第5章　データ分析入門

この章について

この章では、データの分析について解説します。データの分析というと少々敷居が高いように思われるかもしれません。しかし、データの収集とその分析は、社会科学・人文科学・理学・工学を問わずほぼすべての学問分野で、また広く社会で利用されている一般的な技術です。

データの分析を裏打ちするのは「ルベーグ積分」に基づく確率論などの高等数学です。しかし、ここでは分析の道具として利用することを目的にしています。上で「技術」と述べたように、どのような学生であっても教養として身につけているべきデータの分析手法を学習しますので、実際に目指すのは分析方法の選択と分析結果の解釈ということになります。この際、コンピューターを利用することで複雑な統計的計算の詳細な手順を追うことは避けますが、興味を持った学生諸君は参考文献などに当たって学習を深めることをお勧めします。

実際のデータ分析ではデータそのものの収集やデータハンドリング（データの加工）なども、分析の重要な一部です。また、これを自分の展開する議論の論拠として利用するためには分析結果を解読し、グラフや表のような分かりやすい形で表現する必要もあります。

以上のような考え方に従って、この章ではコンピューターがコンピューターと呼ばれる所以である機能、つまり「計算」についての、基礎的な考え方を学習します。

5.1　データの種類

分析しようとしている対象であるデータには、様々な種類のものがあります。このようなデータを数値として分析しようとするとき、これに当てはめるモノサシのことを尺度といいます。

例えば、データは次のようなものである可能性があります。

- 英語の小テストの得点
- 好きな曜日
- 郵便番号
- A 市と B 市の合併に対する賛否
- 長さ
- 重量
- 気温
- 価格（円、ドルなど）

これらはいずれも、何かの特徴を表すものです。これらを大きく 2 つに分けると、曜日や賛否のように大小や順序は関係のない**質的データ**と、大小や順序がある**量的データ**があることがわかります。郵便番号のように、数だからといって量的データではなく質的データのものもあることに注意しましょう。この場合、数値ではなく「数字」と呼びます。

また、データそのものにより分類可能なデータ（得点、曜日、賛否、価格など）と、ある範囲を定めて分類することになるデータ（長さ、重量、気温など）があります。

前者は観測される値そのもので分類できますが、例えば性別で女を 1 に、男を 2 にするなど数値を対応させる（これをコーディングと言います）こともできます。

後者については、当てはめるモノサシの精度によって観測される値が変わることに注意してください。例えば長さを計測する場合、最小の目盛りが 1mm のモノサシを利用するのと 0.1mm のモノサシを利用するのとでは、計測される結果が変わります。これは、仮にデジタルのモノサシや温度計を利用しても同じことで、この場合はセンサーの精度によって測定される値が変わります。つまり、この場合はどのような手段で測定したとしても真の値を計測することはできず、近似値しか得られないということになります。

このように、近似値しか得られないか、それともそのものズバリの値を得られるかでデータを分類することも可能です。長さや重量、温度は連続量であり、得点は離散量です。

これらのデータを、尺度という基準でとらえ直して考えることにしましょう。ここで尺度とは、「データに値を対応させる基準」と定義しておきます。尺度には、名義尺度、順序尺度、間隔尺度、比尺度があります。言い換えると、これらはデータの特性を意味しており、どの尺度であるかによって図示の仕方（グラフの作成方法）も異なり、また計算することのできる統計量も変わることに注意しましょう。

5.1.1　名義尺度（Nominal Scale）

名義尺度は、値の順序や値同士の差が意味を持たないというデータの分類基準です。例えば、郵便番号同士を引き算しても、何も意味のある結果を得ることはできません。データが文字で表される場合もありますし（例えば性別など）、仮に数で表現されていたとしても、数値というよりは数字としての性格を持ったデータということになります。このような尺度を、名義尺度といいます。

5.1.2 順序尺度（Ordinal Scale）

値の差に意味はないものの、順序には意味があるというのが順序尺度です。アンケートによく見られるような、(1) 悪い (2) やや悪い (3) どちらでもない (4) やや良い (5) 良いなどのようなものです。

これらは本来データが等間隔で並んでいることを仮定することはできません。上の例で、選択肢である数値である 5 を 1 で割って、5 は 1 より 5 倍良いとか、差を取って 4 の差があるとは言えないということです。また、例えば好き嫌いを 5 段階に分類して、これを-2、-1、0、1、2 というように表現しても、1、5、15、25、55 というように表現しても、意味としては同じことです。もっとも、通常は 1 からの連続値（1、2、3、4、5 など）を割り当てます[1]。

教育学や心理学、マーケティングなどの分野では、これらのデータの間隔が一定であることを仮定して分析を行うことがしばしばあります。つまり、5.1.3 「間隔尺度」で述べる、間隔尺度であるとみなして分析する場合があります。

順序尺度では平均値を計算することにはあまり意味はありませんが、中央値（小さい方から値を並べて真中になる値）には意味があります。

5.1.3 間隔尺度（Interval Scale）

温度のように、間隔に意味があるというデータです。データの間隔に意味があるということは、つまり差に意味があるということです。例えば 100 度と 80 度では、20 度の差があるということになります。ただし、この場合 80 度は 100 度の 5 分の 4 である、という意味にはならないことに注意しましょう。

温度は、例えばセ氏の場合、氷点を 0 ℃、沸点を 100 ℃というように、人為的な基準によって相対的に測定した値に過ぎません。言い換えれば、0 ℃という基準点に数値という観点からは絶対的な意味はないのです。例えば、100 ℃は 0 ℃の何倍の熱さであるかを考えてみれば容易に理解ができるでしょう。このように、厳密に言えば間隔尺度では値と値の比に意味がないということになります。

間隔尺度については平均値や標準偏差など、主要な統計量のほとんどを計算することが可能です。

5.1.4 比尺度（Ratio Scale）

比尺度は長さや重量など、ゼロにも意味がある（つまり値と値の比に意味がある）ようなデータを言います。例えば、10kg は 5kg の 2 倍の重量があると言えますし、5kg の差があるとも言えます。

比尺度についても、主要な統計量のほとんどを計算することが可能です。

5.2 Calc の基礎と基本構成

ここでは、数値データの集計や分析に良く用いられている表計算ソフト（スプレッドシート）の簡単な使い方を説明します。表計算ソフトによって、数値や文字を入力してこれを集計などして表にまとめ、あるいはグラフとして視覚化することができます。また、「関数」を用いて計算を行うことや、同じ計算を繰り返し行うのに便利な機能を持っていることから、特にビジネスの分野では広く用いられています。

[1] このようにすることで、順位相関や順位和検定などの「ノンパラメトリック」な統計手法を利用することができます。

5.2.1　行と列、セル

ここでは、OpenOffice.org Calc（以下 Calc）の操作の基本を学習しながら、前節で学習した各種データを実際に入力し、表を作り、グラフを作成してみることにしましょう。まず、Calc を起動してください。

「表計算ソフト」とも呼ばれるスプレッドシートソフトウェアですが、どのソフトウェアにも共通しているのが、行と列、そしてセルと呼ばれるものです。

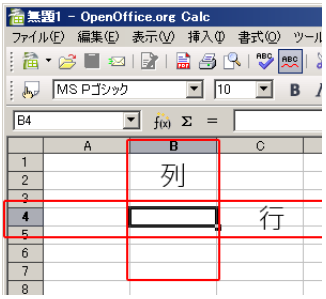

図 5.1: 行と列、セル

スプレッドシートの基本は、セルです。スプレッドシートには、行と列があります。セルとは、この行と列が交わったところをいいます。これをカレントセルといいますが、現在編集の対象になっているセルを意味します。

さて、たくさんあるセルを区別するには、セルに住所のようなものを割り当てます。これを「セル番地」ないし単に「番地」といいます。これは、列文字と行の数を組み合わせて使います。例えば、一番左で一番上のセルは、行が「A」で列は「1」ですから、「A1」という番地が割り当てられています。同様に A2、A3、B1、B2 などのセルがあります。なお、Calc では 256 列× 65,536 行を扱うことができます。A〜Z 列の次は AA〜AZ 列、BA〜BZ 列などの列があり、最後は IV 列までがあります[2]。カレントセルの番地は、左上に「B4」と表示されています。

このセルが、スプレッドシートの基本です。すべてのデータはここに記入されます。

5.2.2　カレントセルの移動、データの入力と編集

データを入力するためには、まずカレントセルを自分の好きな場所にもっていく必要があります。いくつか方法がありますが、もっとも簡単な方法は矢印キー（↑ ↓←→）を利用することです。矢印キーを押すと、カレントセルが移動します。

好みの場所にカレントセルを移したら、好きな文字を入力してみてください。データーを入力することができます。データは文字でも数字でもかまいません。入力したら、エンターキーを押します。このエンターキーを押すことで、入力が確定されます。

入力されたデータが文字の場合、左寄せで表示されるはずです。これが数字の場合、右に寄せて表示されます。また、一定の形式のデータを入力した場合、特別な扱いを受けることもあります。例えば、「10/1」と入力すると、10 月 1 日だと（ある意味では勝手に）解釈して、そのような形式で取り扱いを受けます。

入力の確定には 3 種類ほど方法があり、その方法によってカレントセルの動き方が変わります。まず、キーボードのエンターキーを押した場合、カレントセルは 1 つ下に動いているはずです。次に、

[2]Microsoft Office2007 ではより広範囲な行と列を利用することができるなど、この制限はソフトウェアによってまちまちですが、表計算ソフトは、あまり巨大なデータを操作するのには向きません。

タブキーを押した場合は、カレントセルは 1 つ右に動いているはずです。最後に、テンキーの エンターキー（普通はキーボードの一番右下にあるキー）を押した場合、カレントセルは移動しません。

スプレッドシートを扱っている場合、同じ方向（右とか下とか）にデータを入力続けていくことが多いため、覚えておくと便利です。

なお、カレントセルの移動はマウスでも行うことができます。単にクリックすれば、そこがカレントセルとなります。また、一定の範囲をドラッグすると、その範囲のセルの色が反転します。この状態で入力をすると、セルはその選択された範囲内のみを行き来するようになります。

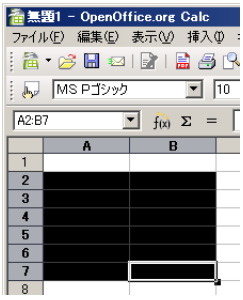

図 5.2: 複数のセルを選択した状態

図 5.2 では、A2 から B7 までの範囲が選択されています。カレントセルは、B7 です。この状態でエンターキーを押すと、次のカレントセルは A2 になります。

セルの範囲選択は、マウスだけでなくキーボードでもできます。シフトキーを押しながら矢印キーを押すと、範囲が選択されます。細かく範囲を選択したい時には、キーボードのほうが確実です。

なお、いったんセルに入力したデータを修正するには、修正したいセルまで移動してから、「F2」キーを押します。

5.2.3 式

スプレッドシートでこれがなくては始まらないというのが、式と関数です。スプレッドシートは単なる表としての利用も可能ですが、式を利用するとコンピューターのコンピューターたるゆえんと便利さを理解することができます。

式を入力するには、カレントセルでまず最初に「=」から入力を開始します。そして、数字や記号などを使って入力していきます。なお、ここで利用するのはいわゆる「半角」の数字や文字であることに注意してください。

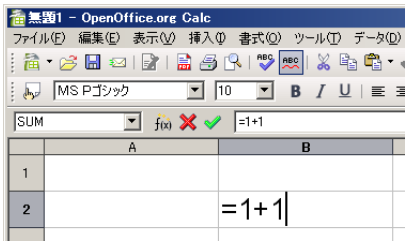

図 5.3: 式の入力

図 5.3 では、「=11+」と入力しています。入力が終わったら、エンターキーを押してみてください。結果として「2」が表示されるはずです。

足し算と引き算はそれぞれ「+」「-」で表されます。乗算と除算はそれぞれ「*」と「/」です。

式はこのような単純なものだけでなく、複雑なものも利用することが可能です。式は、算数で勉強したように掛け算と割り算が優先され、括弧が利用された場合、その中が先に計算されます。例えば、次のような式を考えましょう。

```
=20*(6+4)/2-9*(4+8)/2-5
```

この式では、まず括弧内が計算され、式は次のように変換されます。

```
=20*10/2-9*12/2-5
```

括弧が外れたら、左から順に掛け算および割り算優先で計算していきます。

```
=200/2-108/2-5
=100-54-5
=41
```

5.2.4　関数

関数と呼ばれるものも便利に利用できます。関数とは、これが与えられれば値が定まるという、パラメータを与えることで成立する式です。パラメータのことを引数（ひきすう）ともいいます。例えば、2 の平方根を考えてみましょう。平方根は、その 2 という数を与えられてはじめて計算できるわけです。言い換えれば、その 2 という値が決まればこの関数の値が一意に定まります。ここでその「2」がパラメーターないし引数と呼ばれるわけです。2 の平方根を計算するためには、次のように入力します。

```
=sqrt(2)
```

ちなみに sqrt は SQuare RooT の略です。引数には任意の数値を入れられますから、色々試してみてください。関数は、引数を与えられると「戻り値」を返します。

関数の引数は、sqrt のように 1 つしか取らないものもあれば、複数取るものもあります。引数の数が決まっているものもあれば、決まっていないものもあります。例えば、合計という関数は sum ですが、この引数は 1 以上 30 以下となっています。なお、30 というのは Calc の制限です。

```
=sum(1,2,3,4,5,6,7,8,9,10)
```

sum 関数では、カンマ（,）で区切って複数の引数を指定します。これらの引数が合計を計算する対象になります。

関数には様々な種類があります。大まかな分類と内容を、表 5.1 に示しますが、数百の関数があり、また独自に自分で関数を作成することすら可能です。

関数の中には複雑な計算を行うものも含まれるのですが、ここで注意したいのは、関数を利用する際にはそれを「ブラックボックス」であると考えてはいけないということです。つまり、どのように計算されているのかということについて無自覚ではいけないのです。どのような表計算ソフトであれ、入力された、あるいは計算された結果としてのデーターに責任を持つのは、あなた自身なのです。

いずれにしても、これら全部覚えるのはほとんど不可能であり、無駄といってもいいでしょう。基本的なものは別として、自分に必要なものだけをその都度覚えていくのが正しいアプローチであり、覚えておくべきなのは関数の使い方をどうやって調べるか、です。

日付と時刻	日付と時刻に関わる計算
財務	金利や減価償却などの計算
情報	セルの状態に関する情報
論理	論理計算
数学	絶対値や対数、三角関数など
行列	行列演算や会期分析など
統計	平均や分散、確率密度分布、検定など
表計算ドキュメント	ドキュメント中の文字列検索など
文字列	文字列の抽出、変換など
データベース	特定の範囲にあるデータに関する情報の計算

表 5.1: Calc における関数の分類

すべての関数については、Calc の「ヘルプ (H)」メニューから「OpenOffice.org ヘルプ」を選択することで使い方と簡単な意味を調べることができます。「検索キー (S)」に調べたい語を入力すれば、関数をはじめとする Calc の使い方を検索することができます。しかし、例えば「分散」を VAR 関数などで計算することができるということについては示してくれますが、そもそも分散が何を意味するのかは示してくれないのです。

5.2.5 式・関数におけるセル番地の利用

これまでは式や関数の中で直接数字を指定していました。ここでは、数字の代わりににセルを指定してみましょう。

今まで、例えば 2 の平方根は直接数字を指定していましたが、ここでは次の図のように、A1 のセルに 2 を入力し、B1 のセルに A1 のセル内容の平方根を計算するような指定をしてみます。

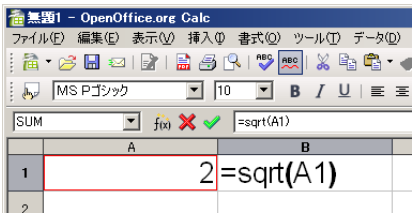

図 5.4: セル番地を引数とする関数の入力

直接数字を指定したのと同じ結果が表示されるはずです。ここで、画面上の「A1」はセル番地「A1」を示しています。

次に、この A1 に入力されている「2」を、別な数字に置き換えてみてください。3 でも 4 でも、100 でも 100000 でもかまいません。A1 に入力する数値が変化すると、A2 の平方根も変化します。このように、スプレッドシートではセル番地を参照させることで、様々な条件に対応した計算を行うことができるのです。

最初の PC 向けスプレッドシートプログラムはハーバード大学ビジネススクールの学生が作成した VisiCalc というプログラムでしたが、その動機になったのは、このような機能を求めてのことだったのです。以前は、手計算をできるだけ容易にするための集計用紙とかグラフ用紙というものが文房具屋で売られていましたが、経理や財務などのシミュレーションには膨大な計算が必要になります。

例えば、売上高が 1% 変化したら最終的な利益はいくらになるかとか、宣伝広告費を徐々に変化させた場合、売上高にどれほどのインパクトがあって、それが利益にどのような影響があるかといったことを、繰り返し計算する必要があります。つまり、計算式は同じでも式に別のパラメータを与えた結果というのが欲しいわけで、このような用途に表計算ソフトはぴったりだったというわけです。

5.2.6　相対参照と絶対参照

次は、式にセル番地を利用することについて、もう少し別の便利な面を見てみましょう。図 5.5 のように、A1 から A5 までにあらかじめ、1 から 5 の数字を入力しておきます。これらの数字それぞれに対して平方根を取ると、どのようになるかということを一覧したいとしましょう。

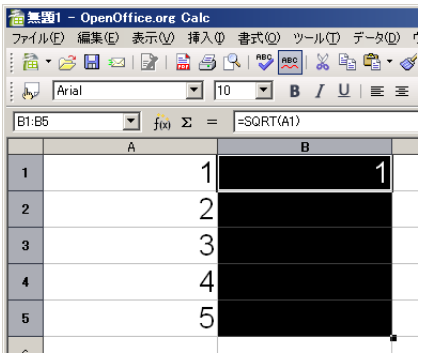

図 5.5: コピーの準備

5 回ほど式を入力して、「=sqrt(1)」、「=sqrt(2)」... とやるのもいいのですが、もう少し賢い方法があります。上の図のように A1 から A5 に数字だけ入力したら、B1 に「=sqrt(a1)」と入力して、エンターキーを押しておきます。次に、セル範囲 B1〜B5 を選択します。マウスでドラッグしてもかまいませんし、シフトキーと矢印キーを利用して範囲を選択するのでも構いません。

次に、「編集」メニューから「連続データ」→「下へ」を選択します。B2 から B5 までに、平方根が計算されているはずです。

カレントセルを B2 や B3 に移動させ、その際の数式バーを検証してください。自分自身では実際に入力していませんが、B2 には「=SQRT(A2)」という式が入力されています (図 5.6)。B3、B4、B5 もそれぞれ参照すると、やはり SQRT の引数がそれぞれ「ずれて」入力されています。

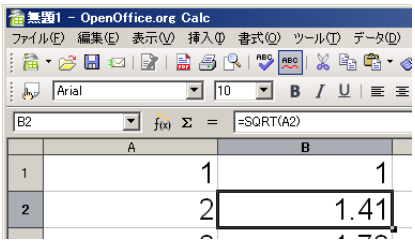

図 5.6: コピーされた式

これは確かに便利で、我々が望んだこともあるのですが、このようになる理屈は次のようなものです。つまり、この「=SQRT(A1)」という式は、相対参照という方式でセルを参照しているために、コピーした際に相対的に式がずれていくのです。

この相対参照とは、セル同士の相対的な位置関係の参照なのです。つまり、B1 のセルは A1 のセルを参照しています。しかし、B1 は A1 のことを直接参照しているわけではなく、単に「1 つ左隣」としてしか認識していないのです。ですから、式をコピーして下方向に持っていった場合、コピーされた下の式も同じように「1 つ左隣」を参照します。つまり、かなり乱暴に図示すると、Calc 内部では、次の図のような内容が入力され、またコピーされていたということになります。

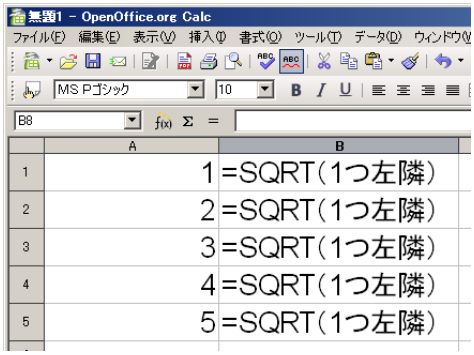

図 5.7: 相対参照コピーの考え方

B1 に入力された「=sqrt(A1)」という式の段階で、Calc が保持している本当の内容は、「=sqrt(**1 つ左隣**)」なのです。ですから、いくらこれを下方にコピーしても、1 つ左隣が参照され続けるということになります。

このような仕組みで、コピーするだけで B2 は A2 を、B3 は A3 を参照するようになるわけです。この相対参照は右とか左だけではなく、上とか下、またそれらを組み合わせて参照することも可能です。この相対参照によって、我々は同じ式を何度も入力する必要がないというわけです。

さて、そうはいっても相対参照だけでは困ることもままあります。そこで、絶対参照というものも用意されています。a1 のように、列と行とに「$」マークをつけることで絶対参照になります。また、$a1 とか a$1 のように、列だけ絶対参照で行は相対参照、またその逆というのも利用できます。

5.2.7 参照における範囲の指定

sum という関数があることは既に説明しましたが、sum という関数は 30 までの引数しか取ることができませんでした。では 30 のセルまでしか合計を取ることができないのかというと、それは違います。より効率的な指定の仕方があります。

例えば、

```
=sum(A1,A2,A3,A4,A5)
```

として A1、A2、A3、A4、A5 と個別に指定する代わりに、次のように A1:A5 と、「:」で範囲として指定することが可能です。

```
=sum(A1:A5)
```

この範囲の指定の仕方は 1 列とか 1 行に限定されるものではなく、自由に決めることができます。行や列をまたがってもかまいません。左上のセルと右下のセルをそれぞれ頂点とする長方形が範囲となります。

5.2.8　論文で利用するデータの計算に表計算ソフトを使う前に

ここで紹介した程度の処理ならば Microsoft Excel や Calc でも可能ですが、アカデミックなレポートや論文に利用するには、心もとないのです[3]。というのも、Excel や Calc は、小数点以下の数値を正確に計算するように作られていないからです。図 5.8 を参照してください。ここでは、セルの表示形式を数値として、小数点以下 20 桁まで表示するようにして計算しています。A1 および A2 のセルにはそれぞれ 3.2 および 3.3 という数値を入力してあり、A3 のセルには「=A2-A1」という式を入力してあります。

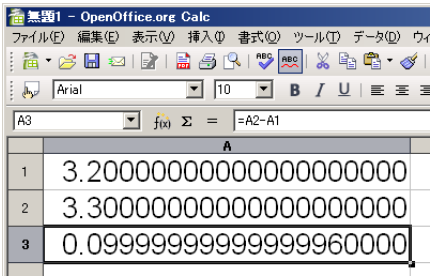

図 5.8: Calc による小数点以下の計算

ここでは Calc を用いましたが、Excel でもまったく同じ結果が得られます。詳細な説明は省きますが、表計算ソフトでは小数点以下の計算には必ず誤差があり、表示されている計算結果は誤差を丸めた結果としてのものです。そのため、計算方法や計算結果、その表示方法によって誤差が見えたり見えなかったりします。ごくわずかな誤差かもしれませんが、多くの場合で本来ユーザの意図していない不正確な計算が行われているのは事実であり、わずかな計算結果の差が問題になるような計算に利用するのは危険ですらあります。

また、例えば Excel には回帰分析や分散分析などの高度な統計計算機能も用意されていますが、この計算結果について過去に多くの問題点が報告されています。利用する前に道具としての妥当性を検証しておきましょう。International Association for Statistical Computing という学会の論文誌、「Computational Statistics and Data Analysis 52 (2008)」で、「Special section on Microsoft Excel 2007」と題して、特集で Microsoft Excel 2007 を計算の道具として利用することが妥当かどうか、５本の論文で集中的に検討されています。それぞれの論文の結論はここには書きませんが、電子ジャーナルとして早稲田大学のコンピューター教室からも閲覧できますので、興味のある諸君は検索して読んでみて下さい。

いずれにしても、プログラムは人間が記述している以上、間違いが含まれている可能性は避けられません。それはどのようなプログラムでも同じです。コンピュータを利用した計算に間違いが混入する可能性が常にあるということは、念頭に置いておく必要があるでしょう。

一方で、多くの専門家が開発に参加し、多数のユーザによって検証が行われている、計算結果を一定度、信頼することができるソフトウェアもあります。表計算ソフトウェアで誤差のない計算をするためのノウハウも存在しますが、そのようなことを気にするよりは、あらかじめ計算精度の保証されているソフトウェアで計算するべきです。

具体的には、統計計算・数値計算・グラフ化に強い「R」、数式処理ソフトウェアの「Maxima」、数値計算・モデリング・シミュレーションに強いソフトウェアの「Octave」など、誰でも無償でダウンロードして利用することのできる数値計算用のソフトウェアがあります。なお、「R」は早稲田大学のほぼ全てのコンピューター教室で利用することができます。

[3] ビジネスも含めた他の用途についても同様です。

早稲田大学では全学の学生が利用できるような形で SAS や SPSS といった商用ソフトウェアも用意していますが、卒業してなお SAS や SPSS を使い続けることのできる恵まれた環境が得られるとは限りません。論文で何かしらの統計処理や数値計算を取り入れたいと考えている諸君は、これらのフリーソフトウェアを積極的に活用すると良いでしょう。

5.3 演習問題

この章では特に演習問題を設けません。早稲田大学メディアネットワークセンターでは、全学副専攻科目として「ソフトウェア学」の他に「データ解析」というコースを持っています。コンピューターを利用したデーターの分析に興味を持った学生諸君は、これらの科目について調べてみて下さい。

コンピューターでデーターを分析することの大切さは論を待ちませんが、参考に次の動画を参考にしてください。

http://www.ted.com/talks/lang/eng/conrad_wolfram_teaching_kids_real_math_with_computers.html

少々 URL が長いですが、Web 検索で「Conrad Wolfram TED」というキーワードで検索しても見つけることができるはずです。英語のスピーチですが、日本語の字幕を付けることも可能です。