２０２０．２．４（火）

(NPO)ビジネスサポート・Bingo

２月度定例研修会

「データサイエンティストへの道」

フルカワ技研　古川　昇

場所：福山職業能力開発短大

1　データサイエンスとは

データサイエンス(data science)とは、データを用いて新たな科学的および社会に有益な知見を引き出そうとするアプローチのことであり、その中でデータを扱う手法である情報科学、統計学、アルゴリズムなどを横断的に扱う。1)

1)Wikipedia;https://ja.wikipedia.org/wiki/データサイエンス

1　データサイエンスとは

データサイエンスで使用される手法は多岐にわたり、分野として
数学、統計学、計算機科学、情報工学、パターン認識、機械学習、データマイニング、データベース、可視化などと関係する。1)

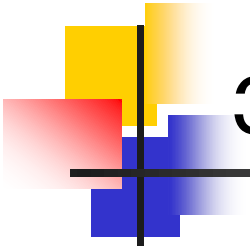
統計学、機械学習、IT(プログラミング含)のスキルを使って実装し、業務の課題を解決する。

1)Wikipedia;https://ja.wikipedia.org/wiki/データサイエンス

２ 不足するデータサイエンティスト

データサイエンスは高校生、大学生、大学院生、社会人のために、生涯の宝となる技術。データサイエンティストの人材が不足。2010年代後半から世界的にデータサイエンティストが不足しているため、高度な知識を持たないユーザーでも解析が出来るシステムの開発が進んでいる。1)2)

1)Wikipedia;https://ja.wikipedia.org/wiki/データサイエンス
2)NEC;業務システムにおける大規模データ予測を自動化する「予測分析自動化技術」を開発 (2016)



３　データサイエンティストとなるために

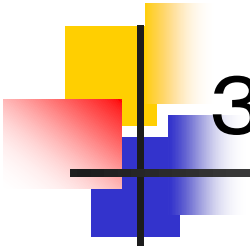
３－１　Python言語の習得

３－２　基礎統計量の把握

データの可視化(ヒストグラム、散布図、箱ひげ図)

３－３　Python言語の拡張ライブラリの利用

Numpy(ナンパイ)、Scipy(サイパイ)、Pandas、Matplotlib



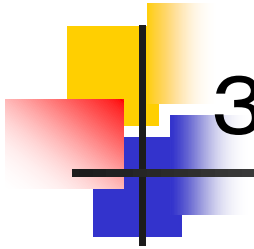
３－１　Python言語の基礎

Pythonはオブジェクト指向言語
操作手順よりも操作対象に重点を置く。
オブジェクト:物、対象、目的
　　英文法では目的語のこと。

クラス(Class)とインスタンスの関係
クラスはオブジェクトのひな型
インスタンスはクラスからできあがる実体
「たい焼き」を作るイメージ

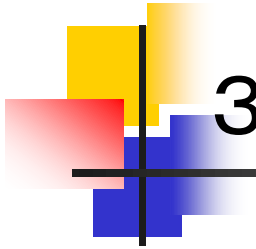
３－１　Python言語の基礎

```
import numpy as np
import matplotlib.pyplot as plt
# x2:0から99までの整数
# y2:0から1のランダムな100個の配列に、x2を掛ける
x2 = np.arange(100)
y2 = x2 * np.random.rand(100)
```



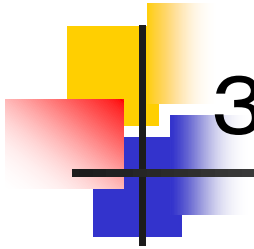
３－１　Python言語の基礎

```
# グラフの表示
#　　散布図
plt.scatter(x2,y2)
#　　ヒストグラム
# plt.hist(y2,bins=5)
# 箱ひげ図
# plt.boxplot(y2)
```



３－１　Python言語の基礎

```
import numpy as np
import matplotlib.pyplot as plt
# ワインに関するデータをロードしてdataに入れる
from sklearn.datasets import load_wine
data = load_wine()
#　インデックス0はアルコール度数、インデックス9は色彩の強さ
x3 = data.data[:,[0]]
y3 = data.data[:,[9]]
```



３－１　Python言語の基礎

```
# 散布図
plt.scatter(x3,y3)
```

３－２　基礎統計量の把握

箱ひげ図とはデータのばらつきをわかりやすく表現するための統計図である。主に多くの水準からなる分布を視覚的に要約し、比較するために用いる。ジョン・テューキーが1970年代に提唱した。様々な分野で利用されるが、特に品質管理で盛んに用いられる。**箱**（box）と、その両側に出た**ひげ**（whisker）で表現される。3)

3)wikipedia:
https://ja.wikipedia.org/wiki/箱ひげ図

箱ひげ図の四分位数

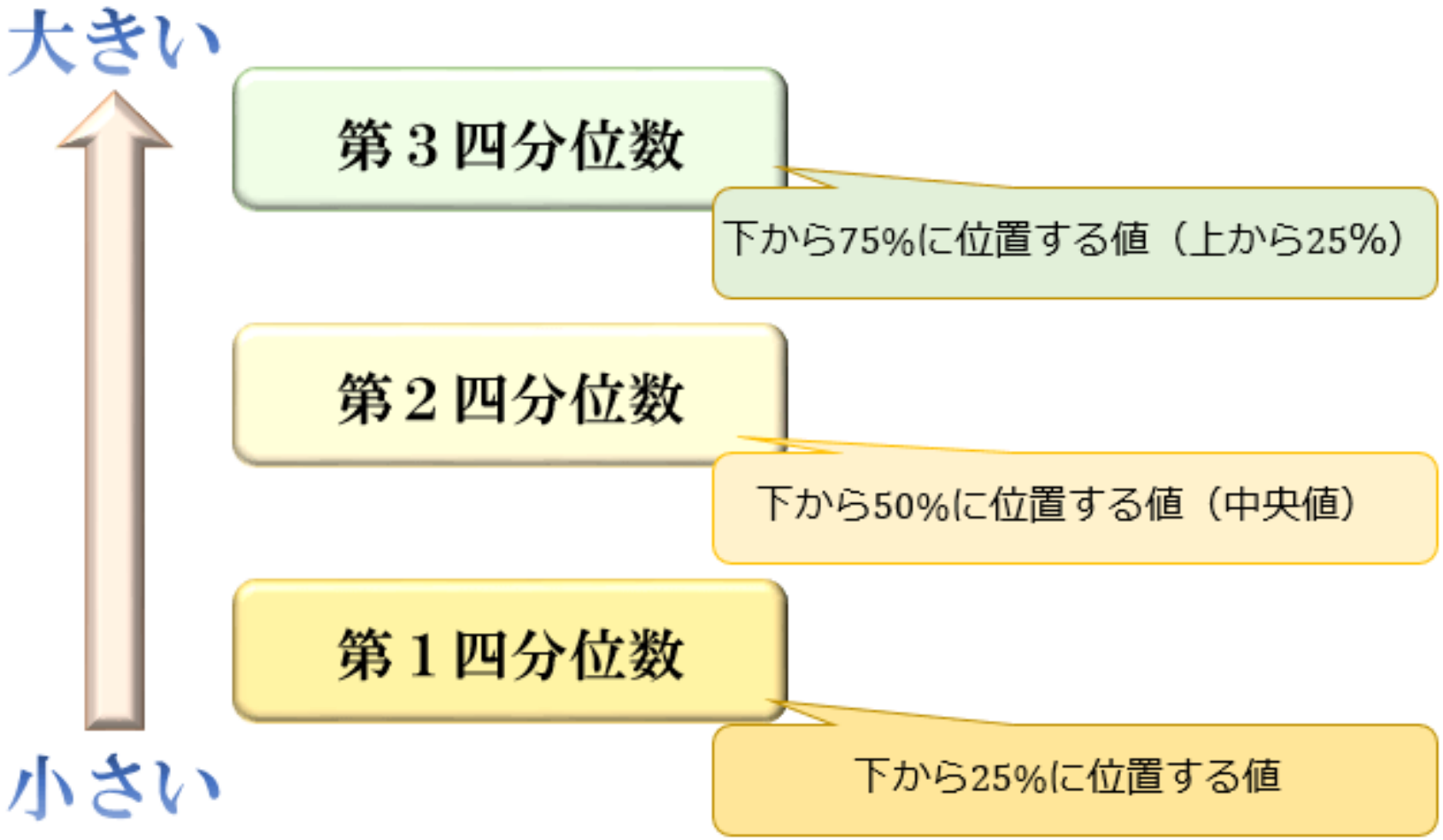


図１　箱ひげ図の四分位数

ToodaYuuto:https://atarimae.biz/archives/19162

中央値の求め方

中央値や四分位数の求め方は、データの数が偶数個か奇数個かによって変わる。

中央値：

(1)　データを小さい順に並べる。

(2)　データの数が奇数個か偶数個か。

(3)　奇数個ならちょうど真ん中の数、偶数個なら真ん中の２つの数の平均。

四分位数(ヒンジ)

(1)　データを小さい順に並べる。
(2)　中央値を求める。(第２四分位数)
(3)　中央値より小さい「前半データ」と中央値より大きい「後半データ」に分ける。
(4)　前半データ内での中央値が第１四分位数、後半データ内での中央値が第３四分位数
(5)　四分位範囲：第３四分位数－第１四分位数
(6)　四分位偏差：四分位範囲／２

箱ひげ図の実例

2020年大学入試センター試験（数学I）から引用。

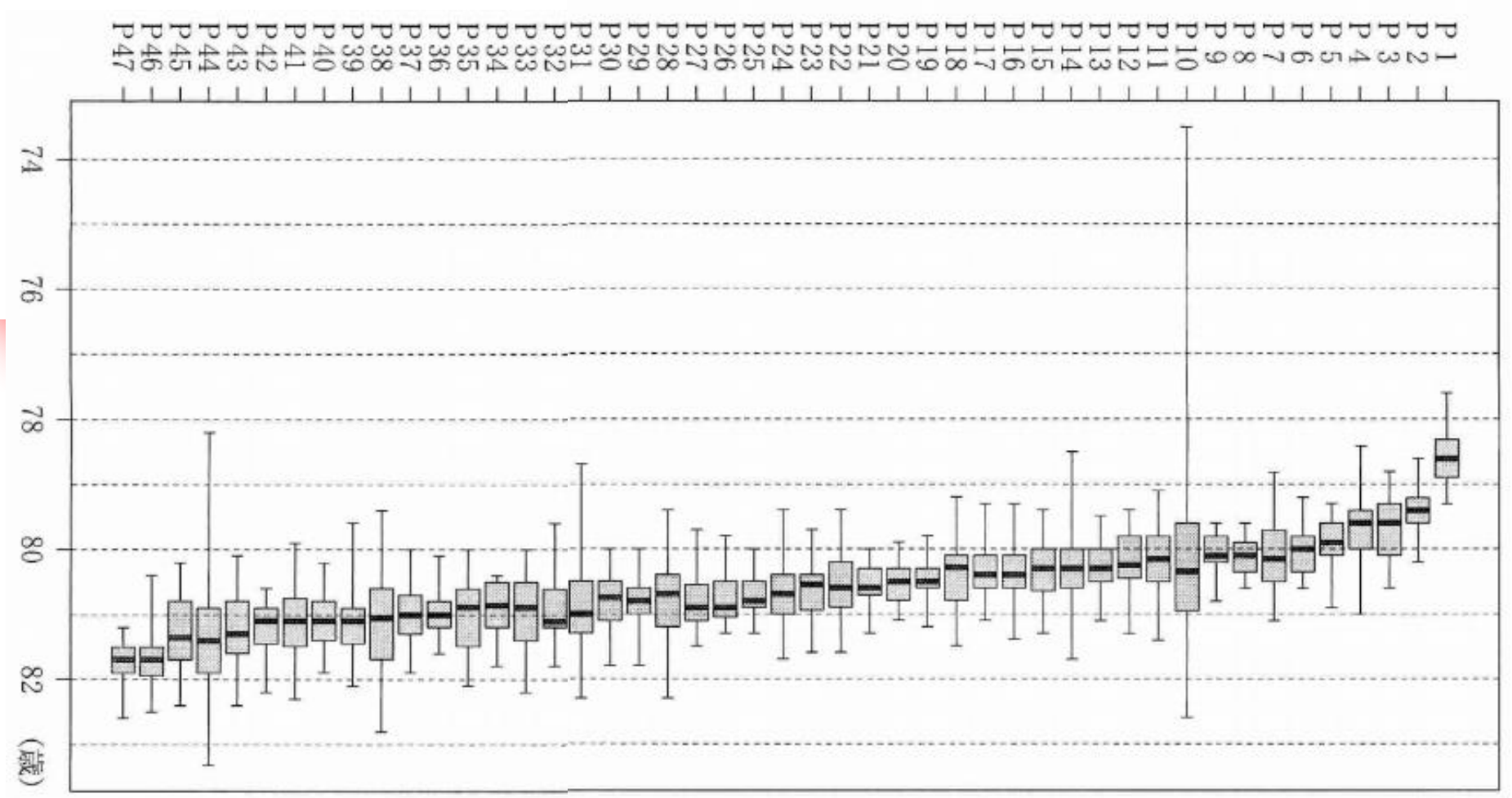
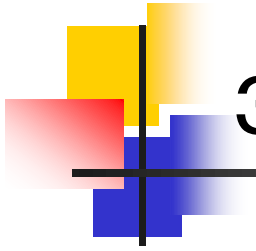


図1　平均寿命　男性の市区町村別平均寿命の箱ひげ図
(東大医学部生の相談室https://todai-counseling.com/?p=2241)

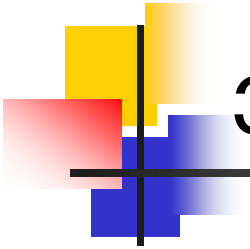


３－３　NumPy、Scipy

NumPy:Pythonで数値計算を効率的に行うための拡張モジュール。

Scipy:配列オブジェクトとその他の基本的な機能を備えた**NumPy**を基礎にしている。SciPyは統計、最適化、積分、線形代数、フーリエ変換、信号・イメージ処理、遺伝的アルゴリズム、ODE (常微分方程式) ソルバ、特殊関数、その他のモジュールを提供する。

https://ja.wikipedia.org/wiki/SciPy



３－３　Pandas、Matplotlib

Pandas:プログラミング言語Pythonにおいて、データ解析を支援する機能を提供。

Matplotlib:データの可視化機能

2020年に注目されている科学技術

Hayabusa2 Departs from Asteroid Ryugu, Expected Back on Earth by End-2020　3)4)

「はやぶさ2」は小惑星リュウグウから出発し、2020年末までに地球に戻る予定3)4)

3)話してみようJapaneseライフ、産経新聞,2020/1/29 13面

4)Takeo Kusaka :November 20, 2019 10:13 pm

https://japan-forward.com/hayabusa2-departs-from-asteroid-ryugu-expected-back-on-earth-by-end-2020/

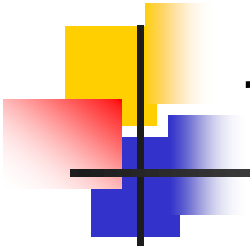
４　ディープラーニングではない手法

主成分分析:principal component analysis; PCA）は、相関のある多数の変数から相関のない少数で全体のばらつきを最もよく表す主成分と呼ばれる変数を合成する多変量解析の一手法。データの次元を削減するために用いられる。

https://ja.wikipedia.org/wiki/主成分文分析

データサイエンティストの業務

データサイエンティストは日々、大量のデータ分析を行っている。
どういうふうに？



まとめ

１　データサイエンティストは今後、ますます不足する。
２　箱ひげ図の確認など、データの前処理が重要。
３　言語はPythonが主流。
４　数値計算用モジュールは無料のものが多い。積極的に利用する。
５　ディープラーニングではないものにも、秀逸な技術がある。

ご清聴、ありがとうございました。