交通計画の学術文献における統計的有意性の使用と誤用 —離散選択モデルを中心として—

パラディ　ジアンカルロス[1]・アックスハウゼン　カイ[2]

[1]正会員　東京大学大学院　工学系研究科（〒113-8656 東京都文京区本郷 7-3-1）
E-mail: gtroncoso@ut.t.u-tokyo.ac.jp

[2]非会員　ETH Zurich Institute for Transport Planning and Systems (Stefano-Franscini-Platz 5)
E-mail: axhausen@ivt.baug.ethz.ch

本研究では，2014 年から 2018 年にかけて離散選択モデルを中心として交通計画の学術文献をレビューし，実証分析における統計的有意性の使用と誤用を評価した．結果として，39%の研究は係数の符号のみに基づいてモデル結果を説明しており，67%の研究は結論で統計的有意性を経済的，政策的，科学的重要性と区別していない．さらに，レビューした研究の中で検定の統計的検出力を考慮したものはなかった．これらの結果に基づいて，統計的有意性から効果量の適切かつ包括的な評価へと焦点を移すように改善策を提案する．

Key Words: *discrete choice models, effect size, statistical significance, statistical power, policy-making*

1. はじめに

交通計画の学術研究の究極目的は，交通に関する人間行動をよりよく理解し，交通政策立案とその実施に貢献することである[1]．統計モデルは，安価な計算力により，交通関連現象を説明するために重要な道具となっている．しかし，定量的分析に頼る多くの分野において，統計モデルの普及とともに，任意の要因の実質的な重要性を考慮せず排他的に統計的有意性による評価というレシピ的な使い方も広まった．McCloskey & Ziliak[2]は，経済学の分野において，統計的有意性の度重なる誤用を明らかにした．しかし，これは決して経済学に限ることなく，交通計画の分野においても同様な誤用は少なくないであろうが，体系的に評価されていない．本研究では，McCloskey and Ziliak[2]の 19 の質問を離散選択モデル中心に交通計画分野の学術文献に適用し，実証分析における統計的有意性の使用と誤用を評価する．本研究は，モデルの検証に関する Parady, Ory & Walker[1]の研究を補完するものであり，当分野におけるより正確なモデリングの実践の推進を目指す．

2. McCLOSKEY と ZILIAK の主な結果

経済学における偏りのないベストプラクティス集として，McCloskey and Ziliak[2] は，1980 年代に American Economic Review 誌に掲載された回帰分析を用いたすべての論文をレビューした．対象論文 182 本に対して，統計的有意性の使用における 19 の質問を問うた．各質問に対して，論文を３つに分類した．健全な統計学の実践の場合「はい」，不健全な統計学の実践の場合「いいえ」，そして該当しない場合「該当なし」と分類した．主な結果は以下の通りである：

1) 70%の論文は，統計的有意性と経済的，政策的，科学的な重要性を区別していない．
2) 3 分の 1 の論文は，論文に変数を入れるための基準として，t-統計量と F-統計量のみを使用している．
3) 72%の論文は，係数の規模が「大きい」か「小さい」かと判断できるように科学的な根拠について議論していない．
4) 59%の論文は，「有意」という言葉を，あるとき帰無仮説と統計的に異なるという意味で，あるとき実質的に重要であるという意味で，曖昧に使っている．
5) 32%の論文は，モデルから変数を除外するための唯一の基準として，統計的有意性を利用したことを明示している．
6) 検定の統計的検出力を考慮した論文は，わずか 4%であった．

7) 69%の論文は，モデルに含めた変数の記述統計を報告していない.

3. 交通計画の学術文献における統計的有意性の使用に関する評価

(1) 論文選定基準

本研究では，現状をより包括的に把握するため，この分野での「ベストプラクティスの集」ではなく，分野全体にレビューの範囲を広げた． Clarivate Analytics が管理する Web of Science Core Collection を利用して，2014 年から 2018 年の間に掲載された離散選択モデルを用いた交通計画における学術論文をレビューした． 選定基準は Parady, Ory and Walker[1]に従ったため，モデル検証に関する彼らの知見を補完するものである． 主な違いは，本研究で表明選好調査（SP 調査）を用いた論文をレビューから除外しなかったことである． 具体的に以下の基準に基づいて論文を選定した：

1) 2014 年から 2018 年にかけて掲載された査読付き論文
2) 離散選択モデルを用いた論文
3) 目的地選択，手段選択または，経路選択を対象とした論文
4) 他の行動を対象とした場合は，3）の対象行動のうち少なくとも 1 つを対象とした論文
5) Web of Science データベースの検索キーワード：「Destination choice」，「Mode choice」，「Route choice」
6) Web of Science データベースの分野：Transportation, Transportation science and technology, Economics, Civil engineering
7) 陸上交通と日常的な交通行動を対象とした論文（観光，避難行動，貨物輸送を対象とした論文は対象外）
8) 数値シミュレーションのみを用いた研究ではない論文
9) 理論的な論文の場合は，実証分析を含めた論文.（ただし，新たな基準として，政策に関連する変数を明示的に対象としなかった論文は除く）
10) 離散選択モデルはより大きなモデルの下位要素に過ぎない確率的利用者均衡（SUE）モデルに関する論文は除く.

レビュー対象の論文数は，該当する全 283 本から無作為に抽出された 95 本であった． これは，該当する全論文の 34%を占める.

⑵ 交通分野における 15 の質問

本研究は McCloskey と Ziliak[2]の調査に基づいているが，交通分野の特殊を反映させるために， 質問項目を変更した． また， いくつかの質問では， 「はい」と「いいえ」の二択を拡大し，部分的に健全な統計学の実践も考慮する． また， 可能な限り評価基準を一致させるようとしたが， いくつかの設問において評価方法は明確ではなかったため， 評価基準が異なる可能性がある． 次節で比較する際には， 留意すべきである.

質問項目は， 著者の経験に基づき， 事前に作成した質問項目から開始し， 該当論文集合からランダムに選定した論文に対してプレテストを行い， 反復的に質問を修正し， 以下の 15 の質問項目を決定した．レビュー対象の最終論文集合は，プレテストで利用した論文 29 本は含まない．比較のため， McCloskey and Ziliak の 19 の質問項目を MZ-1, … , MZ-19 と表記する.

Q1. モデルに使う変数の記述統計や単位を報告したか.

この質問は MZ-2 と同等であり，モデルの結果を適切に解釈するために，変数の単位と記述統計（少なくとも，量的変数の平均値と質的変数の相対頻度）の情報が必要である． 論文を 「網羅的に報告している」，「部分的に報告している」及び「全く報告していない 」と分類した.

Q2. 「効果がどの程度大きいか」の問に応える弾力性，限界効果，または他の関心のある指標を報告したか.

この質問は MZ-3 に相当するが，回帰分析と異なり，係数を直接に解釈できない離散選択モデルにおいて，より重要である． 交通計画分野では，弾力性や限界効果に加え，限界代替率（交通時間価値など）も通常に報告される． 厳密な意味で限界代替率は効果量ではないが，政策に対して重要な指標であるため，対象範囲に入る.

論文を 3 つに分類した． 理想的な状態であり，論文中のほとんどの変数，または著者による主要変数であると明示したものについて，効果量を報告した場合は「網羅的に報告している 」と分類した． 必ずしも全ての主要変数ではなく，1 つまたはいくつかの変数について効果量を報告した場合は「部分的に報告している 」と分類しており，それ以外の場合は「全く報告していない」と分類した.

係数の報告は分野の慣例であるため，この質問では，係数の報告の有無を考慮していない． つまり，モデル係数に加えて，すべての変数について効果量を報告した論文と，効果量のみを報告した論文は，どちらも 「網羅的に報告している」と分類した.

Q3. 標準誤差，t-値及び尤度比をすべて報告したか.

この質問は「*有意性の検定が適切かどうかにかかわらず，t-統計量や F-統計量，標準誤差をすべて報告することを回避したか*」を問う MZ-6 の代替質問である． 当分

野では研究の目的に応じて帰無仮説が暗黙に立てられるのが慣例であり，係数，*t-統計量*や適合度統計量を含むモデルの結果が報告されるのが普通である．著者の経験上で，そうでなければ査読者に要求されることが多い．従って、 分野の標準を基づいてモデル報告の仕方を評価する質問に変更したが、これらは効果量や政策評価に直接関係するアウトプットよりも二次的なものであることを強調したい。

Q4. 検定の検出力を考慮したか.

この質問は MZ-8 に相当するものであり，検定の統計的検出力を指す.

Q5. その場合は，検出力をどう扱ったか.

この質問は MZ-9 に相当する.

Q6. 「Asterisk econometrics」と呼ばれる検定統計量の絶対値の大きさによって係数をランク付けすることが回避したか.

この質問は MZ-10 に相当する． これは t-統計量の絶対値の大きさによって係数をランク付けすることを指しているが，当分野では慣例的なものではないため，あまり起こらないと予想している。

Q7. モデル結果の節で，「sign econometrics」と呼ばれる効果量を考慮せず係数の符号を解釈することを回避したか.

この質問は MZ-11 に相当しており、効果量が実質的に重要であるかを考慮せず，係数の符号に基づいて（多くの場合 t-統計量の大きさに加えて）モデルを解釈することを指す． 本研究では，*sign econometrics* の観点からモデルを完全に論じた後，効果量に着目する別の分析を行うことが十分あり得ることを考慮し，この質問の範囲をモデルの結果が最初に紹介される節に限定した.

論文を３つに分類した． 論文中のほとんどの変数について *sign econometrics* を回避した場合「ほぼ避けている」と分類された． 1 つあるいは，いくつかの変数のみについて回避しているが，必ずしもすべての重要変数についてしていない場合は「部分的に避けている 」と分類しており，それ以外の場合は「全く避けていない 」と分類された.

Q8. 効果量について議論したか.

Q9. 効果量に対して，実質的な重要性について判断したか. つまり，実質的に効果のある要因とそうでもない要因を指摘しているか.

この質問は，MZ-12 に関連するものであり，推定した効果に対して，統計的有意性ではなく実質的な重要性に注目するものである． McCloskey and Ziliak の質問では，特に係数について言及しているが，以上述べたように，係数は直接に解釈できないため，この質問では，シミュレーションを含む効果量または他の関心のある指標に着目するあらゆる分析を対象としたうえで、プレテストの結果による 2 つの質問に分けた． Q8 は，弾力性，限界効果，限界代替率など，あるいはシミュレーションの分析結果を用いて，効果量の解釈や比較などの議論があったかどうかを問う． 一方，Q9 はある効果が「大きい」か「小さい」か，または「実質的に重要である」か「実質的に重要でない」かと，任意の基準を基づいて明示的に判断しているかを問うものである． 任意の２つの効果の中で相対的により大きな効果があっても，その効果が実質的に重要でないことがあるため，この区別は重要である． 効果の程度の判断は定量的分析において極めて重要であるが，必ずしも簡単な判断ではない.

この 2 つの質問に対して，論文を３つに分類した．論文中のほとんどの変数，または著者による主要変数であると明示したものについて効果量の議論と実質的な重要性の判断をしている場合はそれぞれ「網羅的に議論している」，「網羅的に判断している」と分類した． 必ずしも全ての主要変数ではなく，1 つまたはいくつかの変数について効果量の議論と実質的な重要性の判断をしている場合は，それぞれ「部分的に議論している」，「部分的に判断している」と分類しており、それ以外の場合は「全くしていない」と分類した.

Q10. 効果量に対する実質的な重要性を判断するための科学的な根拠について議論したか.

この質問は MZ-13 に相当しており、著者が推定した効果量を，既往研究で報告されている値または分野で一般的に認識されている値と比較しているかを問うものである． 少なくとも 1 つの効果量を比較した場合は，「はい」と，それ以外の場合「いいえ」と分類した.

Q11. 統計的有意性のみに基づく変数の選択を回避したか.

この質問は MZ-14 と同じであり，著者が効果量を無視し，統計的有意性のみに基づいてモデルから変数を除外したことを明示しているかどうかを問うものである．ステップワイズ法を用いた論文も 「いいえ」と分類した。ただし、Ziliak & McCloskey[2]と同様に，著者が明示した場合のみ 「いいえ」と分類したため，下限として考えるべきである.

Q12. 推定した効果の妥当性の判断，または効果量をより良く説明するために，シミュレーションを行ったか ．

この質問は MZ-17 に相当しており，政策シミュレーシ

ョンも含む．ただし、効果量を推定するためにシミュレーションが必要であるダミー変数の限界効果や open-form モデルの弾力性などの報告は対象外である．

Q13. 結論と政策への示唆の節では，統計的有意性と経済的，政策的及び科学的な重要性を区別したか

この質問は MZ-18 に相当する．例えば，統計的に有意な変数を結論としてまとめており，その結果から政策提言を推論する論文は，統計的有意性と実質的な重要性を混在しているとして「いいえ」と分類した．なお，質問の範囲は，該当節の中で，モデル結果を直接に言及している部分に限定した．

Q14. 推定，結論，政策への示唆の節では，あるときは「統計的」な意味で，別のときは「政策や科学で重要になる程大きい」という意味で，「有意」という言葉を曖昧に使うことを回避したか．

この質問は MZ-19 と同等であるが、Ziliak & McCloskey[2] と異なり、推定，結論，政策への示唆の節に範囲を限定した．

Q15. 効果量の信頼区間を報告した且つ，統計的有意性の点推定値の代わりではなく，実質的な重要性を解釈するために使用したか

この問は 19 の質問に含まれていないが，振り返して McCloskey & Ziliak が追加すべき質問であると述べた [3]．本研究では，論文を「網羅的に報告している」，「部分的に報告している」と「全く報告していない」の３類に分類した．信頼区間を報告しているが，議論には用いていない場合は「部分的に報告している」と分類した．

(3) 除外した質問

プレテストの段階で次の質問を除外した。MZ-1 は，統計的有意差が，単にサンプル数が大きいことによる結果にならないように，少数のサンプル数を用いたかを問うものである．少数のサンプル数とは何かという疑問を避けるために，この質問を除外した．ただし，その代わりに，使用された最小サンプル数について後述する．

MZ-4 は「適切な帰無仮説を立てているか」を問うものである．研究者の関心のある帰無値はゼロではないのに，ゼロに対して検定するという間違いがよくある [2]．しかし，当分野では，研究の目的から帰無仮説を暗黙に立て，明示しないことが一般的である．例えば，Khan ら [4]は，「物的環境が非動力系の交通手段への影響を評価すること」を目的として分析を行った．このように係数について先験的な仮説を明示することはほとんどない（例外としては、ネステッドロジットモデルのスケールパラメータなど，理論的に定められているパラメータである）．そのため，このような質問はあまり意義的ではないため除外した．

MZ-5 の「*係数の解釈は慎重に行ったか*」については，Q2，Q8 と重複する部分が多いため，除外した．「*最初の使用に統計的有意性を唯一の基準としたか*」を問う MZ-7 と，「*クレッシェンドの後に統計的有意性を唯一の基準とすることを避けたか*」を問う MZ-15 は，Q7～Q10 でカバーできる内容が多いため，除外した．最後に，「統計的有意性が議論を終えるほど決定的であるか」を問う MZ-16 は，曖昧で，評価基準の決定は困難であった上で，その内容も Q 7～Q10 でカバーできるため，除外した．

4. 主な調査結果

主な調査結果は以下の通りであり（**表 1**），括弧内の数値は McCloskey & Ziliak[2]が報告した数値を示す．

1) 67% (MZ:70%)の論文は，統計的有意性と経済的，政策的，科学的な重要性を区別していない．
2) 86% (MZ: 72%)の論文は，ある効果量が「大きい」か「小さい」かと判断できるように科学的な根拠について議論していない．
3) 62% (MZ: 59%)の論文は，「有意」という言葉を，あるときは帰無仮説と統計的に異なるという意味で，またあるときは実質的に重要であるという意味で，曖昧に使っている．
4) 39% (MZ: 53%) の論文は，係数の符号のみに基づいてモデル結果を説明している．
5) 24% (MZ: 32%) の論文は，モデルから変数を除外するための唯一の基準として，統計的有意性を利用したことを明示している．
6) 検定の統計的検出力を考慮した論文はなかった（MZ: 4%）．
7) 効果量の信頼区間を報告し，経済的や政策的な重要性を解釈するために用いた研究はなかった．しかし，7%はこれらの信頼区間を報告しているが，議論に明示的に使用していない．

(1) 効果量の報告と議論について

効果量に関する質問については，望ましい結果が 2 つあった．一つ目は，Asterisk econometrics，すなわち，検定統計量の絶対値によって係数をランク付けすることを完全に回避されていることである．もうひとつは，記述統計の報告率が高くて、79%(MZ は 31%)に上っている．そのうち 65%は網羅的に，14%は部分的に報告している．先述のように，この情報は，モデルの結果を適切に解釈するために不可欠である．また，77%の研究は，通常に報告されている係数，有意性と適合度をすべて報告し

ている．しかし，報告すること自体に反対しないが，議論をこれらより実質的な重要性に注目すべきであることを強調したい．この点に関して，65%（MZ：67%）の研究が効果量の指標である弾力性、限界効果や他の政策に繋がる評価指標を求めて報告しており，そのうち45%が網羅的に，20%が部分的にしている．同様に，64%（MZ：80.2%）の研究が推定した効果量やその他の評価指標について明示的に議論しており、そのうち34%が網羅的に，31%が部分的にしている（差分は四捨五入による）．つまり，著者は，効果量やその他の評価指標を報告するだけでなく，それらに基づいてモデルの結果を議論することである．例えば，Aziz ら[5]は，徒歩と自転車インフラが交通手段選択に与える影響に関する研究で，*「自己弾力性は，自宅と職場の国勢調査区における自転車レーンの整備割合が 1%増加すると，自転車の選択率が 1.13%増加することを示す」*と標準的に報告している．Heinen & Ogilvie[6]は，イギリスのケンブリッジにおける新たなガイドウェイバスの導入効果に関する研究で「*この結果は，例えば，ガイドウェイバスから 4km 離れて住んでいる人は，アクティブトラベル分担率の（20%以上の）増加可能性が，9km 離れて住んでいる人に比べて，60%から 70%程高いことを示す*」と述べている．

なお、Q2 で「はい、部分的に」と分類された論文のうち、多くの論文が交通時間価値だけを報告して、それ以外の変数については係数の符号のみを報告している．実は，前述のように，39%（MZ：53%）の研究は効果量を考慮せず，係数の符号のみに基づいてモデル結果を説明している．ただし、符号のみでモデルを説明した後，効果量やその他の評価指標が妥当かどうか，あるいは政策効果を評価するためにシミュレーションを行った研究もあった．とはいえ，22%の論文は効果量またはその他の評価指標を全く報告していない．つまり，議論を専ら係数の符号と統計的有意性に限っている．

「効果がどの程度大きいか」という質問に関しては，63%の研究が効果量について明確な判断していない．効果量について議論したかを問う Q8 と異なる点として，多くの場合は著者が，効果量を相対的に議論しても，絶対的に「大きい」か「小さい」，または「実質的に重要」か「実質的に重要ではない」という判断まで至っていない．この点については，Khan ら[4]が，「*(0.5 マイル以内の十字路として定義した）ネットワークの接続性が大きな影響を与える：この変数の標準偏差 1 つの増加は，徒歩の選択率を 34%増加させると推定される*」と述べて，さらに「*駐車料金や無料駐車場の有無の変数は，あまり効果がないことが分かった」と述べており*，効果量を明確に判断している．これらから著者の判断により，ネットワークの接続性は実質的に重要な変数であることが明らかである．さらに，注目すべき点として，独立変数の標準偏差 1 つの増加に対する被説明変数の変化率として効果量を推定していることである。説明変数の 1%の変化に対する被説明変数の変化率と定義される通常の弾力性には重要な限界がある．それは，政策変数群の中、1%を変化させるための政策導入コストが変数による異なる（変えにくい変数と変えやすい変数がある）ことであり、それを考慮するため Khan ら[4]がこの効果量の求め方を考案したものである。

de Luca and Di Pace[7]は，交通時間価値の推定値の議論の中で駐車場立地の実質的な重要性の判断を明確にしており，「*イタリアにおける別の研究で報告された推定値と同程度であり，駐車場の立地が極めて重要であることを示している．片道の平均交通費を 3 ユーロと仮定すると，10 分間の歩行時間（時速 4km で約 700m）は，総交通費の半分以上となる.*」と述べている．関心の効果を経済的に解釈し，明確に判断したうえで既往研究と比較して理想的な形である．

しかし，「効果がどの程度大きいか」という問いは，即答できない難問である．「小さい」や「大きい」という概念そのものが定量的に定義し難くて，ある程度慣例が必要かもしれない．検出力に関する代表的な著作の中で，Jacob Cohen[9]*は，「すべての慣例は恣意的であるため，不合理でないことを要求するだけでよい」*と主張した．さらに，単位のばらつきがなく，様々な研究課題や統計モデルに適用できる普遍的な効果量を特徴づける尺度の利用が望ましいと指摘した一方，*「究極のところ，効果量の意味は，埋め込まれた文脈に依存する」と述べた*．したがって，どの程度大きいかという問題に取り組むため，研究の科学的文脈を明確に理解することが必要である．この点について，86%（MZ：72%）の研究は，効果量や関心のある指標が「大きい」か「小さい」かと判断できるように，既往研究で報告された値に踏まえた科学的な根拠について議論していない．Allard & Moura [10] は，研究対象である長距離インターモーダル交通サービスにおける交通時間価値と支払意思額の推定値を既往研究で報告されている値と比較をして科学的な文脈を考慮している。

変数によっては，効果量の判断がそれほど単純ではなく，むしろ経済的に議論することが不可能なものもある．特に潜在的な構成概念の場合は，単位変化や変化率の意味が明確でないため，効果量の判断が難しい．Hess ら[11]は，潜在的態度の構成概念に関してこの問題を取り上げて，構成概念の尺度は意味を持たないため，その変数の変化率に着目することは無意味である．代わりに，すべての人の態度が特定の属性層の態度と同様になったらどうなるかを検討した方が有意義であろうと主張している．

表 1 交通計画分野における 15 の質問に対する答え

研究は…	該当論文数	はい	そのうち	
			網羅的に	部分的に
Q4.検定の検出力が考慮したか	94	0.00	-	-
Q5.その場合は，検出力をどう扱ったか	0	-	-	-
Q15.効果量の信頼区間を報告した且つ，統計的有意性の点推定値の代わりではなく実質的な重要性を解釈するために使用したか	95	7.37	0	7.37
Q10.効果量に対する実質的な重要性を判断するための科学的な根拠について議論したか	95	13.68	-	-
Q12.効果の妥当性の判断，または効果量をより良く説明するために，シミュレーションを行ったか	95	29.47	-	-
Q13.結論と政策への示唆の節で，統計的有意性と経済的，政策的及び科学的な重要性を区別したか	95	32.63	-	-
Q9.効果量に対して，実質的な重要性について判断したか．つまり，実質的に効果のある要因とそうでもない要因を指摘しているか	95	36.84	13.68	23.16
Q14. 推定，結論，政策への示唆の節では，あるときは「統計的」な意味で，別のときは「政策や科学で重要になるほど大きい」という意味で，「有意」という言葉を曖昧に使うことを回避したか.	93	37.63	-	-
Q7. モデル結果の節で，「sign econometrics」と呼ばれる効果量を考慮せず係数の符号を解釈することを回避したか.	94	60.64	27.66	32.98
Q8. 効果量について議論したか.	95	64.21	33.68	30.53
Q2. 「効果がどの程度大きいか」の問に応える弾力性，限界効果，または他の関心のある指標を報告したか.	95	65.26	45.26	20.00
Q11. 統計的有意性のみに基づく変数の選択を回避したか.	94	75.53	-	-
Q3. 標準誤差，t-値及び尤度比をすべて報告したか.	95	76.84	-	-
Q1.モデルに使う変数の記述統計や単位を報告したか.	95	78.95	65.26	13.68
Q6. 「Asterisk econometrics」と呼ばれる検定統計量の絶対値の大きさによって係数をランク付けすることが回避したか.	94	100.00	-	-

Q14 については，「有意」の言葉を使っていない論文は NA と登録した.

Q4，Q6，Q7，Q11 については，モデル結果を示していない（別途で報告している）論文については NA と登録した.

(2) 統計的検出力について

もう一つの注目すべき結果は，どの研究も検定の統計的検出力を考慮していないことである（MZ：4%）．統計的検出力における包括的な説明は本稿の範囲外であるが，議論を喚起する目的で，検出力をめぐる問題を簡略に説明する．

検定の統計的検出力は，帰無仮説が実際に偽であるときに，検定が帰無仮説を正しく棄却する確率を示す．つまり，第二種の過誤（偽陰性）を回避する確率である．検出力は，サンプル数，統計的有意性そして，より重要である効果量の関数である． 最も一般的に用いられる目的は，既存の研究において統計的検定の検出力の計算と，予想される効果量と検出力に対して必要なサンプル数の計算である[9]. すなわち，次の 2 つの質問に答えるため用いられる．一つ目は，サンプル数nに対して，対象とする効果が実際に存在し，その効果量はmであると仮定し，有意水準をαとし，その効果を検出できる（つまり，帰無仮説を正しく棄却できる）確率はどれくらいか. 二つ目は，有意水準を α と検出力水準をbとし，m程度の効果量を検出するために，必要なサンプル数はどれくらいか.

具体的な例を挙げると，0.2 の検出力を持つ検定は，研究者が 5 回中 4 回誤って帰無仮説を受け入れることを意味する． Ziliak & McCloskey[3]の言葉を借りると，「*検出力は，研究者のナイーブさを抑制するものである*」．検出するためにより大きなサンプルを必要とする「小さい」効果の場合は，特に重要である．

特に，複数の検定が行われる多変量解析について，Maxwell[12]は，心理学における検出力不足の研究の分析で，任意の検出力レベルに対して（a）事前に指定した任意の 1 つの効果，（b）少なくとも 1 つの効果，（c）すべての効果，のいずれかを検出するために必要なサンプル数が異なることを示した． たとえば，5 つの予測子による多変量線形回帰（各予測子と他の予測子および被説明変数との相関 =0.3，n =400，および $\alpha = 0.05$ の場合，少なくとも 1 つの効果を正しく検出するための検出力は >.99 であったが，すべての効果を正しく検出するための検出力は，わずか0.22であった． 既往研究の検出力が不足と示され続けていることを考えると，たとえ検定集合の中で一つの効果を検出するのに十分な検出力を持っていても，個々の検定には十分な検出力が持たない研究が多いと Maxwell が指摘した．交通計画の分野においては，大規模な交通調査（PT 調査等）の通常のサンプル数の場合，検出力はあまり問題にならないかもしれないが，本研究でレビューした論文のサンプル数の中央値は 1,404（選択回数）であり，20 パーセンタイルは 527 である（複数のサンプルが報告されている場合は最小値を用いた）． Maxwell の結果が離散選択モデルに適用可能であることを主張しないが，交通計画の分野においても検出力をめぐる課題を対処する必要性を強調したい． 心理学[13,14]や教育学[15]の分野においては，検出力の研究を多く行われているが，知っている限り交通の分野においてそのような分析は行われておらず，検出力における実態は不明である． しかし，心理学の分野においては，単位に依存しない「小」,「中」,「大」の効果を特徴づける尺度を定義した Cohen[9,13]は主に平均のt検定，相関係数，比例差，線形回帰に焦点を当てたが，離散選択モデルに対する統計力に関する研究は少ないことに留意すべき．

最後に，サンプル数の決定について，離散選択実験のためのサンプル数の求め方に関する文献が多い[18,19]一方，既存の理論は検出力のことを殆ど無視している[17].

(3) 統計的有意性と実質的な重要性の混同

効果量に関する議論の背景には，統計的有意性と実質的な重要性の混同という誤った焦点がある．これは，上述の通り 22%の論文は効果量を全く報告や議論をしなかったという事実からも明らかである． つまり，論文 5 本中のうち 1 本は所見の重要性を完全に統計的有意性に基づいて定めていることである． 24% (MZ: 32%)の研究は，統計的有意性のみに基づいて明示的に変数を落としており，63% (MZ: 59%)の研究は，推定、議論、政策への示唆または結論の節で、少なくとも 1 回に「有意」という言葉を，実質的な重要性と混同しているか，著者がどちらの意味を指しているか判別できない形に使っている． また，t-統計量の大きさが，効果量として解釈した研究もあった． 意思決定過程における社会的相互作用の影響に関する研究で，Kamargianni et al.[17]は，歩行嗜好の潜在的構成要素について，*「この構成要素が最も統計的に有意な変数であり…親が子どもの歩行に対する態度の発達に強い影響を与えることを示している」*と述べており，大きな t-統計量を大きいな効果量と誤解している．同様に，Qin et al.[18]は交通手段転換に関する研究で，*「バスのサービスレベルが最も有意な正の t-値を有しており，バスのサービス水準を向上させることで，自動車利用者に対するバスへの転換率を有意に増加させることができることを示している」*と論じている． ここで，交通手段転換率の「有意な」増加とは，実質的な増加を意味するか，統計的に帰無仮説と異なる差に過ぎないかが不明である．

最後に，67%の論文は議論と結論の節で，統計的有意性と実質的な重要性を混在しており．最も一般的な慣例は，統計的有意性に基づいて，一連の変数が重要であると報告し，それらの実質的な重要性，すなわち，効果量の観点から適切に議論していないことであった．

(4) トップジャーナルはどうなっているか

この問いに答えるため，査読付き論文を総合的に評価できるシンプルなスコアを求めた．このスコアは，各質問に対して，「はい，網羅的に」を 1 点，「はい，部分的に」を 0.5 点，「いいえ」を 0 点とし，各論文の有効な質問数を平均し，100 点に正規化したものである．さらに、論文はジャーナルのランクによる区分した．トップジャーナルは Scimago Journal Ranking の交通分野の第 1 四分位のジャーナルと定義する．結果として、トップジャーナルの論文（N：69，平均：44.4，標準偏差:40)のスコアは，非トップジャーナルの論文（N:26，平均:42.6，標準偏差:39.3)よりわずかに優れているが非常に些細な違いであることが分かった．

5. 分野における改善策の提案

以上の議論を踏まえ，統計的有意性から効果量に基づいた政策立案に寄与できる評価への転換を目指して提言を行う．

(1) 効果量とその信頼区間の報告を義務付ける．

統計的有意性は様々な評価基準の中で，たかが一つであり，最も重要な基準であってはならない．そこで，統計モデルの議論は，効果量または他の政策に関連する指標に焦点を当てるべきである．効果量の信頼区間は，Maxwell [10]の言葉を借りると，係数の右上のアスタリスクの有無が伝える「決定的な雰囲気」を伝えず，効果量の推定値に対する不確実性の程度をより明確に把握できる．モデル係数を報告することには反対はしないが，直接的に解釈できず，論文の付録とすることで十分であろう．

(2) 可能な限り，著者が考える効果や関心のある指標の大きさを伝える判断したうえで，その判断の根拠を報告する．

これは必ずしも簡単な判断ではないが，どの程度の効果が政策に対して重要になるか，その重要性をどう評価するかの議論と共に，政策変数の操作をかかるコスト（政策導入コスト）を議論すべきである．

(3) 可能な限り，推定した効果や関心のある指標を既往研究と比較する．

交通時間価値 [23–25]等よく報告されている値の場合は，既往研究との比較への障壁はないが，よく報告されていない値の場合，変数の定義及び測り方のばらつきにより，比較することが難しくなる時もある．ただし，効果量の報告が普及されたらこの問題が解消される．

(4) 新規研究については，対象効果を十分な検出力で検出できることを保証するために，サンプルサイズを決める際に統計的検出力を考慮すること．また，二次データ（例：パーゾントリップ調査など）を用いた研究の場合は，事後に各検定の検出力を求めて報告する．

離散選択モデルにおける検出力に関する文献は非常に少ないが，参考文献として deBekker-Grob ら [17]の研究がある．

6. 結論

本研究では，2014 年から 2018 年にかけて離散選択モデルを中心として交通計画の学術文献をレビューし，実証分析における統計的有意性の使用と誤用を評価した．その結果，多くの研究は，統計的有意性の使用における誤りを繰り返し，効果量に明確な焦点を当てていないことを明らかにした．

交通計画の研究の究極の目的は，交通政策立案とその実施に貢献することであり，そのためには効果量とその実質的な意味について適切な議論が必要であることを改めて強調しておきたい．そこで，本稿の目的は，分野を批判することではなく，分野が究極の目的とより一致するように改善策を提案することである．

謝辞

本研究は，日本学術振興会科研費 20H02266 の助成を受けたものである．

参考文献

1) Parady，G.，D. Ory，and J. Walker. The Overreliance on Statistical Goodness-of-Fit and under-Reliance on Model Validation in Discrete Choice Models: A Review of Validation Practices in the Transportation Academic Literature. *Journal of Choice Modelling*，Vol. 38，No. November 2020，2021，p. 100257. https://doi.org/10.1016/j.jocm.2020.100257.
2) Mccloskey，D. N.，and S. T. Ziliak. The Standard Error of Regressions. *Journal of Economic Literature*，Vol. 34，No. 1，1996，pp. 97–114.
3) Ziliak，S.，and D. McCloskey. *The Cult of Statistical Significance. How the Standard Error Cost Us Jobs，Justice and Lives*. The University of Michigan Press，2007.
4) Khan，M.，K. M. Kockelman，and X. Xiong. Models for Anticipating Non-Motorized Travel Choices，and the Role of the Built Environment. *Transport Policy*，Vol. 35，2014，pp. 117–126. https://doi.org/10.1016/j.tranpol.2014.05.008.
5) Aziz，H. M. A.，N. N. Nagle，A. M. Morton，M. R. Hilliard，D. A. White，and R. N. Stewart. Exploring the Impact of Walk–Bike

Infrastructure, Safety Perception, and Built-Environment on Active Transportation Mode Choice: A Random Parameter Model Using New York City Commuter Data. *Transportation*, Vol. 45, No. 5, 2018, pp. 1207–1229. https://doi.org/10.1007/s11116-017-9760-8.

6) Heinen, E., and D. Ogilvie. Variability in Baseline Travel Behaviour as a Predictor of Changes in Commuting by Active Travel, Car and Public Transport: A Natural Experimental Study. *Journal of Transport and Health*, Vol. 3, No. 1, 2016, pp. 77–85. https://doi.org/10.1016/j.jth.2015.11.002.

7) de Luca, S., and R. Di Pace. Modelling Users' Behaviour in Inter-Urban Carsharing Program: A Stated Preference Approach. *Transportation Research Part A: Policy and Practice*, Vol. 71, 2015, pp. 59–76. https://doi.org/10.1016/j.tra.2014.11.001.

8) Cantarella, G. E., and S. de Luca. Multilayer Feedforward Networks for Transportation Mode Choice Analysis: An Analysis and a Comparison with Random Utility Models. *Transportation Research Part C: Emerging Technologies*, Vol. 13, No. 2, 2005, pp. 121–155. https://doi.org/10.1016/j.trc.2005.04.002.

9) Jacob, C. *Statistical Power Analysis for the Behavioral Sciences*. Psycology Press, 1988.

10) Allard, R. F., and F. Moura. Effect of Transport Transfer Quality on Intercity Passenger Mode Choice. *Transportation Research Part A: Policy and Practice*, Vol. 109, No. January, 2018, pp. 89–107. https://doi.org/10.1016/j.tra.2018.01.018.

11) Hess, S., G. Spitz, M. Bradley, and M. Coogan. Analysis of Mode Choice for Intercity Travel: Application of a Hybrid Choice Model to Two Distinct US Corridors. *Transportation Research Part A: Policy and Practice*, Vol. 116, No. April 2016, 2018, pp. 547–567. https://doi.org/10.1016/j.tra.2018.05.019.

12) Maxwell, S. E. The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, Vol. 9, No. 2, 2004, pp. 147–163. https://doi.org/10.1037/1082-989X.9.2.147.

13) Jacob, C. The Statistical Power of Abnormal-Social Psychological Research: A Review. *The Journal of Abnormal and Social Psycology*, Vol. 65, No. 3, 1962, pp. 145–153.

14) Rossi, J. S. Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology*, Vol. 58, No. 5, 1990, pp. 646–656. https://doi.org/10.1037/0022-006x.58.5.646.

15) Ziliak, S. T. How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough. *American Statistician*, Vol. 73, No. sup1, 2019, pp. 281–290. https://doi.org/10.1080/00031305.2018.1514325.

16) Rose, J. M., M. C. J. Bliemer, D. A. Hensher, and A. T. Collins. Designing Efficient Stated Choice Experiments in the Presence of Reference Alternatives. *Transportation Research Part B: Methodological*, Vol. 42, No. 4, 2008, pp. 395–406. https://doi.org/10.1016/j.trb.2007.09.002.

17) Rose, J. M., and M. C. J. Bliemer. Sample Size Requirements for Stated Choice Experiments. *Transportation*, Vol. 40, No. 5, 2013, pp. 1021–1041. https://doi.org/10.1007/s11116-013-9451-z.

18) de Bekker-Grob, E. W., B. Donkers, M. F. Jonker, and E. A. Stolk. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: A Practical Guide. *Patient*, Vol. 8, No. 5, 2015, pp. 373–384. https://doi.org/10.1007/s40271-015-0118-z.

19) Kamargianni, M., M. Ben-Akiva, and A. Polydoropoulou. Incorporating Social Interaction into Hybrid Choice Models. *Transportation*, Vol. 41, No. 6, 2014, pp. 1263–1285. https://doi.org/10.1007/s11116-014-9550-5.

20) Qin, H., J. Gao, H. Guan, and H. Chi. Estimating Heterogeneity of Car Travelers on Mode Shifting Behavior Based on Discrete Choice Models. *Transportation Planning and Technology*, Vol. 40, No. 8, 2017, pp. 914–927. https://doi.org/10.1080/03081060.2017.1355886.