統計学II

早稲田大学政治経済学術院
西郷 浩

本日の目標

- 標本平均の性質
  - 大数の法則
    - サンプルサイズ n が大きくなるにつれて、標本平均 $\bar{X}$ の値が母平均 μ の値に近づく。
  - 中心極限定理
    - サンプルサイズ n が大きくなるにつれて、標本平均 $\bar{X}$ の確率分布（標本分布）が正規分布で近似できる。
      - $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \to_d \ N(0,1)$
    - X_i の分布が指定されていない点が、著しい特徴になる。
    - 2項分布が正規分布で近似できることも意味する。

利用データについて

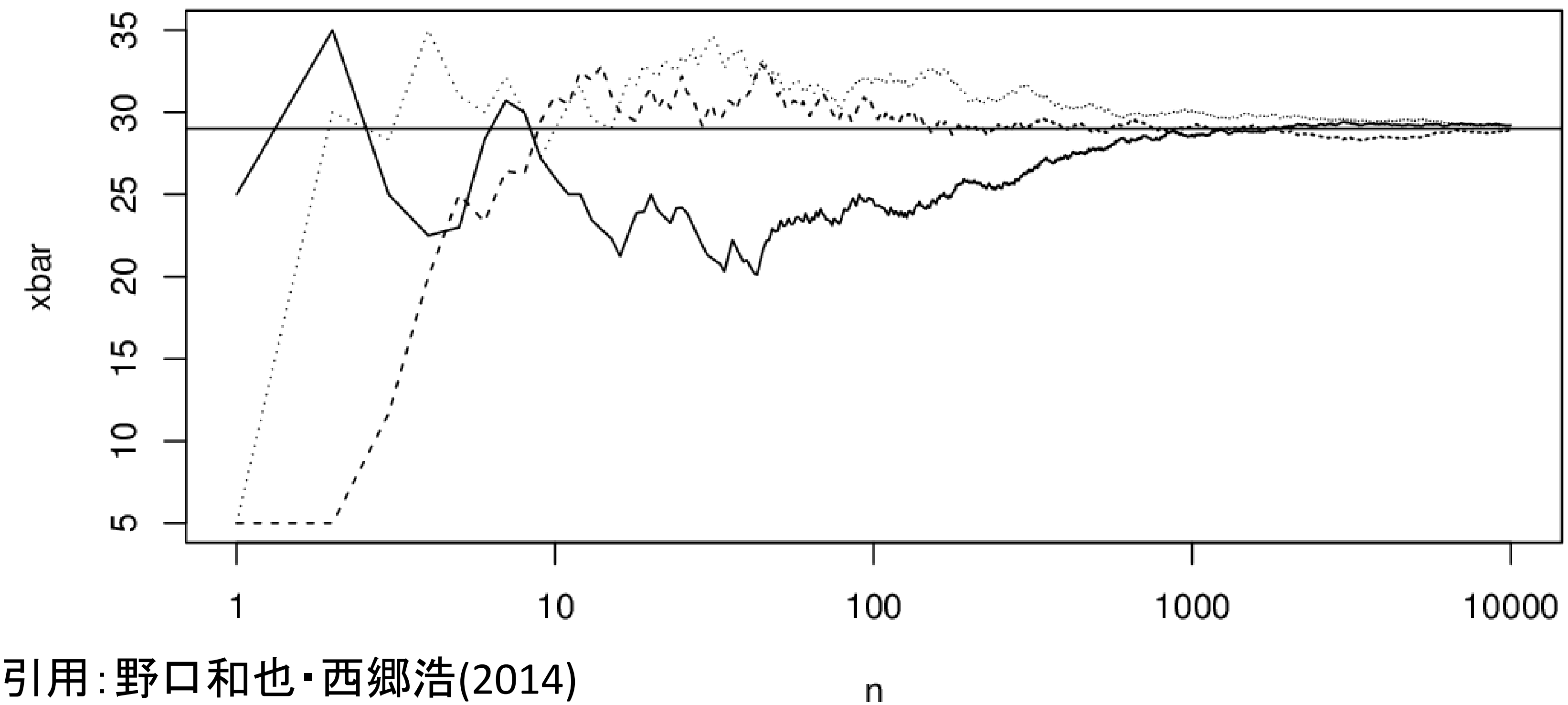
- JGSS2010年
  - 抽出実験に用いたデータ（図4）は、東京大学社会科学研究所付属社会調査・データアーカイブ研究センターSSJデータアーカイブのリモート集計システムを利用し、同データアーカイブが所蔵する[JGSS2010]の個票をデータを集計したものである。

大数の法則(1)

- 実験
  - [5, 15, 25, 45, 55] から復元抽出によって抜き取った数値 $X_1, X_2, \ldots$ の算術平均 $(1/n)\sum_{i=1}^{n} X_i$ を逐次計算して、$(n, (1/n)\sum_{i=1}^{n} X_i)$ を $n = 10{,}000$ まで描画する。
  - この実験を3回繰り返す(結果の安定性を確認するため)。
- 結果
  - 3回とも、5つの数字の平均(母平均μ)に近づいていくように見える。
  - このことから、以下のことが結論できそうである。
    - サンプルサイズ n が大きくなるにつれて、標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ が母平均 μ のごく近辺に出現しやすくなる。

大数の法則(2)

図1：復元抽出の反復実験



引用：野口和也・西郷浩(2014)
『基本 統計学』培風館 p. 124

大数の法則(3)

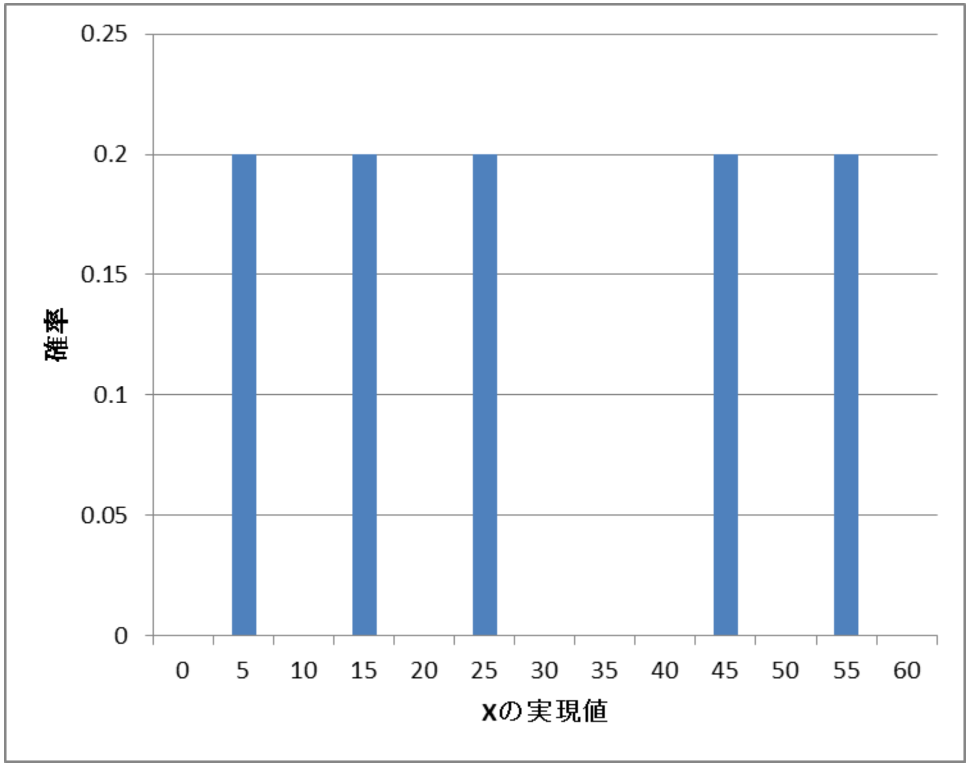
- 大数の（弱）法則
  - 前提：$X_i \sim_{iid} (\mu, \sigma^2),\ i = 1, 2, \ldots.$
  - 結論：任意の $\varepsilon > 0$ について、nが大きくなるにつれて $P(|\bar{X} - \mu| > \varepsilon) \to 0.$
  - 理由（証明）：
    - $V(\bar{X}) = \sigma^2/n$
    - チェビシェフの不等式により、
      $P(|\bar{X} - \mu| > \varepsilon) \leq V(\bar{X})/\varepsilon^2 = \frac{\sigma^2/\varepsilon^2}{n} \to 0$ as $n \to \infty.$
  - 記法
    - 任意の$\varepsilon > 0$について $\lim_{n\to\infty} P(|\bar{X} - \mu| > \varepsilon) = 0.$
    - $\bar{X} \to_p \ \mu$
    - $\operatorname*{plim}_{n\to\infty} \bar{X} = \mu$

中心極限定理(1)

- 実験1（厳密な確率分布）
  - [5, 15, 25, 45, 55] から復元抽出によって抜き取った数値 $X_1, X_2, \ldots, X_n$ の標本平均の標本分布（標本抽出にともなって発生する確率分布）
    - 標本の大きさ $n = 2, 4, 8.$
    - 標本の大きさ n が大きくなるにつれ、正規分布に近づいているように見える（図2）。

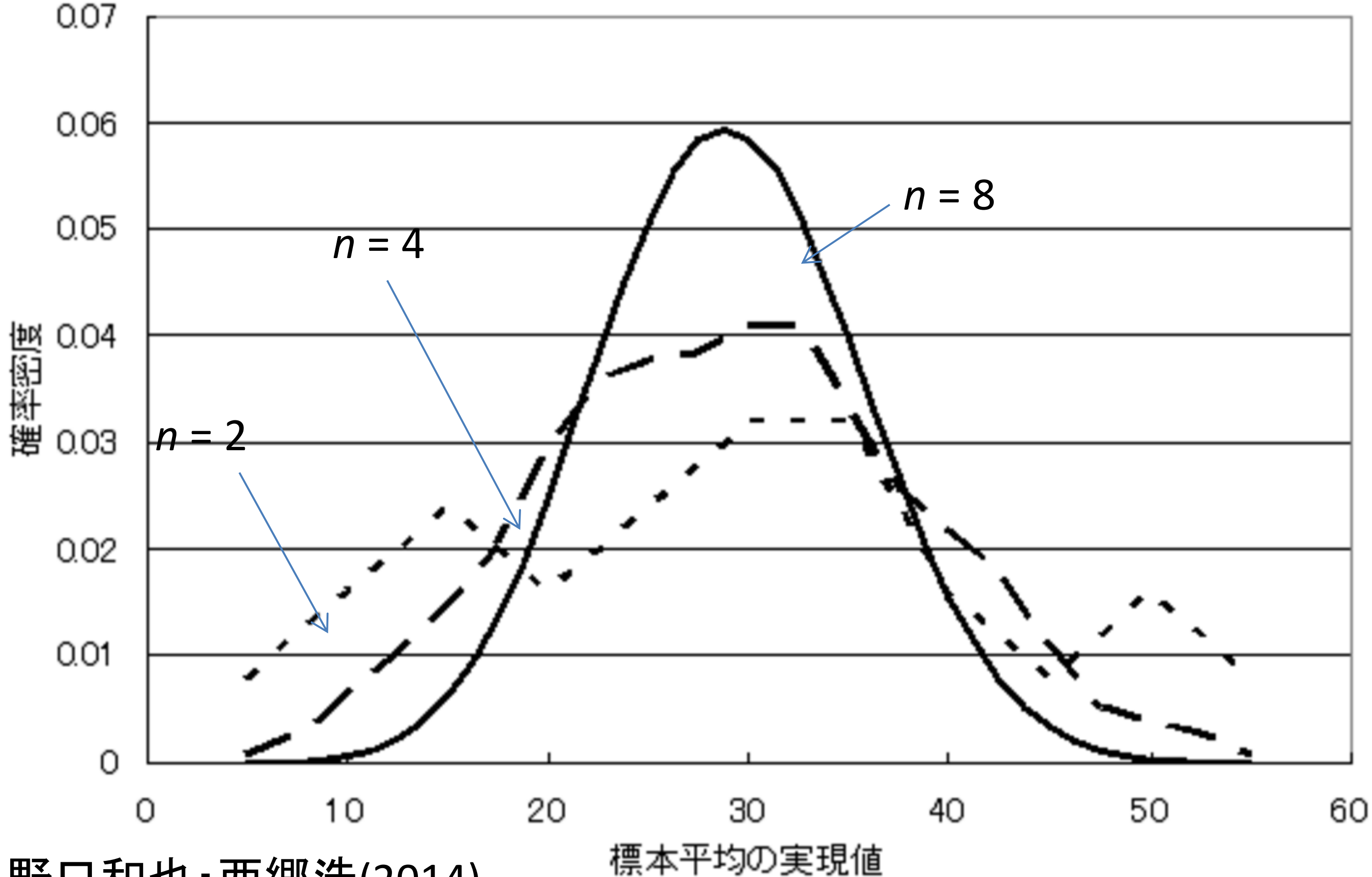
中心極限定理(2)

図2：実験1の母集団分布（確率分布）



中心極限定理(3)

図3：復元抽出における標本平均の標本分布



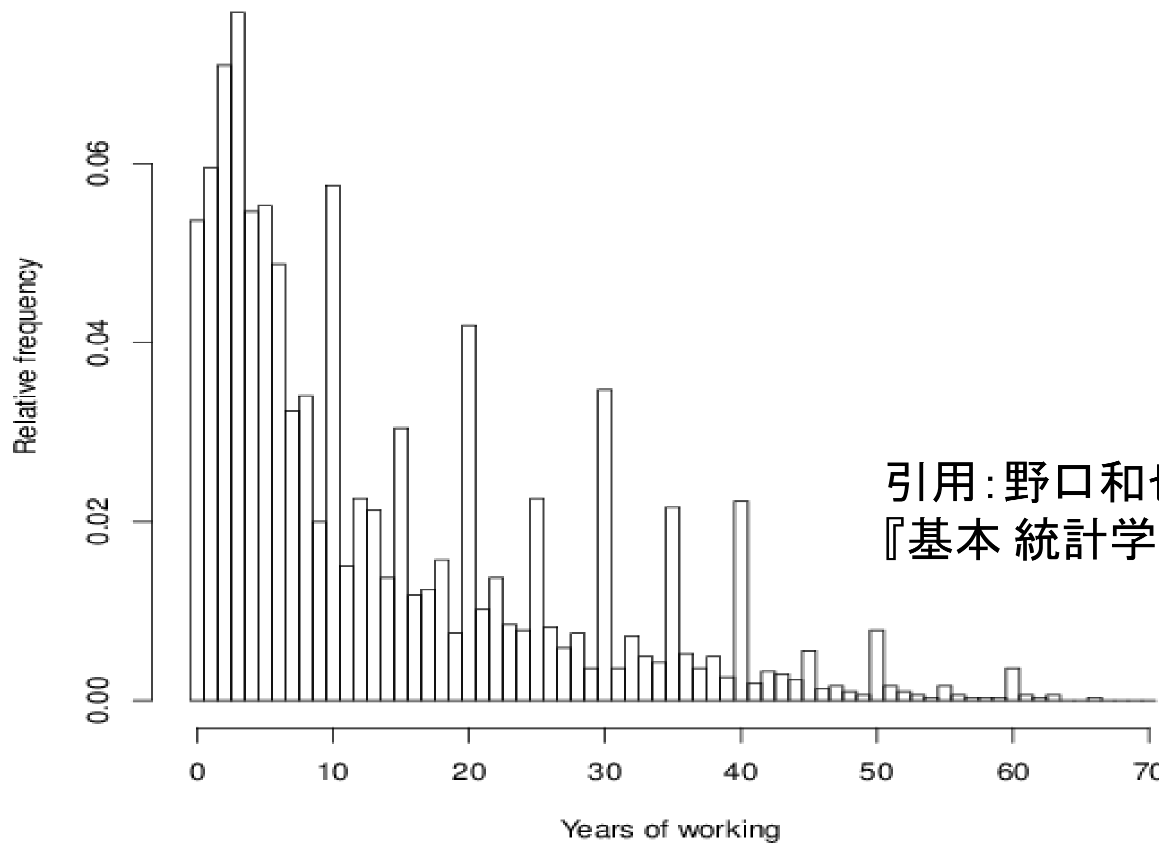
引用：野口和也・西郷浩(2014)
『基本 統計学』培風館 p. 127

中心極限定理(4)

- 実験2（シミュレーションによる近似）
  - JGSS2010の就労年数（母集団の大きさ *N* = 3,056）からの標本抽出
    - $n = 2, 4, 8, 16, 32.$
    - 実験を多数回（100,000回）反復して、
    - 復元抽出の場合（図3）
    - 非復元抽出の場合（図4）
    - 標本の大きさ n が大きくなるにつれ、正規分布に近づいているように見える。

中心極限定理(5)

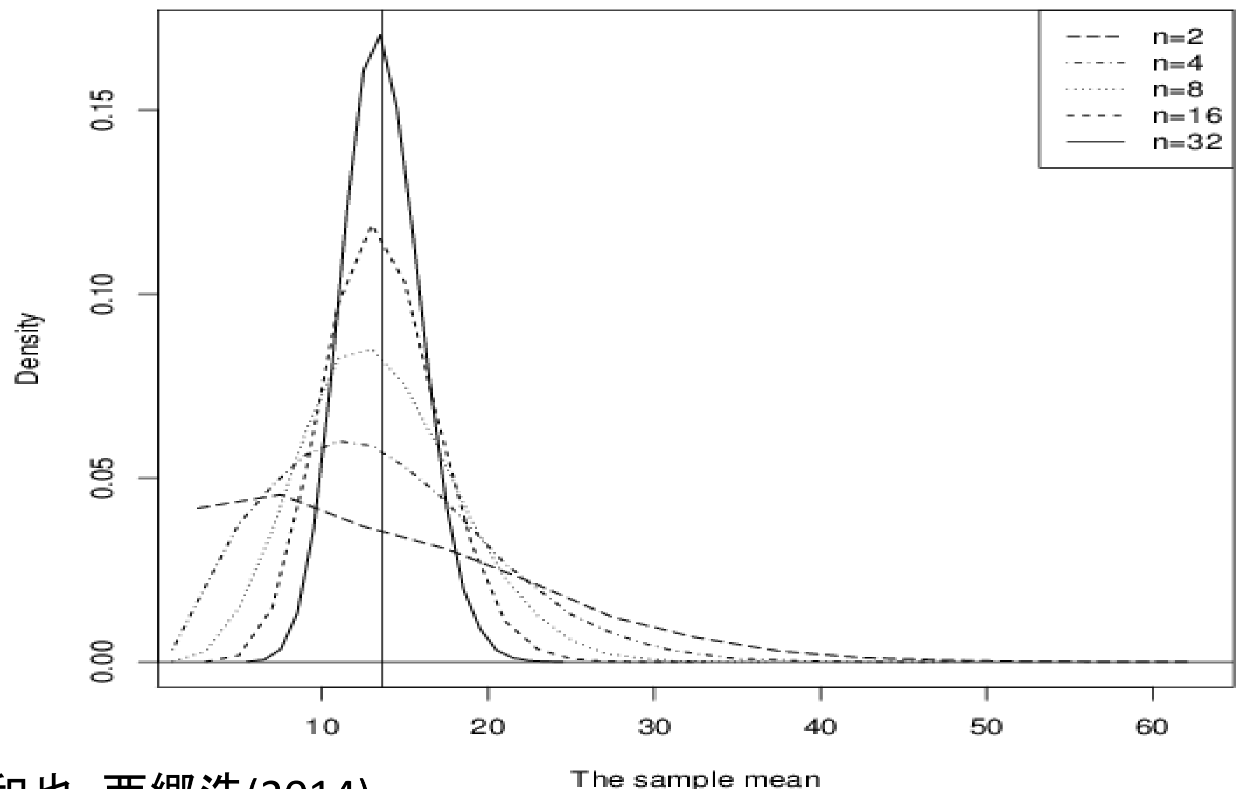
図4：実験2の母集団分布（就労年数の度数分布）



引用：野口和也・西郷浩(2014)
『基本 統計学』培風館 p. 128

中心極限定理(6)

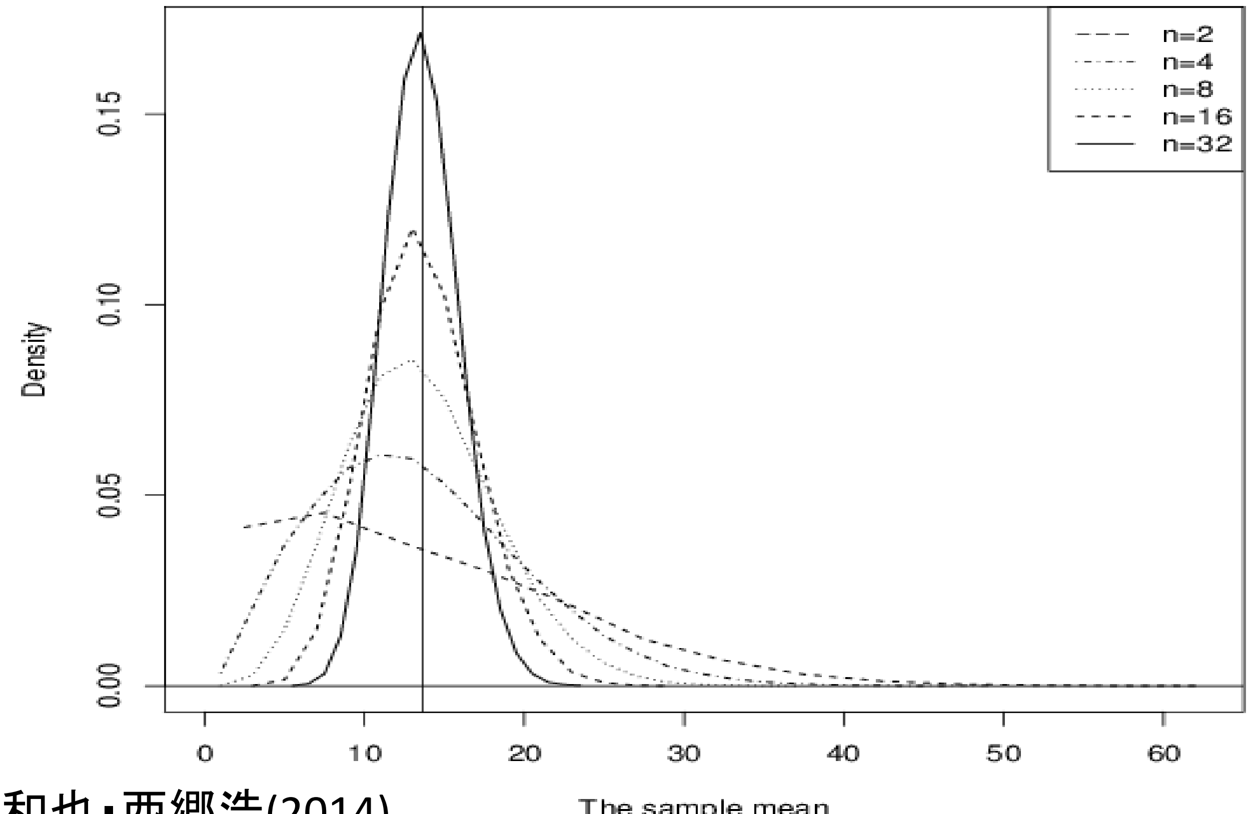
図5：実験2における標本平均の標本分布（復元抽出）



引用：野口和也・西郷浩(2014)
『基本 統計学』培風館 p. 129

中心極限定理(7)

図6：実験2における標本平均の標本分布（非復元抽出）



引用：野口和也・西郷浩(2014)
『基本 統計学』培風館 p. 129

2項分布の正規近似(1)

- ベルヌーイ確率変数への中心極限定理の応用
  - $X_i \sim_{iid} Bernoulli(p)$ つまり
    $$X_i = \begin{cases} 1 & P(X_i = 1) = p \\ 0 & P(X_i = 0) = 1 - p \end{cases}$$
    のときにも、$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \to_d \; N(0,1)$
  - ベルヌーイ確率変数については、
    - $\mu = E(X_i) = p, \sigma^2 = V(X_i) = p(1-p)$
    - $X = \sum_{i=1}^{n} X_i = n\bar{X} \sim Binomial(n, p)$

2項分布の正規近似(2)

– 両者を合わせると

- $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X}-p}{\sqrt{p(1-p)/n}} = \frac{X-np}{\sqrt{np(1-p)}} \to_d \; N(0,1)$
- つまり、$Binomial(n,p)$ が $N\big(np, np(1-p)\big)$ で近似できる。
  – ただし、以下の条件が必要
    » サンプルサイズ n が大きい。
    » $Binomial(n,p)$ が強く歪んでない。

2項分布の正規近似(3)

- 数値例
  - $X \sim Binomial(n, p)$ のとき、$P(X \leq 0.3n) = P\left(\frac{X}{n} \leq 0.3\right)$ を2項分布（正確）と正規近似とで計算して結果を比較する。

2項分布の正規近似(4)

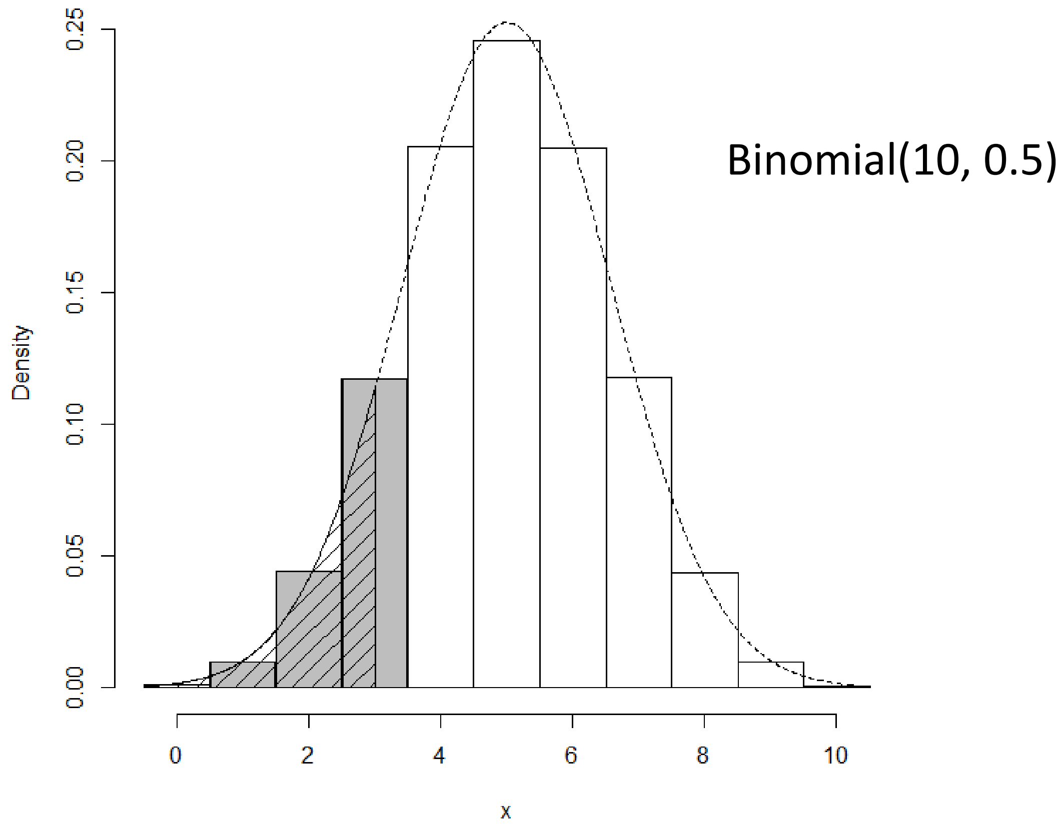
- 数値例（つづき）

表1：2項分布の正規近似（$p = 0.5$）

n	Binomial	Normal
10	0.1718750	0.1029516
20	0.0576591	0.0368191
40	0.0082945	0.0057060
80	0.0002258	0.0001733
160	2.255E-07	2.1E-07

2項分布の正規近似(5)

図7：2項分布の正規近似（半数補正なし）



2項分布の正規近似(6)

- 2項分布の正規近似

補正あり

$$P(X \le a) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \le \frac{a-np}{\sqrt{np(1-p)}}\right) \approx \begin{cases} P\left(Z \le \dfrac{a+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) \\ P\left(Z \le \dfrac{a-np}{\sqrt{np(1-p)}}\right) \end{cases}$$

補正なし

  - ただし、
    - $a = 1, 2, \ldots, n,\ Z \sim N(0,1)$
- 境界線の柱が、計算すべき面積に含められるかどうかで、補正の方法（1/2 を足すか引くか）が異なることに注意する。
  - 例：$P(X < a) \approx P\left(Z < \frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$
  - 図を描くのが最善の対処方法である。

2項分布の正規近似(7)

- 数値例(つづき)

表1:2項分布の正規近似($p = 0.5$)　　　　（半数補正）

n	Binomial	Normal	Normal_end
10	0.1718750	0.1029516	0.1713909
20	0.0576591	0.0368191	0.0587624
40	0.0082945	0.0057060	0.0088530
80	0.0002258	0.0001733	0.0002642
160	2.255E-07	2.1E-07	3.17E-07