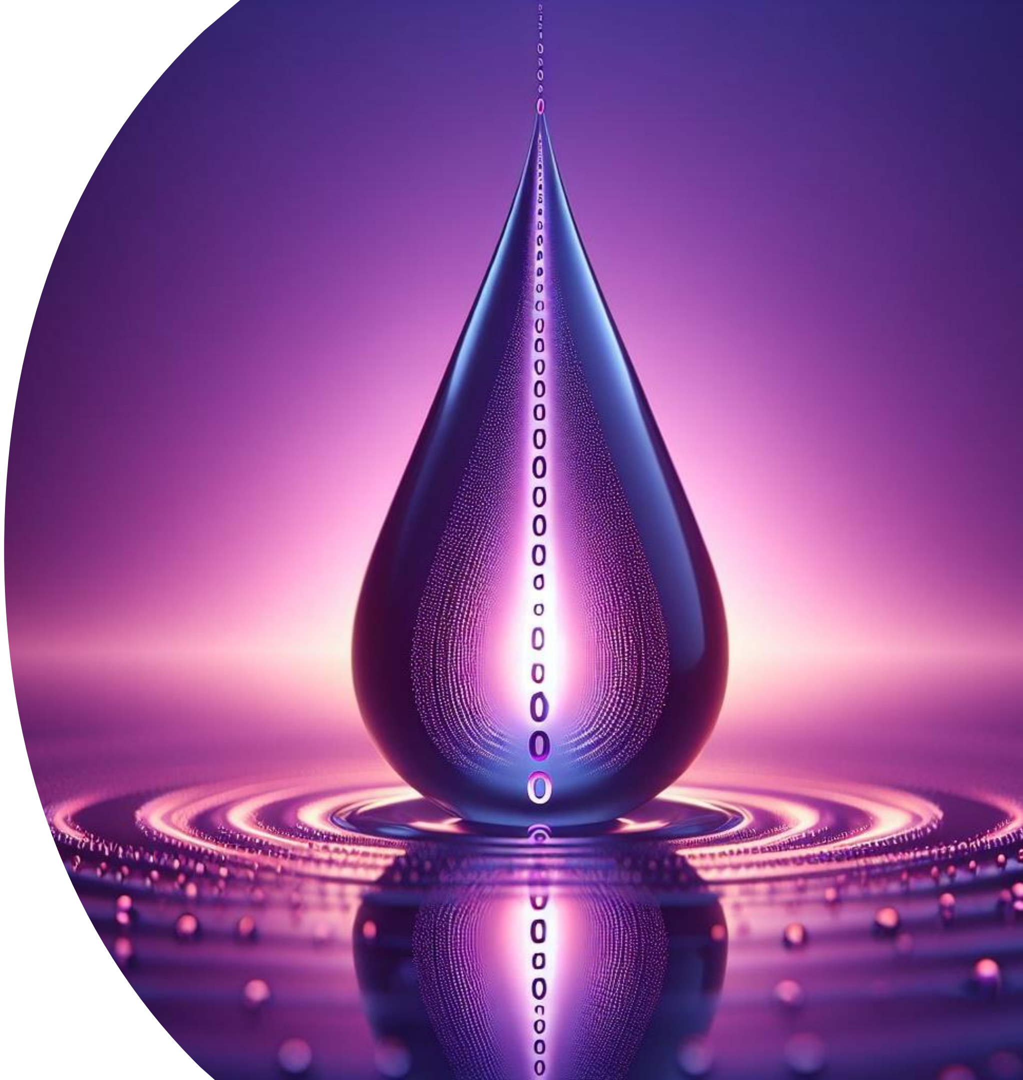
sanofi

Analyzing Zero-Inflated Continuous Data in Limited Sample Size

Andreas Schulz
Non Clinical Efficacy and Safety Biostatistics

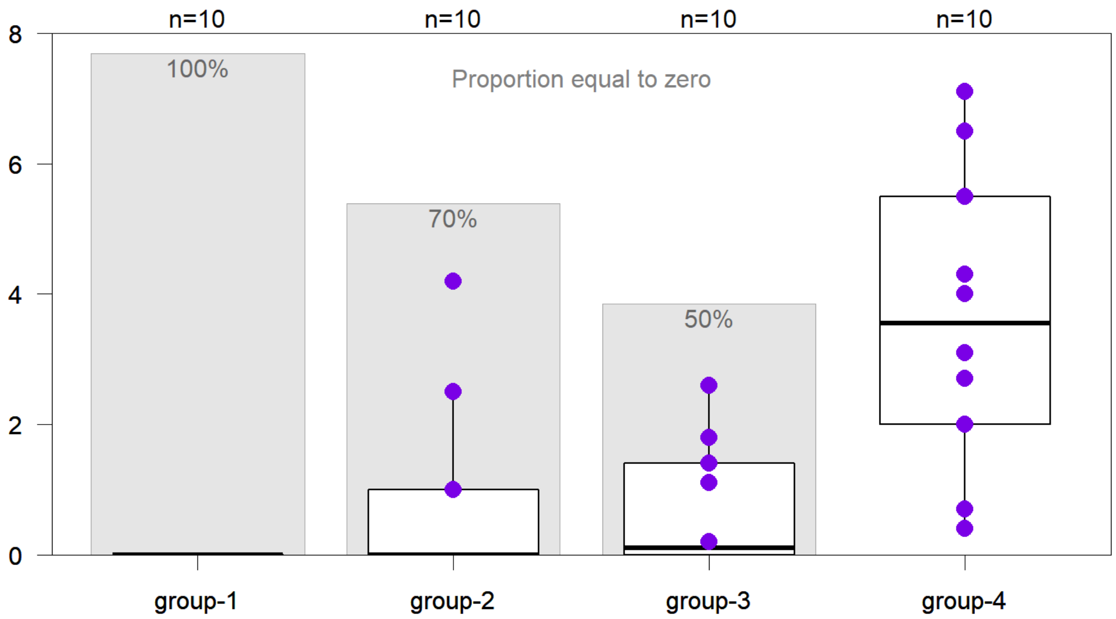
NCS | Non-Clinical Statistics Conference
Wiesbaden, DE / 25-27 September, 2024



Introduction

sanofi

Motivation example



sanofi

Summarizing the problem

Challenges:

- Point mass at zero (many ties)
- Skewed distribution (extreme values)
- Small sample size (partly non-informative)

May be also:

- Perfect separation
- Unbalanced sample size
- Variance heterogeneity

Narrowing the problem

To refine the scope of the investigation, certain assumptions were made.

Assumptions:

- Two sample problem
- Independent observations
- The zero values are real, no censoring or LOD
- Sample size <=20 per group (small)

Methods

Non-parametric test

Presented are top 4 out of ~20 investigated methods.

BM - Brunner-Munzel test

Also known as the generalized Wilcoxon test.

The Brunner-Munzel test is a non-parametric rank-based, very robust test that can handle ties and heterogeneity of variance.

Neubert, Karin & Brunner, Edgar, 2007. A studentized permutation test for the non-parametric Behrens-Fisher problem. Computational Statistics & Data Analysis, Elsevier, vol. 51(10), pages 5192-5204, June.

sanofi

Permutation tests

Simple permutation tests are known to be very robust, except in the case of variance heterogeneity, as they assume exchangeability.

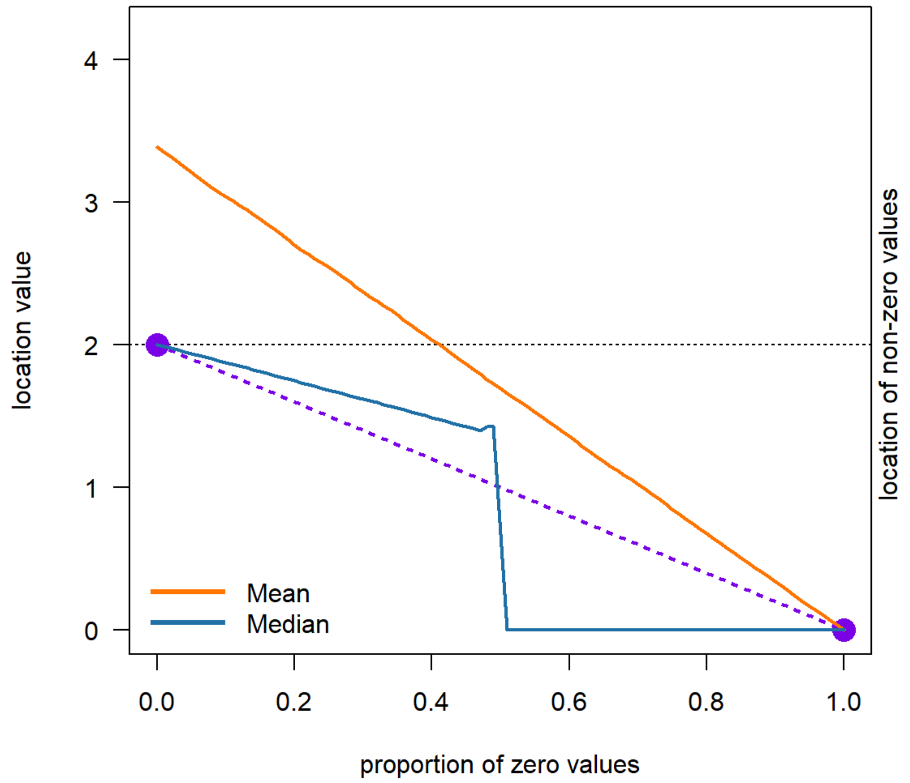
ZWM-p permutation test with zero weighted median

$$ZWM(x) = \frac{median(x_{pos.}) * \ N_{pos.}}{N}$$

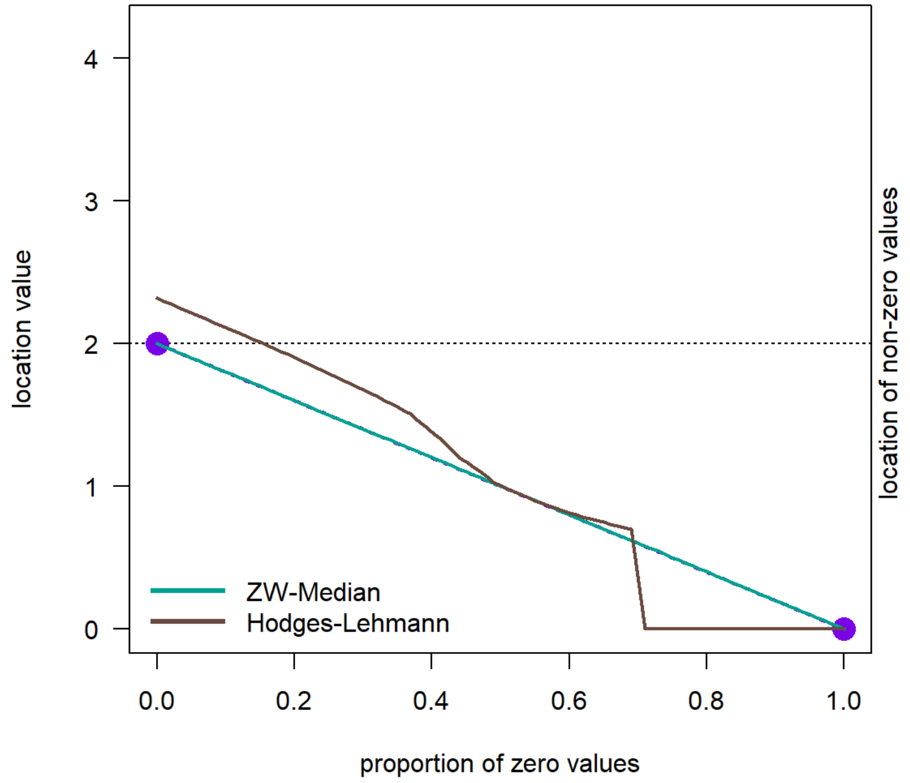
HL-p permutation test with Hodges–Lehmann estimator

Hodges, J.L., and Lehmann, E.L. (1963), Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, **34**, 598–611.

Location parameter



Location parameter



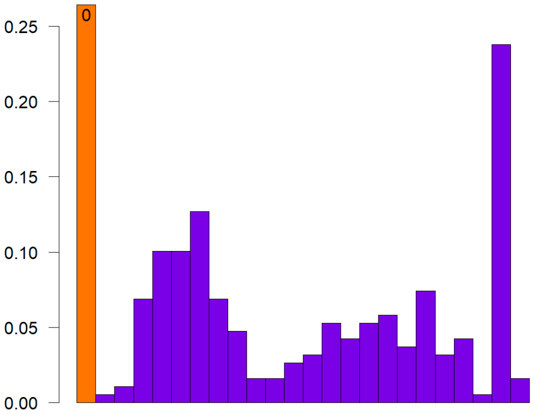
sanofi

Multiple Transformation

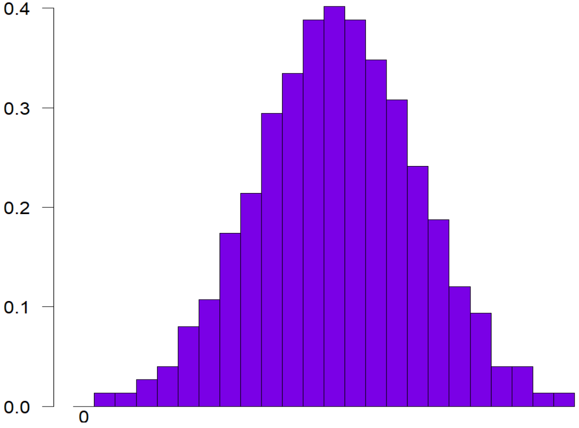
MINT - Multiple Inverse Normal Transformation + Welch's t-test (100 times)

$$INT_r(x) = \Phi^{-1}\left\{\frac{rank(x, ties = random) - \frac{1}{2}}{n}\right\}$$

p-values of 100 tests are combined with simple median rule.



→

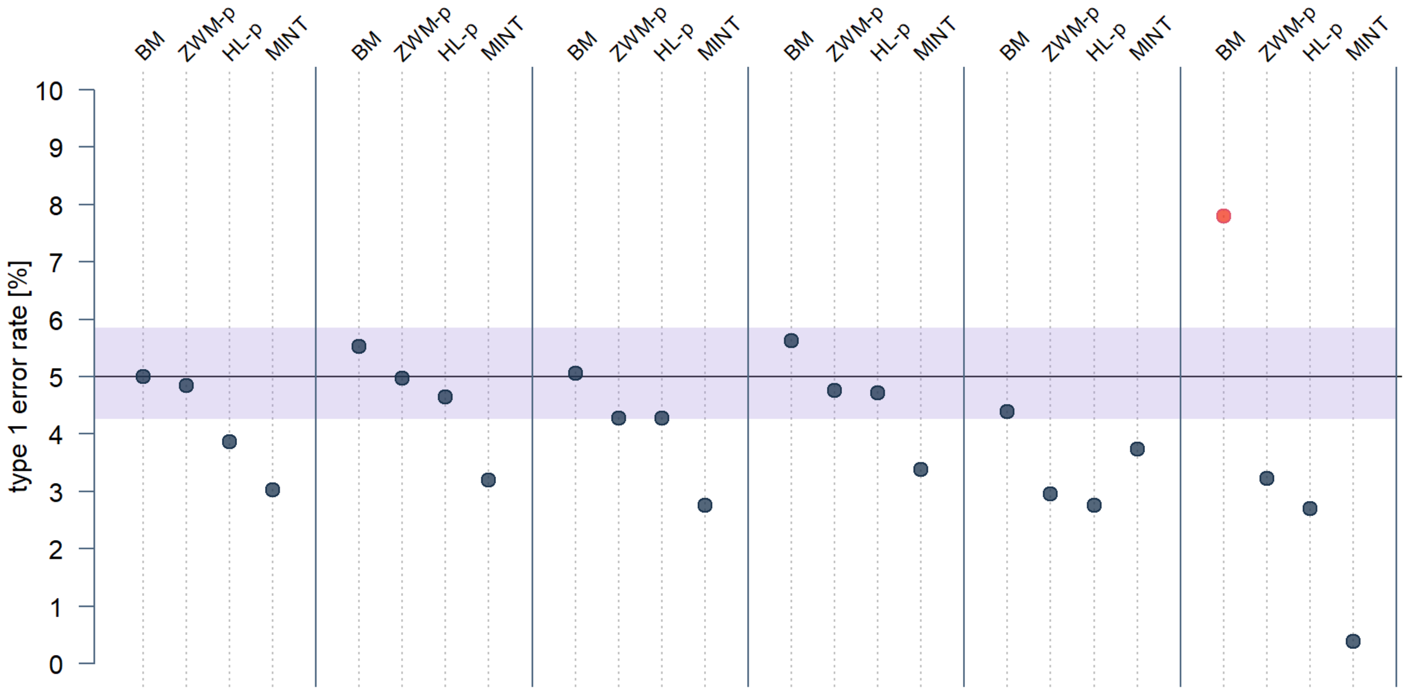


sanofi

Simulation Results

sanofi

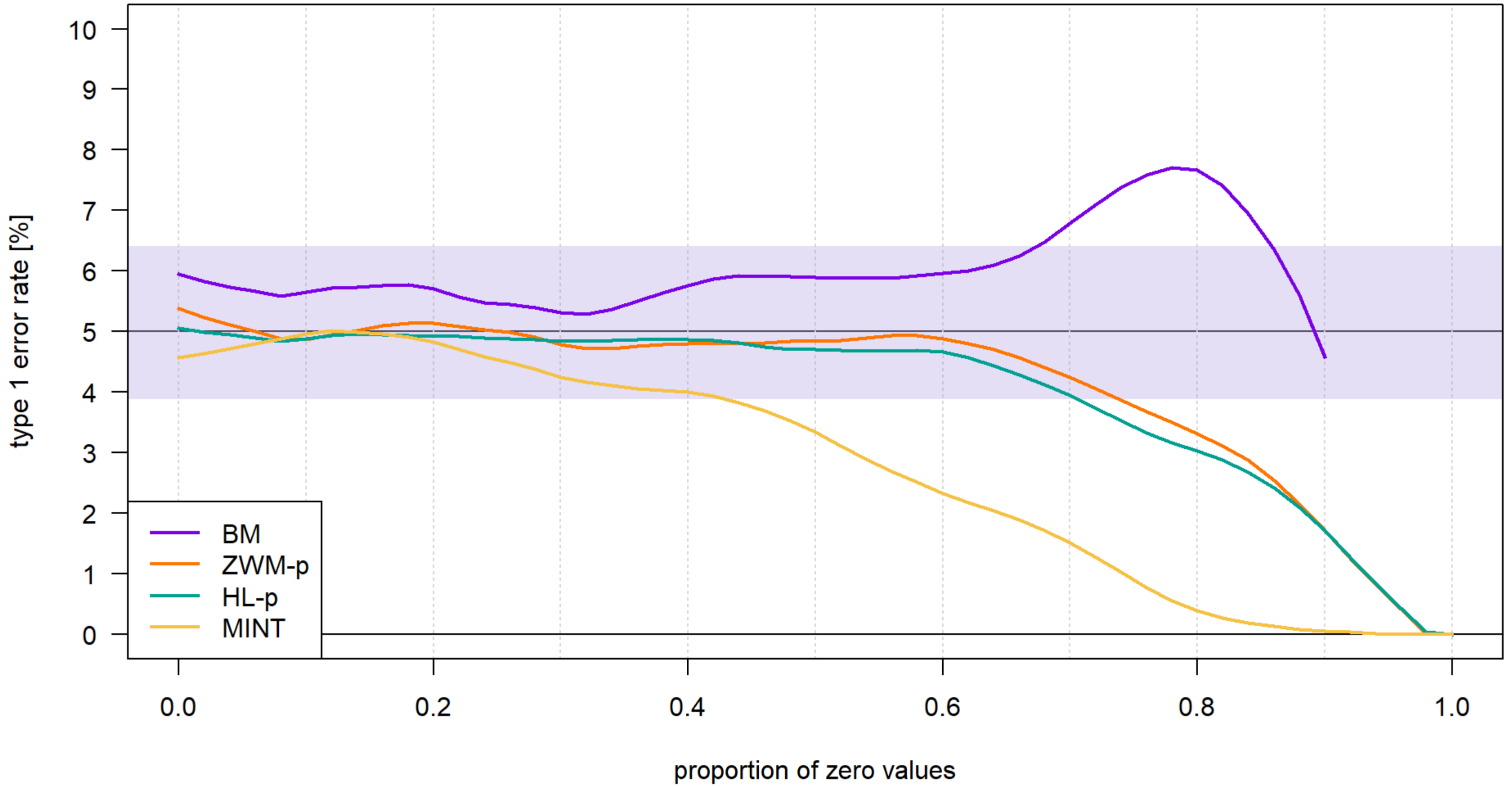
Type 1 error rate



Skewness	0	2.1	0	2.1	0.3	0.4
N	18/18	18/18	18/9	18/9	6/6	18/9
Zeros	30%-70%	30%-70%	30%-70%	30%-70%	30%-70%	75%-85%

$N_{sim}=5000$

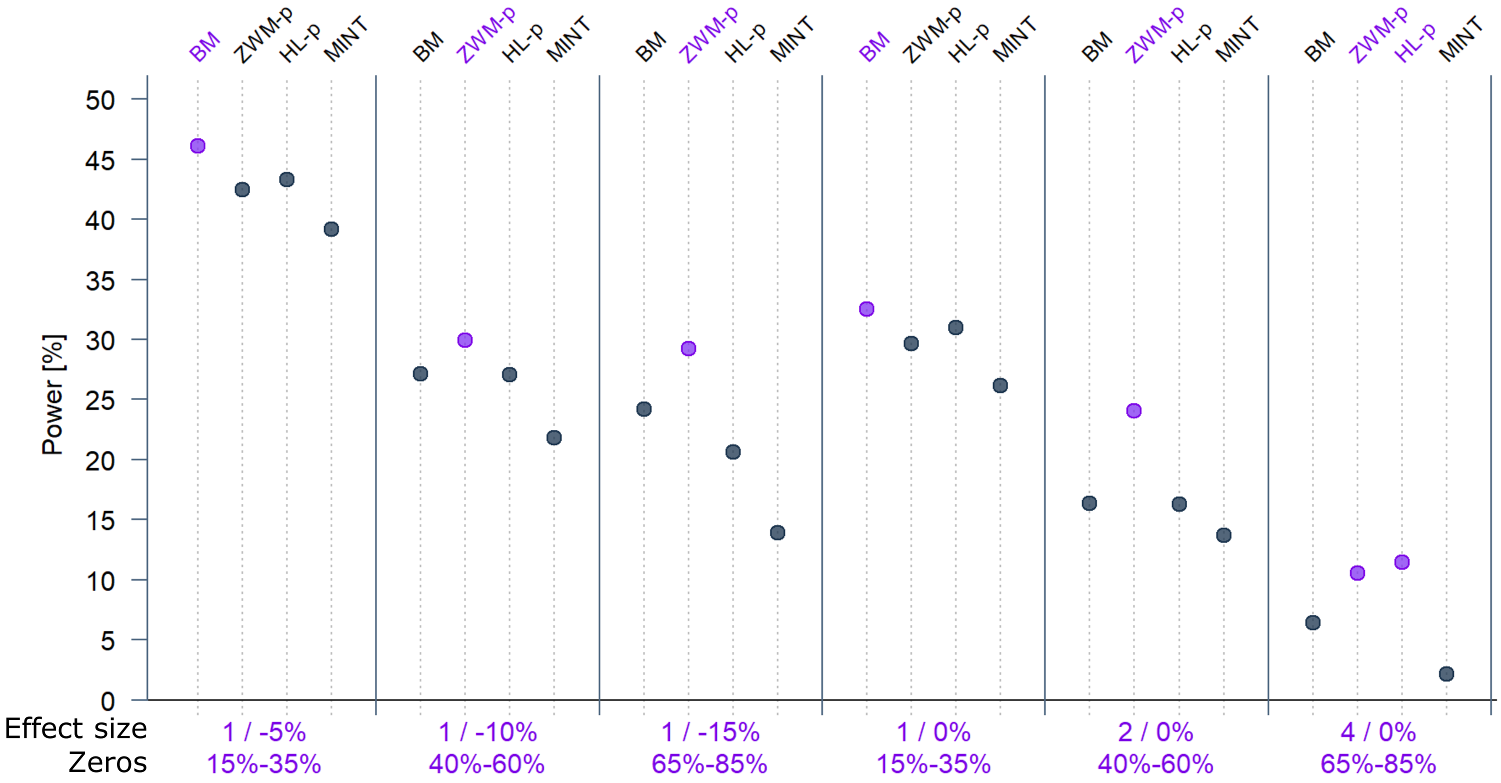
Type 1 error rate



N_{sim}=2000
N = 18/9
Skewness ~ 1

sanofi

Power



N_{sim}=5000
N = 18/9
Skewness ~ 1

sanofi

Conclusion

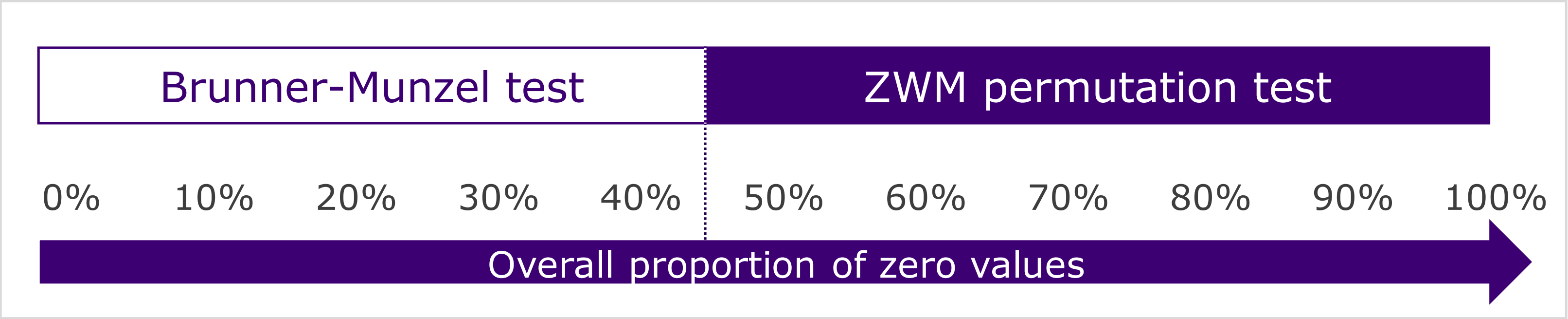
sanofi

Conclusion

The Brunner-Munzel test is most effective when there is a moderate number of zero values in the data.
In contrast, the ZWM-p test performs better when there are many zeros.

All 4 methods can be generalized to more than 2 groups

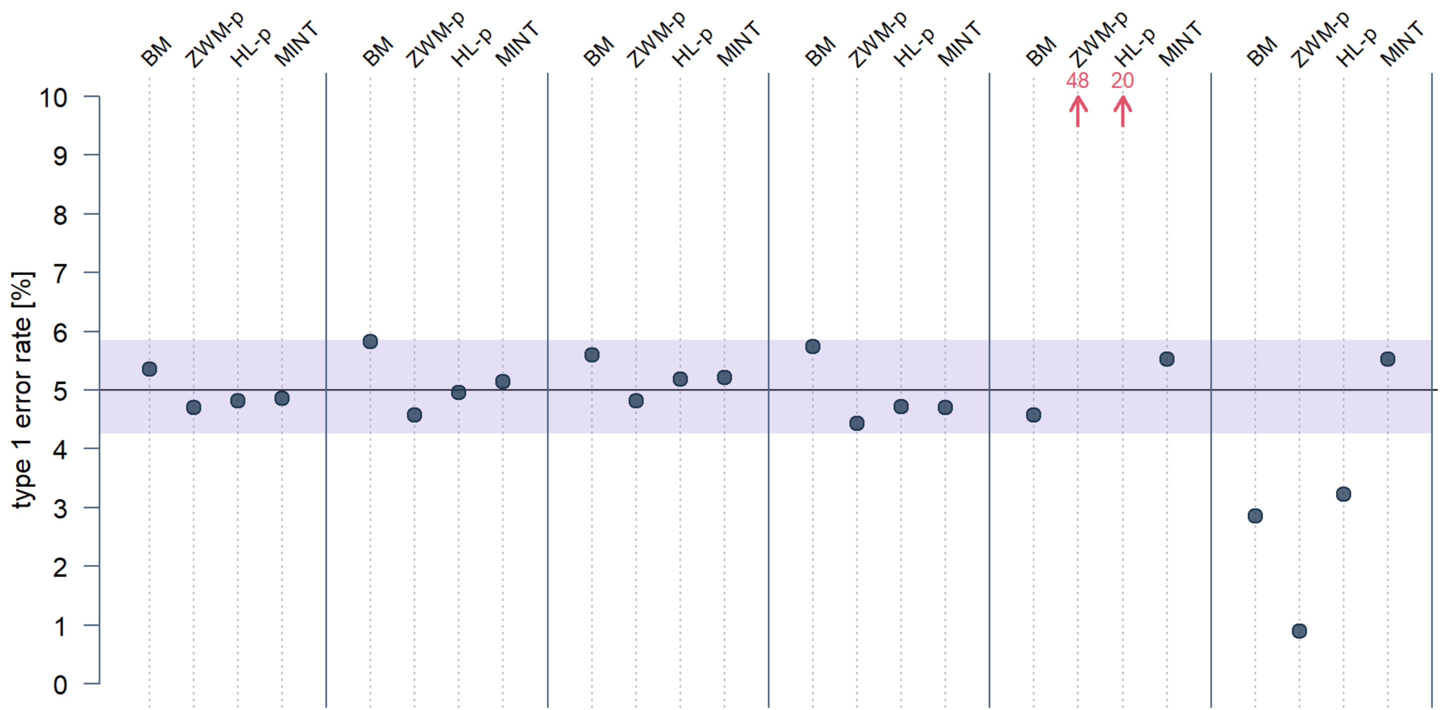
In case of complex design, MINT method may be as well useful, as it is applicable in combination with most parametric methods.



Thank you

sanofi

Type 1 error rate (without zero values)



SKW	0	0	2.3	2.3	0	0
VHF	1	1	1	1	5	1
ZP	0%	0%	0%	0%	0%	0%
N	10/10	10/6	10/10	10/6	10/6	4/4

sanofi

Artificial data example

Random data generation using g-and-h + binomial distribution.