役立つ統計学

「統計学」は「統計解析」・「統計分析」
と呼ばれることもあります。同じです。

講座の進め方

2022年秋頃、コロナ禍が収まればリアルで講座を開講します。それまではオンラインでPPの講座となります。

①ウェブ会議ツール{ZOOM]を利用した講座
②定期的な講座ではありません
③教材は各自で印刷をしてください

統計学を学ぶ理由

■「実学」とは・・・「転用可能なスキル」、つまり、どの分野にでも応用できる知識・技術のことです。 例えば「英語」「プログラミング」などです。生涯にわたり21世紀のスキルとして、グローバル世界に役立つのが「統計学」です。

■大学でも「データ・サイエンス」の習得を全員に推奨しています。

2023年からは高等学校から導入されます

学習内容

Part1

1. 度数分布表
2. ヒストグラム
3. 相関と散布図
4. 平均と標準偏差
5. 分散
6. 標準化
7. 共分散

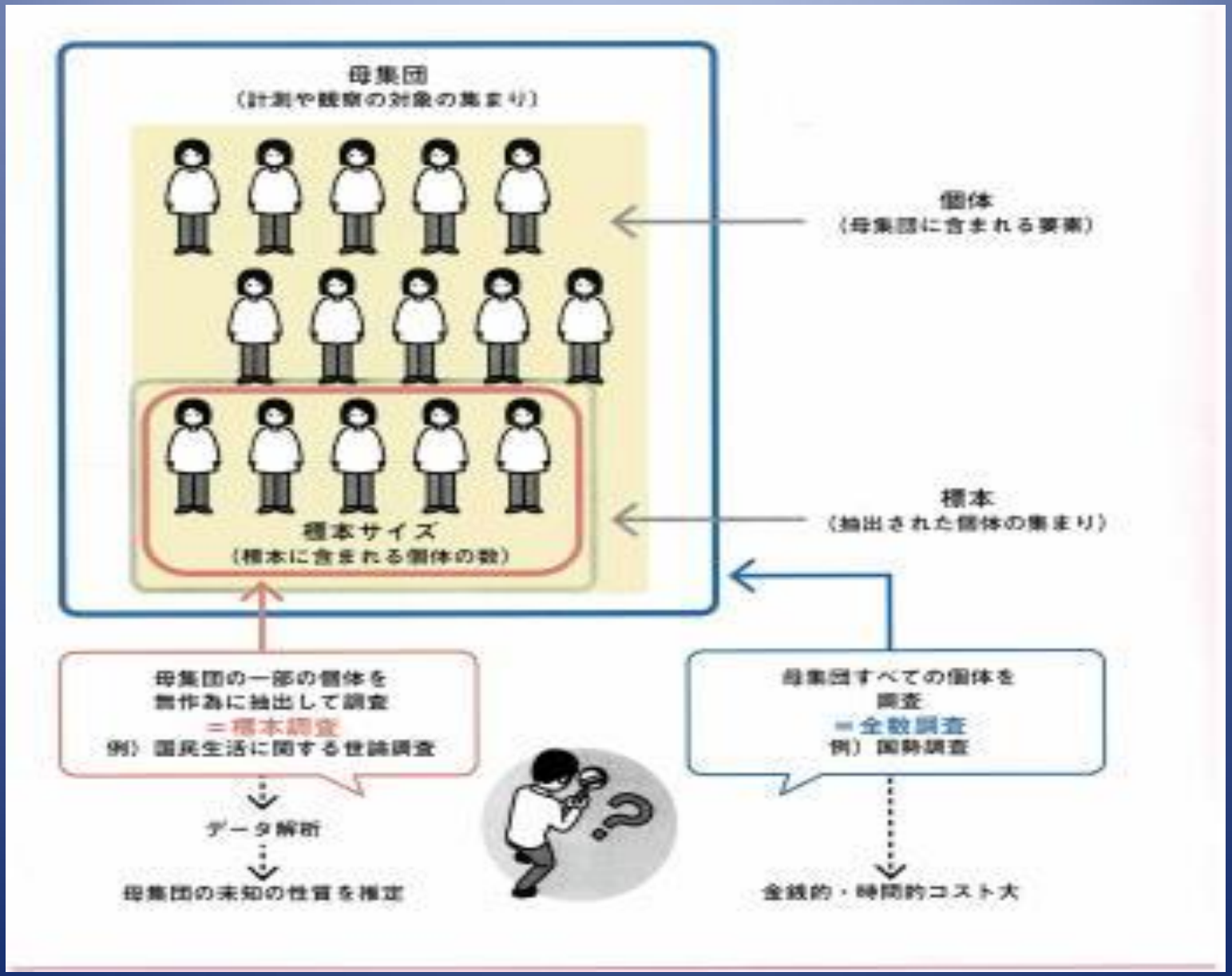
Part2

1. 母集団と標本
2. 確率分布と確率変数
3. 正規分布

Part3

1. 標本平均と標本分散
2. 標本平均の分布
3. 点推定と区間推定
4. 母平均の区間推定
5. 母比率の推定
6. 統計的仮説検定

データ解析の目的


母集団
（計測や観察の対象の集まり）
個体
（母集団に含まれる要素）
標本
（抽出された個体の集まり）
標本サイズ
（標本に含まれる個体の数）
母集団の一部の個体を
無作為に抽出して調査
＝標本調査
例）国民生活に関する世論調査
データ解析
母集団の未知の性質を推定
母集団すべての個体を
調査
＝全数調査
例）国勢調査
金銭的・時間的コスト大

度数分布表

表１　各都道府県の家賃データ

都道府県	公営賃貸住宅家賃	都道府県	公営賃貸住宅家賃	都道府県	公営賃貸住宅家賃	都道府県	公営賃貸住宅家賃	都道府県	公営賃貸住宅家賃	都道府県	公営賃貸住宅家賃
北海道	1,333	栃木県	1,296	石川県	1,176	滋賀県	1,725	岡山県	859	佐賀県	1,125
青森県	995	群馬県	1,231	福井県	1,123	京都府	2,132	広島県	1,220	長崎県	1,276
岩手県	1,062	埼玉県	2,651	山梨県	1,269	大阪府	1,954	山口県	913	熊本県	1,272
宮城県	1,487	千葉県	2,854	長野県	1,273	兵庫県	2,255	徳島県	946	大分県	1,203
秋田県	1,120	東京都	3,623	岐阜県	930	奈良県	2,491	香川県	1,107	宮崎県	1,012
山形県	1,254	神奈川県	3,346	静岡県	1,506	和歌山県	1,475	愛媛県	907	鹿児島県	1,278
福島県	1,067	新潟県	1,274	愛知県	2,044	鳥取県	893	高知県	1,015	沖縄県	1,393
茨城県	1,136	富山県	1,082	三重県	964	島根県	1,014	福岡県	1,883	円（1ヵ月 3.3m² 当たり）	

データを値の大きさで分類したもの

その階級までの度数の累計値

階級に属するデータの数

表２　度数分布表

階級 以上	階級 未満	階級値	度数	相対度数	累積度数
800	1000	900	8	0.170	8
1000	1200	1100	12	0.255	20
1200	1400	1300	13	0.277	33
1400	1600	1500	3	0.064	36
1600	1800	1700	1	0.021	37
1800	2000	1900	2	0.043	39
2000	2200	2100	2	0.043	41
2200	2400	2300	1	0.021	42
2400	2600	2500	1	0.021	43
2600	2800	2700	1	0.021	44
2800	3000	2900	1	0.021	45
3000	3200	3100	0	0.000	45
3200	3400	3300	1	0.021	46
3400	3600	3500	0	0.000	46
3600	3800	3700	1	0.021	47
計			47	1.000	

度数分布表にすると、さまざまな情報を読み取れるようになる！

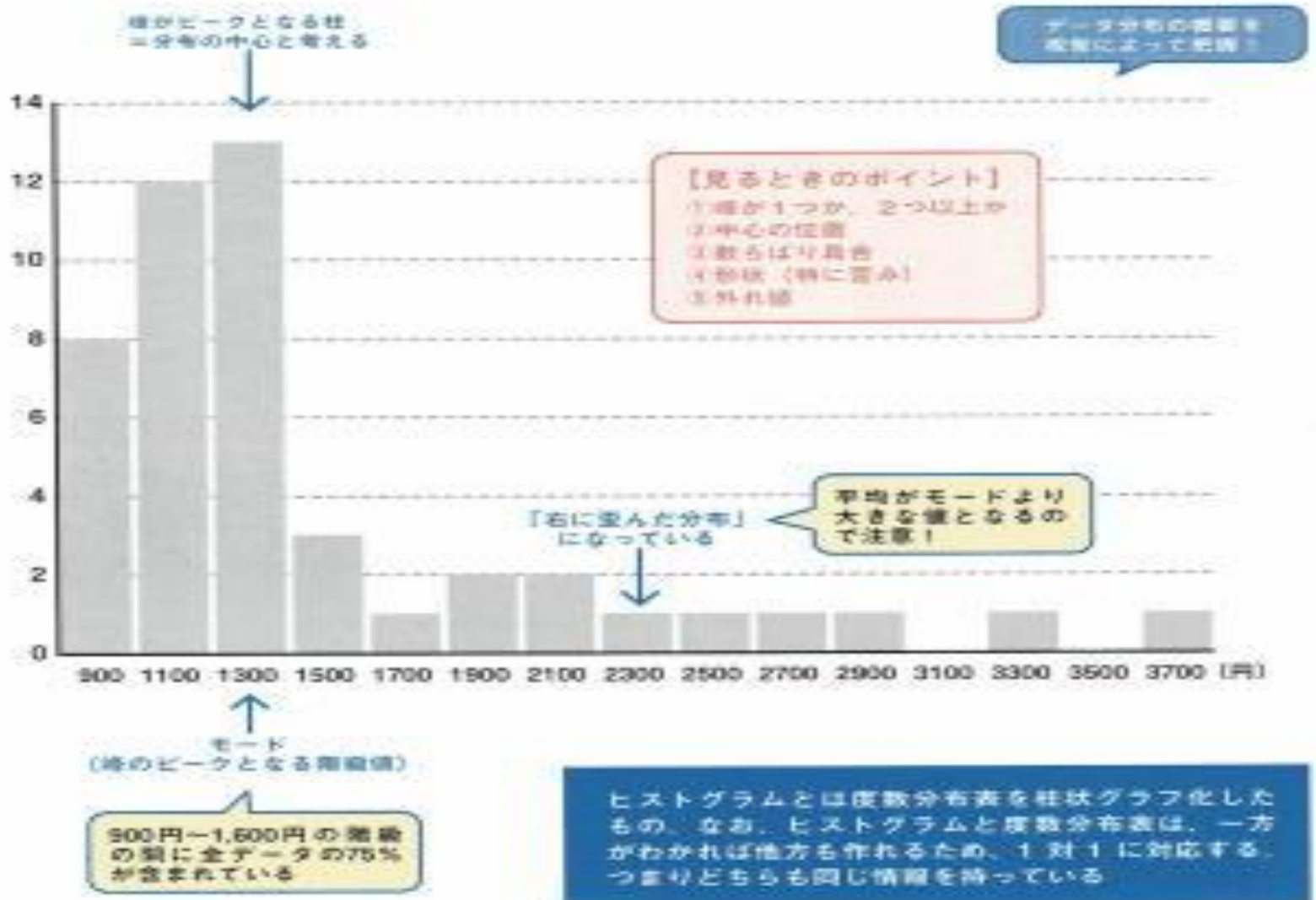
各階級を代表する値で階級の中間値

各階級の度数がデータ数に占める割合

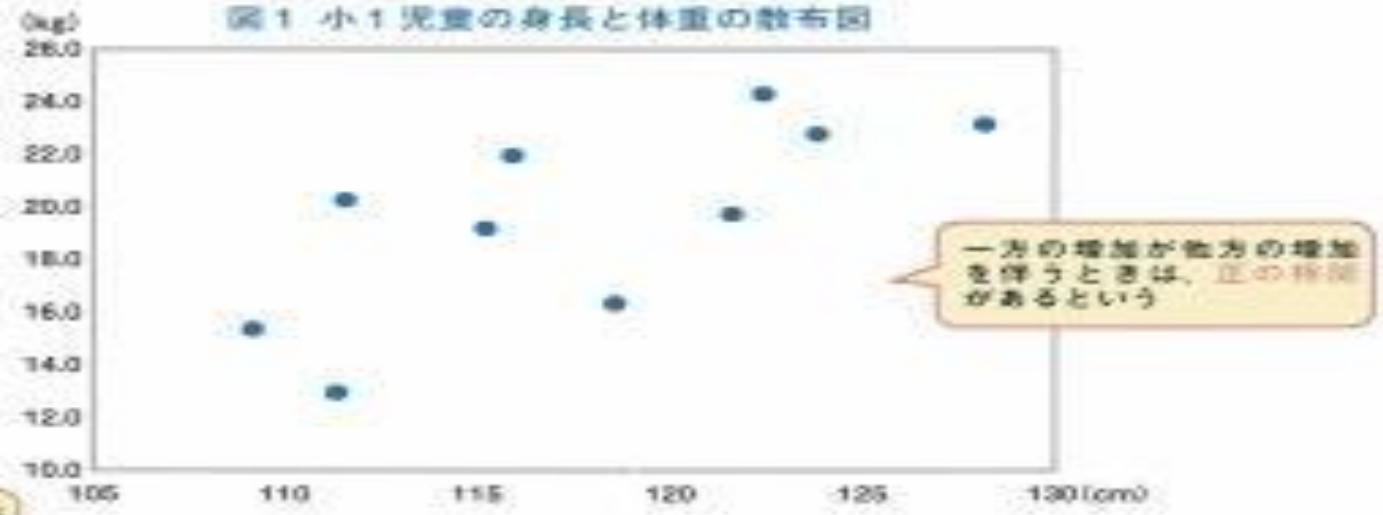
モード（最頻値）
度数が最大となる階級の階級値

「社会生活統計指標－都道府県の指標－2015」（総務省統計局）より
＊データは2013年のもの

ヒストグラム


峰がピークとなる柱
＝分布の中心と考える
データ分布の傾向を
視覚によって把握！
14
12
10
8
6
4
2
0
【見るときのポイント】
①峰が1つか、2つ以上か
②中心の位置
③散らばり具合
④形状（特に歪み）
⑤外れ値
「右に歪んだ分布」
になっている
平均がモードより
大きな値となるの
で注意！
900 1100 1300 1500 1700 1900 2100 2300 2500 2700 2900 3100 3300 3500 3700 (円)
モード
（峰のピークとなる階級値）
900円～1,600円の階級
の間に全データの75%
が含まれている
ヒストグラムとは度数分布表を柱状グラフ化したもの。なお、ヒストグラムと度数分布表は、一方がわかれば他方も作れるため、1対1に対応する。つまりどちらも同じ情報を持っている

相関と相関図


(kg)
26.0
24.0
22.0
20.0
18.0
16.0
14.0
12.0
10.0
105
110
115
120
125
130(cm)
一方の増加が他方の増加を伴うときは、正の相関があるという


一方の増加が他方の減少を伴うとき


図２ 負の相関


図３ 無相関

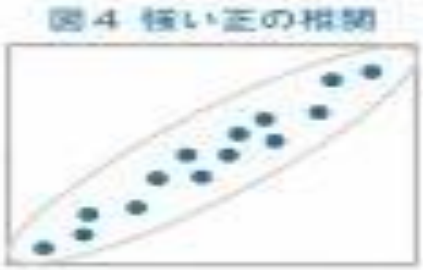


図４ 強い正の相関


直線に近いとき

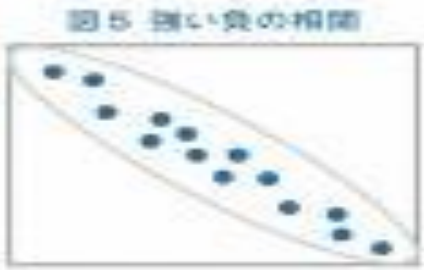


図５ 強い負の相関

平　均

表1　５人の生徒が100点満点の試験を行ったところ成績は以下であった

番号	得点（点）
1	90
2	80
3	70
4	60
5	50

表2　５人の小１児童の身長を測定したところ結果は以下であった

番号	身長（cm）
1	111.6
2	122.5
3	123.9
4	109.2
5	115.9

「平均」は，数値による要約・整理の方法の１つ

$$\text{平均} = \frac{\text{データの総和}}{\text{データの数}}$$

元のデータと同じ単位

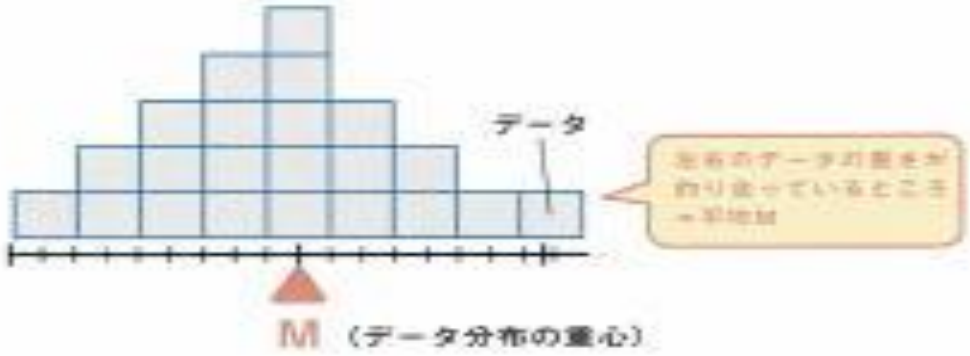
$$= \frac{(90+80+70+60+50)}{5} = \frac{350}{5} = 70\text{点}$$

定義

n 個の測定値からなるデータを $x_1, x_2, \cdots, x_n$ とあらわすと平均（M）は？

$$M = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

各データを同じ重さをもつブロックと見ると？



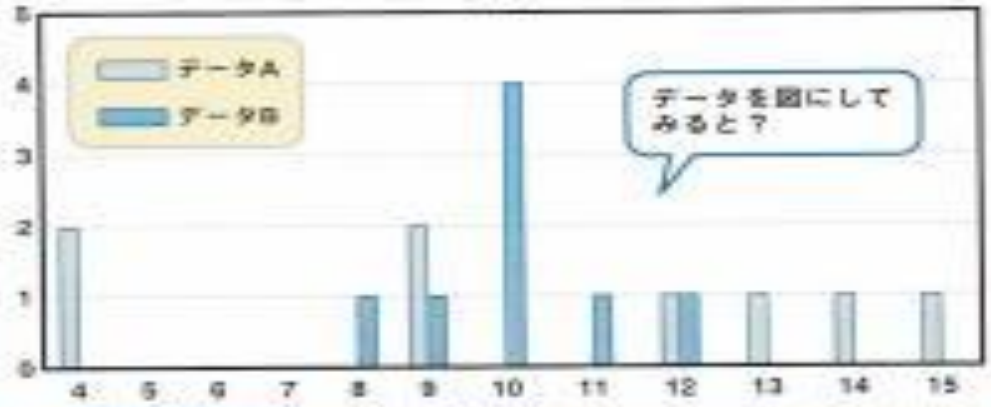
標準偏差

表1 データ

データA	4	4	9	9	12	13	14	15	平均＝10
データB	8	9	10	10	10	10	11	12	平均＝10

← A も B も平均は 10 で分布の中心の位置は同じ

図



平均からのデータの散らばり具合をより明らかにするには？

?

表2 偏差

データA	−6	−6	−1	−1	2	3	4	5
データB	−2	−1	0	0	0	0	1	2

各測定値と平均との差（偏差）を調べる

絶対偏差の平均値のこと

表3 絶対偏差

									平均偏差
データA	6	6	1	1	2	3	4	5	3.5
データB	2	1	0	0	0	0	1	2	0.75

「＋」「−」をとって絶対値にする

平均偏差を見るとデータAのほうが散らばりが大きいことがわかる

絶対偏差が小さいほどデータは平均の周りに集まっている

分　散

前回表3　絶対偏差

データA	6	6	1	1	2	3	4	5
データB	2	1	0	0	0	0	1	2

2乗を取った数字（偏差2乗）

36	36	1	1	4	9	16	25
4	1	0	0	0	0	1	4

データAの

$$\text{分散} = \frac{36+36+1+1+4+9+16+25}{8} = \frac{128}{8} = 16$$

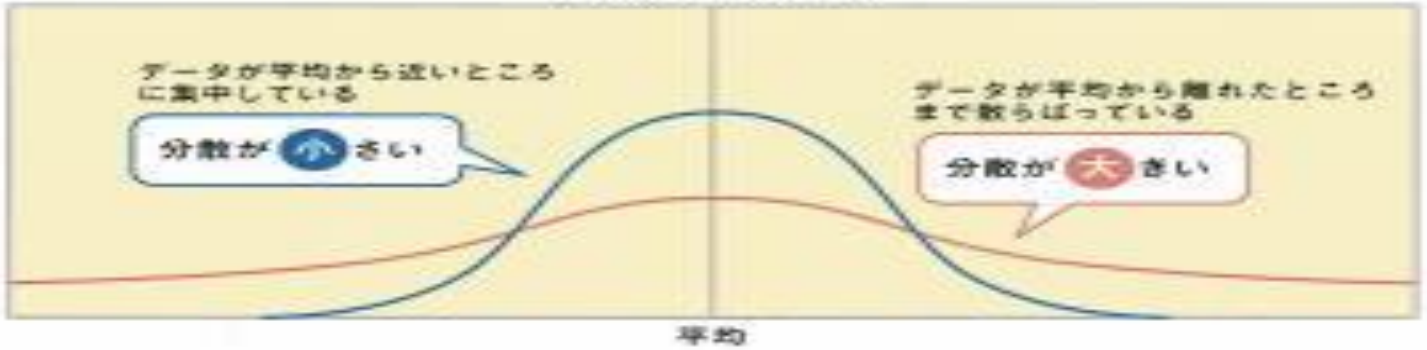
大

データBの

$$\text{分散} = \frac{4+1+0+0+0+0+1+4}{8} = \frac{10}{8} = 1.25$$

小

分散の視覚イメージ



データを $x_1, x_2, \cdots, x_n$ で表し（nはデータ数）その平均をMとしたとき

$$\text{分散} = \frac{1}{n}\left\{(x_1-M)^2+(x_2-M)^2+\cdots+(x_n-M)^2\right\}$$

↑各測定値の偏差の2乗

標準偏差

標準偏差 分散の正の平方根のこと

$$S=\sqrt{\frac{1}{n}\left\{(x_1-M)^2+(x_2-M)^2+\cdots+(x_n-M)^2\right\}}$$

式を使って一般的に書くと？

ポイント 元のデータと同じ単位にするため分散の平方根を取る

表1 2016年3月の最高気温(℃)データ

1日	10.3	9日	14.9	17日	20.4	25日	14.4
2日	12.4	10日	8.5	18日	21.4	26日	12.5
3日	16.4	11日	6.1	19日	16.7	27日	16.1
4日	15.9	12日	8.4	20日	19.0	28日	16.8
5日	16.3	13日	9.2	21日	13.8	29日	18.6
6日	17.3	14日	6.9	22日	17.5	30日	20.0
7日	15.5	15日	14.1	23日	16.0	31日	20.2
8日	20.8	16日	13.9	24日	9.5		

気象庁より

正規分布とシグマ範囲の図

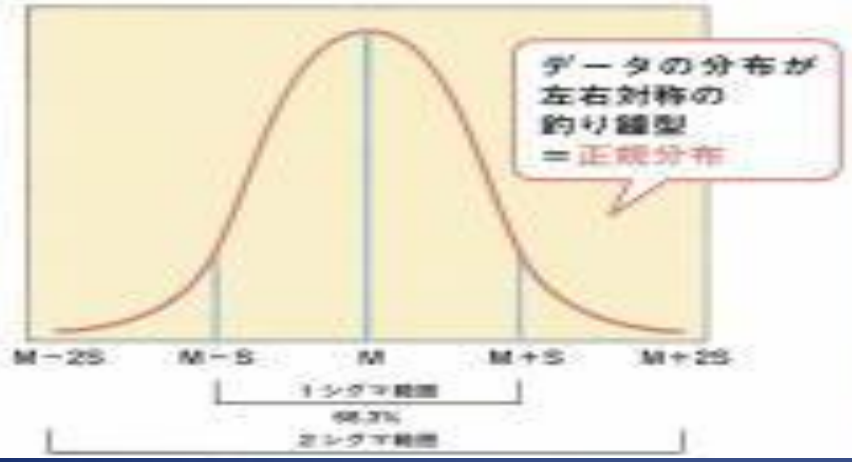


表2

平均	14.9
標準偏差	4.2
1シグマ範囲	10.7以上
	19.1以下
2シグマ範囲	6.5以上
	23.3以下
3シグマ範囲	2.3以上
	27.5以下
最小値	6.1
最大値	21.4

標準化

表1 高校生の体重

番号	体重(kg)	標準化
1	58.5	−0.15
2	61.5	0.15
3	64.0	0.40
4	46.0	−1.40
5	56.8	−0.32
6	55.4	−0.46
7	84.5	2.45
8	48.0	−1.20
9	64.1	0.41
10	61.0	0.10

平均M＝60.0kg
標準偏差S＝10.0kg

$\frac{\text{測定値} - \text{M}}{\text{S}}$ で計算した数値（次項で説明します）

A君の体重

平均と1kgの違いだが、
平均と大差はない

61.0＝60.0＋1.0(kg)

表2 新生児の体重

番号	体重(kg)	標準化
1	2.9	−0.20
2	3.1	0.20
3	2.8	−0.40
4	3.2	0.40
5	3.4	0.80
6	2.2	−1.60
7	2.8	−0.40
8	2.7	−0.60
9	2.8	−0.40
10	4.0	2.00

平均M＝3.0kg
標準偏差S＝0.5kg

同じ1kgでも標準偏差は
20倍の違い

B君の体重

4.0＝3.00＋1.0(kg)

平均と1kgの違いだが、
平均からの隔たりが大きい

共分散

表　小１児童の身長と体重

番号	身長(cm)	体重(kg)	偏差積
1	111.6	20.1	−3.72
2	122.5	24.3	22.56
3	123.9	22.7	19.52
4	109.2	15.3	36.12
5	115.9	21.8	−4.37
6	128.3	23.2	36.85
7	115.3	19.1	1.00
8	111.4	12.8	42.88
9	121.7	19.7	0.76
10	118.6	16.2	−2.64
平均	117.8	19.5	
標準偏差	5.9	3.6	

偏差積の計算方法

偏差×偏差
＝(身長－身長の平均)×(体重－体重の平均)
＝(122.5－117.8)×(24.3－19.5)＝22.56

共分散の計算方法

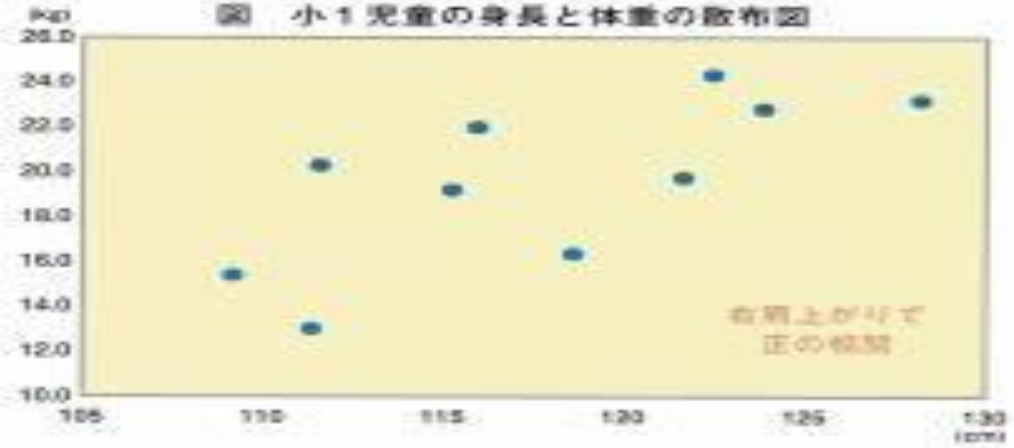
偏差積の平均

$$\frac{[-3.72+22.56+\cdots+(-2.64)]}{10}=15.1$$

正の値をとっている（共分散＞0）

正の相関のとき，
共分散＞0
負の相関のとき，
共分散＜0
となる

図　小１児童の身長と体重の散布図



共分散と相関係数

表　小1児童の身長と体重

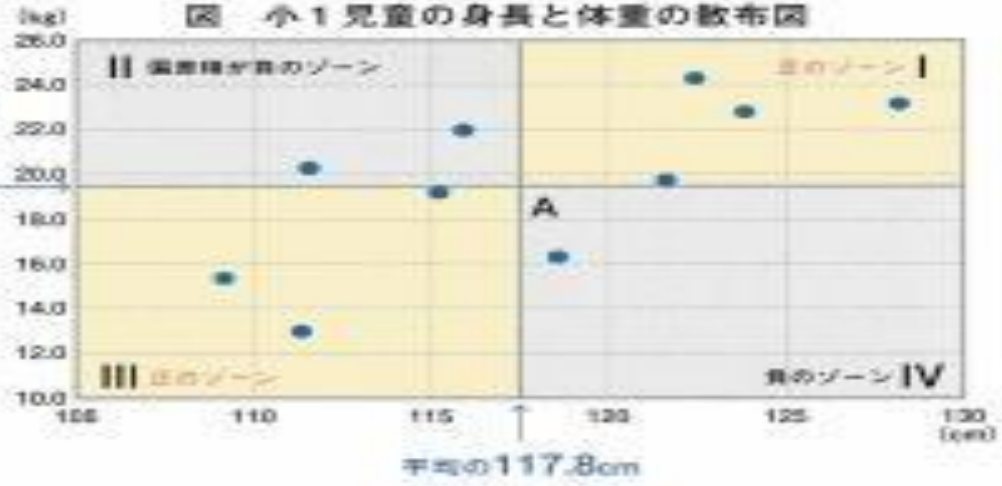
番号	身長(cm)	体重(kg)	身長の偏差	体重の偏差	偏差積
1	111.6	20.1	−6.2	0.6	−3.72
2	122.5	24.3	4.7	4.8	22.56
3	123.9	22.7	6.1	3.2	19.52
4	109.2	15.3	−8.6	−4.2	36.12
5	115.9	21.8	−1.9	2.3	−4.37
6	126.3	23.2	10.5	3.7	38.85
7	115.3	19.1	−2.5	−0.4	1.00
8	111.4	12.8	−6.4	−6.7	42.88
9	121.7	19.7	3.9	0.2	0.78
10	116.6	16.2	0.8	−3.3	−2.64
平均	117.8	19.5		共分散	15.1
標準偏差	5.9	3.6			

← 共分散は偏差積の平均、この例では正となる

表裏関係にして、身長と体重の平均値で区切ってみると？

?

図　小1児童の身長と体重の散布図



単位に依存しない相関の指標

相関係数の算式

$$\frac{\text{身長と体重の共分散}}{\text{身長の標準偏差} \times \text{体重の標準偏差}}$$

分母は必ず正なので

相関係数の正負と共分数の正負は一致することがわかる

Part1
終