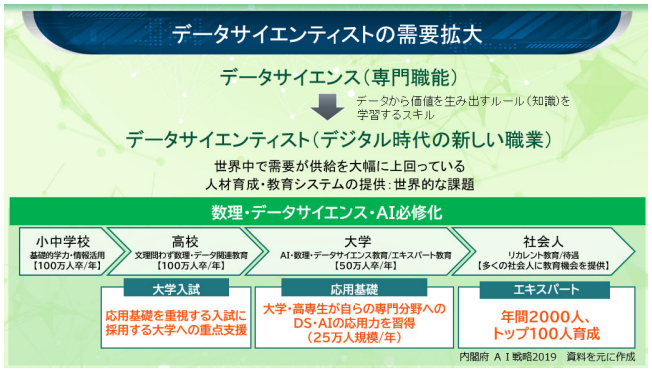
データ対話型アナリティクススキル評価：統計検定 CBT データサイエンス基礎

渡辺 美智子
慶應義塾大学大学院健康マネジメント研究科
watanabe_michiko@nifty.com

1. 100 万人のデータサイエンス

社会の AI・デジタル化が急速に進行し、社会のほぼすべての領域でデータの利活用とその可能性への期待がかつてない高まりを見せている。その中で、デジタル時代の新しい職能として、課題をとらえ、定量的データに基づいて問題を理解・分析・解釈し、組織の意思決定の高度化を先導するデータサイエンティストの需要は大きく、人材育成と教育システムの構築は、どの国においても喫緊の課題となっている。

日本においても、「AI 戦略 2019」(2019 年 6 月閣議決定) において、数理・データサイエンス・AI 必修化として、小中高等学校から大学・大学院、社会人のリカレント教育に至る教育の体系化が示されており、全ての生徒が高等学校卒業時までに、数理・DS・AI の基礎的なリテラシーを習得することを掲げている。かつて日本人が苦手とする英語教育の重要性から、「100 万人の英語」が叫ばれた時代があったが、「100 万人のデータサイエンス」時代の教育と質評価が社会で望まれている。

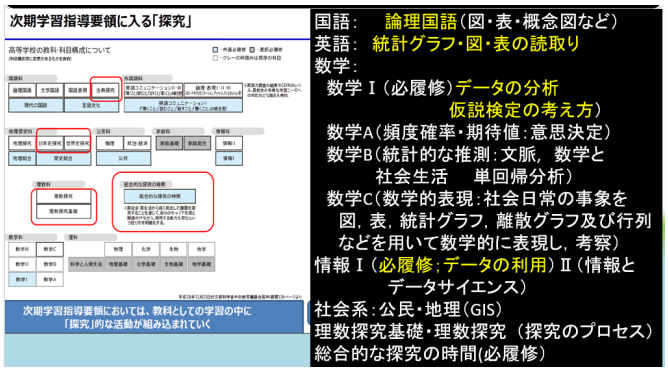


2. 高校でのデータサイエンス探究力強化

新学習指導要領で最も大きく変わるのは 2022 年から年次進行で実施予定の高校教育である。既にその重要性から共通必履修科目として新設された「総合的な探究の時間」が 2019 年から前倒しで先行実施されている。新課程ではさらに、地理探究、数学探究、理科探究、理数探究とほぼすべての教科に、探究が位置づけられ、実社会・実生活から自ら見出した課題の探究活動（アクティブラーニング）を通して、探究能力を育むことが意図されているが、ここでの探究は、調査・実験・観察データに基づく科学的探究を指す。

そのため、科学的探究の方法論である統計・データサイエンスの分析手法の理解と実践スキル習得は、数学Ⅰ「データの分析」、数学 B「統計的推測」と情報Ⅰ「情報通信ネットワークとデータの活用」、情報Ⅱ「データサイエンス」が連携して、主に担う。具体的に取り扱われる内容は、相関、単回帰、重回帰、クロス集計、仮説検定、信頼区間に加え、情報Ⅱ「データサイエンス」単元には、ビッグデータ整形処理をはじめ代表的な多変量解析や機械学習系の手法、テキストマイニングなどが Python や R のコードと共に取り扱われる。

一見すると驚くほどの高度な手法ではあるが、インターフェースに優れたツールの普及で、既にスーパーサイエンスハイスクールなど全国大会では、AI、機械学習、多変量解析を応用した高校生ならではの創意あふれる分析実践例が多数みられる。

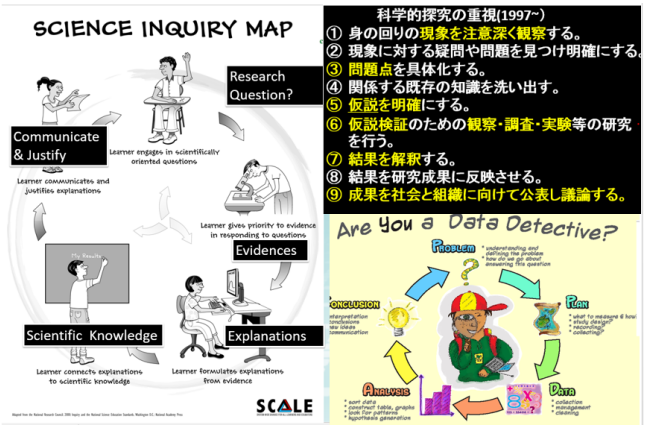


3. データサイエンスの国際的評価のフレーム

(1) AP Statistics

1990 年代初頭より、OECD など国際社会は、予測困難な変化に向かう 21 世紀の人材スキルとして、従来の専門知識・技能(ハードスキル)から問題解決力・科学的思考力・コミュニケーション力・連携力などの業種を超えて転用可能なパフォーマンス(行動特性)基盤型ソフトスキルを 21 世型スキルとしてより重視す

るという方向を示し、初中等から高等教育、社会人教育の枠組みと評価システムの確立を行ってきた。統計教育に関しても問題解決型を指向し、Real Data, Real Problem, Real Learning を軸に、問題解決のサイクル PPDAC: Problem-Plan-Data- Analysis- Conclusion を広く普及させ、1992 年には高校生を対象とした Advanced Placement Statistics の高校 1 年間の教育プログラムと統一試験を始め、現在、毎年 20 万人を超える高校生が現実の文脈の中で統計分析の設定から解釈までを行う問題の試験を受験している。



AP Statistics では、計算の公式は与えられ、単純な計算問題や公式を覚えているかどうかなどの問題ではなく、ブートストラップなどのリサンプリングや仮説検定、ベイズ統計の考え方を理解し応用する問題が統計研究者によって開発・出題されている。

AP 統計が新しい統計能力の評価として登場してから 30 年近くが経過し、実際に 21 世紀を迎え冒頭で述べたデジタル化社会への変革が迫られている現在、AI とアナリティクスの普及は、専門職能とその基礎として、分析結果の解釈方法とともにビッグデータ・スモールデータの収集方法と解析方法、データハンドリング・解析アルゴリズムの機能を理解し問題解決を通して社会の価値に結び付けるデータサイエンスのパフォーマンス評価の段階に至っている。

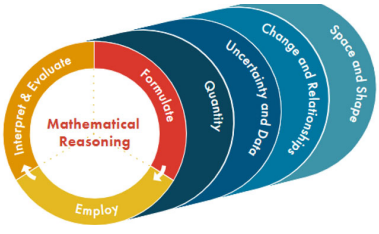
(2) PISA2021 数理リテラシーの新フレーム

OECD が実施する生徒の学習到達度調査 PISA は、義務教育修了時点（15 歳）で身に付けてきた知識や技能を実生活の課題解決に活用できるかを評価する試験として 2000 年から 3 年ごとに、読解力、数学リテラシー、科学リテラシーの主要 3 分野で実施されている。この時点から世界的に、「学力」論は知識・技能を使いこなす「資質・能力」の育成へと変化した。PISA2015 では、先述の 21 世型スキルも、21 世型学力(世界共通の力)として導入され測定が行われた。昨年末に、PISA 2018 の結果が日本でも公表され、「数学的応用力」で 6 位、「科学的応用力」では 5 位と数学、科学はトップレベルを維持したと報道されたが、果たしてこれが次回も達成できるのか。

OECD は既に、劇的にデジタル化した現代社会の変容に合わせ、次回 2021 年の PISA 数学の評価フレームを数学リテラシーの定義、内容、測定方法とも大幅に改訂したことを発表している。主な変更点は、データサイエンスを中心に据えたこと。測定方法も表計算機能を使って段階的にデータ分析的思考で問題を解いていく Computer based assessment（CBA）方式に本格的に変更した。内容領域は「数量」「不確実性とデータ」「変化と関係」「空間と形」で、これまでトップレベルと言われながらも領域的に日本が弱かった「不確実性とデータ」の部分が強調されている。AI 時代の予測モデルの基礎である散布図や相関、関数の適合、集計から得られる条件付確率を基に意思決定するなどの問題がコンピュータベースでインタラクティブに課される。数式を中心とした旧来の紙ベースのセンター試験の形式では測れない、データ対話型のアナリティカルな思考とスキルである。

表計算ソフトは、米国の調査では 6 割強の大人がデータ処理に日常で活用している。子供の未来のために、入試改革も内容だけではなく評価方式もアナリティクスを組み込むなどのデジタル化も急務となっている。日本では、PISA は高校 1 年の春に実施されているが、実は PISA2021 の散布図・相関・クロス集計表から条件付確率を推測して意思決定する等は、次期の指導要領でも高校生の内容である。つまり、義務教育段階の数学(統計)内容が改訂しても既に世界とずれているのである。

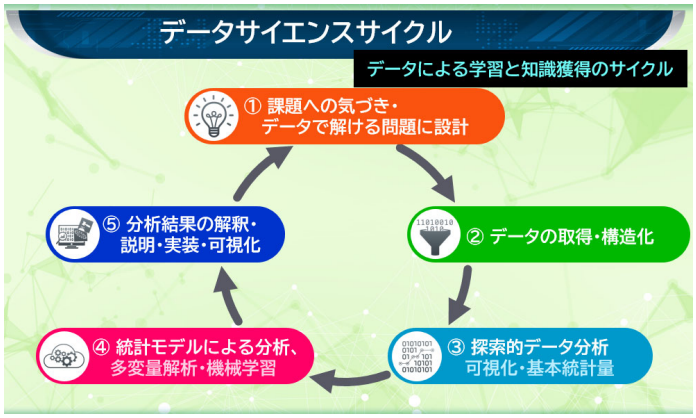
更に、数学(統計)内容は、具体的な文脈（身の回り、職業、社会、科学）において問題解決サイクルのプロセス:「定式化」→「活用」→「解釈と評価」の中に組み込まれる形で問題が開発されており，プロセスごとの評価もできるようになっ



ている。

(3) IDSSP (International DS in School Project)

2018 年に、国際統計協会や米国統計学会、王立統計学会、国際コンピュータ科学学会(ACM)、オーストラリア統計局、Google などの支援で、AP 統計受講後の高校の 2 年間分のデータサイエンス入門コースのカリキュラムチーム IDSSP が国際的に組織されているが、2019 年 9 月に公開されたガイドラインでも、先ず、データに基づく学習(知識獲得)のサイクルを理解し、具体的な課題に応じてこのデータサイエンスサイクルをまわすために、5 つの各ステップでどのようなスキルが必要とされるのかを実践的に学ぶ構成が示されている。



4. 統計検定 CBT-データサイエンス基礎

これまで述べたように、社会のデジタル革新が急速に進行する中で、データサイエンス人材の育成は喫緊の課題であり、その際、日本においても規模の拡大だけではなく、国内外のガイドラインに沿ったデータサイエンス人材の質評価のシステムを確立する必要がある。このため、一般財団法人統計質保証推進協会は、統計検定に新たに CBT『データサイエンス』を創設し、来年度から順次、「基礎」、「発展」、「応用」の 3 水準でその能力評価を実施することとなった。ここでは、「基礎」についての紹介を行う。

データサイエンス基礎試験は、高校における新学習指導要領および国際高校データサイエンスカリキュラム (IDSSP) を参照した内容構成となっている。また、これまでの「統計検定」では、コンピュータ上での実際のデータ処理能力の評価は含まれていなかったが、PISA2021 数学のフレームが重視しているように、具体的な文脈での課題とデータ、所与のデータと情報を中心に対話型で行う「データアナリティクス」の思考力は、AI・データサイエンススキルの主要な能力である。

そこで、データサイエンス基礎試験では、新たに、実際に、具体的なデータセットをコンピュータ上に提示して、分析目的に応じて、解析手法を選択し、データの前処理から解析の実践、出力から必要な情報を適切に読み取り、当初の問題の解決のための解釈や判断を行う、一連の問題解決能力を「データサイエンス基礎」として評価することとした。取り扱う分析手法は、新学習指導要領の数学と情報の両科目の内容レベルとしたが、先述の通り実用的な広い範囲を一通り含んでいる。データの文脈(コンテキスト)は、身近な生活から地域課題に関わる内容とし、さらに、ビジネスな医療・健康・スポーツ・文化等も含めることで、社会人のリカレント教育にも資することも目的としているので、対象は、生徒、大学生、社会人一般である。コンピュータでのデータ処理技術に関しては、Excel によるデータ処理、分析スキルが基本となる。ここでは、Excel 関数、ピボットテーブル、アドインツールである「データ分析」や「ソルバー」の機能も含まれる。これらの対話的な操作を踏まえて、データサイエンス基礎で評価するキーコンピテンシーの構造は、

データサイエンス基礎
＝＞ データアナリティクス基礎
＝＞
分析目的の把握
× データハンドリング × データ解析
× 適切な解釈
＝＞ 意思決定
である。

デジタル社会の課題解決に、上記のスキルを活かす能力をすべての国民が身に着けるべき『情報活用能力』と位置付けている。

下記は、統計検定センターWeb サイト
http://www.toukei-kentei.jp/cbt/cbt_about/grade11/
で公開されている実施趣旨、試験内容、出題範囲表、サンプル問題の情報である。

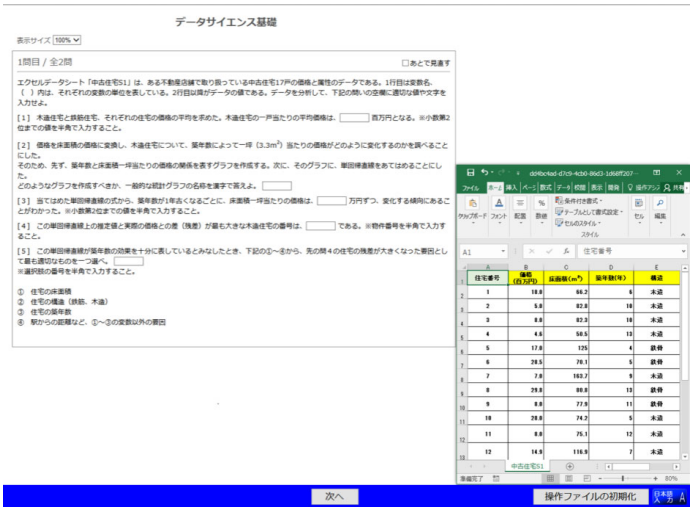
統計検定CBT方式　データサイエンス基礎

実施趣旨

急速に進展したデジタル社会では、規模の大小に係わらず多種多様なデータを処理し、目的に応じた問題解決的思考に基づくデータアナリティクス能力が要求されます。「統計検定」では問題解決に資する統計思考力と活用力を評価する各級として確立してきました。この試験ではCBT方式である機能を活かし、具体的なデータセットをコンピュータ上に提示して、分析目的に応じて、解析手法を選択し、表計算ソフトExcelによるデータの前処理から解析の実践、出力から必要な情報を適切に読み取り、当初の問題の解決のための解釈を行う一連の能力を「データサイエンス基礎」として評価・認証します。既に学校教育では、プログラミングと統計教育の拡充と必修化による生徒のデータ活用能力の育成を掲げた新学習指導要領が告示され、CBT方式による大学入学者選抜試験も実施される可能性が高まっております。

この試験では、新学習指導要領で共通必履修化された数学科と情報科の両科目における「データの分析」・「データの活用」の単元を中心に大学入試までの内容レベルとしますが、同時に社会人が業務の身近な課題をデータ処理するに必須の内容レベルでもあるので、就職や採用時の自身のデータアナリティクス・スキル資格、社内での社員資格等に活用できるものです。

「データサイエンス基礎」試験は、データサイエンスとその応用分野を専門とする大学教員と専門実務家が活用力を重視した問題を開発し、生徒・学生・一般を問わず、AI・デジタル社会の共通スキル「データサイエンス基礎」力を評価し、認証するための検定試験となっています。



試験内容

「データサイエンス基礎」試験で評価するキーコンピテンシーを**データアナリティクス基礎**とし、
(1) データハンドリング技能
(2) データ解析技能
(3) 解析結果の適切な解釈
の3観点を新学習指導要領（平成29・30年改訂）に対応した大学入試までの内容構成で出題します。主に高等学校では、数学I「データの分析」、数学B「統計的な推測」、「数学と社会生活」、数学A「場合の数と確率」、数学C「数学的な表現の工夫」、情報I「情報通信ネットワークとデータの活用」、情報II「情報とデータサイエンス」、理数探究基礎、理数探究等が関係します。

【具体的な内容】
データマネジメント（層別・水準化・変数変換）、データセットマネジメント（欠測値、外れ値処理、データセットの結合や構造化、抽出）、質的データの分析、量的データの分析、記述統計的手法、推測統計的手法、クロス集計分析、相関・回帰分析等

【出題の特徴】
(1) 実際のデータセットを目的に応じてハンドリングし、その結果を問う問題
(2) 分析を実行しその結果を問う問題
(3) 分析結果を読み取り、文脈に応じた適切な解釈を問う問題

2019.4.11 時点

統計検定 CBT「データサイエンス基礎」出題範囲表

大項目	小項目	ねらい	項目（学習しておくべき用語）	主な Excel 操作（項目に渡るものは省略）
社会におけるデータサイエンス	社会におけるデータサイエンス	デジタル社会におけるデータサイエンスやビッグデータの役割、インターネットとデータサイエンスの歴史、個人データに関する情報倫理等を理解する。	インターネットとその歴史、Society5.0、問題解決、エビデンス、オープンデータ、官民データ活用推進基本法、データサイエンティスト、ビッグデータの特徴、機械学習手法の分類(教師あり、教師無し)、個人データ、情報倫理、AI(人工知能)手法の特徴、EBM(証拠に基づく医療)、EBPM(証拠に基づく政策立案)、IoT(モノのインターネット)	
データベース・データマネジメント	データベースマネジメント	分析目的に応じた構造化データの構築やデータ形式の変換、データ抽出等の簡単なデータの整理・整形ができる。	構造化データ(レコード×フィールド、ケース×変数)、欠測値、データの結合、データ形式(ロングフォーマット⇔ワイドフォーマット)、データ抽出(ランダムサンプリング、無作為標本抽出）、乱数	データのソート(並び替え) ピボットテーブル RAND 関数 データの分析 四則演算
	データマネジメント	データの種類や尺度を理解し、層別、水準(レベル)化、変数変換等のデータ処理ができる。	質的データ、量的データ、データの尺度、層別、水準(レベル)化、変数変換、Z 変換(標準化)、偏差値	IF 関数
データの可視化	データの可視化	データを目的に応じて可視化するための統計グラフの作成と解釈ができる。	円グラフ、棒グラフ、折れ線グラフ、帯グラフ、ツリーマップ、パレート図、ヒストグラム、箱ひげ図等	グラフの作成
質的データの分析	1 変量の質的データの分析	質的データを用いて、問題の可視化や現状分析のためのパレート分析(ABC 分析)ができる。	パレート表、パレート図、構成割合(確率)、累積度数(累積相対度数、累積確率)	SUM 関数 ピボットテーブル CHISQ.TEST 関数
	2 変量以上の質的データの分析	2つ以上の質的データを用いて、連関分析や要因探索のためのクロス集計表の分析ができる。	クロス集計表、行(列)比率、セル比率、期待度数とカイ 2 乗統計量、連関係数、特化係数、多重クロス表	
量的データの分析	1 変量の量的データの分析	量的データを用いて、問題の可視化や現状分析のためにデータの分布構造を分析できる。	階級、階級値、標準階級幅、度数分布表、ヒストグラム、基本統計量(平均、標準偏差、分散、四分位数、パーセント点)、箱ひげ図、変動係数、管理図、外れ値	データの分析 AVERAGE 関数 VAR 関数 STDEV 関数
	2 変量以上の質的・量的データの分析	2つ以上の質的データや量的データを用いて、要因探索のための分布の比較や相関分析、単(重)回帰分析による予測モデル構築ができる。	層別ヒストグラム、並列箱ひげ図、相関、相関係数、散布図、単回帰直線、重回帰モデル、寄与率、回帰係数、残差	CORREL 関数
確率による意思決定	確率と確率分布	確率と確率分布による推測の考え方を理解し、シミュレーションを実行できる。	場合の数、確率、ベイズの定理、尤度、事後確率、期待値、2項分布、正規分布、 確率的シミュレーション	BINOM.DIST 関数 NORM.DIST 関数 NORM.S.DIST 関数
	推定	標本変動と誤差を理解し、母集団特性値の推定ができる。	信頼区間、信頼率、標本誤差、標準誤差 母平均、母比率	NORM.INV 関数 Z.TEST 関数
	検定	仮説検定の考え方を理解し、文脈に応じた検定を行い結果の適切な解釈ができる。	帰無仮説、対立仮説、有意水準(危険率)、有意確率(p 値)、第一種の過誤、第二種の過誤、帰無仮説の棄却、2 項検定、Z 検定、t検定、$\chi 2$ 検定、ABC 分析テスト	T.TEST 関数 CHI.DIST 関数 CHI.INV 関数 データの分析
時系列データの分析	時系列データの分析	時系列データの構造を理解し、特徴を分析できる。	指数、移動平均、伸び率、成長率	AVERAGE 関数
テキストマイニング	テキストマイニング	テキストマイニングの意味を知り、単語や品詞の出現頻度を分析できる。		

まとめ

AI・デジタル社会を迎え、多くの全国的な試験が将来的には CBT 化されることが予想される。また、そうならなければ、データ中心のコンピュータ対話型試験は実行できず、モデリングやアナリティクスの真に測らなければいけない能力が測定できない。

海外では、AI やデータ解析、Python などの単純な知識を問う、4 択や 5 択のクイズ形式のオンライン試験や資格は、有償・無償を含め多くインターネット上に存在し、簡単に試験が開始できその場で合格証が手に入る。これらは入り口に過ぎない。

OECD の PISA や AP、IDSSP、日本の大学入試改革などは、問題開発に困難は伴っても、学協会の経験と知識、ICT の技術を使って、今後求められる高度に知識処理ができる人材の質保証にチャレンジしている。AI・データサイエンスは、社会実装化が大学等高等教育の現行課程を追い越し、はるかに速いスピードで進んでいる。評価問題の開発や提供においても、学協会と産官の連携が望まれる。