

クレジット:

UTokyo Online Education データマイニング入門 2018 森 純一郎

ライセンス:

利用者は、本講義資料を、教育的な目的に限ってページ単位で利用することができます。特に記載のない限り、本講義資料はページ単位でクリエイティブ・コモンズ 表示-非営利-改変禁止 ライセンスの下に提供されています。

<http://creativecommons.org/licenses/by-nc-nd/4.0/>

本講義資料内には、東京大学が第三者より許諾を得て利用している画像等や、各種ライセンスによって提供されている画像等が含まれています。個々の画像等を本講義資料から切り離して利用することはできません。個々の画像等の利用については、それぞれの権利者の定めるところに従ってください。



# データマイニング入門 第4回

2018年度

# 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

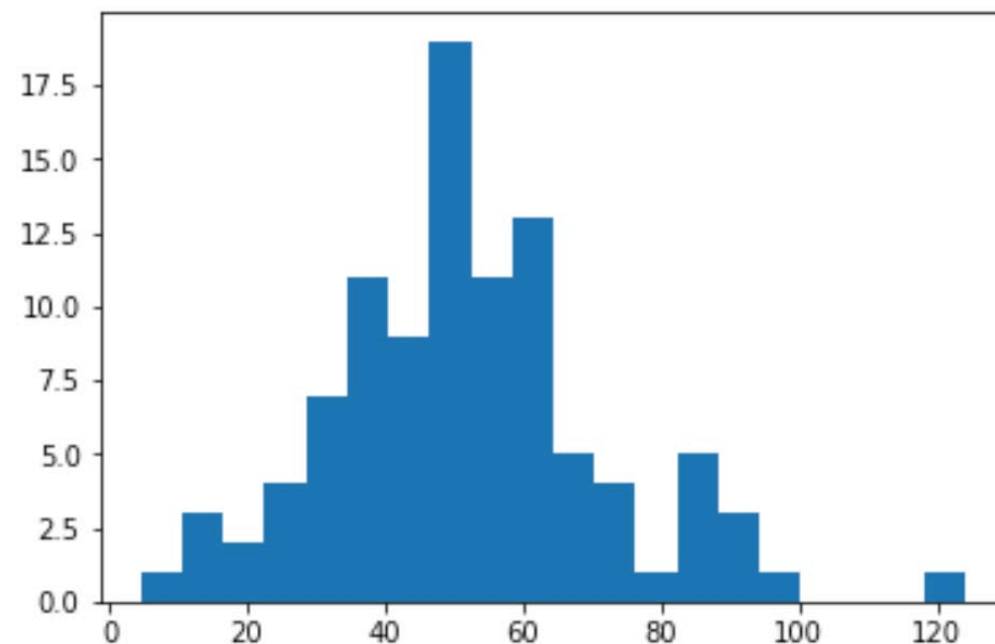
## 階級数と階級幅

- 値の区分の数と区分幅

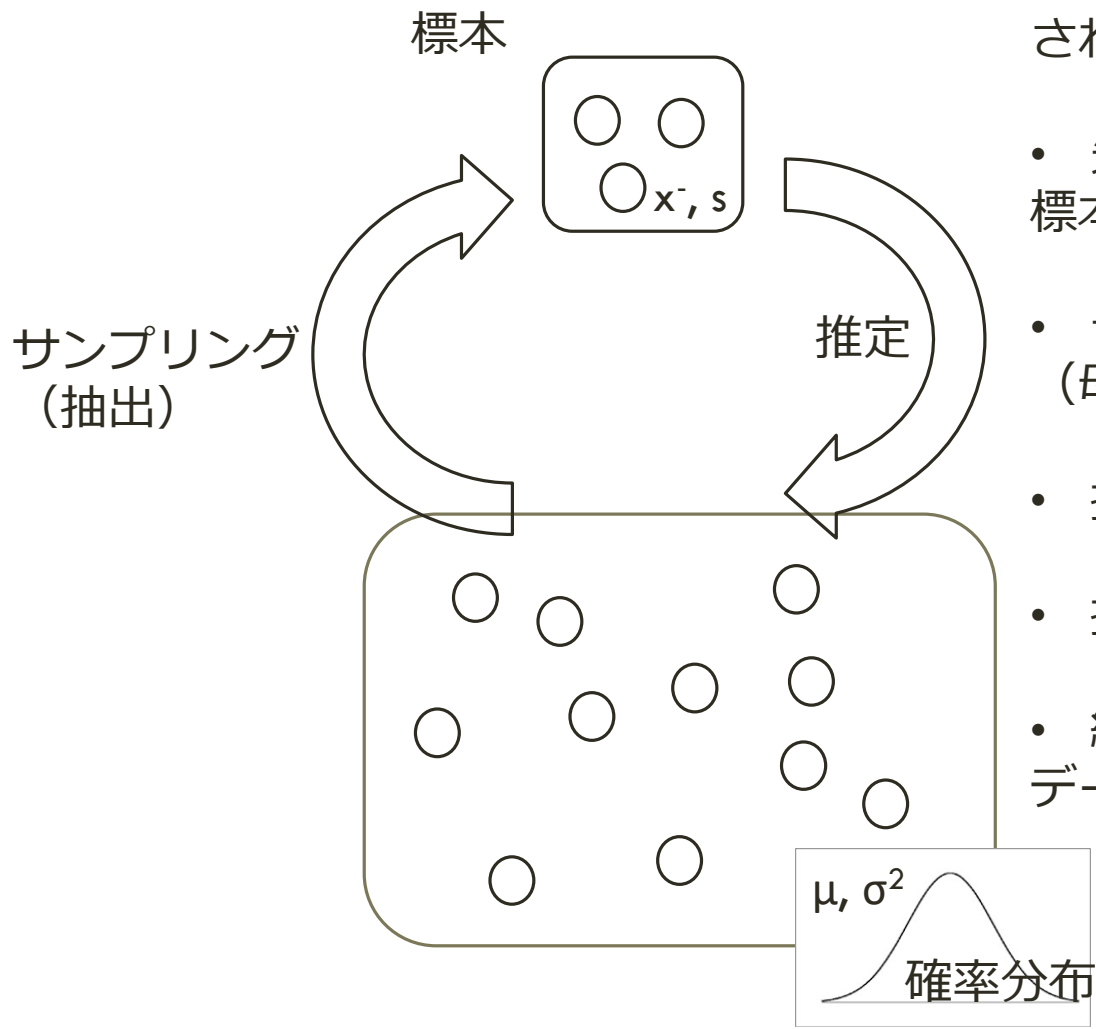
## 度数

- 各階級に属するレコードの数

\*課題で用いてる得点データは厳密には正規分布ではなくワイブル分布に従います



# 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

# 確率変数としてのデータ

統計では、データを確率変数 $X$ の実現値とみなす

- 確率変数

- 離散型

- 確率変数 $X$ のとり値の集合 $\{x_1, \dots, x_i, \dots\}$
    - 確率変数 $X$ が $x_i$ という値をとる確率  $f(x_i) = P(X=x_i) = p_i$
    - 確率分布：確率変数の値とその確率の対応

- 連続型

- 確率密度関数： $f(x)$
    - 確率変数 $X$ が $a \leq X \leq b$ の区間に入る確率

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

# 確率変数としてのデータ

母集団は確率分布（確率密度関数） $f(x)$ を持っている

- $f(x)$ は連続型でも離散型でもよい

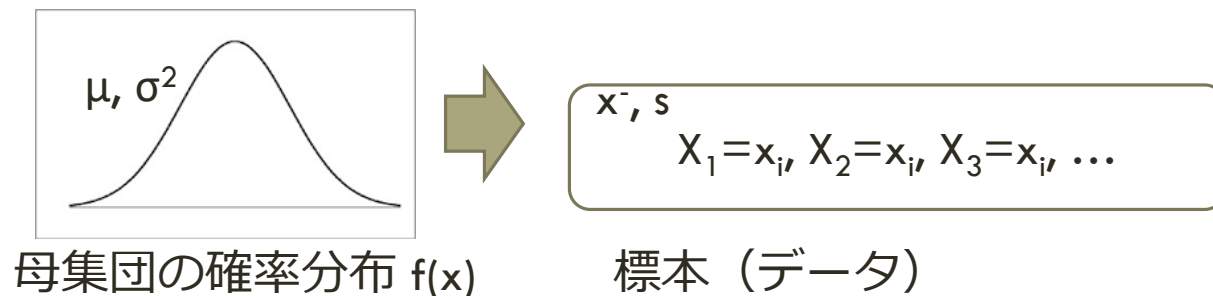
母集団から標本（データ） $\{x_1, x_2, \dots\}$ をランダムに選ぶので、各データ $x_i$ は母集団の確率分布 $f(x)$ に従う確率変数

- データは同一の確率分布に従う独立な確率変数（独立同分布）

統計的推測では標本の特徴から元の母集団の特徴を推測したい

母集団は確率分布を持っているのでその確率分布（パラメータ）を知ればよい

- 母平均 $\mu$ 、母分散 $\sigma^2$ は代表的なパラメータ



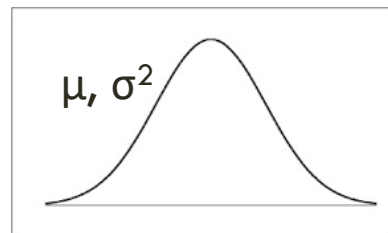
# 統計的推測

## パラメトリックな統計的推測

- 事前に母集団の確率分布がある知られた確率分布であり、その分布を決定するパラメータがわかれば母集団分布について知ることができる
- このパラメータを母数と呼ぶ
- 母集団の確率分布の代表的な母数
  - 母平均 $\mu$ 、母分散 $\sigma^2$
  - 特に正規分布では母平均（期待値） $\mu$ と母分散 $\sigma^2$ の2つの母数によって確率分布が決定される

## 大数の（強）法則

- 確率変数が独立同分布であれば、標本数が十分大きいとき、標本平均 $\bar{x}$ 、標本分散 $s$ は母集団の真の平均 $\mu$ と分散 $\sigma^2$ に収束する
- 標本平均・分散は母平均・分散を知るための手がかり



母集団の確率分布  $f(x)$



$$\bar{x}, s$$
$$X_1=x_1, X_2=x_1, X_3=x_1, \dots$$

標本（データ）

# 正規分布

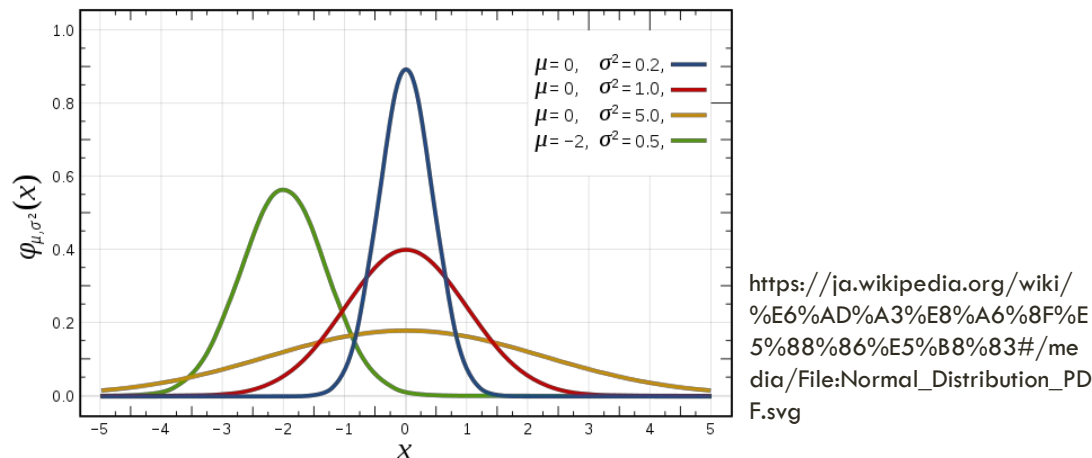
確率変数が正規分布に従うとき、その確率密度関数は

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

正規分布に従う確率変数の平均（期待値） $E(X)$ 、分散 $V(X)$ は

- $E(X)=\mu, V(X)=\sigma^2$

正規分布は平均、分散の2つの母数で決まる



確率変数の平均（期待値）と分散

- 離散型
  - 平均
    - 確率変数のとりうる値とその確率の積

$$E(X) = \sum_i x_i P(X = x_i)$$

- 分散

$$V(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$$

- 連続型

- 平均

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- 分散

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$



# パラメータ（母数）の推定

母集団のパラメータ（母数） $\theta$ を標本から推定する

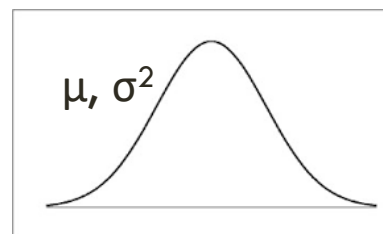
- 母数にはいろいろなものがある
  - 母平均、母分散、...

## 点推定

- 母数 $\theta$ をある特定の値 $\theta^{\wedge}$ で推定する方法
  - 母平均 $\mu$ を標本平均 $\bar{x}$ で推定

## 推定量

- 母数を推定するために標本から求めた統計量
  - 標本平均、標本分散、など
- (最低限) 不偏性と一致性を満たす必要がある
  - 推定量の期待値=母数の値
  - 標本数が十分大きいとき: 推定量の値=母数の値



母集団の確率分布  $f(x)$



$$\bar{x}, s$$
$$X_1 = x_i, X_2 = x_i, X_3 = x_i, \dots$$

標本（データ）

# 最尤法による点推定

## 最尤法による点推定

- 最尤原理
  - 現実の標本は確率最大のものが実現した
- 尤度関数
  - 母数 $\theta$ の元で、標本の値がどの程度起こりうるかを表す関数
  - 確率分布（確率密度関数）の積を母数 $\theta$ の関数とみなしたものの

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

- 尤度関数を（母数がとる値の集合の空間で）最大にするものを推定量とする
- 対数尤度
  - 尤度関数の対数をとって和の形にしたもの
  - 対数尤度を最大にする推定量：最尤推定量

$$\log L(\theta) = \sum_{i=1}^n \log f(x_i|\theta) \quad \frac{\partial \log L(\theta)}{\partial \theta} = 0$$

# 正規分布のパラメータの点推定

## 正規分布の点推定

- 正規分布は平均、分散の2つの母数で決まる確率分布であった
- 尤度関数と対数尤度
  - 確率変数である標本（データ） $\{x_1, x_2, \dots, x_n\}$ の実現値が $\{x_1, x_2, \dots, x_n\}$ の時

$$L(\mu, \sigma^2) = \prod_{i=1}^n f(x_i | \mu, \sigma^2) \quad L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$\log L(\mu, \sigma^2) = -n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}$$

- 対数尤度を最大する平均と分散の最尤推定量を求める

$$\frac{\partial \log L(\mu)}{\partial \mu} = \sum_{i=1}^n \frac{(x_i - \mu)}{\sigma^2} = 0 \quad \frac{\partial \log L(\sigma^2)}{\partial \sigma^2} = \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0$$

- 標本平均・分散が母数（母平均・分散）の推定量となっている \*正確には不偏分散を使ったほうがよい

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

# データの相関

## 2次元データ

- 2つの変数（フィールド）を持つデータ

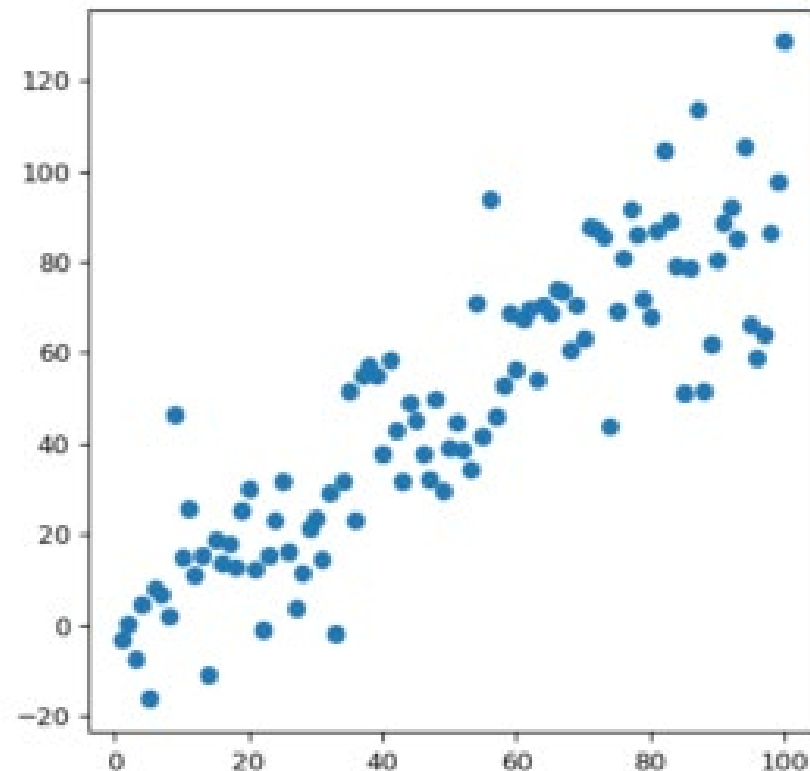
## 関係

- 観測したデータにおける2変数の関係
  - 相関
    - 2変数間の相互関係
  - 回帰
    - 一方の変数が他方の変数から決定される関係

## 散布図

- 縦軸、横軸に2つの変数の数を対応させて、それらの関係を点でプロット

<https://python-graph-gallery.com/130-basic-matplotlib-scatterplot/>  
Copyright © 2017 The python graph gallery (ref. 11 Dec 2018)



# データの相関

## 共分散

- 2つの変数の関係性を見るとき使われ以下の特徴を持つ
  - 共分散が正
    - 片方の変数が大きい値をとれば、もう片方も大きい値をとる
  - 共分散が負
    - 片方の変数が大きい値をとれば、もう片方は小さい値をとる
  - 共分散が0
    - データ間に関係性がない
- 具体的には各変数の平均からの差の積を取る

$$(x \text{ と } y \text{ の共分散}) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

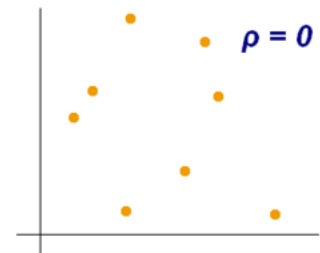
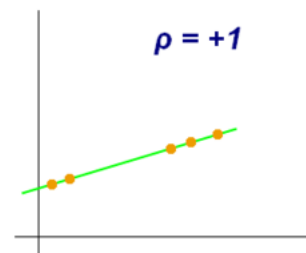
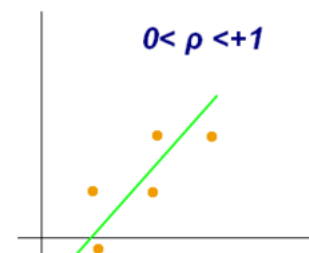
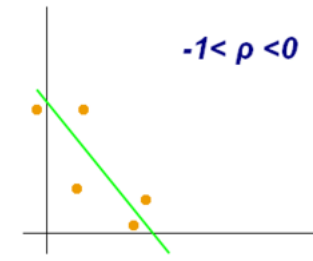
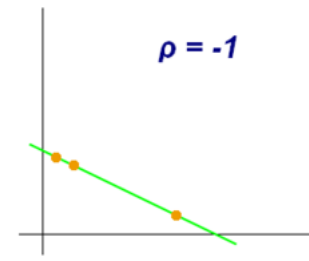
$$(x \text{ の分散}) = \frac{\sum(x_i - \bar{x})^2}{n}$$

# データの相関

## 相関係数（ピアソンの積率相関係数）

- 変数間の相関の度合いを示す指標
- 共分散を、-1~1の範囲に正規化したもの
  - 正の相関を持つほど相関係数が1
  - 負の相関を持つほど相関係数が-1

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left( \left( \sum_{i=1}^n (x_i - \bar{x})^2 \right) \left( \sum_{i=1}^n (y_i - \bar{y})^2 \right) \right)^{1/2}}$$



# データの相関

## 分散共分散行列

- 複数の変数において、分散と共分散の一覧を行列の形でまとめたもの
- 2変数においては以下のようにかける

$$\begin{array}{cc} & \begin{array}{c} x \\ y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} [x\text{の分散}] & [x\text{と}y\text{の共分散}] \\ [x\text{と}y\text{の共分散}] & [y\text{の共分散}] \end{bmatrix} \end{array}$$

## 相関行列

$$\begin{array}{cc} & \begin{array}{c} x \\ y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} [1] & [x\text{と}y\text{の相関係数}] \\ [x\text{と}y\text{の相関係数}] & [1] \end{bmatrix} \end{array}$$

# データマイニングのプロセス

## データの同定・収集・選択

- 対象領域の理解
- データ集合からマイニングに必要なデータセットを選択

## 前処理

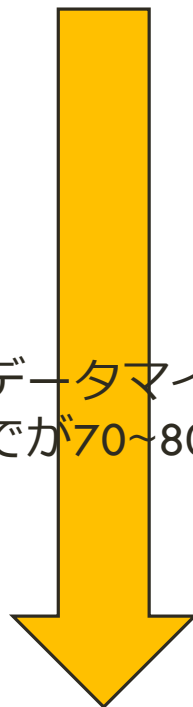
- データセットからノイズや異常値をクレンジング
  - 連続値の離散化・カテゴリー化、欠損の補完、外れ値・重複除去
  - 複数データの統合、正規化、表記・単位の統一、矛盾解消 など

## 変換

- 前処理されたデータをアルゴリズム入力用に変換
  - 属性選択、属性抽出、属性構築・統合、事例選択
  - サンプリング、相関、次元削減、外部知識利用 など
  - 構造データと非構造データ

## マイニング（パターン発見）

- 統計分析、機械学習、データマイニング手法



実際のデータマイニングでは  
変換までが70~80%と言われている



# データの前処理

## “きれい”ではないデータ

- ノイズ
  - センサーにおける計測ノイズなど
- エラー
  - 音声認識、OCR、人手作業によるエラーなど
- データ保護
  - プライバシー、秘匿性による一部データの欠損など
- データ欠損
  - データ収集の不完全性による一部データの欠損など

# データの 前処理 欠損値

欠損値が発生するなんらかの原因によるデータ行列の要素の値の欠損

データを観察し欠損値が発生するパターンを理解した上で欠損の削除・補完を行う

- 欠損値は特定の変数に起因するものか？
- 他の変数の影響を受けているか？
- 欠損値の削除
  - 欠損値を含むレコードの特定と削除
- 欠損値の補完
  - 手動で補完
  - 定数で補完
  - 類似レコードで補完
  - 自動推定
- 欠損値を許容する分析手法

	kokugo	shakai	sugaku	rika
0	30.0	43.0	51	NaN
1	39.0	21.0	50	56.0
2	NaN	NaN	23	57.0
3	29.0	87.0	77	100.0
4	70.0	71.0	78	67.0
5	66.0	NaN	53	NaN
6	29.0	26.0	44	52.0
7	NaN	54.0	37	59.0
8	45.0	NaN	7	44.0
9	68.0	41.0	29	81.0

# データの 前処理

## 欠損値

### 欠損値の補完

- 平均、中央値、最頻値など
  - 欠損値を含む変数の記述統計（平均、中央値、最頻値）で欠損値を補完する
- 確率分布の推定
  - 欠損値を含む変数の確率分布を仮定し、非欠損値からその確率分布のパラメータ) を推定し、推定した確率分布に基づいて欠損値を補完する
- 回帰
  - 欠損値を含む変数をその他の変数から回帰する問題を考えて、回帰モデルに基づいて欠損値を補完する

# データの 前処理 外れ値

## 外れ・異常値の処理

- 矛盾の利用
  - 変数自体の制約の利用など
- ドメイン知識の利用
  - 変数間関係による制約の利用など
- 統計的・数的処理
  - 平均、中央値、最頻値
  - 確率分布の推定
  - データの空間におけるデータ間距離、データの密度、クラスタリングなど
    - 特に多次元データで変数が多い場合

# データの処理 外れ値

## 5数要約

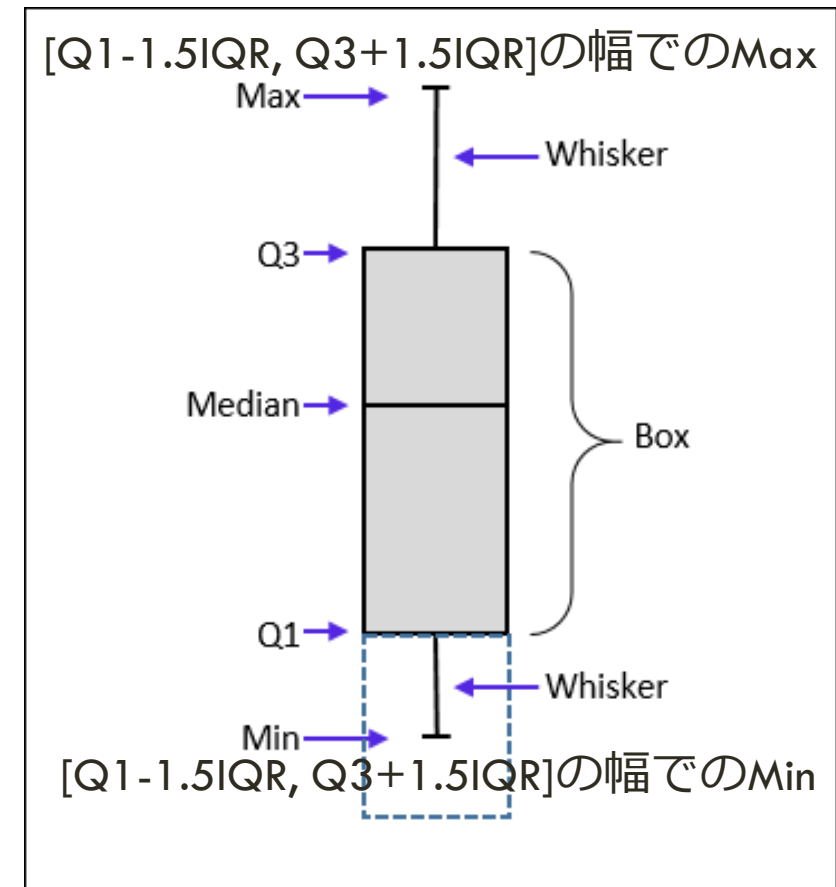
- データのばらつきを5数で表現
  - 最小値、第1四分位、中央値、第3四分位、最大値

## 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

## 外れ値

- $Q1 - 1.5IQR$ より小さい値
- $Q3 + 1.5IQR$ より大きい値



# データの 前処理 外れ値

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](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

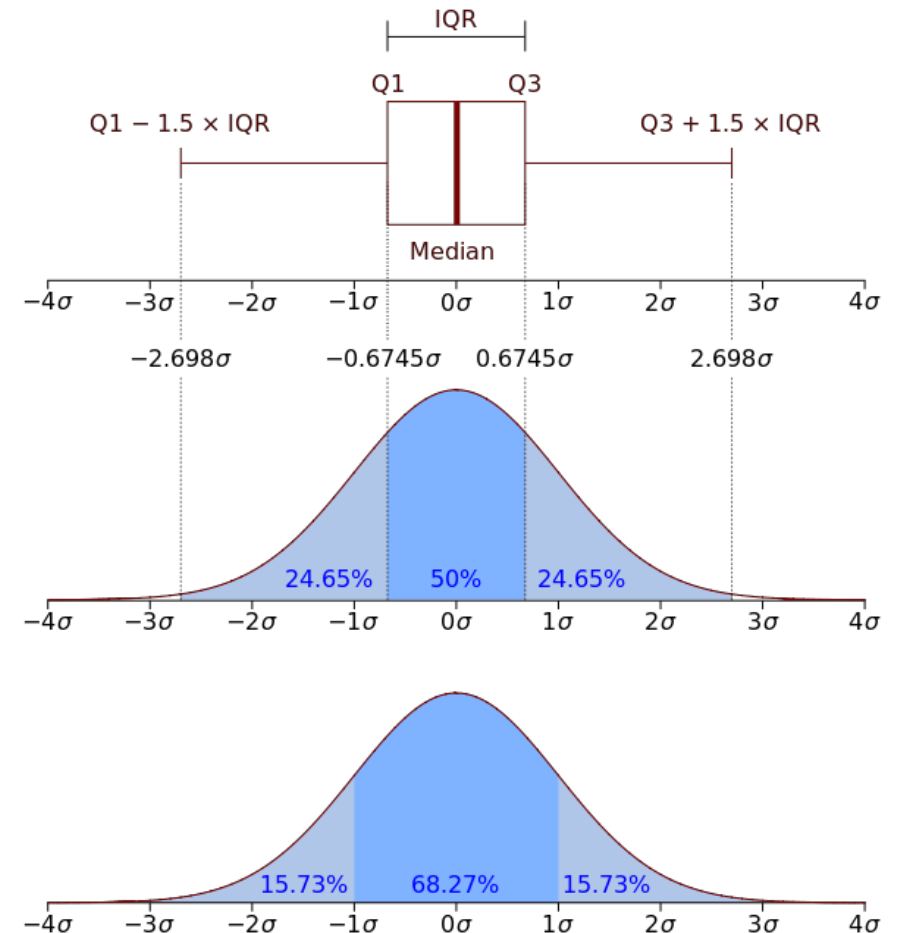
平均 $\mu$ 、標準偏差 $\sigma$ の正規分布では

- $[\mu - \sigma, \mu + \sigma]$ の範囲には全体の約68.27%のデータが含まれる
- $[\mu - 2\sigma, \mu + 2\sigma]$ の範囲には全体の約95.45%のデータが含まれる
- $[\mu - 3\sigma, \mu + 3\sigma]$ 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

- チェビシエフの不等式により
- $[\mu - \sigma, \mu + \sigma]$ の範囲には少なくとも全体の75%のデータが含まれる
- $[\mu - 2\sigma, \mu + 2\sigma]$ 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
- $[\mu - 3\sigma, \mu + 3\sigma]$ 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



# データの前処理 標準化・スケーリング

## 標準化

- データの値から平均を引き、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

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

## スケーリング

- min-maxスケーリング
  - データの値から最小値を引き、最大値と最小値の差で割る
    - $(x - \text{最小値}) / (\text{最大値} - \text{最小値})$
  - スケーリングされたデータは最大値が1、最小値は0となる

# データの変換

現実的なデータ行列は一般的に巨大で疎（多くの0要素）

データ行列を情報量をなるべく減らさずより小さくしたい

- 計算コストを削減
- さまざまな分析手法を適用可能になる

データ削減のアプローチ

- データサンプリング
  - レコード数を削減
- 特徴量（属性）選択
  - フィールド数を削減
    - 相関、情報量など
- 行列分解
  - データ行列の次元を削減
    - 特異値分解など
- データタイプの変換
  - 連続値から離散値のようにフィールドの値を変換