

## 数理統計学まとめ (その1)

### 1 データの整理

#### 1. 度数分布表

- (a) データの幅：(データの最大値) - (データの最小値) から階級の幅  $c$  と階級の数  $k$  を決める。  
 $k$  は 10 ~ 15 くらいが見やすい。  
最初と最後の階級を除いて階級の幅は  $c$  にする。
- (b) 階級の上端  $a_j$  と下端  $b_j$  : 測定値より一桁落とす。(重なりを防ぐ)
- (c) 階級の代表値  $x_j$  :  $a_j \sim b_j$  の階級では  $\frac{a_j+b_j}{2}$  .
- (d) 階級の度数  $f_j$  : 各階級に入るデータの数

#### 2. ヒストグラムと棒グラフ

- (a) 棒グラフ : 横軸は階級、縦軸は度数
- (b) ヒストグラム : 横軸は階級とその幅、単位は面積 .

#### 3. その他

- 累積度数 : 下から  $j$  番目の階級までの度数を足したもの .

$$F_j = f_1 + \cdots + f_j$$

- 累積比率 :  $F_j$  をデータの総数  $n$  で割ったもの .
- 度数多角形 : 折れ線グラフ  
(それぞれの階級の上端  $b_j$  のところで折れ曲がる)

## 2 分布の特性値

### 1. 代表値

- (a) 平均：標本平均．データが  $n$  個  $x_1, x_2, \dots, x_n$  とある時は平均は

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j,$$

度数分布表が与えられたときは

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k f_j \cdot x_j$$

と計算．

- (b) メディアン (中央値)：データが  $x_1, x_2, \dots, x_n$  とある時, これを小さい方から増大順に並べ直した  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  を考え, メディアンは  $n/2$  番目の値．

- $n = 2k + 1$  なら  $k + 1$  番目の値,

$$Me = x_{(k+1)}$$

- $n = 2k$  なら  $k$  番目と  $k + 1$  番目の値の平均

$$Me = \frac{x_{(k)} + x_{(k+1)}}{2}$$

度数分布表が与えられたとき,  $F_{m-1} \leq \frac{n}{2} < F_m$  となる  $m$  を取り,

$$Me = a_m + c \frac{\frac{n}{2} - F_{m-1}}{f_m}$$

と比例配分する．

- (c) モード (最頻値)： $f_m$  が最大になる階級を取り, 簡単にはその階級の代表値  $x_m$  を取る．もう少し詳しくする場合は

$$Mo = a_m + c \frac{f_m - f_{m-1}}{f_m - f_{m-1} + f_m - f_{m+1}}$$

という式が使われる．

(d) 幾何平均  $G_m$  , 調和平均  $H_m$  :  $n$  個のデータがある場合 .

$$G_m = \sqrt[n]{\prod_{j=1}^n x_j}, \quad \therefore \log G_m = \frac{1}{n} \sum_{j=1}^n \log x_j,$$

$$H_m = \frac{n}{\sum_{j=1}^n \frac{1}{x_j}}$$

## 2. 散布度

(a) 分散  $s^2$  :  $n$  個のデータがある場合

$$s^2 = \frac{1}{n} \{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$$

(定義式)

$$s^2 = \frac{1}{n} \sum_{j=1}^n x_j^2 - \bar{x}^2$$

(計算式)

度数分布表が与えられたとき

$$s^2 = \frac{1}{n} \sum_{j=1}^k f_j (x_j - \bar{x})^2$$

(定義式)

$$s^2 = \frac{1}{n} \sum_{j=1}^k f_j x_j^2 - \bar{x}^2$$

(計算式)

ただし , 階級  $a_j \sim b_j$  の代表値が  $x_j$  , 度数が  $f_j$  で , このときの平均  $\bar{x}$  は度数分布表を基に計算したもの .

(b) 標準偏差  $s$  :

$$s = \sqrt{s^2}$$

(c) 四分位偏差  $Q$  : メディアンで全体の中央の値を取ったが , 全体の , 下から  $1/4$  の位置にある点  $Q_1$  , メディアン  $Me$  , 上から  $1/4$  の位置にある点  $Q_3$  を取り ,

$$Q = Q_3 - Q_1$$

と置く . (簡単なばらつきの指標)

(d) モーメント :  $k = 1, 2, 3, \dots$  に対して

$$m'_k = \frac{1}{n} \sum_{j=1}^n x_j^k \quad ((\text{原点積率}))$$

$$m_k = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^k \quad ((\text{中心積率}))$$

$s^2 = m_2$  であることに注意 .

- $m_3$  を  $s^3$  で割ったものを 歪度とよび ,  $g_3$  で書く .  $g_3 > 0$  なら分布の裾野が右に広い .  $g_3 < 0$  なら左に広い .
- $g_4 = m_4/s^4$  を 尖度 と呼ぶ . 正規分布と比較して  $g_3 > 3$  のとき , 山が急に立ち上がる .

3. 標準得点  $x_j$  自身を見るのではなく ,

$$z_j = \frac{x_j - \bar{x}}{s}$$

を見ることがよくある .  $\sum_{j=1}^n z_j = 0$  なので , 平均は 0 になるように変形しており , 標準偏差  $s$  で割ることにより , 他のデータの分布と比較しやすい .  $z_j$  と変形する事を標準化といい , この標準化されたデータを  $z$ -スコアまたは標準得点という .

これをさらに変形して

$$SS_j = 10z_j + 50$$

としたものが 偏差値 .