「大学における数理・データサイエンス教育の全国展開」

# データサイエンス教育に関する
# スキルセット及び学修目標

第１次報告

（リテラシーレベル）

令和元年（２０１９年）１１月２１日

数理・データサイエンス教育強化拠点コンソーシアム

カリキュラム分科会

# 要　　旨

## １　作成の背景と経緯

○　数理・データサイエンス教育強化拠点コンソーシアムでは、「大学の数理・データサイエンス教育強化方策について」（平成 28 年 12 月 数理及びデータサイエンス教育の強化に関する懇談会）を受け、文理を問わず、全国的なモデルとなる標準カリキュラムを検討。

○　本報告は、リテラシーレベルのモデルカリキュラム検討の基礎とする観点から、カリキュラム分科会の第 1 次報告として、リテラシーレベルのスキルセット及び学修目標を整理したもの。

○　コンソーシアムでは、今後、本報告を活用しつつ、産業界や公・私立大学等関係者の参画を得てリテラシーレベルのモデルカリキュラムの検討を継続。

## ２　スキルセット及び学修目標（第１次報告）の性格と役割

○　第 1 次報告は、習得しておくことが望ましい標準的なスキルを網羅的に示したもの。各大学が一律に整備することを意図したものではなく、各大学の教育目的、分野の特性、個々の学生の学習歴等に応じて、必要に応じて上位互換や既習確認等を行いながら、適切かつ柔軟に選択・抽出されることを想定。

○　高校までの学習内容にばらつきがあることを踏まえ、高校での学習内容（数学、情報）を含め高等学校学習指導要領との対応を図るなど、高大接続を考慮。

○　各大学の学生の特性や文系・理系の違い考慮し、データサイエンスを学ぶ意義、重要性の理解や学修の動機付け等を図るために、大項目として「データサイエンスを学ぶ意義」を設定。

○　本報告では、概ね 1.5 単位分程度と考えられる内容を特にコアとして推奨。なお、各大学において「データサイエンス入門」のような科目を設置する場合には、「データサイエンスを学ぶ意義」及びコアに各大学の事情に応じて内容を追加することで、2 単位や 4 単位の講義を設計できるなど、一定の自由度を持つ。

○　本報告を活用し、いくつかの大学の協力を得て実践モデルの試行、グッドプラクティスの紹介等に取り組むことで、データサイエンス教育の全国への普及・展開を促進。

## ３　各大学におけるカリキュラムの実装にあたっての期待

○　「なぜ数理・データサイエンス・ＡＩを学ぶのか」、「社会でどのように活用されているのか」など、学修の動機付けが重要。その際、実際の事例を題材としたケーススタディ、グループワーク等を効果的に取り入れるなど、単に座学としてではなく、各大学・教員が、何らかの形で現実問題に触れるような機会を作るように工夫することにより、学生の興味や関心を喚起しつつ、実務・実践で活用できる能力を

育成することを期待。

## ４　スキルセット及び学修目標（第１次報告）の構成

○　「データサイエンスを学ぶ意義」、「データの法規・倫理」、「データの記述・可視化」、「データの取得・管理・加工」、「統計基礎」、「数学基礎」、「計算基礎」、「モデリングと評価」の大分類、その下に中分類・小分類を置く階層構造。

○　小分類を単位として、スキルセットと学修目標を整理。スキルセットは、データサイエンスの習得すべきスキル・理解すべき概念のセット。学修目標は、スキルや概念のイメージを文章でまとめたもの。

○　大分類等に対応させて科目設計を行うことは必ずしも意図していない。スキルセットと学修目標を参照し、各大学の教育の特徴を活かしつつ、教育資源等も考慮しながら、各大学において、科目編成やシラバス作成等を行うことを想定。

**大分類・中分類・小分類の構成（リテラシーレベル）**　　☆はコアの小分類、★はコア以外の小分類を表す。

| 大分類 | 中分類 | 小分類 |
|---|---|---|
| データサイエンスを学ぶ意義 | (データサイエンスを学ぶ意義、重要性の理解)<br>データ駆動型社会　データに基づく課題解決、意志決定の支援　データサイエンスのサイクル　データサイエンスの様々な事例 | |
| データの法規・倫理 | 情報倫理、情報セキュリティ | ★情報倫理・関連法規　★情報セキュリティ |
| | データに関連する法律・規制 | ☆個人のデータに関連する法規　☆統計法 |
| | データサイエンスの倫理 | ☆倫理に配慮したデータ収集と利活用　☆データの匿名化<br>☆データサイエンスに関する様々なバイアス　★公表バイアス、出版バイアス |
| データの記述・可視化 | データの記述 | ☆種々のデータ　☆基本統計量　☆相関関係 |
| | データの可視化 | ☆グラフの構成要素　☆統計グラフ　☆分布の統計グラフ |
| データの取得・管理・加工 | データ取得とオープンデータ | ☆日本や世界のオープンデータ　☆オープンデータの取得 |
| | データ管理とデータ形式 | ☆表形式のデータ　★その他のデータ形式　★データベース |
| | データの前処理 | ★データクレンジング　★外れ値、異常値、欠損値<br>★データ加工 |
| 統計基礎 | 確率と確率分布 | ☆場合の数、順列・組み合わせ　☆確率　☆確率分布、確率変数　☆主要な確率分布 |
| | データ収集法と確率構造 | ☆標本調査　★ランダム化比較試験 |
| | 統計的推測 | ★統計的モデル　★標本分布　★点推定、区間推定<br>★仮説検定 |
| 数学基礎 | 線形代数 | ☆ベクトル　★行列 |
| | 微積分 | ☆多項式関数　★初等関数　★1 変数関数の微分法<br>★1 変数関数の積分法　★2 変数関数の微分法<br>★2 変数関数の積分法 |
| | 数列 | ☆データと数列　★数列 |
| | 演習 | ★線形代数演習　★微積分演習 |
| 計算基礎 | 情報、コンピュータの仕組み | ★数と表現　★有効数字、計算の誤差　★データ量の単位<br>★デジタル化　★文字の表現　★集合と命題　★論理演算 |
| | データ構造 | ★配列、リスト　★連想配列 |
| | アルゴリズムとプログラミング | ★アルゴリズム　★プログラミング |
| モデリングと評価 | 教師あり学習 | ★回帰分析　★ロジスティック回帰 |
| | 教師なし学習 | ★クラスタリング（k-平均法）　★階層クラスタリング |
| | モデルの評価 | ★評価指標　★訓練データとテストデータ |

# まえがき

情報通信・計測技術の飛躍的発展によって、従来とは質・量ともに全く異なるビッグデータが得られるようになった。ビッグデータの活用領域は予測、意思決定、異常検出、自動化、最適化など多岐に亘り急速に拡大しており、自動運転、画像認識、医療診断、防犯、コンピュータゲームなどデータの活用が従来の社会常識を大きく変えつつある例は枚挙に暇がない。企業の株価総額ランキングの急激な変化が示唆するように、経済成長の要因も従来の労働力・資本・技術革新からデータやデータから価値を生み出す技術へ大きくシフトしている。その結果、数理・データサイエンス・ＡＩなどのサイバー世界の情報処理技術が今後の社会発展の鍵となっている。

このようなデータ駆動型社会への転換を推進していくためには、データから価値を創出できる専門人材であるデータサイエンティストの育成が不可欠であるが、現状ではデータサイエンティストは圧倒的に不足しており、世界的にデータサイエンティストの育成が急速に進められている。専門人材の育成と同時に国民全体のリテラシーレベルの向上も不可欠である。我が国の近代化実現において国民の高いリテラシーが果たした役割はしばしば指摘されているが、データ駆動型社会への転換においては、データリテラシーをすべての国民が備えるべき新たな素養としてその涵養に努める必要がある。

しかしながら、急速にデータサイエンスの教育体制整備が行われた欧米先進国と異なり、我が国ではその教育体制の整備が著しく立ち遅れている状況が続いてきた。このような状況を早急に解消しデータサイエンス教育の普及を目指して、2017 年度に 6 大学が数理・データサイエンス教育強化拠点校として文部科学省に選定され、それぞれの大学において全学データサイエンス教育の普及に向けた取り組みを実施してきた。この取り組みの重要な点は、単に自大学での数理・データサイエンス教育の充実を実現するだけでなく、分野を問わず全ての学生を想定して、学生が習得すべき標準的な内容を示したカリキュラム、教材、教育用のデータベースの整備を目指していることにある。そのために、6 拠点校はコンソーシアムを形成して、3 つの分科会のもとで全国の大学へのデータサイエンス教育の普及に向けた活動を行ってきた。さらに、2019 年度からは新たに 20 大学が協力校として加わり、全大学・高専への普及を加速する活動が開始されている。

今回の報告書は、分野を問わず全国の全ての大学生・高専生を想定したリテラシーレベルのモデルカリキュラム検討の基礎とする観点から、コンソーシアム カリキュラム分科会の第 1 次報告として、リテラシーレベルのスキルセットと学修目標を整理したものである。本報告の作成過程においては、文部科学省、経済産業省、データサイエンス関係団体、放送大学及びデータサイエンス教育について先駆的な取組を有する公・私立大学等との意見交換を行い貴重な意見を頂戴した。コンソーシアムでは、引き続き産業界や公・私立大学等関係者の参画を得てリテラシーレベルのモデルカリキュラムの検討を進めるとともに、より上位のレベルのモデルカリキュラムの公開も行う予定である。

数理・データサイエンス教育強化拠点コンソーシアム カリキュラム分科会

# 目　　次

# Ⅰ　スキルセット及び学修目標（第１次報告）について

## １　スキルセット及び学修目標の位置づけ

数理・データサイエンス教育強化拠点コンソーシアムでは、「大学の数理・データサイエンス教育強化方策について」（平成 28 年 12 月 数理及びデータサイエンス教育の強化に関する懇談会）を受け、コンソーシアム内に設置したカリキュラム分科会において、文理を問わず、全国的なモデルとなる標準カリキュラムの検討を進めてきた。

第 1 次報告では、リテラシーレベルのモデルカリキュラム検討の基礎とする観点から、リテラシーレベルのスキルセット及び学修目標を整理した。なお、これはＡＩ戦略 2019 に掲げられた具体目標「文理を問わず、全ての大学・高専生（約 50 万人卒/年）が、課程にて初級レベルの数理・データサイエンス・ＡＩを習得」への対応を見据えた取り組みの一環でもある。

本報告は、習得しておくことが望ましい標準的なスキルを網羅的に示したものである。したがって、各大学・高専（以下「大学等」という。）が一律に整備することを意図したものではなく、各大学等の教育目的、分野の特性、個々の学生の学習歴等に応じて、必要に応じて上位互換や既習確認等を行いながら、適切かつ柔軟に選択・抽出されることを想定している。

コンソーシアムでは、スキルセット及び学修目標やこれに続くモデルカリキュラムの検討と並行して、放送大学や MOOC（Massive Open Online Course）との連携や、e ラーニング教材、教科書及び教育用データベース等の活用を見据えた検討を進めており、必要に応じてこれらの教育資源を併用し教育内容を補完するなど、各大学等の主体的な取り組みにより、データサイエンス教育の推進・強化が図られることを期待している。

今後、本報告を活用しつつ、産業界や公・私立大学等関係者の参画を得てリテラシーレベルのモデルカリキュラムの検討を継続するとともに、いくつかの大学の協力を得て実践モデルの試行、グッドプラクティスの紹介等に取り組むことで、データサイエンス教育の全国への普及・展開を図る。また、その過程において継続してご意見を頂きながら必要な見直しを行い、応用基礎レベルを含めたモデルカリキュラムを作成する予定である。

また、今回報告するスキルセット及び学修目標は、現在在籍する学生を視野に作成している。高大接続の観点から「情報Ｉ」の必修化等の高等学校学習指導要領改訂を踏まえ、見直すことを前提としている*。

数理・データサイエンス・ＡＩ分野の発展は日進月歩であり、ビジネスや社会との関わりが急速に深化している状況に鑑み、様々な社会的課題の解決に向けて、産業界や国・地方公共団体等との連携を図りつつ、実際の事例を題材としたケーススタディ、グループワーク等を効果的に取り入れるなど、単に座学としてではなく、各大学等・教員が何らかの形で現実問題に触れるような機会を作るように工夫することにより、実務・実践で活用できる能力を育成することが期待される。加えて、概論等の講義を通じて、データサイエンスを専門としない学生に対しても「なぜ数理・データサイエンス・ＡＩを学ぶのか」、「社会でどのように活用されているのか」など、学修の動機付けにも留意が必要である。

---

* 高等学校学習指導要領解説とスキルセット及び学修目標との対応については、附属資料 1 を参照のこと。

各大学等・教員の創意工夫によって Society 5.0、ＡＩ時代に対応した人材育成を強化・推進することを期待している。

## ２　第１次報告の構成

第１次報告は、リテラシーレベルのスキルセット及び学修目標からなる。スキルセットは、データサイエンスの習得すべきスキル・理解すべき概念のセットである。学修目標は、全ての大学等へのデータサイエンス教育の普及・展開に向けて、習得すべき内容を文章でまとめたものである。それぞれ、「データサイエンスを学ぶ意義」、「データの法規・倫理」、「データの記述・可視化」、「データの取得・管理・加工」、「統計基礎」、「数学基礎」、「計算基礎」、「モデリングと評価」の分野（大分類）で構成している。

なお、分野（大分類）等に順序性はなく、各大学等の実情に応じてカリキュラム編成等が行われることを想定している。

## ３　参考資料等

カリキュラム分科会では、スキルセット及び学修目標の検討に当たり下記の資料を参考とした。特に、分野（大分類）の設定等にあたっては、米国の産業界、大学関係者の参画により取りまとめられた報告書「The National Academies of Sciences：Data Science for Undergraduates Opportunities and Options」（2018 年５月）を参考とした。

さらに、文部科学省、経済産業省、データサイエンス関係団体、放送大学及びデータサイエンス教育について先駆的な取組を有する公・私立大学との意見交換を行うなど、多方面からの意見収集に努めた。

- データサイエンティスト協会：データサイエンススキルセット
- 統計教育連携ネットワーク：コアカリキュラム
- R. D. De Veaux et al.: Curriculum Guidelines for Undergraduate Programs in Data Science, Annual Review of Statistics and Its Application, pp.15-30 (2017)
- American Statistical Association: ASA Statement on The Role of Statistics in Data Science (2015)
- 日本学術会議: 大学教育の分野別質保証のための教育課程編成上の参照基準：数理科学分野（2013）
- 日本学術会議: 大学教育の分野別質保証のための教育課程編成上の参照基準：統計学分野（2015）
- 日本学術会議: 大学教育の分野別質保証のための教育課程編成上の参照基準：情報学分野（2016）
- 日本学術会議: ビッグデータ時代に対応する人材の育成（2014）
- 情報・システム研究機構　ビッグデータの利活用に係る専門人材育成に向けた産学官懇談会: ビッグデータの利活用のための専門人材育成について（2015）

# Ⅱ　データサイエンス

## １　データサイエンス教育

データサイエンスは科学や産業に革命を起こす新しいアプローチとして注目されている。今や多くの研究分野のみならず社会の至るところで大規模で網羅的なデータが急速に蓄積されるようになり、データやその分析結果を利用することが可能になりつつある。その結果、今後はデータこそが経済、社会及び日常生活の発展と新しい価値創造の源泉になると考えられている。データサイエンスはそのようなデータ中心主義に立脚した科学的方法論である。アメリカ国立科学財団（National Science Foundation, NSF）の StatSNSF[†] によればデータサイエンスは、「データの設計、取得、管理、解析及びデータからの推論を目指す科学」とされている。

産学官懇談会[‡]の報告書で指摘されたように、データから新しい価値を生み出すことができる「棟梁レベル」をはじめとするデータサイエンスの専門人材を育成することの重要性は言うまでもない。同時に、専門人材が生み出す成果の受け手となる国民全体のデータリテラシーを醸成することが、我が国においてデータサイエンス、ＡＩを活かして超スマート社会を実現していくために必要不可欠だと考えられる。したがって棟梁レベルからリテラシーレベルまで裾野の広い人材育成が望まれる。

特にデータリテラシーとしては、データサイエンスやＡＩの不適切な利用や、個人データの利用から生じうる倫理的問題を認識できるようになることが重要である。何故ならば、その認識が今後のデータ駆動型社会を安全・安心に過ごすために不可欠だからである。もちろんデータサイエンスの専門人材も、そのような側面を十分に認識した上で価値を創造することが強く期待される。不適切な利用の代表例としては、誤判断につながる不適切な可視化、相関と因果の混同が挙げられる。

【不適切な可視化】

データのグラフ（棒グラフ、折れ線グラフ、円グラフなど）により、データの重要な特徴を、比較的容易に他者に伝えることができる。そのために、グラフは学術論文や技術文書だけではなく、テレビを始めとするマスメディアで頻繁に登場する。ただし、発信元が予め想定した主張をサポートするために、しばしばバイアスを持って提示される。その代表例として、縦軸をゼロから始めずに長さを切り詰めた棒グラフが挙げられる。

【相関と因果の混同】

単なる X と Y の相関関係が、もっともらしいデータ分析結果に基づいて因果関係（原因 X が結果 Y に影響を与える）のようにマスメディアで主張される事例は枚挙に暇がない。そのような主張に影響された個人が X を変えたとき、Y に期待された効果が得られないだけでなく、副作用により健康に害を及ぼすかもしれない。また、法人の意思決定や政

---

† https://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf

‡ 2015 年 7 月 30 日　情報・システム研究機構　ビッグデータの利活用に係る専門人材育成に向けた産学官懇談会 報告書「ビッグデータの利活用のための専門人材育成について」

策形成の場で誤認識された因果関係が用いられれば、大きな利益損失や税金の無駄遣いを招くであろう。

また、倫理的問題としては、「プライバシー保護のために匿名化されたデータベースにおいても、外部データベースとの照合などの操作により、個人が再識別される可能性」や「個人が再識別されない場合でも、データ分析の結果の解釈によって、年齢・民族・性などで層別された特定のグループについて、秘密が曝露されたり、差別が引き起こされる可能性」などが挙げられる。人権が侵害されたり個人の尊厳が損なわれることのないデータサイエンスやＡＩの利用を目指した国際的な取り組みは始まったばかりである（2019 年 6 月につくば市で開催された 20 カ国・地域（G20）貿易・デジタル経済相会合など）。我が国でも 2019 年 3 月に統合イノベーション戦略推進会議で「人間中心の AI 社会原則」[§]が決定された。特に教育に関する項で、全ての人間がデータに関する倫理の基礎を学ぶことの重要性が指摘されている。

今後の社会はますますデータ駆動型となるために、データから価値を創造できるデータサイエンスの専門家への需要は高まる一方である。さらに上記のように、データサイエンティストとして主体的にデータから価値を創出する業務に携わらない人にとっても、データサイエンスのリテラシーが重要である。結局、全ての大学等の卒業生に期待される能力やスキルにデータサイエンスの素養が新たに加わったのである。

## ２　データサイエンスのサイクルと重要な分野

データサイエンスは課題解決に向けたデータに基づく知識発展のサイクルと捉えることが重要である。日本学術会議提言（2014)[**] や R. D. De Veaux et al. (2017)[††]で指摘されているように、データサイエンスにおいては

- 課題抽出と定式化
- データの取得・管理・加工
- 探索的データ解析
- データ解析と推論
- 結果の共有・伝達、課題解決に向けた提案

---

§ 人間中心の AI 社会原則（2019 年 3 月統合イノベーション戦略推進会議決定、2019 年 6 月 G20 貿易・デジタル経済相会議閣僚声明）で述べられている人間に期待される能力及び役割等は以下のとおり。
○ 人間に期待される能力及び役割
・ AI の長所・短所の理解，AI の情報リソースとなるデータ，アルゴリズム，又はその双方にはバイアスが含まれること及びそれらを望ましくない目的のために利用する者がいることの認識
・ AI やデータのバイアスの認識（統計的バイアス，社会の様態によって生じるバイアス，AI 利用者の悪意によるバイアス）
○ 以下の原則に沿う教育環境が全ての人に提供される重要性
・ 誰でも AI，数理，DS の素養を身につけられる教育システム
・ 全ての人が文理の境界を超えて学ぶ必要
・ 以下のことを認識させるリテラシー教育
データにバイアスが含まれることや使い方によってバイアスを生じさせる可能性があること等の AI・データの特性があること
AI・データの持つ公平性・公正性，プライバシー保護に関わる課題があること

** 提言 ビッグデータ時代に対応する人材の育成 http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t198-2.pdf

†† R. D. De Veaux et al. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, p.15-30
https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-060116-053930

というプロセスがあり、課題解決と同時にデータからの知識発見や新たな課題発見が起こるために、知識発展のサイクルが生じる（図 1）。

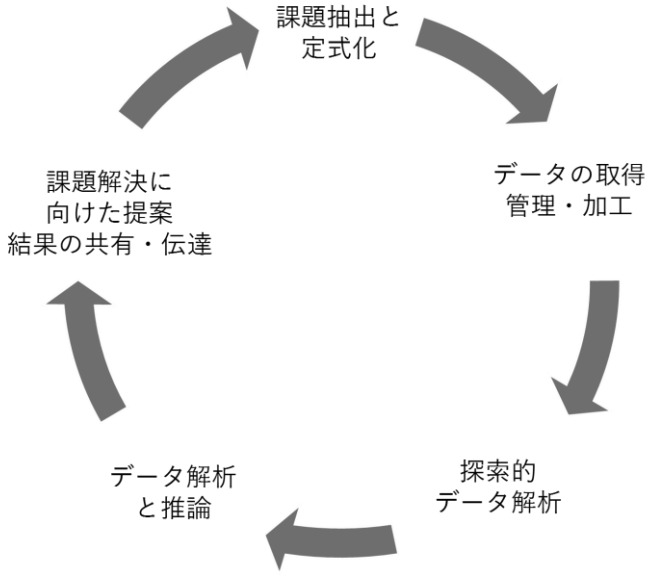



**図 1　データサイエンスのサイクル**

そのようなデータサイエンスのサイクルを理解して実行する能力は、Google の Dr. Diane Lambert の言葉を借りれば「データと共に思考する能力」[#]と呼ばれる。また、アメリカの The National Academies of Sciences, Engineering, and Medicine によるデータサイエンスの学部教育に関する報告書「Data Science for Undergraduates Opportunities and Options」（以下、「DSU 報告書」という。）では「データ洞察力」と定義された能力に対応する。

DSU 報告書では、データ洞察力を学部教育において養うために、以下の 10 分野が重要としている。

- Ethical problem solving 倫理に配慮した課題解決
- Data description and visualization データの記述・可視化
- Data management and curation データの取得・管理・加工
- Statistical foundations 統計基礎
- Mathematical foundations 数学基礎
- Computational foundations 計算基礎
- Data modeling and assessment モデリングと評価
- Domain-specific considerations ドメイン知識の考慮
- Communication and teamwork コミュニケーションとチームワーク
- Workflow and reproducibility ワークフローと再現性

[#] Curriculum Guidelines for Undergraduate Programs in Statistical Science (2014) https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf

これらの 10 分野はサイクルの中で以下のように位置づけられる（図 2）。ただし、「倫理に配慮した課題解決」は全てのプロセスにおいて重要であり、また最後の「ワークフローと再現性」は、サイクルを持続させていくための作業手順の保持と再現性のことであり、各ステップには割り振られない。

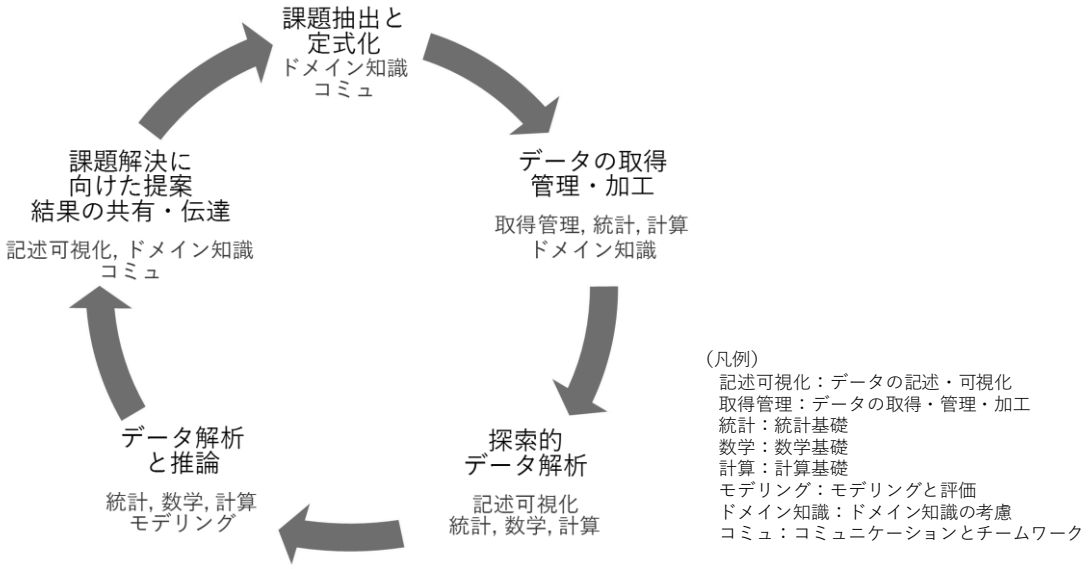



**図 2　データサイエンスのサイクル**

## ３　カリキュラム分科会の取り組み

データサイエンス教育で先行する欧米では、600 を超えるデータサイエンス学部や教育プログラムがスタートしている。日本でも 2017 年度スタートの滋賀大学を皮切りに、データサイエンス学部を設置する大学がいくつか現れている。これらのデータサイエンス教育を推進する学部においては、データサイエンスのサイクル全般にわたって活躍するデータサイエンティストを養成する観点から、DSU 報告書で言及された 10 分野全てをカバーした教育が展開されることが重要であろう。一方で、データ駆動型社会のリテラシーとして、全国の全ての大学生・高専生もデータサイエンスの基礎を学ぶことが期待されている。時間的な制約などから、リテラシーレベルの教育内容として全ての分野を考慮に入れるのは現実的ではない。しかしながら、データサイエンスのサイクルに主体的に関わるような職業に就かない場合でも、全ての国民がデータ駆動型社会の裾野となるとともに、データ分析の受け手としてデータサイエンスの成果を享受しつつ人生を生きていけるようになることが重要である。そのような観点から、数理・データサイエンス教育強化拠点コンソーシアム カリキュラム分科会では、第 1 次報告として身につけておくべきデータサイエンスのリテラシーについて提案する。

リテラシーレベルでは、特に以下の 7 分野に注目する。ただし「倫理に配慮した課題解決」については、法律・規制を含めて、「データの法規・倫理」とした。

- データの法規・倫理
- データの記述・可視化
- データの取得・管理・加工
- 統計基礎
- 数学基礎
- 計算基礎
- モデリングと評価

7 分野をそれぞれ大分類とみなし、その下に中分類・小分類と階層構造を設けている。例えば「データの法規・倫理」では、

大分類　データの法規・倫理

中分類　情報倫理、情報セキュリティ

小分類　情報倫理・関連法規／情報セキュリティ

という構造である。

そして、小分類を単位として、リテラシーとして身につけるべき事柄を**スキルセット**と**学修目標**として整理している。学修目標はスキルや概念のイメージが湧くように具体性を意識して文章で記述したものである。

なお、大分類等に対応させて科目設計を行うことは必ずしも意図していない。スキルセットと学修目標を参照し、各大学等の教育の特徴を活かしつつ、教育資源等も考慮しながら、各大学等において、科目編成やシラバス作成等を行うことを想定したものである。

【「データサイエンスを学ぶ意義」とコアの設定】

学生が興味や関心を持って学修に取り組むためには、データサイエンスを学ぶ意義、重要性の理解や学修の動機付け等を図ることが極めて重要である。このために、7 分野に加えて「データサイエンスを学ぶ意義」を設けた。各大学等の実情に応じて参照されたい。

また、各大学等において学生の高校での数学・情報科目の学習歴等を考慮した必修科目の検討や入門科目を設置する場合の参考となるように特に推奨する小分類をコアと位置づける。以下ではコア（☆）とそれ以外（★）で区別する。

なお、「データサイエンスを学ぶ意義」とコアは概ね 1.5 単位分程度であると想定しており、各大学等において「データサイエンス入門」のような科目を設置する場合には、「データサイエンスを学ぶ意義」及びコアに各大学等の事情に応じて内容を追加することで、2 単位や 4 単位の講義を設計できるなど、一定の自由度を持つものであると考えている。

この他、既存の関連科目にコアがどの程度含まれるかを確認するような用途も想定できる。ただし、高校での学習内容も含まれるため、各大学等の学生の実情（特に高校までの学習歴）によっては省略できる項目もあることに注意されたい。

**【中分類・小分類のリスト】**

リテラシーレベルのスキルを含む中分類・小分類のリストは以下のとおりである。特に☆でコアの小分類、★でコア以外の小分類を表す。

| 大分類 | 中分類 | 小分類 |
|---|---|---|
| データサイエンスを学ぶ意義 | (データサイエンスを学ぶ意義、重要性の理解)<br>データ駆動型社会　データに基づく課題解決、意志決定の支援　データサイエンスのサイクル　データサイエンスの様々な事例 | |
| データの法規・倫理 | 情報倫理、情報セキュリティ | ★情報倫理・関連法規　★情報セキュリティ |
| | データに関連する法律・規制 | ☆個人のデータに関連する法規　☆統計法 |
| | データサイエンスの倫理 | ☆倫理に配慮したデータ収集と利活用　☆データの匿名化　☆データサイエンスに関する様々なバイアス　★公表バイアス、出版バイアス |
| データの記述・可視化 | データの記述 | ☆種々のデータ　☆基本統計量　☆相関関係 |
| | データの可視化 | ☆グラフの構成要素　☆統計グラフ　☆分布の統計グラフ |
| データの取得・管理・加工 | データ取得とオープンデータ | ☆日本や世界のオープンデータ　☆オープンデータの取得 |
| | データ管理とデータ形式 | ☆表形式のデータ　★その他のデータ形式　★データベース |
| | データの前処理 | ★データクレンジング　★外れ値、異常値、欠損値<br>★データ加工 |
| 統計基礎 | 確率と確率分布 | ☆場合の数、順列・組み合わせ　☆確率　☆確率分布、確率変数　☆主要な確率分布 |
| | データ収集法と確率構造 | ☆標本調査　★ランダム化比較試験 |
| | 統計的推測 | ★統計的モデル　★標本分布　★点推定、区間推定<br>★仮説検定 |
| 数学基礎 | 線形代数 | ☆ベクトル　★行列 |
| | 微積分 | ☆多項式関数　★初等関数　★1 変数関数の微分法<br>★1 変数関数の積分法　★2 変数関数の微分法<br>★2 変数関数の積分法 |
| | 数列 | ☆データと数列　★数列 |
| | 演習 | ★線形代数演習　★微積分演習 |
| 計算基礎 | 情報、コンピュータの仕組み | ★数と表現　★有効数字、計算の誤差　★データ量の単位　★デジタル化　★文字の表現　★集合と命題　★論理演算 |
| | データ構造 | ★配列、リスト　★連想配列 |
| | アルゴリズムとプログラミング | ★アルゴリズム　★プログラミング |
| モデリングと評価 | 教師あり学習 | ★回帰分析　★ロジスティック回帰 |
| | 教師なし学習 | ★クラスタリング（k-平均法）　★階層クラスタリング |
| | モデルの評価 | ★評価指標　★訓練データとテストデータ |

## 4　スキルセットと学修目標

本節では、スキルセットと学修目標の位置づけや構成等について説明する。なお、ＡＩ戦略 2019 では、デジタル社会の基礎知識（いわゆる「読み・書き・そろばん」的な要素）である「数理・データサイエンス・ＡＩ」に関する知識・技能などを全ての国民が育み、社会のあらゆる分野で人材が活躍するために「文理を問わず、全ての大学・高専生（約 50 万人卒／年）が、課程にて初級レベルの数理・データサイエンス・ＡＩを習得」すること

が具体目標に設定された。当該分野・領域とビジネスや社会との関わりが急速に深化している状況に鑑み、実際の事例を題材としたケーススタディ、グループワーク等を効果的に取り入れるなど、単に座学としてではなく、各大学等・教員が何らかの形で現実問題に触れるような機会を作るように工夫することにより、実務・実践で活用できる能力を育成することが期待される。また、学生が学ぶ専門分野や学習歴等を踏まえながら、個々の学生が数理・データサイエンス・ＡＩについて学ぶ事柄について全体像を俯瞰し、「なぜ数理・データサイエンス・ＡＩを学ぶのか」、「社会でどのように活用されているのか」を学生が理解し、興味や関心を持って各スキルセットの内容を学修できるよう、学修の動機付けを含め教育内容・方法の創意工夫が期待される。

### （1）スキルセット

スキルセットは、データサイエンスの習得すべきスキルや理解すべき概念を、小分類を単位として関連するキーワードやレベルと合わせて整理したものである。レベルについては、初級（リテラシーレベル）から上級のレベル別に以下の 3 段階で整理する予定であるが、第1次報告ではリテラシーレベル（☆又は★）のスキルを提示している。

| | |
|---|---|
| ☆又は★ | 専門を問わず、全ての大学生・高専生が身につけるべきこと（リテラシーレベル） |
| ★★ | 数理・データサイエンス教育を推進する大学等の一般教育又は専門基礎教育レベル |
| ★★★ | 数理・データサイエンス教育を推進する大学等の専門教育レベル |

例えば、大分類「データの法規・倫理」中分類「データに関連する法律・規制」のスキルセットは以下のとおりであり、各小分類に複数個のスキルや概念が対応し、関連するキーワードとレベルと合わせて整理されている。

| 小分類 | 概要 | キーワード | レベル |
|---|---|---|---|
| 個人のデータに関連する法規 | 個人情報保護法の概要 | 個人情報保護法，改正個人情報保護法（第三者提供，個人情報の定義の拡充） | ☆ |
| | 個人のデータに関連する国際的な動き | EU 一般データ保護規則(GDPR)，忘れられる権利 | ☆ |
| 統計法 | 統計法の概要 | 統計法の基本事項と意義，基幹統計調査 | ☆ |

キーワードは概要を補足し分かりやすくするために挙げており、そのレベルは必ずしも星の数と対応しない。例えば、☆の項目にリテラシーレベルを超えるキーワードが含まれることがある。

なお、リテラシーレベルを超える内容（★★以上）を含めたスキルセットのイメージを3つの大分類（「数学基礎」、「計算基礎」、「統計基礎」）について試作し、参考として本報告に付した。

### （2）学修目標

学修目標では、全大学生・高専生へのデータサイエンス教育の普及・展開に向けて、リテラシーとして習得すべきスキルを文章で記述している。中分類について、当該中分類の

データサイエンス全体における意義や重要性を記述し、各小分類に対して 1 つ以上のスキルを提示している。例えば、大分類「データの法規・倫理」中分類「データに関連する法律・規制」に対する意義や重要性、当該中分類の小分類「個人のデータに関連する法規」に関するスキルは以下のとおりである。小分類は、コア（☆）とそれ以外（★）とで区別している。

大分類「データの法規・倫理」中分類「データに関連する法律・規制」に対する意義や重要性

（改正）個人情報保護法と統計法はデータに関する基本的な法規である。前者について、高度情報化社会において個人の情報・利益の保護と個人情報の有用性のバランスを取ることの重要性を学ぶ。後者について、公的統計が行政利用だけではなく、社会全体で利用される情報基盤として極めて重要であり、根拠に基づく（データに基づく）政策立案の基礎であることを学ぶ。

中分類「データに関連する法律・規制」小分類「個人のデータに関連する法規」に関するスキル

☆個人のデータに関連する法規

1. 我が国において、個人情報保護法が策定された背景を理解している。
2. 近年の飛躍的な情報通信技術の進展に伴い生じた、保護すべき個人情報の範囲の精緻化や、個人情報の適切な利活用の手続き、国境を越えた個人情報の流通への対応を目的とした個人情報保護法の改正の意義を理解している。
3. 改正個人情報保護法における、個人情報取扱事業者が個人情報を第三者に提供する場合の本人同意の必要性の意義、及び他の根拠法令を含めた例外規定の意義を理解している。
4. 改正個人情報保護法の「個人情報」「個人識別符号」「要配慮個人情報」「匿名加工情報」の概念を理解している。
5. EU 一般データ保護規則（GDPR）や「忘れられる権利」など国際的な動きを知っている。

学修目標では、習得すべきスキルのイメージが湧くように具体性を意識して記述した。そのために学修目標の１項目とスキルセットの１項目は必ずしも対応しない。

## ５　高大接続

データサイエンス教育を全国の全ての大学等に展開するにあたり、高校までの学習内容に大きなばらつきがあるのが問題となる。そのためにスキルセットと学修目標には高校での学習内容（数学、情報）をできるだけ含むようにして、高大接続を図っている。

### 数学

モデルカリキュラムの当面のターゲットは、現行の高校学習指導要領の下で学んだ学生である。そのために現行の数学科目の単元との対応をつけたリストを提供する。ただし、数学の単元は有機的に繋がっており、非対応の単元がデータサイエンスにとって重要でないことを意味しないことに注意されたい。

### 情報

現行の学習指導要領において、普通教育における教科「情報」は「社会と情報」と「情報の科学」の選択式である。データサイエンスに関連する内容は「情報の科学」に多く含

まれるが、高校では「社会と情報」の方が圧倒的に教えられている。また入試で「情報」科目の試験を行う大学等は少なく、これらの理由から現行の学習指導要領の教科「情報」と対応付けるのは適当ではない。

一方で、周知されているとおり次期の学習指導要領における「情報 I」には大きな注目が集まっている。全ての高校生の必修となり、大学入学共通テストでの出題が検討されている。さらにデータサイエンスのリテラシーにあたる内容を多く含んでいる。既に述べたようにモデルカリキュラムの当面のターゲットは、現行の高校学習指導要領の下で学んだ学生であるが、数年後の高校生が必修として学ぶ内容は現在の大学生・高専生にも不可欠な学習内容であろう。そのような観点から次期学習指導要領「情報 I」との対応をつける。

高校学習指導要領とスキルセット及び学修目標との対応の詳細は、附属資料 1 を参照されたい。

# Ⅲ　スキルセット

１　スキルセット　リテラシーレベル（コア☆のみ）

２　スキルセット　リテラシーレベル（☆及び★）

【参考】より高度なスキルを含むレベル別のスキルセット（イメージ）

## 1　スキルセット　リテラシーレベル（コア☆のみ）

データサイエンス教育を全国の全ての大学等に展開するにあたり、各大学等の学生の特性や文系・理系の違いを考慮する必要がある。このため概ね 1.5 単位分程度と考えられる内容を特にコアとして推奨した。

さらに、データサイエンスを学ぶ意義、重要性の理解や学修の動機付け等を図るために、7 分野に加えて「データサイエンスを学ぶ意義」を設けた。各大学等の実情に応じて参照されたい。

なお、各大学等において「データサイエンス入門」のような科目を設置する場合には、「データサイエンスを学ぶ意義」及びコアに各大学等の事情に応じて内容を追加することで、2 単位や 4 単位の講義を設計できるなど、一定の自由度を持つものであると考えている。

※中分類、小分類に付記した番号は、学修目標との対応を示す。

| 大分類 | 中分類 | 小分類 | 概要 | キーワード | レベル |
|---|---|---|---|---|---|
| データサイエンスを学ぶ意義 | （データサイエンスを学ぶ意義、重要性の理解）<br>1. データ駆動型社会<br>2. データに基づく課題解決、意志決定の支援<br>3. データサイエンスのサイクル<br>4. データサイエンスの様々な事例 | | | | ☆ |
| データの法規・倫理 | 2. データに関連する法律・規制 | 2.1 個人のデータに関連する法規 | 個人情報保護法の概要 | 個人情報保護法，改正個人情報保護法（第三者提供，個人情報の定義の拡充） | ☆ |
| | | | 個人のデータに関連する国際的な動き | EU 一般データ保護規則(GDPR)，忘れられる権利 | ☆ |
| | | 2.2 統計法 | 統計法の概要 | 統計法の基本事項と意義，基幹統計調査 | ☆ |
| | 3. データサイエンスの倫理 | 3.1 倫理に配慮したデータ収集と利活用 | データ収集対象者からの同意の必要性，オプトアウトの仕組み | インフォームドコンセント，オプトアウト | ☆ |
| | | 3.2 データの匿名化 | 匿名化と仮名化の違い，個人の再識別の可能性 | 匿名化，仮名化，再識別，秘密の曝露や差別の誘引 | ☆ |
| | | 3.3 データサイエンスに関する様々なバイアス | データサイエンスに関するバイアス | データバイアス，アノテーションバイアス，標本選択バイアス，帰納バイアス | ☆ |
| データの記述・可視化 | 1. データの記述 | 1.1 種々のデータ | データの分類 | 質的変数と量的変数．名義尺度，順序尺度，間隔尺度，比例尺度 | ☆ |
| | | | 時系列データ | 時系列データ，季節調整，移動平均 | ☆ |
| | | 1.2 基本統計量 | 分布の中心とばらつきに関する基本統計量 | 最頻値，中央値，平均値，四分位範囲，分散と標準偏差 | ☆ |
| | | | 分布における相対的な位置の概念 | 標準化，標準得点（Z 得点），偏差値，外れ値 | ☆ |
| | | 1.3 相関関係 | 2 つの量的変数の相関関係 | 散布図，共分散，相関係数 | ☆ |
| | | | 多変数データ（3 つ以上の量的変数）の 2 変数間の相関関係 | 散布図行列，相関係数行列 | ☆ |
| | | | 2 つ以上の質的変数の相関関係 | 分割表，クロス集計表，リスク比，リスク差，オッズ，オッズ比 | ☆ |
| | | | 相関関係と因果関係の違い | 相関と因果，交絡，偏相関係数 | ☆ |
| | | | 説明変数が一つの単回帰分析 | 最小二乗法，回帰分析，説明変数，目的変数，予測値，残差，決定係数 | ☆ |
| | 2. データの可視化 | 2.1 グラフの構成要素 | グラフの構成要素 | 色の効果や特徴，点の色・大きさ・形状への配慮，線の太さと様々な破線 | ☆ |
| | | 2.2 統計グラフ | データ全体の傾向や特徴を読み取るためのグラフ | 棒グラフ，折れ線グラフ，円グラフ，帯グラフ | ☆ |
| | | | グラフの作り方，見方のリテラシー | チャートジャンク，不必要な視覚的要素 | ☆ |
| | | 2.3 分布の統計グラフ | ヒストグラム，箱ひげ図，散布図などのデータの分布を示すグラフ | ヒストグラム，箱ひげ図，散布図（２変数），ヒートマップ（３変数以上） | ☆ |

| | | | | | |
|---|---|---|---|---|---|
| データの取得・管理・加工 | 1. データ取得とオープンデータ | 1.1 日本や世界のオープンデータ | 日本政府のオープンデータの取り組みとオープンデータの意義 | 二次利用可能なルール，機械判読への適性，国，地方公共団体，民間事業者の取り組み | ☆ |
| | | | オープンデータの普及に向けた国際的な動き | 2013 年の G8 サミットで合意された「オープンデータ憲章」 | ☆ |
| | | 1.2 オープンデータの取得 | 統計ポータルサイトからのデータのダウンロード | e-Stat，DATA.GO.JP データカタログサイト，Open Knowledge International | ☆ |
| | | | メタデータの検討 | メタデータ，データ取得の経緯，変数の特徴 | ☆ |
| | 2 データ管理とデータ形式 | 2.1 表形式のデータ | 表形式のデータ | csv，tsv | ☆ |
| 統計基礎 | 1. 確率と確率分布 | 1.1 場合の数，順列・組み合わせ | 場合の数，順列・組み合わせ | 場合の数，順列・組み合わせ，階乗 | ☆ |
| | | 1.2 確率 | 確率の概念．加法定理などの基本的な定理．ベイズの定理 | 加法定理，条件付き確率，ベイズの定理 | ☆ |
| | | 1.3 確率分布、確率変数 | 確率変数や確率分布の概念．離散型・連続型の確率分布 | 確率変数，確率関数，確率密度関数，累積分布関数 | ☆ |
| | | | 確率変数の平均（期待値）と分散 | 平均，期待値，分散 | ☆ |
| | | | 複数の確率変数に関する諸概念 | 同時分布，周辺分布，条件付分布，共分散と相関，独立 | ☆ |
| | | 1.4 主要な確率分布 | ベルヌーイ分布，2 項分布 | ベルヌーイ分布，2 項分布，期待値と分散 | ☆ |
| | | | 正規分布 | 正規分布．期待値と分散．線形変換による正規性の保存 | ☆ |
| | | | その他の主要な 1 変量確率分布（１） | ポアソン分布，指数分布，一様分布 | ☆ |
| | | | ２変量正規分布の視覚的理解 | ２変量正規分布，相関 | ☆ |
| | 2. データ収集法と確率構造 | 2.1 標本調査 | 標本を抽出して調査し、その結果から母集団の性質を推測する標本調査 | 単純無作為抽出法，クラスター抽出法，2 段抽出法 | ☆ |
| | | | 統計量を手がかりとした母集団の特性の推測 | 母数と統計量（母平均と標本平均）の区別，標本分布 | ☆ |
| 数学基礎 | 1. 線形代数 | 1.1 ベクトル | ベクトル | n 個の数値の組としての n 次元ベクトル（n=2,3,,,） | ☆ |
| | | | ベクトルの基本的な演算 | ベクトルの和，スカラー倍，線形結合 | ☆ |
| | | | 内積 | 内積，ベクトルの直交性，ノルム，単位ベクトル | ☆ |
| | 2. 微積分 | 2.1 多項式関数 | 二次関数，多項式と基本性質 | 二次関数や多項式のグラフ，四則演算 | ☆ |
| | | | 二項定理と二項係数 | 二項定理，パスカルの三角形 | ☆ |
| | 3. 数列 | 3.1 データと数列 | データの標本平均や分散の表現 | 標本平均，加重平均，標本分散，ベクトルの内積 | ☆ |

## 2　スキルセット　リテラシーレベル（☆及び★）

※中分類、小分類に付記した番号は、学修目標との対応を示す。

| 大分類 | 中分類 | 小分類 | 概要 | キーワード | レベル |
|---|---|---|---|---|---|
| データサイエンスを学ぶ意義 | （データサイエンスを学ぶ意義、重要性の理解）<br>1. データ駆動型社会<br>2. データに基づく課題解決、意志決定の支援<br>3. データサイエンスのサイクル<br>4. データサイエンスの様々な事例 | | | | ☆ |
| データの法規・倫理 | 1. 情報倫理、情報セキュリティ | 1.1 情報倫理・関連法規 | 情報倫理と関連法規の基礎 | 関連する法律（知的財産，個人情報の保護，不正アクセス行為の禁止等） | ★ |
| | | | 情報社会で生活する上でのマナー、モラル、倫理の意義 | マナー，モラル，倫理 | ★ |
| | | 1.2 情報セキュリティ | 情報セキュリティの 3 要素 | 情報セキュリティの 3 要素　機密性・完全性・可用性 | ★ |
| | | | 情報セキュリティを確保するための対策 | パスワード，個人認証，ソフトウェアのセキュリティ更新プログラム | ★ |
| | 2. データに関連する法律・規制 | 2.1 個人のデータに関連する法規 | 個人情報保護法の概要 | 個人情報保護法，改正個人情報保護法（第三者提供，個人情報の定義の拡充） | ☆ |
| | | | 個人のデータに関連する国際的な動き | EU 一般データ保護規則(GDPR)，忘れられる権利 | ☆ |
| | | 2.2 統計法 | 統計法の概要 | 統計法の基本事項と意義，基幹統計調査 | ☆ |
| | 3. データサイエンスの倫理 | 3.1 倫理に配慮したデータ収集と利活用 | データ収集対象者からの同意の必要性，オプトアウトの仕組み | インフォームドコンセント，オプトアウト | ☆ |
| | | 3.2 データの匿名化 | 匿名化と仮名化の違い，個人の再識別の可能性 | 匿名化，仮名化，再識別，秘密の曝露や差別の誘引 | ☆ |
| | | 3.3 データサイエンスに関する様々なバイアス | データサイエンスに関するバイアス | データバイアス，アノテーションバイアス，標本選択バイアス，帰納バイアス | ☆ |
| | | 3.4 公表バイアス | 公表バイアス，出版バイアス | 公表バイアスの危険性，公表バイアスを減らす取り組み | ★ |
| データの記述・可視化 | 1. データの記述 | 1.1 種々のデータ | データの分類 | 質的変数と量的変数．名義尺度，順序尺度，間隔尺度，比例尺度 | ☆ |
| | | | 時系列データ | 時系列データ，季節調整，移動平均 | ☆ |
| | | 1.2 基本統計量 | 分布の中心とばらつきに関する基本統計量 | 最頻値，中央値，平均値，四分位範囲，分散と標準偏差 | ☆ |
| | | | 分布における相対的な位置の概念 | 標準化，標準得点（Z 得点），偏差値，外れ値 | ☆ |
| | | 1.3 相関関係 | 2 つの量的変数の相関関係 | 散布図，共分散，相関係数 | ☆ |
| | | | 多変数データ（3 つ以上の量的変数）の 2 変数間の相関関係 | 散布図行列，相関係数行列 | ☆ |
| | | | 2 つ以上の質的変数の相関関係 | 分割表，クロス集計表，リスク比，リスク差，オッズ，オッズ比 | ☆ |
| | | | 相関関係と因果関係の違い | 相関と因果，交絡，偏相関係数 | ☆ |
| | | | 説明変数が一つの単回帰分析 | 最小二乗法，回帰分析，説明変数，目的変数，予測値，残差，決定係数 | ☆ |
| | 2. データの可視化 | 2.1 グラフの構成要素 | グラフの構成要素 | 色の効果や特徴，点の色・大きさ・形状への配慮，線の太さと様々な破線 | ☆ |
| | | 2.2 統計グラフ | データ全体の傾向や特徴を読み取るためのグラフ | 棒グラフ，折れ線グラフ，円グラフ，帯グラフ | ☆ |
| | | | グラフの作り方，見方のリテラシー | チャートジャンク，不必要な視覚的要素 | ☆ |
| | | 2.3 分布の統計グラフ | ヒストグラム，箱ひげ図，散布図などのデータの分布を示すグラフ | ヒストグラム，箱ひげ図，散布図（２変数），ヒートマップ（３変数以上） | ☆ |

| | | | | | |
|---|---|---|---|---|---|
| データの取得・管理・加工 | 1. データ取得とオープンデータ | 1.1 日本や世界のオープンデータ | 日本政府のオープンデータの取り組みとオープンデータの意義 | 二次利用可能なルール，機械判読への適性，国，地方公共団体，民間事業者の取り組み | ☆ |
| | | | オープンデータの普及に向けた国際的な動き | 2013 年の G8 サミットで合意された「オープンデータ憲章」 | ☆ |
| | | 1.2 オープンデータの取得 | 統計ポータルサイトからのデータのダウンロード | e-Stat，DATA.GO.JP データカタログサイト，Open Knowledge International | ☆ |
| | | | メタデータの検討 | メタデータ，データ取得の経緯，変数の特徴 | ☆ |
| | 2. データ管理とデータ形式 | 2.1 表形式のデータ | 表形式のデータ | csv，tsv | ☆ |
| | | 2.2 その他のデータ形式 | データ記述言語，SNS のつながりや鉄道路線などのデータ | XML，JASON，離散グラフ，キー・バリュー形式である隣接リスト | ★ |
| | | 2.3 データベース | データベースの意義と運用・管理 | データベース管理システム(DBMS) | ★ |
| | | | リレーショナルデータベース | 正規化，選択，射影，結合，SQL | ★ |
| | | | キー・バリュー型に対応したデータベース | NoSQL，キー・バリュー型に対応したデータベース | ★ |
| | 3. データの前処理 | 3.1 データクレンジング | データクレンジング | 文字列の表記の揺れの吸収（文字列，数字，日付，時刻） | ★ |
| | | 3.2 外れ値、異常値、欠損値 | 外れ値、異常値、欠損値 | 外れ値，異常値（箱ひげ図，標準化得点），データ欠損の要因，欠損値の補完 | ★ |
| | | 3.3 データ加工 | データ加工 | 部分集合の抽出，行の並び替え，新しい列の追加 | ★ |
| 統計基礎 | 1. 確率と確率分布 | 1.1 場合の数，順列・組み合わせ | 場合の数，順列・組み合わせ | 場合の数，順列・組み合わせ，階乗 | ☆ |
| | | 1.2 確率 | 確率の概念．加法定理などの基本的な定理．ベイズの定理 | 加法定理，条件付き確率，ベイズの定理 | ☆ |
| | | 1.3 確率分布、確率変数 | 確率変数や確率分布の概念．離散型・連続型の確率分布 | 確率変数，確率関数，確率密度関数，累積分布関数 | ☆ |
| | | | 確率変数の平均（期待値）と分散 | 平均，期待値，分散 | ☆ |
| | | | 複数の確率変数に関する諸概念 | 同時分布，周辺分布，条件付分布，共分散と相関，独立 | ☆ |
| | | 1.4 主要な確率分布 | ベルヌーイ分布，2 項分布 | ベルヌーイ分布，2 項分布，期待値と分散 | ☆ |
| | | | 正規分布 | 正規分布．期待値と分散．線形変換による正規性の保存 | ☆ |
| | | | その他の主要な 1 変量確率分布（1） | ポアソン分布，指数分布，一様分布 | ☆ |
| | | | ２変量正規分布の視覚的理解 | ２変量正規分布，相関 | ☆ |
| | 2. データ収集法と確率構造 | 2.1 標本調査 | 標本を抽出して調査し、その結果から母集団の性質を推測する標本調査 | 単純無作為抽出法，クラスター抽出法，2 段抽出法 | ☆ |
| | | | 統計量を手がかりとした母集団の特性の推測 | 母数と統計量（母平均と標本平均）の区別，標本分布 | ☆ |
| | | 2.2 ランダム化比較試験 | ランダム化比較試験，根拠に基づく医療．根拠に基づく政策立案．A/B テスト | 実験研究と観察研究，ランダム化比較試験．A/B テスト | ★ |
| | 3. 統計的推測 | 3.1 統計的モデル | 母集団に対して複数の確率分布の候補を表現する統計的モデル，分布をコントロールする母数（パラメータ） | 統計的モデル，母数，パラメータ | ★ |
| | | 3.2 標本分布 | 独立同一分布（i.i.d.）に従う確率変数の概念．その標本平均の期待値と分散，標本分散の期待値 | 独立同一分布，標本平均，標本分散 | ★ |
| | | 3.3 点推定，区間推定 | 点推定．特にモーメント法，最尤法による推定量の構成 | モーメント法，最尤法 | ★ |
| | | | 推定量の性質．特にバイアス，不偏性，平均二乗誤差 | バイアス，不偏推定量，平均二乗誤差，バイアス分散分解 | ★ |
| | | | 信頼区間の構成．信頼係数の理解 | 信頼区間と信頼係数，平均の区間推定，比率の区間推定 | ★ |
| | | 3.4 仮説検定の概念 | 二項分布の成功確率の検定などの基本的な仮説検定を通じた仮説検定のロジック | 帰無仮説，対立仮説，2 種の誤り | ★ |

| | | | | | |
|---|---|---|---|---|---|
| | | | 有意水準，検出力，p 値など検定に関する指標．帰無仮説のもとで検定統計量の従う分布 | 有意水準，検出力，p 値，検定統計量 | ★ |
| 数学基礎 | 1. 線形代数 | 1.1 ベクトル | ベクトル | n 個の数値の組としての n 次元ベクトル（n=2,3,,,） | ☆ |
| | | | ベクトルの基本的な演算 | ベクトルの和，スカラー倍，線形結合 | ☆ |
| | | | 内積 | 内積，ベクトルの直交性，ノルム，単位ベクトル | ☆ |
| | | 1.2 行列 | 行列の定義と基本的な演算 | 行列の和とスカラー倍 | ★ |
| | | | 行列の積 | 行列の積．積の非可換性．結合則，分配則 | ★ |
| | 2. 微積分 | 2.1 多項式関数 | 二次関数，多項式と基本性質 | 二次関数や多項式のグラフ，四則演算 | ☆ |
| | | | 二項定理と二項係数 | 二項定理，パスカルの三角形 | ☆ |
| | | 2.2 初等関数 | 指数関数，対数関数と基本性質 | 指数関数の形の把握，指数法則．対数関数の形の把握，対数法則 | ★ |
| | | | 三角関数と基本性質 | 三角関数の形の把握，三角関数の公式 | ★ |
| | | | 自然対数の底と指数関数 | 自然対数の底，ネイピア数 | ★ |
| | | | 関数の極限の定義と計算 | 関数の極限 | ★ |
| | | 2.3 １変数関数の微分法 | 導関数と関数の増減 | 平均変化率，微分係数と接線．導関数 | ★ |
| | | | 初等関数の導関数 | 初等関数の導関数 | ★ |
| | | | 微分法の基本的な関係 | 合成関数の微分法，積の微分法 | ★ |
| | | | 導関数と関数の極値と最大・最小 | 関数の最大・最小，極大・極小と導関数 | ★ |
| | | 2.4 １変数関数の積分法 | 微分の逆演算と原始関数 | 原始関数，不定積分 | ★ |
| | | | 定積分の計算 | 定積分と微分積分学の基本定理 | ★ |
| | | | 積分の基本的な関係式 | 部分積分，置換積分 | ★ |
| | | 2.5 ２変数関数の微分法 | 偏微分の定義 | 偏微分，偏導関数，接平面 | ★ |
| | | | 2 変数関数の最大最小問題 | 最大最小問題，極値と偏導関数の関係，ヘッセ行列 | ★ |
| | | 2.6 ２変数関数の積分法 | 長方形領域での重積分 | 重積分，重積分（長方形領域），累次積分 | ★ |
| | 3. 数列 | 3.1 データと数列 | データの標本平均や分散の表現 | 標本平均，加重平均，標本分散，ベクトルの内積 | ☆ |
| | | 3.2 数列 | 基本的な数列 | 等比数列，等差数列 | ★ |
| | | | 数列および数列の和の極限 | 数列の極限，数列の和の極限 | ★ |
| | 4. 演習 | 4.1 線形代数演習 | ベクトルや行列を対象とした計算機演習 | R や Python を用いた計算機演習 | ★ |
| | | 4.2 微積分演習 | 微積分に関する計算機演習 | Excel，R，Python などの既存の関数，ライブラリを用いた数値計算数式処理の演習（Python の Sympy パッケージなど） | ★ |
| 計算基礎 | 1. 情報，コンピュータの仕組み | 1.1 数と表現 | 2 進数の表現，基数変換の方法，負の数の表現，2 進数の加算や減算，表現可能な数値の範囲など | 2 進数，2 の補数表現 | ★ |
| | | 1.2 有効数字，計算の誤差 | コンピュータが計算処理を有限の桁数で行うことから生じる誤差．浮動小数点演算と誤差．丸め誤差 | 有効数字，仮数部，指数部 | ★ |
| | | 1.3 データ量の単位 | ビット，バイトなどのデータ量の単位や，SI 接頭辞を使った表現 | ビット，バイトなどの情報量，SI 接頭辞(k,M,G,T,P,m,μ,n,p など) | ★ |
| | | 1.4 デジタル化 | アナログとデジタルの特徴，デジタル化のプロセス，画像データ | 量子化，標本化，符号化，画像データ（ラスタ，ベクタ） | ★ |
| | | 1.5 文字の表現 | コンピュータの内部では，文字を数値で表現していることの理解．文字コード | ASCII，Unicode | ★ |
| | | 1.6 集合と命題 | 集合と命題，ベン図，真理値表などの基本的な考え方 | 真偽値，否定，論理和，論理積 | ★ |
| | | 1.7 論理演算 | 論理演算の考え方と基本的な演算，及び真理値表の利用方法 | 論理回路，論理ゲート，AND，OR，NOT，NAND，XOR | ★ |

| | | | | | |
|---|---|---|---|---|---|
| | 2. データ構造 | 2.1 配列，リスト | サイズが既知の複数の要素（値）の集合を格納・管理する配列，サイズが不明のデータを柔軟に格納するリスト | メモリ，ベクトル，行列，アドレス | ★ |
| | | 2.2 連想配列 | 項目（キー）と値（バリュー）を組とするデータを格納するデータ構造 | 連想，辞書，ハッシュ | ★ |
| | 3. アルゴリズムとプログラミング | 3.1 アルゴリズム | フローチャートの記号と処理手順の表現方法，疑似コード，アクティビティ図 | 端子，処理，判断，矢印 | ★ |
| | | | アルゴリズムの基本構造である順次構造，選択構造（分岐構造），繰返し構造（ループ構造） | 代入，計算，順次構造，選択構造，繰返し構造 | ★ |
| | | | 合計，探索，併合（マージ），整列（ソート）などの基本アルゴリズム | バブルソート，幅優先探索，深さ優先探索 | ★ |
| | | 3.2 プログラミング | Python や R などのインタプリタ型言語 | ソースコード，解釈実行，対話的実行環境 | ★ |
| | | | プログラムの書き方の規則である構文と，構文要素の解釈を定める意味論の区別 | 変数，代入，文，計算 | ★ |
| | | | データ型，扱うデータを格納するためその種類に応じた変数の宣言 | 文字型，整数型，浮動小数点型 | ★ |
| | | | 演算子と式．それぞれのデータ型に対して適用できる演算の種類とその表記法 | 四則演算，論理演算，式 | ★ |
| | | | 繰り返し処理や場合に応じた処理を記述する制御構造 | for，while，if | ★ |
| | | | 計算処理単位をまとめる関数．その定義と呼出し | 関数，引数，返値，再帰呼出し | ★ |
| | | | ファイルや標準入出力などに対するデータの読み書き | 標準入出力，ターミナル，ファイル，print，scanf | ★ |
| | | | プログラムの誤り（バグ）の有無を検証するデバッグ，テスト | デバッグ，テスト | ★ |
| | | | ツールやライブラリを用いて具体的な問題を実際に解く応用プログラミング | 表計算，テキスト処理，Excel，ライブラリ，API | ★ |
| モデリングと評価 | 1. 教師あり学習 | 1.1 回帰分析 | 回帰分析の手続き | 重回帰，多項式回帰，最小二乗法，係数の最小二乗推定値，決定係数，統計ソフトウェアによる実データを用いたデータ解析 | ★ |
| | | 1.2 ロジスティック回帰 | ロジスティック回帰の手続き | 最尤法によるパラメータ推定，推定された係数の解釈，統計ソフトウェアによる実データを用いたデータ解析 | ★ |
| | 2. 教師なし学習 | 2.1 クラスタリング(k-平均法) | k-平均法の手続き | k-平均法の手続き，結果の頑健性（初期値，距離），統計ソフトウェアによる実データを用いたデータ解析 | ★ |
| | | 2.2 階層クラスタリング | 階層クラスタリングの手続き | 階層クラスタリングの手続き，樹形図，結果の頑健性（類似度，併合法），統計ソフトウェアによる実データを用いたデータ解析 | ★ |
| | 3. モデルの評価 | 3.1 評価指標 | 回帰問題における平均二乗誤差 | 平均二乗誤差 | ★ |
| | | | 二値分類問題における混同行列に基づく評価 | 正解率(accuracy)，適合率(precision)，再現率(recall)，混同行列 | ★ |
| | | 3.2 訓練データとテストデータ | 訓練データとテストデータ | 訓練データとテストデータ，汎化誤差，過学習，適合不足 | ★ |
| | | | 交差検証法 | 交差検証法，leave-one-out | ★ |

## 【参考】より高度なレベルを含むレベル別のスキルセット（イメージ）

※第１次報告では、「統計基礎」、「数学基礎」、「計算基礎」の３つの大分類について試作。

| 大分類 | 中分類 | 小分類 | 概要 | キーワード | レベル |
|---|---|---|---|---|---|
| 統計基礎 | 確率と確率分布 | 場合の数，順列・組み合わせ | 場合の数，順列・組み合わせ | 場合の数，順列・組み合わせ，階乗 | ☆ |
| | | 確率 | 確率の概念．加法定理などの基本的な定理．ベイズの定理 | 加法定理，条件付き確率，ベイズの定理 | ☆ |
| | | 確率分布、確率変数 | 確率変数や確率分布の概念．離散型・連続型の確率分布 | 確率変数，確率関数，確率密度関数，累積分布関数 | ☆ |
| | | | 確率変数の平均（期待値）と分散 | 平均，期待値，分散 | ☆ |
| | | | 積率に関する不等式や高次の積率の概念 | チェビシェフの不等式，尖度，歪度，積率母関数 | ★★ |
| | | | 複数の確率変数に関する諸概念 | 同時分布，周辺分布，条件付分布，共分散と相関，独立 | ☆ |
| | | | 独立な確率変数の和の性質 | 独立な確率変数の和，期待値と分散、たたみ込みの密度関数 | ★★ |
| | | 主要な確率分布 | ベルヌーイ分布，２項分布 | ベルヌーイ分布，２項分布，期待値と分散 | ☆ |
| | | | 正規分布 | 正規分布．期待値と分散．線形変換による正規性の保存 | ☆ |
| | | | その他の主要な１変量確率分布（１） | ポアソン分布，指数分布，一様分布 | ☆ |
| | | | その他の主要な１変量確率分布（２） | 対数正規分布，ガンマ分布，ベータ分布 | ★★ |
| | | | ２変量正規分布の視覚的理解 | ２変量正規分布，相関 | ☆ |
| | | | 多次元確率分布の基礎である多変量正規分布の性質 | 多変量正規分布，周辺分布と条件付き分布の正規性、平均ベクトルと共分散行列. | ★★ |
| | | 確率変数の漸近的性質 | 確率論における重要な極限定理である大数の弱法則と中心極限定理 | 大数の法則，中心極限定理，確率収束，分布収束 | ★★★ |
| | データ収集法と確率構造 | 標本調査 | 標本を抽出して調査し、その結果から母集団の性質を推測する標本調査 | 単純無作為抽出法，クラスター抽出法，２段抽出法 | ☆ |
| | | | 統計量を手がかりとした母集団の特性の推測 | 母数と統計量（母平均と標本平均）の区別，標本分布 | ☆ |
| | | ランダム化比較試験 | ランダム化比較試験，根拠に基づく医療．根拠に基づく政策立案．A/B テスト | 実験研究と観察研究，ランダム化比較試験．A/B テスト | ★ |
| | 統計的推測 | 統計的モデル | 母集団に対して複数の確率分布の候補を表現する統計的モデル，分布をコントロールする母数（パラメータ） | 統計的モデル，母数，パラメータ | ★ |
| | | 標本分布 | 独立同一分布（i.i.d.）に従う確率変数の概念．その標本平均の期待値と分散，標本分散の期待値 | 独立同一分布，標本平均，標本分散 | ★ |
| | | | i.i.d.正規確率変数における標本平均，標本分散の従う分布．標本平均と標本分散の独立性 | カイ二乗分布，標本平均と標本分散の独立性 | ★★ |
| | | | i.i.d.正規確率変数の標本平均，標本分散から派生する分布であるスチューデントのｔ分布，スネデカーのＦ分布 | ｔ分布，Ｆ分布 | ★★ |
| | | 点推定，区間推定 | 点推定．特にモーメント法，最尤法による推定量の構成 | モーメント法，最尤法 | ★ |
| | | | 推定量の性質．特にバイアス，不偏性，平均二乗誤差 | バイアス，不偏推定量，平均二乗誤差，バイアス分散分解 | ★ |
| | | | 信頼区間の構成．信頼係数の理解 | 信頼区間と信頼係数，平均の区間推定，比率の区間推定 | ★ |
| | | 仮説検定の概念 | 二項分布の成功確率の検定などの基本的な仮説検定を通じた仮説検定のロジック | 帰無仮説，対立仮説，２種の誤り | ★ |

| | | | | | |
|---|---|---|---|---|---|
| | | | 有意水準，検出力，p 値など検定に関する指標．帰無仮説のもとで検定統計量の従う分布 | 有意水準，検出力，p 値，検定統計量 | ★ |
| | | 汎用的な検定 | パラメトリックモデルにおける尤度比検定 | 尤度比検定 | ★★ |
| | | | 様々なノンパラメトリック検定 | ウィルコクソン検定，順位和，並べ替え検定 | ★★ |
| | | 種々の検定 | 母集団の平均がある特定の値であるか（平均値の検定），２つの母集団の平均に違いがあるか（平均値の差の検定） | t 検定（群間の対応あり，なしを含む） | ★★ |
| | | | ３つ以上の母集団に平均の違いがあるか | 一元配置分散分析 | ★★ |
| | | | 母集団がある特定の分布であるか（適合度検定），分割表の２つの属性が独立であるか（独立性の検定） | 適合度検定，独立性の検定 | ★★ |
| | | 多重比較 | 検定を複数回繰り返すことによる第一種の過誤の確率の増加．その補正方法であるボンフェロニ補正 | ボンフェロニ補正，ホルム補正 | ★★ |
| | | ベイズ統計的推測 | 共役事前分布に対する事後分布．明示的な事後平均の表記による事前分布の影響の理解． | 事前分布，共役事前分布，事後分布 | ★★★ |
| | 統計的推測（回帰分析，ロジスティック回帰） | 回帰分析（理論） | 回帰分析における最小二乗推定量の従う分布，残差平方和の従う分布，t 統計量の従う分布 | 正規分布，カイ二乗分布，t 分布，最小二乗推定量 | ★★ |
| | | | 回帰係数や回帰係数ベクトルの有意性検定 | t 分布，F 分布，標準誤差，回帰係数の有意性検定 | ★★ |
| | | 回帰診断 | 残差プロットをもとに回帰分析の諸仮定の検討．多重共線性など推測の信頼性に関わる問題 | 多重共線性，外れ値 | ★★ |
| | | 変数選択・モデル選択 | 重回帰分析を題材とした変数選択，モデル選択 | 自由度調整決定係数．交差検証法．AIC | ★★ |
| | | ロジスティック回帰 | 目的変数が 2 値の場合に使われるロジスティック回帰．最尤法に基づく統計的推測 | 対数オッズ，最尤推定，尤度比検定，ワルド検定 | ★★ |
| | 計算統計 | ブートストラップ | 統計理論を数値計算，シミュレーションで置き換えるブートストラップ．より一般にリサンプリングに基づく手法 | 復元抽出，経験分布，リサンプリング，ブートストラップ | ★★★ |
| | | 疑似乱数 | 決定論的なコンピュータに確率変数を発生させる疑似乱数．一様乱数 | 疑似乱数，合同法，メルセンヌ・ツイスタ法 | ★★★ |
| | | サンプリング | 目標分布からの（独立な）疑似乱数生成．逆変換法，棄却法 | 逆変換法，棄却法 | ★★★ |
| | | | 目標分布からの（サンプル間の相関を許す）疑似乱数生成．メトロポリス・ヘイスティングス法，ギブスサンプリング法 | マルコフ連鎖，詳細釣り合い条件，事後分布 | ★★★ |
| | | モンテカルロ積分 | サンプリングされた疑似乱数を用いた期待値や確率密度の正規化定数の計算 | モンテカルロ積分 | ★★★ |
| 数学基礎 | 線形代数 | ベクトル | ベクトル | n 個の数値の組としての n 次元ベクトル（n=2,3,,,） | ☆ |
| | | | ベクトルの基本的な演算 | ベクトルの和，スカラー倍，線形結合 | ☆ |
| | | | 内積 | 内積，ベクトルの直交性，ノルム，単位ベクトル | ☆ |
| | | 行列 | 行列の定義と基本的な演算 | 行列の和とスカラー倍 | ★ |
| | | | 行列の積 | 行列の積．積の非可換性．結合則，分配則 | ★ |
| | | | 特別な行列とその性質 | 正方行列，単位行列，転置行列，対称行列，三角行列，直交行列 | ★★ |
| | | | 正方行列の逆行列と求め方 | 正方行列の逆行列，行列とその逆行列の積の可換性 | ★★ |
| | | | 行列の基本変形とランク | 基本変形，ランク，簡約な行列 | ★★ |

| | | | | | |
|---|---|---|---|---|---|
| | | | 行列を特徴付ける量とその性質 | 行列式，トレース，正則性 | ★★ |
| | | データ記述と線形代数 | 分散，共分散とベクトルの内積 | all-ones ベクトル，偏差ベクトル，２つの偏差ベクトルの内積 | ★★ |
| | | | 回帰分析と射影操作 | 射影行列，回帰分析における予測値ベクトルと残差ベクトル | ★★ |
| | | 固有値と固有ベクトル | 対称行列の固有値と固有ベクトル | 対称行列の固有値，固有ベクトル | ★★ |
| | | | 対称行列の直交行列による対角化 | 対称行列の対角化，スペクトル分解，直交行列 | ★★ |
| | | | 二次形式と正定値性 | 二次形式と（半）正定値行列 | ★★ |
| | | | 特異値分解の定義と構成 | 特異値分解 | ★★ |
| | | n 次元ユークリッド空間 | n 次元ベクトルと n 次元空間上の点との対応 | n 次元ベクトル，n 次元空間上の点の表現 | ★★ |
| | | | ベクトルの線形独立性，線形従属性 | 線形独立，線形従属 | ★★ |
| | | | 複数のベクトルの張る線形部分空間 | 線形空間，線形部分空間 | ★★ |
| | | | 線形空間，線形部分空間の基底と次元 | 線形部分空間と基底・次元．行列のランクとその列空間の次元 | ★★ |
| | | | 連立一次方程式の解の判定と解法 | 同次方程式．係数行列，拡大係数行列，解空間．解の一意性 | ★★ |
| | | | 正規直交基底とその構成法 | 正規直交基底，シュミットの直交化．QR 分解 | ★★ |
| | | | 射影操作 | 射影と直交成分．直交補空間．射影行列，べき等性 | ★★ |
| | | | 線形不等式 | 線形不等式論（Farkas の補題，線形計画の双対性） | ★★★ |
| | | 数値計算と線形代数 | 連立方程式の数値計算における行列の性質 | LU 分解，QR 分解 | ★★ |
| | | | 大規模連立方程式の解法，固有値解析の解法 | 反復法など | ★★★ |
| | 微積分 | 多項式関数 | 二次関数，多項式と基本性質 | 二次関数や多項式のグラフ，四則演算 | ☆ |
| | | | 二項定理と二項係数 | 二項定理，パスカルの三角形 | ☆ |
| | | 初等関数 | 指数関数，対数関数と基本性質 | 指数関数の形の把握，指数法則．対数関数の形の把握，対数法則 | ★ |
| | | | 三角関数と基本性質 | 三角関数の形の把握，三角関数の公式 | ★ |
| | | | 自然対数の底と指数関数 | 自然対数の底，ネイピア数 | ★ |
| | | | 三角関数の逆関数 | arctan などの逆三角関数 | ★★ |
| | | | 関数の極限の定義と計算 | 関数の極限 | ★ |
| | | １変数関数の微分法 | 導関数と関数の増減 | 平均変化率，微分係数と接線．導関数 | ★ |
| | | | 初等関数の導関数 | 初等関数の導関数 | ★ |
| | | | 微分法の基本的な関係 | 合成関数の微分法，積の微分法 | ★ |
| | | | 導関数と関数の極値と最大・最小 | 関数の最大・最小，極大・極小と導関数 | ★ |
| | | | 初等関数のテイラー展開 | テイラー展開 | ★★ |
| | | | 方程式の数値的な解法 | 方程式の数値的解法，二分法，ニュートン法 | ★★ |
| | | １変数関数の積分法 | 微分の逆演算と原始関数 | 原始関数，不定積分 | ★ |
| | | | 定積分の計算 | 定積分と微分積分学の基本定理 | ★ |
| | | | 積分の基本的な関係式 | 部分積分，置換積分 | ★ |
| | | | 広義積分 | 広義積分 | ★★ |
| | | | 広義積分によって定義される関数 | ガンマ関数，ベータ関数 | ★★ |
| | | ２変数関数の微分法 | 偏微分の定義 | 偏微分，偏導関数，接平面 | ★ |
| | | | 2 変数関数の最大最小問題 | 最大最小問題，極値と偏導関数の関係，ヘッセ行列 | ★ |
| | | | 2 変数関数のテイラー展開 | テイラー展開 | ★★ |
| | | | 最急降下法と最適化手法 | 最急降下法，（最適化手法としての）ニュートン法，準ニュートン法 | ★★ |
| | | | 多変数の非線形方程式の数値解法 | （非線形方程式の解法としての）ニュートン法 | ★★ |

| | | | | | |
|---|---|---|---|---|---|
| | | | 条件付き極値問題の解法 | ラグランジュ乗数法，条件付き極値問題 | ★★★ |
| | | | 不等式による条件付き極値問題 | Karush-Kuhn-Tucker（KKT）条件 | ★★★ |
| | | | 凸関数条件 | 凸関数（定義，ヘッセ行列の（半）正定値性との関係，最適性条件） | ★★ |
| | | ２変数関数の積分法 | 長方形領域での重積分 | 重積分，重積分（長方形領域），累次積分 | ★ |
| | | | 一般の領域での重積分 | 一般の領域での重積分（縦線領域，横線領域の重積分） | ★★ |
| | | | 重積分における置換積分 | 変数変換とヤコビアン | ★★ |
| | | | 広義重積分 | 広義重積分，ガウス積分，正規分布 | ★★ |
| | | | 一般の多変数変数における重積分 | 多重積分，極座標変換 | ★★★ |
| | | 数値積分 | 基本的な数値積分の手法 | 台形則，シンプソン法 | ★★ |
| | | | モンテカルロ法による定積分計算 | モンテカルロ法 | ★★★ |
| | 数列 | データと数列 | データの標本平均や分散の表現 | 標本平均，加重平均，標本分散，ベクトルの内積 | ☆ |
| | | 数列 | 基本的な数列 | 等比数列，等差数列 | ★ |
| | | | 数列の収束の速さの違い | 収束の速さ | ★★ |
| | | | 数列および数列の和の極限 | 数列の極限，数列の和の極限 | ★ |
| | 演習 | 線形代数演習 | ベクトルや行列を対象とした手計算演習 | ベクトルや行列を対象とした手計算演習（内積，逆行列，固有値） | ★★ |
| | | | ベクトルや行列を対象とした計算機演習 | R や Python を用いた計算機演習 | ★ |
| | | 微積分演習 | 微積分に関する手計算演習 | 解析的な計算が可能な問題の手計算による演習 | ★★ |
| | | | 微積分に関する計算機演習 | Excel，R，Python などの既存の関数，ライブラリを用いた数値計算数式処理の演習（Python の Sympy パッケージなど） | ★ |
| 計算基礎 | 情報，コンピュータの仕組み | 数と表現 | 2 進数の表現，基数変換の方法，負の数の表現，2 進数の加算や減算，表現可能な数値の範囲など | 2 進数，2 の補数表現 | ★ |
| | | 有効数字，計算の誤差 | コンピュータが計算処理を有限の桁数で行うことから生じる誤差．浮動小数点演算と誤差．丸め誤差 | 有効数字，仮数部，指数部 | ★ |
| | | データ量の単位 | ビット，バイトなどのデータ量の単位や，SI 接頭辞を使った表現 | ビット，バイトなどの情報量，SI 接頭辞(k,M,G,T,P,m,$\mu$,n,p など) | ★ |
| | | デジタル化 | アナログとデジタルの特徴，デジタル化のプロセス，画像データ | 量子化，標本化，符号化，画像データ（ラスタ，ベクタ） | ★ |
| | | 文字の表現 | コンピュータの内部では，文字を数値で表現していることの理解．文字コード | ASCII，Unicode | ★ |
| | | 集合と命題 | 集合と命題，ベン図，真理値表などの基本的な考え方 | 真偽値，否定，論理和，論理積 | ★ |
| | | 論理演算 | 論理演算の考え方と基本的な演算，及び真理値表の利用方法 | 論理回路，論理ゲート，AND，OR，NOT，NAND，XOR | ★ |
| | データ構造 | 配列，リスト | サイズが既知の複数の要素（値）の集合を格納・管理する配列，サイズが不明のデータを柔軟に格納するリスト | メモリ，ベクトル，行列，アドレス | ★ |
| | | 連想配列 | 項目（キー）と値（バリュー）を組とするデータを格納するデータ構造 | 連想，辞書，ハッシュ | ★ |
| | | スタック，キュー | 後入れ先出しに従ってデータを格納するスタック，先入れ先出しに従ってデータを格納するキュー | push，pop，enqueue，dequeue | ★★ |
| | | グラフ，木構造 | ノード（頂点）とノード間の連結を表すエッジにて表現されるグラフ．階層性を有するグラフである木構造 | ネットワーク，可視化，分類，二分木，ヒープ | ★★★ |
| | アルゴリズムとプログラミング | アルゴリズム | フローチャートの記号と処理手順の表現方法，疑似コード，アクティビティ図 | 端子，処理，判断，矢印 | ★ |

| | | | | | |
|---|---|---|---|---|---|
| | | | アルゴリズムの基本構造である順次構造，選択構造（分岐構造），繰返し構造（ループ構造） | 代入，計算，順次構造，選択構造，繰返し構造 | ★ |
| | | | 合計，探索，併合（マージ），整列（ソート）などの基本アルゴリズム | バブルソート，幅優先探索，深さ優先探索 | ★ |
| | | | 貪欲法，分割統治法，動的計画法，再帰的アルゴリズムなどの応用アルゴリズム | 貪欲法，分割統治法，動的計画法，再帰的アルゴリズム | ★★ |
| | | | 計算量，計算量のオーダー（線形オーダー，べき乗オーダー）とその表現，計算回数 | ビッグ O 記法，入力データ量，計算時間，ステップ数，最大次数 | ★★ |
| | | プログラミング | Python や R などのインタプリタ型言語 | ソースコード，解釈実行，対話的実行環境 | ★ |
| | | | C/C++などのコンパイラ型言語 | ネイティブコード（機械語），翻訳，実行 | ★★ |
| | | | プログラムの書き方の規則である構文と，構文要素の解釈を定める意味論の区別 | 変数，代入，文，計算 | ★ |
| | | | データ型，扱うデータを格納するためその種類に応じた変数の宣言 | 文字型，整数型，浮動小数点型 | ★ |
| | | | 演算子と式．それぞれのデータ型に対して適用できる演算の種類とその表記法 | 四則演算，論理演算，式 | ★ |
| | | | 繰り返し処理や場合に応じた処理を記述する制御構造 | for，while，if | ★ |
| | | | 計算処理単位をまとめる関数．その定義と呼出し | 関数，引数，返値，再帰呼出し | ★ |
| | | | ファイルや標準入出力などに対するデータの読み書き | 標準入出力，ターミナル，ファイル，print，scanf | ★ |
| | | | 再利用可能な定型処理を発見し，関数として括り出して記述するリファクタリング | 計算の構造化，モジュール化，リファクタリング | ★★ |
| | | | データとそれに対する操作が結び付けられたオブジェクトを用いるオブジェクト指向 | オブジェクト指向，オブジェクト，クラス，カプセル化，継承，多相性 | ★★★ |
| | | | プログラムの誤り（バグ）の有無を検証するデバッグ，テスト | デバッグ，テスト | ★ |
| | | | ツールやライブラリを用いて具体的な問題を実際に解く応用プログラミング | 表計算，テキスト処理，Excel，ライブラリ，API | ★ |

# Ⅳ　学修目標

○　データサイエンスを学ぶ意義

○　データの法規・倫理
○　データの記述・可視化
○　データの取得・管理・加工
○　統計基礎
○　数学基礎
○　計算基礎
○　モデリングと評価

# データサイエンスを学ぶ意義

インターネットの社会への広範囲な浸透、情報通信・計測技術の飛躍的発展、ＡＩ技術の活用領域の拡大により、大規模で網羅的なデータが急速に蓄積され、データそのものやその分析結果を利用することが可能になっている。その結果、日常生活の中でも、販売記録、購買履歴、満足度やランキングなどのデータを元に意思決定が行われるようになっている。また、そうした個人の意志決定とその結果そのものが企業にとっては重要な情報であり、次回の意思決定に利用される情報が更新される。このような情報循環は、社会を発展させ個人の日常生活を快適にする。学術の世界においても、専門分野に関わらずデータサイエンス、ＡＩ技術の活用が加速しており、分野横断研究の活性化や新領域の創出が進んでいる。スポーツの世界においても、多くの種目でデータの収集・分析を元にした戦略構築が進んでいる。

データ駆動型への転換が急速に進む今日において、数理・データサイエンス・ＡＩに関するリテラシーは、ＡＩ時代を担う全ての学生にとって必要不可欠な基礎的素養となった。データサイエンスの必要性を事例を通じて理解し、継続的な学習に繋げていくことが極めて重要である。一方でそのようなデータが個人情報と紐付いている場合には、倫理的な問題が生じることも少なくない。こうした問題にどう対応すべきかについてはデータサイエンス教育の最初の段階から意識しておく必要がある。

このように、データサイエンスへの導入として、産業界・学術界等における実際の事例を参照しながら、データサイエンスは社会に多大な便益をもたらすこと、その一方で倫理的問題などの注意すべき側面があること、またそれらを深く理解するために（これから学ぶことになる）基礎が必要であることを学生に認識させる。

## 1.　データ駆動型社会

1. 大規模で詳細なデータが急速に蓄積されるようになっている実例を挙げられる。
2. データやその分析結果の活用によって、経済、社会及び日常生活の様々な側面が大きく変化していることを理解している。
3. 学生の興味に応じて、また専攻分野によって、学術におけるデータ駆動型の成果の例を挙げられる。

## 2.　データに基づく課題解決、意志決定の支援

1. 自分の生活の中で、データに基づいて意思決定した例を挙げられる。
2. 政策立案やビジネスの現場における、データを用いた課題解決や意志決定の意義を理解している。

## 3.　データサイエンスのサイクル

1. データサイエンスは、目的の達成のために、課題設定、データ取得から始まる課題解決に至るまでのプロセスであることを理解している。特に課題解決と同時にデー

タからの知識発見や課題発見が起こるために、プロセスはサイクルをなすことを理解している。

2. データサイエンスのプロセスの各ステップの重要性を理解している。
3. サイクルの全体を通じて、倫理的問題を検討することの重要性を理解している。
4. サイクルの全体を通じて、再現性を担保することの重要性を理解している。
5. 他者と協同して課題解決を行う場面で、データやデータ分析に基づいた（根拠に基づいた）コミュニケーションの重要性や有効性を認識できる。

## 4. データサイエンスの様々な事例

1. データサイエンスを用いた取り組みの成功例を挙げられる。
2. 個人のデータの匿名化を行っても、再び識別され深刻な倫理的な問題が引き起こされる可能性があることを知っている。
3. バイアスのあるサンプルが誤った予測や分類に繋がった例を挙げられる。
4. 不適切なグラフ、可視化が誤解をもたらす可能性があることを理解している。
5. 相関関係と因果関係の違いを理解している。

**※ 概論や授業の導入等に活用できる様々な事例については、附属資料 2 を参照されたい。**

# データの法規・倫理

## 1.　情報倫理、情報セキュリティ

データサイエンスを学ぶ際に、数学や統計などの理論を身につけるだけではなく、コンピュータで実データを分析する能力やその際の注意点を習得することが重要である。また、ネットワークを介したグループでの共同作業により、データサイエンスの課題解決を導くことも想定される。ネットワークに繋がったコンピュータを扱うにあたり情報倫理の基礎を学び、情報セキュリティへの理解を深める。

### 1.1　★情報倫理・関連法規

1. 知的財産に関する法律、個人情報の保護に関する法律、不正アクセス行為の禁止等に関する法律などを含めた法規が情報社会で果している役割を理解している。
2. 情報社会で生活する上でのマナー、モラル、倫理の意義や、法を遵守することの重要性を理解している。
3. 現代情報社会では技術革新のスピードが非常に早く、法規や制度の対応が間に合わないケースがあること、そのために個人のモラルや倫理に基づく正しい対応が期待されることを理解している。

### 1.2　★情報セキュリティ

1. 情報セキュリティの 3 要素である機密性・完全性・可用性の重要性を理解している。
2. 情報セキュリティを確保するための対策として、推測されにくいパスワードや生体認証などの個人認証の必要性、ソフトウェアのセキュリティ更新プログラムを適用する必要性、その提供が終了したソフトウェアを使い続けることの危険性を理解している。

## 2.　データに関連する法律・規制

（改正）個人情報保護法と統計法はデータに関する基本的な法規である。前者については、高度情報化社会において個人の情報・利益の保護と個人情報の有用性のバランスを取ることの重要性を学ぶ。後者については、公的統計が行政利用だけではなく、社会全体で利用される情報基盤として極めて重要であり、根拠に基づく（データに基づく）政策立案の基礎であることを学ぶ。

### 2.1　☆個人のデータに関連する法規

1. 我が国において、個人情報保護法が策定された背景を理解している。
2. 近年の飛躍的な情報通信技術の進展に伴い生じた、保護すべき個人情報の範囲の精緻化や、個人情報の適切な利活用の手続き、国境を越えた個人情報の流通への対応を目的とした個人情報保護法の改正の意義を理解している。

3. 改正個人情報保護法における、個人情報取扱事業者が個人情報を第三者に提供する場合の本人同意の必要性の意義、及び他の根拠法令を含めた例外規定の意義を理解している。
4. 改正個人情報保護法の「個人情報」「個人識別符号」「要配慮個人情報」「匿名加工情報」の概念を理解している。
5. EU 一般データ保護規則（GDPR）や「忘れられる権利」などの国際的な動きを知っている。

### 2.2 ☆統計法

1. 統計法が、公的統計の作成及び提供に関する基本事項を定めていることを知っている。
2. 統計法の目的が、公的統計の体系的かつ効率的な整備及びその有用性の確保を図り、国民経済の健全な発展及び国民生活の向上に寄与することであることを知っている。
3. 公的統計が行政利用だけではなく、社会全体で利用される情報基盤として位置付けられることの意義を理解している。
4. 国勢調査などの基幹統計調査は、公的統計の中核となる基幹統計を作成するための特に重要な統計調査であり、正確な統計を作成する必要性が特に高いことを理解している。

## 3. データサイエンスの倫理

データサイエンスを応用し多種多様で大量のデータを分析した成果は、社会に多大なる便益をもたらす。その一方で、深刻な倫理的な問題を引き起こす可能性があることが指摘されている。例えば、個人情報の曝露や、プロファイリング等の自動処理による年齢・民族・性などで層別された特定のグループへの差別や人権侵害である。ここでは特に、倫理に配慮したデータ収集や利活用、匿名化されたデータであっても個人情報を曝露するリスクがあること、またデータサイエンスに関連する様々な偏り（バイアス）について具体例を通じて理解を深める。

### 3.1 ☆倫理に配慮したデータ収集と利活用

1. データ収集の目的、データの活用方法がデータ収集対象者に説明されたうえで同意を得ていることが倫理的に望ましいことを理解している。
2. データ収集対象者が分析対象から外れることを希望したときにオプトアウトを実現できる仕組みの必要性を理解している。
3. データの利活用において、調査客体が他者に知られたくない変数（調査項目）が国や文化によって異なること、より一般に倫理的基準が国や文化によって異なること、を理解している。

### 3.2 ☆データの匿名化

1. 個人情報を含んだデータの匿名化と仮名化の違いを理解している。

2. 氏名などの個人情報の代わりに研究用 ID などを用いて仮名化されたとしても、外部データベースとの照合などの操作により、個人が再識別されうることを具体例をもとに理解している。
3. 個人が再識別されない場合でも、データ分析の結果の解釈によって、年齢・民族・性などで層別された特定のグループについて、秘密が曝露されたり、差別が引き起こされる場合があることを理解している。

### 3.3 ☆データサイエンスに関する様々なバイアス

1. データバイアス、アノテーションバイアス
    a. データサイエンスにおいて、人手でタグをつけたデータを正解として教師あり学習が適用される場合があることを知っている。その際にタグ付けをする人の認知バイアスなどにより、特定のグループに対する差別が生じうることを理解している。
    b. 推薦システムにおいて、あるアイテムに関する評点をつける場合、画面に表示されているその時点での平均点などに影響される傾向が見られること、またその繰り返しにより将来の平均点などにバイアスが生じることを理解している。
2. 標本選択バイアス
    a. 得られた標本が母集団から偏りを持って抽出されている場合には、たとえ適切なデータモデリング手法で分析されても、信頼できない結果しか得られないことを理解している。
    b. 既に取得されたデータは過去の状態を反映していることを理解している。特に将来の母集団の特性値を予測したい場合に、過去のデータが未来の母集団からの偏りのないサンプルと見なせるかどうかの検討が重要であることを理解している。
    c. 大学入試の成績と入学後の学業成績の関係を調べる場合に、不合格で入学していない個人のデータは原理的に欠損していることなどを例に、標本のバイアスや分析の限界を理解できる。
    d. 幹部社員の男女比が極端に偏っているような企業における採用活動で、現社員に基づくデータ分析をもとに就職希望者の適性を判断するような例を題材として、過去のデータ分析結果を学習した場合は年齢・民族・性などで層別された特定のグループに対する差別が生じうることを理解している。
3. 帰納バイアス
    a. データサイエンスにおけるモデルは様々な仮定のもとに設計されること、それゆえに現実の近似に過ぎないことを理解している。
    b. 解釈の容易さのために相対的に単純な（いくつかの変数を含まない）モデルが好まれること、また相対的に単純なモデルの方が汎化誤差（訓練データで推定したモデルをテストデータに適用したときの誤差）を小さくする傾向があることを知っている。

c. b で述べた特質のために、弱い効果しか持てないマイノリティの影響が無視されるなど、特定のグループに対する差別が生じうることを理解している。

### 3.4 ★公表バイアス

1. 解析結果の意義が小さいと判断されたときには、公表されないことも多いため、第三者が触れることができる解析結果は、研究者等によって意義が大きいと判断されたものに偏りがちであるという公表バイアス（出版バイアス）の存在を理解している。
2. 公表バイアスのために治療法や薬の有効性と安全性に関する誤認が起こり、多くの人の健康に悪影響を及ぼす危険性があることを理解している。
3. 解析データやロジックの公開、事前に調査内容を登録しておく制度（臨床試験登録制度）など公表バイアスを減らす国際的な取り組みが進んでいることを知っている。

# データの記述・可視化

## 1.　データの記述

データを適切な方法で要約して記述することは、データの分布の特徴を理解するための基礎的スキルである。ここでは、データには量的変数と質的変数などいくつかの型があること、一変量データに関する標本平均や標準偏差などの基本統計量の概念、多変量データの相関関係を相関係数や回帰分析によって記述する方法などを学ぶ。

### 1.1　☆種々のデータ

1. データの型として、質的変数、量的変数の違いを具体例を踏まえて理解している。さらにより細かい分類である名義尺度、順序尺度、間隔尺度、比例尺度の特徴を、具体例を踏まえて理解している。
2. 時系列データの具体例を挙げられる。季節調整の基礎である移動平均の考え方を理解している。また GDP など多くの公的統計では、原系列から季節調整を行って得られる数値が公表されていることを知っている。

### 1.2　☆基本統計量

1. 平均値、中央値、標準偏差、四分位数、四分位範囲などの基本統計量の概念を理解している。
2. 標準得点（Z 得点）や偏差値など、分布における相対的な位置の概念を理解している。
3. 何らかの統計ソフトウエアでそれらの基本統計量を計算できる。

### 1.3　☆相関関係

1. ２つの量的変数の線形関係の強さの指標である相関係数について、正負の符号の意味や値の取りうる範囲などの特徴を理解している。
2. ３つ以上の量的変数に関する相関係数行列の各成分の意味を理解している。
3. 単回帰分析について、係数の最小二乗値の意味を散布図に当てはめた回帰直線に基づいて理解できる。
4. ２つの質的変数（名義尺度、順序尺度）について、データセットから分割表を作成できる。また、オッズ比などを踏まえて変数間の関係を理解できる。
5. 統計ソフトウェアを用いて、２つの質的変数から分割表（クロス集計表）を出力できる。また２つの量的変数の相関係数を計算できる。さらに単回帰分析を実行できる。
6. 相関と因果の相違点、交絡の概念を身近な例を踏まえて理解している。

## 2. データの可視化

ビッグデータ時代において、データの可視化は大規模データから本質を取り出し、データ駆動型の意思決定を行うための重要なスキルである。効果的な可視化により、言葉では

伝えるのが難しいデータの重要な特徴やパターンを、他者に伝えることができる。またデータサイエンスのリテラシーとして、データの特徴を正確に可視化するための技術、及び他者の作った不必要な視覚的要素などを含むグラフに騙されないためのポイントを身につける。

### 2.1 ☆グラフの構成要素

1. グラフにおける次のような色の効果、特徴を理解している。
    a. 順序を持たない離散的な項目を区別するための質的な色の利用
    b. 同系色の濃淡などによる、連続的な量的変数あるいは順序を持つ離散的な項目の表現
    c. 特徴的な個体などを強調するための鮮明な色と不鮮明な色の対比
2. 色覚特性に配慮した配色でグラフを描ける。
3. 点の色、大きさ、形状の違いに注意できる。（２カテゴリー以上に分かれる２変量データを同時に散布図に描くときなど）
4. 線の太さと様々な破線の使い分けができる。（複数の折れ線グラフを同時に描くときなど）

### 2.2 ☆統計グラフ

1. 【棒グラフ】棒グラフは比例尺度の量を表すためのグラフであること、また棒の長さ（面積）が量に比例するように描くことを理解している。
2. 【棒グラフ】棒の長さ（面積）を量に比例させる原則のために、縦軸をゼロから始めずに長さを切り詰めることが誤解を生むことを理解している。
3. 【折れ線グラフ】特に時系列データについて、折れ線グラフが（横軸の）時間とともに数量が変わる様子を表現できることを理解している。
4. 【折れ線グラフ】気温などの具体例を想定しながら、縦軸を必ずしも 0 点から始める必要はないことを理解している。
5. 【円グラフ】円グラフは扇形の中心角の大きさで各カテゴリーの全体における割合を表すことを理解している。
6. 【帯グラフ】割合のカテゴリー間の比較には、円グラフよりも帯グラフの方が望ましいことを理解している。
7. 【チャートジャンク】影、グラデーション、3D、矢印などの不必要な視覚的要素（チャートジャンク）が、誤解を招くことを知っている。
8. 【チャートジャンク】チャートジャンクを含むグラフに遭遇したとき、必要に応じて元データを参照するなどの判断ができる。

### 2.3 ☆分布の統計グラフ

1. 【ヒストグラム】ヒストグラムに基づいて、「単峰か否か」「歪みがあるか」などのデータの分布の特徴を理解できる。
2. 【ヒストグラム】ビンの個数、階級の幅によって印象が変わることを理解している。

3. 【箱ひげ図】代表的な分位点により分布を要約する箱ひげ図を用いて、データの分布の特徴を読み取ることができる。
4. 【箱ひげ図】少数の要約統計量に基づいているために、例えば異なるヒストグラムの形となる２つのデータセットから、同じ箱ひげ図が作られる可能性があることを理解している。
5. 【散布図】２つの量的変数に対し、散布図と相関係数の大きさ・絶対値を対応させることができる。
6. 【散布図】散布図をもとに、二つの変数が線形な関係を持つか否か、関係は強いか否か等の基本的な検討ができる。
7. 【散布図】２カテゴリー以上に分かれる２変量データを同時に散布図に描くときに、点の色、大きさ、形状に注意できる。
8. 【散布図】３つ以上の量的変数に関する散布図行列について、各成分の意味を理解している。また相関係数行列に基づくヒートマップから、強い線形関係を持つ２変数の組を抽出できる。

# データの取得・管理・加工

## 1.　データ取得とオープンデータ

データの取得はデータサイエンスの第一歩である。オープンデータを巡る国内外の動きを理解して、関心のあるデータ、課題解決のために有用なデータを探し出してダウンロードできるようになる。同時にメタデータなどからデータ取得の経緯や変数の特徴を理解して、課題解決の目的に適したデータであるか検討できることも重要である。

### 1.1　☆日本や世界のオープンデータ

1. 日本政府のオープンデータの取り組みを知り、オープンデータの意義を理解している。特に、「二次利用可能なルール」「機械可読性」の重要性を理解している。
2. オープンデータを利用した国、地方公共団体、民間事業者の取り組みを一つ以上挙げられる。
3. 2013 年の G8 サミットで合意された「オープンデータ憲章」など、オープンデータの普及に向けた国際的な動きを知っている。

### 1.2　☆オープンデータの取得

1. e-Stat、DATA.GO.JP データカタログサイト（日本）、Open Knowledge International、国際連合（世界）などを通じて、関心のあるデータセット（csv 形式）をダウンロードできる。
2. ダウンロードしたデータのメタデータなどに基づいてデータの取得された経緯や変数の特徴などを理解できる。

## 2　データ管理とデータ形式

データサイエンスでは、課題の発見からその解決に至るプロセスで、データが常に中心に据えられる。そのようなデータをファイルとして蓄積するためのデータの様々な形式を理解し、適切に管理する方法を学ぶ。

### 2.1　☆表形式のデータ

1. csv 形式が行と列からなる表形式のデータを表現できることを理解している。
2. e-Stat 等からダウンロードされたデータを、メタデータ部分を取り除くなどの前処理により、何らかのソフトウェアに読み込める csv ファイルに編集できる。

### 2.2　★その他のデータ形式

1. Web API（Application Programming Interface）によるデータが提供される場合の標準的なデータ形式である、XML 形式（属性と要素の組み合せ）や JSON 形式（JavaScript のデータ定義文をベースとした、簡易的なデータ定義言語）を知っている。

2. データ間の関係性やつながり表す場合など、表形式以外の方法でデータを表現した方が扱いやすいことを理解している。
3. SNS での人のつながりや鉄道の路線などのような関係性について、頂点と辺で構成された離散グラフとして表現できることや、離散グラフがキー・バリュー形式である隣接リストで表現できることを理解している。

### 2.3 ★データベース

1. 【一般】データベースとは、ある目的のために収集した大量の情報を一定の規則に従って体系的に整理・蓄積し、検索や抽出に利用するための仕組みであることを理解している。
2. 【一般】システムの障害や不正アクセスから守るために、データベースを運用・管理するソフトウェアとしてデータベース管理システム（DBMS）が必要であることを理解している。
3. 【リレーショナルデータベース】レコード（行）とフィールド（列）からなる表形式で表現されたテーブル（表）の集合としてデータを管理し、それらのテーブルを共通項目で関連づけられる（リレーションシップ）ことを理解している。
4. 【リレーショナルデータベース】テーブル（表）の分割をもとにデータの矛盾や重複などを無くす正規化により、効率的にデータ管理ができることを理解している。正規化されたテーブルでは、そのテーブルの中で一意に行を識別する主キーを持つことを理解している。
5. 【リレーショナルデータベース】テーブル（表）に対する３つの操作（関係演算）である「選択」・「射影」・「結合」を知っている。これらの操作は SQL（Structured Query Language）と呼ばれる言語を利用して行うことを理解している。
6. 【キー・バリュー型データベース】リレーショナルデータベース以外の代表的なデータベースとしてキー・バリュー型のデータベースがあり、特に大規模データの処理に活用されていることを知っている。また、他のデータ形式に対応したデータベースを含めて NoSQL と呼ばれること、NoSQL がビッグデータの活用と関連して注目されていることを知っている。

## 3. データの前処理

データ分析する前の前処理が、分析結果の精度や効率性に影響する。そのために適切な前処理がデータサイエンスにとって必要不可欠な工程であることを認識する。

### 3.1 ★データクレンジング

1. 名字や住所などを例として、文字列の表記揺れ対策の必要性を理解している。
2. 数字の表記の揺れを吸収する必要性を理解している。（整数部分と少数部分の区切りに「,」と「.」が使われることや、桁区切り文字の地域による違いなど）
3. 時系列データにおける日付、時刻の表記の揺れを吸収する必要性を理解している。

### 3.2 ★外れ値、異常値、欠損値

1. 箱ひげ図、標準化得点などをもとに、外れ値や異常値について検討できる。
2. 必要に応じて、データが取得された状況に立ち返って入力ミスなどを検討する必要性を理解している。
3. データ欠損が起こる要因をいくつか具体例をもとに理解している。（観測されるはずであったその値の大小が欠損に影響する場合など）
4. 標本平均などを用いた、欠損値の（自動）補完について検討できる。

### 3.3 ★データ加工

1. レコード（行）とフィールド（列）からなる表形式で表現されたデータ行列から、ある変数に注目し、その値に基づいてデータ行列の部分集合を抽出できる。
2. 行を並び替えられる（例えばある変数に注目して昇順にする）。
3. 変数の絞り込みによりデータ行列の部分集合を抽出できる。
4. 既存の列に四則演算を含む簡単な関数を適用して、新しい列を追加したデータ行列を作成できる。

# 統計基礎

## 1. 確率と確率分布

データを確率的誤差を持つものとして理解するという統計的推測の基本原則により、研究課題の考察、処方箋において不確実性の定量的な提示が可能となる。そのようなデータの背後に仮定される確率構造の基礎を学ぶ。

### 1.1 ☆場合の数、順列・組み合わせ

1. 場合の数、順列・組み合わせの概念を理解している。またそれらに関する基本的な例題を計算できる。

### 1.2 ☆確率

1. 様々な事象の起こりやすさを数値で表す確率の概念を理解している。また加法定理、条件付き確率、ベイズの定理に関する基本的な例題を計算できる。

### 1.3 ☆確率分布、確率変数

1. 確率変数の意味を理解している。
2. 離散確率変数を特徴づける確率関数、連続確率変数を特徴づける確率密度関数の概念や性質を理解している。
3. 確率変数の期待値（平均）や分散の定義を理解している。また簡単な例で期待値や分散を計算できる。
4. ２つの確率変数の共分散・相関係数・独立性の概念を理解している。

### 1.4 ☆主要な確率分布

1. ２項分布の定義を理解している。またその期待値と分散を知っている。
2. 正規分布のパラメータ（平均と分散）に対応して描かれた確率密度関数の複数のグラフをもとに、正規分布の分布の位置やばらつきの違いを理解できる。
3. その他の重要な分布（ポアソン分布、指数分布、一様分布など）について、確率関数、確率密度関数のグラフに基づいて、その性質を理解している。またそれらの分布に従うと考えられる事象の具体例を挙げられる。
4. ２変量正規分布の確率密度関数の等高線や鳥瞰図に基づいて、２変数の相関の強さについて理解できる。

## 2. データ収集法と確率構造

データを収集するとき、調査の対象となる母集団をもれなく調べる全数調査は多くの場合不可能である。そこで、標本を抽出して調査し、その結果から母集団の性質を推測する標本調査が必要となる。そこに現れる確率構造といくつかの抽出方法を学ぶ。また、データ収集に関連して、「根拠に基づく医療」、「根拠に基づく政策立案」のためのデータ収集デザインであるランダム化比較試験の考え方を学ぶ。

### 2.1 ☆標本調査

1. 抽出された標本の調査結果から、母集団の状況をできるだけ正確に推測するという標本調査の目的を理解している。
2. 母集団の平均と標本の平均が必ずしも一致しないことをシミュレーションによって確認できる。
3. 単純無作為抽出法の原理と意義を理解している。

### 2.2 ★ランダム化比較試験

1. ある要因の効果を測定する場合に、ランダム化比較試験のデザインが注目している要因以外の影響を確率的に取り除く原理を理解している。
2. ランダム化比較試験では倫理的に許されないケースがあることを知っている。（喫煙の健康に与える影響を測定するために、ランダムに分けた２グループに対して、「１０年間喫煙を強いる」「１０年間禁煙を強いる」など）
3. インターネットマーケティング分野で普及しているA/Bテストは、ランダム化比較試験の考え方を基礎にしていることを知っている。

## 3. 統計的推測

データサイエンスにおいては、しばしば統計的推測の立場からデータに基づく推論が行われる。統計的推測の考え方は、データを確率的誤差込みのものと捉えて、複数の確率分布を含む候補の集合の中からデータと整合的な分布を推測するというものである。その候補の集合はパラメータによって表現されており、統計的モデルと呼ばれる。データに基づいて統計的モデルのパラメータの値を推測する代表的な方法が、点推定、区間推定、仮説検定である。

### 3.1 ★統計的モデル

1. データの背後に複数の確率分布の候補を表現するという統計的モデルの考え方を理解している。

### 3.2 ★標本分布

1. 統計的モデルのもとでは、標本平均、標本分散などの統計量が確率変数であり固定した値ではないことを理解している。

### 3.3 ★点推定、区間推定

1. 母平均の推定において、標本平均や中央値など複数の推定量が考えられることを理解している。
2. 複数の推定量が考えられる状況で、不偏性や平均二乗誤差といった推定量の良さに関する性質をもとに使うべき推定量を選択できる。
3. 区間推定の概念及び信頼係数の意味を理解している。

### 3.4 ★仮説検定の概念

1. 二項分布において帰無仮説「成功確率が 1/2 である」対立仮説「1/2 でない」の場合の P 値を計算できる。また予め与えられた有意水準と比較して、帰無仮説を棄却するかどうかの判断ができる。
2. 正規分布の母平均の検定において、また片側検定・両側検定の両方で、正規分布の密度関数のグラフに基づいて P 値を理解している。また予め与えられた有意水準と比較して、帰無仮説を棄却するかどうかの判断ができる。

# 数学基礎

## 1. 線形代数

多変数のデータに関する基本操作を学ぶ。データサイエンスでは、ベクトルによってデータを表現し、表形式のデータをデータ行列と見なして様々な演算を行う。特にリテラシーとして、データの記述やモデリングの表現で頻出するベクトルや行列の和・スカラー倍、ベクトルの内積、行列の積に慣れる。

### 1.1 ☆ベクトル

1. 複数個の数値の組（例：ある学生のｐ個の科目の試験の点数、ある科目のｎ人の学生の試験の点数）をベクトルと見なせることを理解している。
2. ベクトルの和とスカラー倍を計算できる。
3. 2 つのベクトルの内積を計算できる。
4. 与えられたｎ次元データベクトル（サンプルサイズｎの一変量データ）に対して、標本平均や加重平均がデータベクトルと加重係数ベクトルの内積と見なせることを理解している。

### 1.2 ★行列

1. 縦×横の表形式で並んだ複数個の数値を行列と見なせることを理解している。
2. ｎ × ｐ行列がｐ本のｎ次元列ベクトル（縦ベクトル）の組、ｎ本のｐ次元行ベクトル（横ベクトル）の組と見なせることを理解している。
3. 行列の和とスカラー倍を計算できる。
4. 様々なサイズの２つの行列の積を計算できる。特に、2 の観点から積の成分が２つのベクトルの内積で表されることを理解している。
5. ｎ人のクラスのｐ個の科目のテストの点数からなる表形式のｎ × ｐデータ行列について、各人の合計点のｎ次元ベクトルや各科目の平均点のｐ次元ベクトルが、行列とベクトルの積によって得られることを理解している。

## 2. 微積分

確率・統計及びデータモデリングに現れる関数についての主要な操作を学ぶ。１変数関数は１変量の確率・統計の基本となり、２変数関数は多変量の確率・統計の基礎となる。微分はデータモデリングに使われる関数の局所的振る舞いを理解するための道具であり、積分は確率変数の期待値計算や区間領域の確率など大域的な性質を考察するための道具となる。

### 2.1 ☆多項式関数

1. 二次関数について理解し、二次関数を用いて数量の関係や変化を表現できる。
2. 多項式関数の四則演算を計算できる。また与えられた多項式のグラフを描ける。
3. 二項係数と二項定理を理解している。

### 2.2 ★初等関数

1. 指数関数（ネイピア数（e）を底とする指数関数を含む）、対数関数の基本性質を理解している。またそれらのグラフを描ける。
2. 多項式と指数関数の、増加あるいは減少のオーダーの違いを理解している。
3. 三角関数の基本性質を理解している。またそれらのグラフを描ける。
4. 関数値の極限の概念を理解し、lim の記法を身につけ、簡単な例で極限値を計算できる。

### 2.3 ★1 変数関数の微分法

1. 関数の微分の意味を理解し、初等関数の導関数を計算できる。
2. 関数の定数倍、和及び差の導関数を求めることができる。
3. 導関数を用いると、関数の値の増減や極大・極小を調べたり、グラフの概形を描けることを理解している。

### 2.4 ★1 変数関数の積分法

1. 不定積分及び定積分の意味を理解している。
2. 初等関数の定数倍、和及び差の不定積分や定積分を計算できる。
3. 直線や関数のグラフで囲まれた図形の面積を定積分として計算できる。

### 2.5 ★2 変数関数の微分法

1. 2 変数関数のグラフが 3 次元空間内の曲面として表現できることを理解している。
2. 2 変数関数の偏導関数を計算できる。また 2 変数関数の鳥瞰図において偏導関数の意味を理解している。
3. 極値と偏導関数の関係を理解している。また 2 変数関数の鳥瞰図が与えられたとき、極大・極小・鞍点（停留点）のいずれであるか判定できる。

### 2.6 ★2 変数関数の積分法

1. 長方形領域での 2 変数関数の重積分について、鳥瞰図に基づいた積分値の説明を理解している。

## 3. 数列

ベクトルや行列の演算を成分毎に記述する場合や、一般にデータサイエンスの手法を記述する際に、数列の和の記号 Σ が頻出する。記法に慣れることで概念を深く理解することが期待される。

### 3.1 ☆データと数列

1. 数列の和の記号 Σ の意味を理解している。
2. 「データの記述」の標本平均、標本分散、標本相関係数などを Σ を使って書ける。

### 3.2 ★数列

1. 等差数列や等比数列などの簡単な数列について、一般項を表現したり第 n 項までの和を求めることができる。
2. 数列の極限や数列の和の極限を計算できる。

## 4. 演習

データサイエンスでは、行列演算、関数の最大化など線形代数、微積分に関する様々な計算が登場する。計算方法を理論的に理解するだけでなく、学生が手元の計算機で実際に計算できることが重要である。例えば Excel、R、Python の利用が想定される。

### 4.1 ★線形代数演習

1. 行列やベクトルを、配列として表現できる。
2. 既存のプログラミング言語やアプリケーションで定義された関数を用いて、行列やベクトルの和、スカラー倍、ベクトルの内積、行列の積を計算できる。

### 4.2 ★微積分演習

1. 既存のプログラミング言語やアプリケーションで定義された関数を用いて、数値積分や数値的最適化を実行できる。

# 計算基礎

## 1. 情報、コンピュータの仕組み

数値やデータをコンピュータで扱う場合の基礎として、2 進数による数の表現やデータ量の表し方、集合と論理演算の基本的な考え方などについて理解する。また情報のデジタル化を科学的に理解し、目的や状況に応じて適切にデータを取り扱う力を身につける。またコンピュータを用いて数値計算やデータ解析を行う際に、コンピュータ内部では 2 進数を大きな桁で非常に高速に論理演算することにより計算されること、また原理的に計算誤差が生じうること、などを知っていることは強みになる。

### 1.1 ★数と表現

1. 数を 2 進数で表現できる。2 進数の加減算ができる。

### 1.2 ★有効数字、計算の誤差

1. コンピュータでは定められたビット数のデータが扱われていることを理解している。
2. コンピュータでは表現できる値の範囲や精度が有限であることを理解している。
3. コンピュータは計算処理を有限の桁数で行うこと、またそのために計算処理には誤差が生じうること、を理解している。

### 1.3 ★データ量の単位

1. ビット、バイトなどのデータ量の単位や、SI 接頭辞（k,M,G,T,P,m, $\mu$ ,n,p など）を使ったデータ量の表現を理解している。

### 1.4 ★デジタル化

1. アナログとデジタルの特徴を理解している。
2. コンピュータで扱われる画像データには、ラスタデータとベクタデータがあることを知っている。またそれぞれの特徴を理解している。
3. 標本化の精度や量子化のレベルによって、画像などのファイルサイズが違うことを理解している。

### 1.5 ★文字の表現

1. コンピュータの内部では、文字も記号も 2 進法のデータとして扱われ，文字コードと呼ばれる固有の番号が割り当てられていることを知っている。

### 1.6 ★集合と命題

1. 集合と命題に関する基本的な概念（用語、記法、包含関係、和集合、積集合など）を理解している。
2. ベン図を用いて集合と命題に関する基本的な概念を理解している。

### 1.7 ★論理演算

1. 真理値表の基本的な考え方と利用方法を理解している。
2. 論理演算の考え方が分かり、真理値表を用いて基本的な演算を実行できる。

## 2. データ構造

コンピュータ内では、データを変数へ格納したり（代入）、変数から値を取り出す（参照）操作を繰り返しながら、計算処理を進めていく。特に複雑な計算プログラムを作成する場合に、簡易かつ効率的に変数を取り扱うために適切なデータ構造があることを理解する。

### 2.1 ★配列、リスト

1. 複数のデータを一つの変数名と添え字で管理する記述を理解している。
2. 探索アルゴリズムなどに応じて適切なデータ構造があることを理解している。

### 2.2 ★連想配列

1. 項目（キー）と値（バリュー）を組とするデータ構造を理解している。
2. 項目・値形式のデータが、データ間の関係を見出したり、データを検索する処理に有用であることを理解している。さらにデータベースに広く活用されていることを理解している。

## 3. アルゴリズムとプログラミング

データサイエンスにおいては、アルゴリズム的な課題解決が不可欠である。明確に要件定義し、問題を分解し、アルゴリズム的な解法に至る効率的な戦略が必要となる。さらに適切な高水準言語によるプログラミングにより、解法を実装する必要がある。そのためにアルゴリズムとプログラミングの基礎的素養を身につける。またデータサイエンスで広く使われるExcel（VBA）、R、Pythonの基礎的スキルを身につける。

### 3.1 ★アルゴリズム

1. 文章、フローチャート、アクティビティ図などによって表現されたアルゴリズムを、順次構造・選択構造・繰返し構造といった基本構造を踏まえて理解している。
2. アルゴリズムによって処理の結果や効率に違いが出ることを、代表的なアルゴリズムの例を踏まえて理解している。

### 3.2 ★プログラミング

1. アルゴリズムをプログラムとして表現する方法を理解している。例えばデータ型、演算子、変数、条件分岐、繰り返し、関数、ライブラリ、APIを理解している。
2. 作成したプログラムの動作確認や、不具合の修正の重要性を理解している。
3. Excel（VBA）、R、Pythonのいずれかのプログラミング言語で単純なアルゴリズム（例えば数値の合計、平均、最大、最小）を表現できる。

# モデリングと評価

## 1. 教師あり学習

データサイエンスにおいては、統計学や機械学習の様々な手法が用いられる。その多くは教師あり学習（supervised learning）と教師なし学習（unsupervised learning）に分類される。前者は、何らかの変数（目的変数）を別の変数（説明変数）で予測あるいは説明したい場合のモデリングである。教師あり学習の代表的なモデリング手法として、目的変数が連続変数の場合は線形回帰分析、２値（つまり２つのカテゴリーへの分類に対応）の場合はロジスティック回帰がそれぞれあげられる。回帰分析とロジスティック回帰に習熟することは、より複雑なモデリング手法を理解するための基礎となる。

### 1.1 ★回帰分析

1. ２つの説明変数に対する重回帰分析における最小二乗法、係数の最小二乗推定値を、回帰平面を含む鳥瞰図に基づいて理解できる。
2. 自分で収集したデータあるいは e-Stat などからダウンロードしたデータを、統計ソフトウェアに読み込み、回帰分析を実行して結果の解釈を行うことができる。
3. 散布図において説明変数と目的変数に非線形な関係が見られる場合に、説明変数の多項式関数による重回帰分析（例：切片、$x$、$x^2$による回帰分析）を実行できる。

### 1.2 ★ロジスティック回帰

1. 目的変数が離散値（特に２値）の場合には回帰分析が不適切であることを、散布図を通じて理解できる。また不適切さを回避できるロジスティック回帰モデルの意義を理解している。
2. 係数の標準的な推定法として最尤法が使われることを知っている。
3. 推定された係数の解釈（説明変数が１単位増加した場合の目的変数への影響）を回帰分析の場合との違いを比較しながら理解できる。
4. 自分で収集したデータあるいは e-Stat などからダウンロードしたデータをソフトウェアに読み込み、ロジスティック回帰分析を実行して結果の解釈を行うことができる。

## 2. 教師なし学習

クラスタリングは教師なし学習の代表的な手法であり、共通の特徴を持ったいくつかの集団にデータを分割することを目的とする。類似したデータを集団にしていく階層的方法と、集団の数を決めてからデータを所属させていく非階層的方法がある。代表的な手法として、階層クラスタリングと k-平均法をそれぞれ学ぶ。

### 2.1 ★クラスタリング（k-平均法）

1. クラスタ中心の更新とデータの中心への割り当ての繰り返しという k-平均法の手続きを理解している。

2. k-平均法によるクラスタリングが有効な実例を挙げられる。
3. 自分で収集したデータあるいは e-Stat などからダウンロードしたデータを、R や Python などのソフトウェアに読み込み、k-平均法を実行して結果の解釈を行うことができる。
4. 初期値や距離を変えてみて、結果の頑健性について考察できる。

### 2.2 ★階層クラスタリング

1. データ間の距離（類似度）とクラスタの併合法に基づくクラスタの逐次的な併合の繰り返しという階層クラスタリングの手続きを理解している。
2. 階層クラスタリングが有効な実例を挙げられる。
3. 樹形図（デンドログラム）の読み取り方と意義を理解している。
4. 自分で収集したデータあるいは e-Stat などからダウンロードしたデータを、R や Python などの統計ソフトウェアに読み込み、階層クラスタリングを実行して結果の解釈を行うことができる。
5. データ間の距離（類似度）やクラスタの併合法の選択によって、結果が変わることを確認できる。

## 3. モデルの評価

複数のモデルが考慮される場合、個々のモデルの性能を定量的に評価し、その値に基づいたモデル間での相対比較により、「良い」モデルが選択される。性能の評価指標は一通りではなく複数考えられるが、ここでは回帰問題における平均二乗誤差、二値分類問題における混同行列を学ぶ。また教師あり学習における性能評価の基礎である、訓練データで学習させたモデルの性能を、訓練データとは異なるテストデータで試すという考え方に慣れる。

### 3.1 ★評価指標

1. 回帰問題において、平均二乗誤差が一つの評価指標となることを理解している。
2. 二値分類問題において、混同行列（の真陽性、偽陽性、真陰性、偽陰性）に基づく正解率（accuracy）、適合率（precision）、再現率（recall）などが評価指標となることを理解している。

### 3.2 ★訓練データとテストデータ

1. 訓練データをもとに特定のモデルのパラメータを推定して、それをテストデータにおける評価関数の値によってそのモデルの性能を測るという考え方を理解している。
2. 訓練データにおいて当てはまりが良い複雑なモデルが、テストデータにおいて必ずしも性能が良いとは限らない過学習（overfitting）を理解している。
3. 訓練データの多様性の不足などにより単純すぎるモデルを作ってしまうため、訓練データでもテストデータでも性能が落ちる適合不足（underfitting）を理解している。

4. 交差検証法（cross validation（leave-one-out））の考え方を理解している。例えば、３次式、２次式、１次式による回帰モデルのうち、相対的に優れたモデルを、交差検証法による予測平均二乗誤差の大きさによって選ぶことができる。

附属資料１

# 高校学校学習指導要領とスキルセット及び学修目標の対応

※ 下線は中分類との対応を示す。

## 数学（現行学習指導要領）

### 数学 Ⅰ

**数と式**

数と集合　式　計算基礎：情報、コンピュータの仕組み

**図形と計量**

三角比　図形の計量

**二次関数**

二次関数とそのグラフ　二次関数の値の変化　数学基礎：微積分

**データの分析**

データの散らばり　データの相関　データの記述・可視化：データの記述　データの記述・可視化：データの可視化

### 数学 II

**いろいろな式**

式と証明　高次方程式

**図形と方程式**

直線と円　軌跡と領域

**指数関数・対数関数**

指数関数　対数関数　数学基礎：微積分

**三角関数**

角の拡張　三角関数　三角関数の加法定理　数学基礎：微積分

**微分・積分の考え**

微分の考え　積分の考え　数学基礎：微積分

### 数学 III

**平面上の曲線と複素数平面**

平面上の曲線　複素数平面

**極限**

数列とその極限　関数とその極限　数学基礎：微積分　数学基礎：数列

**微分法**

導関数　導関数の応用　数学基礎：微積分

**積分法**

不定積分と定積分　積分の応用　　数学基礎：微積分

## 数学 A

**場合の数と確率**

場合の数　確率　　統計基礎：確率と確率分布

**整数の性質**

約数と倍数　ユークリッドの互除法　整数の性質の活用

**図形の性質**

平面図形　空間図形

## 数学 B

**確率分布と統計的な推測**

確率分布　正規分布　統計的な推測　統計基礎：確率と確率分布　統計基礎：統計的推測

**数列**

数列とその和　漸化式と数学的帰納法　数学基礎：数列

**ベクトル**

平面上のベクトル　空間座標とベクトル　数学基礎：線形代数

# 情報（次期学習指導要領）

## 情報 I

**情報社会の問題解決**

- 情報やメディアの特性．情報と情報技術を活用した問題の発見・解決
- 情報に関する法規や制度，情報セキュリティ，情報社会における個人の責任及び情報モラル　データの法規・倫理：情報倫理、情報セキュリティ　データの法規・倫理：データに関する法律・規制
- 情報技術が人や社会に果たす役割と及ぼす影響

**コミュニケーションと情報デザイン**

- メディアの特性とコミュニケーション手段　計算基礎：情報、コンピュータの仕組み
- 情報デザインが人や社会に果たしている役割
- 効果的なコミュニケーションを行うための情報デザイン

**コンピュータとプログラミング**

- コンピュータや外部装置の仕組みや特徴，コンピュータでの情報の内部表現と計算に関する限界　計算基礎：情報、コンピュータの仕組み
- アルゴリズムを表現する手段，プログラミングによるコンピュータや情報通信ネットワークの活用　計算基礎：アルゴリズムとプログラミング
- 事象のモデル化，シミュレーションを通したモデル評価と改善

**情報通信ネットワークとデータの活用**

- 情報通信ネットワークの仕組みや構成要素，プロトコルの役割及び情報セキュリティを確保するための方法や技術
- データを蓄積，管理，提供する方法，情報通信ネットワークを介して情報システムがサービスを提供する仕組みと特徴　データの取得・管理・加工：データ管理とデータ形式
- データを表現，蓄積するための表し方．データを収集，整理，分析する方法　データの記述・可視化：データの記述　データの記述・可視化：データの可視化

附属資料２

# 導入用の様々な事例

※　本資料は、概論や授業の導入等において活用できる様々な事例及びキーワードを提示したものである。各事例の詳細は、「●」で示す参考文献や URL 等を参照願いたい。

なお、数理・データサイエンス教育強化拠点コンソーシアム 教育用データベース分科会では、データサイエンス教育に活用可能なデータ（実験データ、調査データ、地域の生データ、ビジネスデータ、ネット情報等）の提供システムを構築中である。身近な話題や企業における実際の事例を題材としたケーススタディやグループワーク等を通して、学びの動機付けや実務・実践で活用できる能力の育成等が図られることを期待している。

## 1　様々な分野の事例

### 1-1　総務省統計局「なるほど統計学園高等部」の事例

○ 「ビジネス・金融」「ICT」「医療・科学」「行政」のどのような場面で統計学、データが使われているかについての具体的なエピソード

○ ビジネス・金融「保険料の算定」、ICT「迷惑メールの判別」、医療・科学「薬の効果を調べる」、行政「地方交付税の配分」など 21 の事例

● 総務省統計局 なるほど統計学園高等部 > 豆知識 > 意外なところに統計学
https://www.stat.go.jp/koukou/trivia/careers/index.html

### 1-2　マーケティング

○ 消費者のニーズ把握とアンケート調査

○ 需要予測と回帰分析

○ 顧客のセグメンテーション(属性に応じたクラスタリング)

○ 広告戦略と A/B テスト

○ 推薦システムと相関係数

● 竹村彰通・姫野哲人・高田聖治編：データサイエンス入門，学術図書出版社(2019) 5.1 節

### 1-3　金融

○ ポートフォリオセレクションと確率、最適化

○ 企業への融資、倒産確率とロジスティック回帰

○ 中途解約などの顧客行動の分析と教師あり学習

○ 保険と確率、リスクの細分化と倫理

● 竹村彰通・姫野哲人・高田聖治編：データサイエンス入門，学術図書出版社(2019) 5.2 節

### 1-4　消費動向指数

○ 民間企業が保有する様々な消費関連情報を活用した消費動向指数の改良に向けた動き

● 竹村彰通：データサイエンス入門，岩波新書(2018) pp.88-89

● 総務省統計局：消費動向指数　https://www.stat.go.jp/data/cti/index.html(最終閲覧日 2019.8.26)

### 1-5　選挙

○ 世論調査による選挙結果予測の成功例と失敗例
○ 英国の EU 離脱、ヒラリークリントンとトランプ
○ ギャラップの世論調査(1936 年のルーズベルトの当選予想)
○ 標本選択バイアスなど
● 今井耕介著，粕谷祐子・原田勝孝・久保浩樹訳：社会科学のためのデータ分析入門（上），岩波書店(2018) 4.1 節
● 統計学習の指導のために（先生向け）補助教材 統計エピソード集～授業の導入部で～ アメリカ大統領選挙の番狂わせ ～ 標本調査における偏り
http://www.stat.go.jp/teacher/c2epi4a.html

### 1-6　2008 年アメリカ大統領選挙におけるオバマ陣営の選挙活動

○ RCT によるメーリングリスト加入率の最適化
● 伊藤公一朗：データ分析の力 因果関係に迫る思考法，光文社新書(2017) p.83

### 1-7　著者推測

○ テキストの計量分析（フェデラリスト、シェークスピア）
○ 自然言語処理、単語の使用頻度指標、ワードクラウド
● 今井耕介著，粕谷祐子・原田勝孝・久保浩樹訳：社会科学のためのデータ分析入門(下)，岩波書店(2018) 5.1 節
● 村上征勝：真贋の科学 -- 計量文献学入門，朝倉書店(1994)

### 1-8　品質管理

○ QCD(Quality, Cost, Delivery)
○ 統計的品質管理とデータサイエンス
● 竹村彰通・姫野哲人・高田聖治編：データサイエンス入門，学術図書出版社(2019) 5.3 節

### 1-9　ネットワークデータ

○ Twitter、SNS におけるフォロー関係の可視化やクラスタリング
● 今井耕介著，粕谷祐子・原田勝孝・久保浩樹訳：社会科学のためのデータ分析入門(下)，岩波書店(2018) 5.2 節

### 1-10　画像処理

○ 画像データの表現
○ データサイエンスと画像処理技術
○ 医療における画像診断支援
● 竹村彰通・姫野哲人・高田聖治編：データサイエンス入門，学術図書出版社(2019) 5.4 節

### 1-11　医学

○ 根拠に基づく医療
○ 新薬の治験とサンプリング法
○ 生物学・分子生物学に基づく医療(連鎖解析、家系解析)
● 竹村彰通・姫野哲人・高田聖治編：データサイエンス入門，学術図書出版社(2019) 5.6 節

### 1-12　医学データに基づく意思決定

○ CAST 試験 心筋梗塞
○ Women's Health Initiative 女性ホルモン補充療法
○ 既存の治療法が間違っていたというエビデンス
● 竹村彰通：データサイエンス入門，岩波新書(2018) pp.83-87

## 2 可視化

### 2-1　ナイチンゲールと統計

○ イギリス政府によって看護師団のリーダーとしてクリミア戦争に派遣されたナイチンゲール
○ イギリス軍の戦死者・傷病者に関する膨大なデータを分析。彼らの多くが戦闘で受けた傷そのものではなく、傷を負った後の治療や病院の衛生状態が十分でないことが原因で死亡したことを解明
○ 統計になじみのうすい国会議員や役人にも分かりやすいように、当時としては珍しかったグラフを用いて、視覚に訴える報告
● 総務省統計局 統計学習の指導のために（先生向け）補助教材 統計エピソード集～授業の導入部で～ http://www.stat.go.jp/teacher/c2epi3.html

### 2-2　不適切なグラフや可視化

○ ねこでも分かる！いかさまグラフにはもうダマされない！！
○ 対話形式で不適切なグラフや可視化について解説
● 第一学習社 ed-ict 執筆：天野由貴 監修：奥村晴彦 http://www.ed-ict.net/entry/neco-demo-wakaru-graph-1

### 2-3　OECD 生徒の学習到達度調査（PISA）の出題例

○ 盗難事件数の棒グラフ：誤解を招くグラフ表現
● 小学校算数・中学校数学・高等学校数学 指導資料 －PISA2003（数学的リテラシー）及び TIMSS2003（算数・数学）結果の分析と指導改善の方向
http://www.mext.go.jp/a_menu/shotou/gakuryoku/siryo/05071101.htm
● PISA2003 調査－数学的リテラシー
http://www.mext.go.jp/a_menu/shotou/gakuryoku/siryo/05071101/001.pdf

## 3　データの倫理

### 3-1　ディオバン事件

○ 高血圧の治療薬であるディオバンの臨床試験に、N 社の社員が大学講師の肩書で統計解析者として関与
○ N 社と大学の利益相反に加えて、臨床試験においてデータの改竄があったとして一連の論文が撤回された
● 竹村彰通：データサイエンス入門，岩波新書(2018) pp.93-94

### 3-2　タスキギー梅毒人体実験 Tuskegee syphilis experiment

○ 参加者の同意なしに、また治療せずに、梅毒の経過を観察
○ 梅毒末期の様々な重篤症状を研究
○ 説明に基づく同意(Informed Consent)の必要性

● 星野一正：タスキギー梅毒人体実験と黒人被害者への大統領の謝罪，時の法令 1570 号，pp.45-51（1998）

### 3-3 Facebook とコーネル大学の心理学実験

○ グループ A に positive な news、グループ B に negative な news を提供
○ Facebook への post が news の positive，negative に引きずられる傾向
○ 意図的な感情操作ではないか？
● 日本経済新聞：米フェイスブックが謝罪、投稿操作し心理実験 論文で発覚（2014.7.2）https://www.nikkei.com/article/DGXNASGM0100I_S4A700C1000000/(最終閲覧日 2019.8.26)
● Adam D. I. Kramer，Jamie E. Guillory，and Jeffrey T. Hancock: Experimental evidence of massive-scale emotional contagion through social networks. PNAS 111（24）8788-8790（2014）

### 3-4 Netflix PRIZE

○ 匿名化データセットにおいて外部データベースとの照合による個人の再識別化
● 山口利恵：ビッグデータの利活用とプライバシー保護の難しさ，映像情報メディア学会誌，69 巻 2 号，pp.155-161(2005)

## 4 相関と因果

### 4-1 就寝時の照明と近視

○ 新聞などメディアが「電気をつけたまま寝かせると子供が近視になる」と報道
○ 親が近視であることが交絡因子であるとのフォローアップ論文
● 伊藤公一朗：データ分析の力 因果関係に迫る思考法，光文社新書(2017) p.45

### 4-2 One Laptop per Child

○ ノートパソコンの保有と成績に正の相関
○ 多くの国、国際機関、企業の協力で発展途上国の子供にノートパソコンを無償支給
○ その後のランダム化比較試験で成績への影響がないと結論付けられる。
● 伊藤公一朗：データ分析の力 因果関係に迫る思考法，光文社新書(2017) p.46

### 4-3 飲食店数と金融機関店舗数

○ 東京 23 区の飲食店数と金融機関店舗数に強い正の相関（相関係数=0.9）
○ 飲食店が多い地域に店舗を作るという戦略を金融機関がとっている（もしくはその逆）ではなく、昼間人口が多い地域に両者が立地しやすい傾向にある結果、見かけ上の相関が発生する。
○ 昼間人口が交絡要因
● 東京大学教養学部統計学教室 編：統計学入門，東京大学出版会（1991）．pp.50-54

### 4-4 教育と失業

○ 大恐慌期間中（1929～1933 年）のデータによると、高学歴の人ほど失業期間が短い傾向にあった。
○ しかし、この時代は若い人ほど学歴が高い傾向にあり、「高学歴だから失業しにくかった」のではなく「若いから失業しにくかった」という可能性も否定できない。

○ すなわち、上のデータから直ちに「学歴があると失業しにくい」と主張するのはミスリーディングとなる可能性がある。

● D. Freedman, R. Pisani, R. Purves: Statistics (4th ed.), Norton & Company (2007). pp.150-151

### 4-5 海難事故の件数とアイスの売上

○ 海難事故の件数とアイスの売上には、正の相関がある。

○ これは、「アイスの売上が増えると海難事故が増える」（もしくはその逆）という因果関係を表すのではなく、単に気温が上がるとアイスの売上が増え、一方で海に行く人が増えることで海難事故が増えるからである（気温が交絡要因となっている）。

● G. James, D. Witten, T. Hastie, R. Tibshirani: An Introduction to Statistical Learning, Springer (2013). p.74

● 落海 浩, 首藤 信通（翻訳）R による統計的学習入門, 朝倉書店 (2018)

### 4-6 チョコレート消費量とノーベル賞受賞者数

○ チョコレート消費量とノーベル賞受賞者数との正の相関

○ 「1 人あたり GDP」などの交絡要因

● 清水昌平：統計的因果探索, 講談社 (2017) 第 1 章

## 5 様々なバイアス

### 5-1 日本政策金融公庫の調査

○ 高校以上の学校に就学している子供を持ち国の教育ローンを利用している世帯において、世帯収入に対して教育費が平均 3 割を占め、年収が 200～400 万円の世帯では 5 割に達するという調査結果

○ ほとんどの全国紙に取り上げられた。

○ 調査は「国の教育ローンを利用している家庭」に対するもので、これを「高校以上の学校に就学している子供を持つ家庭」すべてに当てはめるのは誤り（前者は高い教育費を払っているが故に教育ローンを利用している可能性が高い。つまり、選択バイアスがある）。

● 星野崇宏：調査データの統計科学, 岩波書店 (2009). pp.16-18

### 5-2 シンプソンのパラドックス

○ カリフォルニア大学バークレー校大学院の入学許可率のデータ

○ 調査期間において男子の方が女子より入学許可率が 10%近く高かった。

○ しかし、専攻別にみると男女の入学許可率に偏りはなかった。

○ 女子の方が入学許可率の低い専攻を多く志願していたことが原因（専攻の選択が交絡要因となっている）

● D. Freedman, R. Pisani, R. Purves: Statistics (4th ed.), Norton & Company (2007). pp.17-20

# 数理・データサイエンス教育強化拠点コンソーシアムの概要

どの大学・どの学部に進学しても、全ての学生が今後必要となる数理的思考力とデータ分析・活用能力を体系的に身に付けることが出来る環境の構築を目指す

## 拠点校の役割

- 数理・データサイエンスの全学的な教育の実施，カリキュラムの設計・教材作成等
- 多方面にわたる応用展開を念頭に新たな価値の創出ができる人材育成に向けた教育の実施
- 全国的なモデルとなる標準カリキュラムの作成・普及
- 数理・データサイエンスと社会とのつながりについて教えることができる教員の養成（ＦＤ等の充実）
- 地域や分野における拠点として，取組成果の他大学への展開・波及
- 大学，産業界及び研究機関等と連携したネットワークを形成し，実践的な教育の実施

## コンソーシアムの役割

- 全国的なモデルとなる標準カリキュラム・教材を協働して作成．他大学への普及方策の検討・実施．
- センターの情報交換等を行うための対話の場の設定．
- センターの取組の成果指標や評価方法の検討．

## 協力校の選定と全国展開に向けたブロック化

数理・データサイエンス教育強化の全国展開加速のために2019年度より20大学が協力校として選定され，6ブロック化し分担して活動．

① ブロック会議の開催
- 数理・DS教育の実施状況の把握
- 他大学への展開計画・状況の共有
- 今後の取組の方向性の調整

② ブロック別ワークショップの開催
- ブロック内の全大学が対象
- 実施状況の共有
- 数理・DS教育の取り入れ方の議論

③ 拠点・協力校連絡会議の開催
- 全国の拠点・協力校が参加
- 取組の成果や課題の共有
- 自大学及びブロック内での活動の参考に

④ 数理・データサイエンス分野の人材の拡大
- FD活動を通じ，数理・DSを教えられる人材ネットワークを拡大

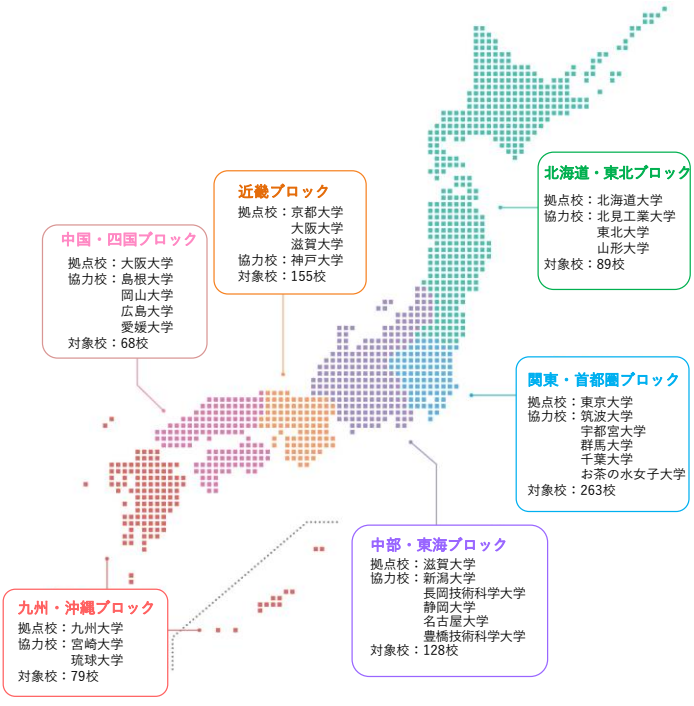
北海道大学 数理・データサイエンス教育研究センター
滋賀大学 データサイエンス教育研究センター
大阪大学 数理・データ科学教育研究センター
東京大学 数理・情報教育研究センター
京都大学 データ科学イノベーション教育研究センター
九州大学 数理・データサイエンス教育研究センター

## 数理・データサイエンス教育強化拠点コンソーシアム

●会合の定期的開催
●3分科会を設置して活動

| | |
|---|---|
| カリキュラム分科会 | 標準カリキュラムの作成 |
| 教材分科会 | 共通教材の作成 |
| 教育用データベース分科会 | 教育用データベースの構築と公開 |

●センターおよびコンソーシアムの成果指標の設定
●各センターのシンポジウム等の主催・後援
●調査活動
●広報活動(ホームページ，ニュースレター)



## 数理・データサイエンス教育強化拠点コンソーシアム　カリキュラム分科会

**委　　員**

| | | |
|---|---|---|
| 主　査 | 丸山　祐造 | 東京大学数理・情報教育研究センター、大学院総合文化研究科教授 |
| 副主査 | 田村　　寛 | 京都大学高等教育院附属データ科学イノベーション教育研究センター特定教授 |
| | 行木　孝夫 | 北海道大学大学院理学研究院教授 |
| | 姫野　哲人 | 滋賀大学 データサイエンス教育研究センター准教授 |
| | 高野　　渉 | 大阪大学数理・データ科学教育研究センター教授 |
| | 大山　　航 | 九州大学大学院システム情報科学研究院准教授<br>（～2019 年 3 月　役職は当時） |
| | 増田　弘毅 | 九州大学大学院数理学研究院教授（2019 年 4 月～） |

**審議経過**　※ 随時メールやワークショップ等で検討を行ったほか以下を開催。

＜分科会会合＞

第 1 回　2018 年 5 月 14 日
- 分科会の目標共有、役割分担等の決定
- スキルセット作成方針についての議論

第 2 回　2019 年 3 月 11 日
- スキルセットについて議論

第 3 回　2019 年 5 月 27 日
- 学修目標、スキルセットの作成方針の確認及び内容の検討

＜意見交換会＞

第 1 回　2019 年 3 月 20 日（水）
参加機関　文部科学省、津田塾大学（総合政策学部）、独立行政法人情報処理推進機構

第 2 回　2019 年 4 月 26 日（金）
参加機関　文部科学省、立教大学（経営学部、総長室教学改革課）、放送大学

第 3 回　2019 年 5 月 24 日（金）
参加機関　文部科学省、経済産業省、武蔵野大学（データサイエンス学部）、データサイエンティスト協会、㈱ブレインパッドアナリティクスサービス本部、独立行政法人情報処理推進機構

第 4 回　2019 年 9 月 12 日（木）
参加機関　文部科学省、立教大学(経営学部)、武蔵野大学(データサイエンス学部)、放送大学