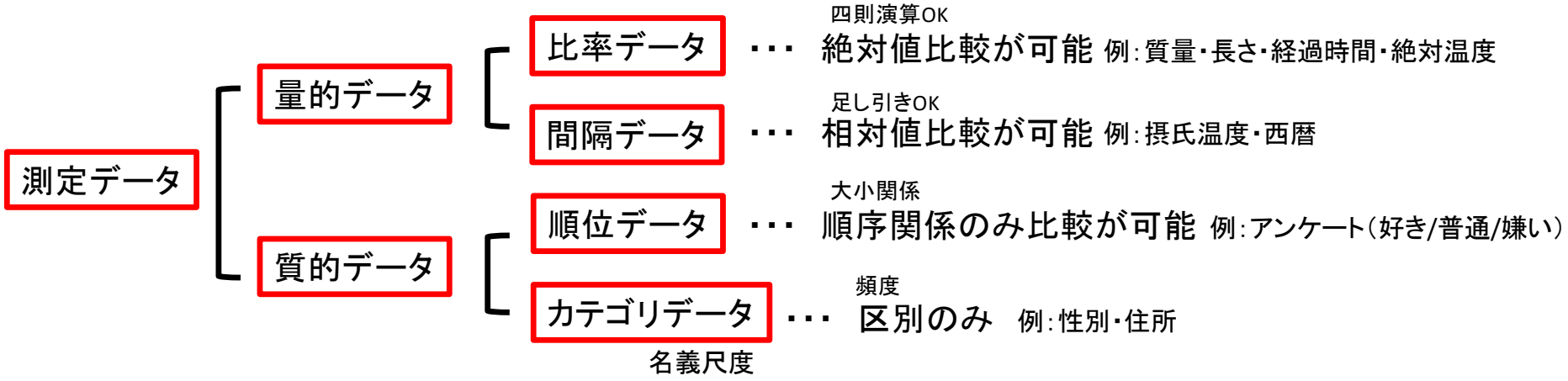
医学統計学
第２回
データの整理:ヒストグラムと確率分布

芳賀　昭弘

:（前期）月曜日6講時　15:30 - 16:30

1-1. 測定尺度

データは、必ず何らかの「尺度」で測定される。数学的水準で分類すると一般に次のように分類できる；



1-2. データの表現

データ

表 1.2　キュウリの収量

ポット番号	栽培法 A(g)	栽培法 B(g)
1	3 063	3 157
2	2 275	2 707
3	2 089	3 270
4	2 855	3 181
5	2 836	3 633
6	3 219	3 404
7	2 817	2 219
8	2 136	2 730
9	2 540	3 408
10	2 263	3 203
11	2 140	2 938
12	1 757	3 286
13	2 499	2 920
14	2 093	3 332
15	2 073	3 478

まとめ

度数分布表（ヒストグラム）

表 1.3　キュウリ収量の度数分布表

総収量(g)	階級値(g)	度数（ポット数）	相対度数(%)	累積相対度数(%)
1 700 以上～2 000 未満	1 850	1	3.3	3.3
2 000 ～ 2 300	2 150	8	26.7	30.0
2 300 ～ 2 600	2 450	2	6.7	36.7
2 600 ～ 2 900	2 750	5	16.7	53.3
2 900 ～ 3 200	3 050	5	16.7	70.0
3 200 ～ 3 500	3 350	8	26.7	96.7
3 500 ～ 3 800	3 650	1	3.3	100.0

可視化

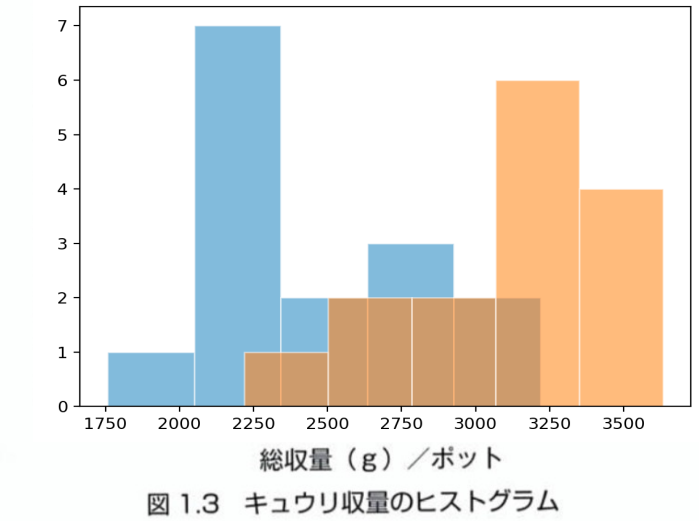


図 1.3　キュウリ収量のヒストグラム

1-2. データの表現

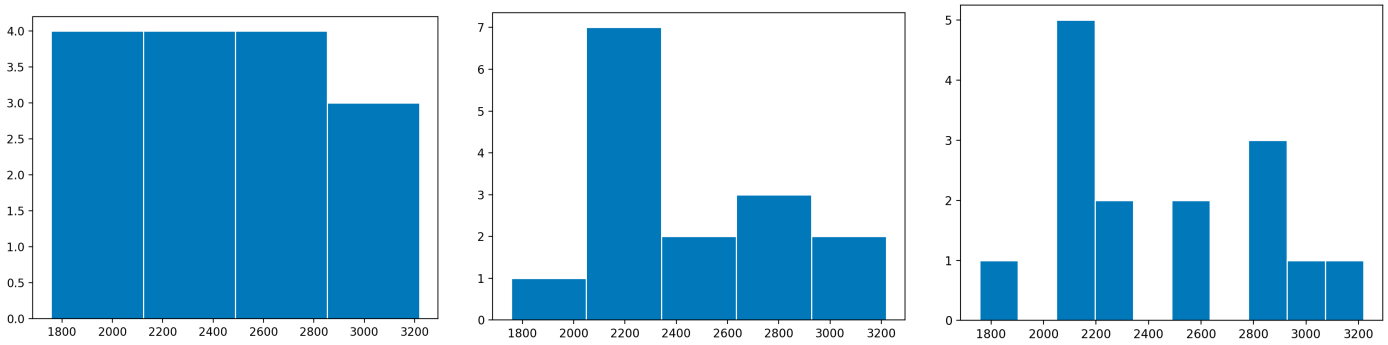
- ヒストグラムを作成することで、データのばらつきが直感的に捉えることができる。
- 一方、区切りの間隔（ビン幅、ビンサイズ）を変えると印象が変わるので注意

データ数をnとすると、ビン幅として

$$\left\lceil \frac{\max(x)-\min(x)}{\kappa} \right\rceil$$

とすることができる（つまり κ をビン幅（ヒストグラムの棒の数）とする）。ビン幅の取り方として

1. $\kappa = \sqrt{n}$
2. $\kappa = 1 + \log_2 n$

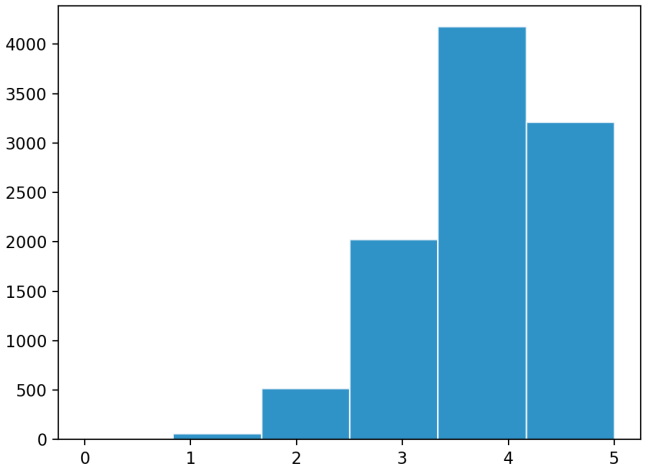


などがある。後者はスタージェスの公式と呼ばれ、後述する二項分布において理論的な根拠を与える。

ヒストグラムと確率

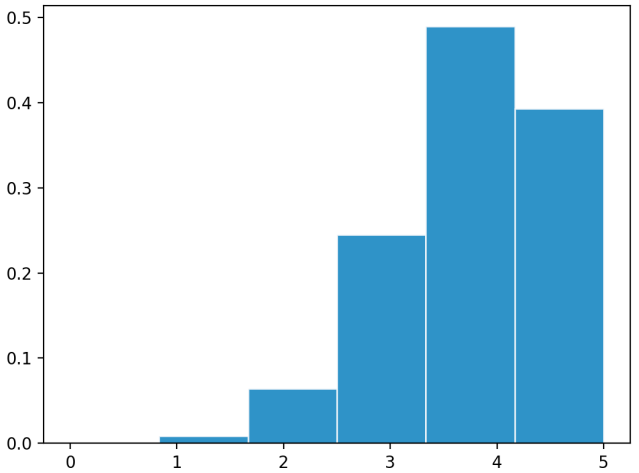
ある病気の患者 5 人のうち 4 人は介護が必要ということが全国調査で分かっているとする。実際にその 病気の患者 5 人を集めてみると、その中に介護が必要な人が何人いるでしょうか？

患者5人を10000グループ集めてみて集計してみました



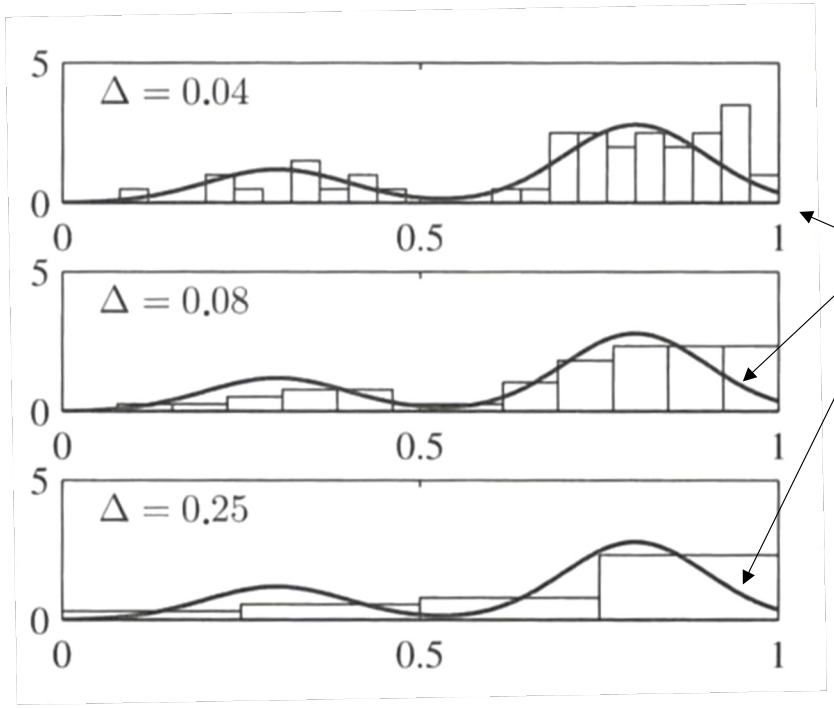
この頻度分布の総和は10000になる

10000で割ろう



確率分布を表している！

確率密度と密度推定



ヒストグラム：　サンプルされたデータが区分分けされた範囲内に入る頻度　（確率分布のこと）

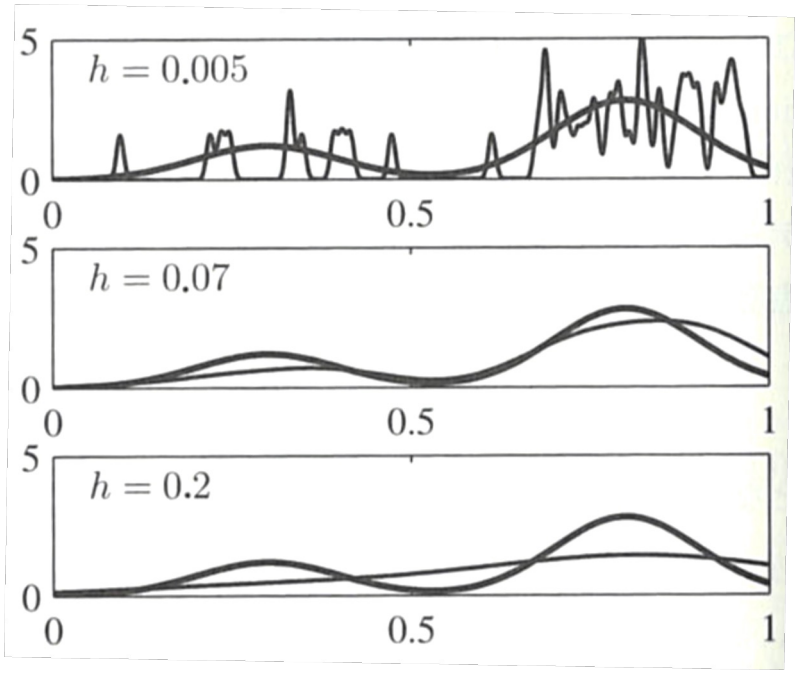
実際の分布が連続だったら？
（「確率密度」という）

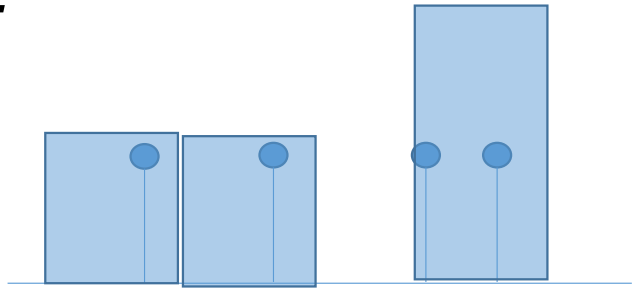
サンプルされて得られたデータから
確率密度分布を推定しよう！

密度推定

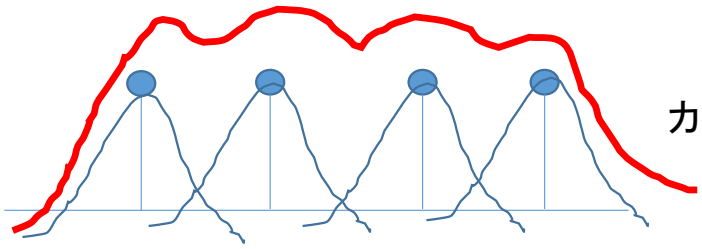
ヒストグラムも密度推定の1つの手法である。
（区分分けした範囲にデータの値が来たらカウントする）

確率密度と密度推定





ヒストグラム密度推定



カーネル密度推定

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$

演習問題　選択肢

1. 比率データ
2. 間隔データ
3. 順位データ
4. カテゴリデータ

として、以下の問題に答えよ。

問題1.
アンケートで「体調が優れている」という問いに対して、回答が
1. 当てはまる, 2. 少し当てはまる, 3. 少し当てはまらない, 4. 当てはまらない
となっている。番号でデータを集めた場合、どの測定尺度であるか？

問題2.
複数の患者の検体から遺伝子変異割合を調べた。このデータは、どの測定尺度であるか？

問題3.
血液のCRP（炎症反応, [mg/dL]）を肝臓がんの患者100名から採取し, ある値以上を
「炎症あり」それ未満を「炎症なし」と分類した。このデータは、どの測定尺度であるか？