

心理学的研究における 統計的有意性検定の適用限界

(葛西俊治 2006)

1) 統計検定の基礎からの解説 (p. 1-15)

なぜ「統計的有意＝真実」とは限らないのか？

2) 統計的有意性検定の適用限界 (p. 16-30)

「聖なる5%」の神話と「無記」とされた母集団の謎



心理統計の限界：「聖なる5%」が隠しているもの

なぜ「有意である」＝「真実である」とは限らないのか

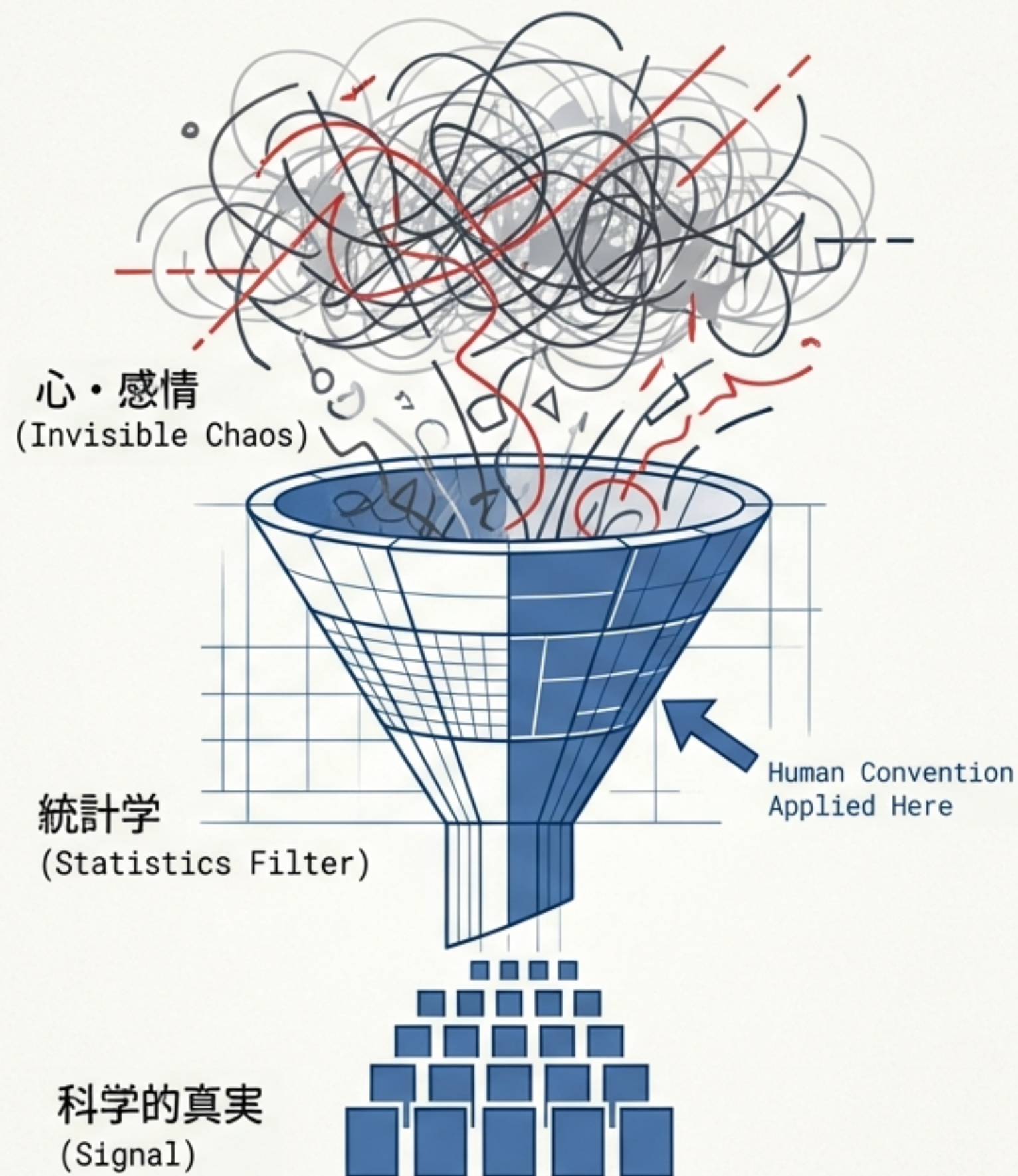
心理学の研究において、私たちは「統計的有意性」を科学的証明の証として信じてきました。しかし、その根拠となる「 $p < 0.05$ (5%水準)」という数字は、本当に絶対的な真理なのでしょうか？本スライドでは、統計的検定が抱える構造的な限界と、私たちが陥っている「数字の神話」を解き明かします。

心理学は「見えないもの」をどう証明するか

物理学と異なり、心理学が扱う「心」や「感情」は直接目に見えません。そこで登場するのが統計学です。

偶然の「ノイズ」と、意味のある「シグナル」を区別するためのフィルターとして統計的検定が使われています。

しかし、このフィルターの設定には、人間的な「決めつけ」が含まれていることを忘れてはいけません。



初心者のための「有意性検定」入門

帰無仮説（Nothing happened）：「その結果は、単なる偶然である」という仮説を立てます。

判定基準：その偶然が起きる確率が極めて低い（通常5%未満）場合、「偶然ではない＝意味がある（有意）」と判断します。



通常（Not Significant）



確率：約2.7%



5%未満なので「偶然」ではないとみなす

「聖なる5%」 という儀式

なぜ「5%」なのか？ それは自然界の法則ではなく、数十年前に決められた慣習に過ぎません。
わずか0.002の差で天国と地獄が分かれるこの現状は、科学というより「儀式」に近い状態です。

$p = 0.049$ (有意)



論文掲載 (Success)

差はわずか 0.002

$p = 0.051$ (有意差なし)



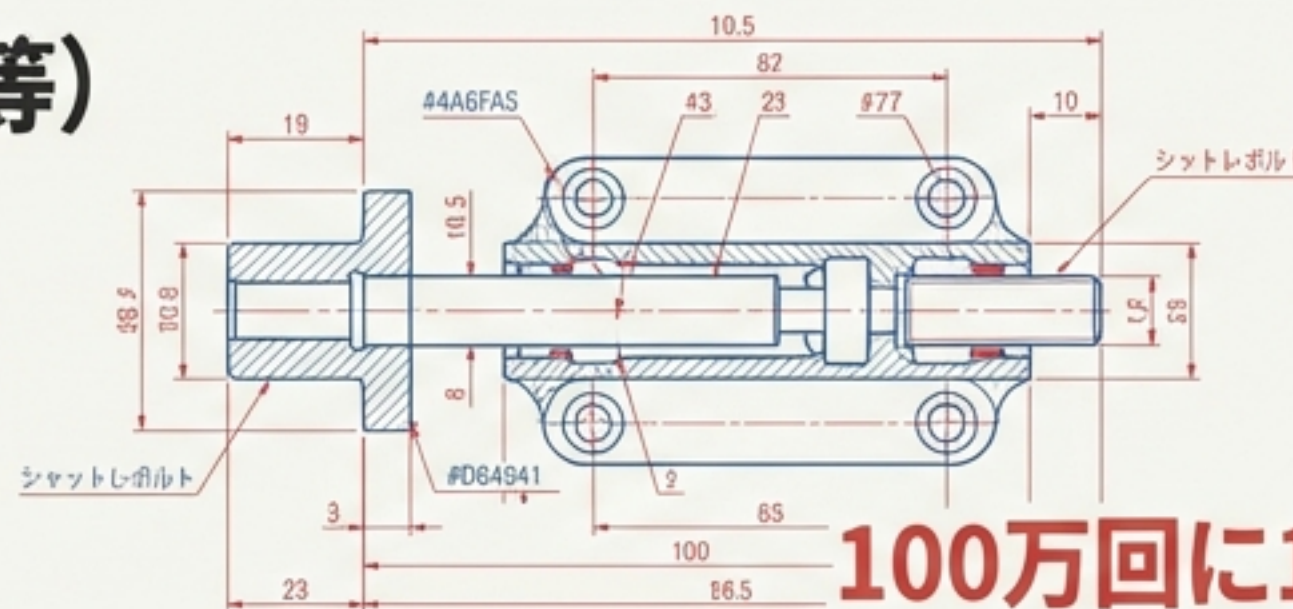
棄却 (Failure)

5%は本当に「厳しい」基準なのか？

Editorial
Architectural

工学（スペースシャトル等）

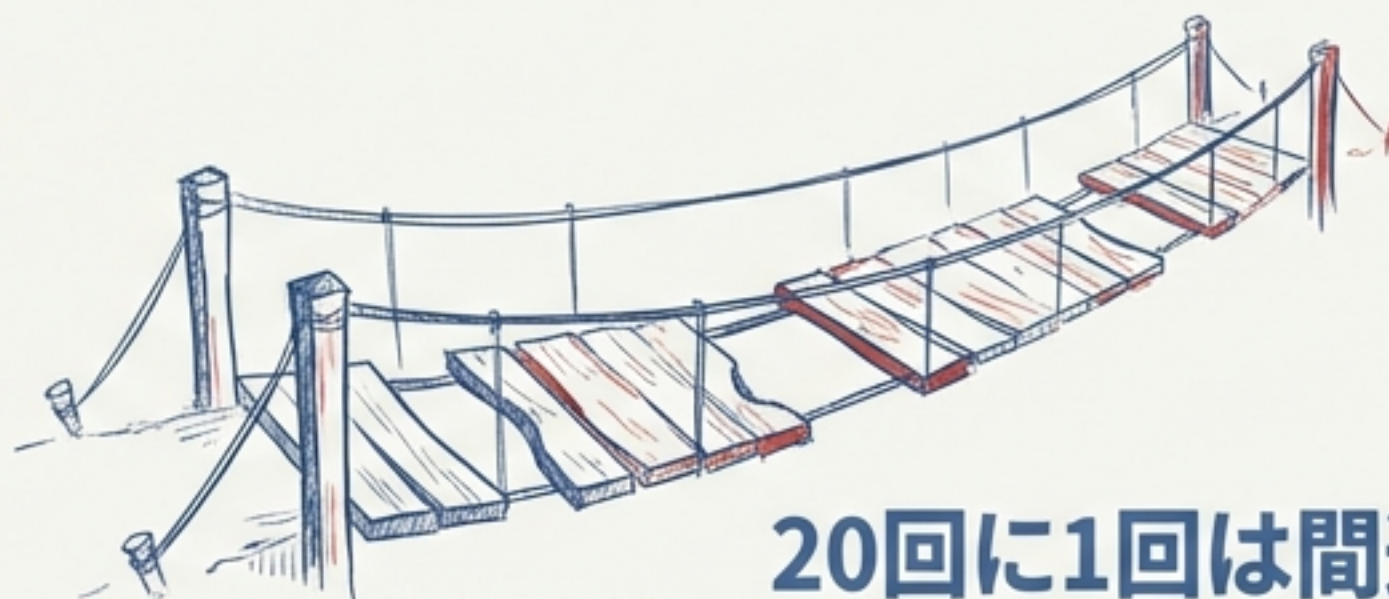
許容誤差 0.0001%



100万回に1回のミスも許されない

心理学

許容誤差 5.0%



20回に1回は間違っている可能性がある

衝撃的な事実: 心理学の専門誌に掲載された論文が20本あれば、確率論的にはそのうち1本は「誤り（偶然を真実と誤認）」である可能性があります。

「鉄球」と「人間」の決定的な違い

物理学の前提

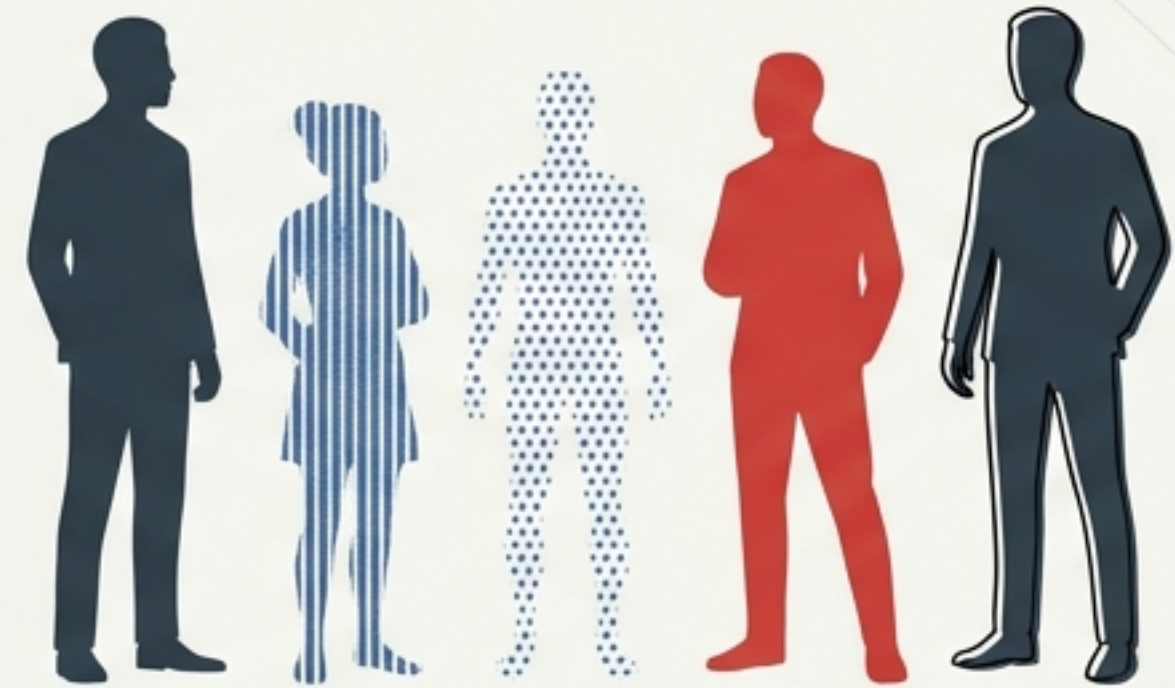
自然の斉一性 (Uniformity of Nature)



いつ、どこで落としても、
重力は同じように作用する。

心理学の現実

人間の多様性 (Human Diversity)



人間には個性、文化、歴史的背景がある。
AさんとBさんは同じ刺激に対しても異なる反応をする。

心理学統計の罫は、人間に対して無理やり「鉄球と同じような均一性」を仮定してしまったことにあります。

「ニュー・ルック」のパラドックス

1940年代の「ニュー・ルック心理学」は、貧しい家の子供はコインを物理的サイズより大きく知覚するなど、知覚が価値観や環境に左右されることを証明しました。

しかし、統計手法は依然として「人間は均一だ」という前提のままです。



研究内容（主観の発見）と研究方法（客観の強制）の矛盾

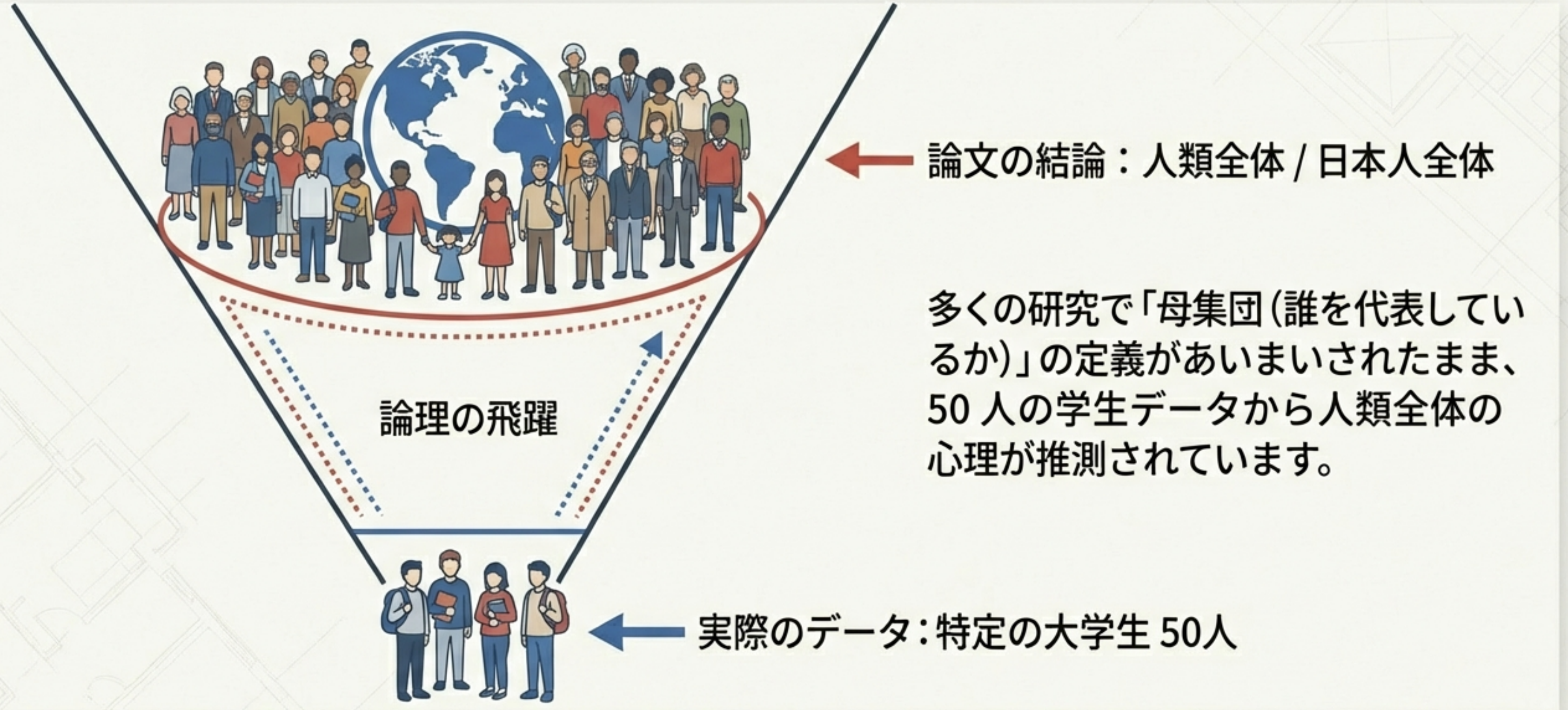
「平均的な人間」など存在しない

統計的検定は「平均値」で物事を判断します。個人の多様性は、統計上では単なる「誤差（エラー）」として切り捨てられます。



統計的有意性を追求するあまり、私たちは個々の人間のリアリティを失っています。

私たちは一体「誰」を調べているのか？



ラベルの罫と「無記 (Non-description)」

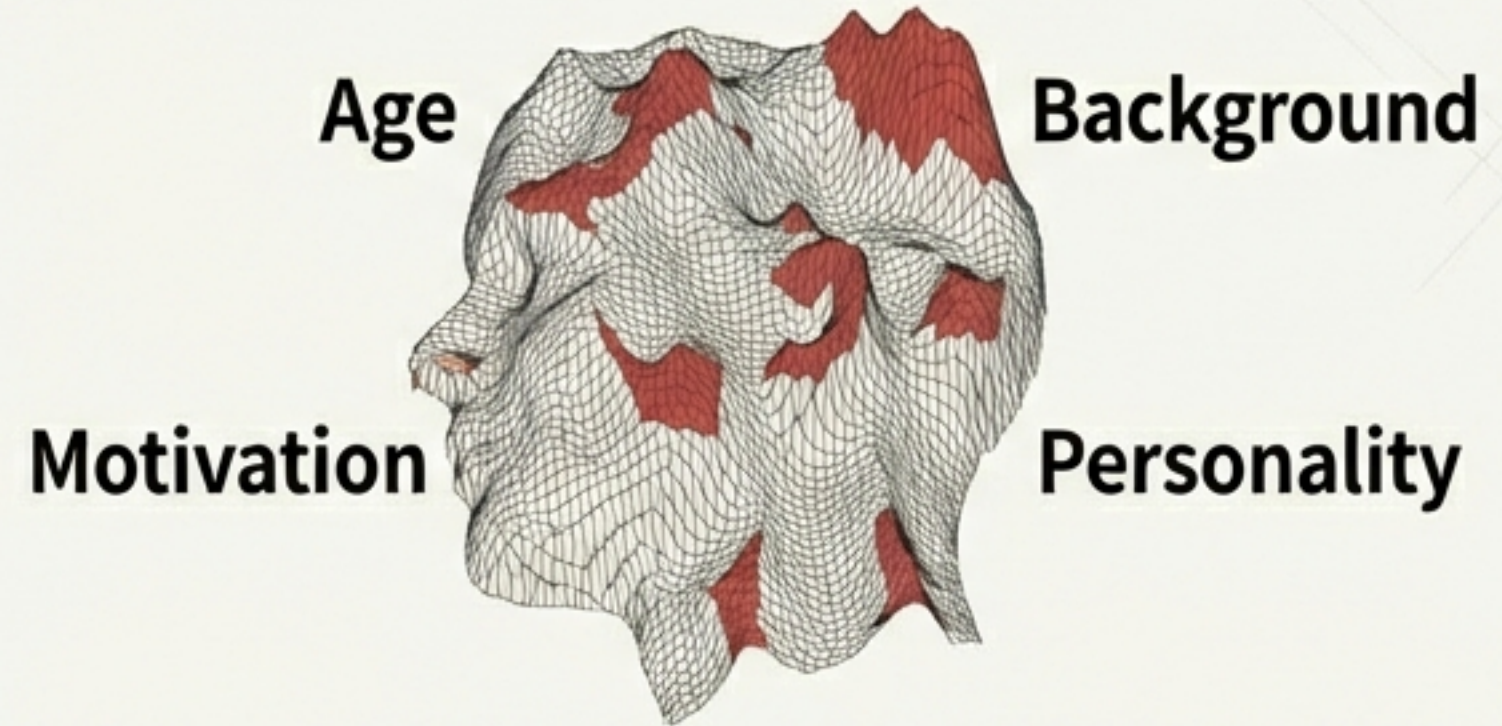
Editorial Architectural

地図は現地ではない (The map is not the territory)

論文上のラベル (The Map)

● 学生 (Student)

実際の人間 (The Territory)



「学生」というラベルは、その人の年齢、背景、動機、性格という広大な「現地」を隠してしまいます。多くの論文では、被験者が具体的にどんな人々かという記述が抜け落ちていきます（無記）。

誤りのシーソー：第1種の過誤と第2種の過誤

統計的決定には2種類のリスクがあります。研究者はこのバランスを意図的に選ぶ必要があります。



第1種の過誤 (Type I Error)

誤警報：何もないのに「ある」とする

厳しくしすぎると...

判定基準
(Significance Level)



第2種の過誤 (Type II Error)

見逃し：あるのに「ない」とする

甘くしすぎると (5%など) ...

解決策①：「無記」からの脱却

統計的検定を正しく使うための第一歩は、対象を明確に定義することです。

Research Paper

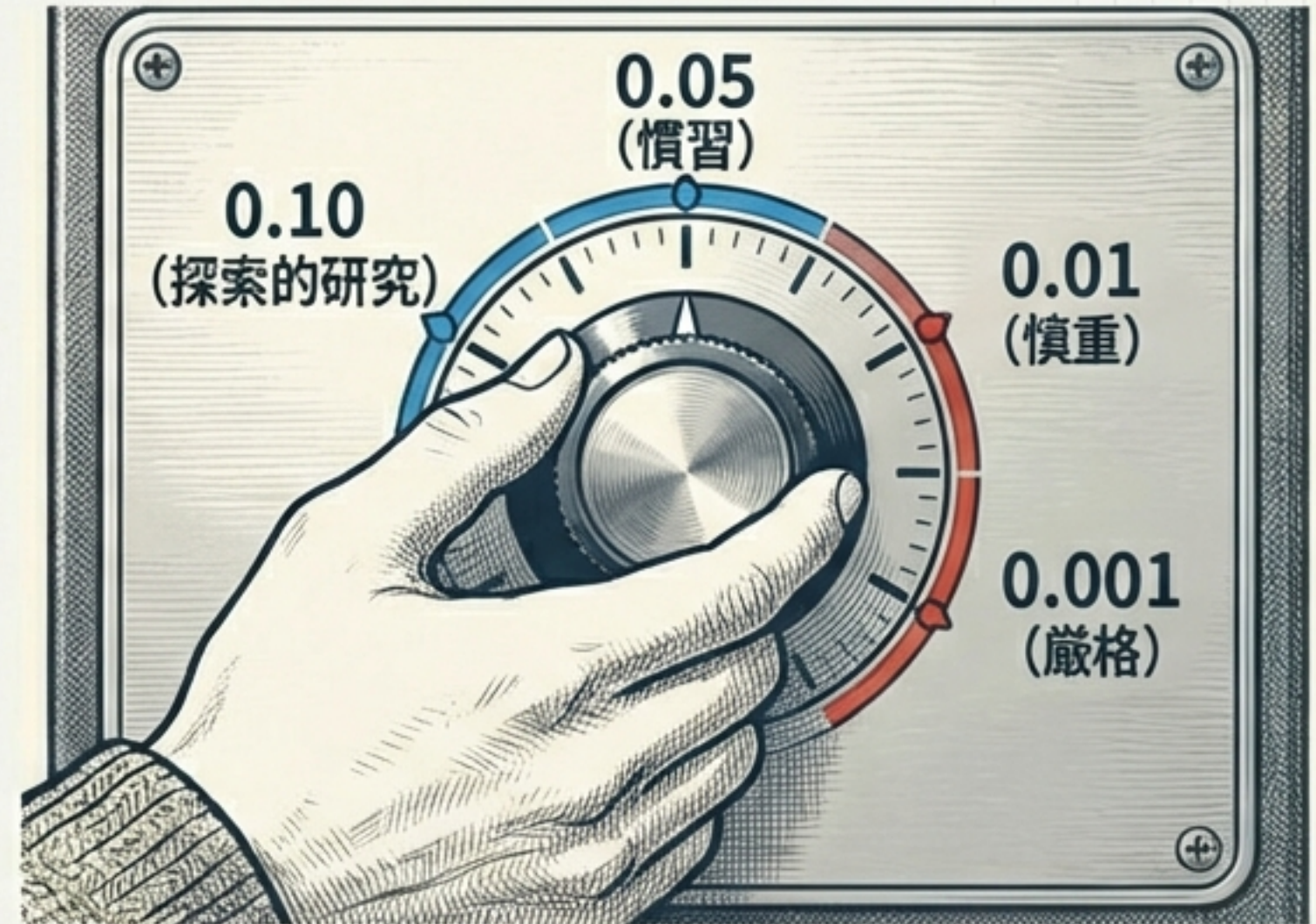
Target Population

~~人間 (Humans)~~ 関東地方の私立大学・文学部2年生
年齢: 19-21歳

- 母集団の明示: 適用範囲を限定する
- 属性の記述: ラベルに頼らず背景を記述する

解決策②：有意水準を「選ぶ」

「とりあえず5%」という思考停止をやめ、研究目的に応じて基準を選ぶべきです。



新薬開発など見逃し厳禁なら → **10%**?

定説を覆すなら → **0.1%**?

「なぜその基準か」を説明できることが重要。

質的研究の復権

数字の限界を補うためにこそ、物語（ナラティブ）が必要です。



量的研究
(Quantitative)

Noto Serif JP Regular
一般化・パターン

質的研究
(Qualitative)

Noto Serif JP Regular
文脈・個人のリアリティ

統計は「宗教」ではなく「道具」である

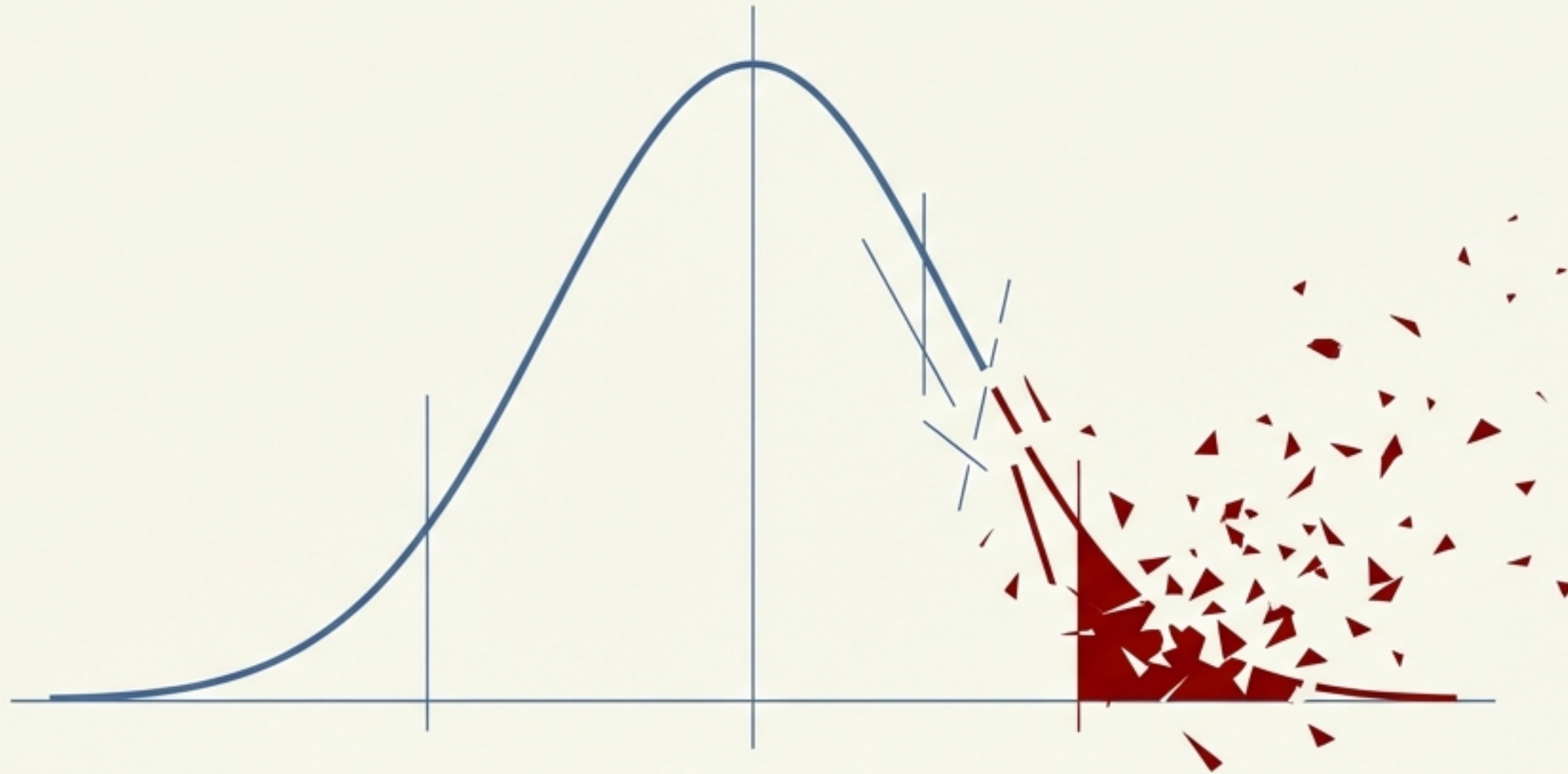
統計的有意性は、真実を自動的に吐き出す魔法の箱ではありません。
それは「ここまでは偶然かもしれない」という境界線を引くための、
人間が作った道具に過ぎません。



数字の背後にいる「**生きた人間**」を見失わないこと。
それが、真に科学的な心理学への道です。

心理学的研究における統計的有意性検定の適用限界

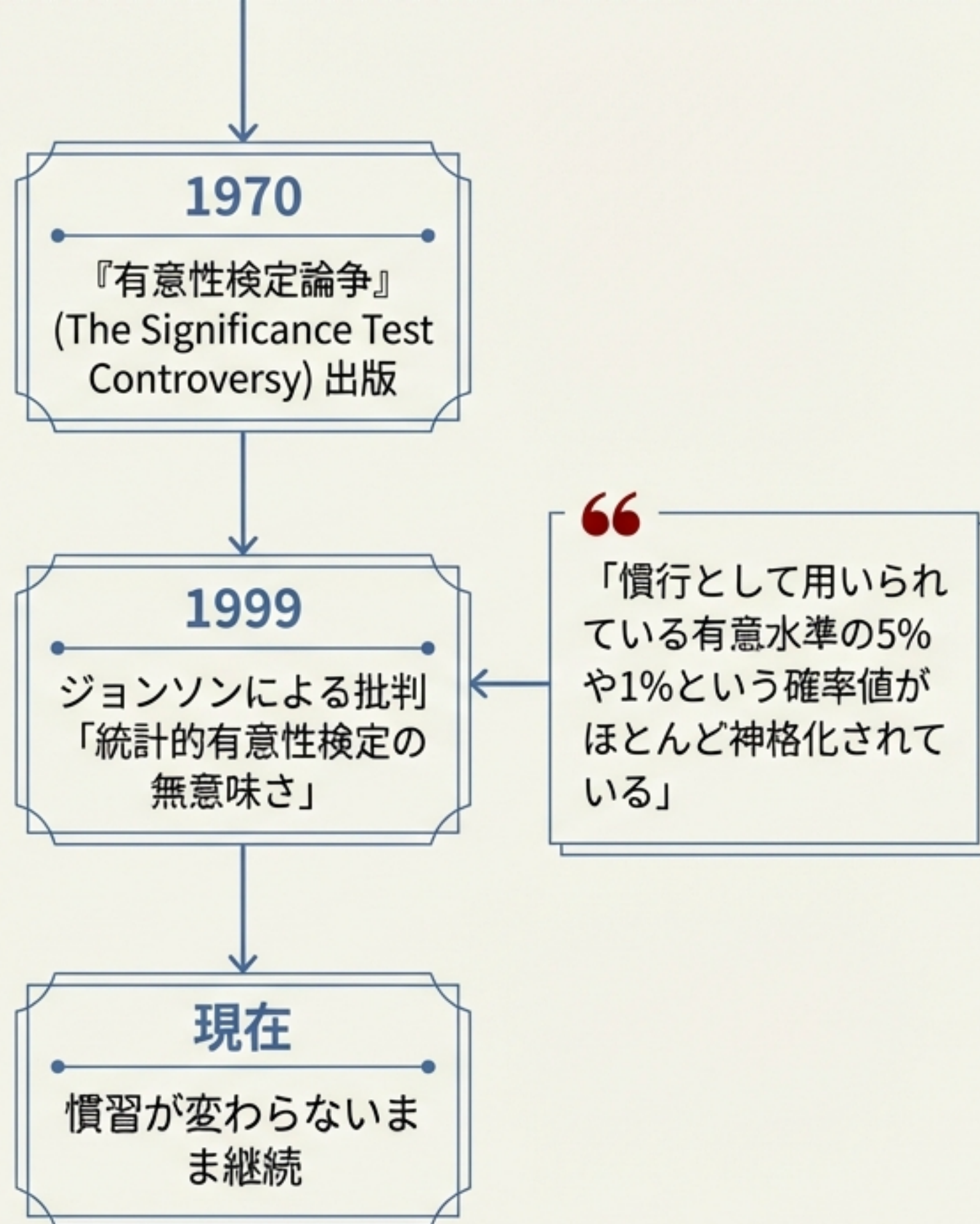
「聖なる5%」の神話と、「無記」された母集団の謎



Based on: 葛西俊治 (2006). 札幌学院大学人文学会紀要, 第79号.

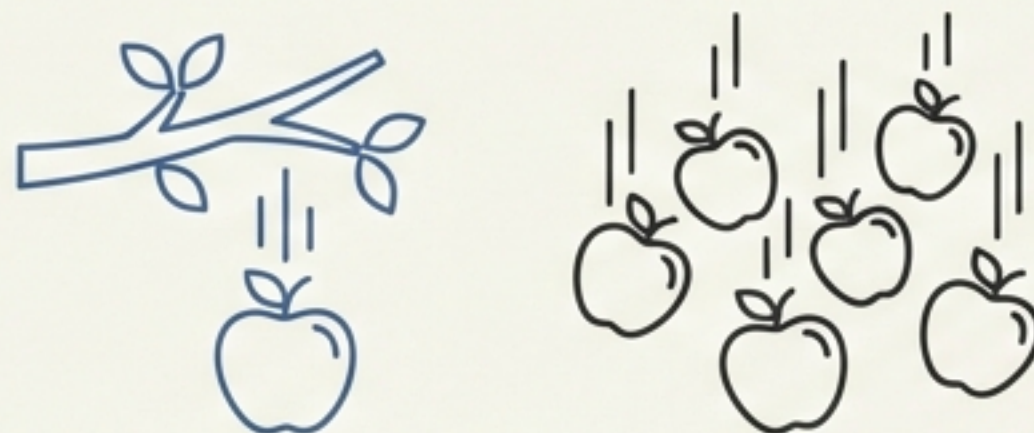
なぜ私たちは「 $p < .05$ 」を信仰するのか？

- 心理学論文の不文律：帰無仮説が5%や1%の水準で棄却されること。
- しかし、この「聖なる数値」への疑義は古くから存在する。
- 1999年、ジョンソンは「**統計的有意性検定の無意味さ**」を指摘したが、乱用は止まらない。
- **Meehl** (1997): 問題は統計学 (Statistics) ではなく、**認識論** (Epistemology) にある。



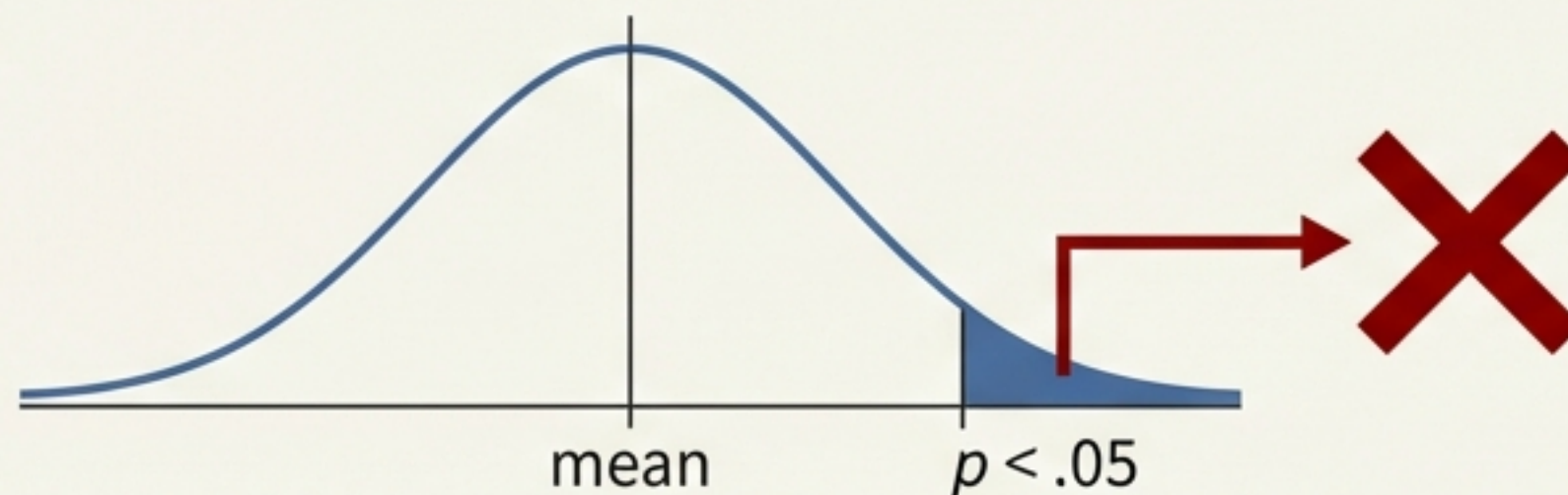
科学的推論の基礎：「自然の斉一性」と統計学

J.S.ミルの論理（自然科学）



公理：「自然の斉一性 (Uniformity of Nature)」
自然は一定不変であるため、少数の実験で一般化が可能。

統計学の論理（心理学）



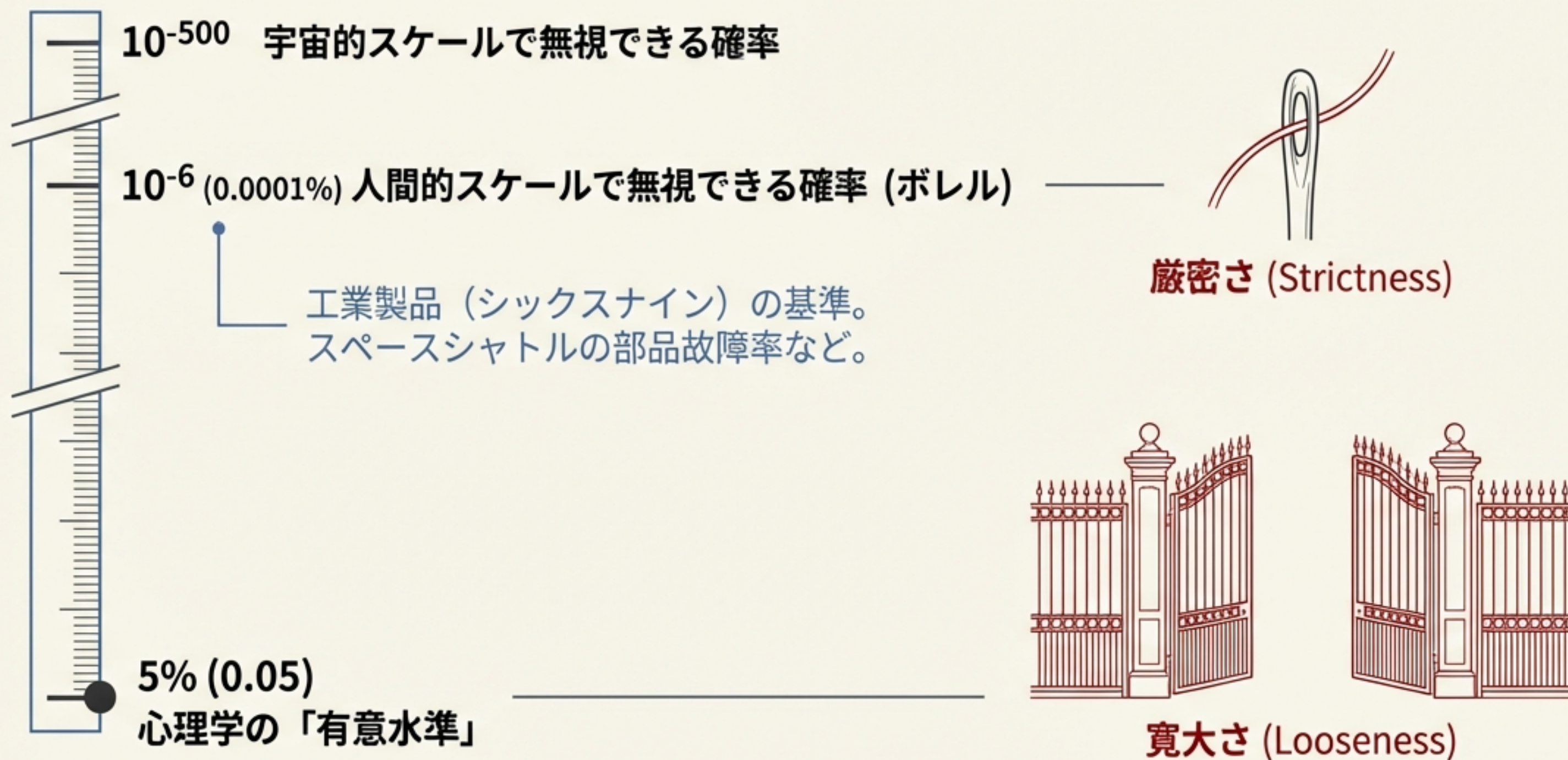
代替案：「確率的判断」

確率が極めて小さいこと（5%未満など）は、起きないものとみなす。

帰納の限界

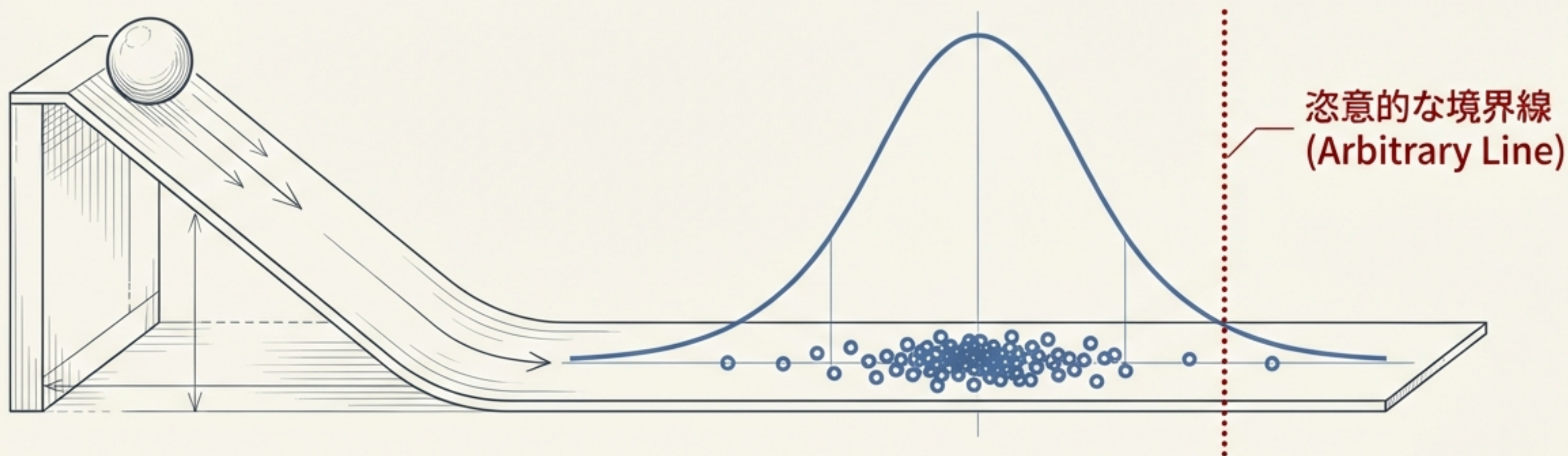
何度実験しても
「次は違う結果」
違う結果」になる
可能性は排除でき
ない（ヒューム／
ミルの問題）。統
計学はこの難問
を「確率」で回避
しようとした。

「5%」は本当に「めったに起きない」確率か？



5% (1/20) という確率は、科学的厳密さからすれば「極めて寛大」で「大雑把」な数値に過ぎない。

鉄球の実験：確率分布の実感



坂から鉄球を転がす実験を考える。平均1メートルで止まるが、毎回誤差がある。

統計的判断：「ずっと遠くまで転がっていく確率は低いから、起きないものとする」。

しかし、その線をどこに引くか？ 5%は慣習に過ぎない。もし「万に一つ（0.01%）」を基準にしたら、従来の心理学研究の成果はほぼ壊滅するだろう。

「慎重さ」の代償：第二種の誤りの危険性

第一種の誤り (α)
(厳しいp値)



第二種の誤り (β)
(見逃しのリスク)

見逃し (Missed Discovery)

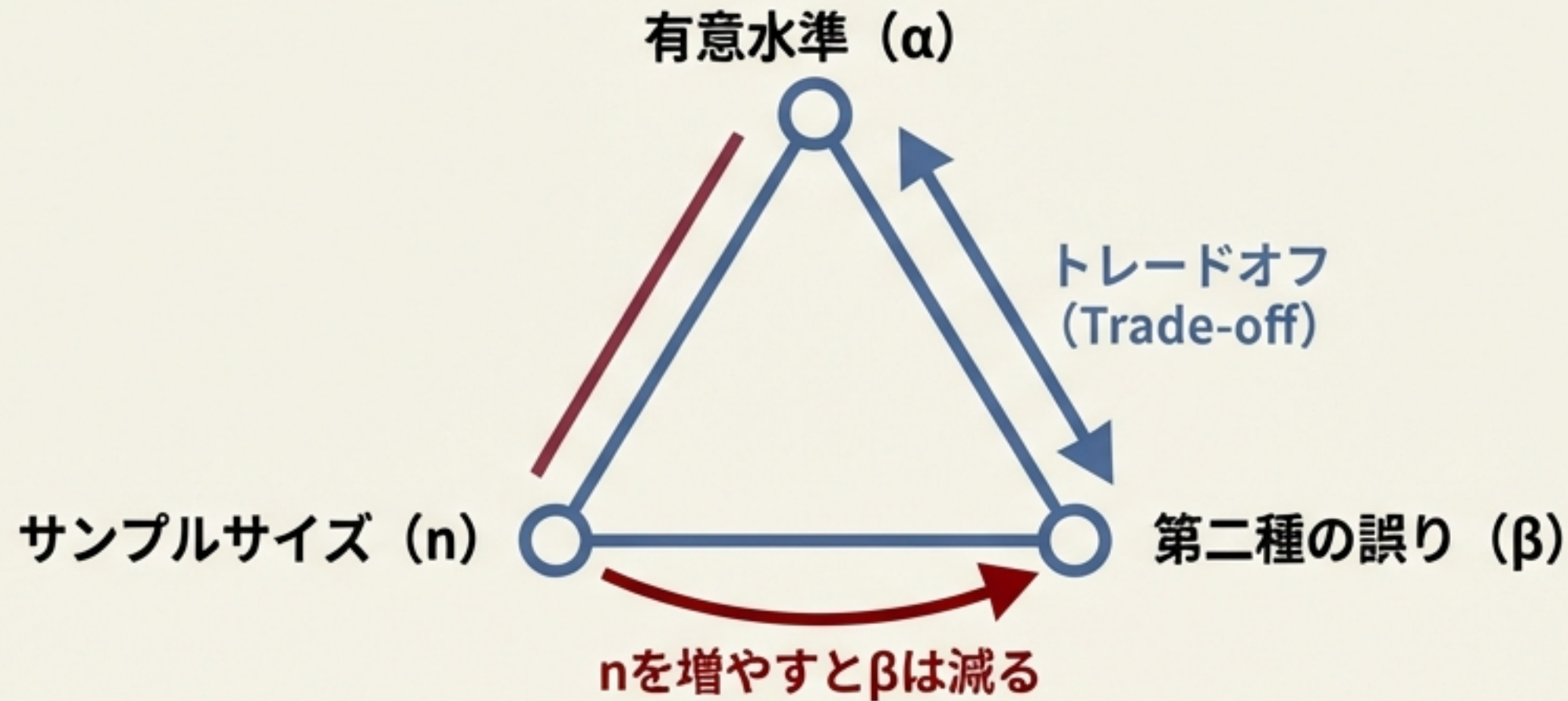
新薬開発のジレンマ：

α を0.001%にして極端に慎重になると、実際に効く薬でも「効果あり」と認められず、開発が中止される (β の増大)。

時には α を10%や30%に設定してでも、可能性を見逃さない方が良い場合がある。

厳密すぎる基準は、真の発見を妨げる可能性がある。

研究者が操作すべき3つのパラメータ



統計的検定は自動的な計算ではない。研究者による「判断」が必要である。

永田（2003）の指摘：

1. α （有意水準）を小さくすると、 β は大きくなる。
2. サンプルサイズ（標本数）を大きくすると、 β は小さくなる。

結論：慣習的な5%に固定する理由はない。研究目的に応じてバランスを設計せよ。

【要請 1】 パラメータ設定の根拠を明示せよ

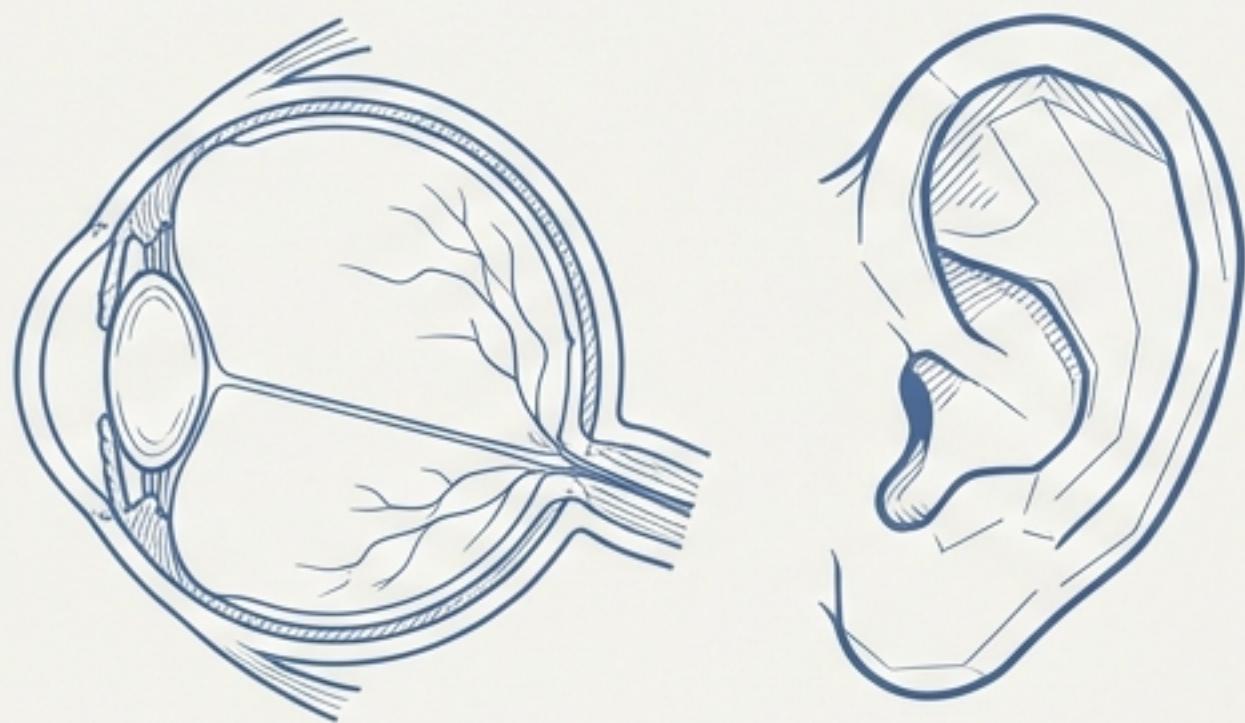
論文記述チェックリスト

- ☐ 有意水準はなぜその値（5%など）なのか？
研究テーマに即して吟味したか？
- ☐ β 過誤（見逃し）への配慮は適切か？
- ☐ 適切なサンプルサイズか？

これらが記述されていない論文は、**妥当性に疑義がある。**

心理学の変遷：感覚から「ニュールック」へ

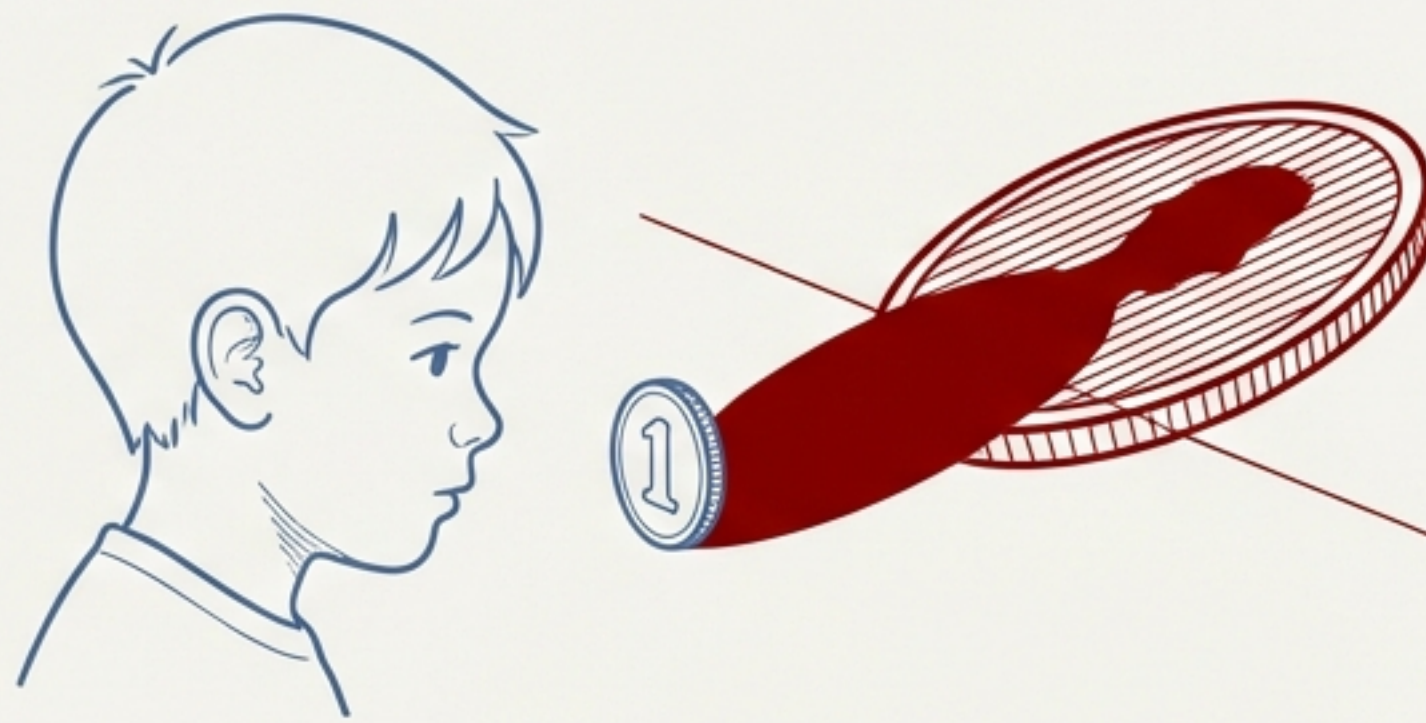
初期：精神物理学 (Fechner, Weber)



対象：視力、聴力、反応速度。

前提：生物としての「ヒト」の器官は共通（斉一性が成り立つ）。統計が機能しやすい。

転換点：ニュールック心理学 (New Look, 1940s)



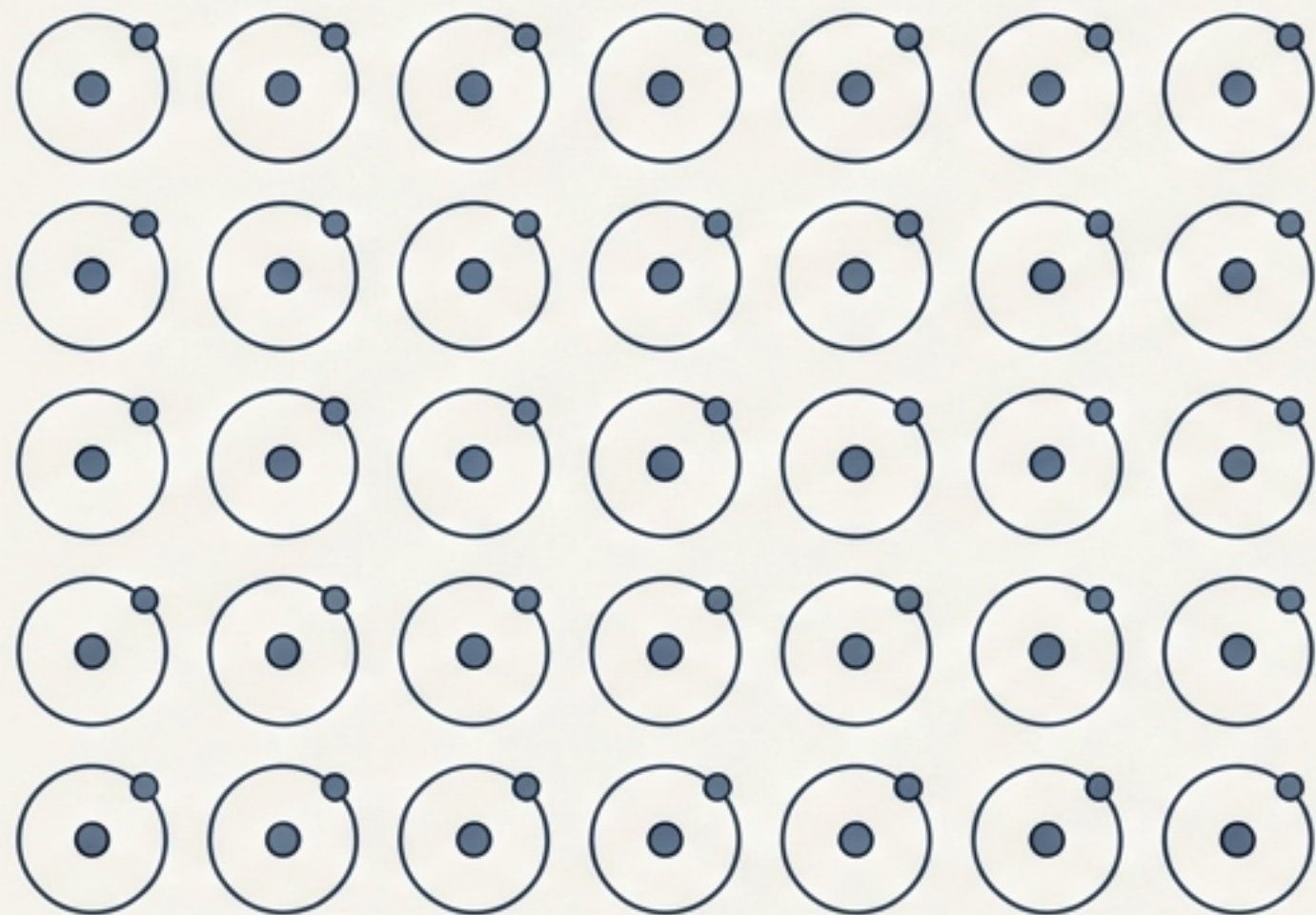
発見：貧しい子どもの方がコインを大きく知覚する。

意味：知覚は「人格的・社会的要因」に影響される（多様性）。

問題：研究対象は「多様」になったのに、方法は「斉一性」を前提とした統計検定のまま止まっている。

最大の誤謬：「人間の斉一性」という幻

自然の斉一性 (J.S. Mill)



水素原子はどこでも同じ振る舞いをする。

人間の多様性 (Human Diversity)



人間は歴史的・文化的・社会的に多様な
「個性記述的 (Idiographic)」存在である。

心理学は「A大学の学生」のデータを「全人類」に適用するという錯覚（人間の斉一性）に陥っている。

母集団は誰か？「50人の学生」の謎



多くの研究は、手近な学生50人のデータから、暗黙のうちに
「人類全体」の真理を語ろうとする。

「無記 (Muki)」：語られない母集団

[illegible]

一般意味論：「地図は現地ではない」。
ラベル（学生）と実態（多様な個人）は違う。

無記の構造：多くの論文は母集団を定義しない。「被験者」というラベルの裏にある、個別の属性（年齢、階層、文化）を捨象する。

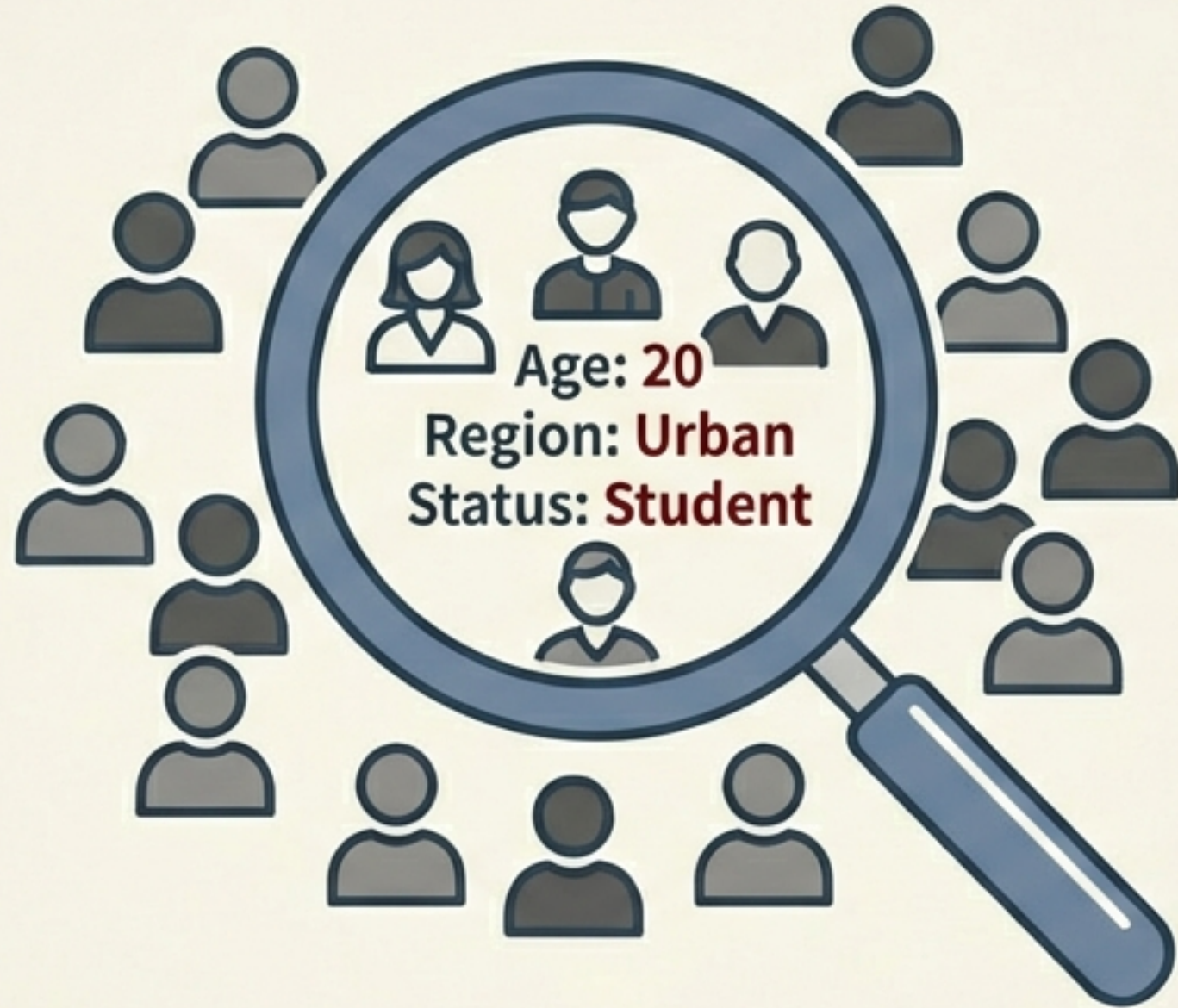
結果：実態のない抽象的な「一般論」が生産される。

「無記」がもたらす砂上の楼閣



比喩的解釈 (Figurative Interpretation)：厳密な母数推定ができないため、結論は科学的真理ではなく「なんとなく人間に当てはまりそうな話」という”たとえ話”に堕してしまう。

【要請2】 属性と母集団を定義せよ



必須記述項目：

1. 回答者の属性（層別化されているか？ 緣故的抽出か？）
2. そのラベル（「学生」など）は実態を反映しているか？
3. この検定は「どの母集団」を推定しているのか？

原則：限定的な標本から得られた知見は、限定的な母集団にのみ適用されるべきである。

自動的な検定から、思考する心理学へ

統計的有意性検定は「万能の真理判定機」ではない。
それは、特定の母集団、特定の $\alpha \cdot \beta$ 設定の下でのみ機能する「限定的なツール」である。

我々の目指す道：

- 「**聖なる5%**」の呪縛を解く
- 「人間の斉一性」を疑い、多様性に向き合う
- 自らの研究の「適用限界」を誠実に語る

