

# 統計の分析と利用

## 3. 母集団と標本



堀田 敬介

2012/11/30, 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
- 🌻  $\chi^2$ 分布

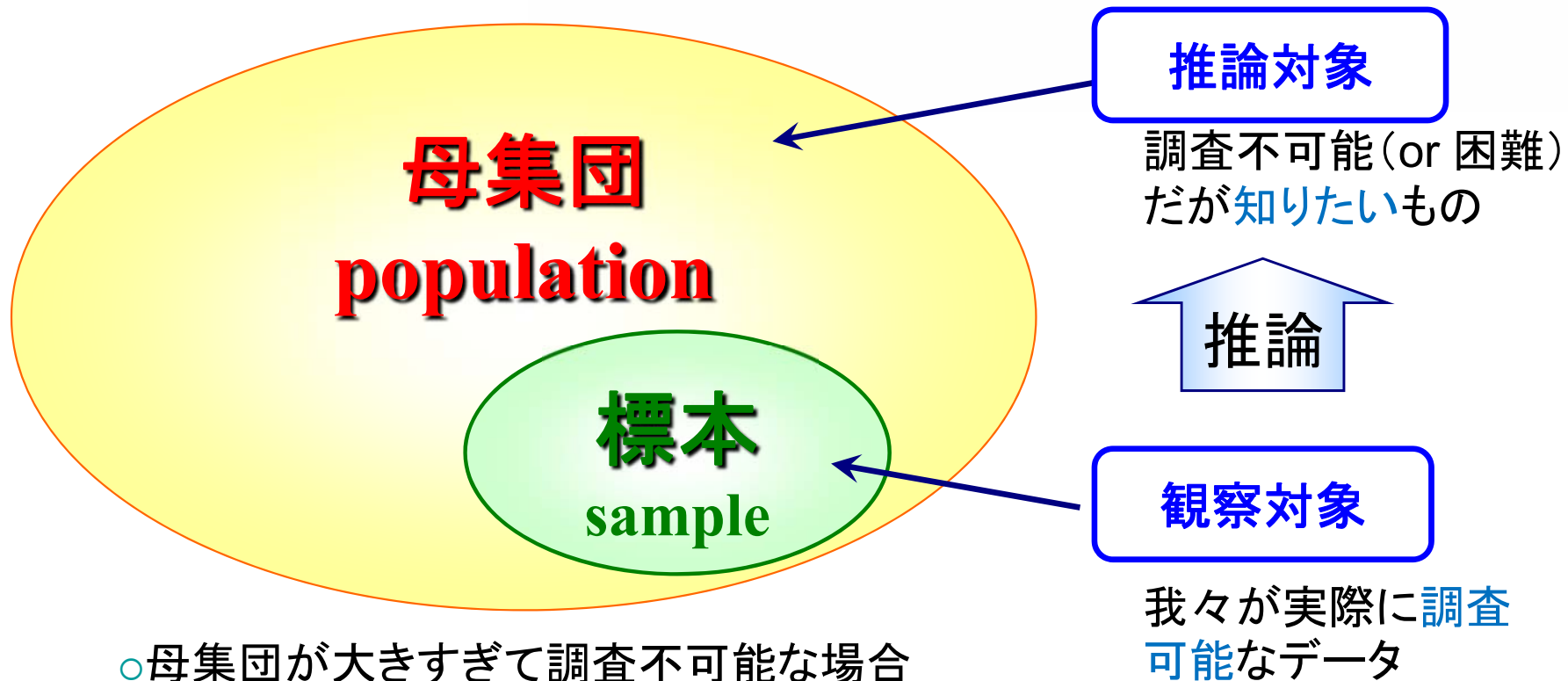
### 🌻 母比率の推測

#### 🌻 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： $\mu$

母分散： $\sigma^2$ （母標準偏差： $\sigma$ ）

## 母集団からの標本

データ  $n$  個を無作為抽出

$$X_1, \dots, X_n$$

無作為抽出には乱数などを利用

$X_1, \dots, X_n$  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、実際にとる値は偶然による

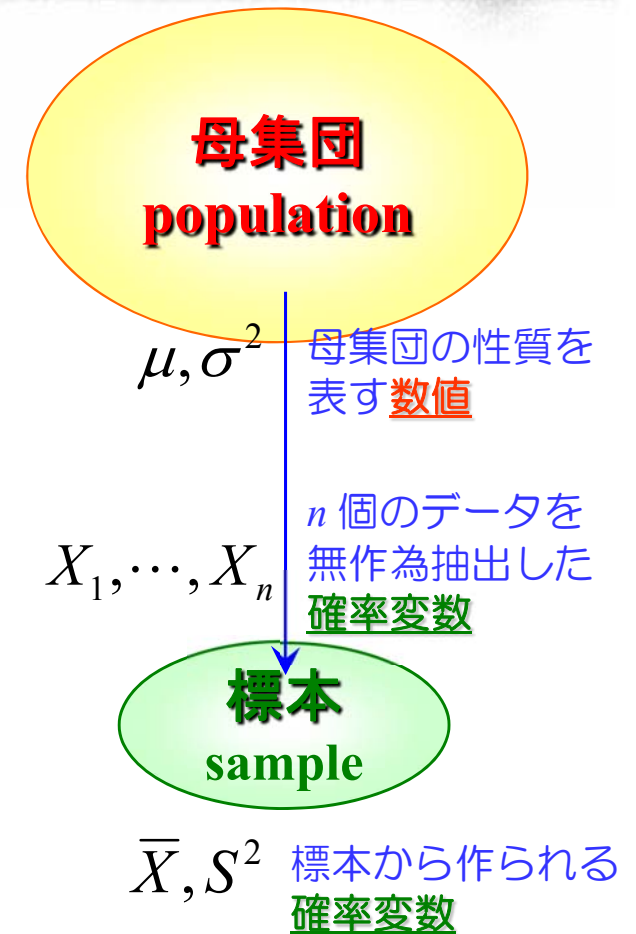
各確率変数  $X_i$  は母集団と同じ分布に従う

$n$  はサンプルサイズ（抽出した標本数）

## 確率変数 $X_1, \dots, X_n$ から作られる確率変数

標本平均： $\bar{X} = \frac{X_1 + \dots + X_n}{n}$

標本分散： $S^2 = \frac{1}{n} \left\{ (X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right\}$

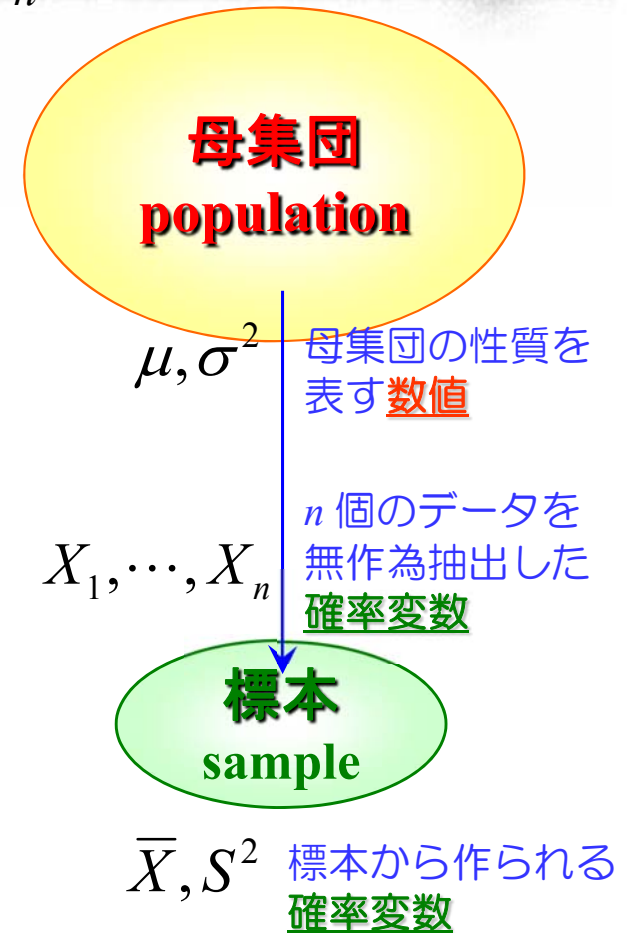


# 標本分布：標本平均



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

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$



注意) 「標本平均」は確率変数 「標本平均値」が標本毎に実際にとる値

# 母集団と標本：標本平均



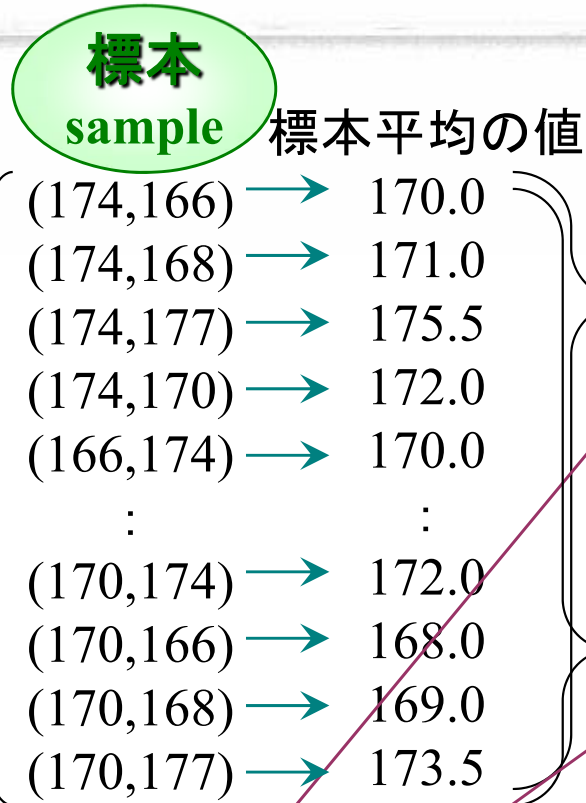
## 標本平均と母平均の関係

例：5人の身長

(170, 174, 166, 168, 177)



2人ずつ  
非復元抽出  
標本数  $n=2$



171.0  
標本平均値  
の平均

6.0  
標本平均値  
の分散

母集団数  $N=5$

母平均  $\mu=171.0$

母分散  $\sigma^2=16.0$

**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)

$$E(\bar{X}) = \mu$$

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

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



# 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）



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

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

$$= \frac{1}{n^2} \left\{ \sum_{i=1}^n E(X_i - E(X_i))^2 + 2\sum_{i<j} (X_i - E(X_i))(X_j - E(X_j)) \right\}$$

$$= \frac{1}{n^2} \left\{ \sum_{i=1}^n V(X_i) + 2\sum_{i<j} Cov(X_i, X_j) \right\}$$

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

$$= \frac{1}{n} \cdot \frac{N-n}{N-1} \sigma^2$$

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

# 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

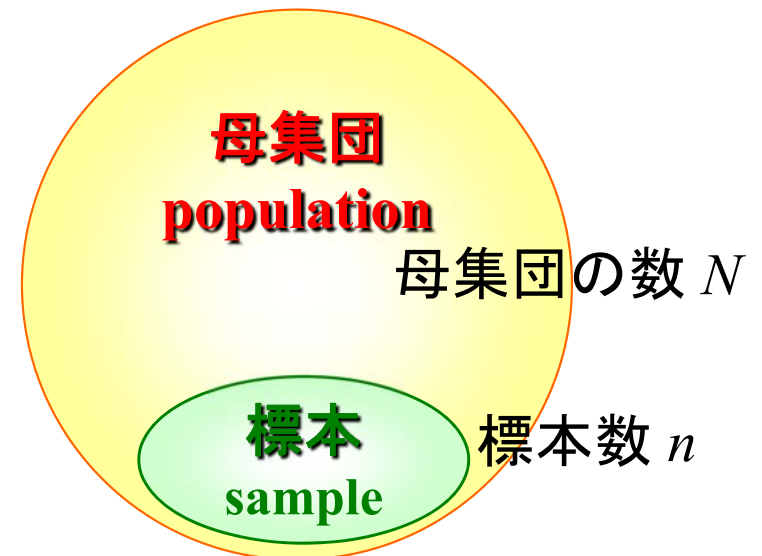
$N$ が余り大きくない場合や、 $n/N$ が大きい場合

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、

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





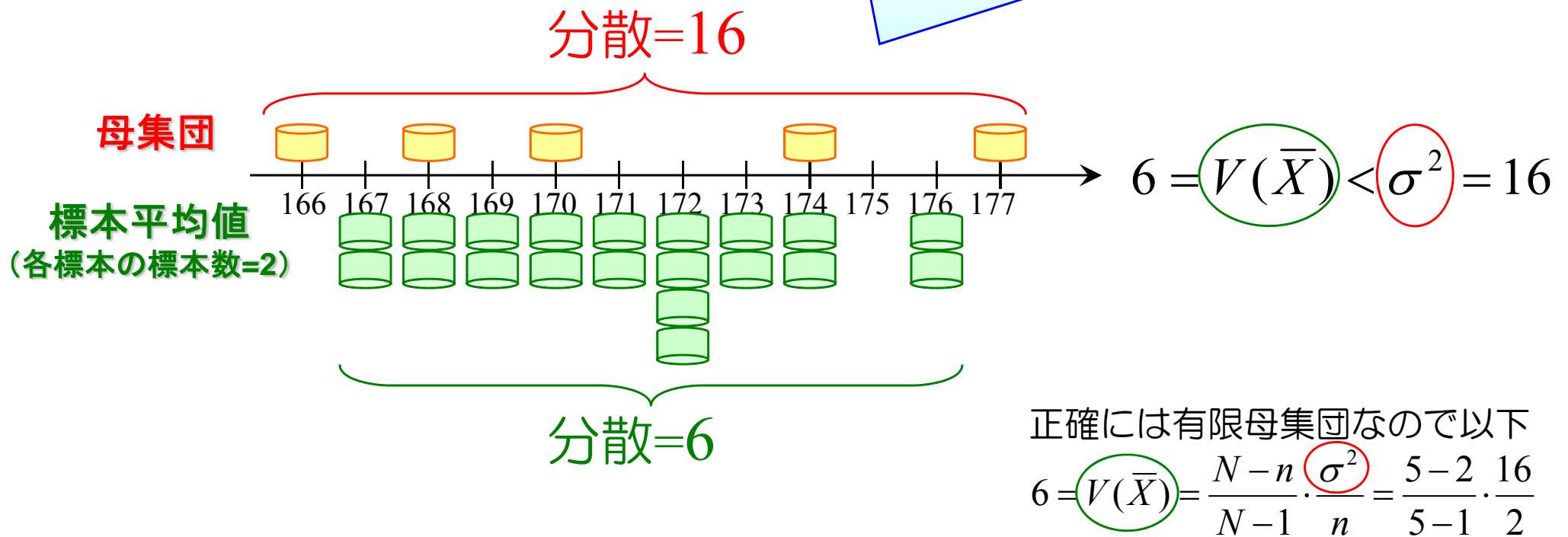
# 補足：母集団と標本：標本平均



🌻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！



# 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

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

- $X_1, \dots, X_n$  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とる値

☀️ 「標本平均  $\bar{X}$  の期待値は母平均  $\mu$  に等しい」  $E(\bar{X}) = \mu$

☀️ 「標本平均  $\bar{X}$  の分散は母分散  $\sigma^2$  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。  $\bar{X} = ?$



3. 1と2の結果から、 $E(\bar{X}) = \mu$  と  $V(\bar{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$  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  $\bar{X}$  の期待値は母平均  $\mu$  に等しい」  $E(\bar{X}) = \mu$
  - 🌸 「標本平均  $\bar{X}$  の分散は母分散  $\sigma^2$  の  $1/n$  に等しい」  $V(\bar{X}) = \frac{\sigma^2}{n}$
- 〔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

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

が母平均  $\mu$  に近い値をとる確率は 1 に近づく。



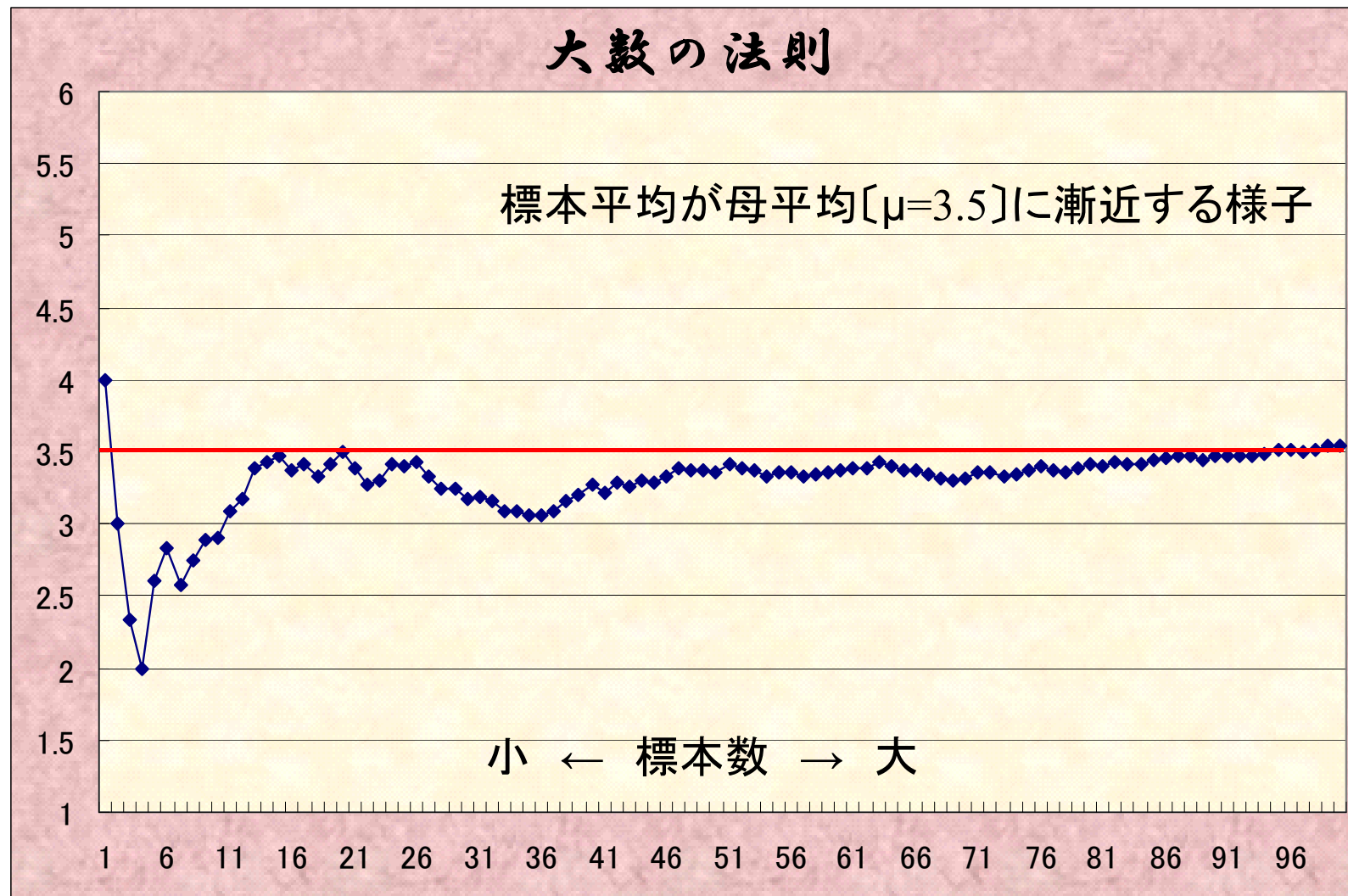
標本数  $n$  が**十分大きければ**、標本は母集団を正しく表すと考えてもよいでしょう。

# 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 [ $\mu=3.5$ ]



# 補足：大数の法則



## 大数の法則

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

☀️証明はチェビシエフの不等式  $P\left(\left|\bar{X} - \mu\right| > k\sigma\right) \leq 1/k^2$  から

∴)  $X_1, \dots, X_n$  は独立で、同じ分布に従う

$$\rightarrow E(X_i) = \mu, V(X_i) = \sigma^2 (i = 1, \dots, n)$$

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

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

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

# 標本分布



🌸 標本平均  $\bar{X}$  が従う確率分布

## 中心極限定理

母平均  $\mu$ , 母分散  $\sigma^2$  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  $\mu$ , 分散  $\sigma^2/n$  の正規分布  $N(\mu, \sigma^2/n)$  に従うとみなすことができる

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

$X_1, \dots, X_n$



# 中心極限定理



**母集団**  
**population**

母平均  $\mu$   
母分散  $\sigma^2$

$n$ 個とってくる

**標本**

**sample**

標本平均  $\bar{X}$   
標本分散  $S^2$

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

...

標本数  $n$  が 十分大きい なら

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

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

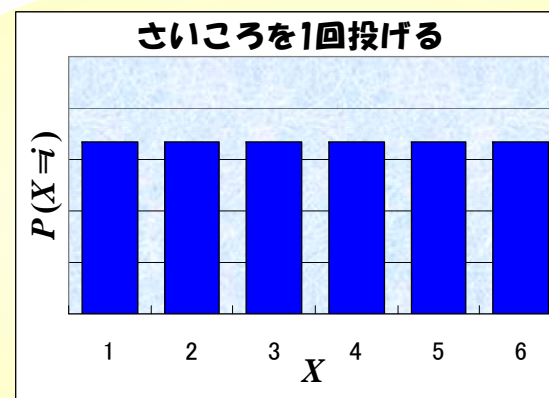


# 中心極限定理



**母集団**  
**population**

母平均  $\mu$   
母分散  $\sigma^2$



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

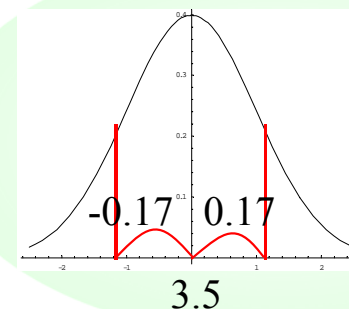
$n$ 個とってくる

**標本**

**sample**

標本平均  $\bar{X}$   
標本分散  $S^2$

標本が十分大きいならば  
サイコロを100回投げる



$$\begin{aligned} \bar{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ &= N\left(\frac{7}{2}, \frac{\frac{35}{12}}{100}\right) \end{aligned}$$

# 補足：中心極限定理



## 中心極限定理

$n \rightarrow \infty$  のとき,

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

が成り立つ。言い換えると,

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

としてよいということ。

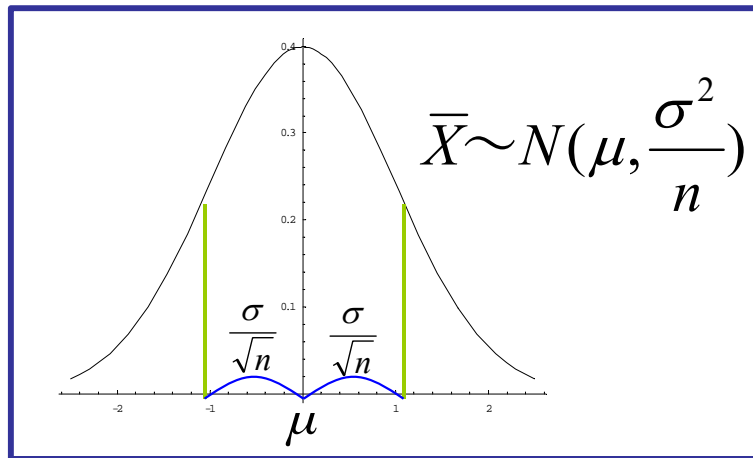
(右辺の $\phi$ は標準正規分布の累積分布関数)

# 標本分布：標本平均の標準化

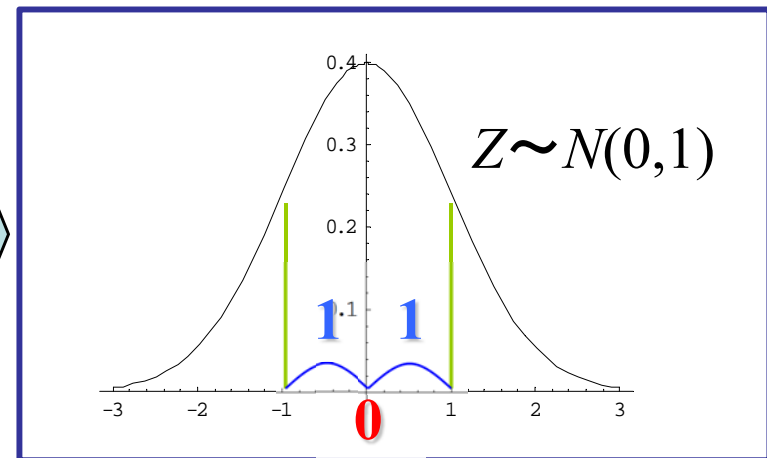


平均 $\mu$ , 分散 $\sigma^2/n$ の標本平均  $\bar{X}$  (確率変数) の標準化

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



標準化  
が  
2つの世界  
の  
架け橋



標本から母平均 $\mu$ 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

# 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： $X=X_1+X_2+\dots+X_n$

二項分布  $Bi(40000, 1/2)$  に従う

$$f(x) = {}_n C_x p^x (1-p)^{n-x} \quad (x=0,1,\dots,n)$$

$$E(X) = np, V(X) = np(1-p)$$

つまり  $P(X > 20400) + P(X < 19600)$  はいくつか？

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

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！

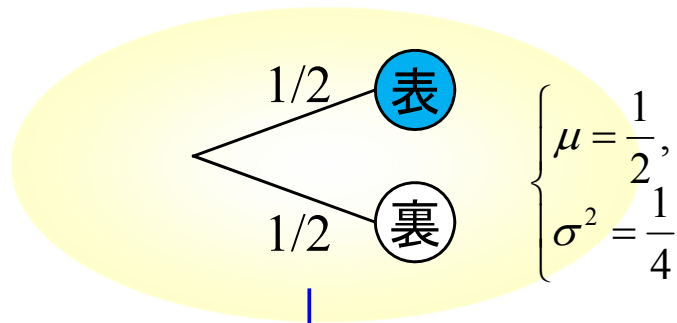
# 中心極限定理の利用



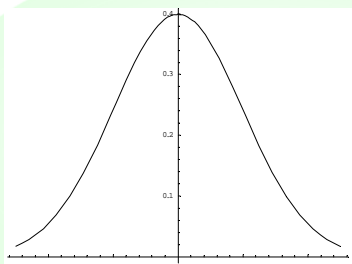
☀ 中心極限定理  $\Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

☀ 標準化  $\Rightarrow Z := \frac{X - \mu}{\sigma} \sim N(0,1) \Rightarrow Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

☀  $X_i \sim Bi(1, 1/2) \Rightarrow \begin{cases} \mu = E(X_i) = n_i p_i = 1 \times 1/2 = 1/2, \\ \sigma^2 = V(X_i) = n_i p_i (1 - p_i) = 1 \times 1/2 \times 1/2 = 1/4 \end{cases}$



↓ 標本  
 $n=40000$ 回



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

表が19600~20400回出る確率を求めたいので、

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

# 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： $X=X_1+X_2+\dots+X_n$

二項分布  $Bi(300, 3/10)$  に従う

$$f(x) = {}_n C_x p^x (1-p)^{n-x} \quad (x=0,1,\dots,n)$$

$$E(X) = np, V(X) = np(1-p)$$



つまり  $P(X \geq 120)$  はいくつか？

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

# 中心極限定理の利用



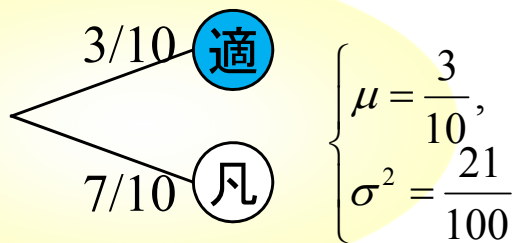
☀ 中心極限定理  $\Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

☀ 標準化  $\Rightarrow Z := \frac{X - \mu}{\sigma} \sim N(0,1) \Rightarrow Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

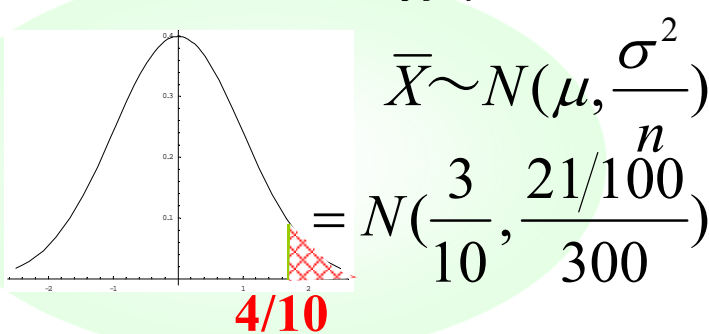
☀  $X_i \sim Bi(1, 3/10) \Rightarrow \begin{cases} \mu = E(X_i) = n_i p_i = 1 \times 3/10 = 3/10, \\ \sigma^2 = V(X_i) = n_i p_i (1 - p_i) = 1 \times 3/10 \times 7/10 = 21/100 \end{cases}$

打率4割以上の確率を求めたいので、

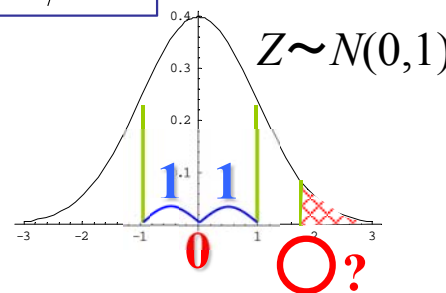
$$\begin{aligned} & P(\bar{X} \geq 4/10) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{4/10 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \geq \frac{\frac{4}{10} - \frac{3}{10}}{\sqrt{\frac{21}{100}} / \sqrt{300}}\right) \\ &= P(Z \geq 3.7796\dots) \\ &= 0.00007853\dots \end{aligned}$$



↓ 標本  
n=300打席



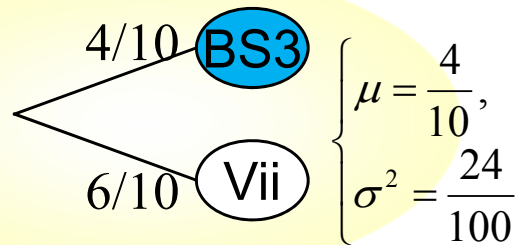
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$



# 中心極限定理の利用

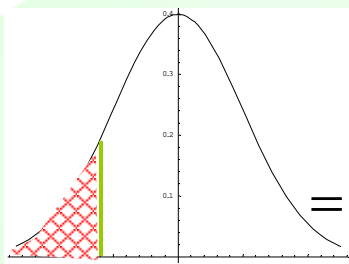


- 例題3：2種類のゲーム機，ソニーのBlainStation3と任天堂のViiの市場シェアはBS3が40%，Viiが60%である．ある店で，どちらかを買いに来た200人の客がいるとき，Viiが110台以上売れる確率は？  
 (ただし，両方買う客はいないとする)



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

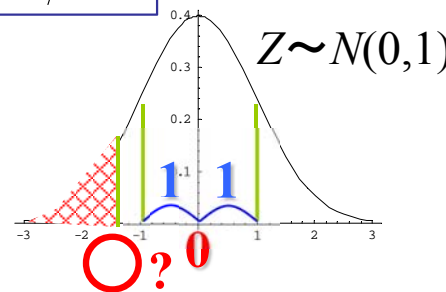
標本  
 $n=200$ 人



9/20

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\frac{4}{10}, \frac{24/100}{200}\right)$$

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



『Viiが110台以上売れる  
 =BS3が90台以上売れない』だから、

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

∴ 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



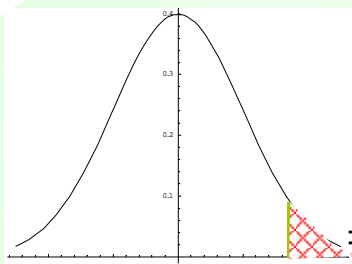
【問題】小学生の1ヶ月の小遣いが、平均2250円、標準偏差360円です。このとき、ランダムに選んだ36人の小学生の小遣い平均が2400円を超える確率は？

## 母集団

母平均  $\mu=2250$ 円

母分散  $\sigma^2=360^2$

↓ 標本  
 $n=36$ 人

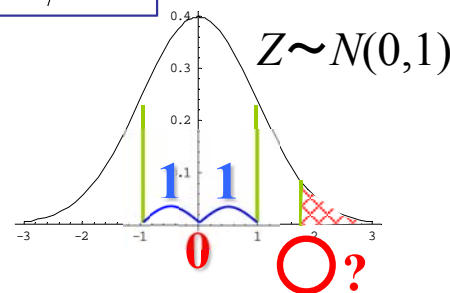


2400

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

$$= N\left(2250, \frac{360^2}{36}\right)$$

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



$$Z \sim N(0,1)$$

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

∴ 答え 0.62%

# 例題：



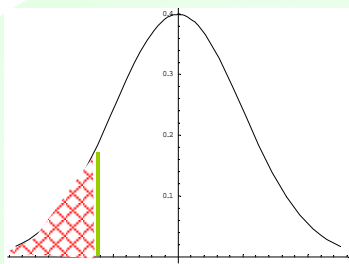
【問題】全国男子大学生の身長が、平均170cm、標準偏差5cmとします。このとき、ランダムに選んだ50人の大学生の平均身長が169cmを下回る確率は？

## 母集団

母平均  $\mu=170\text{cm}$

母分散  $\sigma^2=5^2$

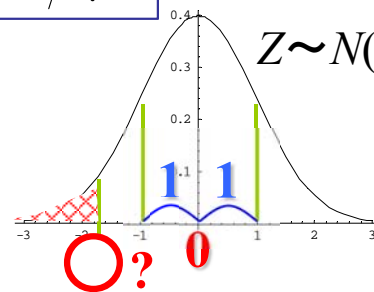
↓ 標本  
 $n=50$ 人



169

$$\begin{aligned}\bar{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ &= N\left(170, \frac{5^2}{50}\right)\end{aligned}$$

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



$Z \sim N(0,1)$

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

∴ 答え 7%

# Coffee Break!



## 100<sup>10</sup>と10<sup>100</sup>はどっちが大きい?

### ☀️ どちらが大きい?

☀️ 100<sup>10</sup> = ? 100<sup>10</sup> = 100,000,000,000,000,000,00 (0が20個)

☀️ 10<sup>100</sup> = ? 10<sup>100</sup> = 100,000,000,000,000,000,000,.....,000 (0が100個)

### ☀️ どちらが大きい?

☀️ 10<sup>100</sup> = ? 10<sup>100</sup> = 100,...,000 (0が100個)

☀️ 100! = ? 100! ≈ 900,...,000 (0が157個)

累乗の増え方は  
凄いな  
階乗の増え方は  
もっと凄いな!

### ☀️ スターリングの公式

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

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

充分大きなNについて、Nの階乗の近似値を与える

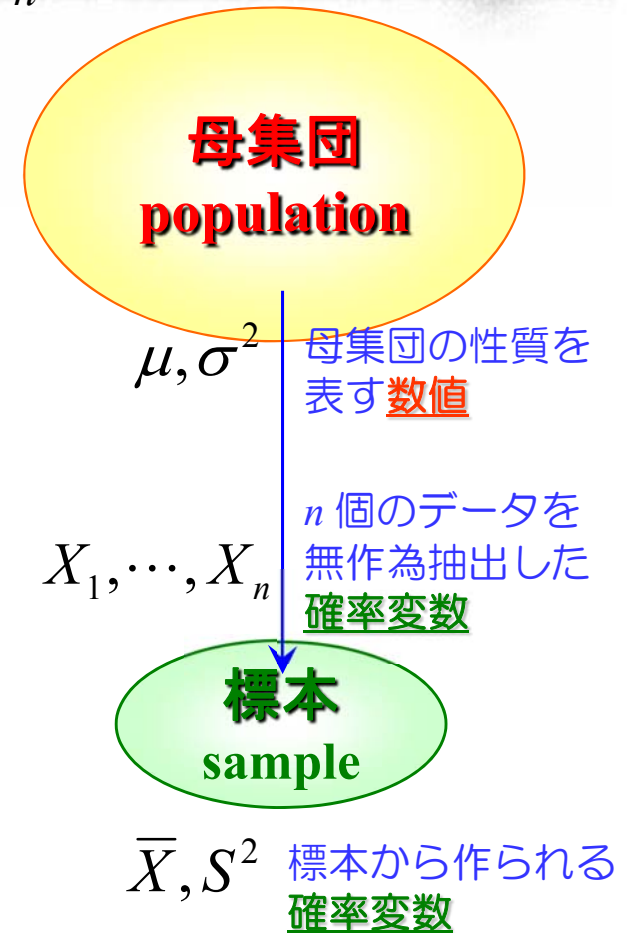


# 標本分布：標本分散



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

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



注意) 「標本分散値」は確率変数 「標本分散」が標本毎に実際にとる値

# 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



母集団数  $N=5$   
母平均  $\mu=171.0$   
母分散  $\sigma^2=16.0$

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)

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

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



# 補足：標本分散の平均と母分散の関係（証明）



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

# 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

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

- $X_1, \dots, X_n$  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：

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

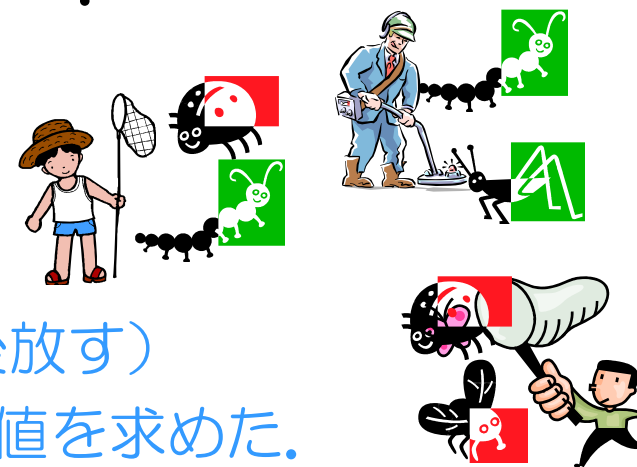


# 演習2：標本分散



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 $S^2$

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

## 🌻 不偏分散 $s^2$

この標本分散は、母分散 $\sigma^2$ の不偏推定量

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

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

有限母集団の場合：

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

$N$ 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？

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

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

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

$\chi^2$ 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  $\chi^2(n-1)$  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？

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

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

**母集団**

母平均  $\mu$

母分散  $\sigma^2$

↓ 標本

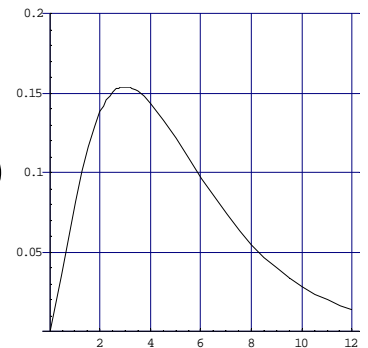
$n$

**標本**

標本平均  $\bar{X}$

標本分散  $S^2$

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



# $\chi^2$ 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  $Z_1, \dots, Z_n$  を考える

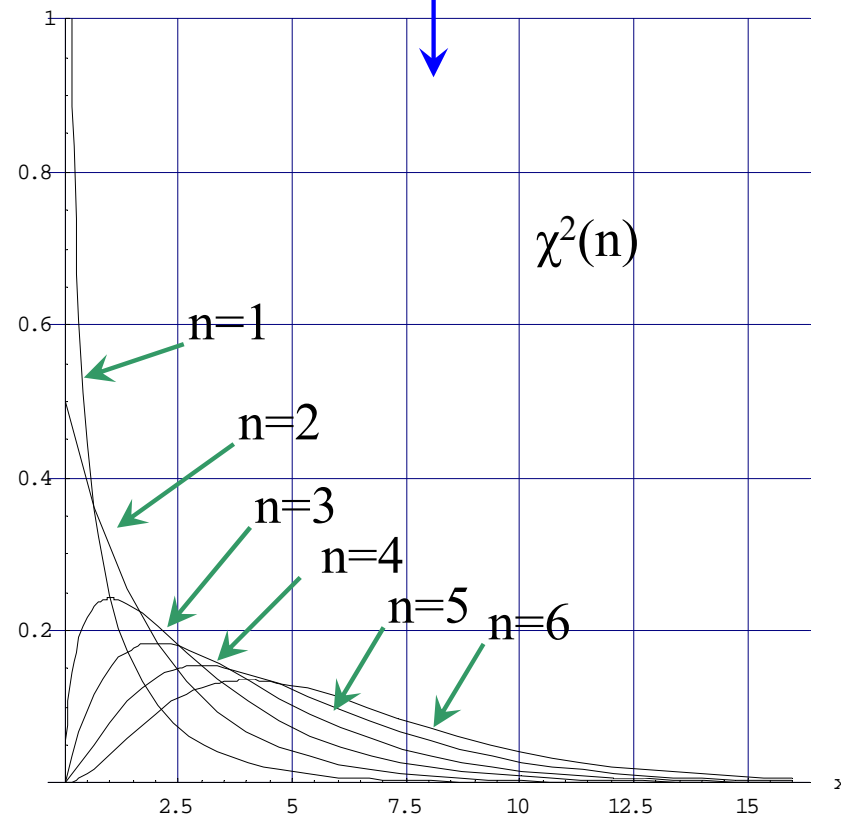
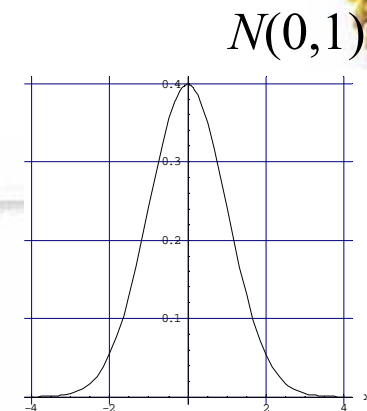
$$\chi^2 = Z_1^2 + \dots + Z_n^2 \quad \leftarrow \text{二乗和をとる}$$

新たな確率変数

この確率変数  $\chi^2$  は，自由度  $n$  の  $\chi^2$  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  $\sigma^2$  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 $\mu=50\text{cm}$ ,分散 $\sigma^2=25$ だとして、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？

$$P(S^2 > 50)$$

$$= P\left(\frac{\chi^2 \sigma^2}{n} > 50\right) \left[ \because \chi^2 = \frac{nS^2}{\sigma^2} \right]$$

$$= P\left(\chi^2 > 50 \frac{n}{\sigma^2}\right)$$

$$= P\left(\chi^2 > 50 \frac{10}{25} = 20\right) \in (0.025, 0.010)$$

$$= 0.017912$$

(Excel関数 CHIDISTより)

**母集団**

母平均  $\mu=50\text{cm}$

母分散  $\sigma^2=25$

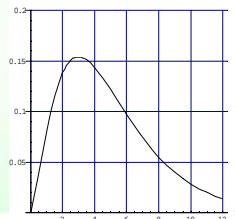
↓ 標本  
 $n=10$ 本

**標本**

標本平均  $\bar{X}$

標本分散  $S^2$

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



自由度9の $\chi^2$ 分布表から

$$P(\chi^2(9) > 19.0228) = 0.025$$

$$P(\chi^2(9) > 21.6660) = 0.010$$

# 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

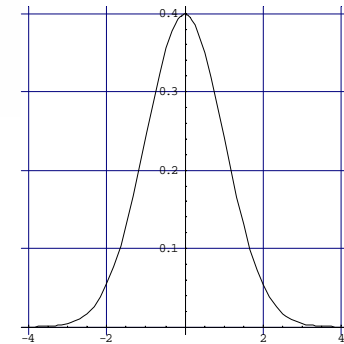
$Y$  : 自由度  $n$  の  $\chi^2$  分布  $\chi^2(n)$  に従う

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

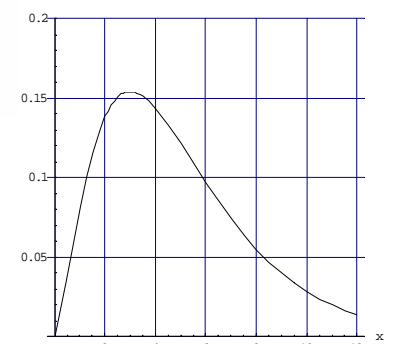
新たな確率変数

確率変数  $T$  は, 自由度  $n$  の  $t$  分布に従う!

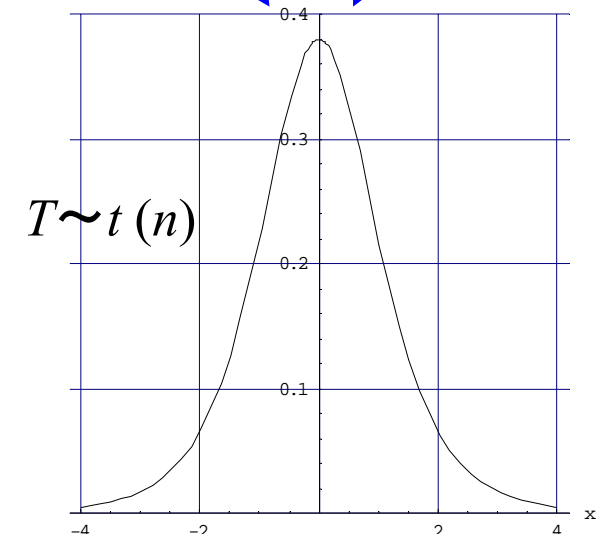
Student の  $t$  分布  
ゴセット (1876-1937)



$X \sim N(0,1)$



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



$T \sim t(n)$

ビール会社ギネスGuinness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

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

標準正規分布  
 $N(0, 1)$  に従う

標本分散  $S^2$  に  $n/\sigma^2$  を掛けた  
確率変数

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

自由度  $n-1$  の  
 $\chi^2$  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  $\mu$  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

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

**母集団**

母平均  $\mu$   
母分散  $\sigma^2$

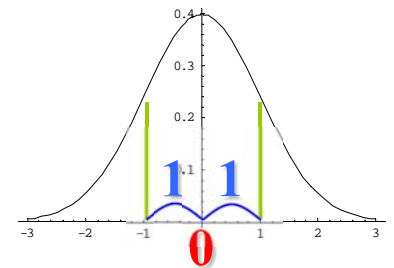
↓ 標本  $n$

**標本**

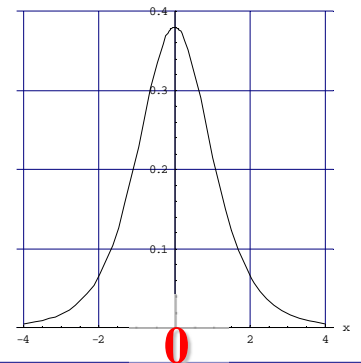
標本平均  $\bar{X}$   
標本分散  $S^2$

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

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



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



# 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

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

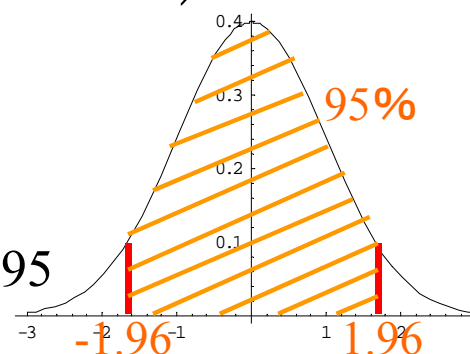
↑ 誤差      ↑ 許容誤差

今、標本平均の従う正規分布から考えて

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

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

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



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 $\varepsilon$ としたとき

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

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

定められた許容誤差 $\varepsilon > 0$ に対し、母集団の大きさ $N$ と母標準偏差 $\sigma$ 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？  $|\bar{X} - \mu| \leq 0.05$

$N$ が十分大きいので、

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

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

$\sigma^2$ 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）