

散布度

統計基礎の補足資料

2019年6月17日
金沢学院大学経営情報学部
藤本祥二

基本統計量

- 基本統計量: 分布の特徴を表す数値

- 代表値 (分布の中心を表す数値)

- 平均値 (mean, average)
- 中央値 (median)
- 最頻値 (mode)

} 前回やった

- 散布度 (分布のばらつき具合を表す数値)

- 分散 (variance)
- 標準偏差 (standard deviation)
- 範囲 (レンジ, range)
- 四分位範囲 (inter quartile range, IQR)

} 後期にやる (検定3級)

} 今回の話

- その他の分布の形を表す数値

- 歪度 (ひずみ度, わい度, skewness)
- 尖度 (とがり度, せん度, kurtosis)

} 統計検定2級以降

散布度

- 平均とデータの差である偏差 ($d_i = x_i - \bar{x}$) から計算した散布度
(平均値を拡張した考え方)
 - 分散(偏差平方の平均)
 - 標準偏差(分散の平方根)
- データの個数の割合から計算した散布度
(中央値を拡張した考え方)
 - 範囲(レンジ)
 - 四分位範囲

後期にやる(検定3級)

今回の話

範囲(レンジ)(教科書P.110)

- 範囲(range)
数量データの範囲を表す基本統計量

$$\text{範囲} = \text{最大値} - \text{最小値}$$

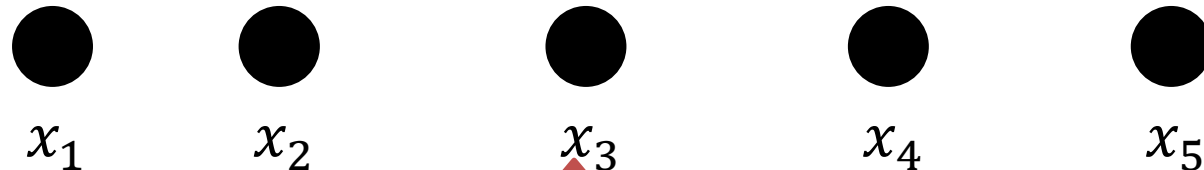
で計算

- 教科書の例1 (P.110)
320, 395, 470, 480, 515, 530, 550, 2000 単位:万円
範囲 = $2000 - 320 = 1680$ 万円
- 教科書の例2 (P.110)
550, 580, 601, 645, 650, 705, 760, 770, 単位:万円
範囲 = $770 - 550 = 220$ 万円

中央値の復習(教科書P.101)

- 中央値
別名: 中位数, 第1二分位数, 第2四分位数,
50パーセント点(50パーセンタイル)
- 要はデータを半分ずつ二分する数値

$n = 5$ の時



中央値 $\tilde{x} = x_3$

$n = 6$ の時

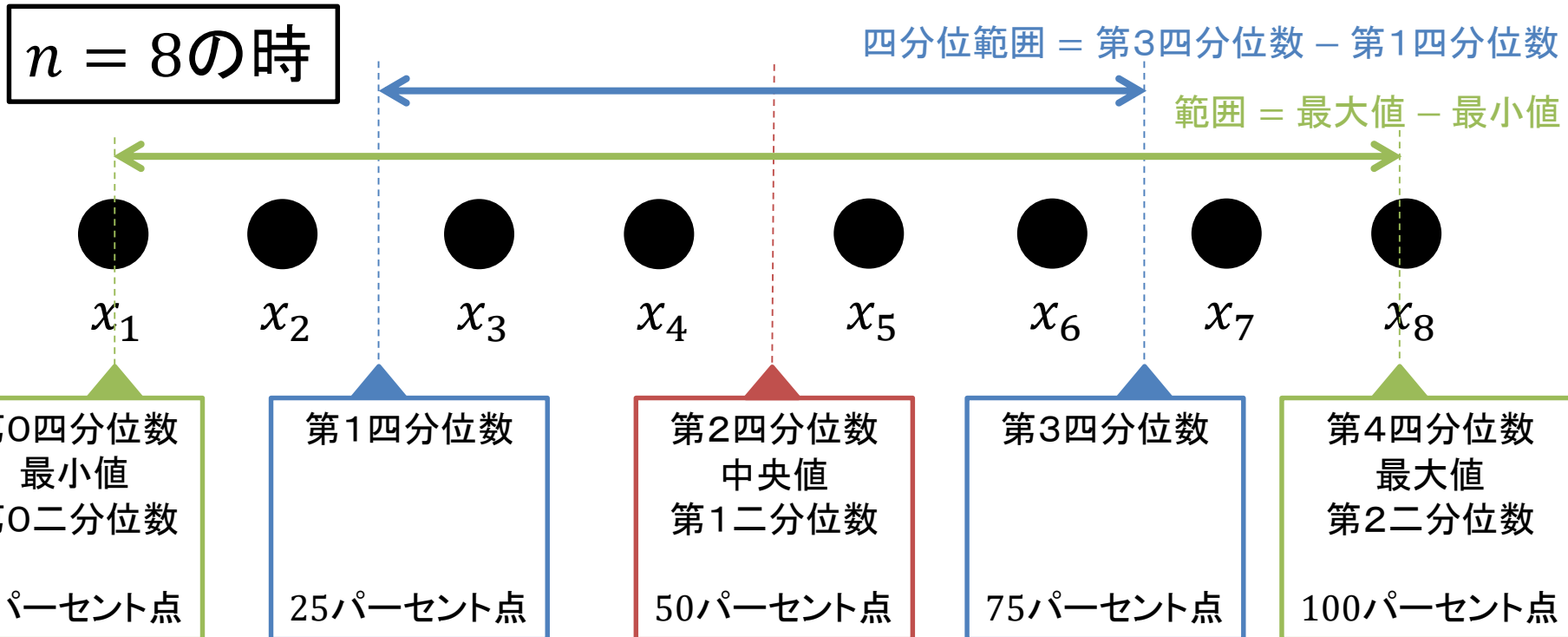


中央値 $\tilde{x} = (x_3 + x_4)/2$

四分位数と四分位範囲

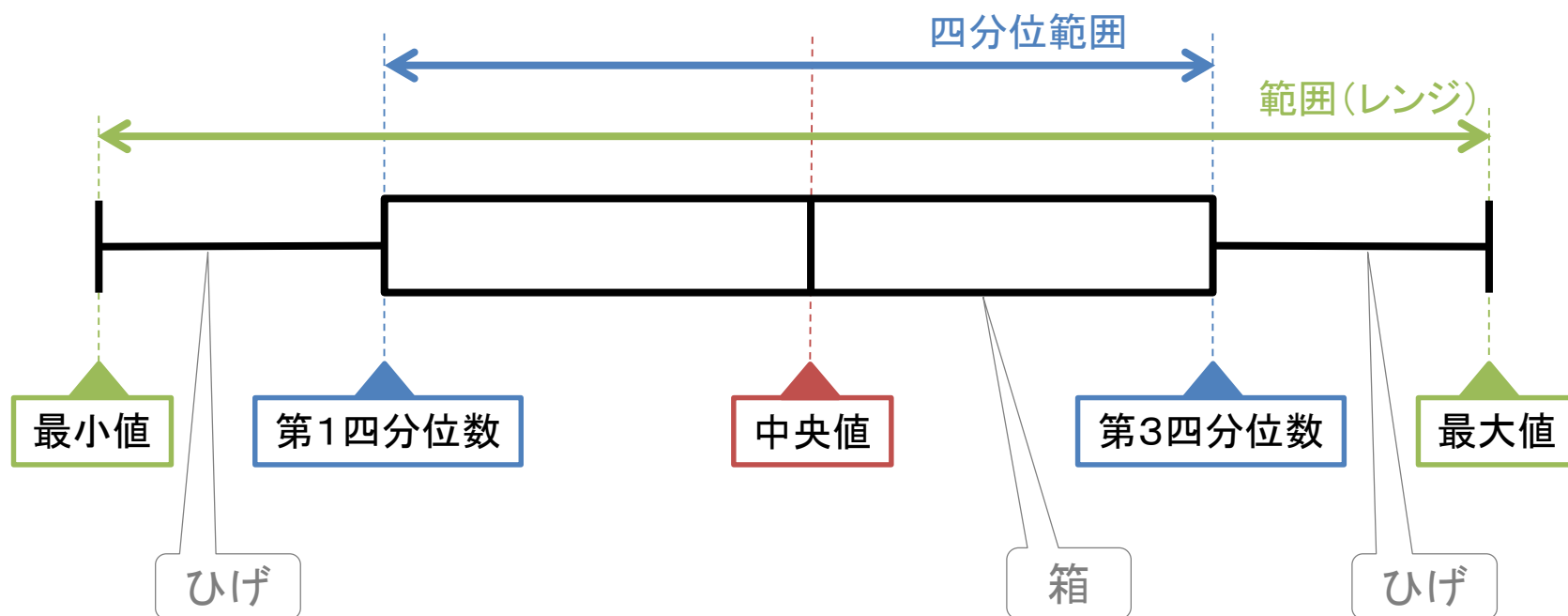
(教科書P.111)

- データを四分割する位置の数値が四分位数
- 第3と第1四分位数の間の範囲が四分位範囲
- 四分位範囲内に50%のデータが入る



5数要約と箱ひげ図(教科書P.116)

- 「最小値」「第1四分位数」「中央値」「第3四分位数」「最大値」の5つの数で分布を表現することを「5数要約」という
- 5数要約を箱ひげ図で表現する



5数の名称と位置

			相対順位的位置		
5数の名称	別名称	別名称	分数	小数	百分率
第0四分位数	第0二分位数	最小値 最小位数	$\frac{0}{4}$	0.00	0%
第1四分位数			$\frac{1}{4}$	0.25	25%
第2四分位数	第1二分位数	中央値 中位数	$\frac{2}{4}$	0.50	50%
第3四分位数			$\frac{3}{4}$	0.75	75%
第4四分位数	第2二分位数	最大値 最大位数	$\frac{4}{4}$	1.00	100%

範囲と四分位範囲

- 範囲(range)

$$\text{範囲} = \overset{100\% \text{点}}{\text{最大値}} - \overset{0\% \text{点}}{\text{最小値}}$$

範囲内には100%の個数(割合)のデータが含まれる

- 四分位範囲(inter quartile range, IQR)

$$\text{四分位範囲} = \overset{75\% \text{点}}{\text{第3四分位数}} - \overset{25\% \text{点}}{\text{第1四分位数}}$$

四分位範囲内には50%の個数(割合)のデータが含まれる

四分位数の求め方

- パーセント点 (percentile) を正確に求める方法を使う
25%の順位に対応する数値(25%点)と
75%の順位に対応する数値(75%点)を求める。
(相対順位という考え方を使って、あらゆるパーセント点を正確に求めることができる。)
- 簡便法を使う(大体の値のこと 概算値を知るにはこちらで十分)
 - 中央値を境に下側と上側のデータセットに分ける。
(データ数が奇数の時は中央値のデータを上下のデータセットに含めるかどうかで2種類の簡便法があり、微妙に結果は異なる。)
 - 下側のデータセットの中央値が第1四分位数
 - 上側のデータセットの中央値が第3四分位数

データ数が多いときは正確なやり方と、簡便法のやり方の結果は殆ど同じ値になる。
やり方を明示すれば、どちらの方法を使ってもよい。

パーセント点を求めて 正確に四分位数を求める方法

パーセント点の求め方

- $\{x_1, x_2, \dots, x_n\}$ の p パーセント点を求める。
- 順位 (rank)

相対順位 $\frac{p}{100}$ に対応する順位が R

$$R = 1 + (n - 1) \frac{p}{100}$$

の、整数部分を r 、小数部分を s とする。

次スライドの例の手順3

- データを小さい順に並べた時の r 番目と $r + 1$ 番目のデータをそれぞれ x_r, x_{r+1} とする。
- $p\%$ のパーセント点は次の式で求められる。

$$x_r + s(x_{r+1} - x_r)$$

次スライドの例の手順4

$$= (1 - s)x_r + sx_{r+1}$$

r 位のデータ x_r と
 $r + 1$ 位のデータ x_{r+1} の
間を $s: (1 - s)$ に内分する点

50%点は中央値と全く同じ式になるので、普通に中央値を求めればよい。

$p = 50$ の時は順位は $R = \frac{n+1}{2}$ になるので、

- n が偶数の時 $r = \frac{n}{2}, s = 0.5$ なので $(x_{n/2} + x_{n/2+1})/2$ が 50% 点
- n が奇数の時 $r = \frac{n+1}{2}, s = 0$ なので $x_{n/2+1}$ が 50% 点

25%点を求める例

{3,5,8,10,12,12,15,17,19,20}の10個のデータの例

25%点は8.5

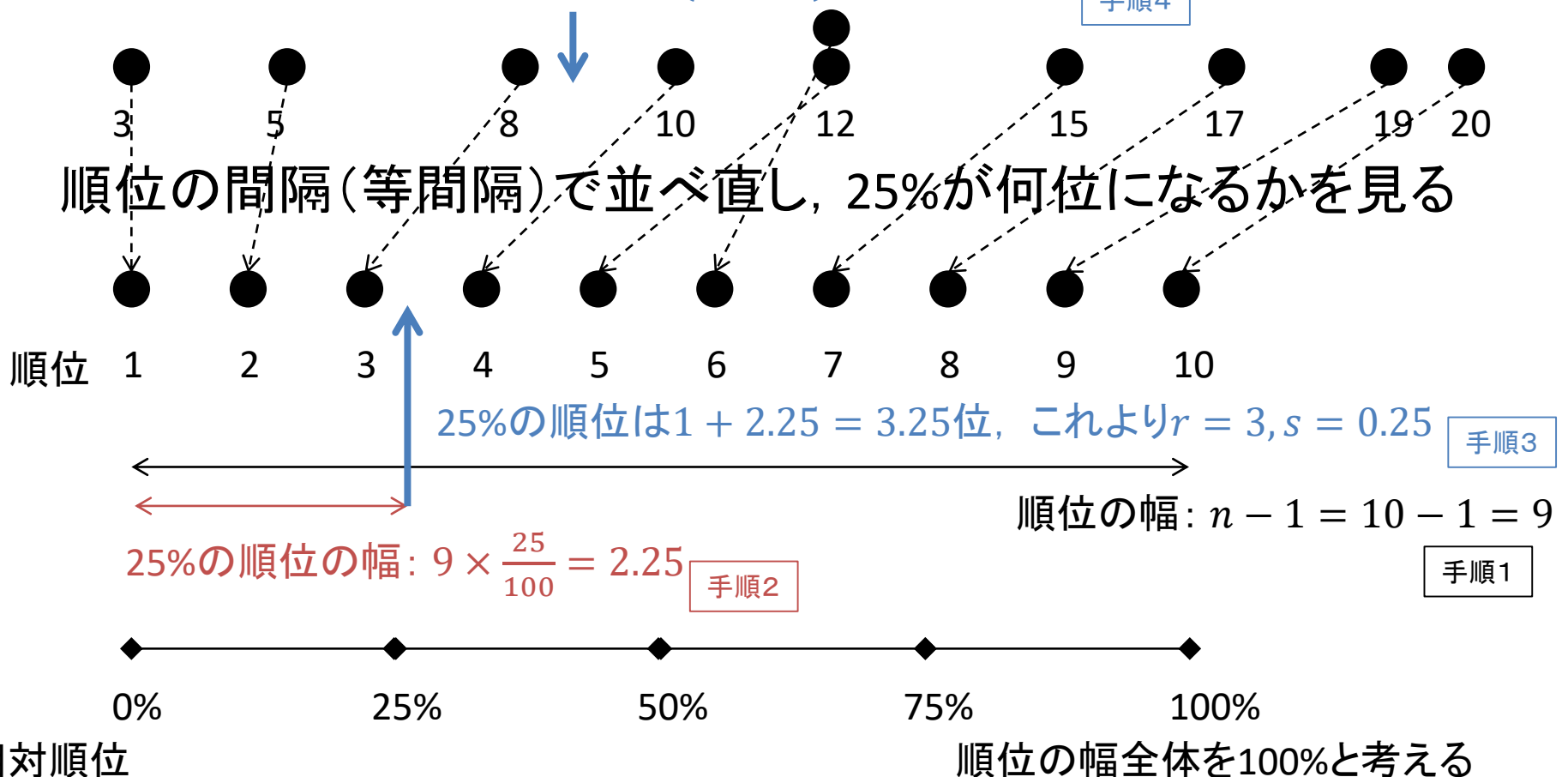
$$x_r + (x_{r+1} - x_r) \times s$$

3位と4位の間を0.25で内分する点

$$25\% \text{点}: 8 + (10 - 8) \times 0.25 = 8.5$$

手順4

順位の間隔(等間隔)で並べ直し, 25%が何位になるかを見る



75%点を求める例

{3,5,8,10,12,12,15,17,19,20}の10個のデータの例

75%点は16.5

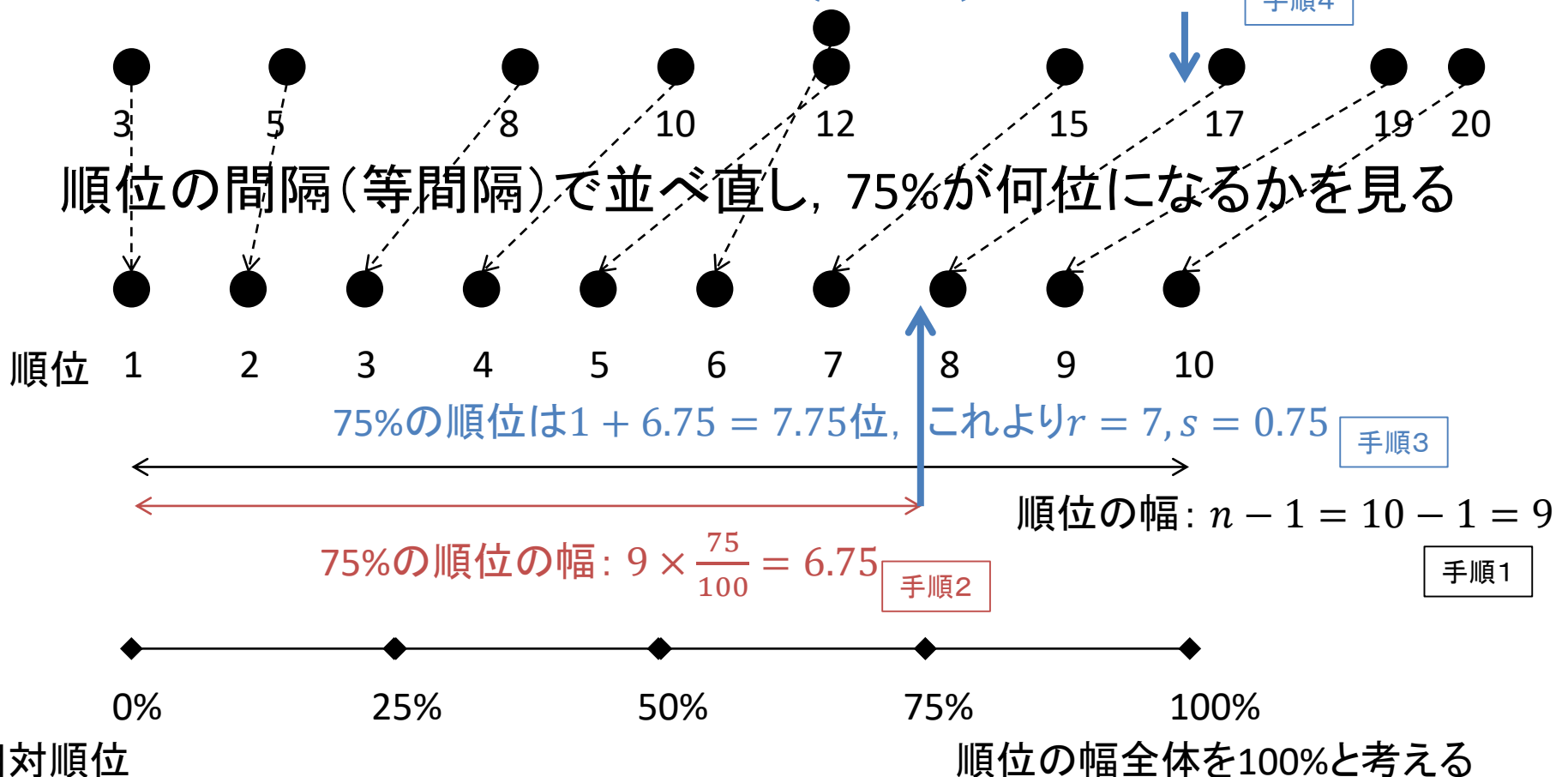
$$x_r + (x_{r+1} - x_r) \times s$$

7位と8位の間を0.75で内分する点

$$75\% \text{点}: 15 + (17 - 15) \times 0.75 = 16.5$$

手順4

順位の間隔(等間隔)で並べ直し, 75%が何位になるかを見る

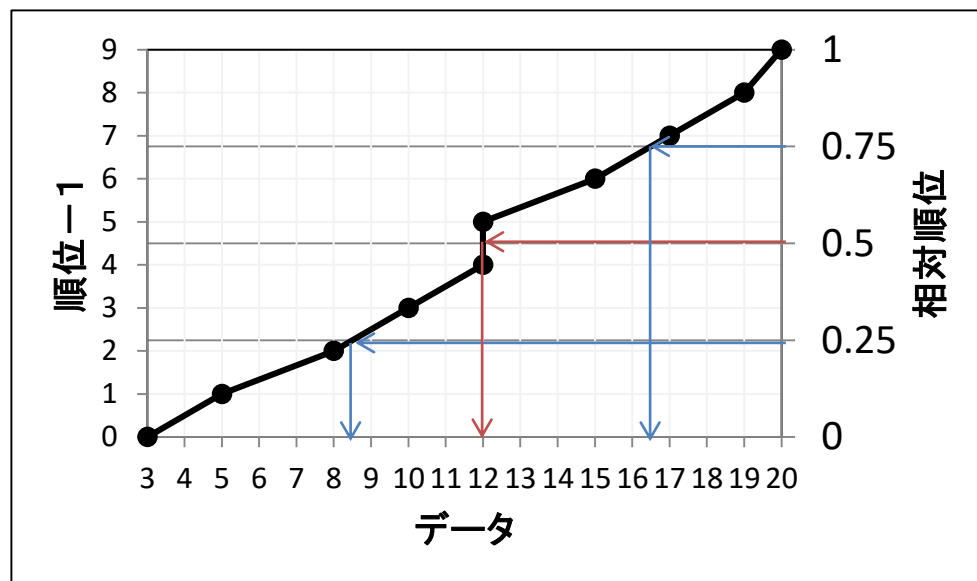


パーセント点を相対順位のグラフを描いて求める方法

データ	累積度数 のようなもの	累積相対度数 のようなもの
	順位-1	相対順位
3	0	$0/9 = 0$ 0%
5	1	$1/9 = 0.11111...$
8	2	$2/9 = 0.22222...$
10	3	$3/9 = 0.33333...$
12	4	$4/9 = 0.44444...$
12	5	$5/9 = 0.55555...$
15	6	$6/9 = 0.66666...$
17	7	$7/9 = 0.77777...$
19	8	$8/9 = 0.88888...$
20	9	$9/9 = 1$ 100%

↑
横軸

↑
2つの縦軸を合わせて
綺麗なグラフを描く



0%点(最小値) : 3
 25%点(第1四分位数) : 8.5
 50%点(中央値) : 12
 75%点(第3四分位数) : 16.5
 100%点(最大値) : 20

範囲 : $20 - 3 = 17$
 四分位範囲 : $16.5 - 8.5 = 8$

Excel 関数を使う方法

- Ver. 2007 以前
 - 関数 QUARTILE
四分位数を求める関数
 - 関数 PERCENTILE
パーセンタイルを求める関数
- Ver. 2010 以降
 - 関数 QUARTILE.INC 関数 PERCENTILE.INC
Ver. 2007 の QUARTILE PERCENTILE と同じ
本スライドと同じやり方
 - 関数 QUARTILE.EXC 関数 PERCENTILE.EXC
相対順位の設定が少し違うやり方

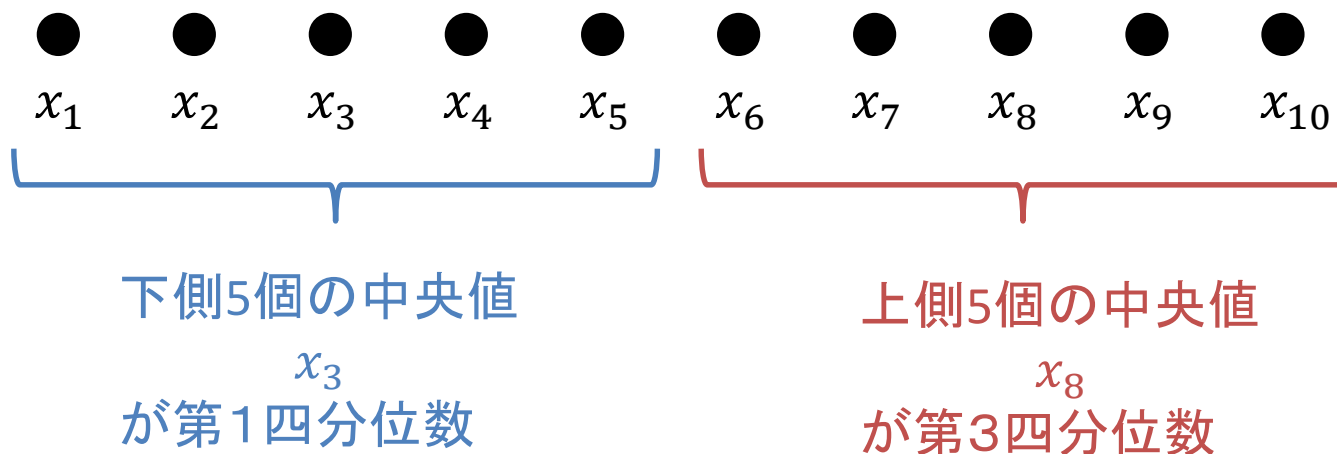
詳しくは[四分位数 パーセンタイル](#)で解説

四分位数を概算する簡便法

四分位数の概算値を求める簡便法の例 (偶数データ)

$n = 10$ (偶数)の時

中央値 $(x_5 + x_6)/2$ を境に下側と上側に分ける



{3,5,8,10,12,12,15,17,19,20}の10個のデータの例では

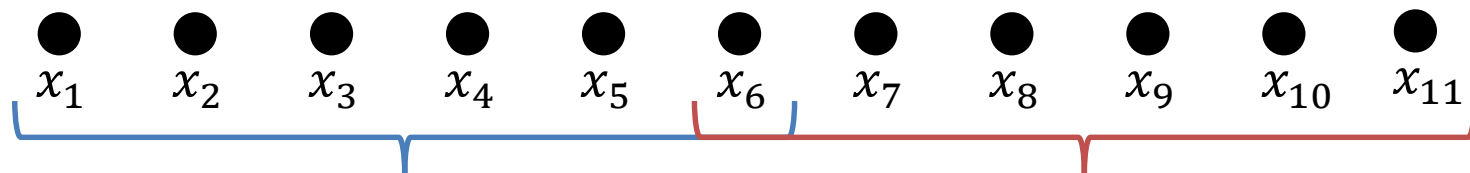
$x_3 = 8$ が第1四分位数, $x_8 = 17$ が第3四分位数

パーセント点を使って正確に求めた8.5, 16.5 と微妙に違うが大体合ってる

四分位数の概算値を求める簡便法の例 (奇数データ)

$n = 11$ (奇数)の時, やり方2種類あるがどちらを使っても良い

中央値 x_6 を境に 下側 と 上側 に分ける (中央値を含めて分けるやり方)



下側6個の中央値

$$(x_3 + x_4) / 2$$

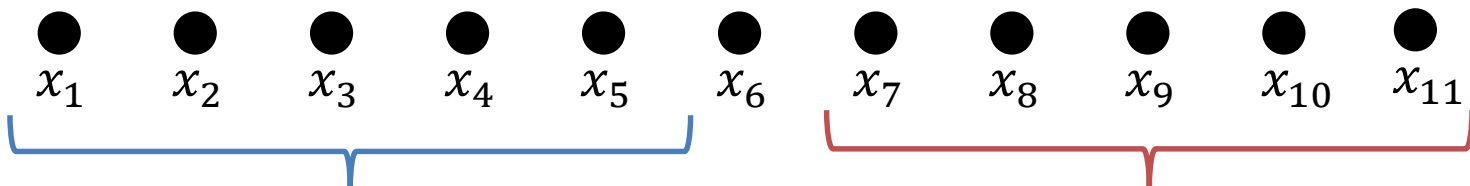
が第1四分位数

上側6個の中央値

$$(x_8 + x_9) / 2$$

が第3四分位数

中央値 x_6 を境に 下側 と 上側 に分ける (中央値を除いて分けるやり方)



下側5個の中央値

$$x_3$$

が第1四分位数

上側5個の中央値

$$x_9$$

が第3四分位数

2019年5月の金沢市 毎日の最高気温データの分析

中央値, 範囲, 四分位範囲
を5月の前後半で比較

2019年5月金沢市の最高気温データ

日にち	1	2	3	4	5
最高気温℃	16	20	22	23	25
2019年5月 金沢の最高 気温を小数 点第1位で 四捨五入し たデータ	6	7	8	9	10
	25	16	21	25	24
	11	12	13	14	15
	23	24	26	24	24
	16	17	18	19	20
	25	25	26	28	29
	21	22	23	24	25
	23	24	27	27	31
	26	27	28	29	30
	30	31	25	22	26
	31	データソース: 気象庁 各種データ・資料 > 過去の気象データ検索 > 日ごとの値 金沢 2019年5月(日ごとの値) 主な要素			
	25				

2019年5月前半の金沢の最高気温

前半1日～15日のデータを並べ替えて5要約数を求める

最小値16

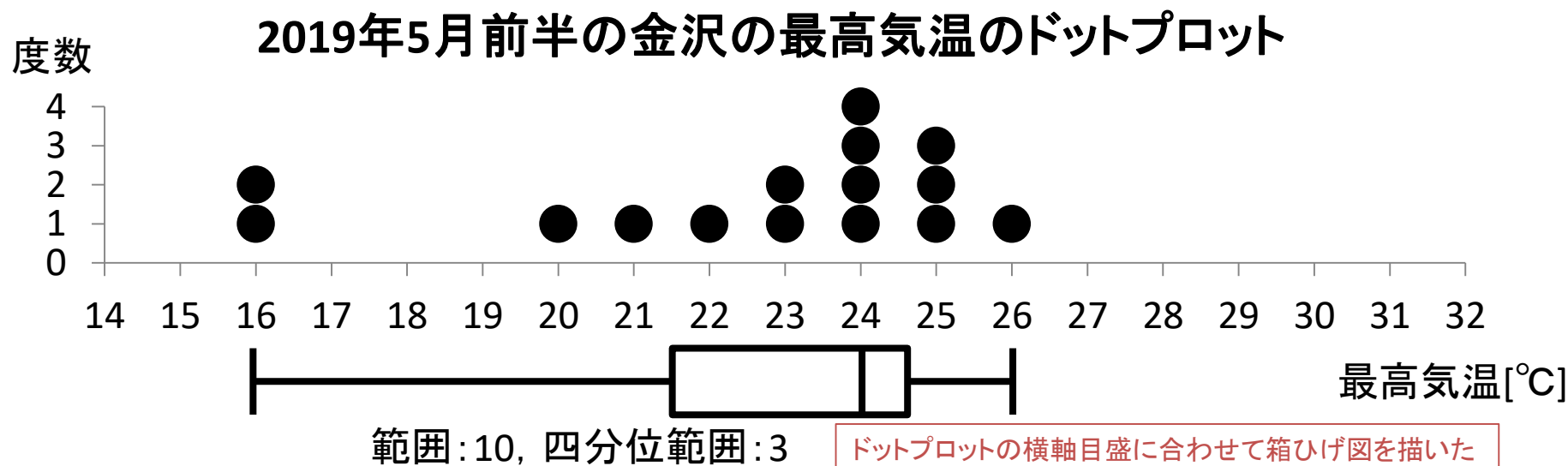
中央値 $x_8 = 24$

最大値26

16, 16, 20, 21, 22, 23, 23, 24, 24, 24, 24, 25, 25, 25, 26

第1 $(x_4 + x_5)/2 = 21.5$

第3 $(x_{11} + x_{12})/2 = 24.5$



2019年5月後半の金沢の最高気温

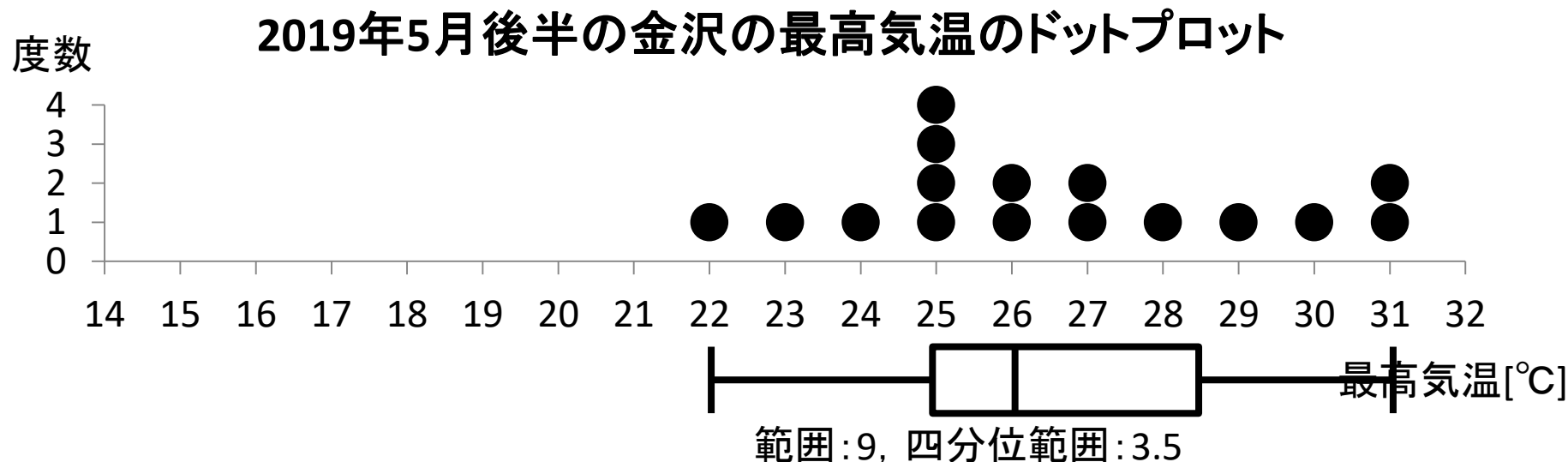
後半16日～31日のデータを並べ替えて5要約数を求める

最小値22 中央値 $(x_8 + x_9)/2 = 26$ 最大値31

22, 23, 24, 25, 25, 25, 25, 26, 26, 27, 27, 28, 29, 30, 31, 31

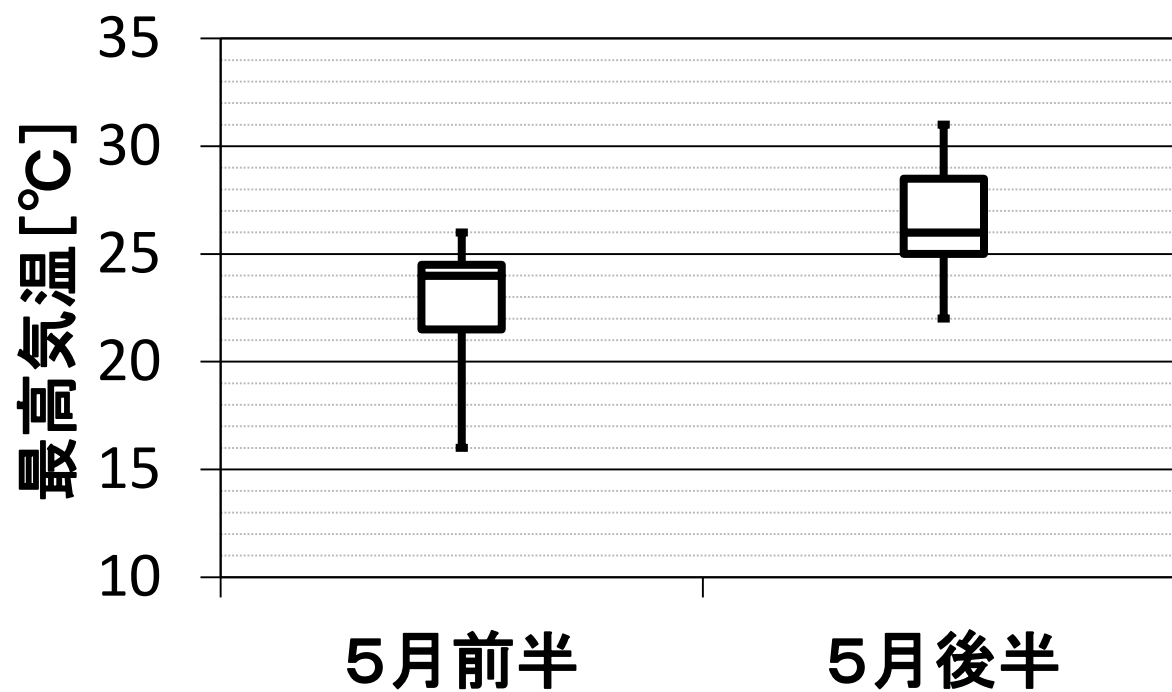
第1 $(x_4 + x_5)/2 = 25$

第3 $(x_{12} + x_{13})/2 = 28.5$



箱ひげ図で比較

金沢市の毎日の最高気温の 2019年5月の前後半の比較



- 箱ひげ図から分布の大体の様子である5数要約が分かる。
- 中央値は後半の方が高いので全体的には後半の方が最高気温は高い。
- 四分位範囲を見ると、前半は21～25℃、後半は25～29℃の日に半分のデータが集中している。

複数の分布を比較する際に、箱ひげ図は便利

5月の金沢市 毎日の最高気温データの分析

中央値, 範囲, 四分位範囲
を2014～2019年で比較

2019年5月の金沢の最高気温

データを並べ替えて5要約数を求める

最小値16

中央値 $x_{16} = 25$

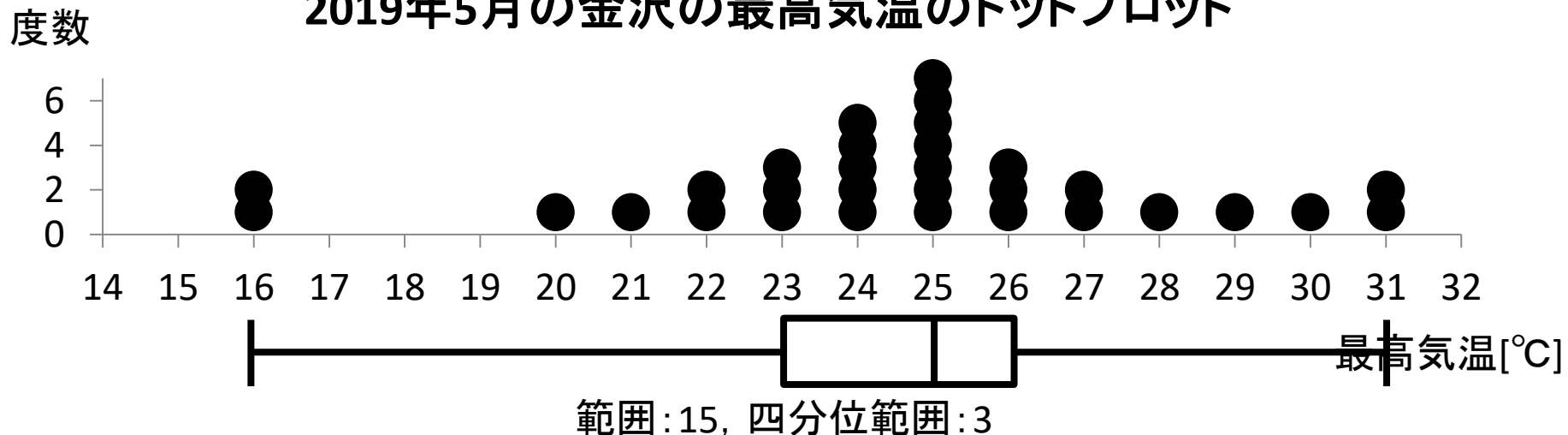
最大値31

16, 16, 20, 21, 22, 22, 23, 23, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 26, 26, 27, 27, 28, 29, 30, 31, 31

第1 $(x_8 + x_9)/2 = 23$

第3 $(x_{23} + x_{24})/2 = 26$

2019年5月の金沢の最高気温のドットプロット



2018年5月金沢市の最高気温データ

日にち	1	2	3	4	5
最高気温℃	25	23	21	18	21
2018年5月 金沢の最高 気温を小数 点第1位で 四捨五入し たデータ	6	7	8	9	10
	26	16	17	17	16
	11	12	13	14	15
	23	23	19	18	27
	16	17	18	19	20
	30	27	32	15	19
	21	22	23	24	25
	24	25	24	20	27
	26	27	28	29	30
	27	24	28	27	23
	31	データソース: 気象庁 各種データ・資料 > 過去の気象データ検索 > 日ごとの値 金沢 2018年5月(日ごとの値) 主な要素			
	24				

2018年5月の金沢の最高気温

データを並べ替えて5要約数を求める

最小値15

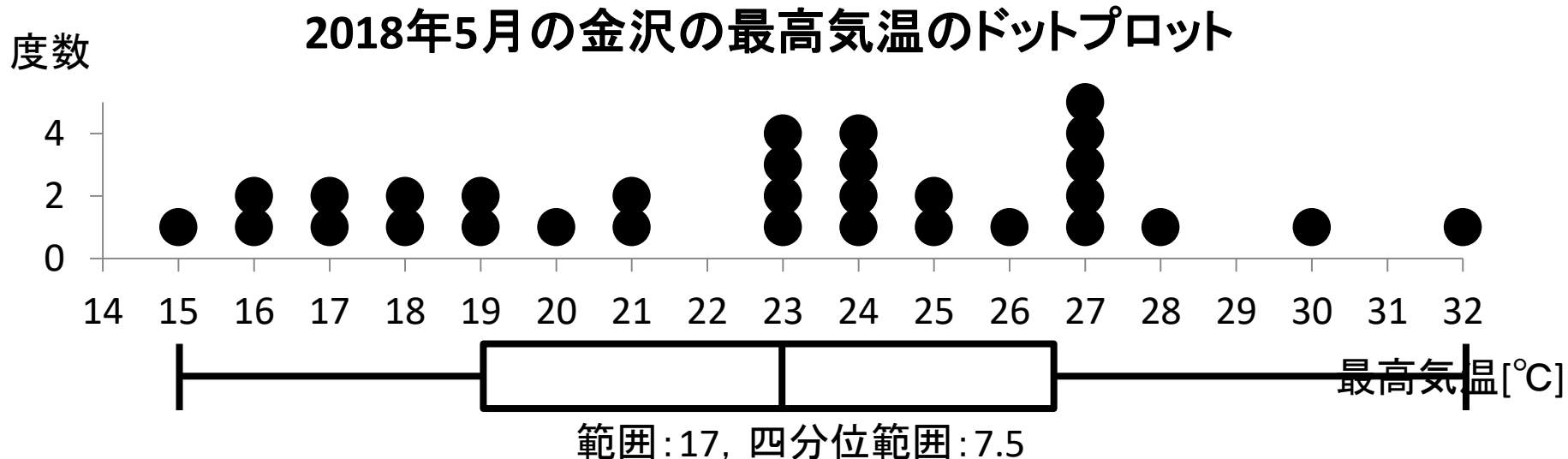
中央値 $x_{16} = 23$

最大値32

15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 21, 21, 23, 23, 23, 23, 24, 24, 24, 24, 25, 25, 26, 27, 27, 27, 27, 28, 30, 32

第1 $(x_8 + x_9)/2 = 19$

第3 $(x_{23} + x_{24})/2 = 26.5$



2017年5月金沢市の最高気温データ

日にち	1	2	3	4	5
最高気温℃	20	20	24	24	25
2017年5月 金沢の最高 気温を小数 点第1位で 四捨五入し たデータ	6	7	8	9	10
	20	21	22	22	23
	11	12	13	14	15
	21	24	21	23	21
	16	17	18	19	20
	20	21	21	25	29
	21	22	23	24	25
	27	27	26	23	24
	26	27	28	29	30
	25	21	22	28	29
	31	データソース: 気象庁 各種データ・資料 > 過去の気象データ検索 > 日ごとの値 金沢 2017年5月(日ごとの値) 主な要素			
	28				

2017年5月の金沢の最高気温

データを並べ替えて5要約数を求める

最小値20

中央値 $x_{16} = 23$

最大値29

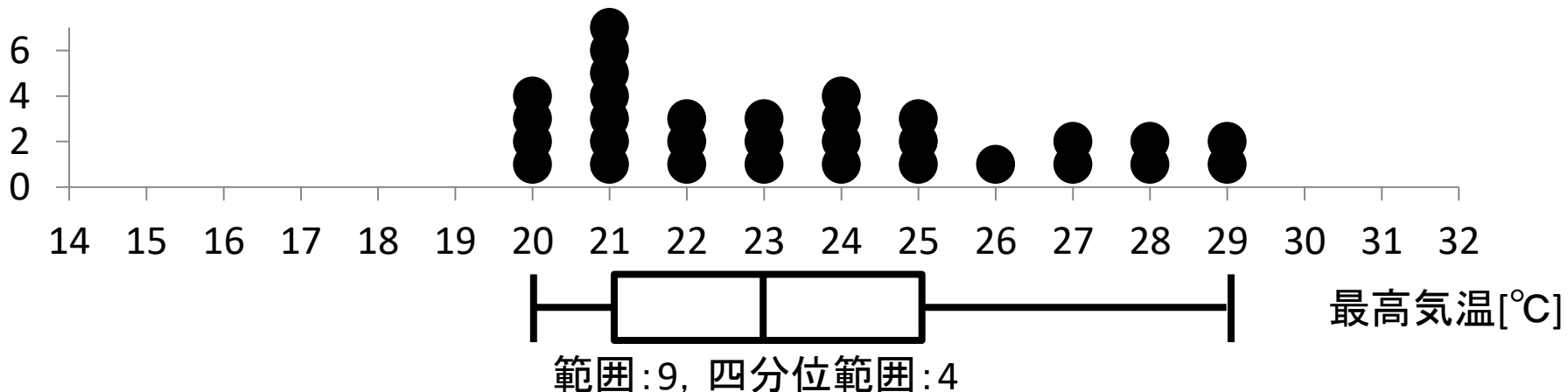
20, 20, 20, 20, 21, 21, 21, 21, 21, 21, 22, 22, 22, 23, 23, 23, 24, 24, 24, 24, 25, 25, 25, 26, 27, 27, 28, 28, 29, 29

第1 $(x_8 + x_9)/2 = 21$

第3 $(x_{23} + x_{24})/2 = 25$

度数

2017年5月の金沢の最高気温のドットプロット



2016年5月金沢市の最高気温データ

日にち	1	2	3	4	5
最高気温℃	21	25	28	25	20
2016年5月 金沢の最高 気温を小数 点第1位で 四捨五入し たデータ	6	7	8	9	10
	25	22	22	22	22
	11	12	13	14	15
	24	19	25	22	27
	16	17	18	19	20
	31	19	25	25	24
	21	22	23	24	25
	23	23	27	31	28
	26	27	28	29	30
	30	25	24	28	21
	31	データソース: 気象庁 各種データ・資料 > 過去の気象データ検索 > 日ごとの値 金沢 2016年5月(日ごとの値) 主な要素			
	25				

2016年5月の金沢の最高気温

データを並べ替えて5要約数を求める

最小值19

中央値 $x_{16} = 25$

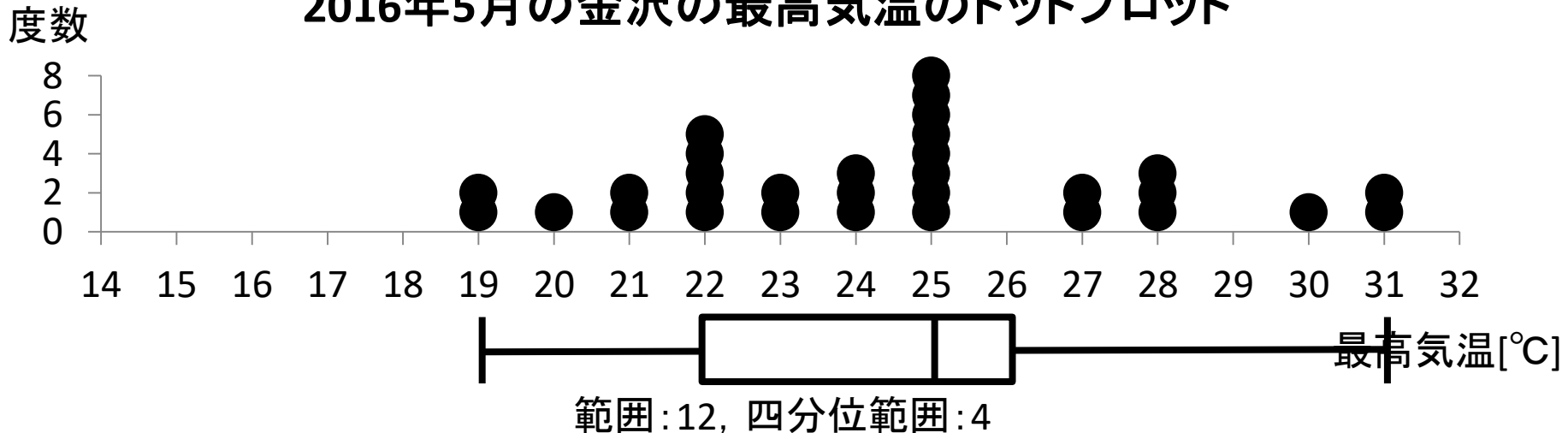
最大值31

19, 19, 20, 21, 21, 22, 22, 22, 22, 23, 23, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 27, 27, 28, 28, 28, 30, 31, 31

第1 $(x_8 + x_9)/2 = 22$

第3 $(x_{23}+x_{24})/2 = 26$

2016年5月の金沢の最高気温のドットプロット



2015年5月金沢市の最高気温データ

日にち	1	2	3	4	5
最高気温℃	24	26	26	24	19
2015年5月 金沢の最高 気温を小数 点第1位で 四捨五入し たデータ	6	7	8	9	10
	23	22	23	22	17
	11	12	13	14	15
	21	28	27	27	29
	16	17	18	19	20
	18	22	26	21	22
	21	22	23	24	25
	21	27	26	25	24
	26	27	28	29	30
	25	27	25	28	25
	31	データソース: 気象庁 各種データ・資料 > 過去の気象データ検索 > 日ごとの値 金沢 2015年5月(日ごとの値) 主な要素			
	25				

2015年5月の金沢の最高気温

データを並べ替えて5要約数を求める

最小値17

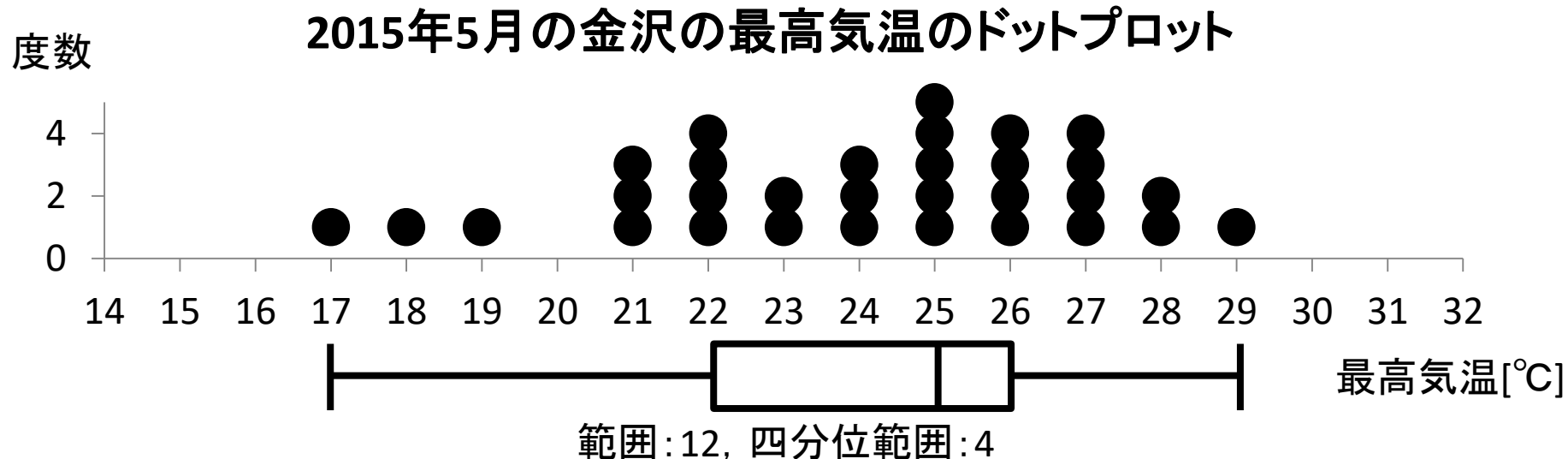
中央値 $x_{16} = 25$

最大値29

17, 18, 19, 21, 21, 21, 22, 22, 22, 23, 23, 24, 24, 24, 25, 25, 25, 25, 25, 26, 26, 26, 26, 27, 27, 27, 27, 28, 28, 29

第1 $(x_8 + x_9)/2 = 22$

第3 $(x_{23} + x_{24})/2 = 26$



2014年5月金沢市の最高気温データ

日にち	1	2	3	4	5
最高気温℃	18	21	22	21	19
2014年5月 金沢の最高 気温を小数 点第1位で 四捨五入し たデータ	6	7	8	9	10
	18	20	27	18	19
	11	12	13	14	15
	24	26	23	29	22
	16	17	18	19	20
	21	18	21	26	24
	21	22	23	24	25
	18	18	17	25	28
	26	27	28	29	30
	22	25	30	30	29
	31	データソース: 気象庁 各種データ・資料 > 過去の気象データ検索 > 日ごとの値 金沢 2014年5月(日ごとの値) 主な要素			
	28				

2014年5月の金沢の最高気温

データを並べ替えて5要約数を求める

最小値17

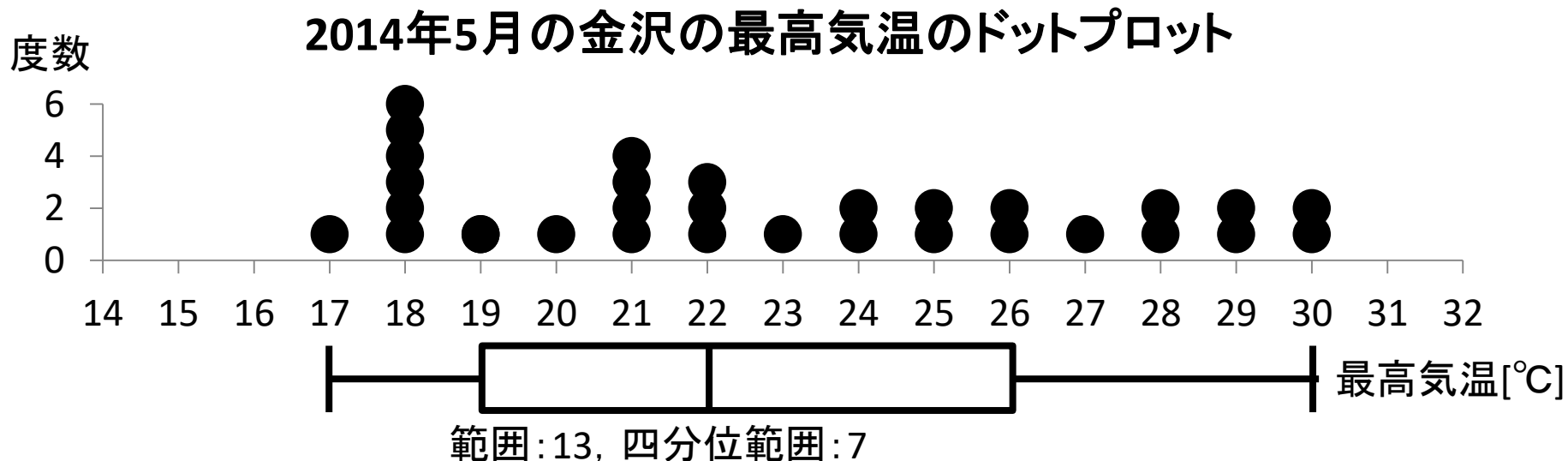
中央値 $x_{16} = 22$

最大値30

17, 18, 18, 18, 18, 18, 18, 18, 19, 19, 20, 21, 21, 21, 21, 22, 22, 22, 23, 24, 24, 25, 25, 26, 26, 27, 28, 28, 29, 29, 30, 30

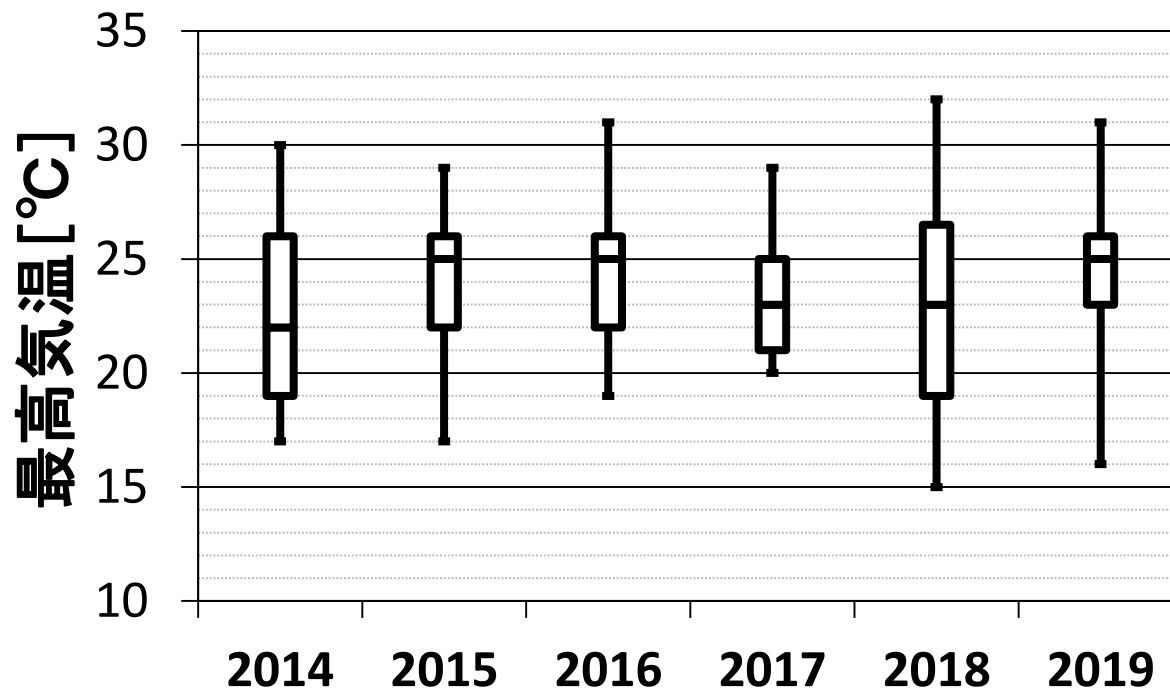
第1 $(x_8 + x_9)/2 = 19$

第3 $(x_{23} + x_{24})/2 = 26$



箱ひげ図で比較

金沢市5月の毎日の最高気温 の2014～2019年の比較



2014年と2015年の比較

- 中央値は2015年の方が高いので全体的には2015年の方が最高気温は高い。
- レンジはどちらも殆ど同じ。
- 四分位範囲は2014年は広く、2015年は狭い。
- 2014年は全体にまんべんなくデータが分布しているのに対して、2015年は25℃付近に集中して存在してる。

複数の分布を比較する際に、箱ひげ図は便利