入門 統計学　第4章

信頼区間の推定

『入門 統計学 第2版　一検定から多変量解析・実験計画法・ベイズ統計学まで一』(オーム社)

※注：本書を購入された方へのサービスですので，教科書指定(参考図書は不可)していない授業での使用はお控えください。

信頼区間の推定とは？

- 点推定の復習：母集団の分布の特性（母平均や母分散）を標本から1つの値で推定
- 点推定の欠点：推定結果がどれぐらい信頼できるかが不明
- 解決策：標本統計量に，信頼性に応じた幅を取って（母数を含む区間を）推定するとわかりやすい
- →それが信頼区間の推定（区間推定と略す）

区間推定に入る前に…

4.1　標本平均に関する2つの定理

- 標本平均は標本サイズが大きくなるに伴い，次のような振る舞いをする

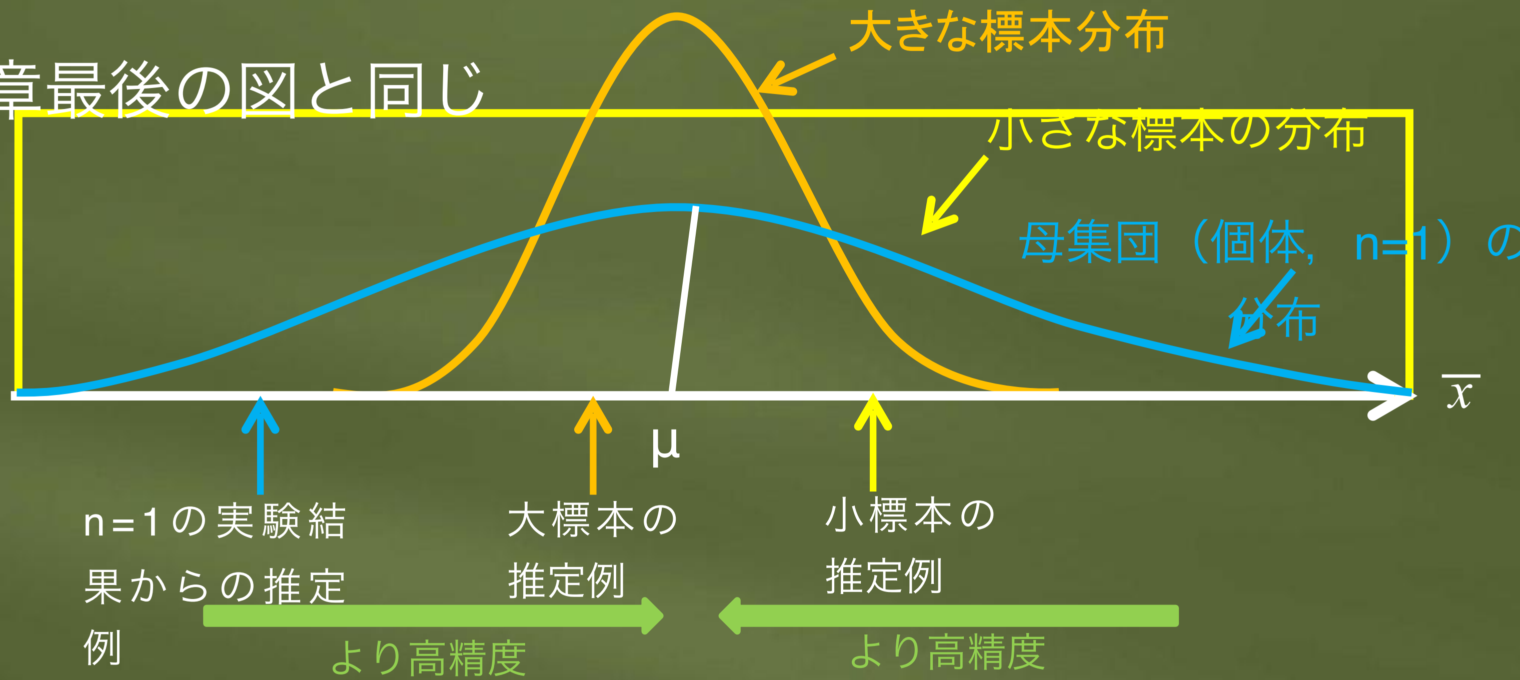
①真の値である母平均に近づく（大数の法則）
→標本サイズが大きな実験ほど推定精度が向上する（前章の復習）

②母集団の分布に関係なく，標本平均は正規分布に近づく（中心極限定理）
→各種推定で正規分布が多く使われるので便利

この2つの法則を土台として母数の信頼区間を推定する

大数の法則

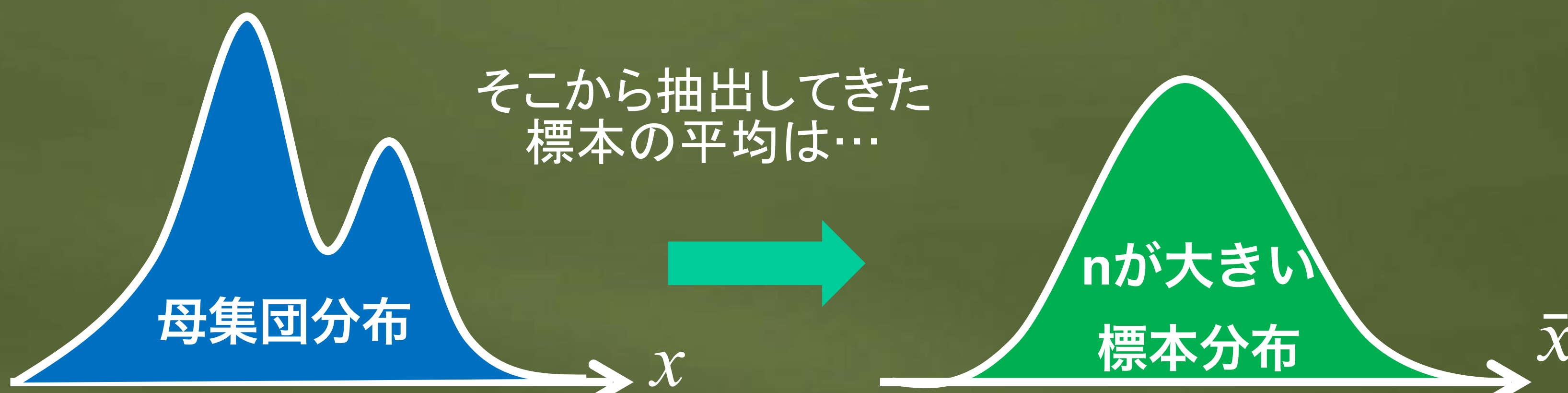
前章最後の図と同じ



より大きい標本から求めた推定値は，真の値に近づく

（母平均の信頼区間の中心に標本平均を使っても良い根拠となる）

中心極限定理



正規分布に従っていなくても…

正規分布に近づく

母集団の分布が何だろうと心配せず，
標本平均が正規分布に従うとして区間推定や検定を実施できる

4.2 区間推定の用語整理

- 区間推定：標本統計量に，推定の信頼性に応じた幅を持たせて，母数を含む信頼区間を推定すること

区間推定の用語の整理
（テントウムシの真の体長の推定例）

信頼区間

××テントウの真の体長は，95%の確率で，○○ ～ ○○の区間に含まれる。

母平均

信頼係数

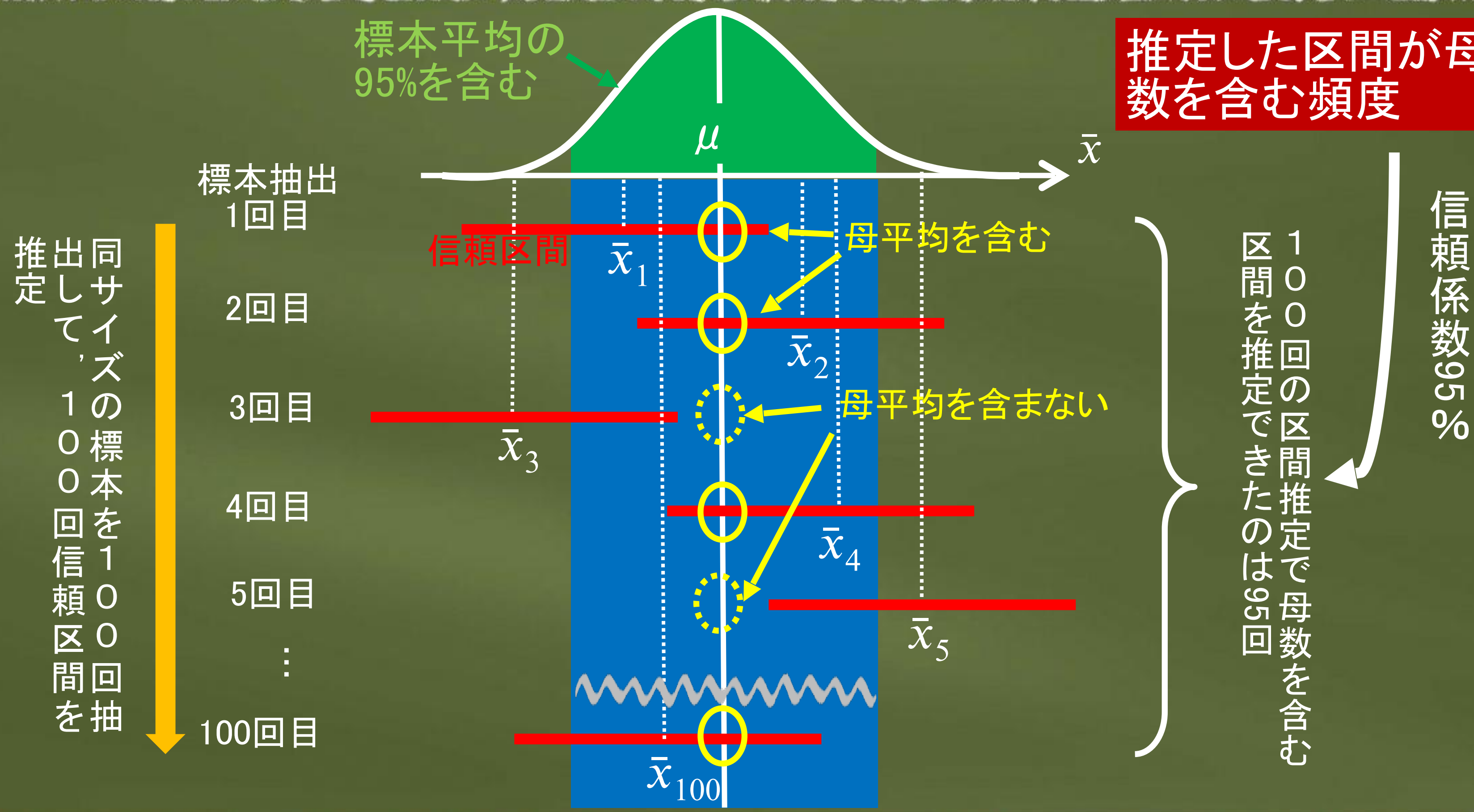
信頼限界
（下限値）

信頼限界
（上限値）

信頼性の高さを表す指標
（重要な概念なので次に解説）

信頼係数

（推定の信頼性の高さ）

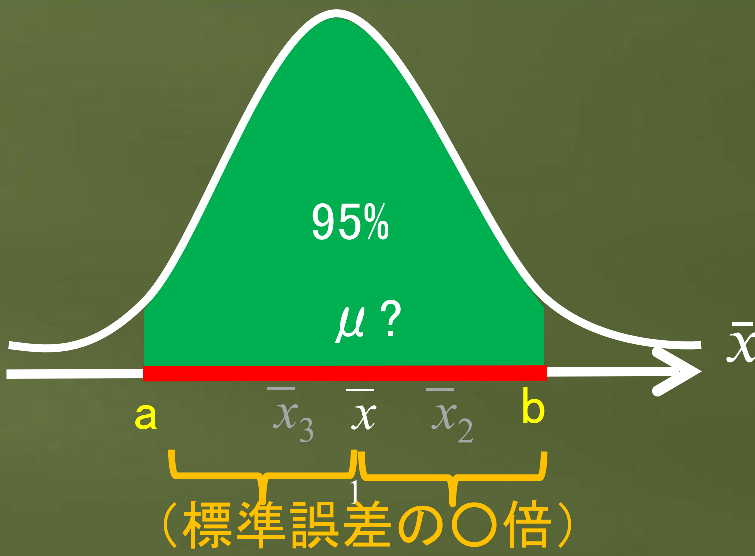


4.3　信頼区間推定の大まかな手順

ステップ①：母平均μの値は分からないので，とりあえず1回目の実験の標本平均$\bar{x}_1$をおく

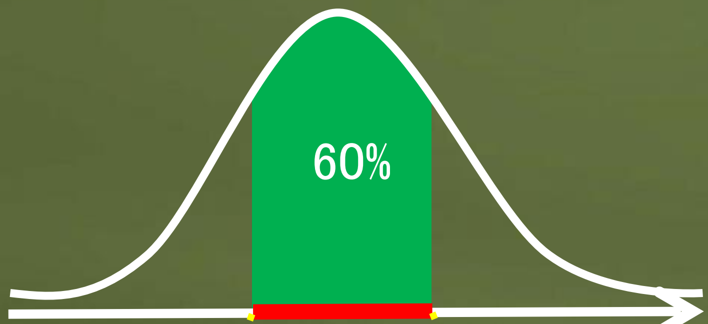
ステップ②：何回か観測しても（実際には1回），それらの標本平均はまあまあ近い値になると考えられる

ステップ③：何回目の観測がより正確かという情報はないので，1回目の標本平均$\bar{x}_1$を中心とした区間を考える



ステップ④：信頼係数に対応した信頼限界の上下値（b,a)を求める　←この後，詳しく解説

ところで，なぜ信頼係数は95％？



狭い区間の方が面白いが，滅多に母数が含まれないのでは意味が無い

とても広い区間には母数が含まれて当然

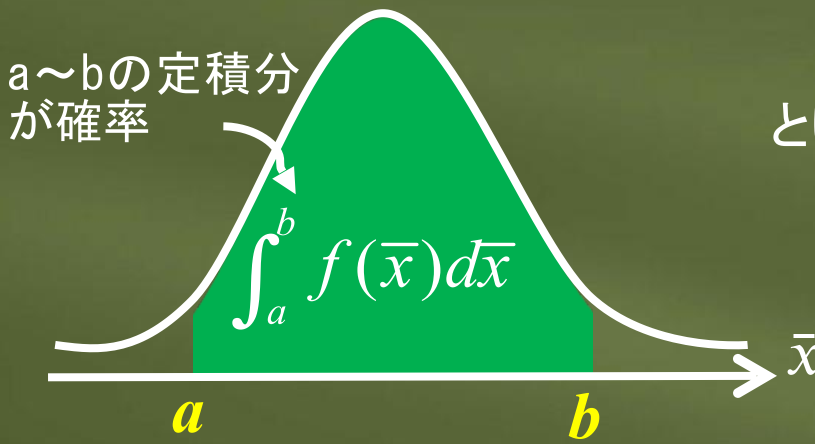
母数が含まれる区間を予想しているのだから幅は狭い方が面白いが，母数が滅多に含まれない（信頼性の低い）区間を推定しても意味が無い

そこで，信頼係数を高くしたいが，あまり高いと区間の幅が広くなりすぎて，母数が含まれて当然の面白くない結果となる

妥協点として，切りの良い95％の信頼係数がもっとも用いられる（標準誤差が小さければ99％）

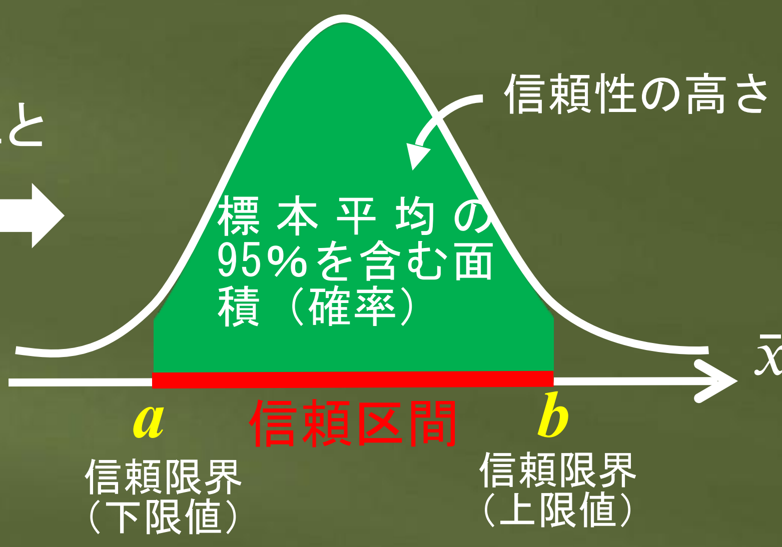
信頼限界の求め方

所与の2点から，それに対応する確率が求まる



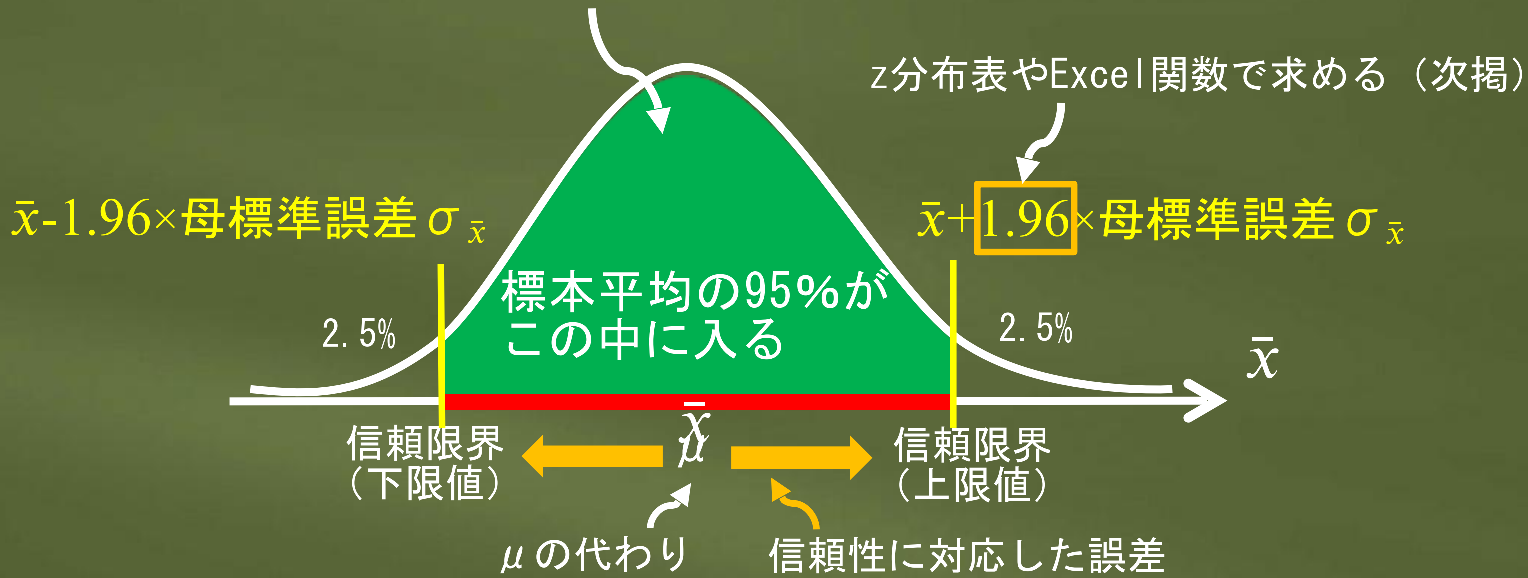
ということは…

所与の確率から，それに対応する2点（信頼限界）が求まる



（標準化していない）正規分布を用いた母平均の区間推定

推定の信頼性の高さである信頼係数が95%とは，標本を繰り返し抽出した場合，その平均の95%が含まれるということ



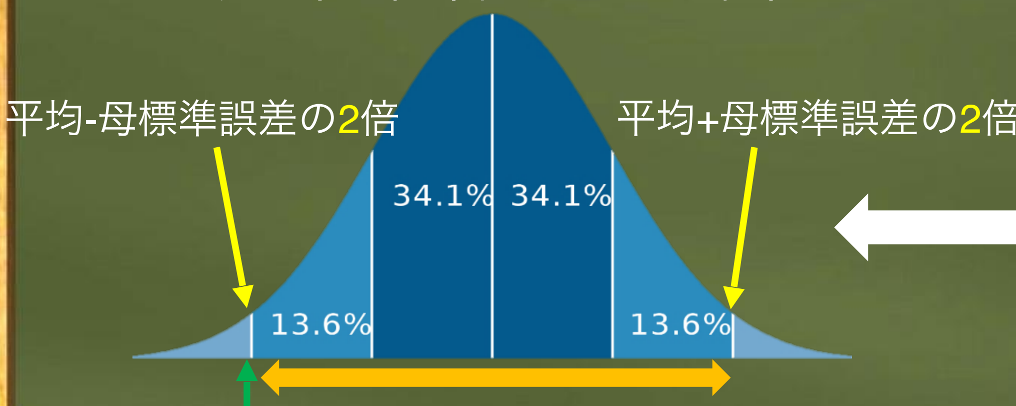
母平均の信頼区間
（母分散が既知）

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

母分散がわかっている必要

母標準誤差の“1.96”倍の見つけ方

正規分布の標準誤差ごとの確率



平均±母標準誤差の２倍の内側の面積（確率）が95.4%

（2倍よりもう少し小さい

Excel関数＝NORM.S.INV(0.025)で，-1.96というz分布（平均0のとき）の下限値が返ってくる

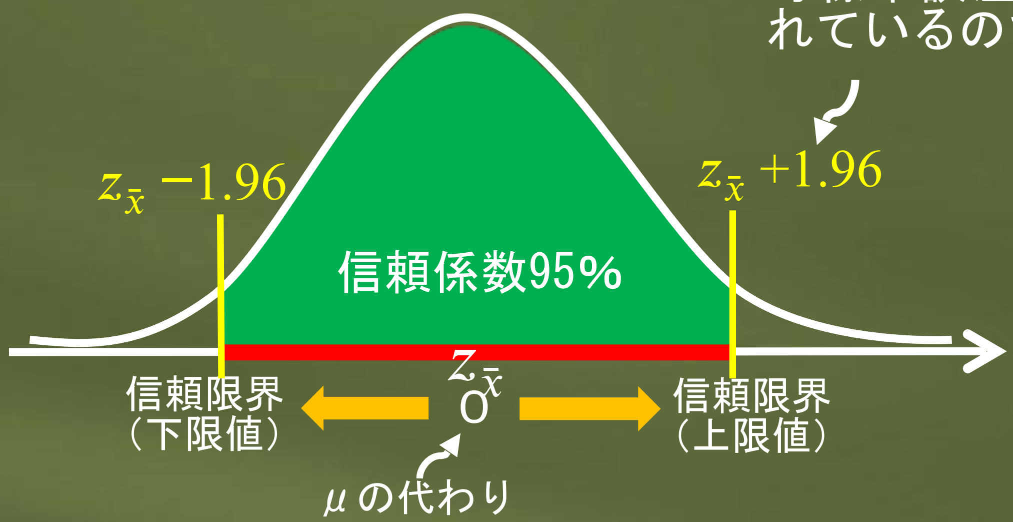
パソコンがなければz分布表から読み取る

z分布表	0.00	0.01	0.02	0.03	0.04	0.05	0.06
...	…	…	…	…	…	…	…
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250

上側確率である表中の値が0.025になるz値（1.9の行と0.06の列がクロスするところ)

標準正規（z）分布を用いた
母平均の区間推定

母標準誤差が1に変換されているので単純になる



母平均

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

母標準誤差（母分散が既知の必要）

母平均の信頼区間（母分散が既知）

$$z_{\bar{x}} - 1.96 < 0 < z_{\bar{x}} + 1.96$$

母平均はゼロに変換されている（わかりにくいので変形してみよう）

（z分布による母平均の）信頼区間の式の変形

$$z_{\bar{x}} - 1.96 < 0 < z_{\bar{x}} + 1.96$$

zの連立不等式にする

$$-1.96 < z_{\bar{x}} < 1.96$$

z式を展開する

$$-1.96 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

母平均μの連立不等式にすると…

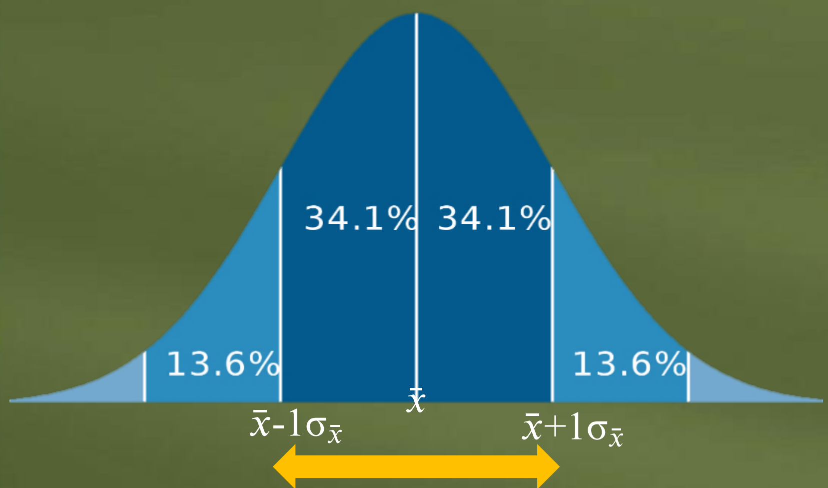
$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

標準化していない
正規分布と同じ

右のように表現すること
もある（［］の場合
もあり）→

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

ということは，前章で学んだ母平均の点推定値に母標準誤差を併記するということは…



標本平均±母標準誤差の1倍

信頼係数68.28%※の信頼区間を推定していることと同じ意味であったことがわかる

→指定した信頼性の高さの下，推定結果の幅を示せる区間推定の方が優れている（わかりやすい）

※注：不偏標準誤差の場合には，自由度に対応して信頼係数の値はやや下がるが，そもそも母分散が不明なので正規分布を前提できない（次節のt分布を考える）

例　題

ある園芸農家が栽培しているカーネーションのなかから16本を抽出して蕾（つぼみ）の直径を調べたら，平均で10.0mmであった。

この農家が栽培している全てのカーネーションの平均直径を信頼係数95％で推定しなさい。ただし，その母分散は36.0㎟だとする。

解：母分散は既知なので，後は信頼係数95％の信頼限界を計算するためのz値を分布表かExcel関数などで得るだけ。

zは1.96なので，下記式から，蕾の平均直径に対する信頼係数95％の信頼区間は（7.06mm,12.94mm）となる。

15mm以上が出荷基準なので，出荷には早いといえる

$$\bar{x} - 1.96\cdot\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \rightarrow \quad 10 - 1.96\frac{\sqrt{36}}{\sqrt{16}} < \mu < 10 + 1.96\frac{\sqrt{36}}{\sqrt{16}}$$

4.4　*t* 分布による区間推定

ー母分散が未知で小標本の場合ー

- （標準）正規分布による区間推定は母分散が分からないと使えない（大標本ならば標本分散を母分散と見なして用いるのもあり）

　→標本データだけから推定する事を考える

　→（スチューデントの）t分布の出番！

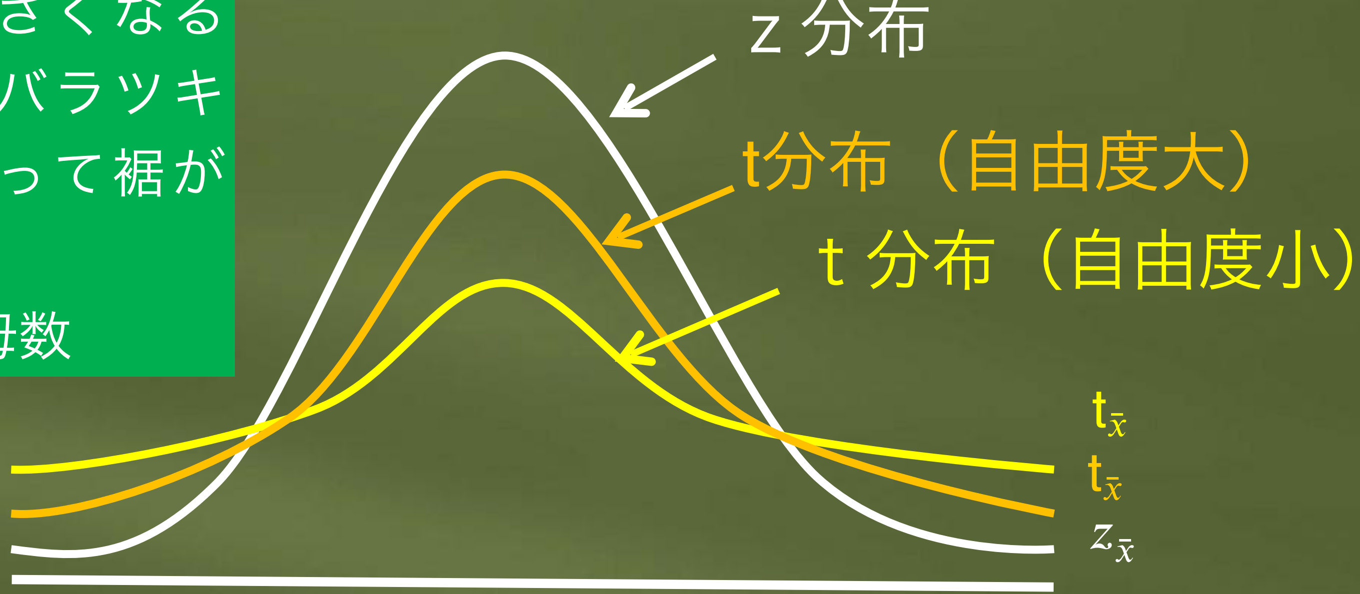
- t 分布：自由度（n-1）に応じてバラツキが変化（小標本のときに誤差を大きく評価）

t分布を考案したゴセット（ペンネームがスチューデント）→

W. S. Gosset
(1876〜1937)

t 分布のバラツキは 自由度（標本サイズn-1)で変化

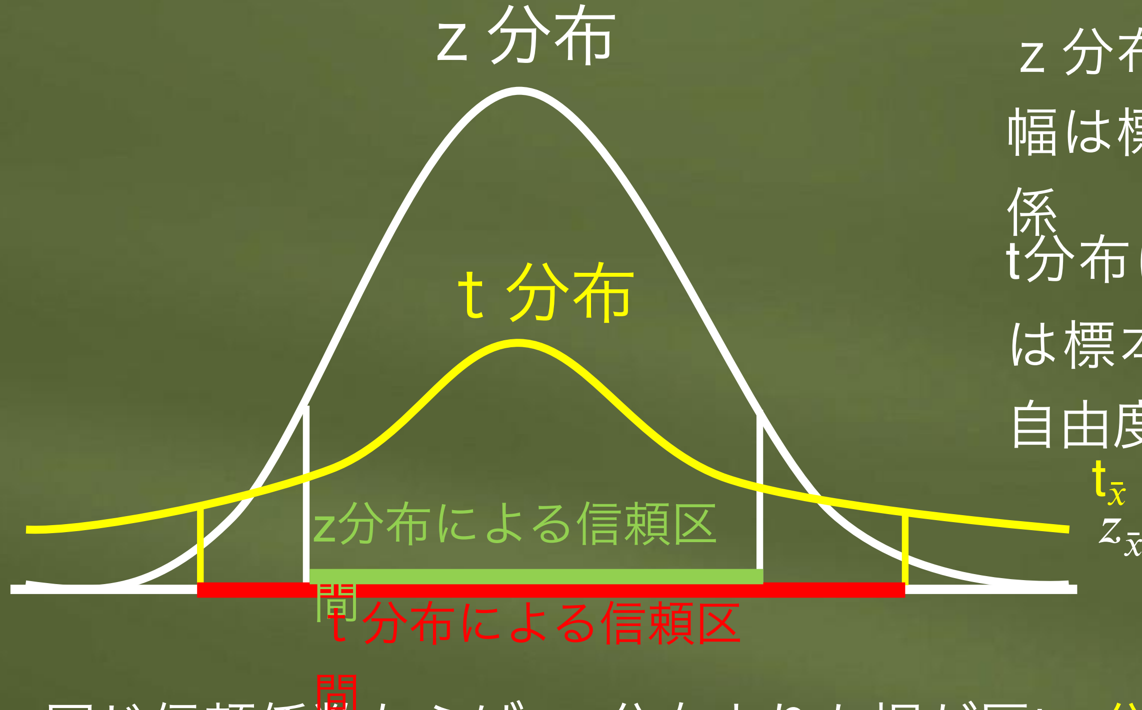
自由度が小さくなるほど分布のバラツキが大きくなって裾が厚くなる
→自由度が母数



z 分布よりもやや尖っていて裾が厚い

自由度（n-1)が大きくなると標準正規分布に近づく

t 分布の信頼区間は広くなる



z 分布による信頼区間の幅は標本サイズとは無関係

t分布による信頼区間の幅は標本サイズ（正確には自由度）に応じて変化

同じ信頼係数ならば，z分布よりも裾が厚いt分布の方が信頼区間は広くなる（保守的な推定となる）

小標本の推定は精度が低いので，それを反映できる

t 分布に従う“ t ”とは？

標準正規分布（z分布）に従うzの復習：

標本平均の
標準化変量

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

標本データからは計算できない

標本平均の
準標準化変量

$$t_{\bar{x}} = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \text{ or } \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

標準化に準じた変換（スチューデント化）を施している

母標準誤差の代わりに標本データから計算できる不偏標準誤差を用いている

t 分布の確率密度関数

（復習） z 分布の確率密度関数

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}$$

nは関係ない

t 分布の確率密度関数

$$f(t)=\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\cdot\left(\frac{1+t^2}{n-1}\right)^{-\frac{n}{2}}$$

自由度n-1が関係

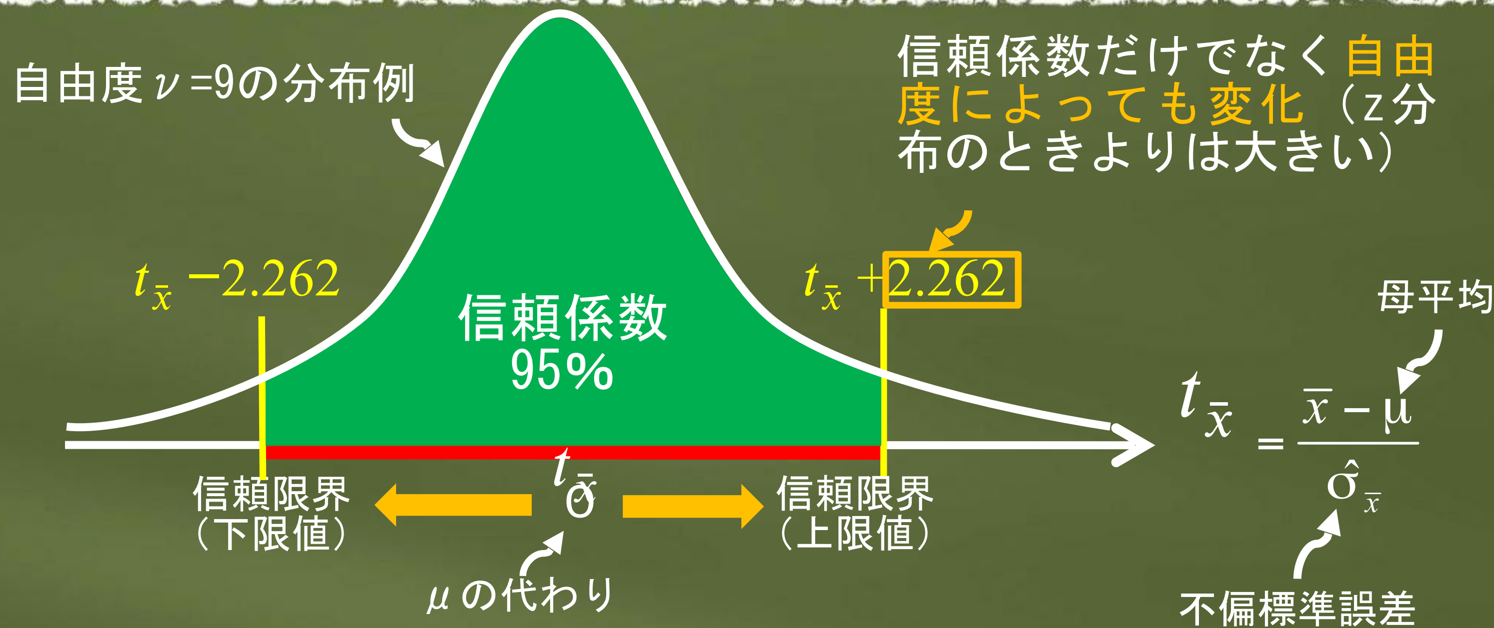
（自由度が母数）

階乗を一般化したガンマ関数

t分布の確率密度を求めるExcel関数

=T.DIST(t値，自由度n-1，確率密度FALSE/下側累積分布TRUE)

t分布を用いた母平均の区間推定



母平均の信頼区間
（母分散が未知でもOK）

$$t_{\bar{x}} - 2.262 < 0 < t_{\bar{x}} + 2.262$$

母平均はゼロに変換されている

（t分布による母平均の）
信頼区間の式の変形

$$t_{\bar{x}} - 2.262 < 0 < t_{\bar{x}} + 2.262$$

tの連立不等式にする

$$-2.262 < t_{\bar{x}} < 2.262$$

t式を展開する

$$-2.262 \leqq \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} \leqq 2.262$$

母平均μの連立不等式にすると…

$$\bar{x} - 2.262\frac{s}{\sqrt{n-1}} \leqq \mu \leqq \bar{x} + 2.262\frac{s}{\sqrt{n-1}}$$

標本サイズが10のとき，母平均に対する信頼係数95%の信頼区間→

$$\left(\bar{x} - 2.262\frac{s}{\sqrt{n-1}},\ \bar{x} + 2.262\frac{s}{\sqrt{n-1}}\right)$$

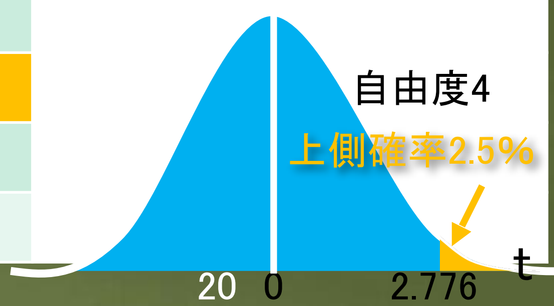
t 分布表の読み方（付録Ⅱ）

上側確率p

自由度
ν=n−1

	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2						
3	1.638	2.353	3.182			
4	1.533	2.132	2.776			
5	1.476	2.015	2.571			
6	1.440	1.943	2.447			

表中の値がt値（z分布表では上側確率p）



自由度ごとに分布表を作成するのは大変なので，よく使う上側確率とtの値だけの対応で1つの表にまとめている

パソコンがあるならば，例えば下記のExcel関数で下側確率に対応するt値が得られる

Excel関数=T.INV(下側確率，自由度)

先ほどの例：=T.INV(0.025，9)→“-2.262”

例　題

ある酪農家が搾乳中のホルスタインのなかから5頭を抽出して乳量を調べたら，平均22.1リットル/日だった。この農家が搾乳している全てのホルスタインの平均乳量を信頼係数95％で推定しなさい。ただし，標本標準偏差は6.5リットル/日であるとする。

解：母分散が未知で小標本なので，t分布で推定する。標本標準偏差は既知なので，後は信頼係数95％の信頼限界のための t 値を分布表かExcel関数などで得ればよい。なお，自由度vはn-1なので，5-1=4となる。すると，t=2.776なので，下記式から，ホルスタインの1日当たりの平均乳量に対する信頼係数95％の信頼区間は（13.1リットル,31.1リットル）となる。

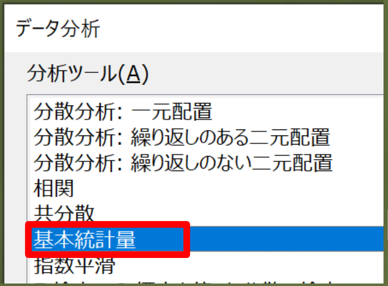
$$\bar{x} - 2.776 \cdot \frac{s}{\sqrt{n-1}} < \mu < \bar{x} + 2.776 \cdot \frac{s}{\sqrt{n-1}} \rightarrow 22.1 - 2.776 \cdot \frac{6.5}{\sqrt{5-1}} < \mu < 22.1 + 2.776 \cdot \frac{6.5}{\sqrt{5-1}}$$

Excel分析ツールによる
母平均の区間推定

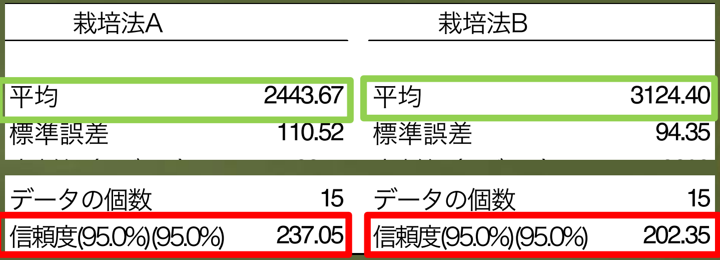
表1.2　キュウリの収量(g)

ポット番号	栽培法A（昼）	栽培法B（夜）
1	3,063	3,157
2	2,275	2,707
3	2,089	3,270
4	2,855	3,181
5	2,836	3,633
6	3,219	3,404
7	2,817	2,219
8	2,136	2,730
9	2,540	3,408
10	2,263	3,203
11	2,140	2,938
12	1,757	3,286
13	2,499	2,920
14	2,093	3,332
15	2,073	3,478

[基本統計量]の中にある



[平均の信頼度の出力]に☑を入れ，
右欄に信頼係数を入力(95%が標準)



A：2443.67±237.05

B：3124.40±202.35

自由度14で信頼係数95%に対応した誤差の大きさ

4.5 母比率の区間推定

比率とは，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※

比率 $p = \dfrac{\text{ある}}{\text{全要素の数}} = \dfrac{x}{n}$

例：視聴率 $= \dfrac{\text{視聴世帯数}}{\text{テレビ所有世帯数}}$

※注：確率と似ているが，確率は事象の起こりやすさ（偶然性の程度）を数値化したもので，比率よりも狭い意味。記号も同じpだが，確率はprobability，比率はproportionの頭文字からきている。

事例として選挙速報の当確予想を考える

選挙速報の「当確」とは？

当選確実の略で，当選者が1名の選挙では，真の得票率（母比率）が0.5を超えると推定された候補者に出される

（ある候補者の）得票率 $p \quad \frac{\text{ある}}{\text{投}} = \frac{x}{n}$

出口調査

推定

標本比率（標本の得票率）$\hat{p}$

母比率（真の得票率）p

標本比率も正規分布に従う

- 投票行動はベルヌーイ試行：「投票する/しない」のいずれかで，誰かの投票行動が他の投票行動に影響を与えず，その候補者に票を入れる確率は誰でも一定

→標本比率は二項分布B(np, np(1-p))に従う

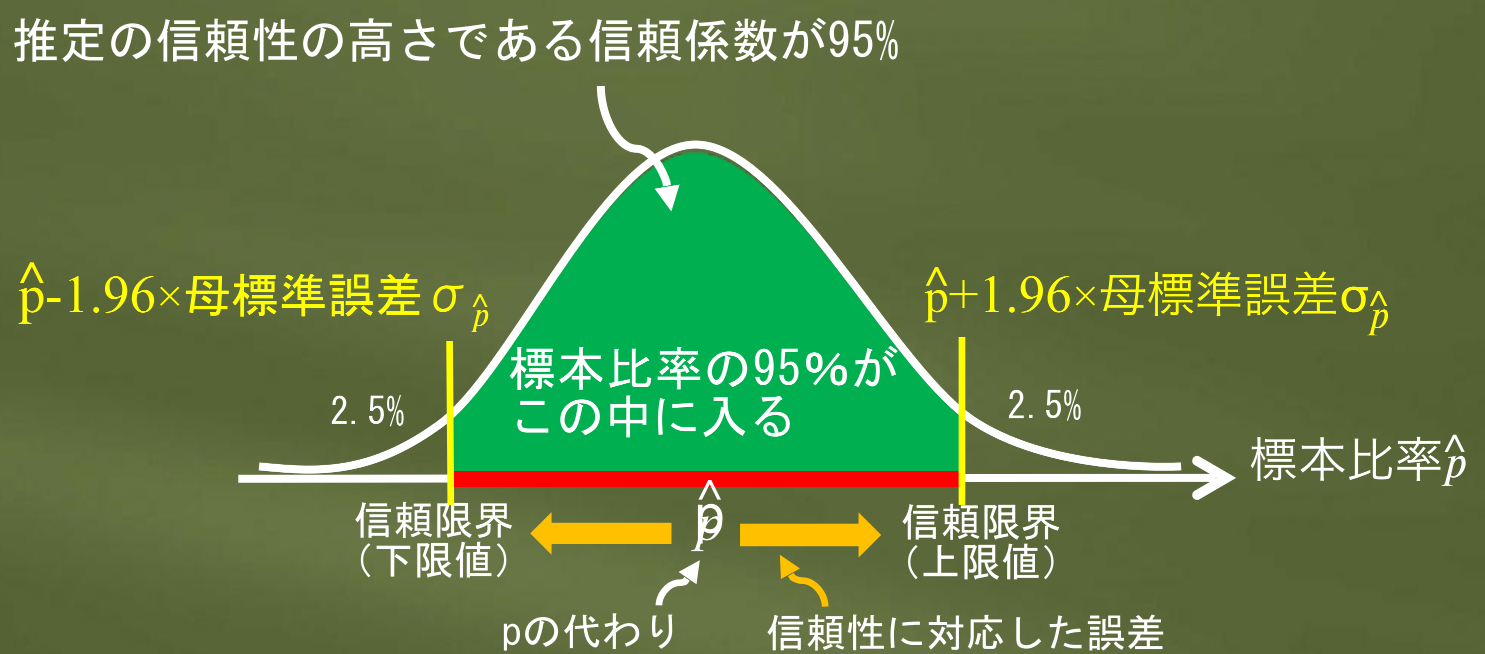
→大標本ならば正規分布と考えて良い(中心極限定理)

ただし，確率変数は比率（試行の成功回数xのままではなく，それを試行回数nで割った“x/n”)であるため，標本比率の母数は，次のようになる

平均：np/nで“p”，　誤差分散：np(1-p）/n^{2}で“p(1-p）/n”

期待値（母比率）

正規分布を用いた
母比率の区間推定（実際は不可能）



母比率の信頼区間
（母分散が既知）

$$\hat{p}-1.96\times\sqrt{\frac{p(1-p)}{n}} < p < p+1.96\times\sqrt{\frac{p(1-p)}{n}}$$

母比率pがわかっている必要

実際の母比率の区間推定

- 大標本（n≧100）ならば，母比率pの代わりに標本比率$\hat{p}$を使う

Wald 法による
母比率の信頼区間（大標本）

$$\hat{p}-1.96\times\sqrt{\frac{\hat{p}(1-p)}{n}}<p<p+1.96\times\sqrt{\frac{p(1-p)}{n}}$$

- 大標本でないときは，標本比率$\hat{p}$を修正した$\hat{p}'$を使う

Agresti-Coull 法による
母比率の信頼区間（小標本）

$$\hat{p}'-1.96\times\sqrt{\frac{\hat{p}'(1-p')}{n}}<p<\hat{p}'+1.96\times\sqrt{\frac{p'(1-p')}{n}}$$

$\frac{x+2}{n+4}$ で標本比率を修正（小標本のときに誤差を大きく評価）

例　題

ある市長選でA氏とB氏が立候補した。50人に対する出口調査の結果，A氏の得票率は70％だったが，A氏に当確を出しても良いだろうか？信頼係数95％で判断せよ。

解：比較的大きな標本なのでWald法で母比率の信頼区間を推定すると，以下のように（57.3%，82.7%）となる。

$$0.7-1.96\times\sqrt{\frac{0.7(1-0.7)}{50}}<p<0.7+1.96\times\sqrt{\frac{0.7(1-0.7)}{50}} \rightarrow 0.573<p \quad 0.827$$

1名のみが当選するので，このうち下限値が0.5を超えていれば当確を出して良いことになる。

A氏の信頼限界の下限値は0.573なので，当確を出せる。

区間推定式の逆算による簡易的な
標本サイズの決め方　比率編

- アンケートなど，母比率の推定が基本となる調査では，どの位の大きさの標本が必要（何人に聞けば良い）？

→母比率の信頼区間の幅を決める式から逆算

→信頼性に対応した誤差の大きさの部分を使う

$$\boxed{\sqrt{\frac{p(1-p)}{n}}}$$

→許容できる誤差の大きさを右辺においてnを求める

$$1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

確保したい信頼性の高さによって変化させる

例　題

A君は卒論でアンケート調査をすることになった。お金も時間も無いので，95％という信頼性なら，母比率の推定誤差は10％に収まれば良いと考えている。さて，A君は何人から回答を得られれば良いだろうか？

解：pは不明なので，誤差が最大になる0.5を入れてnを求めると96となる。よって，A君は96人からの回答を集めれば10％以内の誤差で，母比率に対する信頼係数95％の信頼区間を推定することができる。

$$1.96\sqrt{\frac{p(1-p)}{n}} = 0.$$

⬅ 未知のpには，類似の調査があればその標本比率を，それもなければ保守的な結果となるように，誤差が最大となる0.5を入れておく

標本サイズの決め方　平均編

- 母平均の信頼区間の推定式（の誤差部分）を使っても，事前に標本サイズnを決めることが可能
- 母比率の式よりも単純だが，母分散の予想値が必須

未知の母分散（母標準誤差）には，類似の調査などを利用

$$1.96\frac{\sigma}{\sqrt{n}}=\text{許容できる誤差（実際の値）}$$

例題：新設計の車両の制動距離の信頼区間を知りたい。信頼係数95％で誤差が±1mの信頼区間を計算するには何台（n）からデータを取れば良いか？ただし，類似車両の制動距離の標準偏差（σ）は2mであることがわかっている。

解：1.96×2÷√n=1を逆算するとn=15.3なので，繰り上げて16台となる。

以上で第４章は終了です。