

代理を目標にすると、代理でなくなる

— 繋がりを切るものは四つあり、柵は誤差の裾である — [論文186]

山岸 勇輝 (Yuuki YAMAGISHI)

Yamagishi Complexity Torus Lab, Japan | 山岸勇輝ラボ

2026年8月

— From thinking to syncing. —

要旨

測るものを目標にすると、それは測るものでなくなる——この言い回しは、しばしば「人が制度を出し抜くから」と説明される。本稿は、そこに誰もいなくても起きる 三つがあることを書く。新しい数学定理も新しい統計手法も主張しない。第一に、雑音のある代理で選抜すると、選ぶほど雑音を拾う。雑音の標準偏差が **1.0** のとき、上位 **0.1%** を代理で選ぶと目標の平均は **2.378** にとどまり、目標そのもので選んだ **3.359** に **0.981** 届かなかった（第2節）。第二に、これが柵である。誤差の裾が軽ければ、代理を強く最適化するほど目標も良くなる。ところが裾が重いと、最適な最適化量があり、それを越えると悪くなる——コーシー誤差では上位 **20%** の **0.528** が頂点で、上位 **0.01%** では **0.005**、すなわち無作為と変わらない（第3節）。第三に、相関そのものが裾では効かない。相関 **0.99** の正規対でも、上位 q での一致率は **0.901** から **0.765** へ落ち続けた（第4節）。第四に、これが本稿の芯である。代理は目標を予測するが制御しない。相関 **0.995** の設定で、 P を観測して **2.5** だったときの目標の平均は **2.477**、ところが P を **2.5** に書き込んだときの目標の平均は **-0.0001**——差は値の全部である（第5節）。

本稿の射程：新しい数学定理も新しい統計手法も主張しない——選抜による偏り、正規対に裾の依存が無いこと、観測と介入の区別は、いずれも標準的である。また、いかなる制度・組織・政策についても主張しない——扱うのは、代理量と目標量のあいだの数学だけである。論文**183**との関係：**183**は同じ表が二つの正しい答を持つことを示し、決めるのは矢印だと書いた。本稿の第四の根はその機械をそのまま使う——新しい型ではなく、同じ型が代理量のところにも現れるという報告である。語について：本体系で「指標」はこれまで一貫して表現論の指標であった（論文70・152・181）。本稿の「代理」はそれとは無関係である。加えたのは四つの根を「何が繋がりを切るか」で分けたこと、柵が誤差の大きさではなく裾にあることを測ったこと、裾の依存の無さを五桁にわたって追ったこと、観測と介入の差を数で並べたことである。

1. 記法と位置づけ

本当に良くしたい量を目標、実際に測る量を代理と呼ぶ。以下、目標は標準正規に従うものとし、代理は目標に誤差を足したものとする。数値はすべて自前の抽出であり、標本数を明記する。

位置づけ：論文164は「確率」に四つの住所があると書き、論文183は同じ表が二つの正しい答を持つことを示した。本稿はその隣で、測ることと良くすることの隙間を測る。

2. 雑音のある代理で選ぶと、選ぶほど雑音を拾う

検算 2.1：目標を標準正規、代理をそれに雑音を足したものとし、200 万点から上位を選ぶ。

雑音の標準偏差	上位	代理で選んだ目標の平均	目標で選んだ場合
0.0	0.1%	3.359	3.359
0.5	0.1%	2.996	3.359
1.0	10%	1.241	1.755
1.0	1%	1.890	2.667
1.0	0.1%	2.378	3.359
2.0	0.1%	1.503	3.359

(証明終)

観察 2.2

選抜が厳しいほど、選ばれたものに占める雑音の割合が増える。ここには誰もいない——出し抜こうとする者も、制度も、意図もない。代理に誤差があるというだけで起きる。

3. 柵は、誤差の大きさではなく裾である

検算 3.1：同じ設定で、誤差を正規（軽い裾）とコーシー（重い裾）に取り替え、選抜を強くしていく。

上位	正規の誤差	コーシーの誤差
50%	0.566	0.418
20%	0.991	0.528
10%	1.244	0.336
1%	1.885	0.024
0.1%	2.383	-0.007
0.01%	2.752	0.005

(証明終)

観察 3.2

正規の誤差では、強く最適化するほど目標が良くなる（単調である）。コーシーの誤差では上位 20% が頂点であり、そこから先は下がって、上位 0.01% では 0.005——無作為に選ぶのと変わらない。

観察 3.3

分けているのは誤差の大きさではない。第2節では雑音を 2.0 まで大きくしても単調のままだった。分けているのは裾である——重い裾では、代理の最上位を占めるのが目標の高い個体ではなく、誤差がたまたま巨大だった個体になる。

観察 3.4

したがって「最適化は良いことだ」も「最適化は危険だ」も、どちらも半分の文である。正しくは、どこまで最適化してよいかは誤差の裾が決める。論文157 が最小二乗をコーシー誤差で壊したのと同じ場所に、同じ理由で柵が立っている。

4. 相関そのものが、裾では効かない

検算 4.1：相関 ρ の正規対 (X, Y) を 800 万点取り、 X が上位 q にあるとき Y も上位 q にある割合を測る。

ρ	上位 10%	上位 1%	上位 0.1%	上位 0.01%
0.50	0.325	0.128	0.057	0.015
0.90	0.689	0.542	0.442	0.381
0.99	0.901	0.850	0.810	0.765

いずれの行も標本は 800 以上である。(証明終)

観察 4.2

相関 0.99 でも、先端へ行くほど一致率が落ち続ける。正規対には裾の依存が無い—— ρ が 1 未満である限り、極限では一致率は 0 になる。ただし本稿の計算はその極限に到達していないし、到達したとも言わない。示したのは、五桁進んでもまだ落ち続けているということである。

観察 4.3

相関は真ん中で測られる。最適化はこちらを先端へ連れて行く。測った場所と使う場所が違う——これは誤差の話ですらなく、領域の話である。

5. 芯：代理は予測するが、制御しない

観察 5.1

代理が目標から生じる場合を考える。矢印は目標から代理へ向かい、 $P = T + 0.1\varepsilon$ である。観測上、代理は目標をよく予測する（相関 0.995）。

検算 5.2：400 万点。

P の値	標本数	観測 $E[T \mid P]$	介入 $E[T \mid \text{do}(P)]$
1.0	38645	0.989	−0.0001
2.0	8912	1.977	−0.0001
2.5	2789	2.477	−0.0001

(証明終)

観察 5.3

代理を見て高ければ、目標は高い。代理を書いて高くしても、目標は動かない——矢印が目標から代理へ向かっているからである。差は 2.477 対 −0.0001、すなわち値の全部である。

観察 5.4

この型は本体系の新顔ではない。論文183 は同じ表が二つの正しい答を持つことを示し、決めているのは矢印であって表の欄ではないと書いた。本稿が加えるのは、同じ型が代理量のところにも現れるという一点だけである。

6. 点呼：何が繋がりを切るのか

観察 6.1

根	何が切るか	現れ方	本稿の検算
回帰的	誤差そのもの	選抜が誤差を拾う	第2・3節
極值的	測った領域を出る	裾では相関が効かない	第4節
因果的	矢印の向き	予測するが制御しない	第5節
敵対的	相手がいる	代理を直接動かされる	測っていない

観察 6.2

四つのうち三つは、誰もいなくても起きる。「人が制度を出し抜くから」という説明は、四番目にしか当たらない。前の三つには意図も主体も要らず、誤差・領域・矢印という数学だけで起きる。

観察 6.3

そして対処も四つとも違う。回帰的な根には、誤差を減らすか最適化を止める場所を決めることである（柵は裾）。極値的な根には、使う領域で測り直すことである。因果的な根には、介入で測ることである——観測をいくら増やしても足りない。敵対的な根には、相手の側の話が要る。一語で呼ぶから、対処も一つだと思ってしまう。

7. 結論

1. 選抜は誤差を拾う（検算2.1）：上位 **0.1%** で **2.378** 対 **3.359**。
2. 柵は誤差の裾（検算3.1）：正規では単調、コーシーでは上位 **20%** が頂点。
3. 裾では相関が効かない（検算4.1）： $\rho = 0.99$ でも **0.901** から **0.765** へ。
4. 予測と制御は別（検算5.2）：**2.477** 対 **-0.0001**。
5. 三つは誰もいなくても起きる（観察6.2）。

結び：「この指標は目標にした途端に壊れた」と言われたら、次に問うべきは何が繋がりを切ったのかである。誤差か、領域か、矢印か、相手か。そして最初の三つには、出し抜こうとした者は一人もいない。

参考文献

1. Goodhart, C. A. E. (1975). *Problems of Monetary Management: The U.K. Experience*. Papers in Monetary Economics, Reserve Bank of Australia.
2. Campbell, D. T. (1979). *Assessing the Impact of Planned Social Change*. Evaluation and Program Planning, 2(1).
3. Sibuya, M. (1960). *Bivariate Extreme Statistics*. Annals of the Institute of Statistical Mathematics, 11(3).
4. Capen, E. C., Clapp, R. V., Campbell, W. M. (1971). *Competitive Bidding in High-Risk Situations*. Journal of Petroleum Technology, 23(6).
5. Pearl, J. (1995). *Causal Diagrams for Empirical Research*. Biometrika, 82(4).
6. Manheim, D., Garrabrant, S. (2018). *Categorizing Variants of Goodhart's Law*. arXiv:1803.04585.

7. 山岸勇輝 (2026). 「不確定性原理」は一つの定理ではない [論文129]. Zenodo. 10.5281/zenodo.22079681
8. 山岸勇輝 (2026). カルバック=ライブラーは距離ではない [論文137]. Zenodo. 10.5281/zenodo.22081369
9. 山岸勇輝 (2026). 重なりか、隔たりか [論文139]. Zenodo. 10.5281/zenodo.22081801
10. 山岸勇輝 (2026). 二乗の点呼 [論文157]. Zenodo. 10.5281/zenodo.22091931
11. 山岸勇輝 (2026). 「確率」は四つの住所を持つ [論文164]. Zenodo. 10.5281/zenodo.22092815
12. 山岸勇輝 (2026). 柵は線形性である [論文171]. Zenodo. 10.5281/zenodo.22094287
13. 山岸勇輝 (2026). 「次元の呪い」は一つの呪いではない [論文182]. Zenodo. 10.5281/zenodo.22104170
14. 山岸勇輝 (2026). 同じ表が、二つの正しい答を持つ [論文183]. Zenodo. 10.5281/zenodo.22104230
15. 山岸勇輝 (2026). 両立しなくなるのは、三つ目の選択肢からである [論文184]. Zenodo. 10.5281/zenodo.22105459
16. 山岸勇輝 (2026). 欠陥を分類しているのは行き先の空間である [論文185]. Zenodo. 10.5281/zenodo.22105496

— *From thinking to syncing.* —

山岸 勇輝 (Yuuki YAMAGISHI)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)