データ分析基礎ｃ（情報編集基礎ｃ）

第１回 データの特性とソフトウェア

1. データの種類と分類

データには，様々なものがある。また，その分類の方法も様々な視点が存在するため，一律の基準がない。したがって，一般的な分類は困難ではあるが，ここでは，考えられる分類の基準をいくつか提示して，その基準に基づいたデータの内容について考えてみよう。

1.1. 媒体の違いからみた分類

インターネットの普及により，ネットワークを使ったデータ収集もごく当り前になってきた。しかし，希少なデータ，公開されることを望まない個人情報，刻々と変化する意識，付加価値が高いために入手に代価を要するもの，など，ネットワークには現れることのないデータも依然として存在する。

それらのデータには，ある場合には，別の形で入手できるものもある。たとえば，文献・雑誌・新聞のような**印刷物**の形であるもの，（ネットワークには現れないが）**コンピュータ用のメディア**に記録されたもの，コンピュータ用に意図されてはいないが**電子化されたメディア**に記録されたもの，人の意識や行動パターン・（データ化される前の）自然現象などのように，**物理的な形を持たないもの**，というように分類することは可能であろう。

このような分類は，**コンピュータ処理**との相性といった観点であるとも言える。当然のことながら，コンピュータ用に意図されたものは，そうでないものに比べると扱いやすい。しかし，そうでないものについても記録形式を変えたり，電子化されていないものについても専用の装置を用いたりすることで，コンピュータ処理が可能な形に変換することは可能である。

1.2. データの尺度基準からみた分類

次に，コンピュータ処理に向くデータに限って考えることとしよう。現在のコンピュータは**ディジタル・コンピュータ**であるので，コンピュータ内部のデータはすべて**ディジタル・データ**，すなわち数値表現されたデータである，と言ってよい。これは，文字のような情報でも図形や映像の情報であっても，すべて記号化（あるいはコード化）されているためである。

このように，コンピュータ内部では，すべてのデータが数値化された情報なのであるが，その数値の持つ意味あい（あるいは特性）によって次のような分類をすることがある。このような分類は**尺度基準**による分類と呼ばれており，数値を客観的に扱う分野，すなわち統計学では当然の考え方として広く受け入れられている。

- **名義尺度**
  他の数値との差異のみが重要であるもの
  例）スポーツ選手の背番号 学籍番号
- **順序尺度**
  序列関係（大小関係）のみが重要であるもの
  例）徒競争における順位
- **間隔尺度**
  順序尺度の性質に加えて差の大きさに意味があるもの
  例）（摂氏あるいは華氏による）温度
- **比例尺度**
  間隔尺度に性質に加えて比が意味を持つもの
  例）（一般的に）物理量（長さ・重さなど）

1.3. ファイル形式の違いによる分類

最後に，コンピュータのソフトウェアからみた分類の視点について触れておこう。経験的には，「ワープロのデータは表計算ソフトでは使えない」といった現象のことである，といったら了解できるであろう。

このような現象が起こるのは，それぞれのソフトウェアが必要とするデータが異なること，データの表現については標準化された基準がなくソフトウェアのメーカーやプログラマの裁量に負うところが大きいこと，などが主な理由である。

もちろん，メーカーの異なるコンピュータやソフトウェアどうしでデータのやりとりをする必要があるので，最低限の基準はある。その一つが，**規格化された文字コード**である。大型汎用コンピュータで使用されている EBCDIC やパソコンなどで一般的に使用されている ASCII が，これに相当する。これに加えて，英語圏以外の独自の文字（日本語，中国語，韓国語など）についても規格化されており，日本で使用される漢字は JIS（日本工業規格）で定められたものをベースに，Shift-JIS（Microsoft 漢字），EUC-JIS（Extended Unix Code），Unicode（UTF）などの符号化（Encoding）が使用されている。データ表現をこれらの文字コードのみで表現すれば，原理的にはどのコンピュータでもどのソフトウェアでもデータの相互利用が可能となる。そのような文字表現のみで構成されるデータを**テキスト（形式）データ**と呼び，コンピュータ用の記録メディア（フロッピーディスクやハードディスク，USB フラッシュメモリ，SD カードなどの外部記憶装置）に保管される場合には，**テキスト（形式）ファイル**という呼び方をする。

また，コンピュータの能力が高まったことで，画像，映像（動画），音声のデータも標準化の動きが活発である。これらの情報をテキスト形式で表現することもあるが，無駄な情報を多く含むことになってデータ量が増加してしまうので，通常は，ある決まった方式で

コード化される。代表的なものには，静止画像として JPEG/JFIF，GIF，PNG，など，動画像として，MPEG など，音声データとして，MP などがある。これらのデータは，コンピュータの記憶量の基本単位である**ビット**（bit）あるいは**バイト**（byte）単位で，その符号化の方式が規定されており，総称して**バイナリ（形式）データ**と呼ばれる。

これ以外のデータ表現についても，標準化の動向はあるものの，メーカー間の思惑や独自性の主張のようなことから，実際には，ある特定のソフトウェアでしか利用できないデータ形式が用いられることが多い。このようなデータ（Word，Excel，PowerPoint のオリジナルデータなど）も，その多くはバイナリ（形式）データである。

2. データの種類とソフトウェア

前節で概観したようなデータの種類についての知識（あるいは意識だけでもよい）は,「あるデータを分析したい」という場合に「分析の意図にあった情報か？」「データの記録形式は？」「利用可能なソフトウェアはどれか？」といった意識につながる。そして，この意識は「データ分析基礎（情報編集基礎）」という科目で学ぶべきことがらの最も重要なことがらであるといってよい。

授業では，こうした意識を常に持ちつつ，手軽ではあるけれどもあまり活用されることのないソフトウェアのひとつである**表計算ソフト**によるデータ分析のすすめ方について学習してもらいたい。

また，自己満足で終わるならばともかく，通常の場合，データ分析の結果は，第三者へ正確に伝えることで，その目的が完結する。多用される手法としては，グラフ表現が挙げられるであろう。しかし，その表現様式・データの準備方法については，データの性質や出来上がりの効果 ― つまり伝達したい内容 ― をよく考えた上で選ばないと，時には逆効果にさえなってしまう。したがって，履修の目標として，グラフ表現の技術向上やレポートとしてのまとめ方（**ワープロによる文章構成**）の練習，口頭発表に役立つツール（**プレゼンテーション支援ソフト**）についても言及していくこととする。

さらに，近年注目を浴びている，**ビッグデータ**や**オープンデータ**の解析には，統計処理の専用ソフトによる処理技法を身につけておく必要がある。

これらのソフトウェアの機能を，より深く理解することで，データの収集・ソフトウェアの使い分け・データの加工（分析）・分析結果のまとめ方，といった一連のデータ処理の流れを身につけることが重要である。