クレジット:
UTokyo Online Educationデータマイニング入門 2018 森　純一郎



東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

# データマイニング入門 第3回

2018年度

# 授業計画（全13回予定）

9/26　ガイダンス

10/3　データ分析のためのプログラミング基礎

10/10　データ分析のためのプログラミング基礎

10/17　データの前処理・可視化の基礎

10/24　テキスト分析の基礎

10/31　休講予定

11/7　ネットワーク分析の基礎

11/14　教師なし学習の基礎

11/21　教師なし学習の基礎

11/28　教師あり学習の基礎

12/5　教師あり学習の基礎

12/12　データ分析の実践

12/19　ミニプロジェクト

1/9　発展的な内容と全体まとめ

# PYTHONの基本的な文法

- 式と数学演算子
  - +,-,*,/,//,%,**
- 型と変数
  - int, float, str
- ブール型と比較・ブール演算子
  - TRUE/FALSE, ==/!=/>/<, and/or/not
- フロー制御
  - if, elif, else文, while, break文, for in文
- 関数
  - def, 引数, return
- リスト
  - [], インデックス, スライス, for in, メソッド
- 辞書
  - {}, キーと値, メソッド
- モジュール
  - import
- ファイル入出力（csvファイル）
  - open, csv.reader

# デバッギング

エラーに対処しやすいコードを書くために

- コードに説明のコメントを入れる
- 1行の文字数、インデント、空白などのフォーマットに気をつける
- 変数や関数の名前を適当につけない
- グローバル・ローカル変数に留意する
- コードに固有の"マジックナンバー"を使わず、変数を使う
- コード内でのコピーアンドペーストを避ける
- コード内の不要な処理は削除する
- コードの冗長性を減らすようにする など

関数の単体テストを行う

一つの関数には一つの機能・タスクを持たせるようにする

など

# デバッギング

## 文法エラー

- よくある文法エラーの例：
  - クオーテーションや括弧の閉じ忘れ
  - コロンのつけ忘れ
  - =と==の混同
  - インデントの誤り など
- 文法エラーの対処
  - まず、エラーメッセージを確認しましょう
  - エラーメッセージの最終行を見て、それがSyntaxErrorであることを確認しましょう
  - エラーとなっているコードの行数を確認しましょう
  - そして、当該行付近のコードを注意深く確認しましょう

```
print('This is the error)
```

```
  File "<ipython-input-1-ee5acbb4f977>", line 1
    print('This is the error)
                              ^
SyntaxError: EOL while scanning string literal
```

# デバッギング

## 実行エラー

- よくある実行エラーの例：
  - 文字列やリストの要素エラー
  - 変数名・関数名の打ち間違え
  - 無限の繰り返し
  - 型と処理の不整合
  - ゼロ分割
  - ファイルの入出力誤り など

```
print(1/0)
```

```
---------------------------------------------------------------------------
ZeroDivisionError                         Traceback (most recent call last)
<ipython-input-2-2fc232d1511a> in <module>()
----> 1 print(1/0)

ZeroDivisionError: division by zero
```

- 実行エラーの対処
  - まず、エラーメッセージを確認しましょう
  - エラーメッセージの最終行を見て、そのエラーのタイプを確認しましょう
  - エラーとなっているコードの行数を確認しましょう
  - そして、当該行付近のコードについて、どの部分が実行エラーのタイプに関係しているか確認しましょう。もし複数の原因がありそうであれば、行を分割、改行して再度実行し、エラーを確認しましょう
  - 原因がわからない場合は、print文を挿入して処理の入出力の内容を確認しましょう

# データとは

## レコードとフィールドからなる多次元データ

- レコードの集合
  - レコード：データ（ポイント）、インスタンス、Example、エンティティ、オブジェクト、特徴量ベクトル など
- 各レコードはフィールドの集合からなる
  - フィールド：属性、次元、特徴量、素性、変数 など

## 通常はレコード間は依存関係がなく独立であることを仮定

## レコード間に依存関係があるデータもある

- レコード間の意味的、時間的、空間的な関係
  - 時系列、ネットワーク、文字列・テキスト、空間データなど
- レコード間の関係性を考慮したデータ分析が必要

# データの種類

■菓子

| 画像かも | No. | JANCD | メーカー名 | 商品名称 | 出現日 | 金額PI | PI前週比 | 販売店率 | 平均売価 |
|---|---|---|---|---|---|---|---|---|---|
| 画像 | 1 | 4903333279609 | ロッテ | ロッテ　ラミー　２本 | 09月21日 | 639 | 734% | 77% | 154 |
| 画像 | 2 | 4902555174648 | 不二家 | 不二家　ホームパイ大人のリッチチョコＦＰ２２枚 | 09月18日 | 460 | 654% | 32% | 252 |
| 画像 | 3 | 4903333206759 | ロッテ | ロッテ　バッカス　１２粒 | 09月22日 | 383 | 694% | 76% | 155 |
| 画像 | 4 | 4902555175133 | 不二家 | 不二家　ホームパイ大人のリッチチョコ　１２枚 | 09月21日 | 281 | 856% | 38% | 167 |
| 画像 | 5 | 4549660289388 | バンダイ | バンダイ　仮面ライダーＳＧライドウォッチ０１ | 09月22日 | 278 | 254% | 30% | 495 |
| 画像 | 6 | 4901335175141 | 湖池屋 | 湖池屋　じゃがいも心地　合わせ塩味　５８ｇ | 09月12日 | 260 | 449% | 43% | 97 |
| 画像 | 7 | 4902501209554 | フルタ製菓 | フルタ製菓　チョコエッグワンピース　２０ｇ | 08月23日 | 259 | 94% | 67% | 193 |
| 画像 | 8 | 4901581543039 | 木村屋総本店 | 木村屋　ジャンボむしケーキ大人のチョコ　１個 | 09月01日 | 247 | 194% | 6% | 91 |
| 画像 | 9 | 4902501624524 | フルタ製菓 | フルタ製菓　きなこもちクッキー　２１０ｇ | 09月12日 | 245 | 144% | 17% | 258 |
| 画像 | 10 | 4901870300138 | シジシージャパン | ＣＧＣ　ダブルチョコドーナッツ　８個 | 09月25日 | 234 | 0% | 6% | 258 |
| 画像 | 11 | 4902751332880 | モンテール | モンテール　とろける生ロール・クレームブリュレ４個 | 09月01日 | 229 | 135% | 17% | 276 |
| 画像 | 12 | 4976406130787 | 武蔵製菓 | 武蔵製菓　十五夜団子　１５個 | 09月01日 | 223 | 218% | 11% | 300 |
| 画像 | 13 | 4903333259823 | ロッテ | ロッテ　ＨＷチョコパイパーティーパック　９個 | 08月25日 | 222 | 108% | 80% | 290 |
| 画像 | 14 | 4902751332828 | モンテール | モンテール　ふわもちミニたい焼・あずきミルク　４個 | 09月01日 | 214 | 165% | 12% | 268 |
| 画像 | 15 | 4902751333122 | モンテール | モンテール　ふわもちミニたい焼・カスタード　４個 | 09月01日 | 214 | 162% | 10% | 274 |

株式会社KSP-SP 新商品売れ筋ランキング 2018年39週　菓子
https://www.ksp-sp.com/open_data/ranking/2018/201839.html?yw=201839
( ref. 26 Sep 2018)

# データの種類

レコードのフィールドは主に以下のデータを持つ

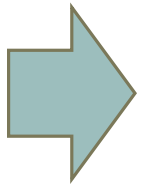
- 量的・数値データ
  - 離散的または連続的
  - 数的・統計的処理が可能
- 順序データ
  - 数の順序・大小を表すデータ
- バイナリ（集合）データ
  - 1か0（出現のありなし）で表現
  - 数的・統計的処理が可能
- カテゴリカルデータ
  - 離散的・順序がないデータ
  - バイナリデータへ変換することで数的・統計的処理が可能
- テキストデータ
  - ベクトルにすることで数的・統計的処理が可能

# データの種類

カテゴリカルデータからバイナリデータへの変換

| メーカー名 |
| --- |
| ロッテ |
| 不二家 |
| ロッテ |
| 不二家 |
| バンダイ |



| ロッテ | 不二家 | バンダイ |
| --- | --- | --- |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

# データの種類（統計解析の変数として）

## 定性的・質的変数

- 名義尺度
  - 数としての意味を持たないが数え上げはできる
    - カテゴリカルな変数
    - バイナリ（2値）変数
- 順序尺度
  - 値に順序・大小関係が意味を持つ変数

## 定量的・量的変数

- 間隔尺度
  - 値の大小関係、間隔・差が意味を持つ変数
    - 温度、西暦　など
- 比例尺度
  - 値の大小関係、間隔・差、比率が意味を持つ変数
    - 長さ、重さ　など

# データ行列

多次元データは行列（あるいはテンソル）とみなすことができる

- 多次元データ $D$
  - レコード数$n$
  - フィールド数 $d$
- $D$は$n$ ×$d$の行列で表現される
  - 行：レコード
  - 列：フィールド

行列処理としてのデータ分析

- データ分析手法の多くはベクトル・行列の計算を用いる
  - 列間の関係
    - フィールド・属性の関係性を見つける
      - 企業と売り上げの関係は？
  - 行間の関係
    - レコード間の関係性を見つける
      - 似ている商品は？

# データの観察
# 記述統計

## データの特徴を指標（統計値）によって記述

- 分析対象データの特徴の理解
- 欠損値や外れ・異常値の処理の準備

## 代表的な記述統計と可視化

- データの中心傾向（代表値）
  - 平均、中央値、最頻値（mode）など
- データのばらつき
  - 最大値、最小値、レンジ、分散、標準偏差、分位数 など
- データ全体の可視化
  - 分布、ヒストグラム、相関、散布図 など

# データの中心傾向

$n$個数のレコードからなる1変数のデータがあり、各レコードの変数の値を$x_i$とする

平均

- データの変数のすべての値を足して、レコードの数で割ったもの
- 平均は外れ・異常値の影響を受けやすいことに注意

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

トリム（調整）平均

- データの最大値と最小値付近の範囲の値を平均の計算から除外
- 外れ値の影響を取り除く
  - 例：上位、下位1%の値を除外

# データの中心傾向

## 中央値（メディアン）

- データの変数の値を大きさの順に並べたとき全体の中央に位置する値
- 平均に比べ外れ値の影響を受けにくい
  - データのレコード数が奇数個の場合は、大きさの順に並び替えた(n+1)/2番目の値
  - データのレコード数が偶数個の場合は、大きさの順に並び替えたn/2番目とn/2+1番目の値の平均
    - [1,2,3,4,5]->3
    - [1,2,3,4]->(2+3)/2=2.5

## 最頻値（モード）

- データの変数の値で最も出現回数が多い値
- カテゴリカルデータで使用することが多い

# データのばらつき

## 範囲（レンジ）
- データの最大値と最小値の差
  - 外れ値の影響を受けやすい

## 四分位
- 第１四分位（Q1）
  - データを小さい方から並べた時、全体を1:3に分ける位置の値
- 第３四分位（Q3）
  - データを大きい方から並べた時、全体を1:3に分ける位置の値
- 第２四分位（Q2）
  - 中央値

## 四分位範囲（IQR: interquartile range)
- 第3四分位数から第1四分位数を引いた値
  - 中央値に対するばらつきを示す指標とみなせる

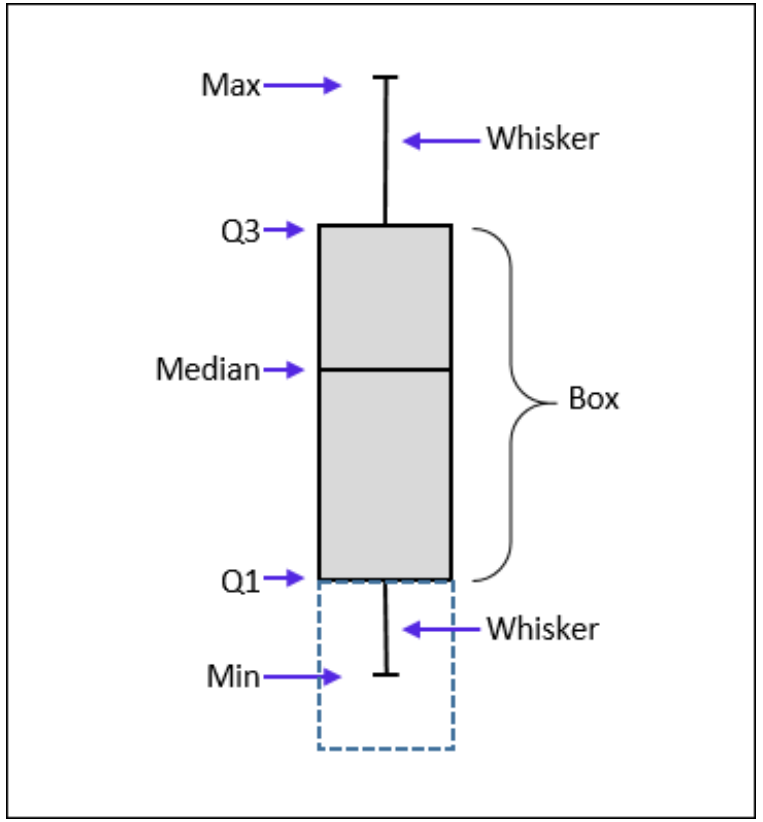
# データのばらつき

## 5数要約

- データのばらつきを5数で表現
  - 最小値、第１四分位、中央値、第3四分位、最大値

## 箱ひげ図

- データの5数要約、分布を可視化
  - 箱の長さは四分位範囲
  - 箱の終端は、それぞれ第１四分位と第3四分位
  - 箱の中央は中央値
  - 箱が乗る直線の終端は、それぞれ最小値と最大値



# データのばらつき

## 分散

- 平均からのデータのばらつきを示す指標
  - データの個々のレコードの変数の値とデータ全体の平均の値の差の2乗和をレコード数で割ったもの
  - 個々のデータとデータ全体の平均の偏差の2乗（偏差の面積）を標準化したものとみなせる

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

## 標準偏差

- 分散と同様にデータのばらつきを示す指標
- 分散の正の平方根
  - 分散の平方根をとることで元のデータの平均と同じ次元となる

# データの正規化

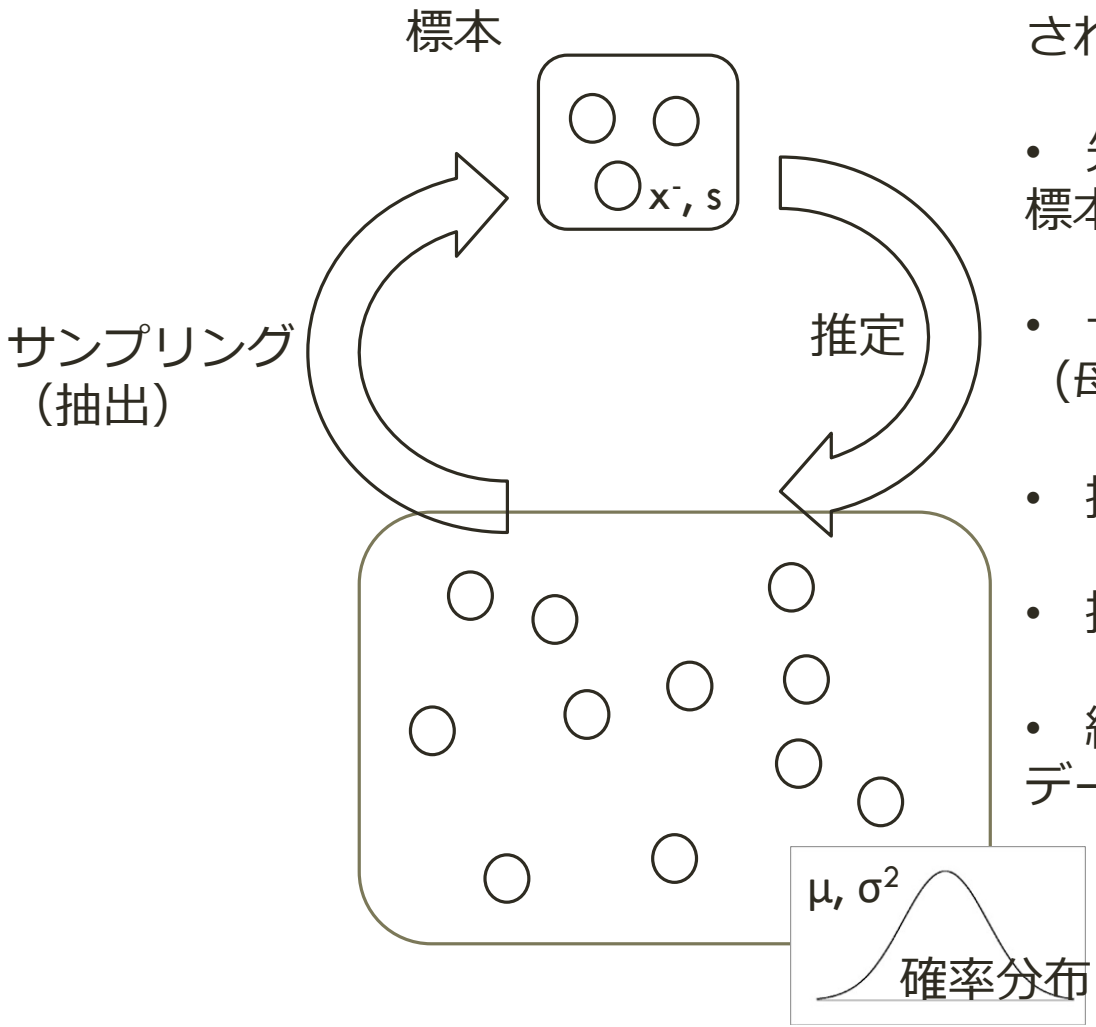
異なるデータをそれぞれの平均や分散によらない統一的な指標に変換

- 標準化
  - データから平均を引き、標準偏差で割る
  - 標準化されたデータは平均が0、標準偏差が1となる

$$\frac{x_i - \bar{x}}{s}$$

# 標本と母集団




- 今日の記述統計の説明では、元の母集団のデータから抽出された標本を「データ」としている
- 先の平均、分散は標本の統計量であり、正確には標本平均、標本分散と呼ぶ
- 一般的に統計では、観測された標本の特徴から母集団の特徴（母数）を推測する（推測統計）
- 推測統計には点推定、区間推定、検定などがあります
- 推測統計については統計データ解析を参照してください
- 統計的機械学習では、一般に学習データ（標本）からデータ全体（母集団）のパラメータ（母数）を推定する

# 標本と母集団

## 不偏分散

- 観測しているデータ（標本）の分散ではなく、そのデータが抽出された元の母集団のデータの分散を推定したいときはnではなくn-1で割ります

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i)^2 - \frac{n}{n-1}\bar{x}^2$$

- n-1は自由度（自由に値を動かせる変数の数）を表している
- 標本の平均は既定されているため自由度が1つ減る
- 母集団の平均が既知の場合は自由度を減らす必要はない

# データの分布と可視化

## 度数分布（ヒストグラム）

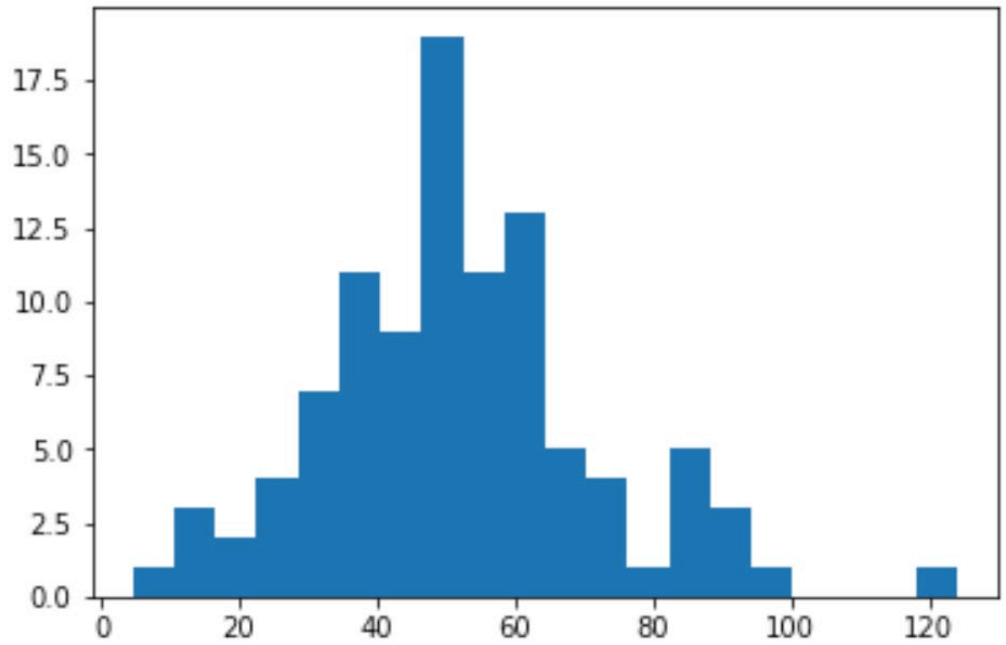
- データの全体的な分布を示す
- データの変数がとる値を区分に分け、区分ごとにその区分内の値をとるレコードを数える

## 階級数と階級幅

- 値の区分の数と区分幅

## 度数

- 各階級に属するレコードの数



# データの分布と可視化

## 確率分布

- 確率変数（データ）の値とその確率の対応

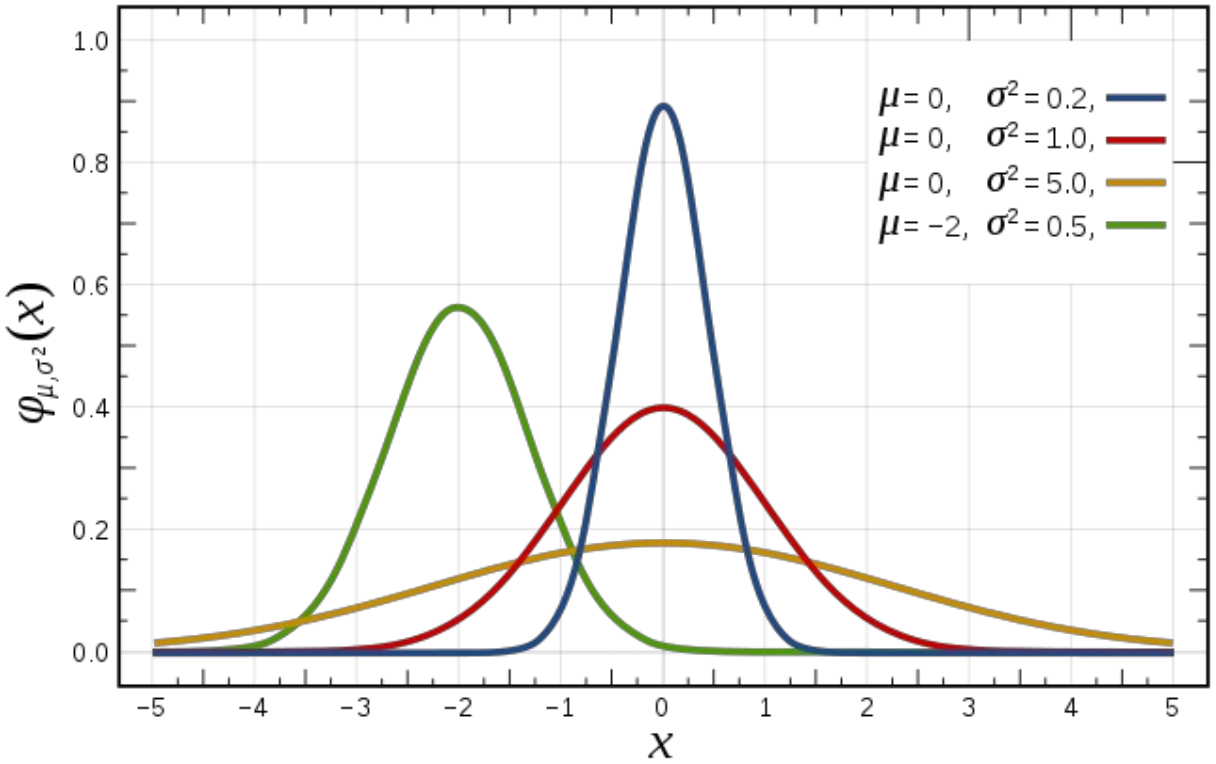
## 正規分布

- 連続確率変数の代表的な分布
- 平均μ、分散$\sigma^2$の正規分布の確率密度関数

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (x \in \mathbb{R})$$

## 標準正規分布

- 任意の正規分布は確率変数を標準化することで
  平均0, 分散1の標準正規分布に変換できる




https://ja.wikipedia.org/wiki/%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83#/media/File:Normal_Distribution_PDF.svg

# データの分布と可視化

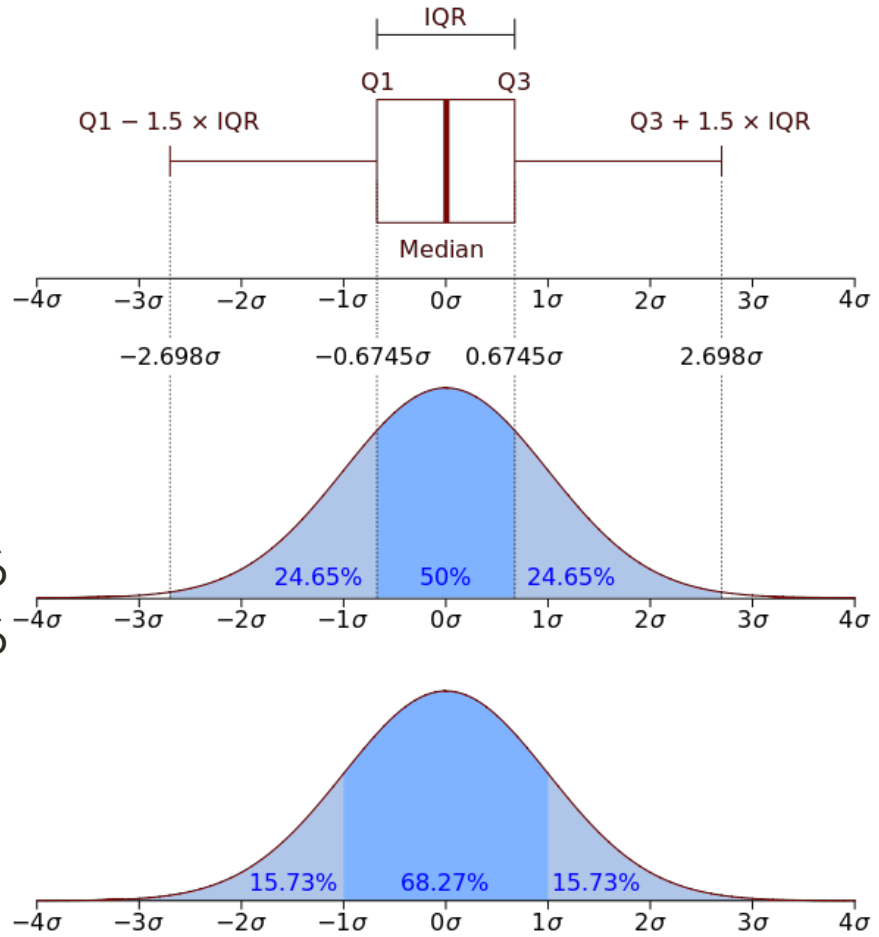
平均μ、標準偏差σの正規分布では

- [μ-σ, μ+σ]の範囲には全体の約68.27%のデータが含まれる
- [μ-2σ, μ+2σ]の範囲には全体の約95.45%のデータが含まれる
- [μ-3σ, μ+3σ]の範囲には全体の約99.73%のデータが含まれる

一般にどのような分布であっても

- チェビシェフの不等式により
- [μ-σ, μ+σ]の範囲には少なくとも全体の75%のデータが含まれる
- [μ-2σ, μ+2σ]の範囲には少なくとも全体の約88.9%のデータが含まれる
- [μ-3σ, μ+3σ]の範囲には少なくとも全体の約93.8%のデータが含まれる

データの外れ値の検出に利用できる




Boxplot vs PDF.svg, https://ja.wikipedia.org/wiki/%E3%83%95%E3%82%A1%E3%82%A4%E3%83%AB:Boxplot_vs_PDF.svg CC BY-SA 2.5

# データの分布と可視化

## 代表的な確率分布

- 離散型
  - ベルヌーイ分布
  - 二項分布
  - ポアソン分布
- 連続型
  - 正規分布（ガウス分布）
  - 指数分布
  - カイ二乗分布
  - t分布
  - F分布