A new interpretation to the t-test

By Tolerance Intervals and Success Probability (probability indexes)

Bernard G Francq, Ron S Kenett
April 2024

GSK

Introduction – Significance Crisis
Traditional null-hypothesis significance-testing…

- *1963: "no longer a sound of fruitful basis for statistical investigation" (Clark)*
- *1978: "radically defective as to be scientifically almost pointless" (Meehl)*
- *1978: "should be eliminated; it is harmful" (Carver)*
- *1987: "despite two decades of attacks, the mystifying doctrine of null hypothesis is still today the Bible" (Gigerenzer and Murray)*
- *1994: "hypothesis testing does not tell us what we want to know… out of desperation, we nevertheless believe that it does" (Cohen)*
- *2003: "null hypothesis testing can actually impede scientific progress" (Kirk)*

Mark Burgman (Imperial College London)
What should applied science journal editors do about statistical controversies?

The debate is quite 'popular' nowadays

- *2016: The **ASA statement** on p-values: context, process, and purposes (Wasserstein and Lazar, The American Statistician)*
- *2018: Statistical Inference as Severe Testing: How to get beyond the Statistics Wars (Mayo)*
- *2019: Moving to a world beyond "p < 0.05" (Wasserstein et al., The American Statistician)*
- *2019: valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values (Greenland, The American Statistician)*
- *2019: Scientists rise up against statistical significance (Amrhein et al., Nature)*
- *2020: "To claim a result to be highly significant, or even just significant, sounds like enthusiastic endorsement, whereas to describe a result as insignificant is surely dismissive" (Sir David Cox,* Annu. Rev. Stat. Appl.*)*

GSK

Medical Research: p-values and confidence intervals (CIs)

p-values and CIs are common in medical research and requested by most of top medical journals

Réduction significative de la mortalité de 17%
HR = 0,83 [0,65 – 1,06], p<0,01

Critical analysis of treatments for COVID-19
(Analyse critique des traitements de la COVID-19)
Youtube video (at 31:34)

HCQ is effective for COVID-19 when used early: real-time meta analysis of 205 studies

Corpus ID: 231610073, Published 2021

• HCQ is effective for COVID-19. The probability that an ineffective treatment generated results as positive as the 205 studies to date is estimated to be 1 in 28 quadrillion (p = 0.0000000000000036).

• Studies from North America are 3.7 times more likely to report negative results than studies from the rest of the world combined, p = 0.00000022.

Researchers proud
to show tiny p-values

GSK

Significance Crisis: our contribution

"A new interpretation to the t-test by tolerance intervals and (Bayesian) Success Probability"

(Francq and Kenett, 2024, under review)

The ASA Statement and related papers propose as alternative solutions:

- Credible intervals
- Prediction intervals
- Compatibility intervals
- s-values
- b-values, d-values
- Probability indexes (degree of overlap)
- CPM (Comparative Probability Metrics)
- …

Our contribution

✓ Success Probability (SP)

✓ A unified framework based on tolerance intervals

✓ Show the one-to-one function with the (frequentist) p-value

✓ Assess the uncertainty of Probability Indexes, CPM, b-value, d-value

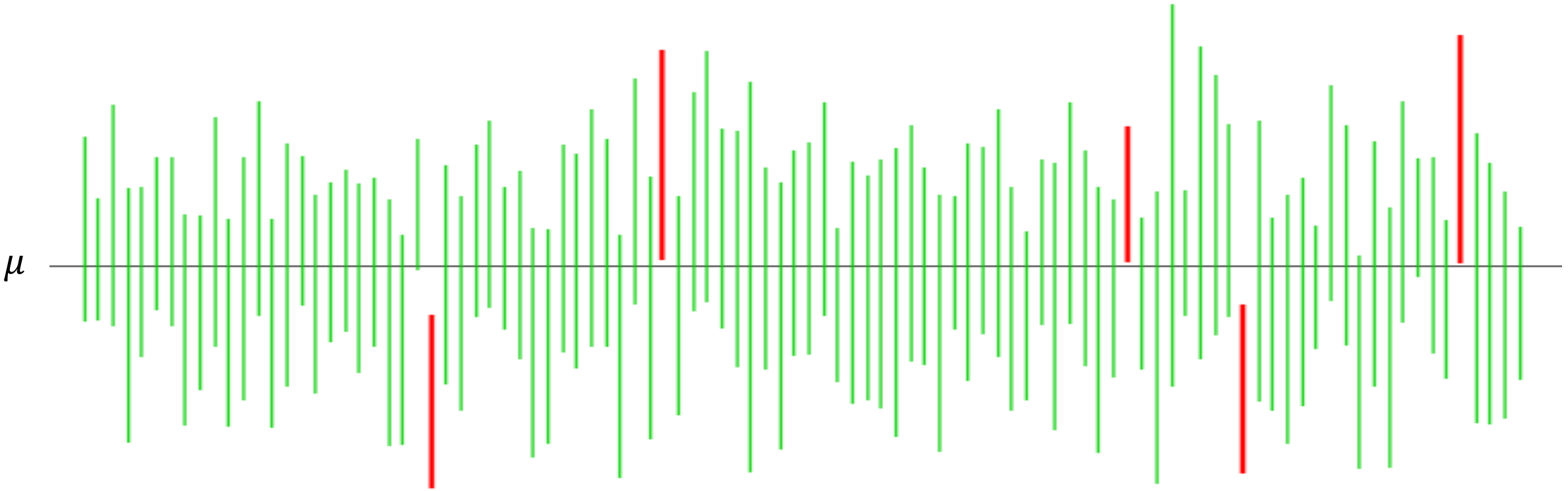
GSK

Statistical Intervals

- Confidence
- Prediction
- Content Tolerance
  - 2-sided
  - 1-sided

Confidence Interval concept

100 simulated 95% CI for the mean μ



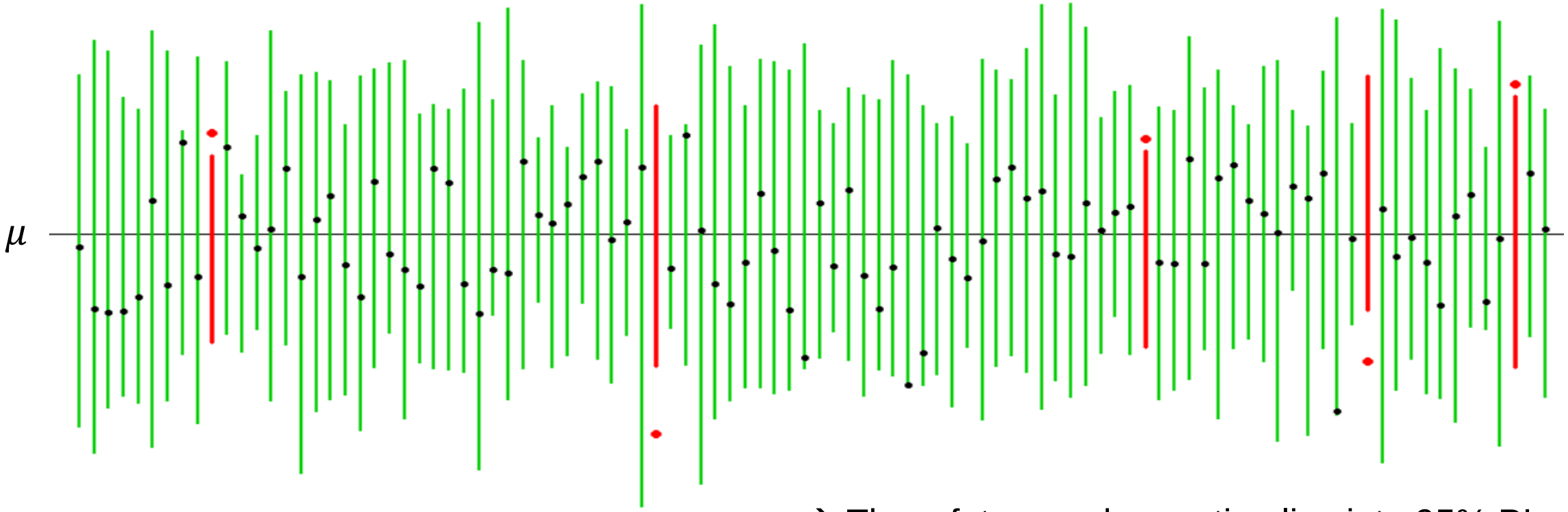
→ The true value, μ, lies in 95% of the CIs

Note: in Bayesian statistics, credible intervals are commonly used

GSK

Prediction Interval concept

100 simulated 95% PI for a future observation



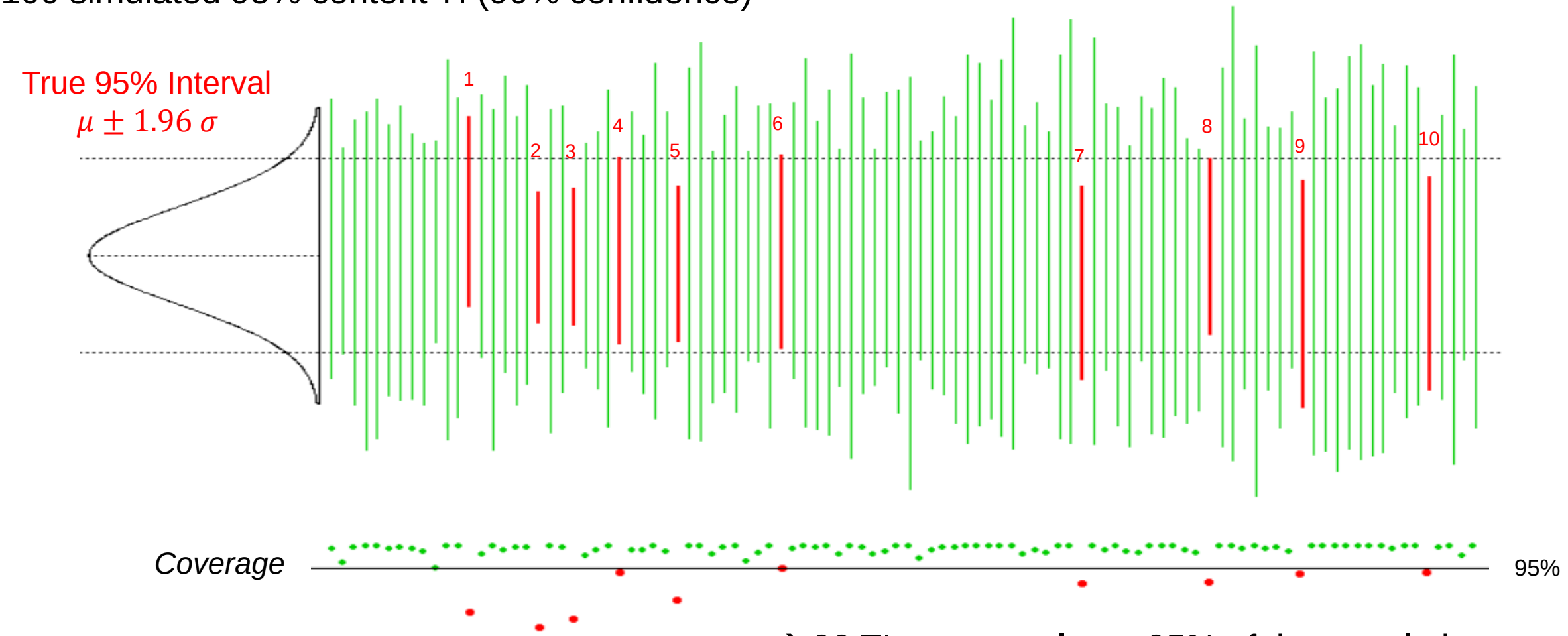
→ The « future » observation lies into 95% PIs

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

GSK

Content Tolerance Interval (type II) concept

100 simulated 95% content TI (90% confidence)



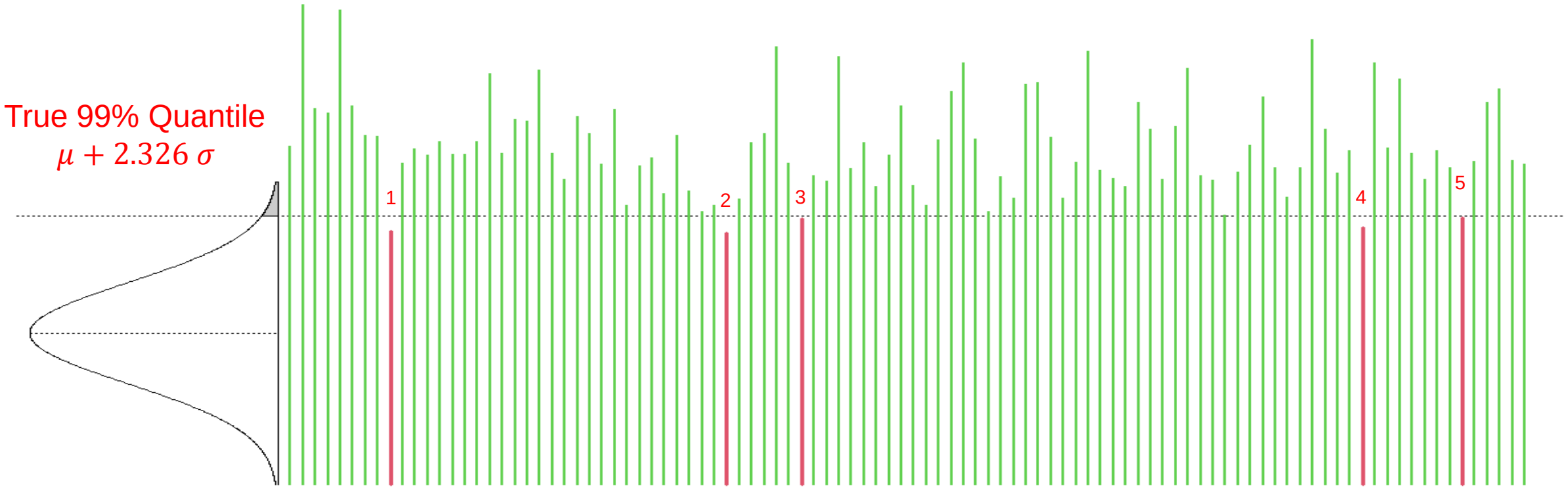
→ 90 TIs cover **at least** 95% of the population

→ 10 TIs cover **at most** 95% of the population

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)

SP is 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 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

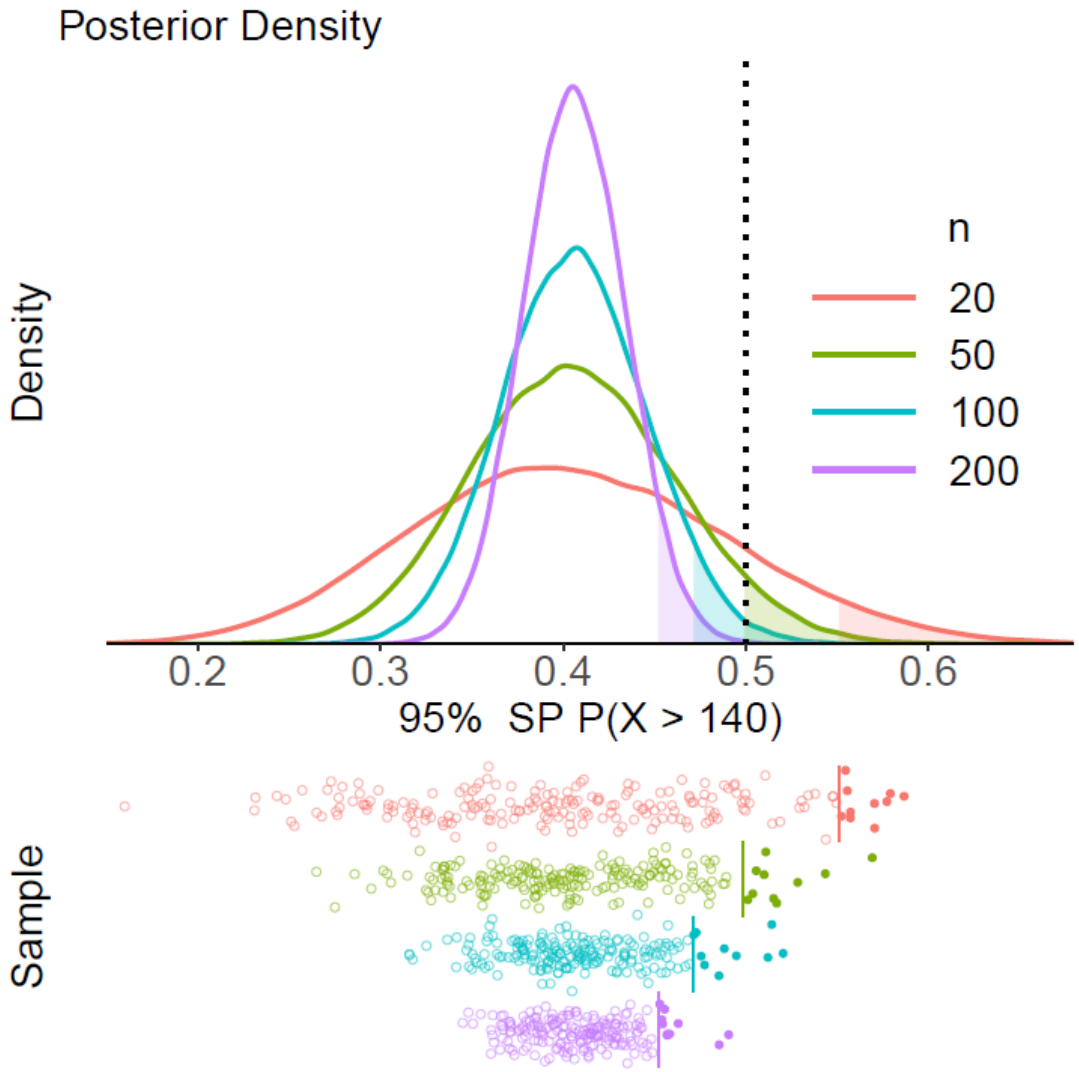
The SP interpretation is straightforward even for big sample sizes (eg $n = 10^3$)

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

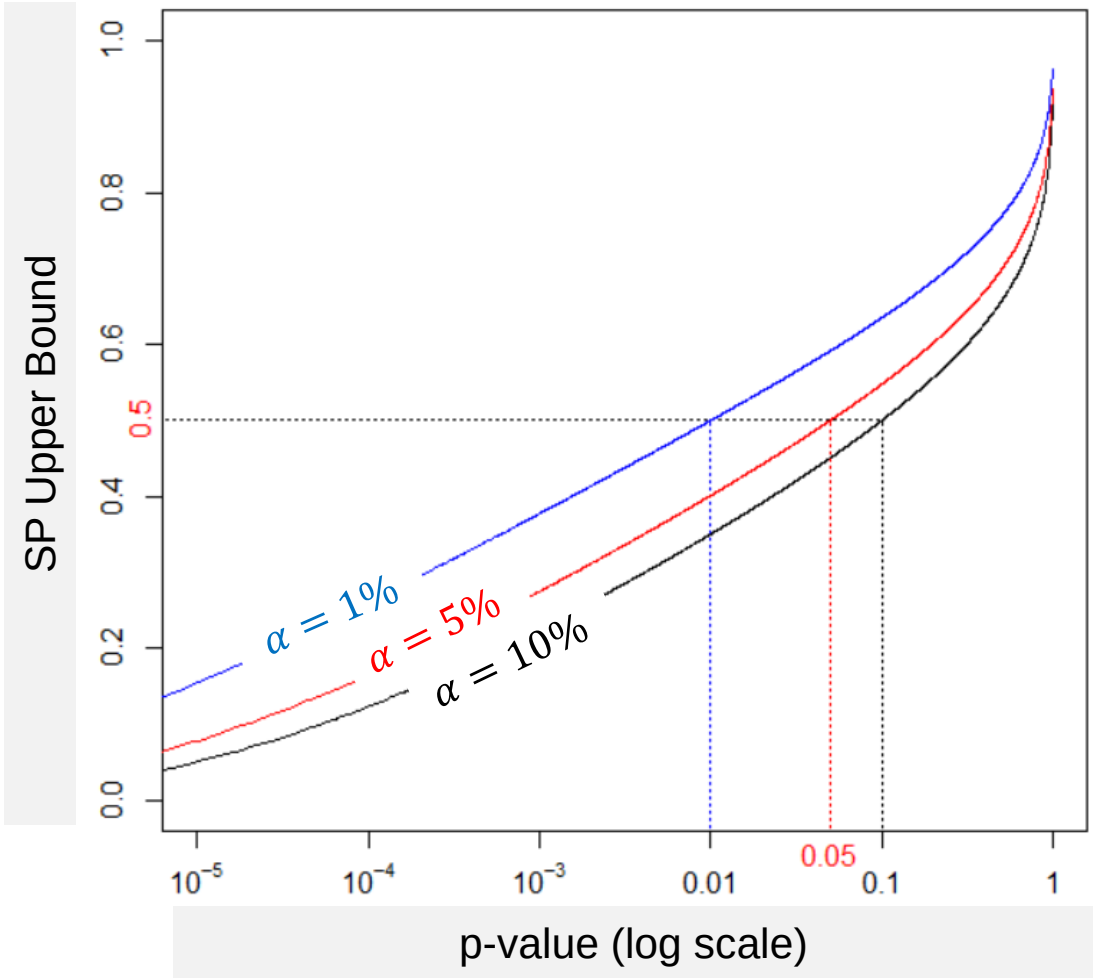
1-sample t-test **Bayesian** synthetic examples



	$H_0 : \mu = 140$ $H_1 : \mu \neq 140$		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

Main 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

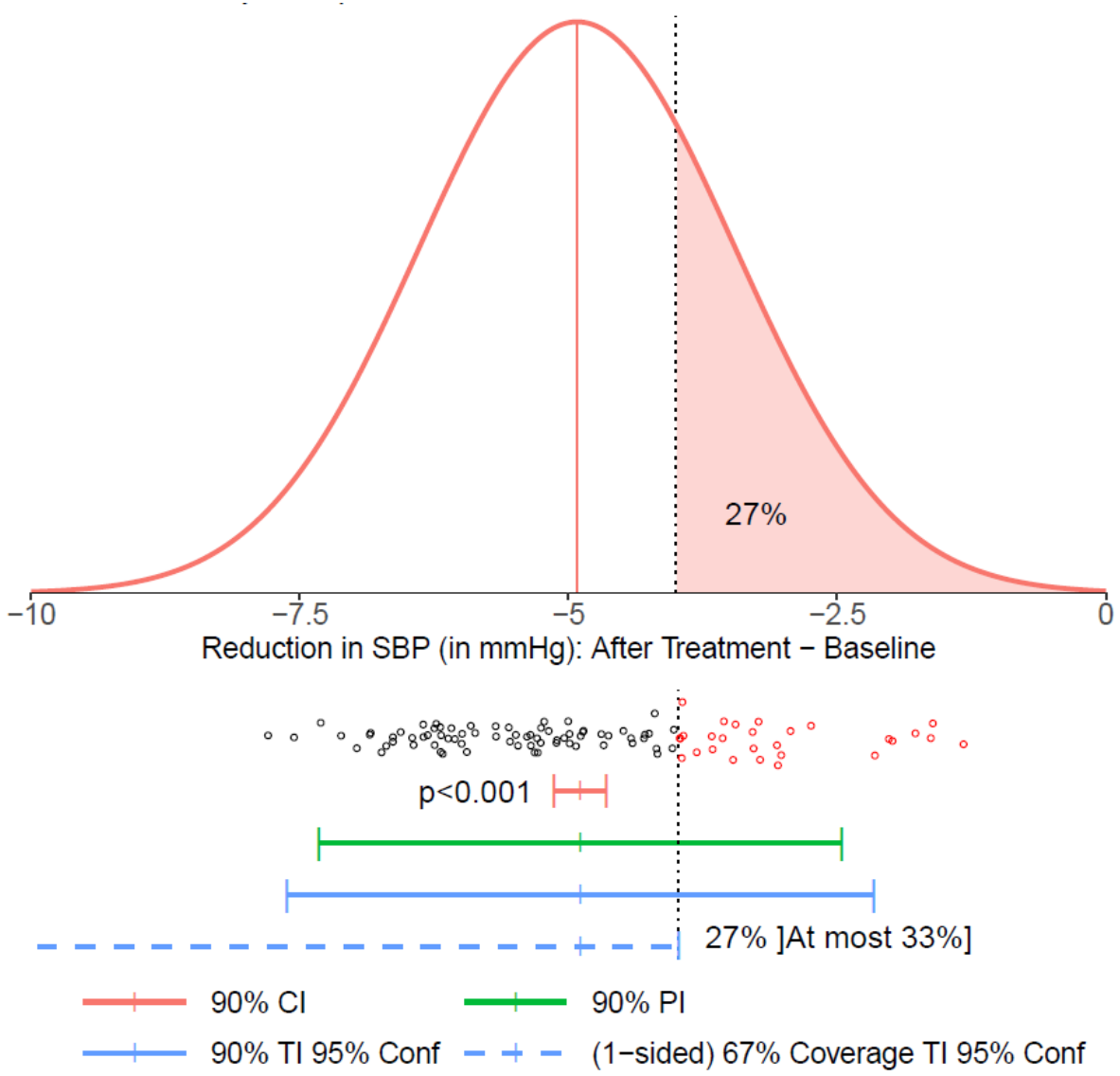
paired t-test

Paired t-test

Toy Example: SBP (mmHg)

- ✓ $n = 100$
- ✓ *Difference = Treatment – Baseline*
- ✓ H_1*: at least 4 units decrease*

- ✓ $P(D < -4) = 74\%$ [at least 67%[
- ✓ $P(D > -4) = 27\%$ [at most 33%[



GSK

2-sample t-test

2-sample t-test: synthetic examples

n	Mean Diff.	Pooled SD	Mean Difference 95% CI	$H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ p-value	# Head	Success Probabilities $P(D < 0)$	$P(D > 0)$
50	0.12	1.41	[-0.27, 0.52]	p=0.54	0.9	53.5	46.5
100	0.12	1.41	[-0.15, 0.40]	p=0.38	1.4	53.5	46.5
500	0.12	1.41	[0, 0.25]	p=0.05	4.3	53.5	46.5
1000	0.12	1.41	[0.04, 0.21]	p=0.006	7.5	53.5	46.5
5000	0.12	1.41	[0.08, 0.16]	p<.001	31	53.5	46.5

p-value collapse

s-value soar

b-value *

d-value *

How to add the 95% CI ?

✓ Reverse the **Tolerance Interval for a Difference** !

✓ Well-established methodology in non-clinical statistics, manufacturing and engineering

GSK

* Eugene Demidenko. The p-value you can't buy. *The American Statistician* 2016; 70: 33 – 38.

2-sample t-test: synthetic examples

n	Mean Difference 95% CI	$H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ p-value	# Head	Success Probability $P(D < 0)$
50	[-0.27, 0.52]	p=0.54	0.9	53.5 [42.5, 64.2]%
100	[-0.15, 0.40]	p=0.38	1.4	53.5 [45.7, 61.2]%
500	[0, 0.25]	p=0.05	4.3	53.5 [50.0, 57.0]%
1000	[0.04, 0.21]	p=0.006	7.5	53.5 [51.0, 56.0]%
5000	[0.08, 0.16]	p<.001	31	53.5 [**52.4**, 54.6]%

p-value collapse

s-value soar

Borderline test
- p-value = 5%
- CI bound = 0
- SP bound = 50%

SP interpretation in 2-arms clinical trials ($n = 5000$)

× *At least 52.4% patients are expected to be better with B (than A) **

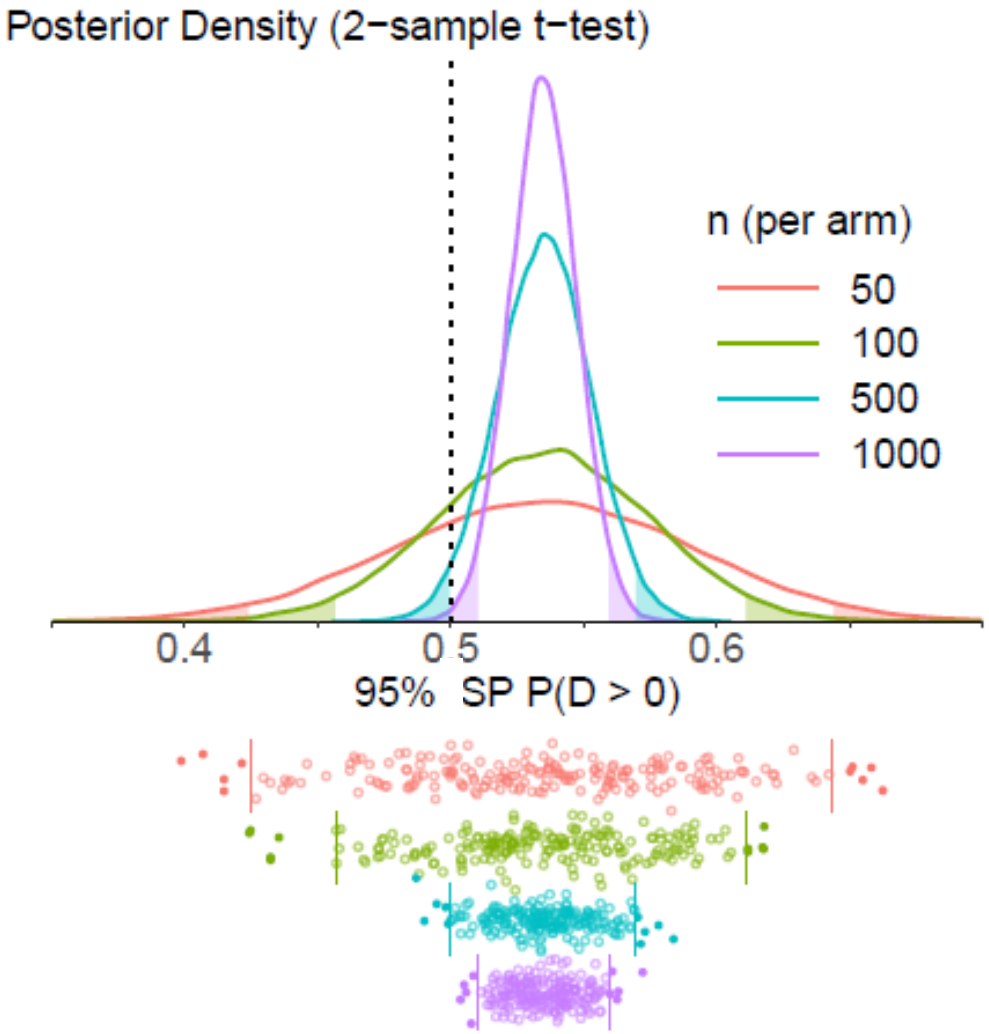
✓ *At least 52.4% patients are expected to get a better clinical outcome with treatment B compared to patients under A*

✓ *By comparing A and B on different patients, B is expected to be better in at least 52.4% of the comparisons*

GSK

* S Greenland *et al*. On Causal Inferences for **Personalized Medicine**: How **Hidden Causal Assumptions Led to Erroneous Causal Claims About the D-Value**. *The American Statistician* 2020; 70: 243 - 248.

2-sample t-test: Bayesian synthetic examples



	$H_0 : \mu_D = 0$ $H_1 : \mu_D \neq 0$		Success Probability (95% confidence) $P(D > 0)$	
n	$\bar{D}$	S_p	Frequentist	Bayesian
50	0.12	1	53.5 [42.5, 64.2]%	53.4 [42.4, 64.3]%
100	0.12	1	53.5 [45.7, 61.2]%	53.5 [45.8, 61.2]%
500	0.12	1	53.5 [50.0, 57.0]%	53.5 [50.0, 57.0]%
1000	0.12	1	53.5 [51.0, 56.0]%	53.5 [51.0, 55.9]%
5000	0.12	1	53.5 [52.4, 54.6]%	53.5 [52.4, 54.6]%

GSK

Demystify a (statistical) urban legend

How do you interpret a slope ?

A slope should only be interpreted for « comparison » of different patients

*→ Interpret coefficients as comparisons, not effects **

→ Like the Tolerance Interval for Differences

GSK

* A Gelman, J Hill, A Vehtari. Regression and Other Stories. *Published by Cambridge University Press in 2020.*

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

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)
- 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

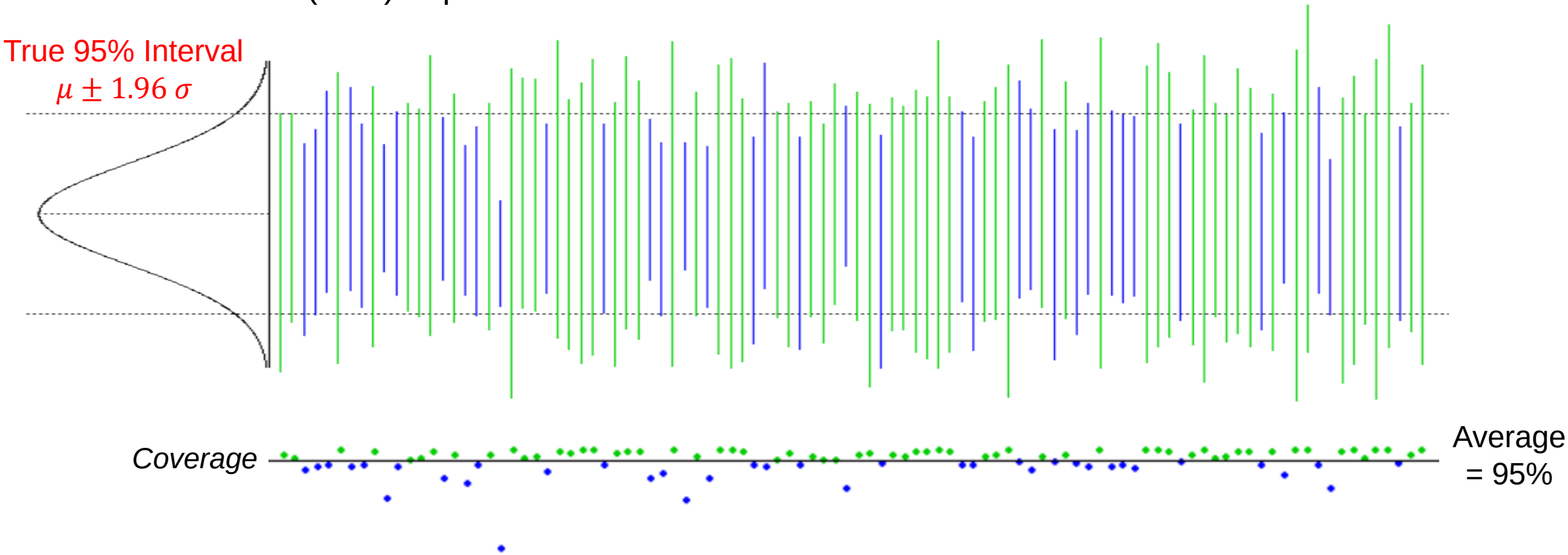
This work was sponsored by GlaxoSmithKline Biologicals SA. BG Francq is employee of the GSK group of companies. RS Kenett is an employee of the KPA group and the Samuel Neaman Institute.

GSK

Back Up slides

Expectation Tolerance Interval (type I) concept

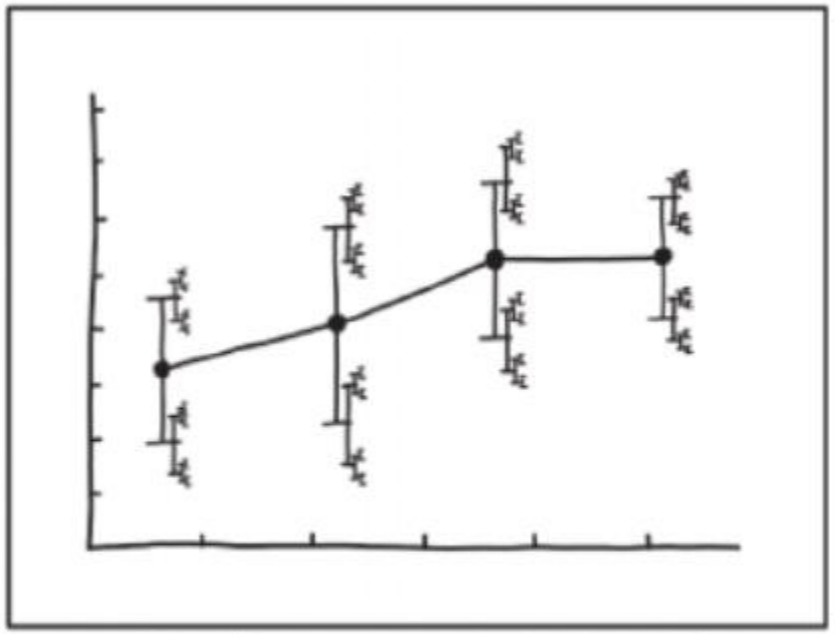
100 simulated 95% (beta)-expectation TI



→ Expectation TI covers 95% of the population, on average

GSK

Confidence Interval of Confidence Interval



https://www.explainxkcd.com/ Error Bars

- Will the PI contain less or more than 95% of future observations?

→ Some researchers calculate the 95% CI for each bound of the 95% PI

- Calculating the CI of a CI is awkward, confusing, misleading
- Unfortunately, widely used in method comparison studies (bridging studies) with Bland-Altman plot (agreement interval)

→ Use the Tolerance Interval type II

To tolerate or to agree: A tutorial on tolerance intervals in method comparison studies with BivRegBLS R Package
BG Francq, M Berger, C Boachie, *Statistics in Medicine*, 2020.

1-sample t-test
p-value, s-value

$H_0: \mu = 140$
$H_1: \mu < 140$

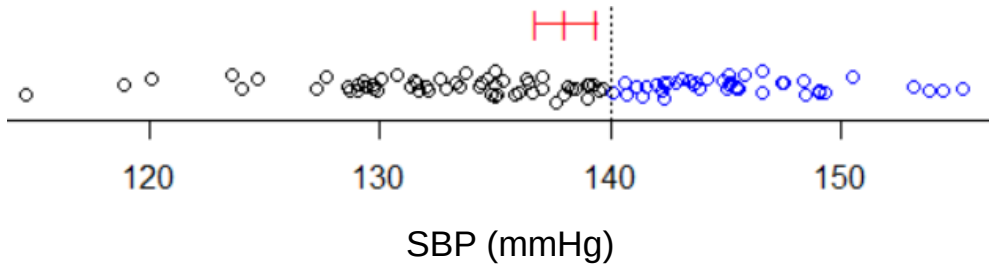
What about the p-value?

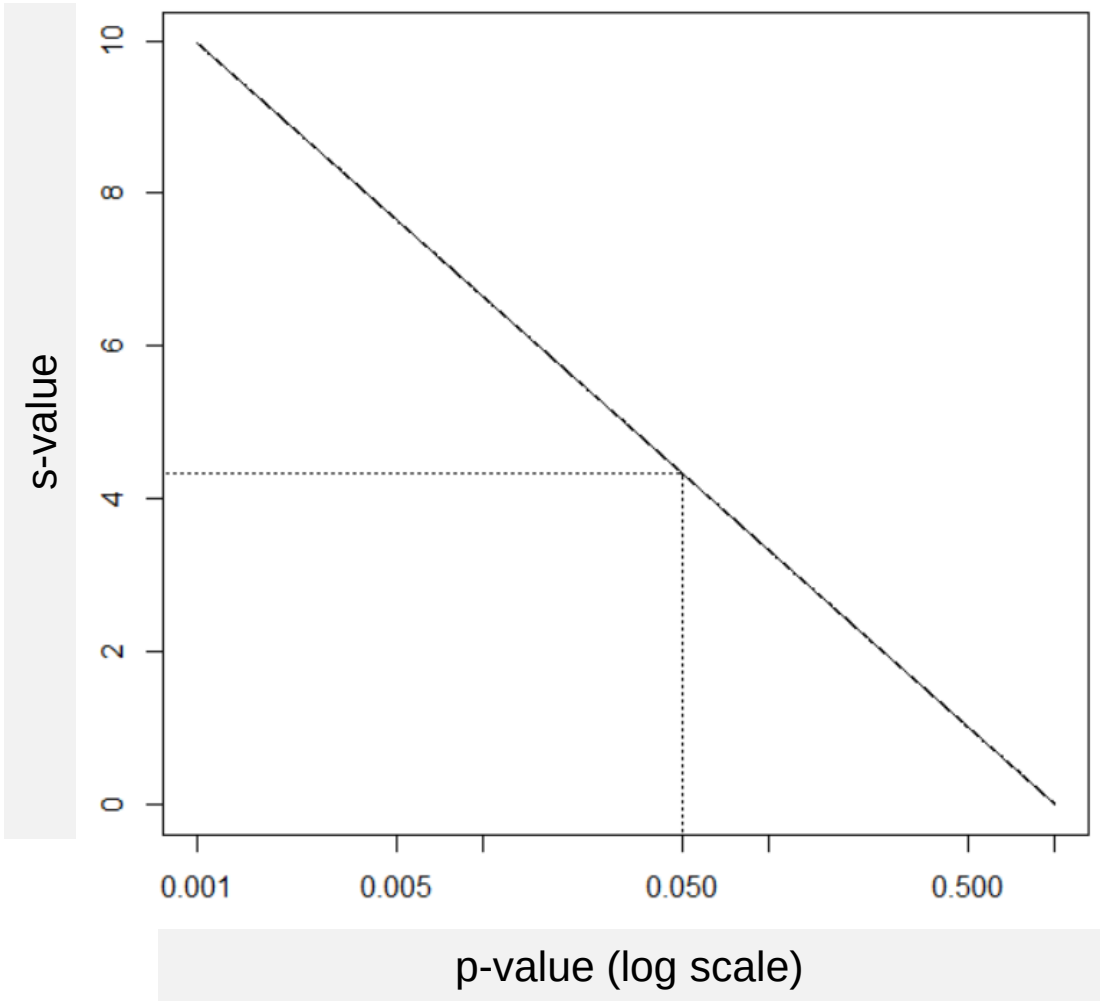
p-value = 0.0098 (significant at $\alpha = 0.05$)

What about the s-value?

s-value = -$\log_2$(p-value) = 6.7

The p-value is equivalent of obtaining more than 6 heads in a row when tossing a fair coin





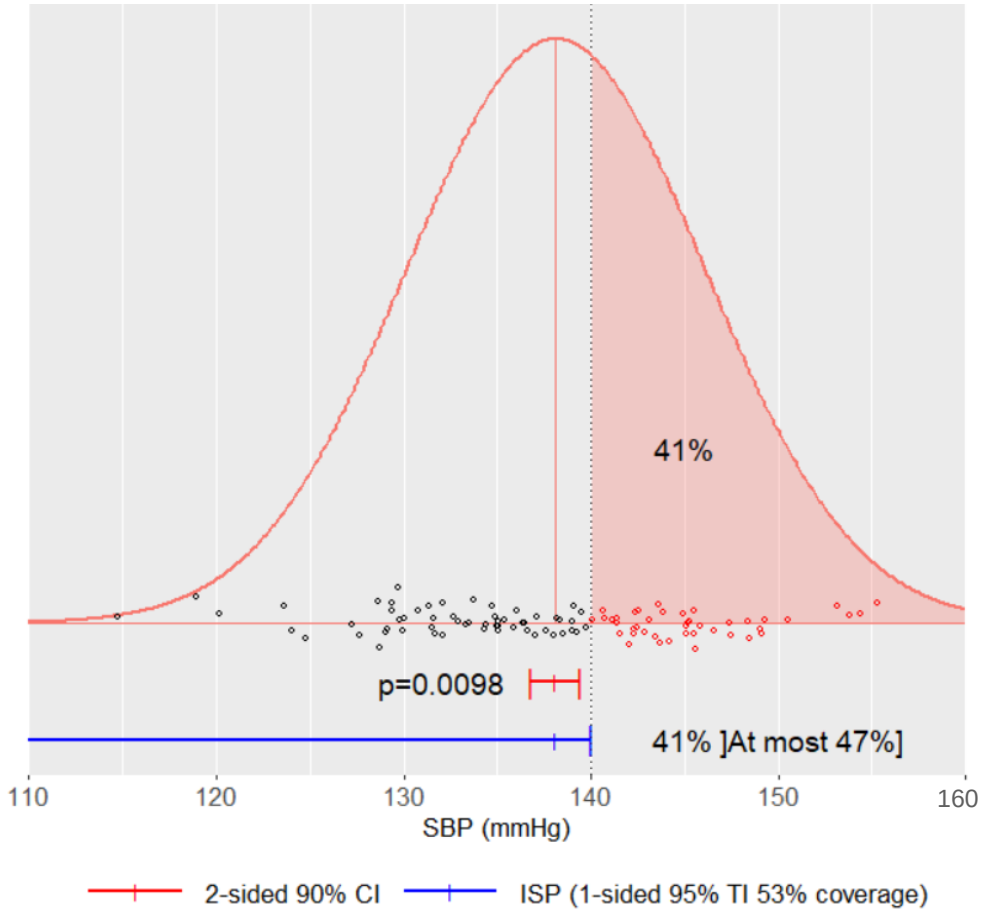
1-sample t-test synthetic examples

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*



GSK

Confidence, Prediction and Tolerance

90% CI		90% PI		98% TI (95% Conf)	
13.92	15.09	11.27	17.74	−∞	19.61

Confidence Interval = CI

- The interpretation is usually confusing and holds only for the average

Prediction Interval = PI

- A future product is expected to be between 11.27 and 17.74 (with 90% confidence)

(β)-expectation Tolerance Interval = TI type I

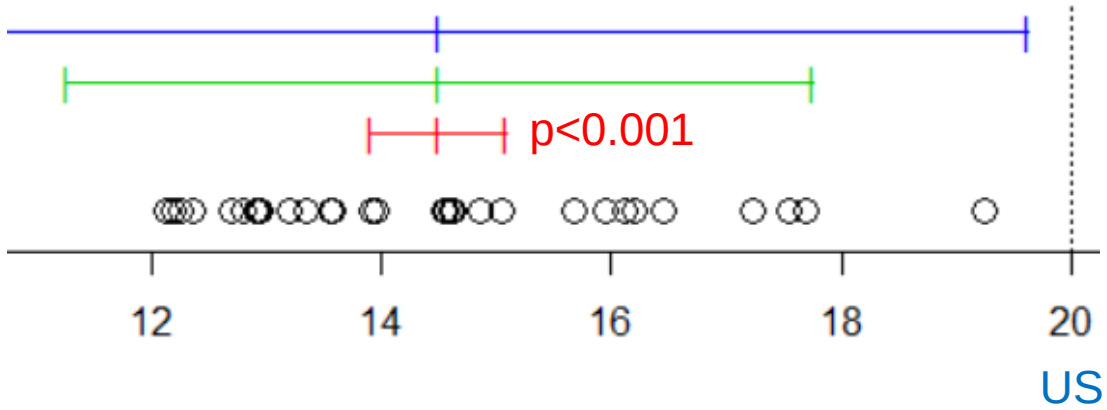
- 90% of the future products are expected to be between 11.27 and 17.74 (on average)

(βγ)-content Tolerance Interval = TI type II

- At least 98% of the future products will be lower than 19.61 (with 95% confidence)

Remarks

- The interpretation of PI and TI is similar in frequentist or Bayesian
- Their interpretation remains identical with/without the log transformation



→ TI < USL
→ The POOS is lower than 2%
→ Smart Risk decision: Go ☺

Can we calculate the POOS?

TOST

Two One-Sided (t)-Tests

TOST: synthetic examples

$H_0 : \mu \notin [11.5, 13]$			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

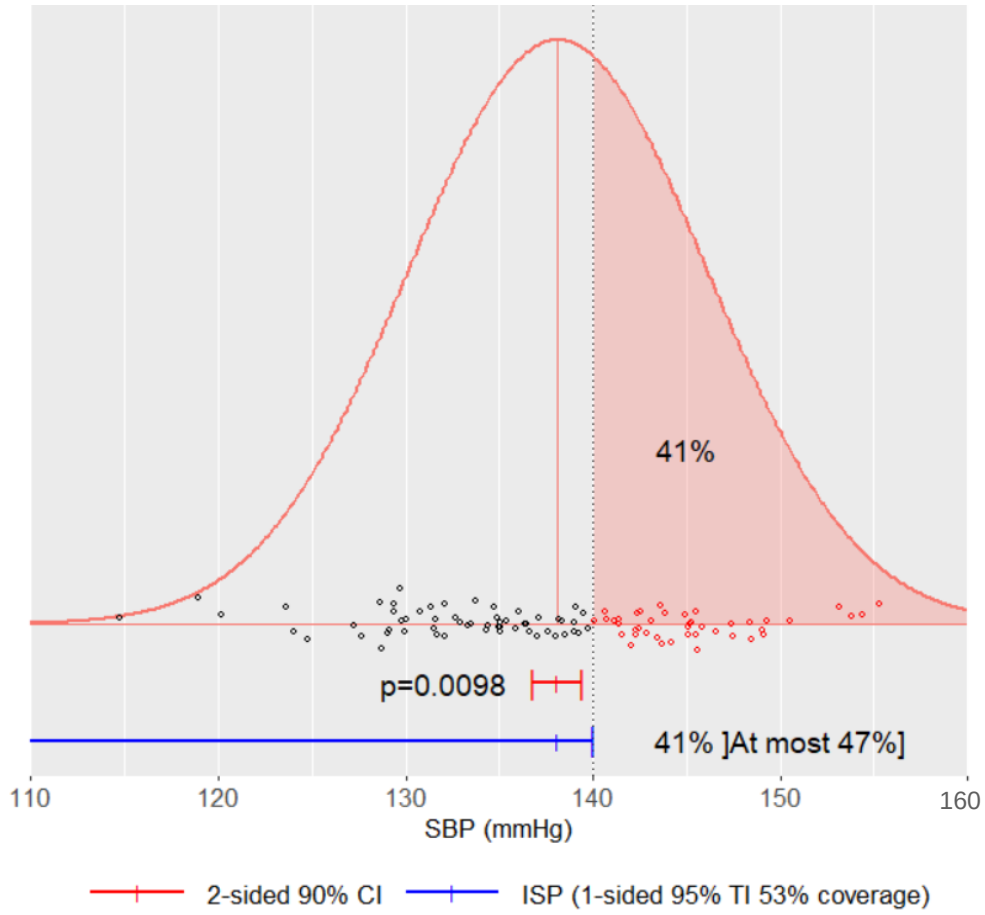
ISP and measurement error

The SBP is certainly measured with some measurement error

→ What is the probability for the 'true' SBP to be > 140 ?
('true' = without measurement error)

Define more precisely, clarify the desired ISP:

- *$P(X_T > 140)$* where X_T is the 'true' value of the next patient
- *$P(X > 140)$* where X is the SBP measured on the next patient



ISP and measurement error: use replicates

- *n = 50* patients, each measured 3 times
- Mixed model by REML method
- Between variance and within variances are the 2 key parameters

```
Individual fixef effect estimates:
             Estimate Std. Error    Lower  Upper
(Intercept) 137.0483  0.9313913 135.1766 138.92

Variance component estimates:
  patient     error
40.619634  8.264555
```

Toy example in R

Covariance matrix variance components

```
            patient      error
patient 77.5986901 -0.4563254
error   -0.4563254  1.3689761
```

ISP and measurement error

The ISP is assessed by the z-score and by using the corresponding variance components.
Example for *P(X > 140)* with the total variance

$$P(X > 140) = 1 - \phi\left(z = \frac{140 - \hat{\mu}}{\hat{\sigma}_T}\right)$$

The lower and/or upper bounds can be obtained by the delta method on the z-score *

$$CI\ \{P(X > 140)\} = 1 - \phi\left(\frac{140 - \hat{\mu}}{\hat{\sigma}_T} \pm z_{0.95}\sqrt{var(z)}\right)$$

If needed (especially for small sample sizes), $z_{0.95}$ can be replaced by the t-distribution with the DF as:

- (Kenward-Roger)
- (Satterthwaite)
- ✓ Francq et al. **

P(X_T > 140) is assessed with the between variance

* 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. BG Francq, O Cartiaux. *Statistics in Medicine* (2016)

** Confidence, Prediction and Tolerance in Linear Mixed Models BG Francq, D Lin, W Hoyer. *Statistics in Medicine (*2019)

ISP, measurement error and Smart Risk

```
Individual fixef effect estimates:
             Estimate Std. Error    Lower  Upper
(Intercept) 137.0483  0.9313913 135.1766 138.92

Variance component estimates:
  patient     error
40.619634  8.264555
```

Toy example in R

Covariance matrix variance components

```
                patient      error
patient 77.5986901 -0.4563254
error   -0.4563254  1.3689761
```

- *P(X > 140) = 33.6]44.0]%* *At most 44% of future patients will have their SBP measured > 140*
- *P(X_T > 140) = 32.2]43.6]%* *At most 43.6% of future patients will have their 'true' SBP > 140*

Smart Risk

What matters is
the probability that a future product has its true (underlying) value outside the spec