講義ノート

統計学

統計的推定と検定

村田 昇

version: 2020 年 6 月 15 日

早稲田大学 先進理工学部
電気・情報生命工学科

はじめに

工学においては「不確定性を伴なって観測されるデータから対象とする現象の数学的なモデルを構築する」といった問題に直面することが多々あります．例えば実験データの解析や，ノイズを伴なって観測されるシステムの入出力関係の同定といった問題です．典型的には

- 「データの生成機構は不確定性を持っている」
- 「不確定性を記述するための (パラメトリック) モデルを作る」
- 「我々が現実の場面で利用できるのは，この生成機構から出てくる有限個のサンプルである」
- 「この有限個のサンプルを使ってモデルのパラメタを推測する」

といった条件で数学的に妥当なモデルを考えることになります．このような場面では不確定性や曖昧さを記述する手法として確率モデルが重要な役割を果します．

この講義では，信号処理・情報理論・制御理論・学習理論といった分野で必要とされる確率モデルを扱うために，確率論と数理統計学の基本的な枠組を知ってもらうことを目的とします．

確率・統計は対にして語られることが多いですが，その考え方は対照的です．確率は演繹的 (deductive) であり，ある確率的な構造をもつシステムから観測されるデータの性質を論じることを目的とします．一方統計は帰納的 (inductive) であり，観測されたデータから，それを生成したシステムの確率的な構造を推測することが目的となり，どちらも重要な工学の基礎概念となります．

本講義で扱う「確率」においては，場合の数を数え上げる組合せ論的なものではなく，確率空間を基礎とする確率の現代的な枠組みを知ってもらい，工学の様々な場面で必要な基本的知識を身につけることを目標とします．特に

- 確率空間
- 様々な確率分布とそれに関わる量
- 独立な確率変数の和に関する定理

といった項目について学んでもらいます．

一方「統計」では数理統計学の基本となる考え方を紹介します．この枠組は学習や制御あるいは信号処理といった分野と関連が深く，特に線形モデルを主体とした統計的手法は本質的に非線形なモデルを扱わなくてはいけない様々な場面の大事な基礎となります．主として

- 推定 (不偏推定，最尤推定，ロバスト推定など)
- 検定

を取り上げ，統計的な考え方を学んでもらいます．また時間が許せば

- ベイズ統計
- 多変量解析 (回帰分析，主成分分析，判別分析など)

といった項目についても簡単に説明します．

教科書は特に指定しませんが，各項の末尾にいくつか参考書を挙げてあるので，必要に応じて参照して下さい．

目次

1 母数の推定

1.1 確率モデルと推定量

これまでは，測度論的な確率の話をしてきたが，これから扱う統計的方法では，問題意識が以下のように異っている．

確率 ある決まった確率法則のもとで，確率変数がどのような性質を持つか論じる

統計 ある確率法則に従うと考えられる確率変数の実現値を観測して，それを生成する確率法則について何らかの推測を行う

工学的な問題としては後者の統計的な問題設定に接する機会が多いであろう．

1.1.1 観測値の生成モデル

例 1.1. 滴定によって水溶液の濃度を求める問題を考える．この問題では同じ条件の元で繰り返し実験データを測定することができるが，データには測定毎に何らかの誤差が生じる．滴定の場合には，上から落とす水溶液の一滴の量のばらつきや，濃度の微妙な不均一から起こるであろう色素の変化のタイミングのばらつきといったものが誤差の原因として考えられるであろう．この偶然に変動するばらつきを確率的なものとして捉えることによって

$$\underset{\text{(確率変数)}}{X_i} = \underset{\text{(未知母数)}}{\theta} + \underset{\text{(誤差・偶然変動)}}{\varepsilon_i}, \qquad i = 1, \ldots, n$$

という確率モデルを設定することができる．滴定問題の場合，未知の母数 (パラメタ; parameter) θ は実験で求めたい水溶液の濃度の真の値に対応し，実験で得られる観測値は，この確率モデルで生成される確率変数の実現値であると考えられる．

さて以下では，未知母数に誤差が加法的に重畳するという上の最も簡単なモデルを取り扱い，実験を何回か行なって得られる観測値 $X_1, X_2, \ldots, X_n$ を用いて未知母数 θ を推定する方法を考える．

数学的に扱い易くするために誤差 ε には以下のような制約を課すことにする．

仮定 0 $\varepsilon_1, \ldots, \varepsilon_n$ はある同時確率分布に従う確率変数である．

これは誤差自体が確率変数であるという上の確率モデルの前提条件を言い換えたものである．

仮定 1 $\varepsilon_1, \ldots, \varepsilon_n$ は互いに独立に分布する．

実験が十分注意深く行われているのであれば，この仮定は合理的であろう．例えば繰り返し実験する際に，前の実験の影響が残るような実験は方法そのものを見直す必要がある．また系統誤差など実験者や実験環境によって測定結果に偏りが出るような測定方法も注意深く避ける必要がある．

仮定 2 $\varepsilon_1, \ldots, \varepsilon_n$ は同じ分布に従う．

各回の実験が完全に同じ条件で行われることを意味している．もしなんらかの理由で毎回条件が異なるのであれば，明示的にモデルに入れることを考える．例えば温度などによって実験結果が左右されることが判っているのであれば，温度変化を考慮した確率モデルを設定し，温度の測定も同時に行うといったことを考えなくてはいけない．

仮定 1 と 2 は多くの推定問題で設けられる標準的なものである．このように観測値が互いに独立に同じ分布に従う場合，「**観測値は** *i.i.d.* (independently, identically distributed) **である**」と言う．

仮定 3 $\mathbb{E}[\varepsilon_i] = 0,\ i = 1, \ldots, n$

この仮定は観測に偏りがないことを主張している．仮に偏りがあったとしてもデータからは知ることができないので，以降の議論を行なう上では最低限必要な仮定である．

仮定 4 $\mathbb{E}[\varepsilon_i^2] < \infty,\ i = 1, \ldots, n$

分散が有限であるというこの最後の仮定は数学上の便宜であり，大数の法則や中心極限定理などを使ってデータの性質を調べる際に必要となる．後ほど分散が発散する場合の例についても触れる．ただし，実際の実験データでは測定値が発散するような状況はほとんどないので，この制約は不合理なものではない．

以上の仮定を踏まえ，しばらくはこの簡単な確率モデルを観測値の生成モデルとして未知母数 θ を推定する方法を考え，推定で重要な概念を説明していく．

1.1.2 推定量と推定値

統計では未知の母数を推定する方式を**推定量** (estimator) と呼び，通常未知母数に $\hat{}$ を付けて表す．つまり

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

は n 個の確率変数 $X_1, X_2, \ldots, X_n$ を観測し，それらを用いて未知母数を推定する方法を表わしている．ここで推定量は確率変数 $X_1, X_2, \ldots, X_n$ の関数であるので，それ自体確率変数となることに注意する．

一方実際に実験をして $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ という観測値が得られたとしよう．これらの確率変数の実現値を上の推定量に代入して得られる

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$$

を**推定値** (estimate) と呼び，統計では推定量と区別することが多い．推定値は確率変数の実現値であり，確率変数ではないことに注意する．確率変数と実現値の区別は文献によっては曖昧にされることもあるが，ここではできるかぎり定義に忠実に書いていくことを試みる．

さて，前述の確率モデルにおける具体的な推定量としては，例えば以下のようなものが考えられ，実際良く用いられている．

推定量.**例** *1* (標本平均)

$$\hat{\theta} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

推定量.**例** *2* (中央値)

$$\hat{\theta} = (X_1, X_2, \ldots, X_n \text{ の中央の値})$$

推定量.**例** *3* (トリム平均) $\{X_i\}$ を小さい順に並べ換えたものを $\{X_{(i)}\}$ として

$$= \frac{1}{n-2m} \sum_{i=m+1}^{n-m} X_{(i)}$$

例 1 は観測される確率変数の算術平均で，**標本平均** (sample mean) と呼ばれる．誤差に偏りがないのであれば直感的にも θ に非常に近い値になるであろうことが予想される．

例 2 は**中央値** (median) と呼ばれるもので，誤差が正の側と負の側に均等に出るのであれば θ をうまく近似することが想像できるであろう．

例 3 は**トリム平均** (trimmed mean) と呼ばれるもので，言い換えると $X_1, X_2, \ldots, X_n$ を小さい順に並べ，小さい方の m 個と大きい方の m 個を捨てた算術平均である．極端に大きな，あるいは小さな値として現れる観測値を捨てることによって，特異的な値による誤差の影響を取り除いて平均を推定しようというものである．[1]

この他にも例えば

[1] 体操，フィギュアスケート，ジャンプといったスポーツ競技の採点法などではこれと似た方法が採られている．

推定量.**例** *4*

$$\hat{\theta} = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n, \quad \sum_{i=1}^{n} c_i = 1$$

推定量.**例** *5*

$$\hat{\theta} = (X_1 \cdot X_2 \cdot \cdots \cdot X_n)^{\frac{1}{n}},$$
$$(\text{ただし } X_1, X_2, \ldots, X_n \text{ は正の数とする})$$

推定量.**例** *6*

$$\hat{\theta} = \frac{n}{1/X_1 + 1/X_2 + \cdots + 1/X_n},$$

といったように，これらが良いかどうかは別として推定量を作ることはできる．例 4 は各観測の重みを変えて平均を取った加重平均であり，例 5 は観測値の幾何平均，例 6 は調和平均である．

1.2 推定量の良さ

上で述べたように，同じ問題においてもいろいろな推定量を構成することができる．以下では，多くの候補の中で望ましい推定量とはどういった性質をもつものであるかを考えていく．

1.2.1 不偏性

推定量に望まれる性質の一つに**不偏性** (unbiasedness) と呼ばれるものがある．

定義 1.2 (不偏性)**.** 推定量 $\hat{\theta}$ が**不偏** (unbiased) であるとは，真の母数が θ であるときに，そこから得られた観測値に基づく推定値の平均値が真の母数 θ に一致すること

$$\mathbb{E}\left[\hat{\theta} \,\middle|\, \theta\right] = \mathbb{E}^{X_1,X_2,\ldots,X_n}\left[\hat{\theta}(X_1, X_2, \ldots, X_n) \,\middle|\, \theta\right] = \theta$$

である．ただし $\mathbb{E}^{X}[\cdot \,|\, \theta]$ は真の母数が θ であるという条件のもとで確率変数 X について平均を取ることを表す．

例えば同じ実験環境で A 君，B 君，...，Z 君が個別に 10 回実験を行ない，それぞれが実験結果

$$\begin{aligned}
\text{A 君: } & x_1^{(A)}, x_2^{(A)}, \ldots, x_{10}^{(A)} \\
\text{B 君: } & x_1^{(B)}, x_2^{(B)}, \ldots, x_{10}^{(B)} \\
& \vdots \\
\text{Z 君: } & x_1^{(Z)}, x_2^{(Z)}, \ldots, x_{10}^{(Z)}
\end{aligned}$$

を得たとする．全員が同じ推定量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_{10})$ を使ったとしても実験結果が異なるのであるから，当然各人の得る推定値

$$\begin{aligned}
\hat{\theta}^{(A)} =& \hat{\theta}(x_1^{(A)}, x_2^{(A)}, \ldots, x_{10}^{(A)}) \\
\hat{\theta}^{(B)} =& \hat{\theta}(x_1^{(B)}, x_2^{(B)}, \ldots, x_{10}^{(B)}) \\
& \vdots \\
\hat{\theta}^{(Z)} =& \hat{\theta}(x_1^{(Z)}, x_2^{(Z)}, \ldots, x_{10}^{(Z)})
\end{aligned}$$

は異なり，これらの値は真の母数 θ のまわりにばらつくことになる．この推定値のばらつき方で推定量の良さを評価しようと考えた場合，不偏性は「推定値が真の母数 θ のまわりに偏りなくばらつく」ことを要請しているのである．

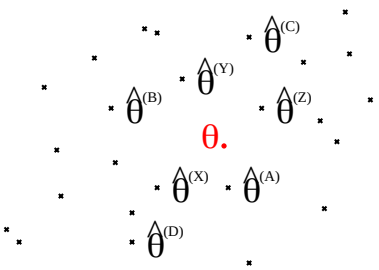

図 1.1: 真の母数と推定値のばらつきの概念図

例 1.3 (標本平均)**.** 先に挙げた推定量:例 1(標本平均) は

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
&= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]}{n} \\
&= \frac{\theta + \theta + \cdots + \theta}{n} \\
&= \theta
\end{aligned}$$

となり，平均値の不偏な推定量になっていることがわかる．

例 1.4 (不偏分散)．ここでは主に平均値の推定の話をしているが，実際の場面では推定値の性質を知るために誤差の分散を知りたいことがある．多くの場合，誤差の分散は未知であることが多いので，観測値から推定する必要がある．平均値の推定方法からの類推される単純な**標本分散** (sample variance)

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

を用いることもあるが，統計の様々な解析においては分散の不偏推定量である**不偏分散** (unbiased variance) を用いることが多い．標本分散の分母である観測値と標本平均の差の 2 乗和の平均を計算すると以下のようになる．

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&\quad = \mathbb{E}\left[\sum_{i=1}^{n}\{X_i - \theta - (\bar{X} - \theta)\}^2\right] \\
&\quad = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \theta)^2 - 2\sum_{i=1}^{n}(X_i - \theta)(\bar{X} - \theta) + \sum_{i=1}^{n}(\bar{X} - \theta)^2\right]
\end{aligned}$$

ここで $\sum_{i=1}^{n}(X_i - \theta) = n(\bar{X} - \theta)$ を用いると

$$\begin{aligned}
&= \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \theta)^2 - n(\bar{X} - \theta)^2\right] \\
&= \sum_{i=1}^{n}\mathrm{Var}(X_i) - n\mathrm{Var}(\bar{X}) \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

したがって分散の不偏推定量は

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

で与えられ，分母が n ではなく $n-1$ となることに注意する．

直観的には観測値の標本分散は良い推定値を与えるように思えるが，これは不偏ではなく

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

となり，若干小さめの値を与えることがわかる．これは n 個の観測値から平均を計算するために観測値一つ分に相当する情報を使ってしまったためとも解釈できる．

1.2.2 推定量の分散

さて平均値が正しい値になっていたとしても，そのばらつき方が大きかったらあまり意味がない．ばらつき方が大きいということ

は，例えば A 君の得た推定値 $\hat{\theta}^{(A)}$ は θ に非常に近いかもしれないが，B 君の得た推定値 $\hat{\theta}^{(B)}$ はとんでもなく離れているかもしれない．つまり当たり外れが大きいということである．実際の実験というのは上のように何人もの人が同じことを繰り返していくつもの推定値を比べられるような状況は稀で，ある一人の人が何回か実験をして推定値を一つだけ出すという一発勝負のようなものである．当然この場合は個々の推定値はできるだけ真の値の傍にいて欲しい訳であり，ばらつきは小さい方が好ましい．

ばらつき方を評価する規準はいろいろ考えられるが，一番簡単なものは推定量の分散

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]$$

であろう．

まず不偏性を要求したとき，推定量には次のような性質があることに注意しておく．

定理 1.5. $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ がいずれも不偏推定量であり，その分散が等しく σ^2 であるとする．すなわち

$$\mathrm{Var}(\hat{\theta}_1) = \cdots = \mathrm{Var}(\hat{\theta}_k) = \sigma^2$$

である．このときこれらの不偏推定量の単純な平均を

$$\hat{\theta}^* = \frac{1}{k}\sum_{i=1}^{k} \hat{\theta}_i$$

とおけば

$$\begin{aligned} \mathbb{E}[\hat{\theta}^*] &= \theta && \text{(不偏性が保存される)} \\ \mathrm{Var}(\hat{\theta}^*) &\leq \sigma^2 && \text{(分散が小さくなる可能性がある)} \end{aligned}$$

が成り立つ．

証明. 平均値の線形性から不偏性は明らか．ここで

$$\begin{aligned} &\sum_i (\hat{\theta}_i - \theta)^2 \\ &\quad = \sum_i (\hat{\theta}_i - \hat{\theta}^* + \hat{\theta}^* - \theta)^2 \\ &\quad = \sum_i (\hat{\theta}_i - \hat{\theta}^*)^2 + 2\sum_i (\hat{\theta}_i - \hat{\theta}^*)(\hat{\theta}^* - \theta) + \sum_i (\hat{\theta}^* - \theta)^2 \end{aligned}$$

$\hat{\theta}^* = \frac{1}{k}\sum_{i=1}^{k} \hat{\theta}_i$ より $\sum_i (\hat{\theta}_i - \hat{\theta}^*) = 0$ となることを用いると

$$= \sum_i (\hat{\theta}_i - \hat{\theta}^*)^2 + k(\hat{\theta}^* - \theta)^2$$

であることに注意して，両辺の平均を取ると

$$\begin{aligned} \sum_i \mathbb{E}[(\hat{\theta}_i - \theta)^2] &= \sum_i \mathbb{E}[(\hat{\theta}_i - \hat{\theta}^*)^2] + k\mathbb{E}[(\hat{\theta}^* - \theta)^2] \\ \sum_i \mathrm{Var}(\hat{\theta}_i) &= \sum_i \mathbb{E}[(\hat{\theta}_i - \hat{\theta}^*)^2] + k\mathrm{Var}(\hat{\theta}^*) \\ k\sigma^2 &= \sum_i \mathbb{E}[(\hat{\theta}_i - \hat{\theta}^*)^2] + k\mathrm{Var}(\hat{\theta}^*) \end{aligned}$$

となる．右辺の第 1 項は 0 または正なので

$$k\sigma^2 \geq k\mathrm{Var}(\hat{\theta}^*)$$

となり，分散が小さくなることがわかる． □

つまりこの定理は，分散によって推定量のばらつきの大きさを評価する場合，異なる不偏推定量があればそれらを併用することによって推定量の性質を良くすることができる可能性があることを示している．

特に推定量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ が $X_1, X_2, \ldots, X_n$ について対称でない場合には，対称化することによって性質を良くすることができる．

定理 1.6. $X_1, X_2, \ldots, X_n$ が互いに独立に同じ分布に従い，$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ を母数 θ の一つの不偏推定量とすると，$X_1, X_2, \ldots, X_n$ の対称な関数として表される不偏推定量 $\hat{\theta}^*$ で，つねに

$$\mathrm{Var}(\hat{\theta}^*) \leq \mathrm{Var}(\hat{\theta})$$

となるものが存在する．

証明. $\hat{\theta}$ を $X_1, X_2, \ldots, X_n$ について対称化，すなわち

$$\hat{\theta}^* = \frac{1}{n!}\sum \hat{\theta}(\{X_1, X_2, \ldots, X_n \text{ のあらゆる並べ替え}\})$$

を考え，前定理を使えばよい． □

したがって，推定量 $\hat{\theta}$ として不偏で分散の小さいものを考えるときには，$X_1, X_2, \ldots, X_n$ について対称なものだけを考えれば良いことがわかる．

既に挙げた推定量の 6 つの例のうち，最初の 3 つは $X_1, X_2, \ldots, X_n$ について対称な推定量になっている．例 2, 3 は一見対称でないように見えるが，結局大きさの順に並べ替えているので $X_1, X_2, \ldots, X_n$ の順によらないことに注意する．例 4 は $c_i = 1/n$ でない限り，あるいは大きさの順に並べ替えるといった操作をしない限りは対称でないことは明らかであろう．例 5，6 は対称ではあるが，ここで考える確率モデルにおいては一般に不偏になっていない．

2 平均値の推定

2.1 誤差の分布が未知の場合

最初に述べた仮定では，誤差の分布の形状については特に言及しなかったが，分布の形がわからない場合には実は観測値の平均 (標本平均) が最も良い推定量となる．

定理 2.1. 分布の形について何ら特別な仮定を入れられない場合

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

は θ の一様最小分散不偏推定量 *(uniformly minimum variance unbiased estimator)* である．

略証.

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \theta$$

なので $\hat{\theta}$ の不偏性は明らかである．

次に $\hat{\theta}$ 以外の $X_1, X_2, \ldots, X_n$ について対称な不偏推定量を $\hat{\theta}^\dagger$ として $T = \hat{\theta}^\dagger - \hat{\theta}$ とすると，不偏性の定義より

$$\mathbb{E}[T] = \mathbb{E}[\hat{\theta}^\dagger] - \mathbb{E}[\hat{\theta}] = \theta - \theta = 0$$

となる．

$X_1, X_2, \ldots, X_n$ の対称な関数 T において，任意の分布について

$$\mathbb{E}[T] = \mathbb{E}[T(X_1, X_2, \ldots, X_n)] = 0$$

となるのものは実は $T = 0$ に限る．これは次のようにして示される．

ここではまず $n = 2$ の場合を考えることにする．任意の分布を考えてよいので，$a < b$ の 2 点だけが，確率 $p, 1-p$ で出るような分布を考える．すると T の平均は

$$\mathbb{E}[T]$$

$$= p^2 T(a,a) + p(1-p)T(a,b) + (1-p)pT(b,a) + (1-p)^2 T(b,b)$$

$$= p^2 T(a,a) + 2p(1-p)T(a,b) + (1-p)^2 T(b,b)$$

$$= 0$$

となるが，これが任意の p について成り立つ必要がある．例えば $p = 0, 0.5, 1$ を代入すれば

$$T(a,a) = 0$$
$$0.25T(a,a) + 0.5T(a,b) + 0.25T(b,b) = 0$$
$$T(b,b) = 0$$

を満たさなくてはいけないが，これを解いて

$$T(a,a) = T(b,b) = T(a,b) = 0$$

であることがわかる．ここで a, b を自由に動かしても成り立たなくてはいけないので，結局 T は恒等的に 0 であることがわかる．

n が 2 以上の場合もこれと同様にして，$a_1 < a_2 < \cdots < a_n$ の離散点に測度 $p_1, p_2, \ldots, p_n$ がある多項分布を考えてみると

$$\mathbb{E}[T]$$

$$= \sum_{m_1+\cdots+m_n=n} \frac{n!}{m_1! \cdots m_n!} p_1^{m_1} \cdots p_n^{m_n} T(a_1 \text{ が } m_1 \text{ 個}, \ldots, a_n \text{ が } m_n \text{ 個})$$

$$= 0$$

が任意の $p_i (\sum p_i = 1)$ について成り立たなくてはいけないが，これは $T = 0$ のときのみ可能であることが言える．逆にある x_1 で $T(x_1, \ldots) \neq 0$ であったとすると，$X_1 = x_1$ にだけ測度がある分布を考えれば $\mathbb{E}[T] \neq 0$ となってしまい，$\mathbb{E}[T] = 0$ とはならなくなるので，$T \neq 0$ となるような点があってはいけないことがわかる．

したがって対称な不偏推定量は $\hat{\theta}$ しかないことがわかり，また前の定理から分散が最小であることが保証される． □

このとき推定量である標本平均の分散は X_i の独立性より

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}\sigma^2$$

となる．もし分散 σ^2 が有限であれば，これは $n \to \infty$ のとき分散が 0 に近づく，すなわち $\hat{\theta} \to \theta$ となることを表わしており，観測値の数が増えれば，その推定精度が良くなり，極限では真の値に一致することを示している．これが仮定 4 を設けた一つの理由である．理想的には無限に沢山の観測をしなくてはいけないのであるが，大雑把に言い換えると一人の人が実験して一つの推定値しか得られなかったとしても，観測値の数が十分大きければほぼ正しい値が推定されることを保証していることになる．このように観測数を多くしたとき正しい値に近づいていく性質をもつ推定量を**一致性** (consistency) があると呼ぶ．

2.2 誤差の分布が既知の場合

これまでは誤差の分布に関して平均 0 分散有限というあまり情報のない場合を考えてきたが，ここでは誤差に関してより具体的な情報としてその分布の形がわかっている場合を考える．すなわち

仮定 5 ε_i は既知の密度関数 $f(\varepsilon)$ を持つ分布に従う.

という仮定を加える.

例 2.2. ある農場で毎朝収穫されるメロンの糖度を調べたい. メロンの糖度 X は平均を θ として

$$X = \theta + \varepsilon$$

という確率モデルに従っていることがわかっている. ここで ε は個体のばらつきを反映した個体ごとに独立な確率変数である. その分布は区間 $[-\sigma, \sigma]$ 上の一様分布に従い, 確率密度関数

$$f(\varepsilon) = \begin{cases} \dfrac{1}{2\sigma}, & (-\sigma < \varepsilon < \sigma) \\ 0, & (\text{それ以外}) \end{cases}$$

で記述されることがわかっているものとする.

さて, 全てのメロンを調べることはできないので, 毎朝 3 個をランダムに抜き取り糖度の測定を行う. 測定結果を X_1, X_2, X_3 で表すこととし, これらを大きさの順に並べ替えたものを

$$X_{(1)}, X_{(2)}, X_{(3)}, \quad (X_{(1)} \leq X_{(2)} \leq X_{(3)})$$

と書くことにする. このとき 3 つの数の並べ替えは 6 通りあることに注意する. 糖度の平均値 θ の推定量としては以下の 3 つを考えることにする.

$$\hat{\theta}_1 = \frac{X_1 + X_2 + X_3}{3} \qquad (\text{標本平均})$$

$$\hat{\theta}_2 = X_{(2)} \qquad (\text{中央値})$$

$$\hat{\theta}_3 = \frac{X_{(1)} + X_{(3)}}{2} \qquad (\text{最小値と最大値の平均値})$$

さて, それぞれの推定量の性質を調べてみることにしよう.

まず標本平均であるが, 誤差 ε は区間 $[-\sigma, \sigma]$ 上の一様分布に従うので

$$\mathbb{E}[\varepsilon] = 0$$
$$\mathrm{Var}(\varepsilon) = \int_{-\sigma}^{\sigma} \frac{x^2}{2\sigma} dx = \frac{\sigma^2}{3}$$

となることに注意すると

$$\mathbb{E}[\hat{\theta}_1] = \mathbb{E}\left[\frac{X_1 + X_2 + X_3}{3}\right] = \theta$$
$$\mathrm{Var}(\hat{\theta}_1) = \frac{1}{3}\mathrm{Var}(\varepsilon) = \frac{\sigma^2}{9}$$

である.

次に中央値であるが, $X_{(2)}$ の確率分布を求めるために, δ を微小な値として区間 $(\gamma, \gamma+\delta)$ に $X_{(2)}$ が含まれる確率 $P(\gamma < X_{(2)} < \gamma+\delta)$

を考える．δ が小さければ，上記の区間に $X_{(1)}$ や $X_{(3)}$ が入る確率は無視できるので，中央値 $X_{(2)}$ の確率分布は，

$$P(\gamma < X_{(2)} < \gamma + \delta) = \sum_{i,j,k \text{ の並べ替え}} P(X_i \leq \gamma < X_j < \gamma + \delta \leq X_k), \quad (\theta - \sigma < \gamma < \theta + \sigma)$$

という関係を満たす．このとき

$$P(X_i \leq \gamma) = \int_{\theta-\sigma}^{\gamma} f(x-\theta)dx = \frac{\gamma - \theta + \sigma}{2\sigma}$$
$$P(\gamma < X_j < \gamma + \delta) = \int_{\gamma}^{\gamma+\delta} f(x-\theta)dx = \frac{\delta}{2\sigma}$$
$$P(\gamma + \delta \leq X_k) = \int_{\gamma+\delta}^{\theta+\sigma} f(x-\theta)dx = \frac{\theta + \sigma - \gamma - \delta}{2\sigma}$$

であるから $X_{(2)}$ の密度関数を $g(x)$ とすると

$$\int_{\gamma}^{\gamma+\delta} g(x)dx = 6 \cdot \frac{\gamma - \theta + \sigma}{2\sigma} \cdot \frac{\delta}{2\sigma} \cdot \frac{\theta + \sigma - \gamma - \delta}{2\sigma}$$

であるが，$\delta \to 0$ の極限を考えると

$$g(x) = \frac{3}{4\sigma^3}\left\{\sigma^2 - (x-\theta)^2\right\} \quad (\theta - \sigma < x < \theta + \sigma)$$

となる．密度関数 $g(x)$ は θ を中心として対称であるから，その平均は

$$\begin{aligned}\mathbb{E}[\hat{\theta}_2] &= \mathbb{E}[X_{(2)}] \\ &= \int_{\theta-\sigma}^{\theta+\sigma} x g(x)dx \\ &= \theta + \int_{\theta-\sigma}^{\theta+\sigma} (x-\theta) g(x)dx = \theta\end{aligned}$$

である．また分散は

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_2) &= \mathbb{E}[(X_{(2)} - \theta)^2] \\ &= \int_{\theta-\sigma}^{\theta+\sigma} (x-\theta)^2 g(x)dx \\ &= \frac{3}{4\sigma^3}\int_{-\sigma}^{\sigma} x^2(\sigma^2 - x^2)dx = \frac{\sigma^2}{5}\end{aligned}$$

となる．

最後に最大値と最小値の平均について調べる．まず観測データ X_1, X_2, X_3 が全て区間 (α, β) に入る確率は

$$P(\alpha < X_1, X_2, X_3 < \beta) = \frac{(\beta - \alpha)^3}{8\sigma^3} \quad (\theta - \sigma < \alpha, \beta < \theta + \sigma)$$

であるが，これは最小値 $X_{(1)}(= x)$ と最大値 $X_{(3)}(= y)$ の同時確率密度を $h(x, y)$ とするとき，

$$P(\alpha < X_1, X_2, X_3 < \beta) = \int_{\alpha}^{\beta}\int_{\alpha}^{\beta} h(x,y)dxdy$$

であるが，大小関係から $x > y$ は起こり得ないため $h(x,y) = 0\ (x > y)$ となることに注意すると

$$P(\alpha < X_1, X_2, X_3 < \beta) = \int_\alpha^\beta \int_\alpha^y h(x,y)dxdy \left(= \int_\alpha^\beta dy \int_\alpha^y dx\ h(x,y) \right)$$

と書くことができる．積分の順序に注意して両辺の微分を考えると

$$\frac{\partial}{\partial\alpha}\frac{\partial}{\partial\beta}P(\alpha < X_1, X_2, X_3 < \beta) = -h(\alpha,\beta)$$

となるので，同時確率密度は

$$h(x,y) = \frac{3}{4\sigma^3}(y-x),\ (x \le y)$$

と求められる．これを用いると推定量 $\hat{\theta}_3$ の平均は

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_3] &= \mathbb{E}\left[\frac{X_{(1)} + X_{(3)}}{2}\right] \\
&= \int_{\theta-\sigma}^{\theta+\sigma} \int_{\theta-\sigma}^{y} \frac{x+y}{2} h(x,y)dxdy \\
&= \theta + \int_{-\sigma}^{\sigma} \int_{-\sigma}^{y} \frac{x+y}{2} h(x,y)dxdy = \theta
\end{aligned}$$

であり，また分散は

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_3) &= \mathbb{E}\left[\left(\frac{X_{(1)} + X_{(3)}}{2} - \theta\right)^2\right] \\
&= \int_{\theta-\sigma}^{\theta+\sigma} \int_{\theta-\sigma}^{y} \left(\frac{x+y}{2} - \theta\right)^2 h(x,y)dxdy \\
&= \int_{-\sigma}^{\sigma} \int_{-\sigma}^{y} \left(\frac{x+y}{2}\right)^2 h(x,y)dxdy = \frac{\sigma^2}{10}
\end{aligned}$$

と求められる．

以上の計算結果から，この問題においては推定量 $\hat{\theta}_3$ が推定量の分散が小さいという意味において最も良い推定量と考えられる．

例で見たように，分布の形の情報を用いると θ の推定量として前節で述べた標本平均より良い推定量を作ることができる場合がある．以下ではより一般の場合を想定して不偏性以外に**位置共変性**という以下の性質を持つ推定量の中でより良いものを考えることにする．

定義 2.3. 任意の c に対して推定量が

$$\hat{\theta}(X_1 + c, X_2 + c, \ldots, X_n + c) = \hat{\theta}(X_1, X_2, \ldots, X_n) + c$$

を満たすとき，これを**位置共変推定量**という．

定義の式は観測値 $X_1, X_2, \ldots, X_n$ が全部同時に c だけずれて $X_1 + c, X_2 + c, \ldots, X_n + c$ となったとき，推定量もちょうど c だけずれることを表しており，今考えている状況においては非常に自然な推定量の性質であると考えられる．

この性質を持つ推定量の族の中では以下のようにして標本平均より良い推定量が作ることができる.

まず n 個の観測値 $X_1, X_2, \ldots, X_n$ から標本平均 $\bar{X}$ だけを推定に用いるということは，観測値の持っている情報の一部だけを使っていることに注意する．このとき捨ててしまっている情報はどのようなものか考えてみるために，例として以下のような 2 つの変数変換を考える.

変換.例 *1*

$$\begin{aligned}\bar{X} &= \frac{X_1 + X_2 + \cdots + X_n}{n}\\ Y_1 &= X_2 - X_1\\ Y_2 &= X_3 - X_1\\ &\vdots\\ Y_{n-1} &= X_n - X_1\end{aligned}$$

変換.例 *2*

$$\begin{aligned}\bar{X} &= \frac{X_1 + X_2 + \cdots + X_n}{n}\\ Y_1' &= \frac{X_1 - X_2}{\sqrt{2}}\\ Y_2' &= \frac{X_1 + X_2 - 2X_3}{\sqrt{6}}\\ &\vdots\\ Y_{n-1}' &= \frac{X_1 + X_2 + \cdots + X_{n-1} - (n-1)X_n}{\sqrt{n(n-1)}}\end{aligned}$$

どちらの変換においても, 当然 $X_1, X_2, \ldots, X_n$ を知れば, $\bar{X}, Y_1, \ldots, Y_{n-1}$ あるいは $\bar{X}, Y_1', \ldots, Y_{n-1}'$ を計算することができるが, 逆に $\bar{X}, Y_1, \ldots, Y_{n-1}$ あるいは $\bar{X}, Y_1', \ldots, Y_{n-1}'$ を知ればもとの観測値 $X_1, X_2, \ldots, X_n$ を復元することができる. つまり $\{X_1, X_2, \ldots, X_n\}, \{\bar{X}, Y_1, \ldots, Y_{n-1}\}, \{\bar{X}, Y_1', \ldots, Y_{n-1}'\}$ というそれぞれ n 個の変数の組は同じ情報を持っていると言え，こうした変換はいくらでも考えることができる. 前節の標本平均 $\bar{X}$ しか用いていない推定量は $Y_1, Y_2, \ldots, Y_{n-1}$ あるいは $Y_1', Y_2', \ldots, Y_{n-1}'$ といった情報を捨てていると考えることができる. 前節での結論は，推定量の分散を小さくするという意味で最も良い推定量は標本平均であるということで，言い換えれば分布に関する知識がない場合には結局 $Y_1, Y_2, \ldots, Y_{n-1}$ といった情報は捨てるしかないということになるが，分布の形がわかっている場合には以下のようにして標本平均を修正し，より良い推定量を作ることができる.

なお例 2 の変換は

$$\begin{aligned}&\mathbb{E}[Y_i'] = 0\\ &\mathbb{E}[Y_i'^2] = \sigma^2\\ &\mathrm{Cov}(\bar{X}, Y_i) = 0, \quad \mathrm{Cov}(Y_i, Y_j) = 0\end{aligned}$$

という特徴を持っているため統計ではしばしば用いられるが，以下では計算上の簡単のため例 1 の変換を用いて議論を進める.

定理 2.4. 真の母数が $\theta = 0$ のときに $Y_1, Y_2, \ldots, Y_{n-1}$ を与えられたもとでの $\bar{X}$ の条件付期待値を

$$\mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]$$

と表すことにする．

$$\hat{\theta} = \bar{X} - \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]$$

とおくと $\hat{\theta}$ は位置共変不偏推定量であって，任意の位置共変推定量 $\hat{\theta}$ に対して

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - \theta)^2]$$

となる．

略証．はじめに $\hat{\theta}$ が位置共変不偏推定量であることを確かめる．
　まず $\mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]$ は $\theta = 0$ のもとで観測される $\bar{X}$ の条件付平均なので，$\bar{X}$ にはよらない $Y_1, Y_2, \ldots, Y_{n-1}$ だけの関数

$$\mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}] = \beta(Y_1, Y_2, \ldots, Y_{n-1})$$

であることに注意する．また観測値 $X_1, X_2, \ldots, X_n$ 全体が c だけ変化し $X_1 + c, X_2 + c, \ldots, X_n + c$ になったとしても，定義から明らかなように $Y_1, Y_2, \ldots, Y_{n-1}$ は変化しない．したがって観測値全体を c だけずらしたとき

$$\hat{\theta}(X_1 + c, X_2 + c, \ldots, X_n + c)$$

$$= \frac{(X_1 + c) + (X_2 + c) + \cdots + (X_n + c)}{n} - \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]$$

$$= \bar{X} + c - \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]$$

$$= \hat{\theta}(X_1, X_2, \ldots, X_n) + c$$

となり，$\hat{\theta}$ は位置共変であることがわかる．また

$$\mathbb{E}^{Y_1, Y_2, \ldots, Y_{n-1}}\left[\mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]\right] = \mathbb{E}_0[\bar{X}] = 0$$

であるから，

$$\begin{aligned}
&\mathbb{E}[\hat{\theta}] \\
&= \mathbb{E}^{X_1, X_2, \ldots, X_n}[\hat{\theta}(X_1, X_2, \ldots, X_n)] \\
&= \mathbb{E}^{X_1, X_2, \ldots, X_n}[\bar{X}] - \mathbb{E}^{X_1, X_2, \ldots, X_n}[\mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]] \\
&= \theta - \mathbb{E}^{\bar{X}}[\mathbb{E}[Y_1, Y_2, \ldots, Y_{n-1}]\mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]] \\
&= \theta - \mathbb{E}^{\bar{X}}[0] \\
&= \theta
\end{aligned}$$

となり不偏である．ここでは変数変換しても全体に対する平均は変わらないこと

$$\mathbb{E}^{X_1,X_2,\ldots,X_n}[\cdot] = \mathbb{E}^{\bar{X}}[\mathbb{E}^{Y_1,Y_2,\ldots,Y_{n-1}}[\cdot]]$$

を用いている．

さて任意の位置共変推定量 $\hat{\theta}$ に対して

$$T = \hat{\theta} - \hat{\theta}$$

とおくと，二つの推定量の位置共変性から

$$T(X_1 + c, X_2 + c, \ldots, X_n + c)$$

$$= \hat{\theta}(X_1 + c, X_2 + c, \ldots, X_n + c) - \hat{\theta}(X_1 + c, X_2 + c, \ldots, X_n + c)$$

$$= \Big\{\hat{\theta}(X_1, X_2, \ldots, X_n) + c\Big\} - \Big\{\hat{\theta}(X_1, X_2, \ldots, X_n) + c\Big\}$$

$$= \hat{\theta}(X_1, X_2, \ldots, X_n) - \hat{\theta}(X_1, X_2, \ldots, X_n)$$

$$= T(X_1, X_2, \ldots, X_n)$$

となる．ここで T を $\bar{X}, Y_1, Y_2, \ldots, Y_{n-1}$ で書き換えると，$\bar{X}$ だけが c ずらす影響を受けることに注意すれば，T は $Y_1, Y_2, \ldots, Y_{n-1}$ のみの関数となることがわかる．すなわち

$$T = T(Y_1, Y_2, \ldots, Y_{n-1})$$

と書かれる．また位置共変推定量の性質から任意の位置共変推定量について

$$\hat{\theta}(X_1, X_2, \ldots, X_n) - \theta = \hat{\theta}(X_1 - \theta, X_2 - \theta, \ldots, X_n - \theta)$$

となるが，

$$X_1' = X_1 - \theta, \quad X_2' = X_2 - \theta, \quad \ldots, \quad X_n' = X_n - \theta$$

と置き換えると，$X_1', X_2', \ldots, X_n'$ が $\theta = 0$ のもとでの観測値に対応することになるので

$$\begin{aligned}\mathbb{E}[(\hat{\theta} - \theta)^2 \mid \theta] &= \mathbb{E}^{X_1', X_2', \ldots, X_n'}[\hat{\theta}(X_1', X_2', \ldots, X_n')^2] \\ &= \mathbb{E}_0[\hat{\theta}^2]\end{aligned}$$

が成り立ち，位置共変推定量の 2 乗誤差は θ によらず一定，つまり真の母数 θ に無関係であることがわかる．したがって $\hat{\theta} = \hat{\theta} + T$ とすれば任意の位置共変推定量 $\hat{\theta}$ の 2 乗誤差は $\theta = 0$ の場合

$$\mathbb{E}_0[\hat{\theta}^2] = \mathbb{E}_0[\hat{\theta}^2] + \mathbb{E}_0[T^2] + 2\mathbb{E}_0[\hat{\theta}T]$$

を評価すればよいことがわかる．条件付確率の性質から

$$\begin{aligned}
&\mathbb{E}_0[\hat{\theta} \mid Y_1, Y_2, \ldots, Y_{n-1}] \\
&= \mathbb{E}_0[\bar{X} - \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}] \mid Y_1, Y_2, \ldots, Y_{n-1}] \\
&= \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}] - \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}] \\
&= 0
\end{aligned}$$

であり，$T = T(Y_1, Y_2, \ldots, Y_{n-1})$ に注意すると

$$\begin{aligned}
\mathbb{E}_0[\hat{\theta}T] &= \mathbb{E}^{Y_1, Y_2, \ldots, Y_{n-1}}[\mathbb{E}_0[\hat{\theta}T \mid Y_1, Y_2, \ldots, Y_{n-1}]] \\
&= \mathbb{E}^{Y_1, Y_2, \ldots, Y_{n-1}}[T\mathbb{E}_0[\hat{\theta} \mid Y_1, Y_2, \ldots, Y_{n-1}]] \\
&= 0
\end{aligned}$$

したがって

$$\begin{aligned}
\mathbb{E}[(\hat{\theta} - \theta)^2] &= \mathbb{E}[(\hat{\theta} - \theta)^2] + \mathbb{E}_0[T^2] \\
&\geq \mathbb{E}[(\hat{\theta} - \theta)^2]
\end{aligned}$$

となり $\hat{\theta}$ の 2 乗誤差，この場合 $\hat{\theta}$ は不偏なので分散の最小性がわかる． □

前節のように分布の形がわからない場合は補正のしようがなかったわけであるが，上の結果は捨ててしまった情報で補正することによってより良い推定量が作られることを主張している．特にこの場合は分散の評価において位置共変性が重要な役割を果たしていることに注意して欲しい．

さて，この推定量の具体形を求めると次のようになる．

定理 2.5. 誤差の確率密度関数が f のもとで最小分散位置共変不偏推定量は

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \frac{\displaystyle\int \theta \prod_{i=1}^{n} f(X_i - \theta) d\theta}{\displaystyle\int \prod_{i=1}^{n} f(X_i - \theta) d\theta}$$

で与えられる．

略証．まず

$$n(\bar{X} - X_1) = (X_2 - X_1) + (X_3 - X_1) + \cdots + (X_n - X_1)$$

であるので $X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1$ で条件付けたとき $\bar{X} - X_1$ は一定の値になることに注意すると

$$\begin{aligned}
&\mathbb{E}_0[X_1 \mid Y_1, Y_2, \ldots, Y_{n-1}] \\
&= \mathbb{E}_0[X_1 \mid X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1] \\
&= \mathbb{E}_0[(X_1 - \bar{X}) + \bar{X} \mid X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1] \\
&= X_1 - \bar{X} + \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}]
\end{aligned}$$

であることがわかる．したがって

$$\begin{aligned}\hat{\theta} &= \bar{X} - \mathbb{E}_0[\bar{X} \mid Y_1, Y_2, \ldots, Y_{n-1}] \\ &= X_1 - \mathbb{E}_0[X_1 \mid Y_1, Y_2, \ldots, Y_{n-1}]\end{aligned}$$

と書き換えられる．

$\theta = 0$ のもとでの $X_1, X_2, \ldots, X_n$ の同時分布の密度関数は，その独立性から

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

で与えられるが，$X_1, Y_1, Y_2, \ldots, Y_{n-1}$ で書き換えると

$$\int_A f(x_1, x_2, \ldots, x_n)dx_1dx_2\ldots dx_n$$

$$= \int_A f(x_1, y_1, y_2\ldots, y_{n-1}) \left|\frac{d(x_1, x_2, \ldots, x_n)}{d(x_1, y_1, \ldots, y_{n-1})}\right| dx_1dy_1dy_2\ldots dy_{n-1}$$

$$= \int_A f(x_1)f(y_1 + x_1)f(y_2 + x_1)\ldots f(y_{n-1} + x_1)dx_1dy_1dy_2\ldots dy_{n-1}$$

となる．ただし変換の Jacobian が

$$\begin{aligned}&\left|\frac{d(x_1, x_2, \ldots, x_n)}{d(x_1, y_1, \ldots, y_{n-1})}\right| \\ &= \begin{vmatrix} \frac{dx_1}{dx_1} & \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_{n-1}} \\ \frac{dx_2}{dx_1} & \frac{dx_2}{dy_1} & \cdots & \frac{dx_2}{dy_{n-1}} \\ \vdots & & & \vdots \\ \frac{dx_n}{dx_1} & \frac{dx_2}{dy_1} & \cdots & \frac{dx_n}{dy_{n-1}} \end{vmatrix} \\ &= \begin{vmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \ldots & 1 \end{vmatrix} = 1\end{aligned}$$

となることを用いた．したがって変換後の確率密度は

$$f(x_1, y_1, y_2\ldots, y_{n-1}) = f(x_1)f(y_1+x_1)f(y_2+x_1)\ldots f(y_{n-1}+x_1)$$

となる．

$Y_1, Y_2, \ldots, Y_{n-1}$ が与えられたもとでの X_1 の条件付確率密度 $f(x_1|y_1, y_2\ldots, y_{n-1})$ は

$$f(x_1, y_1, y_2\ldots, y_{n-1}) = f(x_1|y_1, y_2\ldots, y_{n-1}) \cdot f(y_1, y_2\ldots, y_{n-1})$$

より

$$
\begin{aligned}
&f(x_1|y_1, y_2 \ldots, y_{n-1}) \\
&= \frac{f(x_1, y_1, y_2 \ldots, y_{n-1})}{f(y_1, y_2 \ldots, y_{n-1})} \\
&= \frac{f(x_1, y_1, y_2 \ldots, y_{n-1})}{\displaystyle\int f(x_1, y_1, y_2 \ldots, y_{n-1}) dx_1} \\
&= \frac{f(x_1) f(y_1 + x_1) f(y_2 + x_1) \ldots f(y_{n-1} + x_1)}{\displaystyle\int f(x_1) f(y_1 + x_1) f(y_2 + x_1) \ldots f(y_{n-1} + x_1) dx_1}
\end{aligned}
$$

で与えられる．これより推定量は

$$
\hat{\theta} = X_1 - \frac{\displaystyle\int x_1 f(x_1) f(Y_1 + x_1) f(Y_2 + x_1) \ldots f(Y_{n-1} + x_1) dx_1}{\displaystyle\int f(x_1) f(Y_1 + x_1) f(Y_2 + x_1) \ldots f(Y_{n-1} + x_1) dx_1}
$$

となるが，$\theta = X_1 - x_1$ と変数変換すると

$$
= \frac{\displaystyle\int \theta f(X_1 - \theta) f(X_2 - \theta) \ldots f(X_n - \theta) d\theta}{\displaystyle\int f(X_1 - \theta) f(X_2 - \theta) \ldots f(X_n - \theta) d\theta}
$$

となる． □

この推定量は *Pitman* **推定量**と呼ばれることがある．

例 2.6. 確率モデル

$$X = \theta + \varepsilon$$

において，誤差の分布が $[-0.5, 0.5]$ 上の一様分布，すなわち

$$
f(\varepsilon) = \begin{cases} 1 & |\varepsilon| \leq \frac{1}{2} \\ 0 & \text{それ以外} \end{cases}
$$

のとき，独立な観測値 $X_1, X_2, \ldots, X_n$ に対して Pitman 推定量 $\hat{\theta}$ を求める．

まず X_i の密度関数は母数 θ と ε の密度関数 f を用いて

$$f(x_i - \theta)$$

と表される．この関数はその値として 0 か 1 しかとらないので，n 個の観測値の同時密度関数は全てが 1 のときだけ 1 となり，それ以外は 0 となる．これを θ の関数としてみたとき図 2.1 のようになるが，1 となるのは区間 $[\max_i X_i - 0.5, \min_i X_i + 0.5] = [\alpha, \beta]$

である．したがって Pitman 推定量は

$$\begin{aligned}\hat{\theta} &= \frac{\int_\alpha^\beta \theta d\theta}{\int_\alpha^\beta d\theta} \\ &= \frac{1}{2}\frac{\beta^2-\alpha^2}{\beta-\alpha} \\ &= \frac{1}{2}(\beta+\alpha) \\ &= \frac{1}{2}\left(\max_i X_i + \min_i X_i\right)\end{aligned}$$

となる．

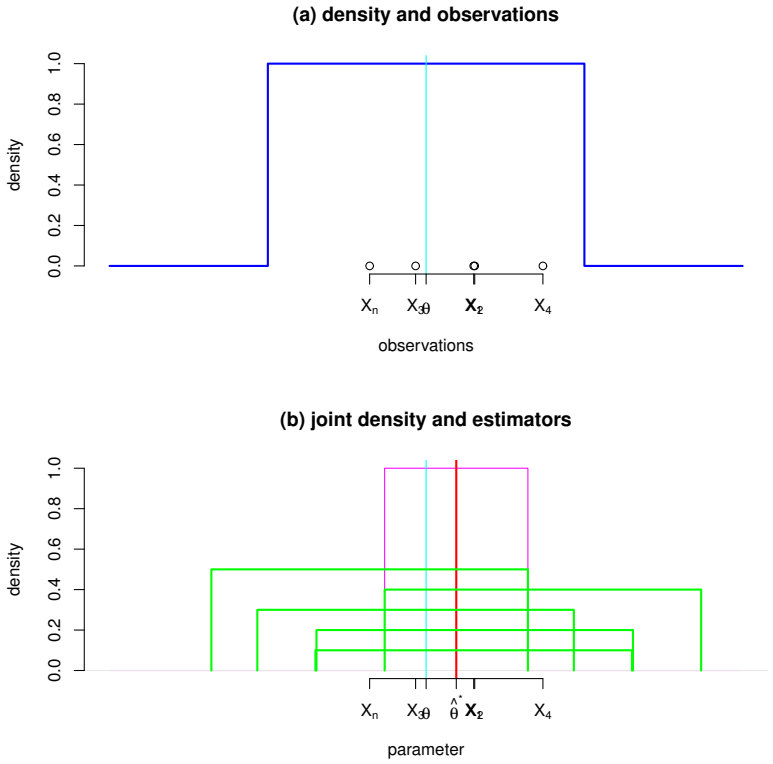

図 2.1: 一様乱数の場合の Pitman 推定量．(a) 一様分布の密度関数と観測データ．(b) 各データに対する可能なパラメタ領域とその交集合．

次に推定量 $\hat{\theta}*$ の平均と分散を求める．このためにはまず n 個の観測値の最小値と最大値の同時分布を求めなくてはいけない．$X_1, X_2, \ldots, X_n$ の最小値，最大値をそれぞれ Y,Z とすると，Y が区間 $[y, y+dy]$ に入り，Z が区間 $[z, z+dz]$ に入る確率は，n 個の一様乱数の内一つが $[y, y+dy]$ に，もう一つが $[z, z+dz]$ に入り，残りの $n-2$ 個が $[y, z]$ に入る確率を考えれば良いので，n 個のどれが最大，最小になるかの場合の数が $n(n-1)$ 通りで，それぞれの区間に入る確率は区間の幅 $\times$ 密度 1 であることから，その

密度は

$$n(n-1)(z-y)^{n-2}dydz$$

と表される. このとき $a \leq Y \leq Z \leq b$ となる確率は $X_1, X_2, \ldots, X_n$ が全て a,b の間に入っている確率に等しいが，確かに

$$\begin{aligned} P(a \leq Y \leq Z \leq b) &= \int_a^b \int_a^z n(n-1)(z-y)^{n-2}dydz \\ &= (b-a)^n \end{aligned}$$

となっていることが確認できる.

注意 2.7. 上のような導き方をせずに，Y と Z の同時確率密度を $f(y,z)$ として

$$\begin{aligned} P(a \leq Y \leq Z \leq b) &= \int_a^b \int_a^z f(y,z)dydz \\ &= (b-a)^n \end{aligned}$$

なることから，これを微分して同時確率密度を求めてもよい.

推定量は

$$\hat{\theta} = \frac{Y+Z}{2}$$

であるが，平均と分散の性質を調べるためには $\theta = 0$ の場合だけ考えれば十分なので，以下では $\theta = 0$ について計算する．この場合 Y,Z は $[-0.5, 0.5]$ で考えることになる．まず平均は

$$\begin{aligned} &\mathbb{E}[\hat{\theta}] \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{z} \frac{y+z}{2} n(n-1)(z-y)^{n-2}dydz \\ &= 0 \end{aligned}$$

したがって不偏であることがわかる．次に分散は

$$\begin{aligned} &\mathrm{Var}(\hat{\theta}) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{z} \left(\frac{y+z}{2}\right)^2 n(n-1)(z-y)^{n-2}dydz \\ &= \frac{1}{2(n+2)(n+1)} \end{aligned}$$

である.

ところで推定量として標本平均を用いた場合，その分散は

$$\mathrm{Var}(\bar{X}) = \frac{1}{n}\sigma^2$$

であったから，$n \geq 3$ のとき確かに Pitman 推定量の方が分散が小さくなっている.

分布が既知の場合，位置共変推定量という限られた推定量の中では Pitman 推定量が一様最小分散不偏推定，すなわちどんな母数に対しても任意の位置共変推定量と比べたとき等しいか，あるいはより小さな分散を持つ推定量となる．しかし位置共変性を持たない推定量にまで範囲を広げた場合にはこれが常に最小分散になるとは限らない．場合によっては真の母数が特定の値のときだけに良い推定値を与えるような推定量を構成することもできる．しかし一般には推定を行う前に母数の含まれる領域を知っているわけではないので，どのような母数に対しても良い推定値を与える推定方式の方が良いと考えるのが自然であろう．推定量の良さをその分散から考える立場では，図の一番下の推定量のように，任意の母数においてどんな不偏推定量よりも小さな分散を持つものが最も望ましい推定量であり，「母数について一様」という意味で一様最小分散不偏推定量と呼ぶ．その具体的な形はもちろん誤差の分布に依存するが，一般の確率モデルに対して一様最小分散不偏推定量が存在するとは限らない．

3 最尤推定量

3.1 尤度と最尤法

これまでは観測値が

$$X_i = \theta + \varepsilon_i, \quad i = 1, \ldots, n$$

という簡単な確率モデルに従う場合を考えてきた．この簡単なモデルに対しても，一般には不偏推定量の中に一様最小分散推定量 (どんな状況でも推定量の分散を比較すると他の推定量と同じかそれより小さくなる推定量) があるとは限らない．しかしながら，不偏推定量に更に条件を加えると具体的に最小分散推定量を構成することができる場合がある．例えば位置共変性を課すと Pitman 推定量が不偏推定量の中で最良となる．実は位置共変性を課さなくても Pitman 推定量は良い (推定の結果が安定している) 推定を与えることは多いが，観測データが多くなると被積分関数が多数の関数の積となり計算が繁雑となり，実際の問題に適用するには問題となることもある．

またより一般の確率モデルとして，観測値 X が平均以外のいくつかの母数を持つ確率密度関数で記述される確率法則に従っていると考え，それらの母数の推定を行いたい場合も多い．例えば観測値 X の起きやすい範囲が 2 つあり，分布の形状が双峰性を持つような場合には，2 つの異なる平均値 θ_1, θ_2 をもつ正規分布を重ね合わせた密度

$$f(x, \theta_1, \theta_2) = \frac{1}{2\sqrt{2\pi}\sigma} e^{\frac{-(x-\theta_1)^2}{2\sigma^2}} + \frac{1}{2\sqrt{2\pi}\sigma} e^{\frac{-(x-\theta_2)^2}{2\sigma^2}}$$

によって分布をモデル化し，母数 θ_1, θ_2 を推定するといった方法がしばしば用いられる．この例では等分散で山の高さも同じであるが，このモデルを多峰 (山が多数) で不均一な混合 (山の高さが

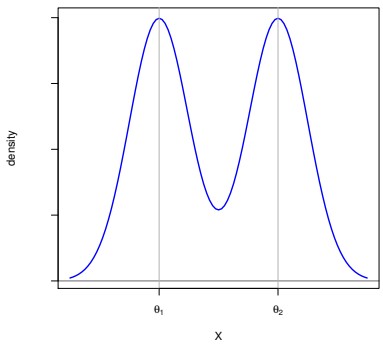

図 3.1: 多峰性のある分布の例．

異なる) に一般化したものは混合正規分布と呼ばれ，データの解析に良く用いられている．

ここでは，一般の確率モデルにおける母数推定の問題を取り上げ，その推定量を与える考え方の一つである最尤法について説明する．後ほど述べるように最尤推定量は一般に不偏ではなく，また最小分散ではない．ここまで述べた基準では良い推定量とはいえないが，観測値の数が大きくなるに従いほぼ不偏で最小分散な推定量になっていく性質があるため，実用上はかなりよい方法となっていることを示す．

3.1.1 尤度関数

確率変数 X が確率密度関数 $f(x,\theta)$ によって表される確率法則に従っているとする．この分布に従う n 個の独立な観測値が得られたとすると，その $X_1,\ldots,X_n$ の同時密度関数は独立性の仮定より

$$f(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}f(x_i,\theta)$$

と書くことができる．直感的には，母数 θ の確率モデルから

$$x_1\le X_1\le x_1+\Delta_1,\ x_2\le X_2\le x_2+\Delta_2,\ \ldots,\ x_n\le X_n\le x_n+\Delta_n$$

の範囲の観測データが得られる確率は同時密度関数を用いて

$$\begin{aligned}&P\left((X_1,X_2,\ldots,X_n)\in[x_1,x_1+\Delta_1]\times[x_2,x_2+\Delta_2]\times\cdots\times[x_n,x_n+\Delta_n]\right)\\&\simeq f(x_1,x_2,\ldots,x_n)\Delta_1\Delta_2\cdots\Delta_n\\&=\prod_{i=1}^{n}f(x_i,\theta)\Delta_i\end{aligned}$$

で表される．つまり，同時密度関数の値は $x_1,x_2,\ldots,x_n$ という観測データの集合が得られる確率に比例する．

一方，同時密度関数を母数 θ の関数として考えてみると，母数 θ を変化させて観測値 $x_1,x_2,\ldots,x_n$ に対する同時密度関数の値を計算することによって，観測値 $x_1,x_2,\ldots,x_n$ を生成した確率モデルとしてどの θ が尤もらしいかを知ることができる．

このように観測値 $x_1,x_2,\ldots,x_n$ に対する同時確率密度

$$f(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}f(x_i,\theta)=L(\theta)$$

を θ の関数 $L(\theta)$ として考えたとき，これを「母数の尤もらしさを測る関数」という意味で**尤度関数** (likelihood function) と呼ぶ．また，母数 θ での尤度関数の値を，観測値 $x_1,x_2,\ldots,x_n$ における母数 θ の**尤度** (likelihood) と呼ぶ．

3.1.2 最尤推定量

最尤推定量は尤度関数を最大にする母数の値として以下のように定義される．

定義 3.1 (最尤推定量). 母数 θ の定義域を Θ とする．尤度関数 $L(\theta)$ を最大にする θ の値 $\hat{\theta}^*$

$$L(\hat{\theta}^*) = \max_{\theta\in\Theta} L(\theta) = \max_{\theta\in\Theta} \prod_{i=1}^{n} f(X_i, \theta)$$

(あるいは以下のように書く場合もある)

$$\hat{\theta}^* = \arg\max_{\theta\in\Theta} L(\theta)$$

を**最尤推定量** (maximum likelihood estimator) という．

最尤推定量の考え方をみるために，再び簡単なモデル

$$X_i = \theta + \varepsilon_i, \quad i = 1, \ldots, n$$

を考えてみることにする (図 3.2 参照)．尤度関数 $L(\theta)$ の最大値を求めるということは密度関数 $f(x-\theta)$ を θ の関数と考え，θ をいろいろ動かして観測値 $x_1, x_2, \ldots, x_n$ に対して一番“良く合う”ところを探すことに対応する (図 3.2(b),(c))．この場合「良く合う」とは独立な確率変数の同時密度関数の定義に従って「密度の高さの積が一番大きくなるところ」と定義される．

3.2 最尤推定量の性質

以下の議論では計算の簡単のため θ は 1 次元の母数として扱うが，多次元の場合も少々繁雑ではあるが同様に計算できる．

3.2.1 一致性

まず最尤推定量の良さを見るために，尤度の対数で定義される

$$\frac{1}{n}\log L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta)$$

という量の性質を考えてみる．$\log f(X, \theta)$ は確率変数 X の関数であるのでそれ自身も確率変数であり，また観測値 $X_1, X_2, \ldots, X_n$ が独立であることに注意すると，$\log L(\theta)$ は独立な確率変数の和であるから，n が大きくなっていくと大数の法則により

$$\frac{1}{n}\log L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta) \to \mathbb{E}_{\theta_0}[\log f(X, \theta)] \quad (n \to \infty)$$

に近づく．ただし，θ_0 は観測値 $X_1, X_2, \ldots, X_n$ の従う分布の真の母数とし，$\mathbb{E}_{\theta_0}[\cdot]$ は θ_0 を母数とする分布のもとでの平均を表すものとする．すなわち

$$\mathbb{E}_{\theta_0}[\cdot] = \int \cdot f(x, \theta_0) dx$$

である．このとき

$$\frac{1}{n}\log L(\theta) - \frac{1}{n}\log L(\theta_0) \to \mathbb{E}_{\theta_0}[\log f(X, \theta)] - \mathbb{E}_{\theta_0}[\log f(X, \theta_0)] \quad (n \to \infty)$$

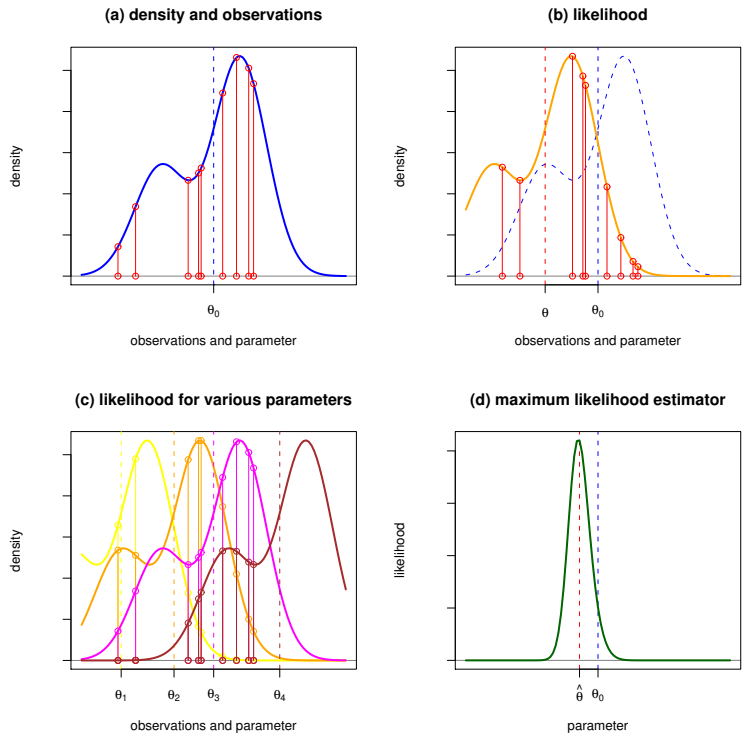

図 3.2: 最尤推定の考え方: (a) X の分布 (密度関数) と観測値．このモデルでは観測値と母数は同じ軸上に描くことができるので，全てのグラフの横軸は同じで，真の母数は θ_0 で表している．(b) ある母数における観測値の密度の積が尤度となる．(c) 尤度をいろいろな母数の値において計算し，観測値と一番良く合うところ (生成確率の高くなるところ) を探す．(d) 計算された尤度関数において尤度の最大値を与える母数 $\hat{\theta}$ を推定値とする．

となるが，右辺は対数の性質により

$$= \mathbb{E}_{\theta_0}\left[\log \frac{f(X,\theta)}{f(X,\theta_0)}\right]$$

となる．また対数関数 $y = \log x$ の $x = 1$ における接線を考える (図 3.3 参照) と $\log x \leq x - 1$ が成り立つことが容易にわかるので

$$\begin{aligned}
&\leq \mathbb{E}_{\theta_0}\left[\frac{f(X,\theta)}{f(X,\theta_0)} - 1\right] \\
&= \int \left(\frac{f(x,\theta)}{f(x,\theta_0)} - 1\right) f(x,\theta_0)dx \\
&= \int f(x,\theta)dx - \int f(x,\theta_0)dx \\
&= 1 - 1 = 0
\end{aligned}$$

となる．最後の積分は確率密度関数の積分が 1 であることを用いている．したがって n が十分大きければ

$$\frac{1}{n}\log L(\theta) \leq \frac{1}{n}\log L(\theta_0)$$

が成り立つことがわかる．このとき等号は x によらず常に

$$\frac{f(x,\theta)}{f(x,\theta_0)} = 1$$

のとき，すなわち

$$f(x,\theta) = f(x,\theta_0)$$

のときにしか成り立たないので，n が大きいと θ_0 と異なる任意の θ について

$$\frac{1}{n}\log L(\theta) < \frac{1}{n}\log L(\theta_0)$$

がほぼ確実に成り立つことがわかる．

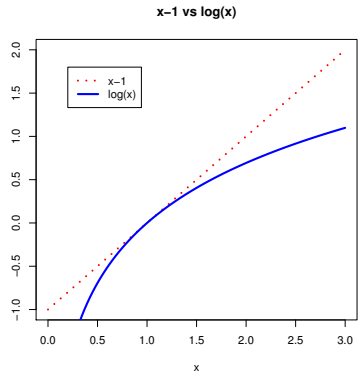

図 3.3: $\log(x)$ と $x - 1$ の関係

対数関数 log は単調増加関数なので，対数をとった尤度関数 $\log L(\theta_0)$ と $\log L(\theta)$ の大小関係はそのまま尤度関数 $L(\theta_0)$ と $L(\theta)$ の大小関係と一致することに注意すると，この不等式は n が大き

いときに真の値 θ_0 から離れた θ が尤度関数 $L(\theta)$ を最大とする確率は小さく，真の値から離れた θ が推定値として選ばれなくなることを意味している．逆に言えば n が大きくなると最尤推定量 $\hat{\theta}^*$ は真の母数の値 θ_0 に近づいていく

$$\hat{\theta}^* \to \theta_0, \quad (n \to \infty)$$

ことがわかる．n が大きくなったときに成り立つ推定量のこのような性質を，一般に**一致性** (consistency) といい，最尤推定量は一致性をもつことがわかる．正確に表現すると以下のようになる．

定理 3.2 (最尤推定量の一致性)**.** 全ての x に対して $f(x) > 0$ で f が連続ならば，最尤推定量 $\hat{\theta}^*$ は**一致推定量** *(consistent estimator)* になる．すなわち，任意の $\varepsilon > 0$ に対して

$$P\left(|\hat{\theta}^* - \theta_0| < \varepsilon \,\middle|\, \theta = \theta_0\right) \to 1 \quad (n \to \infty)$$

が成り立つ．ただし $P(\cdot|\theta = \theta_0)$ は真の母数の値が θ_0 であるという条件のもとでの確率を表すものとする．

3.2.2 漸近正規性

さてもう少し細かく最尤推定量の性質を考えてみることにする．f が連続で2階微分可能とすると，$L(\theta)$ は滑らかな関数で $\hat{\theta}^*$ で最大

$$L(\hat{\theta}^*) = \max_{\theta \in \Theta} L(\theta)$$

となることから

$$\frac{\partial}{\partial \theta} L(\hat{\theta}^*) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i, \hat{\theta}^*) = 0$$

となる．ただし

$$\frac{\partial}{\partial \theta} L(\hat{\theta}^*) = \frac{\partial}{\partial \theta} L(\theta)\Big|_{\theta=\hat{\theta}^*}$$
$$\frac{\partial}{\partial \theta} \log f(X_i, \hat{\theta}^*) = \frac{\partial}{\partial \theta} \log f(X_i, \theta)\Big|_{\theta=\hat{\theta}^*}$$

のことである．この式を θ_0 のまわりで Taylor 展開すると

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i, \theta_0) + (\hat{\theta}^* - \theta_0) \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i, \tilde{\theta}) = 0$$

となる．ただし $\tilde{\theta}$ は $\{X_i\}$ に依存して決まる θ_0 と $\hat{\theta}^*$ との間の値である．これから

$$\sqrt{n}(\hat{\theta}^* - \theta_0) \left\{ -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i, \tilde{\theta}) \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i, \theta_0)$$

と書き換えることができる．

さて，前に述べた最尤推定量の一致性から，n が大きくなると $\hat{\theta}^* \to \theta_0$ となるので，その間にある $\tilde{\theta}$ に関しても $\tilde{\theta} \to \theta_0$ がいえる．したがって大数の法則により

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i,\tilde{\theta}) \to \mathbb{E}_{\theta_0}\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_i,\theta_0)\right] = I(\theta_0)$$

に近づくことがわかる．ここで I は**Fisher 情報量** (Fisher information) と呼ばれる量であり，

$$\begin{aligned} I(\theta_0) &= \mathbb{E}_{\theta_0}\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_i,\theta_0)\right] \\ &= \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f(X_i,\theta_0)\right)^2\right] \end{aligned}$$

と 2 通りの書き方ができる．これについては次節で詳しく説明する．

一方，

$$Y_i = \frac{\partial}{\partial\theta}\log f(X_i,\theta_0)$$

として，新しい確率変数

$$Z = \frac{Y_1+Y_2+\cdots+Y_n}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$$

を考えると

$$\mathbb{E}_{\theta_0}[Y_i] = \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_i,\theta_0)\right] = 0$$

$$\mathrm{Var}_{\theta_0}(Y_i) = \mathrm{Var}_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(X_i,\theta_0)\right) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f(X_i,\theta_0)\right)^2\right] = I(\theta_0)$$

であるから，$n \to \infty$ のとき中心極限定理により Z の分布は平均 0，分散 $I(\theta_0)$ の正規分布 $\mathcal{N}(0, I(\theta_0))$ に近づく，すなわち

$$Z \sim \mathcal{N}(0, I(\theta_0)) \quad (n \to \infty)$$

となる ($\sim$ は左辺の確率変数が右辺の分布に従うことを表す)．

以上の結果を踏まえると，$n \to \infty$ のとき

$$\sqrt{n}I(\theta_0)(\hat{\theta}^* - \theta_0) = Z \sim \mathcal{N}(0, I(\theta_0)) \quad (n \to \infty)$$

であるので

$$\sqrt{n}(\hat{\theta}^* - \theta_0) = \frac{Z}{I(\theta_0)} \sim \mathcal{N}\left(0, \frac{I(\theta_0)}{I(\theta_0)^2}\right) = \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$$

となる．n が大きいときに成り立つことを**漸近的**に成り立つと言い，上の関係は最尤推定量の**漸近正規性** (asymptotic normality) と呼ばれる性質である．纏めると以下のようになる．

定理 3.3 (最尤推定量の漸近正規性)**.** $f(x) > 0$ が連続で 2 階微分可能ならば $\sqrt{n}(\hat{\theta}^* - \theta_0)$ は $n \to \infty$ で正規分布 $\mathcal{N}(0, I(\theta_0)^{-1})$ に近づく．

より直観的には最尤推定量は

$$\mathbb{E}_{\theta_0}[\hat{\theta}^*] = \theta_0 + o\left(\frac{1}{\sqrt{n}}\right)$$
$$\mathrm{Var}_{\theta_0}(\hat{\theta}^*) = \frac{1}{nI(\theta_0)} + o\left(\frac{1}{n}\right)$$

という性質を持つと考えられる．すなわち n が大きくなるとほぼ不偏性が成り立ち，またほぼ $1/nI(\theta_0)$ という分散を持つと考えられる．n が大きくなるにしたがって成り立つ不偏性を**漸近不偏性** (asymptotic unbiasedness) と呼ぶ．また，n が大きくなったときに最尤推定量の持つ分散は，これは次節で述べる Cramér-Rao の不等式の下界に相当し，実は不偏な推定量の中で最小の分散を達成している．これを**漸近有効性** (asymptotic efficiency) という．

3.2.3 Cramér-Rao の不等式*

分布の形がわかっている場合，密度関数が適当な正則条件を満たしているならば推定量の分散の下界に関して次の重要な定理が成り立つ．

定理 3.4 (Cramér-Rao の不等式)．$X_1, X_2, \ldots, X_n$ が互いに独立に母数 θ を含む密度関数 $f(x,\theta)$ を持つ分布に従うとする．このとき *Fisher* 情報量を

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)^2\right] \\ &= \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)\right] \end{aligned}$$

で定義すると，任意の不偏推定量 $\hat{\theta}$ について

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

が成り立つ．

略証．まず $\int f(x,\theta)dx = 1$ の両辺を θ について微分すると

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(x,\theta)dx \\ &= \int \frac{\partial}{\partial\theta} f(x,\theta)\cdot\frac{f(x,\theta)}{f(x,\theta)}dx \\ &= \int \left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right) f(x,\theta)dx \\ &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X,\theta)\right] \end{aligned}$$

得られる．これを更にもう一度微分すると

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\left\{\int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)f(x,\theta)dx\right\} \\
&= \int\left(\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right)f(x,\theta)dx + \int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)\left(\frac{\partial}{\partial\theta}f(x,\theta)\right)dx \\
&= \int\left(\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right)f(x,\theta)dx + \int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2 f(x,\theta)dx \\
&= \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)\right] + \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)^2\right]
\end{aligned}
$$

となり，Fisher 情報量 $I(\theta)$ の 2 通りの定義が等しいことが確認できる．

さて，任意の不偏推定量を $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ とすると，不偏性の定義から

$$\int \hat{\theta}(x_1, x_2, \ldots, x_n)\prod_{i=1}^{n} f(x_i,\theta)dx_i = \theta$$

が成り立つ．両辺を θ で微分すると

$$\int \hat{\theta}(x_1, x_2, \ldots, x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i,\theta)dx_i = 1$$

を得る．一方，密度関数の定義から

$$\int \prod_{i=1}^{n} f(x_i,\theta)dx_i = 1$$

であるので，両辺を θ で微分すると

$$\int \frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i,\theta)dx_i = 0$$

が成り立つ．したがって，上の不偏推定量の関係式と合わせて

$$\int (\hat{\theta}-\theta)\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i,\theta)dx_i = 1$$

を得る．独立な変数の同時密度の微分が

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i,\theta) &= \sum_{j=1}^{n}\frac{\partial}{\partial\theta}f(x_j,\theta)\frac{\prod_{i=1}^{n} f(x_i,\theta)}{f(x_j,\theta)} \\
&= \sum_{j=1}^{n}\frac{\partial}{\partial\theta}\log f(x_j,\theta)\prod_{i=1}^{n} f(x_i,\theta) \\
&= \left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta)\right)\prod_{i=1}^{n} f(x_i,\theta)
\end{aligned}
$$

となることに注意して書き換えると

$$\int (\hat{\theta} - \theta) \left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i, \theta) \right) \prod_{i=1}^{n} f(x_i, \theta) dx_i = 1$$

となる．ところで

$$Z = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i, \theta)$$

と置くと，前に得た

$$\int \left(\frac{\partial}{\partial \theta} \log f(x, \theta) \right) f(x, \theta) dx = \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log f(X, \theta) \right] = 0$$

より，

$$\mathbb{E}_\theta[Z] = \mathbb{E}_\theta \left[\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i, \theta) \right] = 0$$

となるので，上式は $\hat{\theta}$ と Z の共分散が 1 であることを示している．ここで分散・共分散に関する Cauchy-Schwartz の不等式を用いると

$$\mathrm{Var}_\theta(\hat{\theta}) \cdot \mathrm{Var}(\theta) Z \geq \left\{ \mathrm{Cov}_\theta(\hat{\theta}, Z) \right\}^2 = 1$$

すなわち

$$\left\{ \int (\hat{\theta} - \theta)^2 \prod_{i=1}^{n} f(x_i, \theta) dx_i \right\} \left\{ \int \left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i, \theta) \right)^2 \prod_{i=1}^{n} f(x_i, \theta) dx_i \right\}$$
$$\geq \left\{ \int (\hat{\theta} - \theta) \left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i, \theta) \right) \prod_{i=1}^{n} f(x_i, \theta) dx_i \right\}^2 = 1$$

という不等式が得られる．また $X_1, X_2, \ldots, X_n$ の独立性より

$$\begin{aligned}
\mathrm{Var}_\theta(Z) &= \int \left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i, \theta) \right)^2 \prod_{i=1}^{n} f(x_i, \theta) dx_i \\
&= \mathbb{E}_\theta \left[\left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i, \theta) \right)^2 \right] \\
&= \sum_{i=1}^{n} \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X_i, \theta) \right)^2 \right] \\
&= n \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X, \theta) \right)^2 \right] \\
&= n I(\theta)
\end{aligned}$$

となるので，これらを使って纏めると

$$\mathrm{Var}_\theta(\hat{\theta}) \times n I(\theta) \geq 1$$

という不等式が得られ，題意が証明された． □

密度関数に関する適当な条件とは，この場合母数 θ について微分できること，また定理に現れる微分したものの平均が存在しなくてはならないことだと考えておけばよい．これは微分したものが極端に大きな値を取らないことを要請しており，直観的には密度関数が母数 θ に関して滑らかに変化していることを言っている．

ただし一般にこの不等式の下界を達成する不偏推定量が簡単に構成できるとは限らない．

先に述べたように一般に最尤推定量の不偏性は漸近的にしか成り立たないが，その分散は上の定理で示された下界を漸近的に達成している．これを，**漸近有効性** (asymptotic efficiency) という．この意味で最尤推定量はかなり実用的な推定方式であることがわかる．ただし最尤推定は万能ではない．

例えば小数サンプルの場合，外れ値 (outlier) があると過適応 (overfit) しやすいという性質がある．これは対数尤度 $\log f$ が有界でないため，確率密度値が小さく本来あまり起らないはずの見本点が偶々観測された場合，この見本点の影響が非常に強く出てしまうためである．こうした影響を逃れるために縮小推定量や外れ値の判別分析などの併用が行われるが，どういった方法が良いかを一般に論じることは難しい．

また，確率分布に関する事前知識がない場合には尤度関数を決めることができないため，現実の問題においては分布に関して何らかの仮定を置いて尤度関数を考える必要がある．しかしながら，その仮定が正しいかどうかを十分吟味することなく機械的に最尤法を利用するのは危険な場合もあるので，注意しなくてはならない．

いずれにせよ弱点を十分に知った上で，吟味して使えば最尤推定は非常に有効な道具になる．

練習問題

1. 確率モデル

$$X = \theta + \varepsilon, \quad \theta \text{ は未知の母数，} \varepsilon \text{ は誤差}$$

に従う独立な n 個の確率変数 $X_1,\ X_2, \ldots, X_n$ が観測されたとき，これらの観測値を用いて未知の母数 θ を推定することを考える．誤差 ε の確率密度関数 $f(\varepsilon)$ が与えられたとき，観測値 $X_1,\ X_2, \ldots, X_n$ の尤度関数を

$$L(\theta) = \prod_{i=1}^{n} f(X_i - \theta)$$

として，最尤推定量

$$\hat{\theta}_{\mathrm{ML}}(X_1, X_2, \ldots, X_n) = \text{尤度関数} L(\theta) \text{を最大とする} \theta$$

を考える．ただし最尤推定量において尤度を最大とする θ が一つ以上あるときは，可能な全てを答えるものとする．また，記号として定義する

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (標本平均)}, \quad X_{\min} = \min_i X_i \text{ (最小値)}, \quad X_{\max} = \max_i X_i \text{ (最大値)}$$

を適宜用いてできるだけ簡略に答えること．

a) 誤差の分布が $[-1,1]$ 上の一様分布，すなわち誤差の確率密度関数が

$$f(\varepsilon)=\begin{cases}\frac{1}{2} & -1\le\varepsilon\le 1\\ 0 & \text{それ以外}\end{cases}$$

で表されるとき，最尤推定量を求めよ．

b) 誤差の分布が平均 0，分散 1 の正規分布，すなわち誤差の確率密度関数が

$$g(\varepsilon)=\frac{1}{\sqrt{2\pi}}e^{-\frac{\varepsilon^2}{2}},\quad(-\infty<\varepsilon<\infty)$$

で表されるとき，最尤推定量を求めよ．

2. n 個の確率変数 $X_1,\ X_2,\ldots,X_n$ は独立で，いずれも確率モデル

$$X=\theta+\varepsilon,\quad(\theta\text{ は未知の母数，}\varepsilon\text{ は誤差})$$

に従うとする．これらを観測して未知の母数 θ を推定することを考える．誤差 ε の確率密度関数 $f(\varepsilon)$ が与えられたとき，観測値 $X_1,\ X_2,\ldots,X_n$ の尤度関数を

$$L(\theta)=\prod_{i=1}^{n}f(X_i-\theta)$$

として，最尤推定量

$$\hat{\theta}_{\mathrm{ML}}(X_1,X_2,\ldots,X_n)=\text{尤度関数}L(\theta)\text{を最大とする}\theta$$

を考える．ただし最大となる θ が一つ以上存在するときは，全ての可能な値を答えるものとする．また，新たな記号として観測値を小さい順に並べたものを

$$X_{(1)}<X_{(2)}<\cdots<X_{(n)}$$

で定義するので，適宜用いてできるだけ簡略に答えること．

a) 誤差の分布が $[-1,1]$ 上の一様分布で表され，n 個の観測値が得られるときの最尤推定量を求めよ．

b) 誤差の分布が平均 0，分散 1 の正規分布で表され，n 個の観測値が得られるときの最尤推定量を求めよ．

c) 誤差の分布が確率密度関数

$$f(\varepsilon)=\frac{1}{2}e^{-|\varepsilon|}\quad(-\infty<\varepsilon<\infty)$$

で表されるとする．2 個の観測値 X_1,X_2 が得られるときの最尤推定量を求めよ．

d) 前の問で n 個の観測値が得られるときの最尤推定量を，n が奇数のときと偶数のときに分けて求めよ．

3. n 個の確率変数 $X_1,\ X_2, \ldots, X_n$ は独立で，いずれも確率モデル

$$X = \theta + \gamma\varepsilon, \quad (\theta, \gamma \text{ は未知の母数で，} \gamma > 0 \text{ とする．})$$

に従うとする．これらを観測して θ や γ を推定することを考える．確率変数 X の確率密度関数が θ, γ を用いて $f(x, \theta, \gamma)$ で表されるとき，θ および γ の最尤推定量は尤度関数

$$L(\theta, \gamma) = \prod_{i=1}^{n} f(X_i, \theta, \gamma),$$

あるいはその対数をとって

$$\log L(\theta, \gamma) = \sum_{i=1}^{n} \log f(X_i, \theta, \gamma),$$

を最大化する θ と γ の組で与えられる．以下の問いに答えよ．

a) ε の確率密度関数 $g(\varepsilon)$ が与えられたとき，観測値 X の確率密度関数 $f(x, \theta, \gamma)$ を g を用いて表せ．

b) ε が平均 0，分散 1 の正規分布に従うとする．θ の最尤推定量を求めよ．

c) 前問の γ の最尤推定量を求めよ．ただし，前問で求めた θ の最尤推定量は $\hat{\theta}$ と書いてよい．

d) ε が確率密度関数

$$g(\varepsilon) = \frac{1}{2} e^{-|\varepsilon|}, \quad (-\infty < \varepsilon < \infty)$$

で表される分布に従うとする．n が奇数のとき θ の最尤推定量を求めよ．

e) 前問の γ の最尤推定量を求めよ．ただし，前問で求めた θ の最尤推定量は $\hat{\theta}$ と書いてよい．

検定 4

4.1 仮説検定

4.1.1 仮説検定の枠組

母数推定と並んで統計の重要な部分を占める考え方に仮説検定がある．仮説検定は観測されたデータと想定する仮説が矛盾していないかどうかチェックする方法であるが，まずいくつか例を考えてみる．

例 4.1 (平均値の検定)．ある工場で製造される機械の寿命は平均 μ 時間であることが望まれている．n 個の機械をランダムに選び耐久試験をしたところ，寿命はそれぞれ $X_1, X_2, \ldots, X_n$ であった．この機械の平均寿命は μ であると言えるだろうか？

例 4.2 (平均値の差の検定)．A 県と B 県で生産されたの林檎の甘さの違いを調べるために，ある八百屋で売られているそれぞれの林檎一山の糖度を調べたとする．一つ一つの甘さはばらついているので，それぞれの標本平均を計算したところ

$$X_1, X_2, \ldots, X_n \to \hat{\mu}^X = \bar{X}$$
$$Y_1, Y_2, \ldots, Y_m \to \hat{\mu}^Y = \bar{Y}$$

となった．$\hat{\mu}^X$ と $\hat{\mu}^Y$ を比べることによって，本来の平均値が等しい

$$\mu^X = \mu^Y ?$$

と言えるだろうか？

例 4.3. A 社と B 社の開発した 2 つの文字認識機械がある．n 個の文字に対してその性能を調べたところ

	1	2	3	...	n :
A 社	○	○	×	...	○ : 98.1%
B 社	×	○	○	...	○ : 98.0%

のような正答率を示した．このとき A 社の機械は B 社より優れていると言えるだろうか？

これらは典型的な検定の問題となっている．仮説検定とは，確率的なゆらぎを伴なって観測されるデータに基づいて想定している仮説が正しいか否かを統計的に判定し，意思決定を行うための指針を与えることを目的とする．

以下では例 4.2 に述べた 2 つの条件を比較して差があるかどうかを見る問題を考えてみることにする．

まず，2 つの異なる実験を考える．母数推定のところで考えてきたように，観測値は最も単純な確率モデルに従うとする．すな

わち，それぞれの観測値はある母数を中心に誤差が加わっており，得られた 2 通りの観測値の組は

$$X_i = \theta_1 + \varepsilon_{1i}, \quad i = 1, \ldots, n$$
$$Y_j = \theta_2 + \varepsilon_{2j}, \quad j = 1, \ldots, m$$

と記述できるとする．

以降の議論を簡単にするために以下の仮定をおく．

仮定 1 $\varepsilon_{1i}, \varepsilon_{2j}$ は互いに独立に同一の分布に従う．

仮定 2 $\varepsilon_{1i}, \varepsilon_{2j}$ は平均 0，分散 σ^2 である．

$$E(\varepsilon_{1i}) = E(\varepsilon_{2j}) = 0$$
$$E(\varepsilon_{1i}^2) = E(\varepsilon_{2j}^2) = \sigma^2 < \infty$$

仮定 3 $\varepsilon_{1i}, \varepsilon_{2j}$ は正規分布に従う．

さて，ここで問題としているのは $\theta_1 = \theta_2$ とみなしてよいかどうかを考えることである．このとき調べるべき命題「$\theta_1 = \theta_2$」を**統計的仮説** (statistical hypothesis)，あるいは**帰無仮説** (null hypothesis)，あるいは単に**仮説** (hypothesis) という．

まず誤差の分布に正規分布を仮定しているので，θ_1,θ_2 の推定量としてはこれまでの話から

$$\hat{\theta}_1 = \bar{X} = \sum \frac{X_i}{n}$$
$$\hat{\theta}_2 = \bar{Y} = \sum \frac{Y_j}{m}$$

を考えればよいことがわかる．ところで仮説が正しくて $\theta_1 = \theta_2$ であったとしても，観測値から計算される推定値 $\hat{\theta}_1, \hat{\theta}_2$ は推定誤差を含んでいるので，全く同じになることはほどんどないであろう．しかしながら直感的には

$|\hat{\theta}_1 - \hat{\theta}_2|$が小さい $\Rightarrow$ 同じと思って良い

$|\hat{\theta}_1 - \hat{\theta}_2|$が大きい $\Rightarrow$ どうやら違うらしい

と考えられる．この直感をどのように数学的な枠組みに載せるかが問題であるが，以下のように考えることができる．

仮定が正しいならば，母数推定において計算したように

$$\hat{\theta}_1 - \hat{\theta}_2 = \bar{X} - \bar{Y}$$

は，誤差 ε が平均 0，分散 σ^2 の正規分布に従うことに注意すれば

$$\text{平均} \quad E(\bar{X} - \bar{Y}) = 0$$
$$\text{分散} \quad V(\bar{X} - \bar{Y}) = \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2$$

の正規分布に従うことになる．これを正規化 (分散が 1 になるように定数倍) して

$$T = \sqrt{\frac{nm}{n+m}}\frac{\bar{X} - \bar{Y}}{\sigma}$$

という統計量を考えると，T は仮説が正しいとき

$$\begin{aligned} \text{平均} \quad & E(T) = 0 \\ \text{分散} \quad & V(T) = 1 \end{aligned}$$

の正規分布に従うことがわかる．

ここで，正規分布の密度関数はわかっているので $|T| > u_{\alpha/2}$ という事象がおこる確率が α となるような $u_{\alpha/2}$ という値

$$P\left\{|T| > u_{\alpha/2}\right\} = \alpha$$

を考えることにする．この α の値を**有意水準** (level of significance)，あるいは**水準** (level) という．実際によく使われる値は 0.05 とか 0.01 などである．図 4.1 は正規分布における α と $u_{\alpha/2}$ の関係を示している．

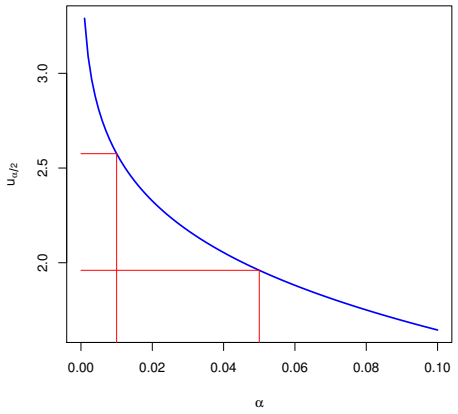

図 4.1: 正規分布における α と $u_{\alpha/2}$ の関係

さて，仮説が正しいのなら $|T| > u_{\alpha/2}$ という事象は確率 α 程度でしか起こらない，すなわち“滅多に起こらない”と言える．したがって $|T| > u_{\alpha/2}$ が起こったのなら，“滅多に起こらないことが起こった”ということであり，仮説を正しいと考えることは疑わしい，すなわち仮説はあやしいと考えられる．統計量 T を値 $u_{\alpha/2}$ と比べ，$|T| > u_{\alpha/2}$ の場合に仮説が正しくないであろうと判定することを**「仮説を (有意) 水準 α で棄却する」**という．逆に仮説が正しいであろうと判定することを**「仮説を受容する」**という．

以上が統計的仮説検定の基本的な考え方である．

一般には以下のような手続きになる．

1. 観測値から計算される検定統計量 T を定める．
2. 帰無仮説が正しいとして T の分布を求める．
3. 十分小さい α を定め，仮説が正しいとき

   $$P\{T \in C_\alpha\} = \alpha$$

   となる領域 C_α を決める．この領域を**棄却域** (critical region) という．

4. 観測された T が C_α に入れば仮説を棄却 (reject)，C_α に入っていなければ仮説を受容 (accept) する．

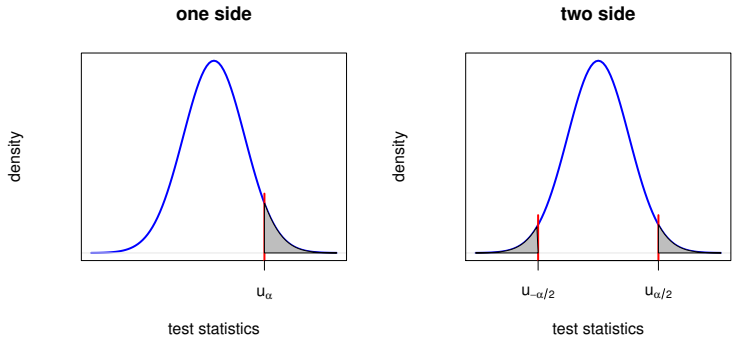

図 4.2: 片側検定と両側検定の棄却域

母数が 1 次元の場合，棄却域の形によって**両側検定**，**片側検定**と区別されることがある．このときの棄却域の形を図 4.2 に示す．
注意しなくてはいけないのは

棄却される $\Rightarrow$ 仮説が正しくない
受容される $\Rightarrow$ 仮説が正しい

と言っているわけではないことである．あくまで

棄却される $\Rightarrow$ 非常に疑わしい
受容される $\Rightarrow$ 正しくないというには証拠不十分

と言っているに過ぎない．一見消極的なようにも思えるが，観測値が確率的である限りは絶対的な判断はできないことを念頭に置かなくてはいけない．

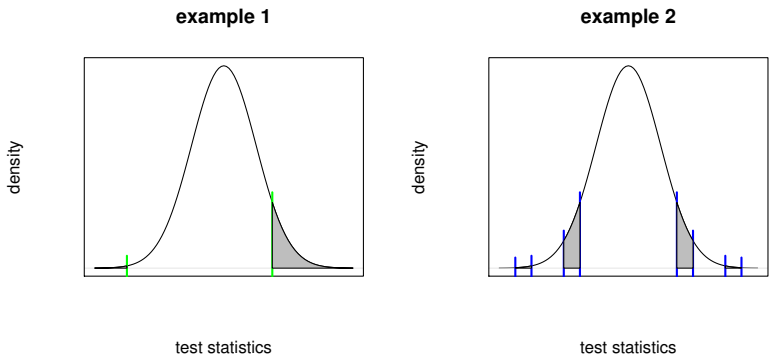

図 4.3: 棄却域は自分で設計してよいので，上記のようなものも可能であるが，これが意味を持つかどうかは吟味する必要がある

例 4.1 についても同様に仮説検定を考えてみる．

例 4.4. 考えるべき問題は，確率モデル

$$X_i = \theta + \varepsilon_i, \quad i = 1, \ldots, n \qquad \varepsilon_i \sim N(0, \sigma^2)$$

のもとで

$$\text{帰無仮説: } \theta = \mu$$

の検定をおこなうことである．ただし分散 σ^2 は既知であるとする．

θ の推定値としては

$$\hat{\theta} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

を用いればよいが，このときこの推定量は

$$\text{平均} \quad E(\hat{\theta}) = \theta$$

$$\text{分散} \quad V(\hat{\theta}) = \frac{\sigma^2}{n}$$

の正規分布にしたがうことは推定量のところで調べたとおりである．ここで

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

という検定統計量を考えると，帰無仮説が正しい，つまり $\theta = \mu$ ならば T は標準正規分布に従う．

例えば $T > 1.96$ という値が出る確率は，帰無仮説が正しい場合には高々 2.5%

$$\int_{1.96}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.025$$

であり，滅多に起こらない事象と言える．これは $T < -1.96$ という状況も同様に考えることができる．したがって，計算した T の値が 1.96 より大きくなったり，あるいは -1.96 より小さくなった場合，すなわち $|T| > 1.96$ の値が出た場合には帰無仮説は怪しいとして棄却する

$$|T| > 1.96 \Rightarrow \text{有意水準 } 0.05 \text{ で帰無仮説は棄却される}$$

というルールを作ることができる．

ただし $|T| > 1.96$ は “起こらない” のではなく，“滅多に起こらない” のだから間違える場合もあることに注意する．これは第 1 種の過誤と呼ばれる．

注意 4.5. ここでは雑音として分散が既知の正規分布を仮定したが，現実問題では分散が未知で，その推定を行なわなくてはならないことが多い．この場合検定統計量として

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}}$$

ただし $\hat{\sigma}$ は不偏分散

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

から求められる推定量である．この時検定統計量は正規分布ではなく t-分布にしたがう．棄却域は t-分布の密度関数を用いて上の場合と同じようにして設定できるが，講義ではこれ以上立ち入らない．

4.1.2 過誤と検出力

推定量のところでその良否を論じたように，検定の方法の良否，すなわち棄却域の良さについても考えることができる．このためには検定すべき**帰無仮説** (null hypothesis) と，その反対である**対立仮説** (alternative hypothesis) の二つを考える必要がある．一般には母数 θ が含まれる領域を用いて

$$\text{帰無仮説: } \theta \in \Theta_0$$
$$\text{対立仮説: } \theta \in \Theta_1$$

のように仮説を定める．このとき次の 2 種類の誤り方が考えられる．

第 1 種の過誤 帰無仮説が正しいのに棄却する

第 2 種の過誤 対立仮説が正しいのに，帰無仮説を受容する

もちろんどちらの過誤の確率も小さいほど良いことは言うまでもない．さて第 1 種の過誤の確率は棄却域の定義より有意水準 α と一致する．一方第 2 種の過誤の確率は棄却域と対立仮説の関係によって決まる．“1-第 2 種の過誤の確率” は，対立仮説が正しいとき帰無仮説を棄却する確率を表すが，これを**検出力** (power) という．通常は同じ有意水準の検定法の中で検出力の大きい，すなわち第 2 種の過誤の確率の小さい方法が良いとされる．

先の例の場合，対立仮説を $\theta_1 \neq \theta_2$ とする．このとき

$$\lambda = \sqrt{\frac{nm}{n+m}}\frac{\theta_1 - \theta_2}{\sigma}$$

とおくと

$$T = \sqrt{\frac{nm}{n+m}}\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sigma}$$

は平均 λ，分散 1 になる．したがって検出力は

$$P(|T| > u_{\alpha/2}|\lambda)$$

で与えられるので，λ に依存して図 4.4 に灰色で示される面積，すなわち棄却域の外側に含まれる対立仮説の確率として計算される．

補足として検定における帰無仮説と対立仮説の様々な設定の仕方について纏めておく．母数が特定の値になるという形で仮説が書かれる場合

$$\theta = \theta_0$$

これを**単純仮説**という．一方母数が一つの値でなく，ある領域に含まれるという形で書かれる場合

$$\theta \in \Theta_0$$

これを**複合仮説**という．**両側検定**，**片側検定**との組み合わせで表 4.1 などの場合がある．

なお「帰無」とは「無に帰す」という意味で，捨て去ってよいかどうか検定するという意味合いがある．例えば

「新しい薬が古い薬の効き目と同じ」
「新しい機械の性能が古い機械の性能と同じ」

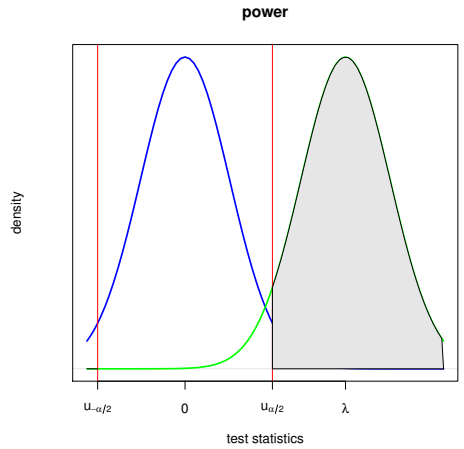

図 4.4: 検出力

	単純帰無仮説			複合帰無仮説		
帰無仮説	$\bar{X} = \mu$	$\bar{X} = \mu$	$\bar{X} = \mu$	$\bar{X} > \mu$	$a < \bar{X} < b$	...
対立仮説	$\bar{X} \neq \mu$	$\bar{X} > \mu$	$\bar{X} < \mu$	$\bar{X} < \mu$	それ以外	...
	両側検定	片側検定				

表 4.1: 仮説の種類

という帰無仮説は当然棄却したいという前提で検定される．本来こうした目的があって仮説検定は発展してきたという事情もあり，歴史的な理由もあって帰無仮説という名称が用いられているが，一般に仮説検定では帰無仮説を捨て去りたいことを示したい場合と捨てられないことを示したい場合とがあることに注意する．

練習問題

1. 例 4.3 の検定の方法について考えてみよ．

2. O 県と Y 県で採れた今年の桃の糖度 (甘さ) を比較したい．糖度を測る測定器の示す値は，

$$\begin{aligned} &\text{O 県の桃の糖度} && X = \theta_1 + \varepsilon \\ &\text{Y 県の桃の糖度} && Y = \theta_2 + \varepsilon \end{aligned}$$

という確率モデルに従っているとする．ここで θ_1, θ_2 は各県の桃の糖度の平均である．また ε は糖度のばらつきを表す確率変数であり，二つの県ともに平均 0，分散 4 の正規分布に従っていることがわかっている．

八百屋さんに行き，O 県産の桃 20 個の糖度 $X_1, X_2, \ldots, X_{20}$ と，Y 県産の桃 25 個の糖度 $Y_1, Y_2, \ldots, Y_{25}$ を測定したとする．θ_1, θ_2 の推定量として標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{20}}{20}, \quad \bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_{25}}{25}$$

を用いるとして，その差について

帰無仮説 H_0 二つの県で糖度は同じ，すなわち $\theta_1 = \theta_2$ である．

対立仮説 H_1 二つの県で糖度は異なる，すなわち $\theta_1 \neq \theta_2$ である．

という仮説を考えて検定を行いたい．このとき，以下の問に答えよ．なお，正規分布に従う独立な確率変数の和は，やはり正規分布に従うことは既知として良い．また，標準正規分布の定積分

$$b = \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

については以下の表を用いよ．($a > 3.1$ では $b = 0$ として計算して良い．)

a	0.96	1.04	1.28	1.64	1.88	1.96	2.04	2.33	2.50	2.58	2.96	3.04
b	0.169	0.149	0.10	0.05	0.03	0.025	0.021	0.01	0.006	0.005	0.0015	0.0012

表 4.2: 正規分布表

a) 標本平均 $\bar{X}$, $\bar{Y}$, およびその差 $\bar{X} - \bar{Y}$ の分散を求めよ．

b) 前問の結果を用いて，帰無仮説が正しいときに標準正規分布に従うような検定統計量 T を，標本平均の差 $\bar{X} - \bar{Y}$ を用いて構成せよ．

c) 検定方法として，$|T| > c$ のとき帰無仮説を棄却するように棄却域を定めることを考える．有為水準が 5%(帰無仮説が正しい時に $|T| > c$ が起こる確率が 5%となる) となるように c を定めよ．

d) 帰無仮説が正しくなく，実は真の値が

$$\theta_1 = 15.4, \quad \theta_2 = 13.6$$

であるときに，前問の検定方法で誤って帰無仮説が受容される確率 (第二種の過誤の確率) はどのくらいか？

e) 実際にある八百屋さんで調べたところ，

$$\bar{X} = 14.2, \quad \bar{Y} = 15.5$$

であった．このとき，上で定めた有為水準 5%の検定で，帰無仮説は棄却されるか，されないか？ また，c の値を変えて有為水準を 1%とした場合は棄却されるか，されないか？

3. ある工場で作った新しいエンジンの性能が，古いエンジンより良くなったのかどうか比較したい．エンジンの性能を示す最大トルクは，

$$X = \theta + \varepsilon \quad (\theta \text{ は最大トルクの平均，} \varepsilon \text{ は誤差})$$

という確率モデルに従っているとする．工場で試作した 16 台のエンジンについて最大トルク $X_1,\ X_2, \ldots, X_{16}$ を測定し，θ の推定量として標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{16}}{16}$$

を用いるとする．古いエンジンについては過去の経験から最大トルクの平均値は 22.4(単位は $kg \cdot m$) であることがわかっている．この仮定のもとで

帰無仮説 H_0 二種のエンジンの性能は同じ，すなわち $\theta = 22.4$ である．

対立仮説 H_1 新しいエンジンの性能は良い，すなわち $\theta > 22.4$ である．

という仮説を考えて検定を行いたい．このとき，以下の問に答えよ．

a) 新しいエンジンの最大トルクの誤差 ε が平均 0，分散 46.24 の正規分布に従っているとわかっているとき，標本平均 $\bar{X}$ の分散を求めよ．

b) 帰無仮説が正しいときに標準正規分布 (平均 0，分散 1 の正規分布) に従うような検定統計量 T を，標本平均 $\bar{X}$ を用いて構成せよ．

c) 適当な c を用いて $T > c$, $T < c$, $|T| > c$ の 3 種類の棄却域 (帰無仮説を棄却する条件を表す式) を考えるものとする．上の対立仮説に対してもっとも適当と思われる棄却域はどれか？ また，それは何故か簡単に説明せよ．

d) 前の問の棄却域において，有為水準が 5%(帰無仮説が正しい時に棄却される確率が 5%となる)，および 1%となるように c を定めよ．

e) 帰無仮説が誤りで，新しいエンジンの最大トルクの真の平均が $\theta = 27.5$ であるときに前の問の検定方法を用いると，帰無仮説が誤って受容される確率 (第二種の過誤の確率) はどのくらいか？

f) 実際に調べたところ，$\bar{X} = 25.4$ であった．このとき，上で定めた有為水準 5%の検定で，帰無仮説は棄却されるか，されないか？ また，c の値を変えて有為水準を 1%とした場合は棄却されるか，されないか？

4. ある大学の 1 年生と 4 年生に英語の統一試験を行って，その成績を比較することを考える．学生の成績はそれぞれ平均を θ_1,θ_2 として

$$1\text{ 年生の成績: } X = \theta_1 + \varepsilon, \quad 4\text{ 年生の成績: } Y = \theta_2 + \varepsilon$$

という確率モデルに従っているとする．ここで ε は成績のばらつき具合を表す確率変数であり，平均 0，分散 60 の正規分布に従っていることがわかっている．1 年生 n 名と 4 年生 m 名を無作為に抽出して観測値 $X_1,\ X_2, \ldots, X_n$，および

$Y_1, Y_2, \ldots, Y_m$ を得たとき，θ_1, θ_2 の推定量として標本平均 $\bar{X}, \bar{Y}$ を用いる．

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}, \quad \bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_m}{m}.$$

帰無仮説 H_0 1年生と4年生で成績は同じ，すなわち $\theta_1 = \theta_2$ である．

対立仮説 H_1 1年生より4年生の方が成績は良い，すなわち $\theta_1 < \theta_2$ である．

という仮説を考えて検定を行いたい．以下の問に答えよ．

a) $n = 10$,$m = 20$ のとき標本平均の差 $\bar{X} - \bar{Y}$ の分散を求めよ．

b) 検定統計量 T として，帰無仮説が正しいときに標準正規分布に従い，また帰無仮説を棄却する棄却域として $T > c$ が適当となるようなものを，$\bar{X}$ と $\bar{Y}$ を用いて構成せよ．

c) 有為水準が5%(帰無仮説が正しい時に $T > c$ が起こる確率が5%) となるように c を定めよ．

d) 実際に調べたところ，$\bar{X} = 54.8$, $\bar{Y} = 59.8$ であった．このとき，上で定めた有為水準5%の検定で，帰無仮説は棄却されるか？

e) 前問において抽出する人数が $n = m = 15$ であった場合，その検定結果はどのようになるか？

またその結果を踏まえて，対立仮説が正しい (実は4年生の方が成績が良い) ときに誤って帰無仮説が受容される確率 (第二種の過誤の確率) を考えると，2つのグループで抽出する合計の人数が一定の場合，各グループで抽出する人数比についてどのようなことが考えられるか簡潔に説明せよ．

5. ある研究室で開発した新型のLED電球が，従来のものより明るくなったかどうか調べることを考える．

新型のLED電球の光量は平均を θ として，

$$X = \theta + \varepsilon$$

という確率モデルに従っていることがわかっているとする．ここで ε は製造過程でのばらつきや測定時に発生するノイズなどの誤差をまとめた確率変数であり，その分布は確率密度関数 $f(\varepsilon)$ で記述されるとする．

検定のためのデータとしては，作成した n 個の試作品の光量を測定することによって観測データ

$$X_1, X_2, \ldots, X_n$$

が得られるものとする．そこで，従来のLED電球の平均光量が θ_0 であるとして，

帰無仮説 H_0**:** 新型の光量は従来と同じ，すなわち $\theta = \theta_0$ である

対立仮説 H_1**:** 新型の光量は従来より多い，すなわち $\theta > \theta_0$ である

という仮説を考えて検定を行いたい．

これらを踏まえて，以下の問に答えなさい．

a) 誤差の確率密度関数が母数 (パラメタ)$\lambda > 0$ を用いて

$$f_1(\varepsilon) = Ce^{-\lambda|\varepsilon|}, \quad (-\infty < \varepsilon < \infty)$$

で表されるとする．このとき，f_1 が確率密度関数となるように C を定めなさい．

b) 誤差の確率密度関数が前問の f_1 のとき，θ の最尤推定量を求めなさい (n の偶奇で場合分けしなさい)．

c) 誤差が平均 0, 分散 σ^2 の正規分布に従っているとき，θ の最尤推定量を求めなさい．

d) 前問で求めた最尤推定量の平均と分散を求めなさい．

e) 試作品の誤差の分布について詳しく調べたところ，平均 0, 分散 σ^2 の正規分布で十分良く近似できることがわかった．そこで以下では誤差は平均 0, 分散 σ^2 の正規分布に従っているとして，検定統計量を構成していく．

以下の 2 つの条件を満たす検定統計量 T を 1 つ構成しなさい．

- 帰無仮説が正しいときに標準正規分布に従う．
- 適当な定数 c を用いて $T > c$ のときに帰無仮説を棄却するための棄却域が定義される．

f) 有意水準が 1%(帰無仮説が正しいときに $T > c$ が起きる確率が 1%) となるように c を定めよ．

g) 実際に調べたところ，従来の光量の平均値は $\theta_0 = 52.5$ であり，観測データは $n = 6$ で

$$X_1 = 59.8, \quad X_2 = 56.6, \quad X_3 = 54.1, \quad X_4 = 59.9, \quad X_5 = 57.1, \quad X_6 = 51.5$$

(単位は 1 ワットあたりのルーメン), 誤差の分散は $\sigma^2 = 24.0$ であった．このとき，上で定めた検定統計量はいくつか？ また，上で定めた有意水準 1%の検定で，帰無仮説は棄却されるか，受容されるか答えなさい．

h) 従来の光量の平均と観測データの誤差の分散は前問と同様とする．新型の光量の真の平均値が $\theta = 56.5$ であり，対立仮説が正しい場合を考える．このとき，帰無仮説を誤って受容する確率 (第 2 種の過誤) が 5%以下となるためには，最低いくつの試作品を調べる必要があると考えられるか答えなさい．

6. ある人気シリーズの新刊発売前になると、A 書店と B 書店は共同で前作を買った人に抽選券を発行し、当たった人は当日並ばなくても優先的に新刊を購入できるというサービスを行っている。抽選券は 1 人に 1 枚だけ配布され、貰った人は名前を書いて A 書店か B 書店のいずれかに提出する。抽選は A 書店と B 書店で別々に行われるため、新刊の仕入

れ状況によってA書店とB書店では抽選に当たる確率は異なるものとする。

今年抽選券を貰ったC君は、A書店とB書店のどちらに出すかを悩んでいたが、前回の抽選結果について

- A書店は5人に1人の割合で当たりが出た。
- B書店に出したC君の友人n人に聞いたところm人が当たっていた。

という情報を得ることができたので、これにもとづいてどちらの書店に出すかを考えることにした。

以上の状況を踏まえて以下の問いに答えなさい。

a) B書店に出した抽選券が当たるかはずれるかを、確率変数Xを用いて

$$X = \begin{cases} 1 & (\text{当たり}) \\ 0 & (\text{はずれ}) \end{cases}$$

で表すことにする。$X = 1$が起きる確率を未知母数θを用いて

$$P(X = 1) = \theta,\ (0 \leq \theta \leq 1)$$

とおくとき、Xの平均と分散をθを用いて表しなさい。

b) C君の友人n人の当たりはずれをそれぞれ$X_1, X_2, \ldots, X_n$で表すことにする。このn人の当たりはずれは互いに独立であるとし、$X_1, X_2, \ldots, X_n$が与えられたときのθの尤度関数を求めなさい。(できるだけ簡単な形で書くこと。)

c) $X_1, X_2, \ldots, X_n$が与えられたときのθの最尤推定量$\hat{\theta}$を求めなさい。

d) A書店とB書店で当たりの確率が変わらないかどうかを考えるために、

帰無仮説 H_0**:** B書店の当たりの確率 $= 0.2$

を設定し検定することを考える。帰無仮説が正しいときに、上で求めた最尤推定量$\hat{\theta}$の平均と分散を求めなさい。

e) 帰無仮説が正しいときに平均が0、分散が1となるような検定統計量の一つ(最も単純なもの)を、最尤推定量$\hat{\theta}$を用いて構成しなさい。

f) 帰無仮説が正しいときに上で構成した検定統計量(以下ではこの検定統計量をTと書くことにする)の特性関数を求めなさい。

g) 平均μ，分散σ^2の正規分布の特性関数を求めなさい．

h) 前々問で$n \to \infty$とするとき，検定統計量Tはどのような分布に収束するか？ その確率密度関数を書き下しなさい．なお，必要であれば

$$\lim_{n\to\infty} \left(1 + \frac{r(x) + s_n(x)}{n}\right)^n = e^{r(x)}, \quad \text{ただし } s_n(x) \text{ は } \lim_{n\to\infty} s_n(x) = 0$$

という計算を使ってよい.

i) 今データの個数 n が十分に大きいとして，上で構成した検定統計量 T は前問の結果得られる確率密度関数を用いて近似して良いものとする．以下では

帰無仮説 H_0**:** B 書店の当たりの確率 $= 0.2$

対立仮説 H_1**:** B 書店の当たりの確率 $\neq 0.2$

を考えるものとする。帰無仮説が正しい時に $|T| > c$ が起こる確率が 5%(有為水準 5%) となるように c を定めよ.

j) 実際に調べたところ，$n = 25$,$m = 3$ であった (n は十分に大きいと考えてよい) とする．このとき，上で定めた有為水準 5%の検定で，帰無仮説は受容されるか？または棄却されるか？

k) B 書店の方が当たりやすいという噂を耳にしたので、

帰無仮説 H_0**:** B 書店の当たりの確率 $= 0.2$

対立仮説 H_1**:** B 書店の当たりの確率 > 0.2

という検定を考えることにする。このとき有為水準が 5%となる適当な棄却域 (その領域に T が入ると帰無仮説を棄却する領域) を設定せよ。

l) 前問で定めた水準では、$n = 36$ のときに m がいくつ以上であったら、B 書店の方が当たりやすいと考えて良いであろうか?

極限定理 A

多数の観測が行われたときに，その和や算術平均などを考えると特徴的な分布が現れる．ここで言う分布とは，多数の観測データを集めた標本ごとに計算される値を複数の標本について計算したとき，それらの値が従う確率法則のことである．これらの性質は，数学的には確率変数の和に関する極限定理として整理されている．この章では，計算機で生成することができる乱数を用いて統計で重要な極限定理のシミュレーションを行う．

A.1 大数の法則

以下では観測される個々のデータ (確率変数)X は有限の分散を持つとする．多数の独立な確率変数の和については，以下の強い主張が成り立つ．

定理 A.1 (大数の強法則)．$\{X_n\}$ が独立で，$\{\mathbb{V}[X_n]\}$ が有界ならば

$$\frac{S_n - \mathbb{E}[S_n]}{n} \to 0 \quad a.s.$$

ただし，$a.s.$ とは *"almost surely"* を意味し，概収束 *(*確率変数としては強い意味での収束にあたる*)* を意味する．

これを**大数の法則** *(law of learge numbers)* という．定理は無限に多くの変数の和について厳密に成り立つと言っているが，実用上は十分に多くの変数の和がその平均に一致すると解釈すれば良い．

特に個々の観測データが同分布に従う場合には以下のようにまとめられる．

定理 A.2 (同分布の場合の大数の強法則)．$\{X_n\}$ が独立で同じ分布に従うとする．このとき

$$\frac{S_n}{n} \to \mathbb{E}[X] \quad a.s.$$

例えば同じ実験を複数回繰り返して算術平均で平均値を求める状況に対応するが，この方法で真の平均値の良い推測が行えることが保証される．

A.2 中心極限定理

大数の法則の収束の仕方をより精密に議論したのが，以下の**中心極限定理** *(central limit theorem)* である．

定理 A.3 (中心極限定理)．$\{X_n\}$ は独立で，$S_n = \sum_{i=1}^{n} X_i$ とする．

$$T_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}}$$

の確率法則は $n \to \infty$ のとき標準正規分布 *(*平均 0, 分散 1 の *Gauss* 分布*)* に近づく.

なお, 上の表現は若干正確さを欠いており, 中心極限定理が成立するためには $\{X_n\}$ に関する適当な条件 (Lingeberg の条件などが有名) が必要となる. こうした条件の内で, 最も単純なものは個々の観測データが同分布に従う場合であり, その主張は以下のように纏められる.

定理 A.4 (同分布の場合の中心極限定理). $\{X_n\}$ が独立で同分布に従う場合

$$T_n = \frac{S_n - n\mathbb{E}[X]}{\sqrt{n\mathbb{V}[X]}}$$

の確率法則は $n \to \infty$ のとき標準正規分布 *(*平均 0, 分散 1 の *Gauss* 分布*)* に近づく.

A.3 重複対数の法則

これ以外の興味深いものとして, 観測データの和の振幅の挙動に関して**重複対数の法則** *(law of the iterated logarithm)* が知られている.

定理 A.5. 中心極限定理よりさらに仮定を強めた条件が必要であるが,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{S_n - \mathbb{E}[S_n]}{(2\mathbb{V}[S_n]\log\log\mathbb{V}[S_n])^{1/2}} &= 1 \quad a.s., \\
\liminf_{n\to\infty} \frac{S_n - \mathbb{E}[S_n]}{(2\mathbb{V}[S_n]\log\log\mathbb{V}[S_n])^{1/2}} &= -1 \quad a.s., \\
\limsup_{n\to\infty} \frac{|S_n - \mathbb{E}[S_n]|}{(2\mathbb{V}[S_n]\log\log\mathbb{V}[S_n])^{1/2}} &= 1 \quad a.s.
\end{aligned}$$

が成り立つ.

A.4 少数の法則

少数の法則 *(law of rare events, Poisson limit theorem)* は和に関する法則ではなく, 滅多に起こらない事象が起こる回数の分布に関する法則である. 例えば, ある製品の不良品率 p はとても小さいとする. 一日に N 個 (非常に多数とする) 生産するとき, 不良品は平均的には $\lambda = pN$ 個発生するが, 日によって不良品の個数には多少のばらつきが生じる. この分布は *Poisson* **分布** *(Poisson distribution)*

$$\Pr(X = k) = \frac{\lambda^k e^{-k}}{k!}$$

に従うことが知られており, 以下のようにまとめられる.

定理 A.6 (少数の法則)**.** $\{X_n\}$ は独立で

$$\Pr(X_k = 1) = p_k, \quad \Pr(X_k = 0) = 1 - p_k$$

とする.

$$\bar{p}_n = \max_k p_k \to 0, \quad p_n = \sum_k p_k \to \lambda \qquad (n \to \infty)$$

とすれば

$$N_n = \sum_{k=1}^{n} X_k$$

の確率法則は $n \to \infty$ のとき λ をパラメタとする *Poisson* 分布に近づく.

簡単な記述統計量 B

記述統計量 (descriptive statistics) とはデータを簡潔に要約して表すための統計値のことで，要約統計量，基本統計量とも言われる．ヒストグラム (あるいは密度関数) や箱ひげ図などのグラフと併用して，その集団全体の特徴を表す重要な指標となる．

B.1 いろいろな記述統計量

ここでは，比較的良く用いられる統計量を，その背景となるモーメント，順序，分布という考え方に基づいて分類する．

B.1.1 モーメントに基づく統計量

N 個のデータ

$$x_1, x_2, \ldots, x_N$$

が得られているとき，その**平均** *(mean)* を

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (\text{平均})$$

で定義する．これを用いて，一般に**平均値まわりの k 次モーメント** *(**積率**; k-th moment)* は

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^k \qquad (k \text{ 次モーメント})$$

で定義される．特に 2 次モーメントを**分散** *(variance)*

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (\text{分散})$$

と言い，その平方根を**標準偏差** *(standard deviation)* という．

正規分布は，平均と分散が決まれば分布を一意に特定できるため，平均 μ と分散 σ^2(または標準偏差 σ) は正規分布に比較的近い分布をもつデータの性質を記述する統計量として基本となる．

正規分布からのずれを知るには平均値まわりの高次モーメントが利用されるが，特に 3 次モーメントを正規化した**歪度** *(skewness)*

$$\gamma = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3 \qquad (\text{歪度})$$

および 4 次モーメントを正規化した**尖度** *(kurtosis)*

$$\kappa = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^4 \qquad (\text{尖度})$$

が重要である．

B.1.2 順序に基づく統計量

N 個のデータを小さい順に並べたものを

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(N)}$$

と書くことにする．最も小さい値 $x_{(1)}$ を**最小値** *(minimum value)*，最も大きい値 $x_{(N)}$ を**最大値** *(maximum value)* と呼ぶ．

並べ変えたときにデータの中央の位置にあるデータを**メディアン** *(***中央値***; median)* といい，

$$x_{((N+1)/2)} = \begin{cases} x_{((N+1)/2)}, & (N \text{ が奇数のとき}) \\ (x_{(N/2)} + x_{(N/2+1)})/2, & (N \text{ が偶数のとき}) \end{cases}$$

と書くことにする．以降，$(\cdot)$ 内の添字が割り切れない場合，すなわち $i = a$(整数) $+ b$(0 以上 1 未満の小数) となる場合は，前後 2 つのデータの重み付き平均 (添字の線形補間，内分点)

$$x_{(i)} = (1-b) \times x_{(a)} + b \times x_{(a+1)}$$

で定義する．

メディアンと同様にデータの 25%，75%の位置にあるデータ (そのデータより小さいデータが全体の何%を占めるかを考える) を**第 *1* 四分位点** *(first quartile)* $x_{((N+3)/4)}$，**第 *3* 四分位点** *(third quartile)* $x_{((3N+1)/4)}$ という．メディアンは 50%の位置にあるデータなので，第 2 四分位点でもある．また，$q(0 \leq q \leq 1)$ の位置にあるデータを q-**分位点** *(quantile)*，また α%の位置にあるデータを α-**百分位点** *(percentile)* と呼ぶこともある．この名称を使えば，最小値は 0-分位点，最大値は 1-分位点などとなる．

データの中で小さいものと大きいものを除いて計算した平均を**トリム平均** *(***刈込平均***; trimmed mean)* という．最小値や最大値は集団内での特殊な値 (異常値，外れ値) となる場合があり，この影響を取り除く効果がある．

最大値と最小値の差を**範囲** *(range)*，第 3 四分位点と第 1 四分位点の差を**四分位範囲** *(interquartile range, IQR)* と呼び，分散や標準偏差のようにデータのばらつきを表す記述統計量として用いる．

B.1.3 頻度に基づく統計量

データの中で最も頻度が高く現れる値を，**モード** *(***最頻値***; mode)* と呼ぶ．モードはデータが有限個の値を取る場合に特に有効であるが，データが連続で無限に多くの値を取ることができる場合には注意が必要である．連続なデータの場合でも有限個の観測データに対してモードは定義できるが，偶々観測値として現れた値なので，その意味はよく考えなくてはならない．必要に応じて，例えば区分的に集計するなどの工夫をすることもとある．

B.2 少数サンプルの性質

集団全体のデータが得られる場合には，上記の記述統計量はその集団の特徴を表す値として重要であるが，現実の問題においては集団の全てを調べることが不可能な場合がある．例えば，工場で

生産される全ての製品の検査するために破壊実験を行うことはできないし，テレビの視聴率を調べるために全ての家庭に調査機器を設置することはできない．このような場合に，対象とする**母集団** *(population)* からランダムに抽出した少数の**標本** *(sample)* を用いて，母集団の真の統計量の推測を行うことがある．少数の標本に基づいて計算された値は一般に母集団の真の値とは一致しないが，ランダムサンプルに基づく統計量の性質は計算機実験で確かめることができる．

B.2.1 標本平均・メディアンの性質

少数の標本から計算される平均とメディアンをそれぞれ**標本平均** *(sample mean)*，**標本メディアン** *(sample median)* と言う．これらの統計的な性質は母集団の分布に大きく左右される．一般に標本平均は平均の良い推定値を与えると考えがちだが，平均から離れた値 (外れ値) が多く出現する裾の重い分布の場合には，外れ値の影響を受け易いので注意が必要である．

B.2.2 標本分散の性質

標本から計算される分散を**標本分散** *(sample variance)* という．標本分散は平均的には真の分散より若干小さめの値を推定することがわかっている．これを補正して平均的に偏りのない推定を行うために**不偏分散** *(unbiased variance)*

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (n \text{ は標本の数})$$

を用いることが多い．

C 標本分布

母集団から無作為抽出を行い統計量を計算するという実験を無限回繰り返したとき，統計量の分布は特徴ある特定の分布となる．例えば前回行った中心極限定理のシミュレーション実験のように，標本の数が十分に大きければ標本平均は正規分布に従っていた．標本から計算される統計量の分布は標本分布 (sampling distribution) と呼ばれ，区間推定や検定など統計において重要な役割を果たす．この章では，特徴的な分布をいくつかを挙げて紹介する．

C.1 正規母集団からの標本分布

まず，理論的に取り扱いが簡単で標本分布を具体的に導出できる場合として，母集団の個々のデータが正規分布に従っている状況を考え，この母集団から無作為に抽出された標本のいくつかの統計量の性質について考えることにする．

C.1.1 正規分布

標本平均の分布は，標本数によらず正規分布に従う．これは独立な正規分布の和は正規分布に従うという，正規分布の特別な性質による．

平均 μ，分散 σ^2 (標準偏差 σ) の正規分布 $\mathcal{N}(\mu, \sigma^2)$ の密度関数は

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

である．

$X_1, X_2, \ldots, X_n$ を正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従う独立な確率変数とする．標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

を考えると，$\bar{X}$ は平均 μ，分散 σ^2/n の正規分布に従うこと

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

が示される．

C.1.2 χ^2-分布

標本分散の分布は χ^2-分布に従う．

$X_1, X_2, \ldots, X_d$ を標準正規分布 $\mathcal{N}(0, 1)$ に従う独立な確率変数とする．このとき

$$Z = X_1^2 + X_2^2 + \cdots + X_d^2$$

の従う分布を自由度 d の χ^2-分布 ($\chi^2(d)$ と書く) と呼ぶ．自由度 d の χ^2-分布の密度関数は

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

ただし，Γ はガンマ関数

$$\Gamma(z) = \int_0^\infty t^{z-1} e^t dt$$

である．

$X_1, X_2, \ldots, X_n$ を正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従う独立な確率変数とする．標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

からの差の平方和の分散に対する比

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

を考えると，S は自由度 $d = n - 1$ の χ^2-分布に従うこと

$$S \sim \chi^2(n-1)$$

が示される．なお，平均の推定のために標本平均を使っているので，偏差の平方和の自由度は (標本数 -1) となることに注意する．この結果を用いると，不偏分散

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

の分布は χ^2-分布を拡大縮小した分布であり，その平均値は真の分散 σ^2 となり，不偏性を持つことがわかる．

χ^2-分布は主に分散の区間推定や検定に用いられる．これ以外に適合度検定や独立性の検定など様々な場面でも用いられる．

C.1.3 t-分布

不偏分散で正規化した標本平均の真の平均からの偏差の分布は t-分布に従う．t-分布とは正規分布に従う確率変数を χ^2-分布に従う独立な確率変数で除したときに現れる分布であり，標本平均が正規分布，不偏分散が χ^2-分布に従うことによる．

X_1 を標準正規分布 $\mathcal{N}(0, 1)$，X_2 を自由度 d の χ^2-分布に従う独立な確率変数とする．このとき

$$Z = \frac{X_1}{\sqrt{X_2/d}}$$

の従う分布を自由度 d の t-分布 ($\mathcal{T}(d)$ と書く) と呼ぶ．自由度 d の t-分布の密度関数は

$$f(x) = \frac{\Gamma((d+1)/2)}{\sqrt{d\pi}\Gamma(d/2)} (1 + x^2/d)^{-(d+1)/2}$$

である．

$X_1, X_2, \ldots, X_n$ を正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従う独立な確率変数とする．標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

と，不偏分散

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

を用いて，変数

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

を考えると，T は自由度 $d = n - 1$ の t-分布に従うこと

$$T \sim \mathcal{T}(n-1)$$

が示される．

t-分布は，特に推定における信頼区間を構成するときに威力を発揮する．

C.1.4 F-分布

2 つの母集団が同じ分散をもつかどうか調べるためには，その比を比べてみればよい．2 つの標本分散の比の分布は F-分布に従う．F-分布とは，独立な 2 つの χ^2-分布の比の分布であり，標本分散が χ^2-分布に従うことによる．

X_1 を自由度 d_1 の χ^2-分布 X_2 を自由度 d_2 の χ^2-分布に従う独立な確率変数とする．このとき

$$Z = \frac{X_1/d_1}{X_2/d_2}$$

の従う分布を自由度 d_1, d_2 の F-分布 ($\mathcal{F}(d_1, d_2)$ と書く) と呼ぶ．自由度 d_1, d_2 の F-分布の密度関数は

$$f(x) = \frac{1}{B(d_1/2, d_2/2)}\left(\frac{d_1 x}{d_1 x + d_2}\right)^{d_1/2}\left(1 - \frac{d_1 x}{d_1 x + d_2}\right)^{d_2/2} x^{-1}$$

ただし，B はベータ関数

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

である．

$X_1, X_2, \ldots, X_m$ は正規分布 $\mathcal{N}(\mu, \sigma^2)$ に，$Y_1, Y_2, \ldots, Y_n$ は正規分布 $\mathcal{N}(\nu, \sigma^2)$ に従う独立な確率変数とする．2 つは平均は異なるが同じ分散を持つことに注意する．このとき，それぞれの不偏分散の比

$$F = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2}{m-1} \Big/ \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}{n-1}$$

を考えると，F は自由度 $d_1 = m-1, d_2 = n-1$ の F-分布に従うこと

$$F \sim \mathcal{F}(m-1, n-1)$$

が示される．

F-分布は，上記のような単純な分散の比較以外に，残差の分散を比較して統計モデルの良否を判断する分散分析と呼ばれる分野などで活躍する．

C.2 その他の標本分布

母集団の分布が正規分布でない場合でも，標本数がある程度に大きければ母集団の分布とは無関係に標本平均は正規分布で十分良く近似できる．これは前章の中心極限定理によるものである．しかしながら母集団の分布が正規分布でない場合は，統計量の分布を書き下すことは一般に難しい．また，母集団が正規分布であっても，メディアンや分位点のようにその標本分布が複雑になって書き下すことが難しい場合もある．こうした場合は計算機上の乱数を用いたシミュレーション実験に頼ることになる．

C.2.1 モンテカルロ法

母集団の分布が具体的にわかっている場合，あるいは適当な分布で十分良く近似されることがわかっている場合には計算機上でその分布に従う乱数を生成することによって標本の抽出と目的の統計量の計算を繰り返し行い，標本分布を近似的に求めることができる．こうした計算機による疑似乱数を用いたシミュレーション実験はモンテカルロ法 (Monte Carlo method) と呼ばれる．

C.2.2 ブートストラップ法

標本抽出とは本来母集団からの無作為抽出でなくてはならないが，母集団の分布に関する情報がない場合には，母集団の分布を用いた上記のようなモンテカルロ法を利用することはできない．このような場合に利用される特別なモンテカルロ法として**ブートストラップ法** *(bootstrap method)* がある．この方法は，得られている 1 つの標本から**復元抽出** *(sampling with replacement)* をすることによって擬似的な標本抽出を行うものである．

なお，ブートストラップ法によって求めた擬似的な標本分布と真の標本分布の間には一般に偏差がある．例えばシミュレーション例でも明らかなように，観測データの標本平均は母集団の平均とは一般に異なるので，ブートストラップによる標本分布の中心は真の標本分布の中心からずれてしまう．したがってブートストラップ法によって求めた分布を標本分布の代わりとして用いる場合には注意が必要である．

D 点推定

母集団から無作為に抽出したデータ (標本) を用いて，母集団の持つ何らかの特性量について推測を行うことを一般に**推定** *(estimation)* と呼ぶ．前回議論した標本から計算される統計量の分布を調べることによって，合理的な推定方法を議論することが本章の目的である．

D.1 点推定と不偏性

母集団の平均値や分散など，母集団の持つ特徴量を 1 つの値で推定する方法を**点推定** *(point estimation)* と呼ぶ．標本を用いて母集団に関する未知の値を推定する方法を**推定量** *(estimator)* と呼び，推定の対象となる量 θ に対して

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

のように $\hat{}$ を付けて表すことが多い．例えば標本平均や標本分散・不偏分散などが点推定における典型的な推定量となる．

推定量一般に要請される基本的な性質として**不偏性** *(unbiasedness)* がある．これは，対象とする特性量に対して，その推定量の平均が真の値と一致する性質

$$\mathbb{E}\left[\hat{\theta}(X_1, X_2, \ldots, X_n)\right] = \theta \text{ (真の値)}$$

を指す．例えば母集団から無作為に抽出したデータを用いて計算される標本平均は明らかに不偏性を持つため，推定された値は真の値のまわりに分布することを意味し，目的とする統計量が偏りなく推定されることを保証している．

D.1.1 平均の点推定

前章で調べたように，母集団が平均 μ，分散 σ^2 の正規分布に従う場合，標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

の分布は平均 μ，分散 σ^2/n の正規分布

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

に従う．

一般の分布でも標本の大きさが十分大きければ，大数の法則によって標本平均が真の平均に次第に近付くことが保証される．標本の大きさが小さい場合には，標本に含まれるデータによって真の平均から離れた値も生じるが，標本分布の平均は真の平均と一致する．したがって標本平均は一般の分布でも不偏であり，不偏

性の観点からは安全な推定量であることがわかる．一方，トリム平均やメディアンはそうとは限らない．例えば母集団の分布が対称であればその平均はメディアンを用いても推定することができるが，分布が非対称であればメディアンを用いると推定に偏りが生じそうなことは容易に想像されるだろう．

D.1.2 分散の点推定

母集団が正規分布に従う場合，分散の推定量である標本分散や不偏分散 (の定数倍) の分布が χ^2-分布に従うことは，前章で確認した．

母集団が一般の分布に従う場合には，標本分散や不偏分散の分布を厳密に論じることは一般に難しいが，それらの推定量の平均を計算することはできる．平均 μ，分散 σ^2 の分布 (正規分布とは限らない) に従う母集団から得られた n 個の標本を用いると

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

という関係が成り立つ．これより，不偏分散が不偏であることがわかる．

D.2 最尤推定

母集団の平均と分散という基本的な特徴量の推定は，比較的単純で直感的にもわかりやすいものであった．以下では，母集団に関する一般の特徴量として母集団の分布の**母数** *(**パラメタ**; parameter)* を推定する方法について考える．

D.2.1 尤度関数

確率変数 X が従う確率法則の密度関数を $f(x;\theta)$ で表すことにする．このとき θ は分布の**母数** *(**パラメタ**)* と呼ばれ，分布の形状を決める役割りを持つ．例えば，正規分布の密度関数は

$$f(x;\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$

であるので，母数 θ は分散が既知の場合には

$$\theta=\mu,$$

分散が未知の場合には

$$\theta=(\mu,\sigma)\quad ;\text{あるいは}\quad \theta=(\mu,\sigma^2)$$

である．上記の場合，θ の第 2 成分は σ または σ^2 の二通りの母数の表し方があることを示している．つまり，変数変換することに

より母数の表現 (parametrization) はさまざまな場合があることに注意する．さて，この分布 $f(x;\theta)$ に従う n 個の独立な観測データ $x_1,\ldots,x_n$ が得られたとすると，独立性の仮定より同時密度関数は個々の密度関数の積

$$f(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}f(x_i;\theta)$$

で表される．密度関数の値は，Δ_i を微小な値として区間

$$x_1\le X_1\le x_1+\Delta_1,\ x_2\le X_2\le x_2+\Delta_2,\ \ldots,\ x_n\le X_n\le x_n+\Delta_n$$

から観測値 $X_1,\ldots,X_n$ が得られる確率が

$$\begin{aligned}&\Pr\{(X_1,X_2,\ldots,X_n)\in[x_1,x_1+\Delta_1]\times[x_2,x_2+\Delta_2]\times\cdots\times[x_n,x_n+\Delta_n]\}\\&\simeq f(x_1,x_2,\ldots,x_n)\Delta_1\Delta_2\cdots\Delta_n\\&=\prod_{i=1}^{n}f(x_i;\theta)\Delta_i\end{aligned}$$

を意味すると考えればよい．つまり，同時密度関数 $\prod_{i=1}^{n}f(x_i;\theta)$ の値は $x_1,x_2,\ldots,x_n$ という観測データの集合が得られる確率に比例する．

一方，同時密度関数を母数 θ の関数として考えると，観測値 $x_1,x_2,\ldots,x_n$ を生成した確率モデルとしてどの θ が尤もらしいかを知ることができる．

このように観測値 $x_1,x_2,\ldots,x_n$ に対する同時確率密度

$$f(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}f(x_i;\theta)=L(\theta)$$

を母数 θ の関数 $L(\theta)$ として考えたとき，これを「母数の尤もらしさを測る関数」という意味で**尤度関数** *(likelihood function)* と呼ぶ．また，母数 θ での尤度関数の値を，観測値 $x_1,x_2,\ldots,x_n$ における母数 θ の**尤度** *(likelihood)* と呼ぶ．

D.2.2 最尤推定量

最尤推定量は尤度関数を最大にする母数の値として以下のように定義される．

母数 θ の定義域を Θ とする．尤度関数 $L(\theta)$ を最大にする θ の値 $\hat{\theta}$

$$L(\hat{\theta})=\max_{\theta\in\Theta}L(\theta)=\max_{\theta\in\Theta}\prod_{i=1}^{n}f(X_i,\theta)$$

(あるいは以下のように書く場合もある)

$$\hat{\theta}=\arg\max_{\theta\in\Theta}L(\theta)$$

を**最尤推定量** *(maximum likelihood estimator)* という．

尤度関数 $L(\theta)$ の最大値を求めるということは密度関数をいろいろ動かして観測データ $x_1,x_2,\ldots,x_n$ に一番“良く合うところ”を探すことに対応する．この場合“良く合うところ”とは同時密度関数の定義によれば「密度の高さの積が一番大きくなり，観測データが最も出易いところ」ということである．

D.2.3 漸近正規性

標本が十分に大きいとき，緩やかな条件のもとで最尤推定量の標本分布は正規分布に従うことが知られている．この性質を**漸近正規性** *(asymptotic normality)* という．条件の細かな記述は割愛するが，直感的には母数の変化に対して分布が滑らかに変化することが要請される．母数が 1 次元の場合には漸近正規性は

$$\hat{\theta} \sim \mathcal{N}(\theta, 1/nI)$$

のように表される．ただし，θ は母数の真の値であり，I は *Fisher* **情報量** *(Fisher information)* と呼ばれる量で，

$$I = \mathbb{E}\left[\left(\frac{d}{d\theta}\log f(X;\theta)\right)^2\right]$$

で定義される．

区間推定 E

前回は，母集団の持つ何らかの特性量について推定を行う際に，代表的な 1 つの値で推定する点推定について学んだ．データにもとづき計算されたこれらの値は一般に真の値と異なるが，推定量の標本分布を考えることによって推定量にある程度の幅を持たせて推測することができる．これを**区間推定** *(interval estimation)* と呼ぶ．今回は標本分布のところで得られた分布を用いて推定量の区間を構成する方法について学ぶ．

E.1 区間推定

対象とする特徴量の真の値を θ とすると分布がわかっている場合には推定量 $\hat{\theta}$ と真の値の差が適当な区間に入っている確率

$$\Pr\left\{a \leq \hat{\theta} - \theta \leq b\right\} = p$$

を求めることができる．これの意味するところは「ある標本から得られた推定量と真の値の差が区間 (a, b) に入る確率は p である」ということである．

一方，この関係を変形すると

$$\Pr\left\{\hat{\theta} - b \leq \theta \leq \hat{\theta} - a\right\} = p$$

となることがわかる．このとき，区間 $(\hat{\theta}-b, \hat{\theta}-a)$ を信頼水準 p(あるいは $100 \times p\%$) の信頼区間という．上の確率の意味するところは「得られた推定量を用いて構成した区間 $(\hat{\theta}-b, \hat{\theta}-a)$ の中に真の値が入る確率は p である」となるが，言い方を換えると「標本を集めて推定を行うという実験を繰り返したとき，上のようにして構成した信頼区間の中に真の値が入ることは 100 回のうち $100 \times p$ 回である」といった意味合いである．このように点で推定するのではなく，区間で推定する方法を**区間推定** *(interval estimation)* と呼ぶ．

一般には 0.05 とか 0.01 とか小さな値 α を考えて $p = 1 - \alpha$ とし，標本分布の上側 $\alpha/2$-分位点 ($1-\alpha/2$-分位点) と上側 $1-\alpha/2$-分位点 ($\alpha/2$-分位点) を用いて信頼区間を構成する．なお，区間推定の場合，通常とは逆向きの分位点を考え，これを「上側分位点」と呼ぶことがあるので注意する必要がある．

E.1.1 分散が既知の正規母集団の平均

標本平均の分布は母集団の平均と分散に依存するが，母集団の分散 σ^2 が既知であれば正規分布に従う．すなわち標本の大きさを n としたとき

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

である．これを利用して信頼水準 $1-\alpha$ の平均の信頼区間は

$$\left[\bar{X}-\frac{\sigma}{\sqrt{n}}z(\alpha/2),\ \bar{X}-\frac{\sigma}{\sqrt{n}}z(1-\alpha/2)\right] = \left[\bar{X}-\frac{\sigma}{\sqrt{n}}z(\alpha/2),\ \bar{X}+\frac{\sigma}{\sqrt{n}}z(\alpha/2)\right]$$

で与えられる．ただし，$z(q)$ は標準正規分布の上側 q 分位点 ($1-q$ 分位点) である．

E.1.2 分散が未知の正規母集団の平均

分散が未知の場合は，推定した分散で正規化した標本平均を考えると，t-分布に従う．すなわち標本の大きさを n としたとき

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\hat{\sigma}} \sim \mathcal{T}(n-1)$$

である．これを利用して信頼水準 $1-\alpha$ の平均の信頼区間は

$$\left[\bar{X}-\frac{\hat{\sigma}}{\sqrt{n}}t_{n-1}(\alpha/2),\ \bar{X}-\frac{\hat{\sigma}}{\sqrt{n}}t_{n-1}(1-\alpha/2)\right] = \left[\bar{X}-\frac{\hat{\sigma}}{\sqrt{n}}t_{n-1}(\alpha/2),\ \bar{X}+\frac{\hat{\sigma}}{\sqrt{n}}t_{n-1}(\alpha/2)\right]$$

で与えられる．ただし，$t_d(q)$ は自由度 d の t-分布の上側 q 分位点，$\hat{\sigma}^2$ は不偏分散である．

E.1.3 正規母集団の分散

不偏分散の標本分布は自由度の χ^2-分布に従う．

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

これを利用して信頼水準 $1-\alpha$ の分散の区間推定は

$$\left[\frac{(n-1)\hat{\sigma}^2}{\chi_{n-1}(\alpha/2)},\ \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1}(1-\alpha/2)}\right]$$

で与えられる．

E.2 ブートストラップ区間推定

母集団が正規分布でない場合には，対象とする統計量の分布を理論的に調べることが難しいことも多い．ここではブートストラップ法を用いて調べた標本分布を利用して，簡単な信頼区間の構成方法を考えてみよう．

E.2.1 ブートストラップ分位点

前章で取り上げたように，ブートストラップ法を用いると統計量 θ の標本分布の形状を推測することができる．ただし，標本の偏りによって分布は左右に移動していることに注意する．多数のブートストラップ標本に対する統計量 θ^* を生成すると，この擬似的な標本分布の q-分位点 $b(q)$ を求めることができる．

$$\Pr\{\theta^* < b(q)\} = q \;\Leftrightarrow\; \left|\{\theta^* \,|\, \theta^* < b(q)\}\right| = q \times R$$

ただし，$|\cdot|$ は集合の個数を表し，R はブートストラップ標本の個数である．

E.2.2 区間の構成法

もとの標本から求めた統計量 $\hat{\theta}$ とこの分位点を用いていくつかの区間推定の方法が提案されている．

- ブートストラップ法で求めた分位点でそのまま区間を構成するもの
- もとの標本の統計量をブートストラップ標本分布の中心 (平均) と考え，上側 q-分位点を $b(1-q) - \hat{\theta}$ で置き換え区間を構成するもの
- ブートストラップ標本分布を正規分布で近似し，正規分布にもとづく区間を構成するもの

がある，それぞれ percentile 法，basic 法，normal 法と呼ばれている．

$$\begin{aligned}
&\Big[b(\alpha/2),\ b(1-\alpha/2)\Big] && \text{(percentile 法)}\\
&\Big[2\hat{\theta} - b(1-\alpha/2),\ 2\hat{\theta} - b(\alpha/2)\Big] && \text{(basic 法)}\\
&\Big[\hat{\theta} - \sigma^* z(\alpha/2),\ \hat{\theta} + \sigma^* z(\alpha/2)\Big] && \text{(normal 法)}
\end{aligned}$$

ただし，normal 法で用いる σ^* はブートストラップ標本から推定された標準偏差であり，正規分布を仮定した区間推定を行っている．さらに正規分布からのズレを推定して補正を行う BC_a 法などもある．

文献に関する補遺

統計学の教科書は多岐に渡っているので一つを選択するのは難しいが，

「現代数理統計学」，竹村彰通，創文社，1991.

などを薦めておく．いずれも数学的な部分をきっちり書いてあるので内容は重いが，更に進んで数理統計を系統立って学びたい者には役に立つと思う．

また例題が豊富でわかりやすく書いてある入門書的なものとして

「数学セミナーリーディングス 統計学初歩」，国沢清典，数学セミナー，1974.

を挙げておく．

この講義資料は主に

「岩波講座 応用数学 統計的方法」，竹内啓，岩波書店，1994.

を参考にし，講義ではいくつか主要な部分をつまみ喰いした形で紹介した．標準的な教科書とは少々趣きを異にした書き方なので初学者には読み難い部分もあるが，もう少し勉強しようと思う者はこの本も参考にすると良いと思う．

統計的な方法を用いようとしたとき陥りやすい誤りを指摘した

「統計的方法のしくみ – 正しく理解するための 30 の急所」，永田靖，日科技連，1996.

は，実際に手法を適用する場面で役に立つであろう．

また検定の個別の方法については講義では扱い切れないので，簡単な実例とともに様々な検定法を網羅的に扱った

「逆引き統計学 実践統計テスト 100」，Gopal K. Kanji (池谷裕二，久我奈穂子，田栗正章 翻訳)，講談社，2009.

を参考文献として挙げておく．

またこれらの本にはいくつか参考図書も紹介されているので，それらも参考にしてもらいたい．

索引