

<https://scholar.google.com/scholar?cluster=17984152132290578827> (<https://scholar.google.com/scholar?cluster=17984152132290578827>)

この論文は、\*\*2019年にSander GreenlandとAlbert Hofmanによって出版された「Multiple comparisons controversies are about context and costs, not frequentism versus Bayesianism」というタイトルの論説です。日本語に訳すと、「多重比較の論争は、頻度主義対ベイズ主義ではなく、文脈とコストに関するものである」\*\*となります。

以下に、この論文の全体を詳しく要約し、特に頻度論とベイズの区別がテクニカルな違いに過ぎず、哲学的違いではないという主張について詳しく説明し強調します。

統計的な検定のような基本的な問題でさえ、300年の歴史の中で合意が得られていないことを考えると、多重比較のようなさらに微妙な問題について合意を期待することは難しいです。基礎についてのコンセンサスがないままでは、多重比較のような複雑なテーマが広く誤解され、対立が生じているのは驚くべきことではありません。

SjölanderとVansteelandt (SV) は、疫学統計学への主要な貢献者ですが、彼らの論文 (SV) は、多重比較の論争の側面を頻度主義統計学に帰し、正式な多重比較調整を行わず、「非公式で質的な」判断に頼ることを推奨しています。このアドバイスは、分析が少数の密接に関連した統計的仮説やパラメーターに焦点を当てた一つの明確な研究課題のみを対象とし、すべての分析と推定値が等しい重点と詳細さで報告される設定においては、現実的に防御可能なヒューリスティックかもしれません。このような単純なケースでは、関連するすべての関連性とその相互作用および不確実性を一度に心に留めておくことができ、「フィッシング」や「ハッキング」といった非難を回避できます。

しかし、SVはこのような「非公式な調整」をすべての多重比較分析に対する一般的な妥協案として提案しているように見えますが、これは実際には妥協案ではなく、正式な多重比較調整を全く行わないという古い極端なアドバイスを繰り返すことに近いのです。このアドバイスはまだ一般的ですが、分析にいくつかの相互依存するパラメーターや仮説がある場合（職業、環境、栄養に関する研究に典型的な複合的暴露や複数のアウトカムの研究など）には賢明ではありません。なぜなら、これらの依存関係を利用して、複数の結果全体の精度を劇的に向上させることができるからです。

さらに悪いことに、正式な調整を提供しないことは、非常に探索的な設定、つまり多くの弱い可能性の中から焦点を絞った努力で追求すべきものを決定することを目的とする設定において、非常に誤解を招く可能性があります。大規模な探索においては、それは非現実的です。これらの設定では、多重比較の問題は現実的であるだけでなく、人間の直感では適切な多重比較調整を適用せずに影響を把握するには遠く及ばないものです。

著者らは、SVやその反対者（Häggström）が記述していない、しかし多重比較の議論と調整にとって極めて重要な問題と広範な発展について概説しています。著者らは、**階層（マルチレベル）回帰が、検定とパラメーターまたは仮説間の依存関係を利用して、報告および意思決定に使用される統計的要約全体の精度を向上させることを強調しています。また、結論や決定を主張するすべての統計的手法に内在する決定コスト（効用または損失関数）の中心的意義を強調しています。**現代統計学の乱用と論争の多くは、これらの損失関数が典型的な研究応用で無視されていることに起因しており、すべての結論と決定はエラーコストに無限に敏感であるにもかかわらずです。この無視による問題は、多重比較の論争においてさらに激化します。

**本論説は、これらの問題が頻度主義とベイズ主義の方法論の違いとは何の関係もなく、統計的手続き（アルゴリズム）の文脈的妥当性とその出力の有効な解釈に関するすべてであることを説明することから始まります。**ここでの妥当性は、頻度主義、ベイズ主義、その他の方法論的理想化の混合を代表する実用的な基準によって判断されます。

「ベイズ」という用語は今日、パラメーターの事後分布を生成するあらゆる手続きに対して使用されますが、その分布が文脈（特に因果モデル）によって情報提供されているかどうかに関わらずです。SVは、広く文学で繰り返されているように、あらゆる技術レベルで共通の誤りを繰り返しています。「ベイズ主義」を文脈情報（「事前」情報）を分析に組み込むことと同一視しています。しかし、後述するように、文脈情報を使用するためのすべてのベイズ的手法には、同じ情報を使用する**並行した頻度主義的手法が存在**します。したがって、情報の組み込みにおける違いは、方法論にとって根本的なものではなく、**むしろ文化的なものです。**ベイズ主義者は曖昧に支持された意見が正確で確固たる設計情報に置き換わることをより寛容に許容する一方で、多くの頻度主義者はすべての統計的仮定に対して物理的な正当化（例えば、独立性の仮定を正当化するためのランダム化）を主張します。

頻度主義とベイズ主義の方法論の違いの誤った特徴付けは、文脈情報を「事前情報」とラベル付けすることが伝統的であったために生じる可能性があり、それがベイズ的方法論を連想させます。しかし繰り返しますが、頻度主義的方法論も同じ情報を取り入れることができます（例えば、経験的ベイズ分析や半ベイズ分析のように、それを階層モデルに組み込むことによって）[7, 3–12, 14, 16]。したがって、多重比較の論争や解決策を頻度主義対ベイズ主義と同一視することは、両システム間の区別を文脈情報を使用する能力の違いと誤って認識することであり、**実際には入出力の定式化の違い**なのです。

**多重比較においては、頻度主義者とベイズ主義者の両方が同じ文脈的基盤を必要とします。**実際、最も一般的な「頻度主義的」多重比較方法のほとんどは、それらが開発された狭い設定以外では健全な根拠がありません。例えば、ボンフェローニ調整は、すべてのステークホルダーにとって真の陽性の事前確率が非常に低く、偽陽性のコストが偽陰性のコストを大幅に上回り、個々の仮説に対する統計的パワーが数パーセントしかなくても許容されるという特別な応用例でのみ正当化される損失関数に基づいています。

このような方法は、多くの統計学者によってコストやパワーの問題、仮説間の依存関係に注意を払わずに精力的に推進されてきました。しかし、その残念な事実は、すべての頻度主義的調整を非難する根拠には全く成りません。したがって、頻度主義もベイズ主義も限られたツールキットに過ぎないと考えているアナリストとして、著者らは多重比較調整の範囲に関するSVの以下の引用が完全に間違っていると考えます。「頻度主義的な視点からは、あらゆる可能な検定のコレクションが調整にとって等しく有効に見え、したがって、これらの間のいかなる選択も完全に恣意的であるように見える。」

これは間違っています！**選択は文脈と探索の対象によって決定されます。各調整は異なる文脈と対象に対応しています。**一つの広範な多重比較の文脈は、探索的スクリーニング（「釣り遠征」）であり、偽のリードや逃した機会のコストを考慮して、どの関連性をさらに研究するかについての決定を対象としています。別の広範な多重比較の文脈は同時推定であり、偏りとランダムエラーのトレードオフを考慮して、関連性のアンサンブル全体に関する総情報の正確な要約を対象としています。

これらの文脈やその他の文脈において、頻度主義とベイズ主義の両方の視点があり、それらは何をどのように調整すべきかについて並行したガイダンスを提供します。これは、観察された関連性に貢献し、研究対象の変数によって共有される要因の認識から始まります。例えば、産業用化学物質が職業的および環境的な疾患との関連性に貢献する要因として、栄養素が健康との食事関連性に貢献する要因として見られることなどです。この種の関連性の因果分析は、完全な事前分布を指定することなく、文脈情報（外部情報）を多重比較調整に入力するための基礎を形成し、頻度主義的およびベイズ的評価の両方で行われます。これは（例えば）なぜ年齢と性別の係数を、職業的または食事効果の多重比較調整探索のための調整セットに含めるべきでないかを説明します。

関連性の因果分析は、文脈と設計情報、情報とエラーのコストを含め、どのような検定と推定のコレクションを調整すべきか、そして調整をどのように行うべきかを決定する側面のほんの一部にすぎません。調整の範囲に関するこのような文脈的入力の必要性は、統計的検定の $\alpha$ レベル（最大許容偽陽性率）を設定するための文脈的入力の必要性、あるいは事前分布の形式、中心、広がりを設定するための必要性和何ら変わりありません。これらの入力の必要性は、頻度主義対ベイズ主義の方法論や哲学とは何の関係もなく、分析によって対処されている質問（例えば、「この研究はこれらの仮説に対してどのような証拠を提供するか？」対「この証拠に直面して我々は何をすべきか？」）にすべて関係しています。具体的な指示のギャップは、文脈的な目的と分析の範囲、そして統計的手法を理解しているアナリストによって埋められることになっています。

この文脈的専門知識の必要性を無視した、すべての回帰係数を調整セットに含めるような不合理な多重比較の議論や分析を見つけることができます。これは、年齢と性別の効果が暴露効果と同様であると期待されているかのようです。これらの悪い実践は、SVが伝えた「頻度主義的な視点からは、あらゆる可能な検定のコレクションが調整にとって等しく有効に見える」という誤った印象を生み出すのに役立ちます。また、ベイズ的な視点からは、すべてを調整すべきであるという印象も生み出します。しかし、文脈への没入こそが、調整なしという貧弱な極端と、すべてを調整するという不合理で不可能な極端の間に、文脈的に *sensible* な点を指定するために不可欠であり、数学的な統計学だけではありません。この点で、方法論が調整セットを正式に指定しないいかなる失敗も、抽象的で文脈から切り離された統計的ルールを使用して *sensible* に答えることができない質問に対する正直な応答です。

多重比較調整の因果的な基盤を認識することで、多重比較調整を避けるというアドバイスは、交絡調整を避けることに似ていると見なすことができます。これは、推定値の精度を向上させるために文脈情報を使用することを避けるというアドバイスです。反対の極端では、すべての比較を他のすべての比較（アナリストによって実行されなかった比較でさえ）について調整すべきだと主張することは、私たちの宇宙の始まりまで遡るすべての考えられる共有された原因について、2つの変数間の観察された関連性のすべての因果分析が調整しなければならぬと主張することと同じくらい現実からかけ離れています。両極端は、すべての応用統計学における文脈の決定的な役割を理解することの失敗、そして結果として文脈情報を分析に適切に統合することの失敗を反映しています。

**私たちのツールキットには、「頻度主義的」と「ベイズ主義的」という極端なものよりもはるかに多くのものがあります。**SVは、両者が共有する限界について頻度主義を非難し、ベイズ主義を免責する中で、それぞれの欠点を解消するために頻度主義とベイズ主義のアイデアを融合させた実用的な方法論の存在を認識していません。Rothmanへの応答で、著者らは経験的ベイズ（EB）、疑似ベイズ、収縮、ランダム係数、階層、マルチレベルモデリングとして知られる代替の多重比較方法論を指摘し、その後、これらの方法論が実際の疫学データでどのように機能するかを詳細な図解で示し [5–7, 11]、さらに主要な研究報告書で展開しました（例えば）。これらの方法論は1970年代までに疫学の例に登場し始め、それ以来応用統計学の書籍や一般的なソフトウェアで広く利用可能になりました。それらは頻度主義的バージョンとベイズ的バージョンの両方で提供されており、部分ベイズ、半ベイズ、準ベイズ、混合モデル、ペナルティ回帰などの多くのハイブリッドも存在します。これらの方法論はすべて、通常の回帰を適合させるのに使用される標準的なソフトウェアを使用して一般的な疫学分析に簡単に適用できます [25–27]、またシミュレーション方法を使用することも可能です。

通常の回帰とその因果的拡張と同様に、階層的な方法論は情報要約と意思決定の両方に使用できます。残念ながら、統計文献（SVを含む）の多くは、これらの二つのタスクを区別していません。これはおそらく、二つのタスクが同じモデリング方法論とコンピュータ出力に依存しているためです。しかし、頻度主義対ベイズ主義とは異なり、**この区別は多重比較の論争の中心にあり**、これは主に意思決定におけるエラーコストの極めて重要な役割によるものです。

**頻度主義-ベイズの区別は、哲学的なものではなく、技術的な違いです。**著者らが見る頻度主義-ベイズの区別に関する議論のあまりに多くは、以下の問題の少なくとも一つ、そして通常はすべてを見落としています。

第一に、**この区別は、多重比較だけでなく、ほとんどの方法論論争において巨大な気晴らしです。**なぜなら、現代統計理論の基礎文献に浸透していた、哲学とツールキットの間の深い混乱に加担しているからです。この混乱は応用分野で依然として風土病です。それにもかかわらず、頻度主義とベイズ主義の両方が、**すべての健全な応用には不十分であるという現実的な意味で、学習理論としても統計学の哲学としても不完全**であることは、様々な言葉で長年にわたり広く認識されてきました [17, 15, 20, 28–38]。働く科学者や統計学者にとって、頻度主義とベイズの方法論は、代わりに**異なる視点から与えられた統計的問題に取り組むツールボックス**であり、提案された解決策の異なる側面に対処します。

第二に、キリスト教に一つの形しかないということがないように、**単一の頻度主義的またはベイズ的哲学や方法論というものは存在しません。**半世紀前、Goodは46,656種類のベイズ主義者の分類スキームを提供し、ほとんどのタイプは誰も持っていないと述べましたが、文献には *dozen* 種類のタイプを見つけることができます—ほとんどは、それが唯一無二の真のベイズ主義であるかのように提示されています。頻度主義についても状況はそれほど単純ではなく、おそらく半ダースのパリエーションがあります。これらの統計的「哲学」内の宗派間の対立は、頻度主義対ベイズ主義という不合理に広範なカテゴリー間の対立（「ユーラシア人」と「北米人」を区別するのと同じくらい情報を提供しない二分法）よりも大きいです。これらの対立において、「統計学の哲学」は、健全な方法論のオープンな探求よりも神学に似ています。しかし、これらの対立のほとんどは、文脈内でしか解決できません。これは、普遍的な推論方法というものが存在しないことを反映しています。

第三に、**統計的方法論は、同じデータで異なる証拠を主張することを完全に合理的に**します。これは実際、ベイズの定理から明らかです。合計証拠または情報が事後分布を通じて測定される場合、それは事前分布と標本分布の両方に敏感です。異なる研究者は異なる事前分布を持ち、異なる事後分布につながります。彼らはまた、適切な標本分布についても異なる見解を持ち、頻度主義者にとっても証拠の異なる評価につながります。この種の対立は、研究者が文脈について非常に異なる見解を持っている場合（例えば、過去の研究についての相反する見解、または分析から導き出される印象、結論、決定に対する相反する利害関係）に常態です。

データが最終的に異なる事前分布を圧倒し、合意を強制するという誤った概念によって、対立問題は時に片付けられます。残念ながら、**データは標本分布のすべての側面を特定できないため、その分布が論争になっている場合に推論についての合意を強制することはできません。**さらに、**データは損失関数についての合意を強制できないため、データや標本モデルについての論争がない場合でも、結論や決定についての合意を強制することはできません。**

低レベルの議論ではまだ見過ごされていますが、高度な教科書ではますます認識されている第四の点として、**すべての統計的方法は、それが提案された頻度主義的手続きであるかのように、そして提案されたベイズ的手続きであるかのように分析することができます。**この**相補性**は、非統計学者にとってはコンピュータサイエンスの視点から見ると理解しやすいかもしれませんが。**データ分析方法論はすべて、データを取り込み数値を出力するデータ処理アルゴリズム（プログラム）と見なすことができます。**これは、アルゴリズムの元の根拠が頻度主義、ベイズ主義、その両方、あるいは全く別のもの（例えば、最小記述長）であったかに関わらずそうです。

これらの出力がどのように解釈されるかは、それを評価する者の目にあり、その解釈は、アルゴリズムの理論的（論理的、数学的）および文脈的な根拠、特にデータ生成プロセスの因果モデルの理解の関数となります。その解釈は、アルゴリズムの論理の誤解、理論的な誤解（例えば、欠陥のある統計教育から生じる）、および文脈的な誤報（例えば、過去の研究の無知、誤報、誤解から生じる）に悩まされる可能性があります。結果として生じる誤解が頻度主義的な形式であろうとベイズ的な形式であろうと、問題の一つの側面（そしてアルゴリズムが両方の視点からの正当化を持っている場合は重要でないかもしれませんが）にすぎず、文脈的な誤報と比較して結果が小さい可能性があります。

**ベイズ統計学は、不正確な文脈的（背景）情報をアルゴリズムに組み込むためのツールに焦点を当ててきました。**これは事前分布（チューニング関数）を介して行われ、プログラム出力がパラメーターまたは仮説の確率の観点から解釈可能になるようにします。**頻度主義統計学は、既知の決定論的およびランダムな形式の入力の下でのアルゴリズムの振る舞いを評価するためのツールに焦点を当ててきました。**これは、データサンプリングモデルに対して方法論を較正することであり（数学的にも、データシミュレーションによっても）。特定の応用においては、これらの視点のそれぞれは、サンプリングモデルが実際のデータ生成メカニズムの振る舞いに関する正確な情報（文脈的であり、主に因果的な形式の情報）を組み込んでいる程度に役立ちます。

頻度主義的な較正は、サンプリングモデルを事前情報とデータに対してチェックを提供し、ベイズデータ分析にとって重要です。ベイズツールもまた、頻度主義的方法論に対して有用なチェックを提供できます。例として、逆ベイズ分析があります。頻度主義的方法論（サンプリングモデルに従って較正されたアルゴリズム）を手渡された場合、アルゴリズムの出力を逆エンジニアリングして、その出力をサンプリングモデルの下での事後要約にするような事前分布を見つけることができます [24, 20, 30, 43–45]。この暗黙の事前分布は、文脈情報と照合してチェックされ、その情報を考慮するように修正できます。事前分布はサンプリングモデルと照合してチェックされるべきです—ただし、較正を維持するためには、そのチェックに基づいた更新は、データの二重カウント（チェックでの最初と更新での二回目）を調整しなければなりません。

方法論はこれらの評価のいずれかに失敗する可能性があります。方法論は、現実的なサンプリングモデルの下で較正が不十分である可能性があります。例えば、サンプリングモデルから描かれたサンプル上でP値が一様でない可能性があり、これはP値の情報を低下させ、したがって効率性（情報使用）を最大化するという頻度主義的要求に失敗します。または、方法論は、応用例に翻訳された場合に受け入れられない事前分布や損失関数を伴う可能性があります。これはほとんどの文脈でのボンフェローニ調整に典型的です。幸いなことに、実際には [25 footnote 4]、コンテキスト的に現実的なサンプリングモデルの下でよく較正され、現実的な事前分布からも導き出せるアルゴリズムを構築することが通常可能です。

このようなデュアル頻度主義-ベイズ（FB）の方法論を生成する比較的簡単な方法は、階層モデリングを介することです [26, 6–8, 12, 22, 23, 50]。これらの方法論はそれ自体では損失関数の問題に対処しませんが、パラメーター（事前）および推定量の依存関係を考慮しており、したがって未調整または伝統的な多重比較調整分析よりも推論のためのより良い情報に基づいた基盤を提供します。BerryとHochbergが書いているように、「一部の統計学者は、ベイズ的視点が、比較ごとの【未調整の】アプローチの提唱者である頻度主義者を支持していると見なしています。我々がセクション2で議論するように、これは一つのパラメーターの（事前および）事後確率が他のパラメーターを考慮しても変化しない場合に真実です。しかし、我々は、より現実的な状況では一般に依存するパラメーターが含まれており、そのような状況では、ベイズ的視点からも調整は正当であり、しばしば必要であると主張します。交換可能な依存パラメーターをモデリングするための便利なアプローチは、階層事前モデルを仮定することです...」。

彼らはその後、頻度主義的な性能基準を満たすように較正されたベイズ的方法論を提唱しています。これは、**それらが頻度主義的方法論でもあるベイズ的方法論である**と言い換えることができます。

fitted モデルは情報要約として機能します。これはモデルが頻度主義的であるかベイズ的であるか、問題が単一比較であるか多重比較であるかに関わらず当てはまります。要約の目標は、研究対象の関係について情報を提供する非ランダムなデータのすべての次元を捉えるモデルを見つけることです。これにより、元のデータセットを fitted モデルに置き換えることができます。この見方は、モデルが多重比較調整を組み込んでいるかいないか、そしてモデルが頻度主義的、ベイズ的、ハイブリッド、またはその他の基準セットを満たす方法論で fitted されているかに関わらず適用されます。正しいモデルを確信できない場合、関連する情報損失を最小限に抑え、誤解を招くモデルを避けるためには、モデルチェックと修正が必要です。例えば、事前分布をサンプリングモデルと比較し、深刻な矛盾がある場合はいずれかまたは両方を修正する必要があります（厳密なベイズ原理に違反しますが）。ただし、較正（したがって圧縮の精度）を維持するためには、その後の要約で修正を考慮する必要があります。

分析の目標が複数の関連性や効果に関する情報を要約することである場合、階層モデリングは、圧縮データ内の有効な情報（シグナル）を最大化し、ランダムなアーチファクト（ノイズ）を最小化するためにモデルを微調整するためのコヒーレントなフレームワークを提供します。標準的な回帰モデルは、モデルに一つのレベルしかなく、個々のアウトカムが個々の共変量（例えば、職業、食物摂取）に対して回帰される特殊なケースです。標準的な結果と階層的な結果の比較は、共変量の係数がそれらの共変量の特長（例えば、職業係数が職業暴露に対して回帰される、食物係数が食物の栄養含有量に対して回帰される）に対して回帰されるもう一つのモデルレベルを追加するだけで、最も劇的な改善が得られる方法を示しています。このアプローチは、個々の共変量効果、および共変量特性の個人への効果の両方について、優れた推定値を提供します。これらは、それらの共変量（第2レベル）特性を、職業的および栄養学的研究で標準的であるように、個々の（第1レベル）共変量であるかのように扱うよりもはるかに優れています。また、ステップワイズ回帰やその他の較正が不十分な一般的な変数選択方法論に対するコヒーレントな代替策も提供します。

**損失関数は、関係についてのいかなる決定的なステートメントを正当化する上で中心的な役割を果たします。**効果がない、害がある、利益があるといった主張は、間違っていることのコスト（常にリスクを伴う）が、非決定的であることや誤った結論を出すことのコストよりも小さいという暗黙の信念を伴います。単に無条件に「さらなる研究が必要である」と述べることで、さらなる研究から得られる情報がコストに見合う価値があるという結論を伴います。

明示的な損失関数を組み込んだ頻度主義およびベイズの方法論は、1950年までに理論的に練り上げられており、厳密な頻度主義的な較正理論と厳密なベイズ的な一貫性理論において、最適な意思決定のためのアルゴリズムが同一であることが示されていました（例えば）。BerryとHochberg が述べたように、多重推論における損失関数の重要性は1960年代までに認識されていました。1980年代までには、損失関数を示す多くの応用書籍があり、方法論は大規模な回帰分析にとって計算上実用的になりつつありました。

残念ながら、明示的な損失関数を使用する方法論は、統計トレーニングのごく一部にすぎず、損失関数は統計の実践と議論で依然として無視されています。これは、モデリングがすべてのステークホルダーにとって受け入れられる情報要約を提供できる場合には許容されるヒューリスティックかもしれません。しかし、実際には、**この無視は、統計分析から導き出されるいかなる結論や決定をも無効にし、方法論的な主張や標準的な統計的実践に内在する価値観を隠蔽します。**

結果として、いかなる統計的方法論の使用を検討する際には、その頻度較正とその（しばしば暗黙の）事前分布だけでなく、**その損失関数についても知る必要があります**。損失関数は隠されている可能性が高いため、逆計算（逆エンジニアリング）する必要があります。論争のあるトピックでこれが行われると、多くの標準的な手続きが特定の一方またはステークホルダーに偏っていることがわかるかもしれません。多重推論問題では、多くの可能なエラーパターンがあるため、損失関数を明らかにすることがさらに重要になります。例えば、多重仮説検定は、アルファ/タイプIエラーとベータ/タイプIIエラーを超えて、すべての仮説にわたるすべてに組み合わせエラーパターンを含みます。

調整戦略（調整なしを含む）の根拠は、結論に重要な利益または損失がかかっている場合（例えば、「発見」が権威ある出版物や法的責任を伴う場合）に、研究者の偏見に対して特に脆弱です。特定の調整（調整なし、「非公式な調整」、ボンフェローニ、またはその他）に基づく結論や決定の要求は、コストに関する暗黙の仮定を使用して統計と科学に価値観を押し付けようとする試みと見なすことができます。したがって、著者らは警告で締めくくっています。**多重比較調整に関する論争を見たら、誰がそのトピックのステークホルダーであるか？ 特定の方法的推奨から誰が得をし、誰が損をするか？ 各推奨が具体的にどのような効果や仮説に対処しているか？ 各側が到達しようとする結論や決定は何か？ 彼らの推奨に暗黙的に含まれる事前分布と損失関数は何か？**と尋ねてください。そうすれば、（著者らがそうであるように）多重比較の論争が異なる目標、価値観、利害から生じており、それらの要因に関する異なる仮定の下で方法論がどのように変化するかを示す以外に解決策はないと判断する可能性が高いです。

付録のBygren et al.の例は、SVが使用した心血管死亡率と祖父母の食料供給の関係に関する報告からのものです。SVが再現した分析では8つの係数のみが含まれており、子供、親、祖父母の性別でデータを層別化しています。この分析にはいくつかの問題があり、その一つは、層別化によって示唆されるモデルが文脈的に不合理であることです。それは、食料供給効果の方向やサイズの変動について、世代別の性別指標によるいかなる制約も全く置いていません。より合理的なモデルを適合させる試みは行われませんでした。SVの表1から計算された8つの係数間の差の検定のP値は約0.2であるにもかかわらずです。Häggström が正当に不平を述べたように、報告書は $P < 0.05$ の1つの推定値を過度に強調するといういつもの誤りを犯しました。観察されたP値0.04は、そのような強調を正当化するには、帰無仮説に対してあまりにも少ない証拠しか示していません（これは帰無仮説に対する情報量が $-\log_2(0.04) = 4.6$ ビットに翻訳されます）。文脈的または統計的に合理的なモデルであれば、唯一の「統計的に有意な」発見は消滅したはずでした。したがって、この例の分析の問題は、多重比較の問題というよりも、**「統計的有意性」を見つけることに意図された者たちに向けられた、無能な（ただし標準的で一般的）モデリングと報告の実践**にあります。

階層モデリングがどのように役立ちうるかを見るために、表1の8つの食料供給係数が、すべてのデータから同時に導出された修正なし（「相互作用なし」）モデルに向かって収縮された可能性があることに注目してください。このモデルは8つの性別組み合わせすべてに対して1つの食料供給係数のみを持ち、ハザード比0.95、95%信頼区間0.73-1.23を生成するように見えます。このアプローチの半ベイズ（ペナルティ付き尤度）バージョンは、文脈的に合理的な修正の範囲（平均係数からの偏差）を表す固定事前分散によって、この単純な定数効果モデルに重み付けを行ったでしょう。より洗練されたバージョンは、世代と性別を第2段階の共変量として使用したでしょう。結果として生じる収縮は、点推定値間の大きな不均衡を排除し、その差をほとんどノイズとして扱い、そうすることで $P=0.04$ という結果を魔法の0.05閾値の上に持ち上げたでしょう。そのような分析は、報告書の重要性和宣伝を抑え、これらの著者だけでなく、研究コミュニティ全体が階層的な多重比較調整を採用することに消極的である理由を容易に理解させます。

要約すると、多重比較に関する論争は、統計的手法（頻度主義 vs ベイズ主義）の選択における根本的な哲学の違いから生じるも