

# 医学統計学演習：資料 6

芳賀昭弘 \*

## 1 信頼区間の推定

本日は次の例題に沿って「信頼区間の推定」について学びます。

### 例題 1：

日本人の成人男性の  $1\mu\text{l}$  あたりの白血球数の平均値が 6300 で標準偏差が 1700 の正規分布をしているとします。ある病気の患者 16 名のデータを取り出して  $1\mu\text{l}$  あたりの白血球数の平均を求めたところ、その数は 7200 であった。この病気の患者の白血球数は、日本人の成人男性の平均と差異があると言えるでしょうか？

### 例題 2：

0 から 10000 番まで番号がついたアンケートを配布した。そのうち 100 枚のアンケートが回収され、番号の平均値が 4135 であった。回収されたアンケートは小さい番号に偏っていると言えるでしょうか？

前回学んだ中心極限定理は、標本抽出を繰り返すことでその平均値が、正規分布を用いて解析できることを示しています。しかし、実際のほとんどの問題において標本抽出を繰り返すことなどできない（しない）でしょう。そんなことが出来るのであれば、母集団全体のデータを調べられるのですから、わざわざ標本から母集団を抽出する必要はありません。実際の状況は、母集団のうち（なんとかして）その一部のデータを手に入れたという状況でしょう。例題 1 と例題 2 でも、得られている結果は一回の標本抽出の結果であり、例題 1 ではある病気の患者 16 名の白血球数の平均、例題 2 では回収された 100 枚のアンケートの番号の平均、がそれぞれの（たった一回の）標本抽出の結果です。手に入れた一回の標本抽出データから、母集団の性質をどうやって調べるのでしょうか？

それは、中心極限定理を別角度から眺めることで実現できることがわかります。標本抽出を繰り返したデータの平均値の分布から母集団の平均値（真の平均値）を得るのではなく、たった 1 回の標本抽出されたデータの平均値  $\bar{x}$  から、真の平均値を推定することを考えるのです。中心極限定理の言っていることは、前節で確かめたように  $\bar{x}$  は真の平均値  $\mu$  の周りに図 1 左のように正規分布に従っているはず、ということです。ここで  $n$  は標本抽出した数のこと（サンプル数）で、十分に大きな数とします（故に母集団の標準偏差  $\sigma$  はこの標本の標準偏差で代用可能とする）。 $\bar{x}$  はこのような確率分布を持って様々な値を取ることができます。1 回だけのサンプリングなので、たまたまこの 95% の区間から外れるような  $\bar{x}$  の値を取ることもあって良いでしょう。このときは、 $\bar{x}$  は真の平均値  $\mu$  から  $1.96\sigma/\sqrt{n}$  以上異なってしまうことを意味します。よって、たった 1 回のデータの平均値  $\bar{x}$  から真の平均値  $\mu$  を推定するのであれば、95% 信頼区間（ここで 95% のこ

\* Electronic address: [haga@tokushima-u.ac.jp](mailto:haga@tokushima-u.ac.jp)

とを信頼係数という)を導入して、

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}, \quad (1)$$

の範囲に真の平均値  $\mu$  が入ってくると考えます。  $n$  個のデータをサンプルして平均を求める作業を 100 回行くと、95 回はこの範囲に真の平均値  $\mu$  が来るということです (図 1 右)。変数を標準化する  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$  で表すと、

$$-1.96 \leq z \leq 1.96, \quad (2)$$

となります。-1.96, 1.96 は、両側 5% (下側と上側それぞれ 2.5%) を除く下限値と上限値で、R では `qnorm(0.025,0,1)` で下限値、`qnorm(0.025,0,1, lower.tail=F)` で上限値を求めることができます。信頼区間の範囲は、標本抽出を 100 回行くと何回この範囲に入ってくるかで制御でき、95 回ではなく 99 回 (100 回に 1 回は推定が外れる) 入ってくる場合には、信頼係数を 99% とした信頼区間 (**99% 信頼区間**) は、 $\pm 2.58$  (`qnorm(0.005,0,1)` と `qnorm(0.005,0,1, lower.tail=F)`) に変えることで得ることができます\*1。

$$\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}, \quad (3)$$

$$-2.58 \leq z \leq 2.58. \quad (4)$$

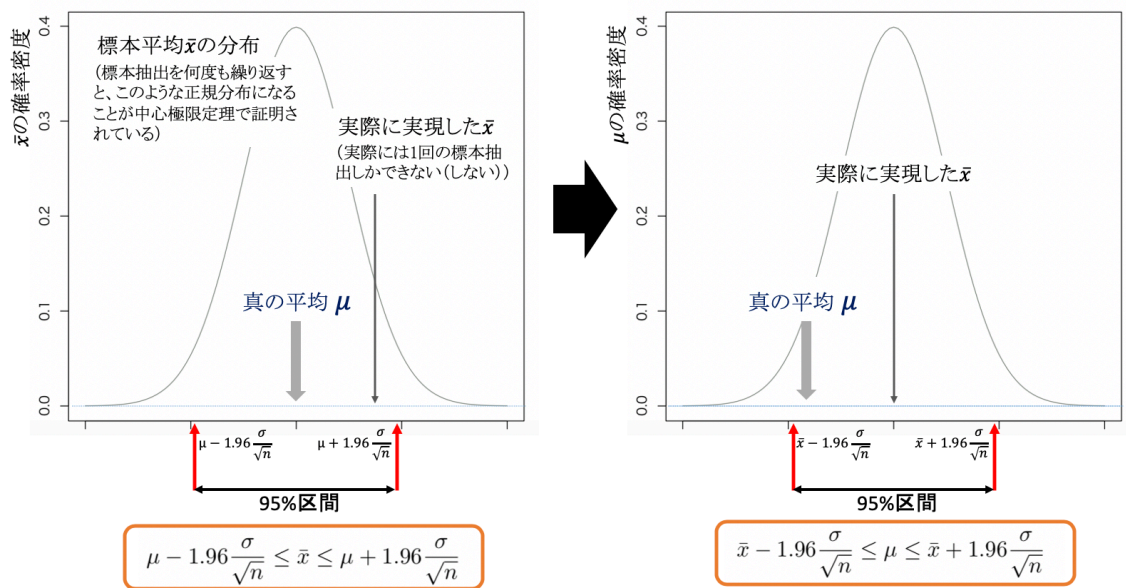
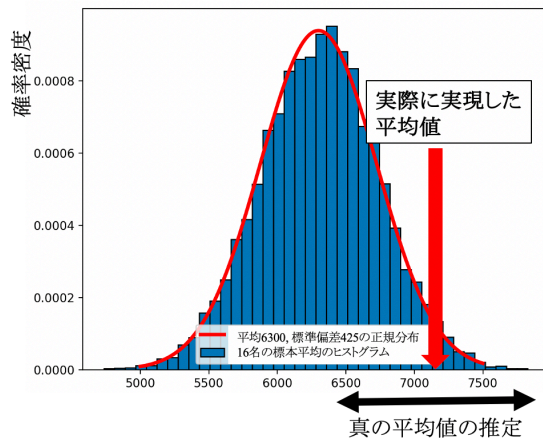


図 1 標本抽出されたデータの平均  $\bar{x}$  の分布 (左) と推定された真の平均値の分布 (右) : 中心極限定理より、標本抽出されたデータの平均値の分布は左のような正規分布となります。そのうちの 1 つが実現された (平均値  $\bar{x}$ ) 場合、真の平均値は逆に実現された平均値の周りに正規分布の確率で存在すると推定することが出来ます。

では、具体的に例題 1 と例題 2 で 95% 信頼区間と 99% 信頼区間を出してみましょう。前回行ったことを思い出しましょう。「もし仮に母集団から標本抽出を繰り返したらどうなるのか?」ということコンピュータ上でシミュレーションしました (MonteCarlo1.R を実行してみましょう)。標本抽出を繰り返したら、その平

\*1 1.96 や 2.58 は、確率分布が正規分布のとき (正確にいうと一次元の正規分布) に使うことができる定数です。標準正規分布をそれぞれ  $[-1.96, +1.96]$ 、 $[-2.58, +2.58]$  の範囲で確率変数を積分することで 0.95, 0.99 となります。

## 例題1



## 例題2

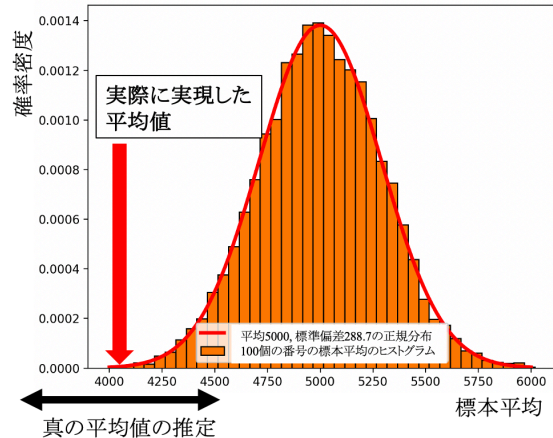


図2 実際に実現された標本の平均値（赤矢印）。図1右のように、実現した値の周りに真の平均値があると考えて黒矢印の範囲を具体的に求めたものが信頼区間であり、95% 信頼区間は式 (1) で、99% 信頼区間は式 (3) で求めることができる。

均値は図2の赤の曲線で示される正規分布になるのが中心極限定理で、確かにそうなるということを示したわけですが、実際に得られた標本の平均はこれらの分布のどのあたりだったのか赤矢印で示したのが図2になります。ランダムに標本抽出されていれば、その標本の平均は分布の中心付近に来る確率が高いはずですが、特に例題2の方では分布の裾野付近が実現されてしまいました。これを今度は図1の右側のように、この実現した平均値の周りに真の分布の平均値がくると視点を変えたのが黒の矢印です。例題1と例題2におけるこの黒矢印の具体的な範囲は、95% 信頼区間に対して式 (1) より、

$$6367 \leq \mu \leq 8033 \quad (\text{例題 1}),$$

$$3569 \leq \mu \leq 4701 \quad (\text{例題 2}),$$

となります。同様に 99% 信頼区間に対して式 (3) より、

$$6104 \leq \mu \leq 8297 \quad (\text{例題 1}),$$

$$3390 \leq \mu \leq 4880 \quad (\text{例題 2}),$$

です。これが「たった一回の標本抽出で求めた標本平均から推定される真の（母集団の）平均値」です。例題1では、真の平均値は6300でした。これは95% 信頼区間に入っていますが、99% 信頼区間からは外れています。例題2の真の平均値5000は、95% 信頼区間にも99% 信頼区間にも入っていません。これをどう解釈すればよいのでしょうか？実は、これにより「標本がランダムにサンプルされているのかどうか」を判定します。ランダムにサンプリングされたのであれば真の平均値に近い値が出るはずなのに、かなり外れた標本平均が得られたということは、何か特別な抽出をしてしまったのではないかと考えるわけです。例題1では、「ある病気」の特定の方を抽出しています。「ある病気」の方だけの白血球数を集めて母集団を形成していれば、その平均値は健康な人と合わせた母集団とは異なる性質（平均値や分散など）を持っている可能性が示唆されたことになります。同じく例題2でも100枚のアンケート番号の平均は異常に低い値であって、どうもランダムにサンプリングされたものではなさそうであることが示唆されています。これで例題1（この病気の患者の白血球数は、日本人の成人男性の平均と差異があると言えるでしょうか？）と例題2（回収されたアンケート番号が若い番号に偏っていると言えるでしょうか？）の問いに対して何らかの回答ができそうです。例えば、例題1

では差異がある確率が高い、例題 2 では若い番号に偏っている確率がかなり高い、などと表現できそうです。このように、何 % の信頼区間に入っているかといった定量値で例題 1 や 2 の問いに回答を与える方法が次に述べる統計的検定になります。次回、ここで行なってきた手順を少し形式張った形で整理したいと思います。

なお、ここまで正規分布でのお話でしたが、確率変数が質的変数であったり連続変数でも正規分布とは異なる確率分布であった場合でも、同じように標本抽出されたデータが実現する確率がそれぞれ 95%, 99% である範囲を 95% 信頼区間や 99% 信頼区間を定義することができます。後の節で  $t$  分布という確率分布が出てきますが、それは**自由度**と呼ばれる変数に応じて信頼区間が異なります。例えば自由度 10 の  $t$  分布の場合、95% 信頼区間は  $[-2.23, +2, 23]$  (`qt(0.025,10)`, `qt(0.025,10,lower.tail=F)`) です\*<sup>2</sup>。信頼区間の考え方がわかってしまえば、他の確率分布への応用も難しくないと思います。中心極限定理があるにも関わらず、正規分布以外の分布への応用も考えなければならない理由は、少数データで且つ母集団が正規分布に従うとは言えない状況があるからです。その時は次週以降に学ぶ  $t$  分布を使うことになります。

## 2 演習

1. 平均 10、標準偏差が 2 の正規分布の 95% 信頼区間と 99% 信頼区間を求めよ。
2. 平均 0、自由度が 7 の  $t$  分布の 95% 信頼区間と 99% 信頼区間を求めよ。
3. 自由度が 1 の  $\chi^2$  分布の下側 95% 信頼区間と 99% 信頼区間を求めよ。
4. 例題 1 で、患者数が 25 名で他は同じ場合の 95% 信頼区間を求めよ。
5. 上記の問題で、実際に 25 名の平均値を求めたところその結果が 5300 だった。この結果をもとに、母集団の平均値の信頼区間を推定せよ。
6. 例題 2 で、50 枚取り出して番号の平均値を求めた時の 95% 信頼区間を求めよ。
7. 上記の問題で、一回だけ 50 枚取り出して番号の平均値を求めたところその結果が 4700 だった。この結果をもとに、母集団の平均値の信頼区間を推定せよ。

各問題で使った R の関数と結果を明示し、解説を付記してください。

---

\*<sup>2</sup> 自由度は標本抽出する数  $n$  と関係しています。この  $n$  が十分大きいとき  $t$  分布は標準正規分布と一致（当然、信頼区間も一致）します。