

Excel によるデータ分析の基礎

－エクセルの初歩から相関関係まで－

土橋 喜

愛知大学現代中国学部

Primary Data Analysis with Excel

－ From Beginning Excel to Correlation －

Konomu Dobashi

Aichi University

目 次

1. はじめに
2. 表計算ソフト Excel 入門
3. Excel を使った計算方法
4. 移動平均と地球の温暖化
5. 高齢化・人口問題・人口ピラミッド
6. データの分散・偏差値・条件判断
7. 度数分布とヒストグラム
8. 2次元データの分析・相関関係
9. 回帰直線・クロス表
10. 乱数とさいころのシミュレーション
11. 社会データの情報源
12. 引用文献

1. はじめに

本稿は文科系や社会科学系の学生を対象にして、表計算ソフトのExcelを活用したデータ分析の入門教材としてまとめたものである。

我々の身のまわりに起こるさまざまな事柄には、現象を数理的かつ視覚的に捉えると理解しやすい場合が多い。例えば人口変動や日々の気温の変化あるいは株価の変動など、さまざまな現象をあげることができるが、これらは数値データだけでなく、グラフ化して直感的に理解しやすく表示している場合も多い。

これらの現象を数理的かつ視覚的に捉えるためには、統計的な分析方法に加えて分析結果を分かりやすく視覚化するグラフ作成の方法も学ぶ必要がある。本稿では統計学の基礎理論を取り上げているが、分析ツールとして表計算ソフトのExcelを使い、数学的に深入りすることなく、データ分析とグラフ化の方法を学べるように試みている。

ところで最近の企業や自治体などの組織には、情報化の進展に伴ってさまざまな形式のデータが大量に蓄積されている。総務省統計局統計センターのホームページには、国勢調査のデータが公開されているのをはじめ、国内外の各種の統計データへのリンク集が作成されている。このように社会に蓄積された各種のデータが、インターネット上にも数多く公開され、手軽に利用できるようになった。

最近では利用できるデータが豊富になったことから、データマイニングと呼ばれる研究が盛んになり、研究成果を発表する国際会議が毎年のように世界各地で開催されるようになった。データマイニングはさまざまな分析手法を活用して大量のデータから一定の傾向を示すパターンを見出し、商品販売などの意思決定に役立ちそうな有益な情報を発見しようとする研究である。新たな分析手法の提案やそれを具体化したソフトウェアの開発が行われており、従来の統計学的なデータ分析をさらに発展させようとするものでもある。データマイニングという言葉や考え方からは新鮮な印象を受けるが、この基礎になっているものの多くはこれまでに研究されてきた統計学的なデータ分析の手法である。

これからの社会人には、情報を収集し、それらを適切に分析してまとめ、発表する能力あるいは情報発信する能力が必要とされるが、本稿では主に統計データの収集とそれらの基礎的な分析方法を身につけることに重点をおいており、言い換えればデータマイニングの基礎を学ぶことでもある。そのため本稿では収集したデータをExcelで分析し、視覚化するための知識とパソコンの操作方法を習得することを目標にし、主にExcelに組み込まれている関数の使い方を学習しながら、入門的なデータ分析の手法を概説した。

従来の統計分析は、自分が考えている仮説が正しいかどうかを確認する仮説検証型の手法であるといわれる。これに対してデータマイニングは、マイニングされたデータのパターンから、新たな問題解決のきっかけやビジネスチャンスを見出す仮説発見型である。

データの収集と分析の両方において、問題発見・問題解決の観点からそれぞれの目的を明確にししながら、自分でデータを収集し分析する意義を考え、問題解決の仮説を立てることが重要であることはいままでのない。

2. 表計算ソフト Excel 入門

表計算ソフトのExcel (Microsoft Excel)を使った基本的なデータ分析を学ぶ。まずExcelの代表的な活用方法を説明する。

(1) 電卓の代用

表に記入された数値の縦や横の計算は、項目が多いほど効果的になる。紙の上での計算や電卓と違い、項目の追加や削除、訂正も可能である。用途としては一般的な計算を始め、現金出納帳、成績処理、見積書、売上集計表などがある。

(2) 自動再計算機能

セルと呼ぶ表のます目の数値を変えると、あらかじめ設定しておいた計算方法で自動的に再計算する。いくつもの数値を変えながら行うシミュレーション分野で活用できる。用途としては、予算管理、資産管理、市場調査分析、在庫管理などがある。

(3) 関数の活用

複雑な計算式を作成しなくても、用意されている各種の関数を使い、いろいろな計算を比較的簡単に行うことができる。用意されている関数は、財務関数、数学／三角関数、検索／行列関数、文字列操作関数、情報関数、日付／時刻関数、統計関数、データベース関数、論理関数などがある。

(4) グラフ作成機能

入力したデータのグラフを作成して分かりやすく表示する。作成できるグラフは、面・横棒・縦棒・折れ線・円・ドーナツ・レーダー・散布図などのグラフおよびいくつかの3次元グラフの作成ができる。

(5) データベース機能

大量のデータを蓄積管理し、それらの中から特定の条件に該当するデータを検索したり抽出したりすることができる。用途としては、住所録、名簿、蔵書管理、顧客情報管理などがある。

(6) マクロ機能

実際に操作した手順をそのまま記録し、それを再実行することによって操作を簡略化できる。プログラムのように手順を書いて、定型的な機能を自動的に処理できるようにする。

2. 1. Excel の起動と終了

(1) 起動

Excel を起動してみよう。Windows のスタートメニューからプログラムを選び、Microsoft Excel をクリックして起動する。図 2.1 のようなウインドウが開くと Excel が使えるようになる。なお各種のアイコンは、メニューバーから「表示」を選び、次に「ツールバー」を選択すると表示するアイコンを変更できる。普段は「標準」をチェックし、「グラフウィザード」や「図形描画」などのアイコンを表示しておくといよい。慣れてきたら必要に応じて、表示するアイコンを変更する。

(2) 終了

終了するときには Excel のメニューバーにある「ファイル」から「終了」を選ぶか、右上の×印の終了ボタンを押す。なお作業中のデータを保存しないで終了しようとする、警告メッセージが表示されるので保存してから終了する。

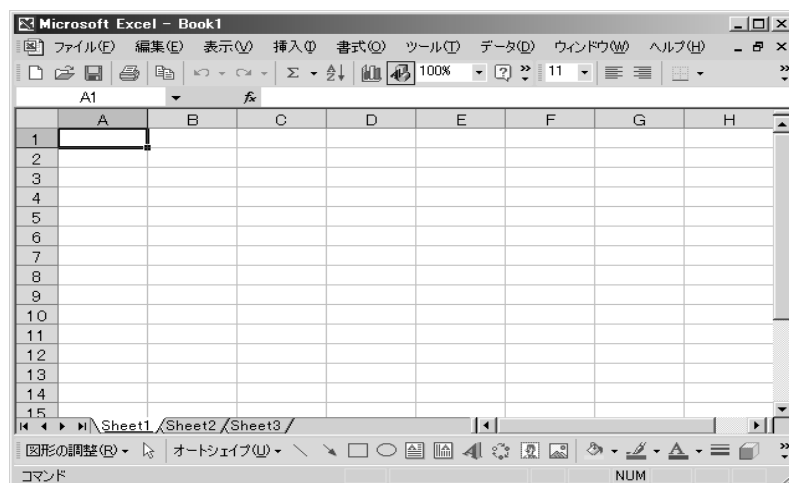


図 2.1 エクセルの起動画面

2. 2. ファイルの読み込みと保存

Excel のファイルの読み込みと保存の基本的操作方法は、Word など他の Windows のアプリケーションと同じである。

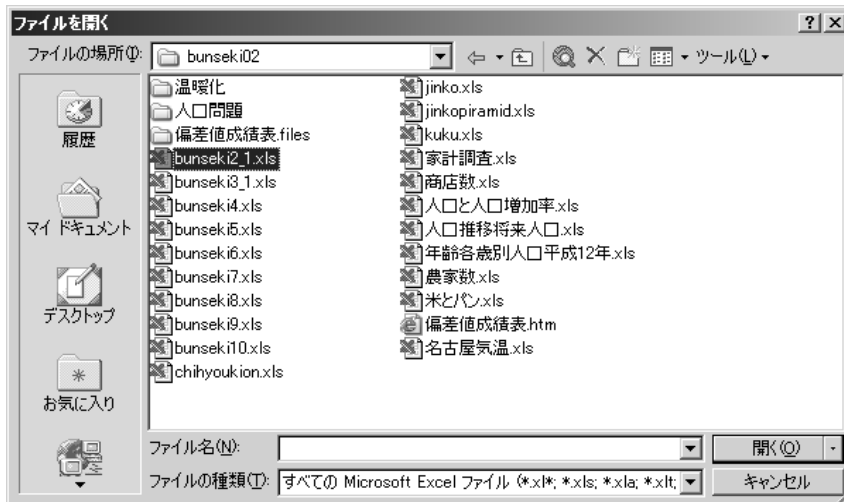


図 2.2 「ファイルを開く」の画面

(1) ファイルの読み込み

メニューバーの「ファイル」から「開く」を選択する。「ファイルを開く」というダイアログボックスが開くので、必要なファイル名をダブルクリックする。またはファイル名の入力欄にファイル名を入力し、「開く」ボタンをクリックすると指定したファイルが読み込まれる（図 2.2）。



図 2.3 「名前を付けて保存」の画面

(2) ファイルの保存

メニューバーの「ファイル」から「名前を付けて保存」を選ぶと、「ファイル名を付けて保存」というダイアログボックスが開くので、「ファイル名」という入力欄に適切なファイル名を付け、「保存」ボタンをクリックする（図 2.3）。ファイル名は Excel のタイトルバー（画面上部）に表示されるので確認できる。

2. 3. 画面構成と機能

ファイル・編集・表示・挿入・書式などのメニューは、Windows ではWord など他のソフトにも使われる共通の名前になっているが、ここでは名前が同じでも各部の機能はExcel 用になっている。メニューの中の項目を全部覚える必要はないが、試行錯誤でどこの部分がどんな働きをするかを少しずつ理解していくとよい。

(1) 数式バー

A, B, C, ... の列見出しの上であり、データの入力や訂正用に利用する。左の端には、現在のアクティブセルが表示される。数式バー上にあるデータはそのまま編集できる。数式バーに表示されるアイコンの機能は次のようになっている。

▼ : 名前ボックス, ☒ : 入力の取り消し, ▣ : 入力中のデータをセルにセット, fx : 数式の編集

(2) シート見出し

他のシートを選択する場合に使う。



(3) ワークシート

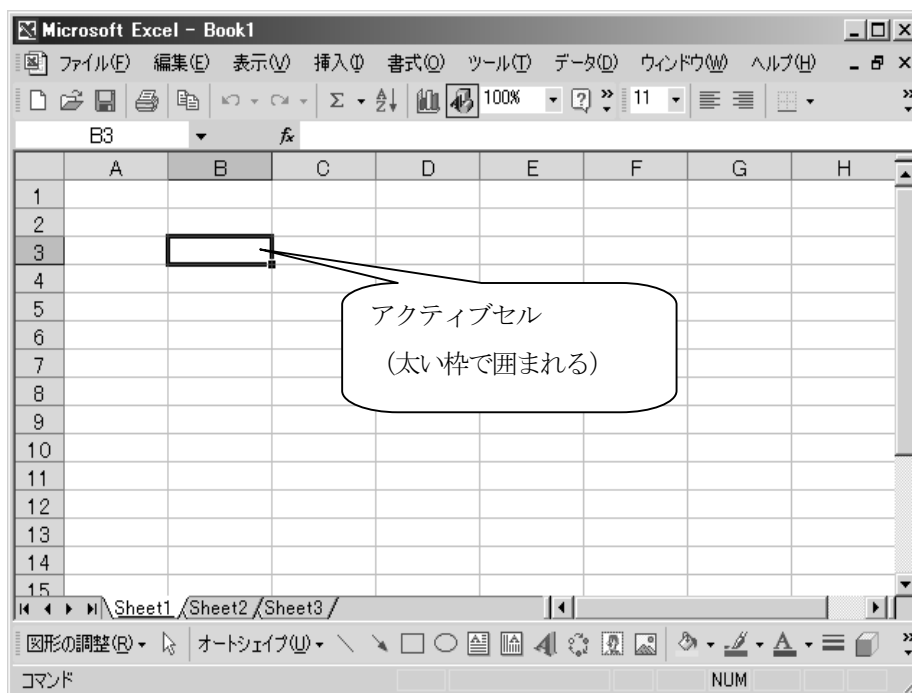


図 2.4 アクティブセル

データを入力する格子状の表のことである。格子状のます目のひとつひとつはセルという。256列×16,384行もある。セル総数は4,194,304になるが、メモリ容量によって一度に利用できるセル総数が限られる。ワークシートは16枚まで使える。それぞれのワークシートにはSheet1, Sheet2…のような名前が自動的に付くが、名前を変更してより適切なものにすることもできる。シート名の変更をするときは、シート見出しの上で右ボタンを押し、名前の変更を選び、適切な名前に変更する。

2.4. データの入力と処理

マウスマウスカーソルを移動して、必要なセルを左クリックすると、そのセルが太い枠で囲まれ、データを入力したり、いろいろな処理の設定をしたりすることができるが、これをアクティブセルと呼ぶ(図2.4)。

(1) ワークシート上の移動

◆ キーボードを使う場合

近くのセルに移動する場合は、矢印キーを操作して移動する。

遠くのセルに移動する場合：

- * PageUp : 1画面戻る
- * Pagedown : 1画面進む
- * Home : A列の先頭に移動
- * Ctrl+Home : A1に移動
- * Ctrl+矢印 : セルがデータで埋まっている場合は、そのエリアの先頭または最後へ移動。セルが空白の場合は、シートの先頭または最後に移動。

◆ マウスを使う場合：スクロールバーなどで移動し、直接マウスでクリックする。

(2) 範囲の指定

◆ 1つのセルの選択

マウスで左クリックして選択する。

◆ 複数のセルの選択

左上のセルから、右下のセルまでドラッグする。連続していない場合は、Ctrlキーを押しながらドラッグする。

◆ 行の選択

行番号をクリックする

◆ 列の選択



列番号をクリックする。

◆ ワークシート全体を選択

 全セル選択ボタン (A 列の左のボタン) をクリックする。

(3) 処理の実行

Excel で処理を実行するときの基本は、他の Windows 上のソフトと同様に、範囲を指定してから処理の命令を実行する。処理の実行方法には次のような方法がある。また処理を実行することを「コマンドを入力する」ということもある。

- ◆ マウスでメニューバーを開き、該当するメニューを左クリックする。
- ◆ Alt+アクセスキーで、該当するメニューを開き、さらにCtrl+アクセスキーで必要な処理を実行する。アクセスキーというのは、画面に表示されているメニューやボタンをクリックする代わりに使うキーのことで、メニューやボタンの名前の後ろに括弧に囲まれている英文字のことである。
- ◆ アンドゥ機能
直前の操作を取り消して、処理を元に戻す機能であり、操作を間違えたときに非常に便利である。この機能は「編集」から「元に戻す」を選択するか、またはCtrl+Zを押すと実行される。あるいはツールボックスのアイコンをクリックする。
- ◆ リPEAT機能
直前に実行したコマンドをもう一度実行する。この機能は「編集」から「繰り返し」を選択するか、またはCtrl+Yを押すと実行される。またはツールボックスのアイコンをクリックする。

(4) 日本語の入力

Word の日本語入力と基本的には同じである。MS-IME を利用して行う。

(5) データの入力と訂正

Excel では、すべてのデータをセルに入力する。それぞれのセルは独立しているので、どこから入力してもかまわない。ひとつのセルにデータの入力が終わったら、Enter キーまたは矢印キーを操作して次のセルに入力する。

入力ミスの訂正をするときは、間違えたセルにカーソルを動かしアクティブにする。その上に直接正しい数値や文字列を入力し、Enter キーを押す。

2. 5. ワークシート上の基本操作

Excel のワークシート上で行う基本的な操作を説明する。

(1) データのコピーと貼り付け

データをワークシートの別な場所にコピーする場合を説明する。例えばA1:A5のデータをB1:B5にコピー

する。

まずコピーしたい行あるいは列を選択する(A1:A5)。次に「編集」メニューから「コピー」を選択する。あるいは右ボタンをクリックして、「コピー」を選択する。

次にコピー先の先頭のセルを選択し、「編集」メニューから「貼り付け」を選択する(図 2.5)。あるいは右ボタンをクリックして、「貼り付け」を選択する。データが正確にコピーされているかどうかを確認する。

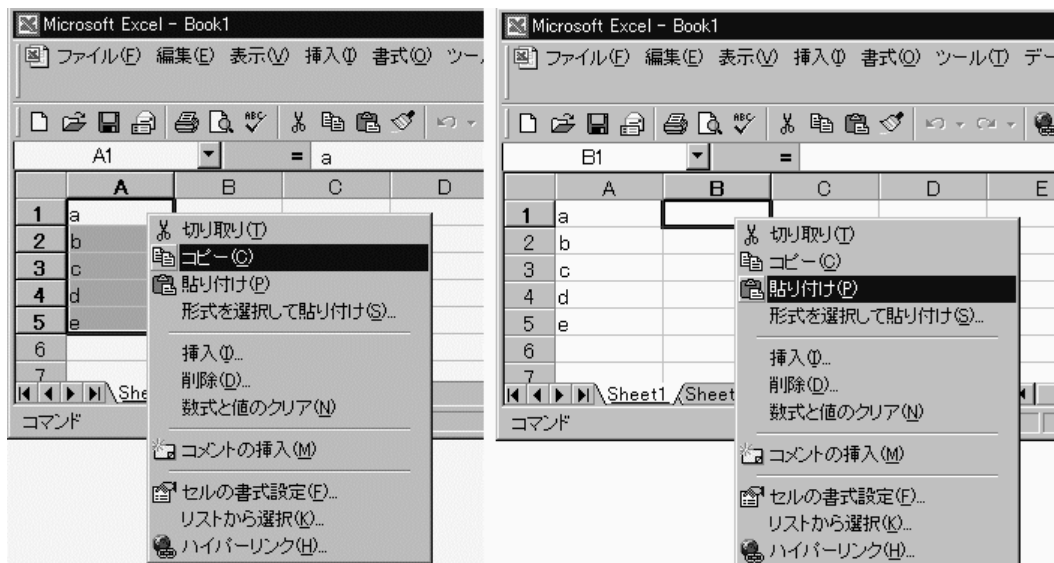


図 2.5 コピー(左図)と貼り付け(右図)

(2) 不連続データの選択

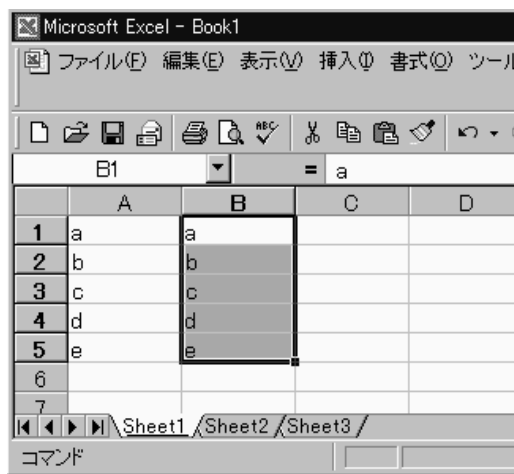


図 2.6 貼り付け完了

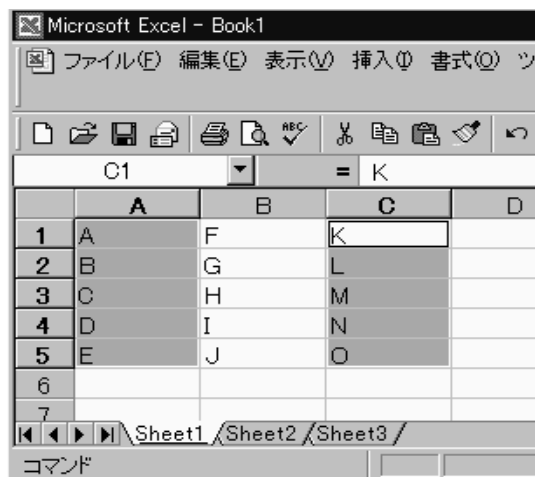


図 2.7 A 列と C 列の不連続データの選択

Excel のワークシートでは、連続していない行や列のデータを選択することができる。例えば A1:A5 および C1:C5 の連続していないデータを選択するときは、Ctrl キーを押しながら A1:A5 および C1:C5 を選択すればよい。選択した後は、コピーしたり貼り付けたりすることもできるが、貼り付けた場合選択していない列は詰められる。

不連続データの選択は、列ごとのデータを使ってグラフ化するときに頻繁に使われる。

(3) 連続データの入力方法(オートフィル機能)

Excel には初期値が与えられると、自動的に連続データを入力するオートフィル機能がある。例えば A1:D1 までのセルに、初期値として数値の「1」、英文字の「xyz」、漢字で「月曜日」、数字と漢字で「1 月」を入力する。

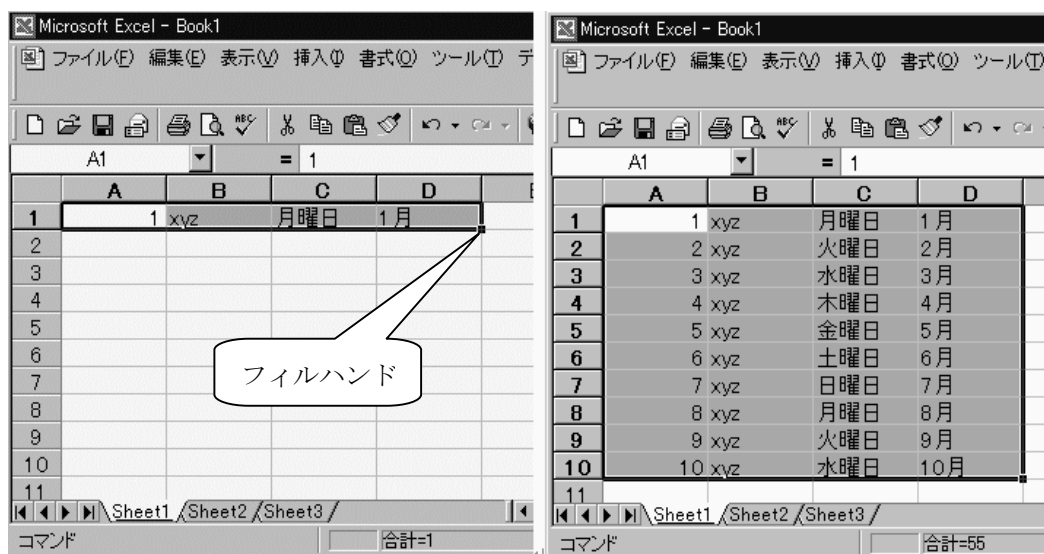


図 2.8 元データの入力と連続データの作成

次に A1:D1 を選択し、マウスを選択した枠の右下に合わせ、ポインタが黒い十字の形に変わることを確認する(この右下の角をフィルハンドルという(図 2.8))。ポインタが十字形に変わったところで、左のボタンを押したまま下にドラッグする。十字形に変わったポインタの右に文字や数値でヒントが表示されるので、適切な値のところまで左ボタンを離すとデータが自動的に入力される(図 2.8)。

この機能では数値など連続データとみなされるものは、初期値を入力すれば続きのデータが自動的に生成される。しかし「xyz」などのように連続データとみなされないものは、変化せず固定されたままの入力となる。しかもこの機能は曜日や月などの連続データにも適用される。

(4) ソート (並べ替え)

データを昇順や降順などに並べ替えを行うと、データの変化を見るときに便利である。例えば 1 から 10 までの数字をランダムに入力して、昇順に並べ替えを行うときは、先にソートするデータを選択する。次にツールバーにある「昇順で並べ替え」をクリックするとソートが実行される(図 2.9)。

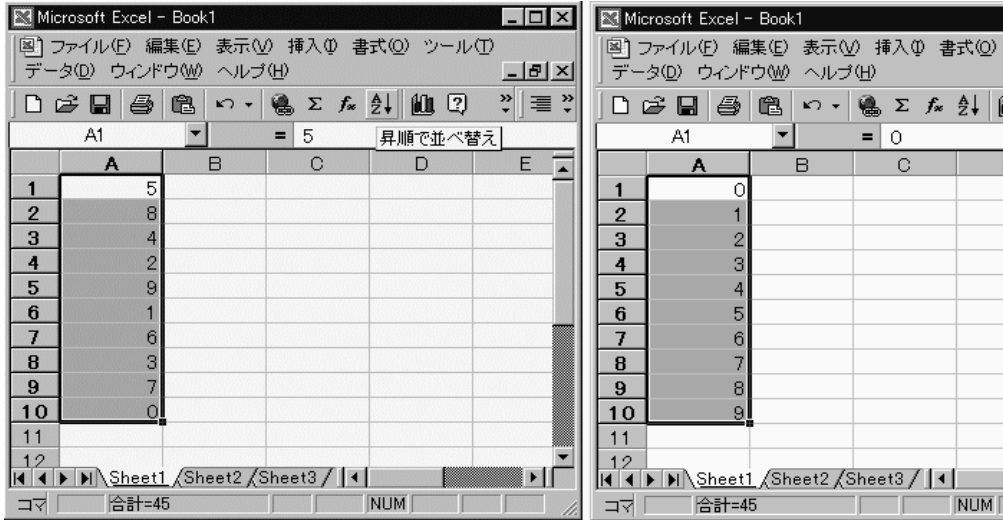


図 2.9 ソートデータの選択(左)と昇順で実行結果

2. 6. グラフの作成

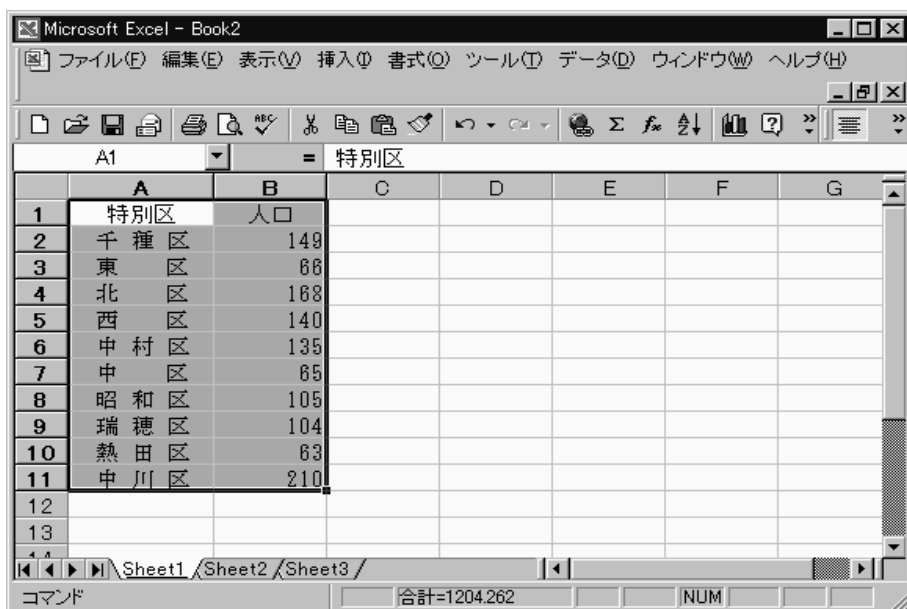
Excel には入力したデータを元にして、さまざまな形式のグラフを作成する機能がある。ここではその手順を説明する。簡単なグラフを作成するときは、グラフウィザードを使うと便利である。Excel を開発した Microsoft は、グラフ作成のような対話形式の自動設定機能にウィザードという名前を付けた。

(1) データの作成

最初にグラフを作成するために必要なデータを入力する(図 2. 10)。

(2) グラフを描く範囲の選択

グラフを描きたい範囲を選択する。項目と数値データを対応させて選択する(図 2. 10)。



The screenshot shows the Microsoft Excel interface with a spreadsheet containing the following data:

	A	B	C	D	E	F	G
1	特別区	人口					
2	千種区	149					
3	東区	66					
4	北区	168					
5	西区	140					
6	中村区	135					
7	中区	65					
8	昭和区	105					
9	瑞穂区	104					
10	熱田区	63					
11	中川区	210					
12							
13							

図 2.10 データの入力と選択

(3) グラフウィザードの起動

ツールバーの「グラフウィザード」ボタンをクリックする(図 2. 11)。

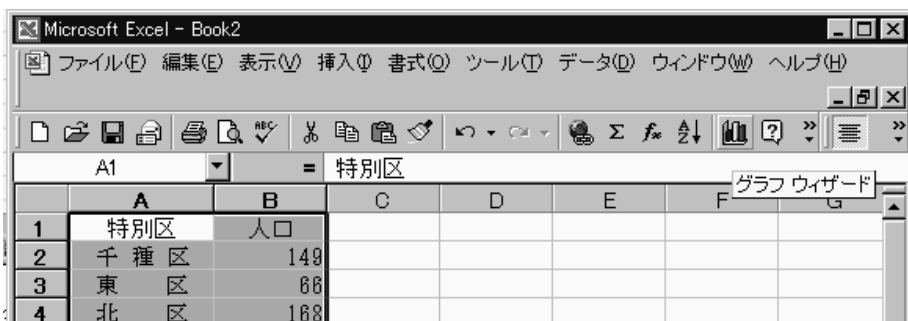


図 2.11 グラフウィザードのアイコン

(4) グラフの種類を選択 (グラフウィザード 1/4)

適切なグラフの種類を選択し、「次へ」をクリックする(図 2.12).

(5) グラフの元データの確認

「データの範囲」や「系列」などを確認し、「次へ」をクリックする(図 2.13).



図 2.12 グラフウィザード(1/4)

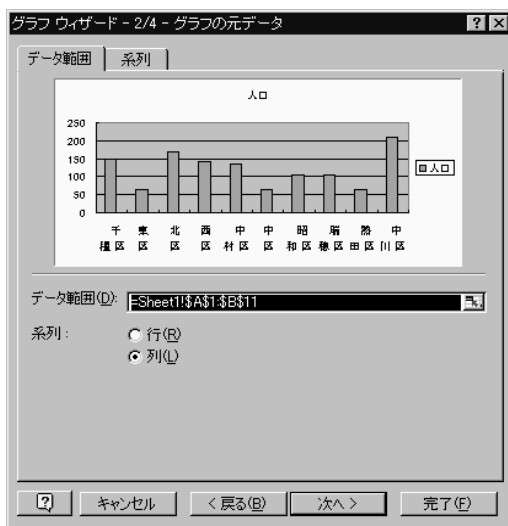


図 2.13 グラフウィザード(2/4)

(6) グラフオプション

グラフのタイトルなどを確認し、「次へ」をクリックする(図 2.14).

(7) グラフの作成場所

グラフの作成場所を確認する。「新しいシート」を選ぶと、グラフ専用シートが作成され、そこにグラフが描かれる。「オブジェクト」を選ぶと、シートの上に描かれる(図 15).

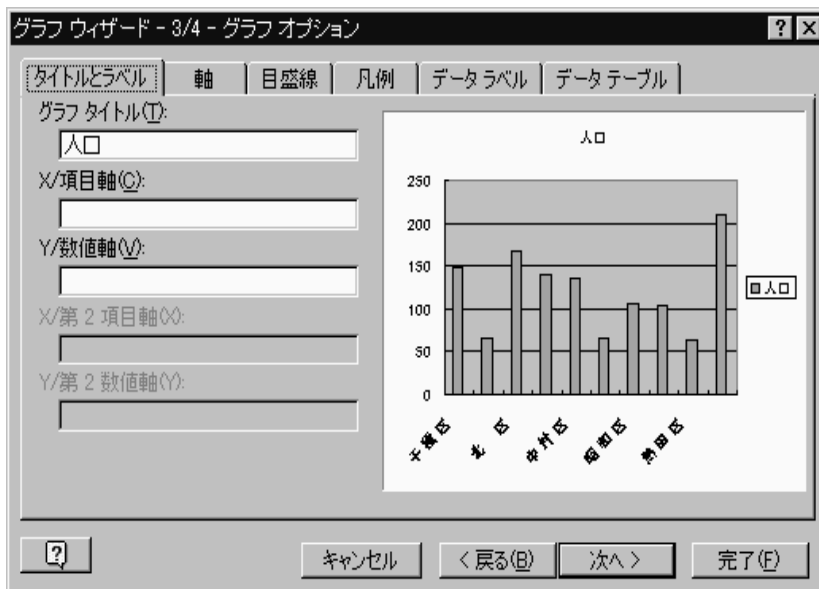


図 2.14 グラフウィザード(3/4)



図 2.15 グラフウィザード(4/4)

(8) グラフの作成

以上の指定を行い、「完了」をクリックするとグラフが作成される(図 2.16)。作成されたグラフの移動は、グラフの上で左ボタンを押してドラッグする。またグラフを左ボタンで選択しておき、グラフエリアの4つの角にマウスを当てると、ポインタが矢印に変わり、グラフの大きさを変更できる。

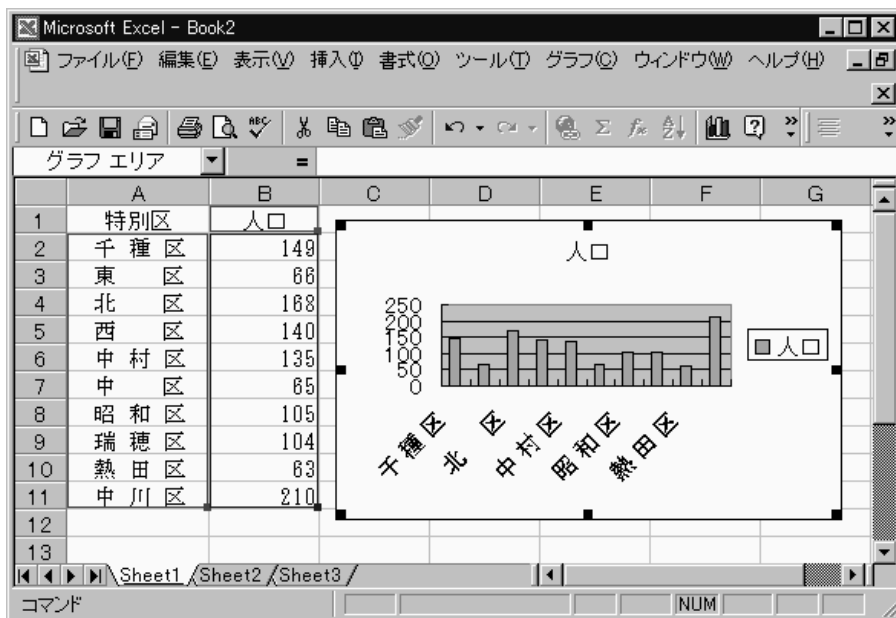


図 2.16 作成されたグラフ

2. 7. テキストの入力

Excel のワークシートにテキストを入力することができる。テキストを入力できると、グラフや表などの注意書きや解説などの文章を作成することができる。

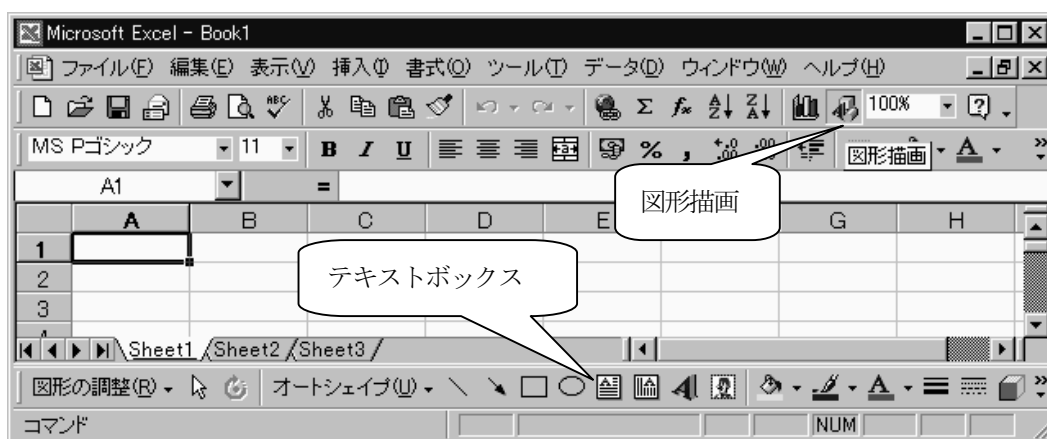


図 2.17 図形描画ボタンとツールの表示

- (1) ツールバーにある「図形描画」ボタンをクリックすると、Excel ウィンドウの下部にツールバーが表示される。
- (2) そこから「テキストボックス」を選択する(図 2.17)。

(3) カーソルで文字を入力する範囲を調節してから入力する(図2.18)。

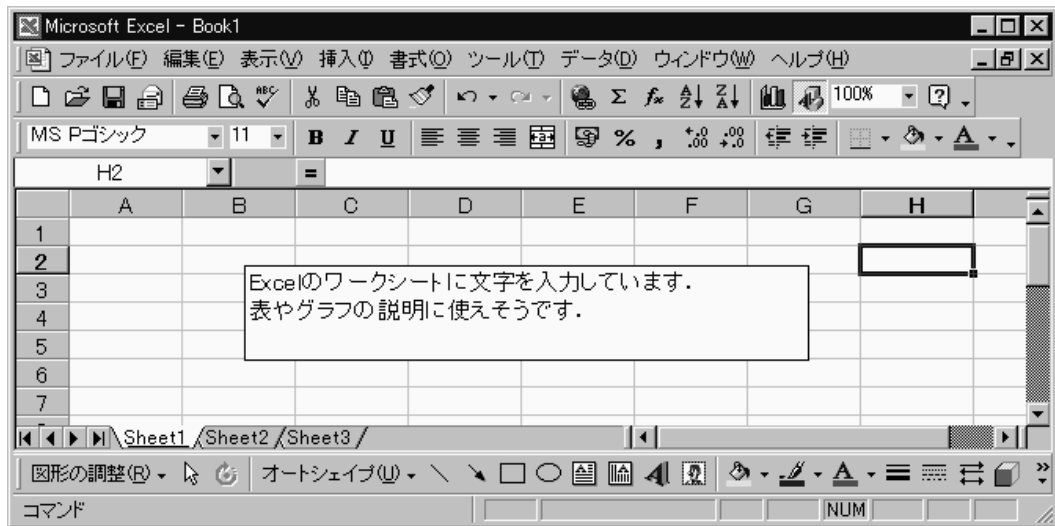


図2.18 ワークシートへのテキストの入力例

2. 8. データの移動

同じワークシート上ではデータを移動できる。例えば、B1:B6 に入力されているデータのうち、B3:B6 までを D4:D7 まで移動する例を示す。

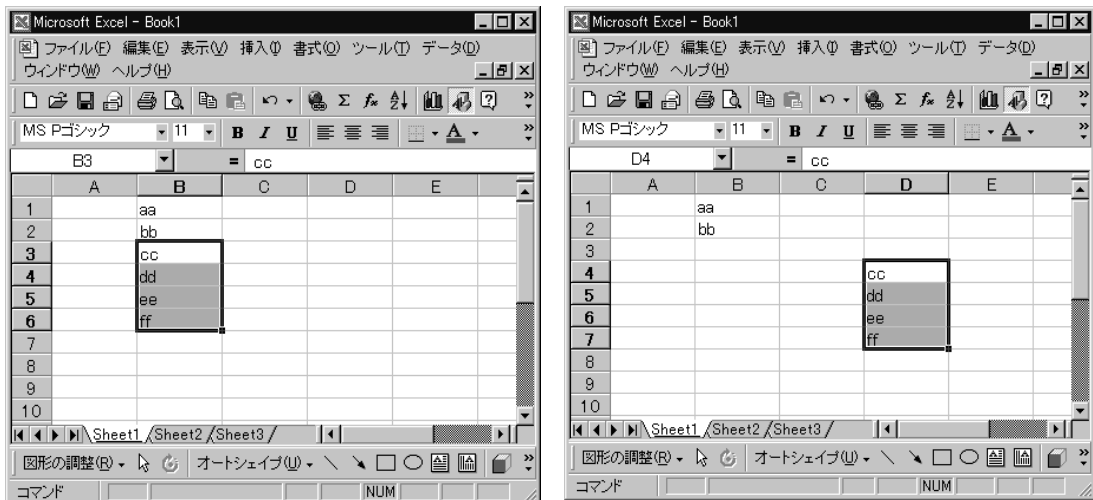


図2.19 データの選択(左図)と移動後

- (1) 先に移動したいデータをカーソルで選択する(図2.19の左図)。
- (2) 次にデータが入力されている外枠にカーソルを合わせ、形が白抜き矢印に変わったところで左ボタンを押す。
- (3) そのまま移動先の先頭のセルまでドラッグする(図2.19の右図)。

(4) 移動先でマウスを開放すると完了する。

2. 9. データの書式

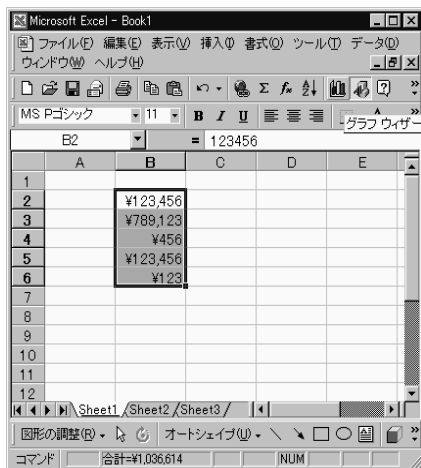
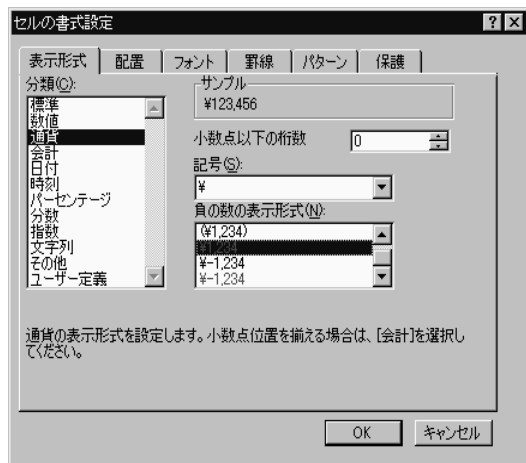
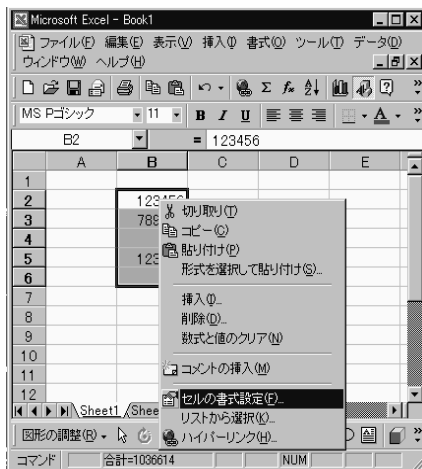


図 2.20 セルの書式設定

- (1) データの選択とセルの「書式設定メニュー」(左上図)。
- (2) セルの書式設定メニュー(右上図)。

Excel にはセルに入力したデータに単位を付けたりして書式を整える機能がある。ここでは数値データに通貨の円マークとカンマを付ける例を示すが、セル内のデータの配置位置やフォント、罫線、セルの網掛けなどの指定もできる (図 2.20)。

2. 10. 枠組みの表示



図 2.21 枠組みを作成するデータと「セルの書式設定メニュー」

表を見やすくするために、枠組みを作成することが多い。次の手順で枠組みの作成を行う。

- (1) 先に枠組みを作成したいセルを選択し、マウスの右ボタンを押すとメニューが表示される(図 2.21 右)。
- (2) 次に「セルの書式設定」から、「罫線」を選択する(図 2.22 左)。
- (3) 外枠の線や内側の線など必要な部分を選択する(図 2.22 左)。
- (4) 「OK ボタン」を押すと枠組みが作成される(図 2.22 右)。



図 2.22 罫線の設定メニューと結果

2.11. 印刷

先に印刷したいExcelの画面を表示しておく。

- (1) 次に「ファイル」メニューから「印刷プレビュー」ボタンをクリックする。
- (2) 印刷結果が望むものがどうか確認をする。
- (3) 確認したら後は印刷ボタンを押す(図 2.23)。

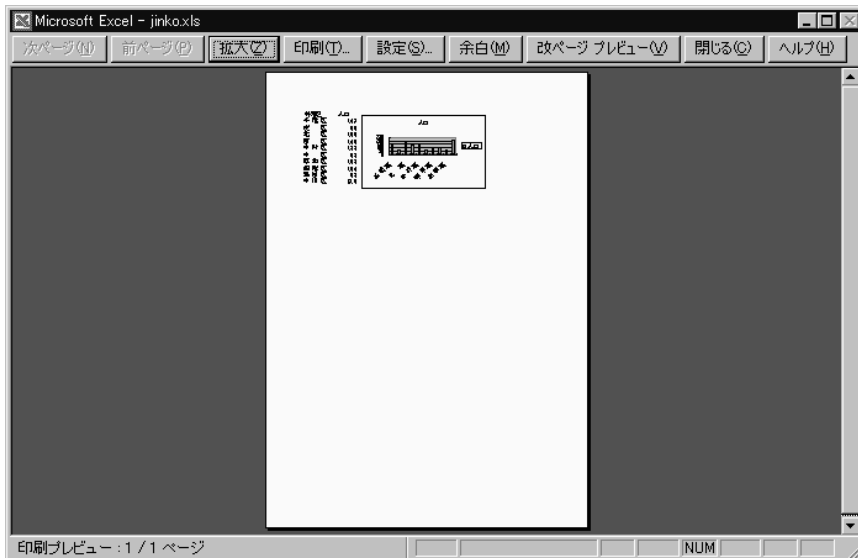
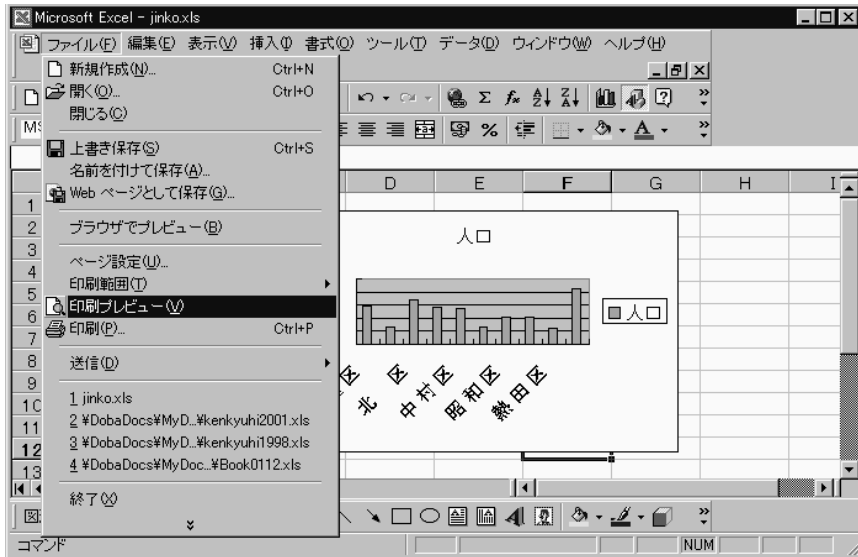


図 2.23 データの表示(上図)と印刷プレビュー(下図)

2. 12. データのダウンロード(取り込み)

総務省統計局統計センターなどでは統計データをダウンロードできるように公開している (<http://www.stat.go.jp/index.htm>)。このようなときは、Internet Explorer などを使い、リンクの上で右ボタンを押すと「対象をファイルに保存」のメニューが出るので、これをクリックするとデータのダウンロード(取り込み)ができる(図2.24)。

データの保存は「ファイル名を付けて保存」というウィンドウが開くので、保存する場所とファイル名を確認して行う(図2.25)。

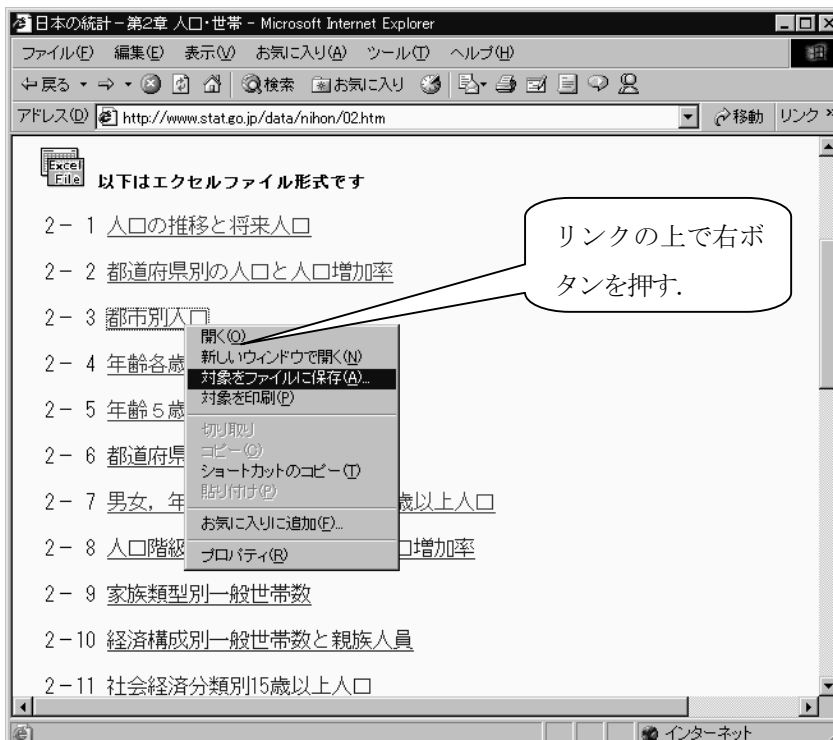


図 2.24 リンクの上で右ボタンを押した画面例



図 2.25 データの保存画面

》》》 演習 2 《《《

以下の演習を行い、結果をファイルに保存せよ。ファイル名はbunseki1.xlsとして保存する。

1. 愛知県内の主な市の人口統計グラフの作成

次のデータは愛知県内の主な市の人口統計である。このデータをもとにしてグラフを作成してみよ。なおデータは総務省統計局統計センターに公開されている「日本の統計、第2章人口・世帯、都市別人口」から一部を引用して作成したものである。「日本の統計」のURLは次のとおりである。

<http://www.stat.go.jp/data/nihon/index.htm>

(単位 1,000 人)

市	人口
豊橋	365
岡崎	337
瀬戸	132
半田	111
春日井	288
豊川	117

豊田	351
新城	36
知立	63
尾張旭	75
豊明	66
日進	70

2. 日本の男女別人口統計グラフの作成

次のデータは日本の人口統計の一部である。このデータをもとにしてグラフを作成してみよ。なおデータは総務省統計局統計センターに公開されている「日本の統計、第2章人口・世帯、人口の推移と将来人口」から一部を引用して作成した。「日本の統計」のURLは上に示したとおりである。

(単位は 1,000 人)

年次	男	女
昭和 25 年	40,812	42,388
30	43,861	45,415
35	45,878	47,541
40	48,244	50,031
45	50,918	52,802
50	55,091	56,849
55	57,594	59,467
60	59,497	61,552
平成 2 年	60,697	62,914
7	61,574	63,996
12	62,111	64,815

3. Excel を使った計算方法

ここではExcel のワークシート上で行う合計(オート SUM), 四則演算, 累積和などの簡単な計算方法や, データを参照する場合に使われる絶対参照と相対参照の仕組みについて学ぶ。

また統計学の基礎概念としての平均, 中央値, 最頻値などを学習する。

3. 1. 合計の計算(オート SUM)

オート SUM 機能を使うと合計を簡単に求めることができる(図 3. 1)。

- (1) 先に合計したいセルを選択しておく。図 3. 1 の例では B2:B8 までが合計を求める範囲である。
- (2) ツールバーのオート SUM ボタンをクリックする。
- (3) 計算結果が選択したデータの下部 (B9 セル) に表示される (行を選択した場合は右)。

	A	B	C	D	E	F	G
1	月	売上					
2	4月	150000					
3	5月	320000					
4	6月	650000					
5	7月	420000					
6	8月	510000					
7	9月	250000					
8	10月	750000					
9	合計	3050000					
10							

図 3.1 オート SUM による合計の計算

3. 2. 四則演算と数式

Excel では数式バーに「=」の後に続けて書いた計算式を数式と呼ぶ。数式の中には, 数値や文字列のような定数または関数などを用いた演算を記述することができる。

数学と同じように, 演算記号で結合したものを項と呼び, 項がいくつもあるときは演算記号を使って結合する。また数式でのカッコの使い方と計算の順番は数学と同じになっている。しかし掛け算や割り算などの演算記号には, コンピュータ特有の記号が使われる。頻繁に使われる演算記号には次のようなものがある。右側はセルを使ったときの計算式の例である。

+	加算	=C2+D2
-	減算	=C3-D3
*	掛け算	=C4*D4
/	割り算	=C5/D5
^	べき乗(累乗)	=C6 ^ D6
&	文字列演算子 (文字の結合)	=C7&D7

Excel で演算記号を入力するときは、/(割り算)、*(掛け算)、-(減算)、+(加算)については、キーボード右側のテンキーからも入力することができる。その際にはNumLock キーを先に押し、数字入力モードにしておく。

なお、数式をいったん設定したセルでは、セルの値を再入力すると自動的に再計算される。これを再計算機能と呼ぶ。

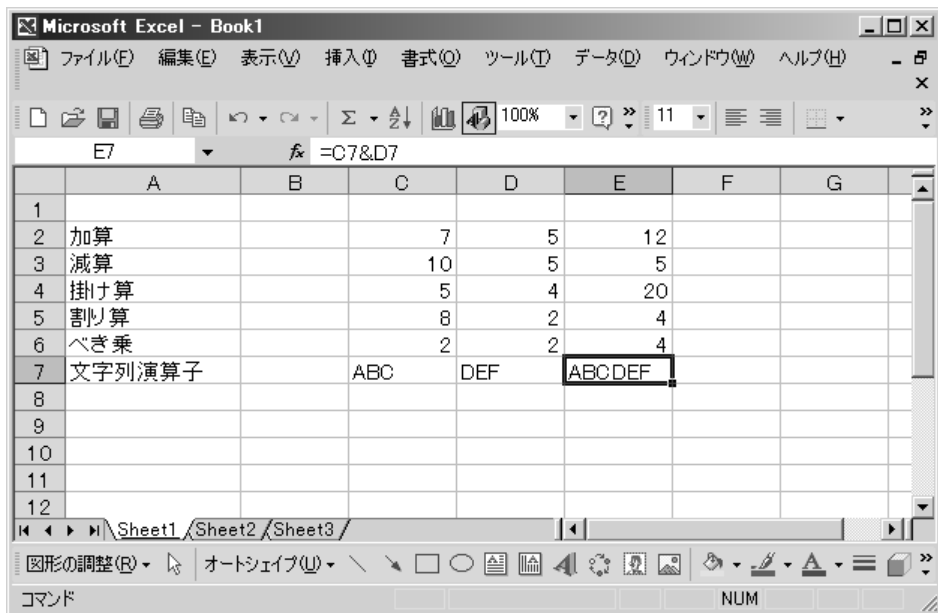


図 3.2 四則演算の例

3. 3. 数式とオートフィル

データを自動的に入力するオートフィル機能は、数式の入力にも使うことができる。図 3.3 はデータを入力し、数式バーに計算式を入力したところである。その後オートフィル機能を使えば、図 3.4 のように計算式がドラッグしたセルに複写され、同時に自動的に計算も行われる。この機能は複数の列に対しても適用され、その場合はそれぞれの列に指定した演算式によって計算される。

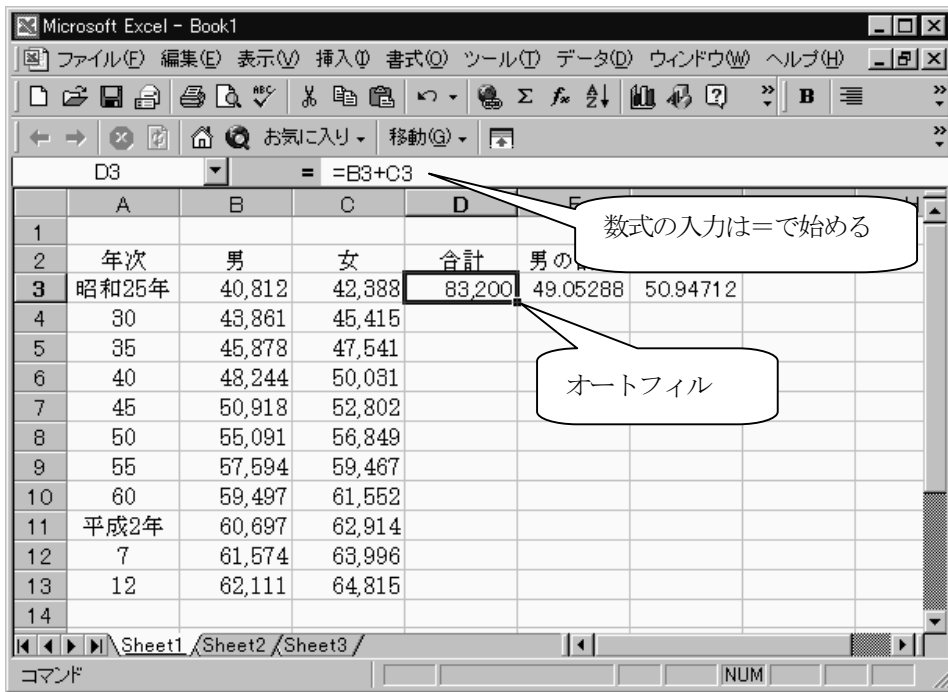


図 3.3 数式バーと計算式の入力(オートフィル実行前)



図 3.4 オートフィルによる自動計算

3. 4. データのソート(並べ替え)と累積和

データの数が多くなると、順番に並んでいないと役に立たないことがある。アルファベット順の辞書、五十音順の名簿や電話帳など、データが順番に並んでいる例はたくさんある。これらは順番に並んでいるからこそ、必要な項目を手作業でも検索できるわけである。



図 3.5 さまざまなデータのソート

データを順番に並べ替えることをソート(並べ替え)という。Excel にもこの機能が備わっており、「昇順で並べ替え」と「降順で並べ替え」のアイコンがツールバーに用意されている。

昇順は小さい順番であり、降順は大きい順番である。数値だけでなく、文字をアルファベット順にソートすることもできる。日本語の漢字はその読み(五十音順)でソートされる。

図 3.5 はアルファベットのソートの例を示している。日本語のかなや漢字あるいは中国語のソートを試してみよう。ここでソートを試す場合は、ソート前のデータと比較するため、元のデータをコピーしてから試すとよい。コピーするときは、元のデータをマウスで選択しその境目にマウスポインタ(カーソル)を置くと形が変わるので、Ctrl キーを押しながら移動先までドラッグする。

3. 5. 累積和

累積和というのは、データを前から順番に足し合わせた合計である。例えば、図 3.6 のサンプルデータを上から順番に合計したもの(2+5+6+...+10=55)が累積和となる。

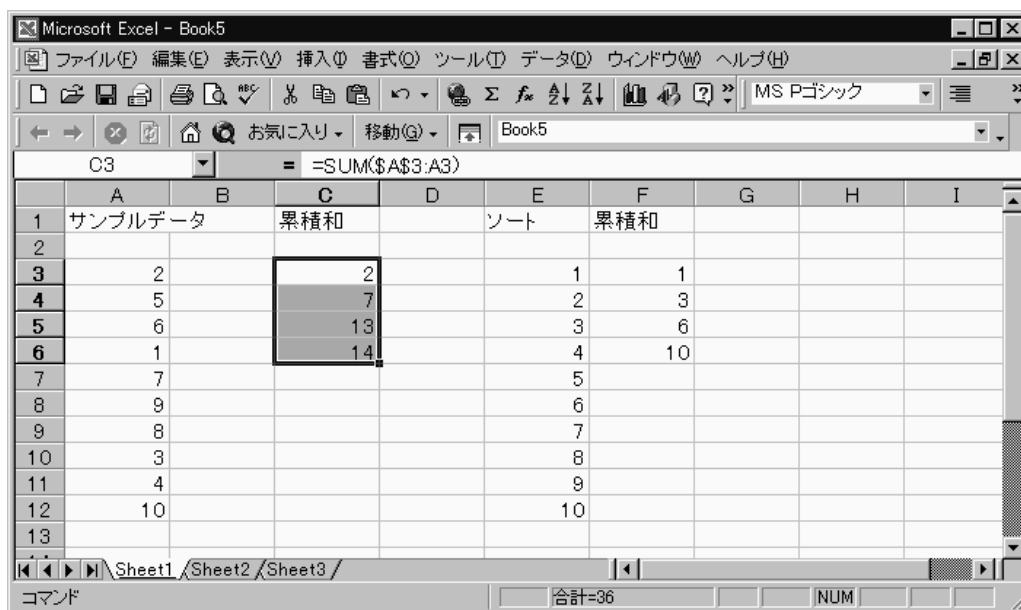


図 3.6 累積和の計算

累積和の計算は、 2 、 $2+5=7$ 、 $7+6=13$ 、…のように前から順番に足し合わせた結果である。

累積和を求めるときは、SUM 関数を使うことができる。図 3.6 の例では、セルの C3 をマウスで選択してアクティブにし、次のように関数を入力し、フィルハンドルにカーソルを合わせてドラッグする(オートフィル機能を使う)。SUM 関数の名前は小文字も使用できる。

$=SUM(\$A\$3:A3)$

3. 6. 絶対参照と相対参照

上の例では、C4, C5, C6…のそれぞれのセルには、順番に $=SUM(\$A\$3:A4)$ 、 $=SUM(\$A\$3:A5)$ 、 $=SUM(\$A\$3:A6)$ 、…と数式が複写され、累積和が計算される。

このとき数式 $=SUM(\$A\$3:A3)$ の中の \$ マークを付けた $\$A\3 はオートフィルでは変化しない。これを 絶対参照 と呼ぶ。

これに対して、\$ が付いていない A3 のほうは、ドラッグするに従い 順次 A4, A5, A6… と変化する。これを 相対参照 と呼ぶ。

3. 7. 平均・中央値・最頻値

データの特徴を表す代表的な値として、平均 (average) , 中央値 (median) , 最頻値 (mode) がある。ここではこれら用語の定義や求め方を説明する。

(1) 平均 (average)

n 個のデータの平均は、データの和をデータの個数で割ったものであり、次の式で求められる。なお平均 M を \bar{x} を使って表すことがある。

$$M = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{数式 3-1})$$

Excel でデータの平均を求めるには、AVERAGE 関数を使う。平均値を表示したいセルを先に選択しておき、ここでは数式バーに AVERAGE 関数をキーボードから手入力する(図 3. 7)。



図 3.7 平均・中央値・最頻値を求める

データは図 3. 7 では、A2 から A11 の範囲に入力されている。D3 セルに=AVERAGE(A2:A11) と入力し、Enter キーを押す。

関数を入力するときは、必ず先頭に=を付ける。データの範囲は A2:A11 のように : (コロン記号) を用いて表す。

式の先頭は=から始まることに注意せよ。

(2) 中央値 (median)

データを大きさの順番に並べたとき、ちょうど真中にくる値のことを中央値またはメディアンあるいは中位数という。データが偶数のときは、中央にくる値が2つになるので、その平均を中央値とする。中央値を求めるにはMEDIAN関数を用いる。図3.7では4と5の平均4.5が中央値になる。

(3) 最頻値 (mode)

データのなかで最も多い値のことをモードまたは最頻値という。モードを求めるには、MODE関数を用いる。図3.7では1のデータが最も多いので、モードは1になる。

》》 演習 3 《《

次の演習を行ってみよ。ファイル名はbunseki3.xlsとし、必要に応じてシートを使い分けよ。

1. 1から100までの累積和を計算してみよ。
2. 掛け算九九の表を作成してみよ(相対参照と絶対参照を工夫して使うこと)(図3.8参照)。

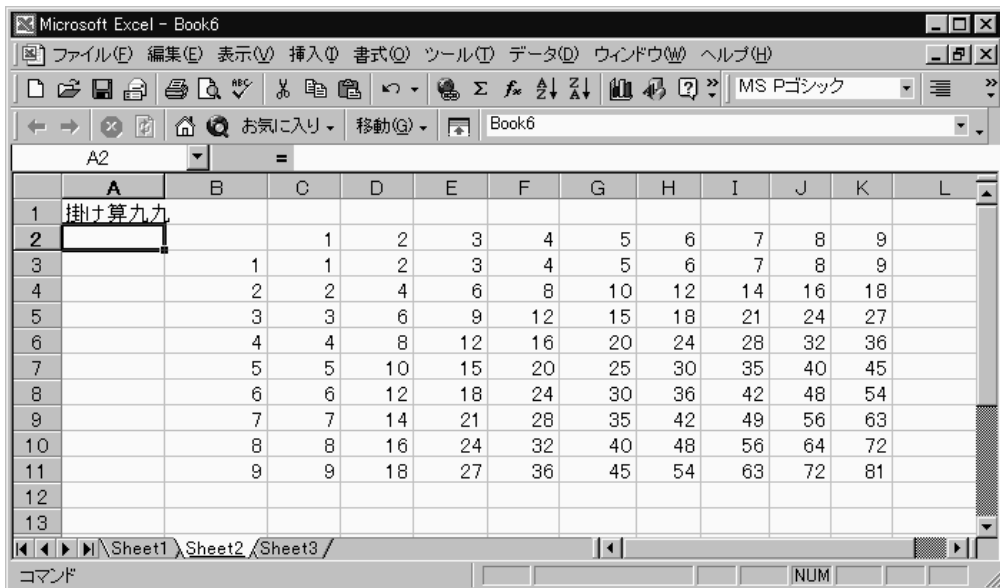


図3.8 掛け算九九の例

3. 愛知県の名古屋を除く市の人口の累積和を求めよ。データは以下のURLから参照する。

<http://www.stat.go.jp/data/nihon/index.htm>

4. 名古屋市のそれぞれの特別区の人口が、名古屋市全体の人口に占める割合を計算してみよ。データは上記のURLを参照せよ。

4. 移動平均

あるデータ x の値に対して、それに対応するデータ y の値があり、 x の変化に対して y の変化が大きいとき、折れ線グラフなどで描いても y の変化を読み取ることが難しいときがある。

このような場合に y の変化を小さくして分かりやすくデータを平滑化する方法があり、移動平均法はその1つである。

例えば日本の3月は、気温の上下を繰り返しながら、しだいに月末に向かって気温が上がっていくことを我々は経験から知っている。しかし毎日の気温の変化が大きいと上昇傾向にあるのか、下降傾向にあるのか変化を読み取ることが意外に難しい時期がある。このようなときは、日付を x 、気温を y として移動平均を取ると、気温の変化が視覚的に分かりやすくなることがある。

また x の値が一定の時間間隔で変化し、 y の値も対応して順次変化するとき、このようなデータのことを時系列データと呼ぶ。例えば気温の変化、株式相場、為替相場などが時系列データの例である。

4. 1. 気温の変化と移動平均

表1はある年の名古屋の気温(3月)を調べたものである。このデータを使って気温の移動平均を求めてみよう。

Excel には移動平均を求める関数が組み込まれているので、その使い方を覚えれば手軽に移動平均を求めることができる。

日付	最高気温	最低気温
3/01	11.7	7.3
3/02	11.0	4.7
3/03	11.4	-0.1
3/04	10.2	6.1
3/05	8.3	0.5
3/06	13.7	2.8
3/07	13.3	6.0
3/08	8.6	0.7
3/09	7.5	-0.4
3/10	7.9	0.2
3/11	8.9	-0.5
3/12	11.6	-1.9
3/13	10.2	0.2
3/14	15.4	-0.1

3/15	15.5	7.8
3/16	14.7	2.9
3/17	8.0	4.6
3/18	15.6	7.9
3/19	18.4	4.3
3/20	20.4	6.1
3/21	21.4	8.1
3/22	19.6	7.3
3/23	18.4	10.3
3/24	20.0	7.6
3/25	17.7	12.2
3/26	17.7	11.8
3/27	15.6	4.5
3/28	14.9	6.6
3/29	12.9	10.0
3/30	11.7	5.0
3/31	11.0	2.6

4. 2. 移動平均の求め方

$x_1, x_2, x_3, \dots, x_n$ という時系列データがあるとき、移動平均は次の式で計算される。ここでは連続する3つのデータの平均をとって移動平均とする例を示しているため、3項移動平均と呼んでいる。

$$\frac{x_1 + x_2 + x_3}{3}, \frac{x_2 + x_3 + x_4}{3}, \frac{x_3 + x_4 + x_5}{3}, \dots, \frac{x_{n-2} + x_{n-1} + x_n}{3} \quad (\text{数式 4-1})$$

なお連続するデータの取り方で、2項移動平均、4項移動平均などもある。

4. 3. 移動平均のグラフ

移動平均のグラフを描くには、まず元になる気温データを使って折れ線グラフを作成する。ここでは最高気温と最低気温の両方をグラフにする。

表1のデータをExcelに入力し、グラフに必要な部分(日付、最高気温、最低気温のデータ)を選択する。次にグラフウィザードを起動し、折れ線グラフを選択してグラフを描く(図4.1)。

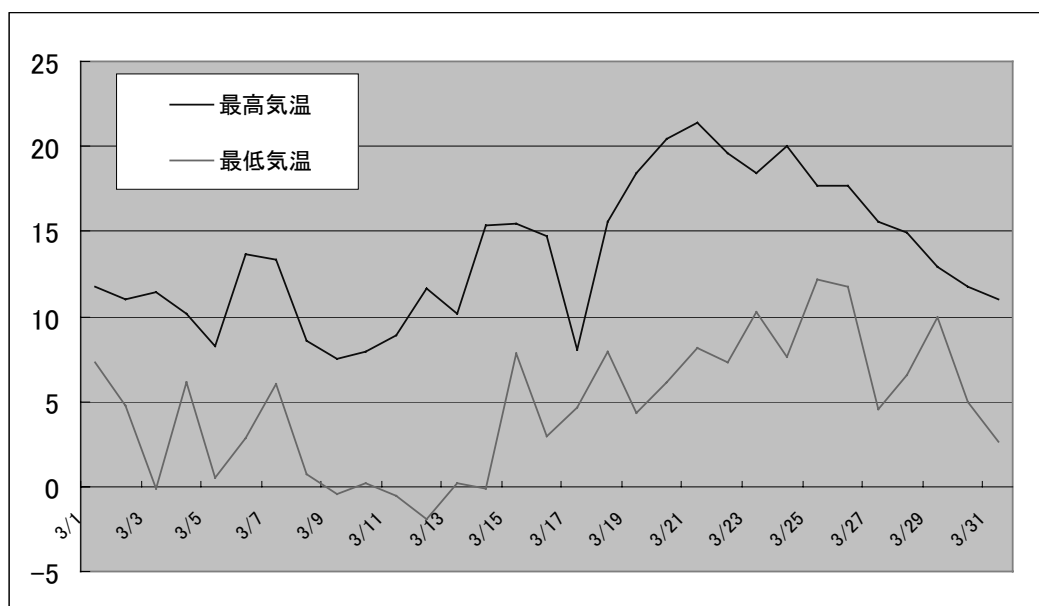


図 4.1 気温の折れ線グラフ

移動平均のグラフを描くためには、次の手順で重ね書きを行う。

- (1) 移動平均のグラフを重ね書きするには、上の手順で気温データの折れ線グラフを先に描いておく。
- (2) プロットエリアで最高気温の折れ線を左ボタンで選択する。
- (3) 折れ線の上に小さな四角形が表示されるので、その上でマウスの右ボタンを押す(図4.2)。
- (4) 「近似曲線の追加」から「移動平均」を選び、「区間」を3にする(図4.3)。
- (5) 「OK」ボタンを押すとグラフが作成される(図4.4)。

(6) 最低気温の移動平均も上と同様に作成できる.

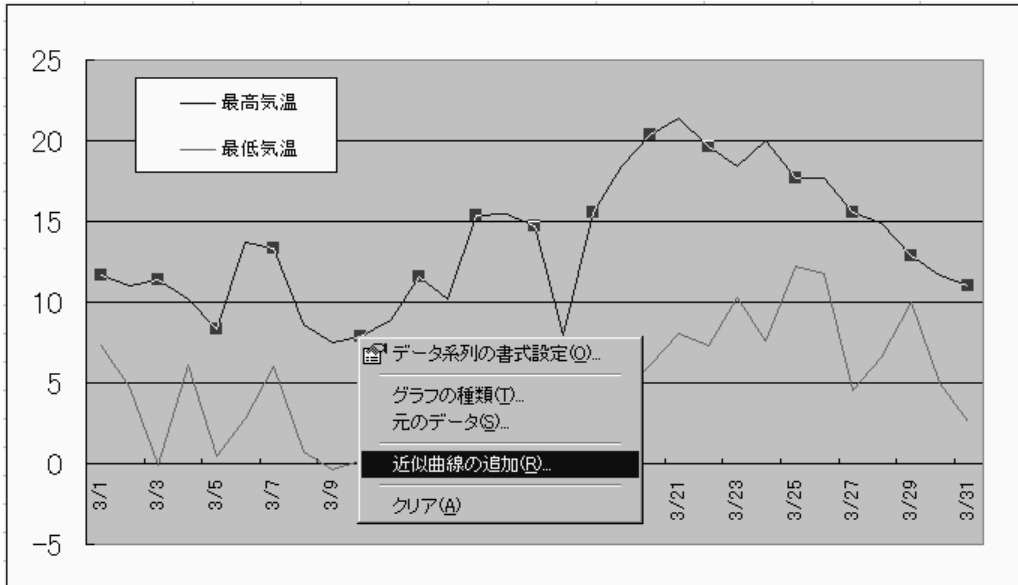


図 4.2 近似曲線の追加メニュー



図 4.3 近似曲線の追加ダイアログボックス

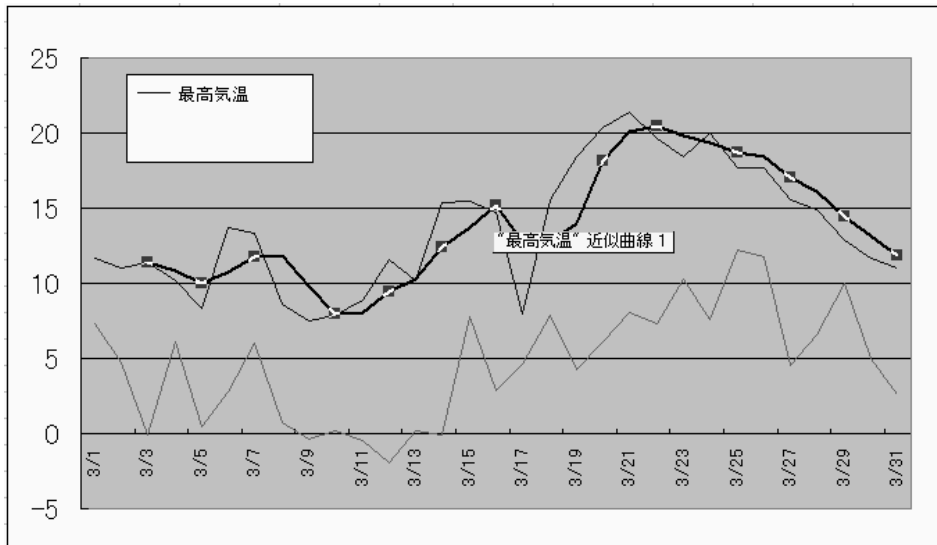


図 4.4 作成された近似曲線

4. 4. 地表の平均気温

最近地球の温暖化が進んでいるといわれている。以下の気温データを使い、地表の平均気温が変化している様子を適切な移動平均グラフに示してみよう。

またこのまま温暖化が進むと、人間生活にどのような影響が現れそうかを検討せよ。

(出典：地球環境データブック，ワールドウォッチ研究所，2001-02，p. 61)

表 2. 地表の平均気温 (1970-1999)	
年	気温
1970	14.03
1971	13.94
1972	14.01
1973	14.11
1974	13.93
1975	13.94
1976	13.81
1977	14.11

1978	14.04
1979	14.09
1980	14.18
1981	14.30
1982	14.09
1983	14.28
1984	14.14
1985	14.10
1986	14.16
1987	14.29
1988	14.33

1989	14.25
1990	14.41
1991	14.38
1992	14.13
1993	14.13
1994	14.23
1995	14.39
1996	14.31
1997	14.41
1998	14.59
1999	14.36

また上のデータでグラフを作成したときに西暦が表示されないときは、次の順番に操作して試みる。
 グラフの上で右ボタンをクリック→元のデータ→項目軸ラベルに使用→(西暦の)セルの範囲を指定→OK を押す。

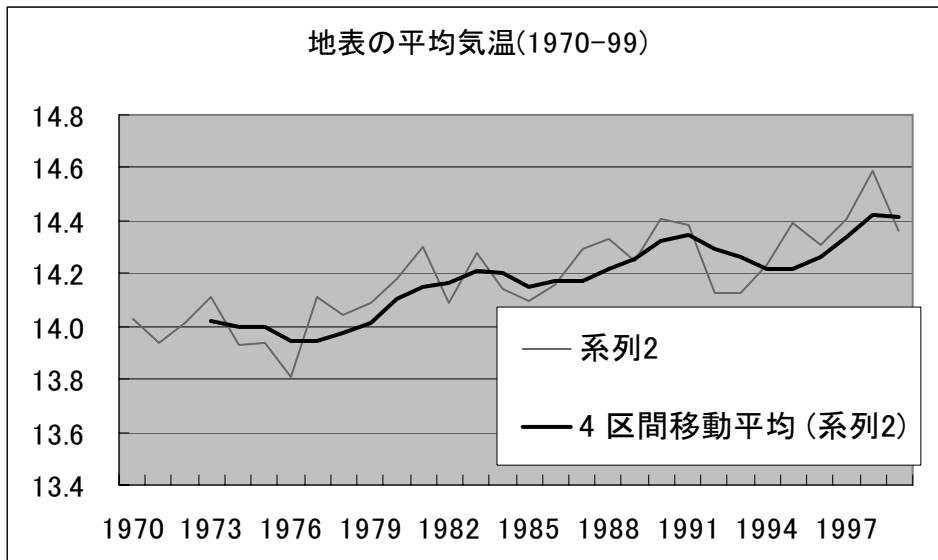


図 4.5 地表の平均気温の変化

4. 5. 地表の平均気温と名古屋の平均気温の比較

上の地表の平均気温のデータと名古屋の気温の変化と比較してみよう。

名古屋市統計年鑑(統計名古屋 Web 版)に年別の平均気温が公開されている。年号の部分を書西暦に変更して作成すると、上で作成したグラフと比較しやすい。平均気温のデータも西暦に対応するように、セルごとにコピーするなど工夫してデータを作成する。

<http://www.city.nagoya.jp/stat/nenkan/nenkan.html>

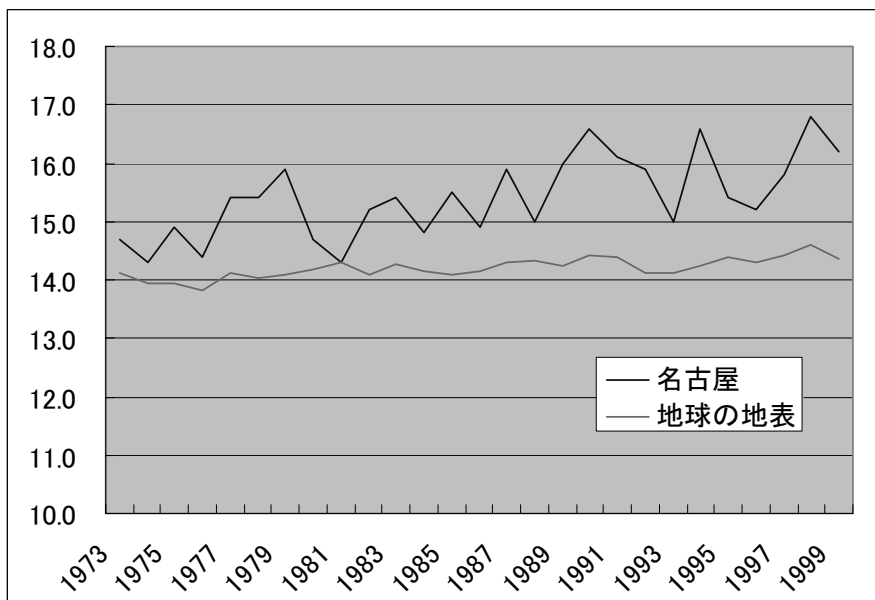


図 4.6 地表の平均気温と名古屋の平均気温の変化

》》》 演習4 《《《

以下の演習を行い、結果をファイルに作成せよ。ファイル名はbunseki4.xlsとする。

1. 上の説明を読み、実際に試してみよ。
2. 上の例題の時系列に対して、3項移動平均のほかに2項移動平均、4項移動平均のグラフを作成してみよ。

5. 高齢化・人口問題

日本は現在高齢化社会に向かっている。当分の間、65歳以上の人口割合が増加し、若年層の人口割合が減少するという状況になりつつある。この状況についてグラフを作成して確認したい。また人口ピラミッドを作成して、人口構成がどのようになっているかを考察しよう。

さらに高齢化社会ではどのような問題や課題が予想されるかを考えてみよう。なおファイル名を bunseki5.xls として保存せよ。

5. 1. 65歳以上老年人口の増加傾向

65歳以上の老年人口が、人口全体に対してどのような増加傾向あるかを把握するためグラフを描く。老年人口と他の世代との関係が分かりやすいグラフを作成したい。

「日本の統計、第2章人口・世帯、2-1人口の推移と将来人口」には、年齢3区分別人口構成比(%)の統計が公開されている。この統計には現在までの人口構成比の推移のほか、将来の推計値も記載されている。

また年齢3区分別人口というのは、0～14歳（年少人口）、15～64歳（生産年齢人口）、65歳以上（老年人口）の3つのことである。

ここではこれらの資料を参考にして面グラフを作成する。その場合に0～14歳（年少人口）、15～64歳（生産年齢人口）、65歳以上（老年人口）の3つの変化が分かるように作成する。将来の推計値を含めても良い。

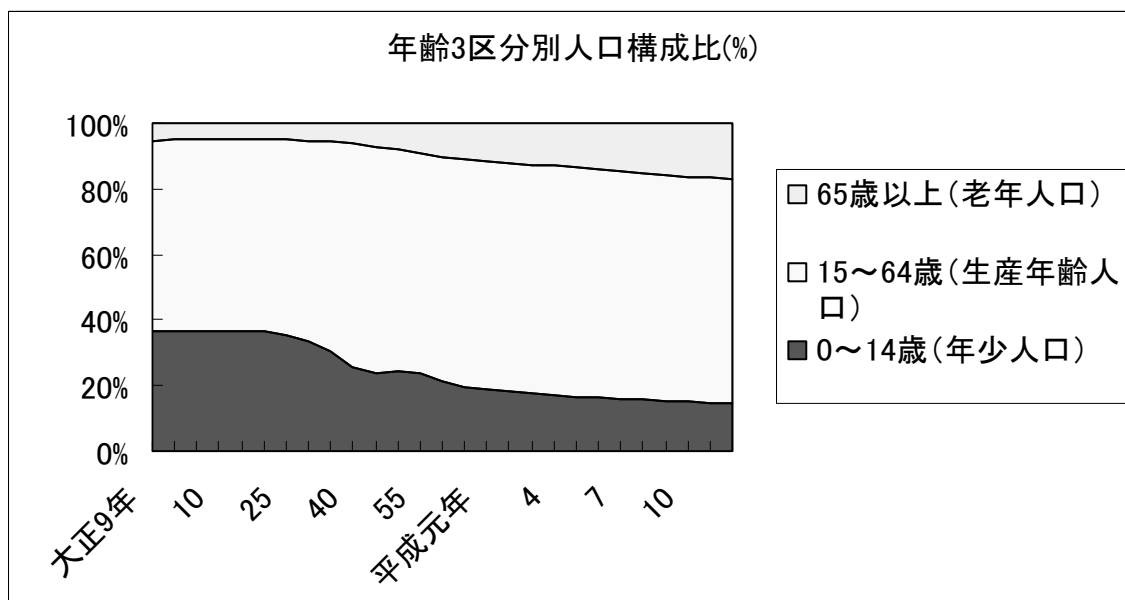


図5.1 年齢3区分別人口構成比 (%)

(参考資料)

<http://www.stat.go.jp/data/nihon/index.htm>

日本の統計 (総務省統計局統計センター)

第2章 人口・世帯

2-1 人口の推移と将来人口

年齢3 区分別人口構成比 (%)

5. 2. グラフ作成の手順

以下におおよその手順を示す。

- (1) インターネットから上記の統計をダウンロードする。
- (2) 必要な部分はどこかを把握する。
- (3) Excel のグラフウィザードを使い、ここでは面グラフに作成してみる。
- (4) 年齢3 区分がうまく表示されるようにグラフの種類を選ぶ。
- (5) グラフのタイトル、項目軸、系列などの表示を行う。
- (6) 作成したグラフの例を図5.1に示す。

5. 3. 年齢各歳別人口(男女別)

総務省統計局統計センターの「日本の統計、第2章、人口・世帯、年齢各歳別人口」によれば、日本人は男性のほうが女性より多く生まれることが明らかになっている。しかし総人口における男女の割合を調べると、近年では女性の人口が2%前後であるが男の人口より多いことも明らかになっている。

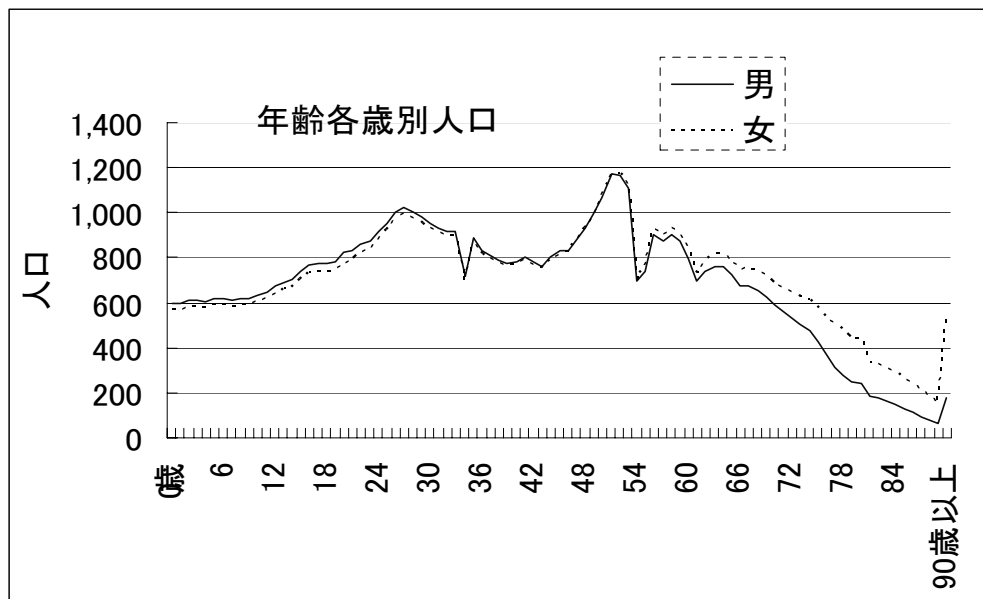


図5.2 年齢各歳別人口 (男女別)

そこでおおよそ何歳ぐらいで男女の人口が逆転するか、グラフを作成して調べてみよう(図5.2)。なお以下の点を考慮して行うこと。

- (1) どのようなデータが必要になるか？
- (2) データは公開されているか？
- (3) どのようなグラフが適当か？

5. 4. 人口ピラミッドの作成

従来から人口の年齢別構成をピラミッドのようなグラフで描くことが行われている。インターネットを検索してみると、さまざまな自治体が「人口ピラミッド」を作成して公開している。

人口ピラミッドという名前は、昔の人口構成グラフがピラミッドの形に描かれたことに由来する。しかし近代になって人口構成にゆがみが生じ、三角形であるはずのピラミッドの形が変形してきた。

ここでは最近における日本の人口ピラミッドを作成し、どのような人口構成に変化しているか、それを視覚的に把握できるようにグラフを作成してみたい。

作成方法はいくつか考えられるが、次の手順を参考に最近のデータを使って人口ピラミッドを作成してみよ。

(参考資料)

日本の統計(総務省統計局統計センター)

<http://www.stat.go.jp/data/nihon/index.htm>

第2章 人口・世帯

2-4 年齢各歳別人口

- (1) 必要なデータを集め、グラフを作成しやすいようにデータの編集する。
- (2) 男女別が分かりやすくする。
- (3) ピラミッドは横の棒グラフで作成するとやりやすい。
- (4) 左に男のデータ、右に女のデータ、中央に年齢構成がくるようにグラフの作成を工夫する。
- (5) 男のデータを左側に書くため、データの前にマイナス符号(-)を付けるなど、男のデータを負の値として扱う工夫をする。例えば男のデータに(-1)をかけ、セルの書式設定を使って、小数点以下を四捨五入すると負の値となる。
- (6) グラフの棒の上で右ボタンをクリックすると、「データ系列の書式設定」というメニューが出る。その中でオプションを選ぶとグラフの棒の重なりを調整することができる。棒の間の隙間をなくすときは「棒の間隔」を0にし、「棒の重なり」を100にする。

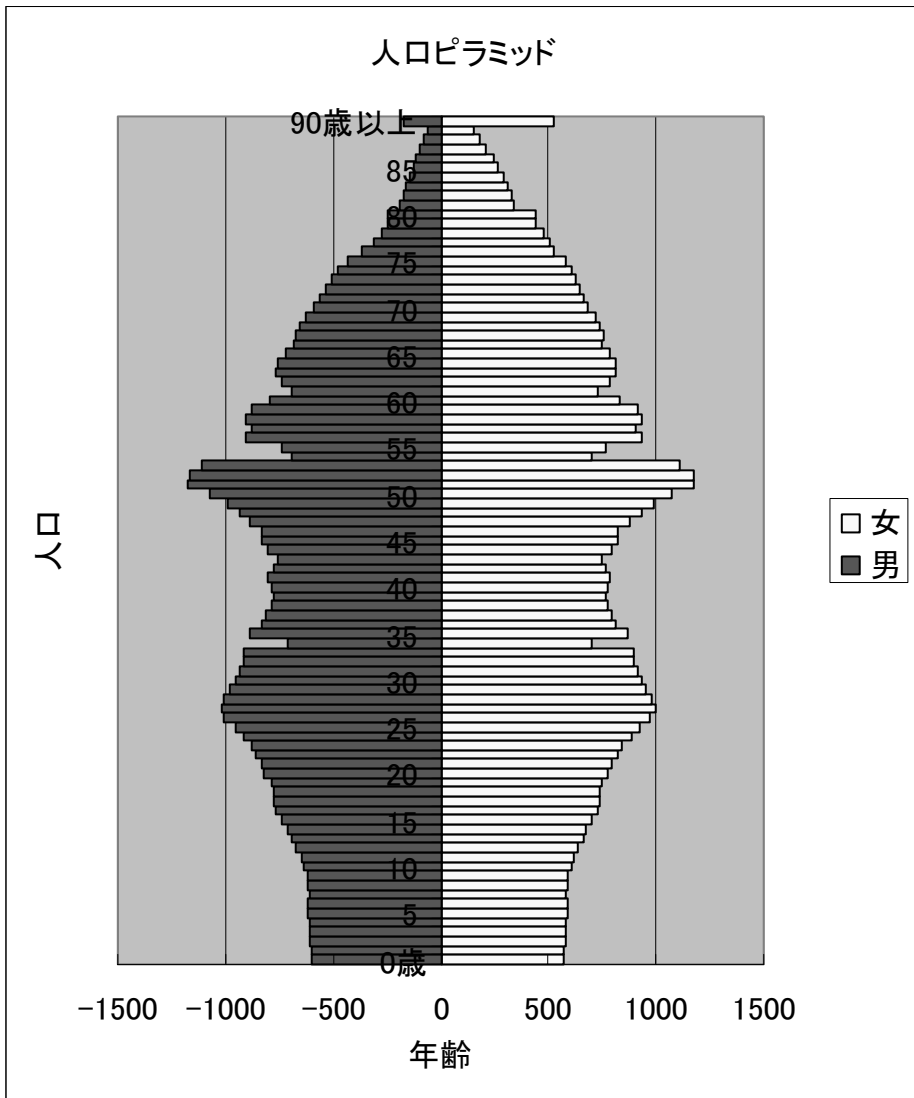


図5.3 人口ピラミッドの例

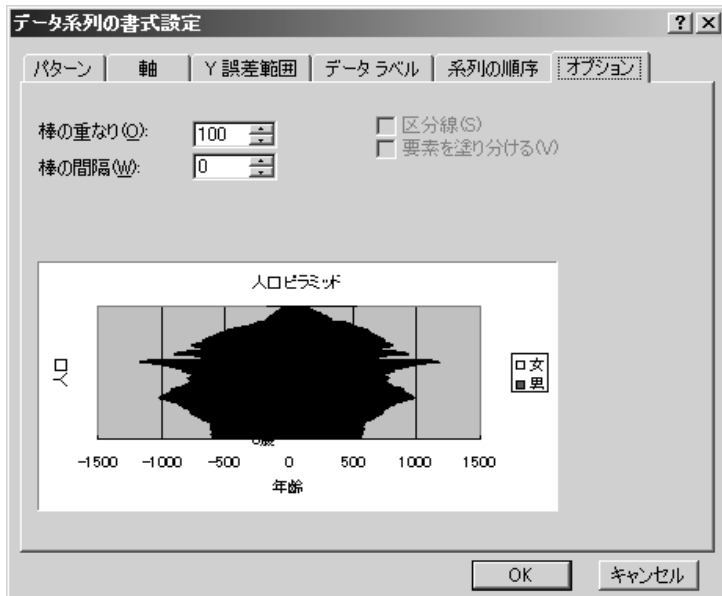


図 5.4 データ系列の書式設定

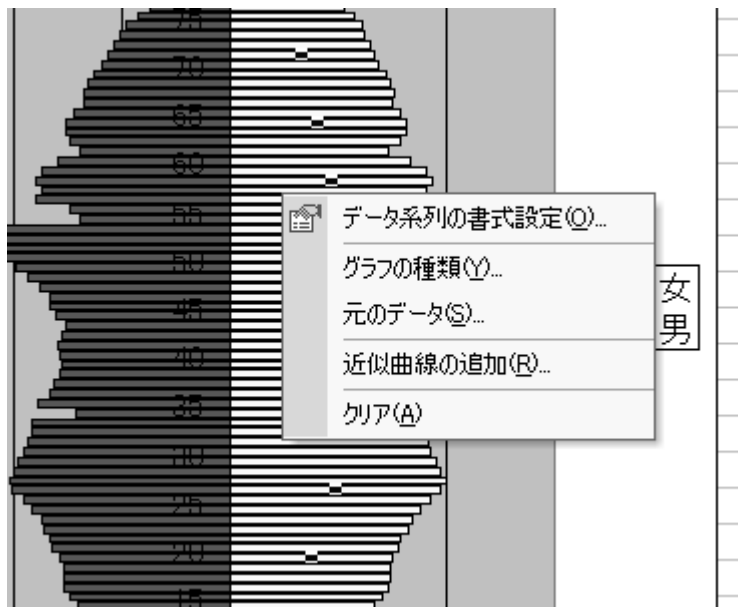


図 5.5 グラフの上で右ボタンをクリックし
データ系列の書式設定を開いたところ

》》 演習 5 《《

上の人口ピラミッドの作成において、年齢の幅を5歳ごとに小計を求め、そのデータにもとづいて人口ピラミッドを作成してみよ。

6. データの分散・偏差値・条件判断

テスト結果などのデータには、統計的にいくつかの特徴を見出すことができる。ここではデータの分散および偏差値などについて取り上げる。また成績判定を例に条件判断の使い方を説明する。

6. 1. データの散らばり

例えば中国語のテストを行ったとき、AクラスとBクラスの平均点が偶然にも同じだったと仮定する。このときAクラスとBクラスの学生は、同じような中国語能力を持っていると判断してよいだろうか。

表 6.1 は 2 つのクラスにおいて、中国語テストの平均を求めたものであるが、平均は両方とも同じになっている。しかし平均はデータの散らばり(度数の分布)を無視して得られる指標であるから、そのことを理解してデータを見る必要がある。

表 6.1 を眺めてみると、Aクラスには得点の高い人もいれば得点の低い人もいて、能力はばらばらである。これに対してBクラスの得点はAクラスよりかなり接近した値になっていることが分かる。

このようにデータをざっと眺めただけでも、2つのクラスは似かよってはいないらしいことが推測できる。

表 6.1 中国語テストの結果 (10点満点)														平均		
Aクラス	3	6	9	1	10	4	8	1	4	3	10	3	5	9	1	5.1
Bクラス	4	8	3	5	6	5	3	6	7	8	3	7	5	3	4	5.1

データのばらつきや散らばりの度合いを表す尺度として、分散、偏差、標準偏差、範囲などがある。これらの尺度を用いて、データの特徴を推定することができる。

(1) 偏差

偏差とはそれぞれのデータと平均との差である。データ x と平均の差は、 $x - \bar{x}$ である。なお \bar{x} は平均を表している。

(2) 分散 S^2

分散はそれぞれのデータが、平均からどのぐらい差があるかという偏差の平均を表す指標である。つまりデータのばらつきの平均といえ、データが平均のまわりにどのように散らばっているかを示す値である。

分散は以下の式で求められる。

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{数式 6-1})$$

また分散が S^2 というように 2 乗で表されるのは、偏差の平均が常に 0 になるためである。つまり偏差の平均を求めると、平均 - 平均 = 0 となる。

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{\sum_{i=1}^n x_i}{n} - \frac{n\bar{x}}{n} = \bar{x} - \bar{x} = 0 \quad (\text{数式6-2})$$

(3) 標準偏差(standard deviation)

標準偏差は統計的な対象となる値が、その平均値からどれだけ広い範囲に分布しているかを計量したものである。標準偏差は分散 S^2 の正の平方根であり、テスト結果などの偏差値の計算に用いられる。

(4) 範囲(レンジ)

データの最大値(maximum)と最小値(minimum)の差を範囲またはレンジ(range)という。

6. 2. 分散を求める

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H
1	中国語テストの結果							
2		Aクラス	Bクラス					
3		3	4		1	3		
4		6	8		1	3		
5		9	3		1	3		
6		1	5		3	3		
7		10	6		3	4		
8		4	5		3	4		
9		8	3		4	5		
10		1	6		4	5		
11		4	7		5	5		
12		3	8		6	6		
13		10	3		8	6		
14		3	7		9	7		
15		5	5		9	7		
16		9	3		10	8		
17		1	4		10	8		
18								
19	平均	5.133333	5.133333					
20	分散	10.24889	3.048889					
21	標準偏差	3.201389	1.746107					
22								

図 6.1 平均値が等しいときの分散

表 6.1 では2つのクラスの平均は同じであるが、分散を求めるとデータのばらつきの度合いを知ること
愛知大学情報処理センター

ができる。分散を求めるときに、母集団全体のデータが得られているときは、VARP 関数を用いる。

表 6.1 のデータを Excel に入力して次の手順で分散を求めてみよう。なおファイル名を buseki6.xls として保存せよ。

- (1) B 列と C 列に縦に、それぞれ A クラスと B クラスの得点を入力する。
- (2) 平均をそれぞれのクラスごとに求める。
- (3) B19 セルに=AVERAGE(B3:B17)を入力し A クラスの平均を求める。次にオートフィルを使って C19 セルに式を複写し、B クラスの平均を求める。
- (4) B20 セルに=VARP(B3:B17)と式を入力し、A クラスの分散を求める。さらにオートフィルを使って C20 セルに式を複写し、B クラスの分散を求める。
VARP(B3:B17)はB3からB17までにあるデータの分散を求めるという意味になる。VARP 関数は、引数(ここではB3:B17のこと)を母集団全体であると見なして分散を求める。指定する数値が母集団の標本(サンプル)である場合は、VAR 関数を使って分散を計算する。
標本は、母集団から選び出した事例や要素の部分的な集合のことである。
- (5) B21 セルに=STDEVP(B3:B17)を入力し、A クラスの標準偏差を求める。次にオートフィルを使って C21 セルに式を複写し、B クラスの標準偏差を求める。
ここでは母集団全体のデータがはっきり得られているため、関数は STDEVP を用いる。

図 6.1 の分散の計算結果から、分散の値が小さいほうがデータのばらつき(散らばり)は小さく、逆に分散の値が大きいほうがデータのばらつきは大きいことが分かる。

A クラスの分散は 10.24889 となり、B クラスの分散は 3.048889 になるので、A クラスのほうがデータのばらつきが大きいことが読み取れる。ここでは A クラスの中国語能力にばらつきがあることを知ることができる。

また 2 つのクラスの標準偏差の値を比較してみると、分散の値が大きいと標準偏差の値も大きく、分散の値が小さいと標準偏差の値も小さい。

6.3. グラフで確認

各クラスのデータを得点の低い順にソートし、グラフを作成して分散の計算結果を視覚的に把握してみよう。次の手順を参考に各自行ってみよ。

- (1) AクラスとBクラスのデータを、それぞれE列とF列にコピーする。
- (2) E列とF列をそれぞれ昇順にソートする。
- (3) グラフウィザードを起動して棒グラフで描く。

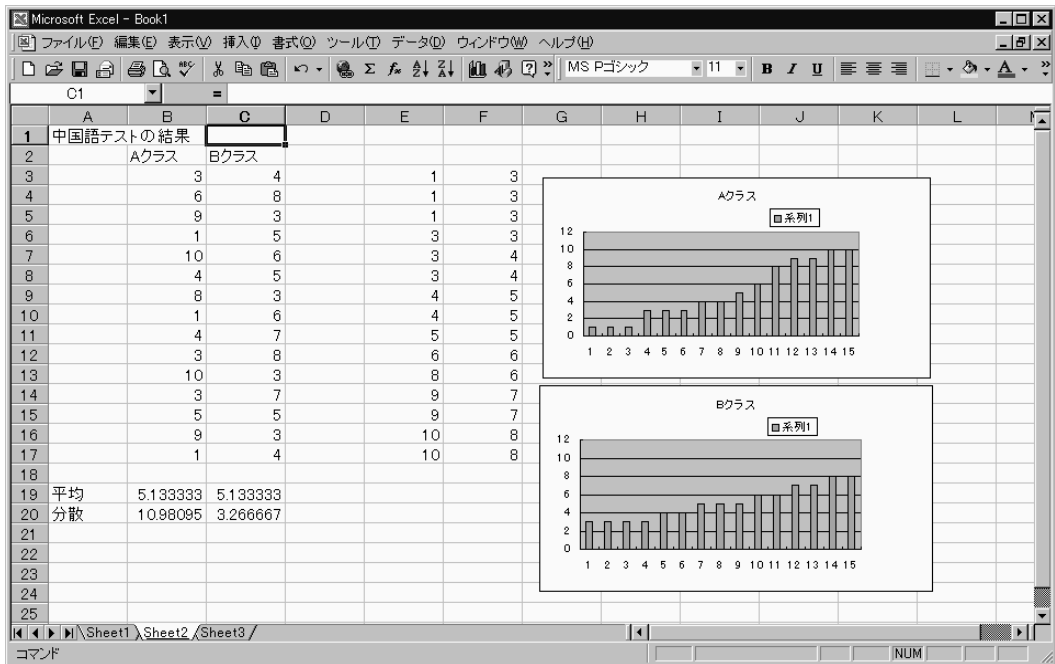


図 6.2 データの散らばりをグラフ化した例

6. 4. グラフの編集

グラフの目盛が同じにならないときは、目盛の表示してある角をクリックし、「軸の書式設定」を開き、「目盛」の値を変更してそろえる。

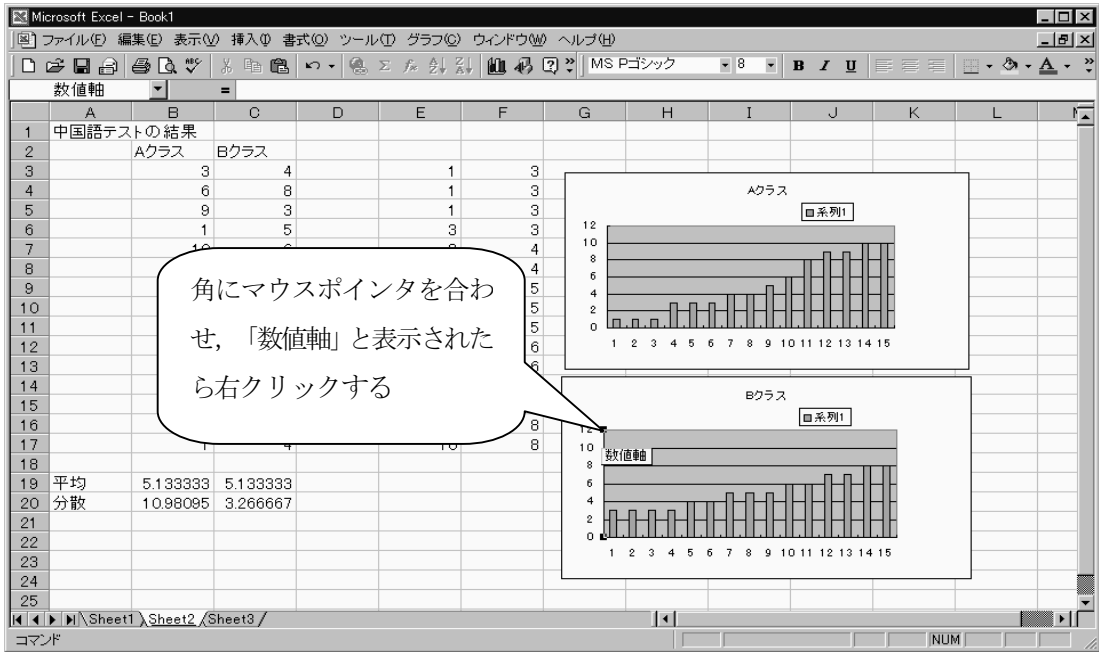


図 6.3 数値軸の編集

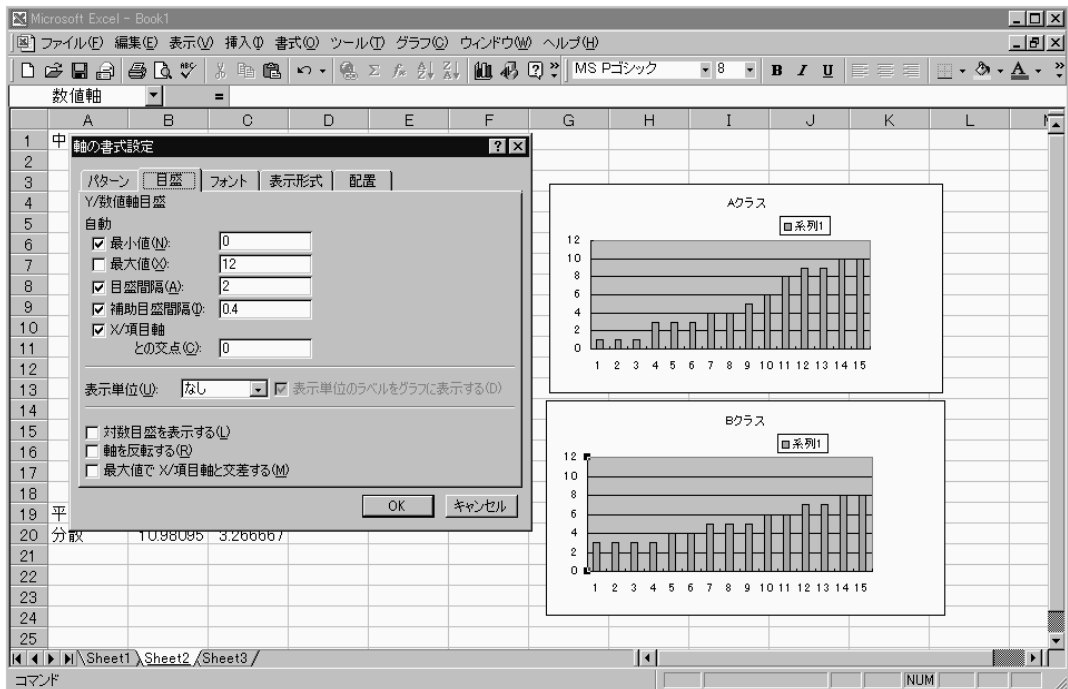


図 6.4 軸の書式設定

6. 5. グラフの重ね書き

グラフの重ね書きを行うと、2つのグラフのばらつきがより視覚的に把握できる。グラフの合成は次のように行う。

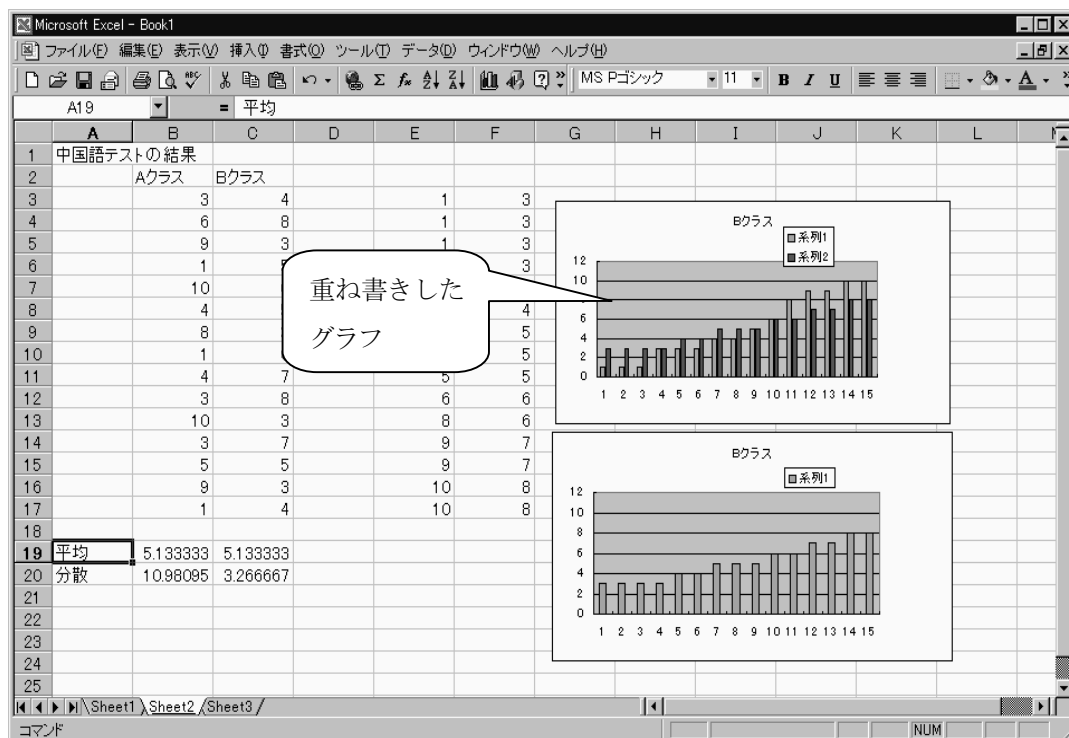


図 6.5 グラフの重ね書き

ここではBクラスのグラフをAクラスのグラフの上に重ねる例を示す。

- (1) Bクラスのグラフの上で右ボタンを押し、「コピー」を選択し、グラフをコピーする。
- (2) Aクラスのグラフの上で右ボタンをクリックし、「貼り付け」を選択し、コピーしておいたBクラスのグラフを重ね書きする。

6. 6. 偏差値と成績表

偏差値はテスト結果で成績の順位を表すのにしばしばもちいられる。テスト結果の素点だけでは、どの程度の順位なのか分かりにくいいため、順位をより分かりやすくしたのが偏差値である。

問題がやさしいときは得点の高い人が多くなり、逆に難しいときは得点の低い人が多くなるのが常識であるが、テストで80点を取っても、問題がやさしかったので80点以上の人がたくさんいたり、反対に難しすぎて80点以上の人がほとんどいなかったりすることもあり、自分の素点だけでは相対的な順位はなかなか分かりにくい。

偏差値は、テストの全ての得点を平均点が 50 点、標準偏差が 10 点となるように、テストの素点を換算したものである。なお偏差値を求める前提として、テストの得点が正規分布になっていることが必要である。

偏差値の求め方は、テストの得点を X 、平均点を m 、標準偏差を s としたとき、次の式で求められる。

$$\frac{X - m}{s} \times 10 + 50 \qquad \text{偏差値} = (\text{得点} - \text{平均点}) \div \text{標準偏差} \times 10 + 50$$

$\frac{X - m}{s}$ の部分は「標準化得点」と呼ばれ、平均からどのぐらい離れているかを示しており、標準偏差 s を

基準に計算したものである。

例えばあるクラスで行われた英語のテストの平均点が 65 点、標準偏差が 7 点だったとする。そのときに A 君の得点は 70 点であるとする、このテストにおける A 君の偏差値は、次のように計算できる。

$$\text{偏差値} = \{ (70 - 65) / 7 \} \times 10 + 50 = 57$$

6. 7. 偏差値を求める

成績表を作り偏差値や順位を求めてみよう。次のように A クラスでレポートとテストを実施したことにしてデータを用意し、上の定義式を入力して偏差値を求めてみよう。

- (1) 学籍番号、レポートとテストの得点など必要な項目を入力する。
- (2) 平均(AVERAGE)、分散(VARP)、標準偏差(STDEVP)、合計点(SUM)などを計算しておく。
- (2) 偏差値を求めたいセルに式を入力し、オートフィルを使って式を複製し、残りを計算する。
- (3) 絶対参照と相対参照の使い方に注意する。B 列、C 列、D 列を使って、E 列、F 列、G 列をそれぞれ計算するので、列を相対参照とし、セルの番号を絶対参照する。

6. 8. 式の入力

ここでは偏差値を計算すると同時に、小数点以下を丸めることも行いたので、ROUND 関数も使って偏差値を求める式を入力する。

(1) 絶対参照の利用

E3 のセルに次のように入力する。なお B\$24 および B\$26 はセルの参照が移動しないように、\$ マークを付けて絶対参照を指定する。

$$=ROUND((B3-B$24)/B$26*10+50, 0)$$

このように列には \$ を付けずに相対参照とし、セルの番号に \$ を付けて絶対参照とすれば、E3 のセルに

指定した計算式をF列とG列にオートフィルで複写できる。

1	Aクラスのレポートと試験の結果															
2	学籍番号	レポート	テスト	合計点	レポート 偏差値	テスト 偏差値	合計点 偏差値	順位	レポート 合否	合計点 SABCDF	S	A	B	C	F	
3	02x1001	70	75	145	51	53	53	7	合格	B						
4	02x1002	70	80	150	51	57	55	6								
5	02x1003	55	40	95	42	27	32	18								
6	02x1004	65	60	125	48	42	44	15								
7	02x1005	80	65	145	57	46	53	7								
8	02x1006	75	80	155	54	57	57	4								
9	02x1007	40	55	95	33	38	32	18								
10	02x1008	90	75	165	63	53	61	3								
11	02x1009	40	85	125	33	61	44	15								
12	02x1010	75	60	135	54	42	48	13								
13	02x1011	95	90	185	66	65	69	1								
14	02x1012	55	75	130	42	53	46	14								
15	02x1013	70	70	140	51	50	51	11								
16	02x1014	45	50	95	36	34	32	18								
17	02x1015	85	60	145	60	42	53	7								
18	02x1016	75	65	140	54	46	51	11								
19	02x1017	50	75	125	39	53	44	15								
20	02x1018	60	85	145	45	61	53	7								
21	02x1019	90	90	180	63	65	67	2								
22	02x1020	80	75	155	57	53	57	4								
23																
24	平均	68.3	70.5	138.8						人数						
25	分散	263.2	172.3	587.2						割合(%)						
26	標準偏差	16.2	13.1	24.2												
27																

図 6.6 成績表の作成と偏差値・順位の計算

(2) 絶対参照とF4 キー

絶対参照を示す\$記号は手で入力してもよいが、F4 キーを押しても指定できる。F4 キーを使うときはまず式を入力し、計算式を数式バーに表示させておく。その数式の上で絶対参照を指定したいセルにカーソルを移動し、F4 キーを押すと以下のB24の部分のように、\$記号を付けたり外したりして絶対参照の指定と解除ができる。B26に指定したいときは、カーソルをB26に移動して行う。

(以下はF4 キーを押して、B24の参照の指定を変更した例)

- =ROUND((B3-B24)/B\$26*10+50, 0) # B列も24セルも相対参照 (F4 キーを押さないとき)
- =ROUND((B3-\$B\$24)/B\$26*10+50, 0) # B列24セルが絶対参照
- =ROUND((B3-B\$24)/B\$26*10+50, 0) # B列は相対参照で24セルが絶対参照

=ROUND((B3-\$B24)/B\$26*10+50, 0)

B列は絶対参照で24は相対参照

(3) ROUND関数と四捨五入

ROUND関数は数値を四捨五入して指定された桁数に四捨五入して丸める関数である。書き方は次のようにする。

ROUND(数値, 桁数)

数値： 四捨五入の対象となる数値を指定する。上の例では計算式を指定している。

桁数： 数値を四捨五入した結果の桁数を指定する。0を指定すると小数点以下を四捨五入する。

6. 9. IF関数による条件判断と合否判定

IF関数は指定した条件による判断を行う関数のひとつである。

例えば成績処理を行っていて、得点が60点以上は合格、それ以外は不合格としたいなどの判断を行うために使われる。

あるいは90点以上はS、80点以上90点未満はA、70点以上80点未満はB、60点以上70点未満はC、60点未満をFと判定したいなどにも応用できる。

IF関数は条件式(論理式)を指定することによって、判断の対象が真(TRUE)であるか、偽(FALSE)であるかを判断する。そして真であるときは「真の場合」に指定した値を返し、偽のときは「偽の場合」に指定した値をそれぞれ返してよこす。

数式の値が真(TRUE)または偽(FALSE)のいずれかであるとき、その数式は論理式であるという。セルの値が特定の条件を満たすときに、別のセルに文字列などを表示させたいときにも論理式を使う。論理式を選択する際にも関数ウィザードを用いる。

IF関数の一般的な書き方は次のようになる。

=IF(論理式, 真の場合, 偽の場合)

上の式の意味は次のようになる。

=IF(条件を指定して, 条件に合っていたときにする処理, 条件に合っていないときにする処理)

(論理式)

(真の場合)

(偽の場合)

例えば次のようなIF関数を使った式をあるセルに設定したとしよう。セルというのはA3などのセルの番号のことである。

=IF(セル>=60, "合格", "不合格")

上の式の意味は、もしセルの値が 60 以上なら、このセルには合格と表示する。もしそうでなかったら不合格と表示するという意味になり、"" (ダブルコーテーション) で囲った文字がセルに表示される。

6. 10. より複雑な条件判断(IF 関数の入れ子構造)

例えば合格と不合格でなく、成績を A, B, C, D などのように同じ列に表示したいときは IF 関数の入れ子構造を使うことがある。例えば次のようにすると、同じ列に A, B, C, D を判断して表示することができる。

= IF(セル>=80, "A", IF(セル>=70, "B", IF(セル>=60, "C", "D")))

つまり、上の式を実行するとセルの値は 80 点以上なら A, 70 点以上なら B, 60 点以上なら C, それ以外は D を表示するようになる。このように上で示したような簡単な IF 関数を入れ子構造にして組み合わせると、より複雑な条件を処理させることができる。

次に示した比較演算子は、左右 2 つの値を比較し、結果として真(TRUE)または偽(FALSE)の論理値を返す(表 6. 2)。

比較演算子の種類	意味と使用例
= (等号)	左辺と右辺が等しい (A1=B1)
> (～より大きい)	左辺が右辺よりも大きい (A1>B1)
< (～より小さい)	左辺が右辺よりも小さい (A1<B1)
>= (～以上)	左辺が右辺以上である (A1>=B1)
<= (～以下)	左辺が右辺以下である (A1<=B1)
<> (不等号)	左辺と右辺が等しくない (A1<>B1)

》》》 演習 6 《《《

1. RANK 関数による順位の求め方

これまでに学習したことをもとにして、合計点の偏差値から RANK 関数を使って順位を求めてみよ。

(1) 学籍番号が最初の人順位の求める

順位を求めるために、ここでは合計点とその偏差値を先に計算しておく。順位は RANK 関数を用いて、20 人中の順位を求めることにする。

まず学籍番号が最初の人順位の求めるため、以下のように式を入力する。

=RANK(G3, G3:G22)

(2) 絶対参照の指定

次に上の式を他のセルにオートフィルで複写するために絶対参照の指定を行う。

(3) オートフィルで式の複写を行い、順位を求める。

2. IF 関数による成績判定

このレジユメのような成績表を作成し、IF 関数と論理式を使い、成績判定を行ってみよ。

(1) 合否の判定

テストの得点から合否を判定してみよ。なお 60 点以上を合格とする。列を挿入して行うとやりやすい。

(2) 成績判定

合計点から S, A, B, C, F のように成績判定を行ってみよ。S は 180 点以上, A は 160 点以上, B は 140 点以上, C は 120 点以上, これら以外は F とする。

(3) 上の成績判定の基準にしたがい、合計点から S, A, B, C, F に該当するそれぞれの人数を求める IF 関数を作成し、実際に求めてみよ。またクラス全体からみて、S, A, B, C, F の人数はそれぞれ何パーセントになるかを計算し、円グラフを作成してみよ (図 6.7)。

	S	A	B	C	F
人数	2	1	9	5	3
割合 (%)	10	5	45	25	15

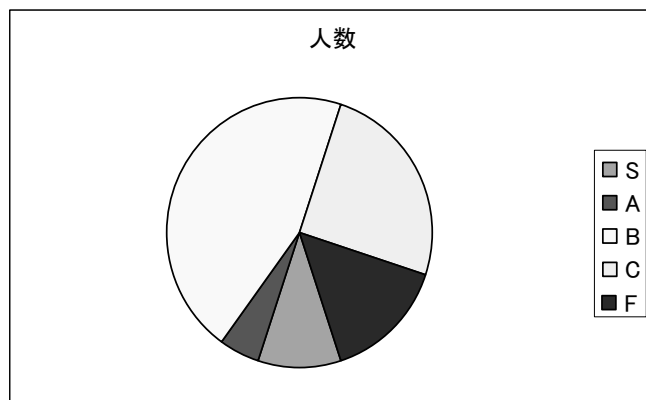


図 6.7 成績の分布を示す円グラフの例

7. 度数分布とヒストグラム

データの重要性を示す指標として度数分布がある。アンケートや実験などによってデータが得られたとき、最初に行う統計的な分析が度数分布表の作成である。表やグラフにするほうが、データ全体の分布の状況がはっきりと把握できる場合が多く、度数分布表のデータから作成したグラフをヒストグラムと呼び、ここではこれらについて取り上げる。

観測されたデータの分析においては、データを整理して役に立つ情報を取り出す方法が重要になるが、これらには基本となる一定の方法があり、記述統計と呼ばれている。

(1) 度数分布表

度数分布表は、データの取りうる値をいくつかの階級 (class) に分け、それぞれの階級に属するデータがいくつあるかを調べて表にしたものである。そのときにそれぞれの階級に属するデータの数が度数 (frequency) である。なおExcel では度数の代わりに頻度ということもある。

(2) 階級値

それぞれの階級を代表する値を階級値という。各階級のなかでは、データは等しく分布していると仮定し、階級の上限値と下限値の中間値を階級値とするのが一般的である。次の式で求められる。

$$\text{階級値} = (\text{下限値} + \text{上限値}) / 2$$

(3) 相対度数 (relative frequency)

データの特徴を把握するため、度数分布表においてそれぞれの階級に属するデータ数の割合を求めることが、多くの場合に重要になる。このようなときにデータ全体の総数 (総度数) を1と見なし、それぞれの階級に属する個数がデータ全体に占める割合を示したものが相対度数である。

$$\text{相対度数} = \text{度数} / \text{総度数}$$

(4) 累積度数と累積相対度数

累積度数 (cumulative frequency) と累積相対度数 (cumulative relative frequency) は、最も下の階級からその階級までの度数または相対度数を順番にそれぞれ加えたものである。

累積度数と累積相対度数はその階級の上限値未満のデータの累積値とその割合を表し、それぞれの最後の階級では、総度数および100%になる。

(5) ヒストグラム

データがどのように分布しているかを見るために、度数分布表をグラフにしたものがヒストグラム (histogram) または柱状グラフである。

(6) 階級の数と幅

度数分布表を作成するときは、階級の数と階級の幅をどのように取るかが極めて重要になる。階級の数を少なくしすぎるとデータが持つ情報の多くは失われることがある。また逆に階級の数を多くしすぎると、それぞれの階級の度数が小さくなりすぎてしまい、データの分析や整理などの本来の目的が果たせなくなる場合もある。

従ってデータ分析の目的をどこにおくかをよく検討して、階級の数と幅を決めるようにしなければならない。また階級を設定するときは、最小値と最大値を調べてから行う。

7. 1. 度数分布表の作成

ここでは前回用いたテスト結果を用いて度数分布表を作成してみよう。なおファイル名は bunseki7.xls として新たに作成する。

なおExcel で度数分布表を作成するときは分析ツールに含まれるヒストグラムを使うと便利である。ツールメニューに「分析ツール」が表示されていないときは、同じツールメニューの中にある「アドイン」から「分析ツール」を選びチェックしておく（図 7.1）。

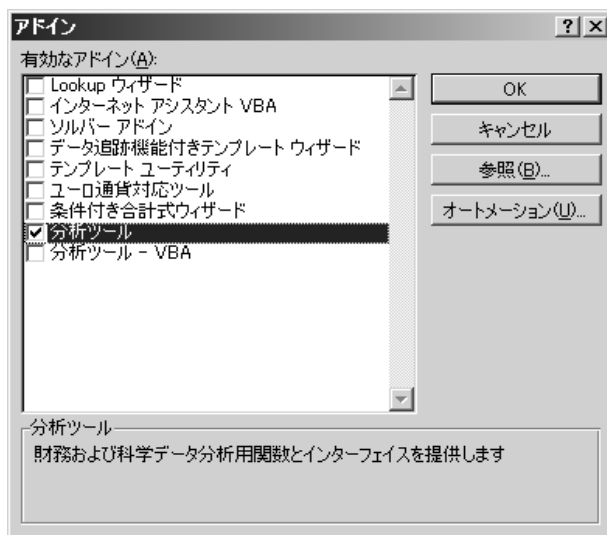


図 7.1 アドインの追加画面

(1) 必要なデータの準備

まず前回のシートから必要なテストの点数部分だけをコピーして、新しいシートに貼り付ける。貼り付けたら上で支持したファイル名を付けて保存しておく。

(2) 必要項目の入力

図 7.2 を参考にしながら、テスト、階級下限、階級上限、階級値、度数、相対度数、累積度数、累積相対

度数、合計などの項目を入力する。

	A	B	C	D	E	F	G	H	I	J	K
1	Aクラスのテスト結果の度数分布表										
2		テスト		階級下限	階級上限	階級値	度数	相対度数	累積度数	累積相対度数	
3		75		0	10						
4		80		10	20						
5		40		20	30						
6		50		30	40						
7		65		40	50						
8		80		50	60						
9		55		60	70						
10		75		70	80						
11		85		80	90						
12		60		90	100						
13		90		100	100						
14		75									
15		70		合計							
16		30									
17		40									
18		50									
19		75									
20		85									
21		75									
22		75									
23											

図 7.2 最初の準備するデータ

(3) 階級の設定

次に分析に必要な階級を設定する。今回は100点満点のテストにしているので、最小値は0（今回はなし）で最大値は100である。従って階級の幅はそれぞれ10とし、オートフィルを使って階級下限と階級上限を入力する。

(4) 階級値を求め方

階級値は次の式をF3セルに入力して求める。

$$=(D3+E3)/2$$

Excelでは階級の上限値を使って度数を計算することができる。階級の下限值はより大きいを意味してその階級には含まれず、上限値は以下を表しておりその階級には含まれる。下限より大きく、上限以下の度数が求められる。

(5) 度数の求め方

度数は分析ツールに含まれているヒストグラムを使って求めることができる。ツールメニューから「分析
愛知大学情報処理センター

ツール」を選び、さらに「ヒストグラム」を選ぶ。

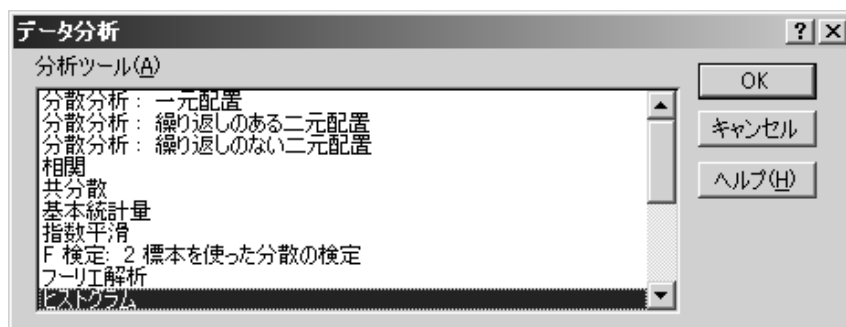


図 7.3 分析ツールの選択画面

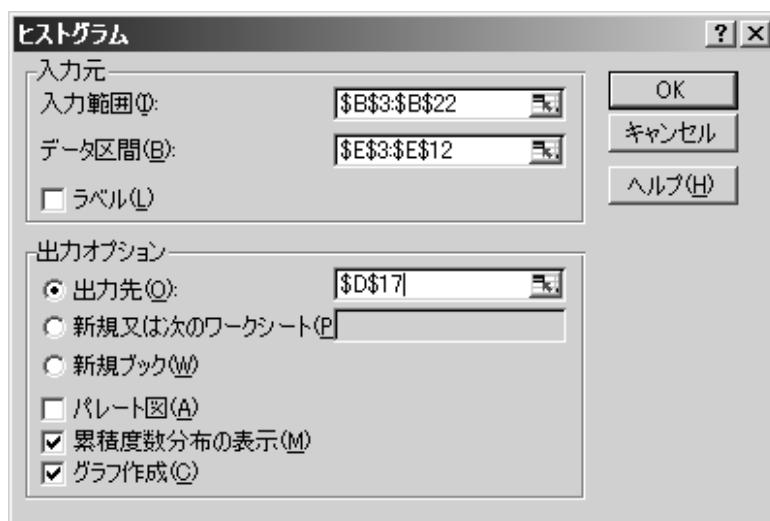


図 7.4 ヒストグラムの設定画面

ヒストグラムの設定画面で、入力範囲、データ区間、出力先を入力する。入力範囲はテストの得点全体とし、データ区間は階級値を指定する。なお出力範囲は同じワークシートにするので、表の下のほうで出力結果が表示される先頭のセルを指定する。新規又は次のワークシートを選ぶと別なワークシートに出力される。

入力範囲： \$B\$3:\$B\$22

データ区間： \$E\$3:\$E\$13

出力先： \$D\$17

(6) 累積度数分布とグラフ作成の指定

ここでは度数と同時に累積度数分布も求め、グラフも作成することにするので、それぞれのチェックボックスをクリックしてから OK ボタンを押して、度数分布表を作成する。

1	Aクラスのテスト結果の度数分布表									
2	テスト	階級下限	階級上限	階級値	度数	相対度数	累積度数	累積相対度数		
3	75	0	10	5	0	0.00	0	.00%		
4	80	10	20	15	0	0.00	0	.00%		
5	40	20	30	25	1	0.05	1	5.00%		
6	50	30	40	35	2	0.10	3	15.00%		
7	65	40	50	45	2	0.10	5	25.00%		
8	80	50	60	55	2	0.10	7	35.00%		
9	55	60	70	65	2	0.10	9	45.00%		
10	75	70	80	75	8	0.40	17	85.00%		
11	85	80	90	85	3	0.15	20	100.00%		
12	60	90	100	95	0	0.00	20	100.00%		
13	90									
14	75	合計			20	1.00				
15	70									
16	30									
17	40	データ区間	頻度	累積%						
18	50	10	0	.00%						
19	75	20	0	.00%						
20	85	30	1	5.00%						
21	75	40	2	15.00%						
22	75	50	2	25.00%						
23		60	2	35.00%						
24		70	2	45.00%						
25		80	8	85.00%						
26		90	3	100.00%						
27		100	0	100.00%						
28		次の級	0	100.00%						
29										

図 7.5 テスト結果の度数分布表

(7) 頻度（度数）と累積相対度数のコピー

表の度数および累積相対度数の列に、頻度と累積%をコピーして貼り付け、表を完成させる（図 7.5）。

(8) ヒストグラムのグラフ

度数分布と累積相対度数のグラフは次のようになる（図 7.6）。

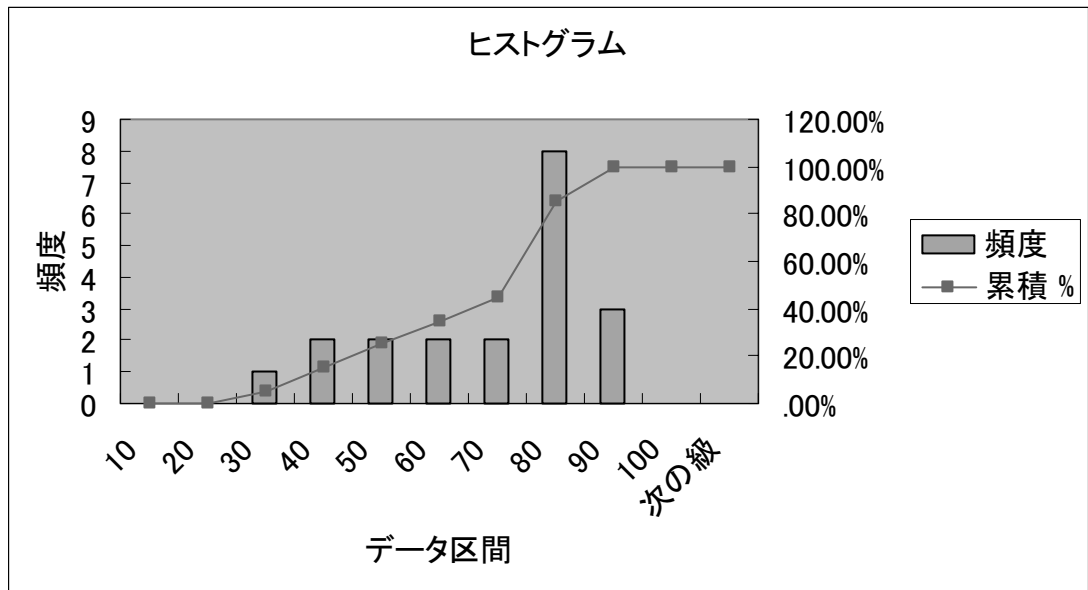


図 7.6 度数分布と累積相対度数のグラフ

8. 2次元データの分析(相関関係)

2次元のデータとは、 x と y の組からなるデータ (x, y) のことである。あるいは観測する対象 i について身長と体重のような2つの変数の観測値 (x_i, y_i) が同時に得られるようなデータのことである。2次元のデータでは2つの変数間の関係がどうなっているかを分析することが重要になる。

また観測値というのは、集めたデータのひとつひとつのことであり、それらを整理してまとめたものがデータである。

8. 1. 相関関係

2つの変数 x, y の関係を考えたとき、 x と y の間に区別をもうけないでお互いに対等に見る方法が相関 (correlation) である。例えば身長と体重などの関係は、身長が高い人は体重も多い人が多く、また身長が低い人は体重も少ない傾向が強いので、相関関係としてとらえるのが適している。しかし年齢と血圧あるいは人口と商業などは、それぞれ相関関係もあるが、年齢とともに血圧が上昇したり、人口の少ないところには商店が少ないなど、ある一方が他方を決定したり左右したりする関係にあるので、相関関係とは見なさないのが普通である。

8. 2. 散布図

表 8.1 はあるクラス 20 人の学生に対して行った数学と物理のテストの得点を個人別に一覧表にまとめたものである。この表を元に数学の得点と物理の得点にはどのような関係があるかを考えてみよう。

学籍番号	数学の得点	物理の得点
02x1001	7	8
02x1002	6	7
02x1003	7	6
02x1004	7	7
02x1005	5	5
02x1006	6	7
02x1007	6	6
02x1008	7	7
02x1009	5	6

02x1010	6	7
02x1011	6	6
02x1012	9	8
02x1013	8	8
02x1014	5	6
02x1015	8	9
02x1016	7	8
02x1017	8	8
02x1018	7	7
02x1019	9	9
02x1020	8	8

表 8.1 からは、数学と物理の得点の関係がどのようになっているかは分かりにくい。このような場合は、2つの変数の値を散布図に作成し、それらの値がどのように分布するかを調べることができる。

散布図 (scattergram, scatter diagram) は横軸に x 、縦軸に y をとった2次元のグラフ上で、 x と y の値がどのような関係になっているかを視覚的に把握しやすく表示するグラフである。

2次元のデータでは、 i 番目の観測対象について2つの変数の観測値 (x_i, y_i) が同時に得られる。2つの観測値がともに量的データのときは、これをグラフに描いてそれらの関係を視覚的に表示して考察すること

ができる。

視覚的な考察によって、2つの変数間に関係がありそうか、あるいはあまり関係がなさそうかなどを考えることができる。一般的な2次元のデータ分析では、散布図の作成が最初に行われる。散布図の上で、点がばらばらに散らばるようならば x と y は関係がなく、逆に点の分布から何らかの傾向が見て取れるなら、 x と y は関係がありそうということがいえる。

散布図を作成するときは、 X 軸にはより根元的な変数や原因となる変数を取り、 Y 軸には説明される変数や結果となる変数をとるのが一般的である。

また x は性別（男女）、 y は学部名（法学部、経営学部、文学部）などのデータのように、値が属している状態やカテゴリーを表すときは散布図に表すことができない。このように男女や所属の学部名のようなデータのことを質的データと呼び、量的データと区別する。

このように質的データの場合には散布図は使えないので、 x と y がとりうる状態によって2次元の表にし、各状態ごとにその度数を数えて集計した分割表（contingency table、クロス表ともいう）を作成することになる。

8.3. 散布図の作成方法

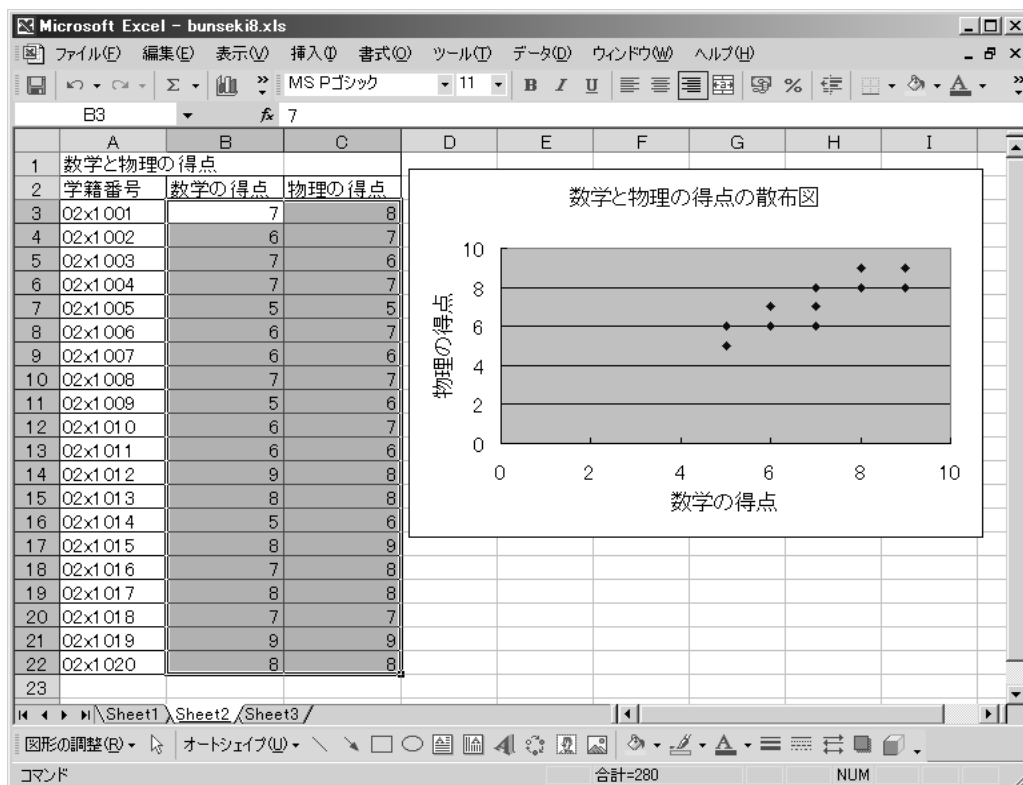


図8.1 数学と物理の得点による散布図

ここでは表 8.1 に示した数学と物理の得点から、散布図を作成してみよう。作成するファイル名は bunseki8.xls として保存する。

- (1) 図 8.1 を参考に学籍番号、数学の得点、物理の得点などをワークシートに入力する。
- (2) カーソルをドラッグして散布図を作成するデータの範囲を選択する。ここでは B3:C22 までを選択する。
- (3) グラフウィザードを起動し、「グラフの種類」から「散布図」を選ぶ。
- (4) 散布図の形式は値の組を比較するものを使う。
- (5) 凡例を表示させないように、「凡例を表示する」のチェックを外す。
- (6) グラフのタイトルに「数学と物理の得点の散布図」と入力し、X 軸や Y 軸の項目などを入力して完了する。
- (7) ワークシートに現れたグラフの大きさなどを適切に変更して完成させる。

上の図 8.1 のように散布図を作成してみると、数学と物理の得点はともに右上がりに分布しており、強い相関関係があることが見て取れる。

8. 4. 相関係数と共分散

統計学の分野では、2 つのデータの関係を相関関係と呼ぶ。2 つの量的データの線形的関係に着目し、直線的な関係があると認められれば、それらの間には相関関係があるという。

2 つの変数の散布図を作成したとき、変数間の関係には次のような場合がある。

- (1) 変数の一方が増加すると、他方の変数も増加する（正の相関関係）。
- (2) 変数の一方が増加すると、他方の変数は減少する（負の相関関係）。
- (3) 上のどちらでもない場合（無相関または無関係）。

また散布図にはデータの分布の状況から、はっきりした直線的な規則性が強く認められる場合と、規則性がかなり弱く認められる場合がある。このようなときの直線的な規則性の度合いは強い、弱いと表現される。つまり強い正の相関関係、弱い正の相関関係、強い負の相関関係、弱い負の相関関係があるといわれる。

相関係数 (correlation coefficient) は、上でまとめたような 2 つの変数間の直線的な関係の度合いを示す。同じように 2 変数間の関係を表し、1 変数の分散に対応するものとして共分散 (covariance) があり、相関係数を求めるために使われる。

x と y の共分散 S_{xy} は x と y の偏差の積 (偏差積) の和を n で割ったものである。

n 個のデータ (x_1, y_1) , (x_2, y_2) , \dots , (x_n, y_n) の共分散 S_{xy} は次の式で定義される。

$$S_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n} \quad (\text{数式 8-1})$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Excel では共分散関数 (COVAR) を使って、x のデータ範囲と y のデータ範囲を指定すれば共分散を計算することができる。共分散の計算は x と y を入れ替えても同じであるため、関数に与える x のデータ範囲と y のデータ範囲を逆にしても同じ結果が得られる。

しかし共分散は値が 1 より大きくなるなど、2 つの変数間の関係を表すものとしてはあまりわかりやすいものとは言えないため、取りうる値の範囲を -1 から +1 の間に基準化したものが相関係数である。

相関係数 r_{xy} は、共分散 S_{xy} を x の標準偏差 S_x と y の標準偏差 S_y で割ったものであり、次の式で定義される。

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{\sum \{(x_i - \bar{x})(y_i - \bar{y})\}}{\{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}\}} = \sum \{(x_i - \bar{x})(y_i - \bar{y})\} / \{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}\}$$

(数式 8-2)

8. 5. 相関係数の特徴

相関係数 r_{xy} については次のような特徴が知られている。

- (1) 常に $-1 \leq r_{xy} \leq 1$ である。
- (2) $|r_{xy}|$ が 1 に近いほど、相関関係が強い。
- (3) $r_{xy} > 0$ のときは、正の相関関係がある。
- (4) $r_{xy} < 0$ のときは、負の相関関係がある。
- (5) $r_{xy} = 1$ のときは、すべてのデータが一直線上にあり、直線の傾きが正の場合 (正の完全相関という)。
- (6) $r_{xy} = -1$ のときは、すべてのデータが一直線上にあり、直線の傾きが負の場合 (負の完全相関)。

8. 6. 相関係数の求め方

Excel を使って相関係数を求めるときは次のように行う。

- (1) CORREL 関数による求め方

相関係数を表示させたいセルをアクティブにしておき、CORREL 関数を使い、x と y にあたるデータの範囲をそれぞれ指定する。

=CORREL(B3:B22, C3:C22)

上の CORREL 関数の指定は、B3:B22, C3:C22 にあるデータの相関係数を求める (図 8.2)。

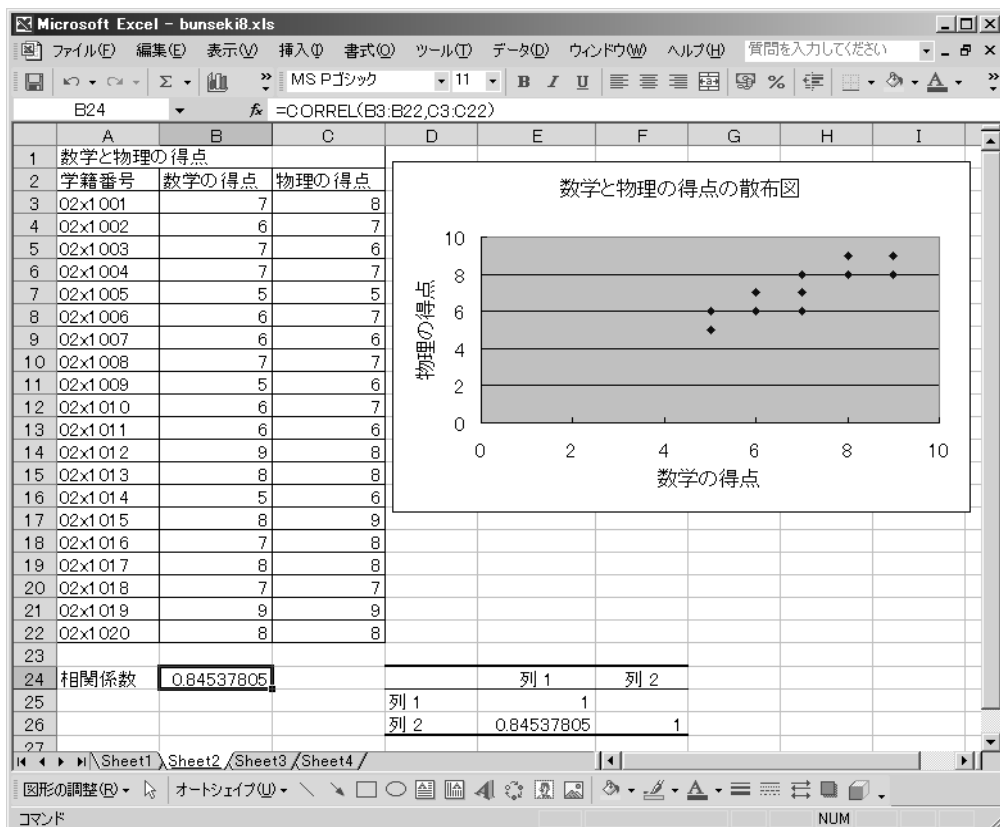


図 8.2 相関係数の求め方

(2) 分析ツールの相関による求め方

分析ツールの相関を使っても相関係数を求められる。「ツール」メニューから「分析ツール」を起動し、「相関」を選び、x と y にあたるデータを指定する (図 8.3)。

入力範囲： \$B\$3:\$C\$22

データ方向： 列

出力先： \$D\$24 (結果の表示を始めたセルを指定)

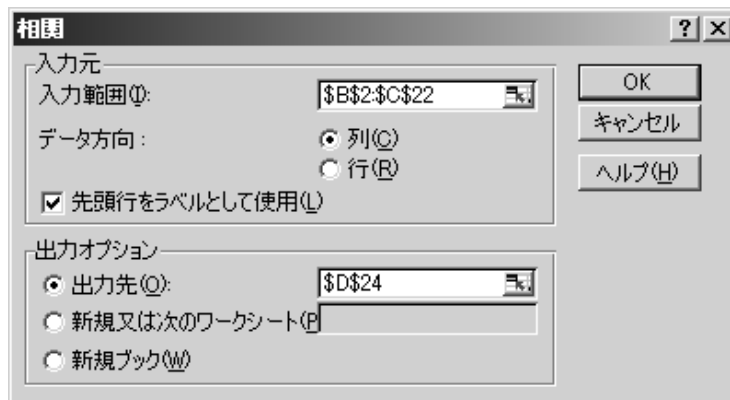


図 8.3 相関の設定画面

なお「先頭行をラベルとして使用」をチェックすると、列番号の代わりにセルに入力してある「数学の得点」などのデータが表示される。

図 8.2 で求めたあるクラスの数学と物理の得点の相関係数は 0.845 となり、かなり強い相関がみとめられる。

8. 7. 共分散の求め方

共分散は以下の関数を使い、 x と y にあたるデータの範囲を指定する。以下の COVAR 関数の指定は、B3:B22, C3:C22 までの共分散を求める。

=COVAR (B3:B22, C3:C22)

また共分散も相関係数と同じように、分析ツールから「共分散」を使っても求められ、操作方法も同様である。

8. 8. いろいろな相関関係

以下は平均気温および社会経済的なデータを散布図にしたものである。散布図を見れば、2 つの変数間の関係が把握しやすい。片方が増加したときに、他方は増加するかまたは減少するかあるいはどちらともいえない場合があることが見えてくる。さらにこれらの関係が強くあらわれているか、または弱くあらわれているかどうかも分かりやすくなる。相関係数の数値を見るより、散布図からはデータの分布状況などの質的側面が観察できる。

(1) 名古屋と地球の地表における平均気温 (正の相関関係) (図 8.4)

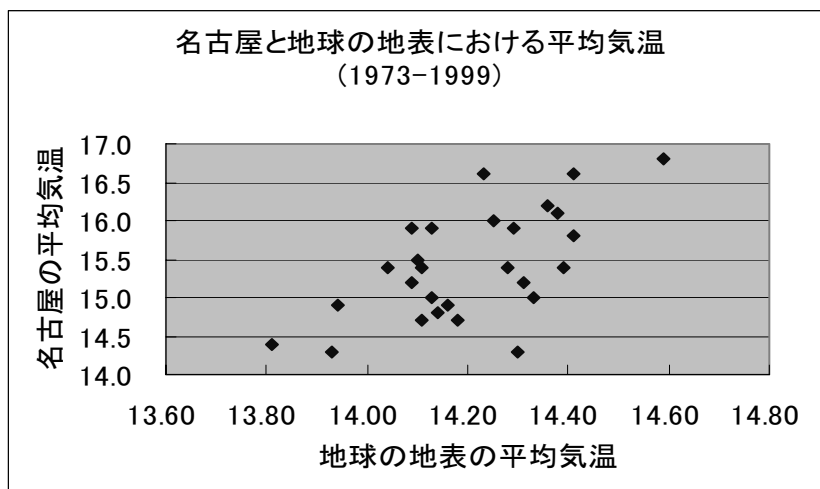


図8.4 名古屋と地球の地表における平均気温 (正の相関関係)

* データの出典

名古屋の平均気温： 名古屋市統計年鑑(統計名古屋Web版, 平成13年版)

<http://www.city.nagoya.jp/stat/nenkan/nenkan.html>

地球の地表の平均気温： 地球環境データブック, ワールドウォッチ研究所, 2001-02, p. 61.

(2) 人口と商店数 (強い正の相関関係) (図8.5)

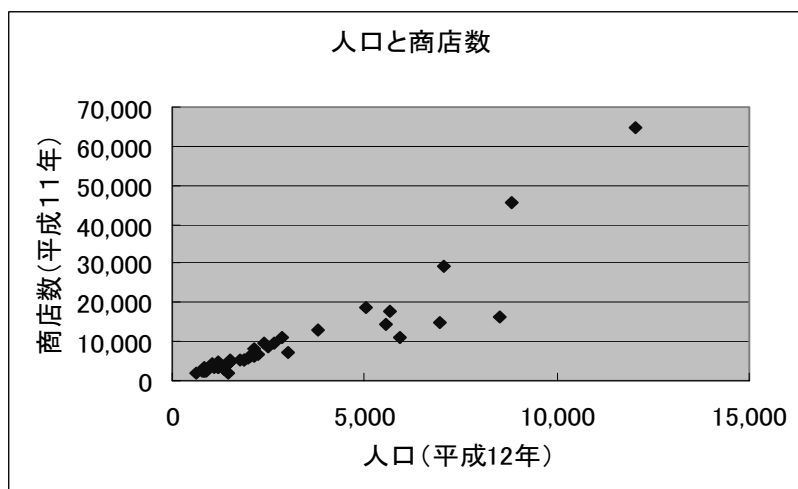


図8.5 人口と商店数 (強い正の相関関係)

* データの出典 (参考文献(1)より作成)

人口（平成 12 年）： 総務省統計局「日本の統計」，第 2 章人口・世帯，都道府県別の人口と人口増加率.

<http://www.stat.go.jp/data/nihon/02.htm>

商店数（平成 11 年）： 総務省統計局「日本統計年鑑」，第 11 章商業・サービス業，都道府県別商店数，従業者数，年間販売額及び売場面積.

<http://www.stat.go.jp/data/nenkan/11.htm>

(3) 1 世帯あたり年間の米とパンの支出金額（無相関）（図 8.6）

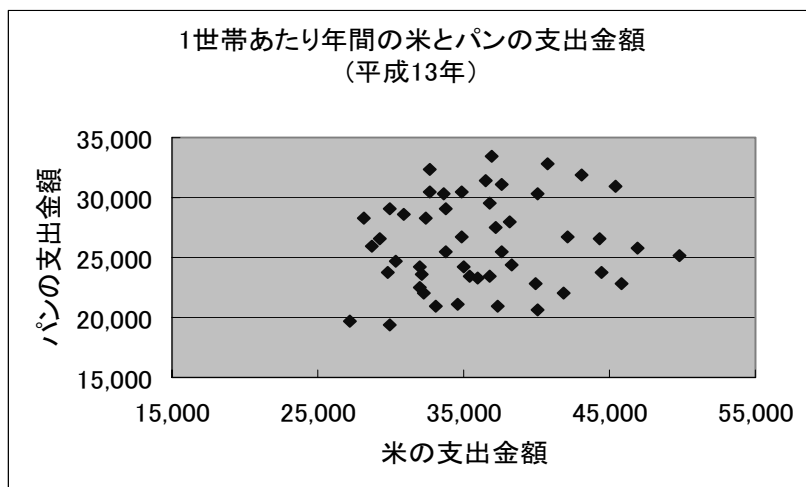


図 8.6 1 世帯あたり年間の米とパンの支出金額（無相関）

* データの出典（参考文献(1)より作成）

都道府県庁所在市別 1 世帯当たり年間の米とパン支出金額（平成 13 年）： 総務省統計局，家計調査平成 13 年年報.

<http://www.stat.go.jp/data/kakei/2001np/zuhyou/2nh1801.xls>

X と Y 軸も目盛りの最小値を 15000 で作成.

(4) 人口密度と総農家数（負の相関関係）（図8.7）

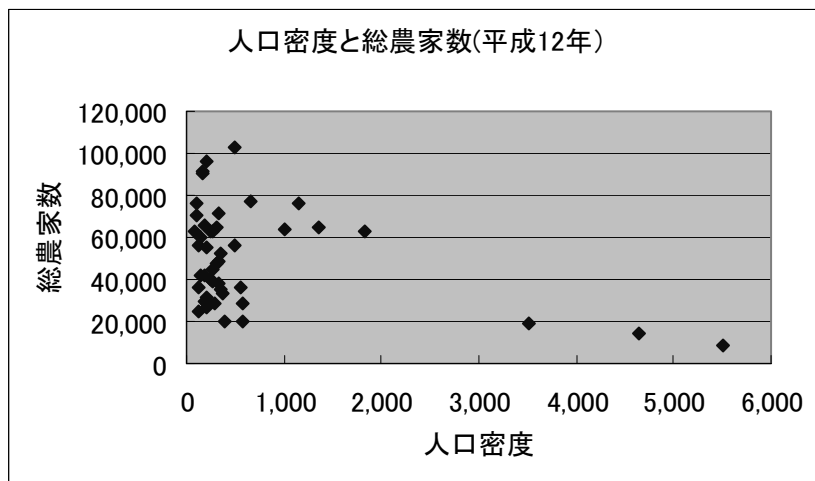


図8.7 人口密度と総農家数（負の相関関係）

* データの出典（参考文献(1)より作成）

人口密度（平成12年）： 総務省統計局，日本の統計，第2章，都道府県別の人口と人口増加率

<http://www.stat.go.jp/data/nihon/02.htm>

総農家数（平成12年）： 総務省統計局，日本の統計，第6章，都道府県別の農家数と農家人口（販売農家）

<http://www.stat.go.jp/data/nihon/06.htm>

》》 演習 8 《《

次の演習を行ってみよ。

1. 地球地表の平均気温と名古屋の平均気温の相関関係を考察せよ。

地球地表の平均気温と名古屋の平均気温のそれぞれの経年変化を使って散布図に表し，次の考察を行ってみよ。

なお以下の作業に必要な地球地表の平均気温のデータは「4. 4. 地表の平均気温」に記載している。また名古屋の平均気温の所在は，「4. 5. 地表の平均気温と名古屋の平均気温の比較」にその出典を記載しているので参照されたい。

- (1) 2つのデータの間にはどのような相関関係があるかどうかを考察せよ。
- (2) X軸とY軸にどちらを取るべきか，X軸とY軸を入れ替えた散布図を作成して考察せよ。
- (3) 相関係数を2つの方法で求めよ。
- (4) 共分散を求めよ。

9. 回帰直線

回帰直線は散布図を作成したときに、その中にどのような傾向があるかを調べるために作成される直線のグラフである。2次元のデータ分析では x と y の組からなるデータ (x_1, y_1) , (x_2, y_2) , \dots , (x_n, y_n) があるとき、 x_i と y_i の間に強い相関関係があれば、これらの点はすべてある直線の近辺に分布することが知られている。この直線を x に対する y の回帰直線 (regression line) と呼んでいる。

9. 1. 回帰直線の作成

回帰直線を描くためには、先に散布図を作成しておき、次のように操作する。

- (1) 散布図のあるグラフエリアをクリックし、グラフの上で右ボタンをクリックする (図9.1)。
- (2) 「近似曲線の追加」を選び、「線形近似」の順に選択する (図9.1)。

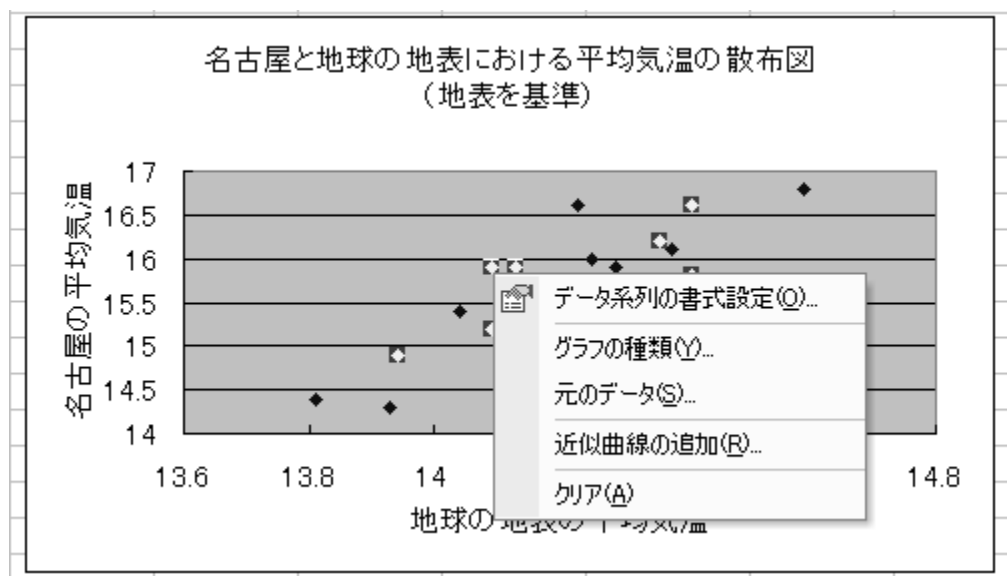


図9.1 回帰直線 (近似曲線) の作成

9. 2. 直線の方程式

直線の方程式を作成したいときは、次のように行う。

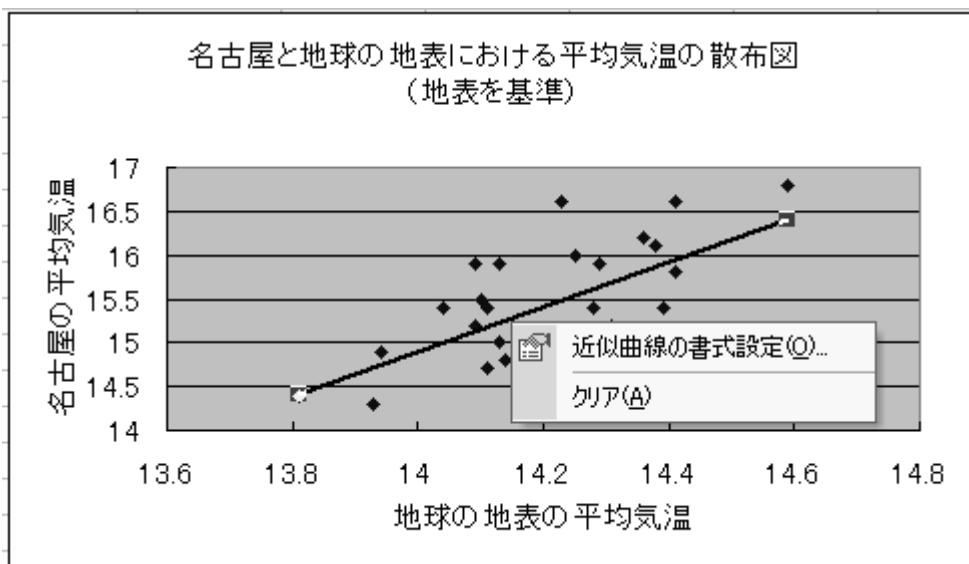


図 9.2 直線の方程式を近似曲線の書式設定で指定

- (1) 近似曲線の上で右クリックし、「近似曲線の書式設定」を呼び出す (図 9.3) .
- (2) 次に「オプション」を選択し、「グラフに数式を表示する」をチェックする (図 9.3) .

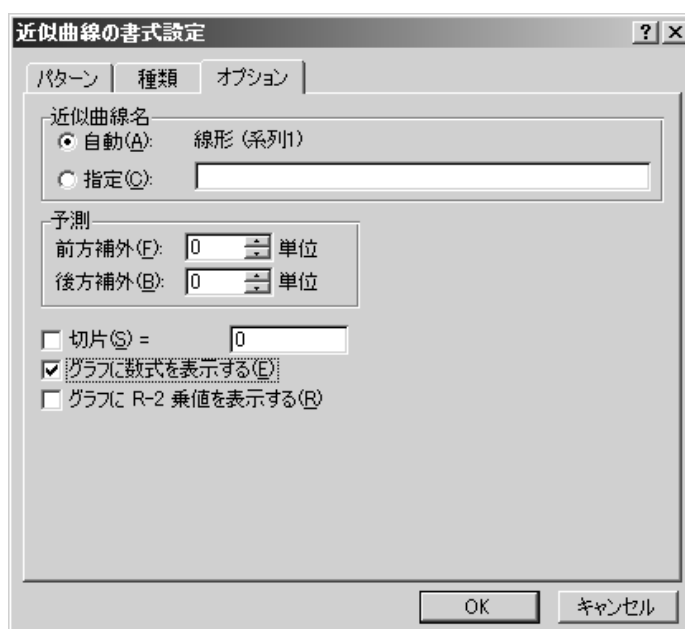


図 9.3 近似曲線の書式設定

9. 3. 回帰直線の考え方

回帰直線は最小 2 乗法にもとづいて作成される。最小 2 乗法は次のように求められる。

(1) 回帰直線は $y = ax + b$ の形式で表されるので、2つの定数 a, b は各データ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ から推定する。

(2) このときに直線との誤差の2乗の総和が最小になるように、 a, b の値を求める。つまり誤差の2乗は $((ax_i + b) - y_i)^2$ で求められるので、総和である $\sum_{i=1}^n ((ax_i + b) - y_i)^2$ が最小になるように求める。 (x_i, y_i) は実際の2次元データである $(x_i, ax_i + b)$ によってあらわされ、回帰直線上の理論値（回帰直線上に分布すると仮定した場合の値）である（図9.4）。

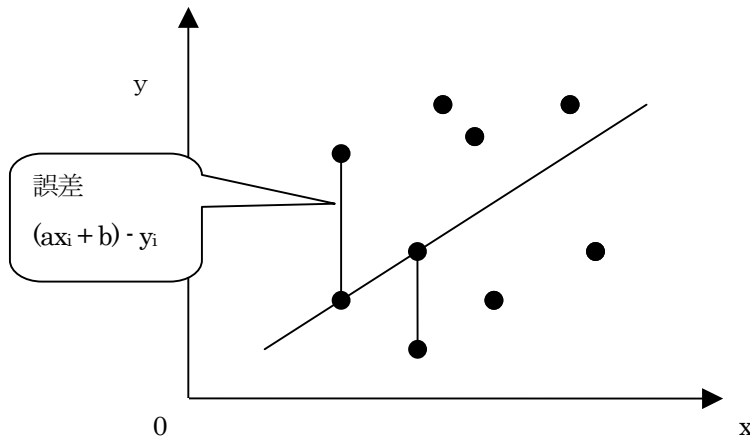


図9.4 回帰直線と誤差の考え方

(3) x_i と y_i の平均をそれぞれ \bar{x}, \bar{y} とし、標準偏差を S_x, S_y 、共分散を S_{xy} とすれば、 a, b の値は次のように求められる。

$$a = \frac{S_{xy}}{S_x^2} \qquad b = \bar{y} - \frac{S_{xy}}{S_x^2} \bar{x}$$

従って回帰直線は次式のようなになる。

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$$

$\frac{S_{xy}}{S_x^2}$ は直線の傾きであり、回帰係数と呼ぶ。

相関係数 $r_{xy} = \frac{S_{xy}}{S_x S_y}$ をもちいれば、 $\frac{S_{xy}}{S_x^2} = r_{xy} \frac{S_y}{S_x}$ となり、この直線は x と y の平均である (\bar{x}, \bar{y})

を通る。

9. 4. クロス表

データが、 x は性別 (男女) , y は学部名 (法学部, 経営学部, 文学部) などのように、値が属している状態やカテゴリーを表すときは散布図に表すことができない。このように男女や出身地のようなデータのことを質的データと呼び、量的データと区別する。

このような質的データの場合には散布図は使えないので、 x と y がとりうる状態によって2次元の表にし、それぞれの状態ごとにその度数を数えて集計したクロス表 (cross table) を作成する。クロス表は分割表 (contingency table) とも呼ばれる。

アンケート調査などで多量のデータを集計する場合、クロス表がしばしば使われる。それぞれの項目ごとの集計だけでなく、2つの項目にまたがって、項目間の関係を見るためにクロス表が作成される。

9. 5. 集計データの準備

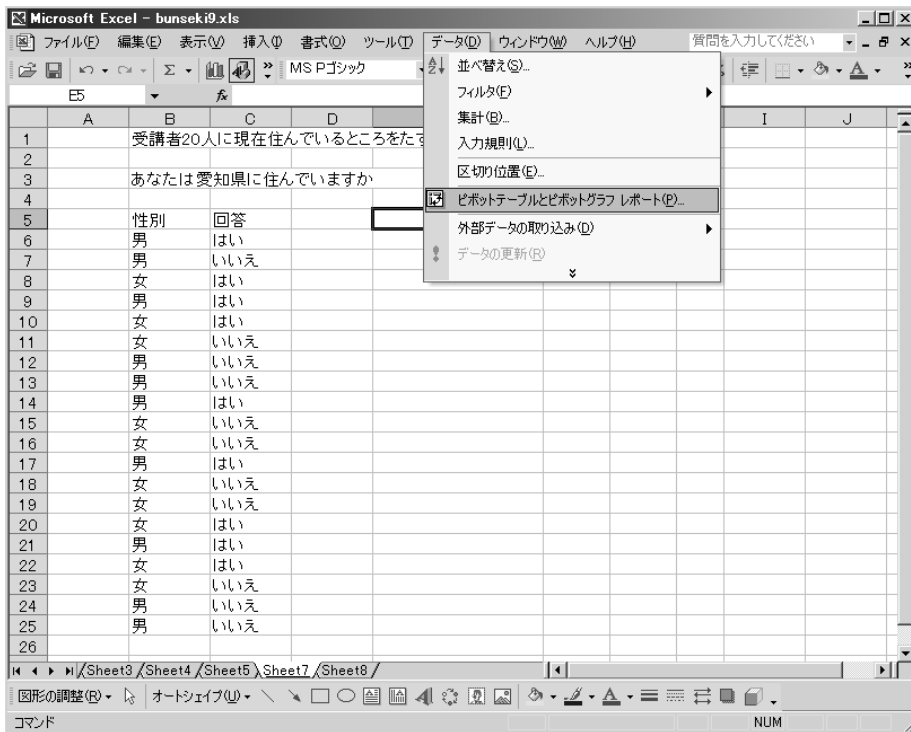


図 9.5 リスト形式のデータ

ピボットテーブルを使うためには、集計するデータをあらかじめ決められた形式でワークシートに用意する必要がある。まずデータを整理したい項目名を列の上部に入力する。次にその項目名に続けてデータの1件1件が、xとyの組みになるように、それぞれのデータを項目名の列に規則正しく並べるようにする。このようなデータをリスト形式とすることがある。

このような形式にデータを入力しておく、並べ替えや検索などの処理にも使えて便利である。

9. 5. ピボットテーブルの作成

ピボットテーブルを使ってクロス表を作成する手順は次のように行う。

- (1) まずワークシート上に集計するデータを作成しておく (図9.5)。
- (2) 入力したセルのいずれでもよいのでアクティブにしておき、「データ」メニューから「ピボットテーブルとピボットグラフレポート」を選ぶ (図9.5)。
- (3) ピボットテーブルウィザードの1/3では、「Excel のリストデータベース」と「ピボットテーブル」をチェックする (図9.6)。

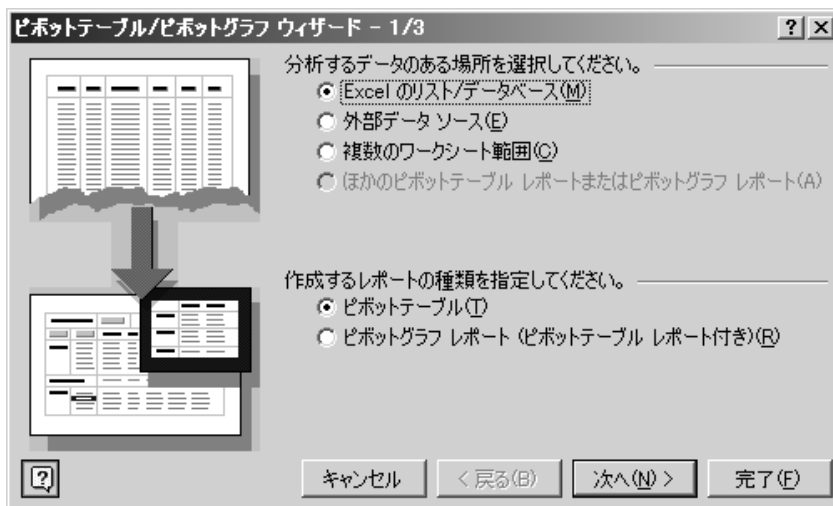


図9.6 ピボットテーブルウィザード 1/3

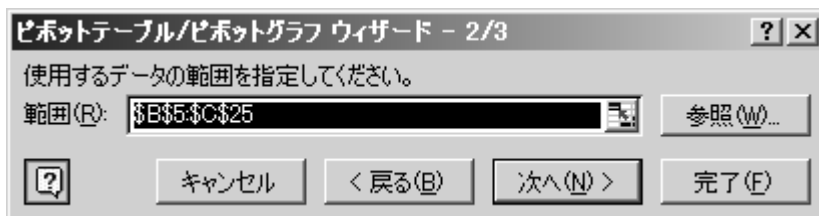


図9.7 ピボットテーブルウィザード 2/3

- (4) ピボットテーブルウィザードの2/3では、使用するデータの範囲を指定する(図9.7)。
- (5) ピボットテーブルウィザードの3/3では、既存のワークシートにピボットテーブルの結果を表示させる指定をする(図9.7)。

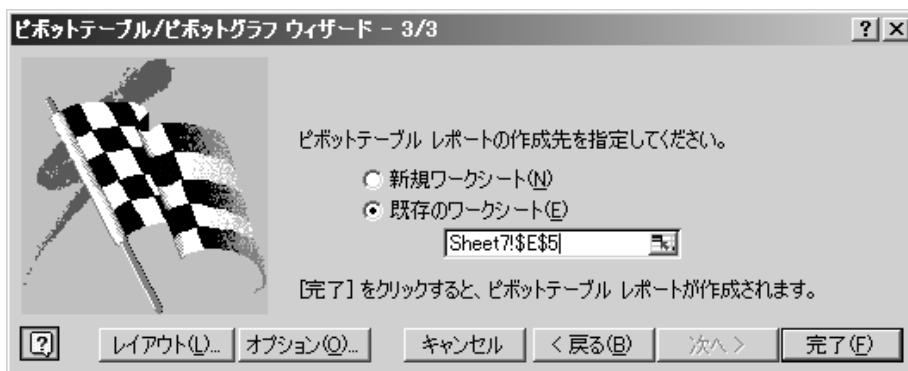


図 9.8 ピボットテーブルウィザード 3/3

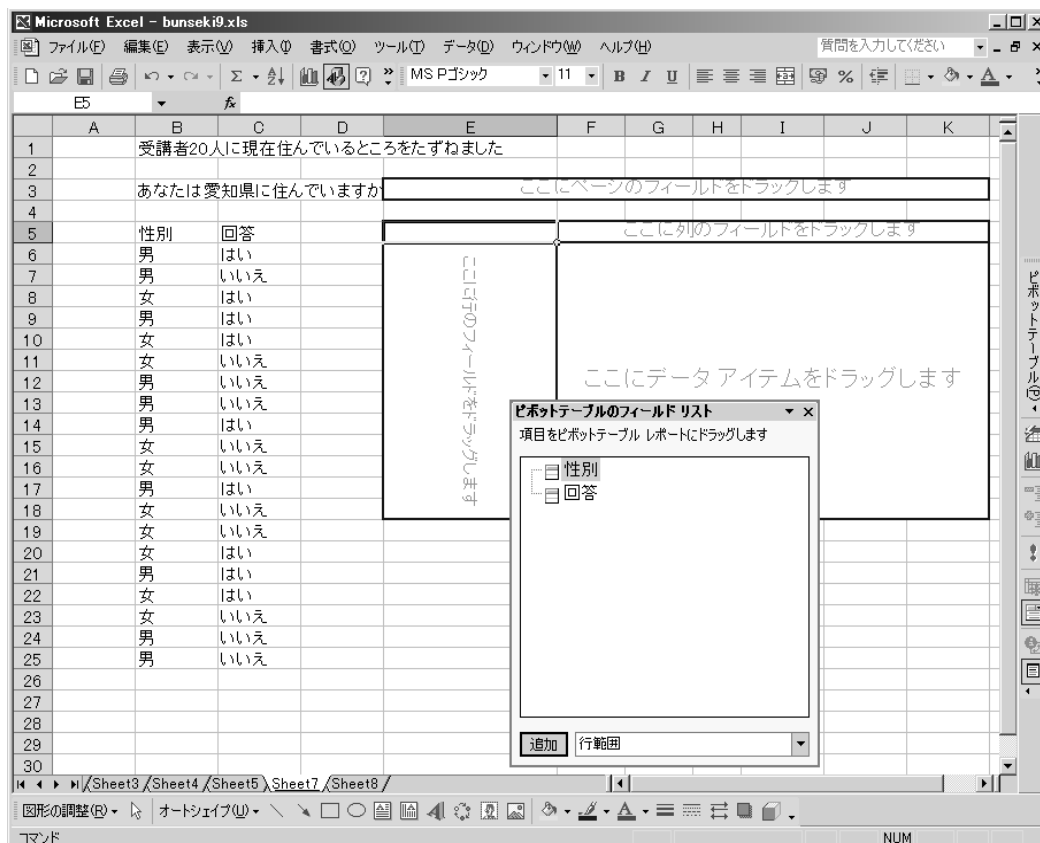


図 9.9 ピボットテーブルのフィールドリスト

(6) ピボットテーブルのフィールドリストが表示されるので、性別をドラッグして「ここに行のフィールドをドラッグします」という表示の上にドロップする。同じように回答をドラッグして、「ここに列のフィールドをドラッグします」という表示の上にドロップする。再度回答をドラッグして「ここにデータアイテムをドラッグします」という表示の上にドロップするとピボットテーブルが作成される (図9.10)。

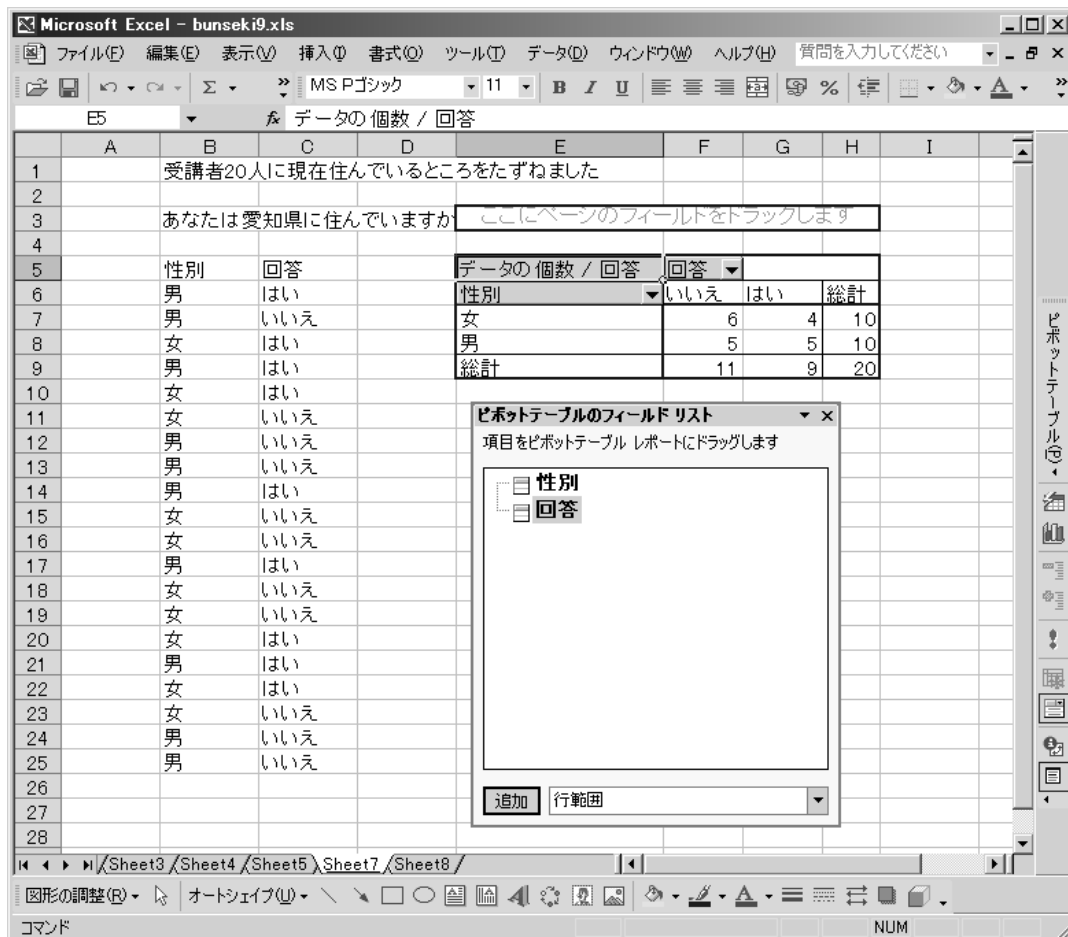


図9.10 ピボットテーブルの完成

》》》 演習 9 《《《

次の演習を行ってみよ。ファイル名は bunseki9.xls として作成する。

1. 「8. 8. いろいろな相関関係」の散布図を作成し、それぞれに回帰直線を求めよ。
2. 次のデータを入力してピボットテーブルを作成してみよ。なおレジюмеでは都合により3段に分けてデータを表示しているので、ワークシートに入力するときは1列に入力する。

履修者全員に所属学部を聞きました	
あなたの学部はどこですか	
性別	回答
男	法
男	法
女	法
男	法
男	法
女	経営
女	経営
男	経営
男	経営
女	経営

女	経営
女	経営
女	経営
女	経営
女	経営
男	経営
女	経営
男	経営
男	経営
男	経営
男	経営
男	経営
男	経営
男	経営
男	経営
男	経営
女	経営
男	経営
男	経営
男	経営

女	経営
男	経営
女	経営
女	経営
女	経営
男	経営
男	経営
男	経営
女	経営
男	経営
男	経営
男	経営
女	現中
女	現中
男	現中
女	現中

10. 乱数とさいころのシミュレーション

乱数を使って、さいころの目の出方をシミュレーション (simulation) してみたい。

シミュレーションとは、一般的には模擬実験のことであるが、コンピュータを使った模擬実験の場合には、コンピュータ・シミュレーション (computer simulation) ともいう。

さいころの目の出方は、1 から 6 までの 6 つの数字がランダムに出現する。それらの数字を出た順番に並べると、規則性のない数の列ができる。このようにしてできる数の並びを乱数という。

乱数は意外に幅広く応用されている。シミュレーションをはじめ、サンプリング、意思決定、ゲームなどでも使われている。

コンピュータでさいころのシミュレーションを行うときは乱数を使用する。Excel には乱数を発生させる関数が備わっているのでここではそれを使って行う。

(1) RAND() 関数

0 以上で 1 より小さい乱数をひとつ発生させる関数である。ワークシートが再計算されるたびに、新しい乱数が返される。

例えば A1 のセルに `= RAND()` と式を入力すれば、乱数がひとつ得られる。乱数を複数個生成する必要があるときは、必要な数だけ数式をドラッグして複写すれば得られる。

次に a と b の範囲で乱数を生成したいときは、次のように数式を作る。

```
=RAND()*(b-a)+a
```

1 から 6 までの整数の乱数を生成させたいときは、上の式のままでは小数点以下も表示されるので、ROUND() 関数を使って小数点以下を四捨五入し、式は次のように作成する。

```
=ROUND(RAND()*(6-1)+1, 0)
```

あるいは INT() 関数を使って小数点以下を切り捨てることができるが、そのときは次のように式を作る。

```
=INT(RAND()*6)+1
```

INT() 関数は、数値の小数点以下を切り捨てる関数であり、例えば INT(6.45) は 6 になる。

(2) RANDBETWEEN()

RANDBETWEEN() 関数は、指定された値の範囲で整数の乱数を生成する。求めたい範囲の最大値と最小値を、

整数で指定する。この関数を使うためには、分析ツールアドインをインストールする必要がある。

例えば1から100までの整数の乱数を求めたいときは、次のように式を作る。

=RANDBETWEEN(1,100)

(3) 乱数の固定

乱数は生成したままでは変動するので、ヒストグラムなどを作成する場合に不便である。固定するためには編集メニューで「コピー」を行い、「形式を選択して貼り付け」を行う。このときに「貼り付け」の中の「値」をチェックすると数値だけが複写される(図10.1)。

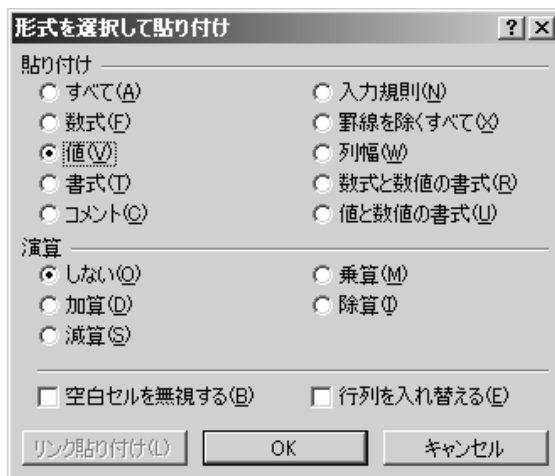


図 10.1 形式を選択して貼り付けの画面

または「ツール」メニューから「オプション」をえらび、次に「計算方法」から「手動」を開きチェックすると、乱数が固定される。ただし同じシートでオートフィルなどは使えなくなるので注意が必要である。

なおひとつだけ乱数を生成して固定するときは、乱数を生成したあと、数式バーをクリックして F9 キーを押せば、値がセルに複写されて固定する。

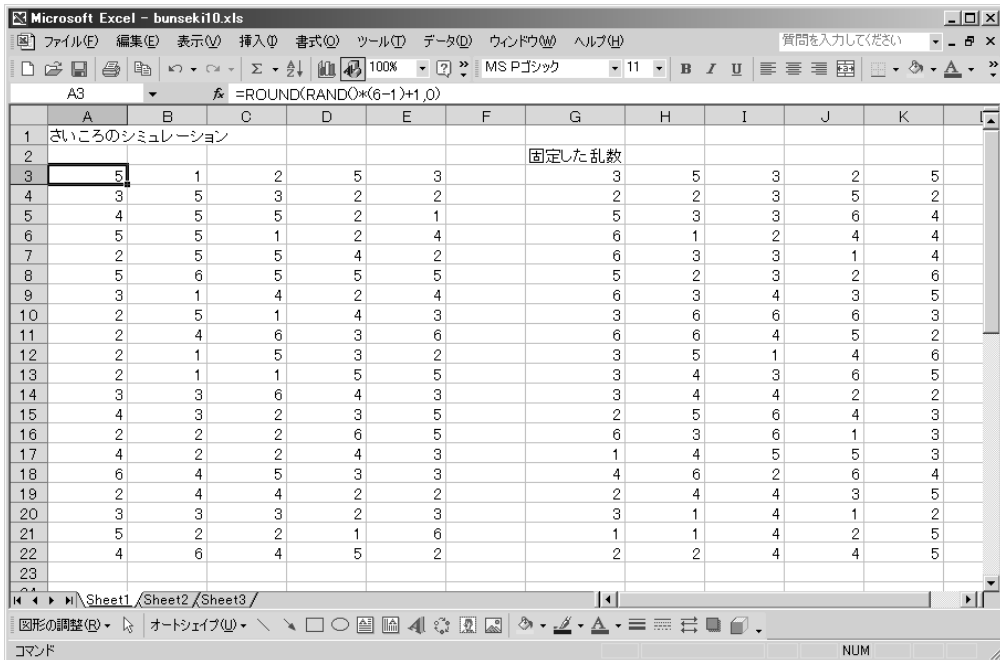


図 10.2 乱数とさいころの目の出方のシミュレーション

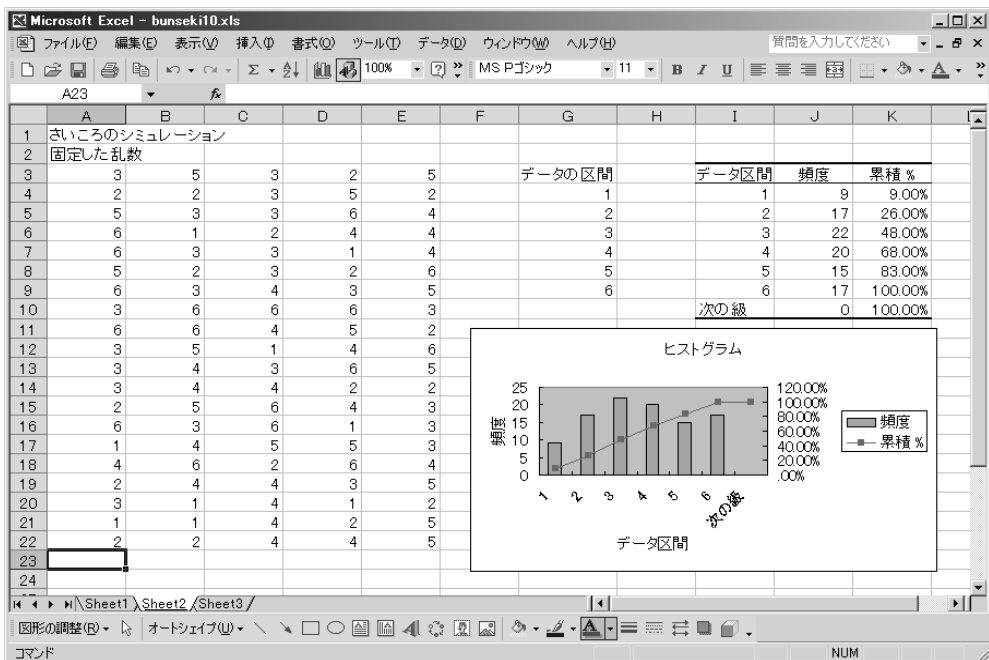


図 10.3 さいころの目の度数分布とヒストグラム

10. 2. 分析ツールによるヒストグラム

さいころの目の数がどのように分布するか、度数分布とヒストグラムを作成してみよう。「分析ツール」の「ヒストグラム」による度数分布の求め方は次のように行う。

- (1) 1から6までの乱数を100個生成する(図10.3)。
- (2) 乱数をコピーしたあと、値を指定して貼り付けを行い、乱数を固定する。
- (3) 「ヒストグラム」で使用するため、「データの区間」を1から6まで作成する。
- (4) 「ツール」メニューから「分析ツール」を選び、次に「ヒストグラム」を実行する。
- (5) 出力先は同じワークシートとし、「累積度数分布の表示」と「グラフ作成」をチェックする。

10.3. COUNTIF()関数

度数を求めるときは、COUNTIF()関数を使うこともできる。COUNTIF()関数は、指定された範囲に含まれるセルから、検索条件に一致するセルの個数を返す。求め方は次のように行えばよい。

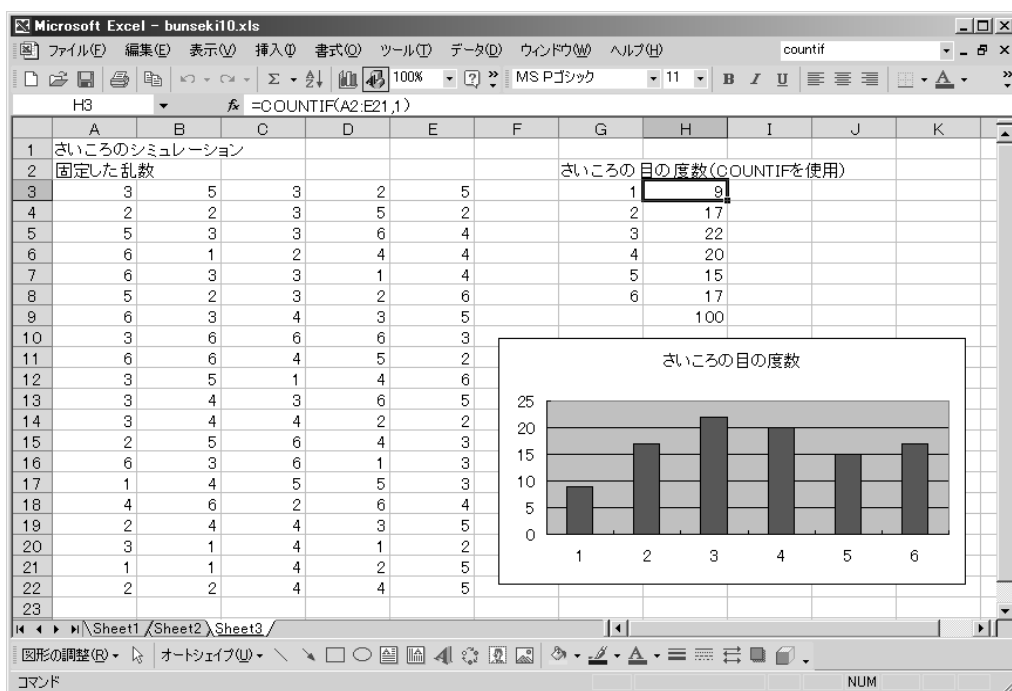


図 10.4 COUNTIF()関数による度数分布のグラフ

- (1) 1から6までのさいころの目の数をセルに入力しておく(図10.4)。
- (2) 1のセルの隣に、次のように式を入力する。A3:E22は検索範囲を示し、1は検索条件である。つまり1に該当するセルをカウントして、その度数を返す。

=COUNTIF(A3:E22, 1)

なお検索条件には、数値のほかセル番号や文字列を指定することができる。その場合にはオートフィルも使用できる。上の例では次のように H3 セルに入力し、オートフィルで式の複写をすれば度数のカウントができる。その場合に範囲の A3:E22 は絶対参照の指定をしておく。

=COUNTIF(\$A\$3:\$E\$22, G3)

(3) さいころの目が 2 の度数を数えるときは、検索条件を 2 にして行う。

(4) グラフの作成はグラフウィザードを起動して行う。

図 10.4 で求めたヒストグラムでは、1 から 6 までのさいころの目の度数分布は、それぞれ 9 から 17 までとなっている。しかしこの結果からさいころの目の出方が正しいかどうかの判断はすべきではない。正しいかどうかは確率分布や検定というような考え方が必要になってくる。

10. 4. アルファベットの文字乱数

Excel の関数に文字を表示する CHAR() という関数がある。この関数に文字コードを表す数値データを与えると、ワークシートのセルに指定された文字を表示する。

CHAR() 関数は、指定した数値を ASCII または JIS コードの番号と見なし、それに対応する文字を返す。例えば適当なセルをアクティブにして、CHAR() 関数の数値に 97 入力すると、文字の a が表示される (図 10.5)。数値に 98 を入力すれば b, 99 を入力すると c が表示される。

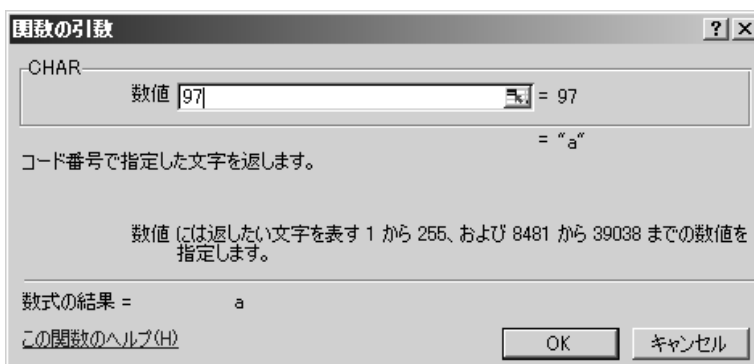


図 10.5 CHAR() に数値 97 を与え a を表示した例

従って与える数値が文字コードの値の範囲でランダムに生成できれば、文字乱数 (文字による乱数) を生成することができる。アルファベットの場合は、97(a) から 26 文字分を加えた 122(z) ままで数値の範囲となる。

文字乱数を生成するときは、乱数を生成する RAND() 関数と小数点以下を切り捨てる INT() 関数、および CHAR() 関数を組み合わせて使用する。

ここでは 100 個の文字乱数を生成し、その度数分布や相対度数を求めてみよう。

10. 5. 文字乱数の生成

文字乱数を生成するときは、1 から 26 までの乱数を生成させ、さらに文字に変換する。

(1) 1 から 26 までの乱数は、INT() 関数と RNAD() 関数を使い、次の式で生成される。26 はアルファベットの個数である。

```
=INT(RAND()*26)+1
```

(2) アルファベットに変換するためには CHAR() 関数を使い、上で作成した式を変更して次の式を作成する。97 は文字 a を示す数値である。

```
=CHAR(INT(RAND()*26)+97)
```

(3) 上で作成した式を A3 のセルに入力し、オートフィルを使って必要な個数の文字乱数を生成する。

10. 6. 文字乱数の度数分布と相対度数

文字乱数の度数分布と相対度数を求めるときは、次のように行う。

- (1) 乱数が変動しないように、あいているセルに値を指定して複写したあと以下の処理を行う (図 10.6)。
- (2) アルファベットの生成に必要なため、1 から 26 までの数値をオートフィルで生成する。
- (3) 度数を求めるアルファベットを次の式で生成する。96 は文字 a のひとつ手前の数値で、M3 は 1 から 26 までの数値が入った列である。

```
=CHAR(96+M3)
```

(4) アルファベットの度数は次の式で求める。\$G\$3:\$K\$22 は複写した文字乱数の範囲を示し、0 (オー)3 は度数を求める a のセルである。

```
=COUNTIF($G$3:$K$22,03)
```

(5) 上の式をオートフィルで複写して、b 以外の度数を求める。

(6) 相対度数はそれぞれの度数を総度数で割って求める。P3 は文字 a のセルを示し、\$P\$29 は総度数のセ

ルであり、絶対指定をしておく。

=P3/\$P\$29

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	アルファベットの文字乱数																
2															文字	度数	相対度数
3	q	z	p	a	v		z	f	u	e	k		1	a	a	4	0.04
4	w	d	n	p	d		l	s	y	m	c		2	b	b	1	0.01
5	g	t	n	w	j		v	c	h	t	q		3	c	c	11	0.11
6	r	p	x	n	q		r	v	t	p	c		4	d	d	3	0.03
7	x	k	w	b	u		s	h	a	g	o		5	e	e	6	0.06
8	d	n	y	r	q		z	o	h	w	h		6	f	f	2	0.02
9	s	p	z	a	h		e	m	d	t	x		7	g	g	2	0.02
10	s	q	o	w	j		g	y	n	o	v		8	h	h	5	0.05
11	q	c	t	j	e		p	x	c	r	p		9	i	i	0	0
12	d	i	g	a	q		j	l	m	o	k		10	j	j	2	0.02
13	e	y	r	h	x		x	j	l	h	q		11	k	k	4	0.04
14	k	y	d	z	o		l	e	p	r	p		12	l	l	9	0.09
15	o	d	s	f	l		t	d	a	c	x		13	m	m	5	0.05
16	t	e	x	o	v		w	q	w	c	k		14	n	n	2	0.02
17	t	r	k	e	p		s	l	f	e	c		15	o	o	4	0.04
18	v	u	o	n	s		d	c	l	l	l		16	p	p	7	0.07
19	e	b	n	k	d		a	a	b	y	y		17	q	q	4	0.04
20	j	y	f	i	l		x	p	p	t	n		18	r	r	3	0.03
21	m	y	e	r	p		c	c	m	q	e		19	s	s	3	0.03
22	t	p	t	t	a		l	e	m	k	c		20	t	t	5	0.05
23													21	u	u	1	0.01
24													22	v	v	3	0.03
25													23	w	w	3	0.03
26													24	x	x	5	0.05
27													25	y	y	4	0.04
28													26	z	z	2	0.02
29													合計			100	1.00
30																	

図 10.6 文字乱数の生成と度数分布・相対度数

演習 10

次の演習を行ってみよ。ファイル名を bunseki10.xls として作成せよ。

- さいころを 200 回投げたことにして、それぞれの目の相対度数とヒストグラムを作成せよ。なお度数の求め方は、分析ツールのヒストグラムによる方法と COUNTIF() 関数による方法の 2 通りで行ってみること。
- アルファベットの文字乱数を 300 個生成し、それぞれの目の相対度数とヒストグラムを作成せよ。

11. 社会データの情報源

インターネットには社会調査のデータが数多く公開されている。ここでは総務省統計局統計センターのWeb ページに収集されているものから紹介する。

11. 1. 総務省統計局統計センター(<http://www.stat.go.jp/>)

国勢調査をはじめ、国内外の各種の統計データが公開されている。公開されている主な統計データには次のようなものがある。

国勢調査	労働力調査	日本統計月報
事業所・企業統計調査	消費者物価指数 (CPI)	日本の統計
人口推計	日本統計年鑑	世界の統計

統計データのほかにも、統計に関するさまざまな情報を提供している。例えば日本の統計制度や統計審議会の議事録、統計に関するQ&Aなどのほか、リンク集にも多くの統計データの情報源が紹介されている。

また分野別一覧では、それぞれの統計データがどのようなものか、簡単な概要が記載されているので、実際に統計を選ぶときに役に立つ。

最新情報のページでは、最近実施された調査の結果が公開されており、最新のデータを得たいときに便利である。

11. 2. 統計情報総合案内(<http://www.stat.go.jp/data/guide/index.htm>)

総務省統計局統計センターのページには、分野別と府省別の統計検索ガイドが公開されている。

分野別の統計検索ガイドからは、それぞれの分野ごとに分類された統計データにアクセスすることができる。具体的な統計データの名前が分からなくても、どのような分野かが分かれば探すことができる。具体的には下表のように分類されている。

国土・気象	人口・世帯	労働・賃金	国民経済計算
企業活動	農林水産業	鉱工業	建設業
エネルギー・水	運輸・通信	商業・サービス業	貿易・国際収支・国際協力
金融・保険	財政	物価・地価	家計
住宅・土地	社会保障	保険衛生	教育
科学技術・文化	公務員・選挙	司法・警察	災害・事故
地域データ	国際統計		

府省別の統計検索ガイドからは、それぞれの府省が公開している統計データを探することができる。府省別の検索ガイドはExcel の表で提供されており、それを見ると公表媒体のところから、印刷物のほかにインターネットなどの電子媒体で提供されているかどうか分かる。以下の府省名から閲覧することができる。

府省名					
内閣府	総務省	公正取引委員会	郵政事業庁	法務省	財務省
国税庁	文部科学省	厚生労働省	農林水産省	経済産業省	国土交通省環境省

11. 3. 統計データのリンク集

総務省統計局統計センターのホームページには、さまざまな統計データのリンク集が作成されている。主なリンク先は次のようなものがある。

(1) 主な地方公共団体

国内で統計データを公開している主な地方自体を集めてある。県だけでなく主な市町村へのリンクもある。

(2) その他の統計関係機関

国内で統計データを公開している主な民間統計を集めたもの。業界団体や大学の研究所なども含まれている。なかでもインターネット提供の民間統計集 (<http://www.nafsa.or.jp/MINKAN.html>) を見ると、数多くの民間統計がインターネットに公開されていることがわかる。

(3) 国際機関等

主な国際機関でインターネットから統計データを利用できる機関を集めたもの、国際通貨基金や国連統計部、世界保健機構、世界銀行などへのリンクがある。

(4) 海外の統計機関

海外の統計機関が地域別にまとめられている。なお国によっては Web ページを表示させるために追加のフォントが必要になる場合もある。

12. 参考・引用文献

東京大学教養学部統計学教室編：統計学入門，東京大学出版会，（基礎統計学 I），pp. 307(1991)。

縄田和満：Excel による統計入門，朝倉，pp. 196(2000)。

上田太一郎：Excel でできるデータマイニング入門，同友館，pp. 340(2001)。

上田太一郎：Excel(エクセル)でできるデータマイニング演習，同友館，pp. 213(2001)。

白石修二：例題で学ぶ Excel 統計入門，森北出版，pp. 145(2001)。

佐藤博樹ほか：社会調査の公開データ -2 次分析への招待-，東京大学出版会，pp. 260(2000)。

福田剛志：データマイニング，共立出版，pp. 169(2001)

木下栄蔵：社会現象の統計分析—手法と実例—，朝倉書店，pp. 197(1998)。