

経済統計学講義ノート No.2

記述統計（1 変量）

蛭川雅之

2022 年 9 月 12 日

目 次

1. $\sum$ （シグマ）記号
2. 度数分布表
3. ヒストグラム
4. 代表値
5. 箱ひげ図

1 $\sum$ (シグマ) 記号

- $\sum$ (シグマ) 記号を使うと、和の表記を簡単にすることができる。

$$\sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$$

■ルール

1. 添え字のない定数 a に対し、以下が成り立つ。

$$\sum_{i=1}^n aX_i = a \sum_{i=1}^n X_i = aX_1 + aX_2 + \cdots + aX_n.$$

$$\sum_{i=1}^n a = \underbrace{a + a + \cdots + a}_{n \text{ 個}} = na.$$

2. 同一の添え字 i のある変数 X_i と Y_i に対し、以下が成り立つ。

$$\begin{aligned}\sum_{i=1}^n (X_i + Y_i) &= \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i \\ &= (X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_n + Y_n).\end{aligned}$$

3. 2 種類の添え字 i と j に対し、和をとる順序を変えることができる。

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^m X_{ij} &= \sum_{j=1}^m \sum_{i=1}^n X_{ij} \\ &= X_{11} + X_{12} + \cdots + X_{1m} + \cdots \\ &\quad + X_{n1} + X_{n2} + \cdots + X_{nm}.\end{aligned}$$

2 度数分布表

2.1 都道府県別映画館スクリーン数のデータ

- Excel ファイル `cinema.xlsx` は、一般社団法人日本映画製作者連盟が発表している 2017 年 12 月末時点の各都道府県内にある映画館のスクリーン合計数等を一覧表にまとめたものである。
 - － 例えば、ある県に 1 つの映画館があり、その映画館に 10 個のスクリーンがある場合、スクリーン数は 10 となる。
- この表を一見して、都道府県別スクリーン数について何か特徴的なことがわかるのか？
 - － スクリーン数はまちまちの値をとる (⇒ 分布の存在)。
 - － 分布の特徴 (特性・癖) までは...??

2.2 度数分布表：見本

- **度数分布表**を作成すれば、データの特徴が分かり易くなる。
 - Excel でどのように作成するのか？
 - 表に使われている用語の意味は何か？

階 級	階級値	度 数	相対度数	累積度数	累積比率
0 ~ 50	25	24	0.5106	24	0.5106
50 ~ 100	75	14	0.2979	38	0.8085
100 ~ 150	125	2	0.0426	40	0.8511
150 ~ 200	175	1	0.0213	41	0.8723
200 ~ 250	225	4	0.0851	45	0.9574
250 ~ 300	275	1	0.0213	46	0.9787
300 ~ 350	325	0	0.0000	46	0.9787
350 ~ 400	375	1	0.0213	47	1.0000
合 計	—	47	1.0000	—	—

2.3 作成手順

(i) 階級を決める。

1. データの最小値と最大値を探し出す。
 - 最小値は 11（鳥取、高知）、最大値は 363（東京）である。
 - 最小値は関数 `min`、最大値は関数 `max` を用いて探す。
2. 階級の範囲を「最小値より小さく切りの良い数字」から「最大値より大きく切りの良い数字」までとする。
 - 範囲を 0～400 に設定する。
3. 階級の範囲を 5～8 の等間隔の小区間（階級）に区切る。
 - 階級数を 8 にとる。
 - 階級幅を $(400 - 0) \div 8 = 50$ に設定する。

※【階級の表記】“50～100”は通常「50 以上 100 未満」を意味する。

(ii) 階級値を決める。

1. 各階級の中央値をその階級を代表する値（**階級値**）とする。
 - 階級値はその階級の左端と右端の値の平均（＝足して2で割った値）である。

(iii) 度数・累積度数を求める。

1. 各階級に含まれるデータ数（**度数**）を求める（数える？）には、その階級までの累積度数から逆算すると安全である。
 - **累積度数**とはその階級までのデータ数の累計である。
2. 各階級の累積度数は関数 `countif` を用いて求められる。
 - **最後の階級の累積度数はデータ数に一致する。**

3. 各階級の度数は

(その階級までの累積度数) - (直前の階級までの累積度数)

で求められる。

- 度数の合計はデータ数に一致する。

(iv) 相対度数・累積比率を求める。

1. 相対度数は (度数) $\div$ (データ数) で求められる。

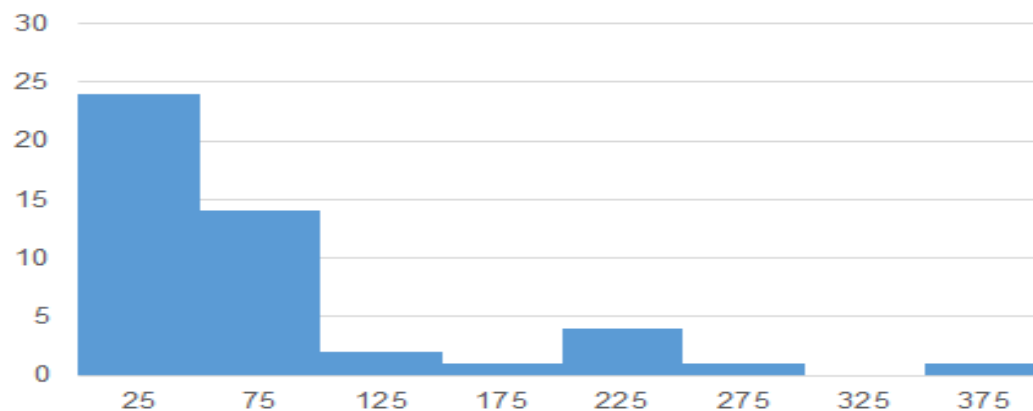
- 相対度数はその階級に含まれるデータ数が全体の何 % にあたるかを示す。
- 相対度数の合計は 1 である。

2. 累積比率 (累積相対度数) は (累積度数) $\div$ (データ数) で求められる。

- 累積比率はその階級までのデータ数が全体の何 % にあたるかを示す。
- 最後の階級の累積比率は 1 である。

3 ヒストグラム

- **ヒストグラム**とは、階級を横軸に、度数を縦軸にとったグラフである。
- Excel では“2-D 縦棒”グラフに手を加えることにより作成できる。
 - － “ヒストグラム”を利用すると階級を自動的に決められてしまうようである。



4 代表値

- 度数分布表やヒストグラムは分布に対する直感的理解を助けてくれる。
- もっとも、
 - 度数分布表やヒストグラムだけを用いて分布の特徴を説明するのは容易でない。
 - 数表・グラフはスペースをとる。
- 分布の特徴を表す数値（**代表値**）を利用して説明すれば、スペースを取らず簡潔に済む。
 - 位置に関する代表値
 - ちらばりの程度に関する代表値
- 以下、 n 個のデータ $X_1, \dots, X_n$ が与えられていると仮定する。

4.1 位置に関する代表値

■ 「中心」を表す代表値

1. 平均値：

- データの総和をデータ数で割って求める。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- 関数 `average` を用いる。

2. 中央値（メディアン）：

- n 個のデータを小さい順に並べ替えた $X_{(1)} \leq \cdots \leq X_{(n)}$ に対し、次のように定義する。

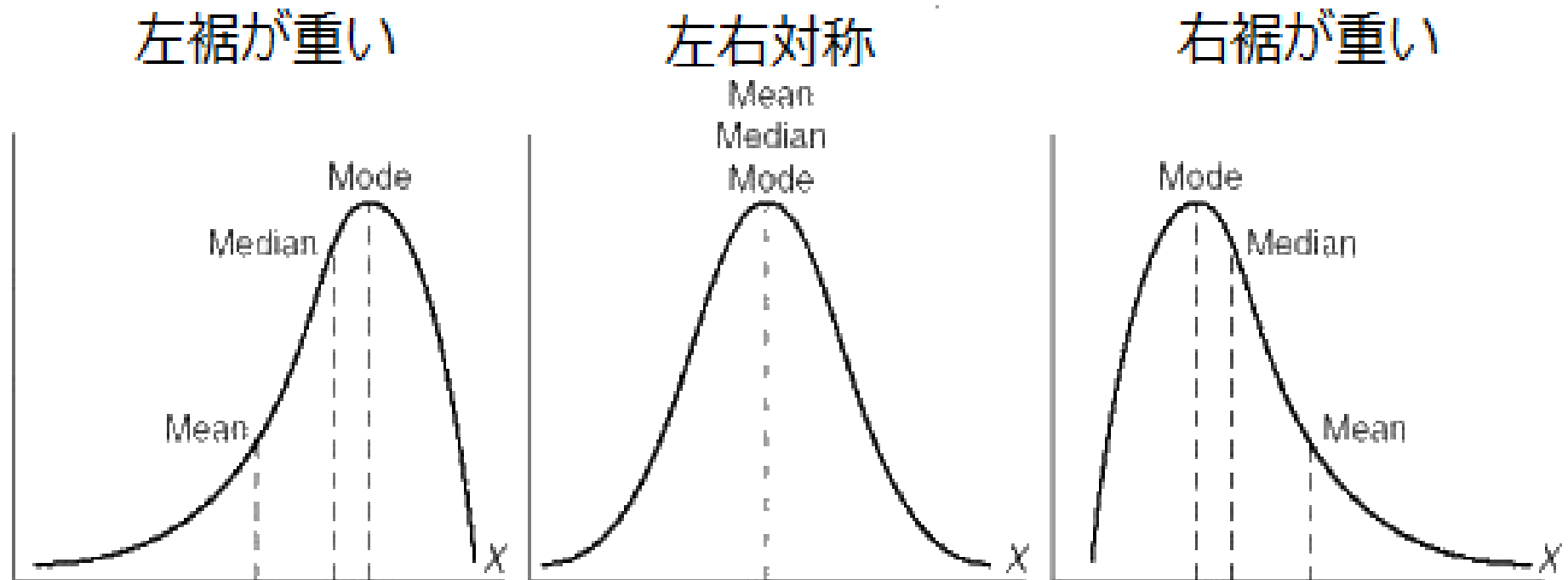
$$\hat{X}_{med} = \begin{cases} X_{(\frac{n+1}{2})} & n : \text{奇数} \\ \frac{1}{2} \left\{ X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)} \right\} & n : \text{偶数} \end{cases}$$

- 関数 `median` を用いる。

3. 最頻値（モード）：

- 最も頻繁に現れる値をいう。
 - － 度数分布表では、（相対）度数が最も大きい階級値を表す。
- 関数 `mode` があるにはある。

■山が一つで左右対称な分布の場合、これら3つの値は一致する。



■その他の代表値

1. 四分位数：

- データを小さい順に並べ替えて四等分した点を指す。
 - 小さいものから順に、第 1 四分位数 (25% 点)、第 2 四分位数 (50% 点) (=中央値)、第 3 四分位数 (75% 点) という。
- 関数 `quartile.inc` を用い、戻り値として 1 (第 1 四分位数)、2 (第 2 四分位数)、もしくは 3 (第 3 四分位数) を指定する。
 - 戻り値として 0 (最小値) および 4 (最大値) も指定できる。
- (第 3 四分位数) – (第 1 四分位数) を四分位範囲 (IQR) という。

4.2 ちらばりの程度に関する代表値

■よく用いられるもの

1. 分散：

- 次の 2 通りの定義がある。

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 : \text{不偏分散}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- s^2 を求めるには関数 `var`、 $\hat{\sigma}^2$ を求めるには関数 `varp` を用いる。

2. 標準偏差：

- 分散の正の平方根として定義される。
 - － 分散の定義に従い、次の 2 通りの定義が可能である。

$$s = \sqrt{s^2} \quad \text{または} \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

- s を求めるには関数 `stdev`（または `sqrt(var)`）、 $\hat{\sigma}$ を求めるには関数 `stdevp`（または `sqrt(varp)`）を用いる。

■その他の代表値

1. 範囲：

- (最大値) – (最小値) で定義される。

2. 変動係数：

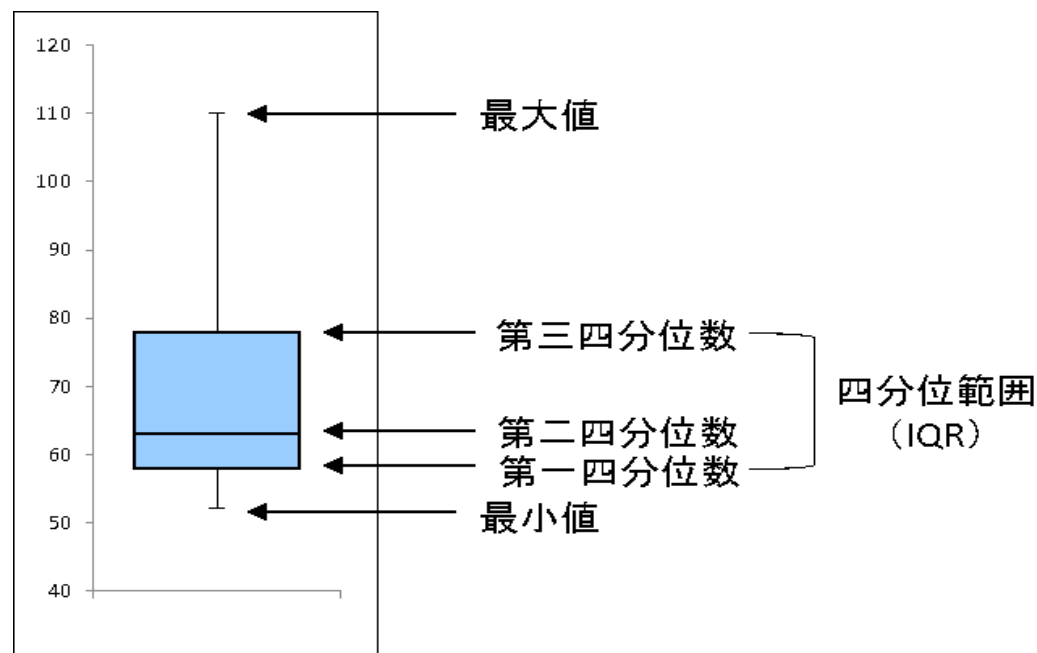
- 標準偏差を平均値で割って求める。

$$CV = \frac{s}{\bar{X}} \quad \text{または} \quad CV = \frac{\hat{\sigma}}{\bar{X}}$$

- 標準偏差が平均値からどの程度ちらばっているかを示す。
 - データの計測単位に依存しないため、直接比較ができる。

5 箱ひげ図

- **箱ひげ図**とは、最小値、第1四分位数、第2四分位数（＝中央値）、第3四分位数、最大値を使い、データのばらつきをわかりやすく表現するためのグラフである。



- Excel 2016 から“統計グラフ”のメニューに箱ひげ図が加わった。
 - － 箱ひげ図オプションでは自動的に「外れ値」が計算・表示されてしまう。
- 旧バージョン Excel の手順に従い、下表に対する積み上げ棒グラフを応用すると外れ値なしの箱ひげ図を作成できる。
 1. 最小値 (Min)、第 1 四分位数 (Q1)、第 2 四分位数 (Q2)、第 3 四分位数 (Q3)、最大値 (Max) を求める。
 2. 1 の結果をもとに次のような表を作成する。

Q1 - Min	15
Q1	26
Q2 - Q1	23
Q3 - Q2	26
Max - Q3	288

- 2 のデータから箱ひげ図を作成する大まかな手順は次の通りである【詳細は実演】。
 1. $Q1$ 、 $Q2 - Q1$ 、 $Q3 - Q2$ の 3 つの数値から積み上げ縦棒グラフを描き、「データの選択」→「行/列の切り替え」により上段・中段・下段の 3 層からなるグラフに変換する。
 2. 下段を「塗りつぶしなし」「線なし」に指定し、「箱」の部分を作成する。
 3. $Q1 - \text{Min}$ ($\text{Max} - Q3$) の数値を使って、以下の手順で下（上）側の「ひげ」を描く。
 - (a) 下段（上段）にカーソルを移動し、グラフ右横の  をクリックする。
 - (b) 「グラフ要素」ウィンドウで「誤差範囲」に ☒ を入れ、横の  から「その他のオプション」を選択する。

- (c) 「誤差範囲の書式」ウィンドウの設定を以下のように行う。
 - i. 方向：「負（正）方向」
 - ii. 終点のスタイル：「キャップあり」
 - iii. 誤差範囲：「ユーザー設定」→「ユーザー設定の誤差」
ウィンドウで $Q1 - \text{Min}$ ($\text{Max} - Q3$) のセルを「負（正）の誤差の値」に指定
- 4. 中段・上段をそれぞれ「塗りつぶしなし」として白抜きにし、
横幅を調整して箱ひげ図を完成させる。

■箱ひげ図【仕上がり見本】

