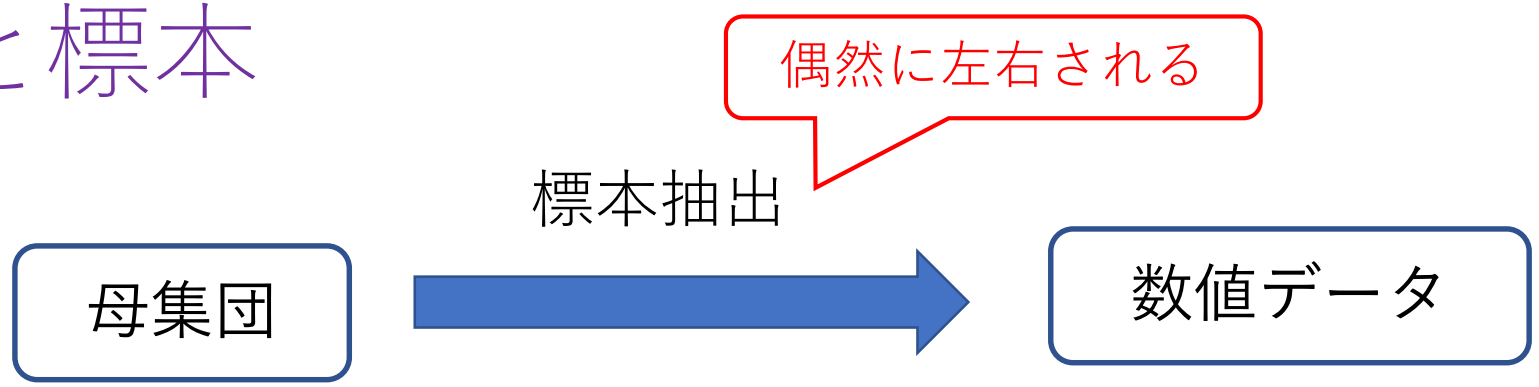
Lecture 6

標本抽出と正規分布

母集団と標本

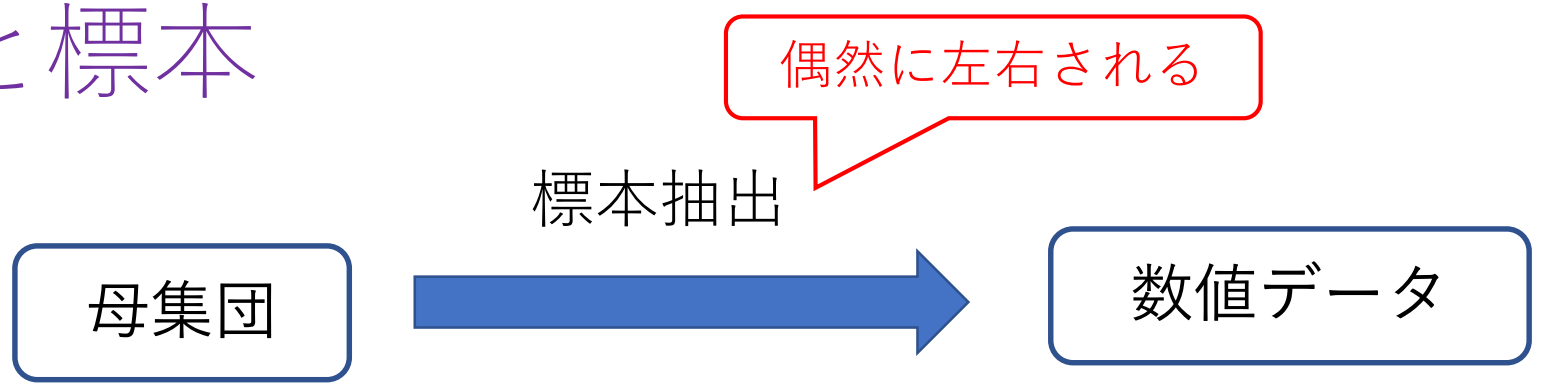


母集団 (population) = 調査対象の全体

実態として存在する集合（世論調査など）
または，理論上想定される（ふつうは無限）集合

課 題　得られた数値データから母集団の性質を信頼度付きで推定する

母集団と標本



標本(sample)の取り出し方の基本：無作為復元抽出

- 母集団のどの要素も偏りなく選ばれる可能性を担保
- 取り出すたびに母集団は変化しない

同一環境下における
繰り返し実験もこれ

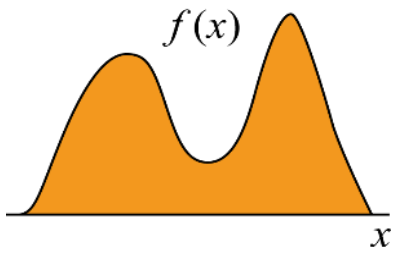
1）母集団から n 個の数値データ $x_1, x_2, \cdots, x_n$ を得る.

2）数値データ x_i は確率変数 X_i の実現値として扱う.

3）無作為復元抽出なので, $X_1, X_2, \cdots, X_n$ はすべて母集団分布に従い,
$X_i \sim$ 母集団分布, かつ独立である.

推測統計の目的

母集団



母集団分布

無作為
復元抽出

標本

$X_1, X_2, \ldots, X_n$

実際に知りたいのは，

母数 θ ＝母集団の統計量

平均値，分散，そのほか

θ を標本 $X_1, X_2, \ldots, X_n$ を用いて

何らかの合理的な計算で求める．

$$\hat{\theta} = T(X_1, X_2, \cdots, X_n)$$

目的　$\hat{\theta}$ の確率分布を調べて，
信頼度付きで θ を推定する．

確率変数の和と積

$\hat{\theta} = T(X_1, X_2, \cdots, X_n)$ のような確率変数の関数を扱う準備

確率変数 X, Y と 定数 a, b に対して

$X + Y$ ：和

XY ：積

aX ：スカラー倍（スカラー積）

$aX + bY$ ：線形和（線形結合）

定理（平均値の線形性）

確率変数 X, Y と 定数 a に対して

（1）$E[aX] = aE[X]$

（2） $E[X + Y] = E[X] + E[Y]$

例（1）X をサイコロの目とすると,

$$E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

Y をサイコロの目を 10 倍して与えられる点数とすると,

$$E[Y] = E[10X] = 10E[X] = 10 \times 3.5 = 35$$

定理　（平均値の線形性）

確率変数 X, Y と 定数 a に対して

（１）$E[aX] = aE[X]$

（２）$E[X + Y] = E[X] + E[Y]$

例　（２）サイコロとコインを同時に投げて，
X をサイコロの目，Y をコインの表(1)裏(0)すると，

$$E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5 \qquad E[Y] = \frac{1}{2}(0 + 1) = 0.5$$

Z をサイコロの目とコインを合わせた点数とすると，

$$E[Z] = E[X + Y] = E[X] + E[Y] = 3.5 + 0.5 = 4$$

平均値の定義

$$E[X] = \sum x_i P(X = x_i)$$

証 明　離散型の場合を扱う（連続型では積分が必要）

（1） X の取りうる値が x_i なら aX の取りうる値は ax_i

$$E[aX] = \sum ax_i P(X = x_i) = a \sum x_i P(X = x_i) = aE[X]$$

（2） X, Y の取りうる値が x_i, y_j なら $X + Y$ の取りうる値は $x_i + y_j$

$$E[X + Y] = \sum (x_i + y_j) P(X = x_i, Y = y_j) = \sum x_i P(X = x_i, Y = y_j) + \sum y_j P(X = x_i, Y = y_j)$$

第 1 項は,

$$\sum_{i,j} x_i P(X = x_i, Y = y_j) = \sum_i x_i \sum_j P(X = x_i, Y = y_j) = \sum_i x_i P(X = x_i) = E[X]$$

第 2 項も同様に変形して,　　$E[X + Y] = E[X] + E[Y]$

定理　確率変数 X, Y と 定数 a に対して,

（１）$V[aX] = a^2 V[X]$

（２）$V[X+Y] = V[X] + V[Y] + 2\mathrm{Cov}(X,Y)$

ただし, $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$ を共分散という.

分散公式

$$V[X] = E[X^2] - E[X]^2$$

証明　（１）

$$\begin{aligned} V[aX] &= E[(aX)^2] - E[aX]^2 \\ &= E[a^2X^2] - (aE[X])^2 \\ &= a^2E[X^2] - a^2E[X]^2 \\ &= a^2(E[X^2] - E[X]^2) \\ &= a^2V[X] \end{aligned}$$

定理　確率変数 X, Y と 定数 a に対して,

（１）$V[aX] = a^2 V[X]$

（２）$V[X+Y] = V[X] + V[Y] + 2\mathrm{Cov}(X,Y)$

ただし, $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$ を共分散という.

分散公式

$$V[X] = E[X^2] - E[X]^2$$

証明　（２）　$V[X+Y] = E[(X+Y)^2] - E[X+Y]^2$

$$= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2$$

$$= E[X^2] + 2E[XY] + E[Y^2] \ -E[X]^2 - 2E[X]E[Y] - E[Y]^2$$

$$= V[X] + V[Y] + 2(E[XY] - E[X]E[Y])$$

$$= V[X] + V[Y] + 2\mathrm{Cov}(X,Y)$$

定理　(独立な確率変数)

確率変数 X, Y が独立であれば,

（1）$E[XY] = E[X]E[Y]$

（2）$V[X+Y] = V[X] + V[Y]$

例　サイコロ2回投げて出た目を X, Y とすると,

$$E[X] = E[Y] = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

Z をサイコロ2回投げて出た目の積とすると,

$$E[Z] = E[XY] = E[X]E[Y] = 3.5 \times 3.5 = 12.25$$

証 明　離散型の場合を扱う（連続型では積分が必要）

（１）X, Y の取りうる値が x_i, y_j なら XY の取りうる値は $x_i y_j$

$$E[XY] = \sum x_i y_j P(X = x_i, Y = y_j)$$

$$= \sum x_i y_j P(X = x_i) P(Y = y_j)$$

X, Y の独立性

$$= \sum x_i P(X = x_i) \sum y_j P(Y = y_j) = E[X]E[Y]$$

（２）　確率変数の和の分散　$V[X + Y] = V[X] + V[Y] + 2\mathrm{Cov}(X, Y)$

共分散　$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$

X, Y が独立であれば，　$E[XY] = E[X]E[Y]$ なので，　$\mathrm{Cov}(X, Y) = 0$

よって，　$V[X + Y] = V[X] + V[Y]$

正規確率変数の線形結合

定理　２つの正規確率変数 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ が独立であるとする.

このとき, 定数 a, b に対して, $aX + bY$ も正規確率変数であって,

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

- $aX + bY$ の分布が正規分布になることを証明するには, 密度関数の 積分を扱う必要がある（参考書等を見よ）
- ここでは, 平均値と分散だけを確認しておく.

$$E[aX + bY] = E[aX] + E[bY] = aE[X] + bE[Y] = a\mu_1 + b\mu_2$$

$$V[aX + bY] = V[aX] + V[bY] = a^2V[X] + b^2V[Y] = a^2\sigma_1^2 + b^2\sigma_2^2$$

標本平均は確率変数である

母集団
母平均 μ
母分散 σ^2

無作為
復元抽出

標本
$X_1, X_2, \ldots, X_n$

$X_1, X_2, \ldots, X_n$ は確率変数になる
（無作為標本ともいう）

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

【重要】　$\bar{X}$ も確率変数である

- 平均値や分散は？
- 確率分布は？

標本平均の平均値と分散

定 理　母平均 μ, 母分散 σ^2 の母集団から取り出した大きさ n の無作為標本

$X_1, X_2, \ldots, X_n$ の標本平均 $\bar{X} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ に対して,

平均値： $E[\bar{X}] = \mu$　　分散： $V[\bar{X}] = \dfrac{\sigma^2}{n}$

証 明

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}E\left[\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$$

$$V[\bar{X}] = V\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}V\left[\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n} V[X_k] = \frac{1}{n^2}\sum_{k=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

標本平均の分布（正規母集団の場合）

正規母集団
$N(\mu, \sigma^2)$

→ 無作為復元抽出 →

標本
$X_1, X_2, \cdots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

定 理

正規母集団 $N(\mu, \sigma^2)$ から取り出した n 個の無作為標本の標本平均 $\bar{X}$ の分布は,

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

標本平均の分布（正規母集団の場合）

正規母集団
$N(\mu, \sigma^2)$

無作為復元抽出

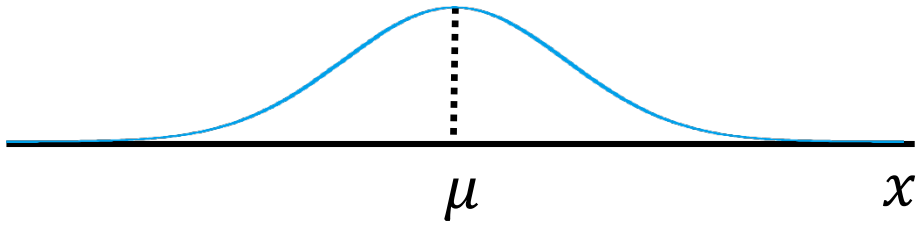
標本
$X_1, X_2, \ldots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

母集団分布

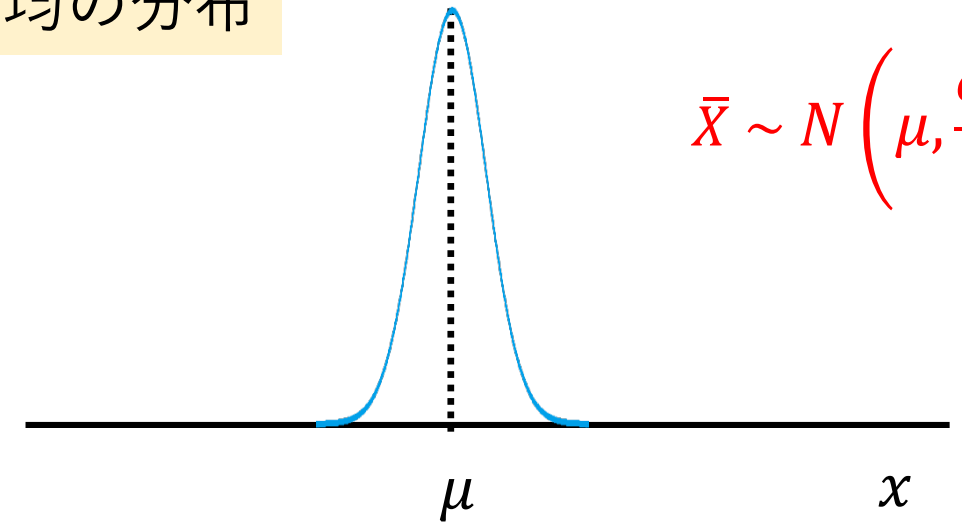
$$X_k \sim N(\mu, \sigma^2)$$



分散 σ^2

標本平均の分布

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



分散 σ^2/n

標本平均の分布（一般の母集団の場合）

一般の母集団
μ, σ^2

無作為復元抽出

標本
$X_1, X_2, \ldots, X_n$

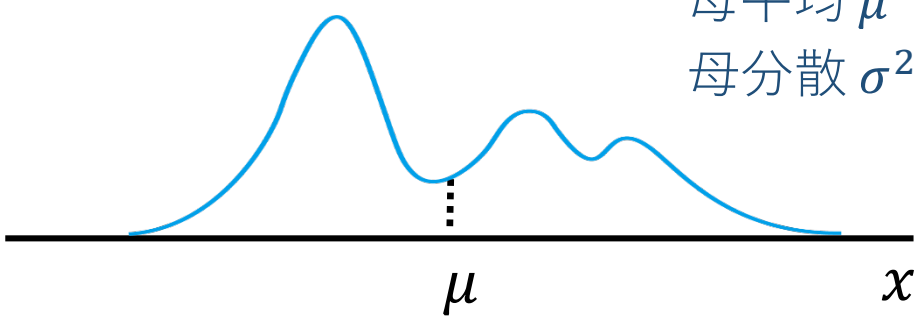
標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

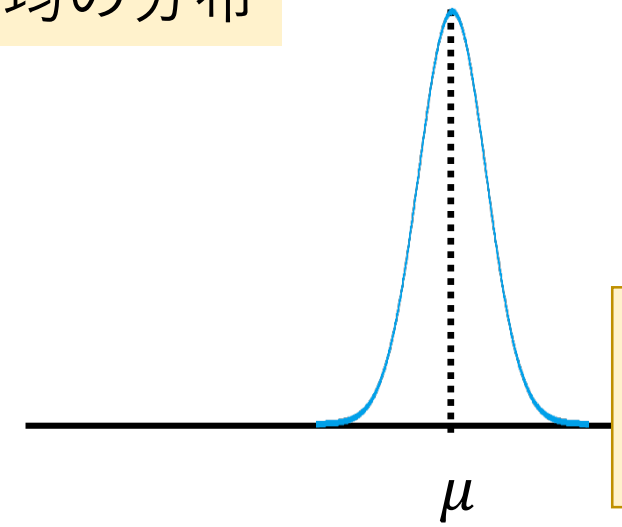
母集団分布

$X_k \sim$ 母集団分布

母平均 μ
母分散 σ^2



標本平均の分布



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

大きな n で近似的に成り立つ

中心極限定理（CLT）

定理　母平均 μ, 母分散 σ^2 のいっぱんの母集団から取り出した大きさ n の

無作為標本 $X_1, X_2, \ldots, X_n$ の標本平均　$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$　に対して,

n が大きいときは近似的に,

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad n \to \infty$$

- 二項母集団のとき：ベルヌーイ, ド・モアブル, …
- 一般の母集団：ラプラス（高度な微積分）

分布の特性関数

≈フーリエ変換 ≈ ラプラス変換 ≈ 連続分布に対する母関数

$$\varphi_X(t) = \int_{-\infty}^{+\infty} f_X(x) e^{itx}\, dx = E\left[e^{itX}\right]$$

- 特性関数は分布を一意的に決める
- 標準正規分布 $N(0,1)$ の特性関数

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} e^{itx}\, dx = e^{-t^2/2}$$

証 明

標本平均 $\bar{X}$ の標準化 Z

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Z_k$$

$$E[Z_k] = 0 \qquad V[Z_k] = E\left[Z_k^2\right] = 1$$

標準化 Z の特性関数

$$\varphi_Z(t) = E\left[e^{itZ}\right]$$

$$= E\left[\prod_{k=1}^{n} e^{it\frac{Z_k}{\sqrt{n}}}\right]$$

$$= \prod_{k=1}^{n} E\left[e^{it\frac{Z_k}{\sqrt{n}}}\right]$$

$$E\left[e^{it\frac{Z_k}{\sqrt{n}}}\right] = E\left[1 + it\frac{Z_k}{\sqrt{n}} + \frac{1}{2}\left(it\frac{Z_k}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right]$$

$$= 1 + 0 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

$$\varphi_Z(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-t^2/2} \quad (n \to \infty)$$

これは標準正規分布 $N(0,1)$ の特性関数

Z の分布 $\to N(0,1)$　　$(n \to \infty)$

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

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

例題 6.1　サイコロを 50 回振るとき, 出目の平均値 $\bar{X}$ はどのような分布に従うと考えられるか.

例題 6.1　サイコロを 50 回振るとき, 出目の平均値 $\bar{X}$ はどのような分布に従うと考えられるか.

母集団
母平均 μ
母分散 σ^2

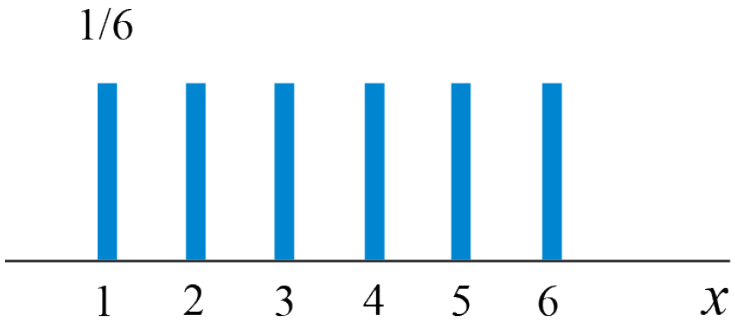
→ 無作為復元抽出 →

標本
$X_1, X_2, \ldots, X_n$

標本平均

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$X_k \sim$ 母集団分布



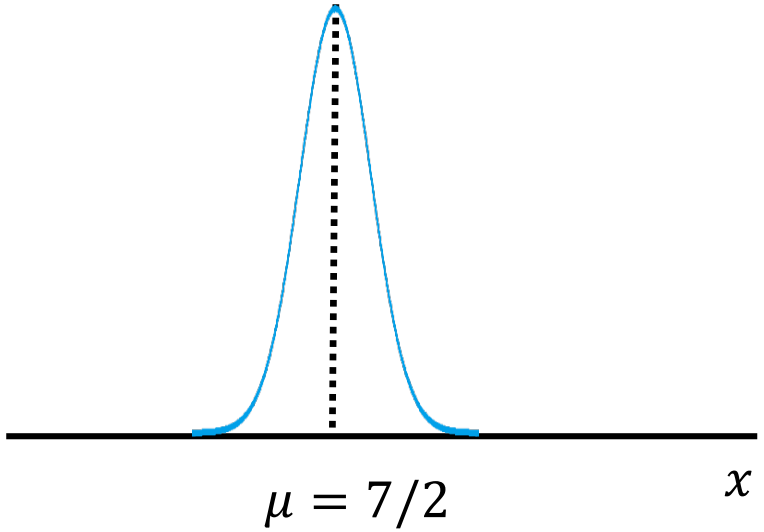
$$\mu = \frac{7}{2} \qquad \sigma^2 = \frac{35}{12}$$

標本平均の分布

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

$$= N\left(\frac{7}{2}, \frac{35}{12n}\right)$$

$$= N\left(\frac{7}{2}, \frac{35}{600}\right)$$



二項分布の正規分布近似

定 理　（ドモアブル-ラプラス）

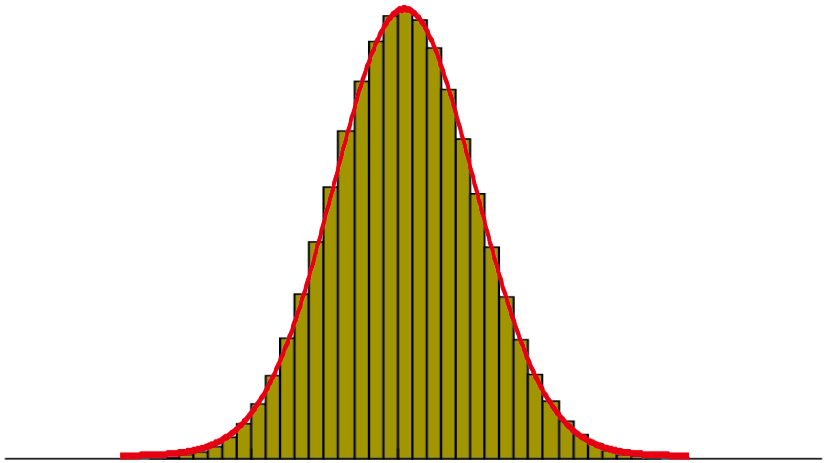
二項分布 $B(n,p)$ は，n が 大きいとき，同じ平均値と分散をもつ正規分布 $N(np\,,np(1-p))$ で近似できる．

証 明　$Z_1, Z_2, \cdots, Z_n, \cdots$ ：ベルヌーイ試行列．　平均値 $= p$，　分散 $= p(1-p)$

$X_n = Z_1 + Z_2 + \cdots + Z_n \sim B(n,p)$　これは定義！

一方，CLT により，　$\displaystyle\sum_{k=1}^{n} Z_k \approx N(np, np(1-p))$

したがって，　$B(n,p) \approx N(np, np(1-p))$



例 題 6.2　公平なコインを400回投げたとき，表が225回以上出る確率を求めよ.

例 題 6.2　公平なコインを400回投げたとき，表が225回以上出る確率を求めよ.

X：表の枚数　$X \sim B(400,0.5) \approx N(200, 10^2)$

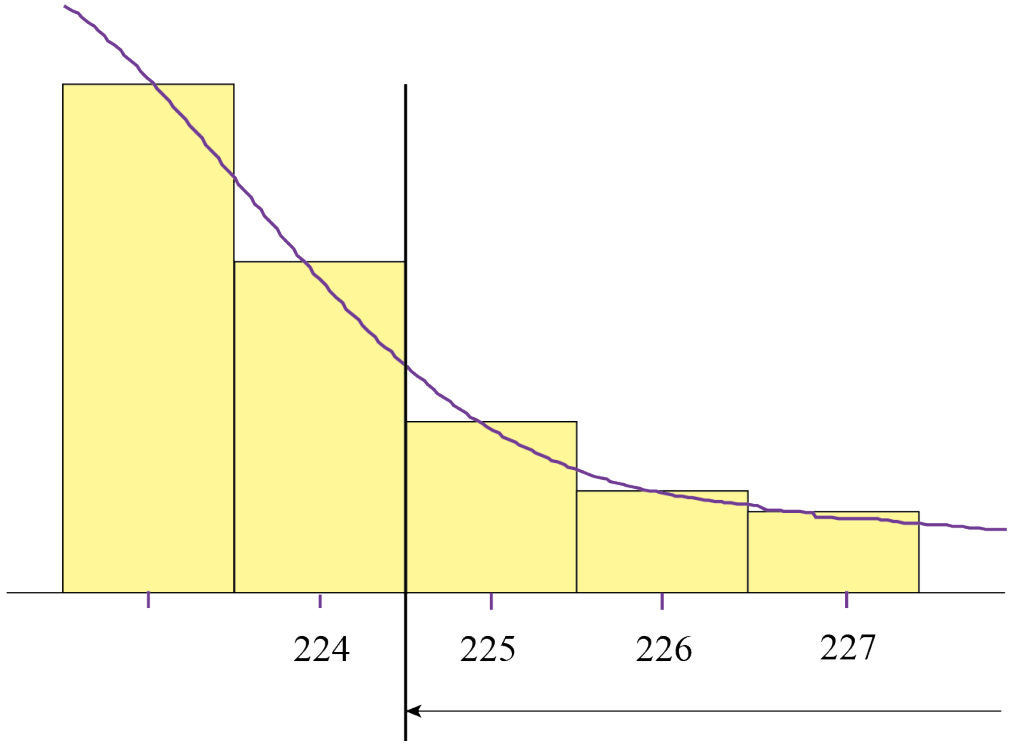
$P(X \geq 225) = P(X \geq 224.5)$

連続補正（半目補正）

$$= P\left(\frac{X-200}{10} \geq \frac{224.5-200}{10}\right)$$

$$= P(Z \geq 2.45)$$

$$= 0.5 - 0.4929 = 0.0071$$



中心極限定理（CLT）の変形

定 理　（中心極限定理＝CLT）

一般の母集団から取り出した n 個の無作為標本の標本平均 $\bar{X}$ の分布は，n が大きいときは近似的に，

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\bar{X}$ の標準化（z-変換）

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

よく使う変形

$$\sum_{k=1}^{n} X_k \approx N(n\mu, n\sigma^2)$$

$$\sum_{k=1}^{n} X_k - n\mu \approx N(0, n\sigma^2)$$

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} \frac{X_k - \mu}{\sigma} \approx N(0,1)$$

$$Z_k = \frac{X_k - \mu}{\sigma}$$　（標準化）

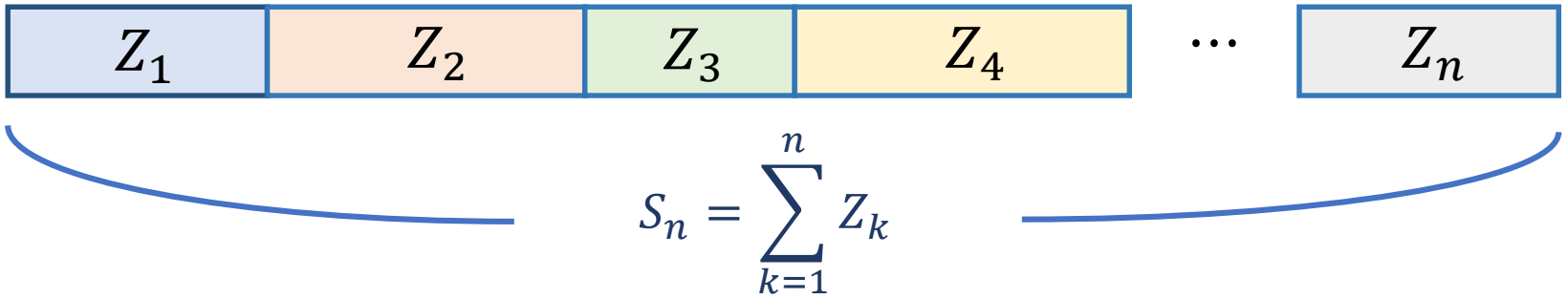
$$E[Z_k] = 0, \quad V[Z_k] = 1$$

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k \approx N(0,1)$$

実世界の揺らぎは細かい誤差の集積

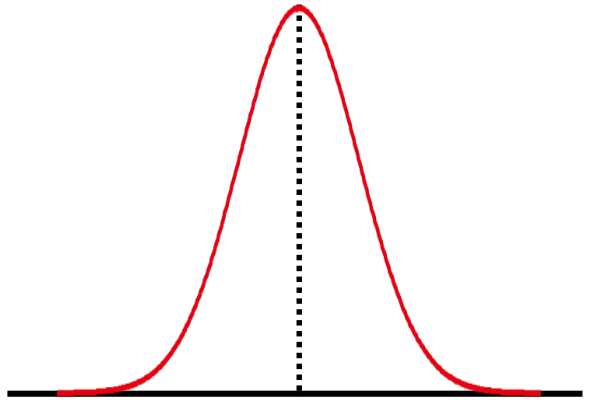
$Z_1, Z_2, \cdots, Z_n, \cdots$　：独立同分布な確率変数列

それぞれの平均値 $= \mu$,　分散 $= \sigma^2$



CLTによって,

$$S_n - n\mu = S_n - E[S_n] \approx N(0, n\sigma^2)$$



つまり，S_n は平均値のまわりに正規分布に従って分布する

大数の法則 (LLN)

定 理　（中心極限定理＝CLT）

母平均 μ，母分散 σ^2 の一般の母集団から取り出した n 個の無作為標本の標本平均 $\bar{X}$ の分布は, n が大きいときは近似的に,

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \;\approx\; N\left(\mu, \frac{\sigma^2}{n}\right) \qquad n \to \infty$$

$n \to \infty$ で分散がゼロになるので, 揺らぎが消えて μ に収束することが示唆される.
このことは, 次のように数学の定理として証明される（やや高度）

定 理　（大数の法則＝LLN）

母平均 μ の一般の母集団から取り出した n 個の無作為標本の標本平均 $\bar{X}$ について,

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \mu\right) = 1$$

コイン投げのシミュレーション

Jakob Bernoulli (1654-1705)

コイン投げ：$Z_1, Z_2, \cdots, Z_n, \cdots$

$Z_k = 1$ (表のとき)，　$Z_k = 0$ (裏のとき)，

母平均と母分散（$= Z_k$の平均値と分散）

$$\mu = E[Z_k] = 1 \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{1}{2}$$

$$\sigma^2 = E\left[Z_k^2\right] - E[Z_k]^2 = \mu - \mu^2 = \frac{1}{4}$$

CLTによって

$$T_n = \boxed{\frac{1}{n}\sum_{k=1}^{n} Z_k} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\frac{1}{2}, \frac{1}{4n}\right)$$

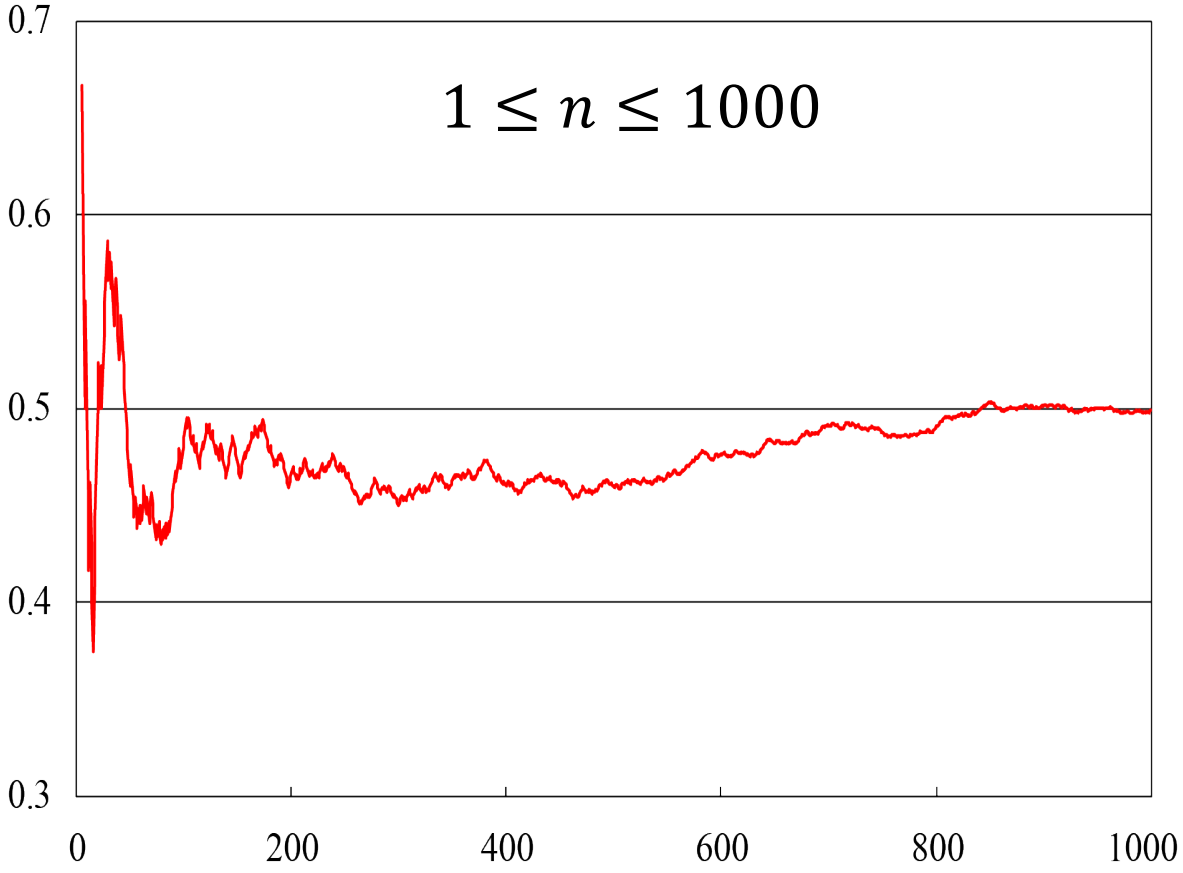
初めの n 回の内，表の回数の相対頻度

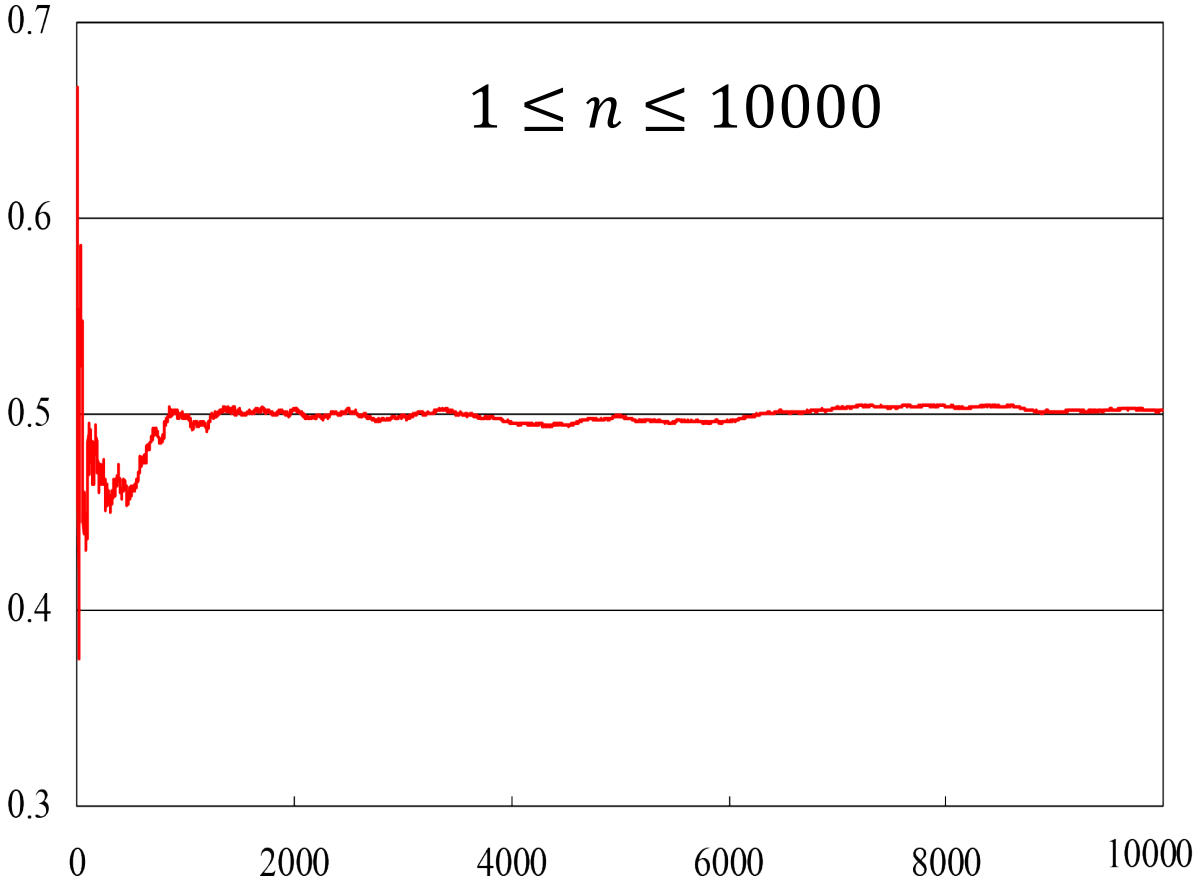
LLN

特に，確率 1 で

$$\lim_{n\to\infty} T_n = \frac{1}{2}$$

$$T_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$$





Python を使ってみる

➢ コイン投げのシミュレーション

・コイン投げ：$Z_1, Z_2, \cdots, Z_n, \cdots$

$$Z_k = \begin{cases} 1, & \text{（表のとき）} \\ 0, & \text{（裏のとき）} \end{cases}$$

・LLNの確認

確率１で　$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Z_k = \frac{1}{2}$

CoinTossLLN - Jupyter Notebook.pdf

➢ 標本分布の実例とCLTの確認

母集団分布として

・ベルヌイ分布（コイン投げ）

・一様分布

CLT-CoinToss - Jupyter Notebook.pdf

CLT-Uniform - Jupyter Notebook.pdf

➢ おまけ：ランダムウォーク

・公平なゼロサム賭け

$$Z_k = \begin{cases} 1, & \text{（勝のとき）} \\ -1, & \text{（負のとき）} \end{cases}$$

$$S_n = \sum_{k=1}^{n} Z_k \quad \text{（期末の利益）}$$

RandomWalk - Jupyter Notebook.pdf

Lecture 6

標本抽出と正規分布

おわり