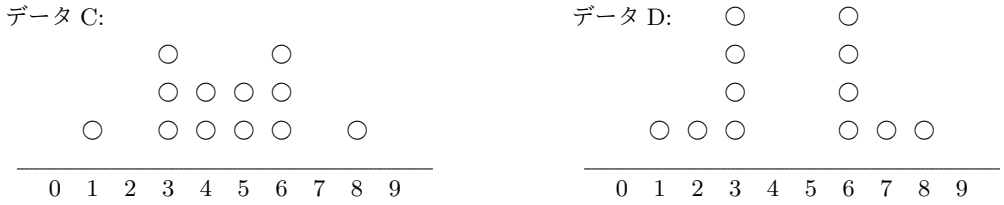
偏差、分散

算術平均、中央値、最頻値、四分位数などに差異が見られないデータは、何を根拠に差別化できるのだろうか。再び、かなり作為的なデータ例ではあるが、検討してみよう。



グラフは、値 x が○の数だけあることを示している。データ C、データ D 共にデータ数は 12、最小値 1、最大値 8 で、

平均値／中央値 := 4.5、最頻値 := 3 と 6、第一・四分位数 := 3、第三・四分位数 := 6

であることはすぐに確認できることだろう。しかし、明らかにデータ C は中央寄りの数値が多く、データ D は左右に寄っている。この見た目の違いを数値化することがひとつの目標である。

それには、各値が中心からどれだけ離れているか見ればよい。データ D は、データ C に比べて中心から離れた値が多い。たとえば、値 1 や値 8 は平均値 4.5 から 3.5 だけ離れている。このとき、値 1 や 8 の**偏差**は 3.5 である、という。これら偏差の総和を求めてみよう。

$$\begin{aligned}
&\text{データ C:} \quad \{|1-4.5|+|3-4.5|\times 3+|4-4.5|\times 2\} \quad \times 2 = \quad 18\\
&\text{データ D:} \quad \{|1-4.5|+|2-4.5|+|3-4.5|\times 4\} \qquad \times 2 = \quad 24
\end{aligned}$$

計算上 2 倍しているのはグラフが左右対称だからで、本来は各値について絶対値をとるものである。値が全体的に中心から離れているデータ D の方が、データ C より大きな数値となるので指標としては機能しているとみてよい。

しかし、中心からの離れ具合を正の数で表すなら、絶対値より平方する方が扱いが簡単である。なおかつ、平方することで離れ具合が強調されることもあって、この場合は

$$\begin{aligned}
&\text{データ C:} \quad \{(1-4.5)^2+(3-4.5)^2\times 3+(4-4.5)^2\times 2\} \quad \times 2 = \quad 39\\
&\text{データ D:} \quad \{(1-4.5)^2+(2-4.5)^2+(3-4.5)^2\times 4\} \qquad \times 2 = \quad 55
\end{aligned}$$

とすればよい。この計算方法で求めた値を**偏差平方和**と呼ぶ。

ただし、偏差平方和だけではばらつきの物差しとして不適当である。なぜなら、データ数が多ければ値が大きくなるのが当然だからだ。たとえば、値 4 と 5 だけが数百個あるデータは、データ D

よりばらつきは少ないように思える。ひとつひとつの偏差は 0.5 であるから、偏差平方は 0.25 でかなり小さい。しかし、データ数が数百あれば偏差平方和は軽く 50 を超えるだろう。この不具合を解消するには、データ 1 個あたりの数値に直すのがよい。つまり、平均をとればよいのである。この値を**分散**と呼び、ばらつきの指標としよう。

n 個のデータ $d_1, d_2, d_3, \ldots, d_n$ の算術平均が m のとき

$$\text{分散} = \frac{(d_1-m)^2+(d_2-m)^2+(d_3-m)^2+\cdots+(d_n-m)^2}{n}$$

標準偏差

数学ではあまり単位について気にしないものだが、データのばらつきを数値化した分散には少し困った点がある。平方和を用いているので単位も平方されることだ。仮に先ほどの x の値が 1cm, 3cm, ... であったら、データ C、D のそれぞれの分散は 3.25cm^2、4.58cm^2 となって、長さのばらつきを測るために面積の値を用いていることになる。それは少々変な話だ。そこで、単位を揃えるために平方根をとって、その値をばらつきを測る物差しとしておこう。それは、それぞれ $\sqrt{\dfrac{39}{12}} \approx 1.80$、$\sqrt{\dfrac{55}{12}} \approx 2.14$ であり、この値を**標準偏差**と呼ぶ。

$$\text{標準偏差} = \sqrt{\frac{(d_1-m)^2+(d_2-m)^2+(d_3-m)^2+\cdots+(d_n-m)^2}{n}}$$

*　　*　　*

早速、公式めいた式が登場したが、丸暗記するにはちょっと複雑な式に見えるかもしれない。公式というものは暗記して使うものではなく、使っているうちに覚えるものである。もし、いつまでたっても覚えられないとしたら、それは使う量が不足しているためだ。公式を覚えたければ、公式を使う練習問題を大量に解こう。

しかし、単に解いているだけでは記憶に残らない。使うときは意味や仕組みを理解しながら解く必要がある。たとえば平均は

$$\text{平均} := \text{値の総和を、値の個数で割る}$$

と理解していれば、何も $(d_1+d_2+d_3+\cdots+d_n)/n$ などという式を呪文のように覚えることはない。実は標準偏差も同じだ。

まず、各値と平均の差（正の値にするため絶対値で考える）は偏差であった。すなわち

$$\text{偏差} := |\ \text{各値} - \text{平均値}\ |$$

である（後で 2 乗するので | | がなくても困らないが）。分散は、偏差平方和の平均だから

$$\text{分散} := \text{偏差平方和、の平均}$$

である。そして、分散の平方根をとったものが標準偏差なので

$$\text{標準偏差} := \sqrt{\text{分散}} := \text{偏差平方和、の平均、の平方根}$$

となる。いたって単純であろう。▌

平均、分散、標準偏差の関係

ここから話は数学っぽくなる。それに合わせて使用する文字も x を中心に据えることにする。

まず、n 個のデータ $x_1, x_2, x_3, \ldots, x_n$ を考える。平均値 $\overline{x}$ は

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \tag{1}$$

である。$\overline{x}$ を用いて分散 s^2 を表すと

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

となる。分散の記号が s^2 であるのは、分散の計算式が平方和であるからなのだが、標準偏差 $= \sqrt{\text{分散}}$ を逆に見て（標準偏差）^{2} = 分散 であることからきている。ちなみに英語で、標準偏差は standard deviation、分散は variance である。

ところで、分散の計算は式自体が複雑なのだが、実際に何らかの値を使って計算すると殊の外（ことのほか）大変であったはずだ。たとえば前出のデータ C やデータ D では、(値 − 平均値)^2 がすべて $\triangle.25$ になったはずだ。そのときはデータが意図的に作られていたので、計算は比較的容易だったかもしれないが、現実的は平均値が $65.4333\cdots$ などとなることは珍しくないものだ。で、その都度 $(\square - 65.4333\cdots)^2$ を計算する？

そこで、ちょっと工夫をしてみよう。

$$\begin{aligned}
s^2 &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} \\
&\quad *\text{平方を展開して…} \\
&= \frac{\{x_1^2 - 2x_1\overline{x} + (\overline{x})^2\} + \{x_2^2 - 2x_2\overline{x} + (\overline{x})^2\} + \cdots + \{x_n^2 - 2x_n\overline{x} + (\overline{x})^2\}}{n} \\
&\quad *\text{それぞれの第 1 項どうし、第 2 項どうし、第 3 項どうしをまとめて…} \\
&= \frac{(x_1^2 + x_2^2 + \cdots + x_n^2) - 2(x_1 + x_2 + \cdots + x_n)\overline{x} + \{(\overline{x})^2 + (\overline{x})^2 + \cdots + (\overline{x})^2\}}{n} \\
&= \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - 2\frac{x_1 + x_2 + \cdots + x_n}{n}\overline{x} + \frac{n \times (\overline{x})^2}{n} \\
&\quad \text{ここで、(1) を用いると…}
\end{aligned}$$

$$= \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - 2(\overline{x})^2 + (\overline{x})^2$$
$$= \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - (\overline{x})^2$$

意外と簡単な式に変形ができて、内容的には

分散 :=（平方和、の平均）−（平均値、の平方）

である。これなら、たとえ平均値が $65.4333\cdots$ などとなっても計算しやすいだろう。とくに、データが整数値なら楽勝に違いない。

ところで、x_1 から x_n までの和を n で割った平均を $\overline{x}$ で表しているが、$\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}$ は、x_1^2 から x_n^2 までの和を n で割った平均であるから、$\overline{x}$ に倣（なら）って $\overline{x^2}$ と書くことにしたい。すると、分散の式はより簡単に書ける。

$$\boxed{\text{分散 } s^2 = \overline{x^2} - (\overline{x})^2}$$

*　　*　　*

現代では電卓が使えるといっても、大量のデータについて平均値や標準偏差を求めるのは大変なことである。その点、Microsoft Excel には統計にまつわる関数が備わっているので、標準偏差も関数式ひとつで済む。ところが、いざ関数を使おうとすると

分散 (VARiance)	標準偏差 (STandard DEViation)
VAR	STDEV
VAR.P	STDEV.P
VAR.S	STDEV.S
VARA	STDA
VARP	STDP
VARPA	STDPA

のようにたくさんの種類があって、どれを使えばよいのか迷ってしまうものだ。結論から言えば、VAR.P と STDDEV.P を使えばよい。前節で使ったデータ C やデータ D に対して使ってみるとよいだろう。▌

大きな値のデータ、小さな値のデータ

データには様々な種類のものがあるので、たとえば家電製品の値段を集めると数千円から数万円のように大きな数値が並ぶだろう。また、100m 走の記録を集めると小数点以下 2 位程度の小さな数値が並ぶに違いない。もちろん、平均値や標準偏差などはデータの大きさに関係なく求められる

のだが、計算機がなければ 2 乗の計算が大きな値や小さな値をより際立たせてしまい、計算がより困難になろう。

平均値であれば、一旦 1/100 倍や 100 倍にすることで、さらにはそこから一定数を引くことで計算をしやすくできる。実際の平均値は、引いた一定数を加えて 100 倍なり 1/100 倍なりすればよいからだ。具体的に 32100, 56400, 78900 の 3 個のデータを考えると、実際の平均は

$$\frac{32100+56400+78900}{3}=55800$$

である。この場合は数値が大きいので、まず 100 で割ってから 500 を引くことにすると、データは $-179, 64, 289$ となる。この平均は 58 であるが、データを変換したときの逆をたどって、500 を足してから 100 倍すると、ちゃんと 558000 となる。

きちんとした理屈を与えるなら、平均の求め方が $\overline{x}=\dfrac{x_1+\cdots+x_n}{n}$ であることから、データの各値を a で割って d を引くと、新たな値の平均 $\tilde{x}$ は

$$\tilde{x}=\frac{(\frac{x_1}{a}-d)+\cdots+(\frac{x_n}{a}-d)}{n}=\frac{\frac{1}{a}(x_1+\cdots+x_n)-n\times d}{n}=\frac{1}{a}\overline{x}-d$$

である。したがって、$\overline{x}=a(\tilde{x}+d)$ が分かるのである。

では、標準偏差はどうだろう。データの各値を a で割って d を引いたとすると、

$$\text{各値は}\quad \frac{x_1}{a}-d,\ \ldots,\frac{x_n}{a}-d,\qquad \text{平均は}\quad \frac{\overline{x}}{a}-d$$

であるから、標準偏差 $\tilde{s}$ は

$$\begin{aligned}\tilde{s} &= \sqrt{\frac{\{(\frac{x_1}{a}-d)-(\frac{\overline{x}}{a}-d)\}^2+\cdots+\{(\frac{x_n}{a}-d)-(\frac{\overline{x}}{a}-d)\}^2}{n}}\\ &= \sqrt{\frac{\frac{1}{a^2}(x_1-\overline{x})^2+\cdots+\frac{1}{a^2}(x_n-\overline{x})^2}{n}}=\frac{1}{|a|}\sqrt{\frac{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}\end{aligned}$$

であるから、本来の標準偏差の $\dfrac{1}{|a|}$ となる。したがって、各値を a で割って d を引いたデータを用いて求めた標準偏差 $\tilde{s}$ から、本来の標準偏差を求めるには、$|a|$ 倍すればよく、引いた値には無関係であることが分かる。

> n 個のデータ $x_1, x_2, \ldots, x_n$ の平均を $\overline{x}$、標準偏差を s とする。
> $x_i \to \dfrac{x_i}{a}-d$ と変換したときの平均 $\tilde{x}$、標準偏差 $\tilde{s}$は、
> $\tilde{x}=\dfrac{1}{a}\overline{x}-d$、　　$\tilde{s}=\dfrac{1}{|a|}s$

先ほどの 3 個のデータを例にして、Microsoft Excel で確かめてみよう。

◇	A	B	C	D	E	F
1	32100		(※ C1)			
2	56400		↓下へコピーする			
3	78900		↓			
4						
5	(※ A5)		(※ C5)			
6						

※ セルの式
(C1)　=A1/100-500
(A5)　=STDEV.P(A1:A3)
(C5)　=STDEV.P(C1:C3)

C5 セルに表示された標準偏差の値は、A5 セルに表示された標準偏差の 1/100 であることが分かる。