データの整理

この世はデータで溢（あふ）れている。データは、漫然と見ているだけでは自分をゴミの山に埋もれさせるのと変わらない。ゴミの山に見えるものは、実際は宝の山である。ゴミの山を宝の山に変えるには、少々手を加えてやる必要がある。手を加えるほどに宝は輝きを増すのだが、まずは汚れを落とす程度のことから始めよう。

データ A:

7	13	14	16	18	18	19	19	19	
20	21	21	22	22	22	23	23	23	24
24	24	25	25	25	25	25	26	27	28
28	28	29	29	29	30	30	34	38	

データ B:

7	7	10	12	12	13	13	14	15	20
20	20	21	21	21	22	22	22	23	23
25	25	25	25	26	27	27	27	28	28
28	32	32	33	33	35	35	36	37	38

データ A とデータ B は、後の説明に都合のよいように選んだ架空のデータである。目的は、これらのデータの差異を探すことである。

平均—とくに**算術平均**—は、数多いデータの性質を表すのにもっともよく利用される。データの値の総計をデータ数で割るだけだから計算は容易である。算術平均とことわったのは、他に**幾何平均**と呼ばれる平均値があるからだ。

$$\text{算術平均} = \frac{d_1 + d_2 + d_3 + \cdots + d_n}{n}$$

$$\text{幾何平均} = \sqrt[n]{d_1 \cdot d_2 \cdot d_3 \cdot \cdots \cdot d_n}$$

データを整理するときに幾何平均を用いることは稀（まれ）だろうから、計算方法を提示するだけにとどめることにする。

さて、データ A とデータ B について算術平均を求めておこう。全部足してデータ数の 38 または 40 で割れば、平均はいずれも 23.5 であることが分かる。平均で見る限り 2 種類のデータに差異はない。

算術平均は、極端に小さいまたは極端に大きい値が含まれると、それらの値に引っ張られて実態を見えにくくすることがある。日本人の所得の平均値などは、極端に所得が大きい人々のために実

態より多く見積もられる傾向にあるのは、そのよい例となっている。そのようなときは、**中央値**や**最頻値**を使うとよい場合がある。

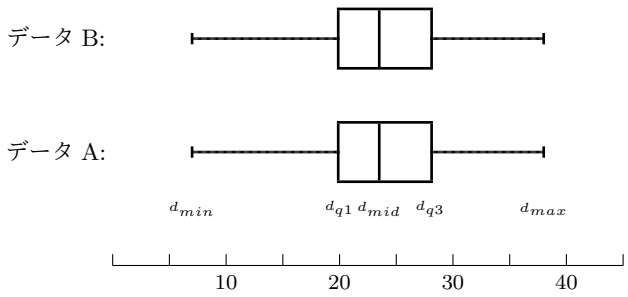
> 中央値 := データのちょうど中央にあたる値
> 最頻値 := データに最も多く含まれる値

データ A はデータ数が 38、データ B はデータ数が 40 であるから、ちょうど中央にあたる位置はそれぞれ 19, 20 番目の間、および 20, 21 番目の間となってしまう。このようなときは、それらの平均をとって中央値とする。したがって、中央値はいずれも値 24 である。また最頻値は、データ A については値 25 が 5 個出現し、データ B についても値 25 が 4 個出現し最も多い。したがって、最頻値はいずれも 25 である。ちなみに、値 25 の他のデータ d_i も同じ個数出現していたら、最頻値は 25 と d_i の 2 種類となる。

ここまでに調べた**代表値**で差異が認められないのは、もちろんそうなるようにデータを選んでいるからで、他に**最小値**も**最大値**も意図的に差異がないようにしてある。

データの見える化

データを見比べる方法は単に計算で代表値を求めるだけでなく、グラフ化すると性格がつかみやすくなるものだ。グラフ化のひとつに**箱ひげ図**がある。まず、データ A とデータ B について箱ひげ図で表してみよう。



まったく同じグラフになったのは、選んだデータのせいである。自然なデータを用いた場合、大抵は差異が目立つ箱ひげ図になるのが普通だ。箱ひげ図は、データの代表値（最小値 d_{min}、**第一・四分位数** d_{q1}、中央値 d_{mid}、**第三・四分位数** d_{q3}、最大値 d_{max}）を線分上に示し、第一・四分位数から第三・四分位数まで厚みを持たせたものである。

第一・四分位数とは中間位置より下位のグループにおける中央値で、第三・四分位数とは中間位置より上位のグループにおける中央値である。データ A もデータ B も中間位置は 20 番目と 21 番目の間であるが、データ A は下位、上位共にデータ数は 19 なので、それらのグループの中間位置はそれぞれ 10 番目と 29 番目である。よって、グループ A の第一・四分位数は 10 番目の値 20、第三・四分位数は 29 番目の値 28 である。

一方、データ B は下位、上位共にデータ数は 20 なので、それらのグループの中間位置はそれぞれ 9, 10 番目の間と 29, 30 番目の間である。よって、グループ B の第一・四分位数は 9, 10 番目の値の平均をとって 20、第三・四分位数も 29, 30 番目の値の平均をとって 28 である。意図的なのが見え見えであろう。

* * *

これまでに示された代表値やグラフは、Microsoft Excel で計算または表示することができる。G 列を図示していないが、そこには C 列と同様の式が記述される。

◇	A	B	C	D	E	F
1	データ A:				データ B:	
2	7	平均値	(※ **C2**)		7	平均値
3	13	最小値	(※ **C3**)		7	最小値
4	14	第一・四分位数	(※ **C4**)		10	第一・四分位数
5	16	中央値	(※ **C5**)		12	中央値
6	18	第三・四分位数	(※ **C6**)		12	第三・四分位数
7	18	最大値	(※ **C7**)		13	最大値
8	19	最頻値	(※ **C8**)		13	最頻値
9	19				14	
10	19				15	
11	20				20	

※ セルの式

(**C2**)	**=AVERAGE(A2:A39)**	(**G2**)	**=AVERAGE(E2:E41)**
(**C3**)	**=MIN(A2:A39)**	(**G3**)	**=MIN(E2:E41)**
(**C4**)	**=QUARTILE(A2:A39,1)**	(**G4**)	**=QUARTILE(E2:E41,1)**
(**C5**)	**=MEDIAN(A2:A39)**	(**G5**)	**=MEDIAN(E2:E41)**
(**C6**)	**=QUARTILE(A2:A39,3)**	(**G6**)	**=QUARTILE(E2:E41,3)**
(**C7**)	**=MAX(A2:A39)**	(**G7**)	**=MAX(E2:E41)**
(**C8**)	**=MODE(A2:A39)**	(**G8**)	**=MODE(E2:E41)**

実際にやってみると、データ A の第一・四分位数と第三・四分位数が、さっき求めた値と微妙に異なる数値になっているはずだ。データ B については同じ数値が表示されていることだろう。原因は、Excel がおそらく次のような計算をしていることにあると思われる。

Excel では、第一・四分位数とはデータの 1/4 の順位にある値のことのようだ。データ数が 38 の場合、その順位は 1 番目と 38 番目の 1/4 にあたる順位、すなわち 1 番と 38 番を 1 : 3 に内分する位置と考えられる。よってその順位は、$\dfrac{3\times 1_{番}+1\times 38_{番}}{1+3}=10.25$ 番となる。いま、データ A の 10 番目の値は 20、11 番目の値は 21 であるが、10.25 番というのは 10 番と 11 番を 1 : 3 に内分する順位と解釈し、そこで得られるであろう値も 10 番目の値と 11 番目の値を 1 : 3 で加重平均した値と捉えている。したがって、$\dfrac{3\times 20+1\times 21}{1+3}=20.25$ となるのであろう。

同様に、第三・四分位数はデータの 3/4 の順位であるから、$\dfrac{1\times 1_{番}+3\times 38_{番}}{3+1}=28.75$ 番となり、28 番目の値 27 と 29 番目の値 28 を加重平均して $\dfrac{1\times 27+3\times 28}{3+1}=27.75$ なのだろう。いかにも計算機的処理と言えよう。

データ B についてはデータ数が 40 であるから、第一・四分位数の順位は 10.75 番、第三・四分位数の順位は 30.25 番である。しかしデータ B では、10, 11 番目は共に値 20、30, 31 番目も共に値 28 であるため、加重平均をとっても変わらない結果となっている。

また、Excel では箱ひげ図も簡単に描ける。データの範囲を選択した上で、メニューからグラフの挿入に関する項目を探って、箱ひげ図のアイコンを選べば一丁上がりだ。ただし、アイコンも実際のグラフも、箱ひげ図は “縦置き” になっているので注意しよう。▌

ヒストグラム

これまでに調べた方法では、データ A とデータ B の差異を見つけることはできなかった。理由は、そうなるようにデータを選んでいたからだが、データの分布を可視化すれば違いがはっきりする。データの羅列を少し整理しておこう。次の表は**度数分布表**である。度数分布表においては、データ範囲の幅を**階級**、出現回数を**度数**と呼んでいる。

データ A:

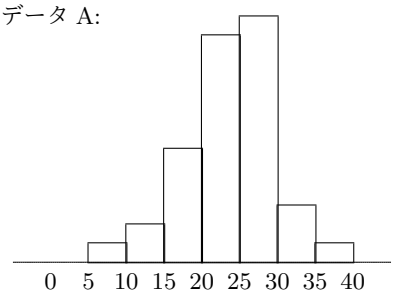
階級	度数
0 以上 – 5 未満	0
5 以上 – 10 未満	1
10 以上 – 15 未満	2
15 以上 – 20 未満	6
20 以上 – 25 未満	12
25 以上 – 30 未満	13
30 以上 – 35 未満	3
35 以上 – 40 未満	1
合計	38

データ B:

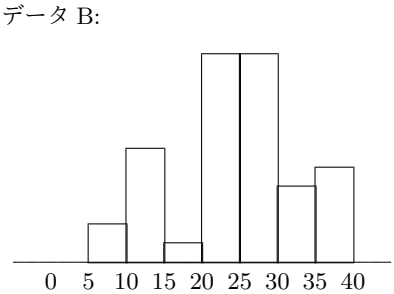
階級	度数
0 以上 – 5 未満	0
5 以上 – 10 未満	2
10 以上 – 15 未満	6
15 以上 – 20 未満	1
20 以上 – 25 未満	11
25 以上 – 30 未満	11
30 以上 – 35 未満	4
35 以上 – 40 未満	5
合計	40

これだけでも少しデータの分布が異なっていることが見てとれよう。そこで、もう一手間かけてグラフ化しておこう。グラフは**ヒストグラム**と呼ばれる。

データ A:



データ B:



こうなればデータ A とデータ B との差異は明白である。最初からヒストグラムを描けばよかったと思うかもしれない。しかし、グラフというものは作成に手間がかかるものである。Excel では簡単に描けるとはいえ、Excel は数値処理の方が優れているものだ。つまり、グラフで見える差異を、どのような数値で表すかが課題なのである。