\{y\}) = 1/6,\ y = 1, \ldots, 6$$

命題 3.1 (離散確率分布の性質)**.** 離散確率変数 X が値域 D と確率分布 *pmf* $p(x)$ を持つとする．このとき確率の定義から，$p(x)$ は明らかに以下の性質を持つ．

1. $p(x) \geq 0,\ \forall x \in D$
2. $\sum_{x \in D} p(x) = 1$

定義 3.9. 離散確率変数 X が確率分布 *pmf* $p(x)$ を持つとする．このとき任意の $x \in \mathcal{R}$ に対して X が x 以下となる確率 $P(X \leq x)$ を x の関数とみて，X の**累積分布関数** ***(cumulative distribution function (cdf))*** あるいは**分布関数**といい，$F(x)$ と記す．

$$F(x) = P(X \leq x) = \sum_{y: y \leq x} p(y)$$

確率変数 X の分布関数は x の単調増加関数であり，右連続の階段関数になる．

3.3.2 確率変数の期待値と分散

前節では，確率変数を導入することで標本空間と数の世界を結びつけ，確率変数の値域である数の集合上に確率分布を導入することができた．ところで，もともと確率が定義されていた一般の標本空間と数の集合の間には，一つ決定的な違いがある．それは，数の集合では「量」あるいは大小関係といった「順序」が定義されるのに対して，一般的な標本空間には順序が存在するとは限らないことである．（例えば，コインの裏は表より「二倍である」とは言わないし，世

論調査における回答で Yes は No より「大きい」とも言わない.）

確率変数 X が確率分布 $p(x) = P(X = x),\ x \in D$ を持つとき，X の値域 D が数の集合でその上に順序が定義されているならば，その確率分布の「範囲」「中心」あるいは「広がり」を考えるのは自然なことである.

3.3.2.1 離散確率変数の期待値

まず，確率分布の「中心」あるいは確率変数が取り得る値の「平均」を考えよう．再びサイコロ投げを例にとると，サイコロの目が同様に確からしいのであればそれぞれの目が出る確率は等しく 1/6 で与えられた．一方，標本平均 (p. 19) は観測値 $x_1, x_2, \ldots, x_n$ に対して $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ であたえられた．このことから類推して，サイコロの目の平均は以下のように与えられる.

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5. \tag{3.7}$$

サイコロの目を $y = 1, \ldots, 6$ とすれば，それぞれの目が出る確率は $P(Y = y) = p(y) = 1/6, y = 1, \ldots, 6$ であったから，(3.7) は以下のように書き直せる.

$$1 \cdot p(1) + 2 \cdot p(2) + \cdots + 6 \cdot p(6) = 3.5$$

すなわち確率変数の平均（これを**期待値**という）は，確率変数の取り得る値をその対応する確率で重みづけた重み付き平均として求められる．一般に，離散確率変数の期待値は以下のように定義される.

定義 3.10. X を離散確率変数，X の値域を D，X の確率分布（確率関数, *pmf*）を $p(x)$ とする．このとき X の**期待値** ***(expected value)*** を $E(X)$, μ_X あるいは μ と記し，以下のように定義する.

$$E(X) = \mu_X = \mu = \sum_{x \in D} x \cdot p(x) \tag{3.8}$$

確率変数の期待値は確率分布の中心であり，確率分布の「位置 (location)」を示す指標と考えられる.

3.3.2.2 離散確率変数の関数の期待値

上では確率変数の期待値を定義したが，場合によっては確率変数の関数の期

待値が関心の対象になることもある．

定義 3.11. X を離散確率変数，X の値域を D，X の *pmf* を $p(x)$ とする．このとき X の関数 $h(X)$ の期待値を $E[h(X)]$ あるいは $\mu_{h(X)}$ と記し，以下のように定義する．

$$E[h(X)] = \mu_{h(X)} = \sum_{x \in D} h(x) \cdot p(x)$$

確率変数の期待値の定義から，以下の性質が導かれる．

命題 3.2. *[期待値の性質 1]* 確率変数 X と任意の定数 a, b に対して，以下が成り立つ．

$$E(aX + b) = aE(X) + b \tag{3.9}$$

Proof. X の関数を $h(X) = aX + b$ とする．定義 3.11 から，

$$\begin{aligned} E[h(X)] = \sum_{x \in D} h(x) \cdot p(x) = \sum_{x \in D} (ax + b) \cdot p(x) &= a \sum_{x \in D} x \cdot p(x) + b \sum_{x \in D} p(x) \\ &= aE(X) + b \end{aligned}$$

但し，上の式一行目最右辺で，期待値の定義から $\sum_{x \in D} x \cdot p(x) = E(X)$，命題 3.1 より $\sum_{x \in D} p(x) = 1$． □

3.3.2.3 離散確率変数の分散と標準偏差

確率変数 X の期待値は，X の確率分布の中心，位置を示すものであった．これに対して確率分布の「広がり (spread)」，「変動 (variability)」あるいは「散らばり (dispersion)」の大きさの尺度として，標本分散 (p. 19) に倣って離散確率変数 X の**分散**と**標準偏差**を以下のように定義する．

定義 3.12. 離散確率変数 X が，値域 D，*pmf* $p(x)$，期待値 μ を持つとする．このとき，X の**分散** ***(variance)*** を $V(X)$, σ_X^2 あるいは σ^2 と記して，以下のように定義する．

$$V(X) = \sigma_X^2 = \sigma^2 = \sum_{x \in D} (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

また，X の**標準偏差** ***(standard deviation, SD)*** を σ_X あるいは σ と記して，以下のように定義する．

$$\sigma_X = \sigma = \sqrt{\sigma_X^2}$$

分散の定義にある $h(X) = (X - \mu)^2$ は，確率変数 X と X の分布の中心である期待値 μ との二乗距離であり，分散は $h(X)$ の期待値である．したがって X の分布の大部分が μ の近くに集中していれば，$h(X)$ が小さな値をとり分散 $V(X)$ の値も小さくなる．逆に，μ から遠くはなれた x に対して $p(x) > 0$ となれば，$V(X)$ の値は大きなものになりうる．

命題 3.3. *[分散と標準偏差の性質]*

1. $V(X) \geq 0,\ \sigma_X \geq 0$
2. $V(X) = E(X^2) - [E(X)]^2$
3. $V(aX + b) = \sigma_{aX+b}^2 = a^2 \cdot \sigma_X^2$
4. $\sigma_{aX+b} = |a| \cdot \sigma_X$

ただし，a, b は任意の定数．

3.3.3　連続確率変数の確率分布

前節では，離散確率変数の確率分布と期待値，分散を考えた．離散確率分布の値域は，有限もしくは可算無限集合であった．本節では取り得る値の集合が実数上の非可算無限集合となる連続確率変数について，確率分布を定義する．

定義 3.13. X を連続確率変数とする．任意の $a \leq b$ に対して，以下の条件を満たす関数 $f(x)$ を X の**確率分布** ***(probability distribution)***，**確率密度関数** ***(probability density function (pdf))*** あるいは単に**密度関数**という．

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{3.10}$$

すなわち，確率変数 X が区間 $[a, b]$ 上に値をとる確率は，$[a, b]$ の上，密度関数 $f(x)$ の下の領域の面積となる．任意の連続確率変数 X に対して，極めて緩やかな条件の下で (3.10) を満たす確率密度関数 $f(x)$ が存在することが証明され

ている．（実際，筆者はデータ解析の場面で密度関数の存在しない連続確率変数というものに，お目にかかったことがない）

命題 3.4 (連続確率分布の性質)**.** 連続確率変数 X が確率密度関数 $f(x)$ を持つとする．このとき確率の定義から，$f(x)$ は明らかに以下の性質を持つ．

1. $f(x) \geq 0,\ \forall x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

また，(3.10) から $b = a$ とすると，

$$P(X = a) = \int_{a}^{a} f(x)dx = 0$$

すなわち，任意の実数 a に対して連続確率変数 X が a に一致する確率は 0 である．離散確率変数の場合と同様，連続確率変数の累積分布関数を以下のように定義する．

定義 3.14. 連続確率変数 X に対して，X の**累積分布関数** ***(cumulative distribution function (cdf))*** あるいは**分布関数** $F(x)$ を以下のように定義する．

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

連続確率変数 X の分布関数は x の単調増加関数であり，以下の性質を持つ

命題 3.5. *[連続分布関数の性質]* 連続確率変数 X が密度関数 $f(x)$ と分布関数 $F(x)$ を持つとする．

1. $\lim_{x\to-\infty} F(x) = 0,\ \lim_{x\to\infty} F(x) = 1$
2. 任意の定数 $a < b$ に対して，

$$P(X > a) = 1 - F(a),\ P(a \leq X \leq b) = F(b) - F(a)$$

3. $f(x) = \frac{d}{dx}F(x)$

分布関数に関連して，標本パーセント点 (p. 23) に倣い以下の概念を定義する．

定義 3.15. p を $0 \leq p \leq 1$ なる任意の実数とする．連続確率変数 X が確率密度関数 $f(x)$ と累積分布関数 $F(x)$ を持つとき，X の $100 \times p$ **パーセント点（*p-th percentile*）**を q と記し，以下のように定義する．

$$p = F(q) = P(X \leq q) = \int_{-\infty}^{q} f(x)dx \tag{3.11}$$

すなわち，確率変数 X の分布の $100 \times p$ パーセント点 q とは，密度関数 $f(x)$ の下の面積のうち $100 \times p$ パーセントが q 以下にあり，$100 \times (1-p)$ パーセントが q 以上であるような点である．

3.3.3.1 連続確率変数の期待値，分散，標準偏差

離散確率変数の場合と同様に，連続確率変数に対しても期待値を定義する．

定義 3.16. X を連続確率変数，X の確率密度関数を $f(x)$ とする．このとき X の**期待値** ***(expected value)*** を，以下のように定義する．

$$E(X) = \mu_X = \mu = \int x \cdot f(x)dx \tag{3.12}$$

離散確率変数の期待値の定義 (3.8) と比較すると，連続確率変数の期待値の定義 (3.12) では，足し算が積分に，確率関数 pmf $p(x)$ が確率密度関数 pdf $f(x)$ にそれぞれ置き換えられていることが分かる．X の関数の期待値，分散，標準偏差も，以下のように定義される．

定義 3.17. X を確率密度関数 $f(x)$ を持つ連続確率変数，$h(X)$ を X の関数とする．このとき $h(X)$ の期待値を以下のように定義する．

$$E[h(X)] = \mu_{h(X)} = \int h(x) \cdot f(x)dx$$

定義 3.18. 連続確率変数 X が，確率密度関数 $f(x)$ と期待値 μ を持つとする．このとき，X の**分散** ***(variance)*** は以下のように定義される．

$$V(X) = \sigma_X^2 = \sigma^2 = \int (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

X の**標準偏差** ***(standard deviation)*** は，以下のように定義される．

$$\sigma_X = \sigma = \sqrt{V(X)}$$

3.3.4 正規分布

第 3.3 節では，離散と連続の確率変数について議論してきた．ここで，確率分布の具体例の一つとして，**正規分布 (normal distribution)** と呼ばれる確率分布を紹介する．正規分布は連続確率分布の一つであり，身長，体重など自然界の多くの変量の分布をよく近似し，また実験などにおける測定誤差をモデル化するのにも用いられる．さらに，仮に個々の確率変数が正規分布に従わなかったとしても，その平均と和の確率分布は比較的緩やかな条件の下で正規分布で近似されるという性質（中心極限定理（p. 88, 定理 3.1））が知られており，正規分布は統計学と確率論の世界で最も重要な確率分布とされるものである．

定義 3.19. 連続確率変数 X の確率密度関数が以下の $f(x;\mu,\sigma^2)$ で与えられるとき，X はパラメター $\mu,\sigma^2(\sigma>0)$ の**正規分布** ***(normal distribution)*** に従うといい，$X \sim N(\mu,\sigma^2)$ と記す．

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},\ -\infty < x < \infty \tag{3.13}$$

ただし，$-\infty<\mu<\infty,\ 0<\sigma<\infty$ である．

命題 3.6. *[正規分布の性質]* ここでは正規分布の性質を，証明なしで述べる．

1. $f(x;\mu,\sigma^2)>0,\ \int_{-\infty}^{\infty} f(x;\mu,\sigma^2)dx = 1.$
2. 正規分布の密度関数 $f(x;\mu,\sigma^2)$ は，パラメター μ,σ^2 で特徴づけられる．
3. $E(X)=\mu,\ V(X)=\sigma^2$
4. 正規分布の密度関数 $f(x;\mu,\sigma^2)$ の形状は，μ を中心に左右対称の釣鐘型（ベル型，*bell-shaped*）である．

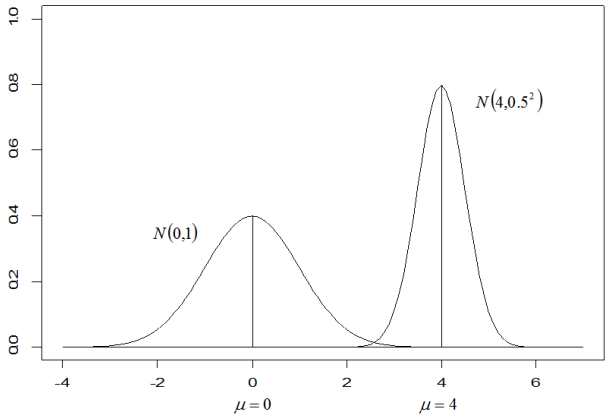
特に，正規分布の期待値が $\mu=0$，分散が $\sigma^2=1$ であるとき，$f(x;0,1)$ を**標準正規分布 (standard normal distribution)** とよぶ．標準正規分布に従う確率変数を標準正規確率変数と呼び，とくに Z と記す．Z の確率密度関数を $\phi(z)$，累積分布関数を $\Phi(z)$ と記す．

$$\phi(z) = f(z;0,1) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2},\ -\infty < z < \infty$$

$$\Phi(z) = \int_{-\infty}^{z} \phi(u)du$$

図 3.5 に，異なる (μ, σ^2) を持つ正規分布の確率密度関数 $f(x; \mu, \sigma^2)$ を示す．正規分布は期待値 μ に関して左右対称であるから，正規分布の中央値（メディアン）$\tilde{\mu}$ は μ に一致する．図 3.5 から，σ^2 が小さいとき $f(x; \mu, \sigma^2)$ のグラフの頂点はより高く，μ 周辺のより狭い範囲に分布することが分かる．

図 3.5 正規分布の密度関数



次に，正規分布に従う正規確率変数の「標準化」と呼ばれる性質を述べる．まず，確率変数一般の標準化とその性質について述べる．

定義 3.20. 確率変数 X が期待値 μ，分散 $\sigma^2 < \infty$ を持つとする．このとき X から μ を引き，σ で割る変換を，X の**標準化** ***(standardization)*** と呼ぶ．すなわち，X の標準化 Z とは，以下の変換である．

$$Z = \frac{X - \mu}{\sigma}$$

命題 3.7. 期待値 μ，分散 σ^2 を持つ任意の確率変数 X に対して，その標準化 Z

は期待値 $E(Z) = 0$，分散 $V(Z) = 1$ を持つ．

Proof.

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma}X + \left(-\frac{\mu}{\sigma}\right)\right) = \frac{1}{\sigma}E(X) + \left(-\frac{\mu}{\sigma}\right) = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0,$$

$$V(Z) = V\left(\frac{1}{\sigma}X + \left(-\frac{\mu}{\sigma}\right)\right) = V\left(\frac{1}{\sigma}X\right) = \left(\frac{1}{\sigma}\right)^2 V(X) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

なお上の導出では，命題 3.2（p. 56）と命題 3.3(3)（p. 57）を用いた． □

命題 3.8. *[正規確率変数の標準化]* いま，確率変数 X が期待値 μ，分散 σ^2 の正規分布に従うとする．$X \sim N(\mu, \sigma^2)$．このとき，正規確率変数 X の標準化 Z は，期待値 *0*，分散 *1* の標準正規分布に従う．$Z \sim N(0, 1)$．

Proof. Z の累積分布関数を $\Phi(z)$ とする．$Z = (X - \mu)/\sigma$ より，

$$\Phi(z) = P(Z \leq z) = P(X \leq \sigma z + \mu) = \int_{-\infty}^{\sigma z+\mu} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

$$\frac{d}{dz}\Phi(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{((\sigma z + \mu) - \mu)^2}{2\sigma^2}\right\} \sigma = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

命題 3.5, 3 (p. 58) によれば，これは Z の確率密度関数が標準正規分布の確率密度関数に等しいことを示す．したがって，$Z \sim N(0, 1)$． □

3.3.4.1 正規分布に関する R コマンド

さて，ここまでは確率変数，特に正規確率変数の一般的な性質について議論してきた．ここからは，特に正規分布に従う確率変数の解析に必要な実践的方法について議論したい．R には，正規分布の確率密度関数 (pdf)，累積分布関数 (cdf)，クォンタイル（quantile, 分位数）関数，および正規分布に従う乱数を生成する以下のコマンドが用意されている．正規分布の期待値と標準偏差は，それぞれ `mean` オプションと `sd` オプションで指定する．`mean` オプションと `sd` オプションのデフォルト値はそれぞれ `mean = 0, sd = 1` であり，その場合（つまり標準正規分布の場合）はこれらのオプションを省略できる．`log, log.p` オプションは `TRUE, FALSE` の論理値をとるオプションで，`TRUE` ならば対数で変換した値が返される（あまり使わない）．`lower.tail` オプションも論理値をとるオ

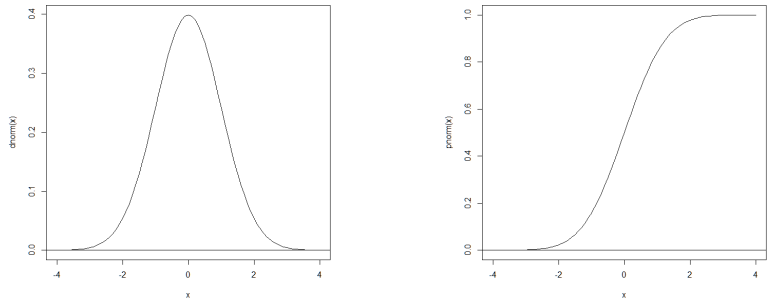
プションで，TRUE ならば下側確率 $P(X \leq x)$ に対応する値を返す．

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

正規分布の確率密度関数 (pdf)　実数もしくはそのベクトル x における正規分布の密度関数の値は，dnorm() コマンドで生成される．標準正規分布の密度関数のグラフを描画するには，以下のようにする．（図 3.6 左参照）

```
x <- seq(-4, 4, 0.1)
plot(x, dnorm(x), type="l")
```

図 3.6



一行目の seq() コマンドは，-4 から 4 までの 0.1 間隔の数列を生成する．dnorm(x) でベクトル x の各点上での密度関数の値を求める．plot() コマンドの type オプションに与えられた値 type="l"は，点と点の間を結ぶ直線 (line) を描画する．図 3.6 左にある通り，正規分布の密度関数は期待値を中心に左右対称であり，その形は釣鐘型（ベル型）である．

正規分布の累積分布関数 (cdf)　確率変数 X と任意の $x \in \mathcal{R}$ に対して，累積分布関数 (cdf) とは $P(X \leq x)$ を x の関数とみたものであった．`pnorm()` は，与えられた期待値と標準偏差を持つ正規分布の cdf の値を返す．標準正規分布の累積分布関数は，以下のようにして描かれる．（図 3.6 右参照）

```
x <- seq(-4, 4, 0.1)
plot(x, pnorm(x), type="l")
```

図 3.6 右にある通り，cdf は一般に単調増加関数であり，下限は 0，上限は 1 である．

正規分布のクォンタイル関数　クォンタイル関数とは累積分布関数の逆関数であり，与えらえた確率 p に対して $p = P(X \leq q)$ となる q を返す関数である．定義 3.15 (p. 58) によれば，クォンタイル関数とは与えらえた確率 p に対して $100 \times p$ パーセント点を返す関数ということになる．`qnorm()` コマンドは，与えられた期待値と標準偏差を持つ正規分布のパーセント点を返すクォンタイル関数である．統計解析でよく用いられる標準正規分布のパーセント点のいくつかを，以下に示す．

```
> qnorm(0.025)# 標準正規分布の 2.5%点
[1] -1.959964
> qnorm(0.05)# 標準正規分布の 5%点
[1] -1.644854
```

場合によっては**上側パーセント点**，すなわち確率 p に対して $p = P(X \geq q)$ となる q が必要となる場合もある．その際は，`qnorm()` コマンドの `lower.tail` オプションの値を `lower.tail=FALSE` とする．

```
> qnorm(0.025, lower.tail=FALSE)# 標準正規分布の上側 2.5%点
[1] 1.959964
```

統計学では，標準正規分布の上側 $100 \times \alpha$ パーセント点を z_α と記す慣習がある．これに従えば，上の標準正規分布の上側 2.5%点は $z_{0.025} = 1.96$ と記される．

標準正規分布は原点 0 を中心に左右対称な分布であるから，標準正規分布の上側パーセント点は下側パーセント点と符号は異なるが絶対値は同じである．

正規乱数の生成 正規分布に従う乱数を生成するには，rnorm() コマンドを用いる．rnorm(5) とすれば，期待値 0，分散 1 の標準正規分布に従う乱数が 5 個生成される．（生成される乱数は，毎回異なった値となる）

```
> rnorm(5)
[1] -1.9793439  0.4809165 -1.4142816  0.5630464 -0.7457991
```

以下に，$n = 1000$ 個の標準正規乱数を生成し，そのヒストグラムと正規確率密度関数を重ね書きする R のプログラムを示す．（図 3.7 左参照）

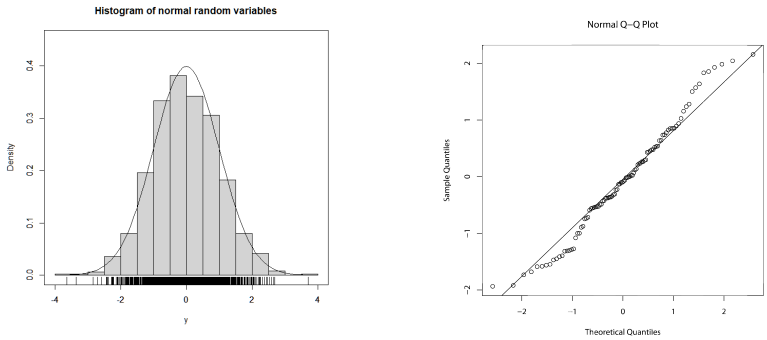
```
n <- 1000
y <- rnorm(n)
hist(y, prob=TRUE, main="Histogram of normal random variables",
xlim=c(-4, 4), ylim=c(0, 0.45))
rug(y)
par(new=TRUE)
x <- seq(-4, 4, 0.1)
plot(x, dnorm(x), type="l", xlim=c(-4, 4), ylim=c(0, 0.45), xlab="", ylab="")
```

hist() コマンドでは，タイトルを指定するために main オプションを，また x, y 軸の範囲を指定するために xlim，ylim オプションを追加している．par(new=TRUE) は，図を重ね描きするためパラメター new を指定している．密度関数を描画するための plot() コマンドは，xlim，ylim オプションで作図の範囲をヒストグラムと合わせ，また，軸のラベルが重なるのを避けるため，xlab="", ylab="" として plot() コマンドでは軸ラベルが出力されないようにしている．

3.3.4.2 QQ-nom plot（normal quantile-quantile plot, 正規確率プロット）

今，n 個のサンプル $x_1, x_2, \ldots, x_n$ が観察されたとする．統計解析で用いられる多くのモデルでは，観察されたサンプルを生み出した母集団の確率分布が正

図 3.7


Histogram of normal random variables
Density
y
Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles

規分布であると仮定されることが多い．前節最後では，rnorm() コマンドを用いて正規分布から乱数を生成する方法を学んだ．本節では，逆に観察されたデータからそのデータを生み出した確率分布が正規分布であるか否かを知る方法を考えよう．まず以下の概念を定義する．

定義 3.21. n 個のサンプルを大きさ順に $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ のように並べなおしたとき，i 番目に小さな観測値 $x_{(i)}$ を $[100(i-0.5)/n]$ **標本パーセント点** ***(****$[100(i-0.5)/n]$***th sample percentile)*** と呼ぶ．

標本パーセント点の概念の本質は，n 個のサンプル中 i 番目に小さなサンプル $x_{(i)}$ を “$100 \times i/n$ パーセント点” と考えることである．すなわち，データ全体を 1 とした時，$x_{(i)}$ が下から何パーセントのところにあるのか相対的な位置を示すのが標本パーセント点である．では，$[100(i-0.5)/n]$ という定義にある “0.5” は何のためにあるのか．理由は単純で，例えばサンプル数が $n = 3$ の場合を考えよう．サンプルを大小順に並べなおせば $x_{(1)} \leq x_{(2)} \leq x_{(3)}$ であるが，0.5 の修正を用いずに $100 \times i/n$ でパーセント点を定義した場合，下から 50%の中央にくる $x_{(2)}$ は $100 \times 2/3 = 66.67$ パーセント点となり具合が悪い．しかし，$[100(i-0.5)/n]$ を定義に用いれば，$100 \times (2-0.5)/3 = 50$ で，サンプル数が奇数の場合に真ん中の大きさのサンプルは 50%標本パーセント点となる．

さて，標本パーセント点の概念を用いてサンプルの正規性を確認するアイデアは，以下の通りである．すなわち，もし n 個のサンプル $x_1, x_2, \ldots, x_n$ が正規分布から生成されたのであれば，その標本パーセント点は正規分布の理論的なパーセント点（定義 3.15 参照）の「すぐそば」にあるはずである．そうであるならば，標本パーセント点とそれに対応する正規分布の理論的なパーセント点を平面上にプロットすれば，プロットされた点は直線の周囲に分布するはずである．このような考えから，以下の QQ-norm plot を定義する．

定義 3.22. n 個の標本が得られたとき, 標準正規分布 *N(0, 1)* の $[100(i-0.5)/n]$ パーセント点と i 番目に小さな観測値=$[100(i-0.5)/n]$ 標本パーセント点のプロットを ***QQ-nom plot*** **（*normal quantile-quantile plot,* 正規確率プロット）** という．

もし観察されたサンプルが，実際に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ か

ら生成されたものであれば，そのサンプルの QQ-norm plot は切片 μ，傾き σ の直線の周囲にプロットされるはずである．

第 2.3 節で，データの視覚的な要約の方法としてヒストグラムとボックスプロットを紹介した．今後は，第三の視覚的要約の方法として QQ-norm plot を付け加えることにしよう．新しいデータが得られたら，ヒストグラムとボックスプロットで分布の形状を要約すると同時に，QQ-norm plot によってデータの正規性を確認することも併せて習慣づけてほしい．

QQ-norm plot を描画するための R コマンドは，`qqnorm()` コマンドである．以下に，標準正規分布から生成した 100 個の乱数を用いた，QQ-norm plot 作成の R のプログラムを示す．（図 3.7 右参照）

```
n <- 100
x <- rnorm(n)
qqnorm(x)
qqline(x)
```

初めの 2 行で 100 個の正規乱数ベクトル `x` を生成し，3 行目の `qqnorm(x)` で QQ-norm plot を作図している．本節で述べたとおり，QQ-norm plot は描かれた点が直線の周囲に分布するかでデータの正規性を判断するものであるから，比較の対象となる直線を付け加えれば便利である．4 行目の `qqline(x)` は，縦軸と横軸の 25%点と 75%点を結ぶ直線を描き加える．図 3.7 右から，正規分布から生成された乱数の QQ-norm plot は（当然ながら）十分直線の近くに分布していることが分かる．

3.4 多次元の確率分布

第 3.3 節では，一次元の確率変数についてその確率分布を考えた．しかし実際のデータ解析の場面では，複数の確率変数が同時に登場する．本節では，まず二つの確率変数の同時確率分布を考え，後にさらに多数の確率変数の同時分布に一般化する．

3.4.1 二次元の確率変数の同時確率分布

3.4.1.1 二次元の離散確率変数の同時確率分布

一つの離散確率変数 X の確率分布あるいは確率関数 (probability mass function, pmf) とは，X の取り得る値 x に対して $(X = x)$ となる確率 $p(x) = P(X = x)$ を対応付けたものであった．二つの離散確率変数 X, Y の同時分布の場合は，可能な確率変数の値のペア (x, y) に対して $(X = x, Y = y)$ となる確率を対応付けたものである．

定義 3.23. X と Y を二つの離散確率変数とする．(X, Y) が任意の点 (x, y) を取る確率 $p(x, y) = P(X = x, Y = y)$ を (x, y) の関数とみて，(X, Y) の**同時確率関数** ***(joint probability mass function)*** と呼ぶ．

$$p(x, y) = P(X = x, Y = y),\ x \in D_X,\ y \in D_Y$$

ただし，D_X, D_Y はそれぞれ X, Y の値域．いま，A を任意の (x, y) からなる集合とする．このとき

$$P[(X, Y) \in A] = \sum_{(x,y)\in A} p(x, y)$$

命題 3.9. 確率の性質から，以下が成り立つ．

$$p(x, y) \geq 0,\ \sum_x \sum_y p(x, y) = 1$$

例題 3.8. いま二つのサイコロを投げた時，出た目を X, Y とする．(X, Y) の同時確率分布は以下のとおりである．

$$p(x, y) = P(X = x, Y = y),\ x, y = 1, 2, \ldots, 6$$

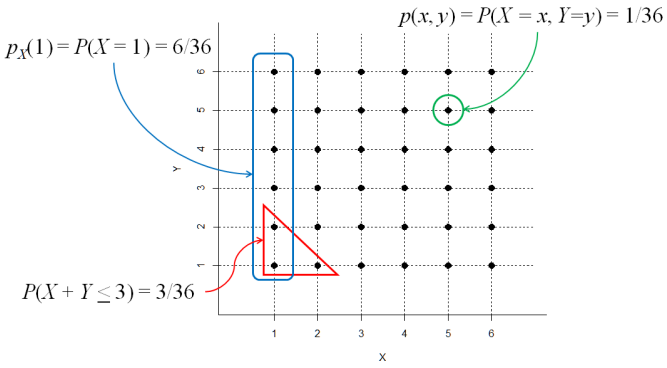
可能な (x, y) の組み合わせは，$(1, 1), (1, 2), \ldots, (6, 6)$ まで計 $6 \times 6 = 36$ 通りある．もし二つのサイコロが公正なものであれば *36* 通りの組み合わせは同様に確からしいから，任意の (x, y) に対して $p(x, y) = 1/36$ となる．（図 *3.8* 参照）

いま集合 A として，「二つのサイコロの目の和が *3* 以下である」$= (X + Y \leq 3)$

となる場合を考える．すなわち，$A = \{(x,y) : x+y \leq 3\} = \{(1,1)\} \cup \{(1,2)\} \cup \{(2,1)\}$．（図 *3.8* $P(X+Y \leq 3)$）このとき

$$P[(X,Y) \in A] = P(X+Y \leq 3) = \sum_{(x,y)\in A} p(x,y)$$
$$= p(1,1) + p(1,2) + p(2,1) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12}$$

図 3.8



二つの離散確率変数の同時確率分布が定義されると，そこから個々の確率変数それぞれの確率分布が導出される．例えば例題 3.8 において，$(X=1)$ となるのは $(X,Y) = (1,1),(1,2),\ldots,(1,6)$ の 6 通りの場合に分けられる．したがって，

$$p_X(1) = P(X=1) = P[(1,1) \cup (1,2) \cup \cdots \cup (1,6)]$$
$$= p(1,1) + p(1,2) + \cdots + p(1,6) = \frac{1}{36} + \frac{1}{36} + \cdots + \frac{1}{36} = \frac{1}{6}$$

すなわち，$p_X(1)$ を求めるには，(x,y) の組み合わせのうち $(x=1)$ を固定した $(1,y), y = 1,\ldots,6$ の形をした全てについて同時確率分布 $p(x,y)$ を足し合わせることになる．（図 3.8 $p_X(1) = P(X=1)$）同様にして X あるいは Y 単独の確率

分布 $p_X(x), p_Y(y)$ を導出することができる．

定義 3.24. X, Y の**周辺確率関数** ***(marginal probability mass function)*** を以下で定義する．

$$p_X(x) = \sum_y p(x,y),\ x \in D_X, \quad p_Y(y) = \sum_x p(x,y),\ y \in D_Y$$

ただし，D_X, D_Y はそれぞれ X, Y の値域．

3.4.1.2 二次元の連続確率変数の同時確率分布

連続な確率変数 X に対して，X が実数上の区間 A（あるいはその和集合）の上に値をとる確率は，確率密度関数 (probability density function, pdf) $f(x)$ の A 上での積分で与えられた．同様に二つの連続確率変数 X, Y のペア (X, Y) が二次元空間上の集合 A の上に値をとる確率は，以下に定義する同時確率密度関数の A 上での積分で与えられる．

定義 3.25. X と Y を二つの連続確率変数とする．任意の集合 A に対し，以下の *(3.14)* を満たす $f(x,y)$ が存在するとき，$f(x,y)$ を (X,Y) の**同時確率密度関数** ***(joint probability density function)*** いう．

$$P[(X,Y) \in A] = \iint_A f(x,y)dxdy \tag{3.14}$$
$$P(a \le X \le b,\ c \le Y \le d) = \int_a^b \int_c^d f(x,y)dxdy$$

命題 3.10. 確率の性質から，以下が成り立つ．

$$f(x,y) \ge 0,\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

離散確率変数の場合と同様に，連続確率変数の同時分布に対しても周辺確率密度関数が定義される．

定義 3.26. X, Y の**周辺確率密度関数** ***(marginal probability density function)*** を以下で定義する．

$$f_X(y) = \int_{-\infty}^{\infty} f(x,y)dy,\ -\infty < x < \infty, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx,\ -\infty < y < \infty$$

3.4.2 条件付き確率分布と確率変数の独立性

第 3.2 節において，標本空間 S 上で定義された事象に対する条件付き確率と独立性について議論した．第 3.3 節において導入した確率変数により，標本空間上の「確率」は確率変数が値をとる数の集合上の「確率分布」に写された．本節では，この数の集合における確率変数の条件付き確率分布と独立性について議論する．

3.4.2.1 条件付き確率分布

まず，二つの離散確率変数に関する条件付き確率分布を考える．X と Y を離散確率変数とし (X,Y) の同時確率関数を $p(x,y)$ とする．$(Y=y)$ なる事象を A，$(X=x)$ なる事象を B とする．このとき条件付き確率の定義式 (3.4)(p. 48) によれば，$(X=x)$ なる条件の下での $(Y=y)$ の条件付き確率 $p(y|x)$ は以下のように与えられる．

$$p(y|x) = P(Y=y|X=x) = P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(X=x, Y=y)}{P(X=x)} = \frac{p(x,y)}{p_X(x)}$$

ただし，y は確率変数 Y の値域 D_Y の任意の値をとり，x は $p_X(x) > 0$ を満たすものとする．X と Y が連続確率変数である場合は，確率関数を確率密度関数に置き換えて条件付き確率分布を定義する．

定義 3.27. X と Y を離散確率変数とし (X,Y) の同時確率関数を $p(x,y)$ とする．Y の値域を D_Y，X の周辺確率関数を $p_X(x)$ とする．$p_X(x) > 0$ となる任意の x に対して，$(X=x)$ **を条件とする** Y **の条件付き確率関数** ***(conditional probability mass function of*** Y ***given*** $X=x$***)*** を以下のように定義する．

$$p(y|x) = \frac{p(x,y)}{p_X(x)},\ y \in D_Y$$

一方，X と Y が連続確率変数の場合，(X,Y) の同時確率密度関数を $f(x,y)$，X の周辺確率密度関数を $f_X(x)$ とする．$f_X(x) > 0$ となる任意の x に対して，

$(X = x)$ を条件とする Y の条件付き確率密度関数 ***(conditional probability density function of*** Y ***given*** $X = x$***)*** を以下のように定義する．

$$f(y|x) = \frac{f(x,y)}{f_X(x)},\ -\infty < y < \infty$$

例題 3.9. いま，二つのコインを投げそれぞれの表を *1*，裏を *0* に対応付けたベルヌーイ確率変数を X, Y とする．X, Y の確率関数 *(probability mass function, pmf)* を，それぞれ $p_X(x)$, $p_Y(y)$ とする．(X, Y) の同時確率分布は以下のように与えられる．

$$p(x,y) = P(X = x, Y = y),\ x = 0, 1,\ y = 0, 1$$
$$p(0,0) = p(0,1) = p(1,0) = p(1,1) = 1/4$$

さてここで，$Z = X + Y$ とする．すなわち Z は二つのコインを投げた時の表の数の合計である．Z の取り得る値域は $D_Z = \{0, 1, 2\}$ であり，Z の *pmf,* $p_Z(z)$ は以下のように与えられる．

$$\begin{aligned}
p_Z(0) &= P(Z = 0) = P(X = 0, Y = 0) = p(0,0) = 1/4 \\
p_Z(1) &= P(Z = 1) = P(X = 1, Y = 0) + P(X = 0, Y = 1) \\
&= p(1,0) + p(0,1) = 1/4 + 1/4 = 1/2 \\
p_Z(2) &= P(Z = 2) = P(X = 1, Y = 1) = p(1,1) = 1/4
\end{aligned}$$

さらに今度は，(X, Z) の同時確率分布を考えてみると，

$$\begin{aligned}
&p(x,z) = P(X = x, Z = z),\ x = 0, 1,\ z = 0, 1, 2 \\
&p(0,0) = P(X = 0, Z = 0) = P(X = 0, Y = 0) = 1/4 \\
&p(0,1) = P(X = 0, Z = 1) = P(X = 0, Y = 1) = 1/4 \\
&p(0,2) = P(X = 0, Z = 2) = 0 \\
&p(1,0) = P(X = 1, Z = 0) = 0 \\
&p(1,1) = P(X = 1, Z = 1) = P(X = 1, Y = 0) = 1/4
\end{aligned}$$

$$p(1,2) = P(X=1, Z=2) = P(X=1, Y=1) = 1/4$$

上で，例えば $p(0,2)$ は，*1* 枚目のコインが裏 $(X=0)$ である時，*2* 枚のコインの表の数が *2*$(Z=2)$ である確率である．もちろんこれは起こりえないことであるから，$(X=0, Z=2)$ は空事象 $\emptyset$ であり $p(0,2) = P(X=0, Z=2) = 0$ である．$p(1,0) = P(X=1, Z=0) = 0$ も同様に示される．

このとき，$(X=x)$ を条件とする Z の条件付確率関数は以下のように与えられる．

$$P(Z=z|X=x) = \frac{p(x,z)}{p_X(x)}$$

$(X=0)$ の場合

$$\begin{aligned} p(0|0) &= \frac{p(0,0)}{p_X(0)} = \frac{1/4}{1/2} = 1/2 \\ p(1|0) &= \frac{p(1,0)}{p_X(0)} = \frac{1/4}{1/2} = 1/2 \\ p(2|0) &= \frac{p(2,0)}{p_X(0)} = \frac{0}{1/2} = 0 \end{aligned}$$

$(X=1)$ の場合

$$\begin{aligned} p(0|1) &= \frac{p(0,1)}{p_X(1)} = \frac{0}{1/2} = 0 \\ p(1|1) &= \frac{p(1,1)}{p_X(1)} = \frac{1/4}{1/2} = 1/2 \\ p(2|1) &= \frac{p(2,1)}{p_X(1)} = \frac{1/4}{1/2} = 1/2 \end{aligned}$$

上の例から明らかなように（また，定義から明らかなように）Z の条件付確率分布は条件付ける確率変数 X の値に依存して異なるものとなる．

条件付き確率分布はそれ自体確率分布であり，条件付き期待値と条件付き分散が定義される．

定義 3.28. X と Y を離散確率変数とし，$(X=x)$ を条件とする Y の条件付き

確率関数を $p(y|x)$ とする．このとき，$(X = x)$ を条件とする Y の条件付き期待値 ***(conditional expectation)*** と条件付き分散 ***(conditional variance)*** を，以下のように定義する．

$$E(Y|X) = E(Y|X = x) = \sum_{y \in D_Y} y \cdot p(y|x) = \mu_{Y|x}$$
$$V(Y|X) = V(Y|X = x) = \sum_{y \in D_Y} (y - \mu_{Y|x})^2 \cdot p(y|x)$$

X と Y が連続確率変数の場合，$(X = x)$ を条件とする Y の条件付き確率密度関数を $f(y|x)$ とすると，$(X = x)$ を条件とする Y の条件付き期待値と条件付き分散は以下のように定義される．

$$E(Y|X) = E(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f(y|x)dy = \mu_{Y|x}$$
$$V(Y|X) = V(Y|X = x) = \int_{-\infty}^{\infty} (y - \mu_{Y|x})^2 \cdot f(y|x)dx$$

条件付き期待値と条件付き分散は，条件付ける確率変数 X の値に依存する．したがって，$(X = x)$ を固定すれば，$E(Y|X = x), V(Y|X = x)$ の値が定まる x の関数と考えることができる．また，X の確率分布を考えれば，$E(Y|X), V(Y|X)$ は X の関数であり，$E(Y|X), V(Y|X)$ 自体が X に依存する確率変数と考えられる．

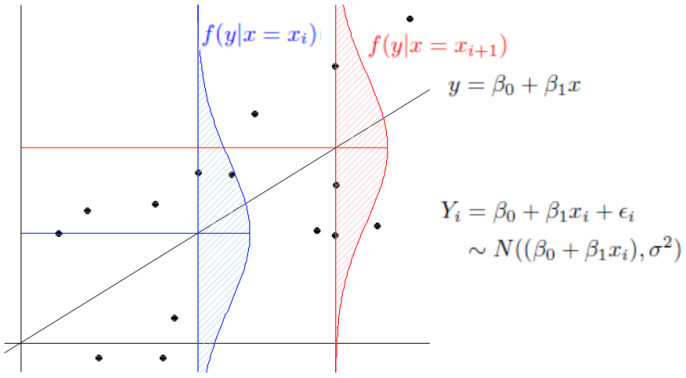
例題 3.10. いま，ある学校の生徒の身長 X が，期待値 μ_X, 分散 σ_X^2 の正規分布に従うとする．$X \sim N(\mu_X, \sigma_X^2)$．一方，この学校の生徒の体重 Y は，以下のように表されると仮定する．

$$Y = \beta_0 + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2) \tag{3.15}$$

ただし，ϵ は期待値 *0*, 分散 σ^2 の正規分布に従うランダムな観測誤差を表し，体重 Y は身長 X の一次式 $y = \beta_0 + \beta_1 x$ で近似されると考える．このとき，$(X = x)$ を条件とする Y の条件付き確率分布は，条件付き期待値 $E(Y|X = x) = \beta_0 + \beta_1 x$，条件付き分散 σ^2 の正規分布 $N(\beta_0 + \beta_1 x, \sigma^2)$ となる．

一般に，例題 3.10 のように複数の確率変数が互いに影響を与え合っている場

図 3.9


$f(y|x = x_i)$
$f(y|x = x_{i+1})$
$y = \beta_0 + \beta_1 x$
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
$\sim N((\beta_0 + \beta_1 x_i), \sigma^2)$

合，一つの確率変数の確率分布や期待値を単独で考えることにはあまり意味がない．例題 3.10 でいえば，体重が身長の影響を受けることが既知であるのに，漠然と「体重 Y の期待値はいくらか?」と考えるのはナンセンスである．「身長が $X = x$cm である時，体重 Y の条件付き期待値はいくらか？」というように Y に影響を与える因子の情報を限定した問いにこそ意味がある．その意味で，条件付き確率分布の概念は，相互に影響しあう因子の間の関係をモデル化するにあたって本質的に重要である．実際，本書の後半で取り上げる 1）回帰分析，2）ロジスティック回帰分析，3）生存時間解析は，説明される変数がそれぞれ 1）連続確率変数，2）0，1 の値をとるベルヌーイ確率変数，3）イベントが起こるまでの時間，である場合の条件付確率分布を具体的に求める手法に他ならない．例題 3.10 式 (3.15) の係数 β_0, β_1 をデータから推定する問題は，第 **??**章回帰分析で扱う．

3.4.2.2 確率変数の独立性

定義 3.4 の条件付確率に続き，定義 3.5 では二つの事象 A，B に対して独立性の概念を定義した．すなわち，二つの事象 A，B は

$$P(A \cap B) = P(A)P(B)$$

が成り立つとき，互いに独立であるという．いま二つの離散確率変数 X と Y に対して，事象 A，B をそれぞれ $(X = x)$, $(Y = y)$ で置き換えれば，$(X = x)$, $(Y = y)$ が互いに独立である時以下が成り立つ．

$$p(x, y) = P(X = x, Y = y) = P(X = x)P(Y = y) = p_X(x)p_Y(y) \qquad (3.16)$$

(3.16) が任意の $x \in D_X, y \in D_Y$ について成り立つとき，離散確率変数 X と Y は互いに独立であるという．連続確率変数の場合も，同様に独立性を定義する．

定義 3.29. X と Y を二つの離散確率変数とする．(X, Y) の同時確率関数を $p(x, y)$，X, Y の周辺確率関数をそれぞれ $p_X(x), p_Y(y)$ とする．このとき，任意の $x \in D_X, y \in D_Y$ について以下が成り立つとき，二つの確率変数 X と Y は互いに**独立** ***(independent)*** であるという．

$$p(x,y) = p_X(x)p_Y(y) \tag{3.17}$$

X と Y が二つの連続確率変数である場合，(X,Y) の同時確率密度関数を $f(x,y)$，X,Y の周辺確率密度関数をそれぞれ $f_X(x), f_Y(y)$ とする．任意の x,y に対して以下が成立するとき，X と Y は互いに**独立** ***(independent)*** であるという．

$$f(x,y) = f_X(x)f_Y(y) \tag{3.18}$$

(3.17) あるいは *(3.18)* が成り立たないとき，X と Y は互いに**従属** ***(dependent)*** であるという．

統計学，確率論では，複数の確率変数が互いに独立に同一の確率分布に従う，と仮定することがある．この「独立かつ同一の分布に従う」という表現を，英語で “independently and identically distributed” と表し **iid** あるいは **IID** と記す．

例題 3.11. いま，確率変数 X と Y が独立かつ同一の標準正規分布に従うとする．$X,Y \overset{iid}{\sim} N(0,1)$．このとき (X,Y) の同時確率密度関数は，以下のように得られる．

$$f(x,y) = f_X(x)f_Y(y) = \phi(x)\phi(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} = \frac{1}{2\pi}\exp\left\{-\frac{x^2+y^2}{2}\right\}$$

3.4.3 三次元以上の確率変数の同時確率分布

第 3.4.1 節では，二次元の確率変数の同時確率分布について考えた．本節では，まず定義 3.23 に倣い三次元以上の確率変数の同時確率分布を定義する．

定義 3.30. $X_1, X_2, \ldots, X_n$ を n 個の離散確率変数とする．このとき $(X_1, X_2, \ldots, X_n)$ の**同時確率関数** ***(joint probability mass function)*** を以下のように定義する．

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

但し，$x_i \in D_{X_i}, i = 1, \ldots, n$，$D_{X_i}$ は X_i の値域である．$X_1, X_2, \ldots, X_n$ を n 個の連続確率変数である場合，$(X_1, X_2, \ldots, X_n)$ の**同時確率密度関数** ***(joint probability density function)*** $f(x_1, x_2, \ldots, x_n)$ は，以下のように定義される．

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

但し，$-\infty < a_i \leq b_i < \infty, i = 1, \ldots, n$.

また定義 3.29 に倣い，三つ以上の確率変数の独立性を以下のように定義する．

定義 3.31. $X_1, X_2, \ldots, X_n$ を n 個の確率変数とする．このとき，$\{X_1, X_2, \ldots, X_n\}$ の任意の部分集合 $\{X_{i_1}, \ldots, X_{i_k}\} \subset \{X_1, X_2, \ldots, X_n\}$ に対して，$(X_{i_1}, \ldots, X_{i_k})$ の同時確率関数（*pmf,* 離散確率変数の場合）あるいは同時確率密度関数（*pdf,* 連続確率変数の場合）が，周辺確率関数あるいは周辺確率密度関数の積として書けるとき $X_1, X_2, \ldots, X_n$ は互いに**独立** ***(independent)*** であるという．

$$\begin{aligned}
&X_1, X_2, \ldots, X_n \text{ は互いに独立} \\
\Longleftrightarrow\ &\forall \{X_{i_1}, \ldots, X_{i_k}\} \subset \{X_1, X_2, \ldots, X_n\}, \\
&p(x_{i_1}, \ldots, x_{i_k}) = p_{X_{i_1}}(x_{i_1}) \ldots p_{X_{i_k}}(x_{i_k}) \quad \text{：離散確率変数の場合} \\
&f(x_{i_1}, \ldots, x_{i_k}) = f_{X_{i_1}}(x_{i_1}) \ldots f_{X_{i_k}}(x_{i_k}) \quad \text{：連続確率変数の場合}
\end{aligned}$$

特に，$X_1, X_2, \ldots, X_n$ は互いに独立である場合，$\{X_{i_1}, \ldots, X_{i_k}\} = \{X_1, X_2, \ldots, X_n\}$ とすれば，

$$\begin{aligned}
&p(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n) \quad \text{：離散確率変数の場合} \\
&f(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n) \quad \text{：連続確率変数の場合}
\end{aligned}$$

さらに，$X_1, X_2, \ldots, X_n$ が独立かつ同一の確率分布に従う (iid) 場合，$p_{X_1} = p_{X_2} = \cdots = p_{X_n} = p$（離散確率変数の場合）あるいは $f_{X_1} = f_{X_2} = \cdots = f_{X_n} = f$（連続確率変数の場合）と書けるから，

$$p(x_1, \ldots, x_n) = p(x_1) \ldots p(x_n) = \prod_{i=1}^{n} p(x_i),\ f(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

となることに注意する．

例題 3.12. いま，$X_1, X_2, \ldots, X_n$ が独立かつ同一の標準正規分布に従うとする．

このとき，$(X_1, \ldots, X_n)$ の同時確率密度関数は，以下のように与えられる．

$$X_1, \ldots, X_n \overset{iid}{\sim} N(0,1),\ f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \phi(x_i) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\}$$

3.4.4 二変量の期待値，母集団共分散，母集団相関係数

定義 3.11，定義 3.17 では，一つの確率変数 X に対して X の関数 $h(X)$ の期待値 $E[h(X)]$ を定義した．本節では，二つの確率変数 X, Y に対して関数 $h(X, Y)$ の期待値を定義しよう．

3.4.4.1 二つの確率変数の関数の期待値

定義 3.32. X, Y を同時に分布する二つの確率変数とする．X, Y が離散確率変数の場合 (X, Y) の同時確率関数 *(joint pmf)* を $p(x, y)$，連続確率変数の場合 (X, Y) の同時確率密度関数 *(joint pdf)* を $f(x, y)$ とする．このとき関数 $h(X, Y)$ の期待値を $E[h(X, Y)]$ あるいは $\mu_{h(X,Y)}$ と記し，以下のように定義する．

$$E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y) \cdot p(x,y) & \textit{(X, Y)} \text{ は離散確率変数} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f(x,y) dxdy & \textit{(X, Y)} \text{ は連続確率変数} \end{cases}$$

定義 3.32 から，以下の期待値の性質が導かれる．

命題 3.11. *[*期待値の性質 *2]* X, Y を二つの確率変数とする．このとき

$$E(X+Y) = E(X) + E(Y)$$

Proof. X, Y が連続確率変数である場合について証明する．(X, Y) の関数を $h(X, Y) = (X + Y)$ とする．定義 3.32 から，

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f(x,y) dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) \cdot f(x,y) dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) dxdy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dxdy \\ &= \int_{-\infty}^{\infty} x \underbrace{\left\{\int_{-\infty}^{\infty} f(x,y) dy\right\}}_{=f_X(x)} dx + \int_{-\infty}^{\infty} y \underbrace{\left\{\int_{-\infty}^{\infty} f(x,y) dx\right\}}_{=f_Y(y)} dy \end{aligned}$$

$$= \int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y f_Y(y)dy = E(X) + E(Y)$$

□

命題 3.12. *[期待値の性質 3]* X, Y を二つの独立な確率変数とする．このとき

$$E(XY) = E(X)E(Y)$$

Proof. これも X, Y が連続確率変数である場合について証明する．(X, Y) の関数を $h(X, Y) = (XY)$ とする．

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) \cdot f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy) \cdot f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy) \cdot f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} y f_Y(y) \underbrace{\left\{\int_{-\infty}^{\infty} x f_X(x)dx\right\}}_{=\mu_X = E(X)} dy = E(X)\int_{-\infty}^{\infty} y f_Y(y)dy = E(X)E(Y) \end{aligned}$$

2 行目の等号は無条件には成立しない．本命題では X, Y は互いに独立であると仮定されており，定義 3.29 式 (3.18) により $f(x, y) = f_X(x)f_Y(y)$ となることから 2 行目の等号は成立している． □

命題 3.2[期待値の性質 1](p. 56) と合わせ，期待値の性質として押さえておこう．

命題 3.13. X_1, X_2 を二つの独立な正規分布に従う確率変数とする．それぞれの期待値を μ_1, μ_2，分散を σ_1^2, σ_2^2 とする．$X_1 \sim N(\mu_1, \sigma_1^2,\ X_2 \sim N(\mu_2, \sigma_2^2)$．このとき $(X_1 + X_2)$ は，期待値 $(\mu_1 + \mu_2)$，分散 $(\sigma_1^2 + \sigma_2^2)$ の正規分布に従う．

$$E(XY) = E(X)E(Y)$$

3.4.4.2 母集団共分散と母集団相関係数

第 2.4.1 で定義した標本共分散と標本相関係数は，x と y の観測値のペアの関係の強さを測る尺度であった．いま，(x_i, y_i) のペアを，同時確率関数 (pmf)

$p(x,y)$ あるいは同時確率密度関数 (pdf) $f(x,y)$ を持つ二次元の確率変数の実現値であると考えれば，標本共分散と標本相関係数に倣い X と Y の関係の強さを測る尺度として，以下の母集団共分散と母集団相関係数が定義できる．

定義 3.33. X,Y を二つの確率変数とする．このとき，X と Y の**母集団共分散** ***(population covariance)***（あるいは単に**共分散** ***(covariance)***）を以下で定義する．

$$
\begin{aligned}
Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\
&= \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x,y) & (X,\ Y) \text{ は離散確率変数} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy & (X,\ Y) \text{ は連続確率変数} \end{cases}
\end{aligned}
$$

$(X-\mu_X)$ と $(Y-\mu_Y)$ は，X と Y のそれぞれの期待値からの乖離である．いま X と Y の間に強い正の関係が存在すれば，$(X-\mu_X)$ と $(Y-\mu_Y)$ は同時に正もしくは同時に負の値をとる可能性が高いから，$(X-\mu_X)(Y-\mu_Y)>0$ となり共分散は正の値をとる傾向がある．逆に，X と Y の間に負の関係があれば，共分散は負の値をとる傾向がある．共分散には，以下の性質が知られている．

命題 3.14. X,Y を二つの確率変数とする．このとき，以下が成立する．

1. 任意の $a,b,c,d \in \mathcal{R},\ ac>0$ に対して，$Cov(aX+b,cX+d) = abCov(X,Y)$
2. $Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y$
3. X と Y は互いに独立であると仮定する．このとき $Cov(X,Y)=0$
4. X と Y は互いに独立であると仮定する．このとき $Var(X+Y) = Var(X) + Var(Y)$

定義 3.34. X,Y を二つの確率変数とする．このとき，X と Y の**母集団相関係数** ***(population correlation coefficient)***（あるいは単に**相関係数** ***(correlation coefficient)***）を $Corr(X,Y), \rho_{XY}, \rho$ と記し，以下で定義する．

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

ただし，σ_X, σ_Y はそれぞれ X,Y の標準偏差．

相関係数の性質として，以下が知られている．

命題 3.15. *1.* ともに正もしくは負の値をとる任意の $a, c \in \mathcal{R}, ac > 0$ と任意の $b, d \in \mathcal{R}$ に対して，$Corr(aX + b, cY + d) = Corr(X, Y)$

2. $-1 \leq Corr(X, Y) \leq 1$

3. もし X と Y が独立ならば，$\rho = 0$．しかし，$\rho = 0$ は X と Y が独立であることを意味しない．

4. $\rho = 1$ もしくは -1 となる必要十分条件は，ある $a \neq 0, b$ が存在して $Y = aX + b$ となることである．

命題 3.15, 1 は相関係数 ρ が，X と Y の測定単位の線形変換に関して不変であることを示している．これはまた，命題 **??**と同様 ρ が X と Y の単位の変換に関して不変であることを示してもいる．命題 3.15, 2, 4 は X と Y の正（あるいは負）の最大の相関は $|\rho| = 1$ で達成され，それは正（あるいは負）の傾きを持つ線形関係であることを示している．

命題 3.15, 3 は X と Y が独立（すなわち，X と Y が互いに影響を与え合わず無相関）であるならば $\rho = 0$，しかしその逆は成り立たないことを示している．$\rho = 0$ が独立を意味しない例として，以下を挙げる．

例題 3.13. 確率変数 X, Y の値を，以下のように定める．

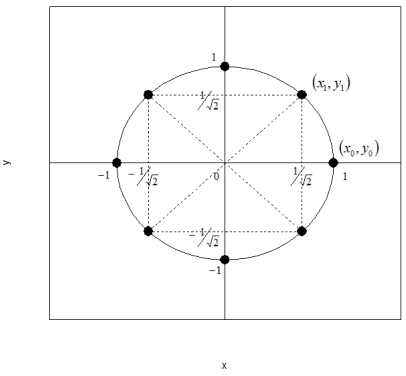
$$x_i = \cos(\pi/4 \times i),\ y_i = \sin(\pi/4 \times i),\ i = 0, \ldots, 7 \tag{3.19}$$

X, Y の値は整数値ではないが，*(3.19)* の定義にある通り X, Y の値域は有限集合であるから X, Y は離散確率変数である．(X, Y) の取り得る値をプロットすると，図 *3.10* 左のようになる．また，$(x_i, y_i), i = 0, \ldots, 7$ が同様に確からしいとすると，図 *3.10* 左の各点の確率は等しく *1/8* であり，X, Y の周辺確率分布は以下のとおり，$E(X) = E(Y) = 0$ となる．

このとき，以下の通り $Cov(X, Y) = 0$ となる．

$$\begin{aligned} Cov(X, Y) &= \frac{1}{8} \sum_{i=0}^{7} x_i y_i \\ &= \frac{1}{8}(0 + 1/2 + 0 - 1/2 + 0 - 1/2 + 0 + 1/2) = 0 \end{aligned}$$

図 3.10



X の周辺確率分布

x	-1	$-1/\sqrt{2}$	0	$1/\sqrt{2}$	1
$p_X(x)$	1/8	2/8	2/8	2/8	1/8

Y の周辺確率分布

y	-1	$-1/\sqrt{2}$	0	$1/\sqrt{2}$	1
$p_Y(y)$	1/8	2/8	2/8	2/8	1/8

したがって，X と Y の相関係数も $\rho = Cov(X,Y)/(\sigma_X\sigma_Y) = 0$ である．しかし *(3.19)* から (X,Y) は原点を中心に半径 *1* の円周上に分布しており，X の値を定めれば Y の値は符号を除いて定まる．よって X と Y は明らかに独立ではない．

3.4.5 標本平均の分布

標本平均 $\bar{X}$ は，母集団平均 μ の推論を行うために用いられる．まず，$T_0 = X_1 + X_2 + \cdots + X_n$ とすると，T_0 について以下の性質が成り立つ．

命題 3.16. $X_1, X_2, \ldots, X_n$ を互いに独立で，等しい期待値 μ と分散 σ^2 を持つとする．このとき，$T_0 = X_1 + X_2 + \cdots + X_n$ に対して，以下が成立する．

$$E(T_0) = n\mu$$
$$V(T_0) = n\sigma^2,\ \sigma_{T_0} = \sqrt{n}\sigma$$

Proof.

$$E(T_0) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \quad (\text{命題 3.11 より})$$
$$= n\mu$$
$$V(T_0) = E[(T_0 - E(T_0))^2] = E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X_i)\right)^2\right] = E\left[\left(\sum_{i=1}^{n}(X_i - E(X_i))\right)^2\right]$$
$$= E\left[\sum_{i=1}^{n}(X_i - E(X_i))^2 + \sum_{i\neq j}(X_i - E(X_i))(X_j - E(X_j))\right]$$
$$= \sum_{i=1}^{n}\underbrace{E(X_i - E(X_i))^2}_{=V(X_i)=\sigma^2} + \sum_{i\neq j}\underbrace{E(X_i - E(X_i))(X_j - E(X_j))}_{\mathrm{Cov}(X_i,X_j)} = n\sigma^2$$

但し，最後の等号は命題 3.12 より，X_i, X_j が互いに独立であることから

$$\mathrm{Cov}(X_i, X_j) = E(X_i - E(X_i))(X_j - E(X_j)) = E(X_i - E(X_i)) \times E(X_j - E(X_j))$$
$$= (E(X_i) - E(X_i)) \times (E(X_j) - E(X_j)) = 0$$

であることを用いた． □

命題 3.16 より，直ちに以下の命題を得る．

命題 3.17. $X_1, X_2, \ldots, X_n$ を互いに独立な確率変数で，等しい期待値 μ と分散 σ^2 を持つとする．このとき以下が成立する．

$$E(\bar{X}) = \mu$$
$$V(\bar{X}) = \sigma^2/n,\ \sigma_{\bar{X}} = \sigma/\sqrt{n}$$

Proof.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E(T_0) = \frac{1}{n}n\mu = \mu$$
$$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 V(T_0) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},\quad \sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

□

命題 3.17, 1 は標本平均 $\bar{X}$ の期待値が推定対象である母集団平均 μ に一致する，すなわち $\bar{X}$ を用いれば μ をバイアスなしに推定できるという**不偏性 (unbiasedness)** と呼ばれる重要な性質である．（p. 97, 定義 4.1 参照）

一方命題 3.17, 2 は，サンプル数 n が無限に増大するとき標本平均の分散は 0 に収束することを意味している．$V(\bar{X}) = \sigma^2/n \to 0\ (n \to \infty)$　これは，多くのサンプルを集め情報量を増やせばより精確に μ を推定できることを保証する，標本平均の持つもう一つの重要な性質を示している．

なお，任意の統計量の標準偏差を**標準誤差 (standard error)** という．特に標本平均の標準誤差 $\sigma/\sqrt{n}$ を，**SEM (Standard Error of Mean)** ということもある．

3.4.5.1 正規母集団からの標本平均の分布

上では，一般に $X_1, X_2, \ldots, X_n$ を互いに独立で，等しい期待値 μ と分散 σ^2 を持つ確率分布に従うと仮定した．本節では，さらに $X_1, X_2, \ldots, X_n$ が同一の正規分布 $N(\mu, \sigma^2)$ に従うとして，標本平均の分布について検討する．

命題 3.18. *1.* 確率変数 X が期待値 μ，分散 σ^2 の正規分布に従うならば，定数 a, b に対して $(aX+b)$ は期待値 $(a\mu+b)$，分散 $b^2\sigma^2$ の正規分布 $N(a\mu+b, b^2\sigma^2)$ に従う．

2. $X_1, X_2, \ldots, X_n$ を，期待値 μ，分散 σ^2 の正規分布に従う独立な正規確率変数とする．このとき任意の n に対して

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

例題 3.14. ある心不全患者の集団の心拍数は，平均 *71.5*，標準偏差 *14.2* の正規分布に従うとする．このとき，この心不全患者集団から抽出した *1000* 人のサンプルの標本平均が，*(70.5, 72.5)* の範囲に含まれる確率を求めよう．

命題 *3.18* より，標本平均 $\bar{X}$ は期待値 $\mu = 71.5$，標準偏差 $\sigma/\sqrt{n} = 14.2/\sqrt{1000} = 0.449$ の正規分布に従う．命題 *3.8[*正規確率変数の標準化*]* によれば，$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$ であるから以下を得る．

$$\begin{aligned} P(70.5 < \bar{X} < 72.5) &= P\left(\frac{70.5 - 71.5}{0.449} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{72.5 - 71.5}{0.449}\right) \\ &= P(-2.227 < Z < 2.227) = P(Z < 2.227) - P(Z < -2.227) \\ &= \Phi(2.227) - \Phi(-2.227) = 0.987 - 0.0130 = 0.974 \end{aligned}$$

なお，正規分布の累積分布関数の値 $P(Z < 2.227), P(Z < -2.227)$ の値は *R* コマンドの *pnorm()(p. 62)* を用いて，以下のように求めた．

```
> pnorm((70.5 - 71.5)/0.449)
[1] 0.01296791
> pnorm((72.5 - 71.5)/0.449)
[1] 0.9870321
> pnorm((72.5 - 71.5)/0.449) - pnorm((70.5 - 71.5)/0.449)
[1] 0.9740642
```

3.4.5.2 中心極限定理

命題 3.18 は，$X_1, \ldots, X_n$ が正規分布に従うとき，任意の n に対して $\bar{X}$ が正

規分布に従うことを示していた．一方，$X_1, \ldots, X_n$ が正規分布に従わない場合でも，n が十分に大きければ $\bar{X}$ の確率分布が正規分布で近似できることが示される．（定理に必要な諸条件は割愛する．）

定理 3.1. *[*中心極限定理 *(The Central Limit Theorem, CLT)]* $X_1, X_2, \ldots, X_n$ を独立かつ同一の分布に従う *(independently and identically distributed, iid)* 確率変数とする．ただし，$E(X_i) = \mu, V(X_i) = \sigma^2, i = 1, \ldots, n$ とする．このとき標本平均の確率分布は正規分布 $N(\mu, \sigma^2/n)$ に収束する．

$$X_1, \ldots, X_n \overset{iid}{\sim} (\mu, \sigma^2) \Rightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to N(\mu, \sigma^2/n) \ (n \to \infty)$$

現実にある様々なランダム現象は，正規分布とは限らない様々な確率分布としてモデル化される．中心極限定理はその多様な確率的現象から出発したとしても，サンプルを増やし十分な情報を集めることができれば，標本平均を通じて「正規分布」という一つの確率分布で現象を解析できることを示している．その意味で，中心極限定理は確率論および統計学において最も重要な定理である．

さらに，中心極限定理（定理 3.1）と正規確率変数の標準化（命題 3.8）を組み合わせると以下の結果を得る．

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1) \ (n \to \infty) \tag{3.20}$$

すなわち，標本平均を標準化した Z の確率分布は $n \to \infty$ となるとき期待値 0，分散 1 の標準正規分布に収束する．したがって (3.20) は，いささか大げさに言えば「宇宙で唯一つ」の標準正規分布という確率分布の性質を十分解明すれば，およそあらゆる確率的現象を解析できることを意味している．これは非常に強力な結果である．

中心極限定理を用いる際，よく聞かれる質問は「サンプル数がいくつ以上ならば，中心極限定理を使えるのか？」というものである．理論的にいくつ以上なら良いと答えることは難しいが，研究者の間の最大公約数的な意見は

中心極限定理を適用するには $n \geq 30$

というものである．するとすぐ返ってくる質問は「では $n = 29$ のときは，どう

するのか？」というものであるが，これはケース・バイ・ケースとしか答えられない．筆者の経験では $n \leq 10$ ではいかにも少なすぎる．$n \geq 20$ なら，中心極限定理の適用を考えてもよいかな，といったものである．

では，サンプル数が少ないときはどうすればよいのだろうか．サンプルが正規分布に従うと仮定できれば，命題 3.18 を用いて問題ない．サンプルが正規分布に従うかどうかは，QQ-norm plot (p. 65) を用いて調べればよい．サンプルが少なく正規分布にも従わないときは，後述するノンパラメトリック法という別の解析法を用いることが出来る場合がある．

例題 3.15. 脳性ナトリウム利尿ペプチド *(brain natriuretic peptide, BNP)* は心臓から分泌されるホルモンで，心不全のマーカーとしてよく用いられる．*BNP* は必ず正の値をとり，その分布は右に歪んでおり正規分布には従わないことが知られている．（例題 *2.2, (p. 28)* 参照）ある心不全患者集団の *BNP* の平均 *=195.9*，標準偏差*=292.4* とする．このとき，この心不全患者集団から抽出した *1000* 人のサンプルの標本平均が，*(190, 200)* の範囲に含まれる確率を求める．

まず，*BNP* の分布は強く右に歪んでおり左右対称でないことから，正規分布には従わないことがわかる．しかし，サンプル数は $n = 1000$ と十分大きいので，中心極限定理を適用可能である．中心極限定理により，標本平均 $\bar{X}$ の確率分布は期待値 $\mu = 195.9$，標準偏差 $\sigma/\sqrt{n} = 292.4/\sqrt{1000} = 9.247$ の正規分布で近似される．

$$
\begin{aligned}
P(190 < \bar{X} < 200) &= P\left(\frac{190 - 195.9}{9.247} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{200 - 195.9}{9.247}\right) \\
&\approx P(-0.638 < Z < 0.443) = P(Z < 0.443) - P(Z < -0.638) \\
&= 0.671 - 0.262 = 0.409
\end{aligned}
$$

第 *2* 行目が等号 *"="* ではなく近似記号 *"≈"* で結ばれていることに留意する．これは第 *1* 行目と第 *2* 行目は正確に等しいのではなく，中心極限定理による近似が行われていることを示している．したがって，サンプル数が少なく元の確率分布が正規分布から離れているほど，この近似は不正確になる．正規分布の累積分布関数の値は *R* の *pnorm()* コマンドで求めた．

```
> pnorm((190 - 195.9)/9.247)
```

```
[1] 0.2617223
> pnorm((200 - 195.9)/9.247)
[1] 0.6712571
> pnorm((200 - 195.9)/9.247) - pnorm((190 - 195.9)/9.247)
[1] 0.4095348
> pnorm(200, mean=195.9, sd=9.247) - pnorm(190, mean=195.9, sd=9.247)
[1] 0.4095348
```

上のように，*pnorm()* コマンドで *mean* オプション，*sd* オプションを使っても同じである．

3.4.6 カイ二乗 (χ^2, chi-squared) 分布，t 分布，F 分布

本章の最後に，本書後半でもよく用いられる正規分布に関連した確率分布，カイ二乗 (χ^2, chi-squared) 分布，t 分布，F 分布を紹介する．これらの確率分布はその確率密度関数を用いて定義されるが，いずれも大変に複雑な形をしている．しかし，それらの確率密度関数自体を計算したりする必要はないので，心配無用である．それよりもそれぞれの確率分布の性質や使い方を理解することのほうがはるかに大切である．

3.4.6.1 カイ二乗 (χ^2, chi-squared) 分布

定義 3.35. 連続確率変数 X の確率密度関数が以下の $f_{\chi^2}(x;\nu)$ で与えられるとき，X は**自由度** ***(degrees of freedom)*** ν の**カイ二乗** ***(χ^2, chi-squared)*** **分布**に従うといい，$X \sim \chi^2(\nu)$ と記す．

$$f_{\chi^2}(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}},\ 0 < x < \infty,\ \nu = 1, 2, 3, \ldots,$$
$$\Gamma(a) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

カイ二乗分布は，以下のような特徴を持つ．

定理 3.2 (χ^2 分布の性質)．X を自由度 ν の χ^2 分布に従う確率変数とする．$X \sim \chi^2(\nu)$

1. カイ二乗分布は非負の確率分布である．$X > 0$.

2. カイ二乗分布の確率密度関数の形状は，右に歪んだ分布である．

3. 期待値 $\mu = \nu$，分散 $\sigma^2 = 2\nu$.

図 3.11 にカイ二乗分布の確率密度関数を示す．実線が自由度 5，点線が自由度 10 のカイ二乗分布の密度関数である．定理 3.2, 1, 2 にある通り，カイ二乗分布の確率密度関数は，非負かつ右に歪んでいることが分かる．また，定理 3.2, 3 にある通り，自由度が大きくなるにつれ分散が大きくなることが分かる．

図 3.11 χ^2 分布の確率密度関数

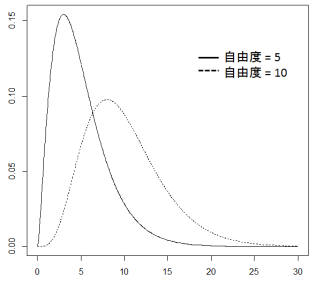
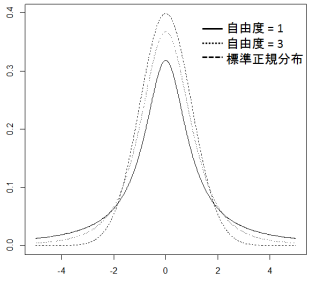


図 3.12 t 分布の確率密度関数



χ^2 分布と正規分布の関係は，以下の定理 3.3 で示される．

定理 3.3. $X_1, X_2, \ldots, X_n$ を独立かつ同一の標準正規分布 $N(0,1)$ に従う確率変数とする．このとき $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ は $\chi^2(n)$ に従う．

χ^2 分布については，以下の定理 3.4 も重要である．

定理 3.4. $X_1, X_2, \ldots, X_n$ を独立かつ同一の正規分布 $N(\mu, \sigma)$ に従う確率変数とする．このとき標本平均 $\bar{X}$ と標本分散 S^2 は独立で，

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

は自由度 $(n-1)$ の χ^2 分布，$\chi^2(n-1)$ に従う．

定理 3.4 は，正規母集団の分散の推測を行う際などに用いられる．また以下に述べる t 分布の性質 (Theorem 3.7) を導くのにも用いられる．定理 3.4 では，正規確率変数の場合 $\bar{X}$ と S^2 が独立であることが示されるが，S^2 の定義に $\bar{X}$ が使われていることを考えると，これは驚くべきことといえる．このような性質は，母集団分布が正規分布である場合に特徴的なものである．

3.4.6.2 t 分布

定義 3.36. 連続確率変数 X の確率密度関数が以下の $f_T(x;\nu)$ で与えられるとき，X は**自由度** ***(degrees of freedom)*** ν の t **分布**に従うといい，$X \sim t(\nu)$ と記す．

$$f_T(x;\nu) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}},\ -\infty < x < \infty,\ \nu = 1, 2, 3, \ldots,$$

t 分布は，以下のような性質を持つ．

定理 3.5 (t 分布の性質)．X を自由度 ν の t 分布に従う確率変数とする．$X \sim t(\nu)$

1. t 分布の確率密度関数の形状は，原点 0 を中心として左右対称のベル型 *(bell-shaped)* である．
2. $E(X) = 0, \nu > 1$. $V(X) = \frac{\nu}{\nu-2}, \nu > 2$.
3. 自由度 ν が $\nu \to \infty$ となるとき，t 分布の確率密度関数は標準正規分布の密度関数に収束する．

図 3.12 に t 分布の確率密度関数を示す．実線が自由度 1，鎖線が自由度 3 の t 分布の密度関数であり，破線が標準正規分布の密度関数である．図 3.12 に見られるとおり，t 分布の確率密度関数の形状は標準正規分布のそれによく似ているが，t 分布の密度関数の方が分散が大きく裾が重い．自由度 1 の t 分布は，期待値，分散共に無限に発散して存在せず，自由度 2 の t 分布は期待値は存在して $E(X) = 0$ であるが分散は存在しないという，いささか変わった確率分布である．定理 3.5, 3 にある通り，t 分布は自由度が大きくなるにしたがって標準正規分布に収束する性質を持つ．t 分布においては，以下の二つの定理が重要である．

定理 3.6. X と Y を独立な確率変数とする．X は標準正規分布 $N(0,1)$，Y は自由度 ν の χ^2 分布 $\chi^2(\nu)$ に従うとする．このとき

$$T = \frac{X}{\sqrt{Y/\nu}}$$

とすると，T は自由度 ν の t 分布 $t(\nu)$ に従う．

定理 3.7. $X_1, X_2, \ldots, X_n$ を独立かつ同一の正規分布 $N(\mu, \sigma^2)$ に従う確率変数とする．$\bar{X}$ を標本平均，S^2 を標本分散とする．このとき

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

とすると，T は自由度 $(n-1)$ の t 分布 $t(n-1)$ に従う．

定理 3.7 は，正規母集団の期待値の推測を行う際に用いられる．

3.4.6.3 F 分布

定義 3.37. 連続確率変数 X の確率密度関数が以下の $f_F(x; \nu_1, \nu_2)$ で与えられるとき，X は**分子の自由度** ***(numerator degrees of freedom)*** ν_1 と**分母の自由度** ***(denominator degrees of freedom)*** ν_2 の F **分布**に従うといい，$X \sim F(\nu_1, \nu_2)$ と記す．

$$f_F(x; \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right) \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_1 x + \nu_2)^{\frac{\nu_1+\nu_2}{2}}},\ 0 < x < \infty, m, n = 1, 2, 3, \ldots.$$

F 分布においては，以下の二つの定理が重要である．

定理 3.8. X と Y は互いに独立で，X は自由度 ν_1 の χ^2 分布 $\chi^2(\nu_1)$，Y は自由度 ν_2 の χ^2 分布 $\chi^2(\nu_2)$ に従うとする．このとき

$$F = \frac{X/\nu_1}{Y/\nu_2} = \frac{\nu_2 X}{\nu_1 Y}$$

とすると，F は自由度 ν_1, ν_2 の F 分布 $F(\nu_1, \nu_2)$ に従う．

定理 3.9. $X_1, X_2, \ldots, X_m$ を独立に正規分布 $N(\mu_1, \sigma_1^2)$ に従う確率変数とす

る．$Y_1, Y_2, \ldots, Y_n$ を独立に正規分布 $N(\mu_2, \sigma_2^2)$ に従う別の確率変数とする．$(X_1, X_2, \ldots, X_m)$ と $(Y_1, Y_2, \ldots, Y_n)$ は，互いに独立であるとし，S_1^2，S_2^2 をそれぞれ X と Y の標本分散とする．このとき

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

とすると，F は自由度 $(\nu_1 = m-1, \nu_2 = n-1)$ の F 分布 $F(m-1, n-1)$ に従う．

3.4.6.4 χ^2 分布，t 分布，F 分布に関する R コマンド

R には正規分布の場合と同様に，χ^2 分布，t 分布，F 分布についても確率密度関数 (pdf)，累積分布関数 (cdf)，クォンタイル（quantile, 分位数，パーセント点）関数，およびそれぞれの分布に従う乱数を生成するコマンドが用意されている．それぞれ `d***()`，`p***()`，`q***()`，`r***()` という形式のコマンド名を持ち，“`***`” の部分には χ^2 分布，t 分布，F 分布を表す “`chisq`”, “`t`”, “`f`” の文字が入る．χ^2 分布と t 分布の自由度は `df` オプションで指定する．F 分布の場合，分子の自由度を `df1` オプションで，分母の自由度を `df2` オプションで指定する．`log`，`log.p` オプションと `lower.tail` オプションは p. 62 の正規分布に関する R コマンドの場合と同様である．各コマンドの `ncp` オプションは「非心度パラメター (noncentrality parameter)」と呼ばれるものを指定するオプションであるが，本書では扱わない．

χ^2 分布

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

t 分布

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
```

```
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

F 分布

```
df(x, df1, df2, ncp, log = FALSE)
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
rf(n, df1, df2, ncp)
```

表 3.1 第 3 章の R コマンドリスト

コマンド名	目的	使い方
dnorm()	正規分布の確率密度関数	dnorm(x)
pnorm()	正規分布の累積分布関数	pnorm(q)
qnorm()	正規分布のクォンタイル関数	qnorm(p), qnorm(0.025) qnorm(0.025, lower.tail=FALSE)
rnorm()	正規分布の乱数生成	rnorm(n), rnorm(100)
seq(a,b,by=c) seq(a,b,length=n)	a から b までの c 間隔の数列を生成する．length=n なら長さ n の等差数列	seq(-4, 4, 0.1) seq(-4, 4, length=81)

第4章　推定，信頼区間，仮説検定

第 1 章 1.2 節で考えたとおり，データ解析の目的は，解析対象である母集団を特徴付けるパラメターについて，何らかの推論を行うことである．しかし一般に母集団に関するすべての情報を得ることは難しいため，母集団の一部分を標本として抽出し，その観測値であるデータから統計量の値を計算してパラメターに関する推論を行う，というのがデータ解析の手順であった（図 1.2 参照, p. 15）．標本から計算される変数 (variable) は確率論における確率変数 (random variable) に相当し，観測されるより前には値が定まらないという不確実性がある．この不確実性が変数の確率分布としてモデル化され，その理論的枠組みを担うのが第 3 章で取り上げた確率論であった．

本章では，観察されたデータに基づく具体的な推論を行う方法論として，点推定，区間推定および仮説検定の一般論を述べる．その例として，最も基本的な一つの母集団の期待値に関する推測（一標本問題）についても解説する．

4.1　点推定

統計的推測あるいはデータ解析は，ほぼ常に母集団における一つ若しくは複数のパラメターの推測を目的としている．パラメターの値は母集団上の変数（確率変数）の確率分布を規定している．たとえば，母集団上の変数が正規分布 $N(\mu, \sigma^2)$ に従うならば，期待値 μ と分散 σ^2 という二つのパラメターの値を求めれば，母集団分布としての正規分布の形を特定できる (命題 3.6，p. 60)．

いま，推定対象のパラメターを一般に θ と記す．母集団平均 μ が推定対象なら $\theta = \mu$，分散 σ^2 が推定対象なら $\theta = \sigma^2$ といった具合である．また，θ を推定するための統計量（あるいは，以下に定義する推定量）を，一般に $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ と記す．母集団平均 μ を推定するのに標本平均 $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ を用いるのであれば，$\hat{\theta} = \bar{X}$ となる．統計量 $\hat{\theta}$ は母集団から抽出されたランダム標本 $X_1, \ldots, X_n$ から求められるが，$X_1, \ldots, X_n$ は n 個の確率変数であるから $\hat{\theta}$ もまた確率変数である．X の観測値 $X = x$ が定まったのち，初めて $\hat{\theta}$ の観測値も一つの実数値として確定する．

定義 4.1. 母集団のパラメター θ を，推定の対象とする．この時，パラメター θ を推定するための統計量を $\hat{\theta}$ と記し，θ の**点推定量** ***(point estimator)*** あるいは単に**推定量** ***(estimator)*** という．推定量の観測値を**点推定値** ***(point estimate)***，あるいは単に**推定値** ***(estimate)*** という．

前述したとおり推定量は確率変数であるのに対して，推定値は例えば $\hat{\theta} = 0.5$ といったように確定したデータから計算された一つの実数値であることに注意する．また，「推定対象」という言葉を **estimand** と称することもある．

推定量 $\hat{\theta}$ は確率変数であるから，その期待値 $E(\hat{\theta})$，分散 $V(\hat{\theta})$ も定義される．$E(\hat{\theta})$ は $\hat{\theta}$ の確率分布の中心であるから，$E(\hat{\theta}) = \theta$ が成り立つならば $\hat{\theta}$ は θ を「偏りなく」推定していることになる．

定義 4.2. パラメター θ の推定量 $\hat{\theta}$ は，任意の可能な θ に対して

$$E(\hat{\theta}) = \theta$$

が成り立つとき，θ の**不偏推定量** ***(unbiased estimator)*** という．また，

$$b(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta$$

を，θ に対する $\hat{\theta}$ の**バイアス** ***(bias)*** という．

命題 4.1. $X_1, X_2, \ldots, X_n$ を，互いに独立で共通の期待値 μ と分散 $\sigma^2 < \infty$ を持つ確率変数の列とする．このとき，標本平均 $\bar{X}$ と標本分散 S^2 は，それぞれ

母集団平均 μ と母集団分散 σ^2 の不偏推定量である．

Proof. まず，

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

ここで，期待値の性質 2 (p. 80) を用いた．$E(\bar{X}) = \mu$ であるから，$E(\bar{X})$ は μ の不偏推定量である．次に，

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\sum_{i=1}^{n} X_i + n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right] \end{aligned}$$

命題 3.3 (p. 57), 2 より $V(Y) = E(Y^2) - [E(Y)]^2 \iff E(Y^2) = V(Y) + [E(Y)]^2$ であったから，Y に X_i あるいは $\sum X_i$ を代入して，以下を得る．

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\left[\sum E(X_i^2) - \frac{1}{n}E(\sum X_i)^2\right] \\ &= \frac{1}{n-1}\left[\sum(\sigma^2 + \mu^2) - \frac{1}{n}\left(V(\sum X_i) + [E(\sum X_i)]^2\right)\right] \\ &= \frac{1}{n-1}\left[(n\sigma^2 + n\mu^2) - \frac{1}{n}(n\sigma^2 + [n\mu]^2)\right] = \frac{1}{n-1}[n\sigma^2 - \sigma^2] = \sigma^2. \end{aligned}$$

上で $V(\sum X_i) = n\sigma^2, E(\sum X_i) = n\mu$ を示すのには，命題 3.16 (p. 85) を用いた． □

一般に**点推定量**（あるいは単に**推定量**）は標本を基に計算され，推定対象のパラメター θ の値についての「良い」推測を与えるものである．命題 4.1 にある通り，標本平均 $\bar{X}$ と標本分散 S^2 は，それぞれ母集団平均 μ と母集団分散 σ^2 の不偏推定量であるという「良い」性質を持っていた．しかし，任意のパラメター θ に対して，その推定量 $\hat{\theta}$ が自明な形で与えられるとは限らない．

そこでここでは，標本から点推定量を「構成」するための二つの方法，**積率法 (the method of moments)** と**最尤法 (the method of maximum likelihood)** について述べる．

4.1.1 積率法 (The Method of Moments)

定義 4.3. $X_1, X_2, \ldots, X_n$ を，確率関数あるいは確率密度関数 $f(x)$ に従う確率変数とする．このとき，$k = 1, 2, 3, \ldots$ に対して，$E(X^k)$ を k **次母集団積率** ***(k-th population moment)***, $(1/n)\sum_{i=1}^{n} X_i^k$ を k **次標本積率** ***(k-th sample moment)*** という．

一般に確率変数の母集団積率は，母集団分布のパラメターの関数として与えられる．積率法とは，標本積率を同じ次数の母集団積率と等置して，未知パラメターについて解くことで推定量を得る方法である．

定義 4.4. $X_1, X_2, \ldots, X_n$ を，確率関数あるいは確率密度関数 $f(x; \theta_1, \ldots, \theta_k)$ に従う確率変数とする．ただし，$\theta_1, \ldots, \theta_k$ は k 個の未知パラメターとする．このとき**積率推定量** ***(moment estimator)*** $\hat{\theta}_1, \ldots, \hat{\theta}_k$ とは，k 次までの標本積率を同じく k 次までの母集団積率と等置して $\theta_1, \ldots, \theta_k$ について解いたものである．

例題 4.1. $X_1, X_2, \ldots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$ を，期待値 μ，分散 σ^2 を持つ確率変数とする．期待値の定義と分散の性質（命題 *3.3, (p. 57)*）から

$$\mu = E(X^1),\ \sigma^2 = E(X^2) - [E(X^1)]^2.$$

1 次と *2* 次の標本積率 $(1/n)\sum X_i, (1/n)\sum X_i^2$ を，それぞれ $E(X^1), E(X^2)$ に代入すると以下を得る．

$$\begin{aligned}
\hat{\mu} &= (1/n)\sum_{i=1}^{n} X_i = \bar{X} \\
\hat{\sigma}^2 &= (1/n)\sum_{i=1}^{n} X_i^2 - \left((1/n)\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_i\right) + n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}\left(\sum_{i=1}^{n} X_i\right) + n\bar{X}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2\bar{X}X_i + \sum_{i=1}^{n} \bar{X}^2\right)
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2\bar{X}X_i + \bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

したがって期待値 μ と分散 σ^2 の積率推定量は，以下のように与えられる．

$$\hat{\mu} = \bar{X},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

ここで分散の積率推定量 $\hat{\sigma}^2$ は，第 *2.2.2.1* 節 *(p. 24)* で定義されこれまで用いてきた標本分散

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

とは形が異なることに注意する．命題 *4.1* で示した通り標本分散 S^2 は母集団分散 σ^2 の不偏推定量であったから，

$$S^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n}{n-1}\hat{\sigma}^2 > \hat{\sigma}^2 \ \Rightarrow E(S^2) = \sigma^2 > E(\hat{\sigma}^2)$$

したがって，分散の積率推定量は負のバイアスを持つ推定量であることが分かる．

4.1.2 最尤法 (The Method of Maximum Likelihood)

前節で考えた積率法は，パラメターが母集団積率で表されることを利用した推定方法であった．しかし，積率法が使えるためにはパラメターが積率について明示的に「解ける」ことが必要であるが，それは常に可能とは限らない．

本節では，点推定量を構成するためのもう一つの方法である「最尤法」について検討する．いま，$X_1, \ldots, X_n$ が同時確率関数 $p(x_1, \ldots, x_n; \theta_1, \ldots, \theta_k)$ を持つ離散確率変数とする．ただし，$\theta_1, \ldots, \theta_k$ は k 個の未知パラメターとする．同時確率関数は，与えられたデータ $(X_1 = x_1, \ldots, X_n = x_n)$ に対してそれが得られる確率を返す関数である（定義 3.30, (p. 78)）．

$$p(x_1, \ldots, x_n; \theta_1, \ldots, \theta_k) = P(X_1 = x_1, \ldots, X_n = x_n)$$

したがって，与えられた $(x_1, \ldots, x_n)$ に対して同時確率関数を最大にする $(\theta_1, \ldots, \theta_k)$ は，現在手元にあるデータが観察できる確率を最大にする「最も

尤もらしい」パラメターの推定値と考えられる．この考え方を連続確率変数の場合にも適用して，「最尤法」による推定の概念を以下のように定義する．

定義 4.5. $X_1, \ldots, X_n$ を，以下の同時確率関数 *(joint pmf)* もしくは同時確率密度関数 *(joint pdf)* を持つ確率変数とする．

$$f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_k) \tag{4.1}$$

ただし，$\theta_1, \ldots, \theta_k$ は k 個の未知パラメターとする．与えられた確率変数の観察値 $(x_1, \ldots, x_n)$ に対して，*(4.1)* を $(\theta_1, \ldots, \theta_k)$ の関数としてみたものを**尤度関数** ***(likelihood function)*** と言い，

$$L = L(\theta_1, \ldots, \theta_k) = f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_k)$$

と記す．$(x_1, \ldots, x_n)$ に対して尤度関数を最大化する $(\theta_1, \ldots, \theta_k)$ を $(\hat{\theta}_1, \ldots, \hat{\theta}_k), \hat{\theta}_j = \hat{\theta}_j(x_1, \ldots, x_n),\ j = 1, \ldots, k$ とする．

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_k) \geq f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_k),\ \forall \theta_j, j = 1, \ldots, k$$

このとき，各 x_i に X_i を代入した $\hat{\theta}_j(X_1, \ldots, X_n)$ を θ_j の**最尤推定量** ***(Maximum Likelihood Estimator, MLE)*** という．尤度関数を対数変換したものを，**対数尤度関数** ***(log likelihood function)*** と言い $\log L$ と記す．

対数関数 log は単調増加関数であるから，尤度関数を最大化することは対数尤度関数を最大化することと同値である．

例題 4.2. $X_1, \ldots, X_n$ が独立かつ同一に正規分布 $N(\mu, \sigma^2)$ に従うとする．

$$\begin{aligned} L(\mu, \sigma^2) = f(x_1, \ldots, x_n; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

このとき $\sigma^2 = \eta > 0$ とすると，対数尤度関数 $\log L$ は以下のようになる．

$$\log L = \log[f(x_1, \ldots, x_n; \mu, \eta)] = -\frac{n}{2}\log(2\pi\eta) - \frac{1}{2\eta}\sum(x_i - \mu)^2 \tag{4.2}$$

$\log L$ を最大化する最尤推定量 $\hat{\mu}, \hat{\eta} = \hat{\sigma}^2$ を求めるため，*(4.2)* を μ, η について微分した偏導関数を *0* と置く．

$$\begin{aligned}
\frac{\partial}{\partial \mu} \log L &= -\frac{1}{2\eta}(-2)\sum(x_i - \mu) = 0 \\
&\Rightarrow \sum(x_i - \mu) = 0 \Rightarrow \sum x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n}\sum X_i = \bar{X} \\
\frac{\partial}{\partial \eta} \log L &= -\frac{n}{2}\frac{1}{2\pi\eta}(2\pi) + \frac{1}{2}\frac{1}{(\eta)^2}\sum(x_i - \mu)^2 = 0 \\
&\Rightarrow -n + \frac{1}{\eta}\sum(x_i - \mu)^2 = 0 \Rightarrow \hat{\eta} = \hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2
\end{aligned}$$

したがって正規分布の期待値 μ と分散 η の最尤推定量は，以下のように与えられる．

$$\hat{\mu} = \bar{X},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

本節のはじめに離散確率変数を例に最尤法の考え方を導入したが，連続確率変数の場合その動機付けはいささか直観的なものであった．しかし，少なくともサンプル数が十分大きいとき，最尤推定量には多くの優れた性質があり，以下に証明なしにそれを述べる．（河田，丸山，鍋谷[7]，Cramér[3], Casella and Berger[1] などを参照）

定理 4.1. *[最尤推定量の性質]* $X_1, \ldots, X_n$ を確率関数あるいは確率密度関数 $f(x;\theta)$ に従う独立な確率変数列とする．$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ を，パラメター θ の最尤推定量とする．このとき，いくつかの条件の下で以下が成立する．

一致性 (consistensy) 任意の $\epsilon > 0$ に対して，

$$\lim_{n\to\infty} P_\theta(|\hat{\theta}_n - \theta| > \epsilon) = 0 \tag{4.3}$$

(4.3) が成り立つ．このとき最尤推定量 $\hat{\theta}_n$ が θ に**確率収束** ***(convergence in probability)*** するといい，$\hat{\theta}_n \xrightarrow{P} \theta$ と記す．$\hat{\theta}_n$ は推定量であるから，確率変数であり確率分布を持つ．*(4.3)* はサンプル数 n が大きくなるに従い，$\hat{\theta}_n$ の分布が推定対象である θ の近傍に収束することを示している．

漸近不偏性 (asymptotic unbiasedness)

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta \tag{4.4}$$

もし $E(\hat{\theta}_n) = \theta$ が成り立つならば，最尤推定量 $\hat{\theta}_n$ は θ の不偏推定量であることを意味する．一般に最尤推定量は不偏推定量であるとは限らないが，*(4.4)* はサンプル数 n が大きくなるに従いバイアス $(E(\hat{\theta}_n) - \theta)$ が *0* に収束することを示している．

漸近正規性 (asymptotic normality)

$$\theta_n \to N(\theta, 1/(nI(\theta))),\ n \to \infty \tag{4.5}$$
$$I(\theta) = E_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(X;\theta) \right)^2 \right]$$

すなわち，サンプル数 n が十分に大きければ最尤推定量の確率分布は正規分布で近似される．また，収束先の正規分布の分散に現れる $I(\theta)$ を，**フィッシャー情報量** ***(Fisher Information)*** と呼ぶ．

漸近有効性 (asymptotic efficiency) パラメター θ の任意の不偏推定量を $\tilde{\theta}$ とするとき，以下が成立する．

$$Var(\tilde{\theta}) \geq \frac{1}{nI(\theta)} \tag{4.6}$$

(4.6) を**クラメール＝ラオの不等式** ***(Cramér-Rao inequality)*** という（尾畑[9]，柳川[13]，竹村[11] などを参照）．最尤推定量の漸近正規性 *(4.5)* は同時に最尤推定量 $\hat{\theta}$ の分散 $Var(\hat{\theta})$ が，サンプル数 n が大きくなるに従い θ のあらゆる不偏推定量の分散の下限を達成することを示している．

最尤推定量の一致性，漸近不偏性と合わせて考えると，サンプル数 n が大きくなるとき最尤推定量 $\hat{\theta}$ はバイアスが *0* に収束し，かつ不偏推定量の中で可能な限り小さい分散で（=最も精確に）推定対象のパラメター θ を推定することを意味している．

不変性 (invariance) $g(\theta) = \eta$ を θ の *1* 対 *1* 関数とする．このとき，$g(\hat{\theta})$ は $g(\theta)$ の最尤推定量である．

例えば，例題 *4.2* によれば，$X_1, \ldots, X_n$ が独立かつ同一に正規分布 $N(\mu, \sigma^2)$ に従うとき，分散 σ^2 の最尤推定量は $\hat{\sigma}^2 = (1/n)\sum(X_i - \bar{X})^2$ で与えられ

た．標準偏差 σ は分散の平方根であるから，σ の最尤推定量 $\hat{\sigma}$ は $\hat{\sigma}^2$ の平方根で与えられる．

$$\hat{\sigma} = \sqrt{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

本書後半で取り上げる，回帰モデル，ロジスティック回帰モデル，生存時間解析などの統計モデルのパラメター一は，すべて最尤法もしくはそれに関連した方法で導出された推定量で推定される．従って，それらの推定量は全て定理 4.1 にある優れた性質を有していることになる．

4.2 信頼区間

前節では，（点）推定量を用いて推定対象のパラメターを推定する方法を議論した．点推定においては，与えられたデータから計算した推定値をもってパラメターの値を数値的に評価することができる．しかし推定値そのものは一個の具体的な実数に過ぎず，推定値の値それ自体は推定の正確さ，信頼度については何の情報も与えない．

例えば，A,B 二つの調査機関が 10 歳児の平均身長を推定したとする．調査機関 A は 10 人の子供の身長を測定し，平均 138.5cm の推定値を得たとする．一方調査機関 B は 10,000 人の身長を測定し，平均 139.7cm の推定値を得たとしよう．このとき A の推定値 138.5cm と B の推定値 139.7cm の，どちらが信頼に足るであろうか．常識に従えば，B の方が多くのサンプルを基に推定しているから，B の結果の方が信頼できそうである．しかし二つの点推定の結果は，138.5 という一つの実数値と 139.7 という別の実数値のみであり，推定値の値そのものからどちらがどれだけ精確かを評価することはできない．

推定の正確さを評価するには，推定の信頼度を示すような，「尤もらしい (plausible) 」推定値の区間（集合）が有用である．

例題 4.3 (一標本問題：正規母集団既知分散)．いま，$X_1, X_2, \ldots, X_n$ が独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うとする．簡単化のために，母集団分散 σ^2 は既知であると仮定する．期待値 μ が未知であるにもかか

わらず，分散 σ^2 が既知であると仮定することはもちろん非現実的であるが，この仮定は後で検討することにする（第 *4.2.4* 節 *p. 114* 参照）．

さて，期待値 μ を推定する場合，最も自然な推定量は標本平均 $\bar{X}$ である．$X_1, X_2, \ldots, X_n$ が独立に $N(\mu, \sigma^2)$ に従うとき，命題 *3.18 (p. 86)* により

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

正規確率変数の標準化（命題 *3.8 (p. 62)*）によれば，標本平均を標準化した Z は期待値 *0*，分散 *1* の標準正規分布 $N(0,1)$ に従う．

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

いま，$z_{\alpha/2}$ を標準正規分布の上側 $\alpha/2$ パーセント点とすれば，以下の式が成り立つ．（図 *4.1* 参照．確率変数のパーセント点については，定義 *3.15*（*p. 58*），標準正規分布の上側パーセント点 z_α については *p. 64* を参照）

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha \tag{4.7}$$

ここで *(4.7)* の左辺括弧内の不等式を，推定対象の μ を挟むように変形することを考える．

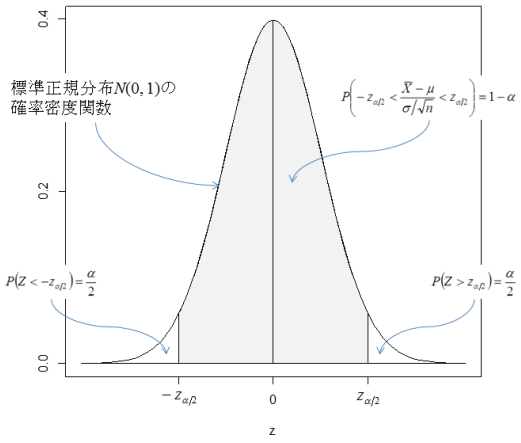
$$\begin{aligned}
& -z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} && \text{(4.7) 左辺括弧内の不等式} \\
\iff & -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}} && \text{各項に } \sigma/\sqrt{n} \text{ を掛ける} \\
\iff & -\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} && \text{各項から } \bar{X} \text{ を引く} \\
\iff & \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} && \text{各項に-1 を掛ける}
\end{aligned} \tag{4.8}$$

上の式変形は論理的に同等であるから，*(4.7)* 左辺の括弧内の不等式が成り立つなら *(4.8)* も成り立つ．したがって（ *(4.8)* の左右を入れ替えて）以下を得る．

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.9}$$

(4.9) を解釈するために，以下の区間を考える．

図 4.1


標準正規分布$N(0,1)$の
確率密度関数
$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$
$P(Z < -z_{\alpha/2}) = \frac{\alpha}{2}$
$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$
0.4
0.2
0.0
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
z

$$(L(X_1,\ldots,X_n),\ U(X_1,\ldots,X_n)) = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad (4.10)$$

(4.10) は，両端にそれぞれ $L(X_1,\ldots,X_n) = \bar{X} - z_{\alpha/2}(\sigma/\sqrt{n})$, $U(X_1,\ldots,X_n) = \bar{X} + z_{\alpha/2}(\sigma/\sqrt{n})$ という二つの確率変数を持ち，$\bar{X}$ を中心に左右 $z_{\alpha/2}(\sigma/\sqrt{n})$ の固定された幅を持つ「ランダム」な区間である．（*(4.10)* の両端が $X_1,\ldots,X_n$ に依存することを強調するために，$L(X_1,\ldots,X_n)$, $U(X_1,\ldots,X_n)$ という表記を用いた．）このとき，*(4.9)* は「ランダムな区間 *(4.10)* が真の μ を含む確率は $(1-\alpha)$ である」と解釈できる．

さて，標本平均 $\bar{X}$ は期待値 μ の推定量であり確率変数であるから，$X_1,\ldots,X_n$ の観測値が確定するまで $\bar{X}$ の値は定まらない．サンプルの値が $(X_1 = x_1,\ldots,X_n = x_n)$ のように観測されたとき，$\bar{X}$ の観測値 $\bar{x}$ が計算され *(4.10)* の両端点も以下のように実数の観測値として定まる．

$$(L(x_1,\ldots,x_n),\ U(x_1,\ldots,x_n)) = \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad (4.11)$$

例題 4.3 に従い，(4.10) のような区間によるパラメターの推定，すなわちパラメターの区間推定による「信頼区間」を以下のように定義する．

定義 4.6. 推定対象のパラメター θ に対して，以下の条件を満たす確率的な区間（二つの確率変数の組）$(L(X_1,\ldots,X_n), U(X_1,\ldots,X_n))$ を考える．

$$P(L(X_1,\ldots,X_n) \leq \theta \leq U(X_1,\ldots,X_n)) \geq (1-\alpha) \quad (4.12)$$

このとき，ランダムな区間 $(L(X_1,\ldots,X_n),\ U(X_1,\ldots,X_n))$ をパラメター θ の**区間推定量** ***(interval estimator)*** という．区間推定量の $(X_1,\ldots,X_n)$ に観測値 $(x_1,\ldots,x_n)$ を代入した以下の区間 *(4.13)* を，θ に対する**信頼水準** ***(confidence level)*** $(1-\alpha)$ の**信頼区間** ***(Confidence Interval, CI)***（あるいは θ の $100\times(1-\alpha)\%$信頼区間）と言う．

$$(L(x_1,\ldots,x_n),\ U(x_1,\ldots,x_n)) \quad (4.13)$$

また，$L(x_1,\ldots,x_n)$, $U(x_1,\ldots,x_n)$ をそれぞれ**下側信頼限界** ***(lower confidence***

bound)*，上側信頼限界 *(upper confidence bound) という．

(4.12) では区間推定量の端点となる $L(X_1,\ldots,X_n), U(X_1,\ldots,X_n)$ として確率変数のペアを考えているが，「片側」区間推定量を考えることもある．例えば $L(X_1,\ldots,X_n)=-\infty$ とすれば左片側区間 $(-\infty, U(X_1,\ldots,X_n))$ を得るし，$U(X_1,\ldots,X_n)=\infty$ とすれば反対の右片側区間 $(L(X_1,\ldots,X_n),\infty)$ を得る．このとき，(4.12) と同様に

$$P(L(X_1,\ldots,X_n)\leq\theta)\geq(1-\alpha)\quad \text{あるいは}\quad P(\theta\leq U(X_1,\ldots,X_n))\geq(1-\alpha)$$

が成立するとき，

$$(L(x_1,\ldots,x_n),\infty)\quad \text{あるいは}\quad (-\infty,U(x_1,\ldots,x_n))$$

を信頼水準 $(1-\alpha)$ の**片側信頼区間 (one-sided confidence interval)** という．（同様に，(4.13) を**両側信頼区間 (two-sided confidence interval)** と呼ぶ．）

信頼水準 $(1-\alpha)$ は理論的には任意であるが，通常よく用いられるのは $1-\alpha=0.95, 0.99, 0.90$ などである．正規母集団の場合，それぞれの信頼水準に対応する標準正規分布 $N(0,1)$ の上側 $\alpha/2$ パーセント点を表 4.1 に示した．

表 4.1 信頼水準 $(1-\alpha)$ と標準正規分布の上側パーセント点

信頼水準 $(1-\alpha)$	0.90	0.95	0.99
$z_{\alpha/2}$	1.645	1.960	2.576

なお，$z_{\alpha/2}$ の計算には R の `qnorm()` コマンドを用いた．

```
> a1 <- 0.1; a2 <- 0.05; a3 <- 0.01
> qnorm(c(a1/2, a2/2, a3/2), lower.tail=F)
[1] 1.644854 1.959964 2.575829
```

4.2.1 一標本問題：正規母集団既知分散の場合の信頼区間

例題 4.3 のように，標本 $X_1, X_2, \ldots, X_n$ が同一の確率分布に従うとき，その確

率分布の期待値，分散その他に関して推測することを「一標本問題 (One sample problem)」という．分散既知の正規母集団における一標本問題の場合，期待値 μ の信頼区間は例題 4.3 (4.11) で与えられる．信頼水準 $(1-\alpha)=0.95$ の場合，標準正規分布の上側 $\alpha/2$%点は $z_{\alpha/2}=1.96$（表 4.1）であるから,μ の 95%信頼区間は以下のように与えられる．

$$\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}},\ \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) \tag{4.14}$$

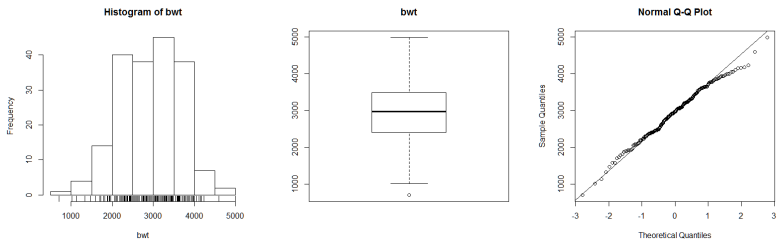
例題 4.4. 正規母集団既知分散の場合の信頼区間の例として，例題 *2.3* で取り上げた *Low Infant Birth Weight* データを考える．*R* の *MASS* パッケージの *birthwt* から，変数 *bwt* は *189* 人の幼児の出生時体重（グラム）である．*bwt* の数量的要約は以下の通り．

表 4.2 Low Birth Weight データ：bwt の数量的要約

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	var	sd
709.00	2414.00	2977.00	2945.00	3487.00	4990.00	531753.5	729.21

また，*bwt* の視覚的要約は，図 *4.2* のようになる．ヒストグラムとボックスプロットから，*bwt* の分布は単峰型で左右対称であることがわかる．また *QQ-norm* プロットがほぼ直線上にあることからから，*bwt* は正規分布に従うと考えられる．

図 4.2 Low Birth Weight データ：bwt の視覚的要約



ここで，*bwt* は独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu,\sigma^2)$ に従うと

し，分散は $\sigma^2 = 730.0^2$ で既知であると仮定する．表 *4.2* より $\bar{x} = 2945.00, n = 189$ であるから，例題 *4.3* に従い期待値 μ の信頼水準 $(1-\alpha) = 0.95$ の信頼区間を求める．

$$\begin{aligned} &(2945.00 - 1.96(730.0/\sqrt{189}),\ 2945.00 + 1.96(730.0/\sqrt{189})) \\ \Longleftrightarrow\ &(2840.925, 3049.075) \end{aligned}$$

また，標準正規分布の上側 α%点は以下のように得られる．

```
> qnorm(alpha, lower.tail=FALSE)
[1] 1.644854
```

したがって期待値 μ の片側信頼区間は，以下のように得られる．

$$(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty) \iff (2945.00 - 1.645(730.0/\sqrt{189}, \infty) \iff (2857.651, \infty)$$

$$(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}) \iff (-\infty, 2945.00 + 1.645(730.0/\sqrt{189}, \infty) \iff (-\infty, 3032.349)$$

4.2.2 信頼区間の解釈

第 4.2 節冒頭で述べたとおり，区間推定量とは推定の信頼度を示すような尤もらしい確率的区間によって，推定対象のパラメター θ を推定するものであった．その区間推定量の観測値である信頼区間の精確さは，信頼水準と信頼区間の幅によって測られる．信頼水準が高く，信頼区間の幅が狭ければパラメターの推定はそれだけ正確であるといえる．

信頼水準 $(1-\alpha)$ とは，ランダムな区間 (4.10) が真の μ を含む確率である $(1-\alpha)$ から受け継がれたものであった．そこで，(4.10) の実現値として観察された信頼区間 (4.11) についても，「信頼区間が μ を含む確率は $(1-\alpha)$ である」と考えたくなる．

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad \Leftarrow \text{ これは誤り!!}$$

しかし，信頼区間 (4.11) に含まれる標本平均の観測値 $\bar{x}$ は，サンプルの観測値 $(X_1 = x_1, \ldots, X_n = x_n)$ から計算された実数値であり，信頼区間に確率変数は

含まれていない．したがって，信頼区間について「確率」を考えることは出来ない．(4.9) の括弧内の事象

$$\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

の確率が $(1-\alpha)$ である時，1 回のサンプリングによって観察された $(X_1 = x_1, \ldots, X_n = x_n)$ によって計算された信頼区間 (4.11) は真の μ を含むかもしれないし含まないかもしれない．しかし，このようなサンプリングを繰り返しそのたびに新たに計算した信頼区間 (4.11) が μ を含むか否かを記録していくと，(4.11) が μ を含む頻度は $(1-\alpha)$ に収束していく．（厳密には「大数法則」という定理により証明される．）信頼水準 $(1-\alpha)$ の信頼区間とは，そのような $(1-\alpha)$ の確率で実現するランダムな区間推定量 (4.10) の，今手元にあるデータにおける実現値と考えることができる [1)]．

分散既知の正規母集団の場合，期待値 μ の信頼水準 $(1-\alpha)$ の信頼区間は

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

で，信頼区間の幅は $2 \times z_{\alpha/2}(\sigma/\sqrt{n})$ であった．したがって標本数 n が無限に大きくなる $(n \to \infty)$ と，信頼区間の両端点は信頼区間の中点の $\bar{x}$ に収束し，信頼区間の幅は 0 に収束する．このことは標本を多く集め情報の量を増やすことで，より幅の狭い信頼区間を得て，より精度の高い推定が可能になることを示す重要な性質である．

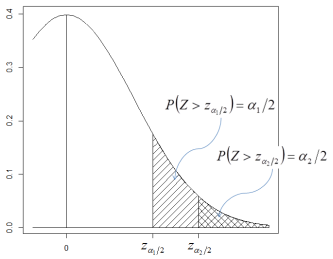
一方，信頼水準 $(1-\alpha)$ はランダムな区間 (4.10) が期待値 μ を含む確率に対応するから，信頼水準は高いほうが望ましい．しかし二つの信頼水準 $(1-\alpha_1) < (1-\alpha_2)$ があったとき，z_α を標準正規分布の上側 $100 \times \alpha$ パーセント点とすると，高い

[1)] 実は，世に行われている「信頼区間」という概念の定義には，混乱がある．教科書によっては，それぞれ 1）ランダムな区間 (4.10)，2）(4.10) とその実現値である (4.11) の両方，3）本書のように (4.11) のみ，を信頼区間の定義とする流儀が行われている．しかし，1）ランダムな区間 (4.10) を信頼区間として定義した場合，(4.9) のように「信頼区間が μ を含む確率は $(1-\alpha)$」という命題には意味があるが，区間の両端点は確率変数になるので信頼区間を実数値の区間として具体的に求めることは出来なくなる．一方 2）(4.10) と (4.11) の両方を信頼区間として定義すると，信頼区間の両端点として確率変数とその実現値という全く別の概念を同一視することになり不都合である．そこで，3）本書では，まずランダムな区間推定量として (4.10) を導入し，その実現値 (4.11) を (4.10) と切り離す形で「信頼区間」として定義する立場をとる．

信頼水準 $(1-\alpha_2)$ に対応する上側 $\alpha/2$ パーセント点 $z_{\alpha_2/2}$ の方が $z_{\alpha_1/2}$ より大きな値をとることが分かる．（図 4.3 および表 4.1 を参照）

図 4.3 上側確率と上側パーセント点の関係

$$
\begin{aligned}
&(1-\alpha_1) < (1-\alpha_2) \\
\Rightarrow &\alpha_2/2 < \alpha_1/2 \\
\Rightarrow &P(Z > z_{\alpha_2/2}) < P(Z > z_{\alpha_1/2}) \\
\Rightarrow &z_{\alpha_1/2} < z_{\alpha_2/2}
\end{aligned}
$$



つまり，信頼水準を高くすると信頼区間の幅 $2 \times z_{\alpha/2}(\sigma/\sqrt{n})$ が広くなってしまうことになる．これは一般に成立する性質で，高い信頼水準と狭い信頼区間の幅を両立するには，結局サンプル数を多くするしかないということになる．

4.2.3 信頼区間の導出

例題 4.3 では，分散既知の正規母集団における期待値 μ の信頼区間を求めた．このときの手順を一般化して，信頼区間を導出するための方法を整理しよう．いま，推定対象のパラメターを θ とし，標本 $(X_1, X_2, \ldots, X_n)$ によって θ の信頼区間を構成するとする．このとき，一般に θ に対する信頼水準 $(1-\alpha)$ の信頼区間は以下の手順によって導出される．

信頼区間の構成

1. 以下の条件を満たす確率変数 $h(X_1, X_2, \ldots, X_n; \theta)$ を見つける．

   (a) $h(X_1, X_2, \ldots, X_n; \theta)$ は，標本 $(X_1, X_2, \ldots, X_n)$ と推定対象の未知パラメター θ を含む．

(b) $h(X_1, X_2, \ldots, X_n; \theta)$ の確率分布は，推定対象のパラメター θ には依存しない．

このような確率変数 $h(X_1, X_2, \ldots, X_n; \theta)$ を，**ピボット確率変数 (pivotal random variable)** あるいは**ピボット (pivot)** と呼ぶ．

2. 以下の条件を満たす，ピボット確率変数 $h(X_1, X_2, \ldots, X_n; \theta)$ のパーセント点 a, b を求める．

$$P(a < h(X_1, X_2, \ldots, X_n; \theta) < b) \geq 1 - \alpha \tag{4.15}$$

3. (4.15) を変形して，θ に対する区間推定量 (L, U) を求める．

$$P(L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n)) \geq (1 - \alpha)$$

4. 標本 $(X_1, \ldots, X_n)$ に観測値 $X_1 = x_1, \ldots, X_n = x_n$ を代入して，信頼水準 $(1 - \alpha)$ の信頼区間 $(L(x_1, \ldots, x_n), U(x_1, \ldots, x_n))$ を導出する．

「信頼区間の構成」の第一段階で定義した「ピボット確率変数」は，推定対象のパラメターの自然な推定量を元に，それを変形してピボットの二つの条件が満たされるように導出される．例 4.3 の正規母集団既知分散の場合，推定対象のパラメターである母集団平均 μ の自然な推定量は標本平均 $\bar{X}$ である．例 4.3 で見たとおり，μ の信頼区間は $\bar{X}$ を標準化した $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ を元に構成された．ここでは，Z がピボットの条件を満たしていることを確認しよう．

1. $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ は，その定義から分子に標本平均 $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ と推定対象のパラメター μ を含む．$\bar{X}$ は，標本 $(X_1, X_2, \ldots, X_n)$ に依存するから，Z はピボット確率変数の条件 (a) を満たす．
2. $X_1, X_2, \ldots, X_n$ が独立かつ同一に正規分布 $N(\mu, \sigma^2)$ に従うことから，$\bar{X} \sim N(\mu, \sigma^2/n)$ であり，$\bar{X}$ を標準化した Z は標準正規分布 $N(0, 1)$ に従う．重要なのは推定対象のパラメター μ の値にかかわらず，Z の確率分布である標準正規分布は不変であり μ に依存しないことである．従って，Z はピボット確率変数の条件 (b) も満たしている．

ピボット確率変数の第二の条件から，信頼区間の構成の第二段階で求めたピボッ

トのパーセント点 a, b は必ず求められることがわかる．なぜなら，ピボット確率変数の確率分布が未知の推定対象に依存しないと言うことは，ピボットの確率分布が何かに依存するとしてもそれは既知のものに限られる，すなわちピボットの確率分布は既知であることを意味するからである．確率分布がわかっているのであれば，そのパーセント点 a, b もまた必ずわかるはずである．

4.2.4 一標本問題：正規母集団未知分散の場合の信頼区間

第 4.2.1 節，例題 4.3 では，$X_1, X_2, \ldots, X_n$ が独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従い σ^2 は既知であると仮定して μ の信頼区間を導出した．しかし，μ が未知であるにもかかわらず，σ^2 を既知であると仮定したのは非現実的であり，あくまで例のための仮定である．本節では仮定を一般化して，$X_1, X_2, \ldots, X_n$ が独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従い，σ^2 も未知であると仮定する．

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),\ \mu, \sigma^2 : \text{未知}$$

この場合，推定対象 (Estimand) が母集団平均 μ であり，その自然な推定量である標本平均 $\bar{X}$ が期待値 μ，分散 σ^2/n の正規分布に従う $\bar{X} \sim N(\mu, \sigma^2/n)$ ところまでは同じである．しかし，例題 4.3 と同様に $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ をピボットとして信頼区間を構成しようとすると，信頼区間の両端に未知の σ が登場してしまうことが不都合である．

ではどうするか．母集団分散 σ^2 が未知であることが困難の元であるならば，それをデータから推定すればよい．よって，$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ の σ に標本標準偏差すなわち標本分散 $S^2 = (n-1)^{-1}\sum(X_i - \bar{X})^2$ の平方根 S を代入することを考える．

$$T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}} \sim t_{n-1}$$

定理 3.7 (p. 93) により，T は自由度 $(n-1)$ の t 分布 $t(n-1)$ に従う．この T を，正規母集団未知分散の場合の一標本問題のピボット確率変数として用いることを考える．まず，「信頼区間の構成」(p. 112) の第一段階に従い，T がピボット確率変数の条件を満たしていることを確認する．

1. $T = (\bar{X} - \mu)/(S/\sqrt{n})$ は，その定義から分子の標本平均 $\bar{X}$ と分母の標本標準偏差 S に $(X_1, X_2, \ldots, X_n)$ を含み，分子に推定対象のパラメター μ を含むから，ピボットの条件 (a) を満たす．
2. $X_1, X_2, \ldots, X_n$ が独立かつ同一に正規分布 $N(\mu, \sigma^2)$ に従うとき，定理 3.7 により T は自由度 $(n-1)$ の t 分布に従う．t 分布は「自由度」というパラメターに依存するが，この場合自由度 $(n-1)$ はサンプル数 n にのみ依存し推定対象のパラメター μ に依存しない．従って，T はピボットの条件 (b) も満たしている．

よって，$T = (\bar{X} - \mu)/(S/\sqrt{n})$ がピボット確率変数として使えることがわかった．次に，「信頼区間の構成」の第二段階に従い，自由度 $(n-1)$ の t 分布の上側 $\alpha/2$ パーセント点 $t_{\alpha/2,n-1}$ と下側 $\alpha/2$ パーセント点 $-t_{\alpha/2,n-1}$ を求める [2)]．第 3.4.6.4 節 (p. 94) より，t 分布のパーセント点は `qt()` コマンドによって求められる．さらに，「信頼区間の構成」の第三段階に従い，例 4.3 と同様の方法により μ の信頼区間 (Confidence interval, CI) を導出する．ただし，$\bar{x}, s$ はそれぞれ標本平均 $\bar{X}$，標本標準偏差 S の観測値．

$$P\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$
$$\mathrm{CI} : (\bar{x} - t_{\alpha/2,n-1} \times (s/\sqrt{n}), \bar{x} + t_{\alpha/2,n-1} \times (s/\sqrt{n})) \tag{4.16}$$

例題 4.5. 本例題では，例題 *4.4* と同様，*Low Birth Weight* データの幼児の出生時体重 `bwt` の信頼区間を求める．本例題では正規母集団未知分散を想定し，母集団分散も未知であると仮定する．このとき，例題 *4.3 (p. 104)* と同様にして，期待値 μ の信頼水準 $(1-\alpha) = 0.95$ の信頼区間を求める．

推定対象 (estimand) 期待値 μ
推定量 (estimator) 標本平均 $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
ピボット確率変数 $T = (\bar{X} - \mu)/(S/\sqrt{n}) \sim t(n-1)$
ピボット確率変数の上側 $\alpha/2$ パーセント点 この場合は，自由度 $(n-1)$ の t

[2)] t 分布は原点 0 を中心として左右対称な分布であるから，下側 $\alpha/2$ パーセント点は上側 $\alpha/2$ パーセント点と絶対値が同じで符号が異なる値となる．

分布の上側 $\alpha/2$ パーセント点 $t_{\alpha/2,n-1} = 1.973$

```
> qt(alpha/2, df=(n-1), lower.tail=FALSE)
[1] 1.972663
```

$100 \times (1-\alpha)$**%信頼区間** *(4.16)* より，信頼区間は以下のように求められる．

$$
\begin{aligned}
&CI:\ (\bar{x} - t_{\alpha/2,n-1}(s/\sqrt{n}), \bar{x} + t_{\alpha/2,n-1}(s/\sqrt{n})) \\
&\quad \Rightarrow (2945.00 - 1.972663(729.21/\sqrt{189}),\ 2945.00 + 1.972663(729.21/\sqrt{189})) \\
&\quad \Rightarrow (2840.366, 3049.634)
\end{aligned}
$$

```
> 2945.00-1.972663*(729.21/sqrt(189))
[1] 2840.366
> 2945.00+1.972663*(729.21/sqrt(189))
[1] 3049.634
```

本例題の，正規母集団未知分散の一標本問題でピボットが t 分布に従う場合の信頼区間は，*R* の *t.test()* コマンドでも求めることが出来る．信頼水準は *conf.level* オプションで指定するが，デフォルトの信頼水準は *0.95* でその場合は省略できる．*t.test()* コマンドによる信頼区間は *(2839.952, 3049.222)* となっていて上で求めたものと若干異なるが，これは小数点の丸めの誤差によるものである．

```
> t.test(bwt, conf.level=0.95)

        One Sample t-test

data:  bwt
t = 55.514, df = 188, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2839.952 3049.222
```

```
sample estimates:
mean of x
 2944.587
```

4.2.5 一標本問題：非正規母集団未知分散で大標本の場合の信頼区間

前節では，$X_1, \ldots, X_n$ が独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従い，かつ σ^2 が未知である場合について μ の信頼区間を導出した．本節では，さらに状況を一般化して，$X_1, \ldots, X_n$ が独立かつ同一に（正規分布に限らず）何らかの確率分布（期待値 μ，分散 σ^2）に従い，σ^2 も未知である場合について μ の信頼区間を導出する．

やはり，推定対象 (Estimand) が母集団平均 μ であり，その自然な推定量が標本平均 $\bar{X}$ で有ることは同じである．ただし，$X_1, \ldots, X_n$ の確率分布が未知であるため，標本平均 $\bar{X}$ の確率分布も不明である．しかし，中心極限定理（第 3.4.5.2 節参照 (p. 88)）により，サンプル数 n が十分大きいとき，標本平均の確率分布は正規分布 $N(\mu, \sigma^2/n)$ に収束する．（サンプル数 n が十分大きいことを，「大標本 (large sample)」であると表現する）さらに，中心極限定理と正規確率変数の標準化を組み合わせれば，$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ が期待値 0 分散 1 の標準正規分布 $N(0, 1)$ に収束することが示される（式 (3.20), (p. 88) 参照）．最後に，やはりサンプル数 n が十分大きいとき標本分散 S^2 は母集団分散 σ^2 に収束するので，以下の命題を得る．（証明は省略．中心極限定理のほか Slutsky の定理という定理を用いる．河田，丸山，鍋谷, p. 161[7]，Casella and Berger, p. 240[1] 参照）

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2) \Rightarrow Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \to N(0, 1), (n \to \infty) \tag{4.17}$$

本節では，この $Z = (\bar{X} - \mu)/(S/\sqrt{n})$ を一標本問題のピボット確率変数として考える．（本節で考える Z は，第 4.2.4 節の T と全く同じ形をしていることに注意しよう．第 4.2.4 節では $X_1, \ldots, X_n$ が正規分布に従うと仮定されピボットは自由度 $(n-1)$ の t 分布にしたがったのに対して，本節では正規性の仮定が外されている点が異なっている）ふたたび「信頼区間の構成」(p. 112) の第一段階に従えば，第 4.2.4 節と同様 Z がピボット確率変数の条件を満たしているこ

とが確認出来る．よって，例 4.3 と同様の方法により μ の信頼区間を導出する．

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1-\alpha,\ n\to\infty \text{ (4.17) による}$$
$$\text{CI} : (\bar{x} - z_{\alpha/2}(s/\sqrt{n}), \bar{x} + z_{\alpha/2}(s/\sqrt{n})) \tag{4.18}$$

例題 4.6. 本例題では，*Low Birth Weight* データの幼児の出生時体重 `bwt` について，非正規母集団大標本の場合の信頼区間を導出する．

推定対象 (estimand) 期待値 μ

推定量 (estimator) 標本平均 $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

ピボット確率変数 $Z = (\bar{X}-\mu)/(S/\sqrt{n}) \approx N(0,1),\ n\to\infty$（中心極限定理による）

ピボット確率変数の上側 $\alpha/2$ パーセント点 標準正規分布の上側 $\alpha/2$ パーセント点 $z_{\alpha/2} = 1.96$

$100\times(1-\alpha)$**%信頼区間** *(4.18)* より，信頼区間は以下のように求められる．

$$\begin{aligned} &\textit{CI: } (\bar{x} - z_{\alpha/2}(s/\sqrt{n}), \bar{x} + z_{\alpha/2}(s/\sqrt{n})) \\ &\quad = (2945.00 - 1.96(729.21/\sqrt{189}),\ 2945.00 + 1.96(729.21/\sqrt{189})) \\ &\quad = (2841.037, 3048.963) \end{aligned}$$

同じ `bwt` の期待値の信頼区間であっても，例題 4.4，4.5，4.6 で母集団分散 σ^2 が既知か未知か，母集団の確率分布が正規分布か否かで，結果が微妙に異なることに注意しよう．

第 2 章で記述統計について議論した際，同じ解析目的に対して複数の解析手法が存在する場合があり，前提条件の違いによって手法を使い分ける必要があることを指摘した．一標本問題の場合，1）$X_1, \ldots, X_n$ が正規分布に従うか否か，2）母集団分散が既知か否か，3）中心極限定理が適用できるか否か，で場合分けされた．標本分布が正規分布に従うか否かは，QQ-norm plot (p. 65) によって検証できる．分散が既知であるか否かは問題によって自明である（通常，分散が既知で有ることは無い）．中心極限定理を適用するためには，標本数が例

えば 20〜30 以上の大標本であることが必要とされている．

4.3 仮説検定

第 4.1 節，第 4.2 節では，点推定，信頼区間について議論した．本節では，統計的推論のもう一つの重要な手法である**統計的仮説検定 (statistical hypothesis testing)** あるいは単に**仮説検定**について議論する．

点推定あるいは区間推定と信頼区間は，与えられたデータから出発して推定対象のパラメター θ にアプローチする推論の方法であった．これに対し仮説検定ではまずパラメター θ に関する何らかの仮説から出発して，手元にあるデータと照らし合わせ，その仮説が妥当であるか検証するという全く別のアプローチをとる．

例えば，ある疾患に対する新薬の効果を検証する臨床試験で，100 例の症例のうち 40 例で効果ありと認められたとき，奏効率を 40%と推定するのが点推定の立場である．これに対して，例えば「新薬の奏効率は 30%より大きい」という仮説を立て，100 例の症例のうち 40 例で効果ありというデータに照らしてこの仮説が妥当であるか否か決定するのが仮説検定の立場である．

伝統的な統計的仮説検定の大きな特徴は，仮説を立てる際に必ず以下の二つの仮説を立てる点にある．

帰無仮説 (null hypothesis) 母集団のパラメターに関する仮説で，検定の当初真であると仮定されるもの．帰無仮説を H_0 と記す．

対立仮説 (alternative hypothesis) 帰無仮説と論理的に対立する仮説．対立仮説を H_1 と記す．

上の例で言えば，「新薬の奏効率は 30 %以下である」という仮説を帰無仮説とすれば，「新薬の奏効率は 30%より大きい」という仮説が対立仮説になる．p を新薬の奏効率とすれば，上記の仮説は以下のように書ける．

$$H_0 : p \leq 0.3 \text{ vs. } H_1 : p > 0.3$$

対立仮説 H_1 は帰無仮説 H_0 と対立する仮説であるから，H_0 と H_1 は論理的に

両立しない二者択一の命題である．H_0 が真であれば H_1 が偽，H_0 が偽であれば H_1 が真であり，H_0 と H_1 が同時に真であることも同時に偽であることもない．帰無仮説を，例えば「新薬の奏効率は 30%である」として仮説を以下のように置くこともある．

$$H_0 : p = 0.3 \text{ vs. } H_1 : p > 0.3$$

これは，例えば従来用いられた薬の奏効率は 30%であり，新薬の奏効率が 30%を下回ることはないと考えられる場合などである．

H_0 と H_1 が置かれると，標本から得られたデータをもとに「H_0 が真であり H_1 が偽である」あるいは「H_0 が偽であり H_1 が真である」と決定することになる．H_0 が真であると決定することを，H_0 を**採択 (accept) する，受け入れる**といい，H_1 が偽であると決定することと同義である．H_0 が偽であり H_1 が真であると決定することを，H_0 を**棄却 (reject) する**という．

統計的仮説検定では，まず帰無仮説 H_0 が真であり正しいと仮定する．その仮定の下で，いま手元にあるデータと同等もしくはより帰無仮説 H_0 に矛盾するようなデータが観察される確率を求める．これが後述する「p 値」と呼ばれるものである．この p 値が事前に定められた定数（後述する「有意水準」）より小さいとき H_0 を棄却するという決定をする．つまり，H_0 が正しいと仮定したときデータから観察されるような事象が起こる確率があまりにも小さいときは，そのような小さな確率の現象が起こったと考えるよりは，初めに「H_0 が正しい」と仮定したことが誤りであるとして H_0 を棄却し H_1 を採択するという決定をするわけである [3)]．以下，仮説検定の定式化をさらに詳しく見ていこう．

定義 4.7. 与えられた帰無仮説 H_0 と対立仮説 H_1 に対して，標本から計算される統計量で，その観測値を用いて H_0 を採択する，あるいは棄却するという決定を行うものを**検定統計量** ***(test statistic)*** という．検定統計量がとりうる値の集合のうち，H_0 が棄却される領域を**棄却域** ***(rejection region)***，H_0 が採択される領域を**採択域** ***(acceptance region)*** という．

本書で取り上げる仮説検定では，検定統計量の構成方法には一定のルールが

[3)] このような仮説検定の考え方は，数学の証明法における「背理法」と大変よく似ている．

ある．それは

> [検定統計量の構成方法]
> 検定統計量の形は，検定の対象となるパラメター—の信頼区間を構成する際のピボット確率変数と「ほとんど」同じ形をしている．ただし検定統計量においては，ピボット確率変数に含まれる未知パラメター—が帰無仮説の下で仮定されたパラメター—の値で置き換えられている．

というものである．（本書における仮説検定の構成は，多分に直感的である．検定統計量の定義を含め，詳細は数理統計学の成書（尾畑[9]，柳川[13]，竹村[11] など）を参照されたい．また，後述するノンパラメトリック法による仮説検定は，本節とは異なる論理で構成される．そちらについても，ノンパラメトリック統計学の教科書（柳川[14]，Hollander, Wolfe and Chicken[5] など）を参照されたい．）

まず，例として第 4.2.1 節，例題 4.3 (p. 104) 正規母集団既知分散の一標本問題を考える．すなわち $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ かつ σ^2 は既知とする．このとき，以下の仮説を考える．

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 (> \mu_0) \tag{4.19}$$

ここで，μ_0, μ_1 というのは一般的な表現で，実際の仮説検定の場面では母集団平均 $\mu_0 = 0, \mu_1 = 1.0$ など μ_0, μ_1 には具体的な実数が代入される．このとき，μ の信頼区間を求めるためのピボット確率変数は $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ であった（p. 113）．よって「検定統計量の構成方法 (p. 121)」により，検定統計量の形は

$$Z = \frac{\bar{X} - \mu_0}{(\sigma/\sqrt{n})} \tag{4.20}$$

となる．この検定統計量の観測値が棄却域に入れば H_0 を棄却し，棄却域に入らなければ H_0 を採択するわけである．それでは，棄却域はどのように定められるのであろうか．それを考えるために，まず，統計的仮説検定において犯しうる過誤について考えてみる．仮説検定における決定が検定統計量によってなされ，かつ検定統計量自体が確率変数でありランダムに変動する存在である以

上，H_0 を受け入れる，あるいは H_0 を棄却するという決定には，誤りの可能性は常に存在する．我々の取りうる決定と仮説の真偽を考えると，我々が直面しうる状況は「帰無仮説 H_0 を受け入れる vs. 帰無仮説 H_0 を棄却する」×「帰無仮説 H_0 が真 vs. 帰無仮説 H_0 が偽」の四通りしかない．表 4.3 参照．

表 4.3

決定／仮説の真偽	H_0 が真（H_1 が偽）	H_1 が真（H_0 が偽）
H_0 を採択する	○	第二種の過誤
H_0 を棄却する	第一種の過誤	○

H_0 が真であるとき H_0 を採択する（表 4.3 左上），H_1 が真（H_0 が偽）であるとき H_0 を棄却し H_1 を受け入れる（表 4.3 右下），はどちらも正しい決定である．H_0 が真であるとき誤って H_0 を棄却することを**第一種の過誤 (type I error)**，H_1 が真で H_0 が偽であるとき誤って H_0 を受け入れることを**第二種の過誤 (type II error)** と呼ぶが，これらはいずれも誤った決定である．

このとき，当然誤りの可能性が限りなく小さくなるように決定を行いたいわけだが，そのような検定を行うことは可能であろうか．

まず，第一種の過誤と第二種の過誤の可能性を同時にゼロにすることは不可能である．もしそのようなことが可能であれば，表 4.3 において誤りを犯す確率がゼロになる．不確実性を含むデータをもとに，100%誤りのない決定をすることはできない．

それでは，第一種の過誤あるいは第二種の過誤のいずれかを犯す可能性をゼロにすることは可能であろうか? 意外なことに，これは可能である．例えば，「いかなるデータを得ようとも，それにかかわりなく H_0 を採択する」という決定方法を考えよう．この場合，どのような場合でも H_0 を採択するわけだから，H_0 が真であるとき誤って H_0 を棄却する第一種の過誤を犯す確率はゼロである．しかし当然のことながら，このような決定方式は以下の二つの理由により全くナンセンスである．

1. この決定方式では，データから得られた情報を全く使用していない．

2. H_0 が真であるときは第一種の過誤を犯す確率がゼロであるが，H_0 が偽（H_1 が真）であるときも必ず H_0 を受け入れるため 100%の確率で第二種の過誤を犯すことになる．

つまり，第一種あるいは第二種の過誤を犯す可能性を同時にゼロにすることはできないし，どちらか片方の可能性をゼロにすることは出来なくはないがナンセンスな決定方法を用いることになりこれも不都合である．

したがって，仮説検定において誤りのない決定をする，というのは「ないものねだり」であり，第一種と第二種の双方の過誤を犯すリスクを受け入れざるを得ない．では，我々にできることとは何か? それは，「秩序をもって」リスクを受け入れる，つまり，どのような過誤を犯す可能性があるのか，その確率を理解したうえでリスクを犯す，ということである．統計的仮説検定においては，第一種の過誤の確率を事前に定めた定数以下にする，という条件の下で第二種の過誤の確率を最小化する，という考え方で仮説検定を定式化する．

定義 4.8. 事前に決められた定数で，第一種の過誤の確率がそれ以下になるように検定が定められるものを**有意水準** ***(significance level)*** と言い α と記す．

つまり，検定は第一種の過誤の確率が有意水準 α 以下になるように定義される．

$$P(\text{type I error}) \leq \alpha \tag{4.21}$$

有意水準 α は理論的には任意であるが，通常 $\alpha = 0.05, 0.01, 0.1$ などとする．

さて改めて，例題 4.3 と同様，$X_1, X_2, \ldots, X_n$ が独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従い，分散 σ^2 は既知であるという仮定の下で．以下の仮説 (4.19) を考える [4]．

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 (> \mu_0)$$

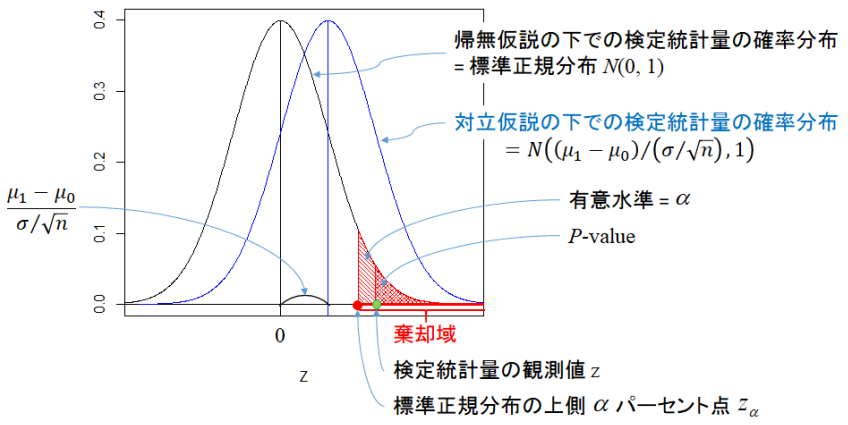
(4.20) で考えた通り，検定統計量は $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ で与えられるが，これを以下のように変形する．

[4] ここからの仮説検定の導出は初見では複雑に思えるが，図 4.4 を見ながらゆっくり考えれば必ず理解できる．がんばろう．

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{\sim N(0,1)} + \underbrace{\frac{\mu - \mu_0}{\sigma/\sqrt{n}}}_{\begin{cases} = 0 \text{ under } H_0 \\ > 0 \text{ under } H_1 \end{cases}} \tag{4.22}$$

(4.22) の右辺第一項は，正規母集団からの標本平均の標準化であるから，常に標準正規分布 $N(0,1)$ に従う．一方，(4.22) 右辺第二項は，帰無仮説の下では $H_0 : \mu = \mu_0$ より 0 に等しいが，対立仮説の下では $H_1 : \mu = \mu_1 (> \mu_0)$ より何か正の定数となる．つまりこの問題における検定統計量 Z は，帰無仮説の下では標準正規分布 $N(0,1)$ に，対立仮説の下では $N(0,1)$ を正の方向にシフトした確率分布 $N((\mu_1 - \mu_0)/(\sigma/\sqrt{n}), 1)$ に従う．（図 4.4 参照）

図 4.4 帰無仮説と対立仮説の下での検定統計量の確率分布



対立仮説 H_1 の下で Z の分布が正の方向にシフトしたということは，$H_1 : \mu = \mu_1 (> \mu_0)$ の下では検定統計量 Z は「より大きな値を取りがちである」ということである．従って，Z の観測値として十分大きな値が得られたときは，対立仮説 H_1 を採択し帰無仮説 H_0 を棄却するのが合理的な決定ということになる．

「棄却域」とは，検定統計量がとりうる値の集合のうち H_0 が棄却される領域のことであったから，$H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1 (> \mu_0)$ という仮説に対しては，検定統計量の値のうち「大きな」値の集合を棄却域とするのが合理的，ということになる．

それでは，検定統計量の値のうち，どの値より大きい集合を棄却域とすればよいか? 検定は第一種の過誤の確率が有意水準 α 未満になるように定義されることから（(4.21), p. 123），帰無仮説の下で検定統計量がある値より大きくなる確率が有意水準 α 未満になるように棄却域を定めればよい．帰無仮説 H_0 の下で検定統計量は標準正規分布 $N(0, 1)$ に従うことから，標準正規分布の上側 $100 \times \alpha$ パーセント点 z_α より大きい領域を検定の棄却域とすればよいことになる．すなわち，検定統計量の観測値 z が z_α より大きいとき帰無仮説を棄却することになる．

$$\text{仮説検定 (4.19) の棄却域 : } \{z : z > z_\alpha\} \tag{4.23}$$

最後に，検定統計量の観測値 z が $z > z_\alpha$ となるとき H_0 を棄却するといっても，z が z_α より僅かに大きな値をとる場合と遥かに大きな値をとる場合では意味が異なることに注意しよう．有意水準 $\alpha =$ 第一種の過誤の確率が小さくなると，棄却域の端点 (critical value) は原点から遠くなる（図 4.3 参照）．すなわち検定統計量の観測値 z がより大きな値をとる場合，有意水準（＝第一種の過誤の確率）をより小さく見積もってもなお帰無仮説 H_0 が棄却されるということになる．

定義 4.9. 与えられた検定統計量の観測値 z に対して，帰無仮説 H_0 を棄却するのに必要な最小の有意水準を **p 値（p-*value*）**という．

言い換えれば， p 値とは帰無仮説 H_0 のもとで，検定統計量がその観測値と同程度かより甚だしい (as extreme or more extreme) 値をとる確率のことである．より小さなは p 値は帰無仮説を棄却するためのより強い証拠であると考えられる．p 値と有意水準の間には，以下の関係が成立する．

[p 値と有意水準の関係]

p 値 $<$ 有意水準 $\alpha \Rightarrow$ 有意水準 α で H_0 を棄却する

p 値 $>$ 有意水準 $\alpha \Rightarrow$ 有意水準 α で H_0 を棄却しない（採択する）

仮説検定の結果を報告する際は，帰無仮説 H_0 を棄却するか否かだけではなく，その検定の p 値も一緒に報告することを習慣としよう．また p 値を記載する際は，小数点以下第 3 位まで記載するのが常識である．（$p = 0.05$ と小数点以下第 2 位までしか報告しない場合，有意水準 $\alpha = 0.05$ と比較して $p = 0.045 < \alpha$ のように α 未満の p 値を切り上げていて統計的に有意に H_0 が棄却されるのか，あるいは $p = 0.054$ のように α より大きい p 値を切り下げていて H_0 を棄却しないのか，判然としないからである．）また，p 値があまりに小さい場合，統計解析ソフトによっては $p = 0$ といった出力を返す場合がある．しかし実際には，確率である p 値が完全に 0 になることは考えづらいので，$p < 0.001$ というように表記することをお勧めする．

以上述べてきた仮説検定の手順を整理すると，以下のようになる．

仮説検定の構成

1. 検定の対象となるパラメターを特定する．
2. 帰無仮説 (null hypothesis) H_0 と，帰無仮説の元で仮定されるパラメターの値 (null value) を決める．
3. 対立仮説 (alternative hypothesis) H_1 を決める．
4. 検定統計量 (test statistic) を定める．検定統計量の形は，検定対象のパラメターの信頼区間を構成するためのピボット確率変数の未知パラメターに，帰無仮説の元で仮定した値 (null value) を代入したものである．
5. 観察されたデータの値を元に，検定統計量の観測値を計算する．
6. 帰無仮説 H_0 の下で，検定統計量の値から p 値を計算する．
7. 検定統計量の観測値が，棄却域の中にあれば帰無仮説を棄却する．p 値と有意水準 α を比較して，帰無仮説を棄却するか否かを決定する．

とくに初めのうちは，検定に当たっては上の手順を厳密に踏襲するべきである．

4.3.1 一標本問題：正規母集団既知分散の場合の仮説検定

正規母集団既知分散の場合の仮説検定の例として，例題 4.4 (p. 109) で取り上げた Low Infant Birth Weight データの幼児の出生児体重 `bwt` を用いる．例題 4.4 と同様，`bwt` は独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うとし，分散は $\sigma^2 = 730.0^2$ で既知であると仮定する．このとき，有意水準 $\alpha = 0.05$ の下で以下の仮説を検定する．

$$H_0 : \mu = 2800 \text{ vs. } H_1; \mu > 2800$$

検定統計量 (test statistic) (4.20) で見た通り，正規母集団既知分散の場合の期待値の検定の検定統計量は $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ で与えられる．($\mu_0 = 2800$ である.) 帰無仮説 $H_0 : \mu = \mu_0$ の下で，検定統計量 Z は標準正規分布 $N(0, 1)$ に従う．

$$\text{検定統計量}: Z = \frac{\bar{X} - \mu_0}{(\sigma/\sqrt{n})} \sim N(0, 1) \text{ under } H_0$$

棄却域 (rejection region) 対立仮説 $H_1 : \mu > 2800 (= \mu_0)$ の下で，棄却域は検定統計量がとりうる値のうち標準正規分布の上側 $100 \times \alpha$ パーセント点 z_α より大きい領域となる．(図 4.4 参照 (p.124)) 今，有意水準が $\alpha = 0.05$ であったから，$z_\alpha = 1.644854$.

$$\text{棄却域}: \{z : z > z_\alpha\} = \{z : z > 1.644854\}$$

```
> alpha <- 0.05
> qnorm(alpha, lower.tail=FALSE)
[1] 1.644854
```

検定統計量の観測値 例題 4.4 より標本平均は $\bar{x} = 2945.00$ であったから，検定統計量の観測値は以下の通り．

$$z = \frac{\bar{x} - \mu_0}{(\sigma/\sqrt{n})} = \frac{2945.00 - 2800}{(730.0/\sqrt{189})} = 2.73$$

決定 検定統計量の観測値 $z = 2.73$ は，棄却域の端点 1.644854 より大きい．す

なわち検定統計量の観測値は棄却域の中にあるので，帰無仮説は有意水準 $\alpha = 0:05$ のもとで棄却される．

p 値 定義 4.9（p.125）より，仮説検定の p 値とは帰無仮説 H_0 を棄却するのに必要な最小の有意水準 α のことであった．検定統計量 Z の観測値を z とすると，対立仮説 $H_1 : \mu > \mu_0$ に対して $z > z_\alpha$ となるとき帰無仮説 $H_0 : \mu = \mu_0$ は棄却された．よって，対立仮説 $H_1 : \mu > \mu_0$ に対する p 値は $p = P(Z > z)$ 等しい．ただし，Z は帰無仮説の元での検定統計量の確率分布である標準正規分布に従い，z は検定統計量の観測値 $z = 2.73$ に等しい．すなわち，この検定の p 値は以下のように得られる．

```
> p <- pnorm(2.73, lower.tail=FALSE)
> p
[1] 0.003166716
```

4.3.2 片側検定と両側検定

前節の (4.20) では，以下の仮説を検定した．

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 (> \mu_0)$$

前節で見たとおり，(4.20) のように対立仮説 H_1 の元で仮定されたパラメターの値 μ_1 が μ_0 より大きい場合，検定統計量の取りうる値の集合の中で「大きい」あるいは「右側」の領域が棄却域となった．逆に，対立仮説の下で $H_1 : \mu = \mu_1 (< \mu_0)$ と仮定された場合，(4.22) 右辺第二項は

$$\frac{\mu - \mu_0}{\sigma/\sqrt{n}} = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} < 0 \text{ under } H_1$$

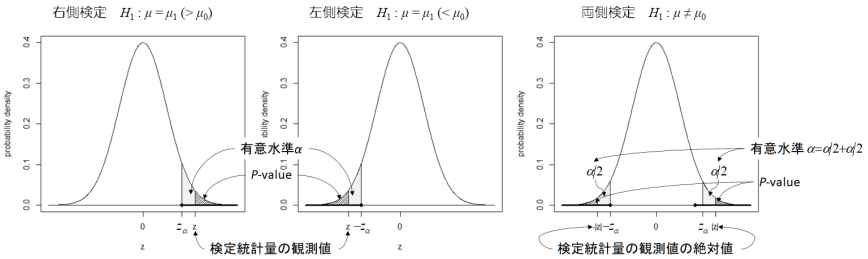
となり，H_1 の下で検定統計量 Z は「より小さな値を取りがち」になる．（帰無仮説 H_0 の下では，Z は依然として標準正規分布に従う．）したがって，対立仮説 $H_1 : \mu = \mu_1 (< \mu_0)$ に対しては，棄却域は検定統計量の取りうる値のうち「小さい」あるいは「左側」の領域になる（図 4.5 中央参照）．

このように，棄却域が右側に来る検定を**右側（上側）検定 (right(upper)-sided test)**，左側に来る検定を**左側（下側）検定 (left(lower)-sided test)**，両方合

わせて**片側検定 (one-sided test)** という．左側検定の場合，p 値は帰無仮説の元で検定統計量がその観測値より小さな値をとる確率に等しい．なお，棄却域が左右どちらの領域にできるかは，対立仮説左右両側に来るの下で仮定される μ_1 の値そのものには依存せず，$\mu_1 > \mu_0$ あるいは $\mu_1 < \mu_0$ という μ_0 との大小関係のみに依存することに注意する．

一方，対立仮説が $H_1 : \mu \neq \mu_0$ となる場合，対立仮説の元で検定統計量は左右いずれか原点から遠い方向に値をとる傾向がある．従って，検定の棄却域は左右両側にそれぞれ帰無仮説の元での確率が $\alpha/2$ となるようにとられる（図 4.5 右参照）．このように棄却域が両側に来る検定を，**両側検定 (two-sided test)** という．両側検定の場合，p 値は帰無仮説の元で検定統計量がその観測値の「絶対値」より大きくなる確率× 2 となる．

図 4.5 仮説検定の棄却域



4.3.3 一標本問題：正規母集団未知分散の場合の仮説検定

第 4.3.1 節では，正規母集団既知分散の一標本問題において仮説検定

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 (> \mu_0)$$

を考えた．本節では，分散既知の仮定を緩めて，第 4.2.4 節に倣って $X_1, X_2, \ldots, X_n$ が独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従い，σ^2 も未知である正規母集団未知分散の場合について検定を考える．

第 4.2.4 節でみた通り，正規母集団既知分散の場合の期待値 μ の信頼区間を

構成するためにピボット確率変数として自由度 $(n-1)$ の t 分布に従う以下の T を用いた．

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

したがって，「検定統計量の構成方法 (p. 121)」により，この場合の検定統計量は μ を μ_0 で置き換えた以下のものになる．

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{ under } H_0$$

(4.22) と同様に，検定統計量 T は帰無仮説 $H_0 : \mu = \mu_0$ の下では自由度 $(n-1)$ の t 分布に従い，対立仮説 $H_1 : \mu = \mu_1 (> \mu_0)$ の下では t 分布を正の方向に $(\mu_1 - \mu_0)/(S/\sqrt{n}) > 0$ だけシフトした分布に従うことが分かる．よって，T は対立仮説 H_1 の下で「大きな」値を取りがちであることから，第 4.3.1 節で考えたのと同様に棄却域は T の取りうる値のうち「大きな」値の集合，具体的には自由度 $(n-1)$ の t 分布の上側 $100 \times \alpha$ パーセント点より大きな集合となる．

$$\text{棄却域} \ : \ \{t : t > t_{\alpha, n-1}\}$$

ただし，$t_{\alpha,n-1}$ は自由度 $(n-1)$ の t 分布の上側 α%点．

例題 4.7. 第 *4.3.1* 節で考えた *Low Infant Birth Weight* データの幼児の出生児体重 `bwt` について，分散 σ^2 を未知と仮定して同様の検定を行ってみよう．

$$H_0 : \mu = 2800 \ \textit{vs.} \ H_1; \mu > 2800$$

検定統計量 (test statistic)

$$\text{検定統計量} : T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \ \textit{under} \ H_0$$

棄却域 (rejection region) 対立仮説 $H_1 : \mu > 2800 (= \mu_0)$ の下で，棄却域は自由度 $(n-1)$ の t 分布の上側 $100 \times \alpha$ パーセント点より大きい領域となる．有意水準が $\alpha = 0.05$ の下で，$t_{\alpha,n-1} = 1.652999$．

$$\text{棄却域} : \{t : t > t_{\alpha,n-1}\} = \{t : t > 1.652999\}$$

```
> n <- length(bwt)
> qt(alpha, df=(n-1), lower.tail=FALSE)
[1] 1.652999
```

検定統計量の観測値 例題 *4.4 (p. 109)* より標本平均は $\bar{x} = 2945.00$ であったから，検定統計量の観測値は以下の通り．

$$t = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})} = \frac{2945.00 - 2800}{(729.21/\sqrt{189})} = 2.725875$$

決定 検定統計量の観測値 $t = 2.725875$ は，棄却域の端点 *1.652999* より大きく観測値 t は棄却域の中にあるので，帰無仮説は有意水準 $\alpha = 0 : 05$ のもとで棄却される．

p 値 対立仮説 $H_1 : \mu > \mu_0$ に対する p 値は，自由度 $(n-1)$ の t 分布に従う確率変数が検定統計量の観測値 $t = 2.725875$ より大きな値をとる確率に等しい．

```
> p <- pt(2.725875, df=(n-1), lower.tail=FALSE)
> p
[1] 0.003509652
```

本例題で考えた，正規母集団未知分散の一標本問題での検定は *R* の *t.test()* コマンドで行うことができる．帰無仮説の下で仮定したパラメターの値は，*mu* オプションで指定する．対立仮説は *alternative* オプションで指定する．（両側検定：*two.sided (default)*，右側検定：*greater*，左側検定：*less*）本例題の場合は，以下のようになる．

```
> t.test(bwt, mu=2800, alternative="greater")

        One Sample t-test

data:  bwt
t = 2.7259, df = 188, p-value = 0.00351
alternative hypothesis: true mean is greater than 2800
95 percent confidence interval:
```

```
 2856.908      Inf
sample estimates:
mean of x
 2944.587
```

なお，本例題のように *t.test()* コマンドで片側検定を行った場合，出力される信頼区間も片側信頼区間（この場合は $(2856.908, \infty)$）になることに注意しよう．

4.3.4 一標本問題 : 非正規母集団未知分散で大標本の場合の仮説検定

本節では，前節からさらに状況を一般化し，第 4.2.5 節に倣って $X_1, X_2, \ldots, X_n$ が独立かつ同一の確率分布に従う場合について，同じく以下の期待値 μ に関する検定を考察する．

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 (> \mu_0)$$

第 4.2.5 節でみた通り，期待値 μ の信頼区間を構成するためのピボット確率変数 $Z = (\bar{X} - \mu)/(S/\sqrt{n})$ の分布は，サンプル数 n が十分大きいとき中心極限定理によって標準正規分布 $N(0,1)$ で近似される．よって，検定統計量は以下のようになる．

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1) \quad (n \to \infty) \text{ under } H_0$$

第 4.3.1 と同様にして，対立仮説 $H_1 : \mu = \mu_1 (> \mu_0)$ に対して棄却域は $\{z : z > z_\alpha\}$ となる．検定統計量の観測値が棄却域に入れば，帰無仮説が棄却されることも同様である．

例題 4.8. 前節，前々節と同様，*Low Infant Birth Weight* データの幼児の出生児体重 *bwt* について，以下の仮説を検定する．

$$H_0 : \mu = 2800 \ \textit{vs.} \ H_1; \mu > 2800$$

検定統計量 (test statistic) *bwt* のサンプル数は $n = 189$ で十分大きいため，中心極限定理が適用可能．

$$\text{検定統計量} : Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1) \quad (n \to \infty) \ \textit{under} \ H_0$$

棄却域 (rejection region) 対立仮説 $H_1 : \mu > 2800 (= \mu_0)$ の下で，標準正規分布の上側 $100 \times \alpha$ パーセント点 z_α より大きい領域となる．

$$\text{棄却域：} \{z : z > z_\alpha\} = \{z : z > 1.644854\}$$

検定統計量の観測値 検定統計量の形は，例題 *4.7* と同じである．

$$z = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})} = \frac{2945.00 - 2800}{(729.21/\sqrt{189})} = 2.725875$$

決定 検定統計量の観測値 $z = 2.725875$ は，棄却域の端点 *1.644854* より大きく観測値 z は棄却域の中にあるので，帰無仮説は有意水準 $\alpha = 0 : 05$ のもとで棄却される．

p 値 対立仮説 $H_1 : \mu > \mu_0$ に対する p 値は，標準正規分布 $N(0, 1)$ に従う確率変数が検定統計量の観測値 $z = 2.725875$ より大きな値をとる確率に等しい．

```
> p <- pnorm(2.725875, lower.tail=FALSE)
> p
[1] 0.003206564
```

4.3.5 一標本問題：ノンパラメトリック検定（ウィルコクソン符号順位検定, Wilcoxon signed rank test）

第 4.3.1, 4.3.3, 4.3.4 節で取り上げた一標本問題の信頼区間や仮説検定は，サンプル $X_1, \ldots, X_n$ が正規分布に従うか，中心極限定理によって標本平均の分布が正規分布で近似できることを前提としていた．しかし，データによっては標本分布が正規分布に従うとは仮定できず，サンプル数も十分大きくない場合がある．そのようなときは，以下に述べる**分布によらない方法 (distribution-free method)** あるいは母集団を特徴付けるパラメターに仮定を置かないという意味で**ノンパラメトリックな方法 (non-parametric method)** と呼ばれる手法で検定が構成される．

いま，$X_1, X_2, \ldots, X_n$ が互いに独立で，中央値 $\tilde{\mu}$ に関して左右対称な連続確率分布に従うと仮定する．（確率分布が左右対称であるとき，期待値 μ は中央値 $\tilde{\mu}$ に一致することに注意しよう）このとき，以下の仮説を検定する．

$$H_0 : \tilde{\mu} = \tilde{\mu}_0 \text{ vs. } H_1 : \tilde{\mu} > \tilde{\mu}_0$$

いま，X_i と $\tilde{\mu}_0$ の差の絶対値 $|X_i - \tilde{\mu}_0|$ を考え，大小順に並べ直したときの順位を R_i とする．つまり，正負によらず中央値 $\tilde{\mu}_0$ に最も近い X_i の順位が $R_i = 1$，$\tilde{\mu}_0$ から最も離れた X_i の順位が $R_i = n$ となる．そのとき，検定統計量を $(X_i - \tilde{\mu}) > 0$ となる場合の順位 R_i（これを**正の符号順位 (positive signed rank** という）の和 $W = \sum R_i$ で定義することにする．

X の分布が中央値 $\tilde{\mu}$ とは,$P(X > \tilde{\mu}) = P(X < \tilde{\mu}) = 1/2$ となる点であるから,帰無仮説 H_0 の下で $(X_1 - \tilde{\mu}_0), \ldots, (X_n - \tilde{\mu}_0)$ はそれぞれ $1/2$ の確率で正または負の値をとるはずである．従って，上で定義した検定統計量 W は帰無仮説の下では $1, \ldots, n$ の順位から $1/2$ の確率で選び出したものの和になっている．この W を用いた検定は，**Wilcoxon 符号順位検定 (Wilcoxon signed rank test)** と呼ばれる．左側検定，両側検定についても，同様に定式化される．検定統計量 W の確率分布と検定の p 値は，R の `wilcox.test()` コマンドで計算される．

例題 4.9. ここでは，本章で繰り返し用いた *Low Infant Birth Weight* データの幼児の出生児体重 *bwt* について， *Wilcoxon* 符号順位検定により検定する．

$$H_0 : \tilde{\mu} = 2800 \textit{ vs. } H_1; \tilde{\mu} > 2800$$

```
> wilcox.test(bwt, mu=2800, alternative="greater", conf.int=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  bwt
V = 11052, p-value = 0.002942
alternative hypothesis: true location is greater than 2800
95 percent confidence interval:
 2863.5    Inf
sample estimates:
(pseudo)median
        2958.5
```

wilcox.test() コマンドの出力から，$p = 0.002942$ で有意水準 α より小さいことから帰無仮説 $H_0 : \tilde{\mu} = 2800$ は棄却される．*wilcox.test()* コマンドに *conf.int=TRUE* オプションを与えると，中央値の *(95%)* 信頼区間（この場合は，$(2863.908, \infty)$ の片側信頼区間）が得られる．

なお，Wilcoxon 符号順位検定を用いるには，X の確率分布が $\tilde{\mu}$ に関して左右対称である必要があったことに注意する [5]．これは，実用上は結構きつい条件である．サンプルのヒストグラムなどで分布の形状を確認し，もし分布が右もしくは左に歪んだ分布であるときは Wilcoxon 符号順位検定は使えないことになる．そのときは，サンプル数を増やして第 4.3.4 節で取り上げた中心極限定理による正規近似に持ち込むのが一つの方法である．

4.3.6 多重仮説検定 (Multiple hypothesis testing)

本節では，ここまで一つの仮説（帰無仮説 H_0 と対立仮説 H_1 のペア）に対して，一標本問題を例に様々な仮説検定を考えてきた．しかし，実際のデータ解析の場面では複数の仮説を一度に検定する「多重仮説検定 (multiple hypothesis testing)」を考えなければならない場合もある．

例題 4.10. 高血圧患者の血圧を下げるために用いられる，降圧剤の効果を調べる試験を行うとしよう．従来用いられた降圧剤は，投与後 *10mmHg* の血圧低下をもたらすとする．いま，A_1 から A_K まで K 種類の新薬の候補があり，従来薬に比べてどの薬の降圧効果が有意に大きいか検定したいとする．新薬候補 A_k を用いたときの血圧低下の期待値を $\mu_k mmHg$ とすると，検定すべき仮説は以下の K 通りになる．

$$H_0^k : \mu_k = 10 \ \textit{vs.}\ H_1^k : \mu_k > 10, \ k = 1, \ldots, K$$

仮説検定の基本的な考え方 (p. 123) は，第一種の過誤 (type I error) の確率が有意水準 α 以下になるという条件の下で，第二種の過誤 (type II error) の確率を最小化する，というものであった．個々の検定の有意水準を $\alpha = 0.05$ とす

[5] 左右対称でない分布に関する中央値の検定も存在するが，本書ではこれ以上触れない．柳川[14]，河田，丸山，鍋谷[7] 等を参照．

ると，これは帰無仮説 H_0^k が正しいにもかかわらず誤って H_0^k を棄却する，すなわち，新薬候補 A_k の降圧効果は従来薬と変わらない $\mu_k = 10\text{mmHg}$ であるにもかかわらず，データに含まれるランダムな誤差のためにあたかも $\mu_k > 10\text{mmHg}$ であるかのように誤断する可能性が 5%有ると言うことを意味する．

それでは，K 個の帰無仮説の集まり（=ファミリー）$\mathcal{F} = \{H_0^1, H_0^2, \ldots, H_0^K\}$ をまとめて考えるときは，どのように検定を行うべきであろうか．この場合，第一種の過誤に相当するのは「全ての帰無仮説 H_0^k が正しいにもかかわらず，誤って何れかの H_0^k が棄却される」過誤ということになる．この，「正しい K 個の帰無仮説 H_0^k のうち，少なくとも一つの H_0^k が誤って棄却される誤り」の確率を **Familywise error rate, FWER** と呼ぶ．多重仮説検定においては，この FWER が有意水準 α 以下になるように検定をデザインする．

いま，E_k を「帰無仮説 H_0^k が正しいとき，H_0^k を誤って棄却する」事象とする ($E_k = \{H_0^k$を誤って棄却する $\}$)．定義から，$P_{H_0^k}(E_k) = \alpha_k$ は H_0^k の検定の有意水準である．ただし，$P_{H_0^k}$ は帰無仮説 H_0^k の下での確率，という意味である．それに対して，FWER は以下のように定義される．

$$\text{Familywise error rate, FWER } = P(E_1 \cup \cdots \cup E_K) \tag{4.24}$$

ただし，(4.24) における確率は，全ての帰無仮説 H_0^k が正しいと言う仮定の下での確率である．FWER (4.24) を有意水準 α 以下になるようにする方法として，以下の二つが知られている．（他にも方法はあるが，ここでは触れない．永田，吉田[8] を参照）

Bonferroni の方法

ボンフェロー二の不等式 (Bonferroni inequality, (p. 46), (3.2)) を (4.24) に当てはめると，以下を得る．

$$\text{FWER} = P(E_1 \cup \cdots \cup E_K) \leq \sum_{k=1}^{K} P_{H_0^k}(E_k) = \sum_{k=1}^{K} \alpha_k \tag{4.25}$$

FWER を α 以下にするためには，個々の検定の有意水準 $P_{H_0^k}(E_k) = \alpha_k$ を $\alpha_k \leq \alpha/K$ とすれば良い．（もちろん，他にも $\sum_{k=1}^{K} \alpha_k < \alpha$ とする α_k の

定め方も存在する）このとき，$\sum_{k=1}^{K} P_{H_0^k}(E_k) \leq K \times (\alpha/K) = \alpha$ となる．個々の検定の p 値を p_k とすれば，「p 値と有意水準の関係 (p. 125)」により，有意水準を $\alpha_k \leq \alpha/K$ とすることは　$p_k \leq \alpha/K \iff p_k^* = Kp_k \leq \alpha$ に対して H_0^k を棄却することと同値である．$p_k^* = Kp_k$ を，ボンフェローニの方法の**修正 p 値 (adjusted p-value)** という．すなわち，ボンフェローニの方法とは K 個の検定の多重仮説検定を行うとき，個々の検定の修正 p 値 $p_k^* = Kp_k$ が有意水準 α 以下になるとき帰無仮説を棄却する方法である．

Holm の方法

本項では，多重仮説検定においてボンフェローニの方法とおなじ条件で使用出来る一方で，ボンフェローニの方法よりも帰無仮説を棄却しやすいホルムの方法 (Holm's method) を紹介する．ホルムの方法の手順は，以下の通りである．

1. 帰無仮説のファミリーを，$\mathcal{F} = \{H_0^1, H_0^2, \ldots, H_0^K\}$ とする．それぞれの検定の p 値を，$p_1, p_2, \ldots, p_K$ とする．
2. 検定の p 値を大小順に並べ直して，$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$ とする．k 番目に小さい p 値である $p_{(k)}$ に対応する帰無仮説を $H_0^{(k)}$ とする．
3. 有意水準を α とするとき，$\alpha_1 = \alpha/K, \alpha_2 = \alpha/(K-1), \ldots, \alpha_K = \alpha$ とする．
4. $p_{(1)} > \alpha_1$ ならば，全ての帰無仮説を棄却せず終了する．$p_{(1)} \leq \alpha_1$ ならば，$H_0^{(1)}$ を棄却して次に進む．
5. $k = 2, \ldots, K$ に対して，$p_{(k)} > \alpha_i$ ならば $H_0^{(k)}, \ldots, H_0^{(K)}$ を全て棄却せずに終了する．$p_{(k)} \leq \alpha_i$ ならば，$H_0^{(k)}$ を棄却して $k \to (k+1)$ とする．これを，$k = K$ となるまで繰り返す．

このとき，FWER (4.24) はやはり有意水準 α 以下になることが示せる．証明は，永田，吉田[8] を参照．$p_{(k)} \leq \alpha_i = \alpha/(K-k+1)$ とすることは，$(K-k+1)p_{(k)} \leq \alpha$ とすることと同値である．$p_{(k)}^{\dagger} = (K-k+1)p_{(k)}$ を，ホルムの方法の**修正 p 値 (adjusted p-value)** という．（ただし，$p_{(k)}^{\dagger} > p_{(k+1)}^{\dagger}$ となるときは，$p_{(k+1)}^{\dagger} = p_{(k)}^{\dagger}$ で置き換える．また，$p_{(k)}^{\dagger} \geq 1$ となるときは，

$p^{\dagger}_{(k)} = 1$ で置き換える.）ホルムの方法とは，多重仮説検定において個々の検定の修正 p 値 $p^{\dagger}_{(k)} = (K-k+1)p_{(k)}$ が有意水準 α 以下になるとき $H_0^{(1)}, \ldots, H_0^{(k)}$ を棄却し，$p^{\dagger}_{(k+1)} > \alpha$ となるとき $H_0^{(k+1)}, \ldots, H_0^{(K)}$ を棄却しない方法である.

多重検定を行うため，R には p.adjust() コマンドが用意されている．p.adjust() コマンドの引数は，K 個の検定の p 値を集めた数値ベクトルと多重検定の方法を指定する method オプションであり，返値はそれぞれの方法による修正 p 値である.

例題 4.11. $K = 10$ 個の仮説のファミリー $\mathcal{F} = \{H_0^1, H_0^2, \ldots, H_0^K\}$ の，多重仮説検定を考える．それぞれの検定の p 値を $p_k = 0.001 \times k, k = 1, \ldots, 10$ とする．このとき，ボンフェローニの方法とホルムの方法の修正 p 値は，以下のように与えられる.

```
> p <- seq(from=0.001, to=0.010, by=0.001) # p-values
> p
 [1] 0.001 0.002 0.003 0.004 0.005 0.006 0.007 0.008 0.009 0.010
>
> p.adjust(p, method="bonferroni") # Bonferroni's method
 [1] 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10
>
> p.adjust(p, method="holm") # Holm's method
 [1] 0.010 0.018 0.024 0.028 0.030 0.030 0.030 0.030 0.030 0.030
```

有意水準 $\alpha = 0.05$ とすると，個々の検定の p 値 p_k は全て *0.05* 以下であるから一つずつの検定は有意に帰無仮説を棄却する．ボンフェローニの方法の場合，修正 p 値は $p_k^* = 0.01 \times k$ であるから，$H_0^1, \ldots, H_0^5$ までは $p_k^* \leq \alpha$ であるから棄却されるが，$H_0^6, \ldots, H_0^{10}$ は棄却されないことになる.

一方，*Holm* の方法の場合，全ての修正 p 値が $p_k^{\dagger} \leq \alpha$ であるから，$H_0^1, \ldots, H_0^{10}$ は全て棄却される.

ボンフェローニの方法とホルムの方法の修正 p 値を比較すると，$p_k^* = Kp_k \geq$

$p_k^\dagger = (K-k+1)p_{(k)}$ であるから一般にホルムの方法の方が帰無仮説を棄却しやすく，その意味でボンフェローニの方法の改良になっている．一方ボンフェローニの方法は，検定の数 K が大きくなると α/K が極端に小さくなり帰無仮説を棄却しにくくなるため，過度に**保守的 (conservative)** な方法とも呼ばれる．実際，R の `p.adjust()` コマンドのオンラインヘルプには，以下のような記載がある．

> There seems no reason to use the unmodified Bonferroni correction because it is dominated by Holm's method, which is also valid under arbitrary assumptions.

ボンフェローニの方法は導出は容易であるがあくまで教育目的であり，実際のデータ解析ではホルムの方法を優先することをお勧めする．

4.3.7 p 値と検出力とサンプルサイズの設計

本節では，ここまで統計的仮説検定の概念とその一般的な手順について議論してきた．最後に，実際に研究をデザインする際に目的とする差を検出するのに必要にして十分なサンプル数を求める「サンプルサイズの設計」と，そのために必要な「検出力」の概念について検討する．

4.3.7.1 検出力

まず，本節の最初に検討した一標本問題の正規母集団既知分散の場合をもう一度考えてみよう．サンプル $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ に対して，以下の仮説 (p. 121) を考える．

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1 (> \mu_0) \tag{4.19}$$

この時，検定統計量は $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ (4.20) で与えられた．対立仮説 $H_1 : \mu = \mu_1 (> \mu_0)$ に対して，検定の棄却域は $\{z : z > z_\alpha\}$ (4.23, p. 125) ただし z_α は標準正規分布 $N(0, \sigma^2)$ の上側 α パーセント点となる．さらに検定の p 値は，帰無仮説 H_0 のもとで検定統計量がその観測値と同程度かより甚だしい (as extreme or more extreme) 値をとる確率のことであったから，検定統計量の

観測値を $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ とすれば，H_0 のもとで検定統計量 Z は標準正規分布に従うから，

$$\text{仮説検定 (4.19) の } p \text{ 値} = P\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \qquad (4.26)$$

ただし，$Z \sim N(0,1)$ で $\Phi(z)$ は標準正規分布の累積分布関数である．ここで，これまであまり表立って考えてこなかった第二種の過誤　(type II error) についても考えてみよう．第二種の過誤とは帰無仮説 H_0 が偽であって対立仮説 H_1 が真であるとき，誤って H_0 を受け入れ H_1 を棄却する過誤であった．(4.22, p. 124) によれば対立仮説 H_1 の下で，検定統計量 Z は期待値 $(\mu_1 - \mu_0)/(\sigma/\sqrt{n})$，分散 1 の正規分布に従う．この時，$H_0$ を受け入れるとは Z が採択域 $\{z : z \leq z_\alpha\}$ に値をとることであるから，第二種の過誤の確率を β とすると β は以下のように得られる．

$$\begin{aligned}
\beta &= P_{H_1:\mu=\mu_1}(Z \leq z_\alpha) = P_{H_1}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha\right) = P_{H_1}\left(\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \\
&= P_{H_1}\left(\bar{X} - \mu_1 \leq \mu_0 - \mu_1 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = P_{H_1}\left(\underbrace{\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}}}_{\sim N(0,1)} \leq z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) \\
&= P(Z \leq z_\alpha - \sqrt{n}(\mu_1 - \mu_0)/\sigma) \quad : \bar{X} \sim N(\mu_1, \sigma/\sqrt{n}) \text{ under } H_1 \\
&= \Phi\left(z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(z_\alpha - \sqrt{n}\Delta\right) \qquad (4.27)
\end{aligned}$$

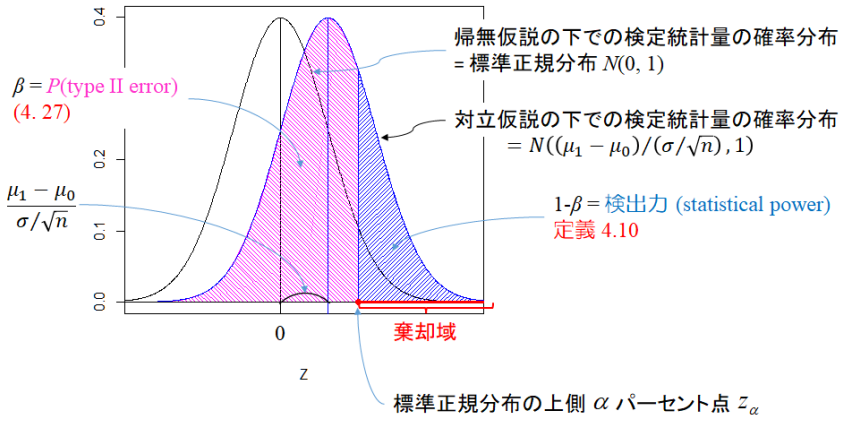
ただし，$\Delta = (\mu_1 - \mu_0)/\sigma$ であり，$\Phi(z)$ は標準正規分布の累積分布関数である．このとき，以下の概念を定める．

定義 4.10. いま，$\beta = P(\textit{tyep II error})$ = 第二種の過誤の確率とする．この時，$(1 - \beta)$ を**検出力** ***(power)*** と呼ぶ．

定義 4.11. $\Delta = (\mu_1 - \mu_0)/\sigma = \delta/\sigma,\ \delta = (\mu_1 - \mu_0)$ を，検定問題 *(p. 121)* の**効果量** ***(effect size)*** と呼ぶ．

すなわち，検出力 $(1 - \beta)$ とは対立仮説 H_1 が真であるとき，帰無仮説を棄却し正しく対立仮説を受け入れる確率のことである．また効果量 Δ とは，帰無

図 **4.6** 検出力 (statistical power)



仮説と対立仮説の乖離あるいは帰無仮説が正しくない程度を量的に測る尺度である．

まず, (4.26) からサンプル数 n が小さいと検定統計量の観測値 $z = \sqrt{n}(\bar{x}-\mu_0)/\sigma$ も小さくなり，p 値が大きくなる．「p 値と有意水準の関係 (p. 125)」と合わせて考えれば，サンプル数 n が小さいと帰無仮説 H_0 を棄却しにくくなり有意差を検出しづらくなる．

また，(4.27) から $\beta = P(\text{type II error})$ は，1）$\Delta = (\mu_1 - \mu_0)/\sigma$ が小さいとき，2）サンプル数 n が小さいとき，のいずれの場合も大きな値をとり，同時に検出力 $(1-\beta)$ は小さくなる．

すなわち，十分な大きさのサンプル数 n を確保できなければ，帰無仮説を棄却して有意差を見いだすことは困難になり，また第二種の過誤の確率も大きくなる．同時に，実験をデザインする際に見出したい差 $\Delta = (\mu_1 - \mu_0)/\sigma$ が小さければ，その分大きな n が必要になると言うことでもある．

それでは，サンプル数 n は大きければ大きいほどよいのであろうか．じつは，帰無仮説を棄却し大きな検出力を確保するには十分大きいサンプル数 n が必要である一方で，大きすぎる n には以下のような問題がある．

1. サンプルの収集には多くの費用と時間が必要であり，多すぎるサンプルを集めることが困難となる場合がある．
2. サンプル数を多くすれば一般的に p 値が小さくなり，帰無仮説を棄却しやすくなる．その結果，臨床的に意味のない差異を統計学的に有意な差として検出してしまう可能性がある．
3. 例えば，新薬の効能を調べる臨床試験などの場合，事前には予期できないような副作用が発生する場合がある．不必要に多くのサンプルを集めることは，実験に参加する被験者に必要以上のリスクを負わせる面がある．

以上の両面から，実験をデザインする際には必要にして十分な数のサンプルサイズを設計する必要がある．

4.3.7.2 サンプルサイズの設計

それでは，検定 (p. 121) で有意水準 α，検出力 $(1-\beta)$，効果量 $\Delta = (\mu_1 - \mu_0)/\sigma$

が与えられたとき，必要なサンプル数 n を求めよう．(4.27) より，検出力 $(1-\beta)$ にたいして

$$\text{Power} = 1-\beta = 1-\Phi\left(z_\alpha - \sqrt{n}\Delta\right) \iff \beta = \Phi\left(z_\alpha - \sqrt{n}\Delta\right)$$

これは，$(z_\alpha - \sqrt{n}\Delta)$ が標準正規分布 $N(0,1)$ の下側 β パーセント点 $-z_\beta$ に等しいことと同義である．（z_β は $N(0,1)$ の上側 β パーセント点）すなわち

$$\begin{aligned} -z_\beta = z_\alpha - \sqrt{n}\Delta &\iff \sqrt{n}\Delta = \sqrt{n}\frac{\delta}{\sigma} = (z_\alpha + z_\beta) \\ &\iff \sqrt{n} = \frac{z_\alpha + z_\beta}{\delta/\sigma} \iff n = \left(\frac{z_\alpha + z_\beta}{\delta/\sigma}\right)^2 \end{aligned}$$

例題 4.12. 第 *4.3.1* 節 *(p. 127)* では，*Low Infant Birth Weight* データを用いて幼児の出生体重に対する正規母集団既知分散の一標本問題の仮説検定を考えた．ここでは，帰無仮説，対立仮説を以下のように固定して考える．

$$H_0 : \mu = \mu_0 = 2800 \textit{ vs. } H_1; \mu = \mu_1 = 3000$$

第 *4.3.1* 節と同様，`bwt` は独立かつ同一に期待値 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うとし，分散は $\sigma^2 = 730.0^2$ で既知であると仮定する．このとき，有意水準 $\alpha = 0.05$，検出力 $(1-\beta) = 0.8$ とすると，

$$\begin{aligned} \Delta = \delta/\sigma = (\mu_1 - \mu_0)/\sigma = (3000 - 2800)/730 \\ n = \left(\frac{z_\alpha + z_\beta}{\delta/\sigma}\right)^2 \end{aligned}$$

```
> alpha <- 0.05# significance level #
> beta <- 0.2# (1 - power) #
> mu1 <- 3000
> mu0 <- 2800
> delta <- (mu1 - mu0)
> sigma <- 730
> n <- ((qnorm(alpha, lower.tail=F) + qnorm(beta, lower.tail=F))/(delta/sigma))^2
```

```
> n
[1] 82.36712
```

よって，必要なサンプル数は整数値に切り上げて $n = 83$ となる．

一般に，サンプルサイズの設計には以下の 5 項目を特定する必要がある．

有意水準 帰無仮説 H_0 が真であるとき，H_0 を棄却する第一種の過誤 (type I error) の確率 α．多くの場合，$\alpha = 0.05$ が用いられる．片側検定の場合には，$\alpha = 0.025$ が用いられることもある．これは，$\alpha = 0.025$ とした場合の片側検定の棄却域が，$\alpha = 0.05$ とした場合の両側検定の棄却域の片側と同じ領域になるためである．

検出力 対立仮説 H_1 が真であるとき，H_0 を棄却し H_1 を採択する確率．すなわち第二種の過誤 (type II error) の確率 β に対して，検出力=$(1-\beta)$．一般的には，検出力として 0.8[6] 以上が推奨されることが多いようであるが，場合によって異なる値が用いられる場合もある．

効果量 効果量 Δ は，帰無仮説 H_0 と対立仮説 H_1 の乖離 δ を標準化したものである．効果量は，臨床的に意味のある差異として設定されることが重要である．効果量を操作すれば必要なサンプル数も操作することが可能であるが，あくまで臨床的に意味のある効果量が先にあり，それを検出するのに必要なサンプル数を求める，という立場を堅守することが必要である．

分散 サンプルの分散 σ^2（あるいは標準偏差 σ）は，効果量を定義するのに必要である．σ^2 は先行研究や小規模な実験 (pilot study) によって設定されることが望ましいが，そのような情報が得られない場合は常識的に許容できる値を慎重に定める必要がある．

脱落率 ヒトを対象とした臨床試験などの場合，当初予定したサンプル収集の過程で登録者が脱落することが起こりうる．必要なサンプル数 n に対して k%の脱落が想定されるのであれば，実際には $n^* = n/((100-k)/100)$ のサンプルを集める必要がある．例題 4.12 であれば，必要サンプル数 $n = 83$ に対して 20%の脱落が想定されるのであれば，実際には $n^* = 83/((100-20)/100) =$

[6] 例えば，Cohen[2], p. 56 を参照

103.75 ≃ 104 例のサンプルが必要になる．そうすれば，仮に 20%が脱落しても $104 \times 0.8 = 83.2$ となり，必要なサンプル数 $n = 83$ が確保されることになる．

4.3.7.3 p 値の意義と限界

付　　録：R の導入と使い方

この付録では，フリーの統計解析ソフトウエア R の導入と使い方の紹介を行う．ただし，R の使い方については本書を読むのに必要な最小限の説明にとどめる．この付録の内容では R を使いこなすには全く不足するので，全般的な R の詳細については他書（船尾[4]，あるいは船尾[4] の簡約版がインターネット上で無料で公開されている R-Tips : http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html など）を参照されたい．

ソフトウエア R は統計計算とグラフィックスのための言語･環境であり，R と付属文書は GNU general public licence (http://www.gnu.org/licenses/gpl.html) の下に自由に配布されている．R は Windows, Mac OS X, Linux などの OS 上にインストール可能であるが，本書は Windows の利用を前提とする．他の OS を使用する読者は，例えば RjpWiki (http://www.okadajp.org/RWiki/) の該当部分などを参照し適宜読み替えていただきたい．本書執筆時点での，R のヴァージョンは R 3.4.3 である（2017 年 11 月 30 日現在）．R のヴァージョンは頻繁に更新されるが，さほど最新版にこだわる必要はない．

.1　R のインストール（Windows）

R のソースコードおよびそのバイナリは Comprehensive R Archive Network (CRAN) のウェブサイト (http://www.r-project.org/) あるいはそのミラーサイトから入手することが出来る．R のセットアッププログラムを入手するため，インターネットに接続した PC 上でブラウザを起動し，以下の URL にアクセスする．

https://cloud.r-project.org/

図 1 が表示されるので，**Download R for Windows** を選択する．次に，図 2 が表示されるので，**Subdirectionaries:** の下の **base** を選択する．

図 1

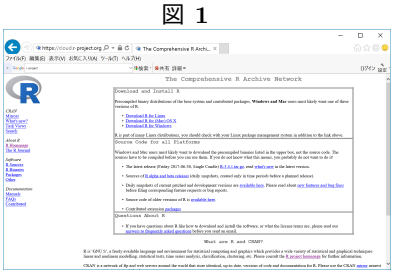
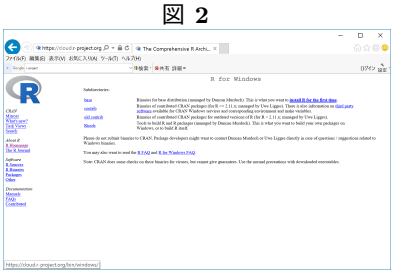


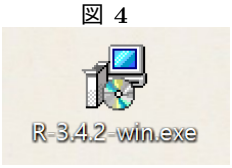
図 2



Windows 用 R セットアップファイルダウンロード画面（図 3）で，**Download R 3.4.3 for Windows** をクリックする．セットアップファイル (R-3.4.3-win.exe) を選択し適当なフォルダ（例えばデスクトップに）にダウンロードする．（図 4）

図 3

図 4



セットアップファイル (R-3.4.3-win.exe) をダブルクリックすると，「ユーザーアカウント制御」ダイアログボックスが表示されるので，「はい」を選択する．「セットアップに使用する言語の選択」ダイアログボックスが開くので，「日本語」を選択し「OK」を押す．「R for Windows 3.4.3 セットアップ」ダイアログ

ボックスと「GNU GENERAL PUBLIC LICENSE」に関する説明が表示されるので，どちらも「次へ」を押す．次の「インストール先の指定」「コンポーネントの選択」ダイアログボックスも，特に理由がなければそのまま「次へ」「次へ」を押す．「起動時オプション」の選択，「プログラムグループの指定」も，そのまま「次へ」を押す．

「追加タスクの選択」（図 5）では，必要な追加タスクを選択する．デフォルトでは「クイック起動アイコンを作成する」はチェックされていないが，これも選択しておくと便利である．「次へ」を押すと，R のインストールが開始される．インストールが終了したら，「完了」ボタンを押してインストールを完了する．デスクトップには R のアイコン（図 6）が出来ているはずである．

図 5

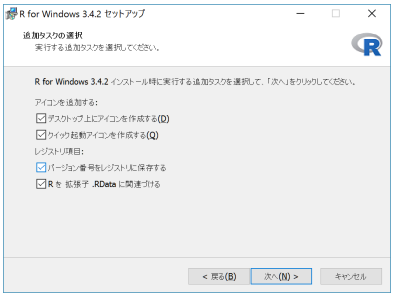
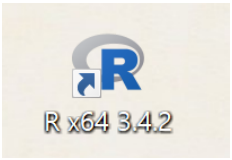


図 6



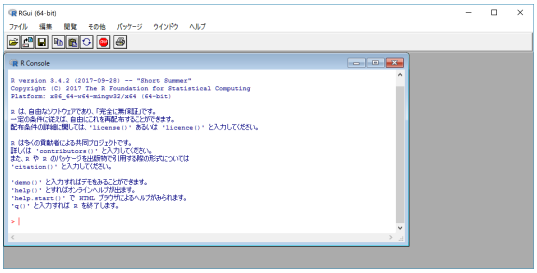
.2 R の起動と終了

さて，R がインストールされたら，早速 R を使ってみよう．R を起動するには，以下の方法がある．

- デスクトップ上の R のアイコンを，ダブルクリックする．
- 「スタート」ボタンから，「すべてのプログラム」→「R」→「R 3.4.3」をクリックする．

R を起動すると，起動直後の画面は図 7 のようになる．初期画面の中には「R

図 7 R の初期画面



Console」というウィンドウがある．この R Console ウィンドウの中にコマンドを入力すると，同じ画面に計算結果が出力される．その意味で，R は「対話型プログラム」と呼ばれる．例えば log(10) を計算すれば，以下の結果を得る．

```
> log(10)
[1] 2.302585
```

2 行目の先頭にある “[1]” は，出力結果の要素の 1 番目，という意味である．R では，四則演算や上記の対数関数 log() の他，指数関数 exp()，三角関数 sin(), cos(), tan()，逆三角関数 asin(), acos(), atan() などの初等関数の値を求めるコマンドがある．さて，R を使い始めたばかりであるが，ここでいったん R を終了してみよう．R を終了するには，以下の方法がある．

- 「ファイル」メニューから「終了」を選択する．
- R の画面の右上隅の「×」印を押す．
- R Console ウィンドウの中で，終了コマンド q() を実行する．

```
> q()
```

R 終了時には「質問」という小さなウィンドウが開き，「作業スペースを保存しますか？」と聞かれる．「はい (Y)」を選択すればそれまでの作業を保存し次回再利用することが出来る．しかし，**「いいえ (N)」**を選択し何も保存しないこ

とを，強くお薦めする．いたずらに作業を保存しても，次回までに保存した内容を忘れてしまい要らぬ混乱を招くのが関の山である．作業内容を保存したいときは，第.6.3 に有るようにプログラムを作成して保存する．もし R 内部に保存されたものが分からなくなったときは，「その他」メニューから「その他」→「すべてのオブジェクトの消去」と辿ると「質問」ダイアログボックスで「本気ですか？」と聞かれるので，「はい」を押して保存されたものをすべて消去してしまう．R の中は，いつもきれいに掃除しておこう．

.3 R の操作

.3.1 オブジェクトと付値

それでは，ふたたび R を起動しよう．R において扱われる様々な変数，関数，データなどの「もの」は，すべて**オブジェクト (object)** と呼ばれる．R において最も基本的なオブジェクトは，**ベクトル (vector)** である．R においてベクトルは，R のコマンド `c()` によって定義される．例えば，以下の構文を R Console ウィンドウに入力することで，要素の数が 5 の実数値ベクトルを定義することが出来る．

```
> x <- c(3, 4, 10, 6, 5)
> x
[1]  3  4 10  6  5
```

1 行目の “<-” は左向き矢印を意味し、「右辺」(この場合は `c(3, 4, 10, 6, 5)`) の実行結果を「左辺」のオブジェクト (この場合は `x`) に代入し保存するための記号である．このようにあるオブジェクトの値を他のオブジェクトに代入することを，**付値 (assign)** する，という．上の例の場合，`c()` コマンドの実行結果が付値されたためオブジェクト `x` の値もまた，要素数 5 の実数値ベクトル $(3, 4, 10, 6, 5)$ となったことを示している．

.3.2 データの型

ベクトルの要素となるデータの型（属性）には，上に挙げたオブジェクト `x` の

ような**数値型 (numeric)** の他，**文字列型 (character)**，**論理型 (logical)**，**因子型 (factor)** などがある．文字列型ベクトルは，引用符（" " あるいは ' '）で囲まれた文字列を要素として持つベクトルである．例えば，以下のベクトル y は長さ 5 の文字列型ベクトルで，最初の要素は 4 文字のアルファベットからなる “jack” という文字列である．

```
> y <- c("jack", "jack", "ginger", "eric", "eric")
> y
[1] "jack"   "jack"   "ginger" "eric"   "eric"
```

論理型ベクトルは，TRUE もしくは FALSE の論理値を要素に持つベクトルである．上に定義したオブジェクト x を用いて，以下の論理式を実行してみよう．

```
> (x > 4.5)
[1] FALSE FALSE  TRUE  TRUE  TRUE
```

x の要素は 3, 4, 10, 6, 5 であるが，その一つ一つに対して「4.5 より大きい $(x > 4.5)$」という命題が評価され，例えば x の最初の要素である 3 は 4.5 より大きくはないため **FALSE (偽)** という論理値が，三番目の要素である 10 は 4.5 より大きいため **TRUE (真)** という論理値が返されている．TRUE，FALSE は，それぞれ T，F と省略することが可能である．

もう一つのデータの型である因子型ベクトルは，同じ長さを持つ別のベクトルの要素のグループ化を行うデータ型であり，factor() コマンドによって定義される．以下の例を考えよう．

```
> y.f <- factor(y)
> y.f
[1] jack   jack   ginger eric   eric
Levels: eric ginger jack
```

上で定義した y は，文字型ベクトルであった．y を表示すると，y の要素が "jack" のように文字列であることを示す引用符 "" 付きで示されることが分かる．一方，y に factor() コマンドを作用させた y.f は因子型に**型変換**され，因子型ベ

クトルになっている．y.f を表示すると，y.f が文字型でない証拠に y.f の要素には引用符（" "）がついていない．また y.f は要素のグループ分けとして eric，ginger，jack の三つの**レベル (Level)** を持つことが分かる．レベルの名前（**ラベル (Label)**）を変えるには，labels オプションで指定する．例えば、eric，ginger，jack を ERIC，GINGER，JACK に変えるには以下のようにする．

```
> y.f <- factor(y, labels=c("ERIC", "GINGER", "JACK"))
> y.f
[1] JACK   JACK   GINGER ERIC   ERIC
Levels: ERIC GINGER JACK
```

因子型のレベルは，数字の大小順あるいはアルファベット順に順序を持つ．もしそれ以外の順序を指定する際は、以下のように levels オプションで指定する。

```
> y.f <- factor(y, levels=c("jack", "eric", "ginger"))
> y.f
[1] jack   jack   ginger eric   eric
Levels: jack eric ginger
```

このようにすると、z.f のレベルがアルファベット順ではなく，指定されたとおり jack，eric，ginger となっていることが分かる．

.3.3 行列，リスト，データフレーム

ベクトルは，行列の形にまとめることも出来る．R で行列を定義するコマンドは，matrix(), cbind(), rbind() などがある．matrix() は入力されたベクトルを行列の形に変形する．以下にその例を示す．

```
> mtx <- matrix(1:8, nrow=2, ncol=4)
> mtx
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
> mtx2 <- matrix(1:8, nrow=2, ncol=4, byrow=TRUE)
```

```
> mtx2
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
```

ここで，**1:8** の ":" は $1, 2, \ldots, 8$ のような数列を生成する関数である．また，`matrix()` コマンドの中の `nrow`, `ncol` はそれぞれ行数，列数を指定するオプションである．上の例にあるとおり，`matrix()` コマンドは入力されたベクトルの要素を列優先で並べるが，`byrow=TRUE` オプションを与えれば行優先で並べられる．これに対して，`cbind()`, `rbind()` は column bind, row bind の意であり，同じ長さの複数のベクトルをそれぞれ列ごとあるいは行ごとに行列の形にまとめる．

```
> x1 <- 1:3
> x2 <- 5:7
> cbind(x1, x2)
     x1 x2
[1,]  1  5
[2,]  2  6
[3,]  3  7
> rbind(x1, x2)
   [,1] [,2] [,3]
x1    1    2    3
x2    5    6    7
```

これまで述べてきたベクトル，行列は，その要素がすべて同じデータの型をしている必要があった．これに対して，**リスト (list)** は異なる型と異なる長さを持ったベクトルや行列等のオブジェクトを一纏めにしたものである．リスト型オブジェクトは `list()` コマンドによって定義される．これまで作ったオブジェクトを使って，以下の例を考えよう．

```
> L <- list(x1, y, y.f)
> L
```

```
[[1]]
[1] 1 2 3

[[2]]
[1] "jack"   "jack"   "ginger" "eric"   "eric"

[[3]]
[1] jack   jack   ginger eric   eric
Levels: jack eric ginger
```

ここでリスト L は，最初の要素が長さ 3 の数値型ベクトル，2 番目の要素が長さ 5 の文字列型ベクトル，等，長さも属性も異なる要素を持ったリストになっている．リスト型オブジェクトの特別な場合として，データフレーム（data frame）と呼ばれるものがある．データフレームとは，長さの等しいベクトルを纏めたリストであり，個々のベクトルの型は異なっていてもかまわない．データフレームは，data.frame() コマンドによって定義される．

```
> Data <- data.frame(id = 1:3,
                     group = c("A", "A", "B"),
                     x = c(5,2,3))
> Data
  id group x
1  1     A 5
2  2     A 2
3  3     B 3
```

データフレームの形は行列と同様であるが，各列ごとに「変数名」に当たるベクトル名（id，group，x）が付いている．また，上の例では Data オブジェクトの二番目の要素 group は group = c("A", "A", "B") のように文字型として定義されているが，data.frame() コマンドでデータフレームの要素として定義された後は因子型に型変換されている（引用符 "" がない）ことに注意する．

.4 要素の取出し

ここまでに導入したベクトル，行列は，数字，文字などのいくつかの要素をまとめたものである．また，リスト，データフレームはいくつかのベクトル，行列等をまとめたものであった．データを操作する際これらのオブジェクトの一部を取り出すには，その要素を指定 (indexing) することで行われる．ベクトル，行列の要素の指定は，オブジェクト名の後ろに四角括弧 `[...]` を付け，1) 要素の場所を数字で指定する，あるいは 2) 論理値により，取り出す要素を指定する．例えば上の例でオブジェクト `x` は，長さ 5 の数値型ベクトルである．

```
> x
[1]  3  4 10  6  5
```

数字で要素の場所を指定する，例えば `x` の 3 番目と 5 番目の要素を取り出すには，`x` の後に付けた四角括弧 `[...]` の中に取り出す要素の位置を示す数値ベクトルを入れて，以下のようにする．

```
> x[c(3, 5)]
[1] 10  5
```

行列の場合，第 i 行を取り出すなら `mtx[i,]`，第 j 列を取り出すなら `mtx[,j]`，(i, j) 要素を取り出すなら，`mtx[i, j]` のように指定する．複数の行あるいは列を指定することも出来る．

```
> mtx
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
> mtx[1,]
[1] 1 3 5 7
> mtx[,3]
[1] 5 6
> mtx[2,4]
```

```
[1] 8
> mtx[(1:2), c(2,4)]
     [,1] [,2]
[1,]    3    7
[2,]    4    8
```

ベクトルから論理値を用いて要素を取り出す場合は，ベクトルと同じ長さの論理ベクトルを用意する．論理ベクトルの要素が TRUE（真）である場所に対応する要素が取り出される．

```
> x > 4.5
[1] FALSE FALSE  TRUE  TRUE  TRUE
> x[x > 4.5]
[1] 10  6  5
```

一行目の `x > 4.5` は `x` と同じ長さの論理ベクトルであるが，その下では $(x > 4.5)$ なる条件が満たされた要素のみが（すなわち，TRUE に対応する要素のみが）返されていることが分かる．行列の場合も同様に，論理ベクトルを使って取り出す行および列を指定することができる．

リスト，データフレームの場合，その要素はベクトル，行列などのオブジェクトである．リスト，データフレームの要素には二重四角括弧 `[[...]]` を使ってアクセスする．前に定義したリスト `L` の 1 番目の要素であれば `L[[1]]`，1 番目の要素のベクトルの 2 番目の要素であれば `L[[1]][2]` で取り出せる．

```
> L[[1]]
[1]  3  4 10  6  5
> L[[1]][2]
[1] 4
```

リスト，データフレームの要素（列）が名前を持っていれば，1) `[[...]]` の中で名前を指定する，2) `$`の後ろに名前を続ける，の 2 通りの方法で要素の名前を指定できる．また，データフレームは形式上行列と同じ形をしており，行あるいは列を指定することも出来る．以下の 4 通りの方法は，データフレーム `Data`

の 3 つ目の要素であるベクトル x を指定している．

```
> Data[[3]]
[1] 5 2 3 8 1
> Data[["x"]]
[1] 5 2 3 8 1
> Data$x
[1] 5 2 3 8 1
> Data[,3]
[1] 5 2 3 8 1
```

.5 外部データの入出力

実際の解析において用いられるデータの多くは，Excel などの表計算ソフトのワークシートとして保存されている．本節では，外部データの形で保存されたデータを R に入力する，あるいは R 内部で作られたデータを外部に出力する方法を解説する．

.5.1 データの準備

本節では一度 R を離れ，まず Excel などを用いて元のデータを作成，保存するところから始める．どのような解析を行うにせよ，データを作成する際は必ずやらなければならないこと，逆に絶対にやってはいけないことがある．最低限やるべきことをルーチンワークとして行うだけで，データ解析のミスを大幅に減らすことができる．

1. **データの形は長方形**: データを入力する際は，第一行目に変数名を記入する．多くのソフトウエアは日本語入力に対応しているが，**全角文字は避ける**方が無難である．第二行目以降にデータを記入するが，元データにはグラフや解析の結果を張り付けたりはしない．そうすると，元データの形は，以下のような「長方形」になるはずである．
2. **元データは絶対に改変しない**: データを解析する際，変数を変換したり新

ID	sex	age	height	bodyweight
1	M	64	173	75.4
2	M	69	164	72
3	M	78	155.2	47.2
4	M	83	159.1	60
5	F	73	147.6	40.5

しい変数を定義したりする必要が出てくることがある．このとき，元データを改変してデータを上書きしたり，新しい変数を付け加えたりしてはいけない．データを改変したときは，必ず**新しいファイル名で保存**する．元データを改変した場合，解析を進めるうちに元データが何であったかわからなくなることがある．元データが分からなくなれば，**意図せざるデータの捏造**まであと一歩である．

3. **個人情報は記載しない**: ヒトのデータを扱う場合，個人情報の保護の重要性は改めて述べるまでもないが，残念ながらいまだに氏名やカルテ番号など個人を特定できる情報を記載したままでデータをやり取りする例がみられる．個人情報はデータ解析の立場からは何の意味もないが，万が一外部に流出した場合，データ元の方のプライバシーと研究そのものに重大な被害を与えることになる．解析データの個人情報は**削除**もしくは**匿名化**する，を徹底する必要がある．

.5.2 R への外部データの読み込み

前節のようにしてデータを作成したら，それを Excel で保存する．本書では，サンプルデータとして **appA_Rintro_Data.xlsx** を用意した．これには $x1$, $x2$ の二つの変数，各 100 個のデータが含まれている．データの内容を R に読み込むには，以下のようにする．

1. **タブ区切りテキストファイル形式の保存**: R は Excel ファイルを含め様々なファイル形式のデータを読み込むことができるが，もっとも簡単なのはテキストファイル形式である．テキストファイルにも「タブ区切りテキスト」「CSV（カンマ区切り）」などがあるが，ここでは「タブ区切りテキスト」形式に絞って説明する．

   Excel でファイルをタブ区切りテキスト形式で保存するには、ファイルメニューから「ファイル」→ 「名前を付けて保存」を選択し、「ファイルの種類 (T)」から「テキスト（タブ区切り）(*.txt)」を指定して「保存 (S)」ボタンを押す。
2. **作業ディレクトリ (working directory) の指定**: R に外部データを読

み込むため，R を起動する．R では外部データを保存したり，逆にデータやグラフなどの出力先となるディレクトリを**作業ディレクトリ (working directory)** と呼ぶ．R 起動時の作業ディレクトリは，getwd() コマンドで確認できる．

```
> getwd()
[1] "C:/Users/＊＊＊＊＊/Documents"
```

R で作業ディレクトリを指定するには，以下の二通りの方法がある．

- R Console ウィンドウをアクティブにする．（R Console ウィンドウの上で，マウスをクリックする） 「ファイル」メニューから「ファイル」→「ディレクトリの変更...」を選択しする．「フォルダーの選択」ダイアログボックスが表示されるので，作業ディレクトリを探して指定する．
- R Console ウィンドウから，**setwd()** コマンドで指定する．setwd() コマンドには，作業ディレクトリの絶対パスを入力する．絶対パスは，エクスプローラーの上部にあるアドレスバーをクリックすれば表示される．Windows ではパス名の中のフォルダーの区切りはバックスラッシュ（“¥” あるいは “\”）であるが，R ではスラッシュ “/” であることに注意する．

```
> setwd("C:/Rdataset")
> getwd()
[1] "C:/Rdataset"
```

ここでは例として，C ドライブの直下に “Rdataset” フォルダーを作り，これを作業ディレクトリとする．フォルダーを作る際は，Windows エクスプローラーを起動し C ドライブの直下に移動したうえで，ファイルメニューから「ファイル」→「新規作成」→「フォルダー」を選択する．フォルダー名（たとえば Rdataset）を指定すれば完成である．今作成した **C:¥Rdataset** フォルダーに，上で作成したタブ区切りテキストファイル appA_Rintro_Data.txt を保存する．

3. **read.table() コマンドによる外部データの読み込み**: 作業フォルダー

に保存したタブ区切りテキストファイルを R に読み込むコマンドは，read.table() コマンドである．read.table() の引数は引用符""で囲んだファイル名であり，いくつかのオプションが付く．

```
> Data <- read.table("appA_Rintro_Data.txt", header=TRUE, sep="\t")
```

上では，右辺で read.table() コマンドを実行し，読み込まれたデータを左辺の “Data” オブジェクトに保存している．read.table() コマンドの “header” オプションは，元データの 1 行目に変数名がある（ヘッダーがある）場合には header=TRUE，変数名がないときは header=FALSE となる．（“TRUE”, “FALSE” は，それぞれ “T”, “F” と省略できる．）“sep” オプションは区切り文字 (separator) を指定するオプションで，sep="\t"はタブ区切りであることを示している．（ここで sep=","とすれば，区切り文字にカンマ “,” を指定して CSV ファイルを読み込むことができる．）読み込まれたデータは，データフレーム形式で保存される．最後に，dim() コマンドで Data オブジェクトの行数と列数（次元, dimension）を，head() コマンドで Data オブジェクトの最初の数行を確認しておこう．

```
> dim(Data)
[1] 100   2
> head(Data, n=2)
      x1     x2
1 -2.671  0.862
2 -2.275  1.429
```

ここまでは Windows 上での R の場合であるが、Mac OS の場合，以下のように fileEncoding="CP932"オプションを指定する必要がある．

```
> Data <- read.table("appA_Rintro_Data.txt", header=T, sep="\t", fileEncoding="CP932")
```

.5.3 R から外部へのデータの書き出し

本節の最後に，今度は R 内部のデータを外部の作業ディレクトリに書きだす

方法について説明する．まずサンプルデータとして，以下の例を考える．

```
> n <- nrow(Data)
> n
[1] 100
> y <- runif(n)
>
> Data2 <- cbind.data.frame(Data, y)
> head(Data2, n=2)
      x1     x2          y
1 -2.671  0.862 0.75868476
2 -2.275  1.429 0.98027072
```

最初の nrow() コマンドは行 (row) 数を数えるコマンドである．（同様に列 (column) 数を数える ncol() コマンドもある．）Data は 100 行 2 列のデータフレームであるから，n の値は 100 である．次の runif() コマンドは (0, 1) の間の乱数を生成するコマンドで，引数の数 ($n = 100$) だけ乱数を返す．オブジェクト y は，100 の乱数からなる数値型ベクトルである．次の cbind.data.frame() コマンドは cbind() コマンドに似ているが，データフレーム，行列，ベクトルなどを列方向に結合して，データフレームを作るコマンドである．新しくできたデータフレームを，Data2 オブジェクトに保存している．head() コマンドで Data2 オブジェクトの先頭 2 行を確認すると，$x1, x2$ の隣に新しい列 y ができているのがわかる．

R 内部で生成した行列、データフレームなどのオブジェクトを外部の作業ディレクトリに書きだすコマンドは，write.table() コマンドである．write.table() の第一引数はオブジェクト名，第二引数はデータが書き込まれるファイル名で，さらにいくつかのオプションが付く．

```
> write.table(Data2, "appA_Rintro_Data2.txt", quote=FALSE,
+ row.names=FALSE, col.names=TRUE, sep="\t", append=FALSE)
```

write.table() のデフォルトでは，書き出されたデータは引用符"で囲まれるが，

"が必要ないときは quote=FALSE とする．データフレームの行名，列名を出力するか否かを制御するのが row.names，col.names オプションである．上の例では，行名は必要ないので row.names=FALSE，$x1$, $x2$, y の列名は必要なので，col.names=TRUE としている．sep オプションは出力ファイルの区切り文字を指定するオプションで，タブ区切りで出力するため sep="\t"としている．最後の append オプションは，既存のファイルに書き加えるか上書きするかを指定するオプションで，ファイルに書き加える場合は append=TRUE とする．append のデフォルトは新規作成あるいは上書きする append=FALSE であり，その場合は省略可能である．

.6 その他

本節では，これまで述べてきたもの以外の個別のトピックについて述べる．

.6.1 大文字，小文字，全角文字

R では，アルファベットの大文字と小文字は別のものとして扱われる．"A"と"a"は，全く別のオブジェクトである．R では，ひらがなや漢字で使われる全角文字も使用可能ではある．ただし，現在の R では全角文字を使用する際，表示に問題が起こるようである．格別の理由がない限り，**全角文字は使わない**ことをお勧めする．

.6.2 コメント

R では，「#」以降改行までの同じ行に入力された文字は，コメントとして無視される．以下の例では#より後にある「足し算」という言葉は，R から無視される．

```
> 1 + 2       # 足し算
[1] 3
```

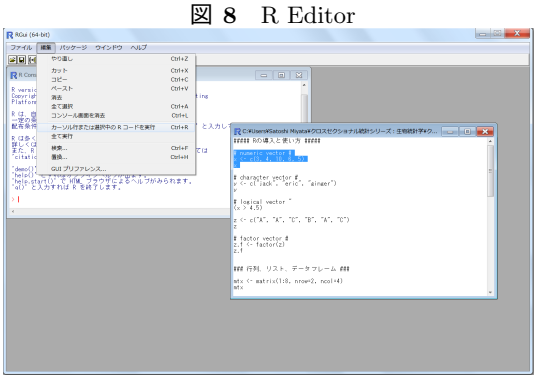
R の入力，あるいは次節に述べる R プログラムには，適宜コメントを挿入することで後で見返したとき理解の助けになる．

.6.3 R プログラムと R Editor

これまでの説明では，R のコマンドや命令を R Console ウィンドウに直接入力してきた．しかしこの方法では，R を終了すると解析の内容が保存されないまま消去されてしまう．解析の内容を保存するためには，R に入力したコマンドや命令をプログラムの形で保存することが望ましい．そのためには，「メモ帳」などの通常のテキストエディタでプログラムを書いてもよいが，R の中で “R Editor” というものを使う方法もある．

R Editor を用いるには，「ファイル」メニューから「ファイル」→ 「新しいスクリプト」を選択する．すると「R Editor」ウィンドウが開く．R Editor の中には R の命令を記入することができる．（図 8）

図 8 R Editor



記入した命令を実行するには，以下の手順に従う．

1. R Editor をアクティブにする（R Editor ウィンドウの上で，マウスをクリックする）
2. 実行したい部分を，マウスで範囲選択する．
3. 以下のいずれかを実行する（どちらでも，同じ結果になる）

   - 「編集」メニューから「編集」→ 「カーソル行または選択中の R コードを実行」を選択する．

- 範囲選択した実行部分の上で，マウスを右クリックする．現れたメニューから「カーソル行または選択中の R コードを実行」を選択する．

作成した R プログラムを保存する際は，「ファイル」メニューから「ファイル」→「名前を付けて保存」あるいは「上書き保存」を選択する．R プログラムの拡張子は，必ず “*.r” とする．このファイルは通常のテキストファイルであるので，「メモ帳」などのテキストエディタやワードプロセッサーで開くことができる．R Editor において，R のプログラムを新規作成するのではなく，すでにある R プログラムを開くときは，「ファイル」メニューから「ファイル」→「スクリプトを開く...」を選び，既存の R プログラムを選択する．

.6.4 グラフのコピーと保存

R の中でグラフを作成した場合，R Graphics ウィンドウにグラフが出力される．レポート作成などのため，このグラフを保存したり，Word などのワープロソフトや PowerPoint などのプレゼンテーションソフトにコピーしたりしたい場合がある．グラフを保存あるいはコピーするには，以下の手順に従う．（図 9）

1. R Graphics ウィンドウをアクティブにする．（R Graphics ウィンドウの上で，マウスをクリックする）
2. 以下のいずれかを実行する（どちらでも，同じ結果になる）

- R Graphics ウィンドウの上で，マウスを右クリックする．

  コピー「メタファイルにコピー...」あるいは「ビットマップにコピー...」を選択する．

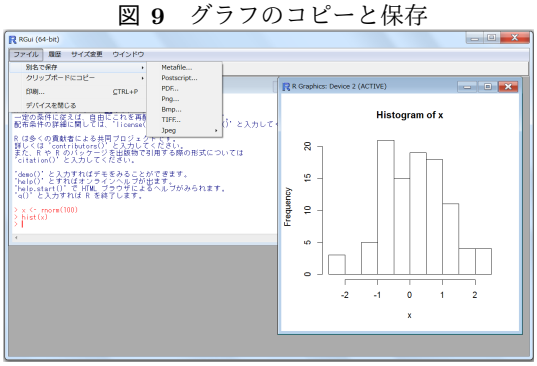
  保存「メタファイルに保存...」あるいは「ポストスクリプトに保存...」を選択する．

- 「ファイル」メニューから

  保存 「別名で保存」 → 保存する画像ファイル形式を選択する．

  コピー 「クリップボードにコピー」 →「ビットマップとして」あるいは「メタファイルとして」を選択する．

図 9 グラフのコピーと保存



あるいは，R で作成した画像を BMP，JPEG，PNG，TIFF 等の多様な画像ファイル形式で保存することも可能である．画像ファイルを保存するコマンドは，ファイル形式にしたがってそれぞれ bmp，jpeg，png，tiff 等である．例えば，以下のプログラムを実行すると作業ディレクトリ上にヒストグラムの画像が保存される．

```
n <- 100
y <- runif(n)

png("histogram.png")
hist(y)
dev.off()
```

png 等のコマンドは，保存するファイル名（ここでは histogram.png）を引数として持つ．オプションにより画像の幅，高さなどを指定することも出来るが詳細はオンラインヘルプを参照して欲しい．png 等のコマンドを使用したら，描画後に必ず dev.off() コマンドによりグラフィックスデバイスをオフにする．

コピーしたグラフは，Word, PowerPoint 等に張り付けることができる．保存したグラフも，同様に挿入することができる．コピーあるいは保存する画像ファイルの形式は，適宜選択する．

.6.5 拡張パッケージのインストールとロード

R の大きな利点の一つに，基本となるシステムの他に多くの研究者によって開発された統計モデルが「拡張パッケージ」として提供されている点が挙げられる．拡張パッケージは元々のシステムにはインストールされていないので，インターネットを通じてダウンロード，インストールする必要がある．

拡張パッケージをインストールするには，インターネットに接続された状態で，「パッケージ」メニューから「パッケージ」 → 「パッケージのインストール...」を選択する．“CRAN mirror” というウィンドウが開くので（図 10），一番上の “0-Cloud” か日本国内のミラーサイト，その他を選択する．次に，利用可能な拡張パッケージのリスト（図 11）が表示されるので，インストールしたいパッケージを選択し「OK」を押す．

図 10
パッケージのインストール

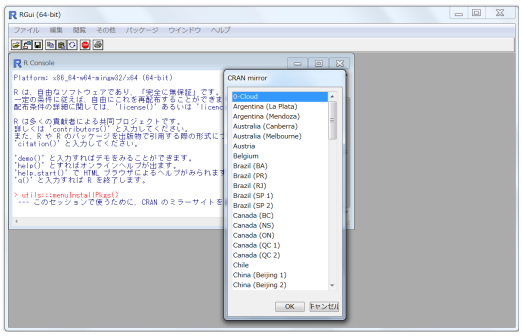
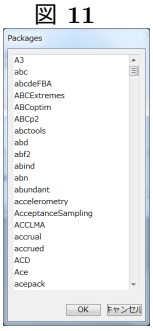


図 11



例えば “MASS” というパッケージをインストールした場合，「パッケージ ‘MASS’ は無事に展開され，MD5 サムもチェックされました」と表示されればインストールは成功である．

拡張パッケージを使用するときは，以下の二通りの方法がある．

- 「パッケージ」メニューから「パッケージ」 → 「パッケージの読み込み...」を選択し，“Select one” ウィンドウから使いたいパッケージを選択する．

- R Console ウィンドウで library() コマンドを使って，使いたいパッケージ名（例えば “MASS”）を指定する．

```
> library(MASS)
```

.6.6 プロキシの設定

基本的には以上の手順により，インストールとパッケージの利用が可能なはずである．ただし，大学などプロキシが設定された環境では，Sys.setenv() コマンドでプロキシサーバーのサーバーアドレスとポート番号を指定する必要がある．R に入力するコマンドは

```
> Sys.setenv(http_proxy="http://******:8080")
```

ただし，http://******はサーバーアドレス，8080 はポート番号であるので，実際のアドレスはシステム管理者に確認していただきたい．

R の R Console ウィンドウの中で，コマンド名の前に “?” をつけて入力すると詳細なオンラインヘルプが得られる．英語であるが，がんばって読んでみよう．例えば，write.table() コマンドのヘルプを得るには，以下のように入力する．

```
> ?write.table
```

表 .1 R の導入と使い方のコマンドリスト

コマンド名	目的	使い方
q()	R を終了する	「作業スペースを保存しますか？」には「いいえ」
c()	オブジェクトを結合する	c(3, 4, 10, 6, 5)
log(), exp() sin(), cos(), tan() asin(), acos(), atan()	対数関数，指数関数 三角関数 逆三角関数	log(10), etc.
factor()	因子型ベクトルを定義する	factor(1:3) ⇒ 数値型ベクトル (1, 2, 3) を因子型に変換する
matrix()	行列を定義する	matrix(1:8, nrow=2, ncol=4) ⇒ 2 行 4 列の行列を定義する
cbind()	列ベクトルを束ねる	cbind(1:3, 5:7)
rbind()	行ベクトルを束ねる	rbind(1:3, 5:7)
list()	リスト型オブジェクトを定義する	list(1:3, c("A", "B"))
data.frame()	データフレームを定義する	data.frame(x=1:3, y=c("A", "B", "C"))
getwd()	現在の作業ディレクトリを示す	getwd()
setwd()	作業ディレクトリを変更する	setwd("C:/Rdataset")
read.table()	外部ファイルを読み込む	read.table("appA_Rintro_Data.txt", header=T, sep="¥t")
write.table()	外部ファイルに書き出す	write.table(Data2, "appA_Rintro_Data2.txt",) quote=F, row.names=F, col.names=T, sep="¥t"))
bmp, jpeg, png, tiff	画像ファイルを保存する	png("histogram.png")

参考文献

1) George Casella and Roger L. Berger. *Statistical inference*. Duxbury Press, North Scituate, MA, 2 edition, 2002.
2) Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.
3) Harald Cramér. *Mathematical methods of statistics*. Princeton University Press, Princeton, 1999.
4) 舟尾暢男. The R Tips 第 3 版: データ解析環境 R の基本技・グラフィックス活用集. オーム社, 2016.
5) Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, New York; Chichester, 3 edition, 2013.
6) David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, New York; Chichester, 3 edition, 2013.
7) 河田敬義, 丸山文行, 鍋谷清治. 大学演習 数理統計［復刊］. 裳華房, 第 11 版, 2011.
8) 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.
9) 尾畑伸明. 数理統計学の基礎 (クロスセクショナル統計シリーズ 1). 共立出版, 2014.
10) Student. The probable error of the mean. *Biometrika*, Vol. 6, No. 20, pp. 1–25, 1908.
11) 竹村彰通. 現代数理統計学. 創文社, 1991.
12) W. N. Venables and Brian D. Ripley. *Modern applied statistics with S*. Springer-Verlag Inc, Berlin; New York, 4 edition, 2002.

13) 柳川堯. 統計数学. 近代科学社, 1990.
14) 柳川堯. ノンパラメトリック法. 培風館, 1997.