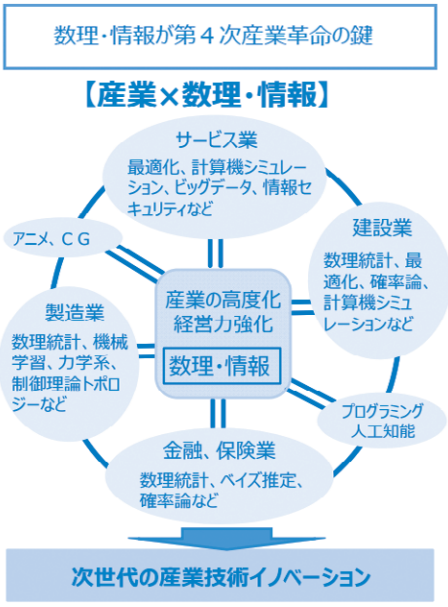
第1部 統計的探究のプロセス

1　いま、求められる“統計・データサイエンス力”とは？

（1）　第４次産業革命の鍵となる統計

21世紀に入り、人工知能を搭載したコンピュータソフトが囲碁や将棋の名人に勝利したり、ロボットが会話したり、自動車が自動運転したり、モノとモノとがインターネットを介して直接データをやりとりするなど、これまでSF の世界とされていたことが、私たちの身の回りで急速に現実化している。この背景には、状況を示す膨大なデータからコンピュータが次々とルール（法則）を学習して、最適な予測や判断を行うことを可能にした統計的なデータ分析技術の進歩がある。

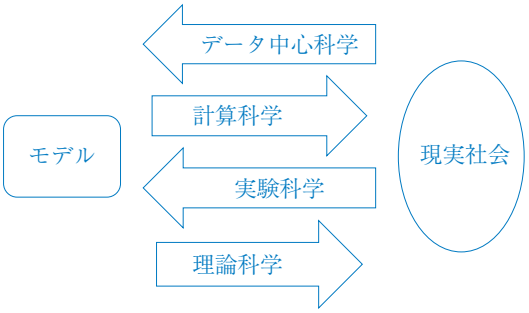
この技術は、既に迷惑メールの検知、クレジットカードの不正使用の検知、数字や顔画像の認識、商品購入のレコメンデーション、医療診断、信用リスクの予測、自然言語処理など、社会で広く応用されているため、現在はデータを中心とする科学技術で第４次産業革命が到来したとまで言われている。



資料：文部科学省「第４次産業革命に向けた人材育成総合イニシアチブ」関連資料（2016年４月）

これまでの産業革命を振り返ってみよう！
　まずは、18世紀後半の工業化の黎明期を語る第１次産業革命。これは、蒸気機関による自動化の時代。
　次に、19世紀後半の大量生産と文明化を語る第２次産業革命。これは、電気による自動化の時代。
　続いて、20世紀後半の電子化による製品・生産設備システムの進化を語る第３次産業革命。
　これは、コンピュータによる自動化の時代。
　そして現代は、第４次産業革命。データ駆動型サービスによる自動化の時代に突入したんだ。

（２）　第４の科学のパラダイム：データ中心主義



データを中心とした変革が進んでいるのは、産業界だけのことではない。大学や大学院に入ってからの研究の方法についても、これまでは、知識の蓄積を背景とした「理論科学」や「実験科学」の方法、コンピュータの発達による「計算科学」の方法が中心であったが、新しく第４の科学のパラダイム（The Forth Paradigm）として、膨大なデータから直接、社会・自然・経済・人間行動等のルール（法則）を発見する「データ中心科学」が急速に広まってきている。

この「データ中心科学」によって、医学、健康科学、生物学、物理学、地学、経営学、経済学、社会学、教育学、スポーツ科学などの多くの領域で、データを活用した創造的な研究成果が生まれてきている（Tony Hey、2009）。

調べてみよう！

データが研究にどのように役立てられているか具体的に調べてみよう

（例）　遺伝子データの医学での活用

遺伝子データが医学研究に活用されるようになって、いろいろな病気の発生リスクがどのようなタイプの遺伝子型と関連しているのかについて研究が進められ、病気の予測や治療方法の選択に役立てられている

大学に「データサイエンス学部」が創設

文部科学省は、平成28年に「大学の数理・データサイエンス教育強化方策について」を公表し、国立大学法人の拠点大学として下記の６大学を選定しています。この背景には、データが豊富に入手できる時代となっているなかで、データとアナリティクスを用いた意思決定を行う企業の割合が世界平均61％であるのに対し、日本は40％と低い状況であること、今後、世界ではますますデータを利活用した新産業創出や企業の経営力・競争力強化がなされるという予想があります。 このため、数理的思考力とデータ分析・活用能力を持つ人材の育成と社会に価値やサービスを生み出すという目的に合致した大学教育システムの構築を目指しています。 拠点大学は、最初は６大学ですが、今後、日本の多くの大学で、統計・データサイエンスの教育拡充が進められる方向です。

数理およびデータサイエンスに係る教育強化拠点大学選定校一覧（国立大学法人が対象）

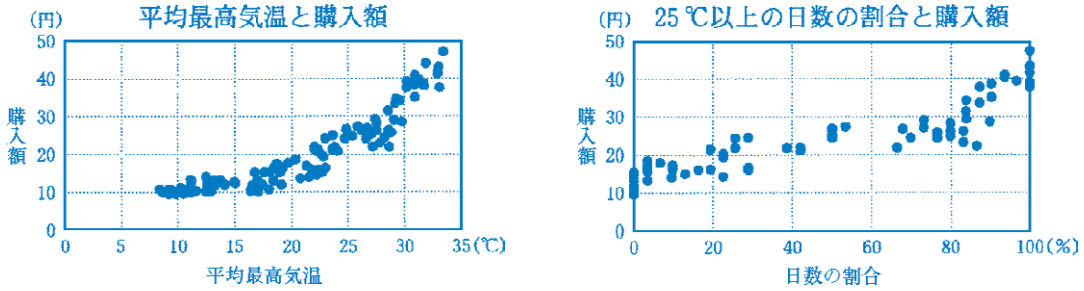
NO	大学名	事業名
1	北海道大学	数理的データ活用能力育成特別教育プログラム ～データサイエンスセンター（仮称）の設置～
2	東京大学	数理・情報教育研究センターの設立
3	滋賀大学	データサイエンス教育の全学・全国への展開 －データリテラシーを備えた人材の育成に向けたカリキュラム・教材の開発－
4	京都大学	データ科学イノベーション教育研究センター構想 －21世紀のイノベーションを支える人材育成－
5	大阪大学	数理・データ科学の教育拠点形成
6	九州大学	九州大学「数理・データサイエンス教育研究センター（仮称）」構想

（３） 統計的探究とデータサイエンスの考え方（センター試験の問題を例に）

データが社会の中心となるなかで、いま、私たちには、身近なデータを活用して新しい知識を創造する探究力が求められている。そのためには、探究のための科学的な思考の方法と統計的にデータを分析する方法を理解し、身に付けなければならない。

社会課題を科学的な思考で取り組むとは、身の回りの課題や地域・社会の課題をいろいろな側面から検討し、広い視野（マクロな視点）で捉えた上で、課題を複数の具体的な現象の関わり合いとして絞り込み、各現象にデータを対応させ、現象間の関連性のルール（法則）を統計的な分析の方法で検証していくという方法をいう。

2016年度の大学入試センター試験の数学Ⅰの出題問題から見てみよう。アイスクリームの購入金額という現象と気温や湿度といった現象の関係を探究して、何がアイスクリームの消費を促すかの法則を見出す設定になっている。そのため、2003年から2012年までの10年間の東京都の月別データとして、１世帯当たりのアイスクリーム消費額（家計調査）と気象庁が公開する気温等のデータを集め、次のような散布図を作成している。



散布図上で、世帯のアイスクリームの消費額という日別や月別で値が変動する変数、気温というやはり値が一定しない変数との関連を具体的なデータで示すことが、特定の品目（アイスクリーム）の消費を気温で評価し、予測するためのエビデンスを得ることに繋がる。また、消費額と気温の関係に、曲線や直線のモデル式を当てはめれば、更に具体的に、消費に与える気温の効果を数量で見積もることもできる。これは、アイスクリームの製造会社が販売量や利益を考える上で、非常に有用な情報（新しい知識）にもなる。

考えてみよう！
① アイスクリームの消費に関係する気温以外の現象は？
② 気温と関係する他の消費品目は？
③ 予測に役立ちそうな２つの関連する現象の他の例は？
④ これが予測できると、こんな課題解決に役立つのでは、と思われる例は？

このように、企業などで何かの判断や決定をするときには、客観的で信頼のおける情報が必要になる。直感や口コミの情報ではなく、できるだけ信頼性のある公的な統計データを使うこと、また、データの数についてもより多くのデータを使うことで、対象とした現象に関して何が起きているのか、その傾向を俯瞰することができる。

現象を表すデータをいろいろな方向から考え、その変動を説明する要因を探し出すことで、予測ができたり、問題解決に繋がる効果的な介入策を考えたりすることができる。これが提案や判断の根拠に繋がる。統計・データサイエンス力とは、統計グラフの種類や個別の分析手法の知識を覚えるだけではなく、現実の社会の課題を捉え、問題を明確にし、その問題を解決する方策を検討し、適切な意思決定を行う力のことである。

次の節で、そのための方法として、PPDAC メソッドを学習しよう。

2　課題発見と問題解決のフレーム：PPDAC メソッドの活用

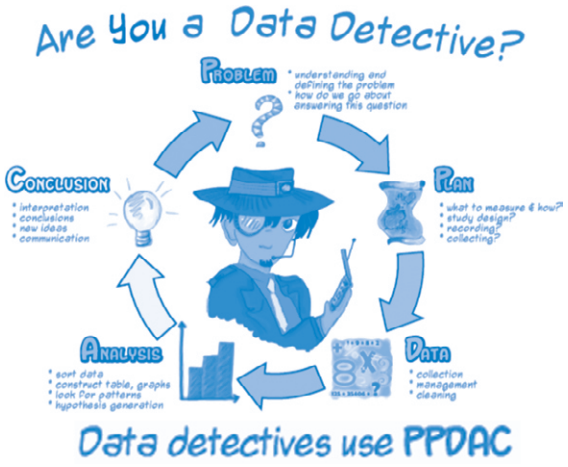
PPDAC メソッドとは、カナダ・米国・ニュージーランド等の学校教育で使用されている科学的探究の手順を示したもので、漠然とした課題をデータで解決可能な問題に落とし込んだ上で統計分析し、元の課題の内容に照らして、状況を判断したり、解決策を提案したりする次の一連の探究活動のフレームをいう。

P	Problem 問題の設定 とらえる	①関心のあるテーマを決め、そこでの**課題**を考える ≪概念図や俯瞰図を作成する≫ ②課題から**問題の構造（原因系と結果系の現象）**を明確にする ≪ロジックツリーや特性要因図・要因連関図などを作成する≫ ③具体的な**研究仮説（リサーチクエッション）**を設定する
P	Plan 計画 みとおす	・問題の重要度を測る指標、その変動に影響を与える要因系の指標など、計測すべき変数（データ）データや統計資料を決め、その収集計画を立てる ・研究仮説を明らかにするための分析の計画を立てる ・分析結果の見通しを立てる
D	Data データ あつめる	・情報（データや統計資料等）を実際に取得し、整理する。データの取得方法（実験か質問紙調査か観察・記録なのかの区別）を意識する。
A	Analysis 分析 まとめる よみとる	表やグラフを作成したり、代表値を計算したりして、データや統計資料を分析する。下記は主な分析の視点である。 ・**全体の傾向（分布）**を見る ・条件の違いやグループに分けて、**比較**する ・指標間の**関連性**を見る ・指標間の**因果関係**を見る ・**時間経過による変化**を見る ・対象を**分類**する
C	Conclusion 結論 いかす	最初に立てた研究仮説に対して判断や結論を示す。同時に、元の課題の内容に戻り、分析に基づいた考察や提言をし、新たな研究課題の提起から次の探究サイクル PPDAC へと繋いでいく。

考えてみよう！

右の図は、ニュージーランドの学校で使用されている PPDAC サイクルのポスターある。それぞれのステップで、どのような内容が書かれているか、英語を訳して考えてみよう。

（例）　Problem
・Define the problem
・Investigative Question



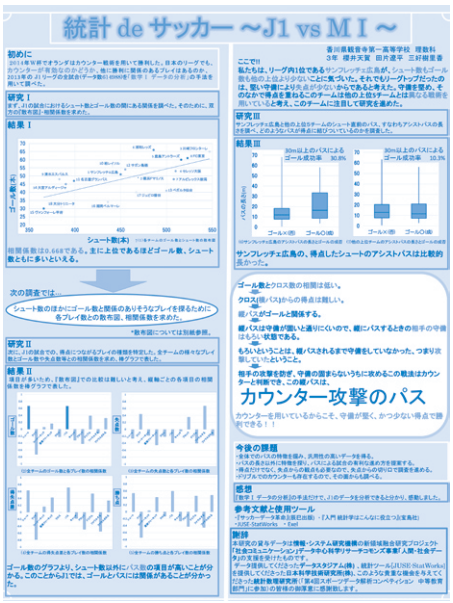
資料：CensusAtSchool NZ

3　高校生の事例で学ぶ PPDAC メソッドの活用

PPDAC メソッドで具体的にどんな分析ができるのか、高校生たちが行った実際の課題研究を例に見てみよう。

◎ Ｊリーグ　チームの強さとプレイの相関分析

【香川県立観音寺第一高等学校】



右のポスターは、2015年度スーパーサイエンスハイスクール生徒研究発表会において、生徒投票賞を受賞した香川県の高校生３名のグループによる研究作品である。2013年のＪ１の全試合データを分析して、リーグ１位のサンフレッチェ広島の強さの要因を解き明かしている。

ここで使用している分析手法は、高校１年生で全員が学習する数学Ⅰの「データの分析」に出てくる、散布図、相関係数、箱ひげ図である。そのため、タイトルに数学Ⅰを表すＭⅠが入っている。最後の「感想」に、「数学Ⅰ　データの分析」の手法だけで、Ｊ１のデータを分析できると分かり、感動しました。」とあるように、基本的なグラフと分析手法だけでも、目的に沿ってそれらを組み合わせることで、オリジナルな研究成果をあげ、新しい知見を見出すことができるという優れた研究事例となっている。では、どういう統計的探究のプロセスをたどったのか、PPDAC のプロセスに沿って考えてみよう。

Problem

◇ テーマと対象、課題の設定

「2014年Ｗ杯でオランダはカウンター戦術を用いて勝利した。日本のリーグでも、カウンターが有効なのかどうか、他に勝利に関係のあるプレイはあるのか、2013年のＪ１リーグの全試合‥を用いて調べた。」とあるように、サッカーＪ１リーグのチームの試合を対象として、何がチームの勝敗を分けるのかがテーマである。

統計的探究における「対象」とは、具体的な観察対象のことで、ある前提条件の下で、複数の対象が観察（測定）可能なものを指す。

「課題」とは何か？

課題とは、「対象」に対しての理想の状態を想定し、現実とのギャップを意識することから見いだされる。サッカーのチームを対象にした今回の場合、理想を“勝利”や“優勝”で捉え、その上で、現実を対比させると、負けもあれば、ランキングで下位になるチームもある。優勝や勝利とのギャップが解くべき課題

＊何がチームを優勝に導くのか

＊何が試合を勝利に導くのか

に繋がる。

■ 課題から問題の構造を見出し、データや統計で解ける研究仮説に落とし込む

ア）評価指標の決定

目的とした「勝利するチーム」という定性的な性質（言葉や感覚で決めた概念）を定量的に測る指標として、まずは決める必要がある。指標が決まらなければ、具体的なデータがとれない。

ここでは、チームの強さを“試合でのゴール数”で計測することとしている。他にも、試合での“得失点差”で測るなどいろいろ考えられる。このように、目的となる指標で、かつ、具体的にデータとして入手できる指標をター

ゲット指標または、最重要評価指標（Key Performance Indicator：KPI）、最重要目的指標（Key Goal Indicator：KGI）といっている。

イ）問題解決は原因分析

データや統計で現実の問題を解くとは、このターゲット指標の値の変動の要因を明らかにし、その値を理想の方向に変える条件や方策を考察することである。

ウ）原因と結果の法則から研究仮説（Research Question）を立てる

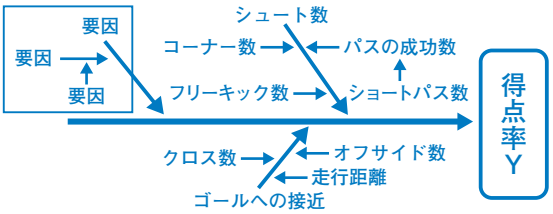
どの要因が最も効果的にターゲットとする指標を変化させるのか？　やみくもにいろいろなデータや資料を集めるのではなく、できるだけ原因と結果の関係に見通し（仮説）を立てた上で、データや統計資料の収集をする必要がある。

仮説を立てる上で、対象に関連するいろいろな現象間の関連性を俯瞰する論理図（**特性要因図**、連関図、ロジックツリーなど）を予め作成することが、統計的探究活動では重要な作業となる。また、このような俯瞰図は、分析の結果を発表する際にも、自分がどのように分析の背景全体を捉えたかを示すためにも、欠かせない重要な資料となる。

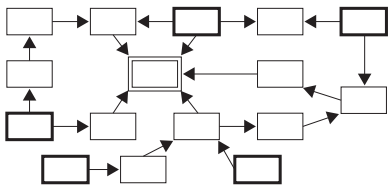
考えてみよう！

ブレーンストーミングのツールを使って、ターゲット指標と要因指標の関係を構造化してみよう

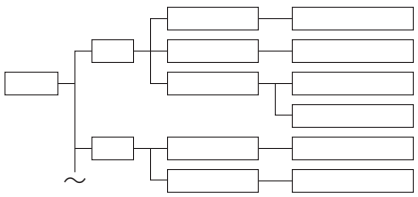
（例）　特性要因図を使うと…



＊連関図を使うと…



＊ロジックツリーを使うと…



ここでは、「試合中のさまざまなプレイがゴール数に関係する」という研究仮説を立てている。

Plan

◇ 仮説の検証に必要なデータや分析の方法を考える

Ｊ１リーグ全チームの2013年の試合における各プレー数、ゴール数等のデータを収集して、散布図や相関係数で関連性を分析する。

Data

◇ データや統計資料を集める・データシートの作成

ターゲット指標に加え、要因系の指標も含めて各チームのデータを１つの行でまとめた、リスト形式と言われる次のようなデータシートを作成する。

順位	チーム名	シュート数	クリア	コーナーキック	直接フリーキック	クロス	パス	キャッチ	ブロック	ドリブル	ファウルする数	ファウルされる数	クリアされる	スルーパス	ゴール数	失点数	得失点差	勝ち点
1	サンフィレッチェ広島	450	748	180	946	648	8397	250	555	557	330	442	746	443	51	29	22	63
2	横浜Fマリノス	469	753	179	543	488	7282	260	618	495	397	534	913	465	49	31	18	62
3	川崎フロンターレ	547	824	177	437	390	8025	311	607	565	419	430	767	600	65	51	14	60
4	セレッソ大阪	510	795	155	371	489	7427	445	554	539	465	360	784	427	53	32	21	59
5	鹿島アントラーズ	516	781	138	450	466	7182	325	550	514	467	437	735	468	60	52	8	59
6	浦和レッズ	486	637	169	448	458	8628	259	489	541	454	438	697	462	66	56	10	58
7	アルビレックス新潟	508	828	220	439	637	6557	322	695	429	448	427	776	579	48	42	6	55
8	FC東京	523	734	148	448	547	7674	313	531	382	396	448	880	577	61	47	14	54
9	清水エスパルス	383	801	152	461	529	6159	315	602	451	435	448	756	406	48	57	-9	50
10	柏レイソル	441	831	176	457	634	7360	240	645	347	494	448	947	533	56	59	-3	48
11	名古屋グランパス	399	839	126	395	565	7860	273	621	370	469	388	798	426	47	48	-1	47
12	サガン鳥栖	455	869	140	448	443	6178	317	699	334	482	440	906	296	54	63	-9	46
13	ベガルタ仙台	500	768	172	421	637	6767	333	608	393	365	413	887	409	41	38	3	45
14	大宮アルディージャ	390	830	157	358	534	6871	291	640	390	430	350	781	529	45	48	-3	45
15	ヴァンフォーレ甲府	366	918	133	429	446	6477	352	542	411	456	427	772	401	30	41	-11	37
16	湘南ベルマーレ	436	1003	144	419	491	6070	331	659	396	455	404	756	457	34	62	-28	25
17	ジュビロ磐田	467	826	222	348	666	7427	314	656	482	477	338	973	469	40	56	-16	23
18	大分トリニータ	385	956	157	406	510	5961	361	638	417	483	405	767	322	31	67	-36	14

Analysis

◇ データを研究仮説に沿って分析する

観音寺第一高校の分析では、一つひとつの分析結果から仮説を次々と進化させ、分析を深めている。

最初の仮説

ゴールに最も結びつくプレーはシュートである。仮説として「シュート数がゴール数に影響する」とする。

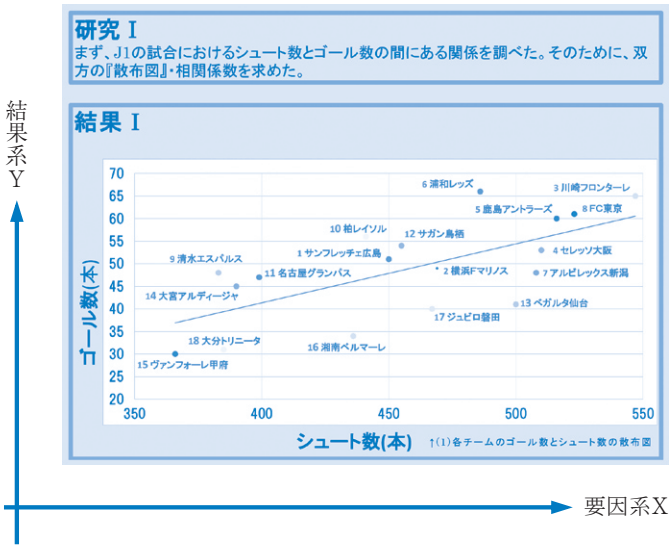
分析の方法

散布図と相関係数

分析結果

シュート数を横軸、ゴール数を縦軸とした散布図では、シュート数が多いチームほどゴール数も多く、逆に、シュート数が少ないチームはゴール数も少ないという正の相関関係（相関係数r＝0.68）が見られた。

その中で、１位のサンフレッチェ広島は、他の上位集団に比べて、シュート数が少ない箇所に位置している。また、ゴール数も特に多いというわけではない。この理由はどこにあるのか？そこで、次の分析の仮説を考えた。



↑(1)各チームのゴール数とシュート数の散布図

次の仮説

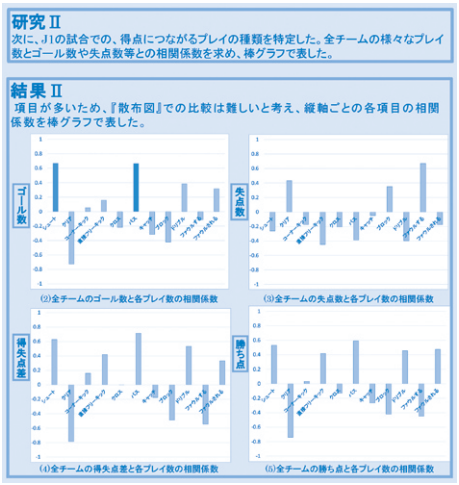
ゴール数以外の勝利に結びつくターゲット指標が存在する。また、シュート以外のプレイも、それらと関係する。

分析の方法

＊勝利に結びつくゴール数以外の指標の洗い出し
＊それらに繋がるプレイの洗い出し
＊各プレイ数とターゲット指標との相関係数の算出
＊プレイごとの相関係数の比較を棒グラフで表現

分析の結果

ゴール数、得失点差、勝点などの指標に対して、シュート数以外に、パス数の相関係数が高い。



覚えておこう！　散布図と相関係数

a　ヒストグラム、箱ひげ図と並ぶ統計３大グラフの１つ
b　２つの数量変数Ｘ と Ｙ の関係のパターンを分析するグラフ
　＊直線傾向の場合（相関関係）
　＊曲線的な傾向もある
c　散布図上で、各対象のポジショニング（位置）が分析できる
　＊傾向に沿う対象
　＊傾向から外れる対象
d　相関傾向の強弱は相関係数ｒで計量化できる
　＊ｒは、－１から＋１の間の値
　＊０に近いほど、相関関係は弱くなる
　＊負の値‥負の相関
　　正の値‥正の相関
　＊絶対値 |r| が１に近いほど、相関関係（直線傾向）が強い、Ｙ の予測モデル（直線）の誤差が小さくなる、Ｙ の変動を Ｘ の変動で説明する説明力が高くなる

Conclusion

◇ 結論とそこから生じる新たな課題を考えるステップ

サッカーの試合に対して、勝利に貢献するプレイとしてシュートがゴール数と関係があることを示しただけではなく、他のプレイも関係すること、そのなかでもパスが重要であることを相関係数によって示している。また、シュート数、ゴール数で特にに大きな特徴がないサンフレッチェ広島の優勝要因が何であるのか、新たな探究課題を見出している。

覚えておこう！　分析を成功させるイロハ

イ）局所管理する

対象の種類の違いが分析結果に影響を与えることを避けるため、対象としているデータはできるだけ同質なものの集合とする

ロ）比較対照をおく

一般的な傾向か固有の傾向かの区別をするため、対照集団（ベンチマーク）をおいて比較する

ハ）繰り返し測定（データの数）する

分析結果の差が単純な標本変動でないことを示すためには、データの数がある程度大きいことが必要

Next Problem

リーグ優勝したサンフレッチェ広島と他の上位チームが、どこで明暗を分けたのか、そこに、サンフレッチェ広島独自の戦術があるのではないか、というテーマを設定し、パスを観測対象とした統計的探究を行う。

Plan

ここで、パスの中でも具体的に、ゴールに結びつくシュート直前のアシストパスに限定して（局所管理）、そのパスによって「シュートが成功したか失敗したか」をターゲット指標に、パスの長さをその関連要因として比較分析することを計画する。

Data

J1の試合で上位チームのアシストパス1本を1行に、下のようなデータシートを作成する。

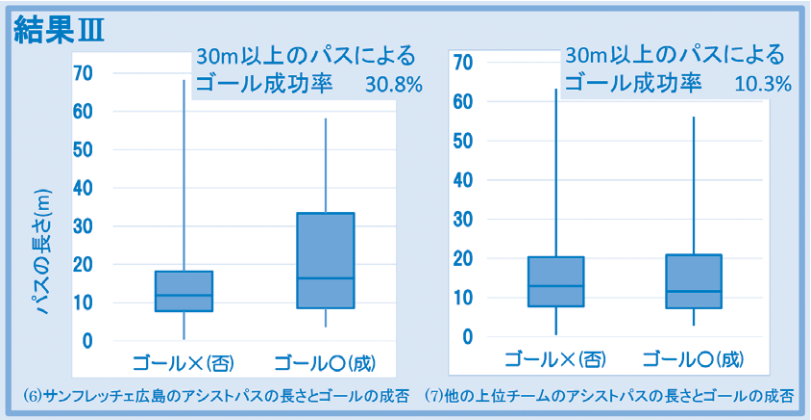
アシストパスID	チーム名	ゴール	パスの距離
1	サンフレッチェ広島	○	38.57
2	サンフレッチェ広島	×	18.34
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・
109	川崎フロンターレ	×	21.08
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・

Analysis

ゴールが成功した場合としなかった場合のそれぞれの「アシストパスの長さ」（量的変数）の分布の比較を五数要約と箱ひげ図で示している。また、その傾向をサンフレッチェ広島とその他の上位チームと比較している。

サンフレッチェ広島		
	ゴール×	ゴール○
最大値	68.27	58.16
第三四分位数	18.14	33.34
中央値	11.89	16.37
第一四分位数	7.79	8.58
最小値	0.37	3.58
アシストパス数	96	12

その他の上位5チーム		
	ゴール×	ゴール○
最大値	63.29	56.14
第三四分位数	20.36	20.86
中央値	13.00	11.60
第一四分位数	7.78	7.32
最小値	0.53	2.87
アシストパス数	595	45

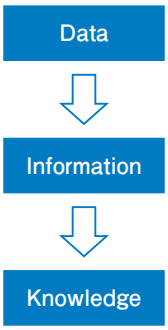


Conclusion

これまでの分析により、サンフレッチェ広島は、ゴール成功時のアシストパスの長さが失敗時に比べて長い傾向にあり、このパスの長さとゴールの成否の関係の傾向は、他の上位チームには、見られないことが分かる。この分析結果から、優勝の要因は、ワールドカップでのオランダチームの優勝要因といわれる、「カウンター攻撃にある」と段階を追った推察で結論付けている。

4　周囲を説得！　できる分析レポートの構成

（1）「データ」→「情報」→「知識」創造を支える分析力



散布図や箱ひげ図といった基本的統計グラフの作成、相関係数や四分位数などの統計量の計算だけでも、「データ」から「情報」を得ることができ、それが背景の文脈の枠組みのなかで考察されることで、その領域固有の「知識」となる。客観的なデータに基づいて得られた「情報」や「知識」は、個人や組織において、提案や提言に説得性をもたせる基礎資料になったり、判断や意思決定の信頼できる基準となる。ここで取り上げた事例は、スポーツデータの分析だったが、PPDAC メソッドの手順は、スポーツ以外のいろいろな課題の発見と問題解決に応用できる。

（2）　説得力をあげる分析レポート：競争優位なレポートとは？

ここでは、どのような視点で分析を加えていくと、レポートの優位性が出てくるのかについて一般的な段階を追って見てみよう：

レベル０：何が起きたのか？　起きたことだけを報告した基礎レポート
「試合に負けた。テストで80点をとった。海岸には空き缶ゴミが捨てられている。給食の残飯が多い。…」など、起きたことだけを報告するレポートは、レベル０。
「サイコロを振ったら２が出た」としているだけで、「サイコロを振ったときに出る目」という現象全体を確率的現象として理解していないことに相当する。

レベル１：どこで、いつ、どうしたら、など、５Ｗ１Ｈに回数や確度を報告した調査レポート
全体の起きうる事象を洗い出し、その起きる回数や確度（相対度数・統計的確率）を調査した上で、現在、起きていることが滅多に起きないことか、よく起きることかを考察したレポート。「サイコロの目は１から６までの数字があり、それぞれ１/６の確率で起きる。その中で２が出た。」というように、試合の結果やテストの得点など、身の回りの現象に分布を対応させて考えることが大切である。

レベル２：そのようなことが起きた問題はどこにあるのか？　考察を加えたレポート

レベル３：取り急ぎ解決に必要なアクションを示し、対策まで加えたレポート

レベル４：統計的に関連性を分析し、なぜそれが起きたかの仮説を加えた基礎統計分析レポート

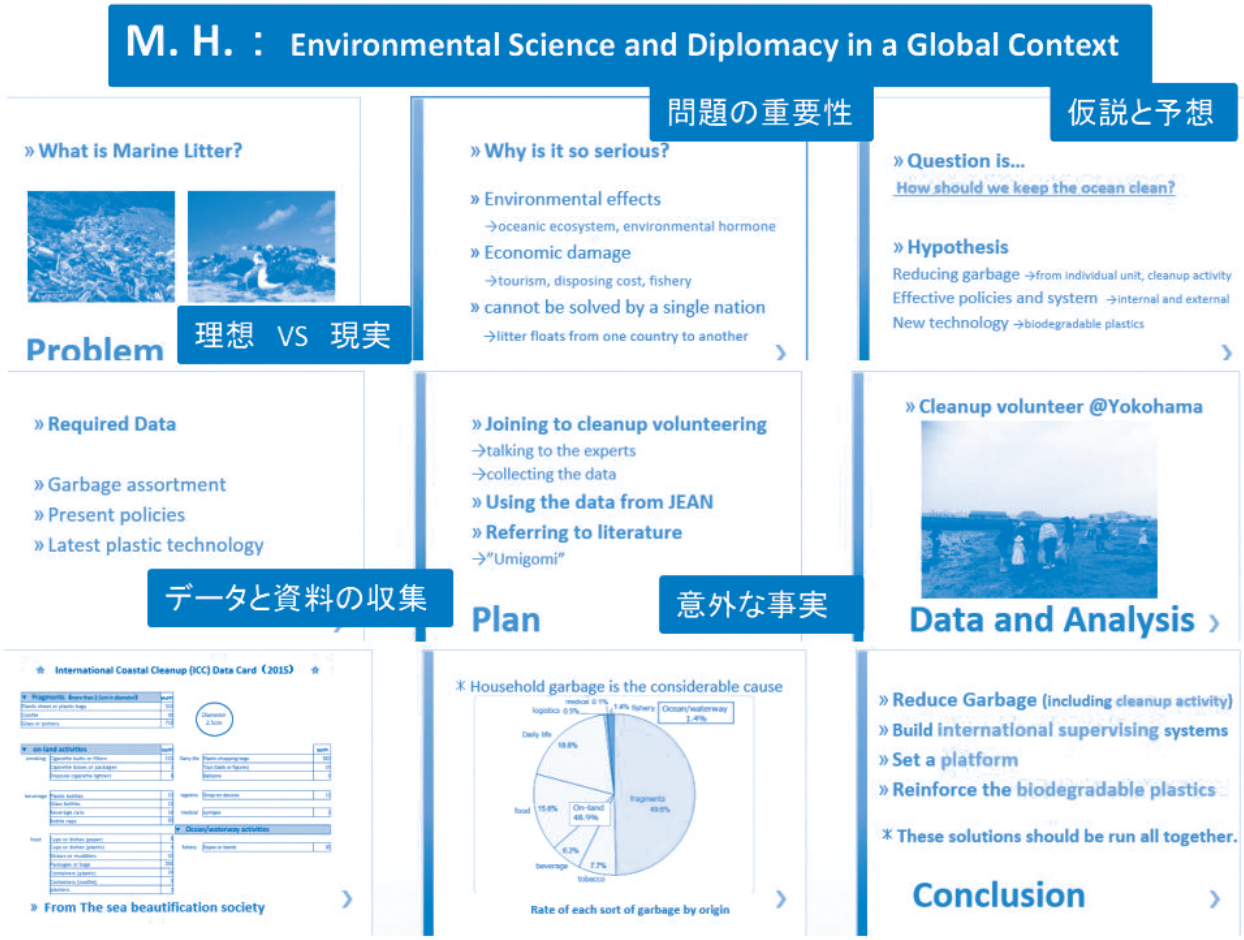
レベル５：この傾向が続けばどうなるのか、予測を示した統計分析レポート

レベル６：複数の要因に対して、それぞれの要因を動かしたらどうなるのかの分析を加たレポート（予測モデルなど高度な統計分析）

レベル７：要因分析を踏まえた上で、取るべき最適な戦略を示した高度な統計分析レポート

下記は、高校生が、PPDAC メソッドに沿って、海岸線のごみの問題を探究した研究成果である。

理想と現実の対比から問題を捉え、問題の重要性を示し、問題解決に有用な方策を想定し、実際のデータを収集・分析によって得られた事実に基づき、最終的な提言を行っている。



引用：スーパーグローバルハイスクール
平成27年度第２回 SGH 連絡会分科会資料
「課題解決に向けた PPDAC メソッド活用」
資料１より抜粋
http://www.sghc.jp/wp/wp-content/uploads/2015/12/2.pdf

本書は、このような PPDAC メソッドを軸にして、具体的な課題発見と問題解決のスキルが身につくように構成されているので、実際に、探究する気持ちで学習してみよう。