

学術分野における論文および統計学論文の 引用状況について

張 菱軒¹・潘 建興²・中野 純司^{3,4}

(受付 2019 年 6 月 21 日；改訂 12 月 22 日；採択 2020 年 1 月 6 日)

要 旨

ビッグデータ分析や機械学習などの出現により、統計学は近年大きな注目を集めている。そして学術論文においては、これまでもデータを正しく分析し新しい知見を裏付けるために統計学が広く使われてきた。ただ現代社会においては非常に多くの学術分野があり、それらの間の競争はますます激しくなっている。そのような競争の中で統計学が生き残るためには、他の学術分野論文への統計学論文の影響を客観的に測定することによって統計学の重要性を示すことが重要である。本研究では、各学術分野内の論文引用状況とそこでの統計学論文の引用状況进行分析する。そのために学術論文データベース Web of Science を利用して学術分野を定義し、分析に必要な引用数を集計した。

キーワード：論文引用解析、学術分野、Web of Science.

1. はじめに

統計学は、データの分析、解釈、生成過程のモデリングおよび推論を行う分野である。そのため、統計学はデータを用いるほとんどの分野で利用される。学術論文においては実験データの客観的な処理のために利用される。また、実用的にも広く利用されている。例えば、工学分野では、実験計画法や品質管理の統計手法を使って、生産物の品質の維持・改善が行われる。薬学では新薬の効果を測るために、臨床試験の結果を統計解析してその効果を分析する。マーケティングでは、A/B テストの結果が統計解析され、広告などの改良に利用される。金融分野や電子商取引の購入動向の分析のためには時系列解析がよく用いられ、予測や行動決定に利用される。

近年、研究分野間の競争は激しくなっており、そのため Institutional Research と呼ばれる分野で学問業績の客観的評価が重要になっている。したがって、統計学がどのように利用されているかを定量的に分析することは統計学にとって重要である。ただ、これまで他学術分野において統計学がどれくらい利用されているかを定量的に調べた研究は少ない。学術論文の影響を調べる一つの方法は、引用論文の状況を調べることである。当然のことながら、論文は自分と同じ分野の論文を多く引用する。論文の目的は新しい知見を発表することであり、そのために

¹ 総合研究大学院大学 複合科学研究科統計科学専攻：〒190-8562 東京都立川市緑町 10-3

² 中央研究院 統計科学研究所：〒11529 台北市南港區研究院路二段 128 號

³ 中央大学 国際経営学部：〒192-0393 東京都八王子市東中野 742-1

⁴ 統計数理研究所：〒190-8562 東京都立川市緑町 10-3

はこれまでその分野で知られていることを共有することが必要で、それには同分野の論文を引用しなければならないからである。しかしながら、異なる分野の論文が引用論文に含まれることも多い。その理由の一つは、同じようなテーマが複数の分野にわたって研究されていることがあるからである。特定のテーマの論文を検索すると、複数の分野の論文にたどり着くことがある。特に新しい分野はその傾向が強く、例えば機械学習は、コンピュータサイエンスと統計学の両方の知識が必要な手法であり、両分野の論文が機械学習の論文に多数引用される。このように、学術分野間には複雑な関係があり、その境界は曖昧であることも多い。

学術分野の引用情報の解析はこれまでもいくつか行われている。特に統計学分野の学術誌における引用情報の解析は Varin et al. (2016)で行われている。多くの学術分野に対しては、Leydesdorff (2004), Zhang et al. (2010)のような解析が行われている。

本論文では、学術分野の論文が統計学論文を引用する頻度に着目する。それはその分野でどのくらい統計が必要とされるかを示すことになるからである。また、その値を指標として基準化するために分野内における引用状況も調べる。分野の分類と引用-被引用関係の集計を行うために、学術論文データベース Web of Science (WoS) を利用した。また解析のために、統計解析システム R (R Core Team, 2019)を使用した。

本論文の構成は以下の通りである。まず、2 節で WoS とそこでの分野分類について説明する。3 節で各分野における論文引用状況を調べる。次に 4 節で、各分野における統計学論文引用状況を調べる。最後に 5 節でまとめと注意を述べる。

2. Web of Science と学術分野

Web of Science (WoS, 2018)は、論文情報、著者情報、および学術誌情報などの学術情報に関する商用データベースであり、Clarivate Analytics 社により開発・維持されている。われわれは 1981 年から 2016 年までの WoS データを用いる。

WoS では各学術誌に 1 つまたは複数の分野が割り当てられている。ただし、140 の学術誌にはどの分野も割り当てられていない(例えば、International Review of Connective Tissue Research)。これらの分野のうち明らかな重複(例えば Legal Medicine と Medicine, Legal)と学術誌名の明らかな重複(例えば 2D Materials と 2D MATERIALS)を除く処理などを行うと、分野の数は 266 であり、全体で 19138 の学術誌と 45769924 の論文が含まれる。

割り当てられている分野数とそれに対する学術誌数および論文数は表 1 で示される。われわれが興味のある統計学は Statistics & Probability として分類されているが、これが単独で割り当てられている学術誌はなく、159 の学術誌がこれを割り当てとして含む。それに対して学術誌数と論文数を見ると、表 2 のようになる。Statistics & Probability を含む 2 つの分野が割り

表 1. 複数分野に割り当てられる学術誌数と論文数.

割り当てられる分野数	1	2	3	4	5	6	7	8	9	10
学術誌数	6792	6794	3062	1474	600	188	55	23	7	4
論文数	16270044	16674268	6618622	3818521	1540125	492942	183419	48700	25634	38414

表 2. 統計学に割り当てられる学術誌数と論文数.

割り当てられる分野数	2	3	4	5	6	7	8	9
学術誌数	74	20	17	14	13	1	3	2
論文数	104229	14890	26234	17769	20204	1158	13817	4997

当てられている 74 の学術誌はすべて Statistics & Probability と Mathematics に割り当てられている。なお、Mathematics だけに割り当てられている学術誌数は 220 である。

われわれは統計学分野に属する論文としては Statistics & Probability が割り当てられている学術誌に掲載された論文をすべて含めることにする。同様に他分野の論文として、その分野が含まれた学術誌に掲載された論文をすべて考える。そのため、例えば学術誌が 5 個の分野に割り当てられていれば、それに掲載された 1 つの論文が 5 個の分野で論文として集計される。

集計のために、われわれは 1981 年から 2016 年までの WoS データを用いて統計数理研究所で構築されたネットワークデータベースを利用する。このネットワークデータベースには 1981 年より前の論文は含まれていないので、例えば 1981 年の論文から引用される論文数はほとんど 0 と集計されることに注意する。

3. 各分野における論文引用特性

本節では各分野ごとの論文引用の状況をいくつかの指標を用いて分析する。最初に、各分野における 1 論文あたりの被引用数を見定める。

$$\text{平均被引用数} = \frac{\text{被引用論文数}}{\text{総論文数}}$$

図 1 は分野ごとの平均被引用数を多い順に示したものである。図 1 は細部が見にくいので、図 2 では特に興味のあるランキングの上位 20 分野と下位 20 分野を拡大して示した。ひとつの論文が際立って多く引用されている分野は Astronomy & Astrophysics で平均 19.08 回引用されている。そのほかの分野では多くても 13.16 回である。Statistics & Probability の平均被引用数は 5.42 回である。この値はかなり少ないが、その理由は利用できるデータが 1981 年から 2016 年までのものなので、1980 年以前の論文を引用しても、それは引用回数としては数えられないからである。従って平均被引用数が上位の分野は同分野の論文を多く引用し、かつ新しい論文をよく引用し、古い論文をあまり引用しない分野と考えられる。下位 20 分野はほぼ文学・芸術関係の分野であり、平均被引用回数は 1 以下である。これらの分野では過去の論文を多く引用するか、論文をあまり引用しないと考えられる。

次に 1 回以上引用されている論文の割合を見る。

$$1 \text{ 回以上引用されている論文の割合} = \frac{1 \text{ 回以上引用されている論文数}}{\text{総論文数}}$$

図 3 は、分野における 1 回以上引用されている論文の割合を示し、図 4 はランキングの上位 20 分野、下位 20 分野を拡大して示したものである。この値は孤立していない論文の割合を示している。半分以上の分野で半分以上の論文が少なくとも 1 回引用されている。ここでも Astronomy & Astrophysics の値が最も高く、84% の論文が少なくとも一回引用されている。やはり文学・芸術関係の分野は孤立した論文が多いようである。

また、図 2 と図 4 の上位 20 分野を見ると、順位は少し異なっている。図 2 では、Neurosciences, Neurosciences & Neurology, Biochemistry & Molecular Biology, Virology, Geochemistry & Geophysics, Management と Psychology が上位 20 分野に入っているが、図 4 には入っていない。これは論文の引用頻度は多いが、それに較べて孤立している論文が多いことを示している。逆に図 4 では Chemistry, Inorganic & Nuclear, Parasitology, Polymer Science, Chemistry, Analytical, Fisheries, Oceanography と Materials Science, Biomaterials は上位 20 分野に入ったものの、図 2 には入っていない。これは論文の引用数の割には、孤立した論文が少ないと考えられる。

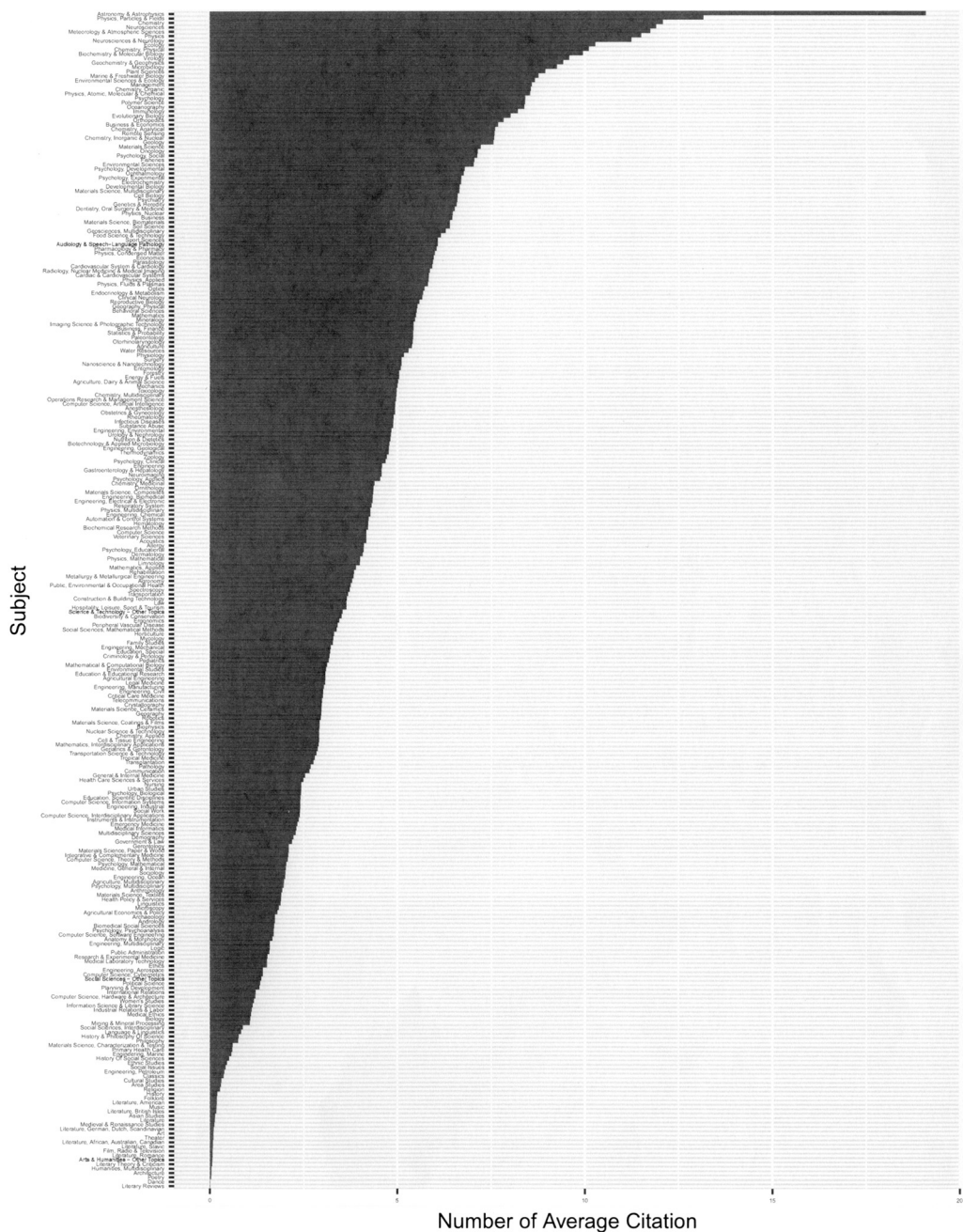




図 2. 平均被引用数の上位 20 分野と下位 20 分野.

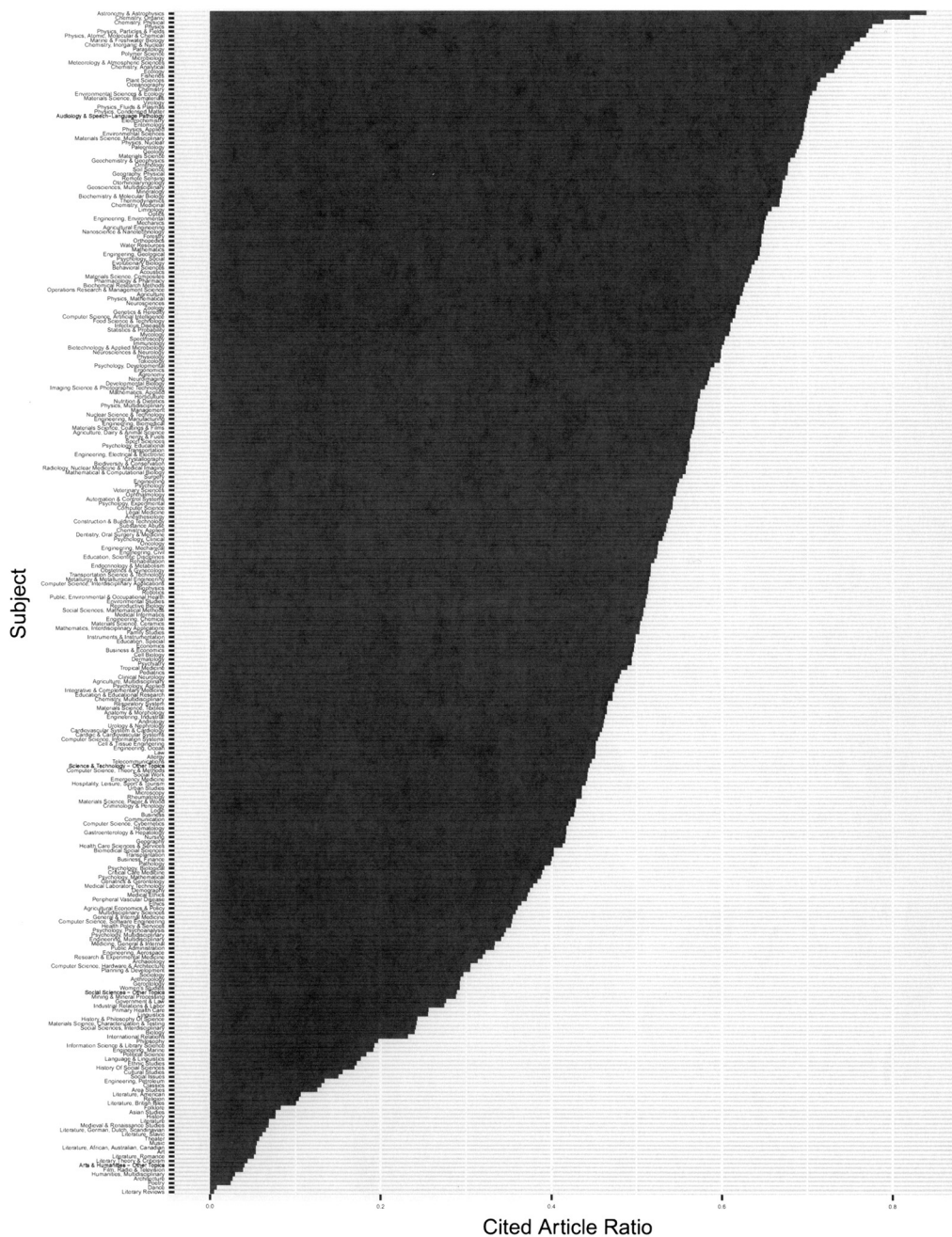


図 3. すべての分野における 1 回以上引用されている論文の割合.



図 4. 1 回以上引用されている論文の割合が上位 20 分野と下位 20 分野.

4. 各分野における統計学論文引用特性

最初に、各分野で 1 論文当たり何件の統計学論文が引用されているかを調べる。

$$\text{統計学論文平均被引用数} = \frac{\text{被引用統計学論文数}}{\text{総論文数}}$$

これはその分野で統計がどれくらい使われているかの一つのわかりやすい指標である。

図 5 は、すべての分野における 1 論文当たりの統計学論文被引用数である。図 6 はそのランキングの上位 20 分野、下位 20 分野を示した。Social Sciences, Mathematical Methods が最も多く統計学論文を引用している分野であり、一つの論文は平均的に 2 本以上の統計学論文を引用している。上位 20 分野の中では 6 分野が数学関連の分野である。下位 20 分野は文学・芸術関係の分野である。なお、近年注目されている Computer Science, Artificial Intelligence は 9 位であり、この分野で統計的手法がよく利用されていることがわかる。

しかしながら明らかに、この指標は各分野の平均論文被引用数に影響をうける。すなわち、もともと引用論文数が少ない場合は引用される統計学論文数も少ない。それで各分野の論文被引用数で基準化する。

$$\text{引用論文の中での統計学論文の割合} = \frac{\text{統計学論文被引用数}}{\text{論文被引用数}}$$

図 7 は、すべての分野における統計学論文引用率である。図 8 はランキングの上位 20 分野、下位 20 分野を示した。この場合、Mathematical & Computational Biology が 1 位となり、引用論文のうち約 80% が統計学論文である。また、図 6 の上位 20 位に入っていない Biology, Computer Science, Cybernetics, Research & Experimental Medicine, Industrial Relations & Labor が図 8 で上位 20 分野に入った。図 6 では上位 20 分野に入った Business & Economics, Evolutionary Biology, Automation & Control Systems と Mathematics, Applied は図 8 では入っていない。これらの分野では統計学論文を引用してはいるが、それは自分の分野の論文と比べるとそれほど多くないということを示す。

次に、どれくらいの異なる統計学論文が引用されているかを考える。すなわち分野ごとに、1 回以上引用された統計学論文数を総論文数で割って基準化した値を考え、図 9, 図 10 に示す。

図 10 において、Computer Science, Cybernetics, Computer Science, Information Systems, Ergonomics と Management は上位 20 分野に入っているが、図 6 の上位 20 分野には含まれない。それはこの二つの分野では多くの異なる統計学論文を引用しているが、引用数自体はそれほど多くないことを示す。逆に、Economics, Business & Economics, Evolutionary Biology と Biochemical Research Methods は図 6 の上位 20 分野に入ったものの、図 10 では圏外になる。それは引用数は多いが、引用している論文の種類は多くないことを示しており、引用されている統計学論文が特定の論文に集中する傾向があることを示している。

次に基準化を行うときに、総論文数ではなくその分野で 1 回以上引用されている(孤立していない)論文数を利用する。この指標を考える理由は、孤立した論文はその分野における傍流あるいは例外と考えられるので、それらと統計学論文の関係を考慮する必要はないと思われるからである。

図 11, 図 12 はその値を示す。ほとんどの分野では、同分野内の論文を引用することが多いのでこの指標の値は 1 より小さい。しかし、Social Sciences, Mathematical Methods だけは 1 より大きく、統計学論文の比重が大きいことがわかる。また、上位 20 分野において、やはり 6 分野が数学関連分野である。図 8 では上位に入っていない Biology, Demography と Computer Science, Software Engineering がここでは上位 20 分野に入っている。この分野では引用論文の

数が少ないが、その割には引用されている統計学論文の種類が多いことがわかる。図 10 では上位に入った Computer Science, Ergonomics と Management はここに入っていないので自分の分野を論文引用数と比べて、統計学論文の引用種類は多くないことを示す。また図 8 では上位に入っていない Automation & Control Systems, Demography, Computer Science, Information Systems, Computer Science, Software Engineering と Mathematics, Applied がここでは上位 20 分野に入っている。この分野では統計学論文の引用数は多くないが、引用する統計学論文の種類は多い事がわかる。逆に、図 8 では上位に入った Economics, Computer Science, Research & Experiment Medicine, Industrial Relations & Labor と Biochemical Research Methods は図 12 では入っていない。それは引用される統計学論文が集中する傾向があることを示す。

5. Conclusion

本論文では、最初に分野における引用状況を調べた。さらに特定の分野 (Statistics & Probability) がその他の分野に引用される状況を調べた。その結果、統計学論文が他分野に与える状況が数値的に示された。

ここでは、統計学分野と他分野の関係を調べたが、同様の分析は他のすべての分野に関しても行える。特に数学、物理学のような基礎学術分野 (統計学もその一つである) を分析することは興味深い。そして、このような分析は各分野が他の分野とどのように相互作用しているかを見ることにもなる。ここで計算したような指標は将来の共同研究の可能性や潜在的な異分野融合の指標として使用することもできる。明らかに、高い相互引用は分野間の強い関係を示すからである。また、論文の異分野融合の評価の指標として使う事も可能であろう。

謝 辞

本論文で使用したデータは Clarivate Analytics から提供されたものである。また、統計数理研究所の URA 室の本多啓介博士と濱田ひろか氏はデータを neo4j データベースに変換してくれた。それを使うことで本論文の分析が可能となった。栗木哲教授にはこの研究を遂行するためのよい環境を整えていただくとともに、研究上の助言もいただいた。査読者の方からは有益なコメントを頂いた。非常に感謝している。

参 考 文 献

- Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in journal citation reports, *Journal of Documentation*, **60**(4), 371–427, DOI: 10.1108/00220410410548144.
- R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, <https://www.R-project.org/>.
- Varin, C., Cattelan, M. and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals, *Journal of the Royal Statistical Society, Series A*, **179**(1), 1–63.
- Web of Science (2018). Clarivate Analytics, <http://www.webofknowledge.com/>.
- Zhang, L., Glänzel, W. and Janssens, F. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research, *Scientometrics*, **82**(3), 687–706, DOI: 10.1007/s11192-010-0180-1.

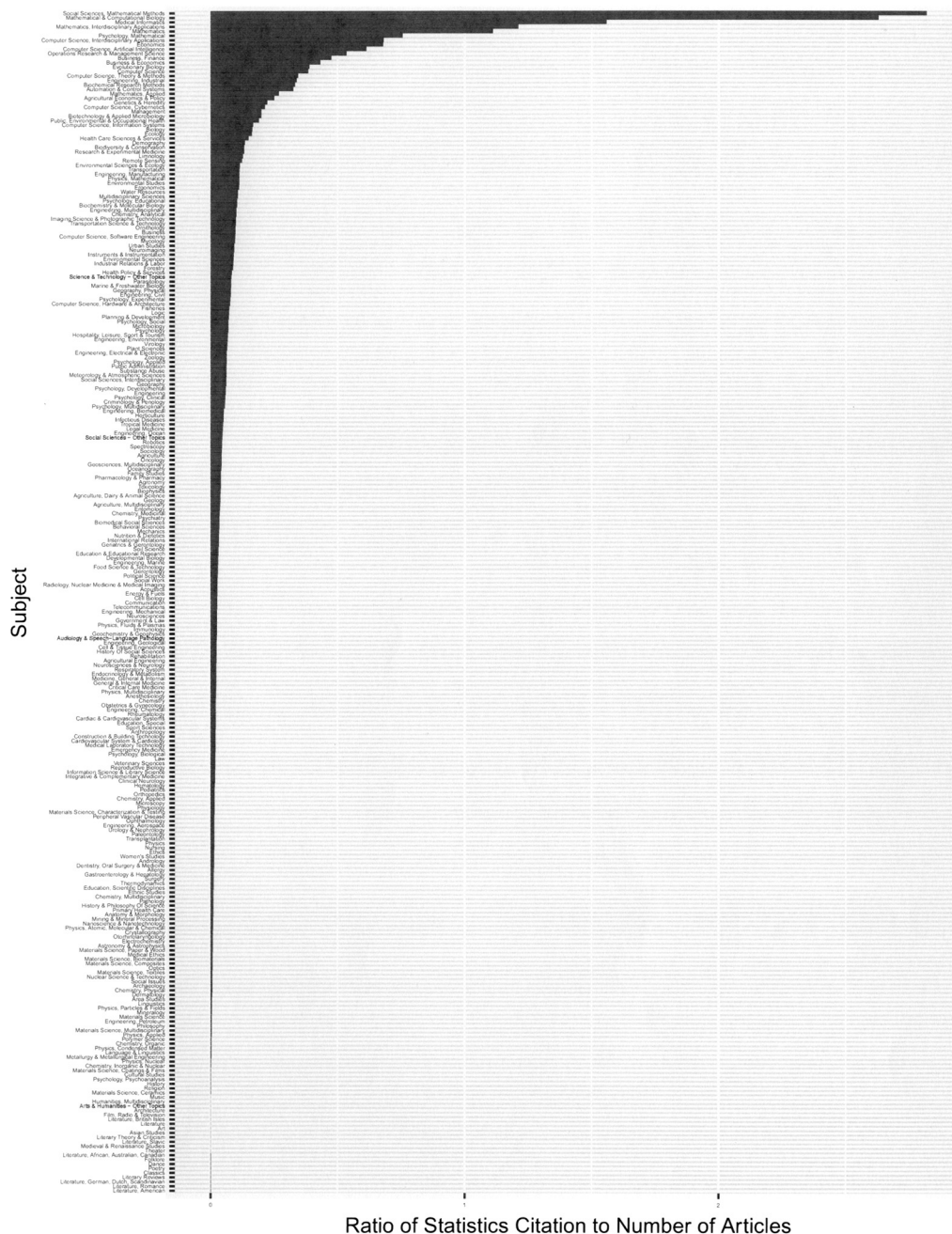


図 5. すべての分野における統計学論文平均被引用数.

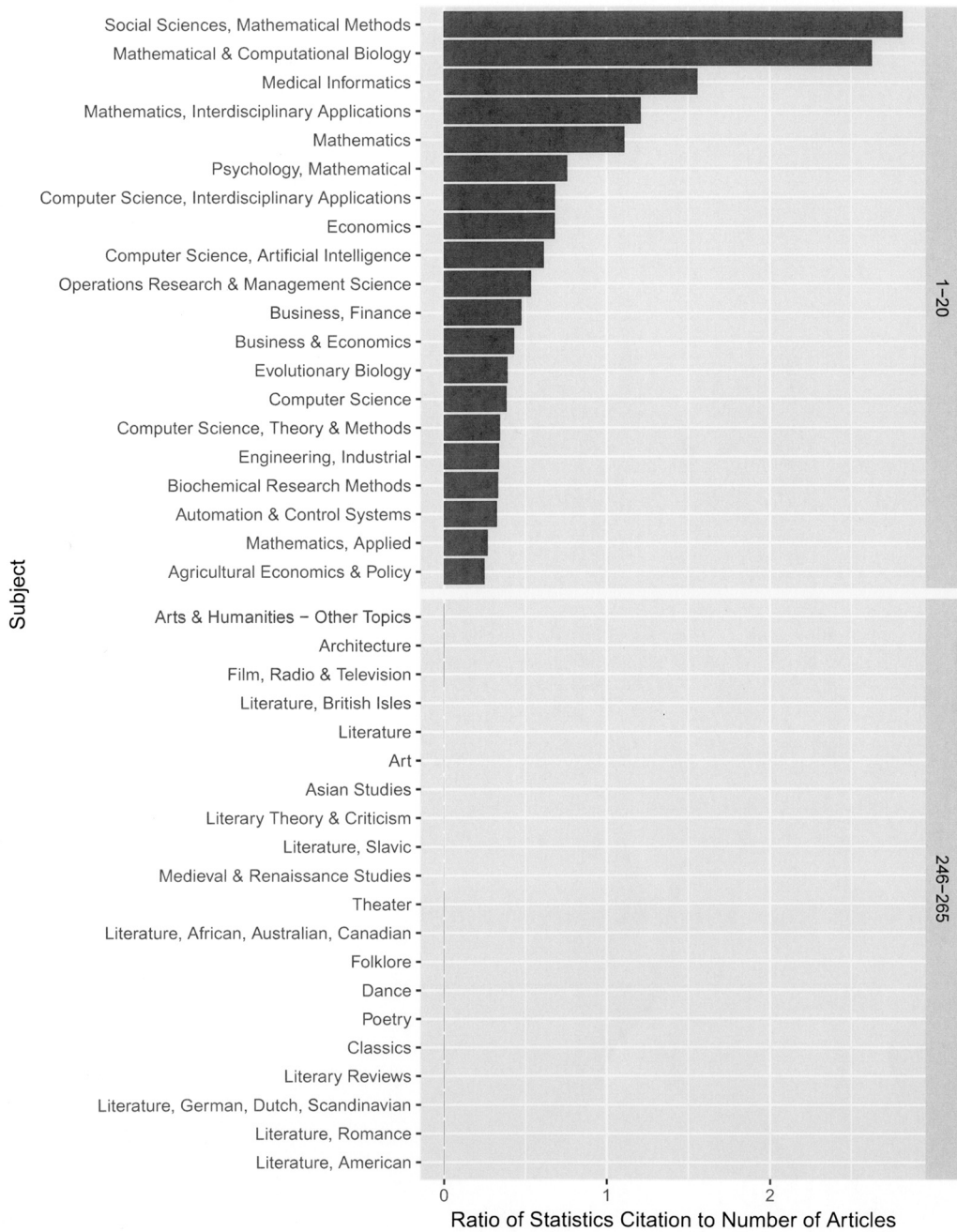


図 6. 統計学論文平均被引用数が上位 20 分野と下位 20 分野.



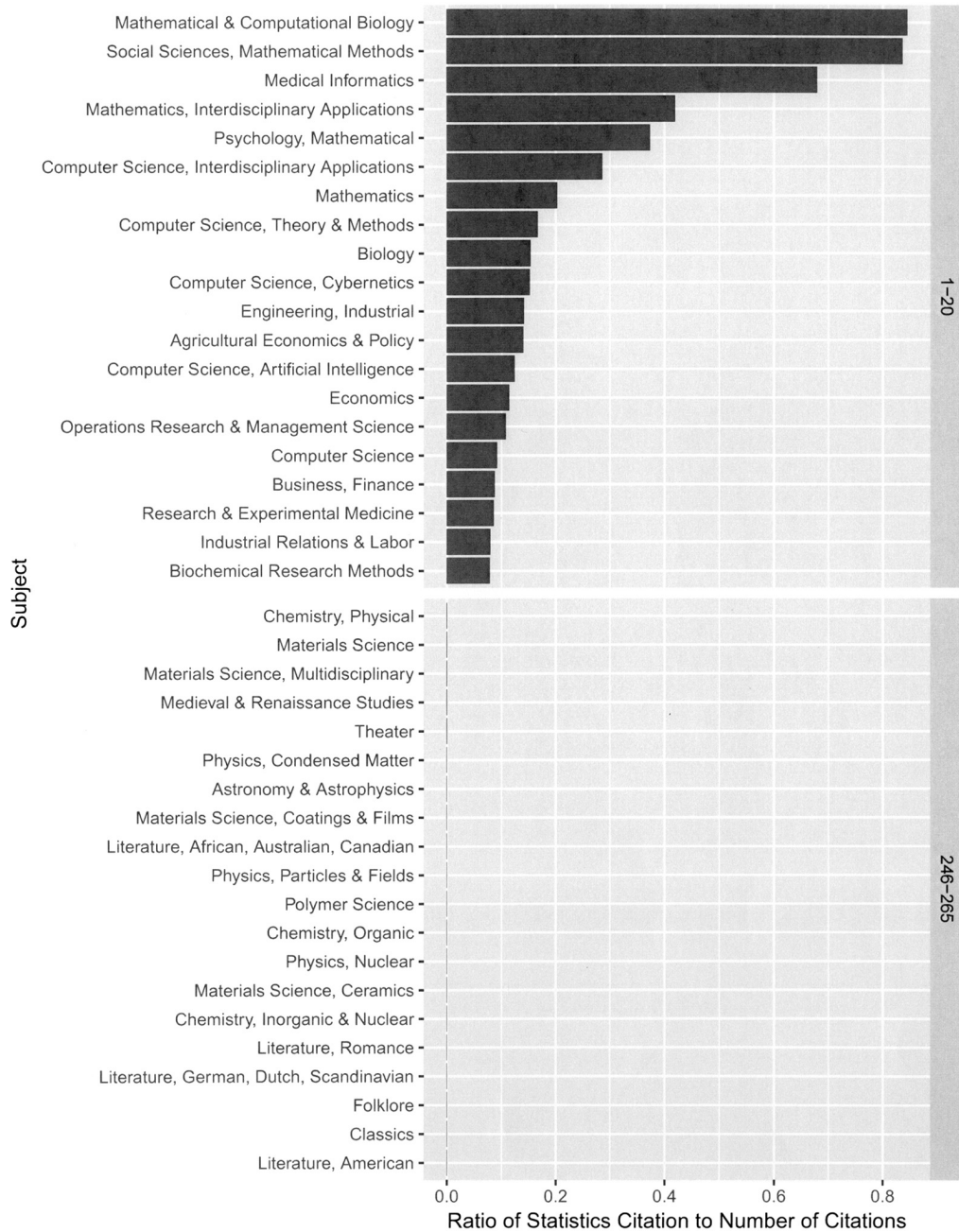


図 8. 引用論文の中での統計学論文の割合が上位 20 分野と下位 20 分野.

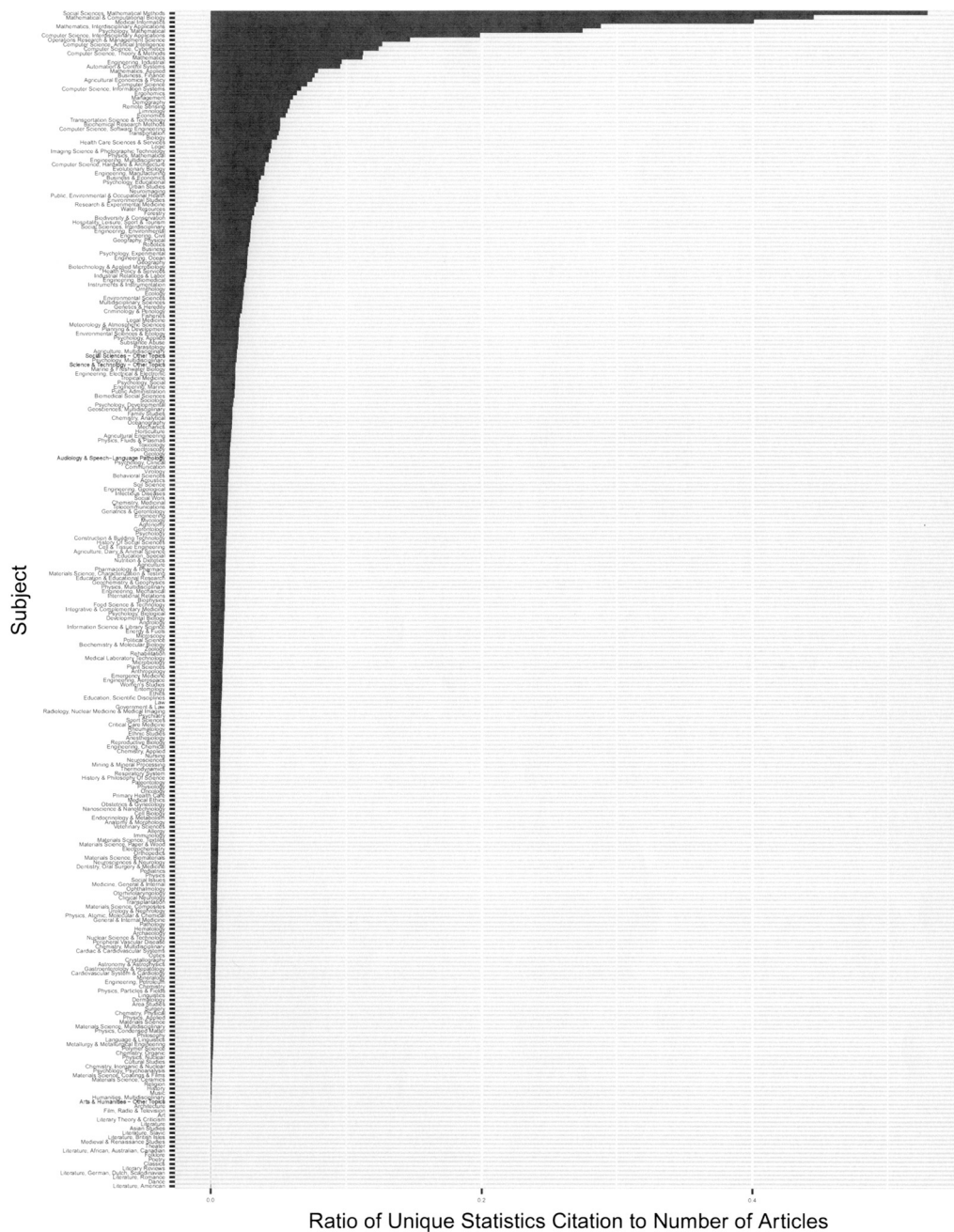


図 9. すべての分野における, 1 回以上引用された統計学論文数(各分野の総論文数で基準化).

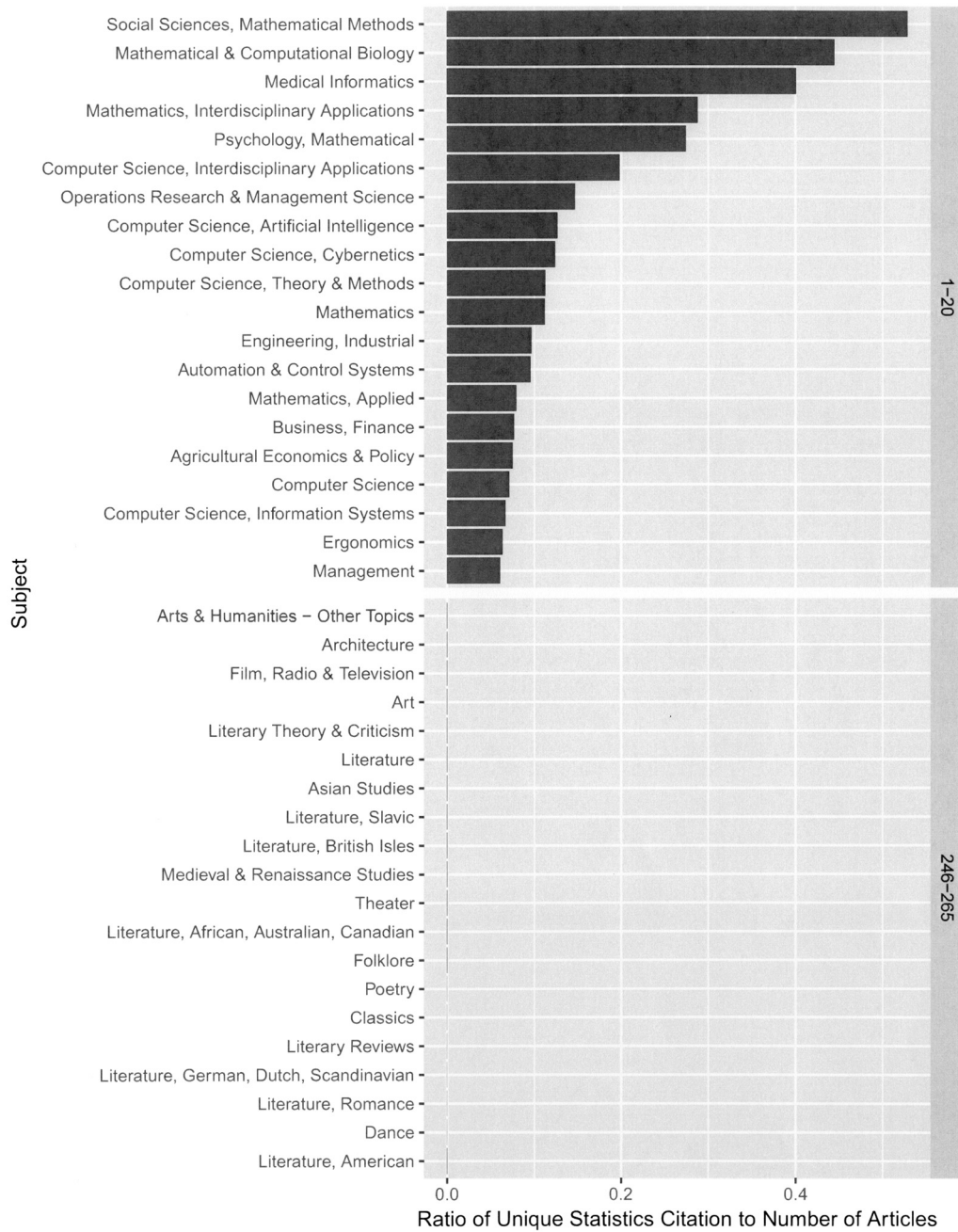


図 10. 1 回以上引用された統計学論文数(総論文数で基準化)が上位 20 分野と下位 20 分野.

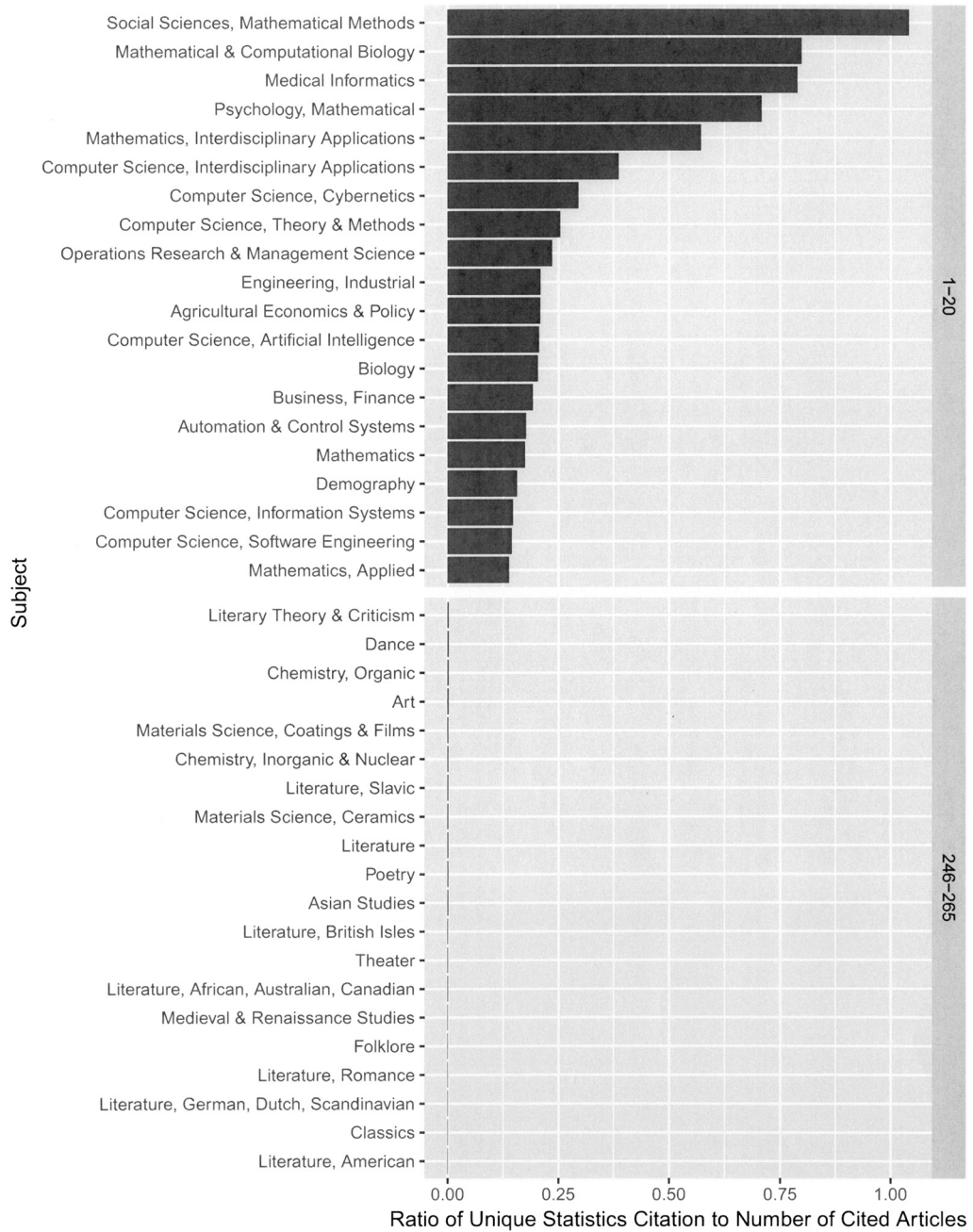


図 12. 1 回以上引用された統計学論文数(1 回以上引用されている論文数で基準化)が上位 20 分野と下位 20 分野.

Citations of Academic Articles and Statistical Articles in Fields of Sciences

Livia Lin-Hsuan Chang¹, Frederick Kin Hing Phoa² and Junji Nakano^{3,4}

¹Department of Statistical Science, School of Multidisciplinary Sciences,
Graduate University for Advanced Studies

²Institute of Statistical Science, Academia Sinica

³Department of Global Management, Chuo University

⁴The Institute of Statistical Mathematics

Statistics has obtained more attention in recent years due to the rise of big data analysis and machine learning. Statistics are widely used in academic studies that require statistical analysis to objectively support their conclusions. In modern society, there exist many academic fields, and competition among them is severe. In order for statistics to survive such competitions, it is important for statisticians to measure the influence of articles in the field of statistics relative to those in other academic fields. In this work, we analyze citations within each academic field, focusing on citations of statistical articles. We used a database of academic articles from “Web of Science” to define academic fields and to count the required numbers of citations in the study.