Data Visualisation & Open Source Technology

(Sample) size matters! – demonstrating sample size calculations across software

Agnieszka Tomczyk & Lyn Taylor

(On behalf of CAMIS WG)

11th June 2025

Agenda

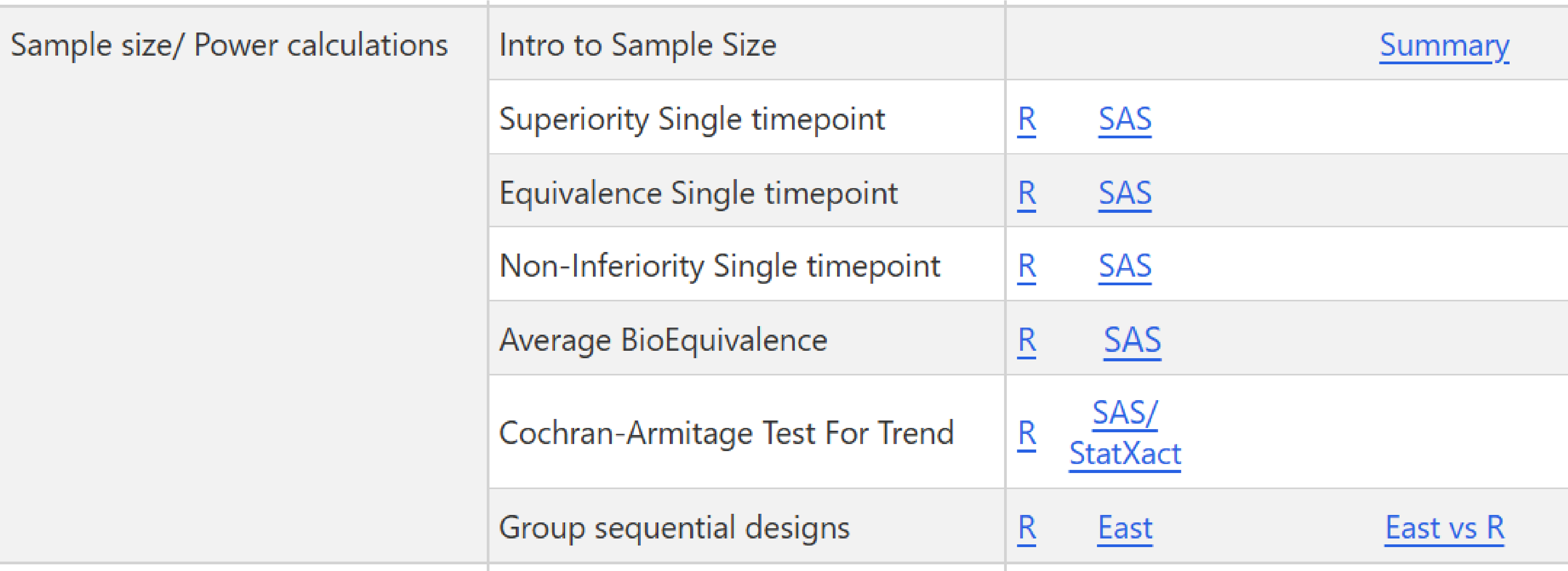
- Introduction to CAMIS project
- Project objectives
- Progress to date
- Selected statistical tests
- Comparison of analyses
- Comparison of results
- Main reasons for discrepancies

What is CAMIS & what are the project objectives?

- CAMIS: Comparing Analysis Method Implementations in Software
- To increase understanding and awareness of analysis result discrepancies across software (R, SAS®, Python etc)
- To demonstrate the methodology through examples
- To document in open GitHub repository
- To grow the repository, increasing quality and quantity of information

Repository location:
https://psiaims.github.io/CAMIS/

Current progress (in the sample size research)



Sample size/ Power calculations	Intro to Sample Size			Summary
	Superiority Single timepoint	R	SAS	
	Equivalence Single timepoint	R	SAS	
	Non-Inferiority Single timepoint	R	SAS	
	Average BioEquivalence	R	SAS	
	Cochran-Armitage Test For Trend	R	SAS/ StatXact	
	Group sequential designs	R	East	East vs R

Considered software:

- R
- SAS
- StatXact
- EAST

Selected statistical tests

- Comparing means in superiority/ non-inferiority/ equivalence studies

  Checking whether the test therapy is better/ as effective as/ difference between the test therapy and the standard therapy is of no clinical importance - considering in each case parallel (unpaired) and cross-over designs

- Average BioEquivalence (BE)

  Two one-sided tests (TOST) to determine whether the average values of the test vs. The standard therapy are comparable. For BE, the 90% CI for the ratio of the averages should fall within a BE limit, usually 80-125%

- Cochran-Armitage Test For Trend

  Testing whether there is a linear trend when the response is binary

- Group sequential designs

  Based on the example of a time-to-event endpoints - phase III oncology trial comparing the test therapy to the standard in terms of progression-free survival (PFS) and overall survival (OS)

Comparison of supported analyses (R/SAS)

Analysis	Supported in R	Supported in SAS	Notes
Means comparison - superiority	YES	YES	Results are matching. For different SD per group in parallel design SAS applies Satterthwaite t-test (Welch's t-test) only, R supports classical t-test as well (MASS). samplesize doesn't support balanced designs.
Means comparison - non-inferiority	YES	YES	Results are matching. Cross-over design is supported only in TrialSize.
Means comparison - equivalence	YES	NO	Results are matching only if the true mean difference = 0. Otherwise, TOST is performed in SAS.
Average Bioequivalence	YES	YES	Results are matching between R (PowerTOST) and SAS – several, more complex approximations are applied (Owen's Q function). TrialSize uses only normal approximation and only for means difference (not ratio).

Comparison of supported analyses (R/SAS)

Analysis	Supported in R	Supported in SAS	Notes
Cochran-Armitage Test	YES	NO, but in StatXact	Supported in R (multiCA) and StatXact only. StatXact is calculating the (exact or asymptotic) power of the exact CA test, and not the exact power of the asymptotic test. That is not implemented in multiCA package in R. R supports both multinomial and binary outcome; StatXact only the latter. Results often differ.
Group sequential designs	YES	NO, but in EAST	Both LF and KT methods are implemented in gsDesign, while KT method is implemented in EAST and rpact. gsDesign2 uses a modification of the LF method while applying an average hazard ratio (AHR) approach. Usage of different log hazard ratio variance assumptions: EAST uses the variance under the null hypothesis and provides an option for using the variance under the alternative hypothesis. gsDesign, on the other hand, is using both of these variances as suggested by LF. gsDesign2 can do either.

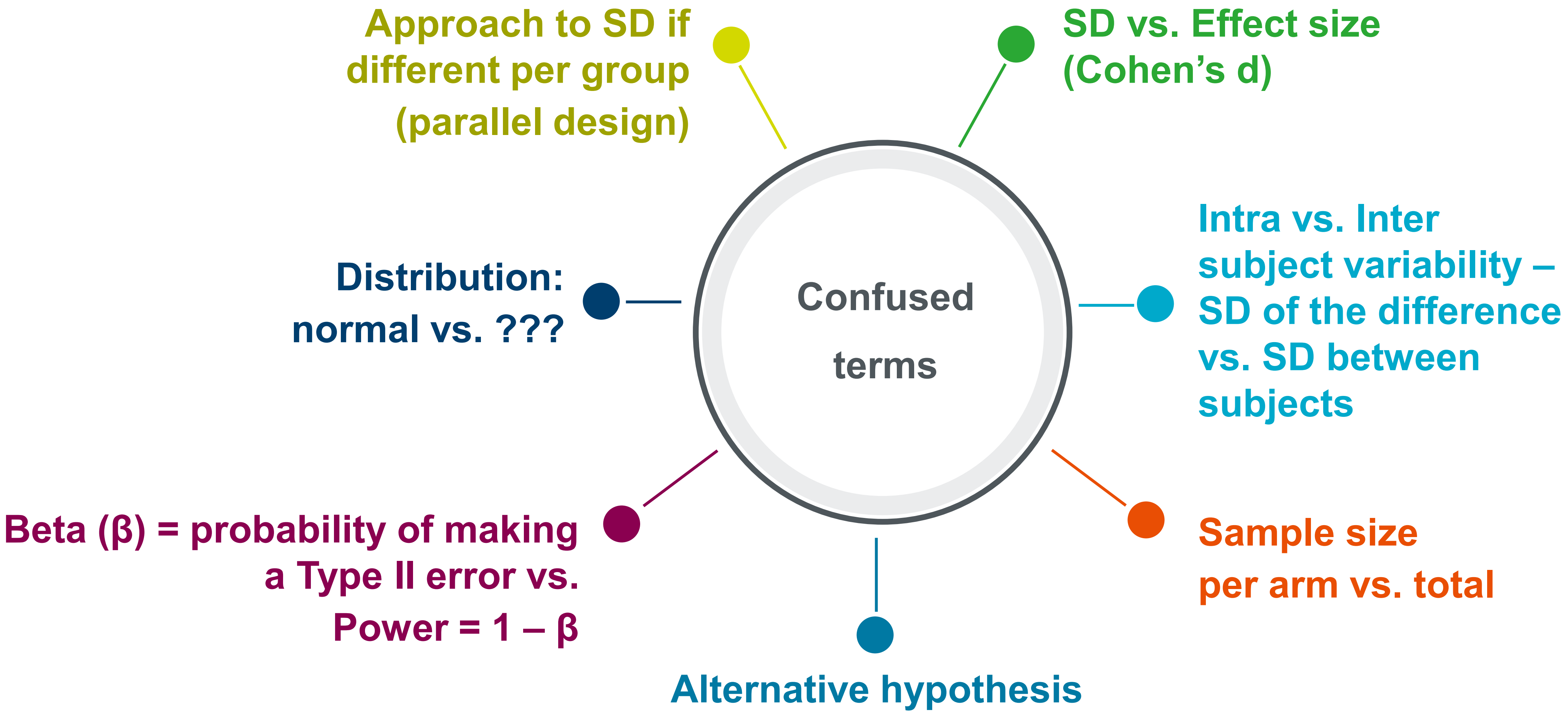
Sample size results comparison

Analysis	Details	R	SAS (StatXact/EAST)
means - superiority	SD1 = SD2, parallel	samplesize: 57 stats: 56.16413 pwr: 56.164	114 total = 57 per arm
	SD1 =/= SD2, parallel	MESS: 328 = 164 + 164 samplesize: 252 = 56 + 196	330 = 165 + 165
means - equivalence	parallel, true mean diff = 0	TrialSize: 43.8469 SampleSize4ClinicalTrials: 44	90 total = 45 per arm
	parallel, true mean diff ≠0	TrialSize: 107.0481 SampleSize4ClinicalTrials: 108	140 total = 70 per arm
	cross-over, true mean diff = 0	TrialSize: 7.612309	8 per arm
avg. bioequivalence	2x2 crossover, mean ratio (log scale)	PowerTOST: 40 total	38 total
	2x2 crossover, mean diff	PowerTOST: 70 total TrialSize: 21 per arm (42 total)	69 total
C-A trend test		multiCA: 526.2628 total = 106 per arm	StatXact: asymptotic: 104 per arm exact: 108 per arm
Group sequential designs		gsDesign gsDesign2 rpact	EAST

Reasons for discrepancies

- Different analysis method
  - Or even statistical test! (equivalence vs. TOST)
  - Confusing terms/parameters
- Rounding…
- Bug in the software

Reasons for discrepancies



Confused terms

Approach to SD if different per group (parallel design)

SD vs. Effect size (Cohen’s d)

Distribution: normal vs. ???

Intra vs. Inter subject variability – SD of the difference vs. SD between subjects

Beta (β) = probability of making a Type II error vs. Power = 1 – β

Sample size per arm vs. total

Alternative hypothesis

Lyn

Agnieszka

To contribute to CAMIS & join our monthly meetings please contact Lyn at: lyn.taylor@parexel.com

For collaboration on sample size & power calculation please contact me at: agnieszka.tomczyk@parexel.com

Meet us at the „SIG at the Bar” session! 🥂

Thank you!