

2002 年 4 月 22 日

総合科目 数理・自然 6 「統計学入門」講義レジュメ (今野)

2.1 度数分布表とヒストグラム

1. 度数分布表とは、観測値の取りうる値をいくつかの階級 (区間) に分け、それぞれの階級に観測値がいくつ含まれるか度数を数え、作成した表である.
2. 度数分布表の作成

(a) n 個の観測値を

$$x_1, x_2, \dots, x_n$$

とする.

(b) データの最大値と最小値の間を k 個の (等) 区間に分割する:

$$a_0 < a_1 < \dots < a_k$$

各区間 $(a_{i-1}, a_i]$, $i = 1, 2, \dots, k$ を 階級 とよぶことにする.

(c) それぞれの階級の中央値

$$m_i = \frac{a_{i-1} + a_i}{2}, \quad i = 1, 2, \dots, k$$

を 階級値 とよび、階級を代表する値とする.(d) 各階級 $(a_{i-1}, a_i]$, $i = 1, 2, \dots, k$ に属する個体数を数える. それを f_i と記す. f_i を 度数 と呼ぶ. さらに,

$$\frac{\text{度数}}{\text{観測数}} = \frac{f_i}{n}$$

を 相対度数 と呼ぶ. 定義から

$$\sum_{i=1}^n \frac{f_i}{n} = 1$$

となる.

(e) i 番目の階級 (値をしたから観て) の上側の境界値 a_i 以下の観測値をもつ個体数

$$F_i = f_1 + f_2 + \dots + f_i = \sum_{j=1}^i f_j$$

を 累積度数 という. さらに, F_i/n を 累積相対度数 と呼ぶ. 定義から $F_k = n$ である.(f) 階級, 階級値, 度数, 相対度数, 累積度数, 累積相対度数をまとめた表を 度数分布表 といい, これを柱状のグラフにしたものを ヒストグラム とよぶ.

3. ヒストグラムから観察できること

- (a) 分布の峰がひとつか？それとも峰が複数あるか？ — 単峰型か双峰型？
 双峰型の場合には，層別化
- (b) 峰がひとつの分布の場合，
- i. 左右対称の山型の分布
 - ii. 右に歪んだ分布 — 右側に長く裾を引く分布か
 - iii. 左に歪んだ分布 — 左側に長く裾を引く分布か

分布の特徴を表す指標

1. 代表値 — 分布を代表する値のことである．

(a) 算術平均

観測値を $x_1, x_2, \dots, x_n$ とする．算術平均は

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\text{観測値の総和}}{\text{観測数}}$$

(b) 中央値 (median)

観測値を $x_1, x_2, \dots, x_n$ を小さいものから順に並び替えたものを

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

と記す．

$$\text{中央値} = \begin{cases} x_{(m+1)} & n \text{ が奇数のとき } (n = 2m + 1) \\ \frac{x_{(m)} + x_{(m+1)}}{2} & n \text{ が偶数のとき } (n = 2m) \end{cases}$$

(c) 最頻値 (モード)

分布の峰に対応する値

(d) 百分位点〔第一四分位点・メデアン・第三四分位点〕

(e) 平均・中央値・モードと分布の形の関係 (単峰型の分布のとき)

i. 左右対称の山型の分布

おおむね一致する

ii. 右に歪んだ分布

モード < 中央値 < 平均

iii. 左に歪んだ分布

平均 < 中央値 < モード

2. 散らばりの尺度

(a) 範囲

(b) 四分位偏差

(c) 平均偏差

(d) 分散と標準偏差

3. 変動係数

4. 標準化得点