統計の分析と利用
（旧カリ：データ分布と予測）
3．母集団と標本

堀田　敬介

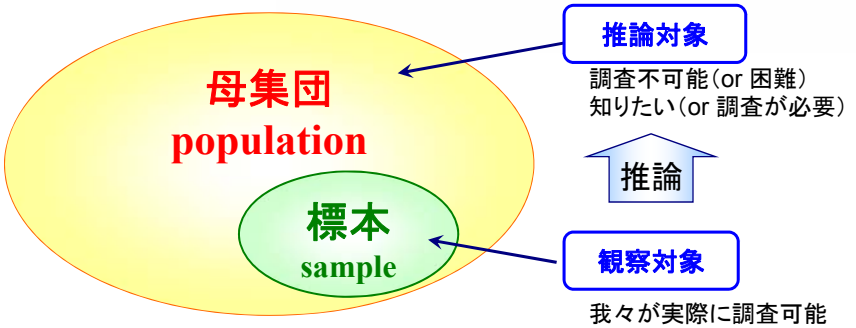
2010/11/26,Fri.～

Contents

- 母集団と標本
  - 母平均，母分散の推測
    - 標本平均
      - 標本平均の従う確率分布
      - 大数の法則，中心極限定理
      - 標準正規分布，t分布
    - 標本分散
      - 標本分散の従う確率分布
      - χ^2分布
  - 母比率の推測
    - 標本比率

母集団と標本：統計的推論

- 推測統計学 statistical estimate / statistical inference

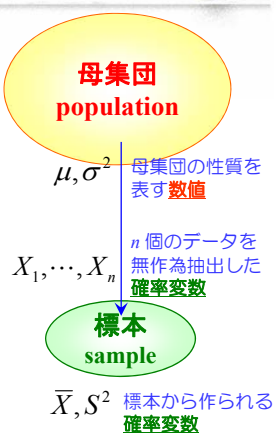


- 母集団が大きすぎて調査不可能な場合
  - 全国大学生の身長
- 全数調査(悉皆調査)がそもそも不可能な場合
  - 品質検査
  - 料理の味見

注意：今後特に断りのない限り，無限母集団を考える．

母集団と標本：統計的推論

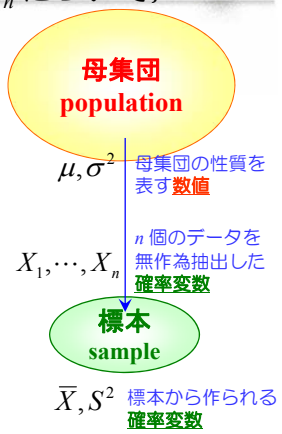
- 母集団の性質を表す数値
  - 母平均：μ
  - 母分散：σ^2（母標準偏差：σ）
- 母集団からの標本
  - データ n 個を無作為抽出 $X_1,\cdots,X_n$
    - 無作為抽出には乱数などを利用
    - $X_1,\ldots,X_n$ は互いに独立な確率変数
      - 標本調査は試行：無作為抽出により，実際に取る値は偶然による
    - 各確率変数 X_i は母集団と同じ分布に従う
    - n はサンプルサイズ（抽出した標本数）
- 確率変数 $X_1,\ldots,X_n$ から作られる確率変数
  - 標本平均：$\overline{X}=\dfrac{X_1+\cdots+X_n}{n}$
  - 標本分散：$S^2=\dfrac{1}{n}\left\{(X_1-\overline{X})^2+\cdots+(X_n-\overline{X})^2\right\}$



標本分布：標本平均

母集団から抽出した標本数 n の標本 $X_1,\ldots X_n$ について，以下の確率変数を標本平均 $\overline{X}$ という

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$



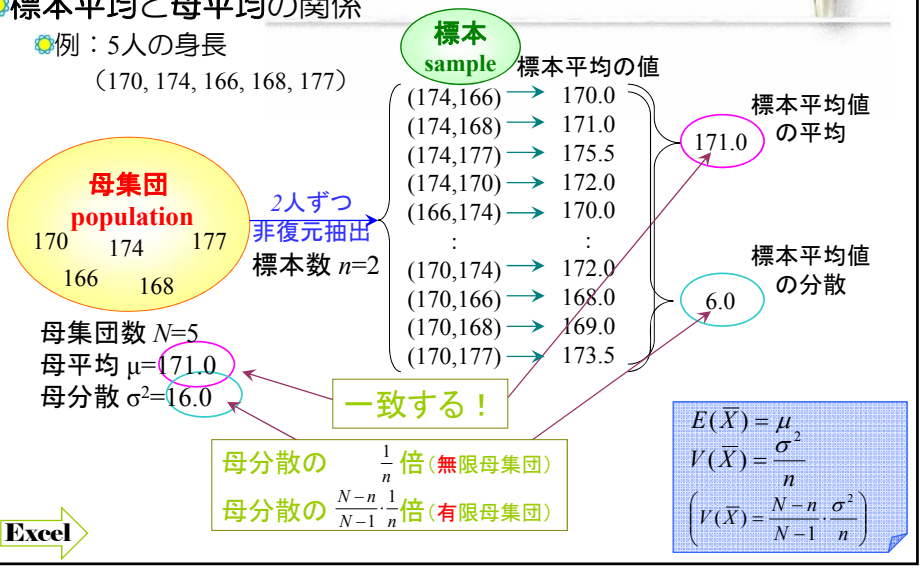
注意）「標本平均」は確率変数「標本平均値」が標本毎に実際に取る値

母集団と標本：標本平均

標本平均と母平均の関係

例：5人の身長

（170, 174, 166, 168, 177）



$$E(\overline{X}) = \mu$$
$$V(\overline{X}) = \frac{\sigma^2}{n}$$
$$\left(V(\overline{X}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\right)$$

補足：標本平均の平均と母平均・標本平均の分散と母分散の関係（証明）

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n\mu = \mu$$

$$\begin{aligned}
V(\bar{X}) &= E(\bar{X} - E(\bar{X}))^2 \\
&= E\left(\frac{X_1 + \cdots + X_n}{n} - E(\bar{X})\right)^2 \\
&= \frac{1}{n^2} E(\{X_1 - E(X_1)\} + \cdots + \{X_n - E(X_n)\})^2 \\
&= \frac{1}{n^2} E\left(\{X_1 - E(X_1)\}^2 + \cdots + \{X_n - E(X_n)\}^2 + 2\sum_{i<j}(X_i - E(X_i))(X_j - E(X_j))\right) \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{n} E(X_i - E(X_i))^2 + 2\sum_{i<j}(X_i - E(X_i))(X_j - E(X_j))\right\} \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{n} V(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\right\} \\
&= \frac{1}{n^2}\left\{n\sigma^2 - 2 \cdot \frac{n(n-1)}{2} \cdot \left(-\frac{1}{N-1}\sigma^2\right)\right\} \\
&= \frac{1}{n} \cdot \frac{N-n}{N-1}\sigma^2
\end{aligned}$$

$$\begin{aligned}
Cov(X_i, X_j) &= E\left((X_i - E(X_i))(X_j - E(X_j))\right) \\
&= E\left((X_i - \mu)(X_j - \mu)\right) \\
&= \frac{1}{N(N-1)}(x_1 - \mu)(x_2 - \mu) + \cdots + \frac{1}{N(N-1)}(x_N - \mu)(x_{N-1} - \mu) \\
&= \frac{1}{N(N-1)}\left(\{(x_1 - \mu) + \cdots + (x_N - \mu)\}^2 - \{(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2\}\right) \\
&= \frac{1}{N(N-1)}\left(\left\{\frac{x_1 + \cdots + x_N}{N} - \mu\right\}^2 - \{(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2\}\right) \\
&= \frac{1}{N(N-1)}\left(0^2 - N\sigma^2\right) = -\frac{1}{N-1}\sigma^2
\end{aligned}$$

補足：有限母集団修正

- 母集団が有限の場合
  - 標本平均の分散と母分散の関係は，

$$V(\bar{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

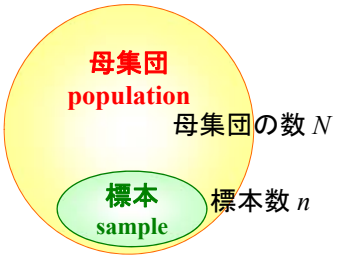
有限修正項

Nが余り大きくない場合や，n/Nが大きい場合

標本数nに比べて母集団の数Nが大きくないとき，有限修正項を考慮する．
無限母集団（Nが十分大きい）時は，有限修正項は1となるので無視して良い．

- 母集団が無限の場合
  - 標本平均の分散と母分散の関係は，

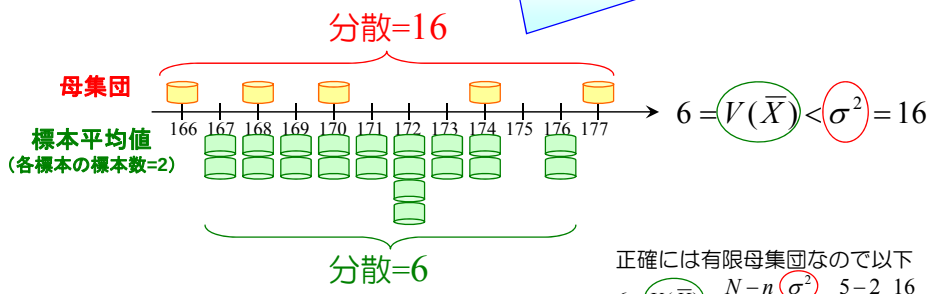
$$V(\bar{X}) = \frac{\sigma^2}{n}$$



補足：母集団と標本：標本平均

- なぜ「標本平均の分散」が「母分散」より小さくなるのか？
  〔即ち，なぜ $V(\bar{X}) < \sigma^2$ なのか？〕
  - 例：5人の身長
    （174, 166, 168, 177, 170）

「**標本平均値**の**散らばり具合**」の方が，
「**母集団**の**散らばり具合**」より小さい！



正確には有限母集団なので以下

$$6 = V(\bar{X}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n} = \frac{5-2}{5-1}\cdot\frac{16}{2}$$

母集団と標本：標本平均（まとめ）

- 標本平均

母集団からn個
無作為抽出

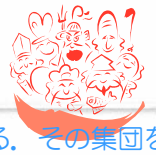
$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

- $X_1,\ldots,X_n$はそれぞれ確率変数
- それから作られる標本平均も確率変数

- 注意：「標本平均」と「標本平均値」は意味が違う
  - 標本平均 … 上で定義される確率変数
  - 標本平均値 … 確率変数「標本平均」が標本ごとに実際に取る値
- 「標本平均 $\bar{X}$ の期待値は母平均μに等しい」 $E(\bar{X}) = \mu$
- 「標本平均 $\bar{X}$ の分散は母分散σ^2の$1/n$に等しい」 $V(\bar{X}) = \frac{\sigma^2}{n}$

有限母集団の場合：$V(\bar{X}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$

演習１：標本平均



1．世界に4匹しかいない貴重な昆虫がいる．その集団を母集団としよう．
- 神様はこの4匹の全長を全て知っており，それぞれ (2, 6, 7, 5) である．
- 神様は母平均の値を求めた．いくつか？ $\mu = ?$
- 神様は母分散の値を求めた．いくつか？ $\sigma^2 = ?$

2．探検家は2匹捕まえる．それが標本となる．
- 各探検家は重複なく2匹を捕まえた．
（つまり，非復元抽出で2匹捕らえ，全長測定後放す）
- 各探検家は自分が捕まえた2匹の標本の平均値を求めた．
- それぞれ，いくつか？ 全ての組合せについて計算せよ．$\overline{X} = ?$

3．１と２の結果から，$E(\overline{X}) = \mu$ と $V(\overline{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$ が成立していることを確認しよう．
ただし，Nは母集団の大きさ，nは標本の大きさである．

母集団と標本：大数の法則

- 「標本平均 $\overline{X}$ の期待値は母平均 μ に等しい」 $E(\overline{X}) = \mu$
- 「標本平均 $\overline{X}$ の分散は母分散 σ^2 の1/nに等しい」 $V(\overline{X}) = \frac{\sigma^2}{n}$

有限母集団の場合 $\frac{N-n}{N-1} \cdot \frac{1}{n}$ 倍

大数の法則

標本数 n が大きくなるにつれて，標本平均

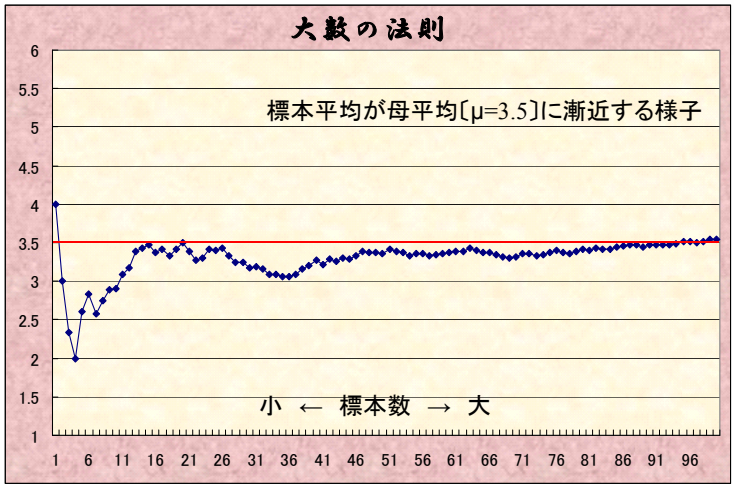
$$\overline{X} = \frac{1}{n}(X_1 + \cdots X_n)$$

が母平均μに近い値をとる確率は 1 に近づく．

標本数 n が十分大きければ，標本は母集団を正しく表すと考えてもよいでしょう．

母集団と標本：大数の法則

- 大数の法則
  - 例：サイコロを振って出た目の平均〔μ=3.5〕



Excel

補足：大数の法則

大数の法則

$$P\left(\left|\overline{X}-\mu\right|<\varepsilon\right)\to 1 \quad (n\to\infty)$$

証明はチェビシェフの不等式 $P\left(\left|\overline{X}-\mu\right|>k\sigma\right)\le 1/k^2$ から

∵) $X_1,\ldots,X_n$ は独立で，同じ分布に従う

$$\to \quad E(X_i)=\mu, V(X_i)=\sigma^2 (i=1,\cdots,n)$$

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \quad \text{とすると} \quad E(\overline{X})=\mu, V(\overline{X})=\frac{\sigma^2}{n}$$

ここで，チェビシェフの不等式から，$k\sigma:=\varepsilon$とおくと （$\sigma^2:=\sigma^2/n$）

$$P\left(\left|\overline{X}-\mu\right|>\varepsilon\right)\le \sigma^2/n\varepsilon^2\to 0 \quad (n\to\infty) \ \blacksquare$$

標本分布

- 標本平均 $\bar{X}$ はどんな確率分布に従うのか？
  - 母集団分布が正規分布 $N(\mu,\sigma^2)$ の場合〔母平均 μ 母分散 σ^2〕
    - ⇨ 標本平均 $\bar{X}$ は正規分布 $N(\mu,\sigma^2/n)$ に従う
  - 母集団分布が正規分布ではない場合〔母平均 μ 母分散 σ^2〕
    - ⇨ 標本平均 $\bar{X}$ は正規分布 $N(\mu,\sigma^2/n)$ に従う

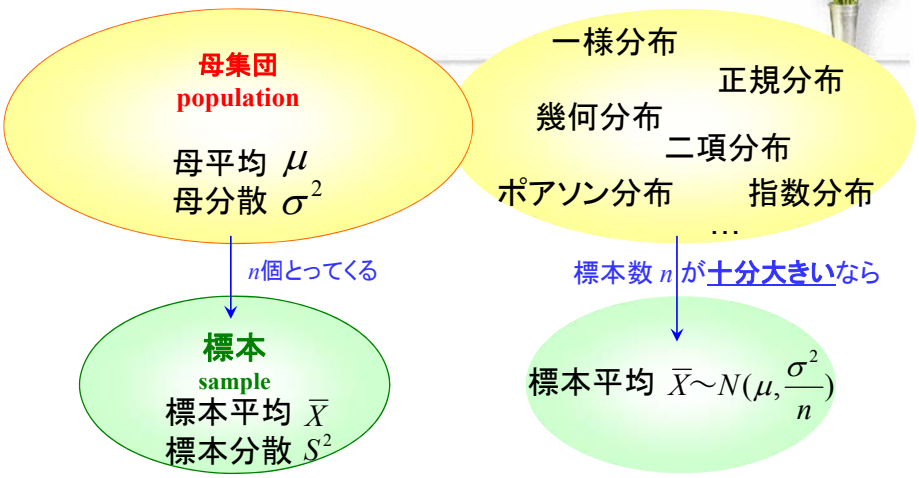
中心極限定理

$X_1,\cdots,X_n$

母平均 μ，母分散 σ^2 の母集団から大きさ n の標本を無作為に抽出した時，n が十分大きければ，母集団の従う確率分布に関係なく，標本平均 $\bar{X}$ は平均 μ，分散 σ^2/n の正規分布 $N(\mu,\sigma^2/n)$ に従うとみなすことができる

$$\begin{cases} X_1+\cdots+X_n \sim N(n\mu,n\sigma^2) \\ \bar{X}=\dfrac{1}{n}(X_1+\cdots+X_n) \sim N(\mu,\dfrac{\sigma^2}{n}) \end{cases}$$

中心極限定理



中心極限定理は，母集団分布がなんであっても（正規分布でなくても），標本数 n が十分大きければ，標本平均 $\bar{X}$ は，近似的に正規分布 に従う，と述べている

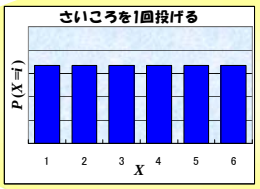
中心極限定理

母集団
population

母平均 μ
母分散 σ^2

n個とってくる

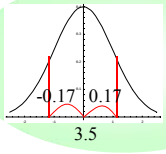
標本
sample
標本平均 $\overline{X}$
標本分散 S^2



$$\begin{cases}\mu=\frac{7}{2},\\ \sigma^2=\frac{35}{12}\end{cases}$$

標本が十分大きいならば

サイコロを100回投げる



$$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$$
$$=N(\frac{7}{2},\frac{\frac{35}{12}}{100})$$

補足：中心極限定理

中心極限定理

$n\to\infty$のとき,

$$P\left(a\le(X_1+\cdots+X_n-n\mu)/\sqrt{n}\sigma\le b\right)\to\int_a^b\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

が成りたつ．言い換えると,

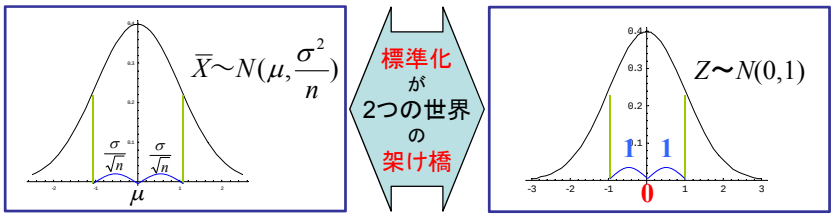
$$P\left(a\le\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\le b\right)\approx\phi(b)-\phi(a)$$

としてよいということ.
(右辺のφは標準正規分布の累積分布関数)

標本分布：標本平均の標準化

- 平均μ, 分散σ^2/nの標本平均 $\bar{X}$（確率変数）の標準化

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



標本から母平均μを推定
「Z推定」・「Z検定」
に利用する

標本平均 $\bar{X}$ が，正規分布 $N(\mu,\sigma^2/n)$ に従うとき，
標準化確率変数 Z は，標準正規分布 $N(0, 1)$ に従う

正規近似

- 二項分布から正規分布へ…
  - 試行回数nを大きくすると，二項分布は正規分布に近づく

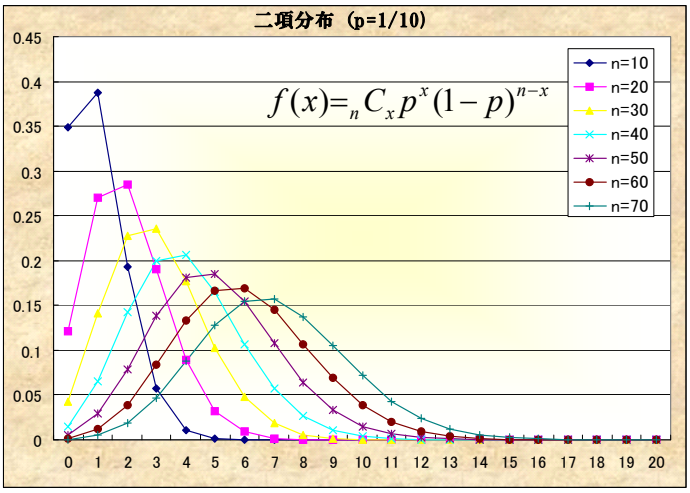
$$Bi(n,p) \xrightarrow{n \to \infty} N(\mu,\sigma^2) \quad \begin{cases} \mu = np \\ \sigma^2 = np(1-p) \end{cases}$$

  - 試行回数nが一定の時に，確率pを0.5に近づけると，二項分布は正規分布に近づく

$$Bi(n,p) \xrightarrow[p \to 0.5]{n = c,} N(\mu,\sigma^2)$$

正規近似

- 二項分布から正規分布へ…
  - 試行回数nを大きくすると，二項分布は正規分布に近づく



正規近似

- 正規分布による二項分布の近似
  - 例：内閣支持率
    - 500人の人に内閣を支持するかどうか聞いたところ，275人が指示すると答えた．

$$\text{内閣支持率：}\ p=\frac{275}{500}=0.55$$

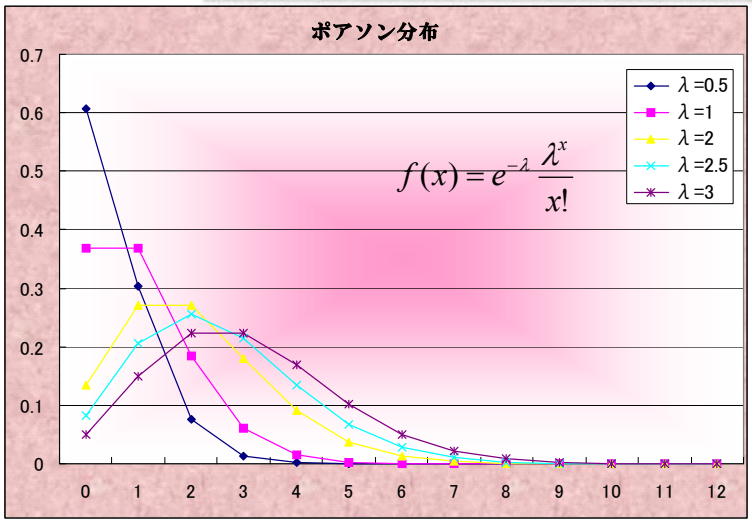
- 内閣支持率を p (不支持率 q = 1-p) とすると，これは二項分布となる．
- 点推定では内閣支持率は55%である．正規分布近似を考えると，$\bar{x}=np=500\times0.55=275$

$$\sigma=\sqrt{npq}=\sqrt{500\times0.55\times0.45}\cong\sqrt{124}\cong11$$

- より，95%信頼区間における区間推定では，内閣支持率は $275\pm1.96\times11 \Leftrightarrow 253\le x\le297$ より50.6%～59.4%

正規近似

- ポアソン分布から正規分布へ…



中心極限定理の利用

平均20,000回で，
400回は±2%の誤差！
ありふれたことだろう．．．

- 例題1： 表裏が等確率で出るコインを40,000回投げる．表が19,600回～20,400回出る確率は？
  - i 回目：X_i=1,0（1：表，0：裏）
  - 表の出る回数：$X=X_1+X_2+\ldots+X_n$

二項分布 $Bi(40000, 1/2)$ に従う

$$f(x)={}_nC_x p^x(1-p)^{n-x} \quad (x=0,1,\cdots,n)$$
$$E(X)=np, V(X)=np(1-p)$$

つまり $P(X>20400)+P(X<19600)$ はいくつか?

$1-\sum_{x=19600}^{20400} {}_{40000}C_x(1/2)^x(1/2)^{40000-x}$ を計算すればよい！

ところが ${}_{40000}C_x$を計算するのは困難！

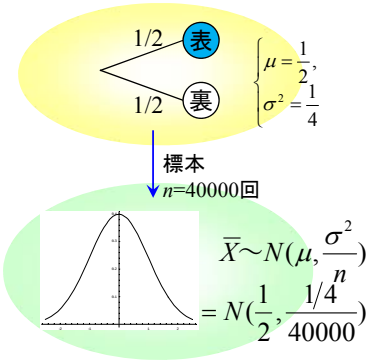
例えば，Excel2003で${}_{40000}C_{19600}$を計算すると，…

#NUM! =COMBIN(40000,19600)

計算不能！

中心極限定理の利用

- 中心極限定理 → $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- 標準化 → $Z := \frac{X-\mu}{\sigma} \sim N(0,1)$ → $Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- $X_i \sim Bi(1, 1/2)$ → $\begin{cases} \mu = E(X_i) = n_i p_i = 1 \times 1/2 = 1/2, \\ \sigma^2 = V(X_i) = n_i p_i (1-p_i) = 1 \times 1/2 \times 1/2 = 1/4 \end{cases}$



表が19600～20400回出る確率を求めたいので,

$$
\begin{aligned}
&P(19600 \le X_1 + \cdots + X_n \le 20400) \\
&= P\left(\frac{19600}{n} \le \frac{X_1 + \cdots + X_n}{n} \le \frac{20400}{n}\right) \\
&= P\left(\frac{\frac{19600}{n} - \mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{\frac{20400}{n} - \mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{\frac{19600}{40000} - \frac{1}{2}}{\sqrt{\frac{1}{4}}/\sqrt{40000}} \le Z \le \frac{\frac{20400}{40000} - \frac{1}{2}}{\sqrt{\frac{1}{4}}/\sqrt{40000}}\right) \\
&= P(-4 \le Z \le 4) \\
&= 0.99993\cdots
\end{aligned}
$$

中心極限定理の利用

- 例題2： 昨シーズン打率3割の打者が，今シーズン300回打席にたった．今シーズンの打率が4割以上となる確率は？
  - i 回目：X_i=1,0（1：ヒット，0：凡打）
  - ヒット数：$X=X_1+X_2+\ldots+X_n$

二項分布 $Bi(300, 3/10)$ に従う

$$f(x) = {}_n C_x p^x (1-p)^{n-x} \quad (x = 0,1,\cdots,n)$$
$$E(X) = np, V(X) = np(1-p)$$

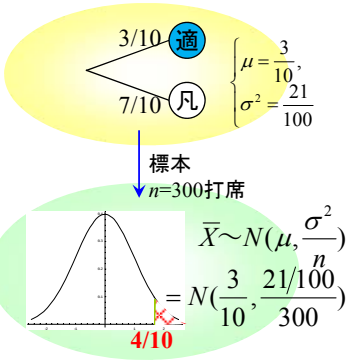
⇩

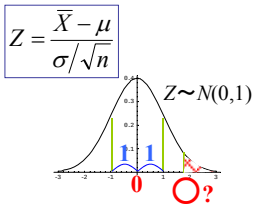
つまり $P(X > 120)$ はいくつか?

$$\sum_{x=120}^{300} {}_{300} C_x (3/10)^x (7/10)^{300-x}$$ を計算すればよい！

中心極限定理の利用

- 中心極限定理 → $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- 標準化 → $Z := \frac{X-\mu}{\sigma} \sim N(0,1)$ → $Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- $X_i \sim Bi(1, 3/10)$ → $\begin{cases} \mu = E(X_i) = n_i p_i = 1 \times 3/10 = 3/10, \\ \sigma^2 = V(X_i) = n_i p_i (1-p_i) = 1 \times 3/10 \times 7/10 = 21/100 \end{cases}$



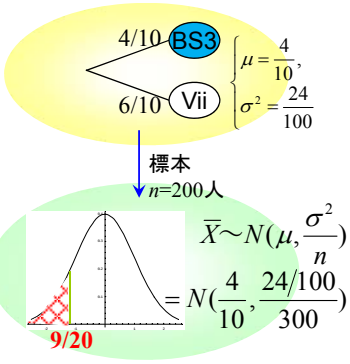


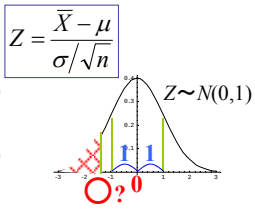
打率4割以上の確率を求めたいので，

$$
\begin{aligned}
&P(\bar{X} \geq 4/10) \\
&= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq \frac{4/10-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z \geq \frac{\frac{4}{10}-\frac{3}{10}}{\sqrt{\frac{21}{100}}/\sqrt{300}}\right) \\
&= P(Z \geq 3.7796\cdots) \\
&= 0.00007853\cdots
\end{aligned}
$$

中心極限定理の利用

- 例題3：2種類のゲーム機，ソニーのBlainStation3と任天童のViiの市場シェアはBS3が40%，Viiが60%である．ある店で，どちらかを買いに来た200人の客がいるとき，Viiが110台以上売れる確率は？（ただし，両方買う客はいないとする）





『Viiが110台以上売れる
＝BS3が90台以上売れない』だから，

$$
\begin{aligned}
&P(\bar{X} \leq 9/20) \\
&= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{9/20-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z \leq \frac{\frac{9}{20}-\frac{4}{10}}{\sqrt{\frac{24}{100}}/\sqrt{200}}\right) \\
&= P(Z \leq 0.8333\cdots) \\
&= 0.20327\cdots
\end{aligned}
$$

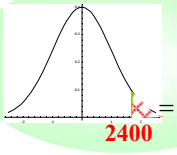
∴ 答え 20.3%

例題：出展 技術評論社「確率・統計の仕組みがわかる本」 例7.2

【問題】小学生の1ヶ月の小遣いが，平均2250円，標準偏差360円です．このとき，ランダムに選んだ36人の小学生の小遣い平均が2400円を超える確率は？

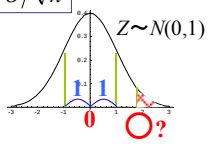
母集団
母平均 μ=2250円
母分散 σ^2=360^2

標本
n=36人



$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(2250, \frac{360^2}{36})$$

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



$$P(\bar{X} > 2400)$$
$$= P(\bar{X} - \mu > 2400 - \mu)$$
$$= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{2400-\mu}{\sigma/\sqrt{n}}\right)$$
$$= P\left(z > \frac{2400-2250}{360/\sqrt{36}}\right)$$
$$= P(z > -2.50)$$
$$\cong 0.0062097$$

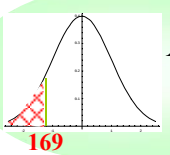
∴ 答え 0.62%

例題：

【問題】全国男子大学生の身長が，平均170cm，標準偏差5cmとします．このとき，ランダムに選んだ50人の大学生の平均身長が169cmを下回る確率は？

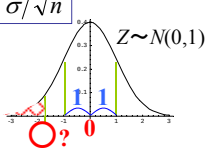
母集団
母平均 μ=170cm
母分散 σ^2=5^2

標本
n=50人



$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(170, \frac{5^2}{50})$$

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



$$P(\bar{X} < 169)$$
$$= P(\bar{X} - \mu < 169 - \mu)$$
$$= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{169-\mu}{\sigma/\sqrt{n}}\right)$$
$$= P\left(z < \frac{169-170}{5/\sqrt{50}}\right)$$
$$= P(z < -1.4142)$$
$$\cong 0.079270$$

∴ 答え 7%

Coffee Break!
10^{100}と100^{10}はどっちが大きい?

- どちらが大きい？
  - 10^{100} ＝ ?
  - 100^{10} ＝ ?
- どちらが大きい？
  - 10^{100} ＝ ?
  - $100!$ ＝ ?
- スターリングの公式

$$N! \approx (N/e)^N \sqrt{2\pi N}$$

$$\left(\lim_{N \to +\infty} \frac{N!}{(N/e)^N \sqrt{2\pi N}} = 1 \right)$$

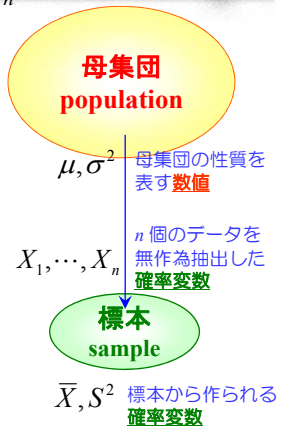
充分大きなNについて，Nの階乗の近似値を与える

累乗の計算も大変だけど，階乗の計算はとんでもなく大変ね!

標本分布：標本分散

- 母集団から抽出した標本数 n の標本 $X_1,\ldots X_n$ について，以下の確率変数を標本分散 S^2 という

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

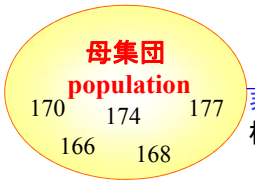


注意）「標本分散値」は確率変数「標本分散」が標本毎に実際に取る値

母集団と標本：標本分散値の平均

母分散と標本分散の関係

例：5人の身長



2人ずつ
非復元抽出
標本数 n=2

標本
sample

	標本分散値
(174,166)	16.0
(174,168)	9.0
(174,177)	2.3
(174,170)	4.0
(166,174)	16.0
:	:
(170,174)	4.0
(170,166)	4.0
(170,168)	1.0
(170,177)	12.3

標本分散値の平均 10.0

母集団数 N=5
母平均 μ=171.0
母分散 σ^2=16.0

母分散の $\frac{n-1}{n}$ 倍（無限母集団）
母分散の $\frac{N}{N-1}\cdot\frac{n-1}{n}$ 倍（有限母集団）

$$E(S^2)=\frac{n-1}{n}\sigma^2$$
$$\left(E(S^2)=\frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2\right)$$

Excel

補足：標本分散の平均と母分散の関係（証明）

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n}\left\{(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2\right\}\right)\\
&= \frac{1}{n}E\left(\left\{(X_1-\mu)-(\bar{X}-\mu)\right\}^2+\cdots+\left\{(X_n-\mu)-(\bar{X}-\mu)\right\}^2\right)\\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}\left\{(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\right\}\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}E(X_i-\mu)^2-2E\left(\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu)\right)+\sum_{i=1}^{n}E(\bar{X}-\mu)^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}V(X_i)-2E\left(n(\frac{X_1+\cdots+X_n}{n}-\mu)(\bar{X}-\mu)\right)+nE(\bar{X}-\mu)^2\right)\\
&= \frac{1}{n}\left(n\sigma^2-2nE(\bar{X}-\mu)^2+nE(\bar{X}-\mu)^2\right)\\
&= \sigma^2-V(\bar{X})\\
&= \sigma^2-\frac{N-n}{N-1}\cdot\frac{1}{n}\sigma^2\\
&= \frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2
\end{aligned}$$

補足：有限母集団修正

- 母集団が有限の場合
  - 標本分散の平均と母分散の関係は，

$$E(S^2)=\frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2$$

有限修正項

母集団の要素数Nが大きくないとき，有限修正項を考慮．
無限母集団（Nが十分大きい）時は，有限修正項は1となるので無視．

- 母集団が無限の場合
  - 標本分散の平均と母分散の関係は，

$$E(S^2)=\frac{n-1}{n}\sigma^2$$

母集団と標本：標本分散（まとめ）

- 標本分散 S^2

母集団からn個
無作為抽出

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

- $X_1,\ldots,X_n$はそれぞれ確率変数
- それから作られる標本平均も確率変数
- よって，それから作られる標本分散も確率変数

- 注意：「**標本平均の分散** $V(\bar{X})$」と「**標本分散の平均** $E(S^2)$」を混同しないこと！

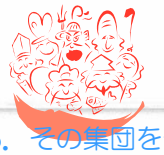
「標本分散値の平均」と「母分散」の関係

$$E(S^2)=\frac{n-1}{n}\sigma^2$$

有限母集団の場合：

$$E(S^2)=\frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2$$

演習２：標本分散



1．世界に4匹しかいない貴重な昆虫がいる．その集団を母集団としよう．
- 神様はこの4匹の全長を全て知っており，それぞれ (2, 6, 7, 5) である．
- 神様は母分散の値を求めた．いくつか？ $\sigma^2 = ?$

2．探検家は2匹捕まえる．それが標本となる．
- 各探検家は重複なく2匹を捕まえた．
（つまり，非復元抽出で2匹捕らえ，全長測定後放す）
- 各探検家は自分が捕まえた2匹の標本の分散の値を求めた．
- それぞれ，いくつか？ 全ての組合せについて計算せよ．$S^2 = ?$

3．１と２の結果から，$E(S^2) = \frac{N}{N-1} \cdot \frac{n-1}{n} \sigma^2$ が成立することを確認しよう．

ただし，Nは母集団の大きさ，nは標本の大きさである．

標本分布：標本分散と不偏分散

- 標本分散 S^2

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

- 不偏分散 s^2 ← この標本分散は，母分散σ^2の不偏推定量

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

$$E(S^2) = \frac{n-1}{n}\sigma^2 \qquad E(s^2) = \sigma^2$$

有限母集団の場合：

$$E(S^2) = \frac{N}{N-1} \cdot \frac{n-1}{n}\sigma^2 \qquad E(s^2) = \frac{N}{N-1}\sigma^2$$

Nが充分大きいならば，$N/(N-1)$は1と考えて良い．

標本分布：標本分散の従う確率分布

- 標本分散S^2 はどんな確率分布に従うのか？

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

$$\rightarrow \frac{n}{\sigma^2}\cdot S^2 = \frac{n}{\sigma^2}\cdot\frac{1}{n}\left\{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\}$$

$$= \left(\frac{X_1 - \bar{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \bar{X}}{\sigma}\right)^2$$

…

χ^2分布に従う ← n個の$N(0,1)$に従う確率変数の二乗和

$\sum(X_i - \bar{X}) = 0$ という制限のため，自由に動ける変数の個数はn-1となる．

- 母集団が正規分布 $N(\mu,\sigma^2)$ に従うとみなせる時，確率変数 $\frac{nS^2}{\sigma^2}$ は自由度n-1のχ^2 (n-1)分布に従う．

標本分布：標本分散の従う確率分布

- 標本分散S^2 はどんな確率分布に従うのか？

$$\chi^2 = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

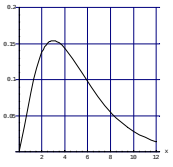
$$S^2 = \frac{1}{n}\left\{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right\}$$

母集団
母平均 μ
母分散 σ^2

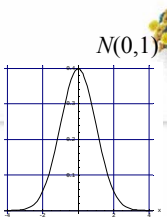
標本 n

標本
標本平均 $\bar{X}$
標本分散 S^2

$$\chi^2 = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$



χ^2 分布とは？



標準正規分布 $N(0,1)$ に従う，互いに独立な n個の確率変数 $Z_1,\ldots,Z_n$ を考える

$$\chi^2 = Z_1^2 + \cdots + Z_n^2$$ ← 二乗和をとる

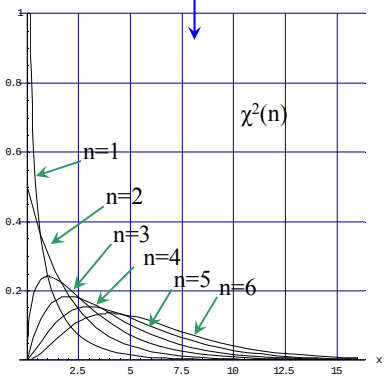
新たな確率変数

⇩

この確率変数χは，自由度n のχ^2分布に従う！

互いに自由に値をとることが出来る確率変数の個数

標本から母分散σ^2を推定
「カイ二乗推定」「カイ二乗検定」



標本分布：標本分散

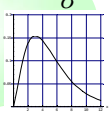
例題：道ばたの雑草の背丈の平均μ=50cm,分散σ^2=25だとしよう．標本として10本の雑草を抜いて調べたとき，その分散が50を超える確率は？

母集団
母平均 μ=50cm
母分散 σ^2=25

⇩ 標本 n=10本

標本
標本平均 $\overline{X}$
標本分散 S^2

$$\chi^2 = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$



$$P(S^2 > 50)$$
$$= P(\frac{\chi^2\sigma^2}{n} > 50) \quad \left[\because \chi^2 = \frac{nS^2}{\sigma^2}\right]$$
$$= P(\chi^2 > 50\frac{n}{\sigma^2})$$
$$= P(\chi^2 > 50\frac{10}{25} = 20) \in (0.025, 0.010)$$

自由度9のχ^2分布表から
$P(\chi^2(9)>19.0228)=0.025$
$P(\chi^2(9)>21.6660)=0.010$

＝0.017912
(Excel関数 CHIDISTより)

t 分布とは？

ギネスビールとは？
1756年創業のビール醸造会社
〔ダブリン（アイルランド）〕
ギネスビール（黒スタウト）を製造

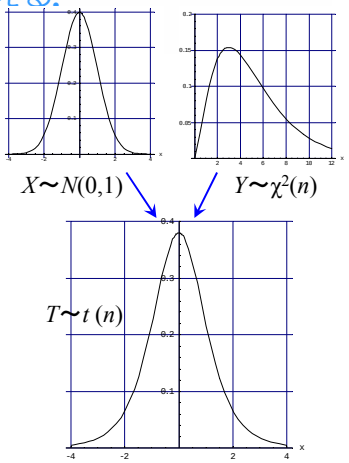
- 2個の互いに独立な確率変数 X, Y を考える.
  - X ： 標準正規分布 $N(0,1)$ に従う
  - Y ： 自由度 n のχ^2分布 $\chi^2(n)$ に従う

$$T := \frac{X}{\sqrt{Y/n}}$$

新たな確率変数

確率変数 T は，自由度 n の t 分布に従う！

Student の t分布
ゴセット (1876-1937)



ビール会社ギネスGuinessでビールの品質管理
標本が小さいとき，分散の値が(正規分布では上手くいかない…)
→ ***t*** 分布の発見（"Student"[W.S.Gossett] '*The probable error of a mean*',Biometrika vol.6,1908）

標本分布：標本平均と標本分散

- 標本平均 $\overline{X}$ の標準化

$$\overline{X} \to Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

標準正規分布
$N(0, 1)$ に従う

- 標本分散S^2にn/σ^2を掛けた確率変数

$$nS^2/\sigma^2$$

自由度 n-1 の
χ^2分布 に従う

$$T = \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) \cdot \frac{1}{\sqrt{\frac{1}{n-1}\cdot\frac{nS^2}{\sigma^2}}} = \frac{\overline{X}-\mu}{S/\sqrt{n-1}}$$

自由度 n-1 の
t 分布 に従う

標本から母平均μを推定
「t推定」「t検定」

標本分布：確率変数Tの従う分布

確率変数Tは，自由度 n-1 の t 分布 に従う

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n-1}}\sim t(n-1)$$

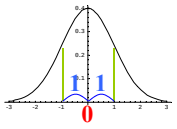
母集団
母平均 μ
母分散 σ^2

標本n

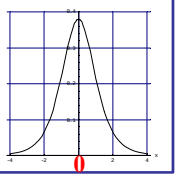
標本
標本平均 $\overline{X}$
標本分散 S^2

$$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$$

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$



$$T=\frac{\overline{X}-\mu}{S/\sqrt{n-1}}\sim t(n-1)$$



標本分布：標本平均（その2）

例題：350ml缶のジュース雑草の背丈の平均μ=50cm,分散σ^2=25だったとしよう．標本として20本のジュースを調べたら，標本分散が5

母集団
母平均 μ=50cm
母分散 σ^2=25

標本
n=20本

標本
標本平均 $\overline{X}$
標本分散 S^2

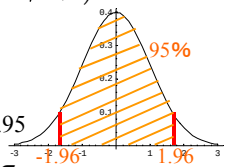
$$\begin{aligned}&P(\overline{X}<345)\\&=P(\overline{X}-\mu<345-\mu)\\&=P\left(\frac{\overline{X}-\mu}{S/\sqrt{n-1}}<\frac{345-\mu}{S/\sqrt{n-1}}\right)\\&=P\left(T<\frac{345-\mu}{S/\sqrt{n-1}}\right)\end{aligned}$$

補足：必要な標本の大きさ

- 標本平均の実現値を母平均の推定値とする場合

$$|\bar{X}-\mu| \le \varepsilon \quad \left(\bar{X} \sim N(\mu, \sigma^2/n)\right)$$

誤差　許容誤差



今，標本平均の従う正規分布から考えて

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow P(-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96) = 0.95$$

$$\Leftrightarrow P(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$\Leftrightarrow P(|\bar{X}-\mu| \le 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

従って，許容誤差をεとしたとき

$$\Rightarrow 1.96\frac{\sigma}{\sqrt{n}} \le \varepsilon$$

$$\Leftrightarrow n \ge \frac{(1.96\sigma)^2}{\varepsilon^2}$$

参考：
有限母集団の場合

$$n \ge = \frac{1}{\frac{\varepsilon^2}{4\sigma^2}\left(1-\frac{1}{N}\right)+\frac{1}{N}}$$

$$\left(S^2 = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\right)$$

定められた許容誤差ε>0に対し，母集団の大きさNと母標準偏差σが既知の場合，単純無作為抽出の大きさnを，左不等式を満たすようにとれば，95%以上の確率で，誤差を許容誤差より小さくできる．

補足：必要な標本の大きさ

- 例題：大きさ6000万の母集団の母比率pを，95%の確率で誤差が0.05以下になるようにしたい．必要な単純無作為抽出の大きさnはいくらか？ $|\bar{X}-\mu| \le 0.05$

Nが十分大きいので，

$$n \ge \frac{(1.96)^2\sigma^2}{\varepsilon^2} \ge \frac{(1.96)^2}{4\varepsilon^2} = \frac{(1.96)^2}{4(0.05)^2} \approx 384.16$$

$$\left(\sigma^2 = p(1-p) = -(p-\frac{1}{2})^2 + \frac{1}{4} \le \frac{1}{4}\right)$$

σ^2の最大値は
0.25(p=0.5の時)

参考文献

- 東京大学教養学部統計学教室編「統計学入門」東京大学出版会（1991）
- 村上雅人「なるほど統計学」海鳴社（2002）
- 田栗正章他「やさしい統計入門」講談社（2007）
- 鈴木達三・高橋宏一「標本抽出の計画と方法」放送大学（1991）
- 永田靖「サンプルサイズの決め方」朝倉書店（2003）

- 高橋信[著]・トレンドプロ「マンガでわかる統計学」オーム社（2004）
- 丹慶勝市「図解雑学 統計解析」ナツメ社（2003）
- 白石修二「例題で学ぶ Excel統計入門」森北出版（2001）
- 東京大学教養学部統計学教室編「自然科学の統計学」東京大学出版会（1992）