A new interpretation to the equivalence test (TOST)

by (Bayesian) Success Probabilities

Bernard G Francq, Ron S Kenett
September 2024

GSK

ENBIS Webinar

3rd February 2022, 2:00 – 3:30 PM (CET)

Statistical Significance and p-values

https://enbis.org/

- Neyman-Pearson framework
- Frequentists and Bayesian Intervals
- Trends Towards Significance
- Fallacies and practical approaches

Learn from our expert panellists

Daniël Lakens
Associate Professor
Eindhoven University
of Technology
Netherlands

Bernard G Francq
Lead Statistician
GSK Vaccines
Belgium

Stephen Senn
Statistical Consultant
Edinburgh, Scotland,
United Kingdom

Ron S Kenett
Chairman of the
KPA Group
Israel

Moderator: Prof Jean-Michel Poggi (France)

A **new interpretation** to the TOST by **Success Probabilities**

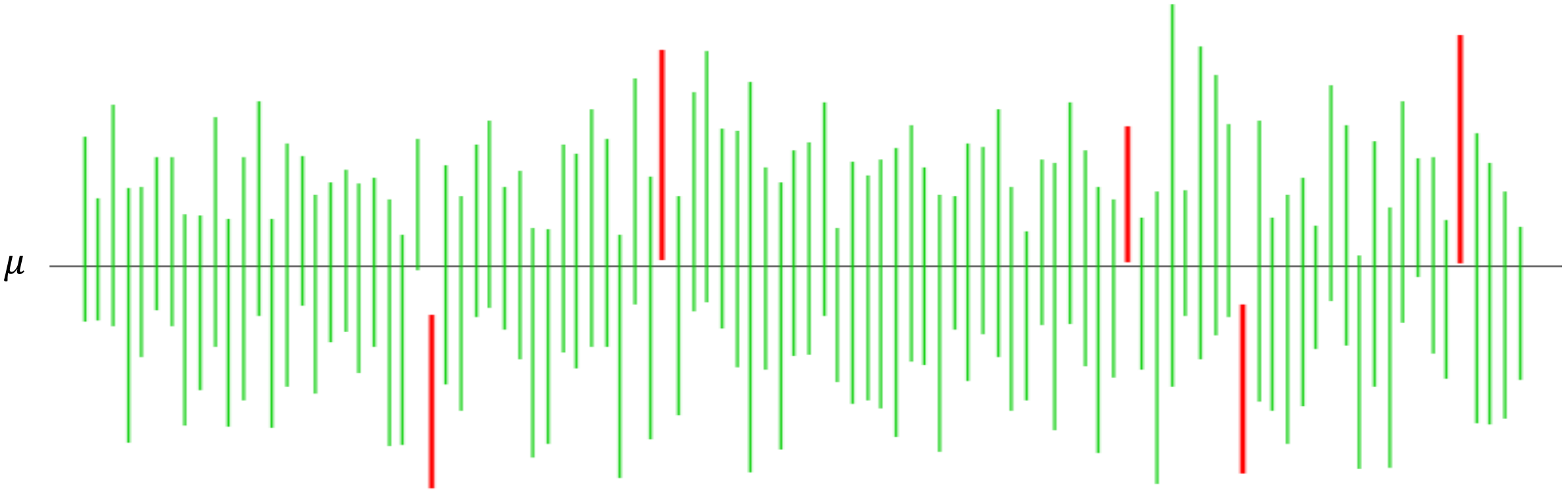
This research is about interpreting results (an alternative to the p-values and confidence intervals for mean, mean difference or mean predictions), not model selection.

Statistical Intervals

- Confidence
- Prediction
- Tolerance

Confidence Interval concept

100 simulated 95% CI for the mean μ



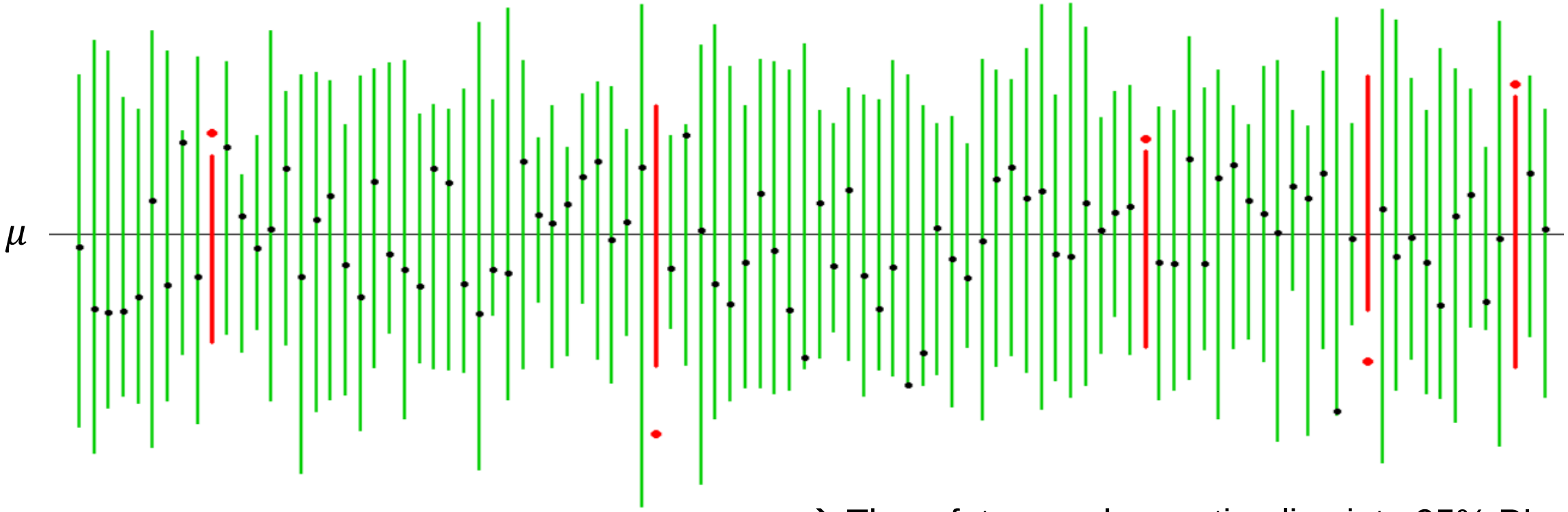
→ The true value, μ, lies in 95% of the CIs

Note: in Bayesian statistics, credible intervals are commonly used

GSK

Prediction Interval concept

100 simulated 95% PI for a future observation



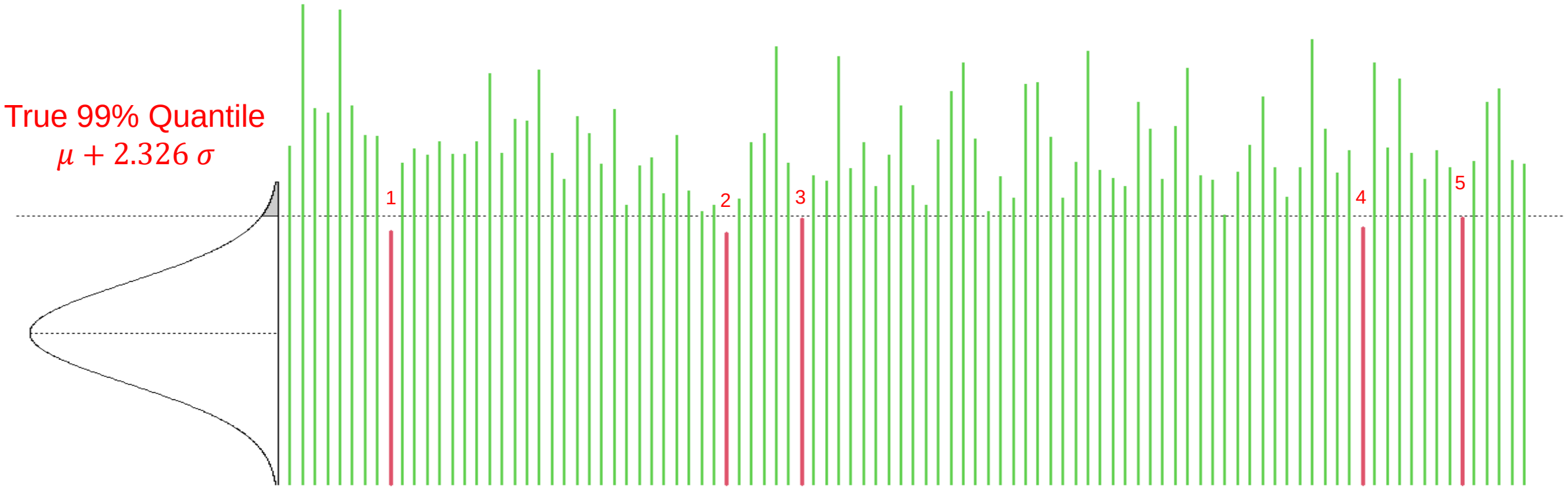
→ The « future » observation lies into 95% PIs

Note: in Bayesian statistics, PI can be obtained by credible intervals from the posterior distribution

GSK

1-sided Content Tolerance Interval - concept

100 simulated (upper) 1-sided 99% content TI (95% confidence)



→ 95 TIs cover **at least** 99% of the population
5 TIs cover **at most** 99% of the population

A 1-sided TI is **identical** to calculating a 1-sided Confidence Interval on a quantile

GSK

Exact 1-sided Tolerance Intervals

TIs encompass a given proportion of the population with a given confidence level

The exact 1-sided TI is given by the non-central t-distribution

$$\bar{X} \pm t_{conf,n-1,\boldsymbol{z_{pred}\sqrt{n}}} \frac{S}{\sqrt{n}}$$

- $-$ or $+$ must be chosen according to the context
- S is the sample standard deviation, $\bar{X}$ the estimated mean, n the sample size
- $conf$ is the desired confidence level
- $pred$ is the desired prediction level (coverage)
- $n-1$ are the degrees of freedom
- $\boldsymbol{z_{pred}\sqrt{n}}$ is the non-centrality parameter
- $\boldsymbol{z_{pred}}$ is the quantile of the standardized normal distribution

GSK

1-sample t-test

1-sample t-test synthetic examples

What if the sample size increases (with identical mean and SD)?

Toy example on SBP (mmHg)

Success Probabilities constant

				$H_1: \mu < 140$		SP (Probability Index)	
n	$\overline{X}$	S	90% CI	p-value	s-value # Head	$P(X < 140)$	$P(X > 140)$
20	138.11	7.97	[135.0, 141.2]	p=0.15	2.7	59.4%	40.6%
50	138.11	7.97	[136.2, 140.0]	p=0.05	4.3	59.4%	40.6%
100	138.11	7.97	[136.8, 139.4]	p=0.0098	6.7	59.4%	40.6%
200	138.11	7.97	[137.2, 139.0]	p=5E-4	11	59.4%	40.6%
10^3	138.11	7.97	[137.7, 138.5]	p=7E-14	44	59.4%	40.6%

p<0.001

p-values collapse, s-values soar

Add the confidence bounds by using the TI methodology

GSK

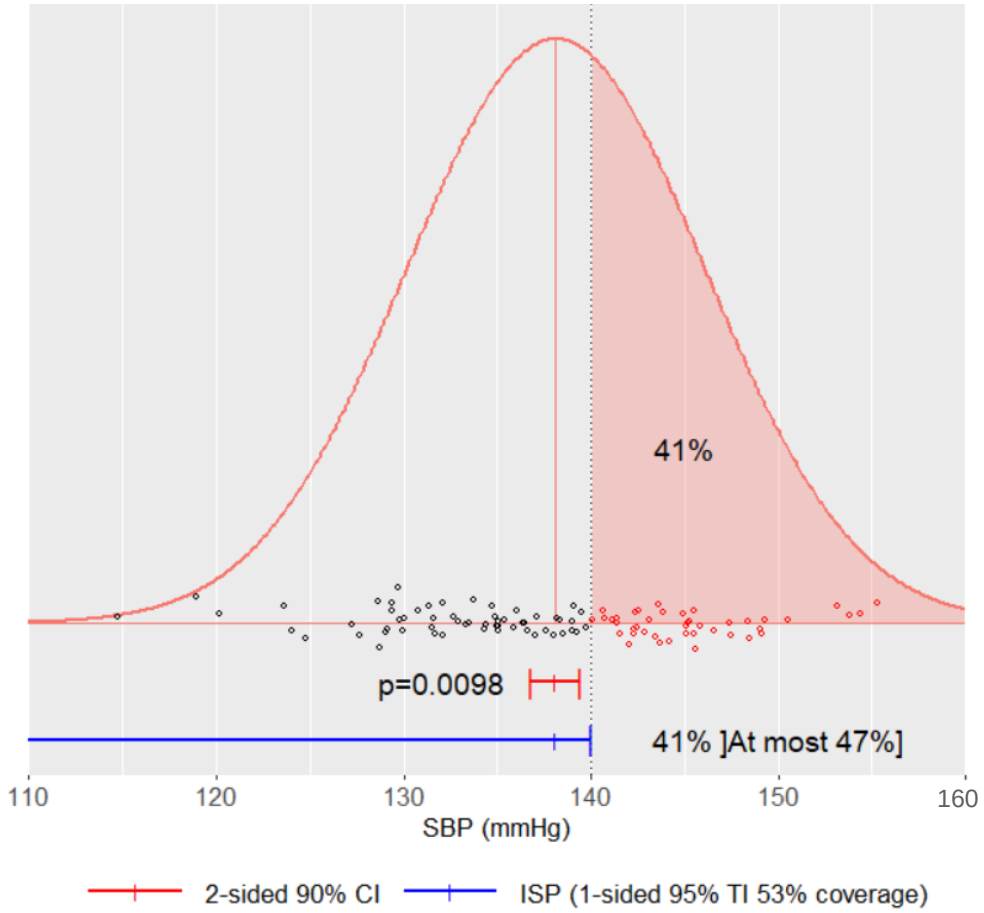
1-sample t-test by Success Probability

What should be the value of the prediction level (coverage) for the TI to be equal to 140 ?

$$138.11 + t_{0.95,100-1,z_{pred}\sqrt{100}} \frac{7.97}{\sqrt{100}} = 140$$

- *At most 47% of the patients have a SBP > 140*

→ *This is the 95% upper (lower) bound for the SP*



1-sample t-test synthetic examples

				$H_1: \mu < 140$		SP (95% CI)	
n	$\bar{X}$	S	90% CI	p-value	s-value # Head	$P(X < 140)$	$P(X > 140)$
20	138.11	7.97	[135.0, 141.2]	p=0.15	2.7	59.4 [44.5[%	40.6]55.5]%
50	138.11	7.97	[136.2, 140.0]	p=0.05	4.3	59.4 [50.0[%	40.6]50.0]%
100	138.11	7.97	[136.8, 139.4]	p=0.0098	6.7	59.4 [52.8[%	40.6]47.2]%
200	138.11	7.97	[137.2, 139.0]	p<0.001	11	59.4 [54.7[%	40.6]45.3]%
10^3	138.11	7.97	[137.7, 138.5]	p<0.001	44	59.4 [**57.3**[%	40.6]**42.7**]%

CI and p-value might be confusing

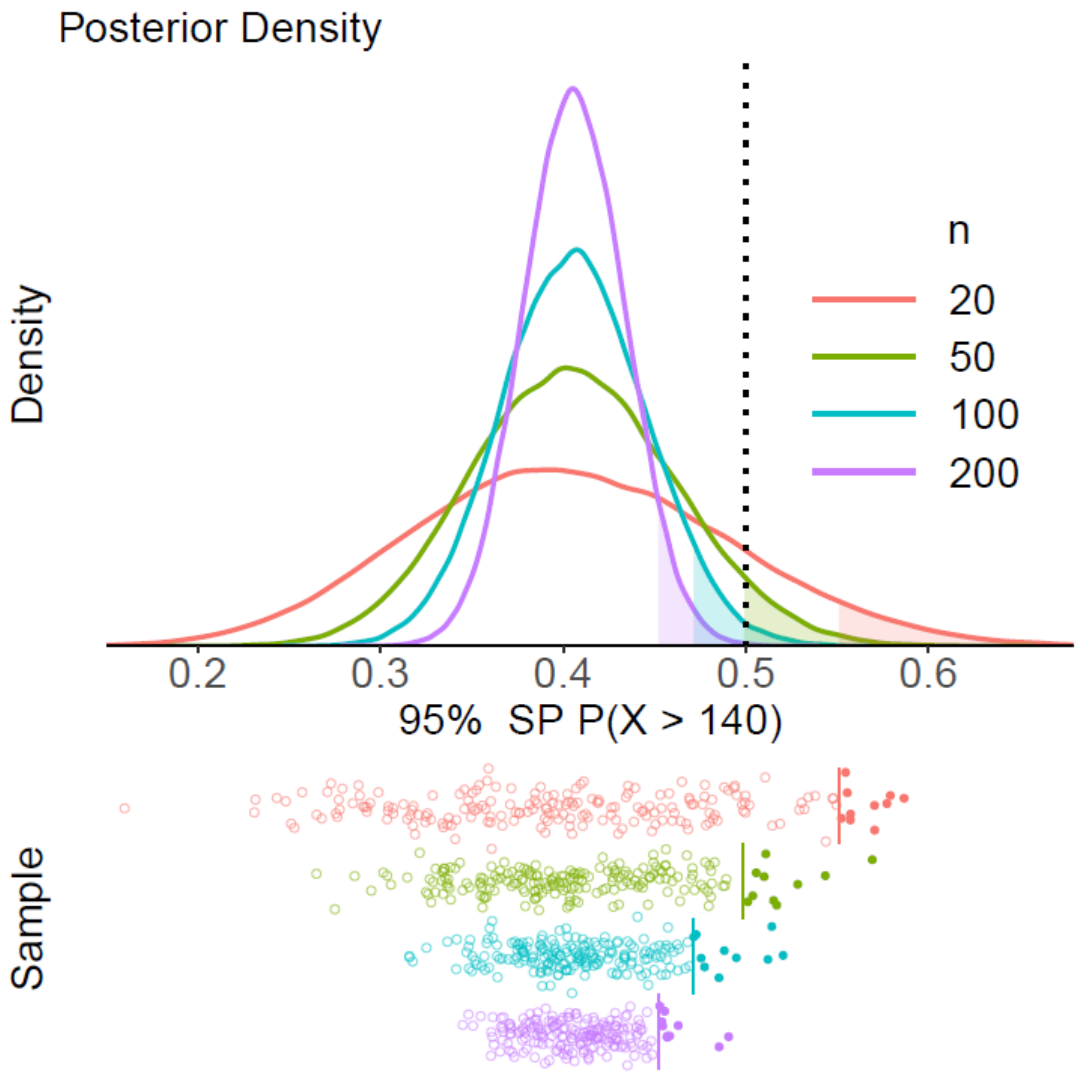
The SP interpretation is straightforward even for big sample sizes (eg $n = 10^3$)

95% confidence that

- ✓ ***At least*** *57.3% of the (new) patients will have a SBP <140 mmHg (success)*
- ✓ ***At most*** *42.7% of the (new) patients will have a SBP >140 mmHg (failure)*

GSK

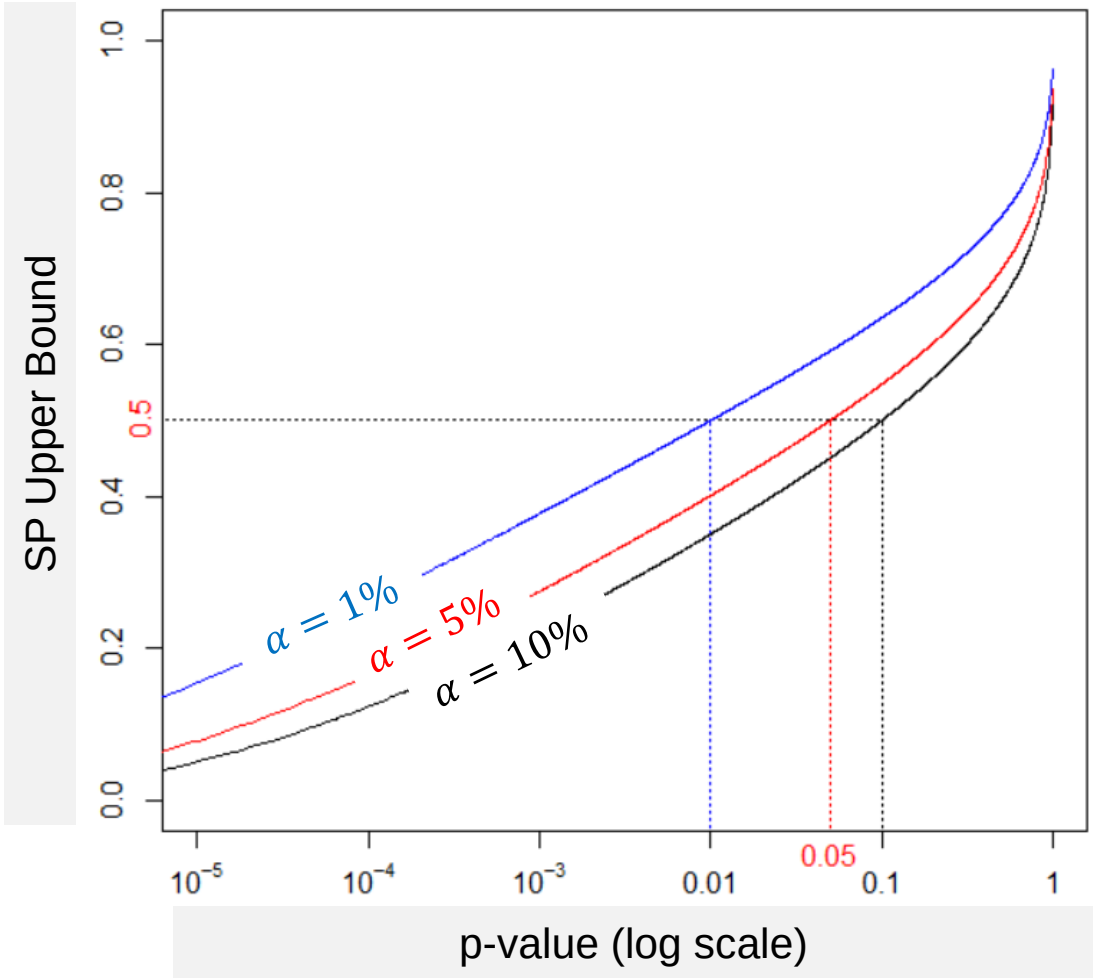
1-sample t-test **Bayesian** synthetic examples



	$H_0 : \mu = 140$ $H_1 : \mu \neq 140$		Success Probability (95% confidence) $P(X > 140)$	
n	Mean	SD	Frequentist	Bayesian
20	138.5	7.56	42.1 [26.2, 59.7]	42.1 [25.9, 59.4]%
50	138.5	7.56	42.1 [31.7, 53.3]	42.1 [31.5, 53.2]%
100	138.5	7.56	42.1 [34.6, 50.0]	42.1 [34.6, 49.9]%
200	138.5	7.56	42.1 [36.8, 47.7]	42.1 [36.7, 47.6]%
1000	138.5	7.56	42.1 [39.7, 44.6]	42.1 [39.7, 44.6]%

GSK

One-to-one function SP & p-value



$$X \sim N(\mu = 145, \sigma = 5)$$
$$n = 10$$
$$H_0: \mu = 140,\ H_1: \mu > 140$$

The (upper bound) SP is
a one-to-one function with the p-value

Advantages of the SP over the p-value

- ✓ Easy to interpret
- ✓ No tiny values
- ✓ No need to use sophisticated rounding rules
- ✓ Realistic and pragmatic interpretation
- ✓ Similar interpretation *frequentist* and *Bayesian*
- ✓ Identical interpretation for log or no-log data
- ✓ The cut-off value is 50% (the middle of the probability scale), an intuitive threshold, whatever the type I error

GSK

TOST

Two One-Sided (t)-Tests

TOST: synthetic examples

$$\begin{cases} \bar{X} - t_{1-\alpha\prime,n-1,z_{\gamma_1}} \sqrt{n} \dfrac{S}{\sqrt{n}} = \Delta_1 \\ \bar{X} - t_{1-\alpha\prime,n-1,z_{\gamma\prime_2}} \sqrt{n} \dfrac{S}{\sqrt{n}} = \Delta_2 \\ \bar{X} + t_{1-\alpha\prime,n-1,z_{\gamma\prime_1}} \sqrt{n} \dfrac{S}{\sqrt{n}} = \Delta_1 \\ \bar{X} + t_{1-\alpha\prime,n-1,z_{\gamma_2}} \sqrt{n} \dfrac{S}{\sqrt{n}} = \Delta_2 \end{cases}$$

$H_0: \mu \notin [\Delta_1, \Delta_2], H_1: \mu \in [\Delta_1, \Delta_2]$

Solving the equations for

- γ_1, γ_2, γ_1' and γ_2' will give the exact bounds of the success probabilities
- Bayesian results are similar (vague prior)
- Advantage Bayesian: Prior information might be used

$H_0 : \mu \notin [11.5, 13]$			Classical TOST		Success Probability (90% confidence)		
$H_1 : \mu \in [11.5, 13]$			Mean		$P(X < 11.5)$	$P(X > 13)$	
n	Mean	SD	90% CI	p-value	Frequentist	Frequentist	Bayesian
20	12.5	3.01	[11.3, 13.7]	p=0.23	37.0 [24.0, 52.0]%	43.4 [29.7, 58.2]	43.3 [29.3, 57.8]%
50	12.5	3.01	[11.8, 13.2]	p=0.12	37.0 [28.4, 46.4]%	43.4 [34.5, 52.8]	43.3 [34.3, 52.7]%
100	12.5	3.01	[12.0, 13.0]	p=0.05	37.0 [30.8, 43.6]%	43.4 [37.0, 50.0]	43.4 [37.0, 50.0]%
200	12.5	3.01	[12.1, 12.9]	p=9.9E-3	37.0 [32.6, 41.6]%	43.4 [38.9, 48.1]	43.4 [38.9, 48.1]%
1000	12.5	3.01	[12.3, 12.7]	p=9.3E-8	37.0 [35.0, 39.0]%	43.4 [41.4, 45.5]	43.4 [41.4, 45.5]%

GSK

DOE

Success Probabilities *by* TOST

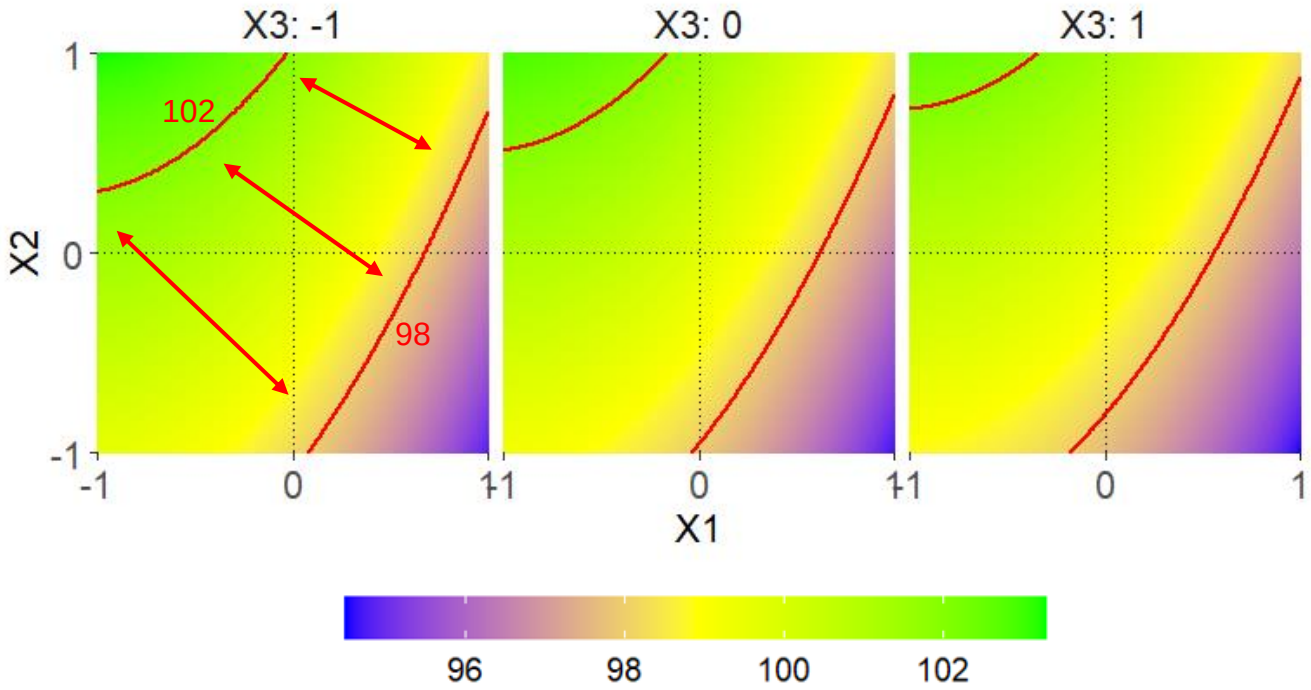
TOST in DOE

Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$
- Find the equivalence region

- Warning
  The area between 98 and 102 is not $H_1: \mu_Y \in [98, 102]$ because the uncertainties are not taken into account

Heatmap of Predictions



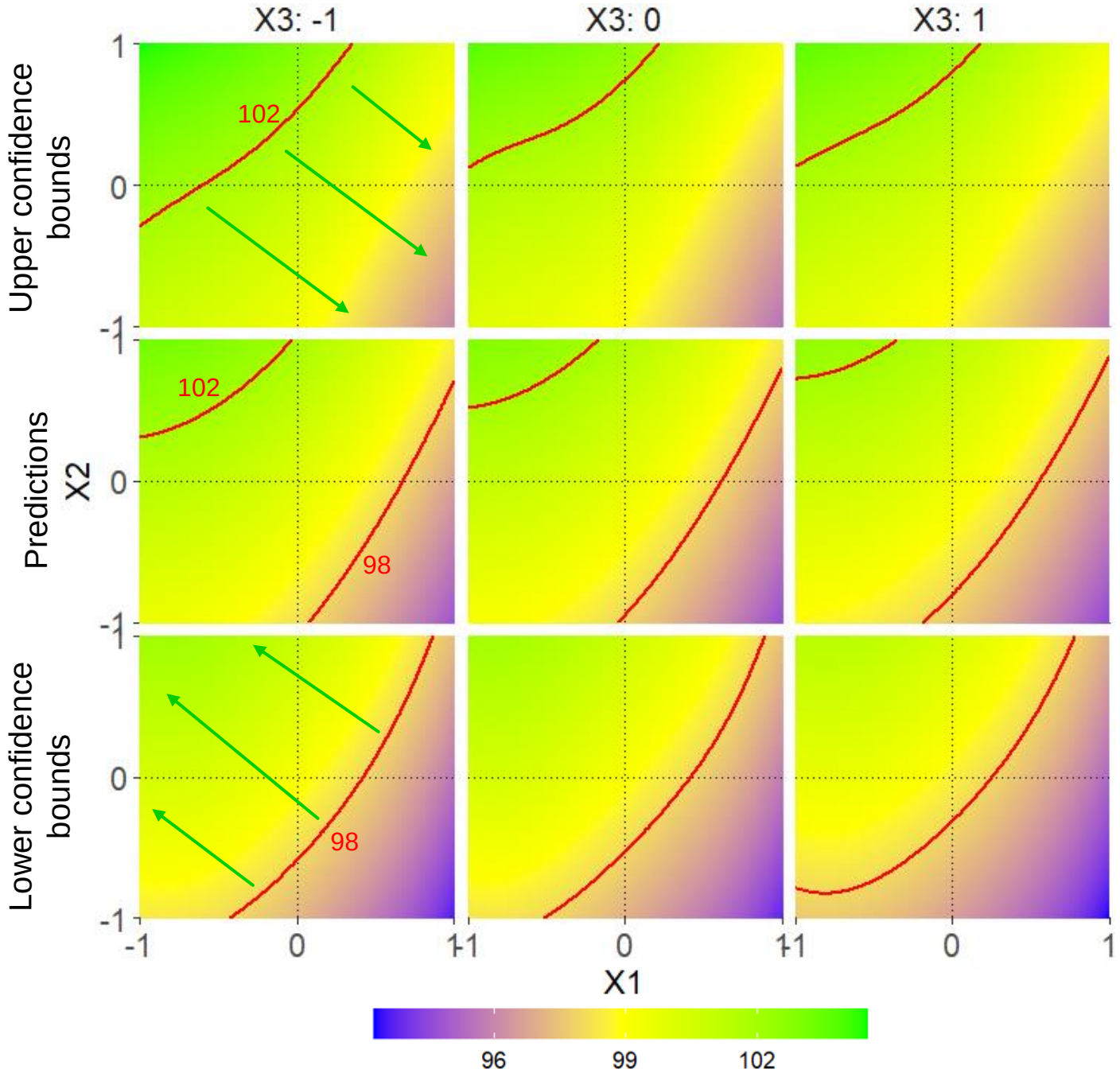
TOST in DOE

Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$
- Find the equivalence region

Heatmap

- The equivalence region must combine the confidence bounds



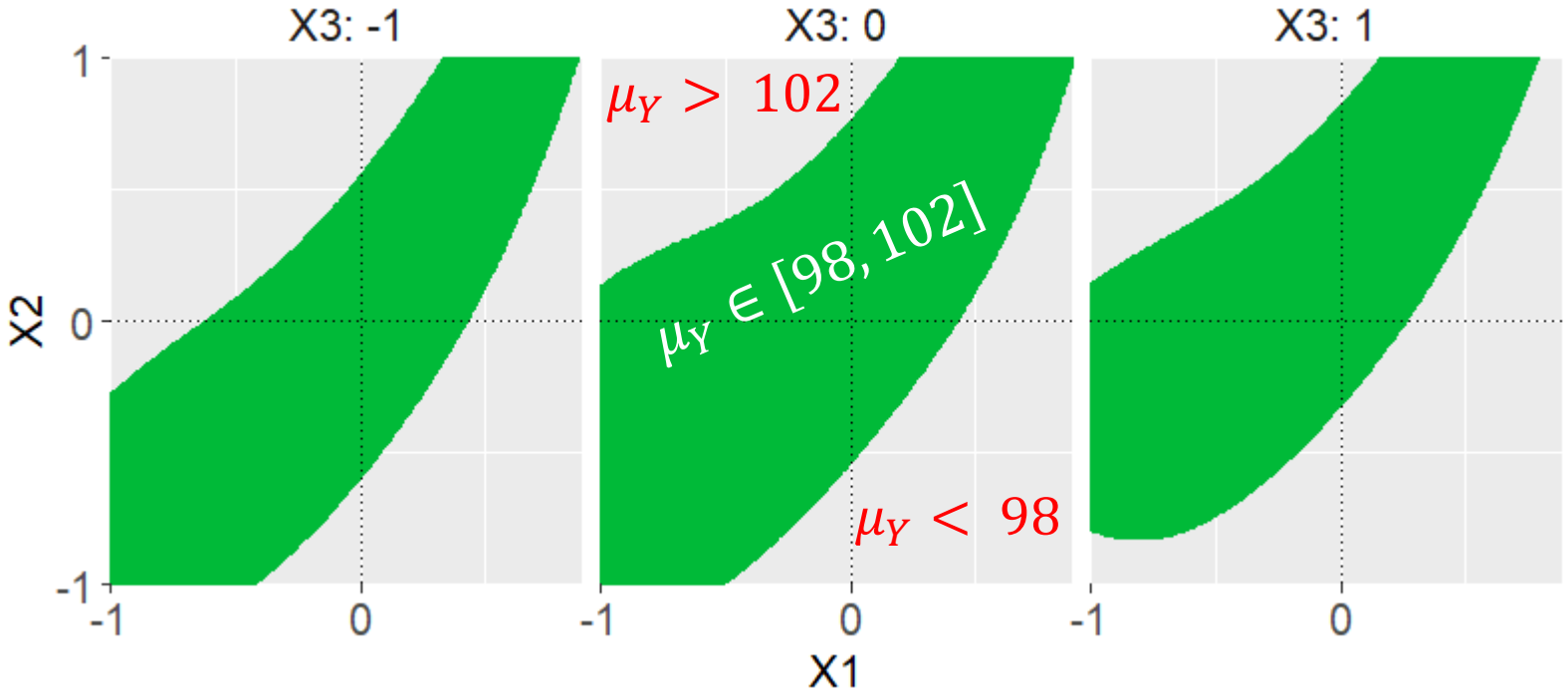
TOST in DOE by Confidence Intervals

Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$

- The equivalence region comes from the joint areas $H_1: \mu_Y > 98$ and $H_1: \mu_Y < 102$

- However, it does not give any further view on the probability to be within the specifications



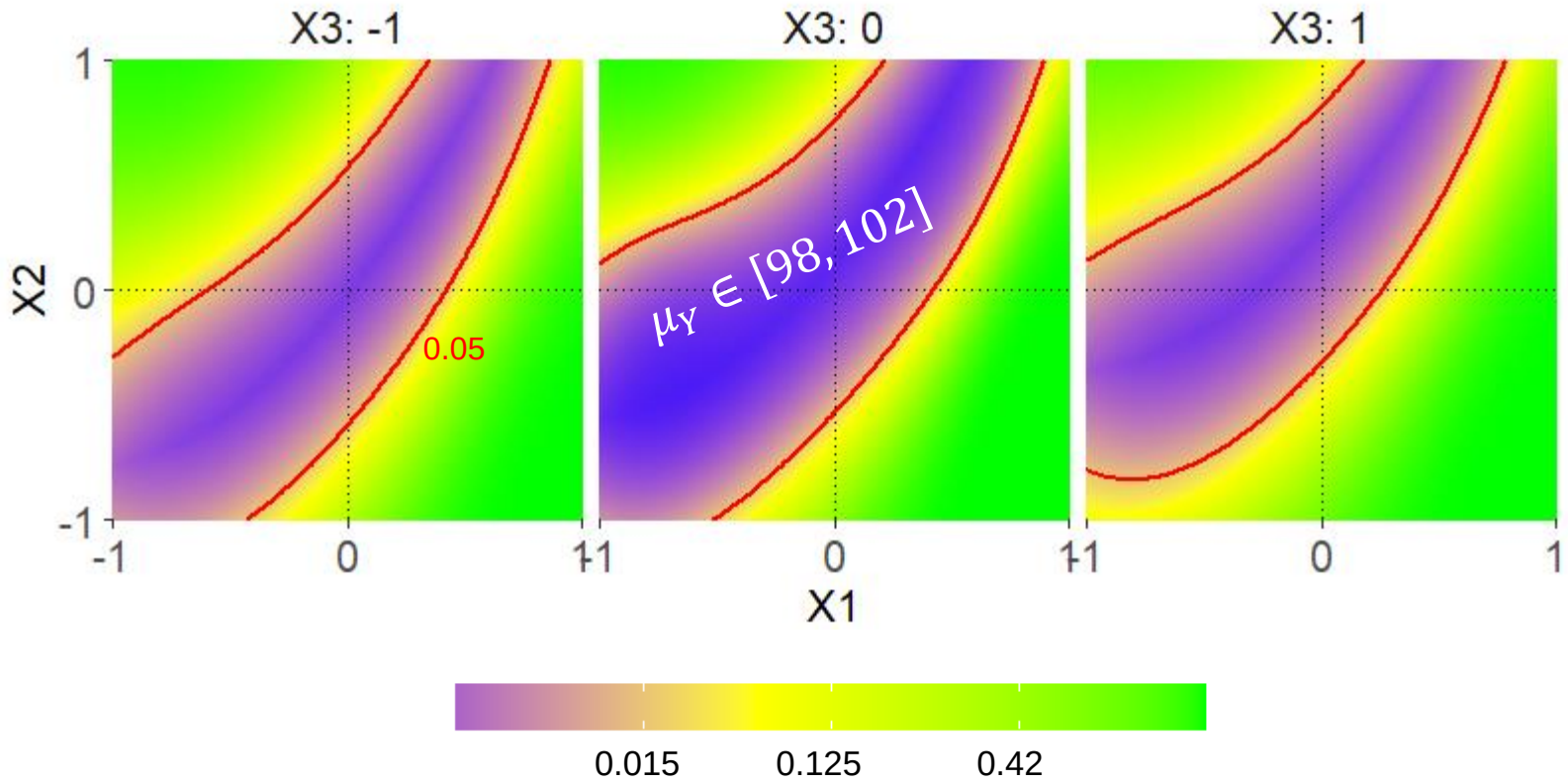
TOST in DOE by equivalence p-values

Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$

- The p-value is quite complex to interpret !
- The threshold (5%) is not intuitive !
- Skewed distribution

Heatmap of TOST p-values (target < 0.05)



Remark: The p-value scale is not linear…

TOST in DOE by **Success Probabilities**

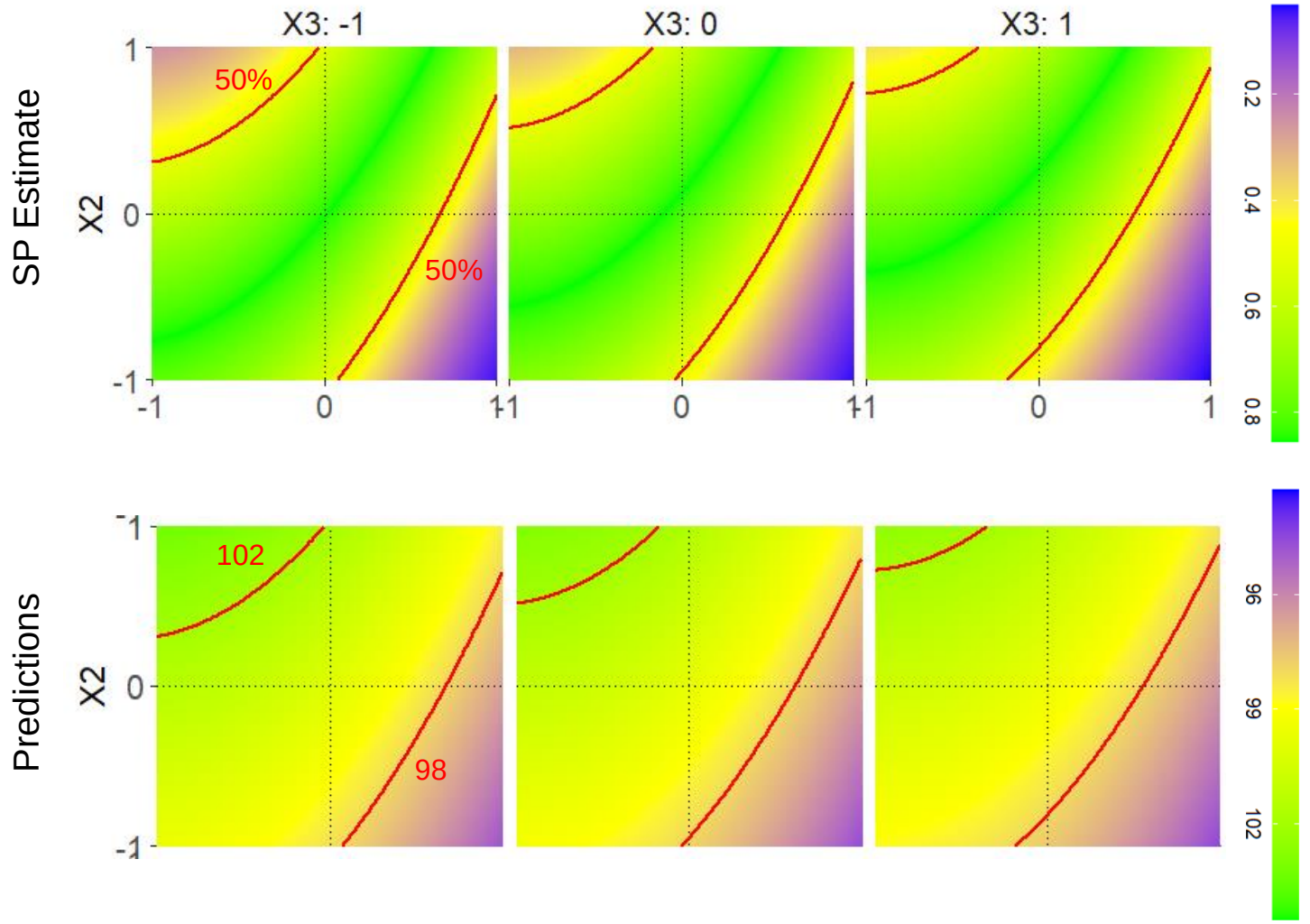
Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$

- Success Probabilities are quite straightforward to interpret

- At least 50% of the products are expected to be compliant in the equivalence region
- This is identical to the predictions
- The uncertainties are not taken into account

Heatmap of TOST
Success Probabilities (SP) Estimate (target > 0.5)



TOST in DOE by **Success Probabilities**

Consider a DOE with 3-X continuous variables

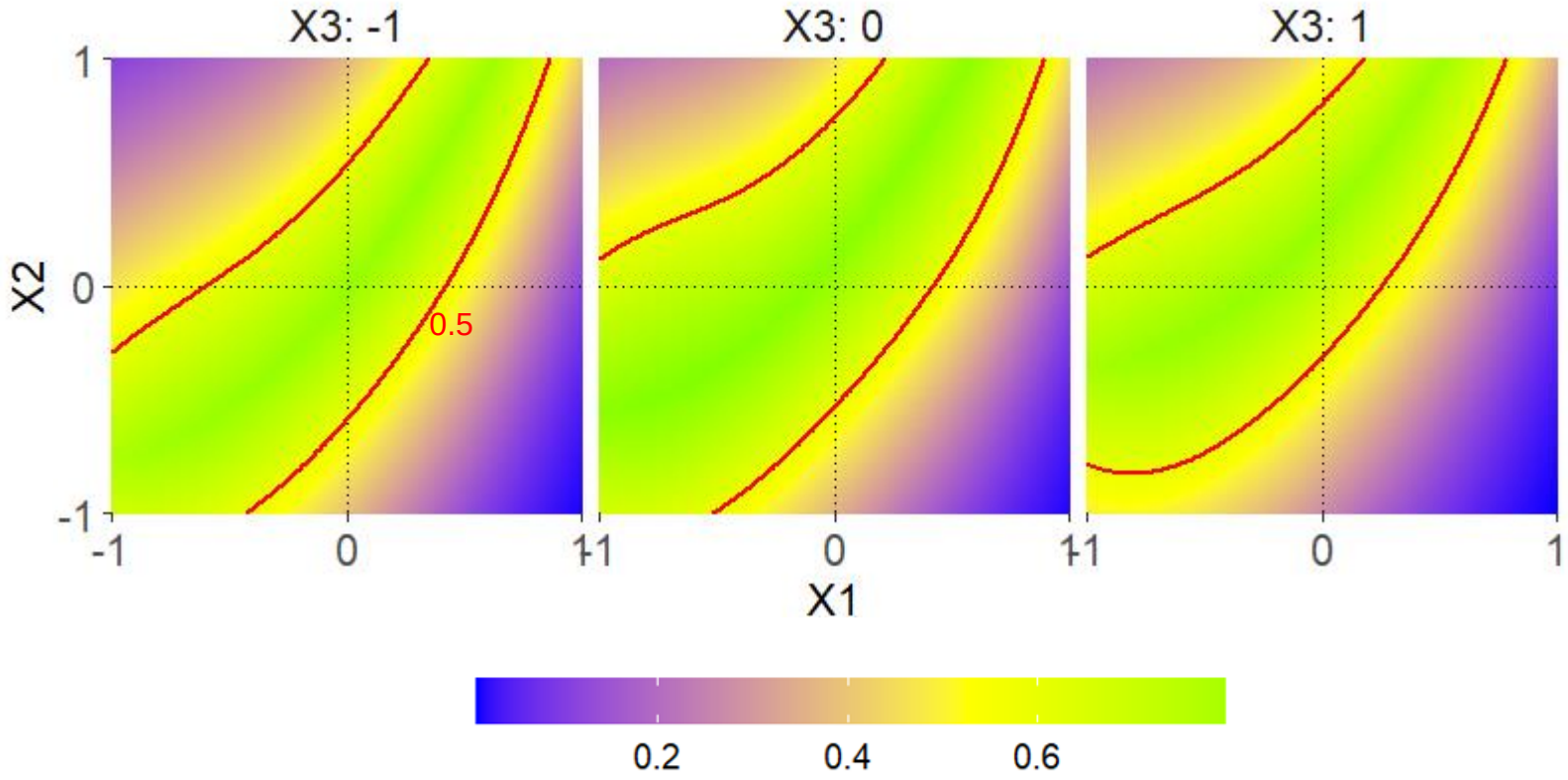
- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$

- Success Probabilities are quite straightforward to interpret
- Intuitive cutoff value 50%

- 95% confident that at least 50% of the future products (in the equivalence region) will be compliant
  - At least ~70% compliance at target
  - At most ~30% non-compliant

- The non-compliant products proportion might soar to ~90% outside the equivalence region

Heatmap of TOST
Success Probabilities (SP) Confidence Bound (target > 0.5)



TOST in DOE - **Summary**

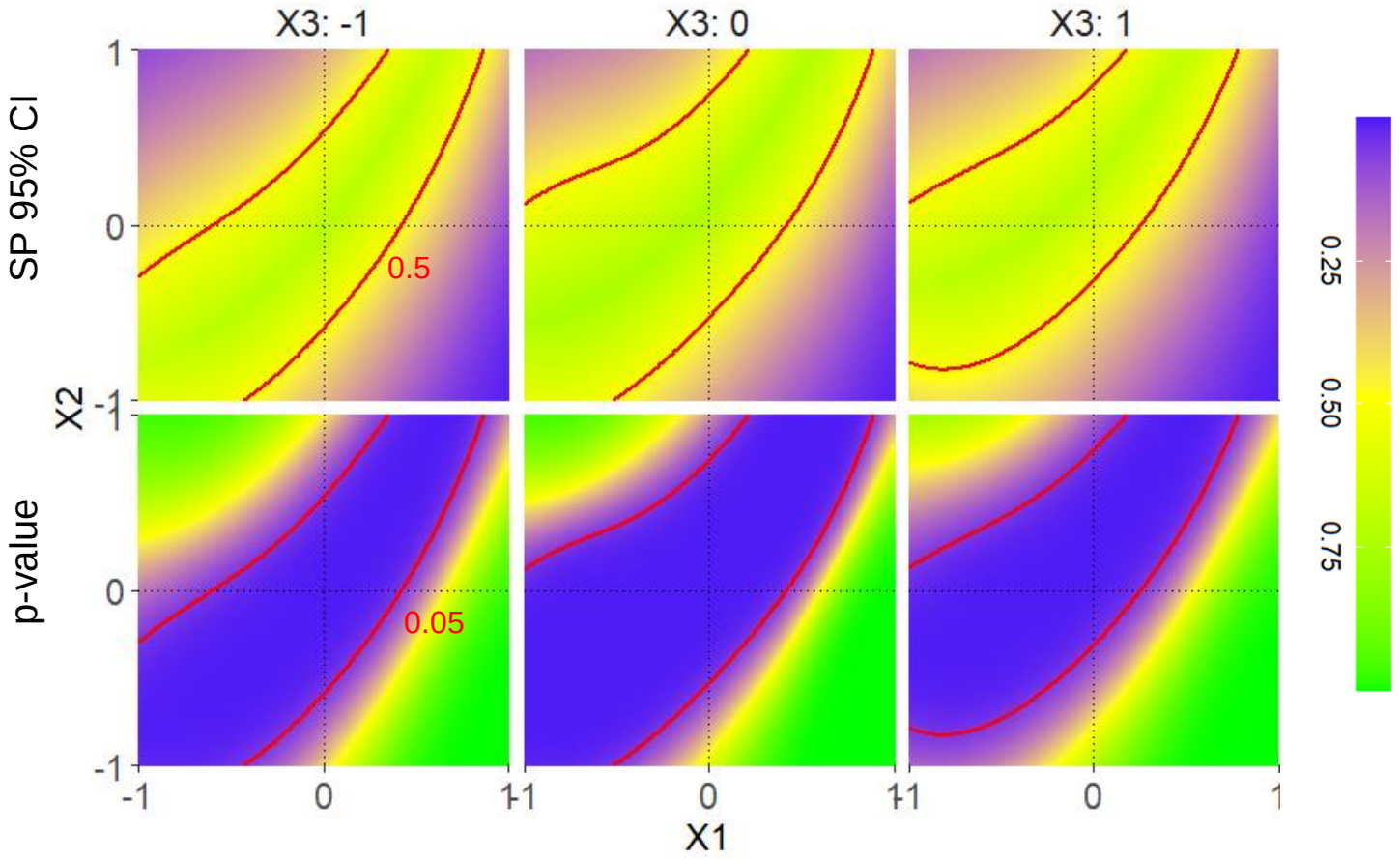
Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$

- Success Probabilities
  - straightforward to interpret
  - Values « well-distributed »
  - Intuitive with 50% cutoff value

- p-values
  - ‘uninterpretable’
  - Very skewed distribution
  - Not intuitive

Heatmap of TOST - Equivalence Region



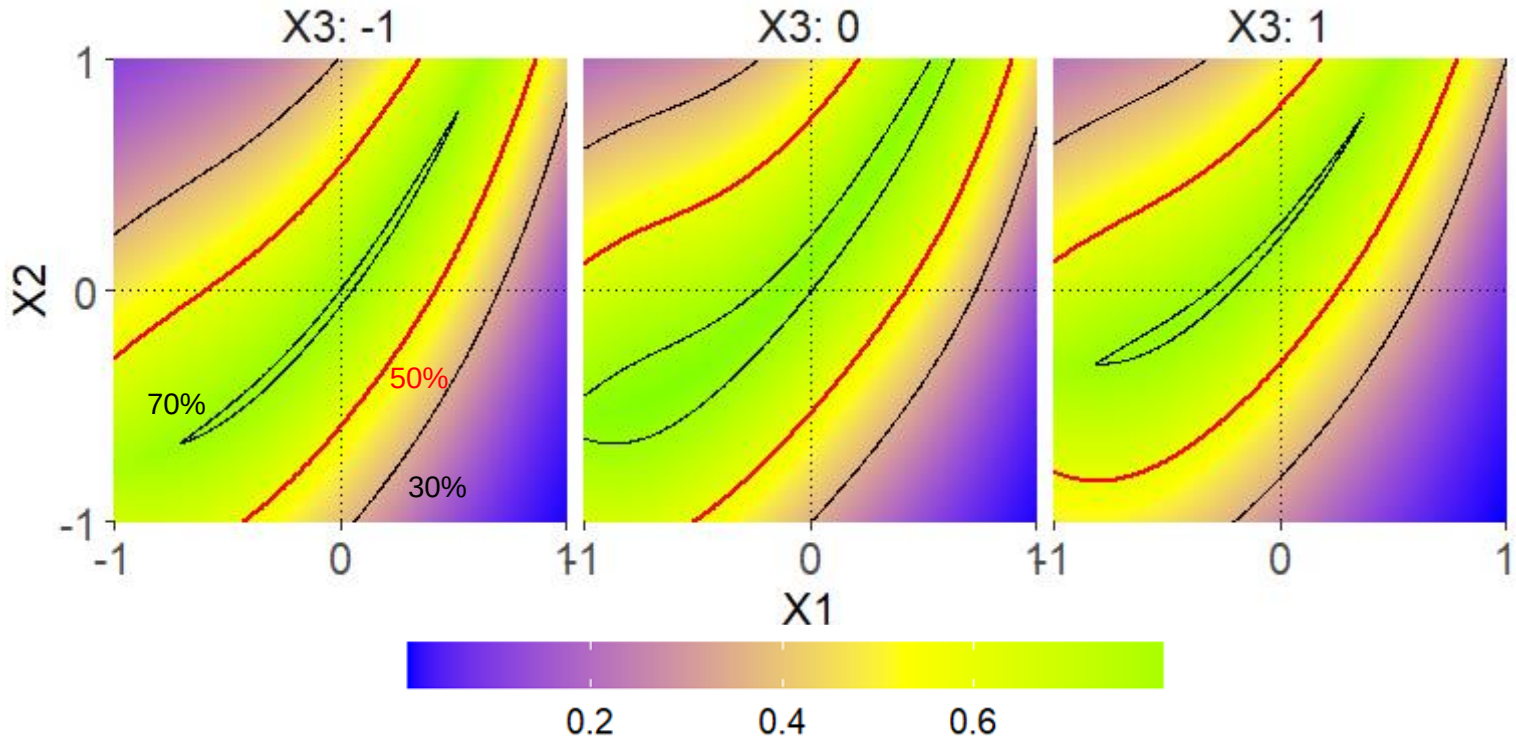
TOST in DOE by **Success Probabilities**

Consider a DOE with 3-X continuous variables

- Target = 100 ± 2
- $H_0: \mu_Y \notin [98, 102], H_1: \mu_Y \in [98, 102]$

- Any contour of Success Probabilities can be displayed
- 95% confident that at least 70% of the future products (in the equivalence region) will be compliant
- No need of sophisticated Bayesian techniques

Heatmap of TOST
Success Probabilities (SP) Confidence Bound (target > 0.5)



When you bike, do you mainly use the front break or the rear one ?

Front brake

Success Probabilities, Bayesian

Rear brake

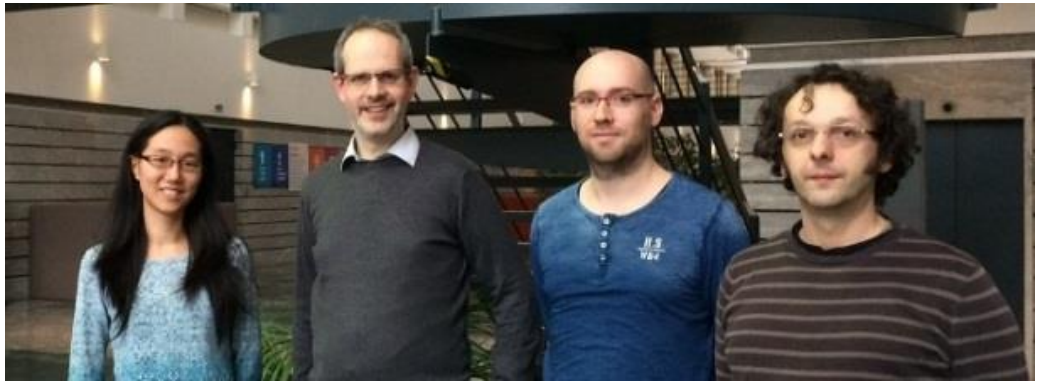
Frequentist CI for mean p-values

Majority of people mainly use the rear brake, because we learnt it.
We actually have to use the front brake !

GSK

In Memory of Stéphane Laurent, PhD

- All our thoughts go to Stéphane tragically passed away on August 28, 2024
- Stéphane was writing scientific software for GSK CMC Stats in Technical R&D, a function currently known as CMC Applied Data Sciences. Stéphane was a brilliant mathematician (PhD) and seasoned R package developer, having attained **legendary status** in reputed online scientific communities such as **Stack Overflow**, where he had earned 17 gold badges (!) and 133 silver ones for his contributions in R, R Shiny, and many other specialized topics. He authored and maintained more than 50 R packages published on CRAN. His personal blog offers a glimpse into the man's intellectual interests above and beyond R or software development and hints at the genius behind a generally reserved, socially atypical individual (https://laustep.github.io/stlahblog/).
- We had a lot of friendly and fruitful discussions on many topics (especially on (multiple)-prediction intervals, tolerance intervals or bridging studies). He gave valuable comments on this research topic and co-authored many of my talks.



Dan Lin, Walter Hoyer, Bernard Francq, Stéphane Laurent

Stéphane liked to quote:

“ Life is hard… Mathematics is harder ”

Last but not least

References

- Francq, Hoyer, Cartiaux, Kenett: A New Interpretation To The The T-Test By Tolerance Intervals and (Bayesian) Success Probability. (2024) (under review)
- Kenett, Francq: Helping reviewers assess statistical analysis: A case study from analytic methods. Analytical Science Advances (2022) ***
- Francq, Berger, Boachie: To Tolerate or To Agree: A Tutorial on Tolerance Intervals in Method Comparison Studies with BivRegBLS R Package. Statistics in Medicine (2020)
- Francq, Lin, Hoyer: Confidence and Prediction in Linear Mixed Models: Do Not Concatenate the Random Effects. Application in an Assay Qualification Study. Statistics in Biopharmaceutical research (2020)
- Francq, Lin, Hoyer. Confidence, Prediction and Tolerance in Linear Mixed Models. Statistics in Medicine (2019) ***
- Francq, Cartiaux. Delta Method and Bootstrap in Linear Mixed Models to Estimate a Proportion When no Event is Observed: Application to Intralesional Resection in Bone Tumor Surgery. Statistics in Medicine (2016)

Acknowledgment

Projects CMC Stat Team at GSK

Conflict of interest

This work was sponsored by GlaxoSmithKline Biologicals SA. BG Francq is employee of the GSK group of companies. RS Kenett is an employee of the KPA group and the Samuel Neaman Institute.

*** *Top 10% most downloaded papers published by Wiley*