データの分析

- ［数Ⅰ］基本用語
- ［数Ⅰ］データの整理 … 階級編
- ［数Ⅰ］データの整理 … 四分位編
- ［数Ⅰ］四分位数の求め方
- ［数Ⅰ］データの分析
- ［数Ⅰ］多次元データの分析
- ［数Ⅰ数Ⅱ補足］重要公式のまとめ
- ［数Ⅰ］度数分布表と統計値
- ［数Ⅰ補足］平均値と中央値
- ［補足］補足事項

Keywords

データ　変量　代表値
平均値 mean $\overline{x}$　中央値 median
最頻値 mode　範囲 range　階級
階級値　階級の幅　度数　相対度数
度数分布の表現　度数分布表　ヒストグラム
四分位数　四分位範囲　四分位偏差
箱ひげ図　偏差　平均偏差　分散 s^2
標準偏差 s　多次元データ　散布図
相関　共分散 s_{xy}　相関係数 r

http://www.zacky.ninja-x.jp
printed：2014.11.04(Tue) 18:13

基本用語

I A II B III C 補

データ

複数の事象・数値の集まり。

データ

x_1	x_2	x_3	$\cdots$
$\cdots$	x_i	$\cdots$	x_n

変量

ある特性を表す値。

添え字(suffix)を用いて表すことも多い。

「データ $\longrightarrow$ 集合，変量 $\longrightarrow$ 要素」と読み替えても良い。

代表値

データの特性を表す値。

平均値 mean $\overline{x}$

データの値の総和を、データの値の個数で割った値。

中央値 median

データを小さい順に並べたとき、中央にある値。

※四分位数の求め方とセットで覚えること。

i. データの値が奇数 $(2n+1)$ 個のとき、
中央値 = $(n+1)$ 個目の値

ii. データの値が偶数 $2n$ 個のとき、
中央値 = n個目と$(n+1)$ 個目の値の平均

最頻値 mode

最も頻度が高く登場した値。

度数分布表で、度数が最も大きい階級値。

範囲 range

データの最大値と最小値の差。

データの整理　…　階級編

I A II B III C 補

階級

データを区切った区間。

階級値

階級の上端と下端の平均値。

階級の幅

階級の上端と下端の差。

度数

各階級に属する変量の個数。

相対度数

全体に対し、ある階級の変量の個数の割合。

度数分布の表現

度数分布表

階級ごとに対応する度数をまとめた表。

ヒストグラム

度数分布表を表した棒グラフ。

度数分布表

階級	階級値	度数	相対度数	階級値×相対度
$0 \leqq x < 10$	5	0	0.0	0.0
$10 \leqq x < 20$	15	2	0.2	3.0
$20 \leqq x < 30$	25	2	0.2	5.0
$30 \leqq x < 40$	35	5	0.5	17.5
$40 \leqq x \leqq 50$	45	1	0.1	4.5
計	×	10	1.0	30.0

平均値:30.0 ， 中央値:35 ， 最頻値:35

問題への取り組み方

与えられたデータの形式により、解答の方針が変わる。

データの形式	代表値の解答
すべてのデータを具体的に書き示す	個別の変量の値
度数分布表で示す	階級値(両端の平均値)

データの整理　…　四分位編

I A II B III C 補

四分位数

データの値を小さい順に並べ、個数ごとに4つに区切る。その境となる値を、小さい方から順番に"第1四分位数"、"第2四分位数"(＝中央値)、"第3四分位数"という。

分位数quantileの頭文字から、Q_n と表記することもある。

四分位範囲

第1四分位数と第3四分位数の差。

四分位偏差

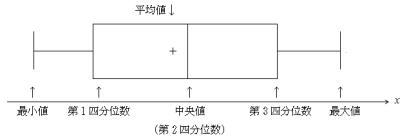
第1四分位数と第2四分位数、第2四分位数と第3四分位数の差の平均。

$$\text{四分位偏差} = \frac{(Q_2 - Q_1) + (Q_3 - Q_2)}{2} = \frac{Q_3 - Q_1}{2} = \frac{\text{四分位範囲}}{2}$$

箱ひげ図

データの四分位数を四角形(箱)、データの最大値・最小値を線(ひげ)で表した図。

平均値を"＋"で示すこともある。



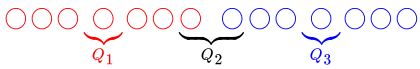
四分位数の求め方

1. データを小さい順に並べる
2. 中央値(第2四分位数)を求める
3. 中央値で区切られた両側において、それぞれの中央値(第1・3四分位数)を求める
   ※第2四分位数の導出に用いた値を、第1・3四分位数の導出で用いるか否かに注意

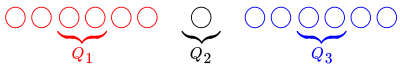
図解編

i. データの要素が $4n+3$ 個のとき、

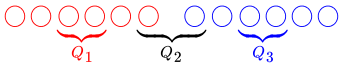
ii. データの要素が $4n+2$ 個のとき、



iii. データの要素が $4n+1$ 個のとき、



iv. データの要素が $4n$ 個のとき、



偏差

変量と平均値の差。

$$\text{偏差} = x_i - \overline{x}$$

平均偏差

偏差の平均

$$\begin{aligned}\text{平均偏差} &= \frac{1}{n}\left\{\left(x_1-\overline{x}\right)+\left(x_2-\overline{x}\right)+\left(x_3-\overline{x}\right)+\cdots+\left(x_n-\overline{x}\right)\right\}\\ &= \frac{\text{データの合計}}{n}-\frac{n\cdot\overline{x}}{n} \quad =0\end{aligned}$$

分散 s^2

偏差の2乗平均。

$$s^2 = \frac{1}{n}\left\{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\left(x_3-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2\right\}$$

標準偏差 s

分散の正の平方根。

$$s = \sqrt{s^2}$$

平均偏差と分散

偏差（平均値からの離れ具合）を見ることで、データ全体の散らばり具合を見ることができる。しかし、その散らばり方が正負に対し均一になると、その離れ具合を互いに打ち消し合い、平均偏差が0に近づき、散らばり具合を正しく見ることができなくなる。
したがって、その変量の散らばり方が互いに打ち消し合わないように偏差を2乗して見たものが分散である。表2において、平均偏差ではバラツキを知ることはできないが、分散ではバラツキを知ることができる。

（例） 得点の散らばり

表1 均一な得点

生徒	得点［点］	偏差	偏差 ^{2}
A	10	0	0
B	10	0	0
C	10	0	0
D	10	0	0
平均	10	0	0

表2 均一でない得点

生徒	得点［点］	偏差	偏差 ^{2}
A	14	4	16
B	9	−1	1
C	7	−3	9
D	10	0	0
平均	10	0	6.5

多次元データの分析

I A II B III C 補

多次元データ

複数の値の組を変量とするデータ。

n 個の値の組からなる変量を持つデータを、“ n 次元データ” という。

散布図

2次元データの変量 (x, y) を座標とした点を描いた図。

(座標から点を打つことを、プロットする という)

相関

散布図の点の集まり方による、そのデータの関係性のこと。

共分散 s_{xy}

多次元データにおける、変量の偏差の積の平均。

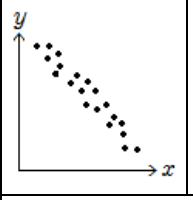
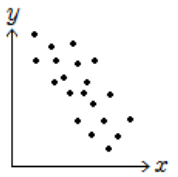
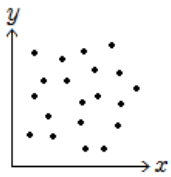
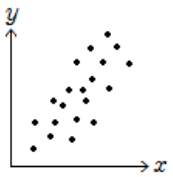
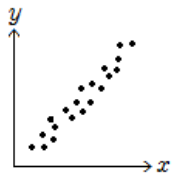
$$s_{xy} = \frac{1}{n}\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + (x_3 - \overline{x})(y_3 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\}$$

相関係数 r

2次元データ (x, y) の各変量について、それぞれの標準偏差 s_x, s_y 、共分散 s_{xy} とするとき、次の値を"相関係数"という。

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

2次元データの相関

正負	負の相関がある		相関がない	正の相関がある	
強さ	相関が強い	相関が弱い		相関が弱い	相関が強い
表現	負の強い相関がある	負の弱い相関がある		正の弱い相関がある	正の強い相関がある
散布図					
相関係数	$r = -1$	⟵	$r = 0$	⟶	$r = 1$

［参考］相関の具体的な表現

相関係数 $\lvert r \rvert$	0.0～0.2	0.2～0.4	0.4～0.7	0.7～1.0
相関の表現	ほとんど相関なし	弱い相関あり	やや相関あり	強い相関あり

重要公式のまとめ

x_i, y_i :変量の値、n :変量の個数、f :度数(頻度を表すfreaquencyの頭文字)、c :階級の数 とおく。

平均値	$\overline{x}$	$=$	$\dfrac{1}{n}\sum_{i=1}^{n} x_i$	
		$=$	$\dfrac{1}{n}\sum_{i=1}^{c} x_i \cdot f_i$	(階級値と度数)
		$=$	$\sum_{i=1}^{c} x_i \cdot \dfrac{f_i}{n}$	(階級値と相対度数)
度数の和	n	$=$	$\sum_{i=1}^{c} f_i$	
相対度数の和	1	$=$	$\sum_{i=1}^{c} \dfrac{f_i}{n}$	
平均偏差			$\dfrac{1}{n}\sum_{i=1}^{n} \lvert x_i - \overline{x} \rvert$	
分散	s^2	$=$	$\dfrac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2$	
		$=$	$\dfrac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \overline{x}^2$	(前式より導出)
標準偏差	s	$=$	$\sqrt{s^2}$	
共分散	s_{xy}	$=$	$\dfrac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$	
		$=$	$\dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$	(前式より導出)
相関係数	r	$=$	$\dfrac{s_{xy}}{s_x s_y}$	

参考:数列の和 … 数Ⅱ「数列」

度数分布表と統計値

度数分布表

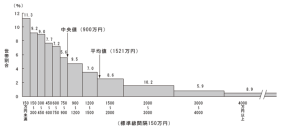
階級	階級値	度数	相対度数	階級値×度数	階級値×相対度数	偏差	偏差^2
○$\leqq x <$○	○$_1$	f_1	$\dfrac{f_1}{\text{度数の和}}$	○$_1 \times f_1$	○$_1 \times \dfrac{f_1}{\text{度数の和}}$	$x_1 - \overline{x}$	$(x_1 - \overline{x})^2$
○$\leqq x <$○	○$_2$	f_2	$\dfrac{f_2}{\text{度数の和}}$	○$_2 \times f_2$	○$_2 \times \dfrac{f_2}{\text{度数の和}}$	$x_2 - \overline{x}$	$(x_2 - \overline{x})^2$
⋮							
○$\leqq x <$○	○$_i$	f_i	$\dfrac{f_i}{\text{度数の和}}$	○$_i \times f_i$	○$_i \times \dfrac{f_i}{\text{度数の和}}$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
⋮							
○$\leqq x \leqq$○	○$_n$	f_n	$\dfrac{f_n}{\text{度数の和}}$	○$_n \times f_n$	○$_n \times \dfrac{f_n}{\text{度数の和}}$	$x_n - \overline{x}$	$(x_n - \overline{x})^2$
計(和)	×	変量の数	1.0	変量の合計	平均値	偏差の合計 =0	偏差^{2}の合計
平均$= \dfrac{\text{計(和)}}{\text{変量の数}}$	×	×	×	平均値	×	0	分散

平均値と中央値

平均値も中央値も「全体の値のど真ん中」を求めているように見えます。しかし、平均値は、全体としては少数でも飛び抜けて大きい・小さい値が少しでもあると、そちらに影響を受けた値が出てしまいます。中央値であれば個数に着目するため、少数のはずれ値は無視することができます。

一部の変量により平均値が引き上げられる例

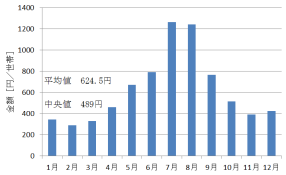
平均金額の貯蓄がある世帯は、全体の3割程度しかいません。
少数の富裕層が平均を引き上げているためです。
全体の平均値だけではなく、世帯数(データの個数)に注目した"中央値"を見る必要があります。



2人以上の世帯の貯蓄額
（総務省統計局 平成21年全国消費実態調査）

変量の値が常に変化する例

季節により、値が大きく変動しています。
平均値・中央値共に、夏のアイスクリームが売れるシーズンに引き上げられています。
状況により常に均等なデータが得られる訳ではないので、データ全体の値ではなく月ごとに範囲(最大値・最小値)・四分位範囲といった幅からデータを見る必要があります。



1世帯当たりアイスクリーム支出金額
（日本アイスクリーム協会 2011年調査）

補足事項

仮平均

計算する値が大きくなってしまうときに、ある基準(=“仮平均”)からの偏差を用いることがあります。

はずれ値

現実の統計では、「テストでみんな高得点なのに、欠席をした1人だけが0点」・「実験で測定ミスがあって変な値が出た」といったような イレギュラー な値が含まれることがあります。

このような値を“はずれ値”といい、データの分析をする際には 無視 をします。

線形性(線形変換)

2つのデータ X と $aX+b$ について、次の関係が成り立ちます。

$$\begin{aligned} E(aX+b) &= aE(X)+b \\ V(aX+b) &= aV(X)+b \end{aligned}$$

※仮平均も線形変換の一種($a=1, b=-$仮平均)といえます。

正規化・標準化

平均を0、分散が1となるようにデータを線形変換すること。

$$\text{標準化得点} = \frac{x_i - \overline{x}}{s}$$

偏差値

平均を50、分散が10となるように標準化した得点のこと。

$$\text{偏差値} = \frac{x_i - \overline{x}}{s} \times 10 + 50$$