

統計復習01

データの特性

計量経済分析

1

データの特性

1. データの特性とは？

- ◆ 中心地 *Location*
- ◆ 散らばり具合 *Dispersion*
- ◆ 相関 *Correlation*

2. データの分類

- ◆ 度数分布 *Frequency Distribution*
- ◆ ヒストグラム *Histogram*

計量経済分析

2

様々な中心 Central location

◆ 算術平均 *Arithmetic mean*

◆ 加重平均 *Weighted mean*

$$\bar{X}_w = \frac{(w_1 X_1 + w_2 X_2 + \dots + w_n X_n)}{(w_1 + w_2 + \dots + w_n)}$$

◆ 幾何平均 *Geometric mean*

$$GM = \sqrt[n]{(X_1)(X_2)(X_3)\dots(X_n)}$$

◆ 中位値 *Median*

◆ 最頻値 *Mode*

計量経済分析

3

様々な中心 Central location

◆ 母平均 *population mean*

$$\mu = \frac{\sum X_i}{N}$$

◆ 標本平均 *sample mean*

$$\bar{X} = \frac{\sum X_i}{n}$$

◆ 算術平均の特徴

- すべてのデータセットは唯一の平均を持つ
- すべての観測値が計算過程で使用される
- はずれ値の影響を受けやすい
- 中心を測る指標で唯一、偏差和が0となる

計量経済分析

4

散らばり具合 Dispersion

◆ 範囲 *Range*

◆ 平均偏差 *Mean deviation*

$$MD = \frac{\sum |X_i - \bar{X}|}{n}$$

◆ 分散 *Variance*

■ 母分散

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

■ 標本分散

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

◆ 標準偏差 *Standard deviation (s.d.)*

◆ 変動係数 *coefficient of variation* $CV = \frac{s}{\bar{X}} (100\%)$

計量経済分析

5

分散・標準偏差の特徴

◆ 分散の特徴

- 計算にすべての観測値を使用
- はずれ値にあまり影響されない
- 単位は元の2乗

◆ 標準偏差の特徴

- 母標準偏差 σ 、標本標準偏差 s はそれぞれ分散の1/2乗
- 単位は元の単位と同じ

計量経済分析

6

相関Correlation

◆相関係数Coefficient of Correlation (r)

- 2変数間の線形関係 (**linear** relationship)

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

- $-1 \leq r \leq 1$
- -1 or 1 は完全相関
- 0 に近いほど弱い相関
- $r < 0$ は負の相関、 $r > 0$ は正の相関

計量経済分析

7

データの分類

◆度数分布 frequency distribution

- データを各階級ごとに分類して度数を表示

◆相対度数分布 relative frequency distribution

- 各階級の度数を相対的に表示

◆例: 学生30人の勉強時間 (/週)

15.0, 23.7, 19.7, 15.4, 18.3, 23.0, 14.2, 20.8, 13.5, 20.7, 17.4, 18.6, 12.9, 20.3, 13.7, 21.4, 18.3, 29.8, 17.1, 18.9, 10.3, 26.1, 15.7, 14.0, 17.8, 33.8, 23.2, 12.9, 27.1, 16.6.

このままではデータの特徴が分かりにくい

計量経済分析

8

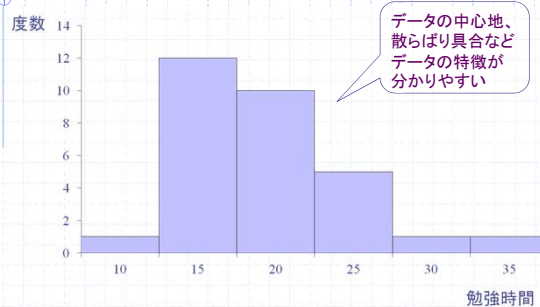
例: 度数分布 (勉強時間/週)

Hours	Frequency f	Relative Frequency
7.5 up to 12.5	1	1/30=.0333
12.5 up to 17.5	12	12/30=.400
17.5 up to 22.5	10	10/30=.333
22.5 up to 27.5	5	5/30=.1667
27.5 up to 32.5	1	1/30=.0333
32.5 up to 37.5	1	1/30=.0333
TOTAL	30	30/30=1

計量経済分析

9

例: ヒストグラム



計量経済分析

10

統計復習02

確率分布

計量経済分析

11

確率分布

1. 確率分布 Probability Distributions

- ランダム変数と確率分布
- 確率分布の平均、分散etc.

2. 正規分布 Normal Distribution

- 正規分布と標準正規分布
- z 値と確率

計量経済分析

12

確率変数と確率分布

- ◆ 確率変数 *random variable*
 - 確率的要素を持つ変数
- ◆ 確率分布 *probability distribution*
 - 変数の値(実現値)とその確率との関係を図示したもの
 - 各実現値の確率 p は $0 < p < 1$
 - 各実現値の確率の合計は1

計量経済分析

13

確率変数の平均と分散

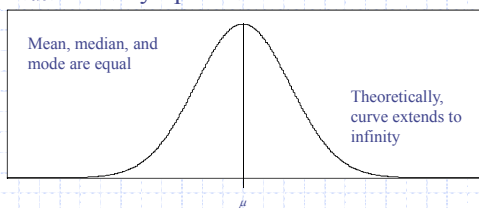
- ◆ 平均(期待値 *expected value*)
 - 確率変数の平均値(確率分布の期待値)
 - $\mu = \sum [xP(x)] = E[x]$
 - ◆ ここで μ は期待値、 $P(x)$ は実現値 x の確率
- ◆ 分散
 - 確率分布の分散
 - $\sigma^2 = \sum [(x - \mu)^2 P(x)] = E[(x - \mu)^2] = \text{Var}[x]$

計量経済分析

14

正規分布の特徴

- ◆ 釣鐘型 (bell-shaped)
- ◆ 平均を中心に左右対称 (symmetrical)
- ◆ 漸近的 (asymptotic)



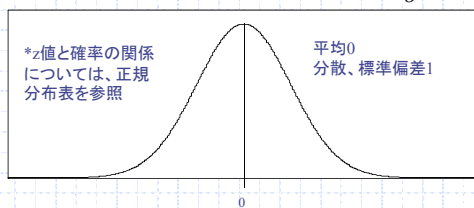
計量経済分析

15

標準正規分布

- ◆ 標準正規分布 *standard normal distribution*

- 平均0、分散1となる正規分布
- 標準化指数(z値 *z-value*): $z = \frac{X - \mu}{\sigma}$



計量経済分析

16

例1

近年の大卒月収が平均20万円、標準偏差2万円の正規分布に従うとき、月収22万円である人のz値はいくらか？ また16万円の場合はどうか？

$$z = \frac{X - \mu}{\sigma} = \frac{22 - 20}{2} = 1$$

$$z = \frac{X - \mu}{\sigma} = \frac{16 - 20}{2} = -2$$

それぞれ上位何%に位置しているか？

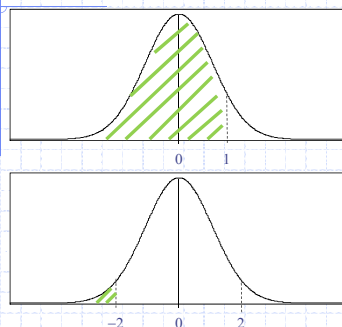
月収22万円のケース: 下位から0.8413 → 上位(約)15.9%

月収16万円のケース: 下位から0.0228 → 上位(約)97.7%

計量経済分析

17

例1 cont.



- ◆ 1以下の確率(左図の斜線部分)は分布表から0.8413

- ◆ 1以上の確率は0.1587(=1-0.8413)

- ◆ -2以上の確率(左図の白部分)は分布が左右対称ゆえ、2以下と同じ

- ◆ 2以下の確率は0.9772

計量経済分析

18

例2

◆ある市の1人当たり水道使用量(/日)は、平均80l、標準偏差20lの正規分布に従う。

1. このとき1日に80-96l使用する人は全体の何%に当たるか？

$$z_{80} = \frac{X - \mu}{\sigma} = \frac{80 - 80}{20} = 0$$

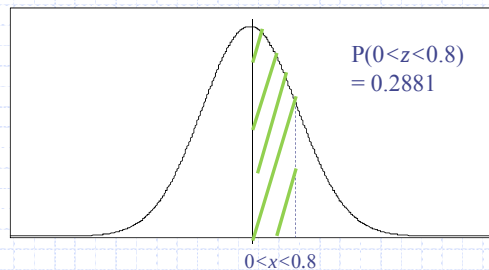
$$z_{96} = \frac{X - \mu}{\sigma} = \frac{96 - 80}{20} = 0.8$$

正規分布表より28.81%($z < 0$ の確率は0.5、 $z < 0.8$ の確率は0.7881。したがって $p = 0.7881 - 0.5 = 0.2881$ 。次図参照。)

計量経済分析

19

例2 cont.



計量経済分析

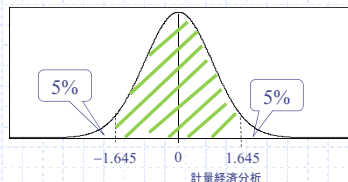
20

例2 cont.

2. 上位5%と下位5%の人は、1日にそれぞれ何l使用するか？

$$z_{5\%} = \frac{X - \mu}{\sigma} = \frac{X - 80}{20} = \pm 1.645$$

$$X = 80 \pm 1.645 \times 20 = (47.1, 112.9)$$



正規分布表から、上位5%のz値は1.645

計量経済分析

21

統計復習03

サンプリングと中心極限定理

計量経済分析

22

サンプリングと中心極限定理

1. 標本平均の分散
2. 中心極限定理 *Central Limit Theorem*
 - 中心極限定理の応用

計量経済分析

23

1. 標本平均の分布

◆標本平均の分布(平均と分散)

- 母集団から抽出された標本について、それぞれの平均を計算し、それをプロットした確率分布
- どの標本が母集団から抽出されるか事前に分からないため、標本平均は確率変数

$$E(\bar{X}) = E\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n} \sum E(X_i) = \frac{n\mu}{n} = \mu$$

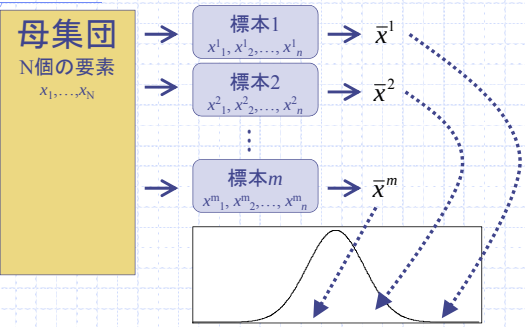
$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

*in the case of finite population, $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

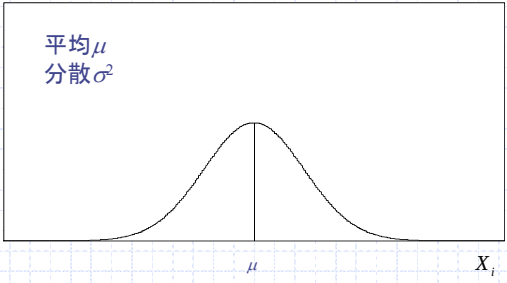
計量経済分析

24

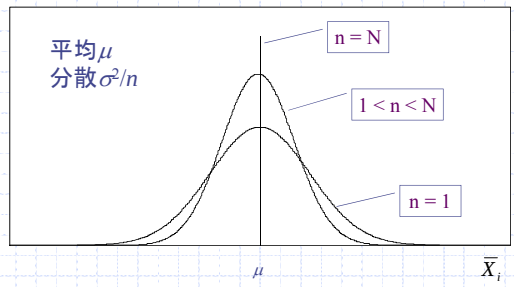
「標本平均の分布」のイメージ



母集団の分布



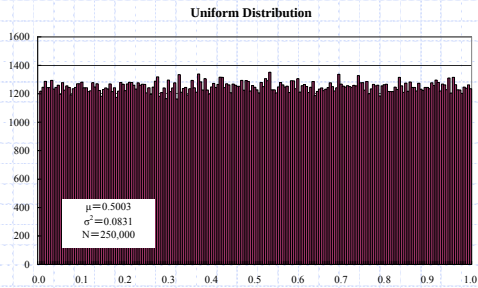
標本平均の分布



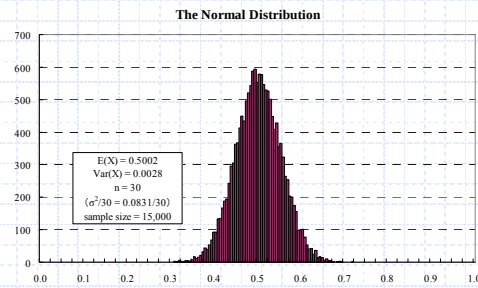
2. 中心極限定理 (CLT)

- ◆ 平均 μ 、分散 σ^2 である母集団において、
 - そこから抽出した標本平均の分布は、平均 μ 分散 σ^2/n となる
 - ◆ どのような母集団でもこのようになるのがポイント
 - ◆ 母集団が正規分布なら、標本平均分布も正規分布
 - 中心極限定理: 母集団がいかなる分布であっても、サンプルサイズ n が大きくなれば、標本平均の分布は正規分布に近づく
 - ◆ どのような母集団でも正規分布になるのがポイント
 - ◆ これにより確率の計算が可能

CLT: 母集団



CLT: 標本平均の分布



CLT 0 the case of $X \sim N(\mu, \sigma^2)$

◆母集団が正規分布であるなら、いかなるサンプルサイズであっても標本平均の分布は正規分布に従う

- 母集団: $X \sim N(\mu, \sigma^2)$
- このとき標本平均の分布は

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

計量経済分析

31

CLT 1 the case of $X \sim \text{non-normal dist.}(\mu, \sigma^2)$

◆母集団が正規分布でないとき(ただし分散は既知)、サンプルサイズが大きくなれば、標本平均の分布は正規分布に従う

- 母集団: $X \sim \text{non-normal dist.}(\mu, \sigma^2)$
- (サンプルサイズが大きいときの)標本平均の分布は

$$\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right) \quad \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0, 1)$$

計量経済分析

32

CLT 2 the case of $X \sim N(\mu, \text{unknown } \sigma^2)$

◆母集団が正規分布であるが、その分散が未知であるとき、標本平均の分布はt分布に従う。(ただしサンプルサイズが増えれば正規分布に)

- 母集団: $X \sim N(\mu, \text{unknown } \sigma^2)$
- 標本平均の分布

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(d.f.) \quad t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \xrightarrow{d} N(0, 1)$$

計量経済分析

33

CLT 3 the case of $X \sim \text{non-N}(\mu, \text{unknown } \sigma^2)$

◆母集団が正規分布でなく、その分散も未知であれば、サンプルサイズが大きくなったとき、標本平均の分布は正規分布になる。

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \xrightarrow{d} N(0, 1)$$

計量経済分析

34