目次

第１部　データ分析の基礎知識

ここでは、初級編で学んだ内容を踏まえ、データ分析に必要な基礎知識について学びましょう。

Ⅰ　様々なグラフ表現

１．統計グラフの特徴

初級編で紹介してきたグラフの特徴は以下の通りです。

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。 棒の高さがそれぞれのカテゴリの量を表している。
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。

この他にも様々な統計グラフが用いられます。

①　幹葉図

幹葉図は、データの大きさnが比較的小さい場合に用いられるグラフ表現で、数値データのばらつきを表す際に用いられます。

例えば、下の表はあるテストの20人分の成績をまとめたものです。

49	71	64	93	80	66	79	58	68	69
80	54	74	75	78	86	85	65	73	86

この数値だけを見て特徴を見つけることは難しいですが、これを幹葉図で表すことで、数値のばらつきの様子を把握することができます。

4	9
5	4 8
6	4 5 6 8 9
7	1 3 4 5 8 9
8	0 0 5 6 6
9	3

幹葉図では、左側の幹の部分に成績の10の位の数値を表示し、右側の葉の部分に成績の1の位を並べています。このグラフでは、60点台、70点台、80点台の数値が多くみられ、40点台、50点台、90点台は少ないことが分かると同時に具体的数値もつかむことができ

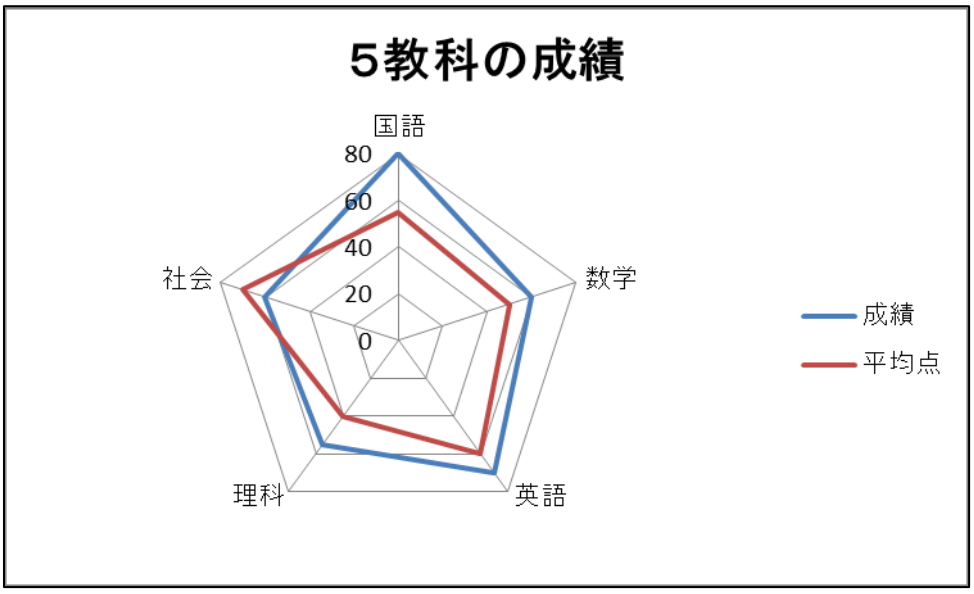
ます。

nが小さいときは手書きでも簡単に書くことができますが、nが大きくなると複雑になり過ぎるため、nが小さいときに適した表現です。また、幹葉図を左に90度回転すると、ヒストグラムと対応します。

②　レーダーチャート

レーダーチャートは複数の値をまとめて表すときに用いられるグラフです。

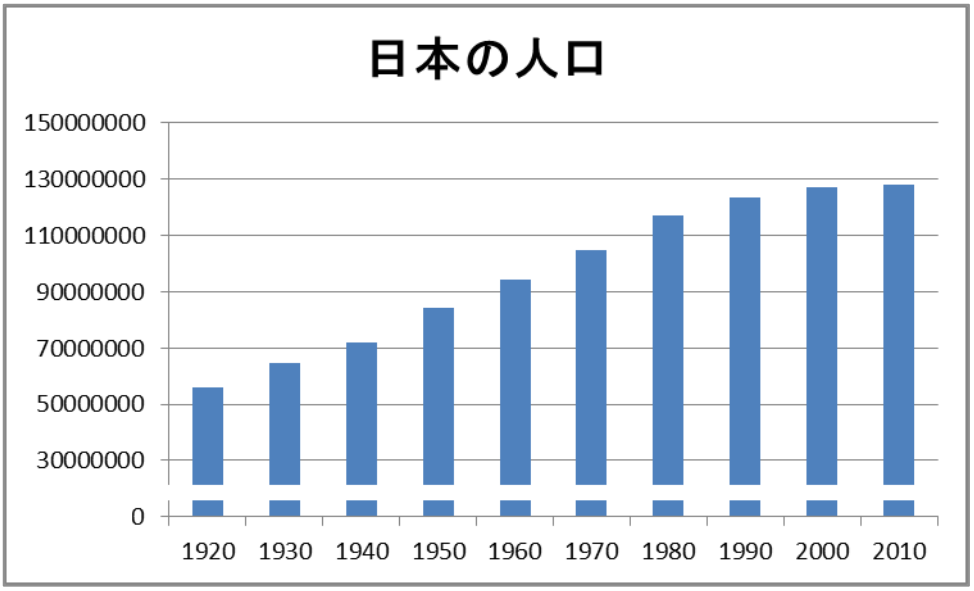
下のグラフは、ある生徒の五つの教科の成績を表しています。このグラフを見ることで、教科のバランスが判断できます。教科によってテストの難易度が異なるときは、クラスの平均点をグラフの中に表示することによって、クラスの平均点とその生徒の成績の関係を示すことができます。



この他、箱ひげ図や散布図といったグラフもありますが、これらについては、後で詳しく説明します。

2. 誤解を招きやすいグラフ表現

下のグラフは日本の人口の推移を表しています。数が大きいため、普通に棒グラフで表すと年ごとの変化が分かりにくくなります。そのため、棒の一部を省略する形でグラフにしています。



　このようなグラフの工夫自体は途中が省略されていることを明確に示していれば、かまいませんが、省略されていないことを明確にしていないと誤解を招く恐れがあります。また、グラフを解釈する場合には、途中が省略されていることを意識する必要があります。

Ⅱ　データの分布をみる

ヒストグラムや度数分布表を用いてデータの分布を見る方法については、初級編で説明しましたが、この章では分布の形を表現するその他の方法について紹介します。

１．分位数と５数要約

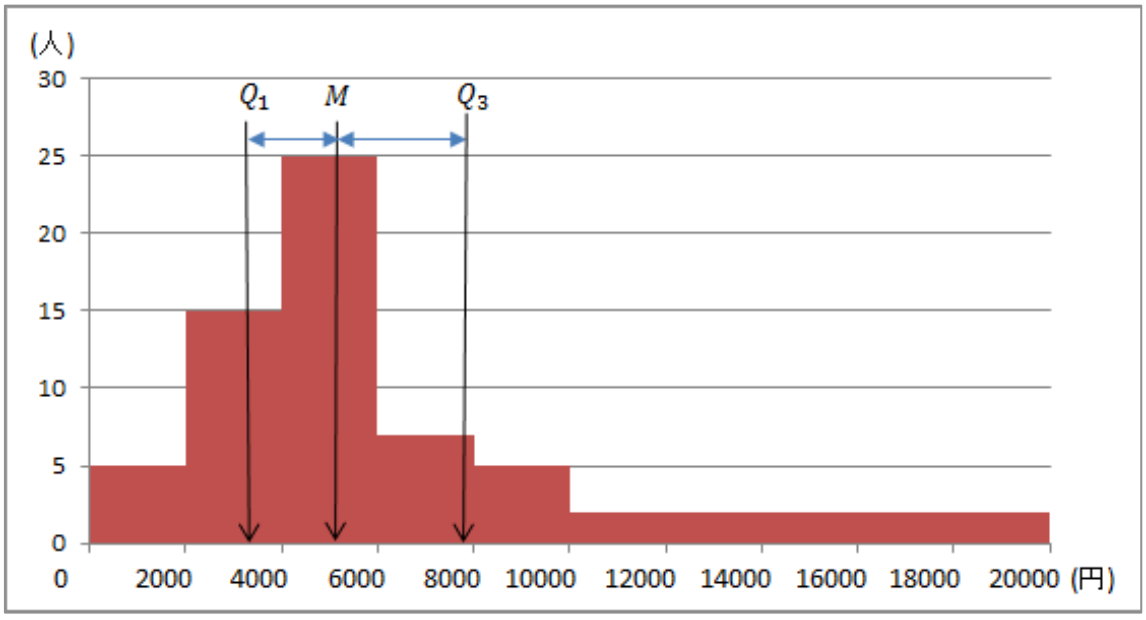
分布を表す指標として、初級編では代表値(平均値、中央値、最頻値)や範囲(レンジ)について説明しましたが、もう少し詳細に分布の形状を明らかにするためには、分位数(分位点)が用いられます。

分位数とは、データを大きさの順に並べ、データ全体をいくつかのグループに等分したときの境界となる値のことをいいます。よく使われるものとしては、4等分した四分位数があります。

最初の境界値を第1四分位数(Q_1)、次の境界値を第2四分位数(中央値Mと同値)、更に次の境界値を第3分位数(Q_3)と呼びます。また、データ全体を100等分する場合は、それぞれ1パーセント点、99パーセント点などと呼ばれます。

なお、最小値、第1四分位数、第2四分位数(中央値)、第3四分位数、最大値の五つの数をまとめて、5数要約と呼び、分布の形状を判断するために用いられます。

対称な分布ではQ_1、Q_3からMまでの距離はほぼ等しくなり、極端な外れ値が存在しなければ最大値と最小値もMに関して左右対称に近い位置にあることが期待されます。$Q_3 - M$が$M - Q_1$よりも大きい場合は、右の裾が長い分布であると予想されます。

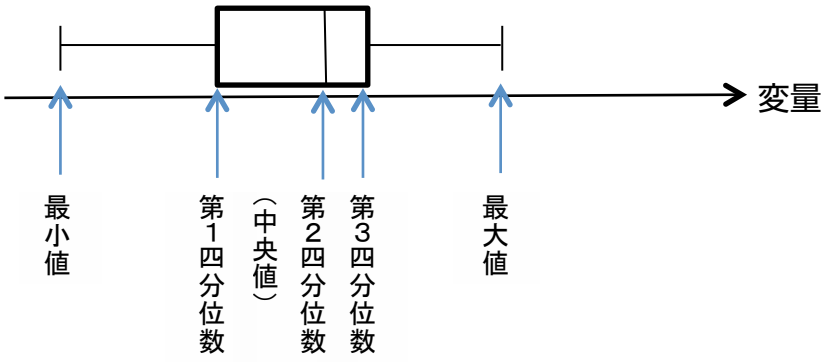


２．複数のデータの分布を比較する

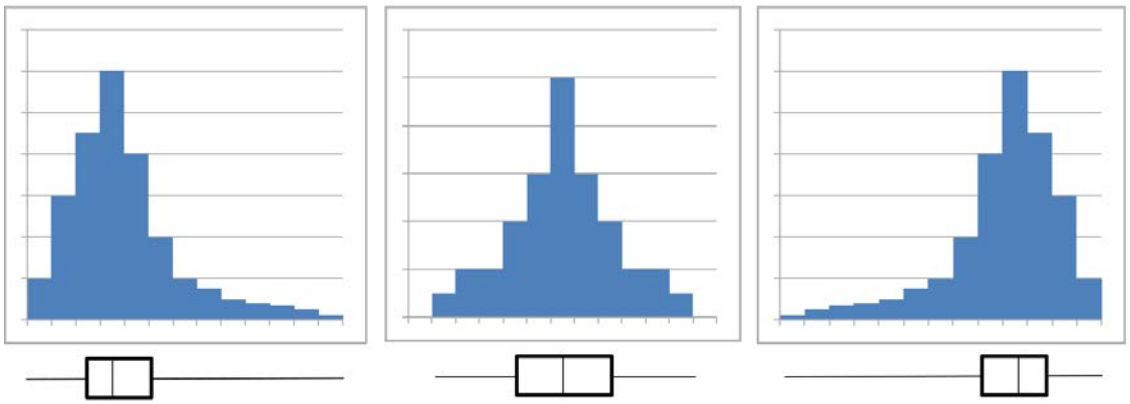
複数のデータの分布を比較する場合、ヒストグラムを複数個描いて比較するのは大変で

す。そのような場合には箱ひげ図と呼ばれるグラフが有用です。基本的な箱ひげ図は、最小値と最大値でひげの端を、第1四分位数と第3四分位数で箱の両端をそれぞれ表すグラフで、ヒストグラムと同様の情報を簡略化して表したものです。

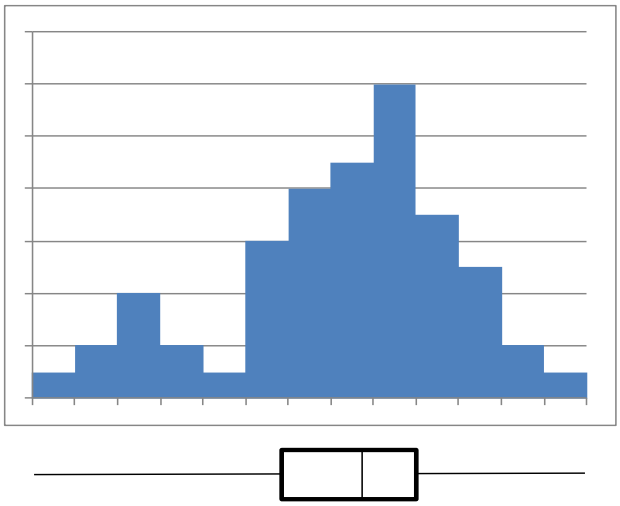
同じ目盛りを用いて複数の箱ひげ図を並べて書くことによって、多数の異なるデータの比較が可能になります。



ひげの両端の間の長さが範囲を表し、箱の長さが後で説明する四分位範囲を表します。分布の形によるヒストグラムと箱ひげ図の対応は下図のとおりです。



なお、箱ひげ図はヒストグラムと異なり、複数の山を持つ分布を適切に表すことができないため、注意が必要です。たとえば、下図のように山が二つの分布の場合、箱ひげ図では、十分な情報を集約できません。



3. データの散らばりを考える

データの散らばり(ばらつき)を表す指標として、初級編では範囲(レンジ)について説明しましたが、ここでは、その他の指標について説明します。

① 四分位範囲

あるファーストフードチェーンのSサイズのドリンクは150mlですが、実際にA店とB店でそれぞれ30個を調べたところ、次の表のようなデータが得られました。

	A 店	B 店
最小値	121	140
第1四分位数	138	146
第2四分位数	148	149
平均	150	150
第3四分位数	164	153
最大値	182	156

平均値はいずれも150mlですが、データの散らばりの程度は異なっています。A店の範囲は　$R = 182 - 121 = 61$　、B店の範囲は　$R = 156 - 140 = 16$となります。

範囲は極端な観測値(外れ値)があると大きく影響されるため、そのような場合には、$Q_3 - Q_1$をちらばりの程度を表す指標として用います。これは四分位範囲(IQR:Inter Quartile Range)と呼ばれます。

A店の四分位範囲は、

$$\mathrm{IQR} = 164 - 138 = 26$$

B店の四分位範囲は、

$$\mathrm{IQR} = 153 - 146 = 7$$

となります。

② 平均偏差

観測値の散らばりを考えるために、観測値からデータの平均を引いた差を考えます。この値は偏差と呼ばれます。変数をxとするとき、i番目の観測値の偏差は

$$\text{偏差}=\text{観測値} - \text{平均値}=x_i - \bar{x}$$

となります。

偏差はそれぞれの観測値と平均値の差を表し、偏差が正の値のときは$x_i > \bar{x}$、負の値のときは $x_i < \bar{x}$ を意味します。また偏差の合計(和)は0となります。そこで、ばらつきの大きさをみるために、偏差の絶対値をとって平均したものが平均偏差(M. D.)です。

$$\mathrm{M.D.} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

③ 分散

偏差の絶対値をとる代わりにその2乗値をとって平均したものが分散(S^2)です。

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

④ 標準偏差

分散の単位は観測値の平方(2乗)となり、平均とは単位が異なって解釈しにくいため、分散の正の平方根をとったものが標準偏差(S)です。

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

⑤ 変動係数

分布の中心の位置が著しく異なる場合には、分散(標準偏差)で分布の散らばり具合を比較することはできません。

たとえば、ある企業の従業者の年収を考えたとき、管理職の年収の標準偏差が450万円、

平均値が2千万円、アルバイトの年収の標準偏差が30万円、平均値が100万円であったとします。このとき、管理職とアルバイトではどちらのばらつきが大きいと考えるのでしょう。標準偏差をみると、管理職のほうがはるかに大きく、15倍ですが、平均値も20倍です。このようなときは、標準偏差を平均値で割った指標を用いることがあります。この指標を変動係数(C. V.)といいます。

$$\mathrm{C.V.} = \frac{S_x}{\bar{x}}$$

この例では、管理職の変動係数は、$450 \div 2000 = 0.225(22.5\%)$、アルバイトの変動係数は、$30 \div 100 = 0.3(30\%)$となり、ばらつきの程度はアルバイトのほうが大きいことが分かります。

これらは、その値が大きいほど観測値が散らばっていることを意味し、値が小さいほど狭い範囲に観測値が集まっていることを意味します。このうち、最も多く使われるのは、分散と標準偏差です。

練習問題

（解答は P.42 です）

問１　ある小学校の卒業生を対象に、卒業までに図書館から借りた本の冊数を調査した結果、次のデータを得た（仮想データ）。

最小値	1冊
第1四分位数	9冊
第2四分位数	12冊
平均	18冊
第3四分位数	23冊
最大値	126冊

この結果から次の2つのことを考えた。
A：卒業までに半数の児童が18冊以上の本を図書館から借りている。
B：借りた本の冊数は平均よりも少なかった児童が過半数である。
このとき、2つの考えについて適切な組み合わせは次の①～④のうちどれか。

① AもBも正しい
② Aのみ正しい
③ Bのみ正しい
④ AもBも正しくない

問２ 次の2つの度数分布表について、下の①～④のうちから最も適切なものを一つ選べ。

個数	A の度数	B の度数
1	30	10
2	20	20
3	10	30
4	0	0
5	0	0
6	10	30
7	20	20
8	30	10

Ⅰ：AとBの平均値は等しい
Ⅱ：AとBの範囲は等しい
Ⅲ：AとBの分散は等しい

① Ⅰのみ正しい
② ⅠとⅡのみ正しい
③ ⅠとⅢのみ正しい
④ すべて正しくない

Ⅲ 観測値の標準化と外れ値

1. 観測値の標準化

複数のデータを比較する場合、平均値や標準偏差が大きく異なると比較することは難しくなります。また、測定の単位が異なる場合も同様の問題が生じます。このような場合、データに標準化又は基準化と呼ばれる処理を行い、統一した基準で比較することがあります。

観測値の標準化とは、各観測値$x_i(i = 1, \dots, n)$に対して、平均を差し引き、標準偏差で割ることをいい、次の式で表されます。

$$z_i = \frac{\text{観測値} - \text{平均値}}{\text{標準偏差}} = \frac{x_i - \bar{x}}{S}$$

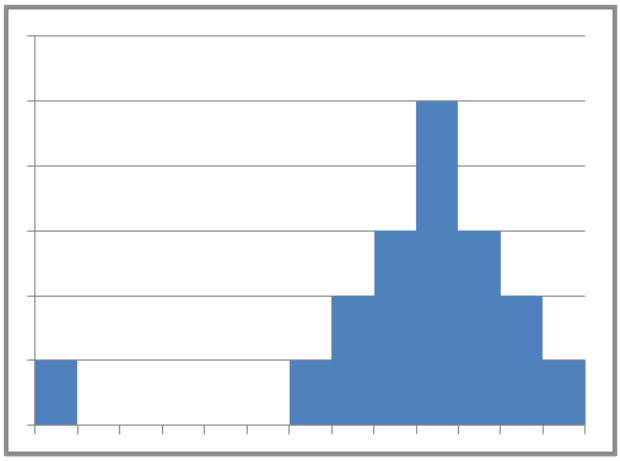
この処理によって、平均 $\bar{z} = 0$ 、標準偏差は $S_z = 1$ にそろえられたことになり、標準化された値はz値又はzスコアと呼ばれます。

成績で用いられる偏差値は、平均50、標準偏差10になるように変換したものです。

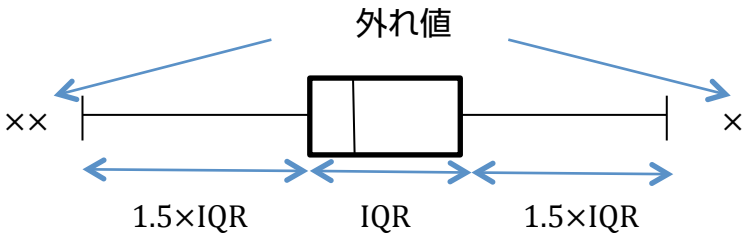
2. データの外れ値とその検出

調査や実験によって得られたデータの分布を確認せずに平均値や標準偏差を求めることは誤った解釈につながる恐れがあるため、注意が必要です。データが得られたら、まず、ヒストグラムや箱ひげ図などの統計グラフを用いて、データ全体の分布を確認することが大切です。それにより、複数の分布が混ざったデータになっていないか、他の観測値から大きくかけ離れた観測値がないかなどを検証し、場合によっては、外れた観測値を除いて計算するなど適切なデータ分析が可能になります。

たとえば、下の図のヒストグラムのように他の観測値と大きく離れた観測値があった場合には、この観測値を除いて考えるか、このような外れた値の影響を受けづらい指標を用いることを考える必要があります。このような他の観測値と比べ大きく外れた観測値を外れ値と呼びます。しかし、一般的にはどの観測値を外れ値とするかの判断は容易ではありません。たとえば、平均 $\bar{x}$ から標準偏差sの3倍以上離れた値を外れ値とすると、そもそも外れ値が存在するデータは$\bar{x}$ もsも大きくなるため、外れ値が見つからないこともあります。



箱ひげ図は、外れ値を検出するための簡易な手法であり、次のように外れ値を定義します。下の図のように、箱の両端から箱の長さ(四分位範囲=IQR)の1.5倍よりも外側に離れている観測値を外れ値と呼びます。



練習問題　（解答は P.42 です）

問１　あるクラスの試験において、以下の3人を点数で小さい順に並べるとどうなるか。下の①～④のうちから最も適切なものを一つ選べ。

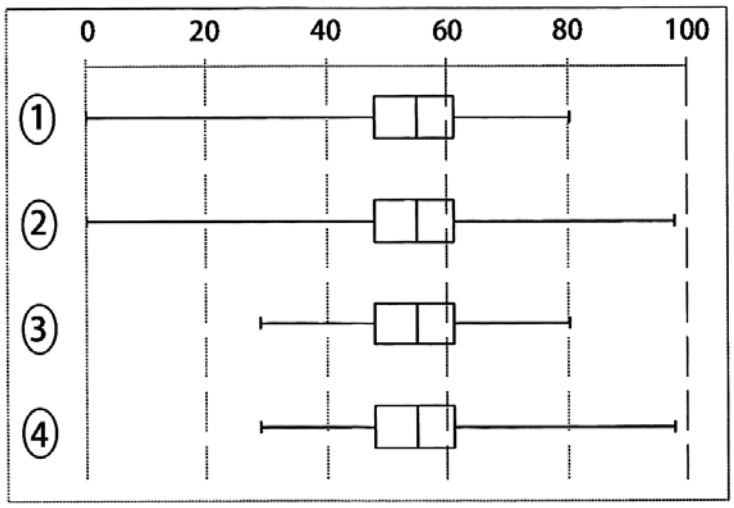
Aさん:　クラスの平均値と標準偏差で点数を標準化して求めたところ値が1となった。
Bさん:　点数がちょうどクラスの点数の第1四分位数と一致した。
Cさん:　点数がちょうどクラスの点数の平均値と一致した。
なお今回の試験におけるクラスの点数の分布は平均値を中心に左右対称なひと山型の分布で平均値と中央値はほぼ一致した。

① A→B→Cの順
② B→A→Cの順
③ B→C→Aの順
④ この情報だけでは求められない。

問２　生徒30人のクラスのある日の登校時間(分)を調べたところ、次のデータを得た。

29	32	35	44	45	46	46	48	50	52
52	53	53	54	55	55	56	57	58	58
59	59	61	65	68	75	76	78	90	98

このデータでは最小値29分、第1四分位数48分、第2四分位数55分、平均値56.9分、第3四分位数61分、最大値98分となっている。第1四分位数－1.5×四分位範囲より小さい、または第3分位数＋1.5×四分位範囲より大きい観測地を外れ値としたき、このデータの適切な箱ひげ図はどれか(グラフははずれ値を取り除いた場合の基本箱ひげ図である)。次の図の①～④のうち最も適切なものを一つ選べ。



Ⅳ 関係の分析

1. 二つの変数の関係

これまでは一つの変数の見方について説明してきましたが、この章では二つの変数を同時に考え、その関係を分析する手法について説明します。

① クロス集計表

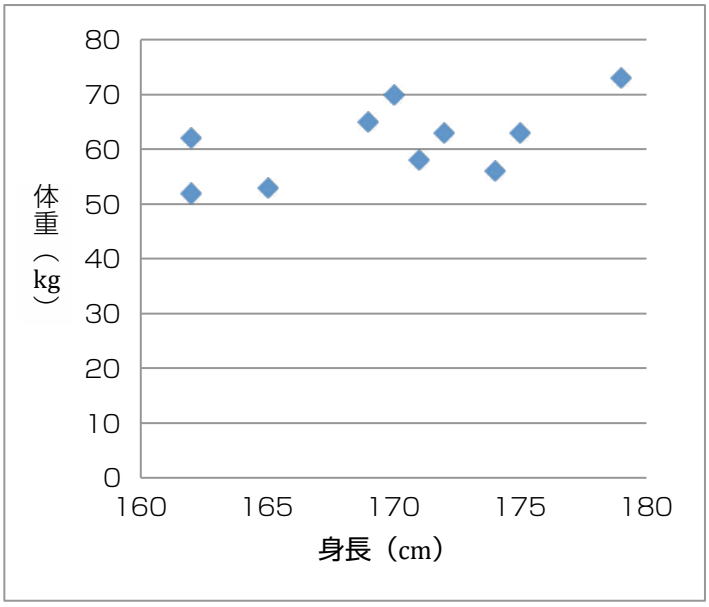
性別や所属クラスのような質的データ同士の関係を考える場合には、クロス集計表を用います。たとえば、下の表は大学生の住所について性別にまとめたクロス集計表ですが、表をみると女子学生は男子学生に比べて自宅通学の比率が高いという特徴が読み取れます。

	下宿	自宅
男	110	214
女	30	290

② 散布図

身長と体重のような量的データ同士の関係を考える場合、2変数であれば、x軸とy軸に二つの変数の数値を対応させて図を描くと、視覚的にxとyがどのような関係になっているか把握することができます。このような図は散布図と呼ばれ、2変数のデータを分析する際には、まず、この散布図をプロットしてみます。

身長（cm）	体重（kg）
162	52
170	70
169	65
175	63
179	73
171	58
162	62
174	56
165	53
172	63



散布図において、一つの変数の値が増えたときに、他方の変数の値も増える傾向にあるとき、2変数間には正の相関関係があるといいます。逆に一つの変数が増えたときに、他方の変数が減る傾向にあるときは負の相関関係があるといいます。また、それらの関係が見られなかったときは、相関関係がない、もしくは無相関といいます。

正の相関関係　　無相関　　負の相関関係

相関の強さは直線的な関係の強さによって、直線に近いときは強い、そうでないときは弱いといいます。

2. 相関係数

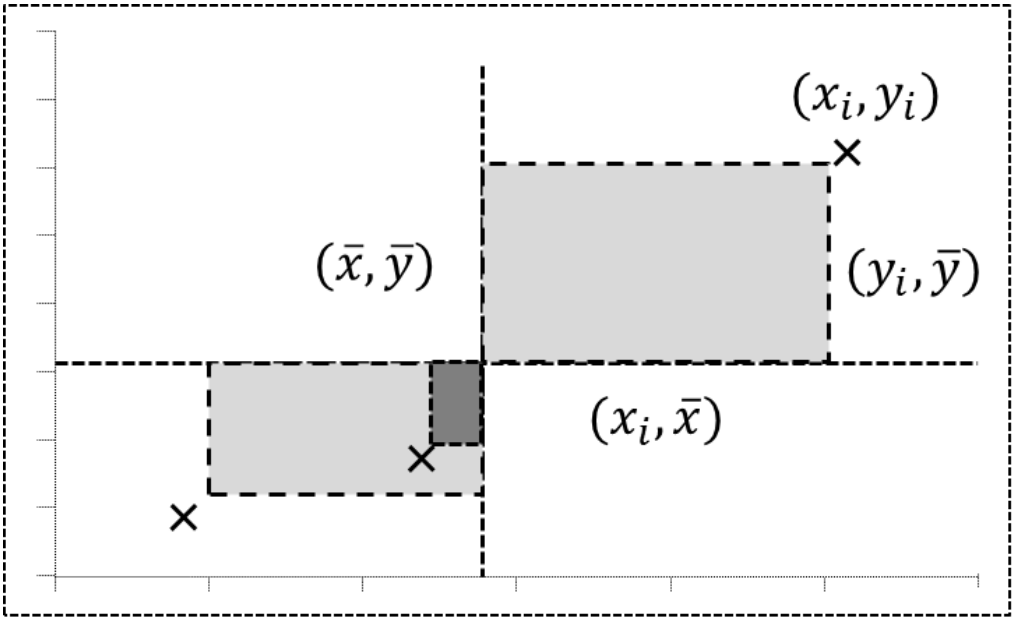
散布図を用いると2変数間の相関関係を視覚的に見ることができます。しかし、散布図では軸や縦横比の描き方によっては、情報を読み間違える可能性があります。そこで2変数の関係を数値として表す指標を考えます。

① 共分散

x, yの観測値の組からなるデータを $(x_1, y_1),\ldots,\left(x_n,\ y_n\right)$ とすると、2変数の共分散(S_{xy})は以下の式で定義されます。

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

共分散は、下の図のように2変数のそれぞれの平均値と観測値の偏差を求め、それらで作る長方形の面積の総和を観測値の個数nで割ったものです。ただし、偏差の定義から右上と左下は正の面積、左上と右下は負の面積として求めます。



これにより、平均値に対して右上と左下に偏って観測値が分布している場合、共分散の値は大きな正の値となり、逆に左上と右下に偏って観測値が分布している場合、共分散の値は大きな負の値になります。平均値を中心に左右上下にまんべんなく散らばっている場合、共分散の値は0に近づきます。このことから、共分散は正の相関のときは正の値、負の相関のときは負の値をとることが分かります。

② 相関係数

共分散により二つの変数の関係の強さを測ることができますが、共分散の値は変数の単位に依存して変化します。この点を修正して相関関係を測る指標として、相関係数があります。相関係数は、2変数の共分散をそれぞれの標準偏差を掛け合わせたもので割った値であり、xの標準偏差をS_x、yの標準偏差をS_y、2変数の共分散をS_{xy}とするとき、相関係数rは以下の式で定義されます。

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

なお、相関係数は

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)\left(\frac{y_i - \bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right)$$

と式を変形することができます。

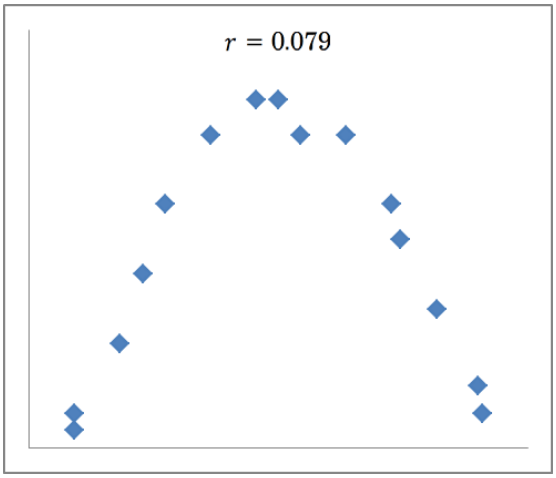
ここで、xとyを標準化して、$u_i = \frac{x_i - \bar{x}}{s_x}$ 、 $v_i = \frac{y_i - \bar{y}}{s_y}$ とおくと、uとvの共分散は、

$$s_{uv} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x s_y}$$

となります。つまり、相関係数rはxとyを標準化したuとvの共分散であることから、xやyを何倍かしたり、定数を加えて単位を変換しても、相関係数は変化しないことが分かります。

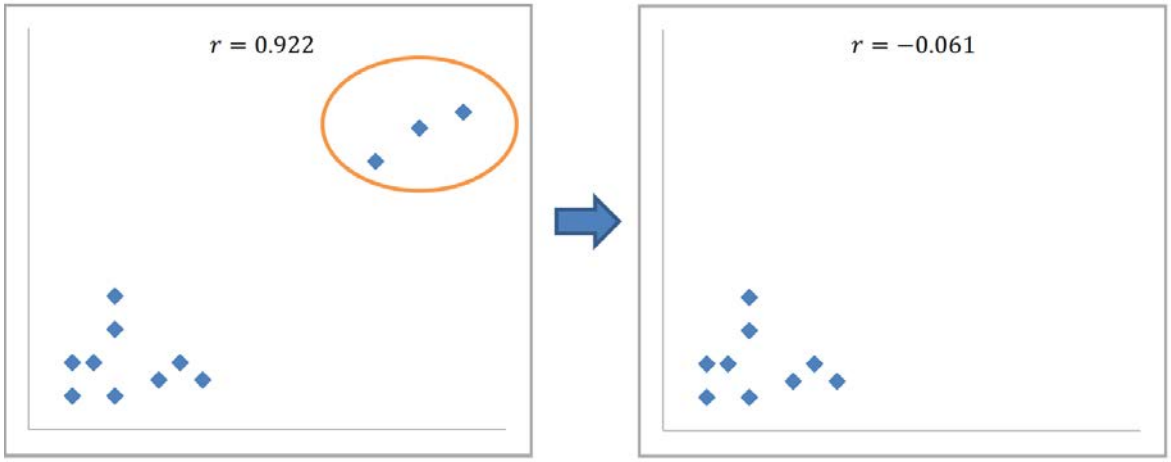
相関係数は-1から1の値を取り、直線に近い関係になるほど絶対値が1に近づきます。

なお、相関係数は直線状の関係を測る尺度であり、2変数間の関係が直線状でない場合はその強さを適切に測ることはできません。たとえば、下の図のように左右対称の2次曲線状の関係が見られる場合の相関係数は0に近い値になります。



また、相関係数は、外れ値の影響を強く受けます。たとえば、下の左側の図のデータで相関係数を求めると、$r = 0.922$と正の強い相関といえますが、散布図から、他の観測値から大きく離れた三つの観測値を除いて相関係数を求めると、$r = -0.061$とほとんど相関関係がないことになります。

このように相関関係を考えるときには、必ず散布図をみることが大切です。



練習問題

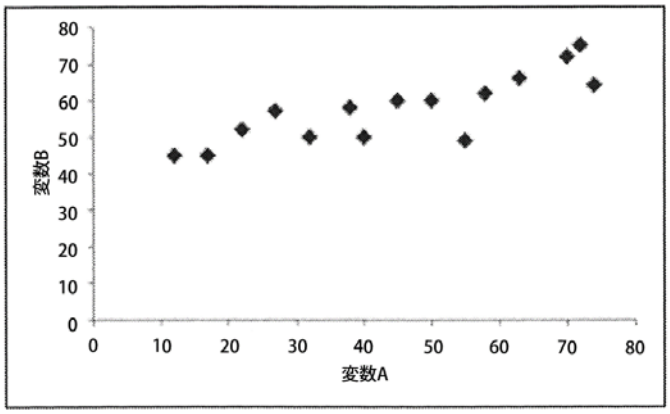
（解答は P.42 です）

問１ あるクラスで中間試験と期末試験を実施したとき、すべての人が中間試験の点数に20点加えた点数を期末試験でとった場合、このクラスの中間試験と期末試験の相関関係はどうなるか。次の①～④のうちから最も適切なものを一つ選びなさい。なお中間試験と期末試験では同じ人が受け、当日の欠席はなかったとする。

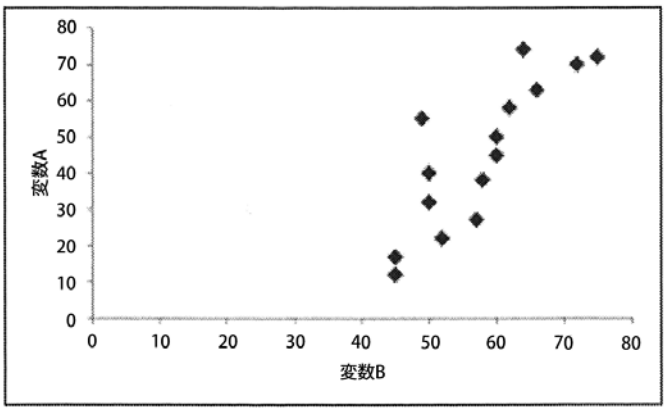
① 正の相関関係を持つ
② 相関関係はない（無相関）
③ 負の相関関係をもつ
④ この情報だけでは相関関係はわからない

問２ 2つの変数A、Bについての観測値$(a_1,\ b_1),\ldots,(a_n,b_n)$が求められたとき、以下の3つの散布図を次の手順で作成した。
（1）は横軸にa、縦軸にbを取った図
（2）は縦軸にa、横軸にbを取った図
（3）は横軸に$100\times a$、縦軸に$100\times b$を取った図

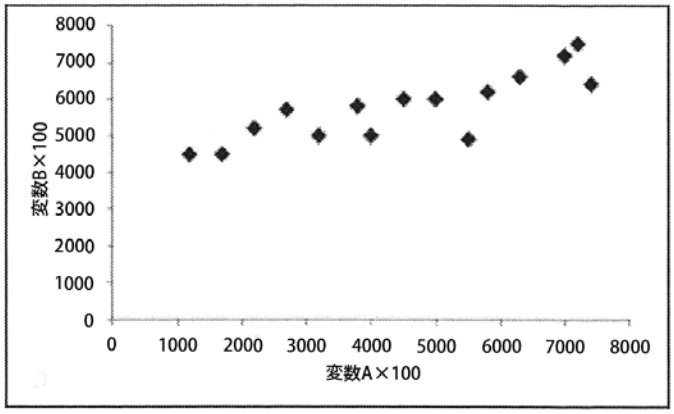
(1)



(2)



(3)



このとき上の散布図の中で相関係数が最も大きいものはどれか。次の①～④のうちから最も適切なものを一つ選べ。

① (1)の散布図
② (2)の散布図
③ (3)の散布図
④ (1), (2), (3)の相関係数は同じになる

V 確率

初級編では、確率の基礎について説明しましたが、この章では確率についてもう少し詳しく紹介します。

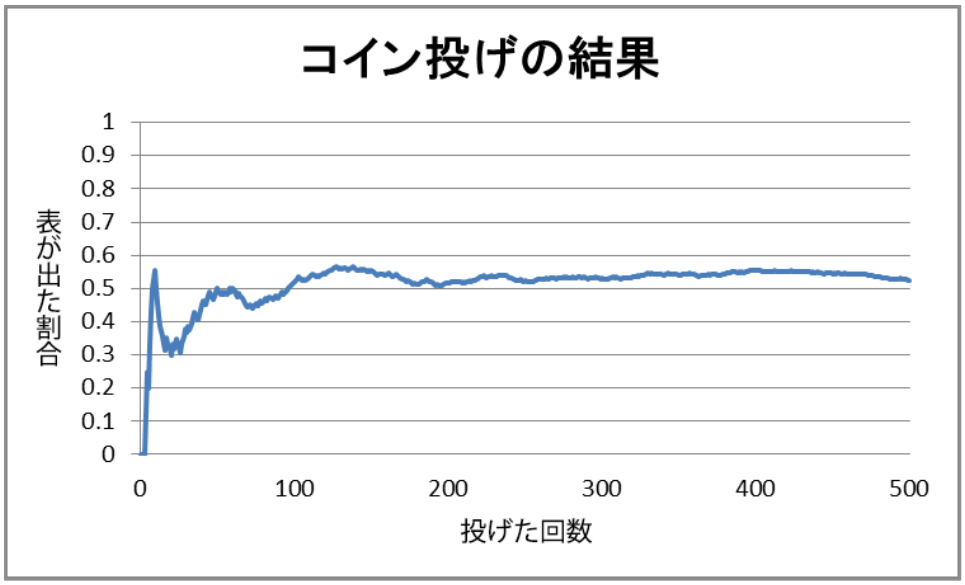
1. 確率の意味

私たちの生活の中では、まだ実際には起こっていない事柄や情報が不足しているために不確かな事柄についても判断をしていく必要があります。たとえば、朝出かける前に傘を持っていくのかどうか判断するには、その日雨が降るかどうかを考えます。このような事柄を事象と呼び、不確かな事象について、その起こりやすさの程度を表す数値を、その事象の確率といいます。

① 経験的確率

繰り返し実験が可能な場合については、ある程度大きな回数の実験を行い、その結果に基づいて事象の起こりやすさを判断することができます。

たとえば、下の図は、コインを500回投げるという実験を行い、横軸を投げた回数、縦軸をそれまでに表が出た割合としてグラフを描いたものです。



コイン投げの場合、回数が少ないときには表が出た割合は大きく変化しますが、投げる回数を増やしていくと、表が出た割合はある値(0.5)に近づいていきます。この実験結果から、コインの表が出る確率を

$$P(\text{表}) = 0.526$$

と求めることができます。

② 理論的確率

先ほどは、コインの表が出る確率を実験で求めましたが、コインのように表裏がほぼ同

じ可能性で出ると仮定できる場合には、そのことを利用して確率を求めることができます。起こりうるいくつかの事象について、それらが起こる可能性が等しいとき、同様に確からしいといいます。

同様に確からしいと仮定できる起こりうる場合の数がn通りあり、ある事象Aに含まれる場合の数がk通りあるとき、Aの起こる確率$P(A)$は

$$P(A) = \frac{k}{n}$$

と定義されます。

たとえば、サイコロを投げたときに偶数の目が出る確率を考えましょう。ゆがみのないサイコロは1から6の目が同じ確率で出ると考えられます。このとき、起こりうる結果は1から6の6通りあります。そのうち、偶数の目の場合は、2、4、6の目が出る場合で3通りです。このことから、偶数の目が出る確率は、$\frac{3}{6} = \frac{1}{2}$ となります。

2. 事象と確率

白と赤の2つのサイコロを投げる例を考えてみましょう。白と赤のサイコロを投げた結果をその順番に$(1,1)$というように表すと、可能な結果は、

$$(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)$$

$$(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)$$

$$(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)$$

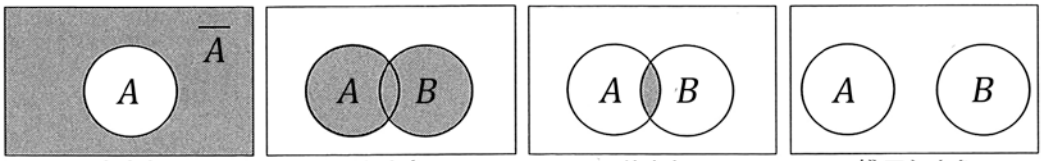
$$(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)$$

$$(5,1), (5,2), (5,3), (5,4), (5,5), (5,6)$$

$$(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)$$

の全部で36通りです。これらの事象はこれ以上分解できないため、基本事象と呼ばれることがあります。ゆがみのないサイコロやコインを投げるときは、それぞれの基本事象の確率は等しいと想定します。

いくつかの事象を組み合わせた事象も考察の対象となります。たとえば、白のサイコロの目が偶数で、赤のサイコロの目が奇数となる事象などが考えられます。事象の組合せを下の図のように整理します。



余事象　和事象　積事象　排反な事象

事象AとBのいずれかが起こることを事象の和と呼び$A \cup B$(エーカップビー)と表します。これを和事象といいます。事象AとBの両方が起こることは事象の積と呼び、$A \cap B$(エーキャップビー)又は単にABと表します。これを積事象といいます。Aが起きないという事象を余事象と呼び、$\bar{A}$(エーバー)と表します。

また、「A:白いサイコロの目が6」と「B:白いサイコロの目が4以下」のように、AとBの両方が同時には起こらない場合、「これらの事象は互いに排反である」といいます。記号では、$A \cap B = \phi$(ファイ)と表します。ϕは起こりえない事象に対応するもので空事象と呼び、その確率Pはゼロとなります。互いに排反である事象A、Bのいずれかが起こるとき、その確率$P(A \cup B)$は、

$$P(A \cup B) = P(A) + P(B)$$

となります。これは排反事象の加法定理と呼ばれます。

3. 事象の独立性

白と赤の2つのサイコロを投げたとき、次の三つの事象の確率を考えてみましょう。
A:白のサイコロの目が3である。
B:赤のサイコロの目が2である。
C:白のサイコロの目が3で、赤のサイコロの目が2である。

赤と白のサイコロの目の組合せは36通りあり、これらは全て同確率と考えます。このとき、事象Aには赤のサイコロの目の出方が6通りあり、事象Bも白のサイコロの目の出方が6通りあるので、どちらの確率も$P(A) = P(B) = \frac{6}{36} = \frac{1}{6}$ となります。一方、事象Cのような目の出方は1通りであるので、$P(C) = \frac{1}{36}$ となります。

事象Aは白のサイコロだけの結果に関係し、事象Bに影響されません。同様に事象Bは赤いサイコロだけの結果に関係する事象であり、事象Aに影響されません。このような場合、二つの事象AとBは独立であるといいます。

一方、事象Cは事象Aと事象Bが両方起こる場合であり、$C = A \cap B$と表すことができます。Cの確率を求めると、Aの確率とBの確率を掛け合わせたものとなっており、

$$P(A \cap B) = P(A)P(B)$$

という関係が成り立っています。このような関係が成り立つとき、二つの事象は独立であると定義します。

4. 反復試行

コイン投げやサイコロ投げのように、同じ条件の下で繰り返すことができるような実験や観測を試行といいます。上の例では、白いサイコロを投げる試行と赤いサイコロを投げる試行の二つの試行を行っていることになります。このように二つの試行T_1、T_2に対して、T_1によって決まる全ての事象と、T_2によって決まる全ての事象が独立であるとき、T_1とT_2は独立であるといいます。

ある独立な試行を繰り返し行うとき、それらの試行を反復試行といいます。
たとえば、コイン投げを5回繰り返す場合を考えると、これらは反復試行となります。
では、コインを5回投げて3回表が出る確率を考えてみましょう。

表が3回出るためには、1回目、2回目、3回目に表が出てもよいし、1回目、3回目、5回目に表が出てもかまいません。このうちの一つ、表、表、表、裏、裏という順序で起こる場合を考えてみます。

$P(\text{表}) = \frac{1}{2}$ であるので、$P(\text{裏}) = 1 - \frac{1}{2} = \frac{1}{2}$ となり、各回の試行は独立であるため、この確率は$P(\text{表})^3 \times P(\text{裏})^2 = (\frac{1}{2})^3 \times (\frac{1}{2})^2 = \frac{1}{32}$ となります。

この同時確率は表の出る順序が変わっても常に同一です。

表が出る順序の組合せは、

(表，表，表，裏，裏)
(表，表，裏，表，裏)
(表，表，裏，裏，表)
(表，裏，表，表，裏)
(表，裏，表，裏，表)
(表，裏，裏，表，表)
(裏，表，表，表，裏)
(裏，表，表，裏，表)
(裏，表，裏，表，表)
(裏，裏，表，表，表)

の10通りあります。

そして、これらの順序は互いに排反であるので、コインを5回投げて3回表が出る確率は$10 \times \frac{1}{32} = \frac{10}{32} = \frac{5}{16}$ となります。

表が出る順序の組合せは、言い換えると、五つの数字の中から三つの数字を選ぶ組合せになります。n個の異なる数字の中からk個を選ぶ組合せの数は、一般に${}_nC_k$と表し、

$$ {}_nC_k = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \times \cdots \times 2 \times 1} $$

で計算できます。

1回の試行である事象Aが起こる確率をpとし、同じ試行をn回独立に繰り返したときに、事象Aがk回起こる確率は、${}_nC_k p^k(1-p)^{n-k}$となります。

5. 条件付き確率

ここでは、ある条件が満たされているときの確率を考えます。

たとえば、ある高校のクラスで生徒を性別と出身中学校で分けると次の表のようになっているとします。

	A 中学校	B 中学校	C 中学校	合計
男子	10	7	5	22
女子	5	7	6	18
合計	15	14	11	40

この40人の中から一人を無作為に選ぶとき、男子である確率は$\frac{22}{40}=\frac{11}{20}$となります。もし、選ばれた生徒がA中学校であることが分かっているときには、15人の中から選ばれることになり、男子の確率は$\frac{10}{15}=\frac{2}{3}$ となります。このようにある条件をつけたときの確率を条件付き確率といいます。

一般に、事象Aが与えられたときの事象Bの条件付き確率$P(B|A)$は

$$P(B|A)=\frac{P(A\cap B)}{P(A)}$$

と定義されます。

上の例で事象Aを「選ばれた生徒がA中学校出身である」とし、事象Bを「男子である」とすると、$P(A)=\frac{15}{40}$、$P(A\cap B)=\frac{10}{40}$であるから、条件付き確率は

$$P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{10/40}{15/40}=\frac{2}{3}$$

となります。

条件付き確率の定義を変形すると、次の式のようになります。

$$P(A\cap B)=P(A)P(B|A)$$

この式は、乗法定理と呼ばれます。

(補足)順列・組合せ

サイコロ投げ、コイン投げやカードの抜き取りなど、同様に確からしい場合に基づいて確率を計算する問題では、場合の数を数えることが必要となります。場合の数を数える方法として、初級編では樹形図を紹介しましたが、組合せの数が大きいとき、樹形図で数えるのは大変です。その場合に順列・組合せの考え方を使うことができます。

全て異なる数字が記されているn枚のカードから1枚を抜き出すとき、異なる結果はn通りあります。順番に2枚を抜き出し並べるとき、異なる結果は、1枚目はn通り、2枚目は1枚抜き出した後なので、$n-1$通りとなるため、$n\times(n-1)$通りです。ここでは同じ数字の2枚(組合せ)のカードであっても、順番が違えば異なる結果とみなしています。例えば、(1，2)も(2，1)組合せとしては同じですが、並べ方としては異なる結果と考えています。

一般に、n枚のカードから順番にk枚を抜き出して並べると、異なる結果は$n\times(n-1)\times(n-2)\times\ldots\times(n-k+1)$通りとなります。これを順列$({}_nP_k)$と呼び、

$$ {}_nP_k = n\times(n-1)\times(n-2)\times\ldots\times(n-k+1) $$

と定義されます。

たとえば52枚のカードから2枚を順に抜き出す場合には、${}_{52}P_2 = 52\times51$となります。

特にn枚のカードを全て順番に抜き出すときは、${}_nP_n = n\times(n-1)\times(n-2)\times\ldots\times2\times1 = n!$となります。$n!$を階乗と呼びます。

一方、n枚のカードからk枚のカードを抜き出すとき、そのカードの組合せの数は、組合せ$({}_nC_k)$と表します。この場合は、(1，2)と(2，1)は同じ組合せと考えます。

抜き出したk枚の並べ方は$k!$通りあります。この場合、抜き出した2枚のカードの並べ方は2×1通りあります。順列には、組合せが同じものも含まれているので、最終的に求める組合せの数は、${}_nP_k$をkの順列の数$k!$で割り、

$$ {}_nC_k = \frac{{}_nP_k}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k!} $$

となります。

この場合は、$\frac{(52\times51)}{(2\times1)} = 1{,}326$ 通りとなります。

練習問題

（解答は P.43 です）

問１　袋の中に赤いカードが20枚、青いカードが15枚、黄色いカードが15枚入っている。よくかき混ぜて、この50枚のカードの中から1枚を選ぶとき青いカードを選ぶ確率を、次の①～④のうちから一つ選べ。

① 0.15

② 0.2

③ 0.3

④ 0.4

問２　ある病気にかかる確率は、喫煙者と非喫煙者で異なり、喫煙者では0.3%、非喫煙者では0.1%とする。もし、ある集団の喫煙者の割合が20%であるとき、病気にかかった人が喫煙者である確率を、次の①～④のうちから一つ選べ。

① $\frac{3}{5000}$

② $\frac{1}{3}$

③ $\frac{3}{7}$

④ $\frac{12}{13}$

Ⅵ 標本調査

１．全数調査と標本調査

私たちの社会の中では、様々な調査が行われています。これらの調査の結果は、政策を決定するための基礎資料として用いられたり、企業における製品の開発や出荷量の決定などの資料として利用されています。

ある集団について知りたいと考え調査を行う際に、対象とする集団を全て調査するものを全数調査あるいは悉皆（しっかい）調査といいます。これに対して、対象とする集団の一部について調査するものを標本調査といいます。

全数調査を行えば、集団についての情報を全て得ることができるため、その意味では全数調査が望ましいですが、実際には、対象とする集団が大きくなると、費用や手間が莫大になるため全数調査を行うことは難しくなります。そのため、全数調査に基づくものは、国勢調査などごく一部の調査に限られており、多くの場合、標本調査が行われます。

標本調査が行われる理由としては、次のようなものが考えられます。

1）製品の寿命調査のように、調査を実施するとその製品が使えなくなる場合。
2）短い期間での時間的な変化をみるため、短時間での調査・分析が必要な場合。
3）全数調査を実施するには莫大や費用がかかる場合。

２．母集団と標本

特徴や傾向などを知りたいと考える集団全体を母集団といいます。標本調査とは、母集団の特徴を知るためにその一部を選び出し、調査を行う方法であり、実際に調査を実施する母集団の一部を標本、選び出すことを標本抽出といいます。また、標本として選び出される個体数を標本の大きさといいます。

標本調査から母集団の性質を正しく推計するためには、母集団の情報が標本に正しく反映されていなければなりません。つまり、標本が母集団の「縮図」になっていることが望ましいと考えられます。

たとえば、日本全体でのコンピュータの利用割合を知りたいときに、インターネット調査で調べたとします。この場合、標本がインターネットを利用している人に限定されるわけですから、利用割合は知りたいと考えた母集団での利用割合よりも高くなるでしょう。このように母集団と標本の傾向が異なる場合には、標本に偏りがあるといいます。

偏りなく標本を抽出する方法として、くじ引きのような形で無作為に抽出する無作為抽出法（ランダム・サンプリング）があります。

３．無作為抽出法

無作為抽出法とは、くじ引きのような形で、母集団に含まれている固体が同じ確率で抽出される方法のことをいいます。具体的には、母集団に含まれる固体全てに異なる番号をつけて、その番号を確率的に抽出します。この方法を単純無作為抽出法といい、最も基本的な抽出法です。

番号を確率的に選ぶ方法としては、次のようなものがあります。

① サイコロやくじびき

たとえば、0から99までの番号のついたくじを準備して、その中から1つ選ぶ方法や正二十面体の各面に0から9の数字のうちの一つを書いて、0から9までの数字が2面ずつあるサイコロを使って、数字を選ぶ方法などがあります。

② 乱数表

あらかじめ①のような方法で作成した数字を表にしたものを乱数表といいます。この乱数表の数字の中から一つ選んで、その場所をスタートしてある方向に数字を順番に選んでいく方法が用いられます。

③ コンピュータで乱数を発生

①や②の方法では、数多くの番号を抽出することは難しいため、そのような場合には、コンピュータで、乱数とよく似た傾向を持つ数字の列を発生させる関数を用いることがよくあります。

標本調査では、単純無作為抽出法などの方法で標本を偏りなく抽出することによって、母集団に比べて少ない数で母集団の傾向を捉えることができます。

練習問題　（解答は P.43 です）

問１　標本調査について述べた次の記述のうち、誤っているものを、次の①～④のうちから一つ選べ。

①　標本調査は、母集団の一部を対象に行われる調査である。

②　母集団から適切に標本を選ぶことによって、母集団の特徴や傾向を予想することができる。

③　標本を選ぶ方法としては、無作為抽出法が望ましい。

④　調査の目的は、標本の特徴や傾向を知ることである。

問２ ある企業の顧客として登録されている人の中から無作為に1.000名を選び、この1,000名に電話をかけて、小学生の子どものいる人600名に子どものお小遣いに関する調査を行った。
このお小遣いの調査で、母集団と標本について述べた次の記述のうち、正しいものを、次の①～④のうちから選べ。

① 母集団は、ある企業に顧客として登録されている人全体であり、標本は電話をかけた1,000名のうち、小学生の子どものいる600名である。

② 母集団は、ある企業に顧客として登録されている人のなかで小学生の子どもを持つ人であり、標本は電話をかけた1,000名のうち小学生の子どもを持つ600名である。

③ 母集団は、ある企業に顧客として登録されている人全体であり、標本は電話をかけた1,000名である。

④ 母集団は、ある企業に顧客として登録されている人のなかで小学生の子どもを持つ人であり、標本は電話をかけた1,000名である。

第 2 部　調査の計画と結果の統計的な解釈

第 1 部ではデータ分析に必要な基礎知識について説明しましたが、ここでは、その基礎知識をもとに、問題を解決するために調査を計画したり、調査結果を統計的に解釈するための方法について学びましょう。

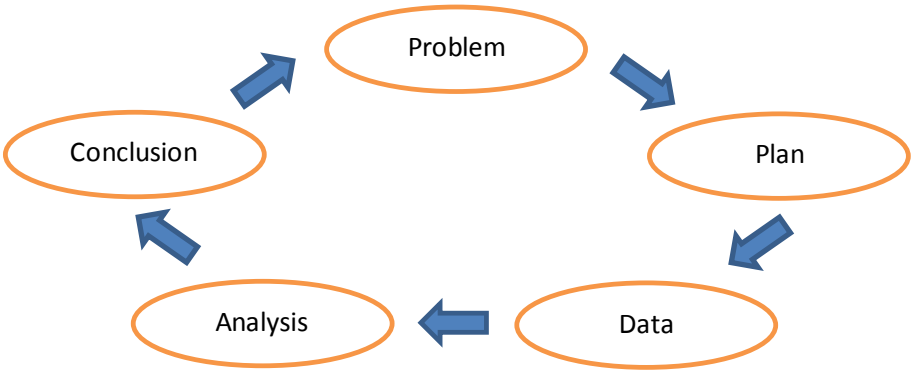
Ⅰ　問題解決のプロセス

1. 統計的問題解決

統計的な分析というと、あらかじめデータが与えられているものと考える人も多いと思いますが、本来は、目的に応じデータを収集するところから始まります。このデータ収集の段階でミスをすると、いくらデータを分析しても本来の目的に対する適切な結果を導くことは難しくなってしまいます。そのため、統計的な問題解決を行う際には、データ分析の知識を身につけるだけではなく、データ収集のための計画やデータ整理の方法なども考える必要があります。

2. PPDAC サイクル

問題の解決に至るプロセスは、必ずしも1回の実験や調査で行われるものではなく、何度も実験や調査を繰り返すなかでより良い結論を得ることが一般的です。そのため、この繰り返し行われる問題解決のプロセスとして、巡回型のプロセスが提案されています。ここでは、その中の一つであるPPDACサイクルを紹介します。PPDACサイクルは下の図のように五つのステップを繰り返し行うものですが、その基礎となったのは、品質管理の分野で用いられてきたPDCAサイクルです。



① Problem　問題の明確化

問題を理解・明確化し、その問題に答えるためにどうすべきか考えます。一般に問題解決のプロセスといっても、ほとんどの場合、最初の段階では問題そのものがそれほど明確になっていません。たとえば、「この勉強法を使えば頭がよくなる」という記述について検討する場合を考えます。このとき「この勉強法」が何を指しているのか、「頭がよくなる」とはどういう意味なのか、という点を明確に定義しなければ、実際に調査を実施する

ことも難しいでしょうし、データを分析した際の解釈も曖昧になってしまう可能性があります。この段階では、ある程度統計的なデータを集めることによって確かめることができるような問題へと集約させていくことが大切です。

② Plan　実験・調査の計画

測定すべきものは何かを考え、設計・記録・収集の方法を考えます。

Problemで明確になった問題に対して、どのように実験や調査を実施するのかを決める段階です。ここでは、誰に対してどのような測定を行うのか、という点が重要です。実験であれば、どのような環境で測定を行うのか、どのような測定方法を用いるのか、ということを考える必要があります。調査票などを用いた調査の場合には、どのような形で質問を行うのか、対象者に対してどのような特性（年齢、性別なども含む）を聞くのか、という点が必要です。対象者の抽出においても、どのような対象者を考え、その対象者をどのように確保するのか、という点を考えておく必要があります。

③ Data　データの収集

データの収集・管理・クリーニングを行います。

Planで策定した計画に基づいて、データの収集を行います。また、データ収集の際に生じる欠測値の問題や回答誤りなどに対しても適切に対応する必要があります。測定値の有効桁数の設定や測定に際して生じる誤りの修正などについても考慮する必要があります。

④ Analysis　データの分析

データを分類し、表やグラフを作成し、パターンをみつけ、仮説を立てます。

収集されたデータについて、集計した結果を表としてまとめたり、グラフを使って表現したりする段階です。もちろん、この段階でも最初に設定した問題を意識しながら、その分析方法について検討する必要があります。

⑤ Conclusion　問題の解決

解釈したり、結論付けたり、新しいアイデアを出したり、コミュニケーションをとったりします。

データの分析結果に基づいて、Problemで考えた問題について判断します。その際には、データの収集の方法や実際の測定の状況等を考慮して解釈する必要があります。また、一つのサイクルだけで問題が解決するとは限りません。問題に対して明確な判断ができない場合には、更に次の問題を考える必要があります。

■ 事例紹介

1）Problem　問題の明確化

学校生活の中での落し物に焦点を当てて、次のような問題を考えます。

学校での落し物が多い、改善することはできないだろうか。

2) Plan　実験・調査の計画

実際にどのような落し物があるのかを把握するためにデータを取る必要があります。落し物は担当の教員に届けられるため、その教員にデータを記録してもらうことにします。記録のための項目、記録用紙の様式など、チェックシートにまとめます。

3) Data　データの収集

作成したチェックシートを担当の教員に渡し、記録をお願いします。一定期間後、その記録用紙を回収し、データを記録します。また、記録用紙の項目にない事項の扱いなどを考えます。

4) Analysis　データの分析

集めたデータを集計し、分析します。たとえば、パレート図にまとめ、どのような落し物が多いのか、落し物の多い場所などの状況を把握します。

5) Conclusion　問題の解決

データの分析結果に基づいて、改善に向けての対策案を探ります。例えば、文具の落し物が多いのであれば授業の終わりに文具の数の確認をする、廊下での落し物が多いのであれば、廊下を走らないようにする、などの対策案を考えてみます。そして、得られた改善案を実際に実行してみて、その効果を探ります。効果の有無は、改善案実施後に同様に調査を行い、まとめて比較してみると分かりやすいかもしれません。

練習問題

（解答は P.43 です）

問 1　次のア～オは、問題解決のサイクルの5つの内容を簡潔に述べたものである。

ア.　データを集計した結果をまとめたり、グラフで表現したりする。
イ.　実験や調査を実施する方法について決定する。
ウ.　漠然としている問題を明確する。
エ.　データを収集する。
オ.　データに基づいて問題を解決したり、問題を再検討したりする。

問題解決のサイクルの順番として正しいものを次の①～④のうちから一つ選べ。

① ウ → エ → ア → オ → イ → ウ
② ウ → イ → エ → ア → オ → ウ
③ イ → ウ → エ → ア → オ → イ
④ イ → エ → ウ → ア → オ → イ

Ⅱ 実験・調査の計画

1. 問題の明確化

前章では、PPDACサイクルについて説明しましたが、ここでは、その中の「Problem問題の明確化」について、更に詳しく考えていきます。

私たちが調査や研究を行うときの最初の段階では、漠然としたアイデアから始まることも往々にしてあります。たとえば、「小さいときにこうしておけば頭がよくなる」とか、「この運動をすると健康になる」というような記述が正しいのか、という問題意識からスタートしたとします。

しかし、これらの記述は、具体的にそれが本当に成り立つかどうかをデータで示すことは困難です。「この運動をする」とはどういうことなのか、「毎日3時間以上する」のか、それとも「週1回1時間程度の運動」でよいのか、というように、運動そのものを定義する必要があるでしょう。また、「健康になる」ということの意味も明確にする必要があります。「治療中の病気がなければ健康」なのか、「メタボリック症候群の疑いがあった場合には健康とみなさない」のかというように、健康をどう定義するのかによって、問題は大きく違ってきます。

それでは、どの程度、問題を明確にすればよいのでしょう。その一つの答えは、その問題に対して、「調査したデータで結論が出せる」というレベルまで問題を具体化することです。この部分が曖昧だと、次のPlanの段階で実験・調査の計画を決めることができません。

その結果、最初にイメージしていた問題をある程度限定したものに変える必要が出てくるかもしれません。例えば、最終の目標として「頭がよい」ことの意味として、人間力や生きるための力というようなものをイメージしていたとしても、実際に測定するためには、ペーパーテストで問うことでできるものに限定することが必要になるかもしれません。

この点に関しては、自分たちで問題解決のサイクルに取り組む場合だけでなく、研究や調査の結果を読む場合においても気をつけておく必要があります。

2. 実験研究と観察研究

統計的な実験・調査は、大きく分けると、実験研究と観察研究に分けることができます。

① 実験研究

実験研究は、対象者にある種の介入を行う研究です。ここで介入とは、たとえば、対象者を二つのグループに分けて、一方のグループには禁煙指導を受けてもらい、もう一方のグループには別の指導を行うというように、ある部分について介入を行うことを想定しています。そのため、介入している部分以外については、二つのグループ間の違いをなるべく小さくする必要があり、対象者の年齢や性別などを合わせるといった工夫を行います。

② 観察研究

観察研究は、対象者に介入を行うことなく、自然の状態を観察する研究です。たとえば、日本の平均寿命を考える場合には、それぞれの人の生死の情報を収集することで求めることができます。また、アンケート調査のように、その時点の対象者の意識や状態を記入してもらうことによって、データを収集する場合もあります。観察研究では、二つの因子の因果関係を考えるときに、原因の部分をコントロールできないため、対象者がなぜそのような選択をしたのか、という点が問題となる場合があります。たとえば、健康教室に通い始めた人は、健康のために通い始めたのか、何らかの病気になったために通い始めたのかによって意味が異なってきます。これらの点は解釈する際に気をつける必要が出てきます。

3. 実験・調査の計画を立てる

最初に考えた問題に対して、実験・調査の計画を立てる際には次の三つを考える必要があります。

① どのような研究方法をとるのか

実験的な研究を行うのか、観察的な研究を行うのかをまず考えます。

実験的な研究であれば、どのような介入を行うのか、どのような条件をコントロールするのかを検討する必要があります。

観察的な研究であれば、1時点での状況を把握するのか、追跡調査を実施するのか、どのくらいの期間追跡するのかなどを検討する必要があります。

② 対象者としてどのような人を選ぶのか

どのような人を対象として選ぶのかということを考えます。高校生を対象とする研究など、研究の目的の中である程度限定される場合もありますが、研究を進める上で更に限定する必要が生じる場合もあります。また、想定している集団を全て調べることが難しい場合には、標本調査を計画する必要も生じます。

③ どのような測定を行うのか

実際に測定するためには、測定の方法を明確にする必要があります。たとえば、「文章を読む速さ」を考えたとき、具体的にどの文章を用いるのか、どのくらいの長さで調査を実施するのかなどを具体的に決める必要があります。

練習問題　（解答は P.44 です）

問 1　「ある食品を摂取することで健康になるかどうか」を調べたい。この問題を明確化するために必要なことを述べた次の①～④のうち、適切でないものを一つ選べ。

①　どの程度食品を摂取するのかを明確に決めることが必要である。

② 食品の摂取方法については、こちらから指示するよりも個人の自由意思に任せた方がよい。
③ 健康かどうかを判断する指標を明確にする必要がある。
④ 健康かどうかを判断する指標を測定する際には、できるだけ条件を揃えておいたほうがよい。

Ⅲ データを解釈する

ここでは、問題の設定やデータの収集方法がデータ分析に及ぼす影響について紹介します。

1. 問題の設定とデータの分析

まず、問題の設定がデータ分析に影響する場合として、次のような例を考えてみましょう。

➢ 例　ある日の気温の変化

下の図は、ある時点の1時間後との気温を幹葉図に表したものです。

8	4 9
9	1 1 4 9
10	2 7 7 7 8 8
11	
12	1 4
13	5 5
14	0 8 9
15	4 5 8
16	5
17	
18	2

左端に1の位までの値を、右側には小数点以下第1位の値を表示しています。

代表値としては、一般に平均値が用いられることが多いですが、この日の平均気温は、12.30℃で、その付近の観測値はあまり多くありません。12℃台を記録しているのは、午前9時と午後8時の2回だけとなっています。また、中央値を計算しても、11.45℃で、その付近の観測値も少なくなっています。

これは、夜間の気温と日中の気温で二つに分かれていることからこのようなことが起こっていると考えられます。

このように1日の気温を考えた場合には、平均値や中央値ではなく、最小値や最大値が生活の中で必要であり、天気予報で最高気温や最低気温が報じられることの意味が分かるでしょう。

2. データの収集法とデータの分析

次に、データの収集方法がデータ分析に影響する場合として、次のような例を考えてみましょう。

➢ 例　スポーツ教室の健康効果

ある保健所で、高齢者の健康の維持を図るために毎週自由参加のスポーツ教室を行っています。その教室の効果を調べるために、年度当初と年度末の2回、教室に参加した人について、体力の変化状況を調べました。

測定値としては、5mの歩行時間を測定し、年度当初からの変化を調べました。分析の方法としては、年度末の歩行時間と年度当初の歩行時間の差の分布を調べて、全体的な傾向を見ることにします。

このような調査では、全体的に歩行時間が長くなっていなければ、健康状態が維持されていると考え、効果があったと判断します。

しかし、この結果を見る場合には、データの収集方法にも気をつける必要があります。分析対象としているのは、年度末と年度当初の2回の測定をともに行った参加者となります。ところが、スポーツ教室は自由参加であるため、年度当初にスポーツ教室に参加した人が全て年度末のスポーツ教室に参加しているわけではありません。もちろん、年度末に参加しなかった人たちが、単に、この日都合が悪かったため参加できなかったのであれば、問題はありませんが、スポーツ教室に参加している期間の途中で体調が悪化したため、参加できなくなった場合には、解釈が難しくなります。要するに、データを測定できた人たちは、その前提として体力が維持されスポーツ教室に参加できたことが条件となります。もし、参加しなかった人たちの測定が可能であれば、その人たちの測定結果は悪い結果となることが予想されます。

このように、どのような方法でデータが収集され、その結果、調査した集団がどのような集団になっているのか、をしっかり把握しておくことが必要になります。

3. 結果の解釈と新しい問題の設定

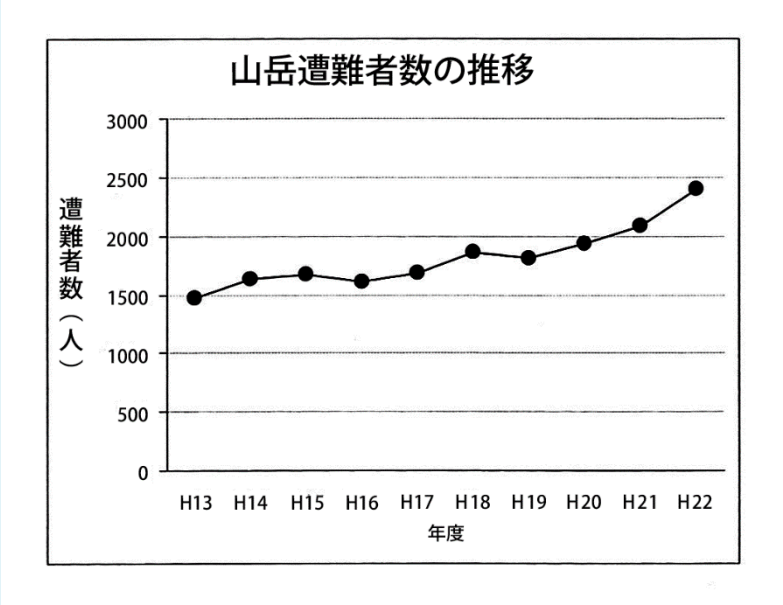
データ分析の結果は、統計的な数値はグラフを解釈するだけではなく、それらの解釈を通して、本来の問題に対する答えが出せるかどうかを検討する必要があります。

たとえば、あるクラスの1週間の読書時間を調べたとします。1か月後に再度調査したら、クラスの平均時間が伸びたことが分かりました。その場合、なぜ平均時間が伸びたのか、なにか対策を講じたことと関係があるのかという新しい問題が浮かび上がります。それを分析するには、次にどのような調査が必要であるかを検討する必要があります。

練習問題

（解答は P.44 です）

問 1　次の図は、平成22年度までの10年間の山岳遭難者の推移を表している。



この資料からもわかるように、この10年間の山岳遭難者数は増加の傾向がみられる。平成18年度以降の60歳以上の遭難者数をみると、次の表のようになっている。

年度	H18	H19	H20	H21	H22
60歳以上の遭難者	909	871	1004	1040	1198

この結果からわかることとして適切なものを、次の①～④のうちから一つ選べ。

① 60歳以上の登山者は遭難する割合が高い。
② 60歳以上の遭難者数は、平成19年度以降だんだん増加している。
③ 遭難者に占める60歳以上の遭難者の割合は年々増加している。
④ 60歳以上の人口が増えているので、60歳以上の登山者数も増えている。

Ⅳ 新聞記事や報告書を読む

1. 私たちの身の回りの統計を探してみよう

私たちの生活の中では、様々な統計データが用いられています。これまでの問題解決のためのプロセスでは、実際に調査を計画するところから、分析し、結論をまとめるところまでを考えてきました。もちろん、実際にこのプロセスを行うことができるようになることは重要ですが、それと同時に新聞記事や報告書等を調べて、そこから正しく情報を把握できるようになることも重要です。

2. 読む際のポイント

ここでは、新聞記事や報告書を読む際に気をつけるべきポイントをまとめておきましょう。

① 記事の基になっているものは何か

統計データに基づいた新聞記事は、新聞社自身が調査を行っている場合もありますが、多くは何らかの調査研究の結果に基づいて記事が書かれています。そのため、どのような調査に基づいて記事が書かれているのかについて、まず調べましょう。

② 調査の実施者は誰か

新聞記事の基となった調査研究を実施している調査者は、誰なのか、どのような立場で調査を行っているのかを確認しましょう。

調査の実施者が必ずしも中立的な立場であるとは限りません。調査実施者はある目的をもって、それぞれの立場で調査を行っています。もちろん、自分たちの問題意識に基づいて、その根拠となるデータを集めることが目的ですが、しっかりした調査者であれば、調査結果の信ぴょう性を確保するために、調査計画段階で公平な計画を立てているでしょうし、その計画を公表しているでしょう。

③ 調査の対象者をどのように選択したのか

調査対象者を選択する方法は、データの分析結果に大きな影響を与えます。そのため、標本調査であれば、その抽出方法や調査の対象を確認することが大切です。また、抽出方法だけではなく、回答を拒否した人の割合や回答拒否の影響が検討されているかなども確認する必要があります。

報告書の場合には、調査の目的に関するデータだけではなく、年齢や性別などの属性の分布データも公表されていることが多いため、その分布を見ることによって、調査に回答した集団が偏ったものになっていないかを確認することができます。

④ どのように測定されたのか

研究の目的に合わせて、測定方法についても検討する必要があります。測定方法によって結論が変わる場合もあります。特に調査票による調査や面接による調査では、どのように問いかけたのかによって、回答が異なる場合があります。

たとえば、「あなたの支持している政党はどこですか」という問いに対して回答してもらう場合、「あなたの支持している政党は、強いていえば、どこですか」と聞くことによって、それぞれの政党の支持率は上がる可能性があるでしょう。

⑤ 比較している場合どのようなグループの比較か

統計的な実験によって、ある方法の効果を調べる場合には、グループ間での比較が必要になりますが、その場合、比較する集団の違いを把握することが重要です。新聞記事に詳細が触れられていない場合には、基になっている調査の報告書等に当たってみることも必要です。また、グループ間の違いが、その他の因子についても生じていないかどうかを確認しましょう。

解答と解説

■練習問題　データの分布をみる　（問題は P.10）

問１　③ Bのみ正しい

Aは第2四分位数が12冊なので、借り出した本の冊数が12冊以下である児童が半数以上いることになるから、間違い。またBは同様に考え、正しいことが分かる。このことから③が正解。

問２　② ⅠとⅡのみ正しい

実際に度数分布における平均値や範囲、分散を求めてもよいが、定義からも平均値や範囲が等しいこと、また分散はAの方が大きいことが分かる。したがって、②が正解。

■練習問題　観測値の標準化と外れ値　（問題は P.13）

問１　③ B→C→Aの順

Cさんの点数は与えられた情報より、中央値、第2四分位数と等しいため、B→Cの順である。またCさんの点数は平均値であることから標準化すると0になり、Aさんの点数は標準化すると1となるため、C→Aの順である。すなわちB→C→Aとなる。したがって③が正解。

問２　③

はずれ値を第3四分位数+1.5×四分位範囲で確認すると、四分位範囲は64−48=13分より、61+1.5×13=80.5分となるため、90と98ははずれ値となる。したがって、大きい方のひげの端は78分となる。また最小値も同様に考え、はずれ値はないため、小さい方のひげの端は29分となる。これらのことから、③の箱ひげ図が適切である。

■練習問題　相関と散布図、相関係数　（問題は P.19）

問１　①

すべての人が中間試験の点数を+20=期末試験の点数となるため、散布図で中間試験と期末試験の点数を書くと右上がりの直線となる。したがって定義から正の相関関係といえる。したがって、解答は①。

問２　④ (1), (2), (3)の相関係数は同じになる

相関係数は定義より測定の単位の影響を受けず、また横軸、縦軸を入れ替えても変わらない。したがって両方の記述は間違っているので④が正解。

■練習問題　確率の基本的な性質、反復試行と条件付き確率（問題は P.27）

問１　③　0.3

50枚のカードは同じ確率で選ばれると仮定すると、青いカードは15枚で、全体は50枚であるから、確率は0.3となるので、③が答えとなる。

問２　③　$\frac{3}{7}$

まず、喫煙者で病気にかかる確率を求めると、$0.2\times0.003=0.0006$となる。非喫煙者で病気にかかる確率は、同様に$0.8\times0.001=0.0008$となる。よって、トータルで病気にかかる確率は$0.0006+0.0008=0.0014$となる。病気にかかったという条件の下で、喫煙者である確率は、$\frac{0.0006}{0.0014}=\frac{3}{7}$ となる。よって、③が正解である。

■練習問題　標本調査（問題は P.29）

問１　④

標本調査は、母数団の一部を対象に行われる調査である。①は適切である。標本が適切に選ばれれば、推定は偏りなくできるので②も適切である。標本を選ぶ際には、偏りを避けるために無作為抽出が望ましい。調査の目的は標本の特徴をつかむことではなく、母集団の特徴や傾向を知ることであるので、④は適切ではない。

問２　②

電話をかけたのはある企業に顧客として登録されている人であるが、小学生の子どもがいない人は調査から除外されているので、ここでの母集団は、「ある企業に顧客として登録されていて小学生の子どもがいる人」全体であるが、標本は、電話をかけた中で小学生の子どもがいる600名となるので、②が適切である。

■練習問題　問題解決のプロセス（問題は P.33）

問１　②　ウ → イ → エ → ア → オ → ウ

アはデータの解析(Analysis)、イは実験・調査の計画(Plan)、ウは問題の明確化(Problem)、エはデータの収集(Data)、オは問題の解決(Conclusion)を表しており、問題解決のサイクルは、

問題の明確化 → 実験・調査の計画 → データの収集 → データの解析 → 課題の解決

の順番で進むので、②が正しい。

■練習問題　実験・調査の計画（問題は P.35）

問 1　②

　食品の摂取方法を原因と個人の自由意思で決定すると、その時の健康状態によって摂取方法が異なることも考えられるため、できるだけ食品の摂取方法については研究実施者の方で割り当てたほうがよいので、②が誤りである。

■練習問題　データを解釈する（問題は P.39）

問 1　②

　①については、60歳以上の登山者が遭難する割合を調べるには、60歳以上の登山者数や60歳未満の登山者数も必要である。③については、遭難者数も増加しているため、必ずしも60歳以上の遭難者の割合が高くなっているとは限らない(実際には、H20が一番高い)。④については、60歳以上の登山者数がわからないので、このデータからは分からない。②については上の表から判断することができるので、答えは②である。