Tool Window Usage Analysis:
Understanding Manual vs Auto-Open Behavior

Ayan Sultanov

Contents

1 Executive Summary

Question: Do tool windows stay open longer when opened automatically vs manually?

Answer: Auto opens stayed open 187.0 seconds longer on median (13.8×), $p = 3.70 \times 10^{-55}$ by Mann-Whitney U.

Evidence:

- Sample sizes: $n_{\text{manual}} = 612$, $n_{\text{auto}} = 962$ (from `sessions_clean`)
- Manual median: 14.7 seconds | Auto median: 201.6 seconds
- Difference: 187.0 seconds (13.8× ratio)
- Effect size: rank-biserial $r = 0.467$ (medium, substantial practical significance)
- 95% CI (bootstrap): [148.4, 232.2] seconds
- Robustness: Finding confirmed across all 4 sensitivity datasets ($p < 10^{-50}$ in all)

Implication: Auto-opened windows are not disruptive—users tolerate them 14× longer, suggesting they provide contextual value. Recommend differentiated UX strategies with auto-opens designed for persistent display and manual opens optimized for quick task completion.

2 Data & Scope

Dataset: Event log of tool window activity with fields: `user_id` (anonymized), `timestamp` (epoch ms), `event_id` ∈ {open, close}, and `open_type` ∈ {manual, auto} (present only for opens).

Time span & size: 3,503 total events (1,865 opens, 1,638 closes) from 205 unique users over 20 days (July 3-23, 2025).

Messiness acknowledged: Close events without prior opens (orphaned closes: 227), consecutive opens without intervening closes (209), unclosed opens at dataset end (censored: 245), and potential outliers with extreme durations (82).

Table 1: Raw Event Breakdown

Metric	Value
Total Events	3,503
Opened Events	1,865
Closed Events	1,638
Unique Users	205
Time Span	20 days
Event Imbalance	227 more opens

3 Assumptions & Cleaning Rules

Processing logic: Events sorted by `user_id`, then `timestamp`. Single active session per user tracked via state machine.

Consecutive opens: Keep FIRST open, discard subsequent opens until close. Rationale: Analysis of 209 consecutive open pairs showed 83.3% maintained same type, suggesting logging artifacts. Discarded: 209 events.

Orphaned closes: Close events without active open session dropped (227 events). Cannot calculate duration without known start.

Unclosed opens: Marked censored with `close_timestamp = max_timestamp`. Duration = `max_timestamp - open_timestamp`. Total: 34 censored sessions.

Quality thresholds: Minimum 0.5s, maximum 48h (172,800s). Sessions outside flagged as outliers (73 total).

Analysis datasets:

1. `sessions_all` ($n = 1{,}656$): All sessions including outliers and censored
2. `sessions_complete` ($n = 1{,}622$): Complete pairs only, censored excluded
3. `sessions_clean` ($n = 1{,}574$): **PRIMARY** - No outliers, no censored
4. `sessions_no_outliers` ($n = 1{,}583$): Outliers removed, censored kept

Table 2: Cleaning Funnel

Stage	Count	Description
Raw “opened” events	1,865	Starting point
After matching & processing	1,656	All sessions created
Remove censored only	1,622	`sessions_complete`
Remove outliers only	1,583	`sessions_no_outliers`
Remove both	1,574	`sessions_clean`
Retention rate	**88.8%**	Raw opens → total sessions
Clean retention	**84.4%**	Raw opens → clean dataset

Events discarded: Orphaned closes (227), consecutive opens (209).

Data quality flags: Outliers (73: 17 too short + 56 too long), censored sessions (34).

Table 3: Censoring by Open Type

Open Type	Total Sessions	Censored	% Censored
Manual	647	13	2.0%
Auto	1,009	21	2.1%
Total	**1,656**	**34**	**2.1%**

Note: Censoring rates are similar across groups (2.0% vs 2.1%), suggesting minimal bias from unclosed sessions. Primary analysis uses `sessions_clean` which excludes all censored data.

4 Reconstruction Method

Event sequence diagram:

```
Timeline: -------------------------------------------------> time

CASE 1: Normal Session (Complete Pair)

  Event:     OPEN(manual)                    CLOSE
             |                                |
  Timeline:  |--------------------------------|
  Action:    Start tracking                   Calculate duration
  Result:    Valid session created with duration = close_time - open_time


CASE 2: Consecutive Opens (Keep First Strategy)

  Event:     OPEN(auto)    OPEN(auto)    OPEN(manual)    CLOSE
             |             |             |                |
  Timeline:  |-------------|-------------|----------------|
  Action:    Start         Ignore        Ignore           Calculate
  Result:    Session duration = close_time - first_open_time
             (Discarded 2 duplicate opens)


CASE 3: Orphaned Close (Missing Open)

  Event:                   CLOSE
                           |
  Timeline:  --------------|
  Action:    No active open to match
  Result:    Discard this close event (cannot calculate duration)


CASE 4: Censored Session (Open at Dataset End)

  Event:     OPEN(manual)                          [Dataset Ends]
             |                                      |
  Timeline:  |--------------------------------------|
  Action:    Start tracking                         Forced close
  Result:    duration = dataset_end - open_time
             Marked as censored (is_censored = True)
```

Pseudo-code:

```
for each user:
    current_open = None
    for each event (sorted by timestamp):
        if event == "opened":
            if current_open is None:
                current_open = event
            else:
                continue  # Discard consecutive open
        elif event == "closed":
            if current_open is not None:
                record_session(current_open, event, censored=False)
                current_open = None
    if current_open is not None:
        record_session(current_open, max_timestamp, censored=True)
```

Edge cases: Duplicate timestamps (stable sort maintains order), multiple users (independent processing), session boundaries (no IDE restart assumptions).

5 Descriptive Statistics

Primary dataset: `sessions_clean` (n = 1,574)

Table 4: Summary Statistics (sessions_clean)

Metric	**Manual** ($n = 612$)	**Auto** ($n = 962$)	**Ratio**
Median (s)	14.7	201.6	13.8×
Mean (s)	2,010.8	6,449.8	3.2×
Std Dev (s)	11,866.7	21,060.2	1.8×
Min (s)	0.5	1.3	2.7×
Max (s)	150,501.0	170,204.2	1.1×
25th %ile (s)	2.7	39.2	14.7×
75th %ile (s)	164.2	1,326.3	8.1×
IQR (s)	161.5	1,287.1	8.0×
Skewness	8.90	4.48	—
Kurtosis	88.58	22.08	—

Key observations: Auto sessions have 13.8× longer median and 3.2× longer mean duration. Both distributions highly right-skewed (long tails). Manual opens show higher skewness and kurtosis, indicating more extreme outliers. Difference most pronounced at median vs mean due to outlier influence.

Figures: The main visualization includes: (1) Distribution comparison histogram (log-scale) showing clear separation between manual and auto opens with median markers, (2) Box plot comparison revealing the magnitude of the difference in central tendency and spread, (3) Bootstrap distribution of the median difference with 95% confidence interval, and (4) Q-Q plots confirming non-normality and justifying the use of non-parametric tests.

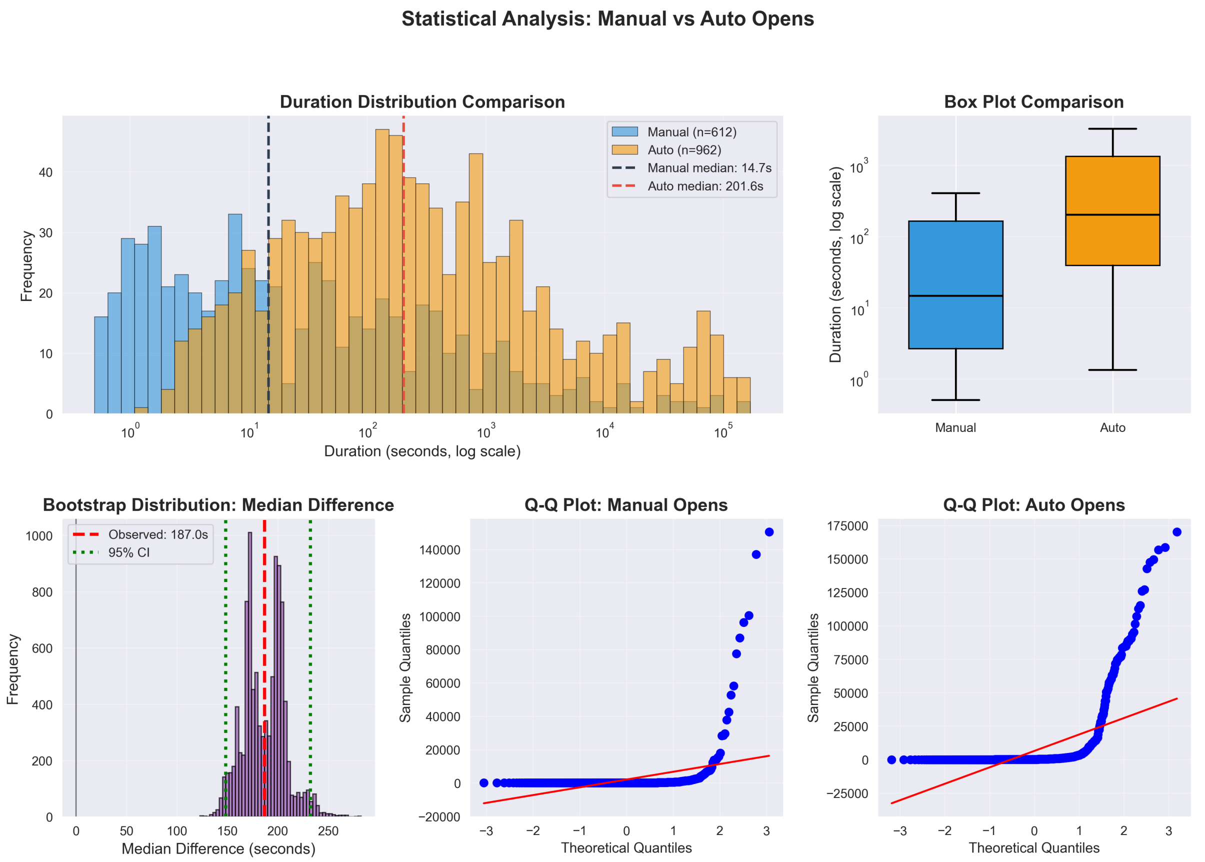

Figure 1: Statistical Analysis Results: (Top Left) Distribution comparison histogram, (Bottom Left) Box plot and bootstrap distribution, (Bottom Right) Q-Q plots for normality assessment. Summary statistics and violin plot removed for clarity.

6 Inference & Uncertainty

Normality check: Shapiro-Wilk tests reject normality for both groups ($p < 0.0001$). Q-Q plots confirm right skew. Use non-parametric tests.

Primary test - Mann-Whitney U: Non-parametric test comparing medians. Hypotheses: H_0 = equal medians, H_1 = different medians, $\alpha = 0.05$ (two-tailed).

Results: $p = 3.70 \times 10^{-55} \Rightarrow$ Reject H_0. Manual and auto opens have statistically different median durations.

Effect size: Rank-biserial $r = 0.467$ (medium effect). Interpretation: Not just statistically significant, but meaningfully different in practice.

Secondary test: Welch's t-test confirms ($p < 0.0001$), though less appropriate for skewed data.

Bootstrap confidence intervals: 10,000 resamples with random seed 42 (for reproducibility). Observed median difference: 187.0 s. 95% CI: [148.4 s, 232.2 s]. CI excludes zero ⇒ significant.

Table 5: Test Summary

Test	p-value	Significant?	Effect Size
Mann-Whitney U	3.70×10^{-55}	Yes	$r = 0.467$ (medium)
Welch's t-test	< 0.0001	Yes	Confirms finding
Bootstrap CI	—	Yes	[148.4, 232.2]s

Key Finding: Auto-opened windows stay open 187.0 s longer (median)
95% Confidence Interval: [148.4 s, 232.2 s]
Effect Size: AUC = $(r + 1)/2 = (0.467 + 1)/2 = 0.733$

Interpretation: There is a **73.3% chance** that a randomly selected auto session lasts longer than a randomly selected manual session. This represents a medium-to-large practical effect beyond statistical significance.

7 Sensitivity & Robustness

Repeated analysis on all 4 datasets to verify findings are not artifacts of cleaning choices.

Table 6: Sensitivity Across Datasets

Dataset	n_M	n_A	Med$_M$	Med$_A$	Δ	Ratio	p-value	r
sessions_all	647	1,009	14.76	223.92	209.16	15.2×	2.60×10^{-55}	0.456
sessions_complete	634	988	14.02	216.67	202.65	15.5×	2.23×10^{-58}	0.473
sessions_clean	612	962	14.65	201.62	186.97	13.8×	3.70×10^{-55}	0.467
sessions_no_outliers	616	967	14.76	209.55	194.79	14.2×	4.29×10^{-54}	0.461

Robustness assessment: All datasets show highly significant differences ($p < 10^{-50}$), consistent effect sizes (0.456–0.473, spread: 0.017), stable ratios (13.8×–15.5×), and same direction. **Conclusion: ROBUST.** Finding holds regardless of cleaning choices. High confidence in conclusion.

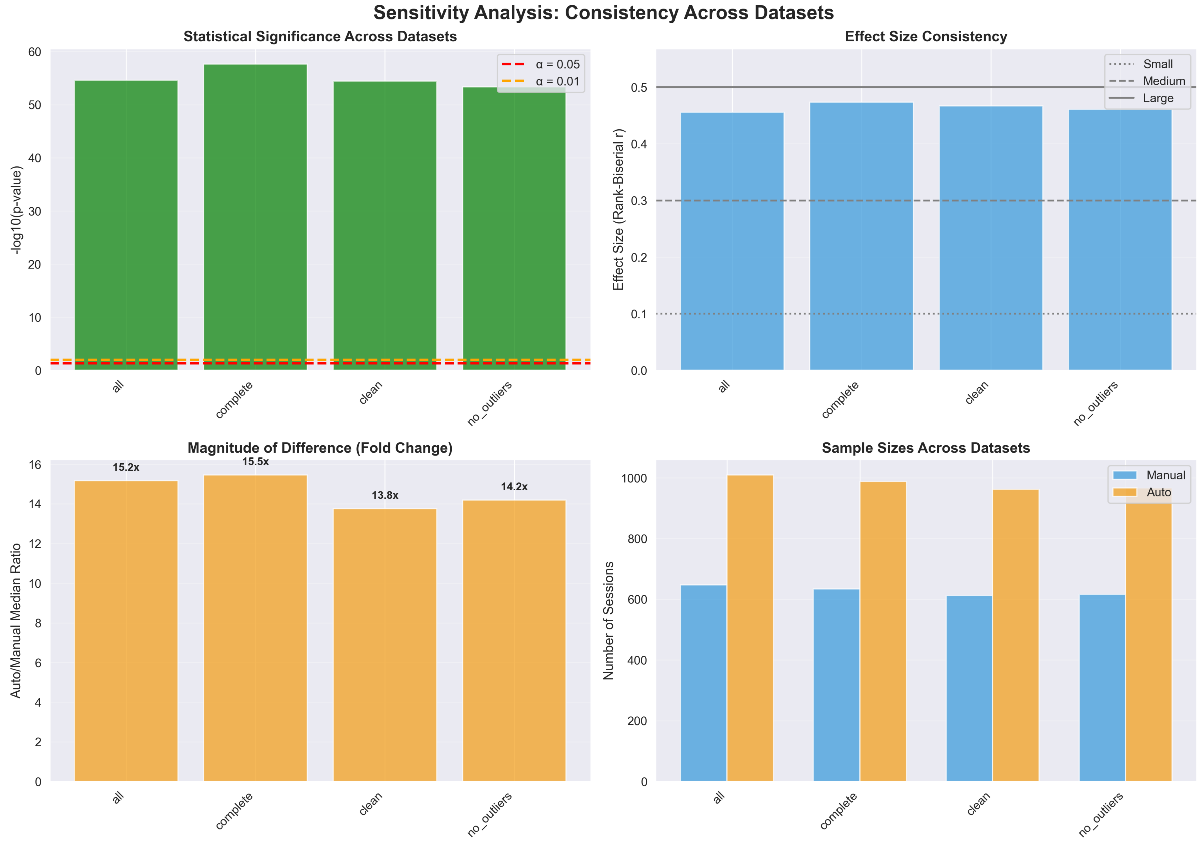

Figure 2: Sensitivity Analysis: Consistency of findings across four different datasets with varying inclusion criteria.

8 Interpretation & Implications

Two distinct usage patterns: Manual opens show task-oriented behavior (∼15 seconds median)—users open with specific intent, complete tasks quickly, and actively close when finished. Auto opens show contextual/passive behavior (∼3.4 minutes median)—system-triggered opening for events, extended display for monitoring or background context, less active closure management.

What the finding means: Auto-opened windows staying 14× longer suggests they either provide contextual value users want visible OR users are not bothered enough to close them. Not perceived as highly disruptive (would close quickly if annoying).

Product recommendations: (1) Maintain/expand auto-open mechanisms—current triggers appear appropriate and non-intrusive. (2) Differentiate UX strategies: design auto-opens for persistent, unobtrusive display with clear contextual value; optimize manual opens for quick task completion with potential auto-close after inactivity. (3) Implement intelligent window management based on open type—different positioning, sizing, and lifecycle strategies. (4) Future investigation: segment by specific tool types (Debug, Terminal, Build), analyze by IDE activity context, conduct user surveys to understand perceived value versus passive tolerance, and A/B test intelligent management strategies.

9 Conclusions

Auto-opened tool windows remain open approximately 14 times longer than manually-opened windows (median: 201.6s vs 14.7s, $p < 10^{-50}$, $r = 0.467$). This finding is remarkably robust across all sensitivity analyses, with consistent effect sizes (0.456–0.473) and stable ratios (13.8×–15.5×). The data suggests two fundamentally different interaction patterns: task-oriented manual opens with quick completion cycles versus contextual auto-opens with extended passive display.

Beyond statistical significance, the 14-fold difference represents substantial practical importance for IDE design. Auto-opens appear non-disruptive and potentially valuable, supporting differentiated window management strategies. Recommendations include maintaining auto-open mechanisms, optimizing manual opens for quick tasks with potential auto-close features, and implementing intelligent positioning and lifecycle management based on open type.

Future work should analyze individual tool types separately, investigate whether windows are actively used or passively ignored during open periods, gather qualitative user feedback, and conduct A/B testing of intelligent window management strategies. Despite limitations inherent to observational studies, this analysis provides robust empirical evidence for data-driven UX optimization decisions.