第8回：二大定理と標本分布

尚　晋
大学院経済学研究科　助教

2025年6月3日

今日のポイント

1. 大数の法則と中心極限定理
2. 標本分布

1　大数の法則と中心極限定理

確率論の大定理である「大数の法則」および［中心極限定理」は，統計学の上でも大きな役割を果す．中心極限定理は統計理論のしくみの最小限の理解には，とばしてもよいが，現実の応用では欠かせない．

確率変数 $X_1, X_2, \ldots, X_n$ があるとき，その和 $X_1 + X_2 + \cdots + X_n$，およびその平均 $\frac{X_1+X_2+\cdots+X_n}{n}$ は，統計学の中心的な考え方である．特に，「たくさん観測すれば、平均は真の値に近づく」という性質は，あらゆる統計的推測の基礎になっている．これは確率論の定理である「大数の法則(law of large numbers)」および「中心極限定理(central limit theorem)」によって理解される．

1.1　大数の法則

コイン投げを通じて「真の値への集中」を直感的理解:

- 正しいコインを $n = 10$ 回投げるとする．各試行は「表」「裏」のどちらかが起きる2値試行(ベルヌーイ試行)，表が出たら 1，裏が出たら 0 として，i 回目のコイン投げ結果を確率変数 X_i とすると，表が出た回数 r は以下で与えられる：

$$r = x_1 + x_2 + \cdots + x_n \tag{1}$$

- 表が出た回数の割合$\hat{p} = r/10$は観測された成功率であって，$\hat{p} = 0, 0.1, 0.2, ...$となる．一般に，$n$をコイン投げの回数とするとき，そして，割合$\frac{r}{n}$は成功率(相対頻度)。

- rは確率変数で、$n = 10, p = 0.5$の二項分布$Bi(10, 0.5)$,即ち、$f_{10}(x) = {}_{10}C_x(0.5)^{10}$に従い、$r$の期待値，分散は，

$$E(r) = np = 5, \quad V(r) = np(1-p) = 2.5 \tag{2}$$

- 観測された成功率である割合$\frac{r}{n}$ の期待値と分散は以下になる：

$$E(\frac{r}{n}) = p = 0.5, \quad V\left(\frac{r}{n}\right) = \frac{p(1-p)}{n} = 0.025 \tag{3}$$

ここで，$p = 0.5$は真の成功率.

- 成功の割合が $x/10$ となる確率は$f_{10}(x)$を実際に計算すると：

観測成功率$x/10$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
確率$f_{10}(x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

- 観測された成功率 $\frac{r}{n}$ が $0.4 \leq \frac{r}{n} \leq 0.6$ となる確率は約 65% 程度.

ここで，コイン投げの回数を $n = 10$ から増やして，r/n の期待値 $E(r/n) = p$, およびその周辺が発生する確率がどのようになっていくか調べてみよう．期待値 $p = 0.5$ の周辺としては，0.5 ± 0.1 の範囲をとり，0.4 から 0.6 までとする．コイン投げの回数は，10 回から 20 回，30 回，40 回，50 回，100 回と増やしていくと，この確率は

$$P(0.4 \leq r/10 \leq 0.6) = \sum_{x=4}^{6} f_{10}(x) = 0.65625$$

$$P(0.4 \leq r/20 \leq 0.6) = \sum_{x=8}^{12} f_{10}(x) = 0.73682$$

$$P(0.4 \leq r/30 \leq 0.6) = \sum_{x=12}^{18} f_{10}(x) = 0.79951$$

$$P(0.4 \leq r/40 \leq 0.6) = \sum_{x=16}^{24} f_{10}(x) = 0.84614$$

$$P(0.4 \leq r/50 \leq 0.6) = \sum_{x=20}^{30} f_{10}(x) = 0.88108$$

$$P(0.4 \leq r/100 \leq 0.6) = \sum_{x=40}^{60} f_{10}(x) = 0.96780$$

このように，n を増やしていくと確率は上り，$n = 100$では，表の観測された成功率 $p = r/n$ が 0.4 から 0.6 までの確率は 96%を越え，ほとんどの値が真の成功率 $p = 0.5$ の周囲に集中する.

実際，上の結果を式の形で表現すると

$$\lim_{n\to\infty} P\left(\left|\frac{r}{n} - 0.5\right| \leq 0.1\right) = 1 \tag{4}$$

ということである．一般に，εがどのように小さい（正の）数であっても

$$\lim_{n\to\infty} P\left(\left|\frac{r}{n} - 0.5\right| \leq \varepsilon\right) = 1 \quad (\forall \varepsilon > 0) \tag{5}$$

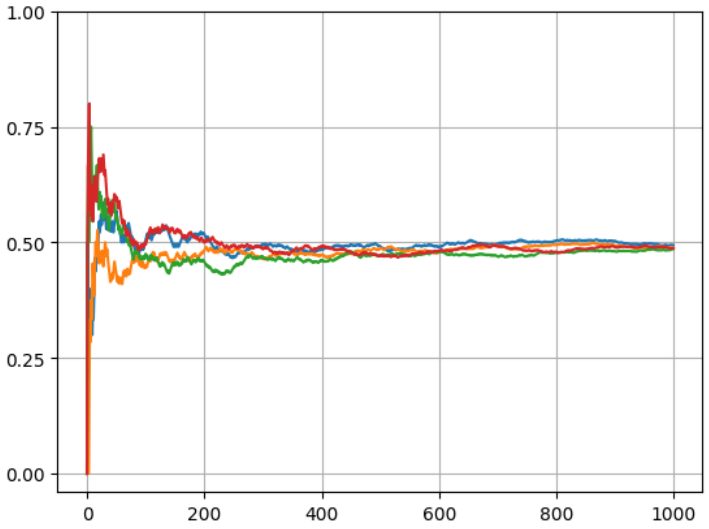
となることが保証されるが[1]，これが大数の法則 law of large numbers と呼ばれるものの一つの形である．標本サイズが無限大になった時、

Nをどんどん増やしていくと、観測値の平均値(観測された表が出る成功率)は期待値0.5に近づいていく、という**大数の法則**をPythonコードを用いてシミュレーション結果(N=1000とし、4回シミュレーション実施)を観察してみよう。どのミュレーションパスもNが大きくなればなうほど、0.5に近づいているのが分かる。(サイコロ投げの場合はどうなるかもやってみてください。)

[1] 式5は$\lim_{n\to\infty} P\left(\left|\frac{r}{n} - 0.5\right| \geq \varepsilon\right) = 0 \quad (\forall \varepsilon > 0)$と書き換えられる

Run 1: ’実行してみてください’

```
#When copy-pasting, please ensure the indent is correct.

#cell 1

import numpy as np
import matplotlib.pyplot as plt

calc_times = 1000
sample_arrary = np.array([0,1])
number = np.arange(1, calc_times + 1)

#generate 4 path
for i in range(4):
    p = np.random.choice(sample_arrary, calc_times).cumsum()
    plt.plot(p / number)

plt.grid(True)
plt.yticks([0, 0.25, 0.5, 0.75, 1.0])
plt.show()
```



統計学上の意義：

(i) 大数の法則は，十分な大きさの標本を調べれば，もとの集団（後に「母集団」といわれる）の様々な特性をかなり正確に知ることができるという認識につながり，統計的推測の理論を生み出すことになった.

(ii) 大数の法則は，一般的に，**大標本では，観察された標本平均を母集団の真の平均（母平均）とみなしてよい**という常識を提供した。(数学的に厳密に証明するにはチェビシェフの不等式[a]を用いる.)

[a] いかなる確率変数Xに対しても、$P(|X-\mu| \geq k\sigma) \leq 1/k^2, \quad (k>0)$が成り立つ。ただし、$\mu = E(X), \sigma^2 = V(X)$である

応用例:社会調査法　「ギャラップ調査」などで有名な，部分（無作為標本）から社会全体の意見（世論）を調べる原理も，大数の法則によっている.

たとえば，「xxx政策に賛成ですか，反対ですか」という設問に対する意見. $X_1, X_2, ..., X_{100}$を，賛否$(1,0)$ の比率が 70 対 30 であるような集団からランダムにとった 100人の回答としよう. これらは独立で，すべてのiについて同一の確率分布 $P(X_i=1)=0.7, P(X_i=0)=0.3$に従う. $X_1+X_2+...+X_{100}$は 100 人の中の 1 (賛成）の人数であり，$\bar{X}$はその相対頻度である. X_iは二項分布 $Bi(1,0.7)$ に従う.計算すれば、$E(X_i)=p=0.7, V(X_i)=p(1-p)=0.21$,「独

立確率変数の和」の内容よって、平均値$\bar{X}$は

$$E(\bar{X}) = 0.7, \quad V(\bar{X}) = 0.0021, \quad D(\bar{X}) = 0.045$$

つまり，$\bar{X}$の範囲は $70 \pm 4.5\%$となる．nが大きいほど$\bar{X}$はもとの比率pの近くに分布するから，この原理を逆に利用して，全体母集団での比率pを部分（標本）の$\bar{X}$からある程度正確に見積ることができる．

1.2　中心極限定理

確率論の大定理である大数の法則は，統計学に応用されるものとしては，標本の大きさnが十分大ならば，標本平均$\bar{X} = (X_1 + ... + X_n)/n$の確率分布は母集団確率分布の平均（母平均）$\mu$の近くに集中していることを保証している．

サイコロ投げを通じて「和の正規性」を直感的理解:

- 次に紹介する**中心極限定理 central limit theorem**は，大数の法則よりくわしい大定理であり，ごく大まかにいえば，**母集団分布が何であっても，nが大なるときには，和$X_1 + X_2 + ... + X_n$の確率分布の形は，大略正規分布と考えてよい**ということである．

- 図8.7~図8.10で見るように，母集団分布の平均，分散（母平均，母分散）をμ, σ^2とすると，母集団分布が何であっても，標本の大きさnが大なるときは，大略

$$\begin{aligned} S_n &= X_1 + X_2 + ... + X_n \quad \text{は} \quad N(n\mu, n\sigma^2) \quad \text{に}, \\ \bar{X} &= (X_1 + X_2 + ... + X_n)/n \quad \text{は} \quad N(\mu, \sigma^2/n) \quad \text{に}, \end{aligned} \tag{6}$$

  従うと考えてよい．

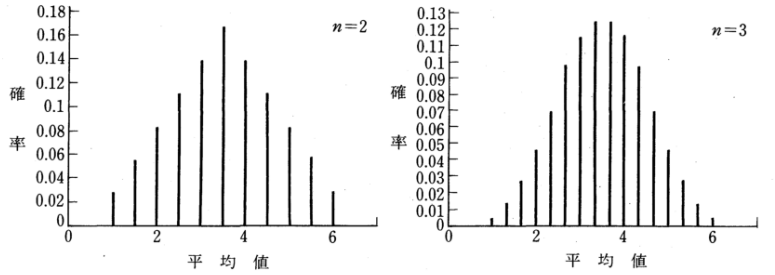


図8.7

図8.8

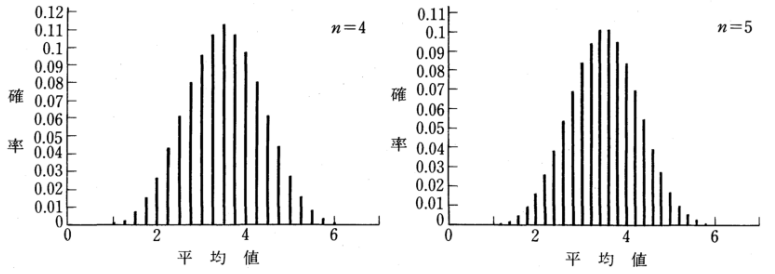


図8.9

図8.10

n 個のさいころの目の平均値の確率分布(図 8.7-図 8.10)
さいころのように日常親しんでいるものからも，中心極限定理を使うと正規分布が生じる．

- 中心極限定理を一応厳密に表すと，$n \to \infty$の時：

$$P\left(a \leq \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leq b\right) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{7}$$

が成り立つということである．いいかえればnが大きければ，

$$P\left(a \leq \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leq b\right) \fallingdotseq \Phi(b) - \Phi(a) \tag{8}$$

としてよい．ここでのΦは第六回のところで見た標準正規分布の累積分布関数である．
左辺は標準化変数の形で表すと：

$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) \fallingdotseq \Phi(b) - \Phi(a) \tag{9}$$

さらに、一般的に、標本平均$\bar{X}$を標準化したものの確率分布は、$n \to \infty$の時、標準正規分布$N(0,1)$に分布収束:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1), \quad (n \to \infty) \tag{10}$$

で表す。記号$\xrightarrow{d}$は左側の確率変数の分布は右側の確率分布に分布収束という。収束先の分布は漸近分布という。

- さいころの目$(1 \sim 6)$の出方の確率分布は離散型の一様分布であり，$\mu = 7/2, \sigma^2 = 35/12$である。この確率分布は，正規分布とは相当に違っている．
- この母集団からランダムに$n = 2$の標本X_1, X_2を取りだしたときの標本平均$(X_1+X_2)/2$とは，つまり，さいころを2回振ったときに出る目の平均値である．さいころを2回振ったときの目の出方は$6X6 = 36$通りあるが，平均値にはいくつか同じものが出てくるので，整理すると次のようになる．

平均値	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
確率	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

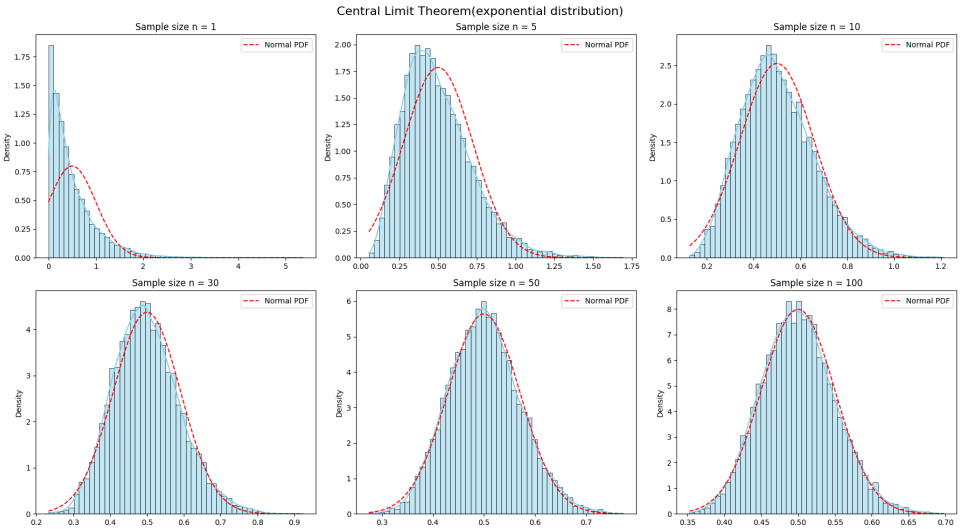
これをグラフにしたものが，図8.7である.1回のときは峰（モード）がないが，2回のときは，それができており，中心極限定理の様子が既にあらわれている．中心極限定理はnを大きくすれば，和や標本平均の分布は正規分布に近づいていくというものだった． $n = 3, 4, 5 \cdots$ と振る回数はnを増やしていくと、出目の和や平均の分布は明確に釣鐘型（正規分布型）に近づいていくことがわかる．これは中心極限定理の述べている内容である．

Nが増えれば増えるほど、母集団は指数分布に従う場合、標本平均が正規分布の形になっていく法則を下記のPythonコードを用いてを見てみよう。

Run 2: ’実行してみてください’

```
#When copy-pasting, please ensure the indent is correct.
#cell 1

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm

np.random.seed(0)
sample_sizes = [1, 5, 10, 30, 50, 100]
num_trials = 10000  # trials numbers
lambda_param = 2.0  # the parameter λ of exponential distribution (mean: 1/λ)

fig, axes = plt.subplots(2, 3, figsize=(18, 10))
axes = axes.flatten() #Converts a 2D matrix to a 1D list.

for idx, n in enumerate(sample_sizes):
    means = []
```

```
    for _ in range(num_trials):
        sample = np.random.exponential(scale=1 / lambda_param, size=n)
        means.append(np.mean(sample))

    sns.histplot(means, bins=50, kde=True, stat='density', ax=axes[idx],
        color='skyblue', edgecolor='black')

    # comparison with normal distribution using the same mean
    mu = 1 / lambda_param
    sigma = mu / np.sqrt(n)
    x = np.linspace(min(means), max(means), 1000)
    y = norm.pdf(x, mu, sigma)
    axes[idx].plot(x, y, color='red', linestyle='--', label='Normal PDF')

    axes[idx].set_title(f'Sample size n = {n}')
    axes[idx].legend()

plt.suptitle('Central Limit Theorem(exponential distribution)', fontsize=16)
plt.tight_layout()
plt.show()
```



中心極限定理の応用例1:

- 例1：40,000 回コインを投げて 20,400 回以上，あるいは 19,600 回以下，表$(X_i = 1)$が出ることは，どの程度の確率であろうか.
- 各回の表，裏を$X_i = 1, 0$として，n回中の1の総回数$X = X_1 + X_2 + ... + X_n$の確率分布を求めればよい．$n = 40000$で，$X$は二項分布$Bi(40000, 1/2)$に従うが，$\Sigma_{19600}^{20400}{}_{40000}C_x(1/2)^x(1/2)^{40000-x}$を計算することは不可能である．そこで，中心極限定理を使う.
- X_iは二項分布$Bi(1, 1/2)$に従うから，$\mu = E(X_i) = 1/2, \sigma^2 = V(X_i) = p(1-p) = 1/4, n\mu = 20000, \sqrt{n}\sigma = 100$、従って：

$$
\begin{aligned}
&P(19600 \leq X_1 + X_2 + ... + X_{40000} \leq 20400) \\
&\quad = P(-4 \leq (X_1 + X_2 + ... + X_{40000} - 20000)/100 \leq 4) \\
&\quad = \Phi(4) - \Phi(-4) \qquad (\Phi\text{は標準正規分布の累積分布関数}) \\
&\quad = 0.9999
\end{aligned}
\tag{11}
$$

よって，求める確率$(P(19600 \leq X_1 + X_2 + ... + X_40000 \leq 20400))$は $1/10000$ 程度であり，事実上ありえないこととなる.

2　標本分布

2.1　統計的推測：母集団・母数と標本

統計的推測　全体（母集団）から一部（標本）を抽出・分析し、全体の特徴（母数）を推測する手法を、**統計的推測**と呼ぶ。統計的推測の構成要素：**母集団・母数と標本**

母集団　分析者が興味のある対象「全体」を**母集団**。有限個の個体から成る母集団を**有限母集団**という．無限個の個体から成る母集団を**無限母集団**という．母集団の例：「サラリーマン全体」の年収平均、「自社製の電球全体」について、100時間以内に切れる確率。

■しかし、母集団全体を把握するのは、ほぼ不可能！例：日本中のサラリーマン全員を調査する、電球の寿命を全部調べるのはムリ。しかし「一部」を観測・分析し、「全体」の特徴を推測するのは？

標本　母数の推測のために母集団から抽出（サンプリング）したデータ，$X_1, X_2, ..., X_n$を**標本**と呼び，代表してX_iと表記（nは標本の大きさ、sample size)。例：サラリーマン$n = 1000$人に年収をアンケート調査で、標本平均を求める。電球$n = 500$個を一斉に点灯する実験で、100時間以内に切れた割合を求める。

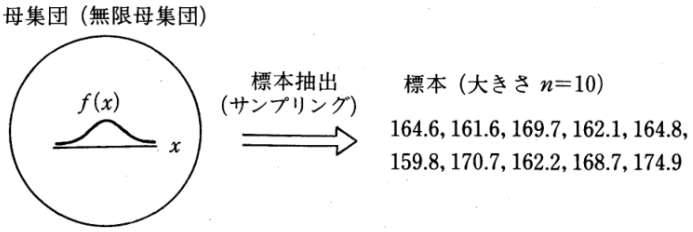


図 9.2　日本人(の身長)という母集団と標本
「日本人」は有限母集団だが，無限母集団と考えて差支えない．母集団分布$f(x)$は正規分布$\mathrm{N}(\mu, \sigma^2)$を考えるのがふつうである．標本はこの確率分布に従う確率変数であるが，現実に扱うのはその実際の値(実現値)，観測値である．

標本抽出　(i) 母集団から標本を取り出すことを**標本抽出**という．

(ii) 取り出した個体を母集団に戻しながら繰り返す抽出を**復元抽出**sampling with replacementという．

(iii) 取り出した個体を母集団に戻さずに繰り返す抽出を**非復元抽出**sampling without replacementという．

(iv) 母集団の各要素が標本に含まれる確率(抽出率)が等しく[2]、即ち、等確率で取り出される抽出を**（単純）無作為抽出**という．

(v) 無作為抽出した標本を**無作為標本**という．

(vi) 復元抽出した無作為標本の各個体を確率変数で表すと，それらは**独立かつ同一に（independent and identically distributed, iid）**母集団分布にしたがう．

母集団分布と母数　(i) 母集団の確率的な特性を表す確率分布を**母集団分布**という。

(ii) 母集団分布の特性を表す定数を**母数**（パラメーター）という．

(iii) 母集団分布の平均を**母平均**という．

(iv) 母集団分布の分散を**母分散**という．

(v) 母集団分布がある知られた確率分布であることが，理論的・経験的にわかっている場合で，有限個の母数で表せる分布を**パラメトリックな分布**という．

[2]母集団に含まれる要素の数（母集団の大きさ）をN,標本として取り出す要素の数（標本の大きさ）をnとし、抽出率が等しくn/Nとするもの.

(vi) 母集団分布の具体的な形が，事前に知られていない場合で，有限個の母数で表せない分布を**ノンパラメトリックな分布**という．たとえば，世界各国の面積や人口の分布などがこの例である．

2.2 統計量と標本分布

統計量 一般に、母集団の母数の推測に使われる、標本から要約したものを**統計量**という。標本平均と標本分散は最も重要な二つの統計量である。

標本分布 確率的な標本抽出にともなう統計量の確率分布をその統計量の**標本分布**という．

■母集団の一部でしかない標本から，母集団の属性についての推測を行うのであるから，標本の分析結果はどのような標本を抽出するかに依存する。標本によるばらつきに対応するためには，確率的な扱いが不可欠であり，統計学では**「標本分布」**という考え方をすることになる．
■母集団分布：個々の標本X_iが従う分布。標本分布:標本を要約した統計量の分布。

標本平均 標本$X_1, X_2, ..., X_n$から計算された平均を**標本平均**という。

標本$X_1, X_2, ..., X_n$は母集団(母平均μ、母分散σ^2)に従う独立な確率変数であるが、標本平均$\bar{X}$は

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} \tag{12}$$

で与えられる．「独立な確率変数の和」のところで述べたように、

$$E(\bar{X}) = E(\frac{X_1 + X_2 + ... + X_n}{n}) = \frac{n\mu}{n} = \mu \tag{13}$$

となって，期待値が母平均μに一致する．この性質は**不偏性**という。

さらに、標本平均$\bar{X}$の分散は

$$V(\bar{X}) = V(\frac{X_1 + X_2 + ... + X_n}{n}) = \frac{1}{n^2}V(X_1 + X_2 + ... + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{14}$$

となって，従って、$n \to \infty$の時$\bar{X}$の分散は0に近づいていき、標本平均$\bar{X}$が母平均μに集中、確率収束[3]していくのである．この性質は**一致性**という。(大数の法則を使っても同じの結論を引き出す．)

標本分散 標本$X_1, X_2, ..., X_n$から計算された分散を**標本分散**という。

(i) 母平均が既知の場合，標本$X_1, X_2, ..., X_n$の標本分散は

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{15}$$

で定義される．$E(\hat{\sigma}^2) = \sigma^2$の性質を満たす(証明は省略).

(ii) 母平均が未知の場合，標本$X_1, X_2, ..., X_n$の標本分散は

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{16}$$

で定義される．$E(s^2) = \sigma^2$の性質を満たす(証明は補足1)．標本分散s^2は，期待値が母分散に一致し，母分散を過大にあるいは過小にではなく不偏に推定する．したがって，上の式のs^2を母分散σ^2の不偏推定量，あるいは**不偏分散** unbiased variance という．

[3] この関係を記号$\xrightarrow{p}$を使って、$\bar{X} \xrightarrow{p} \mu, (n \to \infty)$と表し、$\bar{X}$が$\mu$に確率収束するという

もしここは、'不偏でない'標本分散($S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$)を使うと、$E(S^2) = \frac{n-1}{n}\sigma^2$となり、$n = 10$ならば1割程度の$\sigma^2$の過小評価が起こるので、$S^2$と$s^2$の違いには、要注意。

補足 2.1 不偏分散の期待値の導出：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

であるから、$E(s^2) = \sigma^2$を証明するには、$E(\sum_{i=1}^{n}(X_i - \bar{X})^2) = (n-1)\sigma^2$を証明すればいい。

$$\begin{aligned}
E(\sum_{i=1}^{n}(X_i - \bar{X})^2) &= E(\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2) \\
&= E(\sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\bar{X} - \mu) + n(\bar{X} - \mu)^2) \\
&= E(\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2) \\
&= \sum_{i=1}^{n}E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \\
&= \sum_{i=1}^{n}V(X_i) - nV(\bar{X}) \\
&= n\sigma^2 - n \cdot \frac{\sigma^2}{n} \\
&= (n-1)\sigma^2
\end{aligned}$$

2行目から3行目への変形では$\sum_{i=1}^{n}(X_i - \mu)(\bar{X} - \mu) = n(\bar{X} - \mu)(\bar{X} - \mu) = n(\bar{X} - \mu)^2$を用いた。