Safe Testing: Online, Anytime-Valid Hypothesis Tests

Rosanne J Turner[1,2], Alexander Ly[1,3], Judith ter Schure[1] and Peter D Grünwald[1]

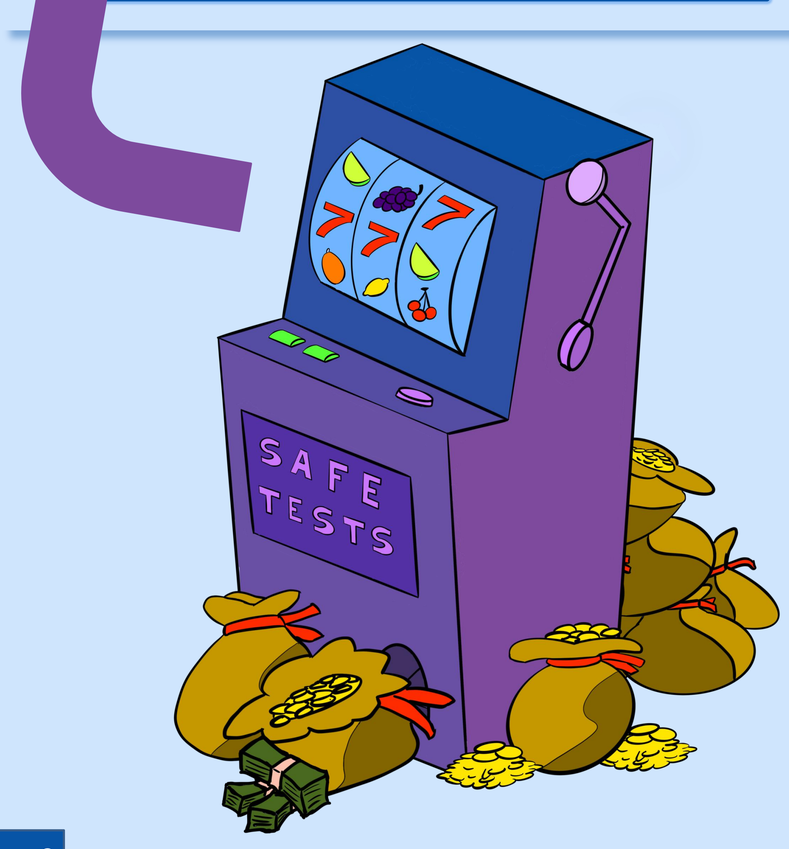
1. Machine Learning Group, Centrum Wiskunde & Infomatica
2. Brain Center, UMC Utrecht
3. Department of Psychological Methods, University of Amsterdam

Hypothesis testing: a recap

- Goal: decide between hypotheses H0 or H1
- No experiment is perfect: need **guarantees** on chance of making **wrong decision**
  - **Type I error:** falsely concluding H1 is true
  - **Type II error:** failing to detect that H1 is really true
- Classical hypothesis tests (p-values) only offer gurantees if **sample size fixed in advance**
- **Problem: p-values cannot be used for continuous, online analysis!**

What are safe tests?

- Hypothesis tests based on s-values, instead of p-values[1]
- Test rule: $s > \alpha^{-1}$? Reject H0 with type I error guarantee **α**
- Interpretation: a **gambling game**
  - **High *s***? Have **won** your "bet" on H1, **keep investing**!
  - **Low *s***? Little evidence, **lost** investment



1. Before study: easily design a safe test

1. Decide on the maximally acceptable **type I error**
   - **Expensive error**: unnecessary follow-up work
   - Corresponds to **significance level α**: often 0.05
2. Decide on the maximally acceptable **type II error rate**
3. Determine the **minimal relevant difference** between H0 and H1 one wants to detect
4. Perform a **power analysis** to determine the maximal study sample size **(Figure 1, blue line)**
   - The actual expected sample size is lower: can monitor results and stop early **(Figure 1, purple line)**

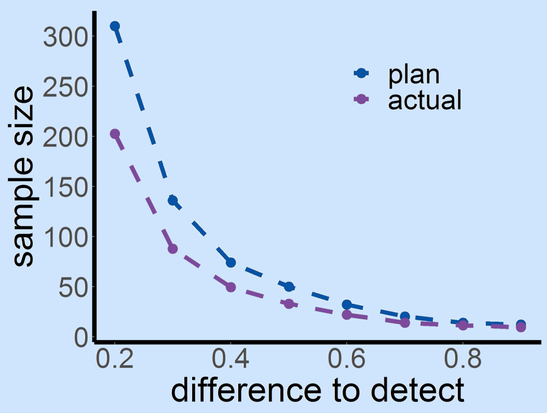


Figure 1. Power analysis for safe two proportions test: number of samples needed with maximal type I error 0.05 and type II error 0.20

2. During study: safe tests allow for monitoring of evidence and early stopping

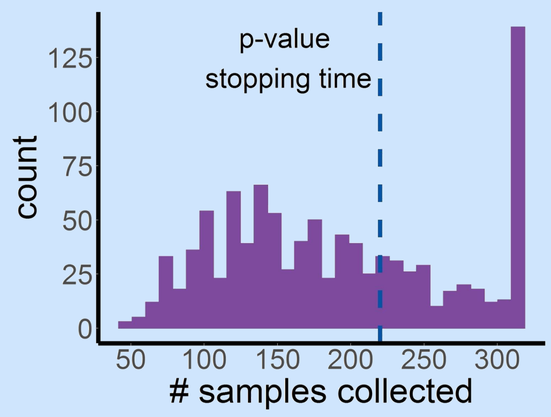


Figure 2. 1000 simulations of stopping times with a safe test for two proportions when the true difference is 0.2

- Safe tests for proportions and t-tests are **anytime-valid**
- Continuous monitoring of evidence is allowed, experiment can be stopped early if *s* exceeds threshold
- Profile of early stopped experiments is shown in **Figure 2**: in 65% of experiments **can stop earlier** than with p-value test
- Monitoring the p-value and early stopping anyway leads to an inflation of the type I error rate (**Figure 3**)
- Scientists used p-values this way because feasible alternatives were lacking: possible cause of the **reproducibility crisis** in science

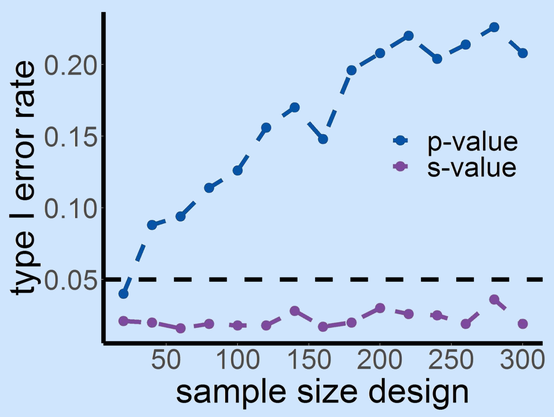


Figure 3. Simulation results illustrating the type I error rate when a p-value (Fisher's exact test) and a safe two proportions test are used in online setting

3. After study: safe tests allow for optionally continuing research

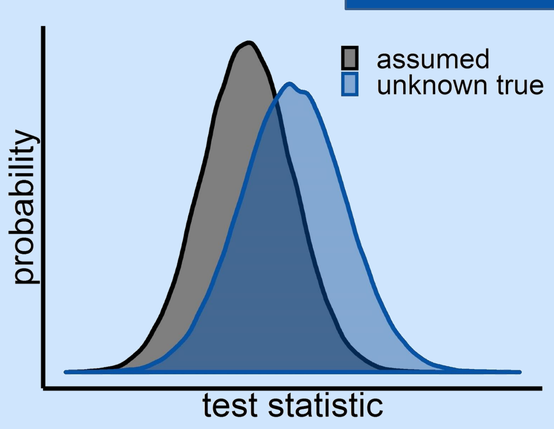


Figure 4. Problem with meta-analysis: distribution needed for calculating p-value is not known (accumulation bias[2])

- Used a **p-value** test? **Final destination**!
  - Combining multiple studies in meta-analysis leads to accumulation bias[2]: cannot calculate p-value (**Figure 4**)
  - Scientists do this anyway, **reproducibility crisis**
- Safe tests allow for **optionally continuing your study**
  - Based on s-value decide to start new study, for example in case of borderline significance
  - **Multiply new and old s-value for one super s-value** that still offers a type I error guarantee[1]
  - Allowed to use different data sources for second s-value

Want to apply anytime-valid hypothesis tests in your own work? Try our tutorial and `R` package `safestats`!

CWI
Centrum Wiskunde & Informatica

rosanne@cwi.nl

References
1) Grünwald, de Heide & Koolen, *Safe Testing*, 2019
2) Ter Schure & Grünwald, *Accumulation bias in meta-analysis: the need to consider time in error control*, 2019

Acknowledgements
Thanks to Bas de Haan for the illustration

This work is part of the project Enabling Personalized Interventions (EPI). The project is supported by NWO in the Commit2Data –Data2Person program under contract 628.011.028.
For more information see: enablingpersonalizedinterventions.nl
Project coordinator: Marloes Bons, bons.marloes@kpmg.nl, +31 6 23593759
Principle Investigators: Prof.dr.ir. C.T.A.M. de Laat, delaat@uva.nl,
Prof. dr. Sander Klous, Klous.Sander@kpmg.nl

Commit2Data