研究アプローチの違いによるデータ分析における従属変数の選択への影響の検討

Differences in selecting dependent variables between hypothesis-driven and data-driven research approaches

松室 美紀 [1*]　三輪 和久 [1]
Miki Matsumuro[1]　Kazuhisa Miwa[1]

[1] 名古屋大学大学院情報学研究科
[1] Graduate School of Informatics, Nagoya University

Abstract: We investigated the effects of hypothesis- and data-driven approaches to selecting dependent variables in data analysis. Many studies have shown that the hypothesis-driven approach is used fundamentally for scientific discovery, most of which focuses on selecting independent variables, or planning experiments. The focus of the instant study is on the data analysis process. The results of our experiments showed that the participants using the data-driven approach analyzed a broad range of variables. The participants who used the hypothesis-driven approach did not analyze a variety of variables as long as their hypotheses were supported. When they were not supported, the participants increased the types of variables used for the analyses even if they had their hypothesis. Additionally, they created new variables to support their hypotheses.

1　はじめに

研究者は，その研究プロセスにおいて，実験を行い，様々なデータを取得し，その分析の結果に基づき，結論を導出する．実験の計画においては，操作する要因を選択し，その値を決定するという独立操作の選択が重要である．実験において，様々なデータの値を取得した後は，研究者は分析対象とする変数を選択し，その分析結果から結論を導出する必要がある．本研究では，このようなデータ分析における従属変数の選択に焦点を当てる．

1.1　仮説と独立変数の選択

多くの科学的研究プロセスに関する先行研究では，独立変数の操作，つまり，実験においてどのような要因をどのように操作するかに焦点が当てられている．それらの研究，特に，探求型学習 (inquiry learning) に関する研究では，特定の仮説に基づき，実験を計画する仮説駆動アプローチが，科学的発見や学習に重要であるとされている [1, 2, 3, 4].

仮説駆動アプローチを用いる場合，研究者や学生は，はじめに先行研究や研究目的に基づき，検討すべき仮説を形成する．そして，実験では，仮説を検証するために必要な変数が操作される．最後に，実験の結果から，仮説が支持されるかどうかが判断される．

上記のプロセスから示される通り，仮説駆動アプローチの利点は，仮説と対応する変数を体系立てて操作し，実験を行えることにあると言える．そのため，探求型学習のような，繰り返し実験を実行できるような状況では，その利点が最大限に生かされ，効率的な学習へとつながったと考えられる [1, 2, 4, 5].

しかし，そのように繰り返して実験を行うことは，とりわけ社会学系の研究においては，実験実施にかかるコストのため非常に困難である．それらの研究では，少数，時には 1 つの実験から結論が導出されている．研究者は，少数の実験で様々な変数の値を測定し，それらに基づき主な結果をサポートし，考察している．

このように多くの変数の値が測定されている場合，どの変数をどのように分析するかが重要となる．結論は分析結果に基づき導出されるため，分析される変数により結論の内容や説得力が変化する [6]．仮説駆動アプローチの利点は，主に，実験の計画にあることを考えると，データ分析においては異なる効果が観察される可能性がある．

そこで，本研究では，近年注目を集めているデータ駆動アプローチを用いた場合とデータ分析中の行動を比較する．データ駆動アプローチは，特定の仮説を持

*連絡先：名古屋大学大学院情報学研究科
〒 464-8601 名古屋市千種区不老町
E-mail: muro@cog.human.nagoya-u.ac.jp

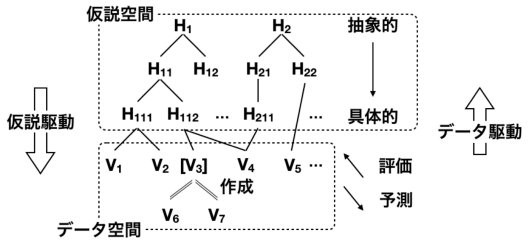


図 1: 仮説空間とデータ空間．H は仮説を，V は変数を表す．

たずに，データの分析を行い，その結果から理論を構築するアプローチである [7, 8]．近年では，遺伝学などでデータ駆動アプローチによる発見がなされている．

1.2 仮説空間とデータ空間

先述した仮説と独立変数の操作の関係は，仮説空間と実験空間の二重空間探索として論じられている [3, 9]．本研究では，仮説とデータの関係を，同様の形で仮説空間とデータ空間の二重空間探索として考える．図 1 にその概念図を示す．なお，本節の考えは，研究者が実験を計画し，関連する様々な変数の値を収集し終わった後のデータ分析を対象としている．

1.2.1 データ空間

データ空間は実験において収集された全てのデータおよび関連データを含む．データは従属変数の候補である各変数とその値の組合せとして定義される [7, 10]．

データ空間の探索はデータ分析を通して行われる．分析の対象として選択された変数が従属変数として選択されたとされる．値が収集されていたとしても，分析の対象として選択されていない変数は，従属変数とならず，探索もされなかったとされる．もし，研究者が特定の変数や類似した変数のみを繰り返し分析していた場合，データ空間の探索が偏っていたと定義される．

さらに，研究者は複数の変数の値を組み合わせ，その値を直接測定することができない新しい変数を作成することができる．たとえば，ポストテストの得点をプレテストの得点で割ることで，テスト得点の変化率を作成することができる．

1.2.2 仮説空間

仮説空間は全ての仮説を含み，それらの仮説は抽象的なものから具体的なものへと階層構造をもつ [2]．抽象的な仮説は一般的で多くの事象を説明するが，そこから特定の変数の値の傾向を予測することはできない．一方，具体的な仮説は厳密で少数の事象しか説明できないが，特定の変数の値の傾向を予測する．そのため，十分に具体的な仮説は，対応する特定の変数の値を分析することにより検証される．

2 研究アプローチとデータ空間探索

2.1 探索の広さ

仮説駆動アプローチを用いた場合，2 つの空間は仮説空間からデータ空間の方向に探索が行われる．研究者はしばしば中程度の抽象度を持つ仮説を検証対象とするため，その子にあたる仮説と対応するデータのみが分析の対象とされる．つまり，探索対象とする仮説空間が制限されることにより，データ空間の探索も制限されると考えられる．

一方で，データ駆動アプローチを用いた場合は，分析の結果に基づき理論を構築するため，データ空間から仮説空間の方向へと探索が行われる．研究者は自身の目的や興味に従って，データ空間全体から分析対象とする変数を選択するため，仮説駆動アプローチを用いる場合のように，データ空間の探索が制限されることはない．

2.2 探索の深さ

仮説駆動アプローチを用いた場合，研究者は仮説を立証する可能性がある変数を予測し，それらを分析することにより自身の仮説を支持する証拠をできるだけ多く入手しようとする．そのため，予測した変数が存在しない場合はその変数を作成すると考えられる．つまり，データ空間の表層にあたる取得データだけではなく，分析に労力を要するデータ空間の深い部分まで探索が行われることが予測される．

一方で，データ駆動アプローチを用いた場合は，研究者は興味を持った全ての変数を分析していくのみである．そのため，新しい変数の作成等の深い分析は必要とされず，データ空間の表層的な探索が行われることが予測される．

2.3 本研究の仮説

以上より，本研究では以下の 2 つの仮説を検証する．

深さ仮説 仮説駆動アプローチはデータ駆動アプローチよりも深い探索を促進する．

広さ仮説 データ駆動アプローチは仮説駆動アプローチよりも広い探索を促進する．

さらに，仮説駆動アプローチをとったときの仮説が，データにより支持されるか否かで，探索の深さと広さが変化するかを検討した．

表 1: データセット中の各変数.

カテゴリ	変数名	変数の説明
学習過程	使用時間	システムを使用した合計時間
	問題数	システムを使用して解いた問題数
	使用回数	システムを使用した回数
	不正解数	システム使用中に間違った回答をした回数
学習結果	全体得点	ポストテスト全体の合計点
	基礎問題得点	ポストテストの基礎問題の合計点
	応用問題得点	ポストテストの応用問題の合計点
質問紙	熟考度	質問「よく考えて問題を解いた」への回答
	動機付け	質問「意欲を持って自宅学習へ取り組んだ」への回答
	難易度	質問「システムを使うのは難しかった」への回答
基礎学力テスト	数学	数学の基礎学力テストの得点
	国語	国語の基礎学力テストの得点
学生調査	兄弟姉妹数	兄弟姉妹合計数
	携帯電話使用時間	1 日に携帯電話を使用する時間
	塾の時間	一週間に塾で学習をしている時間

3　データ分析課題

参加者は，仮説駆動アプローチを用いる仮説駆動条件とデータ駆動アプローチを用いるデータ駆動条件に分けられ，ある架空の実験で収集されたデータを分析するよう求められた.

3.1　シナリオ

データ分析は以下のシナリオ教示のもと行われた. (1) 数学の証明問題用の 2 つの自宅学習用システムが開発された. (2) その評価実験が行われた. (3) 参加者の課題は評価実験で集められたデータを分析することである. 参加者の課題は 2 つのサブタスクからなった. サブタスク 1 はどちらのシステムの学習効果が高いかを検討すること，サブタスク 2 は一方のシステムの学習効果が高いのはなぜかを検討することであった.

仮説駆動条件の参加者は，シナリオ内では大学教授の立場を与えられ，自身の開発した新システムの方が従来の旧システムよりも学習効果が高いという仮説が与えられた. 一方，データ駆動条件の参加者は，開発と関係ない第三者の立場から，新旧システムの名前はそれぞれシステム 1，システム 2 とし分析を行わせた. 以降では，仮説駆動条件に合わせ，新旧システムと言及する. なお，参加者には，分析結果の解釈に利用できるよう，各システムの使用法が教示された.

学習効果の評価実験は以下の手続きでなされたと教示された. 学力が同程度の 30 名の学生に，授業の後，15 名には新システムを，残りの 15 名には旧システムを利用させ 1 週間の自宅学習を行わせた. その後の授業で，基礎問題と応用問題からなる確認テストを実施し，学生にシステムの使用に関する質問紙に回答させたとした.

3.2　データセット

分析対象とされたデータの変数を表 1 に示す. 評価実験での収集データと関連データが含まれていた.

確認テストの得点は，仮説駆動条件の参加者の持つ仮説を支持するように，全体得点と応用問題得点の点数が新システムを使用した条件で有意に高くなるように設定された. さらに，新システムの方が学習効果が高い理由として「新システムを使用するとよく考えて問題を解いていた」という結論が導出されやすいように，各変数の値が設定された. 具体的には，熟考度と不正解数において，新システムを使用した条件で値が有意に大きかった. さらに，使用時間を問題数で割ることにより求められる，一問あたりの思考時間を示す変数も同様に新システムを使用した条件で値が有意に大きかった. ただし，使用時間，問題数の値は 2 つのシステムで有意な差が生じないように設定された.

4　分析支援ツール

本実験のため，従属変数と統計的検定の種類を選択するのみで統計的な検定結果を表示する，分析支援ツールを構築した. 使用できる統計的検定は平均の比較 (t 検定) と相関の有意性の検定であった. さらに，2 つの

変数と演算子を選択することにより，新しい変数の作成を行う機能も実装した．

各条件の分析過程は分析支援ツールにより統制された．仮説駆動条件では，はじめに参加者は具体的な仮説を記述した．その後，仮説を検証するための従属変数と検定方法の記述，検定の実施，仮説が支持されたかどうかの判断を順に行わせた．

データ駆動条件では，参加者は仮説の記述をすることなく，検定の方法を決定，記述した．統計的検定の実施の後に，なぜそのような結果が得られたかに関して仮説や解釈を記述させた．仮説の記述 (仮説駆動)，または，検定の決定 (データ駆動) から検定を繰り返し実行できた．なお，記述のフォームやツールの操作量は，両条件でほぼ同一となるよう調整された．

5　実験 1, 2

2 つの実験を合わせて記述する．実験 1 では，仮説駆動条件の参加者の持つ仮説が支持され，実験 2 では支持されないよう操作した (詳細は後述)．

5.1　方法

5.1.1　参加者

実験 1 には，仮説駆動，データ駆動条件に 23 名ずつ，全 46 名の学部生が参加した．実験 2 には，仮説駆動条件に 23 名，データ駆動条件に 24 名の，全 47 名の学部生が参加した．

5.1.2　手続き

参加者ははじめに，2 つの統計的検定とツールの使い方の簡単な説明を受けた．続いて，自宅学習システムとその評価実験，収集されたデータの説明が与えられた．その後，参加者はサブタスク 1 から 30 分間のデータ分析課題を開始した．学習効果に関して十分な検討が終了したら，参加者はその旨を入力しサブタスク 2 へと進んだ．課題終了後に，分析の結果からわかったことを記述させた．

実験 1 と実験 2 の差異は与えられたデータのみであった．実験 1 では，上述の通り，仮説駆動条件の参加者が持つ仮説が支持されるように値が設定されたデータが与えられた．実験 2 では，実験 1 において，新システムに割り当てられていたデータが旧システムへ，旧システムに割り当てられていたデータが新システムへと割り当てられた．これにより，各変数の値は，仮説駆動条件の参加者が持つ仮説を支持することがなくなった．他の手続き等は，両実験において同一である．

5.2　結果

与えられたデータの内容と一貫する記述を行なった参加者のデータのみが分析に用いられた．実験 1 の仮説駆動条件から 2 名，実験 2 の仮説駆動条件から 5 名，データ駆動条件から 1 名の参加者が除外された．

5.2.1　統計的検定の実施回数

仮説駆動条件に与えられた仮説とサブタスク 1 はともにシステムの学習効果と関連していた．そのため，サブタスク 1 における行動に仮説の影響がある可能性がある．図 2 に全体，サブタスク 1，サブタスク 2 における統計的検定の平均実施回数を示す．

全体の実施回数は実験 1 ではデータ駆動条件の方が多い傾向が見られたが ($t(42) = 1.304, p = .100$)，実験 2 では条件間に有意な差異はなかった ($t(39) = 1.091, p = .141$)．全体に占めるサブタスク 1 のための実施回数の割合を条件間で比較した．その結果，サブタスク 1 のための検定実施割合は，実験 1 では仮説駆動条件 (.280) よりもデータ駆動条件 (.584) において有意に多いが ($t(42) = 3.858, p < .001$)，実験 2 では逆にデータ駆動条件 (.538) よりも仮説駆動条件 (.717) において有意に多かった ($t(39) = 2.228, p = .016$)．

この結果は，自身の持つ学習効果に関する仮説が支持された場合は，学習効果の検討を素早く打切り，支持されなかった場合は長く検討を続けたことを示す．

5.2.2　探索の深さ

探索の深さの指標として，新しく作成された変数の数を用いた．両実験共に，仮説駆動条件において，データ駆動条件よりも有意に多くの変数が作成されていた (実験 1, 0.851 vs. 0.348, $t(42) = 2.343, p = .024$; 実験 2, 0.600 vs. 0.136, $t(39) = 2.310, p = .030$)．ここから，仮説が支持されるか否かにかかわらず，仮説駆動条件の参加者の方がデータ空間の深い探索を行っていたことが示される．

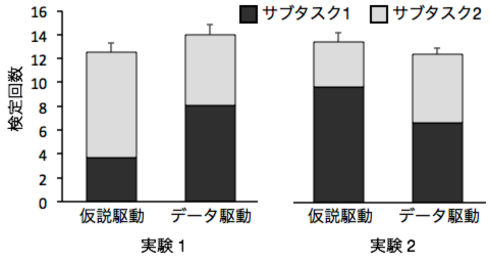


図 2: 統計的検定の平均実施回数 (エラーバーは標準誤差を示す)．

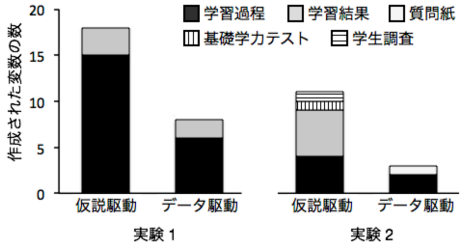


図 3: 作成された変数の種類.

図 3 は新しく作成された変数のカテゴリ (表 1 参照) を示す．変数の作成に用いられた 2 つの変数のカテゴリが同一である場合，新しい変数も同様のカテゴリに分類された．同一でない場合は，参加者が新しい変数につけた名前に基づき，どちらか一方のカテゴリに分類された．仮説駆動条件に着目すると，学習効果に関する仮説が支持された実験 1 では，多くの変数は学習過程へ分類された．一方で，学習効果に関する仮説が支持されない実験 2 では，学習結果へ分類される変数が増加している．

5.2.3 探索の広さ

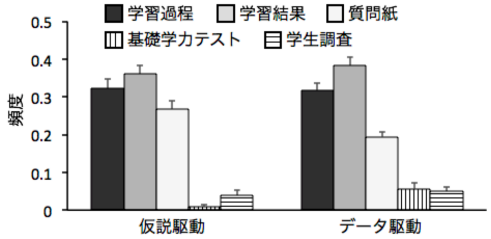
表 1 に示される各カテゴリの変数が検定に用いられた頻度を算出した．その結果を図 4 に示す．カテゴリ間の頻度の偏りを示す指標として，0 から 1 に範囲を調節したエントロピーを下記の式を用い算出した．エントロピーが小さいほど探索が偏っていたことを示す．

$$adjusted{-}H = \left(\sum_{C=category} P(C) \log_2 P(C) \right) / \log_2 5$$

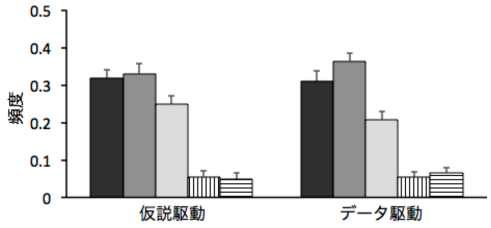
実験 1 では，仮説駆動条件 (.708) において，データ駆動条件 (.768) よりも有意にエントロピーが小さかった ($t(42) = 1.850, p = .035$)．一方，実験 2 では 2 つの条件のエントロピーに有意な差異はなかった (仮説駆動 .765, データ駆動 .769; $t(39) = 0.102, p = .460$)．データ駆動条件では，2 つの実験でエントロピーの値がほぼ同等であったことを考慮に入れると，仮説駆動条件の参加者は，自身の仮説が支持されなかった実験 2 においてデータ空間の探索範囲の偏りを減らし，広い範囲を探索するようになったことが示される．

6 考察

本研究は，2 つの異なる研究アプローチが，データ分析における従属変数の選択に与える影響を検討した．



(a) 実験 1 における検定頻度



(b) 実験 2 における検定頻度

図 4: 各カテゴリに含まれる変数が従属変数として選択された頻度 (エラーバーは標準誤差を示す).

6.1 探索の深さ

本研究ではデータ空間の探索の深さに関して，仮説駆動アプローチはデータ駆動アプローチよりも深い探索を促進するという深さ仮説を検証した．仮説駆動条件の参加者は自身の仮説が支持されるか否かにかかわらず，新しい変数を多く作成したという実験の結果は，深さ仮説を支持する．

作成された変数の種類に実験間で差異が見られたことは，仮説駆動条件の参加者が 2 つの実験において異なる仮説の検証のためにデータ空間の探索を行っていたことを示す．学習効果に関する仮説が支持される実験 1 では，検定回数の結果からも示されるとおり，参加者は学習効果の検証のための検定を短時間で打ち切っていた．ここから，実験 1 では，仮説駆動条件の参加者は，主にサブタスク 2 にあたる学習効果が生じる理由を検討するために変数の作成を行っていたと考えられる．実際に，変数の作成を行った 14 名の参加者のうち 8 名は，実験者が準備した理由を支持する変数である，一問あたりの思考時間にあたる変数を作成していた．

一方，実験 2 では学習効果に関する仮説はデータにより支持されないが，先行研究に示される通り，参加者は仮説をすぐには棄却せず，長い検討を続けた [11]．取得された学習結果に関するデータは仮説を支持しないため，新しく学習結果に関する変数を作成し，仮説を支持する証拠を見つけようとしたと考えられる．これらの結果から，参加者は自身の仮説を支持する証拠が不足していると感じたときに，データ空間の深い探索を行うことが示唆される．

6.2 探索の広さ

データ空間の探索の広さに関しては，広さ仮説において，データ駆動アプローチは仮説駆動アプローチよりも広い探索を促進することを予測した．実験 1 では，データ駆動条件の参加者の方が，仮説区駆動条件よりも偏りの少ない探索を行っており，広さ仮説を支持する結果が得られた．しかし，実験 2 では，仮説駆動条件における探索の偏りは解消され，両条件の探索の広さは同程度となっており，広さ仮説は支持されなかった．

実験 2 の結果は，一見，仮説駆動アプローチをとっていても，仮説が支持されない場合はデータ駆動アプローチへと方略を切り替え，新しい仮説を見つけるという Klahr (2000) の結果と一致しているように見える [12]．しかし，重要な点として，探索の深さは実験 2 においても仮説駆動条件の方が深かった．もし，アプローチを変更しているのであれば，探索の深さは浅くなるはずである．

我々は，実験 2 の仮説駆動条件の参加者は，仮説駆動アプローチを継続して用いていただろうと考える．仮説駆動アプローチを用いた参加者は，最初に探索を行った狭い範囲では，自身の仮説を支持するデータを見つけることができなかった．そこで，アプローチを変更するのではなく，仮説を保持したまま，データ空間の探索範囲を広げることにより，仮説を支持するデータを探そうとしたと考えられる．その結果，実験 2 では両条件の探索の偏りに差異がなくなったのである．

7 結論

データ駆動アプローチを用いた参加者はデータ空間を広く，浅く探索する一方，仮説駆動アプローチを用いた参加者はデータ空間を狭く，深く探索することが示された．ただし，仮説駆動アプローチを用いた場合，自身の仮説が支持されないと，仮説を支持するデータを求め，探索範囲を広げていくことが明らかとされた．これらのアプローチの影響は，学生に科学的研究の方法を教授する上で重要である．今後は，分析結果から導かれる結論の内容も検討する必要がある．

謝辞

この研究に貢献した名古屋大学情報文化学部岡村渓太君に感謝の意を表します．

参考文献

[1] Van Joolingen, W. R., De Jong, T.: Supporting hypothesis generation by learners exploring an interactive computer simulation, *Instructional Science*, Vol. 20, No. 5-6, pp. 389–404 (1991)

[2] Van Joolingen, W. R., De Jong, T.: An extended dual search space model of scientific discovery learning, *Instructional Science*, Vol. 25, No. 5, pp. 307–346 (1997)

[3] Klahr, D., Dunbar, K.: Dual space search during scientific reasoning., *Cognitive Science*, Vol. 12, No. 1, pp. 1–48 (1988)

[4] Shute, V. J., Glaser, R.: A large-scale evaluation of an intelligent discovery world: Smithtown, *Interactive Learning Environments*, Vol. 1, No. 1, pp. 51–77 (1990)

[5] Chen, Z., Klahr, D.: All other things being equal: Acquisition and transfer of the control of variables strategy, *Child Development*, Vol. 70, No. 5, pp. 1098–1120 (1999)

[6] Grolemund, G., Wickham, H.: A cognitive interpretation of data analysis, *International Statistical Review*, Vol. 82, No. 2, pp. 184–204 (2014)

[7] Jolaoso, S., Burtner, R., Endert, A.: Toward a deeper understanding of data analysis, sensemaking, and signature discovery, *Human-Computer Interaction*, pp. 463–478 (2015)

[8] Kell, D. B., Oliver, S. G.: Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era, *Bioessays*, Vol. 26, No. 1, pp. 99–105 (2004)

[9] Simon, H. A., Lea, G.: Problem solving and rule induction: A unified view, In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 105–127) Hillsdale, NJ: Erlbaum (1974)

[10] De Mast, J., Trip, A.: Exploratory data analysis in quality-improvement projects, *Journal of Quality Technology*, Vol. 39, No. 4, pp. 301–311 (2007)

[11] Mason, L.: Responses to anomalous data on controversial topics and theory change, *Learning and Instruction*, Vol. 11, No. 6, pp. 453–483 (2001)

[12] Klahr, D.: *Exploring science: The cognition and development of discovery processes*, Cambridge, MA: The MIT Press (2000)