初心者のためのデータ分析法

データの評価方法

データサイエンス研究所

データの評価方法

英語の研修前後の成績（20人）

	A君の得点	全体の平均点	得点-平均点
研修前	70	58.3	+11.7
研修後	72	58.3	+13.7

- A君の成績の評価は？

全員の成績

＜研修前＞

70、56、89、27、69、57、69
50、33、67、37、49、98、69
68、25、65、67、33、68

＜研修後＞

72、31、95、36、89、88、89
76、28、47、23、28、96、48
51、20、33、91、27、98

データを見る視点は？

全員の成績

＜研修前＞

70、56、89、27、69、57、69
50、33、67、37、49、98、69
68、25、65、67、33、68

＜研修後＞

72、31、95、36、89、88、89
76、28、47、23、28、96、48
51、20、33、91、27、98

平均の信頼性

（例）暑い日に暑い場所で待ち合わせをした。
いつも遅れてくる人が何分後に来るのか予測。

＜過去10回の遅刻データ＞

											平均
①	21	46	8	28	19	34	13	33	19	31	25.2分
②	25	26	23	24	26	27	26	25	24	26	25.2分

①、②それぞれにおける待ち時間の行動は？

①のデータは「バラツキ」が大きい。

「バラツキ」の大きいデータの平均は信頼できない。

A君の成績の評価

全体の平均点は同じ。
自分の得点は上がった。

研修前　3番（20人中）
研修後　9番（20人中）

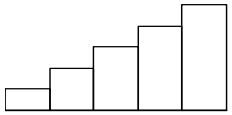
研修前より研修後のバラツキが大きい

成績を比較するためには各データ集団の
バラツキを表す代表値が必要！！

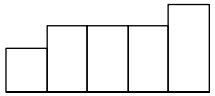
平均値が同じデータ群

A群～D群の平均値はすべて同じ値3

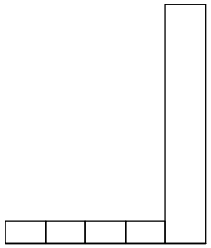
A: {1,2,3,4,5}



B: {2,3,3,3,4}



C: {1,1,1,1,11}



D: {3,3,3,3,3}

平均値は同じであるので同類集団？

各群のバラツキが異なる。　D<B<A<C

バラツキの計算方法の検討

A群のバラツキの部分を抽出する。

A群		平均値		値－平均値
1	－	3	＝	-2
2	－	3	＝	-1
3	－	3	＝	0
4	－	3	＝	1
5	－	3	＝	2
		計	＝	0

合計すると0！

バラツキの部分を2乗し、合計！

A群のバラツキの部分を2乗して合計

	値－平均値	（値－平均値）^{2}
	-2	4
	-1	1
	0	0
	1	1
	2	4
計	0	10

A群のバラツキ＝10

同様に、B、C、D群について計算する。

B群	平均	差	$(差)^2$
2	3	-1	1
3	3	0	0
3	3	0	0
3	3	0	0
4	3	1	1
計		0	2

C群	平均	差	$(差)^2$
1	3	-2	4
1	3	-2	4
1	3	-2	4
1	3	-2	4
11	3	8	64
計		0	80

D群	平均	差	$(差)^2$
3	3	0	0
3	3	0	0
3	3	0	0
3	3	0	0
3	3	0	0
計		0	0

D<B<A<C　　0<2<10<80

「偏差平方和」

E群: {1,1,2,2,3,3,4,4,5,5} の偏差平方和

E群		平均値		差	(差)^2
1	-	3	=	-2	4
1	-	3	=	-2	4
2	-	3	=	-1	1
2	-	3	=	-1	1
3	-	3	=	0	0
3	-	3	=	0	0
4	-	3	=	1	1
4	-	3	=	1	1
5	-	3	=	2	4
5	-	3	=	2	4
計				0	20

➡ E群の偏差平方和=20

E群: {1,1,2,2,3,3,4,4,5,5} の偏差平方和＝20

A群: {1,2,3,4,5} の偏差平方和＝10

A群とE群の構造は同じであるが、
偏差平方和の値はE群の方が大きい。

偏差平方和をデータ数で割ると同じ値となる。

E群：20÷10＝2　　A群：10÷5＝2

バラツキの代表値①
「分散」

F群: {10,20,30,40,50}　　A群: {1,2,3,4,5} の10倍

F群の単位：千円　A群の単位：万円の場合、全く同じデータ。

偏差平方和を計算すると

F群	平均		差	$(差)^2$
10	− 30	=	-20	400
20	− 30	=	-10	100
30	− 30	=	0	0
40	− 30	=	10	100
50	− 30	=	20	400
計			0	1000

F群の偏差平方和： 1000

F群の分散： 200千円^2

A群の分散： 2万円^2

分散の値の比較は困難

分散の平方根は、同じ値となる。

A群：$\sqrt{2}$ ≒ 1.414万円

F群：$\sqrt{200}$ ≒ 14.14千円

⇒ バラツキの代表値②「標準偏差」

VAR.P（分散）

STDEV.P（標準偏差）

A君の成績の評価

	成績	平均	標準偏差
研修前	70	58.3	19.2
研修後	72	58.3	28.6

研修前　　研修後

$$\frac{70-58.3}{19.2} = 0.609 > \frac{72-58.3}{28.6} = 0.479$$

$$\frac{\text{成績と平均値の違い}}{\text{標準偏差}} \Rightarrow \text{Z値}$$

Z値の比較によりデータの評価・比較が可能

Z値×10＋50

偏差値

研修前の偏差値＝0.609×10＋50＝56.09

研修後の偏差値＝0.479×10＋50＝54.79