

データ解析のための統計学基礎

茂木信宏 (Nobuhiro Moteki)

本テキストは、データ解析で日常的に使われている頻度主義統計学の方法論について扱っている。想定読者は理工系大学 3 年生から大学院生くらいまでである。とくに、区間推定と仮説検定の数学的な原理を理解し、結果を正しく解釈できるようにすることを目指している。

質問・コメントは以下メールアドレスまで

現所属（東京都立大）：moteki@tmu.ac.jp

個人：nobuhiro.moteki@gmail.com

はじめに

統計学は実験・観測・モデルのデータ解析を行うための基礎教養として必須であるものの、大学の授業・演習で十分な時間をとって学ぶ機会がない。また、統計的データ解析法についての教科書・web 教材は豊富だが、正規分布性など単純な特徴を持つ理想的なデータの解析を前提とした説明しかなされてないことが多い。

地球惑星科学の研究で扱われる複雑なデータに対しては、理想的なデータを想定して設計された統計解析法をそのまま適用してもよいのか、また、適用したときにどのくらい深刻な誤りを導きうるのかは自明ではない。またその判断基準はデータセットや解析の目的にも依存するので、基礎から丁寧に議論を進めていかないと不安と混乱は拡大するばかりである。

本演習の目標は、観測データの統計解析において、そのような不安と混乱を回避するために必要な考え方と技術を身に着けることである。統計解析の習得のためには、数学的アイデアをおおよそ把握した後、それをプログラムに実装してみることはとても有効である。プログラム化することで理解があいまいなところを漏れなく発見できるからである。演習課題を通じてその学習手順を段階的に進んでいくようにした。

本演習で扱うのは観測量のモデル化としての「確率変数」と、单变量解析における「中心極限定理」と「統計的推定（区間推定・仮説検定）」だけである。これは統計解析の基礎であり、実際に行われる統計解析の 9 割以上を占めていると考える。1 章では、確率変数の概念と算法、観測データのヒストグラム形状に影響を及ぼしている中心極限定理について扱う。統計解析においてデータの母集団分布として正規分布が暗黙に仮定されることが多いとの理由も明らかになる。2 章では、伝統的かつ定石のサブテーマである、正規母集団（確率密度関数が正規分布の母集団）の母平均についての区間推定・仮説検定を扱った。3 章では、任意母集団についての区間推定・仮説検定の方法を扱う。Permutation 検定は、2 つの任意母集団の有意差（母平均、中央値など）や相関の有無を判定するための汎用的な方法である。

0. 統計解析とは

統計解析(statistical analysis)は、観測量について知り得そうな未知パラメータと私たちが取得できる観測データとの関係を確率論をつかって数理モデル化したうえで、未知パラメータを適切なアルゴリズムで推定する手続きである。前半の数理モデル化の段階は解析者の判断に委ねられているため、データが同じであっても導かれる結論は解析者に依存して変わり得る。統計解析では、結果が解析者の主観に全く依存しないということではなく、眞の正解はない。一方、非統計解析では、入力が同じであれば出力は一意的に定まるのが普通である。確率論をつかって観測対象をモデル化する作業は統計解析に固有のものであり、それが統計解析の理解・習得の難易度を上げている。モデル化が済んでしまえば、あとは非統計解析と同じく機械的な計算を実行するだけである。モデル化の流儀は、頻度主義とベイズ主義に分けられる（図 0.1）。

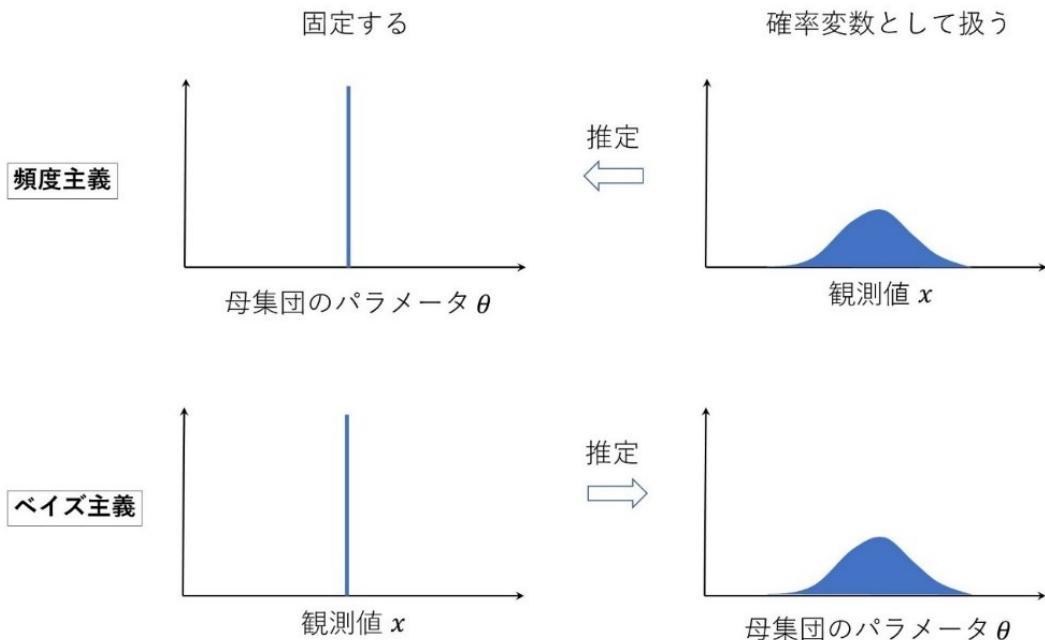


図 0.1. 統計解析における観測量と未知の母集団パラメータの因果関係のモデル化の説明図。頻度主義では、観測量を確率変数とみなし、その母集団を規定するパラメータ (θ : 母平均や母分散など) は固定された数とする立場をとる。得られた観測値は 1 試行における確率変数の実現値とする。一方、ベイズ主義では、観測値は確率変数ではなく固定された数量とみなし、母集団を規定するパラメータ (θ : 母平均や母分散など) を確率変数とみなす。

今回扱う区間推定・仮説検定は頻度主義に属する手法であるので、本資料では専ら頻度主義の立場から説明をする。頻度主義の立場では、観測データ \vec{x} を構成する N 個の観測値 $x_i (i = 1, \dots, N)$ は N 個の確率変数 $X_i (i = 1, \dots, N)$ の実現値とみなされる。

統計データ解析手法は大きく 3 つに分類される。

1. 単変量解析
2. 多変量解析
3. 時系列解析

上から下の順に扱う問題は複雑になる。

単変量解析は、考慮する観測量がただ一つしかなく、その観測量と観測時刻との関係が重要でない場合に用いる。この観測量を X とする。 X は確率変数であり、観測データ \vec{x} の各要素 $x_i (i = 1, \dots, N)$ はある同一母集団から無作為に抽出された N 個の確率変数 $X_i (i = 1, \dots, N)$ の標本であるとみなされる。実際には、データの各要素 x_i を一定時間おきに一つずつ取得していく、観測量の真の母集団は時間に依存して変化しているのかもしれない。しかし、単変量解析では、観測データ各要素と取得時刻 t との間の紐づけを解いてしまって、観測データ \vec{x} をある一つの母集団から同時に得られた N 個の標本値とみなしてから解析を行う。

多変量解析は、考慮する観測量が 2 つ以上あり、それらの観測量間の関係が重要である場合に用いる。観測量が 2 つの場合、それらを確率変数として (X, Y) とおく。データの各要素 $(x_i, y_i) (i = 1, \dots, N)$ を一定時間おきに取得していく、観測量の真の母集団は時間に依存して変化しているのかもしれない。しかし、観測量 X と観測量 Y の紐づけは保つものの、観測量と観測時刻 t との紐づけは解いてしまうものとする。つまり、観測データ (\vec{x}, \vec{y}) を 2 次元の確率密度分布を持つある一つの母集団から同時に得られた N 個の標本とみなして解析を行う。

時系列解析は、観測量と観測時刻との関係が重要である場合に用いる。観測量 X は確率変数とみなして、データの取得時刻 t には誤差がないものとする。観測データ \vec{x} の要素 x_i を一定時間おきに取得し、観測量と観測時刻 t との紐づけを解かないまま、各データ要素 x_i を時刻 t_i におけるある母集団からの標本値とみなして解析を行う。また一般に現在のデータは、過去のデータから影響を受け、かつ未来のデータに影響を及ぼす。母集団分布の時間変動もデータから推察しなければならないので、解析の難易度・多様性は、単変量・多変量解析に比べて格段に増す。時系列解析は、単変量・多変量解析を基礎としてはいるが、ほとんど独立した技術体系として扱われることが多い。実際の観測では時系列データを得ることが多いため、本資料では時系列データの話題にも少し触れる。本演習の最後の 3 回で扱う時系列解析（吉森先生担当）のテーマは、時系列データの周波数成分を抽出するフーリエ解析である。

もくじ

- 1. 確率変数 pp. 1-13
 - 1.1. 離散分布
 - 1.2. 連續分布
 - 1.3. 数値積分
 - 1.4. 統計量
 - 1.5. 乱数
 - 1.6. 中心極限定理
- 2. 区間推定と仮説検定 pp. 14-23
 - 2.1. 方法論
 - 2.2. スチューデントの T
 - 2.3. 正規母集団の母平均の推定
 - 2.3.1. 区間推定と仮説検定の共通事項
 - 2.3.2. 区間推定
 - 2.3.3. 仮説検定
 - 2.3.4. 区間推定と仮説検定の関係
 - 2.3.5. 点推定の標準誤差と区間推定の関係
 - 2.3.6. 2つの正規母集団の母平均の差の推定
- 3. 母集団が未知のときの統計解析 pp. 24-33
 - 3.1. Bootstrap 原理
 - 3.2. BS 区間推定法
 - 3.3. 2 つの未知母集団の比較(Permutation 検定)
- 付録 A-G, 参考文献 pp. 34-46

1. 確率変数

試行のたびに異なる値が実現するものの、その実現確率は決まっている変数のことを確率変数という。本資料では、確率変数はアルファベット大文字で、実現値はその小文字で表す。無限個の実現値の出現個数をヒストグラムにして総和あるいは面積を1に規格化することで得られる関数を、その確率密度関数（あるいは確率密度分布）という。確率変数、確率密度関数、確率変数の実現値をそれぞれ記号 X 、 $f(X)$ 、 x で表す。文脈から明らかな場合は確率密度関数のことを単に”分布”と呼ぶ。確率変数でモデル化された観測量 X について、とり得る全ての実現値 x の集合をその観測量の母集団という。 $f(X)$ を観測量 X の母集団分布と呼ぶこともある。

1.1. 離散分布

確率変数 X の実現値がとびとびの値(例:整数や自然数)をとる場合、その確率密度関数は離散分布となる。

離散分布の規格化条件は、

$$\sum_X f(X) = 1, \quad \cdots (1.1)$$

である。確率変数の下限 $(-\infty)$ からある値 X まで確率の和をとったものを、 $f(X)$ の累積分布関数といい記号 $F(X)$ で表す。

$$F(X) = \sum_{X'=-\infty}^X f(X'), \quad \cdots (1.2)$$

これは、確率変数が X 以下の値をとる確率を表す。規格化条件により、 $X \rightarrow \infty$ の極限で $F(X) = 1$ となる。

確率変数の期待値 $\langle X \rangle$ と分散 $\langle (X - \langle X \rangle)^2 \rangle$ は、それぞれ

$$\langle X \rangle = \sum_{X=-\infty}^{\infty} X f(X), \quad \cdots (1.3)$$

$$\langle (X - \langle X \rangle)^2 \rangle = \sum_{X=-\infty}^{\infty} (X - \langle X \rangle)^2 f(X), \quad \cdots (1.4)$$

と定義される。期待値 $\langle X \rangle$ のことを母平均(記号 μ)、分散 $\langle (X - \langle X \rangle)^2 \rangle$ のことを母分散(記号 σ^2)と呼ぶこともある。離散分布の例は下記のとおり。

離散一様分布

確率変数 X が連続する N 個の整数 i_1, i_2, \dots, i_N を等確率で実現するとき、その分布は離散一様分布

$$f(X) = \begin{cases} \frac{1}{N} & (X = i_1, i_2, \dots, i_N) \\ 0 & (X < i_1, X > i_N) \end{cases} \quad \cdots (1.5)$$

となる。離散一様分布の期待値と分散は、 $\langle X \rangle = (i_1 + i_N)/2$ ， $\langle (X - \langle X \rangle)^2 \rangle = (i_N - i_1)^2/12$ である。

二項分布

一回の試行につき、二つの結果 A, B のうちいずれかがそれぞれ確率 p , $1 - p$ で起こるとする。 N 回の試行で A, B がそれぞれ X 回、 $N - X$ 回実現する確率は、

$$f(X) = \frac{N!}{X!(N-X)!} p^X (1-p)^{N-X} = {}_N C_X p^X (1-p)^{N-X} \quad \dots (1.6)$$

となる。式(1.6)を X について下限 0 から上限 N まで和を取ると、

$$\sum_{X=0}^N f(X) = \sum_{X=0}^N {}_N C_X p^X (1-p)^{N-X} = \{p + (1-p)\}^N = 1$$

となり、式(1.6)の $f(X)$ は規格化条件を満たすので確率密度関数である。これを二項分布といふ。二項分布に従う確率変数の期待値と分散は、それぞれ $\langle X \rangle = Np$ ， $\langle (X - \langle X \rangle)^2 \rangle = Np(1-p)$ である。

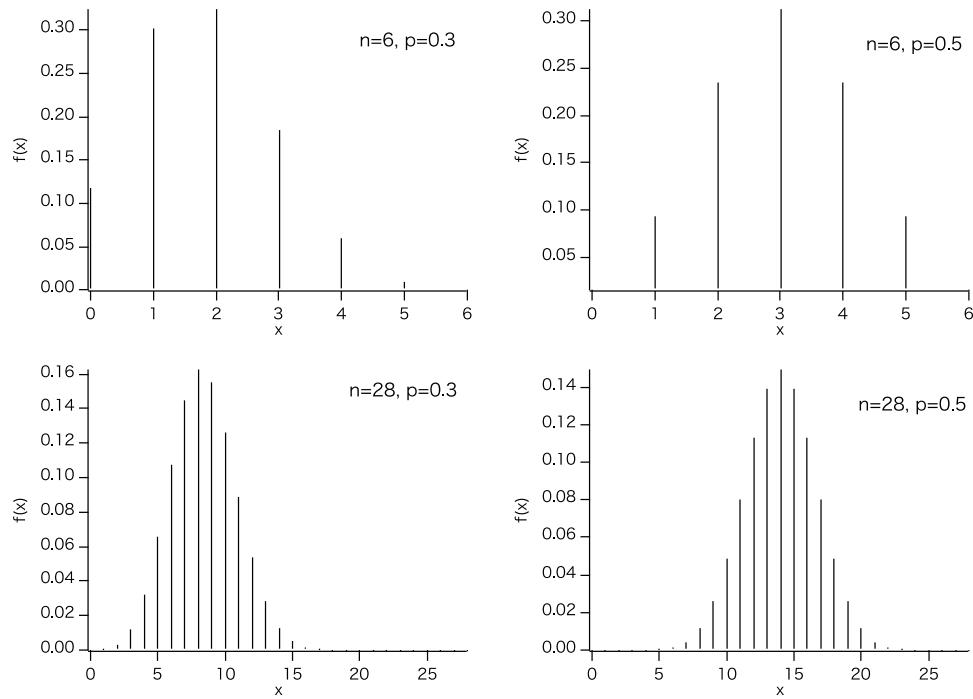


図 1.1. 二項分布の例

例えば、サイコロを 5 回投げたときに、そのうち 3 回で 6 の目が出る確率は、式(1.5)より $f(3) = {}_5 C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 = 0.032$ となる。

ポアソン分布

ごく多数の試行のうちごくまれにしか実現しないが、各試行における実現確率は一定値 p であるような事

象を考える。 N 回の試行における実現回数の平均値を λ とおくと、試行回数 N と実現確率 p の関係は $Np = \lambda$ とかける。そのため、 λ が有限の定数であるとき、 $N \rightarrow \infty$ の極限で $p \rightarrow 0$ となる。この極限において、二項分布の確率密度関数は下記のように近似される。

$$\lim_{N \rightarrow \infty} \lim_{p \rightarrow 0} f(X) = \lim_{N \rightarrow \infty} \lim_{p \rightarrow 0} \frac{N!}{X!(N-X)!} p^X (1-p)^{N-X}$$

$$\approx \lim_{p \rightarrow 0} \frac{N^X}{X!} p^X (1-p)^{N-X} = \lim_{p \rightarrow 0} \frac{\lambda^X}{X!} \frac{(1-p)^{\frac{\lambda}{p}}}{(1-p)^X} \approx \frac{\lambda^X}{X!} e^{-\lambda}$$

この極限における分布を以下のように定義する。

$$f(X) = \frac{\lambda^X}{X!} e^{-\lambda} \quad \dots (1.7)$$

式(1.7)について $X = 0$ から ∞ まで和をとると、

$$\sum_{X=0}^{\infty} f(X) = e^{-\lambda} \sum_{X=0}^{\infty} \frac{\lambda^X}{X!} = e^{-\lambda} e^{\lambda} = 1$$

となり、規格化条件を満たしている。式(1.7)の $f(X)$ をパラメータ λ のポアソン分布関数という。

ポアソン分布に従う確率変数の期待値と分散は、それぞれ $\langle X \rangle = \lambda$, $\langle (X - \langle X \rangle)^2 \rangle = \lambda$ である。

確率変数 X を、空間的に無秩序な事象の出現個数、あるいは時間的に無秩序な事象の生起回数とし、パラメータ λ を、同じ空間内での平均出現個数、あるいは同じ時間内での平均生起回数としたとき、 X の確率分布は式(1.7)のポアソン分布となる。ポアソン分布は、時空間的にランダムな事象の実現回数を記述する確率分布であり、離散分布の中では最も応用範囲が広い。

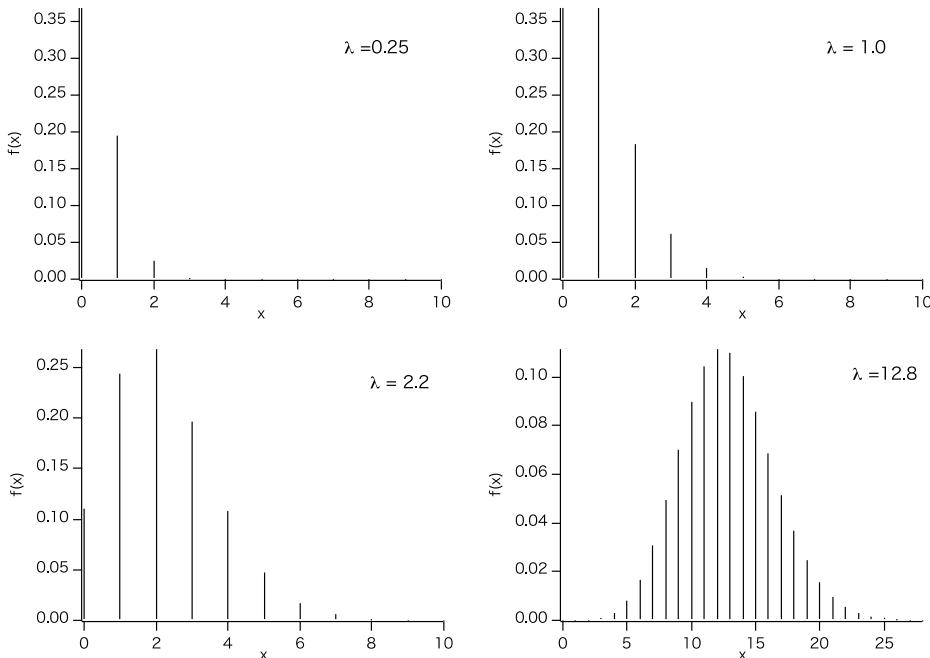


図 1.2. ポアソン分布の例

例題 1-1.

多数のゾウリムシが容積 $V = 1 \text{ m}^3$ のプールに泳いでいる。体積 $v=1 \text{ mm}^3$ の水をプールから採取し、その試料中のゾウリムシの個体数を顕微鏡観察で数える試行を多数回繰り返したところ、試料にゾウリムシが一匹もいない確率が 0.2 であることが分かった。プールの中にいるゾウリムシの総数はいくらか？

回答例：

一試料中のゾウリムシの個体数を X とすると、 X は確率変数でありパラメータ λ のポアソン分布に従う。プールの中のゾウリムシの総数を N とすると、 X の期待値であるパラメータ λ は、

$$\lambda = \text{ゾウリムシの平均数密度} \times \text{試料体積} = \frac{N}{V} \times v \text{ [個]}$$

である。 $X = 0$ の場合の確率が 0.2 であるから、

$$f(0) = \frac{\lambda^0}{0!} e^{-\lambda} = 0.2$$

すなわち

$$\lambda = \ln \frac{1}{0.2} = \ln 5$$

$$N = \frac{\lambda V}{v} = \frac{\ln 5 \times 1 [\text{m}^3]}{1 \times 10^{-9} [\text{m}^3]} = 1.61 \times 10^9 \text{ [個]}$$

■

例題 1-2.

野外調査で採集したある岩石塊から、断面積 1 cm^2 の薄片資料を多数切り出した。薄片資料は不透明なため、観察できるのは切断面のみで体積内部を観察することはできない。光学顕微鏡を用いて、各々の薄片資料の断面積 1 cm^2 の切断面に現れている放散虫化石の断面図形の個数を数えた。その結果、多数の薄片資料のうち、放散虫化石の断面図形が少なくとも 2 個はある資料の数割合は 0.6 であった。放散虫を直径 $100 \mu\text{m}$ の球形と仮定し、上記の観察結果をもとにこの岩石中の放散虫化石の数密度 n [個 m^{-3}]を求めなさい。

回答例：

ある 1 つの薄片資料の切断面で観察される放散虫化石の個数を X とする。 X の期待値を λ とすると、 X はパラメータ λ のポアソン分布に従う。 $X \geq 2$ である確率が 0.6 であることから、規格化条件は

$$1 - (f(0) + f(1)) = 1 - \left(\frac{\lambda^0}{0!} e^{-\lambda} + \frac{\lambda^1}{1!} e^{-\lambda} \right) = 0.6,$$

と書ける。つまり

$$0.4e^\lambda - \lambda - 1 = 0.$$

この方程式をニュートン法で数値的に解くと、 $\lambda = 2.022$ となる。特定の体積内に存在する放散虫の個数を

考えるとき、放散虫の球の中心点でその個数を数えるものとする。薄片資料の断面積を S 、放散虫の半径を r とすると、薄片資料の切削面に現れる放散虫の中心点は、切削面を中心とする厚さ $2r \times$ 断面積 S の体積内に存在しているので、数密度 n は

$$n = \frac{\lambda}{2rS} = \frac{2.022}{2 \cdot 50 \times 10^{-6} \cdot 1 \times 10^{-4}} = 2.0 \times 10^8 \text{ [個 } m^{-3}]$$

■

課題 1-1. (★)

例題 1-2.で、切削面で観察される放散虫化石が 2 個以下の資料の数割合が 0.2 の場合はどうなるか。

1.2. 連続分布

確率変数 X の実現値が連続した実数であるときは、確率密度関数は連続分布となる。 X から $X + dX$ までの値が実現される確率を $f(X)dX$ と表す。連続分布の規格化条件は、

$$\int_{-\infty}^{\infty} f(X)dX = 1, \quad \cdots (1.8)$$

であり、確率変数が X 以下の値を実現する確率である累積分布関数は、

$$F(X) = \int_{-\infty}^X f(X')dX', \quad \cdots (1.9)$$

となる。連続分布に従う確率変数 X の期待値 $\langle X \rangle$ と分散 $\langle (X - \langle X \rangle)^2 \rangle$ は、

$$\langle X \rangle = \int_{-\infty}^{\infty} Xf(X)dX, \quad \cdots (1.10)$$

$$\langle (X - \langle X \rangle)^2 \rangle = \int_{-\infty}^{\infty} (X - \langle X \rangle)^2 f(X)dX, \quad \cdots (1.11)$$

と定義される。期待値 $\langle X \rangle$ のことを母平均（記号 μ ）、分散 $\langle (X - \langle X \rangle)^2 \rangle$ のことを母分散（記号 σ^2 ）と呼ぶこともある。分散の平方根のことを標準偏差という。連続分布の例は下記の通り。

一様分布

確率変数 X が区間 $[a, b]$ の中のあらゆる値を等確率でとるとき、

$$f(X) = \begin{cases} \frac{1}{b-a} & (a \leq X \leq b) \\ 0 & (X < a, b < X) \end{cases} \quad \cdots (1.12)$$

となり、これを区間 $[a, b]$ の一様分布という。一様分布の期待値と分散は、 $\langle X \rangle = (a + b)/2$ 、 $\langle (X - \langle X \rangle)^2 \rangle = (b - a)^2/12$ である。

正規分布

以下の規格化条件を満たす確率密度関数

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(X - \mu)^2}{2\sigma^2}\right\}, \quad \cdots (1.13)$$

を母平均 μ 、母分散 σ^2 （標準偏差 σ ）の正規分布という。期待値と分散は、 $\langle X \rangle = \mu$ 、 $\langle (X - \langle X \rangle)^2 \rangle = \sigma^2$ である。後に説明する中心極限定理により、正規分布は最も重要で応用範囲の広い連続分布である。 $No(\mu, \sigma^2)$ とも書く。正規分布 $No(\mu, \sigma^2)$ は、変数変換

$$Z = \frac{X - \mu}{\sigma}, \quad \cdots (1.14)$$

により、確率変数 Z についての確率密度関数である標準正規分布 $No(0, 1)$

$$f(Z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z^2}{2}\right) \quad \cdots (1.15)$$

と互いに変換できる。式(1.14)のように、期待値を差し引いてから標準偏差で割るという変数変換を、確率変数の標準化という。

証明は省くが、二項分布(1.6)は N が大きくなるにつれ $No(Np, Np(1 - p))$ に近づき、ポアソン分布(1.7)は、 λ が大きくなるにつれ正規分布 $No(\lambda, \lambda)$ に近づく。定性的には、図 1.1, 1.2 からも推察できる。

指数分布

発生確率が一定の突発的事象が次におこるまでの待ち時間の長さ t を確率変数したとき、その確率密度関数を求めよう。単位時間あたりの事象の平均生起回数を λ とすると、任意の微小時間範囲 $t \sim t + dt$ の間に事象が起こる確率は λdt である（ただし $\lambda dt \ll 1$ と仮定した）。経過時間 t の間に事象が一度も起きない確率は、パラメータ λt のポアソン分布において、事象の生起回数 X が 0 のときであるから、

$$\frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \quad \cdots (1.16)$$

である。経過時間 t までに事象が起きず、かつ引き続く dt の微小時間内に事象がおこる確率は、式(1.16)と λdt の積であるから、 $\lambda e^{-\lambda t} dt$ に等しい。経過時間 t を X と書き直すと、

$$f(X) = \lambda e^{-\lambda X}, \quad (X > 0) \quad \cdots (1.17)$$

は確率変数 X についての確率密度関数となる。これをパラメータ λ の指数分布という。期待値と分散は、 $\langle X \rangle = \lambda^{-1}$ 、 $\langle (X - \langle X \rangle)^2 \rangle = \lambda^{-2}$ である。確率変数 X は任意の開始時刻から次に事象がおこるまでの待ち時間である。まとめると、事象がパラメータ λ のポアソン分布に従うとき、その事象が起こるまでの待ち時間 X の分布は

式(1.17)で表される。

しばしば、指数分布は「続いて起こる事象の時間間隔の分布」とも解釈されるが、厳密には上に述べた通り「任意の時刻から待ち始めた場合に事象がおこるまでの待ち時間の分布」であることに注意。待ち始める時刻は、ちょうど事象が起こったときである必要はなく、任意でよい。

課題 1-2. (★★)

ある地域では、いつ起こるか全くわからないが平均で2年に1回の頻度で地震が起こることが観測記録から知られている。この地域において70%の確率で次の地震が起こるのは、今から何年以内であると推測されるか。この問題を解くのにプログラミングは必要ない。(ヒント: 累積分布関数を考える)

自由度 ν のt-分布

確率変数 X についての確率密度関数

$$f(X) = k \left(1 + \frac{X^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad (k \text{は } \nu \text{ に依存した定数}) \quad \cdots (1.18)$$

を自由度 ν のt-分布という。t-分布は、あるデータ \vec{x} の母平均や、二種のデータ \vec{x}, \vec{y} の母平均の差についての区間推定・仮説検定に用いられる。定数 k は規格化条件によって決められる。

t-分布の比例定数 k は、ガンマ関数 Γ をもちいて $k = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}$ と表されるが、この数式表現は忘れても

困らない。なぜなら k の値は規格化条件の数値積分で求められるからである。

1.3. 数値積分

科学技術では、原始関数が存在しない関数について定積分の計算が必要になることが多い。このために数値積分が使われる。数値積分では、 $f(X)$ の定積分を、底辺が ΔX 、高さが $f(X)$ の短冊の面積 $f(X)\Delta X$ の和で近似する。

$$\int_a^b f(X)dX \approx \sum_{i=1}^N f(X_i)\Delta X, \quad \Delta X = \frac{b-a}{N},$$

$$\text{ただし } X_1 = a, \quad X_i = X_{i-1} + \Delta X \quad (i = 2, 3, \dots, N) \quad \cdots (1.19)$$

ΔX が小さいほど、即ち分割区間数 N が大きいほど誤差は小さくなる（ただし計算量は N に比例する。）

課題 1-3. (★)

関数 $\sin(x)$ について区間 $[0, \pi]$ における定積分を数値積分で求めよ。区間の分割数 N （デフォルトでは1000）

をいろいろ変えて、数値積分の誤差がどのように変化するか調べなさい。

課題 1-4. (★)

数値積分法を使って、自由度 $\nu = 10$ の t -分布（式(1.18)）の定数 k の値を求めよ。ただし積分区間 $[-\infty, \infty]$ は

$[-25, 25]$ と近似してよい。厳密式 $k = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}$ の直接計算との一致度を確認してみること。

1.4. 統計量

2つの確率変数 X, Y が互いに独立であるとは、両者の実現値の出現頻度が互いに影響を及ぼさないことを指す。統計解析では、ある観測量を繰り返し測定して得た観測値 x_i ($i = 1, \dots, N$)を、同一分布に従う N 個の互いに独立な確率変数 X_i ($i = 1, \dots, N$)の実現値として扱うことが多い。これはデータの確率論的抽象化の方法の一つであるが、数理的な扱いやすさから実際に多くの解析手法の前提となっている。本演習でもこの前提に従う。同一分布の独立な確率変数 X_i ($i = 1, \dots, N$)から実数值 R への写像

$$R = R(X_1, \dots, X_N), \dots \quad (1.20)$$

のことを統計量(statistic)という。統計量 R は確率変数の関数であるため、それ自身も確率変数である。 N 個の観測値 x_i ($i = 1, \dots, N$)が得られたとき、それを式(1.20)の確率変数 X_i ($i = 1, \dots, N$)に代入すれば、統計量の実現値

$$r = R(x_1, \dots, x_N), \dots \quad (1.20')$$

を得る。統計量はデータ解析において重要な役割を担う。まず、よく使われる統計量として、

標本平均 \bar{X}

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i, \dots \quad (1.21)$$

標本分散 V^2

$$V^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2, \dots \quad (1.22)$$

不偏分散 S^2

$$S^2 = \frac{N}{N-1} V^2, \dots \quad (1.23)$$

を挙げておく。統計量の実現値の呼称は、統計量の名称の末尾に”値”を付け加えたものとする。私たちがいつも“平均”と呼んでいるものは、標本平均 \bar{X} の実現値、つまり標本平均値 \bar{x} のことである。標本平均 \bar{X} の期待値 $\langle \bar{X} \rangle$ 、不偏分散 S^2 の期待値 $\langle S^2 \rangle$ はそれぞれ母平均 $\langle X \rangle$ 、母分散 $\langle (X - \langle X \rangle)^2 \rangle$ に一致する（証明は付録 A）。このことから、統計量 \bar{X} 、 S^2 はそれぞれ母平均、母分散の不偏推定量であるという。

1 次元配列 x に格納された数値データの標本平均値、標本分散値、不偏分散値を算出するには以下のようにすればよい。

```
import numpy as np # numpy ライブラリの使用宣言
(N 個のデータを 1 次元配列 x に格納)

np.mean(x) # x の標本平均値を算出して表示
np.var(x) # x の標本分散値を算出して表示
np.var(x, ddof=1) # x の不偏分散値を算出して表示
```

1.5. 亂数

確率変数とその実現値を模擬するために乱数発生器が用いられる。Python を含めほとんどのプログラミング言語には、指定した確率分布に従う乱数（確率分布乱数）の発生器が備わっている。

乱数発生器を使うコードの冒頭には以下の宣言を記述しておく。

```
import numpy as np # numpy ライブラリの読み込み
rng = np.random.default_rng() # 亂数発生器の定義
```

区間[0,1]の一様乱数を N 個生成して表示するには、例えば以下のようにすればよい。

```
for i in range(N):
    x = rng.random() # 亂数を 1 つ発生
    x # x を画面表示
```

区間[0,1]の一様乱数 N 個を要素とする 1 次元配列 x を生成するには、以下のようにすればよい。

```
x = rng.random(N) # 区間 [0,1] の一様乱数
```

その他の確率分布乱数については付録 G を参照のこと。

課題 1-5. (★)

区間 $[-\pi, \pi]$ の一様乱数を 10000 個発生させ、その頻度分布図（ヒストグラム）を描いて、おおよそ一様であることを確認しなさい。

```
# ヒント： ヒストグラム描画の例
import matplotlib.pyplot as plt # matplotlib ライブラリの使用宣言
(N 個のデータを 1 次元配列 x に格納)
plt.hist(x) # x の要素のヒストグラムの描画
```

課題 1-6. (★)

正規分布 $N(-3, 2^2)$ に従う確率変数の実現値を 10000 個生成し、密度分布図（面積 1 に規格化したヒスト

グラム）と母集団分布の曲線を重ね書きしなさい。乱数の出現確率密度が母集団分布にはほぼ一致していることを確かめなさい。

```
#ヒント： 一様分布乱数の密度ヒストグラムと1次関数y=xのプロットの重ね書きの方法の例
xx= rng.random(10000)
plt.hist(xx,density= True) # densityオプションを指定したヒストグラム
x= np.linspace(0,1,100) # 区間[0,1]の等間隔座標点100個からなる1次元配列の生成
y=x
plt.plot(x,y) # 2つの1次元配列の線形プロット
```

1.6. 中心極限定理

中心極限定理

母平均 μ , 母分散 σ^2 の同一分布に従う N 個の独立な確率変数 X_i ($i = 1, \dots, N$)について, $N \rightarrow \infty$ の極限で, 標本平均 \bar{X} の分布は母平均 μ , 母分散 σ^2/N の正規分布となる（確率変数の和 $\sum_{i=1}^N X_i$ の分布は母平均 $N\mu$, 母分散 $N\sigma^2$ の正規分布となる）。

この中心極限定理によれば, 任意の母集団分布（母平均 μ , 母分散 σ^2 ）もつ観測量について, 多数の標本を含むような時間あるいは空間領域で平均された観測量の確率分布は, 含まれる標本数 N が大きくなるに従い, 母平均 μ , 母分散 σ^2/N の正規分布に近づく。すなわち, 観測量の揺らぎの原因となる素過程スケールよりも広い時空間にわたり平均して観測を行うときはいつも, その観測値の密度ヒストグラムは正規分布に似てくるのである。この定理こそ, しばしばデータ解析において観測量が正規分布に従うと仮定されることの根拠である。（中心極限定理の証明は, 柴田[1995]や Osgood [2019]を参照されたい。）

課題 1-7. (★★)

1秒ごとにランダムに値が変化する物理量 X の時間平均観測を行うことを考える。 X は確率変数であり一様分布 $u[-1,1]$ に従うとする。物理量 X のデータの 10 秒平均観測, 100 秒平均観測, さらに 1000 秒平均観測の各々の条件について, 10000 個の観測値を取得する操作をコンピュータで模擬しなさい。これら 3 条件の観測値の密度ヒストグラムを重ねて描画し, また各条件での不偏分散値を算出しなさい。不偏分散値が平均時間に依存してどのように変わるのが調べ, それが中心極限定理と整合的かどうか理由とともに述べなさい。

上述の中心極限定理では, N 個の独立な確率変数 X_i ($i = 1, \dots, N$)が同一分布に従うことが必要だった。では, 同一分布とは限らないとき, 定理はどのように修正されるだろうか?

一般化中心極限定理

N 個の独立な確率変数 X_i ($i = 1, \dots, N$) それぞれが、母平均 μ_i 、母分散 σ_i^2 の分布に従うとする。 σ_i^2 ($i = 1, \dots, N$) はすべて有限かつ $\lim_{N \rightarrow \infty} \sum_{i=1}^N \sigma_i^2 \rightarrow \infty$ であるとき、 $N \rightarrow \infty$ の極限で、確率変数の標本平均 \bar{X} の分布は母平均 $\frac{1}{N} \sum_{i=1}^N \mu_i$ 、母分散 $\frac{1}{N^2} \sum_{i=1}^N \sigma_i^2$ の正規分布となる（確率変数の和 $\sum_{i=1}^N X_i$ の分布は母平均 $\sum_{i=1}^N \mu_i$ 、母分散 $\sum_{i=1}^N \sigma_i^2$ の正規分布となる）。

この中心極限定理では、各々の確率変数の母分散が総和の母分散に比べて無視できるほど小さいことが前提となっている。言い換えると、分散が突出して大きな少数の確率変数が含まれていなければ、一般化中心極限定理を適用できる。（定理の証明については清水[1976]や蓑谷[2010]を参照されたい。）

系列データのヒストグラム

ある物理量について時間あるいは空間的に連続して取得した観測値のことを系列データとよぼう。時間を空間座標に置き換えれば、空間系列データについても同じであるのでここでは時系列データのみに言及する。不規則ではあるが母平均と母分散が時間変化しない時系列のことを定常確率過程という。定常確率過程に従う時系列データの密度ヒストグラムは、中心極限定理により正規分布に似てくることがある。その状況としては、以下の2種類が考えられる。

1つ目の状況は、各観測値が過去の観測値から全く影響されておらず、かつ、各観測値多くの独立な確率変数の標本平均値となっているときである。例えば、空气中において個々の気体分子の衝突によって生み出される巨視的圧力の1秒平均値の測定時系列データは、極めて分散の小さな正規分布になる（はずである）。

2つ目は、各観測値が近接する観測値から影響されており、時系列データの長さが自己相関距離（影響が及ぶ時間長さ）よりも十分大きいときである。ここでは、過去の観測値から影響される不規則な時系列データの数理モデルとして最も簡単な1次の自己回帰モデル(AR(1))

$$X_i = aX_{i-1} + W_i, \quad \cdots \quad (1.24)$$

を用いる。 X_i は実現値として時刻 i のデータ点を生成する確率変数、 W_i は期待値ゼロ、分散 σ^2 の確率変数（ホワイトノイズ）である。 a は自己回帰係数 ($0 < a < 1$) である。AR(1) モデルの時間 k だけ離れた2つの確率変数 X_i と X_{i+k} の相関の強さ（お互いの影響の強さ）は、自己相関係数

$$R(k) = a^k, \quad \cdots \quad (1.25)$$

で表せる[北川, 2005]。 $R(k) \ll 1$ であれば、 X_i と X_{i+k} は独立な確率変数とみなせる。そこで、ある適当な閾値 ε ($\ll 1$) について、 $R(\forall k > K) < \varepsilon$ となる最小の K を自己相関時間と呼ぼう。図 1.3 は、 W_i が一様分布 $u[-1, 1]$ の AR(1) が生成した時系列データと、期間全体 ($i = 0 \sim 10000$) のデータのヒストグラムを示している。ヒストグラム形状は、自己相関時間が短いとき（ $a = 0.2$ の場合）はホワイトノイズ項の分布形状に近

いが、自己相関時間が長いとき（ $\alpha = 0.95$ の場合）は正規分布に近くなっていることが分かる。図1.3の結果から、「自己相関距離がある程度以上に長い定常確率過程では、ホワイトノイズの分布によらず、長時間取得したデータの分布は正規分布に近づく」ことが推察される。実際の現象について調べてみると面白いかもしれない（船舶の揺れ、波高や風速の揺らぎなど）。

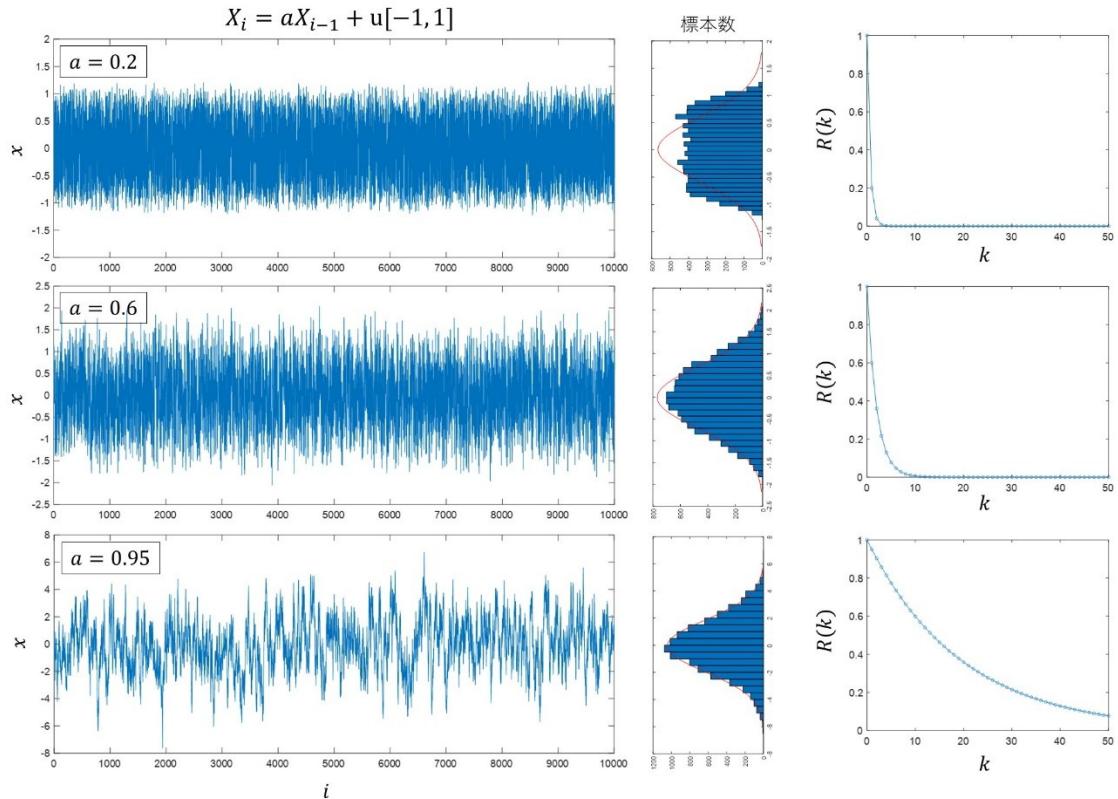


図 1.3. 一次の自己回帰モデルに従い経時変動する観測量 X の実現値 x_i の時系列 ($i = 1 \sim 10000$) とヒストグラム。自己回帰係数 α は上段・中断・下段の各々について時系列図中に示した通りで、ホワイトノイズは一様分布 $u[-1,1]$ とした。右側にはこの時系列モデルの自己相関係数 $R(k)$ をプロットした。ヒストグラムを最もよく近似する正規分布曲線(赤色)も図中に示した。

課題 1-8. (★★★)

図 1.3 の結果から、「自己相関距離がある程度長い定常確率過程では、ホワイトノイズの分布によらず、長時間取得したデータの頻度分布は正規分布に近くなる」ことが推察される。AR(1)モデルの場合について、これはなぜなのか、一般化中心極限定理をつかって説明しなさい。この問題を解くのにプログラミングは必要ない。（ヒント：漸化式を展開してみる）

課題 1-9. (★★★★)

取得した時系列データの密度ヒストグラムが観測量の母集団分布を十分よく再現するためには、時系列データの取得期間が自己相関時間よりも十分長いことが必要である。これはなぜか、AR(1)モデルの場合について説明しなさい。この問題を解くのにプログラミングは必要ない。(参考： $\alpha = 0.95$ の場合に、連続する 500 点のデータのヒストグラム形状は区間に依存して変わってしまうが、連続する 10000 点のデータのヒストグラム形状はいつも同じ正規分布に似る。)

課題 1-10. (★★★)

AR(1) モデル $X_i = \alpha X_{i-1} + W_i$ ($W_i \sim u[-1,1]$) における自己回帰係数 α が 0.2, 0.6, 0.9 のそれぞれの場合について、連続する 10000 点の時系列データの密度ヒストグラムと、一般化中心極限定理による母集団分布関数を重ね描きしなさい。そしてヒストグラムと分布の一一致度の α 依存性の理由を説明しなさい。(参考：ヒストグラムと関数曲線の一一致度を定量的に評価するには、例えば付録 D のコルモゴロフ・スミルノフ検定を使えばよいが、今回は視覚的判断だけでもよい。)

2. 区間推定と仮説検定

2.1. 方法論

单变量のデータ解析では、例えば気温など、いま注目しているただ一つの観測量を確率変数 $X \sim f(X)$ とする。記号 “~” は「分布に従う」を意味する。 N 個の観測値 $\vec{x} = \{x_i; i = 1, \dots, N\}$ は、ある N 個の同一分布かつ独立な確率変数 $\vec{X} = \{X_i; i = 1, \dots, N\}$ のそれぞれの実現値であるとみなされる。確率変数の実現値のことを母集団分布 f からの標本ともいう。データ \vec{x} から、母集団分布 $f(X)$ の X 軸上での位置(location), 広がり(spread), 非対称性(skewness)などについて推論を行うことが統計的推定である。真の分布 f を経験的に知るには無限個の標本を取得する必要がありこれは不可能である。標本 \vec{x} とは別に、何らかの仮定を施し、その条件下で推論を進めることになる。仮定の施し方によって、パラメトリック手法とノンパラメトリック手法に大きく分類される(図 2.1.)。前者は、真の分布 f の近似に適すると考えられる分布族(正規分布など)を先見的に仮定しておき、標本 \vec{x} を使って、その分布族のパラメータを推定する手法である。後者は、真の分布 f を、データ \vec{x} のヒストグラムにもとづく分布の経験的推定 \hat{f} で代用することで推論を進め。経験分布 \hat{f} は各観測値 $x_i (i = 1, \dots, N)$ が等しく $1/N$ の確率で生起する確率密度関数であり、面積を 1 に規格化したヒストグラムと同義である。

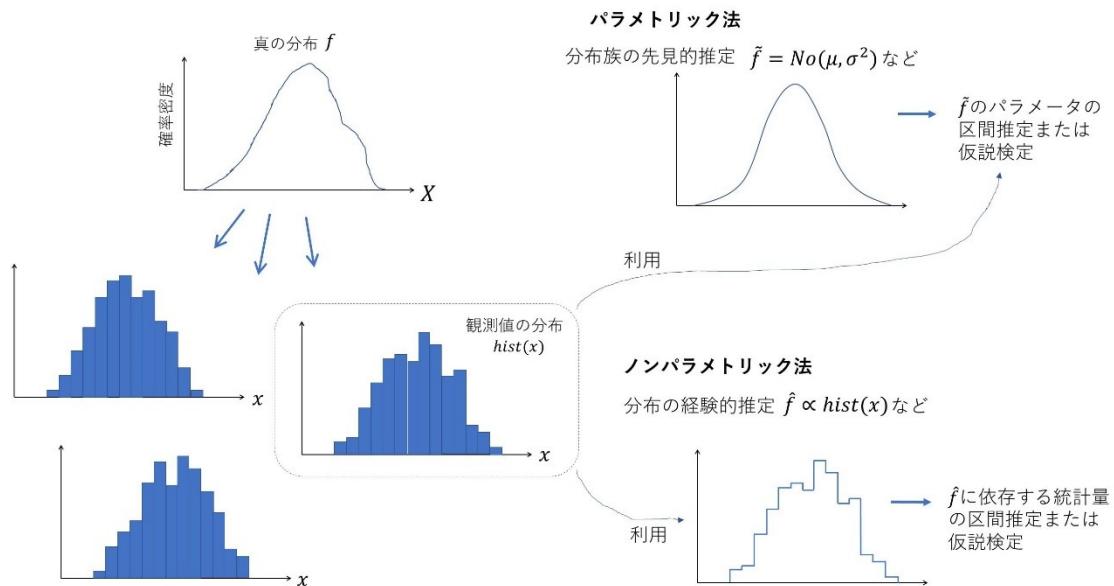


図 2.1. 観測量の真の分布(未知)、そこからの標本として取得された観測値の分布、観測値を用いた統計的推定の関係説明図。パラメトリック法とノンパラメトリック法における観測値の利用目的の違いに注意。

2.2. スチュードントの T

单变量データ解析で重要な統計量として、標本平均 \bar{X} と不偏分散 S^2 のほか、スチュードントの T がある。付録 A に示したように標本平均 \bar{X} の分散 $((\bar{X} - \langle \bar{X} \rangle)^2)$ は、 X の母分散 σ^2 とデータ次元 N との間につきの関係

をもつ。

$$\langle (\bar{X} - \langle \bar{X} \rangle)^2 \rangle = \frac{\sigma^2}{N}, \quad \cdots (2.1)$$

そのため $\langle (\bar{X} - \langle \bar{X} \rangle)^2 \rangle$ の不偏推定量として統計量 S^2/N が用いられる。 S^2/N の平方根で定義される統計量

$$SE = \frac{S}{\sqrt{N}}, \quad \cdots (2.2)$$

を \bar{X} の標準誤差(standard error)という。 \bar{X} の期待値 $\langle \bar{X} \rangle$ からの差を \bar{X} の標準誤差で割った統計量

$$T \equiv \frac{\bar{X} - \langle \bar{X} \rangle}{SE}, \quad \cdots (2.3)$$

はスチューデントの T と呼ばれる。スチューデントの T は、その確率密度関数の位置と広がりが X の分布やデータ次元 N にあまり強く影響されない特性を持つ。

特に、 X の分布が正規分布族 $No(\mu, \sigma^2)$ のとき、式(2.3)において $\langle \bar{X} \rangle = \mu$ であり、統計量 $T \equiv \frac{\bar{X} - \mu}{SE}$ の確率密度関数は解析的に求まり、それは t 分布と呼ばれる[柴田 1995 など]。 t 分布はパラメータ μ, σ^2 に依存しないことから、スチューデントの T は不動統計量(pivotal statistic)と呼ばれる。次の 3 章で述べるように、 T の不動性のおかげで、母平均や母分散が任意(あるいは未知)の正規母集団について、予め計算された数表(統計学の教科書に載っている t 分布表)を用いて X の母平均 μ の区間推定や仮説検定を行うことができる。このように観測量が正規分布に従うことを仮定する古典的な統計解析は、コンピュータが利用できない時代(1970 年代以前)においては必要不可欠な手段であった。 X が正規分布でない任意分布に従うときは、式(2.3)の T の確率密度関数は t 分布にはならず、その分布の解析的表現は一般にはない。ただし、観測値の個数 N が十分大きければ、中心極限定理により統計量 \bar{X} の分布は正規分布に近づくので、式(2.3)の T の確率密度関数は t 分布に近づく。このため、 N が十分大きければ、 $\langle \bar{X} \rangle$ についての区間推定や仮説検定には正規母集団を仮定した古典的方法が目立った誤差なく適用できてしまう。ただし、後で分かるように、 N がどのくらい大きければ十分よい近似なのかは、 X の母集団分布の形状や、区間推定の信頼水準(仮説検定の危険率)に依存してしまう。また、この古典的な区間推定・仮説検定は、標本平均 \bar{X} という特殊な統計量についてしか適用できない。

そこで、1970 年代以降のコンピュータ時代の統計学では、データの母集団の正規分布性や統計量の関数形の仮定を必要としない、一般化した区間推定・仮説検定の方法論の開発が行われた[Efron and Hastie 2016]。式(2.3)の統計量 T は、標本平均 \bar{X} というある特殊な統計量に注目した不動統計量である。この定義を一般化し、標本平均 \bar{X} を、式(1.20)で定義される任意の統計量 R に置き換えたもの

$$T \equiv \frac{R - \langle R \rangle}{SE}, \quad \cdots (2.4)$$

を、一般化したスチューデントの T と呼ぶことにしよう。これは近似的な不動統計量(approximate pivot)である。この一般化された T においては、標準誤差 SE として、 $\langle (\bar{X} - \langle \bar{X} \rangle)^2 \rangle$ の不偏推定量の平方根である式

(2.2)の代わりに、 $\langle(R - \langle R \rangle)^2 \rangle$ の不偏推定量の平方根が採用される。3章で扱うコンピュータ時代の統計学の区間推定法では、式(2.4)の統計量 T が重要な役割を担う。

2.3. 正規母集団の母平均の推定

本 2.3 節では、正規母集団を仮定した古典的な区間推定や仮説検定を扱う。どの統計学の教科書でも扱われている定石の内容であるが、3章の数値的手法との共通点と相違点を把握しやすくするために本資料に含めることにした。ここでは、実用頻度が最も高い、分散 σ^2 が未知であるときの期待値 μ の統計的推定のみを扱う。これは図 2.1.におけるパラメトリック法に属する。区間推定・仮説検定は、互いに独立な手法であるかのように扱われることが多いが、実際には、全く同じ問題に対して見方を変えただけものに過ぎない（例えば竹村 2007）。そのことが実感しやすいように、まず両者の共通事項を説明し、次にそれぞれに固有の事項を説明する。

2.3.1. 区間推定と仮説検定の共通事項

定理 2.1. 統計量 T の母集団分布

$X_i \sim N(\mu, \sigma^2)$ ($i = 1, \dots, N$) であるとき、式(2.3)で定義される統計量 T (スチューデントの T) は自由度 $N - 1$ の t 分布に従う。

この定理 2.1.は標準的な統計学の教科書には必ず載っているものの、その証明は省略されてしまっていることが多い。ここでは定理 2.1.の完全な証明を示す。まず、式(2.3)を以下のように書き直してみる。

$$T = \frac{\left(\frac{\bar{X} - \mu}{\sigma} \right)}{\sqrt{\frac{S^2}{N}}}, \quad \cdots (2.5)$$

式(2.5)の分子は、確率変数 $\bar{X}(N) \sim N(\mu, \frac{\sigma^2}{N})$ を標準化した確率変数であるため、標準正規分布に従う。

$$\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1). \quad \cdots (2.6)$$

では、式(2.5)の分母の統計量はどんな分布に従うだろうか？付録 B に証明を示した「正規母集団の不偏分散の分布定理 B1」を用いると、関係

$$\sqrt{\frac{S^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2(N-1)}{N-1}}, \quad \cdots (2.7)$$

を得る。ここで $\chi^2(N-1)$ は自由度 $N-1$ のカイ二乗分布である。ところで、確率変数 $Z \sim N(0, 1)$ とそれ

と独立な確率変数 $Y \sim \chi^2(N)$ から構成される確率変数 $Z / \sqrt{\frac{Y}{N}}$ は自由度 N の t 分布に従う。つまり

$$\frac{Z}{\sqrt{\frac{Y}{N}}} \sim t(N), \quad \cdots (2.8)$$

式(2.8)の証明はやや複雑であるものの統計学の多くの教科書に記述されている（例えば柴田 1995, pp. 87-89）。さらに付録の定理 B2 により標本平均 \bar{X} と不偏分散 S^2 は互いに独立であるため、式(2.5)の分子と分母は互いに独立である。そのため、式(2.8)左辺で N を $N - 1$ に置き換えた確率変数は、式(2.5)右辺の確率変数と等価である。このことから、統計量 T は $t(N - 1)$ に従うことが分かる。■

$t(N)$ は $N \rightarrow \infty$ において標準正規分布に収束する。厳密な収束証明は複雑である（尾畠 2014, p.140）が、直感的には式(2.5)の分母が $N \rightarrow \infty$ で 1 に収束することと式(2.6)からも推察できる。正規母集団の観測量については、統計量 T の分布は観測量の確率密度関数のパラメータ μ と σ^2 に依存しないため、 T は完全な不動統計量となっている。

データ \vec{x} の各要素 x_i ($i = 1, \dots, N$) を各確率変数 $X_i \sim N(\mu, \sigma^2)$ ($i = 1, \dots, N$) の実現値とする。このとき、式(1.21)-(1.23)の変数 X_i に実現値 x_i ($i = 1, \dots, N$) を代入して、標本平均 \bar{X} の実現値 \bar{x} 、標本分散 S^2 の実現値 s^2 を求め、それらを式(2.3)の変数 \bar{X} 、 S^2 にそれぞれ代入することで、統計量 T の実現値 τ が求まる。

$$\tau = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}, \quad \cdots (2.9)$$

本資料では、確率変数はアルファベット大文字でその実現値は小文字で表すことにしており、統計量 T について本章中では例外的にギリシャ文字 τ を採用した（ t 分布との重複を避けるため）。これらの共通事項に基づいて、母分散が未知である正規母集団の母平均の区間推定・仮説検定の方法を以下に述べる。

2.3.2. 区間推定

信頼水準 $1 - \alpha$ の区間推定を行うためには、まず、図 2.2 のように、統計量 T の確率密度関数 $f(T)$ の左側外縁領域と右側の外縁領域の積分値が互いに等しく $\alpha/2$ となるような、左の境界点 τ_l 、右の境界点 τ_u を定める。 α は小さな値（0.1, 0.05, 0.01 など）に設定される。この左右の境界点はそれぞれ確率密度関数 $f(T)$ の $\alpha/2$ 分位点、 $1 - \alpha/2$ 分位点である。今は特に $f(T) = t(N - 1)$ の場合を考えている。 $f(T)$ は偶関数なので $\tau_l = -\tau_u$ である。分布 $t(N - 1)$ に従う統計量 T の実現値である τ 値は、データ \vec{x} を取得する試行において、確率 $1 - \alpha$ で区間 $[\tau_l, \tau_u]$ の内側に入ることになる。これを式で表すと式(2.9)より、

$$\begin{aligned} 1 - \alpha &= \text{Prob}\{\tau_l \leq \tau \leq \tau_u\} \\ &= \text{Prob}\left\{\tau_l \leq \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \leq \tau_u\right\} \\ &= \text{Prob}\left\{\bar{x} - \tau_u \sqrt{\frac{s^2}{N}} \leq \mu \leq \bar{x} - \tau_l \sqrt{\frac{s^2}{N}}\right\} \end{aligned} \quad \cdots (2.10)$$

式(2.10)に基づき、ある観測データ \vec{x} を取得したとき、観測量 $X_i \sim N(\mu, \sigma^2)$ の母平均 μ の区間推定値は、

$$\bar{x} - \tau_u \sqrt{\frac{s^2}{N}} \leq \mu \leq \bar{x} - \tau_l \sqrt{\frac{s^2}{N}}, \quad \cdots (2.11)$$

として算出される。母平均 μ が区間(2.11)に入る確率は $1 - \alpha$ で、入らない確率は α である。これを言い換えると、「正規母集団に従う観測量について観測データ \bar{x} を取得し、母平均を区間推定する」という試行が、正解を出す確率は $1 - \alpha$ である。

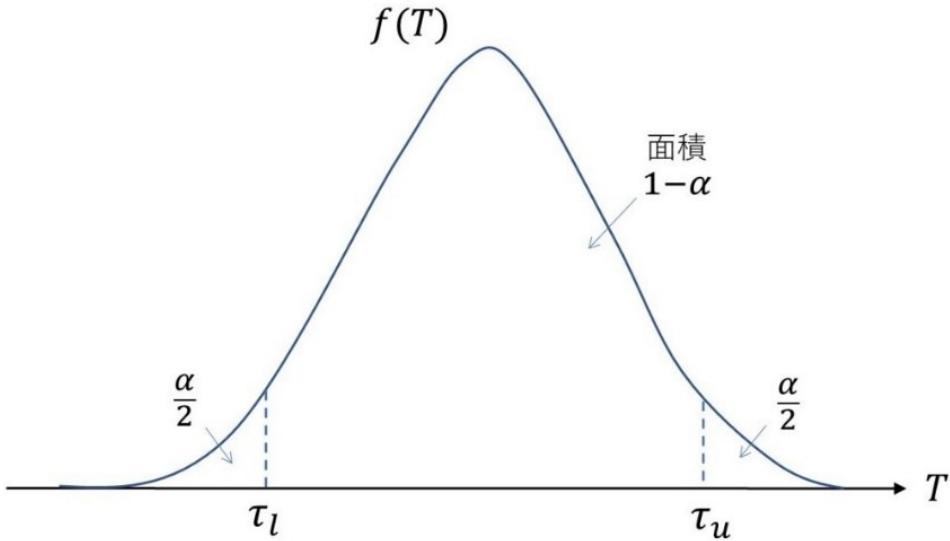


図 2.2. 統計量 T の確率密度関数 $f(T)$ とその信頼水準 $1 - \alpha$ の両側境界値 (τ_l, τ_u) の模式図

課題 2-1. (★★)

自由度 $v = 20$ の t 分布について、信頼水準 $1 - \alpha = 0.95$ の両側境界値 (τ_l, τ_u) を数値積分で計算するコードを書きなさい。自作コードの結果の正誤確認のために、付録 F に記載した scipy モジュールの関数 stat.t.ppf を使うとよい。

課題 2-2. (★★)

サンプル気象データ（付録 E）うち、どれか一つ（例：東京の気温）の 1951~2020 年のデータについて、ヒストグラムを描き、正規母集団を仮定した母平均の区間推定結果を水平線分で重ね描きしなさい。信頼水準は 0.95 とする。scipy モジュールの関数 stat.t.ppf を使ってよい。

2.3.3. 仮説検定

仮説検定の準備として、まず付録 C に示した統計的仮説検定の枠組みを読んでおいて頂きたい。その枠組みに沿って、母分散 σ^2 が未知である条件下での母平均 μ についての仮説検定を説明する。帰無仮説 H_0 とし

て「母平均 μ はある値 μ_0 に等しい」という命題を採用とする。すなわち、

$$P: \mu_0 = \mu, \quad \dots (2.12)$$

として付録Cにおける命題Pを定義する。このとき、仮説Pの検定統計量として式(2.3)の統計量T（スチューデントのT）を $T \equiv (\bar{X} - \mu_0)/\frac{s}{\sqrt{N}}$ として採用すると、帰無分布 $f_0(T)$ は自由度 $N - 1$ のt分布となる。付録Cにおける命題Qを

$$Q: T \sim t(N - 1), \quad \dots (2.13)$$

として定義すると、命題($P \Rightarrow Q$)は恒真であり、その対偶である $(\bar{P} \Leftarrow \bar{Q})$ も恒真である。ここでは危険率 α の両側検定を考える。帰無分布 $t(N - 1)$ における帰無仮説の棄却境界値は、信頼水準 $1 - \alpha$ の区間推定での境界値 (τ_l, τ_u) と等価である。そのため、命題Qが真のとき（式(2.10)と同様の）つぎの関係

$$\begin{aligned} 1 - \alpha &= \text{Prob}\{\tau_l \leq \tau \leq \tau_u\} \\ &= \text{Prob}\left\{\tau_l \leq \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \leq \tau_u\right\}, \quad \dots (2.14) \end{aligned}$$

が成り立ち、統計量Tの実現値 τ が区間 $[\tau_l, \tau_u]$ 内に出現する確率は $1 - \alpha$ であり、区間外（棄却域内）に出現する確率は α となる。そのため、データ x を取得したときもし τ が棄却域に出現したならば、命題 \bar{Q} 「統計量Tは分布 $t(N - 1)$ に従わない」が真であることがデータから示されたとする。このとき命題 \bar{Q} が真であるとする検定結果と恒真命題($P \Leftarrow \bar{Q}$)により、命題Pが真（帰無仮説が偽）であると結論付ける。命題Qが真であるときに τ が棄却域に出現する確率は α なので、この推論が誤る確率は α である。言い換えると、データ x を取得して母平均について仮説検定を行う試行において、正しい帰無仮説が棄却されてしまう確率は α である。これが α を危険率とよぶ所以である。危険率は有意水準とも呼ばれる。

課題 2-3. (★★★)

サンプル気象データ（付録E）うち、どれか一つ（例：東京の気温）の2000~2020年のデータについて、正規母集団を仮定した母平均の仮説検定を行いなさい。 μ_0 としては1951~1999年のデータの標本平均値を用いよ。危険率は0.05とし、両側検定と片側検定を実施し、数値積分でp値も算出しなさい。（回答の自己点検のために、scipyモジュールの関数stats.ttest_1sampを使うよい。）

2.3.4. 区間推定と仮説検定の関係

このように、母平均の両側仮説検定と区間推定は、不動統計量であるスチューデントのTとある水準値 α により定義された数理的問題を、表と裏から見ただけのものである。このことは、乱数標本データを用いてシミュレートした区間推定の正解率と両側仮説検定の棄却率の比較結果（図2.3）からも確認できる。

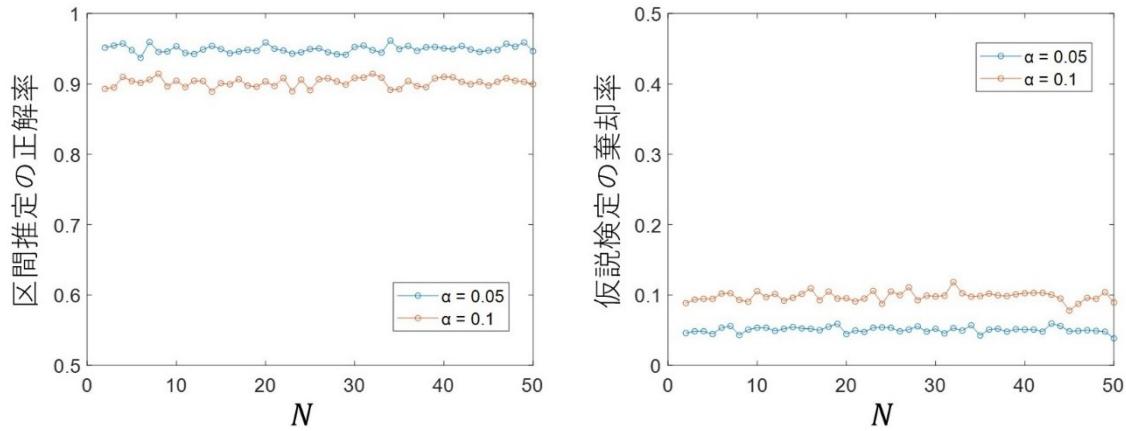


図 2.3 正規母集団 $N(0,1)$ から抽出した乱数データ \vec{x} を用いた期待値の区間推定(信頼水準 $1 - \alpha = 0.95$)の正解率と、期待値の両側仮説検定(帰無仮説 $\mu = 0$ 、危険率 $\alpha = 0.05$)の棄却率を、 \vec{x} の要素数 N に対してプロットしたもの。正解率と棄却率の算出のための試行回数は 2000。

2.3.5. 点推定の標準誤差と区間推定との関係

母平均の点推定量として標本平均 \bar{X} が用いられる。区間推定ではなく点推定のときは、標本平均の誤差を表す統計量として式(2.2)の標準誤差 SE を用い、結果を統計量 $\bar{X} \pm kSE$ の実現値

$$\bar{x} \pm k \cdot se, \quad \dots (2.15)$$

として表すことがある。 k は誤差幅のスケール因子であり、1, 2, 3 などの整数が採用されることが多い。

点推定の表式(2.15)を区間推定の表式(2.11)と比較すると明らかなように、点推定の誤差因子 k は、区間推定における分布 $t(N - 1)$ の分位点 τ_l, τ_u の大きさ（原点からの距離）と等価である。 N が十分大きければこれは標準正規分布の分位点 z_l, z_u の大きさに等しい。大標本条件 ($N \gg 1$) における「点推定の因子 k 」と「区間推定の信頼水準 $1 - \alpha$ 」の対応関係を表 1 に示した。

表 1. 母平均の点推定の誤差因子 k と、母平均の区間推定の信頼水準 $1 - \alpha$ の対応（大標本条件）

k	$1 - \alpha = \text{Prob}\{\bar{x} - z_u \cdot se \leq \mu \leq \bar{x} - z_l \cdot se\}$	α
1	0.683	0.317
1.645	0.900	0.1
1.96	0.950	0.05
2	0.954	0.046
2.575	0.990	0.01
3	0.997	0.003

大標本条件($N \gg 1$)においては、例えば、点推定 $\bar{x} \pm se$ は信頼水準 0.683 の区間推定と等価である。逆に、信頼水準 0.95 の区間推定は、点推定 $\bar{x} \pm 1.96se$ と等価である。

2.3.6. 2つの正規母集団の母平均の差の推定

母分散 σ^2 が互いに等しく母平均が一般に異なる 2 つの正規分布 $No(\mu_1, \sigma^2)$, $No(\mu_2, \sigma^2)$ を考える。これら二つの分布それぞれに従う観測量 X , Y について取得したデータ \vec{x} , \vec{y} から、母平均の差についての区間推定・仮説検定を行う方法を考える。

定理 2.2. 母分散の不偏推定量

互いに独立な確率変数 $X_i \sim No(\mu_1, \sigma^2)$ ($i = 1, \dots, N_1$), $Y_i \sim No(\mu_2, \sigma^2)$ ($i = 1, \dots, N_2$) について、母分散 σ^2 の不偏推定量は $S_{X,Y}^2 = \frac{1}{N_1 + N_2 - 2} [\sum_{i=1}^{N_1} (X_i - \bar{X})^2 + \sum_{i=1}^{N_2} (Y_i - \bar{Y})^2]$ である。

証明 :

不偏分散（式(1.23)）の期待値は母分散 σ^2 に等しいことから、

$$\langle \sum_{i=1}^{N_1} (X_i - \bar{X})^2 + \sum_{i=1}^{N_2} (Y_i - \bar{Y})^2 \rangle = \langle \sum_{i=1}^{N_1} (X_i - \bar{X})^2 \rangle + \langle \sum_{i=1}^{N_2} (Y_i - \bar{Y})^2 \rangle = (N_1 - 1)\sigma^2 + (N_2 - 1)\sigma^2 = (N_1 + N_2 - 2)\sigma^2$$

という結果が導ける。 ■

定理 2.3. 2 つの正規母集団の不偏分散の分布定理

定理 2.2. で定義した統計量 $S_{X,Y}^2$ について、

$$\frac{S_{X,Y}^2}{\sigma^2} \sim \frac{\chi^2(N_1 + N_2 - 2)}{N_1 + N_2 - 2}, \quad \cdots (2.16)$$

が成り立つ。

証明 :

$\vec{X} = \{X_i; i = 1, \dots, N_1\}$, $\vec{Y} = \{Y_i; i = 1, \dots, N_2\}$ をつなげて一つの確率変数ベクトル $\{X_1, \dots, X_{N_1}, Y_1, \dots, Y_{N_2}\}$ とみなし、付録 B の「正規母集団の不偏分散の分布定理」の証明にててくる Helmert 行列を、 X_1, \dots, X_{N_1} , Y_1, \dots, Y_{N_2} それぞれに独立に作用する N_1 次元, N_2 次元の Helmert 行列ブロックからなるブロック対角行列に拡張して定理 B1 の証明と同じ手続きを踏むことにより、統計量

$$(N_1 + N_2 - 2) \frac{S_{X,Y}^2}{\sigma^2} = \sum_{i=1}^{N_1} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^{N_2} \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2$$

が自由度 $N_1 + N_2 - 2$ のカイ二乗分布に従うことが示せる。 ■

本節で推定したい分布パラメータは $\mu_1 - \mu_2$ である。そこで、单变量の母平均の区間推定で用いた統計量 T (式(2.3))の母平均に関する項を、2つの正規母集団の母平均の差に関する項に置き換えてみよう。まず、 $\mu_1 - \mu_2$ の推定量としては統計量 $\bar{X} - \bar{Y}$ を用いるのが適當であろう。統計量 $\bar{X} - \bar{Y}$ の分散は、独立な確率変数の和の分散の公式(付録A1)と正規分布の再生性(付録A3)を用いて、 $\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\sigma^2$ であることが導ける。このことと定理2.2から、統計量 $\bar{X} - \bar{Y}$ の分散の不偏推定量は $\left(\frac{1}{N_1} + \frac{1}{N_2}\right)S_{X,Y}^2$ であることが分かる。

以上の議論から、式(2.3)のスチューデントの T の定義は以下の式で置き換えられる。

$$T \equiv \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)S_{X,Y}^2}}, \quad \dots (2.17)$$

これは式(2.4)で表される一般の統計量 R についてのスチューデントの T の特殊な場合とみなせる($R \equiv \bar{X} - \bar{Y}$)。式(2.17)を少し書き換えると

$$T = \frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\sigma^2}}}{\sqrt{\frac{S_{X,Y}^2}{\sigma^2}}}, \quad \dots (2.18)$$

となる。式(2.18)の分子 $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\sigma^2}}$ は標準正規分布 $No(0,1)$ に従う。式(2.18)の分母 $\sqrt{\frac{S_{X,Y}^2}{\sigma^2}}$ は式(2.16)から

$$\sqrt{\frac{S_{X,Y}^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2(N_1 + N_2 - 2)}{N_1 + N_2 - 2}} \quad \dots (2.19)$$

であることが分かる。定理B2と同じ推論により式(2.18)の分子と分母は互いに独立な確率変数である。これらの結果と定理2.1のときと同じ推論により、式(2.17)で定義されるスチューデントの T について以下の定理が導かれる。

定理2.4. 母平均の差についての T の分布

互いに独立な確率変数 $X_i \sim No(\mu_1, \sigma^2)$ ($i = 1, \dots, N_1$), $Y_i \sim No(\mu_2, \sigma^2)$ ($i = 1, \dots, N_2$)が与えられたとき、式(2.17)の統計量 T の確率密度分布は $T \sim t(N_1 + N_2 - 2)$ となる。

定理2.4.を用いると母平均の差の区間推定を直ちに導出できる。二つの正規母集団 $No(\mu_1, \sigma^2)$, $No(\mu_2, \sigma^2)$ のそれぞれから観測データ $\vec{x}=\{x_i; i = 1, \dots, N_1\}$, $\vec{y}=\{y_i; i = 1, \dots, N_2\}$ を得たとき、信頼水準 $1 - \alpha$ での母平均の差の区間推定は次のように求められる。

$$\begin{aligned}
 1 - \alpha &= \text{Prob}\{\tau_l \leq \tau \leq \tau_u\} = \text{Prob}\left\{\tau_l \leq \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) s_{X,Y}^2}} \leq \tau_u\right\} \\
 &= \text{Prob}\left\{\bar{x} - \bar{y} - \tau_u \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) s_{X,Y}^2} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} - \tau_l \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) s_{X,Y}^2}\right\}, \quad \cdots (2.20)
 \end{aligned}$$

ここで τ_l, τ_u はそれぞれ分布 $t(N_1 + N_2 - 2)$ の $\alpha/2$ 分位点, $1 - \alpha/2$ 分位点である。

つぎに、母平均の差の仮説検定において帰無仮説 P を $\mu_1 = \mu_2$ とした場合について考えよう。この場合、検定統計量は式(2.17)で $\mu_1 - \mu_2 = 0$ と置いた

$$T \equiv \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) S_{X,Y}^2}}, \quad \cdots (2.21)$$

であり、帰無仮説 P のもとにこの統計量 T が従う帰無分布は $t(N_1 + N_2 - 2)$ である。有意水準（危険率）を α としたとき、 T の実現値 τ は以下の確率則を満たす。

$$1 - \alpha = \text{Prob}\{\tau_l \leq \tau \leq \tau_u\} = \text{Prob}\left\{\tau_l \leq \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) s_{X,Y}^2}} \leq \tau_u\right\}, \quad \cdots (2.22)$$

実現値 τ が区間 $[\tau_l, \tau_u]$ の外側（棄却域）に見いだされたとき、帰無仮説は棄却される。片側検定の場合、関係式(2.22)は、下片側検定では $1 - \alpha = \text{Prob}\{\tau \geq \tau_l\}$ 、上片側検定では $1 - \alpha = \text{Prob}\{\tau \leq \tau_u\}$ と表される。

課題 2-4. (★)

サンプル気象データの観測量（気温、年降水量、最大 1h 降水量）のいずれか 1 つについて、東京と前橋の観測値のヒストグラムを重ね描きして視覚的に比較したうえで、「東京と前橋の観測値の母平均の差」を区間推定しなさい。信頼水準は 0.95 とする。

課題 2-5. (★★)

サンプル気象データの観測量（気温、年降水量、最大 1h 降水量）のいずれか 1 つについて、東京と前橋の観測値のヒストグラムを重ね描きして視覚的に比較したうえで、「東京と前橋の観測値の母平均には差がない」という帰無仮説を片側検定しなさい。有意水準（危険率）は 0.05 とする。

3. 母集団分布が未知のときの統計解析

この場合、未知の母集団分布を N 個の観測データ $\vec{x} = \{x_i; i = 1, \dots, N\}$ だけから推定し、それを統計量 R の区間推定・仮説検定に反映しなければならない。本章では、 $N \rightarrow \infty$ の時に厳密な結果を与えることが数学的に示されている実用的な手法として Bootstrap(BS)法を扱う。BS法は数理統計学者 B. Efron によって 1979 年に発表された[Efron 1979]。現在、BS 法は頻度主義統計学におけるノンパラメトリック法（図 2.1）の標準的方法論として広く使われている[Efron and Hastie 2020, 汪・桜井 2011]。

3.1. Bootstrap 原理

BS 法の原理を理解するための準備として、母集団からの標本抽出とそれを用いた統計量の推定手続きについておさらいしておく。観測量 X の真の母集団分布 f から観測データ $\vec{x} = \{x_i; i = 1, \dots, N\}$ を好きなだけ繰り返し無作為抽出できるとする。いま、抽出を B 回繰り返してデータセット $\vec{x}_{(1)}, \vec{x}_{(2)}, \dots, \vec{x}_{(B)}$ を得たとすると、統計量 R の B 個の標本値 $r_{(1)}, r_{(2)}, \dots, r_{(B)}$ が得られる。これらの標本値を用いて統計量 R の分布を推定することができる。統計量 R としては例えば式(2)の標本平均や式(4)の不偏分散などである。統計量 R の分布の特徴量として例えば期待値(R)の点推定値は $\bar{r} = \frac{1}{B} \sum_{j=1}^B r_{(j)}$ である。 $B \rightarrow \infty$ の極限においては、標本平均値 \bar{r} は期待値(R)に一致し、 r の標本密度分布は統計量 R の分布に一致するので、統計量 R について完全に理解できることになる（図 3.1）。

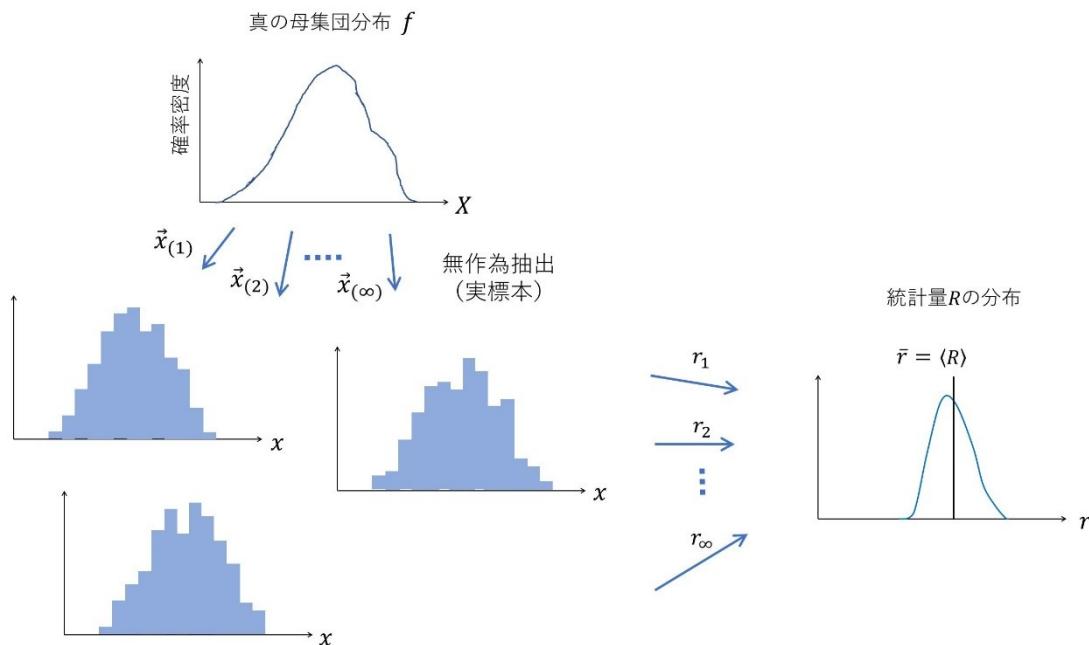


図 3.1. 実世界の理想条件における標本サンプリングと統計量 R の分布の推定

実際の観測では図 3.1.のような $B \rightarrow \infty$ という理想条件は実現できないため、真の期待値(R)や真の統計量 R

の分布を知ることは不可能である。実際の観測ではほとんどの場合、データ \vec{x} を取得できるのは 1 回だけ ($B = 1$) である。ところ 2 章では $B = 1$ だったにも関わらず、標本平均という統計量について期待値(R)とその区間を推定できたのはなぜか？観測量 X の真の母集団分布 f が正規分布族 $No(\mu, \sigma^2)$ であると仮定するパラメトリック法を採用していたからである（図 2.1）。

BS 法は、母集団分布の関数形を何も仮定せず（ノンパラメトリック）に、1 回限りしか取得できないデータ \vec{x} から統計量 R の期待値(R)の信頼区間を推定する方法である。BS 法では、真の母集団分布 f の代わりに、真の母集団分布の推定 \hat{f} から、疑似データ \vec{x}^* の無作為抽出を好きなだけ（ B 回）繰り返してデータセット $\vec{x}_{(1)}^*, \vec{x}_{(2)}^*, \dots, \vec{x}_{(B)}^*$ を得る（図 3.2.）。 \hat{f} から無作為抽出される疑似データとそれに依存する統計量の実現値のことを BS 標本と呼び、実標本と区別するため、上付き添え字*を付ける。

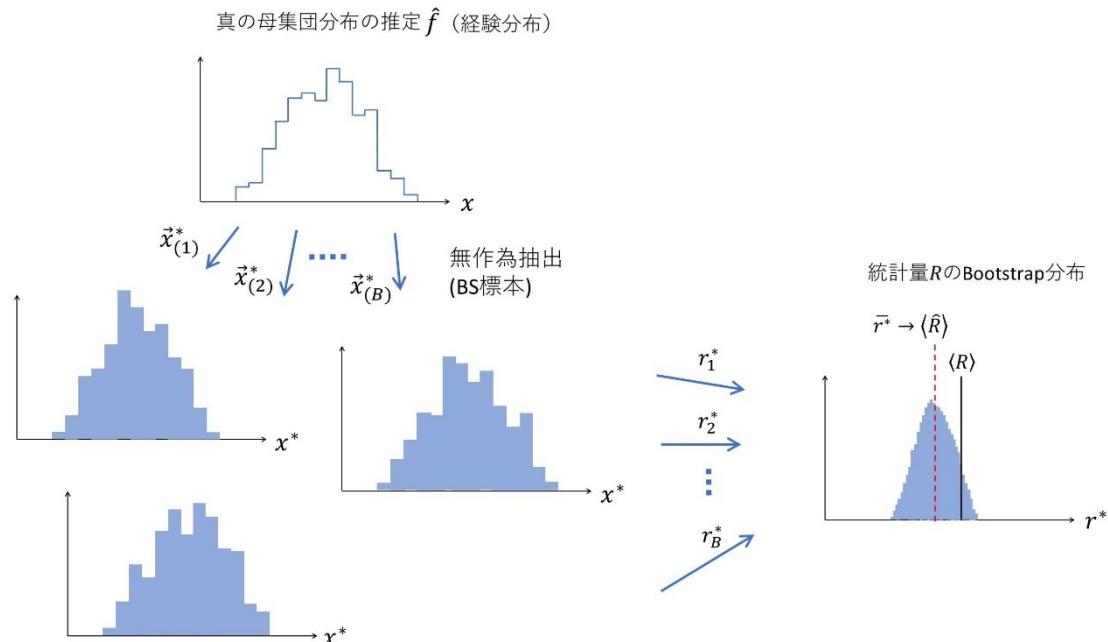


図 3.2. Bootstrap 世界における標本サンプリングと統計量 R の分布の推定

真の母集団分布の推定 \hat{f} は、データ \vec{x} にもとづいて決められるため、経験分布とも呼ばれる。経験分布 \hat{f} の決め方は、通常、次の 2 つのうちいずれかである。

- \vec{x} の要素 x_1, x_2, \dots, x_N がそれぞれ確率 $\frac{1}{N}$ で実現される離散分布
- \vec{x} の要素 x_1, x_2, \dots, x_N の密度推定である滑らかな連続分布（カーネル密度推定等）

本演習では、本家の BS 法である 1 つ目の方法をとる。この経験分布 \hat{f} からある 1 つの BS 標本 \vec{x}_* を抽出する操作は、実標本 \vec{x} の要素 x_1, x_2, \dots, x_N から N 個の要素をランダムに復元抽出する操作と同じである（復元抽出では同じ要素を複数回選べることに注意）。

BS 法では、経験分布 \hat{f} から無作為抽出した BS 標本 $\vec{x}_{(1)}^*, \vec{x}_{(2)}^*, \dots, \vec{x}_{(B)}^*$ から導出される、統計量 R の

BS 標本 $r_{(1)}^*, r_{(2)}^*, \dots, r_{(B)}^*$ に基づいて、経験分布についての統計量 \hat{R} の区間を推定する。一般に、BS 標本にもとづく推定値は真値の不偏推定量ではない。例えば BS 統計量 $\bar{r}^* = \frac{1}{B} \sum_{j=1}^B r_{(j)}^*$ は、 $B \rightarrow \infty$ において経験分布 \hat{f} についての期待値 $\langle \hat{R} \rangle$ と一致するのであって、真の分布 f についての期待値 $\langle R \rangle$ と一致するとは限らない（図 3.2.）。経験分布 \hat{f} における統計量 R の期待値 $\langle \hat{R} \rangle$ は、実標本 \vec{x} についての統計量 R の実現値 r に等しい。実標本 \vec{x} の要素数が大きな極限 $N \rightarrow \infty$ では、経験分布 \hat{f} は真の分布 f に収束するため、BS 期待値 $\langle \hat{R} \rangle$ は真の期待値 $\langle R \rangle$ に一致する。

BS 法による区間推定アルゴリズムは、正規母集団の母平均の区間推定のようにただ一つの決め手があるわけではなく、たくさんのバリエーションがある。 N が大きくなるにつれて区間推定の誤差が減少していくスピードは、真の分布の形状（歪度、尖度など）とアルゴリズムの両者に依存する [Efron and Tibshirani 1993]。

BS 法における復元抽出の組みあわせ総数は、インデックス $1, \dots, N$ でラベルされた元データ要素を N 個の小部屋に割り振る ($N - 1$ 個のしきりで分ける) 組み合わせの数に等しく、 ${}_{2N-1}C_{N-1}$ 通りである。例えば、 $N = 20, N = 100$ のとき、組み合わせ数はそれぞれ $\sim 6.9 \times 10^{10}, \sim 4.5 \times 10^{58}$ にもおよぶ。そのため、多くの場合、すべての BS 標本を網羅することはコンピュータを使っても現実的ではない。そこで「高品質なランダムさ」をもつサンプリングが必要不可欠である。

2 章で扱った正規母集団についての解析的な区間推定法とは異なり、本 3 章での BS 法による数値的な区間推定法は、正規母集団の仮定が必要ないというだけでなく、同じ区間推定アルゴリズムを標本平均以外の統計量 R （例：中央値、刈り込み標本平均など）にもそのまま適用できるという利便性がある。コンピュータが利用できる今、BC 区間推定法はもっと広く用いられるべきである。

3.2. BS 区間推定法

本節では、性能には劣るが最も簡単な BS percentile 法と、性能が優れておりコード実装が簡単な BST 法という、2 種類の区間推定アルゴリズムを紹介する。統計量 R の期待値 $\langle R \rangle$ の区間推定を行うものとし、信頼水準は $1 - \alpha$ 、BS 標本数は $B \geq \sim 10000$ とする。

BS percentile 法

統計量 R の BS 標本の分位値をそのまま算出するだけの簡単な方法である。

アルゴリズム

経験分布 \hat{f} の BS 標本 $\vec{x}_{(1)}^*, \vec{x}_{(2)}^*, \dots, \vec{x}_{(B)}^*$ から算出した統計量 R の BS 標本 $r_{(1)}^*, r_{(2)}^*, \dots, r_{(B)}^*$ を昇順に並

べ替えておき, $\frac{\alpha}{2}B$, $\left(1 - \frac{\alpha}{2}\right)B$ に最も近い整数をそれぞれ L , U とおくと, r^* の分布の $\frac{\alpha}{2}$ 分位値, $1 - \frac{\alpha}{2}$ 分位値がそれぞれ $r_{(L)}^*$, $r_{(U)}^*$ として得られる。これを用いて, 以下のように区間推定を定義する。

$$\begin{aligned} 1 - \alpha &= \text{Prob}\{r_{(L)}^* \leq r^* \leq r_{(U)}^*\} \\ &\approx \text{Prob}\{r_{(L)}^* \leq \langle \hat{R} \rangle \leq r_{(U)}^*\}, \quad \cdots (3.1) \\ &\approx \text{Prob}\{r_{(L)}^* \leq \langle R \rangle \leq r_{(U)}^*\} \end{aligned}$$

式(3.1)の 1 行目から 2 行目では, 統計量 \hat{R} の期待値 $\langle \hat{R} \rangle$ の区間を, 実現値 r^* の区間で代用している（注 a）。2 行目から 3 行目は BS 世界から実世界への移行であり, $N \rightarrow \infty$ で厳密となる。統計量 R の期待値 $\langle R \rangle$ の信頼水準 $1 - \alpha$ での推定区間は, $[r_{(L)}^*, r_{(U)}^*]$ として算出される。

（注 a）この代用の根拠を説明している文献を見つけられなかったので, 茂木の個人的な解釈を述べる。ベイズの定理により, 母数 $\langle \hat{R} \rangle$ を与えたときの実現値 r^* の分布 $\text{Prob}(r^* | \langle \hat{R} \rangle)$ と, r^* が得られたときの未知母数 $\langle \hat{R} \rangle$ の分布 $\text{Prob}(\langle \hat{R} \rangle | r^*)$ との間には, 以下の関係がある。

$$\text{Prob}(\langle \hat{R} \rangle | r^*) \propto \text{Prob}(r^* | \langle \hat{R} \rangle) \text{Prob}(\langle \hat{R} \rangle)$$

ここで $\text{Prob}(\langle \hat{R} \rangle)$ は母数 $\langle \hat{R} \rangle$ の事前分布である。例えば, 事前分布が一様分布(noninformative prior)で, かつ, $\text{Prob}(r^* | \langle \hat{R} \rangle)$ が r^* と $\langle \hat{R} \rangle$ を交換しても変わらない関数形（例： $\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(r^* - \langle \hat{R} \rangle)^2}{2\sigma^2}\right]$ ）であれば, 式(3.1)の 1 行目から 2 行目の代用は厳密であると考えられる。

課題 3-1. (★)

指数分布乱数データを N 個生成してヒストグラムを描きなさい。そのデータについて, BS percentile 法で母平均の区間推定を実施し, 正規母集団を仮定した区間推定と比較しなさい。信頼水準は 0.95 とする。 N の大きさに依存して 2 つの区間推定法の比較結果はどう変わるか観察し, その観察結果の理由を考察しなさい。（ヒント：復元抽出や並べ替え操作のために, 付録 G の numpy モジュールの関数を使ってよい。実数 x に最も近い整数は `int(x)` で得られる。）

BS t 法

BS t 法のアルゴリズムを説明する前に下準備をする。図 3.1 の実世界での標本サンプリングにおいて, 式(2.4)の一般化したスチューデントの T の実現値は

$$t \equiv \frac{r - \langle R \rangle}{\text{se}}, \quad \cdots (3.2)$$

と表せる。ここで r は R の実現値で, se は R の標準誤差(R の不偏分散の平方根)の実現値である。この実世界での T の実現値 t をよく近似する, Bootstrap 世界での T の実現値 t^* はどのように表せばよいだろうか。これは, 経験分布 \hat{f} の BS 標本 $\vec{x}_{(1)}^*, \vec{x}_{(2)}^*, \dots, \vec{x}_{(B)}^*$ から算出した統計量 R の BS 標本 $r_{(1)}^*, r_{(2)}^*, \dots, r_{(B)}^*$ について,

$$t_{(j)}^* \equiv \frac{r_{(j)}^* - \langle \hat{R} \rangle}{\text{se}_{(j)}^*}, \quad j = 1, 2, \dots, B. \quad \cdots (3.3)$$

と表せる。式(3.3)での $\langle \hat{R} \rangle$ は経験分布 \hat{f} における統計量 R の BS 期待値なので、実標本 \vec{x} についての R の実現値 r に等しい。そのため、式(3.3)の Bootstrap 世界での T の実現値は

$$t_{(j)}^* \equiv \frac{r_{(j)}^* - r}{\text{se}_{(j)}^*}, \quad j = 1, 2, \dots, B. \quad \cdots (3.4)$$

と書ける。ここで $\text{se}_{(j)}^*$ は、BS 標本値 $r_{(j)}^*$ を実現させた確率変数の標準誤差の実現値である。 $\text{se}_{(j)}^*$ を評価するには、 j 番目の BS 標本 $\vec{x}_{(j)}^*$ からさらに経験分布 $\hat{f}_{BS(j)}$ を構成し、母集団分布 $\hat{f}_{BS(j)}$ からランダムに復元抽出した K 個の BSBS 標本 $\vec{x}_{(j)(1)}^{**}, \vec{x}_{(j)(2)}^{**}, \dots, \vec{x}_{(j)(K)}^{**}$ を用いて、 $r_{(j)}^*$ を実現させた確率変数の K 個の実現値 $r_{(j)(1)}^{**}, r_{(j)(2)}^{**}, \dots, r_{(j)(K)}^{**}$ を算出する。 $\text{se}_{(j)}^*$ はこれら K 個の値の不偏分散値の平方根

$$\text{se}_{(j)}^* = \sqrt{\frac{1}{K-1} \sum_{k=1}^K (r_{(j)(k)}^{**} - r_{(j)(\cdot)}^{**})^2}, \quad \cdots (3.5)$$

である。ここで $r_{(j)(\cdot)}^{**} \equiv \frac{1}{K} \sum_{k=1}^K r_{(j)(k)}^{**}$ である。BSBS 標本を BS 標本と区別するため、上付き添え字 ** を付けた。

2 章で扱った正規母集団の場合、不動統計量であるスチューデントの T を用いると、母平均や母分散の影響を受けない区間推定法を構成できた。本章で扱っている未知母集団の場合でも近似的に不動性をもつ上記のスチューデントの T を用いることで、母集団の位置や形状の影響を受けにくい安定な区間推定法を構成できる。

アルゴリズム

経験分布 \hat{f} の BS 標本 $\vec{x}_{(1)}^*, \vec{x}_{(2)}^*, \dots, \vec{x}_{(B)}^*$ とそこから算出した統計量 R の BS 標本 $r_{(1)}^*, r_{(2)}^*, \dots, r_{(B)}^*$ を用いて、式(3.4)の Bootstrap T の実現値 t^* を算出する

$$t_{(j)}^* \equiv \frac{r_{(j)}^* - r}{\text{se}_{(j)}^*}, \quad j = 1, 2, \dots, B. \quad \cdots (3.4 \text{ 再掲})$$

$\text{se}_{(j)}^*$ は二重 BS 標本を使って式(3.5)により算出したものである。式(3.5)の $t_{(j)}^*$ を昇順に並べ替えておき $L = \text{integer} \left[\frac{\alpha}{2} B \right]$, $U = \text{integer} \left[\left(1 - \frac{\alpha}{2}\right) B \right]$ とおくと、 t^* の分布の $\frac{\alpha}{2}$ 分位値, $1 - \frac{\alpha}{2}$ 分位値がそれぞれ $t_{(L)}^*$, $t_{(U)}^*$ として得られる。これを用いて、以下のように区間推定を定義する。

$$\begin{aligned}
1 - \alpha &= \text{Prob}\{t_{(L)}^* \leq t^* \leq t_{(U)}^*\} \\
&\approx \text{Prob}\{t_{(L)}^* \leq t \leq t_{(U)}^*\} \\
&= \text{Prob}\left\{t_{(L)}^* \leq \frac{r - \langle R \rangle}{\text{se}} \leq t_{(U)}^*\right\} \\
&= \text{Prob}\{t_{(L)}^* \text{se} \leq r - \langle R \rangle \leq t_{(U)}^* \text{se}\} \\
&= \text{Prob}\{-t_{(U)}^* \text{se} \leq \langle R \rangle - r \leq -t_{(L)}^* \text{se}\} \\
&= \text{Prob}\{r - t_{(U)}^* \text{se} \leq \langle R \rangle \leq r - t_{(L)}^* \text{se}\}, \quad \dots (3.6)
\end{aligned}$$

1 行目から 2 行目への近似は BS 世界から実世界への移行であり(注 b), $N \rightarrow \infty$ で厳密となる。

以上の手続きにより, 統計量 R の期待値 $\langle R \rangle$ の信頼水準 $1 - \alpha$ での推定区間は, $[r - t_{(U)}^* \text{se}, r - t_{(L)}^* \text{se}]$ として算出される。

(注 b) 不動性に優れた統計量 T において BS 世界から実世界への移行を行うことは, 真の分布 f と経験分布 \hat{f} の違いによる誤差を抑える効果がある。これに対して, BS percentile 法では, 一般には不動性を全く持たない生の統計量 R で世界間の移行を行ってしまっている。

上記 BS t 法の実装と実行のコード例を以下に示す。

```
##### BS t 法のコード実装例 (N.Moteki, 2021)
import numpy as np
rng = np.random.default_rng();

# データ 例: 正規分布 No(pop_mean, pop_var) に従う観測量 x
pop_mean = 0.0 # 母平均
pop_var = 2.0 # 母分散
mu = pop_mean
sigma = np.sqrt(pop_var)
N = 20 # 実標本数
x = rng.normal(loc=mu, scale=sigma, size=N); # 観測値 (実標本) の生成

## BS 法による<R>の区間推定 (ここでは R は標本平均, <R>は母平均)
B = 10000 # BS 標本数
K = 25 # BSBS 標本数
r = np.mean(x) # (この例では 統計量 R は 標本平均 としている)
rbs = np.empty(B) # r* 格納用の要素数 B の 1D 配列
tbs = np.empty(B) # t* 格納用の要素数 B の 1D 配列
for j in range(0, B):
    xbs = rng.choice(x, size=N, replace=True) # 実標本の復元抽出により BS 標本を生成
    rbs[j] = np.mean(xbs) # R の BS 標本 (この例では統計量 R は 標本平均 としている)
    rbsbs = np.empty(K)
    for k in range(0, K):
        xbsbs = rng.choice(xbs, size=N, replace=True) # BSBS 標本のサンプリング
        rbsbs[k] = np.mean(xbsbs) # R の BSBS 標本 (この例では統計量 R は 標本平均 としている)
    se_rbs = np.std(rbsbs, ddof=1) # rbs[j] の標準誤差 (式(3.5))
    tbs[j] = (rbs[j] - r) / se_rbs # Bootstrap t value
    se_r = np.std(rbs, ddof=1) # standard error of r

tbs = np.sort(tbs)
alpha = 0.05
L = int(alpha/2*B)
U = int((1-alpha/2)*B)
```

```
#前頁からのつづき
tbs_l= tbs[L] #t*の alpha/2 分位値
tbs_u= tbs[U] #t*の 1-alpha/2 分位値
R_expect_l= r-tbs_u*se_r
R_expect_u= r-tbs_l*se_r
print(R_expect_l,R_expect_u) #BS t 法の区間推定結果
```

図 3.3. BS t 区間推定法のコード実装例

for ループが 2 重になっているのは、1 つの実標本から B 個の BS 標本をサンプリングし、各々の BS 標本から K 個の BSBS 標本をサンプリングしているためである。

BS t 法は標本平均以外の位置統計量(location statistics)についてもうまく機能することが知られている [Efron and Tihshirani 1993]。位置統計量とは、元データ \vec{x} の各要素に定数 c を足したデータの統計量の実現値が $r + c$ になるような統計量のことである。例えば、中央値、その他の分位値、刈り込み平均などは位置統計量である。

課題 3-2. (★)

指数分布乱数データを N 個生成してヒストグラムを描きなさい。そのデータについて、BSt 法で母平均の区間推定を実施し、正規母集団を仮定した区間推定と比較しなさい。信頼水準は 0.95 とする。 N の大きさに依存して 2 つの区間推定法の比較結果はどう変わるか観察し、その観察結果の理由を考察しなさい。2 手法の優劣について考察しなさい。信頼水準は 0.95 とする。BSt 法のコーディングは図 3.3 の例を参考にしてよい。

課題 3-3. (★★)

サンプル気象データ（付録 E）のうちいずれか 1 つの観測量（例：前橋の最大 1h 降水量）について、BS percentile 法、BS t 法、正規母集団を仮定した方法による母平均の区間推定を実施しなさい。標本数 N が多いとき(全データ)と少ないとき(最近 20 年分)の各条件において 3 手法を比較し、3 手法の信頼性の優劣について意見を述べなさい。

課題 3-4. (★★)

指数分布乱数データを N 個生成してヒストグラムを描きなさい。そのデータについて、BS percentile 法、BS t 法により「母中央値」の区間推定を実施しなさい。参考のため、正規母集団を仮定した母平均の区間推定も実施しなさい。上記 3 つの区間推定結果を水平線分でヒストグラムに重ね描きしなさい。標本数 N が多いときと少ないときの各条件で 3 つの結果を比較し、母中央値の区間推定法としての信頼性の優劣について意見を述べなさい。

3.3. 2つの未知母集団の比較(Permutation 検定)

最後に、2つの未知母集団それぞれから観測データ $\vec{x}=\{x_i; i=1, \dots, N_1\}$, $\vec{y}=\{y_i; i=1, \dots, N_2\}$ を得たとき、2つの母集団の関係を推察する方法として Permutation 検定を扱う。2母集団間の関係を記述する任意の統計量を

$$R = R(X_1, \dots, X_{N_1}, Y_1, \dots, Y_{N_2}) = R(\vec{X}, \vec{Y}), \quad \cdots (3.7)$$

とする。 $R(\vec{X}, \vec{Y})$ としては例えば、標本平均の差 $\bar{X} - \bar{Y}$ や、中央値の差、要素数が等しい場合には相関係数 $\rho(\vec{X}, \vec{Y})$ などが考えられる。Permutation 検定は、「 \vec{X} の母集団と \vec{Y} の母集団が、統計量 $R(\vec{X}, \vec{Y})$ の尺度で同一である」という帰無仮説を検定する方法である。通常、差についての Permutation 検定は片側検定で行われ、差がないという帰無仮説が棄却された際には、いずれかの方が大きい（小さい）という対立仮説が採択される。統計量 R の実現値が、 R の帰無分布の片側において有意水準 α 未満の p 値を示したとき、帰無仮説は棄却され、対立仮説が採択される。

差の Permutation 検定

「統計量 R の尺度で 2つの母集団には差がない」という帰無仮説を検定する方法は次の通り。

元データ \vec{x} と \vec{y} の全ての要素をランダムにシャッフルして要素数 N_1 のグループ \vec{x}^* と要素数 N_2 のグループ \vec{y}^* に割り振り、これらのグループについて統計量 R の実現値を算出する試行を多数回繰り返す。その実現値の頻度分布を帰無分布として帰無仮説を検定する。

アルゴリズム

1. 観測データ $\vec{x}=\{x_i; i=1, \dots, N_1\}$, $\vec{y}=\{y_i; i=1, \dots, N_2\}$ から、式(3.7)の統計量 R の実現値 $r = R(\vec{x}, \vec{y})$ を評価する。
 2. \vec{x} , \vec{y} をまとめて 1 次元数値配列 $\{x_1, \dots, x_{N_1}, y_1, \dots, y_{N_2}\} = \{z_i; i=1, \dots, N_1 + N_2\} \equiv \vec{z}$ とする。
 3. \vec{z} の要素の順番をランダムに並べ替える。
 4. \vec{z} の最初の N_1 要素を \vec{x}^* , 残りの N_2 要素を \vec{y}^* として取り出し、統計量 R の実現値 $r^* = R(\vec{x}^*, \vec{y}^*)$ を算出する。
 5. ステップ 3-4 を B 回 ($\geq \sim 10000$) 繰り返し、 r^* の標本 $\{r_{(1)}^*, r_{(2)}^*, \dots, r_{(B)}^*\}$ を取得する。 r^* の標本分布を Permutation 分布と呼び、これを R の帰無分布とする。
 6. 下側検定の場合、 $r^* < r$ となる r^* の標本数 K を数え、 $p \approx K/B$ として検定の p 値を算出する。上側検定の場合、 $r < r^*$ となる r^* の標本数 K を数え、 $p \approx K/B$ として検定の p 値を算出する。
-

差の Permutation 検定は 2 章で扱った 2つの正規母集団の母平均の差の検定を、2つの未知母集団・任意統計量の差の検定に一般化したものである。

上記 Permutation の組みあわせ総数は、 $N_1 + N_2$ 個の要素から N_1 個の要素を非復元抽出するときの組み合わせの数に等しく、 $N_1 + N_2 \text{C}_{N_1}$ 通りである。 $N_1 = N_2 = 10$ のときですら 18 万通りもある。多くの場合、実際に抽出される B 通り ($\geq \sim 10000$) の標本は母集団のごく一部に過ぎない。そのため Permutation 検定では「高いランダムさをもつ並べ替え計算の実行」が必要不可欠である。

相関の Permutation 検定

確率変数 X, Y の実現値として要素数が等しい 2 つのデータ $\vec{x} = \{x_i; i = 1, \dots, N\}$, $\vec{y} = \{y_i; i = 1, \dots, N\}$ を得たとき、両者の相関の強さを表す統計量、例えば標本相関係数

$$R(\vec{x}, \vec{y}) \equiv \rho(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}, \quad \dots (3.8)$$

を尺度として、「2 つの確率変数には相関がない」という帰無仮説を検定する方法は次の通り。

元データ \vec{x}, \vec{y} いづれか一方の全要素をランダムにシャッフルして統計量 R の実現値を算出する試行を多数回繰り返す。その実現値の頻度分布を帰無分布として帰無仮説を検定する。

アルゴリズム

1. 観測データ $\vec{x} = \{x_i; i = 1, \dots, N\}$, $\vec{y} = \{y_i; i = 1, \dots, N\}$ から、式(3.8)の統計量 R の実現値 $r = R(\vec{x}, \vec{y})$ を評価する。
 2. \vec{x} の要素の順番をランダムに並べ替えてそれを \vec{x}^* とする。
 3. 統計量 R の実現値 $r^* = R(\vec{x}^*, \vec{y})$ を算出する。
 4. ステップ 2-3 を B 回 ($\geq \sim 10000$) 繰り返し、 r^* の標本 $\{r_{(1)}^*, r_{(2)}^*, \dots, r_{(B)}^*\}$ を取得する。。 r^* の標本分布を Permutation 分布と呼び、これを R の帰無分布とする。
 5. 下側検定の場合、 $r^* < r$ となる r^* の標本数 K を数え、 $p \approx K/B$ として検定の p 値を算出する。
上側検定の場合、 $r < r^*$ となる r^* の標本数 K を数え、 $p \approx K/B$ として検定の p 値を算出する。
-

Permutation 検定は簡単に実装でき、尺度とする統計量がほぼ任意に選択可能であるため、使いやすく使える場面も多い 2 母集団間または 2 変量間の関係の検定方法である。コンピュータが利用できる場合、これを積極的に使わない理由はないと思われる。統計学の確立に貢献した学者の一人である Fisher (1890-1962) も、ランダム並べ替えの計算を実行できさえすれば、2 母集団の差の t 検定や 2 変量間の相関の t 検定の代わりに、より厳密な Permutation 検定をしたかったとのこと [Hesterberg 2014]。

課題 3-5. (★★)

サンプル気象データ（付録 E）の 3 つの観測量（気温、年降水量、最大 1h 降水量）のいずれか 1 つについて、東京と前橋の観測値のヒストグラムを重ね描きしなさい。つぎに、「東京と前橋の観測量の母平均は等しい」という帰無仮説を、Permutation 検定しなさい。片側検定で危険率は 0.05 とする。このときの p 値を示し、Permutation 分布と実現値 r を重ねて図示すること。

課題 3-6. (★★)

サンプル気象データ（付録 E）の 3 つの観測量（気温、年降水量、最大 1h 降水量）のいずれか 1 つについて、東京と前橋の観測値の散布図を描いて 2 地点のデータの相関関係を視覚化しなさい。「東京と前橋の観測値には相関がない」という帰無仮説を、Permutation 検定しなさい。片側検定で危険率は 0.05 とする。このときの p 値を示し、Permutation 分布と実現値 r を重ねて図示すること。（参考：2 変量データ (x,y) の散布図を描くには付録 G の plt.scatter(x,y) を使えばよい。）

付録

A. 基本公式

A1. 独立な確率変数の和の期待値・分散

N 個の互いに独立な確率変数 X_i ($i = 1, \dots, N$) の期待値と分散がそれぞれ

$$\mu_i \equiv \langle X_i \rangle, \quad \sigma_i^2 \equiv \langle (X_i - \langle X_i \rangle)^2 \rangle$$

($i = 1, \dots, N$) であるとき、これらの確率変数の和 $X(N) = \sum_{i=1}^N X_i$ の期待値と分散はそれぞれ $\sum_{i=1}^N \mu_i$ 、 $\sum_{i=1}^N \sigma_i^2$ となる。証明は例えば柴田 1995 (pp. 54-58) を参照のこと。ここで記号 $\langle X_i \rangle$ は確率変数 X_i の期待値を表す。

上記において各確率変数を $1/N$ 倍した場合の和を考えれば、確率変数の標本平均 $\bar{X}(N) = \frac{1}{N} \sum_{i=1}^N X_i$ の期待値と分散は、それぞれ $\frac{1}{N} \sum_{i=1}^N \mu_i$ 、 $\frac{1}{N^2} \sum_{i=1}^N \sigma_i^2$ になることが示せる。

A2. 標本分散の期待値

N 個の互いに独立な確率変数 X_i ($i = 1, \dots, N$) が同一分布に従い、各確率変数の期待値と分散がそれぞれ

$$\mu \equiv \langle X_i \rangle, \quad \sigma^2 \equiv \langle (X_i - \langle X_i \rangle)^2 \rangle$$

であるとする。標本分散の定義式(1.22)において、両辺の期待値をとると

$$\begin{aligned} \langle V^2 \rangle &= \frac{1}{N} \left\langle \left(\sum_{i=1}^N (X_i - \bar{X})^2 \right) \right\rangle = \frac{1}{N} \sum_{i=1}^N \langle (X_i - \bar{X})^2 \rangle = \frac{1}{N} \sum_{i=1}^N \langle [(X_i - \mu) - (\bar{X} - \mu)]^2 \rangle \\ &= \frac{1}{N} \sum_{i=1}^N \langle (X_i - \mu)^2 \rangle + \frac{1}{N} \sum_{i=1}^N \langle (\bar{X} - \mu)^2 \rangle - \frac{2}{N} \sum_{i=1}^N \langle (X_i - \mu)(\bar{X} - \mu) \rangle \end{aligned}$$

この 2 行目の式の第 1,2,3 項において、それぞれ関係式

$$\langle (X_i - \mu)^2 \rangle = \sigma^2,$$

$$\langle (\bar{X} - \mu)^2 \rangle = \langle (\bar{X} - \langle \bar{X} \rangle)^2 \rangle = \frac{1}{N^2} \langle (X(N) - \langle X(N) \rangle)^2 \rangle = \frac{1}{N^2} N \sigma^2 = \frac{\sigma^2}{N},$$

$$\begin{aligned} -\frac{2}{N} \sum_{i=1}^N \langle (X_i - \mu)(\bar{X} - \mu) \rangle &= -2 \left\langle \left(\frac{1}{N} \sum_{i=1}^N X_i - \frac{1}{N} \sum_{i=1}^N \mu \right) (\bar{X} - \mu) \right\rangle = -2 \left\langle \left(\bar{X} - \frac{1}{N} N \mu \right) (\bar{X} - \mu) \right\rangle \\ &= -2 \langle (\bar{X} - \mu)^2 \rangle = -2 \frac{\sigma^2}{N}, \end{aligned}$$

を用いると、標本分散 V^2 の期待値は

$$\langle V^2 \rangle = \frac{1}{N} N \sigma^2 + \frac{1}{N} N \frac{\sigma^2}{N} - 2 \frac{\sigma^2}{N} = \frac{N-1}{N} \sigma^2$$

となる。

A3. 正規分布の再生性

N 個の互いに独立な確率変数 X_i がそれぞれ正規分布 $No(\mu_i, \sigma_i^2)$ ($i = 1, \dots, N$) に従うとき、それらの確率変数の和 $X(N) = \sum_{i=1}^N X_i$ は、 N によらず、正規分布 $No(\sum_{i=1}^N \mu_i, \sum_{i=1}^N \sigma_i^2)$ となる。証明は例えば柴田 1995 (pp. 54-58) を参照のこと。

B. 正規母集団の不偏分散の分布定理

定理 B1: N 個の互いに独立な確率変数 $X_l \sim No(\mu, \sigma^2)$ ($l = 1, \dots, N$) について、標本平均 \bar{X} 、不偏分散 S^2 をとるとき、統計量

$$(N-1)\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{l=1}^N (X_l - \bar{X})^2, \quad (\text{B.1})$$

は自由度 $N-1$ の χ^2 分布に従う。

定理 B2: N 個の互いに独立な確率変数 $X_l \sim No(\mu, \sigma^2)$ ($l = 1, \dots, N$) の標本平均 \bar{X} と不偏分散 S^2 は互いに独立な確率変数である。

証明: $HH^T = I$ を満たす Helmert 行列

$$H^T = \begin{pmatrix} \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \cdots & \frac{1}{\sqrt{N}} \\ \frac{1}{\sqrt{1 \cdot 2}} & \frac{-1}{\sqrt{1 \cdot 2}} & 0 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{2 \cdot 3}} & \frac{1}{\sqrt{2 \cdot 3}} & \frac{-2}{\sqrt{2 \cdot 3}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3 \cdot 4}} & \frac{1}{\sqrt{3 \cdot 4}} & \frac{1}{\sqrt{3 \cdot 4}} & \frac{-3}{\sqrt{3 \cdot 4}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{N(N-1)}} & \frac{1}{\sqrt{N(N-1)}} & \frac{1}{\sqrt{N(N-1)}} & \frac{1}{\sqrt{N(N-1)}} & \cdots & \frac{-(N-1)}{\sqrt{N(N-1)}} \end{pmatrix}, \quad (\text{B.2})$$

を用いて、 N 個の互いに独立な確率変数 $X_l \sim No(\mu, \sigma^2)$ からなるベクトル $\vec{X} = (X_1, \dots, X_N)^T$ を $\vec{Y} = (Y_1, \dots, Y_N)^T$ に線形変換する。

$$\vec{Y} = H^T \vec{X}$$

これを Helmert 変換という (Basilevsky, A. 2005; Farhadian and Brenton 2020)。要素間の関係は

$$\begin{aligned}
Y_1 &= \frac{1}{\sqrt{N}}(X_1 + X_2 + \cdots + X_N) \\
Y_2 &= \frac{1}{\sqrt{2}}(X_1 - X_2) \\
Y_3 &= \frac{1}{\sqrt{6}}(X_1 + X_2 - 2X_3) \\
Y_4 &= \frac{1}{\sqrt{12}}(X_1 + X_2 + X_3 - 3X_4) \\
&\vdots \\
Y_N &= \frac{1}{\sqrt{N(N-1)}}(X_1 + X_2 + X_3 + \cdots + X_{N-1} - (N-1)X_N)
\end{aligned}, \tag{B.3}$$

となっている。 \vec{Y} の各要素の確率変数 Y_1, \dots, Y_N が互いに独立であることは（共分散 $\text{Cov}(Y_i, Y_j), i \neq j$ が全てゼロとなることから）わかる。式(B.3)から、 \vec{Y} の各要素の母集団分布は、

$$Y_1 \sim N(\sqrt{N}\mu, \sigma^2), \quad Y_2, \dots, Y_N \sim N(0, \sigma^2), \tag{B.4}$$

であることが分かる。また、Helmert 変換におけるノルムの不变性 $\vec{Y}^T \vec{Y} = \vec{X}^T H H^T \vec{X} = \vec{X}^T \vec{X}$ による関係 $\sum_{l=1}^N (Y_l)^2 = \sum_{l=1}^N (X_l)^2$ と、 $\bar{X} = Y_1 / \sqrt{N}$ であることを考慮すると、 \vec{X} の不偏分散 S^2 について

$$(N-1)S^2 = \sum_{l=1}^N (X_l - \bar{X})^2 = \sum_{l=1}^N (X_l)^2 - N(\bar{X})^2 = \sum_{l=1}^N (Y_l)^2 - (Y_1)^2 = \sum_{l=2}^N (Y_l)^2, \tag{B.5}$$

という関係が導かれる。両辺を σ^2 で割ると

$$(N-1) \frac{S^2}{\sigma^2} = \sum_{l=2}^N \left(\frac{Y_l}{\sigma} \right)^2, \tag{B.6}$$

左辺の $\sum_{l=2}^N (Y_l/\sigma)^2$ は、式(B.4)右側の関係により、 $N-1$ 個の独立な標準正規確率変数 $N(0,1)$ の 2 乗和なので、自由度 $N-1$ の χ^2 分布に従う。すなわち

$$(N-1) \frac{S^2}{\sigma^2} \sim \chi^2(N-1), \tag{B.7}$$

である。定理 B1 は示された。また、標本平均 \bar{X} は Y_1 のみに依存し、不偏分散 S^2 は Y_2, \dots, Y_N のみに依存するところが分かる。 Y_1, \dots, Y_N は互いに独立なので、 \bar{X} と S^2 は互いに独立である。定理 B2 は示された。■

C. 統計的仮説検定の枠組み

統計的仮説検定は背理法にもとづいている。背理法とは、「PならばQ」という命題と「QでないならばPでない」という命題（つまり互いに矛盾する命題の対偶）が論理的に同値であること

$$(P \Rightarrow Q) \Leftrightarrow (\bar{P} \Leftarrow \bar{Q}), \tag{C.1}$$

を利用した証明法である。ここで \bar{P} は P の否定を意味する。

統計的仮説検定では、論理式 $(P \Rightarrow Q)$ が真である条件下で、命題 \bar{Q} が真であることが（データから）示されたとき、命題 \bar{P} が真であるという結論が導かれる。一方、命題 \bar{Q} が真であることが示されないとときは、命題 P は肯定も否定もされない。つまり何も結論されず命題 P は無に帰する（このため、命題 P のことを帰無仮説（null hypothesis）と呼ぶ）。

統計的仮説検定では、帰無仮説Pが真であるとき、その確率密度分布 f_0 が一意的に定まるような検定統計量(test statistic) Z を定義しておく。この問題設定により、

$$Q: Z \sim f_0(Z), \quad (C.2)$$

という命題 Q について論理式($P \Rightarrow Q$)は恒真となる。 $f_0(Z)$ は仮説Pの帰無分布(null distribution)と呼ばれる。これと論理式(C.1)により、論理式($\bar{P} \Leftarrow \bar{Q}$)は恒真となっている。仮説検定の結果は、命題 \bar{Q} が真であること（つまり検定統計量 Z が帰無分布 f_0 に従わないこと）をデータから示せるか否かで決まるのであるが、その検定基準は以下の通り。

観測データ x を用いて検定統計量 Z の実現値 z を求める。もしこの z よりも稀なあらゆる実現値が出現する確率の和が極めてちいさいならば、「命題 \bar{Q} が真であることがデータから示された」と判断し、論理式(C.1)により命題 \bar{P} が真であることが示されたとする（帰無仮説 P は棄却される）。逆に、この z よりも稀なあらゆる実現値が出現する確率の和が極めてちいさくはないならば、「命題 \bar{Q} が真であることがデータから示された」と判断できないので、命題 \bar{P} が真であることが示せない（帰無仮説 P は棄却できない）。

帰無仮設Pが真である条件で、ある実現値 z よりも稀なあらゆる実現値が出現する確率の和である p 値は以下のように定義される。

$$p \equiv \begin{cases} \int_{-\infty}^z f_0(Z)dZ, & \text{下側} \\ \int_z^{\infty} f_0(Z)dZ, & \text{上側} \end{cases} \quad (C.3)$$

有意水準（危険率） α の検定において、 $p < \alpha$ であれば「命題 \bar{Q} が真であることがデータから示された」と判断され帰無仮説は棄却される。一方、 $p \geq \alpha$ であればそれが示せないので帰無仮説は棄却されない。個々の検定結果を示すときは、単に帰無仮説が棄却されるか否かだけでなく、 p 値も記載する。

もし $p \gg \alpha$ だったとしても、命題 Q が真である（統計量が帰無分布に従っている）とはいえない。なぜなら、統計量 Z についてのある1つの実現値 z だから、統計量 Z の母集団分布を、仮説 P の帰無分布 f_0 を含めた無数の候補分布から一意的に選択することは不可能だからである。

帰無仮説が棄却されたときに消去法的に採用されることになる「有意差がある」という仮説を、対立仮説とも呼ぶ。片側検定で「差がない」という帰無仮説が棄却された際には、有意差の方向も明示した対立仮説が採用される。有意水準（危険率） α の両側検定では、式(C.3)で算出される下側または上側の p 値を2倍した、「両側 p 値」が用いられる。

D. コルモゴロフ-スマルノフ検定

「観測値の度数分布（ヒストグラム）がある分布に適合している」という帰無仮説を検定するための汎用的な方法として、コルモゴロフ-スマルノフ(K-S)検定[cf. Massay 1951]がある。K-S 検定よりもよく知られた類似の方法として χ^2 分布を用いた適合度検定がある。 χ^2 分布の方法は、ヒストグラムの各階級の度数が少なくとも5は必要である（大標本で有効な近似）なのに対して、K-S 検定は大標本近似を仮定しておら

すこのような制約を受けない。このため、K-S 検定は観測値を（各階級の度数が十分多くなるように）分類する前処理を必要としないという利点がある[Hoel 1978]。

ある確率密度 $f(x)$ の累積分布 $F(x)$ は

$$F(x) = \int_{-\infty}^x f(u)du, \quad (\text{D.1})$$

として定義される。ある累積分布 $F(x)$ をもつ母集団から無作為抽出された標本を $\{x'_i; i = 1, \dots, N\}$ とし、それを小さい順に並べ替えた順序標本を x_1, x_2, \dots, x_N とする。この順序標本に基づく標本分布として、階段関数

$$S_N(x) \equiv \begin{cases} 0, & x < x_1 \\ \frac{k}{N}, & x_k \leq x < x_{k+1} \\ 1, & x \geq x_N \end{cases} \quad (\text{D.2})$$

を定義し、これと理論分布 $F(x)$ との適合性を検定する。これら 2 つの分布の”非適合度”を測るものとして、コルモゴロフ-スミルノフ統計量 D_N

$$D_N \equiv \sup_x |F(x) - S_N(x)|, \quad (\text{D.3})$$

を用いる。これはすべての可能な x における $F(x)$ と $S_N(x)$ のグラフの垂直距離の最大値をあたえる。 $S_N(x)$ は標本ごとに異なるから D_N は確率変数である。 D_N の理論分布とその導出は複雑である[Feller 2015]のでここでは省略する。 D_N の確率密度分布は N のみに依存し分布 $F(x)$ には依存しない。K-S 検定は統計量 D_N についての片側検定である。K-S 検定に用いる D_N 分布の $1 - \alpha$ 分位点（棄却限界値） D_N^α の数表は以下の通り。

表 D1. 自由度 N, 危険率 α の K-S 検定の棄却限界値 D_N^α の数表 [Massay 1951, Table 1]

Sample size (N)	Level of significance (α)				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
over 35	1.07	1.14	1.22	1.36	1.63
	\sqrt{N}	\sqrt{N}	\sqrt{N}	\sqrt{N}	\sqrt{N}

この表の D_N^α は、

$$\text{Prob}\{D_N \leq D_N^\alpha\} = 1 - \alpha, \quad (\text{D.4})$$

を満たす D_N 分布の限界値である。K-S 検定では、帰無仮説下での累積分布関数 $F(x)$ と観測データから統計量 D_N を算出し、 $D_N > D_N^\alpha$ であれば帰無仮説が棄却され、 $D_N \leq D_N^\alpha$ であれば棄却されない。

仮説検定と区間推定は表裏一体であるので、当然、K-S 統計量を用いると未知の累積分布関数 $F(x)$ の区間推定もできる。式(D.3) (D.4) から次のことがいえる。

$$\begin{aligned} 1 - \alpha &= \text{Prob}\left\{\sup_x |F(x) - S_N(x)| \leq D_N^\alpha\right\} \\ &= \text{Prob}\left\{\text{すべての } x \text{ について } |F(x) - S_N(x)| \leq D_N^\alpha\right\} \\ &= \text{Prob}\left\{\text{すべての } x \text{ について } S_N(x) - D_N^\alpha \leq F(x) \leq S_N(x) + D_N^\alpha\right\} \end{aligned}$$

つまり、観測データと表 D1 から定まる 2 つの階段関数 $S_N(x) - D_N^\alpha$ と $S_N(x) + D_N^\alpha$ が、信頼水準 $1 - \alpha$ のもとで、標本が抽出された母集団の未知の分布関数 $F(x)$ の推定区間を与える。

E. 区間推定法の性能比較

区間推定アルゴリズムでは、下側・上側境界それぞれに正確さが要求される。片側境界の外側が真値(R)を被覆する確率が理想値 α からどれだけずれているか(被覆誤差：coverage error)が誤差評価の一つの基準である。以下の説明では下側境界 $\langle R \rangle_{l,est}$ の場合のみ例示するが上側境界 $\langle R \rangle_{u,est}$ でも同様である。区間推定の誤差（ピック・オーライフ）が

$$\text{Prob}\{\langle R \rangle < \langle R \rangle_{l,est}\} = \alpha + O(N^{-1/2}) \quad (\text{E.1})$$

であるアルゴリズムは、1次の確度(first order accurate)といい、

$$\text{Prob}\{\langle R \rangle < \langle R \rangle_{l,est}\} = \alpha + O(N^{-1}) \quad (\text{E.2})$$

であるアルゴリズムは、2次の確度(second order accurate)であるという。例えば、正規母集団を仮定した母平均の区間推定法、BS percentile 法は1次確度、BS t 法は2次確度であることが知られている[Efron and Tibshirani 1993]。確度次数とは別に評価する必要がある重要な性能指標として、 N が小さいときに顕在化する系統誤差(small sample bias)がある。たとえば、BS percentile 法は N が小さいときに区間幅を過小評価することが知られている。

様々な BS 区間推定アルゴリズムの性能を比較した Hesterberg [2014] の Figure20,21 を示しておく。

Figure 20 は正規分布母集団データ、Figure21 は指数分布母集団データについて、式(E.1)左辺の片側被覆確率をデータ要素数 n に対してプロットしたものである。横軸各点において、縦軸値の導出には 10000 回ほどの「実標本抽出→区間推定」の試行を行ったとのことである。Figure 20,21 のプロットレジェンドにおいて、本資料で扱った 3 つの手法との対応は、t: 正規母集団仮定の方法、perc: BS percentile 法、bootT: BS t 法 である。

正規分布母集団からサンプリングされたデータ(Figure20)の区間推定では、当然、正規母集団仮定の方法は正確であり、BS t 法も同等の正確さを示している。BS percentile 法は n が小さいときに被覆確率を顕著に過大評価（すなわち推定区間幅を過小評価）している。左右非対称な分布である指数分布母集団からサンプリングされたデータ(Figure21)の区間推定では、一般に下側と上側で被覆確率の誤差が手法に依存して異なる。正規母集団仮定の方法や、BS percentile 法に比べて、BS t 法の誤差が小さいことが明らかである。Figure 21 における誤差の n 依存性は、正規母集団仮定の方法や、BS percentile 法では式(E.1)と、BS t 法では式(E.2.)と整合的であることが Hesterberg [2014, Figure 22] で示されている。

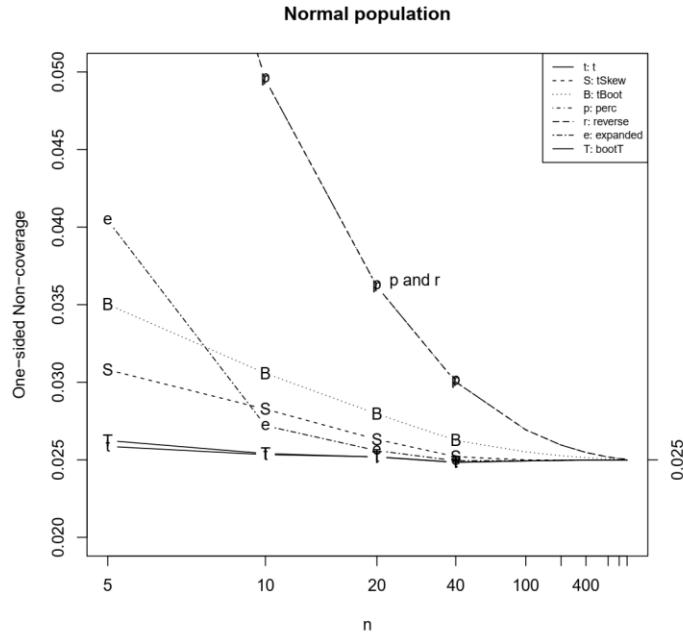


Figure 20: Confidence interval one-sided miss probabilities for normal populations. The intervals are described at the beginning of Section 5.6. Only one side is shown, because non-coverage probabilities are the same on both sides.

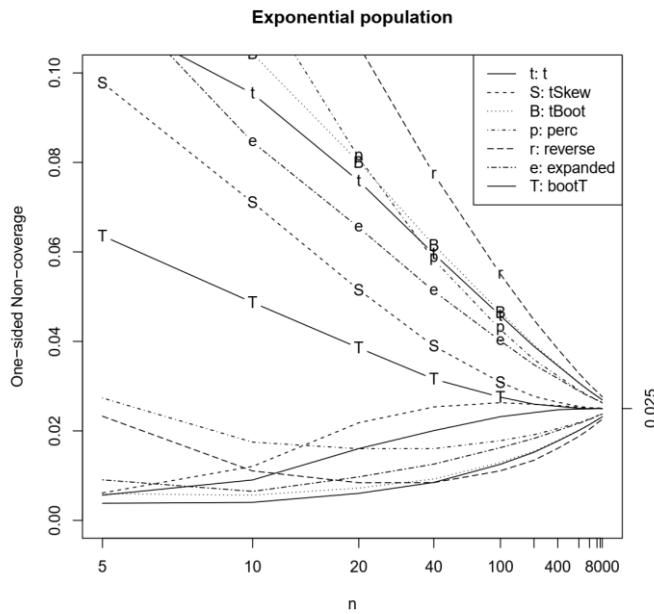


Figure 21: Confidence interval one-sided miss probabilities for gamma populations. The intervals are described at the beginning of Section 5.6. The lines with codes are non-coverage probabilities on the right, where the interval is below θ . The lines without codes correspond to the left side.

F. サンプル気象データ

区間推定や仮説検定の手法を適用してみる観測データの例として、気象庁 HP「過去の気象データ」から取得した、東京と前橋における過去 70 年間(1951~2020 年)の年平均気温、年降水量、年間最大 1h 降水量の時系列値を用意した(配布ファイル SampleData.csv)。T、Pr、MaxHourPr という接頭名のデータ列はそれぞれ年平均気温、年降水量、最大 1h 降水量を表す。

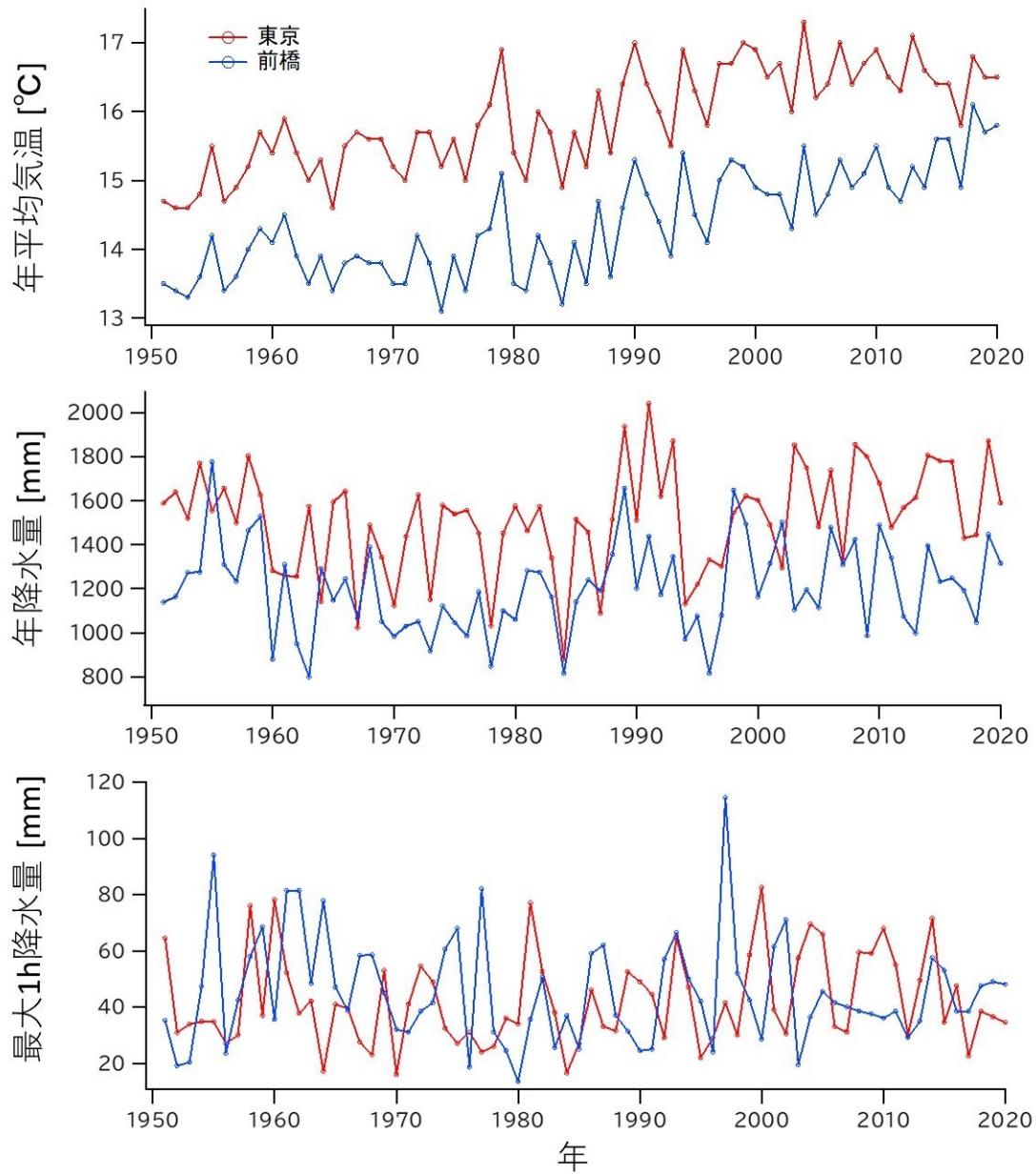


図 E.1. サンプル気象データの時系列図

G. Python 覚書

本演習のサンプルコード・解答で使用した python モジュールのリスト

Math モジュール

```
import math

math.pi           # 円周率π
math.gamma(x)    # ガンマ関数 (引数x)
```

Numpy モジュール

```
import numpy as np

配列
x= np.array(list) # リスト型オブジェクト list を 1D 配列(np.ndarray 型)x に変換
list= x.tolist() # 1D 配列 x をリスト型オブジェクト list に変換
x= np.empty(N) #長さ N の 1D 配列を作成
x= np.append(x,y) # 1D 配列 x の末尾に 1D 配列 y を追加
x= np.sort(x) # 1D 配列 x を昇順に並べ替え
len(x) #配列の長さを取得
```

数学関数 (配列 x の要素ごとの演算)

```
np.exp(x)          #指数関数
np.sin(x)          #正弦関数
np.log(x)          #底 e の対数関数
np.log10(x)        #底 10 の対数関数
```

統計関数

```
np.mean(x)         # 標本平均
np.var(x)          # n で割る方式の分散
np.var(x, ddof=1) # n-1 で割る方式の分散 (不偏分散)
np.std(x)          # n で割る方式の標準偏差
np.std(x, ddof=1) # n-1 で割る方式の標準偏差 (不偏分散の平方根)
```

```
np.median(x)          # 中央値
```

算術

```
np.quantile(x, alpha)      # 1D 配列 x の alpha 分位数 (線形内挿済み)
np.count_nonzero(x > a)    # 1D 配列 x の要素のうち a より大きなものの個数 (これは使用例)
r= np.corrcoef(x, y)[0,1]  # 同じ長さの 2 つの 1D 配列 x, y の標本相関係数 r
```

ランダム関係

```
rng= np.random.default_rng()  # 乱数発生機オブジェクト rng の生成
x= rng.uniform(low=a, high=b, size=n)  # 一様乱数 u[a,b] の 1D 配列 x を生成(要素数 n)
x= rng.normal(loc=mu, scale=sigma, size=n)  # 平均 mu, 標準偏差 sigma の正規分布乱数の 1D 配列 x を生成(要素数 n)
x= rng.exponential(scale=beta, size=n)  # 平均 beta の指数分布乱数の 1D 配列 x を生成(要素数 n)
x= rng.standard_t(df=nyu, size=n)  # 自由度 nyu の t 分布乱数の 1D 配列 x を生成(要素数 n)
rng.shuffle(x)  # 1D 配列 x の要素をランダムに並べ替える(インプレイス)
y= rng.choice(x, size=n, replace=True)  # 1D 配列 x の要素をランダムに復元抽出して 1D 配列 y を生成(要素数 n)
```

Scipy モジュール

```
from scipy import stats
```

t 分布関数の alpha 分位点(alpha*100 パーセンタイル点)の算出

```
stats.t.ppf(q=alpha, df=N) # 自由度 N の t 分布の下側 alpha 分位点(alpha*100 パーセンタイル点)
```

t 検定-1 母集団

```
result= stats.ttest_1samp(x, popmean=mu0) # 「1D 配列 x のデータの正規母集団の母平均が mu0 である」という帰無仮説を両側 t 検定, t 統計量 result.statistic と両側 p 値 result.pvalue が出力される。
```

t 検定-2 母集団

```
stats.ttest_ind(x, y) # 「2 つの 1D 配列 x, y のデータの正規母集団（等母分散）の母平均 が等しい」という帰無仮説を両側 t 検定, t 統計量 result.statistic と両側 p 値 result.pvalue が出力される。
```

コルモゴロフ-スミルノフ検定

```
dstat, pval = stats.kstest(x, stats.norm.cdf, args=(mu, sigma), alternative='two-sided') # 「1D 配列 x のデータが、平均 mu, 標準偏差 sigma の正規分布から抽出された」という帰無仮説を両側 KS 検定、KS 統計量 dstat と p 値 pval が出力される。
```

Matplotlib モジュール

```
import matplotlib.pyplot as plt
```

プロット描画

```
plt.figure(figsize=(12,4)) #図のサイズを横 12、縦 4 に指定

plt.plot(x, y, marker="o") #数値配列 y を数値配列 x に対して線プロット

plt.scatter(x, y) #数値配列 y を数値配列 x に対して散布図プロット

plt.hist(x, density=False) #数値配列 x のヒストグラムプロット (density=True で確率密度)

plt.hist([x1,x2], label=["x1","x2"]) #数値配列 x1,x2 のヒストグラムを重ねて表示。

plt.legend(loc='upper left') #左上にプロットの凡例を表示

plt.xlim(a,b) #x 軸の範囲を [a,b] に設定

plt.ylim(a,b) #y 軸の範囲を [a,b] に設定

plt.title("title") #プロットのタイトルを設定

plt.xlabel("xlabel") #x 軸のラベルを設定

plt.ylabel("ylabel") #y 軸のラベルを設定
```

その他のオプションの指定方法は web で調べてください。

Pandas モジュール

```
import pandas as pd
```

サンプル気象データの読み込み

```
df = pd.read_csv('SampleData.csv') #csv データを DataFrame へ読み込み

df1= df[df['Year']>=2000] # 'Year' が 2000 以上の行のみ抽出して df1 に格納

df= df.tail(n) # 末尾の n 行の抽出

x= df['colname'].values #DataFrame の指定列を 1 次元 numpy 配列 x に変換
```

参考文献

本資料を作成する情報源とした書籍・論文は下記である。これらの多くは文中引用している。

- 柴田文明 (1995), 理工系の基礎数学シリーズ7「確率・統計」, 岩波書店
- Osgood, B. G. (2019). Lectures on the Fourier transform and its applications (Vol. 33). American Mathematical Soc.
- 吉澤康和 (1989) 新しい誤差論, 共立出版
- 清水良一 (1976), 新しい応用の数学シリーズ「中心極限定理」, 教育出版
- 萩谷千鳳彦 (2010), 「統計分布ハンドブック」, 朝倉書店
- 北川源四郎 (2005), 時系列解析入門, 岩波書店
- 尾畠伸明 (2014), クロスセクショナル統計シリーズ1「数理統計学の基礎」, 共立出版
- 杉山将 (2015), 機械学習プロフェッショナルシリーズ「機械学習のための確率と統計」, 講談社
- 竹村彰通 (2020), 「現代数理統計学(新装改訂版)」, 学術図書出版
- Basilevsky, A. (2005), Applied Matrix Algebra in the Statistical Sciences, Dover Publicaiton.
- Farhadian, R., and Brenton C. (2020), A note on the Helmert transformation, *Communications in Statistics-Theory and Methods*, <https://doi.org/10.1080/03610926.2020.1836223>.
- Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*. 7 (1): 1–26.
- Efron and Hastie 著 藤澤・井出 監訳 (2020)「大規模計算時代の統計推論」, 共立出版
- 汪 金芳・桜井裕二 (2011) Rで学ぶデータサイエンス4「ブートストラップ入門」, 共立出版
- Efron and Tihshirani (1993), "An introduction to the Bootstrap", Springer-science + business media.
- Efron, B. (1987), Better Bootstrap confidence interval, *Journal of the American Statistical Association*, 82:397, 171-185.
- Hesterberg, T. C. (2014). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum (longer version), <http://arxiv.org/abs/1411.5279>.
- Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4), 371-386.
- Massay, F. J. Jr. (1951), The Kolmogorov-Smirnov Test for Goodness of Fit, *Journal of the American Statistical Association*, 46:253, 68-78.
- Hoel, P. G. 著, 浅井晃・村上正康(訳) (1978), 「入門数理統計学」, 培風館
- Feller, W. (2015). On the Kolmogorov-Smirnov limit theorems for empirical distributions. In Selected Papers I (pp. 735-749). Springer, Cham.