統計　statistics

1　分散と標準偏差

集団を構成するすべての対象についてのデータを整理し、その中から有意な情報を抽出する統計処方は記述統計と呼ばれる。これに対して、集団中の限られた数の対象についてデータを分析し、それに基づいて全体の有意情報を推測する統計処方は推測統計と呼ばれる。推測統計においては、確率の概念が重要な役割を担う。データの集合があるとき、これらデータは一般にある分布状態を示す。このデータ構造を特徴づける各種のパラメータについて述べる。初めに、データ集合の代表的な値を表す尺度を考える。

算術平均

m 個のデータ $x_1, x_2, x_3, \cdots, x_m$ があるとき、これらデータの総和をデータの個数 m で割って得られる値
すなわち 算術平均 $\overline{x}$ は、データ集合における代表的な値を示すと考えられる。

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ \cdots \ + x_m}{m} = \frac{\sum_{i=1}^{m} x_i}{m} \qquad \left(\text{ただし } \sum_{i=1}^{m} x_i = x_1 + x_2 + x_3 + \ \cdots \ + x_m \right)$$

なお 平均には、算術平均の他に幾何平均や調和平均などがある。

$$\text{幾何平均 } \overline{x}_{\mathrm{G}} \ = \ (x_1 x_2 x_3 \ \cdots \ x_m)^{\frac{1}{m}}$$

$$\text{調和平均 } \overline{x}_{\mathrm{H}} \ = \ \left[\frac{1}{m} \left(\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ \cdots \ + \frac{1}{x_m} \right) \right]^{-1}$$

ただし 幾何平均や調和平均においては、各データ $x_1, x_2, x_3, \cdots, x_m$ がすべて正でなければならない。今後、単に平均と言うときは、算術平均 $\overline{x}$ を指す。

分散

データ集合の広がりの度合を示す尺度について考える。これは各データの散布状況すなわち「ばらつき」の度合を示す尺度を考えることである。
m 個のデータ $x_1, x_2, x_3, \cdots, x_m$ とし、その平均値を $\overline{x}$ とするとき、各データと平均値の差

$$x_1 - \overline{x}, \qquad x_2 - \overline{x}, \qquad x_3 - \overline{x}, \qquad \cdots, \qquad x_m - \overline{x}$$

のことを偏差と呼ぶ。
そこで これら偏差の二乗の総和

$$(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \cdots + (x_m - \overline{x})^2 = \sum_{i=1}^{m} (x_i - \overline{x})^2$$

を定義し、これを偏差平方和と呼ぶ。
偏差平方和は、データ集合の平均値 (中央値) からの「ばらつき」の度合を総体的に表していると考えられる。

［例題 1］

8 個のデータ x_i が、 $(2.0,\ 3.0,\ 4.0,\ 5.0,\ 6.0,\ 7.0,\ 8.0,\ 9.0)$ のとき、偏差の総和と偏差平方和を求めよ。

（解）

平均値 $\overline{x} = (2.0 + 3.0 + 4.0 + 5.0 + 6.0 + 7.0 + 8.0 + 9.0)/8 \ = \ 5.5$

データ x_i	偏差 $(x_i - \overline{x})$	偏差二乗 $(x_i - \overline{x})^2$
2.0	$2.0 - 5.5 = -3.5$	$(2.0 - 5.5)^2 = 12.25$
3.0	$3.0 - 5.5 = -2.5$	$(3.0 - 5.5)^2 = 6.25$
4.0	$4.0 - 5.5 = -1.5$	$(4.0 - 5.5)^2 = 2.25$
5.0	$5.0 - 5.5 = -0.5$	$(5.0 - 5.5)^2 = 0.25$
6.0	$6.0 - 5.5 = 0.5$	$(6.0 - 5.5)^2 = 0.25$
7.0	$7.0 - 5.5 = 1.5$	$(7.0 - 5.5)^2 = 2.25$
8.0	$8.0 - 5.5 = 2.5$	$(8.0 - 5.5)^2 = 6.25$
9.0	$9.0 - 5.5 = 3.5$	$(9.0 - 5.5)^2 = 12.25$
	偏差の総和 $\sum_{i=1}^{8}(x_i - \overline{x}) = 0$	偏差平方和 $\sum_{i=1}^{8}(x_i - \overline{x})^2 = 42.0$

［注］一般に偏差の単なる総和は

$$(x_1 - \overline{x}) + (x_2 - \overline{x}) + (x_3 - \overline{x}) + \ \cdots \ + (x_m - \overline{x}) \ = \ \sum_{i=1}^{m}(x_i - \overline{x}) = 0$$

となるので、データの「ばらつき」度合を表す指標としては定義できない。

それゆえ偏差平方和を定義する必要があったわけである。

一般に偏差平方和は、データの個数が多くなるほど大きな値になっていくので、データの散布状況すなわち「ばらつき」の客観的尺度としては、あまり適切なパラメータではない。そこで偏差平方和をデータの個数 m または $m-1$ で割って得られる値を「ばらつき」の客観的尺度として定義し、これを分散という。

すなわち分散とは

記述統計の立場では

$$v^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ \cdots \ + (x_m - \overline{x})^2}{m} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m}$$

推測統計の立場では

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ \cdots \ + (x_m - \overline{x})^2}{m-1} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m-1}$$

のように定義される。

今後は推測統計の立場で論述する。

［注］分散の定義は次式のように記すこともできる。

$$s^2 = \frac{(x_1^2 + x_2^2 + x_3^2 + \ \cdots \ + x_m^2) - m\overline{x}^2}{m-1}$$

標準偏差

前述した分散の単位は、データのもつ単位の二乗となる。そこでデータの単位と同じ乗数になるような「ばらつき」の尺度として、分散の平方根を定義し、これを標準偏差という。

すなわち標準偏差とは

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ \cdots \ + (x_m - \overline{x})^2}{m-1}} = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m-1}}$$

のように定義される。

［例題 2］

8 個のデータ x_i が (2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0) として与えられたとき、分散 s^2 と標準偏差 s を求めよ。

（解）

$$s^2 = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m-1} \quad \text{において}$$

データの個数 $m = 8$ かつ $\displaystyle\sum_{i=1}^{8}(x_i - \overline{x})^2 = 42.0$ より

$$s^2 = \frac{42.0}{8-1} = 6.0 \ : \ \text{分散}$$

$$s = \sqrt{\frac{42.0}{8-1}} = \sqrt{6.0} \ \approx \ 2.45 \ : \ \text{標準偏差}$$

標準偏差の性質

標準偏差の値は、同一種類の対象 (サンプル) についてのデータ集合に関してのみ大小の比較ができる。
(例えば同一の単位で表せる量についてのデータなど。)
対象の種類が異なっているデータ集合間において、データの「ばらつき」の度合を標準偏差の値だけから即座に判断することは妥当ではない。また データの個数が少ない場合、データ集合内に一つでも突出した値をもつデータが存在すると、その影響により標準偏差が大きくなる。

［例題 3］

大学生と一般住民の各集団内で 8 人の身長と体重についての各データ集合が次の表のように与えられているとき、
各データ集合の標準偏差を求めよ。

(1) 大学生の集団		
学生番号	身長 (m)	体重 (kg)
A1	1.658	55.8
A2	1.605	51.3
A3	1.693	68.6
A4	1.637	58.2
A5	1.614	50.4
A6	1.591	48.1
A7	1.626	55.9
A8	1.682	60.7

(2) 一般住民の集団		
住民番号	身長 (m)	体重 (kg)
B1	1.205	30.2
B2	1.408	38.6
B3	1.653	70.4
B4	1.634	58.1
B5	1.007	25.3
B6	1.592	51.5
B7	1.301	36.7
B8	1.702	65.8

（解）

(1) 大学生の集団について身長と体重の標準偏差を求める。

身長の平均値 $= 1.638$ m,　体重の平均値 $= 56.4$ kg

学生番号	身長の偏差二乗 (m^2)	体重の偏差二乗 (kg^2)
$A1$	$(1.658 - 1.638)^2$	$(55.8 - 56.4)^2$
$A2$	$(1.605 - 1.638)^2$	$(51.3 - 56.4)^2$
$A3$	$(1.693 - 1.638)^2$	$(68.6 - 56.4)^2$
$A4$	$(1.637 - 1.638)^2$	$(58.2 - 56.4)^2$
$A5$	$(1.614 - 1.638)^2$	$(50.4 - 56.4)^2$
$A6$	$(1.591 - 1.638)^2$	$(48.1 - 56.4)^2$
$A7$	$(1.626 - 1.638)^2$	$(55.9 - 56.4)^2$
$A8$	$(1.682 - 1.638)^2$	$(60.7 - 56.4)^2$
分散 $s^2 = \frac{\sum_{i=1}^{8}(x_i - \overline{x})^2}{8-1}$	$0.00134\ \mathrm{m}^2$	$43.1\ \mathrm{kg}^2$
標準偏差 $s = \sqrt{\frac{\sum_{i=1}^{8}(x_i - \overline{x})^2}{8-1}}$	0.0366 m	6.56 kg

(2) 一般住民の集団について身長と体重の標準偏差を求める。

身長の平均値 $=$ 1.438 m , 体重の平均値 $=$ 47.1 kg

住民番号	身長の偏差二乗 (m^2)	体重の偏差二乗 (kg^2)
$B1$	$(1.205-1.438)^2$	$(30.2-47.1)^2$
$B2$	$(1.408-1.438)^2$	$(38.6-47.1)^2$
$B3$	$(1.653-1.438)^2$	$(70.4-47.1)^2$
$B4$	$(1.634-1.438)^2$	$(58.1-47.1)^2$
$B5$	$(1.007-1.438)^2$	$(25.3-47.1)^2$
$B6$	$(1.592-1.438)^2$	$(51.5-47.1)^2$
$B7$	$(1.301-1.438)^2$	$(36.7-47.1)^2$
$B8$	$(1.702-1.438)^2$	$(65.8-47.1)^2$
分散 $s^2=\frac{\sum_{i=1}^{8}(x_i-\overline{x})^2}{8-1}$	$s^2=0.06254$ m^2	$s^2=282.0$ kg^2
標準偏差 $s=\sqrt{\frac{\sum_{i=1}^{8}(x_i-\overline{x})^2}{8-1}}$	$s=0.2501$ m	$s=16.79$ kg

上表の標準偏差の値から一般住民に関する身長データの「ばらつき」の度合は、大学生の身長データの「ばらつき」よりも大きいことが示されている。また体重についても同様の結果が現れている。

変動係数

前述したように標準偏差では、種類や単位などが異なるデータ集合間においてデータの「ばらつき」の度合を相互に比較することはできない。そこで種類が異なるデータ集合間においてデータの「ばらつき」の度合を比較するために、標準偏差 s を平均値 $\overline{x}$ で割ったパラメータを定義し、これを変動係数という。

すなわち変動係数とは

$$\text{変動係数}=\frac{s}{\overline{x}}$$

である。

［例題 4］

前述の例題 3 について変動係数を求めよ。

（解）

(1) 大学生の身長と体重データの「ばらつき」の度合を比較

$$\text{身長データの変動係数}=\frac{0.0366\ \mathrm{m}}{1.638\ \mathrm{m}}\approx 0.0223$$

$$\text{体重データの変動係数}=\frac{6.56\ \mathrm{kg}}{56.1\ \mathrm{kg}}\approx 0.117$$

この結果から体重データの「ばらつき」の度合は、身長データの「ばらつき」に比較して極めて大きいことが示されている。

(2) 一般住民の身長と体重データの「ばらつき」の度合を比較

$$\text{身長データの変動係数}=\frac{0.2501\ \mathrm{m}}{1.438\ \mathrm{m}}\approx 0.174$$

$$\text{体重データの変動係数}=\frac{16.79\ \mathrm{kg}}{47.1\ \mathrm{kg}}\approx 0.357$$

この結果から体重データの「ばらつき」の度合は、身長データの「ばらつき」に比較して若干大きめである。

2　相関

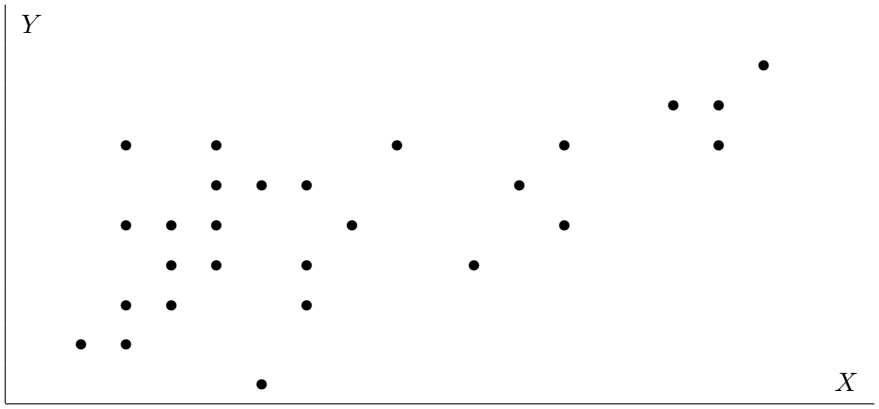
散布図

一般に統計データは次の表のように記せる。

データ数	データ X	データ Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
.	.	.
.	.	.
m	x_m	y_m

ここで X, Y はデータ集合 (サンプル) の種類を表す。二つの異なるデータ集合 X と Y について、それらデータ間の相対的なデータの分布傾向を調べるために、横軸に X，縦軸に Y をとって各データをプロットした図のことを散布図という。

［散布図の例］

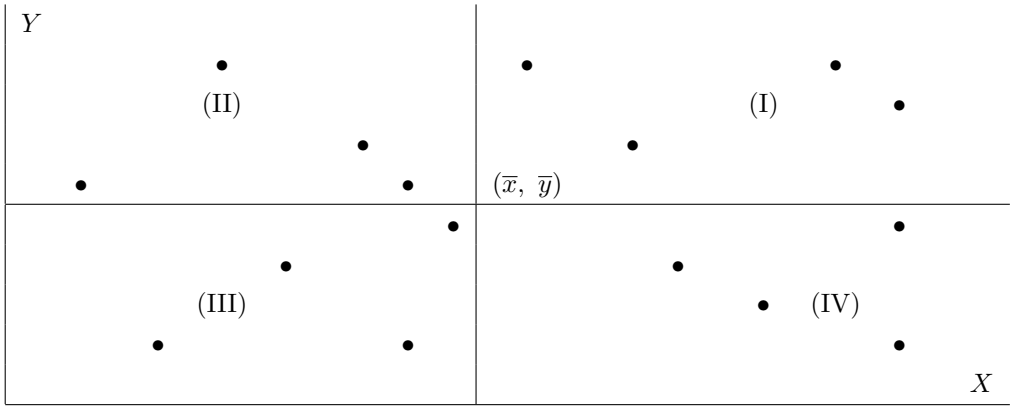


共分散

二つの異なるデータ集合 X, Y が与えられているとき、これらデータ間の相対的なデータ分布傾向を表すパラメータについて考える。

データ数	データ X	データ Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
.	.	.
.	.	.
m	x_m	y_m
	x_i の平均値 $\overline{x}$	y_i の平均値 $\overline{y}$

上表のデータ集合 X, Y を散布図に表し、さらに X, Y 各データの平均値 $(\overline{x}, \overline{y})$ を原点とする座標軸を記入した図を次に示す。



平均値を原点とする座標軸によって四分割された各領域 (I), (II), (III), (IV) の中の各データ (x_i, y_i) について偏差の積をとると、つぎの関係式を満たしている。

領域 (I) にあるとき　$(x_i - \overline{x})(y_i - \overline{y}) > 0$
領域 (II) にあるとき　$(x_i - \overline{x})(y_i - \overline{y}) < 0$
領域 (III) にあるとき　$(x_i - \overline{x})(y_i - \overline{y}) > 0$
領域 (IV) にあるとき　$(x_i - \overline{x})(y_i - \overline{y}) < 0$

したがって偏差の積の総和は、点が領域 (I) と (III) に多く集まるほど正の大きな値となり、領域 (II) と (IV) に集まると負の値になる。そこで二つのデータ集合間の相対的直線的分布状況を示す尺度として、各データの偏差 $(x_i - \overline{x})$ と $(y_i - \overline{y})$ の積の総和を定義し、これを偏差積和と呼ぶ。
すなわち 偏差積和とは

$$(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ \cdots \ + (x_m - \overline{x})(y_m - \overline{y}) = \sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})$$

のように記される。偏差積和が正または負の大きな値をもつほど、散布図上では各データの点が直線状の分布に近づく。
前述の偏差積和はデータ点の個数 m が多くなると、その値 (絶対値) が増加する傾向があるので、点の散布状況の客観的尺度としては不適当である。そこで偏差積和を $m-1$ で割って共分散 s_{sy} を定義する。
すなわち 共分散とは

$$s_{xy} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{m-1}$$

のように記せる。

相関係数

前述の共分散は、データ集合の種類や単位に依存する。そこでデータ集合の種類や単位に依らない最も普遍的なデータ散布状況の尺度として、共分散を各標準偏差 s_x, s_y の積で割って相関係数 r を定義する。
すなわち 相関係数とは

$$r = \frac{s_{xy}}{s_x\ s_y}$$

である。ただし

$$s_{xy} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{m-1}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m-1}}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \overline{y})^2}{m-1}}$$

また 相関係数は次式のようにも記せる。

$$r = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2}\ \sqrt{\sum_{i=1}^{m}(y_i - \overline{y})^2}}$$

相関係数 r は、散布図上においてデータ集合の点の分布状況が、直線的に分布している度合を表す指標となる。

相関係数の性質

相関係数 r は、1 と -1 の間の値をとり、1 または -1 に近づくほど散布図上のデータ点は直線状に分布する。相関係数が正のとき、散布図上の点の集合は右上がりに分布し、このとき二つのデータ集合間には正の相関があるといわれる。相関係数が負のとき、散布図上の点の集合は右下がりに分布し、このとき二つのデータ集合間には負の相関があるといわれる。二つのデータ集合間に相関が認められても、これは二つのデータ集合間に必ずしも論理的な因果関係が存在することを意味するものではない。すなわち着目している事象の間に相関があることと因果関係が存在することとは全く別の問題である。また二つのデータ集合間の相関係数が零であっても、データ集合間に一定の関係が存在する場合もあり得る。例えば 散布図上において、データ点が円周付近に沿って分布しているような場合、二つのデータ集合間には一定の関係が認められるが、相関係数はほぼ零となる。なお 散布図上において、データ点の分布が円内にほぼ一様に分布している場合には、当然ながら二つのデータ集合間には一定の関係は認められず、かつ相関係数もほぼ零である。

［例題 5］

身長と体重についての各データ集合が、次の表のように与えられているとき、身長データと体重データの間の相関係数を求めよ。

データ数	身長 (m)	体重 (kg)
1	1.658	55.8
2	1.605	51.3
3	1.693	68.6
4	1.637	58.2
5	1.614	50.4
6	1.591	48.1
7	1.626	55.9
8	1.682	60.7

（解）

はじめに s_x, s_y, s_{xy} をそれぞれ求める。

身長の平均値 $= 1.638$ m,　体重の平均値 $= 56.4$ kg

数	身長の偏差二乗	体重の偏差二乗	偏差積
1	$(1.658-1.638)^2$	$(55.8-56.4)^2$	$(1.658-1.638)\times(55.8-56.4)$
2	$(1.605-1.638)^2$	$(51.3-56.4)^2$	$(1.605-1.638)\times(51.3-56.4)$
3	$(1.693-1.638)^2$	$(68.6-56.4)^2$	$(1.693-1.638)\times(68.6-56.4)$
4	$(1.637-1.638)^2$	$(58.2-56.4)^2$	$(1.637-1.638)\times(58.2-56.4)$
5	$(1.614-1.638)^2$	$(50.4-56.4)^2$	$(1.614-1.638)\times(50.4-56.4)$
6	$(1.591-1.638)^2$	$(48.1-56.4)^2$	$(1.591-1.638)\times(48.1-56.4)$
7	$(1.626-1.638)^2$	$(55.9-56.4)^2$	$(1.626-1.638)\times(55.9-56.4)$
8	$(1.682-1.638)^2$	$(60.7-56.4)^2$	$(1.682-1.638)\times(60.7-56.4)$
	$s_x^2 = 0.00134$ m^2	$s_y^2 = 43.1$ kg^2	$\sum_{i=1}^{8}(x_i-\overline{x})(y_i-\overline{y}) = 1.555$ kg m
	$s_x = 0.0366$ m	$s_y = 6.56$ kg	$s_{xy} = \frac{\sum_{i=1}^{8}(x_i-\overline{x})(y_i-\overline{y})}{8-1} = 0.2222$ kg m

よって　相関係数 $r = \dfrac{s_{xy}}{s_x\, s_y} = \dfrac{0.2222}{0.0366\times 6.56} = 0.925$

すなわち相関係数の値 0.925 より、身長データと体重データの間には正の相関が認められ、
散布図上では右上がりの直線的なデータ点の分布状況となる。