母集団の統計的推定

心理データ解析演習　2011/07/06

M1　水垣　さなえ

発表の流れ

- はじめに
- 推測統計の基礎事項
- 母集団平均の区間推定 (t分布等)
- 母集団分散の区間推定 (X^2分布)
- 母集団比率の区間推定 (2項分布)
- 区間推定の注意

はじめに　統計的推定とは

例題 (管,2009より)

：Aさんは久しぶりに家に帰り、稲刈りをしました。収穫された稲穂を1本拾い粒数をかぞえると90粒でした。別の1本についても調べると95粒でした。さらに別の稲穂を抜き出し全部で20本の粒数を数えました。稲穂1本あたりの平均粒数は93粒であることがわかりました。このことからAさんは次のことを考えました。

1. 実家の田んぼ全体の稲穂一本あたりの平均粒数を、たった一度の調査である2　0本の平均値から93粒と判断してよいだろうか。
2. 1つの値で言い切るのが難しければ、90粒から95粒の間にあるといった大雑把なことはいえないだろうか

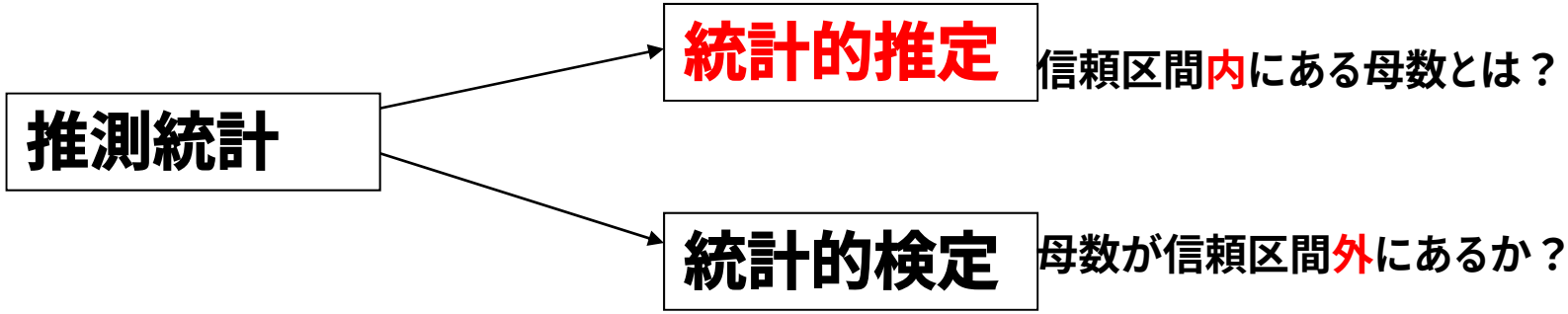
この2つの問いに答える方法が

統計的推定

基礎事項

推測統計

調べたい集団の**一部**のデータから集団**全体**の特色や傾向を明らかにする方法。



統計的推定と統計的検定は”表裏一体”の関係

推計したい集団のこと　＝**母集団**

集団の一部分を対象とする調査　＝**標本調査**

標本調査における対象　＝**標本**or**サンプル**

標本調査における対象の個数　＝**サンプルサイズ**

推測統計の用語

推測統計では標本調査の基本統計量を用いて、母集団を推測する

母集団		標本調査	
サイズ	N	標本サイズ	n
母平均値	μ (ミュー)	標本平均	$\bar{X}$
母標準偏差	σ (シグマ)	標本標準偏差	s
母分散	σ^2	標本分散	s^2
母比率	P	標本比率	$\bar{p}$

推測統計の主な考え方 (確率)

確率事象

ある確率変数 (X) が出るというような事柄に目をつけたとき、その事柄が現れるのは、1回の施行では不確定であるが、その事柄の起こる割合が多数回の実験を試みるとき安定性をもつ、そういう事柄だけについて考えるとき、これを**確率事象**とよぶ。

確率事象は、標本調査の基本統計量間の関係から主に次のようなことが分かっている。

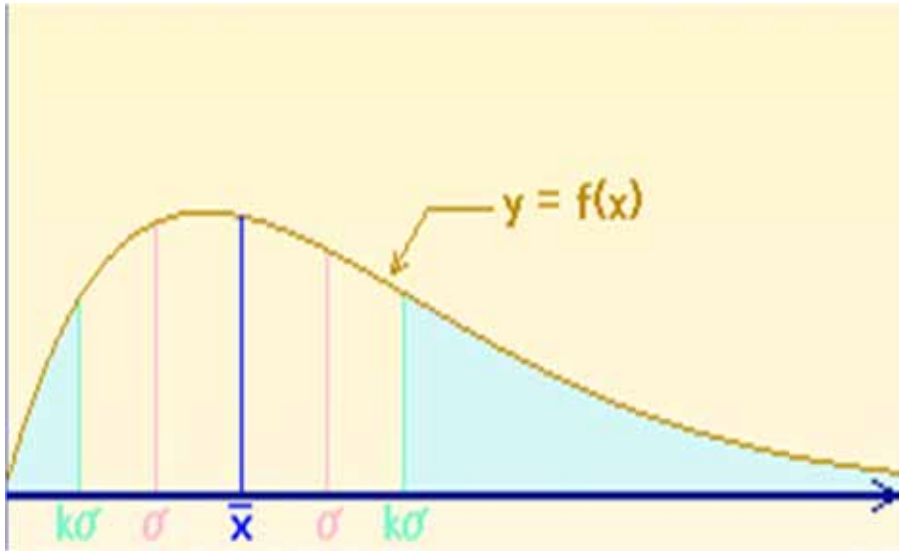
1. **チェビシェフの不等式**
2. **正規分布**
3. **大数の法則**
4. **中心極限定理**

統計的推定の主な考え方（確率）①

チェビシェフの不等式

確率変数Xの平均　E[X]＝μ、分散V[X]＝σ^2が共に有限ならば任意のk（＞0）に対して

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$



http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/cebysev/cebysev.htmより

つまり母平均μ（＝$\overline{X}$）以外の確率変数Xが出る割合（母平均μからのズレの割合）は、母標準偏差σの値で分かるというもの。

式の意味は、母平均μと他の確率変数の差が母標準偏差σのk倍以上になる確率は$1/k^2$以下であるということ。（上の図の青い面積部分の確率）

統計的推定の主な考え方（確率）②

正規分布

理論的に考案された分布

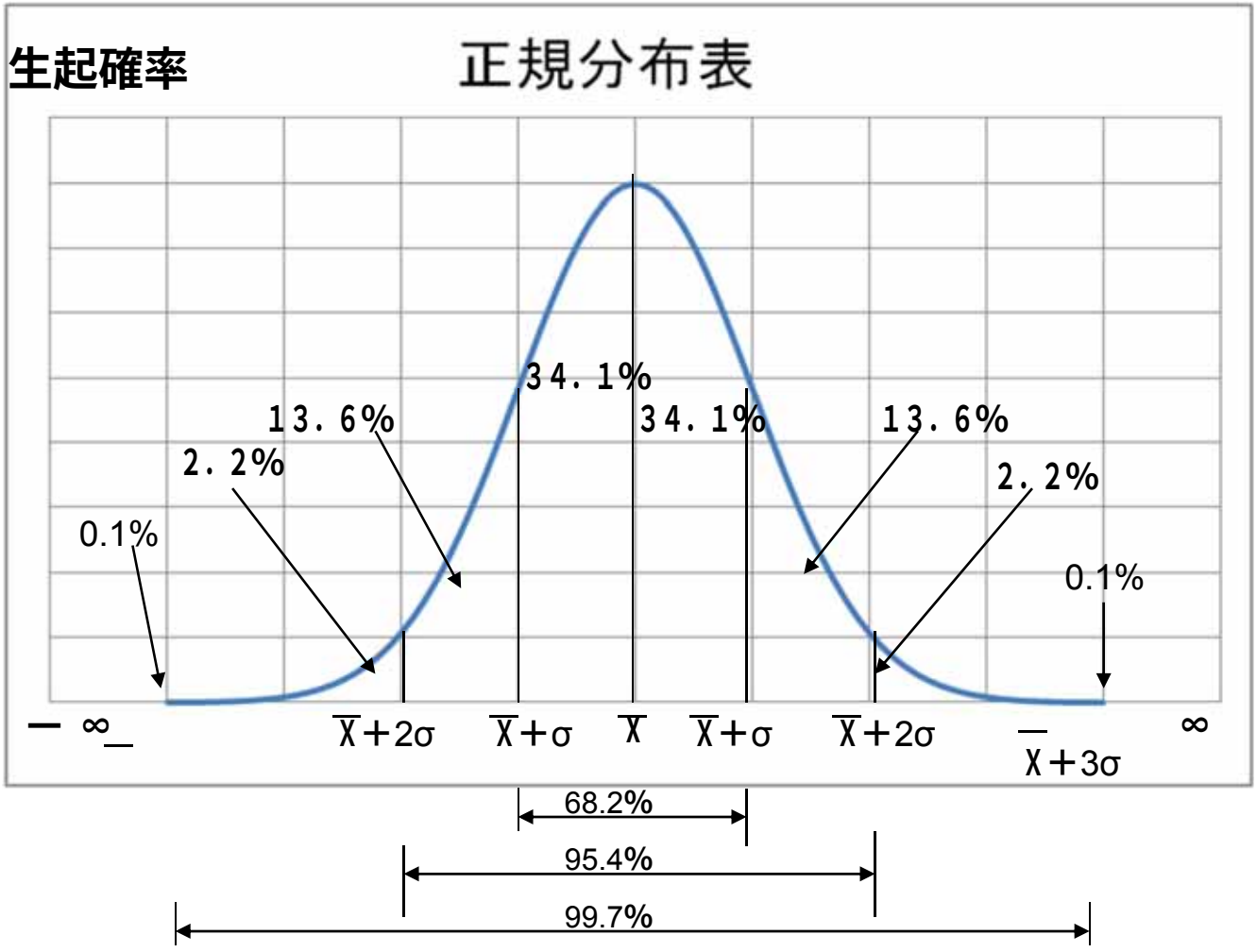
理論的な式によって正規分布に従う変数Xが任意のaからbまでの値をとる確率P（a≦x≦b）が分かる

①左右対称(歪度＝0)

②中央に山が一つ

③両すそがなだらかに広がった（尖度＝3）形

⇒ベルカーブ（釣鐘状）



統計的推定の主な考え方（確率）③

大数の法則

どんなサンプルと母集団にも当てはまる法則

確率変数$X_1 X_2 \cdots X_n$が独立であって、その期待値$E(X_1) = E(X_2) = \cdots E(X_n) = \mu$かつ、一定の定数$\sigma^2$に対して$X_i$の分散$V(X_i) \leqq \sigma^2$ $(i=1,2\cdots,n)$であるならば、任意に与えられた正数$\varepsilon > 0$に対して、

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| < \epsilon\right) = 1$$

となる。

これは、nを十分に大きくとれば確率変数$\overline{X}$は定数μに限りなく近づくという意味

サイコロを一回振って1の目が出る確率＝

=理論的確率

出る目は1か1以外でしかなく、1が⅙回でるということはありえない

でも多数回ふると1が出る割合は頻度の比⅙として確認できる。

http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/taisuu/taisuu.htm

統計的推定の主な考え方（確率）④

中心極限定理

Xが平均μ、標準偏差σのある分布に従うとき、大きさnの無作為標本にもとづく標本平均$\overline{X}$は、nが無限に大きくなるとき、平均μ、標準偏差σ／$\sqrt{n}$の正規分布に近づく

中心極限定理は**母集団の分布に関係なく**、

⇒ その標本平均は、正規分布に従う

http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/tyuusin2/chuusin.htm

推定法

稲穂の例題（管,2009）にもどると

Aさんは久しぶりに家に帰り、稲刈りをしました。収穫された稲穂を1本拾い粒数をかぞえると90粒でした。別の1本についても調べると95粒でした。さらに別の稲穂を抜き出し全部で20本の粒数を数えました。稲穂1本あたりの平均粒数は93粒であることがわかりました。このことからAさんは次のことを考えました。

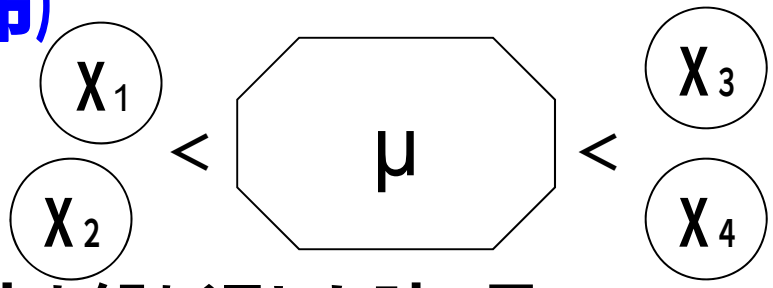
1. 実家の田んぼ全体の稲穂一本あたりの平均粒数を、たった一度の調査である2　0本の平均値から93粒と判断してよいだろうか。
2. 1つの値で言い切るのが難しければ、90粒から95粒の間にあるといった大雑把なことはいえないだろうか

標本分布と母数の関係

標本分布 (標本平均の分布)

- 無作為の標本抽出を多数回それぞれ独立に繰り返したとき、標本抽出に伴う誤差 (偶然) によって標本の平均値Xは、母平均μと完全に一致することはほとんどなく、母平均μを中心にある程度のバラツキ (分散) をもって分布する。

＝標本分布(標本平均の分布)



稲穂の粒数の例でいえば…

もし20本の標本調査を何度も繰り返した時、最初の稲穂の粒数の平均は93粒だったが、別の調査では、95粒だったり、90粒だったりするかもしれない

標本分布と母数の関係

標本分布の分散

母集団が正規分布に従う時、**標本分布**も正規分布に従うことが知られており、その分散は

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

⇒標本サイズnによって標本分布の分散の大きさが変わる

＝nが大きくなれば、よりバラツキの少ない、つまり母平均μの近くに標本の平均値（＝標本分布における確率変数）があつまると考えられる

標準誤差

標本から母集団について推定を行う時、標本分布の分散の大きさに応じて誤差が伴う。このため、標本分布の標準偏差を**標本誤差** (Standard Error：SE)と呼ぶ

母数の推定

標本分布と母集団の関係性から母平均μと母分散σ^2の値を推定していく

1）不偏推定値による母数の推定

2）不偏推定値をもとにした母数の区間推定

不偏推定値による母数の推定①

不偏推定値

ある統計量の期待値がそれに対応する母数（母集団の統計量）と一致する場合、その統計量を母数の**不偏推定値**という。

1）母平均の推定

大数の法則や**中心極限定理**などの考え方から、

標本平均$\bar{x}$は母平均μの**不偏推定値**ということができる。

稲穂の例題　問1への答え

「実家の田んぼ全体の稲穂一本あたりの平均粒数を、たった一度の調査である2　0本の平均値から93粒と判断してよいだろうか。

不偏推定値による母数の推定②

2) 母分散の推定

母分散σ^2も標本分散s^2をそのまま不偏推定値にできるか？

母分散σ^2は、

標本の分散s^2

（$\overline{x}$ に対する確率変数 x のバラツキ具合）

＋

標本分布の分散 $\sigma^2_{\overline{x}}$

（μに対する$\overline{x}$ のバラツキ具合）

だから、母分散$\sigma^2 \neq s^2$で

母分散の不偏推定値は $\hat{\sigma}^2$

$$= \frac{\sum_i^n (X_i - \overline{X})^2}{n} + \frac{\sigma^2}{n}$$

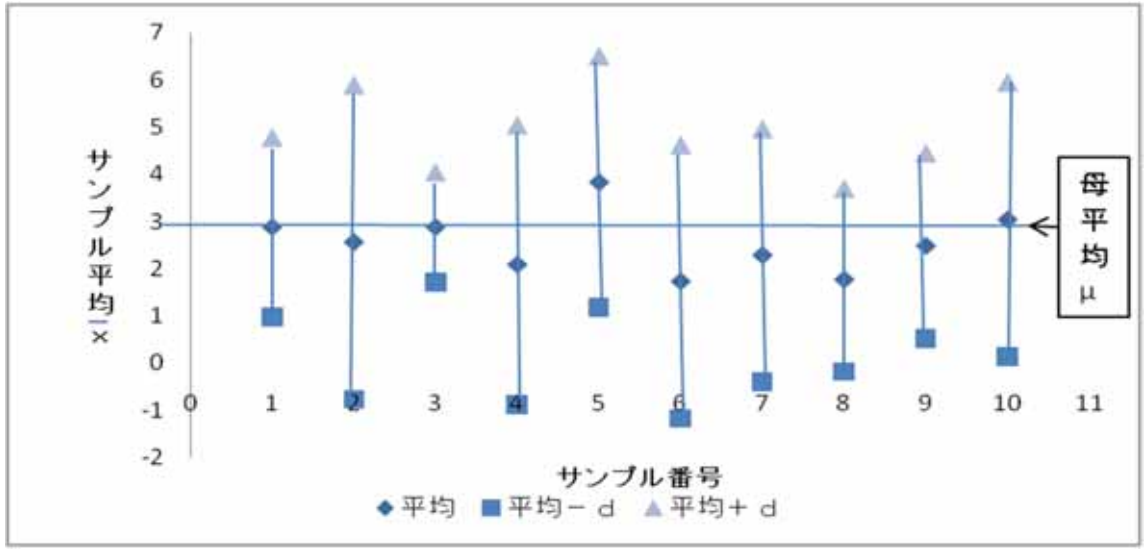
$$= \frac{\sum_i^n (X_i - \overline{X})^2}{n-1}$$

$$= \hat{\sigma}^2$$

不偏推定値をもとにした母数の区間推定

区間推定法

標準誤差（標本抽出による誤差）を考慮して、標本平均 $\bar{X}$ を含むひとつの区間を設定し、その区間の中に母平均μがはいるようにする方法



林、1991を参考に作成

稲穂の例題　問2への答え

1つの値で言い切るのが難しければ、90粒から95粒の間　にあるといった大雑把なことはいえないだろうか

不偏推定値をもとにした母数の区間推定

信頼区間

“母平均μが区間～から～までの間に存在する確率は９５％である”と推定の信頼性が確率的に表現された区間

信頼限界

推定された母数の上限および下限

たとえば母平均μの推定の信頼区間の上限（$\bar{X}u$）および下限（$\bar{X}_L$）は、

上限　$\bar{X}u = \bar{X} + 誤差$　　下限　$\bar{X}_L = \bar{X} - 誤差$

誤差の値の作り方は後述

信頼係数（信頼度）

推定の信頼性を表す確率。

例えば母平均推定の場合、母平均μが信頼区間に含まれる確率を意味する。　この場合信頼係数は１－α

誤差と信頼度は裏腹の関係にある。

誤差が大（信頼区間広い）⇒　信頼度が低い

誤差が小（信頼区間狭い）⇒　信頼度が高い

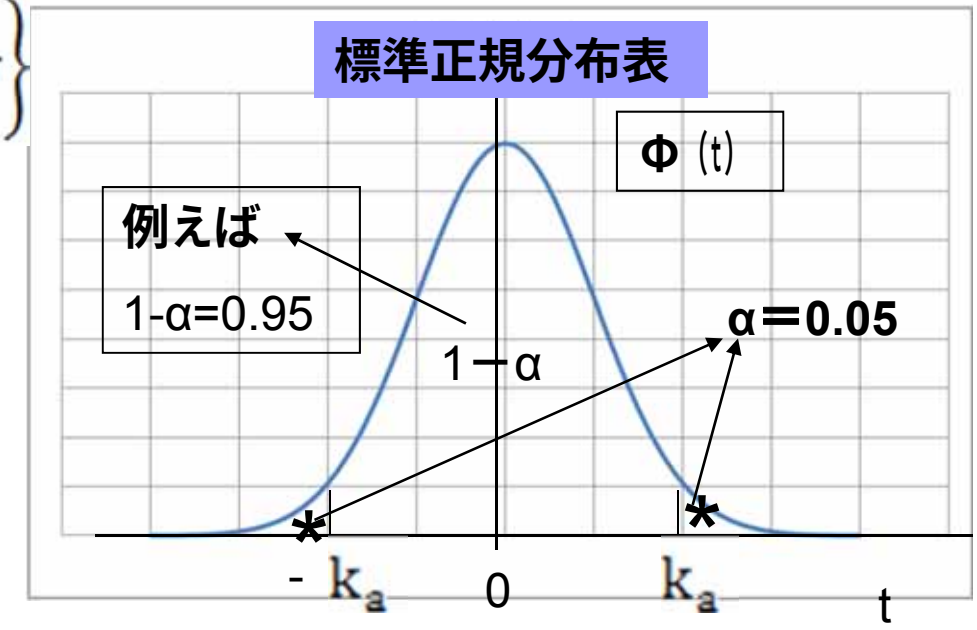
母数の区間推定①　　平均の推定

■ 母集団平均の推定 (1)

母集団が正規分布N $(\bar{x},\sigma^2)$ でσ^2が既知の時

標準正規分布を用いて推定 (実際の調査ではほとんどないが考え方として)

$$P_r\left\{|\bar{x}-\mu|<k_\alpha\frac{\sigma}{\sqrt{n}}\right\}=P_r\left\{\frac{|\bar{x}-\mu|}{\sigma/\sqrt{n}}<k_\alpha\sigma\right\}$$



＝ (Φ (t) のーkaからkaまでの面積)

＝1－α

たとえば信頼度を95%とするなら、

1－α＝0.95でα＝0.05となるz値を

標準正規分布表から探して計算

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

母数の区間推定①　平均の推定

■ 母集団平均の推定 (2)

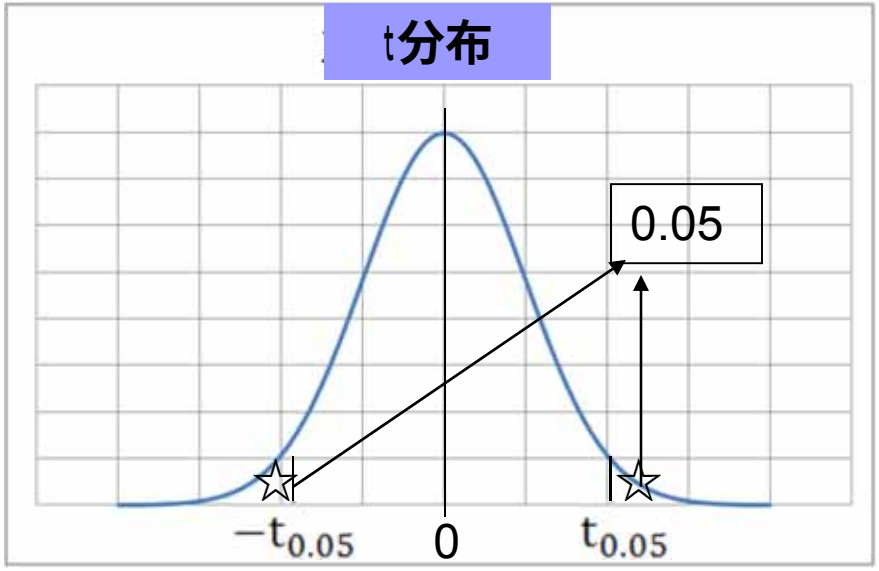
母集団が正規分布N ($\bar{X}$,σ^2) でσ^2が未知の時

t分布を用いて推定 (自由度はn-1)信頼係数95%の時

$$T = \frac{\sqrt{n-1}(\bar{x}-\mu)}{s}$$

$$P_r\{-t_{0.05} < T < t_{0.05}\} = 0.95$$

とすると



となるtの値は$t_{0.05}$ =2.776

$$= \bar{X} - t_\alpha \frac{s}{\sqrt{n-1}} < \mu < \bar{X} + t_\alpha \frac{s}{\sqrt{n-1}}$$

$$= \bar{X} - 2.776\frac{s}{\sqrt{n-1}} < \mu < \bar{X} + 2.776\frac{s}{\sqrt{n-1}}$$

母数の区間推定①　平均の推定

母集団平均の推定 (3)

- **母集団の分布が不明の時**

通常の調査では標本平均 $\bar{X}$ について、nが大きい時 (実用上は大体30以上)

$$E(\bar{X})=\mu \quad , \quad \sigma_{\bar{x}}^2=\frac{\sigma^2}{n}$$

なる正規分布に近似的に従うため、**標準正規分布**で推定

またnが十分大きい時 (大体100以上) は

標本分散s^2を母平均σ^2の代用としてもよい (**中心極限定理**)。

信頼係数は100%としてもよいが、現実的に95%として

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

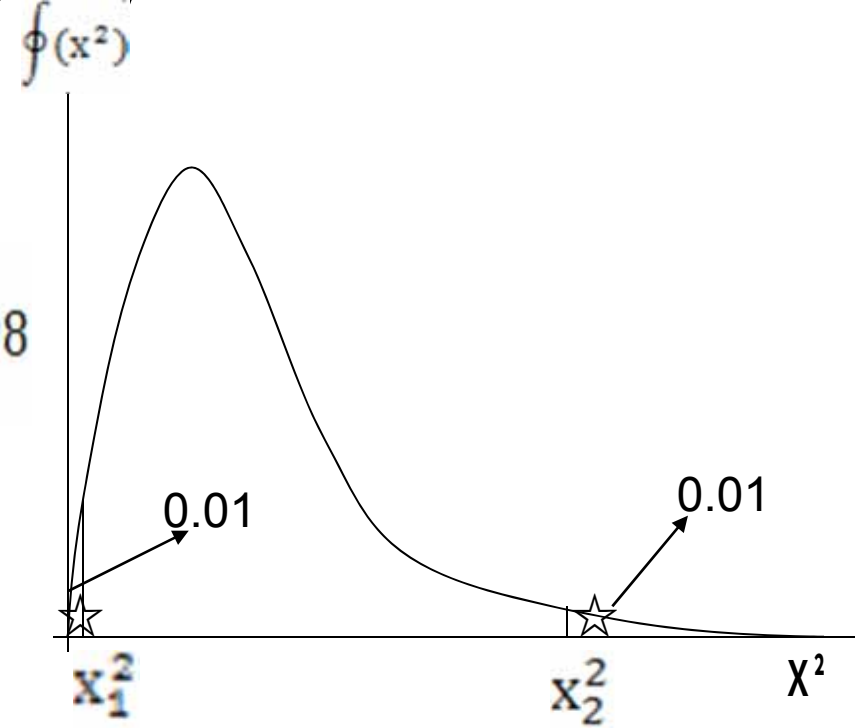
母数の区間推定　分散の推定

母集団分散の推定

- 母集団が正規分布N $(\overline{X},\sigma^2)$ でσ^2が未知の時

  X^2分布を用いて推定（自由度n-1）

$$x^2 = \frac{ns^2}{\sigma^2}$$ とすると



たとえば信頼係数98%の時

$$P_r\left\{x_1^2 < x^2 < x_2^2\right\} = P_r\left\{\frac{ns^2}{x_2^2} < \sigma^2 < \frac{ns^2}{x_1^2}\right\} = 0.98$$

を満足する x_1^2 と x_2^2 をX^2分布より求めて

$$\frac{ns^2}{13.277} < \sigma^2 < \frac{ns^2}{0.297}$$

母数の区間推定　比率の推定

母集団比率の推定

- **母集団が二項分布の時**

  E(p)**標本比率の期待値**=E($\overline{X}$)**標本平均の期待値**=P**母比率**

  このとき母集団の分散＝P(1-P)

  標本の大きさが十分大きい時 (大体100**以上) は標本比率は**p**は近似的に平均**P**,分散** $\frac{P(1-P)}{n}$ **の正規分布に従うから**

$$P_r\left\{|p-P|<k_\alpha\sqrt{\frac{P(1-P)}{n}}\right\}=1-\alpha$$ **より**

  信頼係数1-α**の信頼区間は**

$$p-k_\alpha\sqrt{\frac{P(1-P)}{n}}<P<p+k_\alpha\sqrt{\frac{P(1-P)}{n}}$$

区間推定の注意点

区間推定の注意点

確率モデルを適用するためには

1. **標本調査が無作為 (ランダムサンプリング)**
2. **どのような母集団を推定するか明確に**
3. **母集団分布が正規分布に従うことを仮定**
4. **推定によって得られるのは数学的な結果**

参考文献

- 管　民郎(2009)『らくらく図解　統計分析教室』, ㈱オーム社
- 丹羽時彦　関西学院高等部ホームページより
  「放課後の数学入門」　統計 (数学C)
  http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/toukeihy.htm
- 林　知己夫編　林　文、佐藤良一郎、青山博次郎、林知己夫著 (1991)
  『統計学の基本』　朝倉書店
- 吉田寿夫　(1998)
  『本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本』　㈱北大路書房
- 森　敏昭、吉田寿夫編著(1990)　『心理学のためのデータ解析テクニカルブック』
  ㈱北大路書房