経済統計分析 7
中心極限定理

今日のおはなし.

- 統計的推測 statistical inference へ向けての準備（続き）
  - サンプリングとは
  - 標本分布
  - 大数の法則（LLN: Law of Large Numbers）
  - 中心極限定理（CLT: Central Limit Theorem）

- 今日のタネ
  - 吉田耕作. 2006. 直感的統計学. 日経BP.
  - 中村隆英ほか. 1984. 統計入門. 東大出版会.

なにができるようになりたいか

- ある変数が他の変数に与える効果の大きさの数量化
  - 確率論的な言葉遣いでは,「同時分布の形状を知りたい」
  - あるいは「同時分布の特性値の値を知りたい」
- もし,母集団をすべて調べ上げることができれば…
  - 母集団(population, universe):調査対象の全体
  - 数を数えて同時分布を書いて特性値を計算すればよい
  - 「記述統計」の世界
- 一般には,母集団をすべて調べ上げることはできない
  - 技術的,予算的な制約
  - より「一般的な」母集団を想定している
- 母集団の一部を取り出して,その部分だけから全体を推測
  - サンプル(sample)を抽出して,母数を推測
  - 同時分布の完全な形状は分からない　→特性値(平均・分散)を推測

サンプリング（抽出）

- 母集団からサンプル（標本）を取り出すこと
  - やりかたはいくつか
  - 取り出されたそれぞれの要素や主体（実現値の組合せ）を「観測値 observation」と呼ぶ
  - サンプル sample は観測値の集合（セット）
  - サンプルサイズ = 観測値の数
- ランダムサンプリング（無作為抽出）
  - 母集団のどれが選ばれるかは等確率
  - 「ちゃんと」でたらめになるような方法が使われる：乱数表や電話帳
  - ここではランダムサンプリングを仮定
  - 「ある観測値が選ばれた」ということは，他の観測値が選ばれたかどうか，他の観測値の値がどうか，とは独立
  - 各観測値はi.i.d. (independently and identically distributed)
  - 厳密な意味で成り立つことはそんなに多くない

ランダムでないサンプリング

- 意図的なもの
  - 意図的に「偏り」を作る場合もある
  - 特定の属性を持つものが選ばれにくいことが最初から分かっている場合など
- 意図しないもの，気がつかないもの
  - 統計的推測の結果にも偏り（サンプルセレクション・バイアス）
  - 「ある観測値が選ばれた」ということが，他の観測値が選ばれたかどうか，他の観測値の値がどうか，と相関を持つ可能性
  - 特定の属性を持つものだけが選ばれる／選ばれやすい
  - （例）時系列データ：近接した時点のデータは相関を持ちやすい
  - （例）地域データ：近接した地点のデータは相関を持ちやすい
  - （例）インターネット調査：母集団が「全体」だとすると？
  - （例）平日昼間の電話調査

標本統計量 sample statistics

- 1変数のケース
- 標本平均
  - 記述統計での「算術平均」に対応

$$\bar{Y} = \frac{Y_1 + \ldots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  - 値の総和を観測値数（サンプルサイズsample size）で割ったもの
  - 数学的な扱いの都合上，他の「平均」はさしあたって考えない
- 標本分散，標本標準偏差

$$s_Y^2 = \frac{\left(Y_1 - \bar{Y}\right)^2 + \ldots + \left(Y_n - \bar{Y}\right)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

  - 標本分散の平方根を取ると標本標準偏差を得る
  - 分母が記述統計のときと異なることに注意.

標本統計量は確率変数

- 抽出は確率的
  - どの観測値が選ばれるかはランダム
  - つまり，それぞれの観測値の値は確率変数の実現値
  - もう一度抽出することができたら（仮定法過去），サンプルは異なる
  - マクロデータ・国際データについても「抽出の結果」とみなす

- 標本統計量も確率的
  - 標本統計量の計算のもととなる観測値はランダムに選ばれる
  - だから，標本統計量そのものもランダム
  - つまり，標本統計量は確率変数
  - ただし，通常はサンプルは1セットしか入手できない

標本統計量は確率変数だから…

- 標本分布 sample distribution
  - 標本平均は確率変数なので，確率分布を持つ
  - もしサンプルを何回も抽出することができて，そのたびごとに標本平均を計算することができたら（仮定法過去），その標本平均たちは抽出するたびに異なる値を取るはず
  - 標本分布の平均や分散を考えることができる
  - つまり「平均の平均」や「平均の分散」

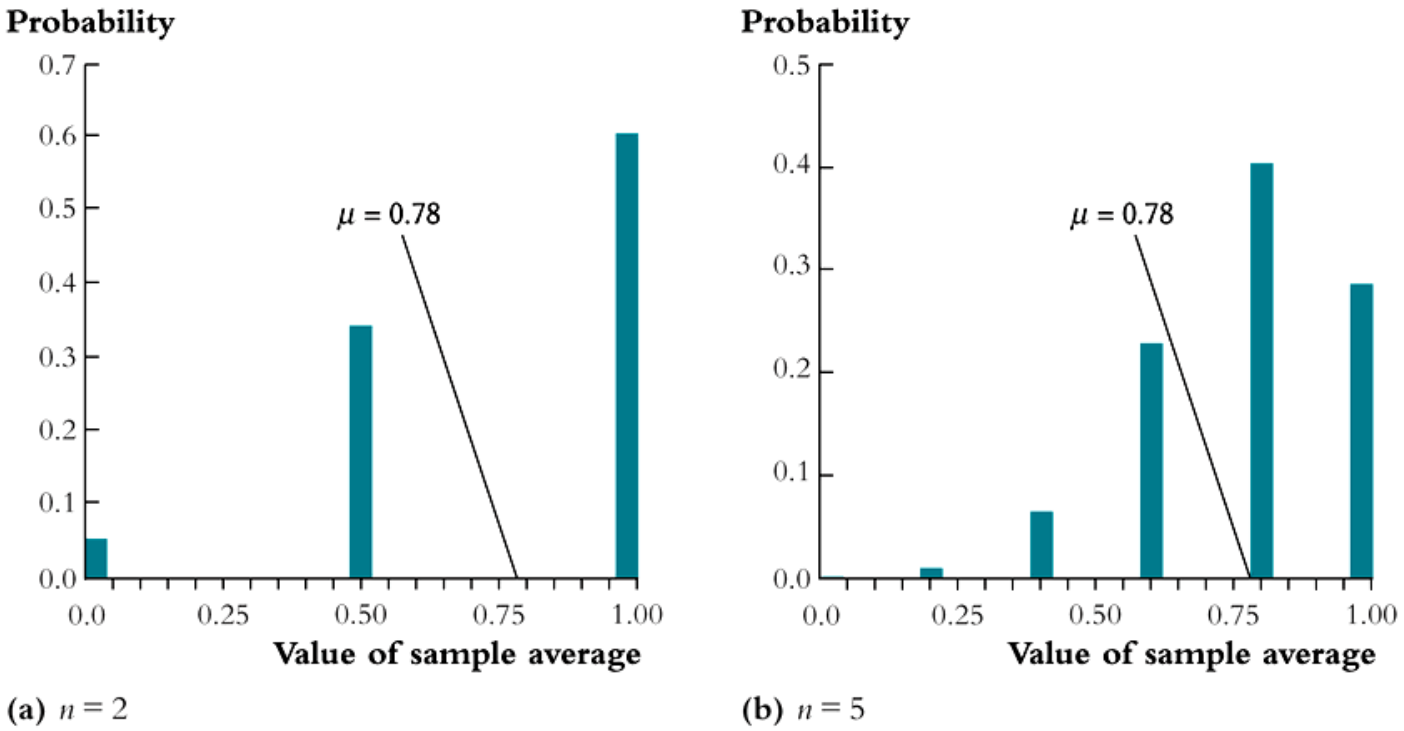
- 標本分布の形状を決めるもの
  - 母集団の分布の形状
  - サンプリングの方法：ここでは無作為抽出を仮定
  - 標本の大きさ

標本分布の例

- コイン投げ
  - 母集団の分布：表が出る（$Y = 1$）確率が 0.78, 裏が出る（$Y = 0$）確率が0.22とする
  - サンプリングの方法：無作為抽出（ふつうのコイン投げ）
- 標本の大きさによって標本平均の分布は異なる
  - 2のとき：確率0.78^2で平均が1. 確率0.22^2で平均が0. それ以外で0.5.
  - 5のとき：取りうる値は, 0, 0.2, 0.4, 0.6, 0.8, 1. たとえば標本平均がゼロになる確率は0.22^5, 1になる確率は0.78^5.
  - 100のとき：取りうる値は, 0, 0.01, …., 0.99, 1. たとえば標本平均がゼロになる確率は0.22^{100}, 1になる確率は0.78^{100}.
  - 100回投げたとしても, 平均値が0.78になるとは限らない

標本分布の例（続き）

FIGURE 2.6 Sampling Distribution of the Sample Average of *n* Bernoulli Random Variables

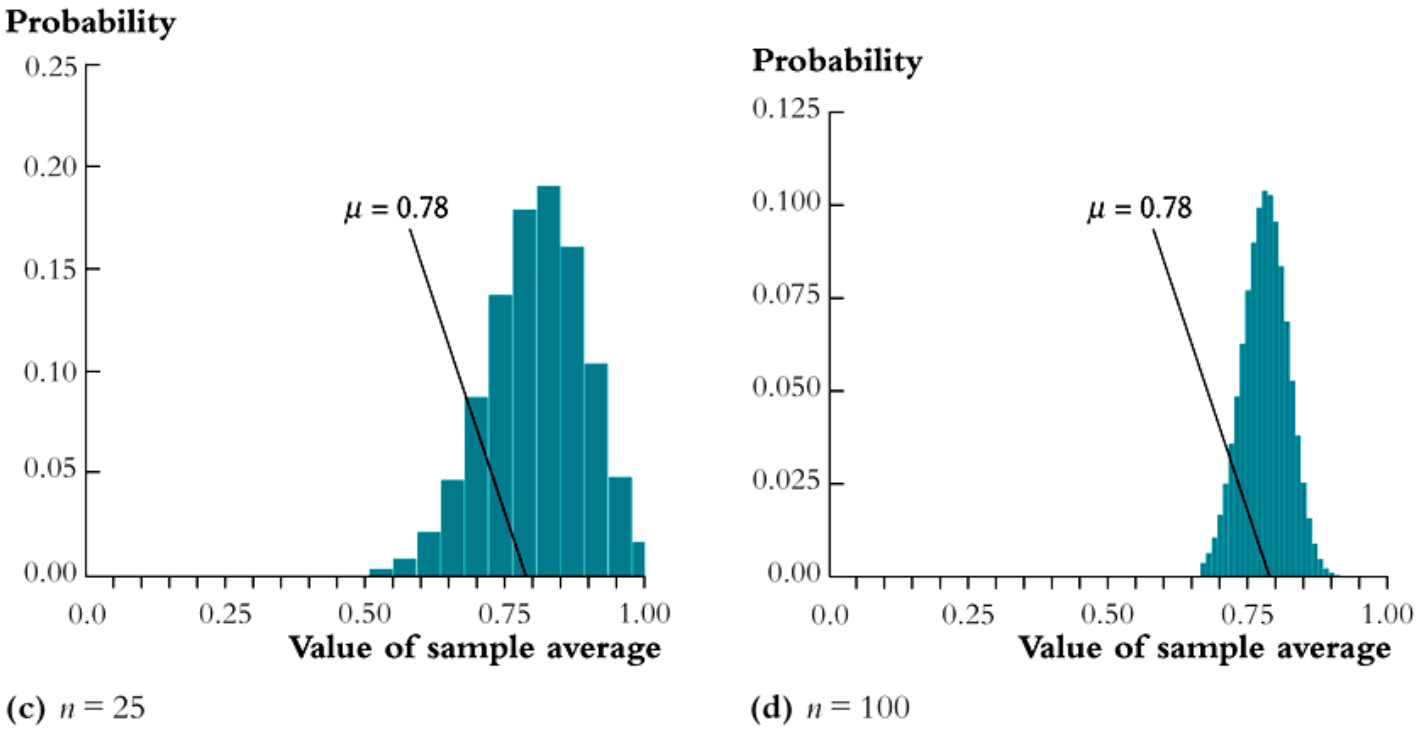


The distributions are the sampling distributions of $\overline{Y}$, the sample average of n independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a fast commute is 78%). The variance of the sampling distribution of $\overline{Y}$ decreases as n gets larger, so the sampling distribution becomes more tightly concentrated around its mean $\mu = 0.78$ as the sample size n increases.

Stock, James H. and Mark W. Watson. 2003. Introduction to Econometrics. International Ed, Prentice Hall. （以下同じ）

標本分布の例（続き）

FIGURE 2.6 Sampling Distribution of the Sample Average of *n* Bernoulli Random Variables



(c) $n = 25$ **(d)** $n = 100$

The distributions are the sampling distributions of $\overline{Y}$, the sample average of n independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a fast commute is 78%). The variance of the sampling distribution of $\overline{Y}$ decreases as n gets larger, so the sampling distribution becomes more tightly concentrated around its mean $\mu = 0.78$ as the sample size n increases.

標本平均と母平均の関係

- 標本平均は母平均に一致しない
  - 確率変数だから.
- 標本平均の分布は簡単に計算できない
  - 母集団の分布の形状
  - サンプリングの方法:ここでは無作為抽出を仮定
  - 標本の大きさ
- 各観測値がi.i.d.のとき, 標本平均の期待値は母平均に一致
  - 標本平均の不偏性（unbiasedness）
  - 各観測値は確率変数とみなされ, その分布は母集団分布そのもの
  - 期待値の線形性から,

$$E\left(\bar{Y}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = E\left(Y\right)$$

標本平均と母平均の関係

- 各観測値がi.i.d.のとき,
- 標本平均の分散は, 標本が大きいほど小さい

$$\mathrm{var}\left(\bar{Y}\right) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\mathrm{var}\left(Y\right)$$

- サンプルサイズが十分に大きければ,
  - 標本平均の分散は非常に小さくなる
  - 母平均の「近く」にないほうがめずらしい
  - 標本平均が「だいたい」母平均とみなしてもよいのでは?

大数の法則 Law of Large Numbers

かなり一般的な条件のもとで，サンプルサイズが十分に大きく，観測値がi.i.d.であれば（ランダムサンプリングであれば），標本平均はかなり高い確率で母集団平均の近くにある

- 平均が同じ確率変数をたくさん集めてその平均を取れば，散らばりが相殺されるので，共通の平均値へと近づく
- 「かなり一般的な条件」は，かなり一般的で，この授業で取り扱うようなデータはすべて条件を満たしているものとみなす
- むしろ問題になるのは「ランダムサンプリング」の仮定
- 少し強い（それでもかなり一般的な）条件のもとで，標本平均についてもっと素敵なことが分かっている

- 「標本平均が母平均に確率収束するconverge in probability」という

中心極限定理 Central Limit Theorem

かなり一般的な条件のもとで，サンプルサイズが十分に大きく，観測値がi.i.d.であれば（ランダムサンプリングであれば），標準化された標本分布は標準正規分布で近似される

- 数式で書くと，

$$\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_Y}{\sigma_Y / n} \xrightarrow{d} N(0,1)$$

- この定理が素敵なのは，結果が母集団の分布に依存しないこと
- 母集団の分布がどのようなものであっても，標本平均について成り立つ
- 「サンプルサイズが十分に大きい」：だいたい100以上が望ましい

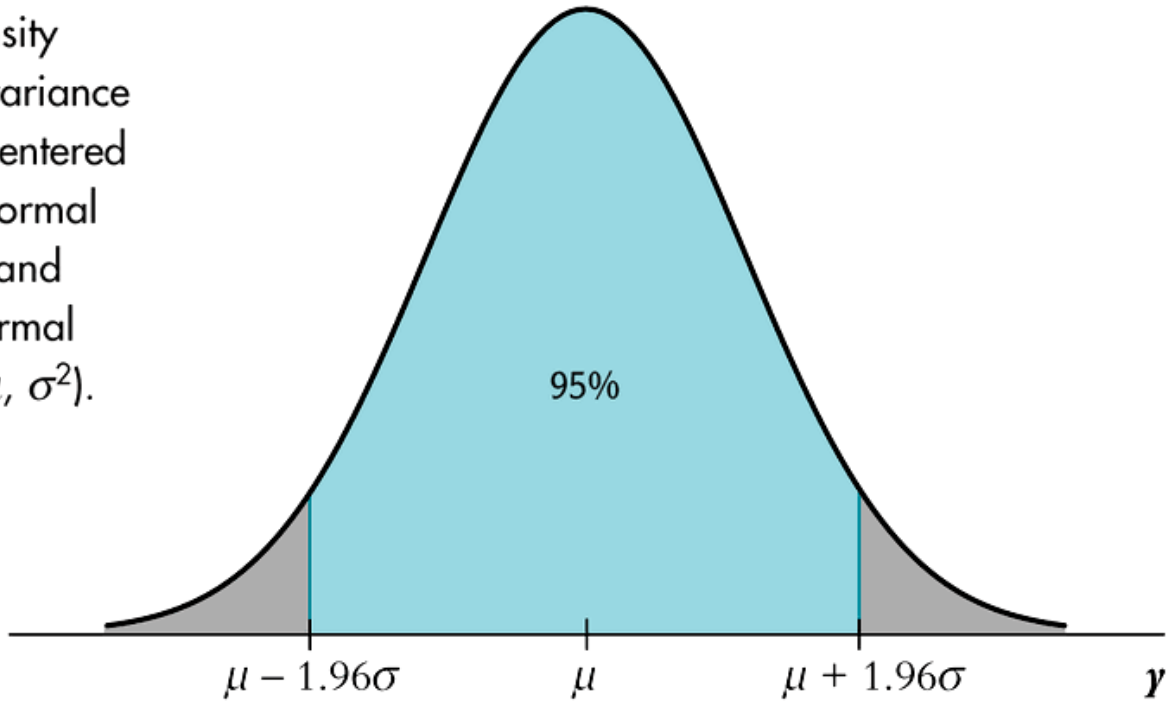
正規分布 Normal distribution

- よくある釣鐘型の分布
  - 多くの自然現象や社会現象が近似できるといわれる
  - 金融データは必ずしもそうでもない？：Black Monday
- 連続変数の左右対称の分布
  - 密度関数が釣鐘型．下の線には接触しない
- 平均と分散だけで形が決まる
  - 一般には，平均と分散が同じでも形状は異なる
  - 平均がμ，分散がσ^2の正規分布を N(μ, σ^2) と表す
  - 平均が0，分散が1の正規分布，N(0, 1) をとくに標準正規分布（standard normal）と呼ぶ
- 多変量正規分布というのもある
  - 例：2変数の同時分布で，2つとも正規分布に従う

正規分布

FIGURE 2.3 The Normal Probability Density

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.



- 平均をはさんで, 標準偏差の1.96倍の幅を両方に取ると, そのあいだの確率が約95%

正規分布の標準化

- 標準化
  - 平均を引いて標準偏差で除して, 新しい確率変数を定義すること
  - 正規分布に従う確率変数を標準化したものは標準正規分布に従う
- 例
  - 平均1, 分散4の確率変数Xを考える
  - Xから1を引いて, 標準偏差2で割って得られる確率変数は標準正規分布に従う
  - だから, 標準正規分布の分布表さえあれば確率分布は求まる

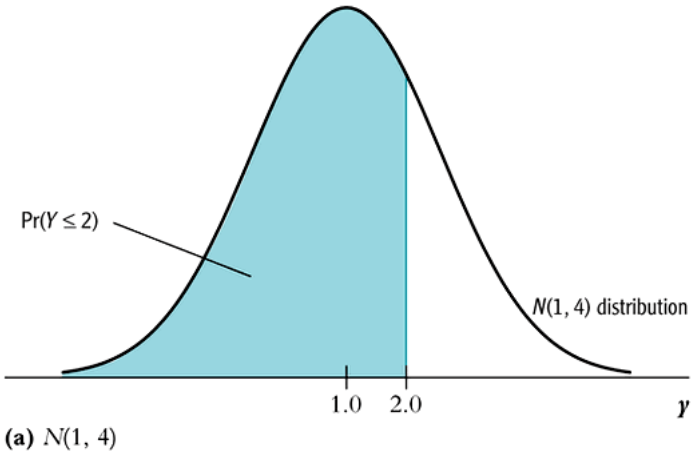
$$\frac{X-1}{\sqrt{4}} \sim N(0,1)$$

$$\Pr(X \leq 2) = \Pr\left(\frac{X-1}{\sqrt{4}} \leq \frac{2-1}{\sqrt{4}}\right) = \Pr\left(\frac{X-1}{\sqrt{4}} \leq \frac{1}{2}\right)$$

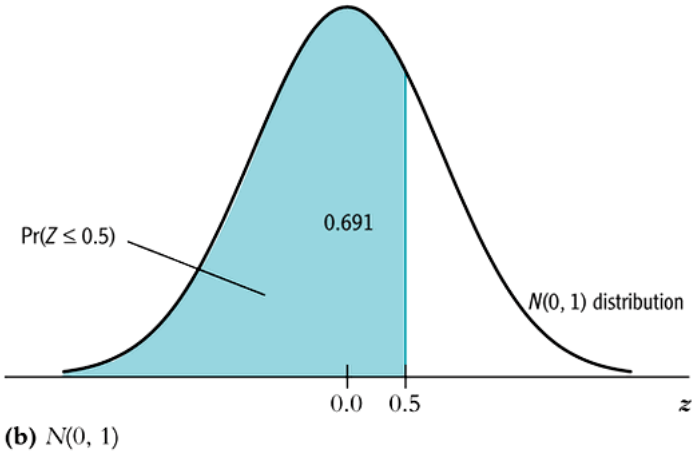
標準化の例

FIGURE 2.4 Calculating the Probability that $Y \le 2$ When Y is Distributed $N(1, 4)$

To calculate $\Pr(Y \le 2)$, standardize Y, then use the standard normal distribution table. Y is standardized by subtracting its mean ($\mu = 1$) and dividing by its standard deviation ($\sigma_Y = 2$). The probability that $Y \le 2$ is shown in Figure 2.4a, and the corresponding probability after standardizing Y is shown in Figure 2.4b. Because the standardized random variable, $\frac{Y-1}{2}$, is a standard normal (Z) random variable, $\Pr(Y \le 2) = \Pr\left(\frac{Y-1}{2} \le \frac{2-1}{2}\right) = \Pr(Z \le 0.5)$. From Appendix Table 1, $\Pr(Z \le 0.5) = 0.691$.



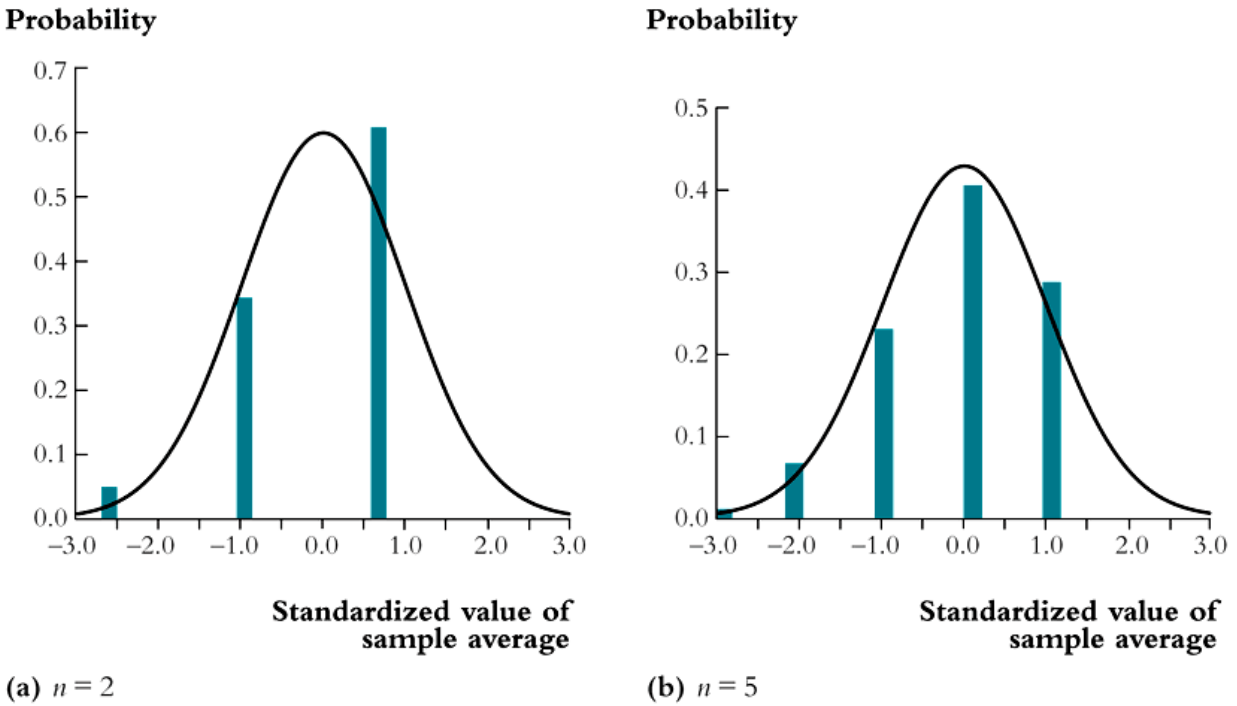
(a) $N(1, 4)$



(b) $N(0, 1)$

中心極限定理の例

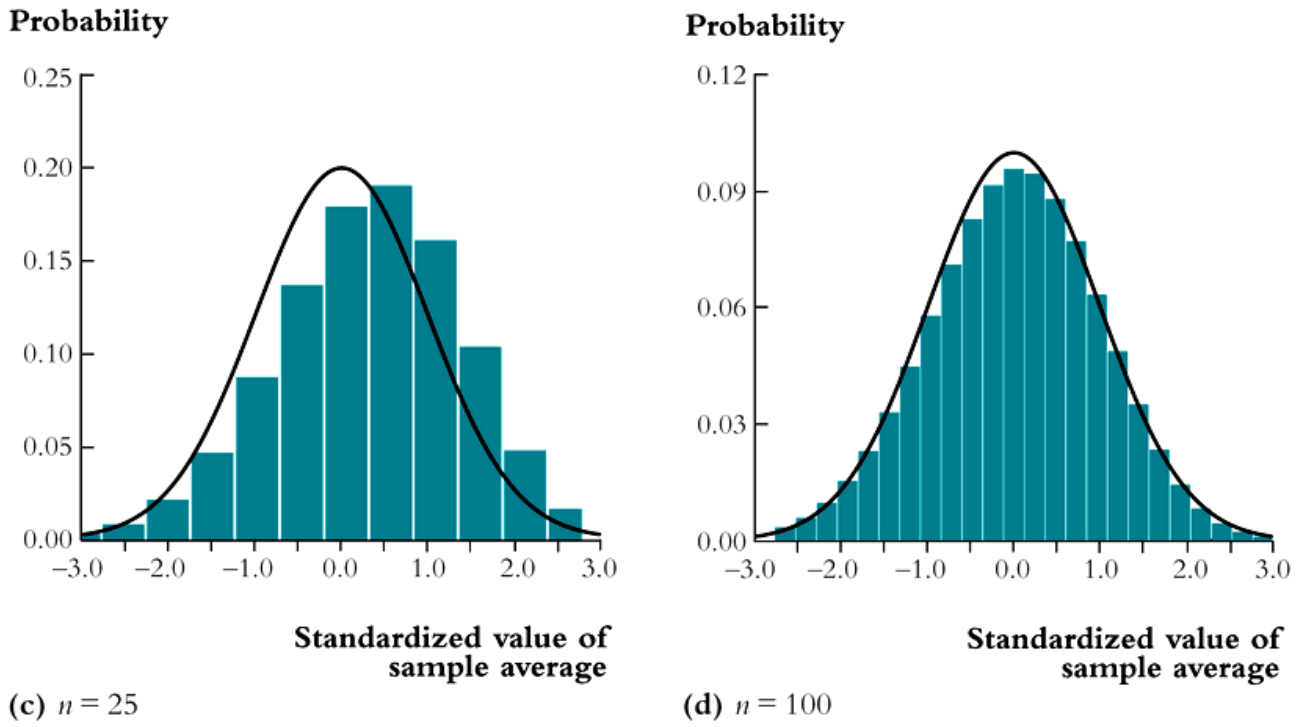
FIGURE 2.7 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = .78$



The sampling distribution of $\overline{Y}$ in Figure 2.6 is plotted here after standardizing $\overline{Y}$. This centers the distributions in Figure 2.6 and magnifies the scale on the horizontal axis by a factor of $\sqrt{n}$. When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem.

中心極限定理の例（続き）

FIGURE 2.7 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = .78$



The sampling distribution of $\overline{Y}$ in Figure 2.6 is plotted here after standardizing $\overline{Y}$. This centers the distributions in Figure 2.6 and magnifies the scale on the horizontal axis by a factor of $\sqrt{n}$. When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem.