

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

有限個の数からなるデータが提示された場合、データを構成する数の総数をデータのサイズ（または大きさ）と言う事にします。

具体的な提示の形式としては単に数の羅列を紙に書き付けたものが基本的ですが、どの数が何個入っているか（＝度数）をまとめて表にした度数分布表や、度数の代わりに全体に占める割合（＝相対度数）を表にした相対度数分布表が便利です。

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

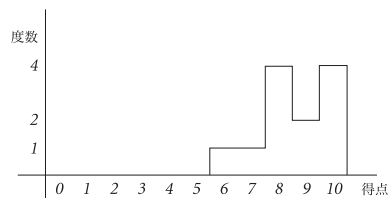
度数分布表

得点	10	9	8	7	6
度数	4	2	4	1	1

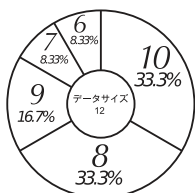
相対度数分布表

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					

またこれらを棒グラフ（ヒストグラム）や円グラフにしたものも使われるでしょう。



ヒストグラム



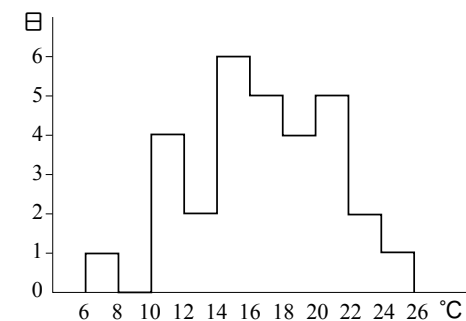
円グラフ

相対度数分布表と円グラフは、各数が全体に占める割合をまとめたものですから例えば 0 と 1 が 1 個ずつのデータでも 0 と 1 が 100 個ずつのデータでも同じになってしまいます。従って相対度数分布表から度数分布表を復元する事は出来ません。この問題を

回避するためにデータのサイズを書き添えておけば度数を復元する事が出来るようになります、他の形式と全く同等になります。

データサイズが大きい場合は、データをいくつかの区間に区切ってその 1 つ 1 つを階級と呼び、各階級にその区間に含まれるデータ数を度数として対応させて度数分布表及びヒストグラムを作成することもあります。

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。

では実数全体は可算無限でしょうか？ つまり全ての実数に『背番号を付ける』ことは出来るでしょうか。

例えば $[0, 1]$ 区間の中の実数全てに番号を付ける事が出来て、0 は特別に最初に置いて $0, r_1, r_2, r_3, \dots$ と並べたしましょう。また、各 r_j を無限小数で表したものを $r_j = 0.r_{j1}r_{j2}r_{j3}\dots$ とします（例えば $\frac{1}{2}$ は 0.5 ではなく $0.4999\dots$ で表します）。

$$r_1 = 0.r_{11}r_{12}r_{13}\dots$$

$$r_2 = 0.r_{21}r_{22}r_{23}\dots$$

$$r_3 = 0.r_{31}r_{32}r_{33}\dots$$

$$\vdots \quad \quad \quad \vdots$$

次に、 r_{11} とは違う 1 から 9 までの数 w_1 を 1 つ取ります。同様に r_{22} とは違う数 w_2 を取り、以降、 r_{jj} とは違う数 w_j を取ってきて並べて出来る無限小数 $w = 0.w_1w_2w_3\dots$

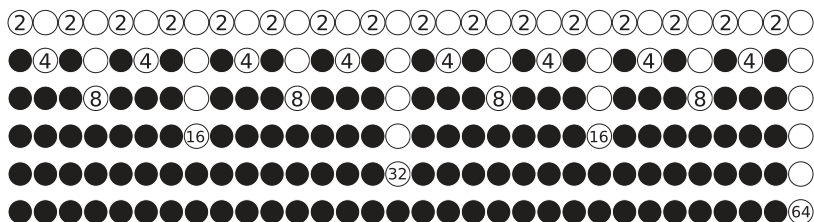
を考えるとこれは必ず $[0, 1]$ 区間内の実数になっており、特に 0 ではありません。また、小数点以下 1 桁目が r_1 とは違いますから w は r_1 とは違う実数です。更に小数点以下 2 桁目が r_2 とは違ってきますからやはり w は r_2 とは違う実数です。同様に、小数点以下 j 桁目が r_j とは違ってきますから w はどの r_j ととも違う実数と云うことになりますね。

これは変ですよ。だって $[0, 1]$ 区間内の実数『全て』を並べたものが $0, r_1, r_2, r_3, \dots$ だったのに、そのどれとも違う $[0, 1]$ 区間内の実数 w が存在してしまいます。これは矛盾です。従って $[0, 1]$ 区間内の全ての実数に $0, r_1, r_2, r_3, \dots$ と番号を付ける事は出来ないと言ふ事が結論されます。

このように、実数は ($[0, 1]$ 区間内ですら) 1 個、2 個、3 個、 $\dots$ と数えた先にあると想像される可算無限よりももっとたくさんあることがわかります。この様に『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

2 4 2 8 2 4 2 16 2 4 2 8 2 4 2 32 2 4 2 8 2 4 2 16 2 4 2 8 2 4 2 64 ...



まず 1 個おきに 2 を並べ、次に空いたところにまた 1 個おきに今度は 4 を並べます。更にまた空いているところに 1 個おきに 8 を並べて $\dots$ これを繰り返して得られる数列を考えます。

明らかにデータのサイズは可算無限ですから、このデータを『羅列』式に全部紙に書く事は人間には出来ませんが、そこは可算無限ですから、『 $\dots$ 』で誤摩化すことは出来ます。

では度数分布表はどうでしょうか。このデータには 2^n の型の整数だけが入っており、どの数も無限個含まれています。従ってこのデータを『度数分布表』で表現しても

数	2	4	8	16	32	$\dots$	2^n	$\dots$
度数	∞	∞	∞	∞	∞	$\dots$	∞	$\dots$

となってしまう要領を得ません。直観的に 2 は 16 より多く含まれているのに『個数』と言うとどちらも無限個ですからその辺りの事情が上手く反映されないのです。この様に無限データに於いては『度数』では上手くデータを表現出来ない事があります。

しかし（このデータに限って言えば）2 は 2 個に 1 個、4 は 4 個に 1 個、8 は 8 個に 1 個入っていますから、相対度数分布表は次のようになります：

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

データの中の数の様子を完璧に表現する事ができます（データのサイズにはあまり意味はありませんので書かなくて良いでしょう）。

勿論いつもこんなに上手くいくわけではなく、度数有限の数と度数無限の数が混在している場合は度数も相対度数も一長一短で必要に応じて使い分けるしかありません。

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。データが非可算無限個ある時点でいわゆる『表』には書けませんし、このようなデータを一体どのように表現したら良いのでしょうか？

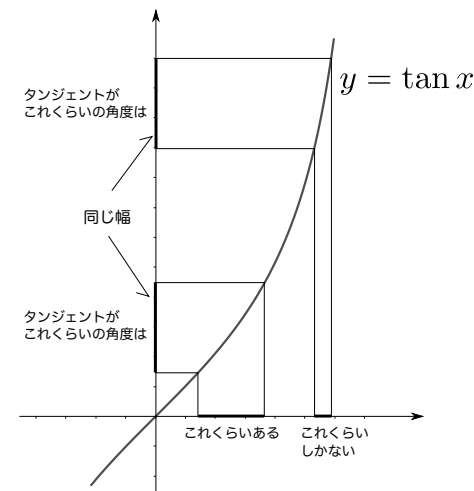
x が $-\frac{\pi}{2}$ から $\frac{\pi}{2}$ まで動くとき $\tan x$ は $-\infty$ から $+\infty$ まで全ての実数値をとります。

また $\tan x$ はこの範囲で単調増加関数ですから、例えばタンジェントの値が 1 となる角度はただ 1 つ $\frac{\pi}{4}$ しかなく、これは全てのタンジェントの値で同様です。

そうするとこのデータの『中身』は、『全ての実数が 1 個ずつ』と云うことになります。

この部分だけを見ると、このデータは偏りのないフラットなデータのように思えますが、右図のように tangent の値が小さい角度の方が tangent の値が大きな角度よりも多く存在します。

このデータの偏りをどうやって表現したら良いのでしょうか。



1.1.5 絶対連続データとその密度関数

このような場合はデータをいくつかの階級に分け、その階級の度数をまとめるのと似たような方法がとられます。ただし、いくら階級に分けても各階級はおそらく区間であり、どんな区間にも非可算無限個の角度のタンジェント値が入りますから、度数を取っても意味がありません。そこで各階級の相対度数を見てみることにします。つまり、たとえば、『 $a \leq \tan \theta \leq b$ となるような角度は全体の何割あるか』を考えてみれば、

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

により、

$$(a \leq \tan \theta \leq b \text{ となるような角度の割合}) = \frac{\text{Tan}^{-1}b - \text{Tan}^{-1}a}{\pi}$$

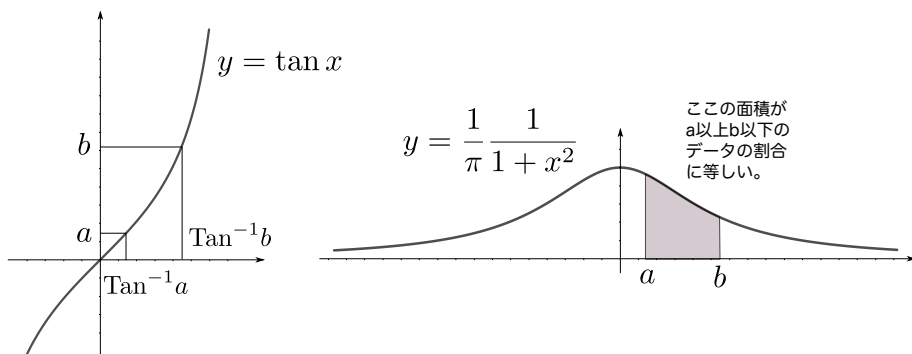
であり、更に

$$\int_a^b \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{1}{\pi} [\text{Tan}^{-1}x]_a^b = \frac{\text{Tan}^{-1}b - \text{Tan}^{-1}a}{\pi}$$

に注意すれば、

$$(\text{全データ中の } a \text{ 以上 } b \text{ 以下の数値の割合}) = \int_a^b \frac{1}{\pi} \frac{1}{1+x^2} dx$$

となっています。この被積分関数は便利ですね。これさえあれば積分することによってデータの中にどの程度の値がどの程度含まれているのか全て分かってしまいます。もちろんデータによってはこのような関数が存在するとは限りませんが、存在する場合はこの関数がデータの分布の全てを表していると考えられます。



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。実際、

$$\begin{aligned} (\text{全データ中の } a \text{ の割合}) &= \lim_{n \rightarrow \infty} \left(a \text{ 以上 } a + \frac{1}{n} \text{ 以下の数値の割合} \right) \\ &= \lim_{n \rightarrow \infty} \int_a^{a+\frac{1}{n}} f(x) dx \quad (\text{密度関数を } f(x) \text{ とする}) \\ &= \int_a^a f(x) dx \\ &= 0 \end{aligned}$$

となってしまうからです。

逆に言えば、1つの数値単独で相対度数が正であるような数値を含む場合は密度関数は存在せず絶対連続ではありません。

今後皆さんが扱う非可算無限データは大抵絶対連続型であり、この密度関数を提示する事によってデータを表現します。

1.2 確率変数としてのデータ

データの数値全体を母集団とし、そこから無作為に1個取り出した（くじ引きだと思って下さい）サンプルの値はランダムに様々な値を取り得るわけですが、この様に

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。

どんな値がどんな確率で出て来るかはデータの中にその値がどの程度の割合で含まれているかで決まります。丁度『相対度数』と『確率』が対応関係にあり、この対応に

よって与えられた数の集まりを（くじ引きの結果としての）確率変数と考える事も出来るし、与えられた確率変数を（その取り得る値の全体としての）数の集まりと考える事も出来ます。

これは1つの対象（＝データ）を別の2つの方向から見ていると考えれば良く、今後はこの2つの見方を自由に行き来しながら話を進めて行きます。

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

数の集まりとしてのデータに X と云う名前をつけたら、断りなく対応する確率変数にも同じ記号 X を使う事にします（逆もあるでしょう）。

密度関数を確率変数の言葉で書けば次の様になります：

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して下さい。

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？

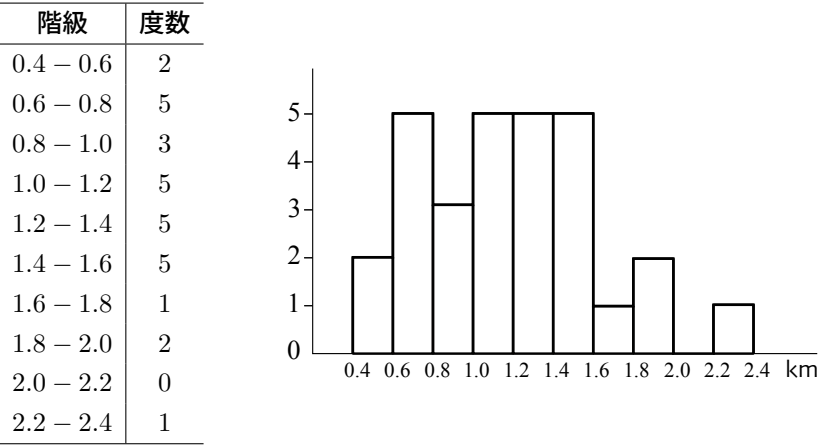
発展演習 1.8 1つの数値単独での相対度数が0でない数値を含むデータの『密度関数』を強引に定義するとしたらどのように考えたら良いでしょうか。

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？

問題演習解答

基本演習 1.1



基本演習 1.2

j	0	1	2
$P[X = j]$	$\left(\frac{5}{6}\right)^5$	$5 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^4$	$\binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$
j	3	4	5
$P[X = j]$	$\binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$	$\binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^1$	$\binom{5}{5} \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^0$

基本演習 1.3 1つの数値単独での相対度数が正である事に注意。

基本演習 1.4

まず 0 をカウントします。

次に、絶対値が 0 より大きく 1 以下の有理数を考えます。これを既約分数（約分し切った状態、分子と分母が互いに素な分数）で表したとき、まず整数になってしまうものは ± 1 のみです。また分母が 2 のものは

$\frac{1}{2}, -\frac{1}{2}$

しかありません。次に分母が 3 のものは

$\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{2}{3}$

が全てです (分子の小さい順に、プラスマイナス交互に並べましょう)。同様に分母を指定すると有限個ずつしかありません。これらを順に並べて行ったものを $r_{11}, r_{12}, r_{13}, \dots$ とします：

$\{r_{1n}\} : 1, -1, \frac{1}{2}, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{2}{3}, \frac{1}{4}, -\frac{1}{4}, \frac{3}{4}, -\frac{3}{4}, \dots$

次に絶対値が 1 より大きく、2 以下の有理数を考えます。同様に既約分数表現して並べて行きます。

$\{r_{2n}\} : 2, -2, \frac{3}{2}, -\frac{3}{2}, \frac{4}{3}, -\frac{4}{3}, \frac{5}{3}, -\frac{5}{3}, \frac{5}{4}, -\frac{5}{4}, \frac{7}{4}, -\frac{7}{4}, \dots$

これを繰り返して全ての有理数が出そろいました：

$$\begin{array}{ccccccc} & 0 & & & & & \\ \{r_{1n}\} & r_{11} & r_{12} & r_{13} & r_{14} & \cdots \\ \{r_{2n}\} & r_{21} & r_{22} & r_{23} & r_{24} & \cdots \\ \{r_{3n}\} & r_{31} & r_{32} & r_{33} & r_{34} & \cdots \\ \{r_{4n}\} & r_{41} & r_{42} & r_{43} & r_{44} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

こうして並べた上で、更に

$0, r_{11}, r_{12}, r_{21}, r_{13}, r_{22}, r_{31}, r_{14}, r_{23}, r_{32}, r_{41}, \dots$

とすれば番号が付けられますね！

発展演習 1.5

- (1) Yes 交互に並べれば良いでしょう。
- (2) Yes 1, 1, 2, 1, 2, 3, 1, 2, 3, 4, ... がヒント。先の発展演習 1.3 の最後の数え方参照。
- (3) No 各項が 0 か 1 であるような数列全体を考えてみましょうか。全部並べたとすると矛盾が出て来る筈。

発展演習 1.6 上の (3) と同様ですね。

発展演習 1.7

- (1) 積分表示を微分するだけ。
- (2) 微分可能性を見ると良いでしょう。
- (3) 存在しない。

☐

発展演習 1.8 色々な考えがあって良いでしょう。デルタ関数とか。

☐

発展演習 1.9 絶対連続であり、密度は

$$\begin{cases} \frac{1}{2} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}.$$

☐

発展演習 1.10 存在しない。

☐