データの活用（add.*確率）

代表値

データに含まれる一つ一つの**標本**値をもとに、データ全体を代表する値として**平均値**（mean）がよく用いられます。通常使われるのは**算術平均**で

$$\boxed{(\text{算術平均}) = \frac{(\text{標本値の総和})}{(\text{標本数})}}$$

と**定義**されています。要するに、全部足して割るのです。算術平均は**相加平均**ともいいます。

わざわざ平均を算術平均と呼ぶのは、別に**幾何平均**（または**相乗平均**）があり、それは

$$\boxed{(\text{標本数 } n \text{ の幾何平均}) = \sqrt[n]{(\text{標本値の総積})}}$$

です。記号 $\sqrt[n]{\ }$ は平方根を一般化したもので n **乗根**と読みます。たとえば $\sqrt[3]{8}$ の意味は

$$\text{同じ数を 3 個掛けて 8 になる、すなわち } \overbrace{\square \times \square \times \square}^{3\text{ 個}} = 8 \text{ となる}\square$$

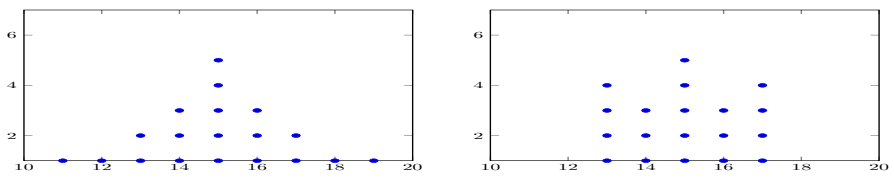
を表す数です。したがって $\sqrt[3]{8} = 2$ です。「$\square^2 = a \quad \Leftrightarrow \quad \square = \sqrt{a}$」と対比させるとよいでしょう。ちなみに記号 $\sqrt{\ }$ は、記号 $\sqrt[2]{\ }$ の省略形です。

代表値は他にも、データを大きさの順に並べたとき、標本数が奇数ならちょうど中央に位置する値を、標本数が偶数なら中央に位置する 2 値の平均を、**中央値**（median）と呼びます。また、データの中で最も多く出現する値を**最頻値**（mode）と呼ぶことは授業で扱っているはずです。

以上の事柄は単純に計算したり調べたりするだけなので、扱いに慣れればすぐに身につくことでしょう。試験でよい成績を得ることが目的なら、目標を達成しやすい分野です。でも、平均などの代表値は生活にも密着しているので、普段から使い慣れることが望ましいと思います。

度数分布

代表値は簡便で分かりやすいのですが、全体を“よく”表しているかというと少し心もとないものがあります。実際、データの分布が



のようであったなら、分布の様子は異なるのに、平均値・中央値・最頻値はいずれも 15 です。

この例はデータの**範囲**も小さく標本数も少ないので、データを表やグラフに起こすのは簡単です。しかし、全校生徒の身長のデータを表やグラフにすることは、標本数が多いことの他に標本値が細かく設定されているため、平均値・中央値・最頻値を求めるのは骨が折れます。コンピュータがなければ工夫が要るでしょう。

そこで、一定の範囲にあるデータはひとまとめに考えて、たとえば 165.0cm 以上 170.0cm 未満は一つのまとまりにして

	階級	(階級値)	度数	(階級値) × 度数
	⋮	⋮	⋮	⋮
160.0 以上	165.0 未満	(162.5)	42	6825.0
165.0 –	170.0	(167.5)	76	12730.0
170.0 –	175.0	(172.5)	47	8107.5
	⋮	⋮	⋮	⋮
	合計		200	33709.0

のような**度数分布表**にしておくと、**ヒストグラム**に起こせます。また、(階級値) × 度数 の合計から、平均値 $\dfrac{33709.0}{200} \approx 168.55$cm を得ることもできます。

度数分布表から求めた平均値は、一人一人の総和を総人数で割った平均値とは異なるでしょうが、それほど見当違いな**誤差**は出ないはずです。このことから、**階級**の範囲をこの例とは異なるように区切っても、似通った結果になることが期待されます。そもそも平均値は代表値の一つであって、そんなに精緻な値ではないのです。

上の表において、たとえば身長の測定日に 188cm の生徒が一人欠席したため、200 人のデータになったとします。本来は 201 人であり、188 の階級値は 187.5 なので

$$(201\text{ 人の平均値}) = \frac{33709.0 + 187.5}{201} = 168.64(\text{cm})$$

となって、200 人の平均値より約 0.1(cm) 高くなります。

もう少し極端な例を挙げるなら、A 組、B 組とも 30 人のクラスを考えてみましょう。数学の前回試験の平均点がどちらも 55 点だったとします。それで今回の試験は、A 組は前回同様の出来だったため平均点は 55 点でした。一方 B 組は、前回平均点並みの二人の生徒が失敗して 30 点でした。このとき B 組の平均点は

$$\frac{55 \times 28 + 30 \times 2}{30} \approx 53.3\,(\text{点})$$

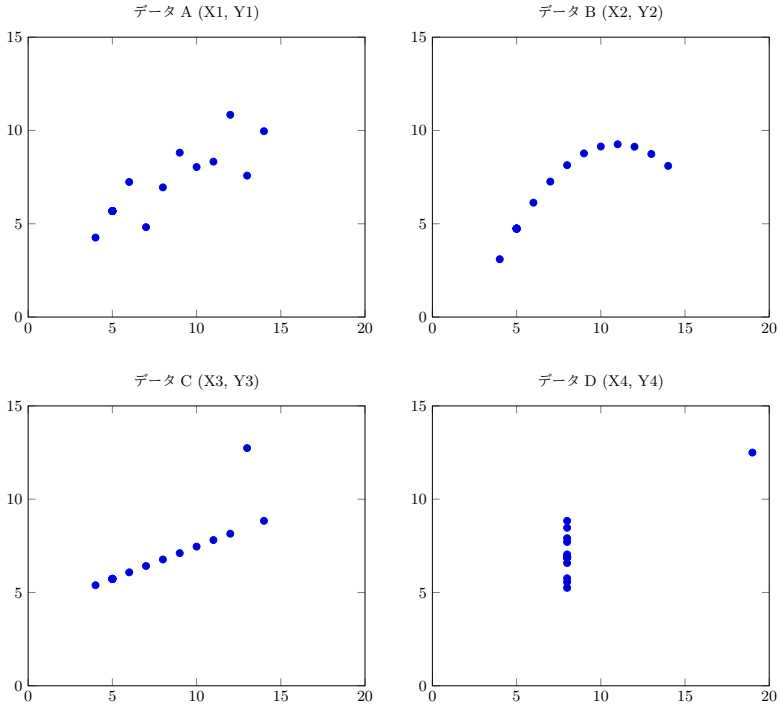
であり、2 点近くクラス平均を下げる結果となります。このように、わずか二人の値が平均より大幅に振れるだけで、全体の平均に少なからず影響します。クラス平均が 10 点近く異なればクラスの差は明らかですが、数点の差であれば一部の**外れ値**が影響した可能性があるものです。

■アンスコムの 4 つ組

アンスコム[*1]の 4 つ組として知られる、いかにも造りました的なデータを紹介します。

項目	A		B		C		D	
	X1	Y1	X2	Y2	X3	Y3	X4	Y4
データ	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
平均	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
標準偏差	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
相関係数	0.82		0.82		0.82		0.82	

平均・**標準偏差**・**相関係数**がすべて同じなのに、実際の分布はまるで異なっています。



とくに、相関係数は X、Y の間にどの程度の関連があるかを示す数値ですが、この例からも関連の様子を相関係数だけに頼るのはよくありません。また相関関係は、互いに関連があるかどうかを表すだけで、実際の**因果関係**は何

[*1] フランク・アンスコム：統計学者。

一つ教えてくれないことを注意しておきます。

コンピュータの利用

さて、せっかくのデータを活用したくても、データの分析は一筋縄ではいきません。とくに標本数が多い場合は、コンピュータの助けなしでは立ち行かなくなるでしょう。平均値・中央値・最頻値ならまだしも、受験に際しよく目にする**偏差値**の計算は手間がかかります。コンピュータのソフトウェアには優れたものが多いので積極的に利用したいものです。

場合によってはプログラミング言語を使って、自分でプログラムを書くことがあるかもしれません。ここでは、データ処理をコンピュータでするとよい、という話をしているので『データ処理のためにプログラミング言語を学ぶ』ことが明確です。そうであれば、何かしらのプログラミング技能が身につく可能性は高いでしょう。しかし、単にプログラム言語を学びたい、というのではおそらく身につく可能性は低いと思います。

プログラム言語とは単に手段に過ぎないので、目的がはっきりしないまま学ぶことは避けましょう。

手段だけ学ぶことは、たとえば外国語の辞書を使って単語や文法だけを学ぶようなものです。それでは会話や文章を読む力はつきません。普通は、会話がしたいとか文章を読みたいとかの**目的**があって外国語を学ぶのです。プログラミング言語も、データ処理をしたいとかゲームを作りたいとかの目的がないと、ただ言語の**命令**と**書式**などを学ぶだけになって、意欲が減退するだけです。私のような "言語フェチ" ならともかく、プログラミング言語を身につけたければ、まず何をしたいかをはっきりさせましょう。

プログラミングはほとんどの人が挫折すると言われています。かりに目的がはっきりしていてプログラミングの勉強を始めたとしても、はじめのうちは**命令**や**構文**がすぐに浮かばないものです。このとき『どうやって覚えたらよいのか？』と聞くようでは、お先真っ暗です。おそらく学校の勉強も、授業で習ったことを覚えて試験で回答するという勉強をしてきたのでしょう。

プログラミングに限らず、勉強は暗記するものではありません。とくに数学はそうです。暗記ではなく、

分からなければ前に戻って調べ、練習問題をたくさん解くこと

です。公式も最初は教科書などの例題を見ながら真似ていればよいのです。そうすれば知らぬ間に覚えてしまうのですから。

プログラミングも同様で、

とにかくコードを書いて、コンピュータで実行すること

です。すると、そのうち正しいコードが書けるようになっています。

偏差値について

偏差値とは

ある集団において、平均値からどれくらい離れているかを表す相対的な数値

のことです。偏差値は平均が 50 になるように調整するので、50 より上なら平均より上、50 より下なら平均より下であることは分かります。しかし、偏差値 52 は "必ず" 偏差値 48 より上とは限りません。なぜなら偏差値は、繰り返しになりますが

ある集団における、相対的な数値

だからです。集団が異なれば、数値どうしの関連はないのです。

たとえば、A さんと B さんが模擬試験で得た偏差値がそれぞれ 52 と 48 だった場合、同じ模擬試験を受験したのであれば間違いなく A さんの成績が上です。しかし、二人がまったく別の模擬試験を受験したのであれば、なんとも言えません。A さんと B さんとでは、属する集団が異なるからです。もし、A さんが難易度がかなり低い模試を受験した上での 52 で、B さんが難易度がかなり高い模試を受験した上での 48 ならば、B さんの実力が上であることは十分考えられます。偏差値の数値だけを聞いて判断するのは意味がありません。

具体的に、度数分布の最初の例で用いた 2 種類のデータ

A: 11, 12, 13, 13, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 17, 17, 18, 19

B: 13, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 17, 17, 17, 17

を使って、コンピュータで検証してみます。ソフトウェアの使い方には触れず、結果のみ示します。

A: 平均 $= 15$、標準偏差 $= 1.946657054$、値 16 の偏差値 ≈ 55.1

B: 平均 $= 15$、標準偏差 $= 1.414213562$、値 16 の偏差値 ≈ 57.1

標準偏差はデータのばらつきの度合いを一つの数値で表したものです[*2]。最初に示した二つのグラフを見比べて、A のばらつきの方が B のばらつきより大きいことから、A の標準偏差の方が B のそれより大きな値にな

[*2] 一般には、データのばらつきの度合いを一つの数値にしたものは**分散**という。しかし分散の単位は計算の性質上、(データの単位)^2 になってしまうため、データの単位にするために平方根で表し、それを標準偏差という。このことと偏差値の計算式から、偏差値は無単位の値となる。

るのです。

そして、ある値 d の偏差値は

$$\boxed{(d\text{の偏差値}) = 50 + \frac{d - (\text{平均値})}{(\text{標準偏差})/10}}$$

と定義されるので、同じ 16 の値であっても偏差値が異なっているのです。ちなみに、偏差値は 25 から 75 程度に収まることが多いものの、ばらつきが小さい集団で、極端に平均から離れた値はその限りではありません。偏差値が -10 とか 105 などはありえます。

*確率の定義

確率の定義は、起こりうるすべての**事象**の数を N、その内ある事象が起こる**場合の数**を a とすると、ある事象が起こる確率 p は

$$\boxed{p = \frac{a}{N}}$$

で定義されることは知っているでしょう。授業などでは（ある事象が起こる確率）$= \dfrac{(\text{ある事象の数})}{(\text{事象の総数})}$ のように、ことばで示していたかもしれません。式の意味を理解することは大事なことですが、文字の意味を理解した上で文字式を使う方が便利だと思います。ここでは、p は probability（確率）、N は必ず自然数（a natural number）ですが総数の雰囲気を出すために大文字で、a は単にいくつかの数を表す文字、という意味で用いました。

*場合の数の数え方

確率の計算が**場合の数**を数えるものだとすると、場合の数さえ数えられればよいことになりますが、それが案外難しいのです。たとえば、ハンバーガーショップでハンバーガーとドリンクを注文するとしましょう。

ハンバーガー
チーズバーガー
フィッシュバーガー
テリヤキバーガー

ドリンク
コーラ
シェイク
ジュース
コーヒー

単品
ポテト
ナゲット

表のように、ハンバーガーは 3 種類、ドリンクは 4 種類の場合、たとえば

> チーズバーガーに対してはコーラ、シェイク、ジュース、コーヒーのいずれかが選べるので、選択肢は 4 通り。同じことがフィッシュバーガーとテリヤキバーガーにもいえるので、ハンバーガー 3 種類の

各々にドリンクの選択肢が 4 通りあるので、3×4（通り）　（※）

のように考えて、すなわち 12（通り）の組み合わせ方があることになります。この考えは**積の法則**と呼ばれ、数え上げの基本となるものです。

積の法則が分かれば、それぞれ何通りかある複数のものの組み合わせ方は、単に掛け算をすればよいことになって、ハンバーガーと単品の組み合わせなら $3 \times 2 = 6$（通り）であることが即座に求められます。

そうなると、積の法則を公式のように使って場合の数を求めることは容易になりますが、ものごとは何でも単純化すればよいわけではないのです。慣れればとくに考えることなく積の法則は使えますが、やはり背景に（※）のような

あるものに対して a 通りの選択肢があり、そのようなものが n 個あるなら場合の数は $n \times a$（通り）

ということが理解できていなければなりません。機械的に掛け算をしていると、間違いなく行き詰まるときが訪れます。気をつけましょう。

*確率の計算

高校から先の確率の勉強は、さまざまな関係式や公式を用いて解くことが多くなります。しかし確率の基本は $\frac{(\text{ある事象})}{(\text{全事象})}$ なので、事象の数を数えることが大事です。中学校で確率を学ぶ場合は、これがすべてでしょう。

すると、問題は数え方になります。ものを数えることは、対象が数多いものだと正確に数えることは難しいでしょう。それでも数える基本は**順列**と**組合せ**です。

・順列：　順番を考慮して数える
・組合せ：順番を無視して数える

という違いがあります。単純な例では、4 種類のドリンク A, B, C, D から

1. 異なる 2 種類を（2 個を）買って、持ち帰る
2. 異なる 2 種類を（2 個を）買って、兄弟に渡す

の場合、1. は組合せ、2. は順列です。1. は単に 2 個のドリンクを買って、たとえば袋に入れてもらうのだから、袋の中でドリンクの優先順位はありません。したがって袋の中が

(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)

となる 6 通りの選択方法があります。

一方 2. は、たとえば (A, B) を買っても ($A_{兄}$, $B_{弟}$) と ($A_{弟}$, $B_{兄}$) は別の選択と考えなくてはなりません。すると、先の 6 通りの組合せのそれぞれに 2 通りが生じるので、$6 \times 2 = 12$ 通りの選択方法となります。

そこで問題となるのが、どこで順列と組合せを見分けるかということです。数学が苦手な人は、何か公式でもないか探すのですが、そんなものはありません。実は、先の例は少し分かりづらい表現をしています。1. は持ち帰って兄弟に渡すのか？　2. はまとめて兄弟に渡すのか、それぞれに渡すのか？　まとめて兄弟に渡すなら、自分は順番を気にする必要はないし、それぞれに渡すなら順番が生じます。

もちろん試験問題であれば、あやふやな表現はできる限り排除するでしょうが、それでも判断に迷う表現はあります。試験問題で判断に迷うなら、それはあなたの練習不足の可能性が高いでしょう。確率などの問題文には、習慣的に順列と組合せを区別する表現があるものです。だったら、その表現を覚えればよいかというと、そうではないのが難しいのです。