確率・統計の基礎

別所俊一郎

慶應義塾大学経済学部

2012 年 4 月

計量経済学とは?

- さまざまな見方があるが…
  - 経済理論の検証、現実への適用
  - 経済関連の指標の予測
  - 数量的な示唆の獲得：理論的に興味深くても…
  - 経済データを分析するために経済理論と統計手法を使う，science and art.
- データと情報
  - データから情報を得る手段の一つが計量経済学
  - Cross-section, Time series, panel data
- Statistical v. identical
- 「1,000 年に 1 度」v.「1,000 年おき」

この講義で扱うこと

「標準的な」計量経済学の基礎

- 確率・統計の基本
- 回帰分析（regression）の基礎としての OLS
- 変数の内生性と操作変数法
- （その他のトピック：実験，パネル，質的データ）
- 確率・統計・計量経済学については Stock and Watson に拠る
  - 最初から大標本理論を使う：自由度修正は気にしない
  - 無作為抽出の仮定：説明変数も確率変数
  - 分散不均一を仮定：DW も扱わない

この講義で扱わないこと

- データの作り方
  - 実査，アンケート，実験の方法
- 計量経済学の「進んだ」トピック
  - 計量経済理論
  - 多変量解析：主成分分析，因子分析，クラスター分析
  - 時系列分析
  - ノンパラメトリック分析

なぜ，計量経済学か

経済理論は定量的関係を示さない：

- 経済理論は、経済変数間の定性的な関係を示唆
- 変数間の定量的な関係を把握するには統計的処理が必要
- 経済理論の裏付けをもった統計的処理の解釈
- 費用便益分析への応用

ただし……

- 定量的な関係がわかったとしても「重要性」の判断は政策立案者のしごと
- 「統計的有意性」は必ずしも政策的重要性を意味しない
- 定量的関係が把握しにくいことも多くある
- 得られる定量的関係にはつねに不確実性：データや手法の限界

因果（causality）と相関（correlation）

- 因果関係の抽出が目的：理想的な実験を想定しよう
- 実験群（treatment group）と対照群（control group）
- 予測には必ずしも因果関係は必要ない

統計の種類

- 統計調査／世論調査，官庁統計／民間統計
- 第一義統計／第二義統計・業務統計
- 民間統計
  - 業界団体のものや，マーケティング目的のもの
  - 回収率，調査設計に問題があることも
- 一次統計／二次統計
  - 一次統計：調査から直接得られる統計
  - 二次統計（加工統計）：一次統計に計算を施して得られる．指数や総合推計．物価指数や国民経済計算が典型．

調査の種類

- 対象
  - 全数調査（センサス）：標本誤差がない，多角的分析が可能
  - 標本調査：調査費用が少ない，非標本誤差が少ない
- 抽出方法
  - 有意抽出：調査の規格者が標本を選定
  - 無作為抽出：偶然性に基づいて標本を選定
- 無作為抽出方法
  - 単純無作為抽出：乱数のみに基く
  - 系統抽出：等間隔抽出，集落抽出，2 段階抽出
  - 不等確率抽出／確率比例抽出

調査・集計方法

- 調査法
  - 回収率・虚偽回答・費用の面などで異なる
  - 固定質問紙法
  - 調査員調査法：面接調査（他計申告）法／留置調査（自計申告）法
  - 郵送調査法：往復メール／メールアウト／メールバック
  - 電話調査法，ファックス調査法，インターネット調査法
  - 会場調査法（座談会など），観察調査法（物価など）
- 集計
  - 内容検査：事実確認の照会，修正補記
  - 自由回答への符号付け：もやしの工場栽培は？
  - データ入力：手入力，OMR，OCR
  - データチェック：論理チェックなど

確率・統計がなぜ必要か

- 実証分析にはさまざまな「でたらめさ」が付随
  - 標本抽出（サンプリング）
  - 考慮できない／捨象された要因たち
  - （決定論的立場には立たない）
- 確率論
  - 「でたらめさ」「リスク」を数量化して扱う数学的手法
  - 回帰分析や計量経済学に必要な部分だけ
- 母集団と標本の区別

確率論の基礎用語

根源事象 outcome： 相互に排他的な，潜在的に起こりうる結果．どれもが同じように起きやすいということではない

標本空間： 起きうる結果すべてから成る集合

事象 event： 標本空間の部分集合．根源事象の組み合わせで表現できる

確率： 事象が起きやすさ．根源事象の確率の和は 1.

確率変数： 事象の数学的表現。標本空間から実数への関数として定義される。(確率変数を大文字で、実現値を小文字で表す)

離散確率変数のばあい

確率分布： 確率変数が取りうるすべての値（根源事象）と，対応する確率のリスト．確率をすべて足すと 1.

各事象の確率： 確率分布から求められる

累積確率分布 c.d.f.： 確率変数がある値より小さい値をとる確率．累積分布関数とも（Table 2.1）

- ベルヌーイ分布
  - 2 値変数のとる分布

$$G = \begin{cases} 1 & \text{確率 } p \\ 0 & \text{確率 } 1-p \end{cases} \tag{2.1}$$

  - ベルヌーイ家は天才を輩出した家系として知られる

TABLE 2.1 Probability of Your Computer Crashing *M* Times

	Outcome (number of crashes)				
	0	1	2	3	4
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

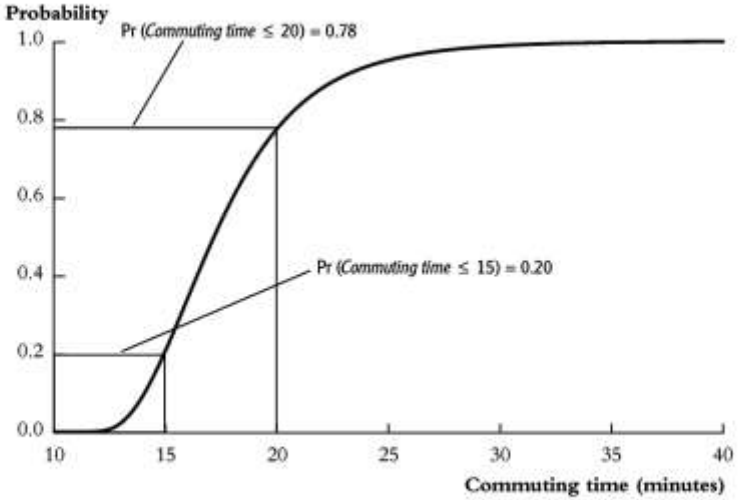
連続確率変数のばあい

累積確率分布： 離散のばあいと同じ．確率変数がある値より小さい値をとる確率

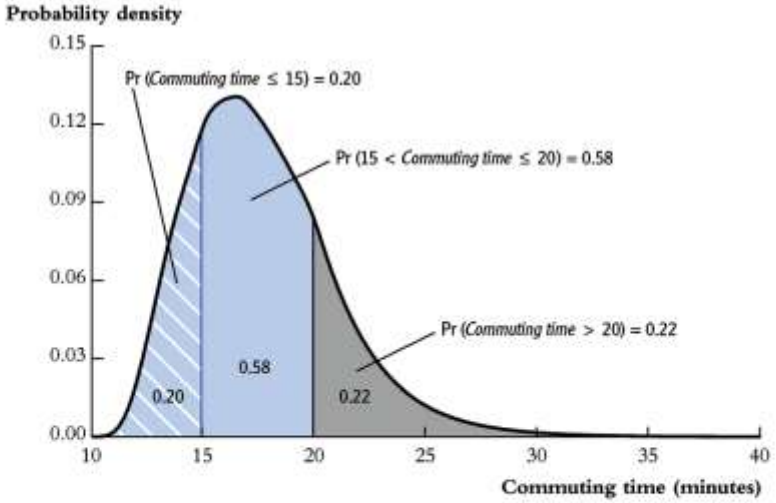
確率密度関数 p.d.f.： 「下」の面積が確率に等しくなるような関数.

- 累積分布を微分したもの全区間を積分すると 1
- 離散変数と同じようには定義できないことに注意
- 例：Figure 2.2.
- 連続変数を扱うケースのほうが多い。

FIGURE 2.2 Cumulative Distribution and Probability Density Functions of Commuting Time



(a) Cumulative distribution function of commuting time



(b) Probability density function of commuting time

Figure 2.2a shows the cumulative probability distribution (or c.d.f.) of commuting times. The probability that a commuting time is less than 15 minutes is 0.20 (or 20%), and the probability that it is less than 20 minutes is 0.78 (78%). Figure 2.2b shows the probability density function (or p.d.f.) of commuting times. Probabilities are given by areas under the p.d.f. The probability that a commuting time is between 15 and 20 minutes is 0.58 (58%), and is given by the area under the curve between 15 and 20 minutes.

期待値・モーメント

- 期待値 $E[Y], \mu_Y$
  - 離散変数のばあいは確率を重みとする加重平均

$$E[Y] = \sum_{i=1}^{k} y_i p_i \tag{2.4}$$

  - 連続変数のばあいも “加重平均” みたいなもの
- モーメント（積率）：累乗の期待値

$$r \text{ 次のモーメント} \equiv E[Y^r] = \sum_{i=1}^{k} y_i^r p_i \tag{2.6}$$

  - 3 次モーメントは歪度 $E[(Y-\mu)^3]/\sigma_Y^3$
  - 4 次モーメントは尖度 $E[(Y-\mu)^4]/\sigma_Y^4$

分散・標準偏差

- 分散：確率分布の「広がり」を表す

$$\text{分散 } \mathrm{var}(Y) \equiv E[(Y-\mu_Y)^2] = \sum_{i=1}^{k}(y_i-\mu_Y)^2 p_i$$

- 標準偏差：分散の平方根

$$\text{標準偏差 } \sigma_Y \equiv \sqrt{\mathrm{var}(Y)} = \sqrt{E[(Y-\mu_Y)^2]}$$

- ベルヌーイ分布のばあい

$$\mathrm{var}(G) = \sigma_G^2 = (0-p)^2(1-p) + (1-p)^2 p = p(1-p) \qquad (2.7)$$

期待値の線形性

確率変数 X，定数 a, b に対して，$Y = a + bX$ とするとき，

$$E[Y] = a + bE[X], \mathrm{var}(Y) = b^2\mathrm{var}(X)$$

[証明]

同時分布・周辺分布

- 同時分布 $\Pr(X = x, Y = y)$
  - 確率変数 X が x の値をとり，かつ，確率変数 Y が y の値をとる確率
  - 全ての組合せについて確率を足すと 1（Table 2.2.）
- 周辺分布 $\Pr(X)$
  - ひとつの確率変数だけに注目したときの分布
  - 同時分布の和として表現でき，

$$\Pr(Y = y) = \sum_{i=1}^{I} \Pr(X = x_i, Y = y) \tag{2.16}$$

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long Commute ($Y = 0$)	0.15	0.07	0.22
Short Commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

条件付き分布

- 条件付き確率 $\Pr(Y = y|X = x)$
  - 確率変数 X の実現値を所与にしたときに確率変数 Y が y の値をとる確率
  - 確率変数 X の実現値の関数となる（Table 2.3.）

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \tag{2.17}$$

- 条件付き期待値，分散

$$E[Y|X = x] = \sum_{i=1}^{k} y_i \Pr(Y = y_i|X = x) \tag{2.18}$$

$$\mathrm{var}(Y|X = x) = \sum_{i=1}^{k} [y_i - E(Y|X = x)]^2 \Pr(Y = y_i|X = x) \tag{2.21}$$

期待値のくりかえしの法則

確率変数 Y の期待値は，確率変数 X の実現値 x で条件付けられた Y の条件付き期待値を X の分布で加重平均したものに等しい．すなわち

$$E[Y] = E[E[Y|X]]$$

[証明]

独立

確率変数 X と Y が互いに独立に分布しているとは

- X の実現値についての情報が Y の条件付き分布についての情報とならない
- 実現値 x, y の全てに対して

$$\Pr(Y = y|X = x) = Pr(Y = y) \quad \forall x, y \tag{2.22}$$

- 条件付き分布の定義式より同時分布について

$$\Pr(Y = y, X = x) = Pr(X = x)Pr(Y = y) \quad \forall x, y \tag{2.23}$$

共分散、相関

- 共分散：2 つの変数が似たような動きをするかどうかを示す．同じ方向に動くときに正の値．

$$\sigma_{XY} \equiv \mathrm{cov}(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{2.24}$$

- 相関係数：共分散の単位をそろえたもの．ゼロであるとき，2 つの確率変数は「無相関である」という
- 「独立」ならば「無相関」

$$\mathrm{corr}(X, Y) \equiv \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.25}$$

- 相関係数は 1 を越えない
- 条件付き平均が無条件平均に等しいとき，共分散はゼロ．逆は必ずしも成り立たない．

確率変数の線形結合

2 つの確率変数 X, Y の和の期待値，和の分散は,

$$E[X+Y] = E[X] + E[Y] \tag{2.28}$$

$$\mathrm{var}(X+Y) = \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}(X,Y) \tag{2.36}$$

[証明]

いくつかの確率分布

- 表現の約束
  - 確率変数 X の分布関数が F で表現されるとき,「確率変数 X は分布 F に従う」といい、「$X \sim F$」と表す
  - 分布関数を大文字（F)、密度関数を小文字（f）で表す
  - 確率変数 X の密度関数を $f_X(x)$ と書くこともある
- よく使う確率分布
  - 検定で用いる
  - 正規分布（normal)，カイ 2 乗分布（Chi-squared)，F 分布，t 分布
  - もちろん，他にもいろいろな名前の分布がある

正規分布 normal distribution

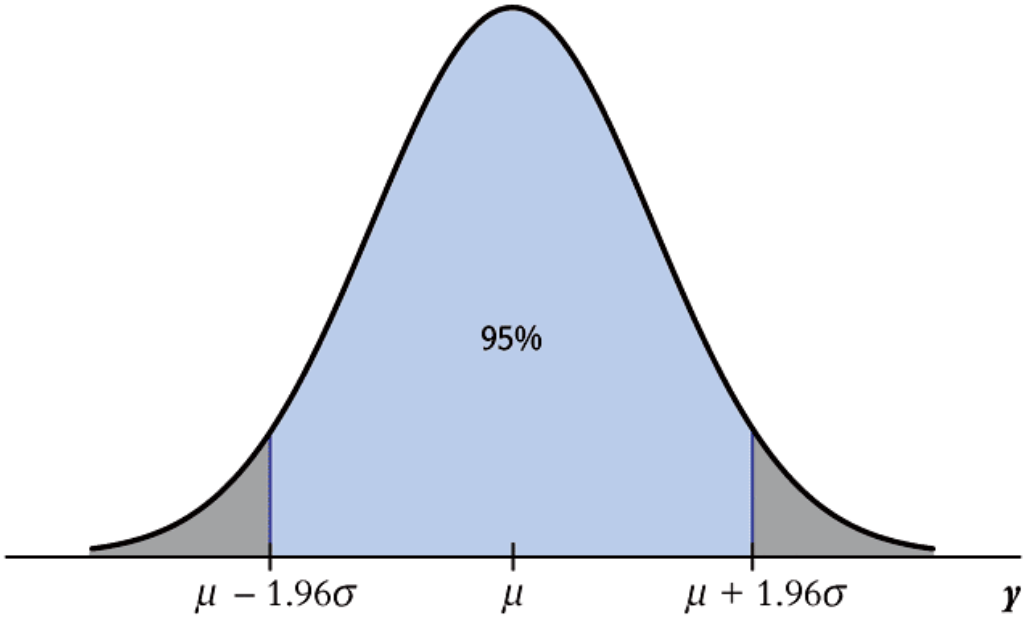
- 釣鐘状の連続分布
- 数式で書くとやや複雑

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- 平均 μ，分散 σ^2 とすると，区間 $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ に 95%が含まれる
- 3 次以上のモーメントが定数になり，平均と分散だけで分布が決まる
- 確率の分野では非常によく用いられる分布

FIGURE 2.5 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 $N(\mu, \sigma^2)$.



正規分布のお約束

- 平均 μ，分散 σ^2 の正規分布を $N(\mu, \sigma^2)$ と表す
- 平均 0，分散 1 の正規分布 $N(0,1)$ のことをとくに「標準正規分布 standard normal」とよび Z で表す．$Z \sim N(0,1)$
- 標準正規分布の分布関数を $\Phi(x)$，密度関数を $\phi(x)$ で表す．定数 c に対して，$\Phi(c) = \Pr(Z < c)$
- 正規分布に従う確率変数から，その平均を引いてその標準偏差で除して新しい確率変数を定義することを「標準化 standardize」と呼ぶ．たとえば $X \sim N(1,4)$ のとき

$$\frac{X-1}{\sqrt{4}} = \frac{1}{2}(X-1) \sim N(0,1)$$

$$\Pr(X \leq 2) = \Pr\left(\frac{1}{2}(X-1) \leq \frac{1}{2}\right) = \Phi\left(\frac{1}{2}\right) \tag{2.41}$$

多変量正規分布

- いくつかの確率変数の同時分布に対して多変量正規分布 (multivariate normal) を考える
- n 変量正規分布は平均の組合せ (n 個) と分散・共分散の組合せ (n^2 個) によって定義
- 確率変数 (X, Y) が 2 変数正規分布に従うとき，定数 a, b に対してその線形結合もまた正規分布に従う

$$aX + bY \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right) \tag{2.42}$$

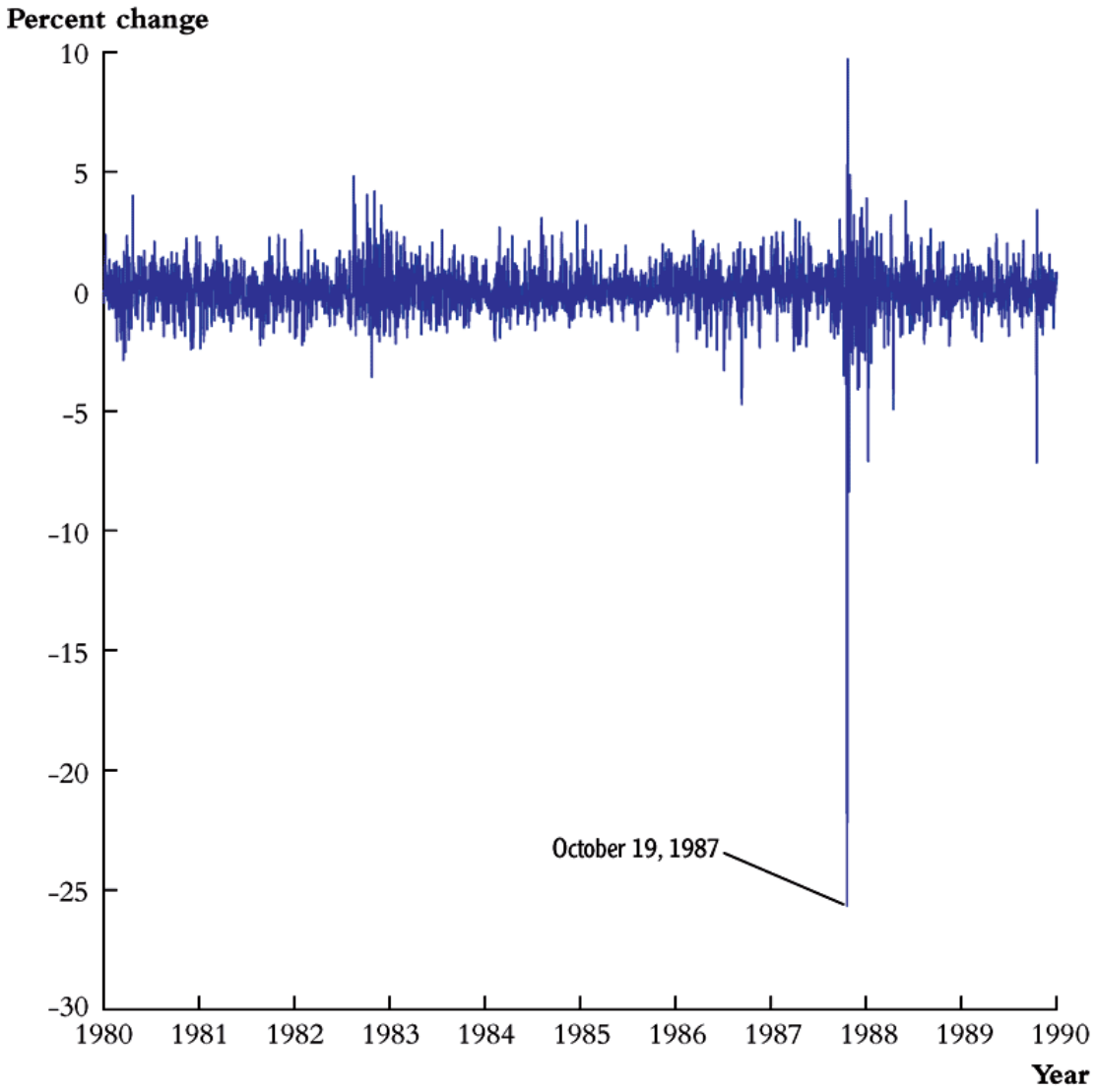
- 一般に n 変量正規分布に従う n 個の確率変数の線形結合もまた正規分布に従う

多変量正規分布

- 一般に，n 変量正規分布に従う n 個の確率変数について，各確率変数の周辺分布もまた正規分布
- n 変量正規分布に従う n 個の確率変数のうち，共分散がゼロであるような 2 個の確率変数は互いに独立
  - 互いに独立である確率変数の共分散は，同時分布の形状に関わらず，ゼロ．
  - 一般には，共分散がゼロであるからといって 2 個の確率変数が独立であるとは限らない
  - この性質は，正規分布では 3 次以上のモーメントが定数であることによる

FIGURE 2.7 Daily Percentage Changes in the Dow Jones Industrial Average in the 1980s

During the 1980s, the average percentage daily change of "the Dow" index was 0.05% and its standard deviation was 1.16%. On October 19, 1987—"Black Monday"—the index fell 25.6%, or more than 22 standard deviations.



カイ 2 乗分布, Chi squared

- 仮説検定で用いられることの多い分布
- m 個の独立した標準正規分布に従う確率変数の 2 乗和の分布を自由度 m のカイ 2 乗分布とよぶ
- 互いに独立な 3 つの確率変数が $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, $Z_3 \sim N(0,1)$ であるとき

$$Z_1^2 + Z_2^2 + Z_3^2 \sim \chi_3^2$$

- 分布表については appendix 参照

F 分布

自由度 m の F 分布

- 自由度 m の χ^2 分布に従う確率変数を m で除して得られる確率変数の分布
- 互いに独立な 3 つの確率変数が $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, $Z_3 \sim N(0,1)$ であるとき

$$\left(Z_1^2 + Z_2^2 + Z_3^2\right)/3 \sim F_{3,\infty}$$

t 分布

自由度 m の（スチューデント）t 分布

- 標準正規分布に従う確率変数を，それとは独立に自由度 m の χ^2 分布に従う確率変数を m で除したもので割って得られる確率変数の分布
- 互いに独立な 2 つの確率変数が $Z \sim N(0,1)$, $W \sim \chi^2_m$ であるとき

$$\frac{Z}{\sqrt{W/m}} \sim t_m$$

- 自由度 m が十分に大きければ t 分布は正規分布に近い．極限では正規分布に一致 $t_\infty = N(0,1)$

無作為抽出 random sampling

母集団と標本

- 母集団（population）：検討の対象とする主体全体
- 標本（sample）：母集団から抽出された観測値（observation）の集合
- 悉皆調査：母集団をすべて調査したもの．センサス．
- 母集団として何を考えているかは重要
- 手許にある標本から母集団の特性を推し量ることが計量経済分析の目的のひとつ

無作為抽出 random sampling

標本の抜き出し方は確率的

- ほとんどのケースで，母集団から標本をどのように抽出（サンプリング）するかは確率的
- 無作為抽出では，ある抽出は次に抽出されるものについての情報をまったくもたない
- 時系列データや地域データなどでは無作為抽出とは考えにくい：系列相関

標本統計量 sample statistics

標本の選び方が確率的であるから，標本 の特性を表す指標（平均など）は確率変数

- 無作為抽出の場合は，各観測値は互いに独立な確率変数
- 観測値から計算される平均や分散もまた確率変数
- 標本統計量：標本から計算される平均など
- 標本から計算される平均値は確率変数だから，標本平均は確率分布を持ち，「平均の分散」を考えることができる
- 標本平均は母平均（母集団の平均）とは一致しないから，標本平均を「評価」するためには標本平均の分散や標準偏差を知る必要

i.i.d.

標本抽出を行うとき，各観測値は確率変数だから，それらの多変量同時分布）を考えることができる

- 各観測値が同じ母集団から抽出され，各観測値の周辺分布が同じであるとき，それら観測値は同じ分布に従う identically distributed という
- 無作為抽出などにより，各観測値の分布が互いに独立であるとき，それらは独立に分布している independently distributed という
- 各観測値が独立に同じ分布に従う independently and identically distributed とき，略して i.i.d. である（独立同一分布）という

標本平均 sample average (mean)

- 標本平均は各観測値の平均で定義される

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.43}$$

- 各観測値は確率変数だから，標本平均も確率変数．標本が変われば，母集団が同じでも，標本平均も変わる
- 標本平均の平均や，標本平均の分散を考えることができる
  - 抽出を繰り返し行い，そのたびごとに標本平均を計算したときの標本平均の平均や分散のこと
  - 標本平均の分布は，標本の大きさにも依存

標本平均の分布

各観測値 $(Y_1, Y_2, \cdots, Y_n)$ が i.i.d. で，母集団の平均と分散を μ_Y, σ_Y^2 とする

- たとえば観測値数が 2 のときの標本平均の期待値は $E[\frac{1}{2}(Y_1 + Y_2)]$
- 一般に標本平均の平均は

$$E[\overline{Y}] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \mu_Y \tag{2.44}$$

であり，母集団の平均に一致

標本平均の分散

各観測値 $(Y_1, Y_2, \cdots, Y_n)$ が i.i.d. で，母集団の平均と分散を μ_Y, σ_Y^2 とする

- たとえば観測値数が 2 のときの標本平均の期待値は $\mathrm{var}(Y_1 + Y_2) = 2\sigma_Y^2$
- 一般に標本平均の分散は

$$\mathrm{var}(\overline{Y}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{\sigma_Y^2}{n} \tag{2.45}$$

であり，標本が大きいほど小さい

標本統計量の大標本近似

- 標本統計量は確率変数になるから，その分布を考えることができる
  - 手許にある標本から母集団の特性を統計的に推測するとき，得られた標本統計量が持つはずの性質・分布を知っておくことは有益
  - 一般には，標本統計量の分布は，母集団の分布，標本の大きさ，サンプリングの方法に依存
  - しかし，ランダムサンプリングで標本の大きさが十分に大きければ，標本統計量の分布はあるいていど推測できる
- 大数の法則と中心極限定理

標本分布の扱い方

Exact アプローチ

- サンプルサイズがどのようなものであっても一般的に成り立つような公式を求める
- exact distribution, finite-sample distribution（小標本分布）と呼ばれる
- 求めにくいことがおおい：母集団の分布が正規分布なら標本平均の分布は正規分布に従う

標本分布の扱い方

Approximate アプローチ

- サンプルサイズが十分に大きい（$n \to \infty$）のときの標本統計量の分布を近似的に用いる
- Asymptotic distribution（漸近分布）と呼ぶ
- 「十分に大きい」とは?: $n = 30$ くらいでもよい
- 「大数の法則」と「中心極限定理」を用いる
- Exact アプローチよりシンプルで、母集団の分布に依存しないので使いやすい

大数の法則 Law of Large Numbers

平均と分散が有限で，サンプルサイズ n が十分に大きく，観測値が i.i.d. 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

- $n \to \infty$ のとき，標本平均 $\overline{Y}$ が真の平均 μ の近くにある確率が 1 に近づく
- 平均が同じ確率変数を多く集めてその平均を取ると，ちらばりが相殺されて共通の平均値へ近づく
- 「確率収束 convergence in probability」と呼び，標本平均が真の平均に確率収束するとき，その標本平均は「一致性 consistency」を持つ、という

$$\overline{Y} \xrightarrow{p} \mu_Y$$

中心極限定理 Central Limit Theorem

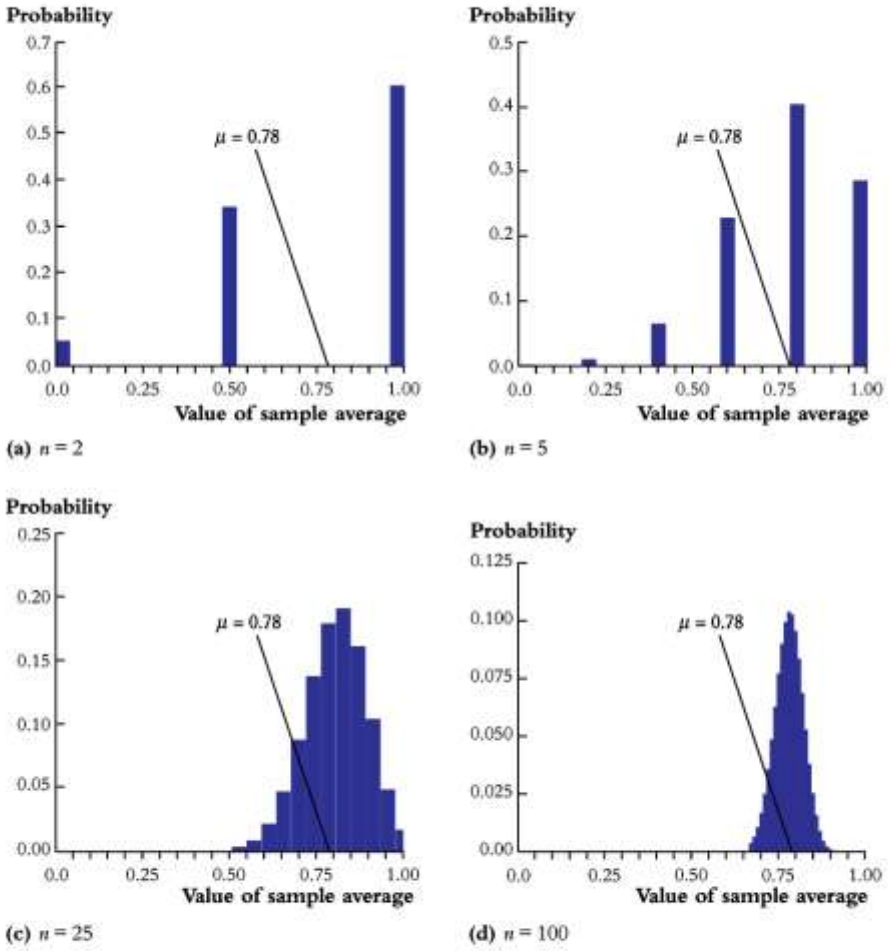
かなり一般的な条件の下で，サンプルサイズ n が十分に大きく，観測値が i.i.d. であれば，標本平均の分布は正規分布で近似される

- 「分布収束 convergence in distribution」とよぶ
- 標準化された標本分布は「漸近的に」標準正規分布に従う

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

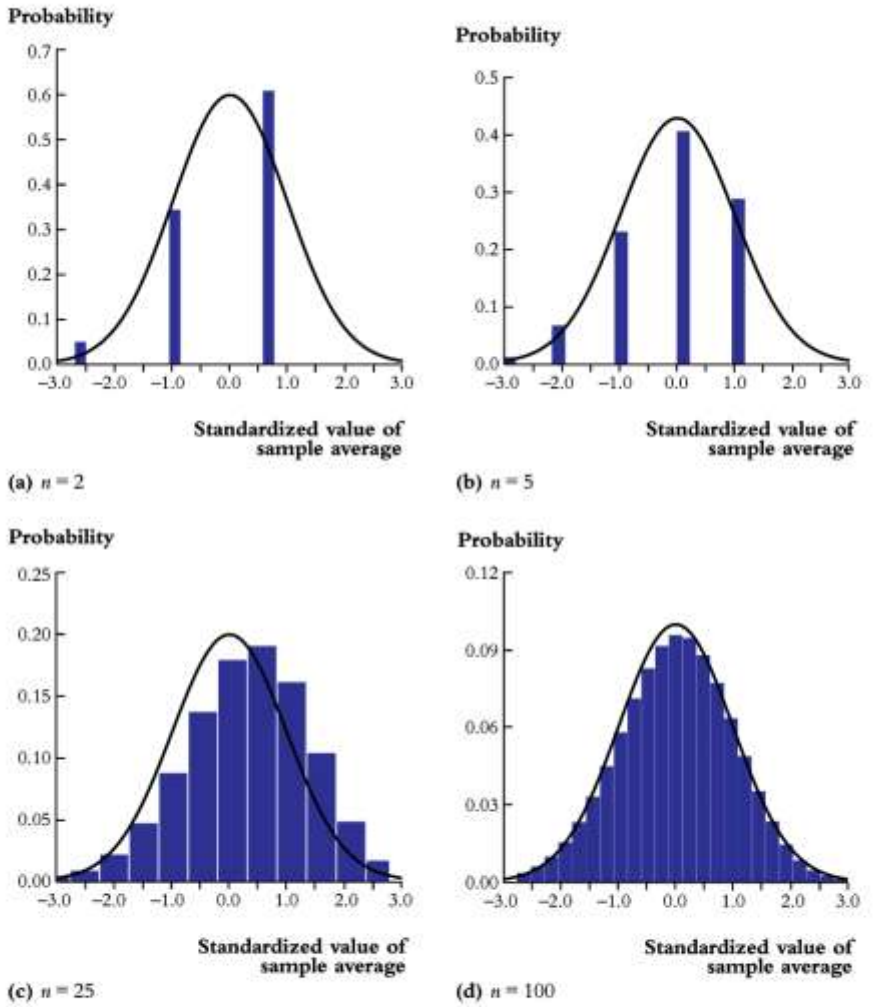
- 母集団の分布が正規分布なら、exact に正規分布になる
  - 母集団がどんな分布でも，漸近的には approximately/asymptotically 正規分布に従う
- そうでないばあいには，$n = 100$ くらいにならないと近似が不十分（Fig. 2.9, 2.10）

FIGURE 2.8 Sampling Distribution of the Sample Average of *n* Bernoulli Random Variables



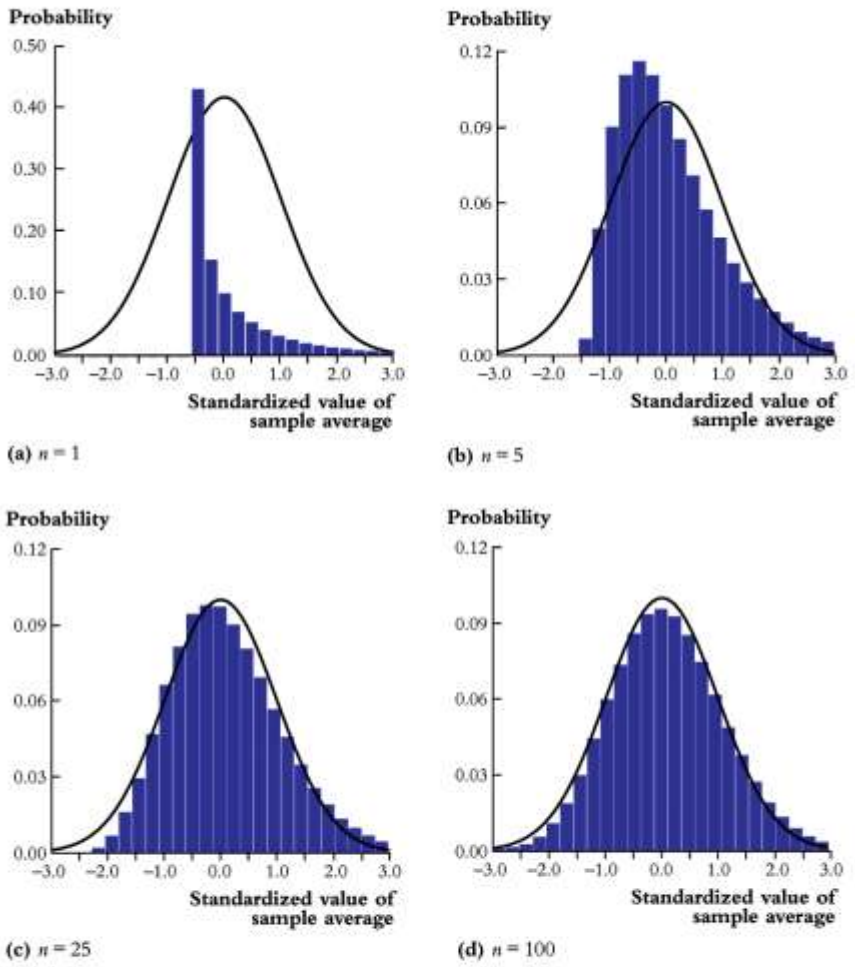
The distributions are the sampling distributions of $\overline{Y}$, the sample average of *n* independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a short commute is 78%). The variance of the sampling distribution of $\overline{Y}$ decreases as *n* gets larger, so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

FIGURE 2.9 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = 0.78$



The sampling distribution of $\overline{Y}$ in Figure 2.8 is plotted here after standardizing $\overline{Y}$. This centers the distributions in Figure 2.8 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. The normal distribution is scaled so that the height of the distributions is approximately the same in all figures.

FIGURE 2.10 Distribution of the Standardized Sample Average of n Draws from a Skewed Distribution



The figures show the sampling distribution of the standardized sample average of n draws from the skewed (asymmetric) population distribution shown in Figure 2.10a. When n is small ($n = 5$), the sampling distribution, like the population distribution, is skewed. But when n is large ($n = 100$), the sampling distribution is well approximated by a standard normal distribution (solid line), as predicted by the central limit theorem. The normal distribution is scaled so that the height of the distributions is approximately the same in all figures.

統計学とは

- データを使って世の中のことを知るための科学
  - 記述統計と推測統計
- 問題にしている母集団の分布の特性（平均や分散など）を知る手がかり
  - 悉皆調査は困難：費用、集計の時間、非回答者の存在
  - 標本を用いた統計的推測 statistical inference
- よく使う統計手法
  - 推定 estimation：未知の特性値の best guess の計算
  - 仮説検定 hypothesis testing：特性値についての真偽の判定
  - 信頼区間 confidence interval：ありそうな区間の推測

推定量の分布

- ある母集団の平均 μ_Y を知りたいとき，サンプルが i.i.d. であればその標本平均 $\overline{Y}$ を計算するのは自然
  - $\overline{Y}$ の，μ_Y の推定量としての性質はどうか？
  - 推定量 estimator：標本から実数への関数のこと
  - 推定値 estimate：得られた数値そのもの
- 推定量の標本分布が持つべき望ましい性質とは何か？
  - 推定量は標本から計算される分布を持つ確率変数
  - 推定量自体は自由に定義できる
  - そのうちでも「何らかの意味で真の値に近づいている」「真の値の近くに分布している」などの「望ましい性質」を持つものを探す

推定量の望ましい性質

不偏性 unbiasedness　：標本分布の平均が母平均に等しい

$$E[\hat{\mu}_Y] = \mu_Y$$

一致性 consistency　：標本の大きさが大きくなるほど，真の値の周りの区間に入る確率が 1 へ近づく．推定量が母平均に確率収束する

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y$$

効率性　：不偏推定量のうち，推定量の分布の分散がより小さいこと．「有効性」とも．他の推定量と比べて分散が小さいかどうかを判定．

標本平均 $\overline{Y}$ の性質

- 不偏性，一致性を持つ

$$E[\overline{Y}] = \mu_Y, \overline{Y} \xrightarrow{p} \mu_Y$$

- 線形不偏推定量のなかでもっとも効率的
  - BLUE

母平均の最小二乗推定量としての標本平均

- 標本平均は，各データとの差の 2 乗の平均を最小化しているという意味でデータにフィット
- 予測誤差の 2 乗和の最小化問題

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2 \tag{3.2}$$

の解を最小二乗推定量 least squares estimator と呼ぶ．FOC は，

$$\frac{d}{dm} \sum_{i=1}^{n} (Y_i - m)^2 = -2 \sum_{i=1}^{n} Y_i + 2nm = 0$$

$$m = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

無作為抽出の重要性

- ランダムでないサンプリングでは，一般に標本平均は不偏性を持たない
- ある特性を持った観測値を overrepresent や oversample している可能性がある
- たとえば「労働力調査」のばあい
  - 層化 2 段抽出法による標本調査
  - 調査区を第 1 次抽出単位，住戸を第 2 次抽出単位として系統抽出

仮説検定の基礎用語

仮説検定 hypothesis testing

- 母集団の分布の特性値について仮説を立て，その真偽を統計的（確率的）に判定すること

帰無仮説 null hypothesis

- 検定したい仮説.「平均値が〜に等しい」といった形で定式化されることが多い

$$H_0 : E[Y] = \mu_{Y,0} \tag{3.3}$$

対立仮説 alternative hypothesis

- 帰無仮説が成り立たないときの仮説

$$H_1 : E[Y] \neq \mu_{Y,0} \text{(両側対立仮説)} \tag{3.4}$$

仮説検定の基礎用語

受容 accept，棄却 reject

- 帰無仮説をさしあたって支持することが受容，支持しないことが棄却
- 受容とは，帰無仮説が真であるという強い宣言ではない
- もとの H_0 自体を棄却されるべきものとして設定することが多い
- 受容・棄却の基準は，検定するひとが決める

検定における過誤

2 種類の過誤

- 第 1 種の過誤 type I error：仮説が正しいのに棄却する過誤
- 第 2 種の過誤 type II error：仮説が間違いなのに受容する過誤
- データが所与であれば，第 1 種の過誤と第 2 種の過誤をともに減らすことはできず，トレードオフの関係にある

	受容 accept	棄却 reject
真 true	(A) OK	(B) type I error
偽 false	(C) type II error	(D) OK

$$\text{検定の size} = \frac{\text{仮説が真で，受容}}{\text{仮説が真}} = \frac{A}{A+B}$$

$$\text{検定の検出力 power} = \frac{\text{仮説が偽で，棄却}}{\text{仮説が偽}} = \frac{C}{C+D}$$

p 値 p-value

実際に計算された標本統計量が仮説と等しいことはほとんどない
$(\overline{Y} \neq \mu_{Y,0} = E[Y])$

- サンプリングによる誤差

p 値

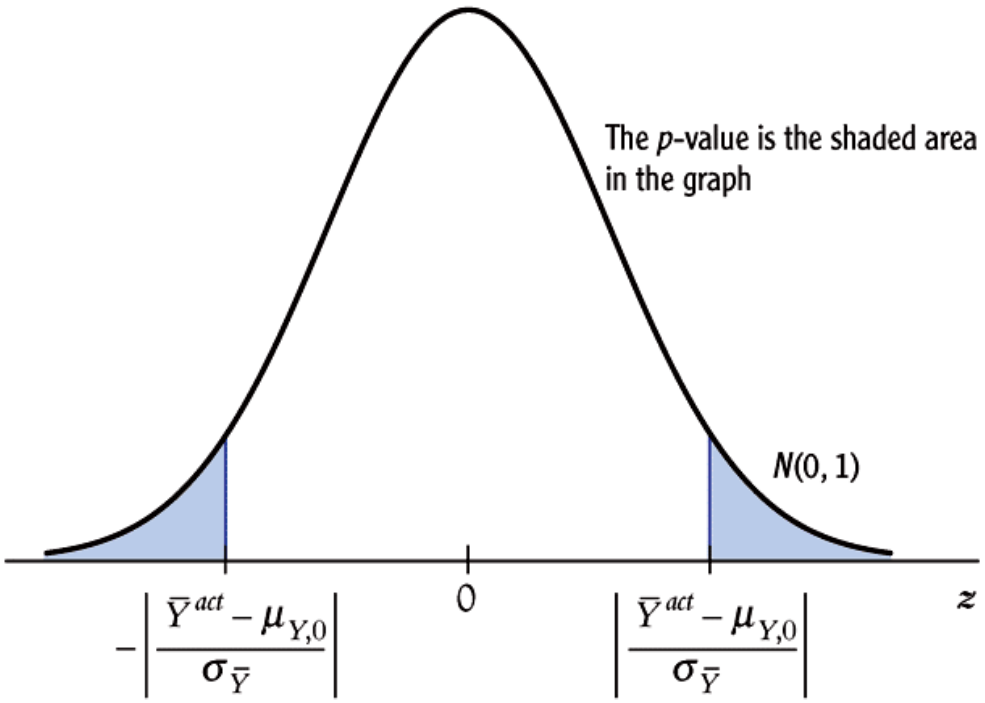
- 帰無仮説が正しいとしたときに，実際に得られた値より「離れた」値が得られる確率

$$p\text{-value} = \Pr\left[\left|\overline{Y} - \mu_{Y,0}\right| > \left|\overline{Y}^{a} - \mu_{Y,0}\right|\right] \tag{3.5}$$

- 第 1 種の過誤 type I error の確率
- p 値が大きい：H_0 のもとで「起こりやすい」こと
- 分布の裾 tail の積分値に等しい

FIGURE 3.1 Calculating a *p*-value

The *p*-value is the probability of drawing a value of $\overline{Y}$ that differs from $\mu_{Y,0}$ by at least as much as $\overline{Y}^{act}$. In large samples, $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma^2_{\overline{Y}})$ under the null hypothesis, so $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ is distributed $N(0, 1)$. Thus the *p*-value is the shaded standard normal tail probability outside $\pm|(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}|$.



p 値の求め方

p 値を計算するには，標本平均の分布関数が必要

- 小標本なら，母集団分布が分かっていても標本分布は複雑
- 大標本なら正規分布で近似可能（中心極限定理）
- $H_0 : \overline{Y} = \mu_{Y,0}$ のもとで，$\overline{Y} \xrightarrow{d} N\left(\mu_{Y,0}, \sigma_Y^2/n\right)$
- 標準化すれば $N(0,1)$ の分布関数さえあれば p 値は求まり，母集団分布の情報は不要

母集団の分散 σ_Y^2 は分かっているのか?

母分散が既知のときの平均値の仮説検定

大標本ならば，$H_0 : \overline{Y} = \mu_{Y,0}$ のもとで，

$$\overline{Y} \sim N\left(\mu_{Y,0}, \sigma_Y^2/n\right) \iff \frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}} \sim N(0,1) = \Phi$$

となるので，

$$\begin{aligned} p\text{-value} &= \Pr\left(\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\overline{Y}^a - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right) \\ &= 2\Phi\left(-\left|\frac{\overline{Y}^a - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right) \end{aligned} \tag{3.6}$$

ただし，一般には母分散は未知.

標本分散，標本標準偏差，標準誤差

標本分散 sample variance s_Y^2：母分散の不偏・一致推定量

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \xrightarrow{p} \sigma_Y^2 \tag{3.7}$$

- n ではなく $n-1$ で除しているのは自由度修正 degree of freedom correction
- 一致性の証明には 4 次モーメントの条件が必要

標本標準偏差 sample standard deviation $s_Y = \sqrt{s_Y^2}$
標準誤差 standard error：標本平均の標準誤差の推定量

$$\hat{\sigma}_{\overline{Y}} = \mathrm{SE}(\overline{Y}) = s_Y/\sqrt{n}$$

母分散が未知のときの平均値の仮説検定

- 標本分散 s_Y^2 は母分散の一致推定量だから，未知の標準偏差の代わりに標本標準偏差を用いると，

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^a - \mu_{Y,0}}{\mathrm{SE}(\overline{Y})}\right|\right) \tag{3.10}$$

- 標準化された標本平均はとくに t 値 t-statistic と呼ばれる

$$t\text{-value} = \frac{\overline{Y}^a - \mu_{Y,0}}{\mathrm{SE}(\overline{Y})} \xrightarrow{d} N(0,1) \tag{3.11}$$

ここで，$\dfrac{Z}{\sqrt{W/m}} \sim t_m$ を思い出そう

t 統計量 t-statistic

- 母集団分布が正規分布であれば，t 値は自由度 $n-1$ の t 分布に従うから，任意の n に対して exact に p 値を計算可能
- ただし，この授業では t 分布を用いない
  - 母集団分布が正規分布であるときしか t 分布は使えない.
  - 経済データがこのような条件を満たすとは考えにくい
  - 正規分布と t 分布の違いは，n が大きい限りにおいて小さいか，無視できる
  - $n > 100$ であれば，漸近正規分布を利用してよい
- 統計ソフトによっては，t 分布の利用が初期値の場合も.

有意水準 significance level

事前に規定した水準より p 値が小さいときに帰無仮説を棄却する，という手続きがしばしば採られてきた

- たとえば，$p < 0.05$ で棄却すると決めておけば，$|t^a| > 1.96$ で帰無仮説を棄却
- このような水準を有意水準とよび,「有意水準 5%で μ_Y は $\mu_{Y,0}$ と統計的に有意に statistically significantly 異なる」という
- ただし，このような言い方では情報が少なくなるので，p 値を報告することがおおい
- 用いる有意水準は 1, 5, 10%など．低い有意水準では帰無仮説を棄却しやすくなる

信頼区間 confidence interval

- 標本の抽出は確率的なので，母平均の正確な値を知ることはできない
- ある確率（信頼水準 confidence level）で真の値が含まれるような集合（信頼集合 confidence set）を見つける
- 1 変数のケースで，信頼集合が上限と下限を持つ実数の集合であるときには信頼区間 confidence interval ともいう
- 「95%両側検定では標本平均より $1.96 \times \mathrm{SE}(\overline{Y})$ 以上遠いと信頼集合から外される」という性質を用いて，

$$\overline{Y} - 1.96 \times \mathrm{SE}(\overline{Y}) \leq \overline{Y} \leq \overline{Y} + 1.96 \times \mathrm{SE}(\overline{Y})$$

を 95%信頼区間とする．

異なる母集団の平均の検定

2 つの異なる母集団の分布の平均の差の検定を考える．2 つのグループをそれぞれ添え字 m, w で表す

- 2 つの母平均の差が d_0 であるという帰無仮説を立てる．母平均が等しいという仮説なら $d_0 = 0$

$$H_0 : \mu_m - \mu_w = d_0 \quad \text{vs.} \quad H_1 : \mu_m - \mu_w \neq d_0 \tag{3.18}$$

- 母平均の差 $\mu_m - \mu_w$ の推定量として $\overline{Y}_m - \overline{Y}_w$ を考える．これを用いて検定を行うにはこの分布を知る必要があるが，中心極限定理より

$$\overline{Y}_m \xrightarrow{d} N\left(\mu_m, \frac{\sigma_m^2}{n_m}\right), \quad \overline{Y}_w \xrightarrow{d} N\left(\mu_w, \frac{\sigma_w^2}{n_w}\right)$$

異なる母集団の平均の検定

$\overline{Y}_m$ と $\overline{Y}_w$ は異なる母集団からの標本から計算された標本平均であるから，互いに独立した確率変数と考えることができ，正規分布の性質を用いると，

$$\overline{Y}_m - \overline{Y}_w \xrightarrow{d} N\left(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right)$$

通常は σ_m^2 と σ_w^2 は未知だから，一致推定量としての標本分散 s_m^2，s_w^2 を用いると，標準誤差は

$$\mathrm{SE}\left(\overline{Y}_m - \overline{Y}_w\right) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} \tag{3.19}$$

異なる母集団の平均の検定

標準化して検定統計量を求めると，H_0 のもとで

$$t = \frac{\left(\overline{Y}_m - \overline{Y}_w\right) - d_0}{\mathrm{SE}\left(\overline{Y}_m - \overline{Y}_w\right)} \xrightarrow{d} N(0,1) \tag{3.20}$$

となるので,

$$p\text{-value} = 2\Phi(-|t|) \tag{3.20}$$

p 値が小さければ，2 つの母平均の差が d_0 であるという帰無仮説は棄却される.
異なる母集団の平均の差の信頼区間も同様に

$$\overline{Y}_m - \overline{Y}_w \pm 1.96 \times \mathrm{SE}(\overline{Y}_m - \overline{Y}_w) \tag{3.21}$$

因果効果 causal effect の推定

- 「理想的な」実験
  - 無作為割り当て実験：randomized controlled experiment
  - 標本を母集団から無作為抽出
  - 実験群 treatment group と対照群 control group に無作為に割り当て
  - 結果の差が「政策」の効果
- 理想的な実験の基本的な考え方
  - 「政策」以外には全く同じ 2 人を連れてきて結果を比べる，というわけではない．そのような 2 人は存在しない
  - 無作為に対象を選んで無作為に「政策」を割り当てることで因果関係が計測できる
- 興味ある因果関係の大きさを知るのに必要な理想的な実験とは？
  - データは理想的な実験から得られたものではない
  - ではどこに違いがあるのか？

因果効果と条件付き期待値

- 「政策」の割り当てが無作為なら,「政策」の水準は結果に影響する他の要因から独立して分布
- 「政策」が 2 値離散変数で表現できるとき，treatment effect は

$$\text{treatment effect} = E[Y|X=1] - E[Y|X=0]$$

- 「政策に効果がない」という帰無仮説を「差の検定」で検定できる
- 心理学・医学・薬学分野では実験はしばしば因果関係の推定に用いられる. 経済学では？

散布図・標本相関係数・標本共分散

散布図： 2 変数の分布を 2 次元平面に表したもの．2 変数の関係を概観するのに適している

標本共分散 s_{XY}： 標本から計算される共分散．自由度修正すると一致性をもつ

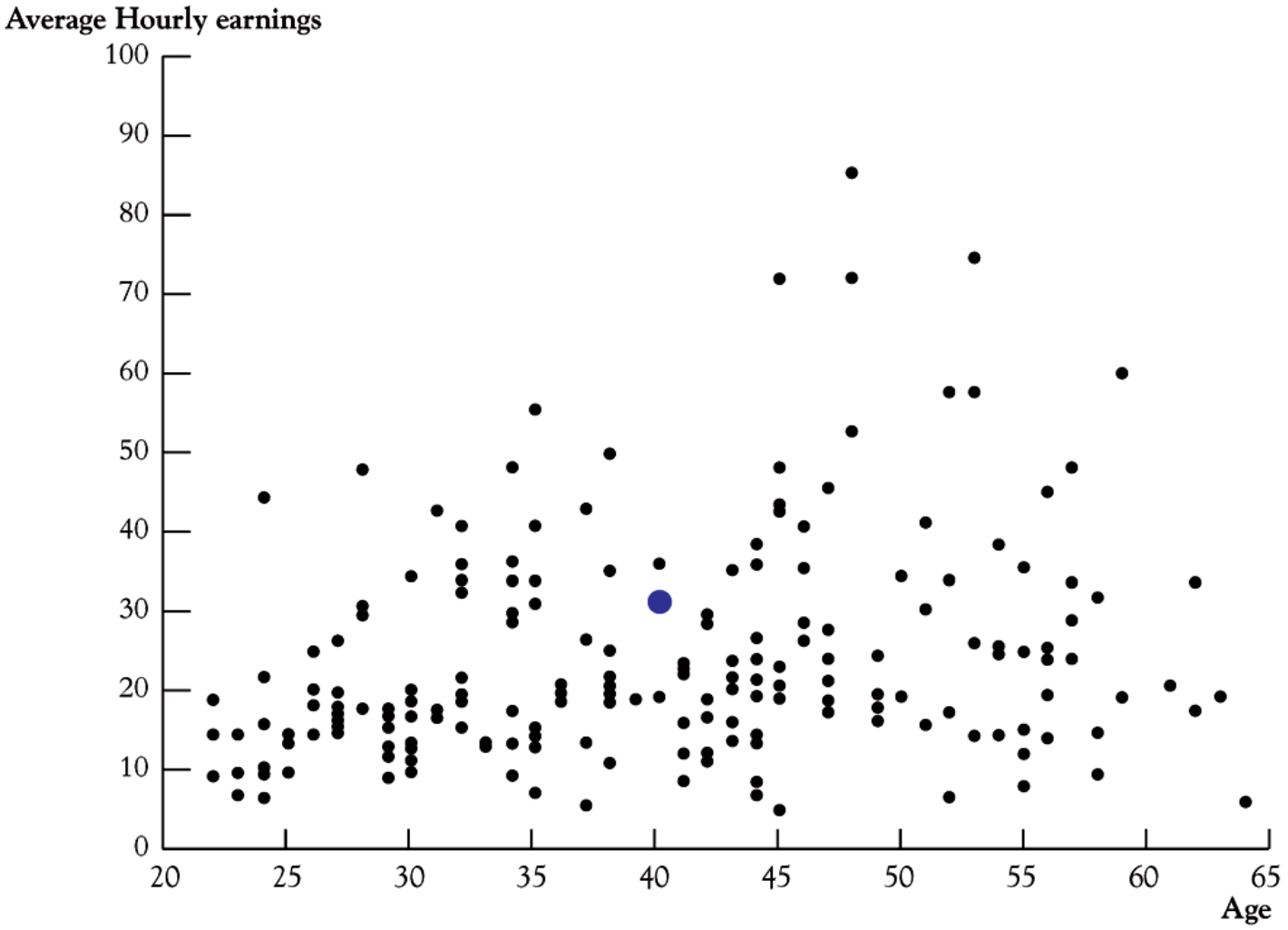
$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \xrightarrow{p} \sigma_{XY} \qquad (3.24,\ 26)$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \xrightarrow{p} \mathrm{corr}(X, Y) \qquad (3.25)$$

散布図・標本相関係数・標本共分散

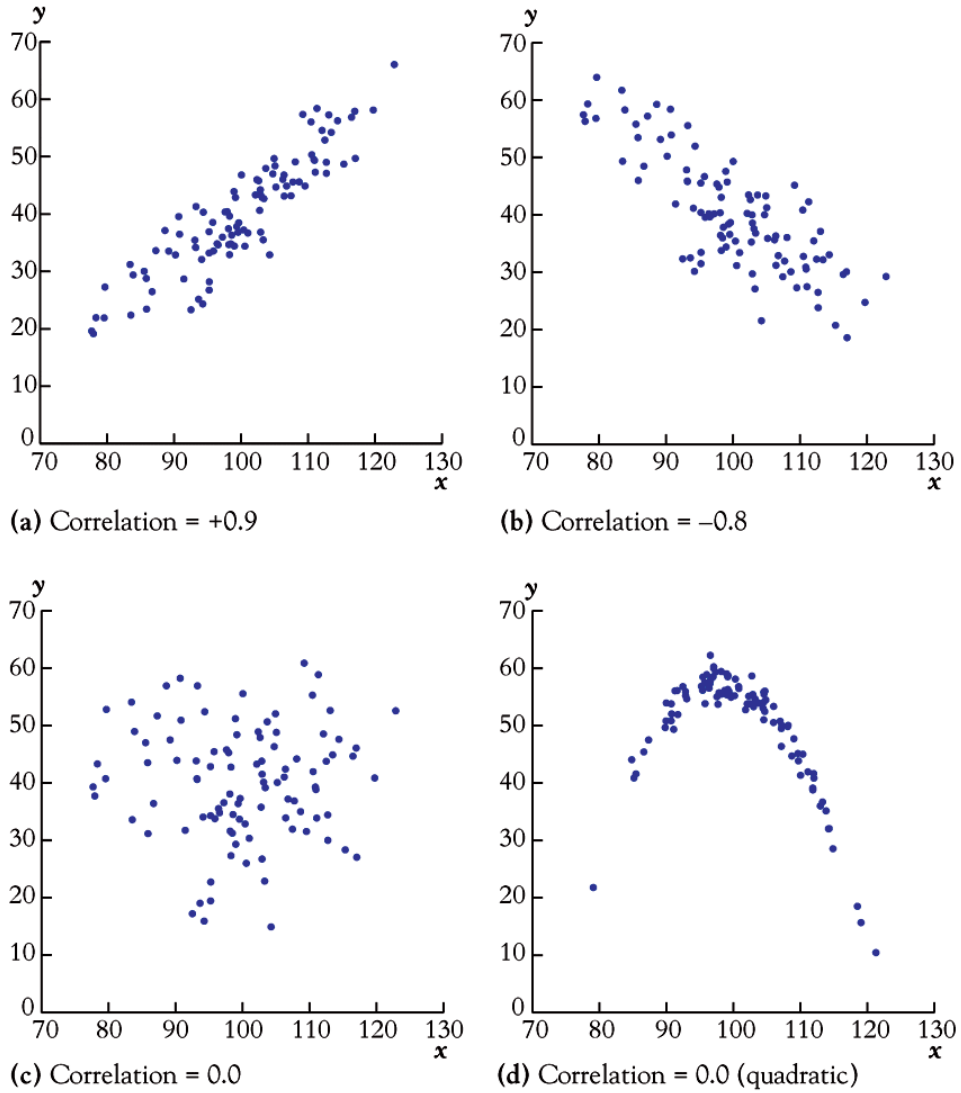
- n 個の観測値の線形のつながりの強さを表す
- $|r_{XY}| \leq 1$ が成り立つ
- $|r_{XY}| = 1$ であれば，散布図上の点は直線上に並ぶ．右上がりなら $r_{XY} = 1$，右下がりなら $r_{XY} = -1$．直線状に近いほど絶対値が 1 に近付く
- 2 変数に非線形な関係があるとき，相関係数がゼロに近くなるケースがある

FIGURE 3.2 Scatterplot of Average Hourly Earnings vs. Age



Each point in the plot represents the age and average earnings of one of the 200 workers in the sample. The colored dot corresponds to a 40-year-old worker who earns $31.25 per hour. The data are for technicians in the information industry from the March 2005 CPS.

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets



(a) Correlation = +0.9

(b) Correlation = –0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between *X* and *Y*. In Figure 3.3c, *X* is independent of *Y* and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.