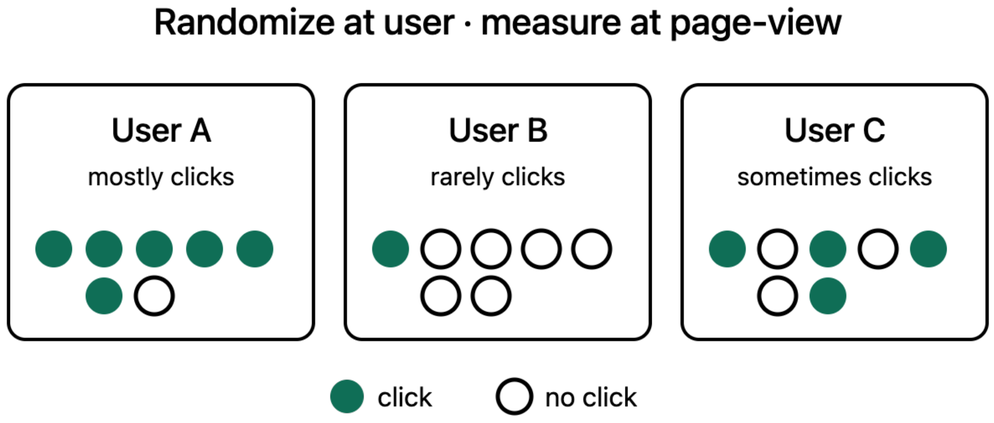
deltatest: Statistical Hypothesis Testing Using the Delta Method for Online A/B Testing

Koji Makiyama (HOXO-M Inc.), Kazuyuki Sano, Daisuke Ichikawa (Kibaroku), Shinichi Takayanagi

Motivation — when the standard Z-test goes anti-conservative

- Online A/B tests randomize at the user level but evaluate ratio metrics at a finer unit — page-views or sessions.
- This mismatch induces within-user correlation, violating the i.i.d. assumption behind the standard Z-test.
- The result: the test becomes anti-conservative — variance is under-estimated, p-values skew small, and false positives inflate above the nominal level.



Method — a delta-method variance correction

- The fix is the variance estimator, not a new test. Following Deng et al. (2018), the delta method yields a variance that accounts for within-user correlation.
- Drop it into the Z-test → a corrected standard error. No resampling, no mixed models.
- The deltatest packages this into a single, tested function — a transparent, package-native workflow instead of per-team copy-paste scripts.

Data & usage — concise formula interface, familiar output

```
data
#> # A tibble: 2,000 × 4
#>    user_id group     clicks pageviews
#>      <int> <chr>      <int>     <int>
#>  1       1 treatment      1         6
#>  2       2 treatment      2        11
#>  3       3 control        0        17
#>  4       4 control        4        12
#>  5       5 control        5        10
#>  6       6 control        1        15
#>  7       7 control        2         6
#>  8       8 treatment      2        11
#>  9       9 treatment      2        16
#> 10      10 control        0        17
#> # ℹ 1,990 more rows
```

Required data structure

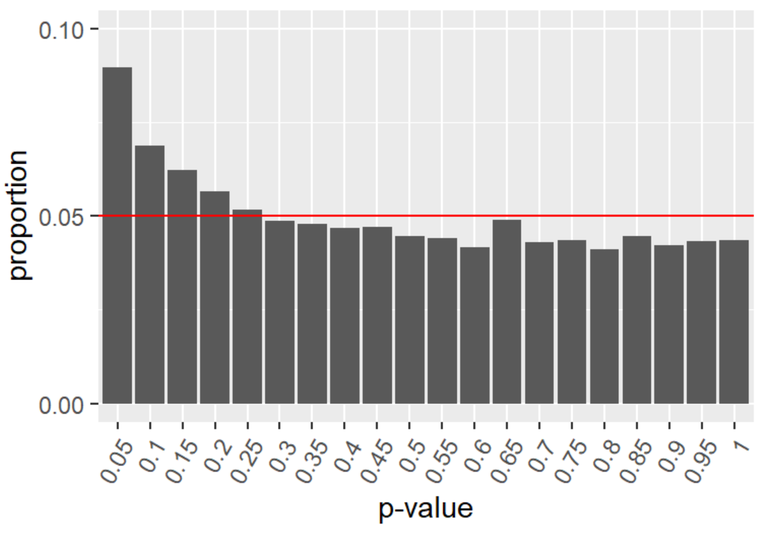
```
library(deltatest)

deltatest(data, clicks / pageviews, by = group)
#>
#>  Two Sample Z-test Using the Delta Method
#>
#> data:  clicks/pageviews by group
#> Z = 0.31437, p-value = 0.7532
#> alternative hypothesis: true difference in means betwee
#> 95 percent confidence interval:
#>  -0.01410593  0.01949536
#> sample estimates:
#>   mean in control mean in treatment        difference
#>       0.245959325       0.248654038       0.002694713
```

Simple usage

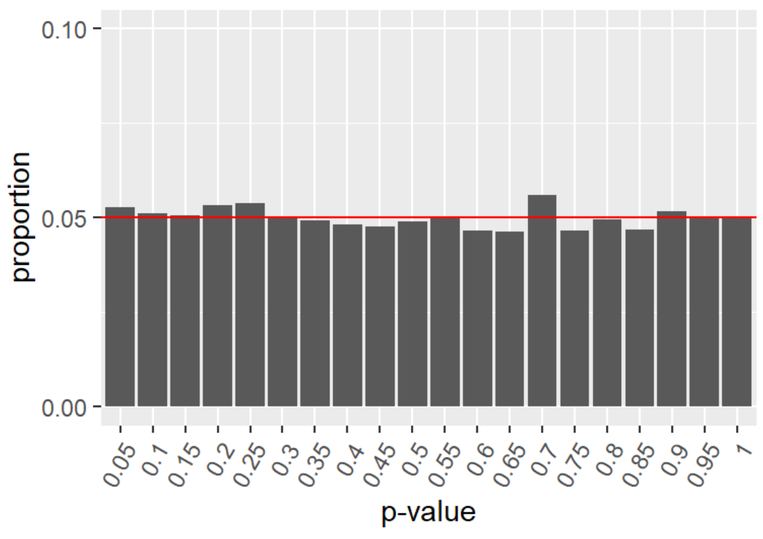
Why this interface

- One-line call, familiar htest output
- Three input styles: formula, lambda, NSE
- absolute or relative change
- broom::tidy() ready
  → drops into tidyverse pipelines

Simulation results — the proof under H_0



naive Z-test → p-values pile up near 0 (Type I error inflated)



deltatest → uniform under H_0 (error back at α)

How we tested it
Simulated null (no true effect), users generate correlated page-views.

Result
- Type I error: **0.09** → **0.05** (α = 0.05)

Correct false-positive control where the naive Z-test silently fails — same null, same data, only the variance estimator changes.

Installation

- The deltatest package is available on CRAN.

```
install.packages("deltatest")
```

- For more details: https://hoxo-m.github.io/deltatest/

• Deng, Knoblich & Lu (2018). Applying the Delta Method in Metric Analytics. KDD '18. doi:10.1145/3219819.3219919