Divergence versus decision P-values A distinction worth making in theory

ご提示いただいた翻訳を使用し、「Divergence versus decision P-values A distinction worth making in theory」（ダイバージェンスP値と意思決定論P値 理論上区別する価値のある違い） の内容について、ソースに基づき詳しく要約します。

この論文の主要な主張は、**観測されたデータセットを生成するプロセスに関する提案された仮説やモデルを評価するための「P値」には、実際には2つの異なる定義が存在し、これらを区別することが重要である**というものです。これらはしばしば混同されますが、一般的には数学的にも意味においても等価ではありません。

1. **ダイバージェンスP値（Divergence P-values）**

    - これは歴史的に見て最初のP値の定義です。その基本的な形は、観測されたデータの、モデルの予測からのずれや**ダイバージェンス**（乖離）を測る統計量（例えば、二乗和や逸脱度統計量など）の、**記述的な幾何学的要約**です。
    - **ダイバージェンスP値**は、モデルとデータから導出された参照分布における、その統計量の**順序上の位置**（順位や分位点）として定義されます。
    - その目的は、意思決定規則を実装することではなく、**証拠を要約すること**です。具体的には、**データとモデルの間の相性の良さ（compatibility）**、整合性（consonance）、一貫性（consistency）を、単位スケールで示す指標として扱われます。
    - 大きな**ダイバージェンスP値**は、観測データが帰無仮説（H0）に対するほとんど証拠を提供しないことを示しますが、それ自体がH0の証拠となるわけではありません。それはデータとH0の整合性、**相性の良さ**、協調性、または一貫性を示すものと解釈されます。
    - **ダイバージェンスP値**が小さいほど、データとモデルの**相性が悪い（incompatible）**ことを示します。これを**相性の悪さ（incompatibility）**または反証（refutation）の指標として捉えるために、負の対数変換（S値 = -logc(p)）を用いることが提案されています。この変換は、P値を確率スケールから外し、後方確率との混同を防ぐのにも役立ちます。
    - **ダイバージェンスP値**は、反証の証拠を首尾一貫して（coherently）示すことができます。これは、**相性の良さの尺度（compatibility measure）**が満たすべき「サブセット一貫性（subset coherence）」のような、**単一標本の一貫性基準（single-sample coherence criteria）**を満たすように定義されるためです。データがモデルの制約と全く矛盾しない場合、**ダイバージェンス**はゼロになり、P値は1になります。より制約的な（入れ子になった）モデルに対するP値は、それを含むより広いモデルに対するP値を超えることはありません。
    - **ダイバージェンスP値**の参照分布は、反復抽出における意思決定の目標や最適化に基づく必要はなく、データとモデル間の**相性の良さ**と対立の首尾一貫した尺度を提供することを目的としています。帰無仮説の下でのP値の均一性は、定義上の、あるいは必須の特性とはみなされません。
    - これらは、参照的（reference）あるいはネオ・フィッシャー流（neo-Fisherian）頻度論の考え方に根ざしており、標本分布と観測データの関係記述に焦点を当てています。
2. **意思決定論P値（Decision P-values）**

    - この定義では、P値は、ファミリFAをさらに使用するために棄却または採択するかどうかの意思決定基準（「検定」）の副産物にすぎません。
    - これは、標本空間全体にわたる離散選択規則をコンパクトに表現するために使用される確率変数、またはその実現値です。観測された**意思決定論P値**pinf(α)は、データyが領域RA(α)に属するときにFAが棄却される最小のαレベルとして定義されます。
    - 目的は、モデルと特定の対立仮説の下で既知の誤差率を持つ意思決定規則を生成することです。ネイマン・ピアソン（NP）理論に基づき、標本空間全体にわたる意思決定を最適化すること（例えば検出力（power）の最適化）に焦点を当てます。
    - **意思決定論P値**は、しばしばデータ記述として誤って表現されますが、実際には意思決定の構成要素です。
    - これは、**単一標本の一貫性基準**に違反することがあります。その定義と評価が反復抽出のみに基づいており、単一標本での実現値の特性に注意を払わないためです。例えば、サブセット一貫性に違反することが示されています（FM ⊂ FA であるにもかかわらず pt;M > pt;A となる場合がある）。これは、検出力を単一標本の一貫性よりも優先したことに起因すると議論されています。
    - NP理論では、P値のランダム変数が帰無仮説の下で単位均一分布に従うこと（またはそれ以上であること）が、検定のサイズ（タイプIエラー率）がαを超えないことを保証する有効性の特性として重視されます。

両者の主な違いと混同

- 多くの場合、**ダイバージェンスP値**は観測量であり、**意思決定論P値**は対応する確率変数であるとだけ認識されているか、あるいは両者が等価であると見なされています。しかし、これは一般的には真実ではありません。特に、**単一標本の一貫性基準**が**ダイバージェンスP値**に課せられ、反復抽出基準のみが**意思決定論P値**を導出するために使用される場合、pd;A > pt;A となることがあります。
- この論文は、P値を証拠の尺度として使うことに対する批判が、**意思決定論P値**とその基準に基づいていることを示し、これが**ダイバージェンスP値**には当てはまらないことを詳細に説明しています。批判は、**相性の良さ**と支持（support）の混同にも基づいています。
- **相性の良さ**は支持とは論理的に異なります。データがモデルに適合しないこと（反証）は、それだけでモデルの支持を意味しません。常に多くの他のモデルが同じデータに適合する可能性があり、支持の尺度を生成するには、これらの他のモデルに非常に厳しい制限を課す必要があります。**ダイバージェンスP値**は**相性の良さ**や対立の尺度であり、支持の尺度ではありません。
- NP理論は対立仮説の明確な指定を必要としますが、**ダイバージェンスP値**の導出や解釈には、必ずしも精密な対立仮説は必須ではありません。ただし、どの統計量を選択するかは、感度を高めたい対立仮説に依存します。

結論として

- P値は、モデルと観測データまたは基準モデルの下でフィルタリングされたデータとの**ダイバージェンス**の記述として導出される場合、**単一標本のアノマリー**を回避し、**単一標本の一貫性**を優先するため、データの**相性の良さ**を首尾一貫して示すことができます。
- **ダイバージェンスP値**は、意思決定論的なP値が首尾一貫しなくなる場合（例えば、区間仮説や境界点がある場合）に、それらと異なる可能性があります。
- 統計的推論や意思決定を行うためには、P値だけでなく、外部からの入力（例えば、誤差許容範囲や損失関数）や仮定に関する情報が必要です。研究報告において、統計的要約（P値や区間推定）のみを提供し、意思決定や推論を省略することが正当化される場合もあります。
- 記述的なアプローチは不可欠であり、**ダイバージェンスP値**は、データとモデル間の関係を、統計的な不確実性を考慮しつつ記述するための有用なツールとなります。従来の「有意性（significance）」や「信頼（confidence）」といった用語を「**相性の悪さ（incompatibility）**」や「**相性の良さ（compatibility）**」に置き換えることが推奨されています。

Divergence versus decision P-values A distinction worth making in theory - UMPU

はい、承知いたしました。いただいたソースと会話履歴に基づき、UMPU（一様最強力不偏）に関する事柄について、ご指定の訳語を使用して詳しく要約します。

UMPU (一様最強力不偏) に関する要約

UMPU（一様最強力不偏、uniformly most powerful unbiased）は、主にネーマン・ピアソン（NP）の意思決定理論の枠組みで使用される基準です。この理論の目的は、標本空間全体にわたる意思決定を、繰り返しのサンプリング特性に基づいて最適化することです。

UMPUは、提案された仮説やモデルを「棄却」または「受容」する意思決定ルールの選択基準として機能します。UMPU検定手続きは、「yが棄却域RA(α)にあるならば、FAを棄却する」という意思決定ルールを実装するために、「最適」（UMPU）なRA(α)を見つけようとします。この手続きは、以下の条件を満たすことを目指します:

1. **第一種過誤率の制御**: FAが真である場合（f ∈ FA）、意思決定ルールの第一種過誤率（誤ってFAを棄却する確率 f(RA(α)))が指定されたレベルαを超えない。
2. **不偏性**: 代替分布族Faltが真である場合（f ∈ Falt）、第二種過誤率（誤ってFAを受容する確率）がαを下回らない、あるいは等価的に、棄却する確率 f(RA(α)) が α 以上である。

UMPUテストは、第一種過誤率を制御しつつ、検定力（真の代替仮説の下で正しく棄却する確率）を最適化するように設計されています。

NP理論では、**意思決定論P値**は、UMPU手続きなど、繰り返しのサンプリング基準を満たすように導出された意思決定の中間計算として定義されます。具体的には、**意思決定論P値**は、観測データに対して仮説が棄却される最小のαレベルとして定義される場合があります。NP理論において、**意思決定論P値**が無仮説の下で一様分布に従う（または少なくとも近似的に上から支配する）という特性は、意思決定ルールのサイズ（第一種過誤率）がαを超えないことを保証するため、望ましい、あるいは定義上必要な特性と見なされます。

しかし、UMPU基準は**繰り返しのサンプリング**特性のみに基づいて定義・評価されており、**単一標本の観測値の特性は考慮されていません**。この点が、**意思決定論P値**が証拠の尺度として欠陥を持つ原因であると論じられています。UMPUが単一標本のコヒーレンスよりも検定力を優先することから、結果として得られる**意思決定論P値**が直感的な**単一標本のコヒーレンス基準に違反する**可能性があります。

特に、Schervish (1996) は、有界なパラメータ区間によって定義されるFAとUMPU手続きの場合、結果として得られる**意思決定論P値**が**部分集合コヒーレンス**（あるモデルMがより一般的なモデルAの制約をすべて含み、さらに多くの制約を課す場合、Mに対する証拠の尺度はAに対するそれを超えないべきであるという基準）に違反する可能性があることを示しました。このコヒーレンスの違反は、UMPU手続きが単一標本のコヒーレンスよりも検定力の最適化を優先した結果であるとされています。このような特性から、UMPUに由来する**意思決定論P値**は、データが仮説やモデルに対して提供する証拠の尺度としては不適切であると主張されています。P値を証拠の尺度とすることへの批判は、しばしばこのような意思決定論に基づいたP値定義に向けられています。

ダイバージェンスP値は、単一標本のコヒーレンス基準（例えば、データとモデルの適合度の幾何学的要約に基づく）を満たすように定義されるのに対し、UMPUに基づく**意思決定論P値**は繰り返しのサンプリング基準のみを考慮します。この異なる目標が、両タイプのP値が一般的に対応しない理由です。例えば、区間仮説の検定において、UMPU **意思決定論P値** (pHL)は、区間の端点におけるP値の平均として計算されるのに対し（パラメータが区間の外にある場合）、**ダイバージェンスP値** (pM)はそれらの端点におけるP値の最大値となります。また、パラメータが区間の内部にある場合、pMは1になりますが、pHLはそうならない場合があります。

UMPUのような繰り返しのサンプリング基準の最適化は、**ダイバージェンスP値**のような単一標本の記述的尺度とは異なる目標を持っています。**ダイバージェンスP値**をUMPU最適化された決定ルール（例: 「pM ≤ αなら棄却」）で使用すると、UMPU **意思決定論P値**を使用した場合と比較して検定力が低下する可能性があります。これは、UMPUが単一標本のコヒーレンスを犠牲にして約1ビットの情報（シャノン情報の意味で）を検定力のために得ているトレードオフと見なすことができます。

結論として、コヒーレンス基準（**ダイバージェンスP値**に関連）とUMPU基準（**意思決定論P値**に関連）は、すべての設定で同時に満たされるわけではありません。ソースでは、このような場合、コヒーレンスが優先されるべきだと論じられています。UMPU基準によって導出される**意思決定論P値**は、その定義と評価が繰り返しのサンプリングのみに基づいているため、単一標本に対するデータの証拠性を測る尺度としては欠陥があると考えられています。

Connecting simple and precise P-values to complex and ambiguous realities

ご提示いただいた翻訳を使用し、「Connecting simple and precise P-values to complex and ambiguous realities」（シンプルで精密なP値を複雑で曖昧な現実に接続する）という論文の内容を、ソースに基づき詳しく要約します。

この論文は、前論文「Divergence versus decision P-values: A distinction worth making in theory and keeping in practice」に対する討論者からの意見への返答（rejoinder）として書かれています。著者によると、討論者の誰もが、**ダイバージェンスP値**と**意思決定論P値**を区別するという前論文の主たる主張には同意しているようです。

この論文の核心的な主張は、**現実世界の複雑さと曖昧さに直面した場合、統計モデルと観察データとの間の関係を記述するための「P値」には、2つの異なる概念があり、これらを明確に区別し、特に一方の概念（ダイバージェンスP値）に基づいて、より慎重で無条件的なデータの記述を行うことが重要である**という点にあります。

主な内容は以下の通りです。

1. **統計理論と現実の乖離**
    - 数学は現実世界の問題解決における限定的な要素であり、その結論は、明示的・暗黙的を問わず、すべての仮定が正しい場合にのみ成り立つものにすぎません。
    - 統計的手法は、データ生成、管理、分析、報告における人間の公平性やバイアスがないことなど、多くの暗黙の仮定に依存していますが、これらの仮定はしばしば満たされません。
    - これらの疑わしい仮定に直面すると、P値や「統計的有意性」、**相性の良さ**の区間（confidence interval）といった標準的な統計量の意味合いが問題となります。標準的な解釈（例えば、タイプIエラー率やベイズ信頼区間のベッティングの根拠）は、それらの統計量を導出する際に用いられたすべての仮定（ターゲットモデルM）が正しいことに条件付けられています。これは、通常明記されない、しばしば破られる仮定（選択バイアス、データエラー、データ掘り起こしなど）を含みます。
    - これらの仮定に関する不確実性が大きい文脈では、形式統計量に依拠した推論や意思決定は誤解を招きます。
2. **無条件的な測定としての統計量**
    - 疑わしい仮定がある場合、統計量の教育、用語、解釈を、論理的に無条件的な形式に切り替える必要があります。
    - 「無条件的な」とは、統計量を導出するために使われた仮定が間違っていたとしても、その統計量の再解釈が正しい観察であるという意味です。例として、「私の時計は午後4時と言っています」という観察の記述は、時計が間違っている可能性を許容します。一方、「午後4時です」という主張は、時計が正しいという暗黙の仮定に条件付けられています。
    - 標準的な統計量の解釈は、後者の「午後4時です」という主張に似ており、ターゲットモデルM（および埋め込みモデルA）が正しいことに過度に依存しています。
3. **ダイバージェンスP値の役割**
    - **ダイバージェンスP値**（観察された**ダイバージェンス**から定義されるフィッシャー流またはネオ・フィッシャー流のP値）は、無条件的な記述を提供します。これは、より弱い仮定の集合A（埋め込みモデル）の下でのデータの投影と、より強い仮定の集合M（ターゲットモデル）の下でのデータの投影との間の**ダイバージェンス**を記述します。
    - この記述は、どちらのモデルが正しいかも仮定せず、Mのあらゆる違反の下でのエラー率やカバレッジ率を提供するものではありません。**ダイバージェンス**統計量の妥当性や有用性は文脈に依存し、Mが現実の近似にすぎない場合や、反事実的な世界を参照する場合でも有用でありえます。
    - 著者は、「記述的」（descriptive）や「**ダイバージェンス**」（divergence）という用語を、データやモデル間の関係の単一標本要約として解釈できるP値を指すために用いています。より狭義には、「**ダイバージェンス**」は、幾何学的な**ダイバージェンス**統計量（二乗和や逸脱度など）から導出されるP値を指し、これは幾何学的測定の順序的な記述です。
4. **意思決定論P値との違いと問題点**
    - NP理論に根ざす**意思決定論P値**は、繰り返し抽出の基準のみを評価基準とし、しばしば単一標本のコヒーレンス基準（subset coherenceなど）に違反する可能性があります。
    - **意思決定論P値**は、特定の対立仮説の下でのエラー率を最適化することに焦点を当てていますが、この最適化が単一標本の特性を犠牲にすることがあります。
    - NP理論は、決定規則の構成要素としてP値を使用し、標本空間全体での意思決定を最適化することを目指します。しかし、これはモデルに関する重要な不確実性がある場合には、形式的な決定を疑わしく、誤解を招くものにします。
    - **ダイバージェンスP値**は、コヒーレントな**ダイバージェンス**尺度から出発するため、ターゲットとする制約集合やモデルファミリーとデータとの**相性の良さ**を首尾一貫して示すことができます。
5. **相性の良さ・相性の悪さと支持**
    - P値はデータとモデルの間の**相性の良さ**や整合性の尺度であり、支持（support）の尺度ではありません。
    - データがモデルに適合すること（または矛盾しないこと）は、それ自体ではそのモデルへの支持を意味しません。常に多くの他のモデルが同じデータに適合する可能性があるためです。支持の尺度を生成するためには、これらの他のモデルに非常に厳しい制約を課す必要があります。
    - **相性の良さ**は支持よりも論理的に弱い概念です。支持は**相性の良さ**を意味しますが、**相性の良さ**は支持を意味しません。
    - P値を**相性の悪さ**または反証の尺度として扱うために、負の対数変換（**S値** = -logc(p)）が推奨されます。これにより、確率スケールから離れ、後方確率との混同を防ぎ、より等間隔な反証の尺度となります。
6. **用語の再考**

- 「統計的有意性レベル」としてP値を扱うことや、「信頼区間」を**相性の良さ**の区間と見なすことは、研究報告の歪みの主な原因であり、結論における過大な確実性（uncertainty laundering）につながります。
- P値や**相性の良さ**の区間を、「有意性宣言」「統計的決定」「信頼」といった混乱を招く追加要素から切り離し、文脈に関連する**ダイバージェンス**尺度の連続的な記述子として扱うべきです。
- 「有意性」を「**相性の悪さ**」に、「信頼」や「確信度」を「**相性の良さ**」に置き換えることは、より慎重な表現です。

7. **統計量の限界と解釈**

- P値は、モデルが想定したものと何かが一致しないことを示す予備的なチェックまたは診断にすぎません。
- どのような対立仮説に対する感度を高めたいかに応じて、統計量の選択は変わります。しかし、特定の対立仮説に焦点を当てすぎると、他の重要なモデル違反に対する感度が低下します。
- **意思決定論P値**の導出やNP理論が要求する、正確な対立仮説の指定は、現実世界の広範な不確実性においては誤解を招く可能性があります。
- 単一のスカラー量（P値を含む）では、**相性の良さ**や**相性の悪さ**といった多次元的な現実世界の概念を適切に捉えることはできません。
- 現実世界の思慮深い意思決定手続きには、P値以上のものが必要です。例えば、外部からの入力（許容誤差や損失関数など）や、P値の計算に使われた仮定が安全に真実とみなせるかどうかの情報が必要です。
- P値や**相性の良さ**の区間などの統計的要約を報告する際に、意思決定や推論を省略することが正当化される場合もあります。
- **ダイバージェンスP値**は、観察された統計量が仮説された分布のどこに位置するかを示す、**ダイバージェンス**尺度の順序的な要約です。
- これらの統計量は、真のデータ生成メカニズムに関する因果関係の説明（ランダム化など）から明示的に導出され、物理的なメカニズムの有効で公開された文書に基づいている場合にのみ、文脈に応じて解釈可能かつ信頼できるものとなります。このような保証がない場合、従来の「推論統計」の解釈は有害なフィクションとなり、データとモデルの関係に関するより慎重な記述に置き換える必要があります。

8. **結論として**

- P値に関する批判の多くは、**意思決定論P値**とその基準に基づいています。
- しかし、**ダイバージェンスP値**は、データの**相性の良さ**または**相性の悪さ**をコヒーレントに示すための有用なツールとなり得ます。
- 記述的なアプローチは不可欠であり、統計的な不確実性を考慮しつつ、観察結果とモデルや理論の関係を記述するための道具としてP値を捉え直すことが推奨されています。