第 6 章

無作為抽出と標本分布

ここからが**推測統計**の始まりである．前章まで問題にしてきた確率も，多数回の試行によって意味を持つものであるが，基本的には個々の事象の期待値に対する理論である．

推測統計の理論は，未知の大きな集合（母集団）の中から取り出した小さな集合（標本）を調べて，親の集合の特性を確率論的に調べるものである．その関係を理解しよう．

6.1 無作為標本抽出

図 6.1 に**無作為標本抽出 (random sampling)** の操作にかかわる概念と諸量の関係を示した．ここで，**母集団 (population)** は未知の**母平均 (population mean)** と**母分散 (population variance)** をもつ「知りたい集合」である．それをより小さな**標本 (sample)** から知ろうというのが，ここからの目標だ．

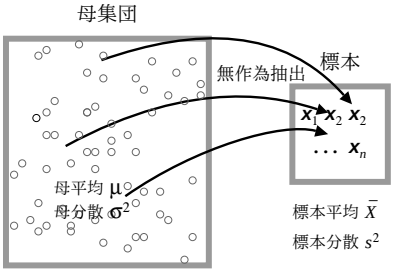

図 6.1 無作為抽出の操作の概念．古典的な統計学のすべては，未知かつ知りたい量である μ, σ^2 と，手のひらの上にある $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}$ から計算できる $\overline{X}, s^2$ を結びつけるところにある．

6.1.1　母集団と標本

母集団 (population) は調査分析したいすべてのサンプルを含む集合である．母集団に含まれる要素の数は有限であっても無限[*1]であってもよい．

母集団の要素すべてについてデータを集めたとして，その平均を**母平均**, **分散**を**母分散**という．これ以降は**母平均**を μ, **母分散**を σ^2 で表すことにする．また，これらの母集団の特性を表すパラメータをまとめて**母数 (population parameters)** と呼ぶ．

母集団のすべてのデータについて集計する，**全数調査**あるいは**悉皆調査**は多くの場合困難である[*2]．また，調査や分析によって標本そのものが破壊されたり変化を蒙ってしまうような場合には，全数調査を行うことは本来の目的を果たせなくなってしまうことになる[*3]．したがって，母集団から有限個の**標本 (sample)** を取り出して，定量的な調査，分析を行うことが多い．取り出される標本の数のことを標本の**大きさ (size)** あるいは**サイズ**とそのまま いう．以後，「大きい標本」，「小さい標本」という表現をしばしば使うが，要するに標本として取り出したデータの数が大きいか小さいかを表していると思ってほしい．

6.1.2　乱数による無作為抽出

母集団から一部の標本を取り出すことを，**標本抽出 (sampling)**，あるいは**抜き取り**という．英語をそのまま使って**サンプリング**ともよくいう．標本抽出に当たっては，母集団の特性が標本の集合にもなるべく正確に反映されるようにしなければならない．そこで行われるのが**無作為標本抽出 (random sampling)** である．すなわち，母集団のどの要素についても，同じ確率で標本として抜き取られるように，作為を交えずにそれぞれを選ぶのである．

無作為抽出には，**離散型一様分布**に従う**乱数**が使われる．現代ではコンピュータを使って乱数を発生させることがほとんどである[*4]．

[*1] ここで無限といっているのは，二つの場合がある．ひとつはきわめて多数であって実際上無限と近似してかまわないもの，もうひとつはいくらでも繰り返される事象，たとえばコインの裏表の出方を調べるとか，工場のラインから製品が毎日生産されるといった場合も，無限とみなせる．

[*2] 国勢調査は最も大規模な全数調査である．この場合には，もっとも完全な形で母集団の統計情報が得られるが，それに要する費用と時間は莫大なものである．それでも真にすべての要素について調べつくすことは，現実には不可能である．

[*3] たとえば電球の寿命の検査をある工場の製品すべてについて行ったとしたら，切れた電球しか出荷できなくなってしまうであろう．

[*4] 「離散型一様乱数」は，どの数も等しい確率で出現し，次に何が現れるかを完全に予測できない，完璧なサイコロのような乱数だが，コンピュータの有限手数のアルゴリズムでは原理的に作れない．実際に作られるのは**擬似乱数**という，「かなり乱数ぽい数列」である．とはいえ，現在のコンピュータで利用されている擬似乱数は，実用上は正しい乱数と考えてもかまわない．

表 6.1 乱数表の例

98474	**71279**	**63082**	**78829**	**42648**	**14443**	**69985**	**58505**	**73760**	**96835**
37252	**88586**	**62283**	**71713**	**61004**	**62979**	**29684**	**15151**	**41589**	**44958**
43215	**04177**	**61654**	**95413**	**43685**	**95877**	**61315**	**09869**	**46923**	**85614**
76004	**67425**	**09426**	**72476**	**52651**	**44729**	**98959**	**10064**	**09796**	**98117**
60610	**70770**	**57281**	**67053**	**19024**	**01629**	**41143**	**01965**	**07339**	**99938**
29309	**69622**	**63555**	**86700**	**03750**	**39202**	**84902**	**06042**	**74703**	**02108**
80801	**28750**	**82589**	**28729**	**15136**	**88027**	**03250**	**15225**	**78384**	**25588**
22125	**23483**	**80242**	**76254**	**93014**	**67361**	**03408**	**69128**	**47009**	**48339**
09106	**73507**	**67285**	**93722**	**35009**	**67651**	**95285**	**00497**	**76141**	**58511**
84030	**37979**	**89450**	**30578**	**64083**	**12380**	**12603**	**51943**	**37857**	**46401**

表 6.1 に乱数表の例を示した．乱数表では，どのように数を拾っていっても特定の傾向が現れることはないようになっている．たとえば，この表から一桁の乱数の系列を得るのに，5 桁の数を下へと追いながら，9,8,4,7,4,3,7,2,5,2,4,3,2,1,5, .. としてもよいし，あるいは 5 列目の縦の系列の 5 桁目だけを拾って，4,6,4,5,1,0,1,9,3,6,... などとしてもよい[*5]．

抽出には，母集団から同じ標本を繰り返して抽出することを許す**復元抽出**と，同じ標本は重複させないでとる**非復元抽出**の二通りのやり方がある．

復元抽出の例：

赤と白の球の入った袋から 1 個取り出しては，また元の袋に戻して次の球を取り出す

アンケート対象者を乱数で決めたとき，たまたま同じ人が 2 回選ばれても，その人に 2 回尋ねることにする

サイコロで当選者を決めていく．二度選ばれても構わない．

非復元抽出の例：

赤と白の球の入った袋から 1 個取り出したら，取り出した球を戻さないで次の球を取り出す

アンケート対象者を乱数で決めたとき，たまたま同じ人が 2 回選ばれたら，重複しないように別の人を乱数で選ぶ

サイコロで当選者を決めていく．二度選ばれたら外す．

例題 6–1 表 6.2 に掲げた高校生の体重の一覧表から，乱数を使って 10 人の標本を抽出せよ．

[*5] 乱数表を使う場合には，「愚直に」数を拾わなければならない．たとえばある特定の数字が連続したりすると，それを排除してもっともらしくしようとする心理が働くものであるが，そのことで必ず何らかの傾向が生じてしまうことになる．ある数字が繰り返されることに対する心理的抵抗は大きいものだが，ちょっと確率の計算をすれば，同じ数字が連続して現れる頻度は，かなり高いことが分かるもので，それを避けることは意味がないのである．ただし，意図的に非復元抽出を行うような場合には，重複を排除するようにして抽出を行なう．

表 6.2　100 人の男子高校生の体重/kg(再掲)

43.6	45.2	45.4	45.8	47.2	47.8	48.2	48.7	48.8	48.9
49.0	49.0	49.4	49.5	49.8	50.4	50.5	50.9	50.9	51.2
51.2	51.2	51.3	51.3	51.6	51.7	51.7	51.8	52.0	52.0
52.1	52.1	52.1	52.2	52.3	52.7	52.7	52.8	52.9	52.9
53.1	53.1	53.8	54.0	54.5	54.5	54.6	54.7	54.7	54.7
54.8	54.9	55.1	55.1	55.2	55.3	55.4	55.4	55.4	55.6
55.7	55.8	55.9	56.1	56.3	56.3	56.3	56.4	56.5	56.7
56.8	57.0	57.1	57.1	57.2	57.3	57.6	57.7	57.8	58.1
58.4	58.6	58.7	58.7	58.7	58.7	59.1	59.3	59.9	60.0
60.1	60.3	60.5	60.6	60.6	60.7	61.3	62.7	64.2	64.6

母集団は 100 個のデータからなるので，それらに 0 – 99 の番号をつけることができる．そして乱数として 2 桁のものを用いれば，母集団からランダムに抽出を行うことができる．ここでは表 6.1 の乱数表の 1 行目の数字を取り出していっては，2 桁ずつに区切って得られる数字を用いることにしよう．すると次の数列が得られる．

$$98, 47, 47, 12, 79, 63, 08, 27, 88, 29, 42, 64, 81, 44, ...$$

ところがここで，47 という値は 2 番と 3 番目に重複して出現しているから，その通りにデータを拾うと，47 番の生徒のデータを二度拾うことになる．このように重複を許して抽出するやり方が**復元抽出**である．この場合，

$$98, 47, 47, 12, 79, 63, 08, 27, 88, 29$$

の番号の生徒の体重が次のように抽出される．

$$64.2, 54.7, 54.7, 49.4, 58.1, 56.1, 48.8, 51.8, 59.9, 52.0$$

一方，重複を許さずに抽出するとすれば，2 度目の 47 は飛ばして，

$$98, 47, 12, 79, 63, 08, 27, 88, 29, 42$$

という番号でデータを拾えばよい．すると次のデータが抽出される．

$$64.2, 54.7, 49.4, 58.1, 56.1, 48.8, 51.8, 59.9, 52.0, 53.8$$

このような抽出の仕方が**非復元抽出**である．ただし母集団が十分に大きいときには，同じものが重複して抽出される確率は低い．したがって，復元抽出でも非復元抽出でも結果にほとんど違いはなくなる．そこで，これ以降では特に断りがない限り，標本を抽出するときには復元抽出を行うものとして進めることにする．

6.2 標本平均の分布

6.2.1 抽出された標本は信頼できるか

図 6.1 をもう一度見てほしい．母集団から抽出された n 個の標本のデータを

$$X_1, X_2, \ldots, X_n$$

としよう．X_1, X_2 等は，サイコロの目の数とか乱数で選んだ高校生の体重といった，何らかの予測できない数で，確率変数であり，特にこの場合には**標本確率変数**と呼ばれる．サイコロを振るときを考えれば，復元抽出であれば，$X_1, X_2, ..., X_n$ は互いに独立であることがわかる．

これから，次の 2 つの重要な量, **標本平均**および**標本分散**を求めることができる．

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \text{ (標本平均)} \tag{6.1}$$

$$s^2 = \frac{1}{n}\left(\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2\right) \text{ (標本分散)} \tag{6.2}$$

当然，標本分散の正の平方根 s は**標本標準偏差**と呼ばれる．

標本平均と母平均の関係を考えてみよう．n が小さいと，標本平均は母集団の要素のごく一部の平均になるのであるから，両者にはずれが大きいだろう．逆に n がきわめて大きければ，つまり大きな標本であれば，標本平均は母平均とよい精度で一致するだろう．これは，私たちが日ごろからほとんど無意識にそう感じていることだ．図 6.2 は，そのことをシミュレーションによってざっと調べたものである．

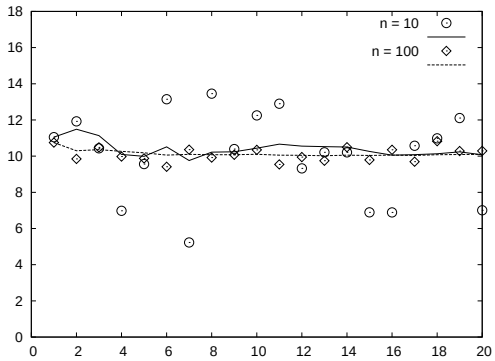

図 6.2 $[0, 20]$ の区間で連続一様分布している母集団から，大きさが 10 と 100 の標本を抽出する操作を 20 回ずつ行った結果．○ と ◇ はそれぞれ $n = 10, 100$ の場合の $\overline{X}$, 実線と破線は $\overline{X}$ のそこまでの平均値．

図から分かるように標本の大きさが大きいと，標本平均 $\overline{X}$ の幅は小さくなり，またその平均値は母平均 $\mu = 10$ に急速に落ち着いていく．

6.2.2　標本平均の分布に関する最も重要な式

上記の定性的な事実を数式できちんとまとめることにする．まず，もしも標本抽出を多数回繰り返していったとすると，その標本平均 $\overline{X}$ の平均は母平均に限りなく一致していくにちがいない．つまり，**標本平均の期待値は母平均に一致する．**

次に，$\overline{X}$ の「広がり」，つまり分散は，上のシミュレーションからも想像されるように，n が大きいほど小さい．また，もしも母分散 σ^2 が小さければ，$\overline{X}$ の広がりも狭く絞り込まれるにちがいない．以上の傾向をまとめたのが次の式である．

$$E[\overline{X}] = \mu \tag{6.3}$$

$$V[\overline{X}] = \frac{\sigma^2}{n} \tag{6.4}$$

これらの詳しい導出は A.5 節 (p.152) に示した．ここで式 (6.4) で表される $V[\overline{X}]$ の平方根 $\frac{\sigma}{\sqrt{n}}$ は**標準誤差 (standard error)** と呼ばれ，抽出された標本の平均の分布の幅を表す重要な指標である．標準誤差はよく SE と略して使われる．

今後の展開において，標本平均と標準誤差はきわめて大きな役割を果たす．

6.3　標本分散の分布

6.3.1　標本分散の平均

標本平均 $\overline{X}$ の分布については上のシンプルな関係式が出てきたが，標本分散 s^2 の分布についてはどうだろうか．

この場合，一般的な関係式は s^2 の期待値についてだけ存在する．この結果も重要で，とてもよく使われる．式の導き方は付録 A.6 に示した (p.152)．

$$E[s^2] = \frac{n-1}{n}\sigma^2 \tag{6.5}$$

例題 6–2　復元無作為抽出によって 6 人の組を何組も選んで体重の平均と分散をとった．その結果，分散の平均値は $(4.32\ \mathrm{kg})^2$ となった．母集団の標準偏差を推定せよ．

上で求められた分散 (s^2) の平均値 (期待値と読み換えてよい) というのは，式 (6.5) の $E[s^2]$ であり，求めたいのは母分散 σ^2 の平方根 σ なのであるから，

$$\sigma^2 = \frac{n}{n-1}E[s^2] = \frac{6}{5} \times 4.32^2$$

$$\sigma = \sqrt{\frac{6}{5}} \times 4.32 = 4.73$$

より，母標準偏差は 4.73 kg となる．

■標本分散はどうして母分散よりも小さめなのか？

式 (6.5) が定性的に語っている[*6]ことは，「標本分散は，平均として母分散よりも少し小さめの値をとる確率変数である」ということだ．どうして小さめになるのだろうか？そのわけは次のとおりだ．

s^2 の定義は次のようになっている．

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

これと次の式を比べてみよう．

$$\frac{1}{n}\left((X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2\right)$$

こちらの式は標本のデータ $X_1, X_2, \cdots$ のそれぞれから μ を引いたものの 2 乗の平均であり，その期待値は σ^2 になる[*7]．一方，上の s^2 の方では，μ ではなく $\overline{X}$ を引いている．

仮に取り出されるデータが，たまたま μ に比べて大きい方に偏っていたとしよう．すると，$\overline{X}$ はそれらの平均なのだから，引きずられて大きい方にずれることになる．結果，s^2 の定義の中の $(X_1 - \overline{X})^2$ 等は $(X_1 - \mu)^2$ 等に比べて小さくなる．そのため，s^2 は σ^2 よりも小さい値を取る．また，n が小さいほど，そのような偏りが生じる確率が大きくなるので，この傾向は著しくなる．

この傾向は，サイズの小さい標本で著しくなり，後の t-分布のところで大事な意味を持つことになる．

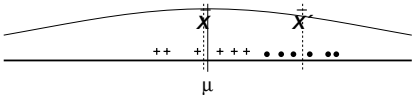

図 6.3　大きさ 6 の標本で，データが母平均の周りに分散している状況 (6 個の＋) と全体として正の方向に偏っている状況 (6 個の●)．

6.3.2　標本分散の分散については？

さて，組み合わせとして最後に残っている標本分散の分散 $V[s^2]$ についてはどうだろうか．その場合の一般的な関係式は存在せず，正規母集団から抽出された標本について，χ^2-分布が現れる．それについては後で取り扱う．

[*6] もちろん「数式は語る」のです．耳をすまそう．

[*7] 仮に標本のサイズを 1 として，X_1 しかない標本を無限に取り出すことを考えると，$X_1 - \mu$ の 2 乗の和の期待値なので σ^2 になることがわかる．

6.4　正規母集団

6.4.1　正規母集団と標本平均の分布

以上で，標本平均 $\overline{X}$ の期待値と分散の**分布**が，母集団の期待値と分散とどう結び付けられるかがわかった．しかし $\overline{X}$ がどのような確率分布に従うかは，まだわかっていない．その点を見ていこう．大きく整理すると次のようになる．

母集団が正規分布している場合：　n の大きさに関わらず $\overline{X}$ は正規分布に従う．つまり**「正規分布は，子どもたちもやっぱり正規分布」**なのである．実際 $n = 1$ のとき，この標本は母集団の中のデータそのものであり，十分な回数の抽出を行ってみれば元の正規分布が現れることは明らかである．
さらに，2 つの正規分布に従う確率変数があるとき，それらの和も正規分布に従うので，n が複数の場合にも $\overline{X}$ は正規分布する．

母集団が正規分布していない場合：　この時には，n が十分大きいならば，$\overline{X}$ は正規分布になる．つまり，**「非正規分布の子でも，たくさん集めて平均したら正規分布になる」**わけだ．これを保証しているのが中心極限定理である．

しかし，n が少数の時には，$\overline{X}$ がどのような分布に従うかについて一般的な法則はなく，個別に解決すべき問題になる．

以上のように，母集団が正規分布している時には，標本平均の分布についても，正規分布の性質を使って非常に強力な手法が使える．その意味で，正規分布に従う母集団のことを**正規母集団**と呼ぶ．

正規分布はまた，世論調査のように非正規母集団から大きな標本を取り出す場合にも利用できて，統計的な扱いにおいて中心的な役割を果たすことができるのである．

例題 6–3　有権者の中の内閣支持率が 30% であったとしよう．1000 人をランダムに選んでアンケートをとった時，そこから得られる支持率の期待値と分散を求めよ．

母集団中の支持を 1, 不支持を 0 とする．この母集団は，0 と 1 しか含まない極端な集合であって，正規母集団ではないが，以下のような取り扱いが可能である．

まず，母平均と母分散を求めておく．母集団中の支持の割合を $p = 0.3$ としよう．つまり，母集団の中で 1 つのデータを見た時にそれが 1 である確率は p, 0 である確率は $1 - p$ だ．したがって式 (3.10) より．

$$\mu = 1 \times p + 0 \times (1 - p) = p$$

となり，また式 (3.11) より，

$$\sigma^2 = (1-\mu)^2 \times p + (0-\mu)^2 \times (1-p) = p(1-p)$$

となる．よって $\mu = 0.3$, $\sigma^2 = 0.21$.

これから，標本平均の期待値と分散は，それぞれ $\mu = 0.3$ と $\sigma^2/n = 0.21/1000 = 0.00021$ となる．分散がきわめて小さい感じがするが，これから標準偏差は 0.0145 となるので，μ と比較して妥当な値になっている．

■カテゴリカル変数と二項分布

支持と不支持，性別といったカテゴリー分け（分類）は，数値的な変数ではないので，**カテゴリカル変数**[*8]とか**名義変数**と呼ばれる．一般に社会調査などにおいてはカテゴリカル変数は非常によく登場する．

カテゴリカル変数であっても適当な数に対応させることで確率変数として取り扱うことができ，特に 2 通りのカテゴリーしかない場合にはそれらを 0 と 1 に対応づけることで，このように単純な定量的な取り扱いが可能になる．

6.4.2　標本平均の分布

■標準化変換された標本平均の分布

$N[\mu, \sigma^2]$ に従う正規母集団から，大きさ n の標本を抽出したとする．すると前節の議論と式 (6.3),(6.4) から，$\overline{X}$ は平均 μ と分散 σ^2/n (標準誤差 $\sigma/\sqrt{n}$) をもつ正規分布に従う．図 6.4 に，そのようすを示した．この図も今後しばしば登場する．標準誤差が n の平方根の逆数に比例することから，標本が大きいほど $\overline{X}$ の広がりは狭まって，分布がシャープになることをしっかり見ておこう．

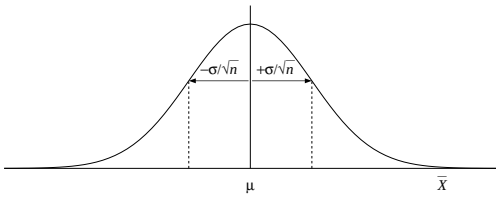

図 6.4　$N[\mu, \sigma^2]$ に従う正規母集団から大きさ n の標本を抽出したときの $\overline{X}$ の分布.

さてここで，今後のために 図 6.4 を標準化変換してみよう．

[*8] カテゴリー変数ともいう.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \tag{6.6}$$

こうして得られた Z は標準正規分布 $N[0,1]$ に従うことを第 5 章ですでに見た (p.71). このように，現実の確率変数に適切な変換を施すことで，なんらかの標準的な確率分布に従うようにするという手続きは，一般化を助けてくれるものだ．次の例に進もう．

6.4.3　母分散なしで標本平均の分布を表現する

ここで，標本抽出のもっとリアルな状況を想定してみよう．

標本平均 $\overline{X}$ の分布を標準化変換して得られた式 (6.6) は，$\overline{X}$ が式 (6.3)，(6.4) で決まる期待値と分散をもつ正規分布に従うことから導かれている．これらのパラメータは母平均 μ と母分散 σ^2 から導かれているわけだ．

しかし，私たちが現実に手にしているのは，あくまで標本のデータだけしかないのである．未知の，というよりむしろ，「知りたい」パラメータである母集団の統計量をなるべく使わないで $\overline{X}$ の分布を決めることができたら，応用はずっと広がるにちがいない．

■標本のサイズが大きい場合

式 (6.5) で示されるように，s^2 の期待値は母分散 σ^2 に係数 $\frac{n-1}{n}$ を掛けたものに等しい．式を再掲しておく．

$$E[s^2] = \frac{n-1}{n}\sigma^2$$

これは，十分な回数サンプルを繰り返すと，s^2 の平均は $\sigma^2 \times \frac{n-1}{n}$ に等しくなるとも言い換えられるわけだ．そこで，上の式の期待値の括弧をえいやっと外して等式にしてしまって，少しいじってやると，

$$\sigma^2 = \frac{n}{n-1}s^2$$

となる．つまり，この右辺をもって未知の標本平均の「代用品」に使おうというのが，標本のサイズが大きい場合のアイディアである．

それではさっそく，この場合についての標本平均 $\overline{X}$ の分布を描いてみよう．まず標本が大きいのだから，母集団が正規分布しているか否かに関わらず $\overline{X}$ は正規分布しているとしてよい（中心極限定理）．また，一つ前にやった，母分散が既知の場合と同様に，$\overline{X}$ の標準偏差 (標準誤差) は $\frac{\sigma}{\sqrt{n}}$ の形だが，ここでは σ を $\sqrt{\frac{n}{n-1}}s$ で置き換えることになるので，

$$\frac{\sqrt{\frac{n}{n-1}}s}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

となる．結局 $\overline{X}$ は平均 μ と分散 $s^2/(n-1)$ (標準偏差 $\frac{s}{\sqrt{n-1}}$) をもつ正規分布に従うことになる．

ここでも，この結果を標準化変換によって一般化しておこう．式 (6.6) とまったく同様にやればよい．次のようにして Z を決めると，

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{6.7}$$

Z は標準正規分布 $N[0,1]$ に従う．

■標本のサイズが小さい場合—スチューデントの t-分布

今度は，正規母集団から抽出された標本のサイズが小さい場合を考える．ポイントは式 (6.7) だ．これを特別扱いするために，Z を T に変更しよう．

$$T = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{6.8}$$

大きな標本ではこの s をあまり動かないものとして扱ったのが，前の扱いになる．そのため Z(ここでは T) は標準正規分布になった[*9]．ところが，標本のサイズが小さい時には，そのことが言えなくなる．

6.3 節で説明したように，標本分散 s^2 は n が小さい時には母分散 σ^2 よりも平均的には小さく，かつ変動の幅が大きい確率変数として振る舞う．標本標準偏差 s も，もちろんそうなる．図 6.5 に，そのことをシミュレーションで確かめた結果を示した．

ということは，式 (6.8) の分母の s は，時として大きな値を取ることになる．それはどういう結果をもたらすかというと，関数を横に引き伸ばす効果をもつのだ．つまり，正規分布の形を左右に引き伸ばしたような分布が作られることになる．これが**スチューデントの t-分布**，あるいは単に **t-分布**とか**スチューデント分布**とよばれる確率分布である[*10]．

t-分布は式 (6.9) で表され，正規分布のような形をしているが，図 6.6 のようにやや裾の広がった形になる．また**自由度**というパラメータ ν があって，その値によって形が異なる．

$$f_\nu(T) = c\left(1 + \frac{T^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad (\nu = 1, 2, 3, ...) \tag{6.9}$$

ここで c は $f_\nu(t)$ の全面積が 1 であるようにするための定数であり， ν は自由度と呼ばれている．

[*9] 厳密にはこの説明は不十分で，s の分布の幅が小さく，かつ正規分布に近い対称な分布になっていることが，T が正規分布する理由なのだが，それを議論すると数学的な細部に突っ込みすぎるだろう．

[*10] Student というのは，この分布を提案した数学者のペンネームで，本名はウイリアム・ゴセット (Willam S. Gosset)．ゴセットはビールで有名なギネス社の社員で，ビールの品質管理に彼が考案した新しい統計的方法を適用して，会社の業績発展に多大な寄与をした．しかし会社は社員が社外で研究発表を行うことを禁止していた．そこでスチューデントというペンネームで論文を投稿していたのである．

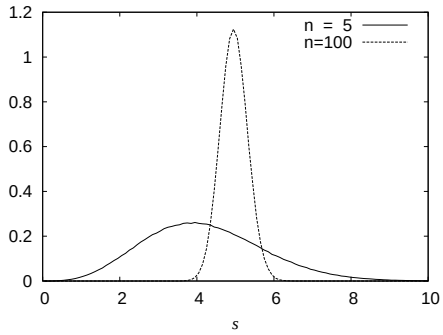

図 6.5　標準正規分布 $N[0,25]$ を母集団として抽出された標本標準偏差 s の分布：$n=5$ の場合には $n=100$ に比べて分布が大きく広がっており，特に裾が右側に大きく伸びていることが特徴的である．抽出はいずれも 100 万回行った．

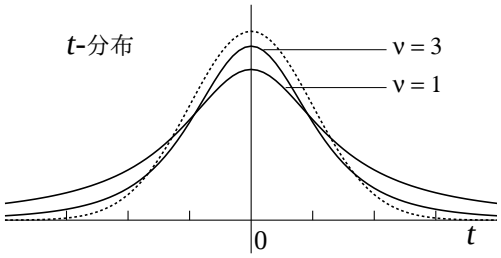

図 6.6　$\nu = 1, 3$ のときの t-分布の形: 破線は標準正規分布曲線

ここで，小さな標本の場合の標本平均の分布がどのように t-分布によって表されるかを見てまとめとしよう．

$N[\mu, \sigma^2]$ に従う正規母集団から，大きさ n の標本を無作為抽出して得られる標本平均 $\overline{X}$ と標本分散 s^2 を得たとしよう．このとき，

$$T = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s} \tag{6.10}$$

で定義される変数 T は，自由度 $n-1$ の t-分布に従う．

6.5 正規母集団と χ^2 分布

6.5.1 標本分散から得られる情報

前節までに，標本分散 s^2 が母集団の統計量とどのような関係にあるかについて，ほんのわずかの結論しか出していない．すなわち，式 (6.5) で，$E[s^2]$ という標本分散の期待値と母分散の関係が与えられているだけである．

しかし私たちが，標本抽出の結果から母集団の分散がどうなっているかを自信をもって知るためには，単に s^2 の平均（期待値）だけではなく，分散がどうなっているかも知っておかないといけない．抽出によって得られた標本分散を使って，母集団の分散についてどの程度のことが言えるのか，それを知らないと危うい結論しか出せないからである．

たとえば 84 ページの例においては，標本分散の複数回の測定から母集団の分散を推定しているわけだが，1 回の抽出ごとの標本分散の値はかなりばらつくので，それらの平均を期待値として安心して使っていいかという問題が残っている[*11]．

この問題については，しかしながら，前節までのような一般的な結論は存在しない．ただし，正規分布をしている母集団，すなわち**正規母集団**の場合については，以下のような結論が導かれている[*12]．これはまず $\mu = 0, \sigma^2 = 1$ であるような正規母集団 $N[0,1]$ について次のように表される．

$N[0,1]$ の正規分布をしている母集団から，n 個の無作為抽出を行ったとき，

$$\boldsymbol{Z = X_1^2 + X_2^2 + ... + X_n^2} \tag{6.11}$$

なる Z は，

$$\boldsymbol{T_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}} \tag{6.12}$$

という確率分布に従う．ここで式 (6.12) は自由度 n の χ^2 分布と呼ばれる確率密度関数である．

$n = 1, 2, ..., 7$ についての χ^2 分布関数の具体的な関数表現は次のとおりである[*13]．図 6.7 にはそのグラフを示した．

[*11] 期待値と平均値は数式の上では同じ形である．しかし期待値というのは無限に回数を重ねたとして，その平均がどこに近づくかという仮想的なものである．一方，実際に実現した事象から計算される平均値は，期待値のまわりに広がって分布する数値になるのである．

[*12] ただし，大きな母集団が正規分布をする傾向は，中心極限定理で保障されているので，正規母集団についての結論は広い一般性をもつものである．

[*13] この関数の形は複雑であるが，覚えておく必要は全くない．グラフで関数の概形を知っておけばよい．

$$T_1(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \tag{6.13}$$

$$T_2(x) = \frac{1}{2} e^{-x/2} \tag{6.14}$$

$$T_3(x) = \frac{1}{\sqrt{2\pi}} x^{1/2} e^{-x/2} \tag{6.15}$$

$$T_4(x) = \frac{1}{4} x e^{-x/2} \tag{6.16}$$

$$T_5(x) = \frac{1}{3\sqrt{2\pi}} x^{3/2} e^{-x/2} \tag{6.17}$$

$$T_6(x) = \frac{1}{4} x^2 e^{-x/2} \tag{6.18}$$

$$T_7(x) = \frac{1}{15\sqrt{2\pi}} x^{5/2} e^{-x/2} \tag{6.19}$$

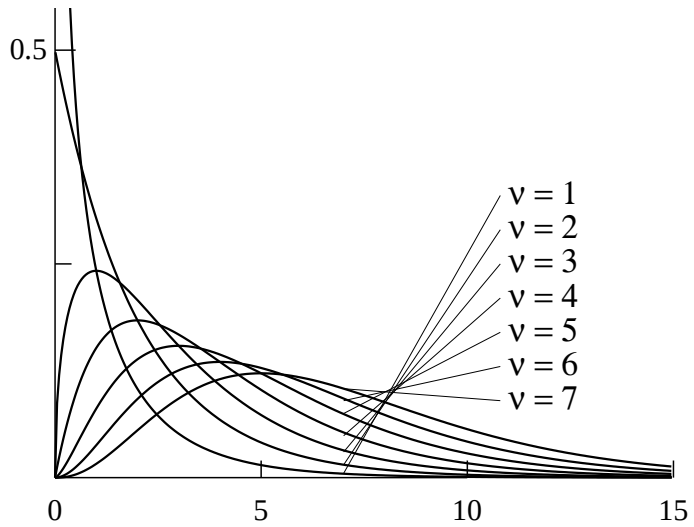

図 6.7　$\nu = 1, 2, ..., 7$ に対する χ^2 分布の密度関数

式 (6.12) に登場する $\Gamma(x)$ はガンマ関数と呼ばれ，階乗 $n!$ を 連続関数に拡張したものである[*14]．なお，この式では n の代わりに ν を使うこともよくあるので，表を引いたりする時には迷わないこと．

さらに，標準正規分布でない一般の正規分布については，次のように言い換えられる．

母集団は平均値 $\boldsymbol{\mu}$，分散 $\boldsymbol{\sigma^2}$ をもつ正規分布 $\boldsymbol{N[\mu, \sigma^2]}$ をしているとする．そこから

[*14] 具体的には，x が正の整数のときには，$\Gamma(x) = (x-1)!$ で，また半整数に対しては，$\Gamma(1/2) = \sqrt{\pi}, \Gamma(3/2) = \frac{1}{2}\sqrt{\pi}, \Gamma(5/2) = \frac{3}{4}\sqrt{\pi}, \Gamma(7/2) = \frac{15}{8}\sqrt{\pi}, \Gamma(9/2) = \frac{105}{16}\sqrt{\pi}, ...$ となる．

n 個の無作為抽出を行ったとき，

$$\boldsymbol{Z = \frac{1}{\sigma^2}\left((X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_n-\mu)^2\right)}\ {}^{*15} \tag{6.20}$$

なる Z は，自由度 n の χ^2 分布 に従う．

上の命題によれば，もしも母集団の平均値 μ が既知であるならば，何回も抽出を繰り返すことによって，母分散 σ^2 を推定できることになる．

さらにもうひとつ，実用的な結果を述べておこう．

母集団は平均値 μ，分散 σ^2 をもつ正規分布 $N[\mu,\sigma^2]$ をしているとする．そこから n 個の無作為抽出を行ったとき，

$$\boldsymbol{Z = \frac{1}{\sigma^2}\left((X_1-\overline{X})^2+(X_2-\overline{X})^2+\ldots+(X_n-\overline{X})^2\right) = \frac{ns^2}{\sigma^2}} \tag{6.21}$$

なる Z は，自由度 $n-1$ の χ^2 分布 に従う．ただし，ここで s^2 は式 (6.2) で表される標本分散である．

上の命題によれば，何度も標本抽出を繰り返して，その標本分散 s^2 がどのような分布をしているかを見れば，母分散 σ^2 が推定できることになる．

6.5.2 χ^2 分布表の利用

χ^2 分布を統計的な評価に使うときに必要なのは，この関数の，与えられた区間にわたる面積である．図 6.8 を見てほしい．ここで影をつけてある部分の面積 α と t の関係が求められれば，無作為抽出による分散について様々な計算を行うことができる．そこで χ^2 分布 $T_n(x)$ について，α の代表的な値ごとに，それに相当する t を計算したものを数表化したものが用意されている．

χ^2 分布の表の一部を下に示す．

α	0.995	0.975	0.950	0.900	0.500	0.05	0.025	0.01	0.005
$\nu=6$	0.676	1.24	1.64	2.20	5.35	12.59	14.45	16.81	18.55
$\nu=9$	1.73	2.70	3.33	4.17	8.34	16.92	19.02	21.67	23.59
$\nu=10$	2.16	3.25	3.94	4.87	9.34	18.31	20.48	23.21	25.19
$\nu=11$	2.60	3.82	4.57	5.58	10.34	19.68	21.92	24.73	26.76

この表の意味は，例えば $T_6(x)$ において α が 0.995 になるような $x=t$ の位置は 0.676 の点であるということである．

*15 式 (6.11) の $\boldsymbol{X_1, X_2, \ldots}$ を標準化した $\frac{(X_1-\mu)}{\sigma/\sqrt{n}}$ で置き換えることで，この式が得られる．

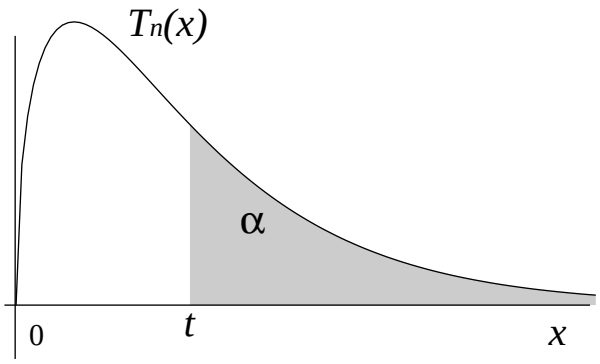

図 6.8　χ^2 分布の数表の意味

たとえば，母集団 $N[0,1]$ から取った大きさ 6 の標本を取ったとしよう，この時，式 (6.11) で与えられる $Z = \frac{1}{6}(X_1^2 + \cdots + X_6^2)$ の分散が 0.676 よりも大きい確率は 99.5% であるということになる．

なお，χ^2 分布の表は正規分布表と異なって，1%, 5%, 90% などのように代表的な α についてのみ t の値を引けるようになっている．実用上はこれで十分だからである．

例題 6–4　($\boldsymbol{\chi^2}$ 分布の実験的検証) 82 ページの表 6.2 のデータから，大きさが 10 の標本を 30 回抽出してみたところ，得られた標本分散は次のようになった（便宜のために結果は昇順に並べ替えてある）．

> 7.02, 8.03, 8.53, 9.34, 13.12, 13.65, 14.17, 14.24, 15.77, 15.83, 16.13, 16.30, 16.52, 16.56, 16.89, 17.41, 17.47, 17.77, 18.25, 18.36, 19.21, 19.48, 20.68, 21.25, 22.91, 24.33, 25.54, 26.24, 26.80, 41.87

この結果から，50% の標本分散が含まれる「切れ目」を求めよ．さらに χ^2 分布を使って，母分散を求めてみよ．

示されているデータを見ると，ある値以上に標本分散 s^2 の 50% 以上が含まれる「切れ目」は，16.89 と 17.41 の間であるから，17.15 である．一方，式 (6.21) より，$Z = ns^2/\sigma^2$ は，自由度 $n-1=9$ の χ^2 分布に従うから，表を見ると，その値が 8.34 のところが「切れ目」に相当している．したがって，$Z = \frac{ns^2}{\sigma^2} = 10 \times 17.15/\sigma^2 = 8.34$．よって母分散は 20.6 となる．一方，9 ページの問題 1–1 の解答にあるように，この母集団の母分散は 17.84 である．こうやってみると，それほど近い値が得られているわけではない．ほどほどの一致というところである．

例題 6–5　($\boldsymbol{\mu}$ が既知の場合) $\boldsymbol{N[3,\sigma^2]}$ に従う母集団から 6 個の標本を多数回無作為抽

出した．その結果，抽出回数の 50% で $\boldsymbol{(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_n-\mu)^2}$ が 21 を超えていた．母分散 $\boldsymbol{\sigma^2}$ を求めよ．

自由度 6 の χ^2 分布の表から，$\alpha=0.5$ となる t の値は 5.35 となることが分かる．つまり式 (6.20) の $Z=\frac{1}{\sigma^2}\left((X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_6-\mu)^2\right)$ が 5.35 のところで，α が 0.5 になるわけである．そこで，$5.35=\frac{1}{\sigma^2}\times 21$ より，$\sigma^2=3.92$ となる．

例題 6–6 ($\boldsymbol{\mu}$ が未知なので標本分散を使う場合) 母平均が未知の正規母集団から大きさ 11 の標本を何回も無作為抽出した．その結果，標本分散の 90% が 12.5 を超えていた．母分散 $\boldsymbol{\sigma^2}$ を求めよ．

この場合には，μ が知られていないので，式 (6.20) は使えない．しかし標本分散は抽出の結果としていつでも求められるものであるから，式 (6.21) を使うことができる．

このとき注意しなければならないのは，標本の大きさ n に対して，その標本分散は**自由度 $\boldsymbol{n-1}$ の $\boldsymbol{\chi^2}$ 分布に従う**ということである．したがってここでは，$\nu=10$ として表を引くことになる．そうすると，$\alpha=0.900$ となるのは $t=4.87$ の時である．この時に，標本分散はちょうど 12.5 になっているというのが題意である．

そこで，式 (6.21) より $4.87=\frac{ns^2}{\sigma^2}$，これに $n=11$, $s^2=12.5$ を代入して $\sigma^2=28.1$ を得る．

【章末問題】

問題 6–1　ある肥料の作物への効果を確かめるために，1 ヘクタール ($10000\mathrm{m}^2$) の畑に 10 m 置きに線を引いて 1 アール ($100\mathrm{m}^2$) の区画を 100 個作った．そうしておいて，5 つおきに区画を選んでその肥料を施して栽培し，他の区画は従来通りの栽培を行うこととした．この選び方は正しいか．

問題 6–2　血液型性格判断，つまり血液型が性格と関係するという考えにもとづく性格判断は，1970 年代にある人とその息子とが相次いで本を書いたことで広まったとされている．著者たちは本に読者アンケートのはがきを付けて返送してもらい，次の本ではそのアンケートにもとづいて所説を展開した．このアンケートの取り方は妥当かどうかを述べなさい．

問題 6–3　ある工場で生産している部品の質量は，平均が 54.2 g，標準偏差は 0.22 g である．この製品を 10 個抜き取って質量を測ったとき，その平均値が 54.1 g〜 54.3 g の間にある確率を求めなさい．

問題 6–4　ある養鶏場から出荷された卵から，12 個ずつ無作為にとって秤量し， これを 30 回繰り返して，その質量 (g) の標本分散を求めたところ，次のように分布していた。結果は小さい順に並べ替えてある．

> 8.76, 9.47, 9.99, 11.85, 12.59, 13.23, 14.79, 18.83, 20.32, 20.74, 21.00, 21.11, 22.40, 23.43, 24.61, 26.14, 27.41, 29.53, 32.22, 33.51, 41.81, 50.57

例題 6–4 にならい，母平均が未知であるものとして，この養鶏場の卵の質量の母分散を χ^2 分布を使って推定せよ（小数第 1 位まで記入）．ただし卵の質量が正規分布していると仮定する．