統計を始める前に
（主に言語研究のために）

Fumiaki Nishihara〔西原 史暁〕
2012年5月5日〔こどもの日〕

目次

第 1 章

言語研究と統計

現実世界は時間的にも空間的にも大きく広がっている。このため、人間の持つ能力で世界の隅々まで精細に捉えるのは難しい。しかしそれでも世界をうまく捉えなくては科学的研究を成功させることはできない。

世界をうまく捉える手法には様々なものがあるが、中でも重要なのは、現実世界にある大量のデータから知見を得る計量的な手法である。統計は、大量のデータを処理するときに有用なツールであり、このことは言語研究においても同じである。

この章では、統計とはどのようなものか簡単に説明した後、言語研究と統計の関係を見ていきたい。

1.1

統計とは何か

There are three kinds of lies: lies, damned lies, and statistics.

Attributed to BENJAMIN DISRAELI

言語研究と統計の関係を見る前に、統計はどこからきたのか、そして統計で何ができるのかということについて簡単に紹介する。

1.1.1 統計の語源

他の多くの学問と同じく、統計は主にヨーロッパで発展してきた学問である。日本に統計が本格的に導入されたのは、明治に入ってからである。以下では、「統計」という言葉の語源をたどることで、統計がどのようにして生まれたかについて見ていく。

1.1.1.1 英語での「統計」の由来

英語では、「統計学」も「統計」も、“statistics” と表される。

参考 最後の “s” を忘れないようにする必要がある。“statistic” は「統計量」という別の用語となるので注意しなくてはならない。また、“statistics” は単数扱いされる。

英語の “statistics” は、ドイツ語の *Statistik* に由来する。これは、元来 *Staat*（英：state）、つまり国家の状況を知る学問のことであった。つまり、かつての統計は、国力をはかることを目的としていた。

参考 もちろん、現在では統計は、政治的な分野以外にも使われる。統計の歴史については、Hacking (1990) に詳しい。

1.1.1.2 日本語での「統計」の由来

日本には近代になってはじめて統計（学）が導入された。明治初めには、英語の “statistics” に対して、「統計学」以外の訳語もあった。例えば、「国勢学」・「知国学」・「政表学」といったものがある。これらの訳語には、統計がもともと国力を知るために使われてきたということが反映されている。

図 1.1　箕作麟祥（みつくりりんしょう）

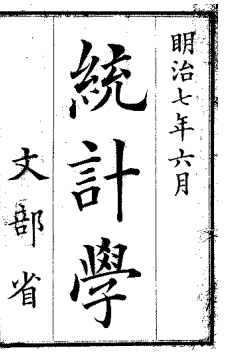

図 1.2　『統計学』(1874)

「統計」の初期の使用例として、1871 年 7 月 27 日に設置された「統計司」という役所の名前がある。なお、1872 年 8 月 10 日に統計司は、「統計寮」と改称されている。学問としての「統計学」という用語が最初に出てきたのは、1874 年 6 月に、箕作麟祥（みつくりりんしょう）が翻訳出版した『統計学』という書籍においてである。

> **『統計学・凡例』**
> 此学原名ヲ スタチスチック ト云ヒ其ノ説ク所ハ皆算数ヲ以テ国内百般ノ事ヲ表明シ治国安民ノ為メ最モ緊要ノ者タリ（傍線原書）

> **「統計学」（『法窓夜話』穂積陳重）**
> スタチスチックスの訳名が「統計学」と定まるまでには多少の沿革がある。始め慶応三年四月に出版せられた神田孝平氏訳「経済小学」の序には、スタチスチックスを訳して「会計学」としてあるが、明治三年二月発布の「大学規則」には「国勢学」とある。これは、欧洲において中世より第十八世紀の始めに至るまでは、この語原の示す如く、国家の状態を研究する学問となっていたとのことであるから、その後の沿革を知らずに、二百年前の用例をそのままに「国勢学」と邦訳したのであろう。同年十月の大学南校規則にも「国務学」となっている。世良太一君の直話に拠れば、国勢学を一時「知国学」ともいうたことがあるが、これは多分杉亨二（こうじ）先生の案出であろうとのことである。津田真道先生がオランダのシモン・ヒッセリングの著書を訳して明治七年十月に太政官の政表課から出版せられたものに「表紀提綱一名政表学論」というの

> がある。「西周伝」に拠れば、津田先生は学名としては「綜紀学」という語を用いられたようである。世良太一君の話に拠ると、「政表」という語は、この後明治十年頃までも用いられたということである。
>
> かくの如く「スタチスチックス」に対する訳字が従来区々であったので、むしろ原語そのままを用いた方が好かろうということで、明治九年頃、杉亨二博士・世良太一氏らの創められた学会には、「スタチスチックス」社という名称を附し、「スタチスチックス」雑誌というのを発刊せられたが、当時「スタチスチックス」という原語に宛てるために寸多知寸知久（スタチスチク）という漢字をも案出創造せられたということである。また始め神田氏の用いられた「会計学」という名称も、その字義からいえば至極穏当のようではあるが、「会計」は他の意義に用いられているから、「統計学」の方が適当であろう。しからばこの「統計学」という名称の創始者はそもそも何人であろうか。
>
> 明治四年七月二十七日大蔵省の中に始めて置かれた役所に統計司というのがある。これは翌八月十日に至って統計寮と改められたが、官署の名に「統計」の名を附したのはこれが初めてである。この「統計」の二字は、恐らくは「英華字典」にスタチスチックに対して「統紀」という訳字を用いておったのに拠って案出したものであろう。この後ち明治七年六月になって、箕作麟祥博士が仏人モロー・ド・ジョンネの著書を翻訳して文部省から出版せられたものには「統計学一名国勢略論」という標題を用いられた。学名として「統計学」という各称を用いたのは、けだしこの書をもって初めとなすべきである。そして前にも述べた如く、この後にも「国勢学」「知国学」「政表学」または「表紀」「寸多チ寸チ久（スタチスチク）」などの名称が存在したにもかかわらず、後には「統計学」という名称が一般に行われて、終に学名と定まるに至ったのである。

1.1.2 統計で何をするか

ところで、統計を使ってどういうことができるのだろうか。統計でできることは多岐にわたるが、大まかに言えば、以下の 4 点にまとめられる。

1. データの集め方を考えること
2. データを集めること
3. データを整理すること
4. データから予測すること

データを集め終わってから統計的な分析が始まると考える人は多い。しかし、これは、ありがちな間違いである。本来はデータを集める前から、統計的に考える必要があるのである。また、一般の統計の入門書などでは、上に挙げた 4 点の内、「データから予測すること」が一番重要な地位を占めることが多く、このことの説明が中心となりがちである。だからと言って、他の点が重要でないわけではない。実際の研究では、データを集めるところから始めることになるので、「データから予測すること」ばかりを考えるのではなく、他のことにも気を配ることが必要である。

1.1.2.1 データの集め方を考える

実際の研究では、まずは自分がどういう仮説を想定しているのかをしっかりと見極める必要がある。そして、その仮説が、統計の言葉で言えば、どのように表現されるかを考えていかなくてはならない。さらに、それを統計的に検証するには、どういうデータを集めれば良いのかということも考えていかなくてはならない。

最初にこのようなことを行わなければ、後で困ることになる。データの集め方を考えるときから、統計のことをしっかり考えていかなくてはならないのである。

1.1.2.2 データを集める

データは、ただやみくもに集めれば良いというものではなく、しっかりとした計画のもとで集めていかなくてはならない。計画をたてずに、やみくもにデータを集めると、良質なデータは得られなくなる。良質なデータが得られなければ、良質な分析はできなくなってしまう。

このため、データの収集にあたっては、細心の注意を払い、良質なデータが得られるように努力する必要がある。ここにも実は統計が関わってくるのである。

1.1.2.3 データを整理する

集められたデータは、そのままでは複雑すぎて、うまく扱うことはできないことがほとんどである。データに対する分析を行う際は、まず最初にデータを取り扱いやすい形にする必要があるのである。

例えば、集計結果を表にまとめれば、データの全体をうまくまとめることができる。さらに、グラフを作れば、データの特徴を視覚的に捉えることが

できる。この他、平均を出したりすることでも、データの様子を調べられる。

1.1.2.4 データから予測する

データが整理されたら、そのデータに基づき、仮説が正しいか検証する必要がある。また、データから色々と推論することもある。こういったときにも統計は大きな力を果たす。例えば、検定という手法を用いれば、自分の仮説が妥当かどうかを調べることができる。

データの予測が済めば、それで終わりというわけではない。仮説の検証結果をもとに、別の仮説を立てたりさらにデータを集めたりする必要が出てくることもありうる。こういった場合、必要に応じて、データの集め方を考えるところに戻るべきである。

1.2

統計と言語研究

どうして君は他人の報告を信じるばかりで自分の
目で観察したり見たりしなかったのですか？

ガリレオ・ガリレイ『天文対話』

この節では、統計と言語研究の関係について簡単に説明する。
今や言語研究は統計を無視することができなくなった。もちろん、統計を使わない手法で言語を研究する人も多いのだが、そういった人でも他の人の研究を見るときには統計の知識が必要な場合が出てくる。このため、これから言語研究を志す人にとって、統計の知識は必要不可欠となる。ここでは、統計が言語研究とどう関わってくるのかについて触れる。

どういう言語研究で、統計が必要となるのだろうか。

伝統的には、言語研究において統計が使われることはなかった。しかし、徐々に言語研究にも統計的手法が使われるようになった。それでは、統計を用いる言語研究とそうでない言語研究の間にはどのような違いがあるのだろうか。

以下、まず、言語研究にはさまざまなものがあることを確認し、その後統計を用いない言語研究、統計を使用する言語研究の順で紹介していきたい。

1.2.1 言語研究とは

言語研究と統計の間の関係を考える前に、言語研究はどういうものかということを見直すことから始めよう。一口に言語研究といっても、さまざまな下位分野があり、分野によって用いられる手法が異なっている。手法として統計がよく用いられる分野もあるし、統計的手法を用いるのが適切でない分野も存在する。このため、自分が研究しようと考えている分野で、統計が必要かどうか、しっかり考えていく必要がある。

問 1–1　言語学の下位分野
一口に言語学といっても、統語論や意味論といったさまざまな下位分野がある。こういった言語学の下位分野を思いつくだけ挙げてみよう。

問 1–2　言語学以外の言語研究
言語学以外に、言語を研究している学問はあるだろうか。あるとしたら、それは言語学と何が違うのだろうか。ないとしたら、なぜ言語学以外の言語研究は成立しないのか。

言語研究 (language studies) は、言語学と同義語ではない。言語学は言語を研究する学問の 1 つであり、言語学以外にも言語を研究する学問はある。例えば、言語哲学では哲学の手法を用いて言語について研究していく。また、言語社会学は、社会学の手法を用いて言語について研究する。自然言語処理では、工学的な興味から言語を取り扱っている。

1.2.2　統計を用いない言語研究

どういう言語研究で統計が必要になるかを考える前に、統計的手法を用いない言語研究について考えてみよう。

言語研究において必ず統計を使うわけではない。統計を使わない言語研究も少なくない。実際、構造主義言語学や（伝統的な）生成文法などでは、統計を使うことがほとんどなかった。これらの研究では、言語は等質であると仮定している。つまり、ある言語を話す母語話者はみな同じようなものであり、多少の違いはあったとしても、それは無視できると考えていたのである。

例えば、伝統的な統語論では、議論に用いる証拠として文法性判断を用いることが多い。文法性判断というのは、ある文が文法的に成立するかを母語話者が判断することである。例えば、以下の 3 つの文で、1 番目と 2 番目は自然な文だが、3 番目の文は日本語としておかしい。 こういった文法性判断を積み重ねていくことで、議論を進めていくのである。

1. 太郎が歩道を歩いた。

2. 花子が太郎を歩かせた。
3. *花子が太郎を歩道を歩かせた。

こういった文法性判断を行うとき、普通は多くの人に判断してもらうことはしない。研究者が母語話者であれば、研究者自身の判断で済ませてしまうことが多い。このように済ませられるのは、母語話者が等質であって、この研究者も他の母語話者と同じような判断をするだろうと仮定しているからである。

> **問 1–3　内省による文法性判断**
> 統語論の論文では、論文の著者が研究対象の言語の母語話者である場合、しばしば著者自身が内省を行うことで、文の文法性を決めることがある。例えば、日本語の母語話者である山本氏は、自身の著作の中で「太郎が友達をぼこぼこに五人殴った」という文を挙げ、山本氏自身の内省によりこの文を非文であると決めた。このような自身の内省に頼った文法性判断には、どういう問題点があると考えられるだろうか。

1.2.3　統計が必要な言語研究

統計が必要となる理由には、**変異**の存在、および**大量のデータ**の処理という 2 つの側面がある。

参考 変異は、バリエーション (vatiation) や変動と言い換えても良い。

1.2.3.1 言語に含まれる変異に着目する

先に述べたように、ある種の言語研究では言語を等質なものとして捉えてきた。しかし、言語は必ずしも等質ではない。母語話者による文法性判断の場合、男と女で判断が違うかもしれないし、年代によっても判断は異なるかもしれない。実際、性別や年齢によって用いる言葉は違うわけで、この違いに着目しようとする研究、つまり言語の中の変異に注目した研究も存在している。このような変異を重視する言語研究では統計が重要な働きを果たすことになる。

参考 変異があったら、必ず統計が必要かと言えば、そうではない。変異が無視で

きる状況であったら、統計は必要ない。

このことは「**変異なくして統計の必要なし**」(*If no variation, no need for statistics*) という言葉で端的に言明される。つまり、統計が必要な場合とは、研究対象に何らかの変異が含まれる場合である。言語研究においても同様で、統計が必要な場合とは、言語そのものや言葉を話す人の変異に注目する場合なのである。

例えば、以下の言語研究は、みな変異が無視できない研究であり、統計が必要となる場合が多い。

- 実験言語学
- 談話研究
- 社会言語学
- 言語教育研究
- コーパス言語学

参考 なお、上記の研究分野は一例であって、他の分野でも統計を使うことはあり得る。逆にこれらの分野でも統計を使わない研究手法を使うこともできる。

1.2.3.2 大量のデータを処理する

変異をあまり重視しない場合でも、大量のデータを処理するならば統計が必要になる。

言葉というものはみながみな使っているものであって、音声の形にせよ文字の形にせよ、毎日大量の言葉が紡ぎ出されている。しかし、人間の持つ能力では、世界の隅々まで精細に捉えるのは難しい。統計は、大量のものを捉える手法として有用であり、言語データの分析にも十分役立つ。

伝統的な言語研究では、必要最小限の例文だけ挙げて議論することが多かった。しかし、少ない例文からではどうしても取りこぼしてしまう言語現象が出てくる。こういったとき有効なのが、大量の言語データを分析する手法である。こういった大量のデータを扱う言語研究には、以下のようなものがある。

- コーパス言語学
- 自然言語処理

なお、先ほど述べた実験言語学や談話研究などのバリエーションを重視す

る研究でも、大量のデータを扱うことがある。そういった場合は二重の意味で統計が必要になる。

1.2.4 言語研究での変異

先に見たように、言語研究の中には、変異を想定しなくてはならない分野が存在する。それでは、そもそもどういうものが変異として捉えることができるのだろうか。

1.2.4.1 言語教育研究の場合

> **問 1–4 言語教育研究における変異**
> 言語教育研究の分野ではどういう変異が想定できるだろうか？

言語教育の研究では、例えば、以下のものが変異として捉えられうる。

- 学習者の本来の能力の差という変異
- 教師の教え方の差という変異
- 授業の規模という変異
- 成績評価という変異
- カリキュラムという変異
- そもそもどんな言語を教えるかという変異

もちろん、上に述べたもの以外にも、さまざまな変異が想定できる。いずれにせよ、さまざまなことが変異として捉えることができ、結局**変異は多数挙げることができる**。

1.2.4.2 実験言語学の場合

> **問 1–5 実験言語学における変異**
> 実験言語学の分野ではどういう変異が想定できるだろうか？

実験言語学の研究では、例えば、以下のものが変異として捉えられうる。

- 被験者という変異

  - 性別という変異
  - 年齢という変異
  - 出身地という変異
- 実験者という変異
- 実験に用いる言語材料という変異
- 実験順序という変異
- 実験機材という変異
- 実験場所という変異
- 実験日の気温という変異
- 実験日の天気という変異
- 実験日の株価という変異
- 実験日の交通事故数という変異

株価や交通事故数が言葉に影響を与えることはあるだろうか？ 株価が下がれば、被験者が落ち込んで、うまく回答できなくなるかもしれない。しかし、常識的に考えてそういう影響はほとんど無いだろう。

先に述べたように、変異はさまざまなものを挙げることができる。しかし、すべての変異が言語研究に関わってくると言うわけではない。さらに、あまりに変異の種類が多すぎると、処理しきれなくなる。このため、役に立ちそうにない変異は切り捨てることも必要となってくる。

問 1–6　さまざまな研究分野における変異

以下の言語研究の分野で、どのようなものが変異としてしばしば想定されるか考えよ。

1. 自然言語処理
2. コーパス言語学
3. 音声学

1.2.5　分野ごとの統計との関わり方の違い

注意しなくてはならないのは、言語研究と統計の関わり方は一通りでないということだ。言語研究にはさまざまな研究対象と研究手法があり、対象と手法に応じて、統計の使い方も変わってくる。

思うに、言語研究と統計の関わり方は、大まかに言って 3 つのパターンがある。1 番目はコーパスに代表される大量のテクストデータを処理する研究における統計の使用である。2 番目は実験処理における統計の使用である。3 番目は教育データにおける統計の使用である。

この 3 つのパターンはそれぞれ必要となる統計的知識が違ってくるので注意が必要である。もちろん基礎は同じであるが、何を学ぶべきかは少しずつ異なってくる。

1.2.5.1 大量のテキストデータと統計

コーパス言語学や自然言語処理では、コーパスと呼ばれる大量の言語データを集めたデータベースを用いて研究する。

こういった研究では、ある単語がデータの中に何回出てきたかを調べるといったことも行われる。単に数えるだけとはいえ、これも立派な統計処理である。出現頻度を集計し、どの単語がよく出てくるかを見るだけでも、色々なことが分かってくるのだ。

例えば、英語で未来を表す助動詞には “will” と “shall” の 2 通りある。昔はこの 2 つを使い分けていたのだが、現代に近づくにつれ、“shall” を用いずにもっぱら “will” を用いるようになったと言われている。古い時代から現代にいたるまでのテキストデータを集め、そこでの “will” と “shall” の頻度を見れば、大まかなことが分かる。以下の図 1.3 は、18 世紀から現在に至るまでのアメリカの大統領の就任演説の中で、“will” と “shall” がどれだけ用いられてきたかを示した図である。青い線が引いてあるところでそれぞれの単語が使われている。青が濃いところが、その単語が頻繁に使われているところである。図の右側のほうが現代に近いとみなしてよい。

図 1.3 を見れば、“will” はずっと使われているが、“shall” は最近になって使用頻度が急減したということが分かる。このように、単に頻度を見て、グラフにするだけでも色々なことが分かるし、そこには統計が用いられている。

また、テキストを分類しようとすることがある。例えば、ある文書が与えられたとき、それが医学に関する文書なのか、法律に関する文書なのか知りたい場合があるだろう。こういった分類の基礎となるのが、大量のテキストデータの分析であり、こういった目的のためには多変量解析という手法が用いられる。

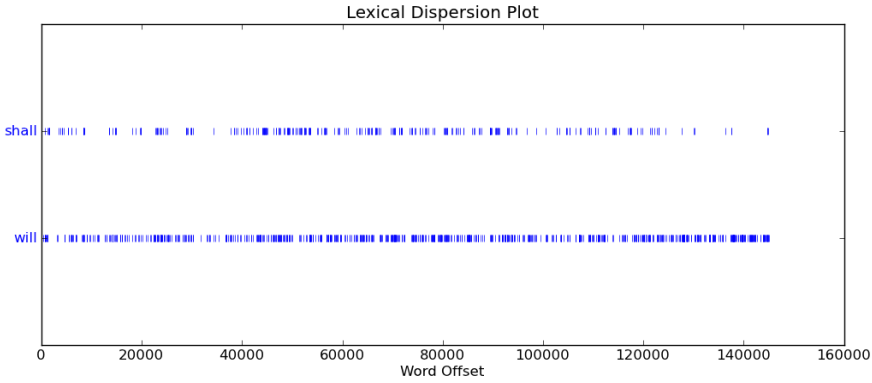

図 1.3　アメリカ大統領就任演説での “shall”（上側）と “will”（下側）の分布。右側にあるほど現代に近い。

1.2.5.2 言語実験と統計

実験をする場合は、大量のテキストデータの分析する場合とは違った統計手法を用いることとなる。

言語研究での実験と言われても想像が付きにくいかも知れない。典型的には、実験を受けてもらう人に来てもらって、その人たちに言語データを提示してどういう反応をするかをみたり、言語データを産出してもらったりする。例えば、ディスプレー上に文を表示させてその文を読み終えるまで何秒かかるかを見て、文の複雑さを調べる場合がある。また、地域によるアクセントの違いを見るために、色んな地域の人に文章を読み上げてもらってその音を録音するといったことも行われる。

実験の目的はさまざまだが、ありがちなのがグループごとに差があるかどうかを調べるものである。例えば、男と女で違いはあるのか、出身地によって違いはあるのかといったことが問題になる。差の有無を調べるには、検定と呼ばれる統計手法が用いられる。検定を用いると、例えば「こちらのグループとあちらのグループでは、99% の確率で反応時間が違う」といった主張が示せる。

参考 「こちらのグループとあちらのグループでは、99% の確率で反応時間が違う」というのはあまり正確な表現ではないのだが、ここでは説明のためにあえて

このような表現を用いた。

差の有無の調査は、実験の後に行われるものだが、実験を始める前にも統計の知識は重要になってくる。どのような順番で実験を行えば良いのか、どのような人に実験に参加してもらえば良いのかについても統計の知識で対応できる。統計の知識無しで適当に実験の計画を立ててしまうと、後で実験の結果をまとめるときにうまくいかなくなってしまう。

1.2.5.3 言語教育研究と統計

教育に関する研究がやりづらいのは、教えたり学んだりするプロセスに関わる要因があまりにも多すぎるからである。例えば、外国語を学習する場合、年齢・性別によって学び方は違うだろうし、学習環境であるとか、親の教育観であるとか、本人の性格によっても成果は大きく変わってくるだろう。言語実験の場合、実験者がある程度統制できる面があるが、教育データの場合、統制が難しい。

例えば、「小学校で英語を教えた方が良いか、それとも小学校では英語を教えずに国語をしっかり学ばせた方が良いか」という問題を考えてみよう。このことを調べるためには、単純に言えば、5 歳ぐらいの子どもを集めて、小学校で英語を学ばせるグループと学ばせないグループにアトランダムに分けて、将来どうなるかを見れば良い。しかし、このように分けることが許されるだろうか？　本人や親の希望を無視してアトランダムに決めてしまうのは、相当問題があることである。また、たとえできたとしても、英語を学ばせないグループの子どもの親が、子どもを英会話教室に通わせたりするかもしれない。そうなったら、せっかく学ばせるグループとそうでないグループに分けたのが台無しである。

要するに、教育データは、統制が困難という側面がある。普通の統計手法は、統制がしっかり行われている場合に最大の効果を発揮するので、普通の統計手法をそのまま教育データに適用するのは難しい。なお、実はこういった場合に対応するための統計手法も色々開発されている。教育データを扱う場合はそういった手法に通じている必要がある。

第 2 章

統計学のための基礎数学

この章では、統計学を扱うために必要な数学の基礎知識について紹介する。統計学には、数学の知識が必須であり、これを避けて通ることはできない。

「数式が出てくると分かりづらい」と考えてしまう人は少なくない。しかし、このような考えは誤りである。実際には、数式を用いることで、普通の言葉では説明しづらいことを分かりやすく示すことができるのだ。数式は、別に読者を混乱させるために存在しているのではない。もし数式が分かりづらいと思うのならば、それは数式に対する予備知識が足りないだけのことである。しっかり学びさえすれば、決して分かりづらいものではない。

統計の入門書のたぐいでは、「数式が出てこない」ことを宣伝文句としている場合がある。こういった書籍で統計を学ぶと、結局、なぜそうなるのかということが理解できないまま終わってしまうだろう。また、数式がなければ、統計処理の部分がブラックボックスと化してしまう。すべてのことを完全に理解することは難しいが、最初から全く避けてしまうのは研究者の態度として問題がある。やはり、しっかりと前向きに数学的知識を蓄積した上で、統計を理解していく必要がある。

2.1

基礎的な文字・記号

We summarize with this, the most remarkable formula in mathematics:

$$e^{i\theta} = \cos\theta + i\sin\theta$$

This is our jewel.

Richard Phillips Feynman

最初に、統計の専門書などでよく用いられる文字・記号を紹介する。いずれも基礎的なものであり、後で何度も出てくるものである。これらの文字・記号を熟悉しているのならば読みとばしても良いだろう。もし、これらの文字・記号を見慣れていないのであるならば、ここでしっかり把握しておく必要がある。なお、四則演算については、特に計算機での表示法を紹介している。これらは計算機で計算を行うときにしばしば用いるものであり、統計を使うとき以外にも役立つであろう。

2.1.1 ギリシャ文字

統計では、ギリシャ文字がいくつか使われるので、この際覚えておこう。表 2.1 には、統計学の教科書などでよく使われるギリシャ文字が書かれている。太字のものは特によく使うものなのでしっかり覚えておこう。なお、表 2.1 の読み方は日本語のカタカナにしたときの代表的な読み方であり、他の読み方をすることもある。α, β, γ は、ギリシャ文字の先頭の 3 文字であり、ローマ字の a, b, c のように、第 1、第 2、第 3 の数を代表させるために使うことがある。

参考 言語学でも、統語論での “θ-role” や意味論での “λ-abstraction” といったように、ギリシャ文字が出てくることは多々ある。覚えていない人は早いうちに覚えるようにしたい。

また、表 2.2 には、ギリシャ文字の一覧が、LaTeX での記法とともに掲載されているので参照されたい。LaTeX でギリシャ文字を入力する際には、数式モードの中で、英語の綴りの前に “\” をつけたコマンドにする。ただし、

表 2.1　統計学でよく使われるギリシャ文字

	読み方	ローマ字転写	大小	主な用途
α	アルファ	alpha	小文字	有意水準
β	ベータ	beta	小文字	$(1-\beta)$ で検出力
Γ	ガンマ	gamma	大文字	Γ 函数
γ	ガンマ	gamma	小文字	—
θ	シータ	theta	小文字	確率分布を示すパラメータ
λ	ラムダ	lambda	小文字	ポアソン分布のパラメータ
μ	ミュー	mu	小文字	母集団での平均
ρ	ロー	rho	小文字	母集団での相関係数
π	パイ	pi	小文字	円周率
Σ	シグマ	sigma	大文字	総和記号
σ	シグマ	sigma	小文字	母集団の標準偏差
Φ	ファイ	phi	大文字	正規分布の累積分布函数
ϕ	ファイ	phi	小文字	正規分布の確率密度函数
χ	カイ	chi	小文字	χ^2 分布という確率分布

大文字のアルファや小文字のオミクロンなどは、ラテン文字と同じ形をしているのでコマンドは用意されていない。

参考　数式の中では、小文字はイタリックにし、大文字は立体にするのが通例である。

2.1.2　自然対数の底 e

2.1.2.1 自然対数の底の定義

統計に限らず、数学でよく出てくる定数として**自然対数の底**（てい）と呼ばれるものがある。自然対数の底は、e という記号で書かれる。ネイピア数 (Napier's constant) とも呼ばれる。

表 2.2　ギリシャ文字の一覧

			LaTeX での書き方	
大文字	小文字	読み方	大文字	小文字
A	α	アルファ	`A`	`\alpha`
B	β	ベータ	`B`	`\beta`
Γ	γ	ガンマ	`\Gamma`	`\gamma`
Δ	δ	デルタ	`\Delta`	`\delta`
E	ϵ	イプシロン	`E`	`\epsilon`
Z	ζ	ゼータ	`Z`	`\zeta`
H	η	イータ	`H`	`\eta`
Θ	θ	シータ	`\Theta`	`\theta`
I	ι	イオタ	`I`	`\iota`
K	κ	カッパ	`K`	`\kappa`
Λ	λ	ラムダ	`\Lambda`	`\lambda`
M	μ	ミュー	`M`	`\mu`
N	ν	ニュー	`N`	`\nu`
Ξ	ξ	クシー	`\Xi`	`\xi`
O	o	オミクロン	`O`	`o`
Π	π	パイ	`\Pi`	`\pi`
P	ρ	ロー	`P`	`\rho`
Σ	σ	シグマ	`\Sigma`	`\sigma`
T	τ	タウ	`T`	`\tau`
Y	υ	ウプシロン	`\Upsilon`	`\upsilon`
Φ	ϕ	ファイ	`\Phi`	`\phi`
X	χ	カイ	`X`	`\chi`
Ψ	ψ	プサイ	`\Psi`	`\psi`
Ω	ω	オメガ	`\Omega`	`\omega`

定義 1 自然対数の底

自然対数の底 e とは以下の数式を満たす定数である。

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

参考 e の値はおよそ 2.71828 である。

2.1.2.2 自然対数の底の性質

自然対数の底 e の性質としては、以下のようなものがある。

- 函数 e^x は微分しても形が変わらない。すなわち、$\frac{d}{dx}e^x = e^x$ となる。この性質があるために函数 e^x は扱いやすい。
- 積分すると、$\int e^x = e^x + C$ となる。なお、C は積分定数である。
- e^x は、単に $\exp x$ と書くこともある。
- $\exp x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$
- $e^{i\theta} = \cos\theta + i\sin\theta$ となる (Euler's formula)。よって、$e^{i\pi} = -1$ となる。なお、ここで i は虚数単位である。

2.1.3 四則演算

2.1.3.1 かけ算の記号 ×

かけ算の記号 × は、日常生活ではよく使うが、数学の教科書ではほとんど使わない。

かけ算を表す時は、以下の例のように、数字同士なら、"×" のかわりに "·" を使う。また、文字同士、または数字と文字の間のかけ算の場合は、"×" を省略する。

$$\begin{aligned} 3 \times 5 &\to 3 \cdot 5 \\ 7 \times 10 &\to 7 \cdot 10 \\ a \times b &\to ab \\ x \times y &\to xy \\ 3 \times b &\to 3b \\ 7 \times x &\to 7x \end{aligned}$$

2.1.3.2 計算機における四則演算の記号

かけ算と割り算の記号は、パソコンなどの計算機上においては、日常生活で用いるものと違ったものを使う。R などの統計ソフトに計算させる場合も、計算機用の記号を用いなくてはならない。

計算機上では、かけ算の記号としては、*（アスタリスク）を用いる。また、割り算は/（スラッシュ）、累乗は^（サーカムフレックス）で表される。足し算や引き算は日常生活と同じく、+（プラス）と-（マイナス）を用いる。

以下に、計算機使用時の特殊な記号の使い方の例を掲げる。左側が日常生活での書き方で、右側が計算機用の書き方である。

$$10 \times 6 \to 10 * 6$$
$$24 \div 12 \to 24 / 12$$
$$7^4 \to 7 \hat{} 4$$

参考 R では、累乗を表す際に、^（サーカムフレックス）を使うほかに、アスタリスクを 2 つ並べたもの（**）を用いることもできる。つまり、7^4 は、`7 ^ 4` としても良いし、`7 ** 4` としても良い。

参考 累乗を表す^（サーカムフレックス）を複数用いるときは、その演算順序に注意する必要がある。例えば、`9 ^ 0.5 ^ 3` は、右側の^（サーカムフレックス）から計算され、次に左側の^（サーカムフレックス）が計算される。つまり、まず 0.5 の 3 乗を行って 0.125 を得て、しかる後に、9 の 0.125 乗を計算し、およそ 1.31607401 を得る。つまり、`9 ^ 0.5 ^ 3` は、`9 ^ (0.5 ^ 3)` を意味するのであって、`(9 ^ 0.5) ^ 3` を意味するのではない。

参考 R では、$1+3$ をポーランド記法のように、`"+"(1,3)` のように書くことをできる。また、$(1+3) \times 5$ は、`"*"("+"(1,3),5)` と書くことができる。
R では二項演算子といえども普通の函数と同じなので、函数名を先に書いてその後に引数を置くという形で書けるわけである。

問 2–1　計算機における四則演算

次の数式を計算機用の書き方に改めよ。

1. 20×31
2. $731 \div 64$
3. 10^8
4. $20 \times 32 + 412$

5 $3^9 - 24 \div 6^4$
6 $(74 + 81) \times 15^2 \div (24 - 16)$

2.1.4 絶対値

絶対値 (absolute value) は、ある数が 0 からどれだけ離れているかを示した値であり、x の絶対値 $|x|$ は、以下のように定義される。

定義 2 絶対値

$$|x| = \begin{cases} x & (x \geq 0) \\ -x & (x < 0) \end{cases}$$

要するに、負数の場合、マイナスを外したものが絶対値となる。

- $|-2| = 2$

また、0 あるいは正数の場合は何も変わらない。

- $|0| = 0$
- $|5| = 5$

問 2–2 絶対値
次の式を計算せよ。
1 $|-7|$
2 $|-5|$
3 $|3|$
4 $|8|$
5 $|5-7|$
6 $|-13+5|$
7 $|-7.5|$
8 $|3.9|$
9 $|-12.8|$

2.1.4.1 R での絶対値

絶対値を求める場合、R では、abs という函数を用いる。

```
> abs(-6)
[1] 6
> abs(5)
[1] 5
> x <- c(-7, 8, 27, 0, -12, -3)
> abs(x)
[1]  7  8 27  0 12  3
```

2.2

対数

Why do mathematicians like national parks?
— Because of the natural logs!

A mathematical joke

この節では、**対数** (logarithm) の基本的な扱い方を学ぶ。

言語データを分析する際、極端な値があるなどの理由で、得られたデータがそのままでは処理しにくいときがある。こういったとき、しばしば得られたデータを対数の形に変換して議論する。このため、対数の形で数値を見るという考え方になれておく必要がある。

また、与えられた数値を対数に変換すると、元々乗算で表されていたものが、加算の形に変わる。乗算に比べると加算の方がずっと計算しやすいので、処理が簡単になるのである。

2.2.1 指数演算

対数について知るためには、先に**指数演算**を知っておく必要がある。指数演算と対数演算は、1 枚の紙の表裏のようなものであって互いの関係は非常に深い。

2.2.1.1 指数とは何か

指数というと大げさな言い方のように聞こえるが、要するに累乗のときに右肩に付ける数のことである。例えば、以下の数式で、一番左側の 2 の右肩に付いている 4 が指数である。

$$2^4 = 2 \cdot 2 \cdot 2 \cdot 2 = 16$$

逆に指数を乗せている数値のことを**底**という。2^4 の底は、2 である。

問 2–3　指数演算

次の式を計算せよ。また、指数と底はそれぞれいくつか。

1. 10^3
2. 2^8
3. 5^3

◈ 負数が関わる指数演算

負の数が関わる指数演算の時には、何が底であるかをよく見極める必要がある。例えば、$(-7)^2$ の底は -7 である。よって、$(-7)^2 = (-7) \cdot (-7) = 49$ となる。これに対して、-7^2 というのは、7^2（底は 7）に、マイナスの記号がついているだけであるから、$-7^2 = -(7 \cdot 7) = -49$ となる。

問 2–4　負数が関わる指数演算

次の式を計算せよ。

1. $(-3)^2$
2. -3^2
3. $(-6)^5$
4. -6^5
5. $(-3)^6$
6. -8^4

2.2.1.2 特別な指数演算

今までの指数の計算は、指数が 2 以上の整数である場合のみを扱ってきた。指数の計算は、指数が 1、0、あるいは負の整数である場合にも拡張することができる。以下で、指数が 1 以下の整数である場合の計算方法について述べよう。

参考　下記のように拡張する理由は、こう拡張することで、1 以下の整数の場合でも、2 以上の整数の場合と同じように、後述の指数法則（定理 3）が満たされるからである。

◈ 指数が 1 の場合

指数が 1 ならば、底と同じ値が出てくる。

$$a^1 = a$$

例えば、$5^1 = 5$ となる。

◈ 指数が 0 の場合

指数が 0 なら、底が 0 でさえなければ、どんなときにも 1 が出てくる。

$$a^0 = 1$$

例えば、$10^0 = 1$ となる。

参考 0^0 は通常定義されない。

◈ 指数が負の場合

指数が負のときは、指数が正の場合の逆数をとる。

$$a^{-n} = \frac{1}{\underbrace{a \times a \times a \cdots \times a}_{n}}$$

例えば、$2^{-5} = \frac{1}{2^5} = \frac{1}{2 \cdot 2 \cdot 2 \cdot 2 \cdot 2} = \frac{1}{32}$ となる。

問 2–5　特別な指数演算

次の式を計算せよ。

1. 25^1
2. 312^1
3. $(-23)^1$
4. 34^0
5. 82^0
6. $(-132)^0$
7. 5^{-3}
8. 8^{-2}
9. $(-2)^{-3}$
10. $(-3)^{-4}$

参考 以上の例では、指数が整数 ($\mathbb{Z}$) の場合しか扱わなかったが、指数は実数 ($\mathbb{R}$) の範囲まで拡張できる。よって、$5^{3.4}$ や $7^{\sqrt{2}}$ といったものも計算することができるのである。

2.2.1.3 指数法則

指数に関する計算で役立つのが、以下に示す指数法則 (exponential law) である。

定理 3 指数法則

$a > 0, b > 0$ であり、$r, s \in \mathbb{R}$ であるとき、以下の法則が成り立つ。

$$a^{r+s} = a^r a^s$$
$$(a^r)^s = a^{rs}$$
$$(ab)^r = a^r b^r$$

参考 $r, s \in \mathbb{R}$ とは、r, s が実数であることを示している。$\mathbb{R}$ は “Real Numbers” の頭文字である。

問 2–6 指数法則の確認

以下の各式の左辺と右辺を別個に計算し、指数法則が成り立っていることを確かめよ。

1. $3^5 = 3^2 \cdot 3^3$
2. $6^4 = 6^3 \cdot 6^1$
3. $\left(7^3\right)^2 = 7^6$
4. $\left(9^2\right)^2 = 9^4$
5. $12^5 = 3^5 \cdot 4^5$
6. $(-6)^3 = (-2)^3 \cdot (3)^3$

指数法則を使うと、以下のように計算が楽になる場合がある。

$$71^8 \cdot 71^{-7} = 71^{8-7} = 71^1 = 71$$

問 2–7 指数法則

指数法則を用いて、次の式を計算せよ。

1. $12^5 \cdot 12^{-3}$
2. $76^8 \cdot 76^{-8}$
3. $81^{-12} \cdot 81^{11}$

4 $\left(\frac{1}{8}\right)^5 \cdot 16^5$
5 $\left(24^8\right)^0$
6 $\left(83^{12}\right)^0$

2.2.2 対数の定義

対数(たいすう)は、指数を裏返したものである。

定義 4 対数

a を底とする x の対数 p とは、$x = a^p$ となるような p のことである。

対数は、$p = \log_a x$ の形で log を使って表す。

参考 a は 1 以外の正の実数、x は正の実数でなくてはならない。p の値は、正・負・ゼロのいずれもありうる。

参考 x は**真数**(しんすう)と呼ばれる。

例えば、2 を底とするとき、16 の対数は、$2^4 = 16$ なので、4 である。すなわち、

$$\log_2 16 = 4$$

となる。

参考 対数は指数の逆演算として捉えることができる。この他の逆演算の事例としては、
- 足し算に対して引き算（$+ \leftrightarrow -$）
- かけ算に対してわり算（$\times \leftrightarrow \div$）
- 平方に対して平方根（$y = x^2 \leftrightarrow x = \sqrt{y}$）

などがある。

参考 逆演算と逆函数 (inverse function) は異なった概念である。

問 2–8 対数
次の式を計算せよ。

1 $\log_5 125$
2 $\log_3 243$
3 $\log_7 2401$

2.2.3 よく使われる対数

対数の底はどんな数でも良いのだが、2, 10, e がよく用いられる。

参考 対数の底とは、$p = \log_a x$ で、a に相当する数のことである。

2.2.3.1 二進対数

2 を底とする対数のことを**二進対数** (binary logarithm) と呼ぶ。エントロピーや相互情報量の計算でしばしば用いられる。

参考 $2^{10} = 1024$ であるので、$\log_2 1024 = 10$ となる。つまり、$\log_2 1000$ もおよそ 10 になる。要するに、1000 倍になるごとに、二進対数はおよそ 10 倍になる。

2.2.3.2 常用対数

常用対数 (common logarithm) とは、10 を底とする対数のことである。

◈ 底の省略

常用対数は、頻繁に用いられるので、log 記号を使うときに、底の 10 を省略することが多い。つまり、単に $\log x$ と書けば、$\log_{10} x$ という意味になる。

問 2–9 常用対数
次の常用対数を計算せよ。
1 $\log 100$
2 $\log 1000$
3 $\log 10$
4 $\log 1$

参考 単純に言えば、ある数 x の常用対数の整数部分は、x の桁数から 1 を引いた

ものになる。例えば、850 は 3 桁の数だが、$\log 850 \approx 2.929$ となる。また、20456 は 5 桁の数だが、$\log 820456 \approx 4.311$ となっている。

2.2.3.3 自然対数

自然対数 (natural logarithm) とは、e を底とする対数のことである。自然対数は、統計処理そのものではあまり出てこないが、数学上扱いやすい対数である。

◈ **自然対数の表し方**

一般に、x の自然対数を $\ln x$ と表す。もちろん、$\log_e x$ と表しても構わない。

参考 ln は、ラテン語の *logarithmus naturalis* の略である。

2.2.4 計算機で対数を計算する方法

2.2.4.1 R

R では、`log` というコマンドで対数を計算することができる。このコマンドで、a を底とする x の対数を求めるには、`log(x, a)` と入力すればよい。以下の例では、3 を底とするときの 81 の対数（すなわち、$\log_3 81$）を求めている。

```
1 > log(81, 3)
2 [1] 4
```

なお、`log` コマンドで底を省略した場合、自然対数が返される。また、二進対数と常用対数には、それぞれ `log2`、`log10` という特別な函数が用意されている。

```
1 > log(81)
2 [1] 4.394449
3 > log2(81)
4 [1] 6.33985
5 > log10(81)
6 [1] 1.908485
```

2.2.4.2 Google の検索窓

Google の検索欄を使うことでも、常用対数・自然対数を求めることができる。

◈ **常用対数の求め方**

540 の常用対数を求めたい場合、Google の検索欄に、`log(540)` という文字列を入れて検索すると、`log(540) = 2.73239376` という結果を返す。

◈ **自然対数の求め方**

302 の自然対数を求めたい場合、Google の検索欄に、`ln(302)` という文字列を入れて検索すると、`ln(302) = 5.71042702` という結果を返す。

問 2–10　計算機を用いた対数の計算

計算機を用いて次の式を計算せよ。

1. $\log_2 74$
2. $\log_2 2983$
3. $\log_2 382729$
4. $\log_{10} 3579$
5. $\log_{10} 4237$
6. $\log_{10} 92135$
7. $\log_e 36$
8. $\log_e 450$
9. $\log_e 354357$

2.2.5　対数変換

データの数値を対数の形に書き換えることを**対数変換**と呼ぶ。多くの場合、常用対数を用いる。

例えば、1,000 と 100,000 をそれぞれ常用対数で対数変換すると、3 と 5 になる。なぜならば、$\log_{10} 1000 = 3, \log_{10} 1000000 = 5$ だからである。キリが良い数でなくても、例えば、5823 なら、$\log_{10} 5823 = 3.765$ なので、3.765 が常用対数で対数変換した結果になる。

言語研究で良くあるのが、単語の出現頻度を対数に変換する例だろう。普通ののコーパスでは、上位の単語の頻度と下位の単語の頻度は極端に違う。例えば、1 番よく使われている “the” は 32 万回出現しているのに、1200 番目によく使われている単語は、1000 回しか出現しないといったように、桁からして違うということがある。対数変換を行うと、32 万の常用対数はお

よそ 5.5 であり、1000 の常用対数は 3 になり、極端に違うという印象はだいぶ薄くなる。

2.2.6 対数でかけ算を足し算にする

足し算とかけ算を比べると、足し算の方が簡単である。このため、色々と楽をするために、かけ算を足し算に変換できるとうれしい。実は、対数を使うと、かけ算を足し算にすることができる。

積の対数には、

$$\log xy = \log x + \log y$$

という性質がある。具体的な例でこのことを確かめてみよう。

> **問 2–11　対数の基本公式の確認**
> $\log_{10} 100 \cdot 1000 = \log_{10} 100 + \log_{10} 1000$ となることを確かめよ。まず、$100 \cdot 1000$ がいくつになるか計算し、その常用対数がいくつかを求めることで左辺の値を確認せよ。次に、右辺の各項、すなわち、$\log_{10} 100$ と $\log_{10} 1000$ を計算し、その和を求めることで右辺の値を確認せよ。

結局、$\log xy = \log x + \log y$ という性質があるので、x と y のかけ算の対数をとると、$\log x$ と $\log y$ の足し算に変形できることが分かる。この意味で、対数をとって、かけ算を足し算にできるのである。

統計に限らず、数式を処理する際には**対数をとって、かけ算を足し算にする操作**がよく行われる。統計の教科書で数式の羅列が続いているときに、この操作が何の注意もなしに行われることがある。そういったときに、「ああ、ここは対数をとって、ややこしいかけ算を分かりやすい足し算にしているんだな」と気づけるようになれば良い。

対数をとってかけ算を足し算にするテクニックは、「x, y という 2 つの正の変数があるとき、xy の最大値を求めなさい」といった問題で使える。x, y ともに正なので、xy の最大値を求めることは、$\log xy$ の最大値を求めることと同じことである。$\log xy = \log x + \log y$ であるから、結局この問題は、$\log x + \log y$ の最大値を求めること同じということになる。かけ算の最大値

を求めることよりも、足し算の最大値を求める方が楽なので、問題が解きやすくなる。

参考 今の段階では、対数をとってかけ算から足し算にする操作について、具体的なイメージが想起しにくいかもしれない。後に触れる最尤推定の時などに、この操作が出てくるので、その際にこのことを思い出せれば良いだろう。

2.3 比率

全体の数が異なるデータを比べる際には、**比率**をもって比べるのが便利である。例えば、あるデータで 10 回、別のデータでは 20 回出現した場合、後者のほうがよく出現していると言えるだろうか。必ずしもそうではない。100 回中 10 回と、1000 回中 20 回だったら、前者の方がよく出現していると言えそうである。なぜならば、前者における出現比率が 0.1 であるのに対し、後者における出現比率は 0.02 であるからである。

ただし、比率はそのままでは統計的に取り扱い場合がある。比率の取り扱いにくさに対処する方法として、**逆正弦変換**という手法を紹介する。

2.3.1 比率の報告

論文などで比率を報告するときは、比率だけを報告しないようにし、何回中何回であったと書くようにする。例えば、「調査した結果、80% の用例で A 構文が用いられており、残りはすべて B 構文が用いられていた」と書くのは良くない。極端な話、5 個しか用例を見ず、そのうち 4 回がたまたま A 構文であった可能性も出てきてしまう。この場合は、「今回の調査対象となった 1026 の用例のうち、821 例 (80%) において A 構文が用いられており、残りの 205 例 (20%) においてはすべて B 構文が用いられていた」といったように、実数も書くべきである。

2.3.2 逆正弦変換

比率の形で表されたデータはそのままでは取り扱いにくいので、**逆正弦変換** (arcsine transformation) を行うことがある。この変換は、角変換 (angular transformation) とも呼ばれる。

逆正弦変換を行うには、まずもともとの比率の平方根を求め、出てきた値の逆正弦を求めればよい。つまり、比率 p に対して、逆正弦変換を行った値は $\arcsin \sqrt{p}$ となる。

参考 ここでは、逆正弦を求める際に、ラジアンに基づいて計算した。度に基づいて計算してもよい。

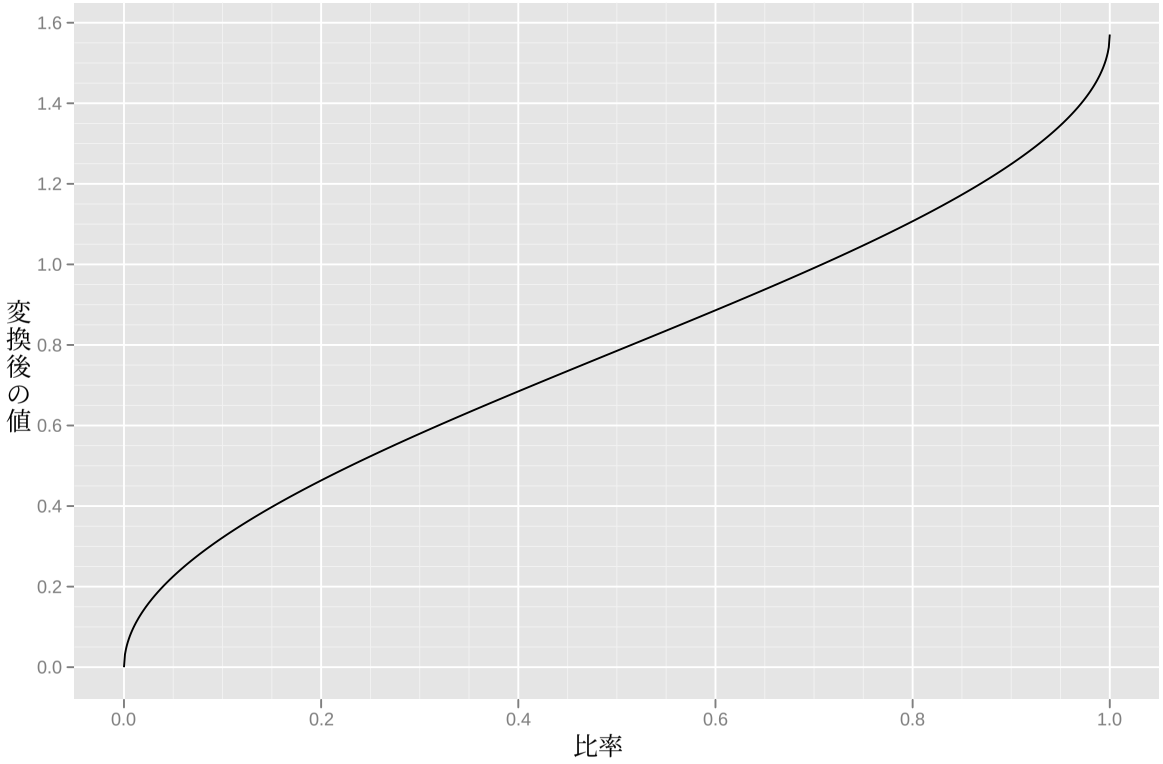

図 2.1　逆正弦変換後の値のグラフ

表 2.3　逆正弦変換の数表

比率 p	逆正弦変換の結果 $\arcsin\sqrt{p}$	比率 p	逆正弦変換の結果 $\arcsin\sqrt{p}$
		0.05	0.2255134
0.10	0.3217506	0.15	0.3976994
0.20	0.4636476	0.25	0.5235988
0.30	0.5796397	0.35	0.6330518
0.40	0.6847192	0.45	0.7353145
0.50	0.7853982	0.55	0.8354819
0.60	0.8860771	0.65	0.9377445
0.70	0.9911566	0.75	1.0471976
0.80	1.1071487	0.85	1.1730969
0.90	1.2490458	0.95	1.3452829

逆正弦変換は、比率の値のみに依存し、全体のデータの数に依存しない。例えば、10 人中 5 人でも、10000 人中 5000 人でも、比率はどちらも 0.5 (=50%) なので、逆正弦変換の結果はおよそ 0.7853982 で同じである。

2.3.2.1 比率が 0 か 1 の場合

比率が、0 か 1（すなわち 0% か 100%）の場合はそのまま逆正弦変換はせず、補正を行う必要がある。この補正は全体のデータ数に基づいて行われる。このため、先に逆正弦変換は全体のデータ数に依存しないと言ったが、比率が 0 か 1 の場合は、全体のデータ数に依存することになる。

ここで全体のデータ数を n とする。本来の比率が 0(=0%) の場合は、$\frac{1}{4n}$ を補正後の比率とする。本来の比率が 1(=100%) の場合は、$1-\frac{1}{4n}$ を補正後の比率とする。

例えば、10 人中 0 人ならば、$\frac{1}{4\cdot 10} = 0.025$ が補正後の比率となり、これに逆正弦変換をかけることで、0.1587802 を得る。20 人中 20 人ならば、$1-\frac{1}{4\cdot 20} = 0.9875$ が補正後の比率となり、これに逆正弦変換をかけることで、1.458759 を得る。

参考　ここでは有効数字を厳密に考えていない。

2.3.2.2 逆正弦変換を行う理由

2 つのものから選択する場合、比率が p だとしたら、$\frac{p(1-p)}{n}$ が分散となってしまう。つまり、p が異なれば分散も変わってしまう。逆正弦変換をすることで、分散が等しくないのを改善することができる。

2.3.2.3 逆正弦変換の計算方法

R では、例えば、`asin(sqrt(0.5))` のように入力すると、0.5(=50%) に対し逆正弦変換を行った結果として、`0.7853982` という値が返される。

Excel では、`=asin(sqrt(0.5))` のように関数を入力すれば、逆正弦変換できる。

Google の検索欄に、`arcsin(sqrt(0.5))` という文字列を入れて検索すると、`arcsin(sqrt(0.5)) = 0.785398163` という結果を返す。

問 2–12　逆正弦変換

ある人が、異なった試験を 9 つ受けた。以下の表は、それらの試験の結果である。各試験での正答率を計算し、それを逆正弦変換せよ。

問題数	100	50	18	50	200	25	100	20	7
正答数	82	26	18	47	164	13	64	12	0

2.4

総和

総和 (summation) とは与えられた数をすべて足しあわせることである。総和の記号としては、$\sum$（大文字のシグマ）が用いられる。

統計では、多数のデータに対して、1 つ 1 つ同じ操作を行うといったようなことを頻繁に行う。この「同じ操作」をデータの数だけ書くのはあまりにも大変なので、「同じ操作」は 1 回書くだけで済ませたくなる。総和の記号を使うと、こういった「同じ操作」を簡単に表すことができる。例えば、平均の定義（定義**??**）にも総和が含まれている。このため、総和の操作に慣れておくと、統計の理解に役立つだろう。

2.4.1 総和の記号

先に述べたように、総和の記号としては、$\sum$ を用いる。

総和記号 $\sum$ を使った例を 1 つ挙げる。

$$\sum_{i=3}^{7} 2i = 2\cdot 3 + 2\cdot 4 + 2\cdot 5 + 2\cdot 6 + 2\cdot 7$$

総和の記号は、$\sum$ の下、上、右の順で読んでいくと分かりやすい。この例は、i を 3 から 7 まで（1 つずつ）増やしていったときの $2i$ をすべて足しあわせたものという意味になる。ここで、i のことを**インデックス** (index) と呼ぶ。

参考 インデックスを表す文字としては、i がもっともよく使われる。もし i が先に用いられていた場合は、j や k が用いられる。

もちろん文字を使って、さらに抽象的に書くこともできる。

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n$$

参考 上式での n は任意の自然数を代表している。ここに限らず、単に n と書いた場合、1 つの自然数を指していることが多い。なお、統計では、しばしばデータの個数を n で表す。

問 2–13　**総和**

次の総和を、上の例に従い、総和記号 $\sum$ を用いない形で表せ。最後まで計算する必要はない。

1. $\sum_{i=1}^{6} 3i$
2. $\sum_{i=2}^{8} i^2$
3. $\sum_{i=5}^{9} 2i^2 + 3i$

2.4.1.1 添字の省略

i を 1 から始めて n まで足していくのはあまりによく行われることなので、$\sum$ 記号の上下の添字はしばしば省略される。また、下側のインデックスだけを残す場合もある。

$$\sum_{i=1}^{n} X_i = \sum_{i} X_i = \sum X_i$$

2.4.2 総和の性質

2.4.2.1 総和の分解

$\sum$ の右側の式の項は分けることができる。

$$\sum_{i=s}^{t} [a_i + b_i] = \sum_{i=s}^{t} a_i + \sum_{i=s}^{t} b_i$$

マイナスでつながれていてもプラスでつながれている場合と同様である。

$$\sum_{i=s}^{t} [a_i - b_i] = \sum_{i=s}^{t} a_i - \sum_{i=s}^{t} b_i$$

例えば、以下のように分解できる。

$$\sum_{i=1}^{5} \left[i^2 + i \right] = \sum_{i=1}^{5} i^2 + \sum_{i=1}^{5} i$$

2.4.2.2 総和の中の定数倍

$\sum$ の右側の式にある定数倍 c は $\sum$ の外に出すことができる。

$$\sum_{i=s}^{t} cX_i = c\sum_{i=s}^{t} X_i$$

参考 ここで、c を「定数」と呼んでいる理由は、i の値によって変動しないからである。

例えば、以下では、定数倍の 5 を取り出している。

$$\sum_{i=7}^{20} 5i^2 = 5\sum_{i=7}^{20} i^2$$

問 2–14　総和の分解

上記の例に従い、総和記号 $\sum$ を分けよ。

1. $\sum(i^5 + i^3)$
2. $3\sum i^3 - 4\sum i^2 + 5\sum i$
3. $7\sum i^4 + \sum 2i^2 - 2\sum 3i$
4. $4\sum\left[2i^5 + 3i^3 - 2i^2\right] - 2\sum 3i$

問 2–15　総和の合成

総和記号 $\sum$ が 1 つしかない形に改めよ。

1. $\sum i^4 + \sum i^3$
2. $3\sum i^3 - 4\sum i^2 + 5\sum i$
3. $7\sum i^4 + \sum 2i^2 - 2\sum 3i$
4. $4\sum\left[2i^5 + 3i^3 - 2i^2\right] - 2\sum 3i$

2.4.3　総和の公式

総和の計算を簡単にする公式をいくつか紹介する。

2.4.3.1 定数の総和

まずは、定数を繰り返し足す操作について見てみよう。

$$\sum_{i=1}^{n} c = nc$$

上記の操作は、定数 c を n 回足しあわせることに他ならず、かけ算になる。

参考 c の中には、インデックスが含まれていない。

以下の式は、定数である 3 を $i = 1 \ldots 5$ までの 5 回足しあわせることを示している。

$$\sum_{i=1}^{5} 3 = \underbrace{3+3+3+3+3}_{5} = 5 \cdot 3 = 15$$

2.4.3.2 n までの整数の総和

1 から n までの整数を 1 つずつすべて足しあわせた数は以下の公式で求められる。

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

1 から 10 までの整数の総和、すなわち $1+2+3+\cdots+9+10$ は、以下のように求められる。

$$\sum_{i=1}^{10} i = \frac{10 \cdot (10+1)}{2} = \frac{110}{2} = 55$$

1 から始まる場合でなくても、以下のように工夫することで簡単に計算することができる。例えば、11 から 20 までの整数の総和は、1 から 20 までの整数の総和から、1 から 10 までの整数の総和を差し引くことで求めることができる。

$$\sum_{i=11}^{20} i = \sum_{i=1}^{20} i - \sum_{i=1}^{10} i = \frac{20 \cdot (20+1)}{2} - \frac{10 \cdot (10+1)}{2} = 210 - 55 = 155$$

なお、一般に、t から s までの整数をすべて足しあわせた数は以下の公式で求められる。

$$\sum_{i=t}^{s} i = \frac{(s-t+1)(s+t)}{2}$$

2.4.3.3 平方の総和

1 から n までの整数をそれぞれ平方して、すべて足しあわせた数は以下の公式で求められる。

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

参考 **平方**とは 2 乗のことを示す。

1 から 5 までの平方の総和は以下のように求められる。

$$\sum_{i=1}^{5} i^2 = \frac{5(5+1)(2\cdot 5+1)}{6} = 55$$

2.4.3.4 立方の総和

1 から n までの整数をそれぞれ立方して、すべて足しあわせた数は以下の公式で求められる。

$$\sum_{i=1}^{n} i^3 = \left(\frac{n(n+1)}{2}\right)^2$$

参考 **立方**とは 3 乗のことを示す。

問 2–16　総和の公式

上記の公式に基づき、以下の式を計算せよ。

1. $\sum_{i=1}^{23} 7$
2. $\sum_{i=1}^{10} i$
3. $\sum_{i=1}^{10} i^2$
4. $\sum_{i=1}^{10} \left[i^2 + i\right]$
5. $\sum_{i=1}^{8} \left[3i^2 + 4i\right]$
6. $\sum_{i=1}^{4} \left[4i^3 - 2i^2 + 9i + 4\right]$

2.4.3.5 調和数

1 から n までの整数の逆数の総和 H_n を、n 番目の調和数 (harmonic number) と呼ぶ。すなわち、

$$\begin{aligned} H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \\ &= \sum_{k=1}^{n} \frac{1}{k} \end{aligned}$$

のように定義される。

$$\begin{aligned} H_n &= \int_0^1 \frac{1-x^n}{1-x}\, dx \\ &= \int_0^1 \frac{1-(1-u)^n}{u}\, du \\ &= \int_0^1 \left[\sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} u^{k-1} \right] du \\ &= \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} \int_0^1 u^{k-1}\, du \\ &= \sum_{k=1}^{n} (-1)^{k-1} \frac{1}{k} \binom{n}{k} \end{aligned}$$

2.4.4 R での総和

R で総和を求めるには、`sum` コマンドを使えばよい。また、行列・リストなどで、各系列ごとに総和を求めるには `mapply(sum, data)` のようにすればよい。

```
> sum(7, 20, 6, 4)
[1] 37

> data <- c(list(1:10), list(11:20), list(21:30))
> data
[[1]]
 [1]  1  2  3  4  5  6  7  8  9 10

[[2]]
 [1] 11 12 13 14 15 16 17 18 19 20

[[3]]
```

```
 [1] 21 22 23 24 25 26 27 28 29 30

> mapply(sum, data)
[1]  55 155 255
```

平方の総和や逆数の総和が必要なときは、sum(x^2) や sum(1/x) とすれば良い。また、以下のようにすれば、行列・リストなどで、各系列ごとに平方や逆数の総和を求めることができる。

```
> data <- c(list(1:10), list(11:20), list(21:30))
> mapply(function(x) sum(x^2), data)
[1]  385 2485 6585
> mapply(function(x) sum(1/x), data)
[1] 2.9289683 0.6687714 0.3972475
```

2.4.4.1 NA や NaN がある場合

sum の引数に NA（データの欠落）や NaN（非数）が含まれている場合、NA, NaN が出力されてしまう。

```
> sum(c(5, 3, 2))
[1] 10
> sum(c(5, 3, 2, NaN))
[1] NaN
> sum(c(5, 3, 2, NA))
[1] NA
# NA is stronger than NaN.
> sum(c(5, 3, 2, NaN, NA))
[1] NA
```

NA や NaN を無視した方が良い場合は、na.rm=TRUE オプションを使う。このオプションを使うと、NA, NaN を外して計算する。

```
> sum(c(5, 3, 2, NaN), na.rm=TRUE)
[1] 10
> sum(c(5, 3, 2, NA), na.rm=TRUE)
[1] 10
> sum(c(5, 3, 2, NaN, NA), na.rm=TRUE)
[1] 10
```

なお、当然のことながら、Inf を含む場合は、以下のようになる。

```
> sum(c(5, 3, 2, Inf))
[1] Inf
> sum(c(5, 3, 2, -Inf, Inf))
[1] NaN
```

na.rm=TRUE オプションもここでは関係ない。

```
> sum(c(5, 3, Inf), na.rm=TRUE)
[1] Inf
> sum(c(5, 3, -Inf, Inf), na.rm=TRUE)
[1] NaN
```

2.4.4.2 論理値の総和

R では、論理値は自動的に数値として扱われるので、sum は論理値にも適用できる。

```
> sum(c(TRUE, TRUE, FALSE, TRUE, FALSE))
[1] 3
> sum(TRUE, 5)
[1] 6
> sum(FALSE, 5)
[1] 5
```

参考 数値として扱われるとき、TRUE は 1、FALSE は 0 に置き換えられる。

このことを利用して条件に当てはまるものの個数を調べることができる。

```
> x <- c(6, 4, 3, 7, 8)
> x < 5
[1] FALSE  TRUE  TRUE FALSE FALSE
> sum(x < 5)
[1] 2
```

最後のコマンド sum(x < 5) は、ベクトル x の要素の中で、5 より小さいものがいくつあるかを調べ、2 個あるという結果を得ている。

2.4.5 総和記号の複数利用

総和記号は、複数重ね合わせて使うことができる。以下の例では、$\sum_{j=1}^{n}$ の内側に、$\sum_{i=1}^{m}$ が埋め込まれている。

$$\sum_{j=1}^{n}\sum_{i=1}^{m}\left[3i^2+2ij-4j^2\right]$$

普通は、内側の総和記号のインデックスの方がアルファベット順で先になるようにする。つまり、内側から、i, j, k とインデックスを振っていく。よっ

て、インデックスが省略されている場合は、次のように復元することができる。

$$\sum\sum\sum X_{ijk} = \sum_{k}\sum_{j}\sum_{i} X_{ijk}$$

総和記号を複数利用している場合でも、以下のように総和記号を用いない形で表すことができる。

$$\begin{aligned}
&\sum_{j=1}^{3}\sum_{i=5}^{10} ij \\
&= \sum_{j=1}^{3} 5j + \sum_{j=1}^{3} 6j + \sum_{j=1}^{3} 7j + \sum_{j=1}^{3} 8j + \sum_{j=1}^{3} 9j + \sum_{j=1}^{3} 10j \\
&= (5\cdot 1 + 5\cdot 2 + 5\cdot 3) + (6\cdot 1 + 6\cdot 2 + 6\cdot 3) \\
&\quad + (7\cdot 1 + 7\cdot 2 + 7\cdot 3) + (8\cdot 1 + 8\cdot 2 + 8\cdot 3) \\
&\quad + (9\cdot 1 + 9\cdot 2 + 9\cdot 3) + (10\cdot 1 + 10\cdot 2 + 10\cdot 3)
\end{aligned}$$

問 2–17　複数の総和記号

次の総和を、総和記号 $\sum$ を用いない形で表せ。

1. $\displaystyle\sum_{j=3}^{4}\sum_{i=1}^{6} 2ij$
2. $\displaystyle\sum_{j=2}^{5}\sum_{i=1}^{3} i^2 + j^2$
3. $\displaystyle\sum_{k=1}^{2}\sum_{j=12}^{17}\sum_{i=1}^{2} 3ijk$

総和記号の複数使用と複数のグループ

いくつかのグループに分けられるデータに対して用いられるとき、総和記号が複数用いられることが多い。まず、グループに分けられないデータとして、n 人の大学生に英語の試験を行った時の得点を考える（表 2.4）。この表では、1 番目の受験者の得点が X_1、2 番目の受験者の得点が X_2、n 番目の受験者の得点が X_n のように表されている。ここで、全受験者の得点の合計は、$\displaystyle\sum_{i=1}^{n} X_i$ とい

う形で、総和記号を 1 つだけ用いて表すことができる。

しかし、普通はデータを集めるときは、いくつかのグループに分けて考えた方が、グループ間の比較などもできて有用である。グループに分けられるデータとして、大学生に英語の試験を行った結果を所属学部ごとに分けた例を考えよう（表 2.5）。なお、各学部とも受験者は n 人ずつであるとする。表 2.5 では、法学部の受験者の 1 番目の人の得点が X_{11}、法学部の受験者の 2 番目の人の得点が X_{21} のように表されている。これに対して、文学部の受験者の 1 番目の人の得点が X_{12}、理学部の受験者の 1 番目の人の得点が X_{12} のように表されている。抽象化して言うと、表 2.5 のデータは、X_{ij} の形で表すことができる。ここで、j がどのグループに属しているかを表し、i がグループの中で何番目のデータ化と言うことを示していることになる。ここのデータでは、$j = 1$ ならば法学部、$j = 2$ ならば文学部、$j = 3$ ならば理学部ということになる。そして、このデータにおいては、全受験者の得点の合計は、$\sum_{j=1}^{3}\sum_{i=1}^{n} X_{ij}$ という形で、総和記号を 2 つ用いて表すことができる。

表 2.4　1 次元データ

得点	$X_1 X_2 X_3 X_4 \cdots X_{n-1} X_n$

表 2.5　2 次元データ

法学部の得点	$X_{11} X_{21} X_{31} X_{41} \cdots X_{n-1\,1} X_{n1}$
文学部の得点	$X_{12} X_{22} X_{32} X_{42} \cdots X_{n-1\,2} X_{n2}$
理学部の得点	$X_{13} X_{23} X_{33} X_{43} \cdots X_{n-1\,3} X_{n3}$

2.5

組み合わせ論

統計学の基礎として重要なものの 1 つとして、確率論がある。確率については、次の 2.6 節で詳しく見る。この節では、確率に入る前に、まず確率の基礎付けとなる組み合わせ論、特に条件に当てはまるものがどれだけあるか数え上げる方法について見る。この節では数え上げる方法として、階乗・順列・組み合わせを紹介する。

どういう組み合わせがあり得るかは、実験を用いた研究で重要な観点となってくる。例えば、被験者をどういう順番で並べるか、実験に用いる言語材料をどう組み合わせるかといったことを考える際、組み合わせ論の考え方が必要になってくる。

2.5.1 階乗

n の階乗 $n!$ は、1 から n までの自然数を全てかけあわせた値として定義される。

定義 5 階乗

$$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot \cdots \cdot (n-1) \cdot n$$

なお、$n = 0$ のときは、$0! = 1$ と定義する。

参考 なぜ $0! = 1$ のように規定されているのだろうか？ すぐ後で見るように、階乗は n 個のものを全て使って並べる場合の並べ方の総数に等しい。つまり、0 の階乗は、「0 個のものを並べる」場合の並べ方の総数ということになる。「0 個のものを並べる」というのは、「何も並べない」という並べ方しかあり得ない。 よって、「0 個のものを並べる」方法は 1 つしかないので、$0! = 1$ となることが正当化される。また、0 の階乗を定義しないと、いろいろなところで例外を考える必要が出てきて不便になる。

0 から 10 までの階乗の値は以下の通りである。一見して分かるように、急速な勢いで値が大きくなっている。

$$
\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \\
3! &= 6 \\
4! &= 24 \\
5! &= 120 \\
6! &= 720 \\
7! &= 5040 \\
8! &= 40320 \\
9! &= 362880 \\
10! &= 3628800
\end{aligned}
$$

参考 興味深いことに、階乗の逆数の総和は、自然対数の底と等しくなる。

$$\sum_{n=0}^{\infty} \frac{1}{n!} = 2.71828182845904523536028747... = e$$

2.5.2 順列

順列 (permutation) とは、ある集合（ものの集まり）から、一定の数の要素を取り出し、なおかつ順番に意味を持たせる場合の数である。

n 個のものから、r 個のものを順番を考えて取り出す場合の数である**順列** ${}_n\mathrm{P}_r$ は、以下のように定義される。ただし、$n \geq r$ である tosuru 。

定義 6 順列

$$ {}_n\mathrm{P}_r = \underbrace{n(n-1)(n-2)\cdots(n-r+1)}_{r} = \frac{n!}{(n-r)!} $$

例えば、アン (A)・ビル (B)・チャーリー (C)・ドロシー (D) の 4 人がいて、4 人の中から 3 人を呼び出して個別に面接を行う場合の数を考えよう。順序も考慮に入れると、24 通りになる。その組み合わせは以下のとおりである。

A-B-C / A-B-D / A-C-B / A-C-D / A-D-B / A-D-C
B-A-C / B-A-D / B-C-A / B-C-D / B-D-A / B-D-C

C-A-B / C-A-D / C-B-A / C-B-D / C-D-A / C-D-B
D-A-B / D-A-C / D-B-A / D-B-C / D-C-A / D-C-B

また、$n = 4, r = 3$ なので、${}_4\mathrm{P}_3 = \underbrace{4 \cdot 3 \cdot 2}_{3\text{ 個}} = 24$ となっており、上記の結果と一致する。

2.5.2.1 0 がかかわる場合

$n = 0$ あるいは $r = 0$ の場合も、定義 6 に基づいて計算すれば、きちんと考えることができる。一般に、

$$
\begin{aligned}
{}_n\mathrm{P}_0 &= 1 \\
{}_0\mathrm{P}_0 &= 1
\end{aligned}
$$

が成り立つ。

2.5.2.2 すべてを使った順列

n 個のものすべてを使って順列を作るときの場合の数はどうなるだろうか？ すべてを使うということは、n 個のものから n 個を選ぶということなので、順列の定義より、以下の等式が成り立つ。

$$
{}_n\mathrm{P}_n = n!
$$

問 2–18　順列

定義 6 に基づき、以下の式を計算せよ。

1. ${}_{10}\mathrm{P}_5$
2. ${}_8\mathrm{P}_4$
3. ${}_7\mathrm{P}_2$
4. ${}_{23}\mathrm{P}_0$

問 2–19　順列に基づく場合の数の計算

以下の場合の数を求めよ。

1. 赤玉、青玉、緑玉、黒玉、白玉がそれぞれ 1 つずつある。ここから、3 つの玉を取り出して並べるとき、その並べ

方は何通り存在するか。

2 25 人の中から、準備をする人 1 人、またそれとは別に片付けをする人 1 人を選びたい。考えられる人選は何通りあるか。

3 今、10 人の被験者がいる。これから行う実験は 1 人ずつでしか参加できないので、10 人の順番を考える必要がある。この時、10 人の順番として考えられるのは何通りあるか。

2.5.3 組み合わせ

組み合わせ (combination) とは、ある集合（ものの集まり）から、一定の数の要素を取り出したときの場合の数である。順列と違って、順番には意味がない。

n 個のものから、r 個のもの（ただし、$n \geq r$ である）を順番を考えずに取り出す場合の数である**組み合わせ** $\binom{n}{r}$ は、以下のように定義される。

定義 7 組み合わせ

$$\binom{n}{r} = \frac{{}_n\mathrm{P}_r}{r!} = \frac{n!}{r!(n-r)!} = \frac{\overbrace{n(n-1)\cdots(n-r+1)}^{n\text{ 個}}}{\underbrace{r(r-1)\cdots 1}_{\mathrm{r}\text{ 個}}}$$

参考 組み合わせの記号は、$\binom{n}{r}$ と書くだけでなく、${}_n\mathrm{C}_r$ という書き方もある。

参考 組み合わせでは、順列と違って、順番を考えないので、A-B-C、A-C-B、B-A-C、B-C-A、C-A-B、C-B-A は全く同じことになる。

例えば、アン (A)・ビル (B)・チャーリー (C)・ドロシー (D) の 4 人がいる場合を考えよう。このとき、4 人の中から 3 人を同時に呼び出す場合の数は、4 通りになる。その組み合わせは以下のとおりである。

A-B-C / A-B-D / A-C-D / B-C-D

また、$n=4, r=3$ なので、$\binom{4}{3} = \overbrace{\frac{4\cdot 3\cdot 2}{3\cdot 2\cdot 1}}^{3\text{ 個}}{}_{3\text{ 個}} = 4$ となり、上記の結果と一致する。

2.5.3.1 組み合わせの性質

◈ 選び出さない場合の数

n 個のものから r 個だけ選び出す場合の数は、$n-r$ 個だけ選び出さない場合の数と同じである。すなわち、

$$\binom{n}{r} = \binom{n}{n-r}$$

が成り立つ。

例えば、アン・ビル・チャーリー・ドロシーの 4 人がおり、このうち 3 人を選び出す場合の数を考えよう。4 人の内、3 人が選び出されるのであるから、逆に言うと、$4-3=1$ 人だけ選び出されないことになる。選び出されない場合は、アンが選ばれない場合、ビルが選ばれない場合、チャーリーが選ばれない場合、ドロシーが選ばれない場合の 4 通りである。ここで、

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!1!} = 4$$
$$\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4!}{1!3!} = 4$$

となっていることに注意せよ。

◈ 特殊な r の値

r の値が、0 や 1 となる場合や、$r=n$ となる場合は簡単に計算することができる。

- $\binom{n}{n} = 1$

参考 $\binom{n}{n}$ は全部取り出す場合なので、「全部」という 1 通りしかない。

- $\binom{n}{1} = n$
- $\binom{n}{0} = 1$

参考 $\binom{n}{0}$ というのは何も取り出さない場合である。

問 2–20　組み合わせ

以下の式を計算せよ。

1. $\binom{10}{5}$
2. $\binom{8}{4}$
3. $\binom{48}{6}$
4. $\binom{32}{1}$
5. $\binom{23}{0}$
6. $\binom{81}{79}$

問 2–21　組み合わせに基づく場合の数の計算

以下の場合の数を求めよ。

1. 赤玉、青玉、緑玉、黒玉、白玉がそれぞれ 1 つずつある。ここから、3 つの玉を取り出し方の組み合わせは何通り存在するか。
2. ある実験を 32 人に対して行った。さらに脳波を測定しようと考えているが、機材の都合上、3 人に対してしか脳波測定をすることができない。脳波測定を受ける被験者の組み合わせは何通り存在するか。

問 2–22　順列と組み合わせの比較

ある人工言語 L_1 には、5 の単語があり、それらの単語を組合わせることで文を作る。ただし、この言語には語順というものがなく、単語をどのように並べても、使っている単語の組み合わせさえ同じであれば、意味は全く同じである。これに対して、人工言語 L_2 は、L_1 と同様に、5 の単語があり、それらの単語を組み合わせることで文を作る。しかし、L_2 には語順の考えがあり、たとえ同じ単語の組み合わせであったとしても、語順が 1 箇所でも異なると意味が変ってしまう。なお、L_1, L_2 ともに、文の長さは 1 から 5 単語までで、1 つの文の中では同じ単語を重複して使え

ないものとする。

1. L_1 の文は最大何通りの意味を表しうるか。
2. L_2 の文は最大何通りの意味を表しうるか。
3. ここから順列と組合せの数の大きさの違いについて簡単に考察せよ。

問 2–23　順列と組み合わせとの間の大小関係

任意の非負整数 n, r について、$n \geq r$ とするとき、

$$\binom{n}{r} \leq {}_n\mathrm{P}_r \leq n^r$$

となることを確かめよ。また等号が成り立つのはどのような場合か。なお、上記の不等式は、n, r が与えられたとき、組み合わせの数が最も少なく、重複を許さない順列の数がそれに次ぎ、重複を許す順列の数が最も多いということを示している。

問 2–24　文字の並び替え

片仮名が 1 つだけ書かれているようなカードを考えよう。こうしたカードを複数持ってきた場合、それら全てを使って並べた文字列の数がいくつになるかを考えよう。例えば、「シ」「ミ」「ロ」というカードが与えられた場合、あり得る並びには、「シミロ」・「シロミ」・「ミシロ」・「ミロシ」・「ロシミ」・「ロミシ」の合わせて 6 通りが存在する。なお、1 枚のカードを複数回使うことはできないものとする。先ほどの例では「シシロ」や「ミロミ」などは許されない。ただし、最初に与えられたカードに元から重複があった場合、

1. 「ア」「オ」「サ」「シ」という 4 枚のカードが与えられたとき、これらを全て使って並べた文字列は何種類になるか。

2「ス」「モ」「モ」という 3 枚のカードが与えられたとき、これらを全て使って並べた文字列は何種類になるか。

3「イ」「イ」「カ」「カ」「キ」「キ」という 6 枚のカードが与えられたとき、これらを全て使って並べた文字列は何種類になるか。

4「ン」と書かれたカードは文字列の先頭に来られないものとする。例えば、「ンカロ」という文字列は許されない。このとき、「シ」「ミ」「ン」という 3 枚のカードを全て使って並べた文字列は何種類になるか。

5「ン」と書かれたカードは文字列の先頭に来られないだけでなく、連続することもないとする。例えば、「カンン」という文字列は許されない。このとき、「ア」「カ」「コ」「ン」「ン」という 5 枚のカードを全て使って並べた文字列は何種類になるか。

2.5.4 R で組み合わせなどを計算する方法

2.5.4.1 階乗

R では、`factorial` というコマンドで階乗を計算することができる。

```
> factorial(5)
[1] 120
```

参考 大きな数の階乗を求めようとすると、以下のように “`value out of range in 'gammafn'`” というエラーが出る。R では階乗を計算するときに、Γ 函数と呼ばれる函数を用いているのだが、このエラーは、あまりにも数が大きすぎて Γ 函数の計算範囲を超えていることを示している。

```
> factorial(171)
[1] Inf
Warning message:
In factorial(171) : value out of range in 'gammafn'
```

2.5.4.2 順列

R には、順列を直接計算する函数はない。このため、順列を求める必要がある場合は、自分で函数を組み合わせて計算する必要がある。以下の例では、総乗の函数 prod を用いるのがよいだろう。${}_nP_r$ を計算したい場合、prod((n-r+1):n) とすればよい（ただし、この方法では $r = 0$ のときうまくいかない）。${}_5P_3$ を求めたい場合は、以下のようにすればよい。

```
> n <- 5
> r <- 3
> prod((n-r+1):n)
[1] 60
```

なお、e1071 というパッケージに含まれる permutations という函数は、引数にとった数までの順列のパターンをすべて表示してくれる。例えば、以下のように、permutations(3) とすることで、1 から 3 までの数をすべて 1 回ずつ使った場合、どういう並べ方があるかを教えてくれる。

```
> library(e1071)
> permutations(3)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    1    3
[3,]    2    3    1
[4,]    1    3    2
[5,]    3    1    2
[6,]    3    2    1
```

2.5.4.3 組み合わせ

組み合わせの数を求めるには、choose という函数を用いる。$\binom{n}{k}$ を求めたければ、choose(n, k) と R でコマンドを入力すればよい。例えば、以下では、$\binom{5}{3}$ を計算している。

```
> choose(5, 3)
[1] 10
```

なお、combn という函数を用いると、ありうる組み合わせのすべてを行列の形で表示させることができる。以下の例では、4 人の中から 3 人を選ぶ組み合わせを作成している。各列が 1 つの組み合わせに対応している。

```
> people <- c("Ann", "Bill", "Charlie", "Dorothy")
> combn(people, 3)
```

```
     [,1]      [,2]      [,3]      [,4]
[1,] "Ann"     "Ann"     "Ann"     "Bill"
[2,] "Bill"    "Bill"    "Charlie" "Charlie"
[3,] "Charlie" "Dorothy" "Dorothy" "Dorothy"
```

2.5.5 Google の検索欄を用いた組み合わせなどの計算

◈ 階乗

Google の検索欄で、階乗を調べたい場合は、数式をそのまま入れれば計算結果を返す。例えば、5! の値を知りたければ、単に 5! という文字列を入れて検索すれば、5 ! = 120 という結果が返ってくる。

◈ 組み合わせ

例えば、$\binom{10}{7}$ を調べたいときに、Google の検索欄に、10 choose 7 という文字列を入れて検索すると、10 choose 7 = 120 という結果を返す。

2.6

確率

Der Alte wurfelt nicht.
神はさいころをふらない

— ALBERT EINSTEIN

確率 (probability) の理論は、統計学の基礎となっているので、きちんと把握しておく必要がある。

言語現象を確率的に捉える試みも行われている。例えば、*Probabilistic Grammar* と言った用語があるぐらいである。

この節では、まず確率の素朴な定義を考えた後、経験的確率の考え方について触れ、条件付き確率や確率に関する基礎的な定理などを紹介する。

2.6.1 確率の定義

2.6.1.1 初歩的な定義

確率を厳密に定義するのは難しい。ここでは、とりあえず素朴な定義を考えるところから始めよう。

◈ 試行と事象

確率の定義の前に、いくつかの用語を定義する。

試行 (trial) とは、同条件のもとで、実験や観察を繰り返すことを言う。試行を行ったことで出てきた結果のことを**事象** (event) と呼ぶ。

参考 事象はしばしばローマ字の大文字で表される。例えば、「さいころを 1 回ふって、奇数の目が出る事象を A とし、さいころを 1 回ふって、偶数の目が出る事象」を B とする」などど書く。

◈ 根源事象

これ以上分けることのできない事象として、**根源事象** (elementary event) というものを考えることができる。

次のような事例を考えると、事象と根源事象の違いが分かりやすいだろう。さいころを 1 回ふって、5 以上の目が出る事象を A と表す。このとき、

5 以上の目は、5, 6 の 2 通りある。5 の目が出ることは根源事象であるし、6 の目が出ることも根源事象となる。

1 つの状況に着目するとき、根源事象が起こる可能性は同様に確からしい。すなわち、根源事象の起きる確率は同じということである。さいころを 1 回ふったとき、根源事象は、1 の目が出ること、2 の目が出ること、3 の目が出ること、4 の目が出ること、5 の目が出ること、6 の目が出ることである。どの目が出る可能性も同じである。

参考 事象は、試行結果の集合として捉えることができる。また、根源事象は単集合 (singleton) と見なすことができる。

◈ 確率の初歩的な定義

以上のことを踏まえ、確率のもっとも初歩的な定義をしよう。

定義 8 確率（初歩的な定義）

事象 A が起こる確率を、$P(A)$ と表し、以下のように定義する。

$$P(A) = \frac{\text{事象 } A \text{ の起こる場合の数}}{\text{全体の場合の数}}$$

参考 ここで言う場合の数は、根源事象の数を数えたものである。

2.6.1.2 確率の具体例

上記の確率の初歩的な定義（定義 8）を踏まえ、いくつかの事例を通じて、確率に関する理解を深めていこう。

◈ コイン

コインを 1 回投げて、表が出る事象を A と表す。このときの $P(A)$ について考えよう。ここで、$P(A)$ は、コインを 1 回投げたとき表が出る確率のことである。さて、コインは表が出るか裏が出るかの 2 通りしかないので、全体の場合の数は 2 通りである。事象 A が起こる場合の数は、当然 1 通りである。ゆえに、$P(A) = \frac{1}{2}$ となる。

◈ さいころ

さいころを 1 回ふって、5 以上の目が出る事象を A と表す。このときの $P(A)$ について考える。さいころは 1, 2, 3, 4, 5, 6 の 6 通りの出目がある

ので、全体の場合の数は、6 通りであると言える。5 以上の目は、5, 6 の 2 通りであるので、事象 A が起こる場合の数は、2 通りである。ゆえに、$P(A) = \frac{2}{6} = \frac{1}{3}$ となる。

◈ コインその 2

コインを 2 つ投げて、両方とも表が出る事象を A と表す。このときの $P(A)$ について考える。ここで、(表・表)、(表・裏)、(裏・表)、(裏・裏) の 4 通りの根源事象がある。よって、$P(A) = \frac{1}{4}$ となる。

参考 「2 枚とも表」、「片方が表で片方が裏」、「2 枚とも裏」の 3 通りの事象があるから、$P(A) = \frac{1}{3}$ となると考えてはならない。「片方が表で片方が裏」というのは、実は 2 つの根源事象から成り立っている。すなわち実際には (表・裏)、(裏・表) の 2 つのパターンがある。

問 2–25　簡単な確率

以下の事象の確率を求めよ。

1. さいころを 1 回ふって、2 以上の目が出る確率
2. さいころを 1 回ふって、奇数の目が出る確率
3. コインを 3 つ投げて、どれも表が出る確率

2.6.1.3 経験的確率

◈ 初歩的な確率の定義の問題点

定義 8 で示された初歩的な確率の定義には大きな問題がある。初歩的な定義は、根源事象に基づいて考えている。しかし、根源事象には、等確率ということがその前提に含まれている。つまり、確率を定義するために確率を用いなくてはならないということであり、初歩的な定義は循環的な定義となっているのである。

このため、初歩的な定義を離れ、より適切な処理が必要になる。確率論では集合論などをもとにした研究が行われているが、これは難解であるため、ここでは省略する。

◈ 確率の経験的定義

初歩的な確率の定義の代わりに、経験的確率というものを考えることができる。これは、実際に試行を繰り返すことを通じて確率を捉えるもので

ある。

定義 9 経験的確率

ある試行を行ったとき、事象 A が起こる確率を、$P(A)$ と表し、以下のように定義する。

$$P(A) = \frac{\text{事象 } A \text{ が起きた回数}}{\text{試行を行った回数}}$$

経験的確率に基づいて考える場合、コインを投げたとき表が出る確率は、実際に何度もコインを投げることで計算される。例えば、コインを 1 万回投げて、そのうち 4992 回だけ表が出たとしたら、表が出る確率は $\frac{4992}{10000} = 0.4992$ となる。

実際の言語研究で確率を算出する事例が出てきた場合、それは経験的確率を求めていることが多い。

問 2–26　経験的確率

定義 9 における経験的確率の考え方に基づき、以下の確率を求めよ。

1. あるコインを投げて表が出る確率。なお、このコインを 100 回投げたところ、55 回表が出た。
2. あるさいころを 1000 回ふったところ、出た目は以下の通りであった。この結果を踏まえ、以下の確率を求めよ。

出目	1	2	3	4	5	6
観測回数	172	184	163	169	165	147

(a) 3 の目が出る確率
(b) 4 の目が出る確率
(c) 5 以上の目が出る確率
(d) 偶数の目が出る確率

2.6.2 確率の基本的性質

2.6.2.1 確率の範囲

どんな確率でも、値は、0 以上 1 以下になる。すなわち、任意の事象 A に対して、

$$0 \leq P(A) \leq 1$$

が成り立つ。ここより、確率の最大値は 1、最小値は 0 ということになる。

- $P(A) = 1$ は、ある事象 A が必ず起きることを示す。
- $P(A) = 0$ は、ある事象 A が全く起きないことを示す。

◈ 確率の最大値・最小値の事例

さいころをふったときに 1 から 6 までの目のいずれかが出てくる事象を A としよう。また、7 以上の目が出てくる事象を B とする。特殊なさいころを使っていない限り、必ず 1 から 6 までの目のいずれかが出てくるし、7 以上の目が出てくることはありえない。よって、事象 A は必ず起きるから $P(A) = 1$ となり、事象 B は絶対に起きないので、$P(B) = 0$ となる。

問 2–27　確率の最大値・最小値

以下の事象の確率を求めよ。

1. コインを 1 つ投げて、表か裏が出る確率
2. さいころを 1 回ふって、奇数か偶数の目が出る確率
3. さいころを 1 回ふって、8 以下の目が出る確率
4. さいころを 1 回ふって、0 以下の目が出る確率

2.6.2.2 余事象

事象 A が起きない事象のことを A の**余事象** (complementary event) と呼ぶ。A の余事象は、$\overline{A}$ と表す。

参考 $\overline{A}$ を略記するために A' と書いては**ならない**。統語論では、$\overline{\mathrm{X}}$ を X' と書くことがあるが、これはあくまでも統語論での約束事であって、確率の時には関係ない。

事象 A の余事象 $\overline{A}$ の起こる確率は、以下のようにして求められる。

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

さいころを 1 回ふって、5 以上の目が出る事象を A と表す。このときの $P(\overline{A})$、すなわち 5 以上の目が出ない事象について考えてみよう。

まず、余事象の確率を求める式を用いないで、確率を考える。$\overline{A}$ とは、4 以下の目が出る場合のことであり、1, 2, 3, 4 の 4 通りの場合がある。全体の場合の数は、6 通りであったので、$P(\overline{A}) = \frac{4}{6} = \frac{2}{3}$ となる。

次に、余事象の考え方を用いて計算しよう。$P(A) = \frac{1}{3}$ だったので、先に挙げた余事象の確率を求める式に基づいて、$P(\overline{A}) = 1 - P(A) = 1 - \frac{1}{3} = \frac{2}{3}$ となることがわかる。余事象の確率を求める式を用いなかった場合と同じ結果になったことが分かるだろう。

問 2–28　余事象の確率

以下の事象の確率を求めよ。

1. さいころを 1 回ふって、3 の目が出ない確率
2. さいころを 1 回ふって、2 以下の目が出ない確率

ギャンブラーの蹉跌

あるギャンブラーが、サイコロを 4 回ふって、そのうち 1 回でも 6 の目が出れば勝ちという賭け事を行った。このギャンブラーは、1 回ふって 6 の目が出る確率が $\frac{1}{6}$ であることから、4 回ふるならば、$\frac{4}{6} = \frac{2}{3}$ の確率で勝てると考えたのである。実際、この賭けで、このギャンブラーは儲けることができた。

次にこのギャンブラーは、サイコロ 2 つを同時にふることを 24 回繰り返し、そのうち 1 回でも 6 のぞろ目が出れば勝ちという賭け事を始めた。6 のぞろ目が出る確率は、$\frac{1}{36}$ であることから、24 回ふれば、、$\frac{24}{36} = \frac{2}{3}$ の確率となり、先ほどの賭けと同じぐらい勝てると考えたのである。しかし、この 2 番目の賭けを行ったところ、このギャンブラーは損を出すこととなった。

最初の賭けをもう少しちゃんと考えてみよう。6 以外の目が出る確率は $\frac{5}{6}$ であり、4 回ふってすべて 6 以外の目が出る確率は、$\left(\frac{5}{6}\right)^4 = 0.4823$ である。ここから余事象の確率の考え方を使え

ば、6 が少なくとも 1 回出る確率は、$1 - 0.4823 = 0.5177$ となる。半分以上の確率で勝つわけだから、長く続ければ儲けることができる。

同様に 2 つ目の賭けを考えてみよう。6 のぞろ目以外が出る確率は、$\frac{35}{36}$ であり、24 回ふってすべて 6 のぞろ目以外が出る確率は、$\left(\frac{35}{36}\right)^{24} = 0.5086$ である。ここから、6 のぞろ目が 1 回でも出る確率は、$1 - 0.5086 = 0.4914$ となる。これでは、長い目で見れば損をすることとなる。

2.6.3 条件付き確率

日本語の文章で「阜」という漢字が出てくることはほとんどない。つまり、「阜」の出現確率は極めて低い。しかし、「岐」という漢字が出てきた直後では、「阜」が出てくる可能性は非常に高くなる。なぜならば、地名で「岐阜」という漢字の組み合わせが用いられるからである。要するに、「岐」が直前に出現したという条件の下では、「阜」の出現確率は非常に高くなるのだ。

「阜」の出現確率のように、全体的に見れば確率が極めて低いにも関わらず、ある一定の条件の下では確率が高くなることがある。また、その逆で、全体的に見れば確率が高いのに、一定の条件の下では確率が低くなる例もある。いずれにしても、「ある条件の下での確率」というものを考えた方が、うまく説明できる場合がある。「ある条件の下での確率」が、ここで扱う「条件付き確率」のことである。

参考 全体的に見れば確率が高いのに、一定の条件の下では確率が低くなる例を 1 つ紹介しよう。英語では 'i' という文字は、出現する確率が高い。しかし、直前に 'q' が出現したという条件の下では、'i' の出現確率は極めて低いものとなる。なぜなら、英語では 'q' の直後には、普通 'u' が出現するからである。

2.6.3.1 同時確率

条件付き確率を扱う前に、その前提となる同時確率について紹介する。

事象 A と事象 B がどちらも起きることを $A \cap B$ と表す。要するに、A が起きてなおかつ B も起きるようなことを $A \cap B$ と表しているのである。

参考 $A \cap B$ は、事象 A と事象 B が「**同時**に起こる」とも言うが、ここでの「同時」は時間的に同じという意味ではない。

$A \cap B$ が起こる確率 $P(A \cap B)$ を A と B の**同時確率**あるいは**結合確率** (joint probability) と呼ぶ。A と B の同時確率は $P(A, B)$ とも表記される。

◈ 同時確率の事例

さいころを 1 回ふったとき、3 以下の目が出る事象を A、奇数の目が出る事象を B としよう。このとき、$A \cap B$ とは、3 以下の目が出て、なおかつ奇数の目が出る事象のことである。$A \cap B$ となるのは、1 の目が出る場合と 3 の目が出る場合の 2 通りしかない。よって、$P(A \cap B) = \frac{2}{6} = \frac{1}{3}$ となる。

参考 良くある誤りとして、$P(A \cap B) = P(A)P(B)$ と考えてしまうことがある。つまり、上記の例で言うと、$P(A) = \frac{1}{2}, P(B) = \frac{1}{2}$ であることから、$P(A \cap B) = P(A)P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ というふうに考えてしまうのである。これは明らかに誤りである。

参考 ただし、定理 14 に記されているように、事象が独立である場合、$P(A \cap B) = P(A)P(B)$ となるので注意が必要である。

問 2–29　同時確率

以下の問いに答えよ。

1. さいころを 1 回ふって、2 以上の目が出る事象を A、5 以下の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。
2. さいころを 1 回ふって、偶数が出る事象を A、3 以下の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。
3. さいころを 1 回ふって、3 以上の目が出る事象を A、2 以下の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。
4. さいころを 1 回ふって、偶数の目が出る事象を A、奇数の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。
5. さいころを 1 回ふって、奇数の目が出る事象を A、3 以上の目が出る事象を B、4 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

⑥ さいころを 1 回ふって、偶数の目が出る事象を A、2 以上の目が出る事象を B、5 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

⑦ さいころを 1 回ふって、奇数の目が出る事象を A、5 以上の目が出る事象を B、2 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

⑧ さいころを 1 回ふって、奇数の目が出る事象を A、3 以上の目が出る事象を B、6 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

2.6.3.2 条件付き確率の定義

先に見たように、ある事象 B が起きたとき、事象 A が起きる確率というものを考えることができる。B が起きたときに A が起きる確率のことを B を条件とする A の**条件付き確率** (conditional probability) と呼ぶ。

定義 10 条件付き確率

B を条件とする A の条件付き確率 $P(A \mid B)$ は、以下のように定義される。

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

ただし、$P(B) \neq 0$ である。

参考 $P(B) \neq 0$ としたのは、0 で除算することを防ぐためである。

◈ 条件付き確率の事例

先ほどの同時確率の事例と同じく、さいころを 1 回ふったとき、3 以下の目が出る事象を A、奇数の目が出る事象を B としよう。このとき、B を条件とする A の条件付き確率 $P(A \mid B)$ を考えよう。

さて、$P(A \mid B)$ とは、奇数の目が出たという条件の下での、3 以下の目が出る確率のことである。とりあえず、上述の条件付き確率の定義（定義 10）は気にしないで、素朴に $P(A \mid B)$ の意味を考えてみよう。

まず、「奇数の目が出たという条件」について考えてみよう。さいころで奇数の目は、1, 3, 5 の 3 通りである。次に、1, 3, 5 のうち、3 以下の目であ

るのは、1 と 3 の 2 通りである。つまり、奇数の目が出たという条件の下での全体の場合の数は 3 通りで、3 以下の目が出る場合の数は 2 通りということになる。よって、$P(A \mid B) = \frac{2}{3}$ と考えることができる。

上記の計算は、条件付き確率の定義と合致しているだろうか。定義 10 によれば、$P(A \mid B)$ の値を計算するには、$P(A \cap B)$ と $P(B)$ の値を計算する必要がある。$P(A \cap B)$ とは A と B の同時確率である。先に見たように、$A \cap B$ となるのは、1 の目が出る場合と 3 の目が出る場合の 2 通りしかなかった。すなわち、$P(A \cap B) = \frac{2}{6} = \frac{1}{3}$ である。$P(B)$ とは、条件を何も加えない場合の B の確率である。事象 B は、奇数の目が出ることであったから、$P(B) = \frac{1}{2}$ である。

これらを踏まえて考えると、B を条件とする A の条件付き確率は、以下のように計算できる。

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

参考 $P(A \mid B) = P(B \mid A)$ は、普通は等しくない。事実、上述の例で、$P(A \mid B) = \frac{2}{3}$ であったが、$P(B \mid A) = \frac{2}{3}$

問 2–30　条件付き確率

以下の問いに答えよ。

1. さいころを 1 回ふって、2 以上の目が出る事象を A、5 以下の目が出る事象を B とする。このとき、$P(A \mid B)$ を求めよ。
2. さいころを 1 回ふって、偶数が出る事象を A、3 以下の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。
3. さいころを 1 回ふって、3 以上の目が出る事象を A、2 以下の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。
4. さいころを 1 回ふって、偶数の目が出る事象を A、奇数の目が出る事象を B とする。このとき、$P(A \cap B)$ を求めよ。

5 さいころを 1 回ふって、奇数の目が出る事象を A、3 以上の目が出る事象を B、4 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

6 さいころを 1 回ふって、偶数の目が出る事象を A、2 以上の目が出る事象を B、5 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

7 さいころを 1 回ふって、奇数の目が出る事象を A、5 以上の目が出る事象を B、2 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

8 さいころを 1 回ふって、奇数の目が出る事象を A、3 以上の目が出る事象を B、6 以下の目が出る事象を C とする。このとき、$P(A \cap B \cap C)$ を求めよ。

2.6.3.3 周辺確率

条件付き確率に対して、条件のない普通の確率を周辺確率 (marginal probability) と呼ぶ。例えば、事象 A の周辺確率は $P(A)$ である。これは 2.6.1 で述べられていた $P(A)$ と全く同じものであり、条件付き確率と対比させるために別の呼び名をつけただけのことである。

2.6.3.4 確率の乗法定理

定義 10 よりただちに、確率の乗法定理が成立する。

定理 11 確率の乗法定理

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

2.6.3.5 ベイズの定理

定理 11 を変形すると、以下のベイズの定理 (Bayes' theorem) が得られる。

定理 12 ベイズの定理

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

参考 $P(B \mid A)$ を直接求めるのは難しいが、$P(A \mid B)$ を直接求めることができる場合、ベイズの定理を使うと、$P(B \mid A)$ を求めることができる。

問 2–31　ベイズの定理

100 冊の本があり、そのうち 60 冊が男性に、40 冊が女性によって書かれている。また、100 冊のうち、70 冊は小説で、30 冊はエッセイである。なお、100 冊のうち、男性が書いた小説は 50 冊である。これらの 100 冊の本から、任意の 1 冊を取り出したとき、男性に書かれた本である事象を A、小説である事象を B とする。このとき、以下の問いに答えよ。

1. $P(A)$ および $P(B)$ を求めよ。
2. $A \cap B$ とはどのような事象か。
3. $P(A \cap B)$ を求めよ。
4. 条件付き確率の定義 (定義 10) に基づき、$P(A \mid B)$ および $P(B \mid A)$ を求めよ。
5. $P(A)$、$P(B)$、$P(A \mid B)$ の計算結果を用いて、ベイズの定理（定理 12）に基づいて $P(A \cap B)$ を求めよ。
6. $\overline{A}$ および $\overline{B}$ とはどのような事象か。
7. $P(B \mid \overline{A})$ を求めよ。条件付き確率の定義 (定義 10) に基づいて考えてもよいし、ベイズの定理（定理 12）に基づいて考えてもよい。

ベイズ統計学という、ベイズの定理に基づく統計理論がある。現代の統計学は、ベイズ統計学の占めるシェアが大きく、統計の勉強を進めていけば、いずれどこかで出会うだろう。ベイズ統計学は様々な分野に応用されている。有名なところでは、迷惑メールの判定にベイズ統計学の知見を生かすということがある。

「事前確率」は政治的に問題？

ベイズ統計学には、条件付き確率の一種として、事前確率・事後確率といった術語が用いられる。この「事前確率」という術語に

ついて面白い逸話があるので紹介したい。

Andrew Gelman という人が、*Data Analysis Using Regression and Multilevel/Hierarchical Models* というベイズ統計学に関する書籍を書いた。この本のリプリントを中国で出版しようという話があったのだが、中国側の出版社から「政治的」理由から出版いたしかねるというメールが届いたそうだ。これは、事前確率の「事前」が中国の旧体制を想起させるためではないかと言われている。むろん、ここでの「事前」はそういう意味ではない。

2.6.4 独立と排反

確率論に関する重要な概念として、独立と排反を紹介する。独立と排反は全く意味の違う用語なので、間違えないように注意が必要である。

2.6.4.1 独立

ある事象 A が起こる確率が、他の事象 B に影響されない時、A と B とは**独立** (Event A and B are independent.) であると定義する。

定義 13 独立（仮の定義）

事象 A と B が独立であるとは、

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

となることである。

しかし、定義 13 では、$P(B) = 0$ のときに、独立が定義できなくなる。なぜならば、定義 13 には、条件付き確率 $P(A \mid B)$ が含まれているからである。条件付き確率は $P(B) = 0$ のとき定義されないないため、独立も定義できなくなるのである。

よって、$P(B) = 0$ のときに問題がないように、定義 14 で、独立を再定義する。

定義 14 独立（新しい定義）

事象 A と B が独立であるとは、

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

となることである。

参考 この定義は、定義 13 を確率の乗法定理（定理 11）に代入することで得られる。

参考 実際には独立でないのに、$P(A \cap B) = P(A)P(B)$ として計算してしまう過ちは犯しやすく、注意が必要である。

◈ コイン投げと独立

100 円玉を 1 枚と 500 円玉を 1 枚投げる場合を考える。100 円玉が表になる事象を A とし、500 円玉が表になる事象を B とする。ここで、A と A は独立である。なぜならば、100 円玉が表になろうとも裏になろうとも、その結果は 500 円玉の裏表に全く影響を与えないからである。

ここで、$P(A) = 0.5, P(B) = 0.5$ であり、$P(A \cap B) = 0.25$ となっている。

問 2–32　事象の独立性

次の各問で挙げた 2 つの事象が独立であるか考察せよ。

1. さいころを 2 回ふって、1 回目で 2 の目が出る事象と、2 回目で 4 の目が出る事象。
2. 袋の中に、赤玉が 3 つ、白玉が 2 つ入っており、その中から 1 つ玉を取り出したとき赤玉がでる事象と、取り出した球を戻さないまま、もう 1 回玉を取り出したとき赤玉がでる事象。
3. 袋の中に、赤玉が 3 つ、白玉が 2 つ入っており、その中から 1 つ玉を取り出したとき赤玉がでる事象と、取り出した球を戻したあと、もう 1 回玉を取り出したとき赤玉がでる事象。

問 2–33　コイン投げ

太郎と次郎が、それぞれ 8 回ずつコインを投げたところ、以下のような結果となった。太郎の結果の方が珍しいか、それとも次郎の結果の方が珍しいか、議論せよ。

太郎	表	表	表	表	表	表	表	表
次郎	裏	表	表	裏	表	裏	裏	表

ルーレット

カジノのルーレットでは、数字が赤 (rouge) と黒 (noir) の 2 通りに分かれている。そして、赤が出る場合と黒の出る場合は半々である。仮に、今まで 5 回の試行で、ずっと赤が出てきたとしよう。この次は、赤と黒のいずれに賭けるべきだろうか。ずっと赤が出てきたのだから、次こそは黒が出ると考えるかもしれない。しかし、ルーレットの試行は、おのおの独立である。つまり、前に赤が出ようと黒が出ようと、それは次の試行に影響を与えない。だから、赤と黒のいずれに賭けても当たる可能性はやはり半々である。

なお、実際は、ディーラーが数字を狙うことができるので、このように独立だとは言えない。また、0 と 00 の 2 つの数字は赤と黒のいずれにも属していないので、赤と黒の出る場合が正確に半々とは言えない。

2.6.4.2 排反

「事象 A も事象 B も両方起こること」がないことを、排反 (Event A and B are mutually exclusive.) と呼ぶ。

定義 15 排反

事象 A と B が排反であるとは、

$$P(A \cap B) = 0$$

となることである。

◈ さいころと排反

さいころを 1 回だけ振ったとき、1 の目が出ることと 6 の目が出ることとは同時には起きない。よって、1 の目が出る事象と、6 の目が出る事象は排反であると言える。

2.6.5 確率の加法定理

次の確率の加法定理は、簡単に言えば、あることと別のことのどちらか少なくとも一方が起きる確率はいくらになるかということを教えてくれる定理である。

定理 16 確率の加法定理

事象 A と事象 B の少なくとも一方が起きる確率 $P(A \cup B)$ について、

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

が成り立つ。特に、A と B が排反であるとき、$P(A \cap B) = 0$ であるので、

$$P(A \cup B) = P(A) + P(B)$$

となる。

参考 実際には排反でないのに、$P(A \cup B) = P(A) + P(B)$ として計算してしまう過ちは犯しやすく、注意が必要である。

◈ さいころと確率の加法定理

さいころを 1 回ふったとき、奇数の目が出る事象を A とし、5 以上の目が出る事象を B とする。このとき、$P(A) = \frac{1}{2}$ であり、$P(B) = \frac{1}{3}$ である。また、事象 A と事象 B が同時に起こるのは、5 の目が出るときだけであるので、$P(A \cap B) = \frac{1}{6}$ となる。加法定理に当てはめると、

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{1}{2} + \frac{1}{3} - \frac{1}{6} \\ &= \frac{2}{3} \end{aligned}$$

となるので、A か B のいずれかが起きる確率 $P(A \cup B)$ は $\frac{2}{3}$ である。

さらに、排反となる例を考えてみよう。さいころで 2 以下の目が出る事象を C とする。このとき、明らかに 5 以上の目が出ることと 2 以下の目が出ることは両立し得ないので、事象 B と C は排反である。このとき、B か C のいずれかが起きる確率 $P(B \cup C)$ は、$P(B) + P(C) = \frac{2}{3}$ となる。

2.7

有効数字と丸め

寸ニシテ之ヲ度(はか)レバ、丈ニ至リテ必ズ差(たが)フ

『淮南子・泰族訓』

現実に手に入るデータは必ずしも、有効数字の桁数がどこでも同じというわけではない。このため、有効数字の桁数を何らかの方法でそろえる必要がある。また、計算の途中で、割り切れないといったことが起きる。例えば、$\frac{12}{36}$ を計算すると、0.33333333$\cdots$ と 3 が延々と続くが、どこかで打ち切って、キリの良い数値にする必要がある。有効数字の桁数をそろえたり、キリの良い数値にするために行われる操作が、**丸め** (rounding) である。

丸めにはさまざまな手法があり、状況に応じて適切な手法を用いる必要がある。ただし、他に理由がない限り、最近接偶数への丸め (2.7.6 節) を用いるのが良い。また、計算機の出力結果は非明示的に丸めが行われている場合もある。このため、計算機でどのような丸めが行われているかも気をつける必要が出てくる。

2.7.1 有効数字

有効数字をはっきりと示すためには、1 以上 10 未満の小数と 10 の累乗との積の形で表す。例えば、342 は $3.42 \cdot 10^2$ となり、0.7985 は $7.985 \cdot 10^{-1}$ となる。この例で、3.42 や 7.985 が有効数字であり、10^2 や 10^1 は位を調整するために掛け合わせているものである。

5600 と書いたとき、下 2 桁の 00 が、位取りの 0 なのか、有効数字の 0 なのかは分からない。このため、これが有効数字が 2 桁であるか、3 桁であるか、はたまた 4 桁であるかは、分からない。これをはっきりさせるためには、小数と 10 の累乗との積の形で表せばよい。このような数値の表記方法を**科学表記**という。

- 有効数字が 2 桁： $5.6 \cdot 10^3$
- 有効数字が 3 桁： $5.60 \cdot 10^3$
- 有効数字が 4 桁： $5.600 \cdot 10^3$

2.7.2 丸め

丸めの定義として、JIS の「数値の丸め方」の記述を引用しよう。

> **JIS8401:1999「数値の丸め方」**
> 丸めるとは、与えられた数値を、ある一定の丸めの幅の整数倍が作る系列の中から選んだ数値に置き換えることである。この置き換えた数値を丸めた数値とよぶ。

例えば、丸めの幅が 100 なら、その整数倍である $\ldots, -200, -100, 0, 100, 200, 300, \ldots, 51$ のいずれかに数値を置き換えれば良いのである。また、丸めの幅が 0.1 なら、その整数倍の例として、$\ldots, 8.0, 8.1, 8.2, 8.3, 8.4, \ldots$ などがあり、そのいずれかに数値を置き換えれば良い。

丸めを行うのは 1 回だけにする。例えば、83.48 を四捨五入して整数にする場合を考えよう。図 2.2 のように、83.48 を 83.5 とし、さらにこの 83.5 を 84 にするという二段階の四捨五入をしてはならない。四捨五入して整数にしたいのならば、図 2.3 のように、一気に四捨五入をし、83.48 から 83 にしなくてはならない。

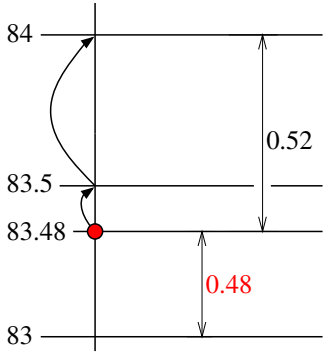

図 2.2　誤った丸め

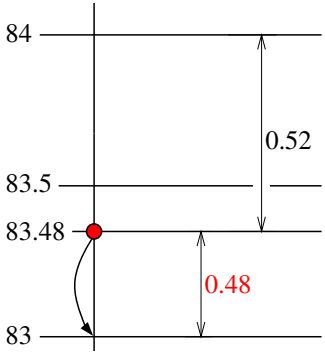

図 2.3　正しい丸め

上記の例はあまりにもばかげていると考える人もいるかもしれない。しかし、注意していないと気づかぬうちに 2 回以上丸めてしまう過ちを犯してしまう場合が少なくない。

2.7.3 切り捨て

切り捨ては、有効数字以下の数値を無くす丸めである。整数になるように切り捨てる場合は、整数部分に変化はなく、小数部分が落とされる。

例えば、83.249 を整数になるように切り捨てたら、83 になる。また、54,783 を有効数字が 3 桁になるように切り捨てたら、$5.47 \cdot 10^4 (= 54,700)$ となる。

2.7.4 切り上げ

切り上げは、有効数字以下の数値を無くした上で、有効数字の末尾の桁を 1 つ上げる丸めである。整数になるように切り上げる場合は、小数部分を落とした後、整数部分に 1 を加えることになる。

例えば、83.249 を整数になるように切り上げたら、84 になる。また、54,783 を有効数字が 3 桁になるように切り上げたら、$5.48 \cdot 10^4 (= 54,800)$ となる。

参考 切り捨ても切り上げも丸めによって誤差が増える。切り捨てを続けると全体としては、値が小さくなる。同様に切り上げを続けると全体としては値が大きくなる。このため、切り捨てや切り上げを用いる機会はそれほど多くない。ただし、慎重で保守的な結果を出したいときに切り捨てや切り上げを用いることもある。

参考 切り捨ても切り上げも、負数の場合はそれぞれ 2 通りの解釈が生じうる。詳しくは、2.7.7 節を参照せよ。

2.7.5 四捨五入

四捨五入とは、一番近いキリの良い数値に近づける丸めの手法である。四捨五入の場合、丸めの対象となる桁の数字が 1 から 4 の場合には切り捨て（捨）、5 から 9 の場合には切り上げる（入）ことになる。

参考 四捨五入は日常生活ではよく用いられるが、偏りのある手法なので、研究では使わないようにしなくてはならい。丸めが必要になったら 2.7.6 節で紹介する最近接偶数への丸めを使うのがよい。最近接偶数への丸めは、ほとんどの場合は四捨五入と同じように丸める。しかし、最近接偶数への丸めは、5 を必ず切り上げる（五入）のではなく、切り下げる場合（五捨）もあり得る。

2.7.5.1 四捨五入の考え方

一番キリの良い数値に近づけるとはどういうことだろうか。実際の例を通じて考えてみよう。

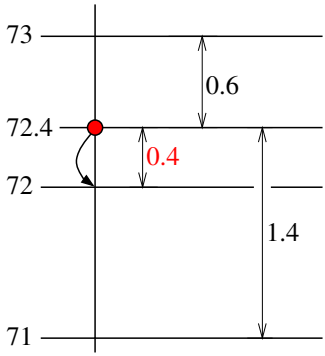

図 2.4　72.4 の四捨五入

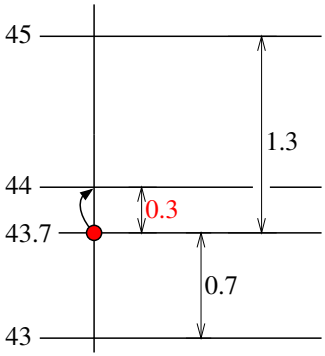

図 2.5　43.7 の四捨五入

まずは、72.4 を整数にするために四捨五入する例を考えよう（図 2.4）。整数の中で、72.4 に最も近いのは、72 である。72 は 72.4 から 0.4 しか離れていない。他の整数は 72.4 からもっと離れている。例えば、73 は 72.4 から 0.6 も離れている。71 は 72.4 からさらに遠く、1.4 も離れている。結局、72 が一番近い整数だから、72.4 の四捨五入の結果は 72 になる。これは、72.4 に対し切り捨てを行った場合と同じ結果である。

次に、43.7 を整数にするために四捨五入する例を考えよう（図 2.5）。整数の中で、43.7 に最も近いのは、44 である。このため、43.7 の四捨五入の結果は 44 になる。つまり、43.7 に対し切り上げを行った場合と同じ結果になる。

もう 1 つ、843,738 を有効数字が 3 桁になるように四捨五入する例を考えよう。有効数字が 3 桁になる数で、843,738 に最も近いのは、$8.44 \cdot 10^5$ $(= 844,000)$ である。この数は、843,738 から 262 だけ離れている。有効数字が 3 桁の他の数はもっと 843,738 から離れている。例えば、$8.43 \cdot 10^5$ $(= 843,000)$ は、843,738 から 738 も離れている。結局、$8.44 \cdot 10^5$ $(= 844,000)$ が 843,738 に最も近いので、これが四捨五入した結果となる。この結果は、843,738 に対し切り上げを行った場合と同じである。

2.7.5.2 四捨五入の問題点

日常生活で行う四捨五入では、四捨五入の対象となる桁が 5 の場合は必ず切り上げる（五入）。このことを踏まえ、四捨五入すべき桁の数字に応じて、切り捨て（捨）と切り上げ（入）がどう分かれるかを見てみると、次のようになる。

- 0 → 変らず
- 1 → 捨
- 2 → 捨
- 3 → 捨
- 4 → 捨
- 5 → 入
- 6 → 入
- 7 → 入
- 8 → 入
- 9 → 入

結局、4 回切り捨て（捨）を、5 回切り上げ（入）をしている。これでは、切り上げのほうが多くなってしまい、偏ってしまう。

また、四捨五入の対象となる桁が 5 の場合、困ったことが起きる。例えば、72.5 が整数になるように四捨五入する例を考えてみよう。五は必ず切り上げる（五入）という日常の規則を無視し、一番近いキリの良い数値に近づけるということだけを考えよう。72.5 に最も近い整数は、実は 2 つある。72 と 73 である。72 も 73 も、72.5 から 0.5 だけ離れている。つまり、一番近いキリの良い数値が 2 つあるのである。つまり、72.5 のような場合、どちらを選べば良いのか本来は悩むべきなのんである。しかし、日常生活で行われる四捨五入では、無批判に切り上げて（五入）73 にしてしまっている。これは問題である。

2.7.6 最近接偶数への丸め

上述の四捨五入の問題点を解決する丸めの手法が、最近接偶数への丸め (rounding to the nearest even) である。

最近接偶数への丸めは、基本的には四捨五入と同じように、一番近いキリ

の良い数値手法である。さらに、最近接偶数への丸めは、切り上げの方が多くなって偏ってしまうという四捨五入の短所を持たない。それでは、どうやって、偏りを防いでいるのであろうか。四捨五入で問題となったのは、丸めの対象が 5 である場合であった。5 の時に切り上げて（五入）ばかりいたのが問題だったのである。結局、状況に応じて、五入だけでなく、切り捨てる（五捨）ことも必要なのである。

ここで、以下の規則を追加する。

- 一番近いキリの良い数値が 2 つある場合、末尾が偶数のものを優先する。

この規則を追加することで、五捨が起こることもある。先ほどの 72.5 の四捨五入の場合も五捨が起き、72 となる（図 2.6）。なぜならば、73 の末尾の桁は 3 で奇数であるのに対し、72 の末尾の桁は 2 で偶数であるからである。

この規則の導入で以下の 2 つの問題が解決される。

- 切り上げのほうが多くなることが回避される。つまり、四捨五入による偏りがなくなる。
- 一番近いキリの良い数値が 2 つある場合にどちらを選べばよいかという問題がなくなる。

なお、JIS や ISO はこの方式の丸めを採用しているので、**JIS 式丸め**や**ISO 式丸め**と呼ぶ人もいる。

他の例も見てみよう。96.5 を整数にするべく最近接偶数への丸めを行う場合、五捨されて 96 となる。96.5 に一番近いキリの良い数値は、96 と 97 である。このうち、97 は末尾が奇数であるが、96 は末尾が偶数である。このため、96 が最近接偶数への丸めの結果になるのである。これに対し、47.5 を整数にするべく最近接偶数への丸めする場合、五入されて 48 となる（図 2.7）。47.5 に一番近いキリの良い数値は、47 と 48 である。このうち、47 は末尾が奇数であるが、48 は末尾が偶数である。このため、48 が最近接偶数への丸めの結果になる。

次に 245 を有効数字が 2 桁になるように最近接偶数への丸めを行う例を考えよう。ここで、245 に一番近いキリの良い数値は $2.4 \cdot 10^2 (= 240)$ と $2.5 \cdot 10^2 (= 250)$ である。ここで、有効数字の末尾が偶数であるのは、

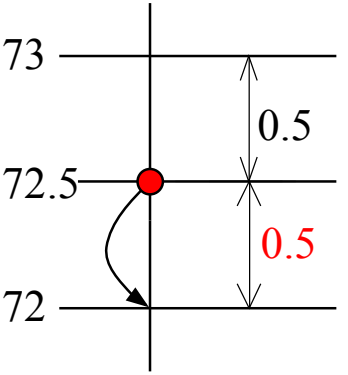

図 2.6 「五捨」の例

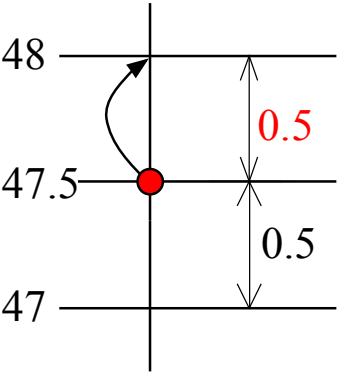

図 2.7 「五入」の例

$2.4 \cdot 10^2 (= 240)$ の方である。このため、最近接偶数への丸めの結果は、$2.4 \cdot 10^2 (= 240)$ になる。

参考 240 も 250 も偶数であると悩まないようにしたい。数値全体が偶数であるかどうかよりも、有効数字の末尾の桁が偶数であるかどうかを見ればよいのである。

最後に、72.53 に対して最近接偶数への丸めを行って整数になるようにする例を考えよう。この例では 72.5 のように切り捨ててはならない。72.5 の時は 72 と 73 が同じ近さであったため、末尾が偶数の方を優先した。しかし、72.53 の場合、72 と 73 は同じ近さではなく、どちらかと言えば、73 の方が近い。このため、ここでは 73 が最近接偶数への丸めの結果になる。末尾が偶数であるものを優先するというのは、あくまでも一番近いキリの良い数が 2 つある時だけに適用される規則であり、一番近いキリの良い数が 1 つしかないときには考える必要はない。

問 2–34 最近接偶数への丸め

次の数値を有効数字が 2 桁になるように最近接偶数への丸めを行え。

1 5321

2 7469
3 48.73
4 0.4234
5 3.558
6 0.0425
7 375
8 4650
9 0.635
10 12501
11 2153
12 0.7458

2.7.7 R での丸め

R には、丸めを行う函数が複数ある。以下に一覧を掲げる。

- `trunc`
- `floor`
- `ceiling`
- `round`
- `signif`
- `zapsmall`

それぞれ、引数としてベクトルを取ることができる。

2.7.7.1 切り捨て・切り上げ

`floor` は切り捨て、`ceiling` は切り上げを行う。`trunc` はやや特殊で、0 に近くなるようにキリの良い数値を求める。要するに、`trunc` は正の数に対しては切り捨てを、負の数に対しては切り上げを行う。よって、正の数を引数に取るときは、`trunc` と `floor` の結果は全く同じものになる。これに対して、負の数を引数に取る場合は、`trunc` と `ceiling` が同じ結果となる。図 2.8 を見れば明らかであろう。

正数に `trunc`, `floor`, `ceiling` を適用した例は以下の通りである。

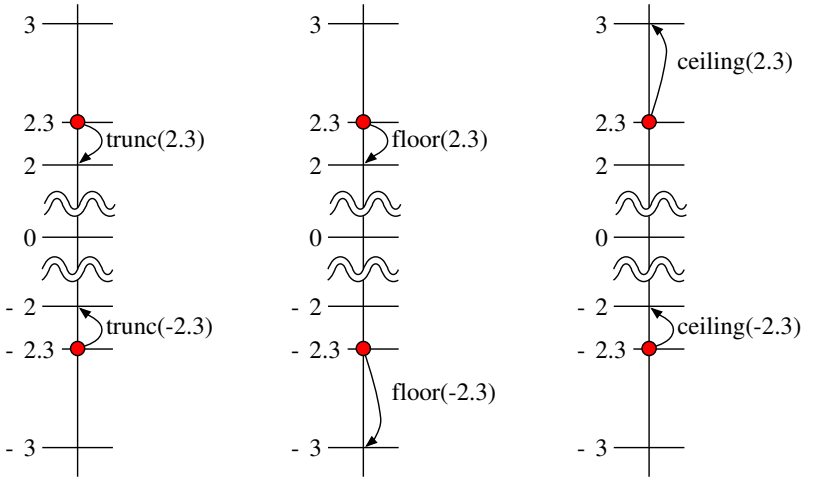

図 2.8 trunc, floor, ceiling

```
> trunc(2.3)
[1] 2
> floor(2.3)
[1] 2
> ceiling(2.3)
[1] 3
```

また、負数に trunc, floor, ceiling を適用した例は以下の通りである。

```
> trunc(-2.3)
[1] -3
> floor(-2.3)
[1] -3
> ceiling(-2.3)
[1] -2
```

2.7.7.2 最近接偶数への丸め

最近接偶数への丸めを行うには、round を使う。

参考 四捨五入を行う函数はデフォルトでは存在しない。だが、四捨五入は偏りがある丸めの手法であるから、そもそも四捨五入を使う必要はない。

参考 どうしても四捨五入を行いたいのならば、floor(x+0.5) のようにすれば、x を四捨五入できる。

デフォルトでは、round は小数点以下を丸め、結果は整数になる。

```
> round(2.3)
[1] 3
> round(5.7)
[1] 6
> round(2.5)
[1] 2
> round(3.5)
[1] 4
```

小数点以下何桁までを残すかを変えたければ、digits の値を変更すれば良い。デフォルトでは、digits = 0 である。

```
> round(3.14159265)
[1] 3
> round(3.14159265, digits=1)
[1] 3.1
> round(3.14159265, digits=2)
[1] 3.14
> round(3.14159265, digits=3)
[1] 3.142
> round(3.14159265, digits=4)
[1] 3.1416
> round(3.14159265, digits=5)
[1] 3.14159
```

digits の値を負にすることで、小数点以上の桁で丸めを行うことができる。

```
> round(123456789)
[1] 123456789
> round(123456789, digits=-1)
[1] 123456790
> round(123456789, digits=-2)
[1] 123456800
> round(123456789, digits=-3)
[1] 123457000
> round(123456789, digits=-4)
[1] 123460000
> round(123456789, digits=-5)
[1] 123500000
```

2.7.7.3 signif, zapsmall

R には、signif という関数があり、これも最近接偶数への丸めを行う。ただし、round とは意味合いが異なっている。signif は、指定した桁数

(digits) の分だけ有効数字を残す函数である。これに対し、指定した桁数 (digits) の分だけ小数点以下の桁を残す函数である。

参考 なお、signif で、digits の値を指定しない場合、digits=6、すなわち 6 桁分残すものとして扱われる。

```
> signif(123456789, digits=1)
[1] 1e+08
> signif(123456789, digits=2)
[1] 1.2e+08
> signif(123456789, digits=3)
[1] 1.23e+08
> signif(123456789, digits=4)
[1] 123500000
> signif(123456789, digits=5)
[1] 123460000
> signif(123456789)
[1] 123457000
> signif(123456789, digits=7)
[1] 123456800
> signif(123456789, digits=8)
[1] 123456790
> signif(123456789, digits=9)
[1] 123456789
> signif(123456789, digits=10)
[1] 123456789
> signif(123456789, digits=11)
[1] 123456789
```