統計学 01

早稲田大学政治経済学部

第6回

西郷 浩

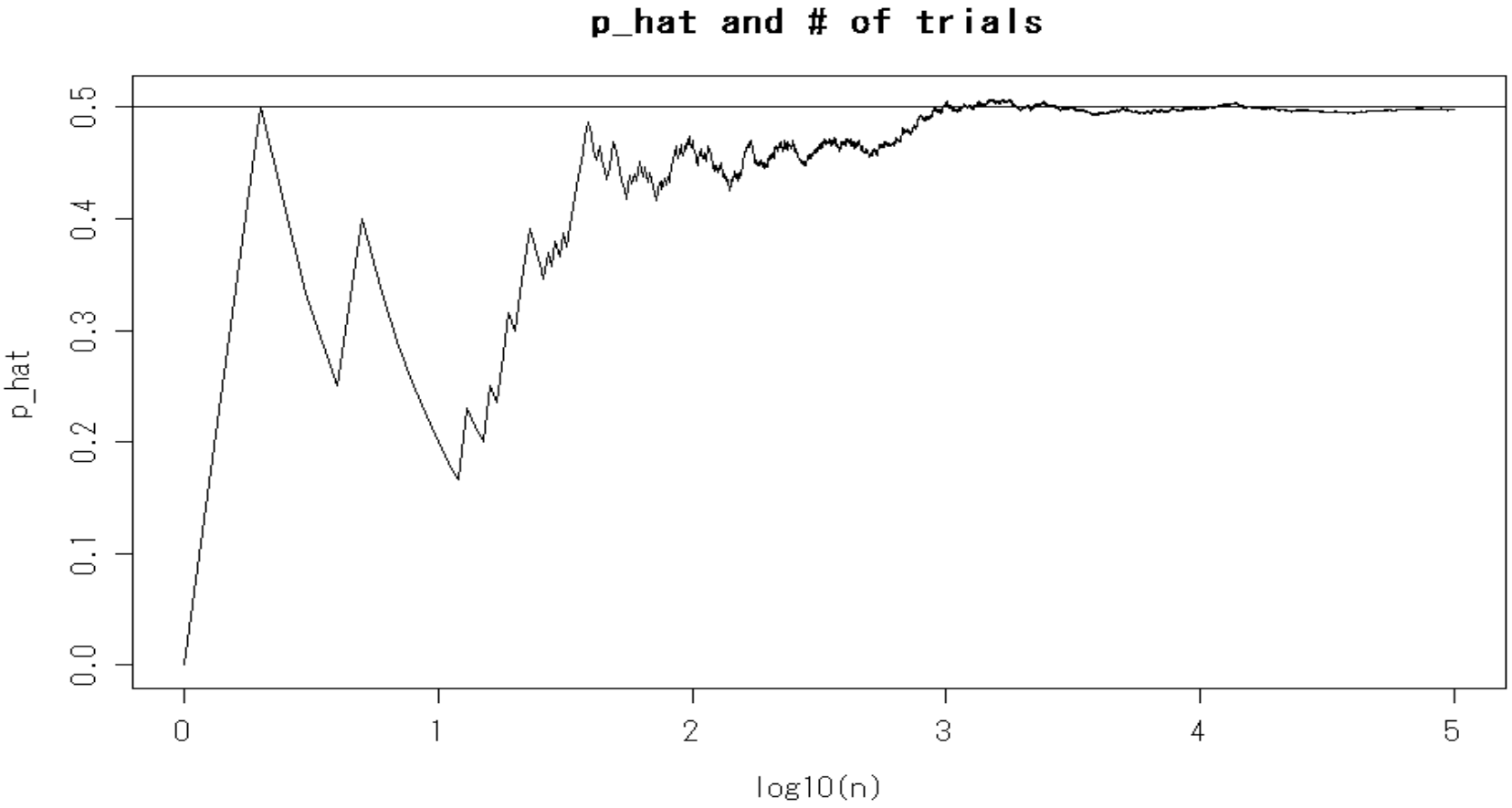
本日の目標

- 第8章 大数の法則と中心極限定理
  - 8.1 大数の法則
  - 8.2 中心極限定理
  - 8.3 中心極限定理の応用

大数の法則(1)

- 大数の法則
  - n 回の（独立な）コイントスの結果、表が出た割合 $\hat{p}$
  - n を大きくしていくと $\hat{p}$ は 1/2 （1回のコイントスで表が出る確率）に近づいていくか？
    - Yes ← 大数の法則
  - シミュレーション
    - 実際に n 回のコイントスと同じ確率的な仕組みをコンピューターの中に作り、結果を観察する。
    - 傍証にはなるが、証明にはならない。

大数の法則(2)



大数の法則(3)

- 大数の(弱)法則の正確な表現

  任意の$\varepsilon > 0$に対し　$P\left(\left|\hat{p} - 1/2\right| < \varepsilon\right) \to 1 \quad (n \to \infty)$

- 大数の(弱)法則の証明:チェビシェフの不等式の利用

$$P\left(\left|Y - E(Y)\right| < k\sqrt{V(Y)}\right) > 1 - 1/k^2$$

$$\Leftrightarrow P\left(\left|Y - E(Y)\right| < \varepsilon\right) > 1 - V(Y)/\varepsilon^2$$

$$\hat{p} = X/n \quad \text{ただし} \quad X \sim \mathrm{Bi}(n, 1/2)$$

$$E(\hat{p}) = E(X)/n = (n \times 1/2)/n = 1/2$$

$$V(\hat{p}) = V(X)/n^2 = (n \times 1/2 \times (1 - 1/2))/n^2 = 1/(4n)$$

$$P\left(\left|\hat{p} - 1/2\right| < \varepsilon\right) > 1 - 1/(4n\varepsilon^2) \to 1 \quad (n \to \infty)$$

大数の法則(4)

- 一般に
  - X_i (i=1,2,…,n) が同一分布からの独立な確率変数
  - $E(X_i) = \mu$, $V(X_i) = \sigma^2$

であれば、任意の正数εに対して、

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

$$\because \quad P\left(\left|\overline{X} - \mu\right| < \varepsilon\right)$$

$$= P\left(\left|\overline{X} - E\left(\overline{X}\right)\right| < \varepsilon\right) > 1 - \sigma^2/\left(n\varepsilon^2\right)$$

$$\to 1 \quad (n \to \infty)$$

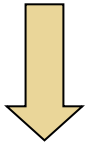
大数の法則(5)

- 大数の法則の意味
  - 観測数（サンプルサイズ）n が大きくなるにつれて、観察値の平均が真の平均の近辺に発生する確率がいくらでも1に近づく（確率収束する）。
    - 「観測数（サンプルサイズ）を大きくした方が正確な値がえられる」という直観に対する理論的な根拠をあたえる。

中心極限定理(1)

- X_i (i=1,2,…,n)：同一分布からの独立な確率変数
- 平均値：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- このとき、$E(\overline{X}) = \mu$ (ただし $\mu = E(X_i)$)

  $V(\overline{X}) = \sigma^2/n$ (ただし $\sigma^2 = V(X_i)$)
- 確率変数としての $\overline{X}$ の確率分布は？
  - $\overline{X}$ の確率分布 ← X_i の発生もとの確率分布



$n \to \infty$

正規分布

中心極限定理(2)

- 厳密な表現

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \xrightarrow[D]{} N(0,1)$$

あるいは（本当は不正確な表現）

$$\bar{X} \xrightarrow[D]{} N\left(\mu, \sigma^2/n\right)$$

  - 要は「発生もとの確率分布がどのようなものであっても、（標本）平均の確率分布は正規分布で近似できる」と覚えておけば（実用上は）間違いない。

中心極限定理(3)

- 例：X_i が相互に独立に Bernoulli(0.1)にしたがう。

$$X_i = \begin{cases} 1 & (\text{確率 } 0.1) \\ 0 & (\text{確率 } 0.9) \end{cases}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (= \hat{p})$$

n=10, 20, 30, 40 の場合の $\overline{X}$ の確率分布

  - 次のスライド
  - n が 10→20→30→40 となるにつれて正規分布に近づいていく様子がわかる。

中心極限定理(4)

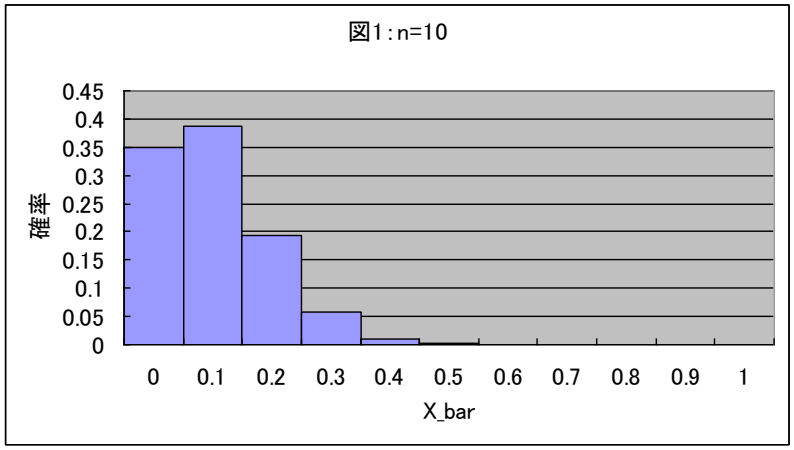


図1：n=10
確率
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
X_bar

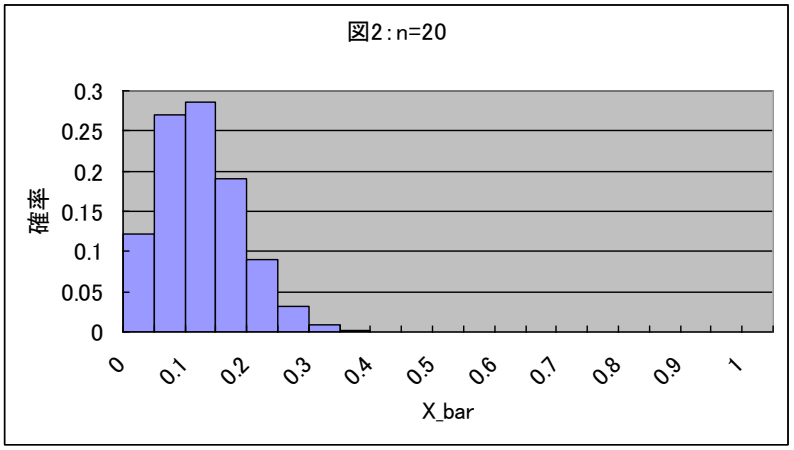


図2：n=20
確率
0.3
0.25
0.2
0.15
0.1
0.05
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
X_bar

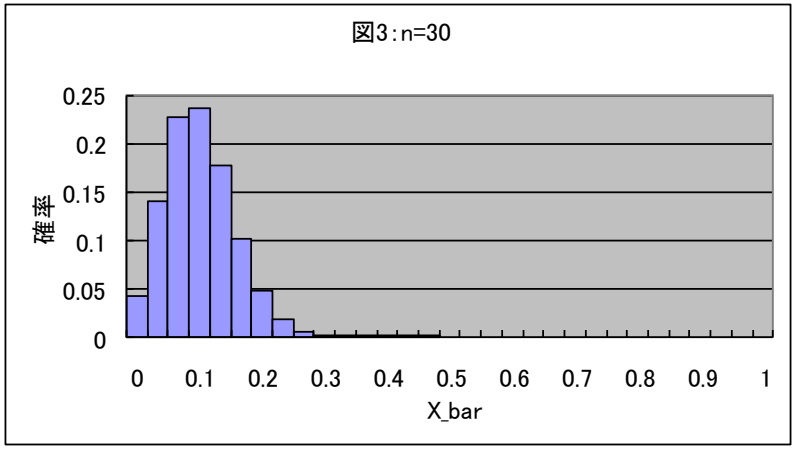


図3：n=30
確率
0.25
0.2
0.15
0.1
0.05
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
X_bar

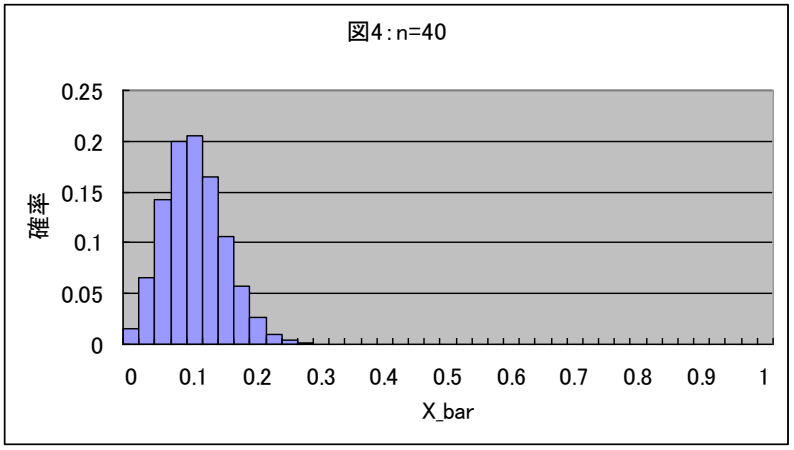


図4：n=40
確率
0.25
0.2
0.15
0.1
0.05
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
X_bar

中心極限定理(5)

- 限定的ながら厳密な証明：教科書 pp. 164-165.
- 発生もとの分布が指数分布のときのシミュレーション：教科書 図8.18-8.21（p. 168）

中心極限定理の応用(1)

- 二項分布の正規近似
  - 確率変数の和の確率分布

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n\sigma^2}}\underset{D}{\rightarrow}N(0,1)$$

つまり、（少し不正確ながら）

$$\sum_{i=1}^{n}X_i\underset{D}{\rightarrow}N\left(n\mu,n\sigma^2\right)$$

つまり、和の分布も正規分布で近似できる。

中心極限定理の応用(2)

- とくに X_i ～ Bernoulli(p) のとき、$X = \sum_i X_i$ ～Bi(n,p). したがって、n がある程度大きいとき、

$$P(X \le k) = P\left(\frac{X - np}{\sqrt{np(1-p)}} \le \frac{k - np}{\sqrt{np(1-p)}} \right)$$

$$\approx P\left(Z \le \frac{k - np}{\sqrt{np(1-p)}} \right) \quad \text{ただし、} \quad Z \sim N(0,1)$$

$$= \Phi\left(\frac{k - np}{\sqrt{np(1-p)}} \right) \quad \text{標準正規分布の下側確率}$$

中心極限定理の応用(3)

- 例：n 回のコイントスのうち、表の出た回数をXとする。
  - $n = 10$ で $X \leqq 4$となる確率
  - $n = 100$ で $X \leqq 40$となる確率
  - $n = 1000$ で $X \leqq 400$となる確率
  - $n = 10000$ で $X \leqq 4000$となる確率
  - 上記の例においてすべて $\hat{p} = X/n \leq 0.4$
  - 計算の方針
    - 正規近似を利用して計算する。
    - 二項分布による正確な確率も計算する。

中心極限定理の応用(4)

- 結果
  - *n* が大きくなるにつれて、P(*p_hat* ≦0.4)となる確率はほとんどなくなる。
  - *n* が小さいときには近似の精度が悪い。
    - 簡単な補正によって近似の精度を向上させることができる。

$$P(X \le k) \approx \Phi\left(\frac{k+1/2-np}{\sqrt{np(1-p)}}\right)$$

表1:二項分布の正規近似(補正なし)

n	正規近似	二項分布
10	0.2635	0.3770
100	0.0228	0.0284
1000	0.0000	0.0000
10000	0.0000	0.0000

注:P(*p_hat* ≦0.4 となる確率)

表2:二項分布の正規近似(補正あり)

n	正規近似	二項分布
10	0.3759	0.3770
100	0.0287	0.0284
1000	0.0000	0.0000
10000	0.0000	0.0000

注:P(*p_hat* ≦0.4 となる確率)

中心極限定理の応用(5)

- 一様分布を利用した正規乱数の発生
  - (0,1)の実数値一様乱数 $X \sim U(0,1)$
    - 0から1までの数が一様に出現する乱数
    - 確率密度関数は $f(x) = 1$ (if $x \in (0,1)$) or 0 (otherwise)
    - $E(X) = 1/2$, $V(X) = 1/12$
  - 正規乱数の発生
    - 中心極限定理（和の確率分布が正規分布で近似できる）を援用する。
    - $n = 12$ として $Z = \sum_i X_i - 6$ ただし、相互に独立に　$X_i \sim U(0,1)$

中心極限定理の応用(6)

$$E(Z)=E\left(\sum_{i=1}^{12}X_i-6\right)=\sum_{i=1}^{12}E(X_i)-6=12\times 1/2-6=0$$

$$V(Z)=V\left(\sum_{i=1}^{12}X_i-6\right)=V\left(\sum_{i=1}^{12}X_i\right)=\sum_{i=1}^{12}V(X_i)=12\times 1/12=1$$

中心極限定理が働くので、

$P(Z\le z)\approx\Phi(z)$　(標準正規分布の下側確率)

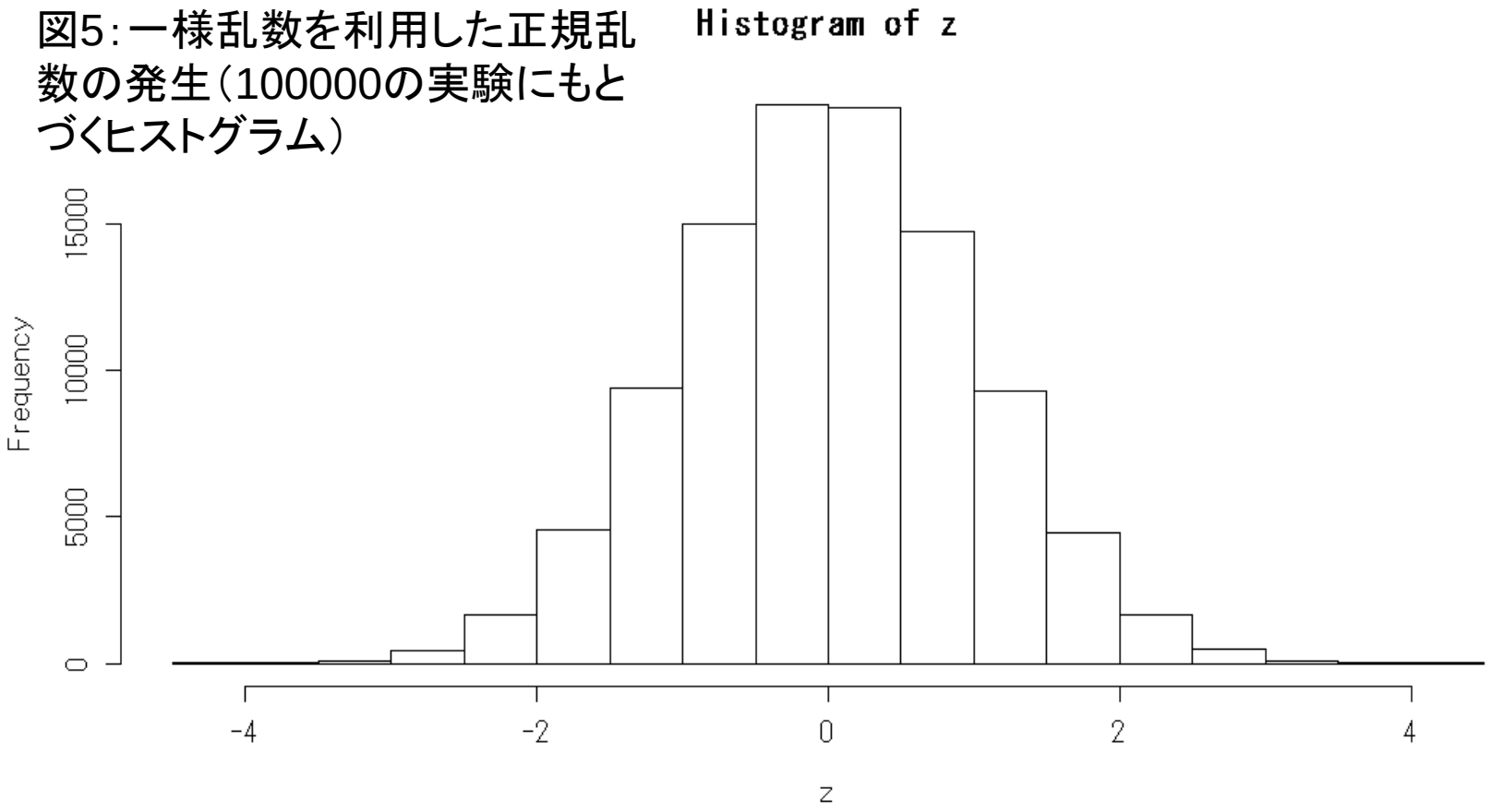
つまり:

　Zは標準正規分布に近い確率変数

である（はず）。

中心極限定理の応用(7)

図5:一様乱数を利用した正規乱数の発生(100000の実験にもとづくヒストグラム)



中心極限定理の応用(8)

- 現在はもっと効率的な正規乱数の発生方法が使われている。