

# データサイエンス入門B

## 第8回 ヒストグラム・箱ひげ図

高田 美樹

# 目次

---

- ▶ データの分類
- ▶ 質的データと量的データ
- ▶ データの分類
- ▶ データのばらつき（分布）の記述
- ▶ データの収集（気象庁）
- ▶ データの整形
- ▶ ヒストグラム
- ▶ 基本統計量
- ▶ 累積相対度数
- ▶ 四分位と箱ひげ図
- ▶ ヒストグラムの形状

# データの分類

データ

定型データ

非定型データ

時系列データ

クロスセクションデータ

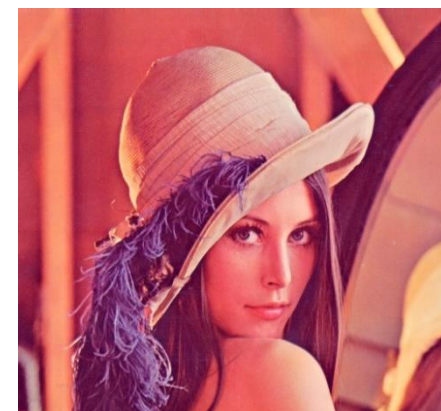
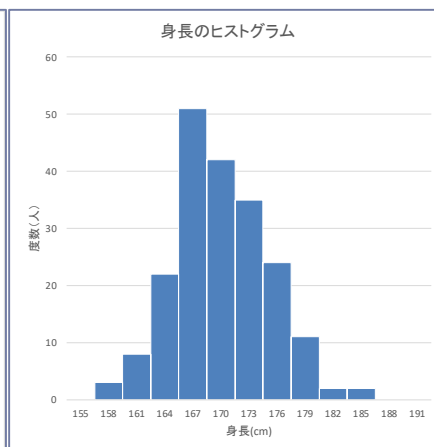
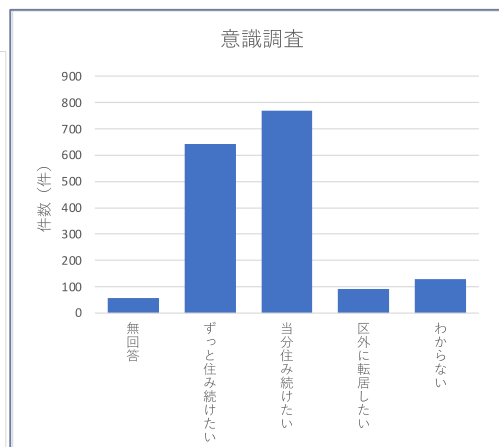
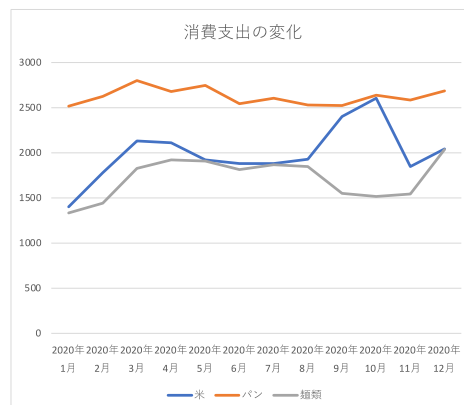
パネルデータ

画像・音・文書

質的データ

量的データ

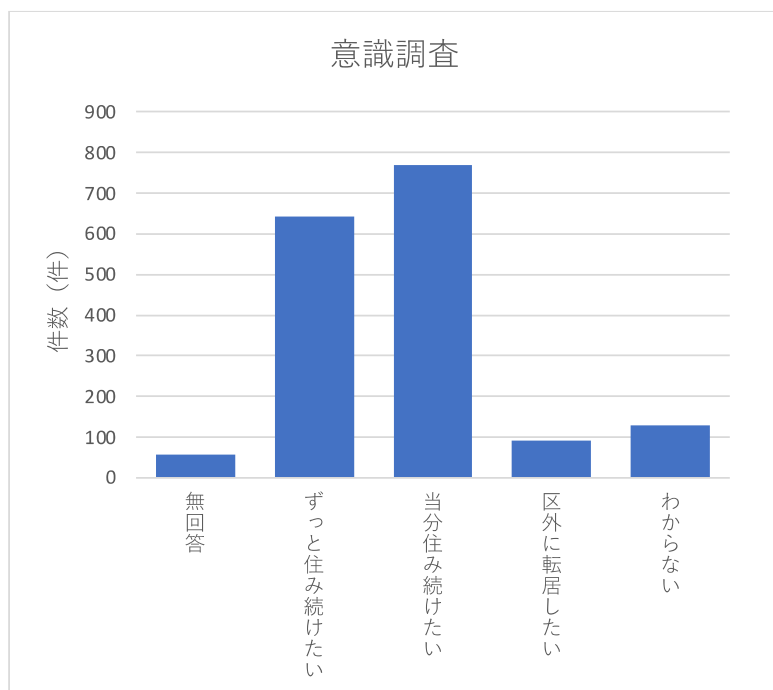
時系列データ＋  
クロスセクションデータ



# クロスセクションデータ

## 質的データ

少数個のカテゴリのどれかが応答値となる  
観測項目



性別 (男性・女性・無回答)

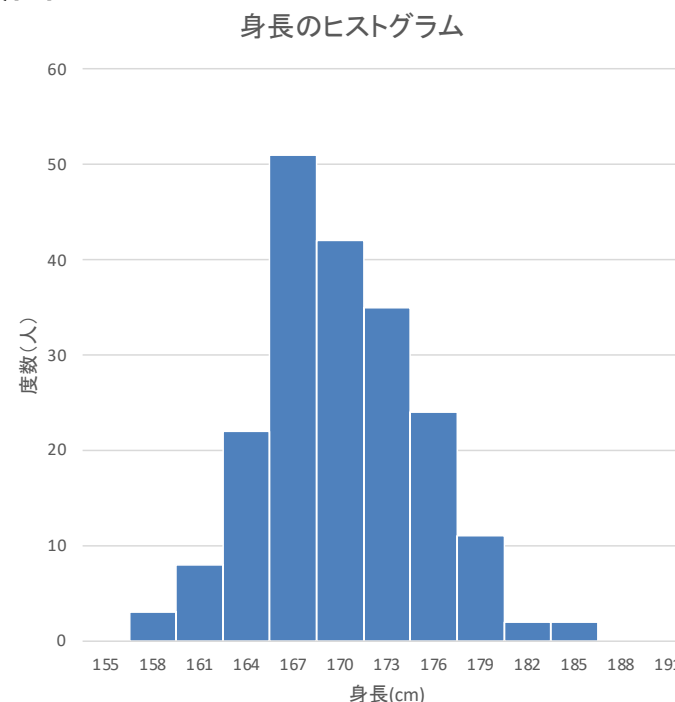
職種 (技術・営業・事務)

好感度

▶ 4 (好き・やや好き・どちらでもない・やや嫌い・嫌い)

## 量的データ

数量値 (数直線上の値) として、記録される  
観測項目



売上高・歩留まり

給与額・在庫量

# 質的データと量的データ

	A	B	C	D	E	F	G	H	I	J	レコード
1	11 月受注一覧										
2	受注日	会員番号	氏名	商品番号	商品名	個数	単価	税込み金額	税率	送料	請求額
3	10月11日	907	田中 美佐枝	107	ハート抜型10個入り	1	4,910	5,401	10%	234	5,635
4	10月11日	902	三浦 鞠子	102	薄力粉(1Kg)	3	429	1,390	8%	324	1,714
5	10月12日	909	相沢 恵	109	粉ふるい大	2	1,771	3,896	10%	945	4,841
6	10月12日	905	本間 佐知子	104	こしあん(500g)	8	475	4,104	8%	420	4,524
7	10月13日	903	佐藤 雅恵	108	星抜型10個入り	1	5,170	5,687	10%	324	6,011
8	10月13日	910	吉川 浩子	102	薄力粉(1Kg)	3	429	1,390	8%	324	1,714
9	10月13日	906	佐々木 真希	107	ハート抜型10個入り	2	4,910	10,802	10%	234	11,036
10	10月13日	901	山田 絵里子	109	粉ふるい大	1	1,771	1,948	10%	450	2,398
11	10月14日	908	内田 誠子	105	紅茶 アールグレイ(150g)	3	907	2,939	8%	945	3,884
12	10月15日	902	三浦 鞠子	105	紅茶 アールグレイ(150g)	1	907	980	8%	324	1,304
13											

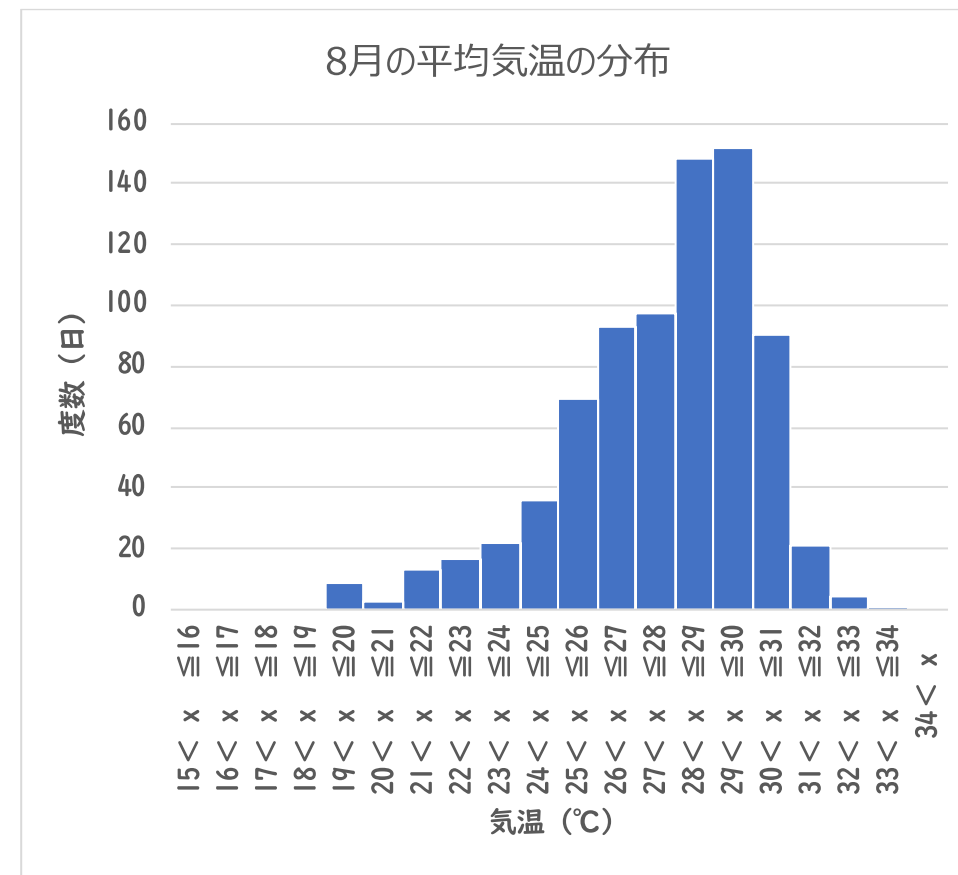
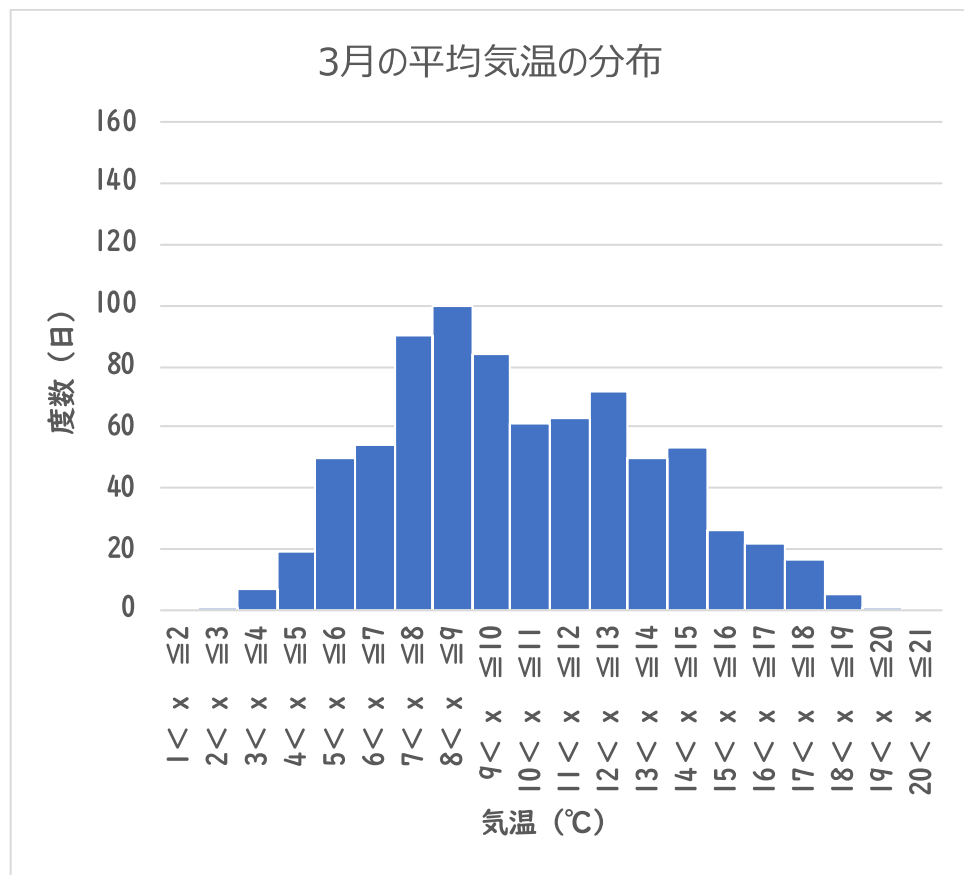
質的データ

量的データ

# データの分類

	尺度	性質	例
質的データ	名義尺度	区別のみ 順序はない	性別（男性・女性・無回答） 職業区分 （営業・技術・事務）
	順序尺度	区別 + 順序 差は取れない	2 値（はい・いいえ） 5 段階評価 （満足・やや満足・どちらでもない・やや不満・不満）
量的データ	間隔尺度	差の大きさに意味がある 比には意味がない	気温
	比例尺度	差も比も意味がある （絶対ゼロがある）	降水量

# データのばらつき（分布）の記述



# データの収集（気象庁）

国土交通省 気象庁 Japan Meteorological Agency

ホーム 防災情報 **各種データ・資料** 地域の情報 知識・解説 各種申請・ご案内

コンテンツの閲覧方法について（よくお寄せいただくご質問）

防災情報 天気 キキクル（危険度分布） 大雨・台風 地震・火山

被災地域等への支援情報

【お知らせ】5月24日以降、気象庁HPの閲覧に際しては、Internet Explorer以外のブラウザをご利用ください。

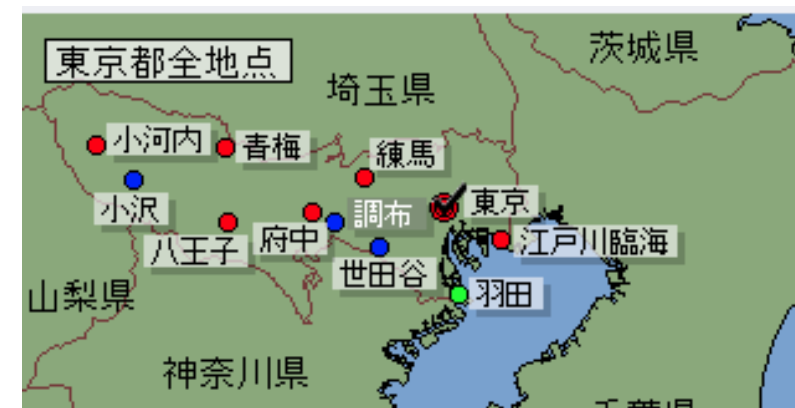
<https://www.jma.go.jp/jma/index.html>

地点を選ぶ 項目を選ぶ 期間を選ぶ 表示オプションを選ぶ

すべての選択済みの地点をクリア

一度にリクエストできるデータ量には上限があります（右上棒グラフ参照）。  
 アクセス集中時や一回のデータ量が多い場合、繋がらないことやデータ取得に時間がかかることがあります。繋がらない場合は時間を置いて再度お試しください。  
 アクセス集中の原因となりますので、自動化ツール等による過度のアクセスはお控えいただくようお願いいたします。

まず、都道府県を選んでください



- 気象**
  - 気象観測データ
    - 最新の気象データ
    - 過去の気象データ検索
    - 過去の地点気象データ・ダウンロード**
    - 過去の地域平均気象データ検索
    - 過去の天気図 / 日々の天気図
    - 過去の台風資料
    - 梅雨入りと梅雨明け 確定値
- 地球環境・気候**
  - 地球環境・気候
    - 地球環境・気候情報の総合ページ
    - 地球温暖化情報ポータル
    - 異常気象
      - 日本の異常気象
      - 世界の異常気象
    - 天候の特徴や見通し
      - 気候系の監視・診断
      - 日本の天候の特徴と見通し
- 海洋**
  - 海洋
    - 海洋の情報 波浪 / 海水温・海流 / 海氷
    - 海洋の健康診断表
    - 海洋の実況や見通し
      - 日本沿岸の潮位
      - オホーツク海の海氷
      - 日本近海の海面水温 月概況 / 旬の状況と今後の見通し
      - 日本近海の海流 月概況 / 旬の状況と今後の見通し
- 地震・津波・火山**
  - 地震の活動状況
    - 最新の活動状況(速報データ)
    - 最近1週間程度の活動状況
    - 各月の地震活動のまとめ
  - 地震・津波の観測・解析データ
    - 震源リスト
    - 震度データベース検索
    - 発震機構解
    - 国内の地震の解析結果
    - 海外の地震の解析結果

# データの収集（気象庁）

地点を選ぶ 項目を選ぶ 期間を選ぶ 表示オプションを選ぶ

データの種類 ? 詳細

- ☐ 時別値
- ☒ 日別値
- ☐ 2日別値
- ☐ 半旬別値
- ☐ 旬別値
- ☐ 月別値
- ☐ 3か月別値※

最初に選択してください

過去の平均値との比較オプション

- ☐ 平年値も表示
- ☐ 平年値からの差(比)も表示  
(平年値:1991年から2020年の30年平均値)
- ☐ 前年までの1年平均も表示
- ☐ 前年までの1年平均からの差(比)も表示

項目 気温 降水 日照/日射 積雪/降雪 風 湿度/気圧 雲量/天気

☒ 日平均気温

- ☐ 日最高気温の日平均
- ☐ 日最低気温の日平均
- ☐ 日最高気温
- ☐ 日最低気温
- ☐ 日最高気温の日最低※
- ☐ 日最低気温の日最高※

- ☐ 日平均気温 25℃以上の日数(日)
- ☐ 日平均気温 0℃未満の日数(日)
- ☐ 日最高気温 25℃以上の日数(日)
- ☐ 日最高気温 0℃未満の日数(日)
- ☐ 日最低気温 25℃以上の日数(日)
- ☐ 日最低気温 0℃未満の日数(日)

※官署(気象台等)のみ値があります

地点を選ぶ 項目を選ぶ 期間を選ぶ 表示オプションを選ぶ

期間

☒ 連続した期間で表示する

最近1年 最近1か月

1998年 8月 1日から  
2023年 7月 31日までの日別値を表示

☐ 特定の期間を複数年分、表示する

8月 1日から 7月 31日の値を  
1998年から 2023年まで表示

画面に表示 ▶

CSVファイルをダウンロード ▶

選択地点・項目をクリア

選択された地点 観測項目  
東京

選択された項目  
日平均気温

☒ 連続した期間で表示する

最近1年

最近1か月

1998年 8月 1日から

2023年 7月 31日までの日別値を表示

# データの収集（気象庁）

	A	B	C	D	E
1	ダウンロードした時刻：2023/08/23 03:38:37				
2					
3		東京	東京	東京	削除
4	年月日	平均気温(°C)	平均気温(°C)	平均気温(°C)	
5			品質情報	均質番号	
6	1998/8/1	26.4	8	1	
7	1998/8/2	29.5	8	1	
8	1998/8/3	29.5	8	1	
9	1998/8/4	28.2	8	1	
10	1998/8/5	28.2	8	1	
11	1998/8/6	27.5	8	1	
12	1998/8/7	27.6	8	1	
13	1998/8/8	23.7	8	1	
14	1998/8/9	22.2	8	1	
15	#####	25.4	8	1	
16	#####	27.7	8	1	
17	#####	28.8	8	1	
18	#####	29	8	1	



	A	B	C	D
1	ダウンロードした時刻：2023/08/23 03:38:37			
2				
3		東京		
4	年月日	平均気温(°C)		
5	1998/8/1	26.4		
6	1998/8/2	29.5		
7	1998/8/3	29.5		
8	1998/8/4	28.2		
9	1998/8/5	28.2		
10	1998/8/6	27.5		
11	1998/8/7	27.6		
12	1998/8/8	23.7		
13	1998/8/9	22.2		
14	1998/8/10	25.4		
15	1998/8/11	27.7		
16	1998/8/12	28.8		
17	1998/8/13	29		
18	1998/8/14	27.5		

幅が足りない

# データの整形

右クリック  
挿入

=MONTH(A5)

「月」を追加

年月日	月	東京 平均気温(°C)
1998/8/1	8	26.4
1998/8/2	8	29.5
1998/8/3	8	29.5
1998/8/4	8	28.2
1998/8/5	8	28.2
1998/8/6	8	27.5
1998/8/7	8	27.6
1998/8/8	8	23.7
1998/8/9	8	22.2
1998/8/10	8	25.4
1998/8/11	8	27.7
1998/8/12	8	28.8
1998/8/13	8	29
1998/8/14	8	27.5
1998/8/15	8	26.8
1998/8/16	8	30.6
1998/8/17	8	29.3
1998/8/18	8	25.7
1998/8/19	8	25.3

データ

フィルター

クリック

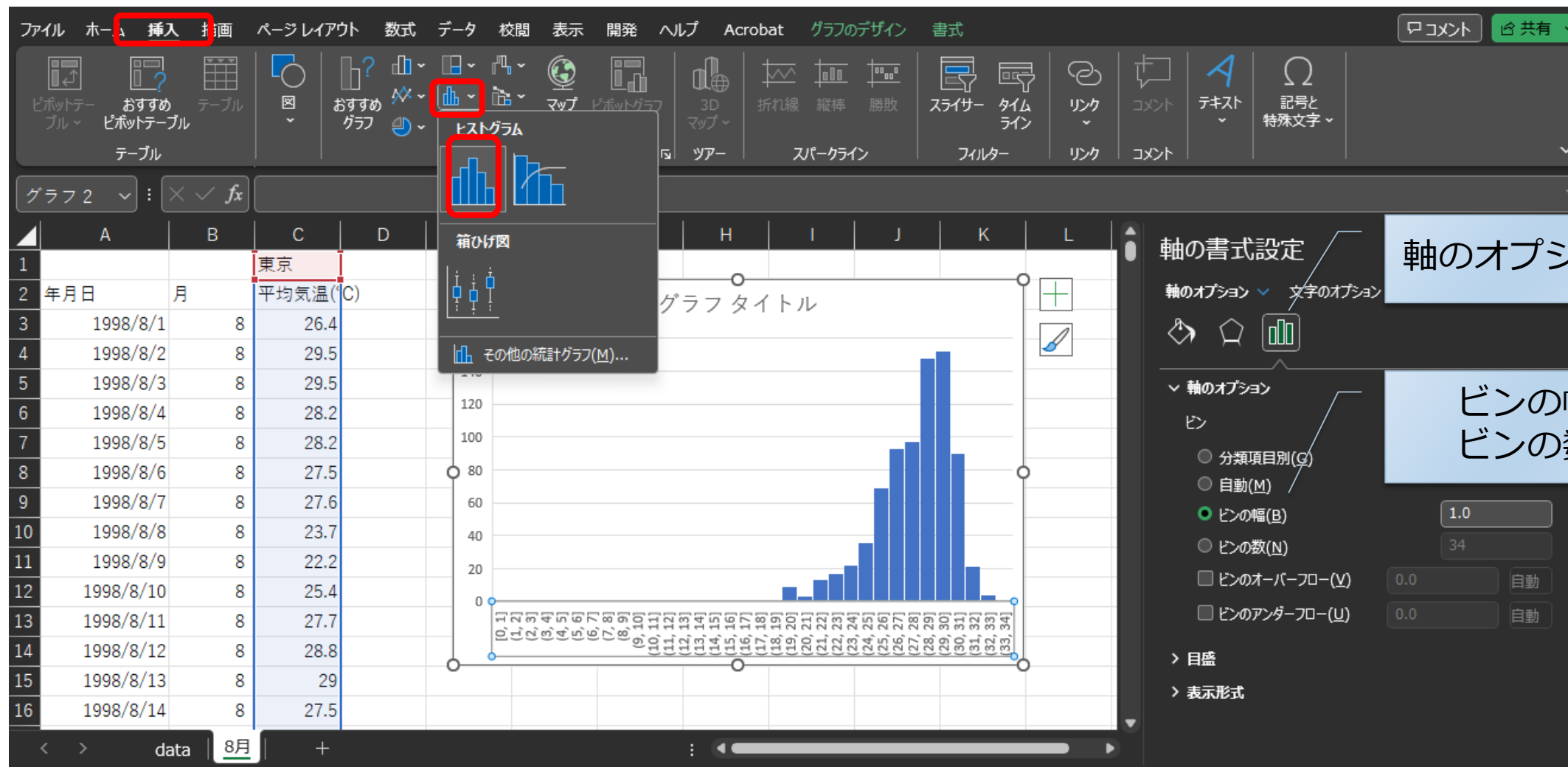
「8月」を抽出

「8」だけ選択

年月日	月	東京 平均気温(°C)
1998/8/1	8	26.4
1998/8/2	8	29.5
1998/8/3	8	29.5
1998/8/4	8	28.2
1998/8/5	8	28.2
1998/8/6	8	27.5
1998/8/7	8	27.6
1998/8/8	8	23.7
1998/8/9	8	22.2
1998/8/10	8	25.4
1998/8/11	8	27.7
1998/8/12	8	28.8
1998/8/13	8	29
1998/8/14	8	27.5
1998/8/15	8	26.8
1998/8/16	8	30.6
1998/8/17	8	29.3
1998/8/18	8	25.7
1998/8/19	8	25.3

# ヒストグラム

挿入



# ヒストグラム

## ▶ スタージャスの公式

$$m = 1 + \log_2 n$$

$m$ :階級数  
 $n$ :データ数

$n$	775
$m$	10.5980525

=COUNT(B1:B775)

=1+LOG(K25,2)

# 基本統計量

データ

データ分析



データ分析

分析ツール(A)

- 分散分析: 一元配置
- 分散分析: 繰り返しのある二元配置
- 分散分析: 繰り返しのない二元配置
- 相関
- 共分散
- 基本統計量**
- 指数平滑
- F 検定: 2 標本を使った分散の検定
- フーリエ解析
- ヒストグラム

OK

キャンセル

ヘルプ(H)

基本統計量

入力元

入力範囲(I) **\$C\$3:\$C\$777**

データ方向:

- ☒ 列(C)
- ☐ 行(R)

☐ 先頭行をラベルとして使用(L)

出力オプション

- ☒ 出力先(O): **\$E\$17**
- ☐ 新規ワークシート(P):
- ☐ 新規ブック(W):

☒ **統計情報(S)**

統計情報

- ☐ 平均の信頼度の出力(N): 95 %
- ☐ K 番目に大きな値(A): 1
- ☐ K 番目に小さな値(M): 1

	A	B	C	D	E	F	G	H	I	J	K
24	1998/8/22	8	25.4								
25	1998/8/23	8	26.4								
26	1998/8/24	8	27.9								
27	1998/8/25	8	29.2								
28	1998/8/26	8	29.2								

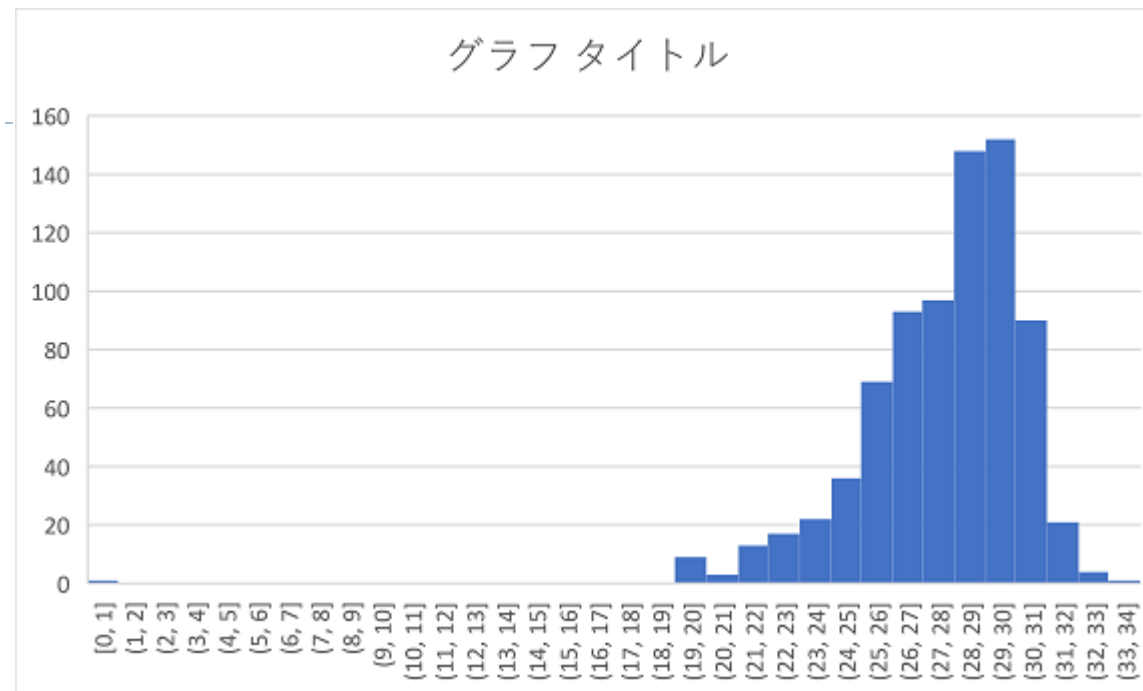
基本統計量

列1	
平均	27.7351
標準誤差	0.087659
中央値 (メジアン)	28.2
最頻値 (モード)	28.7
標準偏差	2.440317
分散	5.955149
尖度	0.899331
歪度	-0.94271
範囲	14.1
最小	19.1
最大	33.2
合計	21494.7
データの個数	775

# ヒストグラム

## ▶ ビンの幅は

平均	27.7351	
標準誤差	0.087659	
中央値 (メジアン)	28.2	
最頻値 (モード)	28.7	
標準偏差	2.440317	
分散	5.955149	
尖度	0.899331	
歪度	-0.94271	
範囲	14.1	
最小	19.1	10.59805
最大	33.2	1.330433
合計	21494.7	
データの個数	775	



ビンの個数 =  $1 + \text{LOG}(\text{データ個数}, 2)$

ビンの幅 =  $(\text{最大} - \text{最小}) / \text{ビンの個数}$

# ヒストグラム

データ

データ分析



データ分析

分析ツール(A)

- 相関
- 共分散
- 基本統計量
- 指数平滑
- F 検定: 2 標本を使った分散の検定
- フーリト解析
- ヒストグラム**
- 移動平均
- 乱数発生
- 順位と百分位数

	A	B	C	D	E	F	G	H	I	J	K	L	M
33	1998/8/31	8	23.5										
34	1999/8/1	8	29.2										
35	1999/8/2	8	29										
36	1999/8/3	8	29.5										
37	1999/8/4	8	29.3										
38	1999/8/5	8	28.8										
39	1999/8/6	8	28.6										
40	1999/8/7	8	28.8										
41	1999/8/8	8	29.5										
42	1999/8/9	8	29.1										
43	1999/8/10	8	26.3										
44	1999/8/11	8	29.1										
45	1999/8/12	8	29.4										
46	1999/8/13	8	27.9										
47	1999/8/14	8	26.3										
48	1999/8/15	8	28.2										
49	1999/8/16	8	28.6										
50	1999/8/17	8	29.8										

ヒストグラム

入力元

入力範囲(I): \$C\$2:\$C\$777

データ区間(B): \$F\$34:\$F\$51

☒ ラベル(L)

出力オプション

☒ 出力先(Q): \$G\$34

☐ 新規ワークシート(P):

☐ 新規ブック(W)

☐ パレート図(A)

☐ 累積度数分布の表示(M)

☐ グラフ作成(C)

OK

キャンセル

ヘルプ(H)

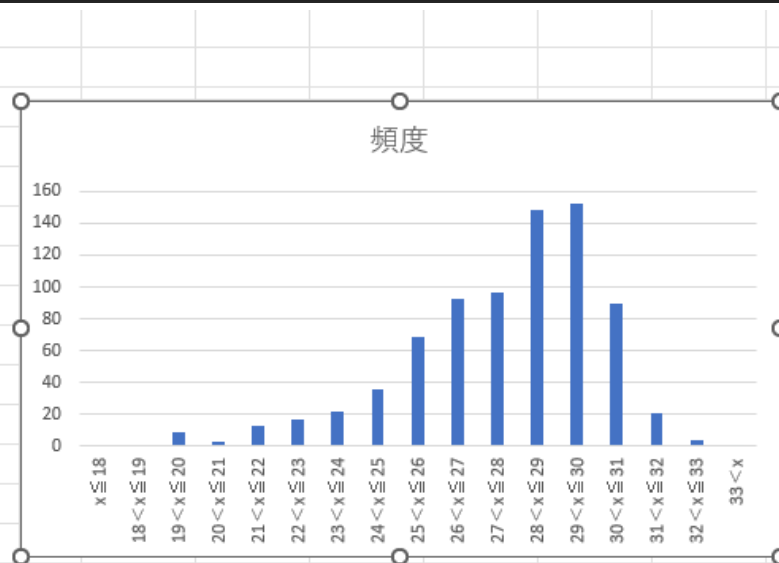
# ヒストグラム

挿入

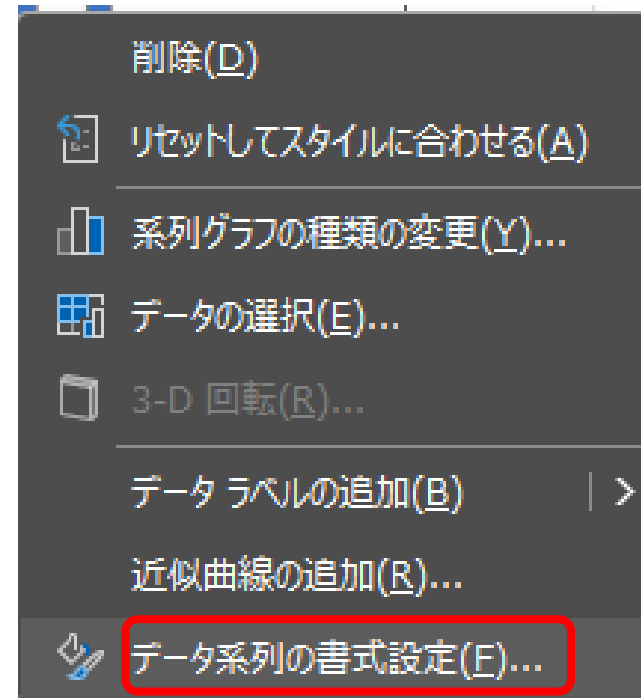


区間	区間	区間	頻度
$x \leq 18$	18	18	0
$18 < x \leq 19$	19	19	0
$19 < x \leq 20$	20	20	9
$20 < x \leq 21$	21	21	3
$21 < x \leq 22$	22	22	13
$22 < x \leq 23$	23	23	17
$23 < x \leq 24$	24	24	2
$24 < x \leq 25$	25	25	6
$25 < x \leq 26$	26	26	9
$26 < x \leq 27$	27	27	33
$27 < x \leq 28$	28	28	97
$28 < x \leq 29$	29	29	148
$29 < x \leq 30$	30	30	152
$30 < x \leq 31$	31	31	90
$31 < x \leq 32$	32	32	21
$32 < x \leq 33$	33	33	4
$33 < x$	34	34	1
次の級			0

=F35 & "< x ≤ "& F36



度数は求まっているので棒グラフ



# ヒストグラム

データ系列の書式設定

系列のオプション

系列のオプション

使用する軸

主軸 (下/左側)(P)

第 2 軸 (上/右側)(S)

系列の重なり(O) -27%

要素の間隔(W) 0%

データ系列の書式設定

系列のオプション

塗りつぶし

枠線

線なし(N)

線 (単色)(S)

線 (グラデーション)(G)

自動(U)

色(C)

透明度(I) 0%

幅(W) 0.75 pt

一重線/多重線(C)

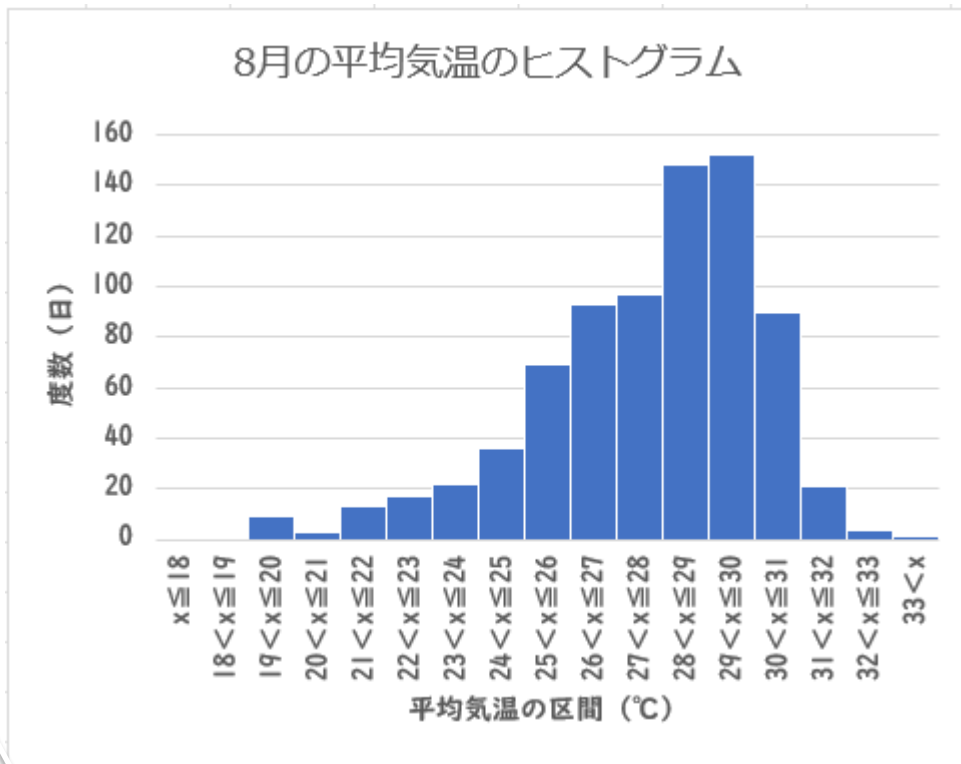
実線/点線(D)

線の先端(A) フラット

線の結合点(J) 丸

ビンをくっつける

枠線は白



# 累積相対度数

区間	区間	区間	頻度	相対度数	累積相対度数
$x \leq 18$	18	18	0	0.0%	0.0%
$18 < x \leq 19$	19	19	0	0.0%	0.0%
$19 < x \leq 20$	20	20	9	1.2%	1.2%
$20 < x \leq 21$	21	21	3	0.4%	1.5%
$21 < x \leq 22$	22	22	13	1.7%	3.2%
$22 < x \leq 23$	23	23	17	2.2%	5.4%
$23 < x \leq 24$	24	24	22	2.8%	8.3%
$24 < x \leq 25$	25	25	36	4.6%	12.9%
$25 < x \leq 26$	26	26	69	8.9%	21.8%
$26 < x \leq 27$	27	27	93	12.0%	33.8%
$27 < x \leq 28$	28	28	97	12.5%	46.3%
$28 < x \leq 29$	29	29	148	19.1%	65.4%
$29 < x \leq 30$	30	30	152	19.6%	85.0%
$30 < x \leq 31$	31	31	90	11.6%	96.6%
$31 < x \leq 32$	32	32	21	2.7%	99.4%
$32 < x \leq 33$	33	33	4	0.5%	99.9%
$33 < x$	34	34	1	0.1%	100.0%
		次の級	0		
		合計	775		

=I35

=H35/\$H\$53

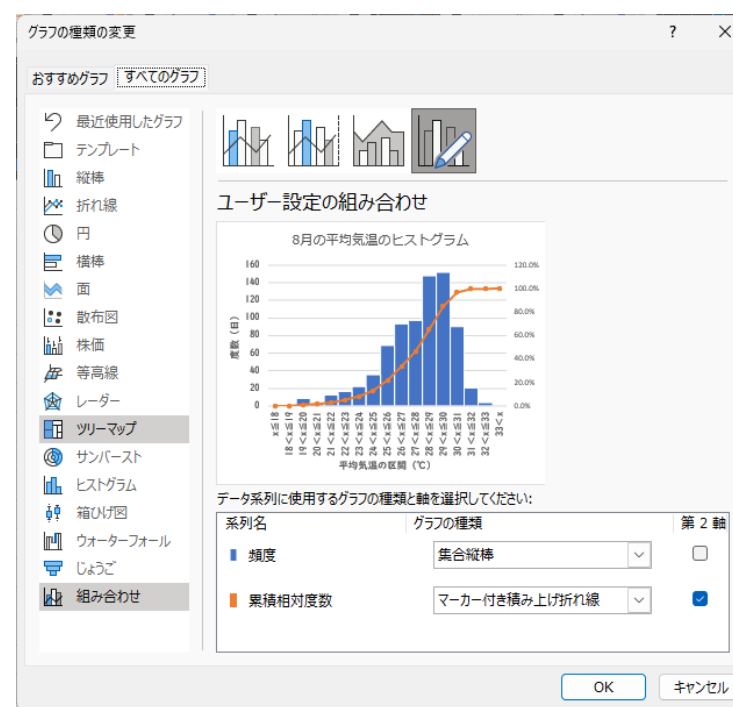
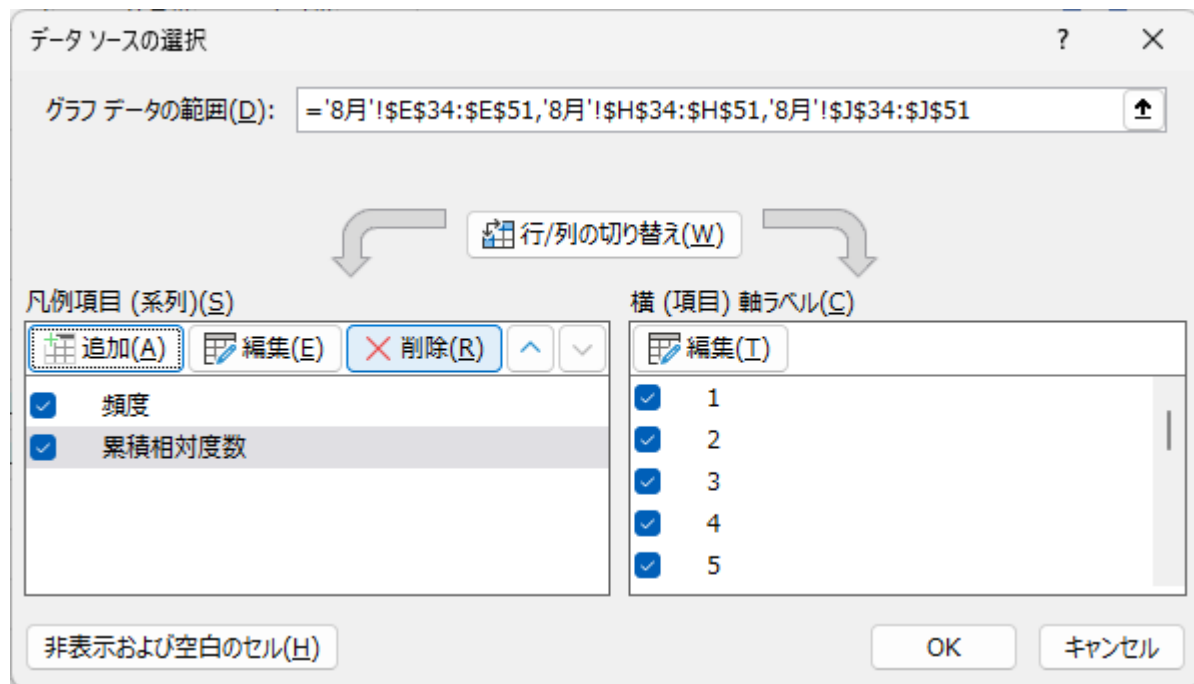
=J35+I36

=SUM(H35:H51)

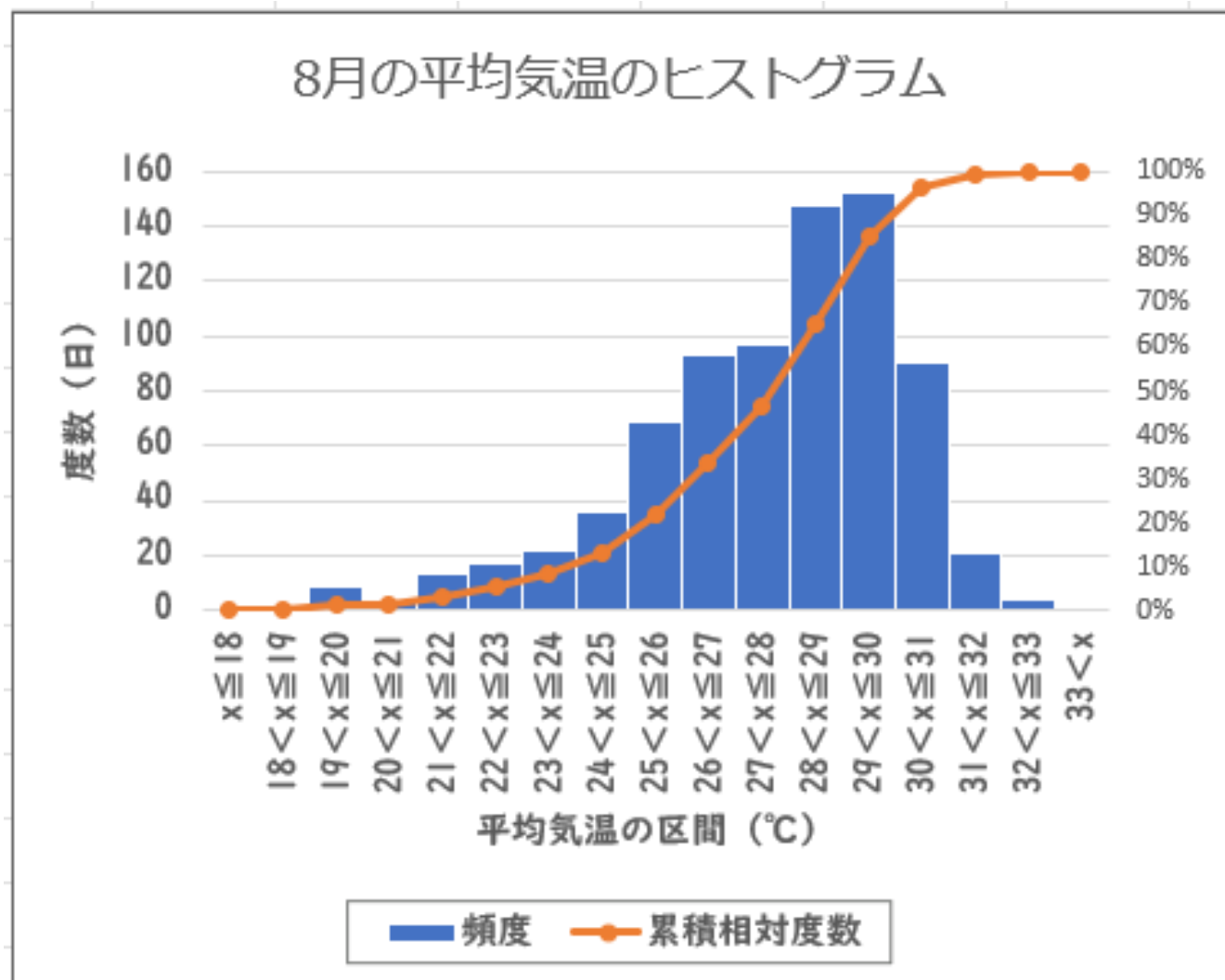


# 累積相対度数のグラフ

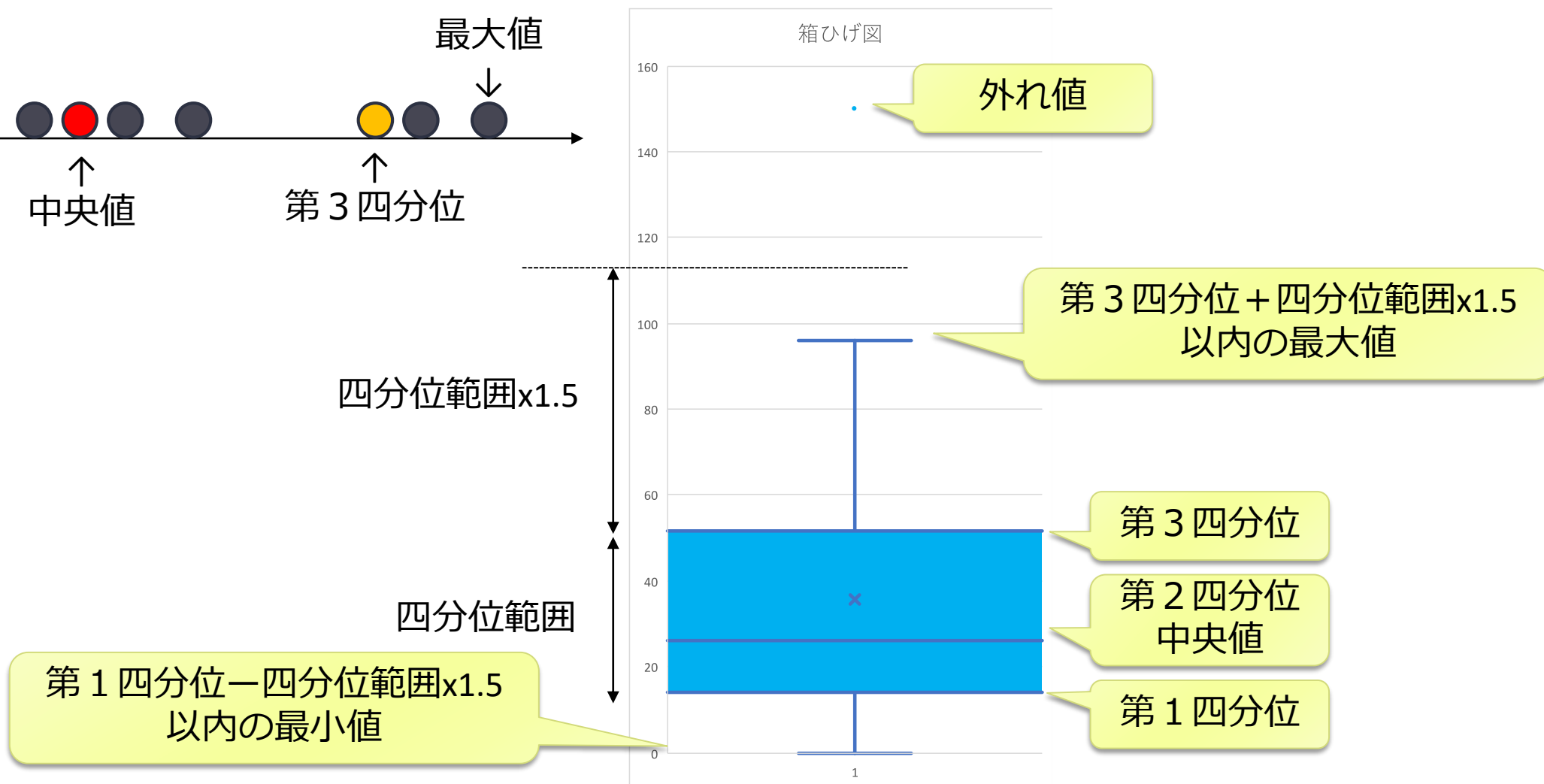
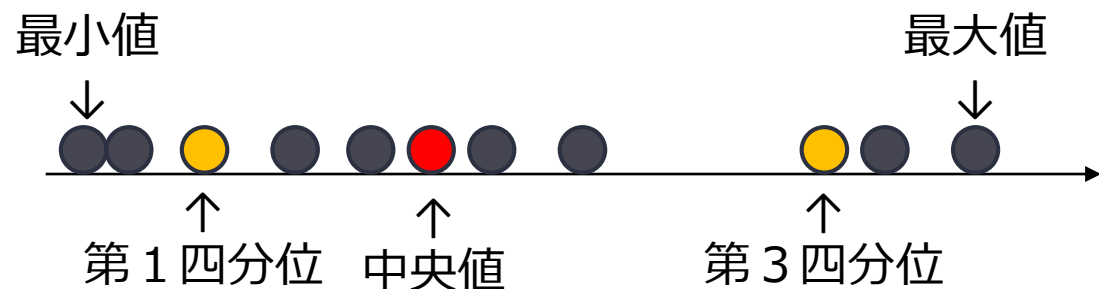
挿入



# 複合グラフ

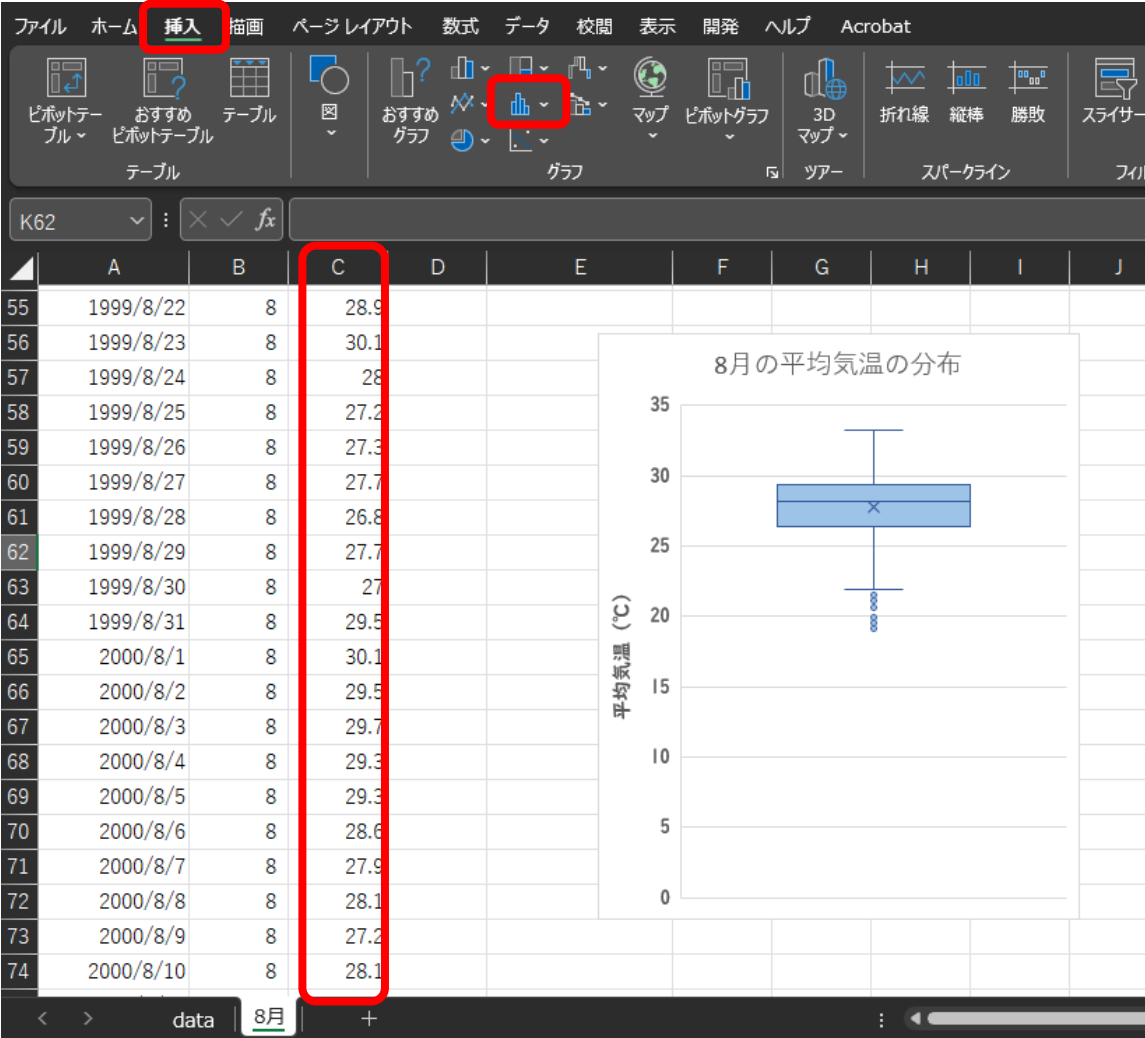


# 四分位と箱ひげ図



# 箱ひげ図と四分位

挿入



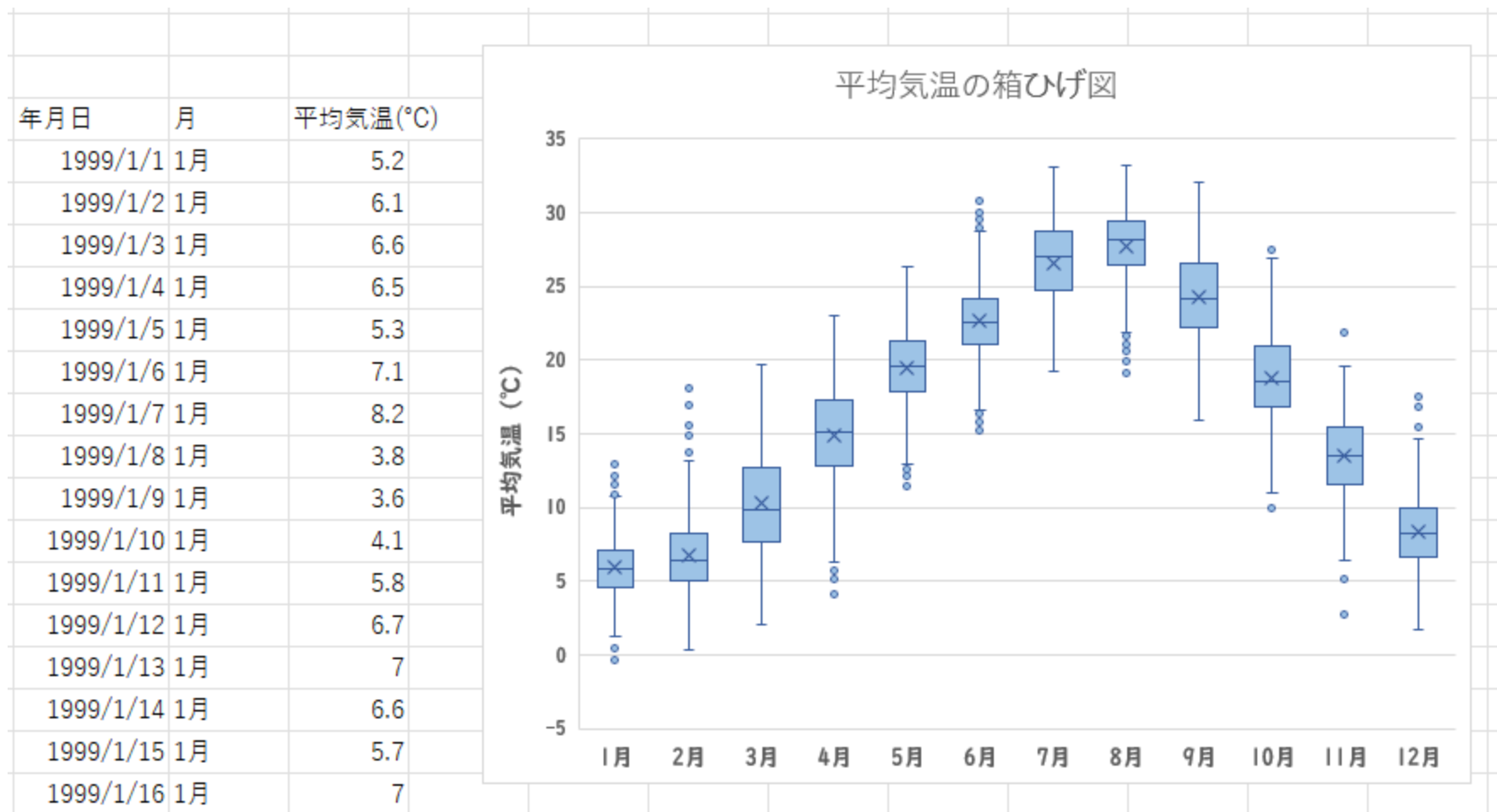
最小値	19.1	=QUARTILE.INC(\$C\$3:\$C\$777,0)
第 1 四分位	26.4	=QUARTILE.INC(\$C\$3:\$C\$777,1)
中央値	28.2	=QUARTILE.INC(\$C\$3:\$C\$777,2)
第 3 四分位	29.4	=QUARTILE.INC(\$C\$3:\$C\$777,3)
最大値	33.2	=QUARTILE.INC(\$C\$3:\$C\$777,4)
四分位範囲	3	=第 3 四分位 - 第 1 四分位

=第 3 四分位 + 四分位 範囲\*1.5

	四分位範囲*1.5	上下限	
ひげの上限	33.9	33.2	=最大値
ひげの下限	21.9	21.9	=下限値

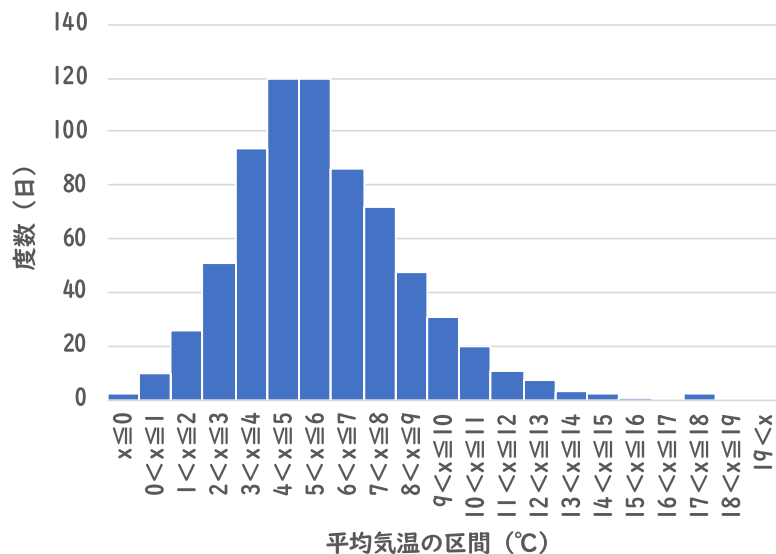
=第 1 四分位 - 四分位 範囲\*1.5

# 多種の箱ひげ図

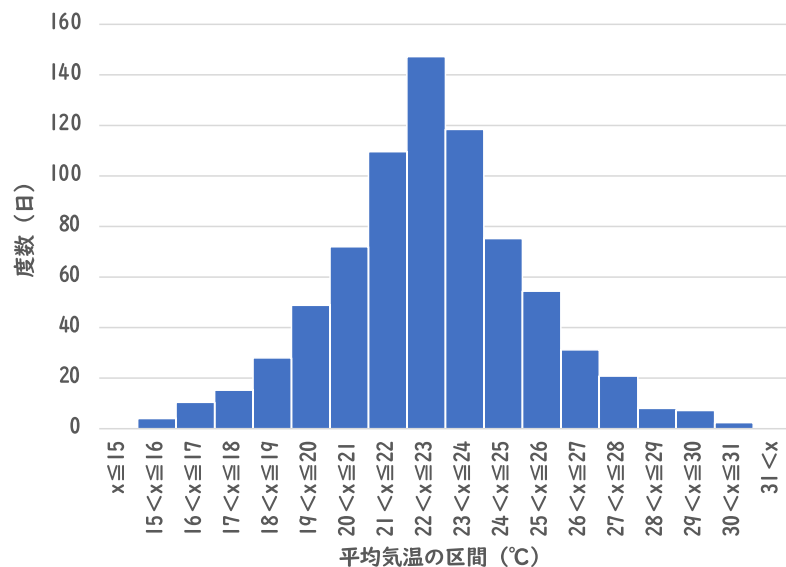


# ヒストグラムの形状

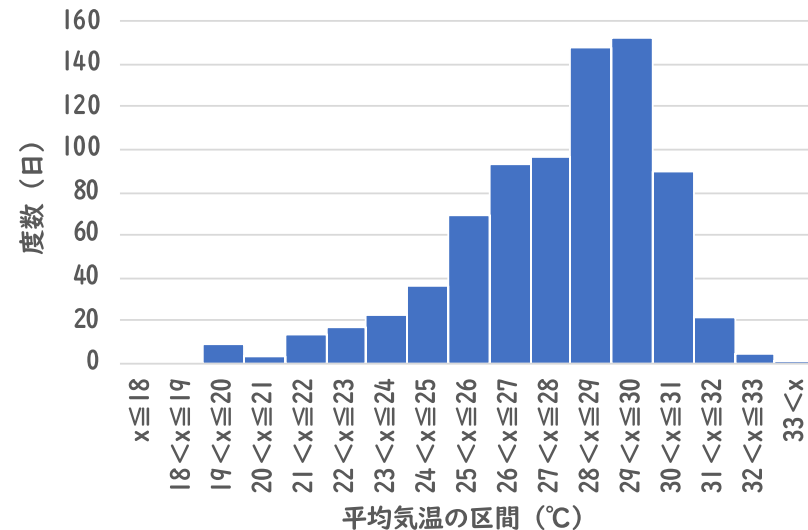
2月の平均気温のヒストグラム



6月の平均気温のヒストグラム



8月の平均気温のヒストグラム



右に歪んだグラフ

平均	6.79
中央値 (メジアン)	6.4
最頻値 (モード)	6

最頻値 < 中央値 < 平均値

左右対称なグラフ

平均	22.66
中央値 (メジアン)	22.6
最頻値 (モード)	22.2

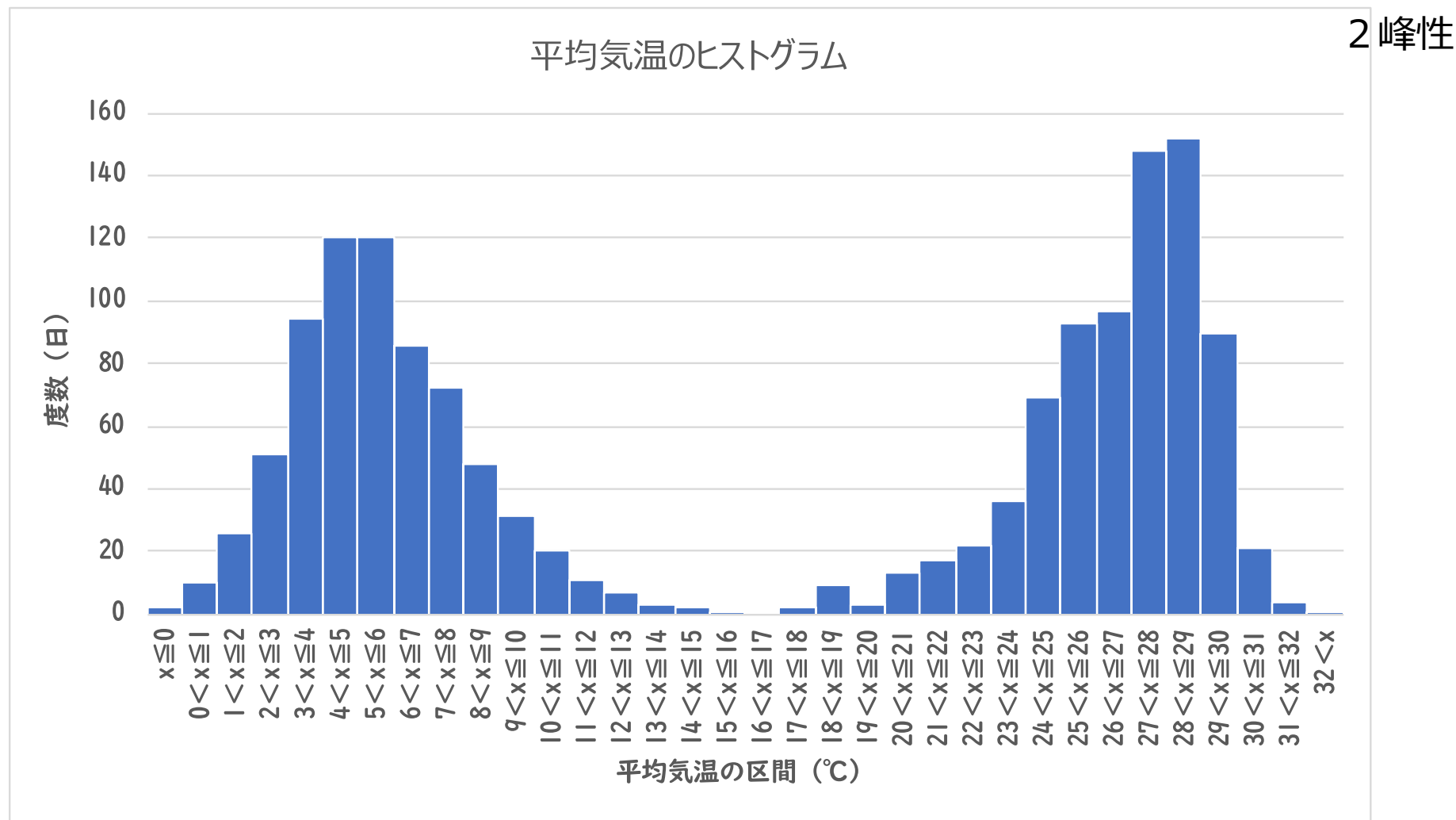
最頻値  $\approx$  平均値  $\approx$  中央値

左に歪んだグラフ

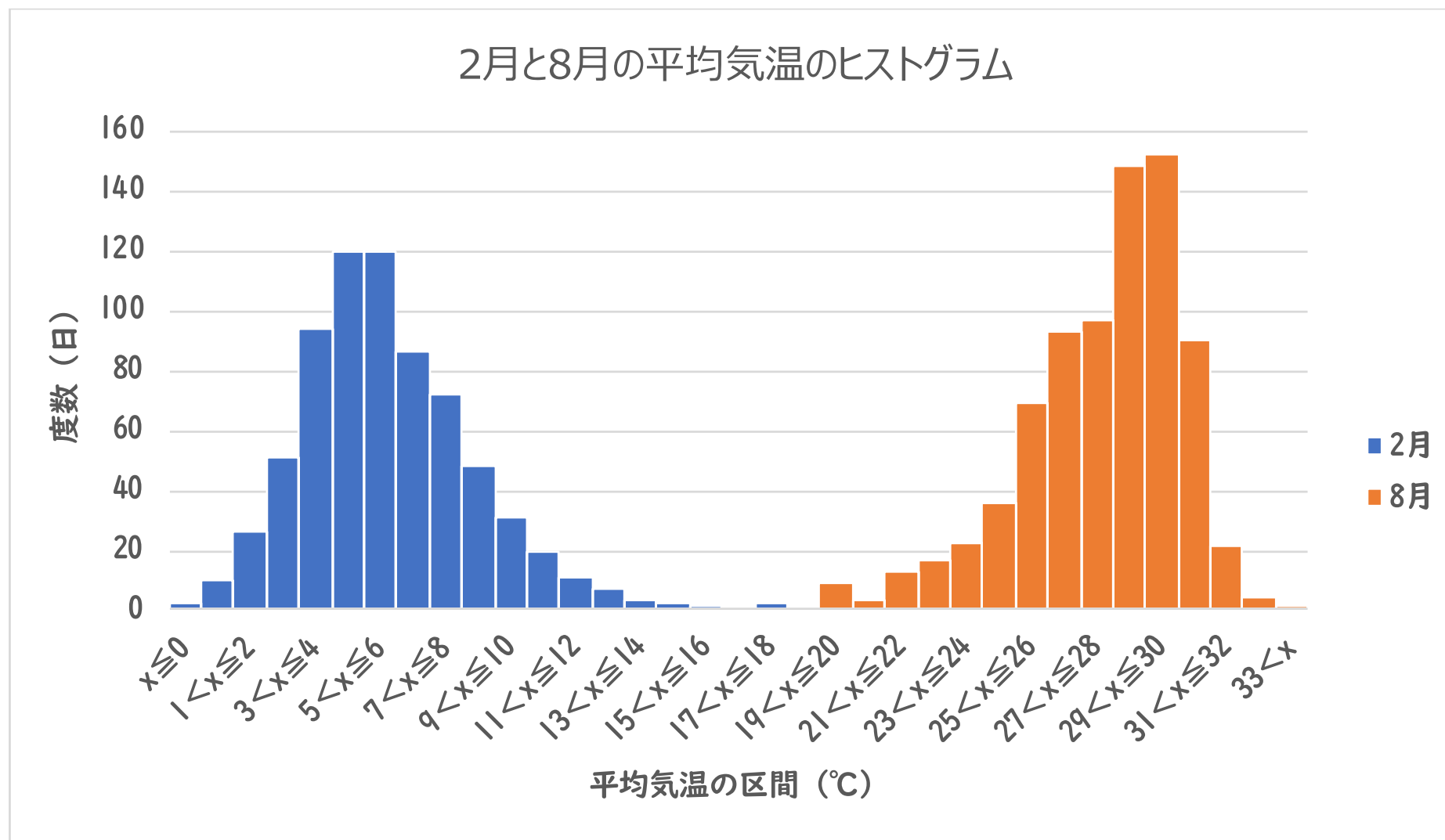
平均	27.74
中央値 (メジアン)	28.2
最頻値 (モード)	28.7

平均値 < 中央値 < 最頻値

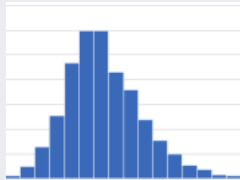
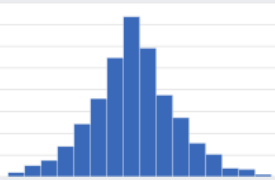
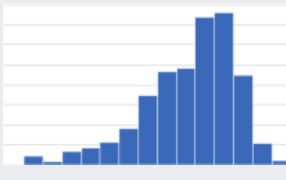
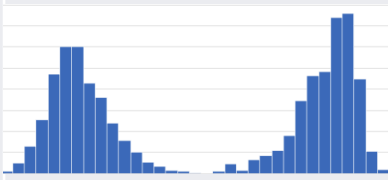
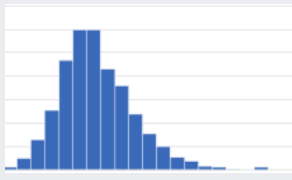
# ヒストグラムの形状



# ヒストグラムの形状



# まとめ

項目	内容				
クロスセクションデータの分類		尺度	性質		
	質的データ	名義尺度	区別のみ 順序はない		
		順序尺度	区別 + 順序 差は取れない		
	量的データ	間隔尺度	差の大きさに意味がある 比には意味がない		
		比例尺度	差も比も意味がある 絶対ゼロがある		
ヒストグラム	ビンの数・ビンの幅によって形が変わる				
ヒストグラムの形状	右に歪む	左右対称	左に歪む	多峰	外れ値
					
箱ひげ図	四分位				

## 参考文献

---

- ▶ 「身近な統計」 石崎克也・渡辺美智子 放送大学教育振興会
- ▶ 「誰でも使える統計オープンデータ」 オフィシャル スタディノート  
総務省統計局