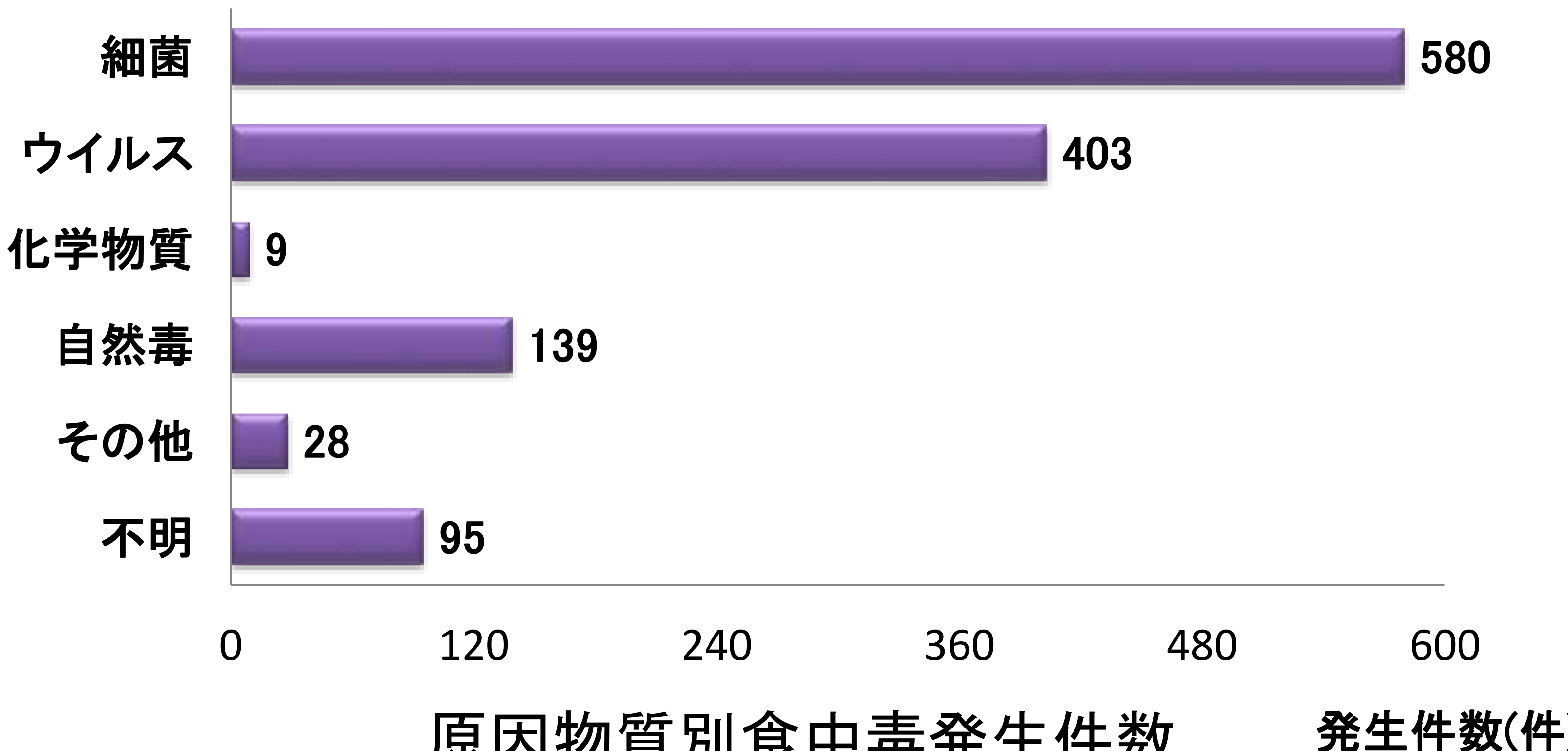
質的データの集計-度数分布表

- 度数
  - それぞれの選択肢が選択された数
- 比率(相対度数)
  - データ全体(ケース数)を分母とした時の各選択肢が選択された数(度数)の割合

度数分布表の例
原因物質別食中毒発生件数

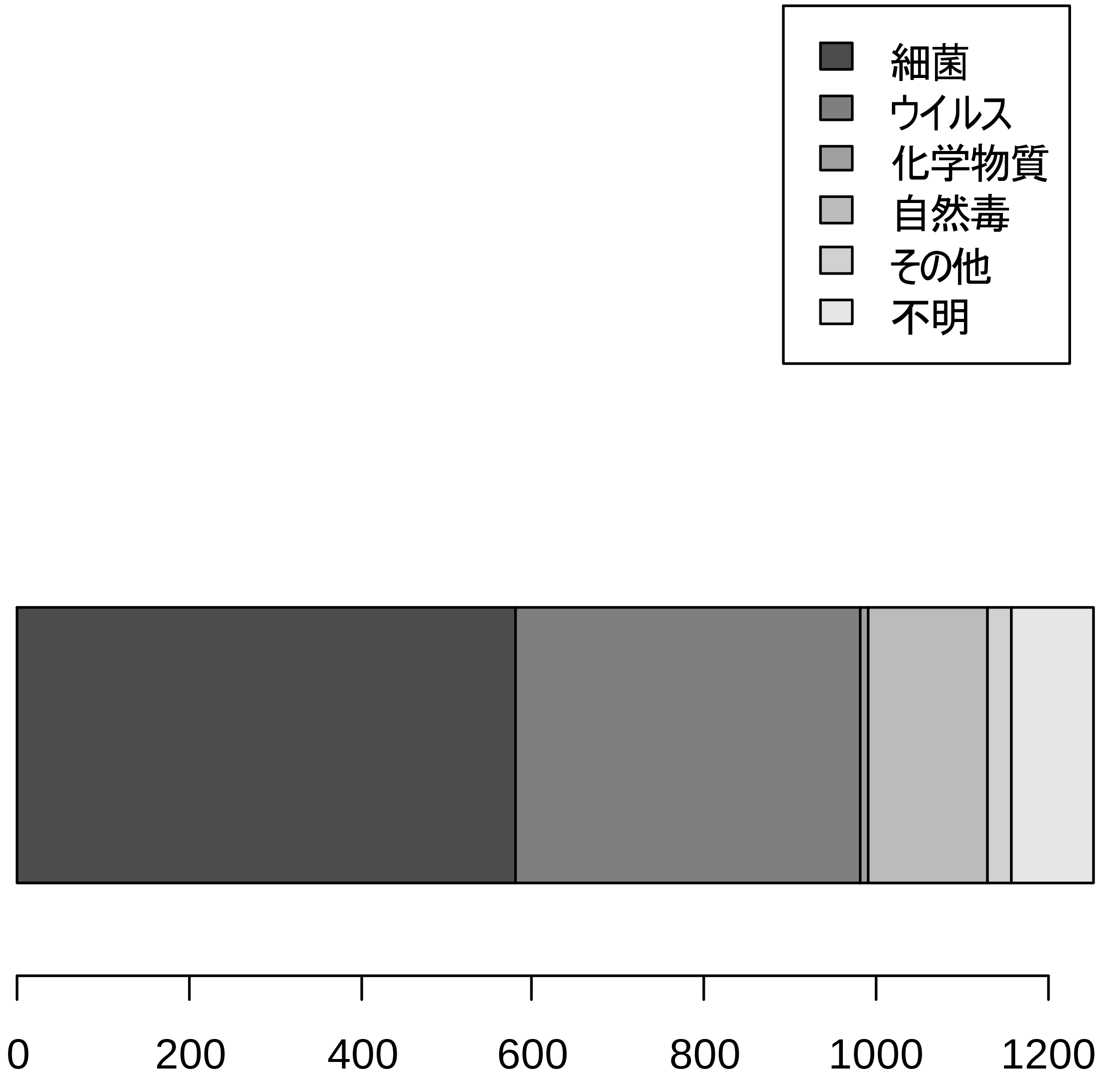
原因物質	発生件数(度数)	発生割合(比率)
細菌	580	46.3
ウイルス	403	32.1
化学物質	9	0.72
自然毒	139	11.1
その他	28	2.23
不明	95	7.58
総数	1254	100

度数分布表のグラフ化-棒グラフ



原因物質別食中毒発生件数

度数分布表のグラフ化-帯グラフ



量的変数の度数分布(1)

身長(cm)	度数	比率(%)
144	1	0.01
145	0	0
146	1	0.01
147	0	0
148	1	0.01
…	…	…
189	5	0.05
190	5	0.05
191	3	0.03
192	0	0
193	2	0.02

値の数が多く表が膨大になる
→値を適当な幅(階級)で区切って集計すると見やすくなる

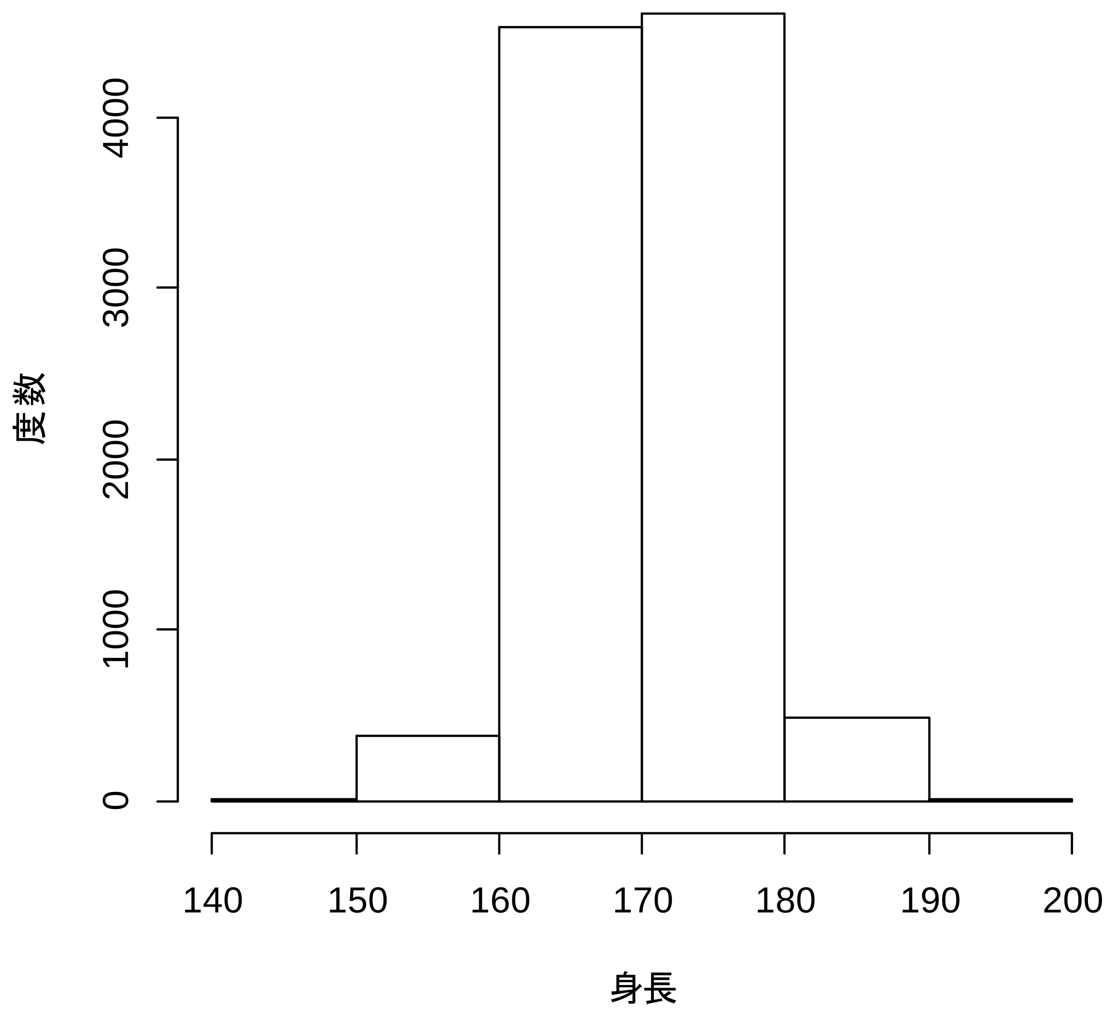
出典: 学校保健統計調査(2011)

量的変数の度数分布(2)

身長(cm)	度数	比率(%)	累積度数	累積比率
141-150	4	0.04	4	0.04
151-160	379	3.79	383	3.83
161-170	4525	45.25	4908	49.08
171-180	4602	46.02	9510	95.1
181-190	485	4.85	9995	99.95
191-200	5	0.05	10000	100
合計	10000	100		

累積度数: その階級以下の度数の合計
累積比率: その階級以下の比率の合計

グラフ(=ヒストグラム)にすると



量的変数の特徴の指標－代表値

- 平均値（算術平均値）
  - 全てのデータを足し合わせ、足し合わせた数で割ることで求められる値
  - はずれ値の影響を受ける
  - 比尺度や間隔尺度の場合
- 中央値
  - 全てのデータを小さい順に並べた時に真ん中に来る値のこと
    - データ数(ケース数)が偶数(2n)の場合は真ん中の値が1つに決まらないため、n番目のケースとn+1番目のケースの値を足して2で割った値が中央値
  - 歪んだ分布やはずれ値が多い分布では中央値を用いることが多い
  - 順序尺度を用いる場合
- 最頻値
  - データの出現率が最大の値
  - 名義尺度のときなど

平均値と中央値と最頻値

元のデータ

店	価格
A	9800
B	10080
C	9980
D	10150
E	9700
F	9800
G	9980
H	9800
I	9990

価格順に並び替えたデータ

店	価格
E	9700
A	9800
F	9800
H	9800
C	9980
G	9980
I	9990
B	10080
D	10150

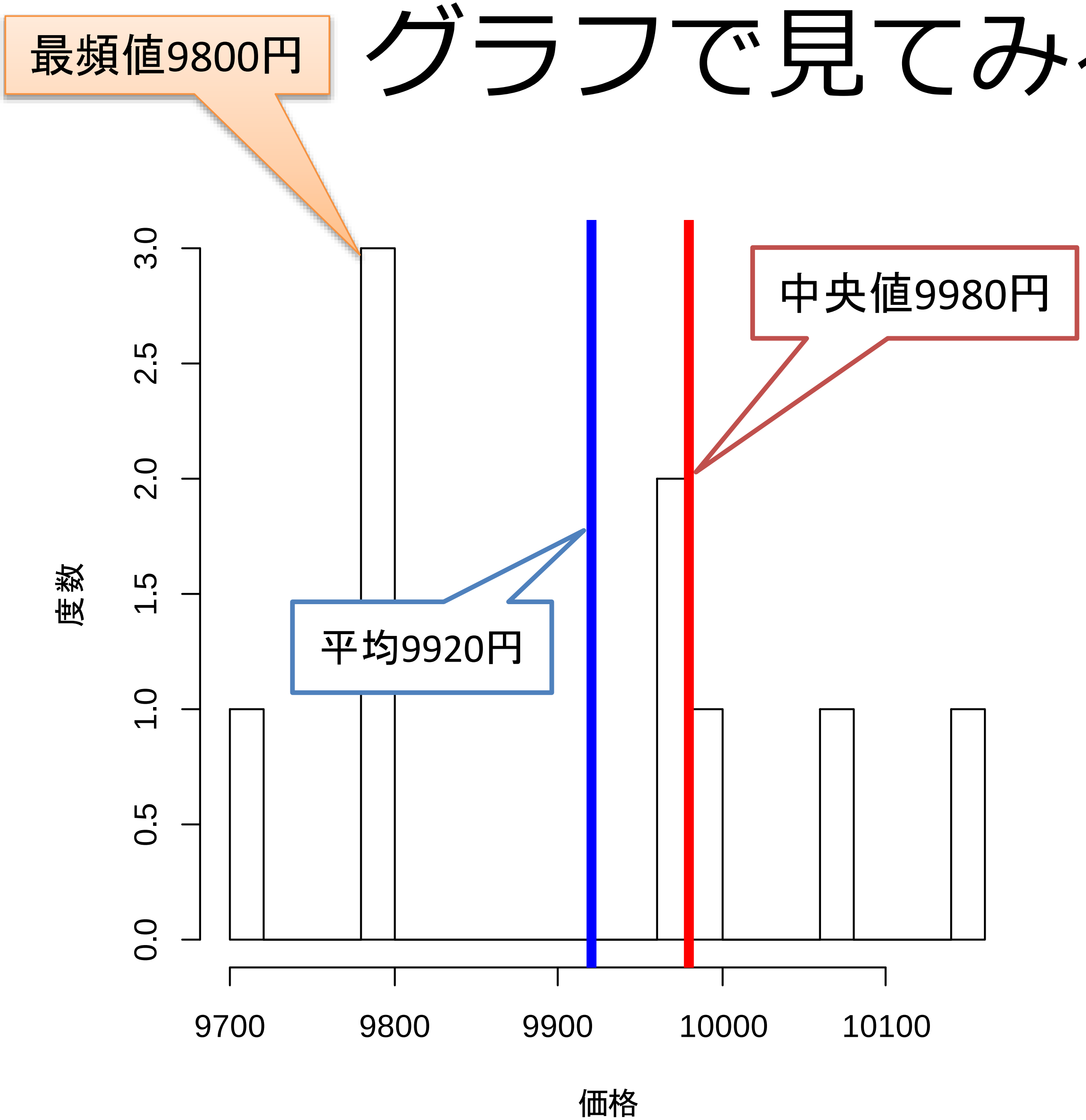
度数分布表

価格	件
9700	1
9800	3
9980	2
9990	1
10080	1
10150	1

平均値
=(9800+10080+9980+10150+9700+9800+9980+9800+9990)÷9=9920円

中央値=9980円
最頻値=9800円

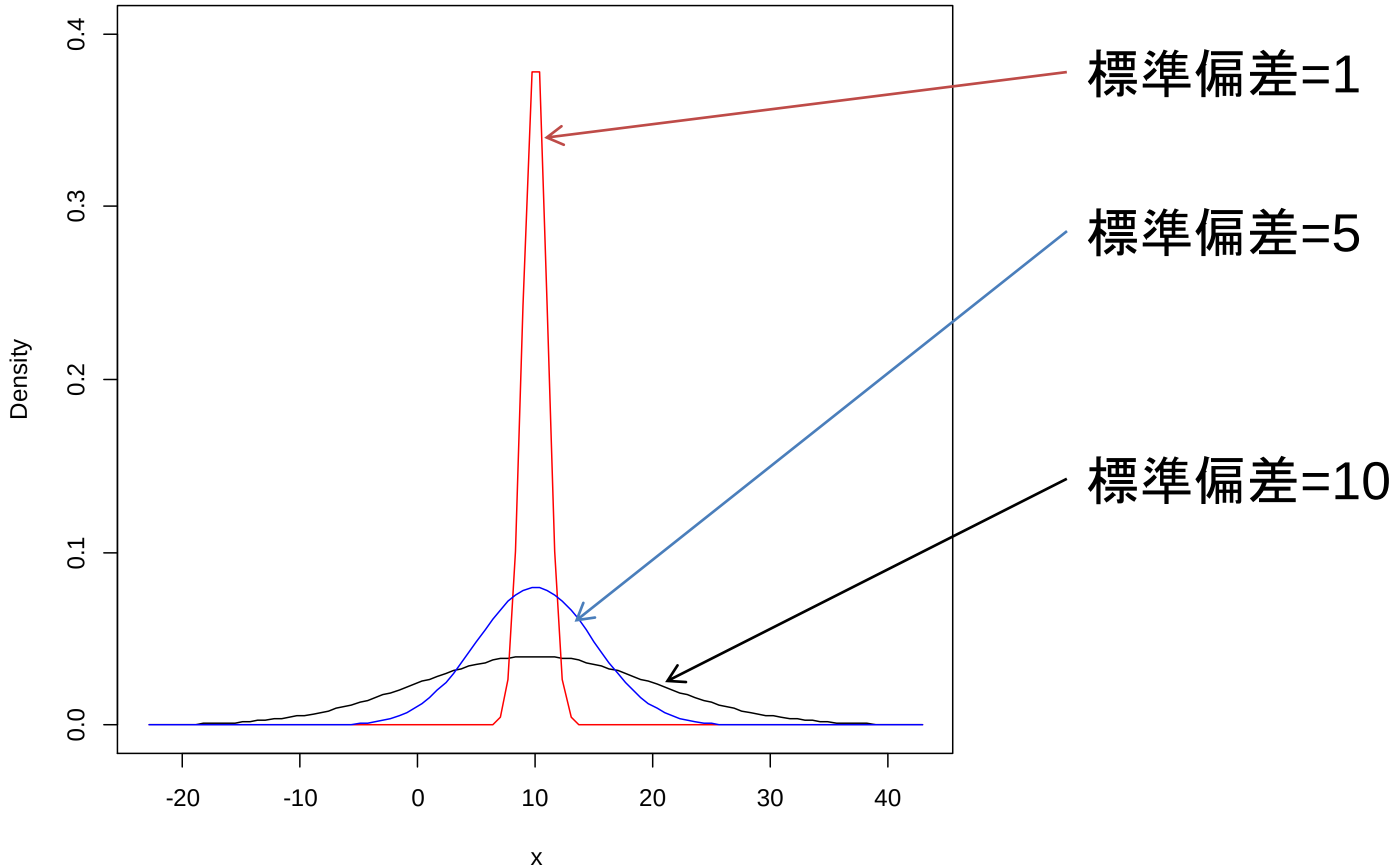
グラフで見てみる



価格	件
9700	1
9800	3
9980	2
9990	1
10080	1
10150	1

量的変数の特徴の指標－ばらつき

- 平均が同じでも分布の形状は異なる
  →ばらつきにも注目して分布の特徴を捉える



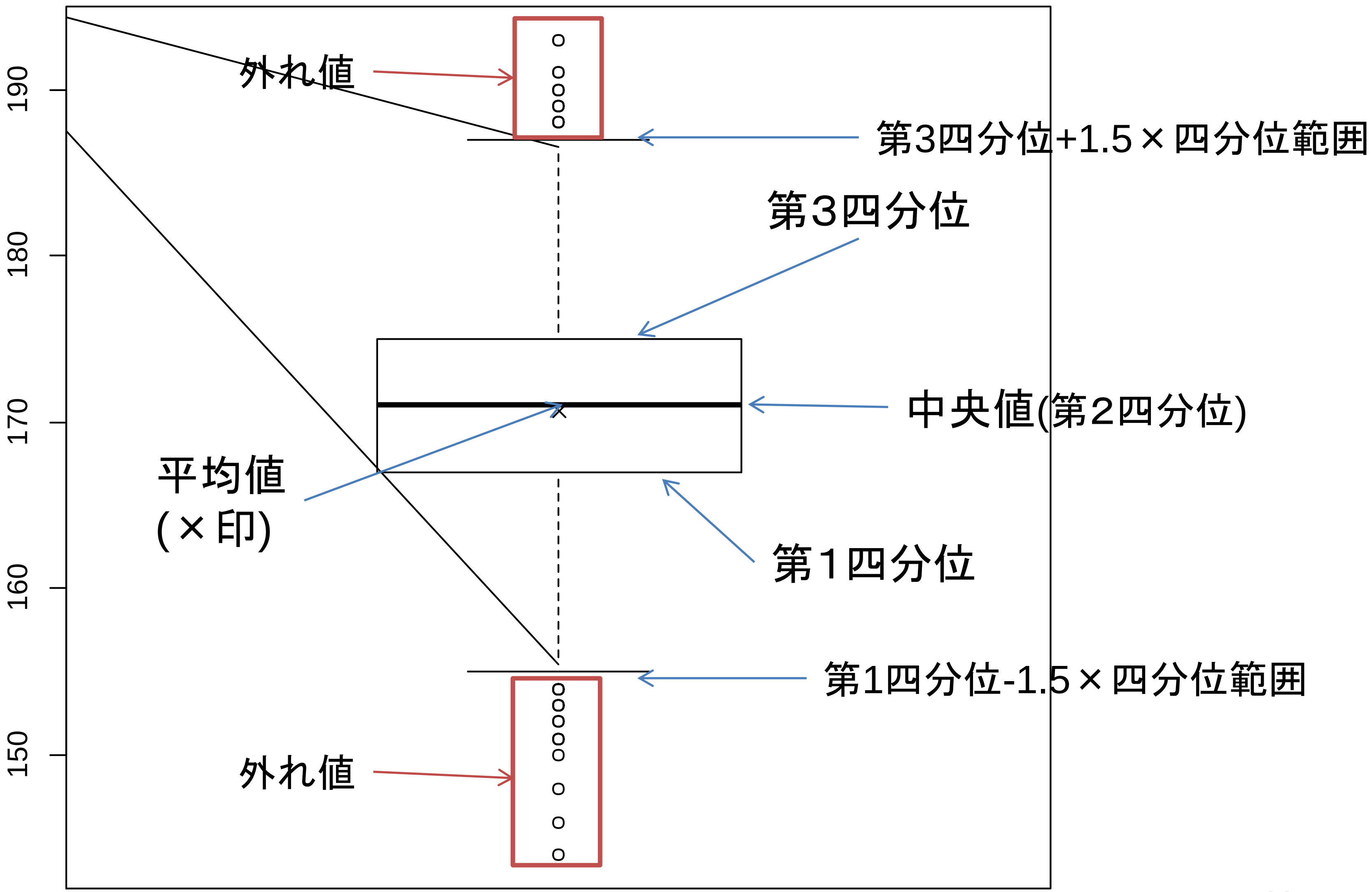
ばらつきの指標(1) – 範囲と分位数

- 範囲＝最大値-最小値
- 分位数： 変数を値の順番に並べた上で等しいサイズに分割する値
  - 4分の1ずつに区切る4分位や5分の1ずつ区切る5分位がよく使われる
- 第3四分位数と第1四分位数の差→四分位範囲(外れ値の影響を受けにくい)

最小値	144cm
第1四分位数	167cm
中央値(第2四分位数)	171cm
平均値	170.7cm
第3四分位数	175cm
最大値	193cm

範囲: 193-144=49
四分位範囲: 175-167=8

範囲などをまとめて表す－箱ひげ図



ばらつきの指標(2)－標準偏差と分散

- 平均値とデータとの差を「偏差」という
- 偏差の二乗をすべて合計し、ケース数で割る（この場合１０で割る）と分散になる分散のルート（平方根）をとると標準偏差になる
- 一般に結果としては平均値と標準偏差を報告する

学生	得点 (x)	偏差 (x−m)	偏差^2 $(x-m)^2$
A	61	−9	81
B	74	4	16
C	55	−15	225
D	85	15	225
E	68	−2	4
F	72	2	4
G	64	−6	36
H	80	10	100
I	82	12	144
J	59	−11	121
平均（m）	70	分散(σ^2)	95.6
		標準偏差(σ)	9.8

標準化・標準得点・偏差値

- 標準化
  - 平均、分散、単位が異なるデータを直接比較可能にする方法
  - 異なる変数間で平均と標準偏差が同じになるように変換

- 標準得点(z得点)
  - 平均が0、標準偏差が1になるように変換(標準化)した値

- 偏差値
  - 偏差値＝50＋標準得点×10

標準得点、偏差値の例

	得点	標準得点	偏差値
A	61	−0.9	40.8
B	74	0.4	54.1
C	55	−1.5	34.7
D	85	1.5	65.3
E	68	−0.2	48.0
F	72	0.2	52.0
G	64	−0.6	43.9
H	80	1.0	60.2
I	82	1.2	62.2
J	59	−1.1	38.8
平均	70		
標準偏差	9.8		