# 統計入門　数理サプリメント

**講義全体の数理を掘り下げる副教材**

神谷之康

初版 2026 年 6 月 25 日 ／ 更新 2026 年 7 月 6 日

# 目次

# はじめに

本書は、京都大学 1 回生向けの講義「統計入門」（全 13 回）の**数理サプリメント**である。各回のスライドが概念と使い方を「直感→ 具体例 →数式」の順で示すのに対し、本書はその数式の側に一歩踏み込み、公式がなぜそうなるのか、分布どうしがどうつながっているのかを、定義から順を追って確かめる。教科書 R.A. Poldrack『統計的思考』（神谷之康訳、朝倉書店）を補完する位置づけである。文章化・構造化には生成 AI（Claude）を活用したが、内容の企画・判断・最終的な責任は著者にある。

統計の公式は「覚えるもの」に見えやすい。しかし期待値・分散の性質も、正規・カイ二乗・$t$ 分布のつながりも、ベイズの更新も、少数の定義と論理から**導けるもの**である。本書のねらいは、各回で登場した結果を「与えられた事実」から「自分で再構成できる道具」に変えることにある。試験前の確認にも、講義中につまずいた箇所の橋渡しにも使えるよう、各章は独立して読めるが、前から読むほうがつながりは見えやすい。

数式の理解に必要な範囲は高校＋ α にとどめ、積分や行列が出る箇所は最小限の説明を添えた。難しければ結論（枠で囲った式や「要点」）だけ拾っても構わない。

---

## 本書の流れ

**基礎**（講義の中心となる内容）

- **統計の重要ポイント** ―講義全体を貫く概念の背骨。記述統計から推測統計までの考え方を一本の線で整理する。
- **公式の導き方** ―期待値・分散の線形性、標準化、不偏分散の $n-1$ など、数理の「幹」をたどる。
- **確率分布のまとめ** ―主要な分布の意味とつながり。正規・カイ二乗・$t$ の標本分布の一族、ベータ分布・コーシー分布までを 1 枚の地図にする。
- **ベイズ統計の数理** ―事前分布・尤度・事後分布・周辺尤度という 4 つの登場人物の役割分担を、コイン投げの通し例で追う。BIC とベイズファクターにも触れる。

**発展的トピック ――進んだ分析手法**（一歩進んだ内容。まずは考え方の地図として）

- **因果推論の初歩** ―観察データから因果効果を読むための共変量調整。因果グラフ（DAG）にもとづき、交絡は入れ、コライダーは外す、という変数選択を線形モデルの数式で確かめる。
- **機械学習の初歩** ―「推論」と「予測」の違いから出発し、予測誤差のバイアス・バリアンス分解、過学習とクロスバリデーション、AIC・BIC（および WAIC・WBIC）の関係を整理する。

**付録**

- **R コードの読み方** ―コードを「指示の文」として読む。講義で使う base R の書き方を文法から押さえる。

数式は本文中にそのまま表示している。

本書はウェブ版も公開している：https://kamitanilab.github.io/stats-math-supplement/

---

## 著者・参照・ライセンス

**神谷之康**（かみたに・ゆきやす）—— 京都大学大学院情報学研究科教授／ ATR フェロー。研究室ウェブサイト: Kamitani Lab

教科書として参照している R. A. Poldrack『統計的思考』(神谷之康訳、朝倉書店)の原著 *Statistical Thinking for the 21st Century* は、ウェブブックとして無料公開されている（statsthinking21.org）。本書とあわせて参照してほしい。

# Part I.

# 基礎

# 1. 統計の重要ポイント ——講義を貫く概念の背骨

この章は、講義（全 13 回）で扱った内容を、**個々のトピックの寄せ集めとしてではなく、一本の筋でつながった一つの考え方として**見渡すためのものです。統計学には公式や手法がたくさん出てきます。けれども、その奥にあるアイデアは驚くほど少なく、しかも互いに深くつながっています。姉妹編の「公式の導き方」「確率分布のまとめ」が**式の道具箱**だとすれば、この章はそれらを**どういう順番で・なぜ使うのか**という地図です。

> 読み方のコツ
>
> 各ポイントは「**なぜ重要か**」「**一言でいうと**」「**典型的な誤解**」の 3 点セットで書いています。式は意味を補強するときだけ最小限に置きました。まず太字の主張を拾い読みし、気になったところを式まで降りていく、という読み方でかまいません。

---

## 0. 全体地図：13 回はどうつながっているか

統計学の流れは、おおよそ次の一本の物語として読めます（図 1.1）。

この地図の**中心に置かれているのが第 5 回「データ＝モデル＋誤差」**です。記述統計も、推測も、回帰も、すべてこの一行の言い換えだと分かると、13 回がばらばらの手法集ではなく**一つの態度**に見えてきます。まずここから始めます。

---

## 1. データ ＝ モデル ＋誤差（中心思想）

**なぜ重要か**：統計学のほぼすべての手法は、次のたった一行の上に立っています。

$$\text{データ} = \text{モデル} + \text{誤差}$$

- **モデル**とは「データをこう要約・予測する」という**こちらが立てた約束**です。いちばん単純なモデルは「全部を 1 つの数で代表する」、つまり**平均**です。
- **誤差**とは、そのモデルでは説明しきれない**残りのズレ**です。

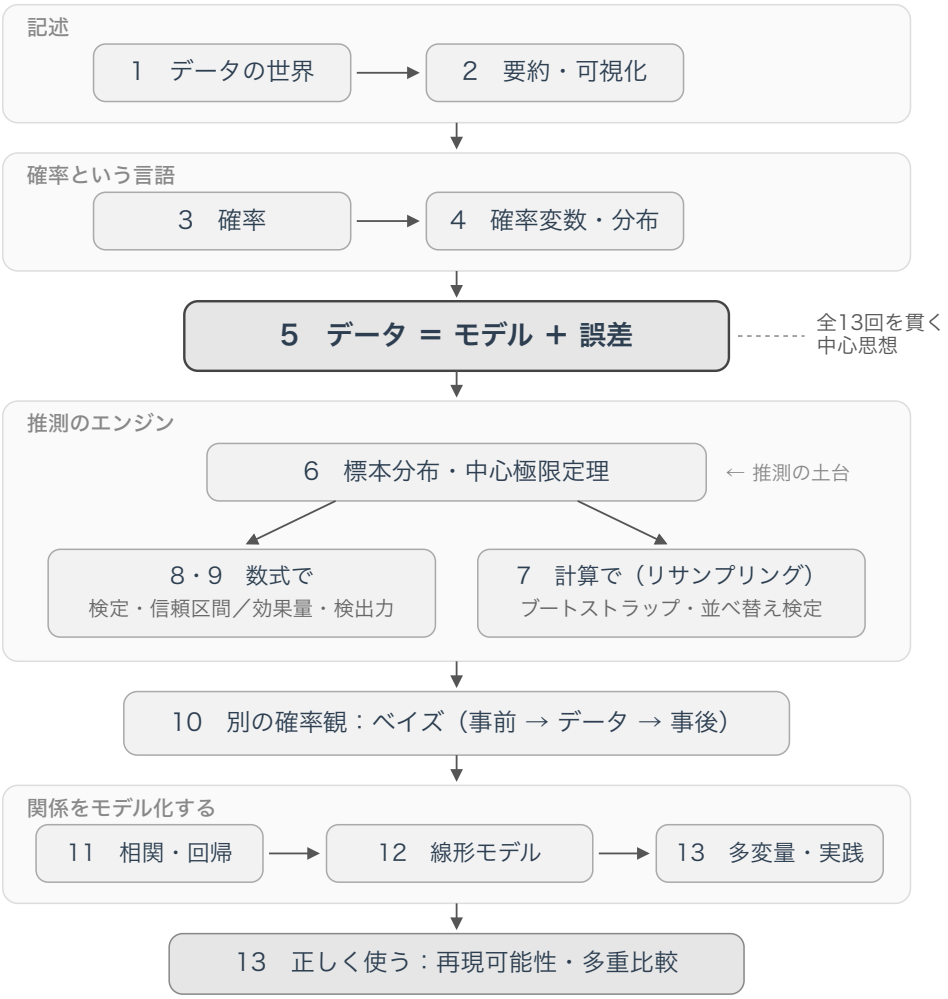



**図 1.1：**講義全 13 回の地図。上から「記述」「確率という言語」を経て、中心思想「データ＝モデル＋誤差」（第 5 回）へ。そこから推測のエンジン（標本分布・中心極限定理）が「数式で」（8・9 回）と「計算で」（7 回）の 2 ルートに分かれ、ベイズ（10 回）、関係のモデリング（11〜13 回）を経て「正しく使う」（13 回）に至る。

データ分析とは、**モデルを工夫して誤差をできるだけ小さく・意味のあるものにしていく作業**にほかなりません。

| 回 | その回の「モデル」 | 「誤差」 |
|---|---|---|
| 2 | 平均(1 つの数で代表) | 各データの平均からのズレ |
| 11 | 回帰直線 $y = a + bx$ | 直線から外れる縦方向のズレ(残差) |
| 12 | 群ごとの平均($t$ 検定・分散分析) | 群内のばらつき |
| 13 | 複数の変数を使った予測式 | それでも残る部分 |

**一言でいうと**:手法が変わっても、やっていることは「モデルを置いて、残った誤差を測る」だけ。第 2 回の平均と第 11 回の回帰は、同じ枠組みの単純な場合と複雑な場合の違いにすぎません。

**典型的な誤解**:「誤差＝間違い・測定ミス」だと思ってしまうこと。ここでの誤差は**モデルでは捉えない自然なばらつき**を含み、消すべきものではなく**理解すべきもの**です。誤差をゼロにしようと変数を足しすぎると、後で出てくる**過学習(オーバーフィッティング)**(13 回)に陥ります。

---

## 2. 母集団と標本、母数と統計量 ——推測の構図

**なぜ重要か**:われわれが本当に知りたいのは、たいてい手元のデータそのものではなく、その背後にある**全体(母集団)**です。日本人全体の平均身長を知りたいが、測れるのは数百人(標本)だけ、という状況です。

- **母集団**:知りたい対象の全体。その特徴を表す数値が**母数(パラメータ)**で、たとえば母平均 $\mu$・母分散 $\sigma^2$。これは**定数**だが、ふつう**未知**。
- **標本**:実際に手に入れた一部のデータ。そこから計算する数値が**統計量**で、標本平均 $\bar{X}$・標本分散 $S^2$ など。これは標本ごとに**変わる**量なので、確率変数として大文字で書きます(観測された具体的な値は小文字 $\bar{x},\ s^2$)。

推測統計とは、**観測できる統計量を手がかりに、観測できない母数を推し量る**営みです。この一方向の矢印—**母数(未知の定数)** → **標本** → **統計量(既知だが揺れる)**—が、第 6 回以降すべての土台になります。

**一言でいうと**:統計量は母数の「影」。影(標本平均)を見て本体(母平均)を当てにいく。

**典型的な誤解**:標本平均 $\bar{X}$ と母平均 $\mu$ を混同すること。$\bar{X}$ は計算できるが標本ごとにブレる量、$\mu$ は 1 つに定まっているが直接は見えない量。両者の橋渡しをするのが次の**標本分布**です。

---

## 3. 記述統計の 3 本柱:中心・ばらつき・形

データを要約するとき、見るべきは大きく 3 つです(第 2 回)。

### 中心（代表値）——平均と中央値

- **平均** $\bar{x} = \frac{1}{n}\sum x_i$ ：全データを反映するが、外れ値に弱い。
- **中央値**：順番に並べた真ん中。外れ値に強い。

両者は無関係ではなく、第 5 回の視点で統一されます。**平均は「誤差の 2 乗和」を最小にする代表値、中央値は「誤差の絶対値の和」を最小にする代表値**です。つまり「どんな誤差を小さくしたいか」で代表値が決まる— これも「データ＝モデル＋誤差」の現れです。

### ばらつき（散布度）—— 分散・標準偏差・なぜ n − 1

- **分散** ＝ 平均からのズレの 2 乗の平均。**標準偏差** ＝ その平方根で、元のデータと単位がそろう。
- **なぜ $n-1$ で割るのか**：標本から母分散を推定するとき、ズレを測る基準に真の母平均 $\mu$ ではなく手元の標本平均 $\bar{x}$ を使う。$\bar{x}$ はそのデータについて 2 乗和を**最小にする点**なので、ばらつきが**少なめに見積もられる**。これを補正するため $n$ より 1 小さい $n-1$ で割って少し大きくする（自由度の調整）。
  - R の `var()` は不偏分散（$n-1$ 版）、`sd()` はその平方根を返します。導出は「公式の導き方」第 8 節。

### 形 ——分布の歪み

- 左右対称か、片側に長い裾を引く（**歪度・ロングテール**）か。年収や都市人口のように右に長い裾を引く分布では、平均が中央値より大きくなり、平均だけ見ると実態を見誤る。

**一言でいうと**：中心・ばらつき・形の *3* 点を見れば、データの「顔つき」がつかめる。1 つの数（平均）だけで語らないこと。

**典型的な誤解**：「平均を見れば十分」。同じ平均でも、ばらつきや形がまるで違うデータはいくらでもあります。可視化（ヒストグラム・箱ひげ図）で形を必ず確かめる、というのが第 2 回の教訓です。

---

## 4. 確率という言語 ——2 つの見方、条件付き確率、独立

推測統計は「不確実さ」を扱うので、その言語である**確率**が必要です（第 3・4 回）。

### 確率の 2 つの見方

- **頻度主義**：確率とは「同じことを無限に繰り返したときの相対頻度」。コインの表が出る確率 0.5 とは「投げ続ければ表が半分」という意味。客観的だが、「1 回きりの出来事」には当てはめにくい。
- **ベイズ主義**：確率とは「ある主張に対する**確信の度合い**」。データを見て確信を更新していく。1 回きりの出来事や未知のパラメータにも確率を与えられる。

この対立は第 8〜9 回（頻度主義の検定）と第 10 回（ベイズ）で再び主役になります。**同じデータでも立場が違えば答えの意味が変わる**— これは欠陥ではなく、確率という概念がそもそも一つではないことの反映です。

### 条件付き確率と独立

- **条件付き確率** $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ :「$B$ が起きたと分かったうえでの $A$ の確率」。情報が増えると確率は更新される、という発想の出発点で、ベイズの定理(第 10 回)に直結します。
- **独立**:$P(A \cap B) = P(A)P(B)$、同じことだが $P(A \mid B) = P(A)$。**$B$ を知っても $A$ の見込みが変わらない**こと。後の「独立な和の分散は足せる」(標準誤差の導出)や「並べ替え検定の前提」で効いてきます。

**一言でいうと**:条件付き確率は「情報で確率を上書きする」操作、独立は「上書きしても変わらない」関係。

**典型的な誤解**:$P(A \mid B)$ と $P(B \mid A)$ を取り違える「**検察官の誤謬**」。「病気なら陽性になる確率」と「陽性だったとき病気である確率」はまったく別物。両者を結ぶのがベイズの定理です(ポイント 10)。

---

## 5. 確率変数・期待値・分散・正規分布・標準化

**なぜ重要か**:データを「確率的に生まれる数」として扱う道具立てです(第 4 回)。

- **確率変数**:偶然で値が決まる変数。サイコロの目、ランダムに選んだ 1 人の身長など。
- **期待値** $E[X]$:その変数の「平均的な値」(値 × 確率を全部足す)。
- **分散** $\mathrm{Var}(X)$:期待値まわりのばらつき。
- **期待値の線形性** $E[aX + b] = aE[X] + b$:最も使う性質で、二項分布の $E[X] = np$ も標準誤差も、すべてここから出る(「公式の導き方」第 1・3〜5 節)。

### 正規分布と標準化($Z$ スコア)

- **正規分布**:左右対称の釣鐘型。**たくさんの独立な量を足す・平均すると現れる**(理由はポイント 6 の中心極限定理)。だから自然界・社会のあちこちに登場する。
- **標準化** $Z = \dfrac{X - \mu}{\sigma}$ :平均を引いて標準偏差で割ると、平均 0・標準偏差 1 にそろう。**単位や尺度の違う量を共通のものさしで比べられる**ようになる。
- 68–95–99.7 ルール:正規分布では $\mu \pm 1\sigma$ に約 68%、$\mu \pm 2\sigma$ に約 95%。検定や信頼区間で出てくる **1.96** はここから来ています。

**一言でいうと**:標準化は「ものさしの統一」。$Z$ スコアにすれば身長と試験点数すら同じ土俵で比べられる。

**典型的な誤解**:「どんなデータも正規分布する」。そうではなく、**たくさん足し合わせた量**が正規分布に近づくのです(次のポイント)。生データ自体は歪んでいることも多い。

---

## 6. 標本分布と中心極限定理 ——推測を可能にするエンジン

**なぜ重要か**：これが**推測統計全体の心臓部**です（第 6 回）。第 8〜10 回の検定・信頼区間・ベイズは、すべてここを土台に立っています。

- **標本分布**：標本平均 $\bar{X}$ は標本ごとに違う値をとる ―つまり $\bar{X}$ **自身が一つの確率変数**で、分布を持つ。これが標本分布。「統計量も揺れる」（ポイント 2）を具体的に表したもの。
- **標準誤差** SEM $= \dfrac{\sigma}{\sqrt{n}}$：標本平均のばらつきの大きさ。$n$ を大きくすると小さくなる。ただし $\sqrt{n}$ なので、**精度を 2 倍にするには標本を 4 倍**必要。
- **中心極限定理（CLT）**：元のデータがどんな形でも、**標本平均の分布は $n$ が大きくなると正規分布に近づく**。

なぜこれが「エンジン」なのか。母集団の形が分からなくても、**標本平均は（$n$ が十分大きければ）正規分布として扱える**。だから「$\bar{x}$ が母平均からどれだけ離れているか」を正規分布の確率で評価でき、検定も信頼区間も計算できる。CLT がなければ、推測のほとんどは成り立ちません。

**つながりの要**：

$$\text{中心極限定理} \longrightarrow \text{検定統計量 } z = \frac{\bar{x} - \mu_0}{\text{SEM}} \longrightarrow \text{仮説検定・信頼区間 (8)・検出力 (9)}$$

**一言でいうと**：個々はバラバラでも、平均すれば正規分布になる。これが「平均」という要約がこれほど信頼される理由。

**典型的な誤解**：「CLT は元のデータを正規分布にする」のではなく、「**標本平均**の分布を正規分布にする」。また「$n$ が大きければばらつきが消える」のではなく、**ばらつきの中心（母平均）に近づく**のが大数の法則、**ばらつきの形が正規になる**のが CLT。

---

## 7. リサンプリング ——数式の近似を「計算」で置き換える

**なぜ重要か**：第 6 回までの結果（SEM、正規近似）は**数式による近似**でした。第 7 回は、コンピュータの反復計算で**同じことを直接シミュレートしてしまう**という発想の転換です。

- **ブートストラップ**：手元の標本から**復元抽出で何度も再標本を作り**、そのたびに統計量を計算する。こうして作った分布が標本分布の代わりになり、標準誤差や信頼区間が**数式なしで**得られる。
- **並べ替え検定（ランダム化検定）**：2 群のラベルをランダムに何度も入れ替えて、「もし群に差がなければ（＝ラベルがどうでもよければ）どんな差が偶然生じるか」を直接作り、実際の差がそのどのあたりに来るかを見る。

両者に共通するのは、「**仮定（正規分布など）に頼る代わりに、データを再利用して必要な分布を作ってしまう**」という考え方です。式が解けない複雑な統計量でも適用できるのが強み。

**一言でいうと**：紙とペンで近似していたものを、*PC* に何千回も実行させて直接見る。CLT と検定の「裏道」。

**典型的な誤解**：「ブートストラップは魔法でデータを増やす」のではない。情報量は元の標本のまま。あくまで**手元のデータが母集団を代表している**という前提のもとで、ばらつきを評価しているだけ。

---

## 8. 仮説検定の論理 ——帰無仮説・p 値・2 種類の誤り

**なぜ重要か**：科学で最も広く使われ、同時に**最も誤解されている**枠組みです（第 8 回）。

論理は背理法に似ています。1. **帰無仮説** $H_0$ を立てる（例：「差はない」「効果はない」）。2.「もし $H_0$ が正しいなら、観測データは**どれくらい起こりにくいか**」を計算する。これが **p 値**。3. p 値が小さければ「$H_0$ のもとでは珍しすぎる」として $H_0$ を棄却する。

**p 値の正しい意味**：帰無仮説が正しいと仮定したとき、観測された結果**以上に極端な**結果が得られる確率。

**2 種類の誤り**：

| | $H_0$ は本当は正しい | $H_0$ は本当は誤り |
|---|---|---|
| $H_0$ を棄却した | **第 1 種の誤り**（あわてんぼう、確率 $\alpha$） | 正しい判断 |
| $H_0$ を棄却しない | 正しい判断 | **第 2 種の誤り**（見逃し、確率 $\beta$） |

検出力（ポイント 9）は「本当に差があるとき、ちゃんと差を見つけられる確率」$= 1 - \beta$ です。

**信頼区間** ——「**有意かどうか**」**を区間で表す**：検定が「ある 1 点 $\mu_0$ を棄却するか」を問うのに対し、信頼区間は**棄却されない値をぜんぶ集めた範囲** を示します。正規近似のもとで 95%信頼区間 $= \bar{x} \pm 1.96\,\mathrm{SEM}$（母標準偏差が未知で標本が小さいときは 1.96 を $t$ 分布の臨界値に置き換える）。差や効果量のように帰無値が 0 の量では、**区間が 0 を含むかどうかが検定の結論とほぼ対応**します。点ではなく **幅**で答えるので、「有意か否か」に加えて「推定の精度」まで一目で伝わるのが利点です。

> **典型的な誤解**：信頼区間を「母平均が 95%の確率でこの区間に入る」と読むこと。母平均は定数で動かない。正しくは「**同じ手続きを繰り返せば、作られる区間の 95%が母平均を含む**」— ランダムなのは母平均ではなく区間のほう。

**一言でいうと**：「偶然だけでこんな結果が出るのは珍しいか?」を確率で測る。

**典型的な誤解（とても重要）**：- p 値は「**帰無仮説が正しい確率**」では**ない**。p 値は $H_0$ を**仮定したうえでのデータの珍しさ**であって、$H_0$ 自体の確率ではない。これを知りたいならベイズ（ポイント 10）が必要。-「p > 0.05 だから効果はない」は誤り。**有意でない＝差がないことの証明ではない**（見逃しかもしれない）。- p 値は効果の**大きさ**を表さない。$n$ が巨大なら、ごく小さな差でも有意になる。だから次のポイントが要る。

---

## 9. 効果量・検出力・サンプルサイズ設計 ——「有意かどうか」を超えて

**なぜ重要か**：p 値だけでは「差があるか／ないか」の Yes/No しか分からない。**どれだけ大きい差か**、そして**その差をちゃんと捉えられる研究になっているか**を語るのが第 9 回です。

- **効果量**（例：Cohen's $d$）：差の**大きさ**を標準偏差を単位として表す。p 値と違い**標本サイズに左右されにくい**。「統計的に有意」と「実質的に重要」は別物だ、という戒め。
- **検出力**：本当に効果があるとき、それを有意と検出できる確率 $1-\beta$。検出力の低い研究は、たとえ有意でも信用しにくく、再現もしにくい。

## サンプルサイズ設計 ——4 つの要素の関係

検出力をめぐっては、次の **4 つの要素**が独立ではなく、**3 つを決めれば残りの 1 つが自動的に決まる**という関係で結ばれています。これを利用して研究を設計します。

| 要素 | 意味 | ふつうの扱い |
|---|---|---|
| 効果量 $d$ | 見つけたい差の大きさ | 過去研究などから見積もる |
| $\alpha$ | 有意水準（第 1 種の誤り） | 0.05 に固定 |
| $1-\beta$ | 検出力（見逃さない確率） | 0.80 を目標に固定 |
| $n$ | 標本サイズ | これを逆算して決める |

研究計画では普通、$\alpha = 0.05$ と検出力 $1-\beta = 0.80$ を**先に決め、期待される効果量** $d$ を見積もったうえで、**必要な標本サイズ** $n$ **を逆算**します。「期待した効果量が本当にあったとき、それを有意にできるだけの $n$ を、データを取る前にあらかじめ用意しておく」— これがサンプルサイズ設計です（R では `power.t.test()` などで計算できる）。直感的には、**効果量** $d$ **が小さいほど、検出力を高く望むほど、必要な** $n$ **は大きく**なります（$n$ はおよそ $1/d^2$ で増える）。

## 注意：標本は大きすぎてもいけない

ここに大事な落とし穴があります。検定統計量は $z = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ のように $\sqrt{n}$ **に比例して大きく** なります。つまり $n$ **を増やしさえすれば、どんなに小さな差でも有意にできてしまう**のです。

- $n$ が **小さすぎる**と、本当にある効果も見逃す（検出力不足）。
- $n$ が **大きすぎる**と、実質的に無意味なほど小さな効果まで「統計的に有意」になり、p 値が独り歩きする。

だからサンプルサイズ設計のねらいは「とにかく大きく」ではなく、**見つける価値のある効果量を、ちょうど検出できるだけの** $n$ **に合わせる** ことです。そして結果が有意になったら必ず**効果量と信頼区間（ポイント 8）を併記**し、「**有意か**」と「**実質的に意味があるほど大きいか**」を分けて読む— これがポイント 8 とあわせた結論です。なお、$\alpha = 0.05$ という水準は自然界の境界ではなく**約束事**にすぎません。「有意か否か」の二分で終わらせず、効果量と信頼区間で**連続的に**語る— これが現代の標準的な作法です。

**一言でいうと**：「有意」かどうかより、「どれくらい大きく、どれだけ確かに、そして適切な $n$ で測れているか」を見る。

---

## 10. ベイズの考え方 —— 事前 → データ →事後

**なぜ重要か**：頻度主義（8〜9 回）とは**別の確率観**から推測を組み立てる枠組みです（第 10 回）。ポイント 8 で「知りたかったのに頻度主義では答えられない」とした「**仮説が正しい確率**」に、ベイズは正面から答えます。

中心はベイズの定理です。

$$\underbrace{P(\text{仮説} \mid \text{データ})}_{\text{事後}} = \frac{\overbrace{P(\text{データ} \mid \text{仮説})}^{\text{尤度}} \overbrace{P(\text{仮説})}^{\text{事前}}}{P(\text{データ})}$$

- **事前分布**：データを見る前の確信。
- **尤度**：その仮説のもとでデータがどれだけ起こりやすいか。
- **事後分布**：データを見た後の更新された確信。

学習とは「**事前にデータの証拠を掛け合わせて事後に更新する**」こと。今日の事後は明日の事前になり、データが増えるほど確信が研がれていきます。

| | 頻度主義（8〜9 回） | ベイズ（10 回） |
|---|---|---|
| 確率とは | 長期的な相対頻度 | 確信の度合い |
| パラメータ | 未知の**定数** | **確率分布**を持つ |
| 答えるもの | p 値・信頼区間 | 事後分布・**信用区間** |
| 区間の意味 | 手続きの 95%が母数を含む | 母数が 95%の確率でこの区間に入る（直感どおり!） |

**一言でいうと**：確信を数値にして、データで更新していく。検査の的中率（ポイント 4 の検察官の誤謬）が直感とずれる理由も、ベイズが説明する。

**典型的な誤解**：「ベイズと頻度主義はどちらかが正しい」。そうではなく、**問いの立て方が違う**。事前分布が結論に与える影響を意識すれば、両者は補い合う道具です。

---

## 11. 関係のモデリング ——相関と回帰、そして相関≠因果

**なぜ重要か**：1 つの変数の要約（〜10 回）から、**2 つ以上の変数の関係**へ進みます（第 11 回）。「データ＝モデル＋誤差」のモデルが、定数（平均）から直線へ格上げされる回です。

- **相関係数** $r$：2 変数が一緒に動く度合い。$-1 \le r \le 1$ で、符号が向き、絶対値が強さ。
- **回帰直線** $y = a + bx$：$x$ から $y$ を予測する直線。傾き $b$ は最小二乗（誤差 2 乗和の最小化）で決まる — **平均を求めたのと同じ原理**（「公式の導き方」第 6・11 節）。

### 相関 ≠因果（この回の核心）

2 つが相関していても、片方がもう片方の**原因とは限らない**。考えられるのは、1. $X \to Y$（本当に原因）、2. $Y \to X$（逆向き）、3. **第 3 の変数 $Z$ が両方の共通原因**（交絡）、4. 偶然。

典型例：アイスクリームの売上と水難事故は相関するが、原因は両方を押し上げる「**気温**」という**交絡変数**。気温で層別すれば見かけの相関は消える（層別すると関係の向きが変わって見える**シンプソンのパラドックス**も、交絡が生む同類の現象です）。

因果を本当に主張するには、**ランダム化比較試験 (RCT)** のように**こちらが介入して操作する**必要がある。観察データだけからは、交絡を完全には排除できない。それでも、交絡を正しく選んで調整すれば、観察データから因果効果に**迫る**ことはできます— その方法が第 5 章「因果推論の初歩」の主題です。

**一言でいうと**：**一緒に動くことと、一方が他方を動かすことは別物**。相関は出発点であって結論ではない。

**典型的な誤解**：$r = 0$ を「無関係」と読むこと。$r$ は**直線的な**関係しか測らない。U 字型のようなきれいな曲線関係でも $r \approx 0$ になりうる。散布図を必ず見ること。

---

## 12. 線形モデル（LM）——ばらばらの手法を 1 つの枠に

**なぜ重要か**：$t$ 検定・分散分析・回帰 —別々の名前で習う手法が、**実は同じ一つのモデルの特殊な場合**だと明かされる回です（第 12 回）。13 回を通じて積み上げてきた「データ＝モデル＋誤差」が、ここで統一的な姿に結晶します。

線形モデル（LM）の骨格は、結局すべて

$$\text{結果} = (\text{説明変数で作った予測}) + \text{誤差}$$

の形です。違うのは**説明変数の種類**だけ。

| 手法 | 説明変数 | 線形モデルでの見え方 |
|---|---|---|
| 1 標本 $t$ 検定 | なし（定数のみ） | 切片だけのモデル |
| 2 標本 $t$ 検定 | 2 値（群 A/群 B） | ダミー変数 1 つの回帰 |
| 分散分析 (ANOVA) | カテゴリ（3 群以上） | ダミー変数複数の回帰 |
| 単回帰 | 連続変数 1 つ | $y = a + bx$ |
| 重回帰 | 連続・カテゴリ複数 | $y = a + b_1 x_1 + b_2 x_2 + \cdots$ |

カテゴリ変数（性別など）は**ダミー変数**（0/1）で数式に入れる。こうすると「群の平均を比べる」検定が「回帰の傾き」として表せる。**$t$ 検定で群差を見ることと、回帰で傾きを見ることは、同じ計算**なのです。

**一言でいうと**：**手法の名前は入口が違うだけ。中身は「予測＋誤差」という一つのモデル**。これを知ると、新しい手法も「あの枠の拡張だ」と見通せる。

**典型的な誤解**：$t$ 検定・分散分析・回帰を**別々に暗記すべき三つの公式**だと思うこと。実際は同じ枠の中で説明変数を取り替えているだけ。覚える量は思うより少ない。

---

## 13. 正しく使うために ——再現可能性・多重比較・p-hacking

**なぜ重要か**：どんなに正しく計算しても、**使い方を誤れば結論はゆがむ**。第 13 回（と第 1 回の問題提起）は、統計を「道具」ではなく「**責任を伴う実践**」として締めくくります。

- **多重比較問題**：たくさんの検定を同時に行うと、本当は差がなくても**偶然どれかが有意**になる。20 回検定すれば、$\alpha = 0.05$ でも平均 1 回は「偶然の有意」が出る。対策はボンフェローニ補正など、有意水準を厳しくすること。
- **p-hacking / 疑わしい研究実践 (QRPs)**：有意になるまでデータを追加したり、たくさん試した中から都合のよい結果だけ報告したりすること。これは多重比較を隠れて行っているのと同じで、**偶然を実力と見せかけてしまう**。
- **再現性の危機**：上のような実践と検出力の低さが重なり、有名な研究の多くが追試で再現されなかった（再現性プロジェクト）。
- **対策**：**事前登録**（分析計画を先に公開）、**データ・コードの共有**、**追試**の尊重。

これらは新しい話題に見えて、実は本章のポイント 8〜9 の裏返しです。**p 値の意味（8）を誤解し、検出力（9）を軽視すると、ここで失敗する**。最後のポイントは、最初に戻ってくるわけです。

**一言でいうと**：統計は「やり方」だけでなく「使い方の誠実さ」まで含めて初めて意味を持つ。

**典型的な誤解**：「有意な結果が出た＝発見」。一度きりの有意は、多重比較や p-hacking の産物かもしれない。**再現されて初めて信頼できる**、というのが現代の統計の到達点です。

---

## まとめ：少数のアイデアがすべてを貫いている

13 回をふり返ると、表面の手法は多彩でも、**奥で働いているアイデアはごく少数**だと分かります。

1. **データ ＝ モデル ＋ 誤差**（1）——平均も回帰も $t$ 検定も、すべてこの一行。モデルを置き、残った誤差を測る。
2. **統計量は母数の影**（2）——観測できる標本から、観測できない母集団を推し量る。
3. **標本分布と中心極限定理**（6）——「平均は正規分布になる」。これが推測（検定・区間・ベイズ）すべての土台。
4. **誤差・ばらつきを正しく評価する**（3, 7, 8, 9）——n − 1、ブートストラップ、信頼区間、効果量・検出力は、いずれも「ばらつきをどう測り、どう伝えるか」の工夫。
5. **確率には 2 つの見方がある**（4, 8, 10）——頻度主義とベイズ。同じデータでも問いの立て方で答えの意味が変わる。
6. **使い方の誠実さ**（11, 12, 13）——相関≠因果、線形モデルという統一視点、多重比較・再現可能性。正しく計算するだけでは足りない。

公式を 1 つずつ暗記するのではなく、**この 6 本の幹のどこに各手法がぶら下がっているか**を地図として持つこと ― それが、統計学を「ばらばらの手続きの集まり」ではなく「**一つの考え方**」として理解した状態です。式の細部は「公式の導き方」「確率分布のまとめ」に戻って、いつでも自分の手で再現できるようにしておきましょう。

# 2. 公式の導き方 ——数理の「幹」をたどる

この章は、講義で扱った主要な公式が「**なぜ・どこから出てくるのか**」を、順を追って確かめるためのものです。公式を丸暗記するのではなく、**少数の道具からすべてが枝分かれしている**という見取り図を持ち、**短い式変形で自分の手で再現できる**ことを目標にしてください。これは試験のための要点集ではなく、講義全体を数理の側から深く理解するための副教材です。

> 読み方のコツ
>
> 各節はおおむね「① 直感（何を言いたいか） → ② 準備（使う道具） → ③導出（式変形） → ④確認（具体例）」の順です（節によっては一部を省きます）。式変形は 1 行ずつ、右側の理由とあわせて追ってください。

---

## 0. 準備：記号と 3 つの道具

### 確率変数とは何か ——普通の変数との違い

これから出てくる式には、$X$ や $\bar{X}$ といった文字がくり返し登場します。これらは中学・高校で習った「普通の変数」とは**意味が違う**ので、まずここをはっきりさせておきます。ここがあいまいなままだと、$E[X]$ や $P(X = x)$ という記号が何を言っているのか分からなくなります。

**普通の変数** $x$：方程式 $x + 3 = 5$ の $x$ のように、「**まだ分からないが、本当は 1 つに決まっている数**」を表す入れもの。解けば $x = 2$ という 1 つの値に確定します。値は 1 つです。

**確率変数** $X$：サイコロを振って出る目のように、「**値が偶然によって決まる量**」。振る前は値が 1 つに決まっておらず、$1, 2, 3, 4, 5, 6$ の**どれもが起こりうる**。確率変数とは 1 つの数ではなく、「**とりうる値の全体**」と「**それぞれの値の起こりやすさ**」をセットで持った量だと考えてください。

この違いを、記号の **大文字・小文字** で書き分けます。

- **大文字** $X$ … 確率変数そのもの（偶然で値が決まる「量」）。
- **小文字** $x$ … その確率変数が**とりうる具体的な値の 1 つ**（普通の数）。

## 表記 $P(X = x)$ の読み方

すると、講義でくり返し出てきた

$$P(X = x)$$

は、「**確率変数 $X$ が、具体的な値 $x$ をとる確率**」と読みます。ここで $X$（大文字）は偶然で揺れる量、$x$（小文字）はそこに代入する 1 つの数で、両者の役割はまったく違います。たとえばサイコロなら

$$P(X = 3) = \frac{1}{6}$$

は「出る目 $X$ が、ちょうど 3 という値になる確率は $1/6$」という意味です。$P(X = x)$ 全体は **1 つの数（確率）**になります。

> よくある混同：$P(X = x)$ を「$P(X)$ が $x$ に等しい」と読んではいけません。正しくは $X = x$（**$X$ が値 $x$ をとる）という出来事の確率** です。$X$ は揺れる量、$x$ は代入する値、と分けて読むのがコツです。

## 確率変数には確率分布が対応する

$x$ を動かしながら $P(X = x)$ をすべて並べたもの —— これが**確率分布**です。確率分布こそが、その確率変数の「正体」をすべて語ります。

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $P(X = x)$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ |

確率分布は必ず次を満たします（確率の決まりごと）：

$$P(X = x) \geq 0, \qquad \sum_{x} P(X = x) = 1.$$

「どの確率も 0 以上」「ぜんぶ足すと 1（必ずどれかは起こる）」というだけのことです。

**確率変数 → 確率分布** という対応が、この章全体の土台です。期待値 $E[X]$ も分散 $\mathrm{Var}(X)$ も、結局は「確率変数 $X$ の確率分布から計算される値」にすぎません。実際、次の道具 1 の期待値の式 $E[X] = \sum_i x_i\, P(X = x_i)$ は、上の表の「値 $x$」と「その確率 $P(X = x)$」を掛けて足しているだけです。

（連続の確率変数では、1 点をとる確率は 0 になるため $P(X = x)$ ではなく**確率密度関数** $f(x)$ で分布を表し、和を積分に置きかえます。詳しくは「確率分布のまとめ」を参照。）

## 和の記号 $\sum$（シグマ）

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

「$i$ を 1 から $n$ まで動かして、ぜんぶ足す」という意味です。

### 道具 1：期待値（平均）の定義

確率変数 $X$ が値 $x_1, x_2, \ldots$ をそれぞれ確率 $P(X = x_i)$ でとるとき、

$$E[X] = \sum_i x_i \, P(X = x_i)$$

これは「**値 × その値が出る確率**」を全部足したもの、つまり**重みつきの平均**です。（連続変数のときは和が積分になります：$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$。$f$ は確率密度関数。）

### 道具 2：分散の定義

$$\mathrm{Var}(X) = E\big[(X - \mu)^2\big] \qquad (\mu = E[X])$$

「平均からのズレ $(X-\mu)$ を 2 乗して、その平均をとったもの」。**ばらつきの大きさ**を表します。2 乗するのは、プラスのズレとマイナスのズレが打ち消し合わないようにするためです。標準偏差は $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$。

### 道具 3：期待値の線形性（最重要）

$$E[aX + b] = aE[X] + b, \qquad E[X + Y] = E[X] + E[Y]$$

- 定数倍 $a$ は外に出せる、定数 $b$ はそのまま足せる。- 2 つの確率変数の和の期待値は、期待値の和（**これは $X, Y$ が独立でなくても成り立つ**）。

この 3 つの道具だけで、以下のほとんどの公式が導けます。

> **記号の約束**：観測された具体的な標本平均は小文字 $\bar{x}$、確率変数として標本ごとに揺れる標本平均は大文字 $\bar{X}$ と書き分けます（母平均 $\mu$ は未知の定数）。

---

## 1. 分散の計算公式　$\mathrm{Var}(X) = E[X^2] - (E[X])^2$

**①直感**：定義どおり「ズレの 2 乗の平均」を計算するのは面倒です。これを「2 乗の平均」から「平均の 2 乗」を引くだけ、という**計算しやすい形**に書き換えます。

**③ 導出**：$\mu = E[X]$ は**定数**であることに注意して、期待値の線形性を使います。

$$\begin{aligned}
\mathrm{Var}(X) &= E\big[(X - \mu)^2\big] \\
&= E\big[X^2 - 2\mu X + \mu^2\big] && \text{(2 乗を展開)} \\
&= E[X^2] - 2\mu\, E[X] + \mu^2 && \text{(線形性：}\mu\text{は定数なので外へ)} \\
&= E[X^2] - 2\mu \cdot \mu + \mu^2 && (E[X] = \mu) \\
&= E[X^2] - \mu^2 = E[X^2] - (E[X])^2.
\end{aligned}$$

④ **確認**：$X$ が 0,1,2,3 を確率 0.1, 0.3, 0.4, 0.2 でとるとき $E[X] = 0.3 + 0.8 + 0.6 = 1.7$、$E[X^2] = 0.3 + 1.6 + 1.8 = 3.7$。$\mathrm{Var}(X) = 3.7 - 1.7^2 = 3.7 - 2.89 = 0.81$。

---

## 2. スケール変換の公式　$\mathbf{Var}(aX+b) = a^2\,\mathbf{Var}(X)$

① **直感**：全体を $a$ 倍に引き伸ばすとばらつきは $a^2$ 倍に。**定数 $b$ を足しても全体が平行移動するだけで、ばらつきは変わらない**（これが $b$ が消える理由）。

③ **導出**：$Y = aX + b$ とおく。まず $E[Y] = aE[X] + b = a\mu + b$。

$$\begin{aligned}
\mathrm{Var}(aX+b) &= E\big[(Y - E[Y])^2\big] \\
&= E\big[(aX + b - a\mu - b)^2\big] \\
&= E\big[(a(X-\mu))^2\big] \qquad (b \text{ が打ち消す}) \\
&= E\big[a^2(X-\mu)^2\big] \\
&= a^2\,E\big[(X-\mu)^2\big] = a^2\,\mathrm{Var}(X).
\end{aligned}$$

④ **確認**：$Z$ スコア $Z = \dfrac{X-\mu}{\sigma}$ は $a = 1/\sigma,\ b = -\mu/\sigma$ の変換。だから $\mathrm{Var}(Z) = \dfrac{1}{\sigma^2}\mathrm{Var}(X) = \dfrac{\sigma^2}{\sigma^2} = 1$、$E[Z] = 0$。→ **標準化すると平均 0・標準偏差 1 になる**ことが式から確認できます。

---

## 3. 独立な和の分散　$\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$

①**直感**：期待値はいつでも足せますが、**分散が足せるのは 2 つが独立なとき**です。互いに無関係な揺れは、足しても打ち消し合いも増幅もしない、というイメージです。

③**導出のあらすじ**：$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$ という関係があり、$X, Y$ が独立なら共分散 $\mathrm{Cov}(X,Y)$（第 11 節で定義します）が 0 になるため、

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

一般に $n$ 個の独立な確率変数では

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

この「**独立な和の分散＝分散の和**」は、次の二項分布と標準誤差の導出で主役になります。

---

## 4. 二項分布の期待値と分散　$E[X]=np,\quad \mathrm{Var}(X)=np(1-p)$

コインを $n$ 回投げて表の回数 $X$ を数える ― この $X$ が従うのが**二項分布**（成功確率 $p$ のベルヌーイ試行を $n$ 回くり返したときの成功回数）です。分布の形そのものは「確率分布のまとめ」第 2・3 節で扱いますが、平均と分散だけなら、分布の式を知らなくても「1 回ごとにバラして足す」だけで求まります。

① **直感**：コインを $n$ 回投げて表の回数を数える―これを「1 回ごとの結果」に**バラして**考えます。各回は「表＝1 点／裏＝0 点」のゲーム（ベルヌーイ試行）で、これを $n$ 回足したものが $X$ です。

② **準備**：$i$ 回目の結果を $X_i$ とする（表なら $X_i=1$、裏なら $X_i=0$、表の確率 $p$）。

$$X=X_1+X_2+\cdots+X_n.$$

③ **導出**：1 回あたりの期待値と分散をまず求める：

$$E[X_i]=1\cdot p+0\cdot(1-p)=p,$$

$$E[X_i^2]=1^2\cdot p+0^2\cdot(1-p)=p\ \Rightarrow\ \mathrm{Var}(X_i)=E[X_i^2]-(E[X_i])^2=p-p^2=p(1-p).$$

あとは足し合わせる：- 期待値（線形性）：$E[X]=\sum_{i=1}^{n}E[X_i]=np$. - 分散（各回は独立 →独立な和の分散）：$\mathrm{Var}(X)=\sum_{i=1}^{n}\mathrm{Var}(X_i)=np(1-p)$.

④ **確認**：$n=10,\ p=0.3$ なら $E[X]=3$、$\mathrm{Var}(X)=10\cdot0.3\cdot0.7=2.1$。

---

## 5. 標準誤差（SEM）　$\mathrm{SEM}=\dfrac{\sigma}{\sqrt{n}}$

① **直感**：標本平均 $\bar{X}$ は、たくさんのデータを平均することで**個々の揺れがならされる**ため、1 個のデータよりばらつきが小さくなります。どれだけ小さくなるかを表すのが標準誤差です。

② **準備**：$X_1,\dots,X_n$ は独立で、それぞれ分散 $\sigma^2$。標本平均は

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

③**導出**：分散のスケール公式（第 2 節）と独立な和の分散（第 3 節）を順に使う。

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \quad (\mathrm{Var}(aX) = a^2\mathrm{Var}(X),\ a = 1/n) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \qquad \text{(独立な和の分散)} \\ &= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

よって標準偏差をとって

$$\mathrm{SD}(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \mathrm{SEM}.$$

④ **確認**：$\sigma = 15$, $n = 25$ なら SEM $= 15/\sqrt{25} = 15/5 = 3$。$n$ を 4 倍（100 人）にすると SEM は半分（$15/10 = 1.5$）に。**精度を 2 倍にするには標本を 4 倍**必要、という $\sqrt{n}$ の効果が見えます。

> ちなみに $E[\bar{X}] = \frac{1}{n}\sum E[X_i] = \frac{1}{n}\cdot n\mu = \mu$。標本平均は平均的には母平均に一致します（不偏性）。

---

## 6.「平均」はどこから来るか：二乗誤差の最小化

① **直感**：「データ＝モデル＋誤差」で、モデルを 1 つの代表値 $m$ にするとき、誤差の 2 乗和を**いちばん小さくする** $m$ が平均値です。これが「平均は最小二乗の意味で最良の代表値」という主張です。

③ **導出**：誤差 2 乗和を $m$ の関数とみる：

$$S(m) = \sum_{i=1}^{n}(x_i - m)^2.$$

$m$ で微分して 0 とおく（最小値の条件）：

$$\frac{dS}{dm} = \sum_{i=1}^{n} 2(x_i - m)\cdot(-1) = -2\sum_{i=1}^{n}(x_i - m) = 0.$$

$$\Rightarrow\ \sum_{i=1}^{n} x_i = nm\ \Rightarrow\ m = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

（微分を使わない場合：$S(m)$ は $m$ の 2 次関数で下に凸なので、頂点が最小。同じ答えになります。）

④ **補足**：誤差の **絶対値の和** $\sum|x_i - m|$ を最小にするのは**中央値**です。だから外れ値に強い代表値がほしいときは中央値を使います。

---

# 7. 最尤推定 ——「データが最も起こりやすい」パラメータを選ぶ

① **直感**：手元のデータ $D$ を最もうまく説明するパラメータ $\theta$ を選びたい。「うまく説明する」を「**その $\theta$ のもとで、いま観測したデータがいちばん出やすい**」と読みかえるのが**最尤推定（maximum likelihood estimation）**です。ここで「データの出やすさ」$p(D \mid \theta)$ は、データが**離散**なら確率、**連続**なら確率密度を指します（連続変数では 1 点をとる確率は 0 なので、確率そのものではなく密度で測ります ― 「確率分布のまとめ」§1）。この出やすさを $\theta$ の関数とみたものが**尤度（likelihood）** $L(\theta) = p(D \mid \theta)$ で、これを最大にする $\theta$ を **最尤推定値** $\hat{\theta}$ とよびます。第 1 章のベイズの定理に出てきた尤度と同じものを、ここでは $\theta$ を動かして最大化する対象として使います。

②**準備**：データが独立なら、尤度は各点の出やすさ（確率または確率密度）の**積**になります。

$$L(\theta) = \prod_{i=1}^{n} p(x_i \mid \theta).$$

積のままでは微分しにくいので、単調増加な対数をとって **対数尤度** $\ell(\theta) = \ln L(\theta) = \sum_i \ln p(x_i \mid \theta)$ に直します。対数は最大になる場所を変えないので、$\ell(\theta)$ を最大にする $\hat{\theta}$ を探せば十分です。あとは §6 と同じ「**微分して 0**」です。

③ **導出（例 1：コイン）**：表が出る確率 $\theta$ のコインを $n$ 回投げ、$k$ 回表が出たとします。尤度は $L(\theta) = \theta^k (1-\theta)^{n-k}$、対数尤度は

$$\ell(\theta) = k \ln \theta + (n-k) \ln (1-\theta).$$

$\theta$ で微分して 0 とおくと

$$\frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \;\Rightarrow\; \hat{\theta} = \frac{k}{n}.$$

最尤推定値は素直に **標本比率** です（第 4 章のコイン例で「7/10 は最尤推定値」と述べたのは、これのことです）。

③ **導出（例 2：正規分布 ＝ 最小二乗法）**：$x_1, \dots, x_n$ が平均 $\mu$・分散 $\sigma^2$ の正規分布から得られたとします。1 点の確率密度は $p(x \mid \mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\!\Big( -\dfrac{(x-\mu)^2}{2\sigma^2} \Big)$ なので、対数尤度は

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \ln (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

まず $\mu$ について最大化します。$\mu$ を含む項は末尾の $-\dfrac{1}{2\sigma^2}\sum_i (x_i-\mu)^2$ だけで、係数 $-\dfrac{1}{2\sigma^2}$ は負ですから、**$\ell$ を $\mu$ について最大にすること**は **$\sum_i (x_i-\mu)^2$ を最小にすること**とまったく同じです。これは §6 でみた二乗誤差の最小化そのもの ― したがって

$$\hat{\mu} = \bar{x}.$$

**「正規分布を仮定した最尤推定」は「最小二乗法」に一致する**のです。§6 で「平均は二乗誤差を最小にする代表値」として求めた平均値が、ここでは「正規誤差のもとで最も尤もらしいパラメータ」として同じ答えで現れます。別々の原理が同じ式に落ち合う ― これが本書でくり返し出会う「幹」の一例です。

次に $\sigma^2$ について最大化すると（$\ell$ を $\sigma^2$ で微分して 0）、

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

となり、$n$ **で割った分散**が出ます($n-1$ ではありません)。最尤推定の手続きに素直に従うと、自然にこの形になります。ただしこの $\hat{\sigma}^2$ は母分散 $\sigma^2$ を平均として**やや小さめに見積もる**（下向きの **偏り**をもつ）ことが知られています。偏差を真の母平均 $\mu$ ではなく手元の $\bar{x}$ を基準に測るぶん、平方和が小さめに出るためです。この偏りをちょうど打ち消すように、分母を $n-1$ に取り替えたものが次節の **不偏分散** です — つまり $n-1$ の側が「補正のために調整した」量で、最尤推定はあくまで素直な結果です。その導出は §8 で扱います。

④**注意**：最尤推定は頻度主義の代表的な推定法で、パラメータを「未知の定数」とみて、それを 1 点に**推定**します。一方ベイズは $\theta$ に事前分布を置き、事後分布**全体**を得ます（第 4 章）。両者は無縁ではなく、**事前分布を一様（無情報）にとると、事後分布を最大にする点（MAP 推定）は最尤推定値に一致**します。第 4 章でコインの事後の最頻値 0.7 が標本比率 $7/10$ と一致したのは、この関係の現れです。

---

## 8. なぜ分散は $n-1$ で割るのか（不偏分散）

① **直感**：標本のばらつきから母集団のばらつき $\sigma^2$ を推定したい。ところが偏差を測る基準に「真の母平均 $\mu$」ではなく「手元の標本平均 $\bar{x}$」を使うと、第 6 節のとおり $\bar{x}$ は二乗和を**最小**にする点なので、ズレが**少なめに**見積もられます。これを補正するため、$n$ ではなく $n-1$ で割って少し大きくします。以下、これを式で確かめます。

② **準備**：$X_1, \dots, X_n$ は独立で、母平均 $\mu$・母分散 $\sigma^2$ をもつとします。標本平均を $\bar{X} = \frac{1}{n}\sum_i X_i$ とおきます。証明では次の 3 つを使います。

- 期待値の線形性：$E\left[\sum_i Z_i\right] = \sum_i E[Z_i]$
- 各データの定義から：$E\left[(X_i-\mu)^2\right] = \mathrm{Var}(X_i) = \sigma^2$
- 標本平均の分散（第 5 節で導出済み）：$E\left[(\bar{X}-\mu)^2\right] = \mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$

③ **証明**：偏差平方和 $\sum_i (X_i - \bar{X})^2$ を、$\mu$ を**経由して**展開するのが鍵です。各項に $-\mu+\mu$ を差し込んで $X_i - \bar{X} = (X_i-\mu) - (\bar{X}-\mu)$ と分け、2 乗を展開します。

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}\left[(X_i-\mu)-(\bar{X}-\mu)\right]^2 \\ &= \sum_{i=1}^{n}(X_i-\mu)^2 \;-\; 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) \;+\; n(\bar{X}-\mu)^2.\end{aligned}$$

ここで真ん中の和は、$\sum_i (X_i-\mu) = \left(\sum_i X_i\right) - n\mu = n\bar{X} - n\mu = n(\bar{X}-\mu)$ なので、

$$-2(\bar{X}-\mu)\cdot n(\bar{X}-\mu) = -2n(\bar{X}-\mu)^2.$$

これを代入すると、最後の 2 項が $-2n(\bar{X}-\mu)^2 + n(\bar{X}-\mu)^2 = -n(\bar{X}-\mu)^2$ にまとまり、

$$\sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}(X_i-\mu)^2 \;-\; n(\bar{X}-\mu)^2.$$

> **ここがポイント**：基準を母平均 $\mu$ から標本平均 $\bar{X}$ に取り替えると、偏差平方和はちょうど $n(\bar{X}-\mu)^2$ だけ**小さくなる**。第 6 節「$\bar{X}$ は二乗和を最小にする点」が、この引き算として目に見える形で現れています。

両辺の期待値をとり、準備の 3 つを使います。

$$E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \sum_{i=1}^{n}\underbrace{E[(X_i-\mu)^2]}_{=\sigma^2} - n\,\underbrace{E[(\bar{X}-\mu)^2]}_{=\sigma^2/n}$$
$$= n\sigma^2 \;-\; n\cdot\frac{\sigma^2}{n} \;=\; n\sigma^2-\sigma^2 \;=\; (n-1)\,\sigma^2.$$

したがって $n-1$ で割れば期待値がちょうど $\sigma^2$ になります（これが不偏性の定義）：

$$\boxed{E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\sigma^2.}$$

もし $n$ で割っていたら期待値は $\dfrac{n-1}{n}\sigma^2$ となり、**真の値より系統的に小さく**なってしまいます。■

④ **確認**：因子 $\dfrac{n-1}{n}$ は $n=2$ で 0.5（2 倍も過小！）、$n=10$ で 0.9、$n=100$ で 0.99。**標本が小さいほど補正が効き、大きくなると $n$ と $n-1$ の差は無視できる**、という直感どおりの振る舞いです。

> 用語の整理：母分散は母集団全体の大きさ $N$ で割る（道具 2 の定義を有限母集団で書いたもの）、標本から母分散を推定する不偏分散は $n-1$ で割る。R の `var()`・`sd()` は $n-1$ 版を返します。「割る数を 1 つ減らした」のは、$\bar{X}$ を使った時点で**自由に動けるデータの数（自由度）が 1 つ減る**ことの現れ、とも説明されます（$\sum_i(X_i-\bar{X})=0$ という 1 本の縛りが入るため）。

---

## 9. 条件付き確率とベイズの定理

① **直感**：モデルから計算できるのは「原因 →結果」という順向きの確率です。しかし本当に知りたいのは、結果を見たあとで「原因は何だったか」を推し量る**逆向き**の確率。この 2 つの橋渡しをするのがベイズの定理です。

② **準備（条件付き確率の定義）**：

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

（最後の形が**乗法定理**。「両方起こる＝ B が起こり、その上で A が起こる」。）

③ **ベイズの定理の導出**：同じ $P(A\cap B)$ を 2 通りに書くと

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

両辺を $P(B)$ で割って

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

分母 $P(B)$ は、$A$ が起こる場合と起こらない場合に分けて（**全確率の定理**）

$$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}).$$

④ **確認（検査の例）**：$A$ =「病気」、$B$ =「陽性」。有病率 $P(A) = 0.01$、感度 $P(B \mid A) = 0.99$、偽陽性率 $P(B \mid \bar{A}) = 0.05$。

$$P(B) = 0.99 \times 0.01 + 0.05 \times 0.99 = 0.0099 + 0.0495 = 0.0594,$$

$$P(A \mid B) = \frac{0.0099}{0.0594} \approx 0.167\ (16.7\%).$$

感度が高くても、有病率が低いと陽性的中率は低い——この「直感とのズレ」がベイズの定理の要点です。

---

## 10. 検定統計量と信頼区間

① **直感**：標本平均 $\bar{x}$ は母平均 $\mu$ のまわりで SEM の幅で揺れます。だから「$\bar{x}$ と仮説の値 $\mu_0$ の差」を「揺れの単位（SEM）」で測れば、差が**何個分の揺れ**にあたるかがわかります。これが検定統計量です。

**検定統計量（$z$ または $t$）**：

$$z = \frac{\bar{x} - \mu_0}{\text{SEM}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

（母標準偏差 $\sigma$ が未知で標本から推定した場合は $t$ 統計量とよび、$t$ 分布を使います。）

**信頼区間（95%）**：標本平均が母平均から $\pm 1.96$ SEM の範囲に 95%の確率で入ることを逆に読むと、母平均の 95%信頼区間は

$$\bar{x} \ \pm \ 1.96 \times \text{SEM}.$$

（$\sigma$ 未知のときは 1.96 を $t$ 分布の臨界値 $t_{\text{crit}}$ に置き換える。）

④ **確認**：$\bar{x} = 165$, SEM $= 3$ のとき - 仮説 $\mu_0 = 160$ に対し $z = (165 - 160)/3 = 5/3 \approx 1.67$。$|1.67| < 1.96$ なので有意水準 5%で棄却されない。- 95%信頼区間：$165 \pm 1.96 \times 3 = [159.12,\ 170.88]$。

> **解釈の注意**：信頼区間は「同じ手続きを繰り返すと 95%の区間が母平均を含む」という意味。母平均は定数で、ランダムなのは区間の方です。

---

# 11. 相関係数と回帰係数

**共分散と相関係数**：2 変数が一緒に動く度合いを測る。

$$\mathrm{Cov}(X,Y) = E\big[(X-\mu_X)(Y-\mu_Y)\big] = E[XY] - E[X]E[Y],$$

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}.$$

$\sigma_X\sigma_Y$ で割って単位を消した量で、コーシー・シュワルツの不等式 $|\mathrm{Cov}(X,Y)| \le \sigma_X\sigma_Y$ から、必ず $-1 \le r \le 1$ に収まります。

**回帰直線の傾き**（最小二乗）：$y = a + bx$ で誤差 2 乗和 $\sum (y_i - a - bx_i)^2$ を最小にすると

$$b = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \qquad a = \bar{y} - b\bar{x}.$$

**どうやって出すか（偏微分 → 0）**：変数が $a, b$ の 2 つあるので、第 6 節の「微分して 0」を**注目する変数ごと**に行います。誤差 2 乗和 $S(a,b) = \sum_i (y_i - a - bx_i)^2$ を、$a$ と $b$ のそれぞれで**偏微分**（他方は定数とみなして微分）し、ゼロと置きます。

$$\frac{\partial S}{\partial a} = -2\sum_i (y_i - a - bx_i) = 0, \qquad \frac{\partial S}{\partial b} = -2\sum_i x_i(y_i - a - bx_i) = 0.$$

この 2 本の一次方程式（**正規方程式**）を解いたものが上の $a, b$ です。変数が何個に増えても手続きは同じ ― 「二乗和を、注目する変数で偏微分してゼロになる点を取る」。これが最小二乗法の計算のすべてです。直線は必ず点 $(\bar{x}, \bar{y})$ を通ります。

**行列でまとめる ――デザイン行列と疑似逆行列**：説明変数が何本に増えても、線形モデルは

$$\mathbf{y} = X\beta + \varepsilon$$

という 1 本の式にまとめられます。$X$ は観測を行に・説明変数を列に並べた**デザイン行列**（切片は「1 を並べた列」として入れる）、$\beta$ は係数を並べたベクトルです。誤差 2 乗和 $\|\mathbf{y} - X\beta\|^2$ を $\beta$ の各成分で偏微分してゼロと置くと、正規方程式 $X^\top X\beta = X^\top \mathbf{y}$ が得られ、その解は

$$\hat{\beta} = (X^\top X)^{-1} X^\top \mathbf{y} = X^{+}\mathbf{y}.$$

ここで $X^{+}$ は **疑似逆行列（ムーア・ペンローズ逆行列）** と呼ばれる「逆行列の代わり」です（$X$ は縦長で普通の逆行列をもたないが、列が一次独立なら $X^{+} = (X^\top X)^{-1}X^\top$）。$t$ 検定・分散分析・重回帰は、デザイン行列 $X$ の**作り方が違うだけ**で（「統計の重要ポイント」ポイント 12）、解はすべてこの 1 本の式から出ます。R の `lm()` が中でやっているのも、本質的にはこの計算です。

---

## まとめ：導出の「幹」

ほとんどの公式は、次の 3 本の幹から枝分かれします。

1. **期待値の線形性** $E[aX+b]=aE[X]+b$ → 分散の公式 $E[X^2]-(E[X])^2$、二項分布の $E[X]=np$
2. **分散のスケール公式と独立な和の分散** $\mathrm{Var}(aX)=a^2\mathrm{Var}(X)$、$\mathrm{Var}(\sum X_i)=\sum \mathrm{Var}(X_i)$ → 二項分布の $np(1-p)$、標準誤差 $\sigma/\sqrt{n}$
3. **2 乗和の最小化**（微分して 0）→ 平均、回帰直線の傾き

この 3 本さえ手が動けば、講義に出てくる公式の大半は自分で再現できます。まずは第 1・2・4・5 節を、何も見ずに最後まで書けるようになることを目指してください。「公式を覚える」のではなく「幹から導ける」状態が、統計を数理として理解したということです。

# 3. 確率分布のまとめ

この章は、講義で登場した主要な確率分布を一望し、それぞれが「何を表し」「どこでつながっているか」を数理の側から整理するためのものです。前章「公式の導き方」の期待値・分散の道具とあわせて読むと、各分布の $E[X]$・$\mathrm{Var}(X)$ が**与えられた事実ではなく導けるもの**として見えてきます。

---

## 1. 分布を表す 3 つの関数

### 確率質量関数（PMF）——離散の確率変数

「ちょうどその値をとる確率」。

$$p(x) = P(X = x), \qquad p(x) \geq 0, \qquad \sum_{x} p(x) = 1.$$

### 確率密度関数（PDF）——連続の確率変数

連続変数では 1 点をとる確率は 0 なので、「区間に入る確率＝面積」で考える。

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1, \qquad P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx.$$

$f(x)$ は確率そのものではなく **密度**。だから $f(x) > 1$ になることもある（面積が 1 ならよい）。

### 累積分布関数（CDF）——離散・連続に共通

「その値以下になる確率」。

$$F(x) = P(X \leq x).$$

連続のときは $F(x) = \int_{-\infty}^{x} f(t)\,dt$、逆に微分すると $f(x) = F'(x)$。

---

## 2. 主要な確率分布のカタログ

| 分布 | 型 | PMF / PDF | パラメータ | $E[X]$ | $\mathrm{Var}(X)$ |
|---|---|---|---|---|---|
| ベルヌーイ | 離散 | $p^x(1-p)^{1-x}\ (x=0,1)$ | $p$ | $p$ | $p(1-p)$ |
| 二項 | 離散 | $\binom{n}{k}p^k(1-p)^{n-k}$ | $n, p$ | $np$ | $np(1-p)$ |
| ポアソン | 離散 | $\dfrac{e^{-\lambda}\lambda^k}{k!}$ | $\lambda$ | $\lambda$ | $\lambda$ |
| 離散一様 | 離散 | $1/n$（値 $1,\dots,n$ が均等） | $n$ | $\dfrac{n+1}{2}$ | $\dfrac{n^2-1}{12}$ |
| 連続一様 | 連続 | $\dfrac{1}{b-a}\ (a\le x\le b)$ | $a, b$ | $\dfrac{a+b}{2}$ | $\dfrac{(b-a)^2}{12}$ |
| 正規 | 連続 | $\dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu, \sigma$ | $\mu$ | $\sigma^2$ |
| 指数 | 連続 | $\lambda e^{-\lambda x}\ (x\ge 0)$ | $\lambda$ | $1/\lambda$ | $1/\lambda^2$ |
| ベータ | 連続 | $\dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\ (0<x<1)$ | $a, b>0$ | $\dfrac{a}{a+b}$ | $\dfrac{ab}{(a+b)^2(a+b+1)}$ |
| カイ二乗 | 連続 | 標準正規の二乗和（§6 で定義） | 自由度 $m$ | $m$ | $2m$ |
| $t$ | 連続 | 標準正規 $\div\sqrt{\text{カイ二乗}/\nu}$（§6 で定義） | 自由度 $\nu$ | $0\ (\nu>1)$ | $\dfrac{\nu}{\nu-2}\ (\nu>2)$ |
| コーシー | 連続 | $\dfrac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$ | $x_0,\ \gamma>0$ | 存在しない | 存在しない |

（カイ二乗分布と $t$ 分布は、データそのものよりも**標本から計算した統計量**を表す「標本分布」です。PDF の式は複雑なので、ここでは式を書かず、定義と意味・つながりを §6 でまとめます。）

---

## 3. それぞれの分布の意味

- **ベルヌーイ分布**：「成功＝ 1 ／失敗＝ 0」の 1 回きりの試行。コイン 1 枚、Yes/No。すべての出発点。
- **二項分布**：成功確率 $p$ のベルヌーイ試行を $n$ 回**独立に繰り返した**ときの成功回数。$n$ 枚のコインの表の数。
- **ポアソン分布**：一定時間・空間に「まれな出来事」が起こる回数。1 時間の来店客数、ある区画の誤植数。二項分布で $n$ が大きく $p$ が小さい極限。
- **一様分布**：どの値も等しく起こりうる。サイコロ（離散）、$[0,1]$ の乱数（連続）。
- **正規分布**：左右対称の釣鐘型。たくさんのものを足す・平均すると現れる（中心極限定理）。
- **指数分布**：「次の出来事までの待ち時間」。ポアソンと表裏の関係。

- **ベータ分布**：0〜1 の値、つまり「割合」や「確率そのもの」を表す連続分布。2 つの形状パラメータ $a, b$ で形が柔軟に変わり、$a = b = 1$ なら一様分布になる。ベイズ統計では二項分布と相性のよい**共役事前分布**として活躍する（→「ベイズ統計の数理」第 5 節）。
- **カイ二乗分布**：標準正規をいくつか二乗して足したもの。正の値だけをとる右に裾を引いた形で、「ばらつき（分散）」を扱う検定に現れる。自由度＝足した本数。
- $t$ **分布**：母標準偏差 $\sigma$ を標本の $s$ で代用したときに現れる「正規分布のいとこ」。左右対称だが正規より裾が重く、自由度が大きいほど正規分布に近づく。平均の検定（$t$ 検定）の土台。
- **コーシー分布**：正規分布に似た左右対称の山だが、**裾が極端に重い**。そのため平均も分散も**存在しない**（定義する積分が発散する）。実は**標準**コーシー分布（$x_0 = 0, \gamma = 1$）は自由度 1 の $t$ 分布に一致する。ベイズ統計では、「めったにないが極端に大きな値」も許容する**弱情報事前分布**としてよく使われる（たとえば効果量の事前分布。→「ベイズ統計の数理」）。裾の重さの実際の違いは図 3.1 を見てください。

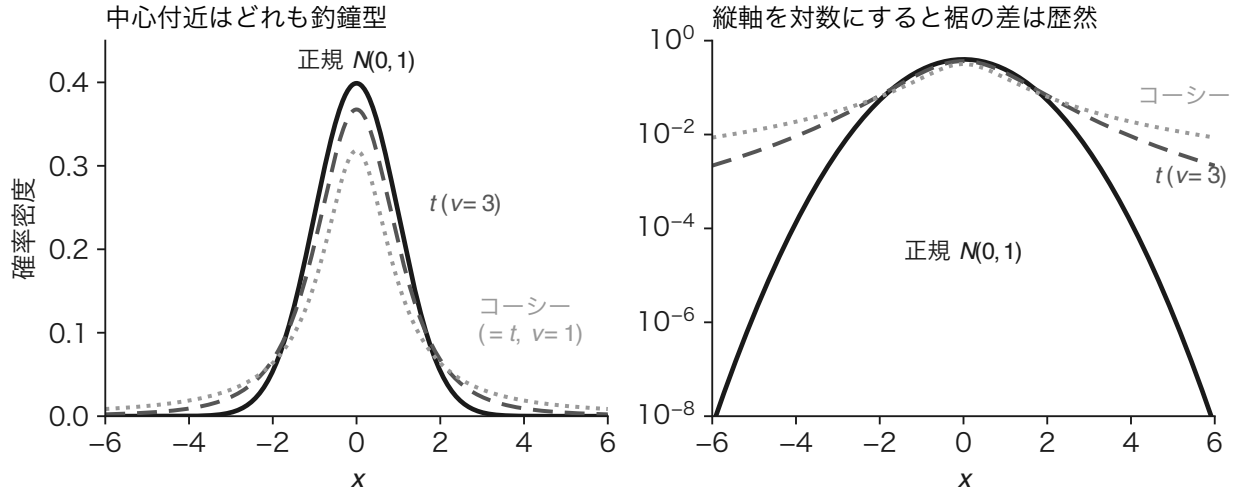



**図 3.1**：正規・$t$ ($\nu = 3$)・コーシーの確率密度。左：中心付近はどれも左右対称の釣鐘型で見分けがつきにくい。右：縦軸を対数にすると裾の差は歴然— 正規は急速に落ちるのに対し、コーシー（＝自由度 1 の $t$）は極端に裾が重く、遠く離れた値もそれなりの確率で出る。

**（導出スケッチ）二項 → ポアソン**：「$n$ が大きく $p$ が小さい極限」を式で確かめておきます。期待値 $np = \lambda$ を保ったまま（$p = \lambda/n$ として）$n \to \infty$ とすると、

$$\binom{n}{k} p^k (1-p)^{n-k} = \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\to 1} \cdot \frac{\lambda^k}{k!} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \longrightarrow \frac{e^{-\lambda}\lambda^k}{k!}.$$

二項分布の PMF が、そのままポアソン分布の PMF に流れ込みます。「まれな出来事の回数」という §3 の意味づけと、§7 の地図の矢印の裏づけです。

---

# 4. 連続分布での期待値・分散の計算（積分の練習）

定義は離散の和を積分に置きかえるだけ：

$$E[X] = \int x\, f(x)\, dx, \qquad \mathrm{Var}(X) = E[X^2] - (E[X])^2.$$

## 例：連続一様分布 $f(x) = \dfrac{1}{b-a}\ (a \le x \le b)$

**期待値**：

$$E[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

（$b^2 - a^2 = (b-a)(b+a)$ で約分。結果は「ちょうど真ん中」で直感どおり。）

**分散**：まず

$$E[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

よって

$$\mathrm{Var}(X) = E[X^2] - \left(\frac{a+b}{2}\right)^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

## 例：正規分布 $N(\mu, \sigma^2)$ の $E[X] = \mu$ と $\mathrm{Var}(X) = \sigma^2$

中心的な分布である正規分布でも、$E$ と Var は「与えられた事実」ではなく導けます。使う道具は、**ガウス積分** $\displaystyle\int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \sqrt{2\pi}$（これだけは既知とします）と置換積分です。$z = \dfrac{x-\mu}{\sigma}$（$x = \mu + \sigma z,\ dx = \sigma\, dz$）と置換し、標準正規の密度を $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ と書くと、

**期待値**：

$$E[X] = \int_{-\infty}^{\infty} x\, \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \int_{-\infty}^{\infty} (\mu + \sigma z)\, \varphi(z)\, dz = \mu \underbrace{\int \varphi(z)\, dz}_{=1} + \sigma \underbrace{\int z\, \varphi(z)\, dz}_{=0} = \mu.$$

（第 2 の積分は、$z\varphi(z)$ が原点対称の**奇関数**なので 0。「平均は対称の中心」という直感の式版です。）

**分散**：$\mathrm{Var}(X) = \sigma^2 E[Z^2]$（スケール公式）なので、$E[Z^2] = \int z^2 \varphi(z)\, dz$ を求めれば十分。$\varphi'(z) = -z\,\varphi(z)$ に気づくと、部分積分で

$$E[Z^2] = \int_{-\infty}^{\infty} z \cdot z\,\varphi(z)\, dz = \Big[-z\,\varphi(z)\Big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \varphi(z)\, dz = 0 + 1 = 1.$$

よって $\mathrm{Var}(X) = \sigma^2$。パラメータ $\mu, \sigma^2$ が期待値と分散**そのもの**であることが、定義から確かめられました。

---

# 5. 正規分布と標準化

正規分布 $N(\mu, \sigma^2)$ は平均 $\mu$ を中心とした左右対称の釣鐘型で、$\sigma$ が大きいほど横に広がる。

標準化 $Z = \dfrac{X - \mu}{\sigma}$ を行うと、平均 0・標準偏差 1 の**標準正規分布** $N(0,1)$ になる。($E[Z] = 0$、$\mathrm{Var}(Z) = \dfrac{1}{\sigma^2}\mathrm{Var}(X) = 1$。導出は「公式の導き方」第 2 節を参照。)

標準化のおかげで、単位や分布の異なるデータを共通のものさしで比較できる。正規分布では、おおよそ次が成り立つ：

| 範囲 | 含まれる割合 |
|---|---|
| $\mu \pm 1\sigma$ | 約 68% |
| $\mu \pm 2\sigma$ | 約 95% |
| $\mu \pm 3\sigma$ | 約 99.7% |

「平均 $\pm$ 約 $2\sigma$ で 95%」は、信頼区間や検定で使う 1.96 の出どころです。

この割合は、正規分布については理論から導かれる性質です。実際のデータに当てはめるときは、データがおおよそ正規分布に従うと想定できる場合、経験則として、標準偏差の何倍の範囲にデータのどれだけが収まるかを見積もることができます。

---

# 6. 標本分布の関係：正規・カイ二乗・$t$

§2〜§5 の分布は「データそのもの」を表すものでした。これに対して、**標本から計算した統計量がどんな分布に従うか**を表すのが **標本分布** です。検定や区間推定の土台になる**正規・カイ二乗・$t$** の 3 つは、別々の分布ではなく、すべて標準正規 $Z$ から作られる「一族」です。順に組み立てていきましょう。

## 6.1 出発点はいつも標準正規 $Z$

母標準偏差 $\sigma$ が **わかっている**とき、標本平均を標準化した量は標準正規になる(正規母集団からの標本なら厳密に、一般の母集団でも標本が大きければ中心極限定理により近似的に)：

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

(分母 $\sigma/\sqrt{n}$ は標本平均の標準偏差＝標準誤差。導出は「公式の導き方」を参照。)この $Z$ がすべての出発点です。

ここで暗黙に使っている土台を、一度だけ定理の形で書いておきます。

**中心極限定理(CLT)**　平均 $\mu$・分散 $\sigma^2$ をもつ同じ分布から独立に取った $X_1, \dots, X_n$ について、

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n\to\infty} N(0,1)$$

（分布の形が標準正規に近づく、という意味）。**元の分布がどんな形でも**成り立ちます。「たくさん平均すると正規が現れる」という本書のあちこちの記述は、すべてこの定理の言い換えです。

## 6.2 カイ二乗分布 —— $Z$ を二乗して足す

独立な標準正規 $Z_1, \dots, Z_m \sim N(0,1)$ の二**乗和**が従う分布を、自由度 $m$ の **カイ二乗分布** といいます。

$$\chi^2_m := Z_1^2 + Z_2^2 + \cdots + Z_m^2, \qquad E[\chi^2_m] = m, \quad \mathrm{Var}(\chi^2_m) = 2m.$$

二乗の和なので**正の値しかとらず**、右に裾を引いた非対称な形です（自由度 $m$ が大きいほど左右対称に近づく）。平均が自由度 $m$ にぴたりと等しいのは、$E[Z_i^2] = \mathrm{Var}(Z_i) = 1$ を $m$ 個足すからです。

統計では「ばらつき（分散）」を測るときに現れます。実際、正規母集団から $n$ 個を抽出したときの不偏分散 $s^2$（$n-1$ で割る版）について

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

が成り立ちます（カテゴリデータの適合度検定でカイ二乗分布が現れるのも、観測度数を標準化すると標準正規の二乗和になるためです）。

なぜ自由度が $n$ ではなく $n-1$ なのか、スケッチだけ描いておきます。「公式の導き方」第 8 節の分解

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

で、右辺第 1 項を $\sigma^2$ で割れば標準正規の二乗和 $n$ 本分（$= \chi^2_n$）、引かれる第 2 項は $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$、つまり標準正規の二乗 **1 本分**です。基準を $\bar{X}$ に取り替えると $\sum_i (X_i - \bar{X}) = 0$ という縛りが 1 本入り、自由に動ける方向が 1 つ減る ―その 1 本分がちょうど抜けて、自由度 $n-1$ のカイ二乗が残る、という勘定になっています（厳密な証明には、正規母集団で $\bar{X}$ と $s^2$ が**独立**になるという特別な事実を使います）。

## 6.3 $t$ 分布 —— $\sigma$ がわからないときの $Z$

現実には母標準偏差 $\sigma$ は普通わかりません。そこで $\sigma$ を標本標準偏差 $s$ で **置きかえる**と、標準正規からわずかにずれます。このずれを正確に表すのが $t$ 分布です。定義は「標準正規 $\div \sqrt{\text{カイ二乗}/\text{自由度}}$」：

$$t_\nu := \frac{Z}{\sqrt{\chi^2_\nu/\nu}} \qquad (Z \text{ と } \chi^2_\nu \text{ は独立}).$$

分母の $\chi^2_\nu$ は、ばらつきの推定 $s^2$ に対応します。$s$ 自体が標本ごとにばらつくぶん、$t$ は正規分布より**裾が重く**（極端な値が出やすく）なります。

**具体例：1 標本 $t$ 検定の統計量**。正規母集団を仮定すると、$\sigma$ を $s$ で置きかえた標準化量は

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

となります。§6.1 の $Z$ で分母の $\sigma$ を $s$ に替えただけ、という対応がそのまま分布に表れています。

**性質**：左右対称で平均は 0（$\nu > 1$）、分散は $\dfrac{\nu}{\nu-2}$（$\nu > 2$）で **1 より大きい**（正規より広がっている）。自由度 $\nu$ が小さいほど裾が重くなります。

### 6.4 自由度を大きくすると正規に戻る

$t$ 分布は自由度 $\nu$ を大きくすると標準正規に一致します：

$$t_\nu \xrightarrow{\nu\to\infty} N(0,1).$$

直感はこうです ——標本が大きいほど $s$ は $\sigma$ に近づき、「$\sigma$ を知っているのと変わらない」状態に近づく。だから大標本では $t$ 検定と $Z$ 検定はほぼ同じ結果になります（経験則として $\nu \gtrsim 30$ で差はごくわずか）。分散 $\dfrac{\nu}{\nu-2} \to 1$ となることからも、$t$ がだんだん $N(0,1)$ に締まっていく様子が読みとれます。
三者の関係を 1 枚の図にまとめます（図 3.2）。

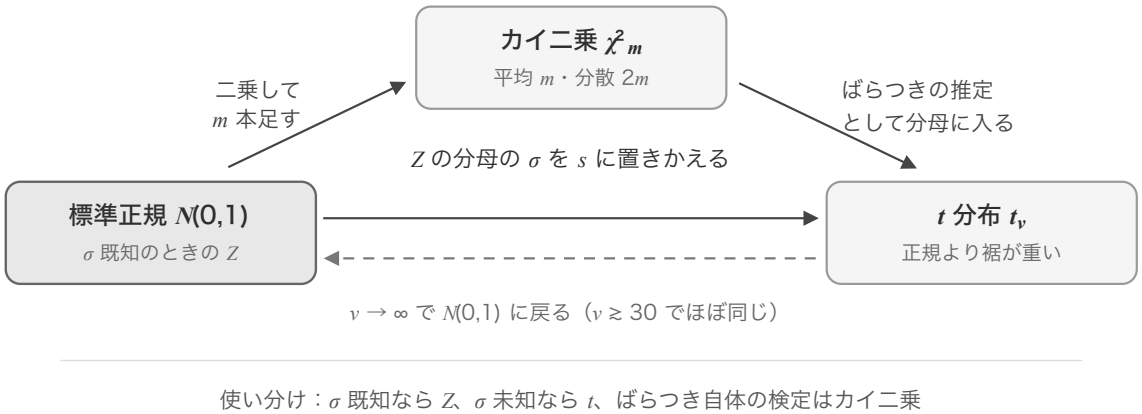



**図 3.2**：標本分布の一族。標準正規 $Z$ を二乗して $m$ 本足すとカイ二乗 $\chi^2_m$、$Z$ を $\sqrt{\chi^2_\nu/\nu}$ で割ると $t_\nu$。$t$ は自由度 $\nu$ を大きくすると標準正規に戻る。

- **正規** … 母体。$\sigma$ が既知なら統計量はそのまま $Z$。
- **カイ二乗** … 正規の「二乗和」。ばらつきそのものを扱う。
- $t$ … 正規 ÷ カイ二乗の平方根。$\sigma$ を $s$ で代用するぶんだけ裾が重い。

検定での使い分けも、この関係そのものです —— **$\sigma$ 既知なら $Z$、$\sigma$ 未知なら $t$、ばらつき自体を検定するならカイ二乗**。

---

## 7. 分布どうしのつながり（地図）

**幹は「ベルヌーイ →二項」**。そこから極限のとり方で正規分布・ポアソン分布が現れます。また二項分布の確率 $p$ を推定するベイズ更新では、その事前・事後分布として **ベータ分布**が現れます（§3、「ベイズ統計の数理」）。そして正規分布を母体に、二乗和でカイ二乗分布、$\sigma$ を $s$ で代用すると $t$ 分布 —— という**標本分布の一族**（§6）

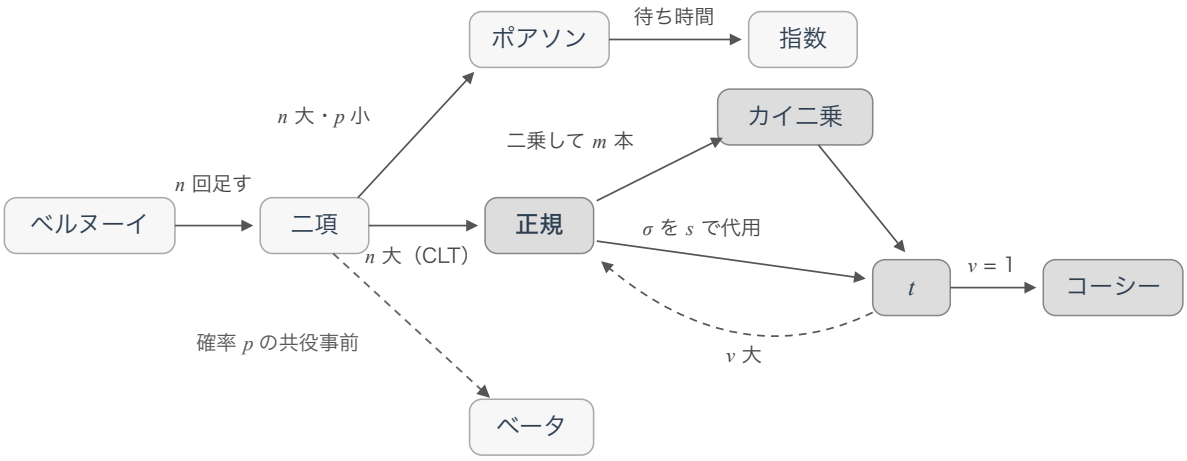



**図 3.3：**主要な確率分布のつながり。幹は「ベルヌーイ→二項」。極限のとり方で正規・ポアソンが現れ、正規を母体に標本分布の一族（濃い塗り：カイ二乗・$t$・コーシー）が枝分かれする。ベータへの点線はベイズ更新（確率 $p$ の事前 → 事後）の関係。

が枝分かれします。$t$ 分布の自由度を 1 まで下げると、裾が極端に重く平均も分散ももたない**コーシー分布**になります。期待値・分散の導出（特に二項分布の $np$, $np(1-p)$）は「公式の導き方」第 4 節を参照してください。

# 4. ベイズ統計の数理

この章は、講義「ベイズ統計」の回で扱う考え方を、数式の側からていねいに追うための読み物です。公式を覚えることが目的ではありません。**事前分布・尤度・事後分布・周辺尤度**という 4 つの登場人物が、ベイズの定理という 1 本の式の上でどう役割分担しているのかを、具体例を通して「腑に落ちる」ところまで持っていくことを目標にします。とくに、つまずきやすい**周辺尤度** に 1 節を割いて掘り下げます。

> 読み方のコツ
>
> 第 1〜3 節で枠組みを作り、第 4 節で周辺尤度を集中的に解説します。第 5 節の通し具体例（コイン投げ）で、抽象的な式が実際の数値になる様子を見てください。数値はすべて R で計算して確かめてあります。

---

## 1. 確率を「信念の度合い」として測る

これまでの講義で確率は、「同じ試行を無限に繰り返したときの相対頻度」として導入されました（**頻度主義**の立場）。ベイズ統計はもう一つの見方をとります。確率を「**ある主張をどれだけ確からしいと考えているか（信念の度合い）**」とみなすのです。

この見方の利点は、「**まだ 1 回しか起きていないこと**」「**これから起きること**」「**未知のパラメータの値**」にも確率を割り当てられる点にあります。たとえば「このコインの表が出る確率 $\theta$ は 0.6 くらいだろう」という主張に、頻度主義では確率を与えられません（$\theta$ は定数で、ランダムに動くものではないから）。ベイズでは $\theta$ 自身を確率変数とみなし、$\theta$ **についての確率分布**を考えます。

ベイズ統計のすべては、次の 1 点に集約されます。

> **データを見る前の信念（事前分布）を、観測したデータによって更新し、データを見た後の信念（事後分布）にする。**

この「更新」を実行する装置が、ベイズの定理です。

---

## 2. ベイズの定理 —— 「逆向き」の確率

「公式の導き方」第 9 節で、ベイズの定理を導きました（条件付き確率の定義から 2 行で出ます）。事象の形では

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

ここで本質的なのは **向きの反転**です。ふつう私たちが模型（モデル）から計算できるのは「原因 $A$ が与えられたとき結果 $B$ が出る確率」$P(B \mid A)$ の方です。ところが本当に知りたいのは、結果 $B$ を観測した後で「原因が $A$ だった確率」$P(A \mid B)$ ― つまり **逆向き**の確率です。ベイズの定理は、計算しやすい順向きの確率から、知りたい逆向きの確率へ橋を架けます。

統計では「原因」がパラメータ $\theta$、「結果」が観測データ $D$ にあたります。記号を置き換えると次の形になります。これがこの章の主役の式です。

$$\underbrace{p(\theta \mid D)}_{\text{事後分布}} = \frac{\overbrace{p(D \mid \theta)}^{\text{尤度}} \; \overbrace{p(\theta)}^{\text{事前分布}}}{\underbrace{p(D)}_{\text{周辺尤度}}}$$

---

## 3. 4 人の登場人物

上の式に出てくる 4 つの量を、役割とともに整理します。$\theta$ は推定したいパラメータ（例：コインの表確率）、$D$ は観測データです。

- **事前分布** $p(\theta)$（**prior**）データを見る **前**の、$\theta$ についての信念。「たぶん 0.5 あたりだろう」「まったく見当がつかない（一様）」などを分布の形で表す。
- **尤度** $p(D \mid \theta)$（**likelihood**）パラメータの値を $\theta$ と **仮に決めたとき**、手元のデータ $D$ がどれだけ出やすいか。$\theta$ を変数とみた「データの出やすさ」。これは確率モデルそのものから計算できる。
- **事後分布** $p(\theta \mid D)$（**posterior**）データを見た**後** の、更新された信念。ベイズ推定の**最終的な答え**。点ではなく分布で得られるのが特徴。
- **周辺尤度** $p(D)$（**marginal likelihood / evidence**）データ $D$ が「（特定の $\theta$ ではなく）このモデルのもとで」どれだけ出やすかったかを表す数。次節で集中的に扱う。

分母の $p(D)$ は $\theta$ を含みません（$\theta$ について平均してしまった後だから）。そのため、$\theta$ の関数として事後分布の **形**を見るときには定数として扱え、しばしば次の比例関係で書かれます。

$$\underbrace{p(\theta \mid D)}_{\text{事後}} \;\propto\; \underbrace{p(D \mid \theta)}_{\text{尤度}} \; \underbrace{p(\theta)}_{\text{事前}}. \qquad (\text{「事後} \;\propto\; \text{尤度} \times \text{事前」})$$

「**事後分布は、尤度と事前分布のかけ算（を規格化したもの）**」― これがベイズ更新の一言まとめです。事前の信念に、データという証拠を「掛け合わせて」修正している、と読めます。

---

## 4. 周辺尤度をていねいに

周辺尤度 $p(D)$ は、ベイズの式の中でいちばん分かりにくい量です。ここを掘り下げます。

## 4.1 定義 ——パラメータについて「平均」する

周辺尤度は、尤度を事前分布で **重みづけ平均** したものです。

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta \qquad (\theta \text{ が離散なら } p(D) = \textstyle\sum_{\theta} p(D \mid \theta)\, p(\theta)).$$

これは「公式の導き方」第 9 節の **全確率の定理** $P(B) = \sum_A P(B \mid A)P(A)$ の連続版にほかなりません。「ありうるすべての $\theta$ について、その $\theta$ の起こりやすさ $p(\theta)$ をかけて、データの出やすさ $p(D \mid \theta)$ を足し上げる」操作です。

「周辺(marginal)」という名前は、同時分布 $p(D, \theta) = p(D \mid \theta)p(\theta)$ から $\theta$ を**積分して消し**、$D$ だけの分布にする — つまり $\theta$ を「周辺化(marginalize)して追い出す」ことから来ています。

## 4.2 役割その 1 ——規格化定数

事後分布は確率分布ですから、$\theta$ について足し合わせると 1 にならなければなりません。

$$\int p(\theta \mid D)\, d\theta = 1.$$

ベイズの式の分子 $p(D \mid \theta)p(\theta)$ だけでは、$\theta$ について積分しても 1 になりません。そこで**分子の総和ぴったりで割って、合計を 1 に合わせる**必要があります。その「割るべき総量」こそ

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta$$

です。つまり周辺尤度の第一の役割は、**事後分布をきちんとした確率分布にするための規格化定数**です。「事後 $\propto$ 尤度 $\times$ 事前」の比例を等式に直すために、分母として必ず現れます。

## 4.3 役割その 2 ——モデルの「データ予測力」の通信簿

規格化のためだけなら $p(D)$ は裏方ですが、$p(D)$ **そのものに意味があります**。$p(D)$ は

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta = \text{「このモデルが、観測される前のデータ } D \text{ に与えていた確率」}$$

です。特定の都合のよい $\theta$ ではなく、**事前分布が予想していた $\theta$ の範囲ぜんぶをならした上で**、現実のデータがどれだけ「ありそう」だったかを 1 つの数で表します。だから周辺尤度は別名**エビデンス(evidence, 証拠)**と呼ばれ、**そのモデルがデータをどれだけうまく予言できたか**の通信簿になります。

ここで自動的に効くのが「**複雑なモデルは自分で自分の首を絞める**」という性質です。周辺尤度 $p(D)$ は、ありうるデータ $D$ 全体について足すと 1 になる確率分布です($\sum_D p(D) = 1$)。そのため、事前分布が $\theta$ を広く取りすぎた(過度に複雑な・なんでも説明できる)モデルは、多種多様なデータを「ありうる」と予言することになり、限られた確率を多くのデータに薄く配るため、実際に観測された 1 つのデータに割り当てる $p(D)$ は小さくなりがちです。逆に、ちょうどよい範囲に賭けていたモデルは $p(D)$ が大きくなります。これは**オッカムの剃刀(無駄に複雑な説明を罰する)**が、追加のルールではなく周辺尤度の定義から自然に出てくる、という美しい事実です。

## 4.4 役割その 3 ——モデル比較とベイズファクター

2 つのモデル $M_1, M_0$ を比べたいとします。それぞれの周辺尤度 $p(D \mid M_1),\ p(D \mid M_0)$ の比を **ベイズファクター（Bayes factor）** と呼びます。

$$\mathrm{BF}_{10} = \frac{p(D \mid M_1)}{p(D \mid M_0)} = \frac{M_1 \text{ のもとでデータが出る確率}}{M_0 \text{ のもとでデータが出る確率}}.$$

$\mathrm{BF}_{10} > 1$ ならデータは $M_1$ を支持し、$< 1$ なら $M_0$ を支持します。仮説検定の $p$ 値が「帰無仮説が正しいと仮定したときの極端さ」しか測れないのに対し、ベイズファクターは **2 つの仮説のどちらがデータをうまく説明したかを直接くらべられる** のが長所です。第 5 節で実際に計算します。

## 4.5 ベイズファクターの手軽な近似 ——BIC

ベイズファクターを正面から求めるには周辺尤度 $p(D) = \int p(D \mid \theta) p(\theta)\, d\theta$ が要りますが、第 7 節で見るとおり、この積分は一般には解けません。そこで実務で広く使われる近似が **BIC（ベイズ情報量規準, Bayesian Information Criterion）** です。**最尤推定値** $\hat{\theta}$（尤度 $L(\theta) = p(D \mid \theta)$ を最大にするパラメータ値。「公式の導き方」第 7 節）での尤度 $L(\hat{\theta}) = p(D \mid \hat{\theta})$、パラメータ数 $k$、標本サイズ $n$ から

$$\mathrm{BIC} = -2 \ln L(\hat{\theta}) + k \ln n$$

で定義します。第 1 項はデータへの当てはまり（小さいほどよく当てはまる）、第 2 項 $k \ln n$ は **パラメータが多いほど重くなる罰則** です。この罰則こそ、§4.3 でみた「複雑なモデルは自分で自分の首を絞める」というオッカムの剃刀を、式の形で取り出したものになっています。

BIC の名前に「ベイズ」が入るのは偶然ではありません。標本サイズ $n$ が大きいとき

$$\mathrm{BIC} \ \approx\ -2 \ln p(D)$$

が成り立ちます（シュワルツ近似。定数項を除いた大標本での近似です）。つまり **BIC は「周辺尤度を $-2\ln$ した量」の近似** です。これを 2 つのモデルで差し引くと、ベイズファクターとの関係が現れます。

$$\mathrm{BIC}_0 - \mathrm{BIC}_1 \ \approx\ 2 \ln \frac{p(D \mid M_1)}{p(D \mid M_0)} = 2 \ln \mathrm{BF}_{10}, \qquad \text{すなわち} \qquad \mathrm{BF}_{10} \ \approx\ \exp\!\Big(\tfrac{\mathrm{BIC}_0 - \mathrm{BIC}_1}{2}\Big).$$

**BIC が小さいモデルほど支持される** わけで、その差がそのまま近似的なベイズファクターに翻訳できます。周辺尤度の積分を解かずに、最尤推定とパラメータ数・標本サイズだけからモデルを比べられる — これが BIC の手軽さです（ただしあくまで大標本での近似で、事前分布の情報を $k \ln n$ という大まかな罰則で置きかえている点には注意）。

> **まとめ**：周辺尤度は「規格化定数」という裏方の顔と、「モデルの予測力＝エビデンス」という主役の顔を併せ持つ。後者がモデル比較・ベイズファクターの土台になり、その手軽な近似が **BIC**（$\mathrm{BF}_{10} \approx e^{(\mathrm{BIC}_0 - \mathrm{BIC}_1)/2}$）である。

---

# 5. 通し具体例：コイン投げ（ベータ – 二項モデル）

抽象論を、計算しきれる具体例で確かめます。コインを $n$ 回投げて表が $k$ 回出たとき、表の確率 $\theta$ を推定します。

## 5.1 尤度（モデル）

表が出る確率を $\theta$ とすると、$n$ 回中 $k$ 回表が出る確率は二項分布です。

$$p(D \mid \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

## 5.2 事前分布

$\theta$ は $[0,1]$ の値なので、$[0,1]$ 上の分布である**ベータ分布** を事前分布に選ぶと便利です。

$$p(\theta) = \frac{1}{B(a,b)}\, \theta^{a-1}(1-\theta)^{b-1}, \qquad 0 \le \theta \le 1.$$

ここで $B(a,b) = \displaystyle\int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta$ は規格化のための **ベータ関数** です。$a = b = 1$ とすると $p(\theta) = 1$、すなわち **一様分布（まったく見当がつかない）** になります。$a = b$ を大きくすると「0.5 付近だろう」という強い事前の信念を表せます（一般には事前分布の平均 $a/(a+b)$ の付近に集中します）。

## 5.3 事後分布 ——「共役」の威力

「事後 $\propto$ 尤度 $\times$ 事前」を計算します。$\theta$ に関係しない定数（$\binom{n}{k}$ や $1/B(a,b)$）はまとめて比例記号に吸収します。

$$\begin{aligned} p(\theta \mid D) &\propto \underbrace{\theta^k (1-\theta)^{n-k}}_{\text{尤度}} \cdot \underbrace{\theta^{a-1}(1-\theta)^{b-1}}_{\text{事前}} \\ &= \theta^{(a+k)-1}(1-\theta)^{(b+n-k)-1}. \end{aligned}$$

最後の式は、パラメータが $a' = a + k,\ b' = b + n - k$ の**ベータ分布の形そのもの** です。つまり

$$\boxed{\theta \mid D \ \sim\ \mathrm{Beta}(a+k,\ b+n-k).}$$

事前がベータなら事後もベータ — このように**事前と事後が同じ分布族になる** 関係を **共役（conjugate）** といいます。共役のときは積分を実行しなくても、パラメータの足し算だけで事後分布が分かります。直感的には、**事前のベータ分布 Beta$(a,b)$ は「あらかじめ表 $a-1$ 回・裏 $b-1$ 回を見たことにする」仮想データ** に相当し、そこに本物のデータ（表 $k$・裏 $n-k$）を足し込んでいる、と読めます。

### 5.4 周辺尤度を実際に計算する

第 4 節の定義どおり、周辺尤度を積分で求めます。$\theta$ に依存する部分だけが積分に残ります。

$$
\begin{aligned}
p(D) &= \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} \cdot \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta \\
&= \frac{\binom{n}{k}}{B(a,b)} \int_0^1 \theta^{(a+k)-1}(1-\theta)^{(b+n-k)-1}\, d\theta \\
&= \binom{n}{k} \frac{B(a+k,\, b+n-k)}{B(a,b)}.
\end{aligned}
$$

（2 行目から 3 行目で、ベータ関数の定義 $\int_0^1 \theta^{a'-1}(1-\theta)^{b'-1}d\theta = B(a',b')$ をそのまま使いました。）この閉じた式を**ベータ – 二項分布の周辺尤度**と呼びます。積分が手で実行できる、数少ない気持ちのよい例です。

### 5.5 数値で確かめる（$n=10,\ k=7$、一様事前 $a=b=1$）

事前を一様 $\mathrm{Beta}(1,1)$ とし、10 回中 7 回表が出たとします。事後は

$$
\theta \mid D \ \sim\ \mathrm{Beta}(1+7,\ 1+3) = \mathrm{Beta}(8,4).
$$

| 量 | 式 | 値 |
|---|---|---|
| 事後平均 | $\dfrac{a'}{a'+b'} = \dfrac{8}{12}$ | $0.667$ |
| 事後最頻値（MAP） | $\dfrac{a'-1}{a'+b'-2} = \dfrac{7}{10}$ | $0.700$ |
| 95% 信用区間 | $\theta$ の事後分布の中央 95% | $[0.390,\ 0.891]$ |
| 周辺尤度 | $\binom{10}{7} B(8,4)/B(1,1)$ | $0.0909\ (=\frac{1}{11})$ |

事後分布の形は 図 4.1 のとおりです。最頻値 0.7 は単純な比 $7/10$（最尤推定値）と一致します。一様な事前は何も主張しないので、答えがデータだけで決まるためです。一方、事後**平均** は 0.667 とわずかに 0.5 寄りになります。これは一様事前が「0.5 付近もまだ十分ありうる」と言い続けているぶん、平均が中央へ引き戻されるからです（**事前による正則化**）。

> **周辺尤度の検算**：一様事前のとき、$p(D)$ は $k$ の値によらず常に $1/(n+1)$ になります。実際 $1/(10+1) = 1/11 = 0.0909$ で表の値と一致。意味はこうです —「表の確率がどれも等しくありそう」と思っているなら、$n$ 回中の表の回数 $k = 0, 1, \ldots, n$ もどれも等しくありそう（全 $n+1$ 通りが等確率）。だから各 $k$ の予測確率は $1/(n+1)$。**周辺尤度の式が、この直感とぴったり合う**ことが確認できます。

### 5.6 信用区間 ≠信頼区間（重要な対比）

上の **95% 信用区間（credible interval）** $[0.390,\ 0.891]$ は、文字どおり

$$
P(0.390 \le \theta \le 0.891 \mid D) = 0.95
$$

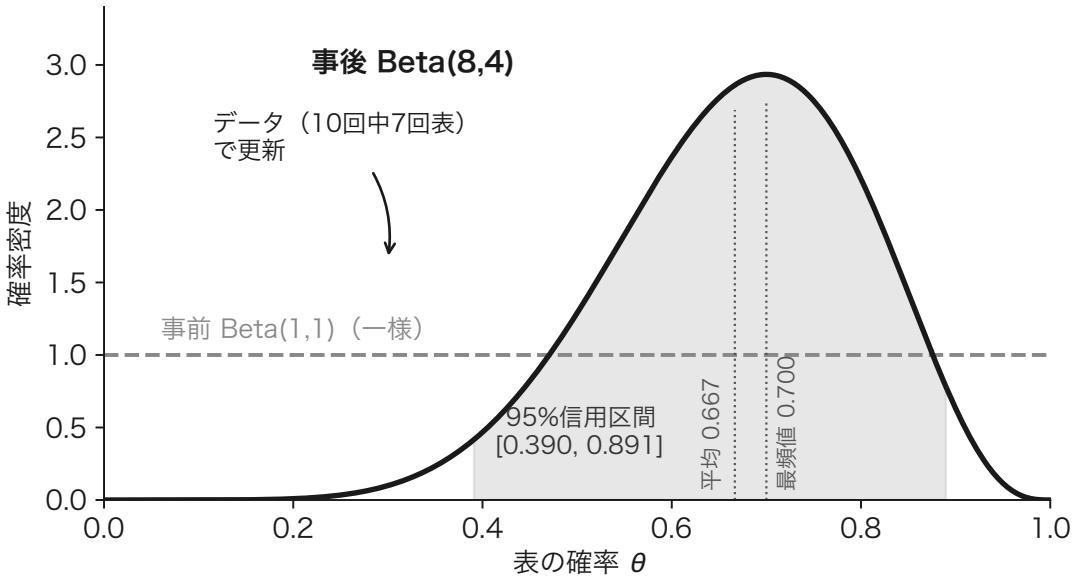



**図 4.1：**ベイズ更新の通し例（$n = 10,\ k = 7$）。一様な事前 Beta(1,1)（点線）が、データによって事後 Beta(8,4)（実線）に更新される。網かけの帯は 95%信用区間 $[0.390,\ 0.891]$、縦の点線は事後平均 0.667 と最頻値 0.700。

と読めます ―「$\theta$ **がこの区間に入る確率が 95%**」（もちろん、事前分布とデータを与えたうえでの確率です）。これはほとんどの人が「信頼区間」に対して期待してしまう（が、頻度主義の信頼区間では成り立たない）解釈です。「公式の導き方」第 10 節で注意したとおり、頻度主義の**信頼区間**は「同じ手続きを繰り返すと 95%の区間が母数を含む」という、区間の側がランダムだという解釈でした。**ベイズでは $\theta$ の側に確率を置く**ので、素直な「$\theta$ がこの中にいる確率」という言い方が許されます。この解釈の自然さがベイズの大きな魅力です。

## 5.7 ベイズファクターでモデルを比べる

「このコインは公正か？」を 2 つのモデルの勝負として定式化します。

- $M_0$（公正）：$\theta = 0.5$ に決め打ち。$p(D \mid M_0) = \binom{10}{7} 0.5^{10} = 0.117.$
- $M_1$（不明）：$\theta \sim \text{Beta}(1, 1)$ の一様事前。$p(D \mid M_1) = 1/11 = 0.0909.$

ベイズファクターは

$$\text{BF}_{10} = \frac{p(D \mid M_1)}{p(D \mid M_0)} = \frac{0.0909}{0.117} = 0.78, \qquad \text{BF}_{01} = \frac{1}{0.78} = 1.29.$$

$\text{BF}_{01} = 1.29 > 1$ なので、**データはむしろ「公正なコイン」をわずかに支持**しています。10 回中 7 回の表は、直感的には「ちょっと表が多い」ように見えても、**公正さを疑う証拠としては弱い**のです。両側検定の $p$ 値もこの例では約 0.34 で「公正さを**棄却できない**」とは言えますが、$p$ 値は原理的に、帰無仮説を**支持する**証拠を与えることができません。周辺尤度の比は、公正さを積極的に支持する程度まで一目で示してくれます。

> **発展：ベイズファクターは事前分布に敏感**　$p(D \mid M_1)$ は尤度を**事前分布で平均した**量なので、$M_1$ の事前を広げるほど（ありそうもない $\theta$ にまで確率を薄く配るほど）$p(D \mid M_1)$ は小さくなり、ベイズファクターは $M_0$ 寄りに動きます。逆に $M_1$ の事前を 0.5 の近くに集中させれば、同じデータでも結論は変わりえます。これは §4.3 のオッカムの剃刀の裏面で、欠陥ではなく仕様ですが、だからこそベイズファクターを報告するときは**どんな事前分布を使ったかを明示し、事前を変えても結論が保つかを確かめる**のが作法です。

### 5.8 次の 1 回を当てる ―事後予測分布

学習の成果は「次はどうなるか」に使えて初めて意味をもちます。事後分布 $\mathrm{Beta}(a+k,\, b+n-k)$ を手にした今、**次の 1 投が表になる確率**は、$\theta$ の不確かさを事後分布で平均して

$$P(\text{次が表} \mid D) = \int_0^1 \theta\, p(\theta \mid D)\, d\theta = E[\theta \mid D] = \frac{a+k}{a+b+n}$$

と求まります。これが **事後予測分布（posterior predictive distribution）**です（表/裏の 2 値なので、この 1 つの数が分布を決めます）。数値例では $\frac{1+7}{2+10} = 0.667$。一様事前（$a=b=1$）のときの $\frac{k+1}{n+2}$ は **ラプラスの継承則**として古くから知られる式です。周辺尤度が「データを見る**前**の予測力」だったのに対し、こちらは「学習した**後**の予測」―この対比は「機械学習の初歩」第 7 節（AIC と BIC）でふたたび主役になります。

---

## 6. 頻度主義との対比（まとめ）

| 観点 | 頻度主義 | ベイズ |
|---|---|---|
| 確率の意味 | 長期的な相対頻度 | 信念の度合い |
| パラメータ $\theta$ | 未知だが **定数** | **確率変数**（分布をもつ） |
| 答えの形 | 点推定＋信頼区間／$p$ 値 | **事後分布**（そこから平均・区間など） |
| 事前情報 | 原則使わない | **事前分布**として明示的に使う |
| 区間の解釈 | 区間がランダム（95%の区間が母数を含む） | $\theta$ がランダム（$\theta$ **が区間に入る確率 95%**） |
| 仮説の比較 | $p$ 値（帰無仮説基準のみ） | **ベイズファクター**（2 仮説を直接比較） |

どちらが正しい・優れているという話ではありません。問いの立て方と、手元にある事前情報をどう扱いたいかによって使い分けます。

---

# 7. なぜベイズは「計算が大変」と言われるのか

第 5 節のコインの例では、共役のおかげで事後分布も周辺尤度も **手で**求まりました。しかし一般のモデルでは、周辺尤度

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta$$

の積分が**解析的に解けません**。とくに $\theta$ が多次元（パラメータがたくさん）になると、この積分は高次元積分となり、まともに計算できなくなります。これがベイズ統計の最大の技術的困難です。

そこで実際には、事後分布から **標本を生成して近似する**方法（**マルコフ連鎖モンテカルロ法, MCMC** など）が使われます。ここでいう「標本」は、これまでの章での標本（母集団から抽出した**データ**の集まり）とは意味が違う点に注意してください。生成するのは**パラメータ $\theta$ の値**です。ベイズでは $\theta$ 自身が分布（事後分布）をもつので、その分布から $\theta$ の値を何千個も引き、ヒストグラムを描くようにして事後分布の形・平均・区間を近似します—データの標本ではなく、**パラメータの標本**です。第 7 回「リサンプリングとシミュレーション」で学んだ「難しい計算を、乱数を使った計算で置きかえる」という発想の、強力な延長線上にあります。規格化定数 $p(D)$ を直接求めずに事後分布の形だけをサンプリングで再現できる、という点が MCMC の巧妙なところです（詳細は本講義の範囲を超えます）。

---

# まとめ：この章の幹

1. ベイズの定理は順向きの確率（尤度）から逆向きの確率（事後）への橋。統計では「**事後** $\propto$ **尤度** $\times$ **事前**」。
2. 登場人物は 4 人：**事前・尤度・事後・周辺尤度**。
3. **周辺尤度** $p(D) = \int p(D \mid \theta) p(\theta)\, d\theta$ は、(i) 事後を確率分布にする **規格化定数**、(ii) モデルの**予測力＝エビデンス**、(iii) **ベイズファクター**によるモデル比較の土台、という三役を担う。大標本では $-2 \ln p(D)$ を **BIC** で近似でき、ベイズファクターは BIC の差から手軽に見積もれる。
4. 共役（ベータ – 二項）なら、事後はパラメータの足し算 $\mathrm{Beta}(a+k,\, b+n-k)$、周辺尤度はベータ関数の比で**手計算できる**。
5. ベイズの区間（**信用区間**）は「$\theta$ がそこに入る確率」と素直に読める。頻度主義の信頼区間との違いを押さえること。
6. 一般には周辺尤度の積分が解けない。事後分布は**シミュレーション（MCMC）**で近似でき、周辺尤度そのものの評価にはブリッジサンプリング等の追加の工夫が要る。

# Part II.

# 発展的トピック —— 進んだ分析手法

# 5. 因果推論の初歩 ——どの変数を「調整」すべきか

> **発展的な話題**　この章は講義の中心から一歩進んだ内容です。1 回生のうちは細部にこだわらず「考え方の地図」をつかむことを目標に、必要になったら戻ってきてください。

「関係のモデリング」「線形モデル」の回で、線形回帰を使えば複数の変数の影響を同時に扱えることを学びました。すると自然にこう思いたくなります—「**手に入る変数は全部モデルに入れておけば安心だろう**」。しかし因果関係を読みたいとき、これは正しくありません。**入れるべき変数と、入れてはいけない変数がある**のです。この章では、線形モデルでの**共変量調整（covariate adjustment）**を題材に、「どの変数を含め、どれを外すか」を**因果グラフ（DAG）**の言葉と簡単な数式で整理します。

> 読み方のコツ
>
> 第 3 節（交絡は入れる）と第 4 節（コライダーは入れない）が核心です。両方とも「線形モデルの係数がどうずれるか」を手計算で確かめます。第 6 節の表が結論の早見表です。

---

## 1. 相関は因果ではない ——そして、その先へ

「アイスの売上 $X$ と水難事故 $Y$ には正の相関がある」。しかしアイスが事故を**起こす**わけではありません。背後に「気温 $C$」という共通の原因があり、暑い日はアイスも売れるし、泳ぐ人も増えて事故も増える。$X$ と $Y$ の相関は $C$ が作った**見かけ**です。

観察データから「$X$ を変えたら $Y$ はどう変わるか」という**因果効果**を知りたいとき、こうした邪魔者を取り除く操作を**調整（adjustment）**または**統制（control）**と呼びます。

**「調整する」とは、線形モデルで何をすることか**：言葉を先にはっきりさせておきます。$Y$ を $X$ で説明する線形モデル

$$Y = \beta_0 + \beta X + \varepsilon$$

に、共変量 $C$ を**説明変数として追加**したモデル

$$Y = \beta_0 + \beta X + \gamma C + \varepsilon$$

を当てはめる — 線形モデルで「$C$ で調整する（$C$ を統制する）」とは、この**追加**のことです。追加すると、$X$ の係数 $\beta$ の意味が変わります。重回帰の $\beta$ は**偏回帰係数**、すなわち「**$C$ の値を同じに保ったまま、$X$ が 1 増えたときの $Y$ の変化**」を表すようになる。つまり $C$ を回帰に入れることは、「$C$ が同じ個体どうしを比べる（$C$ で

層別して比べる)」ことをモデルの上で一挙に行うのと同じで、これによって $C$ 経由の影響を取り除いた $X$ の効果が読めるのです(式での確認は第 3 節で行います)。

問題は、

> **どの変数を回帰に加えるべきか？そして、加えてはいけない変数はどれか？**

この問いに、変数の「足し算」ではなく**因果の構造**から答えるのが因果推論です。まずは全体像を 1 枚の図で示します(図 5.1)。用語は次節以降で説明します。

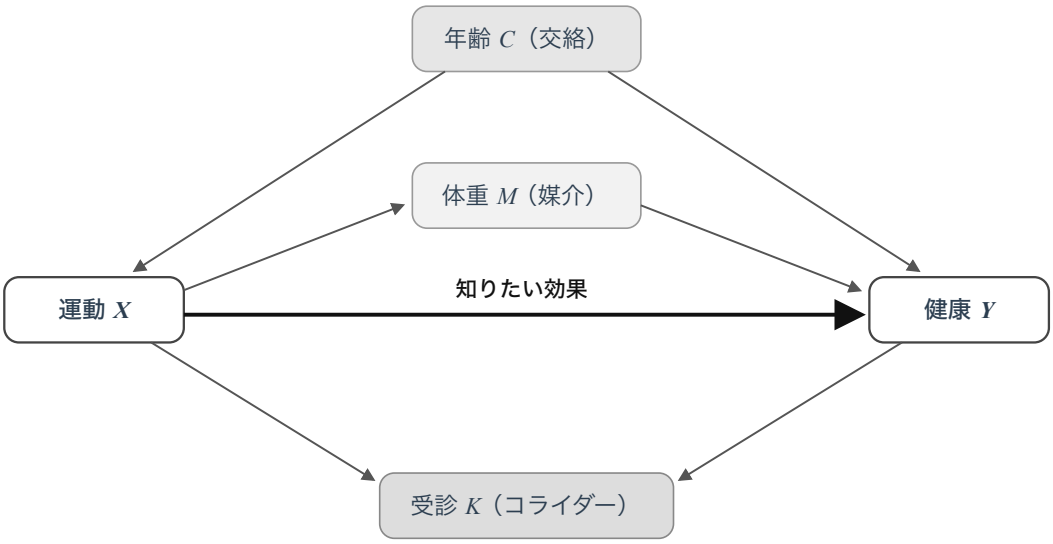



**図 5.1：**この章の全体像。運動 $X$ が健康 $Y$ に与える因果効果(太い矢印)を知りたい。交絡(年齢)は調整に入れ、媒介(体重)とコライダー(受診の有無)は入れない— これがこの章の結論。

> **図の見方**：**交絡**（共通原因。調整に**入れる**）／**媒介**（中間。全効果を見るなら**入れない**）／**コライダー**（共通結果。**入れない**）。なぜそうなるのかを、以下で順に確かめます。

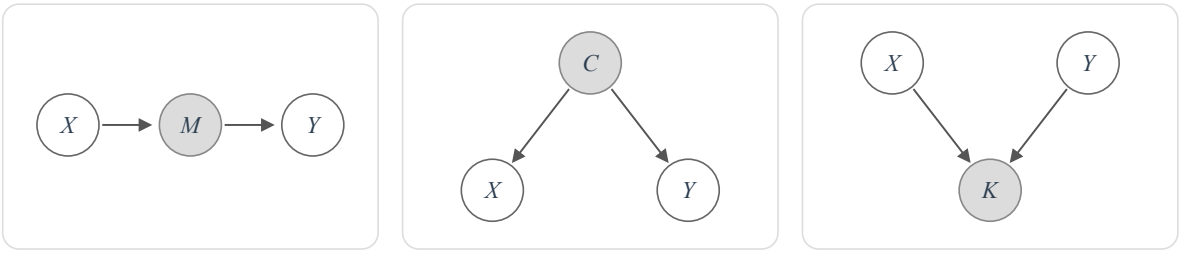
---

## 2. 因果グラフ(DAG)の言葉

変数を点(ノード)、直接の因果関係を矢印で表した図を**因果グラフ**、とくに矢印が一方向で循環しないものを**DAG（directed acyclic graph、有向非巡回グラフ）**と呼びます。矢印 $A \to B$ は「$A$ が $B$ の直接の原因」を意味します。

変数 3 つの並び方には、本質的に次の 3 パターンしかありません(図 5.2)。

- **(a) チェーン**：$M$ は $X$ の効果を $Y$ に伝える「中継ぎ」(媒介)。
- **(b) フォーク**：$C$ は $X$ と $Y$ の「共通の原因」(交絡)。
- **(c) コライダー**：$K$ は $X$ と $Y$ の「共通の結果」(合流点)。




(a) チェーン（媒介）　(b) フォーク（交絡）　(c) コライダー（合流点）

**図 5.2：**変数 3 つの 3 つの基本構造。同じ「$X$ とも $Y$ ともつながる変数」でも、中間にある変数（$M$・$C$・$K$）の役割はまったく違う。

$X$ から $Y$ へ、矢印の向きを無視してたどれる道を **パス**といい、とくに $X$ の原因をさかのぼって $Y$ にいたる「裏口」を**バックドアパス**（例：$X \leftarrow C \rightarrow Y$）と呼びます。因果効果を歪めるのは、このバックドアパスを通じて流れ込む関連です。

ポイントは、ある変数で**条件付ける（＝回帰に入れる・値を固定する）**と、パスの「通り道」が変わることです。

- **チェーン・フォークの中間変数**（$M$ や $C$）で条件付けると、そのパスは**ふさがる（ブロックされる）**。
- **コライダー**（$K$）は逆で、**何もしなければ閉じている**パスが、$K$ で条件付けると **開いてしまう**。

この非対称性が、以下のすべての判断の根拠になります。

---

## 3. 交絡（フォーク）は調整する ——入れないとバイアス

共通原因 $C$（フォーク $X \leftarrow C \rightarrow Y$）は**交絡因子（confounder）**です。これはモデルに**入れるべき**変数の代表で、入れないとバイアスが出ます（図 5.3）。線形モデルで確かめましょう。

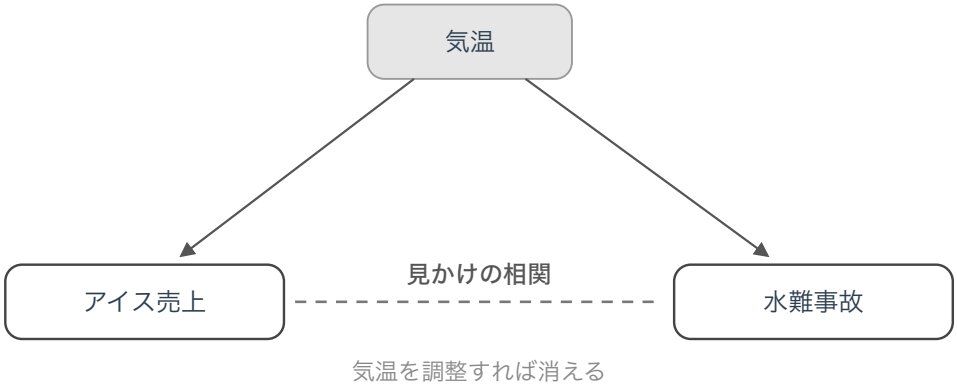



**図 5.3：**交絡の例。気温という共通原因が、アイス売上と水難事故の両方を押し上げる。両者の相関（点線）は気温が作った見かけで、因果ではない。気温を調整すれば消える。

真の構造を次とします（$\varepsilon$ は $X, C$ と無相関のノイズ）。

$$Y = \beta X + \gamma C + \varepsilon, \qquad \gamma \neq 0.$$

$C$ は $X$ の原因でもあるので、観察データでは $X$ と $C$ は相関します（$\mathrm{Cov}(X, C) \neq 0$）。いま $C$ を**省いて** $Y$ を $X$ だけに回帰すると、最小二乗法が与える係数は

$$\hat{\beta}_{\text{naive}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

分子を展開します。

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X,\ \beta X + \gamma C + \varepsilon) = \beta\,\mathrm{Var}(X) + \gamma\,\mathrm{Cov}(X, C).$$

したがって

$$\boxed{\hat{\beta}_{\text{naive}} = \beta + \gamma\,\frac{\mathrm{Cov}(X, C)}{\mathrm{Var}(X)}}$$

第 2 項が **欠落変数バイアス（omitted variable bias）** です。$X$ と $C$ が相関し（$\mathrm{Cov}(X, C) \neq 0$）、$C$ が $Y$ に効く（$\gamma \neq 0$）かぎり、$\hat{\beta}_{\text{naive}}$ は本当の因果効果 $\beta$ から**系統的にずれます**。アイスの例なら、気温 $C$ を無視したぶん、アイスが事故を増やすかのような正の係数が出てしまう、というわけです。

$C$ を入れて重回帰 $Y \sim X + C$ にすれば、$X$ の偏回帰係数は「**$C$ を一定にしたうえでの $X$ の効果**」となり、$\beta$ を回復します。これが「**上流にある交絡（共通原因）は調整に含める**」根拠です。

---

## 4. コライダーは調整してはいけない ——入れると選択バイアス

次は逆向きの注意です。共通の**結果** $K$（コライダー $X \to K \leftarrow Y$）は、モデルに**入れてはいけない**変数の代表です。入れると、本来なかった関連が生まれます。

いちばん鮮やかな例として、$X$ と $Y$ が**まったく独立**（$\mathrm{Cov}(X, Y) = 0$）な場合を考えます（図 5.4）。両方が影響する合流点を

$$K = X + Y$$

とします。

ここで $K$ で条件付ける — すなわち $K = k$ となる個体だけを選んで見る、あるいは $K$ を回帰に加える —と何が起きるか。$X, Y$ を標準正規として、条件付き共分散を計算します。

$$\mathrm{Cov}(X, Y \mid K) = \mathrm{Cov}(X, Y) - \frac{\mathrm{Cov}(X, K)\,\mathrm{Cov}(Y, K)}{\mathrm{Var}(K)} = 0 - \frac{1 \cdot 1}{2} = -\frac{1}{2}.$$

独立だったはずの $X$ と $Y$ が、$K$ で条件付けただけで**負の共分散** $-\frac{1}{2}$ をもってしまいました（$K = X + Y$ を固定すれば $Y = K - X$ なので、条件付き相関はなんと $-1$ —完全な負の連動です）。これは因果関係のない**見かけの関連**です。

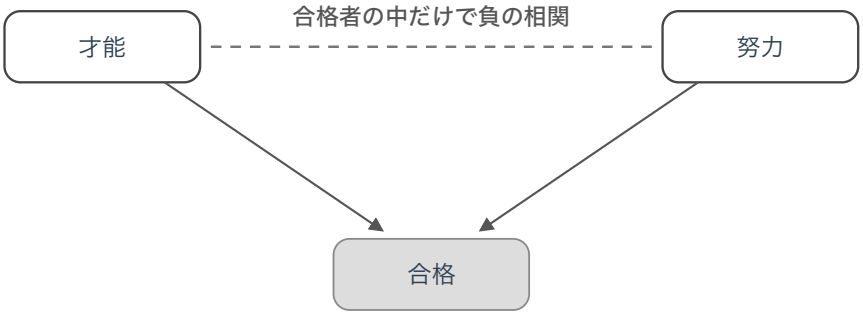



**図 5.4：**コライダーの例。合格は才能と努力の共通の結果。母集団では才能と努力は無関係でも、合格者だけに目を向ける（$K$ で条件付ける）と、両者に負の相関（点線）が生まれる ―これが選択バイアス。

直感はこうです（**バークソンのパラドックス**）。入試の合格 $K$ が「才能 $X$ ＋ 努力 $Y$」で決まるとします。母集団では才能と努力は無関係でも、**合格者だけ**を見れば（$K$ がほぼ一定）、才能が低いのに受かった人は努力が高かったはず ―だから合格者の中では才能と努力が**負に相関**して見えるのです。

「合格者だけを見る」とは、まさに **コライダーの値で標本を選ぶ**ことにほかなりません。だからコライダーでの条件付けが生む歪みを**選択バイアス（selection bias）**と呼びます。回帰に $K$ を入れることは、暗黙にこの選択をしてしまうのと同じで、$X$ と $Y$ に偽の関連を持ち込みます。

> **要点**：交絡（共通原因）とコライダー（共通結果）は、見た目は同じ「$X, Y$ 両方とつながる変数」でも、**調整の作用が正反対**。交絡は入れて初めてパスがふさがり、コライダーは入れるとパスが開く。

---

## 5. 媒介変数・下流の変数も入れない（全効果を見たいとき）

コライダー以外にも、入れると害になる変数があります。

- **媒介変数** $M$（チェーン $X \to M \to Y$）：$M$ は $X$ の効果を $Y$ へ伝える中継ぎです。これを調整すると $X \to M \to Y$ の道がふさがり、$X$ の効果のうち $M$ を経由する分（**間接効果**）が消えてしまいます。$X$ の**全効果**（直接＋間接）を知りたいなら、媒介変数は**入れない**（入れると「過剰調整 over-adjustment」）。
- **結果 $Y$ の下流にある変数**（$Y$ の子孫）：$Y$ の後に起きることを説明変数にすると、結果を使って原因を説明する倒錯になり、やはり関連を歪めます。

逆に、**$Y$ の原因ではあるが $X$ とは無関係な変数**は、因果効果のバイアスには影響せず、入れると推定の**精度**を上げてくれます（分散が小さくなる）。つまり「入れてよい／むしろ入れたい」変数もあります。

実務での大まかな指針は、調整に使う変数を**処置（$X$）より時間的に前に決まっているもの**から選ぶことです。$X$ より後に決まる変数は、媒介かコライダー（またはその子孫）である疑いが濃いからです。ただし「処置前なら常に安全」というわけでもありません。処置前の変数が（測っていない原因たちの）コライダーになっている構造では、条件付けるとかえって閉じていたパスが開きます（**M バイアス**と呼ばれる形です）。最後の判断の根拠は時間順序ではなく、あくまで **DAG（因果の構造）**です。

---

## 6.「全部入れればよい」ではない ——まとめの基準

以上を一般化したのが **バックドア基準（backdoor criterion）** です。ざっくり言えば、調整に使う変数の集合 $Z$ が

1. $X$ と $Y$ の間の**すべてのバックドアパスをふさぎ**（交絡を漏れなく入れ）、かつ
2. $X$ **の下流（子孫）を含まない**（媒介や、$X$ の結果であるコライダーを入れない）

という 2 条件を満たすとき、**線形モデルが正しく指定され、未測定の交絡がないかぎり**、線形モデル $Y \sim X + Z$ の $X$ の係数は、$X$ の **因果効果**に一致します。

| 変数の型（DAG 上の位置） | 例 | 調整に入れる？ | 入れ間違えると |
|---|---|---|---|
| 交絡（共通原因・フォーク） | 気温、年齢 | **入れる** | 欠落変数バイアス（第 3 節） |
| コライダー（共通結果） | 合格・受診の有無 | **入れない** | 選択バイアス（第 4 節） |
| 媒介（中間・チェーン） | 経路上の中間変数 | （全効果なら）**入れない** | 過剰調整（第 5 節） |
| $X$・$Y$ の下流 | 結果の後の指標 | **入れない** | 関連の歪み |
| $Y$ だけの原因（$X$ と無関係） | $Y$ にだけ効く要因 | 入れてよい | （精度が上がる） |

結論はひとつ。**変数を多く入れるほど正確になるわけではありません**。同じ線形回帰でも、因果効果を読みたいなら、入れる共変量を **DAG にもとづいて選ぶ** ことが決定的に重要です ―交絡は迎え入れ、コライダーと媒介は締め出す。「とりあえず全部」は、見えないバイアスを招き入れる近道なのです。

なお、**ランダム化比較試験（RCT）** は、くじ引きで $X$ を割り当てることで、$X$ に入ってくる矢印（$C \to X$ など）を**設計の段階で**すべて断ち切り、バックドアパスをそもそも作らない方法です。観察データでの共変量調整は、この RCT を後追いで模倣しようとする営みだ、と位置づけることができます。

---

## まとめ：この章の幹

1. 観察データで因果効果を読むには、邪魔な関連を**調整（共変量として回帰に追加）**する。だが「何を入れるか」が結論を左右する。
2. 変数の並びは**チェーン（媒介）・フォーク（交絡）・コライダー（合流点）**の 3 型。条件付けの作用は型ごとに異なる。
3. **交絡（共通原因）は入れる**。入れないと欠落変数バイアス $\hat{\beta}_{\text{naive}} = \beta + \gamma\,\mathrm{Cov}(X, C)/\mathrm{Var}(X)$。
4. **コライダー（共通結果）は入れない**。入れると独立な $X, Y$ にも見かけの関連（$\mathrm{Cov}(X, Y \mid K) = -\frac{1}{2}$）＝選択バイアス。

5. **媒介や下流も入れない**（全効果を見たいとき）。$Y$ だけの原因は入れてよい（精度向上）。
6. 指針は **バックドア基準**：すべてのバックドアパスをふさぎ、$X$ の子孫を含めない $Z$ を選ぶ。「全部入れる」は誤り。

# 6. 機械学習の初歩 ―― 「予測」というもうひとつの目標

> **発展的な話題**　この章は講義の中心から一歩進んだ内容です。1 回生のうちは細部にこだわらず「考え方の地図」をつかむことを目標に、必要になったら戻ってきてください。

これまでの章で扱ってきた統計学は、データの背後にある**構造を知る**こと ― 係数の値、その有意性、信頼区間 ― を主な目標にしてきました。これを**推論（inference）**といいます。機械学習はもうひとつの目標を前面に出します。**新しいデータに対する出力をできるだけ正確に当てる**こと、すなわち**予測（prediction）**です。この章では、両者の目的の違いを出発点に、予測の良し悪しを測る考え方 ― **バイアス・バリアンス分解**、**過学習**、**クロスバリデーション** ― を整理します。

> 読み方のコツ
>
> 第 2 節の「誤差＝バイアス² ＋バリアンス＋ノイズ」がこの章の心臓部です。第 3 節でそこから「少しのバイアスを許す」という機械学習らしい発想が出てきます。第 6 節で、前章までの BIC とならぶ AIC が顔を出します。

---

## 1.「推論」と「予測」―― 同じ $y = \beta x$ でも問いが違う

単回帰 $y = \beta x$ を例にとります。同じ式を見ても、二つの立場は**別のこと**を知りたがっています。

- **推論（統計学）**：主役は **係数** $\beta$。「$x$ が 1 増えると $y$ は平均してどれだけ変わるか」「その効果は偶然では説明できないか（$p$ 値）」「どれくらいの幅で信用できるか（標準誤差・信頼区間）」。世界の**仕組み**を語ることが目的。
- **予測（機械学習）**：主役は **出力** $\hat{y}$。新しい $x$ を入れたとき $\hat{y}$ が本当の $y$ にどれだけ近いか**だけ** を気にする。$\beta$ の値そのものや有意性には、必ずしも関心がない。

この違いが、後で見る「**バイアスを少し犠牲にしてでも誤差を下げる**」という機械学習特有の割り切りにつながります。推論ではバイアス（偏り）はあってはならないものですが、予測では「当たればよい」のです。

### 機械学習のモデルには、どんなものがあるか

本章の例は単回帰 $y = \beta x$ ですが、「予測」を目指す機械学習のモデルには、単純なものから複雑なものまで幅があります。イメージのために代表例を挙げておきます。

- **正則化つきの線形モデル（縮小推定）** — **リッジ回帰** や **Lasso（ラッソ）**。線形モデルのまま、係数を 0 の方向へ縮める罰則を加えて予測を安定させます（第 3 節で再登場します）。Lasso は不要な変数の係数をちょうど 0 にするので、変数選択も同時に行えます。
- **決定木とその寄せ集め** —「もし $x_1 > 3$ なら右へ…」という分岐の連なりで予測する**決定木**。1 本では不安定でも、たくさん束ねて平均する**ランダムフォレスト** や **勾配ブースティング**は、表形式データの予測の定番です。
- **$k$ 近傍法** — 新しい点の近くにある $k$ 個のデータの平均（分類なら多数決）をそのまま予測にする、いちばん素朴なモデル。
- **ニューラルネットワーク（NN）** —「線形モデル＋非線形変換」を 1 つの部品とし、それを並べ、重ねたモデル。層を深く重ねたものが**深層ニューラルネットワーク（DNN、深層学習）**で、画像認識や生成 AI の土台です。パラメータ数は数百万〜数十億にもなります。

これらはどれも「新しいデータを正確に当てる」ことを目指す機械学習モデルの一員で、モデルによって**表現力（複雑さ）は大きく異なります**。複雑になれば当てはめる力は上がりますが、そのぶん**ばらつき（バリアンス）も増えます**。自由に形を変えられるモデルは、たまたま手元に来た訓練データの偶然の凸凹まで律儀になぞってしまうので、訓練データを取り直すたびに出来上がる予測が大きく変わってしまう—これが「ばらつきが増える」ということです。だから次節の誤差の分解と、第 4〜5 節の「学習に使っていないデータで評価する」仕組みが、どのモデルにも共通の背骨になります。

---

## 2. 予測誤差の分解 ——バイアスとバリアンス

### まずノイズなしで ——的当てのイメージ

当てたい**真の値**を $f$、手元の訓練データから推定した予測器の出力を $\hat{f}$ とします。訓練データは偶然に左右されるので、データを取り直すたびに $\hat{f}$ は違う値になります — 的当てにたとえると、的の中心が $f$、1 本 1 本の矢が $\hat{f}$ です。このとき矢の「外れ方」は、2 種類に分けられます。

- **狙いのずれ（バイアス）**：矢の**平均的な着地点** $E[\hat{f}]$ が、中心 $f$ から系統的にずれている。
- **ばらつき（バリアンス）**：1 本 1 本の矢が、平均 $E[\hat{f}]$ のまわりで散らばっている。

そして平均二乗誤差は、ちょうどこの 2 つの和に分かれます（図 6.1）。

$$E[(\hat{f}-f)^2] = \underbrace{\left(E[\hat{f}]-f\right)^2}_{\text{バイアス}^2} + \underbrace{E[(\hat{f}-E[\hat{f}])^2]}_{\text{バリアンス}}.$$

仕掛けは「公式の導き方」第 8 節と同じ式変形です。$\hat{f}-f$ に $-E[\hat{f}]+E[\hat{f}]$ を差し込んで 2 乗を展開すると、交差項が $E[\hat{f}-E[\hat{f}]]=0$ でちょうど消えて、この足し算だけが残ります。

### ノイズも入れた完全版

現実のデータには、どんな予測器でも取り除けない偶然のノイズが乗っています。ある点 $x$ での **期待二乗誤差** を考えましょう。真の関係を $y=f(x)+\varepsilon$（$\varepsilon$ は平均 0・分散 $\sigma^2$ のノイズ）とすると、$\hat{f}$ のばらつきに加え


**図 6.1**：的当てで見るバイアスとバリアンス。的の中心が真の値 $f$、各点が訓練データを取り替えるたびの予測 $\hat{f}$。矢のずれは「狙いのずれ（バイアス）」と「平均まわりのばらつき（バリアンス）」に分解でき、2 乗して平均をとると交差項が消えて、この 2 つの和になる。

てこのノイズも誤差に加わり、次の関係が成り立ちます。

$$E[(y-\hat{f}(x))^2] = \underbrace{\left(E[\hat{f}(x)]-f(x)\right)^2}_{\text{バイアス}^2} + \underbrace{E[(\hat{f}(x)-E[\hat{f}(x)])^2]}_{\text{バリアンス}} + \underbrace{\sigma^2}_{\text{除去できないノイズ}}$$

導出は短く、$y = f + \varepsilon$ を代入して $\varepsilon$ が独立であることを使うだけです。

$$E[(f+\varepsilon-\hat{f})^2] = E[(f-\hat{f})^2] + \sigma^2.$$

残った $E[(f-\hat{f})^2]$ は、先ほどの的当ての分解そのもの —バイアス $^2$ ＋バリアンスです。

三つの項の意味は次のとおりです。

- **バイアス**：予測器の平均 $E[\hat{f}]$ が真の値 $f$ からどれだけずれているか。モデルが**単純すぎる**（表現力が足りない）と大きくなる— **過少適合（underfitting）**。
- **バリアンス**：訓練データが変わると $\hat{f}$ がどれだけブレるか。モデルが**複雑すぎる**とデータの偶然の凸凹まで拾い、大きくなる— **過学習（overfitting）**。
- **ノイズ** $\sigma^2$：データ本来の偶然性。どんな予測器でも**取り除けない**下限。

モデルの複雑さを横軸にとると、この 3 項の綱引きが 1 枚の図になります（図 6.2）。

> **発展：分類の場合**　本章の誤差は二乗誤差（回帰）で書いていますが、$k$ 近傍法の「多数決」や画像認識のように**カテゴリを当てる**問題（分類）では、誤差は「外した割合」（0-1 誤差）や交差エントロピーで測ります。損失の式が変わっても、「単純すぎるモデルは系統的に外す（バイアス）」「複雑すぎるモデルは訓練データごとにぶれる（バリアンス）」という構図と、「学習に使っていないデータで評価する」という原則（第 4〜5 節）は、まったく同じように成り立ちます。

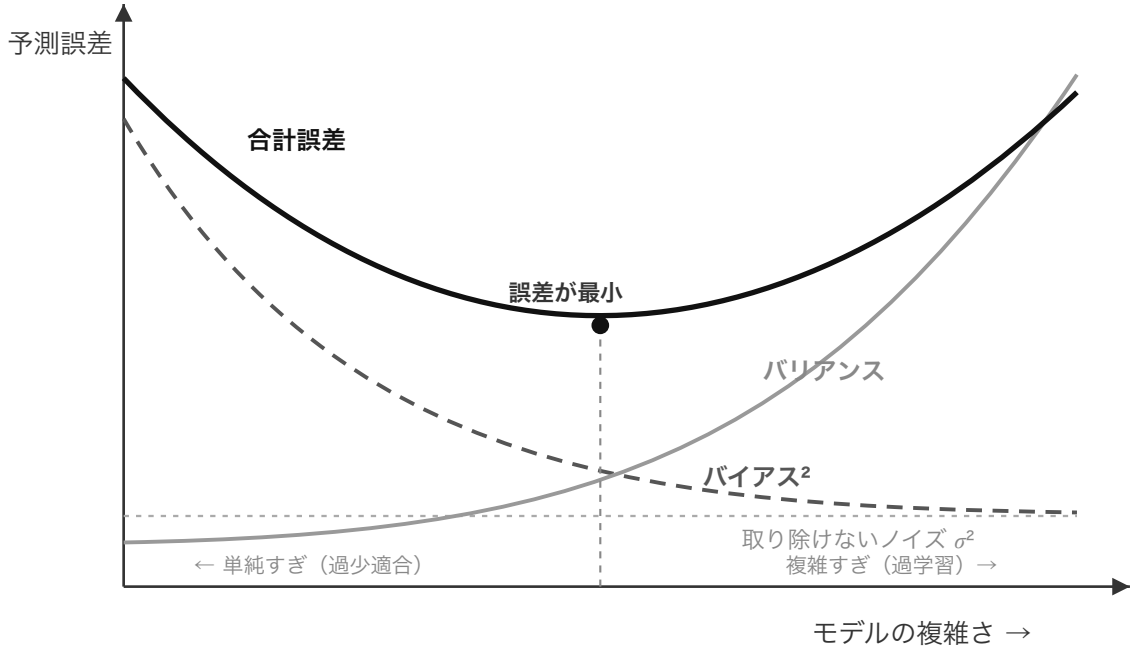



**図 6.2：**モデルを複雑にするほどバイアス² は減るがバリアンスは増える。合計誤差（黒）は両者の和（＋取り除けないノイズ $\sigma^2$）で、**どこかに最小点**をもつ。その点が「ちょうどよい複雑さ」。

---

## 3. バイアス・バリアンス・トレードオフ

ここが推論と予測の分かれ目です。

統計学では**不偏性**（バイアス＝ 0）を理想としてきました（「公式の導き方」第 8 節、なぜ標本分散は $n-1$ で割るか、を思い出してください）。しかし第 2 節の式が教えるのは、**合計誤差を決めるのはバイアスだけではない**ということです。バイアスをゼロに保とうとして複雑なモデルを使うと、バリアンスが膨らんで合計はかえって大きくなることがあります。

そこで機械学習はこう考えます。

> **少しのバイアスを受け入れてでも、それ以上にバリアンスを減らせるなら、合計誤差は小さくなる。**

これが **バイアス・バリアンス・トレードオフ** です。たとえば**リッジ回帰**（係数 $\beta$ をやや 0 へ縮める正則化）は、推定にわざとバイアスを入れますが、そのぶん係数の暴れ（バリアンス）を抑え、新しいデータでの予測を改善します。「不偏だが暴れる推定」より「少し偏るが安定した推定」を選ぶ— これは推定量の平均二乗誤差 $\mathrm{MSE} = \text{バイアス}^2 + \text{バリアンス}$ を最小化する、という一貫した方針の表れです。

---

## 4. 過学習と「新しいデータで試す」

モデルを複雑にすればするほど、**訓練データ**への当てはまり（訓練誤差）はいくらでも良くできます。極端には、全データ点をぴったり通る曲線さえ引けます。しかしそれは、データの偶然の凸凹まで暗記しただけで、**新しいデータ**では大きく外す— これが過学習です。

私たちが本当に知りたいのは、訓練誤差ではなく、まだ見ぬデータに対する誤差、すなわち**汎化誤差（generalization error）**です。訓練誤差は汎化誤差を**楽観的に**見積もってしまうので、当てになりません。

解決の原則はシンプルです。

> **モデルの評価には、学習に使っていないデータを使う。**

いちばん素朴には、データを **訓練用** と **テスト用**に分け（ホールドアウト）、訓練用だけで学習し、テスト用で誤差を測ります。

---

## 5. クロスバリデーション

ホールドアウトには弱点があります。データの一部をテスト用に取り分けるぶん学習に使えるデータが減り、しかも「どう分けたか」で結果が揺れます。これを克服するのが**クロスバリデーション（交差検証, cross-validation, CV）**です。

- **$k$-分割 CV（$k$-fold CV）**：データを $k$ 個のかたまりに分け、1 つをテスト・残り $k-1$ で学習、を全 $k$ 通りで繰り返して誤差を平均する。すべてのデータが一度ずつテストに回るので無駄がない。
- **leave-one-out CV（LOOCV、一個抜き交差検証）**：$k = n$ の極限。**1 個だけ**を抜いて残り $n-1$ 個で学習し、抜いた 1 個で予測する、を $n$ 回くり返して平均する。

$$\mathrm{CV}_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_{-i}(x_i)\right)^2,$$

  ここで $\hat{f}_{-i}$ は「$i$ 番目を除いて学習した予測器」です。

CV が与えるのは、**汎化誤差（予測誤差）の推定値**です。これを使って「どのモデルが、どのハイパーパラメータが、新しいデータでいちばん当たりそうか」を選びます。

ひとつ注意があります。CV の誤差を見て**選んだ**瞬間に、その CV 誤差はもう「新しいデータでの誤差」の公平な見積もりではなくなります。たくさんの候補の中から、たまたま CV 成績のよいものを選んだ —多重比較（「統計の重要ポイント」ポイント 13）と同じ構図だからです。だから本格的な機械学習では、モデル選びに使う**検証データ**と、最後に一度だけ性能を測る**テストデータ**を分けておくのが標準の作法です。

---

## 6. LOOCV と AIC ——別ルートで同じ予測精度へ

予測のためのモデル選択には、CV のほかに **罰則つき尤度**を使う道もあります。その代表が **AIC（赤池情報量規準, Akaike Information Criterion）**です。最大化した尤度 $L(\hat{\theta})$ とパラメータ数 $k$ から

$$\mathrm{AIC} = -2\ln L(\hat{\theta}) + 2k.$$

前章のベイズの式に出てきた **BIC** $= -2\ln L(\hat{\theta}) + k\ln n$（$n$ は標本サイズ）とよく似ていますが、罰則が $2k$（BIC の $k\ln n$ より軽い）である点が違います。そして出自も違います —**AIC は「将来のデータをどれだけうまく予測できるか」を測る規準**として導かれました（情報量にもとづく予測精度の最大化）。

ここで気持ちのよい事実があります。一定の条件（線形回帰＋正規誤差など）のもとで、

$$\mathrm{LOOCV} \;\approx\; \mathrm{AIC}$$

が**漸近的に一致**します（Stone, 1977）。つまり、「データを 1 個ずつ抜いて予測誤差を実測する（LOOCV）」という地道なやり方と、「罰則つき尤度を計算する（AIC）」という解析的なやり方は、**別ルートをたどって同じ予測精度の推定にたどり着く**のです。

こうして AIC は予測のための規準として位置づきます。では前章の BIC とは何が同じで何が違うのか — 次節で整理します。

---

## 7. AIC と BIC を整理する ——予測（事後）か、証拠（事前）か

AIC と BIC は、どちらも「$-2\ln L(\hat{\theta}) + (\textbf{罰則})$」という同じ形をしています。違いは罰則の重さ（AIC は $2k$、BIC は $k\ln n$）だけに見えますが、その背後には「**どちらの予測分布に注目しているか**」という、もっと本質的な差があります。

ベイズには 2 種類の「予測」があります。

- **事前予測分布（prior predictive）**：データを見る **前**の、事前分布だけにもとづく予測。これを 1 つの数にまとめたものが、前章の**周辺尤度** にほかなりません。

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta \qquad (\text{尤度を } \textbf{事前分布} \text{ で平均}).$$

- **事後予測分布（posterior predictive）**：データで学習した**後** の予測。

$$p(\tilde{y} \mid D) = \int p(\tilde{y} \mid \theta)\, p(\theta \mid D)\, d\theta \qquad (\text{尤度を } \textbf{事後分布} \text{ で平均}).$$

この区別が、2 つの規準の性格をきれいに分けます。

- **BIC は事前予測（証拠）の側**。BIC $\approx -2\ln p(D)$ で、$p(D)$ は**事前分布で平均した「見る前の予測力」**＝証拠（エビデンス）です。この $p(D)$ は **事前分布の取り方に敏感**（事前を広げるほど下がる —「ベイズ統

計の数理」§4.3 のオッカムの剃刀）。BIC はその大標本近似で、事前の細部は大標本の近似の過程で落ちてしまいますが、見ているのは**事前で平均した証拠** の側です。

- **AIC は事後予測（汎化）の側**。AIC が見積もるのは、**学習した後に、推定したモデルで将来データをどれだけ当てられるか**＝学習後の予測の良さで、第 6 節のとおり LOOCV に一致します。だから AIC は**学習結果（得られたモデル）による予測に注目した量**であり、データが増えれば事前の影響は薄れていきます。

ひとことでいえば ―「**AIC は学習後の予測（事後予測）を、BIC は学習前の証拠（事前予測＝周辺尤度）を見ている**」。同じ罰則つき尤度の形でも、目指すものが正反対なのです（厳密にベイズの事後分布そのものを使うのは、次に挙げる WAIC です）。

> **発展：WAIC と WBIC（渡辺澄夫）**。AIC・BIC は「正則なモデル」での近似でした。ニューラルネットや混合モデルのように**特異なモデル**（パラメータが一意に定まらない）では、その前提が崩れます。これを乗り越えるのが渡辺によるベイズ版です。**WAIC** は **事後予測分布**から汎化誤差（out-of-sample の予測精度）を推定する **AIC のベイズ拡張**（ベイズ流の LOO-CV に漸近一致）、**WBIC** は**周辺尤度**（ベイズ自由エネルギー）を、温度を調整した特殊な事後分布での 1 回の計算で近似する **BIC のベイズ拡張**です。系統はそのまま受け継がれ ―**WAIC ＝事後予測の系統、WBIC ＝事前予測（証拠）の系統** となります。

整理すると次のとおりです。

| | 予測（事後）の系統 | 証拠（事前）の系統 |
|---|---|---|
| 注目する予測分布 | 事後予測（AIC は $\hat{\theta}$ を代入した近似、WAIC は $p(\tilde{y} \mid D)$ そのもの） | 事前予測＝周辺尤度 $p(D)$ |
| 問い | 学習後、将来データを当てられるか | 見る前にデータを当てられたか |
| 罰則つき尤度（正則モデル） | **AIC**：罰則 $2k$ | **BIC**：罰則 $k \ln n$ |
| ベイズ拡張（特異モデルでも） | **WAIC** | **WBIC** |
| 対応する操作 | LOOCV に一致 | ベイズファクター・自由エネルギー $\mathrm{BF}_{10} \approx e^{(\mathrm{BIC}_0 - \mathrm{BIC}_1)/2}$ |
| 事前分布への感度 | 低い（データが増えると薄れる） | 高い（事前で平均するため） |

（より進んだ整理は Gelman らの解説 *Understanding predictive information criteria for Bayesian models* や、WAIC の Wikipedia 記事 が読みやすいです。）

---

## まとめ：この章の幹

1. **推論**（係数 $\beta$ の意味・有意性）と **予測**（出力 $\hat{y}$ の精度）は別の目標。機械学習は後者を重視する。
2. 予測誤差は $E[(y - \hat{f})^2] = \text{バイアス}^2 + \text{バリアンス} + \sigma^2$ に分解できる。単純すぎ＝高バイアス、複雑すぎ＝高バリアンス。

3. **バイアス・バリアンス・トレードオフ**：少しのバイアスを許してでもバリアンスを下げれば、合計誤差は小さくなる（例：リッジ回帰）。不偏性より MSE 最小。
4. 訓練誤差は汎化誤差を楽観視する。だから**学習に使っていないデータで評価する**（過学習の回避）。
5. **クロスバリデーション**（とくに **LOOCV**）で汎化誤差を推定し、モデルを選ぶ。
6. **LOOCV ≈ AIC**。さらに **AIC は事後予測（予測）**、**BIC は事前予測＝周辺尤度（証拠）** に注目した量 ——「AIC は事後予測、BIC は事前予測を見る」。特異モデルではそれぞれ **WAIC・WBIC** に拡張される。

# A. Rコードの読み方 ——コードは「指示の文」

この章は、講義で実際に使ったRコードを「**読めて、意味がわかる**」ようになるための副教材です。プログラミングが目的ではありません。スライドや配布ノートブックに出てくるコードを見たときに、「**これは何をしている指示なのか**」を自分で説明できることが目標です。

> 読み方のコツ
>
> コードは魔法ではなく、**人間の言葉に直せる短い指示文**です。各行を「何に・何をして・どんな結果が返るか」の3点で読むクセをつけてください。本章はすべて、講義で実際に使ったコードをもとに説明します。

---

## 1. はじめに：コードは怖くない

Rのコードは、つきつめれば次の3種類の文しかありません。

1. **代入** ——「この箱に、この値を入れておく」
2. **関数呼び出し** ——「この道具に材料を渡して、結果を受け取る」
3. **その繰り返し・組み合わせ**

たとえば講義に出てきたこの2行は、

```
x <- c(1, 2, 3, 4, 5)    # 1〜5 の数を並べた「ベクトル」を x に入れる
mean(x)                  # x の平均を計算する（結果は 3）
```

日本語に直すと「**1から5までの数をまとめてxと名づけ、その平均を求めなさい**」というだけの指示です。コードを読むときは、英単語に身構えず、まず「**代入か、関数呼び出しか**」を見分けてください。

---

## 2. Rの基本文法を読む

### 代入<-（左向きの矢印）

```
height <- 170     #「170 を height という名前に入れる」
```

<- は「**右の値を左の名前に入れる**」という矢印です。height = 170 と書いても動きますが、R では慣習として <- を使います(理由は第 9 節)。一度入れた名前は、あとから何度でも呼び出せます。

### ベクトル c() ——値を「ひとまとめ」にする

c() は **combine（つなげる）**の頭文字。複数の値を 1 本の列にまとめます。

```
scores <- c(60, 75, 80, 95)   # 4 人のテスト点をひとまとめにする
```

R では「1 個の数」も「たくさんの数の列」も同じように扱えるのが強みです。mean(scores) のように、**列まるごとを関数に渡せます**。

### 規則的な列を作る seq() と rep()

```
seq(1.5, 12.5, by = 1)   # 1.5 から 12.5 まで「1 きざみ」で並べる(ヒストグラムの境界に使った)
seq(0, 1, length = 100)  # 0 から 1 を「100 点」に等分する(曲線を描くときに使う)
rep(0, 5)                # 0 を 5 回くりかえす → 0 0 0 0 0
```

seq() は **数列を作る道具**、rep() は**同じ値をくり返す道具**です。

### 添字 [ ] ——列の中から取り出す

角かっこ [ ] は「**何番目を取り出すか**」の指定です。**R の番号は 1 から始まります**（第 9 節で再注意）。

```
scores[1]        # 1 番目の値(= 60)
scores[c(1, 3)]  # 1 番目と 3 番目(= 60, 80)
```

#### 論理添字 —— 「条件に合うものだけ」取り出す

[ ] の中に **TRUE/FALSE の列** を入れると、TRUE の位置だけが残ります。これが統計で多用される「条件で絞り込む」操作です。

```
scores[scores >= 80]   # 80 点以上だけを取り出す(= 80, 95)
```

scores >= 80 はまず FALSE FALSE TRUE TRUE という列を作り、[ ] がその TRUE の場所だけを選びます。

### 関数呼び出しの読み方 ——位置引数と名前付き引数

関数は **関数名（材料 1，材料 2，...）** の形です。材料（**引数**といいます）の渡し方には 2 通りあります。

```
hist(NHANES$Age, breaks = seq(-0.5, 80.5, by = 1), col = "lightblue")
```

- `NHANES$Age` … **位置引数**。順番で「最初の材料」と決まる（ここでは「描くデータ」）。
- `breaks = ...`, `col = ...` … **名前付き引数**。名前 `=` 値の形で「どのオプションか」を明示する。

  読むときのコツ：**位置引数は「主役の材料」、名前付き引数は「細かい注文（オプション）」** と思えばよい。名前付き引数は順番を入れ替えても、省略しても動きます（省略時は既定値が使われる）。

### ヘルプの引き方 `?`

意味がわからない関数は、コードの中で次のように調べられます。

```
?hist      # hist 関数の説明（引数の一覧と既定値）を表示する
```

---

## 3. データを読む —— データフレームと NHANES

### データフレーム ——「表」のデータ

この授業で繰り返し使う **NHANES**（米国の健康・栄養調査、約 1 万人分）は**データフレーム**という形式です。Excel の表と同じで、**行が人、列が項目（年齢・身長・性別…）** になっています。

### `$` ——列を 1 本取り出す

```
NHANES$Age       # NHANES という表から「Age（年齢）」の列だけを取り出す
```

`$` は「**表のうち、この名前の列**」という指定です。表名`$`列名 と読みます。

### 行を絞り込む —— 論理フィルタと `subset()`

データフレームの `[ 行の条件 , 列の条件 ]` で、必要な行・列だけを取り出せます。**コンマの左が行、右が列**です。

```
adults <- NHANES[NHANES$Age >= 18, ]    #「年齢 18 以上」の行だけ。コンマ右が空 = 全列を残す
```

同じことを読みやすく書く `subset()` もあります。

```r
adults <- subset(NHANES, Age >= 18)     # 年齢 18 以上の行を取り出す(意味は上と同じ)
```

### NHANES の前処理 —— !duplicated() で重複を消す

NHANES には **同じ人が複数回登場する重複**があります。講義では必ず次の 1 行で重複を消してから使いました。

```r
# duplicated() は「すでに出てきた行」を TRUE にする。! で否定して「重複でない行」だけ残す
NHANES <- NHANES[!duplicated(NHANES$ID), ]
```

- `duplicated(NHANES$ID)` … ID 列を上から見て、2 回目以降に現れた ID を `TRUE` にする。
- `!` … **否定(NOT)**。`TRUE` ↔ `FALSE` を反転させる。だから「重複 **でない**」行が `TRUE` になる。
- これを行の論理添字に使い、**各人を 1 回だけ**残す。

条件を`&`（かつ）でつなげて、重複除去と年齢フィルタを同時にかけることもあります。

```r
nhanes_adult <- NHANES[!duplicated(NHANES$ID) & NHANES$Age >= 18, ]  # 重複除去 かつ 成人
```

---

## 4. 記述統計の関数

データを「要約する数」を返す関数です。引数にはふつう **データの列**を渡します。

| 関数 | 何を返すか | 例 |
|---|---|---|
| `mean(x)` | 平均 | `mean(scores)` |
| `median(x)` | 中央値(まんなかの値) | `median(scores)` |
| `var(x)` | **不偏分散(n − 1 で割る)** | `var(scores)` |
| `sd(x)` | **標準偏差(不偏分散の平方根、n − 1 版)** | `sd(scores)` |
| `quantile(x, p)` | 分位点(下から割合 p の位置) | `quantile(scores, 0.25)` |
| `range(x)` | 最小値と最大値 | `range(scores)` |
| `summary(x)` | 最小・四分位・平均・最大の一覧 | `summary(scores)` |
| `table(x)` | 各値が何回出たかの度数 | `table(NHANES$PhysActive)` |

### 重要：var() と sd() は n − 1 で割る

ここは試験でも講義でも繰り返し強調した点です。R の `var()` と `sd()` は、データの個数 $n$ ではなく $n-1$ で割った値を返します。

$$\texttt{var(x)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \texttt{sd(x)} = \sqrt{\texttt{var(x)}}$$

`var(x)` は標本から母分散を推定するときの **不偏分散** $s^2$ にあたります（$n$ で割る値ではない）。なお `sd(x)` はその平方根で、ばらつきの目安として使いますが、厳密には母標準偏差 $\sigma$ そのものの不偏推定ではありません（この点は本講義の範囲を超えます）。たとえば NHANES の成人身長の標準偏差を計算したコードはこうでした。

```
adults <- NHANES[NHANES$Age >= 18, ]   # 成人だけ
sd(adults$Height, na.rm = TRUE)        # 身長の標準偏差(n-1 版)。na.rm で欠損を除外
```

`na.rm = TRUE` の意味は第 9 節で説明します（欠損値の扱い）。

### `table()` と割合

```
table(NHANES$PhysActive, useNA = "no")         # 各カテゴリの人数を数える(NA は数えない)
prop.table(table(NHANES$PhysActive, useNA = "no"))  # それを割合(合計 1)に直す
```

- `table()` … カテゴリごとの **個数（度数）** を数える。
- `useNA = "no"` … 欠損（NA）を集計に含めない。
- `prop.table()` … 度数を **割合** に変換する。

---

## 5. 図を描く関数とオプションの読み方

base R の作図関数は「**まず 1 枚描く関数**」＋「**上に描き足す関数**」という二段構えです。

### `hist()` ――ヒストグラム

```
hist(NHANES$SleepHrsNight,
     breaks = seq(1.5, 12.5, by = 1),  # ビン(棒)の境界を指定する
     freq   = TRUE,                    # TRUE: 縦軸=度数(人数) / FALSE: 縦軸=密度(面積=1)
     col    = "lightblue",             # 棒の色
     main   = "睡眠時間の分布",          # グラフのタイトル
     xlab   = "睡眠時間", ylab = "人数")  # x 軸・y 軸のラベル
```

- `breaks` … 棒の刻みかた。数値を 1 つ渡せば「だいたいの本数」、列を渡せば「境界そのもの」を指定する。
- `freq` … **TRUE なら個数、FALSE なら密度**。密度ヒストグラムに理論曲線を重ねるときは `FALSE` にする。

### plot() と描き足し lines() / abline()

plot() が土台の 1 枚を描き、lines() や abline() が **その上に線を足します**。

```
x <- seq(140, 200, length = 100)                # 横軸の点を 100 個用意
plot(x, dnorm(x, 170, 5), type = "l",           # type="l": 折れ線で描く
     col = "blue", lwd = 2, ylab = "f(x)")       # col: 色、lwd: 線の太さ(line width)
lines(x, dnorm(x, 170, 10), col = "red", lwd = 2)    # 既存の図に 2 本目の曲線を描き足す
abline(h = 0.04, col = "red", lwd = 2)               # 水平線 y=0.04 を引く(h=横線, v=縦線)
legend("topright", legend = c("σ=5", "σ=10"),        # 凡例(どの線が何か)
       col = c("blue", "red"), lwd = 2)
```

- type = "l"（line)は折れ線、type = "h" は垂直線(棒状)、既定の type = "p" は点。
- abline(h = ...) は水平線、abline(v = ...) は垂直線、abline(a, b) は切片 $a$・傾き $b$ の直線。
- lwd は線の太さ、col は色、lty は線の種類(2 は破線)。

### boxplot() —— 群ごとの箱ひげ図と~ 記法

```
boxplot(Height ~ Gender, data = adults)    #「Height(数値) を Gender(群) ごとに」箱ひげ図にする
```

ここで初めて **チルダ ~** が出てきます。Height ~ Gender は「**左の数値を、右の変数で分けて見る**」という意味の式(formula)で、~ は「〜を〜で説明する／分ける」と読みます(第 8 節で再登場)。

### barplot() —— 棒グラフと描き足し add

```
barplot(rbind(obs, theory), beside = TRUE,    # beside=TRUE: 2 系列を並べて表示
        col = c("gray", "skyblue"))           # 系列ごとの色
```

- beside = TRUE … 複数系列を**横に並べる**（FALSE なら積み上げ)。
- add = TRUE …（一部の作図関数で）**既存の図に重ねて描く**。新しい図にせず上書きする指定。

---

## 6. 確率分布の d / p / q / r 体系

ここが base R の中で最も体系的で、覚えると一気にコードが読めるようになる部分です。R では各分布に対して **接頭辞(先頭の 1 文字)＋分布名**という命名規則があります。

| 接頭辞 | 意味 | 何を返すか | 言いかえ |
|---|---|---|---|
| d | density / mass | 密度 or 確率質量 $f(x)$ | 「**その値の高さ・確率**」 |

| 接頭辞 | 意味 | 何を返すか | 言いかえ |
|---|---|---|---|
| `p` | probability | 累積確率 $P(X \leq x)$ | **「その値以下になる確率」** |
| `q` | quantile | 分位点($p$ の逆) | **「下からこの割合に来る値」** |
| `r` | random | 乱数 | **「その分布からサンプルを生成」** |

分布名は `binom`（二項）, `norm`（正規）, `unif`（一様）, `beta`（ベータ)などです。組み合わせると次のようになります。

| | 二項分布 | 正規分布 |
|---|---|---|
| 密度・確率 `d` | `dbinom(k, n, p)` | `dnorm(x, μ, σ)` |
| 累積確率 `p` | `pbinom(k, n, p)` | `pnorm(x, μ, σ)` |
| 分位点 `q` | `qbinom(prob, n, p)` | `qnorm(prob, μ, σ)` |
| 乱数 `r` | `rbinom(回数, n, p)` | `rnorm(個数, μ, σ)` |

講義で出てきた実例：

```
dbinom(3, 10, 0.5)          # コインを 10 回投げて表が「ちょうど 3 回」出る確率
pnorm(1.96)                 # 標準正規分布で 1.96 以下になる確率(≒ 0.975)
qnorm(0.975)                # 下から 97.5% に来る値(≒ 1.96) ← pnorm の逆向き
rnorm(100, 170, 5)          # 平均 170・標準偏差 5 の正規分布から 100 個の乱数を生成
qbeta(c(0.025, 0.975), alpha_post, beta_post)  # ベータ分布の 2.5%点と 97.5%点(信用区間)
```

> **覚え方**：`p` と `q` は**互いに逆向き**です。`pnorm(値)` → 確率、`qnorm(確率)` → 値。`pnorm(qnorm(0.975))` = `0.975` のように行き来できます。信頼区間で出てくる **1.96** は `qnorm(0.975)` の値です。

---

## 7. ランダムネスとシミュレーション

統計の多くの考え方(標本分布・ブートストラップ・並べ替え検定)は、**コンピュータで何度もくじを引いて確かめる**ことで実感できます。そのための道具です。

### `set.seed()` ——乱数を「再現可能」にする

```
set.seed(12345)    # 乱数の出発点を固定する
rnorm(3)           # → 何度実行しても、同じ 3 つの値が出る
```

乱数は本来「毎回ちがう」ものですが、set.seed(数字) を先に置くと**同じ数字なら毎回まったく同じ乱数列**が出ます。これは結果を**再現できる（他人が確かめられる）**ようにするためです。数字そのものに意味はなく、「どのくじ箱を使うか」の合言葉だと思ってください。

### sample() ——標本を抜き出す（復元・非復元）

```
sample(nrow(nhanes_adult), 250)                    # 行番号から 250 個を「重複なし」で抜く
sample(data, size = length(data), replace = TRUE) # 同じ個数を「復元抽出」（ブートストラップ）
```

- replace = FALSE（既定）… **非復元抽出**。一度選んだものは戻さない（ふつうの標本抽出）。
- replace = TRUE … **復元抽出**。選んだものを戻すので同じ人が何度も出る（**ブートストラップ**で使う）。

### replicate() と for ループ——「同じことを何回もやる」

同じ実験をくり返して結果をためるのが、シミュレーションの中心です。書き方は 2 通り。

```
# 方法 A：replicate(回数, { やること }) —— 結果が自動で集まる
boot_means <- replicate(10000, {
  bs <- sample(samp, length(samp), replace = TRUE)  # 復元抽出で疑似標本を作り
  mean(bs)                                          # その平均を返す
})
# → boot_means には 10000 個の平均が入る（これが標本分布の近似）
```

```
# 方法 B：for ループ —— 1 回ずつ番号 i を動かしながら回す
results <- numeric(1000)                 # 結果を入れる長さ 1000 の空き箱を用意
for (i in 1:1000) {                      # i を 1 から 1000 まで動かす
  results[i] <- mean(sample(data, 30))  # 30 個を抜き出した標本の平均（1 個）を i 番目の箱に入れる
}
```

replicate() は「**回数 → 中身 →結果が並んで返る**」、for は「**番号を動かしながら箱に詰める**」と読みます。中身は同じことができます。

---

## 8. 検定とモデル

ここまでの道具を使って、いよいよ **検定** や**モデルあてはめ** を行う関数です。出力（結果）の読み方も大事です。

### formula 記法 ~ ——「結果 ～説明」

~（チルダ）は「**左を、右で説明する／分ける**」という関係を表します。検定・回帰で共通の書き方です。

| 書き方 | 読み方 |
|---|---|
| `y ~ x` | 「$y$ を $x$ で説明する」（単回帰） |
| `y ~ x1 + x2` | 「$y$ を $x_1$ と $x_2$ で説明する」（重回帰） |
| `y ~ x1 * x2` | 「$x_1$・$x_2$ に加えて、その **交互作用** も入れる」 |
| `数値 ~ 群` | 「数値を群ごとに比べる」（$t$ 検定・箱ひげ図） |

### `t.test()` ——2 群の平均の差を検定する

```
result <- t.test(BPSysAve ~ PhysActive, data = nhanes_sample)  # 血圧を「運動するか」で 2 群比較
```

血圧 ~ 群 という formula で「群ごとに平均を比べる」指示になります。`var.equal = TRUE` を足すと、2 群の分散が等しいと仮定する版（プールした $t$ 検定）になります。

### `cor()` と `cor.test()` ——相関

```
cor(x, y)         # x と y の相関係数だけを返す（-1〜1 の数）
cor.test(x, y)    # 相関係数に加えて、t 統計量・p 値・信頼区間も返す（検定つき）
```

### `lm()` と `summary()` ——線形モデルの出力を読む

`lm()` は **Linear Model（線形モデル）= 直線あてはめ・回帰**の関数です。

```
model <- lm(grade ~ studyTime, data = df)    # 成績を勉強時間で説明する直線をあてはめる
summary(model)                               # 結果（係数・標準誤差・p 値・R^2 など）を一覧表示
```

`summary(model)` の出力は、次のように読みます（数値は例）。

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.000      8.000   3.125   0.0123 *
studyTime      4.500      1.200   3.750   0.0048 **
...
Multiple R-squared:  0.612
```

- **`Estimate`** … 係数の推定値。`(Intercept)` は切片、`studyTime` は「**勉強時間が 1 増えると成績が平均 4.5 増える**」という傾き。

- `Std. Error` … その係数のばらつき（標準誤差）。
- `t value` … 推定値 ÷ 標準誤差。0 からどれだけ離れているか。
- `Pr(>|t|)` … その係数が 0 だとしたら、これくらい以上に極端な値が出る確率（**p 値**）。* が多いほど小さい。
- `R-squared` … モデルがデータのばらつきを説明できた割合（0〜1。例では約 61%）。

`$` を使えば結果の一部だけ取り出せます。

```
t.test(bp ~ act, var.equal = TRUE)$p.value    # 検定結果から p 値だけを取り出す
```

---

## 9. よくあるつまずき

最後に、初学者が必ず一度はひっかかる点をまとめます。

### = と<-

代入はどちらでも動きますが、R では **代入は<-、関数の引数指定は =** と使い分けるのが慣習です。

```
x <- 10              # 代入は <-
hist(x, breaks = 5)  # 引数指定は =（breaks に 5 を渡す）
```

### <-（代入）と==（等しいか判定）

これは別物です。**= 1 個は代入、== 2 個は「等しいか？」の比較**で、TRUE/FALSE を返します。

```
x <- 5      # x に 5 を入れる（代入）
x == 5      # x は 5 と等しいか？ → TRUE（判定）
```

条件で絞り込むときは必ず==（2 個）です。`NHANES$Gender == "female"` のように使います。

### 欠損値 NA と na.rm

データに **欠損（NA）** が 1 つでもあると、`mean()` や `sd()` は **NA を返してしまいます**。

```
mean(c(1, 2, NA))            # → NA（計算できない）
mean(c(1, 2, NA), na.rm = TRUE)  # → 1.5（na.rm=TRUE で欠損を除いて計算）
```

`na.rm` は「**NA を remove（除く）するか**」の指定。NHANES のような実データではほぼ必須です。欠損かどうかの判定は `is.na(x)` を使います（`x == NA` は**使えない**ので注意）。

### 添字は 1 から始まる

多くの言語と違い、**R の番号は 0 ではなく 1 から**始まります。`scores[1]` が最初の要素です。

### `c()` の付け忘れ

複数の値を渡すときは必ず `c()` でまとめます。`col = "blue", "red"` ではなく `col = c("blue", "red")` です。

### 大文字・小文字、全角・半角

R は **大文字と小文字を区別** します（`Mean()` は存在せず `mean()`）。また、かっこ `()` やコンマ `,` は**必ず半角**で。全角スペースや全角かっこはエラーの原因の定番です。

---

> **まとめ**：コードを読むときは、(1) 代入か関数呼び出しか、(2) 主役の材料（位置引数）とオプション（名前付き引数）、(3) 返ってくる結果は何か、の 3 点を口に出して説明してみてください。`d/p/q/r` の体系と `~` の意味、`var/sd` が $n-1$ であることの 3 つを押さえれば、講義のコードはほぼ読めます。

# 更新履歴

- **2026 年 6 月 25 日 ―初版**　全 7 章構成(統計の重要ポイント／公式の導き方／確率分布のまとめ／ベイズ統計の数理／因果推論の初歩／機械学習の初歩／付録：R コードの読み方)。
- **2026 年 7 月 5 日 ― 図版と組版の整備**　図版を SVG で整備(講義全体の地図・標本分布の一族・裾の重さの比較・ベイズ更新・因果グラフなど)。PDF 組版を改良(見出しと柱のゴシック化、欧文フォントの変更、キャプション書式、リンク色の統一)。全文推敲(章間の参照名・用語の統一ほか)。
- **2026 年 7 月 6 日 ―内容の拡充と表記の統一**　和文フォントを原ノ味明朝に変更し、図をグレースケールに統一(的当てによるバイアス・バリアンス分解の図を新設)。用語を「線形モデル(LM)」に統一し、図中の変数をイタリックに。最小二乗法の偏微分の手続きとデザイン行列・疑似逆行列、「調整」の線形モデルでの意味、機械学習モデルの例(Lasso・決定木・NN・DNN など)、MCMC の「パラメータの標本」などの説明を追加。専門家レビューにもとづく改訂として、正規分布の $E$・Var の導出、中心極限定理の明示、二項→ポアソンの極限、ベイズファクターの事前依存性、事後予測分布(ラプラスの継承則)、RCT の位置づけと M バイアス、交差検証と検証／テストデータの分割、分類の場合の注意などを追加。著者情報・原著ウェブブックへの参照・ライセンス(CC BY-NC 4.0)を明記し、GitHub Pages で公開。