統計学

第2回　統計の基礎2

授業内容

- 平均・中央値・最頻値
- 順序統計量と箱ひげ図
- 分散と標準偏差

データの特徴を捉えるには

- グラフ化してその特徴を捉える
  - 統計図表
  - ヒストグラム
- 一つの数字で特徴を代表させる（統計量）
  - 平均
  - 分散、標準偏差、etc…

統計量

データの特徴を一つの数字に要約したもの

- データのどんな類の特徴を要約したいのか？　によって、様々な統計量
  - 平均
  - 分散、標準偏差
  - 中央値、最頻値

代表値

- データの分布の中心を表す数値（統計量）
  - 平均
  - 中央値
  - 最頻値

散布度

- データの分布の広がり具合を表す数値（統計量）
  - 分散
  - 標準偏差
  - レンジ（範囲）
  - 四分位範囲

平均・中央値・最頻値

平均

データの合計をデータ数で割る

n 個のデータがあって、

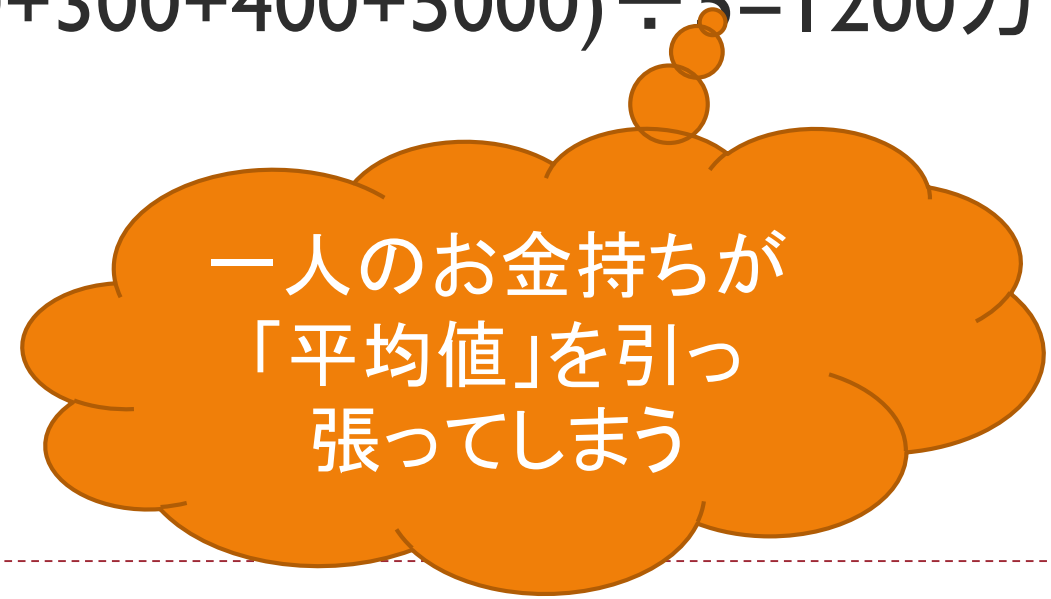
それぞれのデータを $x_1,\ x_2, \ldots, x_{n-1},\ x_n$ とすれば、

算術平均$= (x_1 + x_2 + \ldots + x_{n-1} + x_n) \div n$

$$\left(= \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

平均の計算例

- 100万、200万、300万、400万、500万
  - (100+200+300+400+500)÷5=300万円
- 100万、200万、300万、400万、5000万
  - (100+200+300+400+5000)÷5=1200万円



平均の性質

- ヒストグラムが左右対称である場合、その対称軸の通る点が平均値になる
- 極端に外れている値があると、平均値はその値に「引っ張られて」しまう
- 平均＝データを平らに均した値

中央値と最頻値

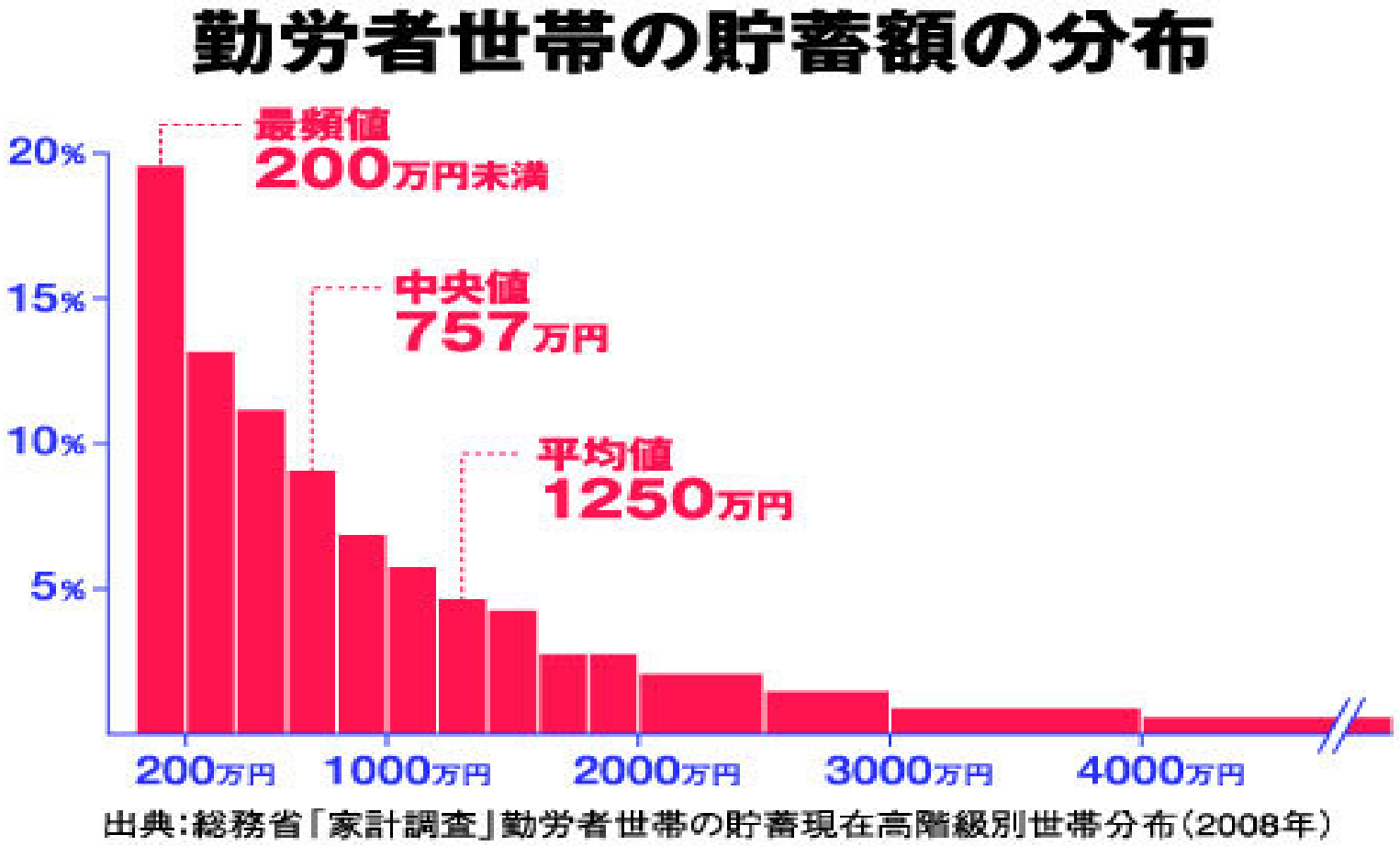
- 中央値：データを小さい順に並べたときに丁度真ん中にくるデータの値
- 最頻値：最も重い頻度で出てくる（階級の度数が高い）データの値

n 個のデータがあって、

それぞれのデータを小さい順に $x_1,\ x_2,\ \ldots,\ x_{n-1},\ x_n$ と並べると、

中央値：$x_{(n+1)/2}$（n が奇数のとき）、$\dfrac{x_{n/2}+x_{(n+2)/2}}{2}$（$n$ が偶数のとき）

例（平均値、中央値、最頻値）


勤労者世帯の貯蓄額の分布
最頻値
200万円未満
中央値
757万円
平均値
1250万円
20%
15%
10%
5%
200万円
1000万円
2000万円
3000万円
4000万円
出典：総務省「家計調査」勤労者世帯の貯蓄現在高階級別世帯分布（2008年）

度数分布曲線（度数折れ線）と代表値

- 平均値：度数分布曲線の重心の横座標（x座標）
  - 重心:度数分布曲線を厚紙に書き写して切ったものを作成したときに、厚紙の釣り合いがとれる場所
- 中央値：度数分布曲線と横軸とで囲まれた部分の面積を考えたとき、面積をちょうど半分にする縦線の横座標（x座標）
- 最頻値：山の頂上の横座標（x座標）

中央値の利用例

- 相対的貧困率：所得の中央値の半分以下の所得しかない世帯の割合
  - 所得＝可処分所得（実際に使える金額）を世帯の人数に応じて調整した額
  - 中央値を使うことで、異常な値の影響を避ける効果
  - 但し、中央値を使う場合、中間層が崩壊すると相対的貧困率は改善されるという弱点がある

相対的貧困率の計算あれこれ

- “トップ2割が金持ち国”の場合
  - 1000万円、120万円、110万円、100万円、90万円
  - 中央値110万円、その半分が55万円、貧困率0%
- “下位2割が貧乏人国”の場合
  - 1000万円、1000万円、980万円、960万円、300万円
  - 中央値980万円、その半分が490万円、貧困率は20%
  - 300万円の人の給料を200万円上げるか、980万円と960万円の人を590万円にすれば、相対的貧困率は0%になる

度数分布表から平均を計算する

- 階級値×度数÷データの総数=平均値
- 但し、生データから計算した平均値と、度数分布表から計算した平均値は若干ずれる
  - 度数分布表はデータの縮約の結果なので
  - でもまあ、実用に耐えうる範囲内

度数分布表から平均の計算例

階級	階級値	度数	階級値×度数
141～145	143	1	143
146～150	148	6	888
151～155	153	19	2907
156～160	158	30	4740
161～165	163	18	2934
166～170	168	6	1008

階級の中に入っているデータは全て階級値の値であるとみて計算

階級値×度数の計:12620

度数分布表からの平均:157.75cm

順序統計量と箱ひげ図

順序統計量

- データを、値の小さい順に並べたときの統計量
  - 最大値
  - 最小値
  - 中央値
  - 四分位値（四分位数）
  - レンジ
  - 四分位範囲

レンジ（範囲）

$$R = \text{最大値} - \text{最小値}$$

データが分布する幅を表す

計算が楽なので、データが小さい時は分散の代わりとして役に立つ。

でも、最大値と最小値以外のデータは無視することになるので、データが大きい時には……。

中央値

- 中央値：データを小さい順に並べたときに丁度真ん中にくるデータの値
  - もちろん、「大きい順に並べたときに丁度真ん中にくる値」でも可

n 個のデータがあって、

それぞれのデータを小さい順に $x_1,\ x_2,\ \ldots,\ x_{n-1},\ x_n$ と並べると、

中央値：$x_{(n+1)/2}$（n が奇数のとき）、$\dfrac{x_{n/2}+x_{(n+2)/2}}{2}$（$n$ が偶数のとき）

四分位値（四分位数）

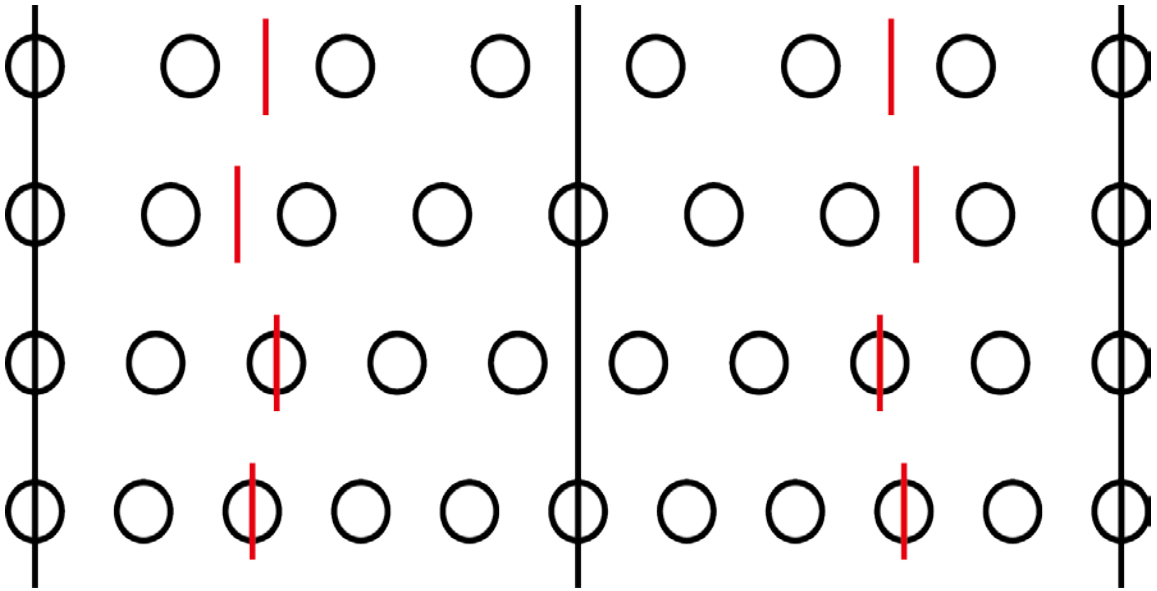
- 四分位値（四分位数）：データを小さい順に並べ、四等分したときの値
- 第３（上側）四分位値：中央値で2つに分けたデータの、上半分の中央値
- 第１（下側）四分位値：中央値で2つに分けたデータの、下半分の中央値
- 四分位範囲：第３（上側）四分位値と第１（下側）四分位値の差

五数要約

- 五数要約：最小値、第1四分位数、中央値、第3四分位数、最小値の5つの統計量でデータを説明すること
  - 最小値：一番小さい値
  - 最大値：一番大きい値
    - 範囲：最大値－最小値
  - 中央値：データを小さい順に並べたときに真ん中にくる値（第2四分位数）
  - 四分位数：データを小さい順に並べ、**四等分**したときの値
    - 第1四分位数：中央値で2つに分けたデータの、下半分の中央値（25%）
    - 第3四分位数：中央値で2つに分けたデータの、上半分の中央値（75%）
    - 四分位範囲：第3（上側）四分位数と第1（下側）四分位数との差
    - 四分位偏差：四分位範囲の半分

高校の教科書では

- データの個数が偶数個の場合、中央値で2つのグループに分け、それぞれのグループの中央値を四分位値にする
- データの個数が奇数個の場合、中央値を除外して2つのグループに分け、それぞれのグループの中央値を四分位値にする



EXCELによる四分位値の計算

- 第1（下側）四分位値
  - QUARTILE(データの範囲,1)
  - PERCENTILE(データの範囲,0.25)
- 第3（上側）四分位値
  - QUARTILE(データの範囲,3
  - PERCENTILE(データの範囲,0.75

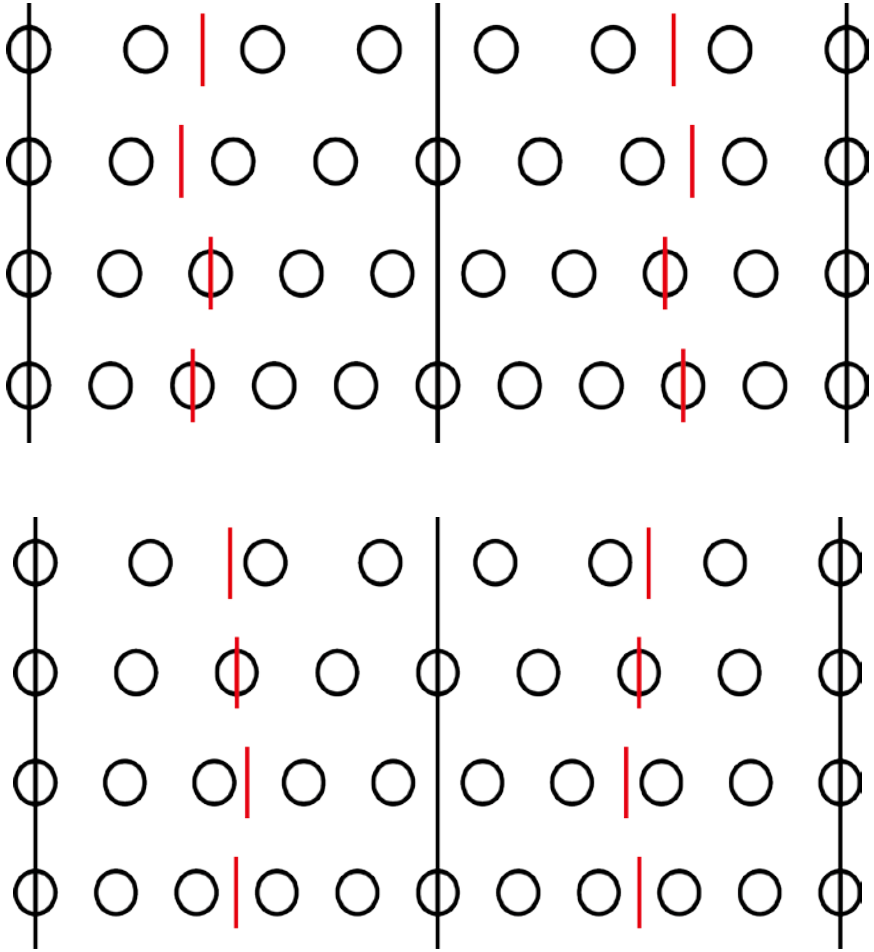
パーセント点の計算と合わせている

n 個のデータ（観測値）$x_i (i = 1, 2, \ldots, n)$ を大きさの順に並べ、小さいほうから j 番目のデータを $x_{(j)}$ で表す。

$$\text{PERCENTILE(データの範囲,}p) = x_{(i)}, \quad i = (n-1)p + 1$$

但し、i が整数にならないときは、i の整数部分を j、小数部分を k として、$x_{(i)} = (1-k) \times x_{(j)} + k \times x_{(j+1)}$

EXCELと数学Iでの４分位の違い



- 上：数学I
- 下：EXCEL

パーセント点（パーセンタイル）

- データを小さい順に並べた時、その値よりも小さな値の割合が指定された割合になる値
  - 10%点：その値より小さいデータが全体の10%になる値
  - 50%点＝中央値
  - 0%点＝最小値、100%点=最大値
  - 上側4分位値＝75%点、下側4分位値=25%点

度数分布表とパーセント点

- 累積相対度数を用いると、4分位値がどこにあるかが分かる

階級	階級値	度数	相対度数	累積度数	累積相対度数
143～147	145	3	0.0375	3	0.0375
148～152	150	10	0.125	13	0.1625
153～157	155	27	0.3375	40	0.5
158～162	160	28	0.35	68	0.85
163～167	165	9	0.1125	77	0.9625
168～172	170	3	0.0375	80	1

最小値

下側4分位値

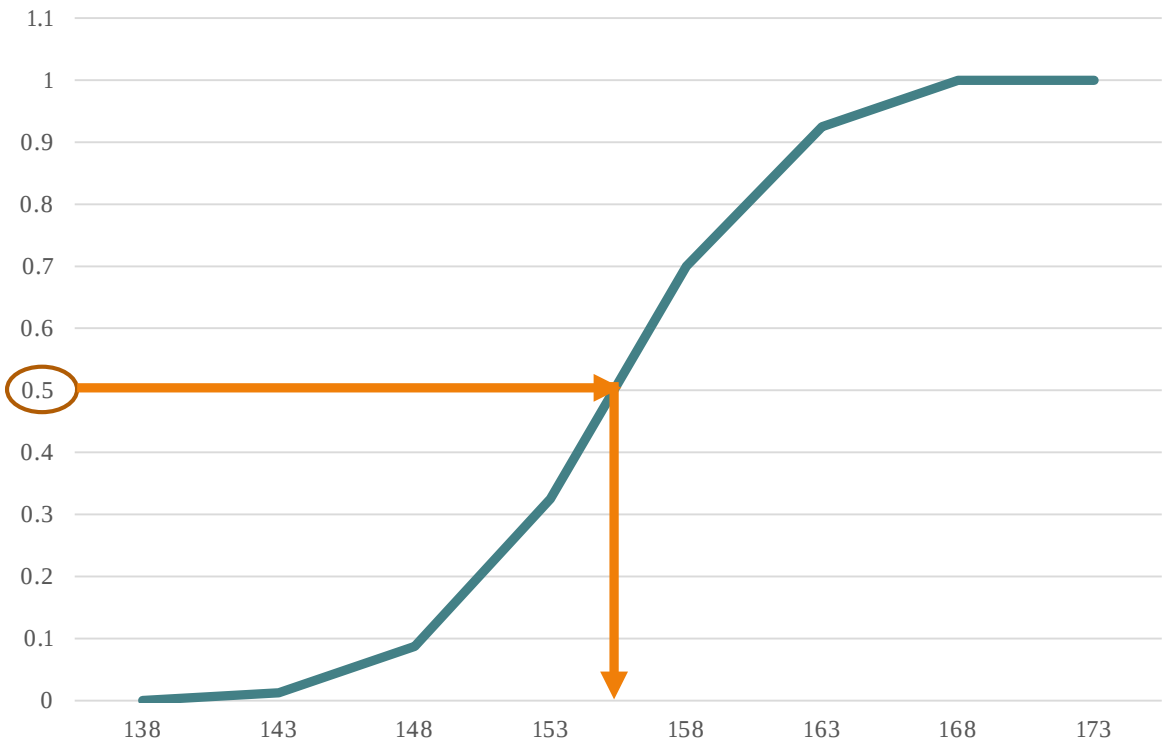
中央値

上側4分位値

最大値

累積相対度数で度数折れ線

- 累積相対度数で度数折れ線を描く→四分位数、パーセント点がどの辺りにあるか分かる

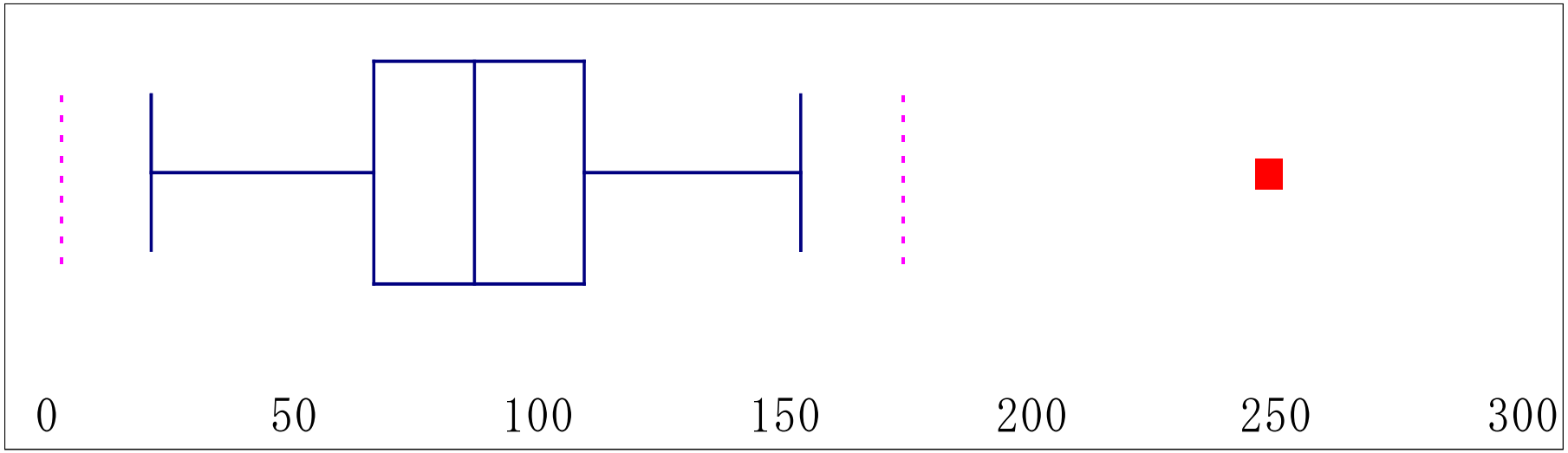


累積
相対度数
↓
箱ひげ図

計算例

- データ：　2,3,5,7,11,13,17,19,23
  - 最小値　2
  - 最大値　23
  - 中央値　11
  - 第1（下側）4分位値　5（8×0.25＋1＝3番目）
    - 高校の教科書通り計算すると、4
  - 第3（上側）4分位値　17（8×0.75＋1＝7番目）
    - 高校の教科書通り計算すると、18
  - レンジ　21　（23-2）
  - 四分位範囲　12　（17-5）

箱ひげ図

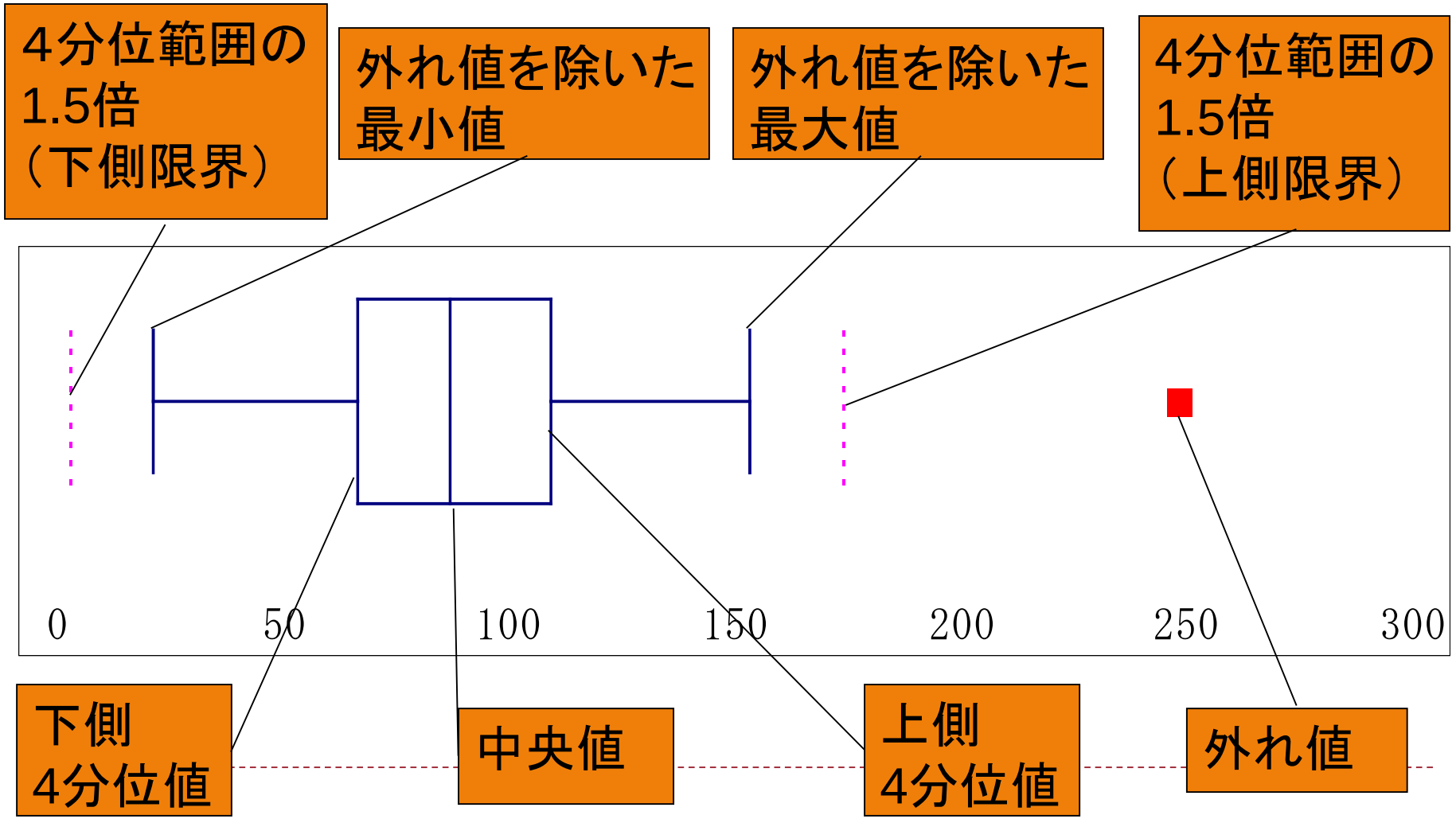


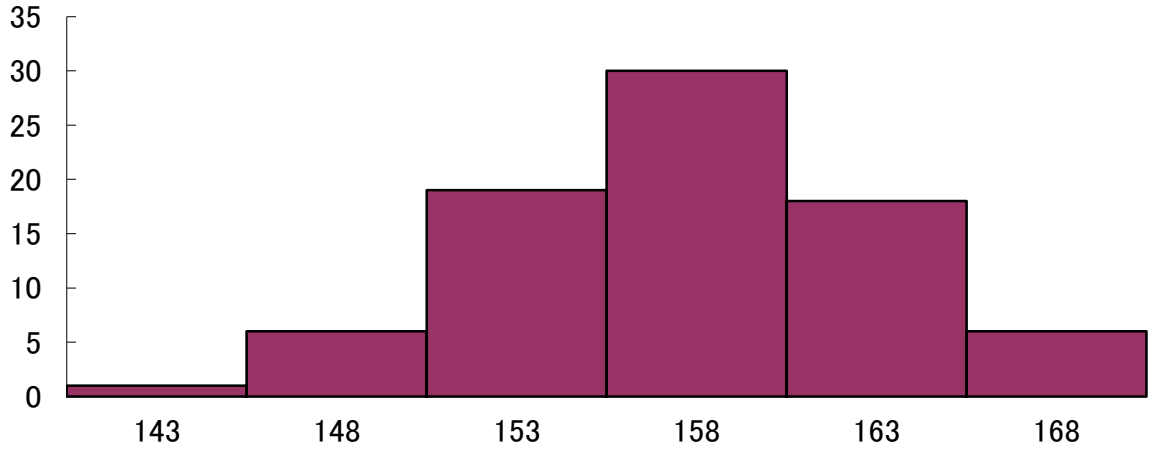
箱ひげ図（Box Wisker Diagram）：順序統計量を用いてデータのばらつきや偏りを表した図

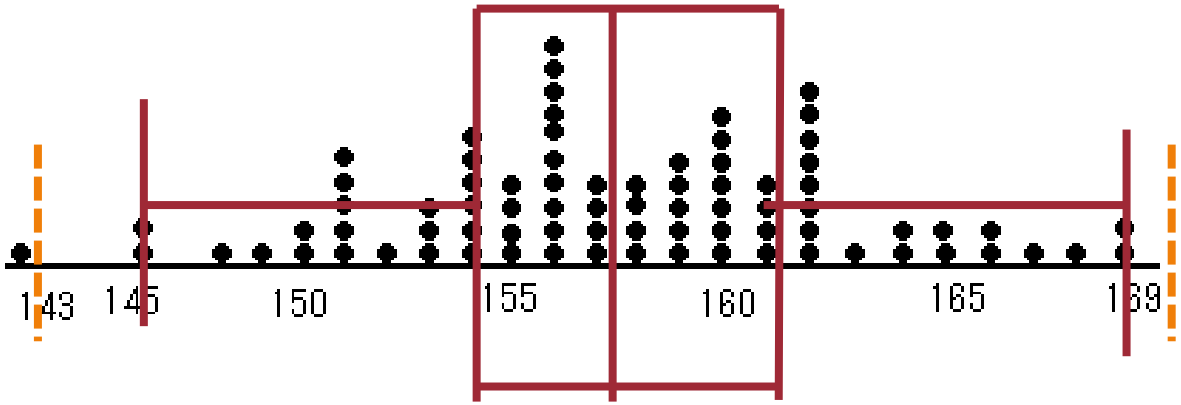
箱ひげ図と四分位

- 箱ひげ図：四分位数を利用して描く
- 色々な描き方がある
  - 四分位の計算方法による違い
  - ひげの伸ばし方
    - 最大値、最小値まで（外れ値を考えない）
    - 四分位範囲の1.5倍を利用（外れ値を考える）
    - パーセント点（データの10・90%、5・95%）

箱ひげ図が表すもの


4分位範囲の
1.5倍
（下側限界）
外れ値を除いた
最小値
外れ値を除いた
最大値
4分位範囲の
1.5倍
（上側限界）
0
50
100
150
200
250
300
下側
4分位値
中央値
上側
4分位値
外れ値





151	154	158	162
154	152	151	167
160	161	155	159
160	160	155	153
163	160	165	146
156	153	165	156
158	155	154	160
156	163	148	151
154	160	169	151
160	159	158	157
154	164	146	151
162	158	166	156
156	150	161	166
162	155	143	159
157	157	156	157
162	161	156	156
162	168	149	159
169	162	162	156
150	153	159	156
162	154	164	161

箱ひげ図を使うと

- データの散らばり具合、集中具合が一目で分かる
- 中央値の位置により、データが対称かそうでないかを知ることができる
- 四分位値と四分位範囲を使った基準により、外れ値を見つけ、表示することができる
- 三つ以上のデータ集合を比較するのに便利
- 2つ以上の項目で層別されたデータを表すのに便利

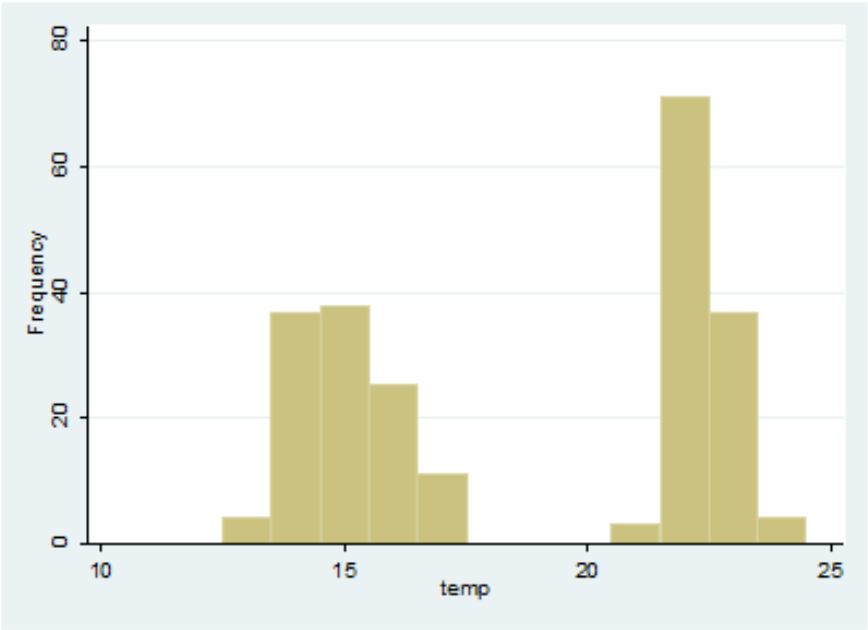
箱ひげ図の見方

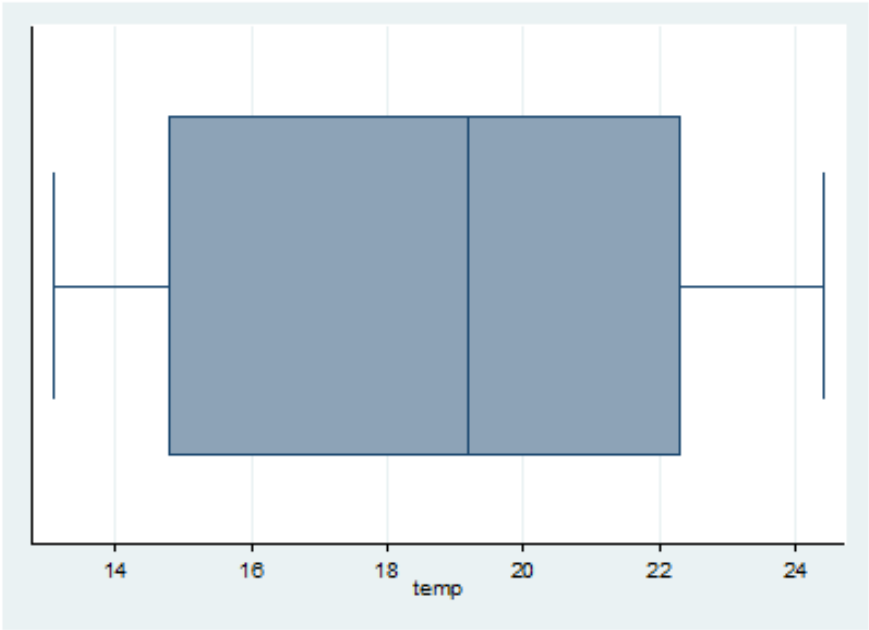
- 箱
  - 大きさ
  - 真ん中の線（中央値）の位置
- ひげ
  - 長さ
  - 対称性
- 全体的に
  - 大きさ
  - 外れ値
  - 対称性

箱ひげ図の強み、ヒストグラムの強み

- 箱ひげ図の強み
  - 順序統計量だけで書ける
    - ヒストグラム：度数分布表（データの集計）が必要
  - 複数のデータ・層別データを比較するときに便利
- ヒストグラムの強み
  - 複数のピーク（山・峰）を持つデータ分布を表現できる
    - 箱ひげ図：箱がひげより長く、いびつなグラフになる
  - データの分布状態（山・最頻値）がより良く分かる

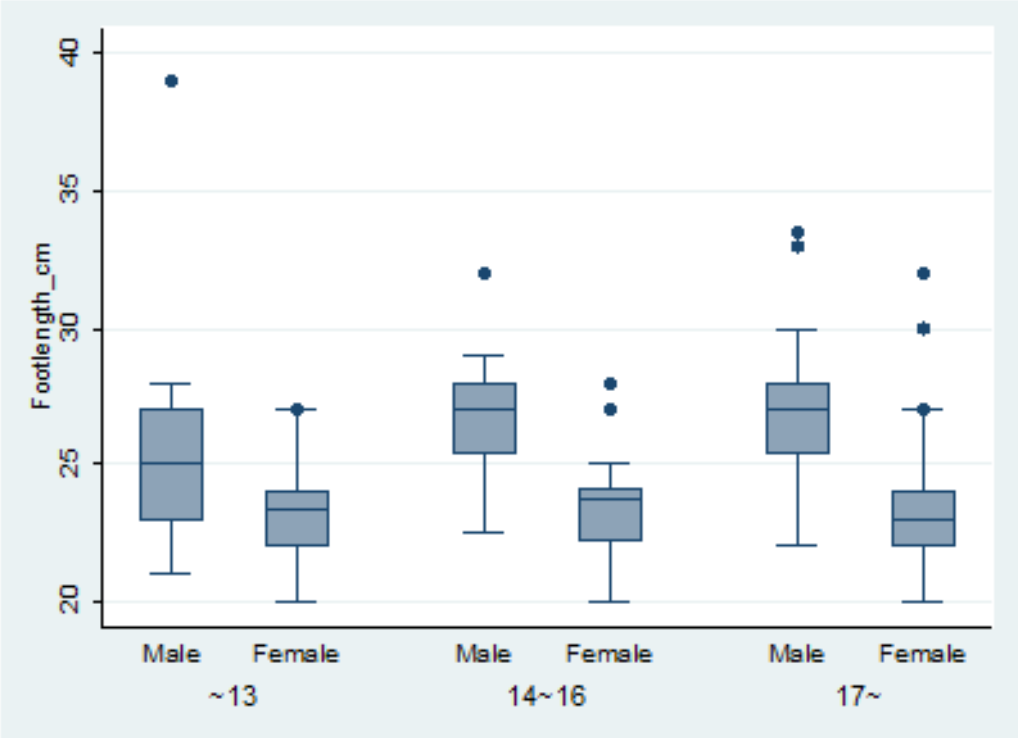
２つの山があるデータ





箱ひげ図とヒストグラム

- 足の大きさを、性別・年齢別で比較
- ヒストグラムを描くと結構面倒
  - ６種類描く
  - 色分けをする



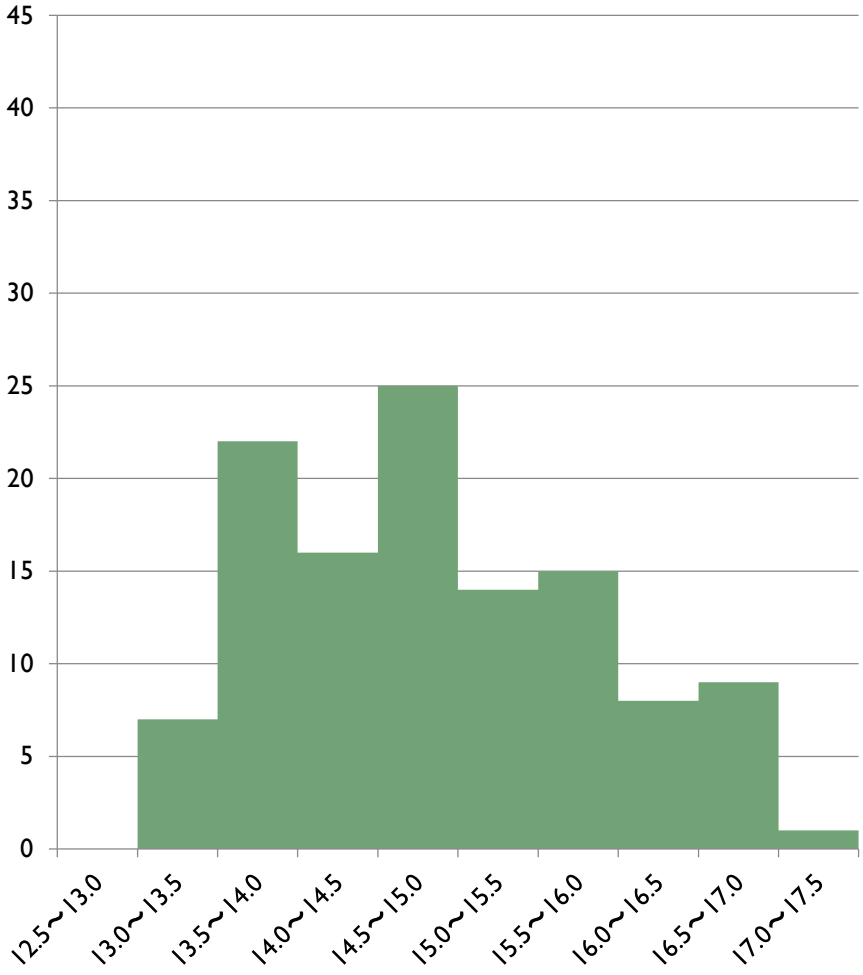
分散と標準偏差

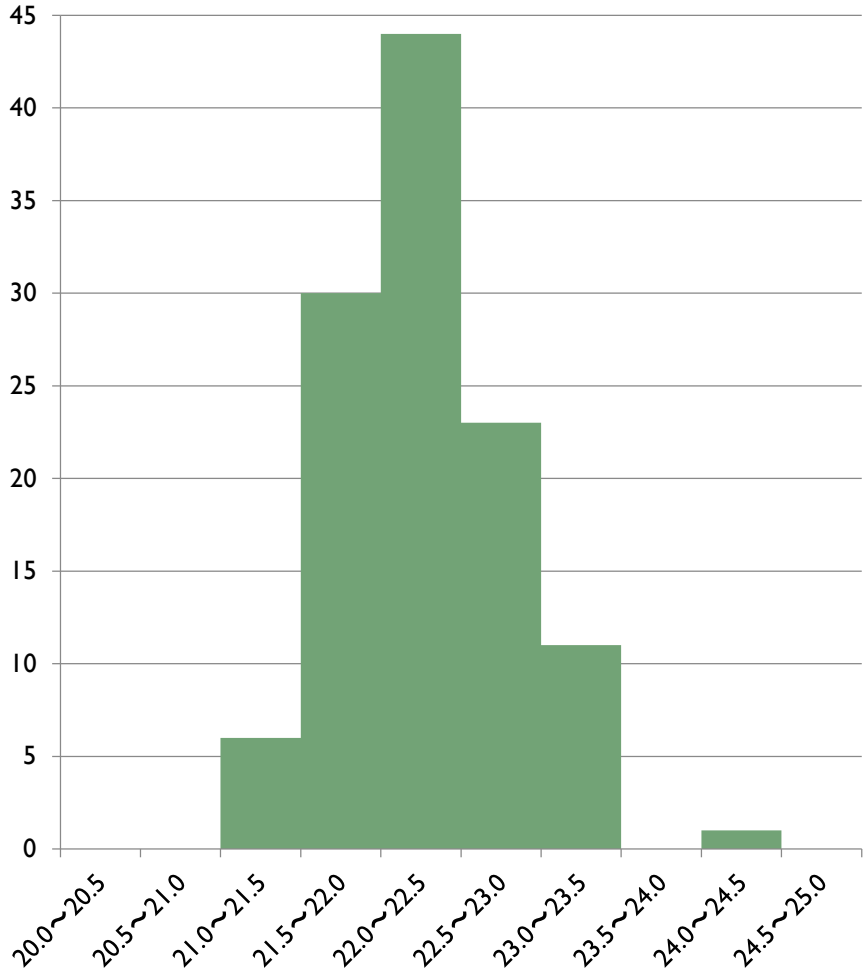
	東京	那覇
1891	14.4	21.8
1892	14	21.9
1893	13.8	21.6
1894	14.8	22.1
1895	13.8	21.7
1896	14	22.3
1897	13.2	22.5
1898	14	22.6
1899	13.8	21.9
1900	13.6	22.3
1901	13.9	21.6
⋮	⋮	⋮

東京と那覇の気温

	東京	那覇
平均 $\bar{x}$	14.91368	22.33304
中央値	14.8	22.3

データ例（ヒストグラム）


45
40
35
30
25
20
15
10
5
0
12.5~13.0
13.0~13.5
13.5~14.0
14.0~14.5
14.5~15.0
15.0~15.5
15.5~16.0
16.0~16.5
16.5~17.0
17.0~17.5


45
40
35
30
25
20
15
10
5
0
20.0~20.5
20.5~21.0
21.0~21.5
21.5~22.0
22.0~22.5
22.5~23.0
23.0~23.5
23.5~24.0
24.0~24.5
24.5~25.0

世の中には

- ヒストグラムを作ると、形が違うデータセットが多々あります
  - 細長かったりぺたんこだったり
  - 「データのバラツキ具合（散布度）」が違う
- そのような『データの違い』を表す統計量はあるだろうか？
  - 一々ヒストグラムで表していたら、スペースがもったいない
  - 平均からの距離（偏差）であらわす
    - プラスマイナスの効果を打ち消すことを忘れずに

偏差平方和（変動）

- データから、データの平均を引いて2乗したものの総和（2乗平均）
- 「平均からのズレ」の総和を表す
  - データからデータの平均を引いたものの総和は0になってしまうので、「平均からのズレ」の指標に適さない
  - 偏差平方和は、データの大きさが大きいほど、大きくなってしまう

$$S = (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

分散

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

x_i：データ（$i = 1, \ldots, n$）、$\bar{x}$：x_i の平均

標準偏差

分散は、データから平均を引いて**2乗**しているから、
測定単位も2乗になる。

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

分散の正の平方根

測定単位が平均と同じになる

度数分布表から分散を計算する

- 度数分布表から分散を計算するには
  - 平均値を計算する
  - 階級値から平均値を引いて2乗し、度数をかける
  - それを合計する
  - データの総数で割る

計算例 1

到着時間	32	27	29	34	33

到着時間	32	27	29	34	33
到着時間の偏差	1	-4	-2	3	2
偏差の２乗	1	16	4	9	4

- 平均：31
- 偏差平方和：34
- 分散：6.8
- 標準偏差：2.61

A：階級値	B：度数	A×B	C：階級値－平均値	C＾2	C＾2×B
1	3	3	-1	1	3
2	5	10	0	0	0
3	1	3	1	1	1
4	1	4	2	4	4
	10	20			8

平均：20÷10=2

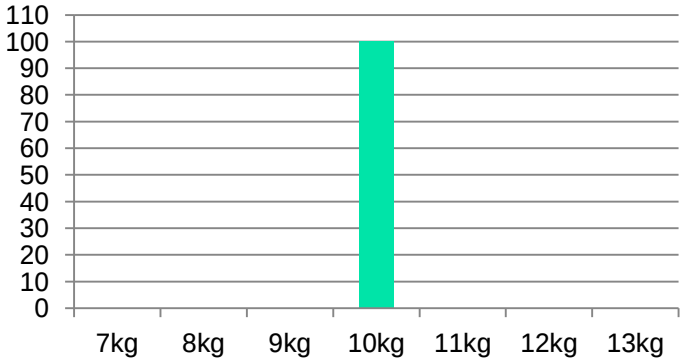
分散：8÷10=0.8

標準偏差：√0.8＝0.894

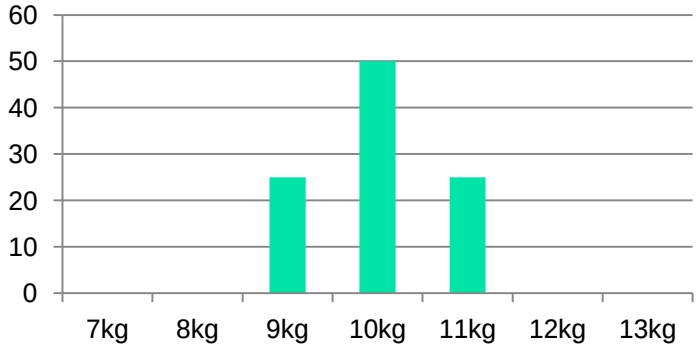
標準偏差の意味

- データの、平均からの散らばりを評価する
  - データの、平均値からの離れ方を平均したもの
  - 大きいほうに離れようが、小さいほうに離れようが、どちらも正の数として評価し、打ち消しあわないように平均している

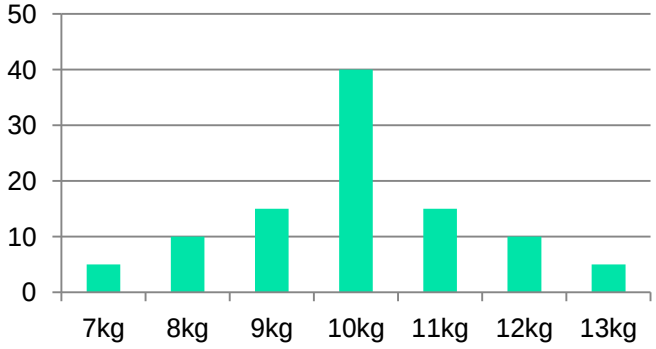
ヒストグラムと分散・標準偏差



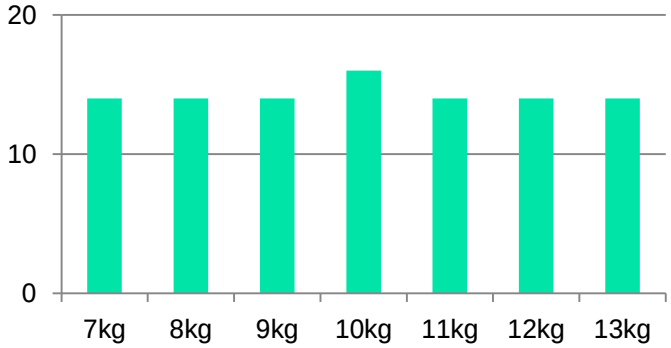
分散0、標準偏差0



分散0.5、標準偏差0.707



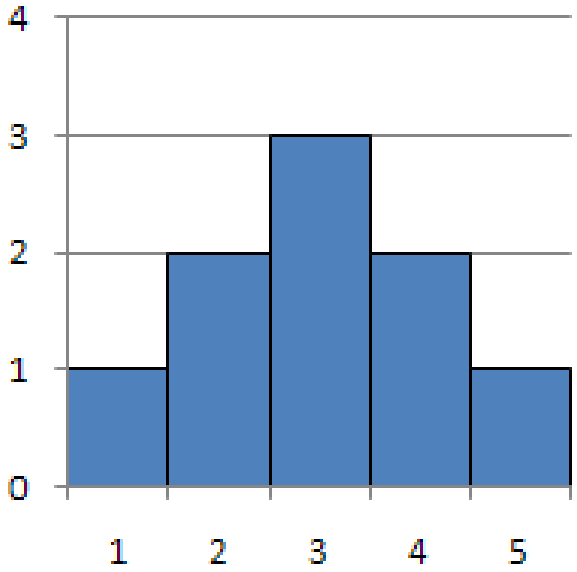
分散2、標準偏差1.414

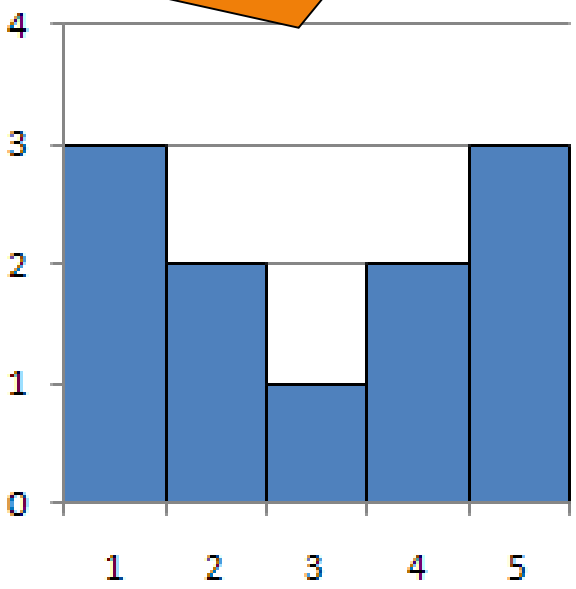


分散3.92、標準偏差1.980

標準偏差　どっちが大きい？

標準偏差が大きい

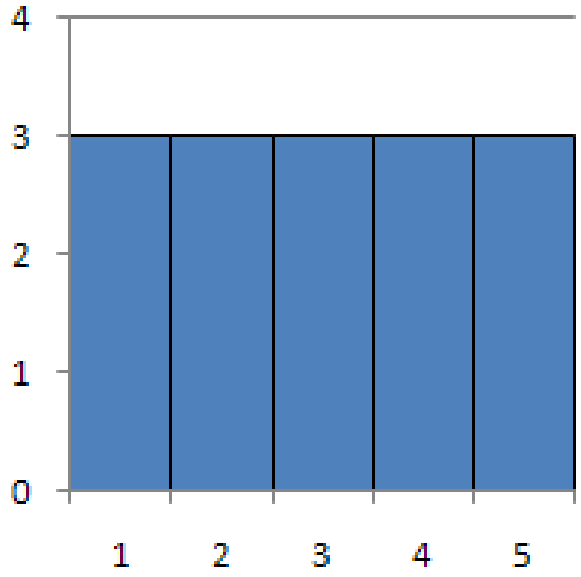


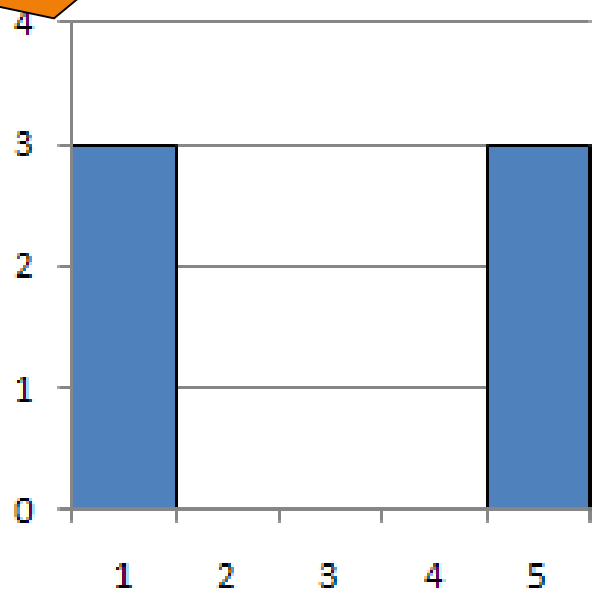


- AもBもデータの分布範囲が一緒
- 左右対称だから平均も一緒
- Bは、-2や+2の偏差を持つデータが多い

標準偏差　どっちが大きい？

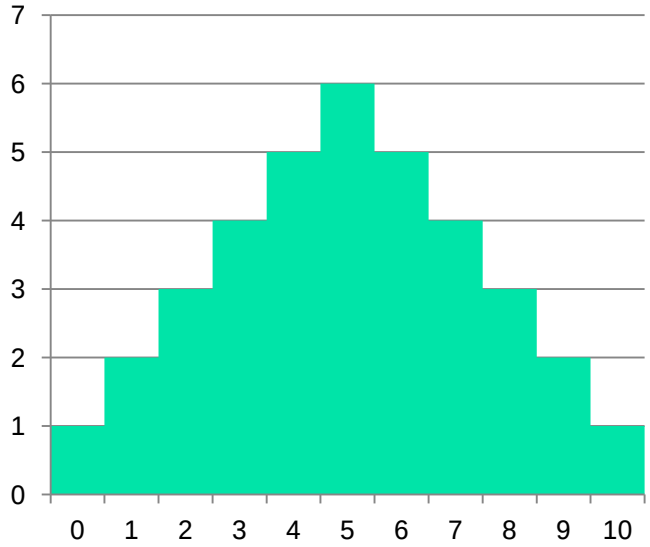
標準偏差が大きい

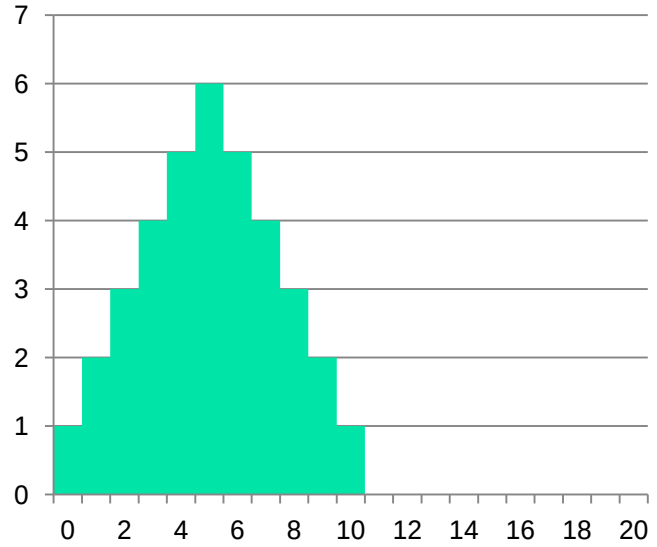




- Bの偏差は、-2または+2
- Aでは、-2または+2の偏差を持つデータは全体の2/5
- 分散は、偏差の2乗の平均

標準偏差　どっちが大きい？



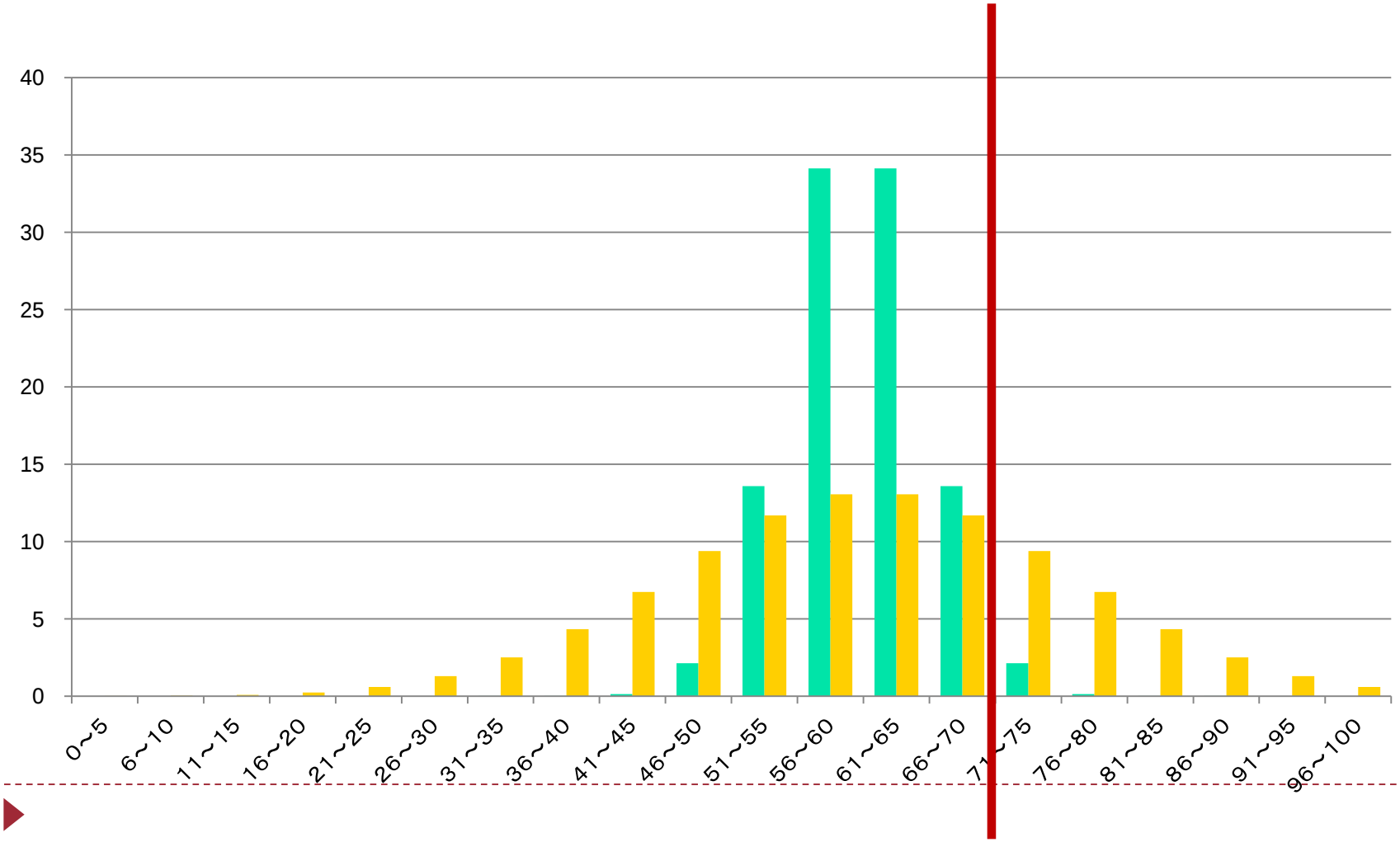


- どちらも同じ
- 同じデータについて、横軸の長さを変えたもの

2標準偏差ルール

- 正規分布に従うデータでは、平均から±2標準偏差の範囲内に収まるデータが95%
- 例：100点満点のテスト
  - 平均は70、標準偏差が5
    - このテストを受けた人の95%は60～80点を取った
    - 80点以上取った人は全体の2.5%しかいない
    - 60点も取れなかった人は全体の2.5%
  - 平均は70、標準偏差が10
    - このテストを受けた人の95%は50～90点を取った
    - 90点以上取った人は全体の2.5%
    - 50点も取れなかった人も2.5%いる

グラフにすると


40
35
30
25
20
15
10
5
0
0~5
6~10
11~15
16~20
21~25
26~30
31~35
36~40
41~45
46~50
51~55
56~60
61~65
66~70
71~75
76~80
81~85
86~90
91~95
96~100

偏差値の話

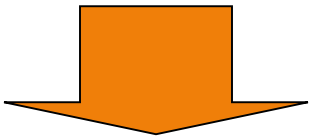
- 偏差値：データを平均が50、標準偏差が10になるように加工したもの

$$\text{偏差値} = \frac{\text{得点}\,x - \text{平均}\,\mu}{\text{標準偏差}\,\sigma} \times 10 + 50$$

標準化

標準化

- データを、（データー平均値）÷標準偏差で加工すると、できたデータの平均と標準偏差は以下のようになる
  - 平均＝0
  - 標準偏差＝1



標準化