

2004 年度 卒業論文

R による統計解析計算の 精度保証

提出日：2005 年 2 月 2 日

指導：大石 進一 教授

早稲田大学 理工学部 情報学科

学籍番号：1G01P112-4

吉岡 毅

目次

1	序論	4
1.1	背景	5
1.2	本論文の目的	7
1.3	本論文の構成	7
2	区間演算	8
2.1	浮動小数点数	9
2.1.1	浮動小数点数	9
2.1.2	浮動小数点数と丸め	11
2.1.3	浮動小数点数の性質	12
2.2	区間解析	13
2.3	区間演算	13
2.3.1	四則演算の定義	13
2.3.2	区間演算の性質	15
2.3.3	区間拡張	16
2.4	機械区間演算	17
3	確率分布	19
3.1	離散分布	20
3.1.1	ベルヌーイ分布	20
3.1.2	二項分布	20

3.2	連続分布	21
3.2.1	正規分布	21
3.2.2	Γ 分布	24
3.3	正規分布から導かれる確率分布	24
3.3.1	χ^2 分布	24
3.3.2	F 分布	25
4	推定	27
4.1	母比率の推定と信頼区間	28
4.2	例題計算	29
4.3	考察	32
5	検定	36
5.1	母比率の検定方法	37
5.2	例題計算	38
5.3	考察	39
6	重回帰分析	40
6.1	重回帰計算	41
6.2	回帰式評価	44
6.2.1	重相関係数	45
6.2.2	標準誤差	45
6.3	計算法の原理	46
6.4	例題計算	46
6.4.1	2 変数回帰	47
6.4.2	10 変数回帰	47
6.4.3	データ	49
6.5	考察	51

7	むすび	52
7.1	まとめ	53
7.2	検証および今後の課題	53
	謝辞	55
	参考文献	56

第 1 章

序論

1.1 背景

数値計算が、数学の応用上に重要な意味をもつことはいうまでもない。もちろん、数学とは数値計算のことと思ったり、逆に数値計算などは純粋数学者の関与するところではない、というような誤解ないしは偏見も、ないではなかったようである。もともと数値計算という分野は、計算の精度と手間という相反する要求を調和させようとして生まれてきたものである。理論的に美しい方法が案外役に立たず、実際の問題を解く必要上、やむを得ず考え出された苦し紛れの方法が、かなりの部分を占めていることは事実である。そして、いずれにせよ、実際に計算を実行するには、おびただしい手間と時間が必要であった。対数表の発見、計算機の発明など、そういった苦勞の緩和に大きな役割を果たしたことは言うまでもないが、科学の進歩につれて、計算の勞力は増加する一方であった。

近年になって、電子計算機の異常なまでの急激な進歩により、こうした事情もかなりかわってきた。もちろん電子計算機といえども、限度がある。現存する世界一の計算機でも、まだ十分でない問題もなる。しかし、かつては夢にすぎなかった大計算が、あいついで現実のものとなってきていることも事実である。もはや電子計算機を抜きにした数値計算論は、過去のものとなったといつてよいであろう。

しかしそのために、これまで問題にならなかった新しい種類の問題が発生してきた。事実人間がみればなんでもなく解ける方程式を、機械的に公式にあてはめて電子計算機にかけると、とんでもない答えがでてくる例もいろいろ報告されている。またこれまで実用には不適当だと考えられていた公式や計算法が、新しく電子計算機のためにみなおされた例もある。なぜなら、四則演算の結果はそのつど四捨五入によって丸められ、極限を含む無限演算はすべて有限演算で近似して行われてしますからだ。したがって計算結果は、常にある程度の誤差をともなっている。そこで、計算結果から正しい結論を得るためには、この計算結果に含まれている誤差を評価して真の値の範囲を確定することが必要となる。

これに対し、真の値を含む区間を数とみなす区間解析の概念が導入され、丸め誤差が存在する浮動小数点数を用いても、数学的に正しい結果を数値計算により導く原理が示された。

この区間解析の例を始め、近年能力が大幅に向上してきた計算機を利用することにより、

連続数学の問題の真の解の存在を証明したり、真の解の誤算範囲が保証された精度のよい近似解を得る数値計算を精度保証付き数値計算という。浮動小数点数演算における丸め誤差を含む演算結果の保証が理論的にも実用的にも高い精度で効率よく実現できることが明らかになり、計算の信頼性という問題は、数学としての数値解析の観点からようやく取り上げられることとなった。これは単に丸め誤差を厳密に評価するというだけでなくとどまらず、科学技術計算の元となった数値解析アルゴリズムそのものに影響を与え、さまざまな数理科学上にあらわれる問題の解を、数学的な厳密さで検証する方法にまで展開しつつある。

このように、精度保証付き数値計算は計算の信頼性の立場から見た今後の科学技術の計算法のあるべき一つの方法として考えられようとしている。近い将来、数値計算に精度保証を付加するのが日常的になることも考えられる。いずれにせよ、精度保証付き数値計算の分野は、いま大きな転換期に面しており、新しい飛躍が望まれているように思われる。

1.2 本論文の目的

統計解析計算は、金融や医療など様々な分野で行われている。解析計算を行う際において、一つの方法としてオープンソースの R というフリーソフトによって解析計算を行う。現代の数値解析技術における精度保証付き数値計算の十分たる必要性は、1.1 節に述べた内容などからも明らかである。

本論文では、統計解析計算の結果を精度保証付きで求め、R での解析結果を検証することを目的とする。

まず、解析計算の丸めを考慮して、区間演算を用いて解析計算を行うプログラムを C++ で作成する。そして、その解析結果と R を使って計算した場合の解析結果の比較を行い、誤差解析を行う。解析計算は、推定、検定、重回帰分析について取り上げる。

1.3 本論文の構成

本論文の構成は以下の通りである。

第 2 章「区間演算」では、精度保証付き数値計算の基礎となる、浮動小数点数システム、区間演算、機械区間演算について論じる。

第 3 章「確率分布」では、第 4 章と第 5 章の準備として、そこで用いるそれぞれの確立分布について、密度関数、平均、分散などの分布の特徴を述べる。

第 4 章「推定」では、母比率の推定方法と信頼区間について述べ、例題計算、それに基づいた考察を論じる。

第 5 章「検定」では、母比率の検定計算方法について述べ、例題計算、それに基づいた考察を論じる。

第 6 章「重回帰分析」では、重回帰分析の計算方法と回帰式の評価値について述べ、連立一次方程式の計算法の原理、例題計算、それに基づいた考察を論じる。

第 7 章「むすび」では、本論文のまとめ、および結果の検証を行い、成果・問題点・今後の課題について論じる。

第 2 章

区間演算

2.1 浮動小数点数

現代の数値計算では、浮動小数点数が標準として用いられている。浮動小数点数は現在では標準化が進められ、ほとんどのパソコンやワークステーション、ベクトル計算機、並列計算機などで同じ形式の浮動小数点体系が用いられている。この体系に基づいて、高速に浮動小数点演算が実行される回路がコンピュータに実装されているため、数値計算は浮動小数点体系上で行われることがほとんどである。そこで、このような浮動小数点数の体系の上での計算の精度を保証することが重要である。

そのために、まず浮動小数点数について説明する。浮動小数点数については IEEE 標準 754 (IEEE standard 754, 以下 IEEE 754 と略記する。) がパソコンやワークステーションなどをはじめとして、多くのコンピュータで標準的に用いられている。本論文では、以下 IEEE 754 に基づく 2 進数浮動小数点数システムを考えることにする。

2.1.1 浮動小数点数

IEEE 754 では 4 つのタイプの数を用意されている。それは規格化 2 進浮動小数点数、零、非規格化 2 進浮動小数点数、NaN (Not a Number, 非数) である。

2 進規格化浮動小数点数

2 進規格化浮動小数点数とは

$$a = \pm \left(\frac{1}{2} + \frac{d_2}{2^2} + \frac{d_3}{2^3} + \cdots + \frac{d_N}{2^N} \right) \times 2^e, \quad (d_i = 0 \text{ または } 1). \quad (2.1)$$

と書ける数をいう。 e_{\min} を負の整数, e_{\max} を正の整数として, e は $e_{\min} \leq e \leq e_{\max}$ となる整数である。

$$m = \frac{1}{2} + \frac{d_2}{2^2} + \frac{d_3}{2^3} + \cdots + \frac{d_N}{2^N}. \quad (2.2)$$

を符号付き仮数 (signed mantissa) といい, e を指数 (exponent) という。指数 e も 2 進数で表される。通常, 単精度, 倍精度 (8 byte = 64 bit), 拡張倍精度 (10 byte = 80 bit) の浮動小数点システムがあるが, それぞれつぎのような浮動小数点システムで

ある.

$$N = 24, \quad (-126 \leq e \leq 127),$$

$$N = 53, \quad (-1022 \leq e \leq 1023),$$

$$N = 64, \quad (-16382 \leq e \leq 16383).$$

規格化 2 進浮動小数点システムにおいて表される数の絶対値の最大値は

$$x_{\max} = \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^N} \right) 2^{e_{\max}}. \quad (2.3)$$

であり, その最小値は

$$x_{\min} = \frac{1}{2} 2^{e_{\min}}. \quad (2.4)$$

である. 倍精度数では (2.3), (2.4) はそれぞれ

$$(2 - 2^{52}) \times 2^{1023} = 1.7976931348623157 \cdots \times 10^{308}, \quad (2.5)$$

$$2^{-1023} = 2.22507358507201 \cdots \times 10^{-308}. \quad (2.6)$$

である. $|x| > x_{\max}$ のときにオーバーフロー (overflow) が生じたという.

倍精度浮動小数点数においては, 仮数部が 53bit である.

$$2^{-53} = 1.1102230246 \cdots \times 10^{-16}. \quad (2.7)$$

より, 倍精度浮動小数点数は 10 進数で約 16 桁の精度がある.

零

零は規格化されて

$$+ \left(\frac{0}{2} + \frac{0}{2^2} + \frac{0}{2^3} + \cdots + \frac{0}{2^N} \right) 2^{e_{\min}}. \quad (2.8)$$

と表される.

非規格化 2 進浮動小数点数

IEEE 754 では浮動小数点数は指数部が e_{\min} となったとき, 仮数部の最初の桁が 1 よ

り小さい数を表すために、デフォルトで最初の桁を 1 とすることをやめ、ここが 0 となる数を置くことを許す規格となっている。これを非規格化数 (denormalized number) という。非規格化数の範囲に数が入ることを、漸近アンダーフロー (gradual underflow) という。このような非規格化浮動小数点数の正で最も小さな数は

$$2^{-1074} = 4.94065645841246544 \cdots \times 10^{-324}. \quad (2.9)$$

である。これ以下の数になると、アンダーフロー (underflow) が生じたという。

NaN

このほかに、IEEE 754 ではつぎのような特別な数が用意されている。

- (a) NaN(Not a Number) は $\sqrt{-5}$, $\frac{\infty}{\infty}$, $+\infty + (-\infty)$ など不当な演算の結果として得られる。
- (b) $\pm\infty$ はオーバーフローの結果や零で割った結果として得られる。
- (c) ± 0 はアンダーフローか $\pm\infty$ での割り算の結果として得られる。

2.1.2 浮動小数点数と丸め

IEEE 754 では、つぎの 4 つの丸めのモードが指定できる。 c を実数 ($c \in R$) とする。

上向きの丸め (round upward)

c 以上の浮動小数点数の中で最も小さい数に丸める。これを $\triangle : R \rightarrow F$ と表す。アルゴリズムでの表記は UP とする。

下向きの丸め (round downward)

c 以下の浮動小数点数の中で最も大きい数に丸める。これを $\nabla : R \rightarrow F$ と表す。アルゴリズムでの表記は DOWN とする。

最近点への丸め (round to nearest)

c に最も近い浮動小数点数に丸める。これを $\cdot : R \rightarrow F$ と表す。アルゴリズムでの

表記は NEAR とする. もし, このような点が 2 点ある場合には, 仮数部の最後のビットが偶数である浮動小数点数に丸める. これを偶数丸め方式 (round to even) という.

切り捨て (round toward 0)

絶対値が c 以下の浮動小数点数の中で, c に最も近いものに丸める. アルゴリズムでの表記は ZERO とする.

2.1.3 浮動小数点数の性質

丸めの演算を写像として $\bigcirc : R \rightarrow F$ と書く. すなわち, \bigcirc は \triangle, ∇ , のいずれかと考える. IEEE 754 では, 丸めの演算はつぎの条件を満たす.

$$\begin{aligned}\bigcirc x &= x, & (\text{任意の } x \in F \text{ について}), \\ x \leq y &\rightarrow \bigcirc x \leq \bigcirc y, & (\text{任意の } x, y \in R \text{ について}).\end{aligned}\tag{2.10}$$

また, $x \in F$ のとき, 符号を変えることにより, $-x$ や $|x|$ が得られるので, これらは正確に計算される.

IEEE 754 では, つぎの性質が成立する.

$$\begin{aligned}(-x) &= -\bigcirc x, & (\text{任意の } x \in R \text{ について}), \\ \triangle(-x) &= -\nabla x, & (\text{任意の } x \in R \text{ について}), \\ \nabla(-x) &= -\triangle x, & (\text{任意の } x \in R \text{ について}).\end{aligned}\tag{2.11}$$

IEEE 754 では, 浮動小数点数演算 (F 上での四則演算) は丸めとの関係により, つぎのように定義されている. $\cdot \in \{+, -, \times, /\}$, $\bigcirc \in \{\triangle, \nabla, \text{ } \}$ のとき

$$x \odot y = \bigcirc(x \cdot y), \quad (\text{任意の } x, y \in R \text{ について}).\tag{2.12}$$

この式は, 左辺の浮動小数点数の四則演算の結果 $x \odot y$ は, 右辺の数学的に正しい (実数としての) 四則演算の結果 $x \cdot y$ を指定された丸めを行って得られた数 $\bigcirc(x \cdot y)$ に一致するように計算することを表している.

また, 平方根も

$$(\sqrt{x})_{fp} = \bigcirc(\sqrt{x}), \quad (\text{任意の } x \in F \text{ について}). \quad (2.13)$$

と, 浮動小数点数演算によって計算された平方根 $(\sqrt{x})_{fp}$ は, 正確な実数演算で計算された平方根 \sqrt{x} を指定された丸めの方向へ丸めた数となる. 注意すべきことは, 指数関数や三角関数などはこのような規格を満たしていない. つまり, 初等関数の値を精度保証付きで求めるためには工夫が必要である.

2.2 区間解析

精度保証付き数値計算の原理は, 実数値で与えられる真の解の上限と下限を浮動小数点演算により計算しようとするものであった. このような考え方は区間解析と呼ばれる数学的理論と密接な関係がある. この区間解析についてまず簡単に述べる.

区間解析では区間を数の拡張と考える. そして, その四則演算が定義される. これを区間演算という. 区間演算を浮動小数点システムの上で展開するものを機械区間演算という.

以下, 区間演算について述べていく.

2.3 区間演算

機械区間演算を定義するのが目標であるが, まず, 実数上での演算を仮定する区間演算の概念を説明するここでの議論において四則演算は実数の四則演算とする.

2.3.1 四則演算の定義

区間解析においては, 区間とは

$$[\underline{x}, \overline{x}] = \{x \in \mathbf{R} \mid \underline{x} \leq x \leq \overline{x}\} \quad (2.14)$$

と表される閉区間であるとする．ただし， $\underline{x} \leq \bar{x} \in R$ で，それぞれの区間の下端，上端とする．以下，記号の節約のため，

$$[x] = [\underline{x}, \bar{x}] \quad (2.15)$$

と区間を表すことにし，区間 $[x]$ と単に書けば，それは $[x] = [\underline{x}, \bar{x}]$ を表すものとする． $\underline{x} = \bar{x}$ となる区間 $[x]$ は点区間という．点区間は実数であるので，点区間 $[x]$ の表す実数を x と書くことにする．

区間 $[x]$ について以下を定義する．

$$d([x]) = \bar{x} - \underline{x}, \quad r([x]) = \frac{\bar{x} - \underline{x}}{2}, \quad m([x]) = \frac{\bar{x} + \underline{x}}{2} \quad (2.16)$$

$d([x]), r([x]), m([x])$ をそれぞれ区間 $[x]$ の直径，半径，中心という．

区間 $[x]$ の最小絶対値と最大絶対値をそれぞれ次で定義する．

$$\begin{aligned} \langle [x] \rangle &= \min\{|x| \mid x \in [x]\} \\ |[x]| &= \max\{|x| \mid x \in [x]\} = \max\{|\underline{x}|, |\bar{x}|\} \end{aligned} \quad (2.17)$$

2つの区間 $[x], [y]$ の距離を

$$q([x]) = \max\{|\underline{x} - \underline{y}|, |\bar{x} - \bar{y}|\} \quad (2.18)$$

で定義する．

2つの区間 $[x], [y]$ が与えられたときその二つの区間の四則演算を次で定義する．

$$[x] \circ [y] = \{x \circ y \mid x \in [x], y \in [y]\} \quad (2.19)$$

ただし， $\circ \in \{+, -, \times, /\}$ ．これを区間演算という．この定義では区間演算は無限回の実数演算をしないと実行できないような印象を与えるが，実は次が成立する．

$$\begin{aligned} [x] + [y] &= [\underline{x} + \underline{y}, \bar{x} + \bar{y}] \\ [x] - [y] &= [\underline{x} - \bar{y}, \bar{x} - \underline{y}] \\ [x] \times [y] &= [\min\{\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}\}] \\ [x]/[y] &= [x] \times \left[\frac{1}{\bar{y}}, \frac{1}{\underline{y}}\right], \quad (0 \notin [y]) \end{aligned} \quad (2.20)$$

これから区間演算が有限回の実数演算で実行できることがわかる．かけ算 (Table 2.1) と割算 (Table 2.2) においては，場合分けにより，より少ない手間で計算する次のような計算規則も簡単に導ける．

	$y \geq 0$	$y \ni 0$	$y \leq 0$
$x \geq 0$	$[\underline{xy}, \overline{xy}]$	$[\overline{xy}, \underline{xy}]$	$[\overline{xy}, \underline{xy}]$
$x \ni 0$	$[\underline{x\overline{y}}, \overline{xy}]$	$[\min\{\underline{x\overline{y}}, \overline{xy}\}, \max\{\underline{xy}, \overline{xy}\}]$	$[\overline{xy}, \underline{xy}]$
$x \leq 0$	$[\underline{x\overline{y}}, \overline{xy}]$	$[\underline{x\overline{y}}, \underline{xy}]$	$[\overline{xy}, \underline{xy}]$

Table 2.1: 区間 $[x], [y]$ の乗算 $[x][y]$

	$y \geq 0$	$y \leq 0$
$x \geq 0$	$[\underline{x/\overline{y}}, \overline{x/y}]$	$[\overline{x/\overline{y}}, \underline{x/y}]$
$x \ni 0$	$[\underline{x/y}, \overline{x/y}]$	$[\overline{x/\overline{y}}, \underline{x/y}]$
$x \leq 0$	$[\underline{x/y}, \overline{x/\overline{y}}]$	$[\overline{x/\overline{y}}, \underline{x/y}]$

Table 2.2: 区間 $[x], [y]$ の除算 $[x]/[y]$

2.3.2 区間演算の性質

区間演算については，包含関係における単調性

$$[x] \subseteq [x'], [y] \subseteq [y'] \Rightarrow [x] \circ [y] \subseteq [x'] \circ [y'], \circ \in \{+, -, \times, /\} \quad (2.21)$$

が成立する．

また，加法と乗法に関し交換則と結合則が成立する．

$$\begin{aligned} [x] \circ [y] &= [y] \circ [x], \circ \in \{+, \times\} \\ [x] \circ ([y] \circ [z]) &= ([x] \circ [y]) \circ [z], \circ \in \{+, \times\} \end{aligned} \quad (2.22)$$

しかし，加法と乗法の逆元は存在しない．すなわち， $-[x] = [-\bar{x}, -\underline{x}]$ であるが

$$\begin{aligned} 0 &= [0] \subseteq [x] - [x] = [\underline{x} - \bar{x}, \bar{x} - \underline{x}] \\ 1 &= [1] \subseteq [x]/[x] \end{aligned} \quad (2.23)$$

となることがわかる．上式で等号は $[x]$ が点区間のときのみ成立する．

また，分配則も区間演算に対しては成立しない．そのかわり次の劣分配則が成立する．

$$[x] \times ([y] + [z]) \subseteq [x] \times [y] + [x] \times [z] \quad (2.24)$$

この式で等号は例えば区間 $[y]$ と $[z]$ が同じ符号をもつときに成立する．

2.3.3 区間拡張

関数 $f : D \subset \mathbf{R} \rightarrow \mathbf{R}$ を領域 D の任意の閉区間の上で連続な初等関数とする．関数 f を次のようにして \mathbf{IR} 上の関数に拡張することができる．

$$f([x]) = \{f(x) \mid x \in [x]\} \quad (2.25)$$

\mathbf{IR} を実数上の有界閉区間の集合とする．関数 $f : D \subset \mathbf{R} \rightarrow \mathbf{R}$ が初等関数の合成や四則演算に基づく演算規則で定義されているとき， f を \mathbf{IR} 上の関数に拡張したい．しかし， $f([x])$ を厳密に計算することは非常な計算時間を要することがある．そこで， f の演算規則を単純に区間演算で置き換えた演算規則によって定義される関数を f の区間拡張といい f で表す．

$$f([x]) \subseteq f \ ([x]) \quad (2.26)$$

が成立する．区間拡張は一意的に定まるわけではなく， f を定義する数式を同等な他の表現形式に書き直したとき，それらの表現に基づく区間拡張が一般に異なる評価を与えることに注意する．

上式で等号が成立するとき，厳密な包み込みができたという．厳密な包み込みは一般に難しいが， f の数学的な表現式に区間値パラメータが現れず，また， $[x]$ が一度しかその表

現に現れなければ，その表現に基づく区間拡張が厳密な包み込みを与えることが知られている．一般に式の表現に $[x]$ が少なく現れれば現れるほど，区間包囲は厳密な包み込みに近くなることが多いことが経験されている．区間 $[x]$ の幅 $d([x])$ が小さいとき，

$$q(f([x]), f([x])) \leq \alpha d([x]) \quad (2.27)$$

となることが知られている．ただし， α は正の $[x]$ によらない定数である．

なお，一般に関数の値域 $f([a, b])$ を $f([a, b]) \subset [c, d]$ と区間 $[c, d]$ で評価するとき，区間 $[c, d]$ のことを $f([a, b])$ の区間包囲という．区間拡張は区間包囲の一種である．

一般の整式の区間包囲は区間を変数として与え計算したもの，すなわち区間拡張したものをいえばよい．初等関数のうち e^x や \sqrt{x} などの単調関数に関しても，区間の両端の計算によって区間包囲が求められる．

2.4 機械区間演算

前節で説明した区間演算においては，区間の両端の数は実数であるとし，厳密な実数演算に基づいて区間演算が定義された．しかし，浮動小数点数システム上で数値計算するときには，厳密な区間演算は実行できない．そこで，数学的な正解が含まれる区間が演算の結果得られることを念頭に区間演算で与えられた浮動小数点数システム R 上に展開する方法を説明する．

$$IR = \{[x] \in IR \mid \underline{x}, \bar{x} \in R\} \quad (2.28)$$

とする． IR の要素を機械区間という．機械区間は区間の定義で上端と下端の数が浮動小数点数で与えられるものをいう．区間 $[a, b] \in IR$ を区間 $[c, d] \in IR$ に丸める区間の丸め $\diamond : IR \rightarrow IR$ を

$$[x] \subset \diamond[x] \quad (2.29)$$

および，次を満たすように定義する．

$$\diamond[x] = [x] \text{ (すべての } [x] \in IR \text{ について)}$$

$$\begin{aligned}
[x] \subset [y] &\Rightarrow \diamond[x] \subset \diamond[y] \text{ (すべての } [x], [y] \in IR \text{ について)} \\
\diamond(-[x]) &= -\diamond[x] \text{ (すべての } [x] \in IR \text{ について)}
\end{aligned}
\tag{2.30}$$

IEEE754 上の浮動小数点数システムにおいては

$$\diamond[x] = [\nabla x, \Delta \bar{x}] \tag{2.31}$$

と定義すれば上記の条件は満たされる．機械区間演算の四則演算を

$$[x] \diamond [y] = \diamond([x] \circ [y]) \tag{2.32}$$

と定義する．ただし， $\circ \in \{+, -, \times, /\}$ である．

具体的には機械区間演算は次のように定義される．

$$\begin{aligned}
[x] + [y] &= [\nabla(\underline{x} + \underline{y}), \Delta(\bar{x} + \bar{y})] \\
[x] - [y] &= [\nabla(\underline{x} - \underline{y}), \Delta(\bar{x} - \bar{y})] \\
[x] * [y] &= [\nabla(\min\{\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}\}), \Delta(\max\{\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}\})] \\
[x]/[y] &= [\nabla(\min\{\underline{x}/\underline{y}, \underline{x}/\bar{y}, \bar{x}/\underline{y}, \bar{x}/\bar{y}\}), \Delta(\max\{\underline{x}/\underline{y}, \underline{x}/\bar{y}, \bar{x}/\underline{y}, \bar{x}/\bar{y}\})]
\end{aligned}
\tag{2.33}$$

解を有限回の四則演算の組合せで求める方法(直接解法という)において，浮動小数点数を機械区間に置き換え，四則演算を機械区間演算に置き換えて計算をすれば，真の解を含む区間が得られる．こうして，数値計算により数学的に正しい結論を導き出すために，直接解法が知られている問題については，このような置き換えを行なうことが盛んに行なわれた．ガウスの消去法において，途中の計算における数を機械区間に，四則演算を機械区間演算に置き換えたアルゴリズムを区間ガウス消去法という．区間ガウス消去法はこのようなアルゴリズムの代表的なものである．

区間ガウス消去法では，実行の途中で現れる区間の幅の爆発が生じてしまうことが多い．したがって，区間ガウス消去法は，高々，数十次元の線形方程式の解法としてのみ有効であると考えられている．

第 3 章

確率分布

3.1 離散分布

確率変数 X のとる値が高々加算無限個のとき, すなわち有限または可算無限集合 $D = \{x_k | k = 1, 2, \dots\}$ が存在して $P\{X \in D\} = 1$ となるとき, 確率変数 X あるいは確率分布 P^X は離散型であるという.

3.1.1 ベルヌーイ分布

確率変数 X が, 確率 θ で値 1 を, 確率 $1 - \theta$ で値 0 を, つまり次のような

$$1 \cdots \theta$$

$$0 \cdots 1 - \theta$$

なる確率分布に従う場合, この確率分布をベルヌーイ分布といい $B_N(1, \theta)$ で表す.

$$E(X) = 1 \times \theta + 0 \times (1 - \theta) = \theta \quad (3.1)$$

$$V(X) = (1 - \theta)^2 \times \theta + (0 - \theta)^2 \times (1 - \theta) = \theta(1 - \theta) \quad (3.2)$$

である.

一般に任意の標本空間に対して 1 つの事象 A を考え, A が起これば 1, 起こらなければ 0 をとる確率変数の確率分布はベルヌーイ分布 $B_N(1, \theta)$ となる. ここに $\theta \equiv P(A)$ で成功の確率と呼ばれる.

3.1.2 二項分布

X_1, X_2, \dots, X_n は独立に $B_N(1, \theta)$ に従う確率変数列とする. $X = X_1 + X_2 + \dots + X_n$ とおく. X は成功の回数を表す. X の確率分布は

$$P(X = x) := {}_n C_x \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, 2, \dots, n \quad (3.3)$$

で与えられる. これを 2 項分布といい $B_N(n, \theta)$ で表わす. X の期待値と分散は次のようになる.

$$\begin{aligned}
 E(X) &= \sum_{x=0}^n x_n C_x p^x (1-p)^{n-x} \\
 &= \sum_{x=1}^n \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
 &= np \sum_{x=1}^n \frac{(n-1)!}{(n-1)!\{(n-1)-(x-1)\}!} p^{x-1} (1-p)^{n-1-(x-1)} \\
 &= np \sum_{y=0}^{n-1} C_y p^y (1-p)^{n-1-y} \\
 &= np
 \end{aligned} \tag{3.4}$$

$$\begin{aligned}
 V(X) &= E(X^2) - \{E(X)\}^2 \\
 &= E(X(X-1)) + E(X) - \{E(X)\}^2 \\
 &= n(n-1)p^2 + np - n^2p^2 \\
 &= np(1-p)
 \end{aligned} \tag{3.5}$$

3.2 連続分布

確率変数 X のとる値が連続量と考えられるとき, X または確率分布 P^X を連続型とよび, 分布関数を F_X とすると

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad \forall x \in \mathbf{R}^1 \tag{3.6}$$

をみたす \mathbf{R}^1 上の非負値関数 $f_X(x)$ を X または P^X の確率密度関数とよぶ.

3.2.1 正規分布

密度関数

$$\begin{aligned}
 f(x; \mu, \sigma^2) &:= \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad -\infty < x < \infty \\
 &\quad -\infty < \mu < \infty, \sigma > 0
 \end{aligned} \tag{3.7}$$

で定義される確率分布を正規分布といい、 $N(\mu, \sigma^2)$ で表わす (図 3.1). $\mu = 0, \sigma^2 = 1$ のとき $N(0, 1)$ を標準正規分布という. X が $N(\mu, \sigma^2)$ に従うとき, その標準化 $Z = \frac{X - \mu}{\sigma}$ は $N(0, 1)$ に従う.

$$\begin{aligned} P(Z \leq z) &= P(X \leq \mu + \sigma z) \\ &= \int_{-\infty}^{\mu + \sigma z} \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} dx \\ &= \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{y^2}{2} \right\} dy \end{aligned} \quad (3.8)$$

逆に Z が $N(0, 1)$ に従うと, $X = \mu + \sigma Z$ は $N(\mu, \sigma^2)$ に従う.

定理 3.2.1 (中心極限定理) $X_n | n = 1, 2, \dots$ は独立で同一分布に従う確率変数列で, $E(X_i) = \mu, V(X_i) = \sigma^2 < \infty, i = 1, 2, \dots$ とする. このとき

$$\lim_{n \rightarrow \infty} P \left\{ \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \leq x \right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (3.9)$$

$$\forall x \in \mathbf{R}^1$$

中心極限定理より, 母平均 μ , 母分散 σ^2 をもつ母集団からとった大きさ n の標本 X_1, X_2, \dots, X_n に対して

$$P \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x \right) \xrightarrow{n \rightarrow \infty} \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2}{2} \right\} dx \quad (3.10)$$

が成立する. したがって, 標本平均 \bar{X} の分布は, n が大きいとき, 正規分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ に従うとみなせる. また和 $S_n = \sum_{i=1}^n X_i$ の分布は正規分布 $N(n\mu, n\sigma^2)$ で近似される.

中心極限定理の特別の場合として, 2項分布 $B_N(n, \theta)$ は n が大きいとき, 正規分布 $N(n\theta, n\theta(1-\theta))$ で近似される. なお, このとき

$$P(a \leq X \leq b) \doteq \int_{(a-\frac{1}{2}-n\theta)/\sqrt{n\theta(1-\theta)}}^{(b+\frac{1}{2}-n\theta)/\sqrt{n\theta(1-\theta)}} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2}{2} \right\} dx \quad (3.11)$$

と, $1/2$ を補うと近似がよい. これを連続補正という.

次に, 正規分布 $N(\mu, \sigma^2)$ の平均, 分散は次のようになる.

$$\begin{aligned}
 E(X) &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
 &= \int_{-\infty}^{\infty} \frac{\mu + \sigma z}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \\
 &= \mu + \sigma \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \\
 &= \mu
 \end{aligned} \tag{3.12}$$

$$\begin{aligned}
 V(X) &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
 &= \int_{-\infty}^{\infty} \frac{\sigma^2 z^2}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \left\{ \left[z \left\{ -\exp\left(-\frac{z^2}{2}\right) \right\} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz \right\} \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \times \sqrt{2\pi} \\
 &= \sigma^2
 \end{aligned} \tag{3.13}$$

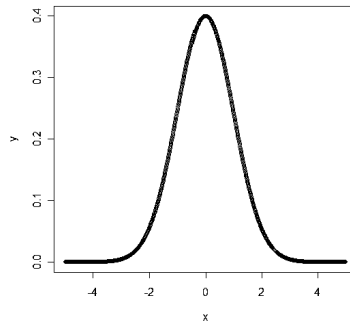


図 3.1: 標準正規分布 $N(0, 1)$, $y = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

3.2.2 Γ 分布

確率変数 X の確率密度関数が

$$f(x; \alpha, \beta) := \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad (3.14)$$
$$\alpha > 0, \beta > 0, x > 0$$

で表されるとき, その確率分布を Γ 分布といい, $\Gamma(\alpha, \beta)$ で表わす (図 3.2).

Γ 分布の平均と分散は

$$\text{平均} = \frac{\alpha}{\beta}, \text{分散} = \frac{\alpha}{\beta^2} \quad (3.15)$$

である.

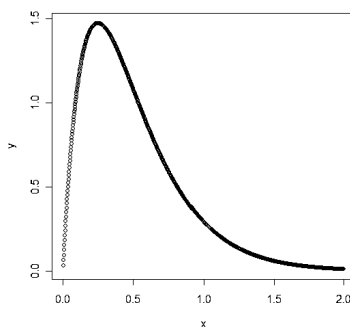


図 3.2: $\Gamma(2, 4)$ の図

3.3 正規分布から導かれる確率分布

3.3.1 χ^2 分布

ガンマ分布 $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, すなわち密度関数が

$$f(x; n) := \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right), x > 0 \quad (3.16)$$

で与えられるとき, この確率分布を自由度 n の χ^2 分布といい, $\chi^2(n)$ で表わす (図 3.3). 自由度 n の χ^2 分布の上側 α 点を $\chi^2(\alpha; n)$ で表わす.

自由度 n の χ^2 分布の平均と分散は,

$$\text{平均} = n, \text{分散} = 2n \quad (3.17)$$

である.

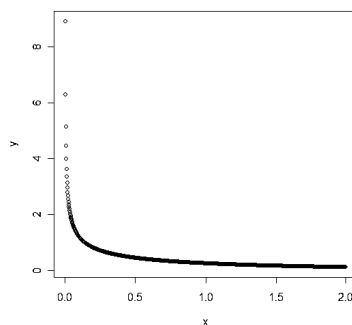


図 3.3: 自由度 1 の χ^2 分布

3.3.2 F 分布

密度関数が,

$$f(x; m, n) := \frac{\Gamma\left(\frac{m+n}{2}\right) m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, x > 0 \quad (3.18)$$

$$m, n = 1, 2, \dots$$

で定義される確率分布を自由度 (m, n) の F 分布といい, $F(m, n)$ で表わす (図 3.4).

平均と分散は次のようになる.

$$\text{平均} = \frac{n}{n-2} \quad (n > 2) \quad (3.19)$$

$$\text{分散} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (n > 4) \quad (3.20)$$

である.

自由度 (m, n) の F 分布の上側 α 点を $F(\alpha; m, n)$ で表わす. F 分布には

$$F(1 - \alpha; n, m) = \frac{1}{F(\alpha; m, n)} \quad (3.21)$$

という関係がある.

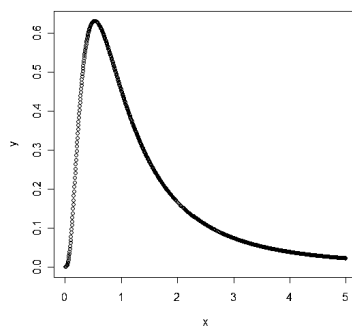


図 3.4: $F(10, 4)$ の図

第 4 章

推定

4.1 母比率の推定と信頼区間

リンカーン法 (Capture-Mark-Recapture) という、生態学で、野原のバツタの数を調べたいときに全数を調べるわけにはいかないの、捕まえてペンキでマークして暫く経ってからまた捕まえてマークされているバツタの割合を求めて、マークした数をそれで割って総数を推定する方法があり、これと同様のことを行う。しかし、この方法で得られる母比率の推定値の信頼性は高くない。このことは、確率が二項分布に従うこと考えれば分かる。

ここで、次のような問題を考える。最初に混入した黒い石の数が 40 個、かき混ぜてから 20 個の石を取り出してみたら黒石 2 個、白いし 18 個だった場合、元の白石の数はいくつと推定されるか？

元の白石の数を x とすると、 $40/(40 + x) = 2/(2 + 18)$ となるので、これを x について解けば、 $x = 360$ が得られる。したがって 360 個と推定される。黒石の母比率が p である容器から 20 個の石を取り出したときに、黒石がちょうど 2 個である確率は、 ${}_{20}C_2 p^2 (1 - p)^{18}$ の二項分布に従う。図 4.1 を見れば分かるように、 $p = 0.1$ で最大であるが、 $p = 0.09$ だろうが $p = 0.11$ であろうが、黒石がちょうど 2 個である確率には大した差はない。だから 360 個という点推定値は、404 個や 324 個に比べて、それほど信頼性は高くない。

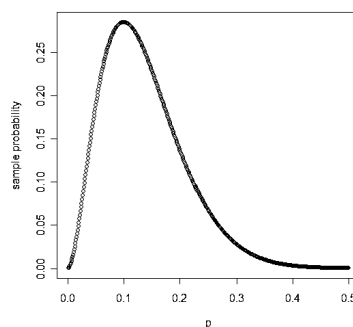


図 4.1: 黒石が 20 個中に 2 個である確立

そこで、ある程度の信頼性が見込める範囲を示すことを考える。この幅のことは「信頼区間」という。二項分布は、 n が大きいときには正規分布で近似できる。母比率 p 、標本サイズ n

で, その中で注目している属性をもつ要素の数を X , 観測比率を $p' = X/n$ とすれば, X が近似的に正規分布 $N(np, np(1-p))$ に従うことになる. 正規分布の 95% のサンプルは, 平均 \pm 標準偏差 $\times 1.96$ に含まれるので,

$$Pr \left[-1.96 \leq \frac{(X - np)}{\sqrt{np(1-p)}} \leq 1.96 \right] = 0.95 \quad (4.1)$$

$$Pr \left[p' - 1.96\sqrt{\frac{p'(1-p')}{n}} \leq p \leq p' + 1.96\sqrt{\frac{p'(1-p')}{n}} \right] = 0.95 \quad (4.2)$$

となり, 母比率 p は 95 % の確率で

$$\left[p' - 1.96\sqrt{\frac{p'(1-p')}{n}}, p' + 1.96\sqrt{\frac{p'(1-p')}{n}} \right] \quad (4.3)$$

の範囲にあるといえる. これが, 母比率 p の 95% 信頼区間となる.

しかし, 分布に歪みやずれのある場合には, 2 項分布を正規分布に近似することができない. その時は, 2 項分布を F 分布に近似することによって信頼区間を求める.

$v_1 = 2(n - X + 1)$, $v_2 = 2r$ を自由度とする F 分布で, 上側確率が 0.975 となる値を F とする. $v'_1 = 2(X + 1)$, $v'_2 = 2(n - X)$ を自由度とする F 分布で, 上側確率が 0.975 となる値を F' とする. 2 項分布を F 分布に近似した場合の信頼区間は

$$\left[\frac{v_2}{v_1 F + v_2}, \frac{v'_1 F'}{v'_1 F' + v'_2} \right] \quad (4.4)$$

となる.

また,

$$\left[0, 1 - \alpha^{\frac{1}{n}} \right] \quad (p' = 0) \quad (4.5)$$

$$\left[\alpha^{\frac{1}{n}}, 1 \right] \quad (p' = 1) \quad (4.6)$$

と定義する.

4.2 例題計算

ここで, 以下の例題について, R と C++ の区間演算プログラムで解析計算を行う.

例題 1

ある番組を関東地区の調査対象の 600 世帯中,137 世帯が見ていた. 関東地区の視聴率はいくつになるか?

点推定値は, $137/300 = 0.228$, つまり 22.8% である.

2 項分布を正規分布に近似した場合の 95% 信頼区間の計算結果は以下ようになる. R statistics (Up) は R での上限値の計算結果,R statistics (Down) は R での下限値の計算結果, C++(Up) は区間演算での上限値の計算結果,C++(Down) は区間演算での下限値の計算結果である.

正規分布近似計算結果

```
R statistics (Up)      = 0.2619210108577094687
R statistics (Down)    = 0.1947456558089571965
C++(Up)               = [0.2619210108577094132,0.2619210108577095242]
C++(Down)             = [0.1947456558089571965,0.1947456558089572521]
```

以上のように,R と区間演算によって求めた信頼区間は

$$R \text{ 計算} = [0.1947456558089571965, \quad 0.2619210108577094687] \quad (4.7)$$

$$\text{区間演算計算} = [0.1947456558089571965, \quad 0.2619210108577095242] \quad (4.8)$$

であり, 区間幅は,

$$R \text{ 計算} = 0.0671753550487522722 \quad (4.9)$$

$$\text{区間演算計算} = 0.0671753550487523277 \quad (4.10)$$

となっている.

区間演算した計算結果は上限値・下限値共に 16 桁目以下で違いが起こっている. よって, 例題の問題では R での計算結果は 15 桁目までの精度が保証され, R の計算結果を信頼してよいと考えることができる.

しかしここで, 信頼区間は 0.5 を含んでおらず, 分布にずれがある. この場合, 2 項分布を正規分布に近似するより F 分布に近似して計算を行うべきである (図 4.2). R において, 用意されている関数 `binom.test` を用いれば, `binom.test(137,600)` を実行することによって F 分布近似計算を行った計算結果を得る.

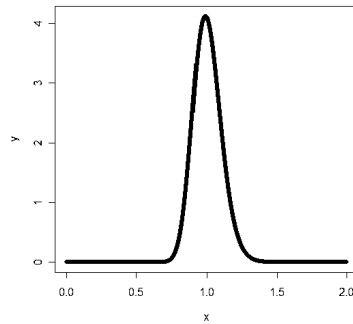


図 4.2: $F(928, 274)$ の図

F 分布近似を行った場合の R と区間演算の計算結果は, 以下のようになった.

F 分布近似計算結果

```
R statistic (Up)      = 0.26404693955446934
R statistic (Down)    = 0.19531920136561551
C++(Up)              = [0.2640469395544692289, 0.2640469395544693954]
C++(Down)            = [0.1953192013656154802, 0.1953192013656155357]
```

以上のように, R と区間演算によって求まった信頼区間は

$$R \text{ 計算} = [0.19531920136561551, \quad 0.26404693955446934] \quad (4.11)$$

$$\text{区間演算計算} = [0.1953192013656154802, \quad 0.2640469395544693954] \quad (4.12)$$

であり, 区間幅は,

$$R \text{ 計算} = 0.06872773818885383 \quad (4.13)$$

$$\text{区間演算計算} = 0.0687277381888539152 \quad (4.14)$$

となっている.

以上のように, 正規分布近似での計算結果と同様に, 区間演算した計算結果は上限値・下限値共に 16 桁目以下で違いが起こっている. よって, 例題の問題では R での計算結果は 15 桁目までの精度が保証され, R の計算結果を信頼してよいと考えることができる.

また、正規分布近似と F 分布近似の計算結果を比較すると、小数点以下 3 桁目で違いが起きている。信頼区間幅は F 分布近似を行った方が大きい。

4.3 考察

以上のように、母比率の信頼区間計算において丸め誤差の影響は少なく、R での解析計算を信頼することができると思われる。

ここでは、信頼区間を計算する際、浮動小数点数の丸め誤差が大きく影響するような場合が存在するかどうかについて、詳しくみていく。

例題において、正規近似の信頼区間幅は 0.067 であり、一般的に標本サイズ n を大きくすると、区間幅は小さくなる。よって、

- n が大きいときには、信頼区間の上限値と下限値が丸め誤差の影響を大きく受けて、考慮すべきではないのか

という仮説について考える。

信頼区間幅は、

$$3.92 \times \sqrt{\frac{p'(1-p')}{n}} \quad (4.15)$$

である。

関数 $g(p')$ を $g(p') = p'(1-p')$ として、まず観測比率 $p' = 0.5$ と固定する。よって、信頼区間幅 $f(n)$ は $f(n) = 3.92 \times \sqrt{0.25/n}$ となる。

以下に、標本サイズ n を x 軸に、 $f(n)$ を y 軸にとった図を示し (図 4.3, 4.4)、一部の値を表にまとめた (Table 4.1)。

次に、観測比率 p' に対する信頼区間幅 $f(p')$ の推移について考える。標本サイズ $n = 100000$ と固定する。よって、 $f(p') = 3.92 \times \sqrt{p'(1-p')/100000}$ となる。

以下に、 p' を x 軸に、 $f(p')$ を y 軸にとった図を示し (図 4.5)、一部の値を表にまとめた (Table 4.2)。

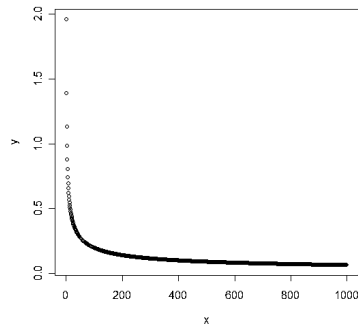


図 4.3: $0 < n < 1000$ の図

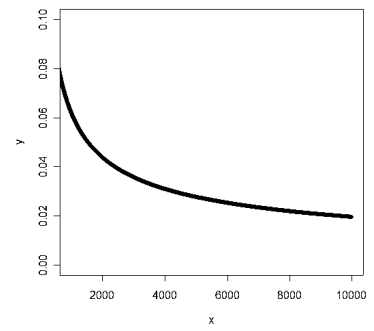


図 4.4: $1000 < n < 10000$ の図

標本サイズ n	信頼区間幅 $f(n)$
1000	0.0619
10000	0.0196
100000	0.00619

Table 4.1: $p' = 0.5$ の時の信頼区間幅推移

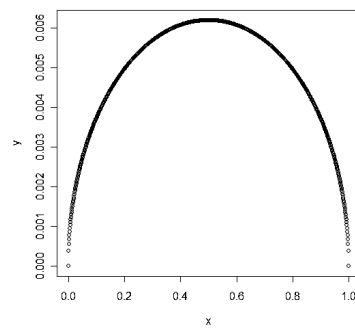


図 4.5: $0 < p' < 1$ に対する信頼区間幅の変化

観測比率 p'	信頼区間幅 $f(p')$
0.01	0.00124
0.001	0.000392
0.0001	0.000124

Table 4.2: $n = 100000$ の時の信頼区間幅推移

以上のことから, 標本サイズが 10 万で観測比率が 0.01% であるような統計においても, 丸め誤差の影響は受けないことが分かる.

実際に, 分布のずれを考えて 2 項分布を F 分布に近似して, $n = 100000, p' = 0.0001$ の場合の信頼区間を R と区間演算で計算すると,

$$R \text{ 計算} = [0.0000479548, \quad 0.0001838958] \quad (4.16)$$

$$\text{区間演算計算} = [0.0000479548, \quad 0.0001838958] \quad (4.17)$$

となっており, 丸め誤差の影響はない. また, 信頼区間幅は正規近似する場合よりも大きい.

よって, 「 n が大きいときには, 信頼区間の上限値と下限値が丸め誤差の影響を大きく受けて, 考慮すべきではないのか」という仮説を棄却することができる.

次に,

- $p' \div 1.96\sqrt{p'(1-p')/n}$ の時, 信頼区間の下限値 $\div 0$ になる. この時, 丸め誤差の影響を大きく受けて, 考慮すべきではないのか

という仮説を考える.

$p' \div 1.96\sqrt{p'(1-p')/n}$ の時, すなわち $p' \div 3.84/(n+3.84)$ の時, 信頼区間は

$$\left[\frac{3.84 - 1.96\sqrt{3.84}}{n + 3.84}, \frac{3.84 + 1.96\sqrt{3.84}}{n + 3.84} \right] \quad (4.18)$$

となり, 下限値は必ず負の値になってしまう. 仮説の場合, 分布の歪みが大きいために正規近似できない. よって, 仮説は棄却される.

仮説の場合を F 分布近似によって R で求めてみると, $n = 10000$, $X = 4$ の時, $[1.089969e-04, 1.0238393e-03]$. $n = 100000$, $X = 4$ の時, $[1.089876e-05, 1.024127e-04]$ となっている.

最後に,

- 上限値 $\div 1$, 又は, 下限値 $\div 0$ のような時, 丸め誤差の影響を大きく受けて, 考慮すべきではないのか

という仮説を考える.

仮説の場合において, 分布はずれや歪みがあるため, 2 項分布は F 分布に近似する. F 分布で上側確率が 0.975 となる F 値の大きさは, 第 1 自由度 $= \infty$, 第 2 自由度 $= 2$ の時に $F = 39.498$, 第 1 自由度 $= 2$, 第 2 自由度 $= \infty$ の時に $F = 3.689$ となっている. よって, (4.4) から, 丸め誤差が影響することはないと考えることができる.

以上のことから, 母比率の 95% 信頼区間を求める際は, 丸め誤差の影響は考慮しなくてよいと考えられる.

第 5 章

検定

5.1 母比率の検定方法

予め母比率について何らかの期待があるとき、標本から推定された母比率がそれと違っていないかを調べたい、ということが起こる。こういう場合の基本的な考え方としては、標本データの度数分布が、母集団について期待される分布と一致するという仮説（帰無仮説）が成り立っている確率を調べて、それが普通では考えられないほど小さい場合（通常は5%未満）に、滅多にないことだから偶然ではない、と考えて帰無仮説を棄却する。

カテゴリ数が全部で n 個あるとき、 i 番目のカテゴリの観測度数が O_i 、期待度数が E_i であるとき、 χ^2 値が、自由度 $n - 1$ のカイ二乗分布に従うことを検定する。

$$\chi^2 := \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (5.1)$$

一般に、 χ^2 が自由度 $n - 1$ のカイ二乗分布の 95% 点よりも大きいときは、統計的に有意であるとみなして、帰無仮説を棄却する。観測された分布が期待される分布と一致している可能性は5%もない場合は、有意水準5%で棄却するということである。自由度とは、標本の数から前もって母数の数を引いた値である。つまり、

$$\chi^2(0.05; n - 1) < \chi^2 \quad (5.2)$$

の時、仮説を棄却する。

有意確率は、自由度を k とすると

$$1 - \int_0^{\chi^2} \frac{1}{2\Gamma\left(\frac{k}{2}\right)} \left(\frac{x}{2}\right)^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right) dx \quad (5.3)$$

によって求められる。

自由度が1($k = 1$)の場合、密度関数は

$$f(x; 1) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \exp\left(-\frac{x}{2}\right) \quad (5.4)$$

であるから、有意確率は

$$1 - \int_0^{\chi^2} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \exp\left(-\frac{x}{2}\right) dx \quad (5.5)$$

である。

5.2 例題計算

ここで、以下の例題について、R と C++ の区間演算プログラムで解析計算を行う。

例題 2

ある病院で生まれた子ども 900 人中、男児は 480 人であった。このデータから、(1) 男女の生まれる比率は半々であるという仮説、(2) 男児 1.06 に対して女児 1 という割合で生まれるという仮説、は指示されるか？

R では、用意されている関数の χ^2 分布の累積確率密度関数 `pchisq` を用いて、有意確率を求めることができる。つまり、 X を χ^2 値、 k を自由度とすると、 $1 - \text{pchisq}(X, k)$ が有意確率となる。

それぞれの仮説 (1)(2) に対して、 χ^2 検定を R と C++ で区間演算を行った結果は以下のようになった。R カイ二乗値は R で計算した χ^2 値、R result は R で計算した有意確率、C++ カイ二乗値は区間演算によって計算した χ^2 値、C++ result は区間演算によって計算した有意確率である。

仮説 (1)

```
R カイ二乗値 = 4.0
R result     = 0.045500263896358528
C++ カイ二乗値 = [4, 4]
C++ result   = [0.04550026389630357216, 0.04550026389635974944]
```

仮説 (2)

```
R カイ二乗値 = 1.26943396226415110
R result     = 0.25987289504626154
C++ カイ二乗値 = [1.269433962264142402, 1.269433962264155725]
C++ result   = [0.2598728950462008136, 0.2598728950462696475]
```

以上のように、仮説 (1) において、区間演算を行った有意確率の計算結果は小数点以下 14 桁目で違いが起きており、R での有意確率の計算結果は 13 桁の精度があると言える。有意確率は 0.05 よりも小さいため、仮説 (1) は棄却される。

同様に、仮説 (2) において、区間演算を行った χ^2 値と有意確率の計算結果は小数点以下 14 桁目で違いが起きており、R での計算結果は 13 桁の精度があると言える。有意確率は 0.05 よりも大きいため、仮説 (2) は採択される。

よって、例題における χ^2 値と有意確率の計算は、R の計算結果を信頼することができる
と考えることができる。

5.3 考察

理想的な検定は、第 1 種の誤りの確率と同時に、第 2 種の誤りを犯す確率も小さいものである。しかしながら、第 2 種の誤りを小さくする検定は、普通第 1 種の誤りの確率を大きくし、逆も成立する。有限個のランダムなデータに基づいて判断を下さざるを得ない理論的限界が背景にある。この二律背反的状况に対して統計学が取る立場は、一度に両方を小さくするのが無理ならせめて、第 1 種の誤りの確率だけでも、確実に有意水準以下にしておこうというものである。よって、理論的限界の結果として、検定は、帰無仮説が棄却された時にのみ、意味を持つ。

そこで、母比率の χ^2 検定計算を区間演算で行った時に、

- 有意確率が $[0.04 \dots, 0.05 \dots]$ のように、計算結果の区間が 0.050 を含み、上限値と下限値で有意判定が異なってくる場合はどのような時か

について考える。

自由度が 1 の場合、このようなことは χ^2 値が

$$[3.8414588206923, \quad 3.8414588206942] \quad (5.6)$$

の時に起こり、この χ^2 値の区間幅は $1.9\text{e-}12$ であり極めて小さい。

自由度が 1 以外でも同様に、0.050 を含んだ区間演算の結果が得られる χ^2 値の区間幅は小さい。よって、無視することができる。と考えることができる。

以上より、 χ^2 検定によって母比率の仮説の有意判定を行う際、丸め誤差の影響はないと考えることができる。

第 6 章

重回帰分析

6.1 重回帰計算

回帰分析においては、 y と x_1, \dots, x_p の間に

$$y = E(y) + \varepsilon = \eta + \varepsilon = f(x_1, \dots, x_p) + \varepsilon \quad (6.1)$$

というモデルを想定する.

$$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (6.2)$$

で表される場合、このモデルを重回帰モデルという. 誤差 ε は互いに独立に、平均 0、分散 σ^2 の正規分布に従う.

説明変数の値が $(x_{\alpha 1}, x_{\alpha 2}, \dots, x_{\alpha p})$ のときの y_α の予測値として

$$\hat{y}_\alpha = \hat{\beta}_0 + \hat{\beta}_1 x_{\alpha 1} + \dots + \hat{\beta}_i x_{\alpha i} + \dots + \hat{\beta}_p x_{\alpha p} \quad (6.3)$$

が得られる.

また、モデル式を変形すると

$$y_\alpha = \beta'_0 + \beta_1(x_{\alpha 1} - \bar{x}_1) + \dots + \beta_p(x_{\alpha p} - \bar{x}_p) + \varepsilon_\alpha \quad (\alpha = 1, \dots, n) \quad (6.4)$$

となる. β'_0 は

$$\beta'_0 = \beta_0 + \beta_1 \bar{x}_1 + \dots + \beta_p \bar{x}_p \quad (6.5)$$

に対応している.

目的変数ベクトル y , 説明変数行列 X , 誤差ベクトル ε , 回帰母数ベクトル β とすると、モデルは

$$y = X\beta + \varepsilon \quad (6.6)$$

と表せ,

$$\begin{cases} E(\boldsymbol{\varepsilon}) = \mathbf{o} \\ V(\boldsymbol{\varepsilon}) = \sigma^2 I \\ E(\mathbf{y}) = X\boldsymbol{\beta} \\ V(\mathbf{y}) = \sigma^2 I \end{cases} \quad (6.7)$$

となる.

実測値 y_α との差

$$e_\alpha = y_\alpha - \hat{y}_\alpha \quad (6.8)$$

を残差という. 残差平方和

$$\begin{aligned} S_e &= \sum_{\alpha}^n e_\alpha^2 \\ &= \sum_{\alpha}^n (y_\alpha - \hat{\beta}_0 - \hat{\beta}_1 x_{\alpha 1} - \cdots - \hat{\beta}_i x_{\alpha i} - \cdots - \hat{\beta}_p x_{\alpha p})^2 \end{aligned} \quad (6.9)$$

が最小となるように, $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ を定める.

ここで S_e を $\hat{\beta}_0$ で偏微分する.

$$\frac{\partial S_e}{\partial \hat{\beta}_0} = 2 \sum_{\alpha}^n (y_\alpha - \hat{\beta}_0 - \hat{\beta}_1 x_{\alpha 1} - \cdots - \hat{\beta}_p x_{\alpha p})(-1) = 0 \quad (6.10)$$

これを整理して, n で割ると

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \cdots - \hat{\beta}_p \bar{x}_p \quad (6.11)$$

が得られる.

以上のことをベクトルを用いて表すと, 残差ベクトル \mathbf{e} は

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad (6.12)$$

と表せる. 残差平方和 S_e は

$$S_e = \mathbf{e}'\mathbf{e} = (\mathbf{y} - X\hat{\boldsymbol{\beta}})'(\mathbf{y} - X\hat{\boldsymbol{\beta}}) \quad (6.13)$$

と表せるから,

$$\begin{aligned}
\frac{\partial S_e}{\partial \hat{\beta}} &= \mathbf{o} \\
-2X'(\mathbf{y} - X\hat{\beta}) &= \mathbf{o} \\
X'X\hat{\beta} &= X'\mathbf{y} \\
\hat{\beta} &= (X'X)^{-1}X'\mathbf{y}
\end{aligned} \tag{6.14}$$

となる.

また, β_0 の推定値 $\hat{\beta}_0$ は, \bar{x} と $\hat{\beta}_1$ を

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}, \hat{\beta}_1 = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \tag{6.15}$$

とすると,

$$\begin{aligned}
\hat{\beta}_0 &= \hat{\beta}'_0 - \hat{\beta}_1\bar{x}_1 - \cdots - \hat{\beta}_p\bar{x}_p \\
&= \bar{y} - \bar{x}'\hat{\beta}_1
\end{aligned} \tag{6.16}$$

となる.

各変数を標準正規分布に従うように基準化する. n 組のデータについての y および各 x_i の標準偏差を

$$\begin{cases} s_y = \sqrt{S_{yy}/(n-1)} \\ s_i = \sqrt{S_{ii}/(n-1)} \end{cases} \quad (i = 1, \dots, p) \tag{6.17}$$

で表すと, 基準化した変数は

$$\begin{cases} y' = \frac{y - \bar{y}}{s_y} \\ x'_i = \frac{x_i - \bar{x}_i}{s_i} \end{cases} \quad (i = 1, \dots, p) \tag{6.18}$$

と書くことができる.

この y' の x'_1, \dots, x'_p に対する推定回帰式を

$$\hat{\eta}' = \hat{\beta}_0' + \hat{\beta}_1' x'_1 + \dots + \hat{\beta}_p' x'_p \quad (6.19)$$

とすれば, $\hat{\beta}_i'$ ($i = 0, \dots, p$) は単位に無関係な標準偏回帰係数となる. $\bar{y}' = \bar{x}_i' = 0$ ($i = 1, \dots, p$) より, 常に $\hat{\beta}_0' = 0$ である.

標準偏回帰係数と普通の偏回帰係数とのあいだには,

$$\hat{\beta}_i = \hat{\beta}_i' \sqrt{\frac{S_{yy}}{S_{ii}}} \quad (i = 1, \dots, p) \quad (6.20)$$

の関係が成り立つ. ここで,

$$\begin{cases} S_{xx} = \sum (x_\alpha - \bar{x})^2 \\ S_{yy} = \sum (y_\alpha - \bar{y})^2 \\ S_{xy} = \sum (x_\alpha - \bar{x})(y_\alpha - \bar{y}) \\ S_{ij} = \sum_\alpha^n (x_{\alpha i} - \bar{x}_i)(x_{\alpha j} - \bar{x}_j) & (i, j = 1, \dots, p) \\ S_{iy} = \sum_\alpha^n (x_{\alpha i} - \bar{x}_i)(y_\alpha - \bar{y}) & (i = 1, \dots, p) \end{cases} \quad (6.21)$$

である.

また, x_i と x_j ($i, j = 1, \dots, p$), あるいは, x_i ($i = 1, \dots, p$) と y の相関係数を

$$\begin{cases} r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} & (i, j = 1, \dots, p) \\ r_{iy} = \frac{S_{iy}}{\sqrt{S_{ii}S_{yy}}} & (i = 1, \dots, p) \end{cases} \quad (6.22)$$

と定義すると, 相関係数行列によって $\hat{\beta}_1', \hat{\beta}_2', \dots, \hat{\beta}_p'$ を求めることができる. よって, 回帰式が決定される.

6.2 回帰式評価

予測した回帰式は, 重相関係数や標準誤差によって評価することができる. 以下に, 重相関係数と標準偏差について説明する.

6.2.1 重相関係数

目的変数 y_α の偏差平方和 $S_T (= S_{yy})$ は, 残差平方和 S_e , 回帰平方和 S_R で表すと,

$$S_T = S_R + S_e \quad (6.23)$$

と書ける. ここで S_R は,

$$S_R = \sum_i^p \hat{\beta}_i S_{iy} = \hat{\beta}_1 S_{1y} + \cdots + \hat{\beta}_p S_{py} \quad (6.24)$$

である.

y_α と予測値 \hat{y}_α の一致の度合いは相関係数によって知ることができる. y と \hat{y} の偏差積和 $S_{y\hat{y}}$ は S_R と同値である.

したがって, 重相関係数 R は

$$R = \frac{S_R}{\sqrt{S_{yy}S_R}} = \sqrt{\frac{S_R}{S_T}} = \sqrt{1 - \frac{S_e}{S_T}} \quad (6.25)$$

である. R^2 は決定係数と呼ばれ, 決定係数が 1 に近いほど当てはまりがよい.

自由度調整済み重相関係数 R^* は,

$$R^* = \sqrt{1 - \frac{S_e/(n-p-1)}{S_T/(n-1)}} \quad (6.26)$$

である. よって, 次式が成り立つ.

$$R^{*2} = R^2 - \frac{p}{n-p-1}(1-R^2) \quad (6.27)$$

6.2.2 標準誤差

標準誤差 σ の推定値は

$$\hat{\sigma} = \sqrt{\frac{1}{n-p} \hat{\epsilon}' \hat{\epsilon}} \quad (6.28)$$

である.

6.3 計算法の原理

重回帰分析のプログラムは, 精度よく掃き出し法を行うために相関係数を用いる. 説明変数の相関係数行列の最終列に目的変数の相関係数を足した行列 X に対して, ガウス・ジョルダン法によって標準偏回帰係数を求める. 行列 X は,

$$X = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} & r_{1y} \\ r_{21} & 1 & \cdots & r_{2p} & r_{2y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 & r_{py} \end{bmatrix} \quad (6.29)$$

で表せる.

まず, (1,1) 要素を枢軸として掃き出し, 次々と (p,p) 要素を枢軸とする掃き出しを実行する. すると標準偏回帰係数 $\hat{\beta}'_1, \hat{\beta}'_2, \dots, \hat{\beta}'_p$ が求まる.

$$X = \begin{bmatrix} 1 & 0 & \cdots & 0 & \hat{\beta}'_1 \\ 0 & 1 & \cdots & 0 & \hat{\beta}'_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \hat{\beta}'_p \end{bmatrix} \quad (6.30)$$

そして, 求まった標準偏回帰係数から回归母数を算出し, 回帰式が求まる.

6.4 例題計算

R においては, 用意されている関数 `lm` によって, 回帰式を求めることができる. また, 関数 `summary` を用いることによって, 回帰式の決定係数や標準誤差などを求めることができる.

以下, 2 つのデータに対して R と C++ の区間演算プログラムで回帰分析を行い, 結果をまとめた. 例題計算として用いた 2 つのデータは, 6.4.3 にまとめてある.

6.4.1 2変数回帰

まず, データ 1 (Table 6.1) について重回帰計算を行う. 身長を第 1 説明変数, 胸囲を第 2 説明変数とし, 体重を目的変数として計算した.R と C++ で区間演算を行った計算結果は, それぞれ以下のようになった.

R 計算

```
定数項    = -82.75193096972550677
第 1 変数  =  0.3461760603811828707
第 2 変数  =  0.9584087708018101948
決定係数   =  0.728972808000567074
標準誤差   =  3.742814385798410637
```

区間演算計算

```
定数項    = [-82.75193096974162188, -82.75193096970940588]
第 1 変数  = [0.3461760603811178117, 0.3461760603812485404]
第 2 変数  = [0.9584087708017544616, 0.9584087708018651509]
決定係数   = [0.7289728080005025702, 0.7289728080006315781]
標準誤差   = [3.742814385769855257, 3.742814385826960244]
```

以上のように, 定数項, 第 1 変数, 第 2 変数, 決定係数は, 区間演算の計算結果では上限値と下限値に 13 桁目以下で違いが起きており, 12 桁の精度であることが分かる. また, 標準誤差は 10 桁の精度である.

よって, データ 1 に対する重回帰解析の結果においては, 丸め誤差の影響はわずかであり, R の解析結果を信頼することができると考えることができる..

6.4.2 10 変数回帰

次に, データ 2 (Table 6.2) について重回帰計算を行う. 第一銘柄から第十銘柄をそれぞれ第 1 から第 10 説明変数とし, 目的変数を TOPIX として計算した.R と C++ で区間演算を行った計算結果は以下のようになった.

R 計算

```
定数項      =91.35466974978476174
第 1 変数    =0.4404082158305101524
第 2 変数    =0.3203327332297490338
第 3 変数    =0.005852338432796612248
第 4 変数    =0.1182360619203007557
第 5 変数    =0.2001831759784096498
第 6 変数    =0.2764294531974435532
第 7 変数    =0.2446665284817033104
第 8 変数    =-0.045854467439678262
第 9 変数    =0.06195105911416100286
第 10 変数   =0.1411876191520624324
決定係数     =0.9827938522381355346
標準誤差     =3.26912598142985722
```

区間演算計算

```
定数項      =[91.35466846876897762,91.35467103080668494]
第 1 変数    =[0.4404082149597670082,0.4404082167012435823]
第 2 変数    =[0.3203327330925113126,0.3203327333669850341]
第 3 変数    =[0.005852338384030813558,0.00585233848156242395]
第 4 変数    =[0.1182360617049353913,0.1182360621356628866]
第 5 変数    =[0.2001831755363172039,0.2001831764204875242]
第 6 変数    =[0.2764294531297597502,0.2764294532651274672]
第 7 変数    =[0.2446665284640607286,0.2446665284993468359]
第 8 変数    =[ -0.04585446748280332802, -0.04585446739654526483]
第 9 変数    =[0.06195105903550109789,0.06195105919284085716]
第 10 変数   =[0.1411876191317758267,0.1411876191723443752]
決定係数     =[0.982793851460310064,0.982793853015959118]
標準誤差     =[3.269123096472783541,3.26912886638835154]
```

以上のように、偏回帰係数の中で一番精度が高いものは 10 桁であり、一番精度が低いものは 8 桁である。定数項の精度は 6 桁、決定係数の精度は 8 桁、標準誤差の精度は 6 桁となっている。データ 1 に対して回帰解析を行った計算結果と比較すると、計算結果の精度がかなり落ちているのが分かる。

回帰式の評価値である決定係数よりも定数項の方が精度が低くなっているのは、定義式によるものである。決定係数の算出には、定数項の値は用いないからである。一方、標準誤差の算出には、定数項の値も用いられている。よって、回帰式の評価は、標準誤差によって評価するのがよい。

以上のことから、回帰式を区間演算によって求めた標準誤差の値は、6 桁目以降に違いが起きているため、丸め誤差の影響は大きくないと考えることができる。よって、R の計算結果を信頼することができると思われる。

6.4.3 データ

6.4.1 と 6.4.2 の例題計算において使用したデータは, 以下の Table 6.1, 6.2 の通りである.

- データ 1 は, 25 人の身長, 胸囲, 体重の測定結果である.
- データ 2 は, 22 日間の株価 10 銘柄と TOPIX の値である.

身長	胸囲	体重	身長	胸囲	体重
171.000000	83.000000	62.000000	172.000000	90.000000	63.000000
180.000000	98.000000	71.000000	175.000000	85.000000	62.000000
169.000000	83.000000	53.000000	163.000000	93.000000	65.000000
173.000000	100.000000	78.000000	163.000000	85.000000	53.000000
173.000000	90.000000	59.000000	162.000000	85.000000	52.000000
163.000000	82.000000	54.000000	167.000000	88.000000	60.000000
168.000000	92.000000	62.000000	170.000000	93.000000	67.000000
176.000000	90.000000	60.000000	173.000000	94.000000	70.000000
173.000000	89.000000	62.000000	170.000000	92.000000	67.000000
174.000000	88.000000	60.000000	170.000000	87.000000	56.000000
164.000000	84.000000	50.000000	175.000000	100.000000	68.000000
171.000000	87.000000	66.000000	170.000000	85.000000	56.000000
172.000000	90.000000	69.000000			

Table 6.1: データ 1

第一	第二	第三	第四	第五	第六	第七	第八	第九	第十	TOPIX
693	530	1100	699	352	950	692	1310	698	3470	1631.71
704	539	1090	714	354	957	687	1320	706	3490	1645.48
693	518	1080	699	346	944	685	1290	699	3430	1615.94
691	520	1140	685	338	927	675	1270	696	3400	1602.33
695	550	1200	700	344	925	680	1250	700	3420	1622.29
679	550	1160	695	335	926	671	1190	703	3350	1608.10
678	599	1130	697	342	925	665	1220	703	3370	1616.32
666	600	1130	695	344	920	660	1200	698	3330	1605.49
668	595	1130	699	345	927	666	1210	691	3330	1616.51
678	600	1170	691	345	930	670	1220	698	3340	1619.90
687	610	1190	697	347	950	676	1230	705	3350	1639.87
681	609	1160	700	354	954	708	1240	709	3350	1642.94
679	610	1120	714	350	957	725	1220	706	3370	1653.27
667	605	1210	715	360	970	720	1220	709	3370	1648.53
663	596	1280	724	354	972	717	1210	707	3370	1647.24
662	597	1250	725	350	969	700	1210	696	3290	1628.01
662	588	1200	727	342	950	683	1210	687	3260	1614.42
659	600	1250	730	342	952	700	1200	680	3240	1614.91
660	591	1290	740	339	958	699	1180	680	3210	1610.35
667	594	1270	711	336	966	685	1180	675	3240	1612.84
665	589	1270	695	332	950	685	1160	665	3200	1596.51
655	580	1340	675	330	935	693	1170	660	3200	1588.22

Table 6.2: データ 2

6.5 考察

以上のように,重回帰分析計算において,偏回帰係数は2変数回帰では12桁の精度で計算され,10変数回帰では8桁の精度で計算されている。また,定数項,決定係数や標準誤差の精度は,偏回帰係数の精度と比較して低くなっている。この原因はそれぞれの値の計算式や定義式から明らかであり,偏回帰係数に含まれている誤差が伝播しているためである。

Rにおいて,回帰母数を求めるために連立一次方程式を解く際,高精度な計算結果を得るために相関行列を利用して計算する。行列中の要素の値の大きさが極端に違うと,数値計算上不利なためである。

ここで,

- 変数の数の増加に伴って,精度はどのように変化するか?

ということについて考える。

変数の数が13であるデータに対して重回帰計算を行なうと,偏回帰係数は6桁の精度,定数項は6桁の精度,標準誤差は4桁の精度で計算された。Rにおいては,表示桁数は指定をしなければ7桁で表示される。よって,変数の数が13以上であるとき,Rの回帰母数の計算結果は丸め誤差の影響を受けていると考えることができる。

よって,変数の数が13以下である場合の重回帰分析を行うとき,Rの回帰母数の計算結果は信頼することができ,丸め誤差の影響は受けていない。

第 7 章

むすび

7.1 まとめ

第2章では、精度保証付き数値計算を論じる上で欠かせない、浮動小数点数、区間演算や機械区間演算について述べた。

第3章では、統計の推定や検定において、確率の近似分布として用いられる様々な確率分布について述べた。

第4章では、母比率の推定と信頼区間について取り上げた。例題を用いながら解析計算結果の丸め誤差解析を区間演算によって行った結果、母比率の推定や信頼区間の解析計算を行う際、丸め誤差の影響は考慮しなくてよいということに帰着した。

第5章では、母比率の検定について取り上げた。例題を用いながら解析計算結果の丸め誤差解析を区間演算によって行った結果、母比率の帰無仮説の有意判定を行う際、丸め誤差の影響は考慮しなくてよいということに帰着した。

第6章では、重回帰分析について取り上げた。例題を用いながら解析計算結果の丸め誤差解析を区間演算によって行った結果、変数の数が13変数以下であるような場合、Rの解析結果を信頼することができるということに帰着した。

7.2 検証および今後の課題

20世紀末、コンピュータの進歩とともに統計学の理論や技術も大きく進展し、同時にパッケージソフトウェアの開発によって、統計学の専門家でなくても統計解析を行うことが可能になった。そのため、統計学は、いまやすべての自然科学や社会科学で適用される科学的分析の技術となっている。統計学をもっとも広い意味で定義するならば、「不確実性を考慮した論理的推論」ということになるだろう。

本論文は、Rを使った統計解析計算結果の精度保証を行うことを目的とした。検証の結果、母比率の推定・信頼区間・検定、重回帰分析においては、コンピュータによる計算誤差の影響は無視できるという結論に帰着した。

計算誤差が少なかった理由として、

- モデル化した確率分布や計算式を用いることによって, 実際の解析計算の計算量が少ない
- 極端に大きさの異なる数値どうしの計算が表れない, 又はそのように計算方法を工夫している

ということが挙げられる.

今後の課題として, 本論文では扱うことができなかった時系列解析や解析計算量が多いフィッシャーの直接確率など, 他の解析計算においても同様に検証することである.

謝辞

この卒業論文の完成にあたってたくさんの方々から多くの助言や指導を頂きました。まず、本研究を進めるに当たり、終始丁寧な御指導及び御激励を賜り、最後まで面倒を見て下さいました大石 進一教授に深く感謝いたします。

大変お忙しい中でも私の質問に対して熱心に答えて下さいました、理工学研究科客員講師、丸山 晃佐氏、荻田 武史氏、中谷 祐介氏に大変感謝いたします。

研究室内や合宿など、日常的な場面でお世話になりました、大石研究室博士課程 1 年尾崎 克久氏、大石研究室修士課程 2 年、大山 博毅氏、森山 敦史氏、並びに大石研究室修士課程 1 年、坂内 太郎氏、島本 誠氏、関本 竜平氏、徳永 克久氏、福地 健氏、細田 幸裕氏、横山 大輔氏に深く感謝いたします。

研究内容だけでなくあらゆる場面で意見の交換や協力などをして下さいました、大石研究室 4 年、有滝 貴広氏、岩田 揚平氏、大成 高顕氏、川崎 文裕氏、栗原 崇氏、佐藤 友規氏、平野 晃氏、ト 詣各氏、水島 裕氏、吉田 昂央氏には様々な助言をもらい大変感謝しています。

参考文献

- [1] 大石進一:精度保証付き数値計算, コロナ社 (2001,11).
- [2] 鈴木武, 山田作太郎:数理統計学, 内田老鶴圃 (2001.6).
- [3] 久米均, 飯塚悦功:回帰分析, 岩波書店 (1994,3).
- [4] R-project:<http://www.r-project.org/>
- [5] CRAN:<http://cran.r-project.org/>
- [6] 間瀬茂, 神保雅一, 鎌倉稔成, 金藤浩司:工学のためのデータサイエンス入門 - フリーな統計環境 R を用いたデータ解析 -, 数理工学社 (2004.3).
- [7] 中澤港著:R による統計解析の基礎, ピアソンエデュケーション (2003.10).