

不確かな測定の不確かさを測る — 不偏分散と t 分布

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

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、標本から推定した分散を使って、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, \dots, X_n 、標本平均を \bar{X} とするとき、不偏分散 s^2 は

$$s^2 = \frac{1}{n-1} \{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2\} \quad (1)$$

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。

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つ（2つしか標本をとらない）とします。母分散は、各データと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。

2つの標本がどちらも母平均から偏ってへだたっているとしても、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。

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。

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。

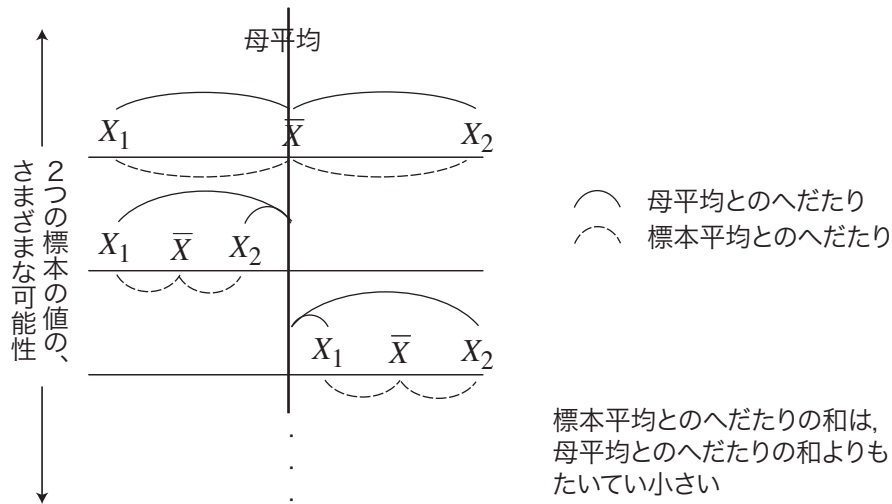


図 1: なぜ $n - 1$ で割るのか？

t 分布と区間推定

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、

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \quad (2)$$

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。

では、母分散 σ^2 が不明であるとしましょう。このとき、(2) 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

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \quad (3)$$

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のでしょうか？

この t 統計量に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。

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、この試験の受験者から 10 人の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 統計量は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 統計量はその値以下になる確率が

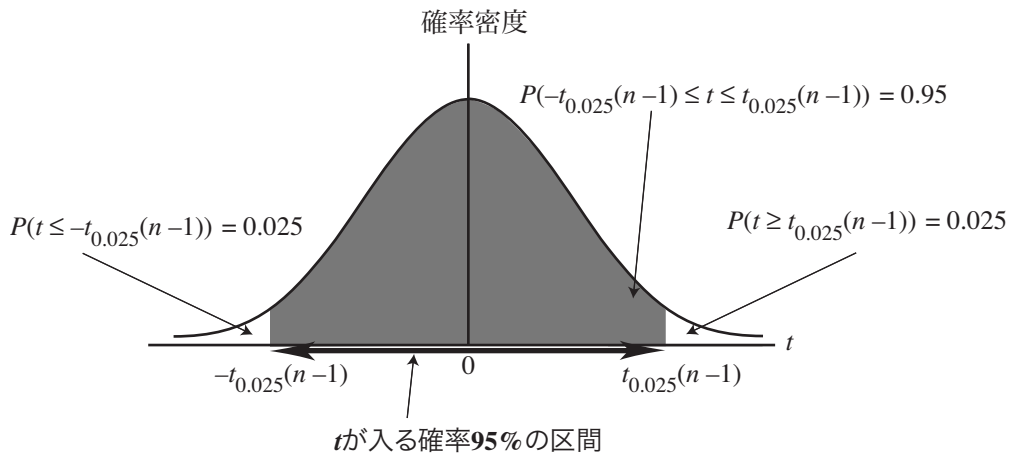


図 2: t 分布と区間推定

0.025 であるような値」とすると

$$P\left(-t_{0.025}(n-1) \leq \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \leq t_{0.025}(n-1)\right) = 0.95 \quad (4)$$

が成り立ちます (図 2)。この式から、

$$P\left(\bar{X} - t_{0.025}(n-1)\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{X} + t_{0.025}(n-1)\sqrt{\frac{s^2}{n}}\right) = 0.95 \quad (5)$$

となりますから、 μ の 95%信頼区間は (5) 式のかっこ内の範囲となります。

$t_\alpha(\nu)$ 、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。

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がっています。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

ある薬品 X の 1 グラムあたりに含まれる物質 Y の量を 10 回測定すると、測定値の平均は 1.0(ミリグラム)、不偏分散は 0.1 ((ミリグラム)²) でした。測定値は真の含有量を平均値とする正規分布にしたがって分布し、また各測定値はそれらから無作為抽出された標本とみなせるとします。このとき、「X の 1 グラムあたりに含まれる物質 Y の量」を信頼係数 95% で区間推定してください。