1.2

母集団

前節でも述べたように、統計学ではまず、調査のために「本当に知りたいこと」を明確にする必要があります。さらには、それを知るためにはどのようなデータを集めなくてはならないのかを適切に設定しなくてはなりません。

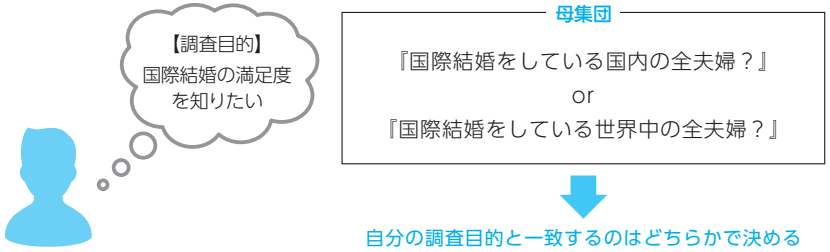
日本人の正確な平均身長を知るためには、日本人1人残らずの全員分のデータが必要になります。ほかには、「国際結婚している夫婦の満足度」を正確に調べるためには、全世界で国際結婚をしている夫婦全員分のデータを集める必要がありますし、「エアコンの平均設定温度」を正確に調べるためには、全家庭のすべてのエアコンを調査する必要があります。

統計学では、本当に知りたいことの真の値（正解値）を知るために、調べなくてはならないデータ全体のことを**母集団**（ぼしゅうだん）と呼び、調査を行ううえでは、まず母集団を定義することが最初の一歩となります。

表 1.1　調査目的と母集団の例

調査の目的	母集団
国際結婚をしている夫婦の満足度	世界全体で国際結婚している全夫婦
日本国内に住む、国際結婚をしている夫婦の満足度	国際結婚をしていて、日本に住んでいる全夫婦
エアコンの設定温度の平均値	世界中の全家庭にあるすべてのエアコン
日本人の未成年の学習時間の調査	日本人の未成年全員

図 1.2　調査目的と母集団の定義



練習問題 1.1

次の (1) ～ (3) の調査を行うための、母集団を定義してください。

(1) 日本の高校生の平均握力。

(2) 日本人が所有するスマートフォンにインストールされているアプリの数。

(3) 癌（がん）の新しい治療薬の効果性。

解答と解説 1.1

(1) は「日本の高校に通っている全生徒」になります。

(2) は「日本人が所有している全スマートフォン」ということになりますが、これは「現在も契約があり利用されているスマートフォン」のみになるでしょうか、それとも「もう契約もなく、使われてはいないが家にずっとあるスマートフォン」も含むのでしょうか。その答えは「調査の目的による」としかいえません。調査者の知りたいことが、契約中のスマートフォンのみであれば前者となりますし、使われていないスマートフォンも含むのであれば後者となります。

このように母集団の定義や範囲は、その目的、すなわち何を明らかにするために調査を行うのかによって変わってきます。

それでは (3) について考えてみましょう。母集団を「現在の全癌患者」とした場合、これから癌にかかってしまう患者には効かないかもしれず、そのありがたみが減ってしまいます。したがって、本来の目的からすると、母集団は「現在の全癌患者に加え、未来の全癌患者」ということになるかと思います。しかし、未来の癌患者に対して治療薬の効果性を調査することは不可能であり、現実問題として、しかたなく「現在の癌患者たち」を対象として調査することになります。このように、いくらお金や時間もあって、すべての人間が協力的であったとしても調査のできないこともありますが、母集団の定義自体は本来の目的に沿って設定した方が良いです。

Excel でやってみよう

Excelを利用して、表2.2形式で記録されているデータの縦横の向きを変換したり（283ページ）、表2.1のデータをテストの点数順に並べ替えたりする（284ページ）操作を行ってみましょう。

練習問題 2.1

小学校のあるクラスの血液型を調べた下記の表について、(1) および (2) の設問に答えてください。

表2.3　一部重複のあるデータ

名前	血液型
相原太郎	A
石川春子	AB
江川次郎	O
岡田夏美	B
岡田夏美	B

(1) 項目数はいくつですか。

(2) レコード数はいくつですか。

解答と解説 2.1

(1) は「名前」と「血液型」の2項目です。また5名分のデータが入っているので、(2) のレコード数は5件です。

ここで1つ気になるポイントがあります。「B型の岡田 夏美」さんというデータが2件あります。これは見た目上、まったく同じデータですので、一見すると間違いなのか、本当に2人いるのか区別できません。ここでは仮に、間違いではなく本当に2人いたとして、さらに「学校から家まで何分掛かるか」というデータを加えたとしましょう。

表2.4　一部重複のあるデータ

名前	血液型	家から学校まで（分）
相原太郎	A	10
石川春子	AB	15
江川次郎	O	15
岡田夏美	B	5
岡田夏美	B	10

2人の「岡田 夏美」さんは別人なので、登校に掛かる時間が異なっていることは不自然なことではありません。では、続けて、「身長」のデータを追加するとします。「岡田 夏美」さんはそれぞれ130cmと132cmでした。それぞれどちらの身長に書き加えるべきか、だんだんとわかりづらくなってきます。また修正する際も、「岡田 夏美さんが引っ越して、家から徒歩7分になった」という指示では不十分で、どちらの岡田さんかわかりません。

この問題のそもそもの原因は、項目の中に、データを完全に区別できるものがなかったことです。もしここに、小学校ではおなじみの、「出席番号」という項目があったらどうでしょうか。

表2.5　主キーのあるデータ

出席番号	名前	血液型	家から学校まで（分）
1	相原太郎	A	10
2	石川春子	AB	15
3	江川次郎	O	15
4	岡田夏美	B	5
5	岡田夏美	B	10

今度は2人の岡田 夏美さんが区別できるようになりました。ここでのポイントは、出席番号はほかと重複することがないことです。このように、ほかと重複することのない項目があると、個々のデータを確実に区別できるようになります。先ほどまでと違い、「出席番号4番のデータに、身長132cmを追加する」、「出席番号5番のデータの、家から学校までの時間を7分に修正する」といったように、出席番号のみで迷わず区別できます。

このように、各レコードを完全に区別するための項目を**主キー**（しゅ）と呼びます。すなわち、表2.5の表では「出席番号」が主キーです。データを記録する際には、必

4.1

度数分布表

クロス集計表は主に、名義尺度や順序尺度などの質的変数を集計する際に、非常に有効的な手法といえます。また変数の数は、1つあるいは2つとなります。しかし、間隔尺度や比例尺度などの量的変数をクロス集計表でまとめようとすると非常に見づらい表となってしまいます。本章では、量的変数についてデータを見やすく集計するための**度数分布表**と呼ばれる表現方法を紹介します。

例として、200名分の身長を調査したデータを考えてみます。前章でも説明したように、『1人目の身長は162.2cm、2人目の身長は151.27cm、3人目は…』と読み上げていく方法は非常にわかりにくいです。一般的に、データというのは元のまま直接見ても、把握しづらいことが多いのです。このデータを説明するための別の方法として「140（以上）～145（未満）cmの人は1人、135～140cmの人は10人…」と伝えると、元データそのまま伝えるよりも大分わかりやすくなります。これを表の形にしたものが度数分布表になります。次の表が、200名分の身長データを示した度数分布表の例です。

表4.1　身長データの度数分布表

階級			度数
130以上	～	135未満 cm	0
135	～	140	0
140	～	145	5
145	～	150	11
150	～	155	18
155	～	160	26
160	～	165	30
165	～	170	41
170	～	175	31
175	～	180	27
180	～	185	7
185	～	190	4
190	～	195	0
195	～	200	0
合計			200

表4.1を見ながら、いくつかの注意点と用語を紹介していきます。

階級というのは「○○～□□の間に」ということを示したものです。表の中では1つ1つに単位を付けると大変なうえに、逆に見づらくなってしまうこともあるので、本例のように、最初の項目にのみ単位が表記されることも多いです。また、135cmの人がどちらに振り分けられるかが明確になるように、「以上・より大きい」や「未満・より小さい」などの記載もしておく方が良いでしょう。あるいは、テストの点数のように、小数点が存在しないことがわかっているデータの場合は、『0～9』で最初の階級、『10～19』で次の階級、とすることも可能で、その場合は「以上」や「未満」などを表記する必要はありません。

度数というのは、その階級に含まれるデータ数のことです。今回の場合は「人数」という見出しでも良さそうですが、データの種類によって「人数」や「個数」など、見出し名を変更することが面倒ですので、度数という用語を用います。通常、度数には「○○人」や「□□個」などの単位は付けずに省略します。また、度数の合計はデータ数と一致するはずですので、検算や確認に利用すると良いでしょう。

分布という言葉は一般的な用語としても使いますが、基本的に、どこにどの程度のデータ数が集まっているかといったことを示します。

以上の言葉を用いて度数分布表を定義すると、『階級ごとの度数を示し、分布を把握するための表』ということになります。

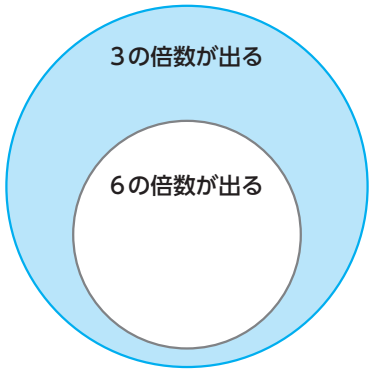
表4.1を確認すると、165～170cmのあたりにデータが集まっていることや、135～140cmあたりにはまったくデータがないことが確認できます。このように、分布の全体像を容易に把握できるようになることが度数分布表の最大の利点となります。そのために、度数分布表を作成する際には、度数順に並べたりするようなことは行わず、必ず階級の小さい順に行を並べるようにしましょう。

度数を確認することで、それぞれの階級にどの程度の人数がいるかわかりますが、それが全体の何割程度であるのかが、少しわかりづらいです。そこで**相対度数**と呼ばれる、各階級の構成比を度数分布表の中に示す方法もあります。表4.2が相対度数付きの度数分布表になります。

になります。逆に「6の倍数」でなかったとすると、1～5のいずれかの目が出たことになるので、「3の倍数」となる確率は$\frac{1}{5}$です。したがって、サイコロを振ってそれが6の倍数であったかどうかによって、3の倍数であるかどうかの確率が変わりますので、独立ではないことになります。

また、このケースでは、6の倍数が出た場合は、必ず3の倍数となります。このように片方が起こると、もう片方の事象も必ず成立するような関係を、**包含関係**と呼び、図で表すと下記のようなイメージとなります。

図7.1 包含関係



(3) については、2つの事象は一見関係するように見え、独立でないように見えます。しかし、もう少し詳しく考えてみましょう。

サイコロを1度振って、「2の倍数が出た」とすると、出た目の可能性としては2, 4 ,6であり、そのうち「3の倍数である確率」は$\frac{1}{3}$です。一方、「2の倍数が出なかった」とすると、可能性としては1, 3, 5であり、そのうち「3の倍数である確率」は$\frac{1}{3}$です。すなわち、2の倍数が出ても出なくても、3の倍数である確率は$\frac{1}{3}$で変わらないことになります。逆の事象から見ても同じで、3の倍数が出ても出なくても、2の倍数となる確率は$\frac{1}{2}$で変わりません。したがって、2つの事象は独立といえます。

7.2 確率の計算

和事象

複数の確率を組み合わせる場合には、「または」あるいは「かつ」で計算を行うことが基本となります。たとえば、サイコロを1度振って、「1か2か3の目が出る確率」を考える場合、それは「1の目が出る」または「2の目が出る」または「3の目が出る」確率を問われていることになります。感覚的に言ってしまえば、それぞれの事象の起こる確率が$\frac{1}{6}$ですので、

$$\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6}=\frac{1}{2}$$

となります。「または」の確率は足し算をすれば良いと、なんとなく認識をしている方もいるかもしれませんが、厳密にいえば、『排反である事象の場合、「または」の確率は足し算をして良い』という規則になります。

図7.2 排反事象の確率

図7.2に示したように、それぞれの事象に重なる部分がないので足し算を行って良いことになります。

続けて、重なる部分がある場合を考えてみましょう。サイコロで「3の倍数

8.2

標準偏差

前節では分散の求め方を学習しました。「散らばり具合を測る式として、前節の計算式で良いのか？　証明できるのか？」と疑問に思われた方がいるかもしれません。分散については、散らばり具合を測るうえで考えられた式ではありますが、正しいか証明のしようはありません。言ってしまえば、「分散とはこのように求める」と誰かが定義したにすぎません。これは平均値も同じで「なぜ合計値をデータ数で割り算するのか」ということに対しては、誰かがそのように定義したからとしか言いようがなく、私たちユーザー（使う側）はその値によって、読み取れることを学習することに努力すべきです。

では、分散についてもう少し深く学習していきましょう。下記のデータは100点満点の英語のテストと、1000点満点の英語のテストの結果で、それぞれについて分散を求めたものです。

表8.6　2つのテストにおける分散値

テストA	テストB
30	300
40	400
50	500
60	600
70	700
分散：200	分散：20000

両テストを見比べるとわかるのですが、テストの基礎（ベース）となる点数が10倍になっているのみで、2つのテストの結果は、ほぼ同等といえるのがわかるでしょうか。**得点率**に変換すると、よりわかりやすいかもしれません。いずれのテストにおいても5名の得点率は30%、40%、50%、60%、70%と変わりません。

それにもかかわらず、2つのテスト結果の分散の値が100倍も異なっています。この理由としては、分散を求める際の計算の過程で、偏差を二乗していることが挙げられます。点数のベースが10倍になると、分散の値は$10^2=100$倍大きくなります。この値を直接見ると、ベースの異なるデータ間で、分散の値を比べにくくなります。二乗したことが原因ですので、分散の$\sqrt{\ }$（ルート）値を求めることで、この差が埋まります。この分散の$\sqrt{\ }$値のことを、**標準偏差**と呼びます。では、表8.6について、それぞれの標準偏差を求めてみます。

表8.7　2つのテストにおける標準偏差

テストA	テストB
30	300
40	400
50	500
60	600
70	700
分散：200	分散：20000
標準偏差：$\sqrt{200} \fallingdotseq 14.142$	標準偏差：$\sqrt{20000} \fallingdotseq 141.42$

$$\text{標準偏差} = \sqrt{\text{分散}}$$
$$\text{分散} = (\text{標準偏差})^2$$

表8.7を見るとわかるように、テストAの標準偏差は約14.142、テストBの標準偏差は約141.42となり、その差が10倍となり、ベースとなる値の倍率と等しくなります。この標準偏差の値は、この後に学習する標準化得点や、あるいは本書では扱わない推測統計の学習には欠かせない値となります。標準偏差にも利点がたくさんありますが、「散らばり具合」を数値化したり比較したりする際には、基本的には分散の値が用いられることが多いので、その点はよく確認しておきましょう。

もう1つ、**変動係数**についても紹介します。先ほどの標準偏差の値を用いると、ベースとなる点数が10倍になると、標準偏差の値も10倍になることがわかりました。ベースとなる値が10倍異なれば散らばり具合の値も10倍異なるということで、特に違和感はありません。しかし、場合によっては、ベースと

8.4

不偏分散

前節では、「偏差の二乗和をデータの数で割り算する」という分散において、母集団の分散値と、標本調査での期待値が一致しないことを確認しました。では、母分散を推測する手段がないのかというと、少し計算方法を変えることで、推測することが可能となります。

では、分散をより正確に推測するための方法を紹介していきます。まず母集団のデータが手元にあるとして、偏差の二乗和まで求めていきます（表8.18）。

表8.18　母集団の偏差の二乗和

名前	身長（cm）	偏差	偏差の二乗
A	150	-20	400
B	160	-10	100
C	170	0	0
D	180	10	100
E	190	20	400
	母平均：170		偏差の二乗和：1000

偏差の二乗和は1000であり、これを標本の大きさである5で割り算したものが分散でした。今回は、5で割り算するのではなく、データ数よりも1つ少ない数、4で割り算します。

$$1000 \div (5-1) = 250$$

偏差の二乗和をデータ数よりも1つ少ない数で割り算した、この値を**不偏分散**と呼びます。

不偏分散＝偏差の二乗和÷（標本の大きさ－1）

したがって、母集団について、母分散は200でしたが、不偏分散の値は250

ということになります。

図8.2　分散と不偏分散

$$\text{分散：}\frac{1000}{5} = 200$$

$$\text{不偏分散：}\frac{1000}{5-1} = 250$$

本来知りたいのは分散の値なのに、この不偏分散の値にどのような意味があるのか、順を追って説明します。

5名中3名の標本調査においても、同様に不偏分散を求めてみます。たとえば、150、160、170cmの3名が標本として選ばれた場合、不偏分散の値は表8.19に示すように100となります。

表8.19　A、B、Cが選ばれた際の不偏分散値

名前	身長（cm）	偏差	偏差の二乗
A	150	-10	100
B	160	0	0
C	170	10	100
	標本平均：160		偏差の二乗和：200

↓

分散：200÷3＝66.66…
不偏分散：200÷（3－1）＝100

残りの9ケースについても、不偏分散の値を求めてまとめたものが表8.20です。

メモ　ほかの9ケースについても、いくつか選んで不偏分散を求めてみると、良い練習になると思います。

表8.20で最も着目したい点は、不偏分散の期待値が250となり、この値が母集団で求めた不偏分散値と一致していることです。すなわち、偶然に小さくなったり大きくなったりすることはありますが、平均的には母集団の値に近い値となることが期待できるということです。

9.1

標準化得点

本章では、基準の異なるデータについて、「どちらが良い／悪いのか」ということを測るための方法について考えます。早速ですが、簡単なデータ例を見ながら少しずつ考えていきましょう。それぞれ100点満点と1000点満点の、同レベル程度の英語のテストがあり、5名ずつが受験した結果を表9.1とします。

表9.1　2つのテストにおける、それぞれ5名の結果

	テスト1における点数
A	30
B	40
C	50
D	60
E	70

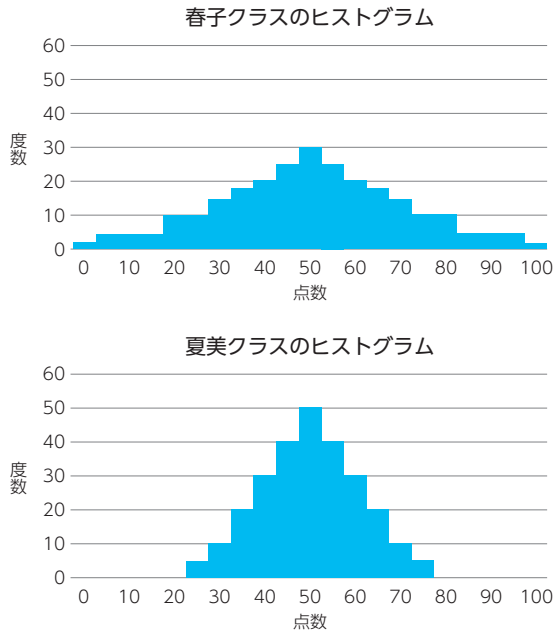
	テスト2における点数
F	300
G	400
H	500
I	600
J	700

テスト1で50点を取ったCさんと、テスト2で500点を取ったHさんでは、どちらの方が良い成績といえるでしょうか。10倍の点数を取ったHさんの方が、はるかに良い成績といえるでしょうか。直感的にもその判断は危険であることがわかるかと思います。その原因は、2つのテストでは基準が異なることにあります。

では、そもそも基準とは何でしょうか。別のいくつかの例で、基準について考えてみましょう。太郎君と次郎君の、学年の少し離れた2人の兄弟がいるとします。太郎君と次郎君は、それぞれの学校での100点満点の算数のテストで、太郎君は50点、次郎君は70点を取りました。このとき、どちらを褒めるべきでしょうか。これ以上に情報がないのであれば、得点率で7割を取っている次郎君を褒めるしかないかもしれません。もしここに、太郎君のクラスの平均点は30点、次郎君のクラスの平均点は80点という情報が加わったらどうでしょうか。今度は太郎君の方が褒められるかもしれません。この判断は、平均値というものが、善し悪しを判断する1つの大きな基準と考えられていることに起因します。

では、次の例を考えてみます。春子さんと夏美さんはそれぞれ別のテストで70点を取りました。両方のテストともに、平均点は50点であったとします。この2人は同等程度の成績の良さといえるでしょうか。今度はヒストグラムで考えてみましょう。図9.1は2人のクラスでのテスト結果を示すヒストグラムであり、どちらも受験者数はほぼ同等です。

図9.1　2人のクラスのヒストグラム



どちらも平均点である50点を中心に、両サイドに左右対称に散らばっていますが、散らばり具合が異なります。上側の春子さんのクラスの方が大きく散らばっていて、山が低く平べったい形です。逆に、下側の夏美さんクラスのヒストグラムは散らばりが小さく、山が高くなっていることが確認できます。

10.2

散布図

本節では2つのデータを比べるうえで、相関関係があるのか、特にその正負や強さを確認するうえで非常に有効なグラフである**散布図**について紹介します。ここではまず、表10.5に示す、5名分の「国語の点数」と「算数の点数」についてのデータで考えてみます。

表10.5 国語の点数と算数の点数

名前	国語の点数	算数の点数
A	70	80
B	60	90
C	50	50
D	90	90
E	30	40

散布図を作成するためには、2種類のデータについて、左側と下側に目盛を振ります。

図10.1 散布図の目盛

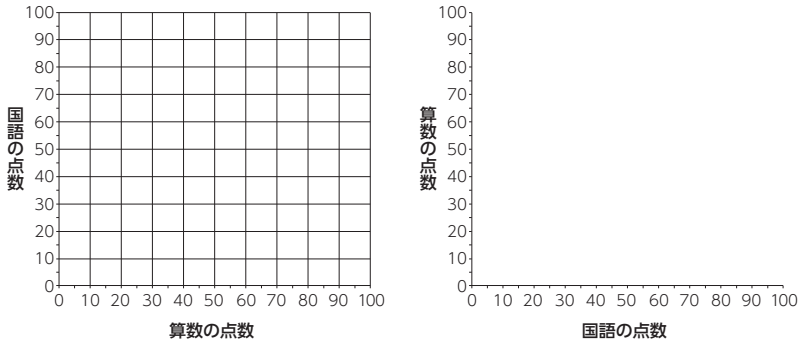


図10.1は、散布図の目盛の振り方の例であり、外観は少し異なりますが2つとも正しい目盛の振り方となっています。「国語の点数」を左側の目盛としても良いですし、「算数の点数」を左側の目盛としても構いません。また目盛の補助線を引いても引かなくても構いません。今回は図10.1の左側の外観（「国語の点数」が左目盛、補助線あり）をもとにして、散布図の作成を説明します。

外観を決めたら、次に**プロット**という作業を行います。プロットとは、グラフ上に各データの位置を示す点を打っていく作業になります。表10.5での1人目のデータ値は、国語が70点、算数が80です。このデータをプロットすると図10.2のようになります。同様に、表10.5の残り4名のデータ値をプロットして完成した散布図が図10.3となります。

図10.2 1人目のデータのプロット

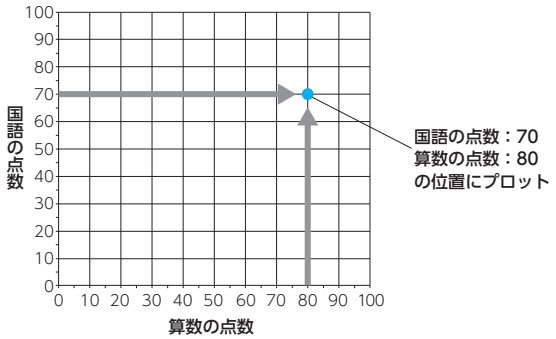
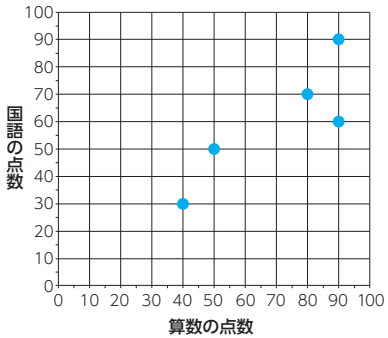


図10.3 国語と算数の点数の散布図



付録

Excelを利用した実践

分析ツールの利用

まず分析をより容易に行うためのExcelのアドイン機能である**分析ツール**の利用方法について紹介します。ただし、分析ツールを利用した具体的な分析手法については、それぞれの学習項目の中で説明しますので、ここでは分析ツールを利用するための設定方法のみを説明します。分析ツールはExcelを標準インストールした状態では、機能としては存在するものの、メニューには表示されていません。分析ツールをExcelのメニューに表示させてみましょう。

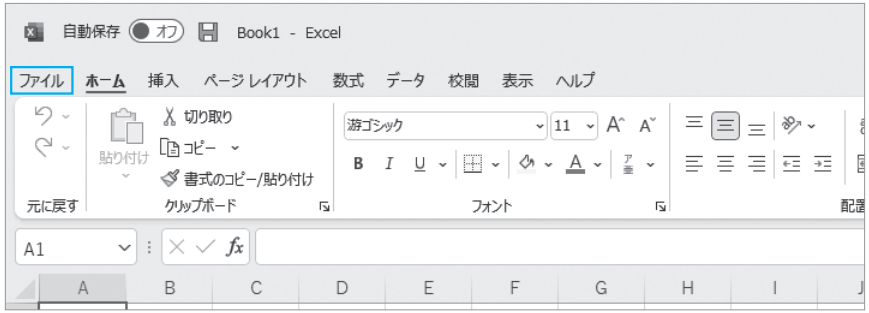
Microsoftのサイトにも同様の手順が紹介されています。macOS版での設定方法についても記載がありますので、必要に応じてこちらもご参照ください。

https://support.microsoft.com/ja-jp/office/excel-で分析ツールを読み込む-6a63e598-cd6d-42e3-9317-6b40ba1a66b4

操作手順

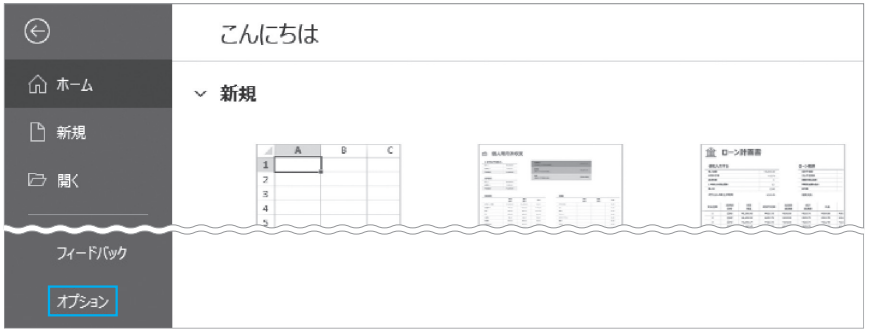
1 Excelのリボン（タブ）から、「ファイル」を選択します（図A.1）。

図A.1 「ファイル」メニューの選択



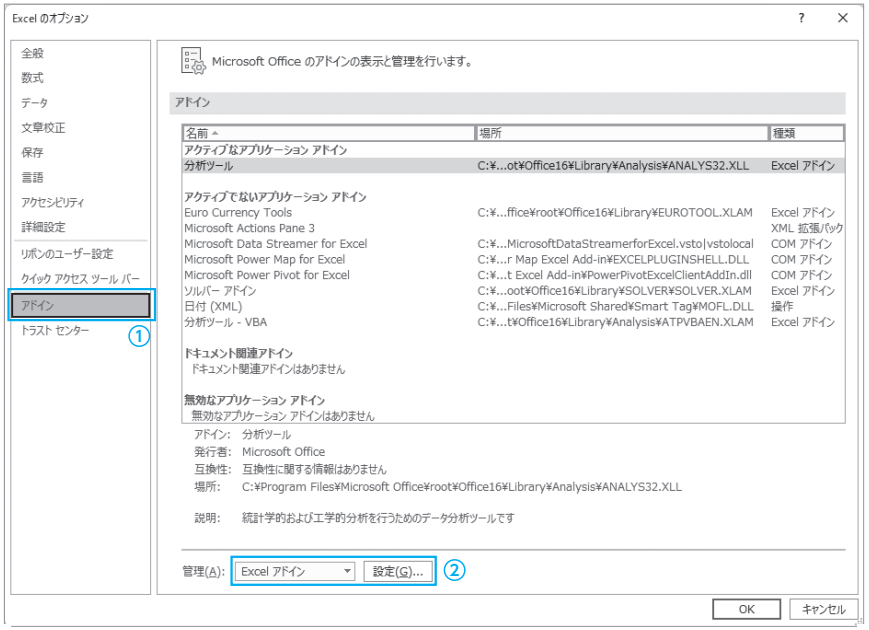
2 左側メニューの下の方にある、「オプション」を選択します（図A.2）。

図A.2 「オプション」の選択



3 大きめのダイアログが開くので、①左側メニューの「アドイン」を選択し、②右側メニューの下の方に表示される項目が「Excelアドイン」になっていることを確認し、「設定」ボタンをクリックします（図A.3）。

図A.3 「Excelアドイン」の「設定」を選択



度数分布表の作成（関数の利用）

292ページでは、分析ツールを利用して度数分布表を作成しましたが、ここでは関数を利用して作成する方法を紹介します。分析ツールでの作成と比較し、多少手間が多くなりますが、累積度数や相対度数なども求めることができ、階級幅が異なっていても対応できるなど、少し手間をかける利点はあります。

元データとしては200名分の身長を利用し、5cm刻みで階級を区切り、相対度数、累積度数、累積相対度数をそれぞれ求めていく手順を解説します。

操作手順

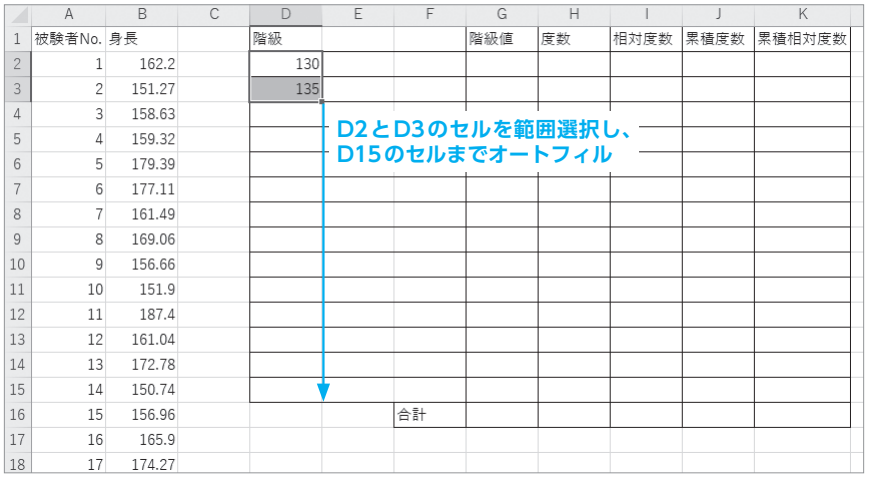
1 データ「06.xlsx」のExcelファイルの「度数分布表」シートを開きます。今回は説明や目標をわかりやすくするために、度数分布表の枠組みを先に用意しています。シートを確認できたら、D2のセルに130を、D3のセルに135をそれぞれ入力します（図A.39）。

図A.39 階級の開始データの入力

	A	B	C	D	E	F	G	H	I	J	K
1	被験者No.	身長		階級			階級値	度数	相対度数	累積度数	累積相対度数
2	1	162.2									
3	2	151.27									
4	3	158.63									
5	4	159.32									
6	5	179.39									
7	6	177.11									
8	7	161.49									
9	8	169.06									
10	9	156.66									
11	10	151.9									
12	11	187.4									
13	12	161.04									
14	13	172.78									
15	14	150.74									
16	15	156.96				合計					
17	16	165.9									

2 D2とD3のセルを範囲選択し、D15のセルまで下方向にオートフィルを行います（図A.40）。D15のセルには195の値が入ります。

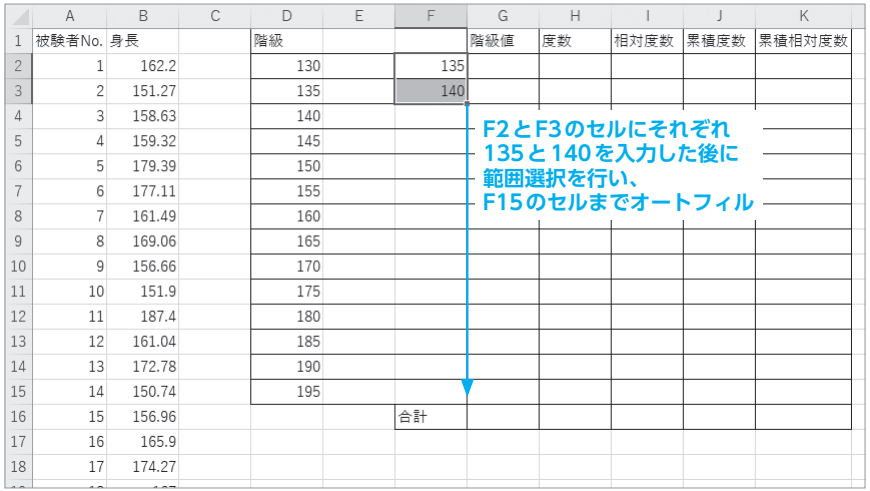
図A.40 階級データのオートフィル



	A	B	C	D	E	F	G	H	I	J	K
1	被験者No.	身長		階級			階級値	度数	相対度数	累積度数	累積相対度数
2	1	162.2		130							
3	2	151.27		135							
4	3	158.63									
5	4	159.32									
6	5	179.39									
7	6	177.11									
8	7	161.49									
9	8	169.06									
10	9	156.66									
11	10	151.9									
12	11	187.4									
13	12	161.04									
14	13	172.78									
15	14	150.74									
16	15	156.96				合計					
17	16	165.9									
18	17	174.27									

3 同様にF2とF3のセルに135と140をそれぞれ入力し、F15のセルまで下方向にオートフィルを行います（図A.41）。D15のセルには200の値が入ります。

図A.41 終点の階級データのオートフィル



	A	B	C	D	E	F	G	H	I	J	K
1	被験者No.	身長		階級			階級値	度数	相対度数	累積度数	累積相対度数
2	1	162.2		130		135					
3	2	151.27		135		140					
4	3	158.63		140							
5	4	159.32		145							
6	5	179.39		150							
7	6	177.11		155							
8	7	161.49		160							
9	8	169.06		165							
10	9	156.66		170							
11	10	151.9		175							
12	11	187.4		180							
13	12	161.04		185							
14	13	172.78		190							
15	14	150.74		195							
16	15	156.96				合計					
17	16	165.9									
18	17	174.27									