RIETI Discussion Paper Series 18-J-032

データ利活用に関する質問票調査を用いた産業別比較

立本 博文
筑波大学

平井 祐理
東京大学政策ビジョン研究センター

渡部 俊也
経済産業研究所

RIETI
Research Institute of Economy, Trade & Industry, IAA
独立行政法人経済産業研究所
https://www.rieti.go.jp/jp/

RIETI Discussion Paper Series 18-J-032
2018 年 12 月

データ利活用に関する質問票調査を用いた産業別比較*

立本 博文（筑波大学大学院ビジネス科学研究科）
平井 祐理（東京大学政策ビジョン研究センター）
渡部 俊也（経済産業研究所）

要　旨

企業がデータ利活用を通じて生産性を高めることや高付加価値の製品サービスを提供することは、日本産業にとって重要な課題となっている。そうした中、IoTやビッグデータ、人工知能といった新しいデータ技術は、産業の生産性を高度に向上させる切り札として期待されている。本研究では、「企業のデータ利活用に関して、産業毎の違いがあるのではないか」とのリサーチ・クエスチョンをもとに、質問票調査を用いて探索的分析を行った。産業毎の特徴値（平均値）の比較を行ったところ、多くの産業で似通ったものであり、データ利活用の成果についても特別に優越している産業が存在しているわけでないことがわかった。また、データ属性や行動属性を入力とし、成果項目を出力としたときに算出される反応係数についてクラスター分析を行ったところ、高度なデータ資源やデータ技術に関して反応係数がマイナスの産業群と、反応係数がプラスの産業群があることがわかった。これらの分析結果から、「データ資源やデータ技術への投資」と「最終的なデータ利活用成果」にはギャップがあることが示唆された。

キーワード：データ利活用、IoT、ビッグデータ、AI、質問票調査、産業別比較
JEL classification: O34

*本稿は、独立行政法人経済産業研究所（RIETI）におけるプロジェクト「企業において発生するデータの管理と活用に関する実証研究」の成果の一部である。本稿の分析に当たっては、RIETI が実施した平成 29 年度「データ利活用に関するアンケート調査」を利用した。また、本稿の原案に対して、橋本正洋教授（東京工業大学）、梶川裕矢教授（東京工業大学）、小川紘一客員研究員（東京大学）、池田毅弁護士（森・濱田松本法律事務所）、日置巴美弁護士（内田・鮫島法律事務所）、古谷真帆客員研究員（東京大学）、二又俊文客員研究員（東京大学）、高野泰朋特任研究員（東京大学）、本プロジェクトのオブザーバー、ならびに経済産業研究所ディスカッション・ペーパー検討会の方々から多くの有益なコメントを頂いた。ここに記して、感謝の意を表したい。

目次

1. はじめに

このディスカッション・ペーパーでは、「企業のデータ利活用に関して、産業毎の違いがあるのではないか」とのリサーチ・クエスチョンをもとに、独立行政法人経済産業研究所（RIETI）が実施した平成 29 年度「データ利活用に関するアンケート調査」を用いて探索的分析を行った結果を報告する。

企業がデータ利活用を通じて、生産性を高めることや高付加価値の製品サービスを提供することは、日本産業にとって重要な課題となっている。1980 年代に始まった IT 化の流れ以降、データを用いた事業活動には長い歴史がある。その一方、近年の IoT デバイスの普及、ビッグデータや人工知能の盛り上がりは、従来のデータ利活用とは異なる次元のものであるとの見解もある。新しいデータ技術は、産業の生産性を高度に向上させる切り札として期待されている。

新しいデータ資源・技術に対する期待や不安は大きい。産業構造の急激な転換をもたらしてしまうのではないか、との意見もある。「従来の産業区分の境界を壊してしまう」とする警告がある。そうなれば競争のルールや事業の収益性は大きく変わってしまう。このため産業によっては、新しいデータ資源や技術への投資が進まない可能性がある。

また、そもそも自社事業に新しいデータ資源・データ技術をどのように適用すればよいのかということがよくわかっていない場合も多い。産業毎にデータ利活用の形態が異なる可能性もある。新しいデータ資源・技術を、どのように用いるのかが不透明な産業では、積極的な投資を躊躇する企業も多いだろう。このような理由から、企業のデータ利活用に関して、産業毎に違いが生じているのではないだろうか。この問いが、本研究のリサーチ・クエスチョンである。

データ利活用が進む産業がある一方、データ利活用が著しく遅れる産業が出現するかもしれない。データ利活用の流れに取り残される産業の存在は、日本産業のボトルネックになる危険性がある。しかしながら、データ利活用に関して産業毎に比較を行った調査はほとんど存在しない。その理由の一つが、そもそもデータ利活用に関する広範なサーベイ・データが存在しないためである。この課題に対処するため、本調査研究では、平成 29 年度「データ利活用に関するアンケート調査」を用いて、企業のデータ利活用に関する産業間の比較を行った結果を報告する。

本ディスカッション・ペーパーの構成は次のとおりである。まず第 1 章の残りで本調査分析の分析枠組みを説明し、分析で取り上げる属性項目と成果項目を紹介する。第 2 章では、属性項目と成果項目の平均が産業毎に異なるのかを、多重比較法を用いて探索的に検討する。第 3 章では、属性項目を入力とし、成果項目を出力としたときに算出される反応係数について、産業毎の傾向を確認するため、クラスター分析を行う。第 4 章では、第 2 章と第 3 章の分析結果を整理し、まとめを行う。

1.1. 分析枠組み

平成 29 年度「データ利活用に関するアンケート調査[i]」の調査票は、①日本における平成 24 年の特許出願件数上位 5000 社から自治体や大学などを除いた企業、②東証一部上場企業、③ビジネス SNS サイト WANTEDLY から抽出した企業、の 3 つの企業リストを用いて、計 6278 社に送付した。調査期間は 2017 年 9 月 15 日～2017 年 11 月 27 日で、562 社から有効回答を得た（有効回答率 9.0%）。

本アンケート調査では 2016 年度の状況について回答を依頼した。アンケート調査票には大分類としての設問は 3 つあり、設問 1 では従業員数や業種といった当該企業のプロフィールについて、設問 2 では当該企業における全社的なデータ利活用について、設問 3 では当該企業においてデータ利活用を行っている事業のうち、データの利活用が最も進んでいる事業を 1 つ選択してもらい、その事業におけるデータ利活用について、それぞれ質問した。

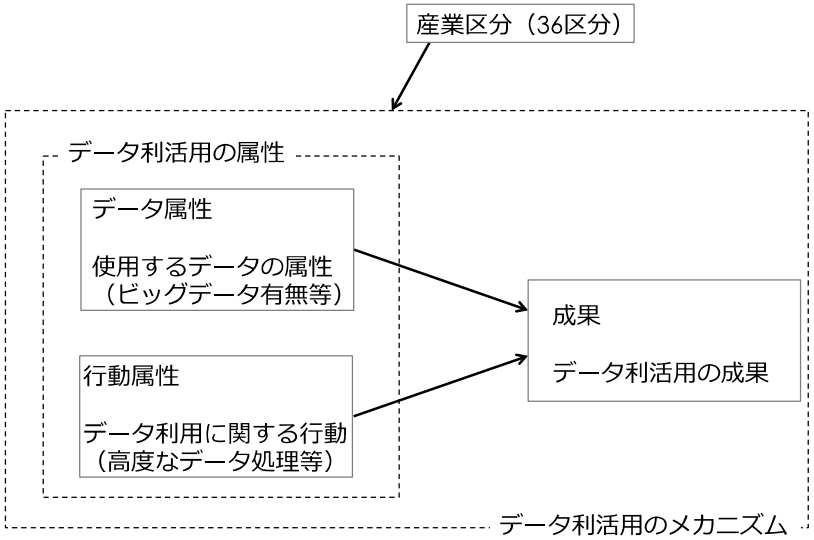

図 1 分析の枠組み

図 1 に分析枠組みを示す。各企業のデータ利活用の成果に対して、データ利活用の属性が影響を及ぼすと考えられる。データ利活用の属性は、事業で使用するデータの属性や、企業のデータ利活用に関する行動で構成される。

データの属性とは、例えば、事業で利用するデータに多くのビッグデータが含まれる、等のことである。データの利活用に関する行動とは、機械学習（ディープラーニングなど）など、高度なデータ利活用を行っているかどうか、等のことである。

事業で扱うデータ属性や、事業データに対してどのような対処をするのか、は、企業のデータ利活用の成果に影響する。このデータ利活用のメカニズムに対して、環境要因として産業特性が影響する、というのが、本分析の基本的な分析枠組みである。

産業特性として、産業の詳細な特徴を明示的に分析モデルに組み込むことも可能である。しかし、本調査では、そのような本格的調査の予備的位置づけとして、まずは、産業区分（業種区分）毎に、このようなデータ利活用のメカニズムに違いがあるのかを探索的に調査する。

1農林水産業	11石油/石炭製品	21電子部品	31金融/保険
2鉱業/採石等	12プラスチック/ゴム製品	22電子応用/計測	32情報/システム/ソフト
3建設業	13窯業/土石製品	23その他電気	33学術/研究開発機関
4食品	14鉄鋼業	24情報通信	34技術/専門サービス
5繊維/衣類	15非鉄金属	25自動車/同付属品	35その他サービス
6パルプ/紙類	16金属製品	26その他輸送用	36その他
7医薬品	17汎用機械器具	27その他製造業	
8総合化学	18生産用機械器具	28卸売/小売	
9油脂/塗料	19業務用機械器具	29通信/放送	
10その他化学	20医療用機械器具	30運輸/物流	

図 2 分析に使用した産業区分

分析には、図 2 で掲示した 36 の産業区分を用いた。この産業区分は、農林水産業、鉱業、金融保険業など、製造業以外にも幅広い産業が含まれている。従来の調査では、製造業にフォーカスした分析が行われていたが、本調査では、このように、より幅広い産業区分に基づいた産業毎の比較を行う。

産業毎の回答数は表 1 のとおりである。回答は重複回答を含んでいるため、回答企業数よりも多くなる。回答企業数は 562 であり、重複回答を含めた産業毎の回答数は 980 である。

業種ラベル	回答数	業種ラベル	回答数	業種ラベル	回答数	業種ラベル	回答数
1農林水産業	10	11石油/石炭製品	9	21電子部品	30	31金融/保険	26
2鉱業/採石等	5	12プラスチック/ゴム製品	36	22電子応用/計測	27	32情報/システム/ソフト	50
3建設業	67	13窯業/土石製品	11	23その他電気	34	33学術/研究開発機関	21
4食品	25	14鉄鋼業	11	24情報通信	20	34技術/専門サービス	36
5繊維/衣類	15	15非鉄金属	17	25自動車/同付属品	47	35その他サービス	32
6パルプ/紙類	10	16金属製品	38	26その他輸送用	20	36その他	27
7医薬品	30	17汎用機械器具	33	27その他製造業	50		
8総合化学	14	18生産用機械器具	58	28卸売/小売	52		
9油脂/塗料	8	19業務用機械器具	36	29通信/放送	5		
10その他化学	29	20医療用機械器具	28	30運輸/物流	13		

表 1 産業毎の回答数

1.2. 分析で取り上げる項目

データ属性
- Q2-1:全社データ総量
- Q2-2:データ利用率
- Q3-1:データ種類
- Q3-10:事業データ総量
- Q3-11:ビッグデータ該当
- Q3-12:データ形式

活動属性
- Q2-3:担当者合計
- Q2-4:契約書
- Q2-5:データ利活用活動
- Q3-14:データ利活用経験
- Q3-17:データイニシアティブ
- Q3-18:高度なデータ処理・解析

成果
- Q2-7:データ利活用成果
- Q3-28:事業競争力貢献

図 3 属性・成果で取り上げる設問項目

図 3 に分析で属性・成果として取り上げる項目を示す。平成 29 年度「データ利活用に関するアンケート調査」は多くの設問から構成される。それら設問項目から、データ属性・活動属性・成果に関する設問項目を抽出し、今回の分析に用いる。

データ属性：

データ属性として、設問 2-1（全社データ総量)、設問 2-2（データ利用率)、設問 3-1(データ種類)、設問 3-10(事業データ総量)、設問 3-11（ビッグデータ該当)、設問 3-12（データ形式）を用いる。

設問 2-1 の全社データ総量とは、調査対象企業が利用権限のあるデータのうち、利活用を行っている（将来利活用を行う見込みがある）データの総容量のことである。「1 台の PC で管理できる程度」「数台のサーバで管理できる程度」「専用のサーバ室、サーバセンターで管理する程度」から選択して回答する。

設問 2-2 のデータ利用率とは、設問 2-1 で回答したデータの総容量のうち、実際に利活用を行っているデータの容量である。「20%未満」「20%以上～40%未満」「40%以上～60%未満」「60%以上～80%未満」「80%以上」から選択して回答する。

設問 3 の各設問では、調査対象企業にデータ利活用が最も進んでいる事業を選んでもらい、その事業におけるデータ利活用の状況を回答してもらっている。

設問 3-1 のデータ種類とは、選んだ事業において、利活用を行っているデータの種類のことである。組織についてのデータ、個人についてのデータ、自然現象・社会現象等についてのデータ、その他から選択する。

組織についてのデータは、自社、グループ組織、顧客組織等の調査対象企業と直接的・間接的に関係のある組織で発生するデータのことである。例えば、自社の販売履歴データ、顧客の生産機器の稼働データなどが含まれる。

個人についてのデータは、自社、グループ組織、顧客組織等の調査対象企業と直接的・間接的に関係のある個人に関するデータを指す。例えば、自社従業員の勤務時行動データ、顧客組織の商業施設の監視カメラから得られた商業施設利用者の行動データなどが含まれる。

自然現象・社会現象等についてのデータは、政府データを含め、気象等の自然現象データ、交通量や住宅分布などの社会現象データ、景気や為替等の金融データを指す。

設問 3-10 の事業データ総量とは、選んだ事業において利活用を行っているデータの総容量のことである。「1 台の PC で管理できる程度」「数台のサーバで管理できる程度」「専用のサーバ室、サーバセンターで管理する程度」から選択して回答する。

設問 3-11 のビッグデータ該当とは、選んだ事業において利活用を行っているデータが、ビッグデータに該当するか否かのことである。ビッグデータとは、規模の大きさ、高い発生頻度、多様な形式という特徴を持ち、従来の技術や分析手法では処理が難しいデータを指す。例えば、画像データや音声データが含まれる。

設問 3-12 のデータ形式とは、選んだ事業において利活用を行っているデータがどのような形式のものかを指している。画像、映像、音声、テキスト、数値、位置情報、その他から選択して回答する。

行動属性：

行動属性として、設問 2-3（担当者合計）、設問 2-4（契約書）、設問 2-5（データ利活用活動）、設問 3-14（データ利活用経験）、設問 3-17（データイニシアティブ）、設問 3-18（高度なデータ処理・解析）を用いる。

設問 2-3 の担当者合計は、全社的なデータ利活用の推進する専門部門の担当者と、既存部門の中の推進担当者の合計人数のことである。

設問 2-4 の契約書は、データ利活用の利害関係者との契約書のひな型の利用実態についてである。企業のデータ利活用に関しては、利活用対象のデータが、自社で発生するデータ以外のことが多く、データ利用に関する契約を社外組織と締結することが多い。このような契約書の利用実

態について、「すでに契約書のひな型があり、それを使いこなしている」「すでに契約書のひな型はあるが、それを使いこなしていない」「契約書のひな型を作成している途中である」「契約書のひな型はない」から選択して回答する。

設問 2-5 のデータ利活用活動は、データ利活用に関する「戦略・方針について」「実施体制について」「人材について」「データについて」の①～⑭の設問に関する回答である。設問 2-5 は企業におけるデータ利活用の詳細な行動属性を問うた設問であり、それぞれについて「1.全くそう思わない」「2.そう思わない」「3.どちらともいえない」「4.そう思う」「5.強くそう思う」から選択して回答してもらった。

設問 3 の各設問では、調査対象企業にデータ利活用が最も進んでいる事業を選んでもらい、その事業におけるデータ利活用の状況を回答してもらっている。

設問 3-14 のデータ利活用経験は、選んだ事業におけるデータ利活用を何年前から行っているのかを、「5 年未満」「5 年以上～10 年未満」「10 年以上～15 年未満」「15 年以上～20 年未満」「20 年以上」から選択して回答してもらったものである。

設問 3-17 のデータイニシアティブは、選んだ事業におけるデータ利活用を企画・推進するに当たり、自社と社外の組織（グループ組織や顧客組織等）のイニシアティブの割合を示したものである。自社と社外組織の割合の合計が 100%になるように回答してもらった。このうち、分析には、自社の割合を用いた。

設問 3-18 の高度なデータ処理・解析は、選んだ事業のデータ利活用におけるデータの解析で、ディープラーニング等の高度なデータの処理・解析を行っているか否かのことである。

成果：

データ利活用の成果について、設問 2-7（データ利活用成果）、設問 3-28（事業競争力貢献)の回答を用いた。設問 2-7 の成果は全社的な成果の評価である。それに対して、設問 3-28 は事業単位の成果の評価である。

設問 2-7 のデータ利活用成果は、調査対象の企業におけるデータ利活用のこれまでの成果について質問したものである。「1.複数の事業で具体的成果（売上やコストダウンといった利益等）が得られている」「2.少なくとも 1 つの事業で具体的成果（売上やコストダウンといった利益等）が得られている」「3.具体的成果は得られていないが、複数の事業で間接的成果（事業活動に役立つノウハウやアイデアの獲得等）が得られている」「4.具体的成果は得られていないが、少なくとも 1 つの事業で間接的効果（事業活動に役立つノウハウやアイデアの獲得等）が得られている」「5.まだ成果は得られていない」の選択肢から得た回答である。オリジナルの選択肢は、1 に近いほど成果が高く、5 に近いほど成果が低いものであった。解釈容易性を考慮し、1 に近いほど成果が低く、5 に近いほど成果が高いという逆転指標を作成し、分析に用いた。

設問 3-28 では、調査対象企業にデータ利活用が最も進んでいる事業を選んでもらい、その事業におけるデータ利活用の状況を回答してもらっている。「1. 十分に貢献している」「2. 貢献している」「3. どちらでもない」「4. 貢献していない」「5. 全く貢献してない」の 5 つの選択肢から答えた回答である。オリジナルの選択肢は 1 に近いほど成果は高く、5 に近いほど成果は低い。設問 2-7 の時と同様に、解釈容易性を考慮して、設問 3-28 でも 1 に近いほど成果が低く、5 に近いほど成果が高いという逆転指標を作成し、分析に用いた。

2. 産業間の平均の比較

本節では分析枠組みで取り上げた設問群について、産業毎に特徴の違いがあるかを把握するために、産業毎の特徴値（平均値）の比較を行った。産業毎の平均値の比較にあたっては、次のような手順で行った。

まず、全設問項目について産業毎の分布に違いがないかを把握するために、全産業の分布と産業毎の分布のヒストグラムを作成した。これにより、大まかに全産業の産業毎の特徴を把握した。作成したヒストグラムは膨大であるが、分析の参考となるので Appendix 1 に示す。

次に、より本分析の分析枠組みに沿った形で、産業毎の特徴値の分析を行った。前項で設定した分析枠組みで取り上げた設問項目について、産業毎に平均値に差があるかどうかを調べた。産業毎の平均値の差を統計的に検定するには多重比較を行うことになる。多重比較では差の検定を多数回行うので第 1 種の過誤が増大する。これに対処するために、本分析では①まずクラスカル・ウォリス(Kruskal-Wallis)検定を行い、群間に差が生じているかどうかを確認し、②チューキー・クラマー法（Tukey-Kramer 法。以下、単に Tukey 法と呼ぶ）による多重比較検定を用いて母平均について群間ですべての対比較の統計的検定を行い、③その結果を CLD 図(compact letter plot)によって確認をした。このような手順は多重比較を用いた分析では一般的なものである[ii]。

2.1. データ属性について

まず、Kruskal-Wallis 検定を用いて、データ属性について産業間差がないことを検証した。有意水準は 5%とした。

設問 No	H 統計量	df	p.value	Pr(H)<0.05
問 2-1:利活用を期待しているデータの総容量	69.546	35	0.000	*
問 2-2:実際に利活用を行っているデータの容量	36.000	35	0.422	
問 3-1:データの種類_組織データの使用有無	71.350	35	0.000	*
問 3-1:データの種類_個人データの使用有無	59.220	35	0.006	*
問 3-1:データの種類_自然データの使用有無	21.204	35	0.968	
問 3-10:選んだ事業のデータの総容量	61.455	35	0.004	*
問 3-11:ビッグデータ区分	65.773	35	0.001	*
問 3-12:データの形式 1 画像データ	37.835	35	0.341	
問 3-12:データの形式 2 映像データ	33.845	35	0.524	
問 3-12:データの形式 3 音声データ	43.148	35	0.162	
問 3-12:データの形式 4 テキストデータ	40.133	35	0.253	
問 3-12 データの形式 5 数値データ	41.490	35	0.209	
問 3-12:データの形式 6 位置情報データ	52.372	35	0.030	*

表 2 データ属性の設問に対する Kruskal-Wallis 検定の結果

表 1 より設問 2-1（利活用を期待しているデータの総容量）、設問 3-1（データの種類_組織データの使用有無）と（データの種類_個人データの使用有無）、設問 3-10（選んだ事業のデータの総容量）、設問 3-11（ビッグデータ区分）、 設問 3-12（データの形式 6 位置情報データ）について、統計的有意に群間差があることが示された。つまり、これら設問以外では、産業間で差があるとは言えない。

次に、産業間で差がある可能性がある設問について、どの産業間で母平均の差が生じているのか検討した。この検討には Tukey の多重比較検定法を用いた。多重比較検定の結果は CLD 図によって表示した。

CLD 図は、各産業の母平均と 95%信頼区間を表示した図に、どの群間で母平均に差が生じているのかの情報を追記したものである。例えば、a,b で示される群があると推定されたときに、a 群と b 群の間には統計的有意に差が生じているが、a 群と ab 群の間の差は統計的有意ではない。同様に、ab 群と b 群の間の差も統計的有意ではない。以下では、このような群間差に関する情報を図 4、図 5 の CLD 図で確認する。

問 2-1:利活用を期待しているデータの総容量、について検討する。12 プラスチック/ゴム製品と 16 金属製品と 34 技術/専門サービスは、31 金融/保険よりも統計的有意に平均が小さいことが報告された。しかし、他の 32 産業間での群間差および、32 業種と前述 4 業種との群間差は統計的有意ではなかった。

問 3-1:データの種類_組織データの使用有無、について検討する。a,b,ab 群の 3 つに分かれることが推定されたが、多くの産業は b 群もしくは ab 群に属すため、群間差は統計的有意ではない。31 金融/保険は 1 産業だけ a 群に属しており、母平均が b 群の産業よりも統計的有意に小さいことが推定された。

問 3-1:データの種類_個人データの使用有無、について検討する。a,b,ab 群の 3 つに分かれることが推定されたが、ほとんどの産業は a 群もしくは ab 群に属すため、群間差は統計的有意ではない。31 金融/保険は 1 産業だけ b 群に属しており、母平均が a 群産業よりも統計的に大きいことが推定された。この推定結果は、問 3-1:データの種類_組織データの使用有無、とはまったく逆のパターンである。

問 3-10:選んだ事業のデータの総容量、について検討する。a,b,ab 群の 3 つに分かれることが推定されたが、ほとんどの産業は a 群もしくは ab 群に属す。31 金融/保険のみが b 群に属し、a 群よりも平均が統計的有意に大きいこと推定された。

問 3-11:ビッグデータ区分、について検討する。a,ab,bc,c,abc 群の 5 つに分かれることが推定されたが、ほとんどの産業は abc 群に属しており、a,ab,bc,c 群の産業と平均の差は統計的有意ではない。31 金融/保険は c 群に属しており、a 群に属す 27 その他製造業や、ab 群に属す 16 金属製品や 25 自動車/同付属品よりも、平均が統計的有意に大きいことが推定された。

問 3-12:データの形式 6 位置情報データ、について検討する。Kruskal-Wallis 検定では産業間で差がある可能性が報告されたが、Tukey の多重比較検定によって、2 産業間の母平均の差の検定をした結果、統計的有意に差が生じている産業ペアは存在しなかった。

データ属性に関する分析結果をまとめると、次のようにいえる。まずデータ属性に関して 36 業種での大きな違いは見られなかった。ほとんどの産業は、他の産業と平均値の差が統計的有意で

はない群に所属している。分析前には、産業によって使用しているデータ属性に大きな違いが生じているのではないか、との予測もあった。例えばビッグデータを使用している産業がある一方で、ほとんどビッグデータを使用していない産業があるのではないか、というような予想である。しかし、そのような予想は、今回のデータ属性の産業間での差の検討からは支持されなかった。

しかし、31 金融/保険は、1 産業だけやや他の産業とは異なる傾向を示していた。問 2-1（利活用を期待しているデータの総容量）や問 3-10（選んだ事業のデータの総容量）は、他の産業よりも大きい。問 3-11（ビッグデータ区分）では、「ビッグデータである」を選択した企業が多い。また、データの種類に関しては、問 3-1（データの種類_組織データの使用有無）と問 3-1（データの種類_個人データの使用有無）では、他の産業と比較して、組織データを使用している企業は少ないが、逆に、個人データを使用している企業は多いことがわかった。このようなデータ属性のパターンは、31 金融/保険に特有のものである。

つまり、データ属性に関して、多くの産業では顕著な違いは見られないものの、31 金融/保険についてはやや特殊であることがわかった。

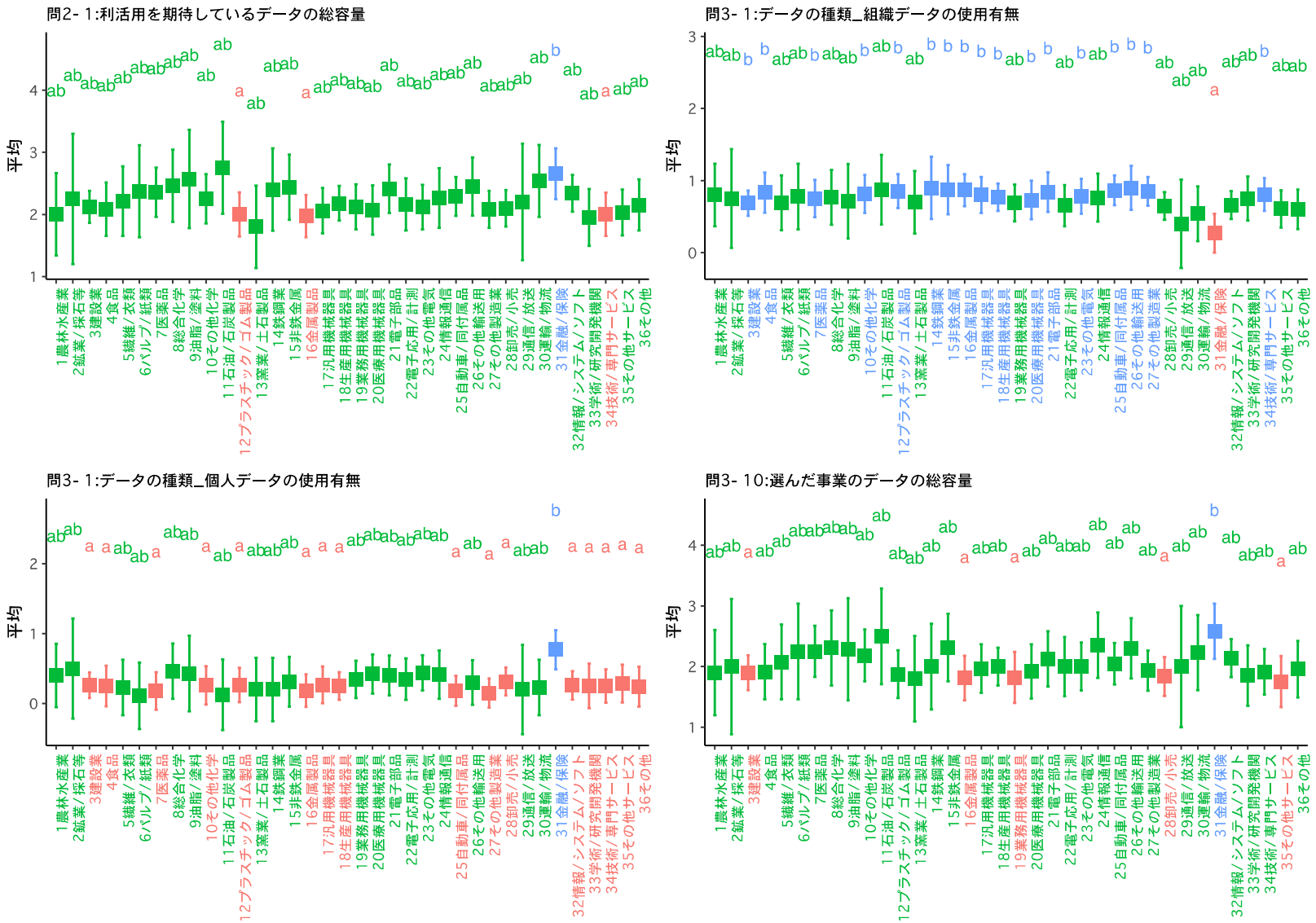

図 4 データ属性の設問に対する多重比較検定結果の CLD 図 1

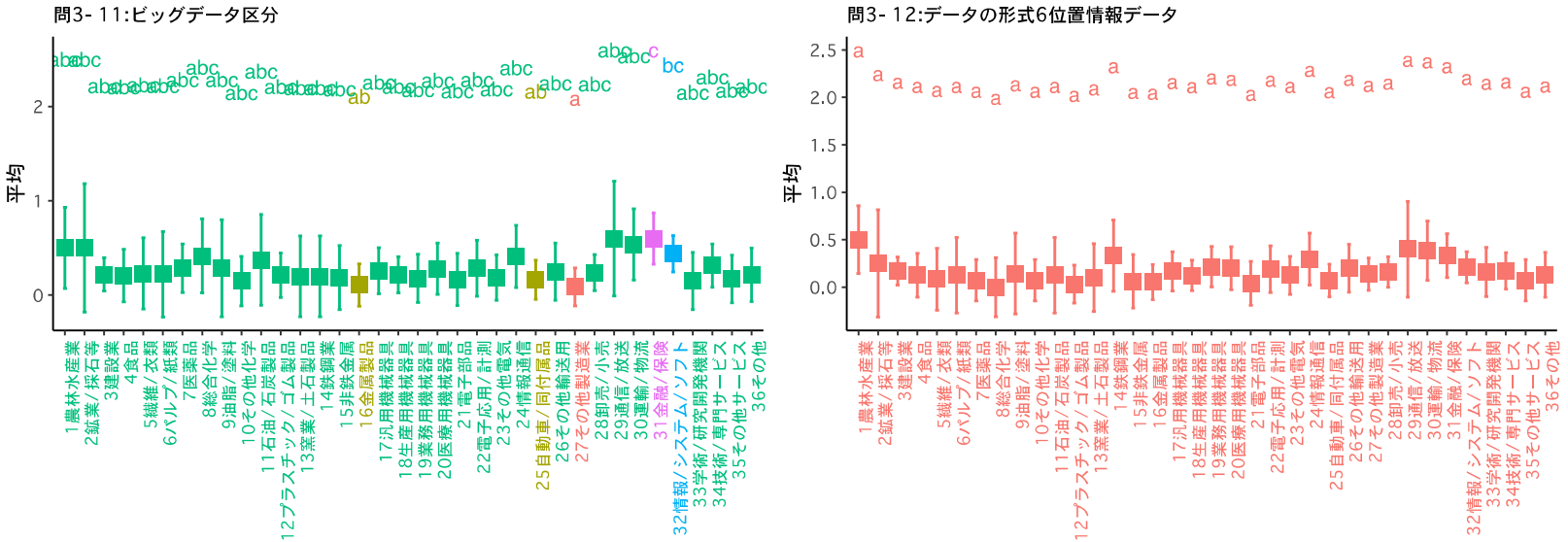

図 5 データ属性の設問に対する多重比較検定結果の CLD 図 2

2.2. 行動属性について

行動属性の設問についても、データ属性の設問と同様の分析を行った。ただし、行動属性の設問のうち、設問 2-5 は複数設問から構成されている。そのため、まず設問 2-5 以外の設問について分析を行い、その後、設問 2-5 の設問群の分析を行う。

設問 No	H 統計量	df	p.value	Pr(H)<0.05
問 2-3:担当者合計	32.461	35	0.591	
問 2-4:契約書のひな型について(1=低い, 4=高い)	82.369	35	0.000	*
問 3-14:データ利活用経験	42.724	35	0.173	
問 3-17:データイニシアティブ	34.587	35	0.488	
問 3-18:高度なデータの処理・解析の使用有無	47.055	35	0.084	

表 3 行動属性の設問に対する Kruskal-Wallis 検定の結果 1

表 3 の Kruskal-Wallis 検定の結果から、行動属性に関する設問のうち、設問 2-4（契約書のひな型について）のみが産業間の差の可能性があることが示唆された。これに基づき、設問 2-4 について、どの産業間で母平均の差が生じているのかを、Tukey の多重比較検定法を用いて分析した。多重比較検定の結果は図 6 の CLD 図によって表示した。Kruskal-Wallis 検定では産業間の差の可能性が示唆されたものの、多重比較検定の結果からは、統計的有意に母平均に差がある産業ペアは見いだせなかった。

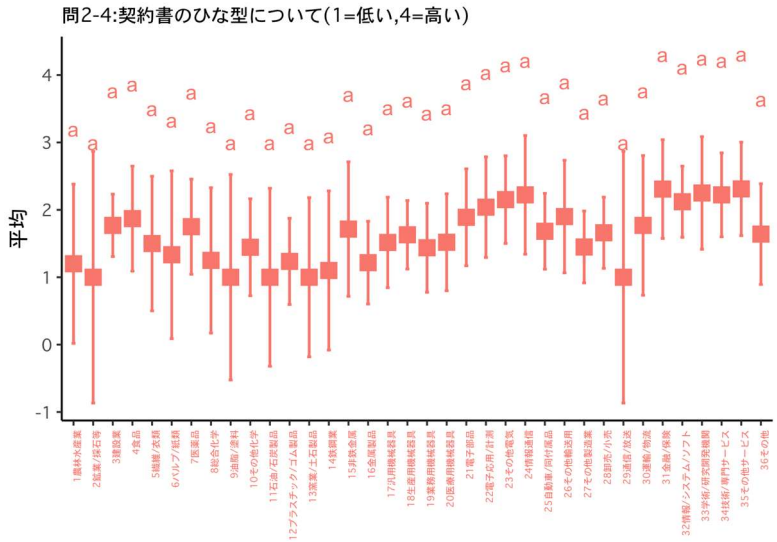

図 6 行動属性の設問に対する多重比較検定結果の CLD 図 1

次に、設問 2-5 について分析を行う。設問 2-5 はデータ利活用に関する企業行動を「戦略・方針」「実施体制」「人材」「データ」というように複数の側面から問うた設問であり、①～⑭の設問から構成されている。本アンケート調査の中で、企業のデータ利活用についての行動に関する主

要な設問である。

設問 No	H 統計量	df	p.value	Pr(H)<0.05
問 2-5:①データ利活用を積極的に推進している	43.471	35	0.154	
問 2-5:②社内全体で連携してデータ利活用を行うことを積極的に推進している	56.210	35	0.013	*
問 2-5:③社外の組織と連携してデータ利活用を行うことを積極的に推進している	82.611	35	0.000	*
問 2-5:④個人情報に該当するデータの利活用を行える体制が整備されている	51.129	35	0.038	*
問 2-5:⑤ビッグデータの利活用を行える体制が整備されている	63.197	35	0.002	*
問 2-5:⑥ディープラーニング等の高度なデータの処理・解析を行える体制が整備されている	71.194	35	0.000	*
問 2-5:⑦データのアクセス権限を共通化する等、社内全体で連携してデータ利活用を行う体制が整備されている	35.025	35	0.467	
問 2-5:⑧データのアクセスを権限に応じて制限する等、社外の組織と連携してデータ利活用を行う体制が整備されている	41.572	35	0.206	
問 2-5:⑨データサイエンティスト等、高度なデータの処理・分析を行える人材を育成、雇用している	69.349	35	0.000	*
問 2-5:⑩ディープラーニング等の高度なデータの処理・解析結果を理解し、事業活動に活かせる人材を育成、雇用している	71.290	35	0.000	*
問 2-5:⑪事業活動に関連するあらゆる情報のデータ化を進めている	65.990	35	0.001	*
問 2-5:⑫事業活動の目的や今後の展開に沿ってどのようなデータが有用か十分に吟味し、データを設計している	56.017	35	0.014	*
問 2-5:⑬複数のデータを組み合わせられるよう、データを設計している	44.582	35	0.129	
問 2-5:⑭データの有用性を評価し、必要なデータを取捨選択する等、定期的にデータ設計の見直しを行っている	60.865	35	0.004	*

表 4 行動属性の設問に対する Kruskal-Wallis 検定の結果 2

表 4 から問 2-5 の①～⑭の設問のうち、②③④⑤⑥⑨⑩⑪⑫⑭の 10 個の設問について産業毎に差がある可能性が示唆された。この 10 個の設問について、Tukey の多重比較検定を行った。その結果を CLD 図としてまとめたものが、図 7、図 8、図 9 である。これらの図を用いて産業毎の差について検討を行う。

問 2-5:②社内全体で連携してデータ利活用を行うことを積極的に推進している、について検討

する。CLD 図の結果から、a,ab,bc,c,abc 群の 5 つの群に分類できることが示唆された。ほとんどの産業は abc 群に属しており、他の産業との平均の差は統計的有意ではない。a 群には 13 窯業/土石製品が属しており、c,bc 群よりも平均が統計的有意に小さいと推定される。c 群には 32 情報/システム/ソフトが属しており、a,ab 群よりも平均が統計的有意に大きいと推定される。

問 2-5:③社外の組織と連携してデータ利活用を行うことを積極的に推進している、について検討する。CLD 図の結果から、a,ab,bc,c,abc 群の 5 つの群に分類できることが示唆された。ほとんどの産業は abc 群に属しており、他の産業との平均の差は統計的有意ではない。a 群には 16 金属製品が属しており、bc,c 群よりも統計的有意に平均が小さいことが推定される。c 群には 32 情報/システム/ソフトが属しており、a,ab 群よりも平均が統計的有意に高い。

問 2-5:④個人情報に該当するデータの利活用を行える体制が整備されている、について検討する。CLD 図より a,ab,b 群の 3 つの群に分類される。ほとんどの産業は ab 群に属しており、他の産業との差は統計的有意ではない。a 群には 13 窯業/土石製品、16 金属製品、27 その他製造業が属する。b 群には 31 金属/保険が属し、a 群の産業よりも平均が統計的有意に大きい。

問 2-5:⑤ビッグデータの利活用を行える体制が整備されている、について検討する。CLD 図より a,b,ab 群の 3 つの群に分類される。ほとんどの産業は ab 群に属しており、他の産業との差は統計的有意ではない。a 群には 3 建設業のほか 6 つの産業が属している。b 群には 32 情報/システム/ソフトが属しており、a 群の産業よりも平均が統計的有意に大きい。

問 2-5:⑥ディープラーニング等の高度なデータの処理・解析を行える体制が整備されている、について検討する。CLD 図より a,ab,b 群の 3 つの群に分類される。ほとんどの産業は a,ab 群に属しており、他の産業との差は統計的有意ではない。b 群には 32 情報/システム/ソフトが唯一属しており、a 群の産業よりも平均が統計的有意に大きい。

問 2-5:⑨データサイエンティスト等、高度なデータの処理・分析を行える人材を育成、雇用している、について検討する。CLD 図より、a,c,ab,bc,abc 群の 5 つの群に分類される。多くの産業は abc 群に属しており、他の産業との差は統計的有意ではない。a 群には 13 窯業/土石製品が唯一属しており、c,bc 群の産業よりも平均値が統計的有意に小さい。c 群には 32 情報/システム/ソフトが唯一属しており、a,ab 群の産業よりも平均が統計的有意に大きい。

問 2-5:⑩ディープラーニング等の高度なデータの処理・解析結果を理解し、事業活動に活かせる人材を育成、雇用している、について検討する。CLD 図より、a,c,ab,bc,abc 群の 5 つの群に分類される。ほとんどの産業は、ab,abc 群に属し、他の産業との差は統計的有意ではない。a 群には 13 窯業/土石製品が属しており、c,bc 群の産業よりも平均値が統計的有意に小さい。c 群には 32 情報/システム/ソフトが属しており、a,ab 群の産業よりも平均値が統計的有意に大きい。

問 2-5:⑪事業活動に関連するあらゆる情報のデータ化を進めている、について検討する。CLD 図より a,b,ab 群の 3 つの群に分類される。ほとんどの産業は ab 群に属しており、他の産業との差は統計的有意ではない。a 群には 12 プラスチック/ゴム製品が属している。b 群には 32 情報/システム/ソフトが属しており、a 群の産業よりも統計的有意に平均値が大きい。

問 2-5:⑫事業活動の目的や今後の展開に沿ってどのようなデータが有用か十分に吟味し、データを設計している、について検討する。CLD 図より、a,b,ab 群の 3 つの群に分類される。ほとんどの産業は ab 群に属しており、他の産業との差は統計的有意ではない。a 群には 3 建設業、12 プラスチック/ゴム製品、13 窯業/土石製品の 3 つの産業が属している。b 群には 32 情報/システム/ソフトが属しており、a 群の産業よりも統計的有意に平均が大きい。

問 2-5:⑭データの有用性を評価し、必要なデータを取捨選択する等、定期的にデータ設計の見直しを行っている、について検討する。CLD 図から a,b,ab 群の 3 つの群に分類される。a 群には 3 建設業、7 医薬品、12 プラスチック/ゴム製品、13 窯業/土石製品、19 業務用機械器具の 5 つの産業が属している。b 群には 32 情報/システム/ソフトが唯一属しており、a 群の産業よりも平均が統計的有意に大きい。

問 2-5 の設問群に関する検定結果をまとめると、次のとおりである。まず、多くの産業では、他の産業と統計的有意な差は生じていない。

例外的に 32 情報/システム/ソフトは、問 2-5 の複数の設問について、低い平均群の産業よりも、統計的有意に平均が大きいことが多かった。しかし、32 情報/システム/ソフトと多くの産業が属する群での平均の差は、統計的有意ではない。つまり、32 情報/システム/ソフトは、データ利活用の行動属性について、得点が低い群の産業よりは統計的有意に平均が大きいものの、多くの産業と比較すると、その差は統計的有意ではない。

一方、これとは逆に、平均の高い群の産業よりは平均が統計的有意に小さいものの、多くの産業との平均値の差は有意ではない産業もみられた。このような産業には、3 建設業、12 プラスチック/ゴム製品、13 窯業/土石製品が含まれる。

まとめると、①多くの産業では他の産業と統計的有意な差が生じているとはいえない。②活動属性が高い産業として 32 情報/システム/ソフトが挙げられる。③活動属性が低い産業として、3 建設業、12 プラスチック/ゴム製品、13 窯業/土石製品が挙げられる。

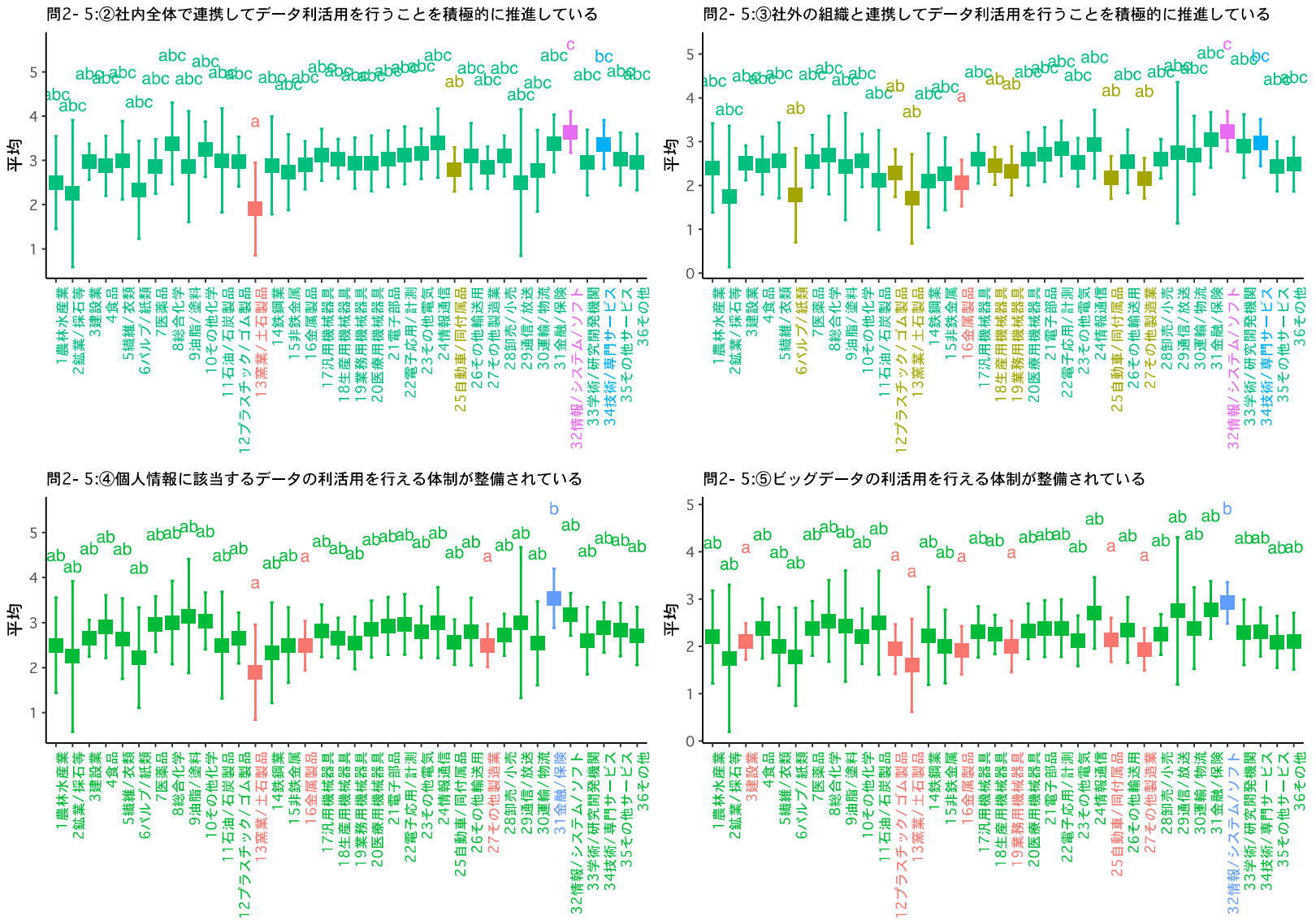

図 7 行動属性の設問（問 2-5）に対する多重比較検定結果の CLD 図 1

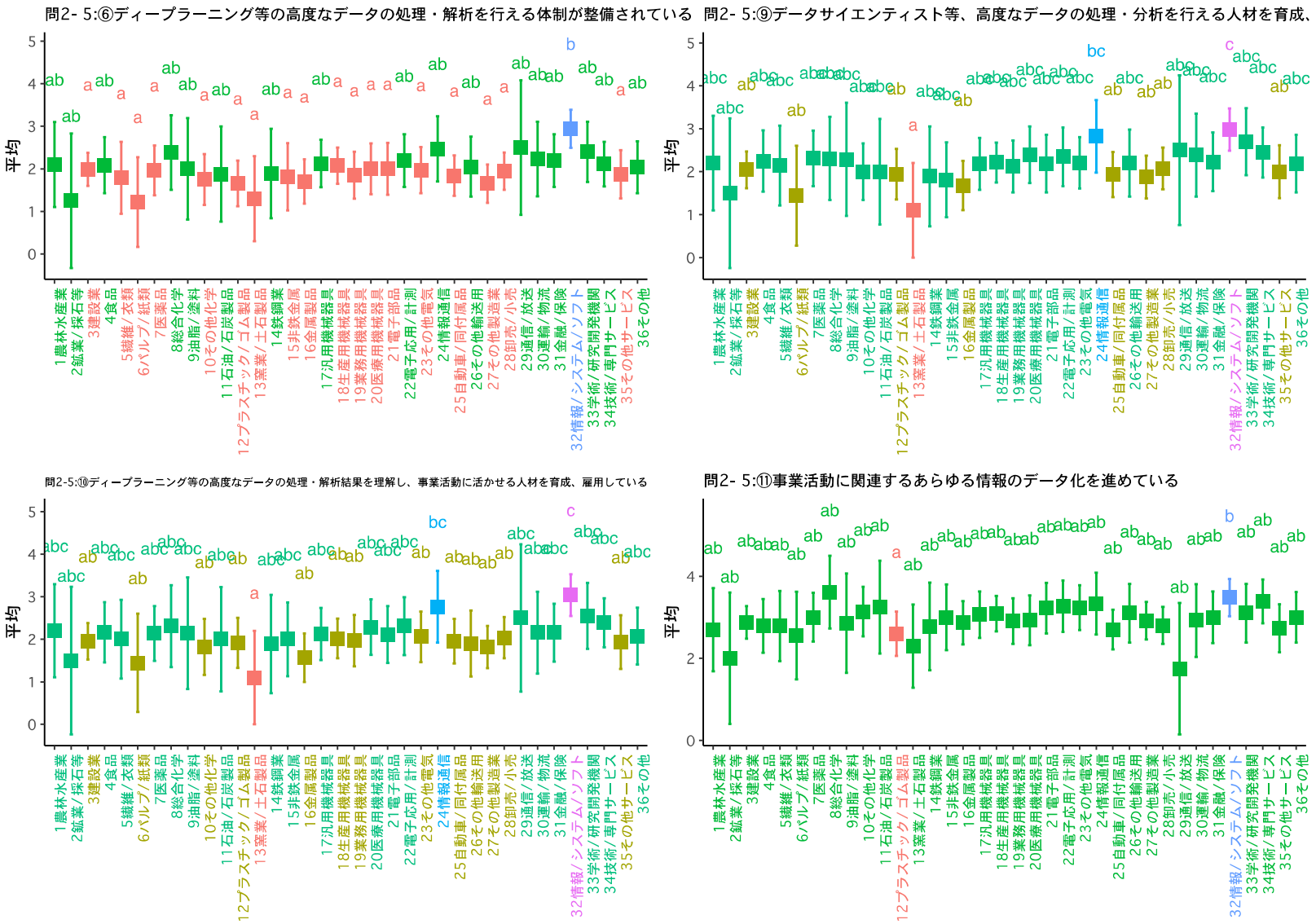

図 8 行動属性の設問に対する多重比較検定結果の CLD 図 2

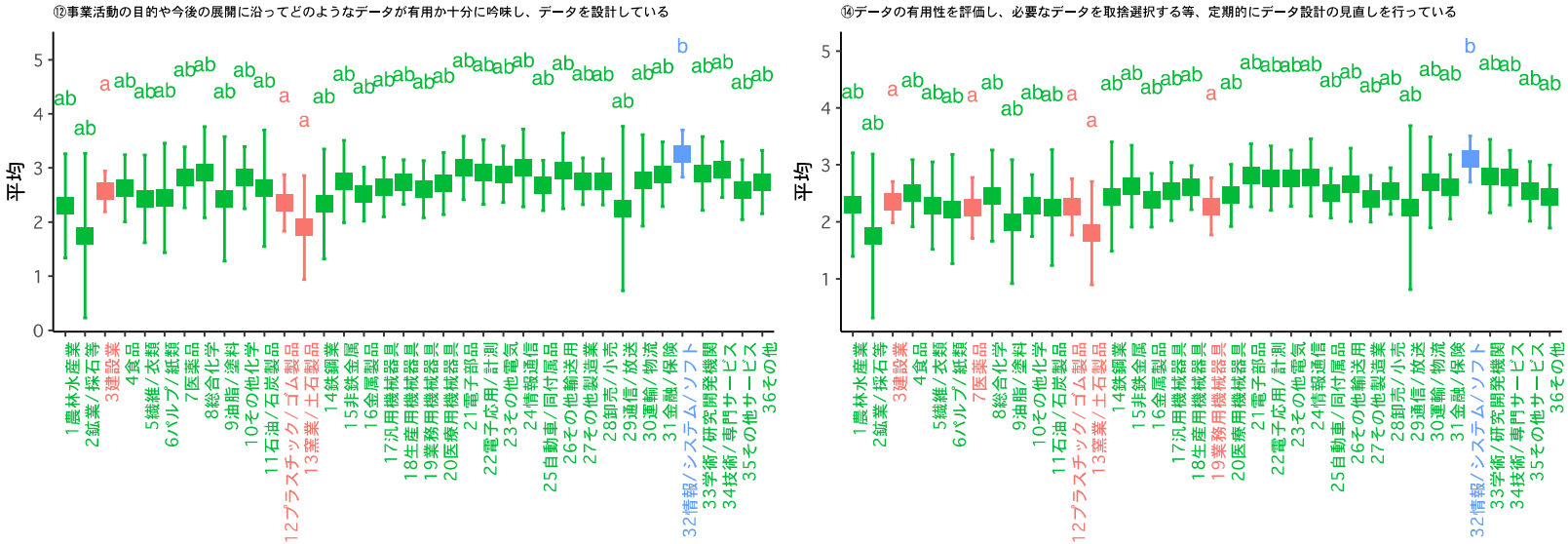

図 9 行動属性の設問に対する多重比較検定結果の CLD 図 3

2.3. 成果について

成果に関する設問についても、データ属性や行動属性の設問と同様の分析を行った。成果に関する設問は、設問 2-7（データ利活用成果）と 設問 3-28（事業競争力貢献）である。設問 2-7 および設問 3-28 については、解釈を容易にするために「1=低い、5=高い」というような変換を行った。

設問 No	H 統計量	df	p.value	Pr(H)<0.05
問 2-7:データ利活用成果(1=低い,5=高い)	41.581	35	0.206	
問 3-28:事業競争力貢献(1=低い,5=高い)	58.313	35	0.008	*

表 5 成果の設問に対する Kruskal-Wallis 検定の結果

表 5 から設問 3-28（事業競争力貢献）について産業間で差がある可能性が示唆された。この結果に基づき、設問 3-28（事業競争力貢献）について、どの産業の間で差が生じているのかを知るために、Tukey の多重比較検定法を用いて分析した。分析結果については CLD 図の図 10 を作成して整理した。表 5 の結果からは設問 3-28（事業競争力貢献）について産業間で差があることが示唆されたが、多重比較の結果をまとめた図 10 からは、統計的有意に母平均に差がある産業ペアは見いだせなかった（この点については文末脚注 ii を参照のこと）。

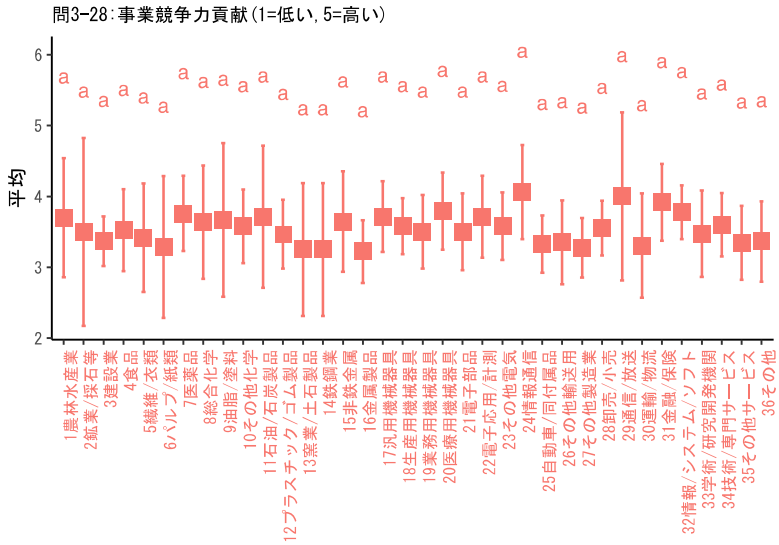

図 10 成果の設問に対する多重比較検定結果の CLD 図

成果に関する設問の分析から、成果に関して全社単位でも、事業単位でも、産業間で統計的有意に差があることは確認できなかった。つまり、今回の分析からは、現時点ではデータ利活用の成果に関して、産業間で顕著な差が生じているとは言えない、と考えられる。

3. 反応係数分析

前章では、データ属性・行動属性・成果の平均について、産業毎に違いがないかを検討した。その結果、産業毎に大きな違いはないとの結果を得た。

本章では、各属性項目を入力とし、成果項目を出力としたときの反応係数の違いについて、産業毎に違いがあるのかを検討する。

属性項目を入力とし、成果項目を出力としたモデル式(1)を想定する。

$$Y = \beta_0 + \beta_1 X \qquad \cdots(1)$$

このとき、反応係数β_1は属性項目Xを投入したときに成果項目Yへの効果を示すものである。例えばXが労働投入でYが製品産出量だとすると、反応係数β_1は（労働）生産性と解釈される。つまり、反応係数は入力に関する変換効率を示すものである。

反応係数β_1は産業の数だけ推定される。説明のため、設問2-1（全社データ総量）と設問2-7（データ利活用成果）についての反応係数を図示したものを図 11 に示す。

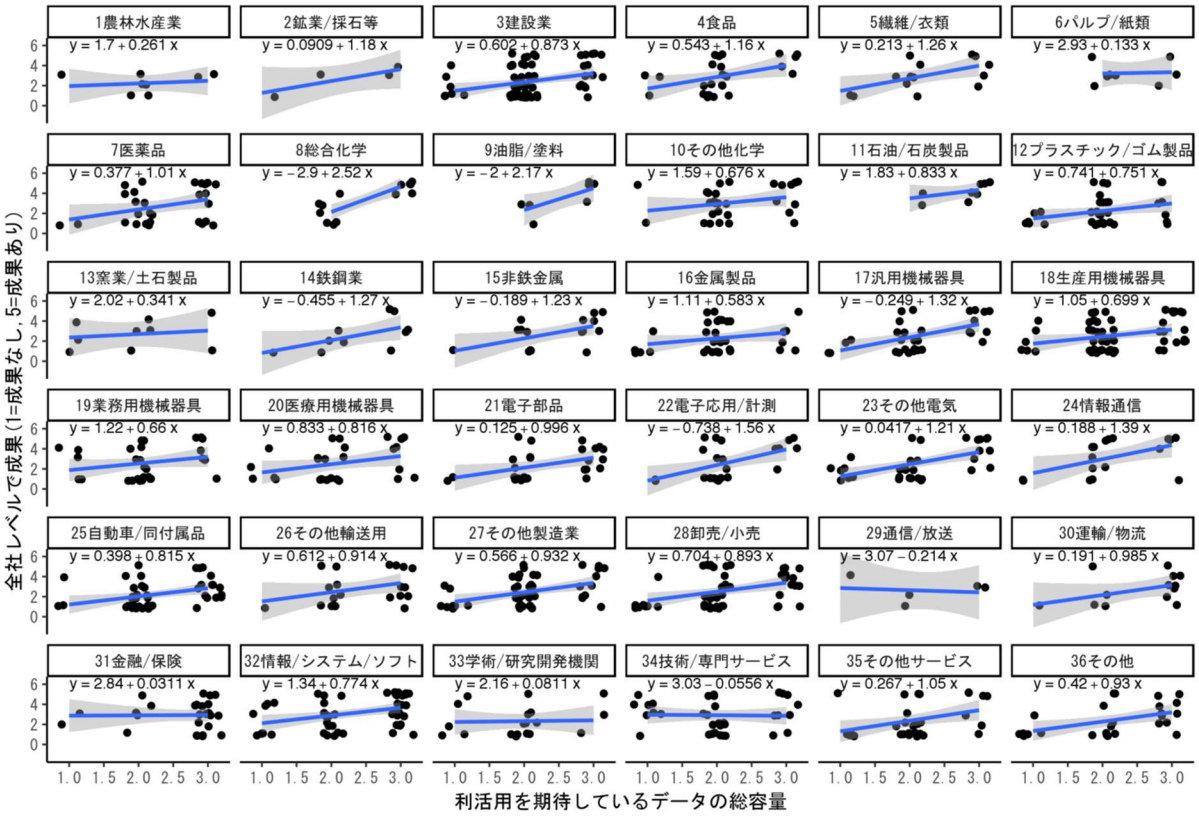

図 11 問2-1と問2-7の反応係数図（産業毎）

利活用を期待しているデータの総量が入力であり、データ利活用の成果（全社レベルでの成果）が出力である。モデル式(1)を使って、β_0とβ_1が産業毎に算出される。例えば、1 農林水産業であれば、

$$y = 1.7 + 0.261x$$

であり、(β_0, β_1)=(1.7,0.261)である。このうち、分析に用いるのはβ_1である。β_1はデータの総容量を入力としたとき、出力であるデータ利活用成果に対する変換効率（＝効果）の大きさと解釈することができる。

比較のために、全産業のデータで同様の反応係数を推定した結果を図 12 に掲げる。全データを用いて反応係数を推定すると 0.701 であり、全産業の傾向としては設問 2-1（全社データ総量）は、設問 2-7（データ利活用成果）に対して正の影響があることがわかる。

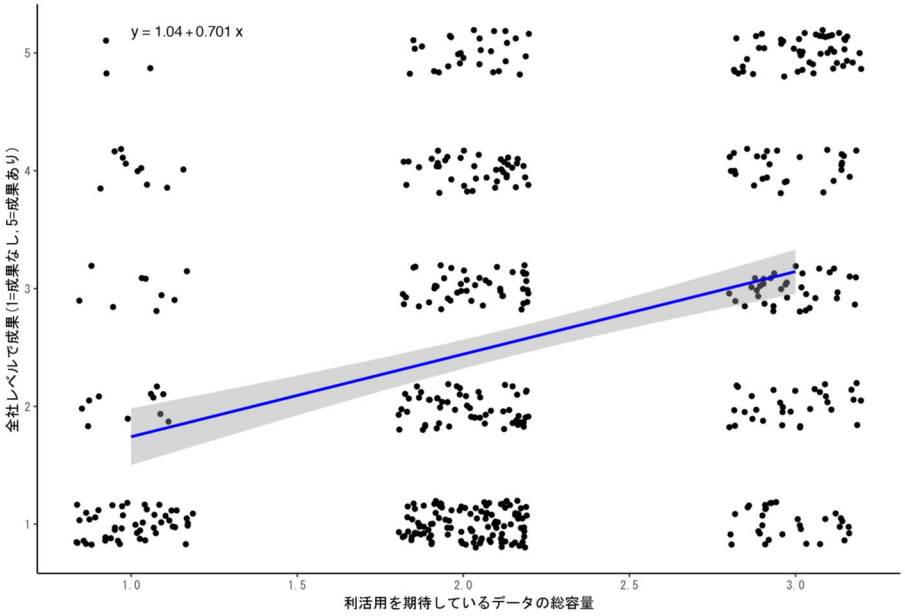

図 12 問 2-1 と問 2-7 の反応係数図（全産業）

しかしながら、すでに図 11 で確認したように反応係数は産業毎にばらついており、マイナスからプラスまでの範囲をとっている。この例では最小の反応係数は-0.214（29 通信/放送）であり、最大の反応係数は 2.52（8 総合化学）である。つまり、データ総量が増加した場合、29 通信/放送ではデータ利活用成果は減少するのに対して、8 総合化学では増加する。反応係数を推定することによって、入力項目である各属性が増加したときに、成果項目に対してどのような影響があるのかを推定することができる。産業毎に反応係数を推定すると、入力項目の影響について産業毎の違いを把握することが可能となる。

3.1. 産業毎の反応係数の推定

産業毎の反応係数の推定は、入力項目として以下の 9 つの属性項目を用いた。

データ属性を示す入力項目として以下の 4 つの設問の回答を用いた。

問 2-1（全社データ総量）

問 2-2（データ利用率）

問 3-10（事業データ総量）

問 3-11（ビッグデータ該当）

行動属性を示す入力項目として以下の 5 つの属性項目を用いた。

問 2-3（担当者合計）

問 2-4（契約書）

問 3-14（データの利活用経験）

問 3-17（データイニシアティブ）

問 3-18（高度なデータの処理・解析）

出力として、全社レベルの成果項目である以下の 1 つの成果項目を用いた。

問 2-7（データ利活用成果）

すべての入力項目に関する産業毎の反応係数はプロット図とともに整理した。推定結果のプロット図は数が多いため、Appendix 2 に掲示した。表 6 は、それらの反応係数について、入力項目の設問毎に記述統計を整理したものである。

設問	平均	分散	平均−標準偏差	平均+標準偏差	最小	最大
問2−1:全社データ総量	0.890	0.313	0.330	1.449	−0.214	2.524
問2−2:データ利用率	0.428	0.035	0.240	0.616	0.139	0.835
問2−3:担当者合計	0.104	0.019	−0.034	0.243	−0.018	0.778
問2−4:契約書	0.476	0.079	0.195	0.758	−0.375	1.042
問3−10:事業データ総量	0.769	0.435	0.110	1.428	−1.500	2.000
問3−11:ビッグデータに該当	0.680	0.813	−0.221	1.582	−1.875	2.367
問3−14:データ利活用経験年	0.293	0.048	0.073	0.513	−0.208	0.907
問3−17:イニシアティブの割合/【自社】	−0.010	0.000	−0.026	0.005	−0.054	0.013
問3−18:高度なデータの処理・解析	1.125	1.487	−0.094	2.345	−2.000	2.853

表 6 問 2-7(データ利活用成果)に対する各属性の反応係数（産業毎）

これら 9 つの入力項目の反応係数の平均は設問 3-17（データイニシアティブ）を除いて、すべてプラスであった。設問 3-17 に関しても、平均の絶対値は小さく、大きくマイナスとは言えない（設問 3-17 の回答は 0〜100 までをとる）。これらの入力項目について、全産業で平均してみると、プラスの効果があることがわかる。

一方、産業毎に反応係数を見てみると、その符号（プラスかマイナスか）は、ばらついていることがわかる。ばらつきは反応係数の範囲を表す最小値と最大値で確認できる。例えば設問 2-1（全社データ総量）では、最小値はマイナス、最大値はプラスとなっている。つまり、反応係数が最小の産業（設問 2-1 の例では 29 通信/放送）では、全社データ総量を増加させると、データ利活用成果を減少させることになる。それとは反対に、反応係数が最大の産業（設問 2-1 の例で

は 8 総合化学）では、全社データ総量を増加させると、データ利活用成果を増加させることになる。この点は、前項で指摘したとおりである。

このように反応係数が最小値と最大値で異なる符号をもつ入力項目は、表 6 に基づくと、9 つの項目のうち、問 2-2（データ利用率）を除く、すべての項目に該当する。多くの入力項目に対して、その効果は産業毎に異なっており、マイナスからプラスまで分布していることがわかる。

最小値と最大値はやや極端な特徴値であるため、平均から 1 標準偏差を減じた値（平均-標準偏差）と平均に 1 標準偏差を加えた値（平均+標準偏差）で、再度、このような反応係数の効果のばらつきを確認する。最小値と最大値を比較したときよりも、反応係数の符号の逆転のケースは少なくなった。しかし、問 2-3（担当者合計）、問 3-11（ビッグデータ該当）、問 3-17（データイニシアティブ）、問 3-18（高度なデータの処理・解析）の 4 つの項目については、符号がマイナスからプラスへと逆転する。つまり、産業毎の反応係数（すなわち入力項目の効果）は、産業毎にばらついており、同じ入力項目でもマイナスの効果を持つ場合もあれば、プラスの効果を保つ場合もあると考えられる。

3.2. 反応係数のクラスター分析

前節では 9 つの属性について反応係数を算出した。複数の反応係数の産業毎のパターンを把握するために階層的クラスター分析を行う。

階層的クラスター分析は以下の手順で行った。

まず 9 つの属性について反応係数行列を作成し基準化を行った。その後、各属性間の距離をユークリッド距離に基づいて算出した。この距離行列に対してウォード法を用いて階層的クラスター分析を行った。その結果をデンドログラムで表示したものが図 13 である。クラスターの分割数については、Calinski-Harabasz 指数（CH 指数）をクラスター数毎に算出し、最も CH 指数が大きくなるクラスター数=2 を採用した。

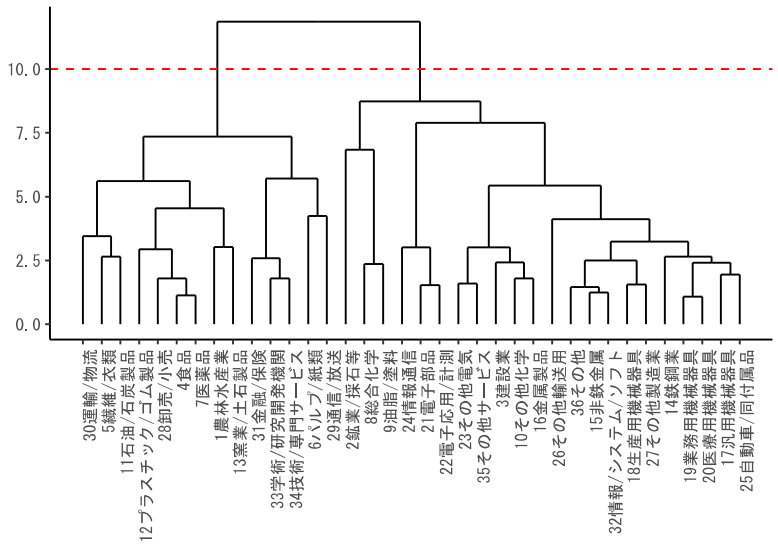

図 13 階層的クラスター分析の結果

階層的クラスター分析の結果、36 の産業は表 7 のように 2 つのグループに分類された。

業種ラベル	clustID	業種ラベル	clustID
1農林水産業	1	2鉱業/採石等	2
4食品	1	3建設業	2
5繊維/衣類	1	8総合化学	2
6パルプ/紙類	1	9油脂/塗料	2
7医薬品	1	10その他化学	2
11石油/石炭製品	1	14鉄鋼業	2
12プラスチック/ゴム製品	1	15非鉄金属	2
13窯業/土石製品	1	16金属製品	2
28卸売/小売	1	17汎用機械器具	2
29通信/放送	1	18生産用機械器具	2
30運輸/物流	1	19業務用機械器具	2
31金融/保険	1	20医療用機械器具	2
33学術/研究開発機関	1	21電子部品	2
34技術/専門サービス	1	22電子応用/計測	2
		23その他電気	2
		24情報通信	2
		25自動車/同付属品	2
		26その他輸送用	2
		27その他製造業	2
		32情報/システム/ソフト	2
		35その他サービス	2
		36その他	2

表 7 各クラスターに含まれる産業リスト

2 つのクラスターの傾向を把握するために、クラスター毎に所属する産業が持つ各反応係数の平均をまとめたものが表 8 である。

clustID	1	2
産業数	14	22
企業数	299	681
Q2-7:データ利活用成果	2.722	2.624
Q2-1:全社データ総量	0.533	1.117
Q2-2:データ利用率	0.373	0.461
Q2-3:担当者合計	0.120	0.095
Q2-4:契約書	0.410	0.510
Q3-10:事業データ総量	0.311	1.061
Q3-11:ビッグデータに該当	-0.139	1.202
Q3-14:データ利活用経験年	0.240	0.327
Q3-17:イニシアティブの割合/【自社】	-0.006	-0.013
Q3-18:高度なデータの処理・解析	-0.114	1.774

表 8 各クラスターの反応係数の平均

表 8 から 2 つのクラスターについて次のような傾向を読み取ることができる。

クラスター1 は 14 の産業が属しており、その企業数は 299 である。これに対してクラスター2 は 22 の産業の 681 企業が属している。成果項目の設問 2-7（データ利活用成果）について、クラスター1 は 2.722 であり、クラスター2 は 2.624 であった。2 つのクラスター間で成果について大

きな違いはない。

クラスター分析の対象となった設問 2-1（全社データ総量）から設問 3-18（高度なデータの処理・解析）を検討すると、設問 3-11（ビッグデータ該当）と設問 3-18（高度なデータの処理・解析）に大きな違いがあった。他の項目は両クラスターで似たような傾向であった。

設問 3-11（ビッグデータ該当）と設問 3-18（高度なデータの処理・解析）は、いずれも高度なデータ技術を積極的に用いるか否かという設問である。クラスター1 は、これらの入力に対して、反応係数がマイナスとなっている。逆に、クラスター2 は、これらの入力に対して、反応係数はプラスとなっている。

つまり、クラスター1 ではビッグデータやディープラーニングなどの高度なデータ技術・資源を増加させると、成果に対してマイナスの効果があると推定される。逆に、クラスター2 は、高度なデータ技術・資源を増加させると、成果に対してプラスの効果があると推定される。2 つのクラスターは高度なデータ技術や資源の入力に対して、対照的な反応をする産業群であると考えられる。

高度なデータ技術やデータ資源がデータ利活用成果につながることは、妥当な因果関係である。この点でクラスター2 の反応係数の傾向は納得できるものである。一方、クラスター1 の反応係数の傾向は、高度なデータ技術やデータ資源の増加が、成果を減少させるというものであり、直感に反するものである。この点について考察すると、(A)(B)の 2 つの可能性が考えられる。

（A）1 つめの可能性は、今回測定した成果指数が主観評価なものであるために、このような現象が生じたというものである。主観評価の場合、回答者の主観尺度で基準化されてしまうため、客観的なデータ利活用成果を反映していない可能性がある。データの利活用について、その成果が厳しく問われている産業では、客観的に同じようなデータ利活用成果が達成されていたとしても、その主観評価は厳しくなりやすい。その逆も当然ありえる。

例えば、前項の産業毎の平均値の比較の分析結果（図 7 と図 8）では、32 情報/システム/ソフトは複数の項目でデータ資源やデータ技術に対して積極的な行動をとっていることがわかった。逆に 13 窯業/土石製品は複数の項目でデータ資源やデータ技術に対して消極的な行動をとっていることがわかった。一方、それら産業のデータ利活用成果については、図 14 で設問 2-7（データ利活用成果）の平均を産業毎比較すると、32 情報/システム/ソフトと 13 窯業/土石製品の間でほとんど差がないことがわかる。

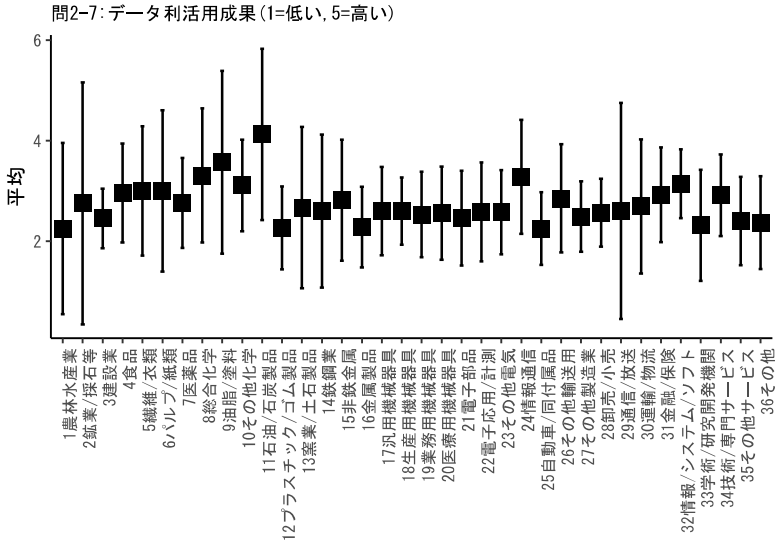

図 14 問 2-7（データ利活用成果）の平均（産業毎）

主観的評価ではなく、客観的評価を導入することで、表 8 のクラスター1 で観察されるような、データ資源やデータ技術の反応係数がマイナスにもかかわらず、データ利活用成果はクラスター2 と比べて高い、というような直感に反する推定結果が改善されるかもしれない。

（B）2 つめの可能性は、成果項目が主観的評価であるにせよ、今回の推定は現象を正しく反映しているというものである。すなわち、表 8 に示されるような反応係数の傾向は正しく、何らかの理由で成果項目である設問 2-7（データ利活用成果）の得点にまで影響していないとの解釈である。影響していない可能性として①②が考えられる。

①レイテンシー（時間遅延）の影響

データ資源やデータ技術への投資は、ある程度の大きさの成果になるまで、時間が必要である、との考えである。

$$Y_s = \beta_0 + \beta_1 X \qquad \cdots(2)$$

反応係数はモデル式(2)で見るように、出力をβ_0と$\beta_1 X$に分解する。β_1は入力に反応して出力を増減させるものであるのに対して、β_0は入力とは独立している潜在的な出力である。つまり、β_0は当該産業がもっている、入力とは独立した潜在的データ利活用成果ということになる。β_0はデータ利活用成果の既存資産と言い換えてもよいだろう。

データ利活用の既存資産（β_0）が大きい産業では、直近のデータ資源やデータ技術への投資は相対的に小さくなる。このため、データ資源やデータ技術の反応係数(β_1)の傾向が、そのまま、最終的なデータ利活用成果に結びついていない、という現象が起こる。

もし、この見解にたてば、現在のデータ利活用に関する所見は次のようになる。データ利活用に関して、表面的にはデータ利活用成果はどの産業でも違いはなく、ほぼ同じようなものである。しかし、水面下ではデータ資源やデータ技術に投資することで、データ利活用成果を積みましている産業群が存在する。時間を経るにつれ、そのような優秀な産業群と劣後する産業群の差は、

明確になってくる、との所見である。

②阻害要因の影響

データ資源やデータ技術への投資が、直接的にデータ利活用成果に結びつかないような何らかの障害が存在している可能性である。クラスター2 は、データ資源やデータ技術に関する反応係数がプラスの産業群である。これら産業では、最終的なデータ利活用成果が高くても良いはずである。しかし、それらの効果を打ち消すようなメカニズムが、「データ資源やデータ技術への投資」と「最終的なデータ利活用成果」の間に存在している可能性である。このような障害は、例えば、いわゆる産業に存在する商習慣や規制かもしれないし、組織的な要因かもしれない。

①と②のケースでは、このような反応係数と最終成果指標とのギャップの問題に対する処方箋が異なる。①のケースであれば、時間を経ることによって、ギャップは自然と解消される。投資に対して、優遇税制を施行したり、補助金を拠出したりすることによって、ギャップ解消の時間は短縮される。しかし、②のケースの場合は、産業内もしくは組織内に存在する障害を取り除かない限り、ギャップの問題は解決しない。

4. まとめ

近年のデータ技術への期待は大きく、停滞している日本産業の生産性を広範囲に向上させるものと考えられている。今回の調査では、狭義の製造業にとどまらない、広い範囲の産業でデータ資源やデータ技術への取組みが見られた。

産業毎の平均値の比較（第２章）からは、多くの産業で似通ったものであることがわかった。すなわち、データ資源やデータ技術に関して、特殊な産業が存在しているわけではないことが示唆された。また、特にデータ利活用の成果についても、特別に優越している産業が存在しているわけでないことがわかった。この点は、昨今の報道にみるような IoT、ビッグデータ、人工知能のバラ色のイメージとはやや異なるものであった。

しかしながら、産業毎の反応係数の分析（第３章）からは、反応係数に 2 つのパターンが生じていることがわかった。反応係数は、データ資源やデータ技術を入力とし、データ利活用成果を出力としたときの効果（変換効率もしくは生産性）と考えることができる。１つめのパターンは高度なデータ資源やデータ技術に関して反応係数がマイナスの産業群であり、2 つめのパターンは反応係数がプラスの産業群である。

データ資源やデータ技術の反応係数がプラスの産業群は、最終的なデータ利活用成果が高いと考えられる。しかし、今回の分析からは、そのようなエビデンスは見いだせなかった。むしろ分析結果から、「データ資源やデータ技術の反応係数がマイナスもしくはプラスである」、ということと、「最終的なデータ利活用成果への影響」の間に乖離が存在する、と推察される。

今回の分析で判明した、『データ資源やデータ技術への投資が最終的なデータ利活用成果に結びつかない』もしくは『「データ資源やデータ技術への投資」と「最終的なデータ利活用成果」にギャップがある』という問題は、いわゆる 1980 年代のソローの生産性パラドックスを想起させる。当然、今回のギャップ問題がソロー・パラドックスと同一のものであるのかについてはより慎重な分析が必要である。

今回の分析結果は今後より精査が必要であろう。しかし、第 3 章の考察で議論したように、判明したギャップ問題が、単なる計測や分析上の問題でない場合、その影響範囲が大きいことを考えると、適切な産業施策や企業での対応が求められることになるだろう。この点は、今後の研究でも引き続き注視されるべき点である。

なお、本研究は探索的な研究であったため、あまり仮定を置かない手法を用いた（多重比較法など）。しかし、より詳細な検討のためには、統御変数などを含んだ回帰モデルで効果を推定したり、階層モデルなどによって分散成分を調整したりするなどの必要があるだろう。この点は今後の課題である。

また、企業が扱うビッグデータは、多くの場合、個人情報であることが多い。この点について、本研究では産業毎の違いの観点から分析できていない。消費者に近い産業と、消費者から遠い産業では、違いが生じると考えるのが自然である。この点についても、今後の分析の課題である。

本稿が今後のデータ技術の産業利用促進の一助になれば幸いである。

5. Appendix 1

Appendix 1 では、属性項目（データ属性と行動属性）と成果項目について、大まかな産業毎の特徴を把握するために、全産業と産業毎のヒストグラムを掲げる。

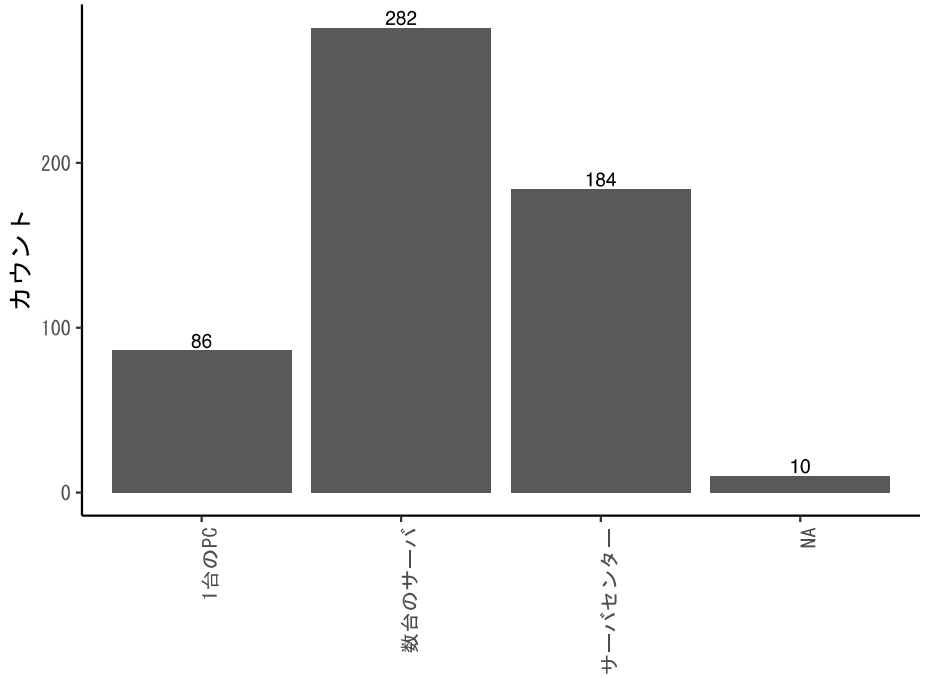

図 15 問 2-1:全社データ総量(全産業)

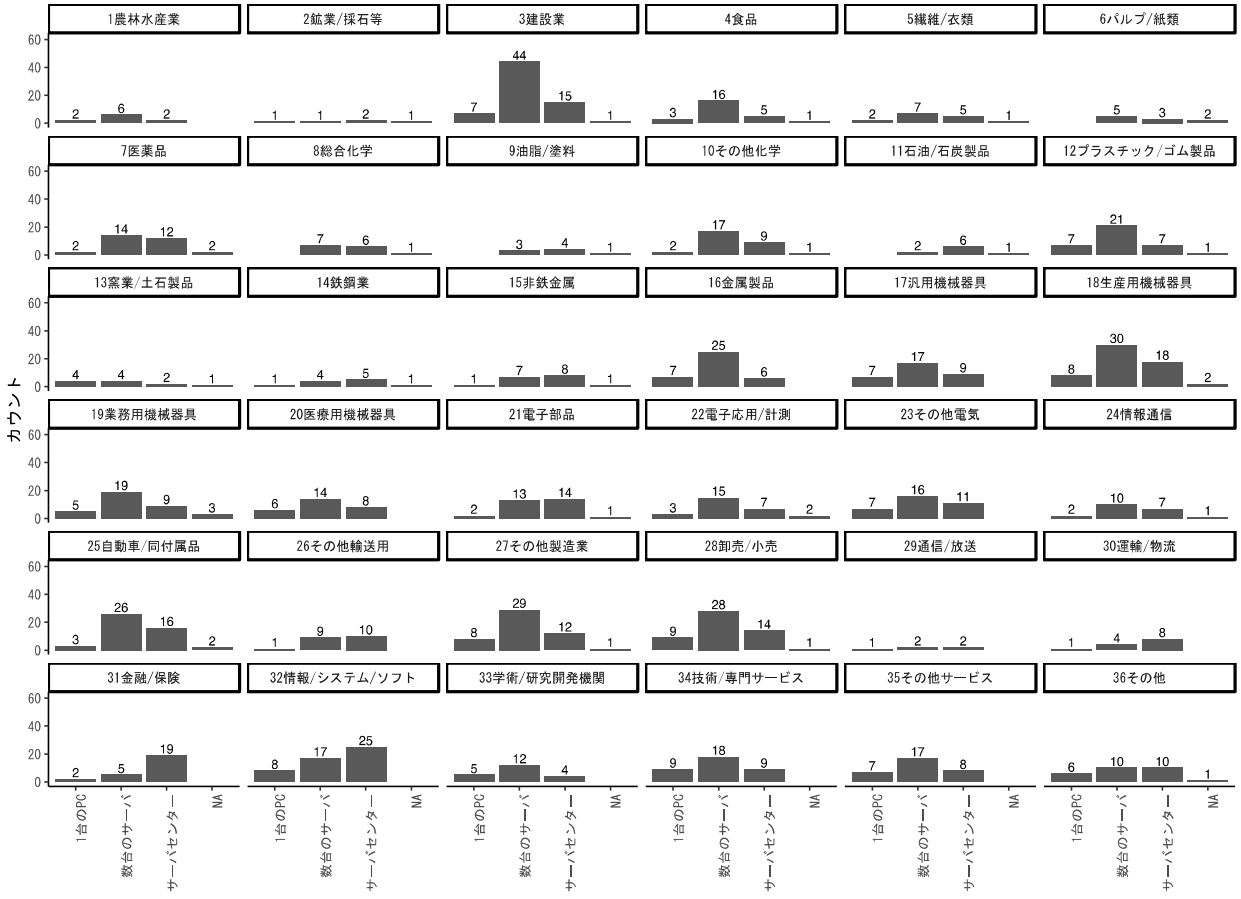

図 16 問 2-1:全社データ総量(産業毎)

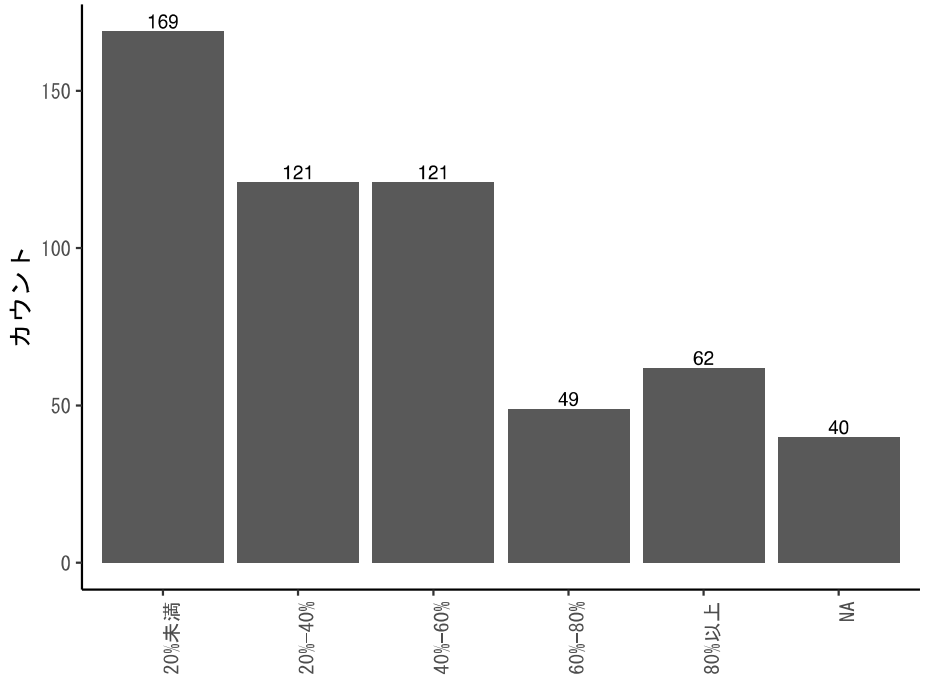

図 17 問 2-2:データ利活用率(全産業)

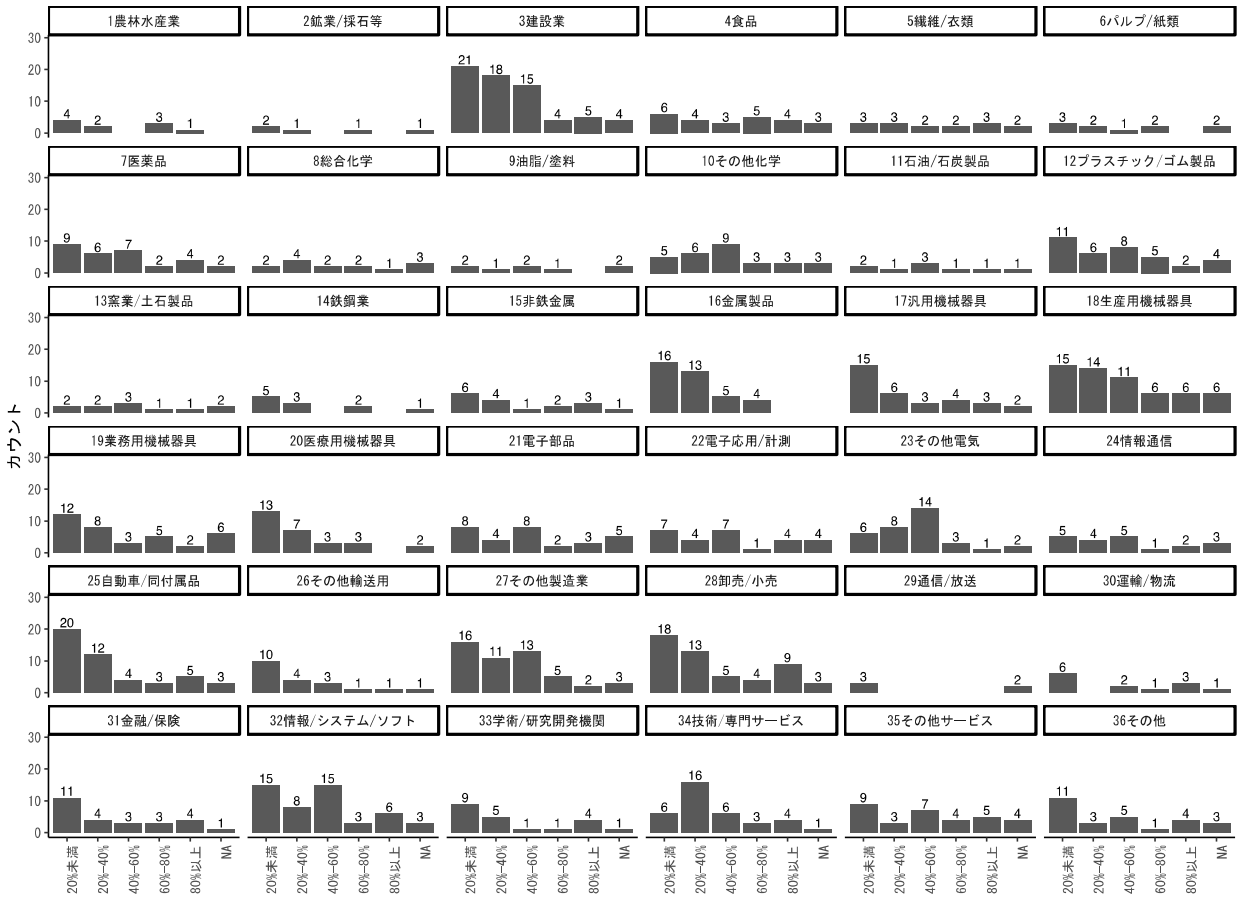

図 18 問 2-2:データ利活用率(産業毎)

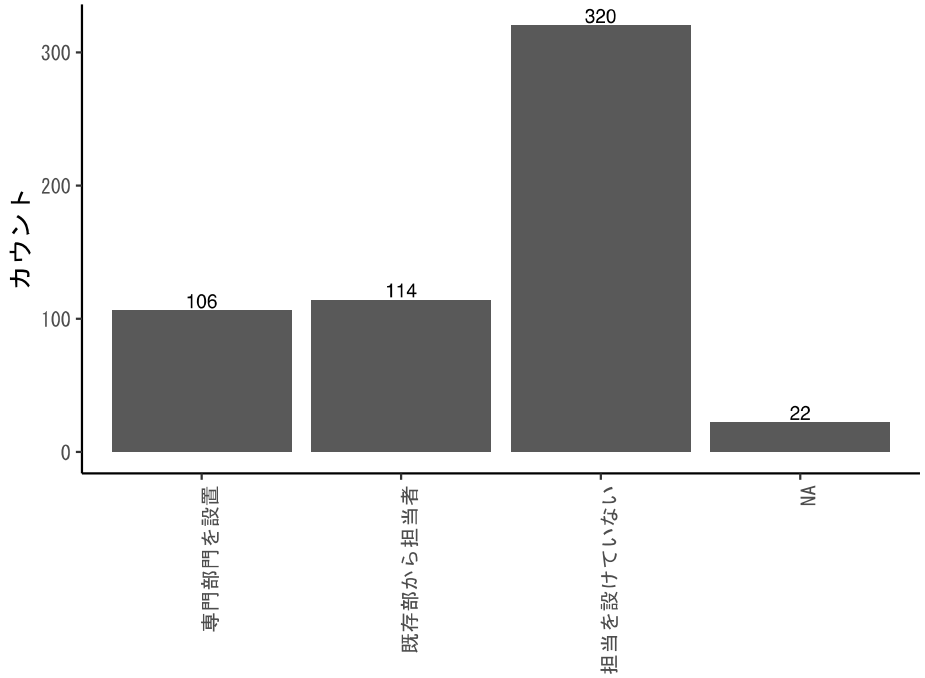

図 19 問 2-3:全社的なデータ利活用を推進する専門部門や担当者(全産業)

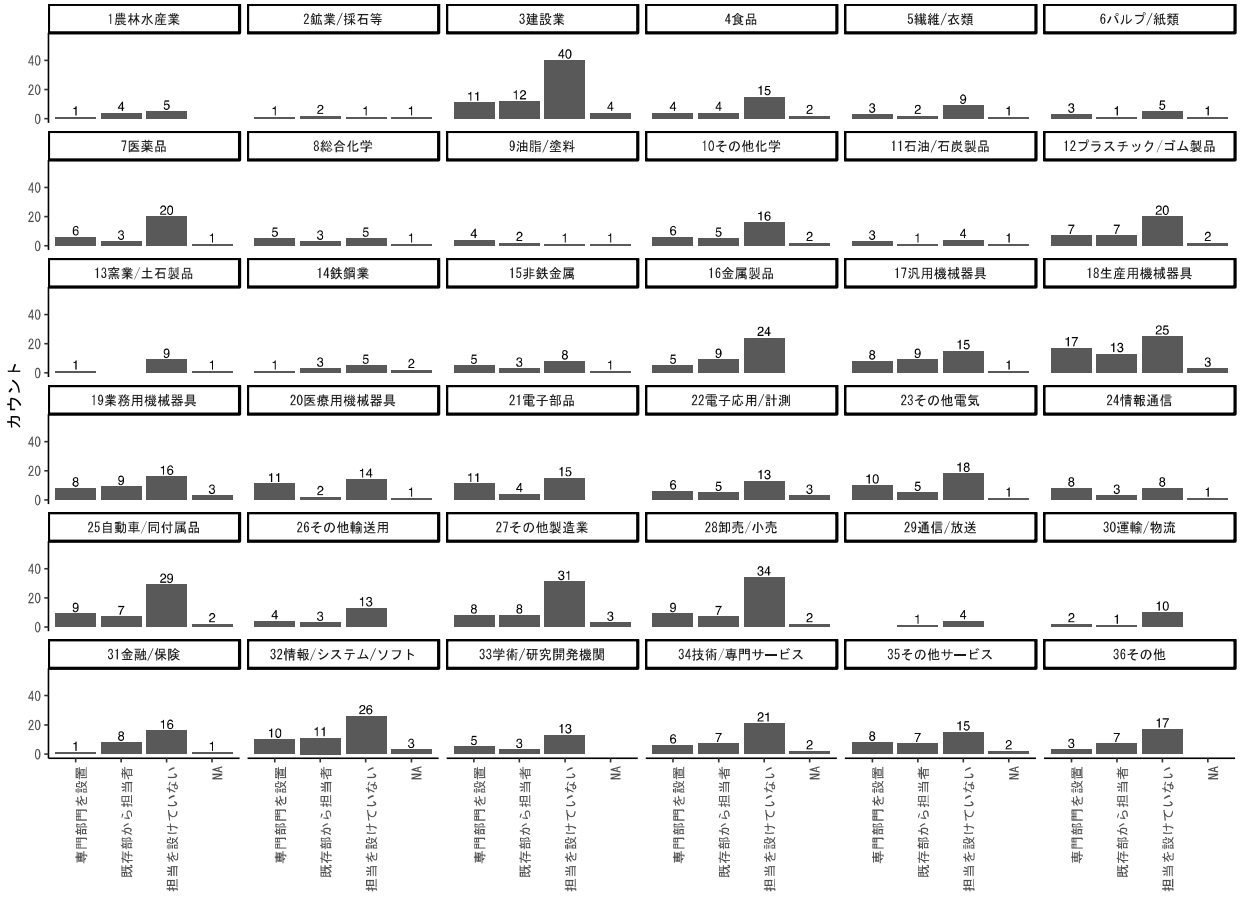

図 20 問 2-3:全社的なデータ利活用を推進する専門部門や担当者(産業毎)

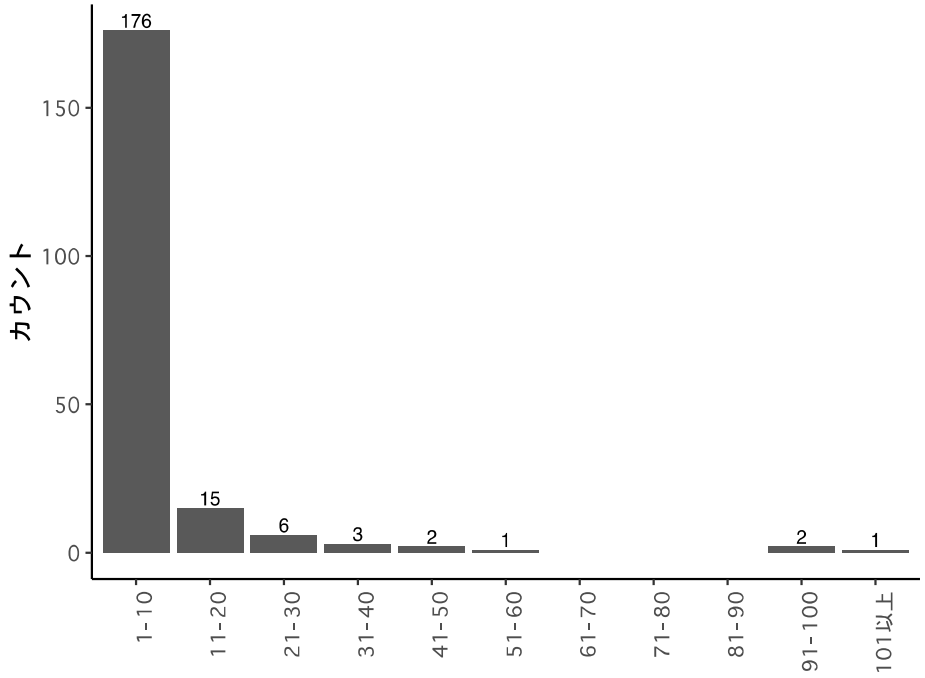

図 21 問 2-3:専門部門と既存部門の担当者数合計(全産業)

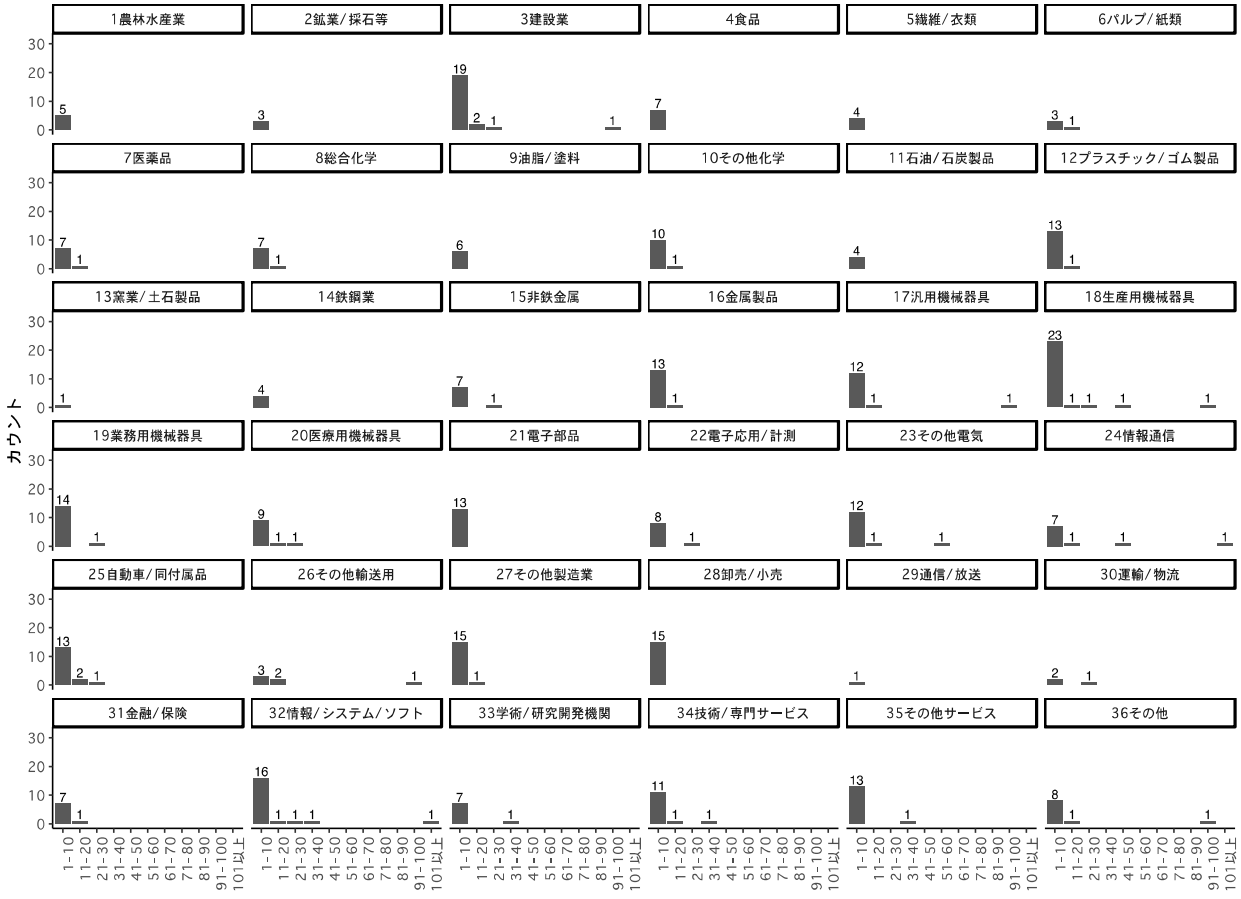

図 22 問 2-3:専門部門と既存部門の担当者数合計 (産業毎)

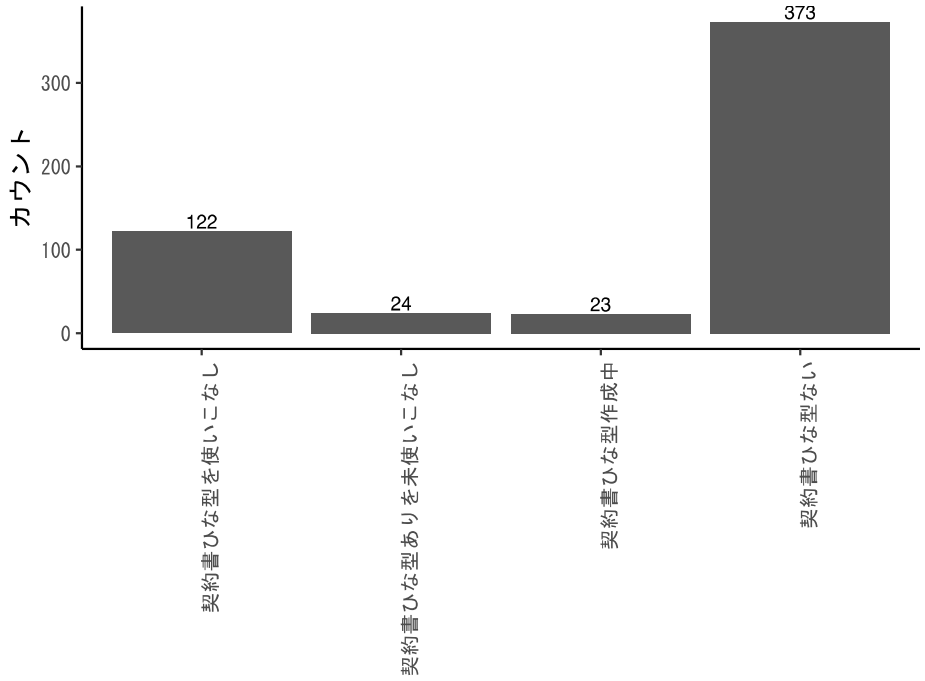

図 23 問 2-4:契約書（全産業）

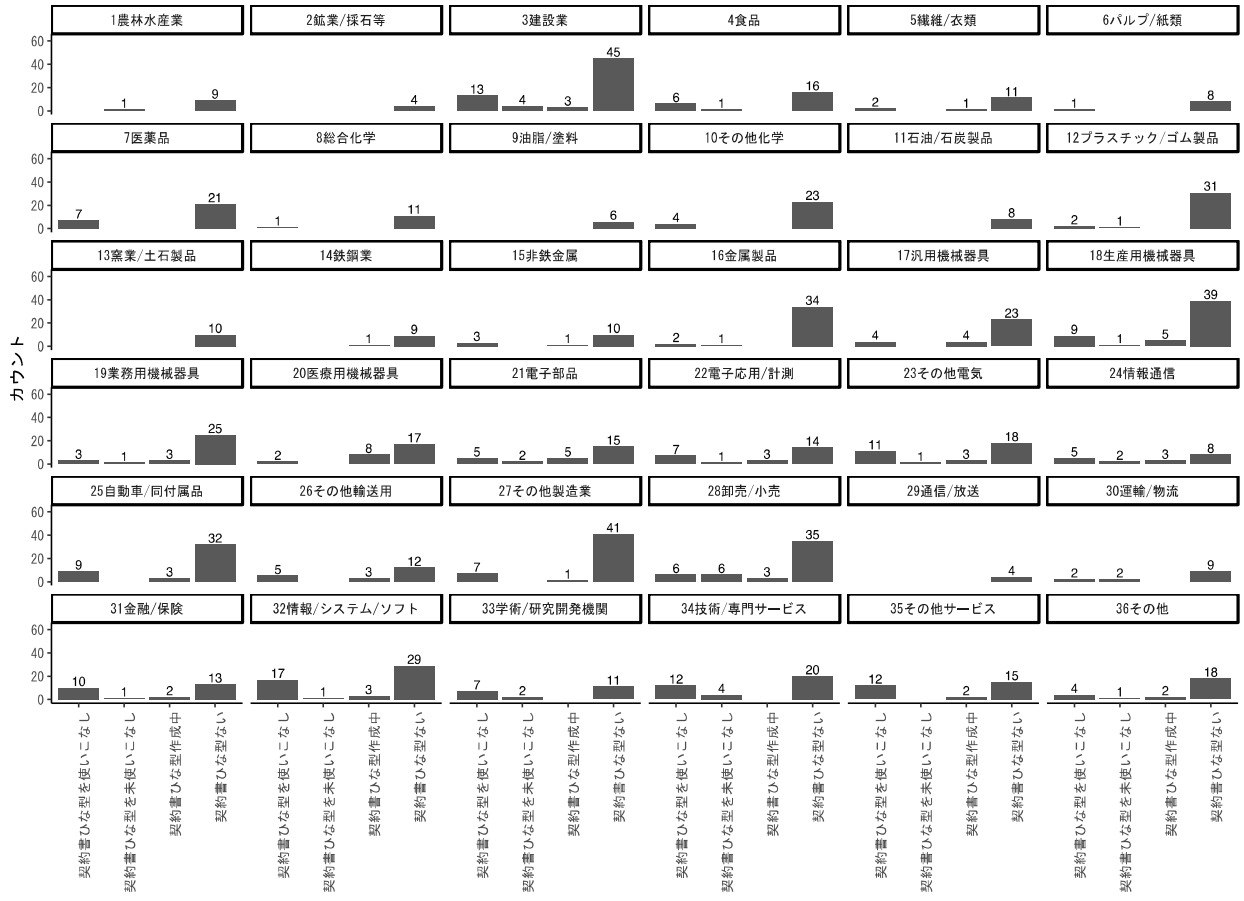

図 24 問 2-4:契約書（産業毎）

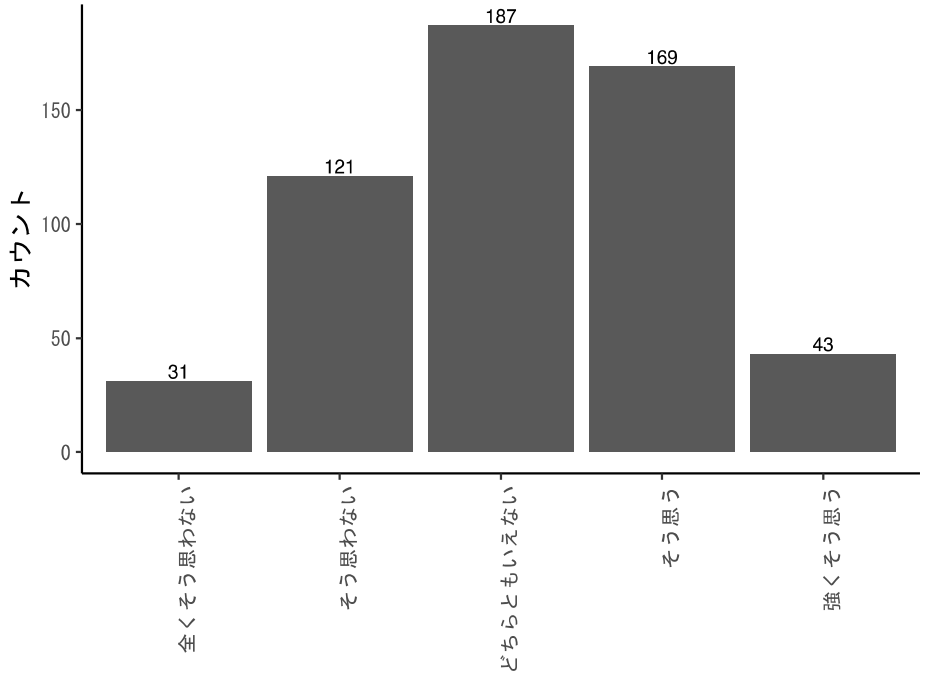

図 25 問 2-5:①データ利活用を積極的に推進している(全産業)

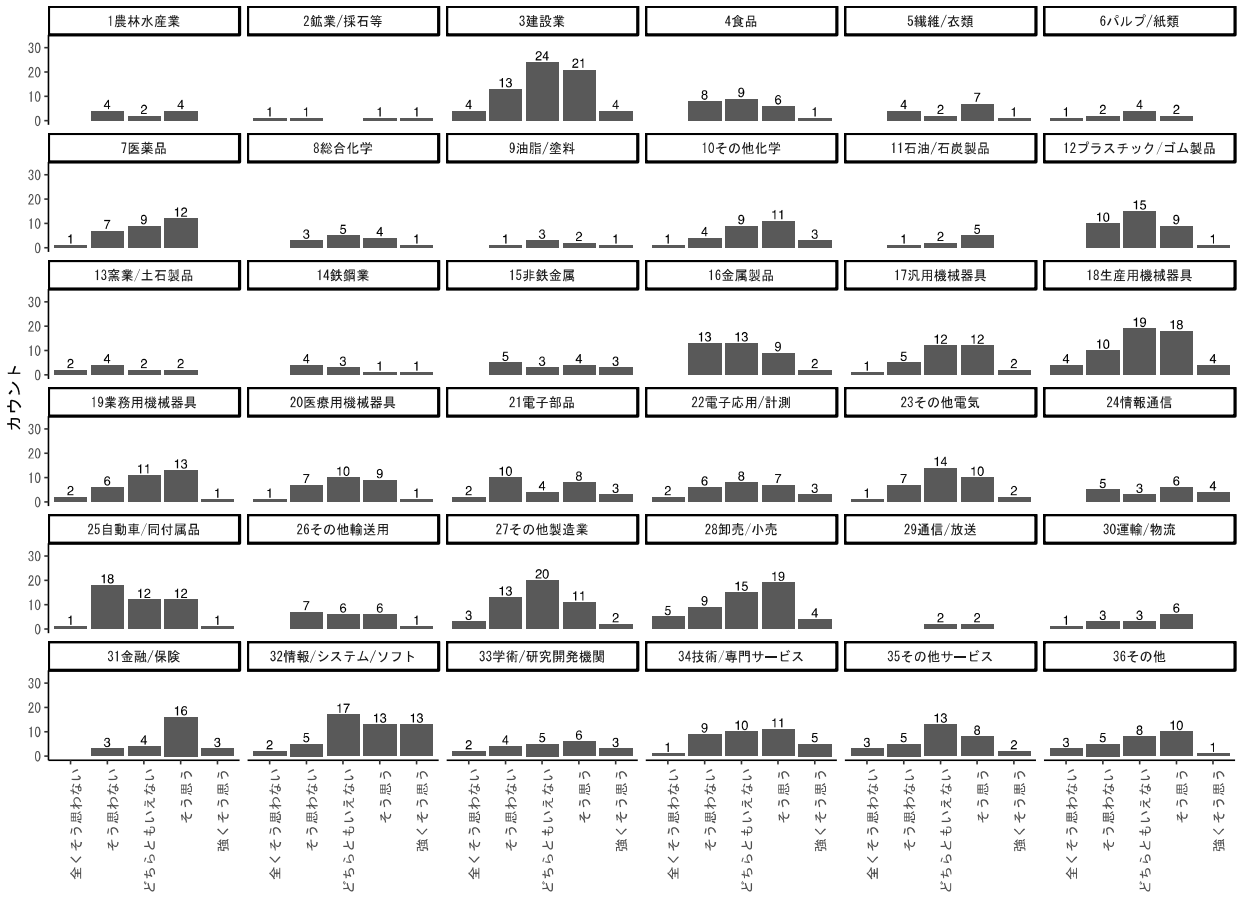

図 26 問 2-5:①データ利活用を積極的に推進している(産業毎)

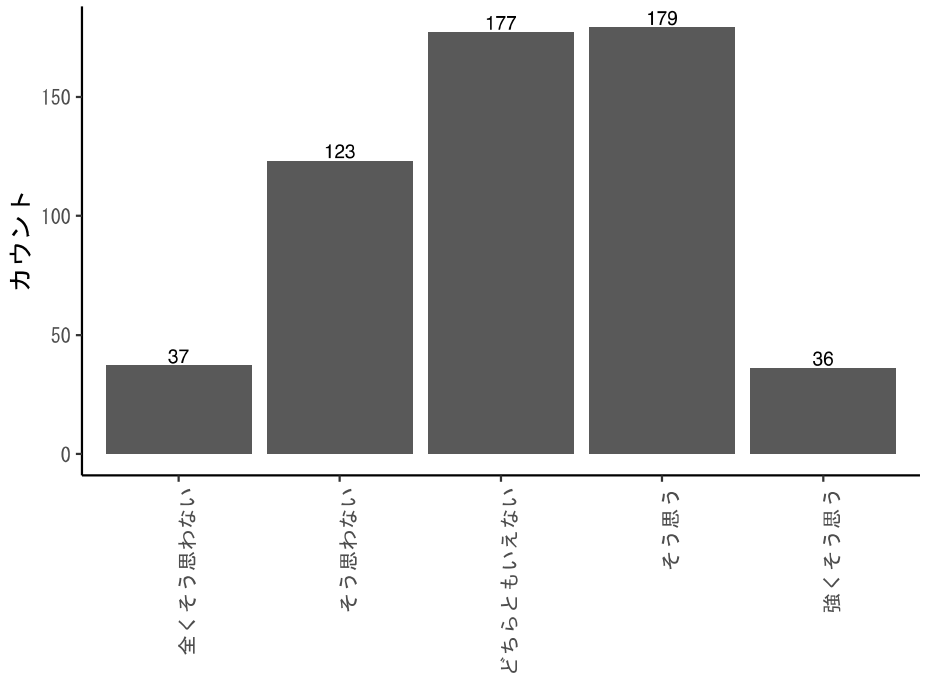

図 27 問 2-5:②社内全体で連携してデータ利活用を行うことを積極的に推進している(全産業)

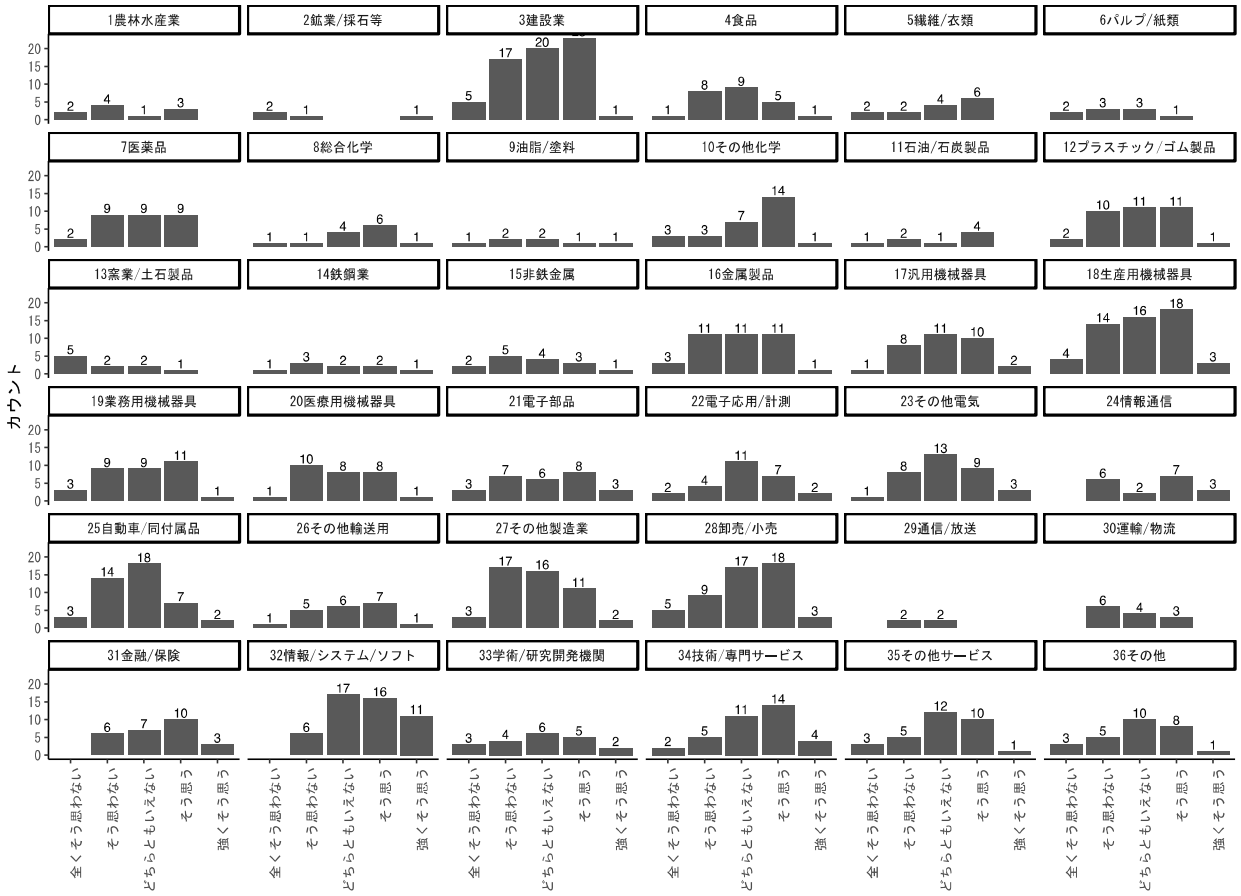

図 28 問 2-5:②社内全体で連携してデータ利活用を行うことを積極的に推進している(産業毎)

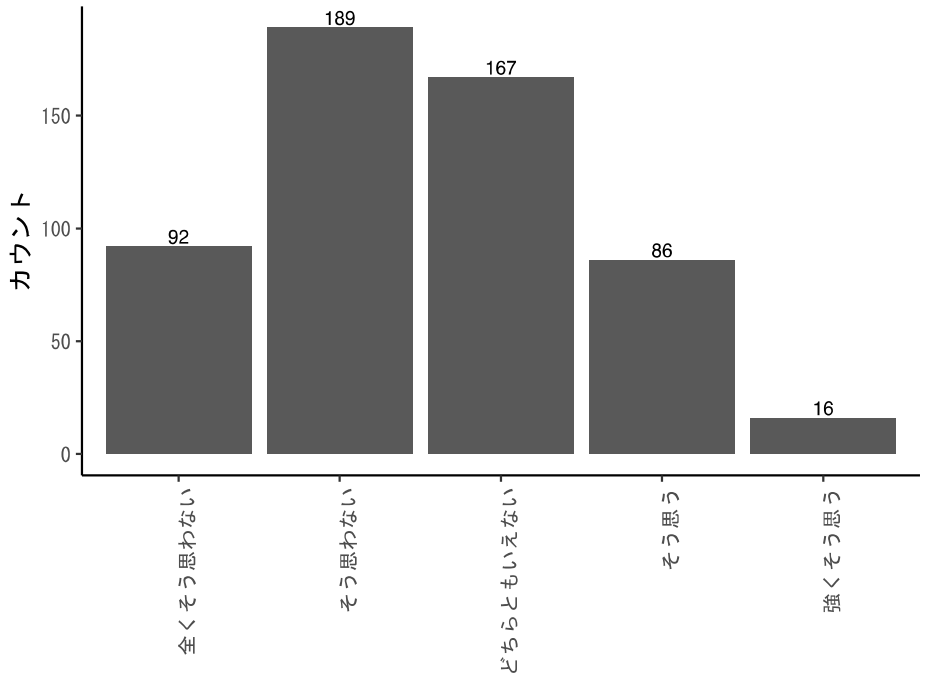

図 29 問 2-5:③社外の組織と連携してデータ利活用を行うことを積極的に推進している(全産業)

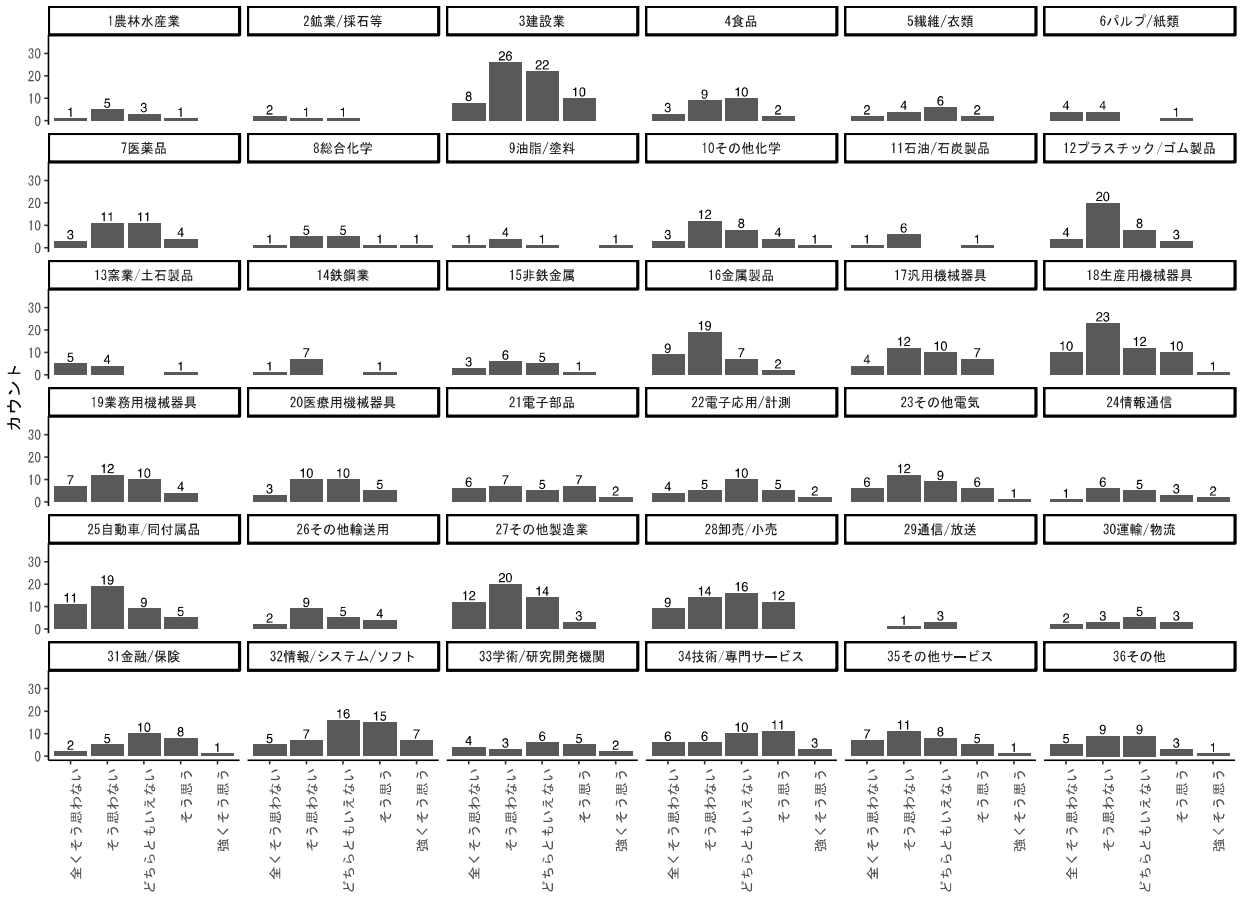

図 30 問 2-5:③社外の組織と連携してデータ利活用を行うことを積極的に推進している(産業毎)

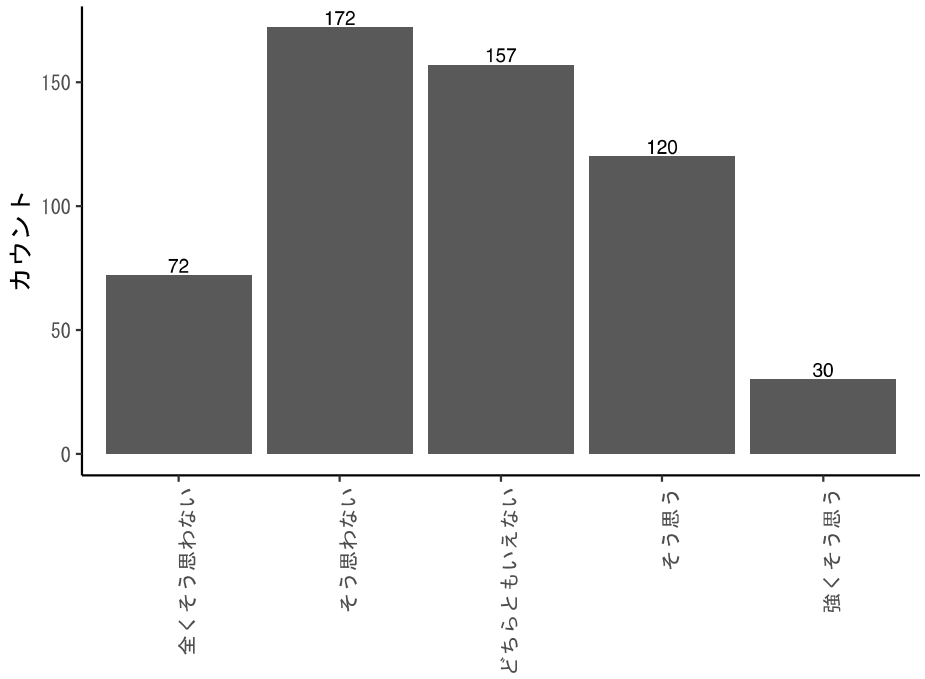

図 31 問 2-5:④個人情報に該当するデータの利活用を行える体制が整備されている(全産業)

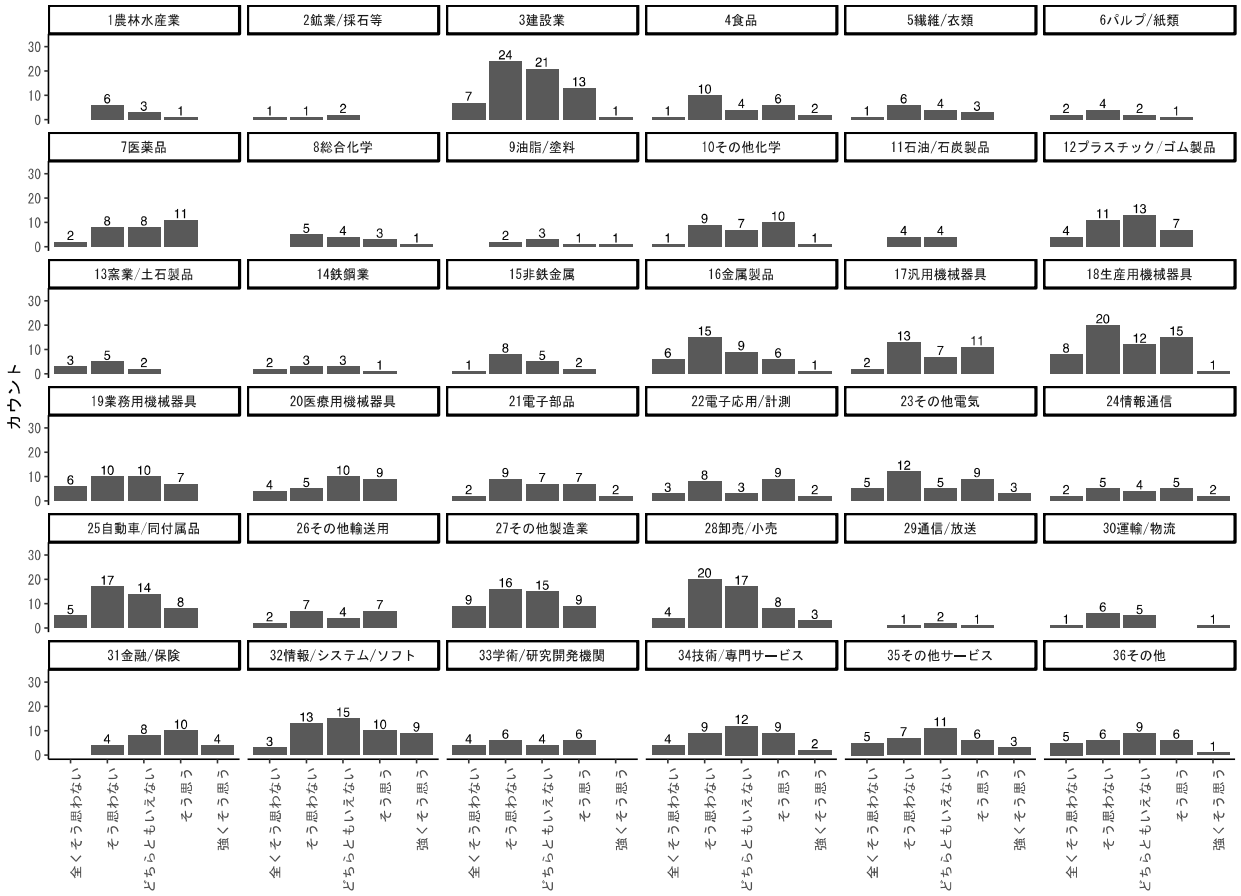

図 32 問 2-5:④個人情報に該当するデータの利活用を行える体制が整備されている(産業毎)

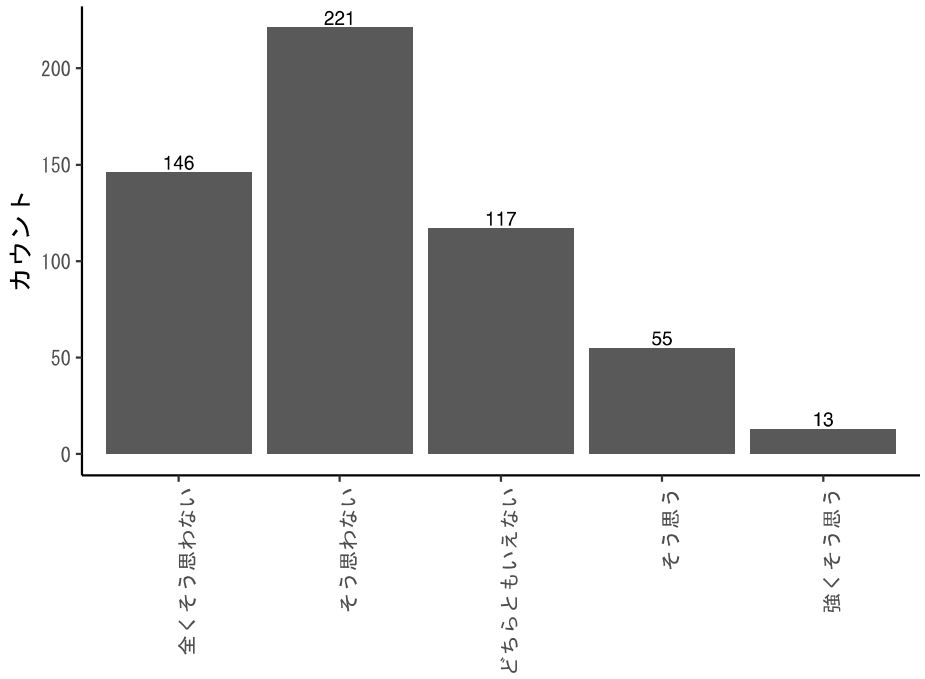

図 33 問 2-5:⑤ビッグデータの利活用を行える体制が整備されている(全産業)

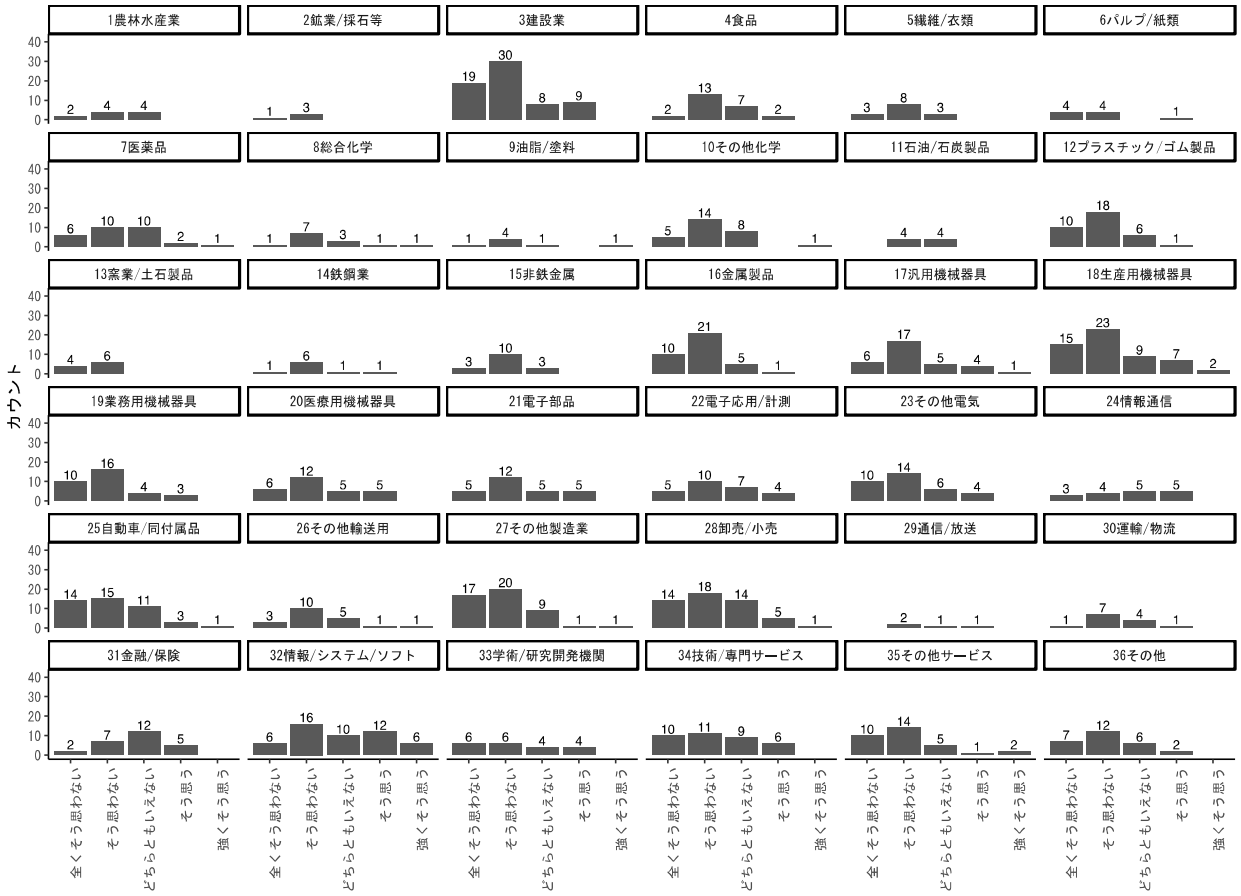

図 34 問 2-5:⑤ビッグデータの利活用を行える体制が整備されている(産業毎)

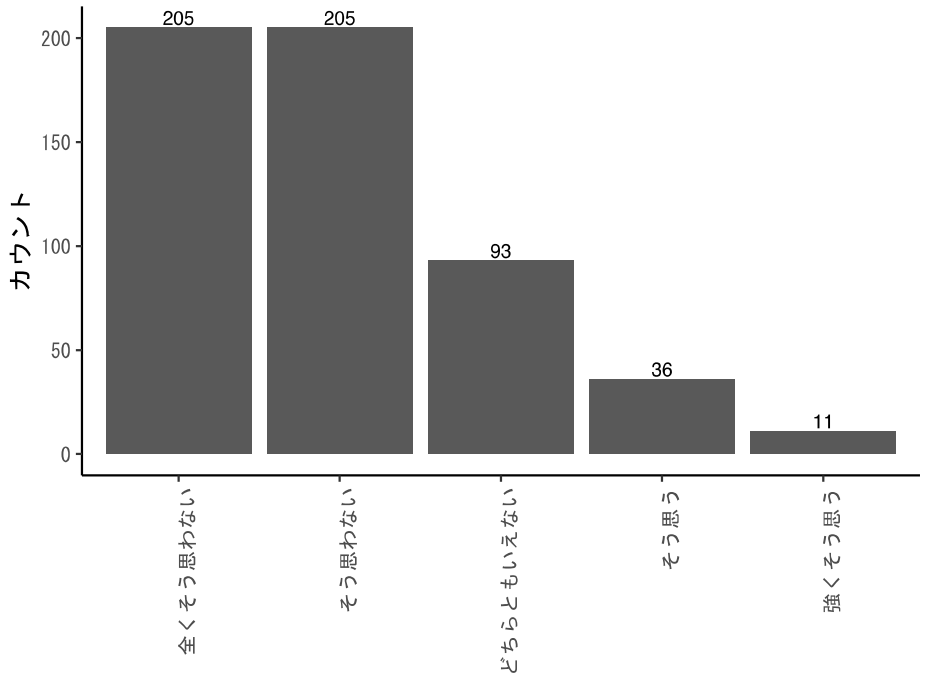

図 35 問 2-5:⑥ディープラーニング等の高度なデータの処理・解析を行える体制が整備されている(全産業)

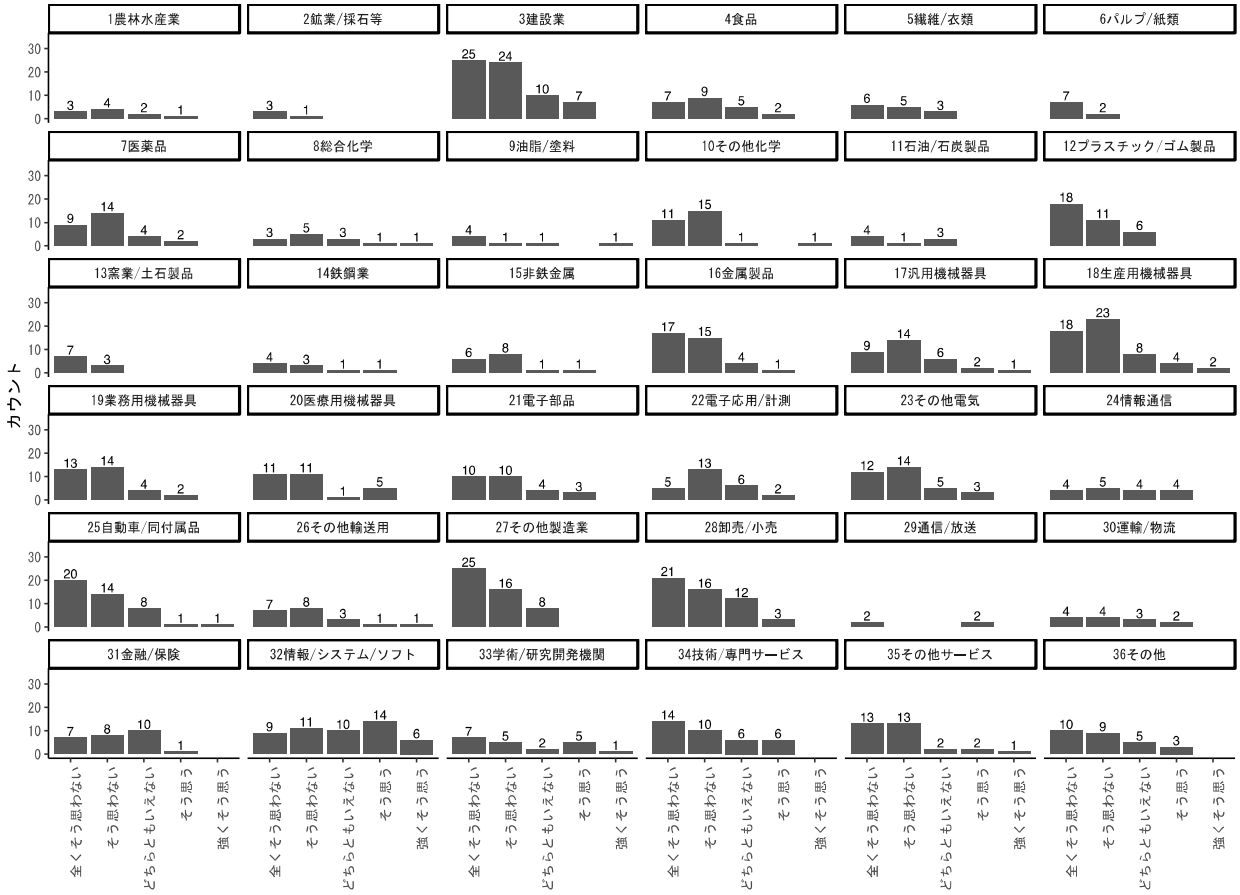

図 36 問 2-5:⑥ディープラーニング等の高度なデータの処理・解析を行える体制が整備されている(産業毎)

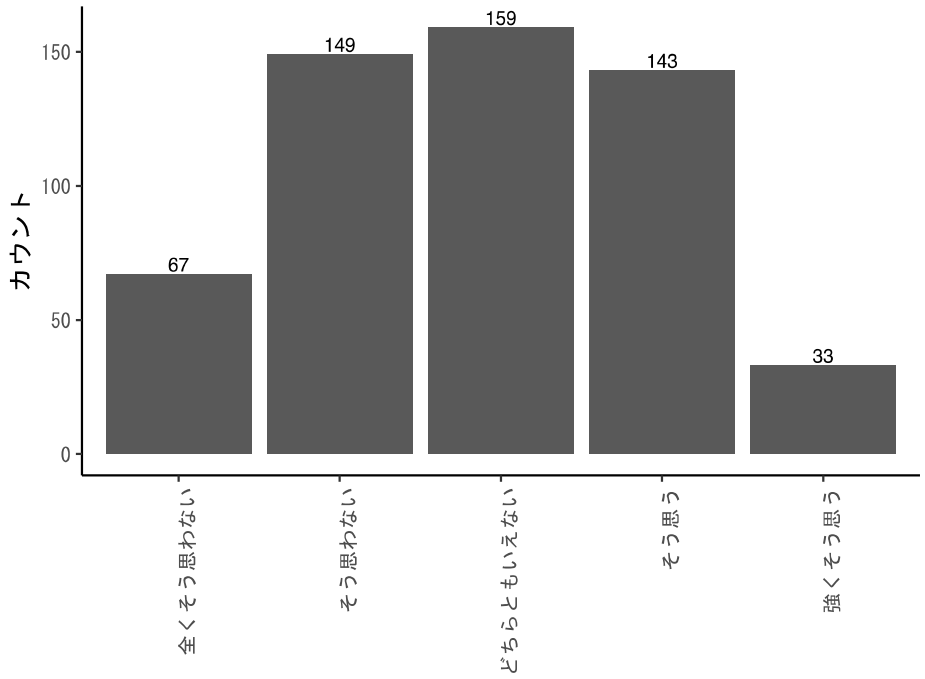

図 37 問 2-5:⑦データのアクセス権限を共通化する等、社内全体で連携してデータ利活用を行う体制が整備されている(全産業)

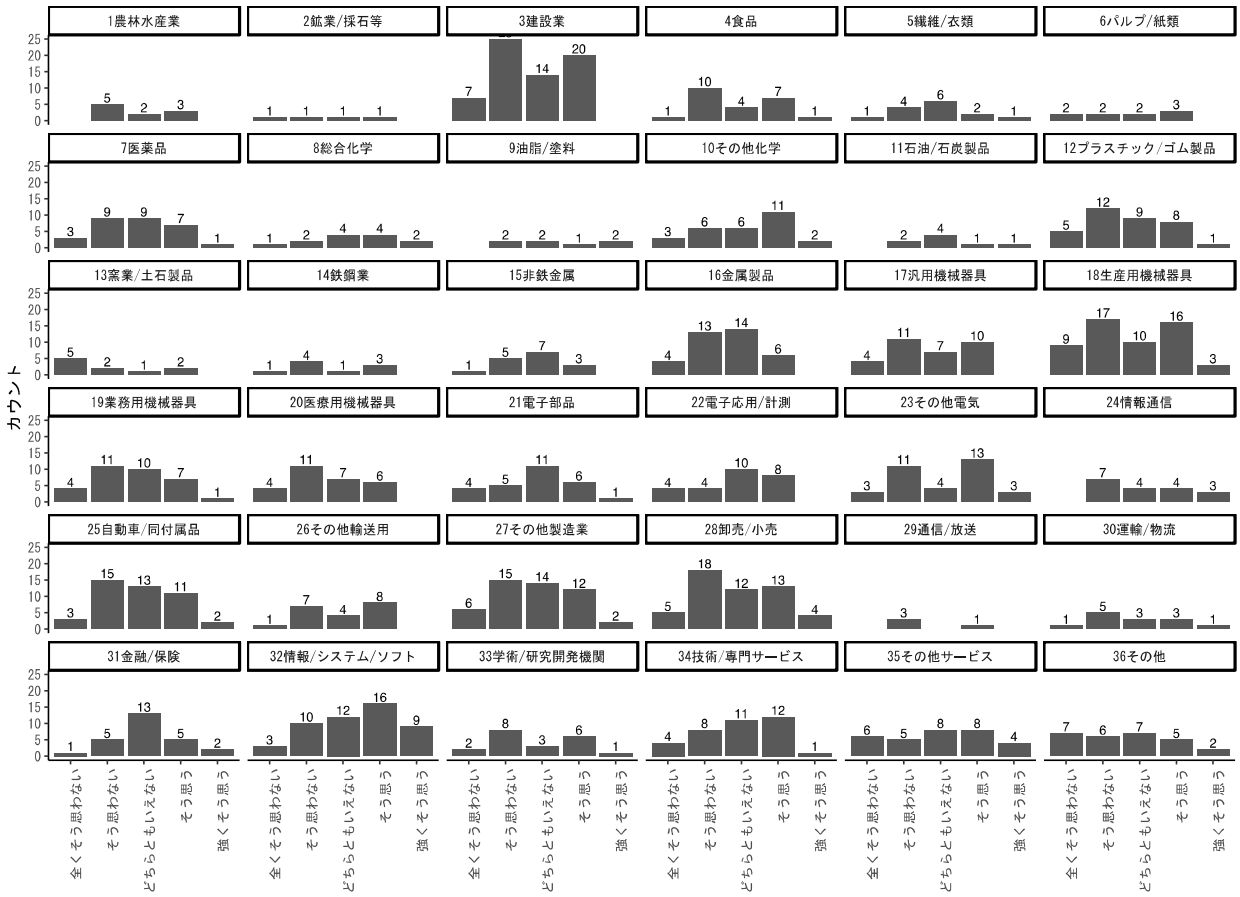

図 38 問 2-5:⑦データのアクセス権限を共通化する等、社内全体で連携してデータ利活用を行う体制が整備されている(産業毎)

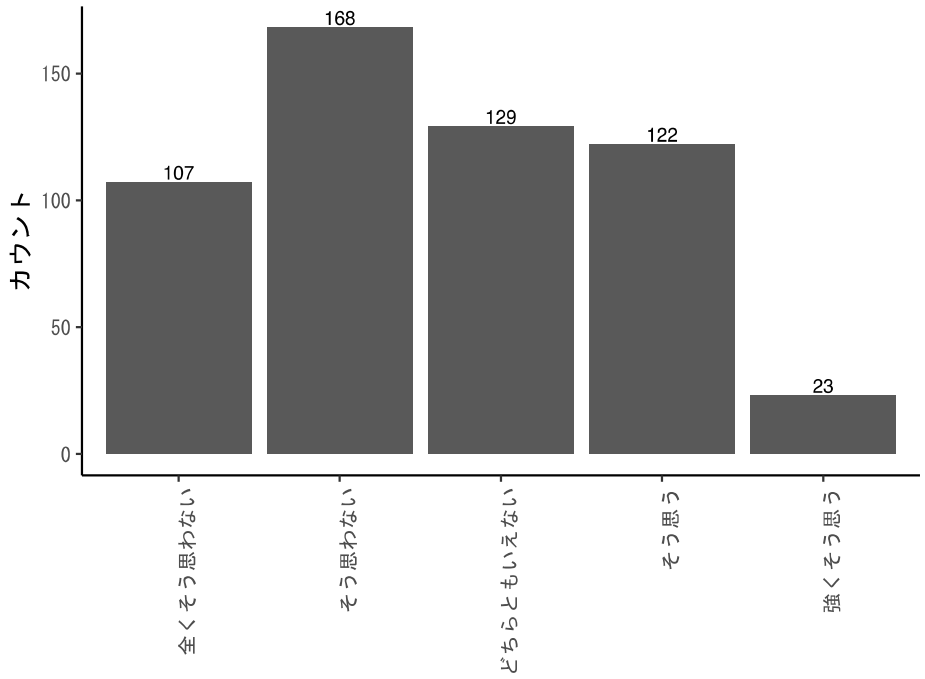

図 39 問 2-5:⑧データのアクセスを権限に応じて制限する等、社外の組織と連携してデータ利活用を行う体制が整備されている(全産業)

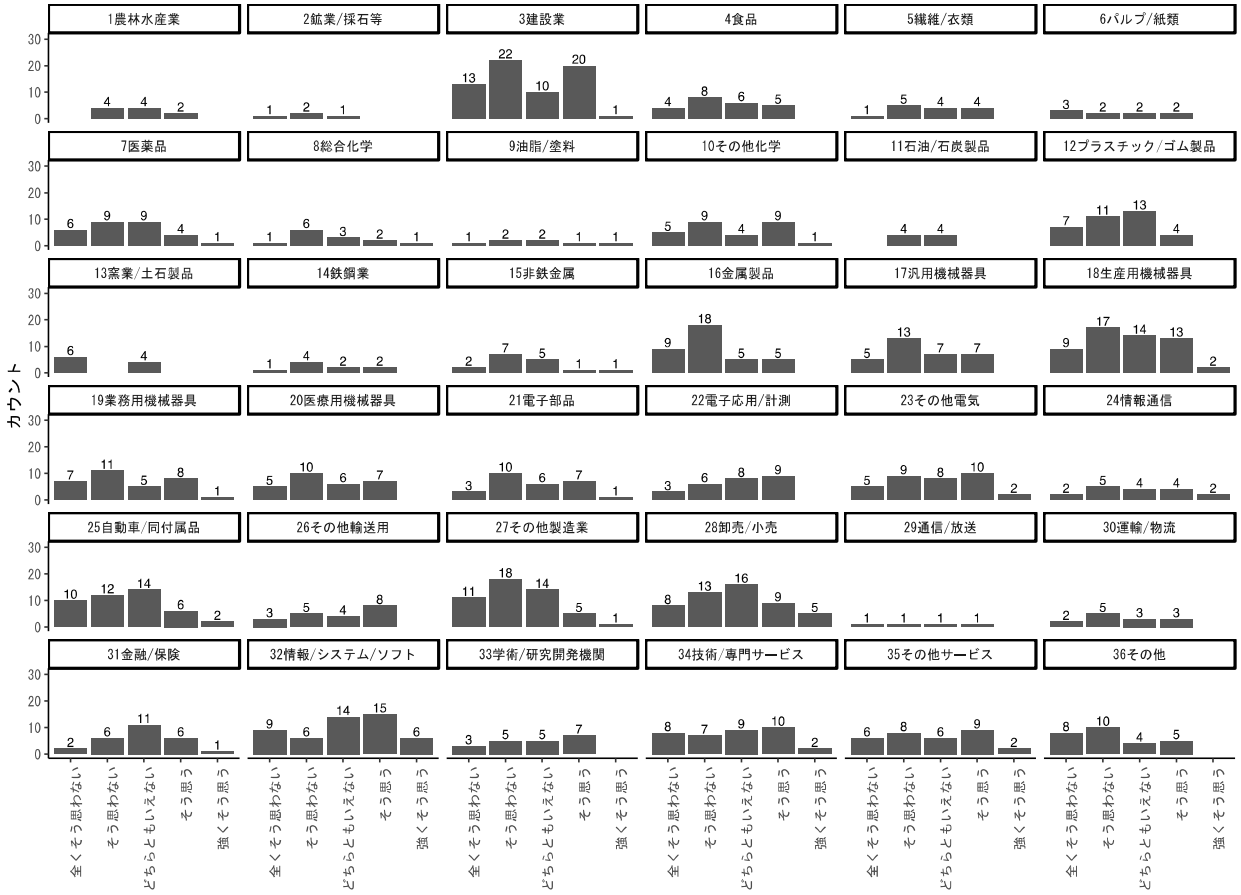

図 40 問 2-5:⑧データのアクセスを権限に応じて制限する等、社外の組織と連携してデータ利活用を行う体制が整備されている(産業毎)

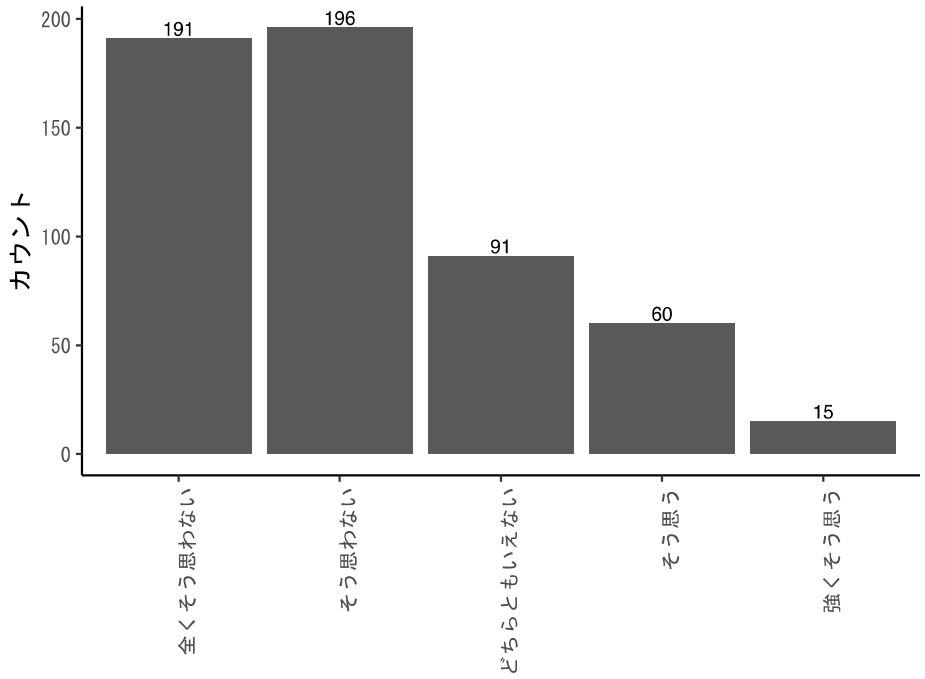

図 41 問 2-5:⑨データサイエンティスト等、高度なデータの処理・分析を行える人材を育成、雇用している(全産業)

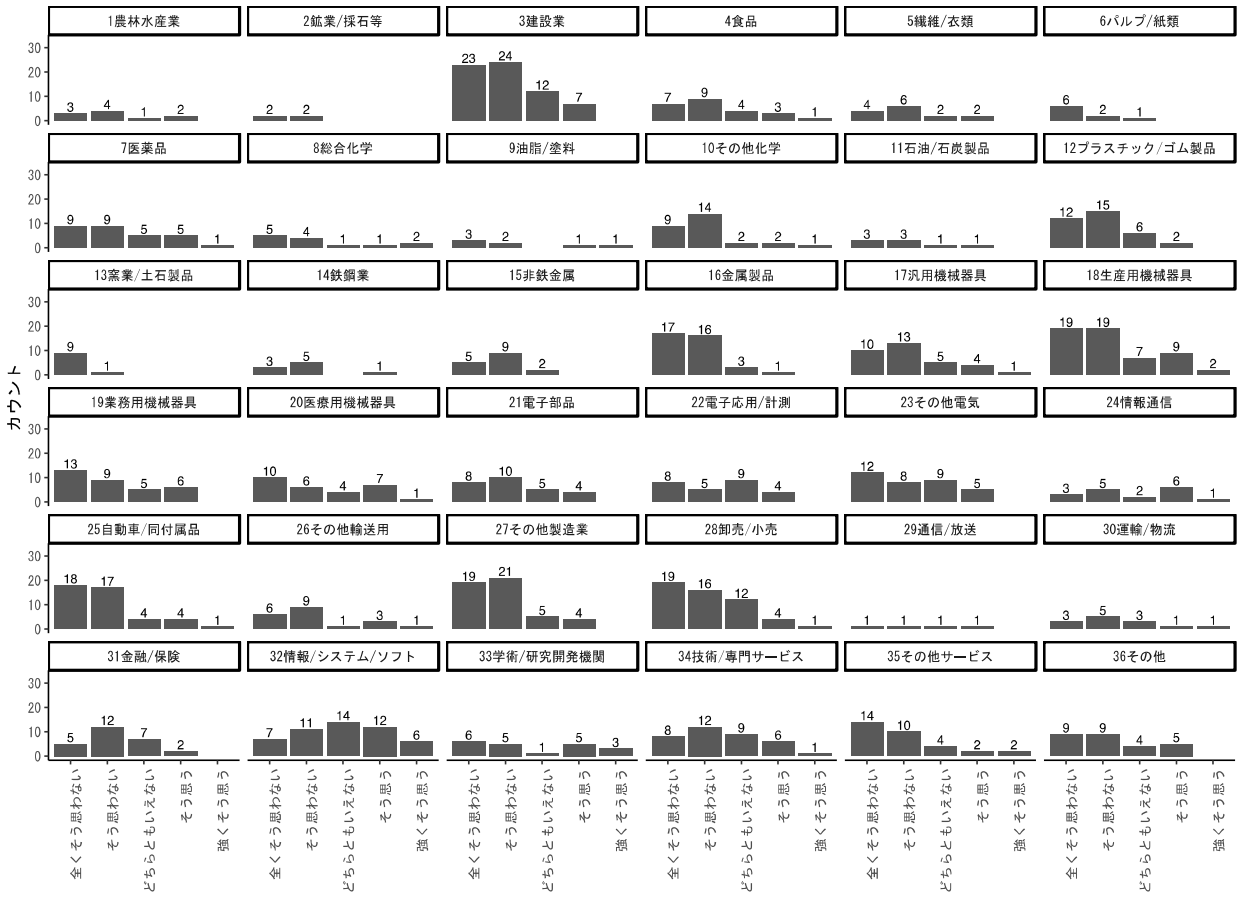

図 42 問 2-5:⑨データサイエンティスト等、高度なデータの処理・分析を行える人材を育成、雇用している(産業毎)

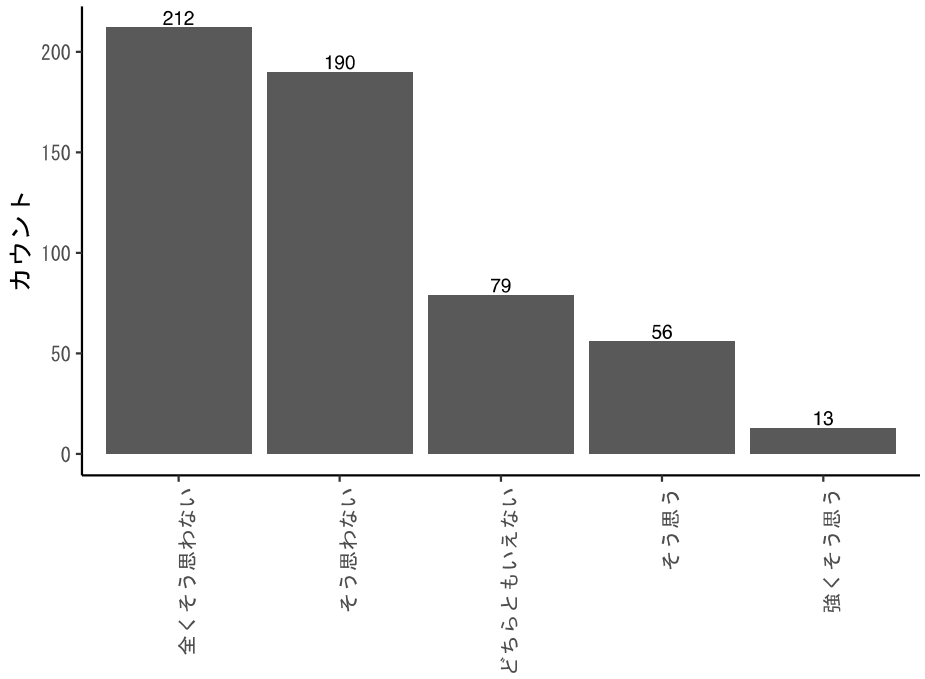

図 43 問 2-5:⑩ディープラーニング等の高度なデータの処理・解析結果を理解し、事業活動に活かせる人材を育成、雇用している(全産業)

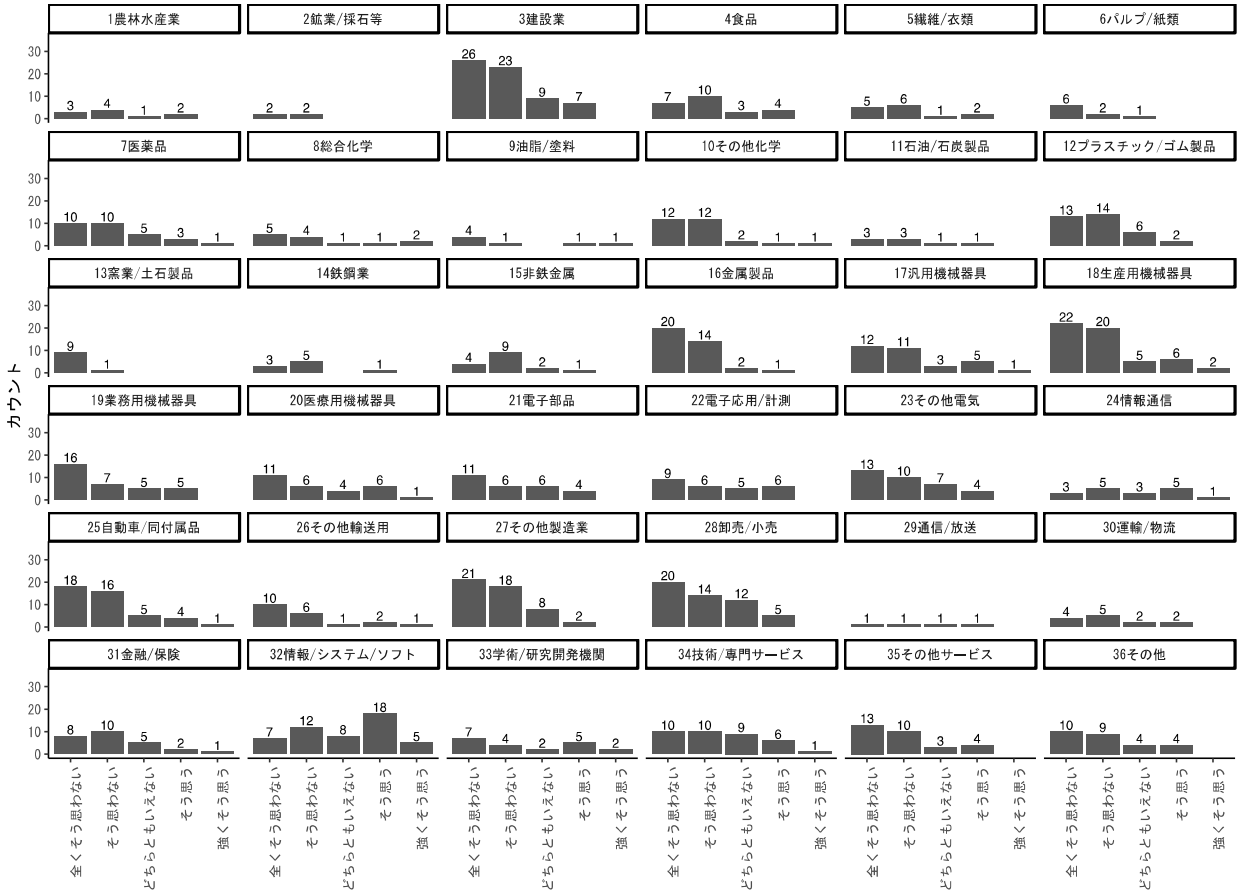

図 44 問 2-5:⑩ディープラーニング等の高度なデータの処理・解析結果を理解し、事業活動に活かせる人材を育成、雇用している(産業毎)

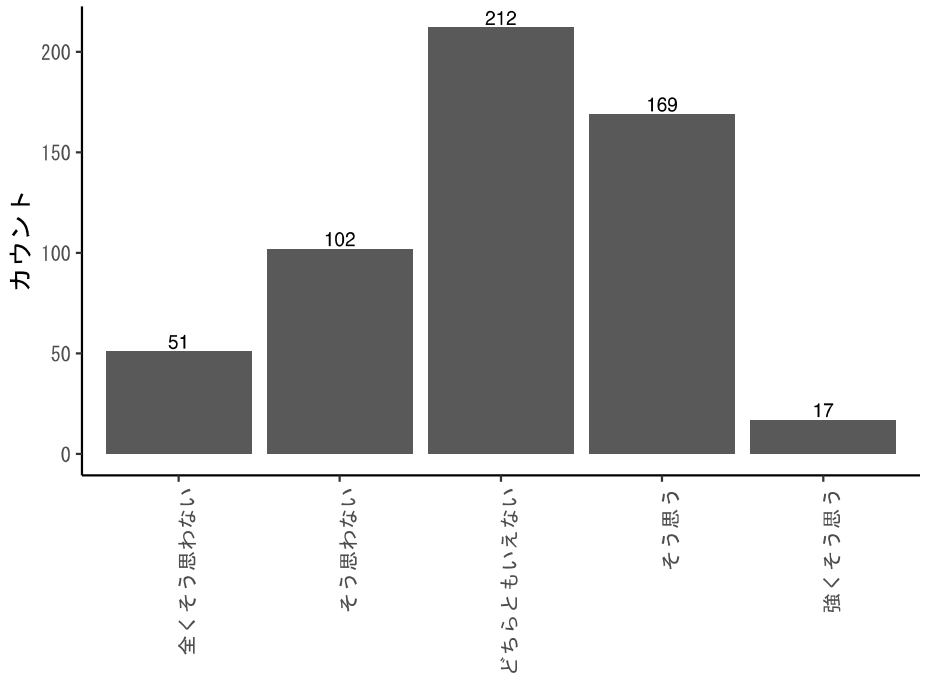

図 45 問 2-5:⑪事業活動に関連するあらゆる情報のデータ化を進めている(全産業)

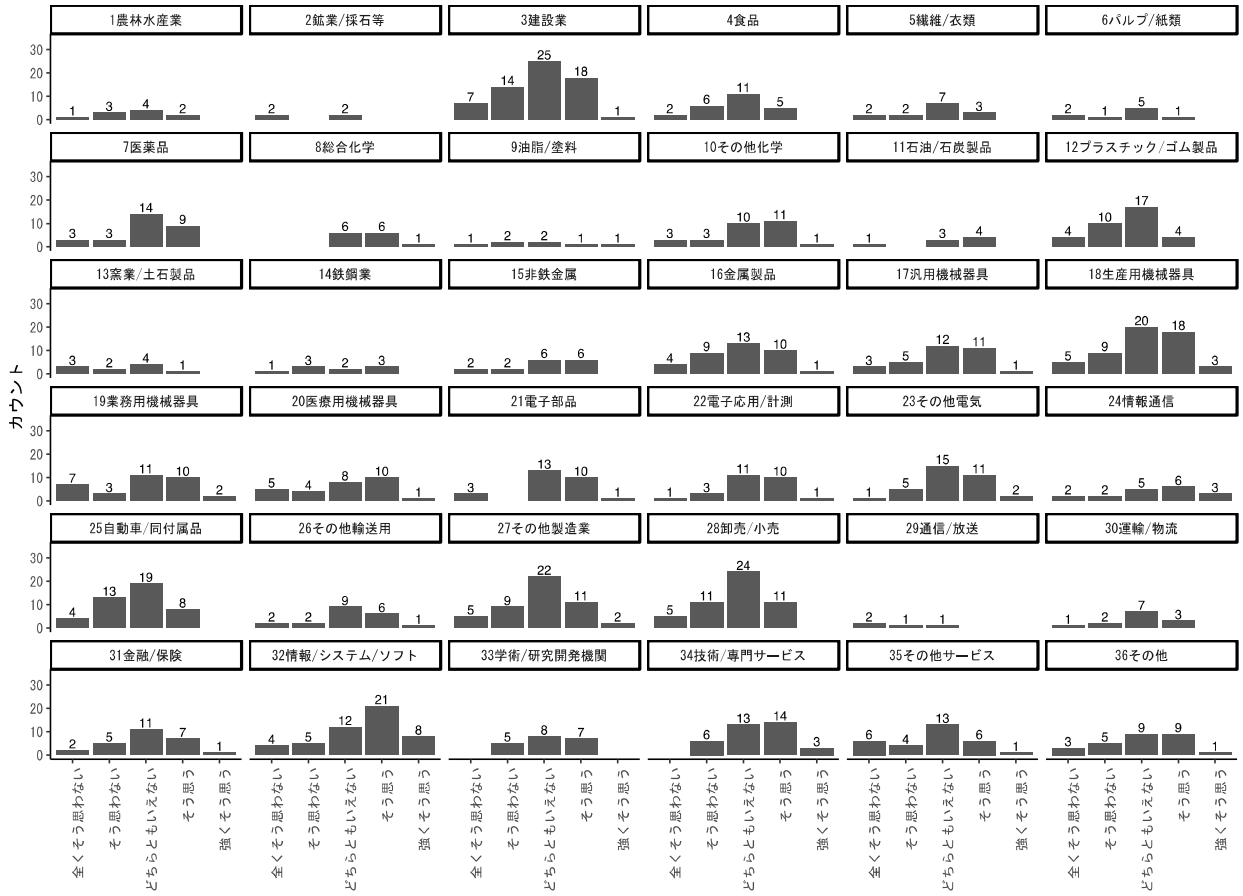

図 46 問 2-5:⑪事業活動に関連するあらゆる情報のデータ化を進めている(産業毎)

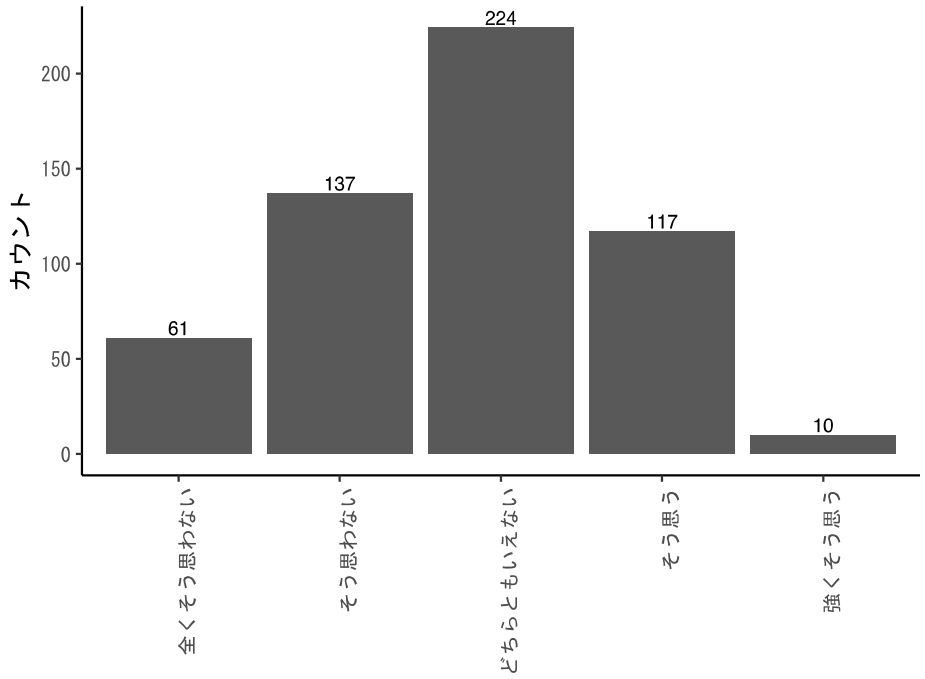

図 47 問 2-5:⑫事業活動の目的や今後の展開に沿ってどのようなデータが有用か十分に吟味し、データを設計している(全産業)

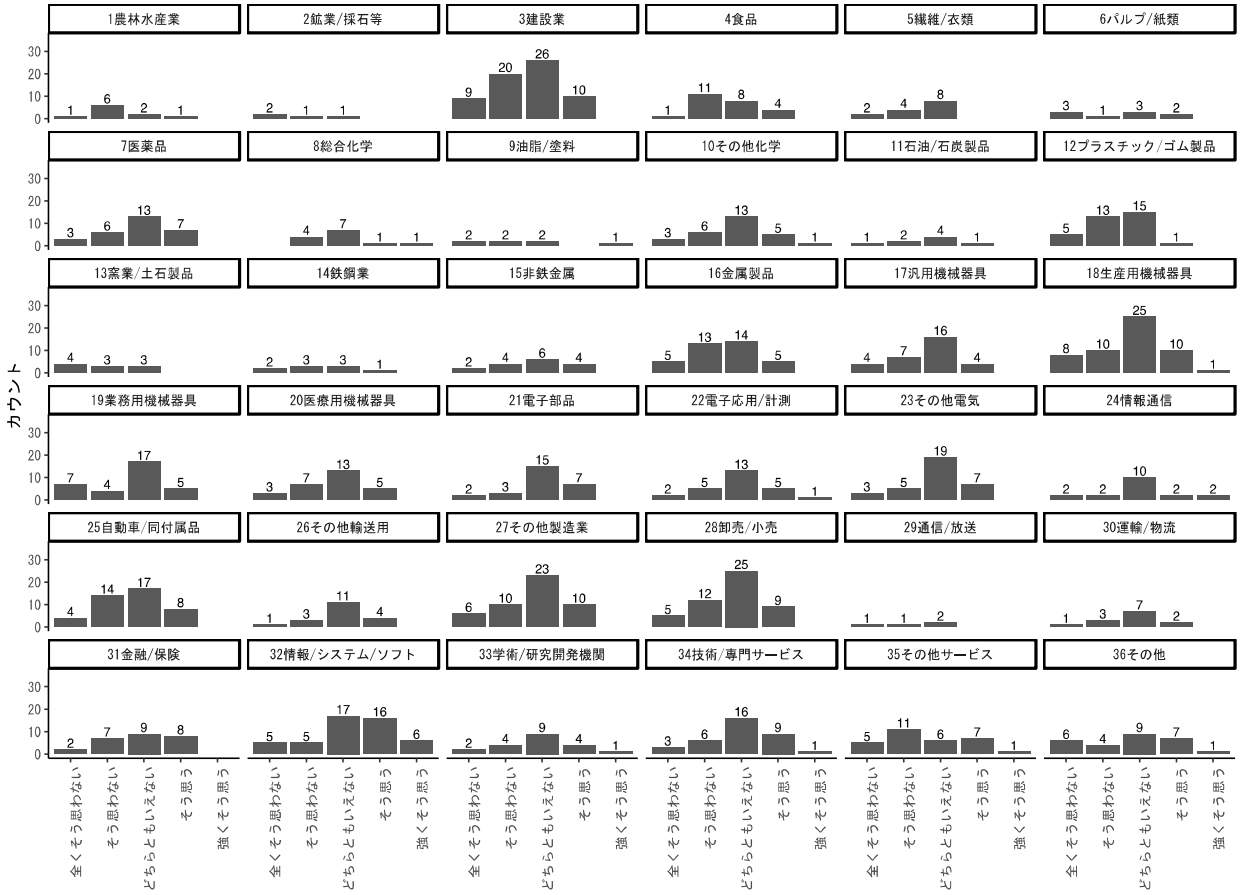

図 48 問 2-5:⑫事業活動の目的や今後の展開に沿ってどのようなデータが有用か十分に吟味し、データを設計している(産業毎)

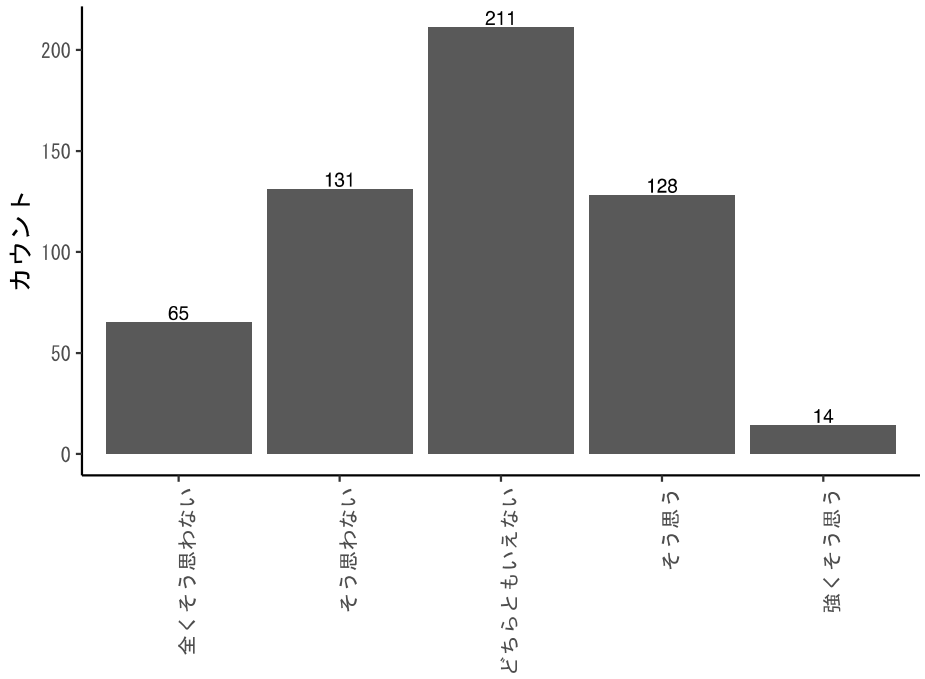

図 49 問 2-5:⑬複数のデータを組み合わせられるよう、データを設計している(全産業)

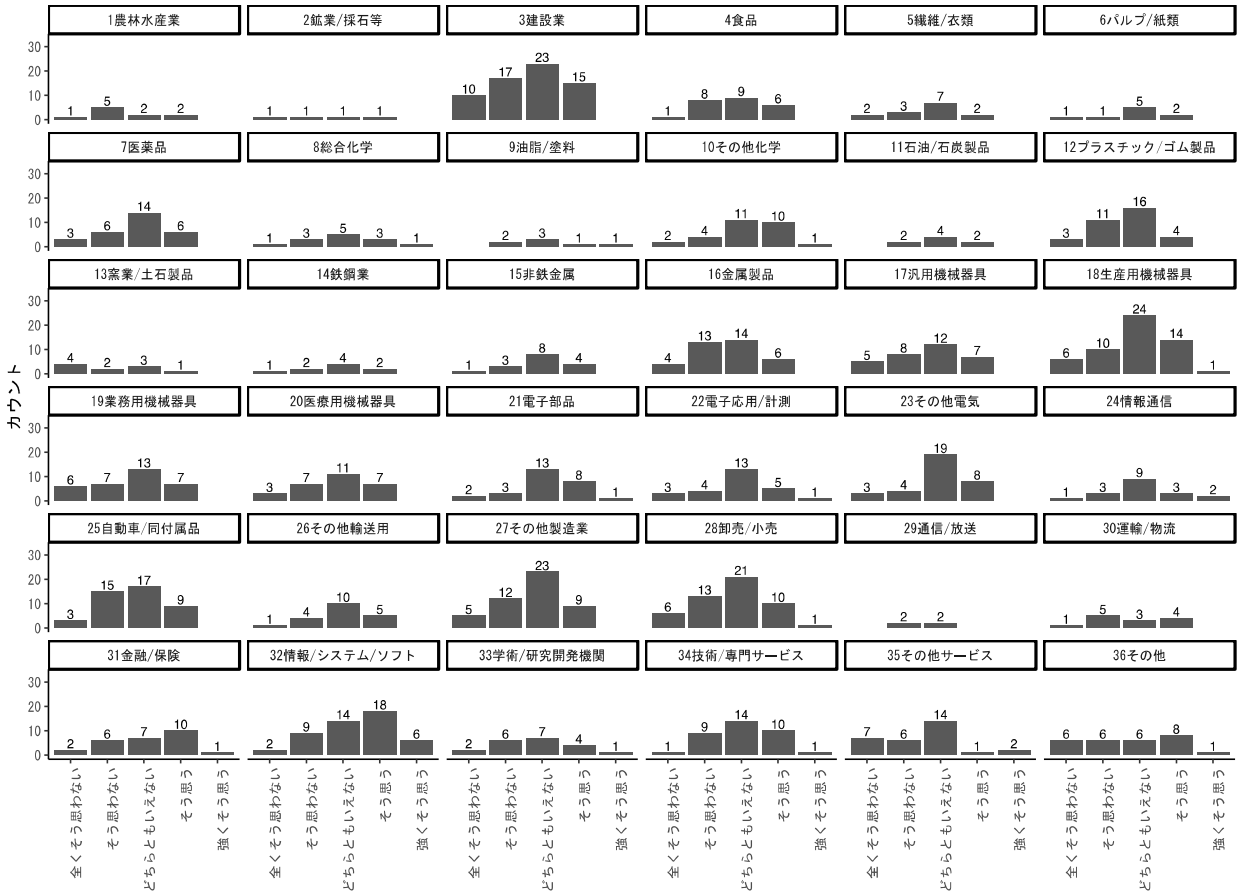

図 50 問 2-5:⑬複数のデータを組み合わせられるよう、データを設計している(産業毎)

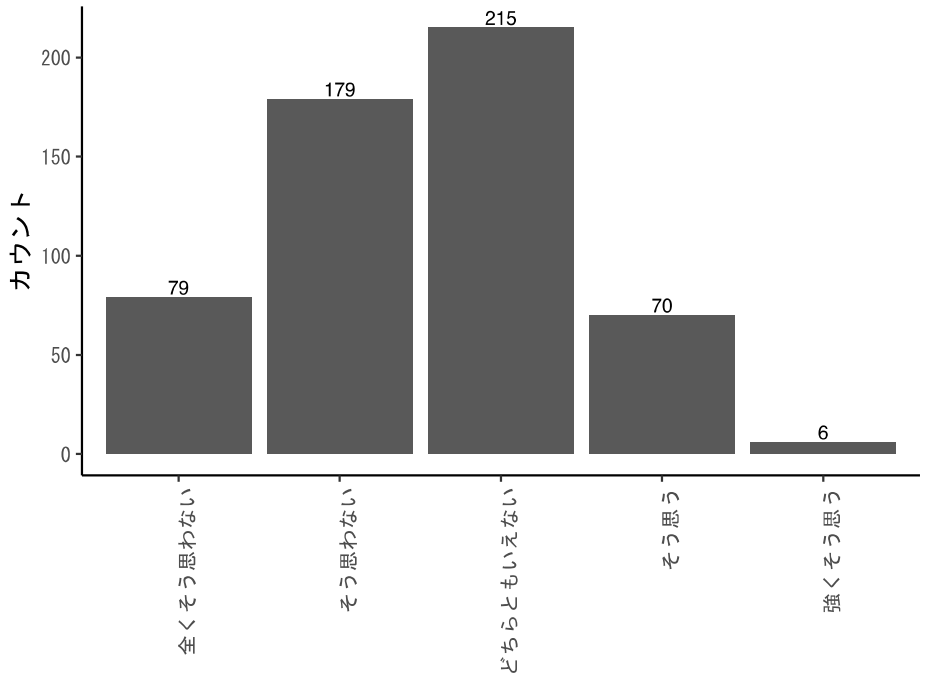

図 51 問 2-5:⑭データの有用性を評価し、必要なデータを取捨選択する等、定期的にデータ設計の見直しを行っている(全産業)

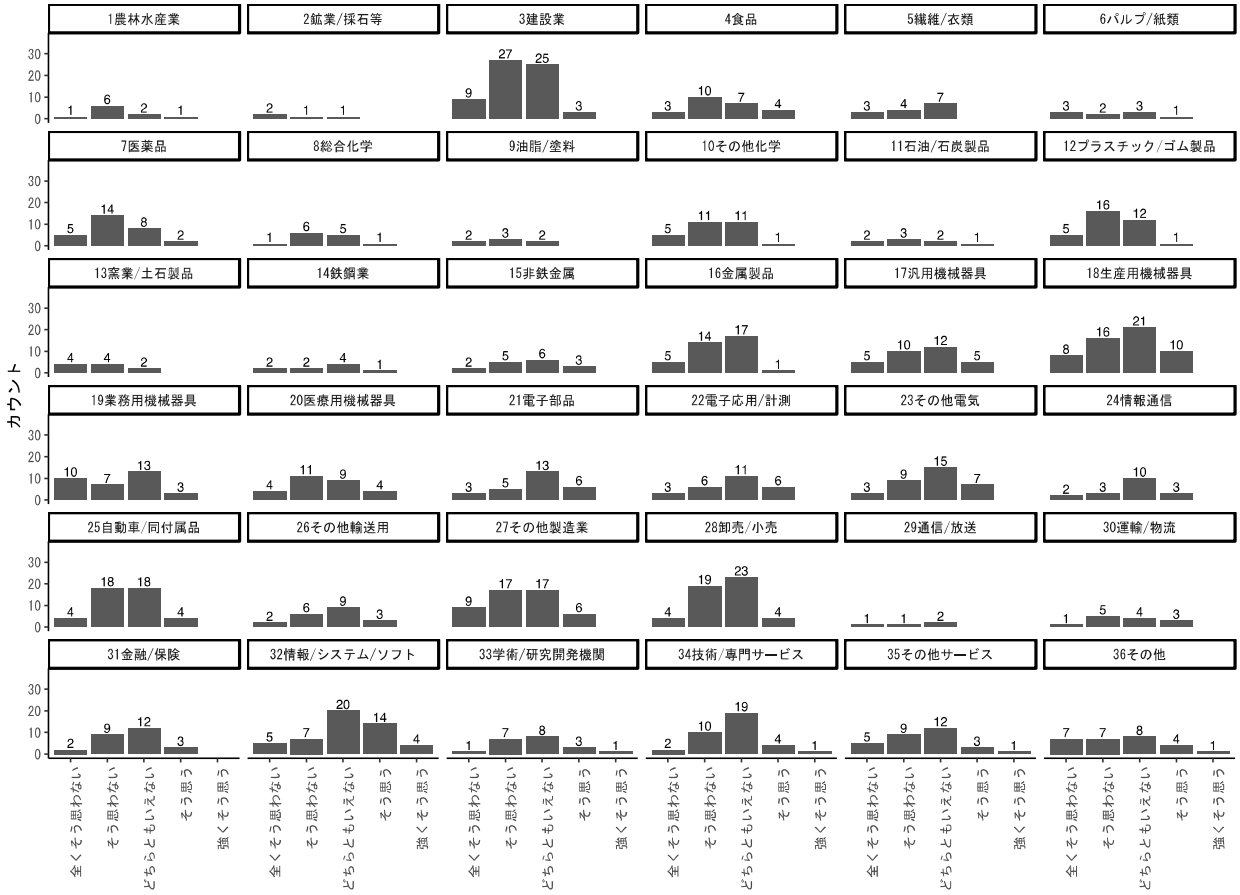

図 52 問 2-5:⑭データの有用性を評価し、必要なデータを取捨選択する等、定期的にデータ設計の見直しを行っている(産業毎)

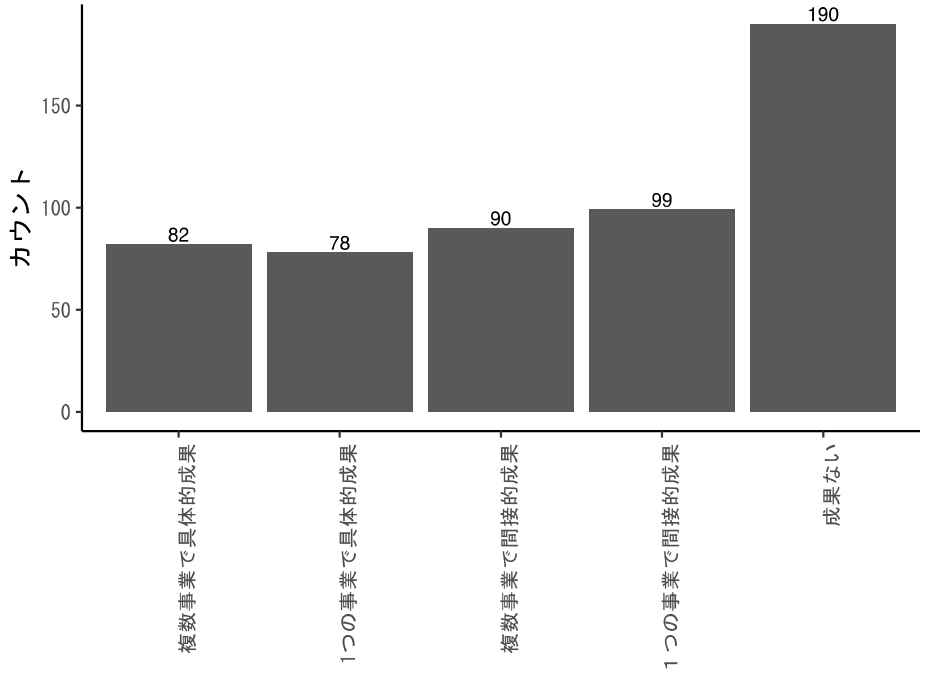

図 53 問 2-7:データ利活用成果(全産業)

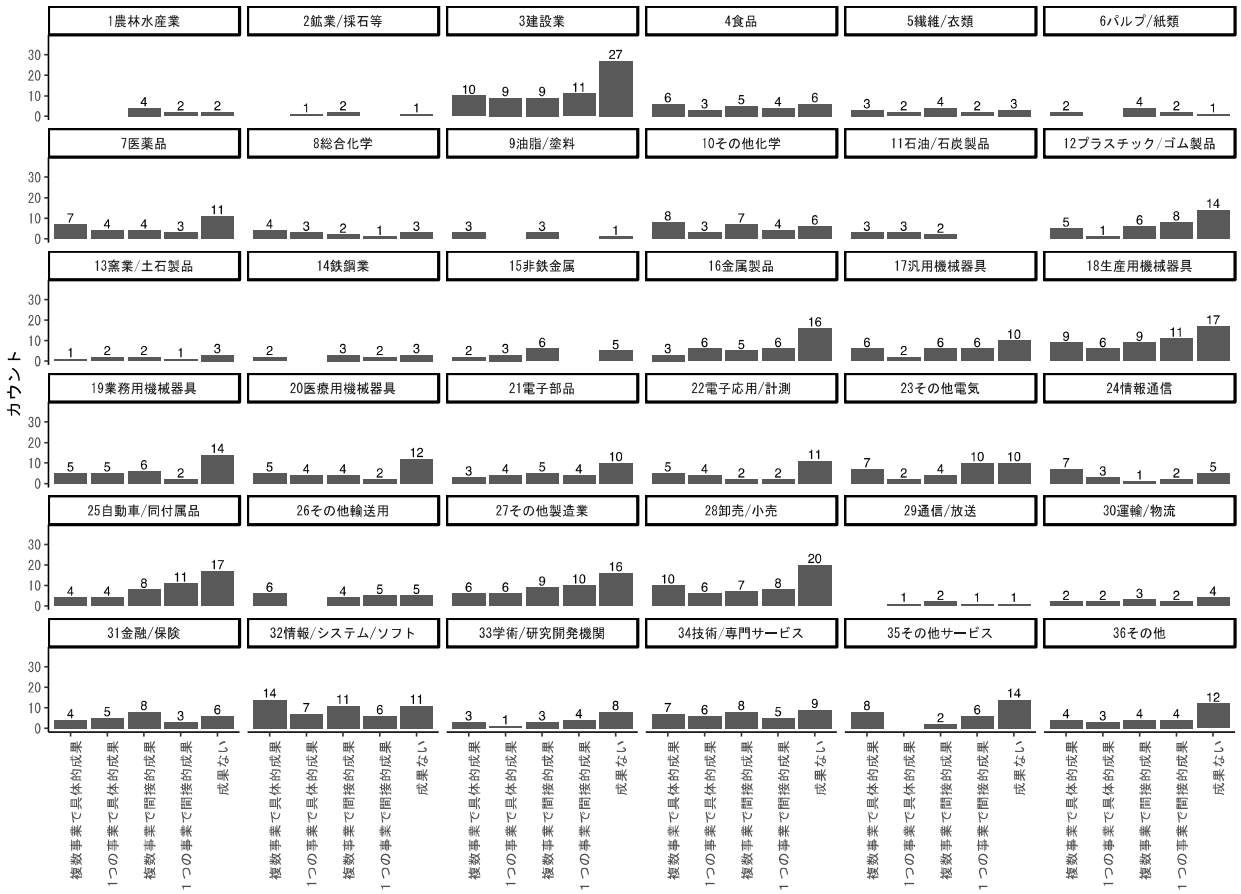

図 54 問 2-7: データ利活用成果(産業毎)

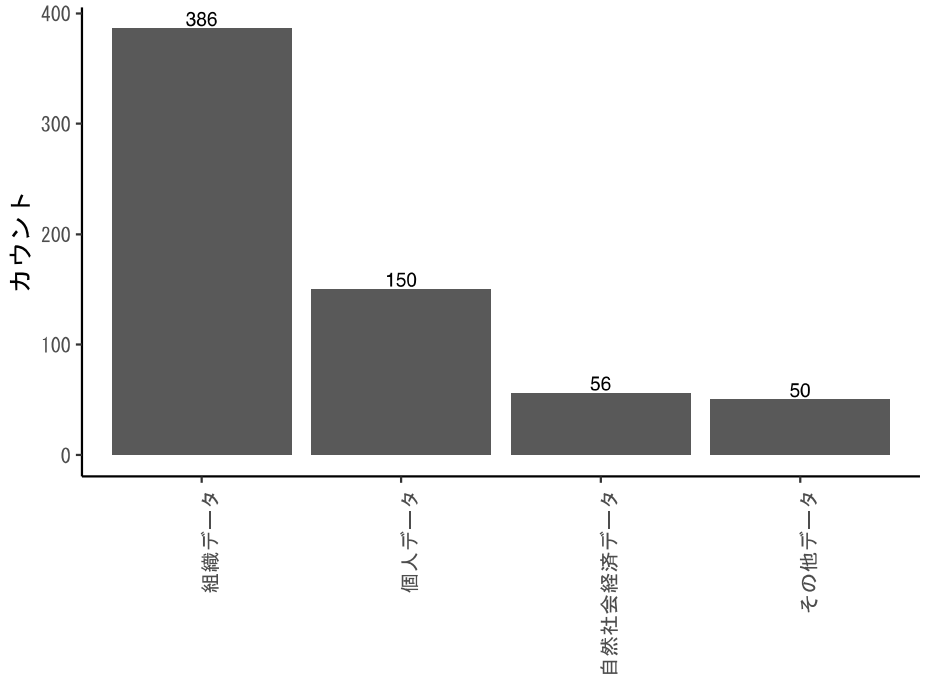

図 55 問 3-1:データ種類(全産業)

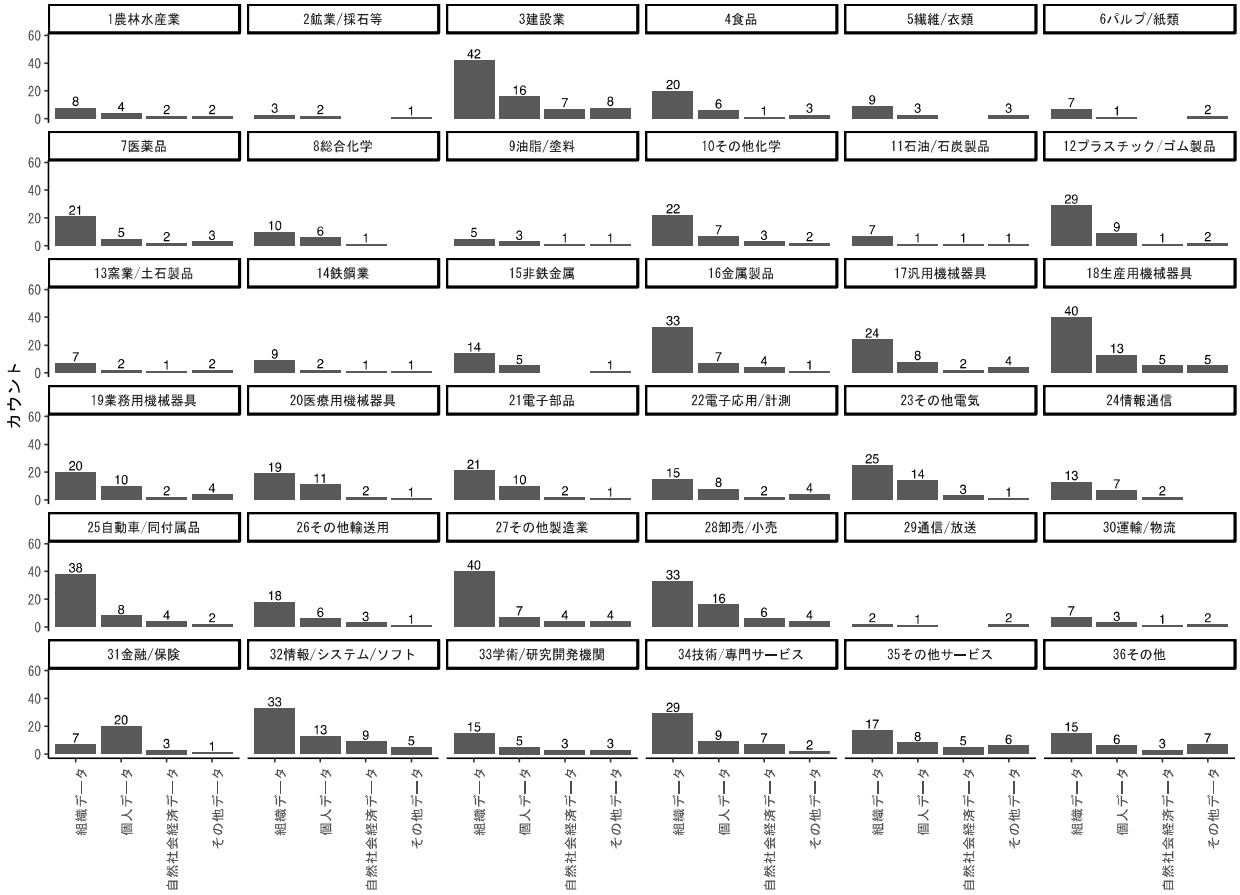

図 56 問 3-1:データ種類(産業毎)

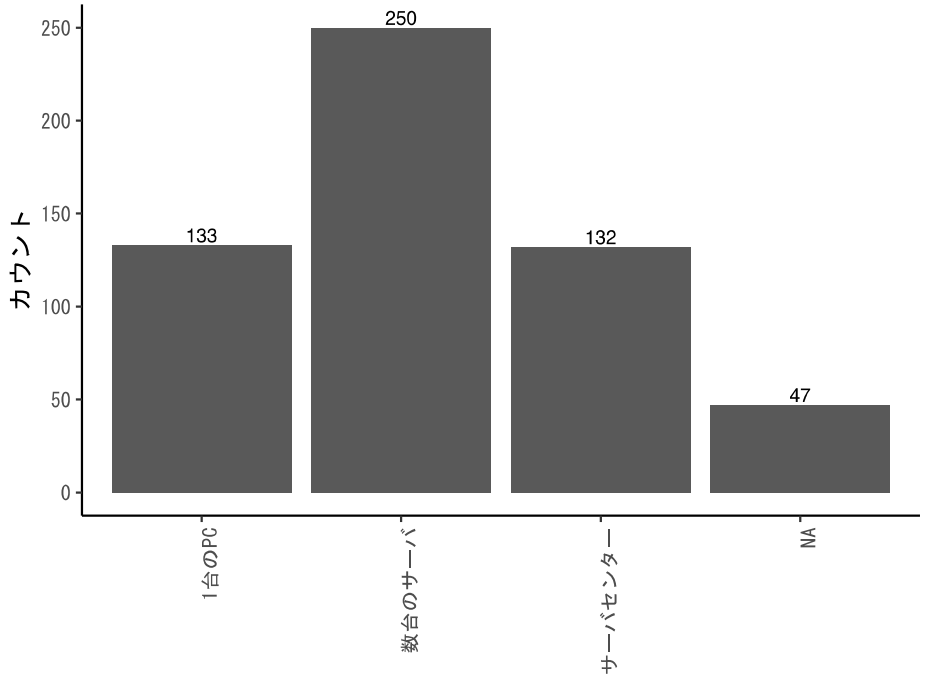

図 57 問 3-10:事業データ総量(全産業)

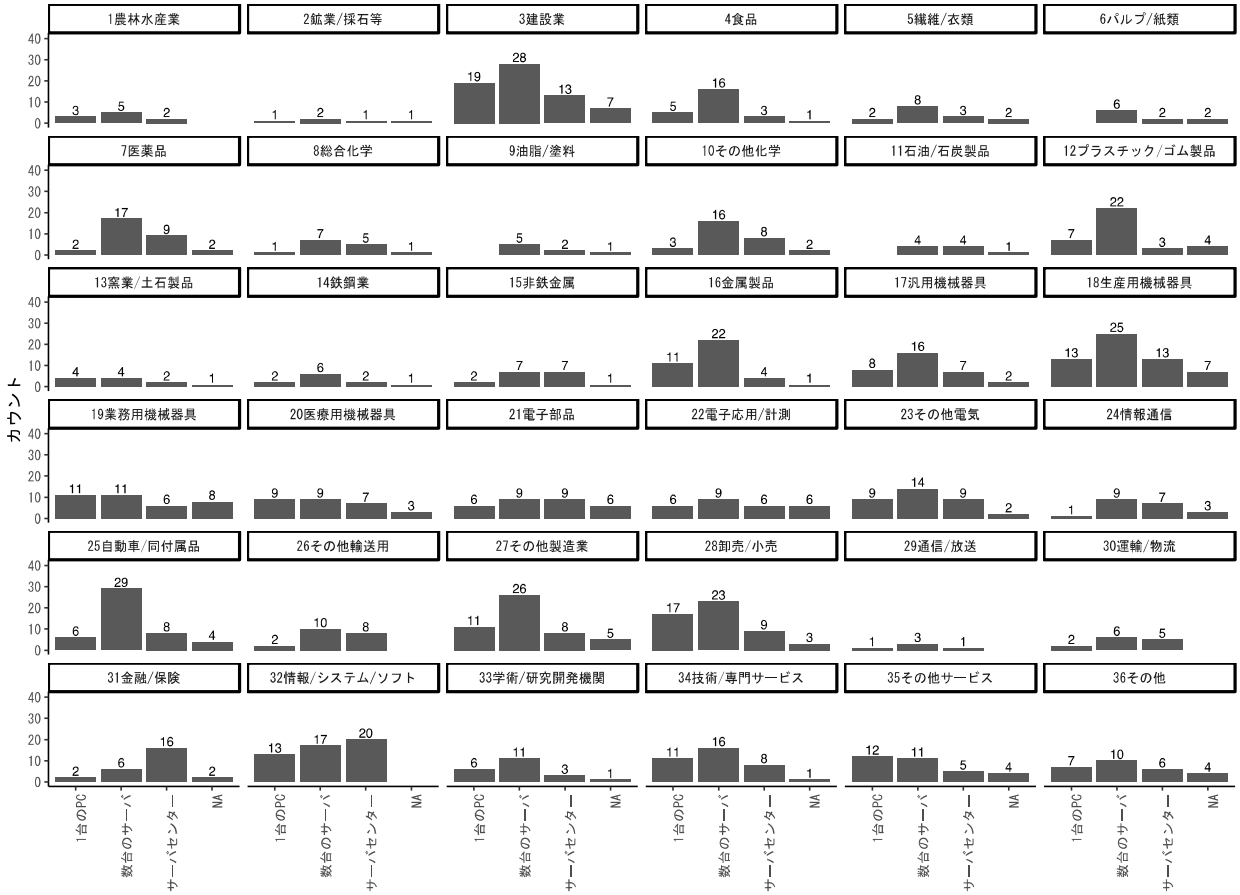

図 58 問 3-10:事業データ総量(産業毎)

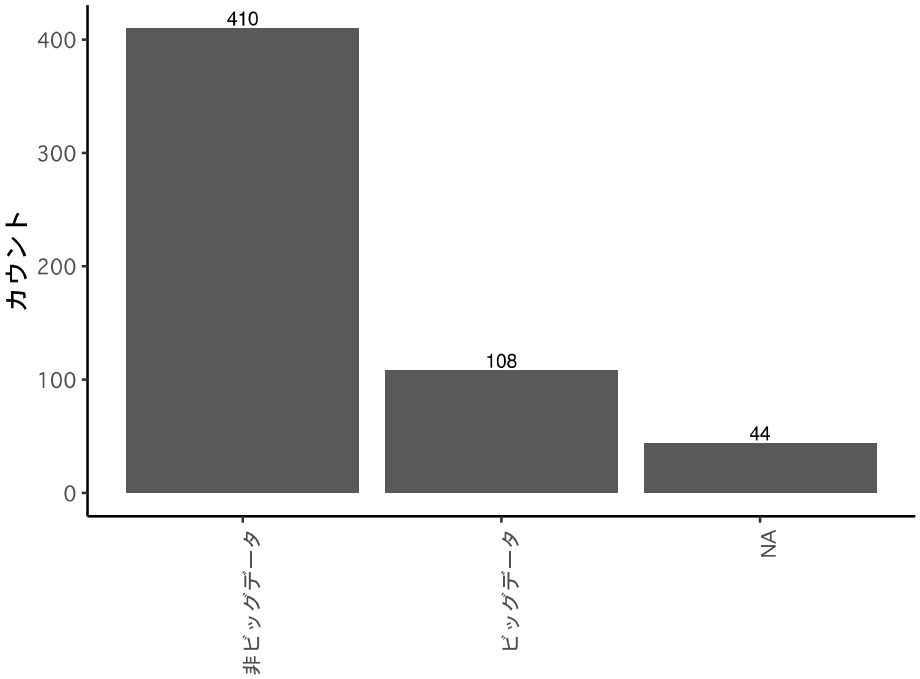

図 59 問 3-11:ビッグデータ該当(産業毎)

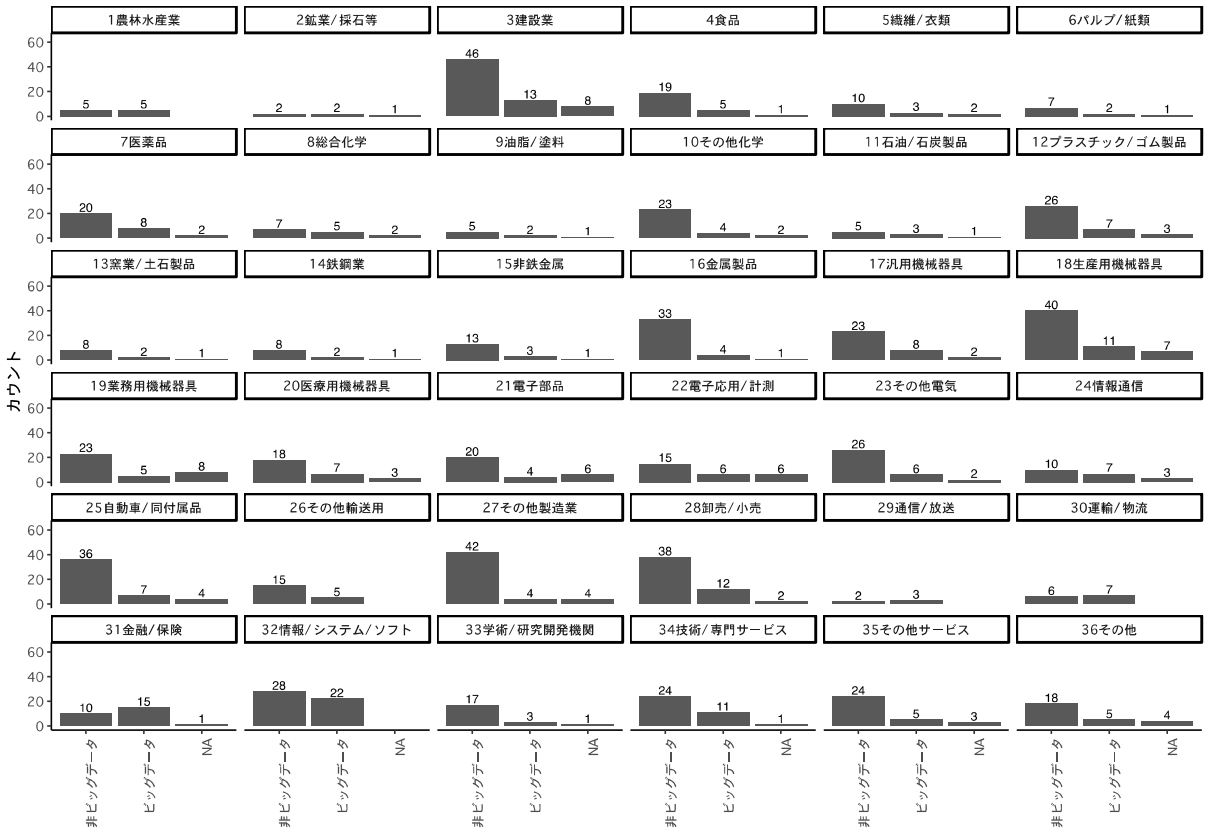

図 60 問 3-11:ビッグデータ該当(産業毎)

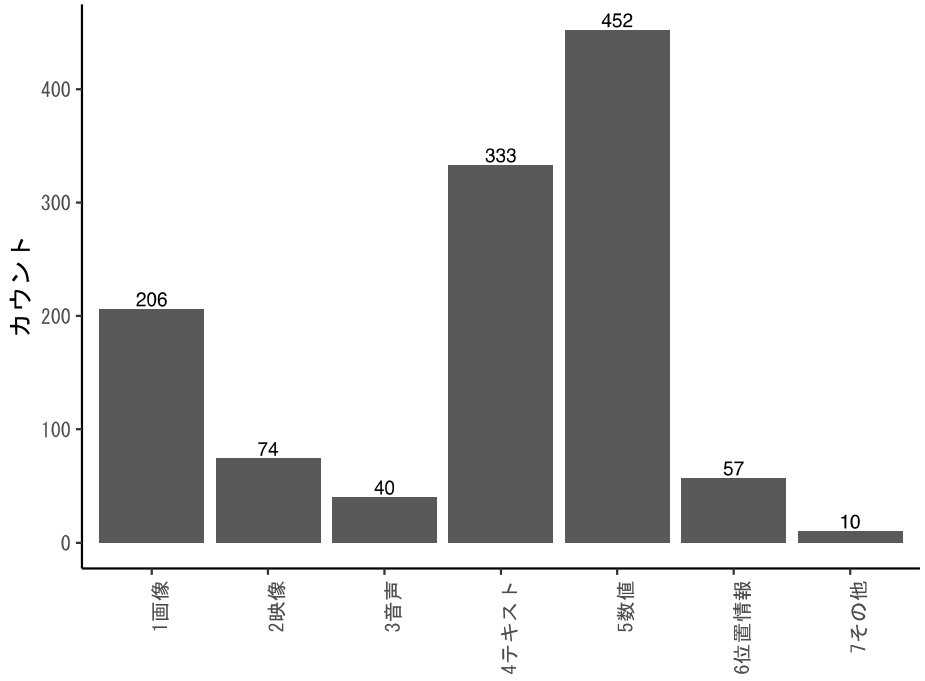

図 61 問 3-12:データ形式(全産業)

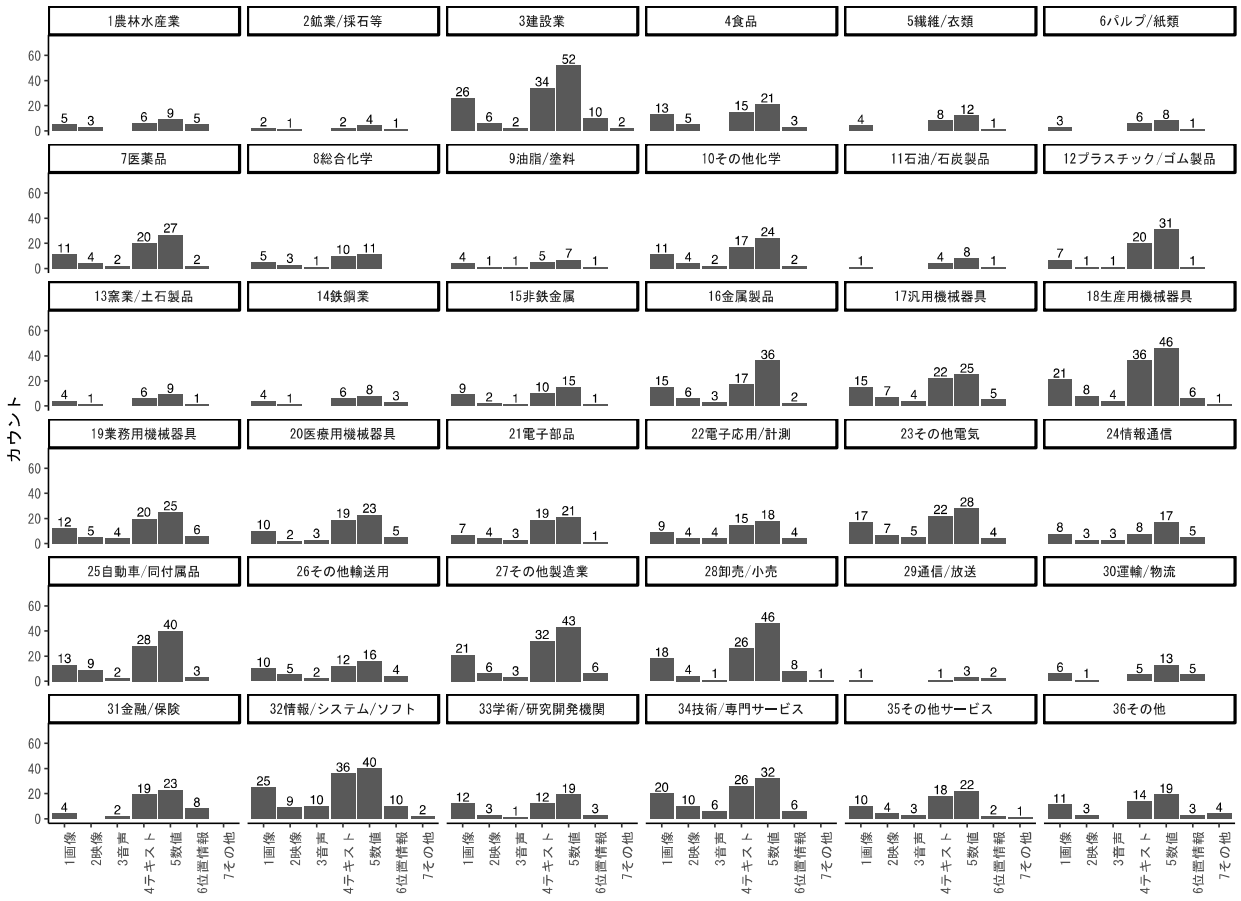

図 62 問 3-12:データ形式(産業毎)

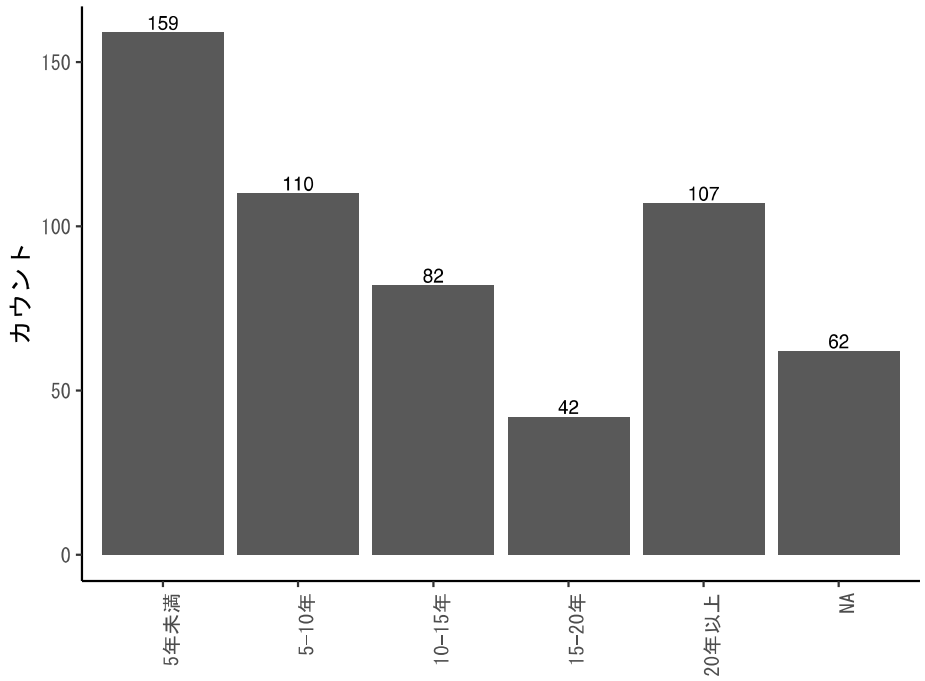

図 63 問 3-14:データ利活用経験(全産業)

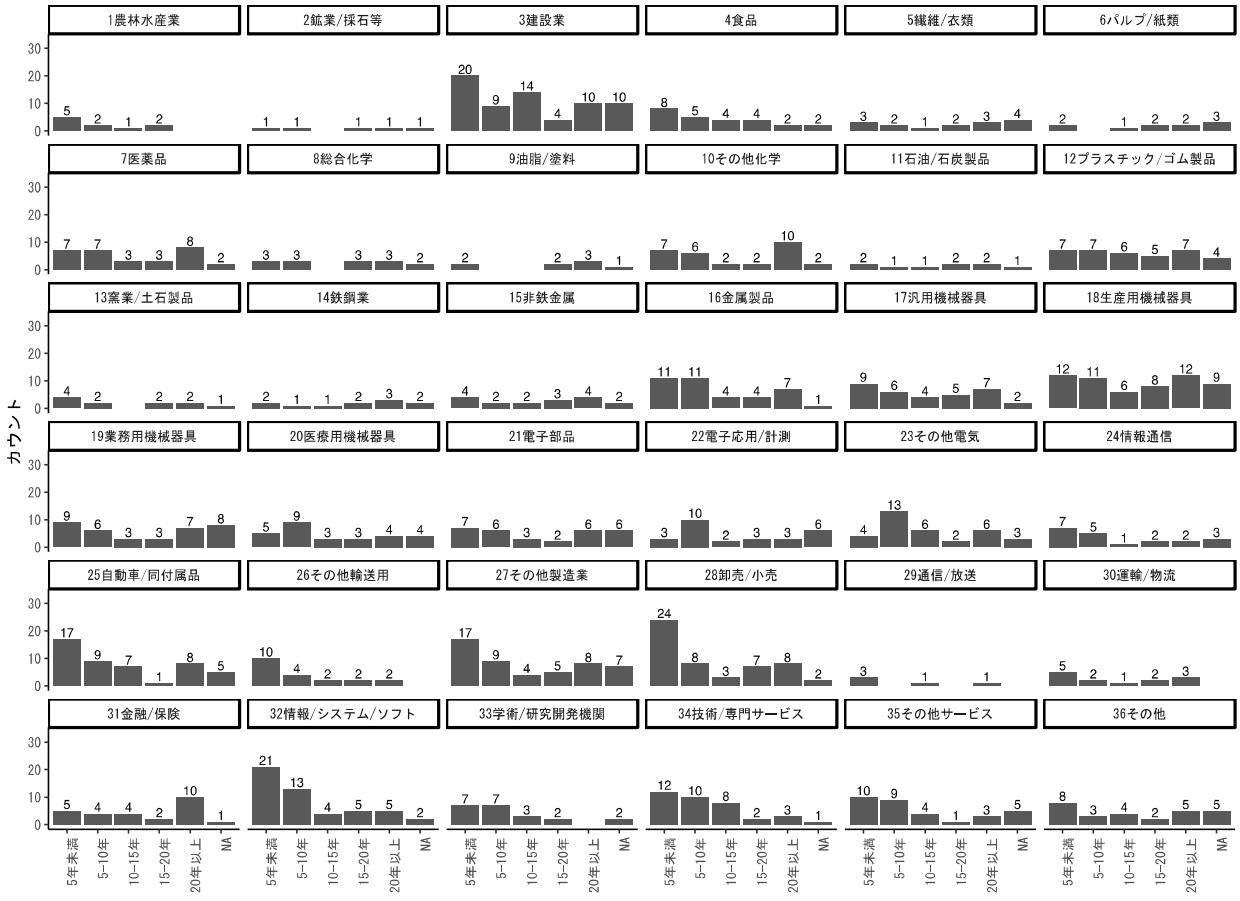

図 64 問 3-14:データ利活用経験(産業毎)

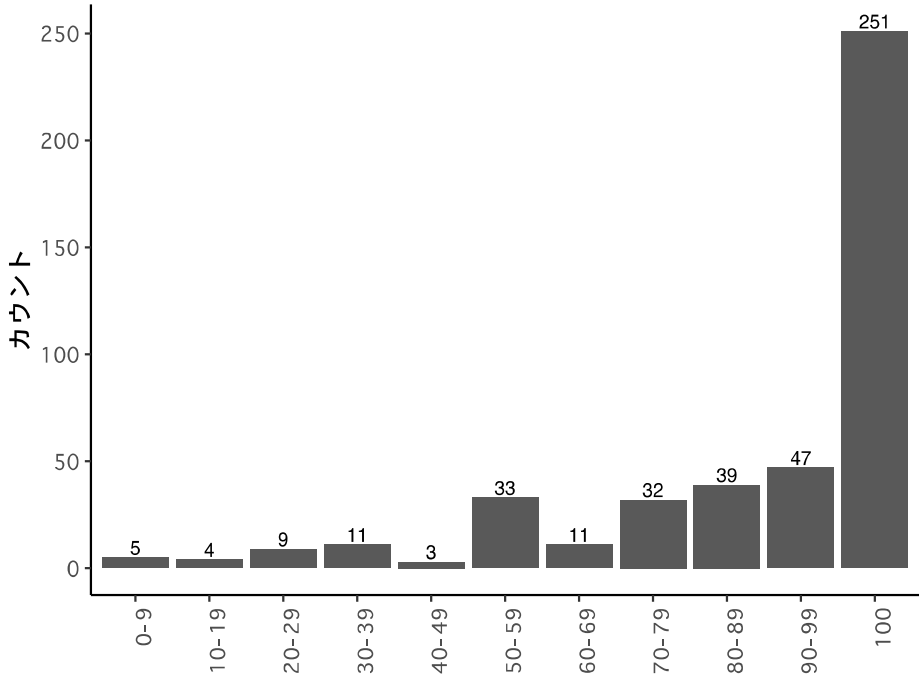

図 65 問 3-17:データイニシアティブ(全産業)

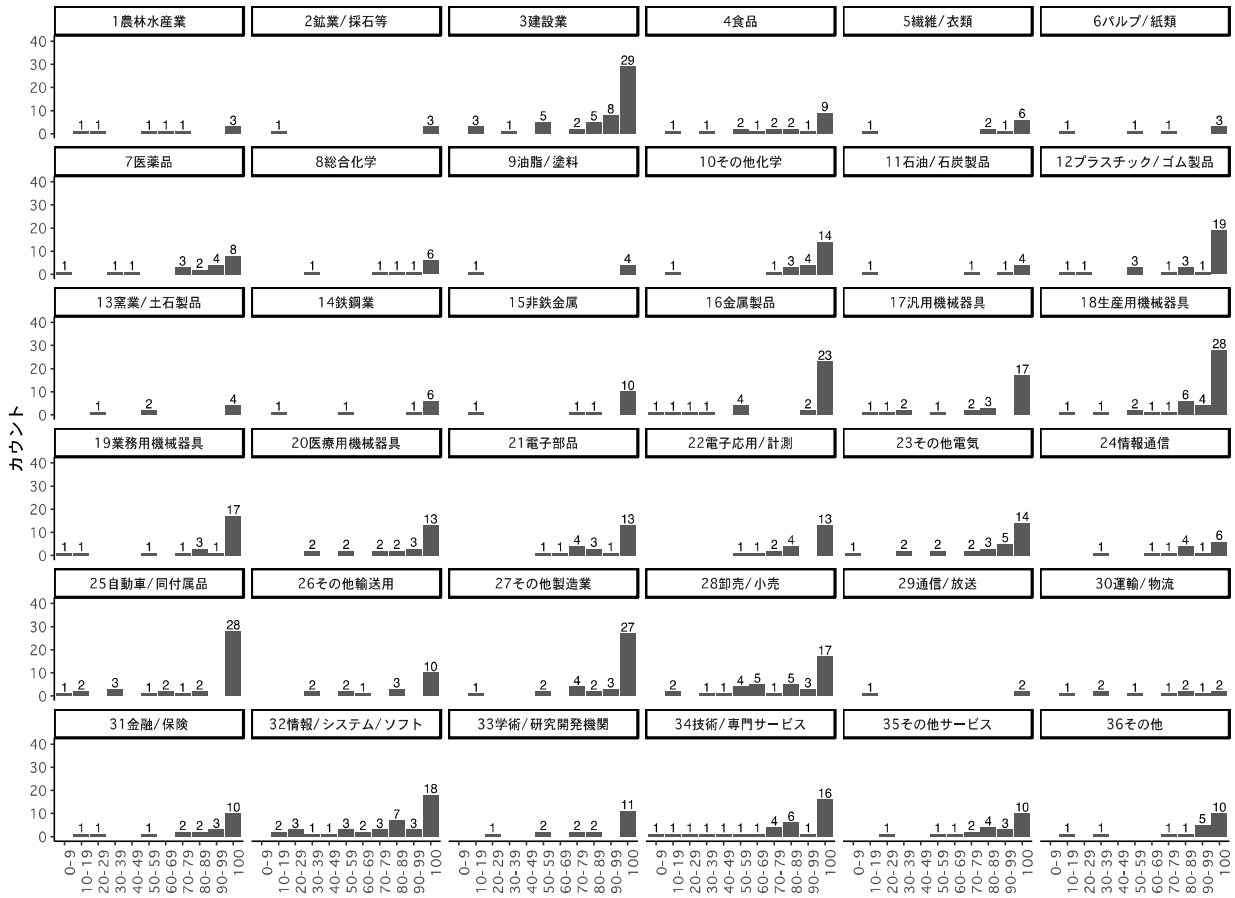

図 66 問 3-17:データイニシアティブ(産業毎)

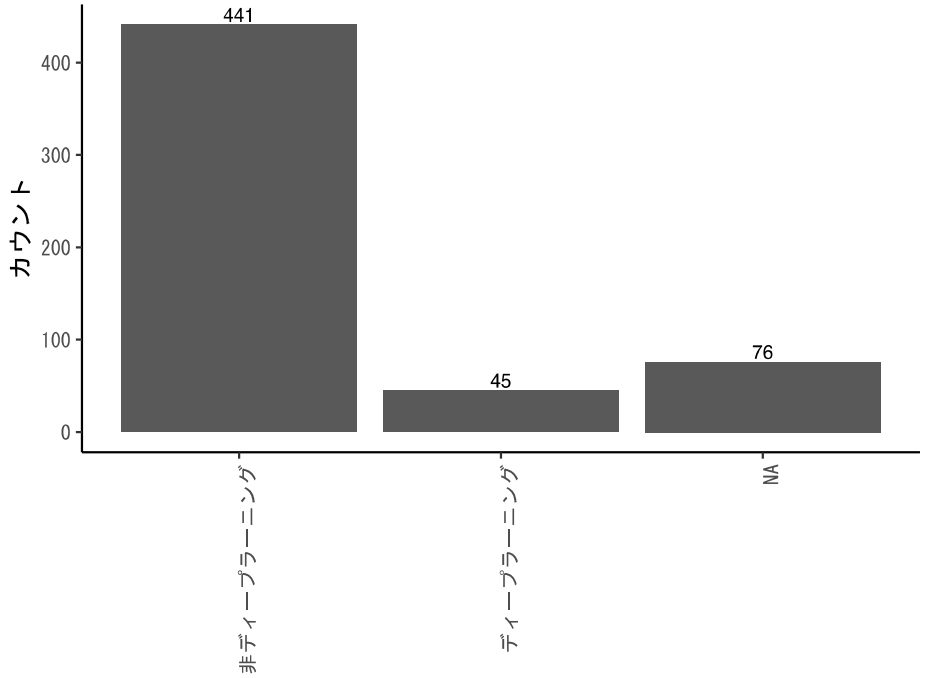

図 67 問 3-18:高度なデータ処理・解析(全産業)

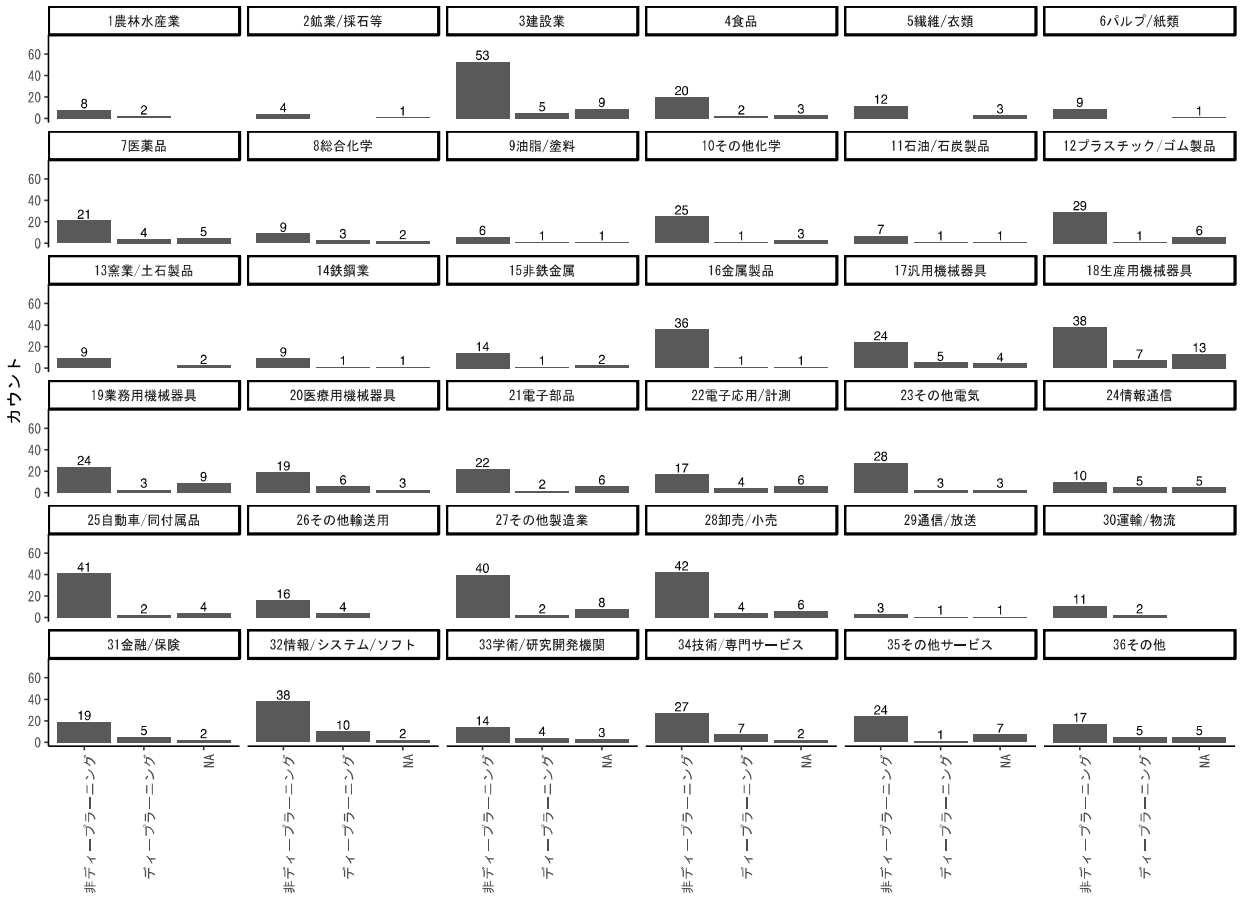

図 68 問 3-18: 高度なデータ処理・解析(産業毎)

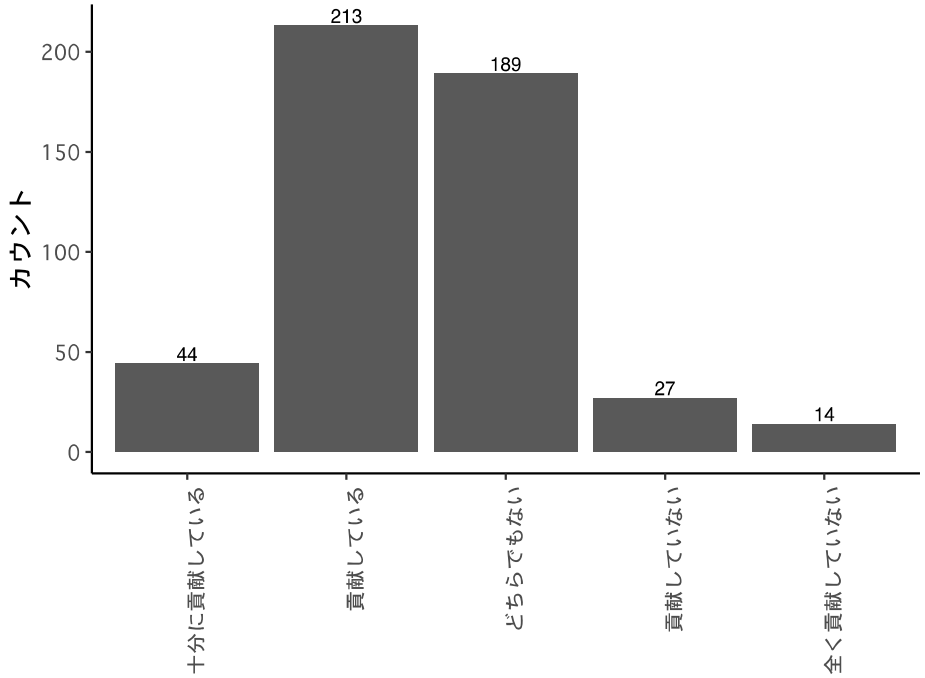

図 69 問 3-28:事業競争力貢献(全産業)

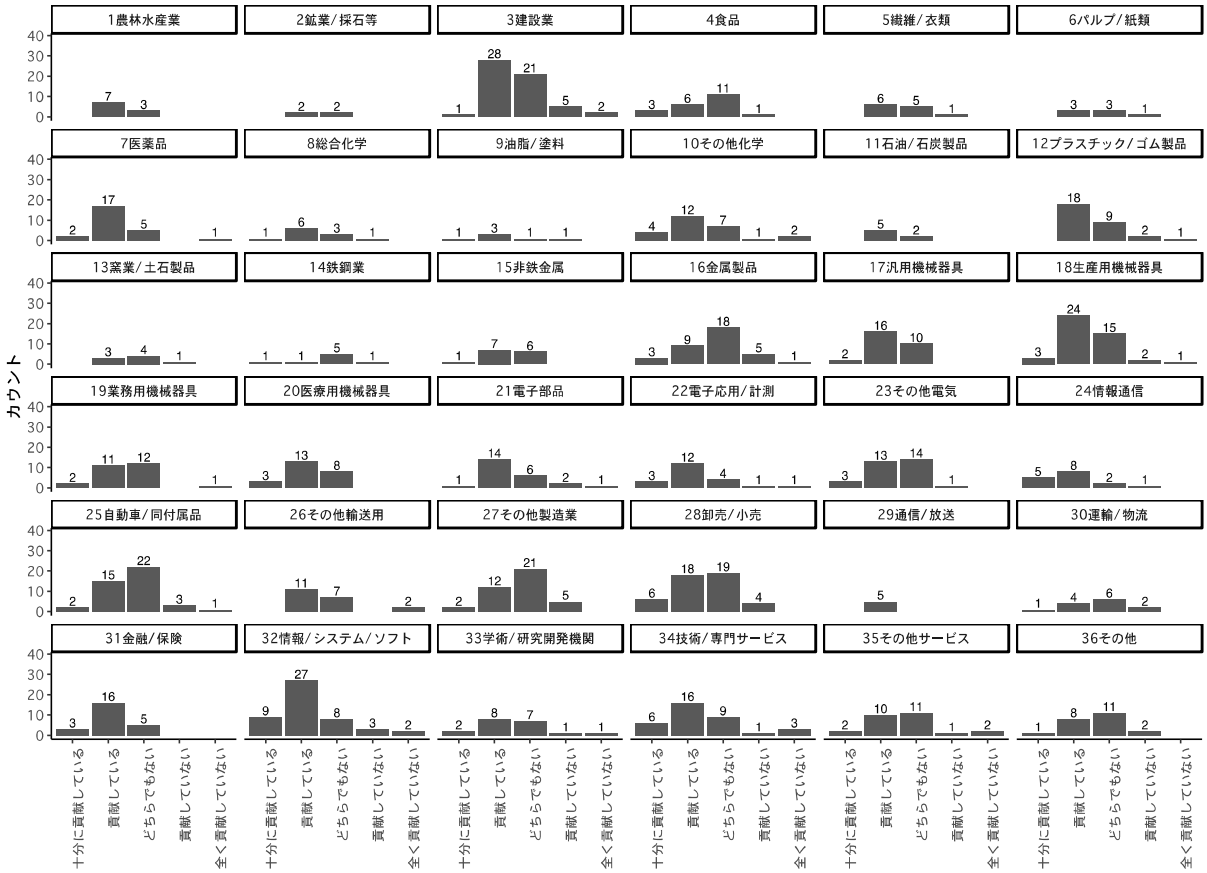

図 70 問 3-28: 事業競争力貢献(産業毎)

6. Appendix 2

Appendix 2 では産業毎に属性項目と成果項目に関するプロット図と反応係数を含む回帰式を掲示する。なお、プロット図を作成するに当たり、表示を見やすくするために、各サンプルにごく僅かな乱数（ジッター）を加えている。このため、算出された回帰式に僅かな誤差を含む。

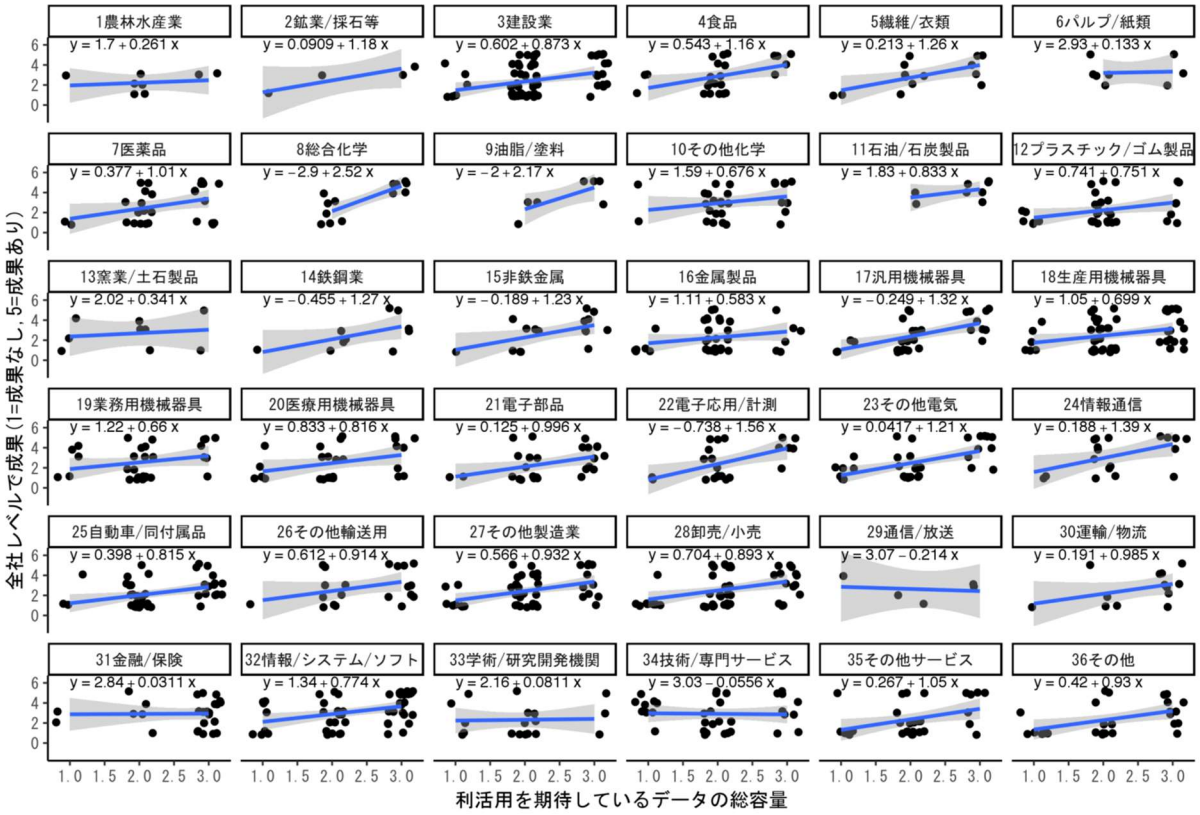

図 71 問 2-7（データ利活用成果）と問 2-1（全社データ総量）の反応係数図（産業毎）

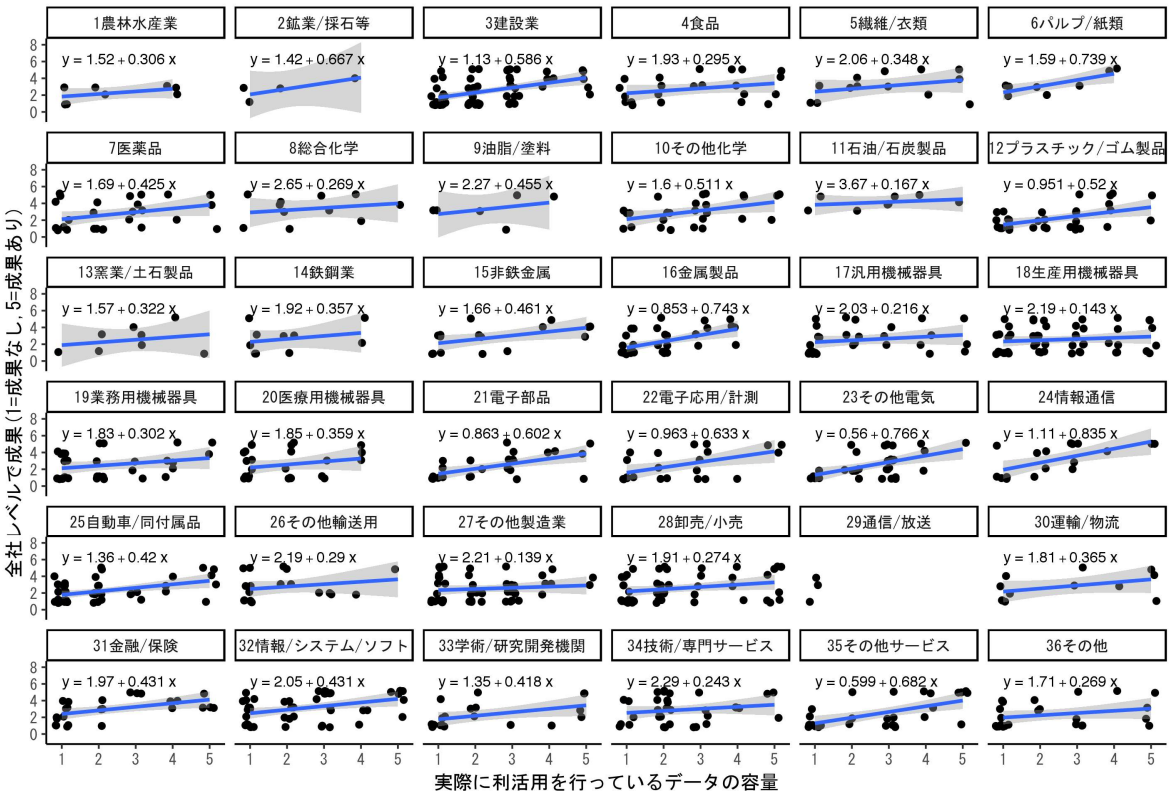

図 72 問 2-7（データ利活用成果）と問 2-2（データ利用率）の反応係数図（産業毎）

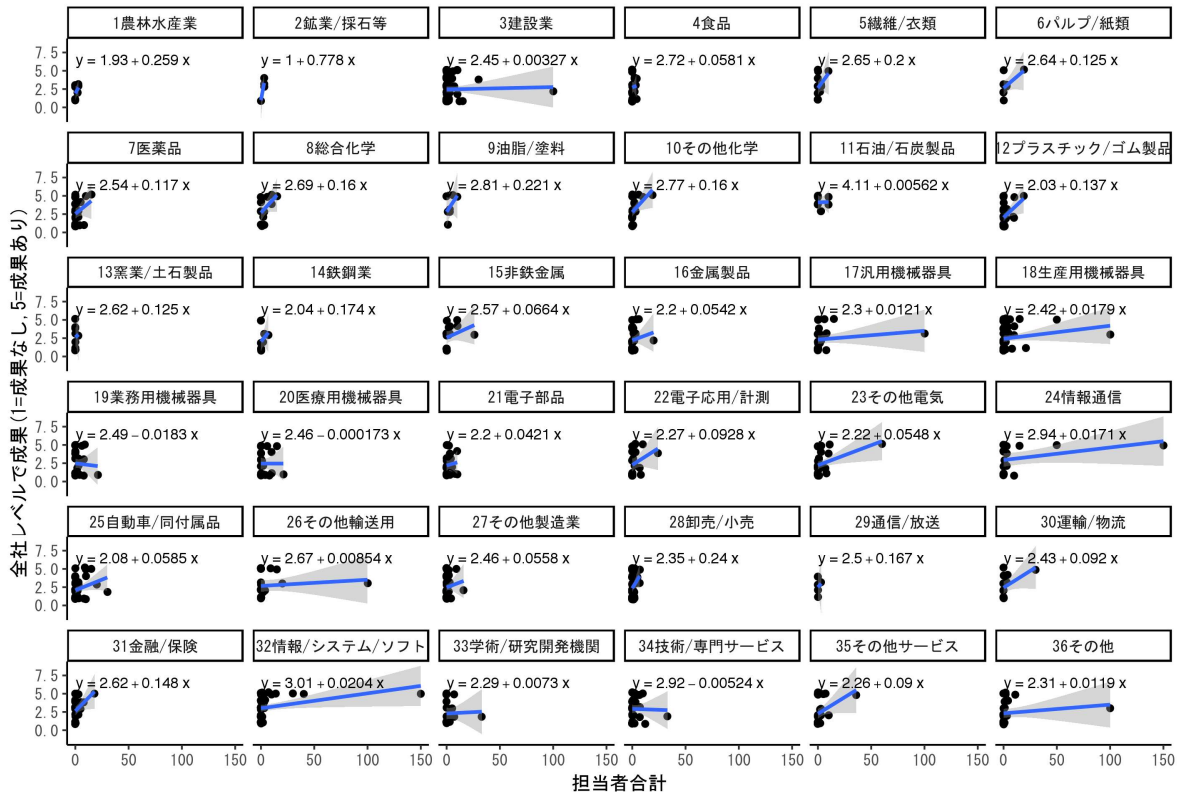

図 73 問 2-7（データ利活用成果）と問 2-3（担当者合計）の反応係数図（産業毎）

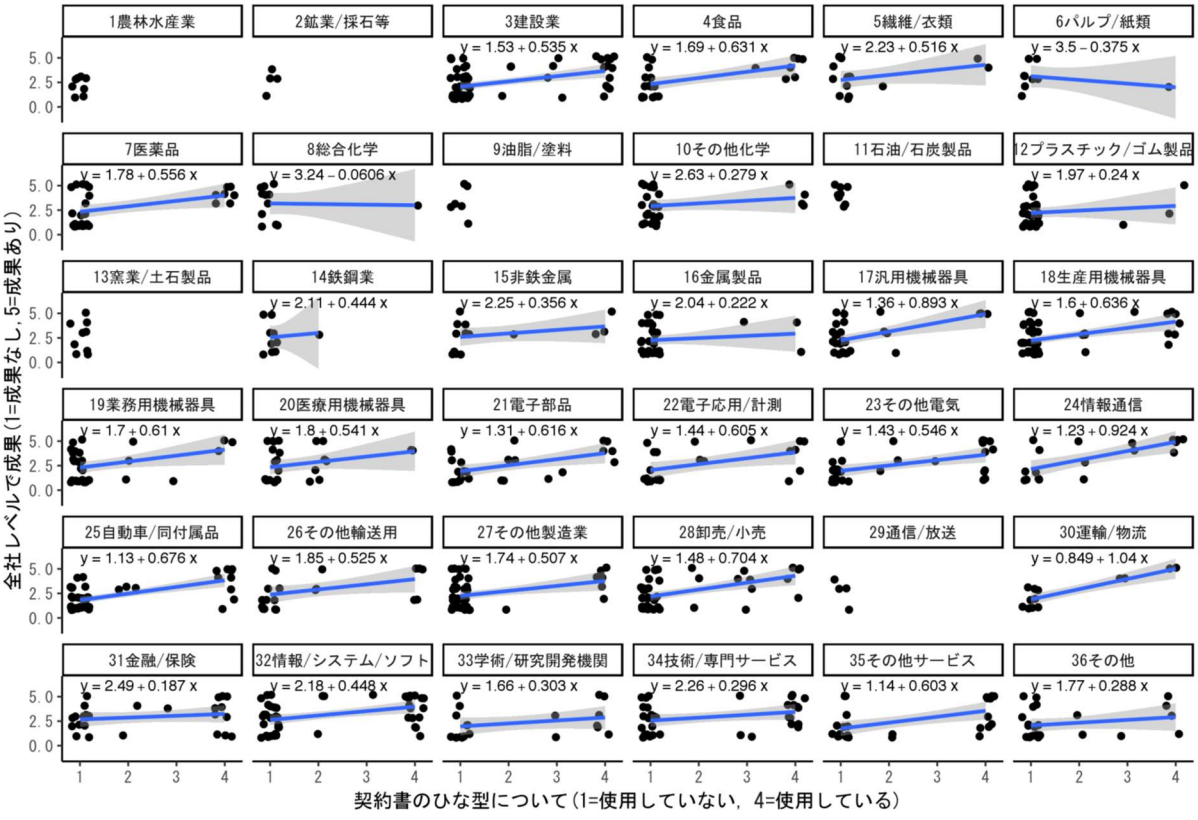

図 74 問 2-7（データ利活用成果）と問 2-4（契約書）の反応係数図（産業毎）

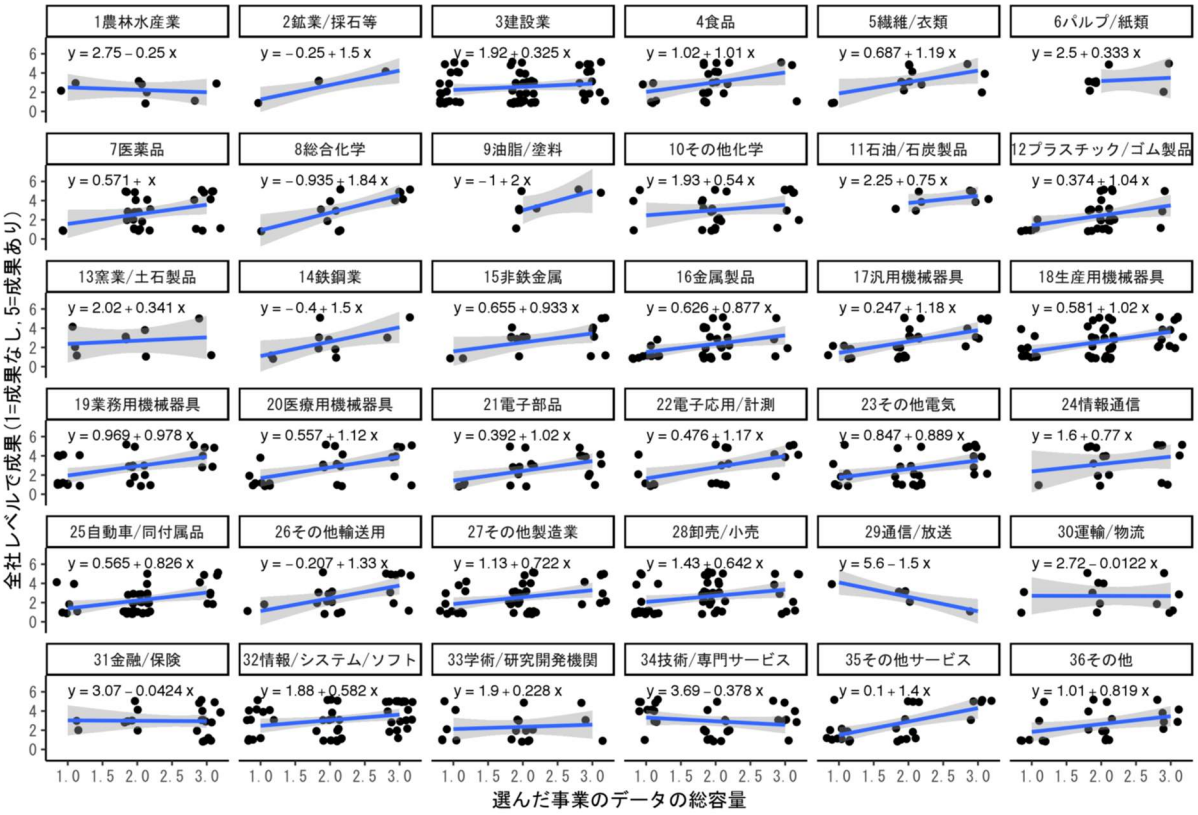

図 75 問 2-7（データ利活用成果）と問 3-10（事業データ総量）の反応係数図（産業毎）

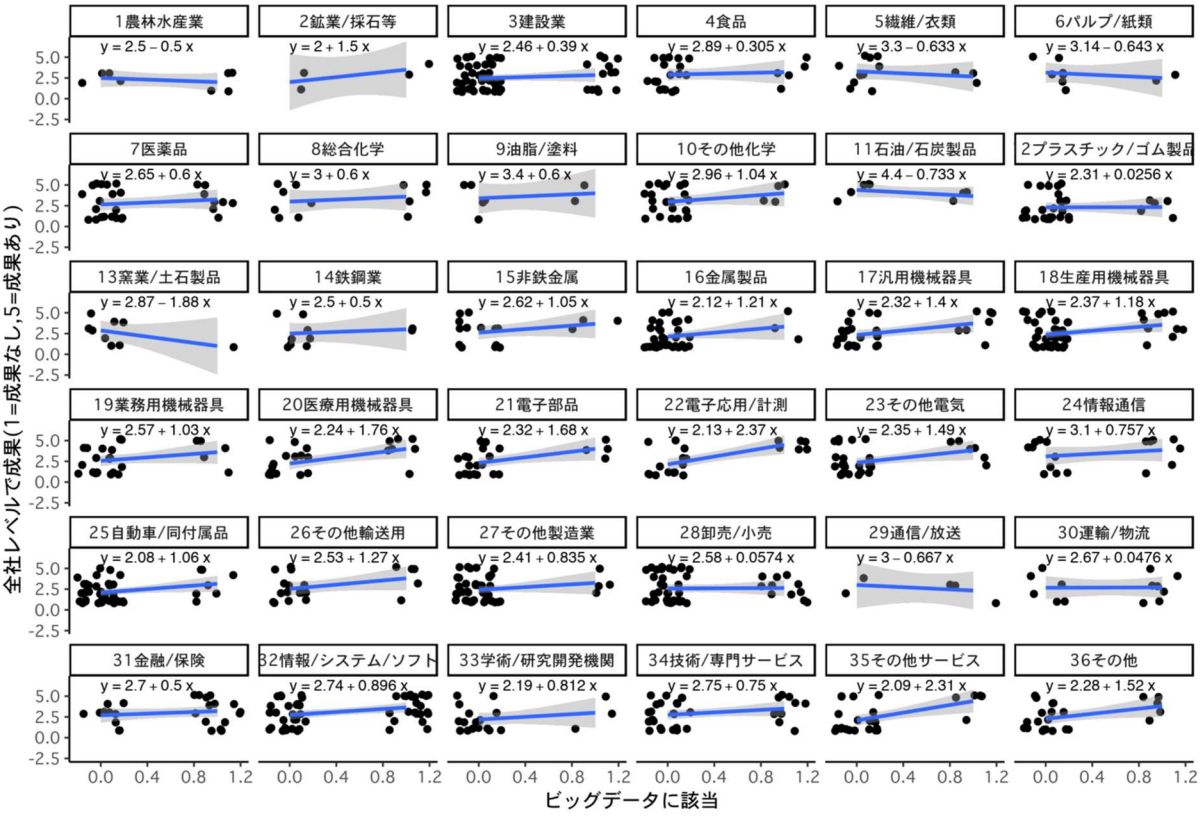

図 76 問 2-7（データ利活用成果）と問 3-11（ビッグデータ該当）の反応係数図（産業毎）

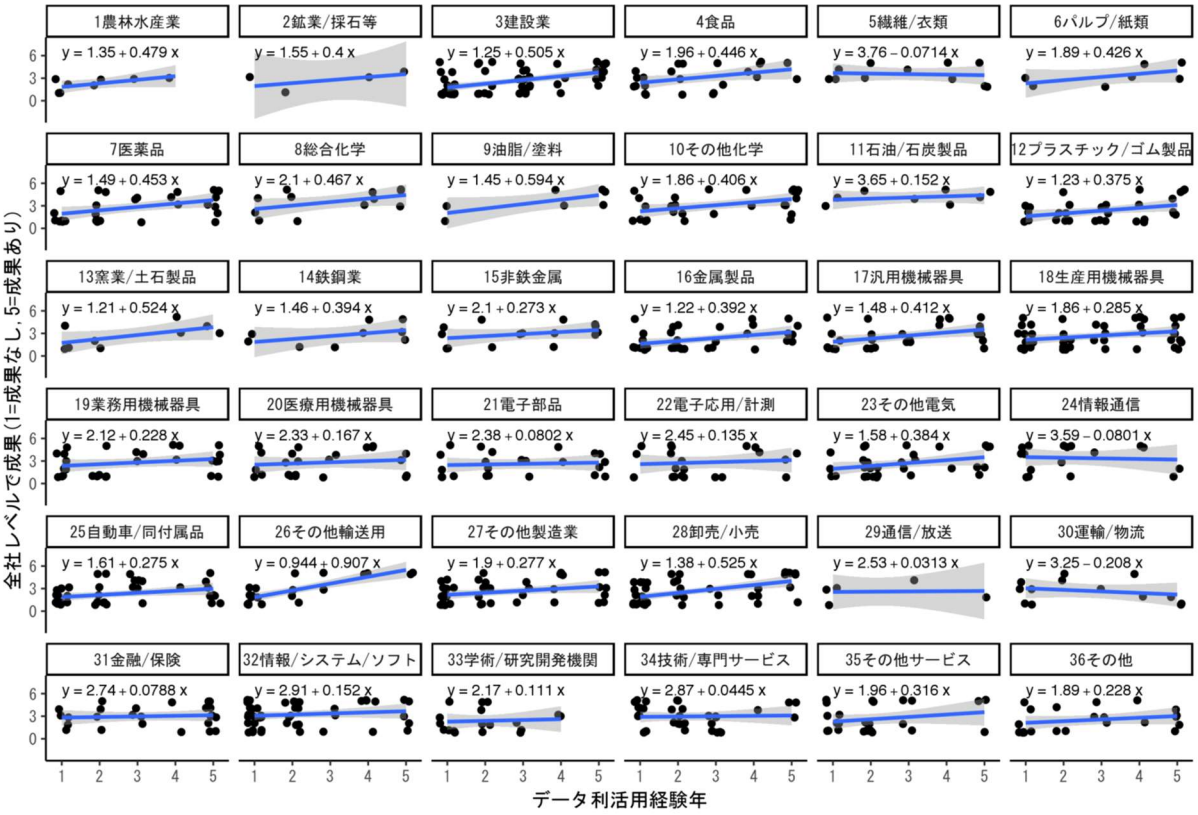

図 77 問 2-7（データ利活用成果）と問 3-14（データの利活用経験）の反応係数図（産業毎）

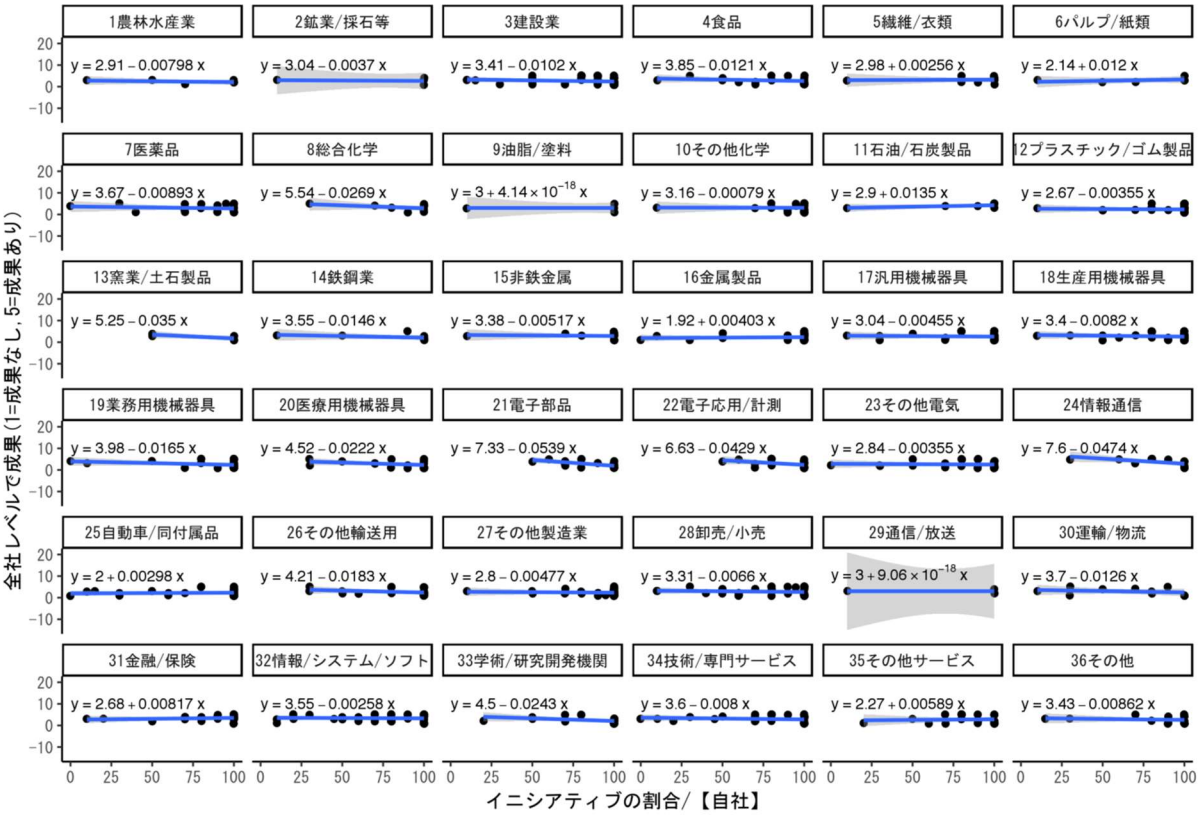

図 78 問 2-7（データ利活用成果）と問 3-17（データイニシアティブ）の反応係数図（産業毎）

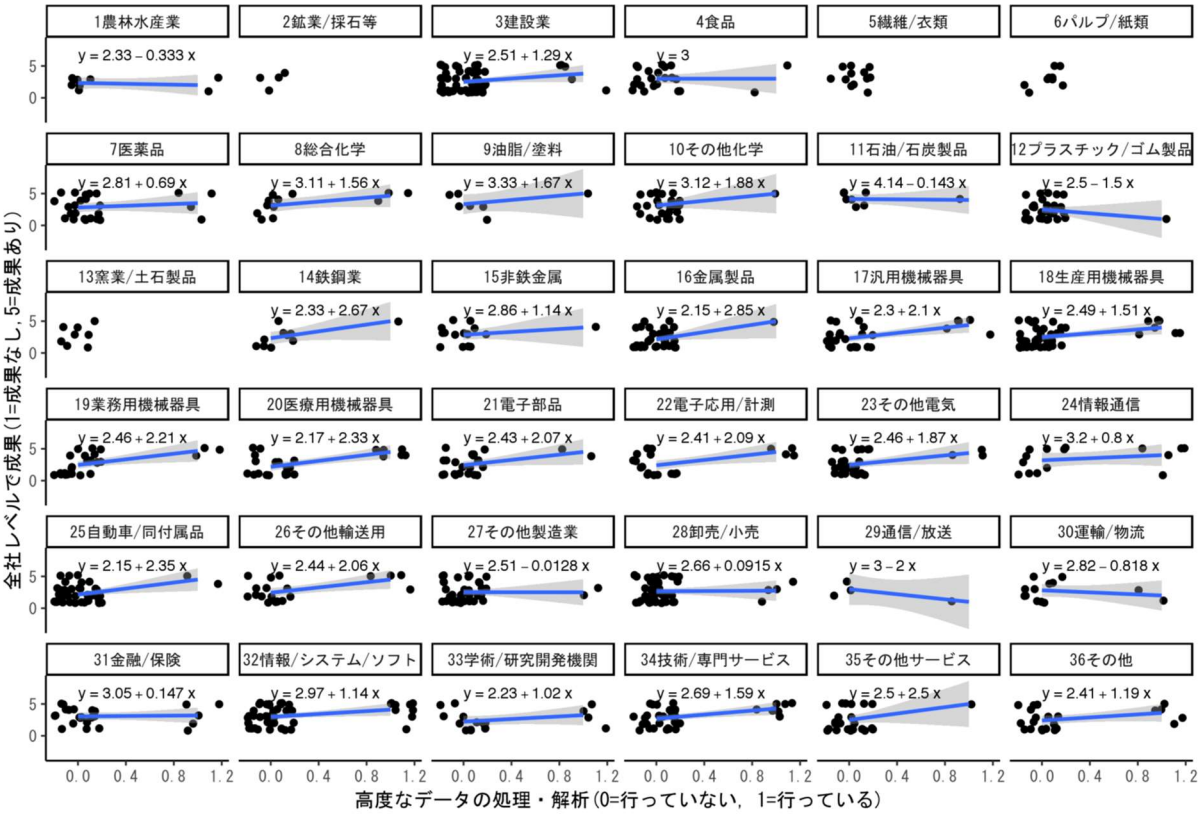

図 79 問 2-7（データ利活用成果）と問 3-18（高度なデータの処理・解析）の反応係数図（産業毎）

---

[i] 調査の詳細は、下記の文献を参照されたい。
渡部俊也, 平井祐理, 阿久津匡美, 日置巴美, & 永井徳人. (2018). 企業において発生するデータの管理と活用に関する研究. RIETI Discussion Paper Series 18-J-028.
< https://www.rieti.go.jp/jp/publications/dp/18j028.pdf >（2018 年 11 月 1 日最終アクセス）

[ii] 多重比較法では、群 1、群 2、…、群 t の群平均$\mu_1, \mu_2, \cdots, \mu_t$に対して、

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_t$$

H_1: 群平均が少なくとも 1 つは異なる

というように帰無仮説と対立仮説を設定し、仮説検定を行う。この時、t 検定などを用いて二群間の平均の差の検定を繰り返し行うと、第 1 種の過誤（偽陽性）の確率が高まる。

例えば、3 つの群が存在する時、第 1 群-第 2 群間（$\mu_1 = \mu_2$）、第 2 群-第 3 群間($\mu_2 = \mu_3$)、第 1 群-第 3 群間($\mu_1 = \mu_3$)について、p 値=0.05 を危険率として各々の群平均に差があるかを t 検定すると、「少なくとも 1 つが有意差あり」となる確率は$1-(1-0.05)^3$で計算され、p 値≒0.14 となる。帰無仮説H_0を 5%の危険率で検定したいのに、各群間での比較をくりかえすと、危険率は 14%に上昇し、第 1 種の過誤が生じやすくなってしまう。このため、多重比較では、何らかの補正を行う。頻繁に使われる多重比較の補正法としてチューキー・クラマー法（Tukey-Kramer 法、以下チューキー法と呼ぶ）がある。本研究もチューキー法を用いて多重比較を行う。

なお、本研究では①クラスカル・ウォリス（Kruskal-Wallis）検定を用いて群間に差が生じているかを推定し、②チューキー法による多重比較検定し、③CLD 図(compact letter plot)によって多重比較の結果を確認する、という 3 ステップをとっている。①で用いたクラスカル・ウォリス検定と②で用いたチューキー法による多重比較検定は、既存研究で頻繁に用いられているが、両検定は厳密には対応していない。そのため、ステップ①で群間差があると報告されたとしても、ステップ②では群間差がある群ペアを検知しない事がある。本研究では既存研究との親和性を重視し、①でクラスカル・ウォリス検定と②でチューキー法による多重比較検定を用いる。