

はじめに

本書の目的

本書の目的は、経済学部1年生程度の経済学初学者に、データ分析の基礎を実習形式で学ぶことを通じて、データ分析の楽しさ、面白さを知ってもらうことである。

大学で経済学を学ぶためには、ある程度のデータ分析力が必要である。自らデータ分析をすることはないとしても、データ分析の結果を読む力は必要であるし、経済学の応用科目を学ぶときはもちろんのこと、ニュースや新聞報道を理解する上でも、統計分析の読み方を知っている必要がある。さらに、就職して企業や行政で働くうえでも、また良き市民として国や地方自治体の政策を評価するうえでも、データ分析力は必要である。

しかし、社会・経済問題には関心があるけれど、数学は苦手な統計数値を扱うのに慣れていないという学生が少なくない。数学が苦手な学生にとって、統計学や計量経済学を学ぶことは少し敷居が高いかもしれない。そこで、統計の基礎的な理論や専門的な分析手法を学ぶ前に、データ分析を体験することで苦手意識を克服してもらいたい。分析は一般的なPCと汎用的な表計算ソフトを使えば簡単にできるので、分析の手順や結果の読み方を学ぶだけで良い。本書の内容を活用できるようになったら、より専門的な学習にステップアップしてほしい。

本書の特徴

以上のような目的から、本書は数学が苦手な学生、統計学や経済理論の知識がない学生でも学べる内容となっている。ただし、統計データを扱うので、簡単な数学（ $+$ $-$ \times \div ）は必要である。また、PCを使って実習をするので、マウス操作、日本語入力、ファイルの保存など、ある程度PCが使えることを前提としている。

本書は経済分析を実習形式で学ぶことを目的としているので、項目ごとの練習問題のほか、章末に課題を用意した。これらの問題や課題に取り組むことで、本文で学んだ分析の手順や結果の表現方法、文書での記述方法などを復習できるようになっている。

本書に取り組むためには数学や統計学の基礎的な知識を必要としないとはいえ、数式の意味を理解したいという読者のために、簡単な統計注を用意した。数学が苦手ではない、統計的な基礎理論も学びたいという場合は、ぜひこちらも参考にしてほしい。

また、本書では、ところどころに表計算ソフトの使い方や経済分析に関する豆知識を盛り込んだ。経済分析を進める上で必ず必要というわけではないが、知っておくと作業の効率が高まったり、本書の内容をよりよく理解できるので、余裕があれば参照してほしい。

使用するデータについて

実習形式のテキストでは練習用に作られたダミー・データを使っている場合が多いように思われるが、本書の練習問題や章末課題で使用するデータはすべて政府の Web ページ等でダウンロードできる公的統計を用いている。したがって、読者は本書の課題に取り組むことにより、現実の日本経済を分析することになる。

データファイルは以下のページからダウンロードできる。また、最新の統計に更新したデータも適宜アップしていく予定である。ご活用いただきたい。

<http://murasemi.com/eco-analysis/>

謝辞

本書は日本大学経済学部における「経済分析入門」の講義レジュメがもとになっている。講義を担当してくださった講師の先生方や受講学生のみなさんからは、誤植の指摘や数多くの有益なコメントをいただいた。この場を借りてお礼申し上げたい。

目次

はじめに	i
・ 本書の目的	i
・ 本書の特徴	i
・ 使用するデータについて	ii
・ 謝辞	ii
1 Excel の基本操作	1
1) 各部の名称	1
2) データの入力と編集	2
3) 数値の計算	3
① 数式を入力する	3
② セル番地を使った数式を入力する	4
③ 関数を使って計算する	5
4) 比率の計算と絶対セル参照	6
5) 第 1 章の課題	9
2 データ特性の表し方 (基本統計量)	11
1) さまざまな統計データ	11
2) さまざまな統計値	12
① 中心の特性値	12
② ばらつきの特性値	17
3) 第 2 章の課題	20
4) 統計学からの補足	20

3	時系列データの分析	23
1)	時系列データとは	23
2)	図表付き分析レポートの作成	24
3)	時系列データの分析	28
4)	第3章の課題	38
4	多変数データの関係	39
1)	多変数データと散布図	39
2)	相関係数	41
3)	散布図を利用したデータの分類	42
4)	第4章の課題	44
5)	統計学からの補足	45
5	単回帰分析	49
1)	回帰分析とは何か	49
2)	Excelによる回帰分析	51
3)	分析結果の読み方	53
4)	第5章の課題	55
5)	統計学からの補足	56
6	重回帰分析	57
1)	重回帰分析とは	57
2)	2次関数による回帰分析	57
3)	ダミー変数	60
4)	重回帰分析をする際の注意点	63
5)	第6章の課題	67
6)	統計学からの補足	68

7	社会調査の基礎	69
1)	社会調査とは	69
2)	経済分析と社会調査	70
3)	社会調査をする際の注意	70
	① 個人情報の保護や機密の保持	70
	② 説明責任とハラスメント回避	70
	③ 調査結果の尊重	71
	④ データのねつ造・盗用	71
4)	調査票調査の進め方	72
	① 研究課題の明確化と作業仮説の設定	72
	② 調査の企画	73
5)	調査票の作成	76
	① 調査方法と質問の分量	77
	② ワーディング	77
	③ 選択肢の設定	80
	④ 調査票全体への配慮	82
	⑤ 予備調査とプリテスト	82
6)	第7章の課題	83
8	アンケート調査の集計と分析	85
1)	調査票の回収とデータの入力およびチェック	85
	① 調査票の回収とチェック	85
	② データの入力	86
	③ データのクリーニング	87
2)	データの集計	88
3)	集計結果の検定	93
4)	第8章の課題	95
5)	統計学からの補足	95

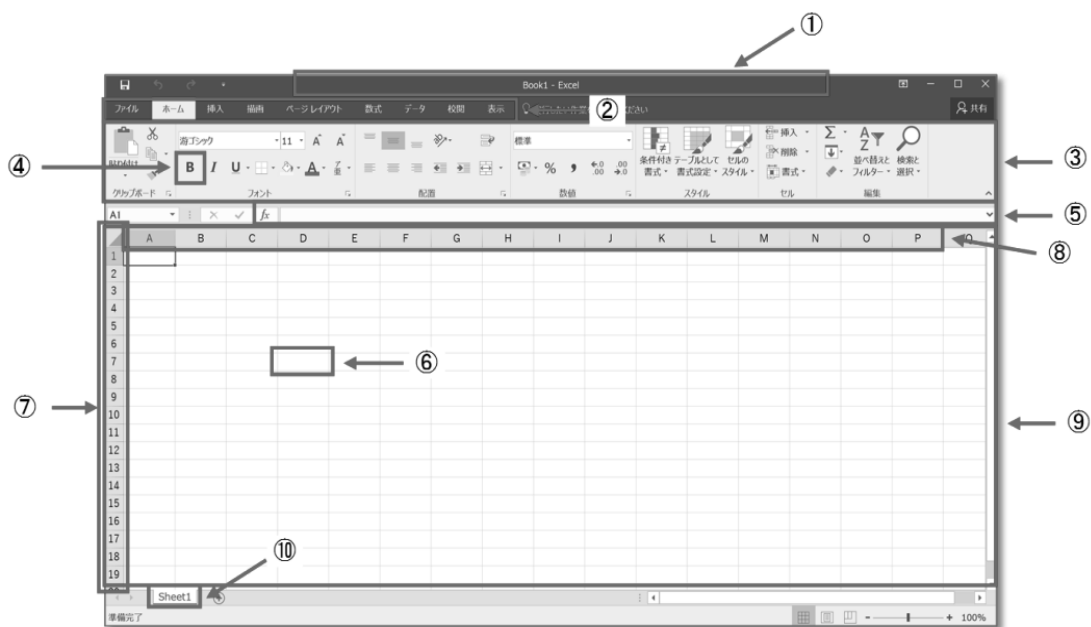
1. Excel の基本操作

1) 各部の名称

本書で使用するのは Excel 2016 for Windows である。使用するバージョンが異なると画面や操作方法も異なることがあるので注意する必要がある。

図 1-1 は Excel の画面の各部の名称を示している。本書で操作方法を説明する際にこの名称を使うので、必ず覚えてほしい。

図 1-1 Excel の画面各部の名称

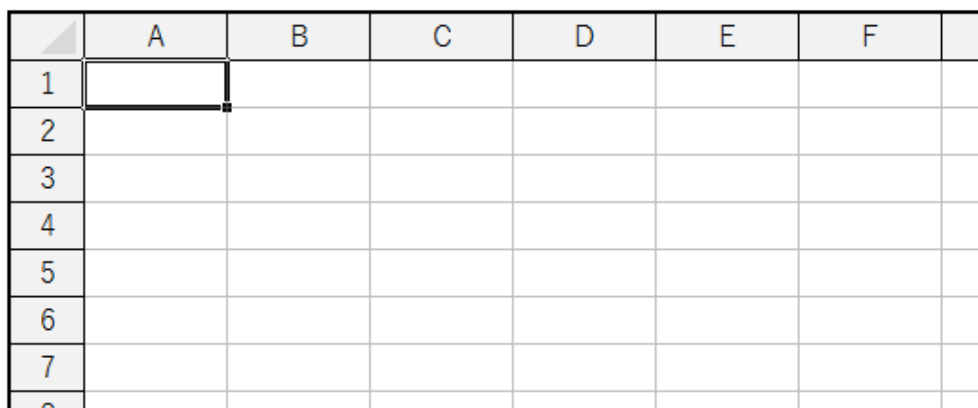


- | | | | |
|---|---------------|---|--------|
| ① | タイトルバー | ⑥ | セル |
| ② | タブ | ⑦ | 行番号 |
| ③ | リボン | ⑧ | 列番号 |
| ④ | コマンドボタン (ボタン) | ⑨ | ワークシート |
| ⑤ | 数式バー | ⑩ | シート見出し |

2) データの入力と編集

Excel は表計算ソフトとよばれるとおり、表形式のワークシートにデータを入力して計算や分析を行う。ワークシートのマスひとつひとつをセル（小部屋とか細胞という意味）と呼ぶ。セルの位置はセル番地で表す。セル番地は列番号+行番号で表す。たとえば、ワークシートの左上のセルは A1 番地となる。セルをクリックするとセルの周りが太い線で囲まれる（図 1-2）。これをセルがアクティブな状態という。アクティブな状態のセルをアクティブ・セルという。この状態でキー入力をするすると、アクティブ・セルに文字が入力できる。

図 1-2 アクティブ・セル



The image shows a portion of an Excel spreadsheet. The columns are labeled A through F, and the rows are labeled 1 through 8. The cell at the intersection of column A and row 1 (cell A1) is highlighted with a thick black border, indicating it is the active cell. The rest of the cells in the grid are empty and have a standard thin border.

豆知識

セルにデータを入力するときに、日本語を入力する場合は日本語入力を ON にし、数値を入力する場合は日本語入力を OFF にしよう。

A1 セルに数値の 10 を入力してみよう。日本語入力を OFF にして、キーボードから 10 をタイプすると、セルの中にカーソルが現れる（図 1-3）。この状態を編集可能状態という。数値を入力して **Enter** キーをタイプすると、セルの中の数値が自動的に右揃えになり、アクティブ・セルが下 (A2 番地) に移動する。なお、文字を入力すると左揃えとなる。

入力ミスをしてしまった場合、アクティブ・セルを A1 に移動し、文字をタイプし直すと文字が上書きされる。文字を部分的に修正したい場合は、**F2** キーをタイプすると編集可能状態となり、矢印キーが使用可能になるので、修正し

たい箇所へ移動して文字や数値を編集し、**Enter** キーで確定する。

図 1-3 文字の入力

	A	B	C	
1	10			
2				
3				
4				

練習問題 1.1

下表の通りデータを入力してみよう。

	A	B	C	D
1	学生番号	(番号を入力)		
2	氏名	(氏名を入力)		
3				
4	300			
5	500			
6	700			
7				

3) 数値の計算

Excel は表計算ソフトといわれるように、ワークシートを使って計算するためのソフトウェアである。計算するには ① 数式を入力する方法、② セル番地を使って数式を入力する方法、③ 関数を入力する方法がある。

① 数式を入力する

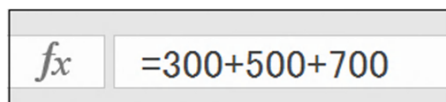
以下の手順で「 $300 + 500 + 700$ 」を計算し、結果を A7 セルに表示してみよう。

- ・ A7 をアクティブにする

- ・ 日本語入力を OFF にして、 $= 300+500+700$ と入力し、**Enter** キーをタイプする
- ・ A7 に計算結果が「1500」と表示される

A7 をアクティブにして数式バーを確認すると (図 1-4)、入力した数式が表示される。これがセルの中身である。このように、Excel はセルの中身と表示が異なる場合があることを理解しよう。

図 1-4 数式バー



足し算と同様に、引き算、かけ算、割り算、べき乗は以下の記号を用いる。

計算	記号 (読み方)	式の例	例の答え
足し算	+ (プラス)	$= 3 + 2$	5
引き算	- (マイナス)	$= 3 - 2$	1
かけ算	* (アスタリスク)	$= 3 * 2$	6
割り算	/ (スラッシュ)	$= 3 / 2$	1.5
べき乗	^ (ハット)	$= 3 ^ 2$	9

② セル番地を使った数式を入力する

数式による計算より便利な方法が、セル番地を使った数式による計算である。こちらの方が Excel らしい計算方法といえる。以下の手順で数式を入力してみよう。

- ・ A7 をアクティブにする
- ・ 日本語入力を OFF にして、「=」をタイプする
- ・ A4 セルをクリックし、「+」をタイプする
- ・ A5 セルをクリックし、「+」をタイプする
- ・ A6 セルをクリックし、**Enter** をタイプする
- ・ A7 に計算結果が「1500」と表示される

A7 をアクティブにして数式バーを見ると「=A4+A5+A6」となっていることが確認できる。これは、「セル A7 の内容は A4、A5、A6 に入力されている数値を加えた結果とする」ということを意味している。

このように、数式を入力する時にセルの内容をセル番地で指定することをセル参照という。ここで、セル A4 の内容を 300 から 500 に修正してみよう。すると、A7 が 1700 に変わる。セル参照を使って計算すると、セルの値が変化したらそれがすぐに計算結果に反映される。

③ 関数を使って計算する

Excel には、よく使う計算がワークシート関数として組み込まれている。数値を合計する場合は sum 関数を使う。以下の手順でセル A4 から A6 の数値を合計する関数を入力してみよう。

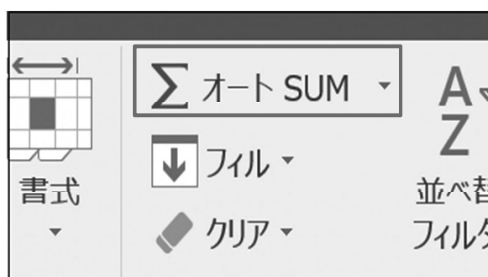
a) 関数を入力する方法

- ・ A8 をアクティブにする
- ・ 日本語入力を OFF にして、「=sum(」とタイプする
- ・ マウスで A4 から A6 をドラッグし、キーボードから「)」をタイプし、 をタイプする
- ・ A8 に計算結果が「1700」と表示される

b) コマンドボタンを使う方法

- ・ A8 をアクティブにする
- ・ 「ホーム」タブの右方にあるオート SUM ボタン (図 1-5) をクリックすると自動的に関数が入力されるので、 をタイプする

図 1-5 オート SUM ボタン



1. Excel の基本操作

(オート SUM を実行すると連続するセルが自動的に認識されるが、セル参照の範囲が正しくない場合は、マウスで正しい範囲をドラッグして指定し直す)

- ・ どちらの方法を使っても、セル A8 には「=sum(A4:A6)」が入力される。セル番地の範囲を指定する場合は、番地と番地を「:」(コロン)でつなげる。

4) 比率の計算と絶対セル参照

足し算の他にもよく使うのが比率の計算である。1950、1980、2015 年の日本の人口データから、以下の手順で年齢階層別の構成比を計算してみよう。

	A	B	C	D	E	F	G
1	人口構成の推移			単位:千人			単位:%
2		1950年	1980年	2015年	1950年	1980年	2015年
3	年少人口(15歳未満)	29786	27507	16096			
4	生産年齢人口(15~39歳)	33070	45128	34413			
5	(40~64歳)	17099	33707	42573			
6	老年人口(65歳以上)	4155	10647	33732			
7	総人口						

豆知識

表の左側の項目名が書かれた部分（上表では年齢階層が書かれた部分）を表側（ひょうそく）、上側の項目名が書かれた部分（上表では年次が書かれた部分）を表頭（ひょうとう）と呼ぶ。

総人口を計算する

- ・ 新規シートを開き、上表の通りデータを入力する (罫線は後で引くので文字と数値だけ)
- ・ B7 をアクティブにする
- ・ オート SUM ボタンを使ってセル B3 から B6 を合計し、1950 年の総人口を計算する

- ・ B7 をアクティブにし、セルの外枠の右下にある四角いフィル・ハンドル (図 1-6) を D7 までドラッグする
- ・ B7 の数式が右方向にコピーされる (数式バーでセル C7、D7 の内容を確認する)

数式の入力されたセルを右方向にコピーすると、C7 (=SUM(C3:C6)) や D7 (=SUM(D3:D6)) のようにセル番地が右方向にずれていることがわかる。このように、式をコピーした時にセル番地もコピーした方向に変化するような参照方法を相対セル参照という。

図 1-6 フィル・ハンドル

A	B	C
〇推移		
	1950年	1980年
15歳未満)	29786	2750
人口(15~39歳)	33070	4512
(40~64歳)	17099	3307
65歳以上)	4155	1064
	84110	

豆知識

セルの入力が終わった時に **Enter** ではなく **Tab** をタイプすると、入力
が確定し、アクティブ・セルは右に移動する。さらに、行末で **Enter** を
タイプすると、アクティブ・セルは次の行の先頭に移動する。

数値の桁が大きくなると読みにくいので、3桁ごとにカンマで区切って読みやすくしよう。エクセルでは、セル内のデータは変えずに、その見え方を表示形式で設定する。表示形式には、「桁区切りスタイル」のほかに「パーセント・スタイル」「小数点の桁数」「負の数のスタイル」「通貨スタイル」「日付スタイル」「時刻スタイル」などがある。

桁区切りを表示させる

- ・ B3 から D7 までをドラッグして選択状態にする

1. Excel の基本操作

- ・ 桁区切りボタン (図 1-7) をクリックする (数式バーでセルの内容と表示形式の違いを確認する)

図 1-7 桁区切りボタン



構成比を計算する

- ・ 日本語入力を OFF にする
- ・ E3 をアクティブにし、「=」をタイプ、B3 をクリック、「/」をタイプ、B7 をクリックして、**Enter** をタイプする
- ・ 結果が「0.354131494」と表示される

豆知識

やや広い範囲のセルを選択するときは、選択範囲の左上のセルをアクティブにし、**Shift** キーを押したまま選択範囲の右下のセルをクリックすればよい。

また、選択範囲の左上をアクティブにしてから、**Shift** + **Ctrl** を同時に押しながら **→** キーをタイプするとデータの右端まで、さらに **↓** キーで下端まで選択できる。

数式のコピー

E4 から E7 までは同様の操作をすればよいので、E3 の数式を下方向にコピーしてみよう。すると「#DIV/0!」と表示される。これは、数値をゼロで割ったというエラーメッセージである。

数式バーで E4 の数式を確認してみると、本来「=B4/B7」と入力すべきであるが、「=B4/B8」となっている。相対セル参照をしているため、分母のセル番地が下方向にずれてしまったためである。

このように、数式をコピーする際にセル番地を固定したい場合がある。このときは以下のように絶対セル参照を使う。

- ・ E3 をアクティブにし、「=」をタイプ、B3 をクリック、「/」をタイプ、B7 をクリックして、F4 キーを 2 回タイプし (セル番地が「B\$7」となる)、 をタイプする
- ・ 結果は同様に「0.354131494」と表示される
- ・ E3 の数式を E7 までコピーすると、結果が正しく表示される
- ・ 同様に、数式を F3 から G7 までコピーする

このように、セル番地の行番号や列番号の前にドル記号「\$」がついていると、数式をコピーしても番地が固定される。

パーセント表示にする

- ・ E3 から G7 を選択状態にし、パーセント・スタイル・ボタンをクリックする。桁下げボタンをクリックして小数点第 1 位まで表示させる

最後に罫線を引いて表を見栄え良く完成させよう。

5) 第 1 章の課題

「雇用形態別被雇用者数」のデータを使って、以下の通り各年の被雇用者の雇用形態別構成比を計算しよう。

- ・ A15 に「被雇用者の雇用形態別構成比」と入力する
- ・ 上の表と同じように表頭、表側を入力する
- ・ 絶対セル参照、相対セル参照を使い分けて、「雇用者（役員を除く）」を 100% とし、その内訳の構成比を計算する式を入力し、式をコピーして表を完成させる
- ・ 上の表と同じように「単位：%」を入力する

1. Excel の基本操作

- ・ 数値の表示形式はパーセント・スタイルとし、小数点第 1 位まで表示させる
- ・ 罫線を引き、見栄えを整える
- ・ 「課題 1」というファイル名で保存する
- ・ 計算結果から、近年の雇用形態の動向について考察しよう

2. データ特性の表し方 (基本統計量)

1) さまざまな統計データ

統計データとは

経済の動向を把握し、その背後にあるメカニズムを理解したり、問題を克服する政策を考えるためには、経済データを分析する必要がある。経済分析で使用する統計データには、次のような特徴がある (田中 [2009] p.2)。

- ① 数字または文字の並びであり、変動的である
- ② 現実の集団現象や集団の構成メンバー (国民、企業、地域など) の特性を観測・計測・調査した結果である

データの種類

統計データは単なる数字や文字の集まりではなく、集団の特性に関する変動する要因 (変数) を計測したものである。このうち、人口や売上、利益、賃金のように数値で表せるデータを量的データといい、それ以外のデータを質的データという。

経済データの計測形態は、大きく時系列データ (タイム・シリーズ・データ) と横断面データ (クロス・セクション・データ) の2つの種類に分けることができる。時系列データは、時間の経過とともに計測されるデータ、横断面データは同一期間内に計測されたデータである。たとえば、2015年の地域別の失業率のデータは横断面データであり、2000年から2015年までの北海道の失業率データは時系列データである。表 2-1 は地域別失業率の推移を示しているので、横断面・時系列データである。

表 2-1: 時系列データと横断面データ

	北海道	東北	南関東	北関東 ・甲信	北陸	東海	近畿	中国	四国	九州 ・沖縄
2000	5.5	4.4	4.8	3.8	3.6	3.7	5.9	3.9	4.1	5.4
2001	5.9	5.0	4.9	4.1	3.9	4.1	6.3	4.2	5.1	5.6
2002	6.0	5.9	5.4	4.4	4.0	4.1	6.7	4.3	5.2	6.1
2003	6.7	5.6	5.1	4.6	4.0	4.0	6.6	4.3	4.8	5.9
2004	5.7	5.4	4.6	4.1	3.7	3.5	5.6	4.3	4.9	5.5
2005	5.3	5.0	4.3	3.7	3.3	3.2	5.2	3.8	4.3	5.3
2006	5.4	4.8	4.0	3.5	3.4	3.0	5.0	3.5	3.9	5.0
2007	5.0	4.7	3.6	3.2	3.4	2.7	4.4	3.6	3.9	4.7
2008	5.1	4.7	3.8	3.5	3.4	2.9	4.5	3.6	4.5	4.6
2009	5.5	6.0	4.8	4.7	4.5	4.6	5.7	4.7	5.0	5.4
2010	5.1	5.7	5.1	4.7	4.2	4.1	5.9	4.2	4.5	5.7
2011	5.2	5.3	4.6	4.4	3.9	3.7	5.0	3.7	4.6	5.2
2012	5.2	4.5	4.4	3.7	3.5	3.5	5.1	3.7	4.2	4.8
2013	4.6	4.0	4.1	3.8	3.4	3.3	4.4	3.8	3.8	5.1
2014	4.1	3.6	3.5	3.2	3.1	2.8	4.1	3.3	3.6	4.8
2015	3.4	3.6	3.3	3.2	2.7	2.6	3.8	3.2	3.2	4.5

出所：総務省統計局「労働力調査」

注：2013年以降の九州・沖縄の値は単純平均値

2) さまざまな統計値

たくさんの数値や文字が並んでいる統計データ全体に含まれる情報や特性を知るために、さまざまな統計値が使われている。そのなかで基本的なものは、中心の特性値とばらつきの特性値である。

① 中心の特性値

統計データは、多くの数値や文字列の並びであり、多数のままではその特性を把握しにくい。データの並びを代表するような数値があれば把握しやすくなる。このような計算をデータの縮約と呼び、縮約によって得られた値を特性値と呼ぶ。

平均 (average)

量的データを縮約するために最も頻繁に使われるのが算術平均、あるいは単に平均と呼ばれる特性値である。平均はデータの合計をデータの個数で割って求める。

$$\text{平均 } (\bar{x}) = \frac{\text{データの合計}}{\text{データの個数}}$$

練習問題 2.1

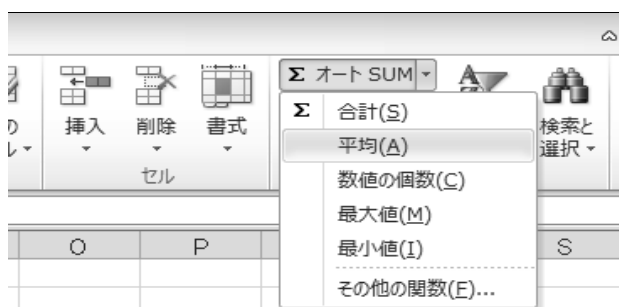
「地域別失業率」のデータを使って、以下の手順で全国 10 地域の平均失業率を計算してみよう。

a) 数式による計算

- ・ L3 に「全国平均」と入力する
- ・ L4 をアクティブにし、オート SUM ボタンをクリックする。データ範囲がおかしいので、マウスでデータ範囲を「B4:K4」に修正する
- ・ F2 キーをタイプして編集可能状態にし、カーソルを行末に移動し、「/10」と入力し、をタイプする

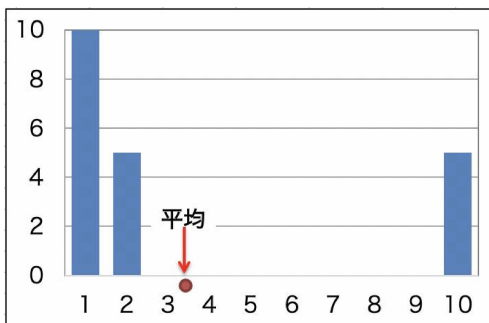
b) 関数による計算

- ・ L4 をアクティブにし、オート SUM ボタンの脇にある ▼ をクリックして「平均」を選択する
- ・ マウスでデータ範囲を「B4:K4」に修正し、をタイプする
- ・ L4 をアクティブにして、フィル・ハンドルをダブルクリックし、数式をコピーする



平均のウソ

平均は、平均点、平均身長、平均所得などよく使われる特性値であるが、極端に大きな数値や小さな数値に影響を受けやすく、データの中心を的確に表さないことがある。たとえば、次の図のようにある試験で1点が10人、2点が5人、10点が5人の場合、平均点は3.5点であるが、これが回答者の分布特性を的確に示しているとは言い難い。



下の表のB社は、10年未満、10年以上ともにA社よりも平均月給が高い。しかし、社員全体の平均月給を計算すると、A社の方が高くなる。このことをもってA社の月収が高いというのは妥当ではない。

勤続年数	A社		B社	
	平均月給	人数	平均月給	人数
10年未満	26万円	50	30万円	100
10年以上	48万円	150	50万円	100
全体	42.5万円	200	40万円	200

(田中 [2009] pp.13-14)

このように、平均値が集団の誤った特性を示してしまうことを平均のウソという。平均値を使用したり、平均に関する記述を読むときは、平均のウソがないかどうか注意する必要がある。

中央値 (median)

量的データを小さい順に並べたとき、中央に位置するデータの値を中央値(メディアン)という。データの数が偶数の場合は、中央の2つの平均とする。たとえば、以下のような6つのデータの場合、平均値は外れ値である90の影響

を受けて 20 となり、中央値は 6 と 7 の平均で 6.5 となる。この場合の中央値は平均値と違い、データの中央を的確に示していることがわかる。

↓
4 5 6 7 8 90

データの分布に偏りがある場合、中心を表す特性値としては平均よりも中央値が適している。たとえば、所得や貯蓄などは、一部の人が極端に高い所得や貯蓄を持っているので、一般に考えられる中所得階層に比べて平均が高めに出る傾向がある。そこで、所得階層の中心を表すには中央値を用いる方が適切である。なお、データの分布が左右対称であれば、平均と中央値は一致する。

練習問題 2.2

「地域別失業率」のデータを使って、M 列に毎年の地域別失業率の中央値を計算してみよう。中央値を計算するワークシート関数は「=MEDIAN(データ範囲)」である。

最頻値 (mode)

データの中で最も頻度が高い値を最頻値 (モード) という。モードは平均や中央値と異なり、質的データにも適用できる。たとえば、ラーメン屋の 1 年間の売上に関するデータで、醤油、塩、味噌、とんこつのうち、味噌ラーメンを注文した人が最も多ければ、モードは味噌ラーメンである。味噌と醤油が同数であれば、モードは 2 つになる。連続的な値を取る量的データの場合は、度数分布表において最大度数を持つ階級の中点を最頻値という。

練習問題 2.3

「地域別失業率」のデータを使って、以下の手順で 2015 年の失業率の度数分布表を作成し、最頻値を計算してみよう。2015 年の地域別失業率は、東海の 2.6 から九州・沖縄の 4.5 まで分布しているので、2.0 から 0.5 刻みの度数分布表を作る。

- ・新しいシートを作成し (シート見出しを右クリックし、[挿入] - [ワークシート])、シート名をダブルクリックして「度数分布表」と入力し、Enter をタイプする

2. データ特性の表し方

- ・ A2 に 2、A3 に 2.5 を入力し、A2 と A3 を選択状態にするフィル・ハンドルを下方向に A8 までドラッグする。これが度数分布の区間を示す「区間配列」となる
- ・ B1 に「度数」と入力する
- ・ B2 がアクティブな状態で、オート SUM ボタンの脇の ▼ をクリックして「その他の関数」を選択する「関数の分類」を統計にし、FREQUENCY 関数を選択して「OK」ボタンをクリックする
- ・ 「データ配列」は「地域別失業率」シートの 2015 年のデータを選択すると「地域別失業率!B34:K34」と入力される
- ・ **Tab** キーをタイプし、「区間配列」に移動し、A2 から A8 をドラッグすると「A2:A8」と入力されるので、**Enter** をタイプして入力を確定させる
- ・ B2 から B8 を選択状態にし、F2 をタイプすると B2 が編集可能状態になるので、そのまま **Ctrl**+**Shift** を押さえながら **Enter** をタイプする

この結果、3.0 より大 3.5 以下の区間が最も度数が高いことがわかった。したがって、モードは

$$\frac{3.0 + 3.5}{2} = 3.25$$

となる。モードは区間の取り方によっても変わってくるうえ、複数存在することもあり、すべての区間で同じ度数となりモードが中心の特性値として意味をなさないこともある。

豆知識

FREQUENCY 関数を探すときは、「関数の分類」を「統計」にしてから、FR と続けてタイプするとすぐに見つけることができる。

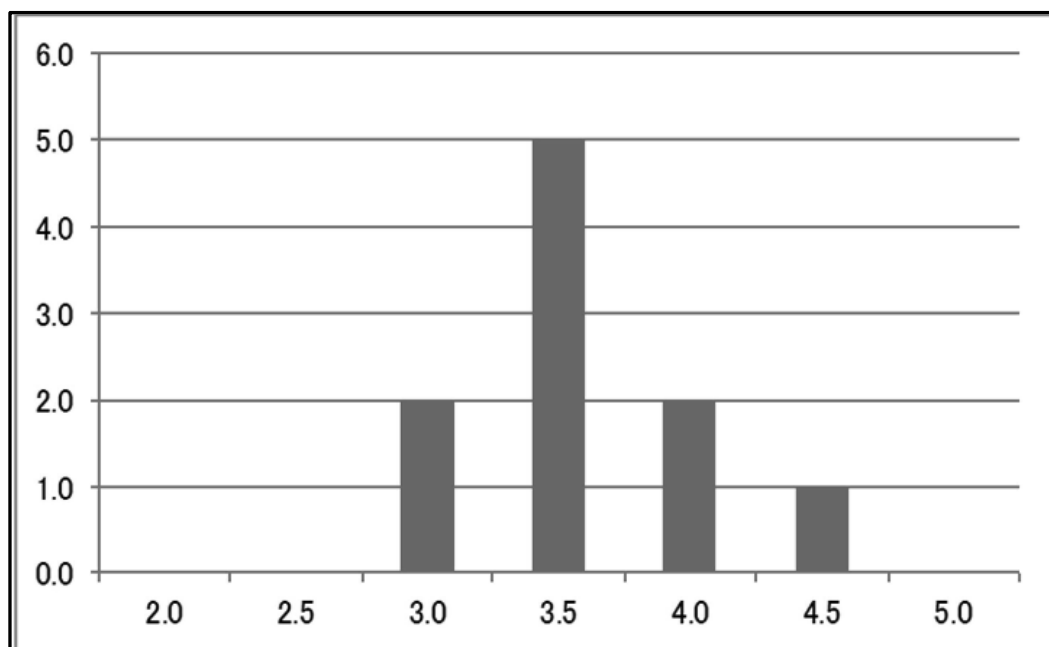
練習問題 2.4

作成した度数分布表をもとに、ヒストグラム（度数分布図）を作ってみよう。

- ・ 「度数分布表」シートの A1 から B8 までを選択状態にする

- ・ 「挿入」 タブの「グラフ」 グループの「縦棒グラフの挿入」 ボタンをクリックし「2-D 縦棒 (集合縦棒)」 を選択する

図 2-1 ヒストグラム



② ばらつきの特性値

量的データでは、中心の特性値とともに、データのばらつき、あるいは散らばりの程度が、データを縮約するうえで重要な特性値となる。

範囲 (range)

データのばらつきを示す最も単純な指標が範囲である。範囲はデータの最大値と最小値の差で表される。データの範囲は、異常値の影響を受けやすいため、データの特性値としては適切ではないが、度数分布表を作成するのに必要な指標である。

練習問題 2.5

「地域別失業率」のデータを使って、N 列に毎年の失業率の範囲を計算してみよう。最大値、最小値を計算するワークシート関数はそれぞれ「=MAX(データ範囲)」、「=MIN(データ範囲)」である。

分散 (variance)

データがデータ全体の平均からどれだけ散らばっているかを示す指標として、分散が用いられることがある。次式の通り、各データと平均との差を2乗した値を合計し、データの個数マイナス1で割った値と定義される。各データと平均との差を2乗しているため、データの測定単位と分散の単位が異なる点に注意する必要がある。

$$\text{分散} = \frac{\sum(\text{各データ} - \text{平均})^2}{\text{データの個数} - 1}$$

標準偏差 (standard deviation)

平均のまわりのばらつきを測る代表的な特性値が標準偏差である。標準偏差は次式の通り分散の平方根(ルート)と定義される。2乗した値の平方根であり、データの測定単位と等しくなるので、ばらつきの指標としてより適切である。

$$\text{標準偏差} = \sqrt{\text{分散}} = \sqrt{\frac{\sum(\text{各データ} - \text{平均})^2}{\text{データの個数} - 1}}$$

練習問題 2.6

「地域別失業率」のデータを使って、2015年の地域別失業率の標準偏差を以下の手順で計算してみよう。

a) 数式による計算

- ・新しいシートを作成し、シート名をダブルクリックして「標準偏差」と入力し、 をタイプする
- ・「地域別失業率」シートの表頭を「標準偏差」シートの2行目にコピーする

- ・ B3 に「地域別失業率」シートの 2015 年の北海道のデータと全国平均の差を 2 乗した値を計算する

このとき平均を絶対セル参照にする

B3 の式は「=(地域別失業率!B34- 地域別失業率!\$L\$34)^2」となる

- ・ B3 の式を K3 までコピーする
- ・ L3 に B3 から K3 間での合計を 9 (=10 地域 -1) で割り、その平方根を計算する

平方根を計算する Excel のワークシート関数は「=SQRT(データ)」である

L3 の式は「=SQRT(SUM(B3:K3)/9)」となる

b) 関数を使った計算

- ・ Excel には標準偏差を計算する関数が組み込まれているので、これを使うと簡単に計算できる「標準偏差」シートの M3 に関数を使って標準偏差を計算し、L3 と同じ値になっているかどうか確認してみよう標準偏差を計算するワークシート関数は「=STDEV.S(データ範囲)」である

豆知識

各データから平均を差し引き、標準偏差で割ると、データは平均 0、標準偏差 1 となる。このような変換をデータの標準化という。標準化されたデータは、測定単位に依存しないデータとなる。いわゆる偏差値は、平均 50、標準偏差 10 となるよう標準化した値である。

変動係数

標準偏差は、測定単位に応じて変化する。そのため、たとえば生産額を 1,000 万円単位で表示した場合は 1 億円単位の標準偏差の 10 倍となり、ばらつきが大きいような印象を与えてしまう。また、平均が大きいほど標準偏差は大きくなる。そこで、測定単位に依存しないよう変換した特性値として変動係数が使われる。ただし、変動係数は一般に平均が正のデータの場合に使われる。

$$\text{変動係数} = \frac{\text{標準偏差}}{\text{平均}}$$

練習問題 2.7

「地域別失業率」のデータを使って、O 列に毎年の標準偏差を、P 列に毎年の変動係数を計算しよう。標準偏差と変動係数の推移を表す折れ線グラフを作成し、失業率のばらつきが 1985 年以降どのように変化したのかを観察してみよう。

3) 第 2 章の課題

「家計消費支出」のデータを使って、支出項目別に年齢階層全体の平均、中央値、範囲、標準偏差、変動係数を計算しよう。

- ・ 数値の表示形式は桁区切りスタイルとし、小数点第 1 位まで表示させる
- ・ 「家計消費支出」シートを「課題 2」というファイル名で保存する
- ・ 分析結果から、どのようなことがわかるかを考察してみよう

4) 統計学からの補足

母集団 (population)

統計データが収集されたもとの集団を**母集団**という。これは統計分析の対象となる全体であり、通常非常に膨大なものである。その一部である統計データは母集団から任意に得られたもの (**任意標本**) で、母集団の情報を代表するものと捉えられる。その情報を精度高く取り出すことが統計分析の役割といえる。

母集団には質的なものと量的なものがあり、前者からは質的データが、後者からは量的データが収集される。本章で学んだ統計値は主に量的母集団に対するものである。量的母集団の中心的な位置を表す代表的な特性量が**母平均**であり、ばらつきの程度を表す特性量が**母分散**と**母標準偏差**である。これらの値も含め、母集団の特性量を推測することが統計分析において求められる。

推定量 (estimator)

母集団の特性量に使われる統計値は**不偏性**と呼ばれる性質をもつことが必要である。不偏推定量は、その平均が対象となる特性量に一致するもので、この性質が推定として使われる根拠となる。

- ・ 本章 2) ① で定義された平均は標本平均とも呼ばれ、母平均の不偏推定量である

これは標本平均の平均が母平均に一致するという性質に基づく

- ・ 本章 2) ② で定義された分散は不偏分散とも呼ばれ、母分散の不偏推定量である

これも不偏分散の平均が母分散に一致するという性質に基づく

- ・ 分散の定義で使われた $\sum(\text{各データ} - \text{平均})^2$ を変動という

変動はデータのばらつきを示す基本的な統計量である

また、分散の定義式の分子 (標本数 - 1) を自由度と呼ぶ

一般に、変動を自由度で割ることにより母分散が推定できる

- ・ 母標準偏差は母分散をルートしたものなので、母標準偏差の

推定量は不偏分散をルートして行う

- ・ データに対する分散には不偏分散とは別に標本分散がある

$$\text{標本分散} = \frac{\text{変動}}{\text{データの個数}}$$

標本分散の平均は母分散からずれるので母分散の推定量ではない。しかし、この分散も種々の分析に使われるので、いずれの分散が使われているのか、確認することが必要である。

Excel の統計関数

ここで学んだ統計値を求める関数が Excel では定義されている。それらも含めて基本的なものをまとめておこう。

*) ここで「範囲」とは、データが入力されているセルの範囲である。

2. データ特性の表し方

Excel 関数	統計値
COUNT(範囲)	数量データの個数 範囲内の数量データの個数 (標本数) を数え上げる
COUNTA(範囲)	質的データの個数 範囲内の全データの個数を数え上げる 質的、数量データともに数え上げ
AVERAGE(範囲)	平均 範囲内の数量データの平均を出す
MEDIAN(範囲)	中央値 範囲内の数量データの中央値を出す
MODE(範囲)	最頻値 範囲内の数量データの最頻値を出す
DEVSQ(範囲)	変動 範囲内の数量データの変動を出す
VAR.S(範囲)	不偏分散 範囲内の数量データの不偏分散を出す
STDEV.S(範囲)	標準偏差 範囲内の数量データの不偏分散に対する 標準偏差を出す
VAR.P(範囲)	標本分散 範囲内の数量データの標本分散を出す
STDEV.P(範囲)	標準偏差 範囲内の数量データの標本分散に対する 標準偏差を出す
MAX(範囲)	最大値 範囲内の数量データの最大値を出す
MIN(範囲)	最小値 範囲内の数量データの最小値を出す