2022 年度秋学期　統計学　第１３回

不確かな測定の不確かさを測る — 不偏分散と t 分布

母平均がわからないのに，母分散がわかっているのはおかしい

前回の講義で，母集団の度数分布（母集団分布）が正規分布であるときに，母集団分布の平均（母平均）の区間推定を行う方法を説明しました。このとき，母集団分布の分散（母分散）があらかじめわかっているものとして，推定の方法を説明しました。

しかし, 母集団分布の平均が不明なのに母集団分布の分散がわかっているというのは, どう考えてもヘンです。母平均が不明ならば，母分散も不明なのが普通でしょう。そこで今回は，母分散も不明なときに，標本から計算される値である「不偏分散」を用いて母平均の推定を行う方法と，そのために用いる「t 分布」という確率分布について説明します。

不偏分散

標本から求めた分散を使う

上で述べたように「母平均が未知なのに母分散が既知」というのは現実にはありえないことで，実際には母平均が未知なら母分散も未知のはずです。つまり「不確かな測定は，その不確かさも不確か」というわけです。

そこで，未知の母分散のかわりに，標本をから求めた分散を使って，母平均を推測することを考えます。分散は「(各データの，期待値（平均）からのへだたり）の２乗の，そのまた期待値（平均)」ですから，これに対応して「(各標本の，標本平均からのへだたり）の２乗の，そのまた平均」を考えます。これを**不偏分散**（不偏標本分散）といい，標本サイズを n，標本を $X_1, X_2, \ldots, X_n$，標本平均を $\bar{X}$ とするとき，不偏分散 s^2 は

$$s^2 = \frac{1}{n-1}\left\{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \cdots + (X_n-\bar{X})^2\right\} \tag{1}$$

となります。標本サイズの n そのものではなく，$n-1$ **で割る**ことに注意してください。

なぜ $n-1$ で割るのか？

不偏分散は，その期待値が母分散に等しくなるように調整された分散です[1]。一言でいえば，「不偏」とは「ひいきをしない」という意味です。同じ母集団から何度もくりかえし標本を取り出して，そのつど不偏分散の値を計算したとすると，取り出される標本は毎回異なるので，不偏分散の値も毎回違います。毎回違いますが，その期待値は母分散と同じです。つまり，不偏分散は，母分散より大きくも小さくも「平等に」外れており，母分散よりたいていいつも大きいとか，たいていいつも小さいということはない，ということです。

なぜ，n ではなく $n-1$ で割るのでしょうか？ それを直観的に理解するために，図１をみてみましょう。図１では標本サイズを２（標本として２つしか数値をとらない）とします。母分散は，各データと

[1] このことを，「不偏分散は母分散の不偏推定量である」といいます。

母平均とのへだたりの 2 乗の平均です。これに対して，標本サイズ n（この場合は 2）で割った標本分散は，標本と標本平均とのへだたりの 2 乗の平均になっています。

標本の 2 つの数値が，どちらも母平均よりも大きく，偏った方向にへだたっていても，標本平均はつねに 2 つの標本の中間にあります。ですから，「標本と標本平均とのへだたり」は「標本と母平均とのへだたり」よりは小さくなります。この違いを調整するために，n ではなく $n-1$ で割っているのです[2]。

上のことを，数式を使って見てみましょう。サイズ 2 の標本を X_1, X_2，標本平均を $\bar{X}$ とするとき，不偏分散 s^2 は

$$s^2 = \frac{1}{2-1}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\right\} \tag{2}$$

となります。{ } 内は 2 つの「へだたり」の 2 乗の和に見えますが，

$$\bar{X} = \frac{X_1 + X_2}{2} \tag{3}$$

ですから，これを (2) 式に代入すると

$$\begin{aligned} s^2 &= \frac{1}{2-1}\left\{(X_1 - \frac{X_1+X_2}{2})^2 + (X_2 - \frac{X_1+X_2}{2})^2\right\} \\ &= \frac{1}{2-1}\left\{\left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2\right\} \\ &= \frac{1}{2-1}\left\{\frac{(X_1 - X_2)^2}{2}\right\} \end{aligned} \tag{4}$$

となり，実際にはひとつの「へだたり」でしかありません。そのため，不偏分散を求めるときにも，2 で割るのではなく $(2-1)=1$ で割っています。もともとは，2 つの変数 X_1, X_2 は自由にいろいろな値をとることができ，それぞれと母平均とのへだたりも別々にあるものです。それが，標本平均とのへだたりを考えるときは，へだたりが 1 つに減ってしまっています。このことを，「自由度」が 2 から 1 に減ったといいます。

また，図 2 は，「標本と標本平均とのへだたり」が，たいてい「標本と母平均とのへだたり」よりも小さくなる，ということを説明する別の例です。ヒストグラムは母集団分布（実際には不明）を表していて，黒い丸は抽出された標本を表しています。(a) のように，標本が母集団分布からまんべんなく選ばれていれば，母平均（μ）と標本平均（$\bar{X}$）はあまり変わりませんから，「標本と標本平均とのへだたり」は「標本と母平均とのへだたり」とほぼ同じです。しかし，(b) のように，標本が偏って選ばれているときは，「標本と標本平均（$\bar{X}$）とのへだたり」は「標本と母平均（μ）とのへだたり」よりも小さくなります。

t 分布と区間推定

母分散がわかっているときは（前回の講義）

前回の例で，母集団分布が母平均 μ，母分散 σ^2 の正規分布で，そこから n 個の標本を取り出したときの標本平均が $\bar{X}$ であるとき，

[2] なぜ $n-1$ で調整するのかは，私の「解析応用」(2013 年度秋学期）第 12 回で説明しています。

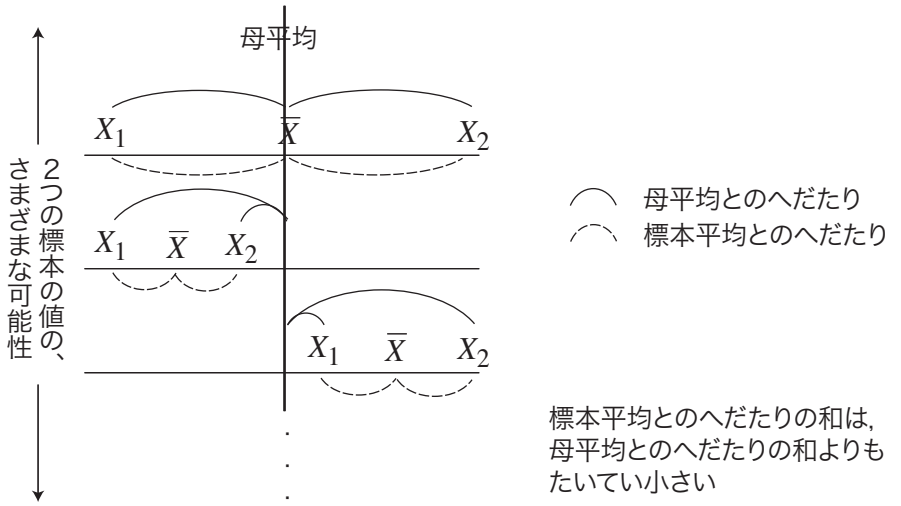

図 1: なぜ $n-1$ で割るのか？

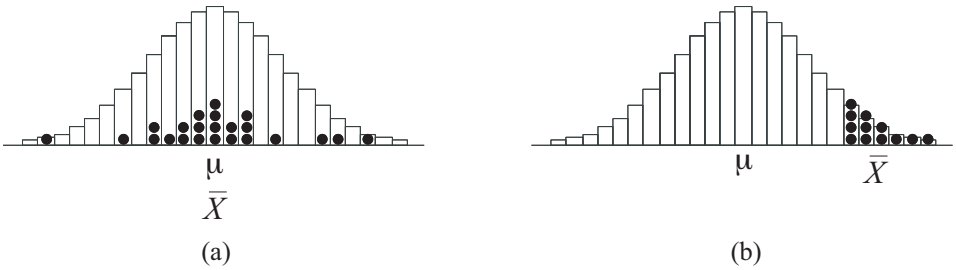

図 2: なぜ $n-1$ で割るのか？（別の説明）

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \tag{5}$$

とおくと，Z は標準正規分布 $N(0,1)$ にしたがうことを説明しました。これまでの例では，Z のこの性質を用いて，母平均 μ の区間推定を行いました。

母分散が不明のときは

では，母分散 σ^2 が不明であるとしましょう。このとき，(5) 式には μ と σ^2 の 2 つの未知の量があるので，μ の区間推定ができません。そこで，母分散 σ^2 を，標本から計算される不偏分散 s^2 でおきかえた

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \tag{6}$$

というものを考えます。この t を t **統計量**といいます。Z は標準正規分布にしたがいますが，t はどのような分布にしたがうでしょうか？

スチューデントの t 分布

この t 統計量がしたがう確率分布は，標準正規分布ではなく，**自由度 $n-1$ の t 分布**（スチューデントの t 分布）という確率分布で，これを $t(n-1)$ と書きます。t 分布の確率密度関数は標準正規分布とよく似ており，$t=0$ を中心とした左右対称の形になっています。

問題例と解き方

t 分布を用いると，母分散が不明の場合でも，標準正規分布の場合と同様に母平均の信頼区間を求めることができます。次の問題を考えてみましょう。

> ある試験の点数の分布は正規分布であるとします。この試験の受験者から，10 人からなる標本を無作為抽出して，この 10 人の点数を平均したところ 50 点で，またこの 10 人の点数の不偏分散が 5^2 でした。このとき，受験者全体の平均点の 95%信頼区間を求めてください。

自由度 $n-1$ の t 分布において，$t_{0.025}(n-1)$ を「t 統計量がその値以上になる確率が 0.025 であるような値」（「2.5 パーセント点」といいます）とし，$-t_{0.025}(n-1)$ を「t 統計量がその値以下になる確率が 0.025 であるような値」とすると

$$P\left(-t_{0.025}(n-1) \leqq \frac{\bar{X}-\mu}{\sqrt{s^2/n}} \leqq t_{0.025}(n-1)\right) = 0.95 \tag{7}$$

が成り立ちます（図 3)。この式から，

$$P\left(\bar{X} - t_{0.025}(n-1)\sqrt{\frac{s^2}{n}} \leqq \mu \leqq \bar{X} + t_{0.025}(n-1)\sqrt{\frac{s^2}{n}}\right) = 0.95 \tag{8}$$

となりますから，μ の 95%信頼区間の下限と上限は，(8) 式のかっこ内の不等式での下限と上限として表されます。

$t_\alpha(\nu)$，すなわち自由度 ν の 100α パーセント点の値を知るには，一緒に配布した数表（t 分布表）を利用することができます。数表では，各自由度 ν（縦軸）と定数 α（横軸）に対して，$t_\alpha(\nu)$ が縦 ν・横 α の交点の値を読むことで求められます。この問題の場合，標本平均 $\bar{X}=50$，不偏分散 $s^2=25$ で，数表から $t_{0.025}(10-1)=2.262$ ですから，μ の 95%信頼区間は「46.4（点）以上 53.6（点）以下」となります。

母分散がわかっている場合と比べてみると

前回の例のように，母分散が 25 とわかっているときには，μ の 95%信頼区間は「46.9（点）以上 53.1（点）以下」でしたから，今回の場合の方が信頼区間が広くなっています。信頼区間が広いということは，推定が不確かであることを意味しています。これは，不偏分散は母分散そのものではなく，母分散を推定した値であるため，不偏分散にはすでに不確かさが入っているためです。

付録：「スチューデント」という名前について

今回説明した「t 分布」は，別名「スチューデントの t 分布」とよばれています。「スチューデント」というのは，t 分布を発見した英国の統計学者ウィリアム・ゴセットのペンネームです。ゴセットはギネス

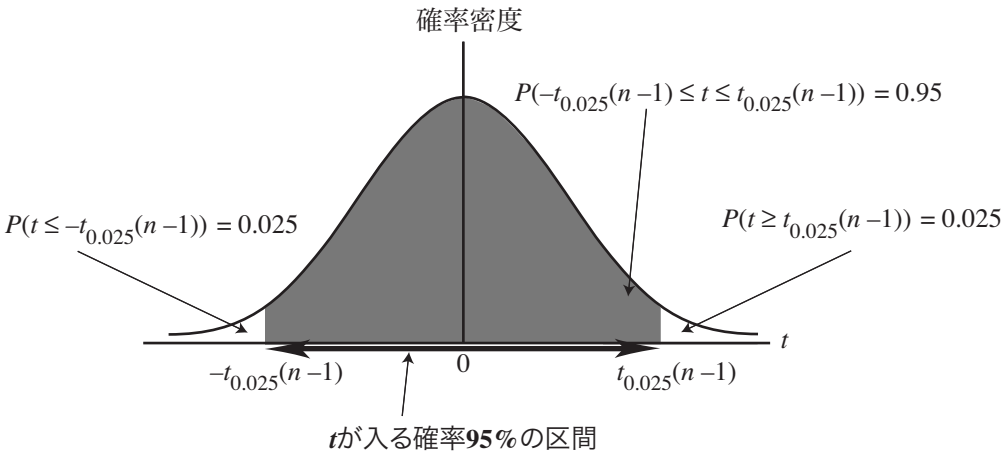

図 3: t 分布と区間推定

ビール社のエンジニアで，会社との契約で本名では論文が出せなかったため，ペンネームで発表したのだそうです。

酒造りなどの醸造業は，昔から統計学が活躍している分野です。それは，酒などができる醗酵という現象は，技術者が自分の手で直接行っているものではなく，無数の微生物の作用や分子の化学反応によって生じるものだからです。技術者は，微生物や分子の作用を，温度や時間などの全体的な量を変化させて調節しているだけです。どう調節すれば，無数の微生物や分子の作用が「全体として」望ましい方向に進むかを知るには，統計学の手法が必要です。

今日の演習

ある製品の長さを 10 回測定したとき, 10 個の測定値の平均は $10.0(\mathrm{cm})$ で, 測定値の不偏分散が $0.04(\mathrm{cm}^2)$ であることがわかっているとします。製品の測定値が，その真の長さを期待値とする正規分布にしたがって分布しているとするとき，真の長さの 95%信頼区間を求めてください。

(答だけ，あるいは式と答だけを書くのではなく，本文の説明を参考に，文章をきちんと書いて答えましょう。また，本来は，何桁の数字で答えるかを「有効数字」を検討して考える必要がありますが，ここでは小数第 2 位までとしておいてください。)