

Lecture 1

1変量データの整理

生データ / ローデータ (raw data)

新生児の体重（測定単位 g ） 標本の大きさ $n = 100$

3110	2500	2770	3010	3000	3000	2740	3040	3060	3410
3100	2620	3910	3650	2840	2480	2790	3720	3520	2850
3140	2780	2270	2700	2830	3020	3160	4060	2620	3390
3050	3190	3710	3460	3200	3260	3040	3610	3360	3280
2480	3440	2970	3050	2590	3320	3580	3820	3450	4150
3300	3020	3360	3140	3300	3600	3330	3300	3300	3170
3340	3250	2880	3560	3060	3320	2740	2380	3590	2460
2960	3170	3000	3250	3140	3220	3160	3730	3460	3360
3160	3540	2890	3060	2900	3040	3220	3590	2680	3150
2770	3220	2970	3300	3560	3520	2760	2740	2820	4180

StatData01_1.csv

生データ/ ローデータ (raw data)

3110	2500	2770	3010	3000	3000	2740	3040	3060	3410
3100	2620	3910	3650	2840	2480	2790	3720	3520	2850
3140	2780	2270	2700	2830	3020	3160	4060	2620	3390
3050	3190	3710	3460	3200	3260	3040	3610	3360	3280
2480	3440	2970	3050	2590	3320	3580	3820	3450	4150
3300	3020	3360	3140	3300	3600	3330	3300	3300	3170
3340	3250	2880	3560	3060	3320	2740	2380	3590	2460
2960	3170	3000	3250	3140	3220	3160	3730	3460	3360
3160	3540	2890	3060	2900	3040	3220	3590	2680	3150
2770	3220	2970	3300	3560	3520	2760	2740	2820	4180



- わかりやすく整理
- 可視化
- 特徴を抽出

度数分布表 (frequency table)

いくつかの階級（クラス）に
分類して表に整理する

① データの範囲

最大値 (max) $\max = 4180$

最小値 (min) $\min = 2270$

範囲 (range) $R = \max - \min = 1910$

② 階級幅と階級数

決め方には自由度がある

目的や見やすさに応じて決める

生データ/ ローデータ (raw data)

3110	2500	2770	3010	3000	3000	2740	3040	3060	3410
3100	2620	3910	3650	2840	2480	2790	3720	3520	2850
3140	2780	2270	2700	2830	3020	3160	4060	2620	3390
3050	3190	3710	3460	3200	3260	3040	3610	3360	3280
2480	3440	2970	3050	2590	3320	3580	3820	3450	4150
3300	3020	3360	3140	3300	3600	3330	3300	3300	3170
3340	3250	2880	3560	3060	3320	2740	2380	3590	2460
2960	3170	3000	3250	3140	3220	3160	3730	3460	3360
3160	3540	2890	3060	2900	3040	3220	3590	2680	3150
2770	3220	2970	3300	3560	3520	2760	2740	2820	4180



- わかりやすく整理
- 可視化
- 特徴を抽出

度数分布表 (frequency table)

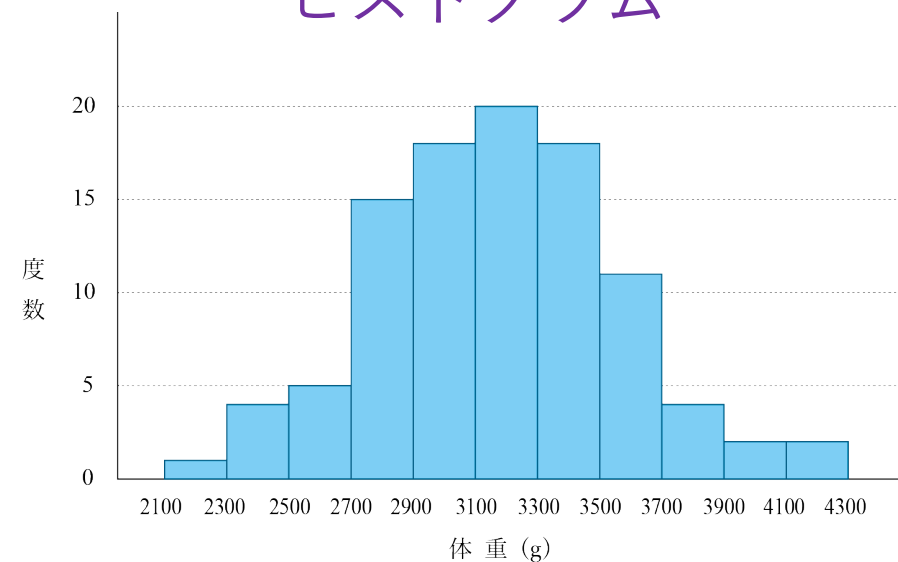
階級番号	階級	階級値	度数
1	2100~2300	2200	1
2	2300~2500	2400	4
3	2500~2700	2600	5
4	2700~2900	2800	15
5	2900~3100	3000	18
6	3100~3300	3200	20
7	3300~3500	3400	18
8	3500~3700	3600	11
9	3700~3900	3800	4
10	3900~4100	4000	2
11	4100~4300	4200	2
計	—	—	100

度数分布表 (frequency table)

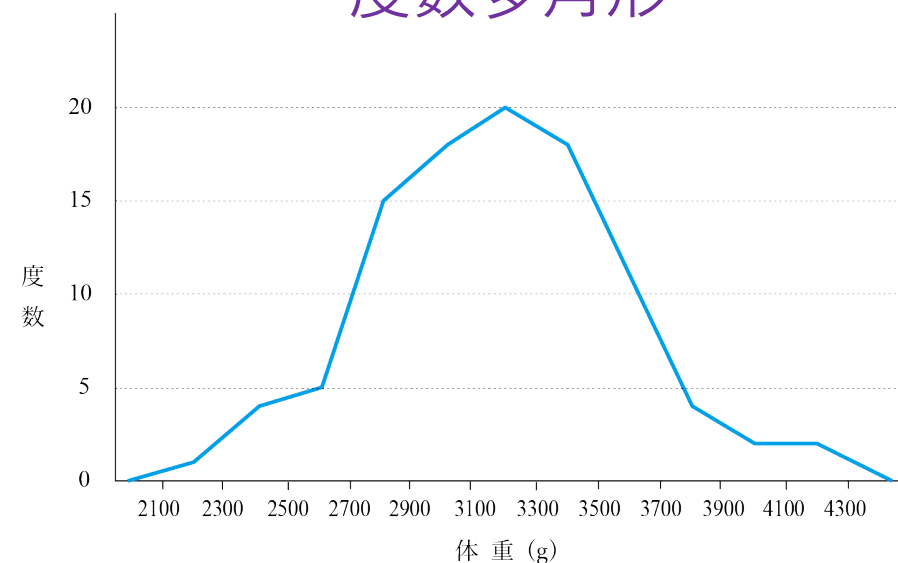
階級番号	階級	階級値	度数
1	2100~2300	2200	1
2	2300~2500	2400	4
3	2500~2700	2600	5
4	2700~2900	2800	15
5	2900~3100	3000	18
6	3100~3300	3200	20
7	3300~3500	3400	18
8	3500~3700	3600	11
9	3700~3900	3800	4
10	3900~4100	4000	2
11	4100~4300	4200	2
計	—	—	100

可視化

ヒストグラム



度数多角形



データから特徴を取り出す：データの代表値

生データ

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

これは単なる（長大な）数列

度数データ

度数分布表

階級値	c_1	c_2	$\dots$	c_j	$\dots$	c_k	計
度数	f_1	f_2	$\dots$	f_j	$\dots$	f_k	n
相対度数	p_1	p_2	$\dots$	p_j	$\dots$	p_k	1

相対度数： $p_j = \frac{f_j}{n}$

データの代表値

平均値 (mean/average)

明らかなときは
 i の範囲を省略

生データの場合：

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum x_i$$

度数データの場合：

階級値	c_1	c_2	...	c_j	...	c_k	計
度数	f_1	f_2	...	f_j	...	f_k	n
相対度数	p_1	p_2	...	p_j	...	p_k	1

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k c_j f_j = \frac{1}{n} \sum c_j f_j = \sum c_j p_j$$

中央値 (median)

生データの場合：データを小さいほうから並べて順位が中央の値

度数データの場合：多少の計算を要す（教科書参照）

最頻値 (mode)

度数データの場合のみ：最大度数をとる階級値

例 題 次の10個のデータの平均値と中央値を求めよ.

2 5 6 7 8 3 1 2 4 24

小さいほうから並べると 1 2 2 3 4 5 6 7 8 24

$$\bar{x} = \frac{1}{10} (1 + \dots + 24) = 6.2$$

$$Me = \frac{1}{2} (4 + 5) = 4.5$$

データのばらつき・広がり

注意（後出）不偏分散

分散 (variance) $s^2 = s_x^2$

生データの場合：
$$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

度数データの場合：

階級値	c_1	c_2	...	c_j	...	c_k	計
度数	f_1	f_2	...	f_j	...	f_k	n
相対度数	p_1	p_2	...	p_j	...	p_k	1

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

標準偏差 (standard variation) $s = \sqrt{s^2}$

基本的な性質（１）

偏差の和はゼロである：

$$\sum_{i=1}^n \overbrace{(x_i - \bar{x})}^{\text{偏差}} = 0$$

証明 平均値の定義 $\bar{x} = \frac{1}{n} \sum x_i$ から $\sum x_i = n\bar{x}$

そうすると,

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

基本的な性質（2）分散公式

分散は2乗の平均値から
平均値の2乗を引いたもの

$$s^2 = \overline{x^2} - \bar{x}^2$$

生データの場合：

$$s^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

度数データの場合：

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

証明 生データの場合（度数データの場合は自分で確認）

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n} \sum_{i=1}^n \bar{x}^2$$

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

平均値 $\bar{x}$

$\bar{x}^2$

例題 次の度数分布表から平均, 分散, 標準偏差を求めよ.

階級値 (c_j)	22	24	26	28	30	32	34	36	38	40	42	計
度数 (f_j)	1	4	5	15	18	20	18	11	4	2	2	100

StatData01_2.csv

例題 次の度数分布表から平均, 分散, 標準偏差を求めよ.

階級値 (c_j)	22	24	26	28	30	32	34	36	38	40	42	計
度数 (f_j)	1	4	5	15	18	20	18	11	4	2	2	100
$c_j f_j$	22	96	130	420	540	640	612	396	152	80	84	3172
c_j^2	484	576	676	784	900	1024	1156	1296	1444	1600	1764	
$c_j^2 f_j$	484	2304	3380	11760	16200	20480	20808	14256	5776	3200	3528	102176

$$\bar{x} = \frac{1}{n} \sum c_j f_j = \frac{1}{100} \times 3172 = 31.72$$

$$\begin{aligned} s^2 &= \overline{x^2} - \bar{x}^2 \\ &= 1021.76 - 31.72^2 = \mathbf{15.60} \end{aligned}$$

$$\overline{x^2} = \frac{1}{n} \sum c_j^2 f_j = \frac{1}{100} \times 102176 = 1021.76$$

$$s = \sqrt{\mathbf{15.60}} = 3.95$$

Python を試してみる

(1) データファイル

- ・ **csv (comma-separated values) が基本**

Excel から読み書き可能

- ・ **授業用のデータファイルは Classroom からダウンロードする**

授業のセクションに準じて StatData01_1.csv のような名称になっている

(2) Python のコード例 (Google Colab または Jupyter Notebook)

- ・ **Classroom からダウンロード可能**

データファイル名に準じて StatData01_1.ipynb のような名称になっている

StatData01_1

StatData01_1.csv

生データ/ ローデータ (raw data)

3110	2500	2770	3010	3000	3000	2740	3040	3060	3410
3100	2620	3910	3650	2840	2480	2790	3720	3520	2850
3140	2780	2270	2700	2830	3020	3160	4060	2620	3390
3050	3190	3710	3460	3200	3260	3040	3610	3360	3280
2480	3440	2970	3050	2590	3320	3580	3820	3450	4150
3300	3020	3360	3140	3300	3600	3330	3300	3300	3170
3340	3250	2880	3560	3060	3320	2740	2380	3590	2460
2960	3170	3000	3250	3140	3220	3160	3730	3460	3360
3160	3540	2890	3060	2900	3040	3220	3590	2680	3150
2770	3220	2970	3300	3560	3520	2760	2740	2820	4180

[StatData01_1 - Jupyter Notebook.pdf](#)

StatData01_2

StatData01_2.csv

階級値 (c_j)	22	24	26	28	30	32	34	36	38	40	42	計
度数 (f_j)	1	4	5	15	18	20	18	11	4	2	2	100

[StatData01_2 - Jupyter Notebook.pdf](#)

Lecture 1

おわり