データマイニングとデータの活用

情報の科学　第22回授業

04情報の蓄積と管理

対応データ　21exp22.xls

今日のテーマ

- 数値で示す
- 数値の「意味」を知る
- わかりやすく示す

（復習）情報分析

☆データに対し、適切な分析方法を理解する

「定量」と「定性」

- 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）
  - テキストマイニングなどで数値化、分析
  - 同じような内容ごとや程度にまとめて数値化
  - 関係性や因果関係、順序などを図解

（復習）定性的なデータを数値化する（1）

- テキストマイニング
  - 大量の文字（＝テキスト）情報を定量（数値）化，分析するための手法
    - アンケート調査
    - TwitterなどのSNS
  - 企業や大学などで広く活用されている
    - マーケティング（売れ筋の商品の分析など）
    - アンケート結果からの心理的な分析など
    - アンケート結果を一目で見るための手法として

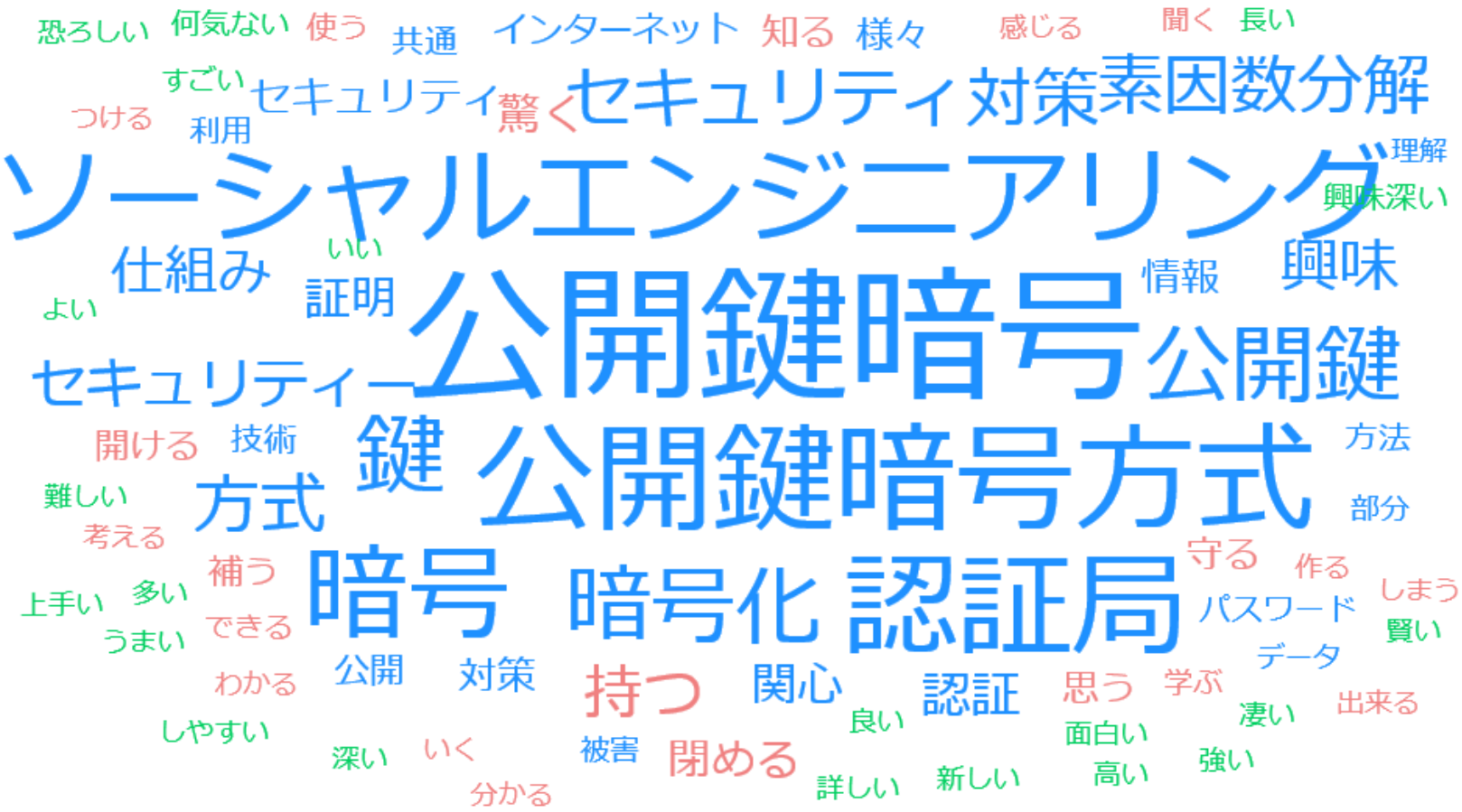
（復習）（例）利用するデータ

- 表計算ソフトで扱えるデータ（.xlsxや.csv）を利用することが多い

C15

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	(2011年のある企業による調査結果より) 実際に被害にあった企業が全体の半分近くいるほど、日本ではソーシャルエンジニアリングが多												
2	「具体的なセキュリティ対策」の部分で、パソコンをすぐに閉じてしまうといつまでも対策プログラムの更新が出来ないということは知												
3	「鍵」の話にに興味を持ちました。インターネットでサイトにアクセスすることはよくあるので、その裏にある仕組みについて知れて面												
4	「公開鍵暗号」 共通鍵暗号の持っている短所を補うものとして効率よく使用できる方法とは何か興味が湧いた。												
5	「認証局」の回で出てきた、共通鍵方式と公開鍵方式のハイブリット方式に興味関心を持った。理由はこの先さらにインターネットが発												
6	1970年代の人の発明がとても賢いと思い、今私たちの暮らしている世界に他にどのような鍵が世界にあったのか気になりました												
7	1つの暗号では不安なので、生体認証、記憶情報、所持情報の二要素認証を行なっている会社があることを知りました。												
8	1番最初に公開鍵暗号方式を思いついて作った人はすごいと思った												
9	2つの暗号方式の長所と短所や、短所を補うための認証局の存在等に興味関心を持ちました。												
10	5回目の暗号化について興味を持ちました。普段何気なく開いているサイトからも情報漏洩の危険があると実感し、ＳＳＬについてしっか												
11	http:// を見たことはあったけれど、意味を知りませんでした。これが鍵だと知り、鍵にもたくさんの種類があるとわかりました。												
12	https://の役割に興味を持ちました。なのでhttps://で始まるかを、意識して確認しいこうと思いました。												
13	RSA公開鍵暗号式のように、情報科学と数学が密接に関わり合っているという事に感心し、他にはどのような事で関わり合っているのか												
14	SSLについて いつも、この鍵のマークはなんだろうと思っていたのでhttpsと鍵のマークで本物のサイトかどうか分かると学んで驚いた。												

（復習）ワードクラウド


恐ろしい
何気ない
使う
共通
インターネット
知る
様々
感じる
聞く
長い
すごい
セキュリティ
驚く
セキュリティ対策
素因数分解
つける
利用
ソーシャルエンジニアリング
理解
興味深い
仕組み
いい
情報
興味
よい
証明
公開鍵暗号
公開鍵
セキュリティー
開ける
技術
鍵
公開鍵暗号方式
方法
難しい
方式
部分
考える
守る
作る
補う
暗号
暗号化
認証局
しまう
上手い
多い
パスワード
賢い
うまい
できる
データ
公開
対策
持つ
関心
認証
思う
学ぶ
わかる
出来る
良い
凄い
しやすい
面白い
深い
いく
被害
閉める
強い
詳しい
新しい
高い
分かる

（復習）単語の出現頻度

名詞	スコア	出現頻度
鍵	298.90	139
興味	133.10	115
暗号	372.59	72
情報	45.18	69
方式	176.10	51
公開鍵暗号	605.71	46
仕組み	108.72	42
対策	47.48	42
公開鍵暗号方式	461.30	36
暗号化	278.39	36
証明	77.63	36
公開	32.58	36
認証局	432.87	34
方法	23.81	34
開心	77.97	33

動詞	スコア	出現頻度
思う	15.45	167
持つ	48.12	136
知る	14.91	79
使う	7.56	58
できる	3.91	56
驚く	23.57	33
開ける	10.74	27
守る	13.44	24
感じる	2.57	22
いく	0.85	21
閉める	25.64	19
考える	1.02	19
聞く	0.88	19
わかる	0.70	19
しまう	0.51	18

形容詞	スコア	出現頻度
面白い	2.55	26
すごい	1.50	26
多い	0.49	13
詳しい	2.11	10
難しい	0.65	9
良い	0.11	9
うまい	0.49	7
興味深い	5.97	6
いい	0.03	6

感動詞	スコア	出現頻度
---	---	---
---	---	---
---	---	---
---	---	---
---	---	---
---	---	---
---	---	---
---	---	---

（復習）単語どうしの係り受けなど

■ 名詞 - ■ 動詞

名詞 - 動詞	スコア	出現頻度
興味 - 持つ	76.69	102
関心 - 持つ	6.35	29
情報 - 守る	7.28	13
鍵 - 持つ（否: 10.00%）	0.80	10（否: 1）
素因数分解 - 使う	1.22	8
鍵 - かける	7.00	7
暗号化 - 持つ	0.41	7
鍵 - 渡す（否: 16.67%）	6.00	6（否: 1）
鍵 - 作る（否: 16.67%）	2.47	6（否: 1）
証明 - できる	0.74	6
暗号 - 使う	0.71	6
部分 - 持つ	0.31	6
鍵 - 使う	0.51	5
方式 - 使う	0.51	5
被害 - 遭う（否: 25.00%）	3.33	4（否: 1）

データの種類

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」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について、

- 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。

分散・標準偏差・正規分布

統計的分析

- 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

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

（数学でもやるのでこの表をよく覚えておくこと）

	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

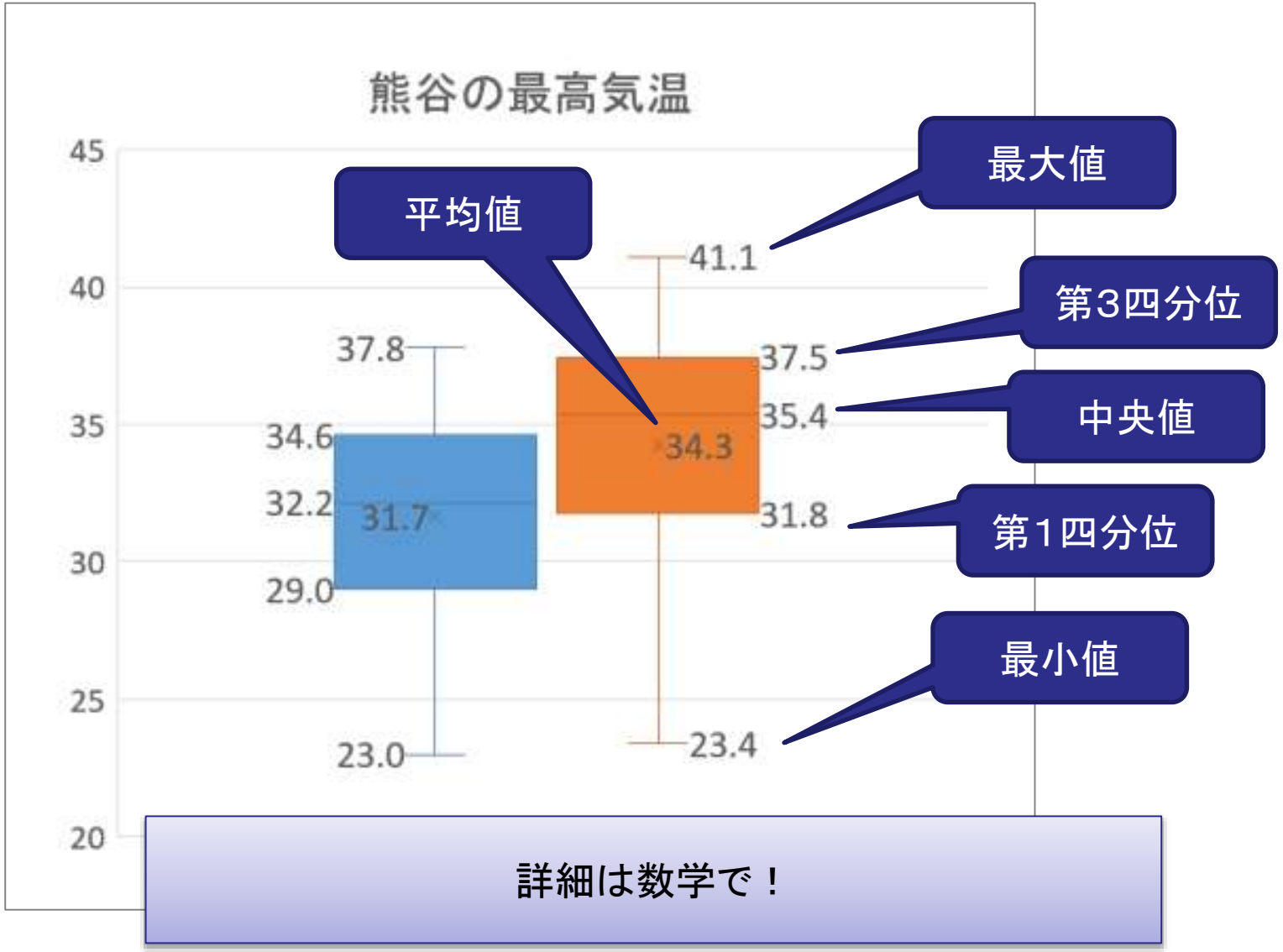
基本統計量

☆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

データのありかを意識

- 「気象庁　最高気温　データ」で検索

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。