

# 統計学 ~ 中心極限定理について

谷崎 久志 (神戸大学大学院経済学研究科)

## 1 はじめに

本稿では，神戸大学経済学部的一年生での必修科目「統計学」(教科書は，豊田・大谷・小川・長谷川・谷崎 (2002) 『基本統計学 (第2版)』東洋経済新報社。以下，豊田他 (2002) とする) から，最も重要な定理の一つである中心極限定理 (豊田他 (2002) の P.80 からの 6.3 節を参照のこと) を取り上げる。この定理は，確率変数  $X_1, X_2, \dots, X_n$  の単純平均 ( $n$  個の確率変数の単純平均という意味で，添え字に  $n$  を付けて， $\bar{X}_n$  と表記する)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

を正規化 (平均 0，分散 1 に変換すること，豊田他 (2002) の P.52 の定理 4.4 を参照のこと) したものを ( $Z_n$  と表記する)

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}}$$

は， $n$  が大きくなるにつれて標準正規分布  $N(0, 1)$  に近づくという内容である (正規分布については，豊田他 (2002) の P.65 からの第 5 章を参照せよ)。標準正規分布の密度関数は  $f(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$  である ( $\exp(x) = e^x$  に注意)。この定理が成り立つためには，個々の確率変数  $X_i$  の平均  $E(X_i)$  と分散  $V(X_i)$  がすべての  $i$  について有限でなければならない。話を簡単にするために， $X_1, X_2, \dots, X_n$

は互いに独立で，しかも，すべての  $i$  について  $E(X_i) = \mu$ ， $V(X_i) = \sigma^2$  としよう。  
このとき，

$$E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu \quad (1)$$

$$V(\bar{X}_n) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{\sigma^2}{n} \quad (2)$$

が得られる。(1) 式について，2 つ目の等式は豊田他 (2002) の P51 の定理 4.1，3 つ目の等式は P57 の定理 4.5 が使われている。(2) 式について，2 つ目の等式は P52 の定理 4.3，3 つ目の等式は P59 の定理 4.8 が用いられている。2 つの式をまとめて，P59 の定理 4.9 に表されている。よって， $n$  が大きくなると，

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

が標準正規分布  $N(0,1)$  に近づくことになる。ポイントは，

- (i)  $X_i$  の分布が何であっても，
- (ii)  $\mu = E(X_i)$  と  $\sigma^2 = V(X_i)$  が有限であれば，
- (iii)  $n$  が大きいとき，
- (iv)  $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$  が成り立つ。ただし， $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  とする。

ということである ( $\sigma^2 = V(X_i)$  が有限であるということは， $E(X_i^2)$  が有限であるということの意味する)。

さらに， $\sigma^2$  をその推定量  $S^2$  で置き換えて，(ii) と (iv) の代わりに，

- (ii)'  $k = 1, 2, 3, 4$  について， $E(X_i^k)$  が有限であれば，
- (iv)'  $T_n = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \rightarrow N(0,1)$  が成り立つ。ただし， $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  とする。

を用いてもよい。(ii)' については， $S^2$  の分散が有限であること，すなわち， $E(X_i^4)$  が有限でなければならないことから生じる条件である。

中心極限定理を数学的に証明することも可能であるが，本稿では，モンテ・カルロ実験によって擬似乱数を生成させてこの定理が成り立つことを確認する。

## 2 モンテ・カルロ実験のセットアップ

まず初めに，前節の (iv) を示すために，下記のような手順で，モンテ・カルロ実験を行う。

- (1)  $X_i$  にある分布を仮定して， $n$  個の乱数を発生させる (具体的にどのような分布を採用したかについては後述する)。発生させた  $n$  個の乱数値を小文字で  $x_1, x_2, \dots, x_n$  と表す。
- (2)  $X_i$  に分布を仮定すると， $\mu = E(X_i)$  と  $\sigma^2 = V(X_i)$  を求めることができるので， $z_n = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$  を計算することができる。ただし， $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$  とする。
- (3) ステップ (1) と (2) を  $M$  回繰り返す。 $M$  個の系列を  $z_n^{(m)}$ ， $m = 1, 2, \dots, M$  と書くことにする。 $M$  個のデータ  $z_n^{(1)}, z_n^{(2)}, \dots, z_n^{(M)}$  をもとにして，ヒストグラムを作成する。ヒストグラムの範囲は  $-4.9 \sim 4.9$  として，それぞれの階級の幅は  $0.2$  とする (すなわち，階級の数  $49$  となる)。

$X_i$  の分布を

- 区間  $(0,1)$  の一様分布 (一様分布とは， $x$  のとる値は  $0 < x < 1$  の実数値で，密度関数は  $f(x) = 1$  となる連続型分布) → 図 1
- $p = 0.3$  のベルヌイ分布 (ベルヌイ分布とは， $x$  のとる値は  $x = 0, 1$  の 2 つで，確率関数は  $f(x) = p^x(1-p)^{1-x}$  となる離散型分布) → 図 2

の 2 つを考える。 $M = 1,000,000$  とした。さらに，一様分布のときは  $n = 2, 3, 10, 100, 1,000$  とし，ベルヌイ分布のときは  $n = 3, 10, 1,000, 10,000, 100,000$  とした。一様分布の場合， $\mu = E(X_i) = 1/2$ ， $\sigma^2 = V(X_i) = 1/12$  となる。ベルヌイ分布の場合， $\mu = E(X_i) = p = 0.3$ ， $\sigma^2 = V(X_i) = p(1-p) = 0.21$  となる。平均  $\mu$ ，分散  $\sigma^2$  の計算の仕方については，豊田他 (2002) の 4.2.1 節と 4.2.2 節を参照せよ。なお，区間  $(0,1)$  の一様乱数について，Excel では `rand()` という関数が用意されていて簡単に生成できる。また，区間  $(0,1)$  の一様乱数 ( $u$  とする) を生成し， $u \leq p$  のとき  $X_i = 1$ ， $u > p$  のとき  $X_i = 0$  とすれば，ベルヌイ乱数を生成することができる。

結果は，一様分布の場合は図 1，ベルヌイ分布の場合は図 2 となった。比較のため，すべてのヒストグラムに標準正規分布の密度関数が実線で描かれてい

図 1: 一様分布

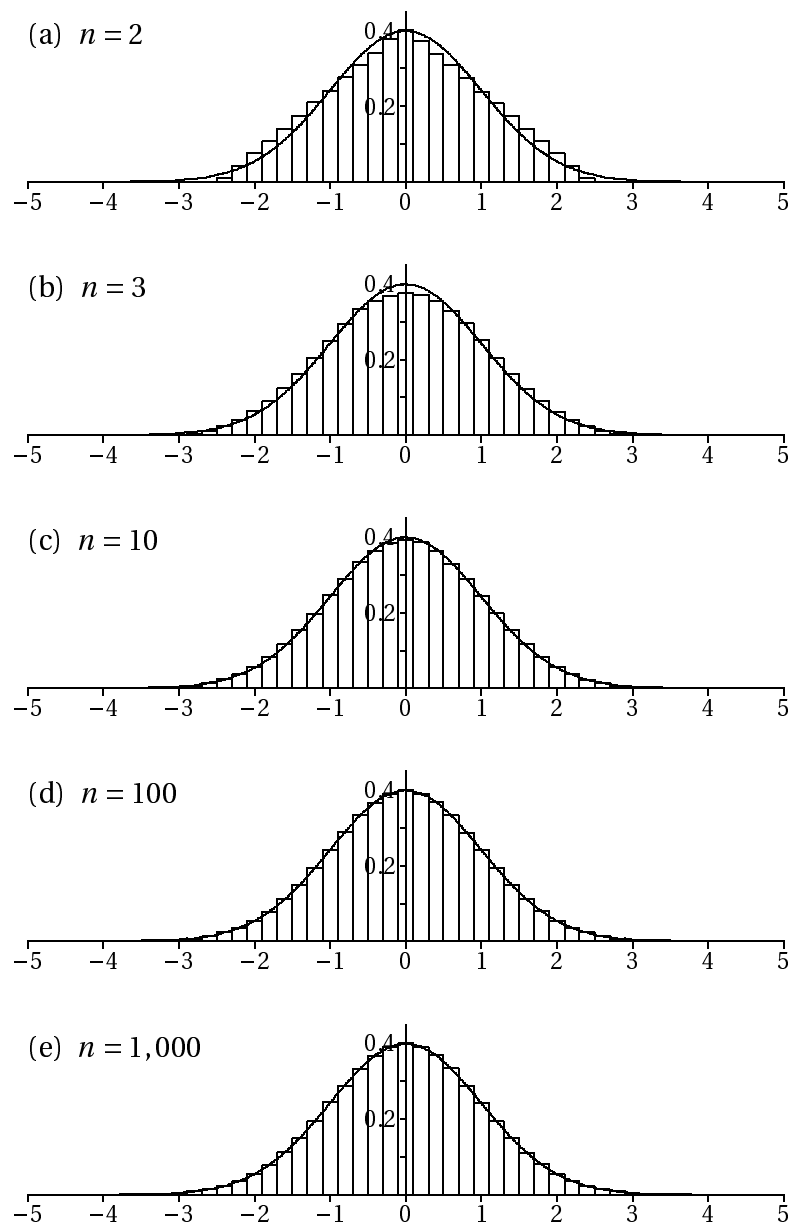
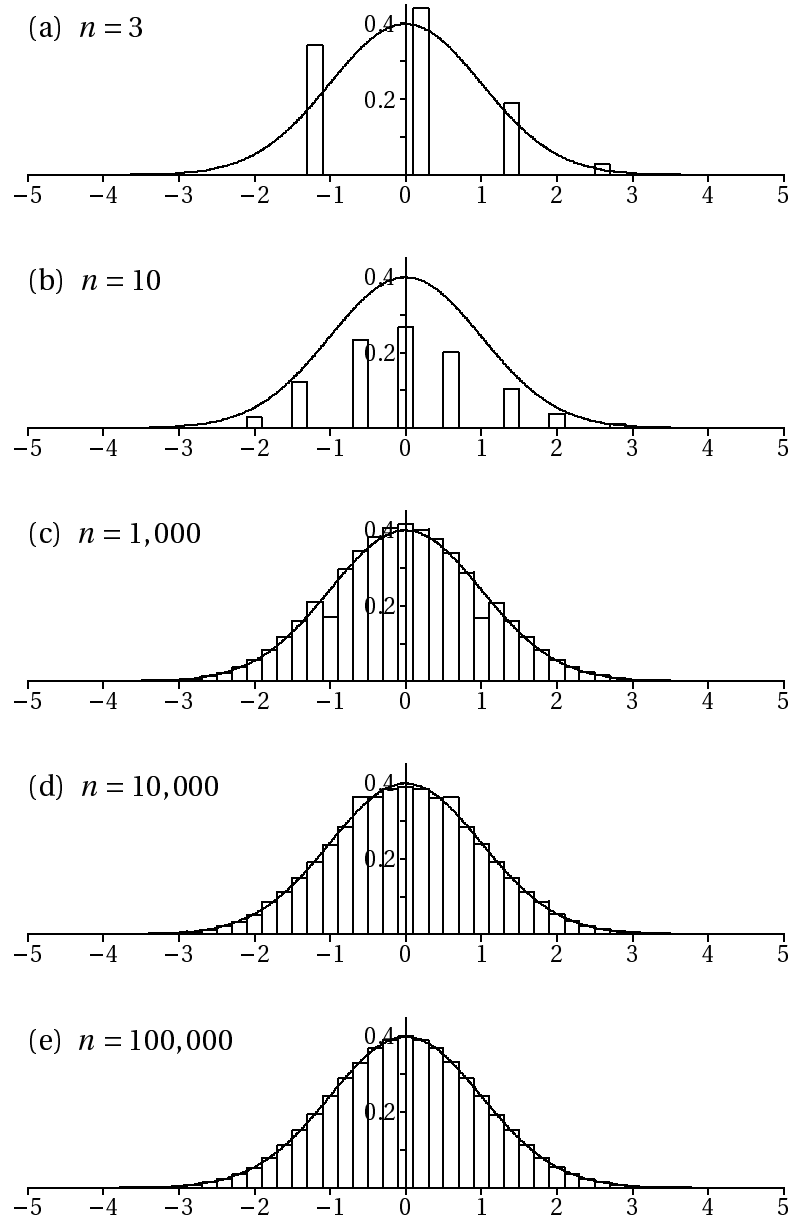


図 2: ベルヌイ分布



る。図 2 の (a)  $n=3$  と (b)  $n=10$  については，離散型分布として描いている（すなわち，それぞれのグラフについて，棒グラフの高さを全部足し合わせれば 1 になる）。 $n$  が大きくなるにつれて，実線とヒストグラムが同じ形になれば（実線が棒グラフのそれぞれの棒の中心を通れば，実線とヒストグラムが同じ形であると言える），中心極限定理が成り立っていると考えることができる。

図 1, 2 を見て分かることを箇条書きに記す。

- 一様分布で  $n=2$  のケース（図 1(a)）は，二等辺三角形の分布になることが数学的に証明できる。グラフからも，それぞれの棒の中心をつなぎ合わせると，0 を中心に二等辺三角形になることが分かる。
- 一様分布のケースもベルヌイ分布のケースも両方とも， $n$  が大きくなるにつれて，それぞれの棒の中心が実線を通り，同じ形の分布になっていることが分かる。したがって，中心極限定理が成り立っていることが確認できる。
- ただし，収束のスピードは元の分布によって大きく異なる。一様分布のケースでは，図 1(c) の  $n=10$  のときには，ヒストグラムと実線は同じように見えることから， $n=10$  で正規近似が可能であると考えてもよいだろう。しかし，ベルヌイ分布のケースでは，図 2(d) の  $n=10,000$  の場合でもヒストグラムが実線で近似できるとは言い難く，図 2(e) の  $n=100,000$  の場合になってようやく正規近似が可能と言えるだろう。

これまでは，第 1 節の (iv) が成り立つことをモンテ・カルロ実験を通して確認した。次に，(iv)' が成り立つかどうかを調べてみる。モンテ・カルロ実験の手順 (1) ~ (3) について，(2) と (3) の代わりに，

(2)'  $t_n = \frac{\bar{x}_n - \mu}{s/\sqrt{n}}$  を計算する。ただし， $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$  とする。

(3)' ステップ (1) と (2)' を  $M$  回繰り返す，得られた  $M$  個のデータ  $t_n^{(1)}, t_n^{(2)}, \dots, t_n^{(M)}$  からヒストグラムを作成する。

を用いて同様の実験を行う。得られた結果は図 3, 4 に描かれている。図からの結果を箇条書きにすると，下記の通りである。

図3: 一様分布 ( $\sigma^2$  の推定)

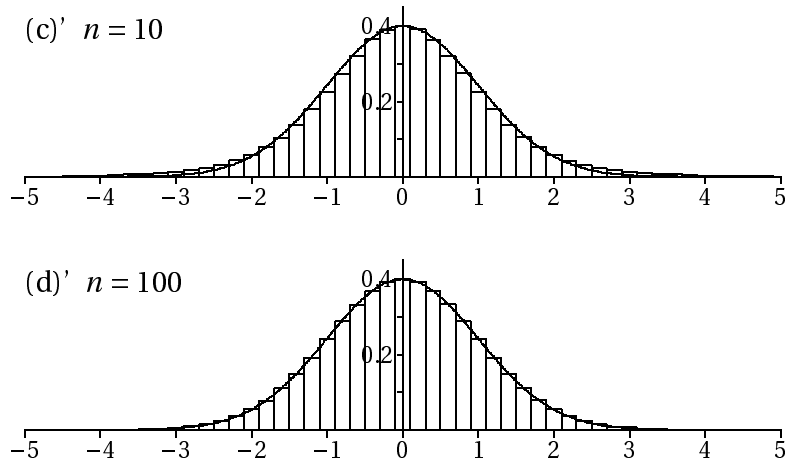
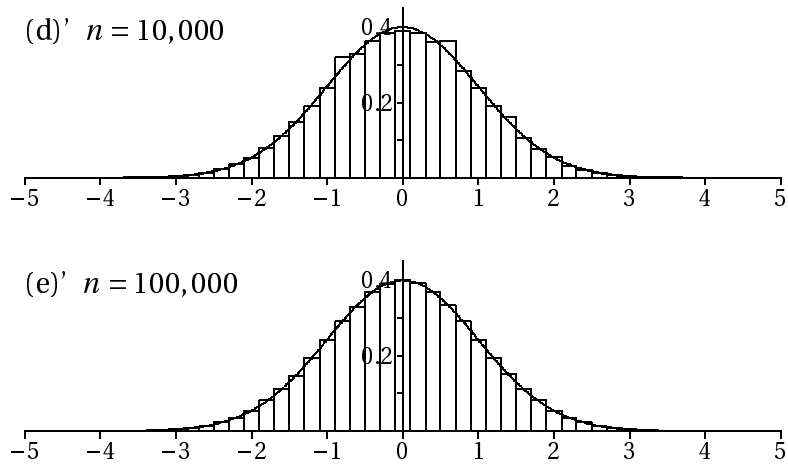


図4: ベルヌイ分布 ( $\sigma^2$  の推定)



- $n = 10$  のケースについて，図 1(c) では正規分布で近似できたが，図 3(c)' ではヒストグラムの裾野が実線の標準正規分布よりも広がっている。しかし， $n = 100$  のケースになると，図 1(d) と図 3(d)' のヒストグラムと実線はほとんど同じである。したがって， $\sigma^2$  の代わりに  $S^2$  を用いると，収束のスピードが遅くなり，正規近似するためにはより大きな  $n$  を必要とすることが分かる。
- 図 2(d)，2(e) と図 4(d)'，(e)' を比べると，それぞれ非常によく似た形のヒストグラムになっていることが分かる。図 1，3 の結果も合わせると， $s^2$  が  $\sigma^2$  と同じ値になればヒストグラムの形は同じになるということから， $n = 100$  以上であれば  $\sigma^2$  も  $s^2$  も値はほとんど同じであると言える。事実， $E(X_i^4)$  が有限であれば， $n$  が大きくなるにつれて， $S^2$  は  $\sigma^2$  に近づくことになる ( $S^2$  は  $\sigma^2$  の一致推定量であるという。一致推定量については，豊田他 (2002) の P.96 の 7.2.2 節参照のこと)。

### 3 おわりに

元の分布が未知の場合であっても，区間推定や検定において標準正規分布が利用される根拠となっているのが中心極限定理である。本稿では，元の分布を一様分布，ベルヌイ分布という簡単によく知られた分布を用いて，中心極限定理が成り立つかどうかをモンテ・カルロ実験によって確認した。すなわち，第 1 節の (i) ~ (iv) は成り立っていることが分かった。ただし，(iv) が成り立つにしても，(i) で選んだ元の分布によっては，(iii) の  $n$  の大きさが大きく異なるということも同時に示した。さらに， $\sigma^2$  を  $S^2$  で置き換えた場合，すなわち，(iv)' が成り立つためにはより大きな  $n$  を必要とするということも示された。