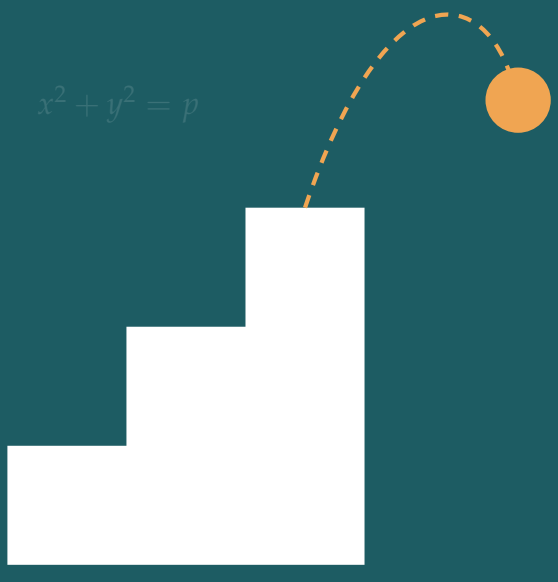
STEP 対策 · 大学の数学

Beyond STEP

第 13 章

世の中の統計とランダムネス



KEITH McNULTY

Beyond STEP

大学数学への入り口としての
STEP 過去問

第 13 章

世の中の統計とランダムネス

Keith McNulty

Beyond STEP：大学数学への入り口としての STEP 過去問

第 13 章：世の中の統計とランダムネス

これは *Beyond STEP* 全 13 章のうちの第 13 章である。本書全体、各章単独の PDF、各節の印刷用ワークシート、そしてすべての STEP の問題の詳しい解答は、`beyondstep.keithmcnulty.org` で無料で閲覧・ダウンロードできる。

2026 年 11 月、オンラインにて初版発行。

本 PDF のうち著者自身の著作物（本文、解答、図、演習問題とその答えを含む）はすべて、クリエイティブ・コモンズ CC0 1.0 全世界（Creative Commons CC0 1.0 Universal）ライセンスのもとでパブリックドメインに提供されている。許可を求めることなく、いかなる目的でも複製、改変、共有してよい。詳しくは `creativecommons.org/publicdomain/zero/1.0` を参照のこと。

ここで扱う STEP の問題の著作権は Cambridge University Press & Assessment に帰属する。問題そのものは転載しておらず、各問題には公式の出典である STEP Support Programme の問題データベース、または OCR の過去問へのリンクを付している。

誤りの指摘は、`github.com/keithmcnulty/step_book_mcnulty` への issue として歓迎する。

章

13

世界における統計とランダムネス

確率はモデルからデータへと進む。統計はデータから世界へと引き返す。

この章の内容

13.1	見出しを読み解く	STEP I 1992
13.2	ハタネズミを数える	STEP II 2000
13.3	ドイツ戦車問題	STEP III 1998
13.4	平均か中央値か？	STEP II 2000
13.5	答えにくい質問をする	STEP II 2023
13.6	積み重なる誤差	STEP II 1993
13.7	相関をつくり出す	STEP III 2010
A	演習問題の答え	p. 53

データから学ぶ

> すべてのモデルは間違っている。だが、役に立つものもある。
>
> — ジョージ・ボックス（George Box）、統計学者

確率（probability）では、さいころにどのような細工がされているかを知ったうえで、何が起こるかを問う。統計学（statistics）では、何が起こったかを見たうえで、さいころにどのような細工がされているはずかを問う。確率は原因から結果へと推論し、統計学は結果から原因へと遡って推論する。人口の推定、世論調査、臨床試験、科学的な測定は、いずれも統計的推測（statistical inference）である。すなわち、限られた、ノイズを含むデータから世界についての結論を引き出すことである。

統計学は、数学の大部分と比べると若い学問である。その現代的な形は、20 世紀初頭にカール・ピアソン（Karl Pearson）、ロナルド・フィッシャー（Ronald Fisher）、イェジ・ネイマン（Jerzy Neyman）らによって作られたが、その動機はしばしば実際的な問題にあった。ビールの醸造（t 検定（t-test）はギネスの醸造所で考案された）、作物の栽培、医薬品の試験などである。データが爆発的に増えた今日、統計学とその近縁である機械学習（machine learning）は、数学の応用のなかでも最も重要なものの一つとなっている。

なぜ統計学が重要なのか

医薬品、公衆衛生、経済政策、気候、司法に関する判断は、いずれも統計的な根拠に基づいている。それを読み誤れば、現実に影響が及ぶ。偽陽性（false positive）に怯える患者（Section 12.2）、偶然の変動に基づいた政策（Section 13.1）、小さな誤差が積み重なっていくコンピュータシステム（Section 13.6）などがその例である。データから何がわかり何がわからないのか、そしてどんな結論にもどれほどの不確かさがあるのかを理解することは、見識ある市民であるための条件の一つであり、科学者、医師、技術者、経済学者にとっては職業上欠かせないものである。

大学で出会うもの

統計学の最初の講義では、推定（estimation、未知の量のよい推定値を求め、その精度を述べること）、仮説検定（hypothesis testing、データが仮説と整合するかどうかを判

断すること）、信頼区間（confidence interval）、回帰（regression、ある変数を他の変数と関係づけること）、そして実験計画（design of experiments）と標本調査の設計を扱う。中心となる考え方は、**標本分布**（sampling distribution、推定量は確率変数（random variable）である）、**偏り**（bias）と**分散**（variance）、そして**尤度**（likelihood）である。さらに進んだ講義では、ベイズ推論（Bayesian inference）、多変量解析（multivariate analysis）、時系列（time series）、そして機械学習の手法が加わる。機械学習とは、巨大な規模で行う統計学にほかならない。

この章を貫く問い

三つの大きな問い

1. **それは本物か、それともノイズか**。偶然によってどの程度の変動が予想されるのか。そして、変化を真剣に受け止めるべきなのはどのようなときか。
2. **どのように推定すべきか**。データが与えられたとき、未知の量の最良の推定値は何か。また、異なる方法をどのように比較すればよいか。
3. **ランダムネスはどのようにデータに入り込むのか**。データの集め方、保護の仕方、処理の仕方は、結論にどのような影響を与えるのか。

本章は、新聞の統計を批判的に読むこと（Section 13.1）から始まり、続いて推定へと進む。捕獲再捕獲法（capture–recapture）によるハタネズミの個体数の推定、通し番号からの総数の推定、そして平均（mean）と中央値（median）の比較である。最後に、ランダムネスがどのようにデータに入り込むかについての三つの問題を扱う。プライバシーを守るためのランダム化された調査、蓄積したり打ち消し合ったりする丸め誤差（rounding error）、そして指定された相関（correlation）をもつ変数の構成である。

本章で扱う考え方

§	題材	考え方
13.1	統計を読む	率、自然な変動、傾向
13.2	捕獲再捕獲法	最尤法（maximum likelihood）
13.3	通し番号	偏り、分散、十分性（sufficiency）
13.4	中心の推定	効率（efficiency）と頑健性（robustness）
13.5	ランダム化回答（randomised response）	混合分布（mixture）と全分散（total variance）
13.6	丸め誤差	系統誤差（systematic error）と偶然誤差（random error）
13.7	相関	コレスキー（Cholesky）による構成

Section 13.7 では共分散（covariance）と相関を用いる。これらは 2019 年に STEP の出題範囲から外されたが、この節では必要な事項を説明している。

本章の読み方

すべての推定量は確率変数であることを忘れないでほしい。データが異なる結果になっていれば、推定値も異なっていたはずである。本章の各手法について、多数回の繰り返しのもとでその分布がどうなるかを問うてみよ。平均的に正しいか、そしてどれほどばらつくか。このただ一つの問いこそが、統計的な考え方の核心である。

13.1 見出しを読み解く

問題

STEP I 1992、問 14

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

`step.maths.org/questions/92-s1-q14`

導入

統計学（statistics）とはデータから学ぶための科学であり、その最初の教訓は「数字はそれ自体では何も語らない」ということである。「6 倍安全」「憂慮すべき増加」「惨事が増えている」。見出しは毎日のように数字を主張に変えている。そしてその主張は、しばしば誤っている。数字そのものが偽りだからではない。比較の相手を間違えていたり、ごく普通の偶然による変動を傾向（trend）と取り違えたりしているからである。

この問題には式がまったく含まれておらず、STEP としては珍しい。求められているのは、統計学者のように考えることである。それぞれの見出しを正当化するには何を知る必要があるのか。偶然だけでどの程度の変動が見込まれるのか。答えにはポアソン分布（Poisson distribution）（Section 12.18）を用いる。そしてここでの推論は、この章の残りすべての土台となる。

取りかかる前に

? **考えてみよう** 死亡が一定の割合で、互いに独立かつランダムに起こるとすれば、年間の総数はどの程度ばらつくはずだろうか。平均が年 1080 人であるとき、年ごとの典型的なずれはどれくらいか。

互いに独立な稀な事象の回数は、近似的にポアソン分布に従う。平均 μ のポアソン分布に従う回数の分散（variance）は μ、標準偏差（standard deviation）は $\sqrt{\mu}$ である。$\mu = 1080$ のとき標準偏差はおよそ 33 であり、年間の総数がおよそ 1015 から 1145 の間にあれば、特に目立つことではない。

順を追って解く

段階 1 「歩くより飛行機のほうが 6 倍安全」

> **考えてみよう** 歩行者の死者は年 1080 人、航空機事故の死者は年 180 人である。正確には、何が 6 倍なのか。

この見出しは死者の総数を比べている。しかし安全性とは**リスク**（risk）、すなわち曝露の単位あたりの死者数の問題である。ほとんどの人は毎日歩くが、飛行機に乗る人は比較的少なく、乗るとしてもたまにである。公正な比較には、移動距離 1 キロメートルあたり、移動時間 1 時間あたり、あるいは 1 回の移動あたりの死者数を用いるべきであり、これらはまったく異なる答えを与えうる（空の旅は 1 キロメートルあたりでは非常に安全だが、1 回の移動あたりではそれほど際立って安全ではない。航空機事故の死者の大部分は離陸時と着陸時に生じるからである）。与えられた数値からは、相対的な安全性について何も「証明」できない。また、「道路交通事故で死亡した歩行者」は狭い区分であり、道路での死者全体はこれよりはるかに多い。

段階 2 「空はより安全に：1990 年の死者はわずか 125 人」

> **考えてみよう** 航空機事故による死亡は、互いに独立な稀な事象のようなものだろうか。年間の総数はどの程度ばらつくと考えられるか。

航空機事故による死亡はまとまって起こる。1 回の墜落で数百人が亡くなることもある。そのため年間の総数は主に墜落の回数で決まり、その回数は小さいので、総数は年ごとに極めて大きくばらつく。平均 180 人に対して 125 人という 1 年があっても、何も変化していないこととまったく矛盾しない。墜落がもう 1 回あれば、数字はがらりと変わっていたはずである。1 年分のデータで傾向を示すことはできない。

段階 3 「道路の惨事が増加：火曜日に 7 人死亡」

> **考えてみよう** 平均すると、1 日に何人の歩行者が亡くなっているか。1 日の死者数がポアソン分布に従うなら、1 日に 7 人以上という日はどのくらいの頻度で見られるか。

平均は 1 日あたり $\frac{1080}{365} \approx 2.96$ 人の歩行者の死亡である。1 日の回数がこの平均のポアソン分布に従うなら $P(\geq 7) \approx 0.031$ であり、純粋に偶然だけで、1 年のうちおよそ 11 日は歩行者の死者が 7 人以上となる。（しかも「道路の惨事」にはおそらく運転者や同乗者も含まれるので、1 日に道路で 7 人が亡くなることはなおさら目立つことではない。）1 日だけでは傾向について何もわからない。この見出しは、ばらつきの大きい系

列から極端な値を拾い出しており、典型的な誤りである。

段階 4 「憂慮すべき増加：1990 年の歩行者の死者は 1350 人」

> **考えてみよう** 1350 をポアソン分布の標準偏差と比べてみよ。これは単なる偶然だろうか。そうでないなら、ほかに何が説明になりうるか。そしてそれは「憂慮すべき増加」と言うに値するか。

ここが本題である。一定のポアソンモデルのもとでは、1350 は平均 1080 より 270 大きく、これは標準偏差のおよそ 8 倍にあたる。偶然である可能性は極めて低い。したがって、何かが変化したと考えられる。しかし、この見出しはいくつかの点で依然として言い過ぎである。

- **平均に隠された傾向。**1080 という数値は 16 年間の平均である。もし死者数が 1974 年のおよそ 900 人から 1989 年のおよそ 1260 人まで着実に増えていたとしても、平均はやはり 1080 となり、1990 年の 1350 は既存の傾向がそのまま続いただけであって、突然の「憂慮すべき」増加ではない（図 13.1.1）。年ごとの数値が必要である。
- **曝露。**人口の増加、歩く機会の増加、交通量の増加、あるいは高齢化は、いずれも 1 回の移動あたりで道路が危険になっていなくても死者を増やす。1 人あたり、あるいは歩行距離 1 キロメートルあたりの割合こそが適切な尺度である。
- **定義。**死亡の記録方法や分類方法の変更（たとえば、事故から何日以内の死亡を数えるか）によって、数値が跳ね上がることがある。
- **過分散**（*over-dispersion*）。現実の年間の回数は、天候や経済活動などの条件が変動するため、ポアソンモデルが示すよりも大きくばらつく。「標準偏差の 8 倍」はある程度慎重に扱うべきである。

増加はおそらく本物である。しかしそれが憂慮すべきものかどうか、そして何が原因かを判断するには、さらにデータが必要である。

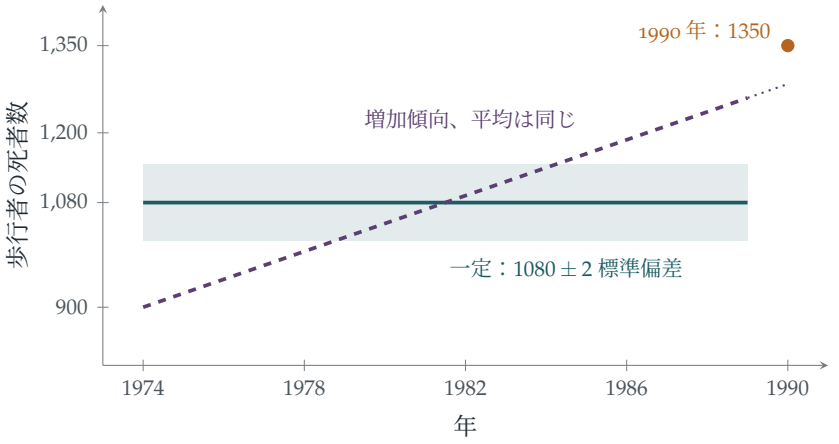


図 13.1.1　16 年間の平均がともに 1080 である、まったく異なる 2 つの推移。死者数が一定であった場合（ティール色、帯はポアソン分布による典型的な変動の範囲）、1990 年の 1350 は劇的な跳ね上がりである。着実に増加していた場合（プラム色）、1350 は傾向をわずかに上回るにすぎない。平均だけではこの 2 つを区別できない。

振り返り

- **同じ条件のものどうしを比べる。**リスクには分母が必要である。1 人あたり、1 キロメートルあたり、あるいは 1 回の移動あたりの死者数である。
- **自然な変動を知る。**ポアソン分布によるばらつき、そしてまとまって起こる事象のはるかに大きなばらつきを知っていれば、どのような変化が注目に値するかがわかる。
- **平均は傾向を隠す。**ただ 1 つの要約値では、ある量が時間とともにどう変化してきたかを記述できない。

背後にある数学

シグナルとノイズ

統計学の問いはすべて、突き詰めれば**シグナル**（signal、本当の効果や傾向）を**ノイズ**（noise、ランダムな変動）から切り分ける問題である。ポアソンモデルは、互いに独立な事象の回数に含まれるノイズを測る物差しを与える。すなわち標準偏差 $\sqrt{\mu}$ である。稀な事象は**相対的な**変動が大きい。9 という回数はおよそ ±3、つまりその大きさの 3 分の 1 ほど変動するが、10 000 という回数の変動はわずか 1% ほどである。小さな

数（1日の死者数、1年の墜落回数）がしばしば過大に解釈されるのも、小さな町でのがんの「集積」がたいてい偶然だと判明するのも、このためである。

平均への回帰

さらにもう1つ落とし穴がある。交通安全カメラは、直前の1年間に異常に事故が多かった場所に設置されることが多い。翌年に事故が減ると、カメラの手柄とされる。しかし異常に悪い年の後には、いずれにせよもっと普通の年が続く可能性が高い。これが**平均への回帰**（regression to the mean）であり、1880年代にゴルトン（Galton）が初めて記述したものである。カメラの効果を平均への回帰から切り分けるには、カメラが設置されなかった同様の場所と比較する必要がある。これが対照実験（controlled experiment）の背後にある考え方である。

ナイチンゲールの統計

クリミア戦争での看護師として名高いナイチンゲール（Florence Nightingale）は、先駆的な統計学者でもあった。1858年の彼女の「鶏頭図（rose diagram）」は、負傷よりも予防可能な病気で亡くなる兵士のほうがはるかに多いことを示し、政府を説得して陸軍の衛生制度を改革させた。彼女は生の総数ではなく割合（兵士1000人あたりの年間死者数）を用いること、そして同じ条件のものどうしを比べることを強く主張した。これはまさにこの問題が問うている点である。彼女は王立統計学会（Royal Statistical Society）のフェローに選ばれた最初の女性であった。

思考のスキル

「何と比べて？」と問う

統計的な主張を受け入れる前に、何と比べているのか、分母は何か、偶然だけでどの程度の変動が見込まれるのかを問え。

大学の統計学の講義では、これらの問いが精密な手法へと形を変える。仮説検定（hypothesis test）は観測結果が偶然と矛盾しないかどうかを問い、信頼区間（confidence interval）はある量がどれほど精密に測定されたかを示し、実験計画（design of experiments）は比較が公正であることを保証する。

次の一歩

1. **ポアソン分布の裾。** X が平均 2.96 のポアソン分布に従うとき、$\mathrm{P}(X \geq 7) \approx 0.031$ であることを確かめよ。歩行者の死者が 9 人以上となる日は、1 年に何日あると期待されるか。

2. **まとまって起こる死亡。** 航空機の墜落が年 1.2 回のポアソン過程（Poisson process）として起こり、1 回の墜落ごとに 150 人が亡くなるとする。航空機事故による年間死者数の平均と標準偏差を求めよ。死者が 125 人以下という年は珍しいか。死者がちょうど 125 人という年はありうるか。

3. **有意な増加か。** 1974–1989 年の数値が 1080 ± 33 で一定であったとすると、ある 1 年に 1350 人以上となる確率を（正規近似（normal approximation）を用いて）求めよ。

4. **よりよい見出しを書く。** 4 つの見出しそれぞれについて、データが実際に裏づける見出しを書け。

さらに学ぶために

- **David Spiegelhalter, *The Art of Statistics***（Pelican）。データから学ぶ方法を、ニュースからの多くの例とともに解説する。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street）。メディアに現れる数字について考えるための 10 のルール。
- **Darrell Huff, *How to Lie with Statistics***（Penguin）。1954 年に書かれた短い古典で、今なお色あせない。

13.2 ハタネズミを数える

問題

STEP II 2000、問 13

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

`step.maths.org/questions/00-s2-q13`

導入

ある島にハタネズミは何匹すんでいるだろうか。すべてを数えることはできない。隠れるし、動き回るし、見た目もよく似ているからである。しかし、何匹かを捕まえて印をつけ、放してから、もう一度何匹かを捕まえることはできる。200 匹に印をつけ、2 回目に捕まえた 200 匹のうち 11 匹に印があったとすれば、個体群のおよそ $\frac{11}{200}$ に印がついていることになるので、個体数は約 $200 \times \frac{200}{11} \approx 3636$ である。

この**標識再捕獲法**（capture–recapture）は、魚、鳥、クジラの個体数の推定に用いられるほか、薬物使用者、ホームレスの人々、戦争の犠牲者の数など、完全には数えきれないあらゆる集団の規模の推定にも使われている。この問題は、直観的な推定値を**最尤法**（maximum likelihood）の原理によって確かな根拠の上に置くものである。最尤法は統計的推定において最も重要な考え方であり、「観測されたデータが最も起こりやすくなるように未知の値を選ぶ」というものである。

取りかかる前に

考えてみよう 個体群には N 匹のハタネズミがいて、そのうち 200 匹に印がついている。無作為に 200 匹の標本を取り出す。ちょうど 11 匹に印がついている確率はいくらか。

印のついた 200 匹から 11 匹を、印のない $N - 200$ 匹から 189 匹を選ぶ。標本は全部で

$\binom{N}{200}$ 通りあり、どれも同様に確からしいので、

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

これが**超幾何分布**（hypergeometric distribution）である。

順を追って解く

段階 1　尤度

N を含む二項係数（binomial coefficients）を書き下すと

$$\binom{N-200}{189} = \frac{(N-200)!}{189!\,(N-389)!}, \qquad \binom{N}{200} = \frac{N!}{200!\,(N-200)!},$$

であるから

$$p_N = \binom{200}{11}\frac{200!}{189!}\cdot\frac{((N-200)!)^2}{N!\,(N-389)!} = k\frac{((N-200)!)^2}{N!\,(N-389)!}. \qquad \blacksquare$$

データを固定し、未知数 N の関数とみなしたとき、p_N を**尤度**（likelihood）という。

段階 2　整数についての最大化

> **考えてみよう**　N は整数なので、微分を使うのは扱いにくい。代わりに p_{N+1} と p_N を比べよう。その比が 1 より大きくなるのはどのようなときか。

階乗のほとんどは約分されて

$$\frac{p_{N+1}}{p_N} = \frac{(N-199)^2}{(N+1)(N-388)}.$$

これが 1 を超えるのは $(N-199)^2 > (N+1)(N-388)$、すなわち $N^2 - 398N + 39601 > N^2 - 387N - 388$、つまり $11N < 39989$、$N < 3635.4$ のときである。したがって p_N は $N = 3636$ まで増加し、その後は減少する。よって最尤推定値（maximum likelihood estimate）は

$$\hat{N} = 3636. \qquad \blacksquare$$

これは $\frac{200\times 200}{11} = 3636.4$ の整数部分であり、直観的な推定値と一致する。

段階 3　3 回目の標本

2 回目の標本を取ったあと、その中の印のない 189 匹にも印をつけるので、印のついたハタネズミは 389 匹になる。$N = 3636$ とすると、印のないものは $3636 - 389 = 3247$

匹である。3 回目に取り出す 200 匹の標本に印のついたものがちょうど j 匹含まれる確率は

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}.$$

$j = 0, 1, \ldots, 200$ についてのこれらの確率の和は 1 でなければならず、これにより求める恒等式が得られる。 ■

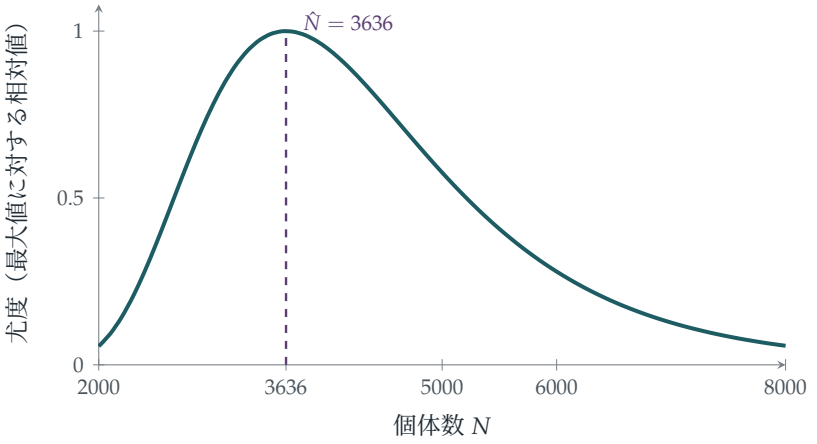


図 13.2.1 2 回目の標本で印のついたハタネズミが 11 匹見つかる尤度 p_N を、個体数 N の関数として、最大値が 1 になるように縮尺したもの。$N = 3636$ で最大となるが、山は幅広く非対称である。およそ 2500 から 6500 までの個体数はいずれもデータとそれなりに整合している。

振り返り

- **未知数の関数としてのデータの確率。**尤度は通常の見方を逆転させる。データを固定し、パラメータを動かすのである。
- **離散的な最大化には比を使う。**隣り合う項を比較することで、整数を変数とする関数の最大値の位置を突き止めた。
- **確率の和は 1 になる。**すべての結果について分布の和をとることで、組合せの恒等式（ヴァンデルモンド（Vandermonde）の恒等式（Vandermonde's identity））を証明した。

背後にある数学

最尤法

最尤法の原理は、1910 年代から 1920 年代にかけてフィッシャー（Ronald Fisher）が発展させたもので、「未知のパラメータを、観測されたデータが最も起こりやすくなる値によって推定せよ」というものである。これは群を抜いて広く使われている推定法であり、データへの曲線のあてはめから、機械学習（machine learning）におけるニューラルネットワーク（neural networks）の訓練にまで及ぶ。後者の「損失関数」（loss functions）は、たいてい対数尤度にマイナスをつけたものである。フィッシャーは、モデルが性質のよいもの（日常的なモデルの大半で成り立つ条件を満たすもの）であれば、標本が大きいとき最尤推定値は近似的に正規分布（normal distribution）に従い、およそ妥当な推定量（estimator）の中で最も精度が高いことを示した。ただし重要な例外もある。次節（Section 13.3）では、未知数がとりうる最大の値、すなわち分布の端点であり、そこでは最尤推定値の振る舞いはまったく異なる。その誤差は近似的に正規分布に従わず、$\frac{1}{\sqrt{n}}$ ではなく $\frac{1}{n}$ のように小さくなっていくのである。

推定はどれほど精確か

図 13.2.1 は尤度の山が幅広いことを示している。11 匹の再捕獲はそれほど多くの情報ではないのである。大まかに計算してみよう。2 回目の標本に含まれる印のついたハタネズミの数は、近似的に平均 $\frac{200\times 200}{N}$ のポアソン分布（Poisson distribution）に従うので、その標準偏差（standard deviation）はおよそ $\sqrt{11} \approx 3.3$ であり、相対誤差は約 30% となる。これはそのまま N の推定値にも引き継がれる。N を 10% 以内の精度で推定するには、生物学者はおよそ 100 匹の再捕獲を必要とするだろう。統計学者はまた、単純な推定値のわずかな偏り（bias）を補正するため、**チャップマン推定量**（Chapman estimator）$\frac{(M+1)(C+1)}{R+1} - 1$ を用いる。ここで M と C は 2 回の捕獲の大きさ、R は再捕獲された数である。

ラプラス、フランスを数える

ラプラス（Laplace）は 1802 年、同じ考え方を用いて、国勢調査を行わずにフランスの人口を推定した。彼は 1 年間に国全体で登録された出生数を知っており、さらにいくつかのコミューンを標本として、その人口と出生数の両方を把握していた。人口と出生数の比がどこでも同じであると仮定して、彼はフランスの人口をおよそ 28 百万人と推定し、さらに不確かさの尺度まで与えた。この方法は、動物の個体数について改めて発見された。1890 年代にデンマークのフィヨルドでツノガレイに標識をつけた C. G. J. Petersen と、1930 年代に鳥の足環を用いた Frederick Lincoln によってである。そのため、今でもしばしば**リンカーン–ピーターセン推定量**（Lincoln–Petersen estimator）と呼ばれている。

思考のスキル

データが最も起こりやすくなるパラメータを選ぶ

未知数を推定するには、観測されたことの確率を未知数の関数として書き、それを最大化せよ。

大学では、最尤法は統計的推論の根幹をなしており、推定値がどれほど精確かを扱う理論（フィッシャー情報量（Fisher information））や、競合するモデルをどう比較するかを扱う理論（尤度比検定（likelihood ratio tests））を伴っている。

次の一歩

1. **一般の公式。** M 匹の動物に印をつけ、2 回目の大きさ C の標本に印のついたものが R 匹含まれていたとする。N の最尤推定値は $\frac{MC}{R}$ の整数部分であることを示せ。
2. **ヴァンデルモンド。** a 人の女性と b 人の男性から c 人の委員会を選ぶ方法を数えることにより、$\sum_j \binom{a}{j}\binom{b}{c-j} = \binom{a+b}{c}$ を証明せよ。
3. **硬貨投げ。** 表の出る確率 p が未知の硬貨を n 回投げたところ、表が h 回出た。p の最尤推定値は $\frac{h}{n}$ であることを示せ。
4. **再捕獲ゼロ。** 2 回目の標本に印のついたハタネズミが 1 匹も含まれていなかった場合、最尤推定値はどうなるか。そのことはこの方法について何を教えてくれるか。

さらに学ぶために

- **David Spiegelhalter, *The Art of Statistics***（Pelican）。推定と不確かさについて。
- **George Seber, *The Estimation of Animal Abundance***（Blackburn）。標識再捕獲法についての標準的な参考書。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。最尤法を含む、数理統計学の入門書。

13.3 ドイツ戦車問題

問題

STEP III 1998、問 14

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

`step.maths.org/questions/98-s3-q14`

導入

第二次世界大戦中、連合国の情報部はドイツが戦車を何両生産しているかを知りたがっていた。スパイの報告によれば、月に約 1400 両であった。統計学者たちは別の方法をとった。鹵獲した戦車のギアボックスなどの部品にはシリアル番号が刻まれており、シリアル番号は事実上 $1, 2, \ldots, N$ からの無作為標本（random sample）とみなせる。番号だけをもとに、彼らは生産数を月に約 250 両と推定した。よく引用される数字では、戦後に調べられたドイツ側の記録は 245 両を示していた。統計学者たちはスパイよりはるかに正確だったのである。

この問題では、シリアル番号から N を推定する二つの方法を比較する。どちらも平均的には正しい。しかし一方は他方よりはるかに効率よくデータを使っており、どちらがそうなのかを調べることで、推定理論の中心的な考え方にたどり着く。すなわち、平均的に正しい推定量のうちでは、ばらつきが最も小さいものを選ぶべきである。

取りかかる前に

考えてみよう M を $[0, N]$ 上の独立な n 個の一様分布に従う値の最大値とするとき、$\mathrm{P}(M \leq m)$ はいくらか。

$M \leq m$ となるのは、すべての $X_i \leq m$ となるときであり、かつそのときに限る。よって $0 \leq m \leq N$ に対して $\mathrm{P}(M \leq m) = \left(\frac{m}{N}\right)^n$ である（Section 12.25 と比較せよ）。これを微分すると、M の確率密度関数（probability density function）は $\frac{nm^{n-1}}{N^n}$ となる。

順を追って解く

段階 1　どちらも不偏である

M について：

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+1}N, \qquad \text{よって} \qquad \mathrm{E}(Z_1) = \frac{n+1}{n}\cdot\frac{n}{n+1}N = N.$$

平均について：各 X_i の期待値は $\frac{N}{2}$ であるから、$\mathrm{E}(Z_2) = \frac{2}{n}\cdot n \cdot \frac{N}{2} = N$ である。 ■

二つの推定量（estimator）はどちらも**不偏**（unbiased）である。すなわち、平均的には正しい答えを与える。係数 $\frac{n+1}{n}$ は、観測された最大のシリアル番号が常に N より少し小さくなることを補正している。

段階 2　それぞれの分散

> **考えてみよう**　$\mathrm{E}(M^2)$ を求め、それを用いて $\mathrm{Var}(M)$ を求めよ。Z_2 については、$[0, N]$ 上の一様分布に従う確率変数一つの分散はいくらか。

$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+2}N^2$ であるから、

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{nN^2}{(n+1)^2(n+2)},$$
$$\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M) = \frac{N^2}{n(n+2)}.$$

$[0, N]$ 上の一様分布（uniform distribution）に従う確率変数の分散（variance）は $\frac{N^2}{12}$ であり、X_i は独立であるから、

$$\mathrm{Var}(Z_2) = \frac{4}{n^2}\cdot n \cdot \frac{N^2}{12} = \frac{N^2}{3n}.$$ ■

段階 3　どちらがよいか

> **考えてみよう**　$\frac{1}{n(n+2)}$ と $\frac{1}{3n}$ を比較せよ。n が大きくなるとき、それぞれはどう振る舞うか。

比は $\frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_2)} = \frac{3}{n+2}$ であり、$n \geq 2$ のとき 1 より小さい（$n = 1$ のときは等しく、このとき二つの推定量は一致する）。したがって Z_1 のほうがよく、n が大きいときにははるかによい。Z_1 の標準偏差（standard deviation）は $\frac{1}{n}$ のように小さくなるが、Z_2 の標準偏差は $\frac{1}{\sqrt{n}}$ のようにしか小さくならない（図 13.3.1）。シリアル番号が $n = 10$ 個あるとき、Z_1 の標準偏差は N の約 9%、Z_2 の標準偏差は約 18% である。

Z_1 を選ぶべき理由はもう一つある。Z_2 はばかげた値になりうるのである。シリアル番号が 10、20、300 のとき、$Z_2 = \frac{2}{3} \cdot 330 = 220$ となり、実際に観測したシリアル番号より小さい。$Z_1 = \frac{4}{3} \cdot 300 = 400$ のほうは、最大の観測値より小さくなることは決してない。 ■

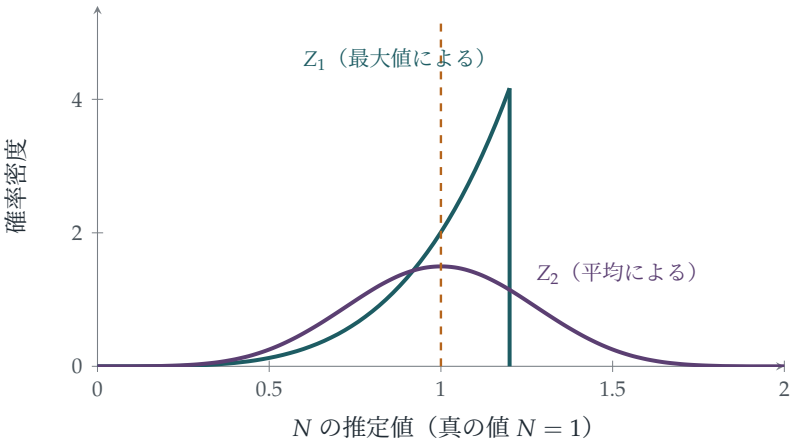


図 13.3.1　シリアル番号が $n = 5$ 個で真の値が $N = 1$ のときの、二つの推定量の標本分布。どちらも平均はちょうど 1 であるが、Z_1（ティール）は Z_2（プラム）よりもはるかに真の値の近くに集中しており、観測された最大のシリアル番号を下回ることは決してない。

振り返り

- **不偏性。**各推定量は、その期待値がちょうど N となるように定数倍されていた。
- **分散を比較する。**不偏推定量のうちでは、分散が小さいほうが信頼できる。
- **最も情報量の多い統計量を使う。**最大値は、標本がもつ N についての情報をすべて含んでいる。平均はその多くを無駄にしてしまう。

背後にある数学

十分性

なぜ最大値はこれほどよいのだろうか。最大値 M が与えられると、残りの値は N が何であっても単に $[0, M]$ 上の一様分布に従う。つまり、それらは N についてそれ以上の情報をもたない。最大値は**十分統計量**（sufficient statistic）なのである。ラオ（C. R. Rao）とブラックウェル（David Blackwell）による定理（1945–47 年）は、どんな不偏推定量も、十分統計量を与えたときの条件付き期待値に置き換えることで改良できる

（少なくとも悪くはならない）ことを述べており、ここではこの操作によって Z_2 が Z_1 に変わる。さらに**レーマン–シェフェの定理**（Lehmann–Scheffé theorem）により、Z_1 が分散の最も小さい不偏推定量であることがわかる。また、Z_1 の分散が $\frac{1}{n^2}$ のように減少し、多くの推定量の $\frac{1}{n}$ より速いことにも注意しよう。これは N が分布の端点であり、データがそこに押し寄せるように集まるからである。

離散版

実際のシリアル番号は整数 $1, 2, \ldots, N$ であり、非復元抽出（sampling without replacement）で得られる。このとき最良の不偏推定値は $\hat{N} = m + \frac{m}{k} - 1$ である。ここで m は観測された k 個のシリアル番号の最大値であり、この式は最大値に観測値どうしの平均的な間隔を加えたものになっている。主にギアボックスと車輪のシリアル番号に基づくこの種の推定を用いて、連合国の統計学者たちはドイツの生産数を月に約 246 両と推定した。これに対し、のちにドイツ側の記録で判明した数は 245 両、情報部の報告が示唆していた数は 1400 両であった。

歴史の中のシリアル番号

この方法はその後も何度も用いられてきた。たとえば、新しい消費者向け製品が何台作られたかを、最初に売られた数台のシリアル番号から分析者が推定した例がある。メーカーもこの教訓を学び、今では多くのメーカーが、単純な連番として読み取れないようにシリアル番号を符号化している。Richard Ruggles と Henry Brodie は、1947 年に *Journal of the American Statistical Association* で戦時中のこの研究について記している。

思考のスキル

推定量は標本分布で評価する

推定量は確率変数である。平均的に正しいか（偏り）、どれだけばらつくか（分散）を問い、誤差の小さいほうを選ぶこと。

大学では、これは効率的な推定の理論へとつながる。クラメール–ラオの下界（Cramér–Rao bound）は不偏推定量の分散がどこまで小さくなりうるかを示し、十分性はその下界にどうすれば到達できるかを説明する。

次の一歩

1. **平均二乗誤差。** 偏りのある推定量 M について、$\mathrm{E}(M-N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M) - N\big)^2$ が成り立つ。これを計算し、$\mathrm{Var}(Z_1)$ と比較せよ。また、$\mathrm{E}(cM-N)^2$ を最小にする定数 c を求めよ。

2. **最小値も。** $(n+1)\min\{X_i\}$ もまた N の不偏推定量であることを示し、その分散を求めよ。これはよい推定量だろうか。

3. **離散版の公式。** シリアル番号 $1,\ldots,N$ から k 個を非復元抽出し、その最大値を m とする。$\mathrm{P}(\max \le m) = \binom{m}{k}/\binom{N}{k}$ を示し、$k=1$ のとき $m + \frac{m}{k} - 1$ が不偏であることを確かめよ。

4. **戦車。** 鹵獲した 5 両の戦車のシリアル番号が 19、40、42、60、98 であるとき、製造された戦車の数を推定せよ。

さらに学ぶために

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**, *Journal of the American Statistical Association* 42 (1947)。当事者による原論文。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。偏り、分散、十分性について。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。

13.4 平均か中央値か？

問題

STEP II 2000、問 14

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

`step.maths.org/questions/00-s2-q14`

導入

測定値の集まりの中心を推定するには、平均（mean）をとることも中央値（median）をとることもできる。どちらがよいのだろうか。答えはデータによって異なる。ときおり大きく外れた誤差を含む測定値では、中央値のほうがはるかに信頼できる。ばかげた値が一つあるだけで平均はどこへでも引きずられてしまうが、中央値はほとんど動かない。性質のよいデータでは、たいてい平均のほうが優れている。この問題は、一様分布に従うデータについて、平均がどれほど優れているかを測るものである。中央値の分散（variance）は平均の分散の 3 倍であり、したがって中央値を平均と同じ精度にするには 3 倍の観測値が必要になる。

この比較を通じて、統計的手法を選ぶうえでの主要な考え方の一つである**効率**（efficiency）と、Section 4.4 にも登場したベータ積分（beta integral）を紹介する。

取りかかる前に

? **考えてみよう** 中央値の密度（density）が $y^n(1-y)^n$ に比例するのはなぜか。（Y が y の近くにあるためには、$2n+1$ 個の値のうちいくつが y より小さく、いくつが y より大きくなければならないか。）

中央値が y にあるためには、値の一つが y にあり、n 個がそれより小さく（それぞれ確率 y）、n 個がそれより大きく（それぞれ確率 $1-y$）なければならない。これにより因子 $y^n(1-y)^n$ が得られる。定数は、どの値がどこに入るかの選び方の数を数えたものである。

順を追って解く

段階 1　定数

密度の積分は 1 でなければならない。

$$k\int_0^1 y^n(1-y)^n\,\mathrm{d}y = k\cdot\frac{n!\,n!}{(2n+1)!} = 1, \qquad k = \frac{(2n+1)!}{(n!)^2}. \qquad \blacksquare$$

段階 2　平均と分散

密度は $\tfrac12$ に関して対称である（y を $1-y$ に置き換えればよい）から、$\mathrm{E}(Y)=\tfrac12$ である。

> **考えてみよう**　ベータ積分を用いて $\mathrm{E}(Y^2) = k\int_0^1 y^{n+2}(1-y)^n\,\mathrm{d}y$ を計算し、階乗を整理せよ。

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

したがって

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac14 = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}. \qquad \blacksquare$$

段階 3　平均との比較

> **考えてみよう**　$2n+1$ 個の独立な一様分布に従う確率変数の平均 $\bar X$ の平均と分散はいくらか。

各 X_i の平均は $\tfrac12$、分散は $\tfrac{1}{12}$ であるから、$\mathrm{E}(\bar X)=\tfrac12$、$\mathrm{Var}(\bar X)=\tfrac{1}{12(2n+1)}$ である。n が大きいとき、

$$\mathrm{Var}(Y)\approx\frac{1}{8n}, \qquad \mathrm{Var}(\bar X)\approx\frac{1}{24n}.$$

> **考えてみよう**　Y と $\bar X$ がともに平均 $\tfrac12$、標準偏差（standard deviation）σ_Y、$\sigma_{\bar X}$ の正規分布に従うとして、両方の確率を標準正規分布の分布関数 Φ を用いて表せ。

$\mathrm{P}\bigl(|Y-\tfrac12|<\tfrac{d}{\sqrt{n}}\bigr) = 2\Phi\bigl(\tfrac{d}{\sqrt{n}\,\sigma_Y}\bigr)-1$ であり、$\sqrt{n}\,\sigma_Y\to\tfrac{1}{\sqrt{8}}$ である。$\bar X$ についても同様で、$\sqrt{n}\,\sigma_{\bar X}\to\tfrac{1}{\sqrt{24}}$ となる。よって二つの確率はそれぞれ $2\Phi(d\sqrt{8})-1$ と $2\Phi(d\sqrt{24})-1$ に近づく。Φ は増加関数であり $\sqrt{24}>\sqrt{8}$ であるから、後者のほうが大きく、したがっ

て十分大きいすべての n について不等式が成り立つ。 ■平均のほうが真の中心に近い値をとりやすいのである。

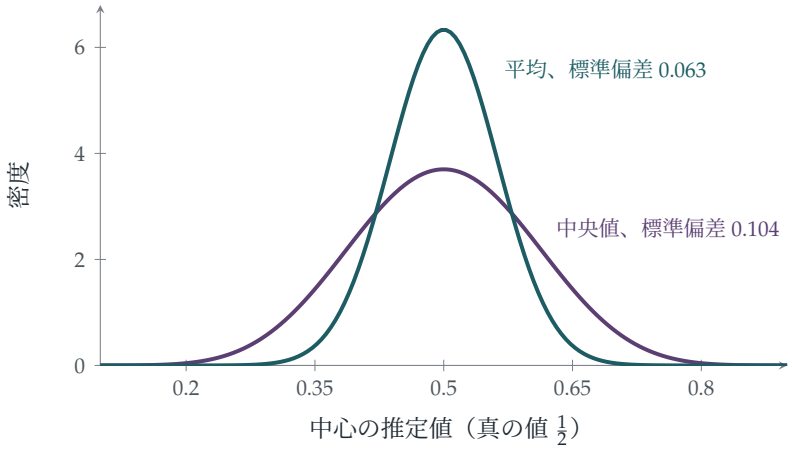


図 13.4.1 21 個の独立な一様分布の観測値の中央値（紫）と平均（青緑）の標本分布。どちらも真の値 $\frac{1}{2}$ を中心としているが、平均のほうがより狭く集中している。その分散は中央値の分散の約 3 分の 1 である。

振り返り

- **ベータ積分。** $\int_0^1 y^r(1-y)^s\,\mathrm{d}y$ という一つの公式から、正規化定数とモーメントの両方が得られた。
- **対称性から平均が得られた。** $\mathrm{E}(Y)$ を求めるのに積分は必要なかった。
- **ばらつきは適切な尺度で比較する。** 標準偏差に $\sqrt{n}$ を掛けることで、比較が n によらないものになった。

背後にある数学

効率

二つの不偏推定量（unbiased estimator）の**相対効率**（relative efficiency）とは、それらの分散の比である。これは、一方が他方に匹敵するには何倍の観測値が必要かを表す。一様分布のデータでは、中央値の平均に対する効率は $\frac{1}{3}$ である。正規分布のデータでは $\frac{2}{\pi} \approx 0.64$ であり、中央値には約 57% 多くの観測値が必要になる。しかし裾の重いデータでは立場が逆転する。コーシー分布（Cauchy distribution、Section 12.22）では平均は役に立たない。n 個の観測値の平均は観測値一つと比べて少しもよくならな

いからである。一方、中央値はうまく機能する。ラプラス分布（Laplace distribution、Section 12.24）では、中央値の効率は平均の 2 倍である。

頑健統計

中央値は**頑健**（robust）である。データのほぼ半分を任意の値に変えても、中央値は大きくは動かない。平均にはそのような防御がなく、誤った値が一つあるだけで台なしになりうる。1960 年代から 1970 年代にかけて、ジョン・テューキー（John Tukey）、Peter Huber らは**頑健統計**（robust statistics）を発展させた。これは、きれいなデータに対しては平均とほぼ同じ効率をもちながら、外れ値（outlier）には強い推定量であり、たとえば**刈り込み平均**（trimmed mean、最大と最小の 10% を捨てて残りを平均したもの）がある。これは飛び込みや体操などの競技の採点に用いられている。

ラプラスと中央値

ラプラス（Laplace）は 1774 年に、誤差が二重指数分布（ラプラス分布）に従うならば、真の値の最良の推定値は中央値であることを示した。ガウス（Gauss）は 1809 年に、誤差が正規分布に従うならば最良の推定値は平均であることを示し、これが最小二乗法（method of least squares）へとつながった。どちらが正しいかは誤差のふるまいによって決まり、これは新しい種類のデータが現れるたびに統計学者がいまなお議論している問題である。

思考のスキル

手法はばらつきで比較する

二つの手法がどちらも平均的には正しいことがある。そのときは、答えがどれだけばらつくか、そしてデータが増えるにつれてそのばらつきがどれだけ速く小さくなるかを比較せよ。

大学では、大標本における分散の比較である漸近効率（asymptotic efficiency）が、モデルが誤っているかもしれない場合のための頑健性とともに、統計的手法を選ぶ際の標準的な道具となっている。

次の一歩

1 **中央値の密度。** 値が一つ $[y, y+\delta y]$ にあり、n 個がそれより小さく、n 個がそれより大きいという数え上げによって、密度 $\frac{(2n+1)!}{(n!)^2} y^n (1-y)^n$ を導け。

2 **三つの観測値。** $n = 1$（一様分布の観測値が三つ）のとき、$\mathrm{Var}(Y)$ と $\mathrm{Var}(\bar{X})$ を正確に求めよ。どちらが小さいか。

3 **ミッドレンジ。** 中心のもう一つの推定量に、ミッドレンジ（midrange）$\frac{1}{2}(\max + \min)$ がある。一様分布のデータで観測値が $m = 2n+1$ 個のとき、その分散が $\frac{1}{2(m+1)(m+2)}$ であることを示せ。平均と比べてどうか。（Section 13.3 と比較せよ。）

4 **外れ値。** 11 個の測定値が 9.8、9.9、10.0、10.0、10.1、10.1、10.2、10.2、10.3、10.4、104（入力ミス）である。平均と中央値を比較せよ。

さらに学ぶために

- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。順序統計量と効率について。
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods***（Springer）。頑健な手法がなぜ重要なのかを、専門家でない読者向けに解説している。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***（Harvard University Press）。集約、平均、中央値を歴史的な視点から論じている。

13.5 答えにくい質問をする

問題

STEP II 2023、問 11

先に進む前に、まず問題を読んでおこう。問題は、OCR が配布している 2023 年の STEP 過去問一式（zip ファイル）のうち、STEP II の問題用紙にある：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

導入

試験でカンニングをした学生や、職場で盗みをはたらいた従業員がどれくらいの割合でいるかを知るにはどうすればよいだろうか。直接たずねれば、多くの人は嘘をつくだろう。1965 年、統計学者の Stanley Warner は巧妙な解決策を考案した。それが**無作為化回答法**（randomised response）である。各人はこっそりと硬貨を投げるかさいころを振り、その結果によって、デリケートな質問に正直に答えるか、その反対の質問に答えるかが決まる。質問者にはどちらなのかがわからないので、個々の回答から罪が明らかになることはない。それでも、無作為化の確率がわかっているので、全体としての割合は推定できるのである。

この問題では、その背後にある数学を探る。各人を**独立に**（independently）無作為化すると（小問 (ii)）、合計は扱いやすい振る舞いをし、割合をきちんと定めることができる。一方、全員を**ただ 1 回の**硬貨投げで無作為化すると（小問 (iii)）、合計の平均（mean）は同じでも、まったく異なる二山の分布になり、ほとんど何も教えてくれない。どこに無作為性を置くかが重要なのである。

取りかかる前に

考えてみよう Y は確率 p で X_1、確率 q で X_2 である。$\mathrm{E}(Y)$ は何か。$\mathrm{E}(Y^2)$ は何か。

どちらの確率変数が選ばれたかで条件付ける（conditioning）と、$\mathrm{E}(Y) = p\mu_1 + q\mu_2$ であり、$\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2)$ である。

順を追って解く

段階 1 混合分布の分散

$$\operatorname{Var}(Y) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2) - (p\mu_1 + q\mu_2)^2.$$

$p - p^2 = pq$ と $q - q^2 = pq$ を用いると、μ を含む項は $p\mu_1^2 + q\mu_2^2 - p^2\mu_1^2 - 2pq\mu_1\mu_2 - q^2\mu_2^2 = pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2$ となる。したがって

$$\operatorname{Var}(Y) = p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2. \qquad \blacksquare$$

混合分布（mixture）の分散（variance）は、各成分内の分散の平均に、成分の平均どうしの**間の**分散を加えたものである。

段階 2 独立な無作為化

> **考えてみよう** $\mathrm{P}(B = 1)$ は何か。n 個の B の値は独立か。

$\mathrm{P}(B = 1) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{5}{6} = \frac{1}{2}$ である。n 個の値は独立で、それぞれ成功確率は $\frac{1}{2}$ なので、$Z_1 \sim \mathrm{B}(n, \frac{1}{2})$ であり、平均は $\frac{n}{2}$、標準偏差（standard deviation）は $\frac{\sqrt{n}}{2}$ である。

平均から 10% 以内であるとは、$|Z_1 - \frac{n}{2}| < \frac{n}{20}$ ということである。標準化すると、これは $|Z| < \frac{n/20}{\sqrt{n}/2} = \frac{\sqrt{n}}{10}$ となり、右辺は無限大に発散するので、確率は 1 に収束する。 $\blacksquare$

段階 3 全員に 1 枚の硬貨

> **考えてみよう** 今度は 1 回の硬貨投げが n 個すべてのさいころについて決める。表が出たとき、Z_2 の分布は何か。裏が出たときはどうか。小問 (i) を用いよ。

表が出たとき、$Z_2 \sim \mathrm{B}(n, \frac{1}{6})$ であり、平均は $\frac{n}{6}$、分散は $\frac{5n}{36}$ である。裏が出たとき、$Z_2 \sim \mathrm{B}(n, \frac{5}{6})$ であり、平均は $\frac{5n}{6}$、分散は $\frac{5n}{36}$ である。$p = q = \frac{1}{2}$ として小問 (i) を用いると

$$\mathrm{E}(Z_2) = \frac{n}{2}, \qquad \operatorname{Var}(Z_2) = \frac{5n}{36} + \frac{1}{4}\left(\frac{4n}{6}\right)^2 = \frac{5n}{36} + \frac{n^2}{9}.$$

> **考えてみよう** 確率の大部分はどこにあるか。平均 $\frac{n}{2}$ の近くに多くあるか。

Z_2 は $\frac{n}{6}$ 付近と $\frac{5n}{6}$ 付近に二つの山をもち、それぞれの幅はおよそ $\sqrt{n}$ である（図 13.5.1）。その間にはほとんど何もない。同じ平均と分散をもつ正規分布（normal distribution）は $\frac{n}{2}$ を中心とする一つの山になるが、そこはまさに Z_2 がほとんど値をとらない場所

である。Z_2 が $\frac{n}{2}$ から 10% 以内にあるためには、$\mathrm{B}(n, \frac{1}{6})$ または $\mathrm{B}(n, \frac{5}{6})$ に従う確率変数が $[0.45n, 0.55n]$ に入らなければならない。$\mathrm{B}(n, \frac{1}{6})$ の場合、これは平均より少なくとも $0.28n$ 大きいということであり、標準偏差にしておよそ $\frac{0.28n}{0.37\sqrt{n}} \approx 0.75\sqrt{n}$ 個分にあたるので、確率は 0 に収束する。$\mathrm{B}(n, \frac{5}{6})$ についても同様である。 ■

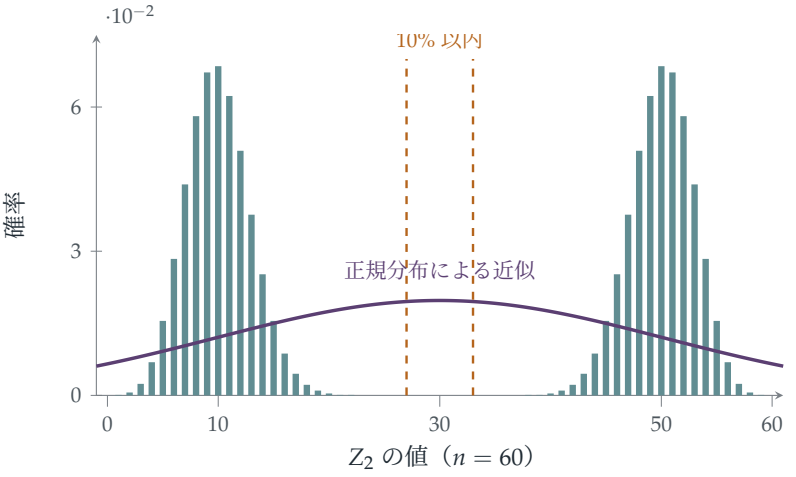


図 13.5.1　$n = 60$ 個のさいころに対する Z_2 の分布（棒）。1 回の硬貨投げの二つの結果に対応して、10 と 50 に二つの山がある。同じ平均（30）と分散をもつ正規分布（紫の曲線）はよく当てはまらず、平均から 10% 以内（琥珀色の線の間）にある確率はほとんどない。

振り返り

- **隠れた選択で条件付ける。**混合分布の平均と分散は、どの成分が選ばれたかで条件付けることで得られた。
- **成分内の分散と成分間の分散の和。**余分な項 $pq(\mu_1 - \mu_2)^2$ は、二つの成分の間の散らばりを測るものであった。
- **平均と分散がすべてではない。**Z_2 は Z_1 と同じ平均をもつが、形はまったく異なっていた。

背後にある数学

無作為化回答法

Warner の方式では、各回答者は確率 p で「あなたはグループ A に属していますか？」に答え、確率 $1 - p$ で「あなたはグループ A に属していないのですか？」に答える。ど

ちらに答えるかは本人だけが知っている。A に属する人の割合を π とすると、「はい」と答える確率は

$$\lambda = p\pi + (1-p)(1-\pi),$$

であるから、観測された「はい」の割合 $\hat{\lambda}$ から、$\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$ と推定できる（ただし $p \neq \frac{1}{2}$ とする）。問題の小問 (ii) はまさにこれであり、各人が自分のさいころで決めている。プライバシーの代償は精度である。$2p-1$ で割ることで分散が大きくなるため、無作為化した調査では直接たずねる調査よりも多くの回答者が必要になる。しかし各人が独立に無作為化するので、大数の法則（law of large numbers）がはたらく。小問 (iii) は、無作為化が共有されると何がまずいのかを示している。ただ 1 枚の硬貨についての不確かさは、決して平均化されて消えることがないのである。

差分プライバシー

無作為化回答法は、**差分プライバシー**（differential privacy）の祖先にあたる。差分プライバシーとは、2006 年に Cynthia Dwork らが導入したプライバシーの数学的な定義であり、ある手続きがプライベートであるとは、任意の 1 人のデータを加えたり取り除いたりしても、どの出力の確率もたかだか小さな倍率しか変わらないことをいう。テクノロジー企業は利用統計の収集にこれを用いており、アメリカ合衆国国勢調査局は 2020 年の国勢調査でこれを用いた。公表する数値に注意深く調整したランダムなノイズを加え、どの個人の回答も復元できないようにしたのである。

全分散の法則

小問 (i) の恒等式は、**全分散の法則**（law of total variance）$\operatorname{Var}(Y) = \operatorname{E}\bigl(\operatorname{Var}(Y \mid C)\bigr) + \operatorname{Var}\bigl(\operatorname{E}(Y \mid C)\bigr)$ の特別な場合である。ここで C は隠れた選択である。この法則は**分散分析**（analysis of variance、ANOVA）の基礎になっている。分散分析は、1920 年代にフィッシャー（Ronald Fisher）がロザムステッド（Rothamsted）での農業実験のために導入したもので、作物の収量のばらつきを、肥料や品種によって説明される部分と、説明されずに残る部分とに分解する。

思考のスキル

要約だけでなく、分布全体を見よ

平均と分散は山が一つの分布をよく記述するが、混合分布に対してはひどく誤解を招くことがある。確率が実際にどこにあるのかを問え。

大学では、いくつかのグループからなる母集団を記述するのに混合モデル（mixture models）が用いられ、全分散の法則は、調査設計から階層モデルに至るまで、統計学の標準的な道具である。

次の一歩

1 **双子。** STEP II 2005、問 12（`step.maths.org/questions/05-s2-q12`）に挑戦せよ。これは姿を変えた無作為化回答法のパズルである。回答が硬貨について何も教えてくれないのはどのようなときか。

2 **π の推定。** $p = \frac{3}{4}$ の Warner 方式の調査で、400 人が回答し、160 人が「はい」と答えた。グループ A に属する人の割合 π を推定し、その推定値のおおよその標準誤差（standard error）を求めよ。

3 **全分散。** いくつかの値をとる離散的な隠れた変数 C について、全分散の法則を証明せよ。

4 **共有された硬貨。** 小問 (iii) において、$Z_2 > \frac{n}{2}$ となる確率は何か。Z_2 が $\frac{n}{2}$ に近い値をとることはほとんどないにもかかわらず、n が大きいときこの確率が $\frac{1}{2}$ に近い理由を説明せよ。

さらに学ぶために

- **Stanley Warner, 「Randomized response: a survey technique for eliminating evasive answer bias」**, *Journal of the American Statistical Association* 60 (1965).
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy***（オンラインで無料公開）。現代の理論であり、最初の数章は読みやすい。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。調査とその落とし穴について。

13.6 積み重なる誤差

問題

STEP II 1993、問 15

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

`step.maths.org/questions/93-s2-q15`

導入

1982 年、バンクーバー証券取引所は 1000 から始まる新しい株価指数を導入した。株価が変化するたびに指数は再計算され、小数第 3 位までで**切り捨て**（truncation）られた。1 回の切り捨てで失われる量は平均 0.0005 であり、無視できるほど小さく見える。しかし再計算は 1 日に約 3000 回行われ、22 か月後には、市場全体は上昇していたにもかかわらず、指数は約 520 になっていた。正しく計算し直したところ、指数は一晩で約 1100 に跳ね上がった。

この問題は、何が起きたのかを説明するものである。切り捨て誤差はすべて同じ向きを持つので、演算回数に比例して着実に積み重なっていく。一方、四捨五入（rounding）による誤差は両方の向きを持ち、大部分が打ち消し合う。その増え方は演算回数の平方根程度にすぎず、これは中心極限定理（central limit theorem、Section 12.24）が予言するランダムウォーク（random walk）の振る舞いである。

取りかかる前に

? **考えてみよう** ある数の小数部分（fractional part）U が $[0,1)$ 上の一様分布（uniform distribution）に従うとき、その数を切り捨てたときの誤差はどうなるか。四捨五入したときはどうか。それぞれの平均（mean）と分散（variance）は何か。

切り捨ては a を $\lfloor a \rfloor$ で置き換えるので、誤差（真の値から近似値を引いたもの）は U であり、$[0,1)$ 上の一様分布に従う。平均は $\frac{1}{2}$、分散は $\frac{1}{12}$ である。四捨五入の誤差は

$(-\frac{1}{2}, \frac{1}{2}]$ 上の一様分布に従い、平均は 0、分散は $\frac{1}{12}$ である。

順を追って解く

段階 1　誤差の総和

1500 個の数を足したあとの誤差の総和は、1500 個の独立な誤差の和であり、中心極限定理により近似的に正規分布（normal distribution）に従う。

- LEP（四捨五入）：平均 0、分散 $\frac{1500}{12} = 125$、標準偏差（standard deviation）$\sqrt{125} \approx 11.18$。
- VOZ（切り捨て）：平均 $\frac{1500}{2} = 750$、標準偏差は同じく 11.18。

段階 2　大きな誤差が生じる確率

> **考えてみよう**　LEP について、$E \sim \mathrm{N}(0, 125)$ のとき $\mathrm{P}(|E| > 15)$ はいくらか。VOZ について、誤差はほぼ確実にどのあたりにあるか。

LEP については $\frac{15}{11.18} \approx 1.342$ であるから、

$$\mathrm{P}(|E| > 15) \approx 2\bigl(1 - \Phi(1.342)\bigr) \approx 2 \times 0.090 = 0.18.$$

VOZ については、誤差はおよそ 750 ± 11 である。その絶対値が 15 未満となる確率は実質的に 0 なので、$\mathrm{P}(|E| > 15) \approx 1$ である。 ■

段階 3　何回足せるか

> **考えてみよう**　n 回の加算のあと、LEP の誤差の標準偏差は $\sqrt{n/12}$ である。$\mathrm{P}(|E| < 10) \geq 0.9$ となるのはどのようなときか。

$\mathrm{P}(|Z| < 1.645) = 0.9$ であるから、$\frac{10}{\sqrt{n/12}} \geq 1.645$ が必要である。すなわち $\frac{n}{12} \leq \left(\frac{10}{1.645}\right)^2 \approx 36.95$ であり、$n \leq 443$ となる。LEP は約 443 回の加算を行うことができる。 ■

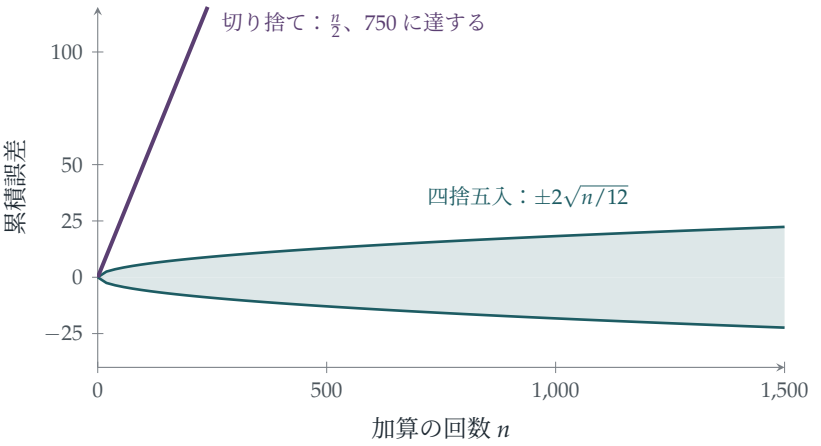


図 13.6.1　誤差の蓄積の様子。切り捨て誤差（紫）はすべて同じ符号を持ち、n に比例して増える。1500 回の加算のあと、総和は約 750 になる。四捨五入の誤差（青緑の帯、0 の両側に標準偏差 2 つ分）は平均的には打ち消し合い、$\sqrt{n}$ 程度にしか増えない。1500 回の加算のあとでも、通常はおよそ ±22 の範囲に収まる。

振り返り

- **個々の誤差をモデル化する。**小数部分が一様分布に従うことから、各誤差は簡単な分布を持つことになった。
- **平均と分散を足し合わせる。**平均は足し合わされ（偏り $\frac{n}{2}$ が生じる）、誤差が独立ならば分散も足し合わされる。
- **正規近似。**中心極限定理によって、1500 個の一様分布に従う誤差の和は正規分布に置き換えられた。

背後にある数学

偏りと分散

切り捨ては**偏りがある**（biased）。すなわち、その誤差の平均は 0 でなく、偏りは線形に蓄積する。四捨五入は**不偏**（unbiased）であり、その誤差はランダムウォークのように振る舞って $\sqrt{n}$ 程度で増える。100 万回の演算のあとでは、その差は 500 000 程度の誤差と 300 程度の誤差の違いになる。これは、実験における系統誤差（systematic error、たとえば校正の狂った測定器）と偶然誤差（random error）の区別と同じである。多数の測定値を平均すれば偶然誤差は小さくなるが、系統誤差にはまったく効果

がない。

浮動小数点

あらゆるコンピュータは実数を有限の精度で保存し、あらゆる算術演算はその結果を丸める。現代のコンピュータは IEEE 754 規格に従っており、表現可能な最も近い数に丸める。ちょうど中間の場合には最後の桁が偶数になる方へ丸めるが、これはまさに丸めが平均的に不偏となるようにするためである。それでも、多数の数を不用意に足し合わせると精度が失われることがある。カハン（William Kahan）の**補償付き総和**（compensated summation）アルゴリズム（1965 年）は、失われた下位の桁を記録しておき、それを足し戻す。1991 年の湾岸戦争では、パトリオットミサイルの砲台が飛来するミサイルの迎撃に失敗した。その時計は時間を 0.1 秒単位で数えていたが、0.1 は 2 進法では正確に表せない数である。100 時間の稼働のあと、切り捨て誤差は約 3 分の 1 秒にまで増え、標的の位置を 0.5 キロメートル以上見誤るのに十分な大きさになっていた。

バンクーバー指数

バンクーバー証券取引所の指数は、1982 年 1 月に 1000.000 という値で創設された。1983 年 11 月までに、市場全体はおおむね上昇していたにもかかわらず、指数は約 524 にまで下がっていた。コンサルタントたちは、指数が 1 日に何千回も再計算され、そのたびに小数第 3 位まで（四捨五入ではなく）切り捨てられていたことを突き止めた。週末をかけて正しく計算し直した結果、指数は約 1098 で取引を再開した。この出来事は、数値計算の講義で定番の戒めとなっている。

思考のスキル

系統誤差と偶然誤差を区別する

符号が一定の誤差は線形に蓄積する。平均 0 の独立な誤差は平方根のように蓄積する。自分が扱っている誤差がどちらの種類なのかを見極めよ。

大学では、数値解析（numerical analysis）で誤差が計算の中をどのように伝わるかを学び、統計学では偏りと分散を区別する。「バイアス–バリアンスのトレードオフ」（bias–variance trade-off）は、現代の機械学習の中心にある考え方である。

次の一歩

1. **VOZ の補正。** VOZ が最後に答えに $\frac{n}{2}$ を加えるとすると、その誤差はどのような分布に従うか。これで LEP と同じくらい良くなったといえるか。

2. **小数第 1 位への四捨五入。** 数を小数第 1 位に四捨五入し、捨てられる桁が一様分布に従うとき、10 000 個の数を足したときの誤差の総和の標準偏差を求めよ。

3. **バンクーバー指数。** 指数が 1 日に 3000 回再計算され、1 回の切り捨てで平均 0.0005 が失われたとすると、取引日 250 日の 1 年間でどれだけ失われるか。

4. **厳密な分布。** 2 つの四捨五入誤差の和について、厳密な確率密度関数（三角形になる）を求めよ。3 つの場合については、その概形を描け。どのくらいの速さで正規分布らしく見えてくるか。

さらに学ぶために

- **Nicholas Higham, *Accuracy and Stability of Numerical Algorithms***（SIAM）。丸め誤差に関する定評ある本であり、第 1 章は読みやすい。

- **David Goldberg, “What every computer scientist should know about floating-point arithmetic”**, *ACM Computing Surveys* 23 (1991)。古典的な論文で、オンラインで無料で読める。

- **Tim Harford, *How to Make the World Add Up***（Bridge Street）。データにおける系統誤差を扱っている。

13.7 相関をつくり出す

問題

STEP III 2010、問 13

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

`step.maths.org/questions/10-s3-q13`

導入

ある銀行が、株式のポートフォリオ（portfolio）で損失を出す可能性がどの程度あるかを知りたいとする。株価は連動して動く。ある株が下がれば、ほかの株も下がる傾向がある。起こりうる将来を何千通りもシミュレーションするには、銀行はちょうど正しい形で相関した乱数を生成しなければならない。独立（independent）な確率変数（random variable）から出発して、指定された相関（correlation）をもつ確率変数をどうやって作ればよいだろうか。

この問題は、3 つの変数についてその作り方を一歩ずつ示している。新しい変数はそれぞれ、それまでの独立な材料の組合せに、分散（variance）を補うための新たな独立な材料を 1 つ加えたものである。この構成は、姿を変えた線形代数の有名な道具である**コレスキー分解**（Cholesky decomposition）であり、また隠れた制約も明らかにする。すなわち、どんな相関の組でも実現できるわけではないのである。

出題範囲についての注意

共分散と相関は、2019 年に STEP の出題範囲から外された。共分散（covariance）は $\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$ であり、独立な変数に対しては 0 となる。また各引数について線形である：$\mathrm{Cov}(aU + bV, W) = a\,\mathrm{Cov}(U,W) + b\,\mathrm{Cov}(V,W)$。さらに $\mathrm{Var}(U) = \mathrm{Cov}(U,U)$ が成り立つ。相関は常に -1 以上 1 以下である。

取りかかる前に

考えてみよう　Z_1 と Z_2 は独立である。その共分散はいくらか。したがって相関はいくらか。

独立な変数は $\mathrm{E}(Z_1Z_2) = \mathrm{E}(Z_1)\,\mathrm{E}(Z_2)$ を満たす。よって共分散は 0 であり、$\mathrm{Corr}(Z_1, Z_2) = 0$ である。

順を追って解く

段階 1　2 つの変数

$\mathrm{E}(Y_2) = 0$ である。Z_1 と Z_2 は独立で分散が 1 なので、

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + (1 - \rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + \sqrt{1 - \rho_{12}^2}\, Z_2) = \rho_{12}.$$

分散はどちらも 1 なので、$\mathrm{Corr}(Y_1, Y_2) = \rho_{12}$ である。 ■

考えてみよう　Z_1 と Z_2 を直交する単位ベクトルとみなそう。Y_2 はどこにあるか。Y_1 と Y_2 のなす角の余弦はいくらか。

$Y_2 = \rho_{12}Z_1 + \sqrt{1 - \rho_{12}^2}\, Z_2$ は、$Z_1 = Y_1$ と角 θ をなす単位ベクトルである。ただし $\cos\theta = \rho_{12}$ である。相関とは、変数をベクトルとみなしたときの、それらのなす角の余弦なのである（図 13.7.1）。

段階 2　3 つ目の変数

考えてみよう　3 つの条件 $\mathrm{Var}(Y_3) = 1, \mathrm{Cov}(Y_1, Y_3) = \rho_{13}, \mathrm{Cov}(Y_2, Y_3) = \rho_{23}$ を a, b, c を用いて書き表せ。

線形性と独立性を用いると、

$$a^2 + b^2 + c^2 = 1, \qquad a = \rho_{13}, \qquad \rho_{12}a + \sqrt{1 - \rho_{12}^2}\, b = \rho_{23}.$$

よって

$$a = \rho_{13}, \qquad b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1 - \rho_{12}^2}}, \qquad c = \sqrt{1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1 - \rho_{12}^2}}.$$ ■

（$c \geq 0$ なので、負でない平方根をとる。）

段階 3 任意の平均と分散

定数倍と平行移動は相関を変えないので、

$$X_i = \mu_i + \sigma_i Y_i \qquad (i = 1, 2, 3)$$

は平均 μ_i、分散 σ_i^2、相関 ρ_{ij} をもつ。 ■

> **手を止めて、やってみよう** c の式では、根号の中の量が負でないことが必要である。$\rho_{12} = \rho_{13} = \rho_{23} = -0.6$ を試してみよう。何がうまくいかないだろうか。また、3 つの変数が互いにどれも強い負の相関をもつことはできない、というのはなぜ納得のいくことなのだろうか。

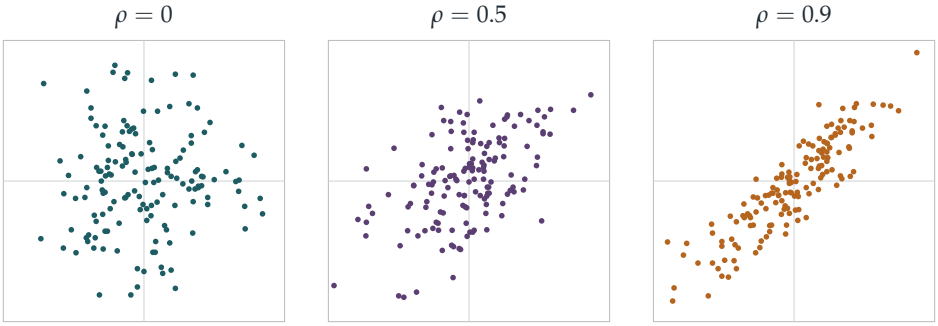


図 13.7.1 独立な標準正規分布（standard normal distribution）に従う Z_1, Z_2 から、$Y_1 = Z_1, Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$ という作り方で得た (Y_1, Y_2) の標本。$\rho = 0, 0.5, 0.9$ の場合である。同じ独立な材料から、好きな相関を作り出すことができる。

振り返り

- **変数を 1 つずつ積み上げる。**新しい変数はそれぞれ、それまでの材料をちょうどよい量だけ取り込み、新たな独立な材料を 1 つ加えたものであった。
- **共分散は線形である。**すべての条件が、係数についての簡単な方程式になった。
- **相関は余弦である。**標準化した変数を単位ベクトルとみなすと、相関は角度になる。

背後にある数学

コレスキー分解

行列の形で書くと、この構成は $\mathbf{Y} = L\mathbf{Z}$ と表される。ここで L は**下三角**（lower triangular）行列

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

であり、$\mathbf{Y}$ の相関行列（correlation matrix）は $R = LL^{\mathsf{T}}$ となる。対称行列を LL^{T} の形に書くことを**コレスキー分解**という。これはフランスの軍の技術者アンドレ＝ルイ・コレスキー（André-Louis Cholesky）が 1910 年ごろ、測地測量の方程式を解くために見いだしたものである（第一次世界大戦で本人が戦死した後に発表された）。この分解は R が**正定値**（positive definite）であればいつでも可能であり、その計算は私たちが手で行ったこととまったく同じである。すなわち、各行を順に解いていくのである。これは、金融、工学、気候モデリングにおいて相関のある変数をシミュレーションする標準的な方法である。

どのような相関が可能か

c の根号の中の数は負であってはならない。展開すると、その条件は

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

となり、これは相関行列の行列式（determinant）が負でないという条件である。幾何学的にいえば、2 つずつのなす角が θ_{ij}（ただし $\cos\theta_{ij} = \rho_{ij}$）である 3 つの単位ベクトルが、実際に空間の中で両立しなければならないということである。もし $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$ ならば、角はすべて約 127° であり、互いに 127° をなす 3 つのベクトルには、和が 360° を超える角が必要になってしまう。これは不可能である。一般に、ある行列が相関行列となりうるのは、それが半正定値（positive semi-definite）、すなわちすべての固有値（eigenvalue）が負でない場合に限られる（Section 7.2）。相関を不完全なデータから個別に推定すると、この条件が成り立たないことがあり、そのとき統計学者は最も近い有効な相関行列を求めなければならない。この問題は 2002 年に Nicholas Higham によって解決された。

ゴルトン、ピアソンと相関

フランシス・ゴルトン（Francis Galton）は 1888 年、親と子の身長がどのように関係しているかを研究する中で相関を導入し、1896 年にはカール・ピアソン（Karl Pearson）がこの問題で用いられている積率（product moment）の公式を与えた。ゴルトンは平均への回帰（regression to the mean）にも気づいていた（Section 13.1）。ピアソン

は1911年、ユニバーシティ・カレッジ・ロンドン（University College London）に世界初の大学の統計学科を創設した。「相関は因果を意味しない」という警句は、相関そのものとほとんど同じくらい古い。

思考のスキル

新たな独立な材料を使って、一歩ずつ構成する

指定された関係をもつ対象を作るには、それらを *1* つずつ加えていき、それぞれを、それまでに作ったものの組合せに新しく独立なものを加えたものとせよ。

これは線形代数のグラム–シュミットの直交化法（Gram–Schmidt process）を逆向きにたどったものであり、またコレスキー分解でもある。大学では、最小二乗法（least squares）、多変量正規分布（multivariate normal distribution）の理論、確率過程のシミュレーションなどに現れる。

次の一歩

1 **不可能な相関。** 3つの確率変数が2つずつすべて相関 ρ をもつことは、$\rho \geq -\frac{1}{2}$ でない限り不可能であることを示せ。n 個の変数の場合、対応する主張はどうなるか。

2 **行列式を確かめる。** R を 3×3 の相関行列とするとき、$1-\rho_{13}^2-b^2\geq 0$ が $\det R\geq 0$ と同値であることを示せ。

3 **ポートフォリオのリスク。** 2つの株の収益率の標準偏差（standard deviation）が $\sigma_1=0.2$ と $\sigma_2=0.3$ で、相関が ρ であるとする。それぞれに半分ずつ投資したポートフォリオの標準偏差を、$\rho=-1,0,1$ のそれぞれについて求めよ。分散投資（diversification）によって何が得られるか。

4 **材料を復元する。** 問題文のように Y_1, Y_2 が与えられたとき、Z_2 を Y_1 と Y_2 を用いて表せ。（これがグラム–シュミットの1ステップである。）

さらに学ぶために

- **Gilbert Strang, *Introduction to Linear Algebra***（Wellesley-Cambridge）。正定値行列と

コレスキー分解について。

- **Paul Glasserman, *Monte Carlo Methods in Financial Engineering***（Springer）。相関のある変数を実際にシミュレーションする方法。高度だが、最初のほうの章は読みやすい。
- **Theodore Porter, *Karl Pearson: The Scientific Life in a Statistical Age***（Princeton University Press）。

学んだこと

この章の初めに、確率論（probability）はモデルからデータへと進み、統計学（statistics）はデータから現実の世界へと戻る、と述べた。そして三つの問いを立てた。それは本物なのか、それとも単なるノイズなのか。どのように推定すべきか。ランダムさはどのようにしてデータに入り込むのか。この章の七つの問題が明らかにしたことを以下にまとめる。

1. 本物か、ノイズか

Section 13.1 では、四つのニュースの見出しが数値を主張に変えていた。分母を考えずに総数を比べること、たった一年や一日の値から傾向を読み取ること、平均の中に傾向を隠してしまうことは、いずれも誤りであった。そしてポアソン（Poisson）モデルは、件数の自然な変動を測る物差しを与えてくれた。標準偏差（standard deviation）の 8 倍にあたる増加はほぼ確実に本物であり、ある火曜日の七人の死亡はほぼ確実に本物ではなかった。平均への回帰（regression to the mean）と、公平な比較の必要性が、この全体像を完成させた。

2. どのように推定すべきか

三つの問題によって推定（estimation）の理論を展開した。

- **最尤法（maximum likelihood）**（Section 13.2）：観測されたデータが最も起こりやすくなるように未知数の値を選ぶ。捕獲再捕獲法（capture–recapture）では、自然な推定値 $\frac{200\times200}{11} \approx 3636$ が得られ、尤度（likelihood）の山が幅広いことから、その推定がどれほど不確かであるかがわかった。
- **バイアスと分散（bias and variance）**（Section 13.3）：潜水艦の数に対する二つの不偏推定量（unbiased estimator）を考えた。一つは最大値に基づくもの、もう一つは平均に基づくものである。十分統計量（sufficient statistic）である最大値のほうがはるかに精度が高く、分散は $\frac{1}{n}$ ではなく $\frac{1}{n^2}$ のように小さくなった。同じ推論が、ドイツ軍の戦車の数について戦時中の諜報機関を出し抜いた。
- **効率と頑健性（efficiency and robustness）**（Section 13.4）：一様分布のデータでは、平均は中央値（median）の 3 倍の効率をもつ。しかし裾の重い（heavy-tailed）分布のデータでは中央値が勝り、しかも中央値は極端な誤差に対して頑健である。

3. ランダムさはどのようにしてデータに入り込むのか

ランダムさは意図的に加えることができる。ランダム化回答法（randomised response）（Section 13.5）は、各回答者が独立にランダム化を行う限り、母集団についての推定を可能にしつつ、一人ひとりのプライバシーを守る。全員が一枚の硬貨を共有してしまうと、二つの山をもつ混合分布（mixture）が生じ、ほとんど何もわからなくなる。ランダムさは処理の過程からも入り込む。丸め誤差（rounding error）は互いに打ち消し合い、$\sqrt{n}$ のようにしか増えないが、切り捨て誤差（truncation error）は線形に蓄積する。バンクーバー証券取引所（Vancouver Stock Exchange）はこのことを身をもって知った（Section 13.6）。さらに、ランダムさは設計することもできる。独立な材料をコレスキー（Cholesky）の構成によって組み合わせれば、妥当な相関（correlation）の組であればどのようなものでももつ確率変数を作り出せる（Section 13.7）。ただし、相関の組がすべて妥当であるわけではない。

4. 統計学の道具箱

手法	使った場面
ポアソン変動	見出し（13.1）
尤度とその最大値	ハタネズミ（13.2）
不偏性と分散	製造番号（13.3）、中央値（13.4）
十分統計量	製造番号（13.3）
正規近似（normal approximation）	13.4、13.5、13.6
全分散の公式（law of total variance）	ランダム化回答法（13.5）
バイアスと偶然誤差	丸め（13.6）
共分散の線形性（linearity of covariance）	相関（13.7）

この先へ

- ベイズの定理（Bayes' theorem）（Section 12.2）はベイズ統計学の基礎であり、ベイズ統計学はこの章で扱った手法に代わる主要な考え方である。
- 中心極限定理（central limit theorem）（Section 12.24）は、この章の随所で用いた正規近似を正当化する。コーシー分布（Cauchy distribution）（Section 12.22）は、それが破綻する場合を示す。
- 相関行列は半正定値行列（positive semi-definite matrix）であり、Section 7.2 の固有ベクトル（eigenvector）を用いて調べられる。

大学入学前にさらに先へ進みたい人のために

- **David Spiegelhalter, *The Art of Statistics***：一般の読者に向けた、統計的思考への最良の入門書である。
- **Tim Harford, *How to Make the World Add Up***：ニュースに出てくる数値の読み方を説く。
- **John Rice, *Mathematical Statistics and Data Analysis***：大学における数理統計学の最初の講義に相当する教科書である。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***：統計学の大きな考え方と、その由来を描く。

ジョージ・ボックス（George Box）は、すべてのモデルは間違っているが、いくつかは役に立つ、と述べた。この章では、モデルを役に立つものにする方法を示そうとしてきた。どの程度の変動を予期すべきかを知ること、多数回の繰り返しのもとでのふるまいによって手法を評価すること、そしてデータがどのようにして得られたのかを考えることである。これこそが、数値を知識に変える習慣である。

最後に

これで本書は終わりである。13 の章を通して、STEP の問題を入口として、大学最初の講義で学ぶ数学へと踏み込んできた。整数論と解析学、微積分と複素数、代数学と線形代数、微分方程式と力学系、組合せ論、力学、そして確率論と統計学である。どの問題も、その背後に何があるのかを問いかけるきっかけであり、ほとんどの場合、その答えは一つの分野まるごとであった。これらの節のうちいくつかでも、鉛筆を手に、余白の問いに取り組み、末尾の問題に挑戦しながら読み進めてきたのであれば、あなたはすでに数学者のように考え始めている。ここから先の旅は、あなた自身のものである。

「次の一歩」の解答

ここに載せているのは完全な解答ではなく、簡潔な答えとヒントである。結果と鍵となるステップを示しているので、自分の解答を確かめるのに使ってほしい。答えを見る前に、まずそれぞれの問題にしっかり取り組むこと。

13.1 見出しを読み解く

1. $1 - \sum_{k=0}^{6} \mathrm{e}^{-2.96} \frac{2.96^k}{k!} \approx 0.0315$ である。したがって、7 人以上が死亡する日は年に約 11 日ある。$\mathrm{P}(X \geq 9) \approx 0.0035$ であり、これは年に約 1 日である。

2. 死者数は $150K$ であり、K は平均（mean）1.2 のポアソン分布（Poisson distribution）に従う。平均は 180、分散（variance）は $150^2 \times 1.2 = 27\,000$、標準偏差（standard deviation）は約 164 である。死者が 125 人以下の年とは墜落事故が 1 件もない年のことであり、その確率は $e^{-1.2} \approx 0.30$ なので、まったく珍しいことではない。しかし、このモデルでは死者がちょうど 125 人になることはありえない。死者数は 150 の倍数でしか生じないからである。平均と標準偏差からはこのことはわからない。集中（clustering）があると分布はとびとびになり、確率はすべて 0、150、300、……に集まるのである。

3. $z = \frac{1349.5 - 1080}{33} \approx 8.2$ であり、確率は 10^{-16} 程度である。一定の発生率では 1350 を説明できない。何かが本当に変化したのである（あるいは死者の数え方が変わったのである）。

4. たとえば次のようになる。(i)「総数だけでは、飛行機と徒歩のどちらが安全かはわからない」、(ii)「1990 年の航空事故死者数は通常の範囲内」、(iii)「火曜日の歩行者死亡 7 人：ひどい日だが、年に約 11 回は起こること」、(iv)「1990 年、歩行者の死者が急増：原因の究明が必要」。

13.2 ハタネズミを数える

1. $\frac{p_N}{p_{N-1}} = \frac{(N-M)(N-C)}{N(N-M-C+R)}$ であり、これが 1 以上となるのは $NR \leq MC$ のとき、かつそのときに限る。したがって p_N は $N = \frac{MC}{R}$ まで増加し、その後は減少する。最大値をとるのは、この値の整数部分においてである。

2. a 人の女性と b 人の男性から c 人の委員会を選ぶ方法を、含まれる女性の人数 j

ごとに数える。各 j について $\binom{a}{j}\binom{b}{c-j}$ 通りである。

3. $\frac{\mathrm{d}}{\mathrm{d}p}\big(h\ln p + (n-h)\ln(1-p)\big) = \frac{h}{p} - \frac{n-h}{1-p} = 0$ より $p = \frac{h}{n}$ が得られる。

4. $R = 0$ のとき p_N はどこまでも増加し続け、推定値（estimate）は無限大になる。この方法は、2 回目の標本に標識された個体が十分に含まれている場合にしか機能しない。再捕獲数が少ないと、推定値は非常に不安定になる。

13.3 ドイツ戦車問題

1. 最大値を M とすると、$\mathrm{E}(M) = \frac{nN}{n+1}$、$\mathrm{Var}(M) = \frac{nN^2}{(n+1)^2(n+2)}$ なので、$\mathrm{E}(M-N)^2 = \frac{2N^2}{(n+1)(n+2)}$ である。これは $\mathrm{Var}(Z_1) = \frac{N^2}{n(n+2)}$ の $\frac{2n}{n+1}$ 倍であり、より悪い。最良の定数倍は $c = \frac{N\mathrm{E}(M)}{\mathrm{E}(M^2)} = \frac{n+2}{n+1}$ であり、そのときの平均二乗誤差（mean squared error）は $\frac{N^2}{(n+1)^2}$ で、Z_1 よりわずかに良い。

2. $\mathrm{E}(\min) = \frac{N}{n+1}$ であり、min は M と同じ分散をもつので、$\mathrm{Var}\big((n+1)\min\big) = \frac{nN^2}{n+2}$ である。これは n が大きくなっても小さくならないので、よくない推定量（estimator）である。

3. 最大値が m 以下となるのは、k 個の数がすべて $\{1,\dots,m\}$ から選ばれるとき、かつそのときに限る。$k = 1$ のとき推定量は $2X - 1$ であり、その平均は $2\cdot\frac{N+1}{2} - 1 = N$ である。

4. $98 + \frac{98}{5} - 1 = 116.6$ であり、戦車は約 117 両である。

13.4 平均か中央値か？

1. どの値を中央値（median）とするか（$2n+1$ 通り）と、残りのうちどの n 個をそれより小さい値とするか（$\binom{2n}{n}$ 通り）を選ぶ。確率は $(2n+1)\binom{2n}{n}y^n(1-y)^n\,\delta y = \frac{(2n+1)!}{(n!)^2}y^n(1-y)^n\,\delta y$ である。

2. $\mathrm{Var}(Y) = \frac{1}{4(2n+3)} = \frac{1}{20}$、$\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)} = \frac{1}{36}$ である。平均のほうが優れている。

3. $\mathrm{Var}(\max) = \mathrm{Var}(\min) = \frac{m}{(m+1)^2(m+2)}$、$\mathrm{Cov}(\max,\min) = \frac{1}{(m+1)^2(m+2)}$ であり、これより $\frac{1}{2(m+1)(m+2)}$ が得られる。これは $\frac{1}{m^2}$ のように小さくなり、平均の $\frac{1}{12m}$ よりはるかに速い。一様分布（uniform distribution）に従うデータでは、戦車の場合と同じく、両端の値が情報を担っているのである。

4. 平均は約 18.6 であり、ただ 1 つの入力ミスによって台無しになっている。中央値は 10.1 であり、影響を受けない。

13.5 答えにくい質問をする

1. 真実の回答が得られる確率を $t = ap + bq$ とする。「はい」の確率は $\frac{1}{2}t + \frac{1}{2}(1-t) = \frac{1}{2}$ なので、$\mathrm{P}(\text{表} \mid \text{はい}) = t$ であり、これが $\frac{1}{2}$ となるのは $2t = 1$ のとき、かつそのときに限る。同じ双子の一方が 2 回答える場合、「はい」の次に「いいえ」となる確率は、硬貨が表であっても裏であっても $pa(1-a) + qb(1-b)$ なので、事後確率（posterior probability）は $\frac{1}{2}$ である。2 回とも「はい」となる確率は、表のとき $pa^2 + qb^2$、裏のとき $p(1-a)^2 + q(1-b)^2$ であり、これらが等しいのは $2(ap + bq) = 1$ のとき、かつそのときに限る。

2. $\mathrm{P}(\text{はい}) = \frac{3}{4}\pi + \frac{1}{4}(1-\pi)$ なので、$\hat{\pi} = \frac{0.4-0.25}{0.5} = 0.3$ である。標準誤差（standard error）は $\frac{\sqrt{0.4 \times 0.6/400}}{0.5} \approx 0.049$ である。

3. $\mathrm{E}(Y^2) = \sum_c \mathrm{P}(C = c)\big(\mathrm{Var}(Y \mid c) + \mathrm{E}(Y \mid c)^2\big)$ である。ここから $\mathrm{E}(Y)^2 = \big(\sum_c \mathrm{P}(C = c)\mathrm{E}(Y \mid c)\big)^2$ を引くと、$\mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$ が得られる。

4. Z_2 はそれぞれ確率 $\frac{1}{2}$ で、$\frac{n}{6}$ に近い値（表）または $\frac{5n}{6}$ に近い値（裏）をとる。したがって $\mathrm{P}(Z_2 > \frac{n}{2}) \to \frac{1}{2}$ となるが、Z_2 が $\frac{n}{2}$ の近くの値をとることはほとんどない。

13.6 積み重なる誤差

1. 切り捨て（truncation）による各誤差は $(-1, 0]$ 上の一様分布に従う。それぞれに $\frac{1}{2}$ を加えると $(-\frac{1}{2}, \frac{1}{2}]$ 上の一様分布となり、四捨五入（rounding）とまったく同じになる。したがって、補正した VOZ は LEP と同等によい。

2. 各誤差は $(-0.05, 0.05)$ 上の一様分布に従い、分散は $\frac{0.1^2}{12}$ なので、合計の標準偏差は $\sqrt{10\,000 \times 0.01/12} \approx 2.9$ である。

3. $3000 \times 250 \times 0.0005 = 375$ 点である。

4. 誤差が 2 つのときは、$[-1, 1]$ 上で $1 - |x|$ となる。3 つのときは、$|x| \le \frac{1}{2}$ で $\frac{3}{4} - x^2$、$\frac{1}{2} \le |x| \le \frac{3}{2}$ で $\frac{1}{2}\big(\frac{3}{2} - |x|\big)^2$ となり、すでに正規分布（normal distribution）の曲線に近い。

13.7 相関をつくり出す

1. 標準化された（standardised）変数に対しては $\mathrm{Var}(X_1 + X_2 + X_3) = 3 + 6\rho \ge 0$ なので、$\rho \ge -\frac{1}{2}$ である。n 個の変数の場合は、$n + n(n-1)\rho \ge 0$ より $\rho \ge -\frac{1}{n-1}$ が得られる。

2. $b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}$ とすると、$(1-\rho_{12}^2)(1-\rho_{13}^2-b^2) = 1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 +$

$2\rho_{12}\rho_{13}\rho_{23} = \det R$ である。$1 - \rho_{12}^2 > 0$ なので、2 つの条件は同値である。

3. $\mathrm{Var} = \frac{1}{4}(0.04 + 0.09 + 0.12\rho)$ であり、$\rho = -1$、0、1 に対して標準偏差はそれぞれ 0.05、0.18、0.25 である。2 つの株が完全に相関（perfectly correlated）していない限り、組み合わせたもののリスクは 2 つのリスクの平均（0.25）より小さい。
4. $Z_2 = \dfrac{Y_2 - \rho_{12} Y_1}{\sqrt{1 - \rho_{12}^2}}$ である。

Beyond STEP

STEP は、ほとんどの中高生がこれまでに受けるなかで最も難しい数学の試験である。そして最も優れた試験の一つでもある。その問題は短く、見慣れず、奥が深い。なかでも最良の問題は、姿を変えた大学数学の小さな断片なのである。

本書は、1987 年の第 1 回試験に先立つ見本問題（specimen papers）から 2025 年の試験問題までのなかから、特に興味深い STEP の問題を選び、その一つひとつを扉として用いる。どの問題も、優れた学生が途中で抱く疑問に触れながら丁寧に解き、そのうえで扉の向こうへと踏み出す。その考え方はどこから来たのか、大学では何へと発展するのか、次に何を読めばよいのか。

本書は、STEP に向けて準備を進める才能ある数学者の卵のうち、過去問演習以上のものを求める人のために書かれている。数学専攻の学部生のように考え始め、その先に何が待っているのかを知ってほしい。

- 純粋数学（pure mathematics）、力学（mechanics）、確率（probability）、統計（statistics）にわたる 13 の章
- STEP の過去問 136 題、それぞれに道筋を示した解答つき
- すべての問題の背後にある大学数学
- 解答つきの発展演習と、さらに学ぶための文献案内

全 13 章のうちの第 13 章。 本書全体、すべての章、すべての解答は、次のサイトで自由に閲覧・ダウンロード・共有できる：

beyondstep.keithmcnulty.org

本文はパブリックドメイン（public domain）に提供されている（CC0 1.0）。STEP の問題 © Cambridge University Press & Assessment.