統計の分析と利用

3. 母集団と標本

堀田　敬介

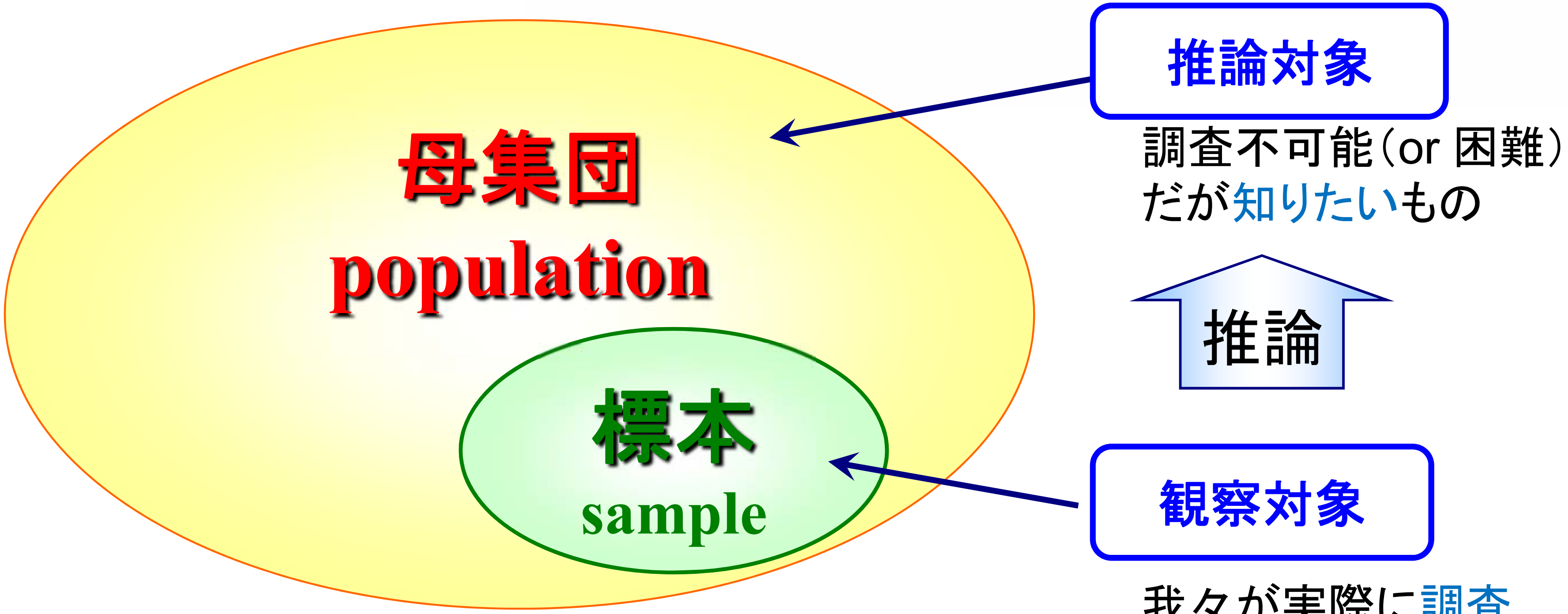
2013/11/22,Fri.～

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



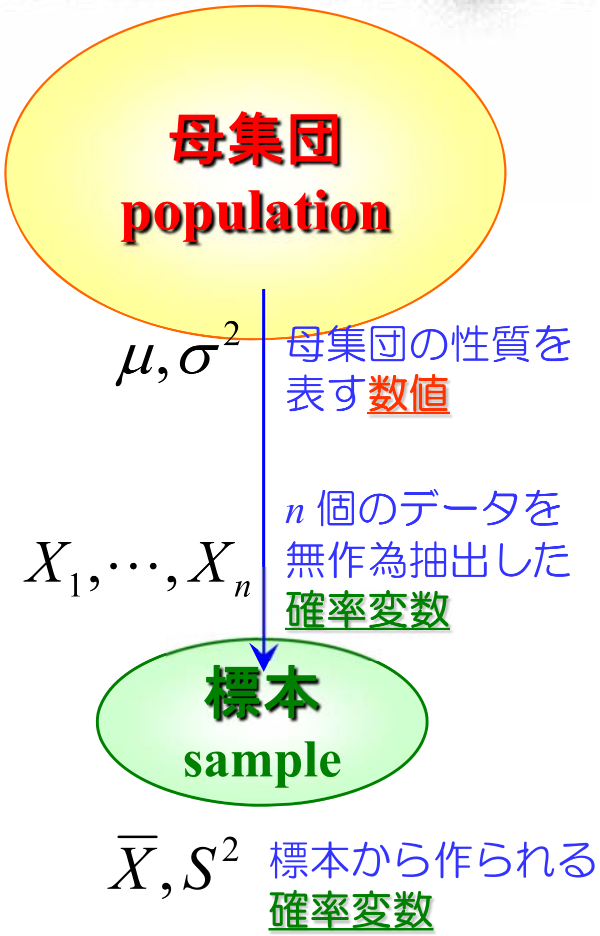
我々が実際に調査可能なデータ

- ○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，<u>無限母集団</u>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

無作為抽出には乱数などを利用

$$X_1, \cdots, X_n$$

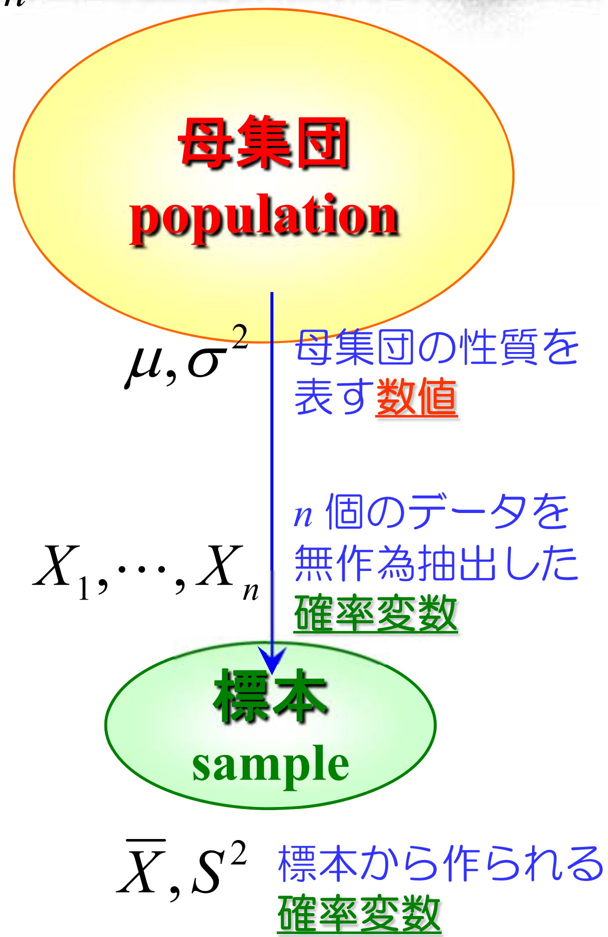
  - $X_1,\ldots,X_n$ は互いに独立な確率変数
    - 標本調査は試行：無作為抽出により，実際に取る値は偶然による
  - 各確率変数 X_i は母集団と同じ分布に従う
  - n はサンプルサイズ（抽出した標本数）

- 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<u>標本平均 $\overline{X}$</u>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）

母集団
population

170　174　177
166　168

2人ずつ
非復元抽出
標本数 n=2

標本
sample

標本平均の値

(174,166) →	170.0
(174,168) →	171.0
(174,177) →	175.5
(174,170) →	172.0
(166,174) →	170.0
:	:
(170,174) →	172.0
(170,166) →	168.0
(170,168) →	169.0
(170,177) →	173.5

標本平均値の平均　171.0

標本平均値の分散　6.0

母集団数 N=5
母平均 μ=171.0
母分散 σ^2=16.0

一致する！

母分散の $\frac{1}{n}$ 倍（無限母集団）

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

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

Excel

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）

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

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

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

補足：有限母集団修正

- 母集団が有限の場合
  - <u>標本平均の分散</u>と<u>母分散</u>の関係は,

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

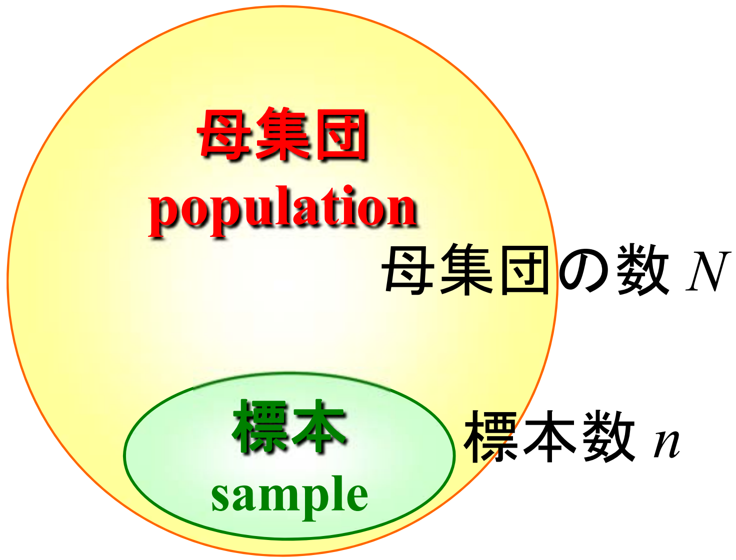
有限修正項

<u>Nが余り大きくない</u>場合や,
<u>n/Nが大きい</u>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
  - <u>標本平均の分散</u>と<u>母分散</u>の関係は,

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

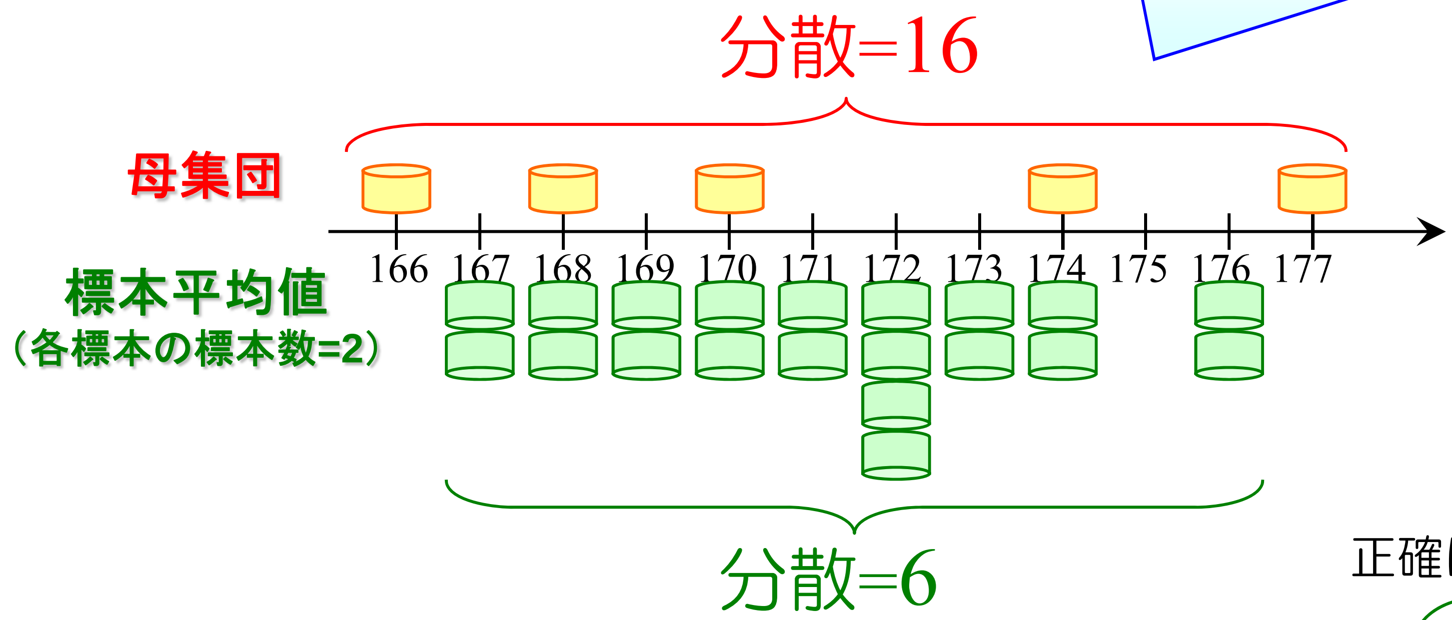


補足：母集団と標本：標本平均

- 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 $V(\overline{X}) < \sigma^2$ 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**！



$$6 = V(\overline{X}) < \sigma^2 = 16$$

正確には有限母集団なので以下

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

母集団と標本：標本平均（まとめ）

母集団からn個
無作為抽出

- 標本平均

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

•$X_1,\ldots,X_n$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

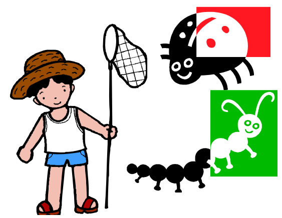
- 注意：「標本平均」と「標本平均値」は意味が違う
  - 標本平均　…　上で定義される確率変数
  - 標本平均値　…　確率変数「標本平均」が標本ごとに実際に取る値
- 「標本平均 $\overline{X}$ の期待値は母平均μに等しい」 $E(\overline{X}) = \mu$
- 「標本平均 $\overline{X}$ の分散は母分散σ^2の$1/n$に等しい」 $V(\overline{X}) = \dfrac{\sigma^2}{n}$

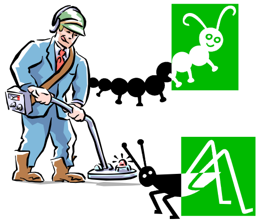
（有限母集団の場合：$V(\overline{X}) = \dfrac{N-n}{N-1} \cdot \dfrac{\sigma^2}{n}$）

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．

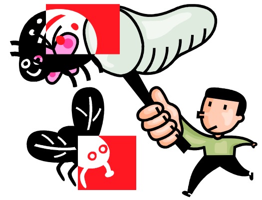
- 神様はこの4匹の全長を全て知っており，それぞれ (2, 6, 7, 5) である．
- 神様は母平均の値を求めた．いくつか？ $\mu = ?$
- 神様は母分散の値を求めた．いくつか？ $\sigma^2 = ?$





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？ <u>全ての組合せについて</u>計算せよ． $\overline{X} = ?$



３．１と２の結果から，$E(\overline{X}) = \mu$ と $V(\overline{X}) = \dfrac{N-n}{N-1} \cdot \dfrac{\sigma^2}{n}$ 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 $\bar{X}$ 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」 $V(\bar{X}) = \dfrac{\sigma^2}{n}$

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

大数の法則

標本数 n が大きくなるにつれて，標本平均

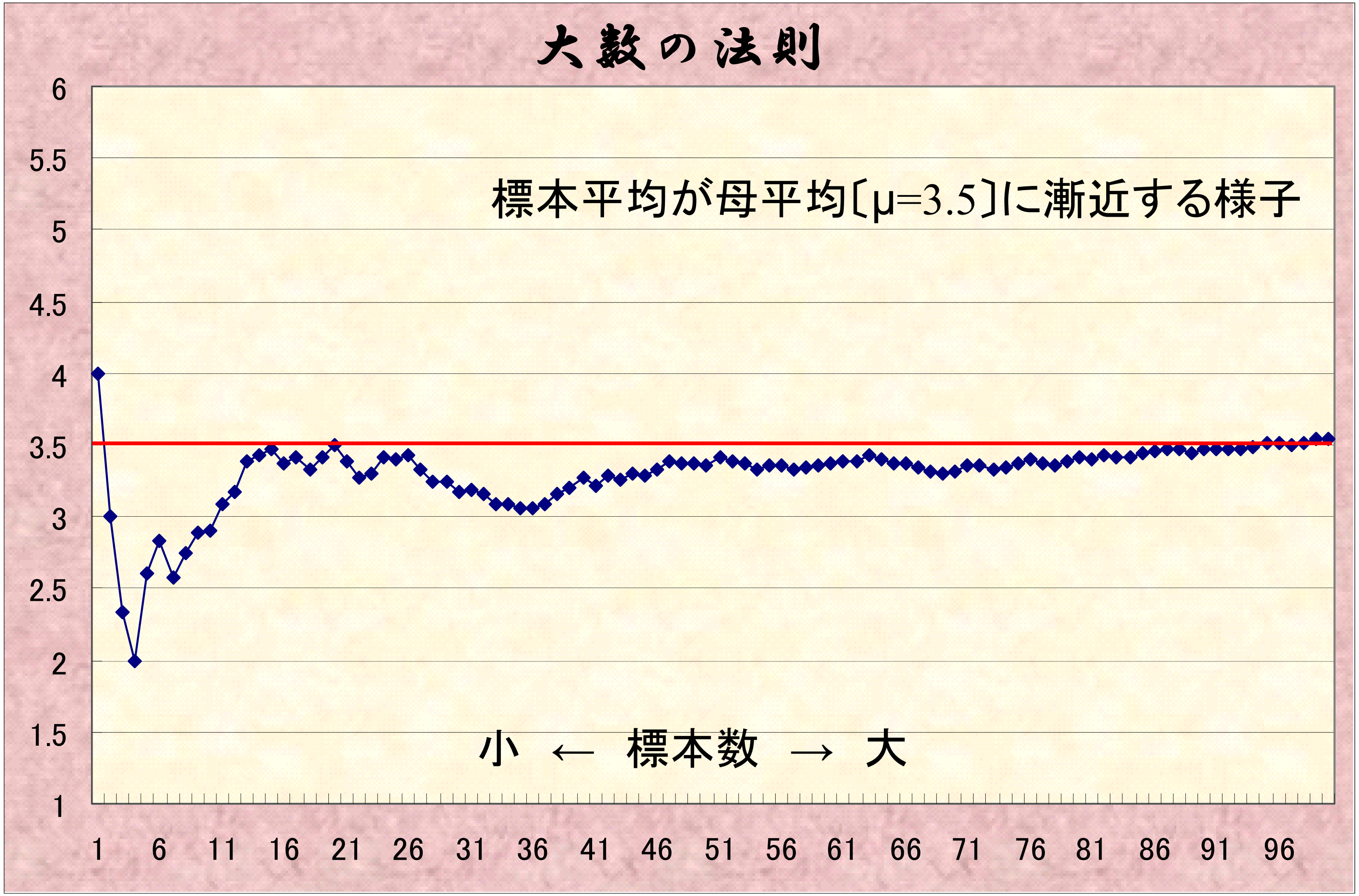
$$\bar{X} = \frac{1}{n}(X_1 + \cdots X_n)$$

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 $P\left(\left|\overline{X}-\mu\right|>k\sigma\right)\leq 1/k^2$ から

∵）$X_1,\ldots,X_n$ は独立で，同じ分布に従う

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

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

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

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

標本分布

- 標本平均 $\overline{X}$ が従う確率分布

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 $\overline{X}$ 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

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

中心極限定理

母集団
population

母平均 μ
母分散 σ^2

一様分布
正規分布
幾何分布
二項分布
ポアソン分布　指数分布
…

n個とってくる

標本数 n が<u>十分大きい</u>なら

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

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

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

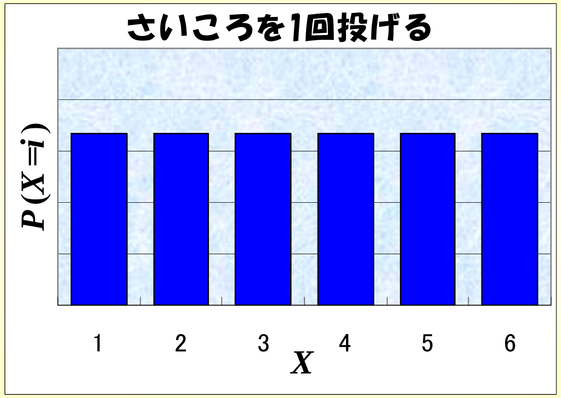
中心極限定理

母集団
population

母平均 μ
母分散 σ^2

↓ n個とってくる

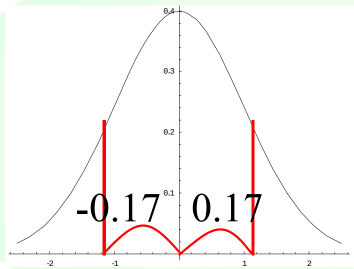
標本
sample
標本平均 $\overline{X}$
標本分散 S^2



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

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，

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

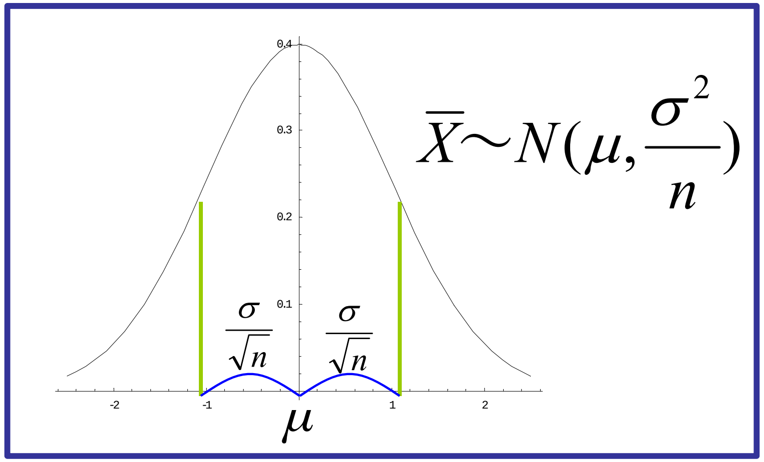
としてよいということ．

(右辺のφは標準正規分布の累積分布関数)

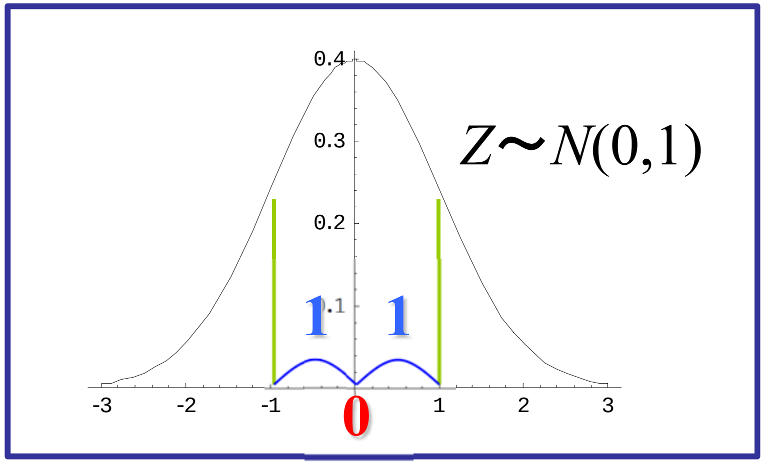
標本分布：標本平均の標準化

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

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



標準化
が
2つの世界
の
架け橋



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 $\overline{X}$ 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

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

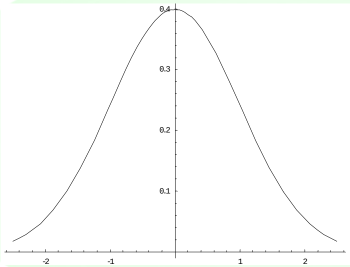
1/2 表

1/2 裏

$$\begin{cases} \mu = \frac{1}{2}, \\ \sigma^2 = \frac{1}{4} \end{cases}$$

標本

n=40000回



$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(\frac{1}{2}, \frac{1/4}{40000})$$

表が19600～20400回出る確率を求めたいので，

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

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

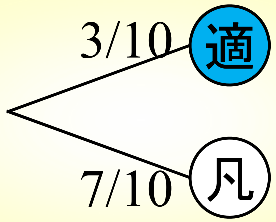
$$f(x)={}_nC_x p^x (1-p)^{n-x} \quad (x=0,1,\cdots,n)$$
$$E(X)=np, V(X)=np(1-p)$$

つまり $P(X \geqq 120)$ 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

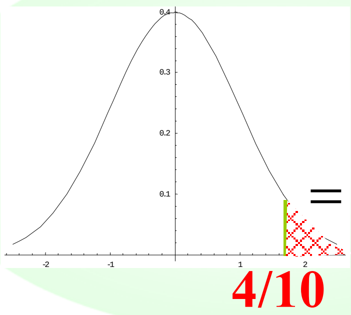
- 中心極限定理 ⇒ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- 標準化 ⇒ $Z := \frac{X-\mu}{\sigma} \sim N(0,1)$ ⇒ $Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- $X_i \sim Bi(1, 3/10)$ ⇒ $\begin{cases} \mu = E(X_i) = n_i p_i = 1 \times 3/10 = 3/10, \\ \sigma^2 = V(X_i) = n_i p_i (1 - p_i) = 1 \times 3/10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



$\begin{cases} \mu = \frac{3}{10}, \\ \sigma^2 = \frac{21}{100} \end{cases}$

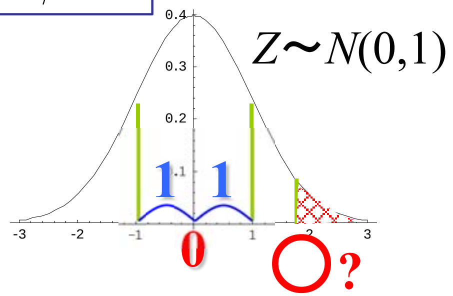
標本

n=300打席



$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(\frac{3}{10}, \frac{21/100}{300})$$

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

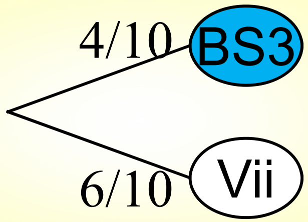


打率4割以上の確率を求めたいので,

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

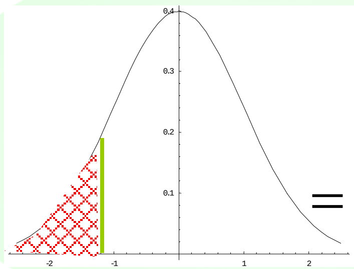
中心極限定理の利用

- 例題3：2種類のゲーム機，ゾニーの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）



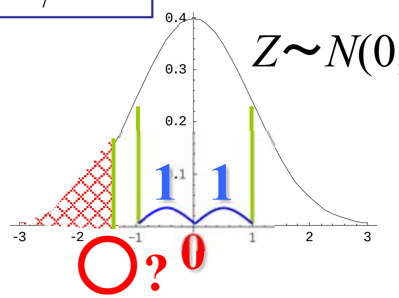
$$\begin{cases} \mu = \dfrac{4}{10}, \\ \sigma^2 = \dfrac{24}{100} \end{cases}$$

標本 n=200人



$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\frac{4}{10}, \frac{24/100}{300}\right)$$

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



$Z \sim N(0,1)$

『Viiが110台以上売れる
＝BS3が90台以上売れない』だから，

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

$\therefore$ 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

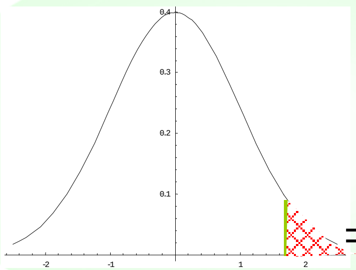
【問題】小学生の1ヶ月の小遣いが，平均2250円，標準偏差360円です．このとき，ランダムに選んだ36人の小学生の小遣い平均が2400円を超える確率は？

母集団

母平均 μ=2250円

母分散 σ^2=360^2

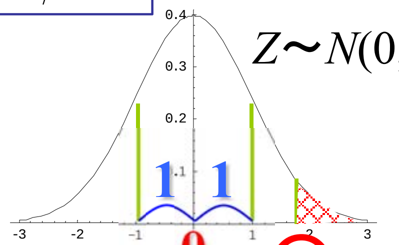
↓ 標本 n=36人



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

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

$Z \sim N(0,1)$



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

∴ 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？

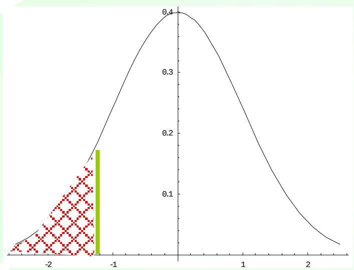
母集団

母平均 μ=170cm

母分散 $\sigma^2=5^2$

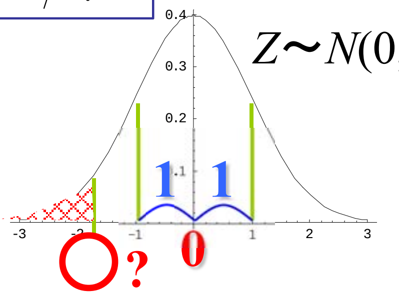
標本

n=50人



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

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



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

$\therefore$ 答え 7%

Coffee Break!

100^{10}と10^{100}はどっちが大きい?

- どちらが大きい？
  - 100^{10} ＝？ $100^{10} = 100{,}000{,}000{,}000{,}000{,}000{,}00$（0が20個）
  - 10^{100} ＝？ $10^{100} = 100{,}000{,}000{,}000{,}000{,}000{,}000{,}\ldots\ldots\ldots\ldots\ldots,000$（0が100個）

- どちらが大きい？
  - 10^{100} ＝？ $10^{100} = 100,\ldots,000$（0が100個）
  - $100!$ ＝？ $100! \simeq 900,\ldots,000$（0が157個）

累乗の増え方は
凄いね
階乗の増え方は
もっと凄いね！

- スターリングの公式

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

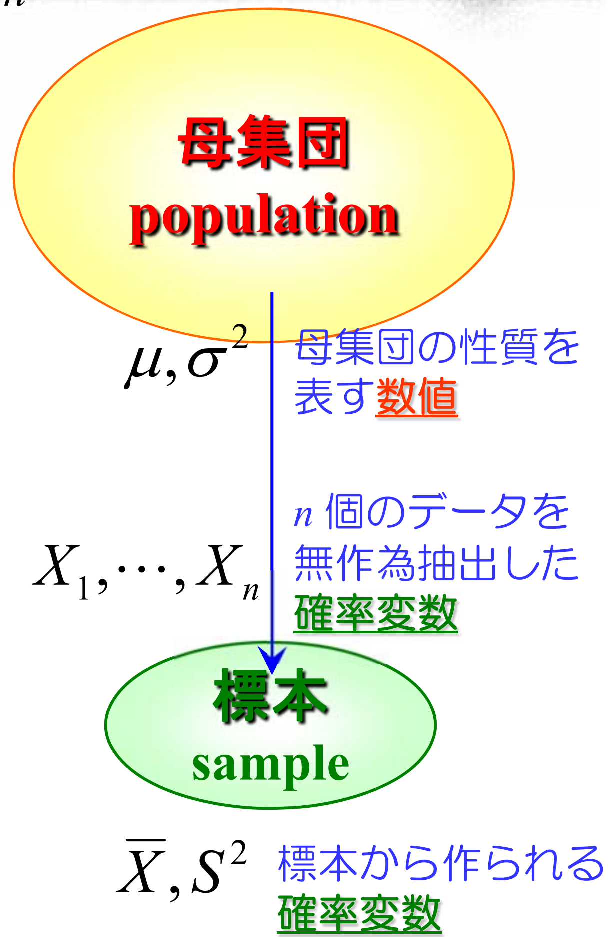
$$\left(\lim_{N\to+\infty} \frac{N!}{(N/e)^N \sqrt{2\pi N}} = 1 \right)$$

充分大きなNについて，Nの階乗の近似値を与える

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

母集団
population

170　174　177
166　168

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）

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

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）

母集団からn個
無作為抽出

- 標本分散 S^2

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

- $X_1,\ldots,X_n$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** $V(\overline{X})$」と「**標本分散**の**平均** $E(S^2)$」を混同しないこと！

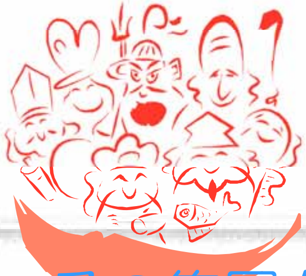
「標本分散値の平均」と「母分散」の関係

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

有限母集団の場合：

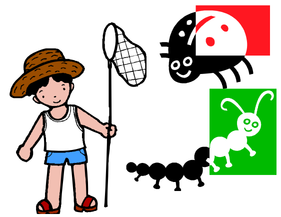
$$E(S^2) = \frac{N}{N-1}\cdot\frac{n-1}{n}\sigma^2$$

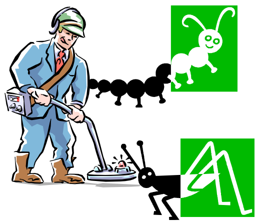
演習２：標本分散



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？ <u>全ての組合せについて</u>計算せよ．$S^2 = ?$

３．１と２の結果から， $E(S^2) = \dfrac{N}{N-1} \cdot \dfrac{n-1}{n} \sigma^2$ 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σ²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？

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

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

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

…

χ^2分布に従う ← n個の$N(0,1)$に従う確率変数の二乗和

$\sum(X_i-\overline{X})=0$ 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 $\dfrac{nS^2}{\sigma^2}$ は<u>自由度n-1のχ^2(n-1)分布</u>に従う．

標本分布：標本分散の従う確率分布

- 標本分散S² はどんな確率分布に従うのか？

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

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

母集団

母平均 μ

母分散 σ^2

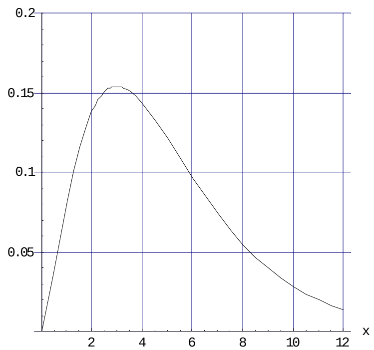
$\downarrow$ 標本 n

標本

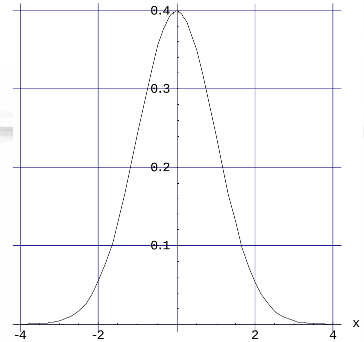
標本平均 $\overline{X}$

標本分散 S^2

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



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

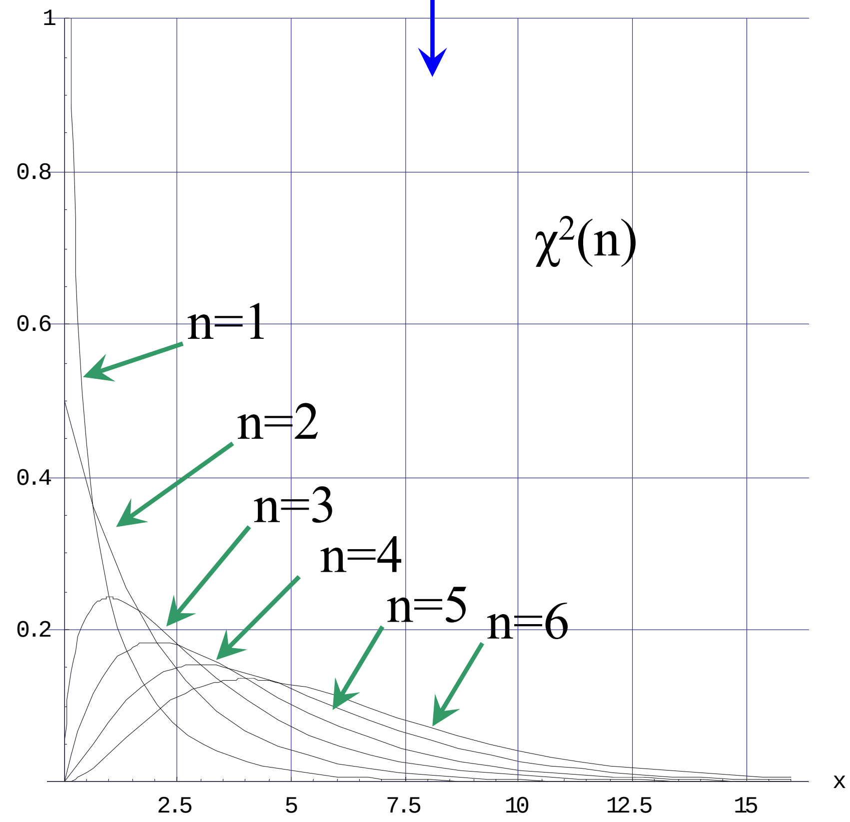
$$\underbrace{\chi^2}_{\text{新たな確率変数}} = Z_1^2 + \cdots + Z_n^2$$

← 二乗和をとる

新たな確率変数

この確率変数 χ^2 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

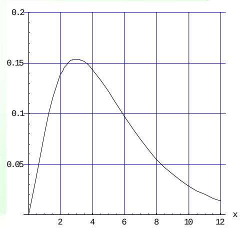
例題：道ばたの雑草の背丈の平均μ=50cm,分散σ^2=25だとしよう．標本として10本の雑草を抜いて調べたとき，その分散が50を超える確率は？

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

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



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

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

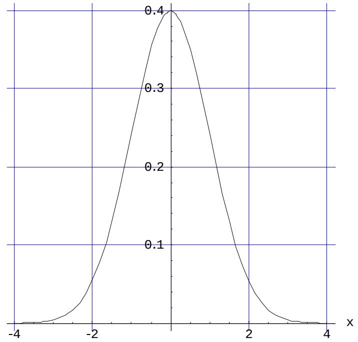
- 2個の互いに独立な確率変数 X, Y を考える．
  - X ： 標準正規分布 $N(0,1)$ に従う
  - Y ： 自由度 n のχ^2分布 $\chi^2(n)$ に従う

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

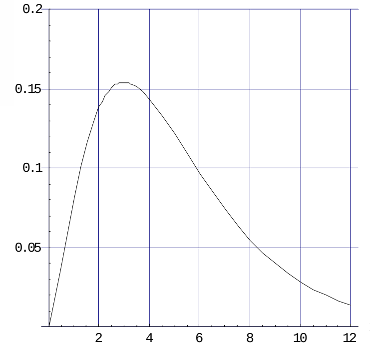
新たな確率変数

確率変数 T は，自由度 n の t 分布に従う！

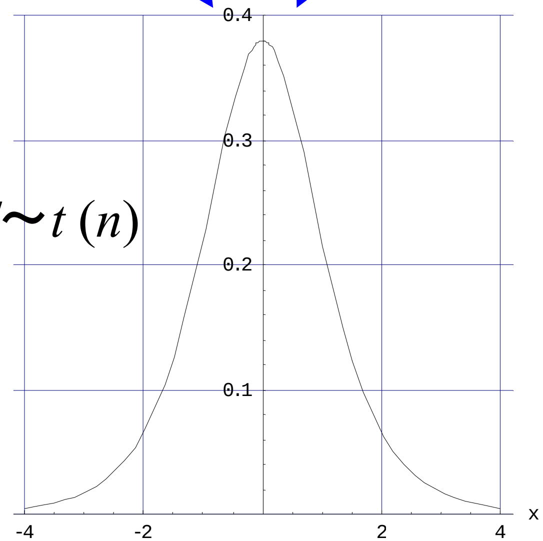
Student の t分布
ゴセット (1876-1937)



$X \sim N(0,1)$



$Y \sim \chi^2(n)$



$T \sim t\,(n)$

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 $\bar{X}$ の標準化

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

標準正規分布
$N(0, 1)$ に従う

- 標本分散S²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

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

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

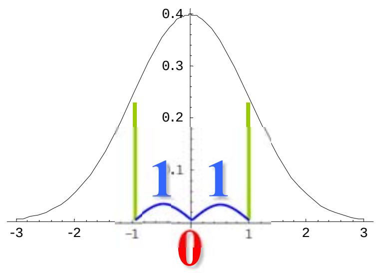
標本

標本平均 $\overline{X}$

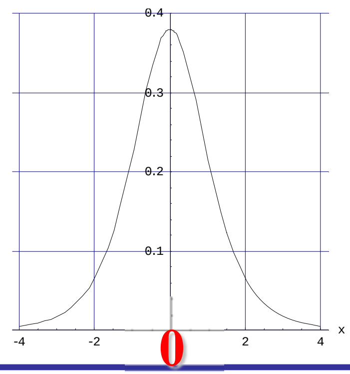
標本分散 S^2

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

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



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



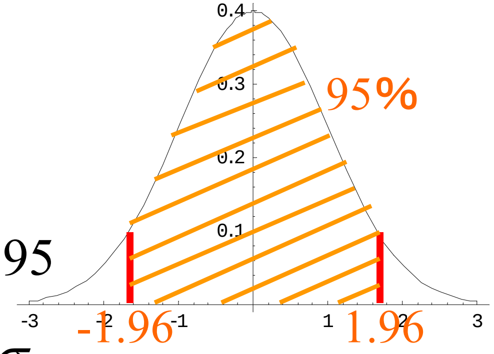
補足：必要な標本の大きさ

標本平均の実現値を母平均の推定値とする場合

$$\left|\underset{\text{誤差}}{\overline{X}-\mu}\right| \le \underset{\text{許容誤差}}{\varepsilon} \qquad \left(\overline{X} \sim N(\mu, \sigma^2/n)\right)$$

今，標本平均の従う正規分布から考えて

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



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

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

参考：
有限母集団の場合

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

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

従って，許容誤差をεとしたとき

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

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

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？

$$\left|\bar{X}-\mu\right| \leq 0.05$$

Nが十分大きいので，

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

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

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）