多変量データの扱い

Rejoui

株式会社 Rejoui
〒151-0053
東京都渋谷区代々木2-30-4
Mail: info@rejoui.co.jp
URL: http://www.rejoui.co.jp

多変量データの扱い①

Agenda

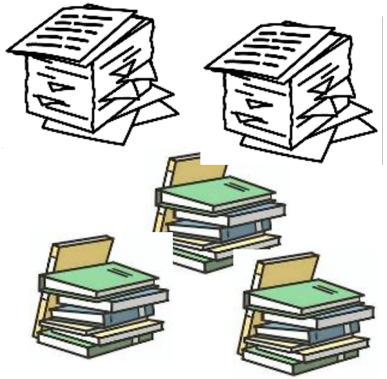
- データとはなにか
- 多変量データとは
- 多変量データを扱うための“プロセス”
- 単変量解析：基本統計量とは
- 単変量解析：時系列解析

データとは何か

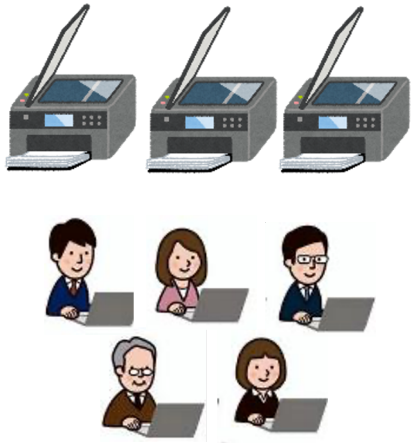
データとは「情報」ですが、21世紀現在においての「データ」とは「機械判読可能」な情報であるかどうか（コンピュータに投入して処理することができるか）、という観点が不可欠です。

紙に記録された情報も“データ”ですが、紙のままではコンピュータに解析させることができません。
デジタル化された情報の活用がここまで叫ばれるのは「情報をデジタルに変換する」「そこから価値を創出する」ことが期待されているからです

書類の山

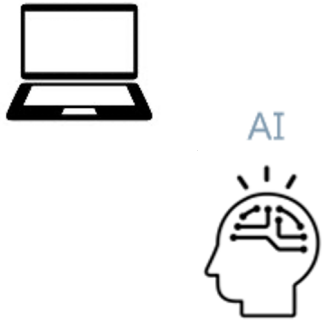


スキャン・打ち込み



電子ファイル

コンピュータで解析可能に



新たな価値に

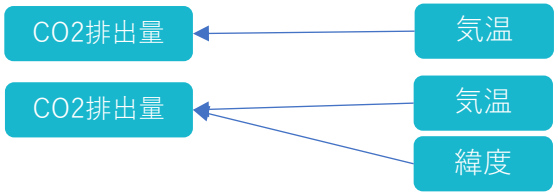
多変量データとは

多変量とは多くの情報（変数に関するデータ）を取り扱うことをいいます。分析者は、何らかの仮説に基づいて多変量のデータ間の関連性を明らかにします。

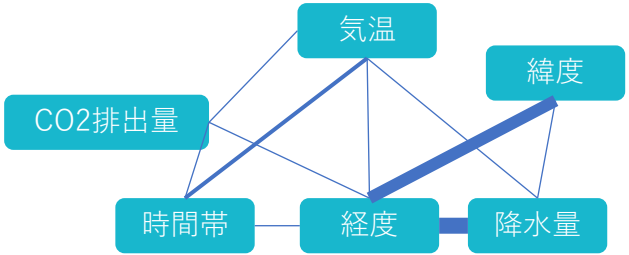
多数の情報（データ）

NO	チョコレートの消費量	CO2排出量	気温	降水量	緯度	経度
000000	2 4 3 5	491	31	3,240	32.6	141.08
000001	3 4 2 3	399	31	1 , 4 40	33.7	137.35
000002	1 1 1 4	462	11	1 ,240	33.3	136.37
000003	9 8 0	514	17	6 ,240	33.0	142.22
000004	3 2 4 5	379	15	3 ,423	32.5	141.13
000005	2 3 4 3	399	10	432	32.3	138.34
000006	9 3 4	378	25	980	32.5	138.97
000007	1 9 3 2	455	28	1,241	32.8	140.22
000008	3 4 3 2	425	15	2,643	33.4	139.29

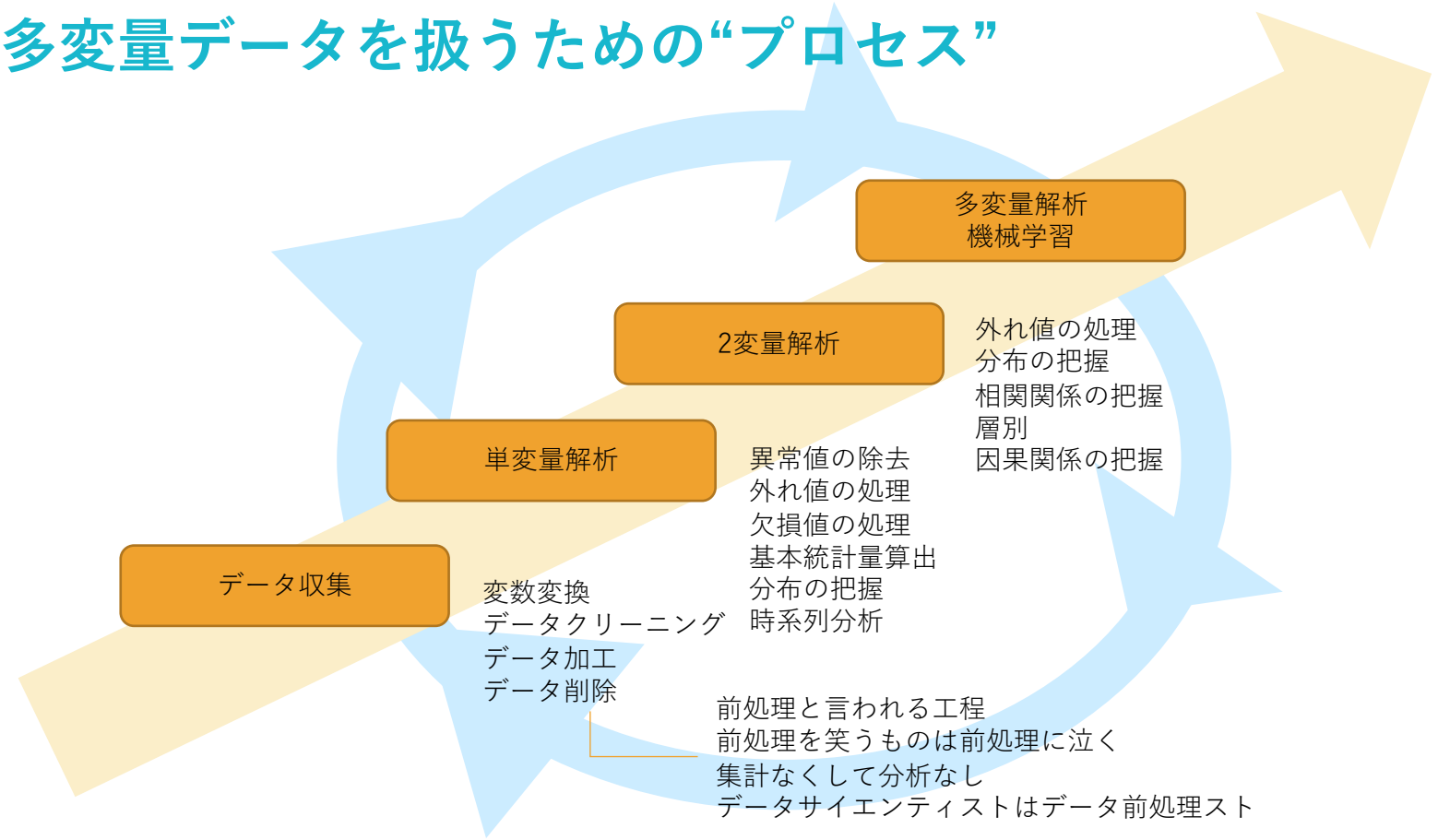
データから他のデータを予測する



データ同士の関係性を把握する



多変量データを扱うための“プロセス”



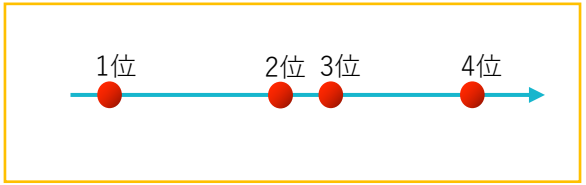
データとは何か　データの尺度 – 1

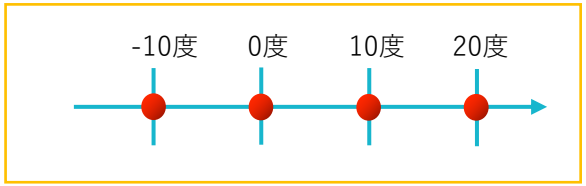
①名義尺度・・・“分類”を目的とした尺度
例：「富士山・剣岳・間ノ岳」「赤・緑・青」など
アンケート：「性別・職業」「所有メーカー」
順序に意味がない

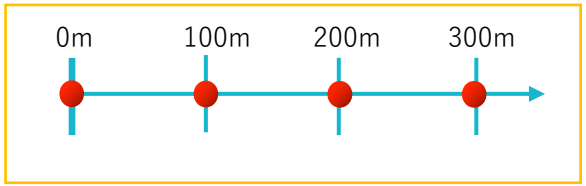
②順序尺度・・・“順序”にも意味がある尺度
例：「金・銀・銅」「1位・2位・3位」
アンケート：「当てはまる～当てはまらない」
間隔は定かではない

③間隔尺度・・・順序尺度で、かつ“間隔”にも意味がある尺度
例：「温度(摂氏)」「暦年(平成、西暦)」
0に意味がない

④比尺度・・・間隔尺度で、かつ“原点”に意味がある尺度
例：「距離・大きさ」「お金」
アンケート：「1年間で旅行した回数」 原点は「ない」0
1番偉い尺度







データとは何か　データの尺度 – 2

①名義尺度・・・“分類”を目的とした尺度
　例：「富士山・剣岳・神之峰」「赤・緑・青」など
　　アンケート：「性別・職業」「所有メーカー」
　順序に意味がない

②順序尺度・・・“順序”にも意味がある尺度
　例：「金・銀・銅」「1位・2位・3位」
　　アンケート：「当てはまる～当てはまらない」
　間隔は定かではない

質的データ

③間隔尺度・・・順序尺度で、かつ“間隔”にも意味がある尺度
　例：「温度(摂氏)」「暦年(平成、西暦)」
　　0に意味がない

④比尺度・・・間隔尺度で、かつ“原点”に意味がある尺度
　例：「距離・大きさ」「お金」
　　アンケート：「1年間で旅行した回数」
　1番偉い尺度

量的データ

単変量解析/2変量解析

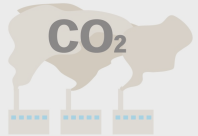
多変量解析は、単変量解析／２変量解析の結果を経て単変量を沢山集めたもの。まずは単変量／２変量解析を行うことが重要です。

単変量

１つの対象に１つのデータ
例：国別チョコレートの消費量

２変量

１つの対象に２つのデータ
例：国別チョコレートの消費量と
二酸化炭素の排出量



NO	チョコレートの消費量	CO2排出量	気温	降水量	緯度	経度
000000	２４３５	491	31	3,240	32.6	141.08
000001	３４２３	399	31	1,440	33.7	137.35
000002	１１１４	462	11	1,240	33.3	136.37
000003	９８０	514	17	6,240	33.0	142.22
000004	３２４５	379	15	3,423	32.5	141.13
000005	２３４３	399	10	432	32.3	138.34
000006	９３４	378	25	980	32.5	138.97
000007	１９３２	455	28	1,241	32.8	140.22
000008	３４３２	425	15	2,643	33.4	139.29

単変量解析/2変量解析

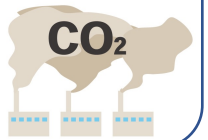
多変量解析は、単変量解析／２変量解析の結果を経て単変量を沢山集めたもの。まずは単変量／２変量解析を行うことが重要です。

単変量

１つの対象に１つのデータ
例：国別チョコレートの消費量

２変量

１つの対象に２つのデータ
例：国別チョコレートの消費量と
二酸化炭素の排出量



NO	チョコレートの消費量	CO2排出量	気温	降水量	緯度	経度
000000	２４３５	491	31	3,240	32.6	141.08
000001	３４２３	399	31	1 , 4 40	33.7	137.35
000002	１１１４	462	11	1 ,240	33.3	136.37
000003	９８０	514	17	6 ,240	33.0	142.22
000004	３２４５	379	15	3 ,423	32.5	141.13
000005	２３４３	399	10	432	32.3	138.34
000006	９３４	378	25	980	32.5	138.97
000007	１９３２	455	28	1,241	32.8	140.22
000008	３４３２	425	15	2,643	33.4	139.29

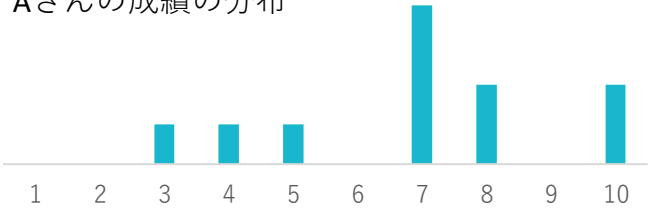
単変量解析：基本統計量

基本統計量はデータのそれぞれの基本的な特徴を表す値です。単変量解析では、まず基本統計量と分布の把握を行います。

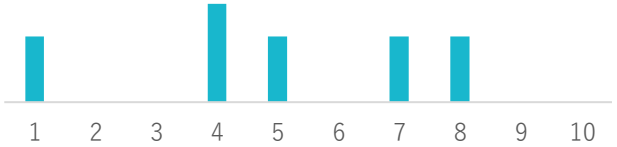
	現代文	古典	地理	数学 II	数学B	政治・経済	英語	音楽	家庭	情報	保健体育
Aさん	8	10	7	10	7	4	8	7	5	7	3
Bさん	7	5	8	5	7	1	4	4	1	4	8
Cさん	4	5	5	3	6	4	6	4	5	5	4

	Aさん	Bさん	Cさん
平均	6.91	4.91	4.64
標準誤差	0.67	0.74	0.28
中央値	7.00	5.00	5.00
最頻値	7.00	4.00	4.00
標準偏差	2.21	2.47	0.92
分散	4.89	6.09	0.85
尖度	-0.31	-0.76	-0.45
歪度	-0.34	-0.36	-0.02
範囲	7	7	3
最小	3	1	3
最大	10	8	6
合計	76	54	51
標本数	11	11	11

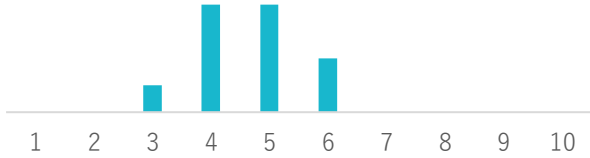
Aさんの成績の分布



Bさんの成績の分布

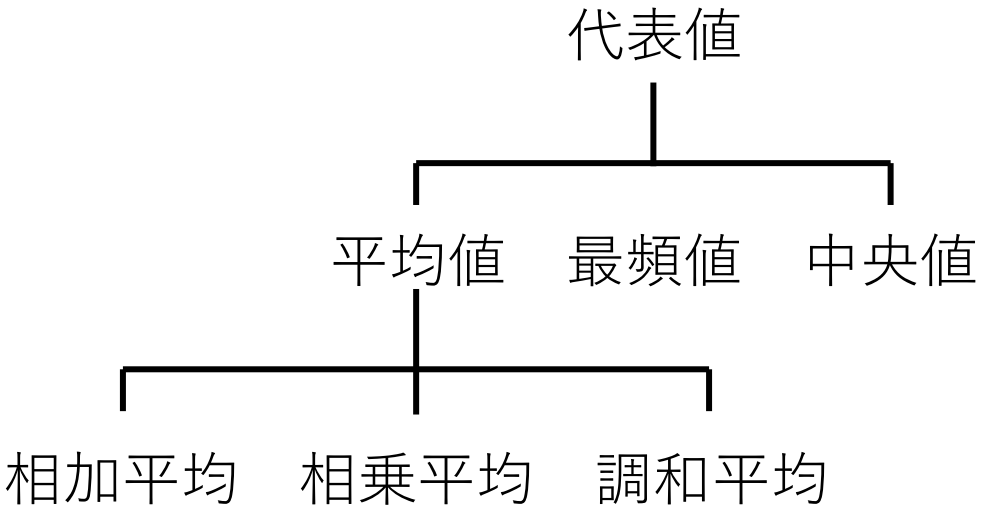


Cさんの成績の分布



代表値とは

統計資料の整理にあたり、多数の数値の分布に関する集団的特徴を表すために用いられる数値で、その値のまわりに分布しているとみられる中心的な位置を示す。



基本統計量

項目	意味
平均	全ての個別データを足しあげて、その合計をデータの個数で割り算した値
標準誤差	推定量の標準偏差であり、標本から得られる推定量そのもののバラつき
中央値	データを大きさの順に並べて、ちょうど真ん中にくる値。
最頻値	もっとも集中しているデータ値
標準偏差	データの散らばり度合いを表す値。値が大きいほど、データにばらつきがある
分散	データの散らばり度合いを表す値。標準偏差の二乗
尖度	分布の尖り具合を表す
歪度	分布の左右のゆがみ具合を表す
範囲	データの取る範囲 最大値-最小値
最小	データの最も小さい値
最大	データの最も大きい値
合計	データの合計
標本数	データの数

様々な分布

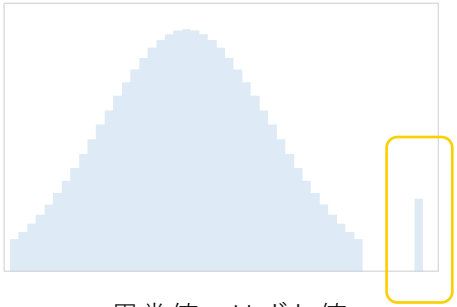
そのデータの分布状況を把握することは、データ分析の基本です。ヒストグラムを描き、データのばらつきや分布を把握します。

正規分布

ひずみのある分布

.ふた山

とがった分布



異常値・はずれ値

平均値・最頻値・中央値の違い

分布の中心（代表値）を表す値

①平均値

全ての個別データを合計し、データの個数で割り算した値

例：チョコレートの消費量全体を足しあげて、国の数で割った数値。

②最頻値

もっとも集中しているデータ値

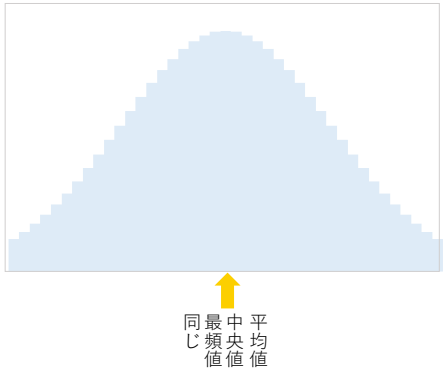
例：チョコレートの消費量を5kg刻みで分割し、1番たくさんの国が入った範囲。

③中央値

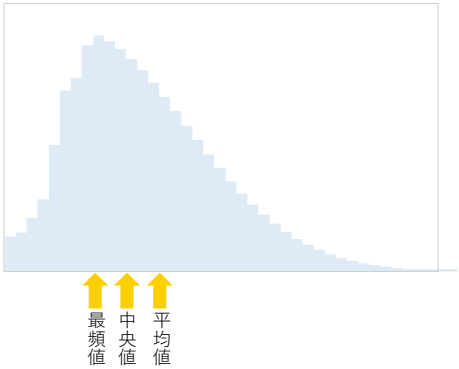
データを大きさの順に並べて、ちょうど真ん中にくる値。

例：チョコレートの消費量をの低い順に並ばせ、ちょうど真ん中に来る国の値。

正規分布

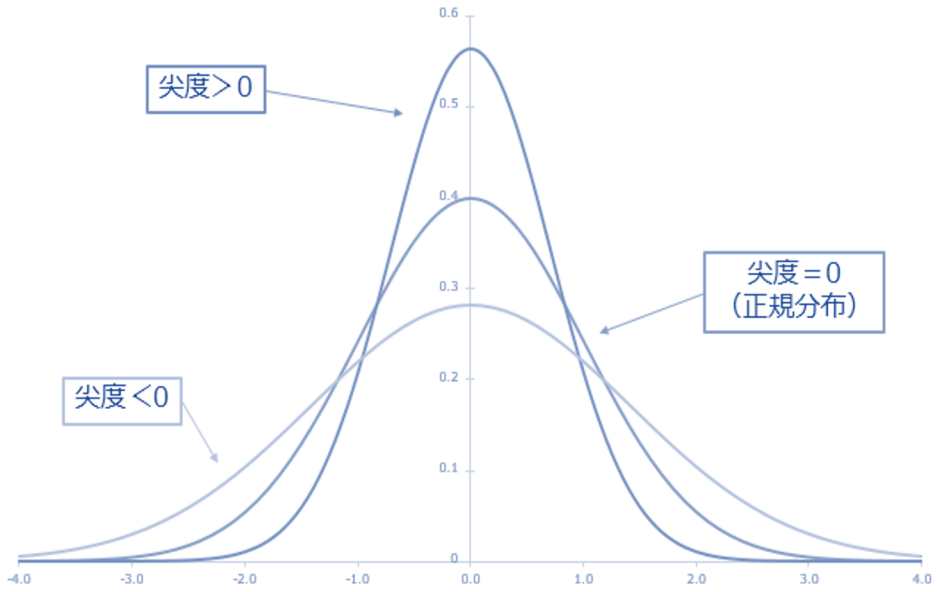


ひずみのある分布



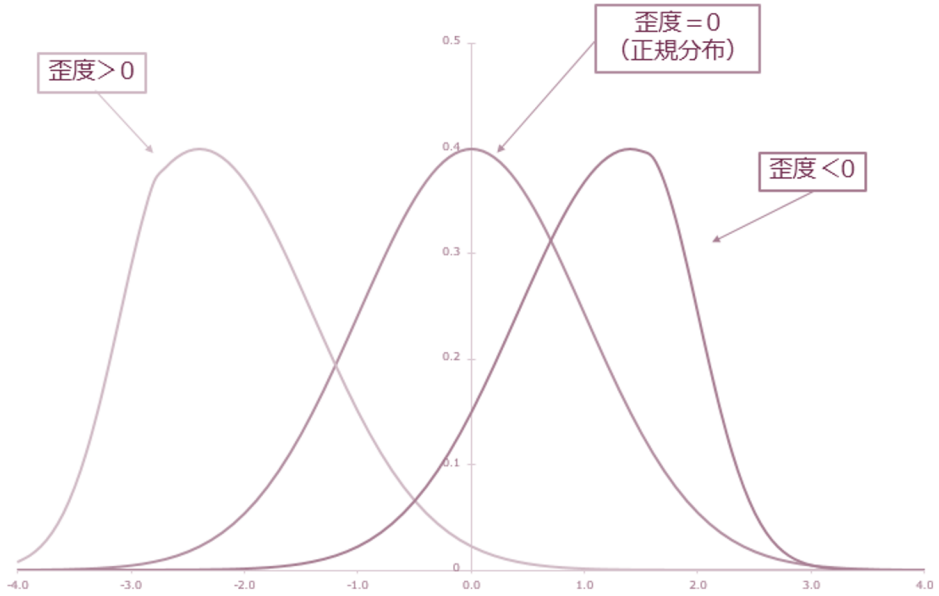
基本統計量（尖度：せんど）

データ分布のとがり具合を表す



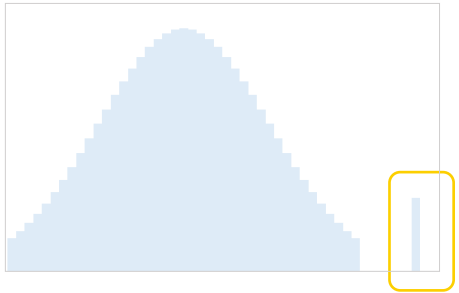
基本統計量（歪度：わいど）

分布の左右対称性（歪み）を表す

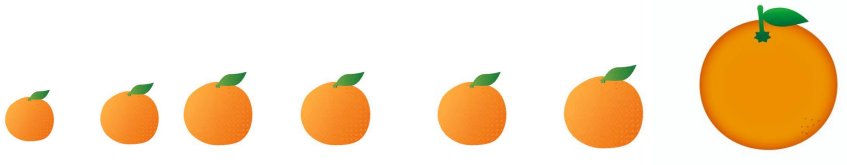


単変量解析（外れ値と異常値）

異常値か外れ値かの判断が必要。異常値は取り除くが外れ値は取り除くかどうか検討が必要



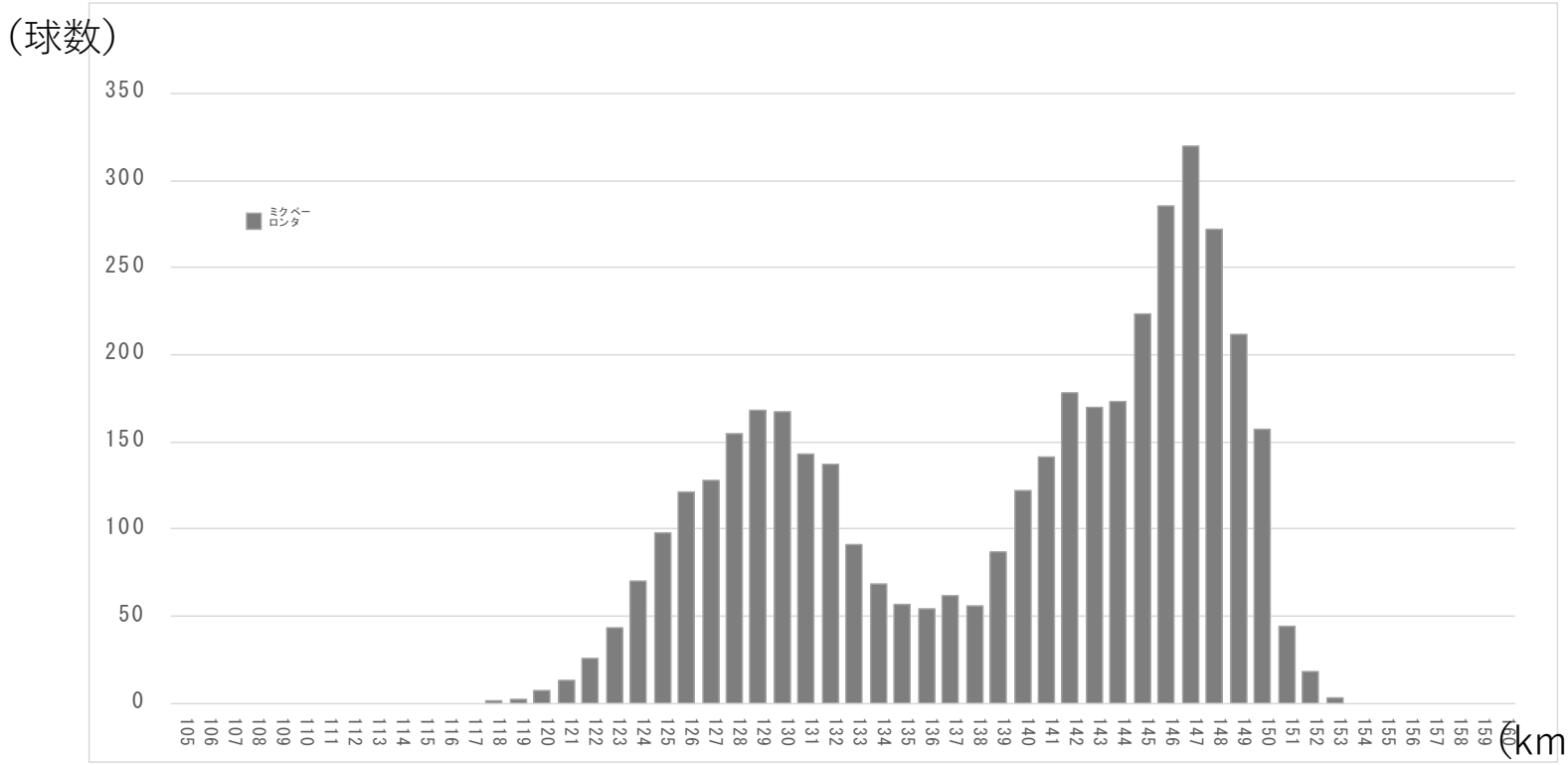
外れ値のなかでも「外れている理由」が分かっているものが「異常値」



時	分			
7	5	12		
8	5	20	35	50
9	5	20	35	50
10	5	20		
11	5	20		
12	5			
13	5			
14	5			
15	5	20		
16	5	20	35	
17	5	20	35	50
18	5	20	35	50
19	5	20	35	
20	5			
21	5			

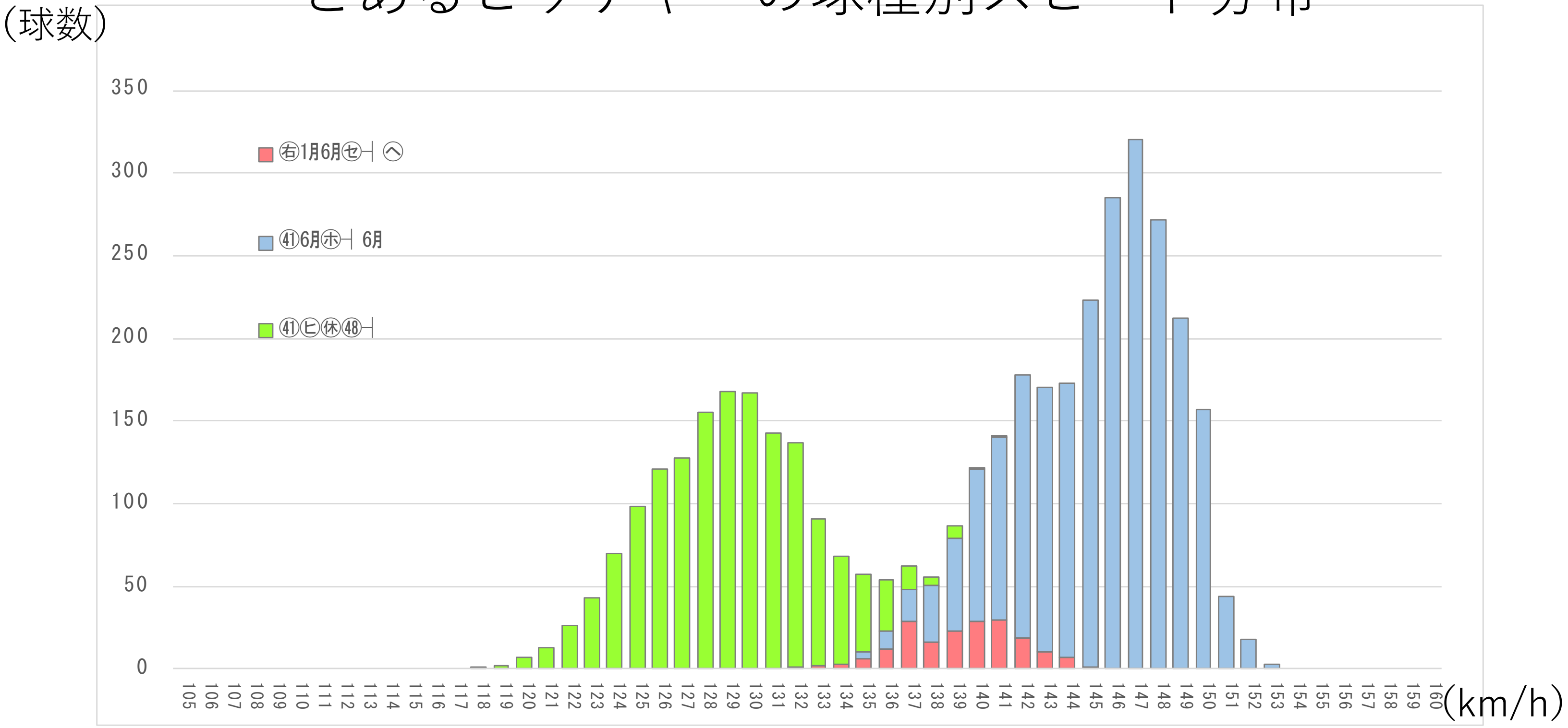
二こぶラクダのような分布の場合、平均にデータが少ないという事も起きます。「多峰性」とも表現します。

多峰性のある分布例


(球数)
350
300
250
200
150
100
50
0
ミクベーロンタ
105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
(km/h)

多峰性のある分布例

とあるピッチャーの球種別スピード分布



ばらつきを定量化する標準偏差と分散

$$分散 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

各データの平均との差を2乗した正方形の面積の平均

$$標準偏差（\sigma） = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

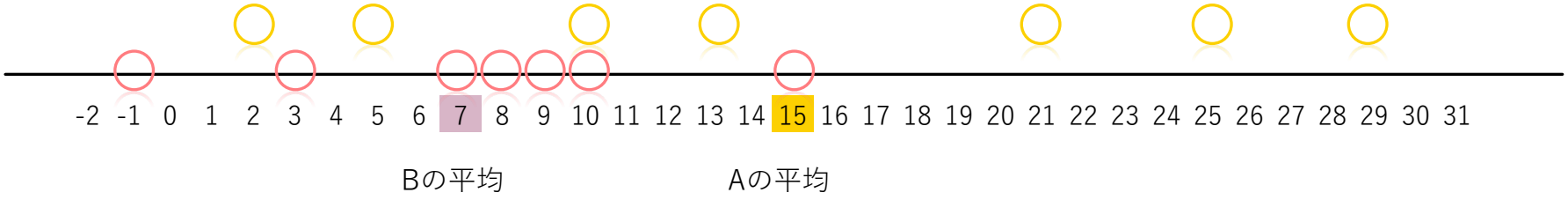
この部分が分散

注）ここでは話しを簡単にするために標本分散を用いる
不偏分散を算出する場合はn-1で割る

どちらの分散が大きいかを考えてみましょう

A（2, 5, 10, 13, 21, 25, 29）

B（-1, 3, 7, 8, 9, 10, 15）



分散の計算

次の7つのデータの分散の算出を考える

（2，5，10，13，21，25，29）

分散の計算

a.平均値

b.データの数

データ	c. 偏差 （データと平均との差）	d. 偏差の2乗 （cの2乗 / 偏差平方）
2		
5		
10		
13		
21		
25		
29		

e. dの合計　（偏差平方和）

e ÷ b　（分散）

分散の計算

a.平均値　15

b.データの数　7

データ	c. 偏差 （データと平均との差）	d. 偏差の2乗 （cの2乗 / 偏差平方）
2		
5		
10		
13		
21		
25		
29		

e. dの合計　　（偏差平方和）

e ÷ b　　（分散）

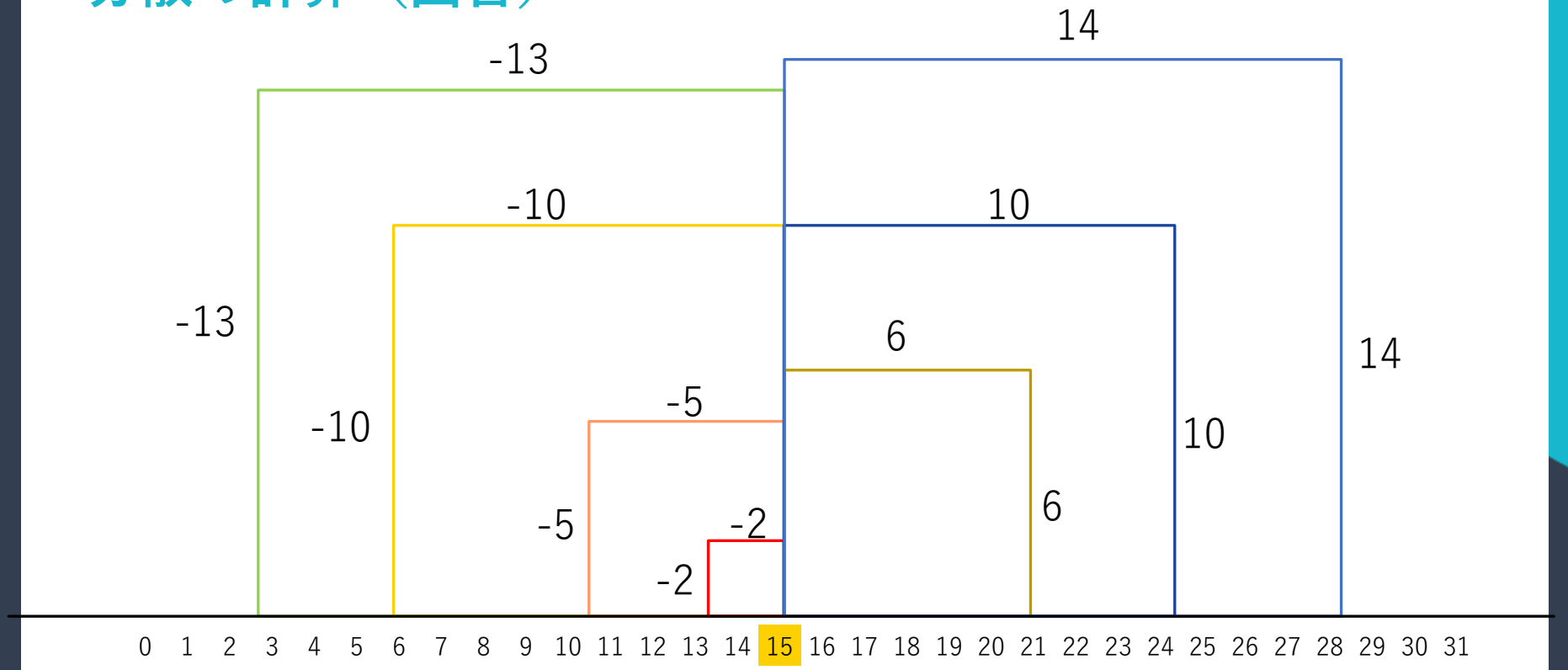
分散の計算

a.平均値	15
b.データの数	7

データ	c. 偏差 （データと平均との差）	d. 偏差の2乗 （cの2乗 / 偏差平方）
2	-13	169
5	-10	100
10	-5	25
13	-2	4
21	6	36
25	10	100
29	14	196

e. dの合計	630	（偏差平方和）
e ÷ b	90	（分散）

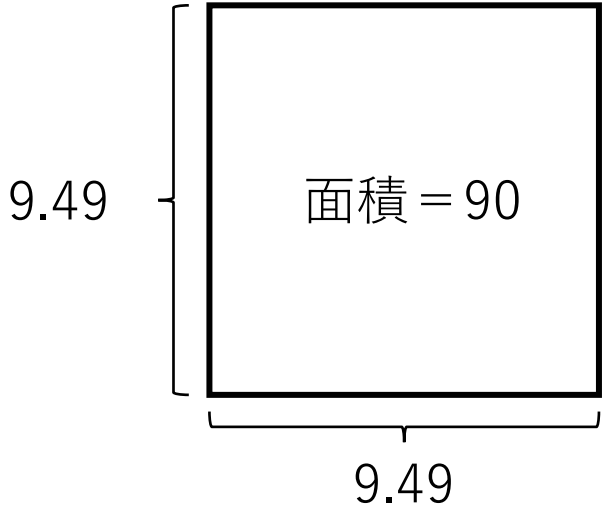
分散の計算（回答）



正方形の面積の平均が「分散」

分散と標準偏差

7つの正方形の平均の面積は90となった。
これが分散。



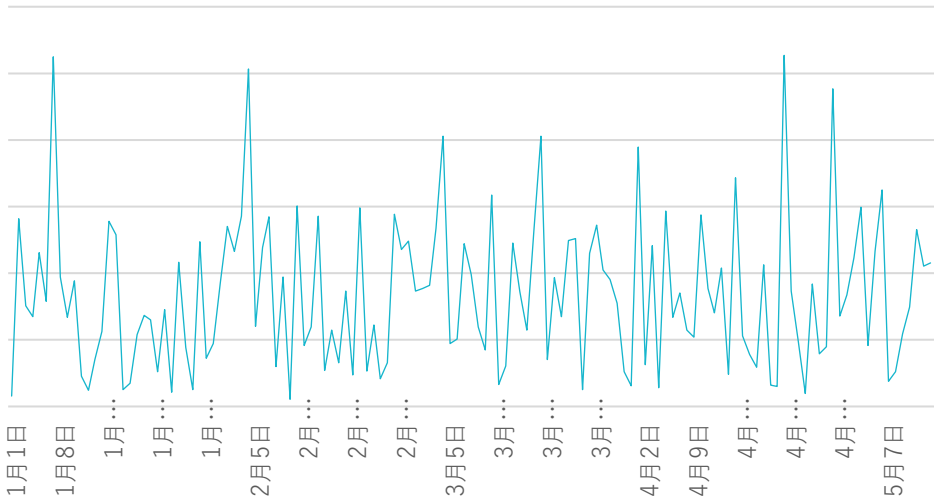
平均の面積（分散＝90）をルートしたもの
すなわち正方形の1辺の長さ9.49が標準偏差。

単変量解析 ー 時系列データ

時系列データとは、時間の経過とともに変化するデータ

・気温など気象情報
・株価など経済現象
・売上高など商業データ
・アクセス数など生活者行動履歴

- 時系列データ
  - 連続時間時系列
  - 離散時間時系列



時系列データは様々な要因が絡み合っている場合が多いので要因分解が課題となる
曜日や時間帯、季節性や商習慣などの法則性が潜むことは多い