2008 年度前期　統計学で考える　第８回

「分布」するデータを扱う **(2)** ー 不偏分散，t 分布と区間推定

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 $X_1, X_2, \ldots, X_n$，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．

> 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]．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，というものです．
>
> 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．いくつかのデータを標本として取り出すとき，図 1 の上のように，母平均のまわりに偏りのないデータが取り出されれば，各データと母平均との隔たりも，標本平均との隔たりも，あまり変わりません．
>
> しかし，実際には図 1 の下のように，母平均からみて偏ったデータが取り出されることがしばしばです．この場合，標本がいずれも母平均から偏ってへだたっていても，標本平均はつねに各標本の中間にあります．この場合,「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．
>
> ですから，標本をつかってそのまま分散を計算すると，その「標本の分散」は，たまたま母分散に近いこともありますが，たいていは母分散よりも小さくなります．この違いを調整す

[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．

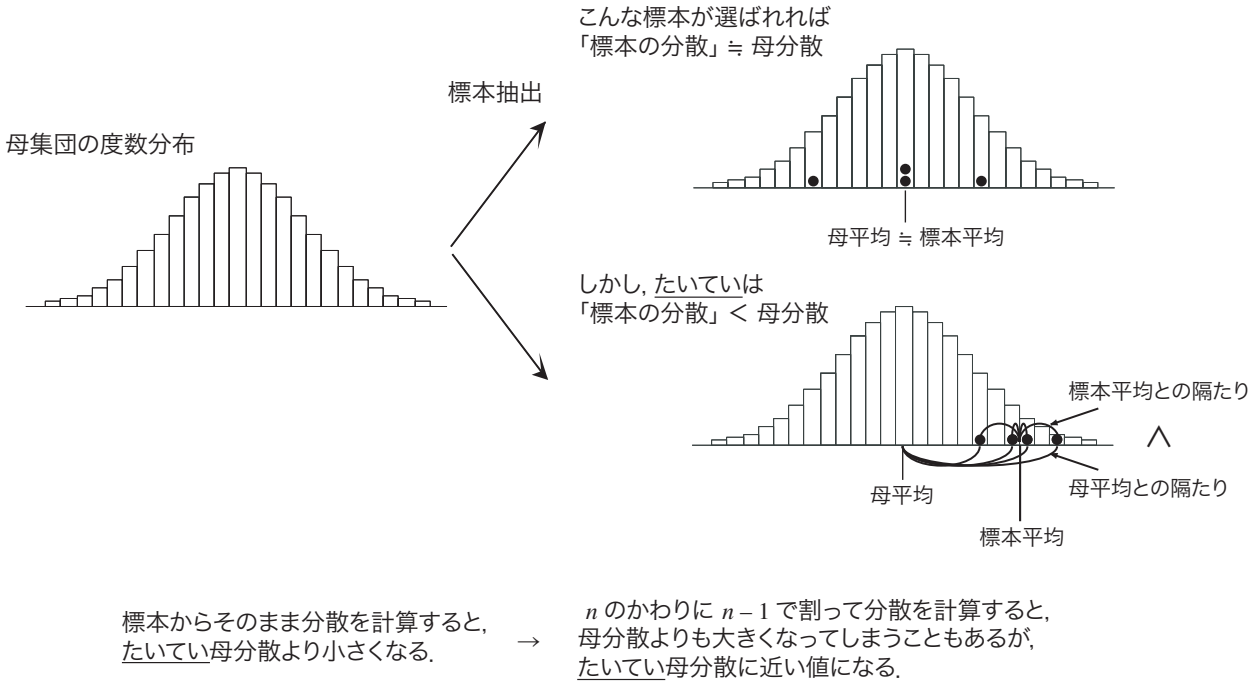

図 1: なぜ $n-1$ で割るのか？

るために，n ではなく $n-1$ で割って，少し大きめにしているのです[2].

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，

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

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

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

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でしょうか？

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．

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．

[2] くわしくは，2006 年度後期の浅野の講義「情報統計学」の第 8 回の講義録をネットで参照してください.

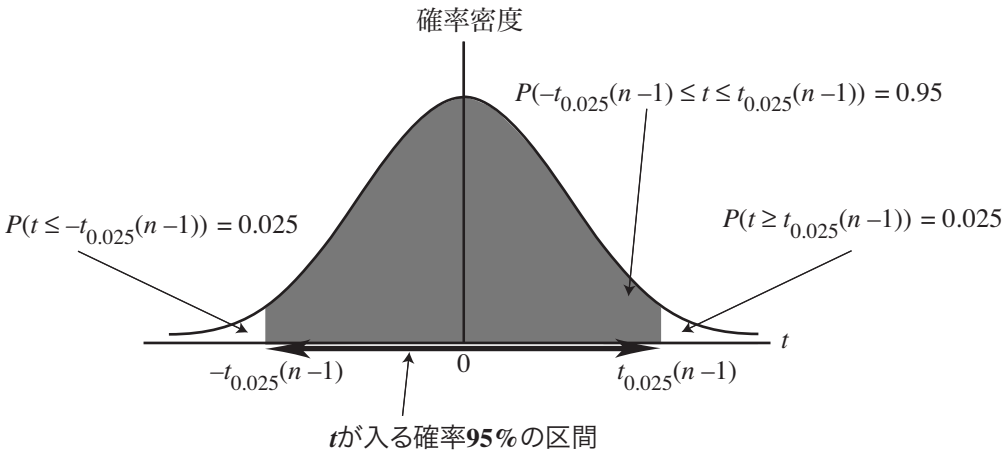

図 2: t 分布と区間推定

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

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

が成り立ちます（図 2）．この式から，

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

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，今日いっしょに配った数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．

---

t 分布と検定

同じ考えで，母平均に関して検定を行なう問題を考えてみましょう．

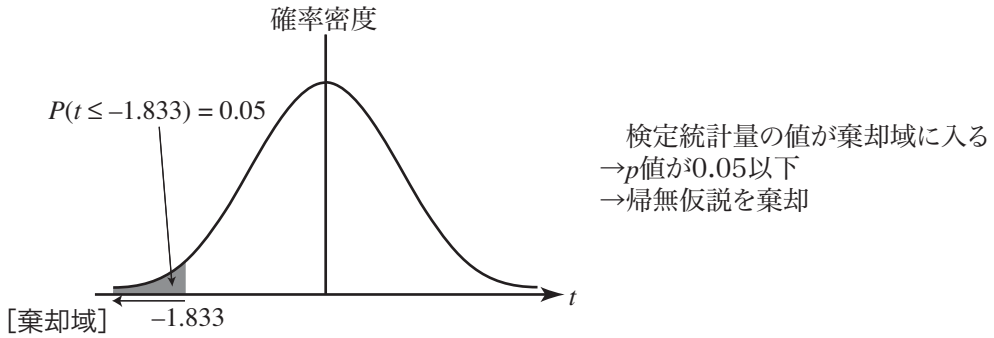

図 3: t 分布と検定

ある試験の点数の分布は正規分布であるとします．この受験者全体から無作為抽出された 10 人の標本の点数の平均は 50 点で，不偏分散は 5^2 でした．このとき,「受験者全体の平均点は 53 点よりも小さい」といえるか，有意水準 5%で検定してください．

問題から，帰無仮説 $H_0 : \mu = 53$, 対立仮説 $H_1 : \mu < 53$ の検定を行います．母平均を μ，標本平均を $\bar{X}$, 不偏分散を s^2，標本サイズを n とするとき，上で述べたとおり，

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

という値は，自由度 $n - 1$ の t 分布にしたがいます．

「対立仮説 $H_1 : \mu < 53$」というのは,「帰無仮説が棄却されたとすると，そのときは $\mu < 53$ という対立仮説を採択する」という意味です．つまり,「帰無仮説が棄却されたとすると，それは『μ は 53 では大きすぎる』からだ」という推論をしたいわけです．μ が大きくなると，(6) 式の t は小さくなります．ですから,「t が小さすぎるとき」帰無仮説が棄却されるように，棄却域を設定します．t 分布の数表から，自由度 $n - 1 = 9$ のとき，(6) 式の t が $-t_{0.05}(9) = -1.833$ 以下である確率すなわち $P(t \le -1.833)$ が 5%であることがわかります．ですから，問題文の数値を入れて (6) 式の t を計算し，その値が −1.833 以下であれば帰無仮説を棄却します．

問題文の $s^2 = 25$，$\bar{X} = 50$，$n = 10$，それに $\mu = 53$ を (6) 式に代入すると $t = -1.897$ ですから，この t は棄却域に入っています．したがって，帰無仮説を棄却して対立仮説を採択し,「受験者全体の平均点は 53 点よりも小さい」と結論します．