

26(2014) 年度 数 理 統 計 学

尾 畑 伸 明

東北大学大学院情報科学研究科

<http://www.math.is.tohoku.ac.jp/~obata>

■ 授業の目的と概要

不確実な予測しかできない状況で合理的な意思決定するときに必要なのが確率・統計のアイデアである。授業では、確率モデルの考え方になじみながら、確率論の基礎概念の理解を深め、統計学への応用を見る。おおむね次のような事柄を取り上げる。学習の到達目標は、基本的な概念や手法を理解し、自ら応用できるようになることにある。

1. 記述統計
2. 確率空間
3. 確率変数とその分布
4. 条件付確率と事象の独立性
5. 標本分布
6. 母数の推定
7. 仮説検定

■ 参考書

1. 拙著「確率統計要論」牧野書店, 2007
過年度の講義を中心に纏めたもの(ここからトピックスを選ぶ予定)。ウェブページには「全学教育ニュース」を設けて、過年度の試験問題・レポート問題などを掲載している。
2. 逆瀬川浩孝「理工基礎 確率とその応用」サイエンス社
この本はなかなか面白い。講義の題材として取り上げる部分もあるかもしれない。
3. F. フェラー(河田龍夫他訳)「確率論とその応用」紀伊国屋
易しいことから高度なことまで、実に内容豊富。世界中の確率論研究者のバイブル。
4. 青木利夫・吉原健一「統計学要論」(改訂版) 培風館, 1985.
5. 吾妻一興・大野芳希・鈴木義也・高木斉「統計学概説」 培風館, 1987.
上の2冊は、標準的な内容をコンパクトにまとめた数理統計学の入門書。
6. P. G. ホーエル「入門数理統計学」 培風館, 1995.
入門とはいえ、結構高度なところまで親切に記述している良書。
7. 竹村彰通「統計」 共立出版, 1997.
数理的な議論はそこそこに、数理統計学の基本をレビューするのに適している。計算機の使用を念頭に、`xlispstat` によるプログラムを示しているが、いささか時代がかっている。フリーソフトなら今では `R` が主流であろう。

8. 鈴木義也・洲之内長一郎「すぐに役立つ統計」 学術図書
数学の予備知識を最小にして統計学の初歩をおさえることを主眼としている。
9. 鈴木武・山田作太郎「数理統計学」 内田老鶴圃, 1996.
比較的高度な内容も含んでいる網羅的な概説書。さらに進んだ数理統計学の内容を眺めることができる。
— 以下, 興味の赴くままに....
10. 楠岡成雄「確率・統計」 森北出版
この本は, 薄くて手軽に見えるが, ベイズ統計に関する記述はかなり興味深い。
11. G. ブロム, L. ホルスト, D. サンデル (森真訳)「確率論へようこそ」 (新装版) シュプリンガー・フェアラーク東京, 2005.
組合せ確率論の面白い問題がたくさん収められていて, クイズ番組を見るようでとても楽しい本。
12. 神永正博「ウソを見破る統計学」 講談社ブルーバックス
気楽な読み物として, 統計学の使われ方を概観する。
13. イアン・ハッキング (石原英樹・重田園江訳)「偶然を飼いならす」 木鐸社, 1999.
「この博物誌的な書物を好奇心に満ちたすべての読者に捧げる」とある。確率統計が 20 世紀の科学に中でいかに成功してきたかを科学的な視点で論ずる。かなり興味深い。
14. 西内啓「統計学が最強の学問である」 ダイヤモンド社
統計リテラシーの必要性をジャーナリストティックに説く話題の本。血が沸き立つような書きぶりではあるが, フィッシャー (20 世紀前半の大統計学者) を超えるのは難しい。

■ 成績評価について

1. 期末試験 (75%) と平常点 (25%) により評価する。60 点が合格基準である。
2. 期末試験: 7 月 23 日 (暫定。後日確定する)。試験は 1 回だけ実施する。病欠などの特別な事情があれば, 公式の手続きによって追試験を行うが, 遅刻・欠席 (クラブ活動等の欠席を含む)・成績不良などを理由に再試験を行うことはない。
3. 平常点は, ミニットペーパーの提出・レポートなどによる。

【ミニットペーパー】 授業中の小問の解答や質問などを書いて提出。— 教師は学生個々の言語レベルに合わせて話すことはできない。数学の専門用語や記号などで分からないことがあれば, その場で質問するのがよい。が, 質問しにくい状況もあるやに思う。というわけで, どんなに稚拙と思われる質問でも遠慮せずに書いてよい。できるだけ, 次回の授業やウェブページで回答する。授業に関係ない質問でも, 気が向けば回答します。質問でなくても (気の利いた) コメントを歓迎する (1~2 点)。ただし, 代筆が判明した場合 (だいたいすぐばれる), 関係者は全員 0 点とする。

第1章 確率を思い出そう



1.1 なぜ確率論を学ぶのか

- (1) 役に立つ
ランダムサンプリングによる推定論の基礎
不十分なデータから状況予測や意思決定を合理的に行う.
- (2) 様々な偶然現象の統一的理解
抽象理論の強み. ある事例のモデルが, 全く違う現象に応用できる.
- (3) 面白い
確率論は, 数学の中にあって比較的若い分野であり, これからも発展する.

1.2 組合せ確率論

起こりうるすべての結果を最小単位 (根元事象) に分解し, 根元事象の全体を Ω とする. 今, 問題になっている事象もまた根元事象の集まり E として表される. つまり, $E \subset \Omega$ である. すべての根元事象が等確率で起こるとき (これは議論の前提), E の起こる確率は

$$P(E) = \frac{|E|}{|\Omega|}$$

で与えられる. $|\cdot|$ は集合の元の個数である. 組合せの個数を数え上げて, 比を取る計算に帰着する. これが組合せ確率論の所以である. 当然, Ω は有限集合であることが前提である.

例題 1.1 (コイントス) ベルヌイ試行ともいう (ヤコブ・ベルヌイ 1654–1705).

最も基本的な確率モデル (ベルヌイ型確率変数). しかし, バカにはできない. 極めて多くの確率モデルはコイントスの組合せで構成される.

例題 1.2 トランプのカード52枚から2枚を同時に抜き取るとき, その2枚がともに絵札 (K, Q, J) である確率を求めよ. [11/221]

例題 1.3 硬貨を10枚投げる時, 表が少なくとも1枚出る確率を求めよ. [1023/1024]

宿題 1 52枚のトランプから2枚を同時に抜き出したとき, 2枚ともエースになる確率を求めよ. また, 52枚のトランプから4枚を同時に抜き出したとき, 4枚ともエースになる確率を求めよ. [1/221, 1/270725]

● 樹形図の応用

例題 1.4 (くじ引き) 10本中あたりが2本含まれているくじがある. このくじを2人が順に引くとき, 1番目に引く人があたる確率, 2番目に引く人があたる確率を求めよ. [2/10]

宿題 2 10本中あたりが2本含まれているくじがある. このくじを順に引くとき, 何番目に引いてもあたりを引く確率は $2/10$ で同じであることを示せ.

宿題 3 A,B,Cがこの順番でコインを投げて, 最初に表を出したものが勝ちとなる. A,B,Cそれぞれが勝つ確率を求めよ.

例題 1.5 (パスカル-フェルマの分割問題) A,Bの2人がゲームをする. これまでの実績からAの勝つ確率は $2/5$, Bの勝つ確率は $3/5$ である. ゲームはどちらかが先に3勝した段階で終わり, 賞金 10000 ユーロを受け取る. Aが2勝, Bが1勝した段階でゲームを中止することとなった. 賞金はどのように配分するのが公平であるか? [仮想的にゲームを継続して勝敗をつけることで配分を決める.] [16 : 9]

例題 1.6 (モンティホールのジレンマ) アメリカのテレビで話題になり議論百出 (映画「ラスベガスをぶっとばせ」にもあった). 3つの扉があり, いずれかの扉の後ろに車 (高級ね) と残りの2つの扉の後ろにはヤギがいる. 君はいずれかの扉を選んで, 当たれば車がもらえる. 君は1つの扉を選んだ. 司会者は, 「ヒントを差し上げましょう」と言って, 1つの扉をあけヤギを逃がした (もちろん司会者は車のある扉を開けたりしない). そしてこう言う「今なら扉を選びなおしてもいいですよ」君ならどうする?

宿題 4 A,Bの2人がゲームをする. これまでの実績からAの勝つ確率は p , Bの勝つ確率は $1-p$ である. ゲームはどちらかが先に4勝した段階で終わり, 賞金 10000 ユーロを受け取る. Aが2勝, Bが1勝した段階でゲームを中止することとなった. 賞金はどのように配分するのが公平であるか?

1.3 確率論小史

カルダノ (1501–1576) ルネッサンスの万能人. 自伝が面白い (ギャンブルは生活の一部).

ガリレオ (1564–1642) サイコロを3回振る場合の確率

パスカル (1623–1662) とフェルマ (1601–1665) の往復書簡 (組合せ論的確率論の祖)

ベルヌイ (1654–1705) 大数の法則の証明

ラプラス (1749–1827) 「確率の解析的理論」 (共立), 「確率の哲学的試論」 (岩波文庫).

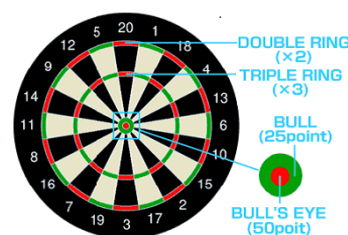
微分積分学との融合による一大発展.

コルモゴロフ (1903–1989) の公理系「確率論の基礎概念」 (根本訳) 東京図書

レヴィ (1886–1971), ウィナー (1894–1964) ブラウン運動, 確率過程

伊藤清 (1915–2008) 時間とともに変化するランダム現象の確率解析 (確率微分方程式)

第2章 確率空間



2.1 確率の与え方

確率を考えるときは, 対象とする「ランダム現象」に対して3つの構成要素を把握するべし.

Ω : 標本空間 (見本空間) = 根元事象 (標本点, 見本点ともいう) の集合

$\mathcal{F}$: 事象の集合 (確率は事象に対して与える)

P : 確率の与え方をできるだけ一般式で表す

記法 事象は標本空間 Ω の部分集合なので, 事象には集合の演算記号を用いる.

- 根元事象 $\omega \in \Omega$
- 部分事象 $E \subset F$ ($E = F$ も許す. 不等式 $a < b$ の使い方とは違うので注意せよ.)
- 空事象 $\emptyset$
- 全事象 Ω
- 余事象 E^c
- 和事象 $E \cup F, E_1 \cup \dots \cup E_n$
- 積事象 $E \cap F, E_1 \cap \dots \cap E_n$
- 排反な事象 $E \cap F = \emptyset$

例題 2.1 (組合せ確率) 根元事象が有限個のときは, Ω が有限集合になる. いずれの根元事象も等確率で起こることが想定されるなら, 事象 E の起こる確率は

$$P(E) = \frac{|E|}{|\Omega|}, \quad |\cdot| \text{ は集合の元の個数.}$$

例: コイントスやサイコロ振りの確率モデル

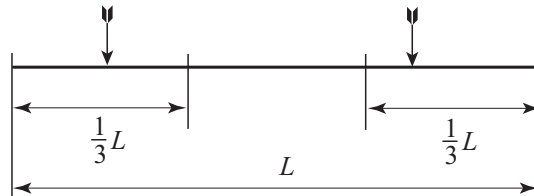
例題 2.2 (Ω が可算集合 (番号付けできる無限集合) のとき) バス停に並んでいる人の数, 栗ようかんに含まれる栗の個数など, 個数を問題にするとき典型的に現れる. 後で議論するが,

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots,$$

が基本的である. ただし, $\lambda > 0$ は定数である. これをパラメータ (または平均) λ のポアソン分布という.

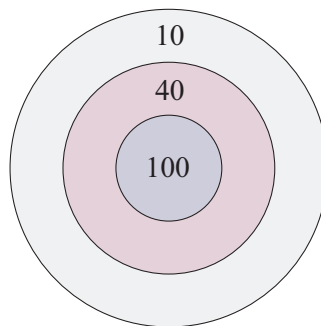
宿題 5 (マメ知識：ポアソン分布では、偶数の出る確率は奇数の出る確率より大きい) バス停に並んでいる客の人数がパラメータ λ のポアソン分布に従うとする。その人数が偶数である確率と奇数である確率とではどちらが大きいのか? [ヒント：指数関数 e^λ と $e^{-\lambda}$ のテーラー展開]

例題 2.3 (Ω が連続無限集合のとき) 棒をランダムに折って長いほうの断片の長さをはかる。長いほうの長さは短いほうの2倍以上になる確率。 [2/3]



● 線分から1点を選ぶとき、どの点も同程度の確からしきで選ばれることを想定すると(民主的ですね)、特定の点を選ばれる確率は0になる。

例題 2.4 (ダーツ) 下手くそなプレイヤーが獲得できる点数の平均値は? ただし、3つの円の半径の比は1:2:3となっている。 [30]



宿題 6 棒をランダムに折って2本の断片を作るとき、長いほうの長さが短いほうの3倍以下になる確率を求めよ。 [1/2]

宿題 7 地図帳で目的地を探すとき、目的地がページの端にあって不便な思いをすることが多くないか? 30cm×40cmの地図で、目的地が周辺5cmの範囲に見つかる確率を求めよ。 [1/2]

● **確率の定め方の原理** Ω における E の占める割合。

$$P(E) = \frac{|E|}{|\Omega|}, \quad |\cdot| \text{ は集合の「大きさ」.}$$

集合の「大きさ」の測り方は様々である。個数、長さ、面積、体積、...

宿題 8 ある2人は正午から午後1時10分前の間に公園に到着し10分間だけそこに滞在するのが日課である。ただし、公園に到着する時刻はお互いに無関係にランダムであるとする。この2人が公園で遭遇する確率を求めよ(どのような確率モデルを作ったかを明確にすることが大事)。 [9/25]

2.2 確率空間

定義 2.5 事象 E に対して数値 $P(E)$ が定められていて、次の3性質をみたすとき、 P を標本空間 Ω 上の確率という。また、 $P(E)$ を E の起こる確率という。

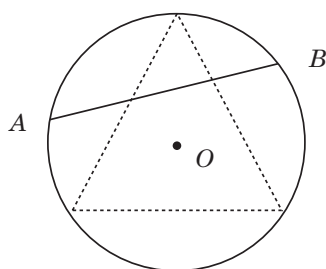
- (i) $0 \leq P(E) \leq 1$.
- (ii) $P(\Omega) = 1$.
- (iii) [可算加法性] $E_1, E_2, \dots \in \mathcal{F}$ が互いに素 (つまり、 $i \neq j$ ならば $E_i \cap E_j = \emptyset$) ならば、

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n).$$

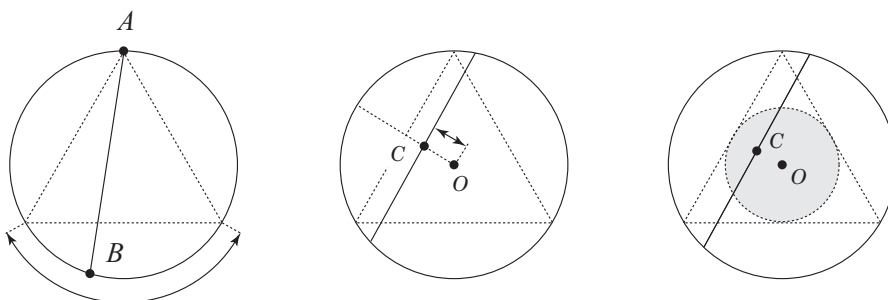
このとき、3点セット $(\Omega, \mathcal{F}, P)$ を**確率空間**という。

2.3 ベルトランのパラドックス

単位円にランダムに引かれた弦 AB の長さが $\sqrt{3}$ を越す確率を求めよ。(なお、 $\sqrt{3}$ は内接する正3角形の辺の長さである.)



ベルトランによる解答 $[1/3, 1/2, 1/4]$



[ポイント] ベルトランのパラドックスにおいて、「単位円にランダムに弦を引く」際に考えるべき根元事象 (標本) および標本空間は何か? ベルトランのパラドックスとは確率モデルの違いに他ならない。今日的にはパラドクスは完全に解消している。

2.4 演習問題

演習 1 52 枚のトランプから同時に 5 枚を抜き出すとき、次の確率を求めよ.

- (1) ロイヤルストレートフラッシュ(同じスートで A,K,Q,J,10) ができている確率
- (2) フォーカードができている確率
- (3) ワンペアができている確率

演習 2 0 から 9 までの数字を 5 個並べて作った乱数 $00000, 00001, \dots, 99999$ のうち 1 つを考える.

- (1) 選ばれた乱数に 9 がちょうど 1 個含まれる確率を求めよ.
- (2) 選ばれた乱数に 9 がちょうど 2 個含まれる確率を求めよ.
- (3) 選ばれた乱数に $0, 1, \dots, 9$ のうち少なくとも 1 つがちょうど 2 個含まれる確率を求めよ.
- (4) 選ばれた乱数に $0, 1, \dots, 9$ のうち少なくとも 2 つがちょうど 1 個含まれる確率を求めよ.

演習 3 (学生さんのコメントから改題) $1\square2\square3\square4\square5$ の $\square$ に $+$ または $-$ のいずれかをランダムに選んで数式を作るとき、その答えが 3 の倍数になる確率を求めよ.

演習 4 サイコロを振り続けて、初めて 1 の目が出るまでに何回サイコロを振ったかを考えよう. 1 の目が出た回は除いて振った回数を k とする. $k \geq 10$ となる確率を求めよ.

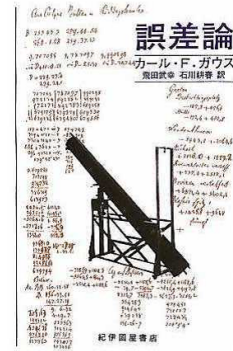
演習 5 直角二等辺三角形の内部に 1 点 P をランダムに選び、直角をはさむ 2 本の等辺に P から垂線を下ろして長方形を作る. この長方形の面積が直角二等辺三角形の面積の $1/3$ 以上になる確率を求めよ.

演習 6 (a, b) を単位正方形 $\{(a, b); 0 \leq a \leq 1, 0 \leq b \leq 1\}$ からランダムに選ぶとき、2 次方程式 $x^2 + ax + b = 0$ が実根をもつ確率を求めよ.

演習 7 (包除原理) 次の等式を証明せよ.

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

第3章 記述統計



3.1 母集団と標本

母集団と個体: 統計的な調査の対象を母集団といい、母集団の各構成要素を個体という。

変数または変量: 各個体に対して何らかの調査や測定が行われ、その特性を表す観測値 (測定値・データともいう) が得られる。観測値は個体ごとに変化するものがふつうであり、そのような観測値をひとまとめにして変数あるいは変量と呼ぶ。

全数調査と標本調査

例題 3.1 (クラス A に所属する学生の身長・体重)

学生番号 \ 項目	身長 (cm)	体重 (kg)
1	166.3	62.0
2	167.6	53.0
3	175.5	67.5
⋮	⋮	⋮
300	164.0	57.0

この場合、母集団 = 個体 = 変数 =

1 変量データ (1 次元データ): 1 つの変数 x に注目して得られた観測値で $x_1, x_2, \dots, x_i, \dots, x_n$ のような実数列。たとえば、身長に注目して得られた観測値を取り出せば、

166.3 167.6 175.5 $\dots$ 164.0

多変量データ (多次元データ): 複数の変数 $x, y, \dots, z$ に注目して得られた観測値は、変数の個数を次元にもつベクトルの列 $(x_1, y_1, \dots, z_1), (x_2, y_2, \dots, z_2), \dots, (x_i, y_i, \dots, z_i), \dots, (x_n, y_n, \dots, z_n)$ 。たとえば、身長と体重をセットにしたものを変数とすれば、2 変量データ (2 次元データ) が得られる:

(166.3, 62.0) (167.6, 53.0) (175.5, 67.5) $\dots$ (164.0, 57.0)

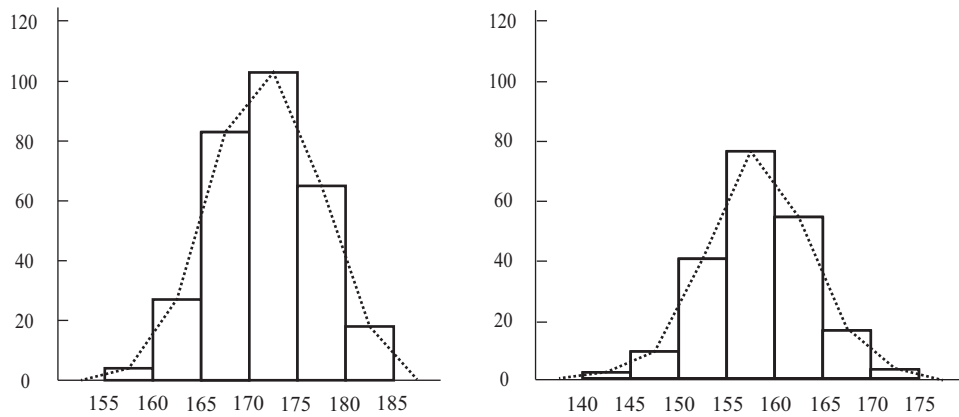
3.2 1 変量データの記述

変数 x の観測値 $x_1, x_2, \dots, x_n$ を整理する.

度数分布と相対度数

階級	150–	155–	160–	165–	170–	175–	180–	合計
度数	0	4	27	83	103	65	18	300
相対度数	0	0.013	0.090	0.277	0.343	0.217	0.060	1.000

ヒストグラムと度数折れ線 (度数多角形)



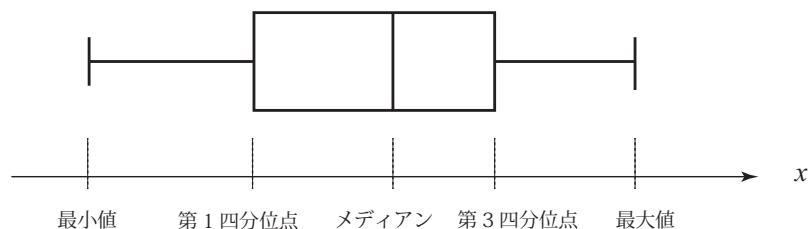
3.3 代表値と分布のばらつき

n 個の観測値 $x_1, x_2, \dots, x_n$ を1つの値で代表させたい.

- 平均値 (算術平均):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- メディアン (中央値): 観測値 $x_1, x_2, \dots, x_n$ を大小順に並び替えたときの真ん中の順位にくる観測値
- 箱ひげ図 (box plot): 順位統計量



- モード (最頻値) 観測値 $x_1, x_2, \dots, x_n$ の中に同じ値が重複して現れる場合は, 現れる回数が最も多い観測値. モードは2つ以上あることもある.

- 分散:

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

- 標準偏差: 分散の正の平方根

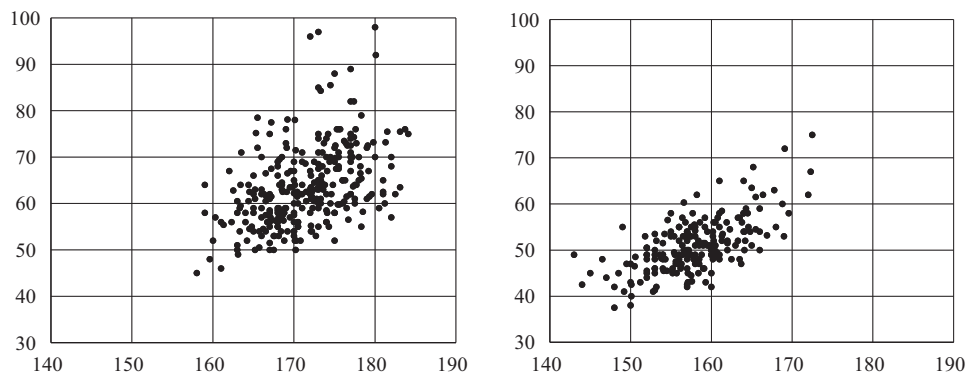
$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

宿題 9 あるクラブには15名のメンバーがいる。このクラブに19歳の若手が1人加わったため平均年齢が3歳(も!)若返ったという。このクラブの構成メンバー16名の年齢構成について考察せよ。

3.4 2変量データと散布図

散布図: 2つの変数 x, y を (x, y) として, 座標平面に図示

例題 3.2 (身長 (x) と体重 (y) の散布図)



2変数の観測値 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ が得られたとしよう。
変数ごとの平均値と分散:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i, & \sigma_x^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i, & \sigma_y^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \end{aligned}$$

共分散

$$\sigma_{xy} = \sigma_{yx} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

相関係数

$$r = r_{xy} = r_{yx} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}} \sqrt{\sigma_{yy}}}$$

正の相関, 負の相関, 無相関

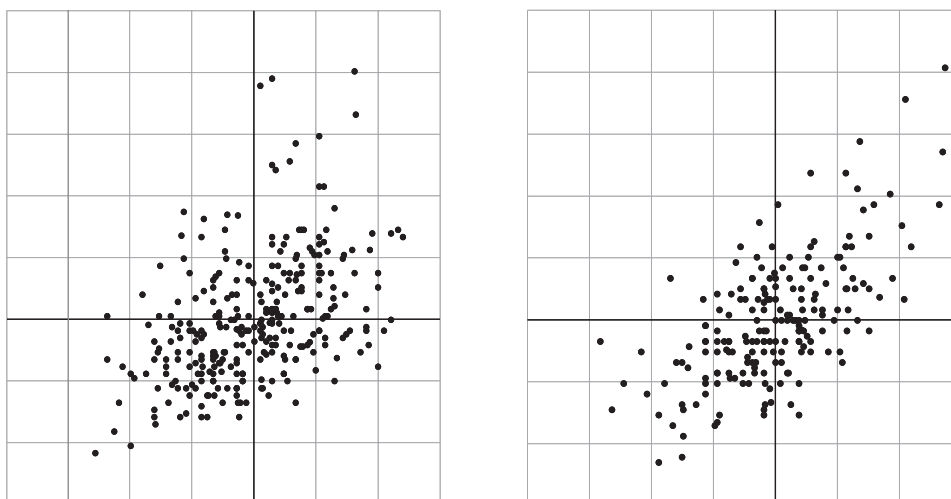
観測値の標準化 (規準化)

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sigma_x}, \quad \tilde{y}_i = \frac{y_i - \bar{y}}{\sigma_y}$$

定理 3.3 2変数 x, y に対して, 規準化された変数を $\tilde{x}, \tilde{y}$ とするとき, $r_{xy} = \sigma_{\tilde{x}\tilde{y}} = r_{\tilde{x}\tilde{y}}$ が成り立つ. つまり, 変数 x, y の相関係数は, それらを規準化した変数の共分散に一致する.

定理 3.4 2変数 x, y の相関係数は $-1 \leq r_{xy} \leq 1$ を満たす.

一般に相関係数の絶対値が 1 に近いとき, 考えている 2 変数について, 相関が強いといい, 0 に近いときは相関が弱いという.



	共分散	相関係数
クラス A	20.1530	0.447
クラス B	20.2323	0.648

宿題 10 $\sum_{i=1}^n \{(t(x_i - \bar{x}) + (y_i - \bar{y}))\}^2 \geq 0$ がすべての実数 t で成り立つことを用いて,

$$|\sigma_{xy}| \leq \sigma_x \sigma_y, \quad -1 \leq r_{xy} \leq 1$$

を導け. ただし, 相関係数を考えるときは $\sigma_x > 0, \sigma_y > 0$ を仮定する.

3.5 回帰分析

説明変数を x , 目的変数を y とするとき, $y = f(x)$ となるような合理的な関数を与える. 特に, 関数の中でもっとも簡単な1次関数で

$$y = ax + b \quad (3.1)$$

のように与えられるものを**線形回帰モデル**または y の x への**回帰直線**という.

最小二乗法 説明変数 x と目的変数 y に1次関数 $y = ax + b$ を想定して, 各観測値 (x_i, y_i) に対して偏差 ϵ_i を

$$y_i = ax_i + b + \epsilon_i$$

によって定義する. 偏差の平方和

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

を最小にするように a, b を定めるのが最小二乗法である.

Q は a, b について2次関数.

$$\frac{\partial Q}{\partial a} = 2an(\sigma_x^2 + \bar{x}^2) - 2n(\sigma_{xy} + \bar{x}\bar{y}) + 2bn\bar{x},$$

$$\frac{\partial Q}{\partial b} = 2bn - 2n\bar{y} + 2an\bar{x},$$

が得られる. 連立方程式 $\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial b} = 0$ を解いて,

$$a_0 = \frac{\sigma_{xy}}{\sigma_x^2}, \quad b_0 = \bar{y} - a_0\bar{x}$$

求めるべき線形回帰モデルは, $y = a_0x + b_0$ となる.

定理 3.5 観測値 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して, x を説明変数, y を目的変数とする線形回帰モデルは

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) = \frac{\sigma_y}{\sigma_x} r(x - \bar{x})$$

で与えられる. また, y を説明変数, x を目的変数とする線形回帰モデルは

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}) = \frac{\sigma_x}{\sigma_y} r(y - \bar{y})$$

で与えられる.

注意: 定理に述べた2つの回帰モデルは, いずれも平均ベクトル $(\bar{x}, \bar{y})$ の定める点を通るが一般には一致しない(説明変数と目的変数は対称的な役割にない).

例題 3.6 (クラス A,B に所属する学生の身長 (x) と体重 (y) の線形回帰モデル) クラス A について,

$$\bar{x} = 171.45, \quad \bar{y} = 63.59, \quad \sigma_x^2 = 27.7557, \quad \sigma_y^2 = 73.3508, \quad \sigma_{xy} = 20.1530$$

となっている. したがって, x を説明変数とする線形回帰モデルは,

$$y = 0.73x - 61.57$$

となる. また, y を説明変数とする線形回帰モデルは

$$x = 0.27y + 154.28$$

となる. 同様にして, クラス B について計算すると,

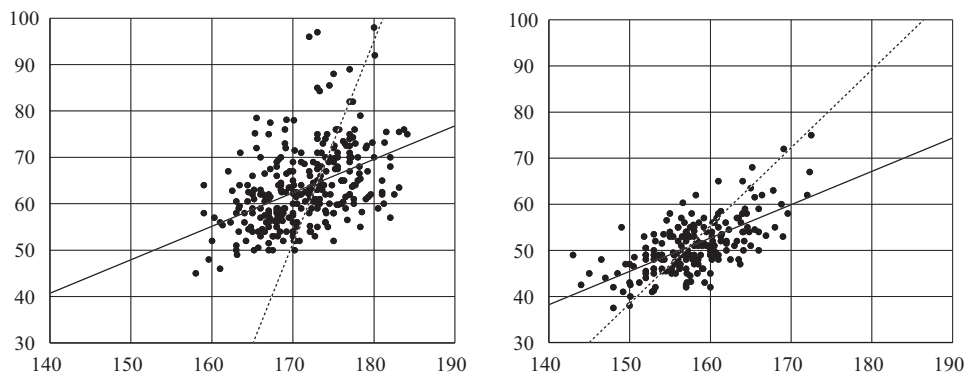
$$\bar{x} = 157.98, \quad \bar{y} = 51.05, \quad \sigma_X^2 = 28.1218, \quad \sigma_Y^2 = 34.6541, \quad \sigma_{XY} = 20.2323$$

から, x を説明変数とする線形回帰モデルは,

$$y = 0.72x - 62.70$$

となり, y を説明変数とする線形回帰モデルは

$$x = 0.58y + 128.18$$



宿題 11 観測値 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して, x を説明変数, y を目的変数とする回帰直線と y を説明変数, x を目的変数とする回帰直線を同一の xy -座標平面に描いたとき, 前者の方が後者よりも傾きの絶対値が小さいことを示せ.

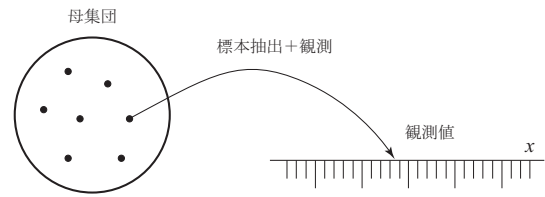
3.6 演習問題

演習 8 平均値・メディアン・モードの大小関係は一般には決まらない. 5 個の観測値を具体的に与えて, さまざまな大小関係が現れることを示せ.

演習 9 観測値 $x_1, x_2, \dots, x_n$ の標準偏差が $\sigma_x = 0$ となるのは, 観測値がすべて一致して $x_i = \bar{x}$ が成り立つときに限ることを示せ.

演習 10 観測値 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ がある直線 $y = ax + b$ ($a \neq 0$) 上にのっているための必要十分条件は $r_{xy} = \pm 1$ であることを示せ.

第4章 確率変数



4.1 確率変数とは？

ランダム現象の観測結果を数値化したもの

- (1) コインを投げて表が出たら 1, 裏が出たら 0 を対応させる.
- (2) サイコロ投げで出た目をそのまま観測値とする.
- (3) コインを 5 回投げて表の出た回数を数える.
- (4) 棒をランダムに折って長いほうの断片の長さをはかる.
- (5) (ランダムサンプリング) 東北大学の学生からランダムに 1 人選んで身長を測定する.

変数と確率変数 ある範囲を動くことが想定されている実数のことを数学用語で (実) 変数という. 文字の使い方は自由ではあるが, 変数には $x, y, z, t, \dots$ などがよく用いられる. たとえば, $0 \leq x \leq 1$ とあれば, x は 0 以上 1 以下の範囲にある実数を代表する変数と理解される. ランダム現象を観測して得られる観測値は, 偶然の影響のために定数ではなく, ある範囲を動く変数となる. しかも, どの値が出やすく, どの値が出にくいといった確率的な傾向が付加されている. このような変数を**確率変数**という. 習慣によって, 確率変数には $X, Y, Z, T, \dots$ のように大文字を用いる.

離散型確率変数と連続型確率変数

4.2 確率変数の分布 (離散型の場合)

確率変数は, ある範囲を動く変数であることに変わらないのだが, どの値が出やすく, どの値が出にくいといった確率的な傾向が与えられているので, そのことを表現する必要がある. そのためのアイデアが「確率変数の分布」である.

例題 4.1 (二項分布 $B(3, 1/2)$) コインを 3 回投げたときに表の出る回数を X とすると, X は $\{0, 1, 2, 3\}$ の範囲を動く確率変数である. このとき,

$$P(X=0) = \frac{1}{8}, \quad P(X=1) = \frac{3}{8}, \quad P(X=2) = \frac{3}{8}, \quad P(X=3) = \frac{1}{8}$$

が成り立つ. このように X の取りうる値それぞれに対して, それをとる確率を与えることで, X の値の出やすさの確率的な傾向 (確率分布) はすっかり明らかにされる.

離散型確率変数の分布 離散型確率変数 X の取りうる値を網羅して $\{a_1, a_2, \dots, a_i, \dots\}$ とするとき,

$$P(X = a_i) = p_i, \quad i = 1, 2, \dots,$$

を与えることで, X の確率分布がわかる. あるいは, このような一覧表を X の**確率分布**という. 当然であるが,

$$p_i \geq 0, \quad \sum_i p_i = 1$$

に注意しよう. ($p_i = 0$ となる a_i を除外しても分布としての本質は同じであるが, $p_i = 0$ を排除しないほうが一般性が保てるので計算が楽).

定義 4.2 離散型確率変数 X のとりうる値を $\{a_1, a_2, \dots\}$ として, $p_i = P(X = a_i)$ とおく. X の**平均値**と**分散**が

$$\begin{aligned} \mathbf{E}[X] &= m = m_X = \sum_i a_i p_i = \sum_x x P(X = x), \\ \mathbf{V}[X] &= \sigma^2 = \sigma_X^2 = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 \\ &= \sum_i (a_i - m)^2 p_i = \sum_i a_i^2 p_i - m^2 \end{aligned}$$

で定義される. 分散の正の平方根を**標準偏差**という.

例題 4.3 コインを3回投げるとき, 3回ともに表なら100円, 2回だけ表なら50円, 1回だけ表なら10円もらえるが, 表が出ないときは80円支払わなければならない. このゲームを1回するとき, もらえる金額の平均値と標準偏差を求めよ. [$m = 25, \sigma^2 = 2400, \sigma = 20\sqrt{6}$]

宿題 12 サイコロ振りで, 初めて1の目が出るまでに振った回数を X とする. (1の目の出た回は含めない. したがって, いきなり1が出たときは $X = 0$ とする.) このとき, X の平均値と分散を求めよ. [$m = 5, \sigma^2 = 30$]

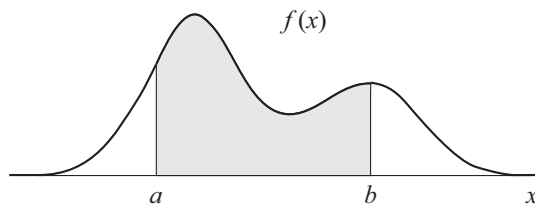
4.3 確率変数の分布 (連続型の場合)

連続型確率変数を全く一般的に扱うためには「測度論」が必要.

定義 4.4 連続型確率変数 X に対して,

$$P(a \leq X \leq b) = \int_a^b f_X(t) dt$$

を満たす関数 $f(x) = f_X(x)$ が存在するとき, それを X の**確率密度関数**という.



定理 4.5 一般に, 密度関数 $f(x)$ は次の性質をもつ.

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

定理 4.6 一般に, 確率変数 X に対して,

$$F_X(x) = P(X \leq x)$$

を X の分布関数という. $F_X(x)$ が連続関数であって区分的に微分可能であれば,

$$f_X(x) = F'_X(x) = \frac{d}{dx} F_X(x)$$

が X の密度関数になる.

定義 4.7 確率密度関数 $f(x)$ であらわされる確率変数 X の平均値と分散は

$$\begin{aligned} \mathbf{E}[X] &= m = m_X = \int_{-\infty}^{\infty} x f(x) dx, \\ \mathbf{V}[X] &= \sigma^2 = \sigma_X^2 = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 \\ &= \int_{-\infty}^{\infty} (x - m)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - m^2 \end{aligned}$$

で定義される. 分散の正の平方根を標準偏差という.

例題 4.8 長さ L の棒をランダムに 2 分割したとき, 長い方の断片の長さを X とする. X の分布関数, 密度関数, 平均, 分散を求めよ.

$$F_X(x) = \begin{cases} 0, & x \leq L/2, \\ \frac{2x - L}{L}, & L/2 \leq x \leq L, \\ 1, & x \geq L, \end{cases} \quad f_X(x) = \begin{cases} \frac{2}{L}, & L/2 \leq x \leq L, \\ 0, & \text{その他}, \end{cases} \quad m = \frac{3L}{4}, \quad \sigma^2 = \frac{L^2}{48}$$

宿題 13 長さ L の棒をランダムに 2 分割したとき, 短い方の断片の長さを Y とする. Y の分布関数, 密度関数, 平均, 分散を求めよ.

定義 4.9 分布関数が連続な確率変数を**連続型**という. 分布関数がジャンプでのみ増加するような確率変数を**離散型**という. (どちらでもないものも存在する.)

例題 4.10 サイコロ投げで得られる出目を X とするとき, X の分布関数を求めグラフで示せ.

定理 4.11 (分布関数の性質) 確率変数 X の分布関数 $F = F_X$ は次の性質をもつ.

- (1) [単調増加] $x_1 \leq x_2$ ならば $F(x_1) \leq F(x_2)$.
- (2) $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$.
- (3) [右連続] $\lim_{\epsilon \rightarrow +0} F(x + \epsilon) = F(x)$.

4.4 確率ベクトル

n 次元確率ベクトル $\mathbf{X} = (X_1, X_2, \dots, X_n)$

定義 4.12 2次元確率ベクトル (X, Y) の共分散が

$$\sigma_{XY} = \text{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

で定義される. $\sigma_{XX} = \sigma_X^2$ は分散である. また, 相関係数が,

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}} \sqrt{\sigma_{YY}}} = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbf{V}[X]} \sqrt{\mathbf{V}[Y]}}$$

で定義される.

相関係数は $-1 \leq r_{XY} \leq 1$ を満たす. $r > 0$ のとき**正の相関**, $r < 0$ のとき**負の相関**, $r = 0$ のときは**無相関**という. また, $|r|$ が1に近いほど相関が強いという.

例題 4.13

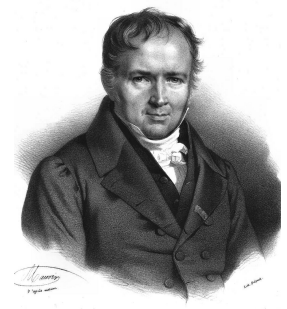
$X \backslash Y$	400	600	合計
200	2/12	5/12	7/12
300	4/12	1/12	5/12
合計	6/12	6/12	1

4.5 演習問題

演習 11 中心を O とする半径 R の円の内部にランダムに1点を選び, その点を通る中心を O とする円の面積を X とする. X の分布関数, 確率密度関数, 平均値, 分散を求めよ.

演習 12 サイコロを2個投げて出た目のうち大きい方 (同じ目のときはその目) を X , 小さい方 (同じ目のときはその目) を Y とする. X, Y の共分散と相関係数を求めよ.

演習 13 サイコロを4回投げるとき, 1の目の出る回数を X と6の目の出る回数を Y とする. X, Y の相関係数の正負について予想せよ. 次に, 計算によって相関係数を求め予想を確かめよ. $[P(X = i, Y = j)$ の一覧表を作るとよい. $\mathbf{E}[X] = \mathbf{E}[Y] = \frac{2}{3}, \mathbf{V}[X] = \mathbf{V}[Y] = \frac{5}{9}, \mathbf{E}[XY] = \frac{1}{3}, \text{Cov}(X, Y) = -\frac{1}{9}, r_{XY} = -\frac{1}{5}]$



第5章 基本的な確率分布

(I) 離散分布

5.1 二項分布

表が出る確率が p であるコインを n 回投げたとき, 表の出る回数 X の分布

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots$$

この分布を成功確率 p の**二項分布**といい, $B(n, p)$ で表す.

特に, $n = 1$ のとき, $B(1, p)$ を成功確率 p の**ベルヌーイ分布**という.

例題 5.1 $B(4, 1/2)$ と $B(4, 1/4)$ を図示せよ.

5.2 幾何分布

表が出る確率が p であるコインを投げ続けるとき, 表が初めて出るまでに表の出た裏の回数 X の分布は

$$P(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, \dots$$

この分布をパラメータ p の**幾何分布**という. (待ち時間の分布として重要)

補注 文献によっては, 表が出る確率が p であるコインを投げ続けるとき, 表が初めて出るまでに要したコイン投げの回数 (表が出た回も 1 回と数える) Y の分布を幾何分布といている.

$$P(Y = k) = p(1-p)^{k-1}, \quad k = 1, 2, \dots$$

5.3 ポアソン分布

確率変数 X がパラメータ $\lambda > 0$ のポアソン分布に従うとは,

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

例題 5.2 $\lambda = 0.5$, $\lambda = 1$, $\lambda = 2$ のときのポアソン分布を図示せよ.

例題 5.3 (栗ようかんに入っている栗の個数) 栗ようかん 1 本には平均 3 個の栗が入っているという. 栗ようかん 1 本にまったく栗が入っていない確率, 栗が 5 個以上入っている確率を求めよ.

定理 5.4 (ポアソンの少数の法則) 二項分布 $B(n, p)$ は $np = \lambda$ を保ったまま, $n \rightarrow \infty, p \rightarrow 0$ とすると, パラメータ λ のポアソン分布に収束する.

例題 5.5 50名のクラスで4月1日生まれの子は何人くらいいるだろうか? 1年を365日として, どの日に生まれる確率も同じと仮定すれば, 4月1日生まれの子の人数 X は $B(40, 1/365)$ に従う.

$$P(X = k) = \binom{40}{k} \left(\frac{1}{365}\right)^k \left(\frac{364}{365}\right)^{40-k}, \quad k = 0, 1, 2, \dots$$

平均値 $np \approx 0.1096$ に注意して, パラメータ 0.1096 のポアソン分布で近似すると,

$$P(X = k) = \frac{0.1096^k}{k!} e^{-0.1096}$$

で与えられる.

k	0	1	2	3	...
近似値	0.8962	0.0982	0.0054	0.0002	...
厳密値	0.8961	0.0985	0.0053	0.0002	...

5.4 平均値と分散

確率分布	平均値 (m)	分散 (σ^2)
ベルヌイ分布 (2点分布) $B(1, p)$	p	$p(1-p)$
二項分布 $B(n, p)$	np	$np(1-p)$
幾何分布 (パラメータ p)	$(1-p)/p$	$(1-p)/p^2$
ポアソン分布 (パラメータ λ)	λ	λ

定理 5.6 (確率母関数) $\{0, 1, 2, \dots\}$ に値をとる離散型確率変数 X に対して,

$$G(z) = \sum_{k=0}^{\infty} z^k P(X = k)$$

を X の (または X の確率分布の) 確率母関数という. このとき,

$$\mathbf{E}(X) = G'(1), \quad \mathbf{E}(X^2) = G''(1) + G'(1), \quad \mathbf{V}(X) = G''(1) + G'(1) - G'(1)^2.$$

例題 5.7 二項分布の平均値と分散を求めよ.

宿題 14 母関数を用いてポアソン分布の平均値と分散を求めよ.



第6章 基本的な確率分布

(II) 連続分布

6.1 一様分布

区間 $[a, b]$ からどの点も同等な確からしきで1点を選ぶときのモデルとして現れる.

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{その他} \end{cases}$$

6.2 指数分布

$\lambda > 0$ を定数として

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

ランダム到着の待ち時間をモデル化するときに現れる.

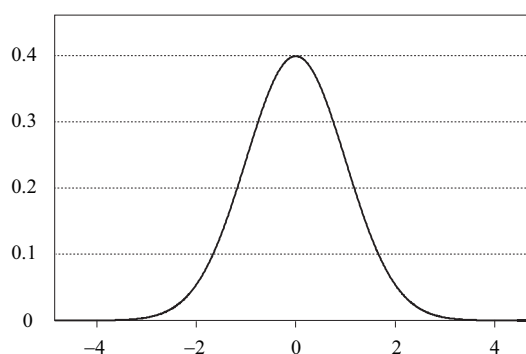
6.3 正規分布 (ガウス分布)

$N(m, \sigma^2)$: 平均 m , 分散 σ^2 の正規分布 (またはガウス分布)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x-m)^2}{2\sigma^2} \right\}$$

$N(0, 1)$: 標準正規分布

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



6.4 平均値と分散

確率分布	平均値 (m)	分散 (σ^2)
$[a, b]$ 上の一様分布	$(a + b)/2$	$(b - a)^2/12$
指数分布 (パラメータ λ)	$1/\lambda$	$1/\lambda^2$
正規分布 $N(m, \sigma^2)$	m	σ^2

宿題 15 積分を計算して, 上の表を確認せよ. ただし, 正規分布に関しては, 次の公式 (重積分の応用として有名) を用いよ.

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

6.5 標準正規分布

例題 6.1 確率変数 Z の分布が標準正規分布である (このことを $Z \sim N(0, 1)$ と書く) とする. 標準正規分布表を用いて,

- (1) $P(Z \leq 1.15)$ と $P(Z \leq -1.23)$ を求めよ.
- (2) 次の等式が成り立つような a を求めよ.

$$P(Z \geq a) = 0.33, \quad P(Z < a) = 0.75, \quad P(|Z| \geq a) = 0.4$$

定理 6.2 (標準化あるいは規準化) $X \sim N(m, \sigma^2)$ のとき,

$$ax + b \sim N(am + b, a^2\sigma^2), \quad \text{特に, } Z = \frac{X - m}{\sigma} \sim N(0, 1)$$

例題 6.3 $X \sim N(2, 5^2)$ のとき, $P(X \leq 0)$ と $P(|X| \geq 4)$ を求めよ.

宿題 16 (1) 確率変数 X が正規分布 $N(20, 4^2)$ に従うとき, $P(X > 17.8)$ を求めよ.

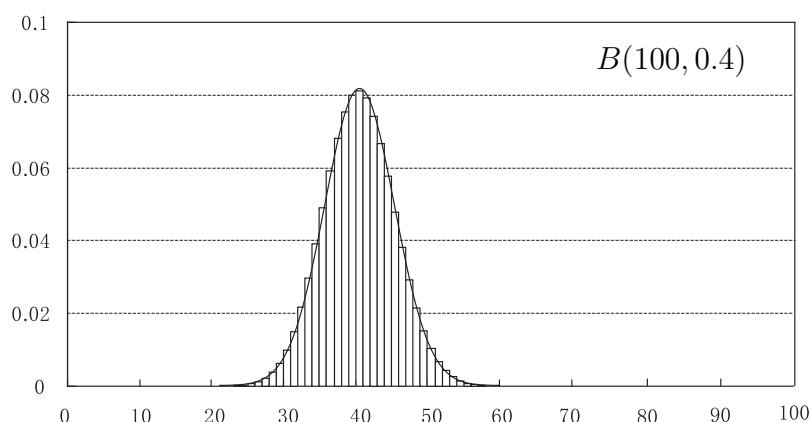
- (2) $X \sim N(50, 10^2)$ のとき, $P(X > a) = 0.985$ を満たす a を求めよ.

宿題 17 大規模な選抜試験が実施され, 上位 5% が合格となる. 試験の結果, 平均点は 68 点, 標準偏差は 8 点であった. 受験者全体の得点分布は正規分布であると仮定して, 合格するための最低点を求めよ.

補足 表にない値は (線形) 補間法で求める. つまり, $x = x_1$ のとき $y = y_1$, $x = x_2$ のとき $y = y_2$ はわかっているとき, $x_1 < x < x_2$ に対する y 値は次式で与える:

$$y = \frac{y_2 - y_1}{x_2 - x_1} (x - x_1) + y_1$$

6.6 ドモアブル・ラプラスの定理



定理 6.4 二項分布は, 同じ平均と分散をもつ正規分布で近似できる.

$$B(n, p) \approx N(np, np(1 - p)), \quad 0 < p < 1, \quad n \rightarrow \infty.$$

例題 6.5 公平なコインを 400 回投げたとき, 表が 225 回以上出る確率を正規分布近似を用いて求めよ (半目補正に注目). このコインは公正であるといえるだろうか? [0.0071]

宿題 18 ある大学では過去のデータによると入学試験の合格者の内, 入学を辞退するものが 4% いるという. 1000 人の定員のところ 1050 人合格としたとき, 定員割れを起こす確率を求めよ. [0.0901]

6.7 演習問題

演習 14 (二項分布のモード) 二項分布 $B(n, p)$ に従う確率変数 X に対して, $P(X = k)$ が最大になるような k を求めよ. [$P(X = k)/P(X = k - 1)$ を調べるとよい.]

演習 15 サイコロを 1 回振り, 引き続いて, 出た目と同じ回数だけコインを投げたとき表の枚数を X とする. X の平均値と分散を求めよ.

演習 16 サイコロを 600 回投げたとき, 1 の目が 120 回以上出る確率を求めよ.

演習 17 $X \sim N(0, 1)$ のとき, X^2 の分布関数

$$F(x) = P(X^2 \leq x)$$

を正規分布を用いて表示せよ. その表示を x で微分することで, X^2 の密度関数が

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

で与えられることを示せ.

標準正規分布表 $I(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-x^2/2} dx$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4773	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

第7章 条件付確率



7.1 条件付確率

定義 7.1 A, B を2つの事象とする. $P(A) > 0$ のとき,

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

を A の下での B の条件付確率という. (事象 A が起こったことを知った上で事象 B が起こる確率と解釈される.)

例題 7.2 サイコロを2個振って出る目のうち大きい方を X , 小さい方を Y とする. ただし, 同じ目が出た場合は $X = Y$ とする. 次の条件付確率を求めよ.

$$P(X \geq 5|Y = 2), \quad P(X + Y \geq 8|X \leq 4),$$

例題 7.3 (くじ引き) 箱の中に10本の三角くじが入っていて, そのうち2本が当たりとなっている. 2人が順番に1本ずつくじを引くとき, 先に引くのが有利か, 後のほうが有利か? 条件付確率を用いて考察せよ.

例題 7.4 (無記憶性) (1) T を幾何分布に従う確率変数とすると, $m, n = 0, 1, 2, \dots$ に対して, $P(T \geq m + n | T \geq m) = P(T \geq n)$ が成り立つ.

(2) X を指数分布に従う確率変数とすると, $a, b \geq 0$ に対して $P(X \geq a + b | X \geq a) = P(X \geq b)$ が成り立つ.

宿題 19 2つの事象 E, F に対して, $P(E) = \frac{1}{3}$, $P(F) = \frac{1}{2}$, $P(E \cup F) = \frac{2}{3}$ がわかっている. 次の確率を求めよ.

$$P(E^c), \quad P(E \cap F^c), \quad P((E \cup F)^c), \quad P(E|F), \quad P(E|F^c), \quad P(E \cap F|E \cup F)$$

7.2 事象の独立性

定義 7.5 2つの事象 A, B が**独立**であるとは,

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

を満たすときにいう. 事象の有限または無限列 $A_1, A_2, \dots$ が**独立**であるとは, そこから取り出した任意有限個の事象 $A_{i_1}, \dots, A_{i_n}$ に対して,

$$P(A_{i_1} \cap \dots \cap A_{i_n}) = P(A_{i_1}) \cdots P(A_{i_n})$$

が成り立つときにいう.

例題 7.6 壺の中に 112, 121, 211, 222 という番号の着いた 4 個の球が入っている. この壺から 1 個の球を取り出して番号を読むとき, 1 位の数字が 1 である事象を A_1 , 10 位の数字が 1 である事象を A_2 , 100 位の数字が 1 である事象を A_3 とする. A_1, A_2, A_3 のいずれの 2 つも独立であるが, 3 つの事象は独立ではない.

宿題 20 A, B, C が独立で, $P(A) = a, P(B) = b, P(C) = c$ とする. 次の事象の確率を a, b, c を用いて表せ.

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

7.3 ベイズの公式

$\Omega = A_1 \cup A_2, A_1 \cap A_2 = \emptyset$ のとき, 任意の事象 B に対して,

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}$$

「結果から原因を知る公式」としての解釈 (事前確率の設定に注意).

例題 7.7 ある国では, 病気 A の感染者は 500 人に 2 人の割合であるという. 検査 B は, 感染者の 95% に陽性反応を示すが, 非感染者の 2% にも陽性反応が出てしまう. ある人がこの検査を受けて陽性反応が出た. この人が感染者である確率を求めよ. 次に, 非感染者の 100 p % にも陽性反応が出ると一般化して, この検査を受けて陽性反応が出た人が感染者である確率が p とともにどのように変化するかを考察せよ.

宿題 21 ある工場では同じ部品を 3 つのメーカー A_1, A_2, A_3 から仕入れる. 仕入れの割合は, A_1 から 30%, A_2 から 20%, A_3 から 50% である. これまでの経験から, A_1, A_2 の部品は 2% が不良で, A_3 からの部品は 1% が不良である. 今, ランダムに取り出した 1 個の部品が不良であったとき, この部品はどのメーカーのものか, 確率を求めて答えよ.

宿題 22 (条件付き確率は直感にあわないかも) 1 から 10 の番号が付いている 10 枚のチケットがある. このうち 1 番と 2 番が当たりくじとなっている. 一郎は 4 枚のチケットを買った.

- (1) 一郎は 1 番をもっていると告げた. このとき, 残りの 6 枚にあたりが残っている確率を求めよ.
- (2) 一郎は少なくとも 1 枚のあたりをもっていると告げた. このとき, 残りの 6 枚にあたりが残っている確率を求めよ.



Jacob Bernoulli (1655–1705)

第8章 標本分布

8.1 標本抽出

調査対象の集団 (母集団) に対して, 全数調査が不可能である場合に, その一部分 (標本) を調査して全体の性質を推定することが重要である. 標本を1個取り出せば, 観測値 x が1つ得られる. 観測値 x は取り出された標本ごとに違った値となるのは当然であるが, 母集団をよくかき混ぜて (無作為に) 標本を選ぶのであれば, 値 x の現れ方に母集団分布が反映する. つまり, 母集団分布に従う確率変数を X としたとき, 観測値 x を確率変数 X の**実現値**とみなされる. この観点にもとづいて統計的推測の理論が構築される.

無作為復元抽出 複数の標本を取り出すときに, 取り出した標本を母集団に戻しながら, 無作為に次々に標本を取り出す仕方を無作為復元抽出という. i 回目 to 得られた観測値を x_i とすれば, n 回の標本抽出によって n 個の観測値

$$x_1, x_2, \dots, x_n$$

が得られる. 上に述べたように各回の観測値は, 母集団分布に従う確率変数の実現値とみなされる. i 回目の標本抽出に対応する確率変数を X_i とすれば,

$$X_1, X_2, \dots, X_n$$

は母集団分布に従う**独立で同分布をもつ確率変数列**になる. これを母集団から得られた**大きさ n の無作為標本**ともいう.

非復元抽出 毎回の標本調査のあと母集団が変化するので事情は複雑になる. しかしながら, 母集団が標本の大きさに比べて大きければ「非復元抽出 $\approx$ 復元抽出」と考えてよい. つまり, 母集団が巨大なら n 個の無作為標本を得たいときに, まとめて n 個を取り出しても誤差は実用上無視してよい.

母数の推定 母集団分布そのものを標本調査によって推定することは困難な問題であり, 実用上知りたいのは母集団分布を特徴づける統計量やパラメータである. そのような量を**母数**と総称する. 特に, 母集団分布の平均値を**母平均**, 分散を**母分散**と呼ぶ. 母平均や母分散などの基本的な母数の推定がこれからのメインテーマである.

8.2 独立同分布の確率変数列

定義 8.1 確率変数列 $X_1, X_2, \dots$ が独立であるとは, すべての $n = 1, 2, 3, \dots$ とすべての実数 $x_1, x_2, \dots, x_n$ に対して

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$$

が成り立つときに言う. 離散型の場合は, $X_i \leq x_i$ を $X_i = x_i$ に置き換えてよい.

公式 一般に, 2つの確率変数 X, Y と定数 α に対して,

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y), \quad \mathbf{E}(\alpha X) = \alpha \mathbf{E}(X), \quad \mathbf{V}(\alpha X) = \alpha^2 \mathbf{V}(X).$$

定理 8.2 確率変数列 $X_1, X_2, \dots, X_n$ が独立であるとする.

- (1) [平均値の乗法性] $\mathbf{E}(X_1 X_2 \cdots X_n) = \mathbf{E}(X_1) \mathbf{E}(X_2) \cdots \mathbf{E}(X_n)$
- (2) [分散の加法性] $\mathbf{V}(X_1 + X_2 + \cdots + X_n) = \mathbf{V}(X_1) + \mathbf{V}(X_2) + \cdots + \mathbf{V}(X_n)$

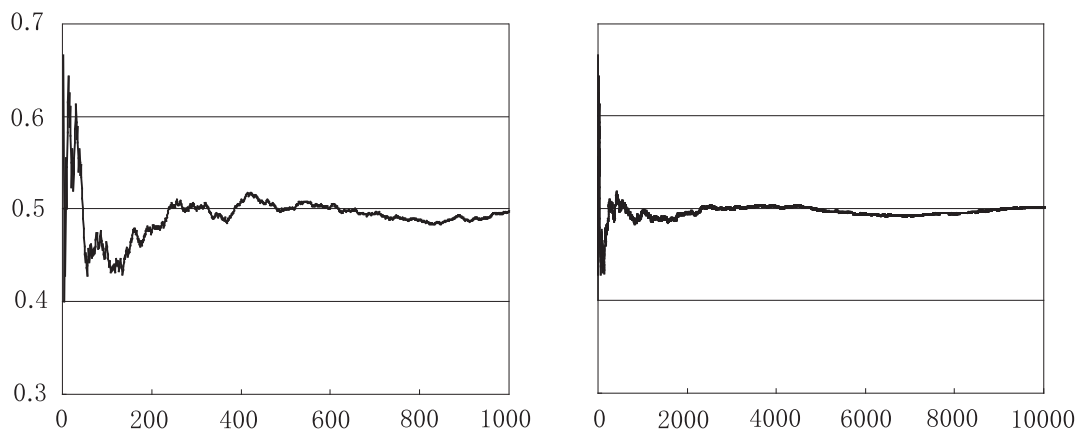
例題 8.3 サイコロを2個投げて出た目の積を X とする. X の平均値と分散を求めよ.

8.3 大数の法則

例題 8.4 (コイン投げのシミュレーション) コイン投げの結果, 表なら $x = 1$, 裏なら $x = 0$ として観測値を定める. コインを投げ続けて, その結果を $x_1, x_2, \dots$ とするとき,

$$t_n = \frac{1}{n} \sum_{k=1}^n x_k$$

は初めの n 回のうち表の出た相対頻度である. 左図は最初の1000回まで, 右図は10000回までの t_n の変化の様子をグラフにしたものである.



定理 8.5 (大数の弱法則) $X_1, X_2, \dots$ を独立で同分布をもつ確率変数列とし, その平均値を m , 分散を σ^2 とする. このとき, 任意の $\epsilon > 0$ に対して次が成り立つ:

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{k=1}^n X_k - m \right| > \epsilon \right) = 0$$

定理 8.6 (チェビシェフの不等式) X を有限な分散 $V[X]$ をもつ確率変数とする. このとき, 任意の $\delta > 0$ に対して, 次が成り立つ:

$$P(|X - \mathbf{E}[X]| \geq \delta) \leq \frac{V[X]}{\delta^2}$$

8.4 中心極限定理

定理 8.7 $X_1, X_2, \dots$ を独立で同分布をもつ確率変数列とし, その平均値を m , 分散を σ^2 とする. このとき,

$$\lim_{n \rightarrow \infty} P \left(a \leq \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - m}{\sigma} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx, \quad a < b.$$

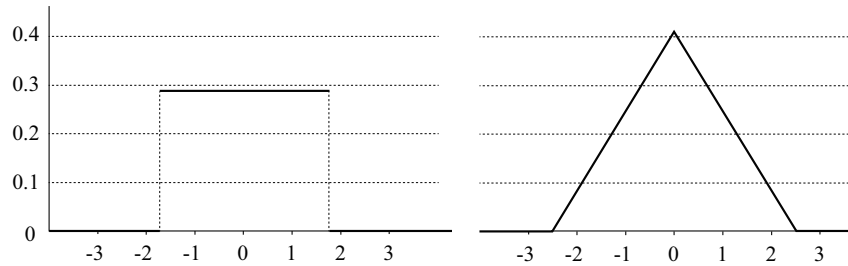
つまり, $\frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - m}{\sigma}$ の分布は, n が大きいとき近似的に標準正規分布 $N(0, 1)$ に従う. したがって,

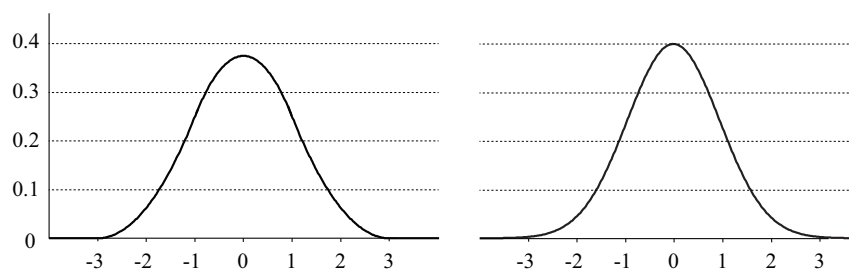
$$\frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - m}{\sigma} \approx N(0, 1) \iff \bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \approx N \left(m, \frac{\sigma^2}{n} \right)$$

注意 $X_k \sim N(m, \sigma^2)$ のとき, 任意の n に対して $\frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - m}{\sigma} \sim N(0, 1)$ (厳密).

例題 8.8 $X_1, X_2, \dots$ を $[-\sqrt{3}, \sqrt{3}]$ 上の一様分布に従う独立同分布な確率変数列とする.

$$Y_1 = X_1, \quad Y_2 = \frac{1}{\sqrt{2}}(X_1 + X_2), \quad Y_3 = \frac{1}{\sqrt{3}}(X_1 + X_2 + X_3), \quad \dots, \quad N(0, 1)$$

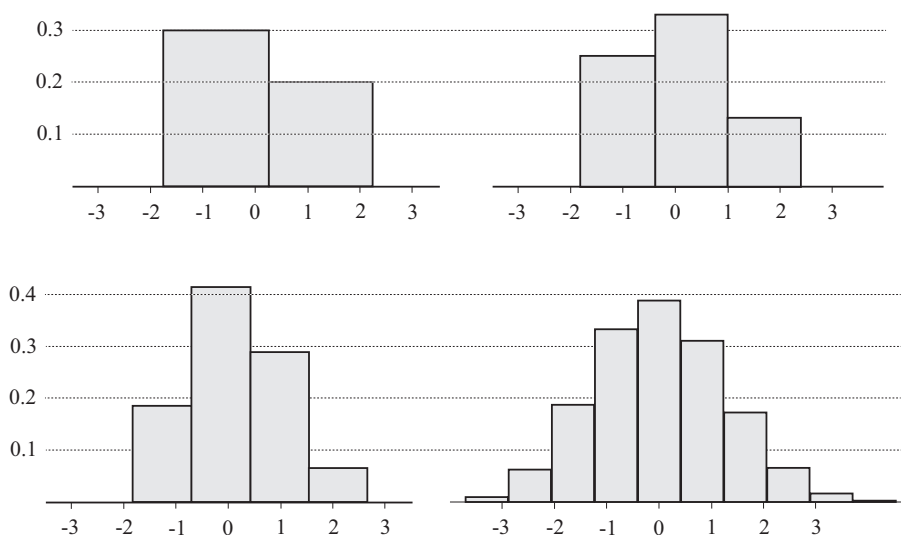




例題 8.9 $X_1, X_2, \dots$ を成功確率 $p = 0.4$ のベルヌーイ分布に従う独立同分布な確率変数列とし, $\tilde{X}_k$ をその標準化とする.

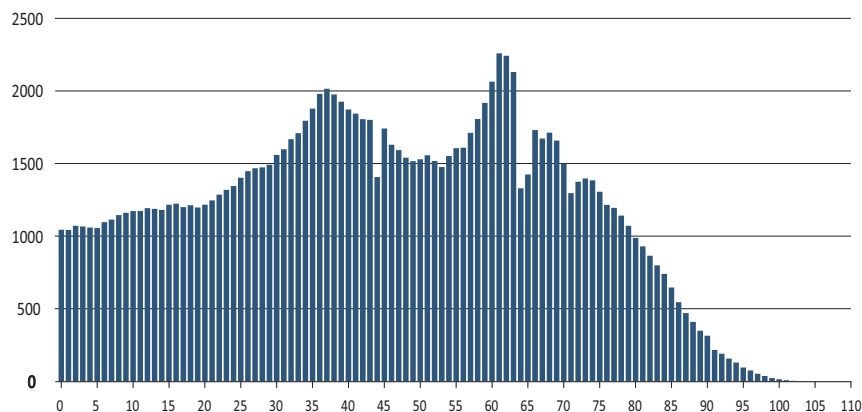
$$Y_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n \tilde{X}_k, \quad \tilde{X}_k = \frac{X_k - 0.4}{\sqrt{0.24}}.$$

グラフは $n = 1, 2, 3, 10$ に対する Y_n の確率分布を示している.



宿題 23 日本人の平均年齢は 44.5 歳, 標準偏差は 23.5 歳である (平成 22 年 10 月). 25 人の日本人を無作為に選んだとき, その 25 人の平均年齢の分布はどうなるか? あるサークルのメンバー 25 名の平均年齢は 32 歳であるが, このサークルは日本人の無作為標本といえるだろうか?

平成 22 年国勢調査人口等基本集計 (総務省統計局)





第9章 母平均の推定

9.1 視聴率調査

テレビ局では視聴率の獲得にしのぎを削っているようである。果たして、コンマ以下の数字に意味はあるのだろうか？

2014年 5月19日(月)～5月25日(日) ドラマ(関東地区) 視聴率ベスト10

番組名	放送局	放送日 放送開始時刻 - 分数	視聴率* (%)
連続テレビ小説・花子とアン	NHK総合	14/05/22(木) 8:00 - 15	24.0
木曜ドラマ・BORDER	テレビ朝日	14/05/22(木) 22:00 - 54	16.7
花咲舞が黙ってない	日本テレビ	14/05/21(水) 22:00 - 60	16.0
日曜劇場・ルーズヴェルト・ゲーム	TBS	14/05/25(日) 21:00 - 54	16.0
軍師官兵衛	NHK総合	14/05/25(日) 20:00 - 45	14.8
土曜ワイド劇場・検事・朝日奈耀子15	テレビ朝日	14/05/24(土) 21:00 - 126	11.9
金曜ドラマ・アリスの棘	TBS	14/05/23(金) 22:00 - 54	11.5
木曜劇場・続・最後から二番目の恋	フジテレビ	14/05/22(木) 22:00 - 54	11.2
月曜ゴールデン特別企画・心に響くサスペンス・隣の女	TBS	14/05/19(月) 21:00 - 114	10.7
ビター・ブラッド・最悪で最強の親子刑事	フジテレビ	14/05/20(火) 21:00 - 54	10.4

* ビデオリサーチ社による番組平均世帯視聴率

日本の放送エリアは全部で32ありますが、それぞれの放送エリアごとに視聴率調査が行なわれています。ビデオリサーチでは、関東地区をはじめ全国27地区の調査エリアで、PMシステムによる調査とオンラインメータシステムによる調査を実施しています。（日本全国をひとつの調査エリアとした視聴率調査は実施していません）また、調査対象世帯数は、PMシステムによる調査の関東地区・関西地区・名古屋地区で600世帯、それ以外のオンラインメータシステムによる調査地区は200世帯です。（ビデオリサーチ社のウェブページから、2014.6現在）

参考：藤平芳紀「視聴率の正しい使い方」（朝日新書）

9.2 母数の推定 — 推測統計学の中心課題

母集団から取り出した無作為標本 $X_1, X_2, \dots, X_n$ から母集団の統計的な性質を推測する。ここで、母集団の統計的な性質とは、典型的には母集団の統計量（平均値・分散・相関係数...）や母集団分布を特徴づけるパラメータ（ポアソン分布の λ 、一様分布の a, b 、正規分布の $m, \sigma^2, \dots$ ）であり、**母数**と総称する。

点推定 無作為標本の関数 $T = T(X_1, X_2, \dots, X_n)$ によって母数 θ を推定する方法。

区間推定 無作為標本からある区間を求めて、母数 θ がその区間（信頼区間）にどのくらいの確率（信頼係数）で含まれるかを示す推定法。誤り確率を明示しているという意味で重用される。

9.3 母平均の点推定

$X_1, X_2, \dots, X_n$: 母集団 (母平均: m ; 母分散: σ^2) から取り出した無作為標本

定義 9.1 (標本平均)

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$$

標本平均が母平均の推定量 (推定値) として妥当な理由

- (i) [不偏性] $\mathbf{E}[\bar{X}] = m$ ($\bar{X}$ は 母平均 m の周りに分布する)
- (ii) [一致性] $P\left(\lim_{n \rightarrow \infty} \bar{X} = m\right) = 1$ ($X_1, X_2, \dots$ は独立同分布確率変数列)
 $\implies$ 大数の法則 $\implies n$ が大きいほど, $\bar{X}$ は高い確率で m に近い

● しかし, これだけでは弱い. 加重平均やメディアンではいけないのか?

定理 9.2 $a_k \geq 0$ かつ $\sum_{i=1}^n a_k = 1$ を満たす定数に対して, 加重平均が次式で定義される:

$$T = \sum_{k=1}^n a_k X_k$$

- (1) $\mathbf{E}[T] = m$ を満たす.
- (2) $\mathbf{E}[(T - m)^2] \geq \mathbf{E}[(\bar{X} - m)^2]$ で等号は $a_1 = \dots = a_n = 1/n$ のときに限り成り立つ. このことを「標本平均は加重平均の中で最も有効である」という.

注意 メディアンの統計的性質は複雑であり, より上級の統計学の内容となる. 大雑把にいうと, メディアンも不偏性を満たすが, 有効性で標本平均より劣る. しかしながら, 母集団分布が「すそ野が厚い」場合は, 高確率で外れ値が現れやすく, 標本平均がなかなか母平均に近づかないため, メディアンによる推定が実用上優れている場合もある.

9.4 母平均の区間推定 (母分散が既知)

標本平均の分布 (中心極限定理) 母平均 m , 母分散 σ^2 の母集団から取り出した無作為標本 $X_1, \dots, X_n$ の標本平均 $\bar{X}$ は近似的に次を満たす.

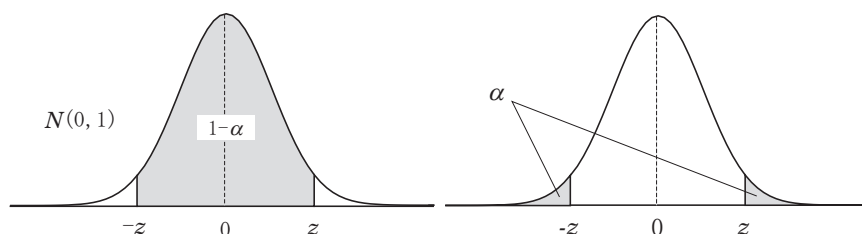
$$\bar{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$$

定義 9.3 (標準正規分布の両側 α 点) $Z \sim N(0, 1)$ (標準正規分布) とする. $\alpha > 0$ に対して,

$$P(|Z| \geq z) = \alpha \iff P(-z \leq Z \leq z) = 1 - \alpha$$

を満たす z を標準正規分布の両側 α 点といい, $z(\alpha)$ と書く (文献によっては, 上側 α 点を $z(\alpha)$ と書いているものもあるので注意).

$z = z(\alpha)$	1.00	1.64	1.96	2.00	2.33	2.58	3.00	3.29
α	0.317	0.100	0.050	0.045	0.020	0.010	0.003	0.001
$1 - \alpha$	0.683	0.900	0.950	0.955	0.980	0.990	0.997	0.999



定義 9.4 (信頼区間) $0 < \alpha < 1$ とする. 母平均 m に対する信頼係数 $1 - \alpha$ の信頼区間が

$$\left[\bar{X} - z \frac{\sigma}{\sqrt{n}}, \bar{X} + z \frac{\sigma}{\sqrt{n}} \right],$$

によって定義される. ただし, z は標準正規分布 $N(0, 1)$ の両側 α 点 (= 上側 $\alpha/2$ 点) である. 区間の端点を**信頼限界**と呼ぶ.

信頼係数として習慣的に用いられる値

90% ($\alpha = 0.1, z = 1.64$) 95% ($\alpha = 0.05, z = 1.96$) 99% ($\alpha = 0.01, z = 2.58$)

信頼区間の意味 標本調査の結果, 観測値 $x_1, \dots, x_n$ が得られたとする. 標本平均 $\bar{x}$ をいつも通り計算し, 上の公式を用いると信頼区間が得られる. 信頼区間は母平均を含んでいるか含んでいないかどちらかである. このことは標本調査ごとに変化するが, $1 - \alpha$ の確率で信頼区間は母平均を含み, α の確率で含まない. 「信頼区間の中点が母比率に近い確率が高い」とか「信頼区間の端の方は母比率から外れている確率が高い」などというのは理論を知らないことさらしているだけだが, 世間には意外と多いので注意.

α	1	大	小	0
信頼係数 ($1 - \alpha$)	0%	小	大	100%
信頼区間の幅	0	小 (シャープな推定)	大 (アバウトな推定)	∞

例題 9.5 ある生産ラインで1万個の製品を作った. ランダムに選んだ40個の製品の平均重量は156gであった. この生産ラインの機械的特性から, 生産される製品の重量の標準偏差は8gである. 生産した1万個の製品の平均重量の信頼区間を求めよ. [95% 信頼区間は 156 ± 2.48]

例題 9.6 例題 9.5 で, 95%信頼区間の幅を1g以下にするためには何個の標本をとる必要があるか? [984]

宿題 24 ある工場のロットから, ランダムに200個の標本を選んで不純物量を測定したとき, 平均2.2gの不純物が含まれていた. この工場の工程から, 不純物量の標準偏差は1.5gであることが経験的に知られている. このロット全体では, 不純物を平均何g含んでいるといえるだろうか? 信頼区間を求めよ. [1.992, 2.408]

9.5 二項母集団の母比率の推定

二項母集団：ある属性 E によって2つの集団に分かれているような母集団.

母比率: 属性 E をもつ集団の比率

$X_1, X_2, \dots, X_n$: 大きさ n の無作為標本

$$X_i = \begin{cases} 1, & i \text{ 番目の標本が属性 } E \text{ をもつ,} \\ 0, & i \text{ 番目の標本が属性 } E \text{ をもたない} \end{cases}$$

$$\hat{p} = \frac{1}{n} \sum_{k=1}^n X_k$$

$\hat{p}$ は一般の文脈では標本平均と呼ばれるが、二項母集団の文脈では**標本比率**と呼ぶ.

二項母集団における母比率の信頼区間 母比率 p に対する信頼係数 $1 - \alpha$ の信頼区間

$$\left[\hat{p} - z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

(1) 母比率 p の二項母集団では、母平均 $= p$, 母分散 $= \sigma^2 = p(1-p)$.

(2) 大きさ n の無作為標本の標本比率について $\hat{p} \sim N\left(p, \frac{\sigma^2}{n}\right) \iff \frac{\hat{p}-p}{\sigma/\sqrt{n}} \sim N(0,1)$

(3) 2次不等式の近似 (詳細は教科書)

$$|\hat{p} - p| \leq z \sqrt{\frac{p(1-p)}{n}} \approx |\hat{p} - p| \leq z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

例題 9.7 (視聴率調査) 標本数 600 から視聴率の推定値 21% が得られた. 信頼係数 95% の信頼区間は,

$$0.21 \pm 1.96 \times \sqrt{\frac{0.21(1-0.21)}{600}} \approx 0.21 \pm 0.033$$

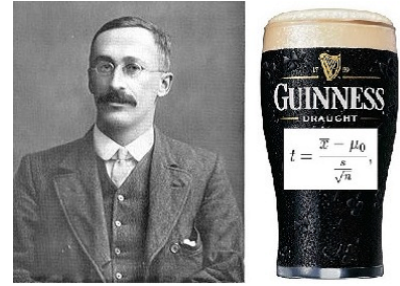
例題 9.8 視聴率調査において、信頼係数 95% の信頼区間の長さが 0.01 以下になるためには、どれほどの標本数が必要か? 信頼係数 90% ではどうか?

宿題 25 商店街のスピードくじを 100 回引いたところ、12 本のあたりを引いた. このスピードくじに含まれている当たりくじの比率の信頼区間を求めよ. [90% 信頼区間は 0.12 ± 0.053]

宿題 26 視聴率調査結果について、信頼区間を求め、その順位について考察せよ.

宿題 27 「就職状況調査によると、2013 年 3 月に卒業を迎えた大学生の就職率は、前年比で 0.3 ポイント改善の 93.9 % だったことが明らかになりました. 調査は全国の大学などから抽出した 112 校・6250 人を対象に実施しました」(厚生労働省および文部科学省 平成 25 年 5 月 17 日発表) 就職率の 95 % 信頼区間を求めよ.

第10章 母数の推定



William Sealy Gosset (1876–1937)

10.1 母分散の点推定

母平均 m , 母分散 σ^2 の母集団から取り出した n 個の標本を $X_1, \dots, X_n$ とするとき,

$$U^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

前者を**不偏分散**, 後者を**標本分散**という. (文献によっては, 前者も標本分散と呼んでいるので, いささか混乱するので注意せよ)

定理 10.1 不偏分散 U^2 は不偏性を満たす: $\mathbf{E}(U^2) = \sigma^2$.

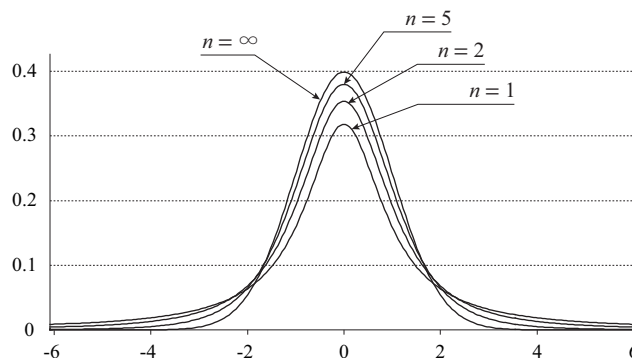
標本分散は不偏性を満たさないで, 母分散の推定量としては不偏分散が優れている. 標本数 n が大きくなれば, S^2 と U^2 の差はわずかであるが, n が小さいときは差は無視できない (t 分布の方法から小標本論へ).

10.2 t 分布

定義 10.2 (自由度 n の t 分布) 密度関数

$$\frac{1}{\sqrt{n} B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n} \Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{1}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

によって与えられる確率分布を自由度 n の t 分布 (またはスチューデントの t 分布) といい, t_n で表す.



(1) Γ はガンマ関数.

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt, \quad a > 0.$$

関数等式 $\Gamma(a+1) = a\Gamma(a)$ が成り立つ. 特に, 自然数 n に対して $\Gamma(n) = (n-1)!$ となる.

(2) B はベータ関数.

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad a > 0, b > 0.$$

(3) $n=1$ の t 分布には平均値が存在しないが, $t \geq 2$ の t 分布の平均値は 0.

(4) 自由度 $n = \infty$ の t 分布は標準正規分布 $N(0, 1)$ を意味する.

(5) 実用上, $n \geq 30$ で標準正規分布 $N(0, 1)$ で代用.

宿題 28

$$\lim_{n \rightarrow \infty} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n} \Gamma(\frac{n}{2}) \Gamma(\frac{1}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

を示せ. [ヒント: $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ とスターリングの公式 $\lim_{z \rightarrow \infty} \frac{\Gamma(z)}{\sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z} = 1$]

定理 10.3 正規母集団 $N(m, \sigma^2)$ から取り出した n 個の標本を $X_1, \dots, X_n$ とする.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ (標本平均)} \quad U^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ (不偏分散)}$$

このとき,

$$T = \frac{\bar{X} - m}{U/\sqrt{n}} \sim t_{n-1} \quad \text{自由度 } (n-1) \text{ の } t \text{ 分布}$$

(T の分布が m や σ^2 に依存しないところが重要!)

● 正規母集団でなくとも, 標本数が大きいときは近似として成り立つ.

10.3 母平均の区間推定 (母分散未知の場合)

定義 10.4 (信頼区間) $\alpha > 0$ とする. 母平均 m , 母分散 σ^2 の母集団から取り出した n 個の無作為標本を $X_1, \dots, X_n$ とするとき, 母平均 m に対する **信頼係数** $1 - \alpha$ の **信頼区間**は

$$\left[\bar{X} - t \frac{U}{\sqrt{n}}, \bar{X} + t \frac{U}{\sqrt{n}} \right], \quad t = t_{n-1}(\alpha),$$

で与えられる. ただし, $t_{n-1}(\alpha)$ は自由度 $n-1$ の t 分布の両側 α 点である.

例題 10.5 ある薬品を精製する実験を同一条件下で8回行ったところ、生成物の重量は次のようになった。この方法で得られる生成物の平均重量の90%信頼区間を求めよ。

32.5 31.8 33.0 32.4 32.2 31.3 32.9 32.1

宿題 29 ある製品を抜き取り調査してその寿命を測定した結果、以下の数値を得た。母集団の平均寿命の95%信頼区間を求めよ。

23 42 33 29 34 41 30 36 34 28

10.4 演習問題

演習 18 (偏差値) 受験者全員の平均点を m 、標準偏差を σ とするとき、

$$(\text{偏差値}) = 50 + 10 \times \frac{x - m}{\sigma}$$

受験者数が多数の時、得点の分布は正規分布に近いと想定されることが多い。偏差値は、20 以下にも 80 以上にもなり得るが、そのような極端な値の出る確率を求めよ。

演習 19 公平なコインを1000回投げたとき、表の出る回数を X とする。 $P(480 \leq X \leq k) = 0.5$ となるような k を求めよ。

演習 20 ある国で内閣支持率を調査したところ 17.5% であった。調査対象は2000人であった。支持率の90%信頼区間を求めよ。

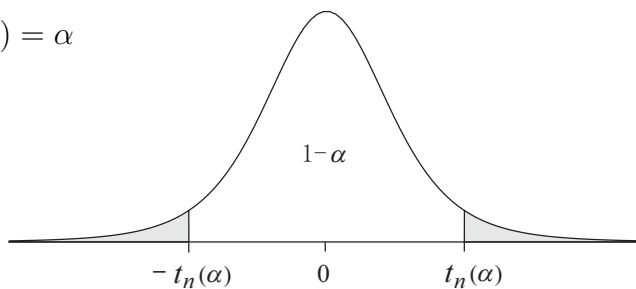
演習 21 コインを100回投げて公平か歪んでいるかを、表の回数が $50 \pm x$ 回の外に出たら歪んでいると判定することにしたい。公平なのに歪んでいると判定する誤り確率を0.01におさえるためには、 x をどのように定めればよいか？

演習 22 10個のサンプルの重量を測定したところ次の値を得た。母平均の95%信頼区間を求めよ。また、母平均の90%信頼区間を求めよ。

14.8 13.6 14.0 12.2 13.8 13.4 12.5 14.1 12.8 13.4

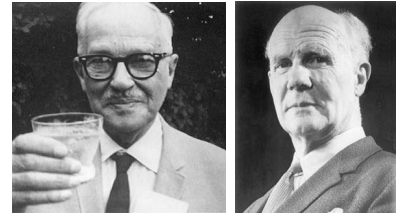
t 分布表

$$P(|T| \geq t_n(\alpha)) = \alpha$$



$n \backslash \alpha$	0.100	0.050	0.020	0.010
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
∞	1.645	1.960	2.326	2.576

第11章 仮説検定(差し替え)



Jerzy Neyman (1894–1981)

Egon S. Pearson (1895–1980)

11.1 仮説検定の方法

例題 11.1 コインを 400 回投げたとき、表が 223 回出た。コインは公平といえるだろうか？

1. 母数に関する帰無仮説 H_0 と対立仮説 H_1 を決める。
2. 関連する確率変数 T (検定統計量) を選び、仮説 H_0 の下で、この確率変数の分布を調べる。
3. 有意水準 $0 < \alpha < 1$ と棄却域 W を決める。
 - 有意水準とは、 H_0 が真なのに誤りと判定してしまう「誤り確率」のこと。慣習では 10%, 5%, 1% などが用いられる。
 - 棄却域とは、 T の実現値として稀と判断される領域で、 T がその値をとる確率がちょうど α になる ($P(T \in W) = \alpha$) ように決める。
4. 標本から T の実現値 t を計算し、 W に落ちる ($t \in W$) かどうかを判断する。
 - $t \in W$ のとき。検定統計量 T は棄却域に落ちるので、有意水準 α で有意 (H_0 から想定される揺らぎを超えて違いが見出されたという意味) である。したがって、 H_0 を棄却し H_1 を採択する。
 - $t \notin W$ のとき。検定統計量 T は棄却域に落ちないので、有意水準 α で有意ではない。したがって、 H_0 を採択する (あるいは H_0 を棄却できないという)。

コインを 400 回を投げて公平かどうかを判断する問題。このコインの表が出る確率を p とする。帰無仮説と対立仮説を

$$H_0 : p = \frac{1}{2} \quad H_1 : p \neq \frac{1}{2}$$

とする。400 回投げたときに、表の出る回数を X とする。 H_0 の下で、 $X \sim B(400, 1/2) \approx N(200, 10^2)$ 。したがって、

$$Z = \frac{X - 200}{10} \sim N(0, 1).$$

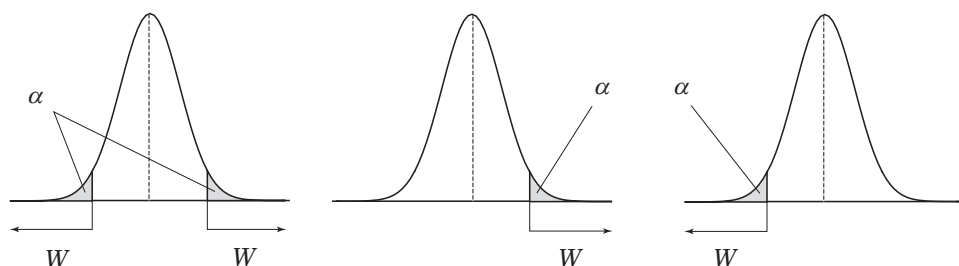
この Z を検定統計量とする。有意水準を $\alpha = 0.05$ とする。棄却域は、正規分布曲線の両側から合わせて 5% 分を切り取ればよい (両側検定)。

$$W : |z| \geq 1.96$$

となる. 観測値 223 から Z の実現値を計算すると, $z = (223 - 200)/10 = 2.3$. これは棄却域に落ちるから, H_0 を棄却する. 有意水準を $\alpha = 0.01$ とすると, 棄却域は $W : |z| \geq 2.58$ となる. Z の実現値は棄却域に落ちないから, H_0 を採択する.

標準正規分布 $N(0, 1)$ における上側 α 点

α	0.1587	0.0500	0.0250	0.0227	0.0100	0.0050	0.0013	0.0005
$z = z(\alpha)$	1.00	1.64	1.96	2.00	2.33	2.58	3.00	3.29
$1 - 2\alpha$	0.6826	0.9000	0.9500	0.9546	0.980	0.990	0.9974	0.9990



例題 11.2 (両側検定) ある機械部品の長さは規格によって 25 mm と定められている. 部品の長さの微小な狂いはやむをえないが, 規格より長すぎても短すぎても困る. ある製造ラインでは, 管理状況から, 部品の長さは標準偏差 0.8 mm の正規分布にしたがっているとしてよい. 16 個のサンプルで実際に長さを調べたところ長さの平均値は 25.45 mm であった. この製造ラインは適正に部品を作っているといえるだろうか?

例題 11.3 (片側検定) ある製造ラインで大量の製品を作っており, その重量は正規分布に従っているとしてよい. ある日に製造された大量の製品から 12 個をサンプリングして重量 (kg) を測定した結果, 平均値 $\bar{x} = 48.6$, 不偏分散 $u^2 = 1.6^2$ を得た. 規定値は 50kg であるが, 50kg を切っているときはラインを直ちに止めて調整する必要がある. この日に生産した製品の平均重量は規定に沿っているか?

宿題 30 コインが公平かどうかを確かめるために, 100 回振ったところ表が 62 回出た. このコインは公平であるといえるか.

宿題 31 ある調味料の製造ラインでは, 各製品の砂糖の含有量は $m = 60$ (g) になるように調整している. しかしながら, 原料の不均一や製造ラインの狂いなどから, m の値は 50 ~ 70 の間を変動するが, これまでの経験から標準偏差は常に一定で $\sigma = 3$ となっている (母分散既知). ある時点で, 製品を 25 個抜き取って, 調査したところ, 砂糖の含有量の平均値は 61.43 であった. その時点で製造ラインは $m = 60$ を保持していると考えてよいか?

宿題 32 ある英語の資格試験の全国平均は 66 点であった. A 塾から 10 名が受験した. 結果は

78 72 65 86 58 64 76 88 74 59

であり, その平均点 72 点が 66 点を大きく上回ると A 塾は主張している. 検定によって A 塾の主張を確認せよ. [有意水準 5% の片側検定で「上回っているとは言えない」]

11.2 2種類の過誤

帰無仮説 H_0 をめぐって、次の4つの場合がある。

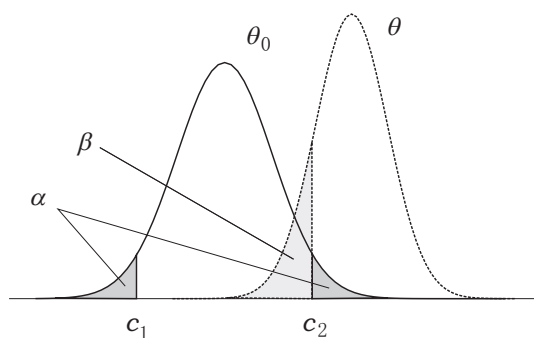
採否 \ 真偽	H_0 は真	H_0 は偽
H_0 を採択	正しい判断	第2種の誤り
H_0 を棄却	第1種の誤り	正しい判断

α : 第1種の誤り確率 = 有意水準 β : 第2種の誤り確率

第1種の誤り = 生産者危険 = “あわて者の間違い”

第2種の誤り = 消費者危険 = “ぼんやり者の間違い” = “疑陽性”

例題 11.4 コインを400回投げたとき、表が215回出た。このコインの公平性に関する仮説検定における第2種誤り確率について考察せよ。特に、 α と β は同時に小さくできないことを説明せよ。



注意 検定統計量の実現値が棄却域に落ちない場合、「 H_0 を採択する」とは言うが、はっきり否定するだけの状況ではないという消極的な採択である。「 H_0 が偽なのに採択している」誤りを犯しているかもしれない、その確率(第2種誤り確率 β)は極めて大きい可能性もあるからだ。そこで、「 H_0 を棄却できない」という表現も多用する。

11.3 P 値(ピーチ)

伝統的な仮説検定では、有意水準 α を示して H_0 の棄却・採択を述べる。が、ユーザーにとって、実現値が帰無仮説 H_0 からどのくらい外れているかを数量的に詳しく知りたいこともある。実現値 t に対して、 H_0 の下で、

P = 実現値 t を含めて、それ以上に起こりにくい実現値が得られる確率

を実現値 t の P 値という。この値をどう判断するかは、個別事情によるもので、数理統計学の枠外の話となる。

例題 11.5 コインを 10 回投げて表が 9 回出たとしよう. 表の出る確率を p として帰無仮説 $H_0: p = 1/2$ の下で検定する. 検定統計量 X として表の出る回数とする.

$$P(X = k) = \binom{10}{k} \left(\frac{1}{2}\right)^{10}$$

となる. X の実現値は $x = 9$ である. これを含めてこれ以上起こりにくいことは $\{X = 0, 1, 9, 10\}$ である. この確率が P 値である.

$$P = \left\{ \binom{10}{0} + \binom{10}{1} + \binom{10}{9} + \binom{10}{10} \right\} \left(\frac{1}{2}\right)^{10} = \frac{22}{1024} = 0.021$$

実際に起こった「表が 9 回出ること」は 2.1% で起こるくらい稀な事象であるというふうに解釈するが, この確率は「表が 9 回出る」確率ではないことに注意せよ. 伝統的な仮説検定では, P 値を有意水準と比較している. 有意水準 5% の両側検定なら $P \leq 0.05$ であることから棄却, 有意水準 1% の両側検定なら $P > 0.01$ であることから採択という判断がされる.

11.4 演習問題

演習 23 ある機械部品の寿命は規格によって 250 時間と定められている. ある製造ラインでは, 管理状況から, 部品の長さは標準偏差 2.25 時間の正規分布にしたがっているとしてよい. 25 個のサンプルで実際に長さを調べたところ長さの平均値は 248.8 時間であった. この製造ラインの部品は規格を満たしているといえるだろうか? [文脈から片側検定]

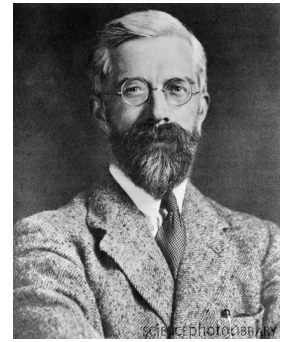
演習 24 正味 500g と書いてある製品を 120 個選んで調べたところ標本平均 498g, 不偏分散 10^2 g であった. この製品は, 明記されたとおりの内容になっているか?

演習 25 コインを 10 回投げて, コインが公正かどうかを判定する仮説検定を考える. コインを 10 回投げたうち表の出る回数を検定統計量 T として, 棄却域として $\{0, 1, 9, 10\}$ をとることにする. 第 2 種誤り確率 β を表が出る確率 p の関数として観察せよ.

定期試験

1. 7 月 23 日 (水)
2. 教科書・参考書・ノート・計算機等の持ち込み不可. 鉛筆と消しゴムだけで解答する.
3. 期末試験は 1 回だけ. 欠席者・成績不良者に対する再試験はなし. やむを得ない事情 (病気、忌引等) で定期試験を欠席し, 追試験を希望する者は正規の手続きに従って取り扱う.
4. 配布プリントの問題を中心に、概念の理解を含めてよく研究しておいてください. なお, 過去問等はウェブページに掲載している.

第12章 その他の推定・検定



Sir Ronald Aylmer Fisher (1890–1962)

12.1 母平均の差の検定

定理 12.1 2つの正規母集団 $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$ から独立に取り出した大きさ n_1 , n_2 の標本平均を $\bar{X}_1$, $\bar{X}_2$ とするとき,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

例題 12.2 ある物質の融点を測定した. 技術者 A は 5 回測定して平均 1264.6 度を得た. 技術者 B は 8 回測定して平均 1263.9 度を得た. 過去の経験によれば A の測定値の標準偏差は 0.7 度, B の測定値の標準偏差は 0.6 度である. さらに 2 人とも測定結果は正規分布に従うとしてよい. 2 人の測定結果に有意の差はあるか検定せよ.

定理 12.3 分散が等しい 2 つの正規母集団 $N(m_1, \sigma^2)$, $N(m_2, \sigma^2)$ から独立に取り出した大きさ n_1 , n_2 の標本平均を $\bar{X}_1$, $\bar{X}_2$, 不偏分散を U_1^2 , U_2^2 とする.

$$U^2 = \frac{(n_1 - 1)U_1^2 + (n_2 - 1)U_2^2}{n_1 + n_2 - 2}$$

とおけば,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) U^2}},$$

は自由度 $n_1 + n_2 - 2$ の t 分布に従う.

例題 12.4 2つの環境 A, B のもとである作物の試験栽培を行った. 環境 A からは 6 個のサンプル, 環境 B からは 8 個のサンプル をとって収穫高を調べた結果は次の通りである.

A : 6.2 6.0 5.9 6.2 6.1 5.8
B : 6.0 5.8 5.7 6.2 6.4 5.9 5.8 6.3

両者の収穫高は同じ分散をもつ正規分布に従うと仮定してよい. 環境 A, B に有意の差はあるか検定せよ.

12.2 分布の適合度検定

観測された頻度分布が理論分布と同じかどうかを検定する.

母集団の属性が $A_1, A_2, \dots, A_k$ の k 種類に分けられている. n 個の標本から, それぞれに属するものが $X_1, X_2, \dots, X_k$ 個得られたとする.

属性	A_1	A_2	$\dots$	A_k	合計
理論分布	p_1	p_2	$\dots$	p_k	1
観測度数	X_1	X_2	$\dots$	X_k	n

観測値から, 各属性の現れる理論分布 $p_1, p_2, \dots, p_k$ が妥当かどうかを検定する.

定理 12.5 $m_i = np_i$ とおくと,

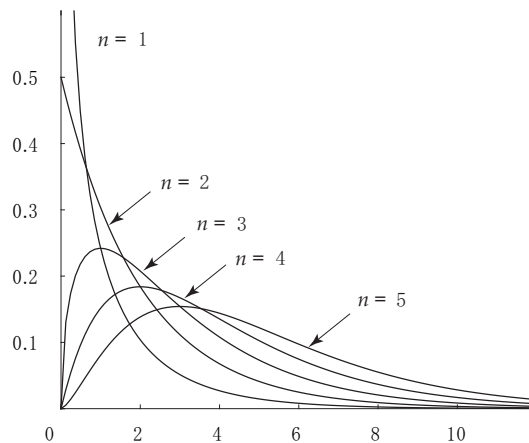
$$\chi_{k-1}^2 = \sum_{i=1}^k \frac{(X_i - m_i)^2}{m_i}$$

は, $m_1, \dots, m_k$ が大きいとき ($m_i = np_i \geq 5$), 自由度 $k-1$ のカイ 2 乗分布に近似的に従う.

密度関数が

$$f_n(x) = \begin{cases} \frac{1}{2^{n/2} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

で与えられる確率分布を **自由度 n のカイ 2 乗分布** (χ^2 -分布) という. (χ^2 は一つの文字として扱う.) 自由度を明記して, χ_n^2 と書くこともある.



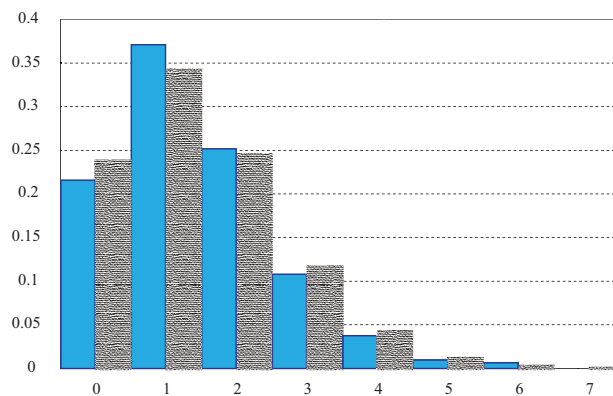
例題 12.6 次の表は, サイコロを 120 回投げて出た目を記録したものである. このサイコロは公正と言えるだろうか?

目	1	2	3	4	5	6	合計
回数	24	18	16	22	23	17	120

例題 12.7 次の表は, サッカーの試合における得点を調べた結果である (2013 年 J リーグ・ディビジョン 1・第 34 節 18 チーム 全 612 試合).

得点	0	1	2	3	4	5	6	7 以上	合計
度数	132	227	154	66	23	6	4	0	612
ポアソン分布	0.2379	0.3416	0.2453	0.1174	0.042	0.0121	0.0029	0.0006	1
同上理論予想	145.6	209.1	150.1	71.8	25.8	7.4	1.8	0.4	612

1 試合当たりの得点について平均値は 1.436, 分散は 1.367 となっている. パラメータ $\lambda = 1.436$ のポアソン分布による理論値を併記した (グラフの網掛け=ポアソン分布による理論値).



- $m_i = np_i \geq 5$ となるように得点を 0, 1, ..., 5, 6 以上の 7 クラスに分ける.
- ポアソン分布特有の事情によって, 自由度 $7 - 1 - 1 = 5$ のカイ 2 乗分布を用いる.

12.3 最大値の推定

例題 12.8 (ドイツ戦車の問題) 1 番から N 番まで通し番号のついた N 台の戦車がある. どの戦車も同等の頻度で街を巡回している. このうち, 異なる番号の n 台が目撃されたとき, その n 台の番号から, 保有台数 N を推定せよ.

$\{1, 2, \dots, N\}$ から無作為非復元抽出で取り出した n 個の標本を $X_1, \dots, X_n$ として,

$$T = \max\{X_1, X_2, \dots, X_n\}$$

とおく. T の分布は,

$$P(T = k) = \binom{N}{n}^{-1} \binom{k-1}{n-1}, \quad n \leq k \leq N.$$

これを用いて,

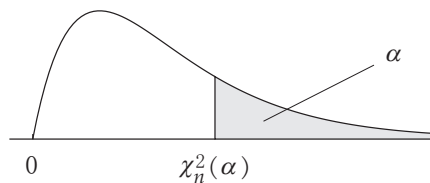
$$\mathbf{E}[T] = \sum_k kP(T = k) = \frac{n}{n+1} (N+1),$$

したがって,

$$\hat{N} = \left(1 + \frac{1}{n}\right) T - 1$$

は $\mathbf{E}[\hat{N}] = N$ を満たす. よって, $\hat{N}$ は N の不偏推定量になる.

カイ・スクエア分布: $P(\chi_n^2 \geq \chi_n^2(\alpha)) = \alpha$



$n \backslash \alpha$	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0 ⁴ 393	0.0 ³ 157	0.0 ³ 982	0.0 ² 393	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

値は小数第4位以下 ($n = 1$ では表示桁未満) を四捨五入してある。