

例題

ある日、ある酪農家が、搾乳中のホルスタイン5頭の乳量を調べたら、1頭あたりの平均乳量は22.1リットル、標本標準偏差は6.5リットルでした。この農家が飼養しているホルスタインの1頭あたりの乳量（／日）を、信頼係数95%で推定してください。

解：

母分散がわからない上に標本サイズも小さい ($n = 5$) ので、正規分布ではなく t 分布を使って推定します。 t 分布表（表4.1）において、自由度は $n - 1$ ですから、 ν が 4 の行と上側確率 p が 0.025 (5% の半分) の列とが交わる 2.776 という値を使って信頼限界を算出します。すると、信頼係数 95% の信頼区間は 22.1 という平均の $\pm 2.776 \times 6.5 \div \sqrt{4}$ という範囲となります。

よって、この酪農家のホルスタイン1頭あたり平均乳量（リットル／日）に対する信頼係数 95% の信頼区間は (13.1, 31.1) となります。

ちなみに、近年はホルスタインの改良も進み、1日あたりの平均乳量は 20~25 リットルといわれています（1970年頃までは 10 リットルにも満たなかったようです）。

4.5 母比率の区間推定 —選挙速報の「当確」とは？—

● 比率と得票率

テレビやラジオで国政選挙などの特番を視聴していると、投票終了直後の開票間もない時点で**当確**が報道されることがあります。ちなみに当確とは、当確実の略で、その候補者が確実に当選しそうな状況にあると予想されただけです。当確が打たれたのに落選してしまうことも昔はありました。では、この当確はどのように判定されるのでしょうか？ 実は、この当確にも信頼区間の推定が用いられているのです。何の信頼区間かというと“母比率”です。

それでは早速、**母比率の区間推定**の方法を解説したいと思いますが、その前に、そもそも比率とは何でしょうか？ 誰もが知っている言葉ではありますが、ここで改めて定義しておきましょう（あとで分布を考えるときに役立ちます）。

比率 (proportion) とは「ある性質について、集団を構成する要素が、それを持つか持たないかのどちらかのとき、その性質を持つ要素の割合」のことです。式ならば、比率 p は、ある性質を持つ要素の数 x が、全要素の数 n に占める割合のこととして、次のように表すことができます。

$$\text{比率 } p = \frac{\text{ある性質を持つ要素の数}}{\text{全要素の数}} = \frac{x}{n}$$

例えば、テレビ番組の視聴率ならば、「その番組を視聴した世帯数 x が、全テレビ所有世帯数 n に占める割合」となります（視聴することが「ある性質を持つ」ということで、世帯が「要素」となります）。なお、比率の記号には proportion の頭文字から p があてられます。確率の p と紛らわしくてすみません……（確率の方は probability の頭文字です）。

当確の話に戻りましょう。この場合の比率とは、ある候補者の得票率のことです。得票率は、正確には相対得票率といい、その候補者の得票数 x が、投票総数 n に占める割合 x/n です。確実に当選したかどうかは、全てを開票して母集団の得票率（母得票率）をみなければわかりませんが、過ちを犯すこと（誤報を打つこと）をある程度許すならば、事前に得た標本の得票率（標本得票率）を使って母得票率の信頼区間を推定することで当確を判定できます。

区間推定に用いる標本得票率は出口調査 (exit poll) によって得ます。出口調査とは、その名の通り、投票所の出口付近で、投票を終えた人たちに「誰に投票したか」を聞く調査です。よって、標本得票率は「ある候補者に投票したと答えた人数が、聞き取りした総人数に占める割合」ということになります。例えば、100人に出口調査を実施し、10人が候補者Aに投票したと答えたたら、A氏の標本得票率は $10/100$ で 0.1 (10%) です。なお、標本比率の記号には、母比率 p の推定量という意味で \hat{p} をあてます。

● 標本比率の確率分布

さて、標本得票率（標本比率 \hat{p} ）の確率分布を使って母得票率（母比率 p ）の信頼区間を推定するわけですが、一体、標本比率はどのような確率分布に従うのでしょうか？それがわからなければ推定できません。

先ほど、比率の定義で分子を「ある性質を“持つか持たないか”的どちらかの性質を持つ要素の数」といいました。投票でいいかえれば、ある候補者に票を入れたか入れなかつたかのどちらかの性質を持つ投票者の数のことです（もちろん得票率とは前者「票を入れた」性質を持つ要素数です）。これを読んで、第2章で何か似たようなことを学んだ記憶がよみがえりませんか？

そうです“ベルヌーイ試行”です。

少し復習しますと（節2.2）、ベルヌーイ試行は、（コイン投げのように）結果が成功・失敗のように2種類のいずれかしかなく、互いの試行の結果が独立

しており、成功確率が試行を通じて一定である実験のことでした。選挙の投票も、ある候補者に投票するか否かしかなく、誰かの投票行動がほかの投票行動に影響を及ぼすことはなく、その候補者に票を入れる確率は誰でも一定であると仮定すれば、やはりベルヌーイ試行なのです。

そして、ベルヌーイ試行を n 回繰り返したとき、成功回数 x を確率変数とする離散型確率分布が二項分布（母平均は np 、母分散は $np(1-p)$ ）であり、二項分布の n が大きいときには、連続型確率分布である正規分布で近似できました（中心極限定理）。ただし、今回の二項分布の確率変数は単なる成功回数 x ではなく、それを試行回数 n で割った比率 x/n なので、母平均は np を n で割った p 、母分散は $np(1-p)$ を n^2 で割った $p(1-p)/n$ となります。

母分散が n でなく n^2 で割られるのは、比率 x/n の分散 $V(x/n)$ を x の分散 $V(x)$ の形にするとき、 $1/n$ を V の外に出すために分散の性質から 2乗が付くからです ($V(x/n) = V(x)/n^2 \leftarrow$ 62ページ補足の性質①)。なお、この母分散は、標本比率という標本統計量の母分散なので、以降は母誤差分散（その平方根は母標準誤差）と呼びましょう。

以上のことから、標本が大きい場合の標本比率 \hat{p} は、母平均が p 、母誤差分散が $p(1-p)/n$ の正規分布に従うことがわかりました。あとは、正規分布による母平均の区間推定と同じ手順を踏めば、母比率の信頼区間も推定できます。

● 母比率の区間推定

図4.10のように、出口調査で得られた標本得票率（標本比率）を \hat{p}_1 すると、信頼係数が 95 % となる母得票率（母比率） p の区間は、 \hat{p}_1 から上下に母標準誤差 ($\sqrt{p(1-p)/n}$) の 1.96 倍の範囲となります（倍数は z 分布表から読み取ります）。信頼係数を 99 % とするならば、1.96 ではなく 2.58 を用います。

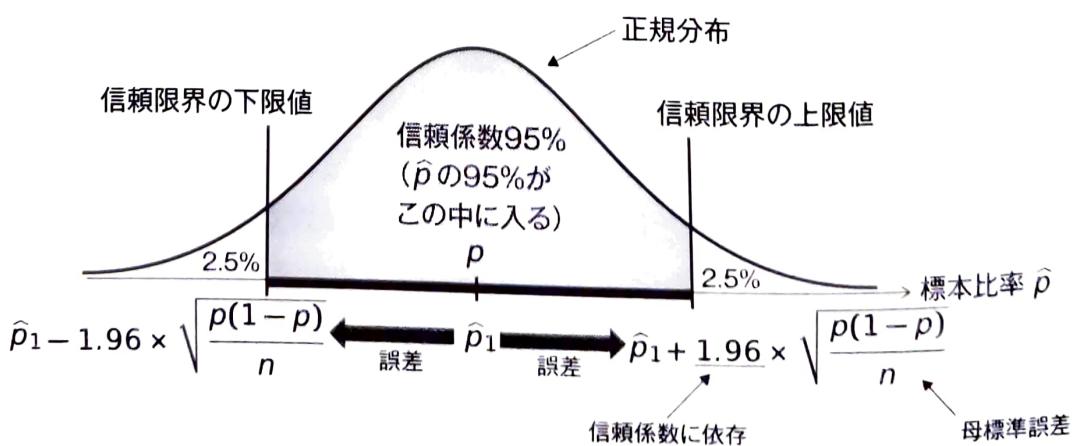


図 4.10 母比率の信頼区間の推定（信頼係数 95 % で母比率 p が既知の場合）

以上、選挙速報の当確判定を事例として、母比率の区間推定を解説してきましたが、残念ながら実際にはこの方法は使えません。なぜならば母標準誤差を求めるのには母比率 p が既知でなければならないからです（母比率の推定をしているのですから既知のはずがありません）。そこで、^{ワルド}Waldという数学者が、標本比率 \hat{p} は母比率 p の推定量なのだから、大きな標本ならば母比率の代わりに標本比率を使用しても、大数の法則から差し支えないだろうと考えました。それが次の Wald 法 (Wald method) と呼ばれる式です。

$$\text{母比率の信頼区間 (大標本)} \quad \hat{p} - 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

この Wald 法は、あくまで大標本のときに使う近似法なので、小標本のときに使うと本来の区間よりも狭く推定してしまいます。自由度によって標準誤差を大きく見積もる t 分布を使って推定すればよいと思われるかもしれませんのが、 t 分布は標本平均の正規分布の代わりの分布なので、標本比率の分布に使うわけにはいきません（そもそも分散の式が異なります）。

そこで、^{アグレステイ}Agresti と ^{クール}Coull という統計学者が、小標本用に次のような修正式を提案しました (**Agresti-Coull 法** ; Agresti-Coull method, あるいは**調整 Wald 法** ; Adjusted Wald method)。一見、Wald 法と同じ見えますが、標本比率 \hat{p} にプライムが付いて \hat{p}' となっている点に注意してください。この標本比率 \hat{p}' は、単なる標本比率の式 x/n を $(x+2)/(n+4)$ に修正したものです。

$$\text{母比率の信頼区間 (小標本)} \quad \hat{p}' - 1.96 \sqrt{\frac{\hat{p}'(1-\hat{p}')}{n}} < p < \hat{p}' + 1.96 \sqrt{\frac{\hat{p}'(1-\hat{p}')}{n}}$$

当確の話に戻りますが、出口調査で観測された標本サイズに合わせてどちらかの式を選び、当確かどうかを知りたい候補者の母得票率の信頼区間を求め、その信頼限界の“下限値”から判定します。なぜならば、どのような選挙でも、下得票率が 0.5 (50 %) を超えれば、その候補者は当選するからです。よって、下得票率が 0.5 を超えていれば、その候補者は過半数の票を 95 % の信頼度で獲得す限値が 0.5 を超えていれば、その候補者は当選できるのです。もちろん議会議員のように複数名当選できる場合には、そのぶん低い下限値でも当確と判断できます。

ただし、実際の報道現場では出口調査だけでなく、地域の担当記者が世帯ごとに取材を行うなど、もっと泥臭い作業も併せて行うことで、当確判定の精度を高めているようです。

例題

ある市長選挙でA氏とB氏が立候補しました。50人に対する出口調査の結果、A氏の得票率は70%でした。この結果からA氏について当確を出してもよいでしょうか？ 信頼係数95%で判断してください。

解：

50人は比較的大きな標本なので、Wald法を使うと、母比率は、

$$0.7 \pm 1.96 \sqrt{\frac{0.7(1 - 0.7)}{50}}$$

の区間に95%の確率で入ると考えられます。

これを計算すると 0.7 ± 0.127 となり、A氏の真の得票率に対する信頼係数95%の信頼区間は（57.3%，82.7%）と推定されます。このうち、下限の信頼限界の値は50%を超えてるので「当確を出してもよい」ということになります。ただし、調査対象者が偏っている（サンプリングバイアスが発生している）と誤判断を招くので注意してください（女性や高齢者、そして農村の投票者は出口調査に非協力的であるという研究論文もあります）。

● 標本サイズの決め方（簡易法）

著者の指導しているゼミでは、毎年、卒論でアンケート調査（意識調査）を実施する学生がいます。そして、その学生からは、決まって「何人ぐらいからアンケートをとればよいですか？」という質問を受けます。もちろん、標本は大きい方がよさそうだというのは学生もわかっているのでしょうかが、街頭調査は骨の折れる作業なので、できるだけ小さい標本（少ない人数への聞き取り）で調査を済ませたいのでしょう。

最後に、区間推定を使った、標本サイズの簡易的な決め方を解説します（検定の標本サイズは難しいので第6章や第10章の検出力分析で扱います）。

さて、アンケート調査では、政策や新商品など、何かに対する意識を聞くことが主な目的となります。ということは、何かに対して賛成の人や、良いなど思った人が全体の何割いるのかという、「比率」に対する統計的な分析がベースとなると考えられます。よって、先ほどまで当確判定で学んだ母比率の区間推定プロセスを使えば、およその標本サイズを求められます。

具体的には、信頼区間の幅を決める $1.96 \times \sqrt{p(1-p)/n}$ の部分を使います。この「（信頼係数95%のz値）×（標本比率の母標準誤差）」は誤差、つまり推

定値と母数とのズレの大きさを表しています。ですから、あらかじめ「この程度の誤差ならば許せる」という適当な許容値を、この誤差に対して設定しておけば、そこから標本サイズ n を逆算できるといわけです。

ただし、母比率 p は未知ですので、類似の調査結果などから想定される標本比率 \hat{p} を使うか、何の手がかりもない場合には、保守的な（大きめの） n を得るために、誤差がもっとも大きくなる 0.5 を入れておきます。 $p \times (1 - p)$ が一番大きくなるのは、 0.5×0.5 のとき（つまり最大は 0.25）であることはわかりますね？

では、事例として、許容できる比率の誤差を5%とした場合の標本サイズを求めてみましょう。 p は不明なので0.5としておくと、95%の確率で信頼できる、母比率の区間推定を可能にするための標本サイズは、次の式から求めることができます。

$$1.96 \times \sqrt{\frac{0.5 \times (1 - 0.5)}{n}} = 0.05$$

$\sqrt{n} = 19.6$ となり、 n は約384となります。よって、384人にアンケートを実施すればよいことになります。例えば384人に対して「現在の内閣を支持しますか?」という質問を行えば、内閣支持率の95%信頼区間を $\pm 5\%$ 以内の誤差で推測することができるというわけです。

もちろん、より高い信頼係数や、より小さい誤差を設定すれば、必要となる標本サイズはもっと大きくなりますし、想定される母比率 p が 0.5 より小さければ、必要となる標本サイズは小さくなります。

なお、母比率ではなく、母平均の区間推定の誤差である「 $1.96 \times \sigma / \sqrt{n}$ 」に許容できる平均の誤差を設定すれば、“母平均”の95%信頼区間の推定を可能にするための標本サイズを得ることができます。ただし、母標準誤差 σ はやはり未知なので、類似の実験結果などから想定される値を用いる必要があります。こちらは標本平均の分布なので、z値（信頼係数95%ならば1.96）ではなくt値（自由度によっても変化する値）を使うこともできますが、そもそも自由度がわからない（というよりこれから決める）状態なのでz値でよいでしょう。

例題

A君は卒論のためにアンケート調査を実施しなければなりませんが、お金も時間もないため、母比率に対する信頼係数95%の信頼区間を推定する際の誤差は10%以内に収まればよいと考えています。さて、A君は何人ぐらいから回答を得られればよいでしょうか？

解：

次の式により、必要な標本サイズ n は約96人になることがわかります。

$$1.96\sqrt{\frac{0.5 \times 0.5}{n}} = 0.1$$

章末問題

問1 ある島に調査に行って、そこに自生しているスギの中から無作為に100本を選び、胸高周囲（成人の胸の高さでの幹周り）を観測したところ、標本平均が2.0 m、標本分散が1.0 m²でした。この島のスギの胸高周囲の母平均に対する信頼区間を信頼係数95%で推定しなさい。

問2 第1章の章末問題で使用した農家データから、この地域の農家の販売金額の母平均に対する信頼区間を信頼係数95%で推定しなさい。

問3 あるペットショップの猫の中から20匹を無作為抽出して血液型を調べたところ、A型の猫が16匹、B型が3匹、AB型が1匹でした（猫の血液型にO型はありません）。このペットショップの猫の中で、血液型がA型である母比率に対する信頼区間を信頼係数95%で推定しなさい。ただし、標本サイズが小さいので、Agresti-Coull法を使うこと。

問4 郵送によるアンケート調査を実施することになりました。とても大事な調査で、かつ予算も十分にあるので、精度の高い調査を実施しようと思います。そこで、誤差が1%となるような調査を目指す場合、何件に調査票を発送すべきかを求めなさい。ただし、返信率は20%とし、返信された回答は全て分析に使えるとする。

χ^2 分布とF分布

χ^2 分布：データの平方和が従う確率分布で、母分散の区間推定や独立性の検定に用いる。

F 分布： χ^2 の比が従う確率分布で、等分散の検定や分散分析に用いる。

5.1 χ^2 分布

● データの平方和

本章では、様々な推定や検定で用いられる統計量である χ^2 （「かいじじょう」と読みます）とF、そしてそれらが従う χ^2 分布とF分布について学んでいきましょう。ボリュームは小さいですが、大変重要な章となります。まずは χ^2 分布から解説しましょう。

前章では、標本平均 \bar{x} が従う正規分布（あるいはt分布）を使って母平均 μ を推定したり、標本比率 \hat{p} が従う正規分布を使って母比率 p を推定しました。

では、母集団の真のバラツキである母分散 σ^2 はどうすれば推定できるでしょうか？ 標本分散 s^2 が従う確率分布を使うのでしょうか？ もし、そうだとしたら、それはどのような確率分布でしょうか？

母分散を推定する場面などあまり想像できないかもしれません。例えば、ある食品工場で生産している製品の袋ごとの容量のバラツキがどのくらいあるのかを知りたいときなど、意外と使うことが多いのです。

答えからいってしまいますと、残念ながら標本分散が従う分布はありません。でも、幸いなことに標本分散（あるいは不偏分散）と比例する統計量が従う分布が考え出されているので、それを使えば母分散を推定できます。それが χ^2 分布（chi-square distribution）です。

χ^2 分布は、F・R・ヘルメルトという、ドイツの測地学（地球の形や大きさを調べる学問）の研究者によって1875年に考え出されました。その後、K・ピ

アソンによって独立性の検定（第11章）などの使い道が見い出され、彼によって命名されました。なお、 χ はアルファベットの x に当たるギリシャ文字で、 χ の2乗などと考えずに χ^2 で1つの単語として扱います（ただの χ という統計量はありません）。

さて、この χ^2 分布に従う χ^2 ですが、実は複数のデータを2乗して全てを足し合わせるだけ（つまり平方和）という、とても単純な統計量です（このあたりが統計量の名称の由来でしょう）。ただし、複数のデータ (x_1, x_2, \dots, x_n) という確率変数に値が入った状態）は、正規分布に従う母集団から、それぞれ独立した試行で得られたものとします（後の図5.2も参照）。

$$\chi^2 \text{ 統計量 (標準化前)} \quad \chi^2 = x_1^2 + x_2^2 + \cdots + x_n^2 = \sum_{i=1}^n x_i^2$$

例えば、-2と5というデータが実験から得られたとすると、 $(-2)^2 + 5^2$ で“29”が χ^2 値となります。このように χ^2 はデータを2乗しているために負にはならず、またそれらを全て足し合わせているのでデータの数（自由度）が多くなるほど値も大きくなりやすい性質を持っています。つまり、 χ^2 が従う χ^2 分布は（t分布と同じように）自由度によって形が変わるので（自由度 ν が χ^2 分布の唯一の母数です）。

χ^2 とはこのような単純な統計量なのですが、対象が異なれば尺度・単位も違ってくるため、たとえデータ数が同じ場合でも、値と確率との対応関係が一律ではなくなり1つの分布表に整理できません。それでは不便なので、どのような対象に対しても同じ分布表を使えるように、普通は、元のデータ x を標準化した標準化変量 z の平方和を χ^2 値とします。

標準化を復習しますと、次のような式で個別のデータ x を変換することで、データ全体（母集団）の平均 μ をゼロ、標準偏差 σ を1に揃えることでした（節2.4）。

$$\text{標準化変量 } z = \frac{x - \mu}{\sigma}$$

この標準化の式を使って、まずは母平均 μ が既知の場合の χ^2 を定義しておきましょう（母平均が未知の場合については次節で扱います）。

χ^2 は、この z に対してデータの数だけ平方和すればよいので、母集団から抽出するデータの数が1つの場合は、次のようになります。

$$\chi_{(1)}^2 = z^2 = \frac{(x - \mu)^2}{\sigma^2}$$

下付きのカッコ内の数字 (1) は自由度 ν を示しているのですが、いまのところ式の中で標本平均 \bar{x} ではなく母平均 μ を使っているため何の制約もかかりず、自由度はデータ数そのものです (自由度 $\nu=n=1$)。

同様にデータが2つ (自由度 $\nu=n=2$) の場合の χ^2 は、次のようにになります。

$$\chi_{(2)}^2 = z_1^2 + z_2^2 = \frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2}$$

自由度が3以上の場合も同様なので、総和記号 \sum を使うと、自由度 n の χ^2 は、次のように表すことができます。

$$\chi^2 \text{ 統計量 (母平均が既知)} \quad \chi_{(n)}^2 = \sum_{i=1}^n z_i^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2}$$

● χ^2 が従う確率分布

χ^2 が従う確率分布を描いたのが図5.1です (とりあえず1, 3, 10という3種類の自由度の分布を描きました)。連続型で右側に歪んだ形をしてします。

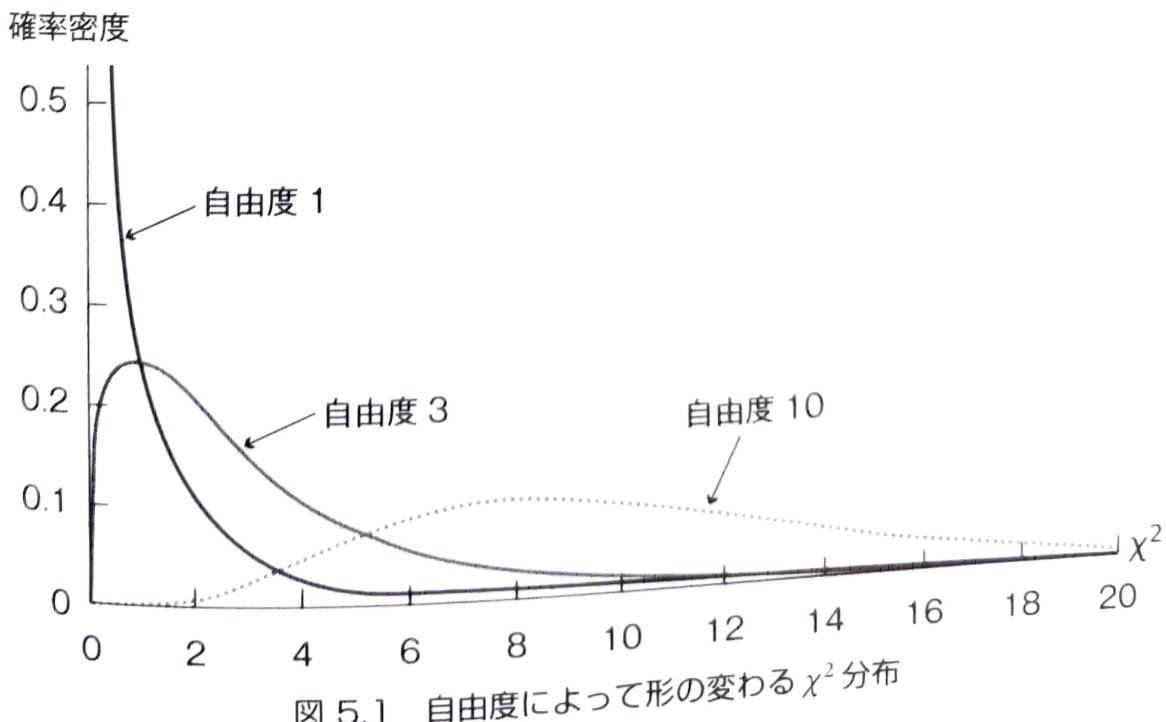


図 5.1 自由度によって形の変わる χ^2 分布

また、 χ^2 分布の確率密度関数式は、次のようになります。

$$\chi^2 \text{ 分布の確率密度関数} \quad f(\chi^2) = \frac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}}$$

t 分布と同様、このような複雑な式を覚える必要は全くありませんが、自由度 ν が式に入っています。それ以外はネイピア数 e や Γ 関数（階乗を一般化したもの）などの定数しかないことから、自由度 ν のみが χ^2 分布の形を決める母数であることが確認できます。

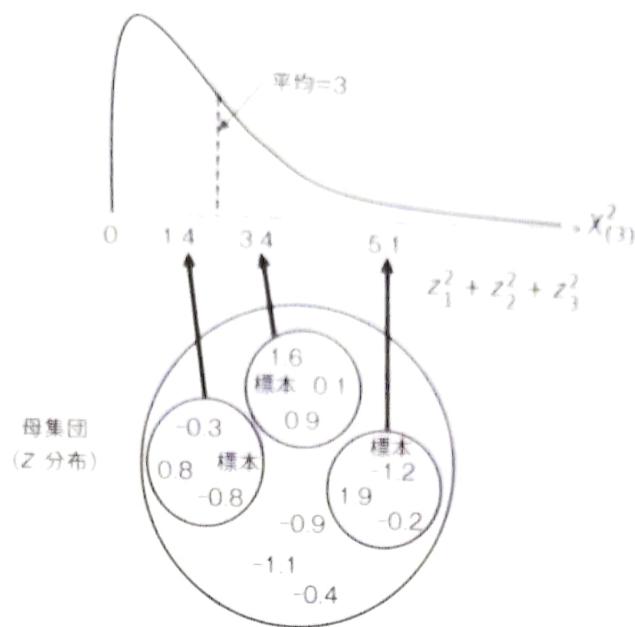
また、 χ^2 分布の平均は自由度 (ν)、分散は自由度の2倍 (2ν) となります。平均も分散も、第2章で積分を使って示した連続確率分布の期待値(平均)と分散の式と、 Γ 関数の性質 ($\Gamma(\nu) = (\nu - 1)\Gamma(\nu - 1)$) を使えば導き出せますが、入門書である本書では省略させていただきます。

● 分布の歪みと平均とのズレ

さて、図 5.1 をもう一度見てください。自由度 1 の χ^2 分布曲線だけが、 $\chi^2=0$ に近い左裾が高い形状になっているのを不思議に感じたかもしれません（実は自由度 2 まではこのような形状をしています）。これは、確率密度関数式の χ^2 の指数部分が自由度 $\nu/2$ から 1 を引く内容になっているからなのです。例えば χ^2 が 0 に近い $\chi^2=0.01$ の場合で考えてみましょう。 $(\chi^2)^{\nu/2-1}$ の部分の値は、自由度 $\nu = 3$ のとき $0.01^{0.5}$ で 0.1 ですが、 $\nu = 2$ では 0.01^0 で 1 となり、 $\nu = 1$ では $0.01^{-0.5}$ で 10 と急に大きくなります。つまり、自由度が 1 や 2 のときの χ^2 の値は、0 に近づくほど大きくなるのです。

また、分布曲線のピークが平均よりもやや左にあることが気になったかもしれません（例えば自由度 10 の分布では平均の 10 ではなく、8あたりにピークが来ています）。これは、分布曲線は単に確率変数値に対応する確率密度を描いているのに対して、平均である期待値は確率密度に確率変数値を乗じたものを積分して足し合わせているからです（第2章の復習）。つまり、正規分布のような左右対称の確率分布ならば分布曲線のピークと平均とが一致しますが、 χ^2 分布のように左右非対称の分布曲線の場合、その形状からは平均の位置を知ることはできないのです。ただし、自由度が大きな χ^2 分布は正規分布に近づくため、ピークと平均とがほぼ一致します。

著者の統計学の授業では、大抵、このあたりで「なぜこのような分布になるのか全くわからない」という反応が返ってきます。みなさんにとって、 χ^2 分布

図 5.2 自由度 3 の χ^2 値（母平均が既知の場合）

がしっくりこない理由ですが、正規分布のように自然界に存在する（というよりは様々な自然・社会の現象の起きる確率を近似できる）ものではなく、推定や検定のために作り出された分布だからでしょう。

ですからダメ押しで、もう少し説明させてください。例として、自由度 3 の χ^2 分布を具体的に考えてみましょう。

母平均 μ が既知の場合の χ^2 の自由度はデータ数（標本サイズ） n ですから、図 5.2 のように z 分布に従う母集団から 3 つのデータが独立に無作為抽出される状況になります。つまり、第 3 章で説明した標本平均と同じように、3 つのデータからなる標本を何度も抽出することができます。そしてこのたくさんの中の標本から、たくさんの χ^2 を得られます。それらは同じ値になる場合もあるでしょうが、微妙に異なることが多いでしょう（図 5.2 では、たくさんは描けないので、3 つの標本から 1.4, 3.4, 5.1 という 3 種類の χ^2 の値が得られた状況を示しています）。つまり、分布するのです。それが χ^2 分布です。

● 食い違い指数

ここまで説明で、 χ^2 分布とは何ぞやということについてはたいたい理解していただけたと思いますが、この z の平方和の分布が母集団のバラツキ（母分散や母標準偏差）の推定に役立つ理由について再確認したいと思います。

複数のデータがあり、それらの標準化変量 z の 2 乗を集めたのが χ^2 ですか。抽出元の母集団のバラツキが小さければ、 z の平均である 0 に近いデータの割合が高くなるため、 χ^2 は大きくなりにくいでしよう。逆に、抽出元のバラ

ツキが大きければ0から離れたデータの割合が高くなるため、 χ^2 は大きくなりやすいでしょう。つまり、 χ^2 は標本のバラツキと連動しているのです（次節では式を展開してこれを証明します）。これが、 χ^2 分布を母分散の推定に使える理由です（標本分散が大きければ、母分散も大きいでしょう）。

また、このように手元のデータが標準値からどれぐらい離れているのかを確率的に評価できる χ^2 の性質は大変便利なため、母分散の推定以外にも様々な検定で用いられています。というのも、比較する標準値を0でなく、何らかの仮説が正しい場合に得られると期待できる値とすれば、現実のデータが、それ（仮説）とどれぐらいズレているのかを捉えることができるからです。ですから χ^2 は「食い違い指数」などとも呼ばれます。とくに第11章の「独立性の検定」では、期待度数（仮説の下で得られるはずの件数）と観測度数（実験で得られた件数）とのズレを捉えるのに χ^2 を用います。仮説や検定の考え方は次章で詳しく解説しますので、とりあえずここでは χ^2 がズレを捉える統計量として、多方面で活躍することを覚えておいてください。

5.2 母分散の区間推定

● 不偏分散との比例関係

第4章では、正規分布や z 分布、 t 分布を使って母平均 μ や母比率 p の信頼区間を推定しました。ここでは、 χ^2 分布を使った母分散 σ^2 の信頼区間の推定について学びます。

前節では、 χ^2 をわかりやすく説明するため、母平均 μ が既知であることを想定して、自由度がデータ数（標本サイズ n ）そのままの χ^2 を使いました。しかし、実際に母分散 σ^2 を推定する場合、母平均 μ は未知でしょうから、 μ の代わりにその標本統計量であり不偏統計量である標本平均 \bar{x} を用いることになります。標本平均 \bar{x} を1つ使用すると、節3.2や節3.6で学んだように、自由度は μ を使う場合に比べて1つ減って $n - 1$ となります。つまり、母分散 σ^2 を推定する場合の（母平均が未知の） χ^2 の式は次のようになります。

$$\chi^2 \text{ 統計量 (母平均が未知)} \quad \chi_{(n-1)}^2 = \frac{\sum(x_i - \bar{x})^2}{\sigma^2}$$

この式をよく見てください。何か見覚えがある式だと思いませんか？

次に示す不偏分散 $\hat{\sigma}^2$ の式（節3.2）と分子の $\sum(x_i - \bar{x})^2$ が同じですね。

$$\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

ということは、両式の分子を整理して、

$$\sigma^2 \times \chi^2_{(n-1)} = (n - 1) \times \hat{\sigma}^2$$

という関係が成り立ちます。これを χ^2 について解けば、

$$\chi^2_{(n-1)} = \frac{(n - 1) \times \hat{\sigma}^2}{\sigma^2}$$

となり、前節で解説したように、 χ^2 が母集団のバラツキの不偏推定量である不偏分散 $\hat{\sigma}^2$ と比例関係にあることが証明されました (σ^2 は定数)。

また、右辺分子の分散において自由度 $n - 1$ を制約がかかる前の n に戻せば、不偏統計量でない標本分散 s^2 とも (χ^2 が) 比例関係にあることもわかります。

$$\chi^2_{(n-1)} = \frac{n \times s^2}{\sigma^2}$$

よって、不偏分散でも標本分散でもよいのですが、両方説明すると紛らわしいので、以降では前者（不偏分散 $\hat{\sigma}^2$ ）を使ってた式で母分散 σ^2 の区間推定を進めたいと思います。

さて、1つ前の不偏分散 $\hat{\sigma}^2$ との比例関係を示した式について、右辺分母である母分散 σ^2 の式に変形すれば、次のようにになります。

$$\sigma^2 = \frac{(n - 1) \times \hat{\sigma}^2}{\chi^2_{(n-1)}}$$

ここで、右辺分母の χ^2 は分布するため幅を持ちますから、分子の不偏分散 $\hat{\sigma}^2$ を標本から計算すれば、母分散 σ^2 を推定できるというわけです。

● 母分散の区間推定

それでは、母分散 σ^2 に対する信頼係数 95 % の信頼区間を推定してみましょう。例として、自由度が 4 (標本サイズ $n = 5$) の場合を考えます (図 5.3)。

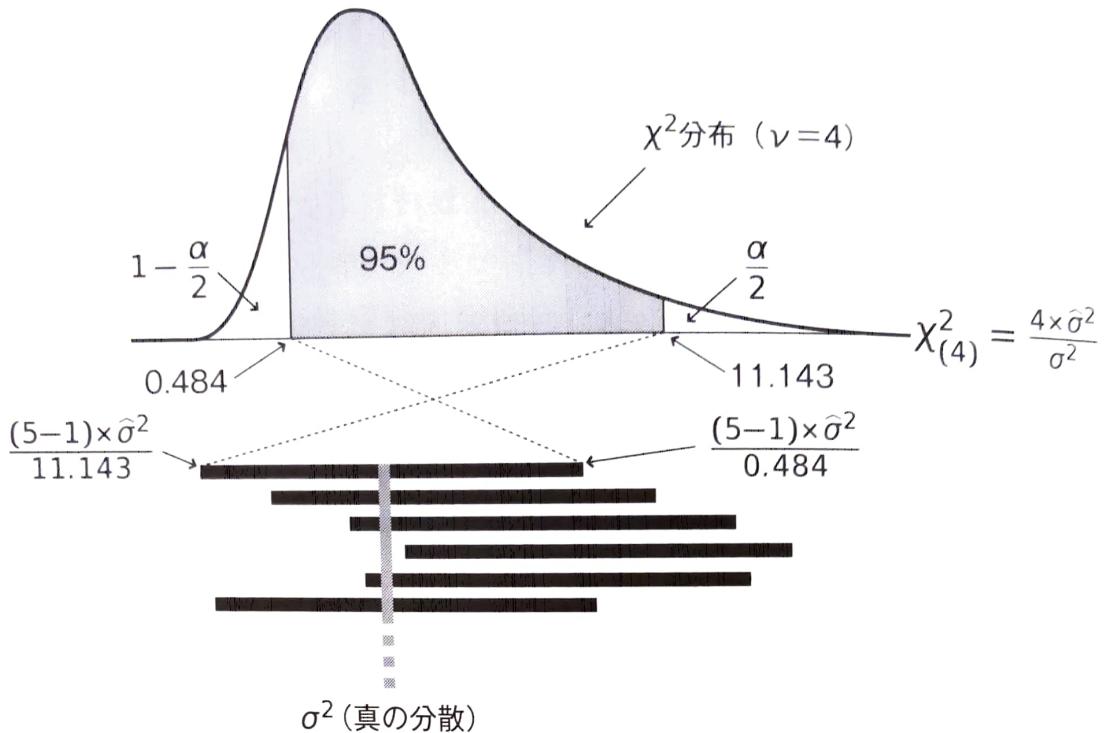


図 5.3 母分散の信頼区間の推定

注： χ^2 分布の横軸がそのまま母分散の信頼区間になるわけではない。

まず、先ほどの母分散 σ^2 の式を、母分散を含むと（信頼係数 $1 - \alpha$ で）期待できる信頼区間を表す連立不等式に変形してみると次のようになります。

$$\text{母分散の信頼区間} \quad \frac{(n-1) \times \hat{\sigma}^2}{\chi_{(n-1, \frac{\alpha}{2})}^2} < \sigma^2 < \frac{(n-1) \times \hat{\sigma}^2}{\chi_{(n-1, 1-\frac{\alpha}{2})}^2}$$

この左辺分母の $\chi_{(n-1, \alpha/2)}^2$ は、母分散に対する信頼区間の信頼限界の下限値（区間左側の小さい方の値）を計算するとき、右辺分母の $\chi_{(n-1, 1-\alpha/2)}^2$ は上限値（区間右側の大きい方の値）を計算するときの χ^2 の値をそれぞれ表しています。不等式において、 χ^2 は分母にあるため、上／下限値の計算では一見、逆に用いているように見えることに注意してください（それが図 5.3 で破線がクロスしている理由です）。添字の α は χ^2 分布の上側の確率を表す記号です。本例では信頼係数 95 % ですから、100 % から 95 % を引いた 5 % ($\alpha = 0.05$) の半分になります。このように、 χ^2 分布は、 z 分布や t 分布と異なり左右非対称なので、上側と下側それぞれの χ^2 の値を読み取る必要があります。

2つの信頼限界を計算するための χ^2 の値は、分布表や Excel 関数で得ることができます。表 5.1 に χ^2 分布表の一部を抜き出しましたので、ここから値を読み取ってみましょう（全体表は巻末に付録 III として掲載）。

表 5.1 χ^2 分布表の一部（表頭の上側確率と表側の自由度に対応する χ^2 値）

ν	p	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1		0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2		0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3		0.072	0.115	0.216				9.348	11.345	12.838	
4		0.207	0.297	0.484				11.143	13.277	14.860	
5		0.412	0.554	0.831				12.833	15.086	16.750	
6		0.676	0.872	1.237					16.812	18.548	
7		0.989	1.239	1.690					18.475	20.278	
8		1.344	1.646	2.180					20.090	21.955	
9		1.735	2.088	2.700							
10		2.156	2.558	3.247	0						

χ^2 分布表には、所与の上側確率 p に対応する χ^2 の値が自由度 ν ごとに掲載されています。さて、2つの信頼限界のうち、上限の計算に使う χ^2 値は、 $1 - \alpha/2$ の確率 p の列、下限は $\alpha/2$ の確率 p の列から読み取ります。事例では α は 5% なので、信頼限界に使う 2 つの χ^2 のうち、上限の計算に使う分布下側の χ^2 の値は、確率 $p(1 - \alpha/2) = 0.975$ の列と自由度 $\nu = 4$ がクロスする 0.484、下限の計算に使う分布上側の χ^2 の値は、確率 $p(\alpha/2) = 0.025$ の列と自由度 $\nu = 4$ がクロスする 11.143 であることが読み取れますね。また、Excel 関数ならば分布下側の χ^2 の値は $\text{CHISQ.INV}(\alpha/2, n-1)$ で、分布上側は $\text{CHISQ.INV.RT}(\alpha/2, n-1)$ で求めることができます。

以上、自由度が 4 (標本サイズが 5) の場合、母分散 σ^2 に対する信頼係数 95% の信頼区間は次のようになります。なお、母標準偏差 σ の信頼区間は、この連立不等式の平方根を取るだけです。

$$\frac{(5-1) \times \hat{\sigma}^2}{11.143} < \sigma^2 < \frac{(5-1) \times \hat{\sigma}^2}{0.484}$$

例題

A という種のテントウムシを 5 匹採集しました。それらの体長は、小さい順に 5 mm, 8 mm, 10 mm, 11 mm, 15 mm でした。このテントウムシの体長の母分散に対する信頼区間を信頼係数 95% で推定してみてください。

解：

母平均の信頼区間の推定と同様に、この例では 5 匹の標本から母分散 (母集団の分散)、つまり誰も知らない真の分散を推測します。

第5章 χ^2 分布と F 分布

まず、採集した5匹の標本から不偏分散を計算すると、節3.2の計算式から、

$$\frac{((5 - 9.8)^2 + (8 - 9.8)^2 + (10 - 9.8)^2 + (11 - 9.8)^2 + (15 - 9.8)^2)}{(5 - 1)} = 13.7$$

となります。また、信頼係数は本文中と同じ95%で自由度も4なので、信頼限界の下限値の計算に使う χ^2 値は11.143、上限値のほうは0.484となり、次の不等式を解けばよいことになります。

$$\frac{(5 - 1) \times 13.7}{11.143} < \sigma^2 < \frac{(5 - 1) \times 13.7}{0.484}$$

すると、このテントウムシの体長の母分散の95%信頼区間は(4.92, 113.22)となります。単位を付けるとすればmm²です。また、これらの値の平方根を取った(2.22, 10.64)は、母標準偏差 σ の95%信頼区間となります(単位はmm)。