

# 定量的なデータの分析 ～統計的な処理

情報 I 第48回授業

06情報通信ネットワークとデータベース

対応データ 23exp48.xls

## 今日のテーマ

- データの種類を知る
- 数値で示す
- 数値の「意味」を知る
- わかりやすく示す

# 【復習】「データ」と「情報」

- データ： 単なる数字や文字の羅列

例) 8890799568    ASAP

→ 価値を見いだしていない状態

- 情報： 意味のある数字や文字の羅列

例) 英数国理社の得点(88,90,79,95,68)

できるだけ早く！(As Soon As Possible)

→ そこに何らかの価値がある(ありそうだ)

※この場合の「価値の大小」については、個人差がある

→ 人によって「情報」か「データ」かが異なる場合がある

# データマイニングとは

膨大なデータから、何らかの役に立ちそうな情報を発見・採掘  
(mining)すること

## ☆ビッグデータの活用

世の中にある、膨大なさまざまなデータを、社会・経済の問題解決  
や業務の効率向上に役立てよう、という考え方。

(ビッグデータ … 数十テラバイト～数ペタバイト  
＝ 単純な半角の文字数にして数十兆から数千兆)

# ビッグデータの活用例

- 膨大な検索語からWebサイトの広告
- 閲覧履歴から「お勧め」を出す(リコメンド)
- SNS等からトレンドを分析し、新商品を開発
- 道路のセンサーから渋滞予測、信号制御
- コンビニエンスストアの売上データから年代別の売れ筋商品を見いだす
- クレジットカードの利用履歴から、不正利用パターンを見つけ犯罪防止に役立てる

# データの収集と整理 (p.192～193)

- オープンデータ
  - 国の行政機関や自治体、研究・教育機関が持つデータ
  - インターネットで広く公開され、容易に入手できる
- データのありかを意識
  - DATA GO JP
  - e-Stat
- 一般的には、分析する前に必要なデータに整理する
  - 必要のある部分のみにデータを削除したり、形式を整えたりする
  - ソフトウェアで活用できるような形式(テキスト・CSV等)に

# 情報分析

☆データに対し、適切な分析方法を理解する  
「量的データ」と「質的データ」

- 数値化されたもの（定量的なデータ）
  - 集計してグラフ化
  - 統計処理
- 数値化されていないもの（定性的なデータ）【次回以降】
  - 関係性や因果関係、順序などを図解
  - 同じような内容ごとや程度にまとめて数値化
  - テキストマイニングなどで数値化、分析

# データの種類 (p.202)

ここに  
注目!

データの種類	尺度	意味	単位	例
質的データ (定性的)	名義尺度	区別しかできない	ない	職業区分、電話番号
	順序尺度	大小比較ができる	ない	優良可の区分、震度
量的データ (定量的)	間隔尺度	差が意味を持つ	ある	気温、偏差値
	比率尺度	比が意味を持つ	ある	長さ、重さ



# データの集計方針(3分)

ワークシートにあるA組とB組それぞれのデータを分析したい。

- 見やすくまとめ、
- 何がわかるかを確かめたい。

どのような方針で行うかを具体的に記入しよう。

# データのグラフ化：見やすくまとめる

- 時系列で示す
  - 折れ線グラフ、有向グラフ など
- 度数分布で表す
  - ヒストグラム、箱ひげ図 など
- 割合で表す
  - 円グラフ、帯グラフ など
- 2つのデータの関連性を表す
  - 散布図 など

# データの統計的分析：数値で示す

- 定量的なデータを、数値(統計量とも言う)を用いて分かりやすく示す
  - 代表的や特徴的な数値を用いる
    - 平均値、中央値、最頻値、最大値、最小値 など
  - 散らばり具合を示す
    - 分散、標準偏差、範囲、四分位偏差 など
  - 偏り具合を示す
    - 尖度(せんど)、歪度(わいど) など
  - 2変数の関係を分かりやすく示す
    - 相関、相関係数、回帰直線
  - 違いを見極める
    - 統計的仮説検定の考え方

# 同じ平均値でも、集団の性質が違う

- 大きい人と小さい人との差が大きいようだ
  - データの「偏り」を客観的に表す必要性
  - 偏りを数値化する必要性

例：A組 最大177. 1 最小153. 2 範囲23. 9

B組 最大180. 3 最小149. 7 範囲30. 6

# データの偏りを表す数字 (数学で扱いましたね)

分散:

- ・それぞれのデータの平均値との差をとり、
- ・その差を二乗し、平均をとったもの

標準偏差:

- ・分散の正の平方根

備考:

偏差値・標準偏差をもとに、平均が50になるように数値化したもの

# 分散と標準偏差

(数学で扱いましたね)

	得点	平均との差	平均との差の2乗
A	67	13	169
B	55	1	1
C	42	-12	144
D	57	3	9
E	49	-5	25
平均	54	0	分散 → 69.6
標準偏差(分散の正の平方根) →			8.342661446

# 分析ツールの活用 ～基本統計量

☆他にも、いろいろな「傾向」を数値で表せる。

中央値（メジアン）

最頻値（モード）

標準偏差

分散

尖度（せんど：ヒストグラムの「とがり具合」）

歪度（わいど：ヒストグラムの「左右対称性」）

範囲（レンジ）

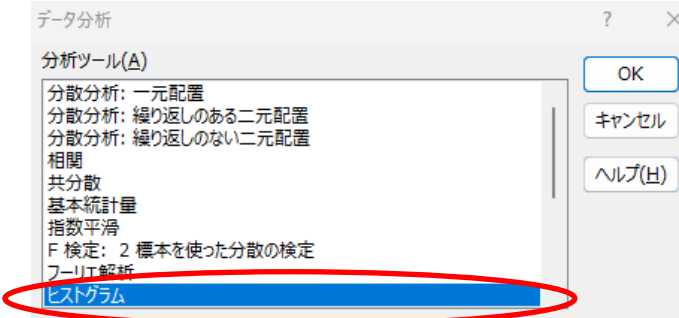
最小

最大

合計

標本数

# 分析ツールの活用 ～ヒストグラム



A組	B組	階級
176.3	177.0	150
157.0	156.3	155
166.0	168.4	160
166.2	166.2	165
175.4	152.1	170
172.1	175.4	175
160.9	170.9	180
164.8	164.8	185
161.4	165.3	
160.3	151.4	
176.4	163.8	
163.8	172.3	
170.1	157.2	
153.2	150.1	
161.5	161.2	
164.7	166.5	
165.3	176.1	
166.1	180.3	
165.2	155.2	
160.4	162.4	
164.7	169.1	
169.1	169.2	
169.1	169.1	
169.2	164.7	
156.5	170.5	
160.0	166.5	
161.2	169.7	
172.9	174.9	
177.1	168.1	
165.1	175.3	
166.1	166.1	
166.5	171.2	
170.4	167.7	
162.7	150.4	
168.4	167.5	
176.5	176.5	
169.7	169.2	
169.2	149.7	
156.8	164.8	
171.2	176.4	

### ヒストグラム

入力元  
入力範囲(I): [ ] ↑

データ区間(B): [ ] ↑

ラベル(L)

出力オプション

出力先(O): [ ] ↑

新規ワークシート(P)

新規ブック(W)

パレート図(A)

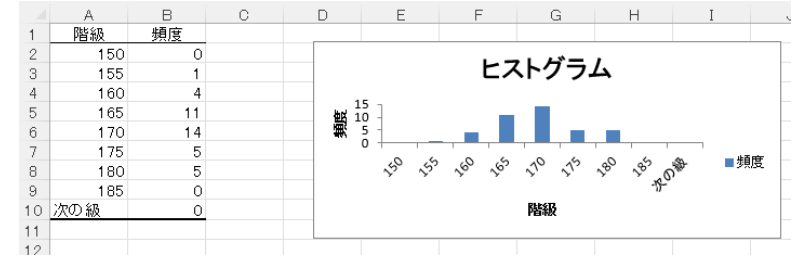
累積度数分布の表示(M)

グラフ作成(C)

※「分析ツールの利用」  
分析ツールは上記のように準備することにより、「データ」タブの中の「分析」  
難しい用語がたくさんでてくるが、今の段階では「基本統計量」と「ヒストグラ

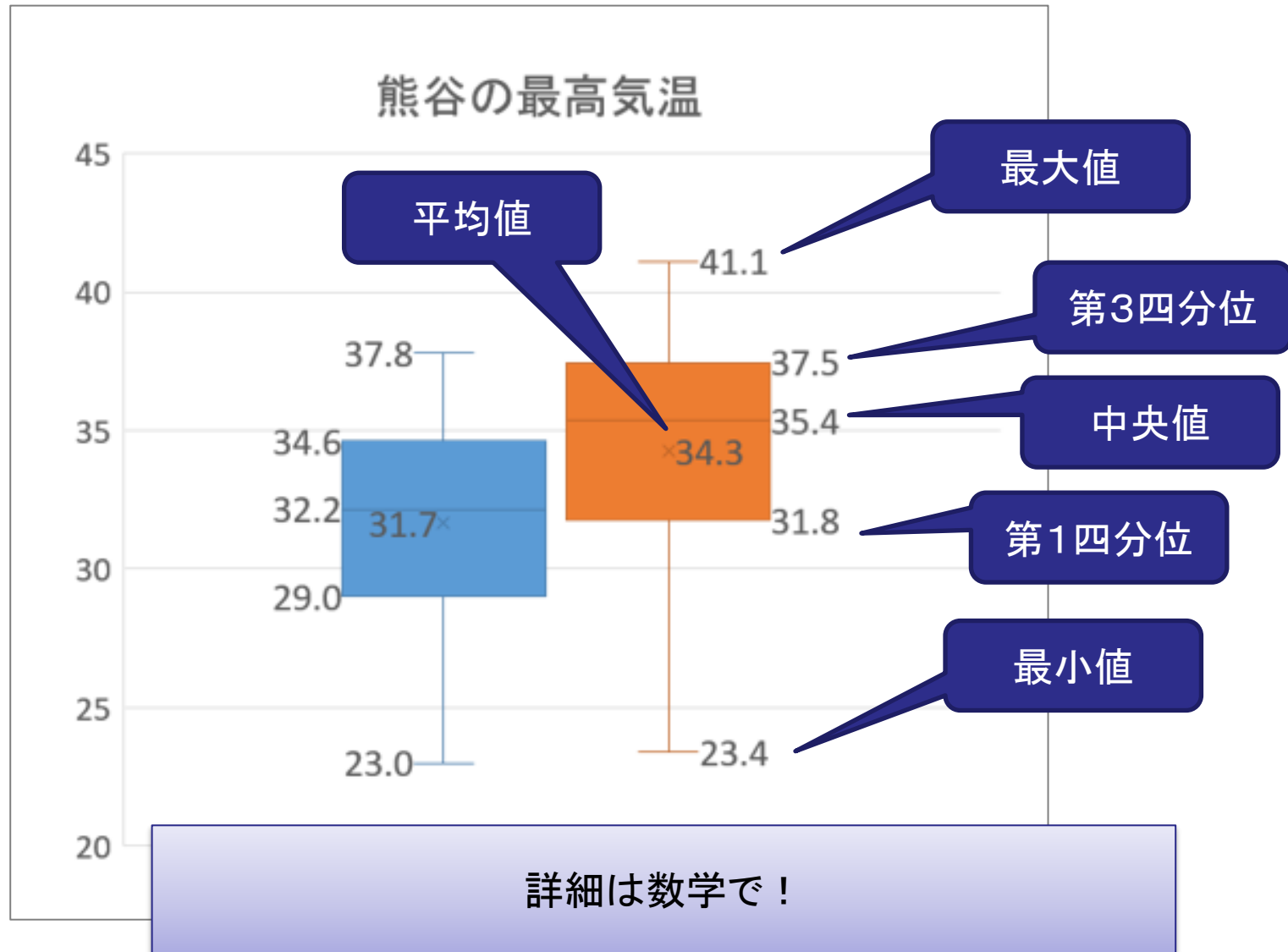
<ヒストグラムの作り方>

- ・「データ」タブから「分析」「データ分析」をクリック
- ・「ヒストグラム」を選択
- ・「入力範囲」には、データとする範囲(A組(B4からB43)またはB組
- ・「データ区間」は、上記「階級」の列(E4からE11)をドラッグして指定
- ・出力オプションは「新規ワークシート」を選び、「グラフ作成」をチェ
- ・新しいワークシートに度数分布表とグラフが作られる





# 箱ひげ図(データを4分割する)



# 演習1

- 2017年と2018年の熊谷の最高気温について、箱ひげ図を描いてみよう。