目次

Ⅰ 統計の役割

1. 統計とは

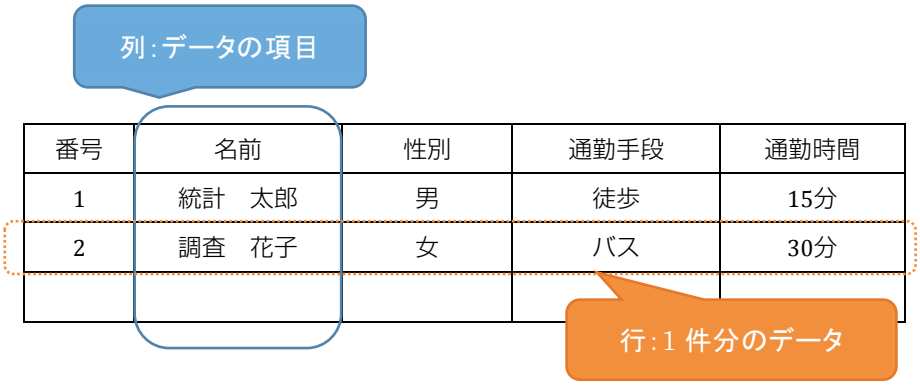
統計とは、「統べて計る」ことをいいます。「統べる」は、「多くのものを一つにまとめる」という意味で、「計る」は「ある基準をもとにして物の度合いを調べる」という意味です。つまり、「多くのものを一つにまとめ、ある基準をもとにして物の度合いを調べる」ということになります。必要な情報の全体を捉え、まとめる方法を統計は提供してくれます。

私たちの身の回りにはたくさんの情報があふれています。その中で私たちは必要な情報を取捨選択し、それをもとに物事を決めています。しかし、多くの場合、私たちは情報の一部しか入手することができなかったり、不確かな情報しか手元にない中で決断を求められます。そのようなときに統計は役に立つたくさんのアイデアを提供してくれます。データとその背景にある文脈を読み、科学的に不確かな問題を解決してくれるのが統計です。

2. 統計分析

統計分析の手法は、様々な分野で利用されています。たとえば、患者さんに薬を処方する場合、医師は直感や雰囲気ではなく、それまでの経験とデータをもとに判断します。降水確率も過去のデータをもとに計算された根拠のある数値です。統計分析の手法を用いれば、客観的で科学的な仮説を立てたり結論を導いたりすることができます。

では、統計分析はどのようにして行うのでしょう。
まず、統計分析にはデータが必要です。データは次のような表の形式に整理できます。



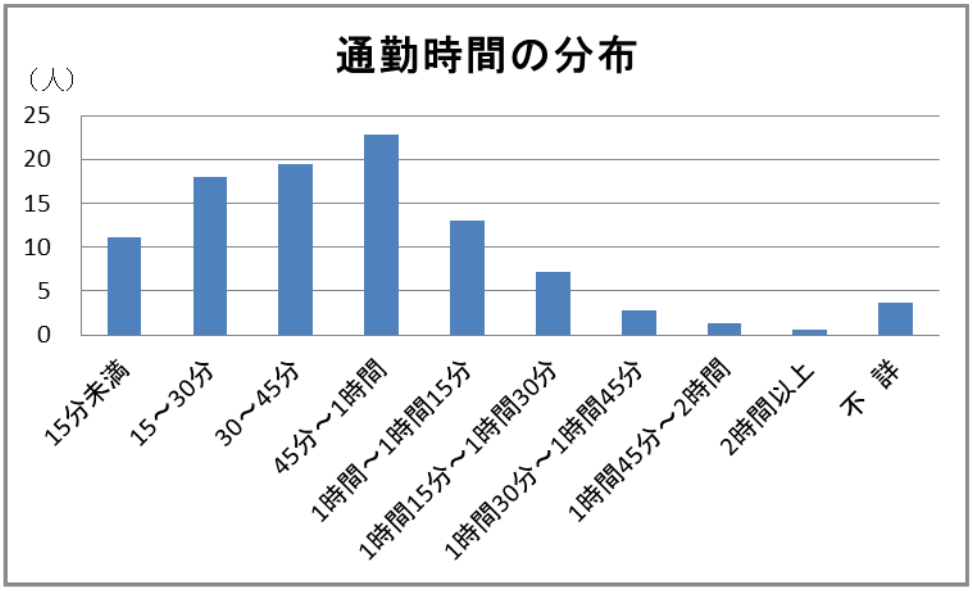
番号	名前	性別	通勤手段	通勤時間
1	統計　太郎	男	徒歩	15分
2	調査　花子	女	バス	30分

ここでは、一人分を1行のデータとし、列ごとに同じ項目のデータが並んでいます。表に整理されたデータの個々の値を統計学では観測値といいます。たとえば、太郎さんの通勤時間の観測値は15分です。

個々のデータはいろいろな値をとります。このデータのばらつきの様子や度合いをグラフや数値で表現することにより、データの全体像を把握することができるため、ばらつきを調べることは大変重要です。

統計では、ばらつきの様子を「分布」といいます。たとえば、下の図はある会社で働く人100人をランダムに選んで調べた通勤時間の分布を表すグラフです。最も多いのは45分～1時間の人で全体の2割強を占めています。多くの人は通勤時間が1時間以内で、7割の人がその範囲に入ります。この結果をもとに、この会社に勤める人全体についても7割程度の人の通勤時間が1時間以内であると推測することもできます。

別の会社や別の地域に勤める人との違いを分析するためには、ばらつきの様子からそれぞれの全体像を捉えて比較することが大切です。このデータのばらつきについては次の章で詳しく説明します。



統計分析は大きく記述統計と推測統計に分類されます。

記述統計：手元にあるデータの持つ情報を明らかにするための分析
推測統計：全体の一部である手元にあるデータから全体を推測する分析

初級編では、まず記述統計の考え方を学んでいきましょう。

3. 基本的なグラフ

グラフはデータ全体の傾向や特徴を見やすくするための道具です。集めたデータを目的に合った形に整理し、グラフ化することにより、状況を的確に捉えやすくなり、分析の糸口が見つかったり、また、他の人にデータの内容を分かりやすく伝えることができます。

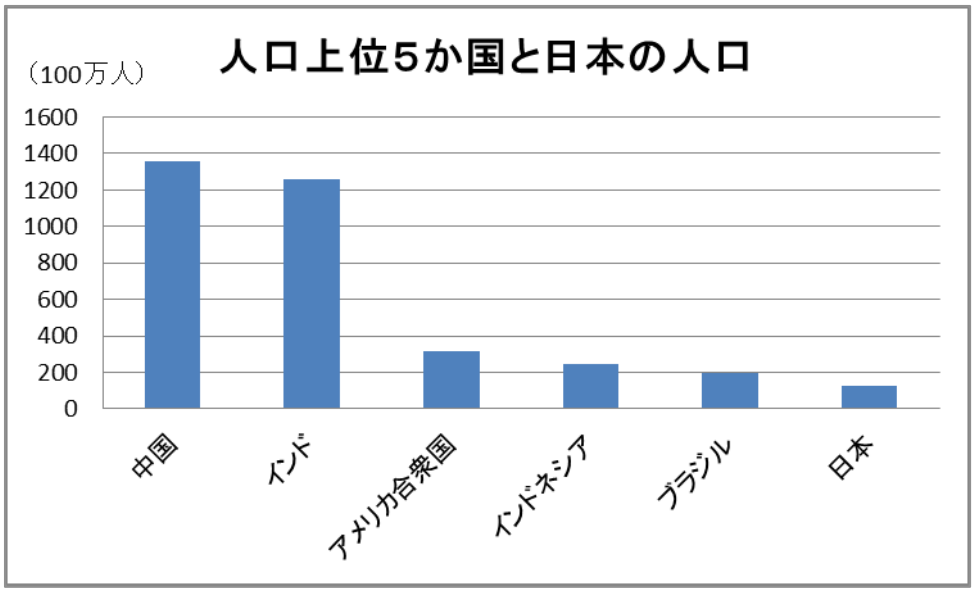
グラフにはいろいろな種類があり、データの内容によって適切なグラフを選択することが重要です。

代表的なグラフの種類とその用途	
棒グラフ	数量の大小を比較する際に用いられる。 棒の高さが量を表している。
折れ線グラフ	数量の時間的な変化を表す際に用いられる。
複合グラフ	棒グラフと折れ線グラフを一つにまとめたグラフ。
円グラフ 帯グラフ	全体に対する割合を表す際に用いられる。

① 棒グラフ

棒グラフは、数量の大小を比較するのに適しています。棒の高さや長さが数量を表すため、比較が簡単にできます。たとえば、世界の国と日本の人口を比べるときは、表で数値を見比べるよりも、グラフにしたほうが、規模の違いが分かりやすく、人口第1位、2位の中国、インドと他の国では随分差があることが一目で分かります。

順位	国	人口(100万人)
1	中国	1,354
2	インド	1,258
3	アメリカ合衆国	316
4	インドネシア	245
5	ブラジル	198
10	日本	127

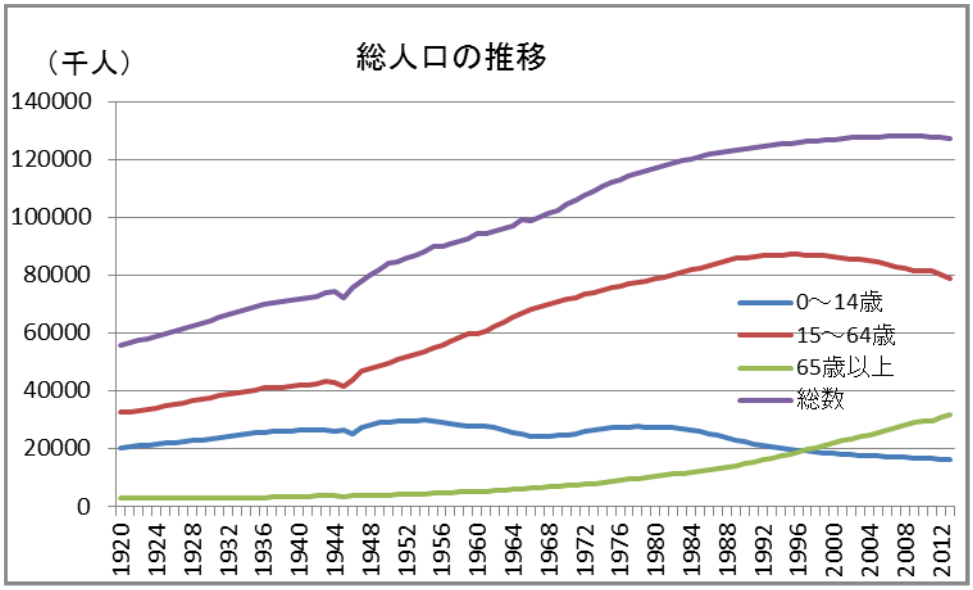


棒グラフには、棒を横向きにした横棒グラフ、1種類の値だけでなく、何種類かの値を同時にグラフ化した複数系列の棒グラフもあります。

② 折れ線グラフ

折れ線グラフは、時間とともに数量が変わる様子を表します。異なる時点で測定された値が、どのように変化するかを見ることができます。

下のグラフは日本の人口の推移を表しています。横軸が年(時間)を、縦軸が人数(数量)を示しています。日本の総人口は、2008年をピークに減少傾向であることが分かります。また、その内訳も同時に記しているため、15～64歳の生産年齢人口が減少し、65歳以上の老齢人口が増加傾向にあることが同時に読み取れます。

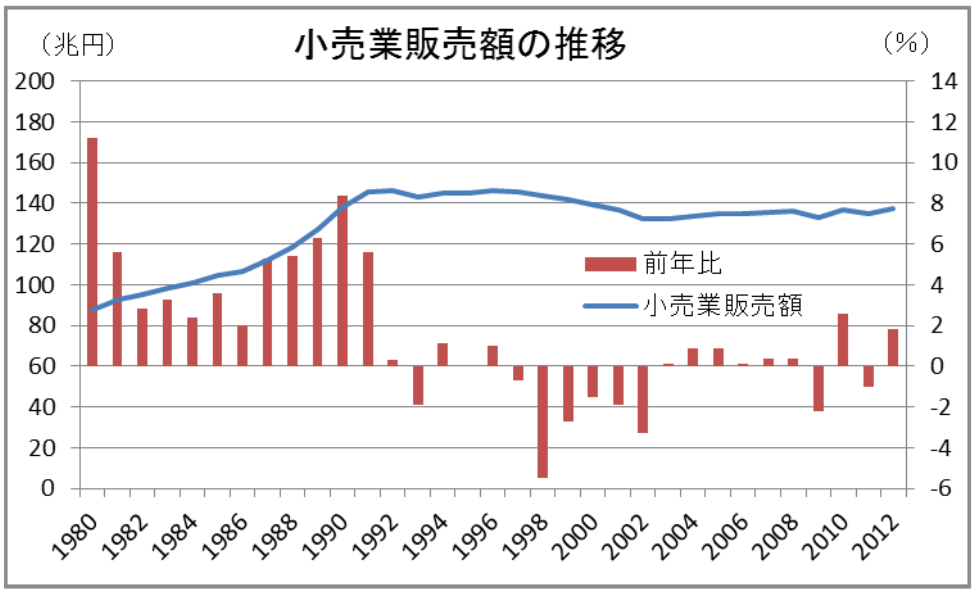


折れ線の傾き方で、変化の大きさを見ることができます。傾きが急な場合は大きく増加(減少)したことを意味し、緩やかな場合は変化が少ないことを表します。横軸の目盛が等間隔に設定されていないと、傾きで比較することができなくなるため、折れ線グラフの横軸は必ず目盛を等間隔に設定しましょう。

また、目盛間隔によって、同じデータでも結果が誇張されたりしますので、グラフを読む際には、グラフだけを見るのではなく、元の数値や単位などにも注意が必要です。

③ 複合グラフ

棒グラフと折れ線グラフを一つにまとめたグラフを複合グラフといいます。下のグラフは小売販売額を折れ線で、前年比を棒で表した複合グラフです。

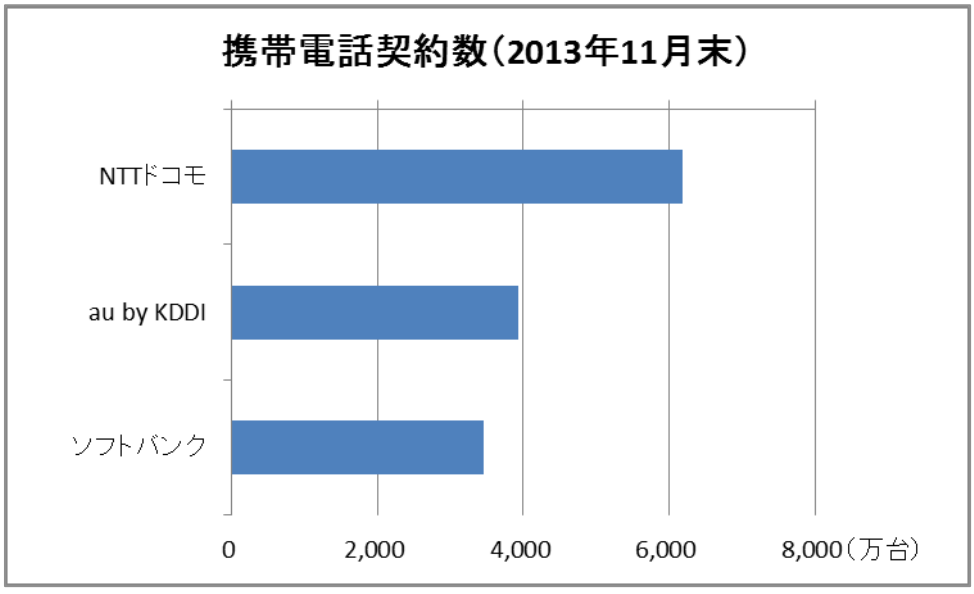


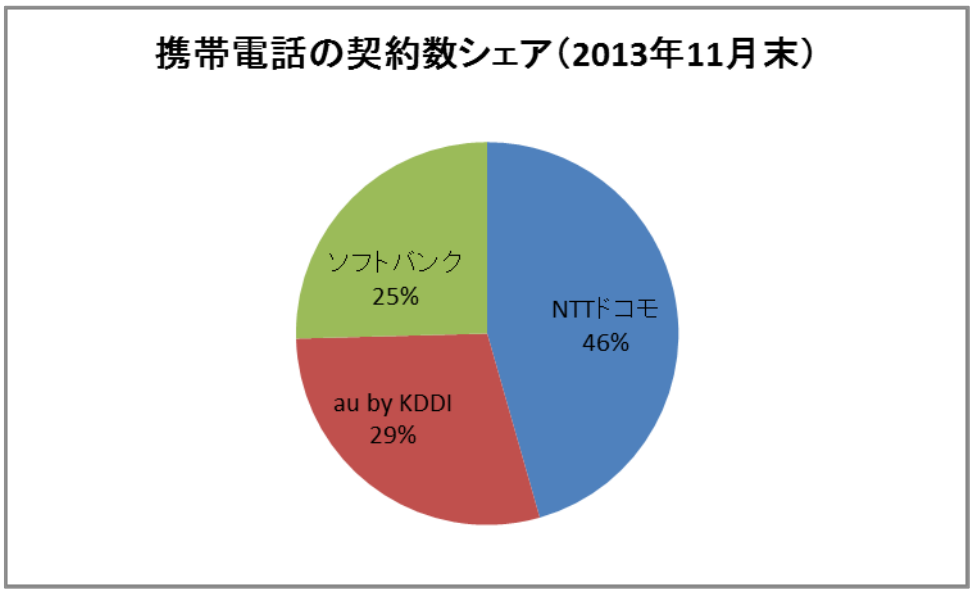
④ 円グラフ

円グラフは、全体に対する割合を視覚的に表現するグラフです。扇形の中心角の大きさで各カテゴリーの割合を表します。

下の横棒グラフは、携帯電話の2013年11月の契約台数を表したグラフです。NTTドコモがトップで他社とどのくらい違うのかが一目で分かります。このように契約数そのものを比較する場合には棒グラフが適しています。

同じデータを使って円グラフを描くと割合に焦点を当てた表現になります。契約台数そのものではなくシェア(市場占有率)を比較することができます。

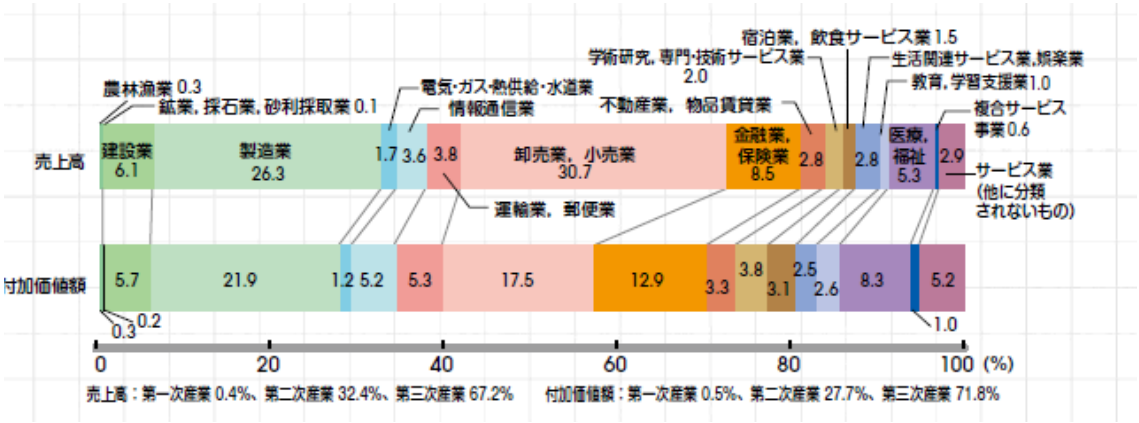




⑤ 帯グラフ

割合の表示には円グラフのほか帯グラフも使用されます。帯グラフでは全体を100%としたときのそれぞれの割合を帯の幅で示します。

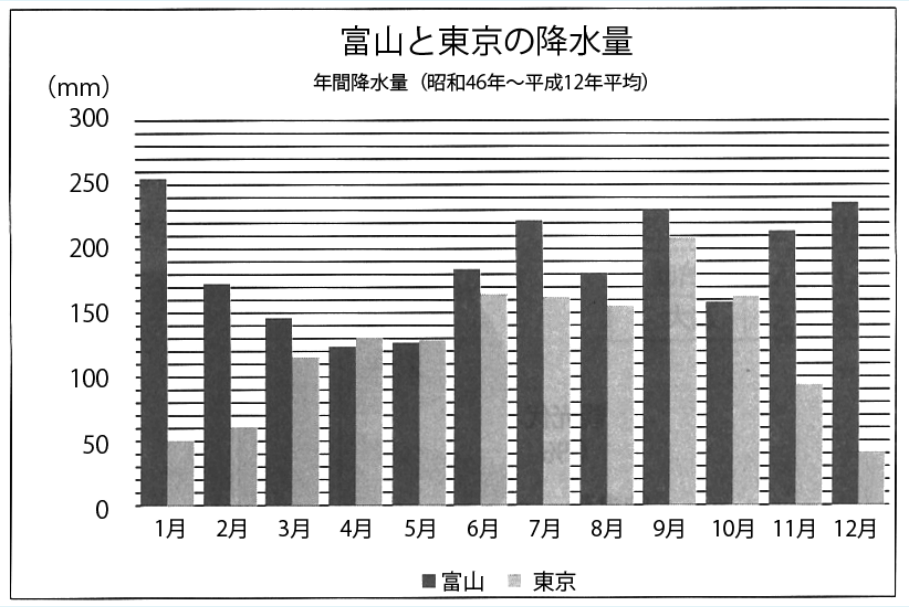
下のグラフは売上高と付加価値額の構成比を産業大分類別に比較したものです。たとえば、売上高では卸売業、小売業が30.7%と全体の3割強を占めていますが、付加価値額では17.5%と2割弱になっていることが分かります。このように、総数の異なる二つのデータを比較するには、割合を計算し、帯グラフを描くと良いでしょう。



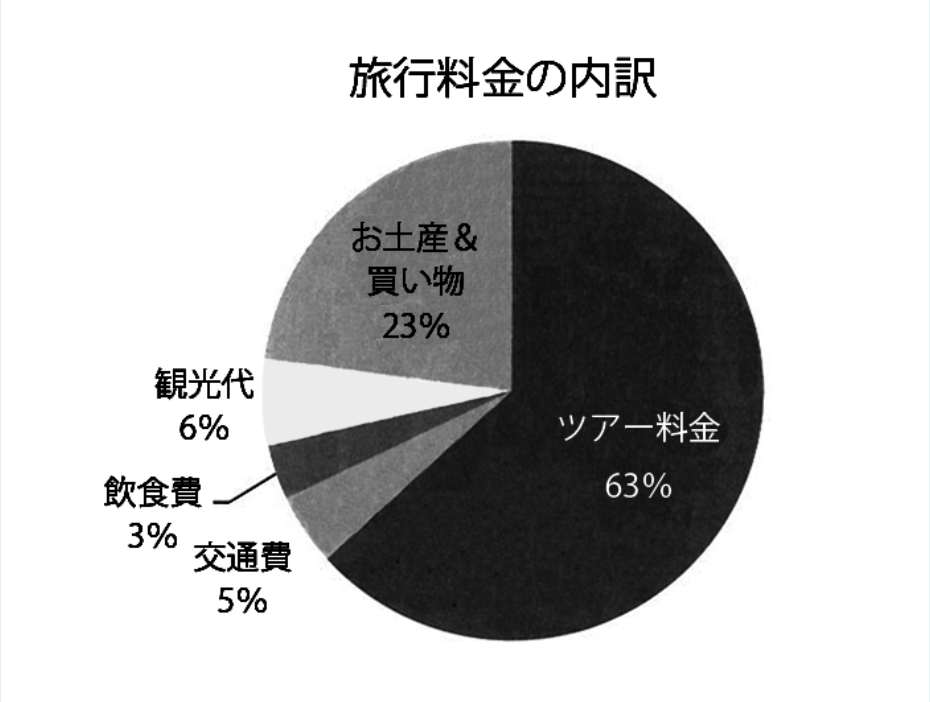
練習問題

（解答は P.40 です）

問１ 次のグラフから、雨や雪がたくさん降った月と降らなかった月の差が大きいのは富山と東京のどちらですか。



問２ ある旅行会社が旅行の費用について調査した結果、費用の内訳は次のグラフに示す通りでした。この結果を参考にあなたが20万円の予算で旅行を計画するとしたら、ツアー料金はいくらぐらいにすればいいでしょう。



Ⅱ データのばらつきの見方

統計では、データのばらつきの様子や度合いを見ることが大変重要です。分析の対象となるデータがどんな値をとり、どのように、どの程度ばらついているかを知ることにより、データの全体的な姿を捉えることができます。このデータのばらつきの様子を「分布」と言います。

個々のデータは、それぞれ異なる値をとります。たとえば、クラスの全員の身長が同じということはありえません。では、この一人ひとり異なる身長のデータを使って「男子のほうが女子より身長が高い」かどうかを確かめるためにはどうしたら良いでしょうか。

男子全体の様子を知るには、男子のデータの分布を調べることが必要です。分布とは、ある範囲にデータがどのくらいあるかを示すものです。性別や血液型のように、分類が決められているデータの場合は、その分類ごとに、A型が何人、B型が何人のように、データの数を数えます。年齢や身長のように数値そのもののデータの場合には、値あるいは値の範囲を決めて、たとえば160～165cmに何人というように、その範囲に含まれるデータの数を数えます。

このように、データのばらつき具合を表すためにはデータの種類によってその数え方を考える必要があります。また、データの種類によって、分析の手法も異なります。そこで、ここでは、データの種類とその分析手法について学びましょう。

1. データの種類

データには、分類(カテゴリー)の違いによって記録される質的データと、大きさや量など、数量として記録される量的データの2種類があります。これら、データの種類によって、データの数え方や分析の手法が異なります。

質的データ:分類や種類(カテゴリー)の違いが記録されるデータ
(例)　国籍、性別、血液型、職業など
量的データ:大きさや量などの数量が記録されるデータ
(例)　身長、体重、気温、年齢など

2. 質的データの分析

質的データとは、そのデータの性質、分類や種類の違いを表すデータです。質的データを分かりやすく表現するには、データの種類ごとにその数を数え、度数分布表を作成します。それを棒グラフなどにするとデータの全体像や分布を捉えることができます。

① 度数分布表

たとえば、ある会社の医務室の利用記録をもとに、利用状況の全体を捉えるにはどうし

たらよいでしょう。

日付	時間	所属	名前	理由
1月4日	10:35	総務課	統計　太郎	ねんざ
1月6日	9:45	営業第二課	調査　花子	頭痛
1月6日	12:30	技術開発部	分析　和夫	発熱
⋮	⋮	⋮	⋮	⋮

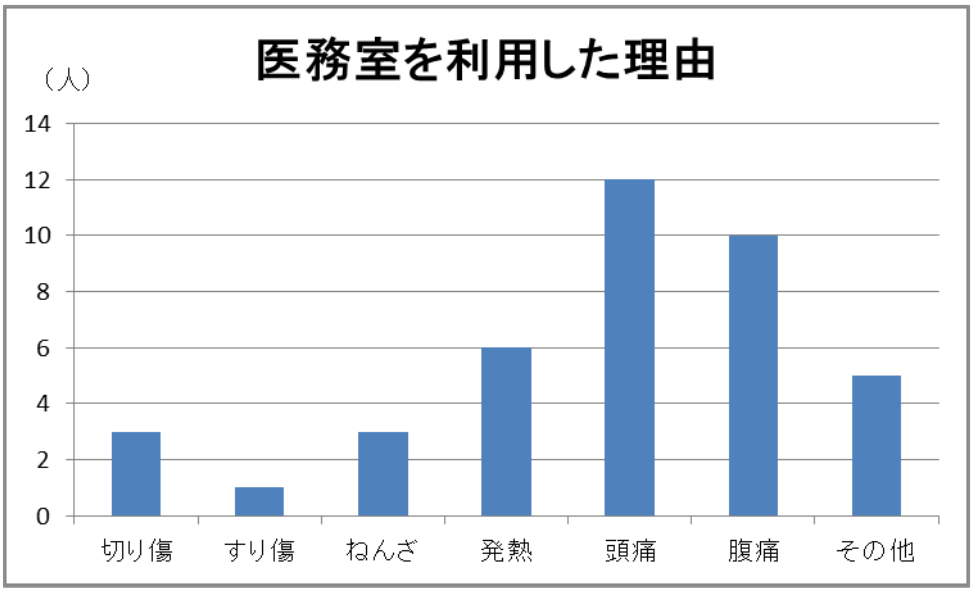
利用状況を順に記録した上の表をみても、データが並んでいるだけで全体の様子を捉えることはできません。このような場合には医務室を利用した理由ごとに利用者の人数を数え、表にまとめます。このような表を度数分布表といいます。

さらに、この表で集計した度数の値を小さい順や大きい順に並べ変えて棒グラフなどで表すと全体の様子がよりわかりやすくなります。

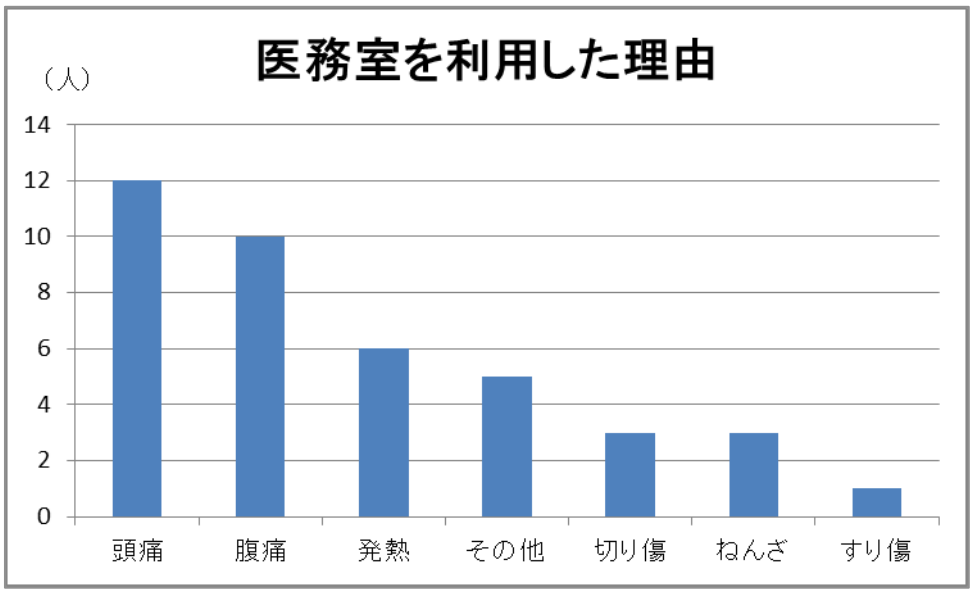
医務室を利用した理由	度数(人数)	相対度数
切り傷	3	3/40=0.075
すり傷	1	1/40=0.025
ねんざ	3	3/40=0.075
発熱	6	6/40=0.150
頭痛	12	12/40=0.300
腹痛	10	10/40=0.250
その他	5	5/40=0.125
計	40	1.000

② 質的データの度数を示す棒グラフ

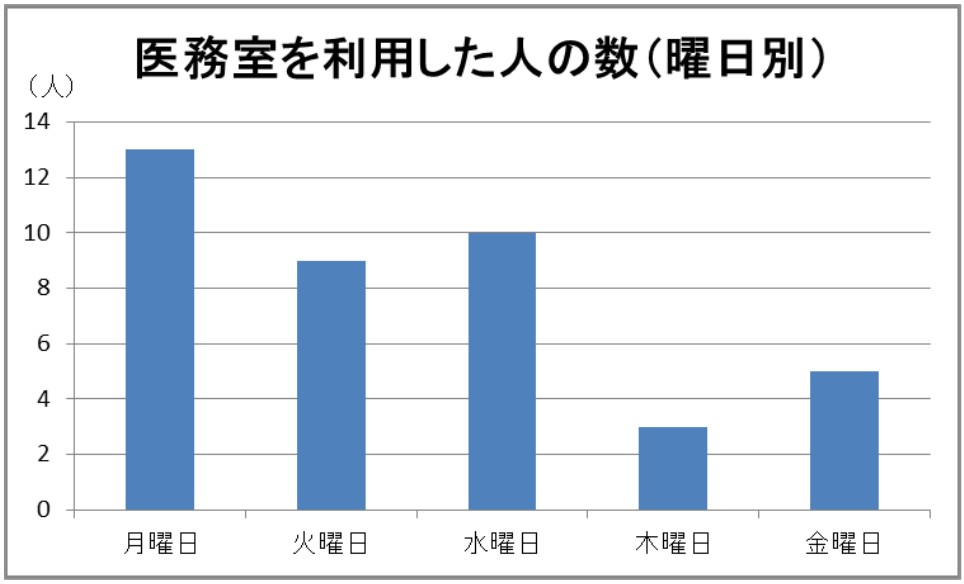
上の度数分布表を棒グラフに表すと全体の様子を分かりやすく表すことができます。



度数の大きい順や小さい順に並べ替えて、見やすさを工夫してみましょう。



ただし、下の図のように医務室の利用者の数を曜日別に集計した場合には、横軸の曜日には順番があるため、度数の大小で並べ替えることはできません。



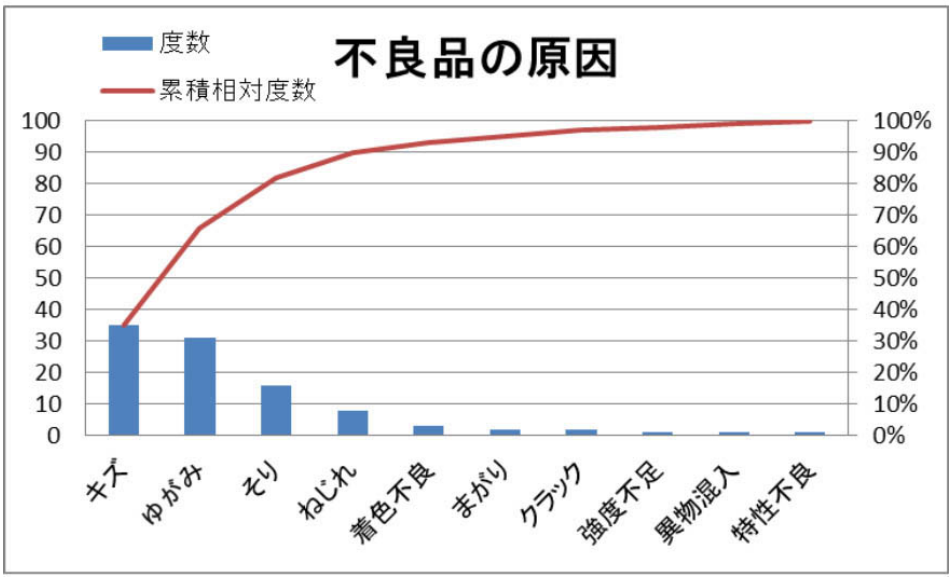
③ パレート図

下の表はある製品の製造工程で不良品が発生した原因をまとめた度数分布表です。不良品の原因ごとに件数が集計されています。累積相対度数とは、相対度数を足し合わせたものです。

不良品の原因	件数(度数)	累積度数	相対度数	累積相対度数
キズ	35	35	35.0%	35.0%
ゆがみ	31	66	31.0%	66.0%
そり	16	82	16.0%	82.0%
ねじれ	8	90	8.0%	90.0%
着色不良	3	93	3.0%	93.0%
まがり	2	95	2.0%	95.0%
クラック	2	97	2.0%	97.0%
強度不足	1	98	1.0%	98.0%
異物混入	1	99	1.0%	99.0%
特性不良	1	100	1.0%	100.0%
合計	100		100.0%	

このような質的データの度数分布表をもとに度数を表す棒グラフと累積相対度数を表す折れ線グラフを合わせて表したグラフをパレート図といいます。この場合、横軸のカテゴリー(不良品の原因)が順序を考慮しなくて良い性質のものなので、棒グラフは度数の大きい順に並べます。これによって注目すべきデータがどれか、それらのデータが全体の何割を占めているかを特定することができます。

このグラフをみると、キズとゆがみで不良品の原因の7割近くを占めていることが分かります。



④ クロス集計表

複数の項目を組み合わせて度数を集計した表をクロス集計表といいます。調査などの結果は、クロス集計表を作成することによって、データの全体像を把握しやすくなります。

たとえば、過去1年間にどんなスポーツをしたかについて、アンケート調査をした場合に、男性と女性では好みが違うだろうと予測し、男女別に結果を集計したとします。その場合、「性別」と「スポーツの種類」のクロス集計表となります。

また、クロス集計表は目的に合わせて項目を変更することにより、様々な分析が可能になります。上記のアンケートの結果も性別だけではなく、年齢階級別や世帯類型に集計することにより違う視点で結果を読むことが可能になります。

練習問題 （解答は P.40 です）

問 1 質的データはどれですか。すべて選びなさい。

① 一番好きな朝食のメニュー
② 携帯電話に毎月かかる額
③ 統計の授業に対する評価（とても良い／良い／良くない）
④ 車のナンバー
⑤ あなたが好きな車の作られた国
⑥ お気に入りのテレビ番組
⑦ あるレストランのサービスに対する満足度（1～5）
・非常に満足している＝1
・満足している＝2
・どちらとも言えない＝3
・満足していない＝4
・非常に満足していない＝5

⑧　1か月の食費
⑨　電話の市外局番

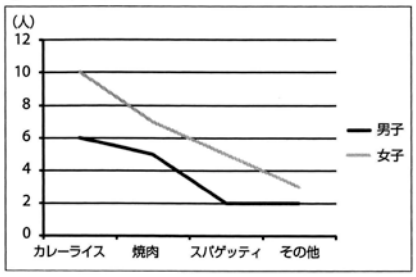
問2　かすみさんのクラスは、男子15人と女子25人の40人です。このクラスで一番好きな食べ物を調査したところ、次のような結果が得られました。

一番好きな食べ物

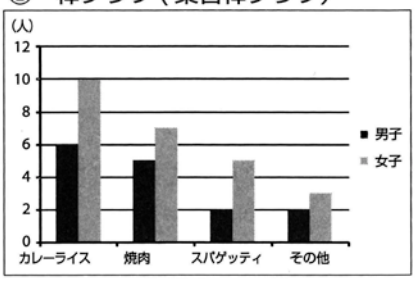
性別	カレーライス	焼肉	スパゲッティ	その他	合計（人）
男子	6	5	2	2	15
女子	10	7	5	3	25
合計（人）	16	12	7	5	40

この表をもとに一番好きな食べ物に男子と女子の間で違いがあるかどうかを調べたいとき、次の①～④のグラフのうちで最も適切なものを一つ選びなさい。

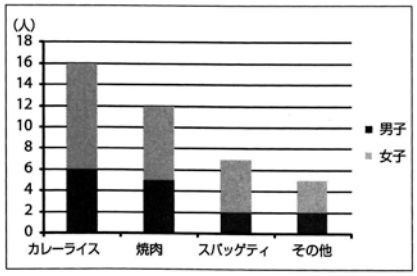
① 折れ線グラフ



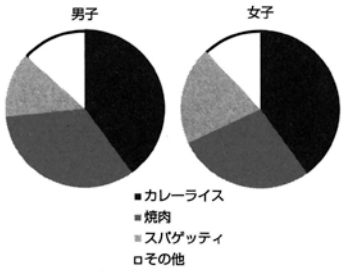
② 棒グラフ（集合棒グラフ）



③ 積み上げ棒グラフ



④ 円グラフ



3. 量的データの分析

量的データとは大きさや量などの数量が記録されるデータです。量的データのばらつきを分析するには、値あるいは値の範囲ごとに対象となるデータの数を数えて度数分布表を作成します。

量的データは、世帯員の数や事業所の従業員数など常に整数の値しかとらない量(離散データ)と身長や体重などの連続的に変化する量(連続データ)に分けられます。

① 度数分布表

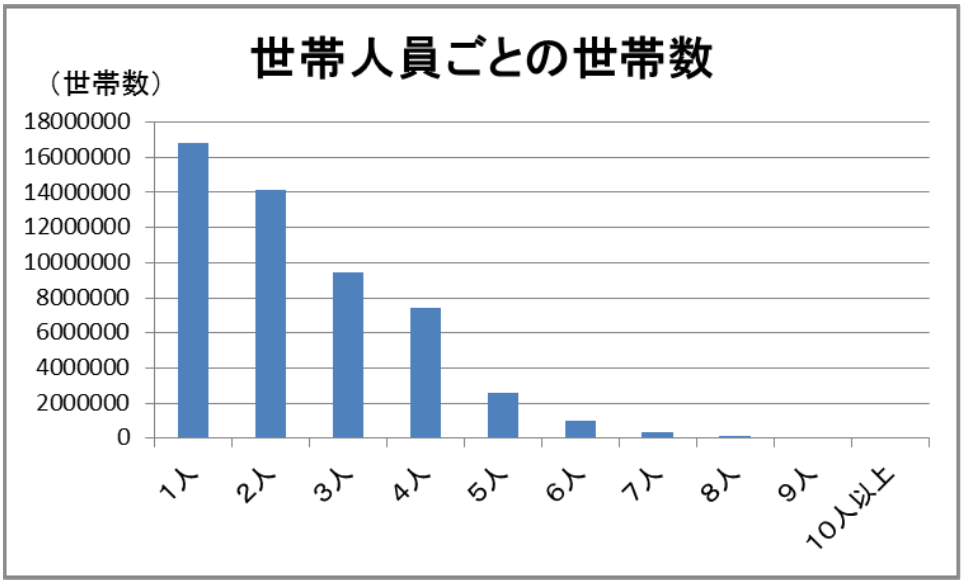
➢離散データの場合

離散データは、整数の値のみを記録したデータです。値の種類が少ない場合には、値ごとに度数分布表を作成します。値の種類が多い場合には、一定の範囲ごとに度数を集計しましょう。

次の表は、世帯人員別の世帯数を表しています。これは世帯数という量的データです。ただし、このデータは0.5世帯とか3.4世帯といった小数点以下の値をとりません。整数の値のみをとる離散型のデータです。

世帯人員	世帯数
1人	16,784,507
2人	14,125,840
3人	9,421,831
4人	7,460,339
5人	2,571,743
6人	984,751
7人	359,325
8人	100,655
9人	23,721
10人以上	9,595
合計	51,842,307

このような場合には、質的データの場合と同様に、世帯人員ごとに世帯数(度数)を集計し、グラフにすると全体を捉えやすくなります。



また、値の種類が多くなる場合は一定の範囲ごとに集計します。

たとえば、年間収入ごとの世帯数であれば、1万円単位で集計するのでは、区分が細かくなり過ぎ、全体としての分布の特徴が読み取りにくくなってしまいます。その場合、50万円間隔や100万円間隔に分類して集計すると、全体の特徴が読み取りやすくなります。

➢連続データの場合

連続データの場合にも、離散データと同様に度数を集計して全体の様子を捉えることができます。また、連続データの場合には、小数点以下の桁数の多い数値が含まれることもありますので、質的データや離散データとは異なり、個々の値でデータの度数を数えることは適切ではありません。そこで、値の範囲をいくつかのクラス(階級)に分けて、その中に入るデータの数を数えます。

例えば、体重データは小数点以下の桁数を含む連続データです。ここでは、中学1年生の体重データを男女別に10kgごとに集計します。

階級	度数(人)	
	男性	女性
10kg 未満	0	0
10kg 以上20kg 未満	0	0
20kg 以上30kg 未満	34	23
30kg 以上40kg 未満	375	350
40kg 以上50kg 未満	386	456
50kg 以上60kg 未満	144	136
60kg 以上70kg 未満	43	26
70kg 以上80kg 未満	13	6
80kg 以上90kg 未満	4	2
90kg 以上100kg 未満	1	0
100kg 以上	0	0

階級とは、量的データの値の範囲をクラス分けしたものです。

階級幅とは、各階級の上限と下限の差を示し、「10kg以上20kg未満」の階級では、階級幅は10kgとなります。

階級値とは階級の上限と下限の中央の値を示します。たとえば、「20kg以上30kg未満」の階級値は25kgです。

➢度数分布表をみるポイント

✧ データの中心を探す

データの最も集中する階級が一つの目安になります。中心が分布のほぼ真ん中に位置する場合や、値の小さい方に偏る場合、値の大きい方に偏る場合などがあります。分布の中心を示す指標については、次章で詳しく説明します。

✧ 全体の約半分を捉える

中心のおおよその位置が分かったら、そこを中心に全体の約半分になるデータの範囲を確認します。

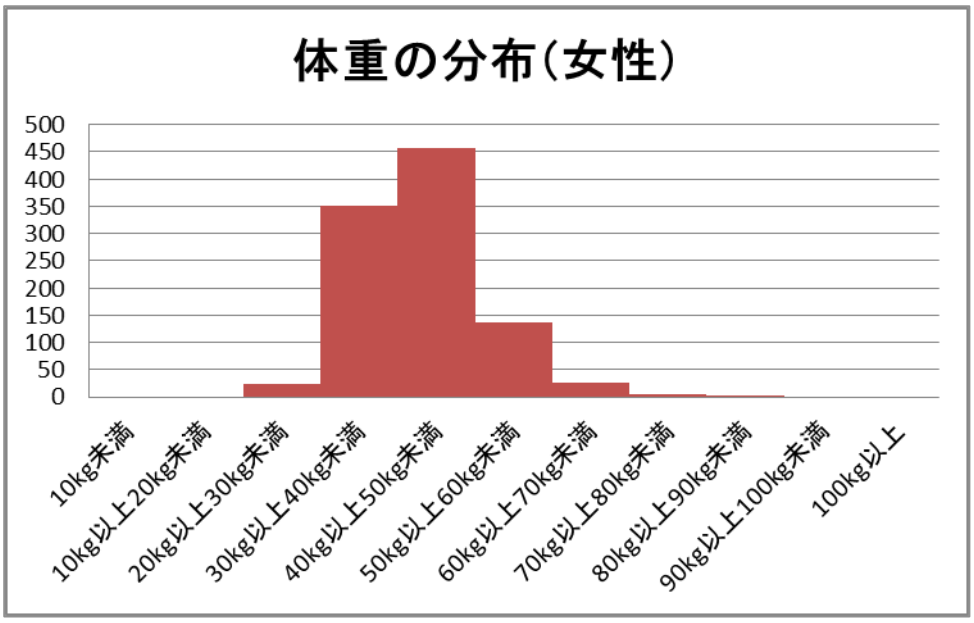
以上の二つのポイントから、上の表をみると、女性は40～50kgの人が最も多く、この階級に50%弱のデータが集中しています。

一方、男性は40～50kgの人が最も多いことは女性と変わりませんが、女性に比べて集中度が低く、右側(体重の重い方)に裾をひく分布となっています。

② ヒストグラム

ヒストグラムとは、連続型の量的データの度数分布表を柱の面積で表したグラフです。

ヒストグラムは棒グラフと異なり、横軸が必ず数値を示します。量のつながり(連続性)を表現するために、柱同士の間隔は空けないで詰めて描きます。



➢ヒストグラムをみるポイント

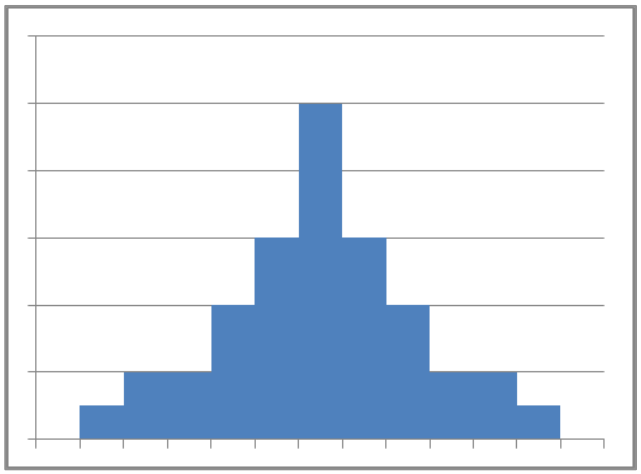
ヒストグラムの形状は、分布の特徴を表すとても重要な情報で山の形などにたとえて表現されます。データが集中している箇所を峰(ピーク)と呼びます。

ピークの数や左右対称かどうかなどを見ることによって、分布の特徴が分かります。

✧ 単峰性、左右対称

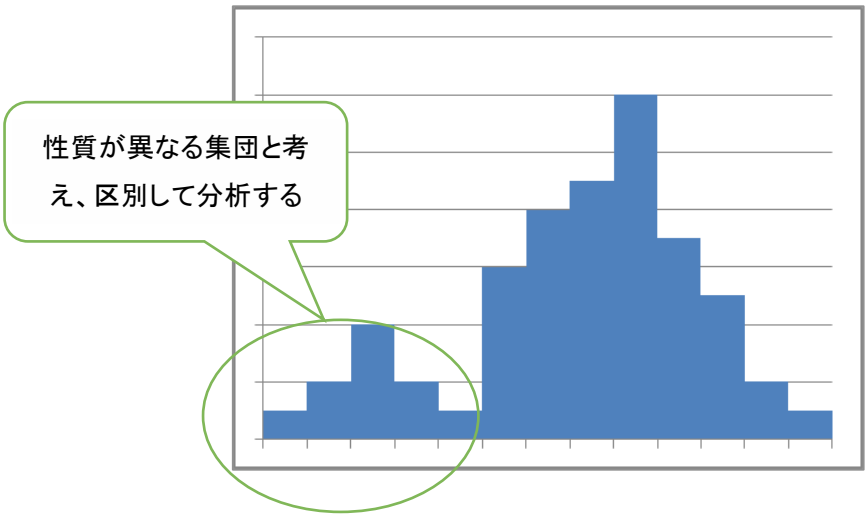
最も基本的なヒストグラムの形はピークが一つ(単峰性)で左右対称です。データのもとになる集団が同じ性質を持っていれば(同質であれば)、山の頂点を中心に左右対称の形状を示すことが多いです。

単峰性のヒストグラムにはとがった山型を示すものや、なだらかな山型を示すものがあります。鋭くとがった山型はばらつきの小さい分布を表し、なだらかな山型はばらつきの大きい分布を表します。



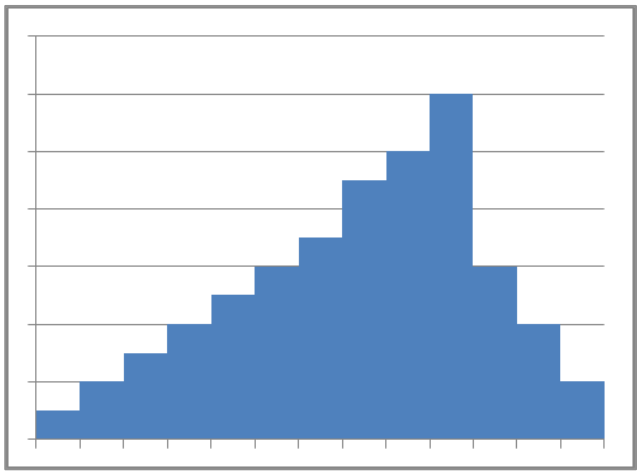
✧多峰性

ピークが2つ以上ある場合は、大人と子供などの異質な集団のデータが混在している可能性があります。その場合は、データを分けて分析するなどの工夫が必要になります。

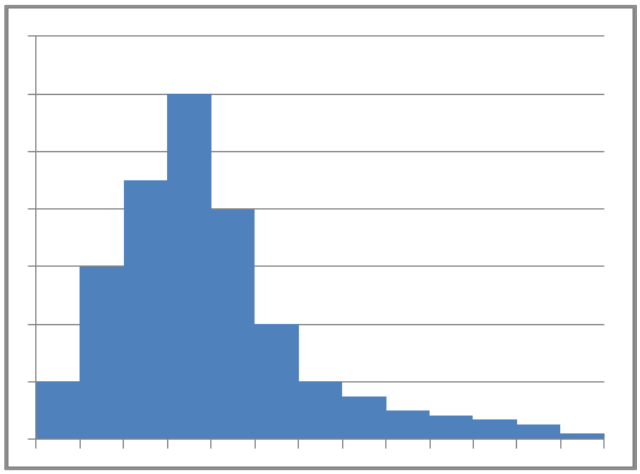


✧ 左右非対称

単峰性のヒストグラムであっても、ピークが右や左に偏り、片側に長く裾を引く場合があります。その原因はいくつかの可能性があり、裾の部分に他とは異なる偏ったデータが混在している場合や、分布そのものが偏った場合が考えられます。



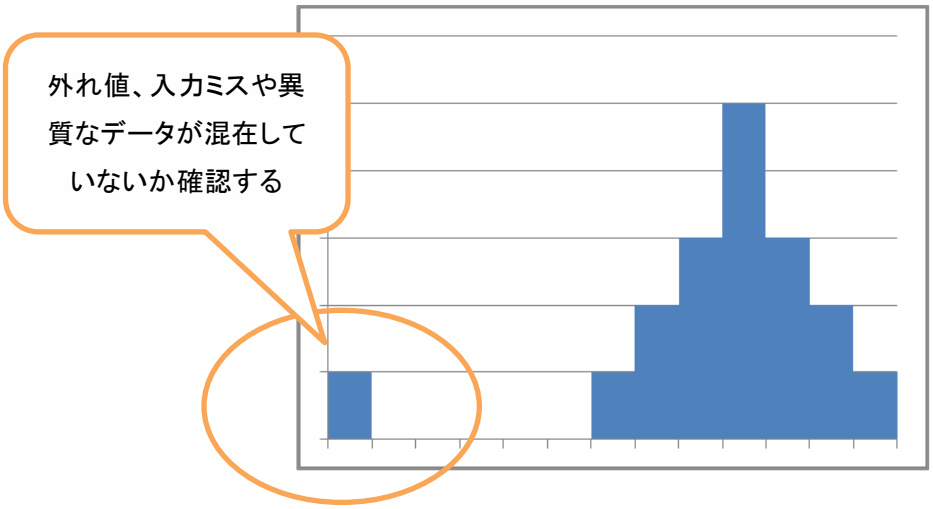
左にゆがんだ分布(左に裾の長い分布)



右にゆがんだ分布(右に裾の長い分布)

✧ 外れ値

集団の多くが示す値と離れたところにある少数のデータを外れ値といいます。外れ値のあるヒストグラムも異質なデータが混在している可能性があります。外れ値が存在する場合には、入力ミスや異質なデータが混在していないかの確認が必要です。



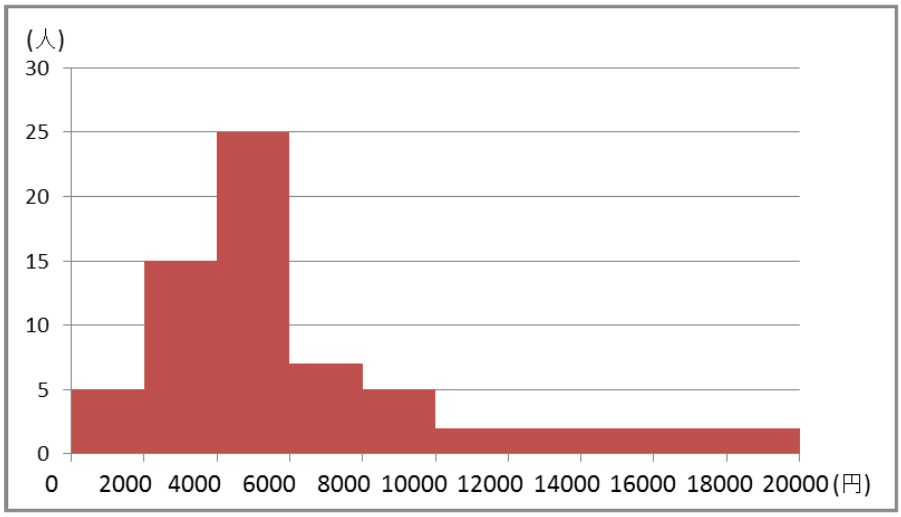
✧階級幅が等しくないヒストグラム

度数分布表の階級の幅が等間隔でない場合には、ヒストグラムの作成に注意が必要です。

たとえば、次のような度数分布表の場合、初めの五つの階級は、階級幅が2000円ですが、最後の階級だけは階級幅が10000円となっています。

小遣い(円)	度数
0円以上2000円未満	5
2000円以上4000円未満	15
4000円以上6000円未満	25
6000円以上8000円未満	7
8000円以上10000円未満	5
10000円以上20000円未満	10

階級幅が等間隔でない場合には、柱の面積が度数に対応するように高さを調整する必要があります。



③ 数値による分布の要約

➢分布の特徴を表す基本統計量

分布の特徴は、ヒストグラムを見るポイントと同様に次のような視点から読み取ります。

✧単峰か多峰か

✧中心の位置、ちらばりの大きさ

✧対称か非対称か

✧外れ値の有無

これらの特徴を数値で表現した指標を基本統計量といいます。ここでは、そのうち中心の位置とちらばりの大きさを見る指標について学びましょう。

➢分布の中心を表す指標

データ全体をある一つの値で代表させるとしたら、それはどのような値でしょう。このようなときに役に立つのが分布の中心を表す指標であり、代表的なものは、平均値(ミーン)、中央値(メジアン)、最頻値(モード)の三つです。

✧ 平均値(ミーン)

平均値は最も良く知られている代表値であり、その中でも特に算術平均と呼ばれる値がよく用いられています。算術平均はn個の観測値を足し上げて個数(n)で割ったものです。

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

✧ 中央値(メジアン)

中央値とは、n個の観測値を大きさの順に並べたときに、ちょうど真ん中に位置する値

で、中位数ともいいます。

nが偶数の場合には、 $\frac{n}{2}$ 番目のデータの値と $\frac{n}{2}+1$ 番目のデータの値の平均が中央値となります。

✧ 最頻値(モード)

最頻値とは、度数の最も大きい値のことをいいます。量的データの場合は、値の種類が多くなることがあるため、個別の値ではなく度数分布表で最も度数の大きい階級又は階級値を最頻値と呼びます。

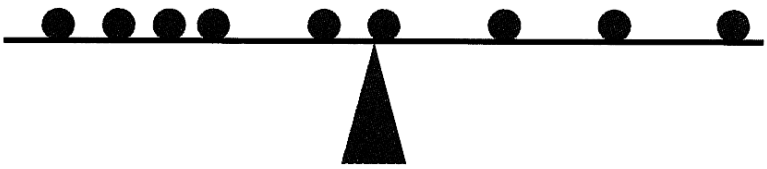
最頻値が複数存在する場合には、異質なデータが混在している可能性がありますので、注意が必要です。

➢ 代表値の特徴

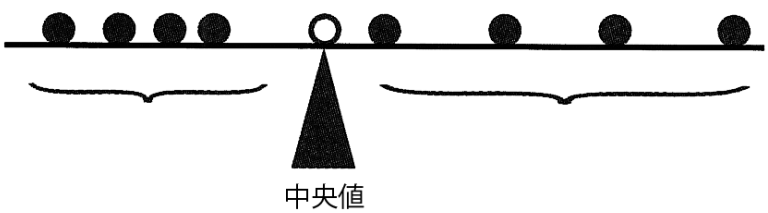
三つの代表値はそれぞれどういう特徴を持っているでしょうか。

平均値は、データの値の場所に同じ重さのおもりを置いたときに、ちょうどつりあう場所であり、データの重心であると言い換えることができます。したがって、平均値は往々にして、実際には観測されていない値となります。

中央値は、その値より大きいデータの数と小さいデータの数が同じ個数になる場所です。



平均値は、データがバランスよく釣り合う位置を示す



中央値は、データの個数を半分に分ける位置を示す

もし、データが単峰性で分布が完全に左右対称の場合には、この三つの代表値は完全に一致します。したがって、ほぼ左右対称の分布の場合にはどの代表値を用い

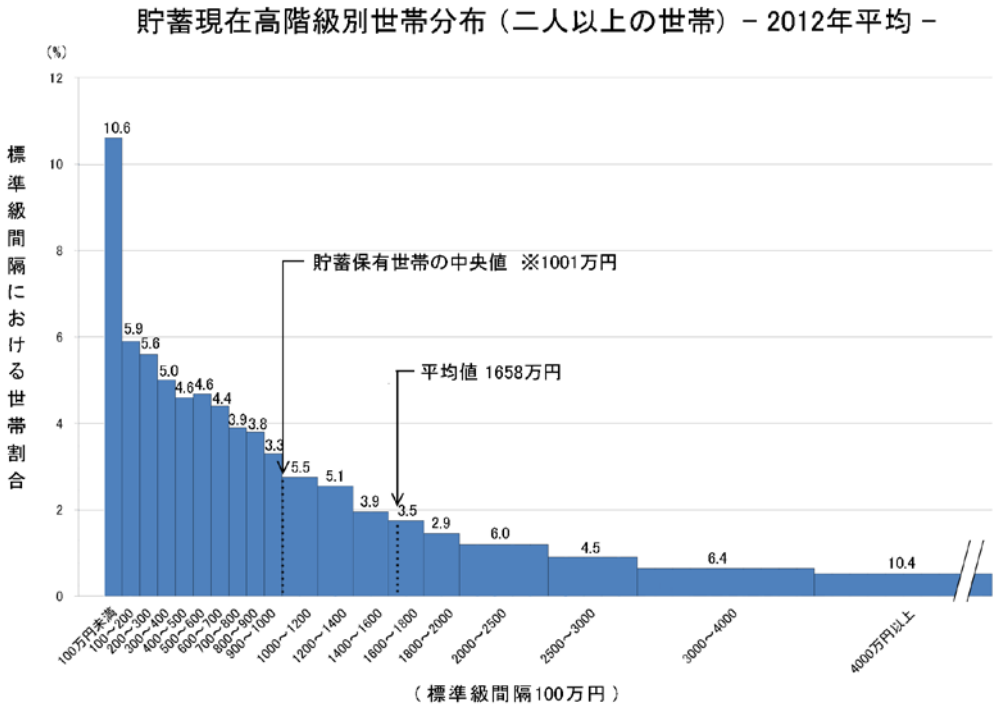
てもそれほど違いはありません。

しかし、左右対称ではなく、片側に外れたデータが存在すると、平均値はその値に引きずられ、外れ値のある方向に偏った値となります。分布が右にゆがんだ分布(右に裾が長い分布)では、一般的に平均、中央値、最頻値の順に大きくなります。

一般に外れ値がある場合やゆがんだ分布の場合には、分布の端の方の極端な値の影響を受けないため、中央値で判断するほうが適切です。

最頻値は、階級の取り方によって値が変わります。また、多峰性の分布では、有効な代表値とはいえません。

この関係を実際の例で見てみましょう。二人以上の世帯の貯蓄現在高を階級別に見た世帯数の相対度数分布のヒストグラムをみると、右に裾の長い、右にゆがんだ分布になっていることが分かります。この場合は、異質なデータが混在しているのではなく、分布そのものがもともとゆがんでいます。一般に収入や貯蓄額などのデータは、中心から外れた高い収入を得ている人や多くの貯蓄を持っている人がいるため、このような分布になります。ここでは平均値は1658万円、中央値は1001万円、最頻値は100万円未満となっています。



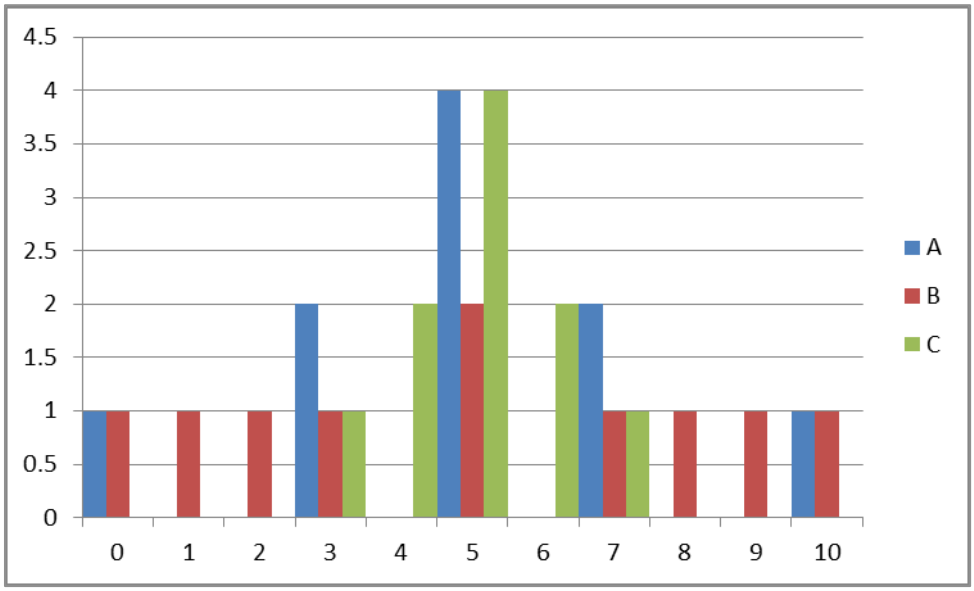
※ 貯蓄（負債）保有世帯の中央値とは、貯蓄（負債）現在高が「0」の世帯を除いた世帯を除いた世帯を貯蓄（負債）現在高の低い方から順番に並べたときに、ちょうど中央に位置する世帯の貯蓄現在高を言います。

➢ ばらつきを表す指標

データ数(n)が同じnの三つのデータA、B、Cが下図のようであったとき、これらの平均、中央値、最頻値はいずれも5であり、同じ値となっています。しかし、分布の形状は異なっています。A、BとCを比べると、Cは分布の幅(範囲)が狭く、固まっ

た分布となっています。AとBは分布の範囲は同じですが、山のとがり方が異なっています。

このように、代表値だけではデータ全体の様子を把握することはできません。中心の位置に加えて分布のばらつき(散らばり)を調べることが、データ全体を知る上で重要です。



✧ 範囲(レンジ)

分布のばらつきを表す指標のうち、最も単純なものは範囲(レンジ)と呼ばれ、分布の存在する範囲を表しています。範囲とは、データの最大値と最小値の差のことです。上の図の場合であれば、AとBの範囲は10、Cの範囲は4となります。

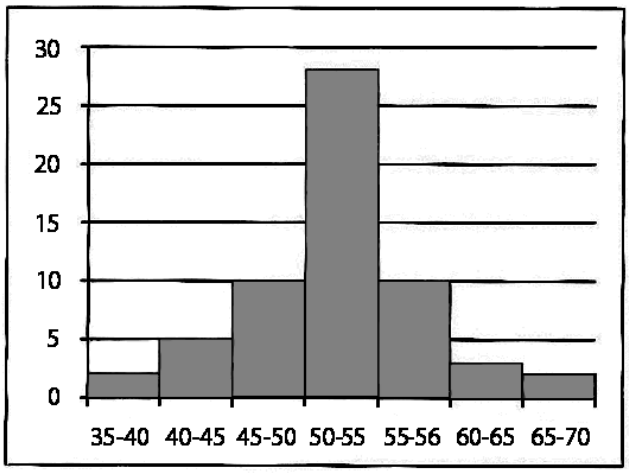
分布のばらつきを表す指標には、範囲のほか、四分位数、分散、標準偏差などがあり、これらの指標と合わせて検討することで、分布の形状を知ることができます。これらの指標については、中級編で説明します。

練習問題

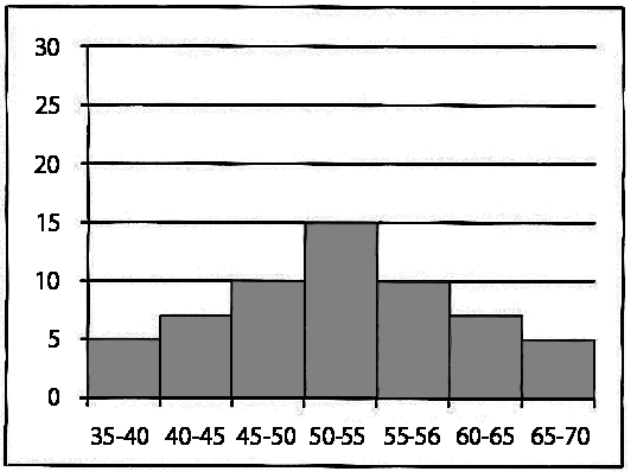
（解答は P.40 です）

問１　次のAとBのヒストグラムについて正しい記述はどれですか。下の1～5のうちからすべて選びなさい。

A



B



① AとBの平均は（ほぼ）等しい
② AとBの中央値は（ほぼ）等しい
③ AとBの最頻値は（ほぼ）等しい
④ AとBの範囲は（ほぼ）等しい
⑤ AとBのちらばりの大きさは等しい

問2　クリスティーナは1日にどのぐらい水を飲むかを、200人にたずねました。結果は次の表に示す通りです。

飲んだ水の量（リットル）	人数
0　より大　0.5以下	8
0.5　より大　1以下	27
1　より大　1.5以下	45
1.5　より大　2以下	50
2　より大　2.5以下	39
2.5　より大　3以下	21
3　より大　3.5以下	7
3.5　より大　4以下	3
合計	200

(1)　最も度数の多い階級はどれですか。

(2)　平均値を求めなさい。

(3)　累積度数を求めなさい。

(4)　医師は、1日に少なくとも1.8リットルの水を飲むように勧めています。この200人のうち水を飲む量が足りていない人は何%ですか。

【出典：IGCSE数学　(2009Specimen Paper4)】

Ⅲ 時系列データの基本的な見方

1. 時系列データ

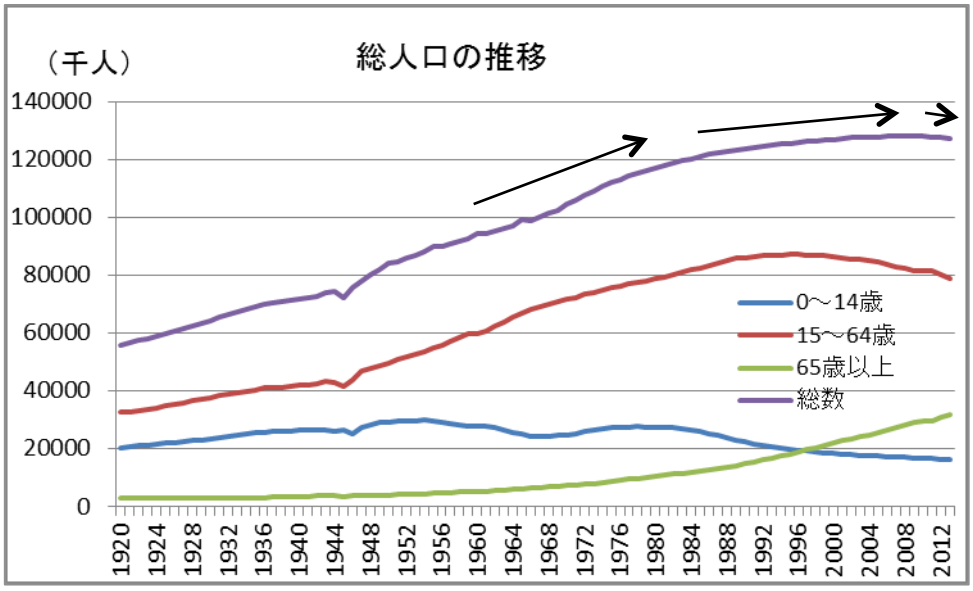
時系列データとは、毎月の気温や月ごとの売上高など、時間に沿って観測されたデータのことをいいます。データは一般に、年、四半期、月ごとに記録され、その期間が等間隔である必要があります。また、時間的な変化を分析するためのデータであることから、各時点で同一の内容・性質のものである必要があります。

時系列データは横軸に時点、縦軸に対象となる変数をとった折れ線グラフで表すと、時間に沿って、左から右にデータがどのように変化するかが見やすくなります。

下のグラフは日本の人口の推移を時系列で表しています。横軸が年(時間)を、縦軸が人数(数量)を示しています。日本の総人口は、1920年から2008年まで、増加していましたが、2008年をピークに最近は減少傾向であることが分かります。

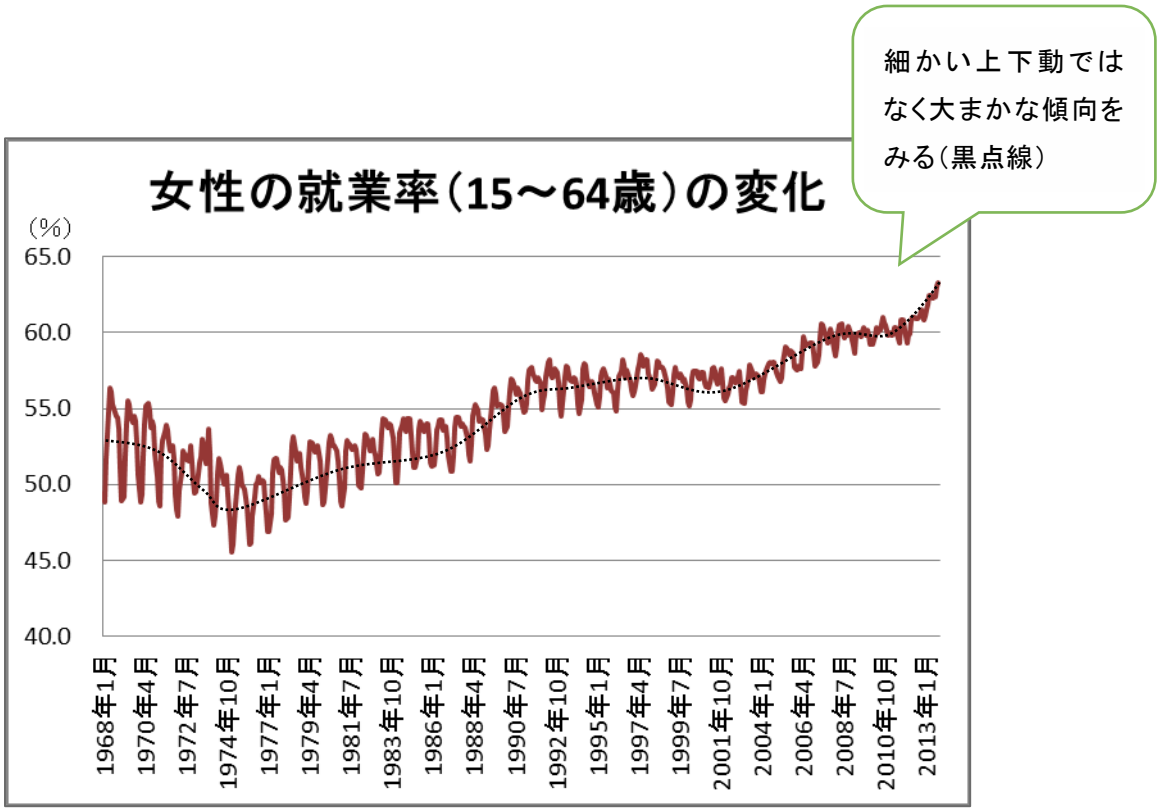
時系列グラフの傾きは、その期間の変化の大きさを表しています。1980年頃を境に人口は増加しているものの以前と比べて、その増加の大きさは小さくなっていることが、傾きが小さくなったことから分かります。

時間の間隔が異なるものを同じ幅で表してしまうと、傾きの意味が違ってしまうため、横軸の間隔は時間軸に沿うように表す必要があります。



時系列データは、①傾向変動(長期変動)、②循環変動(景気変動)、③季節変動、④不規則変動からなります。時系列データの変化を見るときは、短期的な上下の細かい変動ではなく、上昇傾向又は下降傾向、横ばいなど、長期的な傾向変動(トレンド)を見るようにします。

下の図は女性の就業率を月次で表したグラフです。1年ごとに繰り返す規則的な動き(季節変動)がありますが、長期的な傾向としては、1975年頃までは下落傾向、その後は上昇傾向にあり、2013年頃からその傾きが急になっていることが分かります。



2. 移動平均

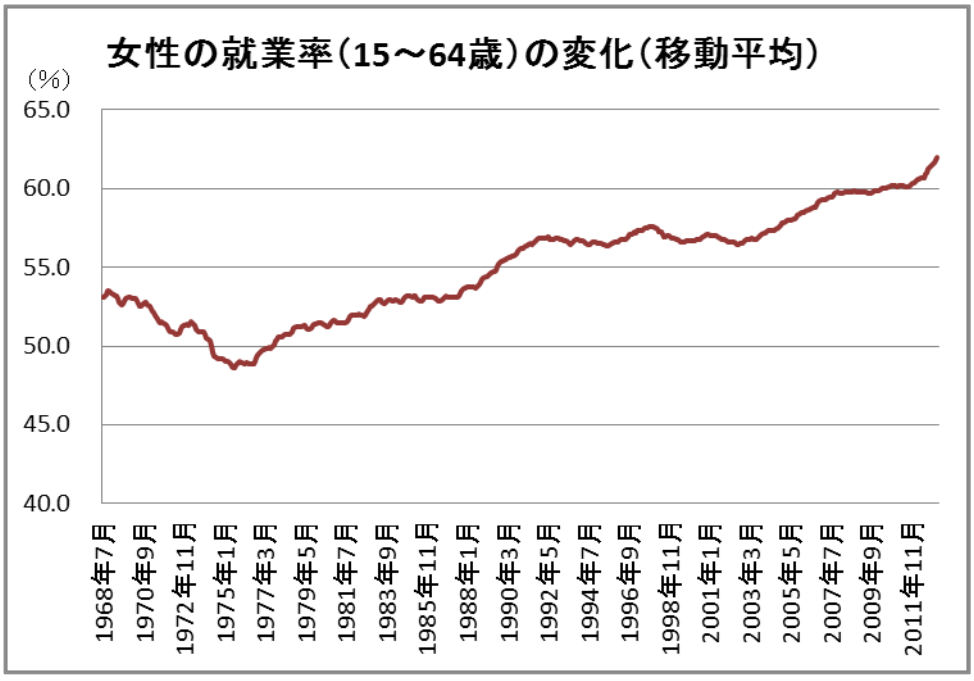
時系列データには、長期的な傾向(傾向変動)や季節的な変化(季節変動)の動きに加えて、その時点ごとに不規則な値の変化(不規則変動)が含まれています。この不規則な値の変化が大きい場合、元のデータから傾向を読み取ることが難しくなります。このような場合、傾向を読みやすくするため、一定の期間ごとにずらしながら平均をとる移動平均と呼ばれる手法が使われます。

下の表は、女性の就業率の13か月移動平均を計算している例です。

下のグラフは、上のグラフを13か月移動平均で描いた例です。季節的な変動や短期的な不規則が打ち消され、傾向変動が見やすくなっていることが分かります。

年月	女性の就業率	移動平均
1968年1月	49.5	
1968年2月	48.9	
1968年3月	51.8	
1968年4月	54.8	
1968年5月	56.4	
1968年6月	56.1	
1968年7月	55.3	53.1
1968年8月	55.0	53.1
1968年9月	54.5	
1968年10月	54.4	
1968年11月	53.6	
1968年12月	51.0	
1969年1月	48.9	
1969年2月	49.2	

1968年1月～1969年1月の平均（13か月移動平均）



3. 指数・増減率

① 指数

時系列データにおける指数とは、基準時点の値を100とし、相対値で表したものです。もとの時系列データを指数に変換することで、基準時点に対しての変化の大きさを読むこ

とが容易になります。

指数の計算式は、

$$比較時点t\text{の指数} = \frac{比較時点t\text{の値}}{基準時点t_0} \times 100$$

となります。

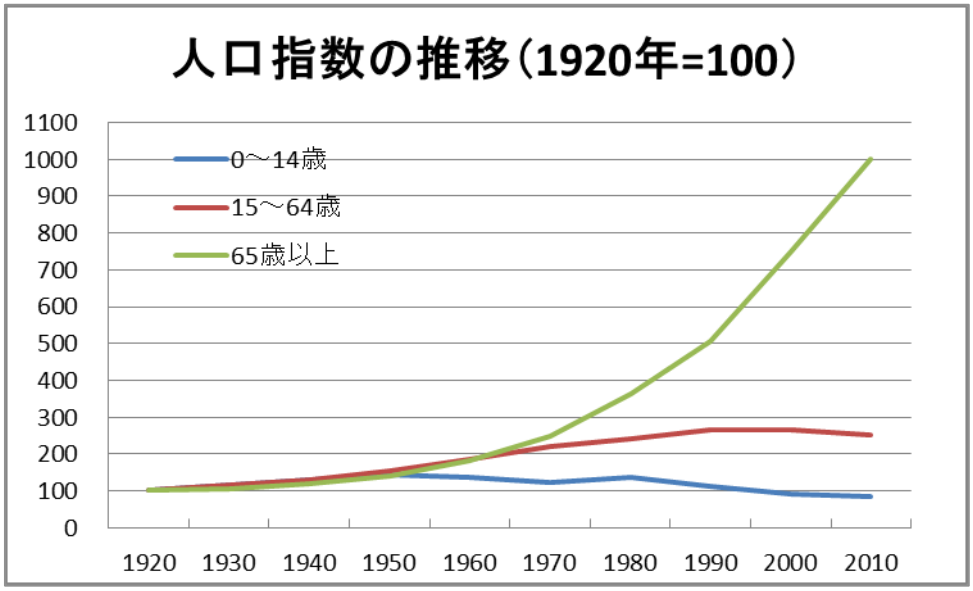
下の表は、1920年の人口を基準に指数化したものです。2010年の総人口は1920年の2.29倍になったことが一目で分かります。

また、指数は複数の時系列データの変化の大きさを比べるときに便利な指標です。

原数値をみると、65歳以上の人口は、15～64歳の人口に比べ、数が少ないため、増えていることは分かりますが、一見するとその変化は生産年齢人口の変化に比べて小さいように見えます。しかし、相対的な変化の大きさに着目すると、2010年で1920年の10倍強と圧倒的な勢いで伸びていることが分かります。

ただし、指数での変化は、飽くまでも基準時点の大きさに対する相対的な変化であり、決して増加数の比較ではないことに注意が必要です。

年	人口(千人)				指数(1920年=100)			
	15歳未満	15～64歳	65歳以上	総数	15歳未満	15～64歳	65歳以上	総数
1920	20416	32605	2941	55963	100.0	100.0	100.0	100.0
1930	23579	37807	3064	64450	115.5	116.0	104.2	115.2
1940	26383	42096	3454	71933	129.2	129.1	117.4	128.5
1950	29430	49661	4109	84115	144.2	152.3	139.7	150.3
1960	28067	60002	5350	94302	137.5	184.0	181.9	168.5
1970	24823	71566	7331	104665	121.6	219.5	249.3	187.0
1980	27524	78884	10653	117060	134.8	241.9	362.2	209.2
1990	22544	86140	14928	123611	110.4	264.2	507.6	220.9
2000	18505	86380	22041	126926	90.6	264.9	749.4	226.8
2010	16839	81735	29484	128057	82.5	250.7	1002.5	228.8



② 増加(減少)率

時系列データの変化の大きさを見るための指標が、増加率(減少率)です。基準時点からの変化量を基準時点の値で除して求めます。

$$\text{比較時点}t\text{の増加(減少)率} = \frac{\text{比較時点}t\text{の値} - \text{基準時点}t_0\text{の値}}{\text{基準時点}t_0\text{の値}} \times 100$$

たとえば、1950年と比べた2010年の日本の総人口の増加率は

$$2010\text{年の人口の増加率} = \frac{128057 - 84115}{84115} \times 100$$

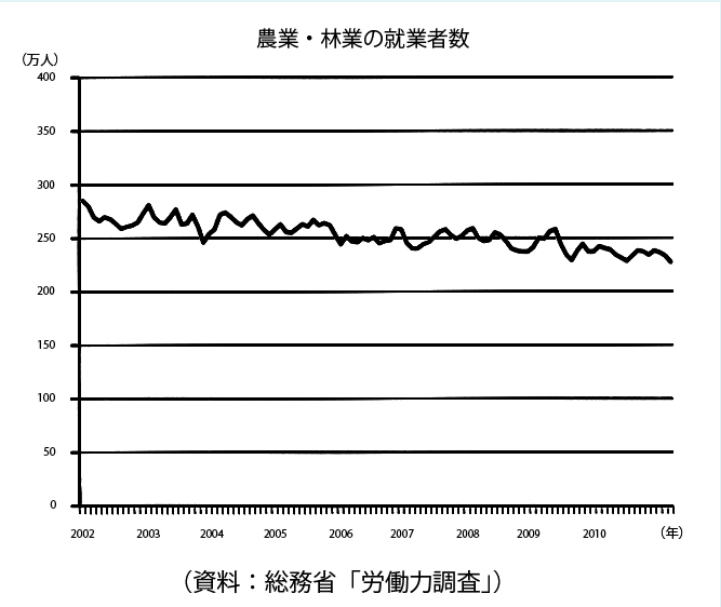
$$= 52.2\%$$

となります。

練習問題

（解答は P.41 です）

問１ 次のグラフは農業•林業で働く人の数を示した折れ線グラフです。



（資料：総務省「労働力調査」）

(1) 2002年から月ごとのデータは上下動を繰り返していて変化の傾向が読み取りにくくなっています。このようなデータにはどのようなデータ加工が効果的でしょうか。

(2) 下の表は2002年から2010年の農業･林業の年平均就業者数を示しています。2002年を基準とした増加率を求めなさい。

年	就業者数 (万人)	増加率 基準：2002年
2002	269	
2003	266	
2004	264	
2005	259	
2006	250	
2007	250	
2008	245	
2009	242	
2010	234	

Ⅳ 確率の基礎

確率とは、ある事柄(事象)が起こるか起こらないか確実には分からないとき、その事象の起こる「起こりやすさ(確からしさ)」を0から1の数字で表したものです。たとえば、公平なサイコロを投げたとき、1の目の出る確率は $\frac{1}{6}$ です。

これを

$$P\left(1\text{の目}\right)=\frac{1}{6}$$

と表します。

スポーツの勝敗などを考える場合も、確率を使って考えることができます。たとえば、明日、AチームとBチームの野球の今年の初試合が予定されているとします。昨年、AチームとBチームは10回試合を行って、その成績が5勝5敗で、今年もそれぞれ実力が同じくらいであれば、Aチームが勝つのは「半々」と考えることができます。半々であることを数字で表すと

$$P(A)=\frac{1}{2}$$

となります。今年はAチームにスター選手がたくさん入団して、Aチームが「十中八九」勝つと考えられれば、それを数字で表すと

$$0.8\leq P(A)\leq 0.9$$

となります。

確率の大きさを計算する方法として、次の三つの考え方があります。

1. 理論的確率(数学的確率)

同様に確からしい事象の起こる場合の数によって数学的に計算される確率のことをいいます。

ある行動の可能な結果がn通りあり、そのうち事象Aが起こる結果がa通り、事象Aが起こらない結果が$n-a$通りあり、これらの結果の全てが同じように起こることが可能で、かつお互いが重複しないとき、事象Aの起こる確率は $\frac{a}{n}$ と定義されます。

ただし、この方法で計算することができるのは、行動の取り得る全ての結果があらかじめ分かっていて、しかもそれらが全て同様に可能であることが認められる場合に限られます。

(例)　コインの表が出る理論的確率　$P\left(\text{表}\right)=\frac{1}{2}$

サイコロで6の目が出る理論的確率　$P(6)=\frac{1}{6}$

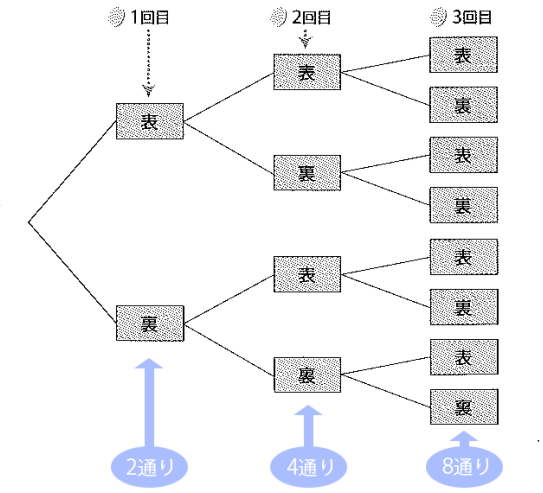
① 場合の数

ある事象が起こるとき、その起こり方の種類の数を場合の数といいます。理論的確率を計算するには、場合の数をもれなく重複なく数えることが必要です。

② 樹形図

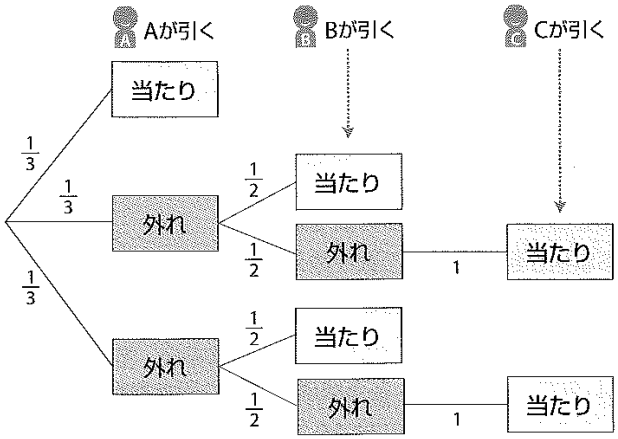
場合の数をもれなく数える方法として、樹形図があります。樹形図とは、それぞれの場合を枝分かれの図で描き表す方法です。

コイン投げの場合には、1回コインを投げると表か裏か2通りの場合があります。例えば、3回投げた場合には、次の図のように、1回目に表と裏の2通り、2回目と3回目にもそれぞれに表と裏があるので、全てを描き出すと樹の枝のようになります。



コイン投げの樹形図

樹形図の応用としてくじ引きを考えてみましょう。今、箱の中に3本のくじがあり、その中の1本が当たりくじとします。この箱からA、B、Cの3人の生徒が順番にくじを引くとします。くじを引く順番で当たりくじを引くことに有利不利が生じるでしょうか。



くじ引きの樹形図

樹形図で考えると

$$P(A) = \frac{1}{3}$$

$$P(B) = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$$

$$P(C) = \frac{2}{3} \times \frac{1}{2} \times 1 = \frac{1}{3}$$

となり、くじを引く順番で有利不利はないことが分かります。

2. 経験的確率(統計的確率)

実験や試行を多数回繰り返した場合に、起こった結果の度数に基づいて推定される確率のことをいいます。

事象Aが起こるかどうかを同じ条件のもとで繰り返して観察します。n回の観察で事象Aがa回起こったとすると、事象Aの相対頻度は $\frac{a}{n}$ です。観察回数nを大きくすると相対頻度が次第に安定し、一定の値に収束する傾向を示すとき、この値を事象Aの確率と考えます。

たとえば、サイコロを実際に25回投げて、1が3回出た場合、この実験における1の目が出る経験的確率は

$$P(1) = \frac{3}{25}$$

です。別の人が25回サイコロを投げると、結果が異なることもあります。つまり、経験的確率は、実験が異なれば値が異なります。ただし、試行の回数を増やしていけば、経験的確率は理論的確率(この場合であれば、$\frac{1}{6}$)に近づいていきます。

この方法では、可能な結果の全てが分かっていなくても確率を求めることができます。しかし、経験したことのない、初めて当面する事象、特に繰り返し観察することができない事象については、この方法で確率を計算することはできません。また、経験的確率は試行の回数が十分に大きいときにのみ使います。

スポーツの記録を分析するときにも、経験的確率の考え方を活用することができます。下の表はある投手の球種別投球数の記録です。

球種	ストレート	スライダー	カーブ	シュート	その他	合計
投球数	532	254	145	97	181	1209
割合	44	21	12	8	15	100

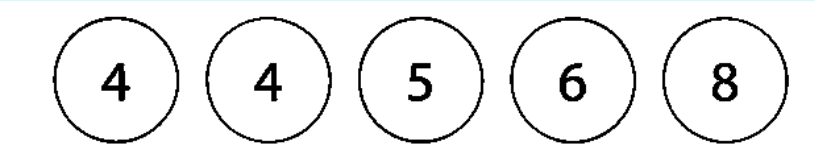
この場合、この投手がストレートを投げる確率は、44%あるといえます。

3. 主観的確率(個人的確率)

経験したことのない事象や1回限りの行動の結果については、1. や2. の方法では確率が求められません。その場合、その事象の確からしさは行動する人の個人的な確信の度合いによって決めるしかありません。この方法では個人の情報、知識、経験などによって確率は異なる可能性があります。主観的確率に基づく統計分析はベイズ統計学と呼ばれます。

練習問題 （解答は P.41 です）

問1



(1) 5つのディスクにはそれぞれ上のような数字が書いてあります。1つのディスクを無作為に選びます。

(i). 最も選ばれやすい数字はどれですか。

(ii). 偶数の書かれているディスクを選ぶ確率はいくつですか。

(iii). 偶数でかつ20の約数が書かれているディスクを選ぶ確率はいくつですか。

(2) 偶数が書かれているディスクから1つのディスクを無作為に選びます。書かれている数字が20の約数である確率はいくつですか。

【出典：IGCSE数学　(November2010Question Paper13)】

問2 あい子さんはサイコロを100回投げ、それぞれの目が何回出たか、記録をすることにしました。このサイコロの目の出方は同様に確からしいものとします。

(1) あい子さんは、6が出る回数を予測したところ、実際に6が出た回数と近くなりました。このとき、あい子さんが予測した値として最も適切なものを、次の①~⑤のうちから一つ選びなさい。

① $\frac{1}{6}$
② 0.6
③ 6
④ 17
⑤ 27

(2) あい子さんはサイコロを投げた結果を次の表のようにまとめました。

目の数	1	2	3	4	5	6
回数(回)	18	15	19	17	16	15

(i). 6の目の相対度数として正しいものを、次の①~⑤のうちから一つ選びなさい。

① $\frac{1}{6}$ %

② 0.6%
③ 16%
④ 0.15%
⑤ 15%

(ii). 奇数の目の相対度数として正しいものを、次の①~⑤のうちから一つ選びなさい。

① 53%
② 47%
③ 47回
④ 53回
⑤ $\frac{1}{6}$ %

(iii). 経験的確率と論理的確率が最も近くなる目の数として正しいものを、次の①~⑤のうちから一つ選びなさい。

① 1
② 2と6
③ 3
④ 4
⑤ 5

解答と解説

■練習問題　基本的なグラフ　（問題は p.8）

問１ 東京

富山県の最大値は255mm、最小値は123mm、東京は最大値209mm、最小値40mm、差が大きいのは東京である。

問２ 126,000円

200,000円×0.63=126,000円

■練習問題　質的データの分析　（問題は p.13）

問１ ①,③,④,⑤,⑥,⑦,⑨

問２ ④円グラフ

■練習問題　量的データの分析　（問題は p.26）

問１ ⑤以外すべて正しい

問２ (1) 1.5リットルより大2リットル以下

(2) 約1.73リットル

$$\frac{0.25\times8+0.75\times27\times1.25\times45+1.75\times50+2.25\times39+2.75\times21+3.25\times7+3.75\times3}{200}=1.7275$$

(3)

飲んだ水の量（リットル）	度数	累積度数
0　より大　0.5以下	8	8
0.5　より大　1以下	27	35
1　より大　1.5以下	45	80
1.5　より大　2以下	50	130
2　より大　2.5以下	39	169
2.5　より大　3以下	21	190
3　より大　3.5以下	7	197
3.5　より大　4以下	3	200
合計	200	

(4) 約55%

■練習問題　時系列データの分析　（問題は p.33）

問1　(1) 移動平均

(2)

年	就業者数	増加率 基準：2002年
2002	269	0.0
2003	266	−1.1
2004	264	−1.9
2005	259	−3.7
2006	250	−7.1
2007	250	−7.1
2008	245	−8.9
2009	242	−10.0
2010	234	−13.0

■練習問題　確率　（問題は p.38）

問1　(1) (i)　4

(ii) $\frac{4}{5}$

(iii) $\frac{2}{5}$

(2) $\frac{1}{2}$

問2　(1) ④

(2) (i)　⑤

(ii) ①

(iii) ④