# 統計検定における考え方と手順

田中 美里　廣安 知之　日和 悟

2015年11月11日

IS Report No. 2015111102

IS Report

Medical Information
System Laboratory

**Abstract**

本レポートは統計検定を少し使用した経験がある学生向けに統計処理の必要性や流れについて説明したレポートである．なお，執筆に際して，市原清志氏の『バイオサイエンスの統計学』（南江堂出版, 2008 年）を参考にした．
**キーワード:** 統計, 検定

# 目 次

# 第 1 章　統計の持つ役割

## 1.1　情報の推定と仮説検定

一般に，統計技術は観測された情報の推定や解析に用いられる．

たとえば，あるパン屋が「1kg のパン」を販売していた場合，店で焼かれたパンは常に 1kg ではなく，ある時は 980g，またある時は 1010g となり，そこには必ずばらつきが存在する．それらのデータ（標本）の中から，店で販売されているものが「1kg のパン」と定量的に言えるかどうかを確かめるため，パンの重さを繰り返し観測して Fig. 1.1 に示すように分布をとり，平均値などの統計量を算出する技法が統計技術である [1]．

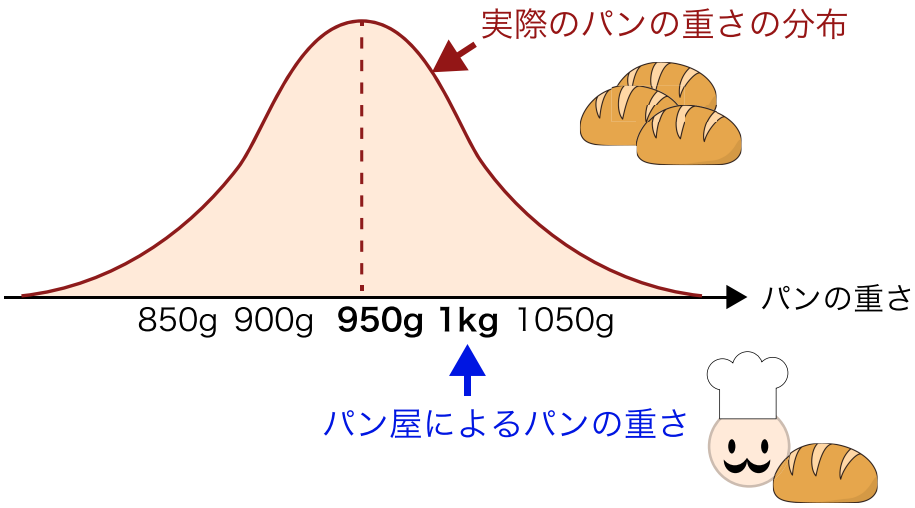


Fig. 1.1　パンの重さの分布から見るパンの平均の重さ（自作）

統計が主に用いられるのは，観測された少数の情報から全体の分布や特徴量を推し量る推定や，または複数の異なる分布のデータが得られたときにそれらに明確な違いがあるかどうかを確かめる検定である．

### 1.1.1　推定／統計的推定

標本（実際に得られたデータ）の持つ統計量から，母集団（対象となる全てのデータ）の特徴量を推し量ることを統計的推定という．この場合，統計量とは平均値や，分散値などを指す．先のパンの重さの例で言えば，母集団はパン屋の焼いた全てのパンの重さのデータであり，標本はその中で購入されて測定されたパンの重さのデータである．

[1] フランスの数学者ポアンカレ（1854-1912，位相幾何学における 3 次元球面の特徴に関するポアンカレ予想を立てた）が，ある店で販売された「1kg のパン」が実際には「950g のパン」であることを証明したのは有名な話である．彼は 1 年に渡って「1kg のパン」を購入し，重さを量り続けた．そして，そのデータからパンの重さの分布を求め，分布の平均が 950g であることを確かめた．

### 1.1.2 検定／統計的仮説検定

実際に得られた標本を基準に，母集団の特徴や状態について何らかの仮説を設け，その妥当性を確率論的に検証することを統計的仮説検定という．たとえば，産地 A と産地 B のどちらのリンゴがより甘いかを確かめるとする．このとき，甘さは糖度によって決められる．糖度を比較するにあたって，産地 A と B でとれた全てのリンゴに対して糖度を測定するわけにはいかない．よって比較用に幾つかリンゴをサンプリングし，その糖度データを標本として計算を行うことで，産地 A と B でとれた全てのリンゴの糖度データ（母集団）を推定して，その差を比較することができる．

## 1.2 情報の構造解析

統計技術は情報の基本構造を解析することにも応用されている．これは，標本から得られた情報を数量的に要約する，および分類することで，見易い，構造的なデータへと変換することができるためである．この主な例が，主成分分析 (Principal Component Analysis: PCA) や判別分析 (Linear Discriminant Analysis: LDA) などの多変量解析の技術である．複数の評価軸を持つ多次元データが存在する場合，これらのデータを一括に比較することは難しい．よって，類似する評価軸をまとめあげるなどの統計的な処理を行うことによって，見易く，比較しやすいデータへと可視化することが出来るようになる．

このように統計には複数の役割が存在する．この中で本レポートでは「統計的仮説検定」に着目し，これを行うにあたって必要な技術や知識について整理する．

# 第 2 章　統計的仮説検定

## 2.1　統計的仮説検定とは

ビジネスや研究においては，あるデータとあるデータの間に差があることを，十分な説得力を持つ形で示さなければならない状況が多く存在する．たとえば，ある研究において提案した手法 A が，既存手法 B よりも良いことを示したいとする．この場合，実験や調査を行って実際に得られた手法 A の結果と手法 B の結果に対し，手法 A の結果の方が手法 B の結果を上回ることを「手法 A の方が良いように見える」という実験者の主観ではなく,「手法 A と手法 B の間には有意な差がある」という定量的な表現によって示さなければならない．この「有意な差がある」，または「有意な差がない」ことを求めるのが，統計的仮説検定である．

本説では統計的仮説検定について詳しい説明を行う前に，予備知識として仮説や有意水準，要因と水準について説明する．

## 2.2　対立仮説と帰無仮説

検定を行うには対立仮説と帰無仮説が必要になる．Fig. 2.1 に対立仮説と帰無仮説の例を示す．帰無仮説は $H_0$，対立仮説は $H_1$ と表現される．両仮説は互いに補足し合う関係になっており，一般的には，帰無仮説を棄却（却下）することによって，その対立仮説が正しいことを推定する．この考えは背理法に基づいており，たとえばある 2 群のデータに「差がある」ことを証明したい場合，2 群間に「差がない」という仮説を立て，その仮説に矛盾を見いだすことで,「差がある」という仮説を採択する．よって，対立仮説と帰無仮説は，両者の間に中間的な条件が一切存在せず，排中律を満たしていることが必要である．よって Fig. 2.1 に示すように「差がない」仮説と「差がある」仮説の両者を合わせれば，全ての状況が説明できることになる．

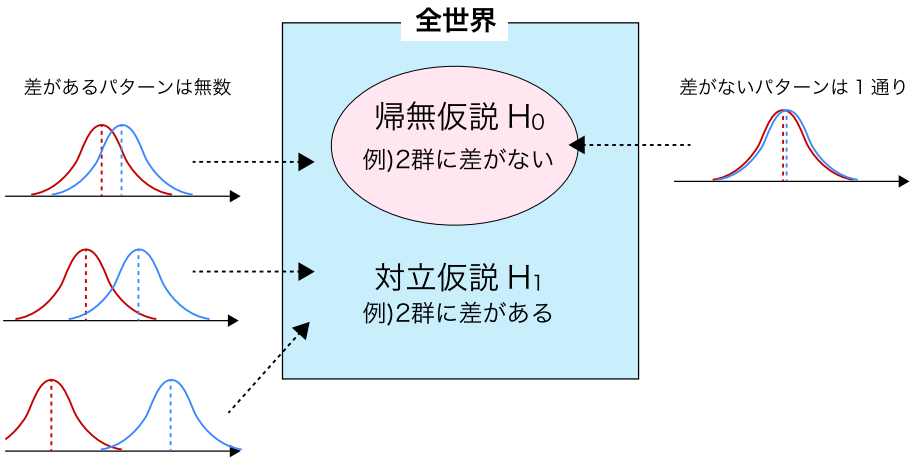


Fig. 2.1　帰無仮説と対立仮説の関係（自作）

一般的に帰無仮説と対立仮説の内容は，帰無仮説の方がより確かめられやすい仮説となるように設定される．たとえば「差がある」仮説と「差がない」仮説では,「差がない」仮説の方が検討しやすい．これは 2 群のデータに「差がない」という状況は 1 通りしかないが，2 群のデータに「差がある」という状況は，Fig. 2.1 に示すように「どのくらい差があるか？」,「どのように差があるか？」によって無数の状況が存在し得ることに基づく．無数に存在する状況の全てが成立するか否かを検証することは不可能であるため,「差がない」仮説が帰無仮説として設定される．この帰無仮説に基づいて統計量を計算し，その統計量が仮説を正しいとするには満たない場合，帰無仮説は棄却される．帰無仮説が棄却されれば,「差がないとは言えない」ことになり，対立仮説である 2 群のデータ間に「差がある」と表現可能ないずれかの状況が発生していることになる．

帰無仮説は棄却されることはあるが，採択されることはない．すなわち，帰無仮説が棄却されれば対立仮説が採用されるが，棄却されない場合に帰無仮説が正しいと証明されるわけではない．これについては次節で述べる有意水準の考え方が関わってくる．

## 2.3 有意水準

前節で述べた帰無仮説の矛盾をどの位厳しく検証するかをコントロールするのが有意水準 (Level of significance) である．帰無仮説の検証では，この仮説に基づいて統計量が計算される．なお，詳細な計算方法については後述する．この対象となる統計量は $p$ 値と呼ばれ，この $p$ 値を有意水準と比較することによって，仮説の妥当性を検証する．

$p$ 値を分かりやすく説明すると「帰無仮説が成立する確率」であり，より厳密に定義すると与えられたデータに対して「帰無仮説が本当は正しい場合に，帰無仮説が正しくないと却下される確率」である．$p$ 値は小さければ小さいほど，帰無仮説を棄却しやすい．とくに前者の定義から捉えると，$p$ 値が有意水準を下回る場合，有意水準で示された確率よりも帰無仮説が成立する可能性が低い，という流れで帰無仮説を却下できる．

後者の厳密な定義について説明する．たとえば 2 群のデータの差を検定する場合，与えられたデータから計算された $p$ 値は,「このデータの分布から互いの母集団を推定した結果，本当は母集団同士には差がない（帰無仮説）のに，誤ってその帰無仮説が却下されて母集団同士に差がある（対立仮説）と判断されてしまう」事象が確率 $p$ で発生するということを示す．統計的仮説検定は，あくまで母集団を推定するため，その推定には誤差が存在し得る．よって，その誤差を踏まえて誤って仮説が棄却，採用される可能性を計算し，その過ちが発生する確率 $p$ が低いデータであれば，帰無仮説を安心して棄却できる，ということになる．逆に確率 $p$ が高ければ，過った判断が発生する確率が高く,「差がない」仮説を棄却はしない方が望ましいと判断する．

このように，有意水準は帰無仮説の棄却が誤って発生し得る確率を示す値であり，帰無仮説が正しい確率そのものではない．このために，$p$ 値が高かったとしても，帰無仮説が正しいことを証明するものではない．よって，帰無仮説は採択されずに判断が保留され,「差があるとはいえない」という表現をされるに留まることに注意する．

## 2.4 要因と水準

統計仮説検定では，要因やカテゴリ，水準という表現がよく用いられる．これらの定義を以下に示す．

- 要因／因子（Factor）
  調べたいデータの変動を説明するための基準，および変数のことをいう．
  例）ある音楽を聴いたときの作業量の違いを見たい．このとき，作業量の違いを生み出す要因としては以下が考えられる．

  **要因 1** 音楽の種類

  **要因 2** 性別

  **要因 3** 年齢

- カテゴリ（Category）／属性（Attribute）／クラス（Class）
  要因内部での条件分けにおいて，定性的な条件で分類する場合に用いられる．

  **例 1** 病気の分類：疾患 A ／疾患 B ／疾患 C

  **例 2** 与える肥料の種類：窒素／リン酸／カリウム

- 水準（Level）
  要因内部での条件分けが，順序尺度[1] 的に行われている場合に用いられる．

  **例 1** 疾患の重症度：重症／中等度／軽症

  **例 2** 与える肥料の量：1g ／ 3g ／ 5g

[1] ある量的な特性の順序関係によって分類する尺度

# 第 3 章　2 群のデータの比較

## 3.1　検定手法の選択とその基準

与えられた 2 群のデータにおいて，平均値の差があるかを検討する場合，その検定の手順は Fig. 3.1 に示されるようなフローチャートに従う．この検定フローでは，2 群の対応の有無，分布の正規性，等分散性によって条件が分岐する．これらの分岐の意味するところを次節より説明していく．なお，本レポートでは各検定手法の詳細については，他の資料，文献に譲る．

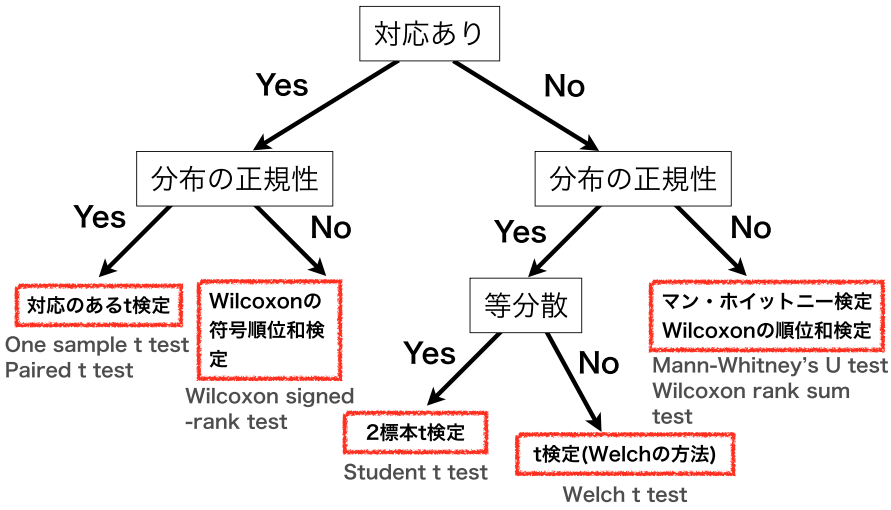


Fig. 3.1　2 群の差における検定手順（自作）

## 3.2　対応の有無について

検定における「対応の有無」とは，与えられたデータが，同一個体から観測されたデータであるか否かを意味している．以下に具体的な例を交えた説明を述べる．

- 対応のある 2 群
  関連 2 群ともいう．同一個体で 2 条件を比較する場合を示す．

  **例 1**　同じ被験者にある薬を投与してダイエットをさせる．時期を分けて，ある時期は薬 A を，別の時期は薬 B を飲ませて各時期の減量具合を見る．

  **例 2**　同じ被験者に快刺激と不快刺激を見せて，その脳活動の差を見る．

- 対応のない 2 群
  独立 2 群ともいう．異なる個体で 2 条件を比較する場合を指す．

**例 1** ある薬 A を飲みながらダイエットをする被験者群と別の薬 B を飲みながらダイエットする被験者群について，それぞれの減量具合を見る．

**例 2** 被験者を年齢によって分ける．高齢者群と若年者群にわけて，それぞれの認知タスクの課題成績を比較する．

対応の有無を考慮するのは，それによって用いられる検定法が異なり，即ち $p$ 値の計算のされ方が異なるためである．例として，あるデータにおいて 2 群に差があることを証明したいケースを挙げる．同じデータに対して，対応のある t 検定と対応のない t 検定で計算を行った結果を下記に示す．

- 対応のない t 検定（2 標本 t 検定，Student t test）
  Fig. 3.2 の左に対応のない t 検定のイメージを示す．サンプル数 $n_1$，$n_2$ の 2 群について，群毎に平均値 $\bar{x_1}$ と $\bar{x_2}$，分散 $s_1$ と $s_2$ を求め，そして合併分散 $s$ を算出する．この場合，帰無仮説 $(H_0)$ はこの $\bar{x_1} - \bar{x_2}$ が 0 ということになる．

  $t$ 値の計算は式 3.1 によって行われる．$p$ 値はこの $t$ 値を変換することによって得られ，t 値が大きければ大きいほど $p$ 値は小さくなる傾向があると考えてよい．

$$t = \frac{\bar{x_1} - \bar{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.1}$$

$$s = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \tag{3.2}$$

- 対応のある t 検定（One sample t test ／ Paired t test）
  Fig. 3.2 の右に対応のある t 検定のイメージを示す．データの分布そのものは左の対応のない t 検定と同じである．n 組（今回は 5 組）のデータに対して，各ペアの差 $d$ を求め，平均 $d_m$ と標準偏差 $d_s$ を出す．帰無仮説 $(H_0)$ は，この $d_m$ が 0 である，ということになる．このとき，t 値の計算は式 3.3 によって求められる．

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{3.3}$$

  対応のない t 検定と比べて，t 値は大きな値となり，$p < 0.01$ において有意差が確認される．

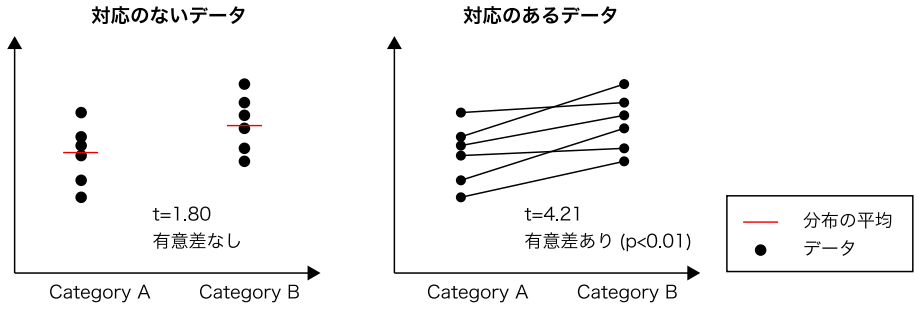


Fig. 3.2　対応の有無による差（自作）

このように両者は計算方法が全く異なり，数値的には同じデータ群に対して全く異なる結果を算出する．多くの場合，対応ありの検定は，対応なしの検定より強力な検定法である可能性が高い．特に，

(1) 条件の間に差はあるが個体差が大きい，(2) 個体差は小さいが条件間の差も小さい，という条件下では対応ありの検定手法の方が有意差を得やすい．

一方で，対応ありの検定は，両方のサンプルが同じ分散（後で説明）を持つ必要がない．対応ありの場合，どちらかの群にデータの欠損があると，別の群の対応するサンプルについても検定から外さなければならない．実験によって対応のありなしを適切に使い分ける必要がある．

## 3.3 分布の正規性

### 3.3.1 パラメトリック検定とノンパラメトリック検定

対応の有無と同様に，分布の正規性についても，それを満たすか否かによって用いられる検定法が異なる．即ち，$p$ 値を出す計算式が，その分布が正規分布であることを前提にしているのか否かによって異なるものとなる．

正規分布に従う（もしくは別の特定の分布に従う）データに対してはパラメトリック検定法，満たさない分布においてはノンパラメトリック検定法が用いられる．両者の説明を以下に示す．

- パラメトリック検定 (parametric test)
  母集団の分布型に仮定があるときに用いられる検定法である．分布としては正規分布，2 項分布，$\chi$ 二乗分布などが挙げられる．
  - 母集団が正規分布であることを前提とする検定法.
    例）t 検定，分散分析
  - 母集団が 2 項分布であることを前提とする検定法.
    例）比率の検定（事象 A と事象 A ’の生起確率がどれだけ偏っているか）
  - 母集団が $\chi$ 二乗分布であることを前提とする検定法.
    例）$\chi$ 二乗検定
- ノンパラメトリック検定 (non-parametric test)
  母集団の分布に仮定が無いときに用いられる検定法である．
  例）Wilcoxon 検定，Mann-Whitney 検定...etc.

基本的に何らかの分布に従うデータであれば，パラメトリック検定の方が有意差は出やすいとされる．

### 3.3.2 分布の修正

先に述べた通り，基本的にはパラメトリック検定の方が有意差を得やすい．よって，正規分布に従わないデータに対して，正規分布へと補正を行うことがある．

対応ありの 2 群の場合，各組合せのデータの差 $d$ の分布が正規分布であれば，対応ありの t 検定に移行できる．これについては式 3.3 を参照すると理解できる．

対応のない 2 群の場合，データを変数変換する必要がある．これにはべき上変換や対数変換が用いられる．Fig. 3.3 にその例を示す．図の左側には偏りのある分布の例が示されている．これらの分布に対して，矢印の上に指定された数式を用いて変数変換することで，図の右側に示すような正規分布

へと変換することができる．例えば，左上の右に偏った分布に対しては，ルートをとることで正規分布に近づけることができる．

ただし，これで変換が可能なのは，分布が単峰性であり，かつ裾広がりが左右どちらか一方の場合のみである．補正がきかなければ，やはりノンパラメトリック検定を用いることになる．

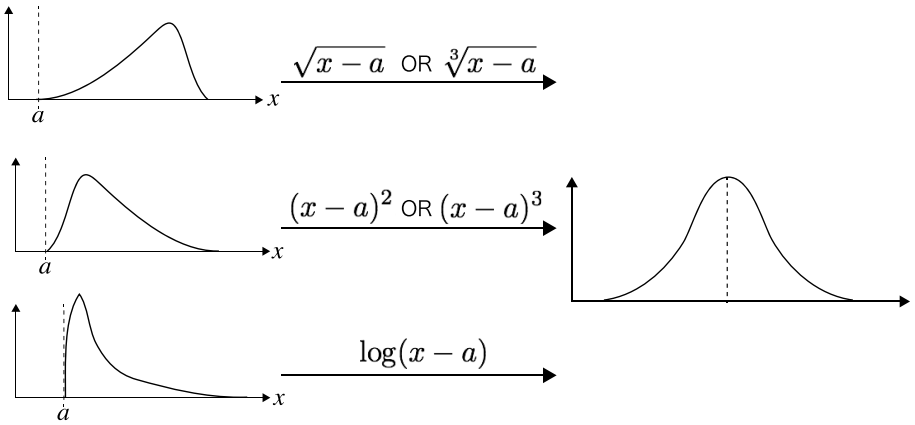


Fig. 3.3　分布の変数変換による正規化（自作）

### 3.3.3　正規性の検定

正規性の検定は,「正規分布である」という帰無仮説を棄却しないことで正規性を証明する（「正規分布ではないとは言えない」ことを証明する）という手順をとる．従って，棄却することによって 2 群の差などを証明する有意差検定と異なるので注意する．下記に，用いられる検定手法の例を示す．

- コルモゴロフ・スミルノフ検定（Kolmogorov-Smirnov test）
- シャピロ・ウィルク検定（Shapiro-Wilk test）
- $\chi$ 2 検定
- 歪度（わいど）と尖度（せんど）による正規性の検定

一般に，サンプルサイズが小さい場合はシャピロ・ウィルク，大きい場合はコルモゴロフ・スミルノフを用いるのが良いと言われている．

## 3.4　等分散性

等分散性は分布の分散が等しいことをいう．これも，検定手法によっては同じ広さの分布を仮定して $p$ 値の計算・比較が行われるため，予め確かめられる必要性がある．

2 標本 t 検定など対応のない 2 群の差の検定では，ほぼ同じ分散を持つ 2 つの分布がどのくらい重なっているかを計算している．式 3.1 より，2 標本 t 検定では，2 群の分散値をまとめて合併分散とした上で，t 値の計算を行っている．従って，各群が極端に異なる分散を持つ場合，t 値の計算が正しく行われない可能性がある．

式 3.4 に等分散性を満たさないときに用いられる t 検定である welch t test の計算式を示す．各変数は式 3.1 の説明を参照されたい．このように分散値は各群のものを使用して，t 値が算出される．

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.4}$$

なお，対応のある 2 群の場合，正規分布であればこの等分散性は気にしなくてよいとされている．

# 第 4 章　まとめ

本レポートには，主に 2 群の差を例題として挙げながら，統計仮説検定における手順と，何故その手順が必要とされるかについて説明した．仮説検定には 2 群の差だけでなく，多群の検定や挙げたものとは異なる分布を扱うものなど，様々な手法がある．本レポートで得た知識をベースに，様々な手法への理解が深めていくことが望ましい．