2-3 データを扱う

東京大学　数理・情報教育研究センター

2020年5月11日
2024年4月30日改訂

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

概要

- データを解析するためのツールの一つであるスプレッドシートを理解し、和や平均の計算などの基本的な使い方を学びます。

- また、データを扱うファイル形式としてよく用いられるcsvファイルについて理解し、スプレッドシートを用いて小規模データ（数百件～数千件レベル）を集計・加工できることを目標とします。

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

本教材の目次

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

データの取得（機械判読可能なデータ）

- データ分析を行うためのデータ収集では、データの正確性に注意するとともに、機械判読可能な形式のデータを取得する必要があります。

- 機械判読可能なデータとは、コンピュータが直接解析や処理が可能な形式で提供されているデータのことを指します。このようなデータは、人間が読むための文書とは異なり、データ分析やAIの学習など、さまざまな自動処理に直接利用することができます。

- 機械判読可能なデータ形式の代表的なものとして、後述のcsv形式のデータが挙げられます。

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

表計算ソフト

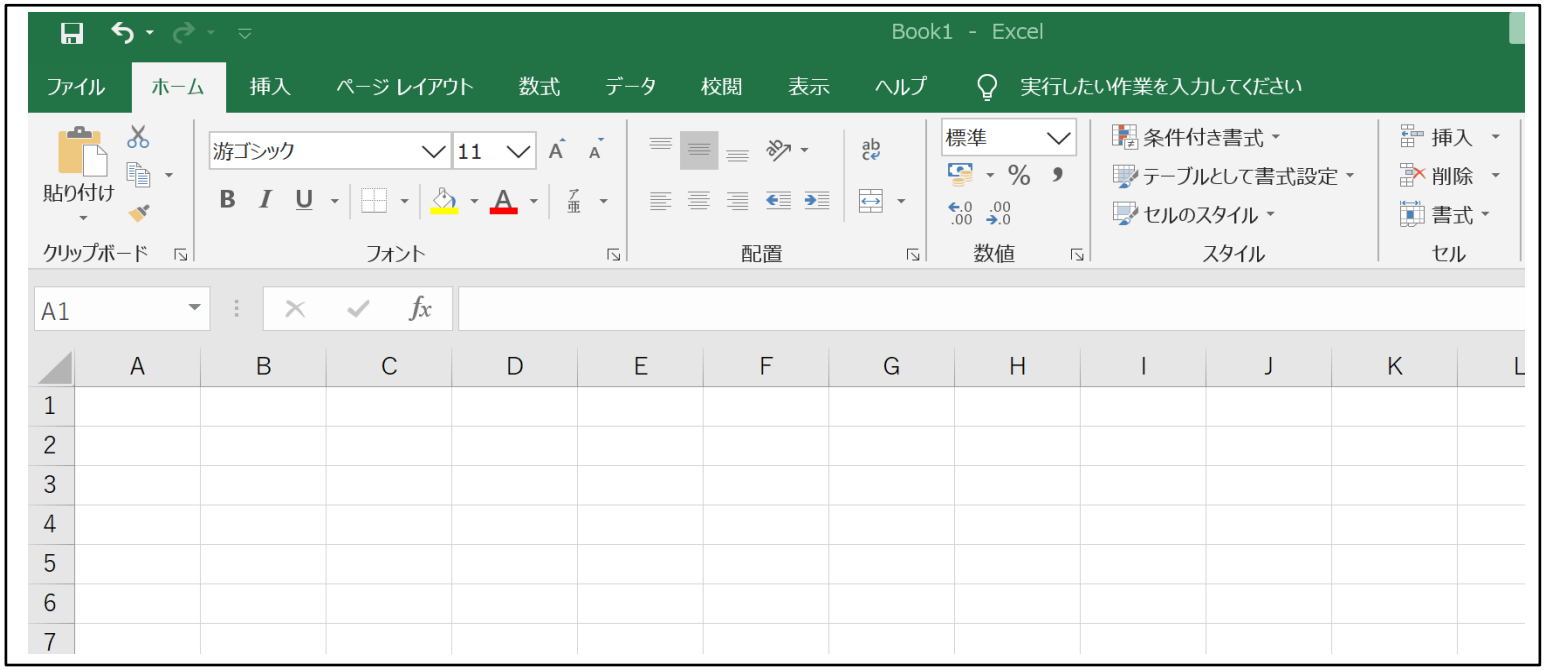
- データを扱う際には、通常は表計算ソフトを用います。
- 表計算ソフトは、表形式のアプリケーション上でデータの入力・表示・集計等の操作を行うことができるソフトウェアです。
- 代表的なものとしてMicrosoft社のエクセルやGoogle社のスプレッドシート等があります。

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

エクセル

- エクセルはMicrosoft社の提供している有償の表計算ソフトであり、代表的な表計算ソフトとして広く用いられています。
- パソコン上にインストールして使用するソフトウェアで、数年おきに新しい機能が追加されたバージョンが販売されています。

＜エクセルの画面例＞



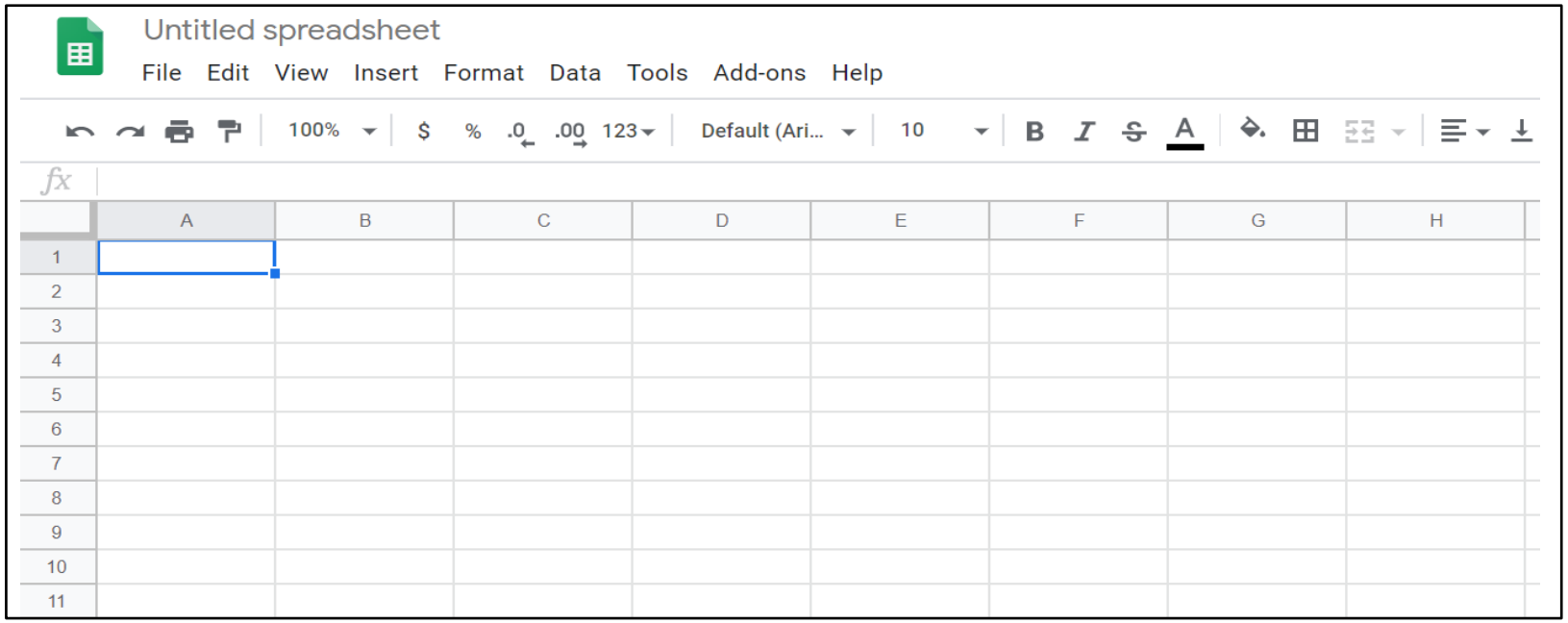
※マイクロソフトの許諾を得て使用しています。

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

Googleスプレッドシート

- Googleスプレッドシートは、Google社が無償で提供している表計算ソフトで、様々なデータに対する操作を行うことができます。
- インターネットを介して使用するWebアプリケーションであり、パソコン上にソフトをインストールする必要がありません。
- 以下ではGoogleスプレッドシートを用いた計算例を見ていきます。

＜Googleスプレッドシートの画面例＞



東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

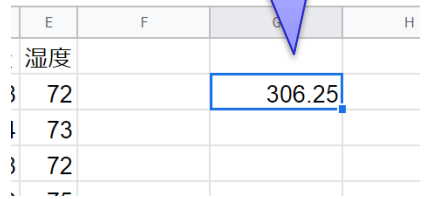
データに対する操作（和、平均）

- <u>和をとる</u>
  - 和をとる範囲を指定してsum関数を用いて計算します。
  - 例えば、右のデータで7月の日射量の合計を計算するときは、適当なセル上で
    =sum(D2:D32)
    等と打つと和を計算します。
- <u>平均をとる</u>
  - 平均をとる範囲を指定してaverage関数で計算します。
  - 右のデータで7月の平均気温を計算するには、
    =average(C2:C32)
    と打ちます。

＜気候データ＞

	A	B	C	D	E
1	月	日	気温	日射量	湿度
2	7	1	28	14.08	72
3	7	2	28.7	14.54	73
4	7	3	28.4	13.18	72
5	7	4	27.5	7.82	75
6	7	5	26.7	3.49	80
7	7	6	21.6	1.42	103
8	7	7	24.9	8.11	81
9	7	8	27.3	10.48	77
10	7	9	27.2	9.08	89

=sum(D2:D32)
と打つと日射量の
合計が計算される

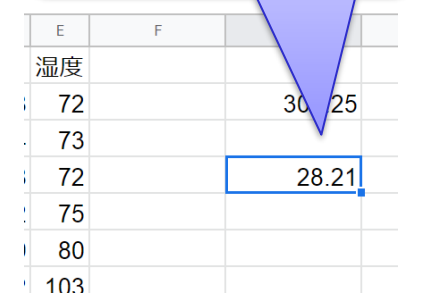


D2からD32の値の和を計算

＜気候データ＞

	A	B	C	D	E
1	月	日	気温	日射量	湿度
2	7	1	28	14.08	72
3	7	2	28.7	14.54	73
4	7	3	28.4	13.18	72
5	7	4	27.5	7.82	75
6	7	5	26.7	3.49	80
7	7	6	21.6	1.42	103
8	7	7	24.9	8.11	81
9	7	8	27.3	10.48	77
10	7	9	27.2	9.08	89

=average(C2:C32)
と打つと気温の平均
が計算される



C2からC32の値の平均を計算

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

データに対する操作（並べ替え、ランキング）

- データ全体を選択し、上部タブの「データ」→「範囲を並べ替え」を選択することでデータのある項目を基準に並べ替えることができます。
- また、並べ替え後のデータに対して番号を入力することでランキングを作成することもできます。

＜数学と国語の点数データ＞

	A	B	C	D	E
1	学生番号	文理	数学	国語	
2	1	理系	93	46	
3	2	理系	48	50	
4	3	文系	41	64	
5	4	文系	28	31	
6	5	理系	75	23	
7	6	文系	68	42	
8	7	理系	63	24	
9	8	理系	19	51	
10	9	文系	39	56	
11	10	理系	60	3	
12	11	文系	12	36	
13	12	理系	80	16	
14	13	理系	15	22	
15	14	理系	1	54	

数学の点数で並べ替え

数学の点数に関するランキングを作成

＜数学の点数ランキング＞

	A	B	C	D	E
1	学生番号	文理	数学	国語	ランキング
2	208	理系	98	59	1
3	209	理系	96	9	2
4	1	理系	93	46	3
5	116	理系	92	57	4
6	212	理系	92	69	5
7	235	理系	91	6	6
8	98	理系	89	32	7
9	34	理系	88	55	8
10	49	理系	87	6	9
11	65	理系	87	61	10
12	165	理系	86	29	11
13	179	理系	85	4	12
14	20	理系	84	33	13
15	156	理系	82	12	14

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

表形式のデータ（csv）

- データを扱うファイルの形式としては、csvと呼ばれる形式のファイルが最もよく使われます。
  - csvファイルはカンマ（, ）で区切られたデータで、スプレッドシート等の表計算ソフトでは、表の形のデータに変換されます。
  - csvファイルは様々なソフトウェアに対応している標準的なファイル形式です。

＜csv形式のデータ＞

```
月, 日, 気温, 日射量, 湿度
7, 1, 28, 14.08, 72
7, 2, 28.7, 14.54, 73
7, 3, 28.4, 13.18, 72
7, 4, 27.5, 7.82, 75
7, 5, 26.7, 3.49, 80
7, 6, 21.6, 1.42, 103
7, 7, 24.9, 8.11, 81
7, 8, 27.3, 10.48, 77
7, 9, 27.2, 9.08, 89
```

スプレッドシート上の表現

	A	B	C	D	E
1	月	日	気温	日射量	湿度
2	7	1	28	14.08	72
3	7	2	28.7	14.54	73
4	7	3	28.4	13.18	72
5	7	4	27.5	7.82	75
6	7	5	26.7	3.49	80
7	7	6	21.6	1.42	103
8	7	7	24.9	8.11	81
9	7	8	27.3	10.48	77
10	7	9	27.2	9.08	89

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA

ビジネス・インテリジェンス・ツール

- ビジネス・インテリジェンス(BI)ツールとは、企業が保有する、顧客データ、ソーシャルメディア、ウェブサイトの分析データなど、多様なソースからの膨大なデータを分析して有益な情報を得るためのソフトウェアやサービスのことです。

- 表計算ソフトも最もシンプルなBIツールの一種と考えられ、企業が保有するデータからデータ分析、データの可視化、レポート作成などを行うことができ、意思決定をサポートする情報を得ることができます。

- より大規模もしくは複雑なデータセットを扱うための、より専門的なBIツールも様々なものが提供されています。

東京大学 数理・情報教育研究センター 荻原哲平 2020 CC BY-NC-SA