

# 2-5 データ加工

東京大学 数理・情報教育研究センター

2021年5月10日

# 概要

- データ加工の基本を学びます
- 数千件～数万件のデータを加工処理するプログラムを作成できるようになることを目指します

# 本教材の目次

1. データ加工とツール	4
2. Excelによる表データ加工	6
3. 計算機で扱うためのデータ加工	13
4. 統計処理のためのデータ加工	17
5. ビッグデータ処理	22
6. まとめ	23
7. 付録	24

# データ加工 (data processing) とは

- 与えられたデータ集合を目的に適った便利な形に変換することです
- 与えられたデータ集合は、そのままでは扱いにくいことが多いです
  - 例えば、Webから表データを収集する際、HTML形式のWebページには、表中の値以外の記号が含まれていて、表データを解析しにくいです
- 単純な統計処理を含みます
  - 統計情報は「便利な形」の1つです

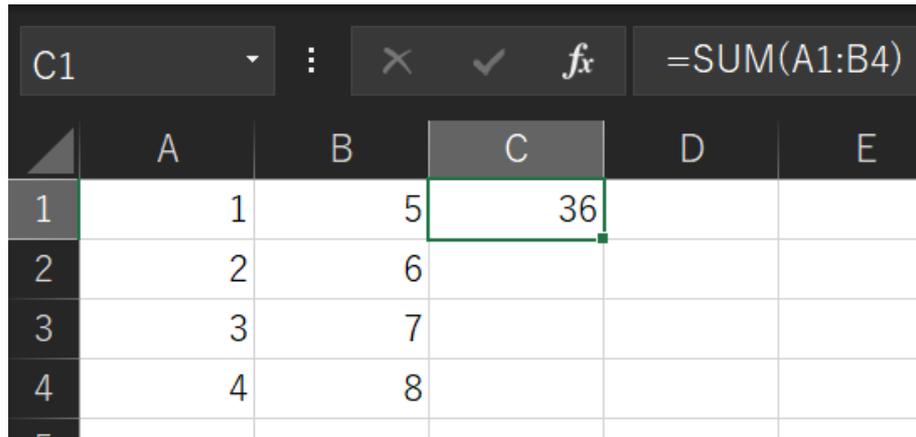
# データ加工のツール

- データ加工に使われるツールは多岐にわたります
- 最良の選択は、入力データの形式や、所望のデータモデル、そこから得たい情報、そのために必要な計算等に応じて変わります
  - しばしば複数のツールをプログラミング環境の下で組み合わせます
  - 万能のツールは有りません
- 表データに限定すればExcelは万能に近いツールです
  - Excelは表計算をするプログラミング環境です
    - 「方眼紙」としての使い方は本来の用途ではありません
  - HTML形式の表データであろうと、自動的に変換してくれます
    - HTMLファイルをExcelで開けば、ページ全体が変換されます
    - Webページ上でコピーして、Excel上で張り付けることもできます
- 以降では、Excelを利用した表データの加工を中心に説明します
  - Microsoft 365準拠

# データ集計

- 総和を入れるセルに =SUM( と入力してから、総和を取るセルの範囲をドラッグで指定し、Enterで確定させます。

• 例：



The screenshot shows an Excel spreadsheet with a formula bar at the top displaying `=SUM(A1:B4)`. The spreadsheet has columns A through E and rows 1 through 4. The values in the cells are as follows:

	A	B	C	D	E
1	1	5	36		
2	2	6			
3	3	7			
4	4	8			

- SUMをAVERAGEに変えると算術平均が計算できます
- STDEV.Pに変えると標準偏差が計算できます

# 四則演算と反復

- セルを参照する計算式を記入できます

- 例：

	A	B	C	D	E
1	1	5	2.5		
2	2	6			
3	3	7			
4	4	8			

- 計算式があるセルの右下の角をドラッグすると反復できます

- 例：

	A	B	C	D	E
1	1	5	2.5		
2	2	6	6		
3	3	7	10.5		
4	4	8	16		

=A4\*B4/2

# 文字列処理

- 文字列を処理する演算子や関数も用意されています
  - 計算式の中で文字列そのもの（リテラル）を書くときには " で囲みます
- & は文字列結合です

• 例：

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	東大	太郎	東大 太郎		
2	John	Doe	John Doe		
3					

The formula bar at the top shows the formula `=A1&" "&B1` for cell C1.

- SUBSTITUTE(文字列, 検索文字列, 置換文字列) は文字列置換です

• 例：

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F
1	データ加工	data加工				
2	データベース	dataベース				

The formula bar at the top shows the formula `=SUBSTITUTE(A1,"データ","data")` for cell B1.

# フィルタとソート

- 表を選択して [データ]»[フィルター] を実行すると、その選択範囲をフィルタ可能になります
  - 選択範囲の1行目は属性名として解釈されます
  - 1行目がドロップダウンになります

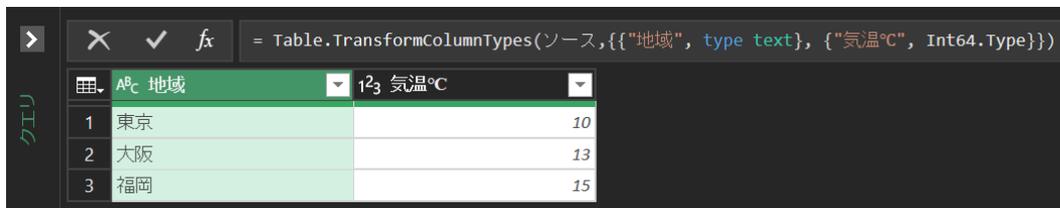
	A	B
1	地域	気温°C
2	東京	10
3	大阪	13
4	福岡	15

- ドロップダウンから、行のフィルターやソートができます



# Power Queryとテーブル

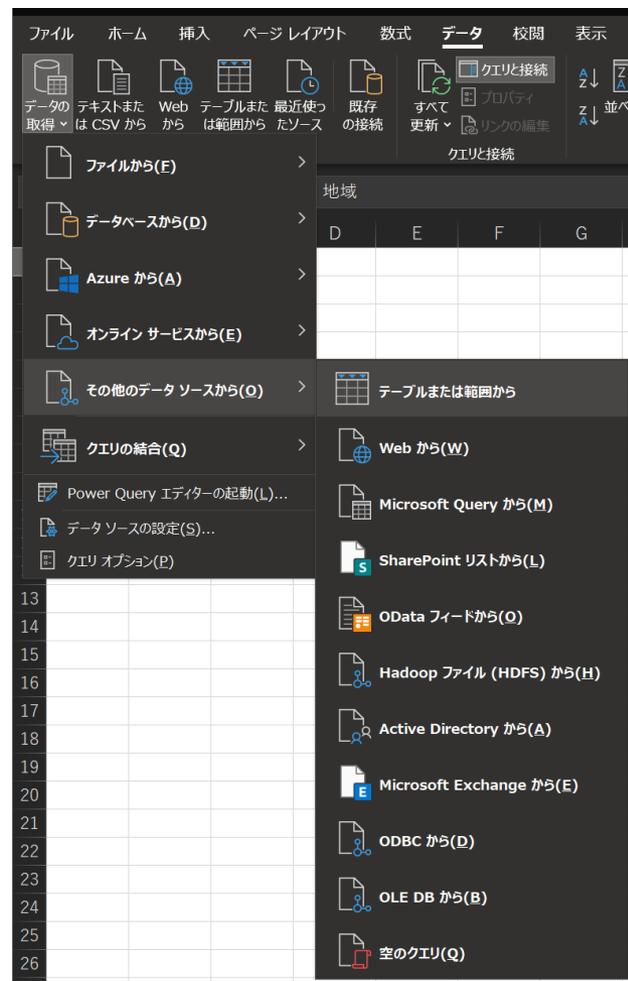
- Power Query機能を使うと表データに対する演算が簡単にできます
  - テーブルと呼ばれる単位で演算します
- [データ]»[データの取得]
  - »[その他のデータソースから]
  - »[テーブルまたは範囲から]で選択範囲をテーブルに変換できます
- Power Queryエディターを[閉じて読み込む]と別のシートに変換後のテーブルが保存されます



The screenshot shows the Power Query Editor interface. The formula bar at the top contains the M code: `= Table.TransformColumnTypes(ソース,{{"地域", type text}, {"気温°C", Int64.Type}})`. Below the formula bar, a table is displayed with two columns: '地域' (Region) and '気温°C' (Temperature in °C). The table contains three rows of data:

	地域	気温°C
1	東京	10
2	大阪	13
3	福岡	15

- [クエリと接続]でテーブル一覧が見えます



# テーブルの結合

- [データの取得]»[クエリの結合]»[マージ]から異なるテーブルを結合して新しいテーブルを作れます
- 結合には複数種類あります
  - 内部：照合列が一致する行だけの結合
  - 左外部：上のテーブル全てと下のテーブルから照合列が一致する行の結合
  - 右外部：下のテーブル全てと上のテーブルから照合列が一致する行の結合
- 外部結合における不一致行のセルは適当にnullで埋まります

マージ

マージされたテーブルを作成するには、テーブルと照合列を選んでください。

テーブル1

地域	気温°C
東京	10
大阪	13
福岡	15

テーブル3

地域	人口(千)
東京	13960
京都	2562
福岡	5119

結合の種類

左外部 (最初の行すべて、および 2 番目の行のうち...)

あいまい一致を使用してマージを実行する

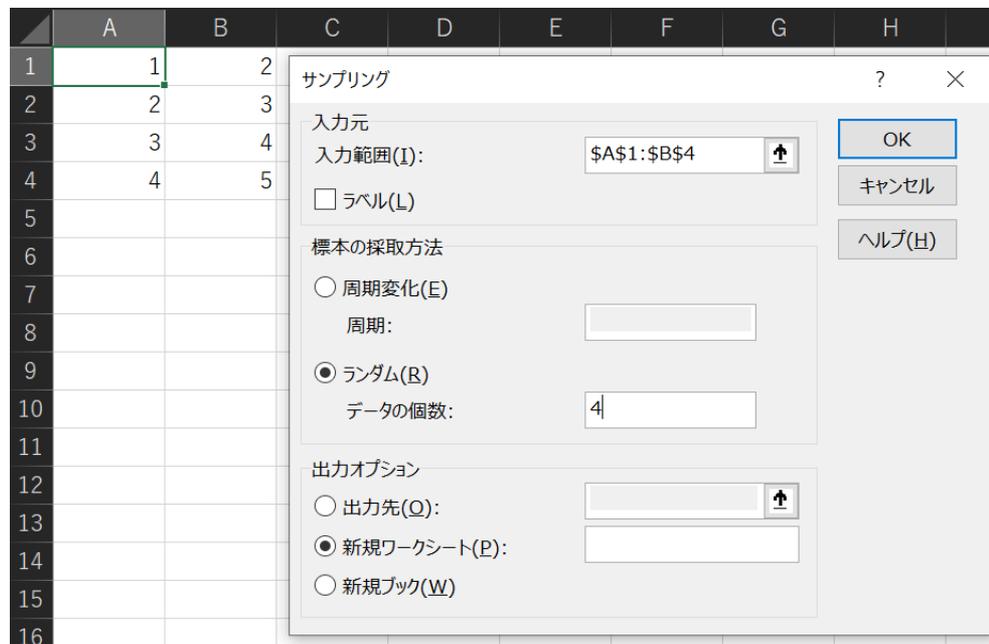
▶ あいまい一致オプション

✓ 選択範囲では、最初のテーブルと 3 行中 2 行が一致しています。

OK キャンセル

# 分析ツールと統計処理

- アドインの分析ツールを使うと統計処理が簡単にできます
  - [ファイル]»[オプション]»[アドイン]»[Excelアドイン]の[管理]»[分析ツール]のチェックを入れて[OK]をすると、アドインを有効にできます
    - Cf. [Excel で分析ツールPak を読み込む](#)
  - [データ]»[データ分析] という項目が追加されます
- 利用例：ランダムサンプリング
  1. [データ分析]»[サンプリング]
  2. 入力範囲を指定
    - セルのドラッグで指定可能
  3. 採取するデータの個数を指定
  4. 出力先を指定
    - デフォルトは新規シート



# 整然データ (tidy data)

- 1行が1つの観測（データ点）に整形された表形式のことです
  - 全ての行が同一の型（定数個の属性の組）となります
  - 整然データでないものを雑然データ（messy data）と呼びます

- 例：整然データ

地域	時間帯	気温°C
東京	午前	10
東京	午後	12
大阪	午前	13
大阪	午後	16

- 例：雑然データ

地域	気温°C (午前)	気温°C (午後)
東京	10	12
大阪	13	16

# 整然データの利点

- 一見すると雑然データの方がコンパクトで見やすく思えます
  - 実際、前頁の例において整然データの方がスペースを多く割いています
- 整然データはフィルタやソートと併用する際に扱いやすいです
- 前頁の例について言えば
  - 同一地域や同一時間帯の観測に限定した表がフィルタで手に入ります
  - 最も気温の高い（低い）地域と時間帯が単純なソートで手に入ります
  - 同一時間帯の平均気温と同一地域の平均気温が殆ど同じ式で計算できます
  - 時間帯を追加したり細分化する際に、表の形を変えなくて済みます
- 整然データは計算機で処理するための形式です

# 名寄せ

- 概念や対象に対する表記を統一することです
  - 例えば「東京」と「東京都」は表記は別ですが同一地域です
  - また「學問のすゝめ」と「学問のすすめ」は同一著作物です
  - 一方、同姓同名の別人には適宜名前を付けて統一的に区別します
- 元々は、複数の銀行口座に渡って同一顧客を追跡して一元管理することを意味していましたが、今ではより広い意味で用いられます
- 適切に名寄せをしないと、加工結果から得る情報が不正確になります
- 表記の同一性に基づいて処理できないデータ形式は、機械的な処理に不向きなので、名寄せは実用上重要です
  - データベースにおける主キーの付与にほぼ対応します
- 異なるデータセットを統合するデータマッピングにおいて、名寄せは中心的なタスクです
  - データマッピングツールは名寄せを支援します

# データ型変換

- 概念に対して特定の値を関連付けることがあります
  - 例えば、日時の表記法は様々ありますが、計算機ではUNIX時間で統一的に時刻を表現することが一般的です
  - UNIX時間：1970-01-01 00:00:00 からの経過秒数
- 文字列型のデータを所定のデータ型に変換する必要があります
  - UNIX時間の場合は符号なし整数への変換です
- 地理オブジェクト名に対して、地理座標（典型的には緯度経度）を付与するジオコーディングも、データ型変換の一種です
  - 文字列の名前から、地理座標を伴った文字列への変換です

# データクレンジング

- 汚れたデータを特定して除去・訂正することです
    - 汚れの例：破損したデータ，不正なデータ，例外的なデータ
  - 統計処理の観点では，外れ値・異常値・欠測値への対処が主です
- 
- ◆ 外れ値：他の測定値から大きく外れた値
    - 観測上の異常です
    - 一般に原因は分かりません
  - ◆ 異常値：外れ値の中で異常さの原因が解明されているもの
    - 本質的な異常です
    - 誤りや故障によって，異常であるべくして異常になったものです
  - ◆ 欠測値：測定データに含まれていない，本来測定されているべき値
    - 例えば，アンケートの「無回答」や記入漏れなどで生じます

# 外れ値の検定

- 外れ値を特定する検定には様々な方法があります
  - 単純かつ代表的な検定は次の2つです
- ◆ 第1と第3四分位数の区間から外れる
    - つまり上位25%と下位25%を外れ値と見做します
  - ◆ 平均からの乖離が $2\sigma \sim 3\sigma$ に達する
    - $\sigma$ は標準偏差を意味します

# 欠測値への対処

- 欠測値は統計処理を不可能にします
  - 欠測値が入ると算術平均すら定義されません
    - 算術平均に基づく統計量が全て定義されません
- 欠測値への典型的な対処法が2つあります
  - ◆ 欠測値を含むケースを丸々除去する
    - つまり、欠測したものは存在しないものとして扱う
  - ◆ 欠測値に非欠測ケースの平均値を代入する
    - つまり、欠測値は平均値に対して無害な値であると仮定する
- これらは統計処理を可能にしますが、統計量（統計モデル）に対してバイアスを生みます

# データの正規化

- 統計量を扱いやすくする幾つかの正規化があります
  - 分布の性質を保存しつつ、データ固有の大きさを捨象します
- 以下では、 $X'$ をデータセット  $X$ を正規化したものとしします
- ◆ Min-max正規化：最小値が0で最大値が1になるよう変換する
  - min-max正規化は  $X' = \left\{ \frac{x - \min(X)}{\max(X) - \min(X)} \mid \forall x \in X \right\}$  で構成できます
- ◆ 標準化：平均が0で分散が1になるよう変換する
  - 標準化は  $X' = \left\{ \frac{x - \text{avg}(X)}{\sigma} \mid \forall x \in X \right\}$  で構成できます
  - ただし、 $\text{avg}(X)$ は $X$ の平均値で、 $\sigma$ は標準偏差です
- 異なるデータの統計量を合算する機械学習において特に有用です

# ダミー変数

- 回帰分析では、独立変数 $X$ と従属変数 $Y$ との間にある定量的な関係を統計的に推定します
  - このとき $Y$ は連続値です
- $X$ が離散的な値を取るとき、そのままでは $X$ に関する関数（例えば線形変換）として $Y$ を定義できません
  - 例えば、 $X$ が曜日を表す場合、数値として演算可能ではありません
- $X$ の定義域 $\text{dom}(X)$ が有限なら、 $x \in \text{dom}(X)$ は2値変数の組  $x_1, \dots, x_k$  で表現できます
  - $x = i \Leftrightarrow x_i = 1 \wedge x_j = 0$ （ただし $j \neq i$ かつ $k = |\text{dom}(X)|$ ）と定義します
- この2値変数  $x_1, \dots, x_k$  を、それぞれ独立変数  $X_1, \dots, X_k$  と見做すことで、関数の形で $Y$ が定義でき、重回帰分析が適用できるようになります
- この $X_1, \dots, X_k$ をダミー変数と呼びます
- ダミー変数の導入は離散データに対する一種の正規化です

# ビッグデータ処理

- ビッグデータとは，伝統的なデータ加工ツールで容易に扱えないほど大きいデータを指します
  - 伝統的なデータベース管理システムに載らない
  - 勿論Excelにも載らない
- 典型的にはビッグデータは1つの計算機のメモリに載りません  
→ 分散データ処理
- ビッグデータ処理ツール
  - Apache Spark [Zaharia et al. '14] <https://spark.apache.org/>
    - 分散データセットを手軽に扱える多目的プログラミング環境
  - Apache Hadoop <https://hadoop.apache.org/>
    - 分散ファイルシステム上の大規模データ処理基盤の古典
    - Sparkへの代替が進んでいる
  - HillView [Budiu et al. '19] <https://github.com/vmware/hillview>
    - Excelのような表計算ツールの分散処理版

# まとめ

- Excelによる表データ処理の基本を紹介しました
- 対象データを計算機で扱いやすくしたり，統計処理を扱いやすくするための様々なデータ変換や正規化を紹介しました
- Excelは強力なプログラミング環境です
  - 表データの処理ならExcelを使いこなせば大抵大丈夫です
  - 使いこなすには相応のプログラミング能力が求められます

# 付録：正規表現

- 文字列集合を記述する言語です
  - 例1： $a^*b$  は「aが任意個連続してbが続く」文字列全体
  - 例2： $(a|b)^*$  は「aかbが任意個連続する」文字列全体
- 文字列の検索や置換などに使われます
  - テキストエディタやコマンドラインツールの中でしばしば使われます
- 正規表現の基本演算
  - 接続： $ab$  (aとbが連続する)
  - 選択： $a|b$  (aもしくはb)
  - 閉包： $a^*$  (aが0回以上反復する)
- 基本演算に加えて様々な拡張が存在します
  - 個々のツール (grepやsedなど) 毎に, 具体的な記法が変わります
- 正規表現の使い方をきちんと理解するにはプログラミング言語と共に学びましょう
  - Cf. Pythonにおける [正規表現HOWTO](#)