シグマ

Σの世界 2009

ヒグマ

立教大学社会学部社会学科
社会学データ実習　副読本

立教大学教授
松本　康

【はじめに】

社会調査法の勉強にとって、統計学は避けて通れません。そして、統計学の初歩を勉強する場合に、総和を表すΣ記号を避けて通ることはできません。

しかし、大学受験で「数学」を避けることができるために、大学に入学してから、Σに出会って、とまどう人が少なくありません。

じっさいには、受験数学でもΣはあまりお目にかかることはなく、文系・理系にかかわらず、最初から勉強し直すようなものです。このプリントでは、統計学に必要な範囲で、Σの計算法を解説します。

まず平均値で考えましょう

【例１】

一番なじみやすい例は、平均値の計算です。表 1 のデータについて平均値を求めましょう。　小学生なら....

表１

ID番号	得点
1	70
2	70
3	56
4	70
5	49
6	35
7	49
8	49
9	21
10	49
11	63
12	56
13	56
14	63
15	63
16	56
17	70
18	63
19	63
20	70

(70+70+56+70+49+35+49+49+21+49
+63+56+56+63+63+56+70+63+63+70) ÷ 20 ＝ 57.05

と計算します。平均値の求め方は、
「全員の得点を足して人数で割る」ですね。

【例２】

データの数は 20 でした。この数は変えずに、一般型を考えましょう。各人の得点は、表２のように表現されます。

ID 番号 1 の得点は、x_1、
ID 番号 2 の得点は、x_2、
…
ID 番号 20 の得点は、x_{20}。

表２

ID番号	得点
1	X1
2	X2
3	X3
4	X4
5	X5
6	X6
7	X7
8	X8
9	X9
10	X10
11	X11
12	X12
13	X13
14	X14
15	X15
16	X16
17	X17
18	X18
19	X19
20	X20

そこで、平均値の計算は、

$$(X_1+X_2+X_3+X_4+X_5+X_6+X_7+X_8+X_9+X_{10}+X_{11}+X_{12}+X_{13}+X_{14}+X_{15}+X_{16}+X_{17}+X_{18}+X_{19}+X_{20}) \div 20 \quad \cdots①$$

となります。

このとき、各データの一般的表記は「X (ID番号) 」ですが、これを x_i と表すことにします。

平均値の計算は、

x_i を i＝1 から 20 まで足して 20 で割ることになりますので、Σ記号を使うと、

$$\frac{1}{20}\sum_{i=1}^{20} x_i \quad \cdots②$$

となります。

この②式は、①式と同じ計算を示しています。

【例 3】

ところで、データの数は 20 とは限りません。そこで、データの数を n としてさらに一般型を考えると、平均値の計算式は、

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1+x_2+....+x_n) \quad \cdots③$$

となります。②式の「20」が、ここでは「n」に一般化されています。

【総和を表す記号Σ】

ここで改めて、総和を表す記号Σについて確認しておきます。平均値の計算から離れて、Σだけに注目すると、③式の n で割らないかたちが見えてきます。

$$\sum_{i=1}^{n} x_i=(x_1+x_2+...+x_n)$$

これがΣの一般的用法です。ここで一段落。

末政ひかる作「たれぱんだ」

【応用】

さて、ここで応用です。表 2 にもうひとつ別の得点を加えて、データを拡充した場合を考えましょう。表 3 の 1 列目と 2 列目は、表 2 と同様です。表 3 では 3 列目が付け加えられています。

表 3

ID番号	得点X	得点Y
1	x1	y1
2	x2	y2
3	x3	y3
4	x4	y4
5	x5	y5
6	x6	y6
7	x7	y7
8	x8	y8
9	x9	y9
10	x10	y10
11	x11	y11
12	x12	y12
13	x13	y13
14	x14	y14
15	x15	y15
16	x16	y16
17	x17	y17
18	x18	y18
19	x19	y19
20	x20	y20

ここで、得点 X の平均値の計算結果を $\bar{x}$ （エックスバーと読みます)、

得点 Y の平均値の計算結果を $\bar{y}$ と表すことにします。

言うまでもなく、X については、

$$\bar{x}=\frac{1}{20}\sum_{i=1}^{20}x_i \qquad \cdots②（先述）$$

一般型は、

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad \cdots③$$

となります。ここまでは、復習。

それでは、Y については、どう表現されますか？

$$\bar{y}=\frac{1}{20}\sum_{i=1}^{20}y_i=\frac{1}{20}(y_1+y_2+\ldots+y_{20})$$

一般型は、

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}(y_1+y_2+\ldots+y_n)$$

となりますね。**xがyに代わった**だけです。

ここでは、平均値の計算結果を、バーを使って表す記号法も導入しました。

では、以上を踏まえて、次のステップに進みましょう。

ダルマ

【平均からの差】

表 1 の各データの「平均からの差」を考えます。この集団の平均値は、57.05 でした。

表 4　平均からの差

ID番号	得点	平均値	平均からの差
1	70	57.05	12.95
2	70	57.05	12.95
3	56	57.05	−1.05
4	70	57.05	12.95
5	49	57.05	−8.05
6	35	57.05	−22.05
7	49	57.05	−8.05
8	49	57.05	−8.05
9	21	57.05	−36.05
10	49	57.05	−8.05
11	63	57.05	5.95
12	56	57.05	−1.05
13	56	57.05	−1.05
14	63	57.05	5.95
15	63	57.05	5.95
16	56	57.05	−1.05
17	70	57.05	12.95
18	63	57.05	5.95
19	63	57.05	5.95
20	70	57.05	12.95

各自の「得点－平均」を「平均からの差」と定義すると、平均値はだれにとっても同じですから、表 4 のように、各自の「平均からの差」が計算されます。

たとえば、ID 番号 1 の人にとっては、平均からの差は、70 － 57.05 ＝ 12.95 となります。

ID 番号 3 のように、平均よりも低い得点の持ち主にとっては、計算結果は負の数になります。

ここで問題です。平均からの差の合計（総和）は、いくつになるでしょう？

直感的に「0」と答えられる人は、かなりセンスがいいです。とりあえず、電卓かエクセルで計算して、0 となることを確かめましょう。

では、なぜ 0 になるのでしょうか？　もういちど、表 4 を見てください。
「平均からの差」の総和は、第 4 列を縦に足したものです。
ところが、そもそも「平均からの差」は、「得点－平均値」ですから、「平均からの差」の総和は、「得点の総和（第 2 列の総和)」－「平均値の総和（第 3 列の総和)」に等しくなります。

ここで、「平均値の総和」とは、57.05 × 20 ですから、「得点の総和」と等しくなります。なぜなら、もともと平均値（57.05）は、得点の総和を 20 で割ったものだからです。

これを一般的に表現すると、どうなりますか？
平均からの差を d、得点を x、サンプル数を n とすると、

$$d_i = x_i - \bar{x}$$

「平均からの差」の総和は、

$$\sum_{i=1}^{n} d_i = d_1+d_2+\ldots+d_n = (x_1-\bar{x})+(x_2-\bar{x})+\ldots+(x_n-\bar{x}) = \sum_{i=1}^{n}(x_i-\bar{x})$$

※左辺に　$d_i = x_i-\bar{x}$　を代入すれば、右辺が得られる。

$$= (x_1+x_2+\ldots+x_n)-(\underbrace{\bar{x}+\bar{x}+\ldots+\bar{x}}_{n\text{個}}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}$$

$$= \sum_{i=1}^{n} x_i - n\bar{x}$$

$$= n\bar{x}-n\bar{x}=0 \qquad ※$$

$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ より、　$n\bar{x} = \sum_{i=1}^{n} x_i$　となることに注意。…（※）

よって、　$\sum d_i = 0$　…④

（Σの上下にある、i=1、n は、しばしば省略される）

以上から、「平均からの差」の総和は、つねに 0 になることが証明されました。
したがって、「平均からの差」の平均も 0 となります。

$$\bar{d} = \frac{1}{n}\sum d_i = 0 \quad \ldots⑤$$

【変動・分散・標準偏差】

データの分布のばらつきを見るのに、「平均からの差」の二乗をもとにした統計量が使われます。

変動は、「平均からの差」の二乗の総和。

分散は、「平均からの差」の二乗の平均（つまり、変動を n で割ったもの）。

標準偏差は、「平均からの差」の二乗の平均の平方根（つまり分散の平方根）。

各ケースの得点（データ）を x_i, データの個数を n、変動を SS、分散を S^2、標準偏差を S として、式で表すと、

変動： $SS_x=\sum d_i^2=\sum(x_i-\bar{x})^2$

分散： $S_x^2=\frac{1}{n}\sum(x_i-\bar{x})^2$

標準偏差： $S_x=\sqrt{S_x^2}=\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}$

【変動】

変動は、平均からの差の自乗の総和である。すべての観測データが平均に一致している場合には、変動は 0。平均値から離れたデータが多ければ多いほど、変動は大きくなるが、最大値は不定。（平均からの差を一辺とする正方形の面積の総和）。

変動の公式を展開すると... $\sum(x_i-\bar{x})^2=\sum x_i^2-n\bar{x}^2$ となる。…⑥

【証明】中学で習った $(a+b)^2=a^2+2ab+b^2$ の展開公式を利用して、

$$\sum(x_i-\bar{x})^2=\sum(x_i^2-2x_i\bar{x}+\bar{x}^2)=\{(x_1^2-2x_1\bar{x}+\bar{x}^2)+(x_2^2-2x_2\bar{x}+\bar{x}^2)+....+(x_n^2-2x_n\bar{x}+\bar{x}^2)\}$$

$$=\{(x_1^2+x_2^2+...+x_n^2)-2(x_1\bar{x}+x_2\bar{x}+...+x_n\bar{x})+(\underbrace{\bar{x}^2+\bar{x}^2+...+\bar{x}^2}_{n個})\}$$

$$=\sum x_i^2-2\bar{x}\sum x_i+n\bar{x}^2$$

$\sum x_i=n\bar{x}$ だから（*）

$$=\sum x_i^2-2\bar{x}\cdot n\bar{x}+n\bar{x}^2$$

$$=\sum x_i^2-n\bar{x}^2$$

【分散】

分散は、平均からの差の自乗の平均、つまり、変動を n で割ったもの。すべての観測データが平均に一致している場合には、分散は 0。平均値から離れたデータが多ければ多いほど、分散は大きくなるが、最大値は不定。(平均からの差を一辺とする正方形の面積の平均)。

⑥の両辺を n で割ると、 $$\frac{1}{n}\sum(x_i-\bar{x})^2=\frac{1}{n}\sum x_i^2-\bar{x}^2 \quad \cdots⑦$$

という式が得られる。

【標準偏差】

標準偏差は、分散の平方根。(平均からの差を一辺とする正方形の面積の平均を、正方形の一辺の長さに変換して表したもの)。

$$S_x=\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}=\sqrt{\frac{1}{n}\sum x^2{}_i-\bar{x}^2} \quad \cdots⑧$$

【問１】　表 4 のデータにもとづき、変動、分散、標準偏差を計算しなさい。

表５

ID番号	得点	平均からの差	…の自乗
1	70	13.0	167.7
2	70	13.0	167.7
3	56	−1.1	1.1
4	70	13.0	167.7
5	49	−8.1	64.8
6	35	−22.1	486.2
7	49	−8.1	64.8
8	49	−8.1	64.8
9	21	−36.1	1299.6
10	49	−8.1	64.8
11	63	6.0	35.4
12	56	−1.1	1.1
13	56	−1.1	1.1
14	63	6.0	35.4
15	63	6.0	35.4
16	56	−1.1	1.1
17	70	13.0	167.7
18	63	6.0	35.4
19	63	6.0	35.4
20	70	13.0	167.7
平均	57.05		
		変動	3065.0
		分散	153.2
		標準偏差	12.4

結果は、左の表 5 のようになる。各自、確かめよ。

【問題２】　次の表 6 のデータにもとづき、変動、分散、標準偏差を計算し、問 1（表 4 のデータ）と比較しなさい。（解答は次頁）

表 6

ID番号	得点
1	86.4
2	86.4
3	81.4
4	81.4
5	87.6
6	72.7
7	83.9
8	81.4
9	45.9
10	86.4
11	72.9
12	86.4
13	85.1
14	86.4
15	87.6
16	70.9
17	90.7
18	72.1
19	82.6
20	73.9
平均	80.09

表 7　　表 6 の計算結果

ID番号	得点	平均からの差	…の自乗
1	86.4	6.3	39.3
2	86.4	6.3	39.3
3	81.4	1.3	1.7
4	81.4	1.3	1.7
5	87.6	7.5	56.4
6	72.7	−7.4	55.3
7	83.9	3.8	14.2
8	81.4	1.3	1.7
9	45.9	−34.2	1168.8
10	86.4	6.3	39.3
11	72.9	−7.2	51.7
12	86.4	6.3	39.3
13	85.1	5.0	25.1
14	86.4	6.3	39.3
15	87.6	7.5	56.4
16	70.9	−9.2	85.3
17	90.7	10.6	112.6
18	72.1	−8.0	63.8
19	82.6	2.5	6.3
20	73.9	−6.2	38.3
平均	80.09		
		変動	1936.2
		分散	96.8
		標準偏差	9.8

表 6 にもとづく変動、分散、標準偏差の計算結果は、表 7 のようになる。

変動、分散、標準偏差は、いずれも表 5 のデータよりも小さい。

データのばらつきを、グラフに表すと、表 5 のデータについては図 1 、表 6 のデータについては図 2 のようになる。

表 5 のデータのほうがばらつきが大きいことが視覚的にわかる。

以上の計算は、相関係数を学ぶ際の基礎となる。

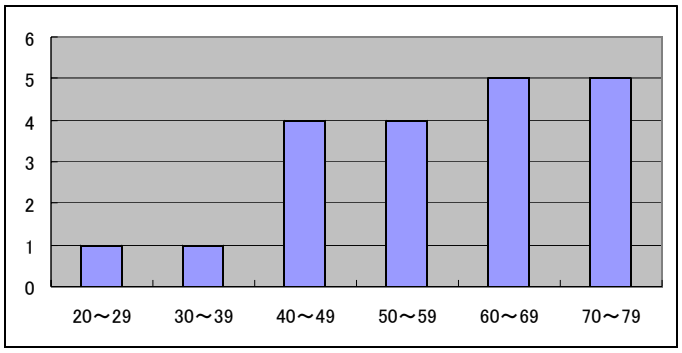

図 1　表 5 のデータの分布を示すヒストグラム

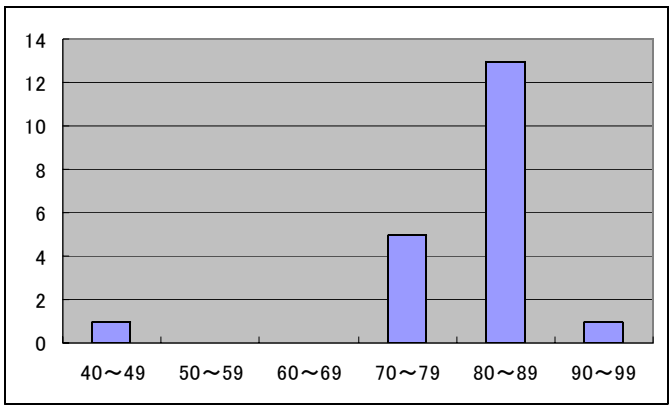

図 2　表 6 のデータの分布を示すヒストグラム

Σ（シグマ）の世界 2009　社会学データ実習　副読本
著者：松本康
発行：2009 年 4 月 20 日
非売品