

A Significance Level Is a Promise per Test, Not per Study

— Twenty Tries Make Something Significant with Probability **0.641514** — Tightening Restores the **0.05** and
Costs You the Findings — [Paper 317]

Yuuki YAMAGISHI (山岸 勇輝)

Yamagishi Complexity Torus Lab, Japan

August 2026

— *From thinking to syncing.* —

Abstract

“Significance at 5%” promises to hold false positives to 5% for one test. This paper asks whether it is still 5% after several tries—the answer is no. No new mathematical theorem and no new law is claimed. First, it breaks with repetition. 5 tries give **0.226219**, 10 give **0.401263** (Section 2). Second, this is the core of the paper. Twenty tries make something significant with probability **0.641514** (Section 2). Third, a hundred tries give **0.994079**. Something will almost certainly turn up (Section 2). Fourth, tightening restores it. At 0.0025 per test the overall false-positive rate is **0.048830** (Section 3). Fifth, what is tightened is also lost. False positives and false negatives cannot both be lowered (Section 4). Sixth, the separator is how many tries were made, and that is nowhere in the data (Section 5).

Scope of this paper (scope note): No new mathematical theorem and no new law is claimed—the definition of a significance level, the family-wise error rate, the Bonferroni correction and the second kind of error are all standard. We do not build statistics—all we use is one power and one division. We do not argue about testing itself—whether significance testing is sound is not treated. We only ask how many tests the promise covers. We do not compare correction methods—the performance of Holm, Benjamini–Hochberg and the rest is not entered. We compute no power—a number for power needs an effect size and a sample size. Only the direction is stated. We do not discuss the replication crisis—the institutional side of publishing is not treated. Relation to earlier papers: Paper 223 showed that “random” means both passing the tests and not being compressible, and exhibited a structured sequence that passes every test—this paper runs the other way. 223 is the side where testing misses real structure; this is the side where testing finds structure that is not there. Paper 164 showed that probability has four addresses—the 0.641514 here lives inside one of them, the frequency address. Paper 183 showed one table with two correct answers—this paper shows one dataset reaching different conclusions depending on a number that was never reported. Paper 255 showed that accuracy is two things and calibration removes only one—here too one operation moves only one of the two errors. Paper 250 showed two instruments giving opposite signs—this paper shows the number of instruments itself changing the answer. What is added is computing the family-wise error rate against the number of tests, giving 0.641514 at twenty and 0.994079 at a hundred, confirming the corrected overall rate as 0.048830, and putting the separator at a count that the data does not carry.

1. Notation and the Question

Setting: Take a test at significance level $\alpha = 0.05$. When the null hypothesis is in fact true, the chance of wrongly rejecting it is 0.05—that is the promise.

Observation 1.1

The promise is written per test. Across k independent tries, the chance of never erring is $(1 - \alpha)^k$ —so the chance of erring at least once is $1 - (1 - \alpha)^k$.

Question: Is it still 5% after several tries?

Observation 1.2

The answer is no—repetition alone breaks the promise.

2. The Core: 0.641514 at Twenty

Check 2.1: $\alpha = 0.05$, k independent tries.

Number of tests k	At least one false positive, $1 - (0.95)^k$
1	0.050000
5	0.226219
10	0.401263
20	0.641514
50	0.923055
100	0.994079

Observation 2.2

This is the core of the paper.**0.641514** at twenty——two times in three, “significance” appears where nothing is.

Observation 2.3

At a hundred it is 0.994079.Something will almost certainly turn up——so its turning up is no evidence at all.

Observation 2.4

The rise is fastest at the start. $1 \rightarrow 5$ multiplies it by 4.5, $5 \rightarrow 20$ by 2.8, $20 \rightarrow 100$ by 1.5——most of the damage is done in the first few tries.

Observation 2.5

A point to write honestly——independence is assumed. Real tests are often correlated, and correlation makes the family-wise rate smaller than this——the table is an upper bound.

3. Tightening Restores It

Setting: The Bonferroni correction lowers the per-test level to α/k .

Check 3.1:

k	Per test $\alpha' = \alpha/k$	Overall $1 - (1 - \alpha')^k$
20	0.002500	0.048830
100	0.000500	0.048782

Observation 3.2

Both come back below 0.05. And the larger k is, the slightly more conservative it becomes——0.048830 against 0.048782.

Observation 3.3

What was paid to keep the promise is the level. 0.05 became 0.0025——tightened by a factor of twenty.

4. What Is Tightened Is Also Lost

Observation 4.1

Tightening the level also makes real effects harder to detect. Lower the false positives (the first kind of error) and the false negatives (the second kind) rise——with the sample unchanged, both cannot fall.

Check 4.2: What each operation moves.

Operation	False positives	False negatives	Cost
Tighten the level	fall	rise	findings are missed
Enlarge the sample	—	fall	time and money
Run fewer tests	fall	—	fewer questions asked

Observation 4.3

Only the second row lowers one without raising the other—and only the second row costs money.

Observation 4.4

Paper 255 wrote that accuracy is two things and calibration removes only one. The same shape appears here—one operation moves only one of them.

5. Separator: The Number of Tries Is Not in the Data

Observation 5.1

Separator—how many tries were made. And that number appears nowhere in the figures that come out.

Check 5.2: One value, $p = 0.03$, changing meaning with context.

Situation	Corrected threshold	$p = 0.03$ is
One test was run	0.05	significant
Twenty tests were run	0.0025	not significant
A hundred tests were run	0.0005	not significant

Observation 5.3

One p -value with three different standings. Same data, same arithmetic—all that differs is a count that was not reported.

Observation 5.4

Paper 183 wrote of one table with two correct answers. This is one step harder—in 183 both answers were inside the table, while here the branch point lies outside the data.

Observation 5.5

Setting the direction beside Paper 223.

	The test	and the content is
Paper 223	lets it through	structured, and it is missed
This paper	rejects	nothing at all, and it is picked up

Observation 5.6

Two ways for the same tool to fail. Paper 300's criteria put two distinct roots under the one word “test”—one is a shortfall of power, the other an unreported count, and the repairs differ.

Observation 5.7

One last thing—0.641514 is not a number produced by misconduct. Everyone testing honestly at $\alpha = 0.05$ by the rules still arrives here—the unit of the promise has simply been mistaken.

6. Roll Call and Closing

1. 0.226219 at five (Check 2.1): It breaks at once.

2. **0.641514** at twenty (Check 2.1): **Two times in three.**
3. **0.994079** at a hundred (Check 2.1): **Almost certain.**
4. **Most damage in the first few** (Observation 2.4): **1 → 5 gives 4.5.**
5. **Independence is assumed** (Observation 2.5): **The table is an upper bound.**
6. **0.048830** after correction (Check 3.1): **Back under 0.05.**
7. **The level is cut twentyfold** (Observation 3.3): **0.05 → 0.0025.**
8. **Both cannot fall** (Check 4.2): **Only a bigger sample is free of that.**
9. **One p , three standings** (Check 5.2): **The count is unreported.**
10. **Opposite to 223** (Observation 5.5): **Missing, and over-picking.**

Closing: “Significance at 5%” is a promise about one test, not about one study. Across k independent tries the chance of erring at least once is $1 - (0.95)^k$ —**0.226219** at five, **0.401263** at ten, **0.641514** at twenty, **0.994079** at a hundred. Two times in three, “significance” appears where nothing is—and at a hundred it is near certain, so its appearance is no evidence at all. Tightening restores it—at 0.0025 per test (a twentieth) the overall rate is **0.048830**, back under 0.05. But what is tightened is also lost—false positives and false negatives cannot both fall with the sample unchanged. Only enlarging the sample avoids the trade, and only that costs money. One thing separates them—how many tries were made. And that number appears nowhere in the figures. The same $p = 0.03$ is significant after one test and not after twenty—same data, same arithmetic, differing only by an unreported count. Paper 223 runs the other way—there the test lets real structure through; here it picks up structure that is not there. Two ways for one tool to fail, and the repairs differ. One last thing—**0.641514** is not a number produced by misconduct. Everyone testing honestly by the rules still arrives here; the unit of the promise has simply been mistaken.

References

1. Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8.
2. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2).
3. Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate. *Journal of the Royal Statistical Society B*, 57(1).
4. Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8).
5. Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd edition. New York: Springer.
6. Yamagishi, Y. (2026). *Probability Has Four Addresses* [Paper 164]. Zenodo. 10.5281/zenodo.22092815
7. Yamagishi, Y. (2026). *One Table, Two Correct Answers* [Paper 183]. Zenodo. 10.5281/zenodo.22104230

8. Yamagishi, Y. (2026). *It Passes Every Test, and the Points Lie on Fifteen Planes* [Paper 223]. Zenodo. 10.5281/zenodo.22120728
9. Yamagishi, Y. (2026). *Two Instruments, Opposite Signs* [Paper 250]. Zenodo. 10.5281/zenodo.22160733
10. Yamagishi, Y. (2026). *“Accuracy” Is Two Things, and Calibration Removes Only One* [Paper 255]. Zenodo. 10.5281/zenodo.22168883
11. Yamagishi, Y. (2026). *The Three Hundredth Census: Whether Two Things Share a Root Is Decidable* [Paper 300]. Zenodo. 10.5281/zenodo.22171914

— *From thinking to syncing.* —

Yuuki YAMAGISHI (山岸 勇輝)

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