2024(令和 6)年度

データサイエンス入門
学習ノート

PART 2

内容

長崎総合科学大学

#08-01

データサイエンス入門 #08
2.1. データを読む

総合情報学部
日當明男

NAS 長崎総合科学大学

#08-02

2.1.1. データの種類(1/1)

種類	尺度水準	概要	例
質的データ	①	分類・区分	性別、血液型、郵便番号
	②	分類・区分での順序、大小関係	震度、ランキング
量的データ	③	間隔に意味のある数値	温度、西暦、知能指数
	④	間隔・比率に意味のある数値、原点あり	身長、速度、収入

尺度水準

① [　　]尺度:観測値に[　　]があり、順序は[　　]

② [　　]尺度:順序に[　　]があるが、間隔は[　　]

③ [　　]尺度:間隔(差)に[　　]があるが、比は[　　]

④ [　　]尺度:観測値やそれらの差や比にも[　　]がある

NAS 長崎総合科学大学 2

#08-03

2.1.2. データの分布と代表値(1/6)

●データ(群)内の一つの共通項目について考える。

例:32車種の燃費項目値[km/L]

データn	車種名	車体価格	・・・	燃費	・・・

8.88	8.88	9.64	9.05	7.90	7.65	6.04	10.31
9.64	8.12	7.52	6.93	7.31	6.42	4.40	4.40
6.21	13.69	12.85	14.33	9.09	6.55	6.42	5.62
8.12	11.54	10.99	12.85	6.68	8.33	6.34	9.05

個々の値ではなく、このようなデータ項目値の群があったとして考える。

NAS 長崎総合科学大学 3

#08-04

2.1.2. データの分布と代表値(2/6)

(a)度数分布とヒストグラム　値(群)の[　　]を把握

まず、項目値の[　　]を含む範囲を
原則として同じ幅の複数の[　　]に分割する。

- [　　]:分割した[　　]
  例:区間「4以上6未満」⇒階級「4～6」
- [　　]:それぞれの[　　]に属する
  データ項目値の[　　]
  例:　階級「4～6」に属するデータ項目値は、
  4.40, 4.40, 5.62の3つ⇒ 階級「4～6」の度数は[　]

#08-05

2.1.2. データの分布と代表値(3/6)

(a)度数分布とヒストグラム

- [　　]:
  [　　]に対する[　　]
  をまとめた表
- [　　]:
  [　　]の[　　]値
- [　　]:
  [　　]に対する
  その階級の[　　]の割合

階級	階級値	度数	相対度数
4～6	5	3	0.09375
6～8	7	12	0.37500
8～10	9	10	0.31250
10～12	11	3	0.09375
12～14	13	3	0.09375
14～16	15	1	0.03125

パーセントで表示することもある

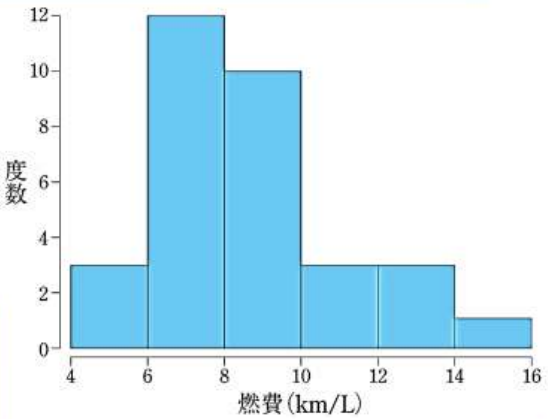
#08-06

2.1.2. データの分布と代表値(4/6)

(a)度数分布とヒストグラム

- [　　]:
  [　　]を横軸に
  [　　]を縦軸にした
  棒グラフ

通常の棒グラフとは異なる
階級区間は連続しているので、
グラフの縦棒は隣と接している



#08-07

2.1.2. データの分布と代表値(5/6)

(b)平均値・中央値・最頻値　代表値:代表の[　　]が重要

- [　　]:分布する値(群)の[　　]に相当する値

例:10個の値(群)「6 2 1 2 4 3 5 5 4 2」の[　　]

$$[\quad\quad] = \frac{\text{値(群)の}[\quad\quad]}{\text{値(群)の}[\quad\quad]}$$

$$= \frac{6+2+1+2+4+3+5+5+4+2}{10}$$

$$= \frac{[\quad\quad]}{10} = [\quad\quad]$$

#08-08

2.1.2. データの分布と代表値(5/6)

(b)平均値・中央値・最頻値　代表値:代表の[　　]が重要

- [　　]:分布する値(群)の[　　]に位置する値

例:10個の値(群)「6 2 1 2 4 3 5 5 4 2」の[　　]

① 値(群)を[　　]に並べる。
　この例では、「1 2 2 2 3 4 4 5 5 6」

② その並びの[　　]に位置する値を見出す。
　値(群)の個数nが[　　]のときには、
　(n/2)番目と(n/2 +1)番目の2つの値の[　　]とする。
　この例では、個数が10個なので、
　5番目と6番目の[　　]:[　　]=3.5

#08-09

2.1.2. データの分布と代表値(6/6)

(b)平均値・中央値・最頻値　代表値:代表の[　　]が重要

- [　　]:分布する値(群)内で[　　]する値

例:10個の値(群)「6 2 1 2 4 3 5 5 4 2」の[　　]

① 値(群)を[　　]に並べる。
【並べると[　　]が分かりやすい】
　この例では、「1 2 2 2 3 4 4 5 5 6」

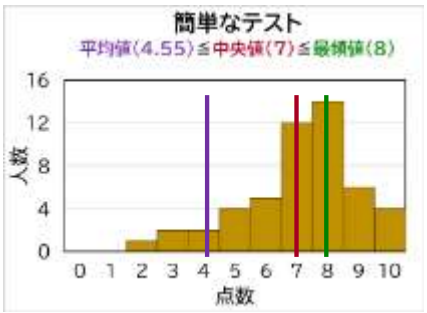
② その並びの中で、[　　]が[　　]値を見出す。
　この例では、値[　]が[　]個で[　　]ので、
　[　　]は[　]

#08-10

2.1.3. 代表値の性質の違い(1/1)

代表値(平均値、中央値、最頻値)が
[　　]ことが多い。

例:テスト(10点満点)の点数分布



#08-11

2.1.4. データのばらつき(1/4)

◎ばらつくデータ⇒[　　]するデータ

- [　　]対象に対する[　　]の観測
  - 例① [　　]生産ラインで製造される[　　]の製品の品質
  - 例② [　　]人がセットする[　　]の髪型
- [　　]対象に対する[　　]テスト
  - 例① [　　]人に対する[　　]測定基準による身長
  - 例② [　　]学生に対する[　　]試験の点数

どの程度？

⇒ ばらつくデータは[　　]の周囲に[　　]。

#08-12

2.1.4. データのばらつき(2/4)

◎バラツキ度合い(平均値からの離れ具合)を測る

データ:$\{a_1, a_2, a_3, \cdots, a_n\}$、$m$:データ$\{a_n\}$の平均値

- 分散var:各データa_iと平均mとの差の2乗$(a_i - m)^2$の平均
- 標準偏差σ:分散の[　　]

$$var = \frac{(a_1-m)^2+(a_2-m)^2+\cdots+(a_n-m)^2}{n}$$

$$\sigma = \sqrt{var}$$

$$var = \frac{1}{n}\sum_{i=1}^{n}(a_i - m)^2$$

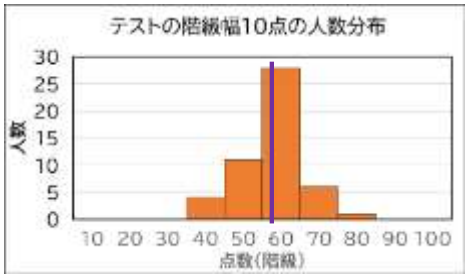
2乗しないで平均すると、[　]になってしまう。

#08-13

2.1.4. データのばらつき(3/4)

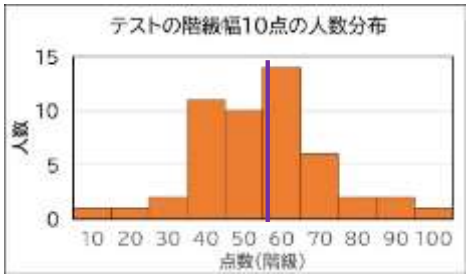
分散・標準偏差が[　　]分布例　分散・標準偏差が[　　]分布例

平均:52.92、標準偏差:8.13



平均の周りに[　　]分布

平均:50.36、標準偏差:17.16



平均の周りに[　　]分布

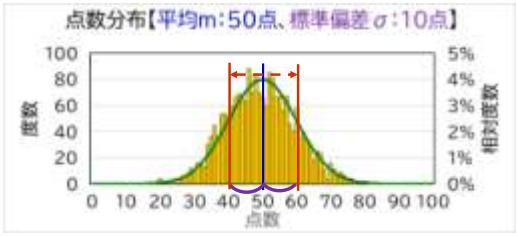
長崎総合科学大学 13

#08-14

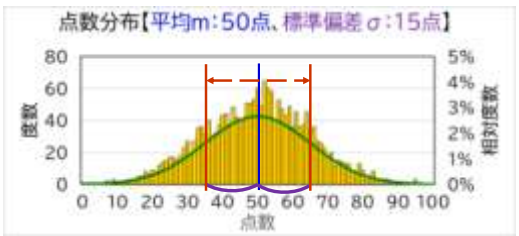
2.1.4. データのばらつき(4/4)

◎平均値の近くのデータ数
分布が、平均値の[　　]と想定される場合。

[　　]にもない



[m−σ, m+σ]=[40,60]内に、68.9%



[m−σ, m+σ]=[35,65]内に、68.8%

[m−σ, m+σ]内に、[　　]のデータが存在すると期待できる

長崎総合科学大学 14

#08-15

2.1.5. 観測データに含まれる誤差の扱い(1/1)

★測定値には、必ず[　　]が含まれる。

観測値＝[　　]＋[　　]

- 機械的誤差:[　　]に起因する誤差
  機械の[　　]や周囲の[　　]による影響
- 人為的誤差:[　　]に起因する誤差
  人間の[　　]による影響
- バイアス(系統)誤差:[　　]に生じる一定量の誤差
  機械の[　　]や測定者の[　　]などによる影響

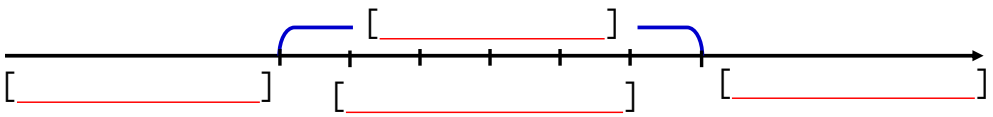
長崎総合科学大学 15

#08-16

2.1.6. 打ち切りや・・・必要なデータ(1/1)

◎打ち切りや脱落を含むデータ

- 打ち切り:事前設定の[＿＿＿＿＿＿]に入らない場合



- 脱落:[＿＿＿＿＿＿＿＿＿＿]における観測値の[＿＿＿]

◎層別の必要なデータ:[＿＿＿＿＿＿＿＿]による分類

データの項目値の分布が[＿＿＿]な場合に有効かも

分類項目が[＿＿＿＿]の場合もあり得る。

#08-17

2.1.7. 相関と因果性(1/3)

- 相関係数r_{xy} :2項目(x,y)間の[＿＿＿＿]な増減関係の尺度

$$r_{xy} = \left\{\left(\frac{x_i - m_x}{\sigma_x}\right) \times \left(\frac{y_i - m_y}{\sigma_y}\right)\right\}_{i=1}^{n} \text{ の[＿＿＿＿]}$$

ここで、m_x : $\{x_i\}_{i=1}^{n}$の平均、σ_x : $\{x_i\}_{i=1}^{n}$の標準偏差

m_y : $\{y_i\}_{i=1}^{n}$の平均、σ_y : $\{y_i\}_{i=1}^{n}$の標準偏差

- 因果性

一つのデータ内で、ある項目の値の[＿＿＿]が、

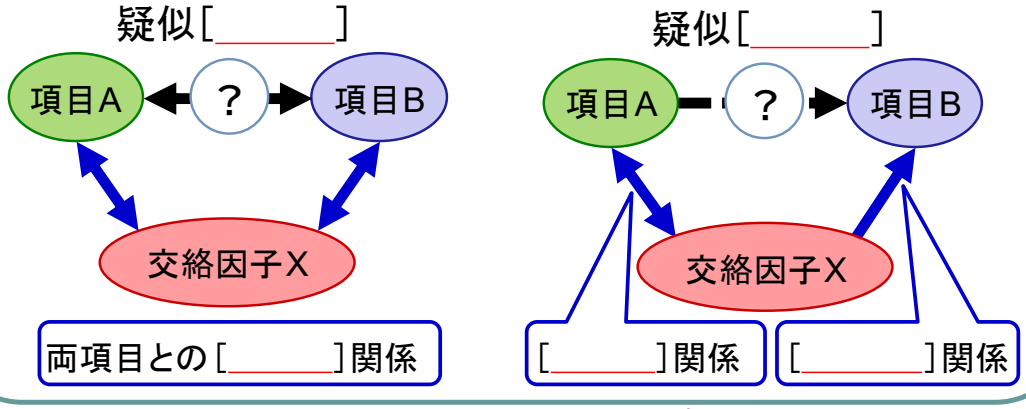
別の項目の値に[＿＿＿＿＿＿＿＿]特性

例: 世帯の[＿＿＿]と[＿＿＿＿]

#08-18

2.1.7. 相関と因果性(2/3)

- 交絡因子:疑似相関や疑似因果の[＿＿＿＿＿＿＿＿]



#08-19

2.1.7. 相関と因果性(3/3)

◎バークレイデータ
性別と大学院合格率に[　　]関係は？

層別にするから[　　]が分かる

学部	合計		男子学生		女子学生	
	志願者	合格率	志願者	合格率	志願者	合格率
全体	12763	41%	8442	44%	4321	35%
A学部	933	64%	825	62%	108	82%
B学部	585	63%	560	63%	25	68%
C学部	918	35%	325	37%	593	34%
D学部	792	34%	417	33%	375	35%
E学部	584	25%	191	28%	393	24%
F学部	714	6%	373	6%	341	7%
:	:	:	:	:	:	:

女子学生の合格率が[　　]

合格率が[　　]学部に、志願者が[　　]

NAS 長崎総合科学大学 19

#08-20

2.1.8. 母集団と標本抽出(1/3)

- 母集団:[　　]全体
  例:国民全体、居住者全体、男性、女性、小学X年生、・・・
- 全数調査:[　　]に対する[　　]調査
  例:国勢調査(居住者全体)
- 標本(サンプル):母集団から[　　]一部
- 標本調査:[　　]に対する[　　]調査
  目的:母集団の[　　]を[　　]に[　　]する
- 標本誤差:[　　]に伴って生じる誤差
  対応:標本誤差の[　　]を図る・・・

どうやって？

NAS 長崎総合科学大学 20

#08-21

2.1.8. 母集団と標本抽出(2/3)

無作為でも、標本には[　　]の[　　]がある

◎標本誤差最小化へ向けて
[　　]と[　　]の検討が必要

- 標本抽出法:[　　]が主流
  - 単純無作為抽出法
  - 層別(層化)抽出:母集団を[　　]
  - 多段抽出法:調査区域も[　　]
- 標本数:[　　]を考慮

作為ある抽出では、標本に[　　]が生じ、標本誤差の[　　]を招く可能性が高い

[　　]などは無作為調査ではない

NAS 長崎総合科学大学 21

#08-22

2.1.8. 母集団と標本抽出(3/3)

- 調査目的による使い分け
  - [＿＿＿＿＿＿＿＿]の対象者の意向を知りたい
    調査方法:[＿＿＿＿＿＿＿＿＿＿]に対する調査
    ⇒[＿＿＿＿＿＿＿＿]によって
    [＿＿＿＿＿＿＿＿]を推定
  - [＿＿＿＿＿＿＿＿]の対象者の意向を知りたい
    調査方法:[＿＿＿＿＿＿＿＿]、[＿＿＿＿＿＿＿＿]、
    [＿＿＿＿＿＿]など
    【前提条件】:調査方法に[＿＿＿＿]を持たない

#08-23

2.1.9. クロス集計表・・・散布図行列(1/2)

(a)クロス集計表
2つ以上の項目(変数)間の[＿＿＿]をある項目の[＿＿＿]に整理した表

表2.1.5
タイタニック号乗船者に関するクロス集計表

等級	性別	死亡	生存
1等	男	118	62
	女	4	141
2等	男	154	25
	女	13	93
3等	男	422	88
	女	106	90
乗員	男	670	192
	女	3	20

#08-24

2.1.9. クロス集計表・・・散布図行列(2/2)

(b)相関係数行列・散布図行列
複数項目(変数)間の[＿＿＿＿]を、
2項目の[＿＿＿＿＿＿]の相関係数や散布図の行列

◎相関関係数行列

	項目1	項目2	・・・	項目n
項目1	<1,1>	<2,1>	・・・	<n,1>
項目2	<1,n>	<2,2>	・・・	<n,2>
:	:	:		:
項目n	<1,n>	<2,n>	・・・	<n,n>

<a,b>:項目aと項目bの相関係数

◎散布図行列

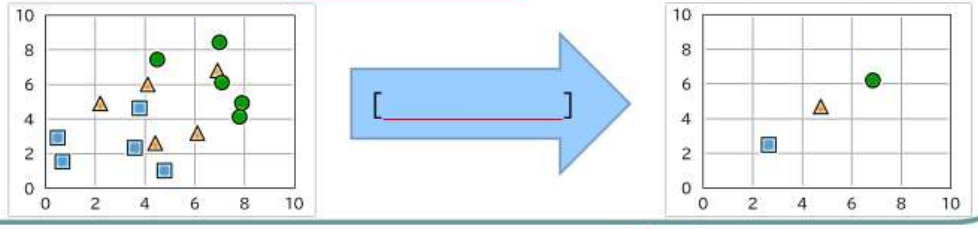
項目1	(2,1)	・・・	(n,1)
(1,n)	項目2	・・・	(n,2)
:	:		:
(1,n)	(2,n)	・・・	項目n

(a,b):項目a[横軸]と項目b[縦軸]の散布図

#08-25

2.1.10. 統計情報の正しい理解(1/1)

- [＿＿＿＿]に注意
  例① A店が20%引き、B店が10%引き。どちらが得か？
  例② 今期は昨期の50%減。来期に今期の50%増で元通り？
- [＿＿＿＿＿＿]的相関
  散布図において、[＿＿＿＿＿＿]を用いると、印象が変わる



NAS 長崎総合科学大学 25

#09-01

データサイエンス入門 #09
2.2. データを説明する

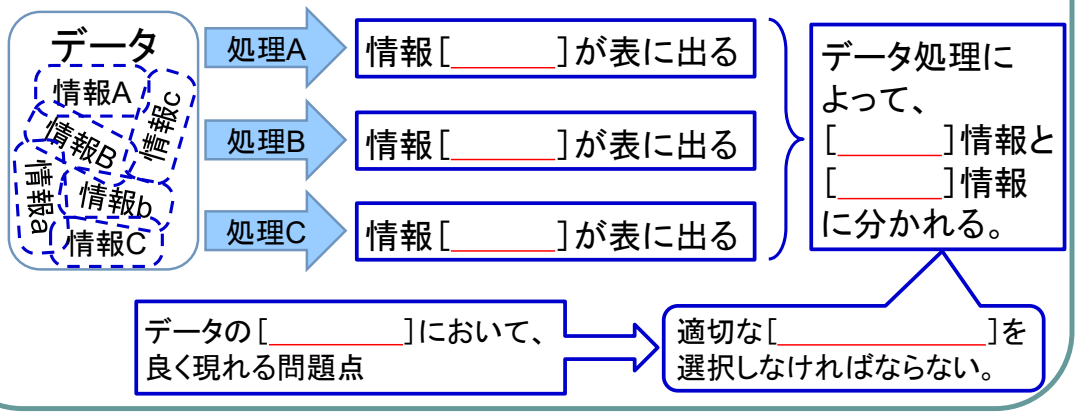
総合情報学部
日當明男

長崎総合科学大学

#09-02

2.2.0. データが持つ情報について(1/1)

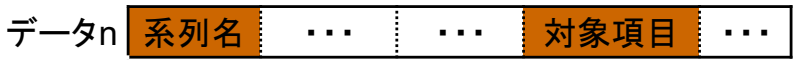
◎ 解釈(処理)される前のデータは、[　　]。



長崎総合科学大学 2

#09-03

2.2.1. データの表現(1/7)



◎データ表現(説明)の視点 ← 適切な[　　]を選ぶため

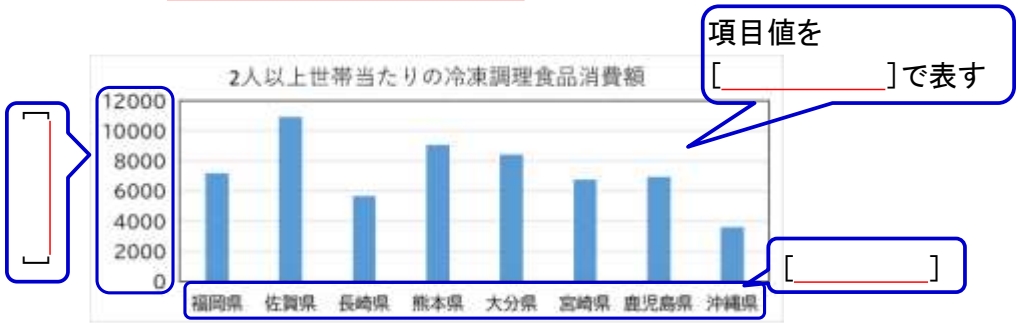
- 目的
  - [　　]
    - [　　]傾向
    - [　　]現象(異常値、外れ値等)
  - [　　]
- [　　]
  - [　　]データ
  - [　　]データ

長崎総合科学大学 3

#09-04

2.2.1. データの表現(2/7)

◎棒グラフ:[　　](質的データ)における
[　　](量的データ)を図示する。



データ出典:SSDSE-C-2021

#09-05

2.2.1. データの表現(3/7)

◎帯グラフ・円グラフ:[　　](質的データ)における
[　　]の[　　]に対する
[　　] (量的データ)を図示する。

データ出典:SSDSE-B-2021

#09-06

2.2.1. データの表現(4/7)

◎柱状グラフ:2つの[　　](質的データ)の組における
[　　] (量的データ)を図示する。

[　　]の[　　]が分かりにくい

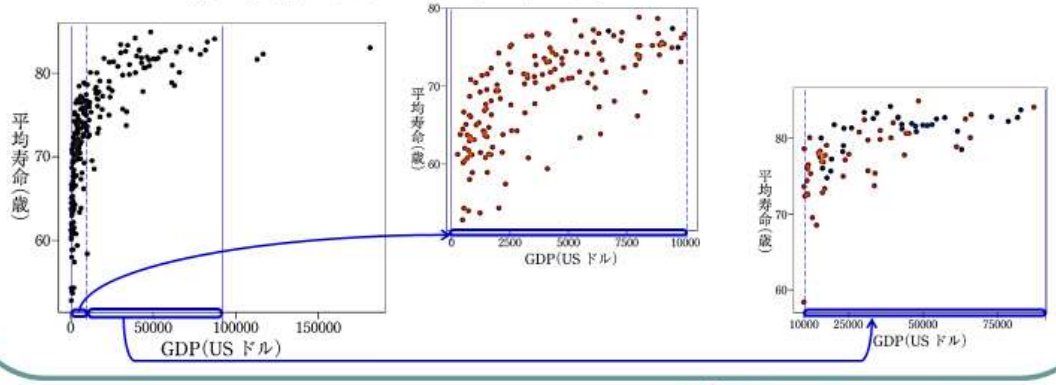
[　　]:
2つ目の系列を[　　]表示

データ出典:SSDSE-D-2021

#09-07

2.2.1. データの表現(5/7)

◎散布図:2つの量的データの組を[　　]と見立てて、縦・横軸平面に配置する。

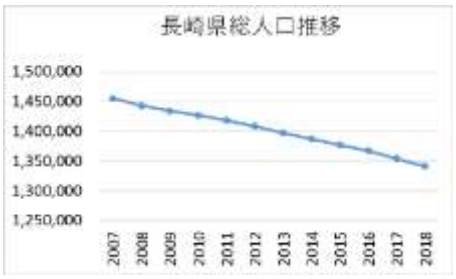


データ出典:SSDSE-D-2021

#09-08

2.2.1. データの表現(6/7)

◎折れ線グラフ:[　　]データの可視化に最適。
対象項目の変化度合いに応じて、[　　]を調整。



データ出典:SSDSE-B-2021

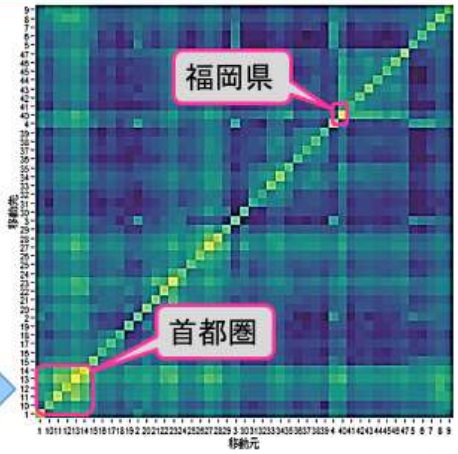
データ出典:Kabutan 時系列データ

#09-09

2.2.1. データの表現(7/7)

◎ヒートマップ:
[　　]の項目の量的・質的データの組に対する別項目の量的データの[　　]を、組に対応する[　　]の[　　]で表現する。

図2.2.11
各都道府県の人口移動数
(明るい色ほど、移動が多い)



#09-10

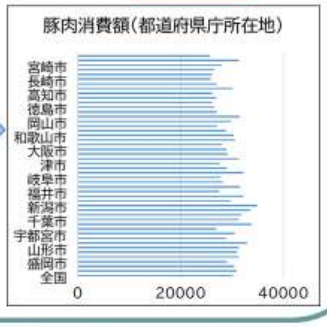
2.2.2. データの図解表現(1/2)

◎グラフ以外の可視化法

- チャート:グラフよりも[＿＿＿＿]があり[＿＿＿＿]な手法。
  [＿＿＿＿]にもできるが、データに対する[＿＿＿]を
  与える[＿＿＿＿]もある。

SSDSE-C
都道府県庁別
豚肉支出額
(≠特化係数)

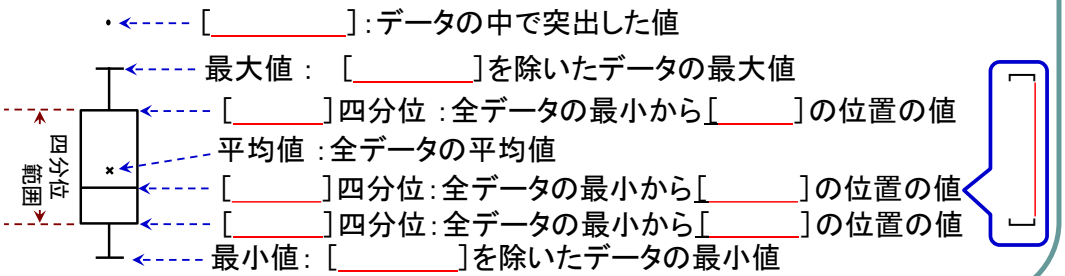
図2.2.12
豚肉消費動向



#09-11

2.2.2. データの図解表現(2/2)

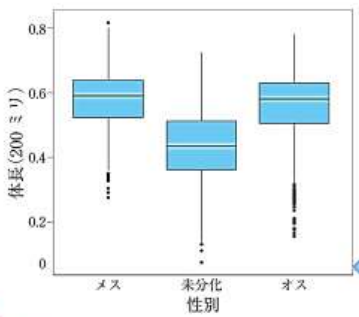
- 箱ひげ図:データの分布を[＿＿＿＿]に表す手法の一つ。
  データ(項目値)の[＿＿＿＿＿]や[＿＿＿]を
  [＿＿＿＿＿]で表す。



#09-12

2.2.3. データの比較(1/2)

- 箱ひげ図を用いた比較
  同じ[＿＿＿]を持つ異なる[＿＿＿＿＿]の分布の比較



サンプル
データ

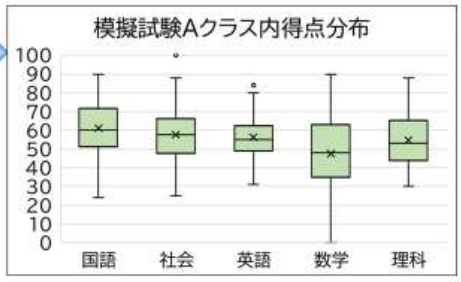
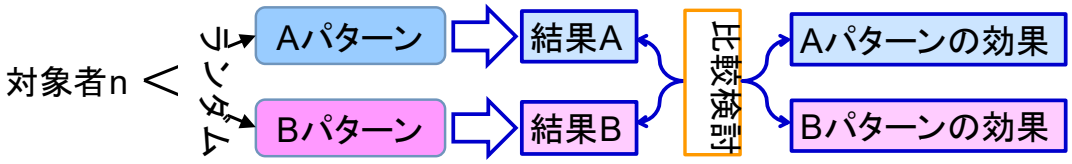


図2.2.14
アワビの体長

#09-13

2.2.3. データの比較(2/2)

- A/Bテスト:措置(A,B)の[＿＿＿＿]を測る手法。
  AパターンとBパターンとを[＿＿＿＿＿＿]に与えて
  その[＿＿＿]を比較して、提供パターンの[＿＿＿]を測る。
  [＿＿＿＿＿＿]にも活用される。



#09-14

2.2.4. 不適切なグラフ表現(1/1)

◎不適切なグラフ表現(可視化)は、[＿＿＿＿]を与える。

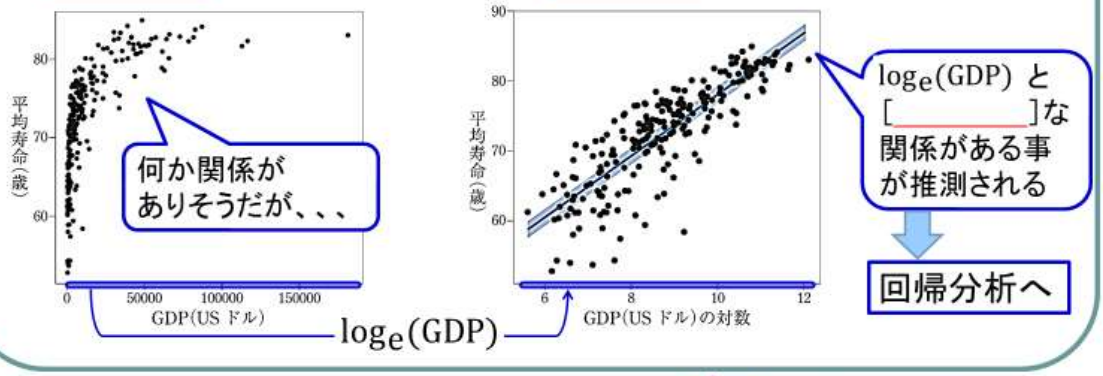
◎特に注意すべき事柄

- [＿＿＿＿＿＿]と[＿＿＿＿＿＿]
  表したい[＿＿＿]にふさわしい[＿＿＿＿＿＿]か？
  誤解を招くような余計な[＿＿＿＿＿＿]を与えないか？
- [＿＿＿＿]手法(グラフ種等)の選択
  適切な[＿＿＿＿]手法か？　[＿＿＿]を与えないか？
- 度数分布・ヒストグラムにおける[＿＿＿＿]
  分布がイメージできる[＿＿＿＿]か？

#09-15

2.2.5. 優れた可視化の例(1/1)

◎可視化手法とデータ処理の[＿＿＿]によって、
[＿＿＿＿]にくかった傾向が[＿＿＿＿]やすくなる事がある。



#10-01

データサイエンス入門 #10
2.3. データを扱う

総合情報学部
日當明男

長崎総合科学大学

#10-02

2.3.1. 表形式のデータ(1/1)

◎CSV形式

[＿＿＿＿＿＿]とも呼ぶ

1行で[＿＿＿＿＿＿]のデータを表し、
[＿＿＿＿＿＿＿＿]を「,(カンマ)」で区切って表現する方法。

都道府県	総人口	年平均気温	降水量	消費支出
青森県	1308265	11.5	1003.5	226214
岩手県	1279594	11.6	1094	286439
宮城県	2333899	13.7	1444.5	267661
秋田県	1023119	12.7	1490.5	249778
山形県	1123891	12.7	1027	308953
福島県	1914039	14.2	1284	321185

CSV形式

図 2.3.1

宮城県,2333899,13.7,1444.5,267661

長崎総合科学大学 2

#10-03

2.3.2. データ解析ツール(1/2)

- Excel:マイクロソフト社の表計算ソフトウェア。
  CSVファイルも[＿＿＿＿＿＿＿＿]できる。

列番号
(「列」は縦方向)

行番号
(「行」は横方向)

	A	B	C	D	E	F
1	都道府県	総人口	年平均気温	降水量	消費支出	
2	青森県	1308265	11.5	1003.5	226214	
3	岩手県	1279594	11.6	1094	286439	
4	宮城県	2333899	13.7	1444.5	267661	
5	秋田県	1023119	12.7	1490.5	249778	
6	山形県	1123891	12.7	1027	308953	
7	福島県	1914039	14.2	1284	321185	
8						
9						

[＿＿＿＿]名

[＿＿＿＿＿＿]

各[＿＿＿＿]
の[＿＿＿＿]

F8セル
(「列」「行」の順)

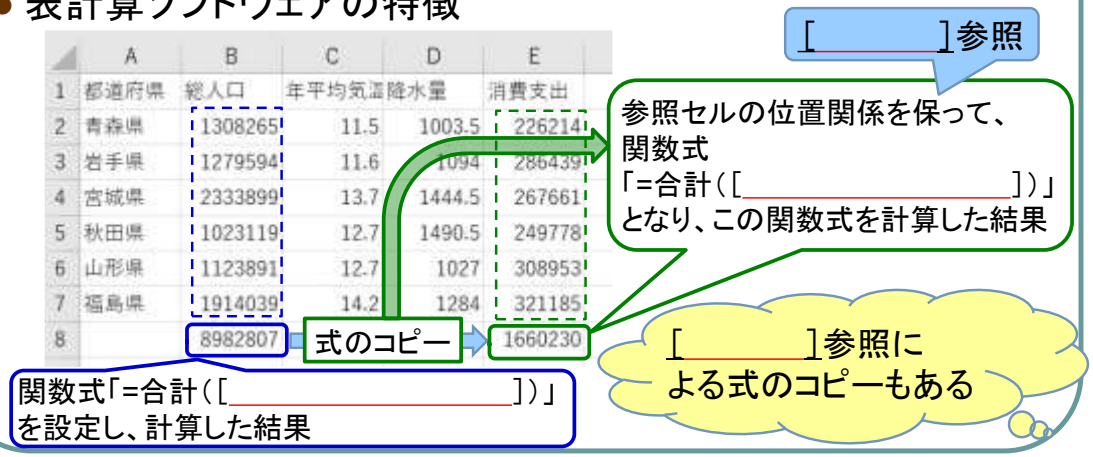
[＿＿＿＿]

長崎総合科学大学 3

#10-04

2.3.2. データ解析ツール(2/2)

- 表計算ソフトウェアの特徴



	A	B	C	D	E
1	都道府県	総人口	年平均気温	降水量	消費支出
2	青森県	1308265	11.5	1003.5	2262141
3	岩手県	1279594	11.6	1094	2864391
4	宮城県	2333899	13.7	1444.5	2676611
5	秋田県	1023119	12.7	1490.5	2497781
6	山形県	1123891	12.7	1027	3089531
7	福島県	1914039	14.2	1284	3211851
8		8982807			1660230

#10-05

2.3.3. SSDSEデータを使う(1/5)

◎SSDSE:DS教育用の汎用教材

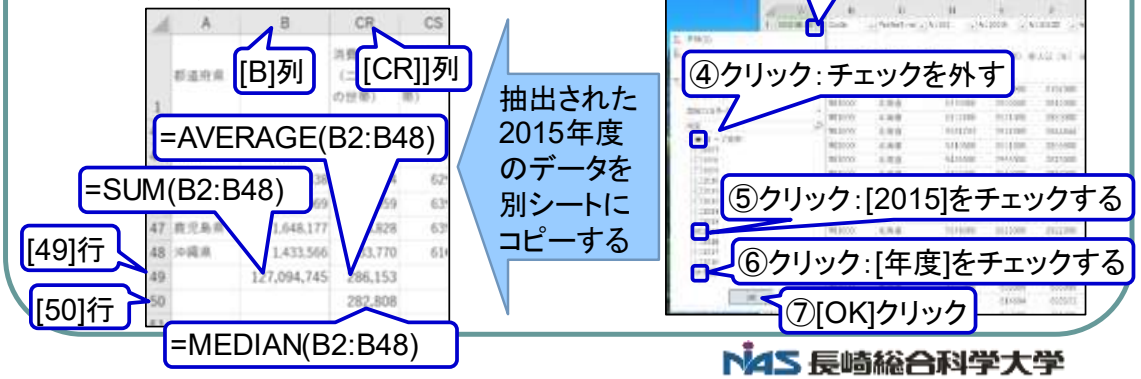
https://www.nstac.go.jp/SSDSE/

- SSDSE-A:全国市区町村のすがた(社会・人口統計体系)
- SSDSE-B:全国都道府県のすがたの推移
- SSDSE-C:全国都道府県庁所在地の家計消費データ
- SSDSE-D:都道府県別の自由時間活動・生活時間データ

#10-06

2.3.3. SSDSEデータを使う(2/5)

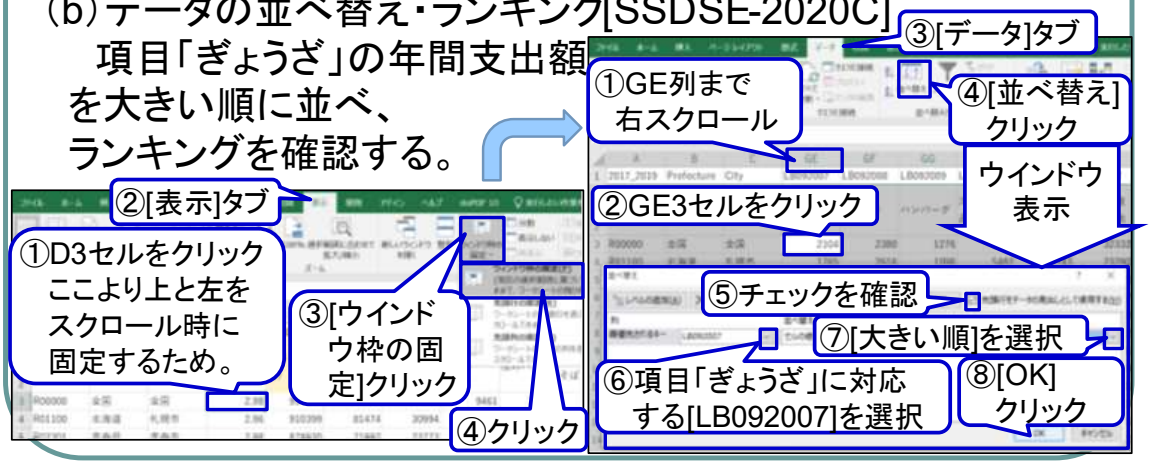
(a)データの集計[SSDSE-2020B]
　2015年度の全都道府県の人口の合計と消費支出の平均を求める。



#10-07

2.3.3. SSDSEデータを使う(3/5)

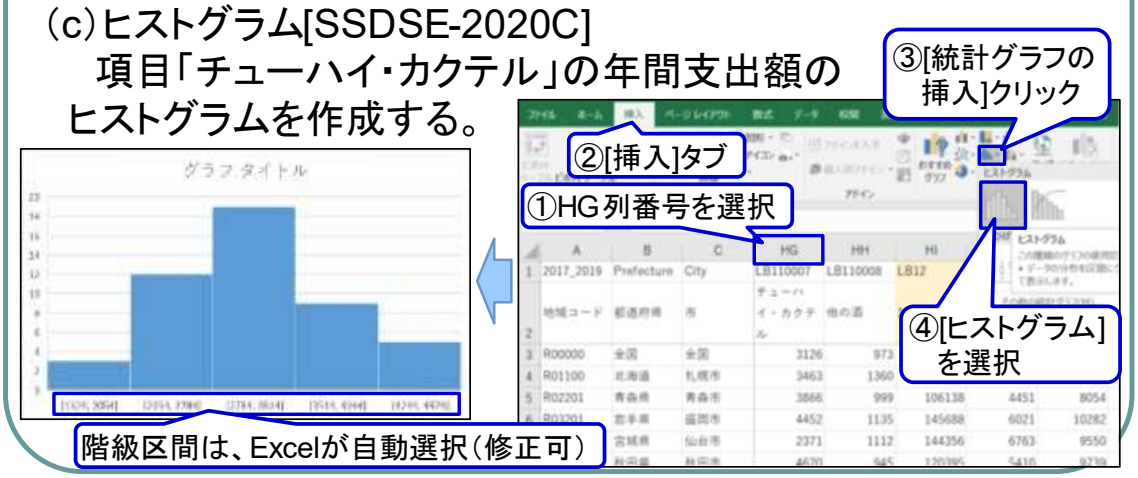
(b)データの並べ替え・ランキング[SSDSE-2020C]
項目「ぎょうざ」の年間支出額を大きい順に並べ、ランキングを確認する。



#10-08

2.3.3. SSDSEデータを使う(4/5)

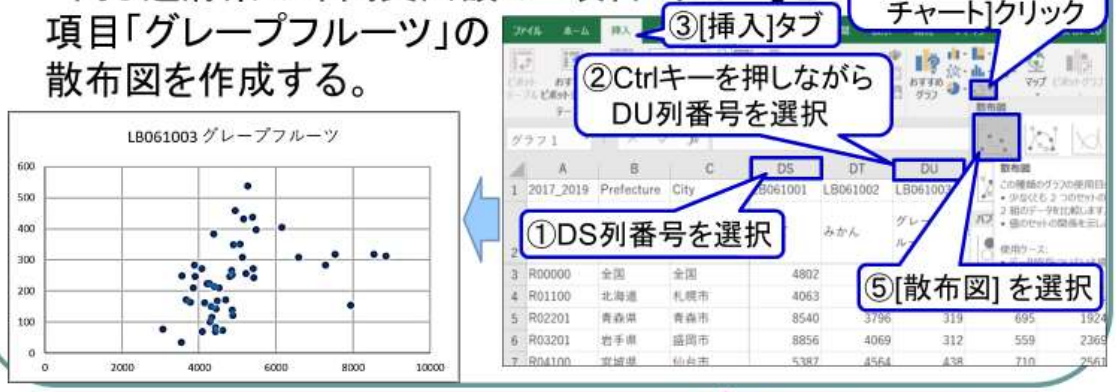
(c)ヒストグラム[SSDSE-2020C]
項目「チューハイ・カクテル」の年間支出額のヒストグラムを作成する。



#10-09

2.3.3. SSDSEデータを使う(5/5)

(d)散布図[SSDSE-2020C]
同じ道府県の年間支出額の　項目「りんご」と項目「グレープフルーツ」の散布図を作成する。



#11-01

データサイエンス入門 #11
3.1. データ・AIを扱う上での留意事項

総合情報学部
日當明男

長崎総合科学大学

#11-02

3.1.1. ELSI（1/1）

◎ELSI（Ethics, Legal and Social Issues）

[　　]的、[　　]的、[　　]的影響
：すべての技術分野で考慮すべき事

- データサイエンス倫理
  ：データ処理及び活用に関わる倫理
- AI倫理
  ：AIの構築（学習）やその活用に関わる倫理
  例：顔認証システムによる誤認識等
- AIサービスの責任論：AIの責任は[　　]？

長崎総合科学大学 2

#11-03

3.1.2. 一般データ保護規則（1/1）

◎GDPR（General Data Protection Regulation）

EU域内の国民の個人データの保護と利用の規則。
他国にも要請して、[　　]で守る。
利用は原則として個人が[　　]範囲内

- GDPRでできる事

①自分の個人情報を最小限のコストで[　　]
②自分の個人情報の[　　]
③機械的生成プロファイルに基づいた判断に[　　]

長崎総合科学大学 3

#11-04

3.1.3. 十分性認定(1/1)

◎十分性認定:

[＿＿＿＿＿＿＿＿＿＿]の個人情報保護環境の整備の認定。

- 十分性認定されない国への、EU域内国民の個人情報の[＿＿＿＿＿＿＿＿＿＿]。
- 十分性認定は、[＿＿＿＿＿＿]もある。
- 日本は2019年に認定。ただ、国内に関連法律が乱立しており、個人情報の[＿＿＿＿＿]を阻害

#11-05

3.1.4. AI倫理(1/1)

◎「AI倫理」についての動き

- アシロマAI原則:現代的なAI倫理の[＿＿＿＿＿]。全23項目。
  研究課題(5)、倫理と価値(13)、長期的な課題(5)
  https://futureoflife.org/2017/08/01/ai-principles-japanese/
- IEEE倫理デザイン(IEEE EAD)第1版、第2版:アメリカ
- 信頼できるAIのガイドライン(EU AI倫理):EU
- 人間中心のAI社会原則:日本

ポイント: AIは人間の[＿＿＿]、[＿＿＿＿＿＿]は人間が行う

#11-06

3.1.5. AI脅威論(1/1)

◎いつの日か、AIが人間の能力を超える(特異点)。
⇒人間に敵対する脅威になりかねない。

現在の見解

アシロマ原則

- 項目10:AIの目標と行動は、[＿＿＿＿]に人間と[＿＿＿]
- 項目19:未来のAIの可能性に[＿＿＿]はない
- 項目22:急激な進展や自己複製AIは、厳重な[＿＿＿＿＿]

「特異点」はまだ先。敵対するか疑問視。⇒ AIへの関心:AI脅威論＜[＿＿＿＿]

#11-07

3.1.6. ブラックボックス化（1/1）

◎ビッグデータの分析結果は何とか理解できても、
AIの［　　］は理解できない。

説明されても、理解できるか疑問

AIの［　　］化

道具として、果たして
［　　］なるか？

活用するには、、、

［　　］
が必要

［　　］ ⇒［　　］
⇒［　　］⇒トラスト

NAS 長崎総合科学大学 7

#11-08

3.1.7. 説明可能性（1/3）

◎現時点で、説明可能性を持つAIは［　　］。

XAI（eXplainable AI）

説明されても［　　］？

人間が介在して、説明可能性＋［　　］を高める。

◎説明可能性
・遡及型：結果が［　　］に説明できる事。
・予見型：［　　］の入力で、変な結果にならない事。
［　　］の段階からこの仕組みを組み込む。

その効用（?）と、説明すべき項目は何？

NAS 長崎総合科学大学 8

#11-09

3.1.7. 説明可能性（2/3）

◎人間が介在することの効用（?）
GDPR22条第1項「機械的生成・・・」の［　　］。

◎説明すべき項目（IEEE EAD ver2で規定）
① 開発・運用企業・そこへの出資者。
学習データの［　　］
② 学習における教師データと実運用時のデータ
③ AIシステムの［　　］

どういう方策があるか。

この説明が最も難しく、
理解してもらうべき箇所

NAS 長崎総合科学大学 9

#11-10

3.1.7. 説明可能性(3/3)

◎2020年時点で、最も現実的な方策:[　　]の活用

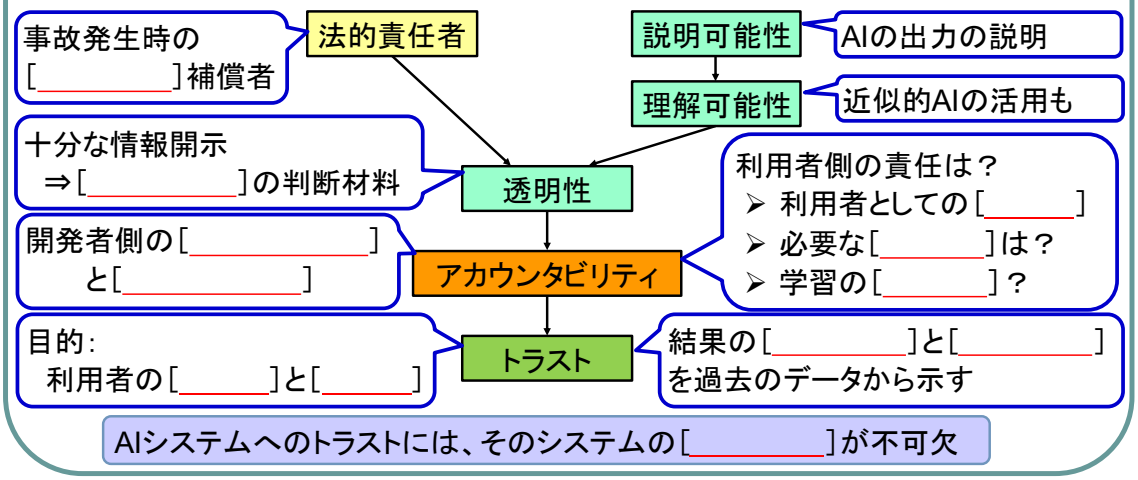
[　　]:元のAI入出力関係を近似的に再現するAI

- if-thenルールや分類木を用いて設計
  ⇒[　　]の向上
- [　　]の出力を説明する

◎カバーできない時の対応
1) 類似した結果で代用する
2) 類似度順に複数の結果(内挿、外挿、平均)を提示
3) 狭い範囲に限定して、分類境界線を線形近似

#11-11

3.1.8. アカウンタビリティ、透明性、トラスト(1/1)



#11-12

3.1.9. 公平性(1/5)

◎AIシステムやデータ処理における公平性

a) 設定した[　　]に関係する項目は平等に扱う
b) 扱う項目に対する[　　]を削除する

- a),b)の説明が理解可能
  この説明による[　　]が公平性につながる
- a),b)の説明が理解不可能
  システム設計者や運用者への[　　]や補償の明記等の[　　]の確保が公平性につながる

#11-13

3.1.9. 公平性(2/5)

◎公平と平等

- 公平
  メンバーの能力や状況に応じて、[　　]に対応を変える。
  能力や状況の評価の[　　]性？
  例：競技会の[　　]は公平。
  [　　]人は報われる。
- 平等
  すべてのメンバーに対して、[　　]に対応する。
  例：競技会の[　　]は平等。
  平等が[　　]を生むことがある。

#11-14

3.1.9. 公平性(3/5)

◎公平性の確保

正当性の維持には、[　　]の解消が不可欠

例えば

アファーマティブアクション

:[　　]を解消する[　　]を設ける

例：女性の定員枠を設ける。
少数民族の代表枠を設ける

[　　]が目的になるかも → 新たな[　　]を生む

アファーマティブアクションは、
[　　]な措置で、
[　　]な解決策ではないかも

#11-15

3.1.9. 公平性(4/5)

◎データ処理における公平性の確保

各種バイアスと不正なデータ操作の[　　]が必要

- バイアス：[　　]
  - データバイアス：無関係な特徴に基づく[　　]
  - アルゴリズムバイアス：点数の[　　]
- 不正なデータ操作：[　　]＋[　　]
  - データ捏造(ねつぞう)：存在しないデータを[　　]
  - データの改竄(かいざん)：都合の良いようにデータを[　　]する
  - データの盗用(とうよう)：他人のデータを[　　]で使用する

#11-16

3.1.9. 公平性(5/5)

◎公平性の維持手法は目的依存

目的	手法
[　　]の平等	アファーマティブアクション
[　　]の平等	データバイアスの削除
[　　]の平等	アルゴリズムバイアスの削除

◎AIシステムにおける公平性

[　　]化したAIの公平性の確保は困難

学習データの[　　]、[　　]なデータで、
公平性は簡単に[　　]。

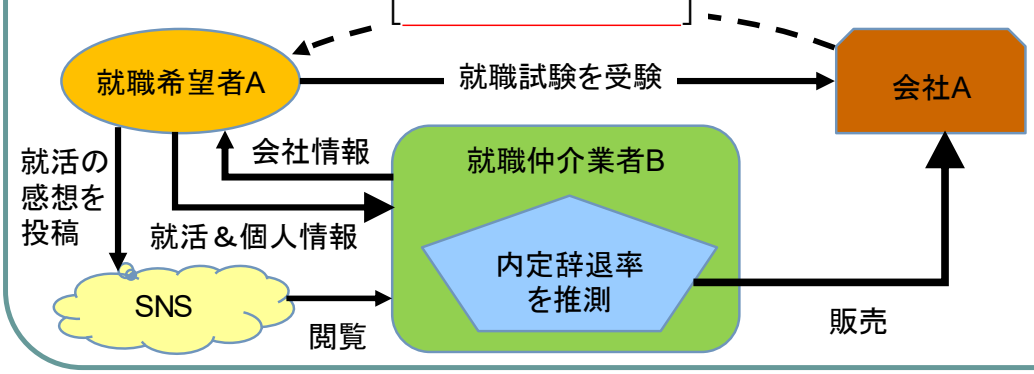
NAS 長崎総合科学大学 16

#11-17

3.1.10. ーデータの悪用・目的外使用ー(1/1)

◎事業者のデータ誤用は、利用者にとってはデータ悪用。

◎2019年の事例

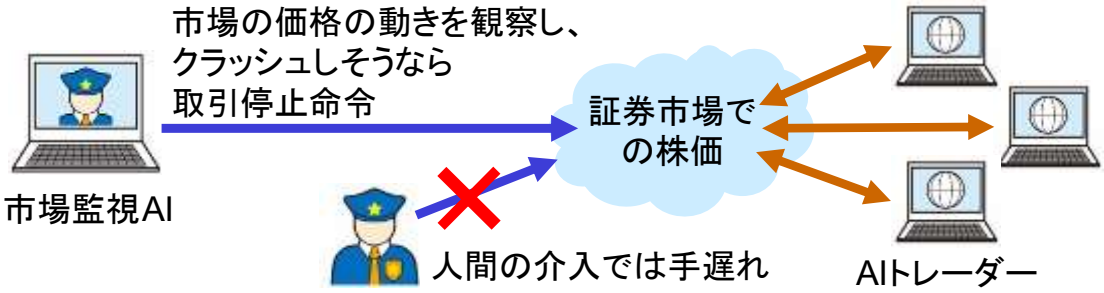


NAS 長崎総合科学大学 17

#11-18

3.1.11. ーフラッシュクラッシューー(1/1)

◎マイクロ秒単位のAIによる株の売買
連鎖反応によって、瞬時に株の乱高下が起こる。



複数のAIが[　　]社会では、このような状況も想定すべきかも。

NAS 長崎総合科学大学 18

#12-01

データサイエンス入門 #12
3.2. データを守る上での留意事項

総合情報学部
日當明男

長崎総合科学大学

#12-02

3.2.1. ・・・セキュリティとプライバシー(1/2)

(a)情報資産の価値

- 情報資産
  [　　]が個人や組織の[　　]に
  つながる情報やシステム全体
- 1次データ
  この有効活用で、[　　]は大きく成長
  個人や組織が[　　]収集したデータ。
  [　　]な情報資産。
  関連するハード、ソフト、[　　]も含む

#12-03

3.2.1. ・・・セキュリティとプライバシー(2/2)

(b)情報資産の活用と保護

- 情報資産の活用
  [　　]、大きな利益になるが、
  [　　]、大きな損失になる。
- 情報資産の保護
  - セキュリティ:外部からのアクセスによる
    情報の[　　]から情報資産を守る事。
  - プライバシー保護:[　　]が相応しい個人の情報を
    漏らさない事

#12-04

3.2.2. データサイエンスと情報セキュリティ(1/7)

(a) 情報セキュリティ

情報の機密性、完全性、可用性を維持する事

【ISO/IEC 27000】

- 機密性:[　　]が、情報にアクセスできる事
- 完全性:情報が[　　]などされずに、
  [　　]に保たれている事。
- 可用性:[　　]が、必要な時に必要な情報に
  [　　]にアクセスできる事。

NAS 長崎総合科学大学 4

#12-05

3.2.2. データサイエンスと情報セキュリティ(2/7)

(b) 機密性

許可者のみが、情報にアクセスできる事。

- 機密性の保持
  カードキー、パスワード等の [　　]が重要
- 物理的保護策
  アクセスデバイスの [　　]と設置場所への [　　]
- ソフト的保護策
  オンラインでの [　　]。特定機能の [　　]
- 情報の [　　]
  許可者が持つ復号キーで復号できる

NAS 長崎総合科学大学 5

#12-06

3.2.2. データサイエンスと情報セキュリティ(3/7)

(c) 完全性

情報が改ざんなどされずに、完全に保たれている事。

- 電子署名
  [　　]と、内容の [　　]事の保証の仕組み

(d) 可用性

許可者が必要な時に確実に情報にアクセスできる事。

- 可用性対策
  サーバーの [　　]や[　　]、
  [　　]など

NAS 長崎総合科学大学 6

#12-07

3.2.2. データサイエンスと情報セキュリティ(4/7)

(e) リスクとインシデント

- リスク
  情報資産の機密性、完全性、可用性が [＿＿＿＿＿＿＿＿] 可能性
- 脅威
  情報資産の機密性、完全性、可用性を [＿＿＿＿＿＿] 要因
- 脆弱性
  脅威によって利用される危険性のある [＿＿＿＿＿]
  [＿＿＿＿＿＿＿＿＿＿] ⇒特定の脅威による [＿＿＿＿＿＿＿＿＿＿]
- インシデント
  リスクが [＿＿＿＿＿＿] した事象。[＿＿＿＿＿＿＿＿] に努める

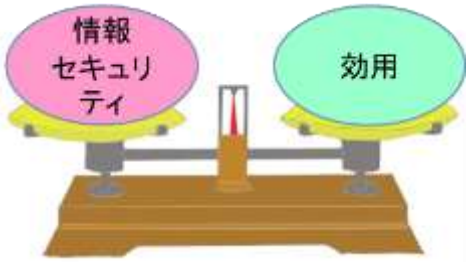
NAS 長崎総合科学大学 7

#12-08

3.2.2. データサイエンスと情報セキュリティ(5/7)

(f) 情報セキュリティと効用のトレードオフ

[＿＿＿＿＿＿＿＿] に応じて、どちらを重視するか決める。



NAS 長崎総合科学大学 8

#12-09

3.2.2. データサイエンスと情報セキュリティ(6/7)

(g) 悪意ある情報搾取の例

(1)外部要因による情報漏洩

2011年プレイステーションネットワーク

サーバーの脆弱性を外部攻撃者に突かれ、
7,700万件以上の個人情報が流出。

- 脆弱性が通知されていたにも拘らず、[＿＿＿＿＿＿]
- 対応の遅れ：公表が8日後。[＿＿＿＿＿＿＿＿]
- 損失：北米ユーザに最大[＿＿＿＿＿＿＿＿＿＿] 補償

NAS 長崎総合科学大学 9

#12-10

3.2.2. データサイエンスと情報セキュリティ(7/7)

(g) 悪意ある情報搾取の例

(2)内部要因による情報漏洩

2014年ベネッセコーポレーション

データベースの保守管理業務の委託先の派遣社員が、個人情報(3500万件)を持ち出し、[　　]に売却

- ベネッセの対応・対策費:[　　]　営業利益も[　　]減
- 派遣社員は[　　]
- 政府も重要情報の[　　]の再検討要請

NAS 長崎総合科学大学 10

#12-11

3.2.3. データサイエンスとプライバシー(1/7)

(a) プライバシー

OECD8原則

1980年公開。日本の個人情報保護法に含まれる

(1) [　　]の原則 (2) [　　]の原則

(3) [　　]の原則 (4) [　　]の原則

(5) [　　]の原則 (6)利用目的等の[　　]の原則

(7) [　　]の原則 (8) [　　]の原則

データ提供者への情報開示　(1)~(7)すべてについて

NAS 長崎総合科学大学 11

#12-12

3.2.3. データサイエンスとプライバシー(2/7)

(b) 個人情報とは　他情報と組み合わせると、[　　]が高まるが、、、

生存する個人の情報であって、・・・

・・・特定の個人を識別できるもの【個人情報保護法】

故人の情報であっても、[　　]の生存者に関わる情報は保護対象

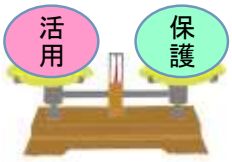
- 基本四情報:氏名、生年月日、性別、住所
- 直接識別情報(個人識別符号)　マイナンバーや学籍番号のようなもの
  :[　　]するために加工された符号。
- 要配慮個人情報:差別に[　　]個人情報
  例:人種、国籍、宗教、犯罪歴

NAS 長崎総合科学大学 12

#12-13

3.2.3. データサイエンスとプライバシー(3/7)

(c) 個人情報の活用と保護のトレードオフ



- 代表的な収集・利用制限

完全性
1) [　　]の内容の保持

OECD 8原則3)
2) [　　]の特定
3) [　　]範囲内での活用

機密性
4) 同意なしに第三者への[　　]
5) 安全管理のための[　　]の実施

#12-14

3.2.3. データサイエンスとプライバシー(4/7)

(d) 仮名化

氡名、住所、個人識別符号等

個人を特定できる情報を[　　]する。

他のデータとの連携で、特定されることもある。

- [　　]は大きく損なわれることはない。
- 仮名加工情報(仮名化された個人情報)
  : 当初の利用目的を[　　]利用できる。

第三者への提供は[　　]

#12-15

3.2.3. データサイエンスとプライバシー(5/7)

(e) 匿名化

個人特定のリスクを[　　]する情報加工。

- k-匿名化:k件以上の元データを含むような[　　]
- 匿名加工情報(一定の条件下で第三者への提供は[　　])
  1) 個人識別 [　　]の削除
  2) 個人識別 [　　]の削除
  3) 他の情報との [　　]の削除
  4) [　　]な記述の削除
  5) 他の個人情報との [　　]で、特定できないような措置

3.2.3. データサイエンスとプライバシー（6/7）

(f) プライバシー侵害の事例

(1) 医療保険情報における特定

- 仮名化医療情報と[＿＿＿＿＿＿＿＿]から特定
  [＿＿＿＿＿＿＿＿＿＿＿＿]が、特定可能性を指摘

(2) 検索履歴情報における特定

- 検索キー内の個人情報のヒントと、[＿＿＿＿＿]から特定
  [＿＿＿＿＿]検索ログの提供を受けた[＿＿＿＿＿]が特定
- コンピュータによる高精度＆網羅的な[＿＿＿＿]が可能

NAS 長崎総合科学大学 16

#12-16

3.2.3. データサイエンスとプライバシー（7/7）

(g) プライバシー・バイ・デザイン

個人情報を扱う情報システムの設計上の7つの基本原則

1) 事後ではなく、[＿＿＿＿]に
2) [＿＿＿＿＿＿＿＿]設定でプライバシー
3) 設計時に[＿＿＿＿＿＿]プライバシー対策
4) ゼロサムではなく、[＿＿＿＿＿＿＿＿＿＿]
5) エンドツーエンドのプライバシー[＿＿＿＿＿＿＿＿]
6) [＿＿＿＿＿]と[＿＿＿＿＿]
7) ユーザプライバシーの尊重

NAS 長崎総合科学大学 17

#12-17