Barkhausen AI

METHOD NOTE · BA-MN-4

Base rates and regression to the mean in visibility claims

Two statistical phenomena manufacture visibility success stories that no intervention produced.

“Every leap begins as a crackle.”

DOCUMENT	BA-MN-4
KIND	Method note
DATE	2026
SUBJECTS	Measurement & statistics
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

Two statistical phenomena manufacture visibility success stories that no intervention produced. The first is base rates: for most entity–need combinations the true probability of appearing in an AI answer sits near zero, so a portfolio-wide scan across many monitored combinations will surface some appearances by chance alone — it samples the lucky tail rather than measuring an effect. The second is regression to the mean: engagements tend to begin just after an unusually poor measurement, and because repeated AI-search measurements are noisy, the next measurement improves on its own, with no intervention. This note works both mechanisms with explicit arithmetic — including a hypothetical portfolio in which chance alone yields roughly four spurious appearances — and states what defeats them: pre-registered cells, interval estimates rather than point comparisons, and the sustainment condition of the Barkhausen Criterion.

CONTENTS

Two statistical mechanisms produce visibility success stories with no underlying effect behind them. Both are selection effects, and both are ordinary enough that a portfolio of AI-visibility measurements will generate apparent wins on its own, before any optimization is attempted. The first arises from base rates across many monitored query-cells; the second from where a measurement series happens to start. This note works each with explicit arithmetic and states what a study must do to keep them out of its conclusions. The numbers in the worked example below are a hypothetical illustration, not a measurement.

1 Base rates: sampling the tail of a portfolio

For most entity–need pairs, the true probability that an entity is mentioned in an AI answer to that need is near zero: a given brand is simply not relevant to the great majority of the questions a category generates. Call that true visibility, for a typical cell, some low value close to zero. A monitoring program that scans many such cells at once is running many low-probability trials in parallel, and across enough of them chance alone will turn up some appearances. A report that surfaces those appearances — "we showed up for query X" — has selected the lucky tail of a portfolio of near-zero cells, not detected an effect. The mechanism is multiplicity: the more cells monitored, the more chance appearances, which is why BA-C-3 requires the number of monitored cells to be disclosed alongside any claim that particular cells moved.

2 A worked example, hypothetical

Suppose a program monitors 200 query-cells, each drawn once, and suppose the true visibility in every one of them is a low 2% — so that by construction no cell carries any real signal above that floor. The number of cells that nonetheless show an appearance on a single draw is a binomial count with mean $np = 200 \times 0.02 = 4$ and standard deviation $\sqrt{np(1-p)} = \sqrt{200 \times 0.02 \times 0.98} \approx 2$. The probability that at least one cell appears is $1 - 0.98^{200} \approx 0.98$. So a scan of this hypothetical portfolio returns, on average, four "appearances," and almost never zero, purely from base rates. A success file assembled by keeping those four cells and discarding the 196 that stayed silent would look like a result while measuring nothing but the size of the portfolio. Doubling the number of monitored cells doubles the expected count of spurious appearances; it does not raise any single cell's true visibility. This is why the count of cells scanned, not just the cells that lit up, is the denominator a reader needs — the same denominator argument that BA-MN-3 makes for selected case studies.

3 Regression to the mean: where the series starts

The second mechanism needs no portfolio; it operates on a single entity. Optimization engagements tend to begin right after an unusually poor measurement — a bad reading is what prompts the effort. But an AI-visibility measurement contains a large transient component: identical prompts issued to the same engine on consecutive days shared only 45–59% of the brands they mentioned — measured across four engines, on the verticals where brands were reliably detected, over a roughly six-week window in a 2026 study [1]. When part of a reading is transient, an extreme low reading is extreme partly by luck, and the next reading tends to fall back toward the entity's typical level whether or not anything was done in between. A before-and-after comparison whose "before" is the measurement that triggered the engagement is therefore biased upward by regression alone, and the recovery is easy to miscredit to whatever was introduced in the interval. This is the temporal sibling of the selection effect described in BA-MN-3: survivorship selects which entities get written up, while regression selects when the baseline was taken. Both hand the analyst a favorable starting point for free.

4 What defeats both

The common cure is to remove the after-the-fact choice that each mechanism exploits. Pre-registering the cells and the outcome measure before collection — naming the entities, queries, and window in advance, then reporting what happened to all of them — prevents keeping the lucky cells and dropping the silent ones, and prevents choosing a baseline for its lowness; BA-R-2026-02 is this series' worked example of a pre-registered design. Reporting each estimate as an interval rather than a point, following the arithmetic in BA-MN-1, exposes a single-draw "appearance" for what it is: an estimate with an interval so wide it stays consistent with near-zero visibility. And the sustainment condition of the Barkhausen Criterion (BA-C-3) requires a shifted level to persist across at least two consecutive windows before it counts as a change — a bar that a chance appearance and a regression rebound both fail, because neither persists. Significance against proper intervals, sustainment, engine-change exclusion, and multiplicity disclosure together are built to reject exactly the artifacts this note describes.

5 Limitations

The worked example is a deliberately simple illustration: it assumes every cell shares one true rate and that draws are independent, whereas real portfolios mix heterogeneous true rates and correlated cells, which changes the exact counts but not the direction of the effect. Neither mechanism shows that an intervention does nothing — a real effect can coexist with both — only that a success file selected after the fact cannot establish that an intervention does something, because base rates and regression would produce the same file with no effect present. The drift figure cited here describes the specific engines and window of one study [1] and should be read as evidence that the transient component is large, not as a universal constant.

REFERENCES

1. Schulte, Bleeker, and Kaufmann. *Don't Measure Once: Measuring Visibility in AI Search (GEO)* (2026). https://arxiv.org/abs/2604.07585 Accessed 2026-07-10. [archived]

HOW TO CITE

Barkhausen AI (2026). Base rates and regression to the mean in visibility claims. https://barkhausen.ai/notes/base-rates-and-regression-to-the-mean/