統計学の基礎 (統計学テキスト/main2.tex)

□□□□□ □□□□□ □□□□□ □□□□□ □□□□□ □□□□□ 2011 年 8 月 6 日版

目次

1　統計を学ぶために

統計とはどのようなものか？

統計は、「ある目的をもって、一定の条件（時間、空間、標識）で定められた集団を対象に、調べ、集めたデータを、集計、加工して得られた数値」ということができます。より具体的には、統計の特徴は、次のようなものです。

1　統計の第 1 の特徴は、架空のものを対象とするのではなく、存在が明確に規定（定義）された具体的な集団を対象とすることです。統計は、社会現象、経済現象、政治現象、自然現象にとどまらず、個人の意識、行動、感情、能力など、この社会の具体的な現象からなる集団を、すべて対象とすることができます。そしてまた、集団の大きさ、構成及び特性など、広い範囲にわたる内容を対象にしています。

2　統計の第 2 の特徴は、集団を構成する各個体の特定の性質（標識）を数値としてとらえ、集団的に把握することです。統計は、数値により集団的に把握をするからこそ、個々にはばらばらになっていて特徴的な傾向や規則性が見えない現象でも、集団として見た場合、さまざまな傾向や規則性が浮かび上ります。これを統計的規則性といいます。また、統計は数値（数量）で表します。その数値は、世界のだれもが共通的に理解できるもので、そのため極めて客観性が高い特徴を持っています。統計が “ 世界の共通語 ” といわれる理由がここにあります。

統計学の歴史から

統計学の源流は国家または社会全体における人口あるいは経済に関する調査にある。東西を問わず古代から現代まで、政権の勢力維持、把握、統制、将来予測のために行われている。国を統治するための基礎資料として活用されてきた歴史がある。学問としては、17 世紀にはイギリスで W. ペティの『政治算術』(Political Arithmetic) などが著述され、その後の社会統計学につながる流れが始まった。またライプニッツ (微積分の創始、稀代の知的巨人) やエドモンド・ハレー（ハレー彗星、保険数理学）による死亡統計の研究も行われた。これらの影響のもと 18 世紀にはドイツのジュースミルヒが『神の秩序』（1741 年）で人口動態にみられる規則性を明らかにしたが、これには文字通り「神の秩序」を数学的に記述する意図があった。

ドイツでは 17 世紀からヨーロッパ各国の国状の比較研究が盛んになったが、1749 年にアッヘンヴァルがこれにドイツ語で Statistik（「国家学」の意味）の名をつけている。19 世紀初頭になるとこれに関して政治算術的なデータの収集と分析が重視されて、Statistik の語は特に「統計学」の意味に用いられ、さらにイギリスやフランスなどでも用いられるようになった。この頃アメリカ、イギリス、フランスなどで国勢調査も行われるようになる。

一方フランスでは、数学者、物理学者、哲学者、思想家、宗教家で早熟の天才であり、その才能は多分野に及んだブレーズ・パスカル (Blaise Pascal, 1623- 1662:「クレオパトラの鼻」、歯車式計算機、自然科学者)、「数論の父」とよばれるピエール・ド・フェルマー（Pierre de Fermat、1607(1608?)-1665, 数論、解析幾何学）に始まった確率論の研究がフランスを中心にして進み、19 世紀初頭にはラプラス (Pierre-Simon Laplace, 1749 - 1827)（フランスのニュートンともよばれる、天体力学、確率論の解析理論、ナポレオン政権の時代の内務大臣、「天体力学概論」と「確率論の解析理論」) によって一応の完成を見ていた。またレオンハルト・オイラー（Leonhard Euler, 1707 - 1783、数学者・物理学者、天文学者（天体物理学者)。18 世紀最大・最高の数学者、「数学のサイクロプス（単眼の巨人)」) による誤差や正規分布についての研究も統計学発展の基礎となった。ラプラスも確率論の社会的な応用を考えたが、この考えを本格的に広めたのが「近代統計学の父」と呼ばれる

アドルフ・ケトレー (天文学、確率の社会研究への応用) であった。彼は『人間について』(1835 年)、『社会物理学』(1869 年) などを著し、自由意志によってばらばらに動くように見える人間の行動も社会全体で平均すれば法則に従っている (「平均人」を中心に正規分布に従う) と考えた。ケトレーの仕事を契機として、19 世紀半ば以降、社会統計学がドイツを中心に、特に経済学と密接な関係を持って発展する。代表的な人物にはアドルフ・ワグナー (ビスマルク期の経済学者、ワグナーの法則)、エルンスト・エンゲル (社会統計学者、エンゲル係数で有名)、ゲオルク・フォン・マイヤー (ドイツの統計学者) がいる。またフローレンス・ナイチンゲール (イギリスの看護師、社会起業家、統計学者、看護教育学者) も、社会医学に統計学を応用した最初期の人物として知られる。

同じく 19 世紀半ばにダーウィンの進化論が発表され、彼の従弟に当たるゴルトンは数量的側面から進化の研究に着手した。これは当時 Biometrics*(生物測定学) と呼ばれ、多数の生物 (ヒトも含めて) を対象として扱う統計学的側面を含んでいる。ゴルトンは回帰の発見で有名であるが、当初生物学的と思われたこの現象は一般の統計学的対象の解析でも重要であることが明らかとなる。ゴルトンの後継者となった数学者カール・ピアソンはこのような生物統計学をさらに数学的に発展させ (数理統計学)、19 世紀終わりから 20 世紀にかけ記述統計学を大成する。(*注：現在の言い方では生物統計学 Biostatistics に当たり、この単語は現在では生体認証という別の意味で使われている。現代でも多くの統計学者が研究論文の発表の場所として Biometrika 誌を活用)

20 世紀に入ると、W. ゴセット (イギリスの統計学者、ギネス醸造技術者、ペンネームの「スチューデント」)、続いて Sir の称号をもつロナルド A. フィッシャー (1890-1962, 統計学者、進化生物学者、遺伝学者で優生学者。推計統計学の確立者とか現代統計学の父とよばれる) が農学の実験計画法研究をきっかけとして数々の統計学的仮説検定法を編み出し、記述統計学から推計統計学の時代に移る。ここでは母集団から抽出された標本を基に、確率論を利用して逆に母集団を推定するという考え方がとられる。続いてネイマン (現代の推計統計学の中心的理論を確立)、エゴン・ピアソン (仮説検定理論、信頼区間の理論) らによって現代の推計統計学の理論体系が構築され、これは社会科学、医学、工学などの様々な分野へ応用されることとなった。K. ピアソンとフィッシャーは論敵というべき仲だった (息子エゴン・ピアソンの代まで持ち越される有名な「けんか物語」話が伝えられている) が、ゴセットはおだやかな人柄で、両者との交友関係を保ち続けた。アメリカのエイブラハム・ワルド (Abraham Wald,1902-1950、「統計的推定論および仮説検定論について」には、後の統計的判定関数のアイデア。47 年の著書「逐次推測」は、第 2 次世界大戦後の品質管理の要望にこたえたもので、逐次推定理論を体系化し、抜き取り検査の実際場面における難点に回答を与えた。彼の不朽の業績は「統計的判定関数」(1950) である。この本では、ゲームの理論に基づき、ミニマック原理を使って損失関数の概念を導入、推定と検定という従来 2 つに分かれていた分野を統一的に考える視点を定式化した。その年の 12 月、標本調査指導のため、インド統計研究所に招かれ、南インド旅行中に飛行機事故で死亡した。) やテューキー (John Wilder Tukey,1915-2000、アメリカの統計学者。プリンストン大学の数理統計学部、AT&T Bell Laboratories、1965 年に作られたプリンストン大学統計学部の最初の学部長。高速フーリエ変換、探索的データ検索法を考案。The future of data analysis (1962)、Mathematics of Computation (1965) Cooley との共著が有名。彼は、単一実験から得られたパラメータ値のセットにおける分散分析および同時推定の問題に多大な貢献をした。また、Software や Bit という造語を作ったことでも知られる) が有名な統計学に貢献した人々である。

`http://ja.wikipedia.org/` にて検索し引用した。

"Statistics"に「統計」という訳語を与えたのは、明治初年の洋学者・神田孝平 (1830-98) といわれているが、一説には、西周 (1829-97) であったともいわれている。「統計」が確定されるまで、「人別、表記、政表、

経国学、会計学、形勢学、国勢学」なども提唱されていた。日本で初めて「近代的な人口調査」が行われたのは 1869 年で、太政官正院・政表課大主記であった杉亨二（1828-1917）が行った。杉は、「政表」を用いていたが、いちいち訳す必要はないとして、「スタチスチクでよかろう」と「漢字を使わないと感じが出ないなら」、と新漢字も作った。しかし、杉を師と仰いで統計思想の普及に努めていたスタチスチク社の訳語不要論に対して、森鴎外（1862-1922）が「統計」が妥当であろうと論陣をはり、森鴎外全集の著作篇第 26 巻には 6 篇におよぶ論争文が収められている。

川北稔「政治算術」（『歴史学事典 13 所有と生産』（弘文堂、2006 年） ISBN 978-4-335-21042-6）

統計調査：

統計を得る方法として第 1 に挙げられるのは、統計調査を実施し、その結果を集計、加工するものです。第 2 は、必ずしも統計の作成を第一義的目的としないで集められた業務資料を集計、加工する方法です。貿易統計はその例です。これらの方法のほか、各種統計を組み合わせ、加工して統計を得ることもあります。国内総生産（GDP*）はその例です。これらの中で、統計調査の実施は有力な方法であり、多用されています。＊国内総生産（こくないそうせいさん、GDP : Gross Domestic Product）とは、一定期間内に国内で産み出された付加価値の総額。ストックに対するフローをあらわす指標であり、経済を総合的に把握する統計である国民経済計算の中の一指標で、GDP の伸び率が経済成長率に値する。

(i) 統計の種類

統計調査にどのようなものがあるのかを見てみます。統計調査にはいろいろあり、区分の仕方もいろいろ考えられますが、ここでは、実施の主体の観点から区分し、国民の実態を把握する目的で行政機関が行う政府統計調査、民間の会社などが行う民間統計調査、そして、研究者が研究仮説を立て、そのことを検証するために行う研究のための統計調査について説明します。

1 **政府統計調査** 国や都道府県、市町村などの行政機関が、経済問題、環境問題など国民の生活にかかわる問題を明らかにするため、経済や環境の状況を数値でとらえるために行う統計調査を政府統計調査といいます。我が国でもっとも大規模に行われる政府統計調査は国勢調査で、我が国の人口の状況を明らかにするために行われています。政府統計調査の仕組みや種類などの概略は、総務省統計局の「統計制度」のページ（http://www.stat.go.jp/index/index.htm）で知ることができます。

2 **民間統計調査** ある目的のために民間団体や企業が行う統計調査を民間統計調査と呼びます。新製品の販売するための市場調査、販売後のアンケート調査、時刻天候などのよる物品の販売量変化の解析など。また新聞社マスコミなどが選挙予測や世論調査もおこなう。

3 **研究のため統計調査** 政府統計調査や民間統計調査が実態を調べるためにそれぞれの機関・組織が行う統計調査であるのに対して、研究統計調査は、研究に従事する個人やグループが行う統計調査です。研究者は研究上重要と思う諸問題に対し、それにかかわる現象を、科学的・客観的にとらえるため、研究の中で仮説を立て、その仮説を検証するために実験や観測によってデータを集めます。

以上は総理府統計局 http://www.stat.go.jp/ http://www.stat.go.jp/howto/case.htm 独立行政法人統計センター (政府統計の総合窓口) http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do より抜粋、編集。

(ii) 統計の手法

データはそれぞれ固有の性質から、量的属性と質的属性に分類できる。さらに量的なものは、離散型と連続型におおきく分けられる。前者は 2 項分布、後者は正規分布が代表的な分布であるが、日常に表れる分布はこのような名前のない一般形をしているものは非常に多い。質的属性は数値として得られるものではないから、わ

れわれが計算し易いよう適当に数量化する処理を施すことが多い。データを得るにはまず、統計調査をしなければならない。どのようにしてデータを集めるかという調査の設計を考え、全数調査か標本調査かを決める。標本調査となれば、標本抽出をする。対象とする集団を統計用語で母集団、標本の個数を標本の大きさとよぶ。もし大きな母集団から、効率よく標本抽出をするための一つの工夫として、地域や階級などの補助情報を用いる。

記述統計: 統計調査の結果得られたデータは、単に数値などの羅列にしか過ぎないが、この観測値の集りを要約した情報として整理をすることが記述統計とよばれるものである。具体的には、この集まりを 1 つの数値でまとめたり、いろいろな見易いグラフで表す工夫をしている。平均とか分散は典型的な位置やひろがり具合を表すための代表値である。また最近では探索的データ解析というデータのいろいろな記述を通して全体の構造を見出すことが行なわれている。

推測統計: 統計調査のデータを記述統計によって集約してみると、そこには新しい発見が見出せるかもしれない。しかし実際、未知の状態を理論として裏付し、認識するには、真の状態を推定したり、いくつかの考えられるものから真の状態を判断する検定がおこなわれる。このように観測値から推定や予測を行い、新しい状況に対して、推測する基礎としてもちいる理論的方法を推測統計とよぶ。母集団から抽出された標本にもとづいた母数（パラメータ）の推定、信頼区間や仮説の検定が具体的な方法である。最近では統計的決定理論として推測問題は発展している。大まかに統計学は記述統計と推測統計とに、上で述べた手法の違いによって分類される。

予測のための解析: 統計理論として 1 次元のデータ、すなわち 1 変量の理論が簡単であるが、実際の分野に統計を応用しようとすると、多変量の理論が多いし、現状は複雑であってどうしても必要となろう。一般に多変量解析とよぶ多くの変量間の相互関連を分析する手法がある。いく組かの標本を抽出して、ある特性に関して有意に異なっているか否かを判断するために、標本の分散を分けて分析する分散分析や、変量の間にモデルとする回帰式をたてて、ある目的とする変数の測定値の変動が、モデルによって説明する変数によって充分満足のいくものであるかどうか、現象の理解や目的変数の予測をすることが回帰分析であり、現代では、心理学、社会学、政治学、生物学、農学、医学、工学など多くの分野で用いられている。

以上のように統計学は、経験的に得られたバラツキのあるデータから、応用数学の手法を用いて数値上の性質や規則性あるいは不規則性を見いだす。統計的手法は、実験計画、データの要約や解釈を行う上での根拠を提供する学問であり、幅広い分野で応用されている。英語で統計または統計学を statistics と言うが、語源はラテン語で「状態」を意味する statisticum であり、この言葉がイタリア語で「国家」を意味するようになり、国家の人力、財力等といった国勢データを比較検討する学問を意味するようになった。現在では、経済学、自然科学、社会科学、医学 (疫学、EBM)、薬学など広い分野で必須の学問となっていることは論をまたない。

(iii) 統計データの使い方

統計データの使い方事例集にはどんなものがあるだろうか。

例 1. 若者と女性の就業状況
例 2. 少子化と将来の人口
例 3. 時間の使い方
例 4. 所得、消費支出、貯蓄《経済的な豊かさをとらえる》
例 5. グローバリゼーション
例 6. 消費の地域差《東西食べ物対決》

これらはすべて統計データとして公表されているデータベースから収集することができ、それをもとに統計の

解析を行うことができます。一番簡単にはインタネット上の公開リソースからが便利でしょう。

統計学に役立つインターネット上のリソース（資源）：

URL(Uniform Resource Locator) とは、インターネット上にある文書や画像などの場所を Web ブラウザーに指し示すための記述方式、サーバ名、フォルダ、ファイル名で構成されている。http(HyperText Transfer Protocol) とはデータの送受信のためのプロトコル（しきたり）。ただしこの URL らは不変ではありません。消えてしまう（削除）ものもありますので承知しておいてください。

◎統計データ・ポータルサイト 政府統計の総合窓口

http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do http://portal.stat.go.jp/ 各都道府県にもたくさんの公開データがあります

◎統計処理プログラム R (Win, Mac, Linux)（フリーウェア）R on Windows: http://plaza.umin.ac.jp/~takeshou/R/

◎情報検索のウェブページ 検索ソフト、グーグル http://www.google.co.jp/ フリー百科事典「ウィキペディア（Wikipedia）」http://ja.wikipedia.org/wiki/

◎ワープロ、表計算、データベースなどオープンオフィス（フリーウェア）http://www.openoffice.org/ http://ja.openoffice.org/

◎「授業案内と資料」のホームページ http://www.math.s.chiba-u.ac.jp/~yasuda/index-j.htm

第 1 回〜第 2 回：統計データの整理と要約第 3 回〜第 4 回：平均と標準偏差、相関と回帰第 5 回〜第 6 回：順列と組合せ、事象と確率第 7 回〜第 8 回：確率変数と確率分布、期待値第 9 回〜第 10 回：２項分布と正規分布第 11 回〜第 12 回：母集団と標本第 13 回〜第 14 回：統計的推測、母数の推定第 15 回：仮説検定

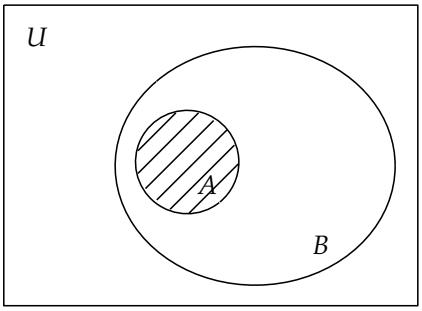
2 集合と確率

2.1 集合の考え

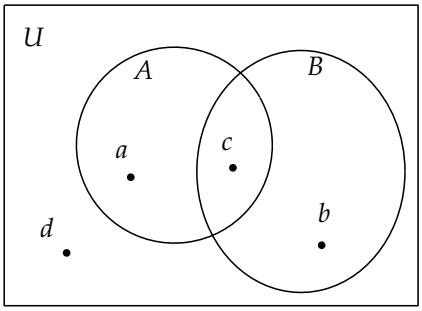
集合 (set) とはある事柄の集まりのうち、定義が具体的に示されているものを集合という。例えば、「自然数」は「$n > 0$ となる整数 n の全体」という定義があるので、集合といえるが、「大きな数」は、どこからが大きな数といえるのかがはっきりしないため、集合とはいえない。ただし、「大きな数」を例えば「1 億以上の整数」と定義すれば集合になりえる。

集合や要素の関係の表し方：点集合とは要素が実数などの数値で与えられるときが多い。要素がある集合に属するか属さないかの二律背反である。

たとえば要素 a は集合 A に属する、要素 b は集合 A に属さないという。a が集合 A の要素であるとする。このとき、a は集合 A に属するといい、記号で、$a \in A$ と表す。また、b が A の要素でないときは、$b \notin A$ と表す。

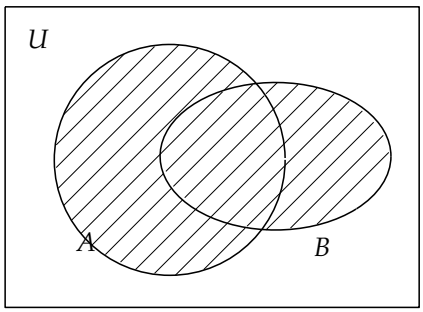


部分集合 $A \subset B$

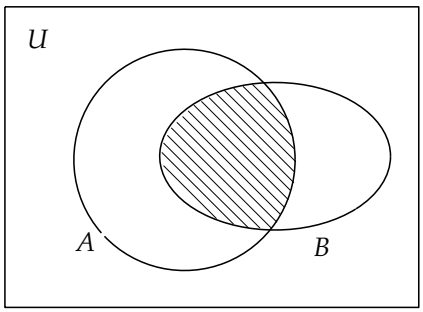


属する／属さない　$(a \in A, a \notin B)$

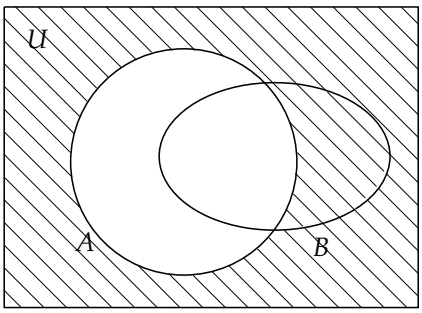
集合を表現するとき、その方法はいくつかあります。　(1) 要素を書き表す;　$\{2, 4, 6, \cdots, 20\}$　(2) 条件で書き表す; $\{x \mid x = 2n, n = 1, 2, \cdots, 10\}$, $\{2n \mid n = 1, 2, \cdots, 10\}$, $\{2n \mid n$ は 1 以上 10 以下の自然数 $\}$ などと表される。



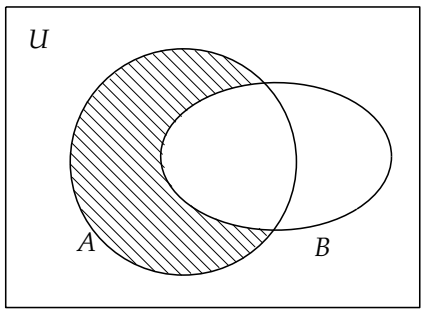
和集合 $A \cup B$



積集合 $A \cap B$



補集合 A^c または $\overline{A}$



差集合 $A \setminus B = A \cap \overline{B}$

集合と集合の関係

2 つの集合 A,B があり、$x \in A$ ならば $x \in B$ が成り立つとき、A は B の部分集合であるといい、「B は A を含む」か「A は B に含まれる」という。この状態を記号で $A \subset B$ や $B \supset A$ と表す。A の部分集合には A 自身もある（$A \subset A$）。また、A,B の集合の要素が同じとき、2 つの関係が同時に成り立つ場合、$A \subset B, B \subset A$ のとき、数値と同じ等号をもちいて $A = B$ と表す。空集合 (empty set) $\{\}$、つまり「要素がなにもない」というのもひとつの集合として考えられる。 これを空集合といい、ギリシャ文字のファイ（ϕ）あるいは空集合の記号 ($\emptyset$) で表される。空集合は全ての集合に含まれる。（例：$\phi \subset A$）

2 つの集合 A, B があるとき それらの両方ともに満たす集合を A と B の**積集合** (intersection) とよび、$A \cap B$ と書く。積集合は 2 つの集合の要素の共通部分である。また、集合 A, B どちらかの条件を満たす集合を A と B の**和集合** (union) とよび、$A \cup B$ と書く。和集合は 2 つの集合の合併を表す。

集合について考えるときは考察の対象とする全ての要素を含む全体を U あるいはギリシャ文字のオメガ Ω とおきます。このときの U を**全体集合** (Universe set) という。また、全体集合 U の中で集合 A に属さないものを U に関する A の**補集合** (complement) という。これを記号で、A^c あるいは $\overline{A}$ と表す。補うということでもととの和集合は全体集合になります。全体集合は全ての要素を含んでいるので、全体集合は空集合ではなく、また、全体集合の補集合を**空集合** (empty set) といいます。

定理 2.1 *(ド・モルガンの法則)* 和集合と積集合、さらに補集合を関係づける式*:*

$$(1)\ \overline{A \cup B} = \overline{A} \cap \overline{B} \qquad (2)\ \overline{A \cap B} = \overline{A} \cup \overline{B}$$

2 つの集合間の関係（2 項演算）はさらにより多くの関係式を導くことができる。数値の演算、加法や乗法と同じように発展させることができます。

定理 2.2 2 つの演算について、結合律、交換律、また配分律という次の性質が成り立つ。

(1) $(A \cap B) \cap C = A \cap (B \cap C)$　結合律　　*(2)* $(A \cup B) \cup C = A \cup (B \cup C)$　結合律

(3) $A \cup B = B \cup A$　交換律　　*(4)* $A \cap B = B \cap A$　交換律

(5) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$　配分律

これらの性質から、次の定義ができる。

定義 2.1 n 個の集合,$A_1, A_2, \cdots, A_n$ について

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$$
$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n$$

問 2.1 一般の n 個の集合 $\{A_i, i = 1, 2, \cdots, n\}$ についてドモルガンの法則を述べよ。

問 2.2 変数 $a, b, c \in \{0, 1\}$ を 集合 A, B, C に対応させると、次の対応が成り立つことを確かめよ。

(i) $1 - a$	補集合 $\overline{A}$;	*(ii)* $a = 1 - (1 - a)$	2 回の補集合 $\overline{(\overline{A})}$;
(iii) 最大値 $\max\{a, b\}$	和集合 $A \cup B$;	*(iv)* 最大値 $\max\{a, b, c\}$	和集合 $A \cup B \cup C$;
(v) 最小値 $\min\{a, b\}$	積集合 $A \cap B$;	*(vi)* 最小値 $\min\{a, b, c\}$	積集合 $A \cap B \cap C$.

問 2.3 要素と集合の関係が " $x \in \bigcup_i A_i$ " とは " 少なくとも一つの番号 i があって、$x \in A_i$ が成り立つ " こと、また " $x \in \bigcap_i A_i$ " とは " すべての番号 i で $x \in A_i$ が成り立つこと " を確かめよ。

[補足]「少なくとも一つ」とは " 適当に番号を選んで " とか、" 番号が存在して " という意味なるから、存在 (exist) の頭文字を逆さまにして記号 "∃"をもちい、また「すべての番号」とは、全部、どんな番号でもということから、すべて (all) の頭文字から記号 "∀"を用いられる。略記号で

$$x \in \bigcup_i A_i \quad \Leftrightarrow \quad x \in A_i, \exists i$$
$$x \in \bigcap_i A_i \quad \Leftrightarrow \quad x \in A_i, \forall i$$

と書く。

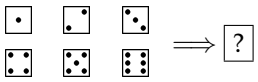
問 2.4 n 個の変数を $x_i \in \{1, 0\}, i = 1, 2, \cdots, n$ とするとき、次を満たすことを示せ。

(i) $\max\{x_1, x_2, \cdots, x_n\} = 1 - (1 - x_1) \times \cdots \times (1 - x_n)$
(ii) $\min\{x_1, x_2, \cdots, x_n\} = x_1 \times x_2 \times \cdots \times x_n$

この関係式とドモルガンの法則とどんな関係があるか考えよ。

2.2 場合の数と確率

サイコロの目数やトランプの手札の並べ方など同じ様な事でありながら、少しずつ違った種類のやり方があり、それらのやり方が全部で何通りあるか考える。このような場合、実際に数えられた数のことを、場合の数という。ある事柄についてそのことが起こりうる場合の数を間違いなく数えることは、 その事柄についてどのことが おこりやすくどのことが起こりづらいかを見分けるための基礎となる。 例えば、ポーカーなどのゲームでは集めることが難しい役は高いランクが与えられているが、これは起こりにくい役が出来るトランプの組み合わせの 現われる確率が小さいことによる。このことは、52 枚のカードから 5 枚を 引いて来たときに全てのカードを引く確率が同じであるとしたとき、ある役に対応する カードの組み合わせを引く場合の数がより少ないことに対応する。 このように、場合の数は事柄が起こりうる確率と密接な関係にある。いいかえれば、すべての起こり得る総数との比率で、事象の確率を定める。全体の総数が膨大で、集めようとする手札場合の数が少なければ、手札を得る確率、事象の起こる確率は小さい。



カードゲームのように確率が具体的に計算できる場合以外にも、 確率の考え方を用いて計算される事柄は多くある。例えば、保険とは ある事柄に値段、価値を売買することであるが、保険を下ろさなくてはならない事柄が起こりにくいと客観的に思われるものほど、そのものの値段が下がるという特徴がある。自動車保険に加入するのに必要な代金は若者では高く、年令を重ねるごとに 低くなっていく。これは、若者は自動車の免許を取得して時間が短い場合が 多く、保険金の支払を必要とする自動車事故をおこす可能性が高いことによる。 一方、年令を重ねたものについては運転の技量が時とともに上達すると 一般に考えられるので保険をかけるための代金は少なくなるのである。 また、同じ若者でも既に何度か事故を重ねたものは同じ年代の他の若者よりも 保険料が高くなる傾向がある。これは、何度か事故を重ねたものは運転の仕方に 何らかの問題がある傾向があり、それによってふたたび事故をおこす可能性が通常のものと比べてより高いと考えられることによる。

金融業界でも変動等のリスクを考慮しながら、確率の考えを用いて高い利益を出すよう実践している。投資から得られる利益とリスクをもとに、リスクがあるが、投資利益の可能性が高い相手に対しては応じた資金を投資し、より安定した利益が得られる場合には、利得は低いがリスクを回避する意味での投資を実行し、手持ちの資金から生じる利益を多くするよう考える。リスクを評価すること、どのような変動の可能性をもつかを考察するためには確率の概念が重要である。

ここでは場合の数と確率の計算法を紹介する。まず先に様々な事柄の場合の数の計算法を扱い、その結果を用いてある事柄が起こる確率を計算する。

2.3 順列・組合せ

場合の数の計算方法として、 n 個の異なったものを並べ換える仕方の数を数える。まず最初に並べるものは n 個、次に並べるものは $(n-1)$ 個、その次に並べるものは $(n-2)$ 個 ... とだんだんと選べるものの数が減って行き、最後には 1 個 しか残らなくなることに注目すると、 この事柄に関する場合の数は $n\times(n-1)\times\cdots\times 2\times 1$ となることが分る。 "エクスラメーションマーク"(exclamation, 感嘆詞記号、ビックリマーク)! をもちいて $n!$ を n の階乗という。階乗（かいじょう、ファクトリアル）の記号とは

$$n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 \ \ (n \geqq 1), \quad 0! = 1$$

とする。10 までの階乗を計算してみると、$3! = 6$, $4! = 24$, $5! = 120$, $6! = 720$, $7! = 5040$, $8! = 40320$, $9! = 362880$, $10! = 3628800$ となっていく。驚くほど速く大きい数になっていくことが分かる。１０の階乗は約３６０万にもなっている。

「r 種類の箱に n 個のボールの中から選んで入れること」が場合の数を数え上げるための基本計算です。

1. 順列: 「区別のある箱」に「区別のあるボール」を入れる
   重複順列「重複を許してひとつの箱に何個入れてもよい場合」という条件で取り出された場合の数。

$$n^r = \overbrace{n \times n \times \cdots n}^{r} \tag{2.1}$$

   順列 (permutation)「重複は許さずにひとつの箱には高々ひとつしか入れない場合」という条件で取り出された場合の数。

$$(n)_r = \overbrace{n(n-1)(n-2)\cdots(n-(r-1))}^{r} = \frac{n!}{(n-r)!} \tag{2.2}$$

2. 組合せ: 「区別のある箱」に「区別のないボール」を入れる
   重複組合せ「重複を許してひとつの箱に何個入れてもよい場合」という条件。

$$\frac{[n]^r}{r!} = \frac{\overbrace{n(n+1)(n+2)\cdots(n+(r-1))}^{r}}{r!} = \binom{n+r-1}{r} = \binom{n+r-1}{n-1} \tag{2.3}$$

   組合せ (combination) 「重複は許さずにひとつの箱には高々ひとつしか入れない場合」という条件。

$$\frac{(n)_r}{r!} = \frac{n(n-1)(n-2)\cdots(n-(r-1))}{r!} = \binom{n}{r} = \binom{n}{n-r} \tag{2.4}$$

階乗に関連した記号を定義しておく。これらはほぼ一般的であるが、すべてに共通して用いられるとは限らないことを注意する。

$$\text{減少階乗積: } (n)_r = \overbrace{n(n-1)(n-2)\cdots(n-r+1)}^{r}$$
$$\text{増加階乗積: } [n]^r = \overbrace{n(n+1)(n+2)\cdots(n+r-1)}^{r}$$

結果の並び順を表す記号：

$$\text{順序のある（区別する）場合を } (a_1, a_2, a_3, \cdots, a_r),$$
$$\text{順序ない（区別しない）場合を } \{a_1, a_2, a_3, \cdots, a_r\}$$

と表すときがあります。括弧の違いに注意します。並び替えは $r!$ 通りありますから、両者の場合の数を比較すると

$$\{ \text{ 区別する場合の数 } \} = r! \times \{ \text{ 区別しない場合の数 } \}$$

です。括弧の違いはたとえば (a, b, c) はベクトルの座標、$\{a, b, c\}$ は集合と関連して覚えてもいいでしょう。

`http://ja.wikibooks.org/wiki/`**高等学校数学 A_場合の数と確率** にはいくつかの簡単な例題が述べられていて練習になる。

問 2.5 それぞれに 1 から 5 までの数字が書かれた 5 枚のカードがある。このカードを順に横一列に並べ換え、5 けたの数をつくる。つぎの計算をせよ。

(i) 作られた数の最大値、最小値　(ii) あらゆる並べ方の総数　(iii) 偶数となる場合の数
(iv) 奇数となる場合の数　(v) 5 の倍数となる場合の数

解答 (i) カードの数が 5 枚でそれぞれが区別できることから、カードの並べ方の数は $5! = 5*4*3*2*1$ となり、120 となる。(ii) 偶数を得るためには一の位である最も右に出るカードが、偶数となればよい。 このようなカードは 2 と 4 であり、それぞれに対して後の 4 枚は自由に選んでよい。 このため、このようなカードの並べ方は、$2*4! = 48$ となる。(iii) 奇数を得るためには一の位である最も右に出るカードが、奇数となればよい。 このようなカードは $1,3,5$ であり、それぞれに対して後の 4 枚は自由に選んでよい。 このため、このようなカードの並べ方は、$3*4! = 72$ となる。一方、5 枚のカードを並べ換えて得られる数は必ず偶数か奇数のどちらかであるので、(i) の結果から (ii) の結果を引くことによっても (iii) の結果は得られるはずだが、実際にそれを計算すると $120 - 48 = 72$ となり、確かにそのようになっている。

問 2.6 数字 $\{0,1,2,3,5\}$ が書かれた 5 枚のカード (4 はない) がある。これを並び換えて数を作る。 初めに 0 が並べられるときは数の桁数には入れないこととする。つぎの場合を計算せよ。

(i) 5 桁の数が得られる数　(ii) 5 桁の偶数が得られる数　(iii) 5 桁の奇数が得られる数
(iv) 5 桁の 5 の倍数が得られる数　(v) 5 桁の 10 の倍数が得られる数

をそれぞれ求めよ。

解答 (i) 先頭が 0 になったときには 5 桁の数にならないことに注意すればよい。求める場合の数は $4*4! = 96$ となる。(ii) 最初が 0 でなく最後が 0 か 2 である数を数えればよい。まず、最後が 0 であるときには、残りの 4 枚は任意であるので $4! = 24$ 通りの組み合わせがある。 次に、最後が 2 であるときには最初は 0 であってはいけないので、$3*3! = 18$ 通りある。 2 つを合わせた数が 5 桁の偶数が得られる場合の数である。答えは、$24 + 18 = 42$ となる。(iii) (i) の結果から (ii) の結果を引けばよいが、ここではその結果が正しいかどうか 確かめるためにも 5 桁の奇数が得られる組み合わせを数え上げてみる。5 桁の奇数を得るためには最後の数は $1,3,5$ のいずれかでなくてはならない。このうちのどの場合についても 5 桁の数を得るためには最初の数が 0 であってはならないのでそれぞれの場合の数は、$3*3*3! = 54$ となりこれが 5 桁の奇数を得る場合の数である。 (ii) の結果と足し合わせると確かに (i) の結果と等しい 96 を得る。(iv) 5 の倍数を得るためには最後の数が 0 か 5 であればよい。このとき最後が 0 になる場合の数は他の 4 つが任意であるため $4! = 24$ 存在する。次に、最後が 5 になる場合の数は最初の数が 0 であってはならないため $3*3! = 18$ だけ存在する。よって答えは $24 + 18 = 42$ となる。

問 2.7 5 個のボールが入ったつぼから *2* つのボールを取りだすとき *(*ボールはそれぞれ区別できるものとする*)*、*2* つのボールの選び方は、 何通りあるか計算せよ。

解答 ボールの取りだし方は組み合わせの数を用いて計算できる。5 つのボールの中から 2 つを取りだすのであるからその場合の数は、$\binom{5}{2} = \dfrac{5\cdot 4}{2!} = 10$ となる。よって、ボールの取りだし方は 10 通りであることがわかる。あるいは取り出して残った結果とは、 1 対 1 の対応があるから、同じ場合の数となる。残りを数え上げると $\binom{5}{3} = \dfrac{5\cdot 4\cdot 3}{3!} = 10$ で同じ $\binom{5}{2} = \binom{5}{3}$ となる。

問 2.8 6 個の互いに区別できるボールが入った箱がある。 この中から *(i) 3* つのボールと *2* つのボールを取りだす方法の場合の数 *(ii) 2* つのボールを取り出すことを *2* 回くり返し、それぞれを別の互いに区別できる 袋

にいれる場合の数 *(iii)* 2 つのボールを取り出すことを 2 回くり返し、それぞれを別の互いに区別できない 袋にいれる場合の数をそれぞれ計算せよ。

解答 (i) 最初にボールを取りだすときには、6 つのボールの中から 3 つのボールを 取りだすことからその場合の数は $\binom{6}{3}$ だけある。また、次にそれを取り除いた中から 2 つのボールを取り除くときには その取りだし方は、$\binom{3}{2}$ だけある。 よって、このときの場合の数は $\binom{6}{3} * \binom{3}{2}$ だけになる。実際この値を計算すると、$20 * 3$ となり、60 通りであることが分かる。(ii) (i) の場合と同様に 6 つのボールの中から 2 つのボールを取りだすことからその場合の数は $\binom{6}{2}$ だけある。また、次にそれを取り除いた中から 2 つのボールを取り除くときにはその取りだし方は、$\binom{4}{2}$ だけある。 よって、このときの場合の数は $\binom{6}{2}\binom{4}{2}$ だけになる。実際この値を計算すると、$15 * 6$ となり、90 通りであることが分かる。(iii) (ii) と同じ計算で値を求めることが出来るが、今はボールをいれた袋が 互いに区別できないことに注意しなくてはならない。 このことによって、起こりうる場合の数は (ii) の場合の半分になるので求める場合の数は 45 通りとなる。

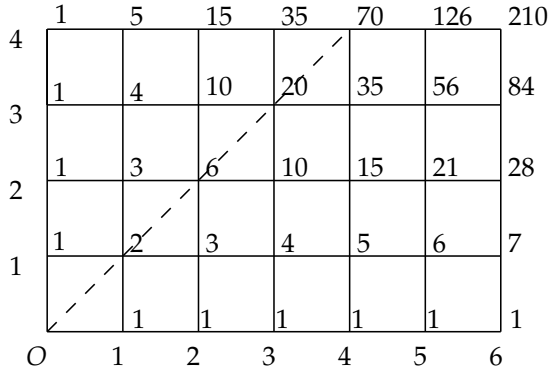
2 項係数のまとめ：整数 n, r に対して $\binom{n}{r}$ はつぎの式が成り立つ。

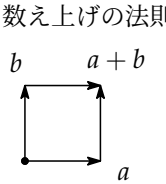
対称性： $\binom{n}{r} = \binom{n}{n-r}$

加法性： $\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}$

2 つ目の式は、"n 個のものから r 個を選ぶ仕方の数は、最初の 1 つを選ばずに他の $n-1$ 個から r 個を選ぶ仕方の数と、最初の 1 つを選んで、他の $n-1$ 個から $r-1$ 個を選ぶ仕方の数の和である"ということを表わしている。さらにここでは $n, r \geq 1$ のみならずボールの個数だけではなく、定義の式では、r は非負整数 ($0! = 1$) として、n はより一般の実数 $\binom{\pi}{3} = \frac{(\pi)_3}{3!} = \frac{\pi * (\pi-1) * (\pi-2)}{6}$ や $\binom{3}{5} = \frac{(3)_5}{5!} = \frac{3 * (3-1) * (3-2) * (3-3) * (3-4)}{5!} = 0$ と計算される。

応用：経路の数え上げ計算　　原点 $O = (0,0)$ から出発して、上 (North) か右 (East) に進む。それらの経路について格子点での経路数を数え上げる。





整数 $m, n = 0, 1, 2, \cdots$ での格子点 (m, n) までの経路数は $N(m,n) = \binom{m+n}{n} = \binom{m+n}{m}$ で与えられる。たとえば、$N(0,3) = 1$, $N(2,3) = N(3,2) = 10$, $N(3,3) = 20$ などとなっている。逆に $N(m,n) = k$ となるよう、k の値から (m,n) を求める。$k = 15$ のときは、$(m,n) = (2,4), (4,2)$ の 2 ヶ所となり、$k = 252$

のときは、$(m,n) = (5,5)$ の 1 ヶ所である。関係式 $N(m,n) + N(m-1,n+1) = N(m,n+1)$ が成り立つ。すなわちこれは 2 項係数の関係式；$\binom{m+n}{n} + \binom{m+n}{n+1} = \binom{m+n+1}{n+1}$ を表している。証明は多項式の展開で $(x+1)^{m+n} \times (x+1) = (x+1)^{m+n+1}$ の x^n 次の係数比較をすればよい。

問 2.9 格子点 (m,n) までの経路数は $N(m,n) = \binom{m+n}{n} = \binom{m+n}{m}$ で与えられることを示せ。

二項定理: 組合わせの計算を多項式の展開に応用することが出来る。変数 a,b の n 次多項式;$(a+b)^n$ を展開する。これは、

$$(a+b)^n = \overbrace{(a+b) \times (a+b) \times \cdots \times (a+b)}^{n}$$

という式を展開したものであり、その係数は "n 個の $(a+b)$ の中からいくつの a または b を選ぶか" で決めることが出来る。 かけ算なので a か b のどちらかを必ず選ばなくてはならない。 この仕方の数は、a を r 個選んだら、b は $n-r$ 個選んでいる。a について r 次の項では、つまり、$a^r b^{b-r}$ の係数は $\binom{n}{r} = \binom{n}{n-r} = \frac{(n)_r}{r!} = \frac{(n)_{n-r}}{(n-r)!}$ に等しい。 よって、次の式が得られる。

$$(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \cdots + \binom{n}{r}a^{n-r}b^r + \cdots + \binom{n}{n-1}a^1 b^{n-1} + \binom{n}{n}a^0 b^n$$

2.4 確率の基本法則

与えられた集合の要素を数え上げることで始めに確率を定める。要素の個数が有限個の集合のことを有限集合 という。要素の数え上げるとき、集合の全体要素数が有限個の場合であれば、集合全体とその部分集合の比率で定まる数値を、その部分集合に対する確率と定める。部分集合を条件を満たすような場合の数と考え、事象とよぶ。このような有限個の集合では、全体集合に対して、部分集合を事象 (event) とよぶ。一般の場合に確率が定まる事象を定義しなければならないが、有限集合のときには、すべての部分集合について、確率が定まるから、事象となる。ある場合の数が、実際に条件を満たす部分集合であれば、比率（割合）のことを確率 (probability) とよぶ。

定義 2.2 同じ確からしさのもつとき、事象の確率定義: 起こりうるすべての場合の数を N、事象 A に含まれる要素数 $\#(A)$ と表すとき、それぞれの要素がすべて同じ程度に確からしいとき、事象 A の起こる確率 $P(A)$ を

$$P(A) = \frac{\#(A)}{N}$$

とくに要素 $a \in A$ では, 1 個の要素からなる集合を $\{a\}$ と表し、$P(\{a\}) = \frac{1}{N}, \forall a$ という関係式が得られる。同様に n 個ならば

$$P(\{a_1, a_2, \cdots, a_n\}) = \frac{n}{N}, \quad \forall a_i \in A$$

確率の計算: ある場合の数が実際に現われる割合はその場合の数を、その事柄に おいて起こり得る全ての事柄の場合の数で割ったものに等しい。 例えば、全く等しい割合で全ての面が出るさいころをふったとき、1 が出る確率はである。 これは 1 が出る場合の数 1 を、1,2,3,4,5,6 のいずれかが出る場合の数 6 で割ったものに等しい。

確率の性質: 確率の定義から、次の性質が得られる。全体集合を Ω(オメガ) とし、その要素数が有限個 $\#(\Omega) < \infty$ とするとき、

(1) どんな事象 A についても、$0 \leq P(A) \leq 1$

(2) 決して起こらない事象 (空事象) $\emptyset$ または ϕ の確率は $P(\phi) = 0$,

(3) 必ず起こる事象（全体事象、全事象）Ω の確率は $P(\Omega) = 1$.

和事象、積事象の確率: 2 つの事象 A,B が同時に起こらないとき、$A \cap B = \phi$ のとき、事象 A と B は互いに排反 (mutually exclusive) である、または互いに素 (mutually disjoint) という。A と B は**排反事象** (exclusive event) であるという。A と B が排反事象ならば、A または B が起こる確率はそれぞれの和に等しい。すなわち、

$$A \cap B = \phi \quad \text{ならば、} \quad P(A \cup B) = P(A) + P(B)$$

同様にもし互いに素であるならば、積事象の確率は

$$A \cap B = \phi \quad \text{ならば、} \quad P(A \cap B) = 0$$

である。和事象と積事象の和はそれぞれの確率の和に等しい。一般の場合には

$$P(A \cup B) + P(A \cap B) = P(A) + P(B) \tag{2.5}$$

積（同時）事象の記号は積に対応するから $a \times b \times c = abc$ などと同じように省略することが多い。数列 $\{a_i\}$ の和 $\sum_i a_i$, 積 $\Pi_i a_i$ と同様な使い方をする。可算の列 $A_1, A_2, \cdots, A_n$ では、$\bigcup_{i=1}^{n} A_i = \bigcup_i A_i = A_1 \cup A_2 \cup A_3 \cup \cdots A_n$ $\bigcap_{i=1}^{n} A_i = \bigcup_i A_i = A_1 \cap A_2 \cap A_3 \cap \cdots A_n = A_1 A_2 A_3 \cdots A_n$

問 2.10 赤玉 *2* 個と白玉 *3* 個が入った袋から、玉を *2* 個同時に取り出す。このとき、*2* 個とも白玉が出る確率を求めよ。

解答 赤白あわせて 5 個の玉から 2 個を取り出す方法は $\binom{5}{2} = 10$（通り）このうち、2 個とも白玉になる場合は $\binom{3}{2} = 3$（通り）よって求める確率は $\binom{3}{2}/\binom{5}{2} = 3/10$

問 2.11 男子 7 人、女子 5 人の中から、くじ引きで 3 人の委員を選ぶとき、3 人とも同性である確率を求めよ。

解答 男子女子の合計 12 人の中から 3 人の委員を選ぶ場合の数は $\binom{12}{3} = 220$（通り）ここで、「3 人とも男子である」事象を A、「3 人とも女子である」事象を B とすると、$P(A) = \binom{7}{3}/\binom{12}{3} = 35/220$, $P(B) = \binom{5}{3}/\binom{12}{3} = 10/220$「3 人とも同性である」事象は、和事象 $A \cup B$ であり、しかも、A と B は排反事象, $A \cap B = \emptyset$ である。よって求める確率は $P(A \cup B) = P(A) + P(B) = 35/220 + 10/220 = 45/220 = 9/44$.

余 (補) 事象の確率: 事象 A に対して、「A でない」「A が起こらない」事象を $\overline{A}$ で表し、A の**余事象（補事象）**(complement) という。記号 A^c も用いられる。A の余事象を $\overline{A}$ とすると

$$P(\overline{A}) = 1 - P(A) \tag{2.6}$$

問 2.12 赤玉 *5* 個、白玉 *3* 個の計 *8* 個入っている袋から *3* 個の玉を取り出すとき、少なくとも *1* 個は白玉である確率を求めよ。

解答 8 個の玉から 3 個の玉を取り出す場合の数は $\binom{8}{3} = 56$（通り）。いま、「少なくとも 1 個は白玉である」事象を A とすると、補事象 $P(\overline{A})$ は「3 個とも赤玉である」という事象だから $P(\overline{A}) = \binom{5}{3}/\binom{8}{3} = 10/56$ よって求める確率は $P(A) = 1 - P(\overline{A}) = 1 - \binom{5}{3}/\binom{8}{3} = 23/28$

問 2.13 3つの事象 A, B, C があるとき、

(1) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - \{P(A\,B)$
(2) $P(A \cup B) = 1 - P(\overline{A \cup B}) = 1 - P(\overline{A} \cap \overline{B}) = 1 - P(\overline{A}\,\overline{B})$
(3) $\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - \{P(A \cap B) + P(B \cap C) + P(C \cap A)\} + P(A \cap B \cap C) \\ &= P(A) + P(B) + P(C) - \{P(A\,B) + P(B\,C) + P(C\,A)\} + P(A\,B\,C) \end{aligned}$
(4) $P(A \cup B \cup C) = 1 - P(\overline{A \cup B \cup C}) = 1 - P(\overline{A} \cap \overline{B} \cap \overline{C}) = 1 - P(\overline{A}\,\overline{B}\,\overline{C})$

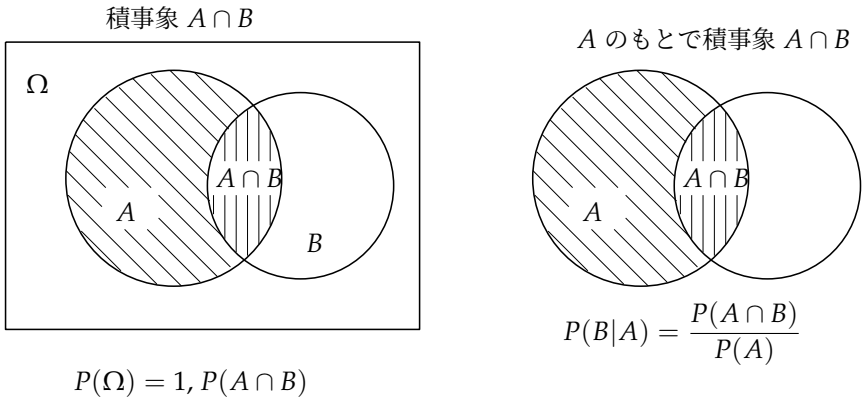
が成り立つ。これを示せ。

2.5 独立な試行と確率計算

繰り返しの試行は独立な試行になりますが、独立の定義をします。そのためには一般的に起こった事象に影響されて、次の結果がおこるという条件付き確率から定義します。

条件付き確率 (conditional probability); 事象 A が起こったという条件のもとで事象 B の起こる確率を，A のもとでの B の **条件付き確率** (conditional probability) といい，$P(B|A)$ で表す。ただし，$P(A) \neq 0$ とする。

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{2.7}$$



条件付き確率は、ベースとなる考えるの中で制限をした比率、すなわち確率であるから、条件付き確率もやはり確率の性質をもつ。また式を変形して

$$P(A \cap B) = P(A)\ P(B \mid A) \tag{2.8}$$

となるから、条件付き確率 $P(B \mid A)$ のほうが計算できるならば、$P(A)$ をかけることで、同時（積）事象の確率が求める方法がよく用いられる。

2 つの事象 A，B があって，A が起こった場合と，起こらなかった場合とで B の起こる条件付き確率が等しいとき，事象 B は事象 A と **独立** (independent) であるという。数式で書けば，

$$P(B \mid A) = P(B \mid \overline{A}) \qquad \Leftrightarrow \qquad P(A \cap B) = P(A)\,P(B)$$

より、事象 A が起こる起こらないに関わらず、条件付き確率が一定の値であることを示している。あるいは $P(B \mid A) = P(B)$ としても同じ命題で、事象 A の起こる条件付きが B のおこる絶対的な確率に等しいことを

求めている。2 個の事象に対して、独立とは

$$P(A \cap B) = P(A)\,P(B) \tag{2.9}$$

と定める。一般の n 個の対しては、少し条件を増やしたほうが取り扱いやすくなる。事象 $\{A_1, A_2, \cdots, A_n\}$ が独立であるとは、

条件 (i)　すべての 2 個ずつの組で, A_i と A_j が独立で

条件 (ii)　$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n)$

が成り立つものとする。条件 (i) のみを**組み毎に独立** (pairwise independent) とよぶ場合がある。

たとえば、1 から 15 までの番号札があり、その 15 枚の札から任意に 1 枚を選ぶ。 このとき、2 の倍数を選ぶという事象を A、3 の倍数を選ぶという事象を B とすると、$A = \{2,4,6,8,10,12,14\}, B = \{3,6,9,12,15\}$ となる。このとき、選び出された札が 2 の倍数であるとわかったとして、それが 3 の倍数である確率 p を考える。 p は、2 の倍数である札 7 枚の中から、6 の倍数である札 2 枚を選ぶ確率であるから $A \cap B = \{6,12\}$. 事象 A が起こったとして、そのときに事象 B の起こる確率を、A が起こったときの B の条件つき確率といい、$P(B|A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{2}{7}$ となる。

事象の集まり $\{B_1, B_2, \cdots, B_n\}$ が全事象 Ω の**分割** (partition) とは,

(i) $B_i \cap B_j = \phi \quad (i \neq j)$,

(ii) $\Omega = \bigcup_i B_i$

のとき、つまり互いに素で少なくとも一つは必ず起こるときをいう。このときには、条件付き確率を分解して表現することができる。いわゆる場合分けをすれば、確率計算が容易になることが多い。

定理 2.3「場合分けの確率」もし $\{B_1, B_2, \cdots, B_n\}$ が全事象 Ω の分割であれば、

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n) \tag{2.10}$$

が成り立つ。

独立な試行と確率: 1 個のさいころを投げるとき、偶数の目が出る事象を A、3 の倍数の目が出る事象を B、4 以上の目が出る事象を C とすると、$A = \{2,4,6\}$, $B = \{3,6\}$, $C = \{4,5,6\}$. このとき , 条件付き確率の定義より、$P(B|A) = P(B)$ が成り立つ。つまり、事象 A が起こることは事象 B が起こることに影響を与えていない。また、$P(C|A) = P(C)$ が成り立つ。つまり、事象 A が起こることは事象 C が起こることに影響を与えている。

2 つの事象 A , B について、事象 A の起こることが事象 B の起こることに影響を与えないとき、A と B は独立であるという。事象 A と B が独立であるとき、$P(B|A) = P(B)$ である。

互いに他の結果に対して影響をおよぼさない操作を繰りかえすとき それぞれの試行は独立であると言う。独立な試行については、ある試行の起こる確率が定められていて、 それを n 回繰りかえしたとき、それらが起こる確率は、 それぞれの試行が起こる確率の積となる。

問 2.14 赤玉 *3* 個、白玉 *2* 個の計 *5* 個入っている袋がある。この中から *1* 個の玉を取り出して色を確かめてから袋に戻し、再び *1* 個を取り出すとき、*1* 回目は赤玉、*2* 回目は白玉を取り出す確率を求めよ。

解答　1 回目に取り出した玉を袋に戻すので、「1 回目に取り出す」試行と「2 回目に取り出す」試行とは互いに独立である。1 回目に取り出した 1 個が赤玉である確率は $\dfrac{3}{5}$, 2 回目に取り出した 1 個が白玉である確率は $\dfrac{2}{5}$. したがって求める確率は $\dfrac{3}{5} \times \dfrac{2}{5} = \dfrac{6}{25}$

反復試行の確率: 同じ試行を何回か繰り返して行うとき、各回の試行は独立である。この一連の独立な試行をまとめて考えるとき、それを反復試行という。反復試行の確率を計算する。ある試行で、事象 E の起こる確率が p であるとする。この試行を n 回繰り返すとき、事象 E がそのうち r 回だけ起こる確率は $\binom{n}{r}p^r(1-p)^{n-r}, (r=0,1,2,\cdots,n)$ となる。この生起回数は 2 項分布 $Binom(n,p)$ にしたがうという。

問 2.15 *1* 個のさいころを *5* 回投げるとき、*3* の倍数の目が *4* 回出る確率を求めよ。

解答 1 個のさいころを 1 回投げるとき、3 の倍数の目が出る確率は $\dfrac{2}{6}=\dfrac{1}{3}$ である。よって、1 個のさいころを 5 回投げるとき、3 の倍数の目が 4 回出る確率は $\binom{5}{4}\left(\dfrac{1}{3}\right)^4\left(1-\dfrac{1}{3}\right)^{5-4}=\dfrac{10}{243}$

問 2.16 ある観光バスの乗客のうち、*60%* が女性で、*42%* が *50* 歳以上の女性である。女性の中から任意に *1* 人を選び出したとき、その人が *50* 歳以上である確率を求めよ。

解答 「女性である」事象を A、「50 歳以上である」事象を B とする。$P(A)=\dfrac{60}{100}=0.6$, $P(A\cap B)=\dfrac{42}{100}=0.42$. よって、求める確率は $P(B\mid A)=\dfrac{P(A\cap B)}{P(A)}=0.7$.
乗法定理
のとき

問 2.17 *5* 本のくじの中に *3* 本の当たりくじがある。*a*、*b* の *2* 人が、引いたくじをもとに戻さないで、*a*、*b* の順に *1* 本ずつくじを引くとき、*2* 人とも当たる確率を求めよ。

解答 a が当たるという事象を A、b が当たるという事象を B とすると、求める確率はである。a が当たったときには、残りは 4 本で、この残りのくじの中に当たりくじが 2 本あるから、$P(B\mid A)=\dfrac{2}{4}=\dfrac{1}{2}$. よって、2 人とも当たる確率は $P(A\cap B)=P(A)\,P(B\mid A)=\dfrac{3}{5}\times\dfrac{1}{2}=\dfrac{3}{10}$.

問 2.18 トランプのハートのカードが *1* 組 13 枚ある。*(1)* 初めに *a* が *1* 枚引き、そのカードをもとに戻さないで、次に *b* が *1* 枚引く場合、*a*、*b* がともに絵札を引く確率を求めよ。*(2)* 初めに *a* が *1* 枚引き、そのカードをもとに戻して、次に *b* が *1* 枚引く場合、*a*、*b* がともに絵札を引く確率を求めよ。

解答 a が絵札を引くという事象を A、b が絵札を引くという事象を B とする。(1) A と B がともに絵札を引くという事象は で表される。　A が絵札を引く確率は　A が絵札を引いたあと、12 枚のカードの中に絵札が 2 枚残っているから、B が絵札を引く確率 PA(B) は、　よって　(2) A が引いたカードは、もとに戻すから、2 つの事象 A、B は互いに独立である。　したがって確率は

2.6 ベイズの定理

ベイズの定理は上で述べた条件付き確率の応用です。

定理 2.4 事象 A と全事象 Ω の分割 $\{B_1,B_2,\cdots,B_n\}$ があるとき、$P(B_i)$,$P(A\mid B_i)$, $(i=1,2,\cdots,n)$ から、条件順序が反転している確率を計算できる：

$$P(B_i\mid A)=\frac{P(A\mid B_i)\,P(B_i)}{\sum_j P(A\mid B_j)P(B_j)}$$

ベイズの定理はこの形のみにとどまらず、広く統計手法をもちいた認識論と考え、これを特に従来の公理的に確率を与えるのではなく、不確実性の認識をめざします。ベイズ統計学とは、数学的な言語である「確率」を用いて、認識論的な不確実性を記述するためのシステムといえます。ベイズ理論の枠組みでは、自然の状態における信憑性 (belief) の程度を特定化しようとしまう。つまり非負の値で、自然の状態すべての信憑度が 1 となるよう基準にします。ベイズ統計学における方法では、「事前」（あらかじめ）の信憑度が存在するとして、それから実際のデータを観測することで「事後」（事がらの起こった後、あるいは観測データを得てから）の信憑度を改訂更新しながら、推測決定のための基盤を与えていくものと考えます。

1763 年、Thomas Bayes(1702-1761) は、帰納的問題（特殊から一般へ）に関する論文を発表しました。現在の表現を用いれば、2 項データ (たとえば、成功と失敗) において、n 回中 r 回の"成功"があった場合に元となっている偶然確率 θ を、試行を重ねながら理解しようとしたのです。このときに鍵となったベイズによる考察の貢献は、θ に関する不確実性の表現するために、確率分布を使おうとしました。将来の出来事に伴う予測不可能性から生じる偶然的な (aleatory) 確率というよりも、むしろ、この未知分布は「認識論的な (epistemological)」不確実性を表現するものであり、たとえば偶然性のあるゲームではよく考えられるものでしょう。

よく知られた公理的（測度論的）な確率の定義とは異なり、現代の「ベイズ統計学」では未だに、未知の量に関する不確実性を表現できる確率分布をどのように定式化したらよいかという問題をかかえています。また多くの事前確率の存在など幾年にもわたり議論がなされていますが、世界的にはベイズ流の統計学が主流になっているといえます。システムのパラメータをどう帰納的に理解することができ、将来の観測 (または予測) がどのようにできるかということです。

ベイズの定理を

$$\begin{aligned} \text{事後確率 } P(B|A) &= \frac{\text{事前確率 } P(B) \times \text{尤度 } P(A|B)}{\text{基準化の定数}} \\ &\propto \text{事前確率 } P(B) \times \text{尤度 } P(A|B) \end{aligned}$$

と考えることが基本になっています。ここで、{尤度} = {ある前提条件に従って結果が出現する場合に、観察結果から逆にみて前提条件が「何々であった」と推測する尤もらしさ（もっともらしさ）を表す数値}，観測データ $B = \{Y = b\}$ のとき、事象 A がおこる条件条件つき確率 $P(A|B) = P(A|Y = b)$ を b の関数とみて事象 A を固定してこれを $L(b)$ と表す。これを尤度 $L(b) = (\text{定数}) \times P(A|Y = b)$ とよび、確率密度関数とは別の概念としています。

3 確率分布と期待値

確率の計算は複雑になることがあるから、より複雑な解析を行うために確率変数を導入する。さらにこの確率変数と付随して確率分布を定める。

3.1 確率変数

確率は事象とともに定めたが、事象の結果を数値で表わすには、確率変数をもちいる。たとえばコイン投げの結果 (集合):$\{H\}$、$\{T\}$ という結果を数値 $\{1,0\}$ に対応させる対応（関数）を**確率変数** (random variable) とよび、大文字 X で表す。H は表 (Head),T は裏 (Tail) の意味。

たとえば 1 枚の硬貨を 2 回続けて投げる試行において、表の出る回数を X で表す。4 通りの結果があり、表の出た回数を X とすると、X のとりうる値は 3 つの場合 $\{0,1,2\}$ が対応する。確率変数は変数というより、結果から数値を対応させるから、関数というほうがよいかも知れないが習慣としてこの呼び方をする。

結果 (1 回目,2 回目)	X 表の回数	確率
(H,H)	2	p^2
(H,T)	1	$p(1-p)$
(T,H)	1	$p(1-p)$
(T,T)	0	$(1-p)^2$

表のでる確率を p とすれば、それぞれが起こる確率は $X=0$ となる確率は $(1-p)^2$、$X=1$ となる確率は $2p(1-p)$、$X=2$ となる確率は p^2 となる。この結果を表にすると、次のようになる。

X の値	0	1	2	計
確率	$(1-p)^2$	$2p(1-p)$	p^2	1

これを確率分布表という。またこれを関数として、X の値 x から表現したもの $p_X(x)$：

$$p_X(x)=\begin{cases}(1-p)^2 & \text{if } x=0\\ 2p(1-p) & \text{if } x=1\\ p^2 & \text{if } x=2\end{cases}$$

左の値は 2 項係数を用いて

$p_X(x)=\binom{2}{x}p^x(1-p)^{2-x},\ x=0,1,2$ と表せる。

となる。コインの枚数、あるいは繰り返す数をより一般に n とおくと、

$$p_X(x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,2,\cdots,n \tag{3.1}$$

このとき、記号で $X\sim Binom(n,p)$ と表し、パラメータ n,p の**二項分布** (Binomial distribution) に従うという。

一般に、確率変数 X が有限個の値 $x_1,x_2,\cdots,x_n$ をとり、確率 $p_i=P(X=x_i), i=1,2,\cdots,n$ が与えられば $(i)\ 0\le p_k,\quad (ii)\ \sum_i^n p_i=1$ を満たす。このとき値 x_i と 確率 p_i の対応は下の表のようになります。

値 x	x_1	x_2	$\cdots$	x_n	計
$P(X=x)$	p_1	p_2	$\cdots$	p_n	1

これを**確率密度関数,pdf**(probability density function)（あるいは離散密度 (discrete density)）という。大

文字と小文字の使い分けにとくに注意する。この対応関係を X の**確率分布** (probability distribution) という。$\{X = x_k\}$ となる確率、離散密度を $p_X(x_k) = p_k, k = 1, 2, \cdots, n$ と書く。

これまでの説明はとり得る値が $\{x_1, x_2, \cdots, x_n\}$ という有限個の点集合であったが、一般に偶然な現象の結果は有限個に限ることとはならない。ある区間の実数の値をとる場合もある。有限個の点集合をとり得る値とするときを離散型確率変数といい、実数の値をとり得る場合には、連続型確率変数とよぶ。和の演算の代わりに積分、点集合の代わりに区間が対応する。また差をとることなどは微分をすることになる。

領域として $\{X \leq x\}$ としたものに対する確率, いわば累積分布に対応した形；

$$F_X(x) = P(X \leq x),\ -\infty < x < \infty \tag{3.2}$$

を**分布関数** (distribution function) とよぶ。この関数は (i) 単調増加, (ii) 0 と 1 の値を取り, (iii) 右側から連続、となる。離散型の確率変数では、階段状の増加関数であるが、もし連続関数でさらに微分できるならば、導関数を小文字で $f_X(x) = \dfrac{d}{dx} F_X(x)$ とする。つまり逆に積分すると

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

まとめると

離散型確率変数： (discrete type) $\quad P(a \leq X \leq b) = \displaystyle\sum_{a \leq x_i \leq b} P(X = x_i) = \sum_{a \leq x_i \leq b} p_X(x_i)$

連続型確率変数： (continuous type) $\quad P(a \leq X \leq b) = \displaystyle\int_{\{x\,:\,a \leq x \leq b\}} f_X(x)\, dx = \int_a^b f_X(x)\, dx$

3.2 確率変数の期待値

偶然をともなうゲームを考えよう。このゲームに参加するには参加料 c 円を支払う必要がある。ゲームの結果、ある金額をもらうものとし、それを X と表す。必ずしも正の値とは限らず、負の値（このときには負けてさらに支払わねばならない）かもしれない。簡単のために、この値は r 通りの $x_1, x_2, \cdots, x_r$ という値をとるものとする。このゲームに参加したほうがよいだろうか、あるいは参加しないほうが賢明なのであろうか。もしこのようなゲームをかなり多くの回数繰り返すことができるとすれば、とり得る値の頻度（度数）から、参加の決定により得られる利得を判断できる。n 回ゲームをすれば、nc 円支払うが、$X_1 + X_2 + \cdots + X_n$ を手にする。この収支金額がちょうどゼロになる c の値は $nc - (X_1 + X_2 + \cdots + X_n) = 0$, つまり参加料 c が $c = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ となる。

各々の値は互いに独立な確率変数である。その密度関数は $P(X = x_i) = f_X(x_i), i = 1, 2, \cdots, r$ とする。いま $N_n(x)$ で値 x をとったゲームの回数、つまり $X_1, X_2, \cdots, X_n$ のうち、値 x となっているものの個数 $N_n(x) = \sum_i \mathbf{1}_{\{X_i = x\}}$ とし、これを総回数で割れば、その比率（割合）が求まる。$X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{r} x_i N_n(x_i)$, すなわち

$$\frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{r} x_i \frac{N_n(x_i)}{n} \tag{3.3}$$

大数の法則[*1]から、$N_n(x)/n \xrightarrow{n \to \infty} f_X(x) = P(X = x)$ であるから、

[*1] 大数の弱法則とは、分布が近づくことであり、大数の強法則とは、確実に（除外される可能性はゼロの確率）近づくという概念で

X の値	x_1	x_2	$\cdots$	x_r	計
比率 $N_n(x)/n$	$N_n(x_1)/n$	$N_n(x_2)/n$	$\cdots$	$N_n(x_r)/n$	1
離散密度 $f_X(x) = P(X = x)$	$f_X(x_1)$	$f_X(x_2)$	$\cdots$	$f_X(x_r)$	1

極限として (3.3) の値は $\mu = \sum_{i=1}^{r} x_i f_X(x_i)$ に近づく。したがって左辺はゲーム 1 回あたりの参加費用 c であるから、こうして $\mu > c$ ならば、利益が得られ、逆の不等式ならば、損失をこうむる。等号の $\mu = c$ ならば、利益、損失もない状態である。このような閾値が確率変数 X の期待値

$$\mu = \sum_i x_i \, P(X = x_i) = \sum_i x_i \, p_X(x_i) \tag{3.4}$$

である。 もし連続型確率変数であれば、和が積分に対応するから、密度関数 $f_X(x) = \dfrac{d}{dx} F_X(x) = \dfrac{d}{dx} P(X \leq x), (-\infty < x < \infty)$ をもちいて

$$\mu = \int_{-\infty}^{\infty} x \, f_X(x) \, dx \tag{3.5}$$

確率変数の平均・分散・標準偏差: 確率変数 X の確率分布が離散型:$\{p_X(x_i), i = 1, 2, \cdots\}$, あるいは連続型:$\{f_X(x), -\infty < x < \infty\}$ で与えられているとする。記号 $\mu = E(X)$ を**期待値** (expectation) という。あるいは**平均** (mean) という。これを確率変数の場合には average といわない。期待値は確率変数、確率分布に対する概念で、観測データでいう「平均」は、いわゆるアベレージ、算術平均をさす。

$$E(X) = \sum_i x_i \, p_X(x_i) : \qquad \text{離散型のとき}$$
$$E(X) = \int_{-\infty}^{\infty} x \, f_X(x) \, dx : \quad \text{連続型のとき}$$

また確率変数の分散はとり得る値 x と平均 m との 2 乗偏差 $(x - m)^2$ の期待値、つまり $E[(X - m)^2]$ をいう。2 乗偏差とは X と m とのへだたりの程度を表している。X は変動するから、m の近くであれば、2 乗値は小さいが、離れる場合が多いならば 2 乗偏差は大きくなる。この値の期待値を確率変数 X の**分散** (variance) といい、$V(X)$ で表す。

$$V(X) = E[(X - m)^2] = \textstyle\sum_i (x_i - m)^2 p_X(x_i) : \quad \text{離散型のとき}$$
$$V(X) = \int_{-\infty}^{\infty} (x - m)^2 \, f_X(x) \, dx : \qquad \text{連続型のとき}$$

また、X の**標準偏差** (standard deviation) とは、$\sigma(X)$ などと表し。

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

のことである。期待値の計算を厳密に行うには、積分の理論が必要となるがここでは結果のみを記すことにする。分散 $V(X)$ を表す式は次のように変形できる。2 次式の展開から、$(X - m)^2 = X^2 - 2mX + m^2$ となるから、$V(X) = E[(X - m)^2] = E[X^2 - 2mX + m^2] = E[X^2] - 2mE(X) + m^2 = E[X^2] - m^2$ あるいは $V(X) = E[X(X - 1)] + m - m^2$ が成り立つ。

ある。証明は平均と分散に関する確率の関係式を述べるチェビシェフの不等式により弱法則についての確率収束（あるいは分布収束ともよばれる）は簡単に証明されるが、一方強法則のいう概収束は多少の確率に関する予備知識が必要となる。

問 3.1 *(1)* 離散型：*1* 個のさいころを投げるとき、出る目の数を X とする。確率変数 X の平均、分散、標準偏差を求めよ。*(2)* 連続型：区間 $[0,1]$ 上で同じ確からしさ、区間の長さに比例した確率をもつとき、区間上の点 X を選ぶとき、この平均、分散、標準偏差を求めよ。この分布を**一様分布** *(Uniform distribution)* とよび、$X \sim Unif\{1,2,\cdots,6\}$, $X \sim Unif[0,1]$ と表す。

和の分布；

数枚のコインを投げる場合では，各コインの結果（表が 1，裏がゼロ）をまとめて，和を取ると，表の出た枚数となります。一般にこれは和の分布を計算することです。独立な確率変数について，

$$P(X_1 + X_2 = a) = \sum_k P(X_1 = k)P(X_2 = a - k)$$

が成り立ちます。たとえば、さいころを 3 回投げたときの総和の分布を考えましょう。縦 (列) には 2 回の和 X_2 を書き並べ、横 (行) には 3 回目の値 X_1 を書き、これらの和をとった値を表にしてみました。例として

$$\begin{aligned} & P(X_1 + X_2 = 7) \\ = \ & \textstyle\sum_k P(X_1 = k)P(X_2 = 7 - k) \\ = \ & P(X_1 = 1)P(X_2 = 7-1) + P(X_1 = 2)P(X_2 = 7-2) + \cdots \\ & \quad + \cdots + P(X_1 = 6)P(X_2 = 7-6) \\ = \ & (1/6)(5/36) + (1/6)(4/36) + (1/6)(3/36) + (1/6)(2/36) + (1/6)(1/36) \\ = \ & 15/216 \end{aligned}$$

となります。表の中では斜めの直線で “7” になる部分を加えていきます。$X_1 = 1,2,\cdots,6$, $X_2 = 2,3,\cdots,12$ を行列形にまとめます。

X_2 の値 (確率):X_1 の値 (確率)	1 (1/6)	2 (1/6)	3 (1/6)	4 (1/6)	5 (1/6)	6 (1/6)
2 (1/36)	3	4	5	6	7	8
3 (2/36)	4	5	6	7	8	9
4 (3/36)	5	6	7	8	9	10
5 (4/36)	6	7	8	9	10	11
6 (5/36)	7	8	9	10	11	12
7 (6/36)	8	9	10	11	12	13
8 (5/36)	9	10	11	12	13	14
9 (4/36)	10	11	12	13	14	15
10 (3/36)	11	12	13	14	15	16
11 (2/36)	12	13	14	15	16	17
12 (1/36)	13	14	15	16	17	18

[確率変数の変換] 確率変数 X と定数 a,b に対して、$Y = aX + b$ とすると、Y も確率変数となり、

(i) $E(Y) = aE(X) + b$　　(ii) $V(Y) = a^2V(X)$ (b によらない)
(iii) $\sigma(Y) = |a|\sigma(X)$

同時分布（結合分布）；確率変数はサイコロ振りなどにおいて、その試行からいろいろな値を考えることができる。たとえば、n 回中の表の出た回数、表がはじめて出た回数（何回目に初めて表が出たか)、表の出た回数と裏の出た回数の差、などいろいろと考えられる。いままでは一つだけしか取り扱わなかったが、いろいろと考えられる確率変数を同時に考えてみる。いま試行の結果、2 つの確率変数 X,Y を考える。2 変量 (X,Y)

の結果について，積事象 $\{X=a\}\cap\{Y=b\}$ を $(X,Y)=(a,b)$ と表し，$f_{X,Y}(a,b)=P(X=a,Y=b)$ を同時密度とよびます。周辺分布とは $f_X(a)=\sum_b f_{X,Y}(a,b), f_Y(b)=\sum_a f_{X,Y}(a,b)$ をいいます。行列の表形式で与えられた 2 次元の度数分布表から周辺度数（縦計や横計）を求めたことに相当します。また和の分布は

$$P(X_1+X_2=a)=\sum_{\{(x,y)|x+y=a\}} f_{X,Y}(x,y) \tag{3.6}$$

という意味で，ここで和 $\sum$ は $x+y=a$ となるすべての (x,y) にわたる和をとります。最後の項が独立であれば，確率の積になります。

以上は離散型確率変数ですが，連続型の場合もほぼ同様です。まとめると、

平均の定義式；

$$E(X)=\begin{cases}\sum_i x_i p_X(x_i) & \text{離散型}\\ \int_{-\infty}^{\infty} x f_X(x)dx & \text{連続型}\end{cases} \tag{3.7}$$

離散型 では $E[(X-a)^2]=\sum_i (x_i-a)^2 p_X(x_i)=\sum_i x_i^2 p_X(x_i)-2a\sum_i x_i p_X(x_i)+a^2$ また連続型では $E[(X-a)^2]=\int_{-\infty}^{\infty}(x-a)^2 f_X(x)dx=\int_{-\infty}^{\infty} x^2 f_X(x)dx-2a\int_{-\infty}^{\infty} x f_X(x)dx+a^2$ となります。
分散 $V(X)$ の定義は，上の式で，$a=E(X)$ とおいたもので，

$$V(X)=\begin{cases}\sum_i x_i^2 p_X(x_i)-\{E(X)\}^2 & \text{離散型}\\ \int_{-\infty}^{\infty} x^2 f_X(x)dx-\{E(X)\}^2 & \text{連続型}\end{cases} \tag{3.8}$$

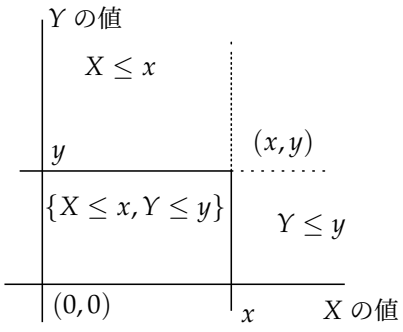
一般に 2 変量確率変数 (X,Y) から定めた実数値関数の確率変数 $h(X,Y)$ の期待値は

$$E[h(X,Y)]=\begin{cases}\sum_i\sum_j h(x_i,y_j)p_{X,Y}(x_i,y_j) & \text{離散型}\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f_{X,Y}(x,y)\,dxdy & \text{連続型}\end{cases} \tag{3.9}$$

同時分布（結合分布）$\{p_X(x_i,y_j)\}$ を取り扱うとき、この関数の値は平面上の点 (x_i,y_j) に高さ $\{p_X(x_i,y_j)\}$ をもつ形である。このとき、一方のみに関する和 $\sum_i p_{X,Y}(x_i,y_j)=p_Y(y_j), \sum_j p_{X,Y}(x_i,y_j)=p_X(x_i)$ となる。この関係式は事象 $A_i=\{X=x_i\}, B_j=\{Y=y_j\}$ を考え、これらが互いに素で、和集合が全事象、すなわち全事象の「分割」になっているからである。これらを**周辺密度** (Marginal distribution) という。単一の変数のときに定義した分布関数に対して、この同時分布における 2 変数の分布関数を定義する。

定義 3.1 同時分布の分布関数：

$$\begin{aligned} & F_{X,Y}(x,y)\\ = \;& P(\{X\le x\}\cap\{Y\le y\})\\ = \;& P(X\le x, Y\le y)\end{aligned}$$



事象の積をカンマで表していることに注意する。

定義 3.2 X と Y の**共分散** *(covariance)* とは

$$cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]=E[XY]-\mu_X\mu_Y$$

ここで μ_X, μ_Y はそれぞれの平均とする。

確率変数の和と積

定義 3.3 確率変数の独立: 確率変数 X のとる任意の値 a と確率変数 Y のとる任意の値 b について、積事象 $\{X=a\}$ かつ $\{Y=b\}$ である確率 $P(\{X=a\}\cap\{Y=b\}) = P(X=a, Y=b)$ がそれぞれの確率の積 $P(X=a)\times P(Y=b)$ に等しいとき、確率変数 X と Y は互いに独立であるという。

定理 3.1 確率変数 X と Y が独立となる必要十分条件は、同時分布関数がそれぞれの積に等しいとき、

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad \forall x,y$$

定理 3.2 確率変数 X,Y について, "独立であってもなくても"

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

式 (3.9) において、$h(x,y) = x+y$ とき、周辺密度より期待値を計算すればよい。一方、積については

定理 3.3 確率変数 X,Y について, "独立であれば"

$$E(X\cdot Y) = E(X)\times E(Y)$$

逆も成り立たないことに注意する。期待値が積になっても一般に独立になるとは限らない。

例題 3.1 *1* 個のさいころを投げるとき、出る目の数を X とする。確率変数 $Y = 2X+3$ の平均、分散、標準偏差を求めよ。

[解] 条件から $P(X=i) = \dfrac{1}{6}, i = 1,2,\cdots,6.$ $E(X) = \sum_{i=1}^{6} i\times\dfrac{1}{6} = (1+2+3+4+5+6)\times\dfrac{1}{6} = \dfrac{7}{2}$
また $V(X) = E(X^2) - \left(\dfrac{7}{2}\right)^2 = (1^2+2^2+3^3+4^2+5^2+6^2)\times\dfrac{1}{6} - \dfrac{49}{4} = \dfrac{91}{6} - \dfrac{49}{4} = \dfrac{35}{12}$ Y の平均は $E(Y) = E(2X+3) = 2E(X)+3 = 2\times\dfrac{7}{2}+3 = 10$, Y の分散は $V(Y) = 2^2V(X) = 4\times\dfrac{35}{12} = \dfrac{35}{3}$.

2 つの確率変数 X,Y の和の分散についても、次のことが成り立つ。

定理 3.4 独立な確率変数の和と差の分散：
(i) 確率変数の和はそれぞれの和に等しい

$$V(X+Y) = V(X)+V(Y)$$

(ii) 差の分散は和になり、分散が小さくなることはない

$$V(X-Y) = V(X)+V(Y)$$

もし独立でないならば,$cov(X,Y)\neq 0$ である。しかし逆にゼロになるとしても、独立になるとは限らない。一般の分散の和、差については共分散の項がともなう。

$$V(X+Y) = V(X)+V(Y)+2cov(X,Y)$$
$$V(X-Y) = V(X)+V(Y)-2cov(X,Y)$$

例題 3.2 *A,B* の *2* 人がそれぞれ *1* 個のさいころを投げる。A は、さいころの目が 3 の倍数ならば 0、3 の倍数

でなければ 1 と記録する。B は、さいころの目が 1 ならば 1、偶数の目ならば 2、1 以外の奇数の目ならば 3 と記録する。A,B の記録する数をそれぞれ X,Y とするとき、$E(X+Y)$ および $E(XY)$ を求めよ。

さいころの目	X	Y	$X+Y$ の値	XY の値
1	1	1	2	1
2	1	2	3	2
3	0	3	3	0
4	1	2	3	2
5	1	3	4	3
6	0	2	2	0

事象	個数	確率
$\{X=1\}$	〒	$4/6=2/3$
$\{X=0\}$	丅	$2/6=1/3$

事象	個数	確率
$\{Y=1\}$	一	$1/6$
$\{Y=2\}$	〒	$3/6=1/2$
$\{Y=3\}$	丅	$2/6=1/3$

$E(X)=1\times 2/3+0\times 1/3=2/3,\ \ V(X)=1^2\times 2/3+0^2\times 1/3-(2/3)^2=2/9$
$E(Y)=1\times 1/6+2\times 1/2+3\times 1/3=13/6,$
$V(Y)=1^2\times 1/6+2^2\times 1/2+3^2\times 1/3-(13/6)^2=17/36$

事象	個数	確率
$\{X+Y=2\}$	丅	$2/6=1/3$
$\{X+Y=3\}$	〒	$3/6=1/2$
$\{X+Y=4\}$	一	$1/6$

事象	個数	確率
$\{XY=0\}$	丅	$2/6=1/3$
$\{XY=1\}$	一	$1/6$
$\{XY=2\}$	丅	$2/6=1/3$
$\{XY=3\}$	一	$1/6$

$E(X+Y)=2\times 1/3+3\times 1/2+4\times 1/6=17/6,$
$V(X+Y)=2^2\times 1/3+3^2\times 1/2+4^2\times 1/6-(17/6)^2==17/36.$
$E(XY)=0\times 1/3+1\times 1/6+2\times 1/3+3\times 1/6=4/3,$
$V(XY)=0^2\times 1/3+1^2\times 1/6+2^2\times 1/3+3^2\times 1/6-(4/3)^2=3-16/9=11/9.$
さらに $cov(X,Y)=E(XY)-E(X)E(Y)=4/3-2/3\times 13/6=-1/9$ より、共分散と関係を調べると、$V(X)+V(Y)+2cov(X,Y)=2/9+17/36+2(-1/9)=17/36$ となり、$V(X+Y)$ と一致している。

問 3.2 大小 2 個のさいころ（区別ができる場合）を同時に投げるとき、それぞれのさいころの出る目を X,Y とする。出る目の和 $X+Y$ の平均、出る目の積 XY の平均、出る目の和 $X+Y$ の分散を求めよ。

3.3 二項分布, 正規分布

1 個のさいころを 3 回投げるとき、1 の目の出る回数を X とすると X は {0,1,2,3} の値をとり、たとえば $\{X=1\}$ は 3 回中 1 回 1 の目が出ることで、1 の目の出る確率は 1/6 で、出ない確率は 5/6 である。3 通りの場合があるから、$3(1/6)(5/6)^2$ となる。このようにしてまとめると、確率変数 X の確率分布は次のようになる。

X の値	0	1	2	3	計
確率	$(5/6)^3$	$3(1/6)(5/6)^2$	$3(5/6)^2(1/6)$	$(1/6)^3$	1

一般に、1 回の試行で事象 A の起こる確率が p であるとき、この試行を n 回行う反復試行において、A の起こる回数を X とすると、確率変数 X の確率分布は次のようになる。ただし、$q=1-p$ である。

X の値	0	1	2	$\cdots$	k	$\cdots$	n	計
確率	$\binom{n}{0}q^n$	$\binom{n}{1}p^1q^{n-1}$	$\binom{n}{2}p^2q^{n-2}$	$\cdots$	$\binom{n}{k}p^kq^{n-k}$	$\cdots$	$\binom{n}{n}p^n$	1

この確率分布を**二項分布** (binomial distribution) といい、Binom(n,p) で表す。

$$P(X=k)=\binom{n}{k}p^kq^{n-k},\ k=0,1,2,\cdots,n \tag{3.10}$$

例題 3.3 *1* 枚の硬貨を *6* 回投げるとき、表が出る回数を X とすると、X は二項分布に従う。二項分布に従う確率変数 X の平均・分散・標準偏差を求めよう。ただし、$q=1-p$ とする。

X の平均は $k\binom{n}{k}=n\binom{n-1}{k-1}, k=1,2,\cdots,n-1$ をもちいると、

$$E(X)=\sum_{k=0}^{n}k\binom{n}{k}p^kq^{n-k}=np\sum_{k=1}^{n-1}\binom{n-1}{k}p^{k-1}q^{n-k}=np$$

また、X^2 の期待値は $X^2=X(X-1)+X$ となるから、$k(k-1)\binom{n}{k}=n(n-1)\binom{n-2}{k-1}, k=2,3,\cdots,n-2$ よって、X の分散は

$$\begin{aligned}V(X) &= \textstyle\sum_{k=0}^{n}k(k-1)\binom{n}{k}p^kq^{n-k}+np=n(n-1)p^2\sum_{k=2}^{n-2}\binom{n-2}{k}p^{k-2}q^{n-k}+np\\ &= n(n-1)p^2+np=npq\end{aligned}$$

一般に、二項分布に従う確率変数について、次のことが成り立つ。

定理 3.5 二項分布の平均・分散確率変数 X が二項分布に従うとき、$q=1-p$ とすると

$$X\ \sim\ \text{Binom}(n,p)\quad\Longrightarrow\quad E(X)=np,\ \ V(X)=npq$$

これは 2 項分布で $Binom(n,p)=\left(8,\frac{3}{10}\right),\left(8,\frac{5}{10}\right),\left(8,\frac{8}{10}\right)$ の 3 通りを書いています。

2 項分布の密度関数:

$$p(k)=\binom{n}{k}p^k(1-p)^{n-k},\ k=0,1,2,\cdots,n$$

において、回数は n は同じ値ですが、p の 3 通りの値によって、得られる値が小さい場合が多い ($p=\frac{3}{10}$), 対称の場合 ($p=\frac{5}{10}=\frac{1}{2}$), 大きい場合が多い ($p=\frac{8}{10}$) となっています。

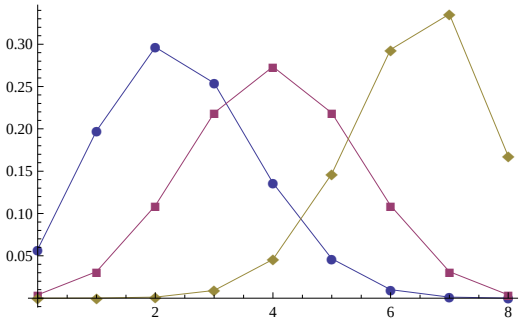
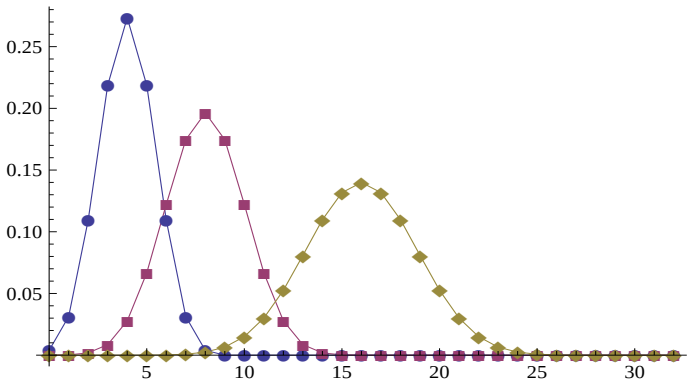


図 1: 2 項分布

次の図は 2 項分布のパラメータで,$p=\frac{1}{2}$ と一定にして繰り返し数 $n=8,16,32$ と大きくした 3 通りのものです。 $B(8,\frac{1}{2}),B(16,\frac{1}{2}),B(32,\frac{1}{2})$ の順によく知られている正規分布の形に近づいています。正規分布は連続型確率分布ですが、後述の中心極限定理です。この命題はそこで述べたいとおもいます。



問 3.3 白玉 7 個と黒玉 3 個が入っている袋から、もとに戻しながら、玉を 100 回取り出す。白玉の出る回数 X の平均、分散を求めよ。

解答 X は繰り返して抽出を行うから二項分布に従う。$X \sim Binom(n,p)$, ここで $n=100$, $p=\dfrac{7}{10}$. $P(X=k)=\binom{100}{k}\left(\frac{7}{10}\right)^k\left(\frac{3}{10}\right)^{100-k}$, $k=0,1,2,\cdots,100$ である。X の平均 $E(X)=\mu_X=\sum_k k\,P(X=k)$ は $k\binom{n}{k}=n\binom{n-1}{k-1}$ より $\mu_X=\sum_{i=0}^{100} k\binom{100}{k}\left(\frac{7}{10}\right)^k\left(\frac{3}{10}\right)^{100-k}=100\sum_{i=1}^{100}\binom{99}{k-1}\left(\frac{7}{10}\right)^k\left(\frac{3}{10}\right)^{100-k}=100\left(\frac{7}{10}\right)\left(\frac{7}{10}+\frac{3}{10}\right)^{99}=70$. また X の分散 $var(X)$ は $var(X)=E(X-\mu_X)^2=\sum_k(k-\mu_X)^2P(X=k)=\sum_k k^2P(X=k)-(\mu_X)^2=E(X^2)-\mu_X^2$ であるが、これよりも $E(X^2)=E(X(X-1))+EX$ と変形しておく。なぜならば $k(k-1)\binom{n}{k}=n(n-1)\binom{n-2}{k-2}$ となるから。これから $E(X(X-1))=\sum_{i=0}^{100}k(k-1)\binom{100}{k}\left(\frac{7}{10}\right)^k\left(\frac{3}{10}\right)^{100-k}=100\cdot99\sum_{i=2}^{100}\binom{98}{k-2}\left(\frac{7}{10}\right)^k\left(\frac{3}{10}\right)^{100-k}=100\cdot99\left(\frac{7}{10}\right)^2\left(\frac{7}{10}+\frac{3}{10}\right)^{98}=n(n-1)p^2=99\times49=4851$. したがって $var(X)=4851+70-70^2=21=np(1-p)$.

ポアソン分布: 起こり難い現象を記述する分布として、ポアソン分布 (Poisson) が知られている。いわゆる稀な事象 (rare event) という意味は 2 項分布において、生起する確率 p が小さいときをさす。もし p がゼロに近く、繰り返しを多数回おこなうとき (n が大きいとき)、得られる分布がポアソン分布である。条件として、ある正の定数 λ があって、$p=p_n\to0, n\to\infty$, さらに積 np_n が $np_n\to\lambda$ を仮定する。ネピア数（自然対数の底）e における関係式 $exp(a)=e^a=\lim_n(1+\dfrac{a}{n})^n$ をもちいる。a の関数として指数関数に他ならない。計算によって $n\to\infty$ とすると $\binom{n}{k}p^k(1-p)^{n-k}\to\dfrac{\lambda^k}{k!}e^{-\lambda}$. これを $\mathrm{Po}(\lambda)$ とあらわす。

$$X\sim \mathrm{Po}(\lambda):\qquad p_X(k)=\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,1,2,\cdots \tag{3.11}$$

2 項分布においては 平均 np、分散 $np(1-p)$ であったから、条件 $np\to\lambda$ から, ポアソン分布の平均は λ で分散のほうは $np(1-p)=np-(np)\times p\to\lambda-\lambda\times0=\lambda$ で、平均と分散が等しくなる。

問 3.4 $X\sim Po(\lambda)$ において、$E(X)=\lambda, var(X)=E(X^2)-\{E(X)\}^2=\lambda$ を示せ。

正規分布; 正規分布は統計学の最も重要な概念となる分布の一つである。2 項分布のパラメータで繰り返し数 n を大きくすると、ひとつの平たい山形が表れてくる。とり得る値が拡がるので適当な横軸の縮約を施すことでゼロを中心とする山形にすることができる。標準正規分布に関する確率変数 $Z\sim N(0,1)$ の密度関数

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right) \tag{3.12}$$

は統計学、確率論では必ず記述される基本の関数である。確率変数 Z の密度関数であれば,$f_Z(z)$ と書くべきかも知れませんが、通常は正規分布の密度関数は $\phi(z)$ (ファイ、フィー) を用います。2 項分布をシュミレーションすることで、正規分布の概形を調べることができる。つまりコイン投げを多数回繰り返した分布がほぼ正規分布に近い。ラプラスが正規分布をみいだす解析の原点である。資料の総数が非常に多いときは、階級の幅を十分細かく分けて、ヒストグラムを作ると、対応する度数折れ線は 1 つの曲線に近づく。その曲線が X の確率分布を表す。X が連続変量である確率変数とする。このとき、次のような性質をもつ曲線 $y = f(x)$ がその分布曲線である。

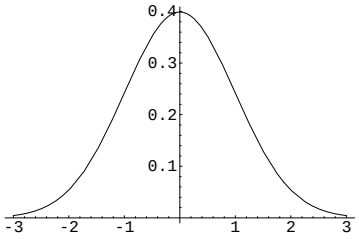


図 2　標準正規分布の密度関数 $N(0,1)$

(1)　非負の関数
(2)　曲線 $y = f(x)$ と x 軸の間の部分の面積は 1 である。
(3)　区間 $a \leq x \leq b$ において、X のとる値が確率 $\int_a^b f(x)dx$ に等しい。

このとき、$f(x)$ を確率変数 X の確率密度関数という。

連続変量 X では、積分で与えられるとするから、$P(X = x) = \lim_{h\to 0} P(x \leq X \leq x+h) = \lim_{h\to 0} \int_x^{x+h} f(x)dx = 0$ となる。このような P のもつ性質を完全加法性とよび予め仮定をしておかなければならない。したがって $P(X = a) = P(X = b) = 0$ であるから、等号があってもなくても確率の値は等しい。つまり $P(a \leq X < b), P(a < X \leq b), P(a < X < b), P(a \leq X \leq b)$, はいずれも等しい値となる。離散型の確率変数とは異なる性質で、とくに補事象の計算において、離散型では「不等号に等号アリ」か、「等号ナシ」に注意しなければならなかったが、連続型確率変数では同じになる。X が連続型確率変数で、その密度関数の曲線が 2 つのパラメータ μ, σ をもちいて、関数

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \tag{3.13}$$

のグラフで表されるとき、X は正規分布 $N(\mu, \sigma^2)$ に従うという。このときはそれぞれ確率変数 X の平均 μ、標準偏差 σ である。密度関数の標記する際、$\frac{1}{\sqrt{2\pi}\sigma}$ と $\frac{1}{\sqrt{2\pi\sigma^2}}$ とは同じであるが、注意しなければならない。

正規分布の平均 μ と標準偏差 σ あるいは分散 σ^2 の計算：$X \sim N(\mu, \sigma^2)$ のとき、

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)dx = \mu, \qquad \mathrm{var}[X] = E(X^2) - E(X)^2 = \int_{-\infty}^{\infty} x^2\, f_X(x)dx - \mu^2 = \sigma^2$$

この計算を行うには微積分の知識（ガンマ関数の計算あるいは 2 重積分の変数変換）が必要となる。X が正規分布に従う確率変数であるとき、関数のグラフを正規分布曲線という。この曲線は、分布曲線の一般な性質のほかに、更に次の性質をもつ。(1)　曲線は直線 $x = \mu$ に関して対称であり、y の値は $x = \mu$ で最大になる。(2)　x 軸を漸近線とする。裾の値は急激に減少する。(3)　標準偏差 σ が大きくなると、曲線は横に広がって山が低くなり、小さくなると、曲線は対称軸 $x = \mu$ の周りに集まって山が高くなる。分散が平均への集中度を表す。また正規分布の山の「とんがり具合」を分布の尖度（せんど）として基準に考える。

X が正規分布 $N(\mu,\sigma^2)$ に従うとき、$Z=\dfrac{X-\mu}{\sigma}$ とすると、標準正規分布 $N(0,1)$ にしたがう。分母には σ で割ることに注意。標準正規分布：$Z\sim N(0,1)$

$$\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\quad -\infty<z<\infty$$

正規分布を標準正規分布 (standard normal distribution) という。

標準正規分布の分布曲線の概形は図 3.3 である。縦軸と横軸の目盛りに注意しておく。縦軸はおよそ 0.4 の値は $\dfrac{1}{\sqrt{2\pi}}=0.3989$ である。

標準正規分布において、確率を

$$\Phi(x)=P(Z\le x)=\int_{-\infty}^{x}\phi(z)dz=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$

で表すとする。いろいろな x の値に対する $\Phi(x)$ の値を表にまとめたものが正規分布表である。

一般の正規分布 $X\sim N(\mu,\sigma^2)$ $\Leftarrow$ $\begin{matrix}Z=\dfrac{X-\mu}{\sigma}\\ X=\sigma Z+\mu\end{matrix}$ $\Rightarrow$ 標準正規分布 $Z\sim N(0,1)$

たとえば、

$(i)\ P\left(-0.5\le Z\le 0.5\right)=0.3829$　$(ii)\ P\left(-1\le Z\le 1\right)=0.6827$
$(iii)\ P\left(-1.645\le Z\le 1.645\right)=0.9000$　$(iv)\ P\left(-1.96\le Z\le 1.96\right)=0.9500$
$(v)\ P\left(-2\le Z\le 2\right)=0.9545$　$(vi)\ P\left(-3\le Z\le 3\right)=0.9973$

であることが知られている。この確率の計算は解析の積分で求めることはできない。数値計算を行わねばならない。表計算ソフトでは変数 z の値から確率は $F(z)=F_X(z)$ は`=NORMSDIST(z)` で求めてから `2*F(z) -1` により、$P\left(-z\le\dfrac{X-\mu}{\sigma}\le z\right)=2*F(z)-1$. また逆に確率 p を与えて、$p=P\left(-z_{p/2}\le\dfrac{X-\mu}{\sigma}\le z_{p/2}\right)=P\left(-z_{p/2}\le Z\le z_{p/2}\right)=P\left(\mid Z\mid\le z_{p/2}\right)$ とする $z_{p/2}$ を求めるためには, `=NORMSINVDIST((1+p)/2)` とすればよい。ここで $z_{p/2}$ としている理由は絶対値での該当範囲が両側あるからです。このような対応値をまとめたものが正規分布表 (Normal distribution table) とよばれる。パソコンではこの算出に表計算ソフトを用いることも便利である。確率の事象における不等号は等号を含んでいてもいなくても、正規分布は連続型確率変数であるから、同じ値になることを注意しておく。また分布の原点に関する対称性: $\Phi(z)=F_Z(z)=P(Z\le z)=1-P(Z>-z)=1-P(-Z<z)=1-P(Z\le -z)=1-F_Z(-z)=1-\Phi(-z)$ も確率の計算には重要である。

問 3.5 Z が標準正規分布に従うとき、正規分布表あるいは表計算ソフトから, つぎの確率を求めよ。
(1) $P(Z<2.5)$　*(2)* $P(1\le Z<2.5)$　*(3)* $P(-0.8<Z)$　*(4)* $P(-0.8<Z<1.5)$

解 (1) $P(Z<2.5)=P(Z\le 2.5)=\Phi(2.5)=0.9938$ (2) $P(1\le Z<2.5)=P(1\le Z\le 2.5)=P(\{Z\le 2.5\}\cap\overline{\{1>Z\}})=P(Z\le 2.5)-P(Z\le 1)=\Phi(2.5)-\Phi(1)=0.9938-0.8413=0.1525$ (3) $P(-0.8<Z)=P(Z>0.8)=1-P(Z\le 0.8)=0.2119$ (4) $P(-0.8<Z<1.5)=P(Z\le 1.5)-P(Z\le -0.8)=\Phi(1.5)-(1-\Phi(0.8))=0.9332+0.7881-1=0.7213$

問 3.6 確率変数 X が正規分布 $N(3,4^2)$ に従うとき、確率 $P(1\le X\le 7)$ を求めよ。

解 X が正規分布 $N(\mu,\sigma^2)$ に従うとき、$Z=\dfrac{X-\mu}{\sigma}$ とすると、標準正規分布 $N(0,1)$ にしたがう。分母には σ で割ることに注意。よって X が $N(3,4^2)$ に従うとき、$Z=\dfrac{X-3}{4}$ が $N(0,1)$ に従う。標準正規分布の確率計算に求められるから、$P(1\leq X\leq 7)=P(\dfrac{1-3}{4}\leq Z\leq \dfrac{7-3}{4})=P(-0.5\leq Z\leq 1)=P(Z\leq 1)-P(Z>-0.5)=\Phi(1)-(1-\Phi(0.5))=0.8413+0.6915-1=0.5328$

3.4 中心極限定理

中心極限定理（ちゅうしんきょくげんていり、英:central limit theorem）CLT とは、確率論・統計学における極限定理の一つで、次のように表現されている。

- どんな確率分布でも、同じ物をたくさん集めて平均を取ると正規分布になる.
- 一言でいうと、誤差の集積の分布は正規分布に近づくというのがその内容である。
- 互いに独立で同一の確率分布に従うような確率変数の標本平均の分布は、正規分布に収束する.
- 母集団分布が正規分布でなくても，標本が大きくなると標本平均値の分布は次第に正規分布に近づく.
- 正規母集団以外であっても，その母平均，母分散をそれぞれ μ，σ^2 として， 大きさ n 個の標本平均値の分布が漸近的に正規分布 $N(\mu,\sigma^2/n)$ になる.
- どんな分布をする集団でも半ば強引に正規分布にしてしまえるこの定理は統計学において極めて重要なもので、ノンパラメトリック (分布に関するパラメータを使わない) を称する検定の多くも統計量が漸近的に正規分布することを利用していたりします.
- 中心極限定理とは、標本の平均の分布が正規分布に近づくということを表した定理です. 元の集団が正規分布ではなくても、標本数が多くなるにつれ、その標本の平均の分布は正規分布に近づきます。
- 平均 μ、分散 σ^2 の任意の確率分布にしたがう母集団から無作為抽出（ランダムにとりだすこと）した n 個の標本からの標本平均 $\bar{x}$ 自体の分布は、$n\to\infty$ のとき、平均 μ、分散 σ^2/n の正規分布になる。元の分布が正規分布である場合には、$n\to\infty$ の条件を必要としない。例えば、$N(\mu,\sigma^2)$ の正規母集団からの標本平均（標本数 n）の分布は $N(\mu,\sigma^2/n)$ になる。正規分布でない一般の分布の場合には、これが中心極限定理として近似的に成立する。母集団の分布は同じものであれば任意でよい（正規分布でなくてもよい)。サンプル数 n が大きくなればなるほど、σ^2/n は小さくなる。したがって、標本平均 $\bar{x}$ が μ に近い確率が高い。
- 大数の法則によると、ある母集団から無作為抽出された標本平均はサンプルのサイズを大きくすると真の平均に近づく。これに対し中心極限定理は標本平均と真の平均との誤差を論ずるものである。多くの場合、母集団の分布がどんな分布であっても、その誤差はサンプルのサイズを大きくしたとき近似的に正規分布に従う。なお、標本の分布に分散が存在しないときには、極限が正規分布と異なる場合もある。統計学における基本定理であり、例えば世論調査における必要サンプルのサイズの算出等に用いられる。出典: フリー百科事典『ウィキペディア（Wikipedia）』 (2009/0 4/25 12:17 UTC 版)
- 標本数が大きくなると標本の分布型によらず，母集団の平均値は正規分布することではありません。いくらたくさん標本をとったところで，ランダム標本を元に戻しながら繰り返して抽出した場合（復元無作為抽出）には、母集団の性質が変わることはありません。中心極限定理とは、どんな母集団からでも無作為抽出され標本サンプルを整理して作った標本平均の分布が正規分布に近づくということです。たとえば一様分布する母集団から標本を得る場合も，対数正規分布に従う母集団から標本を得る場合も，母集団が変化することではなく、その標本の平均値（標本平均、標本の平均値とは算術平均値の意味）

が正規分布に従うということです.

- 元の分布が何であれ、 そこからサンプリングされた標本の平均値は正規分布に従って分布する。にわかには信じがたい話だ.

ド・モアブル Abraham De Moivre(1667-1754) の歴史的なＣＬＴへの貢献は、現在は「二項分布の正規近似」とよばれる。ド・モアブルは 1730 年、”Miscellanea Analytica”での研究題目として出版されている。1733 年”Document of Chance”での結果を現代形で表現するとつぎのようになる：

確率変数が二項分布にしたがうとき、$X \sim \mathrm{Bin}(n,p)$ のとき、

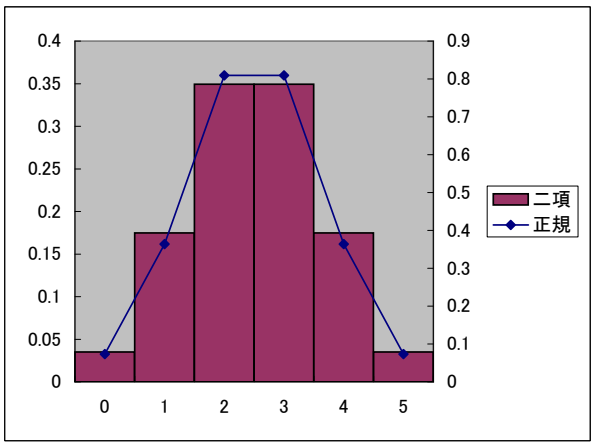
$$P(X=x) = \binom{n}{x} p^x q^{n-x}$$

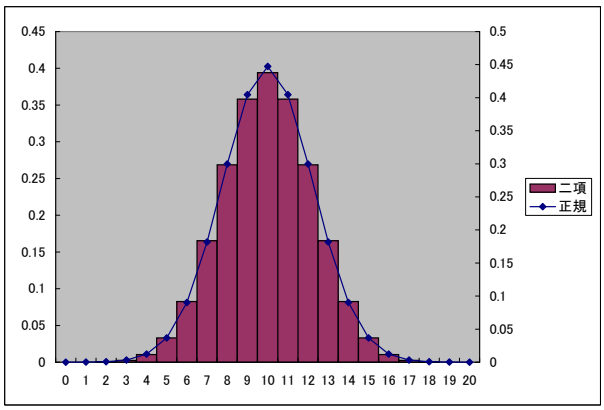
$x = 0,1,2,\cdots,n$ に対して、このとき

$$\sqrt{npq}\binom{n}{x} p^x q^{n-x} \approxeq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-np}{\sqrt{npq}}\right)^2}$$

ただし $q = 1-p$ となる。二項分布の密度と正規分布での密度関数との関係式が示されている。数学の定数 π や e が表れることも極めて興味深い。e は大きな数を扱うために対数で表れるが、π はスターリングによるべき計算 n^n と階乗 $n!$ によるものである。

2 項分布 $Bin(5,1/2)$,$Bin(20,1/2)$ と正規分布について、数値計算の結果: 「表計算ソフト」でグラフを描いたもの。





定理 3.6 確率変数列 $X_1, X_2, \cdots,$ は独立、同一分布にしたがい、有限な平均 μ と分散 $0 < \sigma^2 < \infty$ をもつならば、

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \to N(0,1) \quad (n \to \infty)$$

いいかえると、t について一様に

$$P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \leq t\right) \to \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

問 3.7 ２項分布の確率と正規分布の確率を比較してみよ。

問 3.8 *1* 枚の硬貨を 800 回投げるとき、表が出る回数が 380 回以下である確率を求めよ。

解答 表が出る回数を X とする。X は二項分布に従う。$X \sim \mathrm{Binom}(800, \frac{1}{2})$. X を標準化すると $Z = \dfrac{X - 800 \times \frac{1}{2}}{\sqrt{800 \times \frac{1}{2} \times \frac{1}{2}}} = \dfrac{X - 400}{10\sqrt{2}}$ 800 は十分に大きいので、Z は近似的に標準正規分布 $N(0,1)$ に従うから、$P(X \leq 380) = P(Z \leq \dfrac{380 - 400}{10\sqrt{2}}) = P(Z \leq -\sqrt{2}) = P(Z \leq -1.414) = 1 - \Phi(1.414) = 0.0793$ 実際の値との比較をすると $P(0 \leq X \leq 380) = \sum_{k=0}^{380} \binom{800}{k} 2^{-800} = 0.08395$ である。2 項係数は数が大きいときには電卓では厳しいかも知れない。したがってこのような近似の意味がある。

3.5 正規分布の発見と中心極限定理

どんなに偏りのあるコインであっても，繰り返し投げ続けていくと，すべてが同じ結果，たとえば表ばかりとか裏ばかりに一方的に偏ることがなく，適度な変動をもった結果が集約して得られると想像される。n 枚のコインを投げると（1 枚のコイン投げを n 回繰り返すと），表の出る枚数（回数）は 2 項分布にしたがうことという命題は既に学んだ。この枚数（回数）をどんどん大きくしていくと，2 項分布ではあるが，ある分布に近づくことがわかる。この分布が正規分布である。ラプラス (Pierre Simon de Laplace), 1749-1827 フランスの革命期の数学者、天文学者。数学特に解析学の多くの分野に大きな業績を残したが、古典確率論の大成はその貢献の一つである。

これは古典的な中心極限定理の魁（さきがけ）となっています。中心という言葉は Polya(1920) によるものといわれ、基本的とか「中心的に重要なもの」という意味がこめられている。つまりどんな分布であってもたくさん集めて、その観測値データの算術平均をとれば、その分布は正規分布に近づくという。これを分布収束とよぶ。

定理 3.7 中心極限定理 (Central Limit Theorem) とは $X_1, X_2, \cdots,$ を独立同一分布で共通の平均 μ と分散 σ^2 をもつならば、観測データの和 $S_n = \sum_{i=1}^{n} X_i$ について, $n \to \infty$ とすると,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \to N(0,1)$$

あるいは標本平均 $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ について、

$$\frac{\overline{X_n} - n\mu}{\sigma/\sqrt{n}} \to N(0,1)$$

この中心極限定理は確率論の中心テーマの一つといえる。多くの研究者が取り組んでいるが、とくにロシア（ソビエト連邦）では盛隆であり、拡張された優れた結果が多く発表された。正規分布の歴史は 1730 年ドモアブル（De Moivre）の記事 Miscellanea Analytica（さまざまな解析） 、1733 年 The Doctrine of Chances（偶然の原則）から始まる。さらに 1738 年の The Doctrine of Chances（偶然の原則）第 2 版に 2 項分布を近似する内容が述べられているという。その後ラプラスにより、Analytical Theory of Probabilities（確率の解析的理論）(1812) によって拡張された。

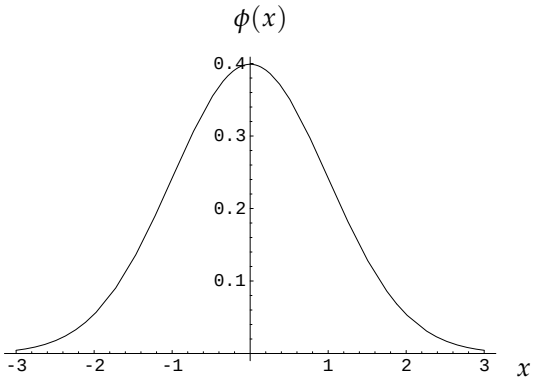
一般のコイン投げを多数回繰り返すと、その分布は正規分布に近い形でことがわかる。つまり 2 項分布の極限は正規分布である。このように 2 項分布を正規分布で近似することは、ドモアブル・ラプラスの定理とよばれる。

確率変数がパラメータ n, p の 2 項分布にしたがうならば、記号では $X \sim Bin(n,p)$ と表し, その密度関数が $P(X=x) = \binom{n}{x} p^x (1-p)^x$ であるならば、

$$P(X=x) \sim \frac{1}{\sqrt{2\pi}} \exp\left\{ -\left(\frac{x-np}{\sqrt{np(1-p)}} \right)^2 /2 \right\} \frac{1}{\sqrt{np(1-p)}}$$

(2 項分布の確率密度)　$\sim$　(正規分布の確率密度関数)

標準正規分布の密度関数の曲線です。滑らかに変化し、ひとつ山の形をしています。



このグラフの式はつぎで与えられるものです。

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2x^2}\right) = \frac{1}{\sqrt{2\pi}}\, e^{-1/2x^2}$$

実際、数値を当てはめてみると比較をしてみます。かなり正確に当てはまることがわかる。いま 2 項確率とは $\sqrt{npq}P(X = x)$ 正規確率とは $\phi(u)$, $u = \dfrac{x - np}{\sqrt{npq}}$, $q = 1 - p$ として計算してみます。たとえば $n = 10$, $p = 0.7$ のとき、

値	0	1	2	3	4	5	6	7	8	9	10
2 項確率	0.000	0.000	0.002	0.013	0.053	0.149	0.290	0.387	0.338	0.175	0.041
正規確率	0.000	0.000	0.001	0.009	0.047	0.154	0.314	0.399	0.314	0.154	0.047

もう少し n を大きくして $n = 30$, $p = 0.7$ のときの数値結果は

値	10	12	14	16	18	20	22	24	26	28	30
2 項確率	0.000	0.001	0.011	0.058	0.188	0.355	0.377	0.208	0.052	0.005	0.000
正規確率	0.000	0.001	0.008	0.055	0.195	0.369	0.369	0.195	0.055	0.008	0.001

証明はすこし技術的な要素が強いですが、スターリングの公式 $m! = \sqrt{2\pi m} m^m e^{-m}$ と対数関数の近似（テーラー展開）$\log(1 + x) = x - \dfrac{x^2}{2} + O(x^3)$ をもちいる。$\sqrt{npq}\binom{n}{x} p^x q^{n-x} \sim \dfrac{1}{2\pi} e^{-\frac{u^2}{2}}$ の両辺の対数をとって示すことが多い。$x/n - p = u\sqrt{\dfrac{pq}{n}}$, $(n - x)/n - q = -u\sqrt{\dfrac{pq}{n}}$ であり、先の公式から

$$\log \frac{x/n}{p} = \log\left(1 + u\sqrt{\frac{q}{np}}\right) = u\sqrt{\frac{q}{np}} - \frac{u^2}{2}\frac{q}{np}$$

,

$$\log \frac{(n - x)/n}{q} = \log\left(1 - u\sqrt{\frac{p}{nq}}\right) = -u\sqrt{\frac{p}{nq}} - \frac{u^2}{2}\frac{p}{nq}$$

が得られ、より高次の n については無視することで近似される。

つぎは Pierr Simon Laplace(1749-1827) による厳密な証明である。彼は特性関数（ラプラス変換と同様）とその逆変換という技法をもちいた。つぎの変換 $\psi(t) = E(e^{itX})$ をおこなって,$n \to \infty$ とした極限関数とが

正規分布のそれと一致することから、対応の一意性によって厳密に証明した。確率論では強力に使える手法である。まず 2 項分布（ベルヌイ分布；1 枚のコイン投げ）の特性関数を計算し、独立な和の分布は積になるから、2 項分布の特性関数が求められる。コインの枚数 (n) を多くするから、これを n の値で評価する。微積分でのテーラー展開でもちいて、極限を計算する。この極限関数が、正規分布の特性関数に対応するものであり、対応が 1 対 1 であることから、証明される。

一方，ガウス (Karl Friedrich Gauss), 1777-1855 が正規分布について極めて大きな貢献を与えた。いわゆる中心極限定理である。ガウスはドイツの数学者、物理学者、天文学者で、幼少時より天才の誉れ高く、数学・物理学に巨大な功績を残している。ラプラスが議論した、２項分布の極限として正規分布を導いた方法とは全く異なり，天体観測の誤差測定を解析し，変動があり得る状況から，微分方程式をたてて，正規分布を導いている。「誤差論」円錐曲線で太陽の回りを回る天体の運動理論—多くの観測結果にもっともよく合う軌道の決定—という論文（１８０９年）で密度関数を示した。

ガウスは、誤差の解析をすることで、微分方程式 $\frac{d}{dx}\phi(x) = -x\phi(x)$ を導き、その解を求める。この微分方程式は変数分離形であるから

$$\log\phi(x) = \int(-x)dx + const. = \{x \text{ の 2 次式 }\} = cx^2\text{の形（条件から）} = \text{（係数）} \times \frac{-1}{2}x^2$$

と求められる。

ガウスの考えかたは（観測値）＝（未知の真の値）＋（誤差分布）ととらえ、もし n 個の観測値があれば、それらを $M_i = z + v_i, i = 1, 2, \cdots n$ とおく。また誤差分布はゼロを中心として対称（正と負の両方の値をとる）で標本 n が多数であればゼロに近いとし、$\frac{M_1 + M_2 + \cdots + M_n}{n} = z$ 分布の確率は密度 $\phi(v)dv$ をもつとする。$v_i = M_i - z$ であるから、独立として積の形になるから $P = \phi(v_1)dv_1 \times \phi(v_2)dv_2 \times \cdots \times \phi(v_n)dv_n$ この P を最大にするには 対数をとって $\log P$ を最大にすればよい、すなわち $\log\phi(v_1) + \log\phi(v_2) + \cdots + \log\phi(v_n)$ を微分してゼロになるようなところを定める。

$$\frac{\partial P}{\partial z} = \frac{1}{\phi(v_1)}\frac{\partial\phi(v_1)}{\partial z_1} + \frac{1}{\phi(v_2)}\frac{\partial\phi(v_2)}{\partial z_2} + \cdots + \frac{1}{\phi(v_n)}\frac{\partial\phi(v_n)}{\partial z_n} = 0$$

合成関数の微分から $\frac{\partial\phi(v)}{\partial z} = \phi'(v)\frac{\partial v}{\partial z}$ であり、いま $\frac{\phi'(v)}{\phi(v)} = \psi(v)$ とおけば

$$\psi(v_1)\frac{\partial v_1}{\partial z_1} + \psi(v_2)\frac{\partial v_2}{\partial z_2} + \cdots + \psi(v_n)\frac{\partial v_n}{\partial z_n} = 0$$

が成り立ち、仮定からすべての i で $v_i = M_i - z$ としたから $\frac{\partial v_i}{\partial z_i} = -1$ で共通だから、

$$\psi(v_1) + \psi(v_2) + \cdots + \psi(v_n) = 0 \quad \cdots\cdots\cdots \quad (a)$$

の形に帰着される。また

$$v_1 + v_2 + \cdots + v_n = 0 \quad \cdots\cdots\cdots \quad (b)$$

である。なぜなら $v_1 + v_2 + \cdots + v_n = (M_1 - z) + (M_2 - z) + \cdots (M_n - z) = M_1 + M_2 + \cdots + M_n - nz = 0$ したがって 2 つの関係式 $(a), (b)$ から $\psi(x) = cv$ (c は定数) の形でなければならない。元の関数で微分がはいった式では $\frac{\phi'(v)}{\phi(v)} = cv$、$\log\phi(v) = c\frac{v^2}{2} + c'(c, c' : const)$。この c, c' は $\phi(v)$ の対称性条件から $\phi(v) = ke^{-h^2v^2}(k, h : const)$ という正規分布の形に結論として得られた。この係数の値は $k = \frac{1}{\sqrt{2\pi}}, h = \frac{1}{\sqrt{2}}$ である。

ラプラスの方法は解析の方法がかなり難しく簡単には述べられない。が、汎用性に富んでいる。これに比べればガウスの方法はシンプルであるといえる。しかし発想には天才による煌きがある。

大数の法則 (Law of Large Numbers): いま，表の出る確率が p, 裏のでる確率が $1-p$ のコインがあったとする。このコインを投げ表が出れば，1点，裏ならば0点として，この点数が表される各回の結果を $X_1, X_2, \cdots, X_n$ とおく。$\Sigma_i X_i$ は1がでた回数で，表の出た比率はデータ平均 $\overline{X}$ となる。大数の法則とは，この表の出た比率は，繰り返し数をどんどん大きくすると，1回投げたときに表れるべき表の確率の値 p に近づくことという命題である。近づくという意味を数学的には，確率収束とか，概収束で正確に書き下される。より一般には，「独立な同一分布にしたがう確率変数列の標本平均（観測値を全体個数で割ったもの，データの平均，算術平均）は，個数を大きくしていくともとの分布の期待値（分布の平均）に確率収束（概収束）する」ということである。この定理は，抽出した標本数を大きくしていけばいくほど，母集団の平均を明らかにすることができるという，統計学の基本定理の一つである。

中心極限定理 (Central Limit Theorem): 大数の法則が，標本の平均（観測値の和をデータ総数で割ったもの）と分布の期待値（平均値）との関係を明らかにしているが，中心極限定理とは，同じようにデータ数を大きくしていった状況の近づきかたを示している。つまり，２項分布の極限が正規分布であるということがひとつの例である。繰り返されるコイン投げの列などを例にして，確率変数 $X_1, X_2, \cdots, X_n, \cdots$ は独立ですべて同じ平均 μ, 分散 σ^2 をもつとします。この n 個の観測結果から，標本平均　$\overline{X}_n = \Sigma_{i=1}^n X_i$ を作る。大数の法則から $\overline{X}_n - \mu$ は 0 に確率収束します。これではどのように近づいていくのか分からないので，$\dfrac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}$ を考える。この分数の分母には，$\overline{X}_n$ の分散の平方，すなわち，標準偏差です。もちろんこの値は n が大きくなれば，ゼロに近づきますが，分子もゼロですから極限の不定形になっています。ラプラスは，この極限分布が正規分布とよばれる分布であることを示しました。数学的に表現すると，任意の実数 a, b に対し，とすれば

$$\lim_{n\to\infty} P\left(a \le \frac{\overline{X_n} - \mu}{\sqrt{\sigma^2/n}} \le b\right) \to \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})\, dx$$

が成り立つ。右辺の積分値のなかに表れている関数が標準正規分布の密度関数であり，ガウスが微分方程式で導いたものと同じである。積分の計算がそのままになっている理由は，原始関数が求められない（つまり積分が求められない）から，形を書いているだけである。もし実際に a, b を入れて，この値を求めるには「正規分布表」とよばれる，区間と対応する確率（積分値）を予め数値計算したものを準備しておかねばならない。簡単に言えば，面積の計算が積分で求められないから数値計算したというわけである。この数値表は，統計学のテキストには必ず掲載されているし，今後の推定や検定にはよく用いられるので手元に準備しておく必要がある。

この結果をいいかえると標本平均 $\sqrt{n}(\overline{X}_n - \mu)$ の漸近分布 (個数が大きく場合の分布）は $N(0, \sigma^2)$ である。ここで $N(\mu, \sigma^2)$ とは平均 μ と分散 σ^2 の一般正規分布を表す。

$$\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \quad -\infty < x < \infty$$

係数は面積が１となるようにしたもので，$\dfrac{1}{\sqrt{2\pi}\sigma}$ とか $\dfrac{1}{\sqrt{2\pi\sigma^2}}$ などとも表すので違いに注意する。グラフの概形は，ベル型をした一つ山で，頂点 $x = \mu$ が極大，２つの変曲点 $x = \mu \pm \sigma$。ここで変曲点とは，上にふくらんだ状態から下にふくらみをもつ状態に変わるところ，極大・極小が１階の微分係数をゼロとして求められ，変曲点は凸と凹の境目で，２階の微分係数をゼロとして求められる。

ガウスの示した中心極限定理のすばらしいことは，もとの分布がコイン投げ（パラメータ p のベルヌーイ分

布）であろうがさいころ振り（６点集合の離散型一様分布）でもカマワナイこと。離散型，連続型のものであっても，極限は連続型の正規分布になる。多少の偏りがあっても，独立な繰り返しとその平均と分散の存在が保証されている限り，どんな分布であっても極限は正規分布になります。非常に強力で有用な定理であることがわかるだろう。

２項分布の正規近似をおこなうときに，ひとつ注意をすることがある。半整数補正とよばれる近似をよくするためのテクニックである。２項分布は整数の値をとる離散型で，一方正規分布は実数値をとる連続型であるから，次のような補正をする。

確率変数 X をパラメータ n, p の２項分布とするとき，中心極限定理から平均 np, 分散 $np(1-p)$ の正規分布 Y で近似される。また標準正規分布を Z とすると，このばあいの半整数補正とは

$$P(X=k) \fallingdotseq P(k-0.5<Y<k+0.5) = P\left(\frac{(k-0.5)-np}{\sqrt{np(1-p)}} < Z < \frac{(k+0.5)-np}{\sqrt{np(1-p)}}\right),$$
$$P(i \le X \le j) \fallingdotseq P(i-0.5<Y<j+0.5) = P\left(\frac{(i-0.5)-np}{\sqrt{np(1-p)}} < Z < \frac{(j+0.5)-np}{\sqrt{np(1-p)}}\right)$$

3.6 正規分布の確率計算

正規分布に関する確率を計算するには，正規分布表をもちいる。この表は平均 0，分散 1 の標準正規分布 Z の分布関数を表にまとめたものです。横軸にとり得る値 x（$-\infty$ から ∞）をとり縦軸に $\Phi(x) = P(Z \le x)$ を対応させます。密度関数の左右対称性から，正の値のみを書いてあるものが多いですが，これからでもすべての値を計算できます。連続型であることから，$P(Z \le x) = P(Z < x)$ が成り立つ，離散型とことなり，等号が含まれていてもいなくても同じ値であることに注意します。

いくつかの点を取り上げると，

x	0.000	1.00	1.282	1.645	1.960	2.00	2.241	2.326	2.576	2.807	3.00
$\Phi(x)$	.5000	.8413	.9000	.9500	.9750	.97725	.98745	.9900	.9950	.9980	.99865

正規分布表は「変数の値」と「(値以下の）確率」の対応を数値計算したものですが、すべては表していません。中間にある対応を求めるには「線形補間」をおこないます。三角形の相似形を応用したものです。確率 $\phi(x_i) = a_i, i = 1, 2$ が与えられたとき、$a_1 < \alpha < a_2$ に対する値は $c = \dfrac{x_2 - x_1}{a_2 - a_1}(\alpha - a_1) + x_1$ で近似されます。

例題

(1) $P(0<Z<1.2) = P(0<Z\le 1.2) = P(0\le Z<1.2) = P(0\le Z\le 1.2) \fallingdotseq 0.3849$
(2) $P(-1.2<Z<0) = P(0<Z<1.2) \fallingdotseq 0.3849$
(3) $P(-1.2<Z<1) = P(0<Z<1.2) + P(0<Z<1) \fallingdotseq 0.3849 + 0.3413 = 0.7262$
(4) $P(Z<-1.2) = P(Z<0) - P(-1.2<Z<0) = 0.5 - P(0<Z<1.2) \fallingdotseq 0.5 - 0.3849 = 0.1151$

問 3.9 確率変数 X が平均 10, 分散 4 の正規分布にしたがうとき，つぎの値をもとめよ。
(1) $P(X \le 13)$ *(2)* $P(X > 11)$ *(3)* $P(9 < X < 12)$ *(4)* $P(X < c) = 0.1$ を満たす c の値

問 3.10 ある測定器具の誤差は平均 0, 標準偏差 0.2(分散 $0.2^2 = 0.04$)mm の正規分布にしたがう。この器具で測定誤差が 0.5mm 以上となる確率はいくつか（ヒント；誤差は両側を考えるから，絶対値として一定以上となる確率を求める）

問 3.11 ポアソン分布について，$P(X=k+1)$ と $P(X=k)$ の関係式 $(k=0,1,2,\cdots)$ を導き，２項分布と同様なアルゴリズムを作りなさい。

問 3.12 コインを 400 回投げるとき，表のでた回数が 180 回以上，210 回以下となる確率はいくつか？半整数補正も加えて，正規分布近似で計算しなさい。

問 3.13 中心極限定理の応用として，正規乱数を求める方法の根拠を説明しなさい。（ヒント；和の分布として平均と分散を計算してみなさい）

20 面体のサイコロでこれを乱数サイとよんでいます。3 個をころがして、その結果を記録するのですが、いまはコンピュータのシュミレーションがふつうに行われます。

4　記述統計学

様々な統計資料の整理とグラフ化や代表値・標準偏差などの基礎概念、また実際の処理がどのように行われるかをコンピュータの表計算ソフトを利用して学習します。表計算のセクションは予め各自使用している表計算ソフトの操作を知っておくとスムーズに学習が進められます。また優秀なフリーウェアもありますから、気軽に利用できます。

4.1　資料の整理

ここでは様々な統計資料を視覚的に分かりやすくなるようにまとめることを具体的な例を用いて学習する。

資料の分布:

以下の資料はある学校の生徒 10 人の体重をまとめた資料である。

出席番号	1	2	3	4	5	6	7	8	9	10
体重（kg）	60.3	57.9	65.4	56.1	53.6	62.7	65.3	55.8	67.1	63.1

個々の生徒の体重は読み取りやすいが全体の傾向は読み取りにくい。以下は上の資料から読み取った値を階級値の 1 つが 62.5kg、その前後 ±1.5kg の 3.0kg 毎に階級の区間を定め、その区間に該当する生徒の人数を記録してまとめたものである。

階級	52.0 以上 ～ 55.0 未満	55.0～58.0	58.0～61.0	61.0～64.0	64.0～67.0	67.0～70.0	計
階級値	53.5	56.5	59.5	62.5	65.5	68.5	
度数	1	3	1	2	2	1	10
相対度数	0.1	0.3	0.1	0.2	0.2	0.1	1.0
累積相対度数	0.1	0.4	0.5	0.7	0.9	1.0	-

このように値をいくつかの区間に区切り全体の傾向を読み取りやすくする時、その区間（ここでは体重）を階級、またその幅を階級の区間という。また、階級の区間の中央にくる値をその区間の階級値（級中央値）という。各階級に該当する資料の個数（ここでは人数）を度数、各階級に度数を組み込んだ表を度数分布表という。

累積度数: それぞれの階級以下、または階級以上の度数を全て加えた和を累積度数といい、それを表にまとめたものを累積度数分布表という。

相対度数: それぞれの階級の度数を資料の総数で割った値をその階級の相対度数といい、それを表にまとめたものを相対度数分布表という。相対度数分布表では各階級の相対度数の総和は 1 となる。つまりパーセントで表示したもの。データ数が異なる集団を比較する場合、大きさが相対的割合となる比べ易くなる。

ヒストグラム: 度数分布表を更に整理して柱状のグラフに表したものをヒストグラムという。各長方形の高さは各階級の度数に比例する。離散データを表す棒グラフにはこのような幅に意味をもたないが、連続データを丸めて階級にしているヒストグラムには幅が意味をもつ。また、ヒストグラムの各長方形の上の辺の中点を結んでできるグラフのことを度数折れ線という。但しこのグラフを作る際は左右両端に度数が 0 である階級があるものとして作図をする。また、同じ目盛幅であればヒストグラムの囲む面積と度数分布表の囲む面積は等しい。

4.2 代表値

資料の分布についてはヒストグラムなどからも得ることができるが全体の特徴を 1 つの数字つまり「尺度」として表すことにより分かりやすくできる。このような値を資料の代表値あるいは統計量という。ここではよく用いられる代表値やその意味を考える。

4.2.1 平均 (平均値)Average

変量が取るいくつかの値がある 1 組の資料でその階級値の総和を資料の個数で割ったものを変量の平均値と言う。資料の平均値たんに平均とは n 個の資料 $\{x_1, x_2, \cdots, x_n\}$ の算術平均をさす。

度数分布表からも、平均値の近似値を求めることができる。このときは、各階級に属する資料の値を丸めているので元のデータとは異なるが、その階級値に等しいと考えて計算する。資料 x の度数分布表で、階級が r 個に分類されたとして、第 i $(i = 1, 2, \cdots, r)$ 番目の階級値を a_i とし、それに対応する度数を f_i とする。

このとき、総度数 n と総和は

$$n = f_1 + f_2 + \cdots + f_r$$
$$x_1 + x_2 + \cdots + x_n = a_1 f_1 + a_2 f_2 + \cdots + a_r f_r = \sum_i a_i f_i$$

a	a_1	a_2	$\cdots$	a_r	計
f	f_1	f_2	$\cdots$	f_r	n

であるから、資料 x の平均は次のようになる。

$$\text{(I)}\quad \text{データの直接計算：}\ \overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k$$
$$\text{(II)}\quad \text{度数分布表からの平均：}\ \overline{X} = \frac{1}{n}\sum_i a_i f_i$$

変量の数が多い時に上記のような計算をすると計算をしなければいけない数が多くなるので、時間がかかったり計算間違いが起こる可能性が少なくない。そこで、計算をより簡単にするための方法を考えてみよう。適当に c を選ぶことで各階級の値 $x_i - c$ が絶対値を小さくして計算しやすい数になるようにすれば、平均値の計算を簡単にすることができる。x_i を新たな値 $y_i = x_i - c$ にすることを変量の変換といい、またその値 c を仮平均という。

問 4.1 仮平均をもちいたとき、変量 X と Y の平均はどういう関係式があるか？

4.2.2 中央値 (中位数, メディアン) Median

資料データ $\{x_1, x_2, \cdots, x_n\}$ を大きさの順に並べ, $\{x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}\}$ とした時、中央の順位にくる数値をその資料の中央値またはメジアンと言う。Me とも表す。資料が偶数個 n の場合ならば、$n/2$ 番目と $(n+1)/2$ 番目の中間にあたり、中央に 2 つの値が並ぶので、そのときは 2 つの数値の算術平均を中央値 $Me = \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}}{2}$ とする。もし奇数であれば、一つだから $Me = x_{(\frac{n+1}{2})}$ となる。中央値は全体の 50% がそれより以下の部分で、以上の部分が 50% ととなる。大きさの順に並んだとき、前後に半分づつとなる真中である。

4.2.3 最頻値 (モード) Mode

度数分布表において度数が最大である階級値をその資料の最頻値またはモードという。Mo とも表す。

度数分布での平均、中央値、最頻値を比較すると、もし分布の形状が L 字型ならば、$Mo \leq Me \leq \overline{X}$ 、逆 L 字型であれば、逆順となる。

箱ひげ図（Box Whisker Chart) では、ひとつのヒストグラムを 1 本の直線形状で表そうとするもの。度数分布表から基本統計量 (1) データ最小値 (min), (2) 25% 点 (第一四分位数:Q_1), (3) 平均 (average), (4) 50% 点 (メジアン;$Me = Q_2$), (5) 75% 点 (第三四分位数:Q_3),(6) データ最大値 (max) を計算する。四分位数の間を箱で表現し、最大、最小までをそれぞれひげのように範囲を示すために伸ばす。時系列データのように、ヒストグラムをいくつも並べることができないときには有効は表現方法である。

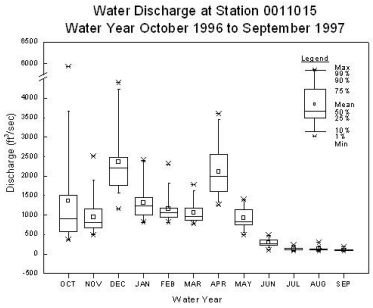


図 3　箱ひげ図の例

4.3　散らばり具合

代表値が同じであってもその分布が代表値近くに密集していたりばらばらであったりと色々なことが考えられる。ここでは資料の散らばり具合の表す量について見てみよう。

4.3.1　**範囲** Range

資料データが変動した最大値から最小値を引いた値をその資料の分布の範囲と言う。$R = \max_i x_i - \min_i x_i$ をさす。資料データを 25% ずつ 4 つに分けたとき、真ん中が中央値であるが、これを含む中央部の 50% の範囲、両裾の 25% を取り去ったものを四分位範囲という。ボックス・チャート (box chart) とは、最大値、最小値、平均をプロットし、四分位範囲を箱で書き表したもので、時間的な変化の様子と表すときによく用いられる。

4.3.2　**平均偏差** Mean Deviation

変数 x のとる値が $x_i, i = 1, 2, \cdots, n$ の n 個あるとき、各値と平均値との差 $|x_i - \overline{X}|$ を、それぞれ平均値からの（絶対）**平均偏差** (MD:mean deviation) という。絶対値を取るから、平均までの距離を表すと考えてよい。資料の下記例で、体重の平均 $\overline{X} = 61.2$ からの絶対偏差は次のようになる。

出席番号	1	2	3	4	5	6	7	8	9	10	計
体重	60.3	57.9	65.4	56.1	53.6	62.7	70.0	55.8	67.1	63.1	-
偏差	-0.9	-3.3	4.2	-5.1	-7.6	1.5	8.8	-5.4	5.9	1.9	0.0
絶対偏差	0.9	3.3	4.2	5.1	7.6	1.5	8.8	5.4	5.9	1.9	44.6

このときの平均偏差は $MD = \dfrac{44.6}{10} = 4.46$。絶対値をとらない単に差を計算して、その平均を計算すれば常に 0 になる。$\sum_i (x_i - \overline{X}) = 0$。この関係式は平均の定義式から導ける。

4.3.3 **分散** Variance **と標準偏差** Standard Deviation

[分散と不偏分散]: 偏差の 2 乗 $(x_i - \overline{X})^2,\ i = 1, 2, \cdots, n$ の平均値を考える。この値を**分散** (variance) という。分散を s^2 で表す。

$$s^2 = \frac{1}{n}\sum_i = \frac{1}{n}\sum_i x_i^2 - \overline{X}^2 = \frac{n\sum_i x_i^2 - (\sum_i x_i)^2}{n^2}$$

また 2 乗和は, 度数分布表からの計算では $x_1^2 + x_2^2 + \cdots + x_n^2 = a_1^2 f_1 + a_2^2 f_2 + \cdots + a_r^2 f_r = \sum_i a_i^2 f_i$ と等しいから、

$$s^2 = \frac{1}{n}\sum_i (a_i - \overline{X})^2 f_i = \frac{1}{n}\sum_i a_i^2 f_i - \overline{X}^2$$

この分散の定義は自然なものであるが、データの個数 n があるが、関係式 $\sum_i (x_i - \overline{X}) = 0$ がある。そこで n で割るのではなく、$n-1$ で割ったものを考えることがある。これを**不偏分散** (unbiased variance) といい、

$$u^2 = \frac{1}{n-1}\sum_i (x_i - \overline{X})^2 = \frac{1}{\binom{n}{2}}\sum_{i<j} \frac{(x_i - x_i)^2}{2}$$

と表す。ここで $\sum_{1 \le i < j \le n}$ とは、条件 $\{(i, j)\ ;\ 1 \le i < j \le n\}$ を満たす組にわたる 2 重和でその個数は $\binom{n}{2}$ あるから、$\dfrac{(x_i - x_i)^2}{2}, i, j = 1, 2, \cdots, n$ を平均したもの。不偏という用語は、標本の分散と母集団の分散とのずれがないという意味で、推定の節でさらに述べる。

分散の単位を考えると、たとえば、データが身長の場合には、その単位は cm であるから、分散の単位は偏差の 2 乗だから、その単位は cm^2 乗になってしまう。そのため、単位を変量と合わせるために、不偏分散 u^2 の正の平方根 $u = \sqrt{u^2}$ を考えることが多い。この u を 変量 x の**標準偏差** (standard deviation) という。

[不偏分散と標準偏差]:

$$u^2 = \frac{1}{n-1}\sum_{i=1}^n x_i^2 - \frac{n}{n-1}\overline{X}^2$$

$$u = \sqrt{u^2} = \sqrt{\frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}}$$

表計算ソフトでは

統計量	命令書式	説明
平均 $\overline{X}$	AVERAGE(データ列)	データ列:$\{x_1, x_2, \cdots, x_n\}$
中央値 (メジアン)Me	MEDIAN(データ列)	
最頻値 (モード)Mo	MODE(データ列)	
四分位数 Q	QUARTILE(データ列、タイプ)	タイプとは 0: 最小値、1:第 1 四分位数 (25%), 2: 第 2 四分位数 (50%)=中央値、3: 第 3 四分位数 (75%)、4:最大値
標準偏差 u^2	STDEV(データ列)	(個数 - 1) の値 = (n-1) 法を使って計算
分散の平方根 s	STDEVP(データ列)	n で割る式
不偏分散 u^2	VAR(データ列)	(n-1) 法
分散 s^2	VARP(データ列)	n で割る式

例題 4.1 つぎの資料（体重）の分散と標準偏差などを求めよう。

体重; *60.3 57.9 65.4 56.1 53.6 62.7 70.0 55.8 67.1 63.1*

平均 $\overline{X}$	61.20	分散 s^2	26.04
標準偏差 (SD) u	5.38	不偏分散 u^2	28.93
中央値 (Me)	61.50	最頻値 (Mo)	Non available
4 分位数 (Q_0)	53.60	(Q_1)	56.55
(Q_2)	61.50	(Q_3)	64.83
(Q_4)	70.00		

4.3.4 **偏差値** Score

$\{x_1, x_2, \cdots, x_n\}$ の中の数値 x_i の偏差値 z_i とは、データに対して線形変換（アフィン変換）を施したものである。線形とは変量 x に対して、定数 a, b をもちいて、$y = ax + b$ とすること。$b = 0$ ならば、線形変換である。

$$y_i = \frac{x_i - \overline{X}}{u},\ i = 1, 2, \cdots, n, : \quad a = \frac{1}{u}, b = \frac{-\overline{X}}{u} \tag{4.1}$$

さらに $z_i = 10y_i + 50$ と変換する。10 とか 50 といった定数は、出てきた数値が直感的にわかりやすい大きさとなるようにしている定数（規格化定数という）であり、直接に意味はない。注目すべきは、この計算式の中に、平均と標準偏差が含まれているということである。つまり、同じ学力を持った人どうしであっても、違う試験を受ければ、試験を受けた他の人たちの動向によって偏差値は大きく変化するということである。そのような数値であるので、少しの変化にあまり一喜一憂しすぎないようにしたい。

平均は中心的な位置（重心）を表し、分散は平均への集中度を表したが、これ以外にも、分布の形状を調べるためにいろいろな尺度がもちいられる。全体での相対的位置関係のために、25% ずつに分割した四分位数 (quantile)、10% ずつ 10 個に分けた値は、十分位数 (decitile)。また標本サイズが大きい場合には、100 等分に分けた百分位数 (percentile) など。さらに範囲 (range) とは最大から最小までをいい、四分位範囲 (quantile range) とは 3/4(75%) から 1/4(25%) までの範囲をさす。左右の対称性を調べるための歪度（わいど,skewness)、とがり具合には正規分布を基準にしてこれよりも山の頂上が尖っているか平らなのかをみるための尖度 (せんど,kurtosis) が基本統計量として表計算ソフトなどで簡単に求められる。

4.4 多変量データ

いままでは 1 種類のデータ（ひとつの測定値）についてのデータ分析を行ってきた。対象とした個体には同時に測定することで、幾つかのデータの組を得ることができる。ここでは 2 種類の変数値があるとして、この両者にはどのような傾向、関係があるか考える。

4.4.1 **相関図**

例として体重と身長の値を組としてデータを得たとする。。

出席番号	1	2	3	4	5	6	7	8	9	10
体重（kg）	60.3	57.9	65.4	56.1	53.6	62.7	70.0	55.8	67.1	63.1
身長（cm）	161.2	154.3	162.8	160.4	155.7	163.5	172.5	166.4	173.2	164.0

例えば、上の資料 7 の体重を x（kg)、身長を y（cm）として、点を座標平面上にとったとする。このよう

に、2 つの変量からなる資料を平面上に図示したものを相関図または散布図という。以下は資料 7 の相関図である。また、点の付近にある数字はその数値に該当する人の出席番号を表す。

一般に、相関図において、2 つのデータの一方が増えるとき、もう一方も増える傾向にある場合、正の相関関係があるという。2 つのデータの一方が増えるとき、もう一方が減る傾向にある場合、負の相関関係があるという。2 つのデータの間に、正の相関関係も負の相関関係もない場合、相関関係はないという。

4.4.2 相関係数

2 つのデータ x , y について、n 個の値の組 $\{(x_i, y_i); i = 1, 2, \cdots, n\}$ を考える。変量 x の平均を $\overline{X}$, 変量 y の平均を $\overline{Y}$ とし、また、x の分散を $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2$、$y$ の分散を $s_y^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2$ とする。また $s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{XY}$ を x と y の共分散とする。

分散は 2 乗をしているから、必ず非負となるが、共分散は積の値がプラスとマイナスの両方がありえる。共分散が正のときは、$(x_i - \overline{X})$ と $(y_i - \overline{Y})$ が同符号で多くて、積が正となるものが、負よりも多い。よって、共分散が正のとき、x と y には正の相関関係がある。共分散が負のときは、それぞれの項が異符号で負のほうが正よりも多い。よって、共分散が負のとき、x と y には負の相関関係がある。

共分散の値は、資料 (x, y) の内容によって大きく値が変わるので、x , y の偏差をそれぞれの標準偏差 s_x,s_y で割った値の積の平均値；つまり標準化した値 $\left(\dfrac{x_i - \overline{X}}{s_X}, \dfrac{y_i - \overline{Y}}{s_Y}\right), i = 1, 2, \cdots, n$ の共分散を考え、この値を資料 x , y の相関係数といい、r で表す。

定義 4.1 x の平均値を、y の平均値をとすると、相関係数 r は

$$r = \frac{s_{XY}}{s_X s_Y} \tag{4.2}$$

相関係数 r は、一般に $|r| \leq 1$ が成り立つ。相関係数 r の値が 1 に近いほど、傾き正の直線状に並んでいて、このとき正の相関が強いという。相関図の点は右上がりに分布する。相関係数 r の値が -1 に近いほど、傾き負の直線上に並ぶが、負の相関が強いという。このとき、相関図の点は右下がりに分布する。相関係数 r の値が 0 に近いときは、相関は弱いという。ではこれを用いて資料の相関関係を求めて散布図に表す。ここで体重と身長のデータから、それぞれの平均、分散 (標準偏差）をもとめて、偏差値；データから平均を引いて、標準偏差で割った値を 10 倍して、５０を加えた値が体重偏差、身長偏差とおいたものである。このような変換をおこなって共分散は変わるが、相関係数は変わらない。

出席番号	1	2	3	4	5	6	7	8	9	10
体重（kg）	60.3	57.9	65.4	56.1	53.6	62.7	70.0	55.8	67.1	63.1
体重偏差	48.24	43.53	58.23	40.01	35.11	52.94	67.25	39.42	61.56	53.72
身長（cm）	161.2	154.3	162.8	160.4	155.7	163.5	172.5	166.4	173.2	164.0
身長偏差	46.25	34.50	48.98	44.89	36.88	50.17	65.50	55.11	66.70	51.02

よって相関係数 r は

$$= CORREL(\text{配列}1, \text{配列}2) = CORREL(\text{体重}, \text{身長}) = CORREL(\text{体重偏差}, \text{身長偏差}) = 0.756$$

となり、この 10 人の身長と体重にはやや強い正の相関関係があることが分かる。

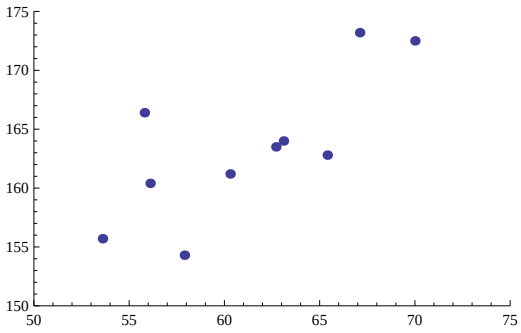


図 4　身長と体重のデータ

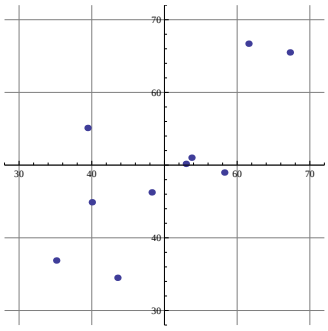


図 5　身長偏差と体重偏差

4.5　不平等の解析、ジニ係数

ジニ係数とは イタリアの経済学者 Corrado Gini が１９１２年に発表した。計算は累積分布関数から求め、公平さと不公平さの違いの程度を表すものと解釈され、0 と 1 の間の値となる。所得の例では変数値の所得金額に対して、その所得金額以下の割合を縦軸に表していく。ジニ係数の値 0 とは、所得金額が公平に配分されている状況を表す。すなわち変数値の金額に対して、割合の増加が直線になる。もう一つの極端な係数値が 1 とは、ある人がすべての所得金額を占めており、その他の人々が所得金額ゼロとなる場合である。このようにジニ係数が低い値であることは、社会の富の配分（所得金額の値）がほぼ平等に分けられていることを示し、数値が高いことは、全体の少数部分が高い富を占有していることとなる。Wikipedia (Gini coefficient) を参照。また 100 倍してパーセント表示にしたものをジニ指数（Gini Index）とよぶ。`http://www.sustainablemiddleclass.com/Gini-Coefficient.html` による近年のジニ指数の値比較：

Japan 24.9	United Kingdom 36.0	Sweden 25.0	Iran 43.0
Germany 28.3	United States 46.6	France 32.7	Argentina 52.2
Pakistan 33.0	Mexico 54.6	Canada 33.1	South Africa 57.8
Switzerland 33.1	Namibia 70.7		

このようなジニ係数 (指数）の結果は、経済学での貧富の指標、公平さの変化を調べるために、社会変化や政治活動のひとつとして公表されている。たとえば、発展国としてヨーロッパ諸国では係数値は 0.24 から 0.36 であるのに対して、アメリカ合衆国では 0.4 を超えている。つまりアメリカとは、貧富の差の大きい国である。また政治哲学や政策の補填の意味でも、このジニ係数は役立つと考えられる。しかし大きな国と小さな国で比較を行なうときには、当然誤解をもたらすことを注意しておかねばならない。世界全体のジニ係数は、およそ 0.56 から 0.66 の間といわれている。Bob Sutcliffe (2007), Postscript to the article ‘ World inequality and globalization ’ (Oxford Review of Economic Policy, Spring 2004), ¡http://siteresources.worldbank.org/INTDECINEQ/Resources/PSBSutcliffe.pdf¿. Retrieved on 2007-12-13. `http://data.worldbank.org/` には The World BANK : Working for a World Free of Poverty として世界中の発展途上国を支援するための財政的かつ技術的なデータが調べられている。

Jonhson、Kots、Balakrishnan,”Continuous Univariate Distribution,Vol.1” からの条件つき期待として計算する。

定義 4.2 非負値の確率変数 X, X_1, X_2 について、独立で同じ分布関数 $F_X(t) = P(X \le t)$、密度関数 $p_X(t)$ をもつとする。また $F_X^{-1}(t) = \inf_x\{x; F_X(x) \ge t\}$ とおく。

1. ジニ平均差 *(Gini mean difference);* $\gamma(X) = E[|X_1 - X_2|]$
2. ローレンツ曲線 *(Lorenz curve);* $L(p) = \dfrac{1}{E(X)} \displaystyle\int_0^p F_X^{-1}(t)dt$
3. ジニ集中度指数 *(Gini concentration index);* $C(X) = 2\displaystyle\int_0^1 \{p - L(p)\}dp = 1 - 2\int_0^1 \{L(p)\}dp$

定理 4.1 次の命題が得られる。

(1) $g = \binom{n}{2}^{-1} \displaystyle\sum_{i<j} |X_i - X_j|$ は $\gamma(X)$ の不偏推定量 *($i < j$ には等号を含まない 2 重和)*

(2) $L(F_X(x)) = \dfrac{E[X \mid X \le x]\, F_X(x)}{E[X]} \le F_X(x)$

(3) $L(p) \le p$, $(0 \le p \le 1)$, $\quad L(0) = 0, L(1) = 1.$

(4) $C(X) = 1 - \dfrac{E[X_1 \mid X_1 \le X_2]}{E[X]} = \dfrac{1}{2}\dfrac{\gamma(X)}{E[X]}$

(5) $\gamma(X) = E[X_1 \mid X_1 \ge X_2] - E[X_1 \mid X_1 \le X_2]$

(証明)　いくつかの証明を与える。非負値で、同一分布、独立、連続型分布であるから、$E[X \mid X \le x]F_X(x) = E[X\,;\, X \le x]$, $P(X_1 \le X_2) = P(X_1 \ge X_2) = \frac{1}{2}$ である。さらに条件付き期待値の性質から、

$$\begin{aligned} E[X] &= E[X \mid X_1 \le X_2]P(X_1 \le X_2) + E[X \mid X_1 \ge X_2]P(X_1 \ge X_2) \\ &= \tfrac{1}{2}\{E[X \mid X_1 \le X_2] + E[X \mid X_1 \ge X_2]\} \end{aligned}$$

が成り立っている。したがって、絶対値を場合に分けて外すと

$$\begin{aligned} &\gamma(X) = E[|X_1 - X_2|] \\ &= E[X_1 - X_2 \mid X_1 \ge X_2]P(X_1 \ge X_2) + E[-(X_1 - X_2) \mid X_1 \le X_2]P(X_1 \le X_2) \\ &= \tfrac{1}{2}E[X_1 \mid X_1 \ge X_2] + \tfrac{1}{2}E[X_2 \mid X_1 \ge X_2] - \tfrac{1}{2}E[X_1 \mid X_1 \le X_2] - \tfrac{1}{2}E[X_2 \mid X_1 \ge X_2] \\ &= E[X_1 \mid X_1 \ge X_2] - E[X_1 \mid X_1 \le X_2] \end{aligned}$$

である。また変数変換 $x = F_X^{-1}(t)$, $p_X(x)dx = dt$, $y = F_X^{-1}(p)$, $p = F_X(y)$, $dp = p_X(y)dy$ であるから、

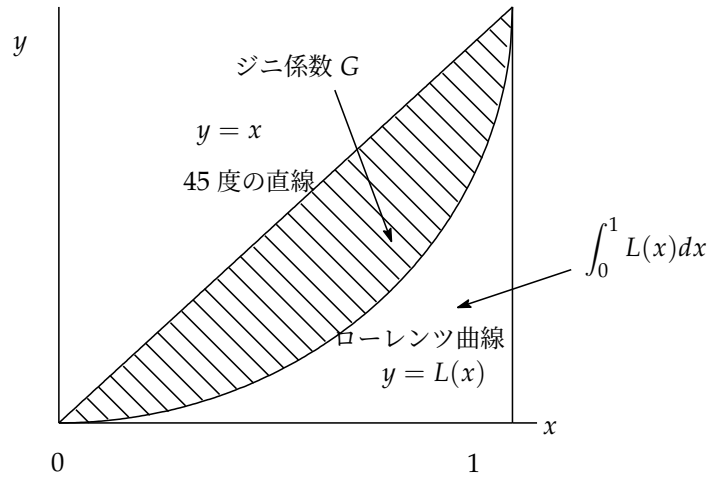
$$\begin{aligned} \int_0^1 L(p)\,dp &= \frac{1}{E(X)}\int_0^1 dp \int_0^p F_X^{-1}(t)\,dt = \frac{1}{E(X)}\int_0^\infty p_X(y)\,dy \int_0^{F_X(y)} F_X^{-1}(t)\,dt \\ &= \frac{1}{E(X)}\int_0^\infty p_X(y)\,dy \int_0^y x\,p_X(x)\,dx = \frac{1}{E[X]}\int_0^\infty E[X\,;\, X \le y]\,dF_X(y) \\ &= \frac{1}{E[X]}\int_0^\infty E[X \mid X \le y]\,F_X(y)\,dF_X(y) = \frac{1}{E[X]}\int_0^\infty \left\{\int_0^x t\,p_X(t)\,dt\right\} p_X(x)dx \\ &= \frac{1}{E[X]}E[X_1 \mid X_1 \le X_2]\,P(X_1 \le X_2) \end{aligned}$$

ここでは非負値、連続型で同一分布、独立性を用いて計算した。また式 (4) は $E[X_1 \mid X_1 \le X_2] = E[X] - \frac{1}{2}\gamma(X)$ から求められる。 □

ジニ係数の計算計算式：

$$G = 1 - 2\int_0^1 L(x)dx = \frac{1/2 - \int_0^1 L(x)dx}{1/2} \quad \text{or} \quad = 1 - 2\sum_i L(x_i) \;\; (i.e. L(x) \text{ の面積})$$

ここで $L(x)$ はローレンツ曲線で、



つぎで定める：

(1) 離散型分布のとき；$i = 1, 2, \cdots, n$

値	x_i	x_1	x_2	$\cdots$	x_n
確率	$p_i = f(x_i)$	p_1	p_2	$\cdots$	p_n
累積確率	F_i	$F_1 = p_1$	$F_2 = p_2 + F_1$	$\cdots$	$F_n = p_n + F_{n-1} = 1$

$$L_i = \sum_{j=1}^{i} x_j f(x_j)/L = \sum_i \frac{L_{i+1} + L_i}{2}(F_{i+1} - F_i) \quad (i = 1, 2, \cdots, n)$$
ただし $L = \sum_{j=1}^{n} x_j f(x_j)$

性質：$L_0 = 0 \leq L_1 \leq L_2 \leq \cdots \leq L_{n-1} \leq L_n = 1$ 増加関数で、下にとつの形をし、0 から 1 まで変化する。

(2) 連続型分布のとき；$-\infty < x < \infty$

値	x
確率	$f(x)$
累積確率	$F(x) = \int_{-\infty}^{x} f(t)dt$

$$L(x) = \int_{-\infty}^{x} tf(t)dt \Big/ L \quad (0 \leq x \leq 1) \text{ ただし } L = \int_{-\infty}^{\infty} tf(t)dt$$

性質：

(i) $L(0) = 0 \leq L(x) \leq \cdots \leq L(y) \leq L(1) = 1, 0 < x < y < 1$

(ii) $0 < L(x) \leq x, 0 < x < 1$

例題：10 人の所得金額を調べると、つぎのデータを得た。この集団におけるジニ係数を計算せよ。

所得	50	100	200	計
人数	4	4	2	10

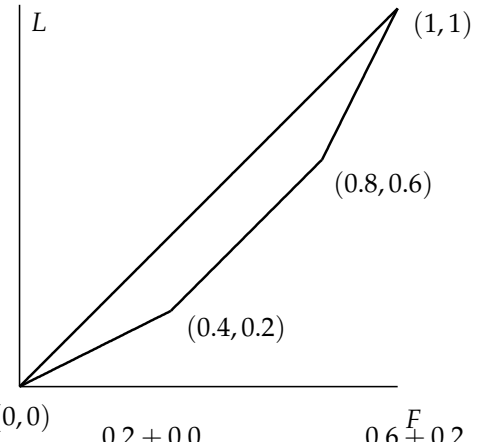
（解）累積相対度数を計算すると、

$n = 3$	$i = 1$	$i = 2$	$i = 3$	
所得×人数	50×4 = 200	100×4=400	200×2 =400	合計 1000
上記の相対値 (L_i)	200/1000 = 0.2	400/1000 = 0.4	400/1000 =0.4	
累積値	0.2	0.2+0.4= 0.6	0.2+0.4+0.4 =1.0	
人数	4	4	2	10
相対累積値 (F_i)	4/10=0.4	0.4 + 4/10 =0.8	0.8+ 2/10=1.0	

人数の累積相対値と所得計の累積相対値の組 $(F_i, L_i) i = 0, 1, 2, \cdots, n$ ただし $(F_0, L_0) = (0, 0), (F_n, L_n) = (1, 1)$ が得られ、これがローレンツ曲線であり、この横軸 (F_i) に対する縦軸 (L_i) の値について、折れ線グラフの面積を

$$\sum_i L(F_i)\Delta F_i = \sum_{i=0}^{n-1} \frac{L_{i+1} + L_i}{2}(F_{i+1} - F_i)$$

で求める。



上記の数字に当てはめると $\sum_i L(F_i)\Delta F_i = \frac{0.2 + 0.0}{2}(0.4 - 0.0) + \frac{0.6 + 0.2}{2}(0.8 - 0.4) + \frac{1.0 + 0.6}{2}(1.0 - 0.8) = 0.04 + 0.16 + 0.16 = 0.36$ がローレンツ曲線の積分に相当する。したがって $G = 1 - 2\sum_i L(F_i)\Delta F_i = 1 - 2 \times 0.36 = 0.28$ (終わり)

参考：Frank Cowell,"Gini,Deprivation and Complaints" 2005, Discussion Paper. 個人的な欠乏に起因する集約測度 (aggregation) の構成。`http://sticerd.lse.ac.uk/DARP/`

ジニ係数 (Gini coefficient, Gini's coefficient) は、主に社会における所得分配の不平等さを測る指標。ローレンツ曲線をもとに、1936 年、イタリアの統計学者コッラド・ジニによって考案されたといわれる。所得分配の不平等さ以外にも、富の偏在性やエネルギー消費における不平等さなどに応用される。係数の範囲は 0 から 1 で、係数の値が 0 に近いほど格差が少ない状態で、1 に近いほど格差が大きい状態であることを意味する。ちなみに、0 のときには完全な「平等」つまり皆同じ所得を得ている状態を示す。（参考：豊田敬「不平等解析 — ジニ係数と変動係数」経営志林第 41 巻 4 号 p.131-135）

ただし留意点としてつぎを注意する。（参考：内閣府、平成 18 年 7 月年次経済財政報告）ジニ係数は不平等さを客観的に分析・比較する際の代表的な指標の一つとなっているが、以下の点には留意する必要がある。同じジニ係数で示される状態であっても、ローレンツ曲線の元の形が著しく違えば、実感として感じる不平等さがまったく変わってくる可能性がある。税金や社会福祉などによって再分配機能が充実した国の場合、初期所得（税引き前の給与）でのジニ係数と、所得再配分後のジニ係数が異なる。調査対象に特定の傾向がある場合は、1 に近いからといって必ずしも不平等が悪いことになるとは限らない。例えば、ある高級住宅地に年収 10 億円の人が 99 人、年収 1 兆円の大富豪が 1 人いるとする。そこでこの高級住宅地に住む 100 人を対象にジニ係数を計算すると約 0.91 となり、非常に格差が大きいが、年収 10 億円でもかなりの高収入であり、この状態が悪いとは一概に言えない。

また富の再分配（とみのさいぶんぱい）または所得再分配（しょとくさいぶんぱい）とは、所得を公平に配分するため、租税制度や社会保障制度、公共事業などを通じて一経済主体から別の経済主体へ所得を移転させることをいう。ジニ係数を使って日本の所得分配の不平等度を計測している統計には、厚生労働省が実施して

いる所得再分配調査がある。このほかにも、家計の所得・支出を調査している家計調査や全国消費実態調査のデータを使って、ジニ係数が計算されている。ジニ係数を計算するためには、個々の家計の所得を使ってローレンツ曲線を描く必要があるが、家計調査や全国消費実態調査などでは、ジニ係数の計算に利用できる公表データが、所得金額ごとや所得金額によって全体を 5 分割ないし 10 分割した世帯の平均値であったりする。こうした階層ごとの平均値を使って求めたジニ係数の近似値は、擬ジニ係数と呼ばれることがある。

日本の所得ジニ係数の推移；所得再分配とジニ係数 (等価再分配所得を基に)

`http://wkp.fresheye.com/wikipedia/`**ジニ係数**

厚生労働省の平成 17 年度所得再分配調査の結果から計算したジニ係数の 1993 - 2005 年までの推移である。「直接税、社会保障給付金、現物支給」の再分配を考慮した所得のジニ係数、「社会保障給付金、現物支給」の再分配を考慮した所得のジニ係数、当初所得のジニ係数、を示している。世帯人員数を考慮に入れた補正を行っている。

なおこの所得再配分調査は、当初所得に公的年金が含まれていないため、他の調査よりもジニ係数が高くなる。公的年金を計算に入れた国民生活基礎調査の結果に基づいて計算すると、ジニ係数は 0.1 ほど小さくなる。また、単身者世帯を調査対象に含まない全国消費実態調査に基づいて計算したジニ係数は、0.2 ほど小さくなる。このように、ジニ係数は所得の定義や世帯人員数への依存度が大きいので注意が必要である。

上記所得再分配調査の結果に寄れば、日本のジニ係数は、当初の高齢化によるとされる急激な上昇分を社会保障の再分配によってほとんど吸収してるが、十分ではなく税による再分配が弱まっているために、ジニ係数の上昇を早めている。原因として、中間所得層に対する税率が OECD 各国に比べて低すぎること、労働年齢層に対する社会保障が少ないことが明らかにされ、養育に対する支援も少ないことで子育て世帯の貧困率を高めている可能性があることが指摘されている。

4.5.1 回帰分析

回帰分析 (かいきぶんせき：regression analysis) について説明します。1 個体に複数の変数を対象として、データが観測されるとき、このようなデータの解析を多変量解析という。その一つに回帰分析が知られている。回帰分析は、因果関係が想像される 2 つの変数間の関係を調べるために用いられる。たとえば、ある現象に対して、起因と考える原因とその結果が一例であり、特に原因となる数値と結果となる数値の関連性を統計的手法により調べる。回帰分析は、多くの分野で応用され、予測や異常値の発見などに用いられる。回帰分析では、原因となる数値（説明変数）と結果となる数値（目的変数）との関係式を求め、目的変数を予測したり説明変数の影響の大きさを評価したりする分析手法のことをいい、要因分析などに用いられます。また目的変数とは従属変数とも呼ばれ、”結果”としてとらえる変数のことであり、要因から影響を受ける変数のことをいいます。一般的には出力特性値などが目的変数にあたります。説明変数とは独立変数とも呼ばれ、目的変数に影響を与える変数のことをいいます。説明変数が 1 つの場合を**単回帰分析**、2 つ以上の場合を**重回帰分析**といい、得られた多項式の各項の係数を偏回帰係数と呼びます。

現在の統計数理学では、単回帰分析の解析には、「被説明変数の平均値と、個々の被説明変数との差の 2 乗」の総和が最小になるような近似直線を求めます。線形モデルとよばれる $y = ax + b$ という形の一次式、すなわち回帰式を考え、観測データからの説明変数と目的変数の関係をこの回帰式で表し、目的変数が説明変数によってどの程度説明できるかを定量的に分析する。Y の X への回帰式 (regression line of Y on X) とは、$y = ax + b$（x:説明変数、y:目的変数）で表される。X の一次関数として、変数 X が与えたときの Y の条件付き平均、あるいは中央値を考えるものである。

一般に回帰とはもとの位置または状態に戻ることをいうが、元来、生物データから見出された現象であり、

その最初はフランシス・ゴルトンにより 1877 年に発表された種子の重量に関する結果である。ゴルトンは 7 組のスイートピーの種子（種子の重量は組により異なるが、組の中では同じにした）を栽培し比較したところ、以下のことを見出した：(1) 子世代の種子重量は親世代と同じく正規分布に従い、また子世代種子の平均直径を親の平均直径に対してプロットすると直線に近い関係がある（現在でいう線形回帰が適用できる）。(2) しかし、子の平均直径は親の直径と比較すると、より全体の平均直径に近づく傾向がある（回帰）。彼は初めこの直線の勾配を「復帰係数 coefficient of reversion」と呼んだ（いわゆる先祖帰りのような生物的現象と考えた）。その後この効果は生物的なものでなくデータの扱いの結果であることを発見し、その名を「回帰係数 coefficient of regression」と変更した。この結果は「有利な形質をもつ個体が生存して子孫を残し、代を重ねるごとにその形質は顕著になる」という当時の進化に関する考えと矛盾するように見えて注目された。実際にはこの種子の大きさは遺伝による部分より偶然的変動が大きかったということである。彼はさらに研究を重ね、1888 年に「相関 co-relation」という言葉を使い、これを表す定数（相関係数）に"r" という字を用いた。また、このような研究をヒトにも適用し、たとえば様々な分野の天才を調べ、彼らの子はほとんど常に親より平均に近くなることを見出した。さらに定量的で客観的な方法として、父親と息子の身長を比較し、やはり特別に高身長の父親でも、特別に低身長の父親でも、息子たちの身長は父親たちの身長より平均に近くなることを見出した。このように元来の意味での「回帰」は、むしろ「相関が低い」ことを表しているのである。

回帰分析は、予測・要因分析等に用いられる。例えば、過去の生産量と製造費用のデータから回帰式を求め、将来の生産量に対する製造費用の予測に活用される。この場合、生産量 (x) に対する製造費用 (y) の過去のデータから回帰式を推定する。生産量 (x) 1 単位当りの製造費用 (y) がどれ程増加するかを示す傾き a は変動費、切片 b は固定費となる。将来の生産量 (x_1) を回帰式に代入すると将来の製造費用 (y_1) が導かれる。また心理学やマーケティングでは、共分散構造分析という重回帰より複雑な関係を適切に説明できるモデルが構築され、普及している。

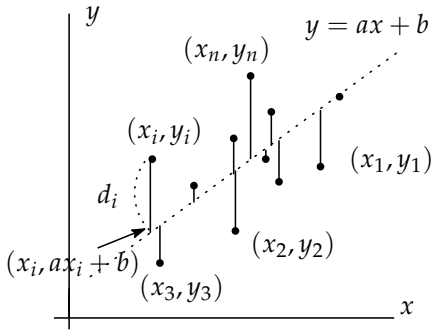
回帰式 $y = ax + b$ を求めるのに変数 a、切片 b を推定する必要がある。この推定には最小二乗法を用いる。回帰直線のあてはめ (最小二乗法) Fitting the Regression Line である。最小二乗法は、n 個の観察された各点 $(x_i, y_i) : i = 1, 2, \cdots, n$ と回帰線上の各点 $(x_i, y) : y = ax_i + b$ との残差 $d_i = y_i - y = y_i - (ax_i + b) : i = 1, 2, \cdots, n$ の平方和 $\sum_i d_i^2$ が最小となる直線を求める方法である。

予測をする際には、回帰式の精度の良さの尺度となる決定係数（0〜1 の値）が 1 に近い（当てはまりが良い）のが望ましい。

また回帰分析は、因果関係が想像される 2 つの変数の関係を調べるのに用いられるが、回帰式は、ある変数が増加（減少）すれば、もう一方の変数が増加（減少）するという関係性を示しているだけで、変数間に因果関係が本当に存在するかは注意して判断しなければならない。

元来は、生物の親と子供の間での属性の対応関係を示す直線、しかもその傾きが 1 より小さいことに大きな意味を持たせている概念であった。歴史的な変遷は後述する。が、現在では本来の Galton, F. (1886) の提唱とは無関係に、 2 変数 x, y 間の対応関係を示す直線

$$y = ax + b$$



を回帰直線、係数 a を回帰係数と呼んでいる。一連の n 組の測定値

$$(x_1, y_1), (x_2, y_2), \cdots (x_i, y_i), \cdots, (x_n, y_n)$$

に最適な直線 (回帰直線) $y = ax + b$ の係数 (回帰係数) a と b を決定するためには、通常、最小 2 乗法が用いられる。x と y とのずれを残差とよび、$y_i - (ax_i + b) = \delta_i, i = 1, 2, \cdots$ とおくと、残差の 2 乗和は

$$\sum_i \delta_i^2 = \sum_i (y_i - ax_i - b)^2$$

と表現できる。δ_i を残差、δ_i^2 を残差の 2 乗、$\sum_i \delta_i^2$ を残差の 2 乗和という。 係数 a と b をうまく定めなければ、残差の 2 乗和はいくらでも大きくできるから、最大値は存在しないが、係数 a と b を適切に決めれば、残差の 2 乗和 $\sum_i \delta_i^2$ は小さくできる。もし、すべてのデータが当てはめた直線上に並ぶならば、残差の 2 乗和は 0 となる。しかし一般には直線上には並んでいないから、この 2 つの値を変化させて、最小になるように定める。つまりその変化率を調べることになる。そのためには微分を行う。したがって最適な係数 a と b を決定することは、残差の 2 乗和を 2 変数関数として微分する。したがって偏微分を施して、その変化が 0 になるような値を考える。

$$\frac{\partial}{\partial a} \sum (y_i - ax_i - b)^2 = (-2) \sum (y_i - ax_i - b) x_i = 0$$
$$\frac{\partial}{\partial b} \sum (y_i - ax_i - b)^2 = (-2) \sum (y_i - ax_i - b) = 0$$

を同時に満足する a と b を決定することと同値である。

$$\begin{cases} (\sum x_i^2) a + (\sum x_i) b = \sum x_i y_i \\ (\sum x_i) a + nb = \sum y_i \end{cases} \tag{4.3}$$

となる。これは、a と b に関する 2 元 1 次連立方程式なので、これを解いて、

$$\begin{aligned} a &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sigma_{xy}}{\sigma_x^2} \\ b &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{({\sigma_x}^2 + \overline{x}^2)\,\overline{y} - \overline{x}\,(\sigma_{xy} + \overline{xy})}{{\sigma_x}^2} = \overline{y} - \overline{x} \frac{\sigma_{xy}}{{\sigma_x}^2} \end{aligned} \tag{4.4}$$

が得られる。これが 2 乗和を最小にすることがわかる。これを y に関する x への回帰直線といい、

$$y - \overline{y} = \frac{\sigma_{xy}}{{\sigma_x}^2} (x - \overline{x})$$

として得られる。この係数 a と b の値を用いてそれぞれの x_i から y_i を推定し、残差の 2 乗和 $\sum \delta_i^2$ を求めると、

$$\sum \delta_i^2 = \sum (y_i - ax_i - b)^2 = \sum y_i^2 - 2b \sum y_i + nb^2 + a^2 \sum x_i^2 + 2ab \sum x_i - 2a \sum x_i y_i$$

となる。

係数 a と b の 2 乗平均誤差 σ_a と σ_b は、

$$\begin{aligned}\sigma_a^2 &= n/(n\sum x_i^2 - (\sum x_i)^2) \times \sum \delta_i^2/(n-2)\\ \sigma_b^2 &= \sum x_i^2/(n\sum x_i^2 - (\sum x_i)^2) \times \sum \delta_i^2/(n-2)\end{aligned} \tag{4.5}$$

となる。確率誤差 r_a と r_b は、それぞれ、σ_a と σ_b を 0.6745 倍すれば、求められる。y と x の平均値、$\overline{y} = \dfrac{\sum y_i}{n}$ と $\overline{x} = \dfrac{\sum x_i}{n}$ は、得られた回帰直線上の点であるから、$y = ax + b$ を満足する。いいかえると $\overline{y} = a\overline{x} + b$ が成り立つ。また平均は必ずこのような関係を満たすので、y の分散

$$\sigma_y^2 = \frac{1}{n}\sum (y_i - \overline{y})^2$$

が小さいほど、推定した回帰直線はモデルとして優れていると考えられる。回帰直線による推定値の平均は $a\overline{x} + b$ なので、回帰直線による推定値 $\hat{y}_i = ax_i + b, i = 1, 2, \cdots, n$ に関する平均値 $a\overline{x} + b$ の分散は

$$\sigma_{\hat{y}}^2 = \frac{1}{n}\sum (\hat{y}_i - \overline{y})^2 = \frac{1}{n}\sum (ax_i + b - a\overline{x} - b)^2 = \frac{a^2}{n}\sum (x_i - \overline{x})^2 = a^2\sigma_x^2$$

と表現される。推定値の分散とデータの分散の比 $\dfrac{a^2\sum (x_i - \overline{x})^2}{\sum (y_i - \overline{y})^2}$ は、決定係数あるいは寄与率とよばれる。回帰係数 $a = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$ を利用すると、

$$\begin{aligned}\frac{a^2\sum (x_i - \overline{x})^2}{\sum (y_i - \overline{y})^2} &= a^2\frac{n\sum x_i^2 - (\sum x_i)^2}{n\sum y_i^2 - (\sum y_i)^2} = \left(\frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}\right)^2 \frac{n\sum x_i^2 - (\sum x_i)^2}{n\sum y_i^2 - (\sum y_i)^2}\\ &= \frac{(n\sum x_i y_i - \sum x_i \sum y_i)^2}{\left\{n\sum x_i^2 - (\sum x_i)^2\right\}\left\{n\sum y_i^2 - (\sum y_i)^2\right\}} = r^2\end{aligned}$$

となり、推定値の分散とデータの分散の比は、相関係数

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}}\sqrt{\sigma_{yy}}} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$$

の 2 乗に等しい。即ち、相関係数の 2 乗 r^2 は、生のデータのもつ平均値の回りの分散のうち回帰直線による推定値によって説明できる比率を意味している。即ち、回帰直線の説明度を示す際には、相関係数 r よりも相関係数の 2 乗 r^2 の方が適している。 また、回帰直線の係数 a と相関係数 r の間には、上で述べた関係式から、

$$a = \frac{\sigma_y}{\sigma_x} r$$

が成り立つ。

（http://www.econ.uiuc.edu/~roger/research/galton/galton.pdf **より**）

5 表計算ソフト

アンケートなど、資料の数が多い場合には手作業で計算をすると膨大な時間がかかる。そこでコンピューターの表計算ソフト（ここでは Microsoft Excel, OpenOffice calc を例に取る）を用いて統計処理を行ってみよう。

5.1 基本コマンド

表計算ソフトではたくさんの長方形の何も書かれていない枠が並んでいる。この枠のことをセルという。また縦方向（1,2,3,･･･）のことを行、横方向（A,B,C,･･･）のことを列という。セルの個々の呼び方は「横列と縦行」の順に表す。例えば横列が C、縦行が 3 であるセルは「C3 のセル」であるという。

	A	B	C	D	E
1	30	2			
2	20	4			
3	40	6			
4	35	5			
5					

問 5.1 *A2* のセルと *B3* のセルに当たる数値をそれぞれ答えよ。また、「*35*」,「*2*」はそれぞれどのセルに入力されているか。

表計算ソフトでは直接セルに計算式（特に関数と言われる：詳しくは後述。）を入力することによって、指定されたセルに対して計算を行い、その実行結果が計算式を入力したセルに反映される。またそのセルを複写すると複写先のセルに応じた計算式となって入力され、その実行結果が表示される。

[数値計算] ; ここでは数値が入力されたセルに対しての計算方法を学ぶ。表計算ソフトによって計算式の種類や入力方法など異なる場合があるので事前に確認しておくこと。またここではオート SUM は割愛する。

[演算子] ; セルに計算式を入力することで様々な計算ができます。また、その計算に必要な記号のことを演算子 `{+,-,*,/,%,^}`といい、セルに入力されている数値の演算は”=”をはじめに書いて、通常の電卓計算のように数式セルに入力します。

`X1` と `Y1` の和・・・	`=X1+Y1`	`X1` から `Y1` を引いた値・・・	`=X1-Y1`
`X1` と `Y1` の積・・・	`=X1*Y1`	`X1` を `Y1` で割った値・・・	`=X1/Y1`
`X1` を `Y1` で割った余り・・・	`=X1%Y1`	`X1` の `Y1` 乗・・・	`=X1^Y1`

5.2 組み込み関数

数学の関数では、x の値を決めると y の値が 1 つに定めますが、コンピュータにおいての関数は用途別に予め用意された計算式（プログラム）のことを表す。`X1` に入力された数値の演算の代表的な例です。変数に相当する部分 `X1` を引数とよびます。

X1 の正の平方根・・・	=SQRT(X1)	X1 の絶対値・・・	=ABS(X1)
X1 を超えない最大の整数・・・	=INT(X1)	X1 を整数値で四捨五入・・・	=ROUND(X1)
X1 を整数値で切り上げ・・・	=ROUNDUP(X1)	X1 を整数値で切り捨て・・・	=ROUNDDOWN(X1)

セルが"X1・X2・X3・・・Xn"という列または行であるとき、この範囲を引数とする演算があります。行列"X1・X2・X3・・・Xn・Y1・Y2・Y3・・・Yn"でも範囲は (X1:Yn) または (X1;Yn) とします。セミコロン（;）とコロン（:）の違いがあります。コロン“:”は Excel の場合で、セミコロン“;”は Openoffice で命令は共通。

全ての和・・・	=SUM(X1:Xn)	平均値・・・	=AVERAGE(X1:Xn)
中央値・・・	=MEDIANE(X1:Xn)	最頻値・・・	=MODE(X1:Xn)

5.3 集計に用いる関数

与えられた統計データを集計するには、総和をとったり、個数を数え上げます。このためにはつぎの関数を用います。あるいは簡便にできるようプログラムを定義しているものもあります。

=if(**論理式**, **値** 1, **値** 2)	論理式が真 (true) のときに値 1、偽（その他）で値 2; 例：=IF(X1>Y1,"**予算超過**","OK")、 X1 と Y1 との大小を比較して (予算超過) か (OK) を出力
=and(**論理式** 1, **論理式** 2,...)	すべての論理式が成り立つ（真）のとき、値 1 を返し、その他では値 2 とする例：=if(and(**論理式** 1, **論理式** 2,...), **値** 1, **値** 2)
=or(**論理式** 1, **論理式** 2,...)	少なくとも一つの論理式が成り立つ（真）のとき、値 1 を返し、その他では値 2 とする。例：=if(or(**論理式** 1, **論理式** 2,...), **値** 1, **値** 2) 例：=IF(X1>=100,SUM(Y1:Yn),""), セル X1 の数値が 100 で以上ある場合は、セル範囲 Y1:Yn の合計が計算され、それ以外は、空白文字列 (””) が返される
=countif(**範囲**, **検索条件**)	範囲のうち、検索条件に一致するセルの個数を返す。特定の文字で始まるすべてのセルや、指定された数値よりも大きい数値または小さい数値を含むすべてのセルをカウントできます。 例：=COUNTIF(X1:Xn,"**合格**")、範囲セルで ” 合格 ” 者数を求める。複数の検索条件の場合には COUNTIFS をもちいる。

単純集計表（度数表）：集められたデータの属性別に整理して、それらの度数を表にまとめたもの。たとえば、データ属性が {1 年、2 年、3 年} と分かれていればそれぞれの該当数を合計する。このような質的データではなく、量的データとして、体重あるいは身長などであれば最大値と最小値から階級別に分類してから、その級値に入る人数の表が度数分布表である。グラフにしたものがヒストグラム (Histgram) という。度数 (count) は頻度 (frequency) ともよばれる。

countif 関数をもちいて集計できる。

簡便に行うために Frequency 関数ももちいるが、これは出力のために配列数式とする。配列数式を入力するには、数式を含むセルを選択し、Ctrl キーと Shift キーを押しながら Enter キーを押すことに注意する。

クロス集計表 (cross tabulation)：クロス集計とは、与えられたデータのうち、2 つないし 3 つ程度の項目に着目してデータの分析や集計を行なう。1 つ (ないし 2 つ) の項目を縦軸に、もう 1 つの項目を横軸において表を作成して集計を行なう。手作業を用いながら行うには、countifs

MS エクセルでは、pivot テーブルをつかう。以前のバージョンの Microsoft Office Excel では、[データ] メニューに◆ [ピボットテーブルとピボットグラフ レポート] コマンドがあり、ピボットテーブル/ピボットグラフ ウィザードを起動できました。Microsoft Office Excel 2007 では、[ピボットテーブルとピボットグラフ レポート] コマンドが次の 2 つのコマンドに分割されています。◆ [ピボットテーブル] コマンド。[ピボットテーブルの作成] ダイアログ ボックスが表示されます。◆ [ピボットグラフ] コマンド。[ピボットグラフ付きピボットテーブルの作成] ダイアログ ボックスが表示されます。

集計結果の例：階級分けした集計

階級	度数	正字の計数	Tally mark	簡易出力
100 – 120	5	正	𝍸	*****
121 – 140	3	下	𝍷	***
141 – 160	16	正正正	𝍸 𝍸 𝍸	**********
161 – 180	14	正正下	𝍸 𝍸 𝍷	**********
181 – 200	8	正下	𝍸 𝍷	********
201 – 220	11	正正丁	𝍸 𝍸 𝍵	**********
221 – 240	3	下	𝍷	***
合計	60			

単純集計表の例：学年別人数

学年	1 年	2 年	3 年	合計
人数	93	79	119	291
割合 (%)	31.9	27.1	40.9	100

クロス集計 1:学年別と文系/理系

学年	1 年	2 年	3 年	合計
文系	45	42	55	142
理系	48	37	64	149
人数	93	79	119	291

クロス集計 2:学年別、男女別数の集計

	男			女			合計
学年	1 年	2 年	3 年	1 年	2 年	3 年	計
文系	23	20	27	22	22	28	142
理系	23	24	27	25	13	37	149
人数	46	44	54	47	35	65	291

★ピボットテーブルで度数分布表をつくる。

[問題] 図のようにセル範囲 A1:A10 に総得点として 9 個のデータ;{ 150, 200, 250, 320, 330, 360, 380, 420, 480} があります。このデータについて、以下の (1) から (4) がそれぞれ何件あるかをピボットテーブルで集計するにはどうすればいいでしょうか？

	A
1	総得点
2	150
3	200
⋮	⋮
9	420
10	480

階級分け:
(1) 200 未満、
(2) 200 以上 300 未満
(3) 300 以上 400 未満
(4) 400 以上 500 未満

データの個数 / 総得点

総得点	集計
＜ 200	1
200-299	2
300-399	4
400-500	2
総計	9

[手順の説明]

1. リスト内（A1:A10）のセルのどれかを選択
2. メニュー［**データ**］-［**ピボットテーブルとピボットグラフレポート**］
3. ［Excel **のリスト／データベース**］と［**ピボットテーブル**］にチェックを確認、［**次へ**］ボタンをクリック
4. ［**範囲**］ボックスに $A$1:$A$10 が入力されていることを確認して［**次へ**］ボタンをクリック
5. ［**レイアウト**］ボタンをクリック
6. ［**行**］に 総得点 をドラッグ、［**データ**］にも 総得点 をドラッグ
7. ［**データ**］にドラッグした データの個数：総得点 をダブルクリック
8. ［**集計の方法**］で データの個数 をダブルクリック
9. ［OK］ボタンをクリック
10. ［**既存のワークシート**］をクリック。例えばセル D1 をクリック。
11. ［**完了**］ボタンをクリック →これで、いったんピボットテーブルが完成
12. セル D2 上で右クリック-［**グループ化**］
13. ［**先頭の値**］を 200 に
14. ［**末尾の値**］を 500 に
15. ［**単位**］が 100 になっていることを確認して［OK］ボタンをクリック

★ frequency 関数をもちいる方法

[手順の説明] (1) 予めデータの最大値=max{**範囲**}と最小値=min{**範囲**}をもとめて (2) データの個数=count{**範囲**}から階級の幅と (3) 各階級の始まりと終わりを求める (4) これを、たとえばセルの範囲を D13 から D17 までに入力（［**連続データの作成**］）(5) 度数 (集計）を記入する E13 から E17 までのセルをアクティブに (6) 数式バーに=frequecy と書き (7) もし組み込み関数 FREQUENCY(**データ配列, 区間配列**）については、詳しく知りたいときに help 命令で検索して参照します。(8) データ配列が A2 から A10 とし、(9) 区間配列には D13 から D17 までとする (10) ここで単に OK をクリックするのではなく、数式配列の入力 CTRL+SHITFT を押しながら OK をクリックする。

5.4 グラフの作成

グラフの作成は [挿入][グラフ] と進んで、グラフウィザードに沿っていけば大まかなグラフは作れます。

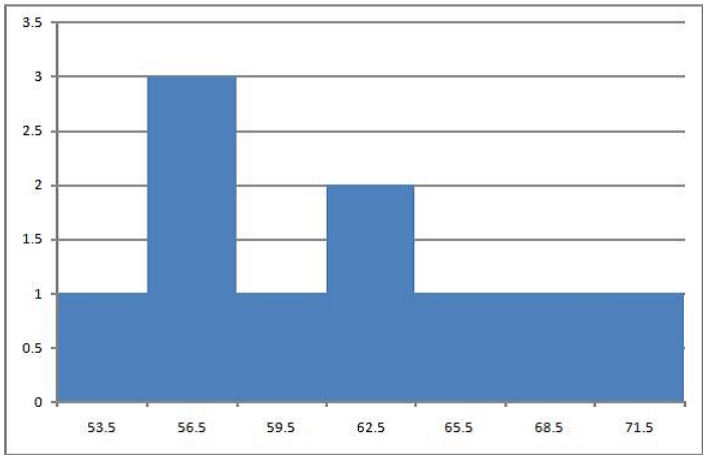
以下の表は資料（左側データ）を表計算ソフトに入力して出力したグラフ、ヒストグラム（右図）である。ただし階級は、52.0kg 以上 55.0kg 未満の階級のことを 52.0-55.0 などと表すことにする。ここで注意することは、棒グラフとヒストグラムの違いで、棒グラフは、質的データや離散変量などを表すために使いますが、例

として挙げているデータは連続変量で、項目間の幅に意味をもちます。ですから取りあえず、グラフウィザードで棒グラフを選んでおいてから、変量間の間隔をゼロに設定します。具体的にはつぎのように行います。

(1) 縦軸に描く系列データ (度数値) を選んで、この範囲のセルをアクティブにする
(2) 挿入 - グラフ選択 - 縦棒 - ２ D 縦棒, とすると大まかなグラフができる
(3) 横軸（項目データ）には指定をしていないので、1,2,3, などと出るのでこれを修正します
(4) グラフエリアで「横（項目）軸データ」(1,2,3, などの部分）をクリック、「データソースの選択」ウィザードで「編集」、「グラフデータの範囲」を階級値として選んでから,「OK」で入力される。ここまでで一応の棒グラフとして完成している
(5) ヒストグラムにするために横軸の幅を調整する。描かれている棒（系列１）にマウスを移動して右クリックする
(6) 「データ系列の書式設定」で「系列のオプション」「要素の間隔」を「なし (0%)」にスライドさせると、間隔がなくなり、ヒストグラムとして完成。

もし描いたグラフを他の利用するためにファイルとして保存するためには、まずグラフエリアをクリックし、「クリップボード」、「貼り付け」、「図」、「図としてコピー」「図のコピー」にて OK とすると、準備完了つぎに「スタート」「すべてのプログラム」「アクセサリー」と進んで「ペイント」を起動して、枠の右下側にて範囲を設定してから、「編集」「貼り付け」とすれば、表示される。それから「名前を付けて保存」とすれば、JPEG ファイルなどとして保存される。

	A	B	C
1	階級	階級値	度数
2	52.0-55.0	53.5	1
3	55.0-58.0	56.5	3
4	58.0-61.0	59.5	1
5	61.0-64.0	62.5	2
6	64.0-67.0	65.5	1
7	67.0-70.0	68.5	1
8	70.0-73.0	71.5	1



問 5.2 表計算ソフトに上記の数値を入力してみよ。また、グラフ作成機能を用いてヒストグラムと度数折れ線を作成してみよ。上のほうの「資料とグラフ」に挙げたようなグラフになるはずである。

度数折れ線は左右両端に度数が 0 である階級があるものとして作図をすると前に述べた。故にこのグラフを表計算ソフトで作成する場合は表 2 の 2 行の前の行に階級値が 50.5 であるもの、8 行の後の行に階級値が 74.5 であるもの（それぞれ度数は 0）を事前に作っておかなければならない。

組み込み関数を組み合わせて、**平均・分散・標準偏差** を計算する。

1 表計算ソフトに上記の表を作成し、D 列にそれぞれの「階級値×度数」を求める式を入力せよ。例えば D2 のセルの値は B2 のセルの値と C2 のセルの値を掛け合わせた数値なので `= B2 * C2` と入力。
2 実習 3 の結果から B11 のセルに平均値を求める式を SUM を使った式で入力せよ。

3　E 列にそれぞれの偏差を求める式を入力せよ。例えば E2 のセルの値は B2 のセルの値から B11 のセルの値を引いた数値なので = B2 - B11 と入力される。

4　F 列にそれぞれ偏差の 2 乗を入力した後、G9 のセルに「偏差の 2 乗×度数」の合計を求める式を入力せよ。また、F2 のセルに= (B2 - $B$11)^2 * C2 と入力し F3～F8 のセルに複写することもできる。「$」については複写時に $ にはさまれている場合は複写しても引用する部分が変わらないことを示す。

5　実習 6 より B12 のセルに分散、B13 のセルに標準偏差をそれぞれ表示させてみよ。

全ての空欄を埋めた表は以下の通りになる。

	A	B	C	D	E	F	G
1	階級	階級値	度数	階級値×度数	偏差	偏差の 2 乗	偏差の 2 乗×度数
2	52.0-55.0	53.5	1	53.5	-7.8	60.84	60.84
3	55.0-58.0	56.5	3	169.5	-4.8	23.04	69.12
4	58.0-61.0	59.5	1	59.5	-1.8	3.24	3.24
5	61.0-64.0	62.5	2	125.0	1.2	1.44	2.88
6	64.0-67.0	65.5	1	65.5	4.2	17.64	17.64
7	67.0-70.0	68.5	1	68.5	7.2	51.84	51.84
8	70.0-73.0	71.5	1	71.5	10.2	104.04	104.04
9	合計		10				309.6
10							
11	平均値	61.3					
12	分散	30.96					
13	標準偏差	5.564					

相関係数

以下の表 4 は資料 7 を表にしたものである。ここでは今まで学んだことを用いて全ての空欄を埋めて欲しい。13 行は表の見やすさのために空けてある。いくつかのセルは結合されているがその手順を以下に示す。以下の例では A1・A2 のセルを結合させる場合を考える。A1 のセルから A2 のセルに向けてドラッグ（逆方向にドラッグしてもよい）し、2 つのセルを選択させた状態にする。選択された範囲内で右クリックし、「セルの書式設定＞配置＞文字の制御」の「セルの結合」の部分にチェックマークを入れる。A1・A2 のセルの間の境界線が無くなり、2 つのセルが結合された状態になる。

表 4

A B C D E F G

1 出席番号体重 身長 2 数値 偏差 偏差の 2 乗 数値 偏差 偏差の 2 乗 3 1 60.3 161.2 4 2 57.9 154.3 5 3 65.4 162.8 6 4 56.1 160.4 7 5 53.6 155.7 8 6 62.7 163.5 9 7 70.0 172.5 10 8 55.8 166.4 11 9 67.1 173.2 12 10 63.1 164.0 13 14 相関係数

全ての空欄を埋めた表は以下の通りである。各々作成した表と見比べ確かめてみるとよい。表 4（完成）

A B C D E F G 1 出席番号体重身長 2 数値 偏差 偏差の 2 乗 数値 偏差 偏差の 2 乗 3 1 60.3 -0.9 0.81 161.2 -2.2 4.84 4 2 57.9 -3.3 10.89 154.3 -9.1 82.81 5 3 65.4 4.2 17.64 162.8 -0.6 0.36 6 4 56.1 -5.1 26.01 160.4 -3 9 7 5 53.6 -7.6 57.76 155.7 -7.7 59.29 8 6 62.7 1.5 2.25 163.5 0.1 0.01 9 7 70.0 8.8 77.44 172.5 9.1 82.81 10 8 55.8 -5.4 29.16 166.4 3 9 11 9 67.1 5.9 34.81 173.2 9.8 96.04 12 10 63.1 1.9 3.61 164.0 0.6 0.36 13 14 相関係数 0.755568

5.5 シミュレーション

擬似乱数（ぎじらんすう、pseudorandom numbers）とは、乱数列（乱数）のように見えるが、実際には一定の規則的な確定的な計算式によって求める数を指す。擬似乱数を生成するアルゴリズムを擬似乱数生成法と呼ぶ。乱数は本来規則性も再現性も無いために予測は不可能だが（例：サイコロを振る時、今までに出た目から次に出る目を予測するのは不可能）、擬似乱数は計算によって作るので、作り方が分かれば理論的には予測可能であり、また内部の初期値（種数シード）が分かれば、先に計算しておくこともできるので完全な乱数とはいえない。何をもって擬似乱数と呼ぶのかは議論があるところだが、暗号理論では擬似乱数をもちいられるが、数の桁サイズが多項式処理時間の計算機では乱数と識別不可能な列出力するものをいう。メルセンヌ・ツイスタ (Mersenne twister) という松本眞と西村拓士によって開発された $2^{19937}-1$ という長い周期となる擬似乱数生成アルゴリズムが有名で、表計算ソフトにも取り入れられるアドインが公開されている。

コイン投げやサイコロ振りを計算機（パソコン）上で実現でき、このような成果のおかげで、ランダム・ウォーク（乱歩、酔歩）などを実際の結果を一目瞭然で理解できる。

一様乱数を基本として、正規分布に従う乱数、正規乱数をつくる。

12 個の一様乱数から、1 個の正規乱数をつくる方法：

定理 5.1 単位区間 $[0,1]$ 上の一様分布 $U_i, i=1,2,\cdots,12$ から、$X=\sum_{i=1}^{12} U_i - 6$ をつくると、$X \sim N(0,1)$ にほぼ等しい。

実際には表計算ソフトで乱数生成命令 $= rand()$ を 12 回繰り返し、このつくった 12 個の乱数を加えて 6 を引けばよい。この理由は中心極限定理によるものです。

6　推測の統計

標本調査・正規分布など自然や社会の仕組みを把握するために必要な統計的方法を学習します。ここでは対象から抽出される標本を確率変数と考え、標本平均・標本標準偏差などの数値を用いて、ある統計的な判断を下せるようにすることが目標です。

6.1　標本調査

[標本の抽出] ; 統計調査には、対象となる集団のすべてを調べる全数調査と、対象となる集団の一部を調べる標本調査がある。標本調査の場合に、調査の対象になるものの全体を母集団、調査のために母集団から取り出されたものを標本、母集団から標本を取り出すことを標本の抽出という。また、母集団に含まれるものの個数を母集団の大きさといい、標本全体が含むもの個数を標本の大きさという。標本調査は、その標本の性質から母集団の性質を推定するのが目的であるから、標本が母集団の性質をよく表すように選ばなければならない。例えば 200 人から 30 人を選ぶとき、かたよりがないように、くじ引きなどを用いて選ぶことがある。このように、かたよりなく取り出すことを無作為抽出といい、そのように抽出された標本を無作為標本という。

標本を抽出するとき、一度抽出した標本をもとに戻してから次の標本を抽出する方法を復元抽出という。これに対して、抽出した標本をもとに戻さずに次の標本を抽出する方法を非復元抽出という。無作為抽出を行うには、乱数さいや乱数表がよく使われる。最近はコンピューターを使って乱数に近い数の列（擬似乱数）をつくらせ、それを使うのが普通になっている。

[統計的手法] ;

記述統計　記述統計とは、収集したデータの要約統計量（平均、分散など）を計算して分布を明らかにする事により、データの示す傾向や性質を知ること。

推測統計　データからその元となっている諸性質を確率論的に推測する分野。推測統計学の項に詳述。

尺度水準　データ（あるいは変数、測定）の尺度はふつう次のような種類（水準）に分類される。尺度水準によって、統計に用いるべき縮約統計量や統計検定法が異なる。**質的データ（カテゴリデータ）**名義尺度：単なる番号で順番の意味はない。電話番号、背番号など。順序尺度：順序が意味を持つ番号。階級や階層など。**量的データ（数値データ）**間隔尺度：順序に加え間隔にも意味がある（単位がある）が、ゼロには絶対的な意味はない。摂氏・華氏温度、知能指数など。比率尺度：ゼロを基準とする絶対的尺度で、間隔だけでなく比率にも意味がある。絶対温度、金額など。

実験計画　データ収集の規模や対象、割付方法をコントロールし、より公正で評価可能なデータが収集できるよう検討すること。統計の世界には Garbage in, garbage out という格言がある。これは「ゴミのようなデータを使っていくら解析しても出てくる結果はゴミばかりだ」という意味であり、データ収集の前にその方法を十分に検討する必要があることを強調したものである。

6.2　標本統計量

標本データとは、母集団から抽出されたデータを意味する。この集まりは、取り出し毎に変動するから、どういう分布をするか調べる必要がある。対象とするデータは母集団から抽出される標本であり、標本から直接算出される統計量は観測（観察）できる確率変数であり、標本の性質を表現する数値である。母集団を母数（未知であり、観測できない）によって特徴づけられる確率分布として仮定し、そこからあるサイズの標本を

ランダムに抽出する。この標本データを解析することで、未知の母集団における母数を推定したり、仮説の検定をおこなう。したがって母集団の分布とそこから抽出した標本の分布との関係を調べ明らかにすることが推定検定の基礎となる。抽出したデータについて、集計計算したものを標本統計量という。たとえば、データの総和を個数で割った算術平均は統計量であり、母集団の平均と深い関係をもつ。このような関係を明らかにすることが統計解析での推測とよばれる。

定理 6.1 母平均 μ、母分散 σ^2、母標準偏差 σ の母集団から復元抽出で無作為に大きさ n の標本 X_i を取り出すとき、統計量 $Z = c_1X_1 + c_2X_2 + \cdots + c_nX_n$ の平均、分散は? ただし $\{c_i\}$ は定数とする。

1. $E(Z) = (c_1 + c_2 + \cdots + c_n)\mu$
2. $V(Z) = (c_1^2 + c_2^2 + \cdots + c_n^2)\sigma^2$

[標本平均の分布と正規分布] ; 一般に、標本平均の分布について、次のことが成り立つ。

定理 6.2 ***[標本平均の分布]:*** 標本平均 $\overline{X}$ の分布;

1. $E(\overline{X}) = \mu$
2. $V(\overline{X}) = \dfrac{\sigma^2}{n}$

中心極限定理から、母平均 μ、母標準偏差 σ の母集団から無作為に抽出した大きさ n の標本平均 $\overline{X}$ の分布は、n が十分大きければ、正規分布に近い。したがって

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

とおくと、Z は近似的に標準正規分布 $N(0,1)$ に従う。また、母集団分布が正規分布の場合には、n の値が何であっても、標本平均の分布は、正規分布となる。

問 6.1 母平均 $\mu = 120$、母標準偏差 $\sigma = 16$ である正規母集団から、大きさ $n = 25$ の標本を無作為に抽出するとき、標本平均 $\overline{X}$ についての確率 *(1)* $P(\overline{X} < 120 + \frac{32}{5})$, *(2)* $P(|\overline{X} - 120| > \frac{32}{5})$ を求めよ。

問 6.2 *(i)* X, Y が独立で正規分布 $N(0, \sigma^2)$ にしたがうとき、$X + Y$ と $X - Y$ の共分散、相関係数を求めよ。
(ii) $\overline{X} = \sum_{i<j} \frac{X_i + X_j}{2}$, $u^2 = \binom{n}{2}^{-1} \sum_{i<j} \frac{(X_i - X_j)^2}{2}$ を示せ。
(iii) 変量 $X_i, i = 1, 2, \cdots, n$ が正規母集団から抽出した無作為標本とするとき、標本平均と標本分散は独立であることを示せ。

6.3 推定

ある母集団において、母平均 m が未知のとき、これを標本調査を通じて得られたデータから推測することを母平均の推定という。母平均 m、母標準偏差 σ(分散 σ^2) の母集団 (正規分布とは限らなくてよい) から、大きさ n の標本 $\{X_1, X_2, \cdots, X_n\}$ を無作為抽出し、その標本平均を $\overline{X} = \dfrac{1}{n}\sum_i X_i$ とする。n が大きいとき、$\overline{X}$ の分布は、中心極限定理によって正規分布 $N(m, \dfrac{\sigma^2}{n})$ に近づくから、これを標準化した $Z = \dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$ は標準正規分布 $N(0,1)$ に近づく。正規分布表を用いると、$P(|Z| \leq c) = P(-c \leq Z \leq c) = 0.9500$ を満たす c の値

は 1.96 である。したがって

$$\{\ |Z| \leq c\ \} = \left\{ |\overline{X} - m| \leq c\frac{\sigma}{\sqrt{n}} \right\} = \left\{ \overline{X} - c\,\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + c\,\frac{\sigma}{\sqrt{n}} \right\}$$

となり、括弧内の式を変形して、未知の m に対する区間、信頼区間 $\left[\overline{X} - c\,\dfrac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + c\,\dfrac{\sigma}{\sqrt{n}} \right]$ が求まる。信頼度 95% の信頼区間は $c = 1.96$ とすればよい。同様に不等式を満たす c の値は 2.58 で 95% であることから、信頼度 99% の信頼区間は、1.96 を 2.58 に変えればよい。

定理 6.3 ***[母平均の推定]:*** 母標準偏差 σ の母集団からとった大きさ n の標本の標本平均が $\overline{X}$ であるとき、母平均 m の信頼区間は信頼度 95% では

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \ \leq\ m\ \leq\ \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

信頼度 99% では

$$\overline{p} - 2.58\frac{\sigma}{\sqrt{n}} \ \leq\ p\ \leq\ \overline{p} + 2.58\frac{\sigma}{\sqrt{n}}$$

母標準偏差 σ の値が既知でないときは、σ の代わりに標本から得られた標本分散 $s^2 = \dfrac{1}{n}\sum_i X_i^2 - \overline{X}^2$ から定まる標準偏差 $s = \sqrt{s^2}$ を用いる。ただし、このとき近似であるから、標本の大きさは $n > 30$ 程度と十分大きくなければならない。

問 6.3 ある県の高校 *1* 年の男子 *1600* 人を無作為に抽出して身長を調べたところ、平均身長が *164cm*、標準偏差が *6cm* であった。この県の高校 *1* 年男子の平均身長 m を、信頼度 *95%* で推定せよ。

母比率の推定: 母集団において、ある性質 A をもつもの、もたないもの区別されるとき、この集団を二項母集団という。全体に対する性質をもつ割合 p を母比率という。母集団から十分大きな標本の大きさ n の標本が得られるとき、2 項分布の正規近似によって推定する。標本のうち性質 A をもつものの個数を X とすると、X は二項分布に従う。よって、X の平均、分散は $m = np$ と $\sigma^2 = np(1-p)$ となる。標本の大きさ n が十分大きいとき、この分布は正規分布に近いので、母平均の推定の考えを用いる。とし、標本割合 $\overline{p} = \dfrac{X}{n}$ とすれば、$Z = \dfrac{X - np}{\sqrt{np(1-p)}} = \dfrac{\overline{p} - p}{\sqrt{p(1-p)/n}} \ \sim\ N(0,1)$ から、

$$\{\ |Z| \leq c\ \} = \left\{ |X - np| \leq c\sqrt{np(1-p)} \right\} = \left\{ \overline{p} - c\sqrt{p(1-p)/n} \leq p \leq \overline{p} + c\sqrt{p(1-p)/n} \right\}$$

となり、括弧内の式において、p に関する式に変形し、$1/n$ を無視して近似的に、

$$\left[\overline{p} - c\,\frac{\overline{p}(1-\overline{p})}{\sqrt{n}} \leq p \leq \overline{p} + c\,\frac{\overline{p}(1-\overline{p})}{\sqrt{n}} \right]$$

が得られる。

定理 6.4 (母比率の推定) 大きさ n の標本の標本比率が $\overline{p}$ のとき、母比率 p の信頼区間は信頼度 95% では

$$\overline{p} - 1.96\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \ \leq\ p\ \leq\ \overline{p} + 1.96\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

信頼度 99% では

$$\overline{p} - 2.58\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \ \leq\ p\ \leq\ \overline{p} + 2.58\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

問 6.4 ある都市の市長選挙のとき、世論調査を行った。有権者の標本として 250 人を無作為抽出してみたところ、110 人が A 候補の支持者であった。有権者全体における A 候補の支持率を信頼度 95% で推定せよ。

6.4 検定

２つの母集団を比較することを考えよう。これらから抽出された標本間に見られる差異が，母集団の間の本質の差異か，単なる偶然変動で少しだけの違いがあるのか。どちらが正しいか，データの統計的な変動の大きさなどから判断する。これが仮説検定である。たとえば，あるテレビ番組の視聴率が，前回の調査時では１２%であったが，今回は１５%と多くなった。この数字をみて，番組の人気が上昇していると判断できるか，という問題がひとつの例である。

帰無仮説と対立仮説　まず「前後の視聴率には変化がない」という仮説を立てる。慎重にあるいは冷静に悲観的にともいうべき立場での，このような仮説を**帰無仮説**（ H_0 と表す）という。これに対して，期待すべき，積極的，楽観的な「変化があることを期待する」ことの仮説を**対立仮説**（ H_1 と表す）という。帰無仮説が否定された場合に成り立つ仮説である。いづれか一つのみが正しいとする 2 者択一の世界である。さまざまな状況において，対立仮説の命題にはいろいろな場合が考えられる。たとえば，血圧降下の新薬を開発しているとき，その効果を調べるときには，帰無仮説が「服用の前後に変化なし（効果なし）」であって，対立仮説は「効果あり（降下している）」であるから，変化のもう一つ「上昇している」ことはあり得ないから，この場合には対立仮説になり得ない。

第 1 種の過誤と第 2 種の過誤　　標本データから仮説が正しいもの (真）と判断することを，仮説を**採択する**といい，それに対して，正しくない（偽）とする判断を，仮説を**棄却する**という。仮説の検定では、判断の誤りが伴う。帰無仮説が真であるとき，これを棄却する誤りを**第 1 種の過誤**といい，帰無仮説が偽であるにもかかわらす，これを採択する誤りを**第 2 種の過誤**という。第 1 種の過誤を犯す確率を**有意水準または危険率**といい，第 2 種の過誤を犯さない確率を**検出力**という。

検定統計量と棄却域　　実際の検定では、判断を下すために，与えられた仮説に対して，適当に選んだ統計量がある領域に含まれるかどうかで棄却と採択を決める。この統計量を**検定統計量**とよび，帰無仮説が棄却されることになる検定統計量の実現値の範囲を**棄却域**という。検定統計量と棄却域を選ぶことが，検定を定めることで，有意水準を一定以下にし，検出力を最大にするよう検定を求める。

検定の一般的な手順

(1)　帰無仮説と対立仮説を定める。
(2)　有意水準の値を決める。
(3)　検定統計量と棄却域を選ぶ。
(4)　与えられた標本データから，検定統計量を計算する。
(5)　もし検定統計量が棄却域に含まれるならば，帰無仮説を棄却（対立仮説を採択する），あるいは棄却域に含まれないならば，帰無仮説を採択する。

例題 1.（平均値の検定）正規母集団 $N(\mu,\sigma^2)$ において，分散 $\sigma^2 = 15^2 = 225$ は既知であるとして，100 個のデータから計算された標本平均が $\bar{x} = 38.0$ であった。このとき，帰無仮説 $H_0 : \mu = 40$，対立仮説 $H_0 : \mu \neq 40$ の有意水準 $\alpha = 0.05$ の検定をおこなえ。

（解）母集団分布は正規分布であり，$\sigma^2 = 15^2 = 225$ は既知として与えられているから，検定統計量は，$z = \dfrac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}} = \dfrac{\overline{x} - 40}{\sqrt{225/100}} = -1.33$, 対立仮説から両側検定を考える。有意水準は $\alpha = 0.05$ で，両側にそれぞれ $\alpha/2 = 0.025 = 2.5\%$, に対する標準正規分布のパーセント点を正規分布表（インターネットのホームページ，統計テキストの付表など）から 読み取ると，$z_{\alpha/2} = z_{0.0025} = 1.960$ である。したがって棄却域は $|z| > 1.960$ を得る。この棄却域と検定統計量の値，すなわち $z = -1.33$ との大小関係を比較すると，棄却域に含まれない（棄却域に落ちない）ので，帰無仮説は棄却されない，いいかえると，帰無仮説は採択される。つまり 100 個の標本平均値 $\overline{x} = 38.0$ を得たが，このデータは正規分布 $N(40, 225)$ からの抽出されたと判断される。（終）

例題２ ある技師がニッケルの融点を 9 回測定して，つぎの値を得た（単位：^oC）。

1475 1420 1433 1452 1441 1466 1432 1453 1414

この結果はニッケルの真の融点とされている 1455^oC に矛盾しないという仮説を有意水準 5% で検定せよ。

（解）分布は明記されていないが，正規分布であるとする。また対立仮説は問題の意図から，両側検定，すなわち

$$H_0 : \mu = 1455 \qquad H_1 : \mu \neq 1455$$

とすることが適当である。与えられたデータから，検定統計量の値を計算する。まず変量を $x_1, x_2, \cdots, x_9$ とおき，$n = 9, \sum_i x_i = 12986, \sum_i^2 = 18740684$ となる。(注意；このデータのように，2 乗すると桁数が多くなる。電卓等の有効計算桁数を正確にするためには，分散の計算は，直接 2 乗するのでなく，変数変換，つまり各データから 1400 を引いて，小さい数字の平均と分散を計算して，もとにもどすこと，を行うべきである。) 題意より，分散が未知の場合であり，$\alpha = 0.05, \nu = 9 - 1$ の t －分布表を用いる。$t_{0.025}(8) = 2.306$ であるから，棄却域は $|t| > 2.306$ つまり，$(-2.306 < t < 2.306)$。与えられた標本値から検定統計量の値は，

$$t = \frac{\overline{x} - \mu_0}{\sqrt{u^2/n}} = \frac{1442.9 - 1455}{\sqrt{416.1/9}} = -1.78$$

この値は棄却域に落ちないから，帰無仮説 H_0 は棄却されず，採択となる。したがってこの実験結果は真の融点とされている値と矛盾していない。(終)

例題３． ある年に行われたレスリングの重量別大会で，36 試合中の勝者の平均体重は 64.5Kg, 標準偏差は 3.2Kg であった。一方敗者はそれぞれ平均 62.8Kg, 標準偏差は 2.5Kg であったという。選手の体重は試合の勝ち負けに影響するといえるか？

(解)「体重は試合の結果に影響しない」という帰無仮説を立て，また選手の体重は正規分布に従うとして，有意水準を 5% として解く（問題には水準が要求されていないから，1% などとして解いてもよい)。帰無仮説をいいかえると，「すべての勝者の平均体重 μ_1 とすべての敗者の平均体重 μ_2 との間には差がない」ということであり，体重の多いほうが有利と考えられるので，ここでは右側検定を用いる。すなわち

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 > 0$$

また $n_1 = n_2 = 36$ だから，大標本による 2 つの平均値の差の検定における統計量と棄却域をもちいる。与えられた標本値から検定統計量の値は，

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{64.5 - 62.8}{\sqrt{\dfrac{3.2^2}{36} + \dfrac{2.5^2}{36}}} = 2.51$$

$\alpha = 0.05$ の右側 (片側) 検定で，$z_0.05 = 1.645$ だから，棄却域は，$\{z : z > 1.645\}$ の区間である。したがって $z = 2.51 > 1.645$ より，帰無仮説は棄却される。体重はレスリングの試合結果に影響すると考えられる。(終)

7　仮説検定の一覧表

参照「統計学演習」村上正康、安田正實, 培風館

仮説	条件	検定統計量	検定統計量の分布	棄却域 (有意水準 $\alpha = 0.05, 0.01$ など)
1. 平均の検定				
(i) $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	母集団分布は正規分布 $N(\mu, \sigma^2)$, σ^2 は既知	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ ただし大標本 ($n \geq 30$) のときは, 母集団分散 σ^2 のかわりに標本不偏分散 u^2 をもちいて $z = \dfrac{\overline{x} - \mu_0}{u/\sqrt{n}}$ としてもよい。	標準正規分布 $N(0,1)$	(i) $\lvert z \rvert > z_{\alpha/2}$
(ii) $H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$				(ii) $z > z_{\alpha/2}$
(iii) $H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$				(iii) $z < -z_{\alpha/2}$
2. 平均の検定 (小標本)				
(i) $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	母集団分布は正規分布 $N(\mu, \sigma^2)$, σ^2 は未知	$t = \dfrac{\overline{x} - \mu_0}{u/\sqrt{n}}$	スチューデントの t 分布 (自由度 $\nu = n - 1$)	(i) $\lvert t \rvert > t_{\frac{\alpha}{2}}(\nu)$
(ii) $H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$				(ii) $t > t_{\frac{\alpha}{2}}(\nu)$
(iii) $H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$				(iii) $t < -t_{\frac{\alpha}{2}}(\nu)$
3. 分散の検定				
(i) $H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	母集団分布は正規分布	$\chi^2 = \dfrac{u^2}{\sigma^2/\nu} = \dfrac{s^2}{\sigma^2/n}$	カイ 2 乗分布 (自由度 $\nu = n - 1$)	(i) $\chi^2 < \chi^2_{1-\frac{\alpha}{2}}(\nu)$, $\chi^2 > \chi^2_{\frac{\alpha}{2}}(\nu)$
(ii) $H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	$N(\mu, \sigma^2)$ $\nu = n - 1$			(ii) $\chi^2 > \chi^2_{\alpha}(\nu)$
4. 比率 (割合) の検定				
(i) $H_0 : p = p_0$ $H_1 : p \neq p_0$	母集団分布はベルヌーイ分布 $B(p)$, 大標本 ($n > 30$) で $np, n(1-p) > 5$	$z = \dfrac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ ただし $\overline{p}$ は標本比率(平均)	標準正規分布 $N(0,1)$ で近似	(i) $\lvert z \rvert > z_{\alpha/2}$
(ii) $H_0 : p = p_0$ $H_1 : p > p_0$				(ii) $z > z_{\alpha/2}$
(iii) $H_0 : p = p_0$ $H_1 : p < p_0$				(iii) $z < -z_{\alpha/2}$
5. 平均の差の検定 (2 標本問題)				
(i) $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	2 つの母集団分布は正規分布 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, 大きさ n_1, n_2 と分散 σ_1^2, σ_2^2 は既知	$z = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{u}}$ ただし $u = \sigma_1^2/n_1 + \sigma_2^2/n_2$	標準正規分布 $N(0,1)$	(i) $\lvert z \rvert > z_{\frac{\alpha}{2}}$
(ii) $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$				(ii) $z > z_{\frac{\alpha}{2}}$
(iii) $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$				(iii) $z < -z_{\frac{\alpha}{2}}$

7.1 統計数値表のための表計算ソフト命令

いままでの多くの統計教科書には、一様乱数、正規分布表、t −分布表、カイ 2 乗分布表、F 分布表が必ず附表として掲載されているが、ここでは表計算ソフトで、これらを求める計算命令を述べることにする。

分布名	Excel 関数名	説明
一様分布; $Unif[0,1]$	RAND()	0 以上で 1 より小さい実数の乱数, $[a,b]$ の範囲では, $RAND()*(b-a)+a$
二項分布; $Binom(n,p)$	BINOMDIST$(k,n,p,T/F)$	k:成功数, n:試行回数, p:成功率, T/F:TRUE は分布関数, FALSE は確率密度関数
正規分布; $N(\mu,\sigma^2)$	NORMDIST$(x,\mu,\sigma,T/F)$	x: 変数, μ:平均, σ:標準偏差, T/F: TRUE は分布関数、FALSE は確率密度関数
正規分布の%点逆引き	NORMINV(p,μ,σ)	NORMDIST の逆関数, p:確率,μ:平均,σ:標準偏差, NORMDIST(x,μ, σ, TRUE) $= p$ となる値 x
標準正規分布; $N(0,1)$	NORMSDIST(z)	z:変数, $P(Z \leq z)$ の値、$Z \sim N(0,1)$, 標準正規分布の分布関数
標準正規分布の%点逆引き	NORMSINV(p)	p:確率, NORMSDIST の逆関数
カイ 2 乗分布; $\chi^2(f)$	CHIDIST(x,f)	x; 変数値, f: 自由度, カイ 2 乗分布の上側 (右裾) 確率:CHIDIST$= P(X > x)$
カイ 2 乗分布の%点逆引き	CHIINV(p,f)	p:確率, f:自由度, CHIDIST の逆関数, p =CHIDIST(x,f), CHIINV$(p,f) = x$ という関係
F 分布; $F(f1,f2)$	FDIST$(x,f1,f2)$	F 分布の上側確率, x:変数値, $(f1,f2)$: 自由度
F 分布の%点逆引き	FINV$(p,f1,f2)$	FDIST の逆関数, FDIST$(x,f1,f2) = p$, FINV$(p,f1,f2) = x$, p:上側確率, $(f1,f2)$: 自由度
t 分布; $t(f)$	TDIST(x,f,B)	x: 変数値, f:自由度, B:片側／両側尾部, 尾部に 1 を指定すると片側分布の値、2 を指定すると両側分布の値。TDIST$(x,f,1) = P(X > x)$ として計算、TDIST$(x,f,2) = P(\|X\| > x) = P(X > x\, or\, X < -x)$。$x < 0$ の場合には, TDIST$(-x,f,1) = 1-$TDIST$(x,f,1) = P(X > -x)$ および TDIST$(-x,f,2)$ =TDIST$(x,f,2) = P(\|X\| > x)$ を使用。
t 分布の%点逆引き	TINV(p,f)	TDIST の逆関数, p:両側確率, f:自由度, TINV は、$P(\|X\| > t) = p$ となる t の値, $P(\|X\| > t) = P(X < -t or X > t)$ 。片側 t 値は、確率に $2*p$ を指定。例：$p = 0.05$、$f = 10$ の場合、両側値は TINV$(0.05,10) = 2.28139$、片側値は TINV$(2*0.05,10) = 1.812462$