情報処理の応用 B

立正大学データサイエンス学部，成塚拓真

最終更新：2022 年 6 月 21 日

目次

第 1 章

イントロダクション

1.1 オーバービュー

1.1.1 講義の概要

本講義で扱う内容は，主に統計学の基礎とデータ解析の基礎である．高校から大学初年度で扱う統計学の入門的な内容を網羅的に学び，簡単なデータ解析ができるようになることを講義全体の目的とする．講義は複数のテーマからなり，基本的には講義 1 回で 1 テーマ（おおよそ本講義ノートの 1 セクションに対応）を扱う．各テーマでは，まず必要な統計学の知識を学んだ後，実データの解析例を示すという構成になっている．DS 学部では，1 年次から python のプログラミングを学ぶので，実データの解析を扱う際には python によるコーディングの例も示すようにする．（ただし，プログラミングの授業ではないので，詳細には立ち入らない．）

統計学に関するより厳密で高度な内容は 2 年次の必修科目である「統計学 I，II」で扱うことになる．よって，本講義はその前段階の橋渡し的な位置づけと捉えても良い．本講義は複数クラス開講科目であり，「情報処理の応用 A」が並列で行われる．メインテキストは A，B ともに同じであるが，細かい授業内容は異なる．レポートや評価基準についてはできる限り統一する．なお，「情報処理の基礎」とのつながりはあまりない．

講義で用いる資料は講義ノートとスライド，jupyter notebook（プログラミングの解説）であり，これらは随時更新する．講義のメインテキストは文献 [1][*1] であり，全体の構成はこれに基づくが，統計学の内容は文献 [2, 5, 6, 7] を参考にしている．

[*1] 以下からダウンロード可能：https://www.soumu.go.jp/toukei_toukatsu/info/guide/stkankyo.htm

1.1.2 講義スタイル

コロナウイルスの蔓延状況によって，対面（パターン I）とオンライン（パターン II）が混在する可能性がある．対面の場合は基本的に板書に基づいて進め，プログラミングの説明などは適宜スライドや jupyter notebook を用いて行う．python が動く PC が手元にあると良い．オンラインの場合は Zoom を用いた同時双方向授業を行い，基本的にはスライドを用意する予定であるが，場合によっては手書きになるかもしれない．

講義に関する連絡や講義ノート・スライドの共有にはポータルサイトの「オンライン授業」を用いる．対面授業に移った場合でも「オンライン授業」を用いる．

1.1.3 評価方法

レポート（60%）と授業への取り組み姿勢（40%）で評価する．試験は行わない．レポートは 2 回くらい？を予定し，講義で扱った実例についてレポートの形にまとめてもらう（予定）．提出されたレポートは添削して返却し，再提出してもらう（予定）．レポートは科学的な文書を書く訓練の意味も兼ねているので，ある程度書き方を指定する（次節参照）．

1.1.4 レポートの書き方

レポートとは，科学的な研究で得られた知見を，研究をやっていない人でも分かるようにまとめた “フォーマルな” 報告書である．よって，**見栄えが整っていることは大前提**であり，その上で内容が伝わるように書く．最低限以下を遵守する：

- **鉄則：“当事者でなくても分かるように書く．”**
- **PPDAC メソッドとレポートの章立ての対応は 1.3 節を参照．**
- Word，LaTeX，などを使って書いても良いが，書式は統一する（フォントや文字サイズを統一する，見出しをつける，数式エディタを使うなど）．
  ソフトを使いこなせないなら手書きのほうが良い．
- レポートの本文は（このテキストのように）フォーマルな文章で過不足なく記述する．（省略記号やメモ書き，パワポのスライドのような記述はレポートとして相応しくない．）
- 結果を表やグラフにまとめる場合は，本文でも詳細を説明する．
- 初めにレポートのタイトル，学籍番号・氏名を書く．
- ファイル形式は PDF（手書きの場合はスキャンして PDF 化する）．
- レポートは 1 つのファイルにまとめ，ファイル名は “**学籍番号_氏名**.pdf” などとする．

1.2 PPDAC メソッド

PPDAC メソッドとは Problem，Plan，Data，Analysis，Conclusion という科学的探求の手順を示したものである．カナダ・アメリカ・ニュージーランド等の学校教育で使用されている．本講義では，各テーマごとにデータ解析の実例を示すが，これらは PPDAC メソッドに沿っている．以下に各ステップで行われる探求プロセスをまとめる．

STEP 1: Problem

第 1 ステップでは，まず関心のあるテーマを決め，そこでの課題を明らかにする．また，課題から問題の構造を明確にし，具体的な研究仮設（リサーチクエスチョン）を設定する．

STEP 2: Plan

第 2 ステップでは，研究仮設を明らかにするための分析の計画を立てる．具体的には，計測すべきデータや統計資料を決め，その収集計画を立てる．

STEP 3: Data

第 3 ステップでは，実際にデータを取得し，整理する．

STEP 4: Analysis

第 4 ステップでは，収集したデータを実際に分析する．分析の具体例としては，以下が挙げられる

- 全体の傾向（分布）を見る
- 条件の違いなどによってデータをグループに分け，比較する
- 指標間の相関関係を見る
- 指標間の因果関係を見る
- 時間経過による変化を見る（時系列解析）
- 対象を分類する（クラスタリング）

STEP 5: Coclusion

第 5 ステップでは，分析結果に基づいた考察や提言を行い，同時に新たな課題を明らかにする．最後に，最初に立てた研究仮設に対して判断や結論を示す．

レポート・論文との対応

PPDAC メソッドによる問題解決によって得られた結果はレポートや論文の形にまとめることになる．通常，レポートや論文は，Introduction（導入），Method（方法），Result（結果），Discussion（考察），Conclusion（結論），という手順でまとめる．PPDAC メソッドとの対応関係はおおよそ以下のようになる：

$$
\begin{aligned}
\text{Problem} &\Longleftrightarrow \text{Introduction（導入，はじめに）} \\
\text{Plan, Data} &\Longleftrightarrow \text{Method（方法）} \\
\text{Analysis} &\Longleftrightarrow \text{Result（結果）} \\
\text{Conclusion} &\Longleftrightarrow \text{Discussion（議論）, Conclusion（結論）}
\end{aligned}
$$

第 2 章

記述統計学

本章では，**記述統計学**について学ぶ．記述統計学とは，調査や実験によって得られたデータを整理・要約し，データの性質や傾向を明らかにするための手法である．一言で言えば，手持ちのデータを集計するための方法である．一方，手持ちのデータをより大きな母集団からの標本と捉え，標本から母集団全体の特徴を調べる方法は**推測統計学**と呼ばれる．推測統計学については 5 章以降で扱う．

2.1　データの整理

2.1.1　質的データ・量的データと尺度

データには**質的データ**と**量的データ**の 2 種類がある．質的データとは，数値で表すことができず，あるカテゴリーに属していることやある状態にあることだけが分かるデータである．例えば，性別（男，女），学歴（大卒，高卒，中卒），天気（晴，曇，雨，雪），などは質的データである．一方，量的データとは，数値で表すことができるデータのことを指す．例えば，長さ，重さ，体積，面積，金額，温度，時間などは量的データである．

質的データは，どのような尺度で測定されたかという基準によって，さらに 2 つに分類できる．まず，「男・女」など，他と区別するためだけに用いる尺度を**名義尺度**と呼び，対応するデータをカテゴリカルデータと呼ぶ．カテゴリカルデータに対しては一切の計算が許されず，唯一できるのは数をカウントすること（度数や最頻値の計算）だけである．一方，「小・中・大」のように大小や前後が決まるような尺度を**順序尺度**と呼び，対応するデータを順序データと呼ぶ．順序データに対しては $>, =$ などの演算が許される．

次に，量的データも測定尺度によって 2 つに分類できる．まず，値の大小関係と値の差だけに意味があるような尺度を**間隔尺度**と呼び，対応するデータは間隔データと呼ばれる．間隔データは値同士の加減が許される．間隔データの代表例は，摂氏・華氏温度や時刻である．例えば，摂氏温度は水の融点を 0 ℃，沸点を 100 ℃としてその間を等分した尺度なので，値の大小関係と差に意味はあるが，比に意味はない．実際，4 ℃と 8 ℃を比較して 4 ℃

暑いということはできるが，2 倍暑いなどということはできない[*1]．一方，値の大小関係と値の差に加えて，値同士の比にも意味があるような尺度を**比率尺度**と呼び，対応するデータは比率データと呼ぶ．比率データは値同士の加減乗除が全て許される．比率データの代表例は身長，体重，年齢などである．例えば，身長 150cm と 180cm には「値の大小関係」があり，「値の差」も 30cm と意味がある．また 100cm と 200cm であれば，「後者は前者の 2 倍」であると解釈でき，比が意味を持つ．K（ケルビン）で表される絶対温度も比例データの例である．実際，絶対温度は値 0 が絶対的な意味を持ち，1K と 2K ではある量が実際に 2 倍になっているので比にも意味がある．

間隔尺度と比例尺度が見分けづらい場合は，「0 の値が相対的な意味しか持たない」場合が間隔尺度，「0 の値が絶対的な意味を持つ」（ある量が無いことを意味する）場合が比率尺度と考えると良い．例えば，摂氏温度や西暦が 0 だったとしてもそれらは無いわけではないが，身長や速度が 0 であるときは本当に無いので，前者は間隔尺度，後者は比例尺度の例である．

2.1.2　量的データの要約

四分位数と五数要約

15 個の量的データがあるとする．これを小さい順に並べたとき，図 2.1 のように 4 等分に分割できる．このとき，アを**最小値**，イを**第 1 四分位数**，ウを**中央値（第 2 四分位数）**，エを**第 3 四分位数**，オを**最大値**と呼ぶ．また，データを小さい順に並べたとき，左半分のデータを下位データ，右半分のデータを上位データと呼ぶ．ただし，データの数が奇数個の場合は中央値を除いて下位・上位に分ける方法を採用する（中央値を両方に含める場合もある）．このとき，第 1 四分位数 Q_1 は下位データの中央値，第 3 四分位数 Q_3 は上位データの中央値である．以上のようにデータのばらつきを 5 つの数で表す方法を**五数要約**と呼ぶ．また，第 3 四分位数と第 1 四分位数の差 $Q_3 - Q_1$ を**四分位範囲**と呼ぶ．なお，四分位は英語で quartile なので，各四分位数（イ，ウ，エ）を Q_1，Q_2，Q_3 と表すことが多い．

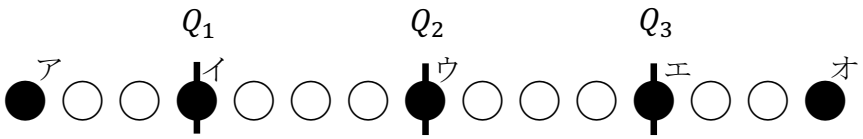

図 2.1. 15 個の量的データを小さい順に並べた例．

五数要約の求め方はデータの数が奇数個の場合と偶数個の場合で異なる．まず，データの数が奇数個の場合は最小値，最大値，中央値が自動的に定まる．一方，中央値を除いて下

[*1] 比例尺度である絶対温度で表すと，4 ℃は 277.15K，8 ℃は 281.15K であり，その比は 2 倍ではない [4]．

位データと上位データに分けると，それぞれが偶数個になる．この場合，下位データ，上位データそれぞれの中央にくる 2 つの値の平均値を第 1 四分位数，第 3 四分位数とする．

次に，データの数が偶数個の場合は中央にくる 2 つの値の平均値を中央値とする．下位データと上位データはそれぞれ奇数個に分かれるので，第 1 四分位数，第 3 四分位数は自動的に求まる．

以下に具体例を示す．

例）データが 9 個（奇数個）の場合

$$2\ 3\ 5\ 5 \mid 6 \mid 8\ 10\ 12\ 15$$

この場合，中央値は $Q_2 = 6$ となるので，これを除いて下位データと上位データに分ける．第 1 四分位数は $Q_1 = (3+5)/2 = 4$，第 3 四分位数は $Q_3 = (10+12)/2 = 11$ と求まる．四分位範囲は $Q_3 - Q_1 = 11 - 4 = 7$ である．

例）データが 10 個（偶数個）の場合

$$2\ 2\ 5\ 6\ 7 \mid 9\ 10\ 13\ 14\ 18$$

この場合，中央値は $Q_2 = (7+9)/2 = 8$ となる．また，第 1 四分位数は $Q_1 = 5$，第 3 四分位数は $Q_3 = 13$ と求まる．四分位範囲は $Q_3 - Q_1 = 13 - 5 = 8$ である．

箱ひげ図

五数要約の結果は図 2.2 のような図によって可視化できる．これを**箱ひげ図**と呼ぶ．箱ひげ図は以下の手順によって描く（これを**テューキーの方式**と呼ぶ）．

1. データの第 1 四分位数から第 3 四分位数の間に箱を描く．
2. 中央値の位置に線を引く．
3. 箱から箱の長さ（四分位範囲）の 1.5 倍を超えて離れた点（外れ値）を白丸で描く．
4. 外れ値ではないものの最大値と最小値から箱まで線（ひげ）を引く．

なお，外れ値を表示しない簡便な描き方もある．また，90 度回転させて横に描くことも多い．

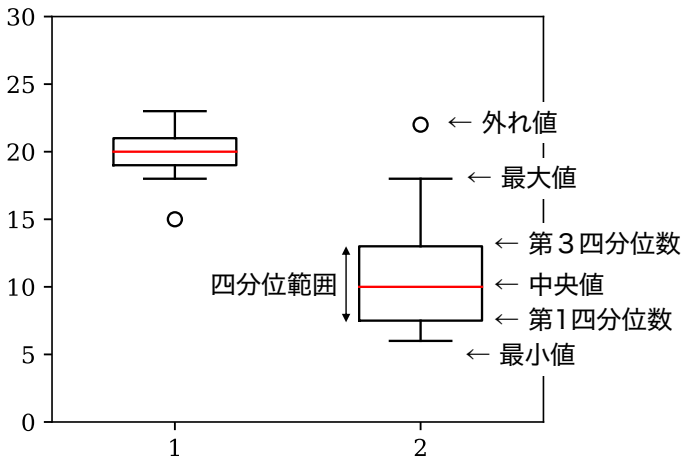

図 2.2. テューキーの方式による箱ひげ図の例.

ヒストグラム

データの分布の傾向（どの値がどのくらいあるか）を表す方法として，箱ひげ図ではデータを要約しすぎてしまい適切にその特徴を表せないことがある．そこで，より詳細に分布の傾向を可視化する方法として度数分布表やそれを可視化した**ヒストグラム（度数分布図）**がある．度数分布とは，値を 0 以上 10 未満，10 以上 20 未満などのいくつかの区間（**階級**，**ビン**）に分けてそれぞれの区間に含まれるデータの個数（度数）をまとめたもので，横軸に階級の代表値（**階級値**），縦軸に度数をとったグラフがヒストグラムである．なお，ヒストグラムの横軸には各階級の最小と最大を表示する場合と，階級値として階級の最小値や中央値を示す場合がある．また，縦軸には度数ではなく相対度数（度数/データ数）を取ることもある．

具体例として，ここでは Iris Dataset[*2] に含まれるアヤメのがく片の長さ，がく片の幅，花弁の長さ，花弁の幅のデータを用いる．表 2.1 はアヤメのがく片の長さのデータに対する度数分布表である．ここでは，各階級の中央値を階級値としている．また，度数，相対度数の他に，相対度数を足し合わせた累積度数も示している．

図 2.3 はアヤメのがく片の長さ，がく片の幅，花弁の長さ，花弁の幅のヒストグラムである．ヒストグラムの中でデータが集中している部分が山のようになっているとき，山が 1 つの場合には単峰性，2 つの場合には双峰性，それ以上の場合には多峰性と呼ぶ．図 2.3 では，がく片のヒストグラムは単峰性，花弁のヒストグラムは双峰性である．特に，多峰性のヒストグラムの場合には箱ひげ図によって可視化するとデータを要約しすぎてしまうため，適切にその特徴を表すことができない．この他にも，値の小さなところにデータが集中していて大きな値のところに少数のデータがあるとき，「右に裾を引いている」という．

[*2] Kaggle のウェブサイトからダウンロード可能：https://www.kaggle.com/uciml/iris

ヒストグラムの階級（ビン）の幅は大きすぎても細かすぎても分かりにくくなる．一般に，階級数は標本の大きさ（サンプルサイズ）N の平方根 $\sqrt{N}$ 程度が良いとされている．例えば，100 個のデータを含む場合は 10 程度の階級が望ましい．また，$1 + \log_2 N$ という公式も存在し，これを**スタージェスの公式**と呼ぶ．図 2.3 では，スタージェスの公式を用いて階級数を決めている．

表 2.1. アヤメのがく片の長さの度数分布表

最小（cm）	最大（cm）	階級値 v_i（cm）	度数 f_i	相対度数（%）	累積相対度数（%）
4.30	4.75	4.53	11	7.3	7.3
4.75	5.20	4.97	30	20.0	27.3
5.20	5.65	5.43	24	16.0	43.3
5.65	6.10	5.88	24	16.0	59.3
6.10	6.55	6.32	31	20.7	80.0
6.55	7.00	6.78	17	11.3	91.3
7.00	7.45	7.22	7	4.7	96.0
7.45	7.90	7.68	6	4.0	100.0

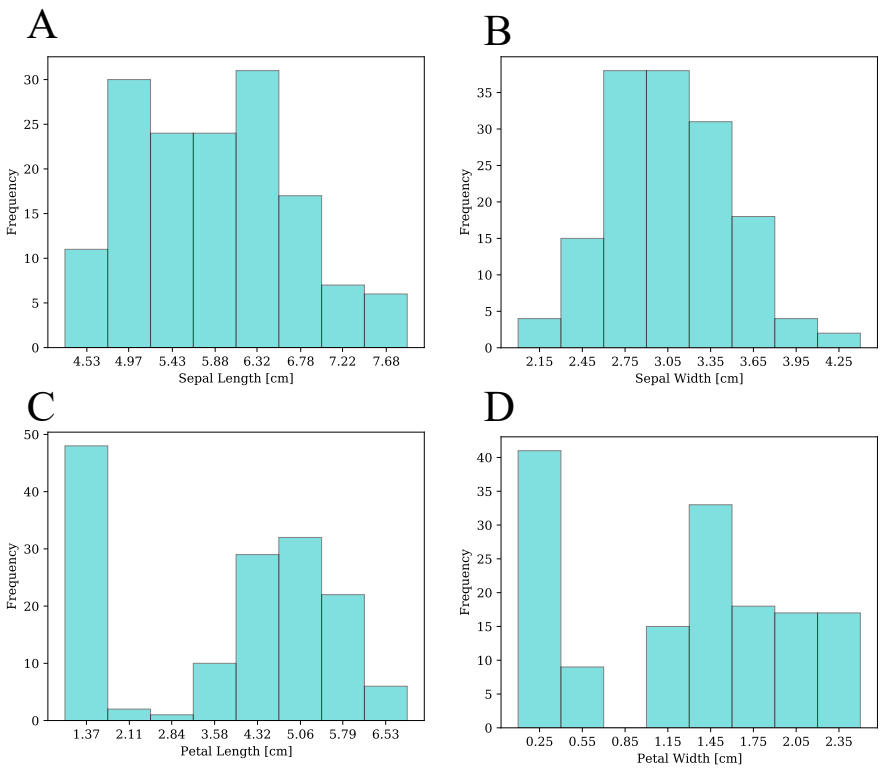

図 2.3. ヒストグラムの例．A:アヤメのがく片の長さ，B：がく片の幅，C：花弁の長さ，D:花弁の幅．

2.1.3 実例：夏の避暑地の気候の特徴〜夏の避暑地が快適な理由は？

日本への外国人旅行者は近年急増しているが，一方で，日本人の国内旅行者の動向を月別に見ると，表 2.2 のように月ごとに変動している．特に，5 月や 8 月は国内旅行者の数が突出して多くなっているが，これはゴールデンウィークや夏休みを利用して旅行する人が多いからである．

実習

表 2.2 のデータから折れ線グラフを作成せよ．

表 2.2. 2015 年の月別国内旅行者数（観光庁「2015 年旅行・観光消費動向調査」より）

月	1	2	3	4	5	6	7	8	9	10	11	12
旅行者数	4315	3620	5331	4456	6322	4693	4458	7177	5707	4647	4794	4952

STEP 1：Problem

ある高校に通う 5 人の高校生は，2015 年の夏休みにそれぞれ別の都市で過ごした．以下は日本の各都市についての気候に関する意見をまとめたものである．

- 軽井沢は東京と比べて過ごしやすかった
- 東京も今年は涼しい日もあったけど，すごく暑い日が多かった
- 熊谷は東京以上に暑かった
- 沖縄は暑かったけど，慣れてしまえば逆に過ごしやすかった
- 札幌は過ごしやすかったけど，大阪は東京と同じように暑かった

それぞれの場所で，本当に暑さに違いはあったのだろうか？特に，日本では，夏に避暑地を訪れる人が多いが，避暑地にはどのような特徴があるのだろうか？

STEP 2：Plan

気象庁の HP（http://www.data.jma.go.jp/gmd/risk/obsdl/index.php）には 1 日の平均気温，最高気温，最低気温，湿度などのデータが掲載されている．ここでは，1 日の最高気温，最低気温，湿度のデータを収集する．

収集したデータは五数要約や箱ひげ図によって傾向を調べる．また，夏の蒸し暑さを定量化した指標である**不快指数**を計算し，各都市の特徴を調べる．不快指数は気温を t，湿度を

H とすると

$$\text{不快指数} = 0.81t + 0.01H(0.99t - 14.3) + 46.3 \tag{2-1}$$

によって求められる．一般に，不快指数が 75 になると人口の約 1 割が不快を感じ，85 になると全員が不快になる（三省堂編集所，大辞林，三省堂 (1988)）．

STEP 3：Data

実習

- 気象庁の HP から 2015 年 8 月の各地点の 1 日の平均気温，最高気温，最低気温，湿度のデータ（csv ファイル）をダウンロードせよ．
- ダウンロードしたデータを python などで解析しやすいように加工せよ．

STEP 4：Analysis

まず，最高気温に着目する．収集したデータは，五数要約や箱ひげ図を使って特徴を整理することができる．

実習

- 各都市の最高気温のデータに対し，五数要約と四分位範囲を求めよ．
- 五数要約の結果から，各都市に対して並行箱ひげ図を作成せよ．

次に最低気温に着目する．

実習

- 各地点の最低気温のデータについて，並行箱ひげ図を作成せよ
- 各地点について，熱帯夜（最低気温が 25 ℃以上の夜）の日数を求めよ

最後に不快指数に着目する．

実習

- 式 (2-1) を用いて，6 地点の 2015 年 8 月 1 日から 31 日までの不快指数を計算せよ
- 各地点の不快指数のデータについて，並行箱ひげ図を作成せよ

STEP 5: Conclusion

解析の結果を基に，各都市の気候についてどのようなことが分かるか考察する．

実習

- 最高気温に対する並行箱ひげ図を基に，各地点の特徴について分かったことを次の観点からまとめよ．
  1. 東京や大阪のような大都市は避暑地と比べて暑い日が多いか？
  2. 避暑地として人気の高い軽井沢は高原にあるが，北海道とどのように違うか？
  3. 熊谷や沖縄は暑い地域として有名だが，それぞれで違いはあるか？
- 熊谷は最高気温は高いが，最低気温は東京や大阪と比べて低い．なぜこのような違いが出るのか考えよ．
- 不快指数を基に，各都市の特徴をまとめよ
- 軽井沢や札幌は夏の避暑地として人気が高い．その理由をまとめよ．
- その他，分析結果を元に自由に考察せよ．

2.2　特性値の活用

データの性質を定量的に表すための統計量（データに対して何らかの統計的な計算をして得られた数値）を**特性値**という．代表値，記述統計量，要約統計量などとも呼ばれる．ここでは代表的な特性値について説明する．なお，以下では n 個のデータを $(x_1, x_2, \cdots, x_n)$ と表す．

2.2.1　データの中心を表す特性値

算術平均

データの中心を表す特性値として最もよく知られ，よく用いられるのが**算術平均**であり，

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2-2}$$

と定義される．通常，与えられたデータから平均値を求めるにはこの式に当てはめれば良い．例えば，アヤメのがく片の長さのデータの場合，平均値は $\bar{x} = 5.843$ cm となる．算術平均は分布形状が左右対称に近いデータの場合にはデータの中心を表す量と捉えられるが，分布形状が極端に非対称な場合にはデータの中心を表す特性値としてふさわしくない．

算術平均は度数分布表から求めることもできる．具体的には，階級値を v_i，対応する度数を f_i，階級の数を k とすると，平均値は

$$\bar{x} = \frac{f_1 v_1 + f_2 v_2 + \cdots + f_k v_k}{f_1 + f_2 + \cdots + f_k} = \frac{1}{n}\sum_{i=1}^{k} f_i v_i \tag{2-3}$$

と表される．例えば，アヤメのがく片の長さのデータ（表 2.1）の場合，

$$\begin{aligned}\bar{x} &= \frac{4.53\times 11 + 4.97\times 30 + 5.43\times 24 + 5.88\times 24 + 6.33\times 31 + 6.78\times 17 + 7.22\times 7 + 7.68\times 6}{11+30+34+34+31+17+7+6} \\ &= 5.854\end{aligned}$$

となる．

以上の例を見て分かるように，度数分布表から求めた平均値はデータから直接求めた平均値と一致しない．これは，度数分布表から求めた平均値が近似値であるからである．度数分布表では，各階級の値を階級値で代表させているためこのようなことが起こるが，各階級の幅を十分小さく取れば近似の度合いは上昇する．

幾何平均（相乗平均）

算術平均に対して，幾何平均というものも存在する．これは，n 個のデータに対して，値の積の n 乗根をとったもので，

$$\bar{x}_g = \sqrt[n]{x_1 \times \cdots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i \right)^{1/n} \tag{2-4}$$

と定義される．定義より，幾何平均は正の数のみしか扱えず，さらに掛け算によって定義されるので比率データにしか適用できない．

幾何平均は成長率や倍率の平均を計算するときに用いられる．例えば，次のような事例を考える：

> 1 年目から 2 年目にかけての物価は対前年比 2 倍になり（100 円のものが 200 円になり），2 年目から 3 年目にかけての物価は対前年比 8 倍となった（200 円のものが 1600 円になった）．では，この 2 年間の物価の対前年比伸び率の平均はいくらか？

まず，算術平均を適用してみる．すると，物価の対前年比伸び率の平均は $(2+8)/2 = 5$ 倍となる．これは，1 年目に 100 円だったものが 2 年目に 500 円になり，さらに 3 年目に 2500 円になることを意味するので，実際よりも過大に見積もってしまう．そこで，次に幾何平均を適用してみる．すると，物価の対前年比伸び率の平均は $\sqrt[2]{2 \times 8} = 4$ 倍となる．これは，1 年目に 100 円だったものが 2 年目に 400 円になり，さらに 3 年目に 1600 円になることを意味するので，実際の金額と一致する．このように，倍率の平均値を計算する場合には，算術平均ではなく幾何平均を用いるのが妥当である．

中央値・最頻値

データの分布形状が歪んでいる場合，算術平均はデータの中心を表す特性値としてふさわしくない．例えば，

$$1, 1, 1, 1, 2, 3, 4, 5, 16, 20$$

のようなデータがあったとき，この算術平均は 5.4 になるが，平均より小さいものが 8 個を占め，残りの 2 個が平均より大きい．これは，少数のデータ（16 と 20）が平均を押し上げている例である．このような場合，分布の中心という意味では既に述べた**中央値（メディアン）**を用いる方が適切である．実際，中央値を用いれば，その値より小さい数と大きい数の個数が等しくなる．

平均値，中央値の他によく用いられる特性値として，**最頻値（モード）**がある．これは，データの中で最も頻出する数であり，度数分布表において度数が最大となる階級の階級値に対応する．ただし，分布形状が双峰性の場合には有効な特性値とならないので注意が必要である．

2.2.2　データのばらつきを表す特性値

データの特性を知りたい場合，中心を表す特性値だけでは情報不足であり，中心からどの程度ばらついているかも考慮しなければならない．例えば，以下の 3 つのデータは中心を表す算術平均，中央値，最頻値がすべて 5 であるが，分布の形状は異なる．

$$A : 0, 3, 3, 5, 5, 5, 5, 7, 7, 10$$
$$B : 0, 1, 2, 3, 5, 5, 7, 8, 9, 10$$
$$C : 3, 4, 4, 5, 5, 5, 5, 6, 6, 7$$

通常，ばらつきを求める際には，算術平均からの距離 $x_i - \bar{x}$ を考える．これを**偏差**と呼ぶ．この偏差を全データに対して平均すれば，ばらつきを表す特性値になりそうであるが，これだと問題が生じる．例えば，データが左右対称に分布している場合，平均より小さい値のデータ（偏差が負）と大きい値のデータ（偏差が正）が同程度あるため，偏差を平均するとほぼ 0 になってしまう．ばらつきが 0 というのは明らかにおかしいため，別の特性値を考える必要がある．以下に代表的な方法を示す．

平均偏差

1 つ目の方法は偏差の絶対値を取ってから平均するというものであり，**平均偏差**と呼ばれる：

$$\text{平均偏差} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{2-5}$$

これは，n 個のデータの 1 個当りの平均からの距離であり，ばらつきの指標として直感的に理解しやすい．しかし，絶対値の扱いが数学的に面倒，分布の中心が $\bar{x}$ ではなく中央値のときに最小になる，平均から大きく外れた値も等しい寄与となる，など問題があるため利用されることは少ない．

分散・標準偏差

2 つ目の方法は，偏差の 2 乗をとってから平均するというものであり，**分散**と呼ばれる：

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2-6}$$

分散は n 個のデータ 1 個当りの平均からの距離の 2 乗であり，平均から大きく離れるほど寄与が大きくなる指標である．また，分散の平方根 s は**標準偏差**と呼ばれる：

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-7}$$

標準偏差はデータの測定単位と同一の単位となるので扱いやすい．通常，データのばらつきを表す特性値としては分散または標準偏差が最もよく用いられる．

なお，平均と同様に分散・標準偏差を度数分布から求めることもできる：

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{k}(v_i-\bar{x})^2 f_i \\
s &= \sqrt{\frac{1}{n}\sum_{i=1}^{k}(v_i-\bar{x})^2 f_i}
\end{aligned}
\tag{2-8}
$$

変動係数

標準偏差は測定単位によって値が変化してしまう．また，データの水準の変化（おおまかには平均値の大きさ）とともに標準偏差の大きさは変化する．そこで，データの水準を平均値によって調整した指標が変動係数（Coefficient of Variation）で，次式で定義される：

$$
\mathrm{CV} = \frac{s}{\bar{x}} \tag{2-9}
$$

変動係数は同じ単位を持つ量同士で割り算をしているので，無次元（単位がない）となり，データの測定単位に寄らない．通常，単位や平均が異なるグループ間でばらつきを比較する際には変動係数が用いられる．

2.2.3　ローレンツ曲線とジニ係数

ある社会を構成する構成員（人物でも市町村でも何でも良い）に対し，所得が対応したデータを考える．この構成員を所得の小さい順に並べて n 個の階級に分け，階級 i のサイズ（人数や個数）を x_i，平均所得を y_i とする．また，サイズの累積相対度数を X_i，平均所得の累積相対度数を Y_i とする．このとき，横軸に構成員の累積相対度数，縦軸に所得の累積相対度数を取ったグラフを**ローレンツ曲線**と呼ぶ．グラフの横軸，縦軸はともに 0 から 1 の範囲であり，曲線上の (X, Y) という点は，社会全体の貧しい側から $X \times 100\%$ の人が全体の $Y \times 100\%$ の富を占めることを表す．例えば，$(0.25, 0.1)$ という点は，社会全体の 25% の人が 10% の富を占めるということを表す．ローレンツ曲線はこれらの点を点線で結んだ折れ線グラフとして表され，必ず両端が $(0, 0)$ と $(1, 1)$ になる．もし，富が全構成員に平等に配分されている場合，ローレンツ曲線は傾き 1 の直線となり，これを**完全平等線**と呼ぶ．一方，富の配分に格差があるほどローレンツ曲線は完全平等線から下にずれていく．このように，ローレンツ曲線はある社会における富の配分格差を可視化したグラフといえる．

社会の格差の度合いはローレンツ曲線が完全平等線から下にずれるほど大きくなる．よって，ローレンツ曲線と完全平等線によって囲まれた部分の面積が，完全平等線，$x = 1$，$y = 0$ で構成される三角形の面積に占める割合を格差の指標と考えることができる．この指標は**ジニ係数**と呼ばれる．階級を n 等分したとき，下から i 番目の階級の平均所得を y_i と

する（$y_1 \leq y_2 \leq \cdots \leq y_n$）．このとき，ジニ係数を以下のように表すこともできる：

$$G = \frac{1}{2n^2\bar{y}} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j| \tag{2-10}$$

ただし，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ は全体の平均所得である．ジニ係数は 0 から 1 の間で定義され，完全に平等な配分のときに $G = 0$，一人がすべての所得を占有しているときに $G = (n-1)/n$ となる（つまり n が大きいときには 1 に近づく）．

例として，$n = 3$ の場合を考える．まず，式 (2-10) を用いてジニ係数を計算すると，

$$\begin{aligned}\sum_{i=1}^{3} \sum_{j=1}^{3} |y_i - y_j| &= 2\{(y_2 - y_1) + (y_3 - y_1) + (y_3 - y_2)\} \\ &= 4(y_3 - y_1) \\ \bar{y} &= \frac{1}{3}(y_1 + y_2 + y_3)\end{aligned}$$

より

$$G = \frac{2}{3} \frac{y_3 - y_1}{y_1 + y_2 + y_3}$$

となる．

次に，面積からジニ係数を計算する．横軸と縦軸の値はそれぞれ表 2.3 のようになる．ローレンツ曲線と $x = 1$，$y = 0$ に囲まれた領域の面積は 1 つの三角形と 2 つの台形から成るので，

$$\frac{1}{2} \cdot \frac{1}{3} \left[\alpha_1 + (\alpha_1 + \alpha_2) + (\alpha_2 + 1)\right] = \frac{1}{6}(2\alpha_1 + 2\alpha_2 + 1)$$

これより，完全平等線と $x = 1$，$y = 0$ に囲まれた領域の面積は

$$\frac{1}{2} - \frac{1}{6}(2\alpha_1 + 2\alpha_2 + 1) = \frac{1}{3}(1 - \alpha_1 - \alpha_2)$$

となるので，ジニ係数は

$$G = \frac{1}{3}(1 - \alpha_1 - \alpha_2) \div \frac{1}{2} = \frac{2}{3}(1 - \alpha_1 - \alpha_2) = \frac{2}{3} \frac{y_3 - y_1}{y_1 + y_2 + y_3}$$

と求まる．

以上より，式 (2-10) から求めたジニ係数と面積から求めたジニ係数が確かに一致することが分かった．なお，一般の n に対する証明は省略する．

表 2.3. $n = 3$ の場合の度数分布表

階級	横軸（構成員）	縦軸（所得）		
	累積相対度数	平均所得 y	累積度数 Y	累積相対度数
I	1/3	y_1	y_1	$\alpha_1 = \frac{y_1}{y_1+y_2+y_3}$
II	2/3	y_2	$y_1 + y_2$	$\alpha_2 = \frac{y_1+y_2}{y_1+y_2+y_3}$
III	1	y_3	$y_1 + y_2 + y_3$	1

表 2.4. ローレンツ曲線作成のためのサンプルデータ

階級	横軸（構成員）			縦軸（所得）		
	人数 x	累積度数 X	累積相対度数	平均所得 y	累積度数 Y	累積相対度数
I	25	25	0.25	15	15	0.0937
II	25	50	0.5	25	40	0.25
III	25	75	0.75	40	80	0.5
IV	25	100	1	80	160	1

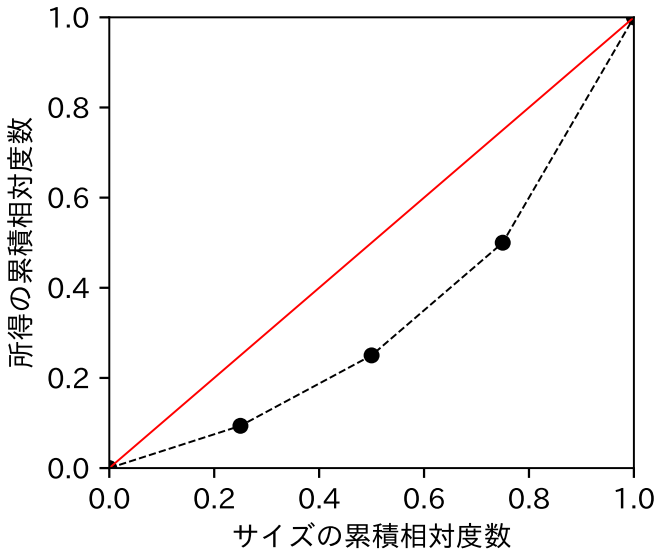

図 2.4. ローレンツ曲線の例（表 2.4 から作成）

2.2.4　実例：地域の豊かさの格差は拡大しているか？

STEP 1: Problem

2015 年の国勢調査では，日本全体の人口が 1920 年の調査開始以来，初めて減少したことが明らかになった．また，都道府県ごとの人口を見ても，5 年前（2010 年）に比べて人口が減少したのは 39 の道府県にのぼる．一方，東京を中心とした大都市には人口が集中し，都市部と地方の格差が広がっているのも事実である．人口は経済・社会の基盤を成すものであり，人口の増減は経済的な豊かさと密接に関わっていると思われる．近年の人口変動によって，地域間で経済的な豊かさの格差は拡大したのだろうか？

STEP 2: Plan

地域ごとの経済的な豊かさを捉える指標として，**都道府県別の 1 人当たり県民所得**に着目する．これは，企業を含めて県民全体の経済水準を表すもので，都道府県間で比較可能な統

計データである．ただし，各都道府県は人口規模が大きく異なるので，地域ごとに比較する際には規模を表す人口等の変数で除した量を用いることが必要となる．そこで，今回用いる 1 人当たり県民所得は，県民所得を県内に居住する人口（「国勢調査」と「人口推計」に準拠）で除して求める．また，格差の大きさは一人当たり県民所得の標準偏差，変動係数，ジニ係数で評価し，ばらつきが 40 年間で拡大しているか否かを調べる．

STEP 3: Data

1 人当たり県民所得は，内閣府「県民経済計算」（https://www.esri.cao.go.jp/jp/sna/data/data_list/kenmin/files/files_kenmin.html）から利用することができる．ただし，年度が同じでも基準（平成 23 年基準や平成 17 年基準など）によって算出された値が異なることに注意する．

STEP 4: Analysis

まず，1 人当たりの県民所得の平均と標準偏差に着目する．

実習

- 1975 年から 2018 年までの 1 人あたり県民所得の csv ファイルを読み込み，各年度に対して平均と標準偏差を求めよ．
- 1975 年〜91 年の標準偏差は一貫して増加しており，格差は拡大しているように見えるが，本当にそう言えるか？平均値の変化と関連付けて考えよ．

次に，標準偏差を平均で割った変動係数の変化を調べる．

実習

- 全年度に対して変動係数を求め，時系列変化を可視化せよ．
- 変動係数の変化から，1975 年〜91 年および全期間にかけて格差が増加しているか考えよ．

次に，所得格差を表すジニ係数に着目する．

実習

- 年度を 1 つ選び，その年度のローレンツ曲線とジニ係数を求めよ．
- 全年度に対してジニ係数を求め，時系列変化を可視化せよ．
- ジニ係数の変化から，1975 年〜91 年および全期間にかけて格差が増加しているか考えよ．

STEP 5: Conclusion

解析の結果を元に，1975 年からの 40 年間で格差が増加したのかどうか考察する．

実習

- 格差を表す指標の 40 年間の推移から地域間で経済的な豊かさの格差が拡大したのかどうか考えよ．
- 格差を表す指標の 40 年間の推移を見ると，細かい時間スケールでの変動が見られる．これらは具体的にどのような出来事を反映していると考えられるか？

2.3　関係の度合い

2.3.1　散布図と相関係数

ここまでは 1 種類の量的データの要約法を扱ってきたが，一方で 2 種類の量的データの間の関係について知りたい場合もある．ここでは，2 種類の量的データを変数 X と Y で表し，これらの間の関係を可視化・定量化する方法を扱う．

まず，２つの量的データの間の関係を可視化するには，それぞれを横軸と縦軸に取ったグラフを描けば一目瞭然である．このようなグラフは**散布図**と呼ばれる．より具体的に散布図とは，n 組のデータ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ に対し，(x_i, y_i) を座標とする点を $X - Y$ 平面上にとったグラフのことである．なお，データは必ず点でプロットし，データ同士を線で結んだりはしない．例として，アヤメデータについて，がく片の長さと幅，花弁の長さと幅の散布図を図 2.5 に示す．まず，がく片の散布図を見ると，長さに対して幅が一定となっており，特に２つの変数に関係はないようである．一方，花弁については右上と左下の区画にデータ点が多く，右上がりの傾向がある．すなわち，花弁が長くなれば，それとともに花弁の幅も大きくなる傾向がある．

なお，もし散布図の中に他の点から極端に外れた点がある場合は外れ値の可能性が高い．このような場合には加工に誤りがないか調べ，データ解析からその値を削除するかどうか検討する．

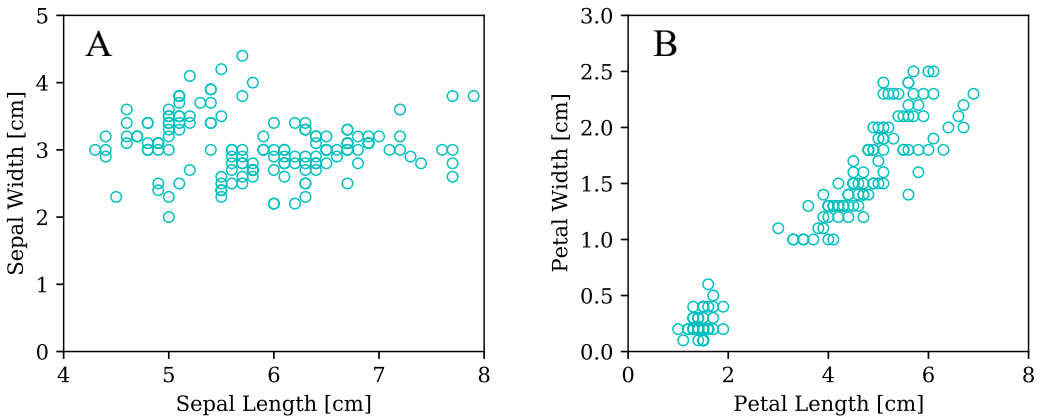

図 2.5. A：アヤメのがく片の長さと幅の散布図．B：アヤメの花弁の長さと幅の散布図．

変数 X, Y の散布図に右上がりの傾向などがある場合には相関関係があるという．このよ

うな相関関係を定量化した量は**ピアソンの相関係数**と呼ばれ，以下の式で定義される：

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{XY}}{s_X s_Y} \tag{2-11}$$

また，分散公式を用いると，

$$r_{XY} = \frac{\overline{xy} - \bar{x}\bar{y}}{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)} \tag{2-12}$$

と表すこともできる．ここで，式 (2-11) の分子に現れた量

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y} \tag{2-13}$$

は変数 X, Y の**共分散**と呼ばれる（$\mathrm{Cov}(X,Y)$ と表すこともある）．共分散は各データごとに X, Y の偏差（平均との差）の積を考え，それらの全データに対する平均を考えている．これにより，散布図が右上がりのときに正，右下がりのときには負となる統計量が得られる．ただし，これだけだと X, Y の単位やばらつきの度合いによって値が大きく異なってしまい，相関の程度を一定の基準で表すことができない．そこで，共分散を X, Y の標準偏差で割って，各変数の単位やばらつきの度合いに依らない量としたのが相関係数である．

実際，相関係数の定義域は $-1 \leq r_{XY} \leq 1$ であり，その絶対値によって相関の強さを表すことができる．また，定義より，相関係数の符号は右上がりの傾向の場合に正，右下がりの傾向の場合に負となる．特に，相関係数が正の値のときに**正の相関**，負の値のときに**負の相関**があるという．また，相関係数が 0 のときには**無相関**という．実際に図 2.5 から相関係数を求めると，がく片については $r_{XY} = -0.11$ なのでほぼ無相関，花弁については $r_{XY} = 0.96$ となるので強い正の相関があることが分かる．

なお，相関係数はあくまでも 2 つの変数の散布図が直線的な関係になる場合だけ意味があることに注意しなければならない．例えば，X, Y の散布図が円状に分布する場合，2 つの変数には何かしらの規則があると考えられるが，相関係数は 0 となり無相関と判断されてしまう．このため，相関係数を調べる際には必ず散布図も併せて描く必要がある．

2.3.2 相関関係と因果関係

2 つの変数 X, Y の間に相関関係があったときに，それらの間に因果関係があるといえるだろうか？つまり，$X \Rightarrow Y$ または $Y \Rightarrow X$ という関係が成り立つだろうか？実は，これは必ずしも成り立つとは限らない．その理由は，第一に，全くの偶然で強い相関関係が現れることがあるからである．例えば，文献 [8] によると，以下の 3 つは全くの偶然で強い相関関係が現れた例である：

- 「ニコラス・ケイジの年間映画出演本数」と「プールの溺死者数」
- 「ミス・アメリカの年齢」と「暖房器具による死亡者数」
- 「商店街における総収入」と「アメリカでのコンピュータサイエンス博士号取得者数」

また，第二に，調べたい 2 つの変数 X, Y それぞれが別の変数 Z と強く相関する場合，X と Y の相関が見かけ上強くなってしまうこともある．このような相関は**疑似相関**と呼ばれ，疑似相関の原因となる変数 Z のことを**第 3 の変数**と呼ぶ．疑似相関では，$Z \Rightarrow X$ および $Z \Rightarrow Y$ という因果関係が成り立つが，$X \Rightarrow Y$ または $Y \Rightarrow X$ という因果関係は成り立たないことに注意する．なお，第 3 の変数のデータは必ずしも手に入るとは限らないが，もし入手できていない場合は**潜在変数**と呼ぶ．疑似相関の例は枚挙にいとまがないが，例えば，「子供の体力」と「子供の学力」の強い相関関係は疑似相関の典型例である．この場合，第 3 の変数は「親の教育熱心さ」であり，親が教育熱心であれば当然学力が高い傾向にあり，また子供にスポーツを習わせるので体力も上がる傾向があるということになる．

第 3 の変数の影響を除く方法はいくつか知られている．1 つ目は第 3 の変数による層別の方法である．これは，第 3 の変数の値が近いものだけでいくつかのグループに分け，グループ内で相関を見る方法である．

2 つ目は第 3 の変数の単位あたりの量に変換する方法である．例えば，第 3 の変数が人口の場合，人口 1 人あたりの X, Y に変換し，これらの相関を見ることで正しい相関関係を調べることができる．

3 つ目は**偏相関係数**を用いる方法である．偏相関係数とは，関係を調べたい 2 つの変数に対して別の変数の影響を取り除いた上で求めた相関係数である．いま，第 3 の変数を Z とし，$(x_1, y_1), \ldots, (x_n, y_n)$ に対して $(z_1, \ldots, z_n)$ の影響を除いた相関係数を考えたい．これには，以下のように回帰直線の考え方を使う．まず，Z による X の予測値を $\hat{x}_i = az_i + b$ として，最小二乗法によって a, b を求める．このとき，Z の影響を除いた X を $\tilde{X}$ とすると，これは残差 $\tilde{x}_i = x_i - \hat{x}_i$ によって与えられる．同様にして，Z の影響を除いた Y を $\tilde{Y}$ とすると，これは予測値 $\hat{y}_i = cz_i + d$ に対して，残差 $\tilde{y}_i = y_i - \hat{y}_i$ によって与えられる．このようにして，Z の影響を除いた $\tilde{X}, \tilde{Y}$ のデータ $(\tilde{x}_i, \tilde{y}_i) = (x_i - \hat{x}_i, y_i - \hat{y}_i)$ $(i = 1, 2, \ldots, n)$ が得られる．偏相関係数は Z の影響を除いた $\tilde{X}, \tilde{Y}$ の相関係数 $r_{\tilde{X},\tilde{Y}}$ として定義されるが，実は $r_{\tilde{X},\tilde{Y}}$ は以下のように変数 X, Y, Z に対する通常の相関係数から求めることができる：

$$r_{\tilde{X},\tilde{Y}} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}} \tag{2-14}$$

証明には次節で扱う最小二乗法が必要となるので省略する（https://manabitimes.jp/math/1400 を参照）．

なお，第 3 の変数を取り除いた結果相関がなくなれば，それは擬似相関であり，2 変数間に因果関係がないことが分かる．一方，第 3 の変数を取り除いても相関が大きいままの場合，相関関係が成立する可能性は高くなるが，2 変数間に因果関係があるかについては何も言えないことに注意する．例えば，偏相関係数が大きくてもそれが因果関係の存在を意味するわけではない．また，第 3 の変数の影響によって見かけ上相関が発生する疑似相関とは逆

に，第 3 の変数の影響によって見かけ上無相関となる**疑似無相関**も存在する．

2.3.3　実例：警察職員数と刑法犯認知件数の関係 [2]

STEP 1: Problem

ある統計によると，警察官の数と犯罪の件数には正の相関関係があると言われている．では，これらの間に因果関係はあるだろうか？

STEP 2, 3: Plan, Data

都道府県別の警察職員数と刑法犯認知件数のデータを用い，これらの相関関係および因果関係の有無を調べることにする．今回用いるデータは以下の通りである：

- 2015 年度の都道府県別刑法犯認知件数のデータ：平成 27 年警察白書（https://www.npa.go.jp/hakusyo/h27/data.html）
- 2015 年度の都道府県別警察職員数のデータ：総務省の HP（https://www.soumu.go.jp/main_sosiki/jichi_gyousei/c-gyousei/teiin/109981data.html）
- 2015 年度の都道府県別人口：e-Stat（https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200241&tstat=000001039591&cycle=7&year=20150&month=0&tclass1=000001039601&result_back=1&tclass2val=0）

図 2.6 は 2015 年度の都道府県別警察職員数と刑法犯認知件数[*3]の散布図を表している．この散布図を見ると確かに両者には正の相関関係があり，相関係数も高くなりそうである．では，このことから警察職員数と刑法犯認知件数に因果関係[*4]があるといえるだろうか？これを調べるため，今回は都道府県の人口を第 3 の変数と仮定して解析を行う．

実習

- 図 2.6 について，相関係数を求めよ．

[*3] 警察等の捜査機関によって犯罪の発生が確認された件数

[*4] 警察職員が増えると刑法犯認知件数が増える／刑法犯認知件数が増えると警察職員数が増える

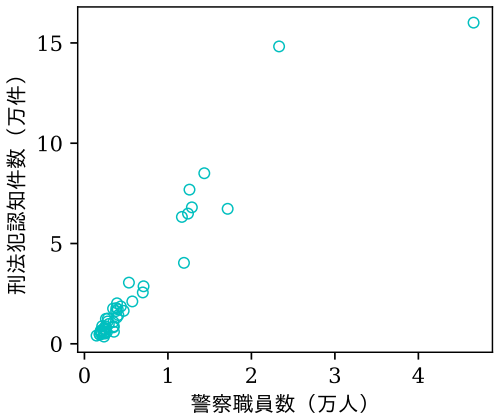

図 2.6. 2015 年の都道府県別警察職員数と刑法犯認知件数の散布図

STEP 4: Analysis

まず，都道府県の人口を第 3 の変数と仮定し，警察職員数と刑法犯認知件数の関係が疑似相関であるか調べる．

実習

- 刑法犯認知件数と警察職員数のそれぞれについて人口との散布図を描け．
- これらの散布図の相関係数を求め，それが何を意味するか考察せよ．

次に，疑似相関の有無についてさらに詳しく考察するため，人口の影響を取り除いた相関関係を調べる．

実習

- 刑法犯認知件数と警察職員数の散布図について，人口が 100 万人未満，100 万人以上 200 万人未満，200 万人以上 500 万人未満で層別し，結果を考察せよ．
- 人口 1000 人あたりの警察職員数と刑法犯罪認知件数に関する散布図を描いてその相関係数を求め，結果を考察せよ．
- 人口の影響を除いた警察職員数と刑法犯認知件数の偏相関係数を求め，結果を考察せよ．

STEP 5: Conclusion

実習

- 解析の結果から，警察職員数と刑法犯認知件数の間に因果関係があるかどうか考察せよ．

2.4　散布図・相関分析による問題解決

2.4.1　回帰直線と最小二乗法

2 つの量的データ X と Y が与えられたとき，変数 X の値から Y の値を予測するための数式のことを**回帰モデル**と呼ぶ．また，X を**説明変数（独立変数）**，Y を**目的変数（従属変数，被説明変数）**と呼ぶ．

例として，X と Y の散布図が図 2.7 のように与えられる場合を考える．このとき，X と Y の間には直線関係が成り立ちそうである．よって，回帰モデルとして，1 次関数

$$\hat{y} = ax + b \tag{2-15}$$

を用いるのが妥当と考えられる．この回帰モデルは**単回帰モデル**と呼ばれ，式 (2-15) の直線のことを**回帰直線**と呼ぶ．なお，a, b は回帰直線の切片と傾きを表すパラメータであり，**回帰係数**と呼ばれる．

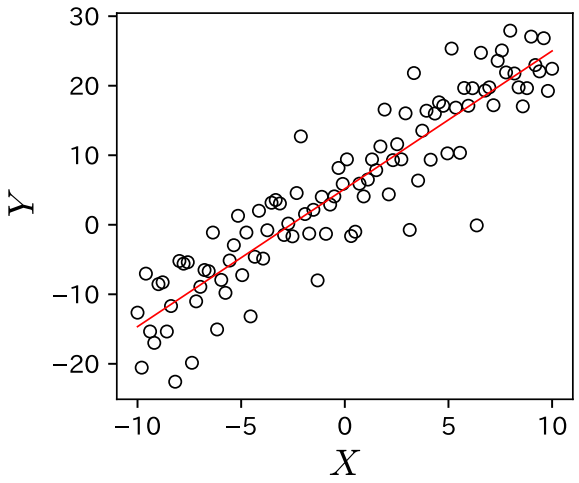

図 2.7. 最小二乗法による回帰直線の例．

最小二乗法

n 組のデータ (x_i, y_i) $(i = 1, 2, \ldots, n)$ が与えられたとき，式 (2-15) を用いてデータから最適な回帰直線を求めることを**単回帰分析**と呼ぶ[*5]．単回帰分析には様々な方法があるが，最も基本的な方法が**最小二乗法**である．最小二乗法の発想は単純であり，予測値 $\hat{y}_i = ax_i + b$

[*5] 説明変数が複数ある場合は重回帰分析と呼ぶ．

と実データ y_i の差の二乗和（残差変動）

$$E = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

が最小となるような a, b を回帰係数とするものである．このための条件は，残差二乗和 E の a, b による偏微分がゼロという式で与えられる：

$$\frac{\partial E}{\partial a} = 0, \qquad \frac{\partial E}{\partial b} = 0$$

実際にこれらの条件を適用すると，a, b は

$$a = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2} = \frac{s_{XY}}{s_X^2}$$

$$b = \bar{y} - a\bar{x} = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i)$$

と表される．なお，傾き a は共分散 s_{XY} を X の分散 s_X^2 で割った形になっている（相関係数の式に似ているが違う）．よって，以下が成り立つ：

共分散 s_{XY} が正，0，負 $\Longleftrightarrow$ 最小二乗法による傾き a が正，0，負

2.4.2 目的変数の変動と決定係数

n 組のデータ (x_i, y_i) $(i = 1, 2, \ldots, n)$ に対して，次の 3 つの変動を考える．

1. 全変動（データ Y のばらつき）：$S_y^2 = \displaystyle\sum_{i=1}^{n} (y_i - \bar{y})^2$
2. 回帰変動（回帰モデルによる予測値のばらつき）：$S_{\hat{y}}^2 = \displaystyle\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
3. 残差変動（実データと予測値のズレ）：$S_e^2 = \displaystyle\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

単回帰分析の場合，以下の関係が成り立つ：

$$S_y^2 = S_{\hat{y}}^2 + S_e^2$$

以上を踏まえ，回帰直線の当てはまりの良さを示す**決定係数**を

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} \tag{2-16}$$

と定義する[*6]．これは，全変動と回帰変動の比として定義されるので，実データの変動のうち，どの程度が回帰モデルで説明できるのかを表す量である．特に，式 (2-16) より定義域は $0 \leq R^2 \leq 1$ であり，残差変動が 0 に近づく（データへの当てはまりが良い）と R^2 は 1 に近づく．一方，残差変動が大きくなる（データへの当てはまりが悪い）と R^2 は 0 に近づく．なお，予測値 $\hat{y}$ が最小二乗法によって決められた場合，決定係数は相関係数の二乗に等しい．

2.4.3　実例：都市の平均気温と緯度の関係

STEP 1: Problem

地球上では，赤道付近は暑く，極地に近づくほど寒くなる．世界の様々な地域の年間平均気温はどのように決まっているのだろうか？

STEP 2: Plan

世界の各地域で年間平均気温は異なっている．各地域での年間平均気温に影響を与える要因は，図 2.8 のような図にまとめることができる．これを**特性要因図**と呼ぶ．ここには，各地域の地球上での位置，都市の自然環境，人間活動が要因として挙げられている．以下では，各都市の地球上での位置に関するデータを収集し，年間平均気温との関係を探る．

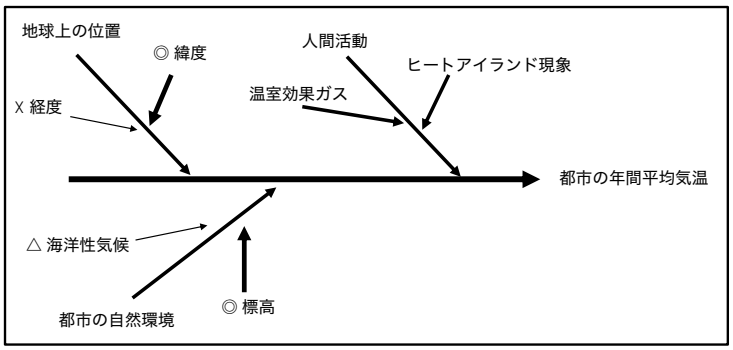

図 2.8. 特性要因図．都市の年間平均気温に影響を与える要因．

STEP 3: Data

世界各都市の年間平均気温は理科年表や気象庁の HP で調べることができる．表 2.5 は理科年表から得られた各都市の年間平均気温，緯度，標高のデータである．

[*6] 他の定義もあるので注意．

表 2.5. 世界の 25 都市の年間平均気温，緯度，標高

地名	平均気温（℃）	緯度（度）	標高（m）		地名	平均気温（℃）	緯度（度）	標高（m）
昭和基地	-10.5	-69.00	18		ドーハ	27.0	25.15	11
メルボルン	14.5	-37.39	132		カイロ	21.7	30.06	116
ブエノスアイレス	17.8	-34.35	25		ケープタウン	16.8	33.58	46
ブリスベン	20.3	-27.23	4		東京	15.4	35.42	25
リオデジャネイロ	23.9	-22.55	5		サンフランシスコ	14.5	37.37	6
リマ	19.3	-12.01	12		北京	12.9	39.56	55
ジャカルタ	28.0	-6.11	8		サラエボ	10.4	43.52	630
シンガポール	27.6	1.22	5		リオン	11.9	45.43	197
ボゴダ	13.4	4.42	2547		チュリッヒ	9.4	47.22	555
コロンボ	27.7	6.54	7		プラハ	8.4	50.06	380
アジスアベベ	16.6	9.02	2354		ダブリン	9.8	53.26	68
チェンマイ	29.0	13.00	13		レイキャビク	4.7	64.08	54
メキシコ	16.7	19.24	2309					

STEP 4: Analysis

散布図

まず，平均気温と他の量との相関関係を視覚的に確認するために散布図を調べることにする．図 2.9 は緯度と平均気温，標高と平均気温の散布図である．

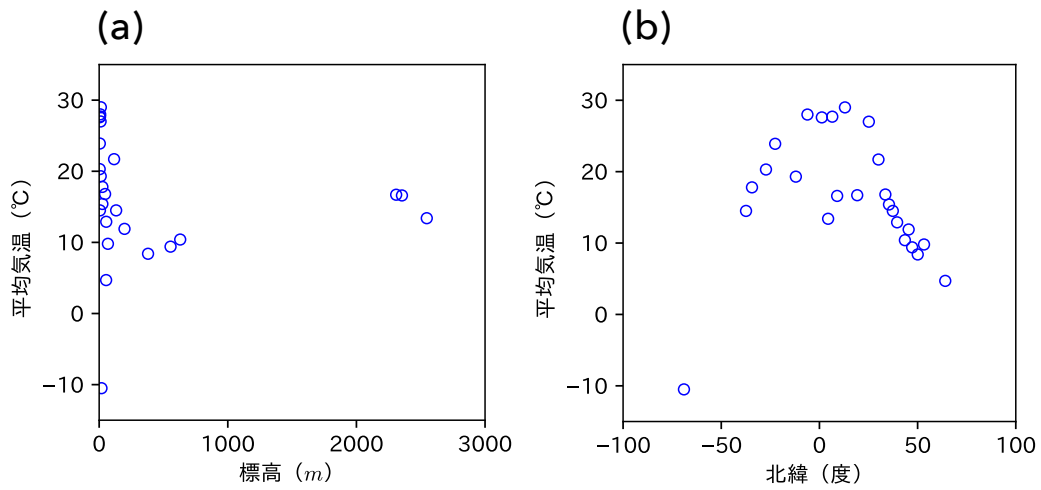

図 2.9. 緯度・標高と平均気温の関係

図 2.9 を見ると，平均気温は緯度に対して上に凸の 2 次関数のような関係となり，かつ赤道（0 度）に対して左右対称になっていることが分かる．一方，標高と平均気温については特定の関数関係はない．このように，散布図がそもそも直線関係となっていない場合には，相関係数を求めるのは不適切である．

緯度と平均気温の関数関係

緯度と平均気温の関数関係を特定すれば，散布図が直線関係を示すような適切な変数変換を導ける．そこで，緯度の値別に複数のグループに分け，グループ内で平均気温，平均緯度などを求めて関係を調べることにする．具体的に，ここでは 25 都市を緯度の絶対値の昇順に 5 都市ずつのグループに分ける：

群 1：シンガポール，ボコタ，ジャカルタ，コロンボ，アジスアベバ
群 2：リマ，チェンマイ，メキシコ，リオデジャネイロ，ドーハ
群 3：ブリスベン，カイロ，ケープタウン，ブエノスアイレス，東京
群 4：サンフランシスコ，メルボルン，北京，サラエボ，リオン
群 5：チューリッヒ，プラハ，ダブリン，レイキャビク，昭和基地

実習

- 各群に対し，絶対緯度の平均，平均気温の平均，平均気温の標準偏差を求めよ
- 平均絶対緯度を横軸，平均気温の平均を縦軸に取った図をエラーバー付きで描け

実際に平均絶対緯度を横軸，平均気温の平均を縦軸に取った図を描くと，絶対緯度と平均気温の関係は 2 次関数的な変化となっていることが分かる．

実習

- 横軸に絶対緯度の 2 乗，縦軸に平均気温を取った散布図を描け
- この散布図に対して，相関係数を求めよ
- この散布図に対して，最小二乗法で回帰直線を求めよ

Step 5: Conclusion

各都市の年間平均気温と緯度の関係を散布図によって調べた結果，緯度の 2 乗に対して直線関係があることが分かった．一方，年間平均気温が緯度の 2 次関数になるということは，緯度が高くなれば気温もいくらでも大きくなることを意味し，やや奇妙である．

実習

- 緯度 θ における太陽エネルギーは $\cos\theta$ に比例することが知られている．これより，年間平均気温と緯度を結びつける，より適切な関数を求めよ．
- テイラー展開の観点から，2 次関数の妥当性を議論せよ．

Step 6: Problem 2

緯度の 2 乗と平均気温の散布図を見ると，直線関係から少し外れる都市がいくつかあることが分かる．これらの都市は，外れ値の大きい順にボコダ，メキシコ，アジスアベバ，である．では，これらの都市はなぜ直線関係から外れるのだろうか？

Step 7: Plan & Data 2

直線から外れている都市について，表 2.5 を見てみると，ある共通点が浮かび上がる．それは，標高が高いことである（いずれも標高 2000m 以上）．一般的に，標高が高くなるほど都市の気温は低くなる．よって，平均気温と緯度の関係を見るためには，標高の影響を調整する必要がある．一般に，標高が 100m 高くなると，気温は 0.6 ℃低くなると言われている．これより，平均気温（T）に対して標高（z）の影響を調整した気温（高度調整済み平均気温）は $T_0 = T + 0.006z$ と表される．

Step 8: Analysis 2

実習

- 緯度の 2 乗と高度調整済み平均気温の散布図を描け
- この散布図から相関係数を求めよ
- この散布図に対して回帰直線を引き，直線の式を求めよ

Step 9: Conclusion 2

年間平均気温と緯度の 2 乗の関係において発生する外れ値は，標高の影響によるものであることが分かった．そこで，標高の影響を調整することで，年間平均気温，緯度，標高に対する適切な関数を推定することができた．

第 3 章

確率と確率変数

3.1　確率の概念

3.1.1　標本空間と事象

サイコロを振ったりコインを投げるなど，同じ条件で繰り返すことができる操作を**試行**と呼び，試行の結果起こりうる事柄を**事象**と呼ぶ．また，ある試行によって起こりうる個々の結果を**標本点** ω，その全体の集合を**標本空間** Ω と表す．

事象は集合を使って整理することができ，より正確には標本空間の**部分集合**によって定義される．ここで，A が B の部分集合であるとは，A のすべての構成要素が B の構成要素である場合をいい，$A \subset B$ と表す．標本点 $\omega_1, \cdots, \omega_n$ から成る事象 A は $A = \{\omega_1, \cdots, \omega_n\}$ のように表す．なお，標本空間 Ω 自体を**全事象**，標本点を 1 つも含まない集合を**空事象** ϕ と呼び，これらも事象と見なす．一般に，標本空間が r 個の標本点から成る場合，事象（部分集合）は 2^r 個存在する．

例えば，コインを 1 回投げる場合，「表」と「裏」が標本点であり，標本空間は

$$\Omega = \{\ \text{表}, \text{裏}\ \}$$

となる．また，標本点の数が $r = 2$ なので，事象は以下のように $2^2 = 4$ 個ある：

$$\{\ \text{表}, \text{裏}\ \}, \{\ \text{表}\ \}, \{\ \text{裏}\ \}, \phi$$

ここで，$\{$ 表, 裏 $\}$ が全事象に対応し，コインを 1 回投げて表または裏が出るという事象を表わす．サイコロを 2 個同時に投げる場合の標本空間は 36 個の標本点から成る：$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$．特に，目の合計が 7 となる事象は

$$A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$$

である．

事象のうち，ただ 1 つの標本点から成りそれ以上分解できないものを**根元事象**と呼び，一方で複数の標本点を含み 2 つ以上の根元事象に分解可能なものは**複合事象（結合事象）**と呼

ぶ．例えば，サイコロを 1 回投げた場合，$\{1\}, \{2\}$ は根元事象であるが，$\{1, 3, 5\}$（偶数の目が出る）は複合事象である．

以上のように，標本点が有限個の点から成る場合もあるが，長さ，重さ，温度など，ある区間内のすべての点を取りうる場合には標本空間は無限個の点から成る無限集合である．例えば，電球の寿命は確率的と考えられるが，標本空間は

$$\Omega = (0, \infty)$$

である．このとき，1000 時間目に電球がまだ正常であるという事象は $A = (1000, \infty)$ である．

事象の演算

事象を表すのに便利な方法として，**ベン図**がある．これは，標本点の数に関わらず標本空間を長方形で表し，各事象を長方形の内部に円で示す方法である．2 つの事象 A, B の関係は以下の 3 つに分類できる（図 3.1）：(a) A は B の部分集合である，(b) A は B の部分集合でないが，A と B は共通部分を持つ，(c) A と B は共通部分を持たない．この中で，(c) の場合，一方が起これば他方は起こらないので，A と B は**排反事象**または**互いに排反**であるという．

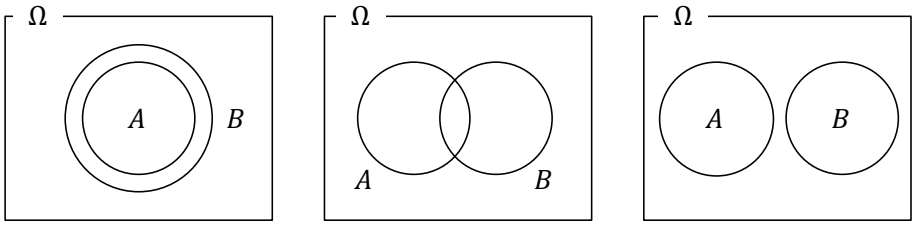

図 3.1. ベン図による 2 つの事象の関係の図示．

A と B の 2 つの事象のうち少なくとも 1 つが起こる（A または B が起こる）という事象を，A と B の**和事象**と呼び，$A \cup B$ で表す（$\cup$ の記号は “または” や “カップ” などと読む）．例えば，サイコロを 1 回投げる場合，A を奇数の目が出る事象，B を 3 以下の目が出る事象とすると，$A \cup B = \{1, 3, 5\} \cup \{1, 2, 3\} = \{1, 2, 3, 5\}$ である．

A と B が同時に起こる（A かつ B が起こる）という事象は，**積事象**と呼ばれ，$A \cap B$ と表される（$\cap$ の記号は “かつ” や “キャップ”，“共通部分” などと読む）．例えば，サイコロを 1 回投げる場合，A を奇数の目が出るという事象，B を 3 以下の目が出るという事象とすると，$A \cap B = \{1, 3, 5\} \cap \{1, 2, 3\} = \{1, 3\}$ である．なお，A と B の共通部分がない場合には $A \cap B = \phi$ である．

3 つの事象 A, B, C に対して，以下の分配法則が成り立つ：

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

事象 A が起こらないという事象は A の**補事象（余事象）**と呼び，A^c または $\bar{A}$ と表す．また，事象 A, B に対し，以下の**ド・モルガンの法則**が成り立つ：

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c$$

例えば，A をスペードが出る，B を絵札が出るという事象とすると，「スペードまたは絵札が出る」（$A \cup B$）の補事象は「スペードが出ず，かつ絵札も出ない」（$A^c \cap B^c$）である．

3.1.2　確率の定義

ラプラスの定義

ある試行について，根元事象が N 個あり（つまり，標本空間の大きさが N），それらが同程度に確からしく起こるとする．このとき，事象 A に対応する根元事象が R 個あれば，事象 A の確率は

$$P(A) = R/N$$

と定義される．例えば，サイコロを 1 回投げて奇数の目が出るという事象を A とすると，その確率は $N = 6$，$R = 3$ なので $P(A) = 3/6 = 1/2$ である．

以上の定義の利点は，確率が標本点の個数，すなわち起こり方の場合の数の数え上げに帰することであり，順列，組み合わせの諸定理が使えることである．一方，問題点は，各標本点が同様に確からしく起こると仮定していることである．例えば，サイコロの場合は各目が同程度の確かさで出現すると仮定しているが，これが正しい保証はない．

頻度による定義

ラプラスの定義はサイコロやコインなどについては有益であるが，各標本点が同程度に確からしく起こると考えられない場合には適用できない．そこで，より実際的な定義として，以下のような頻度による定義（頻度説）がある．ある試行を n 回繰り返し，事象 A が生じた回数（頻度または度数）n_A を数える実験を行う．いま，$n \to \infty$ としたとき，その相対頻度（相対度数）が

$$\lim_{n \to \infty} \frac{n_A}{n} = \alpha$$

となるならば，$P(A) = \alpha$ と定義する．

一般に，相対頻度 n_A/n は真の確率 $P(A)$ と一致せず，試行回数 n が同じでも各回ごとに異なる値が観測される．しかし，いずれの場合も n が大きくなるに従い，相対頻度は真の

値とほとんど等しくなる様子が観察され，これを確率と見なすのが頻度説である．このように，頻度説は無限の試行によって初めて正当化されるので，あくまでも理論上の仮定の上に成り立っている．

公理主義的な定義

確率と事象の関係を規定し，確率を数学的に構成するためには，公理[*1]を設けてそれに基づいて体系的に議論する必要がある．この考えに基づくのが以下に示す確率の公理主義的定義である．

> 標本空間 Ω の事象 A に対して次の 3 つの条件を満たす実数 $P(A)$ が存在するとき，$P(A)$ を事象 A が起こる確率という：
>
> 1. $0 \leq P(A) \leq 1$
> 2. $P(\Omega) = 1$
> 3. 互いに排反な事象 A_1，A_2，... に対し，
>
> $$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

これらの公理のうち，3 つ目は経験的確率の性質を一般的に書いたものである．例えば，サイコロを振って偶数の目が出るという事象 A_1 の確率は $P(A_1) = 3/6$，3 の目が出るという事象 A_2 の確率は $P(A_2) = 1/6$ であり，これらは排反事象である．このとき，偶数の目または 3 の目が出るという事象の確率は $P(A_1 \cup A_2) = 3/6 + 1/6$ によって計算できる．

3.1.3　確率の性質

以下では，確率の公理主義的な定義を前提とする．

加法定理

事象 A と B が互いに排反（共通部分を持たない）であるとき，公理より

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

が成り立つ．

一方，事象 A と B が排反ではなく，共通部分を持つ場合には

$$P(A \cup B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B)$$

が成り立つ．また，$A = (A \cap B^c) \cup (A \cap B)$ であり，$A \cap B^c$ と $A \cap B$ は排反事象なので，

$$P(A) = P(A \cap B^c) + P(A \cap B)$$

[*1] 数学的な体系を構築する上で，無条件で正しいとする前提，仮定

同様にして，

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

である．以上より，

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

が成り立つ．これを**確率の加法定理**と呼ぶ．

例えば，ジョーカーを含むトランプ 53 枚から 1 枚取り出す場合を考える．カードがスペードであるという事象を A とすると，$P(A) = 13/53$ である．また，カードが絵札であるという事象を B とすると，$P(B) = 12/53$ である．さらに，スペードの絵札である確率は $P(A \cap B) = 3/53$ である．よって，加法定理を用いると，カードがスペードかまたは絵札である確率は

$$P(A \cup B) = 13/53 + 12/53 - 3/53 = 22/53$$

となる．

条件付き確率

2 つの事象 A, B があるとき，事象 A が起こったという条件の下で B が起こるという事象を $B|A$ と表す．また，その確率 $P(B|A)$ を条件 A の下での B の**条件付き確率**といい，

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

と定義する（ただし，$P(A) \neq 0$）．

例として，トランプから 1 枚のカードを取り出す場合を考える．カードがスペードであるという事象を A，絵札であるという事象を B とする．このとき，$B|A$ はカードがスペードであると分かった場合にそれが絵札であるという事象を意味する．スペードのカードは 13 枚あり，そのうち絵札は 3 枚なので，$P(B|A) = 3/13$ である．一方，スペードである確率は $P(A) = 13/53$，スペードでありかつ絵札である確率は $P(A \cap B) = 3/53$ であるから，条件付き確率の定義より $P(B|A) = \frac{3/53}{13/53} = 3/13$ となり定義を満たすことが分かる．

なお，条件 B の下での A の条件付き確率 $P(A|B)$ も同様に定義でき，

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

が成り立つ（ただし，$P(B) \neq 0$）．よって，これらの式を変形すれば以下が成り立つ：

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

これを**確率の乗法定理**と呼ぶ．

独立性

条件付き確率 $P(A|B)$ は B が起こったという条件の下で A が起こる確率であるが，B が A に何の影響も及ぼさない場合もある．このようなときは

$$P(A|B) = P(A)$$

が成り立ち，乗法定理は

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

となる．以上 2 つの式が成り立つとき，事象 A と B は**統計的に独立**であるという．なお，これは n 個の事象 A_1，A_2，$\cdots$，A_n についても成り立ち，

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

のときに事象 A_1，A_2，$\cdots$，A_n は独立である．

3.1.4　ベイズの公式

事象 $B_1, B_2, \cdots, B_n$ は互いに排反で，かつ $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$ が成り立つとする．このとき，ある事象 A に対し，乗法公式より

$$P(A|B_i)P(B_i) = P(B_i|A)P(A)$$

両辺の i について和をとれば，

$$\begin{aligned}\sum_i P(A|B_i)P(B_i) &= \sum_i P(B_i|A)P(A) \\ &= P(A)\end{aligned}$$

となる（**全確率の定理**）．これより，

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)}$$

が成り立つ．これを**ベイズの公式**と呼ぶ．

いま，B_i を原因，A を与えられた結果と見なすと，原因 B_i が起こる確率 $P(B_i)$ と条件付き確率 $P(A|B_i)$ が分かれば，結果 A から原因の確率 $P(B_i|A)$ を計算できることを意味する．ここで，$P(B_i)$ は**事前確率**，$P(B_i|A)$ は**事後確率**，$P(A|B_i)$ は**尤度**と呼ばれる．

例 1）ある製品を作る機械 B_1，B_2，B_3 がある．この製品を作るときに，各機械が使われる割合（確率）はそれぞれ $P(B_1) = 0.5$，$P(B_2) = 0.3$，$P(B_3) = 0.2$ である．いま，出来

上がった製品が不良品であるという事象を A とする．各機械が不良品を出す割合がそれぞれ $1\%, 1.5\%, 2\%$ であるとすれば

$$P(A|B_1) = 0.01, \quad P(A|B_2) = 0.015, \quad P(A|B_3) = 0.02$$
$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) = 0.0135$$

となる．このとき，出来上がった製品が不良品であったとき，使われた機械が B_i であった確率はベイズの公式よりそれぞれ

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A)} = 0.3704$$
$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A)} = 0.3333$$
$$P(B_3|A) = \frac{P(A|B_3)P(B_3)}{P(A)} = 0.2963$$

と求まる．すなわち，この場合は機械 B_1 で作られた可能性が最も高い（機械 B_1 の性能は最も高いことに注意).

例 2) 新型コロナウイルスの PCR 検査によって陽性になるという事象を A とする．また，新型コロナウイルスに感染するという事象を B とする．いま，市中感染率は P(B)=0.1% で，PCR 検査の偽陰性率は 30%，偽陽性率は 1% と仮定すれば，尤度と全確率は

$$P(A|B) = 0.7, \quad P(A|\bar{B}) = 0.3$$
$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) = 0.01069$$

となる．このとき，PCR 検査で陽性だったときに実際に感染している確率は

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = 0.7 \times 0.001/0.01069 = 0.06548 \text{（約 } 6.5\%\text{）}$$

と求まる.

3.2 確率の応用

3.2.1 確率変数と確率分布

確率変数

確率を数学的に扱うには，各事象に適当な数値を与えると便利である．例えば，サイコロを投げる場合，目の数字を変数と見なすと，各事象には 1 から 6 までの整数値が与えられることになる．また，雨が降る，降らないといった事象の場合には，「降る」を 1，「降らない」を 0 とすれば変数となり得る．このように，標本空間の根元事象に対して適当な数値を対応させた変数 X を考え，$X = x$ となる確率が定まっているとき，X を**確率変数**と呼ぶ．一般に，確率変数 X は試行を行って初めて値が決まる変数であるので，単なる数値と区別して大文字で表す．また，個々の試行の結果（確率変数の**実現値**）は小文字で表す．もし，確率変数 X が離散的な値 $x_1,\ x_2,\ \cdots$ しか取らないとき（サイコロの目など），X を**離散型確率変数**と呼ぶ．一方，X が連続値を取る場合（重さ，長さ，時間など）は**連続型確率変数**と呼ぶ．

離散型確率分布

離散型確率変数 X が実現値 x_i を取る確率を

$$P(X = x_i) = f(x_i) \quad (i = 1, 2, \cdots)$$

と表す．ただし，f は以下の条件を満たすとする：

$$f(x_i) \geq 0, \quad \text{かつ} \quad \sum_{i=1}^{\infty} f(x_i) = 1$$

このように，確率変数 X の各実現値 $x_1,\ x_2,\ \cdots$ に対してその確率を対応させた関数 f を**確率分布**と呼ぶ．特に，X が離散型確率変数の場合，f を**離散型確率分布**または**確率（質量）関数**と呼ぶ．また，確率変数 X が x 以下である確率を

$$P(X \leq\ x) = F(x) = \sum_{x_i \leq x} f(x_i)$$

と表し，これを**累積分布関数**と呼ぶ．なお，離散型確率変数の累積分布は不連続な関数となる．

なお，記述統計学において，度数分布（ヒストグラム）を扱ったが，これは与えられたデータに対して，階級値と（相対）度数が対応したものであった．データ数 n が十分大きい（$n \to \infty$）ときに相対度数が確率に一致するということを踏まえると，確率分布とはヒストグラムに対する理論的なモデルと捉えることができる．

例として，サイコロを 1 個投げた場合を考える．この場合，確率変数の実現値は $1, 2, 3, 4, 5, 6$ であり，それぞれの確率が $1/6$ なので，確率分布は以下のようになる：

$$f(1) = \frac{1}{6},\ f(2) = \frac{1}{6},\ f(3) = \frac{1}{6},\ f(4) = \frac{1}{6},\ f(5) = \frac{1}{6},\ f(6) = \frac{1}{6}$$

また，この確率分布をグラフに表すと図 3.2 のようになる．なお，このように全ての実現値に対して同じ確率を取るような確率分布を**一様分布**と呼ぶ．

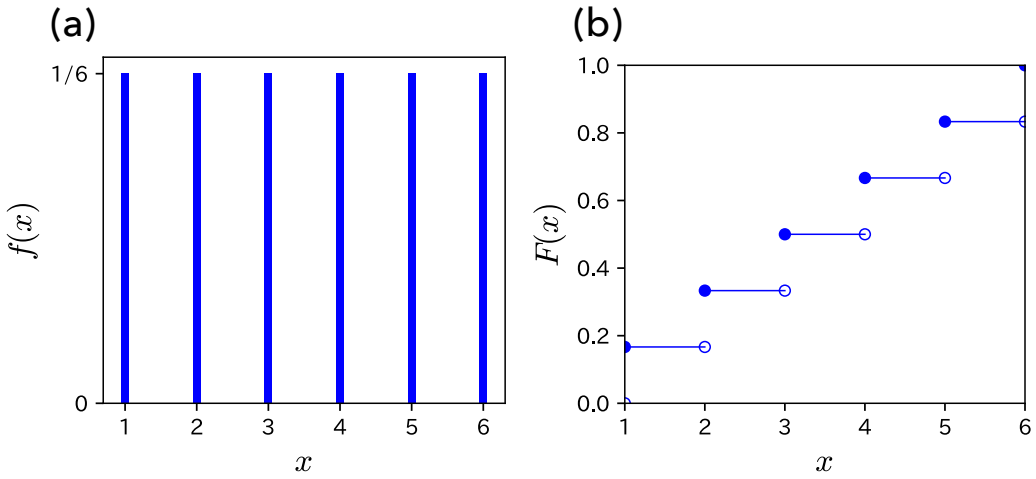

図 3.2. サイコロを 1 個投げた場合の確率分布（一様分布）．(a) 確率関数，(b) 累積分布関数．

連続型確率分布

連続型確率変数の場合，確率変数 X がある実現値 a を取る確率はゼロとなる：

$$P(X = a) = 0$$

そこで，連続型確率変数の場合には，X がある範囲 $a \leq X \leq b$ に入る確率を考え，

$$P(a \leq\ X \leq b) = \int_a^b f(x)dx$$

と表す．ただし，f は以下の条件を満たす：

$$f(x) \geq\ 0, \quad \text{かつ} \quad \int_{-\infty}^{\infty} f(x)dx = 1$$

このとき，関数 $f(x)$ を X の**確率密度関数**と呼ぶ．また，離散型の場合と同様に，X の取る値が x 以下である確率を

$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

と表し，これを**累積分布関数**と呼ぶ．なお，微分積分学の基本定理より，累積分布関数と確率密度関数は

$$f(x) = \frac{dF(x)}{dx}$$

の関係にある．

例として，身長の確率分布 $f(x)$ が図 3.3 のように与えられるとする．このとき，身長が $165 \leq X \leq 175$ の範囲にある確率は

$$P(165 \leq X \leq 175) = \int_{165}^{175} f(x)dx$$

によって求めることができる．これは，図 3.3 の灰色部分の面積を求めていることになる．なお，この確率分布は**正規分布**と呼ばれ，次章で詳細を説明する．

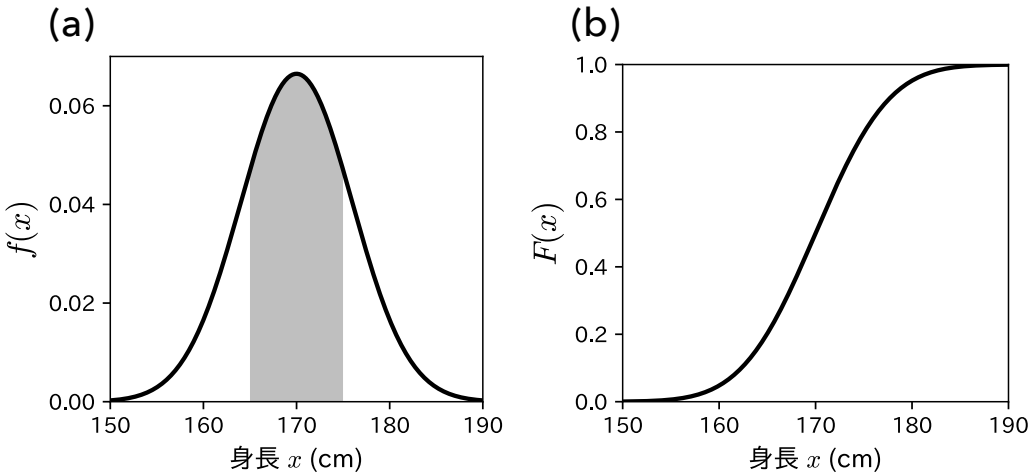

図 3.3. 身長の確率分布（正規分布）．(a) 確率密度関数，(b) 累積分布関数．

3.2.2 期待値と分散

一般に，確率変数が従う分布が与えられると，確率変数の各実現値がどのような確率で生じるか完全に分かる．つまり，確率関数や確率密度関数は確率変数に関する情報をすべて含んでいる．一方，現実の問題では，確率分布に関する細かい情報まで知る必要はなく，大体どの程度の値を取るかが分かれば良い場合も多い．このような場合，確率密度関数や確率分布関数を要約した統計量を用いたほうが良く，その代表例が期待値や分散・標準偏差である．これは，実データからヒストグラムを求めてその要約のために平均や分散を求めたのと同じことである．

期待値

確率分布においてもヒストグラムにおける平均値に対応する量を定義でき，これを**期待値**と呼ぶ．これは，試行の結果期待される値という意味である．一般に，確率変数 X に対する期待値は $E(X)$ または μ と書き，離散型，連続型それぞれに対して以下で定義される：

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} x_i f(x_i) \quad \text{（離散型）} \\ E(X) &= \int_{-\infty}^{\infty} x f(x) dx \quad \text{（連続型）} \end{aligned} \tag{3-1}$$

ここで，離散型の場合はヒストグラムにおける標本平均の式 (2-3) と対応していることが分かる［式 (2-3) では f_i/n が相対頻度を表し，これが式 (3-1) の $f(x_i)$ に対応する］．

例）ある「くじ」から得られる賞金をどれだけ期待できるかを表すの期待金額であり，これが確率変数の期待値の本来的な意味である．いま，1 から 100 までの番号がついた 100 個の玉が入っている箱から（毎回元に戻しながら）玉を 1 個取り出す．このとき，玉の番号に応じて賞金 X の金額が以下のように決まっているとする（単位は千円）：

- 番号が 1 から 60：$x_1 = 0$
- 番号が 61 から 90：$x_2 = 1$
- 番号が 91 から 100：$x_3 = 10$

これより，くじを 1 回引いて $x_1,\ x_2,\ x_3$ という結果が起きる確率（確率分布）は

$$f(x_1) = 0.6,\ f(x_2) = 0.3,\ f(x_3) = 0.1$$

である．このとき，くじを多数回引くときに得られる 1 回当たりの金額が期待値であり，式 (3-1) より以下のように計算される：

$$\begin{aligned} E(X) &= \sum_{i=1}^{3} x_i f(x_i) \\ &= x_1 f(x_1) + x_2 f(x_2) + x_3 f(x_3) \\ &= 0 \times 0.6 + 1 \times 0.3 + 10 \times 0.1 \\ &= 1.3 \text{（千円）} \end{aligned}$$

分散

期待値は分布の重心を表す指標であるが，期待値が同じでも形状が異なる分布はたくさんある．そこで，分布の形状に関するより詳しい情報を得るには，分布のばらつき具合を表す指標が必要となる．これが分散であり，$\mu = E[X]$ に対して

$$V(X) = E[(X - \mu)^2]$$

と定義される．特に，離散型，連続型の確率変数に対して以下のように与えられる：

$$
\begin{aligned}
V(X) &= \sum_i (x_i - \mu)^2 f(x_i) \quad \text{（離散型）} \\
V(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad \text{（連続型）}
\end{aligned}
\tag{3-2}
$$

期待値と同様に，分散の式 (3-2) もヒストグラムにおける分散の式 (2-8) と対応している．なお，確率分布の標準偏差は分散の平方根として定義される．

例）宝くじ

表 3.1 は 1 枚 300 円のある宝くじの賞金と当選確率の関係（確率分布）である．この宝くじの賞金 x に対してその期待値を計算すると，$\mu = 134$ 円となる．宝くじの値段 300 円に対して期待値が 134 円であり，買い手からすると明らかに損をするように見える．しかし，標準偏差の値は約 10 万円でありばらつきも非常に大きいことが分かる．これは，購入枚数が少なければ大勝する可能性がある一方で，購入枚数が増えるほど損をすることを意味する．

表 3.1. 宝くじの賞金・当選確率と期待値および分散の計算

賞	賞金 x（円）	当選確率 $f(x)$
1 等	3×10^8	$1/10^7$
1 等前後	1×10^8	$2/10^7$
2 等	1×10^7	$4/10^7$
3 等	1×10^5	$1/10^4$
4 等	1×10^4	$2/10^3$
5 等	2×10^3	$1/100$
6 等	3×10^2	$1/10$
外れ	0	0.8878993

3.2.3 代表的な離散型確率分布

一様分布： $f(x) = \dfrac{1}{b-a+1} \quad (a \leq x \leq b)$

$a \leq x \leq b$ において, 一定確率を取る分布．サイコロを 1 回投げた場合は $a = 1$, $b = 6$ の一様分布となる．

ベルヌーイ分布： $f(x) = p^x(1-p)^{1-x}$

成功確率が p，失敗確率が $1-p$ の試行を**ベルヌーイ試行**と呼ぶ．1 回のベルヌーイ試行において，成功，失敗を 1 と 0 に対応させた確率変数を X とすると，その確率分布はベルヌーイ分布となる．

二項分布： $f(x) = \dbinom{n}{x} p^x(1-p)^{n-x}$

ベルヌーイ試行を n 回繰り返すとき，成功回数 $\sum_{i=1}^{n} X_i$ を新たな確率変数 X とすると，その確率分布は二項分布となる．詳細は次章で扱う．

ポアソン分布： $f(x) = \dfrac{\mu^x}{x!}\mathrm{e}^{-\mu}$

二項分布において平均を $np = \mu$ とおき，μ を一定に保ったまま $n \to \infty$ とした場合に現れる確率分布．詳細は次章で扱う．

幾何分布： $f(x) = p(1-p)^x$

ベルヌーイ試行を繰り返すとき，初めて成功した時点で失敗した回数を確率変数 X とすると，その分布は幾何分布となる．

超幾何分布： $f(x) = \dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

赤玉 K 個と青玉 $N-K$ 個を混ぜた計 N 個の中から，n 個を取り出すとき，含まれている赤玉の数を確率変数 X とすると，その分布は超幾何分布となる．

負の二項分布： $f(x) = \dbinom{r+x-1}{x} p^r(1-p)^x$

ベルヌーイ試行を繰り返すとき，r 回成功した時点で失敗した回数を確率変数 X とすると，その分布は

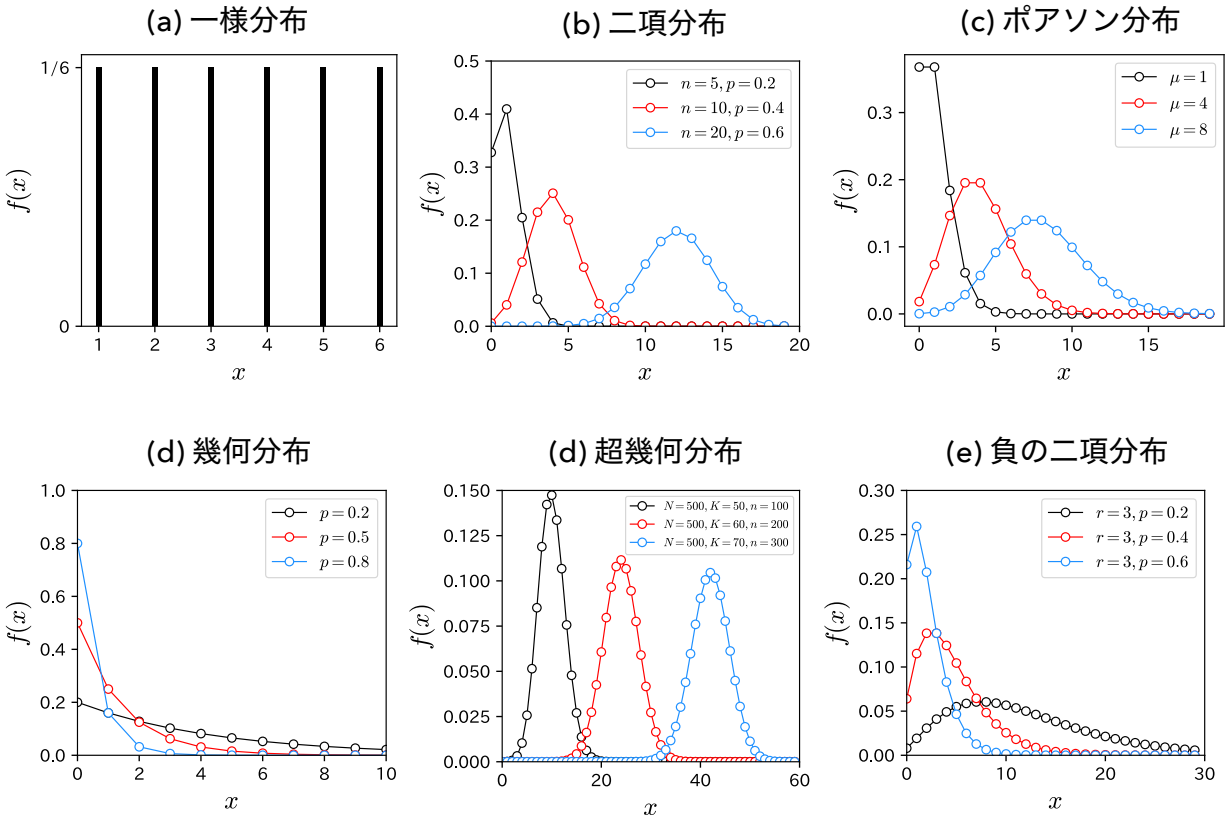

図 3.4. 代表的な離散型確率分布

3.2.4 代表的な連続型確率分布

一様分布： $f(x) = \dfrac{1}{b-a}, \quad (a \leq x \leq b)$

$f(x)$ が確率変数 X の値に依らず，一定値を取る分布．

正規分布： $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]$

様々な自然現象，社会現象において観られる確率分布．正規分布が現れるメカニズムには中心極限定理「任意の分布に従う n 個の確率変数の和の分布が n を大きくしたときに正規分布に近づく」がある．

指数分布： $f(x) = \dfrac{1}{\lambda}\mathrm{e}^{-x/\lambda}, \quad (x \geq 0)$

連続的な時間で事象が独立に一定の確率で生じるような確率過程（ポアソン過程）において，初めて事象が生じるまでの待ち時間分布は指数分布となる．これより，幾何分布の連続版と捉えることができる．

ガンマ分布： $f(x) = \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\mathrm{e}^{-x/\beta}, \quad (x > 0)$

ポアソン過程において，事象が α 回生じるまでの待ち時間はガンマ分布に従う．これは，指数分布に従う確率変数の和の分布がガンマ分布であることを意味する．

対数正規分布： $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\dfrac{(\log x - \mu)^2}{2\sigma^2}\right], \qquad (x > 0)$

確率変数 X の対数変換 $Y = \log X$ が正規分布に従うとき，X は対数正規分布に従う．一般に，確率変数 X_t，α_t に対して，X_t の時間発展が

$$X_t = \alpha_{t-1} X_{t-1}$$

で与えられるときに，$t \to \infty$ における X_t の分布は対数正規分布に従う．これは，ある量 X_{t-1} にランダムな成長率 α_t を掛けたものが次の時刻の値 X_t になるということを意味し，**ランダム乗算過程**と呼ばれる．基本的に，近似的にでもこのようなプロセスで成長する現象では対数正規分布が観られる．例えば，ガラス棒を落として破壊する実験を考える．このとき，時刻 t における破片の大きさを X_t とすると，X_t の時間発展はランダム乗算過程で記述できる[*2]．社会現象における対数正規分布の例としては，高齢者の死亡年齢，児童生徒の身長・体重，駅の降車人数，アニメのキャラクターのサイズ，など枚挙にいとまがない [9]．

べき分布： $f(x) = Cx^{-\alpha}$

べき分布の著しい性質として，分布を特徴づける平均や分散が意味を成さない点がある．実際，べき分布では極端に大きい値を取る確率が無視できないので，平均値は分布の中心を意味せず，分散は発散したりする．しかし，自然現象や社会現象にはべき分布に従う現象が数多く観測されており，一例として，地震のエネルギー，個人の資産，都市の人口，テキスト中に出現する単語の頻度，本や音楽の売上げ，論文の引用回数などが挙げられる．例えば，個人の資産がべき分布に従うということは，一部の人が莫大な資産を有する一方で，大多数の人は平均以下の資産しか持たないことを意味する．べき分布が現れる背景には要素間の複雑な相互作用が存在することが多く，その出現メカニズムは未だに研究対象となっている [9, 10]．

[*2] ただし，落とす高さなどにも依存する

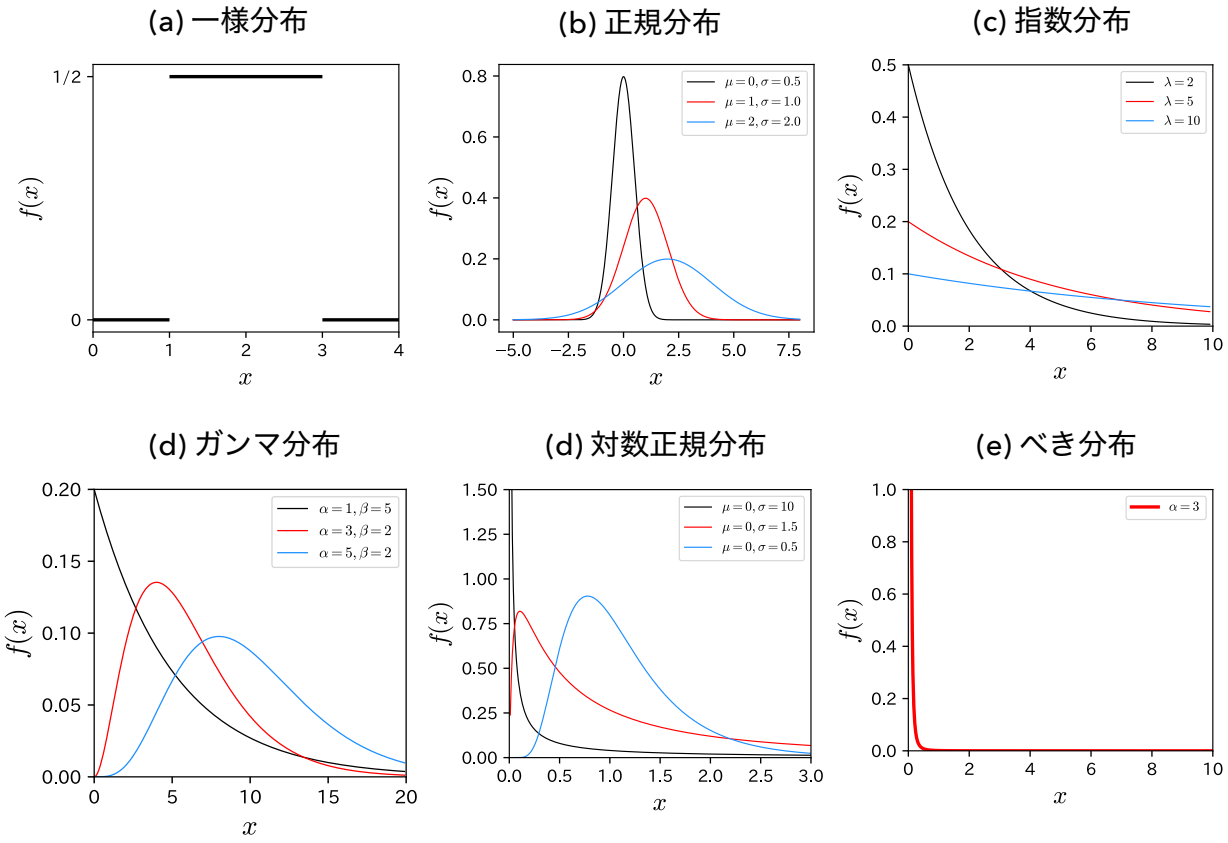

図 3.5. 代表的な連続型確率分布

第 4 章

確率分布の応用

4.1　二項分布から正規分布へ

4.1.1　二項分布

1 回の試行において，起こりうる事象が 2 種類しかない場合，これを**ベルヌーイ試行**と呼ぶ．通常は，2 種類の事象をそれぞれ成功（1），失敗（0）に対応付けた確率変数 U を考え，成功確率を p，失敗確率を $1-p$ とする．このとき，確率分布は

$$P(U=u) = p^u(1-p)^{1-u}$$

となり，これを**ベルヌーイ分布**と呼ぶ．例えば，コイン投げは典型的なベルヌーイ試行である．

いま，ベルヌーイ試行を独立に n 回繰り返したとき，成功の回数 $X = \sum_{i=1}^{n} U_i$ を新たな確率変数とする．このとき，成功が x 回，失敗が $n-x$ 回生じたとすると，その確率分布は

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{4-1}$$

で与えられる．この式において，$p^x(1-p)^{n-x}$ は成功が x 回，失敗が $n-x$ 回生じる確率を意味する．また，$\binom{n}{x}$ は n 個から x を取り出す組み合わせの数 ${}_nC_x$ を表し，n 回の中で何回目に成功するかの場合の数に対応する．この分布は**二項分布**と呼ばれ，離散型確率分布の中でも代表的な分布の 1 つである．なお，$n=1$ の場合は**ベルヌーイ分布**と呼ばれる．

確率変数 X が二項分布に従うとき，その期待値と分散は

$$E(X) = np, \qquad V(X) = np(1-p)$$

で与えられる．$E(X)$ は n 回の試行による成功回数の期待値であるから，試行回数に成功確率を乗じた np となることは直感に合う．

二項分布は試行回数 n と確率 p によって分布の形が決まる．図 4.1(a) は p を固定して n を変えたときの分布の変化である．n が小さいときには左右非対称な離散的分布であるが，n が大きくなるにつれて左右対称で滑らかな分布に近づいていくことが分かる．

$E(X) = np,\ V(X) = np(1-p)$ **の証明**

$$
\begin{aligned}
E[X] &= \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np \sum_{x'=0}^{n-1} \frac{(n-1)!}{x'!(n-1-x')!} p^{x'} (1-p)^{n-1-x'} \\
&= np \sum_{x'=0}^{n-1} \binom{n-1}{x'} p^{x'} (1-p)^{n-1-x'} \\
&= np \\
E[X^2] &= \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} (x^2 - x + x) \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1) \binom{n}{x} p^x (1-p)^{n-x} + \sum_{x=0}^{\infty} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} + np \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x} + np \\
&= n(n-1) \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!\{(n-2)-(x-2)\}!} p^x (1-p)^{n-x} + np \\
&= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!\{(n-2)-(x-2)\}!} p^{x-2} (1-p)^{n-x} + np \\
&= n(n-1)p^2 \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} (1-p)^{(n-2)-(x-2)} + np \\
&= n(n-1)p^2 \sum_{x'=0}^{n-2} \binom{n-2}{x'} p^{x'} (1-p)^{(n-2)-x'} + np \\
&= n(n-1)p^2 + np \\
V[X] &= E[X^2] - E[X]^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= np(1-p)
\end{aligned}
$$

大数の法則

ここまでは成功確率 p のベルヌーイ試行を n 回繰り返したときの成功回数 $X = \sum_{i=1}^{n} U_i$ を確率変数としていたが，以下では成功の割合

$$T = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} U_i$$

を新しい確率変数と考えてみる．（これは，確率変数 U の標本平均と捉えることもできる．）

いま，成功割合 T の確率分布を $g(t)$ とすると，$g(t)$ も二項分布に従う[*1]．図 4.1(b) は成功確率 p を一定値 $p = 0.2$ に固定して，試行回数 n を大きくしたときの成功割合 $T = X/n$ の確率分布である．この図を見ると，n の増加に伴い $x/n = 0.2$ の周りに分布が集中してきて，高さが大きくなる様子が分かる．

この様子を式で見てみる．まず，成功回数 X の期待値は np であるから，成功割合 $T = X/n$ の期待値は p になる．同様にして，成功割合 X/n の分散は $p(1-p)/n$ になる．これより，成功割合 X/n の期待値は n に依らず一定 p で，分散は n とともに 0 に近づくことが分かる．これが，成功割合 X/n の分布が n の増加とともに p の近くに集中する理由であり，特に，$n \to \infty$ のベルヌーイ試行では，成功割合 X/n が理論値 p に一致する．

以上のように，試行回数 $n \to \infty$ の極限で成功割合が理論値 p に一致する性質は**大数の法則**と呼ばれる．なお，ここではベルヌーイ試行を例に説明したが，大数の法則はより一般に成り立つ法則である[*2]．

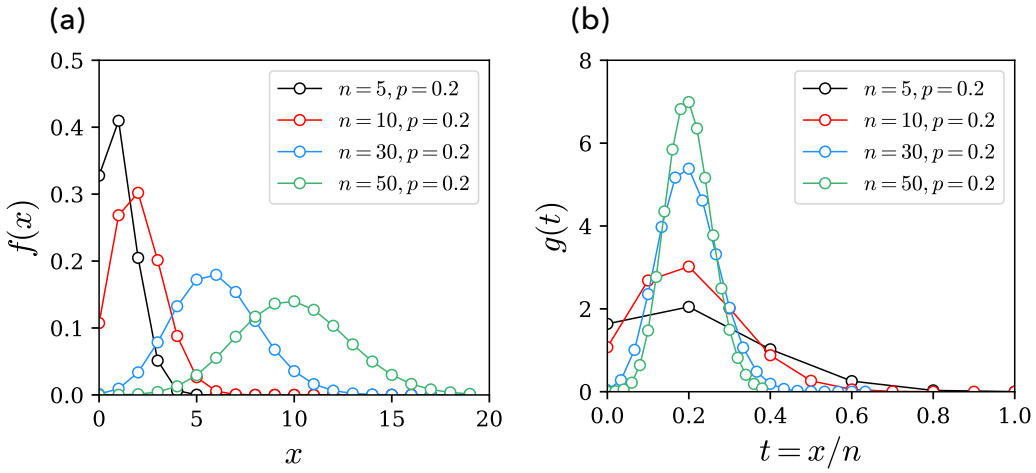

図 4.1. ベルヌーイ試行を n 回繰り返したときの成功回数 X の分布 (a) と成功割合 $T = X/n$ の分布 (b) ．共に二項分布に従う．

[*1] より詳しくは $g(t) = nf(nt)$ の関係にある．

[*2] 独立同分布に従う n 個の確率変数 $U_1, U_2, \ldots, U_n$ の標本平均 $\sum_{i=1}^{n} U_i/n$ が $n \to \infty$ で $E[U_i]$ に一致する．

4.1.2　正規分布

前節では，成功割合 $T = X/n$ の分布が n を大きくしたときに理論値 p の周りに集中し，大数の法則が成り立つことを見た．また，n を大きくしていくとき，成功回数 X や成功割合 T の分布（いずれも二項分布）が左右非対称から左右対称な形へと変化することも見た．実は，n を十分大きくしたときに出現する左右対称で滑らかな分布は**正規分布**と呼ばれており，これは中心極限定理による帰結である．中心極限定理とは，「同じ確率分布に従う n 個の確率変数の和の分布が n を大きくしたときに正規分布に近づく」というものである．今の場合，成功回数を表す確率変数 X は，成功を 1，失敗を 0 とした確率変数 U_i の和 $X = \sum_{i=1}^{n} U_i$ となっているので，次のように言い換えられる

> 「ベルヌーイ分布に従う n 個の確率変数の和 $X = \sum_{i=1}^{n} U_i$ の分布が n を大きくしたときに正規分布に近づく」

一般に，正規分布は連続型確率変数 X の従う確率分布で，確率密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{4-2}$$

で与えられる．ここで，μ と σ^2 は正規分布の期待値と分散に対応する：

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

以下では，この形の正規分布を $N(\mu, \sigma^2)$ と表す．正規分布 $N(\mu, \sigma^2)$ は平均 μ と分散 σ^2 によって形状が図 4.2 のように変わる．特に，平均 μ は分布のピークの位置に対応し，分散 σ^2 は分布の広がりを決める．正規分布は中心極限定理が背景にある多くの自然現象や社会現象において観られるので，統計学の理論上最も重要な分布である．

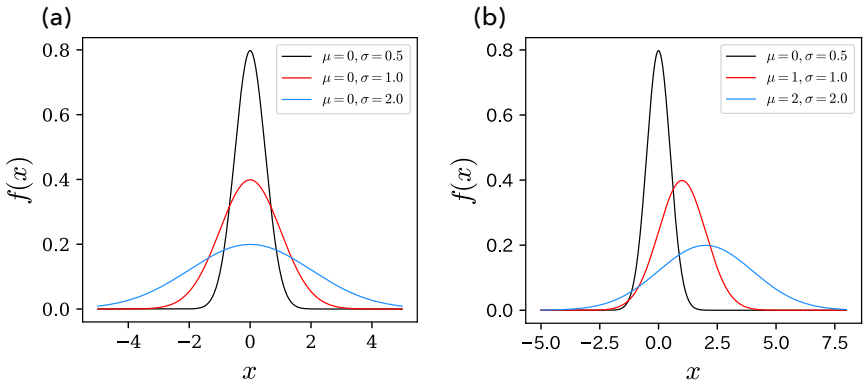

図 4.2. 正規分布の例

$E(X)=\mu,\ V(X)=\sigma^2$ の証明

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx \\
&\left(\frac{x-\mu}{\sigma}=y \text{ とおくと, } dx=\sigma dy\right) \\
&= \int_{-\infty}^{\infty}(\sigma y+\mu)\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{y^2}{2}}\sigma dy \\
&= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y\mathrm{e}^{-\frac{y^2}{2}}dy+\frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{y^2}{2}}dy \\
&\left(\text{ガウス積分}\int_{-\infty}^{\infty}\mathrm{e}^{-ay^2}dy=\sqrt{\frac{\pi}{a}}\right) \\
&= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{d}{dy}\left(-\mathrm{e}^{-\frac{y^2}{2}}\right)dy+\mu \\
&= \frac{\sigma}{\sqrt{2\pi}}\left[-\mathrm{e}^{-\frac{y^2}{2}}\right]_{-\infty}^{\infty}+\mu \\
&= \mu \\
V[X] &= \int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx \\
&\left(\frac{x-\mu}{\sigma}=y \text{ とおくと, } dx=\sigma dy\right) \\
&= \int_{-\infty}^{\infty}\sigma^2y^2\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{y^2}{2}}\sigma dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y^2\mathrm{e}^{-\frac{y^2}{2}}dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}-y\frac{d}{dy}\mathrm{e}^{-\frac{y^2}{2}}dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}}\left[-y\mathrm{e}^{-y^2}\right]_{-\infty}^{\infty}+\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{y^2}{2}}dy \\
&\left(\text{ガウス積分}\int_{-\infty}^{\infty}\mathrm{e}^{-ay^2}dy=\sqrt{\frac{\pi}{a}}\right) \\
&= \sigma^2
\end{aligned}
$$

標準化と標準正規分布

正規分布は連続型確率分布であるので，確率変数がある範囲に存在する確率を求めるには積分を実行する必要がある．しかし，この積分は初等的な方法では求まらないので，通常はコンピュータを使って計算する．正規分布の積分の結果をまとめた表は正規分布表と呼ばれている．ただし，正規分布は μ と σ に依存して形を変えるので，通常は $\mu = 0, \sigma = 1$ に変換した確率変数を考える．この変数変換は**標準化**と呼ばれ，確率変数 X が正規分布 $N(\mu, \sigma^2)$ に従うとき

$$Z = \frac{X - \mu}{\sigma} \tag{4-3}$$

によって与えられる．この式より，Z というのは，X が平均 μ から標準偏差 σ いくつ分ずれた位置にあるかを表している．例えば，$Z = 2$ であれば，X は μ から 2σ だけずれた位置にあることを意味する．標準化した確率変数 Z は**標準化得点**あるいは Z **値**と呼ばれ，$N(0, 1)$ に従う：

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \tag{4-4}$$

平均 0，標準偏差 1 の正規分布 $N(0, 1)$ は**標準正規分布**と呼ばれている．

通常，正規分布表には，標準正規分布の上側累積確率

$$P(Z \geq a) = \int_a^\infty \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] dz$$

がまとめられている．具体例として，$\mu - a\sigma < X < \mu + a\sigma$ となる確率を正規分布表から求める方法は以下のようになる．まず，X を標準化すると，

$$P(\mu - a\sigma \leq X \leq \mu + a\sigma) = P(-a \leq Z \leq a)$$

となる．よって，

$$\begin{aligned} P(-a \leq Z \leq a) &= \int_{-a}^\infty f(z)dz - \int_a^\infty f(z)dz \\ &= 1 - 2\int_a^\infty f(z)dz \\ &= 1 - 2P(Z \geq a) \end{aligned}$$

となるので，正規分布表から $P(Z \geq a)$ を探せば計算できる．以下は，$a = 1,\ 2,\ 3$ の場合である：

- 区間 $(\mu - 1\sigma, \mu + 1\sigma)$ に入る確率は 0.683 である
- 区間 $(\mu - 2\sigma, \mu + 2\sigma)$ に入る確率は 0.954 である
- 区間 $(\mu - 3\sigma, \mu + 3\sigma)$ に入る確率は 0.997 である

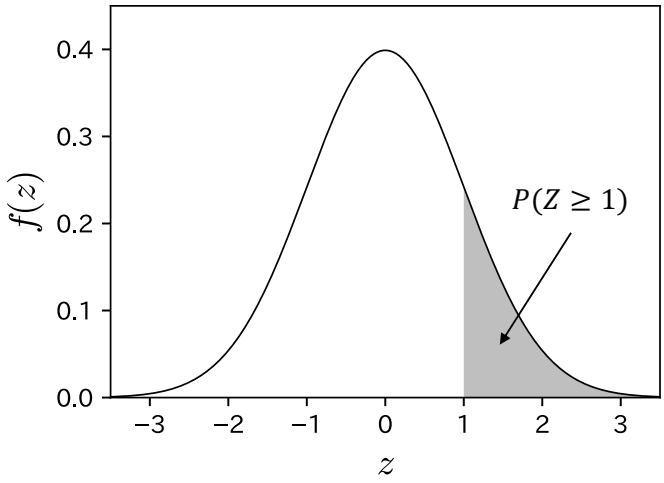

図 4.3. 標準正規分布

偏差値

あるテストの平均値を μ，標準偏差を σ とする．このとき，テストの得点を X とすれば，偏差値 H は Z 値を用いて以下のように定義される：

$$H = 50 + 10 \times Z = 50 + 10 \times \frac{X - \mu}{\sigma} \tag{4-5}$$

この式より，偏差値とはグループの中で平均値からどの程度乖離しているか（平均値から標準偏差いくつ分離れているか）を表す数値である．特に，平均値の場合に 50，そこから 1 標準偏差離れるごとに 10 加算または減算される量であることが分かる．例えば，偏差値 50 の場合にはグループ全体のちょうど真ん中，60 の場合には $\mu + 1\sigma$，70 の場合には $\mu + 2\sigma$ に位置していることになる．もし，テストの得点分布が正規分布 $N(\mu, \sigma)$ に従っていれば，偏差値 50 は上位 50%，偏差値 60 は上位 16%，偏差値 70 は上位 2% 程度に位置することになる．

4.1.3　実例：視聴率調査の仕組みは？

Step 1: Problem

2016 年 10 月 31 日時点で関東地区の世帯数は約 1800 万世帯である．これらの世帯の中で，ある番組を見ている世帯の割合を表したものが番組視聴率である．通常，全世帯の視聴率を完全に把握するには全世帯を調査する必要があるが，それは現実的ではない．そこで，一部の世帯（n 世帯）だけを抽出し（これをサイズ n の標本と呼ぶ），そこでの視聴率調査から全世帯の視聴率を推定する方法が取られる．実際，ビデオリサーチ社の視聴率調査において調査対象となる世帯数は関東地区で 900 世帯（つまり，サイズ 900 の標本）となっている．では，どのような方法で 900 世帯のデータから全体の視聴率を推定しているのだろうか？

Step 2: Plan

以下のような視聴率調査を模した模擬実験を考え，世帯数と視聴率調査の正確性の関係を調べる．まず，黒玉を「番組を見た世帯」，白玉を「番組を見ていない世帯」とする．黒玉は 20 個，白玉は 60 個用意して箱の中に入れる．すなわち，視聴率の理論値は 20/80=25% となる．実験では，まず，箱の中から 4 個の玉を取り出すことを 200 回繰り返し，黒玉の比率（視聴率）のヒストグラムを作成する．次に，標本サイズが $n = 12, 15, 30$ の場合に対して同様のことを繰り返し，ヒストグラムの変化を見る．

理論的には，箱から玉を n 個取り出したときの黒玉の個数を確率変数 X で表すと（つまり，黒玉を 1，白玉を 0 とした確率変数の和が X），X/n の分布は二項分布に従い，n が大きい極限では中心極限定理より正規分布に近づく．よって，今回の実験でも n を増やしていくとヒストグラムは正規分布に近づくと考えられる．

Step 3: Data

$n = 4, 12, 15, 30$ の場合に模擬実験を 200 回繰り返した結果，表 4.1〜表 4.4 のような結果を得た．

表 4.1. $n = 4$ の場合

黒玉の比率（%）	0	25	50	75
頻度	70	80	40	10

表 4.2. $n = 12$ の場合

黒玉の比率（%）	0	8	17	25	33	42	50
頻度	5	30	50	45	50	15	5

表 4.3. $n = 15$ の場合

黒玉の比率（%）	0	7	13	20	27	33	40	47	53
頻度	3	6	30	42	54	36	18	6	5

表 4.4. $n = 30$ の場合

黒玉の比率（%）	3	7	10	13	17	20	23	27	30	33	37	40	43
頻度	2	1	3	5	7	28	46	56	18	14	10	8	2

Step 4: Analysis

実習

- 表 4.1〜4.4 のデータを用いて，各 n に対する黒玉比率のヒストグラムを作成せよ．
- n の増加に応じてヒストグラムの形がどのように変化するか確認せよ．
- 各 n に対して黒玉比率の標本平均と標準偏差を計算せよ．

Step 5: Conclusion

実習

- 実験の結果得られたヒストグラムについて，大数の法則の観点から考察せよ．
- 実験の結果得られたヒストグラムについて，中心極限定理の観点から考察せよ．
- 実験全体を踏まえ，視聴率調査の仕組みについて考察せよ．

4.2 二項分布からポアソン分布へ

4.2.1 ポアソン分布

成功確率 p のベルヌーイ試行を独立に n 回繰り返したとき，成功回数 X は二項分布

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

に従い，期待値（平均値）は np となる．前節では，試行回数 n が大きい場合に二項分布が正規分布に近づくこと（中心極限定理）を見た．ここでは，二項分布から別の極限を取ったときに得られるポアソン分布について説明する．

いま，成功確率 p が小さく，かつ試行回数 n が大きい極限を考える．ただし，極限を取る際に平均値が一定値 $np = \mu$ になるようにする．このような条件で成功回数 X が従う分布は，二項分布の式に $np = \mu$ を代入し，極限 $p \to 0$, $n \to \infty$ を取ることで

$$f(x) = \frac{\mu^x}{x!} \mathrm{e}^{-\mu} \tag{4-6}$$

と求まる．これを**ポアソン分布**と呼ぶ．ポアソン分布は 1 つのパラメータ μ だけで特徴づけられ，期待値と分散はともに μ となる．図 4.4 は μ を変化させた場合のポアソン分布の変化である．この図からも分かるように，μ が大きいときにはポアソン分布は左右対称な分布となり，正規分布に近づく[*3]．

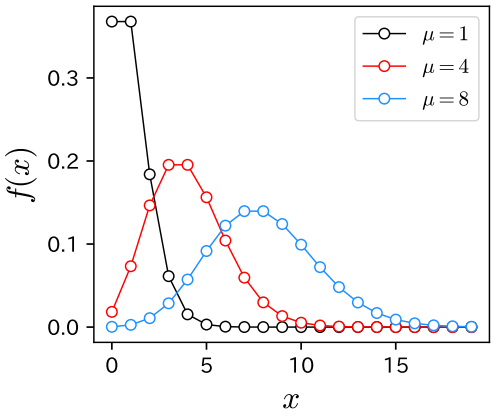

図 4.4. ポアソン分布の例．横軸はある期間内の成功回数．

ポアソン分布は，一定の期間内（例えば 1 時間や 1 日）に，稀な現象（$p \to 0$）を多数回試行（$n \to \infty$）した場合にその発生回数が従う分布である．ポアソン分布が現れる例は無

[*3] 一見すると，ポアソン分布の導出の時点で $n \to \infty$ としているので，中心極限定理が適用できて正規分布になりそうである．しかし，np を一定に保ちながら $p \to 0$, $n \to \infty$ とするような極限は中心極限定理の適用範囲外となる．

数にあり，「1 日の交通事故件数」，「1 分間の放射性元素の崩壊数」，「1 ヶ月の有感地震の回数」，「サッカーの試合における 90 分間の得点数」などは典型例である．

例）プロシア陸軍には 200 の部隊がある．1875 年から 1894 年までの 20 年間に，馬に蹴られて死亡した兵士の数を各部隊について調べると，その頻度分布は表 4.5 のようになる．このデータから，20 年間の 1 部隊あたりの平均死亡者数は 0.61 人である．これを踏まえ，（プロシア陸軍において）20 年間で馬に蹴られて死亡する人の数を確率変数 X とし，X が $\mu = 0.61$ のポアソン分布に従うとする．このとき，死亡者数が x 人である部隊数の理論値 $N(x)$ は

$$N(x) = 200\frac{0.61^x}{x!}\mathrm{e}^{-0.61}$$

となる．これを計算すると，

$$N(0) = 108.7,\ N(1) = 66.3,\ N(2) = 20.2,\ N(3) = 4.1,\ N(4) = 0.6$$

となり，表 4.5 のデータとよく一致する．

表 4.5

死亡者数	0	1	2	3	4
部隊数	109	65	22	3	1

ポアソン分布の導出

$np = \mu$ を式 (4-1) に代入すると，

$$\begin{aligned} f(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x} \\ &= \frac{1(1-\frac{1}{n})\cdots(1-\frac{x-1}{n})}{x!} n^x \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^n \left(1-\frac{\mu}{n}\right)^{-x} \end{aligned}$$

ここで，$n \gg x$ より，$n \to \infty$ において

$$1(1-\frac{1}{n})\cdots(1-\frac{x+1}{n}) \to 1$$

$$\left(1-\frac{\mu}{n}\right)^n \to \mathrm{e}^{-\mu}$$

$$\left(1-\frac{\mu}{n}\right)^{-x} \to 1$$

となる．以上より，

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

が得られる．

4.2.2　実例：サッカーとバスケの得点頻度の違いは？

Step 1: Problem

サッカーとバスケットボールは様々なスポーツの中でも特に高い人気を誇る競技である．これらの競技の特徴は得点がランダムに入る（どちらのチームがいつ得点するかが予測不能）という点であり，このランダム性こそが人々を魅了する理由と考えられる．また，バスケは 1 試合の得点数が比較的大きい競技であるが，サッカーはほとんど点が入らない競技として有名である．では，こうした特徴は，統計的にはどのように定量化できるのだろうか？

Step 2: Plan

チームの強さや試合展開など細かいことはひとまず無視し，サッカーやバスケにおける得点イベントがランダムに発生すると考える．特に，両チームが常に得点を目指し一瞬で得点チャンスが生まれることから，試合中のどの時点においても一定の得点確率があると見なし，得点確率 p の試行を何度も繰り返す現象（$n \to \infty$）と捉えることにする．また，各時点で得点する確率は非常に小さいとする（$p \ll 1$）．以上のような仮定を置くと，サッカーやバスケにおける 1 試合の得点数はポアソン分布に従うことが期待される．

Step 3: Data

web 上で公開されているバスケットボールとサッカーのオープンデータセットを用いる．各データセットの詳細は以下の通りである．

バスケットボール

nba-movement-data [11]（MIT ライセンス）を用いる．データは JSON 形式で，イベントデータ，シュートデータ，トラッキングデータが含まれる．対象試合は NBA2015-16 シーズンの約 500 試合である．イベントデータにはプレーごとの時空間情報（例えば，パスが行われた座標，時刻，関わった選手など），シュートデータには試合中の全てのシュートの情報，トラッキングデータには全選手の 0.04 秒ごとの座標（トラッキングデータ）が含まれる．

サッカー

Pappalardo データセット [12]（CC BY 4.0 ライセンス）を用いる．データは JSON 形式で，イベントデータのみ含む．対象試合はヨーロッパリーグ（イングランド，イタリア，フランス，ドイツ，スペイン）2017-18 シーズンの約 1400 試合である．

Step 4: Analysis

サッカーとバスケの 1 試合の得点頻度を集計したところ，表 4.6 および表 4.7 のような結果を得た．この結果から，1 試合の得点を横軸，その度数（試合数）を縦軸にとってヒストグラムを描くと図 4.5 が得られた．ここで，図中の丸印は，1 試合の平均得点をパラメータとするポアソン分布のモデル値を表している．この結果から，サッカーとバスケの 1 試合の得点分布はおおよそポアソン分布に従っていることが分かる．なお，1 試合の平均得点はサッカーが 1.6 点，バスケが 36 点となった．

表 4.6. サッカー（ドイツリーグ）の 1 試合の得点分布

得点	0	1	2	3	4	5	6
試合数	61	101	79	41	16	4	4

表 4.7. NBA の 1 試合の得点分布

得点	[14, 17]	[18, 21]	[22, 25]	[26, 29]	[30, 33]	[34, 37]	[38, 41]	[42, 45]	[46, 49]	[50, 53]
試合数	1	2	6	45	112	142	129	68	16	2

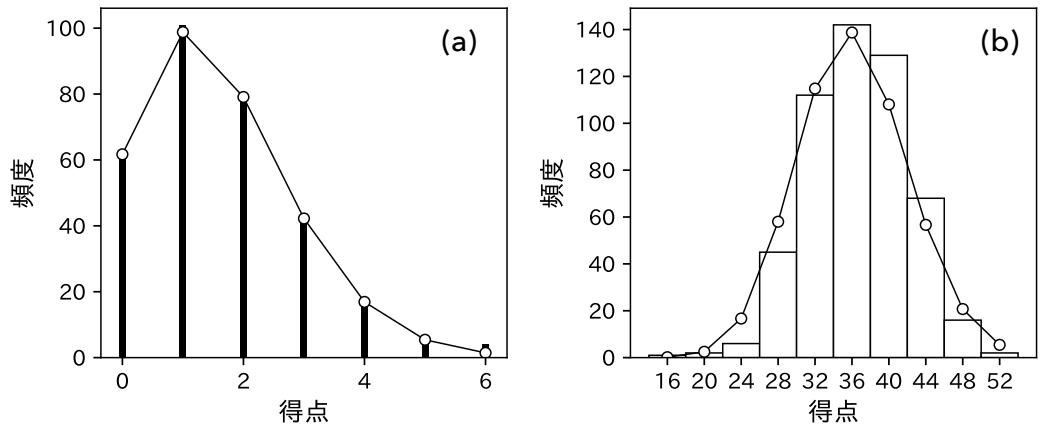

図 4.5. (a) サッカー（ドイツリーグ）の 1 試合の得点分布．(b) NBA の 1 試合の得点分布．丸印は各データの平均得点をパラメータとしたポアソン分布のモデル値．

Step 5: Conclusion

サッカーとバスケの 1 試合の得点はポアソン分布に従っていると考えられる．これにより，サッカーとバスケは得点確率の小さい試行を何度も繰り返すような現象であることが確認できる．一方，1 試合の平均得点はバスケとサッカーで大きく異なった．特に，バスケの

平均得点はサッカーに比べて大きい値となったが，これによりヒストグラムが左右対称となり正規分布による近似が成り立っていることが確認できる．こうした特徴は，バスケにおける促進ルール（24 秒以内にシュートを打たなければならないなど）を反映したものであると考察できる．

参考文献

[1] 総務省政策統括官政策基準部編集，高校からの統計・データサイエンス活用〜上級編〜，日本統計協会，2017，DL 用 URL：https://www.soumu.go.jp/toukei_toukatsu/info/guide/stkankyo.htm.

[2] 竹村彰通・姫野哲人・高田聖治編，データサイエンス入門，学術図書出版社，2019.

[3] 深 KOKYU,「間隔尺度と比例尺度」，https://haru-reha.com/interval_scale_proportional_scale/，2021 年 7 月 8 日最終アクセス.

[4] 間隔尺度と比尺度の違いを説明するための例，https://note.com/celestia1212/n/n4b39b3b36055，2021 年 7 月 8 日最終アクセス.

[5] 東京大学教養学部統計学教室編，統計学入門，東京大学出版会，1991.

[6] 薩摩順吉，確率・統計，岩波書店，2019.

[7] 栗原伸一，入門統計学（第 2 版），オーム社，2021.

[8] 中室牧子，津川友介,「原因と結果」の経済学，ダイヤモンド社，2017.

[9] 松下貢,「統計分布を知れば世界が分かる」，中公新書，2019.

[10] マーク・ブキャナン,「歴史はべき乗則で動く」, ハヤカワ文庫 NF，2009.

[11] “sealneaward/nba-movement-data”, https://github.com/sealneaward/nba-movement-data.

[12] L. Pappalardo et al., A public data set of spatiotemporal match events in soccer competitions, Scientific Data 6, 2019. https://figshare.com/collections/Soccer_match_event_dataset/4415000/2