統計学入門

小波秀雄

June 2022

はじめに

多数のデータから意味のある情報を抽出するのが統計的手法であり，その理論が**統計学 (statistics)** である．統計学は，確率論を基礎にして，不確実性を含む多数のデータから，一定の確実さをもった判断を下すことを目的にしている．

統計学は，社会や人間に関わるさまざまな事象の分析と多数のデータの定量的な取り扱いを可能にすることから，社会科学や医学などの人間集団を相手にした学問研究分野，心理学や教育学などの人間行動の分野，品質管理などの生産現場，保険や経営といったマネージメント分野，また政策決定のための指針作成など，さまざまの分野で広範に活用されている．

自然科学の分野でも，不確実性を含む自然現象は数多く，データの統計的な取り扱いが必要になる．また情報理論の中でも確率論とその応用は重要な一分野である．特に本書で展開される確率分布の理解は情報理論の中でも基本となるものである．

このように，統計学はまさに現代の学問と産業を支えている主要な理論のひとつであるといっても過言ではない．

その反面，確率や統計の誤った解釈や，意図的に捻じ曲げられた解釈によって，誤った指針や主張が導かれることも稀なことではない．嘘をつくための道具として，統計が不審の眼を向けられることも昔からよくあることである．誤った解釈に振り回されたり，統計の嘘にだまされたりしないためにも，統計の理論を基礎から理解することは大切である．

このテキストでは，確率論の入門からはじめて，古典統計学の理論を一通り取り扱う．数式は多いが，高校数学程度の力があれば追えるように，巻末付録に式の誘導を掲載した．あまり多くはないが，理解に必要な例題と練習問題を入れてあるので，それも含めてまじめに取り組んでいただけば，統計学の十分な基礎力を獲得できる．現代的な多変量統計や予測統計はこの先の展開になるが，その勉強のための足がかりにもなるはずである．

表計算アプリケーションの利用

統計処理では数多くのデータを使って多数回の計算を行う．

その労力を省くために，Excel や Numbers, OpenOffice/LibreOffice Calc といった表

計算アプリケーション[*1]を使うと便利だ．セルにデータを打ち込んでから，簡単な数式を使って一斉に同型の計算をさせたり，総和を取ったりできるので，このテキストの問題を解くために活用してみることをお勧めしたい．

ただし，これらを使用する際に注意しておかなければならないのは，特に統計関数を利用したときに，出てきた数字をそのまま信用してしまって，ミスを見逃してしまうことである．たとえば分散および標準偏差を求める関数として VAR と STDEV があるが，これは第 1 章で出てくる標準偏差とは定義と値が異なることを知っておかないとまずい．

統計計算のためのパッケージの利用

本格的な統計処理のためのパッケージとして，オープンソースの統計処理のためのプログラミング言語である R[*2] が開発されて，広く利用されるようになっている．これから何らかの統計処理パッケージを導入する場合には，まず R を使うことをおすすめしたい．単に「R」で検索するだけでダウンロードの仕方も含めて情報が手に入るようになっている．R については，多数の参考書やマニュアルも出版されているので，その意味でも学びやすい環境になっている．巻末に R に関する情報をまとめてあるので参考にしていただきたい．

また Python も数理分野に強く，最近は人気のあるプログラミング言語である．統計処理についても多数の参考書が出ているので，好みに合ったものを使って実務に活用することもお勧めである．

正しくアプリケーションを使うために

車を運転するのに，エンジンの仕組みや道路設計に関する知識は必要ない．それでも，どこに行こうとしてハンドルやアクセルを操作しているのかを分かっていないと，車はあらぬところに到着してしまう．ところが，それでも「目的地に到着しました！」と運転手が宣言する，そんなことがあったら客はどう思うだろうか？

ところが，「コンピュータで統計処理をやりました」といって，これとまったく同様の誤りを犯してしまうことはむしろありふれている．研究や実務に携わる人でさえ，実は統計学について無知なままに手続きだけを覚えて，結果を出しているケースは珍しくない．それを避けるには，統計的なデータ処理の意味をわかっておくことが必須であり，このテキストはそのために書かれている．

統計学を学ぶということは，難しい数学をマスターすることではないし，まして，基本的な定理の証明にまで遡って勉強する必要はないと言える．このテキストでも，ほとんど

[*1] Excel, Numbers, はそれぞれ Microsoft Office, Apple iWork, に含まれる表計算アプリケーション．
[*2] R はフリーソフト財団の GNU プロジェクトとして開発されているので GNU R と呼ぶこともある．

の数式の導出は付録に回して，数学的に納得したい人の便宜を図りながらも，本文では数学的な細部にあまり立ち入らないように留意した．

しかし，数値データを材料として処理を進める以上，その処理が何を意味しているかを理解するためには，最低限の数学的な扱いは必要である．それを押さえた上でアプリケーションの使い方をマスターすれば，安心して，かつ創造的に統計の手法を活用できる人になれるのだ．そのつもりで，本書を学んでほしい．

用語について

統計学は広範な分野で使われているために，分野ごとに，あるいは本によって用語の不統一が目立つ．このテキストでは定評のある英語の教科書とその日本語訳を中心にして，用語の統一を図った．

この本の利用について

この本の PDF ファイルは下からダウンロードできます．

`http://konamih.sakura.ne.jp/Stats/Text/`

ダウンロードは自由に行っていただいてかまいません．利用にあたっては，次の点に留意してください．

- 個人としての利用は許諾なしに行ってください．
- 学校や企業などにおける講義，セミナー等で使う際には，利用の形について著者に教えていただけると幸いです．
- 出版その他のパブリックな媒体への転載，図版の利用等については著者の許諾を得てください．
- ウェブからダウンロードできるようにするときには，古いバージョンがネット上に残ることを避けるため，上の URL へリンクすることとし，転載したファイルを別に置くことは避けてください．
- 内容に関するコメント（誤りの指摘，質問，要望など）がありましたら，メールでお知らせください．

著者連絡先

著者の肩書と連絡先は以下のとおりです．
京都女子大学　名誉教授
小波秀雄
E-mail: konami@kyoto-wu.ac.jp

目次

第 1 章

データの整理と表現

もともと「統計」という言葉は，集めた多数のデータを整理して利用しようという実用的な目的のもとに使われるようになった．そのための手法を**記述統計学 (descriptive statistics)** と呼ぶ．この章では，多数のデータをどのように要約し，どのように表現するかを学ぶ．

1.1　データの集合から統計量を求める

100 人の男子高校生の体重を調べて，表 1.1 のような結果が得られた．

表 1.1　男子高校生 100 人の体重のデータ：単位は kg

43.6, 45.2, 45.4, 45.8, 47.2, 47.8, 48.2, 48.7, 48.8, 48.9, 49.0, 49.0, 49.4, 49.5, 49.8, 50.4, 50.5, 50.9, 50.9, 51.2, 51.2, 51.2, 51.3, 51.3, 51.6, 51.7, 51.7, 51.8, 52.0, 52.0, 52.1, 52.1, 52.1, 52.2, 52.3, 52.7, 52.7, 52.8, 52.9, 52.9, 53.1, 53.1, 53.8, 54.0, 54.5, 54.5, 54.6, 54.7, 54.7, 54.7, 54.8, 54.9, 55.1, 55.1, 55.2, 55.3, 55.4, 55.4, 55.4, 55.6, 55.7, 55.8, 55.9, 56.1, 56.3, 56.3, 56.3, 56.4, 56.5, 56.7, 56.8, 57.0, 57.1, 57.1, 57.2, 57.3, 57.6, 57.7, 57.8, 58.1, 58.4, 58.6, 58.7, 58.7, 58.7, 58.7, 59.1, 59.3, 59.9, 60.0, 60.1, 60.3, 60.5, 60.6, 60.6, 60.7, 61.3, 62.7, 64.2, 64.6

このようなデータの数値の並びをデータ列 $\boldsymbol{x}$ と呼び，次のように表現することにしよう．n はデータの数である．

$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \tag{1.1}$$

1.1.1　平均 $\bar{x}$, μ

これから**平均 (mean**[*1]**)** を求めるには，だれでも知っているように次のように計算すればよい．

$$\frac{1}{100}(43.6 + 45.2 + 45.4 + 45.8 + \cdots + 64.6) = 54.46$$

$\boldsymbol{x}$ の平均は $\bar{x}$ のように表記され，μ が使われることもある[*2]．平均は一般的に次のように定義される．

$$\begin{aligned} \bar{x} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned} \tag{1.2}$$

総和記号 $\sum$ を使った書き方は短くて便利だが，ちょっとむつかしそうに見えるので，それを展開した形を思い浮かべて使うとよい．本書ではなるべく展開した形も併記する．

1.1.2　偏差

統計量そのものではないが，**偏差 (deviation)** もよく使われる量である．**平均偏差**とも呼ぶことがある．偏差は式 (1.3) で表されるように，あるデータが平均値からどれだけずれているかを意味する[*3]．

$$\delta x_i = x_i - \bar{x} \tag{1.3}$$

すべてのデータについての偏差の和はゼロになることが，次のようにして簡単に示せる．

$$\begin{aligned} \delta x_1 + \delta x_2 + \cdots + \delta x_n &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= (x_1 + x_2 + \cdots + x_n) - n\bar{x} \\ &= n \times \frac{1}{n}(x_1 + x_2 + \cdots + x_n) - n\bar{x} = 0 \end{aligned}$$

もっとも，平均よりも大きい分と小さい分が打ち消しあうので総和がゼロになると考えれば，式は見なくても直感的に理解できるだろう．

[*1] average もここで定義される平均の意味で使われるが，メジアン（後述）などデータの「真ん中」を表す他の尺度も含むあいまいな用語である．

[*2] μ はミューと読む．mean の m に相当するギリシャ文字である．

[*3] δ はデルタ．小さな差を表すのによく使われる．

1.1.3 分散 σ^2, 標準偏差 σ

データがどこを「中心」として分布しているのかを表すためには平均や後述するメジアンが使われる．それではデータがどの程度ばらばらに散っているかの目安としては，どのような量を考えればよいのだろうか．

偏差は，それぞれのデータの平均からのずれなので，すべての偏差を平均してみてその大きさで「ばらばら度」の尺度にしてみるという発想ではどうだろうか？しかし，上ですでに指摘したように，偏差の総和は常にゼロになるので，偏差の平均もゼロになってしまう．

そこで，偏差を 2 乗した値の平均として表される σ^2 という量を，データの広がりを表す尺度として定義する[*4]．

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\left((x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \end{aligned} \tag{1.4}$$

σ^2 は**分散 (variance)** と呼ばれ，この値が大きいほどデータはばらばらに散っていることになる．

また，分散の平方根 σ は**標準偏差 (standard deviation)** [*5]と呼ぶ．

$$\sigma = \sqrt{\sigma^2} \tag{1.5}$$

■分散と標準偏差の使い分け

分散をデータの広がりの尺度として導入したが，どうしてわざわざその後で平方根をとった標準偏差というものを持ち込むのだろうか．

今，長さのデータを扱っているものとして，その単位が m であったとする．分散は 2 乗の平均だから，単位は m^2 ということになる．つまり分散はデータそのものとは異なった単位をもっているので，データや平均の値と比較することはできない．「10 m と 100 m^2 とどっちが大きい？」と聞かれても，答えるのは不可能だ．

そこで，分散の正の平方根である標準偏差を考えると，こちらはもとのデータと同じ単位をもっているので，たとえば平均の周りでデータがどのようにばらついているかを考えるには，標準偏差が有効だということになる．つまり，データから直接に計算できるのは分散なのだけれど，標準偏差のほうがデータと比較する尺度としては直観的に分かりやすいものだということになる．

以上を次のようにまとめておこう．

[*4] σ はシグマと呼ぶ．

[*5] SD などと略されることがある．また，RMS (Root Mean Square) と呼ばれることもある．

平均と標準偏差は，分布の中心と広がりをつかむためのワンセット

なお，**標準誤差 (standard error)** というのもある．94 ページを参照のこと．

■平均と分散はもっとも重要な統計量

データの集合の特徴を表す量のことを**代表値 (representative value / descriptive statistics)** という．データの「真ん中」を代表する値には平均かメジアンが使われることが多いが，数学的には平均のほうがずっと扱いやすい．

そこで，平均をデータを代表する統計量，分散をデータのばらつきを表す基本的な統計量として取り扱うことが統計の中心的な作業になる．ただし，データの集団の実態とかあるいは実感といった見方からすると，次節で述べるメジアンや四分位数のほうが，より分かりやすい代表値であるということもしばしばある．

1.1.4　分散，標準偏差を求める別の公式

式 (1.4), 式 (1.5) は，別の形に導くことができ，そのほうが便利なことがある．すなわち，

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{x}^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2=\overline{x^2}-\overline{x}^2 \qquad (1.6)
\end{aligned}$$

最後の式に現われる $\overline{x^2}$ は $\frac{1}{n}(x_1^2+x_2^2+\ldots+x_n^2)$，つまり各データの **2 乗の平均**を意味している．

平均 $\overline{x}$ はデータ全体によってきまる定数だから，総和の記号の外にくくり出せることを利用して，次のように変形を行っている．

$$\sum_{i=1}^{n}\overline{x}x_i=\overline{x}x_1+\overline{x}x_2+\ldots+\overline{x}x_n=\overline{x}\sum_{i=1}^{n}x_i=\overline{x}\times n\overline{x}=n\overline{x}^2$$

$$\sum_{i=1}^{n}\overline{x}^2=\overline{x}^2\overbrace{(1+1+\ldots+1)}^{n}=n\overline{x}^2$$

さて，式 (1.6) は，次のきわめて大事な事実を教えてくれる．

分散 ＝二乗の平均 − 平均の二乗

この関係はしばしば利用される．また，分散を求めるための効率のよいアルゴリズムにもなっている．

問題 1–1　表 1.1 のデータから分散と標準偏差を求めよ．いずれも有効数字 4 桁で答えること．

問題 1–2　0 と 1 が合計 n 個あり，そのうちの 1 の割合が p であるようなデータを考える．このデータの平均と分散と標準偏差を求めよ．なおこの結果は，世論調査の結果の分析などで重要な意味を持つ．

♡　**標準偏差とデータのまとまり — チェビシェフの不等式**　♡

標準偏差がデータのばらつきの尺度であることはすでに説明したが，これについては**チェビシェフの不等式 (Chebyshev's inequality)** と呼ばれる有名な公式がある．数式を使わずに表現するとこうなる．

あるデータの集合の平均 μ と 標準偏差 σ が分かっているとする．そのとき，全体のうち $\mu \pm a\sigma$ の範囲からはみ出すデータの割合は，任意の a について $\frac{1}{a^2}$ 以下しかない．

たとえば，表 1.1 のデータでは，$\mu = 54.46, \sigma = 4.22$ となることが計算してみて分かる (問題 1–1)．そこで $a = 2$ ととってみると，平均の $\pm 2 \times 4.22$ の範囲は $54.46 - 2 \times 4.22 = 46.02$ と $54.46 + 2 \times 4.22 = 62.90$ を両端とする区間だ．定理が教えるのは，この範囲の外には，全部で 100 個あるうちたかだか $1/2^2 = 1/4$ 以下しかないということだ．つまり 25 個以下ということだ．一方，表を見てこの範囲から外れるデータの数を数えると全部で 6 個だから，チェビシェフの不等式と一応合致している．

もっとも，こうやって実際に計算してみると，この不等式による「縛り」は緩すぎて，大してありがたくないように思えるかもしれない．しかしこの定理は，データは平均から遠ざかるほど割合が減少し，その減り方は標準偏差で測られるということを教えてくれるという意味で大切なものである．

★

1.1.5　メジアンと四分位数, median / quartile

■四分位数

データを同数に 4 等分したときに，全体の 1/4, 2/4, 3/4 の位置に相当する値を**四分位数 (quartile)** といい，3 つの値の小さい方から**第 1 四分位数 (first quartile)**，**第 2 四分位数 (second quartile)**，**第 3 四分位数 (third quartile)** という[*6]．ただし第 2 四分位数は次に出てくるメジアンに等しいので，四分位数は第 1 と第 3 についてのみ使うことが多い．これらの正確な計算法は次で述べる．

なお，一般にデータを任意に n 等分した三分位数，五分位数なども考えることができるが，最もよく用いられるのは四分位数である．

■メジアン

すべてのデータを大きさの順に並べたときに，中央に位置するデータの値を**メジアン (median)** または**中央値**という．メジアンは第 2 四分位数であり，平均と同様にデータの集合を代表する最も重要な統計量のひとつである．

■四分位数を計算して求める

データを 4 分割するといっても，データの数 n によって分割の仕方が変わるので，その場合によって計算の仕方が異なることになる．そこで，n を $4m$, $4m+1$, $4m+2$, $4m+3$ $(m = 0, 1, 2, \ldots)$ のように場合分けして考える．

図 1.1 を見てほしい．図中の $x_1, x_2, \ldots, x_n$ は昇順に並べられたデータの値だ．**これらは実際にはばらばらな値をとっているのだが，このように等間隔に配置して計算を進める．**データの数 n を 12 から 15 までと，およびそれらを一般化した $4m$ から $4m+3$ までの 4 通りの場合に分けて，上から順にデータ列の並びを示してある．

この図を使って実際に計算をする段取りは，次のようになる．

1. データの数 n の値によって，使うべき場合を決める．ここでは仮に 12 としよう．すると一番上の $4m$ の場合で行くことになる．
2. $\boldsymbol{Q_1}$ を決める点は，左から $n/4$ 番目と $n/4+1$ 番目，つまり x_3 と x_4 である．
3. 図から $\boldsymbol{Q_1}$ は x_3 と x_4 を 3 : 1 に内分する点だ．したがって次の式で求められる．

$$\frac{1}{4}(x_3 + 3 \times x_4)$$

4. 次に，$\boldsymbol{M}$ を決める点は $n/2 = 6$ 番目と $n/2+1 = 7$ 番目になる．ただし今度は 2

[*6] 1/4 分位数，2/4 分位数，3/4 分位数という呼び方もある．

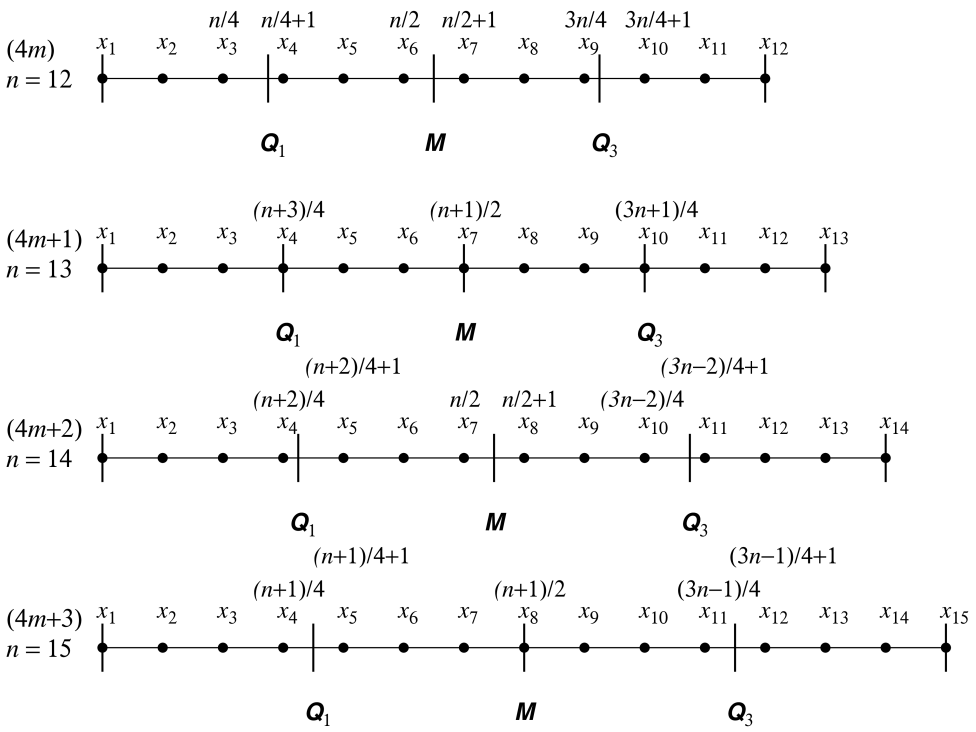


図 1.1　メジアン ($\boldsymbol{M}$), 第 1 四分位数 ($\boldsymbol{Q_1}$), 第 3 四分位数 ($\boldsymbol{Q_3}$) を計算するための場合分けと各分位数の位置．細かい意味は本文を参照のこと．

つの点を等分に内分しているので，次の式で求められる．

$$\frac{1}{2}(x_6 + x_7)$$

5. 最後に，$\boldsymbol{Q_3}$ を決める点は，$3n/4 = 9$ 番目と $3n/4 + 1 = 10$ 番目になる．今度はこれらを 1 : 3 に内分しているので，次の式で求められる．

$$\frac{1}{4}(3 \times x_9 + x_{10})$$

例題 1–1　メジアンと四分位数を求める

表 1.1 のデータから，第 1 四分位数 $\boldsymbol{Q_1}$，メジアン $\boldsymbol{M}$ ，第 3 四分位数 $\boldsymbol{Q_3}$ を求めよ．

図 1.1 を使ったデータの数 $n = 100$ は $4m$ の場合になるから．図を参考にして計算に使う各点の値を決めると次のようになる．

$$n/4 = 25,\ n/4 + 1 = 26,\ n/2 = 50,\ n/2 + 1 = 51,\ 3n/4 = 75,\ 3n/4 + 1 = 76$$

これから次のように計算して結果が得られる．

$\boldsymbol{M} = \frac{1}{2}(x_{50} + x_{51}) = \frac{1}{2}(54.7 + 54.8) = 54.75$
$\boldsymbol{Q_1} = \frac{1}{4}(x_{25} + 3x_{26}) = \frac{1}{4}(51.6 + 3 \times 51.7) = 51.675$
$\boldsymbol{Q_3} = \frac{1}{4}(3x_{75} + x_{76}) = \frac{1}{4}(3 \times 57.2 + 57.3) = 57.225$

例題 1–2　次のデータ列について，第 1 四分位数 $\boldsymbol{Q_1}$，メジアン $\boldsymbol{M}$ ，第 3 四分位数 $\boldsymbol{Q_3}$ を求めよ．解答は (　) 内に記した．

1. $\{3.2, 4.8, 14.0, 17.2, 22.8\}$
   (4.8, 14.0, 17.2)
2. $\{20.5, 30.5, 39.0, 46.5, 57.5, 59.0, 70.5, 80.5\}$
   (36.875, 52.0, 61.875)
3. $\{10.1, 10.7, 10.8, 11.2, 11.8, 12.5, 12.5, 12.8, 13.3, 13.8, 14.0, 14.7, 15.5, 16.3\}$
   (11.35, 12.65, 13.95)
4. $\{80.0, 80.0, 88.0, 92.8, 100.0, 108.8, 118.4, 129.6, 136.0, 144.8,$
   $146.4, 161.6, 176.0, 185.6, 192.0\}$
   (96.4, 129.6, 154.0)

1.1.6　四分位数と関係する用語

■パーセンタイル

四分位数ではデータを 4 つに分割する境目を考えるが．データを 100 分割して，100 分位数に相当する概念も**パーセンタイル (percentile)** と呼ばれてしばしば使われる．四分位数との関係では，第 1 四分位数が 25 パーセンタイル，メジアンが 50 パーセンタイル，第 3 四分位数が 75 パーセンタイルに相当する．

■ヒンジ

四分位数 $\boldsymbol{Q_1}, \boldsymbol{Q_3}$ を求める手順はやや面倒なので，**ヒンジ** と呼ばれる値が使われることもある．この場合にも，次のように下側と上側の 2 つのヒンジがあり，それぞれ $\boldsymbol{Q_1}, \boldsymbol{Q_3}$ と近似的に一致する．

下側ヒンジ (lower hinge)　メジアン以下のデータのメジアンを指す．
上側ヒンジ (upper hinge)　メジアン以上のデータのメジアンを指す．

$\boldsymbol{x} = \{1, 2, 3, 4\}$ の場合, 下側ヒンジ = 1.5, 上側ヒンジ = 3.5 であり，$\boldsymbol{x} = \{1, 2, 3, 4, 5\}$ の場合, 下側ヒンジ = 2, 上側ヒンジ = 4 となる．データ数が偶数の場合，メジアンはデータ点に含まれないので，メジアンよりも小さいデータを使って下側ヒンジを求めてい

る．上側についても同様．

■**五数要約，箱ひげ図**

データの最小値，第 1 四分位数，メジアン，第 3 四分位数，最大値の 5 つをまとめて，**五数要約 (five number summary)** と呼ぶ．これによって，データ全体の幅，中央，全体の半数が入っている領域をつかむことができる．なお，五数要約の定義として**第 1 四分位数と第 3 四分位数の代わりに下側ヒンジと上側ヒンジを使うこともある**．いずれにしても大きな違いは出ないので，実際上の不都合はない．

表 1.1 のデータについては，すでに例題 1–1 で第 1 四分位数，メジアン，第 3 四分位数が求めてあるので，それに最小値と最大値を付け加えて，五数要約は次のようになる．

$$43.6,\ 51.675,\ 54.75,\ 57.225,\ 64.6$$

■**箱ひげ図**

五数要約をグラフィカルに表した**箱ひげ図 (box and whiskers plot, box plot)** がしばしば用いられる．図 1.1.6 に，代表的な箱ひげ図の形とその各部の意味を示した．数値は表 1.1 のデータを用いている．箱ひげ図を使うと，データ集合の分布の様子が視覚的によく分かる．

なお．箱ひげ図の形や表現する内容は統一されてはおらず，形や向きを変えたり，後述する外れ値を表示するなど，使う目的とセンスによってさまざまな描き方がある．

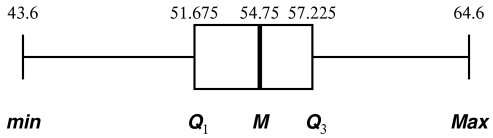


図 1.2　箱ひげ図．第 1 四分位数 $\boldsymbol{Q_1}$，メジアン $\boldsymbol{M}$，第 3 四分位数 $\boldsymbol{Q_3}$，を箱で表し，両端の「ひげ」で最小値 ***min*** 最大値 ***Max*** の位置を表す．

■**四分位範囲 (IQR)**

第 3 四分位数から第 1 四分位数を引いた値を**四分位範囲 (IQR**[*7]**)** といい，その半分の値を**四分位偏差**という．データの半数が含まれる幅を意味する量である．

[*7] Interquartile Range の略．

■外れ値

集団から遠く離れたデータのことを**外れ値 (outlier)** という．外れ値についての一致した数学的定義はなく，いくつかの基準が提唱されている．その中では，四分位数と関連付けた外れ値の定義[*8]がわかりやすく，次のように定義される．

データ x は次の条件のいずれかを満たすときに外れ値という．

$$x < \boldsymbol{Q_1} - k(\boldsymbol{Q_3} - \boldsymbol{Q_1}) \text{ または } x > \boldsymbol{Q_3} + k(\boldsymbol{Q_3} - \boldsymbol{Q_1})$$

言い換えれば，$\boldsymbol{Q_3} - \boldsymbol{Q_1} = \mathbf{IQR}$ だから，データが第 1 四分位数あるいは第 3 四分位数の外側に $\mathbf{IQR}$ の k 倍よりも遠く離れているときに外れ値と定義している． ここで k は必要に応じて 1.5〜3 ととる．

1.1.7　メジアンや分位数は頑健な代表値

■お年玉の金額の分布から

図 1.3 は，小学生 25 人がもらったお年玉の仮想的なデータを使って作った箱ひげ図である．一応現実的なデータに合わせるために現実の調査データを参考にしてある[*9]．計算に使ったデータは下の通りだ (単位 100 円)，

87, 143, 149, 163, 180, 186, 186, 212, 222,
247, 251, 255, 257, 261, 271, 274, 277, 281,
287, 296, 306, 347, 406, 449, 1300

平均では約 2 万 9 千円のところ，13 万円ももらった小学 2 年生がひとり含まれている．やはり子どもの世界でも，お金に関してはごく少数の「持てる者」が突出した金額を手にしているようだ．それをそのままプロットして見たのが，左の **A** である．見てのとおり，極端な外れ値が現れている．

この外れ値をいじって，2 番めの最大値と同じ程度にしてみたのが **B** だ．箱ひげ図を見ると，メジアンも第 1，第 3 四分位数も変化していない．

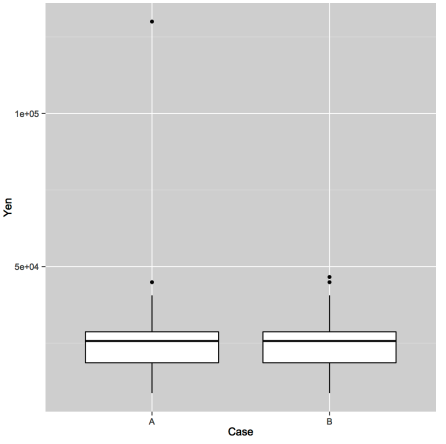


図 1.3　ある市の子どものお年玉の金額の分布を表した箱ひげ図 **A:** 大きな外れ値あり，**B:** 外れ値を修正してみたもの．

[*8] http://people.richland.edu/james/lecture/m170/を参照.
[*9] 川崎信用金庫「お年玉とお正月調査について」(2012)

こんどは，**A** と **B** の平均と標準偏差を比較してみよう．すると，平均値は 29,172 円から 25,836 円へと 3,300 円も下がり，標準偏差は 22,027 円から 8,920 円へと大幅に縮小している．このように外れ値 1 個のために，平均も標準偏差も少なからぬ影響を受けることが分かる．

このように，平均や分散は，大きな外れ値の存在によって敏感に変動する性質をもっている．一方，メジアンや四分位数は外れ値があっても，あまり，場合によっては全く，動かないことが分かる．このように「鈍感」であることを**頑健 (robust)** であると表現することがある．英語の読みのままで**ロバスト**ということもしばしばある．

1.1.8 残る命は何年だろうか

たとえば，ある病気にかかって手術を受けた人がいたとして，予後を知るために医学的な統計データを見たとしよう．データの中には，手術後の生存期間の情報をまとめたものもある．この人が頼りにするべきは，生存期間の平均だろうか，それともメジアンだろうか？

この治療の後で，かなりの人が 10 年程度生存し，15 年，20 年と生きた人もいたとしよう．しかし，2 割の人は 1 年以内に亡くなったものとする．すると余命の平均は約 8 年程度だが，メジアンの方は 12 年というといったケースが起こりうることになる．

こんな状況でこの人はどのように判断するのが賢明だろう？平均よりもメジアンを目安に考えるほうがよいのではないだろうか．「治療後のケアに十分な注意を払って，短命に終わることを避ければ．メジアンのところまでは行けそうだ」—そう考えることと，平均値を見て「あと 8 年の命か」と考えることとを比較すれば，このことは理解できるだろう．

こんなふうに，メジアンは「全体の真ん中あたり」という，いわば「並み」のポジションを表現しているものと考えられる．この後で扱う度数分布においては，このことがさらにはっきりと現れることになる．

1.2　度数分布

1.2.1　度数分布でデータを表す

生の数値を並べただけでは，これらのデータのもつ特徴をそこから直観的に見てとることは難しい．そこで，この種のデータを整理するために，**度数分布表 (frequency distribution table)** がしばしば使われる．度数分布表は，個々の数値を**表 1.2** のように**階級 (class)** に分けて，その**度数 (frequency)** を示したものである．度数は**頻度**ともいう．

また，ある階級までの度数の和の累計を**累積度数**という．

表 1.2　100 人の体重の統計を表す度数分布．表 1.1 のデータを使って構成した．

階級	階級値 (x_i)	度数 (f_i)	累積度数 (F_i)
43.0 – 45.0	44.0	1	1
45.0 – 47.0	46.0	3	4
47.0 – 49.0	48.0	6	10
49.0 – 51.0	50.0	9	19
51.0 – 53.0	52.0	21	40
53.0 – 55.0	54.0	12	52
55.0 – 57.0	56.0	19	71
57.0 – 59.0	58.0	15	86
59.0 – 61.0	60.0	10	96
61.0 – 63.0	62.0	2	98
63.0 – 65.0	64.0	2	100

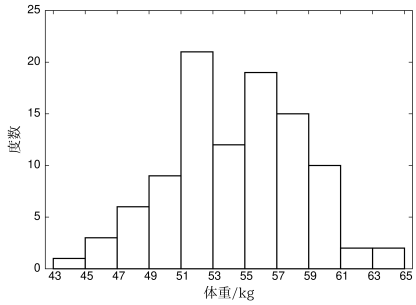


図 1.4　100 人の体重の統計を表すヒストグラム

また，度数分布をグラフで表して視覚的に把握しやすくしたものを**ヒストグラム (histogram)** という．表 1.2 の度数分布からは，図 1.4 の形のヒストグラムが作れる．

度数分布の表やヒストグラムを見ると，この集団の統計的な特徴を大づかみに見て取ることができる．すなわち，このデータによれば，中央付近の階級が大きな度数を持つ分布であり，平均はおよそ 53 から 55 の間に入るのではないかというふうに一目で推測できる．

1.2.2 度数分布から統計量を求める

度数分布表は集団のすべてのメンバーから得たデータを区分によって縮約したものである．その過程でいくらか情報量は失われるが，平均，メジアン，分散（と標準偏差）は，ほぼ正確に求めることが出来る．以下でその方法を考えよう．

■平均

度数分布から平均を求めるにはどうしたらよいだろうか．**表 1.2** を見てみよう．まず，体重の和は次のようにばらして書けることに注意する．

$$\begin{aligned} \text{総体重} &= \overbrace{44.0}^{1} + \overbrace{46.0 + 46.0 + 46.0}^{3} \\ &+ \overbrace{48.0 + 48.0 + 48.0 + 48.0 + 48.0 + 48.0}^{6} + \cdots \end{aligned} \tag{1.7}$$

これから同じ階級値の数値をまとめてやると，平均値は次のようにして計算できる．

$$\frac{\text{総体重}}{\text{総人数}} = \frac{44.0 \times 1 + 46.0 \times 3 + \ldots + 64.0 \times 2}{1 + 3 + \ldots + 2} = \frac{5446}{100} = 54.46 \tag{1.8}$$

式 (1.8) で得られる平均値は，個別のデータではなくて，階級という「塊」にまとめたものを使っているのであるから，幾分かの誤差を含むはずである．しかし多くのデータを扱う場合には，誤差は打ち消しあって十分に小さくなるので，ほぼ正しい平均値が得られる．

ここで式 (1.8) を一般化しておこう．データは k 個の階級に分けられており，階級値を $x_1, x_2, ..., x_k$，その度数を $f_1, f_2, ..., f_k$ とする[*10]．すると，上の例にならって，平均値を次のように表すことができる．

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i = \sum_{i=1}^{k}(x_i \times \frac{f_i}{n}) \tag{1.9}$$

[*10] ここでは x_i が個々のデータの値ではなく，階級値であることに注意.

ここで n は $\sum_{i=1}^{k} f_i$，つまりデータの総数である．式 (1.9) は平均を表す式を 2 通りに表現したもので，2 つ目の表現は次の形をしていることに注意してほしい．確率分布でもこれによく似た形のものが表れる．

$$\text{平均} = (i \text{ 番目の階級値} \times i \text{ 番目の階級の割合}) \text{ の和}$$

■分散と標準偏差

分散は，式 (1.4) の定義を使えば次のようになる．

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 \times f_i}{n} = \sum_{i=1}^{k}(x_i - \overline{x})^2 \times \frac{f_i}{n} \tag{1.10}$$

2 番目の表現はやはり次の形をしている．

$$\text{分散} = (\ i \text{ 番目の階級の偏差の } \mathbf{2} \text{ 乗} \times i \text{ 番目の階級の割合}) \text{ の和}$$

ここでも，分散の計算については，1.1.4 節で扱った場合と同様にして，2 乗の平均から平均の 2 乗を引けば求められる．

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \overline{x}^2 \tag{1.11}$$

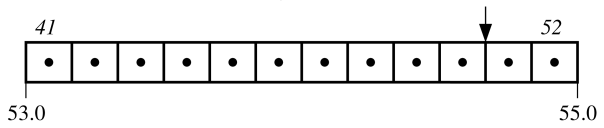
問題 1–3　式 (1.11) を利用して，表 1.2 のデータから体重の分散を求めなさい．

■メジアン

度数分布表から**メジアン**を求めるにはどうしたらよいだろうか．そのためには，ちょうど中央に位置する人の体重（総数が偶数の場合には中央の二人の体重の中間）を推定すればよい．この場合には 50 人目と 51 人目の人のデータの中間を推定したい．

累積度数を目安にして表を見ていくと，階級 (53.0–55.0) に 41 人目から 52 人目までの 12 人がいることがわかる．つまり，53.0 と 55.0 を両端とする区間の中に，12 人が並んでいるわけである．この並び方は等間隔ではないが，仮に等間隔と仮定して計算すればよい．

下のようにこれらの 12 人を並べたとすると，下の図のように考えて，50 人目と 51 人目の人の境目の位置は次の式で計算できる．



$$53.0 + \frac{2.0}{12} \times 10 = 54.666...$$

こうしてメジアンとして 54.67 が得られた．この値は真の値 54.75 にかなり近い．

■モード

この度数分布では，最も多くの度数をもつ階級の階級値は 52.0 である．このとき，この分布の**モード (mode)** あるいは**最頻値**は 52.0 であるという．モードも集団の代表値のひとつである．

1.3　平均，メジアン，モード，どれが全体を代表するのか

多数のデータの代表値としてもっともよく使われる統計量はいうまでもなく平均である．私たちの目に触れる社会の統計データでメジアンやモードという言葉が使われることはめったにない．それでは平均というのは，その集団を代表するパラメータとして他の 2 つの代表値よりも「優れて」いるのだろうか．

1.3.1　平均所得は「ふつうの所得」を意味するか

図 1.5 は，総理府が毎年実施している国民生活基礎調査のレポートから，各世帯ごとの所得金額の分布のヒストグラムを引用したものである．

これを見ると全世帯の平均所得は 552.3 万円であり，メジアン (中央値) はそれより 100 万円以上低い 437 万円となっている．モードは書かれていないが，所得が 200〜300 万円の階級が最も多いので，階級値をとって約 250 万円とみなすことができる．つまり，この所得分布の代表値をみると，モードがもっとも小さく，その上にメジアンが来て，そして平均が最も大きな値をとっている．

この状況をまとめると次のようになるだろう．

- モードから判断すると，年間所得が 200 万円台の階層が最も多数を占める．

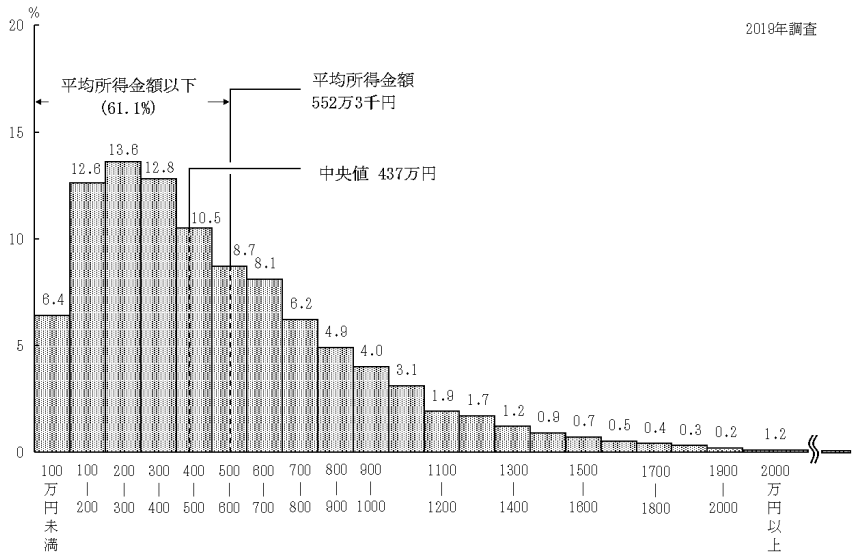


図 1.5　世帯ごとの所得金額の分布：2019 年度国民生活基礎調査

- メジアンから判断すると，真ん中の世帯は 437 万円の所得を得ている．
- 低所得者層から高所得層までを平均した所得は 552.3 万円である．

モードはもっともありふれた階層とみなせるから，これは「ふつうの世帯」としてランキングされるとも言える．またメジアンは，いうまでもなく「真ん中」だ．アンケートなどによく出てくる「中の中」という人々にあたるといってもよい．ところが平均はこれらよりもかなり高い所得のところにあるわけで，国民の実感とはかなりずれているわけだ．このようなことは，平均よりもずっと離れたデータの方が，平均に近いデータよりも，平均値に強く影響を与えることから起きる．それについて，少し考えておこう．

仮に図 1.5 の分布に，年間所得が平均よりも 100 万円多い 652.3 万円の世帯が 0.1% 加わった場合を考えてみよう．それによって平均がどうなるかは，次のように簡単に計算できる．

$$552.3 \times 0.999 + 652.3 \times 0.001 = 552.4$$

単位は 1 万円としてある．結果，1000 円だけ上昇するわけだ．今度は，年間所得が平均よりも 1 億円多い 10,538 万円の世帯が 0.1% 加わった場合で同じ計算をしてみると，552.3 万円，つまり 10 万円上昇することになる．

国民の所得は，最低のゼロから始まって上は何億円という人もいるはずだ[*11]．ということは，平均よりも桁外れに多い所得を得ている人が少数だけ存在しているために，実感とはかけ離れた平均所得が統計に現れることになる．

このように，この種の統計では得てして平均の値がさも一般庶民のものであるかのようにメディアでも政治でも扱われがちだが，それは実は虚像であって，モードやメジアンの方が，国民の実態を反映した代表値であると考えるのが妥当であろう．

1.3.2　試験の成績の分布は

表 1.3 は国立教育政策研究所の平成 14 年度高等学校教育課程実施状況調査[*12] から引用した学力試験の結果の一部である．

まず国語の成績について平均点，メジアン，モードを度数分布から知ることができ，それぞれ 15.9, 17, 19 となっている．この試験では，23 点満点に近い成績が多かったために，平均は下位の得点に引っ張られることになってしまい，真ん中に位置する点数（メジアン），仲間がいちばん多い点数（モード）よりもかなり平均点が低くなっている．この例でも所得の統計のときと同様に，平均の値がどちらか一方に大きくずれた階級があるために，引きずられてしまうということが起きている．

[*11] 筆者はそのへんの想像力もお金の知識もないので，貧しい想像で書いている．

[*12] `http://www.nier.go.jp/kaihatsu/katei_h14/index.htm`

一方，数学の成績について平均点，メジアン，モードを求めると，それぞれ 8.0, 8, 15 となっていて，モードが極端な値をとってしまっている．これは数学のやさしい試験で起こりがちな結果で，一応わかっている人は満点や高得点を取れるが，かなりの数の数学を不得手とする階層がいるために，低得点の側にもうひとつ山ができるのである．このようなときにはモードはほとんど意味をなさない．メジアンは安定して真ん中の人がどの辺なのかを示している．このことは図 1.6 のようにデータの分布をグラフにしてみるとよくわかる．

以上のように，平均は，そこから遠く外れたデータからの影響を受けやすいが，メジアンは「真ん中」がどこかということを的確に示すという意味では，「よい代表値」であると言える．とはいえ，このような分布を得た場合に教育関係者や政治家が考えるべきことは，数学がきわめて苦手な生徒たちの大きな集団が我が国の学校に存在することを認識して，どのような方策をとるべきかということだろう．ヒストグラムの形は，統計的代表値では表しきれない情報を含むことがしばしばある．

表 1.3　全国の高校生対象の学力試験 (2012 年) における国語と数学の得点分布

	国語		数学	
素点	人数	累計	人数	累計
0	54	54	420	420
1	14	68	730	1150
2	27	95	983	2133
3	52	147	1118	3251
4	76	223	1122	4373
5	118	341	1045	5418
6	149	490	950	6368
7	169	6 59	993	7361
8	286	945	948	8309
9	343	1288	889	9198
10	424	1712	937	10135
11	518	2230	907	11042
12	678	2908	873	11915
13	737	3645	943	12858
14	927	4572	999	13857
15	1023	5595	1551	15408
16	1171	6766	–	–
17	1268	8034	–	–
18	1426	9460	–	–
19	1450	10910	–	–
20	1278	12188	–	–
21	1030	13218	–	–
22	569	13787	–	–
23	151	13938	–	–

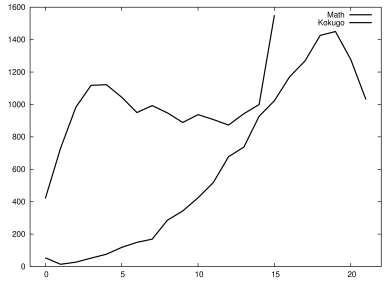


図 1.6　表 1.3 のデータから作成した得点分布．

問題 1–4　図 1.5 で表される世帯所得分布のデータから，メジアンの半分の所得になる世帯が下位何 % に来るのかを概算しなさい[*13].

問題 1–5　ある幼稚園の構成人員の年齢を調べたところ次のようになっていた．平均とメジアンを求めて，これら 2 つの代表値の違いについて考えなさい．

年齢	3	4	5	6	22	25	46	49	70	75
人数	15	28	31	15	1	1	1	1	1	1

[*13] この値は大まかには相対的貧困率に対応するものであるが，公式の相対的貧困率の定義は世帯の可処分所得と世帯人数を使って求めるものであり，これとは異なる値になる．詳しくは貧困統計ホームページ (`https://www.hinkonstat.net/`) の解説やデータを参照のこと．

第 2 章

初等的な確率論

統計学において，集められたデータの信頼性や有効な利用を保証するのは確率論である．

そのためにまず，集合と論理について簡単なまとめをしておく．確率は一般に「事象」を扱うものであるから，事象そのものの「集まり」を概念化して扱うことは有益である．

その後，中学，高校で扱った確率について基本的な概念を掘り下げて，整理しておこう．

2.1 集合と論理代数

2.1.1 集合とは何か

あいまいさなく定義できる**要素 (element)** を集めたものを**集合 (set)** という．要素のことを元ということもある．

要素としてはどんなものでもよいので，たとえば「すべての人間の集合」は，人間の範囲をきちんと定義すれば，集合とみなせる[*1]．「すべての整数の集合」，「すべてのアミノ酸の集合」のように，要素の定義が厳密になされたものの場合には，その集合は数学的に意味がある．

一方，「美人の集合」というのは，数学的な集合としては扱えない．「美人」の定義というのは人によって，文化や時代によって様々であるから，一人の女性（たぶん美人は女性に限ると思う）がその集合に含まれるかどうかはあいまいだからである．同じように，「大きな数の集合」も数学的には意味をなさない[*2]．

[*1] 「人間の範囲をきちんと定義する」というのは実際問題としてはそう単純ではない．生きている人間のことを指すのか，そうだとすればこの瞬間に生まれようとしている人間，死ぬ間際の人間，いったいどこまでを範囲に含めるべきなのかを考え始めると厄介なことになる．そういう境界領域の問題というのは，物事を厳密に考えようとするとしばしば現われる．このように現実的な問題を数学で扱うためには，適切な区切りを付けることが必要になる．

[*2] ただし，このように境界があいまいな集合であっても，「ファジーな集合」という名の下に研究の対象となっている．

2.1.2　集合とその演算

■集合の表現 — ベン図

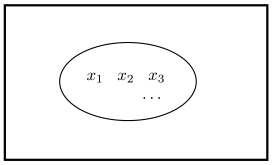
集合を A, その要素を $x_1, x_2, \ldots$ とする．これを次のように表すことにしよう．

$$A = \{x_1, x_2, \ldots\} \tag{2.1}$$

x が A の要素であることは次のように表す．

$$A \ni x, \text{ または } x \in A \tag{2.2}$$

さらに，これを図にすると次のような感じになる．



ただし，以後必要がないときには，集合の要素をいちいち描かないで，単に丸や四角の「囲み」だけで集合を表すことにする．このように集合を閉じた囲みで表し，それらの関係を表現した図を**ベン図 (Venn diagram)** という．[*3].

■空集合

要素をひとつも含まない集合を**空集合 (empty set/null set)** という．何も含まないものを集合というのは変な感じがするかもしれないが，数字の 0 と同じように，空集合という概念を用いることで，集合の体系的な記述が可能になるのである．空集合は，記号 $\emptyset$ で表す[*4].

[*3] イギリスの論理学者 John Venn が 1880 年の論文で集合の論理的な関係の表現にこの種の図形的な方法を用いたことから（彼自身によってではなく），“Venn Diagram” という名前が使われるようになった．ただし，13 世紀にはすでに類似の表現方法が現れ，さらに 17 世紀には著名な数学者ライプニッツが，そして 18 世紀には大数学者オイラーが同様の図的表現を使っている．Venn 自身も，上記の論文の中でオイラーの表現をオイラー図 (Euler Diagram) と呼んで，彼自身の表現にはユークリッド図 (Euclid Diagram) と名を付けていた．Venn の図はその後も改良されていき，1918 年のアメリカの Clarence Irving (C.I.) Lewis の書物に初めて “Venn Diagram” という名称が現れた．この経緯については下の “What is a Venn Diagram” という記事に詳述されている．以上を踏まえて，ここでは集合とそれらの関係を表す図を総称して「ベン図」という呼称を用いることとする．

`http://www.lucidchart.com/pages/tutorial/venn-diagram`

[*4] ギリシャ文字のファイの小文字 ϕ を使う人もいるが，正しくない．

■部分集合

集合 A と集合 B があり，B の任意の要素が A に含まれるならば B は A の**部分集合 (subset)** であるといい，次のように表記する．

$$A \supset B, \text{ または } B \subset A \tag{2.3}$$

空集合 $\emptyset$ あるいは A 自身も A の部分集合である．また，

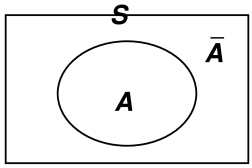
$$A \supset B, \text{ かつ } A \neq B \tag{2.4}$$

のとき，B は A の**真部分集合 (proper subset)** であるという．たとえば，犬全体の集合は哺乳類全体の集合の真部分集合になっている．

■全体集合, 補集合

なんらかの集合 A を想定するとき，同時にA に含まれない要素の全体を暗黙に，あるいは明示的に想定することがしばしばある．このとき，考えられる要素すべてを含む集合を S とし，これを**全体集合**という名で呼ぶ．

S に含まれて，A に含まれない要素の集合を A の**補集合 (complementary set)** といい $\bar{A}$ で表す．$S, A, \bar{A}$ の関係をベン図で表すと下のようになる．

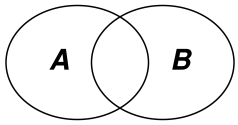


A の補集合は A に含まれない要素の集合を意味するのであるから，論理的には**否定**を意味することになる．

全体集合は，取り扱っている問題によって決まるものである．つまり，たとえば A としてすべての犬の集合をとったとき，S がすべての哺乳類の集合になるのか，すべての動物の集合になるのかは，どんな議論をしているのかによる．

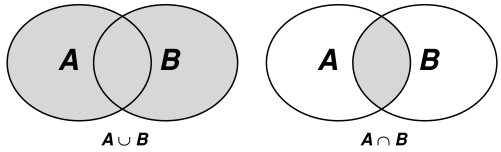
■集合の二項演算

2 つの集合 A, B が与えられており，これらは共通の要素を含んでいるものとしよう．このとき，A, B の関係はベン図で下のように表される．



このとき，A または B の少なくとも一方に含まれている要素の集合を $A \cup B$ と表し，これを 集合 A と B の**和集合 (union)** といい，「A または B」ということもある．

また，A および B の両方に含まれている要素の集合を $A \cap B$ と表し，これを 集合 A と B の**共通部分 (intersection)**[*5]といい，「A かつ B」ということもある．下の 2 つのベン図に和集合と共通部分を示した．



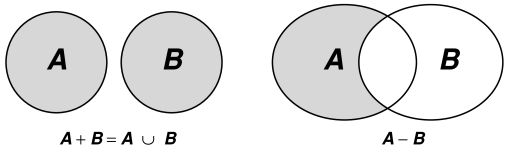
集合 A, B が共通の要素を含まない，つまり

$$A \cap B = \emptyset \tag{2.5}$$

であるとき，これらは互いに**素**であるという．

集合 A, B が互いに素であるとき，$A \cup B$ を $A + B$ で表し，集合 A, B の**直和**という．

また，集合 A の要素から集合 B の要素を除いて得られる要素全部の集合を $A - B$ であらわし，集合の**差** という．

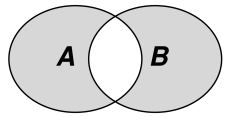


[*5] 積集合ということもあるが，最近では共通部分という言い方が一般的になってきている．

問題 2–1 次の関係をベン図で確かめなさい．ここでは補集合を扱っているので，A, B 両方を含む全体集合 S を考える必要がある．

$$A - B = A \cap \bar{B}$$

問題 2–2 下図のように，集合 A, B の和集合から共通部分を除いた集合を A, B の**排他的論理和 (exclusive or)** という．A, B に対して補集合，和集合，共通部分をとる演算を組み合わせることで，排他的論理和を表しなさい．



2.1.3　集合の演算規則

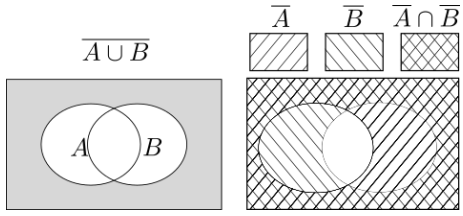
集合 A, B, C が与えられたとき，次の関係式が成り立つ．これらはベン図を描いてみれば直観的に理解できる．

$$\begin{aligned}
&\textbf{交換法則} && A \cup B &&= B \cup A \\
&&& A \cap B &&= B \cap A && (2.6)\\
&\textbf{結合法則} && A \cup (B \cup C) &&= (A \cup B) \cup C \\
&&& A \cap (B \cap C) &&= (A \cap B) \cap C && (2.7)\\
&\textbf{分配法則} && A \cap (B \cup C) &&= (A \cap B) \cup (A \cap C) \\
&&& A \cup (B \cap C) &&= (A \cup B) \cap (A \cup C) && (2.8)
\end{aligned}$$

次の**ド・モルガンの法則 (De Morgan's laws)** もよく知られている．これは論理を扱うときにも非常に重宝な関係式である．

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \tag{2.9}$$
$$\overline{A \cap B} = \bar{A} \cup \bar{B} \tag{2.10}$$

式 (2.9) が成立することを，下のベン図で示す．その次も同様に考えればよい．



ド・モルガンの法則は 3 つ以上の部分集合についても成立する．すなわち，

$$\begin{aligned}
&\overline{A \cup B \cup C \cup \cdots} = \bar{A} \cap \bar{B} \cap \bar{C} \cap \cdots \\
&\overline{A \cap B \cup C \cup \cdots} = \bar{A} \cup \bar{B} \cup \bar{C} \cup \cdots && (2.11)
\end{aligned}$$

このことは確率の計算でもよく使われるので，ベン図で表したイメージをしっかり持っておこう．

問題 2–3　50 人のクラスの中に，眼鏡をかけている人が 27 人 (集合 **A**)，携帯を持っている人が 36 人 (**B**) いる．眼鏡をかけておらず，携帯ももっていない人は 9 人だった．

1. 全体集合を S として，この状況をベン図で描きなさい．
2. 眼鏡をかけていて，かつ携帯をもっている人は何人か．

2.2 集合と確率

2.2.1 経験的確率と数学的確率

一般の数理科学の理論では，定量的な結論が得られれば，それにもとづいた現象はかなり正確に予測できる．たとえば，天体の運動のようにニュートンの運動方程式に従う現象では，次の皆既日食の日時や場所がきわめて精密に予測されている．現実が理論の意味を証言してくれるのである．

確率の場合にも，ある事象が起きる確率をきちんとした数値で与えてくれるという意味で，その結果は定量的である．しかし，その事象の起きる確率が p であったとしても，それが実際に起きるかどうかについての確実な予想は $p = 0$ または $p = 1$ であるとき以外にはできない．つまり確率の値が何を意味するかということと実際に起きることとの間に，正確な対応が付けられるわけではないのである．

したがって，確率という名で与えられる数値が何を意味するのかということは，議論の前提としてあらかじめ考えておく必要がある．

■経験的確率

気象観測は長年行われているので，過去のデータを多数調べることで，たとえば 8 月上旬に台風が何個上陸するかという平均値が得ることができる．平均値に影響するような大規模な気候変動が起きない限りは，その平均値を使うことで，稲が被害を受ける確率を計算して災害に備えることができるだろう．プロ野球選手の打率の数字にしても，それを確率として作戦を立てることが行われている．このように多数回経験した事象から推定される確率のことを，**経験的確率**と呼ぶ．

経験的確率は，私たちには分からない未知の要因によってランダムに現われてくる現象から来るものであり，得られる数字も幅，つまり不確かさを持つことになる．しかし，そのような限界をもつ数値でもあっても，よく吟味して活用することはきわめて有用なことである．

経験的確率の考えをさらに推し進めると，もっと積極的に過去の事象を将来予測に使おうという立場も現れてくる．じゃんけんを例にとるならば，相手が出す手は 3 種類のうちのひとつだから，どの手も 1/3 の確率で出ると考えるのではなく，相手の過去の実績から「くせ」を判断して，より積極的に確率を評価していくという戦略を想定するとよい．その場合，初手に関してはどの手も 1/3 という予測を行うものの，その後は成り行きに従って使うべき確率の値は変化することになる．このような立場に拠って考えられた確率を**ベイズ確率 (Bayesian probability)** という．

ベイズ確率は，現代の IT 技術においてもスパム（迷惑）メールの判定などに活用され

るなど，近年になって大きな進展を見せており，ベイズ統計の分野が新しい発展をみせている．注目される世界なので，関心のある方は情報を探ってもらうと興味深いものが見つかるだろう．

■数学的確率

理想的なサイコロやカードを使ってゲームをすることを考えると，ある事象の起きる確率は正確に計算できる．たとえば，サイコロを振って 1 の目が出る確率は 1/6 である．これは起こり得るすべての事象つまり，「1 の目がでる」，「2 の目が出る」，... ,「6 の目が出る」という 6 通りの事象のうち，ひとつの事象が実現する確率として計算されているわけである．

一般化するために，ある問題について起こりうるすべての事象の集合 S と，その部分集合をなすある特定の事象の集合 A を考えよう．S と A の要素の数をそれぞれ n, m とする． S の要素となる事象がどれも同じぐらいの確かさで起きるとするとき，A に属する事象が起きる確率 $P(A)$ は

$$P(A) = \frac{m}{n} \tag{2.12}$$

で与えられる．このように定義される確率が**数学的確率**である．

ここで，「要素となる事象が同じぐらいの確かさで起きる」という前提がなぜ成立するかということは，数学的な議論の対象にはしない．このようにある与えられた前提を出発点（公理）として，論理な操作によって体系を作り上げるのが，**公理的方法**と呼ばれる数学の基本的な構築の仕方である．

2.2.2　集合論から確率概念へ

確率を扱うときには，起こり得る事象を要素とする集合を考える．この集合のことを**全体集合**といい，また確率論では**標本空間**という言葉を用いる[*6]．2.1.2 節の部分集合を表す Venn 図で，四角の枠で表現されている集合 S は，暗黙に全体集合のイメージを持たせたものである．

確率論は集合論に基礎付けられているので，集合論の用語に対応のつく確率の用語が多い．表 2.1 にそれらの対応関係を示した．

表 2.1　集合と確率の用語の対照表

記号	集合論	確率論
S	全体集合	標本空間
	部分集合	事象
$\emptyset$	空集合	空事象
	補集合	余事象
	和集合	和事象
	共通部分	積事象
	互いに素	排反

[*6] 数学的に厳密な確率論は旧ソ連のコルモゴルフが公理論的に定式化したものであるが，ここでは数学的な厳密性を避けて，直観的にわかりやすい説明を取っている．数学としての確率論に興味のある人は，より専門的な入門書をあたっていただきたい．

2.2.3 基本的な確率の公式

事象を A あるいは $A_1, A_2, \ldots$ で表す．起こり得るすべての事象の集合を標本空間と呼び，S で表すことにしよう．このとき，確率 $P(A)$ は次の定義に従うものとして定める．

■確率の基本定義

標本空間 S は考えられるすべての事象の集合なので，その実現確率は 1 である．

$$P(S) = 1 \tag{2.13}$$

$A_1, A_2, \ldots$ が互いに排反であるとき，確率は足し合わせられる．

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots \tag{2.14}$$

このとき直和の記号を使えば，次のように簡単に書ける．

$$P(A_1 + A_2 + \ldots) = P(A_1) + P(A_2) + \cdots \tag{2.15}$$

ここからいくつかの関係が導かれる．

■空事象の確率

$$P(\emptyset) = 0 \tag{2.16}$$

これは直観的にも明らかだが，背理法で考えれば簡単に証明できる[*7]．

■余事象の確率

次の関係はとても重要で，しょっちゅう使われる．これは $S = A + \bar{A}$ からすぐにわかる．

$$P(\bar{A}) = 1 - P(A) \tag{2.17}$$

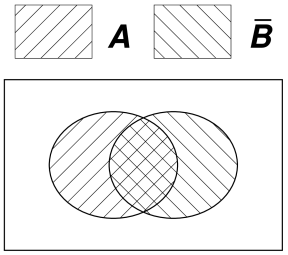
■和事象の確率

事象 A と B が与えられたとき，次の式が成り立つ．

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.18}$$

下のベン図で考えてみよう．

[*7] もし $P(\emptyset) > 0$ であったとしよう．S と $\emptyset$ は共通の要素を持たないので互いに排反である．したがって，式 (2.14) より，$P(S \cup \emptyset) = P(S) + P(\emptyset) > 1$ となり，また $S \cup \emptyset = S$ であるから，$P(S) > 1$ となって，式 (2.13) に反する．



$P(A)+P(B)$ は共通部分 $A\cap B$ を 2 回数えてしまっているので，その部分を引いているわけだ．

2.3 条件付き確率

2.3.1 条件付き確率の定義

ある事象 A が起きているとした上で他の事象 B が起きる確率を $P(B|A)$ で表し，このような確率を**条件付き確率 (conditional probability)** という．$P(B|A)$ については式 (2.19) が成り立つ．図 2.1 を使ってこのことを説明しよう．

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{2.19}$$

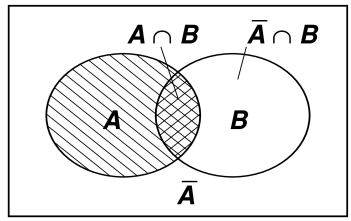


図 2.1 条件付確率の意味を Venn 図で表す

図のそれぞれの集合に含まれている根元事象の数が次のようになっているものとしよう．すなわち，全体集合に n，A に n_1，B に m，$A \cap B$ に m_1，$\bar{A} \cap B$ に m_2 のように．このとき式 (2.19) の左辺は，A に含まれる事象のうちで B に含まれている事象の起こる確率だから，次のようになる．

$$P(B|A) = \frac{m_1}{n_1}$$

右辺の分母と分子はそれぞれ，すべての事象の中で A に含まれる事象の起こる確率，すべての事象の中で $A \cap B$ に含まれる事象の起こる確率を使って，次のように表すことができる．

$$P(A) = \frac{n_1}{n}$$

$$P(A \cap B) = \frac{m_1}{n}$$

したがって，式 (2.19) が成り立つことになる．ただしこの式は，図を見て考えれば直観的に理解できるものである．

なお，条件付き確率が扱われるとき，上の議論における $P(A)$ は，前提となる確率という意味で**事前確率 (prior probability)** と呼ばれ，また $P(B|A))$ は**事後確率 (postrior probability)** と呼ばれることがある．

2.3.2　乗法定理と独立事象の公式

式 (2.19) を次のように変形してみよう．

$$P(A \cap B) = P(A)P(B|A) \tag{2.20}$$

この式は，2 つの事象がいずれも起こる確率を求めるための公式になっており，確率の**乗法定理**と呼ばれる．

ここでもしも次の式が成立するとしよう．

$$P(B) = P(B|A) \tag{2.21}$$

すると，式 (2.20) は次のように書き換えられる．

$$P(A \cap B) = P(A)P(B) \tag{2.22}$$

この意味を考えてみよう．$P(B) = P(B|A)$ が成り立つということは，**事象 $\boldsymbol{A}$ がすでに起きていたときに事象 $\boldsymbol{B}$ が起きる確率 $\boldsymbol{P(B|A)}$ が，$\boldsymbol{A}$ が起きたか否かを問わず $\boldsymbol{B}$ が起きる確率 $\boldsymbol{P(B)}$ と等しい** ということである[*8]．たとえば，2 組のカードのセットがあって，一方から引いたカードがハートであった (事象 A) として，その後で別のセットから引いたカードがハートではない (事象 B) 確率はどうなるだろうか？常識的に考えて，2 組のカードの山同士は無関係なのだからその確率は変わるはずがない．

2 つの事象 A と B があって，A が起きたことが B の起きる確率に影響を与えず，その逆も成り立つものとする．このとき，事象 A と B は**独立**であるという．2 つの独立事象が両方とも起きる確率は，それぞれの起きる確率の積に等しい．

ただし数学的には，式 (2.22) が成立するときに，事象 A, B が独立であるという[*9]．

2.3.3　独立性と因果関係

前節の話を，こんな例を用いて考えよう．

> ある日，犬を見た人が，たまたま風邪を引いたとする．となると，人によっては「自分が風邪を引いたのは，犬を見たことと関係があるのではないか？」という疑

[*8] 論理的には，式 (2.21) だけではなく A と B をひっくり返した形の $P(A) = P(A|B)$ も成り立たないと，A, B が独立であることはいえない．しかし，片方が成り立てばもう一方も同時に成り立つことを，簡単な計算で確かめることができる．

[*9] 確率を現実の現象を扱う道具として使う立場からは，2 つの事象が独立であるかどうかは，それらを引き起こす原因の間に何らかの関係があるかどうかという観点で議論される．たとえば，ある場所で雪が積もることとそこで人が道で転ぶことの間には間違いなく関係がある．しかし，数学はそのような個々の事情から離れて抽象化された議論をするものであるから，「事象 A と B が独立である」という命題を式 (2.22) が成立することの必要十分とみなすのである．

問にとらわれる人も出てくるだろう．それでは，この人にとって犬を見るという事象 (A) と風邪を引くという事象 (B) とは関係があるのかどうか，どうやって判断すればよいのだろうか．

そのためには十分に多数回の経験をしてみて，その上で確率を考えてみることになる．

さて，仮に犬を見ると風邪をひきやすいということがあったとしたら[*10]，次のような関係が成立することになるだろう．左辺は犬を見たとして，風邪をひくという条件付き確率で，右辺は犬を見たかどうかとは無関係に風邪をひく確率だ．

$$P(B|A) > P(B) \tag{2.23}$$

これに式 (2.19) の条件付き確率の定義の式を代入すると，

$$\frac{P(A \cap B)}{P(A)} > P(B)$$

より，

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

となる．つまり，事象 A と B がどちらも起きる確率は，それぞれが起きる確率の積よりも大きいということになる．まさにこれは，2 つの事象の間に何らかの因果関係があることを感じさせる．

一方，式 (2.23) で等号が成立していたらどうだろう．それが意味することは，**犬を見たときに風邪をひく確率と，犬がいようがいまいが風邪をひく確率は変わらない**ということだ．つまり，犬を見るという事象と風邪をひくという事象は独立であるということになる．そしてこのときには，

$$P(A \cap B) = P(A)P(B) \tag{2.24}$$

が成立する．

つまり，**2 つの事象の実現が独立であるときにのみ，事象が両方とも実現する確率がそれぞれの事象が起きる確率の積となる．**

2.4　ベイズの定理

条件付確率については，**ベイズの定理 (Bayes' theorem)** という実用的にも重要な定理が知られている[*11]．

[*10] もちろんそんなことはないだろうが，縁起をかつぐ人はそんなふうに思い込むものだ．

[*11] 「なんとかさんの定理」というと，たいていは著名な数学者の名前から来ているものだが，ベイズという人は 18 世紀初めにイギリスに生まれた牧師で，数学者としては無名だったようだ．彼は条件付き確率の問題の特定の場合について考察を遺しており，死後にそれは発表された．

簡単なケースについて考えてみよう．事象 A, B が互いに排反で，かつ標本空間を尽くしているとする．すなわち，次のように S が A, B の直和になっているとする．

$$S = A + B$$

このとき，A, B それぞれの下に，ある事象 E が起きる確率，

$$P(E|A), P(E|B)$$

が知られているとする．

今，E が起きたとして，それが A によるものである確率 $P(A|E)$ は次のようになる．これがベイズの定理である．この定理は，事象 $A, B, C, \ldots$ と拡張してもそのまま成立する．

$$P(A|E) = \frac{P(A)P(E|A)}{P(A)P(E|A) + P(B)P(E|B)} \tag{2.25}$$

ここで想定している状況について例を挙げて説明しておこう．飲酒の習慣があるか，そうでないかという前提条件の違いは，適当な区切りを設定すれば，お互いに排反な事象とみることができる．そこでこれらをそれぞれ A, B としよう．飲酒の習慣は，当然のこととして健康への影響があると思われるので，ある病気にかかる事象を E とすると，お酒を飲む人がこの病気にかかる確率 $P(E|A)$ と飲まない人がかかる確率 $P(E|B)$ には違いがあるはずだ．たいていの病気ではたぶん $P(E|A) > P(E|B)$ つまり，お酒は体に悪いということになるのかも知れない．

さて，**もしもある人を任意に選んで，その人がこの病気にかかっていたとしよう．この人がお酒のみである確率はどれだけか？**この確率は，条件確率 $P(A|E)$ で与えられる．これがベイズの定理で設定している状況である．

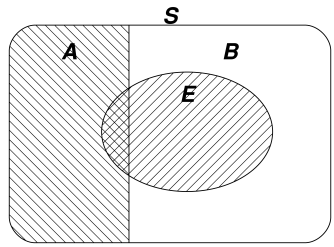


図 2.2　ベイズの定理の説明図

図 2.2 のベン図を使って考えてみよう．状況を見やすくするために，条件付き確率の説明に登場した式 (2.19) の $\bar{A}$ を B に，B を E に置き換えて少し変形すると，次の最初の

式が得られ，2 つめの式も同様に得られる．

$$\begin{aligned} P(A \cap E) &= P(A)P(E|A) \quad (A \text{ と } E \text{ の共通部分が実現する確率}) \\ P(B \cap E) &= P(B)P(E|B) \quad (B \text{ と } E \text{ の共通部分が実現する確率}) \end{aligned}$$

図 2.2 の楕円形で区切りの左側と右側が，それぞれ式 (2.26) の $P(A \cap E)$, $P(B \cap E)$ に相当している．この 2 つを合わせたうちに占める $P(A \cap E)$ の割合が，E が実現しているときにそれが A によるものである確率 $P(E|A)$ であるから，式 (2.25) が得られることになる．

なお，式 (2.25) のベイズの定理は，簡単な計算で次のように変形することができる[*12]．

$$P(A|E) = \frac{P(A)P(E|A)}{P(E)} \tag{2.26}$$

この式のほうが見かけ上簡単であるために，本によってはこちらを紹介しているものもある．しかし，確率 $P(E)$ の値は $P(E|A)$ や $P(E|B)$ のように現実のデータとしては想定しにくいので，必ずしも使いやすいとは言えない．

2.4.1 ベイズの定理の例

A, B の二つの農場でトマトを市場に出荷していて，消費者はそのどちらかだけを購入しているものとする．A 農場では毎日 800 個のトマトを出荷し，そのうちの 5% は虫食いである．B 農場では毎日 2000 個のトマトを生産していて，そのうちの 12% が虫食いであるとする．

今，トマト 1 個を無作為にとって，それが虫食いだったとする．このトマトが A 農場のものである確率はどうなるか．

A 農場のトマトである確率を $P(A)$, A 農場のトマトであって虫食いである確率を $P(E|A)$, 虫食いのトマトが A 農場のものである確率を $P(A|E)$ とする． B 農場のほうについても同様に定義する．すると，ベイズの定理から次のように結果が得られる[*13]．

[*12] $P(E) = P(A)P(E|A) + P(B)P(E|B)$ という関係を使えばよい．この式は**全確率の法則 (law of total probability)** と呼ばれ，$P(A) + P(B) = 1$ の時に，より一般的には $P(E) = P(A)P(E|A) + P(B)P(E|B) + \ldots$，つまり互いに共通部分を持たない事象 $A, B, C, \ldots$ が標本空間を覆うときに成立する．

[*13] この種の計算を行うときには，$\frac{800}{2800}$ のような分数は計算しないで進めること．そうすれば約分できて計算が劇的に簡単になる！

$$\begin{aligned} P(A|E) &= \frac{P(A)P(E|A)}{P(A)P(E|A)+P(B)P(E|B)} \\ &= \frac{\frac{800}{800+2000}\times 0.05}{\frac{800}{800+2000}\times 0.05+\frac{2000}{800+2000}\times 0.12} \\ &= 0.143 \end{aligned}$$

2.4.2　ベイズの式を使わない考え方

ベイズの公式 (式 2.25) は抽象的であってやや理解しづらいし，また逆に式に数値を当てはめるだけで中身も分らないで操作するということにもなりやすい．もっと具体的な考えに立って上の問題を解いてみよう．いわば**頻度**を使った方法である．

問題のような状況で，トマトを 100 個，無作為に買ってきたとしよう[*14]．そのうち，何個ずつが 2 つの農場から出荷されたものかを考えると，その期待値は次のようになる．

$$\begin{aligned} &\text{A 農場産の個数} = 100\times\frac{800}{800+2000} = 28.571 \\ &\text{B 農場産の個数} = 100\times\frac{2000}{800+2000} = 71.429 \end{aligned} \tag{2.27}$$

これらのうちの虫食いの数はそれぞれ，

$$\begin{aligned} &\text{A 農場産のうちの虫食いの個数} = 28.571\times 0.05 = 1.4286 \\ &\text{B 農場産のうちの虫食いの個数} = 71.429\times 0.12 = 8.5715 \end{aligned} \tag{2.28}$$

したがって，虫食いのうちで A 農場産のものである確率は次のようになる．

$$\frac{1.4285}{1.4285+8.5715} = 0.143 \tag{2.29}$$

このように，ベイズの公式に当てはめたのと同じ結果が得られた[*15]．これならベイズの定理を覚えていなくても解ける．このように頻度を用いた計算は，直感的にわかりやすいので便利である．

[*14] 100 のかわりに他の考えやすい数でもかまわない．
[*15] これは当然のことで，このやり方を一般化したのがベイズの定理である．

2.5　医療・疫学と確率統計

現代の医療においては，**根拠に基づく医療**という概念がきわめて重視されている．英語では **Evidence-based medicine** であり，略して **EBM** と書かれることが多い[*16]．

この概念は広く医療の倫理と哲学を含むものであるが，医療統計においては，公正に行われた検査データに正しい統計処理をすることで，ある治療法や薬が真に有効なのかどうかを判定する作業が中心となる．たとえば昔から効くと信じられてきた薬でも，その効果が明瞭でないまま漫然と使われていることがありえて，以下に述べるような手法で真の効果があるかどうかを見直す作業がずっと進められている．

以下では，治療の効果を判断するのに使われるオッズ比の考え方と，潜在的に病気を持っている人を検査によってふるい分けるスクリーニングの手法について述べる．ここでベイズの定理はきわめて重要な道具になっていて，しばしば **2 × 2 分割表**が使われる．

2.5.1　オッズ比—治療の効果を検証する

何かの病気にかかった人にある薬を飲んでもらって，その効果を検証するような場合を想定しよう．このとき，被験者は全員が「薬」を飲まされるが，その中には一部**偽薬**[*17]を飲まされる人が，本物を飲む人に混じっている．本物を飲む事象を A とし，また E の方は，本物であれ偽物であれ，薬を飲んで一定の日数後に治癒したという事象であるとする．

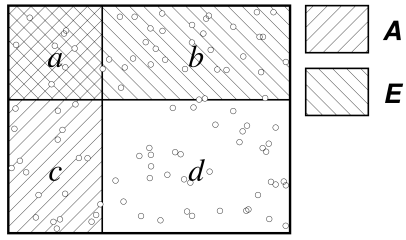


図 2.3　オッズ比と条件付き確率を考えるための図．矩形の領域は，左上から時計回りに $A \cap E$, $\overline{A} \cap E$, $\overline{A} \cap \overline{E}$, $A \cap \overline{E}$ の集合を表し，それぞれに含まれる根元事象の数は a, b, c, d である．

十分に多数の被験者を使って得た，この試験の結果が図 2.3 のようになったとしよう．与えられている a, b, c, d を使うと，次のように確率が得られたことになる．

[*16] EBM に対する補完的な概念として **NBM**（**Narrative-based Medicine**）がある．興味ある人は調べてみていただきたい．

[*17] **プラセボ** (placebo)，あるいはプラシーボともいう．

$$\begin{aligned} P(A) &= \frac{a+c}{a+b+c+d} \\ P(E) &= \frac{a+b}{a+b+c+d} \\ P(A|E) &= \frac{a}{a+b} \\ P(E|A) &= \frac{a}{a+c} \end{aligned} \tag{2.30}$$

もしも，このときに事象 A と事象 E が独立，つまり薬を飲んでいようがいまいが治癒する確率は変わらなかったとすると

$$P(E|A) = P(E)$$

が成立するので[*18]，式 (2.30) 以下を使って整理すると次の式が得られる．

$$\frac{a}{c} = \frac{b}{d} \tag{2.31}$$

この式の意味を考えよう．$\frac{a}{c}$ という値は，本物の薬を飲んでいたとして，治癒した数 a をしなかった数 c で割ったものである．この値を**オッズ (odds)** という[*19]．もう一方の薬を飲まなかった方についても．同じようにオッズ $\frac{b}{d}$ が定義される．

薬を飲んだ場合のオッズと，飲まなかった場合のオッズとの比の値 (前の数を後の数で割った値) を**オッズ比 (odds ratio)** という．つまり，

$$\text{オッズ比} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

である．

ここで扱った場合では薬の効果がなかったことを想定していて，そのときオッズ比は 1 になっている．もしも薬の効果があったとすれば，飲んだ方のオッズは大きくなり，飲まなかったほうは変わらないはずだから，オッズ比は 1 よりも大きくなることになる．

このように，オッズ比が 1 よりも大きいかどうかは，治療薬の客観的な効き目を判定する上で重要な指標になる．

例題 2–1　ある疾患の新しい治療薬の効果を試すために，100 人の罹患者に対して処方したところ，20 人だけが治癒した．そのことをもってこの薬には効果がないと判断してよいか．

[*18] p.36 の乗法定理参照．

[*19] オッズというのはなじみがないかも知れないが，賭博で使われる言葉である．ある勝負があって，それに勝つ確率を負ける確率で割ったものをオッズという．たとえばコインを 2 枚投げて，両方とも表になったら勝ち，そうでなかったら負けというルールだったら，勝つ確率は 1/4 だから，オッズは $\frac{1/4/}{3/4} = \frac{1}{3}$ になる．

一見すると，この薬では治ったのは全体の 1/5 だから，効果がなかったようにも感じられる．ちなみにこの場合のオッズは $20/80 = 0.25$ である．**しかし，薬を飲ませなかった場合にどの程度の人が治癒するかについてのデータがないので，これだけで効果の有無を判断することはできない．**客観的に効果を判定するためには，本物を投薬された人と偽薬を飲まされた人の両方でデータをとってみて，オッズを比較することが必要だ．

その結果が右の **2 × 2 分割表**のようになったとする．

	投薬	偽薬
治癒	20	11
非治癒	80	87

オッズ比を計算してみると，次のように 1 よりもかなり大きい値になるので，この薬は有効であると推測することができる．

$$\frac{20}{80} \div \frac{11}{87} = 1.98$$

実際の臨床試験においては，被験者も実験者も本物の薬と偽薬のどちらを扱っているか分からないように実施する．これを**二重盲検法 (double blind test)** という．

さらにいうと，試験データは罹患者全体を母集団としてそこから抽出して得たものと考えられるので，結果は統計的ゆらぎをもつことになる．そこで真に有効かどうかを判定するための統計的な検定を行う必要がある．それについては第 8 章で詳しく学ぶ．

2.5.2　スクリーニング

ある疾病[*20]ははじめ無症状であるが，放置するといつか重い症状が現れるものとしよう．高血圧はその典型的な例であり，またがんも初期には自覚症状がないのがふつうである．また，この疾病にかかる人が少なくなくて，早期に治療を開始することで健康に生きられるようになるとか，あるいは死亡率を下げられるものとしよう[*21]．

上のような条件を満たす疾病に対して，一般の人を対象にして行われる検査を**スクリーニング (screening)** という．早期に発見して治療することでその人個人の幸福を守り，かつ医療費を引き下げようというわけだ．自治体や職場で実施している健康診断の検査項目の多くはスクリーニングであり，現在は無症状でも，疾病にかかっている可能性のある人をふるい分けるために行なわれている．

自覚症状はないが，検査対象の疾病にかかっている人を**有病者**，そうでない人を**非有病者**とよぶ．

一方，検査の結果がいわゆるクロ，つまり疾病を持っているという判定になることを**陽性 (Positive)** と呼び，反対のシロ，つまり疾病なしと判定されることを**陰性 (Negative)** と呼ぶ．注意してほしいのは，実際には有病者であっても結果は陰性という判定が起こりえて，それを**偽陰性 (False Negative)** という．反対に，実際には非有病者であっても陽性の判定になることを**偽陽性 (False Positive)** という[*22]．

したがって，単に「陽性」と言った場合には，有病者と非有病者の両方を含む可能性があるので，陽性から偽陽性を除いたものを**真の陽性 (True Positive)** と呼び，同様に陰性から偽陰性を除いたものを**真の陰性 (True Negative)** と呼ぶ．

通常の疾患では，スクリーニングの対象になる疾病に罹患している人の割合，つまり**有病率**はすでに調べられて分かっている．

つまり，ランダムに人を選んで検査した場合にその人が実際に疾病を持っている確率は**事前確率**として分かっている．また行なわれる検査についても，この後で出てくる**感度**と**特異度**がすでに調べられていることが前提になる．

スクリーニングの結果を理解するには表 2.2 の 2×2 分割表が便利である．また図 2.4

---

*20 「しっぺい」と読む．

*21 がんについては，この条件に当てはまるものはすべてではないことに注意してほしい．現在国が推奨しているがん検診は，胃がん，子宮頸がん，乳がん，肺がん，大腸がんの 5 つだけである．つまり早期発見して治療することのメリットが，症状のない人が検診をうけることで生じるデメリットを上回るのは，この 5 つだけである．かつて「早期発見で早期治療」ということが盛んに言われたが，上のがんについては確かにそのとおりでも，他のがんについてまで適切とはいえない．詳細は国立がん研究センターのがん情報サービスを参照のこと．

*22 結核菌に感染しているかどうかを調べるためのツベルクリン反応の検査がある．この検査では結核菌の培養液から精製した成分を皮内注射して，抗原抗体反応による発赤の大きさを調べる．発赤が大きければ既に感染したことがあるので陽性，小さいかないかだったら陰性，中間で判断が難しい場合は**疑陽性**とする．疑陽性は偽陽性と読みが同じだが意味はまったく異なることに注意しよう．

に，これらの関係を図として表現した．このとき，検査の結果得られる数値は $a+b$ と $c+d$ で表される陽性か陰性の数だけであって，$a,\ldots,d$ が個別に得られるわけではないことに注意しよう．また，検査によって知りたいのは，本当はどれだけ疾病があるのか，またはないのかを示す $a+c$ と $b+d$ である．

表 2.2 スクリーニングの結果
a： 有病者 かつ 結果が陽性 (真の陽性)
b：非有病者 かつ 結果が陽性 (偽陽性)
c： 有病者 かつ 結果が陰性 (偽陰性)
d：非有病者 かつ 結果が陰性 (真の陰性)

		疾病の有無		
		あり	なし	合計
検査結果	陽性	a	b	$a+b$
	陰性	c	d	$c+d$
	合計	$a+c$	$b+d$	

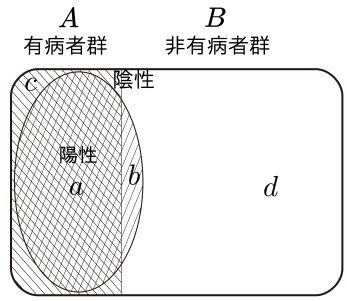


図 2.4 スクリーニングのベン図

2.5.3 スクリーニングの結果の評価

表 2.2 の a,b,c,d はそれぞれ，真の陽性，偽陽性，偽陰性，真の陰性の出現数を意味している．これらから得られる主なパラメータを表 2.3 に示した．これらは，ある検査法の質を考えるのに使われるが，特に次の 2 つは重要である．

感度 有病者のうちで陽性になる割合．**敏感度**，**鋭敏度**ともいう．
感度が高いということは，有病者ならば高い割合で検査で引っかかるということだ．逆にいうと，感度の低い検査では，その疾病に罹患しているのにそのままになってしまうことになりやすい．

特異度 非有病者のうちで陰性になる割合
特異度が高い検査の場合，検査でシロとなったらまず大丈夫だろうと安心できることになる．

理想的な検査法は a,d のみが値をもち b,c はゼロとなるものである．しかし，現実には

表 2.3 スクリーニングで使われる主なパラメータと確率: A,B は図 2.4 参照

$(a+c))/(a+b+c+d)$	有病率, prevalence	$P(A)$
$a/(a+c)$	感度（敏感度，鋭敏度）sensitivity	$P(E\|A)$
$d/(b+d)$	特異度，specifity	$1-P(E\|B)$
$a/(a+b)$	陽性的中率，Predictive value Positive	$P(A\|E)$
$d/(c+d)$	陰性的中率，Predictive value Negative	$P(B\|\overline{E})$

偽陰性や偽陽性のない検査法はなく，図 2.4 のように，陽性の人のほとんどが正しく有病者であり，陰性の人のほとんどは無病であるというものがよい検査法であろう．ただし，50 ページで触れるように，特異度がかなり高くても，陽性的中率は低くなることがあることに注意しないといけない．

■感度と特異度はトレードオフの関係にある

感度も特異度も高い検査は，罹患しているのに救済できない人が少なくて，検査で大丈夫となったら安心できることになる．しかし，普通はそうはいかない．この 2 つが両立するのは難しいのである．そのことを図 2.5 で説明しよう．

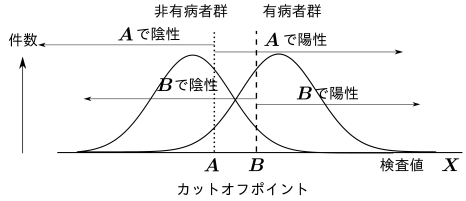


図 2.5　感度，特異度とカットオフポイントの関係

検査においては，被験者の血液や尿の中の特定の成分の濃度といった何らかの数値（検査値）によって判定することが多い．たとえば LDL コレステロールの濃度が 140 mg/dL 未満なら正常，それ以上なら高 LDL コレステロール血症と診断される[*23]．しかし，人間には個人差があるために，非有病者群と有病者群ではまったく検査値が異なるわけではない．図にあるように，非有病者と有病者で検査値はそれぞれの異なる分布をもつものの，分布には重なる部分が出てくる．その場合，ある検査値を示した被験者がどちらに属しているかは決定できないのである．

その状況でも何かの判断基準を置いて，その基準未満であれば陰性，基準以上ならば陽性と判定する[*24]．これがカットオフポイントと呼ばれるものである．

図には $\boldsymbol{A}, \boldsymbol{B}$ 2 通りのカットオフポイントが示されているが，これらの意味を比較してみよう．$\boldsymbol{A}$ を基準とした場合，検査値 $\boldsymbol{X}$ が $\boldsymbol{A}$ よりも小さく出た人は非有病者と判定され，残りが有病者となる．$\boldsymbol{B}$ についても同様だ．この 2 通りのとり方は判定にどう影響するだろうか．

カットオフポイントに $\boldsymbol{A}$ を採用した場合，非有病者群のうちざっと 3 割ほどは $\boldsymbol{A} \leq \boldsymbol{X}$ となり，陽性と判定されることになる．実際には有病でないのだから，これは偽陽性だ．このとき陰性 $(\boldsymbol{A} > \boldsymbol{X})$ と判定された人は，有病者群の左裾のわずかな部分に入ってしまう確率があるものの，おそらく自分は有病ではないと安心してよいだろう．

[*23] 厚生労働省 e-ヘルスネットの記載 (2020) から．
[*24] 疾病や検査によってはこの関係は逆になるかもしれない．

それでは $\boldsymbol{B}$ を採用した場合はどうなるだろうか？この場合には，陰性と判定された人の中にかなりの数の有病者が含まれてしまう．病気のある人を見逃す確率が高いということは，検査の目的からするとまずいことだ．しかし陽性と判定されたら，ほぼ有病者であると考えてよい．

つまり，カットオフポイントが $\boldsymbol{A}$ の場合には，検査の感度は高くなって有病者の多数が陽性と判定される（これはよいことだ）．一方，非有病者でも陽性の判定を受けてしまう（これは迷惑な話だ）．

$\boldsymbol{B}$ の場合には，検査の特異度が高くなり，非有病者の大半は「シロ」判定されるものの，有病者の中には誤って陰性と判定される偽陰性の人が多くなる結果となる．もしも，病気があるのに陰性となってしまったら治療の機会を失うことになるので，それは避けたいところだ．結局，カットオフポイントの決め方としては，上のようなトレードオフをにらんで，なるべく多くの有病者を救済しつつも，偽陽性と判定されて二次検査で無駄足を踏まされたり，余計なコストがかかることのない点を選ぶことになる．

もしも，スクリーニングが図 2.6 のように有病者と非有病者を完全に分けてくれれば，カットオフポイントの設定は中間の適当なところにすればよいのだか，それは医学の進歩によって少しずつ理想に近づくものであろう．

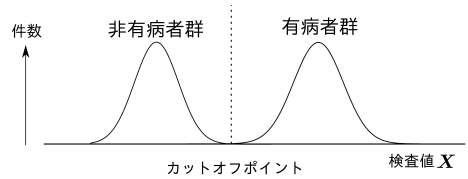


図 2.6　非有病者群と有病者群が完全に分けられる理想的な検査

2.5.4　スクリーニングとベイズの定理

次の問題を考えてみよう．

例題 2–2　スクリーニングを受診した人が，陽性と判定されたとする．その人が実際に病気を持っている確率はどれだけになるだろうか？また，逆に陰性と判定された人が，実は有病者である確率はどれだけか？表 2.3 の a, b, c, d を用いて表わせ．

この問題は，すでに出てきたベイズの定理の問題とまったく同じ形である．

図 2.4 を見て考えると，有病率を**事前確率** $P(A)$ として，条件付き確率 $P(E|A)$ は有病の人が陽性になる確率で，$P(E|B)$ は無病の人が陽性になる確率であるから次のように表わせる．

$$\begin{aligned} P(A) &= \frac{a+c}{a+b+c+d} \\ P(B) &= \frac{b+d}{a+b+c+d} \\ P(E|A) &= \frac{a}{a+c} \\ P(E|B) &= \frac{b}{b+d} \end{aligned} \tag{2.32}$$

よって，陽性になった人が本当に有病である確率は，

$$P(A|E) = \frac{P(A)P(E|A)}{P(A)P(E|A)+P(B)P(E|B)} = \frac{a}{a+b}$$

となる．こうして得られたのが表 2.3 の**陽性的中率**である．

次に，陰性と判定された人が本当に非有病者である確率は

$$P(B|\overline{E}) = \frac{d}{c+d}$$

となり，これは**陰性的中率**と呼ばれる．

問題 2–4　スクリーニングの結果の信頼性: ある病気に感染しているかどうかを調べる検査法がある．この検査によると，被験者が非感染者なのに陽性になる確率が 1.5%，感染者なのに陰性になる確率が 0.5% ある．この病気は人口の 2% が感染しているものとしよう．スクリーニングによって，ある人が陽性と判定された．この人が本当に感染者である確率はどれだけか？

2.5.5　新型コロナウイルス肺炎とスクリーニング

2019 年 12 月，中国の武漢市で原因不明の肺炎患者が発生し，その後短期間で世界中に感染が拡大した．2010 年 3 月，WHO はこの新型肺炎の疾患名を **COVID-19** , 原因のウイルスを **SARS-Cov-2** と命名し，世界がパンデミックの状態にあることを宣言した．このテキストの執筆時点で事態は生々しく進行中であるが，21 世紀の大きな災厄として長く記憶に残ることになるだろう．

さて，日本で感染が広がりつつあった 2020 年の 2 月から 3 月にかけては，しばしばスクリーニングの実施を求める声が挙がったが，それは実施されることなく終わった．このことは，スクリーニングがどんな目的でどんな場合に必要なのかということを考えさせてくれるエピソードとなっているので，ここで詳しく検討しておこう[*25]．

[*25] COVID-19 をめぐる社会情勢や研究は，このテキストを執筆中の 2020 年 3 月から 4 月の間に急速な変貌と進展を見せており，その疫学的な問題を議論できる状況ではない．ここではきわめて限定した内容について議論していることを，読者にはご了解願いたい．いうまでもなく，ここでの議論は変化していく状況を理解する上で役に立つものであるので，考える上で活用していただければ幸いである．

まず，この疾病の特徴について見ておこう[*26]．ウイルスに感染した場合，約 80% の人は特に症状が現れず日常の行動を続けることができる．残り 20% の人が発症して，風邪のような咳や痰などの症状が現れる．さらに，そのうちで全体の 5% 程度の人は重症化して肺炎を起こす．

■低い有病率の段階での難しさ

感染拡大の比較的初期の段階では有病率は非常に低い状態にある．仮に 2000 人の感染者がいたとすると，日本における有病率は人口が 1 億人として 0.002% 程度しかない[*27]．44 ページで挙げたスクリーニングが可能な条件のうち，「疾病にかかる人が少なくない」という条件がここでは成立していない．

新しい病原体の出現で感染者が増え始めているときには，人口比でどれだけの感染者がいるのかが基礎データになる．つまり有病率は，この時点では**事前確率**として知られてはいない．そこでまずはその値を知りたいので，

$$\text{有病率} = \frac{\text{感染者数}}{\text{検査した人の数}}$$

として求めることになる．

ところが，ここで問題が起きる．感染者数を知るためには図 2.4 における $a + c$ の値が必要だ．そのためには，有病者の中の偽陰性者の数と，陽性と判定された中の偽陽性者の数がわからないといけない．結局，ほしい値を得るためのデータとして未来のデータをもってくる必要があるわけで，この段階では決められない．さらに詳しく確定診断を行ってみて，それから推定していることになる．一般に何らかのパラメータを決めるときには，行きつ戻りつのプロセスを経るもので，スクリーニングの基礎データの取得はすぐにできるものではない．

それでもなんとかして有病率を求めたとしよう．流行の初期段階では，その値は小さいはずだ．たとえば 1 万人検査すると，有病者は 0.2 人程度になる．しかしこの数字は無作為抽出の結果であるから，当然ゆらぎが起きる．第 4.3 節 (→ p.69) で触れるが，この値は平均値が 0.2 の**ポアソン分布**に従う確率変数であり，出てきた値から有病率を推定するのは難しい．

■低い感度と高い特異度

COVID-19 の蔓延が始まってしばらくの間に，重要な疫学的データが明らかになってきた．多くの報道で次のことが言われている．診断のために行なわれる PCR(polymerase chain reaction) 検査の感度は 50–70% 程度であり，特異度は 99% である．つまり感度

[*26] これらの知見は 2020 年 3 月におけるものであり，特に 3 月 25 日に都庁で開かれた都知事の記者会見における大曲貴夫・国際感染症センター長の話から．

[*27] 2020 年 4 月 1 日までの集計では，感染が判明した人数は 2450 名である．

はかなり低く，特異度はきわめて高い．この状況で次の問題を考えてみよう．最初とは変わって，有病率がかなり上昇してきている状況になっていることを想定する[*28]．

例題 2–3　ある人が新型コロナウイルスに感染しているかどうか PCR による検査を受けた．結果が陽性であったとして，実際に有病である確率を求めよ．ただし，検査の感度は 60%, 特異度は 99% とし，この時点での有病率は 0.1% とする．

また，有病者のうち陰性と判定される人の割合を求めよ．

最初の方はベイズの定理を使っても簡単に計算できるが，ここでは 40 ページでやったように，頻度を使って概算してみよう． 1 万人を検査したとすると，その中に（期待値で）10 人の有病者がいる．感度が 60% なので，そのうち 6 人が陽性の判定を受ける．残りの非有病者は約 1 万人であり，そのうち 1% つまり 100 人が陽性と判定されるので，実際に有病であるのは 6/106 で約 6% になる．特異度が 99% もあるので偽陽性の人は少ないだろうと思いがちだが，有病率が低くて，非有病者が圧倒的に多い場合にはこういう結果が起きやすい．

なお，あとの方は，単に偽陰性率を出せばいいので，1 から感度を引いて，40% となる．有病者であってもこれだけ見逃されるのでは，個人の病気を探り出す役には立たない．検査法がスクリーニングに向いていないのである．

[*28] これを書いている時点ではそうなっていない．そうならないことを祈っている．

2.6　認識と確率

数学の分野の中では，確率はもっともよく話題にされる．しかしながら，確率に関する人間の直感にはバイアスがかかっていて，しばしば誤ったイメージを持ってしまう．数学的な内容からは逸脱するが，確率をめぐる人間の認識の問題について考えておこう．

2.6.1　一様な確率とは何か

■野球解説者の確率論

「この打者の打率は 2 割 5 分で，これまで 3 打席三振だったから，次はヒットを打つはずですね」という野球解説者はよくいるものだ．それは正しいのだろうか．打者がコンスタントに打率を維持している場合，ヒットとアウトはどのように現われるのだろう．

この疑問を確かめるために，確率 1/4 の割合で H が，残りは O が合計 50 回出現する簡単なシミュレーションを行ってみた．その結果は次の通りである．

```
O H H O O H O O O O O O H H H O O H H O O H O O O H O H O O O H H H
O O H O O O O O O H O H O O O O H O O H O O O O O O O O H O O O O O
O O O O H O O O O O O O O O O O H H O O O O O H O O O O O O O O H
```

これを見てどう思うだろうか．きっと，「案外ヒットがかたまっているものだなあ」と感じるのではないだろうか．逆に後半になると今度はなかなかヒットが出ないスランプ状態も出現している．このように，確率そのものは一定であっても，実際に出現する事象のほうはかなり偏りを見せることになる．人はこれらを，「つき」，「スランプ」，「運に見放された状態」などと呼ぶことがあるわけだ．しかしこのシミュレーションに見るように，これらは確率の自然な表れなのである．

問題 2–5　0 から 9 までの数字を，なるべくでたらめだと思うやりかたで 50 個書き並べなさい．その後，連続して同じ数が出現する確率を計算してみて，自分の「くせ」を検証してみなさい．

今度は別の実験結果を示そう．図 2.7 は 1/10 の確率でマス目に黒石を置き，残りは白石を置いたみたところをシミュレーションでやってみたものである．これを見ると，黒い石の間にまるで引っ張りあう力が働いているかのような「意味ありげなかたまり」がたくさん現われている．

■ランダムさはなぜ意味ありげに振舞うのか

以上のように，ある確率のもとにランダムに起きるはずの現象は，人間にとってはむしろある種のパターンを感じさせることが多い．この意外さはどこからくるのだろうか．

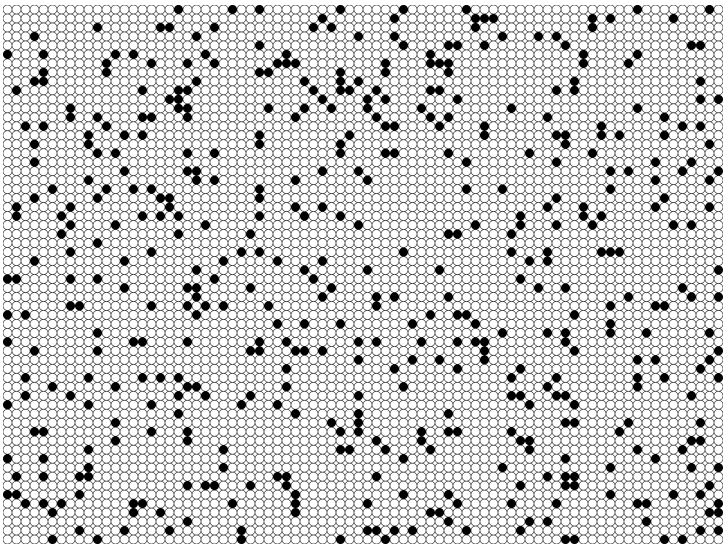


図 2.7　白 9 個に対して黒 1 個の割合で混ざった碁石をランダムに敷き詰めてできたパターン

打率が 0.25 の打者の例を考えてみよう．いま，この打者が 3 連続でアウトになったとする．さて次に起きることとして，次の 2 つの予想のどちらが正しいのだろうか．

確率の「流れ」を考える人　打率 0.25 ということは，4 打席に 1 回はヒットになるということだ．今 3 打席アウトになったところだから，そろそろヒットになってもいいはずだ．

確率は変らないと考える人　打率 0.25 ということは，どんな状況でもその次がヒットになる確率が 0.25 だということだ．だから次もアウトの確率の方が高い．

実際にシミュレーションの結果をながめてみて，どちらの予想の方が的中したのかを調べてみよう．アウトが 3 回連続している場合に，その次がアウトになっている場合とヒットになっている場合を数えてみるのである．調べてみると，アウト 3 回連続の後でアウトになっている回数は 23 回，ヒットになっている回数は 10 回となっていて，どちらかというと後の予測の方が的中している．以上を次のようにまとめておこう．

一様な確率では，次の事象はその前の影響を受けるわけではない．

●確率的に起きる現象は，「つき」，「はずれ」とか意味ありげなまとまり方を見せることが多いが，それは確率的な現象として自然なのである．

●人間は，一様な確率を，規則的な間隔で現象が起きることと勘違いしやすい，

2.6.2 降水確率 70% の意味は

朝のニュースを見ていたら，京都市の午後の降水確率が 70% であるという予報が流された．これが意味することは，次のうちのどれにもっとも近いのだろうか．

1. この日の午後には京都市の約 70% の地域に雨が降る．
2. この日の午後に京都市のどこかで雨が降る確率は 70% である．
3. この日の午後，京都市のどこかにいたとすると，雨に遭う確率は 70% である．

それぞれについて考えてみよう．1. が正しいとしたら，京都市民のうち 70% が雨に降られる（市内の人口密度が均一だったとして）ことになるからなんとなく当たっているような気がするかもしれない．しかし，天気予報では，地域の中で雨が降る場所の面積比率を予測しているわけではない．そもそもそんな観測自体できるはずがない．現在の気象状況と過去のデータから，その市の観測地点で雨が降るであろう確率を計算しているのである．

2. は明らかにおかしい．「京都市のどこかで雨が降る」事象の余事象は，「京都市のどこでも雨が降らない」ということである．これは市内のどこかで 1 滴でも雨が降ったら成り立たなくなるわけだから，少しでも雨模様が予想される日にはその確率はほとんどゼロのはずだ．つまり，ほんの少しでも雨が降りそうな日には，どこかで雨が降る確率はほぼ 1 になってしまう．降水確率が 0 になるのは，雲ひとつない快晴のときだけということになり，70% などという中途半端な確率はほとんど現れなくなるだろう．

3. について考えてみよう．降水確率 70% という予報は，過去の観測データと現在の気象状況，それにシミュレーションによる予測とから，観測地点（気象台や測候所）で降水がある確率が 70% としているのである．大雑把に言えば，ある打者が出たらヒットになる確率がどうなるかを，その打者の打率で考えるのと似たようなものである．降水確率が 70% という予想が出た日を多数回調べれば，そのうちの 70% で予報が的中しているという，経験的確率を，天気予報として発表しているといってもよい．

3. で示された解釈は，観測地点ではなくて市内の任意の場所を考えているので，多少の違いはあるが，予報というものの性格上，観測所と地理的に近い範囲ではほぼ同じ予測を適用できるとしているのだ．だからこの解釈は妥当である．

2.6.3 ホームランボールはだれかに当たる

宝くじで一等をとった人がいると，その人には特別な神通力でもあるかのように思われることがあるし，一等が出た宝くじ売り場には「ここで一等賞が出ました！」という宣伝が貼り出されて，いかにもその売り場が特別であるかのように装うものである．

昔の知り合いに全く予想もつかない偶然で再会したりすると，「不思議な力」を感じたり，「赤い糸」を想像したりするのもよくある話だ．何回かの行事の外出のたびに雨に降られると，「雨男」，「雨女」にされて，その人は雨を呼ぶ力を持っていることになるというのもよくある．これはいわゆる**ジンクス**のたぐいである．

このように低い確率の事象が起きると，何かしら驚きや必然のようなものを感じるのが人間の感覚である．このことについて考えておこう．

■千人にひとりが「必ず」当たる

大勢で勝ち抜きじゃんけんをして，10 回勝ったら景品がもらえるイベントがあったとしよう．あなたは自分がその景品をもらえることを期待するだろうか．確率は 1/1024 であるから，たいていの人は期待しないだろう．しかし，それでも約千人にひとりの「だれか」には必ず当たるのである．各人にとっては期待できない偶然であり，一方全体ではかならず誰かに当たる必然として，景品が当たるという事象がおきるわけだ．

さてこのとき，たまたま当たった人はなんらかの意味で「特別な人」なのだろうか？私たちはそのように思うことがしばしばあるし，そのような人を「運のよい人」とも呼ぶ．

しかし，もともと「運のいい人」がいるわけではない．誰が運のいい人になるのかは，決して事前には予測できないし，最初から決まっているわけでもない．あくまで結果論としてだけ，「運がよかった人」は存在するわけで，その人は確実に現われる（ただしまずあなたではないだろう）．当たった人も，外れたあなたも，事前の状態においてはまったく平等なのである．

> 運がいいとか悪いとか，人はときどき口にするけど，そういうことって確かにあると，あなたを見ててそう思う（さだまさし「無縁坂」より）

「運がいい人」をうらやむ人はいるものだが，たまたま当たった人であっても，自分がまさかそういう立場になれるとは思いもしなかったはずだ．低い確率の幸運など当てにしないで，つまり自分が運のいい人になれるだろうとは期待しないで，自分の行動を組み立てることしか，私たちにできることはない．もちろん，たまたま幸運に恵まれることはたいていの人の人生の中で何回か起きるわけだから，それを活かすか無駄にするかは，努力や心構えの問題だ．

2.6.4　偶然はだれにもコントロールできない

念力とかサイコキネシスと呼ばれるオカルト系の話の小道具がある．たとえば目力（めぢから）で机の上の鉛筆を転がしたり，何ら力を加えないのにスプーンが曲がったりするあれである．これらは全部インチキであるから，信用してはいけない．

それでも幸運を授けてくれる護符とか，あるいは強く信じれば奇跡が起きるという信念

は，この社会に根強いものがある．前世の自分や背後霊が自分の現在に何らかの影響を与えて，それらに祈って働きかけることで自分の運を変えることができるという言説も根強い．それは妥当だろうか？たとえば加持祈祷でガンは治るだろうか．

ガンの治療というのは，手術や放射線などでガン組織を取り除いて，他の病巣が残っていないか，あったとしても制圧可能な程度であることを期待するものだ．これはいわば賭けであり，治る，治らないという 2 つの排反事象の割合を少しでもよい方に動かすべく治療を行うのである．実際にどちらの事象が発現するかは，知りようがない．このような状況に直面して，加持祈祷や護符のようなものを求める人は多い．

そのような行為が不安を鎮めるために役立つことはあるだろう．また死の不安に対してどのような行動をとるかを批判することは，慎みのない行為である．しかしそれでも，人間の情念がいかに強くとも，それだけで事象の発現を左右することはできないのである．ガンを避けるためにできることは，ガンを招く生活要因を避けること，早期発見に努めること，万一ガンが見つかったら最善の治療を選択すること，— これらができることのすべてである．ガンの発生確率や再発の確率を引き下げることができる合理的な努力をするしかない．

それでも不運に見舞われるかもしれないというところが，不安の理由なのだが，それについては，「悪運を受け入れざるを得ないかもしれない」という覚悟が必要なのだ．つまり先人の言葉を借りれば，人事を尽くして天命を待つしかないのである．

以上のように，確率の問題を考えることは，リスクの中に生きる存在である私たちの行動の仕方を問うことでもある．簡単にまとめておこう．

●私たちにできるのは，確率全体を増減するように状況を変えるための合理的な働きかけや努力だけだ．
●その都度どちらの事象が偶然に現われるのかをコントロールする方法はない．
●超自然的に偶然を操る存在（神とか霊とか）もない[*29]**．**

【章末問題】

問題 2–6　式 (2.18) を使って，52 枚のカードから 1 枚を引いたときに，それが黒であるか偶数であるか，少なくともどちらかである確率を求めなさい．ただし，クイーンは偶数

[*29] ここで言っているのは，偶然を操って現実の世界の現象を左右するような存在としての神や霊はないということである．ピッチャーの球のコースは予測不可能だし，打者の身体コントロールがたまたまそれと合致したときにヒットが生まれるわけだから，それには毎回の偶然が大きく左右する．そこでサイコロをちょいと動かしてくれる神やら霊などはないのだ．私たちにできるのは，普段の訓練とその場の集中によって打球がヒットになる事象の集合を広げることだけである．まして祈ったら現世の利益が得られるなどということはない．ただし，人が超自然的な存在に対する観念や信仰をもつことをここで否定しているわけではない．信じることから得られる心の安定や健康への影響を大事だと思う人はいるかも知れない．

には含めないものとする．

問題 2–7　ある学校のクラスの学生で，誕生日が一致する人がいる確率を考える．何人以上のクラスで，この確率が 1/2 を超えるか．

第 3 章

確率分布

確率分布には，確率変数が離散的な場合と連続的な場合がある．ここではもっぱら離散的な場合に限定して解説し，連続的な場合については次章以降で導入する．

3.1 確率変数と確率関数

2 個のコインを投げて，その裏表がどうなるかを考えよう．2 個のコインの区別はしないものとし，表を A, 裏を B として，起こりうる事象 E_1, E_2, E_3 を次のように決める．

E_1：2 枚とも A
E_2：1 枚が A, もう 1 枚が B
E_3：2 枚とも B

の 3 通りであり，それぞれの確率は $1/4, 2/4, 1/4$，すなわち

$$P(E_1) = 1/4$$
$$P(E_2) = 2/4$$
$$P(E_3) = 1/4 \tag{3.1}$$

と書くことができる．現実にコインを投げて試行を重ねると，その結果は，この確率に比例した割合に徐々に近づいていくであろう．

ここで上の表し方を一歩進めて，事象 E_i の代わりに何らかの数値を使うことにすると，取り扱いが便利になる．たとえば 1 回の試行で B が現れる回数を X とすれば，

$$P(X = 0) = 1/4$$
$$P(X = 1) = 2/4$$
$$P(X = 2) = 1/4 \tag{3.2}$$

と書けるであろう．

このように，事象を数値で代表させるようにした場合の変数を**確率変数 (random**

variable または stochastic variable) といい，大文字の X, Y などでしばしば表される．X は，具体的には数値 $x_1, x_2, ...$ という値をとる．このケースのように確率変数が飛び飛びの値をとる場合は，とくに**離散的確率変数 (discrete random variable)** という．いくつか例を挙げてみよう．

- コインの表が出たら 1，裏が出たら 0 と確率変数を決めておくと，可能な値は 0 と 1
- サイコロを 1 回投げる試行では，確率変数はサイコロの目の数で，可能な値は 1 から 6 までの整数
- ランダムに選んできた 20 人の集団の中で特定の血液型をもつ人数を考えるとき，確率変数はその人数，可能な値は 0 から 20 までの整数
- 3000 人を対象の内閣支持率の調査では，確率変数は支持する人の数，可能な値は 0 から 3000 までの整数
- 宝くじでは，確率変数は当選金額，可能な値は末等から一等までの金額.

$X = x_1, x_2, ...$ に対する確率は，**確率関数 (probabilty function)** または**確率密度 (probability density)** と呼ばれる．確率関数はしばしば $f(x_i)$ のように表される．

$$P(X = x_i) = f(x_i)$$

この表記では，上の式 (3.2) の関係は

$$\begin{aligned} f(0) &= 1/4 \\ f(1) &= 2/4 \\ f(2) &= 1/4 \end{aligned} \tag{3.3}$$

となる[*1].

また，度数分布における累積度数に相当する関数は，しばしば $F(X)$ で表され，**分布関数 (distribution function)** と呼ばれる[*2]．ここの例では，

$$\begin{aligned} F(0) &= 1/4 \\ F(1) &= 3/4 \\ F(2) &= 4/4 \end{aligned} \tag{3.4}$$

となる．

[*1] ここでは離散的な確率変数を取り上げており，その場合については，$P(X = x)$ は $f(x)$ と同じものであるから，ここの定義は単なる言い換えに過ぎない．しかし後にみるように，連続的な確率変数の場合には，変数がある有限の範囲にある確率が定義されるので，ある値 x に対して $P(X = x) = f(x)$ という関係は成り立たない．

[*2] ここではいくつかの教科書を参照して，確率関数と分布関数という呼び方を紹介しているが，実際には確率関数という呼称はそれほど一般的ではなく，確率分布ということが多いようだ．

例題 3–1 2 個のサイコロを振る．このときの確率変数としては何を用いるのが適当か．またその値はどのような範囲をとるか．

2 個のサイコロの目の数の和を確率変数とするのが自然である．そのとき確率変数の値は 2 から 12 までの範囲をとる．

なお，この答えは唯一ではない．たとえば目の数の積とか，差でもかまわない．ただしそれぞれに応じて変数の範囲と確率分布は異なる．

3.2 離散的な確率関数の例 — 離散型一様分布

離散的な確率関数で最も基本的なものは，**離散型一様分布 (discrete uniform distribution)** と呼ばれるものである．

コインの投げ上げやサイコロを転がしたとき，起こり得る事象はどれも等しい確率をもつということにしよう．コイン投げであれば表と裏がどちらも 1/2 の確率で，サイコロであればどの目も 1/6 の確率で出るということにしてしまう．

サイコロの場合，目の数を確率変数とすると，

$$\begin{aligned} f(1) &= 1/6 \\ f(2) &= 1/6 \\ &\cdots \\ f(6) &= 1/6 \end{aligned} \tag{3.5}$$

となる．このようにどれも等しい確率で実現するような確率関数を**離散型一様分布**という．一般的には，根元事象の数を n として，確率関数が

$$f(x) = \begin{cases} 1/n & (x = 1, 2, \ldots, n) \\ 0 & (\text{それ以外}) \end{cases}$$

で表されるのが離散型一様分布である．

3.3 離散的な確率変数の性質

離散的な確率変数が，

$$X = x_1, x_2, \ldots, x_n$$

のように n 個の x_i の集合ですべての場合を尽くしているとしよう．すると，確率関数 $f(x)$, 分布関数 $F(x)$ について次の式が成り立つ．

$$\sum_{i=1}^{n} f(x_i) = f(x_1) + \ldots + f(x_n) = 1 \tag{3.6}$$

$$F(x_i) = \sum_{k=1}^{i} f(x_k) \tag{3.7}$$

$$F(-\infty) = 0 \tag{3.8}$$

$$F(\infty) = 1 \tag{3.9}$$

$F(x)$ は，確率（正かゼロの値しかとらない）の足し合わせで定義されているので，減少しない関数であることにも注意しておこう．

例題 3–2　コインを 3 回投げる試行を考える．表が出たコインの数を確率変数 X とする．X のとり得る値と，それらに対する確率を求めて，確率関数を書き下ろしなさい．

$X = 0, 1, 2, 3$ であり，それらに対する確率は，$1/8, 3/8, 3/8, 1/8$ であるから，確率関数は

$$f(0) = 1/8$$
$$f(1) = 3/8$$
$$f(2) = 3/8$$
$$f(3) = 1/8$$

となり，また，分布関数は

$$F(0) = 1/8$$
$$F(1) = 4/8$$
$$F(2) = 7/8$$
$$F(3) = 1$$

と書ける．

3.4 離散的確率変数の期待値と分散

多数の試行の後で，確率変数 X の平均がどの値に収束するかというのが，**平均 (mean, average)** または**期待値 (expectation value)** [*3]である．期待値は $E[X]$ または μ で表され，次のように与えられる．

$$\begin{aligned} E[X] = \mu &= \sum_{i=1}^{n} x_i f(x_i) \\ &= x_1 f(x_1) + \ldots + x_n f(x_n) \end{aligned} \tag{3.10}$$

この形は 1 章 p.17 の式 (1.9) とよく似ていることに注意しよう．

また，確率変数 X の**分散 (variance)** は $V[X]$ または σ^2 で表され，次の式で与えられる．

$$V[X] = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) \tag{3.11}$$

この形を見ると，分散というのは偏差の 2 乗の期待値であるとも言える．なお 1.1.4 節の式 (1.6) にならって分散の式を変形でき，

$$V[X] = E[X^2] - E[X]^2 \tag{3.12}$$

ここでも**分散 = 2 乗の平均 − 平均の 2 乗** となる．

例題 3–3 ある賭けで，確率 0.1, 0.05, 0.01 でそれぞれ 100 円, 200 円, 1000 円儲けることができ，また確率 0.001 で 2000 円損をするものとする．これは離散的確率分布の例であるが，確率変数はどのような値をとるか．また，儲ける金額の期待値を求めよ．この賭けによってあなたは儲けるだろうか．

2000 円の損は −2000 円の儲けと考える．また 0 円の儲けの確率は $1 - 0.1 - 0.05 - 0.01 - 0.001 = 0.839$ である．従って儲けの額は $-2000, 0, 100, 200, 1000$ であり，これらが確率変数の値 $x_1, x_2, \ldots, x_5$ に相当する．確率関数 $f(x)$ は $f(x_1) = 0.001, f(x_2) = 0.839, f(x_3) = 0.1, f(x_4) = 0.05, f(x_5) = 0.01$ である．これから期待値を求めるには，

[*3] 平均と期待値は概念上はまったく異なるものである．すなわち本来，平均というのはすでに存在する複数のデータから算出されるものであり，期待値というのはまだ実現していない事象がどうなるかについて，多数回の試行を行った結果を想定して，その平均を考えるのである．しかし，数学的には，同じように扱っても問題はない．

式 (3.10) を使えばよい．

$$\begin{aligned} E[X] &= x_1 f(x_1) + x_2 f(x_2) + \cdots \\ &= -2000 \times 0.001 + 0 \times 0.839 + 100 \times 0.1 + \cdots + 1000 \times 0.01 \\ &= 28 \end{aligned}$$

結局期待値は 28 円となるから，十分に多数回やっていれば儲かることになる．

例題 3–4　前節のコインを 3 回投げる試行の期待値と分散を求めよ．また，実際にコインを 3 回投げる試行を数十回行ってみて，結果を理論値と比較してみよ．

期待値は，

$$\mu = \frac{1}{8} \cdot 0 + \frac{3}{8} \cdot 1 + \frac{3}{8} \cdot 2 + \frac{1}{8} \cdot 3 = \frac{3}{2} = 1.5$$

すなわち，平均して 1.5 に収束する．次に分散を求めるには，

$$E[X^2] = \frac{1}{8} \cdot 0^2 + \frac{3}{8} \cdot 1^2 + \frac{3}{8} \cdot 2^2 + \frac{1}{8} \cdot 3^2 = 3$$

を求めておいて，

$$E[X^2] - \{E[X]\}^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4} = 0.75$$

となる．

例題 3–5　事象の数が n であるような離散型一様分布の平均値と分散を求めよ．

$$\mu = 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \ldots + n \cdot \frac{1}{n} = \frac{n(n+1)}{2} \frac{1}{n} = \frac{n+1}{2}$$

$$\sigma^2 = E[X^2] - \{E([X]\}^2 = 1^2 \cdot \frac{1}{n} + 2^2 \cdot \frac{1}{n} + \ldots + n^2 \cdot \frac{1}{n} - \frac{(n+1)^2}{2^2} = \frac{n^2 - 1}{12}$$

ここで高校数学の数列の問題で扱われる次の式を使った．

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

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

3.5　確率変数の期待値と分散に関する公式

期待値や分散は確率変数 X の関数であり，よく使われる公式がいくつかある．また 2 つの確率変数 X, Y が存在する場合に，それらの和や積などどのように振舞うかを知りたいこともある．それらについて，基本的な関係式を挙げておく．これらの導出は付録 A.3 に詳しく書いてあるので，数学的に理解したい人は読んでいただきたい．

- 確率変数の一次式の期待値

$$E[aX + b] = aE[X] + b \tag{3.13}$$

- 分散の基本的な性質．分散は二乗の平均から平均の二乗を引いたものである[*4]．

$$V[X] = E[X^2] - E[X]^2 \tag{3.14}$$

- 確率変数の一次式の分散．定数項は現れないことに注意．

$$V[aX + b] = a^2 V[X] \tag{3.15}$$

- 二つの確率変数の和の期待値は，それぞれの期待値の和になる．

$$E[X + Y] = E[X] + E[Y] \tag{3.16}$$

- 二つの確率変数の積の期待値は，**それらが独立なときにだけ**，積の形になる．この式と次の式 (3.18) における条件の成立については，付録 A.3 を参照のこと．

$$E[XY] = E[X]E[Y] \tag{3.17}$$

- 二つの確率変数の和および差の分散は，**それらが独立なときにだけ**，それぞれの分散の和になる．

$$V[X \pm Y] = V[X] + V[Y] \tag{3.18}$$

例題 3–6　駅伝の二つの区間を A, B の両選手がリレーするものとする．最初の区間を A 選手が走った実績は，平均 27.6 分，標準偏差が 1.5 分であり，次の区間を B 選手が走った実績は，平均 16.2 分，標準偏差が 1.2 分であるとする．二人の合計タイムの平均と標準偏差はどうなるか．また，その計算に必要とされる仮定はなにか．

[*4] $E[X]^2$ は $(E[X])^2$ であることに注意．

式 (3.16) から，平均については単に足せばよい．標準偏差については，式 (3.18) に従ってまず 2 乗して分散にしてから足し合わせて，その平方根を合成された標準偏差とる．つまり $\sqrt{1.5^2 + 1.2^2} = 1.92$ 分となる．これが成立するのは A, B 両選手の走りが互いに独立であり，相互に影響し合うことがないものとしたときにのみ成立する．標準偏差には単純な足し算が成立しないことに注意しよう．

第 4 章

二項分布

二項分布は，さまざまの状況で非常にしばしば現れる離散型確率分布である．また，二項分布を連続分布の極限へと延長すると，統計学において中心的な役割を果たす正規分布に導かれるし，「さみだれ式」の分布であるポアソン分布にもつながっている．

4.1 二項分布

4.1.1 二項分布を適用できるケース

血液型が A 型である人の割合は，日本人の場合，ほぼ 40% である．では，無作為に選んだ 10 人のうち 4 人が A 型である確率はどうなるだろうか．

10 人を次々につれてきて，最初から 4 人目までが A 型で，残りの 6 人がそれ以外である確率は，

$$0.4^4 \times (1-0.4)^6$$

となる．他の順序であっても，人数構成が同じであれば，この確率は変わらない．たとえば，最初に A 型が 2 人，次にそれ以外が 6 人，最後に A 型が 2 人でも，上の積の取り方の順序が変わって，

$$0.4^2 \times (1-0.4)^6 \times 0.4^2$$

となるだけで，確率は同じである．人を並べる順序が異なるということは互いに排反であるから，これらの確率は足し算されることになる．すなわち，A 型 4 人とそれ以外の 6 人を並べるやり方が何通りあるかを考えて，それを上の積に掛けてやれば，求めるべき確率が得られることになる．

ここで，A 型の人 4 人とそうでない人 6 人を並べるやり方は，

$${}_{10}C_4 = \frac{10!}{4!\,6!} = 210 \text{ 通り}$$

である．従って，問題の答えは

$$ {}_{10}C_4 \times 0.4^4 \times (1-0.4)^6 = 210 \times 0.4^4 (1-0.4)^6 = 0.251 $$

となる．

この導出にならって，n 回の試行において，確率 p であるような事象が x 回起きる確率関数は，一般に次の式で表される．

$$ f(x) = {}_nC_x p^x (1-p)^{n-x}, \quad (x = 0, 1, ..., n) \tag{4.1} $$

このような確率分布を**二項分布 (binominal distribution)** と呼び，しばしば $B[n,p]$ と記号で表される[*1]．二項分布は現実の現象においてしばしば登場する分布である．

例題 4–1　5 つの解答から正答 1 つを選択する方式の問題が 5 問出題されている．まったくランダムに答えを選択していった場合に，60% 以上の得点が得られる確率はどれほどになるかを計算せよ．

ひとつの問題で正答する確率は $1/5 = 0.2$ である．すると二項分布から，5 問中で正答が 5，4，3 である確率はそれぞれ，

$$ \frac{5!}{5!\,0!}\,0.2^5, \quad \frac{5!}{4!\,1!}\,0.2^4 \times 0.8, \quad \frac{5!}{3!\,2!}\,0.2^3 \times 0.8^2, $$

となるから，これらの和をとって，求める確率は 0.058 となる[*2]．

4.1.2　二項分布の期待値と分散

二項分布 $B[n,p]$ について，期待値と分散は次のようになる．平均については，意味を考えれば，この結果は自明である[*3]．なお，これらの導出は付録に与えてある．

$$ \mu = np \tag{4.2} $$

$$ \sigma^2 = np(1-p) \tag{4.3} $$

二項分布の期待値と分散は，後に出てくる正規分布との関係で非常に重要である．

[*1] $B[n,p]$ という記号の中には，変数であるはずの x は含まれておらず，分布を特徴付ける n と p というパラメータだけが書かれていることに注意しておくこと．

[*2] このケースでは，運よく合格できる確率が 5% は存在することになる．ここでもし問題の数がもっと多くて，たとえば 10 問だったとすると，合格の確率はどうなるだろうか．

[*3] 出現確率 p である現象が n 回の試行の中で平均して何回現れるかということであるから，当然 np という結果になる．もちろん，厳密な数学的導出によっても同じ結果が得られる．

4.1.3　二項分布の形

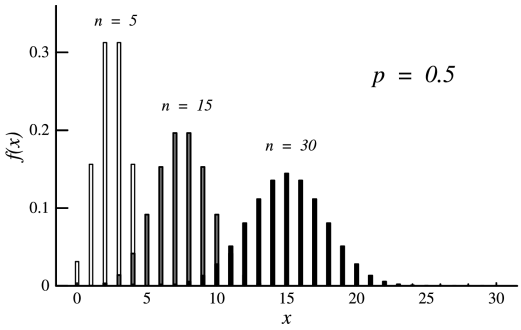


図 4.1　$p = 0.5$ のときの二項分布のようす

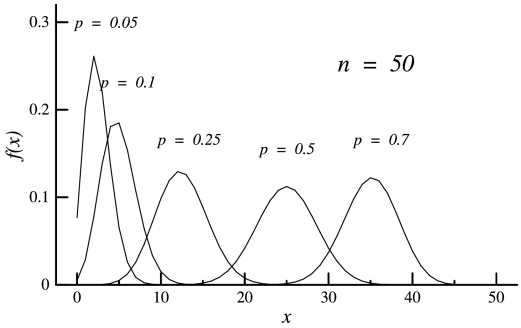


図 4.2　p を変化させたときの二項分布のようす．グラフの重なりを避けるために折れ線グラフで示した．

二項分布が実際にどのような分布になるかを，いくつかのグラフで示しておこう．図 4.1 は，$p = 1/2$ で n を 5, 15, 30 としたときの結果を図示したものである．いずれも中心に山を持ち，左右対称な分布になっている．このグラフは，コインを n 回投げる試行を行っては表の出る率を調べて，それを何度も繰り返してプロットしていったときに得られるものと同じものになるはずである．

図 4.2 は，p を変化させながら 50 回の試行を繰り返したときの二項分布のようすをプロットしたものである．np のところにモードを持つ分布になっていることが分かる．

4.2　多項分布

日本人の血液型分布は，A, O, B, AB 型の人の比率が，およそ 40%, 30%, 20%, 10% となっている．ここで 10 人の日本人がいたとすると，その血液型の構成が 2 人，4 人，2 人，2 人となっている確率はどれほどであるか．— **多項分布 (polynomial distribution)**

は，このようなケースに対して有効である．

この問題は二項分布の説明で述べたのと全く同様に考えればよい．まず，最初の 2 人が A 型，次の 4 人が O 型，続いて 2 人が B 型，最後の 2 人が AB 型となる確率は，

$$0.4^2 \times 0.3^4 \times 0.2^2 \times 0.1^2$$

で与えられる．しかし，この人数構成比になるための人の並び方の順序は，

$$\frac{10!}{2!4!2!2!}$$

通りであるから，これらの積を取れば，求める確率が得られる．

つまり，多項分布というのは，二項分布を多数の事象の場合について拡張したものに他ならない．この話を一般化すると，次のようになる．

排反な事象の完全な組，$E_1, E_2, ..., E_k$ があり, それぞれの実現確率が $p_1, p_2, ..., p_k$, ただし $p_1 + p_2 + ... + p_k = 1$ であるとする．n 回の試行において，それらが $x_1, x_2, ..., x_k$ 回ずつ実現する確率 $f(x_1, x_2, ..., x_k)$ は，次の式で与えられる．

$$f(x_1, x_2, ..., x_k) = \frac{n!}{x_1!\, x_2!\, \ldots x_k!} \times p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \tag{4.4}$$

4.3 ポアソン分布

4.3.1 二項分布からポアソン分布へ

二項分布から導かれるもうひとつの重要な分布は**ポアソン分布 (Poisson distribution)** である．これは，ひとつひとつの確率は低いものの，非常に多数の回数の試行が行われるような現象に対して適用される確率分布である．

例えば，ある窓口に対して，8 時間で平均 24 本の電話がかかってくるようなケースを考えよう．分刻みで見れば，480 分に 24 回の割合でかかってくるわけだから，ある 1 分の刻みの中でかかってくる確率は 0.05 であり，電話が来るかどうかを監視する回数は 480 回というわけである．このようなケースで，たとえば連続した 60 分間に電話が来る確率を計算するとすれば，$B[60, 0.05]$，すなわち $n = 60, p = 0.05$ であるような二項分布を使ってもよい．

しかし実際上は，${}_{60}C_x$ といった因子を計算することは困難であるから，数学的に扱いやすい形でこの問題を扱いたい．そこで登場するのがポアソン分布なのである．

4.3.2 ポアソン分布の導出

上に述べたように，二項分布に対して $n \to \infty, p \to 0$, という極限操作を行う．すると，

$$f(x) = \lim_{n\to\infty, p\to 0} {}_nC_x p^x (1-p)^{n-x} = \frac{\mu^x}{x!} e^{-\mu} \tag{4.5}$$

となる．ただし $\mu = np$ としてある．これがポアソン分布である．この式の導出の詳細は付録に掲載した．ここに見るように，ポアソン分布は平均値 μ だけで関数の形が決まるので，$P[\mu]$ と表記されることもある．

上にも書いたように，ポアソン分布を使う理由は，大きな n について二項分布を計算することが困難であることによると考えてよい．実際，式 (4.5) の計算であれば，関数電卓でも求められるものである[*4]．

4.3.3 ポアソン分布の期待値と分散

▼期待値 ポアソン分布の期待値，つまり平均は μ そのものである．

▼分散 分散は，二項分布の場合の分散を表す式 (4.3) から，次のように簡単に導くことができる．

[*4] Windows に付いている電卓は関数電卓に切り替えて使えるので，この種の計算をこなすことができる．

$$\sigma^2 = \lim_{p\to 0} np(1-p) = np = \mu \tag{4.6}$$

つまり，**ポアソン分布では期待値と分散はいずれも μ に等しい**という面白い性質がある．

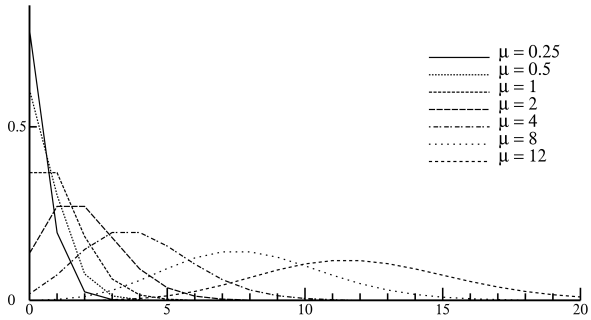


図 4.3　さまざまな μ の値に対するポアソン分布の形：ポアソン分布は離散分布なので確率は x が整数のときのみ値をもつ．しかし，その通りに描くと点だけのグラフになって見にくいので，点をつないでわかりやすくしている．

例題 4–2　60 分に平均して 2 回の電話が掛かってくる窓口がある．60 分の間に，この窓口に 5 本以上の電話が掛かってくる確率を求めよ．

ポアソン分布

$$f(x) = \frac{\mu^x}{x!}e^{-\mu}$$

で，$\mu = 2$ として考える．5 本以上の電話が来る事象は 4 本以内の電話が来るという事象の排反事象であるから，後者の確率すなわち $f(0)$ から $f(4)$ までの和を求めて，それを 1 から引けばよい．従って，

$$\begin{aligned} &1 - (f(0) + f(1) + f(2) + f(3) + f(4)) \\ &= 1 - e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}\right) \\ &= 0.053 \end{aligned}$$

となり，ほぼ 5% の確率となる．

4.3.4　ポアソン分布が適用される問題

上に述べたように，ポアソン分布は二項分布の n が大きく，かつ p が小さくて，平均値 np が有限であるようなケースに適用される．通常はおよそ $n > 50$ であることと，$n \approx 50$ に対して $np \leq 5$ 程度であれば適用可能であるとされている．

ここでは，ポアソン分布のもうひとつの特徴について言及しておこう．ポアソン分布の式，

$$f(x) = \frac{\mu^x}{x!}e^{-\mu}$$

を見ると，ここには n も p も姿を見せておらず，平均値 μ だけしかパラメータとして使われていない．ということは，ポアソン分布においては，多数回の試行や個々の事象は隠されてしまっており，ある事象の発生する平均的な頻度だけが表に出ているのである．このことはこの分布の適用される現象について，広範な一般性を与えることになり，有用な手法を提供することを意味する．

たとえば，夜空で流星が見られる回数について考えてみよう．流星群などが来ている特別な状況を除いて，長期間の観測を行えば，ある平均的な頻度が得られることになる．仮にそれを 1 時間当たり 5 回としよう．

私たちは流星が落ちる原因については何も知らないが，しかし地球の周りの宇宙空間でさまざまな偶然が重なった結果として，たまたま宇宙の塵が大気に突入することになったのだろうという程度の，無難でさして意味のない推定ぐらいはできるであろう．流星を産み出すことに関わる現象は，天空のどこかで時々刻々起きているであろうし，その中で，非常に「運の悪い」塵が叩き落されるとすると，これはポアソン分布を使ってよいケースになる．

流星の発生する仕組みについては全く無知である以上，その仕組みと関係して決まるはずの n や p に関する知識を得ることはできない．せいぜい，n は「時々刻々」の現象ということで非常に大きな数だろうとか，「運が悪いのが落ちる」のだろうから p は小さいだろうということを仮定してもかまわないだろうという程度の知識しかないのである．しかし，観測によって平均値 $\mu = np$ は確実に分かるから，それさえあればポアソン分布の式を利用することが可能になるというところに，この問題の面白さがある．

実際，$\mu = 5$ という値を使えば，たとえば 1 時間に 10 個以上の流星が観測できる確率を知ることができ，式から計算してみると，その確率は 0.03 である．そこでもし，あるときに 1 時間に 10 個の流星が観測されたとすると，それはきわめて稀な現象が起きているというよりは，週に 1,2 度くらいはそんなこともあるだろうという判断ができることになる．

このように，ポアソン分布は，現象の原因やメカニズムがわからなくても，さまざまの自然現象や社会現象の解析において，観測されたデータから確率的な推論を下すための，有力な武器を提供している．

【章末問題】

問題 4–1　ある大学では学生の 2/3 が関西出身である．この大学でランダムに 8 人を選んだとき，関西出身の学生が 4 人未満である確率を求めなさい．

問題 4–2　6 人がじゃんけんして，ぐう，ちょき，ぱあがそれぞれ 2 人ずつになる確率はどれだけか．

問題 4–3　ある中学校のクラスでは，1 週間に平均して 2.5 名の病気欠席者がいる．ある週とその次の週が連続して無欠席になる確率を求めなさい．

第 5 章

正規分布

統計学の中心となる概念は正規分布である．正規分布はいたるところに現れる確率分布で，多くの場合にその性質を使った分析は役に立つ情報を与えてくれる．

5.1 離散的確率分布から連続的確率分布へ

5.1.1 実数データの取り扱い

これまでの確率分布はすべて離散的な確率変数をもつものとして取り扱ってきた．しかし，私たちが現実に取り扱うデータの多くは整数ではなく，小数点を含むものである．たとえば体重の分布を考えてみると，1 章で扱ったように，ふつうは 0.1 kg 刻みのデータとして収録されているであろう．もっと考えると，そもそもある人の体重は，

58.32417081... kg

のように，現実に測定はできないものの，限りなく細かいところまで決まっているはずである[*1]．**多くの数値データは実数として存在している**のである．

すなわち，これまでは離散的な確率変数を扱ってきたのだが，私たちが扱いたい数値データの多くは実数であるので，離散的な変数には乗らないのである．このような**連続的な確率分布** を本章では取り扱う．中でも重要なのは正規分布である．

5.1.2 離散的確率関数の形

離散型一様分布というのは，3.2 節で紹介したように，サイコロの目式の確率分布である．すなわち 1 から n までのいずれかの目が等しい確率 $1/n$ で出現するような確率分布のことである．ちなみにサイコロの目についての確率分布をグラフで表すと図 5.1 のようになるだろう．

[*1] 実際には，こんな細かい値は息をするだけで変動しているであろうが，それでもある一瞬一瞬では相当に細かいところまで確定しているに違いない．

この図の意味するところは，確率関数 $P(X)$ は，$X = 1, 2, ..., 6$ の点でだけ $1/6$ という値をとり，それ以外では至るところゼロであるような，櫛の歯のような関数であるということである．サイコロの目というのは，1 から 6 の目を取るだけで，その間の中間的な値などは取りようもないのだから，これは当然である．このように**櫛の歯型の関数になるというのは，離散型確率関数の特徴である．**

5.1.3　連続領域での確率の特徴

離散型一様分布に対して**連続型一様分布**というのは，ある区間に落ちてくる雨粒の分布のようなものである．今，地面に 2 m の長さの線を引き，落ちてくる雨粒の位置が，その上のどこに来るかということを考えてみよう．雨は広い区域でまんべんなく降るのであるから，この区間の中のどこでも，雨粒が落ちてくる確率が等しいだろうということは，容易に想像できる．このようすを図にするならば，下の図 5.2 のようになると思われるであろう．しかし，このとき，グラフの縦軸の値はどうすればいいのだろうか？

離散的な確率関数の場合，図 5.1 にあるように，縦軸の値が意味するのは，確率変数 X がある値をとるときの確率そのものであった．ところが，図 5.2 のように $[0, 2]$ の連続した区間[*2]にあるすべての実数について一定の確率 p が与えられているとすると，実数というのは有限の区間の中に無限に存在するのだから，それら無数に多くの点についてその確率が適用されることになり，全部の確率を加えた値はかならず無限大になってしまう[*3]．

5.1.4　連続分布の確率は面積で表現する

上述の問題を解決するには，連続的な確率変数の場合の確率の値を，グラフ上の面積で表すということにするとよい．図 5.3 は，長さ 2 m の区間に一様に雨が降ってくる確率を表す関数のグラフである．具体的に式で書くと次のようになる．

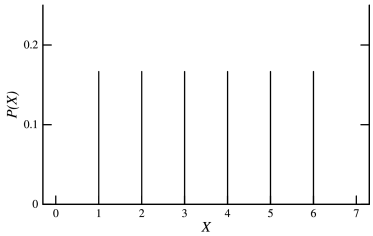


図 5.1　サイコロの目の出方の確率分布関数

[*2] たとえば 1 と 2 と，その間のすべての実数を含む区間のことを，$[1, 2]$ というふうに表す．

[*3] かといって，ある一点にちょうど雨粒が来る確率は限りなく小さいはずだからという理由で，p をゼロとしてしまうと，いくら加えてもゼロのままになって始末におえなくなる．

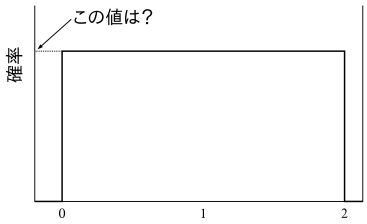


図 5.2　雨粒がある区間に落ちる確率は一様である．その確率の値はどう決めればよいのだろうか

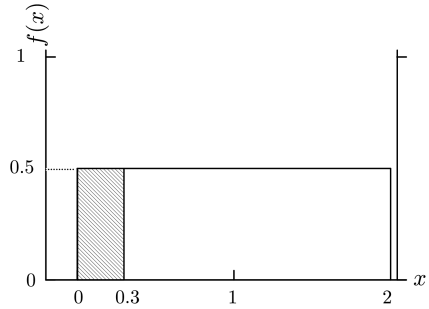


図 5.3　2 m の区間で定義された連続型一様分布の確率密度関数

$$\begin{aligned} f(x) &= 0.5 \quad (0 \leq x \leq 2) \\ f(x) &= 0 \quad (\text{それ以外}) \end{aligned} \tag{5.1}$$

ここで注目しておかねばならないのは，グラフの縦軸の値が 0.5 になっていて，全面積を 1 にするように関数の値が決められているということである．

こうしておけば，この関数の定義域である $[0, 2]$ の区間のグラフの面積はちょうど 1 になる．確率は 1 のときに全事象を覆うのであるから，**面積を確率とみなすことで，連続な分布のときの確率をうまく表現できる**ことになる．したがって，たとえば図の左側の $[0, 0.3]$ の区間に雨が落ちる確率は，斜線で示された部分の面積，$0.3 \times 0.5 = 0.15$ というふうに表現される．

このことを数学的に表しておこう．区間 $[a, b]$ の中で，ある関数 $f(x)$ が描く面積は定積分で表されるから，x がその区間に収まる確率 $P(a \leq X \leq b)$ は，次のように書ける．

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx \tag{5.2}$$

また，起き得るあらゆる場合について確率を足し合わせると 1 になることから，

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{5.3}$$

となる[*4] .

この例に即して書けば，$f(x) = 0.5,\ a = 0,\ b = 0.3$ であるから，

$$P(0 \leq X \leq 0.3) = \int_0^{0.3} 0.5\, dx = [0.5x]_0^{0.3} = 0.15 \tag{5.4}$$

となる．

このように，連続型の確率変数 x に対して，確率の大きさを表す関数 $f(x)$ を**確率密度関数 (probability density function)** という．

また，$f(x)$ の**累積分布関数** $\Phi(z)$ は次のように与えられる関数である．

$$\Phi(z) = \int_{-\infty}^{z} f(x)\ dx \tag{5.5}$$

これは，図 5.4 を見れば分かるように，$f(x)$ のあるところまでの面積で表される値であり，確率の定義から，

$$\Phi(-\infty) = 0 \tag{5.6}$$

$$\Phi(\infty) = 1 \tag{5.7}$$

となっていなければならない．

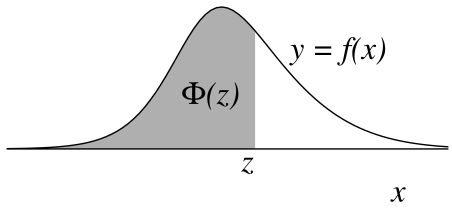


図 5.4　累積分布関数 $\Phi(z)$ は，ある範囲の事象を覆う確率密度関数 $f(x)$ の面積で定義され，その範囲の事象が起きる確率を表現する．

図 5.3 ，式 5.1 で示される確率密度関数については，その累積分布関数は

$$\Phi(z) = \int_{-\infty}^{z} 0.5\ dx = 0.5z, \quad (0 \leq z \leq 2) \tag{5.8}$$

と表せる．

例題 5–1　式 5.1 で示される確率密度関数を使って，一滴の雨粒が端から測って 0.8 m の点から 1.5 m の点の間に落ちる確率を求めよ．

[*4] ここで積分範囲を $[-\infty, \infty]$ にとったのは，X のすべての範囲について積分するという理由からである．しかし，実際には確率が定義されている範囲にわたって積分を計算すればよい．

$f(x) = 1/2$ を単に $x = 0.8$ から $x = 1.5$ まで積分するだけである．従って，

$$\int_{0.8}^{1.5} \frac{1}{2}\, dx = \left[\frac{x}{2}\right]_{0.8}^{1.5} = 0.35$$

もちろん図 5.3 を見て，相当する範囲の面積を求めるだけでもよい．

5.1.5 連続的確率関数の平均，分散

ある確率密度関数が与えられたとき，その平均や分散がどうなるかを導いておこう．その前に，離散型の場合の式 (3.6) に相当する関係式として，

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{5.9}$$

という条件が満たされている必要がある．この条件は，全事象に対する確率の和が 1 であることを意味しており，**規格化条件 (normalizing condition)** という．

さて, 期待値 (平均) を表すための式 (3.10) に相当するのは，次の式である．

$$E[X] = \mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{5.10}$$

さらに，分散を表す式 (3.11) に相当するのは次の式である．

$$V[X] = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \tag{5.11}$$

これらは単に 3.4 節の関係式における総和を積分に変えただけのものである．

式 (3.12) はここでも成立している．すなわち，

$$V[X] = E[X^2] - E[X]^2 \tag{5.12}$$

ただし，

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx \tag{5.13}$$

である．

5.1.6 コンピュータで発生させる一様乱数

コンピュータを使うと区間 $[0, 1]$ で一様な確率密度をもつ乱数を発生させることが簡単にできる．たとえば，下は試しに 10 個の乱数を発生させてみたものである．

0.2747, 0.2288, 0.6893, 0.1855, 0.9086, 0.1876, 0.9291, 0.5324, 0.3335, 0.8568

この乱数は式で表すと次の一様連続分布に従っている．

$$\begin{aligned} f(x) &= 1, \quad (0 \le x \le 1) \\ f(x) &= 0, \quad (\text{それ以外}) \end{aligned} \tag{5.14}$$

この分布の期待値が 1/2 であることはほとんど自明だが，式 (5.10) を適用してみると，

$$\mu = \int_0^1 x \times 1 \; dx = \frac{1}{2} \tag{5.15}$$

となって，その通りになっている．

一方，分散は式 (5.12) より，

$$\sigma^2 = \int_0^1 x^2 \times 1 \; dx - \mu^2 = \frac{1}{12} \tag{5.16}$$

となる．このことについては章末の問題で触れることにする．

コンピュータを使うと，他にも，ある整数よりも小さい正の整数をランダムに発生させたり，一様でない分布に従う乱数を発生させたりすることが容易にできるようになっている．

5.2　二項分布から正規分布へ

二項分布 $B[n,p]$ で，n が非常に大きくなるとどのような分布を描くかを調べてみよう．図 5.5 は，$p=0.4$ として，n を変えてみたときに，二項分布の形がどのようになるかを調べたものである．

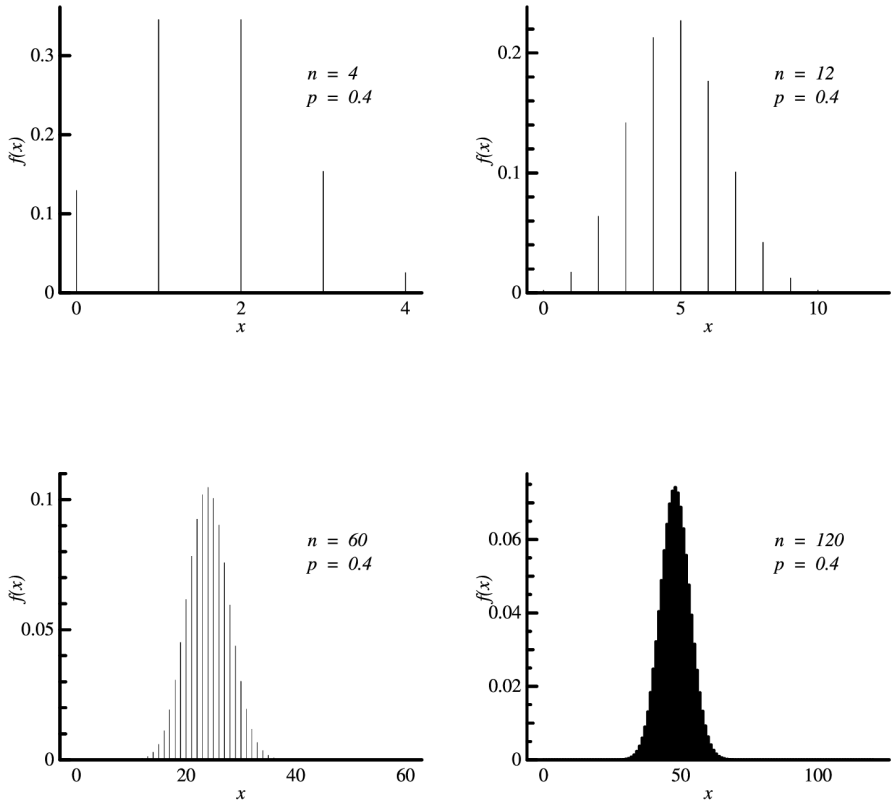


図 5.5　二項分布 $B[n,p]$ の n を増加させていったときの分布の形の変化

このグラフを見ると，二項分布の n が大きくなるに従って，平均値の周りに左右対称な吊鐘形をした分布になっていく．この分布は**正規分布 (normal distribution)** と呼び，確率統計において中心的な役割を果たす確率分布である．

正規分布は，次のような式で表される

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{5.17}$$

ただし，$\exp(a)$ は e^a と同じである[*5]．

ここで μ, σ^2 は，それぞれ正規分布の平均と分散である[*6]．

4.1.2 で見たように，二項分布においては $\mu = np$, $\sigma^2 = np(1-p)$ であるから，上述の二項分布で n を大きくしていったときの正規分布への移行は，

$$B[n,p] = {}_nC_x p^x (1-p)^{n-x} \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{5.18}$$

を意味する．

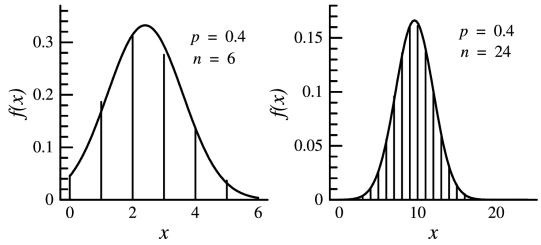


図 5.6　同じ σ と μ をもつ二項分布 (縦の線) と正規分布（実線）を $n = 6, 24$ に対して描いたもの．

この式の導出はやや手順が長いので省略するが，実際に二項分布と正規分布のグラフを重ねてみて，両者がどのように近づくかを見てみよう（図 5.6）．この図を見ると $n = 24$ の二項分布は正規分布ときわめてよく一致していることがわかる．実用上は，$np > 5$ かつ $n(1-p) > 5$ であれば，二項分布を正規分布として扱うことは差し支えないとされるので，$p = 0.5$ であれば $n = 10$ 程度でもよい近似が得られると考えてよい．

正規分布は，後にみるように非常に多くのケースに使われるもっとも重要な確率分布であり，しばしば $N[\mu, \sigma^2]$ と表される．すなわち，

平均 $\boldsymbol{\mu}$, 分散 $\boldsymbol{\sigma^2}$ をもつ正規分布を $\boldsymbol{N[\mu, \sigma^2]}$ と表す．

[*5] exp を使ったほうが，式がかさばらないので，しばしば使われる．

[*6] 正規分布の平均が μ であることは，式 (5.17) の形からすぐに分かる．すなわち，この関数は $x = \mu$ のときに最大値をとり，かつそこを中心に左右対称になっているわけであるから，平均は $x = \mu$ のところになっていなければならない．

5.3　正規分布表の活用

5.3.1　標準正規分布と標準化変換

平均が μ, 分散が σ^2 であるような確率変数 x が正規分布に従うものとしよう．このとき，式 (5.19) で定義される変換をほどこしてみる．

$$z = \frac{x-\mu}{\sigma} \tag{5.19}$$

すると z は式 (5.20) の分布に従う．

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{5.20}$$

これは平均がゼロで標準偏差（分散）が 1 であるような正規分布であり，**標準正規分布**と呼ばれる．また，式 (5.19) で定義される変換を**標準化変換 (standardization)** という．

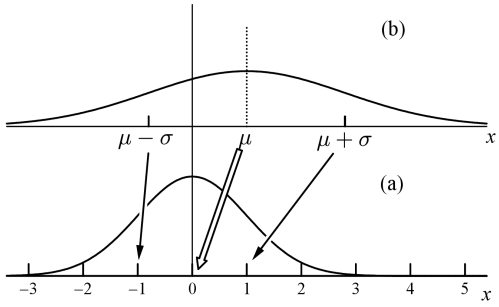


図 5.7　一般の正規分布 $N[\mu, \sigma^2]$ を $z = (x-\mu)/\sigma$ で標準正規分布 $N[0,1]$ に変換する．

図 5.7 に標準化変換の意味を示した．一目でわかるように，一般の正規分布 $N[\mu, \sigma^2]$ に従う確率変数 x が与えられたとすると，それを標準化変換して得られる z は正規分布 $N[0,1]$ に従う．

また，標準正規分布から一般の正規分布への逆の変換も考えられる．これは式 (5.19) からただちに導かれる．

$$x = \sigma z + \mu \tag{5.21}$$

一般の正規分布に従う分布を標準正規分布に変換することで，次に出てくる正規分布表を使ってさまざまな計算を進めることが可能になる．

表 5.1　正規分布表からの抜粋

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.00	0.500000	1.00	0.841345	2.00	0.977250	3.00	0.998650
0.10	0.539828	1.10	0.864334	2.10	0.982136	3.10	0.999032
0.20	0.579260	1.20	0.884930	2.20	0.986097	3.20	0.999313
0.30	0.617911	1.30	0.903200	2.30	0.989276	3.30	0.999517
0.40	0.655422	1.40	0.919243	2.40	0.991802	3.40	0.999663
0.50	0.691462	1.50	0.933193	2.50	0.993790	3.50	0.999767
0.60	0.725747	1.60	0.945201	2.60	0.995339	3.60	0.999841
0.70	0.758036	1.70	0.955435	2.70	0.996533	3.70	0.999892
0.80	0.788145	1.80	0.964070	2.80	0.997445	3.80	0.999928
0.90	0.815940	1.90	0.971283	2.90	0.998134	3.90	0.999952

5.3.2　正規分布表とその意味

正規分布を利用するさいには，**正規分布表**と呼ばれる数表を用いる．この表には，z に対して次の定積分の値 $\Phi(z)$ を計算したものを掲載してある．表 5.1 に抜粋した表を示す．

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \tag{5.22}$$

この積分は，図 5.8 の影をつけて示されている領域の面積に相当するので，標準正規分布に従う確率変数が $x = z$ よりも左側の部分に入る確率を意味することになる．

なお，ここで注意しておいてほしいのは，世の中の正規分布表には積分区間を $[-\infty, z]$ ではなく，$[0, z]$ としたものもあるということである．その場合には $z = 0$ に対して $\Phi(z) = 0$ となり，表の値は 0.5 だけずれることになる．ただし図を見て考えれば，このことは大して面倒なことではない．

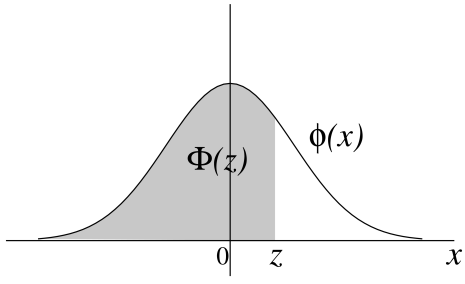


図 5.8　正規分布 $\phi(x) = 1/\sqrt{2\pi}\, \exp(-x^2/2)$ と累積分布関数 $\Phi(z)$ との関係

5.3.3 正規分布の計算の手順

標準正規分布関数 $f(z)$ について，常に次の 2 点を押さえておこう．

- $f(z)$ を $-\infty$ から ∞ まで積分して足し合わせた全面積は 1 である．
- $f(z)$ は左右対称である．

■標準正規分布における確率を求める

標準正規分布 $N[0,1]$ において，$-1 \leq z \leq 1$ の範囲に含まれる面積はどれだけか．

$z < 1$ の部分の面積は，表から 0.841345 である．したがって $z > 1$ に含まれる面積は $1 - 0.841345 = 0.158655$ であり，図 5.8 を見て考えれば，$z < -1$ に含まれる面積もそれに等しいことが分かる[*7]．よって答えは，$1 - 0.158655 \times 2 = 0.68269$

■標準化変換を使う

標準正規分布に従うような事象はあまりなく，一般的な $N[\mu, \sigma^2]$ に従う分布がほとんどであるから，よくある解き方の手順は次のような感じになる．

1. $x \to z$ の標準化変換を行う
2. 正規分布表から z よりも小さい部分の面積を読み取る
3. その値と，上記の 2 つの点を使って，求めるべき部分の面積を計算して確率とする．

例題 5–2 受験者の平均点が 65 点，標準偏差が 12.5 であるような試験があったとする．得点分布が正規分布しているとした場合，この試験で 90 点以上を取る人の割合は，全体のうち 何 % と見ればよいか．

式 (5.19) の変換によって，この成績での 90 点を標準化すると，$z = (90-65)/12.5 = 2.0$ になる．そこで $\Phi(2.0)$ を表から調べると，0.97725 であるから，100 人のうち 97.7 人が 90 点未満のところにいるとしてよい．従って答えは 2.3 % となる．

■半整数補正—二項分布を正規分布で近似する

5.2 節で扱ったように，二項分布は n がそこそこ大きければ正規分布に近似できる．そのため，面倒な組み合わせの計算をすることなく，数表と簡単な計算だけで二項分布に従う事象の確率を計算でき，非常に強力な確率計算の武器になる．

[*7] 図を見てちょっと考えれば，$2\Phi(z) - 1$ が求める答えであることが分かる．それを使うと計算はもっと簡単である．

ただし，離散的な確率分布を連続分布に置き換えて計算する都合上，**半整数補正**と呼ばれる近似手法がよく用いられる．次の例題でそのことを説明しよう．

例題 5–3　日本人で A 型の血液をもつ人は 40% いる．10 人の日本人を集めたときに，A 型の血液を持つ人が 4 人から 6 人の間になる確率を求めて，二項分布による計算の結果と比較せよ．

二項分布では $\mu = np$, $\sigma = \sqrt{np(1-p)}$ である．したがってこの集団については，$\mu = 4$, $\sigma = \sqrt{10 \cdot 0.4 \cdot 0.6} = 1.55$ となる．一方**この二項分布を，連続分布である正規分布とみなすと，4 人から 6 人の間であるということは 3.5 人と 6.5 人の間にあるとみなせる**から，$z_1 = (3.5-4)/1.55 = -0.32$, $z_2 = (6.5-4)/1.55 = 1.61$ の間に入る確率 $\Phi(1.61) - \Phi(-0.32)$ を計算すればよい．図 5.9 にそのことを示した．ここで負の z に対する値は正規分布表にないが，正規分布の形を考えれば $\Phi(-0.32) = 1 - \Phi(0.32) = 0.374$ であり，$\Phi(1.61) = 0.946$ なので，答は 0.572 となる．

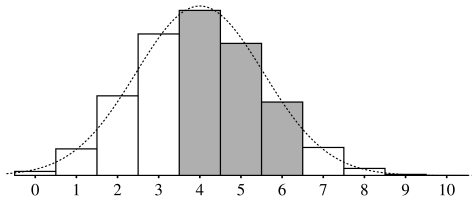


図 5.9　半整数補正の使い方．二項分布 $B[10, 0.4]$ で $x = 4, 5, 6$ に相当する積分範囲は 3.5〜6.5 になる．

▼半整数補正を使う条件　二項分布 $B[n, p]$ が与えられていたとき，$[x_1, x_2]$ という区間の面積を求めたい．ここで x_1, x_2 は整数である．このとき半整数補正を使うかどうかを判断する基準を考えておこう．

段取りを見直すと，まず次のように x_1, x_2 を標準化する．

$$\begin{aligned} z_1 &= \frac{x_1 - \mu}{\sigma} \\ z_2 &= \frac{x_2 - \mu}{\sigma} \end{aligned} \tag{5.23}$$

その後，正規分布表から $\Phi(z_1)$ と $\Phi(z_2)$ を拾って差を求めるわけだが，このままだと無視できない誤差が入ることがある．図 5.9 にあるように，両端の x_1, x_2 が現れる確率はヒストグラムの区間幅が 1 の柱の面積なのだが，区間 $[z_1, z_2]$ の面積を求めたとすると，両端の柱が半分ずつしか含まれなくなる．これが誤差の原因だ．

そこで誤差を補正するために面積を求める区間を修正する．つまり，

$$\begin{aligned} z_1 &= \frac{x_1 - \frac{1}{2} - \mu}{\sigma} \\ z_2 &= \frac{x_2 + \frac{1}{2} - \mu}{\sigma} \end{aligned} \tag{5.24}$$

としてやればよい近似になる．これが半整数補正の内容である．

それでは，半整数補正をやらなくてもよい近似が成立するというのは，どういう条件の下でだろうか．それは式 (5.24) において 1/2 という補正によって z_1, z_2 が影響をほとんど受けないという条件だ．ということは，σ **が** 1/2 **よりもずっと大きい場合**ということになる．

たとえば $n = 400$ で $p = 0.5$ の場合，$\sigma = \sqrt{100} = 10$ である．このときの 1/2 の補正の寄与は 0.5% しかないので，正規分布表を引いたときには意味がなくなる．

【章末問題】

問題 5–1 第一章で扱った 100 人の男子高校生の体重の平均と標準偏差から，平均値の前後の何 kg の範囲をとれば，半分の人数がそこに含まれることになるか計算して予測せよ．さらにその結果を実際のデータと比較してみよ．

問題 5–2 入試や模擬試験の個人成績を表すのによく使われる**偏差値 (standard score)** は，次のようにして算出される．全体の平均点を μ, 標準偏差を σ とした場合，点数が μ に等しいものを偏差値 50 とする．そして得点がそれより $\sigma, 2\sigma, \cdots$ だけ上回る点数を偏差値 $60, 70, \cdots$ とし，下回るほうについても同様に決める．

今，参加者が 12000 人の模擬試験で偏差値 58 を得た受験生がいたとする．この人の成績は全体で何番ぐらいに位置するかを求めなさい．

問題 5–3 日本人で A 型の血液をもつ人は 40% いる．24 人の日本人を集めたときに，A 型の血液を持つ人が 9 人以上 12 人以下である確率を求めよ．

問題 5–4 生まれる赤ちゃんが女の子である確率は 1/2 であるとする．ある年に 4 万人の赤ちゃんが生まれたとして，その男女比が平均からずれて，一方の性の比率が全体の 49% 以下または 51% 以上になる確率を求めなさい．

問題 5–5

1. 内閣支持率を無作為抽出によって調査したい． 真の支持率が 45.0 % であったとして，調査人数を 400 人とした場合に，調査結果として 得られる支持率が 45.0±2.5% の範囲になる確率を求めよ．

2. 上と状態における調査を，調査人数を 2000 人に増やして行いたい．このときに調査結果として得られる支持率が 45.0 ± 2.5 % の範囲になる確率を求めよ．

5.4 中心極限定理

前節で，二項分布の近似として正規分布が使えることがわかったが，それだけにとどまらず，もっと一般的な分布に対しても正規分布が成立することがしばしばある．それは次のような**中心極限定理 (central limit theorem)** によって保証される．

> 確率変数 $X_1, X_2, ..., X_n$ が互いに独立で，平均 μ, 分散 σ^2 をもつ分布に従っているとする．このとき，平均
>
> $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.25}$$
>
> をとり，
>
> $$Z = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) \tag{5.26}$$
>
> とすると，$n \to \infty$ の極限で Z は標準正規分布 $N[0,1]$ に従う．あるいは，式 (5.19) と見比べると，$\overline{X}$ は平均が μ, 分散が σ^2/n となるような正規分布 $N[\mu, \sigma^2/n]$ に従っていると言ってもよい．

この定理では，下線部の条件を満たす確率分布であれば，**もとの分布が正規分布でなくても**，そこからとった独立な多数の変数の平均が正規分布に従うことを保障している．これは驚くべき事実であり，正規分布のような「山なり」の分布でなくても，そこから何個かの確率変数をとって平均したものを集めると，正規分布になっているというのである．たとえば「平らな」分布である一様分布からも，そのようにして正規分布を作ることができることを，次の例で示した．

なおこの定理によって Z が正規分布とみなされるための n の値は，もとの分布の性質にもよるが，10 程度でもかなりよい近似を与える．多くの統計データの処理において正規分布を使った解析が威力を発揮するのは，この中心極限定理によって，確率的に振舞うデータから正規分布が実現することが保証されているからである[*8]．

問題 5–6 5.1.6 節 (77 ページ) に出てきた一様乱数は，分散が 1/12 である (式 (5.16))．一方，2 つの独立な確率変数の分散は，それぞれの分散を足し合わせることで求めること

[*8] このことについて考察しておこう．たとえば日本人の成人男子の身長のデータを正規分布しているとして扱っても，かなりよい近似になる．これは人の身長が，多くの遺伝子が独立に，あるいは共同して関与した結果であるとともに，胎内での発生過程に関わる母親の栄養状態や心理状況，生後の生育環境を構成する食物や気候や運動など，無数の要因が関わったものであり，それらの効果の総和として身長が決まることによる．つまり，独立した多数の要因によって決まる量は中心極限定理によって正規分布するのであるから，身長も正規分布する傾向をもつことになるのは自然なことである．もっとも，体重のデータは身長のおよそ 2 乗から 3 乗に比例するために，身長の場合よりも非対称な分布になる傾向が強くなる．

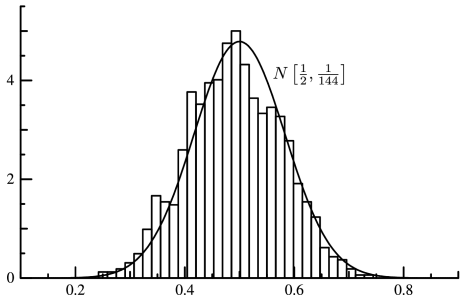


図 5.10　区間 $[0, 1]$ の連続一様分布から 12 個の確率変数を取って作った平均がどのように分布するかを調べたヒストグラムと，中心極限定理で予想される正規分布の曲線.

ができる (式 (3.17)).

このことを利用すると，区間 $[0, 1]$ の一様連続分布に従う乱数を 12 個足し合わせて，その和から 6 を引いた値はかなり標準正規分布に近い分布になる．そのことを説明しなさい.

第 6 章

無作為抽出と標本分布

ここからが**推測統計**の始まりである．前章まで問題にしてきた確率も，多数回の試行によって意味を持つものであるが，基本的には個々の事象の期待値に対する理論である．

推測統計の理論は，未知の大きな集合（母集団）の中から取り出した小さな集合（標本）を調べて，親の集合の特性を確率論的に調べるものである．その関係を理解しよう．

6.1　無作為標本抽出

図 6.1 に**無作為標本抽出 (random sampling)** の操作にかかわる概念と諸量の関係を示した．ここで，**母集団 (population)** は未知の**母平均 (population mean)** と**母分散 (population variance)** をもつ「知りたい集合」である．それをより小さな**標本 (sample)** から知ろうというのが，ここからの目標だ．

6.1.1　母集団と標本

母集団 (population) は調査分析したいすべてのデータを含む集合である． 母集団に

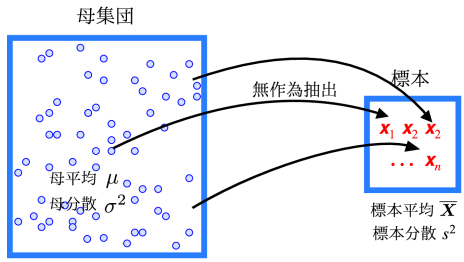


図 6.1　無作為抽出の操作の概念．古典的な統計学は，未知かつ知りたい量である μ, σ^2 と，手のひらの上にある $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}$ から計算できる $\overline{X}, s^2$ を結びつけるところから始まる．

含まれる要素の数は有限であっても無限[*1]であってもよい．

母集団の要素すべてについてデータを集めたとして，その平均を**母平均**, 分散を**母分散**という．これ以降は**母平均を μ, 母分散を σ^2** で表すことにする．また，これらの母集団の特性を表すパラメータを**母数 (population parameters)** と呼ぶ．

母集団のすべてのデータについて集計する，**全数調査**あるいは**悉皆 (しっかい) 調査**は多くの場合困難である[*2]．また，調査や分析によって標本そのものが破壊されたり変化を蒙ってしまうような場合には，全数調査を行うことは本来の目的を果たせなくなってしまうことになる[*3]．したがって，母集団から有限個の**標本 (sample)** を取り出して，定量的な調査，分析を行うことが多い．取り出される標本の数のことを標本の**大きさ (size)** あるいは**サイズ**とそのまま いう[*4]．以後，「大きい標本」，「小さい標本」という表現をしばしば使うが，要するに標本として取り出したデータの数が大きいか小さいかを表していると思ってほしい．

6.1.2　乱数による無作為抽出

母集団から一部の標本を取り出すことを，**標本抽出 (sampling)**，あるいは**抜き取り**という．英語をそのまま使って**サンプリング**ともよくいう．標本抽出に当たっては，母集団の特性が標本の集合にもなるべく正確に反映されるようにしなければならない．そこで行われるのが**無作為標本抽出 (random sampling)** である．すなわち，母集団のどの要素についても，同じ確率で標本として抜き取られるように，作為を交えずにそれぞれを選ぶのである．

無作為抽出には，**離散型一様分布**に従う**乱数**が使われる．現代ではコンピュータを使って乱数を発生させることがほとんどである[*5]．

表 6.1 に乱数表の例を示した．乱数表では，どのように数を拾っていっても特定の傾向が現れることはないようになっている．たとえば，この表から一桁の乱数の系列を得るの

[*1] ここで無限といっているのは，二つの場合がある．ひとつはきわめて多数であって実際上無限と近似してかまわないもの，もうひとつはいくらでも繰り返される事象，たとえばコインの裏表の出方を調べるとか，工場のラインから製品が毎日生産されるといった場合も，無限とみなせる．

[*2] 国勢調査は最も大規模な全数調査である．この場合には，もっとも完全な形で母集団の統計情報が得られるが，それに要する費用と時間は莫大なものである．それでも真にすべての要素について調べつくすことは，現実には不可能である．

[*3] たとえば電球の寿命の検査をある工場の製品すべてについて行ったとしたら，切れた電球しか出荷できなくなってしまうであろう．

[*4] しばしば，標本の大きさのことを「標本（の）数」と呼んでいるケースをみるが，厳密に言うとこれは誤りである．

[*5] 「離散型一様乱数」は，どの数も等しい確率で出現し，次に何が現れるかを完全に予測できない，完璧なサイコロのような乱数だが，コンピュータの有限手数のアルゴリズムでは原理的に作れない．実際に作られるのは**擬似乱数**という，「かなり乱数ぽい数列」である．とはいえ，現在のコンピュータで利用されている擬似乱数は，実用上は正しい乱数と考えてもかまわない．

表 6.1 乱数表の例

98474	**71279**	**63082**	**78829**	**42648**	**14443**	**69985**	**58505**	**73760**	**96835**
37252	**88586**	**62283**	**71713**	**61004**	**62979**	**29684**	**15151**	**41589**	**44958**
43215	**04177**	**61654**	**95413**	**43685**	**95877**	**61315**	**09869**	**46923**	**85614**
76004	**67425**	**09426**	**72476**	**52651**	**44729**	**98959**	**10064**	**09796**	**98117**
60610	**70770**	**57281**	**67053**	**19024**	**01629**	**41143**	**01965**	**07339**	**99938**
29309	**69622**	**63555**	**86700**	**03750**	**39202**	**84902**	**06042**	**74703**	**02108**
80801	**28750**	**82589**	**28729**	**15136**	**88027**	**03250**	**15225**	**78384**	**25588**
22125	**23483**	**80242**	**76254**	**93014**	**67361**	**03408**	**69128**	**47009**	**48339**
09106	**73507**	**67285**	**93722**	**35009**	**67651**	**95285**	**00497**	**76141**	**58511**
84030	**37979**	**89450**	**30578**	**64083**	**12380**	**12603**	**51943**	**37857**	**46401**

に，5 桁の数を下へと追いながら，9,8,4,7,4,3,7,2,5,2,4,3,2,1,5, .. としてもよいし，あるいは 5 列目の縦の系列の 5 桁目だけを拾って，4,6,4,5,1,0,1,9,3,6,... などとしてもよい[*6]．

抽出には，母集団から同じ標本を繰り返して抽出することを許す**復元抽出**と，同じ標本は重複させないでとる**非復元抽出**の二通りのやり方がある．

復元抽出の例：

赤と白の球の入った袋から 1 個取り出しては，また元の袋に戻して次の球を取り出す

アンケート対象者を乱数で決めたとき，たまたま同じ人が 2 回選ばれても，その人に 2 回尋ねることにする

サイコロで当選者を決めていく．二度選ばれても構わない．

非復元抽出の例：

赤と白の球の入った袋から 1 個取り出したら，取り出した球を戻さないで次の球を取り出す

アンケート対象者を乱数で決めたとき，たまたま同じ人が 2 回選ばれたら，重複しないように別の人を乱数で選ぶ

サイコロで当選者を決めていく．二度選ばれたら外す．

例題 6–1 表 6.2 に掲げた高校生の体重の一覧表から，乱数を使って 10 人の標本を抽出せよ．

母集団は 100 個のデータからなるので，それらに 0 – 99 の番号をつけることができる．

[*6] 乱数表を使う場合には，「愚直に」数を拾わなければならない．たとえばある特定の数字が連続したりすると，それを排除してもっともらしくしようとする心理が働くものであるが，そのことで必ず何らかの傾向が生じてしまうことになる．ある数字が繰り返されることに対する心理的抵抗は大きいものだが，ちょっと確率の計算をすれば，同じ数字が連続して現れる頻度は，かなり高いことが分かるもので，それを避けることは意味がないのである．ただし，意図的に非復元抽出を行うような場合には，重複を排除するようにして抽出を行なう．

表 6.2　100 人の男子高校生の体重/kg(再掲)

43.6	45.2	45.4	45.8	47.2	47.8	48.2	48.7	48.8	48.9
49.0	49.0	49.4	49.5	49.8	50.4	50.5	50.9	50.9	51.2
51.2	51.2	51.3	51.3	51.6	51.7	51.7	51.8	52.0	52.0
52.1	52.1	52.1	52.2	52.3	52.7	52.7	52.8	52.9	52.9
53.1	53.1	53.8	54.0	54.5	54.5	54.6	54.7	54.7	54.7
54.8	54.9	55.1	55.1	55.2	55.3	55.4	55.4	55.4	55.6
55.7	55.8	55.9	56.1	56.3	56.3	56.3	56.4	56.5	56.7
56.8	57.0	57.1	57.1	57.2	57.3	57.6	57.7	57.8	58.1
58.4	58.6	58.7	58.7	58.7	58.7	59.1	59.3	59.9	60.0
60.1	60.3	60.5	60.6	60.6	60.7	61.3	62.7	64.2	64.6

そして乱数として 2 桁のものを用いれば，母集団からランダムに抽出を行うことができる．ここでは表 6.1 の乱数表の 1 行目の数字を取り出していっては，2 桁ずつに区切って得られる数字を用いることにしよう．すると次の数列が得られる．

$$98, 47, 47, 12, 79, 63, 08, 27, 88, 29, 42, 64, 81, 44, ...$$

ところがここで，47 という値は 2 番と 3 番目に重複して出現しているから，その通りにデータを拾うと，47 番の生徒のデータを二度拾うことになる．このように重複を許して抽出するやり方が**復元抽出**である．この場合，

$$98, 47, 47, 12, 79, 63, 08, 27, 88, 29$$

の番号の生徒の体重が次のように抽出される．

$$64.2, 54.7, 54.7, 49.4, 58.1, 56.1, 48.8, 51.8, 59.9, 52.0$$

一方，重複を許さずに抽出するとすれば，2 度目の 47 は飛ばして，

$$98, 47, 12, 79, 63, 08, 27, 88, 29, 42$$

という番号でデータを拾えばよい．すると次のデータが抽出される．

$$64.2, 54.7, 49.4, 58.1, 56.1, 48.8, 51.8, 59.9, 52.0, 53.8$$

このような抽出の仕方が**非復元抽出**である．ただし母集団が十分に大きいときには，同じものが重複して抽出される確率は低い．したがって，復元抽出でも非復元抽出でも結果にほとんど違いはなくなる．そこで，これ以降では特に断りがない限り，標本を抽出するときには復元抽出を行うものとして進めることにする．

6.2 標本平均の分布

6.2.1 抽出された標本は信頼できるか

図 6.1 をもう一度見てほしい．母集団から抽出された n 個の標本のデータを

$$X_1, X_2, \ldots, X_n$$

としよう．X_1, X_2 等は，サイコロの目の数とか乱数で選んだ高校生の体重といった，何らかの予測できない数で，確率変数であり，特にこの場合には**標本確率変数**と呼ばれる．サイコロを振るときを考えれば，復元抽出であれば，$X_1, X_2, ..., X_n$ は互いに独立であることがわかる．

これから，次の 2 つの重要な量, **標本平均**および**標本分散**を求めることができる．

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \text{ (標本平均)} \tag{6.1}$$

$$s^2 = \frac{1}{n}\left(\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2\right) \text{ (標本分散)} \tag{6.2}$$

当然，標本分散の正の平方根 s は**標本標準偏差**と呼ばれる．

標本平均と母平均の関係を考えてみよう．n が小さいと，標本平均は母集団の要素のごく一部の平均になるのであるから，両者にはずれが大きいだろう．逆に n がきわめて大きければ，つまり大きな標本であれば，標本平均は母平均とよい精度で一致するだろう．これは，私たちが日ごろからほとんど無意識にそう感じていることだ．図 6.2 は，そのことをシミュレーションによってざっと調べたものである．

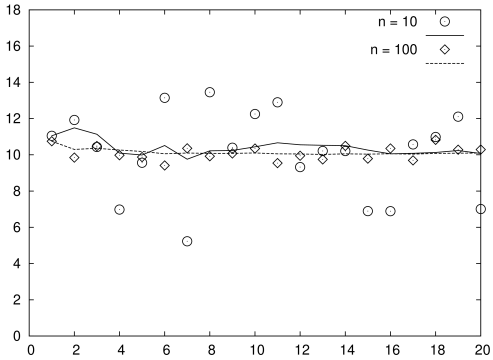


図 6.2 $[0, 20]$ の区間で連続一様分布している母集団から，大きさが 10 と 100 の標本を抽出する操作を 20 回ずつ行った結果．○ と ◇ はそれぞれ $n = 10, 100$ の場合の $\overline{X}$, 実線と破線は $\overline{X}$ のそこまでの平均値．

図から分かるように標本の大きさが大きいと，標本平均 $\overline{X}$ の幅は小さくなり，またその平均値は母平均 $\mu = 10$ に急速に落ち着いていく．

6.2.2　標本平均の分布に関する最も重要な式

上記の定性的な事実を数式できちんとまとめることにする．まず，もしも標本抽出を多数回繰り返していったとすると，その標本平均 $\overline{X}$ の平均は母平均に限りなく一致していくにちがいない．つまり，**標本平均の期待値は母平均に一致する．**

次に，$\overline{X}$ の「広がり」，つまり分散は，上のシミュレーションからも想像されるように，n が大きいほど小さい．また，もしも母分散 σ^2 が小さければ，$\overline{X}$ の広がりも狭く絞り込まれるにちがいない．以上の傾向をまとめたのが次の式である．

$$E[\overline{X}] = \mu \tag{6.3}$$

$$V[\overline{X}] = \frac{\sigma^2}{n} \tag{6.4}$$

これらの詳しい導出は A.6 節 (p.166) に示した．ここで式 (6.4) で表される $V[\overline{X}]$ の平方根 $\frac{\sigma}{\sqrt{n}}$ は**標準誤差 (standard error)** と呼ばれ，抽出された標本の平均の分布の幅を表す重要な指標である．標準誤差はよく SE と略して使われる．

今後の展開において，標本平均と標準誤差はきわめて大きな役割を果たす．

6.3　標本分散の分布

6.3.1　標本分散の平均

標本平均 $\overline{X}$ の分布については上のシンプルな関係式が出てきたが，標本分散 s^2 の分布についてはどうだろうか．

この場合，一般的な関係式は s^2 の期待値についてだけ存在する．この結果も重要で，とてもよく使われる．式の導き方は付録 A.7 に示した (p.166)．

$$E[s^2] = \frac{n-1}{n}\sigma^2 \tag{6.5}$$

例題 6–2　復元無作為抽出によって 6 人の組を何組も選んで体重の平均と分散をとった．その結果，分散の平均値は $(4.32\ \mathrm{kg})^2$ となった．母集団の標準偏差を推定せよ．

上で求められた分散 (s^2) の平均値 (期待値と読み換えてよい) というのは，式 (6.5) の $E[s^2]$ であり，求めたいのは母分散 σ^2 の平方根 σ なのであるから，

$$\sigma^2 = \frac{n}{n-1}E[s^2] = \frac{6}{5} \times 4.32^2$$

$$\sigma = \sqrt{\frac{6}{5}} \times 4.32 = 4.73$$

より，母標準偏差は 4.73 kg となる．

■変数の自由度—標本分散はどうして母分散よりも小さめなのか？

式 (6.5) が定性的に語っている[*7]ことは，「標本分散は，平均として母分散よりも少し小さめの値をとる確率変数である」ということだ．どうして小さめになるのだろうか？そのわけは次のとおりだ．

s^2 の定義は次のようになっている．

$$s^2 = \frac{1}{n}\left(\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2\right) \tag{6.6}$$

これと次の式で表される t^2 を比べてみよう．

$$t^2 = \frac{1}{n}\left((X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2\right) \tag{6.7}$$

t^2 は標本のデータ $X_1, X_2, \cdots$ のそれぞれから μ を引いたものの 2 乗の平均であり，その期待値は σ^2 になる[*8]．一方，上の s^2 の方では，μ ではなく $\overline{X}$ を引いている．

仮に取り出されるデータが，たまたま μ に比べて大きい方に偏っていたとしよう．すると，$\overline{X}$ はそれらの平均なのだから，引きずられて大きい方にずれることになる．結果，s^2 の定義の中の $(X_1 - \overline{X})^2$ 等は $(X_1 - \mu)^2$ 等に比べて小さくなる．そのため， s^2 は σ^2 よりも小さい値を取る．また，標本のサイズ n が小さいほど，そのような偏りが生じる確率が大きくなるので，この傾向は著しくなり，後の t-分布のところで大事な意味を持つことになる．

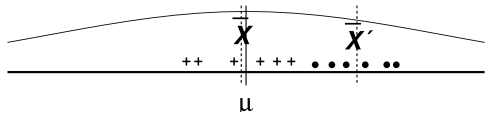


図 6.3　大きさ 6 の標本で，データが母平均の周りに分散している状況 (6 個の＋) と全体として正の方向に偏っている状況 (6 個の●)．

以上の議論から，統計の数学でしばしば現れる**自由度 (degree of freedom)** とは何かということがわかる．既に何度も出てきたように，標本平均は次のように n 個の変数から定義される．

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

[*7] もちろん「数式は語る」のです．耳をすまそう．

[*8] 仮に標本のサイズを 1 として，そこから X_1 を無限回取り出すことを考えると，$X_1 - \mu$ の 2 乗の和の期待値なので σ^2 になることがわかる．

これを与えられたものとすると[*9]，n 個の変数はこの 1 つの関係式によって束縛されて，ある変数たとえば X_n は，次のように $\overline{X}$ と他の $n-1$ 個の変数を使って表されることになる．

$$X_n = n\overline{X} - \sum_{i=1}^{n-1} X_i$$

これを式 (6.6) に代入すると，標本分散 s^2 は $X_1, \ldots, X_{n-1}$ と $\overline{X}$ で表される．結局，実質的な変数の数は $n-1$ となり，これが**自由度**と呼ばれるのである．

以下の章でも，χ^2 分布や t-分布など，標本抽出と関連した統計分布で自由度が登場する．そこでも抽出されたデータの値から標本平均を引き去った量を用いるときには自由度が 1 だけ減るという形になっているのは，ここで説明した理由による．

▼物理における自由度　物理で気体分子の運動を勉強して自由度という用語に接した人がいるのではないだろうか．アルゴンのような単原子分子の気体は三次元の 3 つの運動の自由度をもち，アルゴン分子が 2 個であれば，2 つの原子は束縛されることなく自由に運動できるので，全体の自由度は 6 になる．

一方，酸素のような 2 個の原子からなる分子の場合，それらの座標を $(x_1, y_1, z_1), (x_2, y_2, z_2)$ として，原子同士の結合距離を d としたときに $d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$ という関係が成立するために，分子の運動の自由度は $3 \times 2 = 6$ よりも 1 少ない 5 になる．このことが気体のエネルギーや比熱に効いてくる．このように変数間に束縛の関係があるために自由度が小さくなることは，物理でもよく出てくるありふれた現象である．

6.3.2　標本分散の分散については？

さて，組み合わせとして最後に残っている標本分散の分散 $V[s^2]$ についてはどうだろうか．その場合の一般的な関係式は存在せず，正規母集団から抽出された標本について，χ^2-分布が現れる．それについては後で取り扱う．

[*9] 何しろ真の平均 μ は未知なのだから，私たちは $\overline{X}$ を使うしかない．

6.4 正規母集団

6.4.1 正規母集団と標本平均の分布

以上で，標本平均 $\overline{X}$ の期待値と分散の**分布**が，母集団の期待値と分散とどう結び付けられるかがわかった．しかし $\overline{X}$ がどのような確率分布に従うかは，まだわかっていない．その点を見ていこう．大きく整理すると次のようになる．

母集団が正規分布している場合： n の大きさに関わらず $\overline{X}$ は正規分布に従う．つまり**「正規分布は，子どもたちもやっぱり正規分布」**なのである．実際 $n = 1$ のとき，この標本は母集団の中のデータそのものであり，十分な回数の抽出を行ってみれば元の正規分布が現れることは明らかである．

さらに，2 つの正規分布に従う確率変数があるとき，それらの和も正規分布に従うので，n が複数の場合にも $\overline{X}$ は正規分布する．

母集団が正規分布していない場合： このときには，n が十分大きいならば，$\overline{X}$ は正規分布になる．つまり，**「非正規分布の子でも，たくさん集めて平均したら正規分布になる」**わけだ．これを保証しているのが中心極限定理である．

しかし，n が少数のときには，$\overline{X}$ がどのような分布に従うかについて一般的な法則はなく，個別に解決すべき問題になる．

以上のように，母集団が正規分布しているときには，標本平均の分布についても，正規分布の性質を使って非常に強力な手法が使える．その意味で，正規分布に従う母集団のことを**正規母集団**と呼ぶ．

正規分布はまた，世論調査のように非正規母集団から大きな標本を取り出す場合にも利用できて，統計的な扱いにおいて中心的な役割を果たすことができるのである．

例題 6–3 有権者の中の内閣支持率が 30% であったとしよう．1000 人をランダムに選んでアンケートをとったとき，そこから得られる支持率の期待値と分散を求めよ．

母集団中の支持を 1, 不支持を 0 とする．この母集団は，0 と 1 しか含まない極端な集合であって，正規母集団ではないが，以下のような取り扱いが可能である．

まず，母平均と母分散を求めておく．母集団中の支持の割合を $p = 0.3$ としよう．つまり，母集団の中で 1 つのデータを見たときにそれが 1 である確率は p, 0 である確率は $1 - p$ だ．したがって式 (3.10) より．

$$\mu = 1 \times p + 0 \times (1 - p) = p$$

となり，また式 (3.11) より，

$$\sigma^2 = (1-\mu)^2 \times p + (0-\mu)^2 \times (1-p) = p(1-p)$$

となる．よって $\mu = 0.3$, $\sigma^2 = 0.21$.

これから，標本平均の期待値と分散は，それぞれ $\mu = 0.3$ と $\sigma^2/n = 0.21/1000 = 0.00021$ となる．分散がきわめて小さい感じがするが，これから標準偏差は 0.0145 となるので，μ と比較して妥当な値になっている．

■カテゴリカル変数と二項分布

支持と不支持，性別といったカテゴリー分け（分類）は，数値的な変数ではないので，**カテゴリカル変数**[*10]とか**名義変数**と呼ばれる．一般に社会調査などにおいてはカテゴリカル変数は非常によく登場する．

カテゴリカル変数であっても適当な数に対応させることで確率変数として取り扱うことができ，特に 2 通りのカテゴリーしかない場合にはそれらを 0 と 1 に対応づけることで，このように単純な定量的な取り扱いが可能になる．

6.4.2　標本平均の分布

■標準化変換された標本平均の分布

$N[\mu, \sigma^2]$ に従う正規母集団から，大きさ n の標本を抽出したとする．すると前節の議論と式 (6.3),(6.4) から，$\overline{X}$ は平均 μ と分散 σ^2/n (標準誤差 $\sigma/\sqrt{n}$) をもつ正規分布に従う．図 6.4 に，そのようすを示した．この図も今後しばしば登場する．標準誤差が n の平方根の逆数に比例することから，標本が大きいほど $\overline{X}$ の広がりは狭まって，分布がシャープになることをしっかり見ておこう．

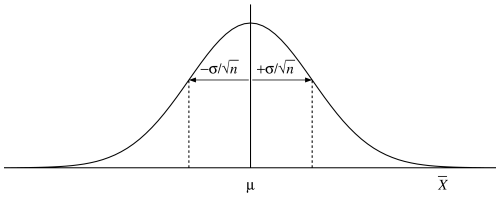


図 6.4　$N[\mu, \sigma^2]$ に従う正規母集団から大きさ n の標本を抽出したときの $\overline{X}$ の分布.

さてここで，今後のために 図 6.4 を標準化変換してみよう．

[*10] カテゴリー変数ともいう.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \tag{6.8}$$

こうして得られた Z は標準正規分布 $N[0,1]$ に従うことを第 5 章ですでに見た (p.81). このように，現実の確率変数に適切な変換を施すことで，なんらかの標準的な確率分布に従うようにするという手続きは，一般化を助けてくれるものだ．次の例に進もう．

6.4.3 母分散なしで標本平均の分布を表現する

ここで，標本抽出のもっとリアルな状況を想定してみよう．

標本平均 $\overline{X}$ の分布を標準化変換して得られた式 (6.8) は，$\overline{X}$ が式 (6.3)，(6.4) で決まる期待値と分散をもつ正規分布に従うことから導かれている．これらのパラメータは母平均 μ と母分散 σ^2 から導かれているわけだ．

しかし，私たちが現実に手にしているのは，あくまで標本のデータだけしかないのである．未知の，というよりむしろ，「知りたい」パラメータである母集団の統計量をなるべく使わないで $\overline{X}$ の分布を決めることができたら，応用はずっと広がるにちがいない．

■標本のサイズが大きい場合

式 (6.5) で示されるように，s^2 の期待値は母分散 σ^2 に係数 $\frac{n-1}{n}$ を掛けたものに等しい．式を再掲しておく．

$$E[s^2] = \frac{n-1}{n}\sigma^2$$

これは，十分な回数サンプルを繰り返すと，s^2 の平均は $\sigma^2 \times \frac{n-1}{n}$ に等しくなるとも言い換えられるわけだ．そこで，上の式の期待値の括弧をえいやっと外して等式にしてしまって，少しいじってやると，

$$\sigma^2 = \frac{n}{n-1}s^2$$

となる．つまり，この右辺をもって未知の標本平均の「代用品」に使おうというのが，標本のサイズが大きい場合のアイディアである．

それではさっそく，この場合についての標本平均 $\overline{X}$ の分布を描いてみよう．まず標本が大きいのだから，母集団が正規分布しているか否かに関わらず $\overline{X}$ は正規分布しているとしてよい（中心極限定理）．また，一つ前にやった，母分散が既知の場合と同様に，$\overline{X}$ の標準偏差 (標準誤差) は $\frac{\sigma}{\sqrt{n}}$ の形だが，ここでは σ を $\sqrt{\frac{n}{n-1}}s$ で置き換えることになるので，

$$\frac{\sqrt{\frac{n}{n-1}}s}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

となる．結局 $\overline{X}$ は平均 μ と分散 $s^2/(n-1)$ (標準偏差 $\frac{s}{\sqrt{n-1}}$) をもつ正規分布に従うことになる．

ここでも，この結果を標準化変換によって一般化しておこう．式 (6.8) とまったく同様にやればよい．次のようにして Z を決めると，

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{6.9}$$

Z は標準正規分布 $N[0,1]$ に従う．

■標本のサイズが小さい場合—スチューデントの t-分布

今度は，正規母集団から抽出された標本のサイズが小さい場合を考える．ポイントは式 (6.9) だ．これを特別扱いするために，Z を T に変更しよう．

$$T = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{6.10}$$

大きな標本ではこの s をあまり動かないものとして扱ったのが，前の扱いになる．そのため Z(ここでは T) は標準正規分布になった[*11]．ところが，標本のサイズが小さいときには，そのことが言えなくなる．

6.3 節で説明したように，標本分散 s^2 は n が小さいときには母分散 σ^2 よりも平均的には小さく，かつ変動の幅が大きい確率変数として振る舞う．標本標準偏差 s も，もちろんそうなる．図 6.5 に，そのことをシミュレーションで確かめた結果を示した．

ということは，式 (6.10) の分母の s は，時として大きな値を取ることになる．それはどういう結果をもたらすかというと，関数を横に引き伸ばす効果をもつのだ．つまり，正規分布の形を左右に引き伸ばしたような分布が作られることになる．これが**スチューデントの t-分布**，あるいは単に **t-分布**とか**スチューデント分布**とよばれる確率分布である[*12]．

t-分布は式 (6.11) で表され，正規分布のような形をしているが，図 6.6 のようにやや裾の広がった形になる．また**自由度**というパラメータ ν があって，その値によって形が異なる．

$$f_\nu(T) = c\left(1 + \frac{T^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad (\nu = 1, 2, 3, ...) \tag{6.11}$$

ここで c は $f_\nu(t)$ の全面積が 1 であるようにするための定数であり，ν は自由度と呼ばれている．

[*11] 厳密にはこの説明は不十分で，s の分布の幅が小さく，かつ正規分布に近い対称な分布になっていることが，T が正規分布する理由なのだが，それを議論すると数学的な細部に突っ込みすぎるだろう．

[*12] Student というのは，この分布を提案した数学者のペンネームで，本名はウイリアム・ゴセット (William S. Gosset)．ゴセットはビールで有名なギネス社の社員で，ビールの品質管理に彼が考案した新しい統計的方法を適用して，会社の業績発展に多大な寄与をした．しかし会社は社員が社外で研究発表を行うことを禁止していた．そこでスチューデントというペンネームで論文を投稿していたのである．

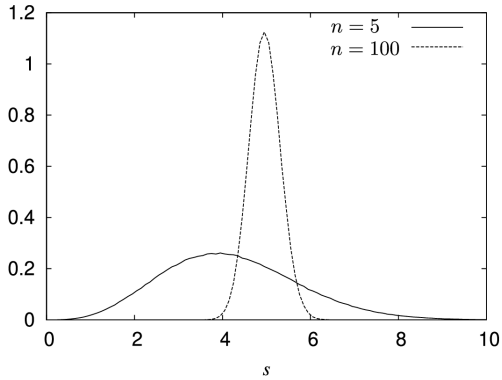


図 6.5　標準正規分布 $N[0, 25]$ を母集団として抽出された標本標準偏差 s の分布：$n = 5$ の場合には $n = 100$ に比べて分布が大きく広がっており，特に裾が右側に大きく伸びていることが特徴的である．抽出はいずれも 100 万回行った．

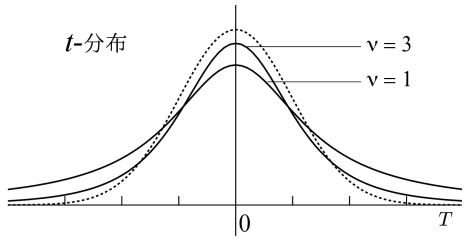


図 6.6　$\nu = 1, 3$ のときの t-分布の形: 破線は標準正規分布曲線

ここで，小さな標本の場合の標本平均の分布がどのように t-分布によって表されるかを見てまとめとしよう．

$N[\mu, \sigma^2]$ に従う正規母集団から，大きさ n の標本を無作為抽出して得られる標本平均 $\overline{X}$ と標本分散 s^2 を得たとしよう．このとき，

$$T = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{6.12}$$

で定義される変数 T は，自由度 $n - 1$ の t-分布に従う．

6.5　正規母集団と χ^2 分布

6.5.1　標本分散から得られる情報

前節までに，標本分散 s^2 が母集団の統計量とどのような関係にあるかについて，ほんのわずかの結論しか出していない．すなわち，式 (6.5) で，$E[s^2]$ という標本分散の期待値と母分散の関係が与えられているだけである．

しかし私たちが，標本抽出の結果から母集団の分散がどうなっているかを自信をもって知るためには，単に s^2 の平均（期待値）だけではなく，分散がどうなっているかも知っておかないといけない．抽出によって得られた標本分散を使って，母集団の分散についてどの程度のことが言えるのか，それを知らないと危うい結論しか出せないからである．

たとえば 94 ページの例においては，標本分散の複数回の測定から母集団の分散を推定しているわけだが，1 回の抽出ごとの標本分散の値はかなりばらつくので，それらの平均を期待値として安心して使っていいかという問題が残っている[*13]．

この問題については，しかしながら，前節までのような一般的な結論は存在しない．ただし，正規分布をしている母集団，すなわち**正規母集団**の場合については，以下のような結論が導かれている[*14]．これはまず $\mu = 0, \sigma^2 = 1$ であるような正規母集団 $N[0,1]$ について次のように表される．

$N[0,1]$ の正規分布をしている母集団から，n 個の無作為抽出を行ったとき，

$$\boldsymbol{Z = X_1^2 + X_2^2 + ... + X_n^2} \tag{6.13}$$

なる Z は，

$$\boldsymbol{T_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}} \tag{6.14}$$

という確率分布に従う．ここで式 (6.14) は自由度 n の χ^2 分布と呼ばれる確率密度関数である．

$n = 1, 2, ..., 7$ についての χ^2 分布関数の具体的な関数表現は次のとおりである[*15]．図 6.7 にはそのグラフを示した．

[*13] 期待値と平均値は数式の上では同じ形である．しかし期待値というのは無限に回数を重ねたとして，その平均がどこに近づくかという仮想的なものである．一方，実際に実現した事象から計算される平均値は，期待値のまわりに広がって分布する数値になるのである．

[*14] ただし，大きな母集団が正規分布をする傾向は，中心極限定理で保障されているので，正規母集団についての結論は広い一般性をもつものである．

[*15] この関数の形は複雑であるが，覚えておく必要は全くない．グラフで関数の概形を知っておけばよい．

$$T_1(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \tag{6.15}$$

$$T_2(x) = \frac{1}{2} e^{-x/2} \tag{6.16}$$

$$T_3(x) = \frac{1}{\sqrt{2\pi}} x^{1/2} e^{-x/2} \tag{6.17}$$

$$T_4(x) = \frac{1}{4} x e^{-x/2} \tag{6.18}$$

$$T_5(x) = \frac{1}{3\sqrt{2\pi}} x^{3/2} e^{-x/2} \tag{6.19}$$

$$T_6(x) = \frac{1}{4} x^2 e^{-x/2} \tag{6.20}$$

$$T_7(x) = \frac{1}{15\sqrt{2\pi}} x^{5/2} e^{-x/2} \tag{6.21}$$

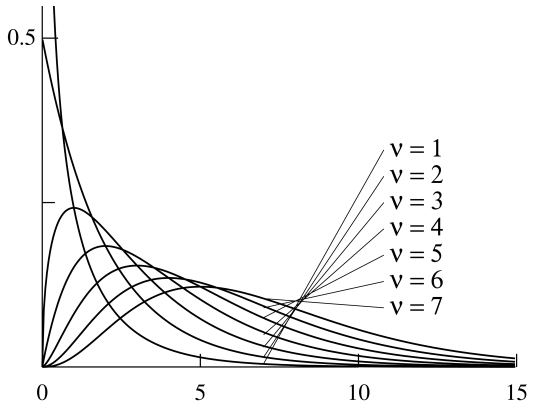


図 6.7 $\nu = 1, 2, ..., 7$ に対する χ^2 分布の密度関数

式 (6.14) に登場する $\Gamma(x)$ はガンマ関数と呼ばれ，階乗 $n!$ を 連続関数に拡張したものである[*16]．なお，この式では n の代わりに ν を使うこともよくあるので，表を引いたりするときには迷わないこと．

さらに，標準正規分布でない一般の正規分布については，次のように言い換えられる．

母集団は平均値 μ，分散 σ^2 をもつ正規分布 $N[\mu, \sigma^2]$ をしているとする．そこから n 個の無作為抽出を行ったとき，

$$\boldsymbol{Z = \frac{1}{\sigma^2}\left((X_1 - \mu)^2 + (X_2 - \mu)^2 + ... + (X_n - \mu)^2\right)} \text{ [*17]} \tag{6.22}$$

[*16] 具体的には，x が正の整数のときには，$\Gamma(x) = (x-1)!$ で，また半整数に対しては，$\Gamma(1/2) = \sqrt{\pi}, \Gamma(3/2) = \frac{1}{2}\sqrt{\pi}, \Gamma(5/2) = \frac{3}{4}\sqrt{\pi}, \Gamma(7/2) = \frac{15}{8}\sqrt{\pi}, \Gamma(9/2) = \frac{105}{16}\sqrt{\pi}, ...$ となる．

[*17] 式 (6.13) の $\boldsymbol{X_1, X_2, ...}$ を標準化した $\frac{(X_1 - \mu)}{\sigma/\sqrt{n}}$ で置き換えることで，この式が得られる．

なる Z は，自由度 n の χ^2 分布 に従う．

上の命題によれば，もしも母集団の平均値 μ が既知であるならば，何回も抽出を繰り返すことによって，母分散 σ^2 を推定できることになる．

さらにもうひとつ，実用的な結果を述べておこう．

母集団は平均値 μ，分散 σ^2 をもつ正規分布 $N[\mu, \sigma^2]$ をしているとする．そこから n 個の無作為抽出を行ったとき，

$$Z = \frac{1}{\sigma^2}\left((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right) = \frac{ns^2}{\sigma^2} \tag{6.23}$$

なる Z は，自由度 $n-1$ の χ^2 分布 に従う．ただし，ここで s^2 は式 (6.2) で表される標本分散である．

上の命題によれば，何度も標本抽出を繰り返して，その標本分散 s^2 がどのような分布をしているかを見れば，母分散 σ^2 が推定できることになる．

6.5.2　χ^2 分布表の利用

χ^2 分布を統計的な評価に使うときに必要なのは，この関数の，与えられた区間にわたる面積である．図 6.8 を見てほしい．ここで影をつけてある部分の面積 α と t の関係が求められれば，無作為抽出による分散について様々な計算を行うことができる．そこで χ^2 分布 $T_n(x)$ について，α の代表的な値ごとに，それに相当する t を計算したものを数表化したものが用意されている．

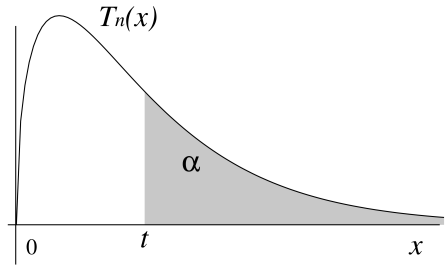


図 6.8　χ^2 分布の数表の意味

χ^2 分布の表の一部を下に示す．

α	0.995	0.975	0.950	0.900	0.500	0.05	0.025	0.01	0.005
$\nu = 6$	0.676	1.24	1.64	2.20	5.35	12.59	14.45	16.81	18.55
$\nu = 9$	1.73	2.70	3.33	4.17	8.34	16.92	19.02	21.67	23.59
$\nu = 10$	2.16	3.25	3.94	4.87	9.34	18.31	20.48	23.21	25.19
$\nu = 11$	2.60	3.82	4.57	5.58	10.34	19.68	21.92	24.73	26.76

この表の意味は，例えば $T_6(x)$ において α が 0.995 になるような $x = t$ の位置は 0.676 の点であるということである．

たとえば，母集団 $N[0,1]$ から取った大きさ 6 の標本を取ったとしよう．このとき，式 (6.13) で与えられる $Z = \frac{1}{6}(X_1^2 + \cdots + X_6^2)$ の分散が 0.676 よりも大きい確率は 99.5% であるということになる．

なお，χ^2 分布の表は正規分布表と異なって，1%, 5%, 90% などのように代表的な α についてのみ t の値を引けるようになっている．実用上はこれで十分だからである．

例題 6–4　($\boldsymbol{\chi^2}$ 分布の実験的検証) 92 ページの表 6.2 のデータから，大きさが 10 の標本を 30 回抽出してみたところ，得られた標本分散は次のようになった（便宜のために結果は昇順に並べ替えてある）．

> 7.02, 8.03, 8.53, 9.34, 13.12, 13.65, 14.17, 14.24, 15.77, 15.83, 16.13, 16.30, 16.52, 16.56, 16.89, 17.41, 17.47, 17.77, 18.25, 18.36, 19.21, 19.48, 20.68, 21.25, 22.91, 24.33, 25.54, 26.24, 26.80, 41.87

この結果から，50% の標本分散が含まれる「切れ目」を求めよ．さらに χ^2 分布を使って，母分散を求めてみよ．

示されているデータを見ると，ある値以上に標本分散 s^2 の 50% 以上が含まれる「切れ目」は，16.89 と 17.41 の間であるから，17.15 である．一方，式 (6.23) より，$Z = ns^2/\sigma^2$ は，自由度 $n-1 = 9$ の χ^2 分布に従うから，表を見ると，その値が 8.34 のところが「切れ目」に相当している．したがって，$Z = \frac{ns^2}{\sigma^2} = 10 \times 17.15/\sigma^2 = 8.34$．よって母分散は 20.6 となる．一方，9 ページの問題 1–1 の解答にあるように，この母集団の母分散は 17.84 である．こうやってみると，それほど近い値が得られているわけではない．ほどほどの一致というところである．

例題 6–5　($\boldsymbol{\mu}$ が既知の場合) $\boldsymbol{N[3, \sigma^2]}$ に従う母集団から 6 個の標本を多数回無作為抽出した．その結果，抽出回数の 50% で $\boldsymbol{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_n - \mu)^2}$ が 21 を超えていた．母分散 $\boldsymbol{\sigma^2}$ を求めよ．

自由度 6 の χ^2 分布の表から，$\alpha = 0.5$ となる t の値は 5.35 となることが分かる．つまり式 (6.22) の $Z = \frac{1}{\sigma^2}\left((X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_6 - \mu)^2\right)$ が 5.35 のところで，α が 0.5 になるわけである．そこで，$5.35 = \frac{1}{\sigma^2} \times 21$ より，$\sigma^2 = 3.93$ となる．

例題 6–6　($\boldsymbol{\mu}$ が未知なので標本分散を使う場合) 母平均が未知の正規母集団から大きさ 11 の標本を何回も無作為抽出した．その結果，標本分散の 90% が 12.5 を超えていた．母分散 $\boldsymbol{\sigma^2}$ を求めよ．

この場合には，μ が知られていないので，式 (6.22) は使えない．しかし標本分散は抽出の結果としていつでも求められるものであるから，式 (6.23) を使うことができる．

このとき注意しなければならないのは，標本の大きさ n に対して，その標本分散は**自由度 $n-1$ の χ^2 分布に従う**ということである．したがってここでは，$\nu = 10$ として表を引くことになる．そうすると，$\alpha = 0.900$ となるのは $t = 4.87$ のときである．このときに，標本分散はちょうど 12.5 になっているというのが題意である．

そこで，式 (6.23) より $4.87 = \frac{ns^2}{\sigma^2}$，これに $n = 11$, $s^2 = 12.5$ を代入して $\sigma^2 = 28.2$ を得る．

【章末問題】

問題 6–1 ある肥料の作物への効果を確かめるために，1 ヘクタール ($10000\mathrm{m}^2$) の畑に 10 m 置きに線を引いて 1 アール ($100\mathrm{m}^2$) の区画を 100 個作った．そうしておいて，5 つおきに区画を選んでその肥料を施して栽培し，他の区画は従来通りの栽培を行うこととした．この選び方は正しいか．

問題 6–2 血液型性格判断，つまり血液型が性格と関係するという考えにもとづく性格判断は，1970 年代にある人とその息子とが相次いで本を書いたことで広まったとされている．著者たちは本に読者アンケートのはがきを付けて返送してもらい，次の本ではそのアンケートにもとづいて所説を展開した．このアンケートの取り方は妥当かどうかを述べなさい．

問題 6–3 ある工場で生産している部品の質量は，平均が 54.2 g，標準偏差は 0.22 g である．この製品を 10 個抜き取って質量を測ったとき，その平均値が 54.1 g〜 54.3 g の間にある確率を求めなさい．

問題 6–4 ある養鶏場から出荷された卵から，12 個ずつ無作為にとって秤量し， これを 30 回繰り返して，その質量 (g) の標本分散を求めたところ，次のように分布していた．結果は小さい順に並べ替えてある．

8.76, 9.47, 9.99, 11.85, 12.59, 13.23, 14.79, 18.83, 20.32, 20.74, 21.00, 21.11, 22.40, 23.43, 24.61, 26.14, 27.41, 29.53, 32.22, 33.51, 41.81, 50.57

例題 6–4 にならい，母平均が未知であるものとして，この養鶏場の卵の質量の母分散を χ^2 分布を使って推定せよ（小数第 1 位まで記入）．ただし卵の質量が正規分布していると仮定する．

第 7 章

推定

標本抽出によって集められたデータは母集団の状況を反映はするものの，そのままの真実を伝えてはくれない．私たちは確率分布を考えることで，データから母集団の真の統計量を推定することができる．その手法を学ぼう．

7.1 点推定と区間推定

世論調査で発表される内閣支持率などの数字は標本平均であり，人々はその数字を無意識のうちに母平均として認識している．しかし，前章で見たように，抽出された標本の平均は抽出のたびごとに異なった値をとる確率変数であり，その期待値（平均）は母平均に一致するが，ある分散をもつ．

つまり $\overline{X}$ を復元抽出によって得られた標本平均とすると，その期待値と分散は

$$E[\overline{X}] = \mu \tag{7.1}$$

$$V[\overline{X}] = \frac{\sigma^2}{n} \tag{7.2}$$

で表されることになる．たとえば内閣支持率が 35% であると発表された場合には，その数値は一定の幅をもっているのである[*1]．

そこで，抽出によって標本平均が得られたとして，それがどの程度信用できるかという問題を考える必要が出てくる．標本平均 $\overline{X}$ は確率変数であるから，母平均とちょうど一致する確率はゼロかきわめて小さいからである．そこで $\overline{X} \pm \alpha$ のようにある範囲をとってやれば，その中に入る確率を指定できることになる．その場合に幅 α を大きくとれば

[*1] 世論調査では 1000 人から 3000 人ほどの無作為抽出標本をもとに統計量が算出されるのが普通だが，「無回答」という層も少なからず存在する．もし，これらの回答をしない人々が一定の傾向を持っているならば，得られた結果はそれを反映した偏りをもつであろう．その種の結果の偏りは，数学的な取り扱いだけでは取り除くことが難しい．

正しい確率は大きくなるが，情報としては意味がなくなってくるし，幅を小さくとれば情報としてシャープになるが，間違う危険は大きくなる．

すなわち，

> **この調査結果が 95% 正しいというためには，どれくらいの幅を持たせておかなければならないか？**

といったことを検討しなければ正しい統計にはならない．

すなわち，「母平均はこれこれの値であると考えられる」というふうに誤差を許容しつつ一点の値で答えるような**推定 (estimation)** を**点推定 (point estimation)** といい，このようにして一点で示された値を**点推定量 (point estimator)** という．

一方，「母平均は $x_1 \sim x_2$ の範囲にある確率が 95% である」というふうに「幅」をもって推定することを**区間推定 (interval estimation)** といい，なされた推定結果を**区間推定量 (interval estimator)** という．言い換えると，区間推定量は**信頼区間 (confidence interval)** を伴った推定量である．

7.1.1　パーセント点と信頼区間

ここで正規分布に関わって，統計的推定でよく用いられる量を定義しておこう．図 7.1 の標準正規分布 $N[0,1]$ で，右端から面積が α となるようにとった点を z_α と書く．この点は通常，**$100(1-\alpha)$ パーセント点**と呼ばれることが多い．たとえば，$\alpha = 0.05$ であれば，この点は $z_{0.05}$，すなわち 95 パーセント点であり，正規分布表を参照して，$\Phi(z) = 0.95$ となる z を見つけて，**95 パーセント点は 1.645** であるということになる．

なおパーセント点は，第 1 章で登場したパーセンタイル（→ p.12）と同じ内容である．

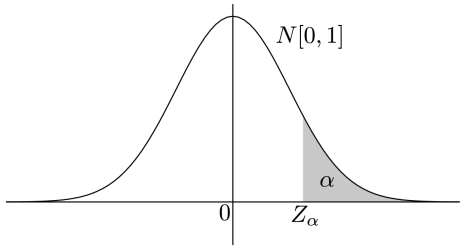


図 7.1　標準正規分布 $N[0,1]$ におけるパーセント点の定義．z_α の点を $100(1-\alpha)$ パーセント点という．

いくつかの重要なパーセント点を表 7.1 に示した．

ここで注意しておかなければならないことは，一般的な正規分布の利用法では，平均値を中心として対称な面積を考えることが多いということである．つまりたとえば 95 パーセント点と 5 パーセント点とが対になって，その内側の 90% の領域を作っていると捉え

るのである．

そこで，z_α ではなくて，$z_{\alpha/2}$ もよく用いられる．これは，その右側の面積が $\alpha/2$ であるような点であるから，$-z_{\alpha/2}$ の左側にも対称に領域を作ったとすると，合わせて α の面積が両端にできることになり，その間には $1-\alpha$ の面積が残ることになる．図 7.2 を参照のこと．

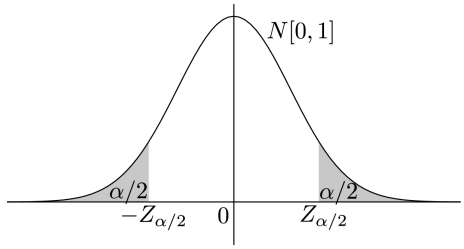


図 7.2　$z_{\alpha/2}$ の意味

▼パーセント点と正規分布表　表 7.1 を使わずに，巻末の正規分布表から 90% 点などを求めよ．

90% 点について求めてみる．正規分布表をみると，$z=1.28$ のときに $\Phi(z)=0.899727$, $z=1.29$ のときに $\Phi(z)=0.901475$ となっている．短い区間であるから直線関係を仮定して，$\Phi(z)=0.9$ となる z を比例配分で求めればよい (比例配分の計算方法については巻末付録参照)．

$$z = 1.28 + \frac{0.9 - 0.899727}{0.901475 - 0.899727} \times (1.29 - 1.28) = 1.28156 \approx 1.282$$

▼z_α と $z_{\alpha/2}$　標準正規分布 $N[0,1]$ がある．平均値を中心に 90% および 95% の面積を含む区間の両端の値を表 7.1 から読み取れ．

90% を含む区間は，**両側の端**に 5% ずつを含むようなとり方をするのであるから，95 パーセント点を使えばよい．したがって，$z_{0.05}$ を選んで，区間の右端は 1.645, 左端は

表 7.1　正規分布の各パーセント点

パーセント	90	95	97.5	99	99.5
α	0.10	0.05	0.025	0.01	0.005
z_α	1.282	1.645	1.960	2.326	2.576

-1.645 となる．同様に 95% の場合には，両端は ± 1.960 となる．これらはよく使われる値であるので気に止めておくとよい．

7.2 不偏推定量

μ や σ^2 のような母集団の統計量 θ が，$\overline{X}$ や s^2 のような標本の統計量 Θ の期待値として推定されるとき，すなわち，

$$\theta = E[\Theta] \tag{7.3}$$

として推定されるとき，Θ を母数 θ の**不偏推定量 (unbiased estimation)** [*2]という．つまり，Θ ($\overline{X}$ や s^2) を多数回標本抽出して得て，それらの平均値を知ることができるならば，θ になるであろうということである．

すでにこの形の式の例は，

$$E[\overline{X}] = \mu \tag{7.1}$$

が与えられていて，これから，**母平均 $\boldsymbol{\mu}$ の不偏推定量は標本平均 $\boldsymbol{\overline{X}}$ である**．また母分散 σ^2 の不偏推定量については，式 (6.5)，すなわち，

$$E[s^2] = \frac{n-1}{n}\sigma^2$$

から，

$$\sigma^2 = E\left[\frac{n}{n-1}s^2\right]$$

と変形でき，**$\frac{n}{(n-1)}s^2$ は母分散の不偏推定量と見なせる**ことが分かる．これは更に，

$$s^2 = \frac{1}{n}\left((X_1-\overline{X})^2 + (X_2-\overline{X})^2 + ... + (X_n-\overline{X})^2\right)$$

という定義を用いれば，次のようになる．

$$\frac{1}{n-1}\left((X_1-\overline{X})^2 + (X_2-\overline{X})^2 + ... + (X_n-\overline{X})^2\right) \tag{7.4}$$

これが σ^2 の不偏推定量であり，式 (7.4) の値を**標本不偏分散**，その平方根を**標本不偏標準偏差**という[*3]．

結局，もしもたった 1 回の標本抽出で母分散を知りたいときには，式 (7.4) を使って計算した値をもって最も確からしい推定値にするのである．このことを理解しておいてほしい．

[*2] unbiased というのはバイアス，つまり偏りが入っていないという意味である．

[*3] Excel で分散，標準偏差を求める関数として使われている VAR,STDEV は本来の分散や標準偏差ではなく，この標本不偏分散および標本不偏標準偏差を指している．そのため，誤解して使っているケースもかなりあるはずである．

7.3　母平均の推定

一般に行われる統計調査で最もよく利用されるのは，平均である．政党や内閣の支持率，番組の視聴率，人の寿命，身体測定のデータ，その他数えきれないほどの統計で平均が登場する．この場合，標本平均 $\overline{X}$ を真の平均，つまり**母平均** μ と考えることを，私たちは暗黙のうちに認めている．そのことを保証しているのは，**標本平均**が母平均の**不偏推定量**であるという事実だ．このことは，母平均の**点推定量**として標本平均を使っていると言い換えてもよい．

一方，知りたい母平均が確率的に見てどんな範囲に広がっているかを知ることも大切だ．そのためには**区間推定**を行うことが必要になる．

そのための理論的な枠組みはすでに第 6 章で取り上げた (p.97, 正規母集団)．それに従って，次の 3 つの**モデル**に分けて考えていこう[*4]．

最初は，母分散 σ^2 が知られているケースである．この場合には，式 (7.1)，(7.2) の関係をそのまま利用して，標本平均 $\overline{X}$ の広がり，つまり分布を知ることができる．最も単純なモデルである．

しかし，母分散が分かっているのに母平均が分からないなどという幸運な，あるいは虫のよい状況はそうそうあるものではない．そんなときでも，標本のサイズが大きければ，$\overline{X}$ が正規分布していることが**中心極限定理**から期待できる．また，94 ページの式 (6.5) を使って標本分散から母分散を推定することができるので，それを使って標本平均の広がりを知ることが可能である．

しかし，標本のサイズが小さい場合には，$\overline{X}$ が正規分布をするという想定はできなくなるので，上記の方法はもはや適用できなくなる．その場合にも一定の条件の下で Student の t-分布を仮定することで，$\overline{X}$ の広がりを考えることができることになる．

7.3.1　母分散が既知の場合

このケースは 6.4.2 節 (p.98) で検討してある．すなわち，$\overline{X}$ を次のように標準化変換したとすると，

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \tag{7.5}$$

式 (7.5) で定義される Z は標準正規分布に従うことになる．

[*4] ここで言っている「モデル」というのは，考えている状況をどのように数学的な表現で記述できるかという事を指す言葉である．

仮に 90% の信頼区間を求めるとしよう．このときには 95% 点を考えればよいことになるが，表 7.1 から，それが 1.645 であることがわかる．結局，上の Z の値が ± 1.645 になるような 2 つの $\overline{X}$ の間が，求める信頼区間だということになる．

そこで，式 (7.5) で $Z = \pm 1.645$ と置いて $\overline{X}$ を求めると，信頼区間の両端が得られ，次のように信頼区間が決まる．

$$\left[\overline{X} - 1.645 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right] \tag{7.6}$$

一般化して，考えるべきパーセント点を λ として信頼区間を書き直そう．

$$\left[\overline{X} - \lambda \times \frac{\sigma}{\sqrt{n}}, \overline{X} + \lambda \times \frac{\sigma}{\sqrt{n}}\right] \tag{7.7}$$

この式から次のことが言える．

母分散 $\boldsymbol{\sigma^2}$ が既知の場合の母平均の信頼区間は，標本平均 $\boldsymbol{\overline{X}}$ を中心として，前後に標準誤差 $\frac{\sigma}{\sqrt{n}}$ の $\boldsymbol{\lambda}$ 倍の幅をもつ．

7.3.2 母分散が未知の場合—大標本

母分散が未知で，標本のサイズが十分に大きい場合については，99 ページで扱った．

結果を再度書くと，次の式で与えられる Z は標準正規分布 $N[0,1]$ に従う．

$$Z = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s}$$

考えるべきパーセント点を λ とすると，ここでも次のようにして信頼区間が表される．

$$\left[\overline{X} - \lambda \times \frac{s}{\sqrt{n-1}}, \overline{X} + \lambda \times \frac{s}{\sqrt{n-1}}\right] \tag{7.8}$$

もしも信頼度を 90% に取るならば，95% 点 1.645 を λ として信頼区間を計算すればよい．

7.3.3 母分散が未知な小標本 — Student の t-分布

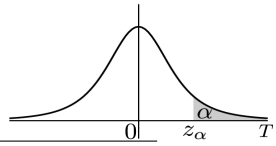
■Student の t-分布を使う

小さな標本の場合には，母分散を標本分散から直接に推定して，それから母集団の平均を区間推定する．これについては，100 ページ以降で扱った．それによれば，次の式で定義される T は，自由度 $n-1$ の t-分布に従う．

$$T = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{7.9}$$

注意しておくと，t-分布は標本のサイズによって分布の形が異なるので，パーセント点も正規分布とは異なった表にまとめられることになる (表 7.2).

表 7.2　t-分布の形とパーセント点：α の意味は図を参照



パーセント	90	95	97.5	99	99.5	99.75
α	0.10	0.05	0.025	0.01	0.005	0.0025
$\nu = 1$	3.078	6.314	12.706	31.821	63.657	
$\nu = 2$	1.886	2.920	4.303	6.965	9.925	14.089
$\nu = 5$	1.476	2.015	2.571	3.365	4.032	4.773
$\nu = 6$	1.440	1.943	2.447	3.143	3.707	4.317
$\nu = 7$	1.415	1.895	2.365	2.998	3.499	4.029
$\nu = 8$	1.397	1.860	2.306	2.896	3.355	3.832
$\nu = 9$	1.383	1.833	2.262	2.821	3.250	3.690
$\nu = 10$	1.372	1.812	2.228	2.764	3.169	3.581

さて，式 (7.10) の形を見ると，式 (7.7) や式 (7.8) と全く同じである，つまり信頼度 90% で区間推定を行うのであれば，ここでも 95% 点を使って (正規分布ではなく t-分布なのでその値は異なることに注意) 区間の両端を求める．これまでと同様に，与えられた信頼度に相当するパーセント点を λ とすると，次のように信頼区間が得られる．λ の値は t-分布のパーセント点の表から探せばよい．

$$\left[\overline{X} - \lambda\frac{s}{\sqrt{n-1}},\ \overline{X} + \lambda\frac{s}{\sqrt{n-1}}\right] \tag{7.10}$$

いくつかの n の値に対する t-分布のパーセント点が表 7.2 に掲げてある．詳しい表は巻末付録に掲載してある．これを用いて，実際に推定を行うことができる．なお，表によっては異なる面積の取り方をしているものもあるが，その場合でも意味を考えれば，それほど混乱することはない．

7.3.4 平均値の区間推定の例

例題 7–1 平均値の区間推定

等級 L の卵の大きな箱から 10 個を抽出して，その質量 (g) を測定したところ，以下のようになった．

65.1, 67.5, 71.5, 68.4, 70.1, 72.2, 68.7, 69.3, 70.6, 67.1

この等級の卵の質量はほぼ正規分布しており，母分散は 4.0 g^2 であることが知られているものとする．この箱の卵の質量の平均値を 90% と 95% の信頼区間で推定せよ．

まず，これらから標本平均を計算すると 69.05，また題意より $n = 10$, $\sigma^2 = 4.0$ である．そこで 90% の信頼区間をとるときには，$\alpha = 0.05$ に相当するから，表 7.1 より，$z_{\alpha/2} = z_{0.05} = 1.645$ を読み取る．したがって式 (7.6) から，$69.05 - 1.645 \times \sqrt{4.0}/\sqrt{10} = 68.0$ と $69.05 + 1.645 \times \sqrt{4.0}/\sqrt{10} = 70.1$ が信頼区間の両端である．

すなわち，この卵の母平均 μ の信頼区間は，90% の信頼水準で $68.0 < \mu < 70.1$ と推定される．また信頼水準を 95% にとったときには，同様の計算により，信頼区間は $67.8 < \mu < 70.3$ となる．

例題 7–2 大標本の平均値の区間推定

ある県の高校生から 40 人を無作為に選んで，体重を測定したところ，標本平均 $\overline{X}$ と標本分散 s^2 はそれぞれ，53.8 kg, 18.81 kg^2 であった．これから母集団の平均値を 90% の信頼区間で推定せよ．

上の議論に従えば，信頼区間は

$$53.8 - 1.645 \times \frac{\sqrt{18.81}}{\sqrt{40-1}} = 52.6575..., \; 53.8 + 1.645 \times \frac{\sqrt{18.81}}{\sqrt{40-1}} = 54.9424...$$

となる．

答え：平均値は 90% の信頼水準で $52.7 < \mu < 54.9$ と推定される．

例題 7–3 小標本の平均値の区間推定

等級 L の卵の大きな箱から 10 個を抽出して，その質量 (g) を測定したところ，

65.1, 67.5, 71.5, 68.4, 70.1, 72.2, 68.7, 69.3, 70.6, 67.1

であった．箱全体の卵の質量は正規分布しているとして，この箱の卵の質量の平均値を 90% と 99% の信頼区間で推定せよ．

この問題は，117 ページの例題と同様に小標本のサンプルに関するものであるが，母分散は知られていない．母分散が分かっているケースというのはむしろまれであるから，この問題のほうが実際的である．

まず標本平均 $\overline{X}$ と標本標準偏差 s を求める．

$$\overline{X} = (65.1 + 67.5 + 71.5 + ... + 67.1)/10 = 69.05$$

$$s^2 = (65.1^2 + 67.5^2 + 71.5^2 + ... + 67.1^2)/10 - \overline{X}^2 = 4.1845$$

$$s = \sqrt{4.1845} = 2.045$$

また，自由度は $\nu = n - 1 = 9$ である．

90% の区間に含まれるためには $\nu = 9$ の時の t-分布の 95 パーセント点 $z_{0.05}$ を表から読みとって，1.833 を得る．したがって求める区間は，

$$\left[69.05 - 1.833 \times \frac{2.045}{\sqrt{9}}, 69.05 + 1.833 \times \frac{2.045}{\sqrt{9}}\right]$$

となって，平均値 μ は $67.8 < \mu < 70.3$ と推定される．

同様にして，信頼水準を 99% にとると，平均値は $66.8 < \mu < 71.3$ と推定される．確かめていただきたい．

7.3.5　平均値の推定のまとめ

■なぜ t-分布を使うのか

母分散 σ^2 が知られていない場合，標本平均 $\overline{X}$ の分散 $V[\overline{X}]$ が σ^2/n で与えられるというおなじみの関係式 (式 (6.4) 参照) は使えない．

そこで母分散の不偏推定量である不偏分散 $\frac{ns^2}{n-1}$ を利用して，$\overline{X}$ の分布の広がりを正規分布で扱おうというのが，7.3.2 節での大標本の取り扱いの基本的な考え方であった．

しかし，標本の大きさが小さくなると，中心極限定理によって標本平均 $\overline{X}$ が正規分布するとみなしてもよいという仮定はもはや成立しなくなる．その場合でも，**もし母集団が正規分布しているのであれば**，標本平均は母平均 μ にピークを持つ山形の分布をするであろう．

この場合，標本が小さいほど分布のすそは正規分布よりも広がるであろうことが直観的に予想される．また，標本が十分大きい極限では，その分布は正規分布に一致するようになるはずである．t-分布はそのような条件を満たすように作られている．このことは図 6.6 を見るとよく分かる．

7.3.6 区間推定におけるモデルの使い分け

ここまでの 3 つの場合の平均値の信頼区間の推定の手法について，表 7.3 にまとめた．ここで表の λ のパーセント点は，正規分布なら区間の幅だけで決まるが，t-分布の場合には標本の大きさ n に対して自由度として $n-1$ を選ぶ必要があることに注意しよう．

それでは，これらのモデルの使い分けについては，どういう基準を設ければよいのだろうか？

まず，大標本と小標本の区別は，標本のサイズ n がどのへんのところを境目にすればよいのだろうか．これについては，大体 $n = 20$ 程度を境にするということで実用的には問題ない．つまり Student の t-分布を使う目安としては抽出するデータの数が 20 程度以下としておけばよい．

また，十分な大きさの大標本の場合，標本不偏分散の値 $\frac{n}{n-1}s^2$ と標本分散 s^2 との比は，ほぼ 1 に等しいとしてもよい．この近似は n が 100 程度でも十分に成り立つ．

$$\frac{n}{n-1} \approx 1 \tag{7.11}$$

このときには，s^2 を母分散と同一してしまうこともできるので，数式の上では母分散が既知の場合と同じ扱いになる．

表 7.3 平均値推定のモデルの使い分け: λ はそれぞれの分布におけるパーセント点

	事前の知識	区間の幅	確率分布関数
大小標本	母分散既知	$\pm\lambda \times \frac{\sigma}{\sqrt{n}}$	正規分布
大標本	母分散未知	$\pm\lambda \times \frac{s}{\sqrt{n-1}}$	正規分布
小標本	母分散未知	$\pm\lambda \times \frac{s}{\sqrt{n-1}}$	t-分布

■3 つのモデルのつながり

ここまで述べてきた 3 通りの扱いについて，その関連を考えておきたい．

母集団から無作為抽出するデータの数が少ない場合，そこから得られる情報の信頼性が低いことはだれでも分かる．「クイズ 5 人に聞きました」という番組タイトルでは，ぜんぜん真実味を感じられないだろう．しかし，その場合でもデータから一定の情報は得たいということがしばしばある．標本平均が正規分布することが保証できないようなケースで，どうやったら定量的な信頼性のある結果を得ることができるのだろうか．

ビール会社で生産管理をしていたゴセットが突き当たったのはその問題だった．そこで彼は，母集団が正規分布している場合には，小標本から式 (6.10) で与えられる T を作ってやると，正規分布よりも信頼区間が広くなるような確率分布である t-分布が成立するこ

とを示して，小標本の統計的推定に道を開いた．

誤解してはいけないのだが，ゴセットがやったことは「少ないデータからでも信頼性の高い推定をできるようにした」というわけではない．むしろ「少ないデータから得られる結果は信頼性が低いけれど，低いなりの信頼性をきちんと明らかにした」と理解したほうがよい．

今，n 個，ただし n は数個程度のデータを，母集団からランダムサンプリングで得たとしよう．そこから計算して得られる標本分散

$$s^2 = \frac{1}{n}\left((x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2\right)$$

は，とても不安定な量であることに注意してほしい．つまり，標本平均 $\overline{X}$ は抽出のたびに大きく変動してしまう．ということは，標本平均を使って導かれる標本分散も，大きく変動することになる．

それでは，標本から得られる分散を真の平均である母平均 μ を使って次の式で定義される

$$\frac{1}{n}\left((x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2\right)$$

こういう「なんとか分散」を計算して使えばよいかというと，そもそも未知の μ を知りたくて推定の手続きを進めようとしているのだから，それが最初から使えれば苦労はない．まったくナンセンスなことになってしまう．

ただし，仮に母平均はわからないのに母分散だけがわかっているというううまい話があったとすると（すごいレアなケースだ！），そこを手がかりにして母平均を推定できる．

しかし普通はどちらの量も分かっていないだろう．そのときでも，もしも標本のサイズが大きければ，標本平均は正規分布することが期待でき，かつ標本不偏分散をもって母分散として計算することができる．

そしてゴセットは，標本のサイズが小さいときに，ただし母集団が正規分布している場合に，標本不偏分散と，t-分布という特別な分布を持ち込むことで，問題を解決した．

【章末問題】

問題 7–1　100 人の有権者を無作為に選び，今の内閣を支持しているかどうかを尋ねたところ，支持率は 30% であった．このときの真の内閣支持率を 95% の信頼区間で推定しなさい．

問題 7–2　母集団が正規分布しているとして，そこから 12 個のデータを無作為抽出したところ標本平均が 16, 標本標準偏差が 4 であった．このとき，母平均 μ について 99% 信頼区間を推定しなさい．

問題 7–3 問題 7–2 で，データの数が 80 であったとし，他の量は同一であったとする．
このとき，母平均 μ について 99% 信頼区間を推定しなさい．

第 8 章

仮説と検定

ある命題が確からしいかどうかを統計学的に判定することを**統計的検定 (statistical test)** または**仮説検定 (hypothetical test)** という．また，判定される命題を**仮説 (hypothesis)** という．統計的検定とは，ある仮説が妥当であるかそうでないかについて，一定の確率的根拠に基づいて判断するために行われる作業である．

8.1 ひょうたん島での仮説検定

どのような状況で仮説検定が行われるのか，その例を不朽の人形劇として今も名前の残る「ひょっこりひょうたん島」の舞台に託して示すことにしよう．断るまでもなく，ひょっこりひょうたん島は故井上ひさし氏の手になる名作テレビドラマであるが，ここで紹介しているのはまったくのパロディであって，オリジナルとは何の関係もない[*1]．

[*1] 実は，小波秀雄が大作家の故井上ひさしに弟子入りして文章道を教わり，小さな作品を激賞されたことがあるというのは，自慢したい事実である．

8.1.1　大統領の陰謀の巻

【ここまでのあらすじ】ひょうたん島の住民たちの間で，ある日次のような事件が発生した．身勝手で尻が軽くて口の達者なひょうたん島大統領ドンガバチョは，トラヒゲデパートの社長トラヒゲに商品の在庫処分を頼まれた．なにせ人口が少ない島のことだから，一度売れた商品は捌(さば)けなくなってしまうのである．そこはガバチョ，「ほいほい，ワタクシの舌先三寸で不良在庫の山なんか軽く一掃してあげますよ．その代わりといっては何ですが，売り上げの半分はワタクシめにくださるでしょうな？」と，トラヒゲの渋い顔を他処に，商品のサンプルの包装紙を別のものに取り替えて，ひょうたん島小学校の前で叩き売りを始めたのである．もちろんトラヒゲだって売り上げをガバチョに渡す気なんか毛頭ない．お互いに相手を利用して一儲けしようというコンビなのだ．

はじまり

ドンガバチョ：　みなさあーーーん，耳寄りの話でございますよ！わが社の最新式の扇風機，ソヨリン X は，従来の製品に比べて 1 時間当たりの電気代がぐぐぐぐっと少ないのでありまーす．ここで買わなきゃ損をすること間違いなし．さあ，さあ，お集まりのほどを！

ハカセ：　（「へんだなあ」と独白）ガバチョさん，だったら持ってきた製品を 10 個出してみて（ガバチョしぶしぶ出す）．さあ，みんな，試供品のソヨリン X をとりあえず 1 時間動かしてみて，エネルギー消費を測ってみてくれないかな．

トラヒゲ：　やめろやめろーい（ガバチョ「おやめなっしゃれー！」）．そんなことしたら俺のデパートの倉庫から持ってきたってことがばれちゃうじゃないかよ（と，言いかけてあわてて）っとっとっと，つまりだな，ようするに人間は正直が一番てことよ．

サンデー先生：　そうよ，トラヒゲさん．いいこと言うわねえ（トラヒゲ，でへへへと照れながら頭を掻く）．

サンデー先生と子どもたち：　♪「正直はすてきなこと」の歌．その間に時間がすぎる．

ハカセ：　（1 時間後）さて，みんなのレポートをまとめてみようか．ええとダンディさんのデータだと，エネルギー消費は 1250 ガンバだったんだね．サンデー先生のは 1305 ガンバと，・・・・ふむふむ，10 人のデータを平均すると 1273 ガンバ，標準偏差は 29.5 ガンバってことだな．・・・・これくらいは暗算でこのとおりスイスイさ，ボクにとってはネ．

ドンガバチョ：　おっほっほっほ，さっすが天才ですなあ，ハカセさん．もう結論が出ましたですね．わがニューソヨリン X はやはりすんばらしい性能でがしょう？

ハカセ：　まだ何にもきまってないじゃないよ，ガバチョさん．えっと，ボクの頭の中の

データベースによると，もともとトラヒゲデパートにあったソヨリンオリジナルだと時間当たりエネルギー消費は 1286 ガンバで，その標準偏差は 35.4 だったはずだ．

トラヒゲ： そーれ見ろよ．やっぱり少なくなってるじゃねえかよ，ハカセ．おいらの言ってることはまちがいないだろう？そうだよな．

サンデー先生： ハカセさん，ガバチョさんだって正直なときは正直なのよ．人を疑うのはよくないわ．1286 ガンバ が 1273 ガンバに減ったってことは，やっぱりソヨリン X って少しはいいんじゃないの？

ハカセ： いやいや，ものごとは疑ってかかることだって必要だと思うな．えっと，つまり，ガバチョさんが持ってきた 10 個の試供品はトラヒゲデパートにあったソヨリンオリジナルの包装だけ変えたんじゃないかってことは十分考えられる．それでも成績はオリジナルよりもよかったってことは，たまたままぐれで成績がよかったのかも知れない．その可能性がありうるかどうかを考えてみないとね．

ドンガバチョ，トラヒゲ： ギョギョギョギョッ！

来週に続く

8.1.2 ハカセは仮説を検定する

■ふたつの仮説

さて，以上のシナリオからみて，一体なにを検証できればガバチョとトラヒゲの悪だくみを看破することができるのだろうか．双方の言い分をもう一度みてみよう．

ドンガバチョが主張しているのは，「ソヨリン X はオリジナルから抜き取られたものではない」ということである．つまり，ソヨリン X のデータはオリジナルよりも優れているのであるから別物だというのである．つまり次の主張だ．

> **仮説 I：ソヨリン X という商品のエネルギー消費率の平均は，「本当に」オリジナルの平均よりも小さい．**

「本当に」というのは，たまたまそうなったのではなく，ソヨリン X がオリジナルとは異なる母集団に属していて，その母平均はオリジナルの母平均よりも小さいということを強調するために入れている．

なお，ここで**仮説 (hypothesis)** と言っているのは，上の主張がこれから検証されるべき命題であって，その真偽はまだニュートラルな状態にあることを意味している．

しかし，通常の統計的検定では次のように上と反対の仮説を設けることが多い．これはハカセの取っている立場である．

> **仮説 II：ソヨリン X という商品のエネルギー消費率の平均は，オリジナルの平均**

に等しい.

ここで注意しておかなければならないのは，ソヨリン X の 10 個のサンプルから計算される標本平均が，母平均，つまりオリジナルの平均と一致しているかどうかを問題にしているのではないということである（1273 ガンバと 1286 ガンバだから，実際一致していない）．標本平均というのは抽出の度に変化する確率変数なのだから，たまたま一致することはまずありえないのだ.

そうではなく，**ソヨリン X はオリジナルから抽出されたものであって，本来の平均はオリジナルのものと一致するといっても確率的にはかまわない**ということを，「等しい」と表現しているのである.

■帰無仮説と対立仮説

上の仮説のうち，仮説 II のようなものは，それが**棄却 (reject)** (= 否定) されたときに意味をもつことになる．つまり，棄却されれば「等しくない」ということになるのだが，それは実際に「ちがう」という積極的な意味を持っている．逆に仮説 II が**採択 (accept)** されたら，ソヨリン X がソヨリンオリジナルと同じ性能を持つからといって，これらが同じものであるとは言えないのだから，せいぜい「ちがうわけではない」という程度の意味しか持たせられない.

このように，棄却されたときに意味を持つような仮説を**帰無仮説 (null hypothesis)** という[*2]．**帰無仮説は棄却されるときに意味をもつ.**

一方，仮説 I のように帰無仮説と対立して，「差がある」という言明を含む仮説のことは，**対立仮説 (alternative hypothesis)** と呼ばれる．統計的検定では，帰無仮説のことを $\boldsymbol{H_0}$, 対立仮説のことを $\boldsymbol{H_1}$ と表記することがしばしばある[*3].

■仮説を検定する

それでは統計的処理をどういうふうに使ったら，これらの仮説の当否を**検定 (test)** できるのだろうか.

この問題では，母集団のソヨリンオリジナルの母平均 μ および母標準偏差 σ は，すでにハカセの頭の中のデータベースに入っていて既知である．もし，そこから $n = 10$ 個の標本を無作為抽出したとしたら，得られる標本平均 $\bar{X}$ の期待値は μ と等しくなり，また $\bar{X}$ の分散は σ^2/n に等しくなるはずである（94 ページの式 (6.3)(6.4) 参照）.

したがって，オリジナルから抽出された大きさ 10 の標本によって得られる標本平均は

[*2] 本によっては，比較対象の差がゼロになるような仮説を帰無仮説と説明しているものがある．これらは微妙に意味が異なるが，実際上は重なる概念である.

[*3] この辺の話はややこしくて，学習者にとっては頭が混乱するところである．しかし，後に見るように，分布の中での位置関係というイメージで考えれば，それほどややこしい概念操作は必要ではないことがわかる.

正規分布 $N[\mu, \sigma^2/n]$，具体的には $N[1286, 125.3]$ に従うことになるはずである[*4]．つまり，1286 ガンバを中心として，標準偏差が 11.2 $(= \sqrt{125.3})$ ガンバの正規分布が，オリジナルからの抽出で得られる大きさ 10 の標本の標本平均のとる分布である．**その分布の中で，今回の調査で得られた 1273 ガンバという値は，明白に小さい側に外れているのだろうか**．これが問題の中心なのである．外れていればドンガバチョの勝ち，そうでなければハカセの疑いには妥当性があるし，もちろんソヨリン X を買う意味はない．

なお，このように対象が分布の左右どちらかに外れているかどうかを見るような検定を**片側検定**あるいはもっと詳しく**左側検定**，**右側検定**と言う．それとちがって分布の中心から遠いかだけを判断するような検定は**両側検定**である．

■分布の中のどの位置にあるかで検定する

さて，ここまで来れば問題の解決は容易である．正規分布 $N[1286, 11.2^2]$ における 1273 という値を標準化変換すればよい．

$$\frac{1273 - 1286}{11.2} = -1.16$$

一方，標準正規分布では，-1.28 以下の部分に入る面積は 10% $(\alpha = 0.1)$ になるので，結局 1273 ガンバという値は，全体の 10% という狭い領域には入っていない．なお，図の影を付けた領域のことを**棄却域 (critical region)** といい，棄却域に入ることを「棄却域に落ちる」と表現する[*5]．ここではもちろん棄却域に落ちてはいない[*6]．

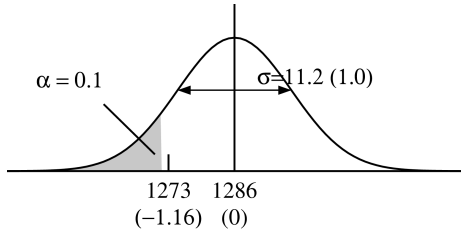


図 8.1　$N[1286, 11.2^2]$ の中での 1273 の位置：カッコ内は標準化された値．

この図から，ソヨリン X の標本平均はソヨリンオリジナルからの抽出標本のものとは**それほど外れていない**ということがわかる．10 回に一度程度おきる現象はさほど確率的にめずらしいものではないが，ドンガバチョが持ち込んだサンプルはその程度の**有意差**を見せなかったのである．このことを次のように表現する．

[*4] 125.3 というのは $35.4^2/10$ を計算すれば出てくる．

[*5] 対立仮説の側から考えると，棄却域に落ちるのは，対立仮説が棄却されないことを意味していて，考える上では厄介である．

[*6] 棄却域に対して，**採択域 (acceptance region)** もあるが，使われる頻度は低い．

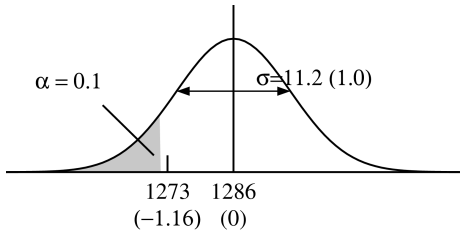


図 8.2　p 値の意味：$z = -1.16$ のデータよりも左の部分の面積 ($= p$ 値) は 0.12 であり，それほど「めずらしい」値ではない．(図 8.1 と同図)

「仮説 II：ソヨリン X という商品のエネルギー消費率の平均は，オリジナルの平均に等しい」は，危険率を 0.1 と取ったときには棄却できない．

ここで使われている**危険率 (risk)** という言葉の意味は，むしろ仮説が棄却されたときのほうがわかりやすい．つまり，「仮説 A は危険率 0.01 で棄却される」は，「仮説 A を棄てても，外れる危険 (リスク) は 0.01 である」ということなのである．逆に棄却できないときには，それ以上の確率で外れるかも知れないから，何も言わないでおこうということになる．

8.1.3　p 値

上の操作でソヨリン X の標本平均を標準化変換して得られた値 $z = -1.16$ は，どの程度の「めずらしさ」をもっているのだろうか．正規分布表で調べると，この値は端から 0.12 の面積のところに位置することがわかる．このように，検定されるべき値が，分布の端からどれだけの面積のところに位置するかを ***p* 値**という名前で表すことにする（図 8.2)．

仮説検定においては，p 値だけを用いて，「この観測値の p 値は 0.015 だから有意に異なると判断できる」という言い方もする．p 値は，いわばデータの「めずらしさの度合い」であるから，一般に値が小さいほど有意であるという判断が下されることになる．

8.1.4　別のシナリオ—贋物を検定する

さて，ひょうたん島でのトラヒゲとドンガバチョの悪だくみは，オリジナルそのものを別の新しい製品であるかのように見せかけようということであった．その意図はハカセによる統計的検定のおかげで見事に見抜かれて，不良在庫を売りつけられる事態は免れたのである．

ところが，世の中いくらでも悪いやつはいるものであるし，悪いやつらには必ず悪事のチャンスが到来するものである．わがひょうたん島を根城にしている海賊たち，ガラクー

タ，ヤッホー，トウヘンボク，それにキッド坊やの 4 人は，ある日孤島の洞窟からニセ金貨の詰まった箱を発見した．これがなんと！ひょうたん島で発行されている 1000 ガバス金貨とそっくりだったのだ．しかもニセ金貨は全部で 25 枚，25000 ガバスといえばトラヒゲデパートの年間売り上げを上回る大金である．

これを海賊たちが見逃すはずがない．さっそくニセ金貨を両替するためにトラヒゲ銀行ひょうたん島本店に持ち込んだ．

しかし，さすがはかつて自分も海賊だった頭取のトラヒゲ — その昔に海賊大学校を落第したことでガラクータたちにひそかにコンプレックスを抱いているのだが — は，海賊たちが突然大金を持ち込んだことに不審をいだいて，持ち込まれた金貨を鑑定してもらうためにハカセを呼んだ．

もちろんハカセの頭脳のデータベースには，ひょうたん島の 1000 ガバス金貨の重量の平均値と標準偏差の数値なんかしっかり入っている．二つ返事で承知すると，持ち込まれた 25 枚の金貨を秤で精密に量ることにした．

ここで再び物語のシナリオが再開される．今度の主な登場人物は銀行頭取のトラヒゲ，海賊の首領ガラクータと子分のトウヘンボクとヤッホー，キャプテンキッドの息子キッド坊や[*7]，そしていうまでもなく天才の名をほしいままにしてきたハカセである．

またはじまり

ガラクータ　（丁寧で上品な女言葉ふうの，いかにもインテリヤクザといった口の利き方がこの海賊の首領の特徴である）ハカセさん，私たちが持ってきたこの金貨はニセモノじゃないかって，疑っておいでのようネ？

ハカセ　もちろん疑ってるヨ（海賊たち「ナニッ！」と気色ばむ）．おっとそんなに焦らないで，海賊さんたち．どんなことでもうたぐってかかるというのが科学者というものなんだからね．ボクとしては，データに基づいて客観的に判定するだけだから，海賊さんたちが本当にホントのことを言っているんだったら，何も心配することはないと思うよ (海賊たち，不安そうに目を見合わせる)．

トラヒゲ　やいハカセ，オレ様はだな，銀行頭取としてだな，ハカセにこの金貨がニセモノかどうかを鑑定してもらいたいんだよ．トラヒゲ銀行が仮にも損をしないようにやってくれなくちゃ困るわけだ．だから頼むからサァ，海賊の野郎どもに遠慮なんかしないでやってくれヨォ！

ハカセ　誤解がないように言っておきたいんだけど，ボクがこれからやろうとしているのはとっても簡単なことなんだ．つまり，海賊さんたちが持ってきた金貨が本物であるという主張，えっとむつかしくいえば仮説というんだけど，その仮説を否定して

[*7] 伝説の海賊であるキャプテンキッドの息子が現代に登場するのはおかしいと思うかもしれないが，キャプテンの残した遺産の箱の中から，数百年の眠りを覚まして飛び出してきたのである．

も大丈夫かどうかという点を，データを使って確かめようというのさ．

ヤッホー　（自分の頭の上でくるくる手を回しながら）ひゃー，ハカセの言ってるのを聞いていたら，頭の中がぐるぐる回って，そんでもってなんかがパチンとはじけたみたいだなあ，おいトウヘンボク，おいらの頭どうなってるか見てくれよ．

トウヘンボク　どれどれ，わあ，ヤッホーの髪の毛のぐるぐるはちょっとおかしいあるネ．右に巻いたり左に巻いたり，わたし，こんなのは昔，動物園のお猿さんの頭で見たことあるヨ（ヤッホー怒ってトウヘンボクをどつく）．あ，私も分らないあるから勘弁してヨ！

ガラクータ　静かになさい！ハカセさんの言っていることがあなたたちに分るはずなんかないじゃないの．ハカセさんはネ，つまり，あなたたちのお頭（つむ）では死んでもわからないぐらいむつかしいことを言ってるのよ．ワタシにだって分らないんだから（一同ずっこける）・・・・でもハカセさん，あなたのおっしゃることってすごくなんだか，えっと，その，あのォ，とっても回りくどーく聞こえるんだけど？

ハカセ　さすがガラクータさん，いいとこに気がついたね．その通り確かに回りくどいんだ．そもそもボクにできることは限られていて，秤で重さを精密に調べるぐらいのことしかできない．そのデータから，「『海賊さんたちの持ってきた金貨が本物である』ということを否定できるかどうか」だけを判断できる．否定できれば海賊さんたちの言っていることは多分ウソだろうということになるし，そうでなければ，ウソだとは言えない — 本当だとも言えないけどね．

一同　ふーん．おいら分んない．へえー，そうなのか，なるほどォ．分ったふりなんかするんじゃないよ！（などと口々に騒ぐ）

ハカセ　ちょっと失礼するよ（天秤で 25 枚の金貨の秤量を始める．♪トラヒゲと海賊「お金を量ってどうなるの？」）．さあ，できた．25 個の金貨の重さはこんな結果になった．

32.97, 36.37, 35.24, 36.03, 34.84, 33.63, 37.94, 33.48, 34.09, 33.74, 34.53, 36.86, 31.79, 35.61, 34.14, 34.51, 35.13, 32.83, 34.89, 32.19, 36.67, 36.01, 37.04, 35.1, 33.73 (単位はポンズ)

ハカセ　一方で，ボクのデータベースによると，ひょうたん島の 1000 ガバス金貨の重さは，平均が 35.03 ポンズ，標準偏差が 0.925 ポンズだ．これでデータは出揃ったぞ．

一同　（顔を見合わせて） さあ，どうなるんだろう？

ハカセ　さて，平均値の方をまず検定してみよう．おっと！これは海賊さんたちの言うことは否定できないかも知れないぞ．ちゃんと計算すると・・・データの平均はええと，34.77 ポンズ．なるほど，これだとこの金貨が本物だって話を否定することはできないな．

海賊たち　(ヤッホー，トウヘンボク) わーい，金貨は本物だってよオ，ハカセが言ってる

ぞーい.（ガラクータ）ハカセさんってやっぱり天才だわねェ.

トラヒゲ （やや憮然として）なんてこった，金貨は本物かも知れないってエのかよ，ハカセ？おいらにゃどうもうさん臭い気がして仕方ないんだけどヨオ.

ハカセ さっき言った通りさ．本物であるという海賊さんたちの話を否定するデータにはなってないんだよ．・・・ ふむふむ，もう少し別の角度からデータを検定するとどうなるんだ？

この後の原稿紛失

8.1.5 平均値の一致を検証する

ハカセがニセ金貨の重量の平均値を得たときに検定しようとした仮説は次の通りである．これは海賊たちの立場に立った主張になっていることに注意してほしい．それを棄却できるかどうかを判断するのが，検定の目的なのである.

仮説：持ってきた 25 枚の金貨の重量の「本来の」平均値は本物の金貨の平均値と等しい.

ここで「本来の」平均値としているのは，25 枚の秤量で得られた標本平均のことではない．海賊たちとしては，持ち込んだ金貨 25 枚が本物の金貨を母集団として抽出されたのだと主張しているのである．この構図は，ソヨリン X のときとはちょうど反対であることに注意しよう.

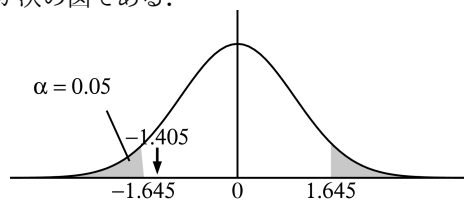
そして，実際に得られた標本平均 $\overline{X}$ はというと，

$$\overline{X} = \frac{32.97 + 36.37 + 35.24 + \cdots + 33.73}{25} = 34.77$$

ソヨリン X の議論のときと同じく，海賊の金貨がひょうたん島の金貨から抽出されたものであるならば，その平均 $\overline{X}$ は分散 $s^2 = \sigma^2/25$(標準誤差 $s = \sigma/\sqrt{n} = 0.185$) をもち，$\overline{X}$ の期待値（「本来」の平均）35.03 にピークを持つ正規分布をなすはずである．これらを使って，得られた平均値 34.77 を標準化変換してみよう.

$$Z = \frac{34.77 - 35.03}{0.185} = -1.405$$

Z の位置を表したのが次の図である.



海賊の金貨の平均値 Z

図を見れば一目瞭然，海賊の持ってきた金貨の質量の平均値は，本物の金貨を 25 個持ってくればあり得たであろう分布の中に入っていて[*8]，これだけの証拠では，**本物ではないということはできない**．このことを次のように表現する．

上の仮説は，危険率 0.1 で棄却されない．

なお，この場合の棄却域，つまり図の斜線をほどこした部分は，両側にとられていることに注意してほしい．これは検定の目的が本物と異なっているかどうかを問題にしているからである．このように，両側に棄却域をとる検定を**両側検定**と呼ぶ．

一方，ソヨリン X の検定のように，サンプルの値が比較対象よりも優れて（劣って）いるかどうかを検定する場合には片側に棄却域がとられるので，**片側検定**と呼ぶ．

8.1.6　分散の一致を検証する—χ^2 検定

さてハカセがつぶやいた「もう少し別の角度」とは何だろうか．失われた原稿に書かれてあったシナリオを考えてみよう．

ハカセは 25 個のデータを使って平均値を計算し，それと本物の平均値および分散を利用して検定を行った．しかし，25 個のデータから得られる量は平均だけではない．分散もそうである．あるいはもっと別の量だって計算できる．

さて，分散とその類似の量がどのような分布をなすかということは，すでに第 6 章 6.5 節で詳しく取り扱った．そこではまず，次のことが示されている (p.103)．

【定理 I】母集団は平均値 μ，分散 σ^2 をもつ正規分布 $N[\mu, \sigma^2]$ をしているとする．そこから n 個の無作為抽出を行ったとき，

$$Z = \frac{1}{\sigma^2}\left((X_1-\mu)^2 + (X_2-\mu)^2 + \cdots + (X_n-\mu)^2\right) \tag{8.1}$$

なる Z は，自由度 n の χ^2 分布 に従う．

また，標本平均 $\overline{X}$，標本分散 s^2 を使って次のような定理も述べられている (p.104)．

【定理 II】母集団は平均値 μ，分散 σ^2 をもつ正規分布 $N[\mu, \sigma^2]$ をしているとする．そこから n 個の無作為抽出を行ったとき，

$$Z = \frac{1}{\sigma^2}\left((X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \cdots + (X_n-\overline{X})^2\right) = \frac{ns^2}{\sigma^2} \tag{8.2}$$

[*8] 図の中で，この平均値はかなりきわどいところにあって，それを暗算で判定したハカセの頭脳はさすがに大したものなのである．

なる Z は，自由度 $n-1$ の χ^2 分布 に従う．ただし，ここで s^2 は式 (6.2) で表される標本分散である．

■母平均と母分散を使って検定する

最初に式 (8.1) を使った検定を試みることにしよう．それにはこの式に現れる Z を，データを使って求める必要がある．ハカセのデータベースによれば，母集団の平均 μ は 35.03 ポンズ，標準偏差 σ は 0.925 ポンズである．なお，母集団といっても，海賊がもってきた 25 枚の金貨がそこから抽出されたわけではないが，仮説としては「25 枚の金貨は本物である」として検定するのだから，ここでは仮にニセ金貨の母集団を本物の 1000 ガバス金貨としているのである．また金貨の質量の数値は，それぞれ 32.97, 36.37, ... , 33.73 となっているから（つまり $X_1 = 32.97, X_2 = 36.37, ...$），$Z$ は次のようにして求めることができる．

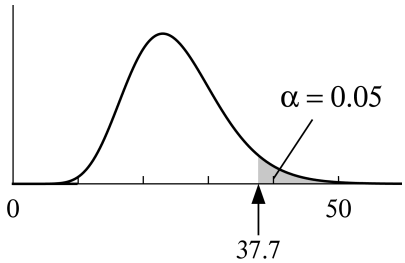
$$Z = \frac{1}{0.925^2}\left((32.97-35.03)^2 + (36.37-35.03)^2 + \cdots + (33.73-35.03)^2\right) = 70.49$$

この種の計算は面倒なので，表計算ソフトをうまく使うとよい．

定理 I によれば，Z は自由度が 25 の χ^2 分布に従う．つまり，ひょうたん島の本物の金貨から 25 個の抽出を何度も行って，上の手順で Z を求めると，自由度 25 の χ^2 分布に従ってあちこちに値が散らばるわけだ．さて問題は，**海賊の持ってきた金貨のデータから得られた 70.49 という値は，その分布の中に現れる可能性がどの程度あるのか**ということである．それを確かめるには χ^2 分布表を見ればよい．

χ^2 分布表からの抜粋

α	0.995	0.975	0.95	0.9	0.5	0.05	0.025	0.01	0.005
$\nu = 23$	9.260	11.688	13.090	14.848	22.337	35.172	38.076	41.638	44.181
$\nu = 24$	9.886	12.401	13.848	15.659	23.337	36.415	39.364	42.980	45.558
$\nu = 25$	10.520	13.120	14.611	16.473	24.336	37.652	40.646	44.314	46.928
$\nu = 26$	11.160	13.844	15.379	17.292	25.336	38.885	41.923	45.642	48.290



$\nu = 25$ のときの χ^2 分布と 5% 点

上の図と表で分るとおり，70.49 という値は 5% 点どころか 0.5% 点をとっても棄却域に落ちてしまい，この分布ではほとんど起こりえない値であることが分る．つまり，χ^2 を使った検定によって，次の結論が得られる．

海賊の持ってきた金貨が本物であるという仮説は危険率 0.5% で棄却される．

要するに，安心してニセモノと決め付けてよいということだ．

■標本分散を使って検定する

さて次に，**定理 II** のほうを使った検定を試みよう．定理 II では，25 枚の金貨の重さから得られる標本分散と標本平均を利用することができ，計算もやや簡単である．

まず標本平均は前に求めたように $\overline{X} = 34.77$ であり，標本分散は 2 乗の平均から平均の 2 乗を引けば得られて

$$s^2 = \frac{32.97^2 + 36.37^2 + \cdots + 33.73^2}{25} - 34.77^2 = 2.347$$

となる．表計算ソフトを使って計算してもよい．これから Z を求めると，

$$Z = \frac{ns^2}{\sigma^2} = \frac{25 \times 2.347}{0.925^2} = 68.57$$

が得られる．この値を今度は自由度が $n - 1 = 24$ の χ^2 分布を使って検定すると，やはり海賊が持ってきた金貨が本物であるという仮説は非常に低い危険率で棄却される．ようするに，海賊たちが持ってきた金貨は，平均の質量こそ本物と見分けがつかないほどに似ていたのだが，作りが悪いために個々の重量がまちまちだったのである．

8.1.7　金貨のできが良過ぎても怪しい—両側検定

ここまでの設定では，海賊たちが持ってきた金貨は質が悪くて，本物に比べて質量のばらつきが大きすぎたところを検定で発見したケースになっている．しかし，悪い奴らがもっとハイテクを駆使できる連中だとしたら，逆のことが起きるかもしれない．次の例題を考えてみよう．

例題 8–1　ひょうたん島銀行に，りゅうとした身なりの紳士が現れて，12 枚の金貨の両替を依頼した．トラヒゲ社長は天性の勘で「こいつは怪しい！」と，紳士を別室で待たせている間に金貨の質量を量って，ハカセに鑑定してもらうことにした．データは次のようになっている（単位 ポンズ）．

$$35.7, 35.03, 35.11, 34.21, 35.08, 34.86, 35.13, 35.09, 34.36, 35.23, 35.24, 35.67$$

ひょうたん銀行発行の本物の金貨の平均質量は 35.03 ポンズ, 標準偏差は 0.925 ポンズである．持ち込まれた金貨は本物であると判定してよいか．

上のデータから定理 I の Z を計算してみよう．

$$Z = \frac{1}{0.925^2}((35.7 - 35.03)^2 + (35.03 - 35.03)^2 + \cdots + (35.67 - 35.03)^2) = 2.451$$

この結果を自由度 12 の χ^2 分布とにらみ合わせて考えることにする．表を見ると，$\alpha = 0.995, 0.005$ の点がそれぞれ 3.074, 28.300 となっている．これらは両端にそれぞれ 0.5% の面積を切り取る Z の値で，データから求めた 2.451 という値はその外にある．ただし前の問題とちがって，大きい方にではなく小さい方に外れている．

つまり，このケースでは，持ち込まれた金貨が本物から抽出されたものであるという仮説は，危険率を 1% にとった場合に棄却されるということになる．つまり贋金だと判定できる．もっとも，分布の左側に外れたということは，分布が狭い範囲にまとまり過ぎているということを意味する．ニセ金貨を作った連中はあまりにも正確に作ってしまったというわけだ．このように，χ^2-分布で**両側検定**を行うことで，意味のある検定結果を得ることができる．

8.2　その他の検定

8.2.1　小標本について平均値を検定する — t-検定

ひょうたん島のソヨリン X 売り込み騒動，ニセ金貨両替詐欺事件では，平均値の検定が使われた（後者では分散の χ^2 検定が絶大な威力を発揮したが）．その場合，母集団の平均値や分散が精密に分っているという条件があって，検定は容易に行われた．

しかし，実は，ソヨリン X のサンプル 10 個を手にしたとき，ハカセの脳のデータベースにはオリジナルのソヨリンのデータは平均値しかなかったとしよう．つまり母分散は知られていないのである．この場合には 8.1.1 節以降のシナリオは変更を余儀なくされるが，7.3.3 節で出てきた t-分布を利用することで検定を行うことができる．

今，正規分布している母集団から抽出された大きさ n の標本があったとして，母平均 μ, 標本平均 $\overline{X}$, 標本標準偏差 s を使って与えられる T という量を次のように求めたとしよう．

$$T = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s}$$

このとき，T は自由度が $n-1$ の t-分布に従う．

さて，ここではすでに論じたように，

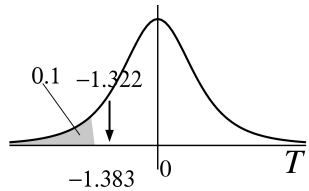
「10 個のソヨリン X の値の平均は，ソヨリンオリジナルのデータの平均と等しい」という帰無仮説が棄却されるかどうかを検定で明らかにしたいのである．そこで，与えられたデータから T を計算してみよう．

$$T = \frac{\sqrt{10-1}(1273 - 1286)}{29.5} = -1.322$$

例によって，この値が分布のどのへんにあるかを表から確かめよう．

α	0.1	0.05	0.025	0.01	0.005	0.0025
$\nu = 8$	1.397	1.860	2.306	2.896	3.355	3.833
$\nu = 9$	1.383	1.833	2.262	2.821	3.250	3.690
$\nu = 10$	1.372	1.812	2.228	2.764	3.169	3.581

採用するのは自由度 $\nu = 9$ のところの値である．また検定の意図は「ソヨリン X のエネルギー消費率がソヨリンオリジナルよりも小さいかどうか」の確認にあるのだから，分布で左端の面積 0.1 の領域を棄却域として設定する．



図を見れば，この場合にも上の仮説は棄却されない．つまり，有意な差があるとはいえないという結論になる．

8.2.2　分割表による検定

χ^2 分布を応用して，ある集団がもついくつかの性質が独立であるかどうかを判定することができる．その種の問題としては，たとえば次のようなものがある．

1. ある集団について，生活習慣と健康状態のアンケートを調べて，それらが独立であるか，あるいは関係があるかどうかを判断する．
2. 読書傾向と好きな色に関する調査結果から，それらの間に相関があるといえるかどうかを判断する．

これは単に集計表を見ただけでもある程度の判断はできるものである．しかし，一見して傾向がありそうなデータであっても，実際には傾向がないのに偶然そういうことがあるという程度のものなのか，本当に傾向があるのかという判断を定量的に行うことは難しい．また集計表の数字を定性的に判断しようとすると，先入観や主観的な期待によって判断に偏りが生じることがしばしばある．そこで客観的にこれらの判断を下すためには，以下に説明する**独立性の χ^2 検定**を行うとよい．

なお，χ^2 分布による検定はきわめて応用が広く，ここで取り上げた以外にもさまざまのタイプの検定がある．詳しくは他の書物を参照してほしい．

■分割表

アンケート調査によって表 8.1 のような表を作成することがしばしばある．

このような表を**分割表 (contingency table)** [*9] という．上の分割表の個別の欄に書き込まれた 39,45,21,83, ... といった数値は，**観測度数 (observed frequency)** という．

表 8.1　分割表の例

好み	赤	青	緑	計
クラシック	39	45	21	105
邦楽ポップス	83	68	47	198
洋楽ポップス	53	51	65	169
歌謡曲	41	32	55	128
計	216	196	188	600

[*9] contingency の形容詞は contingent で，「〜に依存して」，あるいは「偶発的な」という意味をもつ．分割表よりも，こちらの contingency table の方がずっと意味のある用語だ．

たとえばこの分割表を見ると，音楽の好みと色の好みに関連があるかどうかを知ることができるかも知れない．たとえばクラシックの愛好者はどちらかというと青を好み，邦楽ポップスの愛好者は赤を好むようにも見える．しかし，このような関連が本当に存在するのか，通常のデータの揺らぎの範囲なのかは，見ただけではなかなか分らない．

このように，2 つの性質によってグループ分けされた分割表から，異なる性質同士（音楽と色の好み）に関連があるかどうかを知りたいときには，独立性の検定という手続きを行う．

■独立性の検定

ある母集団から大きさ N の標本を抽出して，性質 A, B でそれぞれ m 個，n 個ずつに分類される．その結果として次のような分割表が与えられたとしよう．

性質	$B_1,\ B_2,\ B_3,\ \ldots, B_n$	計
A_1	$x_{11}, x_{12}, x_{13}, \cdots,\ x_{1n}$	a_1
A_2	$x_{21}, x_{22}, x_{23}, \cdots,\ x_{2n}$	a_2
$\cdots$	$\cdots, \cdots, \cdots, \cdots, \cdots$	$\cdots$
A_m	$x_{m1}, x_{m2}, x_{m3}, \cdots, x_{mn}$	a_m
計	$b_1,\ b_2,\ b_3,\ \cdots,\ b_n$	N

このとき，次の命題が成り立つことが知られている．すなわち，**性質 $\boldsymbol{A, B}$ が独立であるならば，次の $\boldsymbol{X}$ は，自由度が $\boldsymbol{(m-1)(n-1)}$ の $\boldsymbol{\chi^2}$ 分布に従う．**

$$
\begin{aligned}
X = & \\
& \frac{(x_{11}-a_1b_1/N)^2}{a_1b_1/N} + \frac{(x_{21}-a_2b_1/N)^2}{a_2b_1/N} + \cdots + \frac{(x_{m1}-a_mb_1/N)^2}{a_mb_1/N} + \\
& \frac{(x_{12}-a_1b_2/N)^2}{a_1b_2/N} + \frac{(x_{22}-a_2b_2/N)^2}{a_2b_2/N} + \cdots + \frac{(x_{m2}-a_mb_2/N)^2}{a_mb_2/N} + \\
& \cdots + \\
& \frac{(x_{1n}-a_1b_n/N)^2}{a_1b_n/N} + \frac{(x_{2n}-a_2b_n/N)^2}{a_2b_n/N} + \cdots + \frac{(x_{mn}-a_mb_n/N)^2}{a_mb_n/N}
\end{aligned}
\tag{8.3}
$$

式 8.3 はいかにも複雑に見えるが，実際にはきわめて単純で機械的な計算を意味している．表 8.1 について，X を計算してみよう．この場合には，$N = 600, m = 4, n = 3, a_1 = 105, a_2 = 198, \ldots, b_1 = 216, b_2 = 196, x_{11} = 39, x_{12} = 45, \cdots, x_{43} = 55$ であ

る．従って，

$$X = \frac{(39-105\times216/600)^2}{105\times216/600} + \frac{(83-198\times216/600)^2}{198\times216/600} + \frac{(53-169\times216/600)^2}{169\times216/600} + \frac{(41-128\times216/600)^2}{128\times216/600} + \cdots + \frac{(55-128\times188/600)^2}{128\times188/600} = 25.9$$

として計算される X は，もし A, B が独立であれば自由度が 6 $(=(4-1)\times(3-1))$ の χ^2 分布に従うから，高い確率で分布曲線の内側に入ってくるはずである．そこで，**もしも X の値が χ^2 分布の曲線の外側の方にはみ出しているならば，A と B が独立であるという仮説は棄却される．**

ただし，「はみ出し」というのは何に対してのはみ出しであるのか，その基準を設けておかないといけない．それが**危険率**である．では，今求められている X の値, 25.9 は，「はみ出し」ているのだろうか．

図 8.3 の χ^2 分布のグラフを見ながら考えてみよう．グラフの陰をつけた部分の面積が α であり，ここでは $\alpha = 0.05$ になるときの X の値は 12.59 であることが示されている (5% 点は 12.59 である)．この値はもちろん χ^2 分布表から拾ったものだ．この値を使うのが，**危険率 5% で検定する**ということである．

この問題の分割表から得られた X の値，25.9 は，あきらかにこの 5% 点の外側に外れている．従って, **音楽の好みと色の好みが独立であるという仮説は危険率 5% で棄却される．**[*10]

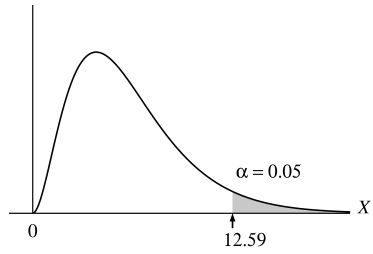


図 8.3　$n = 6$ に対する χ^2 分布関数

このように，分割表から X を求めて，それを χ^2 分布の表と見比べることで，2 つの性質が独立であるかどうかを検定することができる．

【章末問題】

[*10] 老婆心ながら誤解を避けるために付け加えると，この結論はでっち上げデータに基づくものであって，実際的な意味はない．

表 8.2　χ^2 分布の表

α	0.995	0.975	0.950	0.900	0.500	0.05	0.025	0.01	0.005
$\nu = 6$	0.676	1.24	1.64	2.20	5.35	12.59	14.45	16.81	18.55
$\nu = 9$	1.73	2.70	3.33	4.17	8.34	16.92	19.02	21.67	23.59
$\nu = 10$	2.16	3.25	3.94	4.87	9.34	18.31	20.48	23.21	25.19
$\nu = 11$	2.60	3.82	4.57	5.58	10.34	19.68	21.92	24.73	26.76

問題 8–1　トウモロコシの中には黄色と白の 2 色の実が混ざってついているものがある．これはバイカラーと呼ばれ，優性の黄色の実の純系品種と劣性の白い実の純系品種の 1 代交配種であるために，その子（純系品種から見たら孫）である種子がもつ 2 個の遺伝子の黄色，白の組み合わせによって色の違いが生じたものである．この場合，最初の品種が純系であれば黄色と白の実の数の比の期待値はメンデルの法則に従って 3 : 1 となるが，そうでないと，たとえば花粉が別の品種の花から飛んできたものだったりすると，比率の期待値は変化することになる．

いま，あるトウモロコシの実を調べたところ，144 粒の実のうち黄色の実が 119 個，白が 25 であった．この実が純系種の 1 代交配種から作られたという仮説は棄却できるか．危険率を 5% と 1% にとって検定しなさい．

問題 8–2　遺跡から貝殻が 12 個出土したので，その質量のデータを使って現存種の貝と同種のものであるかどうかを検定したい．データは次のようになっている（単位は g）．

11.78, 12.92, 7.55, 14.52, 12.05, 19.0, 11.29, 11.81, 15.38, 9.62, 14.19, 12.62

これと比較したい現存種の貝の質量の平均値は 12.6 g，標準偏差は 1.9 g である．この母集団は正規分布しているものとする．

1. 平均質量の分布を考えて検定を行ったとき，「この貝殻は現存種のものである」という仮説を危険度 5% で棄却できるか．
2. 式 (8.2) で定義される Z を使って χ^2 検定を行ったとき，「この貝殻は現存種のものである」という仮説を危険度 5% で棄却できるか．

問題 8–3　次の表は，ある疾病にその人がかかっているか，また飲酒の習慣があるかということをたずねたアンケートに基づいて作成した分割表である．この結果から，飲酒とその病気にかかることとは独立であるという仮説を危険率 1% で検定しなさい．

	罹患者	正常	計
習慣あり	37	37	74
習慣なし	93	133	226
計	130	170	300

第 9 章

相関と線形回帰

得られたデータから何らかの傾向を見つけ出すことは，統計的なデータ処理における重要な目的のひとつである．現実には散布図を見ただけでも分かるような傾向もあれば，判定に困る微妙なケースも少なくない．そのようなときに客観性をもった判断を行うにはどうしたらよいのだろうか．本章では，そのことについて考えてみる．

9.1　データの相関

一般に背が高い人ほど体重は重い．また，所得が多い人ほど，その人が住む住宅の面積も概して大きいにちがいない．一方，次のような例も考えられる． — 世界の国や地域を比較した場合，一人当たりの台所洗剤の使用量が多い地域ほどガンによる死亡率は高い．— このような事例は，自然現象あるいは人間や社会に関わる現象の中にいくらでも探し出すことができる．

そして以上の例のように，2 種類の数値データが互いに何らかの関連をもっている場合，データの間に**相関 (correlation)** があるという．本章では数値データの相関について統計的に分析する手法を扱う．

9.1.1　相関と因果関係

冒頭に挙げた例のうち，身長と体重の相関や所得と住居面積の相関は，直接的な原因と結果の関係から導かれる．なぜなら，身長が伸びるときには体が大きくなるのだから，当然体重も増加するはずだし，お金持ちは家に大きな費用を支出できるからこそ，広い住居に住むことになる．

しかし，地域の間の比較で，中性洗剤の使用量とガンの発生率について単純な原因と結果があるということはできない．このような傾向が見られることは，むしろ次のように解釈するのが適切である． — 一般に低開発国では中性洗剤の利用は少ない．工業製品が所

得水準にくらべて割高で，低所得の住民には手がとどかないからである．一方，低開発国ほど平均寿命は短く，高齢者の割合が先進国に比べて少ないのである．そのため人口比でみたガンの発生率は低い[*1]．したがって，中性洗剤の使用量とガンの発生率の間に相関が現われることはあっても，それは中性洗剤がガンを引き起こしているというわけではないのである．

2 つの現象のうち一方が他方の結果になっているとき，これらの間には**因果関係 (causality)** があるという．2 つのデータの間に因果関係があれば，なんらかの相関が表れると考えてよい．ただし，その他の要因が加わることで相関が分かりにくくなることもある．

因果関係があれば相関が生じることが多い

複数のデータの間に相関があれば，なんらかの因果関係の存在が示唆される．しかし，相関があっても因果関係は存在しないことも珍しくない．上のガンと洗剤のケースなども，容易に「中性洗剤はガンの原因になる」などというインチキな話にすりかえることができる．このように相関関係を因果関係に結びつける論法は，統計で人をだます誤った推論の代表的なものである．

相関があれば因果関係があるかも知れない．
しかし，必ずしもそうとは限らない

9.1.2　散布図 — 相関をグラフで見る

2 種類のデータの相関を見るために，横軸と縦軸を使ってデータをプロットしたものを**散布図 (scatter diagram)** という．散布図を使うと直観的に相関の有無を判断することができる．

図 9.1 に示された例は，$\boldsymbol{X}, \boldsymbol{Y}$2 系列のデータの間に相関がないものから，完全な相関をもつものまで，4 つの場合について散布図を示した．この図にあるように，**正の相関**はデータ X が増加するときに Y も増加するような相関，**負の相関**はデータ X が増加するときに Y が減少するような相関である．

図 9.1(2) のように弱い相関が見られる場合というのは，それが何らかの必然性によるものであるかどうかの判断は困難である．つまり特に理由がなくても，データのばらつきによって**偽の相関**が生じる確率は無視できない．この問題については後に詳しく議論す

[*1] WHO の資料で見ると，先進国においてはガンが主要な死因のひとつであり，その死者の割合は死者全体の 12.7% を占めるのに対し，低開発国においては，感染症や出産時の死亡，エイズなどが主要な死因を占めていて，ガンによる死者の割合はきわめて低い (2005 年のデータ). `http://www.who.int/mediacentre/factsheets/fs310/en/`

る．しかし (3)(4) のように，明らかに傾向が読み取れるようなケースでは，なんらかの必然的な原因があって，相関が現われていると考えてよい．ただし，繰り返しになるが，この場合でも因果関係があると即断してはいけない．

これらの図の欄外に示されている ρ_{xy} という数値は**相関係数**と呼ばれるもので，相関の強さを表す重要な統計量である．これについては次節で詳しく解説する．

なお，ここでは 2 系列のデータの間に隠れている相関は直線的な関係，つまり 1 次関数で表されるような単純なものであると仮定する．このような相関を 1 次または**線形の相関**という．この章では主に線形の相関について考えることにする．

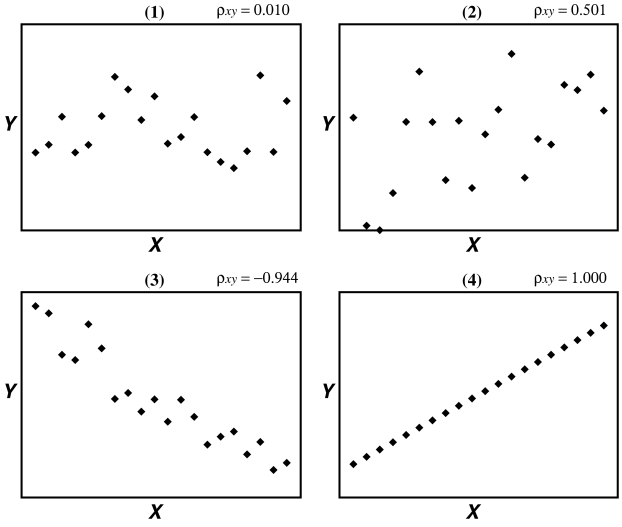


図 9.1　散布図で示された相関の有無: **(1)** 相関なし **(2)** 弱い正の相関 **(3)** 強い負の相関 **(4)** 完全な正の相関

9.2　相関係数と線形回帰

9.2.1　共分散

たとえば国ごとの人口密度と出生率とか，人ひとりずつの喫煙年数とガンによる死亡率，各自治体の教育予算が予算全体に占める割合と大学進学率のように，サンプルごとに 2 つの変数を知ることができるものとする．これらのデータを次のように 2 つの変数列 $\boldsymbol{x}$, $\boldsymbol{y}$ で表すことにする．

$$\boldsymbol{x} = x_1, x_2, x_3, \ldots, x_n$$
$$\boldsymbol{y} = y_1, y_2, y_3, \ldots, y_n$$

さて，次の式 (9.1) の定義で表される σ_{xy} は，$\boldsymbol{x}, \boldsymbol{y}$ の**共分散 (covariance)** と呼ばれる．

$$\begin{aligned} \sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n} \delta x_i \delta y_i \\ &= \frac{1}{n}(\delta x_1 \delta y_1 + \delta x_2 \delta y_2 + \ldots + \delta x_n \delta y_n) \end{aligned} \tag{9.1}$$

ここで δx_i は x_i の偏差 $x_i - \overline{x}$ を意味し，他も同様である (p.6)．

式 (9.1) は次の形をしていることに注目しよう．

共分散 ＝ （x_i の偏差 × y_i の偏差）の平均

共分散の定義は，1 章で出てきた p.7 の分散の定義を拡張したものであり，式 (9.1) で $x_i = y_i$ と置けば，そのまま分散の式が得られる．

例題 9–1　共分散の定義の式で y_i を x_i に置き換えると，分散の定義の式になることを確かめなさい (解答は省略)．

なお，共分散 σ_{xy} は次のように変形され，計算に当たってはこちらのほうが使いやすい．この形は，**積の平均 − 平均の積** が共分散になることを示している．

$$\sigma_{xy} = \frac{1}{n}\sum x_i y_i - \overline{x}\ \overline{y} \tag{9.2}$$

9.2.2　共分散の意味

ここで共分散の意味を考えるために，散布図上で 4 個のデータ点 S_1, S_2, S_3, S_4 が図 9.2 のように長方形の形に分布している状況をまず考えてみよう．この図で S_1 と S_2 の

データの平均値は，x が増加しても増加していない．すなわち，このように長方形の各頂点に 4 点があるようなデータでは，相関はない．

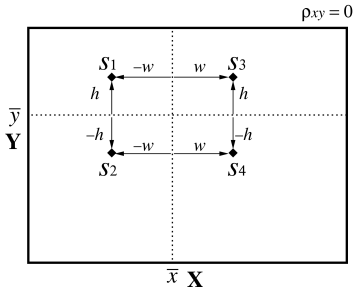


図 9.2　共分散がゼロになるデータの配置．2 つの点線は $\boldsymbol{x}$ と $\boldsymbol{y}$ の平均値を表す．

さて x_i, y_i の偏差というのは，x_i, y_i がそれぞれ平均値からどれだけずれているかを示す値のことであるから，図 9.2 のようなデータの配置では，x, y の偏差はいずれも図のように絶対値が等しい．そのことから，式 9.1 の値がどうなるかを考えてみよう．

$$\sigma_{xy} = \frac{1}{4}\{(-w) \times h + (-w) \times (-h) + w \times h + w \times (-h)\} = 0$$

すなわち共分散 σ_{xy} はゼロになる．

もっと一般の場合には，共分散はどのような値をとるのだろうか．図 9.3 の散布図をみてほしい．ここでは多数のデータがランダムに，つまり相関がないようにして与えてある．このとき，共分散はどのような値をとるかを考えてみよう．

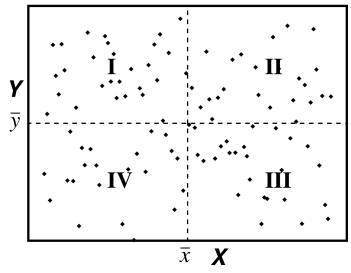


図 9.3　相関のないランダムなデータの散布図．破線は $\boldsymbol{x}$ および $\boldsymbol{y}$ の平均値．

まず $\boldsymbol{x}$ の偏差は，領域 **II** と **III** では正，領域 **I** と **IV** では負である．また $\boldsymbol{y}$ の偏差は，領域 **I** と **II** では正，領域 **III** と **IV** では負になる．したがって，式 (9.1) に現われる偏差の積 $(x_i - \overline{x})(y_i - \overline{y})$ の値は，領域 **II** と **IV** では正に，領域 **I** と **III** では負になることがすぐに分かる．

さて，これらのデータはランダムに与えられているのであるから，**I, II, III, IV** どの領域にも同じ確率で出現し，また図の破線の両側に対称に現われると考えてよいだろうから，4 つの領域のデータの偏差の積を足し合わせたものは，期待値がゼロになるはずである．すなわち，**相関をもたないランダムなデータの場合には，共分散の期待値はゼロになる．**

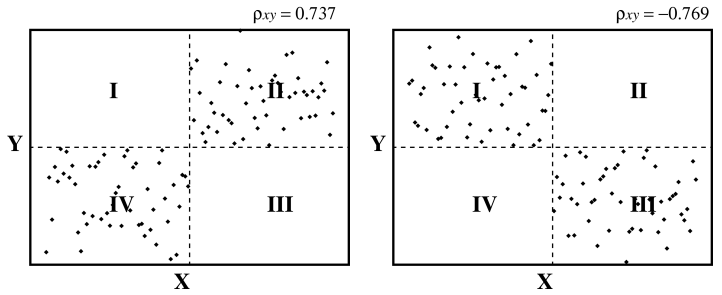


図 9.4　正と負の相関が生じる分布

もしも図 9.4 のようにデータが領域 **II** と **IV** にだけあったとしたら，結果はどうなるだろうか．今度は偏差の積が正の部分だけしかないのだから，その和は正，結局共分散は正になる．このとき，データは右上がりに分布しているから，結局，**データが右上がりに分布しているときには共分散の期待値は正になる．**逆に，**データが右下がりに分布しているときには共分散の期待値は負になる．**

9.2.3　相関係数

ある変数ともうひとつの変数の間の相関がどの程度強いかを表す量として**相関係数 (correlation constant)** がよく用いられる．相関係数は ρ_{xy} で表されることが多く[*2]，式 (9.3) で定義される．

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{9.3}$$

ここで，σ_x, σ_y はそれぞれ変数 $\boldsymbol{x}$, $\boldsymbol{y}$ の標準偏差であり，σ_{xy} は前節で登場した共分散である．この定義と前節での考察から，相関係数の値は，相関がなければゼロに近い値になることがわかる．それでは相関が強いときには相関係数はどのようになるのだろうか．

最も強い相関は図 9.1(4) のように，散布図でデータが直線状に配置されている状態である．ただし，直線が水平または垂直の場合には，相関がないことに注意しよう．つまり，強い相関というのは式 (9.4) のように y_i が x_i の 1 次式で表される状況である．

$$y_i = ax_i + b, \quad (i = 1, 2, \ldots, n) \tag{9.4}$$

[*2] 他に r_{xy} もよく使われる．なお，ρ は「ロー」と読む．

計算を簡単にするために $b=0$ として，次の比例関係が $\boldsymbol{x}$ と $\boldsymbol{y}$ の間にあるものとしよう．

$$y_i = ax_i,\ (i=1,2,\ldots,n) \tag{9.5}$$

このとき共分散 σ_{xy} を求めてみよう．まず δy_i は次のように変形される．

$$\delta y_i = y_i - \overline{y} = ax_i - a\overline{x} = a\delta x_i \tag{9.6}$$

ここですべての y_i が a 倍されると，それらの平均も a 倍されることを使っている．

式 (9.1) にこれを代入すると，

$$\begin{aligned}\sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n}\delta x_i\delta y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta x_i \times a\delta x_i \\ &= a\frac{1}{n}\sum_{i=1}^{n}\delta x_i^2 = a\sigma_x^2\end{aligned} \tag{9.7}$$

次に分散 σ_y^2 について考えてみよう．

$$\begin{aligned}\sigma_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(\delta y_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(a\delta x_i)^2 = a^2\frac{1}{n}\sum_{i=1}^{n}(\delta x_i)^2 \\ &= a^2\sigma_x^2\end{aligned} \tag{9.8}$$

これから $\sigma_y = \sqrt{\sigma_y^2} = |a|\sigma_x$ となり，相関係数は次のように表されることになる．

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{a\sigma_x^2}{|a|\sigma_x^2} = \pm 1,\ (a>0 \text{ のとき } 1,\ a<0 \text{ のとき } -1) \tag{9.9}$$

結局，式 (9.5) のように y_i が x_i に比例するときには，傾きの正負に対応して相関係数は 1 か -1 になることが分かった．この結論は，前節で，相関がないときには共分散の期待値がゼロになり，その結果相関係数の期待値もゼロになることを確かめているので，まとめると，2 つのデータ系列 $\boldsymbol{x}$, $\boldsymbol{y}$ の相関係数 ρ_{xy} とデータの相関の間には次のような関係がある．

相関係数 $=\pm 1$　$y_i = ax_i + b$ で表されるような完全な相関があるとき
相関係数 ≈ 0　相関がないとき

一般には相関係数はこれらの中間的な値をとり，0.9 以上であれば**強い相関**があるといえる．相関がないときの相関係数の期待値はゼロであるが，あるデータ列から得られる相関

係数は一般にゼロを中心としたある種の分布を作る．そのため，たとえば相関係数が 0.4 であった場合に，相関があるという判断を単純に行うことは無理がある．その問題については 9.2.5 節で検討する．

9.2.4 線形回帰 (最小二乗法)

相関のあるデータを散布図にして，その中にデータの変化の傾向を表す直線を引きたいことはよくある．なお，このように全体の傾向を表す関数のことを**モデル**と呼ぶこともある[*3]．

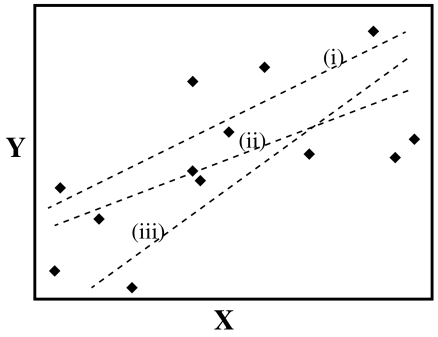


図 9.5 すべてのデータ点を代表する直線はどれがよいか

図 9.5 に描かれた 3 本の直線のうち，適切なものが (ii) であることは，勘で分かる．しかし客観的に最良のモデルとなる直線を決定するにはどうしたらよいだろうか．そのために用いられるのが**線形回帰 (linear regression)** または**最小二乗法** (**least-square method**) と呼ばれる方法である．

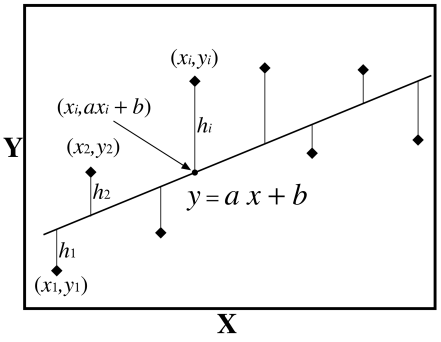


図 9.6 最小 2 乗法の原理: $h_1^2 + h_2^2 + \ldots$ が最小になるように a, b の値を決めてやる．

[*3] モデルという意味は次のようなことだ．— 与えられたデータ点の中に潜む関係としてはいろいろなものが考えられる．その中で単純な線形な関係を，仮にひとつの「モデル」として仮定してデータをそれに当てはめることで，問題を分析したり解釈したりしたい．

図 9.6 のようにデータ点 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられていたとき，$y = ax + b$ で表される直線を引いたとしよう．このとき，i 番目のデータ点と直線の縦のずれを h_i とすると，

$$h_i = y_i - (ax_i + b) \tag{9.10}$$

となる．直線 $y = ax + b$ は，a と b の値を変化させることで，傾きを変えたり平行にずらしたりできる．それでは a, b がどのような値をとったときに，直線はデータ点をもっともよく近似できるだろうか．詳しい導き方は付録（p.167）にゆずり，ここでは結果だけを示す．

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{9.11}$$

$$b = \overline{y} - a\overline{x} \tag{9.12}$$

式 (9.11),(9.12) で得られた a, b を用いて直線を引くと，データの増減を「ほどよく」表した直線が得られる．このような直線を**回帰直線**と呼ぶ．

例題 9–2　成人男子 6 人の靴のサイズと身長を調べたところ，次のようなデータの組が得られた．これらを散布図にプロットし，最小 2 乗法を使って回帰直線の係数を求めて，図に直線を描きなさい．

$(24.5, 165.4), (28.0, 182.7), (26.0, 171.6), (25.5, 173.1), (25.0, 175.1),$
$(24.0, 170.6)$

これらのデータの組を $(x_1, y_1), (x_2, y_2), \ldots$ とすると，式 (1.6) から分散 σ_x^2 が，式 (9.2) から共分散 σ_{xy} が得られる．具体的には x_i，y_i，x_i^2，y_i^2，$x_i y_i$ それぞれの平均，$\bar{x}, \bar{y}, \bar{x^2}, \bar{y^2}, \bar{xy}$ を個別に計算してから分散と共分散を求め，式 (9.12) に代入すればよい．電卓を使ってもかなりめんどうなので，Excel などを利用するとよい．

データの散布図と，このようにして求めた a, b を使って引いた直線を図 9.7 に示した．

以上のようにして求められた回帰直線で示される相関が，真の相関であるのか，あるいは母集団には相関がないのに，抽出によってたまたまある傾向が現れてしまったのかという疑問は，特にデータの点が少ないときには問題になる．このことについては次節で考えてみよう．

9.2.5　相関の有無を検定する

前述のように，母集団が全く相関をもたない場合でも，そこから無作為抽出を行った場合の標本の相関係数は一般にはゼロにならず，ある範囲で分布する．図 9.8 は，ランダム

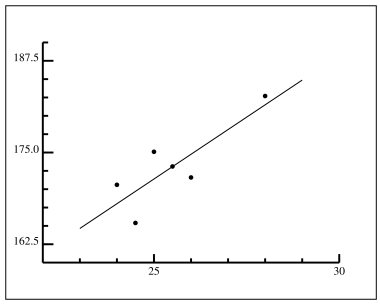


図 9.7　6 組のデータから得られた回帰直線

に 10 個の点を発生させて，その相関係数を計算したものである．真の相関は存在しないはずなのに，場合によってはかなり大きな相関係数が出現してしまうことがわかる．

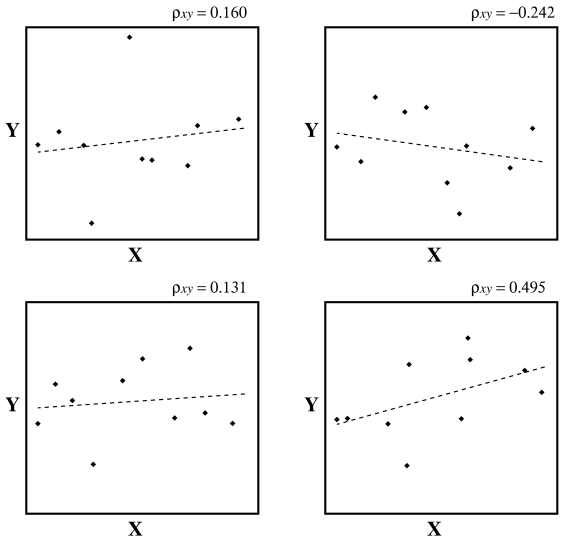


図 9.8　ランダムな操作で作られた偽の相関

この例のように，真の相関はないのに現われてしまう**偽の相関**を何らかの方法で検定して排除するにはどうしたらよいだろうか．ここでは次の定理を使うと，一定の仮定の下に相関の有無を検定することができる．

[定理] 相関を持たない母集団があり，ただしその母集団が 2 次元の正規分布 をしているとしたとき，そこから大きさ n の標本を無作為抽出したとすると，標本の相関係数 ρ_{xy} について，次の式 (9.13) のように定義される量 T は自由度 $n-2$ の t 分布に従う．

$$T = \sqrt{\frac{(n-2)\rho_{xy}^2}{1-\rho_{xy}^2}} \tag{9.13}$$

ここで 2 次元の正規分布というのは，図 9.9 のように 2 つの量がそれぞれ正規分布しているようなものをいう．図のように，2 つの量の間には相関がある場合もない場合もある．上の定理では，図 9.9 の左のように相関が全くない母集団から n 個の標本を抽出して散布図を作って相関係数 ρ_{xy} を求めることを想定している．この場合，何度も抽出を繰り返すと，その度に ρ_{xy} は異なった値をとるが，T のような量を計算してやると，その値は自由度 $n-2$ の t-分布をつくるというわけだ．

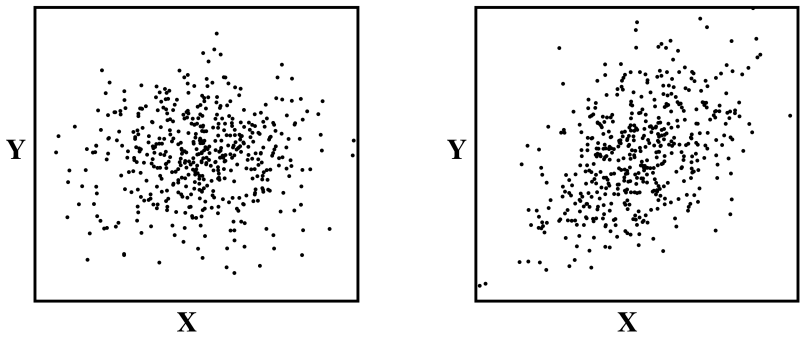


図 9.9　二次元の正規分布をもつデータの散布図：左は相関なし，右には正の相関がある．

ここで 9.2.4 節で出てきた例題をふたたび考えてみよう．図 9.7 を見ると，データ点は回帰直線からかなり離れていて，しかも標本のサイズは 6 しかない．このとき，本当に相関があるのかどうかは，微妙な問題である．そこで上の定理を用いて検定してみよう．

例題 9–3　成人男子 6 人の靴のサイズと身長を調べたところ，次のようなデータの組が得られた．このデータから，靴のサイズと身長には相関があるといえるかどうかを検定しなさい．

$(24.5, 165.4)$, $(28.0, 182.7)$, $(26.0, 171.6)$, $(25.5, 173.1)$, $(25.0, 175.1)$,
$(24.0, 170.6)$

ここでは，靴のサイズと身長には相関がないという仮説，つまり帰無仮説を立てて，それが棄却できるかどうかを検定する．

式 (9.3) を用いれば，このデータにもとづく相関係数は，最小 2 乗法の計算のついでぐらいの手間で求めることができ，実際に計算してみると，その値は $\rho_{xy} = 0.8323$ になる．なお，ここではすべての成人男子を母集団として，そちらのほうの相関係数は ρ_{xy} で表しているので，異なった表記法を使っている．

これから，式 (9.13) を使って，T を計算してみると，$n = 6$ であるから，

$$T = \sqrt{\frac{(6-2)0.8323^2}{1-0.8232^2}} = 3.00$$

となる．ここで t-分布表を見ると，自由度 4 のとき，3.00 という値は $\alpha = 0.01$ と $\alpha = 0.025$ の間に来る．つまり，母集団に相関がないと仮定した場合には，T は 0.025 以下の確率でしか現われないはずである．よって，靴のサイズと身長に相関がないという仮説は危険率 0.025 で棄却される．すなわち相関はある．

ここで注意しておくと，T は負の値をとることはないので，片側検定と同じく分布の正の領域だけで判断すればよい．つまり両側検定のときのように $\alpha = 0.025$ から危険率を 0.05 とすることはない．

現実にデータの間に相関があるかないかを判定したいことは，よくあることだ．散布図を見て感覚的に判定することもあるが，このように「きわどい」データの場合には判定に苦しむことになる．前提として母集団に正規分布が仮定できるというときという制限はあるもの，この検定はそのような場合に客観的な判定法を与えてくれて，実用的な意味が大きい．

9.2.6 線形でない相関

ここまで見てきたように，x_i と y_i が式 (9.4) のような 1 次式の関係にあるときは，線形の相関，あるいは 1 次の相関があるという．しかし，2 つの変数の間のの相関が直線的でない場合もしばしばある．たとえば図 9.10(a) のように変数の関係が指数関数に沿って分布している場合や，図 9.10(b) のように，2 次曲線に沿って分布している場合では，相関係数を求めることは意味がない．図 9.10(b) のような場合には，明確に相関があるにも関わらず相関係数がゼロに近くなってしまう．

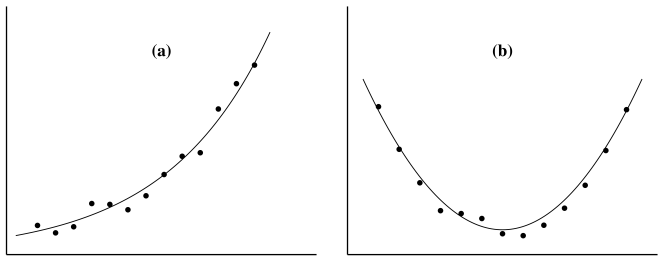


図 9.10 非線形の相関の例： (a) 指数関数に近似できる相関 (b)2 次関数に近似できる相関

このような場合には，**非線形最小 2 乗法**を使うことがある．具体的な手法についてはここでは述べないが，要点だけ解説しておこう．

たとえば，あるデータの傾向が理論モデルにもとづく予測から 2 次関数に従うと予測されているとしよう．だれでも一度は勉強したように，2 次関数は一般に次の形をしている．

$$y = ax^2 + bx + c$$

ここで係数 a, b, c を適当に与えることによって，最小 2 乗法のときと同じように，データと理論曲線の間の偏差から求められた 2 乗和を最小にするように工夫すればよい．このような手続きで，データをある与えられた関数で近似することができる．当然のことながら，これは直線でなく曲線なので，**回帰曲線**という．

観測されたデータがどのような関数の形に従うのかを決定することは，法則性を調べたり因果関係を推定する上で重要な意味をもつことが多い．そのような場合に，非線形の最小 2 乗法を用いて回帰曲線を求めることは，有力な解析の手段になり得る．

【章末問題】

問題 9–1 たまたま大きな卵を割ってみたときに，黄身がそれほど大きくないように思った A さんは，8 個の卵を買ってきて，全体の質量 (x g) と黄身の質量 (y g) を調べてみた．その結果，次のような結果を得た．

全体 x	62.2	42.8	61.8	79.3	63.1	51.4	60.9	69.9
黄身 y	36.7	28.7	32.0	37.7	31.8	31.5	32.3	34.8

1. x, y それぞれの分散と x と y の共分散を求めなさい．
2. x, y の相関係数 ρ_{xy} を求めなさい．
3. このデータを $y = ax + b$ に回帰したときの係数 a, b を求めなさい．
4. 式 (9.13) を使って， T の値を計算しなさい．さらに卵の質量と黄身の質量の間には相関がないという仮説を，危険率 0.01 で検定しなさい．

付録 A

重要な関係式などの導出

A.1　四分位数を求める

10 ページでは 3 つの四分位数（ひとつはメジアン）を決定する手順を示したが，2 個のデータ点を不均等に内分するところの説明は飛ばして天下りに記述してある．ここではデータの数が $n = 4m$ $(m = 1, 2, 3, \ldots)$ の場合についてもう少し詳しく説明する．

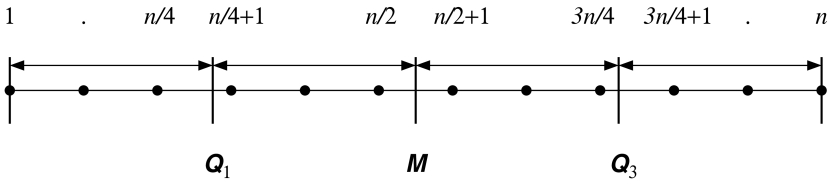


図 A.1　データ数が 4 の倍数の場合に四分位数 $\boldsymbol{Q_1}, \boldsymbol{M}, \boldsymbol{Q_3}$ を求めるための図．データ点 $x_1, x_{n/2}$ 等は添字だけで示してある．

図にはデータ点 $x_1, x_2, \ldots, x_n$ を等間隔で描いてある．左端の座標を 1 としてデータ間の区間の長さを 1 とすると，全体の長さは $n-1$ になる．$n-1$ は 4 では割り切れないので，4 等分した区切りは区間の途中に入る．

$\boldsymbol{Q_1}$ の位置は左端から $\frac{n-1}{4}$ のところにあるので，その座標は $1 + \frac{n-1}{4} = \frac{n}{4} + \frac{3}{4}$ となる．つまり $x_{n/4}$ と $x_{n/4+1}$ のデータを 3 : 1 に内分する点が Q_1 となる．同様にして，$\boldsymbol{Q_3}$ の位置は左端から $\frac{3(n-1)}{4}$ のところにあることから，$x_{3n/4}$ と $x_{3n/4+1}$ のデータを 1 : 3 に内分する点が Q_3 となる．

またメジアン $\boldsymbol{M}$ については，$x_{n/2}$ と $x_{n/2+1}$ を 1 : 1 に内分する点であることが図からすぐに分かる．

A.2　ベイズの定理

事象 A, B, C が互いに排反で，かつ標本空間を尽くしているとする．このとき，A, B, C それぞれの下に，ある事象 E が起きる条件付き確率，

$$P(E|A), P(E|B), P(E|C)$$

が知られているとすると，次の式が成立する．

$$P(A|E) = \frac{P(A)P(E|A)}{P(A)P(E|A) + P(B)P(E|B) + P(C)P(E|C)} \tag{A.1}$$

まず，条件付き確率 $P(A|E)$ は次の式に従う (p.35, 式 (2.19) 参照)．

$$P(A|E) = \frac{P(A \cap E)}{P(E)} \tag{A.2}$$

$P(B|E)$, $P(C|E)$ についても同様の式が成立する．逆に次の形も成立することにも注意しよう．

$$P(E|A) = \frac{P(A \cap E)}{P(A)} \tag{A.3}$$

さらに，事象 A, B, C が排反であり，かつ標本空間を尽くしているから，$P(E)$ は次のようになる．

$$P(E) = P(E \cap A) + P(E \cap B) + P(E \cap C) \tag{A.4}$$

そこで，式 (A.2) に式 (A.3) と式 (A.4) を代入して整理すれば，求める式が得られる．

A.3　確率変数の期待値と分散に関する公式

■式 (3.13) の導出

$X = \{x_1, x_2, \ldots, x_n\}$ とすると，$E[X] = \sum_{i=1}^{n} x_i f(x_i)$(式 (3.10) 参照) より，

$$\begin{aligned} E[aX + b] &= \sum_{i=1}^{n} (ax_i + b) f(x_i) = a \sum_{i=1}^{n} x_i f(x_i) + b \sum_{i=1}^{n} f(x_i) \\ &= aE[X] + b \end{aligned} \tag{A.5}$$

ここで $\sum_{i=1}^{n} f(x_i) = 1$, すなわちすべての確率の和は 1 に等しいことを使っている．

■式 (3.14) の導出

この形は確率統計の数学で頻出し，すでに式 (1.6) で出てきているが，もう少し形式的に扱ってみよう．まず次のように書けることに注意．

$$V[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = E[(X - E[X])^2] \tag{A.6}$$

これから次のように変形すれば得られる．$E[\ \ ]$ の中の $E[X]$ は，X と違って確率変数ではなく平均値という定数になっているので，外にくくり出せることが変形のポイントである．

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] - 2E[X]\cdot E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned} \tag{A.7}$$

■式 (3.15) の導出

式 (A.6) より，

$$\begin{aligned} V[aX+b] &= E[((aX+b) - (aE[X]+b))^2] \\ &= E[a^2X^2 + 2abX + b^2 - (a^2E[X]^2 + 2abE[X] + b^2)] \\ &= a^2E[X^2] - a^2E[X]^2 \\ &= a^2V[X] \end{aligned} \tag{A.8}$$

■式 (3.16) の導出

次のように展開した形で考えればわかりやすい．

$$\begin{aligned} E[X+Y] &= \frac{1}{n}\sum_{i=1}^{n}(x_i f(x_i) + y_i g(y_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i f(x_i) + \frac{1}{n}\sum_{i=1}^{n} y_i g(y_i) \\ &= E[X] + E[Y] \end{aligned} \tag{A.9}$$

■式 (3.17) の導出

ここではまず，2 つの確率変数 X, Y によって次のように定義される**共分散 (covariance)** $Cov(X,Y)$ を導入する．共分散は 2 つの確率変数の間の相関（非独立性）を表し，

X, Y の共分散が 0 ならば，これらは独立な確率変数である．定義をみると X と Y の偏差の積の和になっている．なお，共分散の展開形は，9.2.1 節 (p.146) で与えられている．

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{A.10}$$

この定義を変形してみよう．

$$\begin{aligned} Cov(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]EY] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

よって，

$$E[XY] = E[X]E[Y] + Cov(X,Y) \tag{A.11}$$

これから，X, Y が独立であれば $Cov(X,Y) = 0$ なので，

$$E[XY] = E[X]E[Y] \tag{A.12}$$

となる．

■式 (3.18) の導出

式 (3.14) と式 (A.11) を使えばよい．

$$\begin{aligned} V[X \pm Y] &= = E[((X \pm Y) - E[X \pm Y])^2] \\ &= E[X^2 \pm 2XY + Y^2 - (E[X]^2 \pm E[X]E[Y] + E[Y]^2)] \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \pm 2(E[XY] - E[X]E[Y]) \\ &= V[X] + V[Y] \pm 2Cov(X,Y) \end{aligned} \tag{A.13}$$

よって X ,Y が独立であれば次のようになる．

$$V[X \pm Y] = V[X] + V[Y] \tag{A.14}$$

A.4 二項分布の平均と分散

A.4.1 二項分布の平均

二項分布の平均（期待値）は 61 ページの式 (3.10) で与えられている．

$$\mu = \sum_{x=0}^{n} x \ {}_nC_x p^x q^{n-x} = np, \ \ (q = 1 - p) \tag{A.15}$$

この式を導くには次のようなテクニックを使うのが面白い．まず，二項定理を使って $(p+q)^n$ を展開すると次のようになる．

$$\begin{aligned}(p+q)^n &= \sum_{x=0}^{n} {}_nC_x p^x q^{n-x} \\ &= p^n + np^{n-1}q + \frac{n(n-1)}{2}p^{n-2}q^2 + \cdots \end{aligned} \tag{A.16}$$

ここで式 (A.16) を p で微分してみると，次のようになる．

$$n(p+q)^{n-1} = \sum_{x=0}^{n} x\ {}_nC_x p^{x-1} q^{n-x} \tag{A.17}$$

両辺に p を掛けて，$p+q=1$ を使えば，

$$np = \sum_{x=0}^{n} x\ {}_nC_x p^x q^{n-x} \tag{A.18}$$

となって，求める関係式が得られている．

A.4.2　二項分布の分散

次に分散 σ^2 を求めてみよう．最初に，

分散 ＝ 2 乗の平均 − 平均の 2 乗

であることを思い出しておこう．

式 (A.17) をもう一度 p で微分すると，

$$n(n-1)(p+q)^{n-2} = \sum_{x=0}^{n} x(x-1)_nC_x p^{x-2} q^{n-x} \tag{A.19}$$

が得られる．両辺に p^2 を掛けて，左辺と右辺を展開し，さらに $p+q=1$ と置くと，

$$n^2p^2 - np^2 = \sum_{x=0}^{n} x^2\ {}_nC_x p^x q^{n-x} - \sum_{x=0}^{n} x\ {}_nC_x p^x q^{n-x} \tag{A.20}$$

式 (A.15) より，$np=\mu$．従って左辺第 1 項は平均の 2 乗である．また，右辺第 1 項は 2 乗の平均を，右辺第 2 項は平均を意味している．これらに注意して整理すると，次のように変形して，最後の結果が得られる．

$$\begin{aligned}\sigma^2 &= \sum_{x=0}^{n} x^2\ {}_nC_x p^x q^{n-x} - n^2p^2 \\ &= \sum_{x=0}^{n} x\ {}_nC_x p^x q^{n-x} - np^2 = np - np^2 = npq \end{aligned} \tag{A.21}$$

A.5 ポアソン分布

ポアソン分布は，二項分布において n が非常に大きく，かつ p が非常に小さい極限で成立する確率分布である．つまり，$n \longrightarrow \infty,\ p \longrightarrow 0$ としたときに，

$$ {}_nC_x p^x (1-p)^{n-x} \longrightarrow \frac{\mu^x}{x!}\mathrm{e}^{-\mu} \tag{A.22} $$

となる．なおここで，$\mu = np$ であり，この値は平均値であって，有限の値をもつ．これを次のように段階に分けて証明する．

■x が小さい場合の ${}_nC_x$

$$ \begin{aligned} {}_nC_x &= \frac{n!}{x! \times (n-x)!} \\ &= \frac{n \cdot (n-1) \cdots 2 \cdot 1}{x \cdot (x-1) \cdots 2 \cdot 1 \times (n-x) \cdot (n-x-1) \cdots 2 \cdot 1} \end{aligned} \tag{A.23} $$

この式の分子の方をよく考えよう．n の階乗で，$n > x \geq 0$ だから $n \geq (n-x)$ となることから，次のように書いておく．

$$ n! = n \cdot (n-1) \cdots (n-x+1) \cdot (n-x) \cdot (n-x-1) \cdots 2 \cdot 1 $$

これから，式 (A.23) の分母と分子を通分して次の式を得る．

$$ {}_nC_x = \frac{n \cdot (n-1) \cdots (n-x+1)}{x \cdot (x-1) \cdots 2 \cdot 1} \tag{A.24} $$

ここで分子の $n \cdot (n-1) \cdots (n-x+1)$ は，x 個の因子の掛けあわせであることを押さえておこう[*1]．さらに $n \to \infty$ より $n \gg x$ であるから，$n-1$ や $n-x+1$ などはすべて n と近似的に等しいとみなせて，式 (A.24) は $n \to \infty$ の極限で次のように近似できる．

$$ {}_nC_x \longrightarrow \frac{n^x}{x!} \tag{A.25} $$

■$p^x(1-p)^{n-x}$ の極限値

まず，自然対数の底 (てい) である e (**ネイピア数 (Napier's constant)** ともいう) の定義は次の式で与えられることを心に留めておこう．この定義の式の意味については適当な数学の参考書を見ていただきたい．

[*1] $n-0$ から $n-(x-1)$ までの積であるから，$0, 1, \ldots, (x-1)$ と数え挙げると x 個ある．

$$\mathrm{e} = \lim_{q\to\infty}\left(1+\frac{1}{q}\right)^{q} \tag{A.26}$$

定義 (A.26) から次の式が導けることを証明なしにあげておく．

$$\lim_{q\to\infty}\left(1-\frac{1}{q}\right)^{q} = \frac{1}{\mathrm{e}} \tag{A.27}$$

それでは二項分布の式 (A.22) 左辺の後半に現れる因子[*2]を変形していこう．

$$p^x(1-p)^{n-x} = \left(\frac{p}{1-p}\right)^{x} \times (1-p)^n \tag{A.28}$$

式 (A.28) の右辺の前半の因子については，$p\to 0$ で $(1-p)\to 1$ となるから，

$$\left(\frac{p}{1-p}\right)^{x} \longrightarrow p^x \tag{A.29}$$

となる[*3]．さらに $(1-p)^n$ について考える．この式で $q = 1/p$ と置くと次のように変形でき，これから $p \longrightarrow 0$ における極限値として次の式が得られる．

$$\begin{aligned}(1-p)^n &= (1-p)^{\frac{1}{p}\times np}\\ &= \left(1-\frac{1}{q}\right)^{q\times\mu}\\ &\longrightarrow \mathrm{e}^{-\mu}\end{aligned} \tag{A.30}$$

3 番目の式を得るのには，式 (A.27) が使われている．

さて，ここで式 (A.25) (A.29) (A.30) をまとめると，

$$\begin{aligned}{}_nC_x p^x(1-p)^{n-x} &\longrightarrow \frac{n^x}{x!}\times p^x \times \mathrm{e}^{-\mu}\\ &= \frac{(np)^x}{x!}\mathrm{e}^{-\mu}\\ &= \frac{\mu^x}{x!}\mathrm{e}^{-\mu}\end{aligned} \tag{A.31}$$

というポアソン分布の式を得る．

[*2] 掛け算の形の式があるときに，掛け合わされるものを因子 (英語では factor) という．たとえば $ax(x-1)$ とあるときには，a, x, $x-1$ が因子である．因数ともいう．

[*3] $p \longrightarrow 0$ なら $1-p$ が 1 に近づくときに分子の p も同時にゼロにするべきではないかと思う人もいるだろう．しかし，そうすると全体がゼロになってしまって式としては意味を失う．$p \longrightarrow 0$ ということは，あくまで p は正の実数であって，何がしかの意味をもった値であることには違いはないのである．それでも気になる人は 0.99999 を 1 に近似しても大した違いはないが，0.00001 をゼロにしたらまずいということを考えればわかるはずだ．

A.6　標本平均の平均と分散の関係

$$\begin{aligned} E[\overline{X}] &= E[\frac{1}{n}(X_1 + X_2 + ... + X_n)] \\ &= \frac{1}{n}\left(E[X_1] + E[X_2] + ... + E[X_n]\right) \\ &= \frac{1}{n}(\mu + \mu + ... + \mu) = \mu \end{aligned} \tag{A.32}$$

2 つ目の式を得るために，式 (3.16) が使われている．また， $E[X_1]$ などを μ と置けるのは，母集団から要素を 1 個だけ取り出したときの期待値は母集団の中での平均値，すなわち母平均そのものであるからである．この結果を使うと，式 (6.4) も導かれる．

$$\begin{aligned} V[\overline{X}] =& E[(\overline{X} - E[\overline{X}])^2] \\ =& E\left[\left(\frac{1}{n}(X_1 + X_2 + ... + X_n) - \mu\right)^2\right] \\ =& \frac{1}{n^2}E\left[(X_1 + X_2 + ... + X_n - n\mu)^2\right] \\ =& \frac{1}{n^2}E\left[((X_1 - \mu) + (X_2 - \mu) + ... + (X_n - \mu))^2\right] \\ =& \frac{1}{n^2}E\left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + ... + (X_n - \mu)^2\right] \\ & -\frac{2}{n^2}E\left[((X_1 - \mu)(X_2 - \mu) + (X_1 - \mu)(X_3 - \mu) + ...)\right] \\ =& \frac{1}{n^2}\left(E[(X_1 - \mu)^2] + E[(X_2 - \mu)^2] + \ldots + E[(X_n - \mu)^2]\right) - 0 \\ =& \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{\sigma^2}{n} \end{aligned} \tag{A.33}$$

ここでは，$X_1, X_2, ...$ が独立であるということを前提に，式 (3.17) が使われている．すなわち，標本抽出が無作為かつ復元的に[*4]行われていることが必要である．

A.7　標本分散の平均と母分散の関係

式 (6.5) は，次のようにやや技巧的な導き方で得られる．まず s^2 を変形しておく．

[*4] 有限集団からの非復元抽出では，前の結果が後の結果に影響を及ぼすので，独立性が失われる．

$$
\begin{aligned}
s^2 &= \frac{1}{n}\left((X_1-\overline{X})^2+(X_2-\overline{X})^2+\ldots+(X_n-\overline{X})^2\right) \\
&= \frac{1}{n}\left(X_1^2+X_2^2+\cdots+X_n^2\right)-\overline{X}^2 \ (2\text{ 乗の平均 }-\text{ 平均の 2 乗})
\end{aligned}
$$

ここで，

$$
E[X_1^2] = E[X_2^2] = \cdots = \sigma^2 \tag{A.34}
$$

また，

$$
E[\overline{X}^2] = V[\overline{X}] = \frac{\sigma^2}{n} \tag{A.35}
$$

これらを使って，

$$
\begin{aligned}
E[s^2] &= \frac{1}{n}(\sigma^2+\sigma^2+\cdots)-\frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

この導出の途中では式 (6.3), (6.4) の導出過程をも用いた．

A.8　最小二乗法

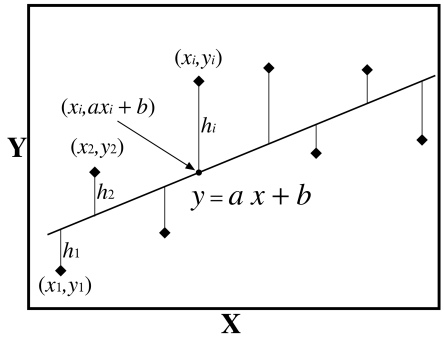


図 A.2　最小二乗法の原理: $h_1^2+h_2^2+\ldots$ が最小になるように a, b の値を決めてやる．

図 A.2 のようにデータ点 $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ が与えられていたとき，$y=ax+b$ で表される直線を引いたとしよう．このとき，i 番目のデータ点と直線の縦のずれ

を h_i とすると，

$$h_i = y_i - (ax_i + b) \tag{A.36}$$

となる．直線 $y = ax + b$ は，a と b の値を変化させることで，傾きを変えたり平行にずらしたりできる．それでは a, b がどのような値をとったときに，直線はデータ点をもっともよく近似できるだろうか．

そのためにまず次の S を定義しておこう．近似がもっともよいときには，S は極小になるはずである．

$$S = \frac{1}{n}(h_1^2 + h_2^2 + \ldots + h_n^2) \tag{A.37}$$

ここで右辺に $\frac{1}{n}$ を掛けているのは，後の計算をうまく処理するためである．

式 (A.37) に式 (A.36) を代入して整理すると，次の式が得られる．

$$S = \overline{y^2} + a^2\overline{x^2} - 2b\overline{y} - 2a\overline{xy} + 2ab\overline{x} + b^2 \tag{A.38}$$

a, b を変化させて S が最小になる条件を求めるには，次の 2 つの偏微分がゼロになればよい．

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2a\overline{x^2} - 2\overline{xy} + 2b\overline{x} = 0 \\ \frac{\partial S}{\partial b} &= -2\overline{y} + 2a\overline{x} + 2b = 0 \end{aligned} \tag{A.39}$$

これを整理して，次のような連立方程式が得られる．ただしここでは a, b が未知数であることに注意！

$$\overline{x^2}a + \overline{x}b = \overline{xy} \tag{A.40}$$

$$\overline{x}a + b = \overline{y} \tag{A.41}$$

これを解くと次の結果が得られる．

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{A.42}$$

$$b = \overline{y} - a\overline{x} \tag{A.43}$$

♡ **偏微分って何？** ♡

上の計算に出てきた**偏微分**を初めて見る人がいるかも知れない．これは複数の変数をもつ関数を微分するのに 1 つの変数だけで微分し，他の変数は定数として扱うものだ．だからある関数の偏微分は，変数の数だけある．例として，x, y を変数とする関数

$$f(x,y) = ax^2 + bxy + cy^2 + dy$$

を考える．このとき次のように x, y それぞれに関する偏微分が存在する．偏微分では，微分記号に通常の d ではなく，∂ を使う．

$$\begin{aligned} \frac{\partial f(x,y)}{\partial x} &= 2ax + by \\ \frac{\partial f(x,y)}{\partial y} &= bx + 2cy + d \end{aligned}$$

x による偏微分では y は定数とみなされるために，$f(x,y)$ の cy^2 と dy はゼロになることに留意してほしい．

★

付録 B

数表

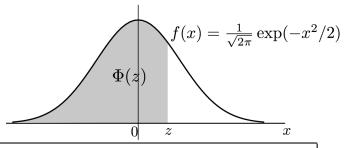
B.1 正規分布のパーセント点

パーセント	90	95	97.5	99	99.5
α	0.10	0.05	0.025	0.01	0.005
z_α	1.282	1.645	1.960	2.326	2.576

B.2 正規分布表

正規分布表には積分範囲が異なるものがあるので注意すること.

$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$



	0	1	2	3	4	5	6	7	8	9
0.0	.50000	.50398	.50797	.51196	.51595	.51993	.52392	.52790	.53188	.53585
0.1	.53982	.54379	.54775	.55171	.55567	.55961	.56355	.56749	.57142	.57534
0.2	.57925	.58316	.58706	.59095	.59483	.59870	.60256	.60641	.61026	.61409
0.3	.61791	.62171	.62551	.62930	.63307	.63683	.64057	.64430	.64802	.65173
0.4	.65542	.65909	.66275	.66640	.67003	.67364	.67724	.68082	.68438	.68793
0.5	.69146	.69497	.69846	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72574	.72906	.73237	.73565	.73891	.74215	.74537	.74857	.75174	.75490
0.7	.75803	.76114	.76423	.76730	.77035	.77337	.77637	.77935	.78230	.78523
0.8	.78814	.79102	.79389	.79673	.79954	.80233	.80510	.80784	.81057	.81326
0.9	.81593	.81858	.82121	.82381	.82639	.82894	.83147	.83397	.83645	.83891
1.0	.84134	.84375	.84613	.84849	.85083	.85314	.85542	.85769	.85992	.86214
1.1	.86433	.86650	.86864	.87076	.87285	.87492	.87697	.87899	.88099	.88297
1.2	.88493	.88686	.88876	.89065	.89251	.89435	.89616	.89795	.89972	.90147
1.3	.90319	.90490	.90658	.90824	.90987	.91149	.91308	.91465	.91620	.91773
1.4	.91924	.92073	.92219	.92364	.92506	.92647	.92785	.92921	.93056	.93188
1.5	.93319	.93447	.93574	.93699	.93821	.93942	.94062	.94179	.94294	.94408
1.6	.94520	.94630	.94738	.94844	.94949	.95052	.95154	.95254	.95352	.95448
1.7	.95543	.95636	.95728	.95818	.95907	.95994	.96079	.96163	.96246	.96327
1.8	.96406	.96485	.96562	.96637	.96711	.96784	.96855	.96925	.96994	.97062
1.9	.97128	.97193	.97257	.97319	.97381	.97441	.97500	.97558	.97614	.97670
2.0	.97724	.97778	.97830	.97882	.97932	.97981	.98030	.98077	.98123	.98169
2.1	.98213	.98257	.98299	.98341	.98382	.98422	.98461	.98499	.98537	.98573
2.2	.98609	.98644	.98679	.98712	.98745	.98777	.98808	.98839	.98869	.98898
2.3	.98927	.98955	.98982	.99009	.99035	.99061	.99086	.99110	.99134	.99157
2.4	.99180	.99202	.99223	.99245	.99265	.99285	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99429	.99445	.99461	.99476	.99491	.99505	.99520
2.6	.99533	.99547	.99560	.99573	.99585	.99597	.99609	.99620	.99631	.99642
2.7	.99653	.99663	.99673	.99683	.99692	.99702	.99710	.99719	.99728	.99736
2.8	.99744	.99752	.99759	.99767	.99774	.99781	.99788	.99794	.99801	.99807
2.9	.99813	.99819	.99824	.99830	.99835	.99841	.99846	.99851	.99855	.99860

(以下次ページ)

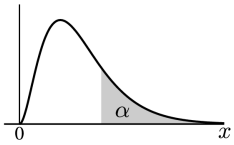
(前ページからの続き)

	0	1	2	3	4	5	6	7	8	9
3.0	.99865	.99869	.99873	.99877	.99881	.99885	.99889	.99892	.99896	.99899
3.1	.$9_3$032	.$9_3$064	.$9_3$095	.$9_3$125	.$9_3$155	.$9_3$183	.$9_3$211	.$9_3$237	.$9_3$263	.$9_3$288
3.2	.$9_3$312	.$9_3$336	.$9_3$359	.$9_3$381	.$9_3$402	.$9_3$422	.$9_3$442	.$9_3$462	.$9_3$480	.$9_3$499
3.3	.$9_3$516	.$9_3$533	.$9_3$549	.$9_3$565	.$9_3$581	.$9_3$595	.$9_3$610	.$9_3$624	.$9_3$637	.$9_3$650
3.4	.$9_3$663	.$9_3$675	.$9_3$686	.$9_3$698	.$9_3$709	.$9_3$719	.$9_3$729	.$9_3$739	.$9_3$749	.$9_3$758
3.5	.$9_3$767	.$9_3$775	.$9_3$784	.$9_3$792	.$9_3$799	.$9_3$807	.$9_3$814	.$9_3$821	.$9_3$828	.$9_3$834
3.6	.$9_3$840	.$9_3$846	.$9_3$852	.$9_3$858	.$9_3$863	.$9_3$868	.$9_3$873	.$9_3$878	.$9_3$883	.$9_3$887
3.7	.$9_3$892	.$9_3$896	.$9_4$003	.$9_4$042	.$9_4$079	.$9_4$115	.$9_4$150	.$9_4$183	.$9_4$215	.$9_4$246
3.8	.$9_4$276	.$9_4$305	.$9_4$332	.$9_4$359	.$9_4$384	.$9_4$409	.$9_4$433	.$9_4$455	.$9_4$477	.$9_4$498
3.9	.$9_4$519	.$9_4$538	.$9_4$557	.$9_4$575	.$9_4$592	.$9_4$609	.$9_4$625	.$9_4$640	.$9_4$655	.$9_4$669
4.0	.$9_4$683	.$9_4$696	.$9_4$709	.$9_4$721	.$9_4$732	.$9_4$743	.$9_4$754	.$9_4$764	.$9_4$774	.$9_4$784
4.1	.$9_4$793	.$9_4$802	.$9_4$810	.$9_4$818	.$9_4$826	.$9_4$833	.$9_4$840	.$9_4$847	.$9_4$854	.$9_4$860
4.2	.$9_4$866	.$9_4$872	.$9_4$877	.$9_4$883	.$9_4$888	.$9_4$893	.$9_4$897	.$9_5$022	.$9_5$065	.$9_5$106
4.3	.$9_5$146	.$9_5$183	.$9_5$219	.$9_5$254	.$9_5$287	.$9_5$319	.$9_5$349	.$9_5$378	.$9_5$406	.$9_5$433
4.4	.$9_5$458	.$9_5$483	.$9_5$506	.$9_5$528	.$9_5$550	.$9_5$570	.$9_5$590	.$9_5$608	.$9_5$626	.$9_5$643
4.5	.$9_5$660	.$9_5$675	.$9_5$690	.$9_5$705	.$9_5$718	.$9_5$731	.$9_5$744	.$9_5$756	.$9_5$767	.$9_5$778
4.6	.$9_5$788	.$9_5$798	.$9_5$808	.$9_5$817	.$9_5$825	.$9_5$834	.$9_5$841	.$9_5$849	.$9_5$856	.$9_5$863
4.7	.$9_5$869	.$9_5$876	.$9_5$882	.$9_5$887	.$9_5$893	.$9_5$898	.$9_5$903	.$9_5$907	.$9_5$912	.$9_5$916
4.8	.$9_5$920	.$9_5$924	.$9_5$928	.$9_5$931	.$9_5$935	.$9_5$938	.$9_5$941	.$9_5$944	.$9_5$946	.$9_5$949
4.9	.$9_5$952	.$9_5$954	.$9_5$956	.$9_5$958	.$9_5$960	.$9_5$962	.$9_5$964	.$9_5$966	.$9_5$968	.$9_5$969
5.0	.$9_5$971	.$9_5$972	.$9_5$974	.$9_5$975	.$9_5$976	.$9_5$977	.$9_5$979	.$9_5$980	.$9_5$981	.$9_5$982

.$9_3$032 とあるのは 0.$\underline{999}$032 の意味．下付の添字で連続する 9 の数を示した．他同様．

B.3 χ^2分布表

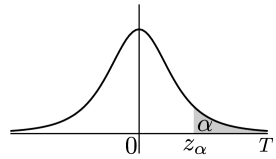
【表の見方】 自由度 ν が 5 のときに α の面積が 0.05 になるような x の値 (x_α) は，表から 11.071 となる．3.93e-5 は 3.93×10^{-5} を表す．



α	0.995	0.975	0.95	0.9	0.5	0.05	0.025	0.01	0.005
$\nu=1$	3.93e-5	9.82e-4	3.93e-3	0.0158	0.455	3.841	5.024	6.635	7.879
$\nu=2$	0.010	0.051	0.103	0.211	1.386	5.991	7.378	9.210	10.597
$\nu=3$	0.072	0.216	0.352	0.584	2.366	7.815	9.348	11.345	12.838
$\nu=4$	0.207	0.484	0.711	1.064	3.357	9.488	11.143	13.277	14.861
$\nu=5$	0.412	0.831	1.145	1.610	4.351	11.071	12.833	15.086	16.750
$\nu=6$	0.676	1.237	1.635	2.204	5.348	12.592	14.449	16.812	18.548
$\nu=7$	0.989	1.690	2.167	2.833	6.346	14.067	16.013	18.475	20.278
$\nu=8$	1.344	2.180	2.733	3.490	7.344	15.507	17.535	20.090	21.955
$\nu=9$	1.735	2.700	3.325	4.168	8.343	16.919	19.023	21.666	23.589
$\nu=10$	2.156	3.247	3.940	4.865	9.342	18.307	20.483	23.209	25.188
$\nu=11$	2.603	3.816	4.575	5.578	10.341	19.675	21.920	24.725	26.757
$\nu=12$	3.074	4.404	5.226	6.304	11.340	21.026	23.337	26.217	28.300
$\nu=13$	3.565	5.009	5.892	7.041	12.340	22.362	24.736	27.688	29.819
$\nu=14$	4.075	5.629	6.571	7.790	13.339	23.685	26.119	29.141	31.319
$\nu=15$	4.601	6.262	7.261	8.547	14.339	24.996	27.488	30.578	32.801
$\nu=16$	5.142	6.908	7.962	9.312	15.338	26.296	28.845	32.000	34.267
$\nu=17$	5.697	7.564	8.672	10.085	16.338	27.587	30.191	33.409	35.718
$\nu=18$	6.265	8.231	9.390	10.865	17.338	28.869	31.526	34.805	37.156
$\nu=19$	6.844	8.906	10.117	11.651	18.338	30.144	32.852	36.191	38.582
$\nu=20$	7.434	9.591	10.851	12.443	19.337	31.410	34.170	37.566	39.997
$\nu=21$	8.034	10.283	11.591	13.239	20.337	32.670	35.479	38.932	41.401
$\nu=22$	8.643	10.982	12.338	14.041	21.337	33.924	36.781	40.289	42.796
$\nu=23$	9.260	11.688	13.090	14.848	22.337	35.172	38.076	41.638	44.181
$\nu=24$	9.886	12.401	13.848	15.659	23.337	36.415	39.364	42.980	45.558
$\nu=25$	10.520	13.120	14.611	16.473	24.336	37.652	40.646	44.314	46.928
$\nu=26$	11.160	13.844	15.379	17.292	25.336	38.885	41.923	45.642	48.290
$\nu=27$	11.807	14.573	16.151	18.114	26.336	40.113	43.194	46.963	49.645
$\nu=28$	12.461	15.308	16.928	18.939	27.336	41.337	44.461	48.278	50.993
$\nu=29$	13.121	16.047	17.708	19.768	28.336	42.557	45.722	49.588	52.336
$\nu=30$	13.787	16.791	18.493	20.599	29.336	43.773	46.979	50.892	53.672
$\nu=40$	20.706	24.433	26.509	29.050	39.335	55.758	59.342	63.691	66.766
$\nu=50$	27.991	32.357	34.764	37.689	49.335	67.505	71.420	76.154	79.490
$\nu=60$	35.534	40.482	43.188	46.459	59.335	79.082	83.298	88.379	91.952
$\nu=80$	51.172	57.153	60.391	64.278	79.334	101.88	106.63	112.33	116.32
$\nu=90$	59.20	65.65	69.13	73.29	89.33	113.15	118.14	124.12	128.30
$\nu=100$	67.33	74.22	77.93	82.36	99.33	124.34	129.56	135.81	140.17

B.4 Student の t-分布表

【表の見方】 自由度 ν が 5 のときに α の面積が 0.05 になるような t の値 (z_α) は 2.015 である.



α	0.1	0.05	0.025	0.01	0.005	0.0025
$\nu = 1$	3.078	6.314	12.706	31.821	63.657	127.321
$\nu = 2$	1.886	2.920	4.303	6.965	9.925	14.089
$\nu = 3$	1.638	2.353	3.182	4.541	5.841	7.453
$\nu = 4$	1.533	2.132	2.776	3.747	4.604	5.598
$\nu = 5$	1.476	2.015	2.571	3.365	4.032	4.773
$\nu = 6$	1.440	1.943	2.447	3.143	3.707	4.317
$\nu = 7$	1.415	1.895	2.365	2.998	3.499	4.029
$\nu = 8$	1.397	1.860	2.306	2.896	3.355	3.833
$\nu = 9$	1.383	1.833	2.262	2.821	3.250	3.690
$\nu = 10$	1.372	1.812	2.228	2.764	3.169	3.581
$\nu = 11$	1.363	1.796	2.201	2.718	3.106	3.497
$\nu = 12$	1.356	1.782	2.179	2.681	3.055	3.428
$\nu = 13$	1.350	1.771	2.160	2.650	3.012	3.372
$\nu = 14$	1.345	1.761	2.145	2.624	2.977	3.326
$\nu = 15$	1.341	1.753	2.131	2.602	2.947	3.286
$\nu = 16$	1.337	1.746	2.120	2.583	2.921	3.252
$\nu = 17$	1.333	1.740	2.110	2.567	2.898	3.222
$\nu = 18$	1.330	1.734	2.101	2.552	2.878	3.197
$\nu = 19$	1.328	1.729	2.093	2.539	2.861	3.174
$\nu = 20$	1.325	1.725	2.086	2.528	2.845	3.153
$\nu = 21$	1.323	1.721	2.080	2.518	2.831	3.135
$\nu = 22$	1.321	1.717	2.074	2.508	2.819	3.119
$\nu = 23$	1.319	1.714	2.069	2.500	2.807	3.104
$\nu = 24$	1.318	1.711	2.064	2.492	2.797	3.091
$\nu = 25$	1.316	1.708	2.060	2.485	2.787	3.078
$\nu = 30$	1.310	1.697	2.042	2.457	2.750	3.030
$\nu = 35$	1.306	1.690	2.030	2.438	2.724	2.996
$\nu = 40$	1.303	1.684	2.021	2.423	2.704	2.971
$\nu = 45$	1.301	1.679	2.014	2.412	2.690	2.952
$\nu = 50$	1.299	1.676	2.009	2.403	2.678	2.937
$\nu = 60$	1.296	1.671	2.000	2.390	2.660	2.915
$\nu = 70$	1.294	1.667	1.994	2.381	2.648	2.899
$\nu = 80$	1.292	1.664	1.990	2.374	2.639	2.887
$\nu = 90$	1.291	1.662	1.987	2.368	2.632	2.878
$\nu = 100$	1.290	1.660	1.984	2.364	2.626	2.871
$\nu = 120$	1.289	1.658	1.980	2.358	2.617	2.860

付録 C

ちょっとした数学的手法

C.1　比例配分によるデータの内挿

あまり変化が急激でない関数 $y = f(x)$ があって，飛び飛びに関数の値が分かっているものとしよう．そのとき，任意の点の関数値を，それを両側から挟む 2 点のデータから概算することができる．

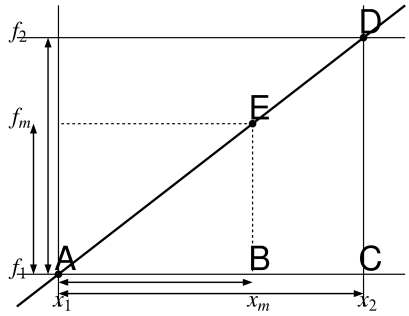


図 C.1　データの比例配分の原理

図 C.1 を見てほしい．斜めの太い直線が，関数 $f(x)$ の一部である．x_1 と x_2 の間は十分に小さいので，$f(x)$ はほとんど直線とみなしてよい．x_1, x_2 における関数の値 $f(x_1)$, $f(x_2)$ をそれぞれ f_1, f_2 とする．x_m はその中間のどこかである．このとき f_m をこの図を使って求めることができる．これを比例配分，あるいは一次の内挿といい，よくある計算手法である．

図をみると三角形 ACD と ABE は相似だから次の式が成り立つ．

$$\frac{x_2 - x_1}{f_2 - f_1} = \frac{x_m - x_1}{f_m - f_1} \tag{C.1}$$

これを変形していけば，次の形が得られる．

$$f_m = f_1 + \frac{(f_2 - f_1)(x_m - x_1)}{x_2 - x_1} \tag{C.2}$$

C.2 有効数字

C.2.1 基本的な考え方

アナログの体重計に乗ってみたところ 52.6 の目盛りと 52.7 の目盛りの真ん中よりやや上のところに針が来た．そこでこれを 52.7 kg として記録した．これは針の位置が 52.7 の目盛りに最も近いから，いいかえれば次の不等式を満たす x を 52.7 としようというのである．

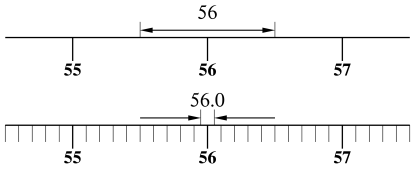


図 C.2 56 と 56.0 のちがい：単に 56 は上の目盛りで読み取った範囲，56.0 なら下の細かい目盛りで読んでいると考えられる．

$$52.65 \leq x < 52.75$$

つまり 52.7 という値が報告された場合には，± 0.05 の幅をもつ量であると考えるべきで，とりあえず 0.7 のところは意味のある (significant な) データと見なしてよい．これが四捨五入という操作の意味である．そして意味のある数字は上の桁から 5,2,7 の 3 桁であるから，これを**有効数字または有効桁数を 3 桁である**という．

次に 56.0 kg というデータが記録されていたとしよう．最後の 0 は何のためについているのだろうか．ちょっと考えると 0 があろうとなかろうとゼロはゼロでしかないのだから，これを 56 kg としても何ら不都合はないように思える．しかし，上の有効数字の考えに照らせば，これらはやはり意味が異なるということがわかる．前者は 0.01 kg の桁を四捨五入したものであり，後者は 0.1 kg の桁を四捨五入したものなので，データの精度が大きく異なるのである (図 C.2)．

あるいは実際に即していえば，56.0 というデータを読み取れる秤（はかり）と 56 kg としか読み取れない秤とでは，目盛りの細かさや針の位置の正確さが異なるといってもよい．

C.2.2 実際のデータの有効数字

表 C.1 に，具体的な数値の表記例と，それらの有効桁数の大きさを示した．0.01234 のように，単に大きさを示すためだけに 0 が先頭についている小数については，0 でない数字からが有効数字に含まれる．

810 の場合には，最後の 0 が小数第 1 位を四捨五入して得られたものなのか，あるいは 1 の位が四捨

表 C.1 有効数字の桁数: *少なくとも 2，多くて 3

表記	有効桁数
12.3	3
12.30	4
0.01234	4
0.0012	2
813	3
810	不明*
810.	3
8.10×10^2	3
6.02×10^{23}	3

五入されて 10 の位の 1 が出てきたのかが分からない．前者であれば有効数字は 3 桁であり，後者なら 2 桁ということになる．精度が問題になるようなケースでは，このようなあいまいな表記は好ましくない． ただし 625, 810, 752 などと複数の数値と一緒に並んでいるのなら，推定はできる．

上の問題点を回避するためには指数表記を行うとよい．表にあるように 8.10×10^2 とあれば有効数字は 3 桁，8.1×10^2 とあれば 2 桁と明確に分かる．

C.3　数値の丸め誤差

今，真の値が $a = 0.505,\ b = 1.05$ であるような 2 つの数の積を計算することを考える．

真の積はもちろん，

$$ab = 0.505 \times 1.05 = 0.53025$$

である．ところでこれらをいずれも有効数字 3 桁目で四捨五入したとして，$a' = 0.51,\ b' = 1.1$ と近似値を使って計算すると，

$$a'b' = 0.561$$

となる．この結果は真の値に比べて，

$$\frac{0.561 - 0.53025}{0.53025} = 0.0580$$

すなわち，6% ほども大きすぎる値になっている．このように，四捨五入は誤差を結果に引き込んでしまう処理であるから，計算に使う前の数値を四捨五入するときには慎重な注意を要する．上の場合に即して言うならば，**もし最終結果で有効数字が 2 桁ほしいのであっても，途中計算は 3 桁以上の数値をそのまま使って進めておいて，最後に得られた値だけを四捨五入によって丸めるのが正しい**のである．

C.4　多数回の計算による丸め誤差の蓄積

確率統計の計算では，丸め誤差が発生する要因がかなりある．たとえば次のような計算を見てみよう．二項分布でしばしば現れる次のようなケースはどうだろうか？

$${}_{12}C_4 \times \left(\frac{1}{7}\right)^4 \left(\frac{6}{7}\right)^8 = \frac{12 \cdot 11 \cdot 10 \cdot 9}{4 \cdot 3 \cdot 2 \cdot 1} \left(\frac{1}{7}\right)^4 \left(\frac{6}{7}\right)^8$$

ここで，$1/7 = 0.14285714... \approx 0.143,\ 6/7 = 0.85714285... \approx 0.857$ として，ひとつずつ掛け合わせて計算してみると，0.060227... が得られ，一方，可能な限り正確に計算してみると，0.060067... となる．つまり，3 桁に丸めて計算した値は，割合で見ればかなり大きく見積もってしまうことになる．

そこで，$1/7 \approx 0.142857$, $6/7 \approx 0.857412$ で，計算してみると，結果は 0.060066.. と，正確な値にかなり近い．

このように，計算を何度も行うときには，丸めによる誤差がどんどん蓄積することになる[*1]ので，それを避けるためには，なるべく長い桁数で計算を進めていく必要がある．電卓のような簡単な道具で計算を進めるときにも，**多数回の計算による誤差が予想される場合には，目的とする数値の桁数よりも 4 桁以上長い桁数で途中計算を進める**ことが望ましい．

実際の統計用アプリケーションや表計算アプリケーションでは，1 万回を超える繰り返し計算も稀ではないので，ずっと高い精度で内部の計算が行われている．

ただし，**結果の数字のほうは，必要に応じた桁数で表現するべきであって，意味もなく長い桁数で数値を記すのは誤りである．**答を求められている課題の意味を考えて，たとえば日常的な問題で出てくる確率を求めるのであれば，せいぜい 2 桁程度の精度で回答すればよい．

[*1] これも確率論の問題として考えることができ，正規分布が現れるが，ここでは定性的な議論にとどめる．

付録 D

電卓とコンピュータを活用する

D.1　電卓で統計計算

家庭用の 300 円ぐらいの電卓でも，簡単な統計の計算に活用することができる．ただし，メモリキー [M+] [MR] が備わっているものでないといけない．ルートキー [√] はあった方が便利だが，なくても平方根の計算は難しくない（後述）．

基本的な使い方

■メモリーキーの意味

[M+] メモリーに加算 (memory plus)
[MR] メモリーを呼び出す (memory recall)
[MC] メモリーをクリア (memory clear)
[MRC] メモリー呼び出し，二度押しでクリア (機種によって MR,MC の代わりにある)

■メモリーを使って次々に加算する

加算を繰り返していくときには，[＋] キーよりもメモリー機能を活用したほうが効率がよい．まずは単純な足し算でメモリの使い方を覚えよう．

計算式： $\mathbf{1+2+3+4+5+6+7+8+9+10}$
キー操作： **1** [M+] **2** [M+] **3** [M+] $\cdots$ **10** [M+]　[MR] $\Rightarrow$ **55**

結果を消去するには，[MC] でクリアする．

メモリー機能は次の積和の計算でさらに威力を発揮する．

計算式： $\mathbf{23.5\times 3+41.2\times 5+51.0\times 6+13.5\times 2}$
キー操作： **23.5** [×] **3** [＝] [M+] **41.2** [×] **5** [＝] [M+] **51.0** [×] **6** [＝] [M+]
13.5 [×] **2** [＝] [M+]　[MR] $\Rightarrow$ **609.5**

■2 乗を計算
計算式: $\mathbf{12^2}$
キー操作: **12** [×] [=] ⇒ **144**

■2 乗の和を求める
計算式: $\mathbf{1^2 + 2^2 + 3^2 + \cdots + 10^2}$
キー操作: **1** [×] [=] [M+] **2** [×] [=] [M+] **3** [×] [=] [M+] ··· **10** [×] [=] [M+] [=] ⇒ **385**

■平方根
2 乗の繰り返し計算で平方根を追い詰めて[*1]，$\sqrt{30}$ を有効数字 3 桁まで求めてみる．ドン臭いやり方に見えるが，回数に比例して有効数字が増え，効率はよい．

- $5^2 = 25$, $6^2 = 36$ だから，**5.5** [×] [=] を試す．⇒ 30.25 ほんの少し大きすぎ．
- **5.4** [×] [=] を試す．⇒ 29.16 戻し過ぎ．
- **5.45** [×] [=] でどうだ？ ⇒ 29.7025 ちょっと足りない．
- **5.47** [×] [=] でどうかな？ ⇒ 29.9209 まだ足りない．
- **5.48** [×] [=] でどうよ？ ⇒ 30.0304 行き過ぎたから戻すか．
- **5.475** [×] [=] ならどうだ？ ⇒ 29.975626 足りない．
- つまり，5.475 と 5.480 の間になる．
- 5.475 を四捨五入して，5.48 でオーケー！

生データからの統計計算

■平均
データ： **23.5, 41.2, 50.1, 32.3**
キー操作 **23.5** [M+] **41.2** [M+] **50.1** [M+] **32.3** [M+] [MR] [÷] **4** [=] ⇒ 36.775

■分散
分散が 2 乗の平均 − 平均の 2 乗 で計算できることを利用する．

キー操作: **2 乗の平均：** **23.5** [×] [=] [M+] **41.2** [×] [=] [M+] ··· [MR] [÷] **4** [=] ⇒ 1450.7475 メモしておく

平均の 2 乗から 2 乗の平均を引く:

36.775 [×] [=] [−] **1450.7475** ⇒ −98.347

符号反転:

符号が反対になっているので， 98.35 が分散．

[*1] 方程式の解を繰り返し計算で求めるニュートン・ラフソン法 (ニュートン法) という計算手法の原理だ．

標準偏差：

さらに標準偏差を求めるには，平方根をとって $\sqrt{98.35} = 9.92$

度数分布から平均と分散を求める

データ：

階級値	**5**	**15**	**25**	**35**	**45**
度数	**2**	**4**	**7**	**4**	**3**
累積度数	**2**	**6**	**13**	**17**	**20**

■平均

[MC]

5 [×] **2** [=] [M+]

15 [×] **4** [=] [M+]

...

45 [×] **3** [=] [M+]

[MR] [÷] **20** [=] ⇒ 26

■分散

まず 2 乗の平均を求める

5 [×] [=] [×] **2** [=] [M+]

15 [×] [=] [×] **4** [=] [M+]

...

45 [×] [=] [×] **3** [=] [M+]

[MR] [÷] **20** [=] ⇒ 815

次にこの値から平均の 2 乗を引くのだが，次のように順序を反対にしたほうが電卓で計算しやすい．

26 [×] [=] [−] **815** [=]

⇒ −139

負号を外して，分散 = 139

標準偏差は $\sqrt{139} = 11.79$

D.2　スプレッドシートで統計計算

統計計算のための機能は，Microsoft の Excel などの表計算ソフトにひと通り備わっている．これらを利用して第 1 章で扱った計算を実際に理解するために，数式に沿って一から計算する方法を示す．

CSV ファイルでデータ入力

第 1 章に出てきた表 1.1(p.5) のデータをスプレッドシートに入力するものとする．このとき，セルに直接打ち込むよりも，エディタで別のファイルに予め打ち込んでから作業するほうが便利な事が多いので，そのやり方を説明する．

データをファイルに入力して保存した後，表計算ソフトに読み込ませる．手順は次の通りである．

- 適当なエディタを開く．Windows なら メモ帳，Mac なら テキストエディットが使える．[*2]
- 数値を**半角で** 1 行に 1 つずつ入力して改行する．
- 最後のデータの後も改行を忘れずに入れる
- 保存する．ファイル名はなんでもよいが，拡張子は `.txt` としておく
- 保存したファイルの拡張子を `.csv` に変更する（警告が出るが気にせず変更する）
- ファイルのアイコンをダブルクリックする
- 表計算ソフトが立ち上がって，右図のように A 列の `A1` から `A100` までのセルにデータが入っている．

	A	B	C	
1	43.6			
2	45.2			
3	45.4			
4	45.8			
5	47.2			
	...			
99	64.2			
100	64.6			

このやり方をとるのは，ふつうのデータ入力では，ウェブ画面や電子ファイルからデータを取り込むことが多いからである．その場合，コピー＆ペーストでテキストファイルにデータを書き込むことで，正確な入力と省力化が可能になる．印刷されたデータを使うのであれば，上の手間を掛けずにデータをセルに直接入力するだけでも構わないのだが，練習のためにこの方式でやってみよう．

[*2] Mac のテキストエディットの初期設定ではリッチテキスト形式になっているので，「環境設定」メニューを開き，「標準テキスト」をチェックして，標準テキストに変更しておくこと．

平均を求める

A101 のセルに `=SUM(A1:A100)/100`　と入力する．次の計算を実行したことになり，平均値が表示される．

$$\mu = (43.6 + 45.2 + \cdots + 64.5)/100$$

あるいは `=AVERAGE(A1:A100)` としてもよい．`AVERAGE` は平均を求める関数である．

分散を求める

- 平均は上のやり方で求めておく
- B1 のセルに `=A1^2` と入力する
- B1 のセルをクリックして「コピー」する
- B2 から B100 までのセルを範囲指定して「貼り付け」する
- B101 のセルに `=SUM(B1:B100)/100` と入力する
- C101 のセルに次のように入力すれば，分散が得られる．
  `=B101-A101^2`
- 標準偏差を求めるには，C102　のセルに，`=SQRT(C101)` と入力すればよい．

分散=2 乗の平均から平均の 2 乗であることを使っている (p.8 を参照).

度数分布表を使う

今度は度数分布表から統計量を計算しよう．第 1 章に出てきた表 1.2(p.16) のデータを，次のように入力する．

- エディタを開く
- 数値データのうち，階級値と度数のペアを，カンマ ’,’ で区切って 1 行に入力する
- テキストファイルとして保存する
- ファイルの拡張子を .csv に変更したあと，ダブルクリックしてスプレッドシートに読み込ませる．これで右図のようにシートに入力されている．

	A	B	C	
1	44	1		
2	46	3		
3	48	6		
4	50	9		
	...			
10	62	2		
11	64	2		

度数分布から平均を求める

- C1 のセルに `=A1*B1` と入力する

- C1 のセルをクリックして「コピー」する
- C2 から C11 までのセルを範囲指定して「貼り付け」する
- C12 のセルに次にように入力すれば，平均が得られる
  `=SUM(C1:C11)/SUM(B1:B11)`

度数分布から分散を求める

- 平均は上のやり方で求めておく
- D1 のセルに `=A1^2*B1` と入力する
- D1 のセルをクリックして「コピー」する
- D2 から D11 までのセルを範囲指定して「貼り付け」する
- D12 のセルに次にように入力すれば，平均が得られる
  `=SUM(D1:D11)/SUM(B1:B11)-C12^2`
- 標準偏差を求めるには D13 のセルに `=SQRT(D12)` と入力すればよい

表計算ソフトの「分散」は標本不偏分散

一般的なスプレッドシートのマニュアルには分散を計算するための関数として VAR が用意されている．これを使うと，たとえば A1 から A100 までのセルのデータの分散は `=VAR(A1:A100)` で求められることになりそうだ．

ところが，実際にその通りに実行してみると，上で紹介したやり方で得られた値よりも微妙に大きな値が得られる．実はこの違いは，一般の表計算アプリケーションにおける「分散」の関数が標本不偏分散を与えるようになっており，また，「標準偏差」を表す関数 STDEV も，標本不偏標準偏差を与える仕様になっている．

このような事情は，専門の統計パッケージでもあることなので，くれぐれも，定義を知らないままで関数を使って，分散には区別があることを知らないで「分散を求めました」という過ちは避けるべきである．

D.3　統計計算のためのフリーソフト

序文でも述べたように，本書は初等的な統計学について，なるべくごまかすことなく解説することを主眼においている．これだけの内容を理解していれば，社会生活において出会うさまざまの確率的な事象，あるいは公表される各種の統計に対して十分な基礎になるはずだ．

しかしながら，研究や実務で実際に統計を駆使しようとする場合，統計処理は膨大な数値計算を必要とするものなので，コンピュータを使った統計処理が必須である．したがって，基礎の理解を前提として，統計処理のためのソフトウェアを活用することが，各分野での応用を試みる人には求められる．

そのさいに，前節で挙げた表計算アプリケーションを使うことは好ましくない．その理由としては，

- これらのアプリケーションは数値演算の精度が低く，統計処理における大きなデータを使う複雑な処理にはもともと適していない．
- これらのアプリケーションが出力するグラフや画像はきわめて品質が低く，プレゼンテーションや印刷に用いるとかなり見劣りがする．特に出版用の図版にはまったく使い物にならないといってよい．

といったことが挙げられよう．最後の点は，本質とは関係ないと思われるかも知れないが，プレゼンテーションの質を高めることは決して軽視すべきではない．

そのような状況を踏まえた上で，統計の実務家にも大きく浸透してきた統計処理パッケージとして R がもっとも勧められる製品である．また，最近では数学分野につよい汎用のプログラミング言語の Python も頭角を現している．

R と Python の共通の特徴

フリーソフトで無料で入手できる　フリーソフトというのは，単に無料のソフトという意味ではなく，ソースが公開されて，自由に複製したり改変したりできるソフトウェアのことだ．もちろん無料で公開されていて，だれでも自由にダウンロードして自分のパソコンにインストールすることができる．

多くのプラットフォームで利用可能　Windows, MacOS, Linux など，一般に使われている多くの OS で使うことができる．

多種多様な統計手法に対応している　R は現代において使われる多種多様な統計処理が可能であり，現代的な統計手法を含むあらゆる分野で利用されている．しかも専門領域によっては，さらに特化したパッケージが開発されて公開されている．

プログラミング言語であり，処理に汎用性がある プログラミング言語であるのできわめて柔軟に処理を組み立てることができる．「プログラミングはむつかしそうで手が出ない」という人もいると思うが，最初のソースを書いて実行するまでが，なれない人には苦労するところである．一旦慣れてしまえば Excel のマクロを組むよりも楽だし，技術の拡張性も高い．

高品質のグラフィックス機能 R はデフォルトのままでも多彩で見栄えのするグラフィックス機能を持っていて，さまざまのグラフや絵を作成することができる．さらに ggplot2 のような優れたグラフィックスを可能にしたパッケージも利用できる．
Python も matplotlib などのグラフィックス用ライブラリを持っていて，きわめて美しい可視化ができる．

■R と Python のちがうところ

R は統計計算に特化して開発された言語であり，特にライブラリを使わなくてもかなりの統計処理ができる．フリーの統計ソフトの定番として使われてきたために，膨大なライブラリの資産が利用できる．

一方，Python は汎用プログラミング言語であり，画像処理などきわめて広範な利用がなされている．統計計算のためには，そのためのライブラリを読み込んで利用するが，それで手間がかかるというわけではない．

R/Python の情報源

R と Python について，初心者が情報を得るためのウェブサイトと本を紹介しておく．これ以外にも膨大な情報があるがまずはこれで十分と思う．

The R Project for Statistical Computing 「R プロジェクト」つまり本家のサイト（英文）．
本体，および各種パッケージのダウンロード
http://www.r-project.org/

The R Manuals 本家サイトで公開されているマニュアル群（英語）．HTML による情報だけでなく PDF ファイルも提供されており，細かい目的ごとに分かれているので，英語が苦手でなければ使いやすい．
http://cran.r-project.org/manuals.html

RjpWiki R に関して日本語で情報交換することを目的とした Wiki サイト．筑波大学大学院人間総合科学研究科の岡田昌史氏によって運営されている．
http://www.okadajp.org/RWiki/

R-Tips 舟尾暢男氏が作成したコンテンツを収録したサイト．農業・食品産業技術総合

研究機構の竹澤邦夫氏によって管理されている．

http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

「The R Tips 第 3 版 データ解析環境 R の基本技・グラフィックス活用集 」 舟尾暢男，オーム社 (2016)

上記の舟尾氏による R-Tips サイトのコンテンツを，著者自身が大幅に加筆したもの．非常に分厚いが，その分きわめて丁寧に書かれていて読みやすい．電子ブックも出版社から購入可能なので，そちらのほうが使いやすいかもしれない．

「R による統計解析」 青木繁伸，オーム社 (2011)

この本はまとまりがよく，最初から読むには適している．電子ブックも出版社サイトから購入可能である．

https://www.ohmsha.co.jp/book/9784274067570/

「R による統計解析の基礎」 中澤港，ピアソン・エデュケーション (2003)

絶版になった本を著者が自分のサイトで公開しているもの．統計の基礎知識とともに演習問題を載せた理論と実用性のバランスのとれた良書．

http://minato.sip21c.org/statlib/stat-all-r9.pdf

Python.org Python のオフィシャルサイト．英文であるが，日本語のドキュメント類も含まれている．

https://www.python.org/

Python 3.x.x ドキュメント 上記サイトにある Python 3[*3] のチュートリアル，ライブラリーリファレンス，言語リファレンス等のコンテンツからなる．日本語で読める

https://docs.python.org/ja/3/index.html

「Python ではじめるプログラミング」 小波秀雄，インプレス (2019)

Python のをセミナー形式で初心者向けに解説してあり，統計分析のためのライブラリにも触れている．

「いまさら聞けない Python でデータ分析 多変量解析，ベイズ統計分析」 岡本安晴，丸善出版 (2019)

Python によるデータ分析の基礎を紹介し，数学ライブラリの使い方，可視化，多変量解析，ベイズ統計分析に触れている．

[*3] 現時点では Python の主要なバージョンは 2 と 3 である．Python3 は 2 よりも優れた仕様と能力をもつ（と小波は信じている）が，OS の中でも Python が使われていて，そちらは古いバージョンになっている．困ったことに両者の互換性は乏しいので，2 つのバージョンをうまく使い分ける必要がある．

付録 E

解答と解説

■第 1 章　データの整理と表現

問題 1–1

▼平均

$$\mu = \frac{1}{100}(43.6 + 45.2 + \cdots + 64.6) = 54.461$$

答え： 54.46

▼分散, 標準偏差　次のように式 (1.6) を使うと，平均からの偏差をいちいち求める必要がなくなり，計算の手間が省ける．

$$\sigma^2 = \frac{1}{100}(43.6^2 + 45.2^2 + \cdots + 64.6^2) - 54.461^2 = 17.841$$

答え： 分散 17.84, 標準偏差 4.224

上の計算では計算途中の数値の桁数を大きく取っていて，最後の答えのときだけ指定された有効数字に丸めていることに注意．有効数字と数値の丸め誤差については付録 C を参照のこと．

分散の定義式 (1.4) をそのまま使うと，次のようになる．結果は同じになるはずだが，若干誤差が出やすい．

$$\sigma^2 = \frac{1}{100}\left\{(43.6 - 54.461)^2 + (45.2 - 54.461)^2 + \cdots + (64.6 - 54.461)^2\right\} = 17.84$$

$$\sigma = \sqrt{17.84} = 4.2237\ldots$$

問題 1–2

0 は $n(1-p)$ 個，0 は np 個ある．つまりデータ列は次のようになるわけだ．

$$\overbrace{0,0,\ldots,0}^{n(1-p)},\overbrace{1,1,\ldots,1}^{np}$$

式 (1.2) を使って平均を求めると，次のようになる．

$$\mu=\frac{1}{n}(\overbrace{0+0+\ldots+0}^{n(1-p)}+\overbrace{1+1+\ldots+1}^{np})=\frac{np}{n}=p$$

あるいはこのデータを 階級値 0, 1 だけをもつ度数分布とみなすと，17 ページにあるように，平均は階級値とその割合の積の和に等しい．0 の割合は $1-p$, 1 の割合は p であるから答えはとても簡単である．

$$\mu=0\times(1-p)+1\times p=p$$

分散を求めるには，2 乗の平均から平均の 2 乗を引くやり方が便利だ．

$$\sigma^2=\frac{1}{n}(\overbrace{0^2+0^2+\ldots+0^2}^{n(1-p)}+\overbrace{1^2+1^2+\ldots+1^2}^{np})-\mu^2=p-p^2$$

標準偏差は正の平方根をとって $\sqrt{p-p^2}$ となる．

問題 1–3

体重の 2 乗の和を求める．式 (1.7) にならって，次のように書いてみよう．

$$2\text{ 乗の和}=44.0^2+\overbrace{46.0^2+46.0^2+46.0^2}^{3}+\overbrace{48.0^2+48.0^2+\cdots+48.0^2}^{6}+\cdots$$

したがって 2 乗の平均は，次のようにして求めることができる．

$$\frac{2\text{ 乗の和}}{\text{総人数}}=\frac{44.0^2\times1+46.0^2\times3+\ldots+64.0^2\times2}{1+3+\ldots+2}=\frac{298308}{100}=2983.08$$

これから先に求めておいた平均の 2 乗を引いてやれば，分散が求められる．

$$2983.08-2965.89=17.19$$

問題 1–4

メジアンは 437 万円であるから，その半分の 218.5 万円の所得のところのパーセンタイルを求めればよい．これは 13.6% の割合を含む 200〜300 万円の階級に属していて，階級

の左端のパーセンタイルは $6.4 + 12.6 = 19.0$ % になっている．また階級の幅は 100 万円である．これらから計算して，次のように求めることができる．

$$19.0 + \frac{13.6(218.5 - 200)}{100} = 21.516$$

つまり，全世帯の所得分布の中でメジアンの半分以下の所得を得ているのは，21.5% となる．

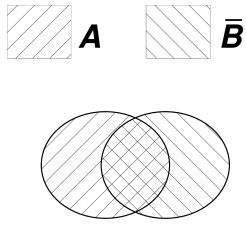
問題 1–5

度数分布表から計算すると，平均は 7.3 歳，メジアンは 5 歳となる．幼稚園に 7.3 歳の人がいることは考えられないので，この平均値は実態を反映しておらず，少数の大人（ゆか先生，あゆみ先生，お母さん先生，園長先生，おばあちゃん先生，おじいちゃん先生というところかも）の年齢に引っ張られている．一方で，むしろメジアンのほうが 5 歳を中心とした構成になっていることをよく表現していることになる．

■第 2 章　初等的な確率論

問題 2–1

下図で A と $\overline{B}$ のいずれにも含まれる領域が $A - B$ であることは一目瞭然．



問題 2–2

ベン図を見ると，A, B の排他的論理和は $A - B$ と $B - A$ の和集合となっている．従って，次のように表される．

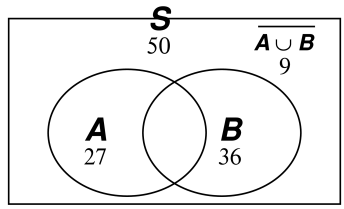
$$(A - B) \cup (B - A)$$

さらに問題 2–1 の式を使えば次のようになる．

$$(A \cap \bar{B}) \cup (\bar{A} \cap B)$$

問題 2–3

1. 図のようになる．



2. 全体集合 S の要素数は 50 である．S に含まれる要素のうち A でも B でもない要素の集合，つまり $\overline{A} \cap \overline{B} = \overline{A \cup B}$ の要素の数は 9 なのだから，$A \cup B$ の要素数は 41．つまり $27 + 36 = 63$ は重なりを含めた数で，真の合計は 41 なのだから，$A \cap B$ に含まれる人数は $63 - 41 = 22$ となる．

問題 2–4

A を感染者であるという事象，B を非感染者である事象とすると，A と B は互いに排反で，かつ標本空間を尽くしている．E を検査で陽性になる事象としよう．問題は $P(A|E)$ を求めることである．

題意より，$P(E|B) = 0.015$, $P(E|A) = 1 - 0.005$, $P(A) = 0.02$, $P(B) = 1 - P(A) = 0.98$ であるから，

$$P(A|E) = \frac{P(A)P(E|A)}{P(A)P(E|A) + P(B)P(E|B)} = \frac{0.02 \times 0.995}{0.02 \times 0.995 + 0.98 \times 0.015}$$
$$= 0.575\ldots$$

よって約 58% ．この段階では本当に感染しているかどうかは半々に近いというわけだ．

問題 2–5

連続して同じ数が出現する確率は 1/10 だから，0 から 9 の数字をランダムに 50 書いたとすると，5 回程度は現れる．3 個連続する確率も 1/2 だから，あってもおかしくはない．

問題 2–6

A を黒である事象，B を偶数 (2,4,6,8,10) である事象とする．ただし J, Q, K のカードは数字札とはみなさないこととする．したがって $P(A) = \frac{1}{2}, P(B) = \frac{5}{13}$．一方，黒でありかつ偶数であるという事象 $A \cap B$ は，スペード，クラブの 2,4,6,8,10 の 10 枚のカードの出現に対応する．よって $A \cap B = \frac{10}{52}$．結局

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{5}{13} - \frac{10}{52} = \frac{36}{52} = \frac{9}{13}$$

問題 2–7

ここでは閏年を考慮しないものとする．クラスに n 人いて，誕生日がだれも一致しない確率を p_n とする．すると一致する人がいる確率は $1-p_n$ で計算できる．

いま，1 人いるところにもう 1 人が来たとしよう．最初の 1 人によって 365 日のうち 1 日は誕生日に当たっているから，2 人目がそれと一致しない確率 p_2 は $364/365$ となる．次にもう一人が来たら，すでに 2 人がいて誕生日が一致していないのだから，3 人目の人も一致しない確率は $p_3 = 363/365$．これを続ければ n 人の誕生日が一致しない確率は次のようになる．

$$p_n = \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \cdots \frac{365-n+1}{365}$$

電卓などを使って実際に計算していくと，$n = 20, 21, 22, 23$ のときの値は $0.589, 0.556, 0.524, 0.493$ となるので，23 人の時に誰か誕生日が一致する人がいる確率は $1 - 0.493 = 0.507$ となって $1/2$ を超える．

ちなみにこの答の人数は，たいていの人が常識的に考えるよりもかなり小さくて意外な感じを与える．そこで「誕生日のパラドックス」と呼ばれることがある．

■第 4 章　二項分布

問題 4–1

二項分布が成立するので，式 (4.1) で $p = 2/3, n = 8$ として計算する．関西出身が 4 人未満だから $x = 0, 1, 2, 3$ について確率を求めて足し合わせればよい．

$$ {}_8C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^8 + {}_8C_1 \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^7 + {}_8C_2 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^6 + {}_8C_3 \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^5 $$
$$ = 0.088 $$

問題 4–2

この問題は，3 通りの事象があって，6 回の試行でそれぞれが 2 回ずつ実現する確率とみることができるので，多項分布の式が使える．式 (4.4) で，$n = 6, x_1 = 2, x_2 = 2, x_3 = 3, p_1 = 1/3, p_2 = 1/3, p_3 = 1/3$ とおいて，

$$f(2,2,2) = \frac{6!}{2!2!2!} \times \left(\frac{1}{3}\right)^6 = \frac{10}{81}$$

と計算すればよい．

問題 4–3

このように比較的小さい期待値で事象が発生する現象にはポアソン分布が有効である．まずある週が無欠席になる確率 p を求めよう．式 (4.5) に $\mu = 2.5, x = 0$ を代入して，

$$p = \frac{2.5^0}{0!}e^{-2.5} = e^{-2.5} = 0.0821.$$

求めるのは $p^2 = 0.0067$ となる．

■第 5 章　正規分布

問題 5–1

正規分布表から，$\Phi(z) = 0.75$ となるのは $z = 0.67$ である．一方，データから計算して $\mu = 54.46,\ \sigma = 4.22$ となるから，$\mu \pm 0.67 \times 4.22 = 51.6, 57.2$ を求めればよい．この範囲に入る人数を実際に数えてみると 49 名となっているから，予測は正しいことが分かる．

問題 5–2

偏差値の定義から，偏差値 58 に相当する点数は，標準正規分布で次の z に相当する．

$$z = \frac{58 - 50}{10} = 0.8$$

正規分布表から $\Phi(z) = 0.788$ が求められるので，この人の上位にいる人の数は割合にして $1 - 0.788 = 0.212$．したがって $12000 \times 0.212 = 2544$ となるから，2540 番目ぐらいに位置することになる．

問題 5–3

二項分布では $\mu = np,\ \sigma = \sqrt{np(1-p)}$ である．従ってこの集団については，$\mu = 9.6,\ \sigma = \sqrt{24 \cdot 0.4 \cdot 0.6} = 2.4$ となる．一方，この二項分布を正規分布とみなすとき，9 人以上 12 人以下は $z_1 = 8.5$ から $z_2 = 12.5$ までの区間とみなされる（半整数近似）．そこで $z_1 = (8.5 - 9.6)/2.4 = -0.458$ $z_2 = (12.5 - 9.6)/2.4 = 1.208$ であるから，正規分布表から $\Phi(0.46) = 0.6772, \Phi(1.21) = 0.8869$ を探す．z_1 と z_2 に挟まれる区間の面積は $(\Phi(0.46) + \Phi(1.21) - 1) = 0.6772 + 0.8869 - 1 = 0.5641 \approx 0.56$ とすればよい．ここでは **$\Phi(1.21) - \Phi(0.46)$ としてしまう誤り**を犯しがちなので注意．

問題 5–4

全数が n, 確率 p がちょうど 1/2 であるから，女（男でもよい）の子の数の期待値は $\mu = np = 40000 \times \frac{1}{2} = 20000$ となる．また，標準偏差は $\sigma = \sqrt{np(1-p)} = 100$．女の子の数 x が全体の 51% であったとすると $x = 40000 \times 0.51 = 20400$ であるから，標準化変換によって

$$z = \frac{x - \mu}{\sigma} = 4$$

となる．同様にして 49% からは $z = -4$ が得られる．正規分布表を調べると，$\Phi(4) = 0.9999683$ であるから，女の子の数が $-4 < z < 4$ の範囲から外れる確率は $(1 - 0.9999683) \times 2 = 6.3 \times 10^{-5}$ となる．

このように，全体の数が大きい場合には，事象が実際に実現する回数は数学的な期待値に近い値になることが確実になる．これは確率における**大数 (たいすう) の法則**と呼ばれている．

なお，実際の男女の出生比率は 1 : 1 ではなく，男の子のほうがやや多い．

問題 5–5

1. 0.71, 2. 0.98

この問題では，400 人のときには半整数補正の効果が 3% 程度ある．2000 人の時にはほとんど影響しない．

問題 5–6

区間 $[0, 1]$ で一様連続分布する確率変数の平均は $1/2$，分散は $1/12$ であるから (5.1.6 節参照)，この分布に従う 12 個の変数の和を S とすると，S の平均は $1/2 \times 12 = 6$, 分散は $1/12 \times 12 = 1$ となる．その理由については 3.5 節参照のこと．

したがって，$S - 6$ という値は平均が 0, 標準偏差が 1 の連続分布になる．一方，多数の独立な確率変数を足し合わせると正規分布に近づくことが中心極限定理で保証されているから，$S - 6$ は $N[0, 1]$ の標準正規分布に従うことになる．

■第 6 章　無作為抽出と標本分布

問題 6–1

このように整然と区画を選ぶと，もともと畑全体に存在していた何らかの傾向，たとえば縁にある区画が余計に選び出されるので作物の生育環境の偏った区画を選んでしまうといった種類の偏りを反映してしまう可能性がある．すべての区画に 1 から 100 までの番号を振っておいて，乱数を使って 20 の区画を選び出すのが正しい方法である．このさい，たとえば隣同士が選び出された場合にそれは避けるといった，事後の人為的な調整も行わない方がよい．

問題 6–2

ある本を買う人は，その本の内容に関心があり，かつそれを肯定的にとらえる傾向があると考えられる．また，その本を読むことによって著者の主張を受け入れる可能性も高い．したがって，読者アンケートによる調査は，信頼できる客観的な調査とはいえない．

ちなみに，血液型が性格と関係があるという説に対しては何人かの心理学者による調査があるが，それを肯定する結果はほとんどなく，また科学的にも関連性があるという根拠

はない．ただし，最近の韓国での研究で，B 型の男性について有意な差がみられるという結果があった．これについては韓国で血液型ブームが起こり，歌や映画で B 型が嫌われる風潮を扱ったりしたことによって，思い込みによる回答が無視できないほどにまでなったことによるものであると考えられている．

問題 6–3

題意から，母平均は $\mu = 54.2$, 母分散は $\sigma = 0.22^2 = 0.0484$ である．標本のサイズは $n = 10$ なので，式 (6.3) より，標本平均 $\bar{X}$ の平均値は 54.2, 式 (6.4) より標本平均の分散は $V[\bar{X}] = \sigma^2/n = 0.00484$ だから，その標準偏差は平方根をとって 0.0696 となる．標本平均の 54.1, 54.3 をこれらを使ってそれぞれ標準化変換すると，

$$z_1 = \frac{54.1 - 54.2}{0.0696} = -1.44, z_2 = \frac{54.3 - 54.2}{0.0696} = 1.44$$

となる．正規分布表から $\Phi(1.44) = 0.92506$ を読み取って，この 2 つの値の間に入る確率を求めると，確率として 0.85 が得られる．

問題 6–4

24.4 g^2 (解法はテキスト参照のこと)

■第 7 章　推定

問題 7–1

この問題には二種類の解法が考えられる．

1. **支持を 1, 不支持を 0 として標本平均と標本分散を求めて計算する方法：**
   支持を 1, 不支持を 0 とすると，100 人の標本平均は $\bar{X} = \frac{1}{100}(1\times30+0\times70) = 0.3$
   標本分散は $s^2 = \frac{1}{100}((1-0.3)^2 \times 30 + (0-0.3)^2 \times 70) = 0.21$.
   したがって母分散の不偏推定値は $\sigma^2 = \frac{n}{n-1}s^2 = 0.2121$ より，$\sigma = 0.4605$．これより $\bar{X}$ の分布の標準偏差は $\sigma/\sqrt{100} = 0.04605$ となる．
   求めるのは平均値を挟む 95% の信頼区間であるから，97.5% 点を使って，$0.3 \pm 0.04605 \times 1.960$ を求めて，内閣支持率は 0.21 〜 0.39 と推定される．
2. **支持者の数に関する二項分布を使う方法：**
   標本における支持率を p とし，$n = 100$ 人のうち支持している人の数を x とすると，x は二項分布に従うから標本平均は $\bar{X} = np = 30$, 標本分散は $s^2 = np(1-p) = 100 \times 0.3 \times 0.7 = 21$ である．これから母分散 σ^2 の不偏推定値は $\sigma^2 = \frac{n}{n-1}s^2 = 21.21$ となる．そこで 95% 信頼区間の幅は

$$\sigma \times 1.960 = \sqrt{21.21} \times 1.960 = 9.027$$

つまり，30 ± 9.027 が信頼区間に入るので，人数の 100 で割って内閣支持率は 0.21 〜 0.39 と推定される．

問題 7–2

標本が小さいので $t-$ 分布を使う．標本の大きさは $n = 12$ であるから，自由度は $\nu = n - 1 = 11$, 99% 信頼区間を計算するには $\alpha = 0.005$ として，3.106 を表から読み取る．ここで Student の変数 T を計算すると，

$$T = \frac{\sqrt{n-1}(x-\mu)}{s} = \frac{\sqrt{11}(16-\mu)}{4}.$$

ここから $T = 3.106$ および $T = -3.106$ として 推定値 μ を求めればよい．結果は 12.3 〜19.7 となる．

問題 7–3

この場合には大標本の取り扱いをしてよい．したがって母平均の推定は 7.3.2 節の議論をもとに式 (7.8) と同様にして考えればよい．99% 信頼区間は次のようになるので，

$$\left[16 - 2.576 \times \frac{4}{\sqrt{80-1}}, 16 + 2.576 \times \frac{4}{\sqrt{80-1}}\right]$$

計算して，14.8〜17.2 を得る．

この結果を問題 7–2 の結果と比べると，推定区間の幅はかなり小さくなっていることがわかる．これは当然で，標本が小さいほど，データから推定できる幅は広く，つまりあいまいにならざるを得ないということを反映しているのである．

■第 8 章　仮説と検定

問題 8–1

実の総数は $n = 144$．このとき 1 代目が黄色と白それぞれの純系種であったとすると，メンデルの法則に従って黄色の実ができる確率は $p = 3/4$ である．したがって黄色の実の数の期待値は $np = 108$，標準偏差は $\sigma = \sqrt{np(1-p)} = 5.196$ となる．実際に得られたのは 119 であるから，これを標準化変換してみると，

$$z = \frac{119 - 108}{5.196} = 2.11$$

となる．この値を正規分布のパーセント点と比較すると，危険率 5% であれば 97.5% 点 (あるいは $\alpha = 0.025$) の 1.960 との比較になるので，仮説は棄却される．つまり，1 代目は純系でなかったという判断になる．もちろんこの判断は 20 回に 1 回程度以内の間違いを含むことを許容しているのである．

一方，危険率 1% であれば 99.5% 点（あるいは $\alpha = 0.005$) の 2.576 との比較になるので，仮説は棄却されない．つまり，1 代目は純系であったという仮説を棄却できない．ただしこの判断は，上の判断よりも厳しい基準を設けてずっと用心深く振る舞っているのであり，2 つの判断の間に矛盾はない．

問題 8–2

1. 母平均 $\mu = 12.6$ g, 母標準偏差 $\sigma = 1.9$ g の母集団から 12 個抽出を行ったときの標本平均の分布は，(標本平均の) 平均が 12.6 g,(標本平均の) 標準偏差が $1.9/\sqrt{12}$ の正規分布になる． これを使って与えられたデータから得られた実際の標本平均 $\bar{X} = 12.725$ を標準化変換すると，0.228 となる．正規分布で $\alpha = 0.025$ となる点は 1.960 であるから，標本平均の値は棄却域に落ちない．つまり，仮説は棄却されない．

2. 式 (8.2) を使って Z を計算してみる．

$$Z = \frac{1}{\sigma^2}\left((11.78 - \bar{X})^2 + (12.92 - \bar{X})^2 + \cdots\right) = 25.62$$

この Z が母集団（現存種）から抽出されたデータによるものであるとした場合，自由度 $\nu = 12 - 1 = 11$ の χ^2 分布に従うことになるはずである．そこで危険率 5%(両側検定なので $\alpha = 0.025$) の棄却域を表から調べると 21.920 である．つまり Z の値は棄却域に落ちるので，仮説は棄却される．

これらの結果をまとめると，遺跡から出た貝は，質量は大体現存種とあっているのだが，個体差がかなり大きいという特徴があるので異なる種なのかも知れない．

問題 8–3

式 (8.3) に従って X を計算すると 1.78 が得られる．一方この値は自由度 $(2-1)(2-1) = 1$ の χ^2 に従う．一方付録の χ^2 分布表で $\nu = 1, \alpha = 0.005$ の値を調べると 7.879 であるから，X は 1% の危険率で棄却域の内側にある．したがって，独立であるという仮説は棄却できない．具体的にいうと，この病気にかかることと飲酒の習慣の間に相関があるとはいえないことになる．

■第 9 章　相関と線形回帰

問題 9–1

1. $\sigma_x^2 = 105.32, \quad \sigma_y^2 = 7.801, \quad \sigma_{xy}^2 = 24.40$
2. $\rho_{xy} = 0.851$
3. $a = 0.23, \quad b = 19.0$
4. $T = 3.98$ となる．t-分布を調べると，自由度 $8 - 2 = 6$ のときに $\alpha = 0.005$ の点は 3.707 である．したがって T は棄却域にあるので，仮説は棄却される．つまり，卵の質量と黄身の質量は独立ではないと検定できる．

索引