

1 データとその表現形式

1.1 数の集まりとしてのデータ

定義 1.1.1 有限個であれ無限個であれ、幾つかの数の集まりをデータと言います。

1.1.1 有限データ

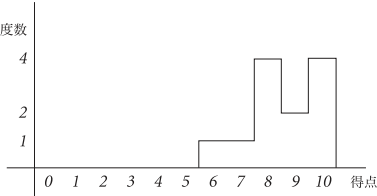
数の羅列 10, 10, 8, 7, 10, 8, 8, 6, 10, 9, 8, 9

度数分布表

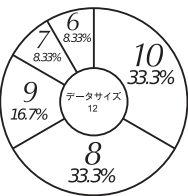
得点	10	9	8	7	6
度数	4	2	4	1	1

相対度数分布表

得点	10	9	8	7	6
相対度数	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{4}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
データのサイズ：12					

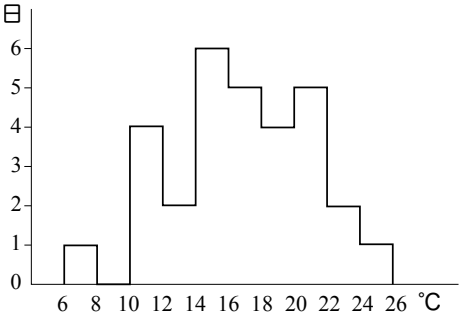


ヒストグラム



円グラフ

階級 (°C)	度数 (日)
6 以上 8 未満	1
8 ~ 10	0
10 ~ 12	4
12 ~ 14	2
14 ~ 16	6
16 ~ 18	5
18 ~ 20	4
20 ~ 22	5
22 ~ 24	2
24 ~ 26	1
計	30

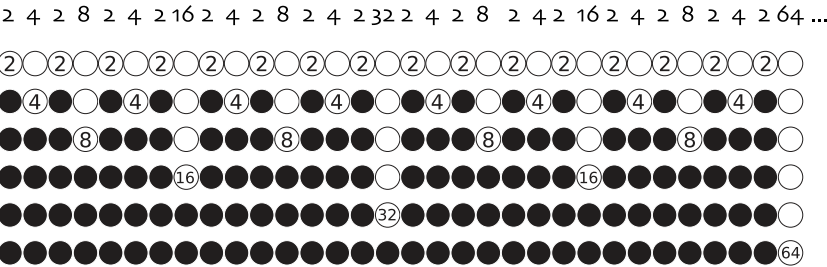


1.1.2 可算無限と非可算無限

正の整数全体のように、1 個、2 個、3 個、・・・と順番に数えていくことで（理屈の上では）全てを網羅できるような無限のことを『可算無限』と言います。

『数えられない』くらい大きな無限のことを非可算無限と呼んでいます。

1.1.3 可算無限データ



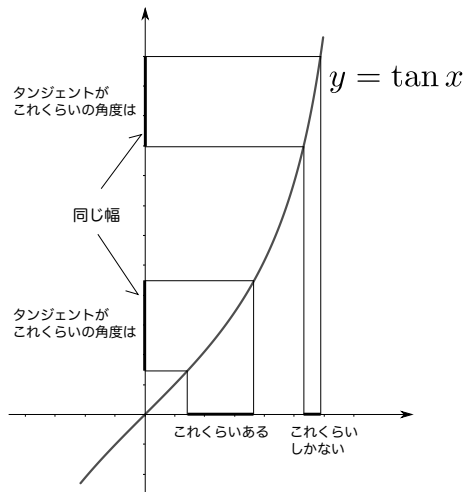
数	2	4	8	16	32	...	2^n	...
度数	∞	∞	∞	∞	∞	...	∞	...

数	2	4	8	16	32	...	2^n	...
相対度数	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	...	$\frac{1}{2^n}$	...

1.1.4 非可算無限データ

$(-\frac{\pi}{2}, \frac{\pi}{2})$ の範囲の全ての角度のタンジェントの値をデータとしましょう。この範囲の角度は非可算無限個ありますからデータのサイズは非可算無限です。

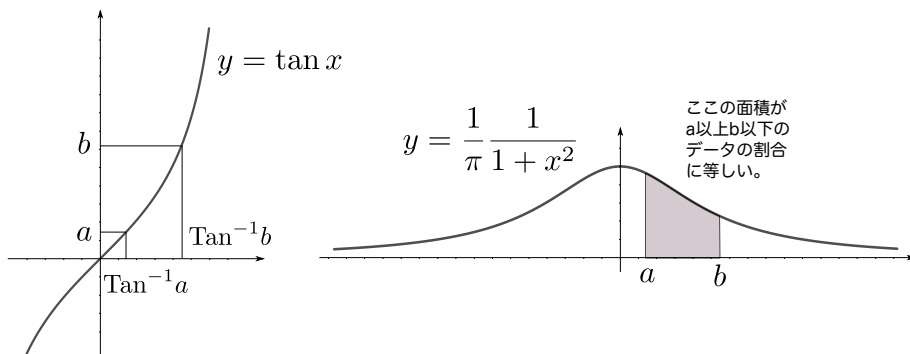
1.1.5 絶対連続データとその密度関数



$$a \leq \tan \theta \leq b \iff \tan^{-1} a \leq \theta \leq \tan^{-1} b$$

(全データ中の a 以上 b 以下の数値の割合) = ($a \leq \tan \theta \leq b$ となるような角度の割合)

$$\begin{aligned} &= \frac{\tan^{-1} b - \tan^{-1} a}{\pi} \\ &= \frac{1}{\pi} [\tan^{-1} x]_a^b \\ &= \int_a^b \frac{1}{\pi} \frac{1}{1+x^2} dx \end{aligned}$$



定義 1.1.2 積分可能な関数 $f(x)$ が存在して、任意の $a \leq b$ に対して

$$(\text{全データ中の } a \text{ 以上 } b \text{ 以下の数値の割合}) = \int_a^b f(x) dx$$

となる時、このデータは絶対連続型であると言い、 $f(x)$ をその密度関数と言います。

絶対連続型データでは、必ず1つの数値単独での相対度数は0になっています。

1.2 確率変数としてのデータ

定義 1.2.1 某かのランダムネスによって様々な値を取りうる関数／変数のようなものを確率変数 (random variable) と呼んでいます。

データの2つの捉え方	
数の集まり	くじ引きの結果
相対度数	確率

定義 1.2.2 確率変数 X があったとき、任意の $a \leq b$ に対して、

$$P[a \leq X \leq b] = \int_a^b f(w) dw \quad (P \text{ は確率を表す})$$

となるような関数 $f(w)$ が存在するならば、これを X の密度関数と言います。

問題演習

基本演習 1.1 次のデータはある鉄道の駅と駅間の距離です：

1.99 2.21 1.45 1.23 1.05 0.79 1.28 0.79 0.95 0.59
 1.09 1.11 0.52 0.76 1.57 0.72 1.12 1.81 1.22 0.86
 1.42 1.32 0.74 1.46 1.21 1.61 1.50 1.19 0.89 (km)

階級を 0.4km から始めて幅 0.2km で区切り、階級分けデータ（同じ階級内のデータは、すべてその階級の中央値であると考えた近似データ）の度数分布表とヒストグラムを作ってください。

基本演習 1.2 1つのサイコロを 5 回振った時に 3 の目が出る回数を表す確率変数を X とします。 X の相対度数分布表（確率分布表）を書いて下さい。ただし 2 項係数 $\binom{n}{k}$ あるいは ${}_nC_k$ は計算せずそのまま結構です。

基本演習 1.3 有限データは密度関数をもたない事を示して下さい。

基本演習 1.4 全ての有理数に 1、2、3、・・・と番号を付けるにはどのようにしたら良いでしょうか。

発展演習 1.5 (1) (可算無限) + (可算無限) は可算無限でしょうか？

(2) (可算無限) + (可算無限) + ... は可算無限でしょうか？

(3) $2^{(\text{可算無限})}$ は可算無限でしょうか？

発展演習 1.6 整数全体 $\mathbb{Z}$ の部分集合は全部で可算無限個でしょうか？

発展演習 1.7 データ／確率変数 X に対して、

$$F(x) = (x \text{ 以下のデータの割合}) = P[X \leq x]$$

で定まる関数 $F(x)$ を X の分布関数と言います。

(1) X の密度関数 $f(x)$ が連続関数であるとき、 $F(x)$ の微分が $f(x)$ である事を示して下さい。

(2) X の密度関数 $f(x)$ が不連続である場合、例えば

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

である場合、 $F(x)$ の微分はどうなるでしょうか？

(3) $F(x)$ が不連続関数になる場合に X の密度関数はどんな風でしょうか？

発展演習 1.8 1つの数値単独での相対度数が 0 でない数値を含むデータの『密度関数』を強引に定義するとしたらどのように考えたら良いでしょうか。

発展演習 1.9 区間 $[0, 2]$ 内の実数全てを 1 つずつ含むデータ X は絶対連続であるかどうか調べ、もしそうなら密度関数を求めて下さい。

発展演習 1.10 全ての実数を 1 つずつ含む偏りのないデータの密度関数はどうなるでしょうか？