p-value, s-value, b-value, d-value,… *what else ?*

The probability of being out of specification !

An Ode to Tolerance

Bernard G Francq, Ron S Kenett
NCS conference 2022

GSK

Introduction – Medical Research

Which of the following sentences would you prefer your surgeon to tell you?

PS: If you struggle to read this text until the end, then you better check your eyes with your ophthalmologist

- Surgery A is significantly better than B, on average ($p<0.001$)
- *If you undergo surgery A, and your friend surgery B,* then there is **at most 20%** chance that you end up with a better clinical outcome
- *If you undergo surgery A instead of B,* there is **at most 30%** chance that you end up with a better clinical outcome

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.*)*

Significance Crisis: our contribution

“Individual Success Probability (ISP): Beyond the t-test and p-values” (2022, under review)

Our work

- ***Smart Risk in CMC and Non-Clinical Statistics***
- *Compare the p-values, s-values, b-values, d-values, Probability Indexes, Generalized Pairwise Comparison*
- *Assess the uncertainty of the b-value and Probability Indexes*
- *Propose the ISP concept by ‘reversing’ the tolerance interval concept (not well known in clinical statistics)*
- *Show the one-to-one function p-value - ISP*

GSK

Overview of presentation

Smart Risk

Medical Research: misinterpretations

1 sample t-test

- Significance tests: Confidence Interval (CI), p-value, s-value
- Prediction Interval (PI)
- Tolerance Interval (TI)
- Individual Success Probability (ISP)
- p-value or ISP

Measurement error

2-arms clinical trials (B-value)

Smart Risk in CMC Statistics
Probability Out Of Specification (POOS)

Question

How to quantify the probability of being out of specification (POOS)?
What is the probability to be greater than the upper limit? ie a future lot, batch, product,…

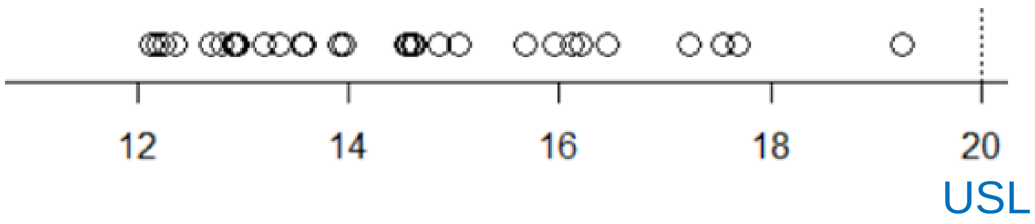
Parameters

- Upper Specification Limit (USL) = 20
- Type I Error = $\alpha = 0.05 = 5\%$
- Maximal *tolerated* OOS = 2%
- Sample Size = $n = 30$

Assumptions

Normality, independence,…

Data Visualisation



What is the proportion of products greater than 20?
Naive approach: P = 0/30 = 0%
but the 95% classical CI assumptions fails: [0,0]

$$P \pm z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

Smart Risk in CMC Statistics
Probability Out Of Specification (POOS)

What could be the proportion of future products greater than 20?

Solution (when $P = 0$)
Use the „rule of 3“: $3/n$
The 95% CI is then [0, 0.1] → at most 10% of products OOS!

Other solutions (in R) ($P = 0$ or $P \neq 0$)

	method	x	n	mean	lower	upper
1	agresti-coull	0	30	0.000	-0.021	0.135
2	asymptotic	0	30	0.000	0.000	0.000
3	bayes	0	30	0.016	0.000	0.062
4	cloglog	0	30	0.000	0.000	0.116
5	exact	0	30	0.000	0.000	0.116
6	logit	0	30	0.000	0.000	0.116
7	probit	0	30	0.000	0.000	0.116
8	profile	0	30	0.000	0.000	0.104
9	lrt	0	30	0.000	0.000	0.062
10	prop.test	0	30	0.000	0.000	0.141
11	wilson	0	30	0.000	0.000	0.114

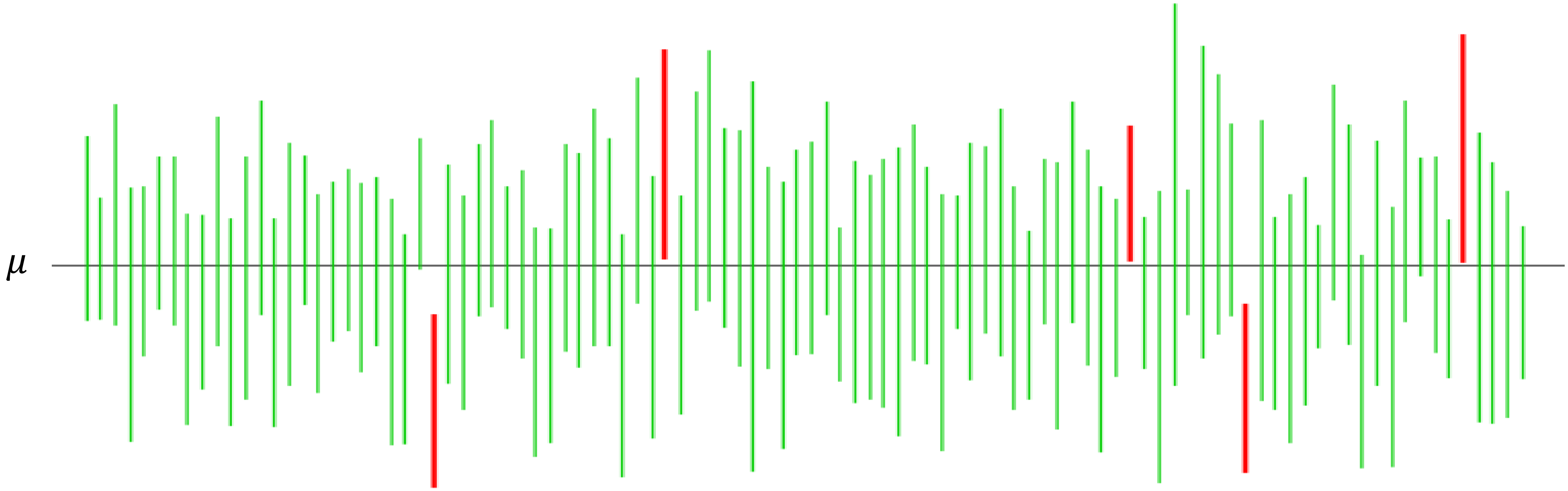
```
library(binom)

USL = 20
x = rnorm(30)
x = (x – mean(x)) / sd(x) * 1.87 + 14.51
binom.confint(x = sum(x > USL), n = length(x), conf.level = 0.95)
```

Interpretation
The upper bounds are all close to 10%
The maximal ‚threshold‘ was 2% → Stop!

Can we do better?
Calculate Statistical Intervals…

Confidence Interval concept

100 simulated 95% CI for the mean μ

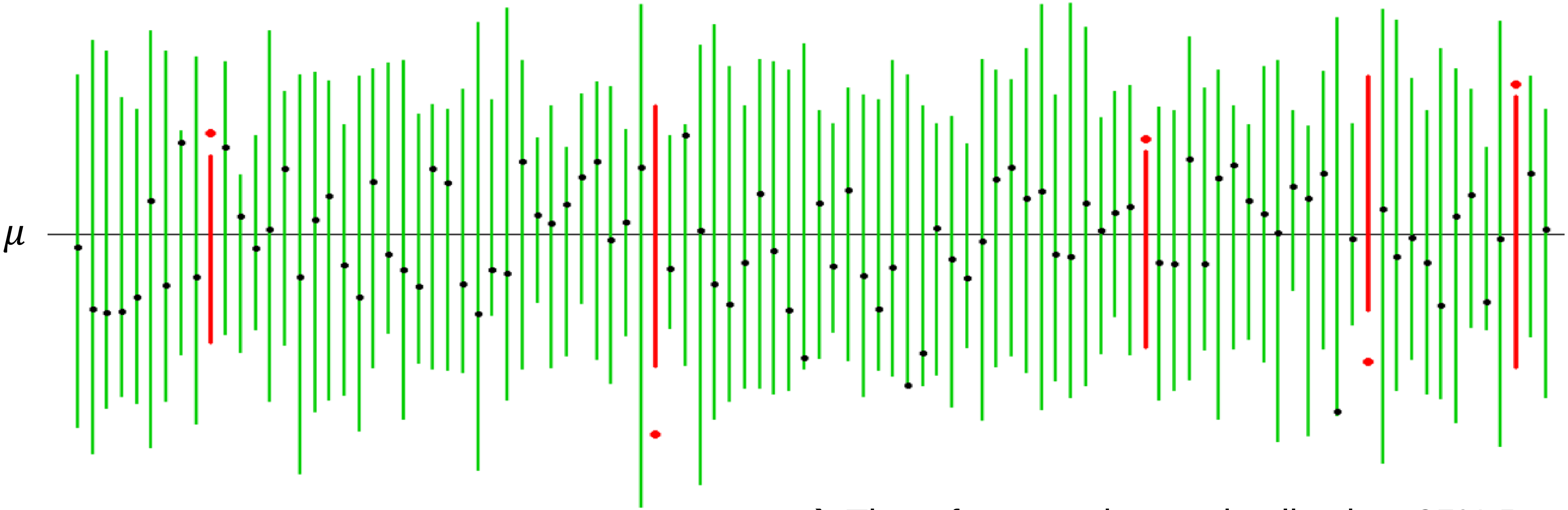


→ The true value, μ, lies in 95% of the CIs

Note: in Bayesian statistics, credible intervals are usually used

Prediction Interval concept

100 simulated 95% PI for a future observation

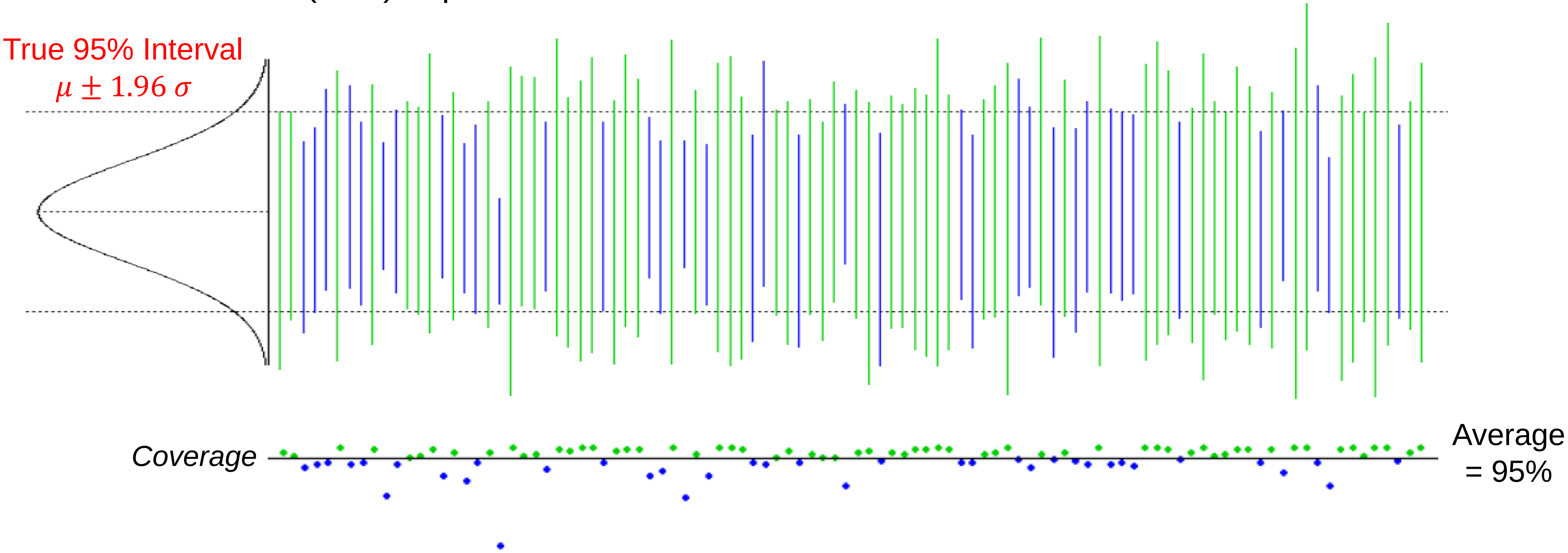


→ The « future » observation lies into 95% PIs

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

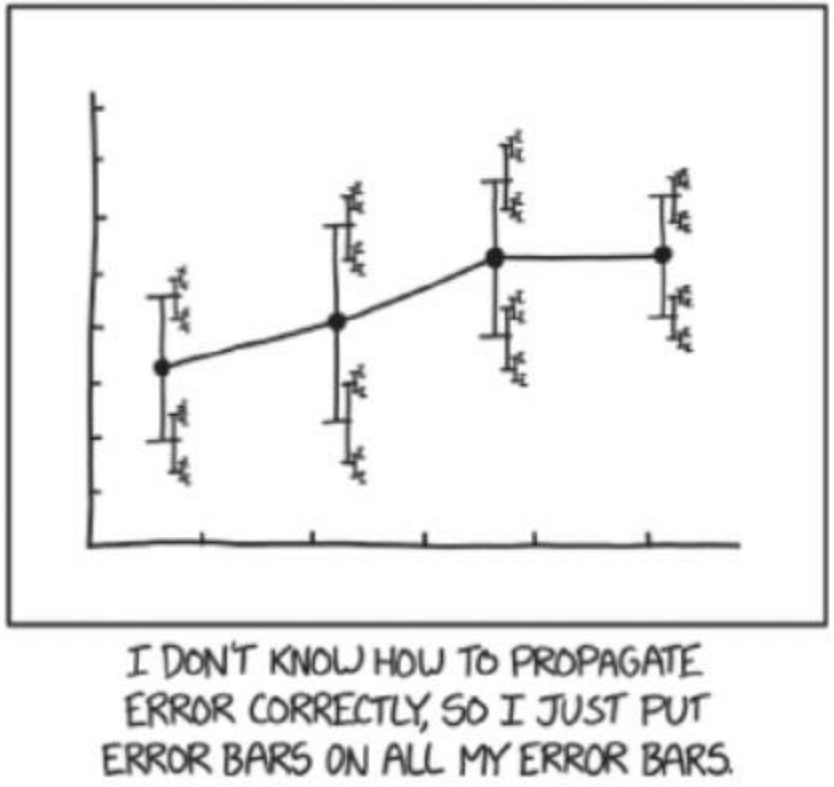
Expectation Tolerance Interval (type I) concept

100 simulated 95% (beta)-expectation TI



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

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

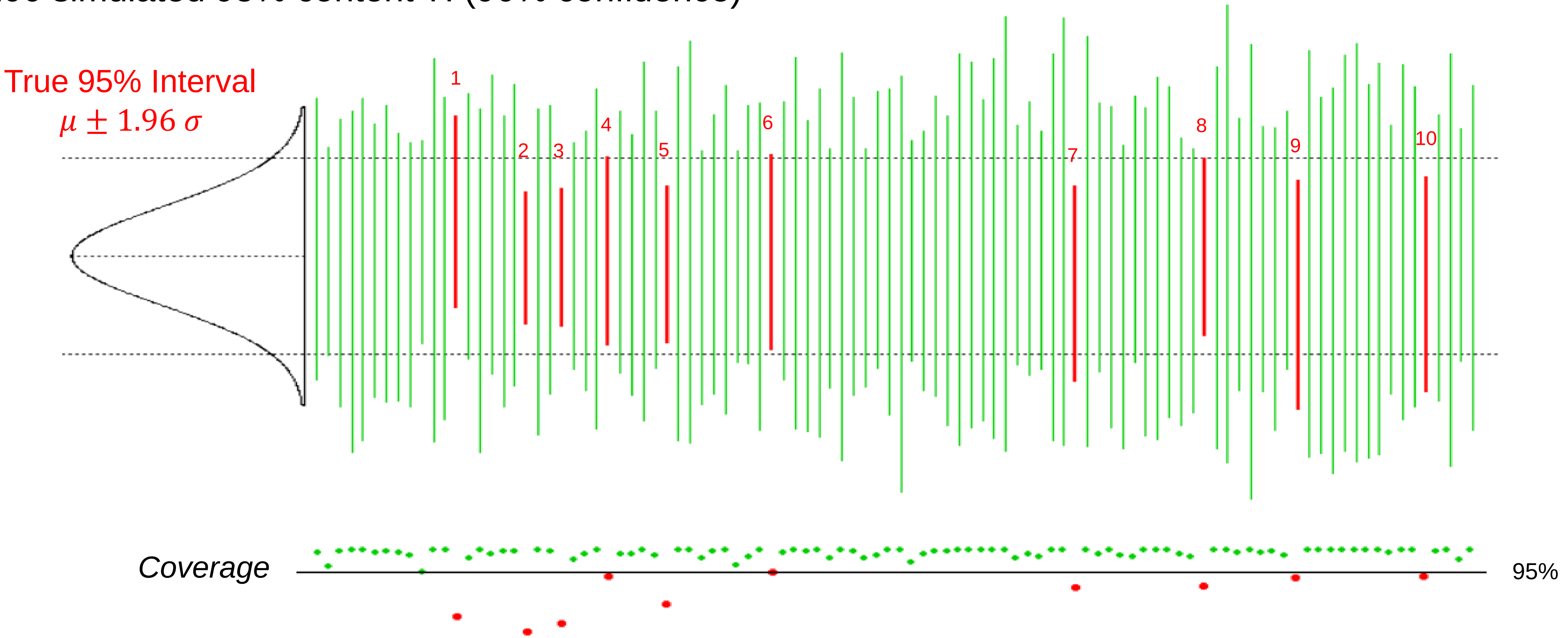
- Calculating the CI of a CI is awkward, confusing, misleading
- Unfortunately, widely used in method comparison studies (bridging studies) with Bland-Altman plot (agreement interval)

→ 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.

Content Tolerance Interval (type II) concept

100 simulated 95% content TI (90% confidence)



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

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

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
- $conf$ 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

TI and **quantiles**

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

Confidence, Prediction and Tolerance

90% CI		90% PI		98% TI (95% Conf)	
13.92	15.09	11.27	17.74	$-\infty$	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

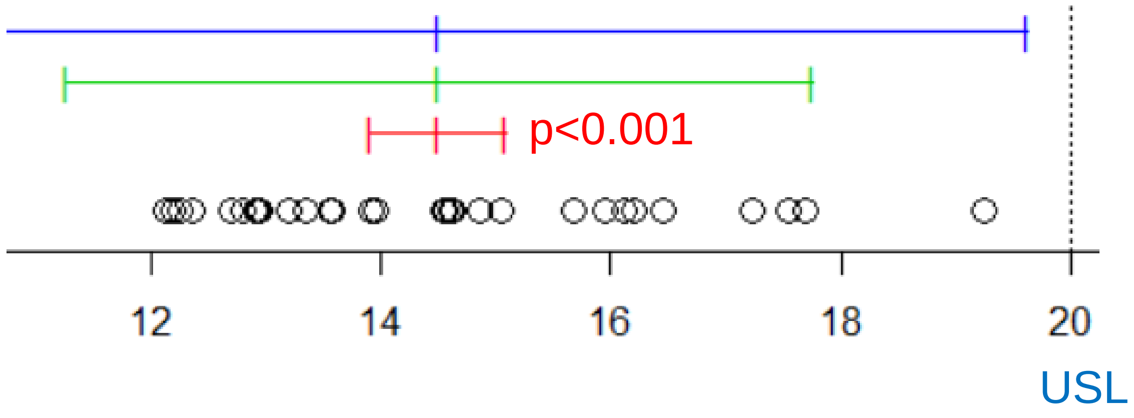
- 90% of the future products are expected to be between 11.27 and 17.74 (on average)

(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

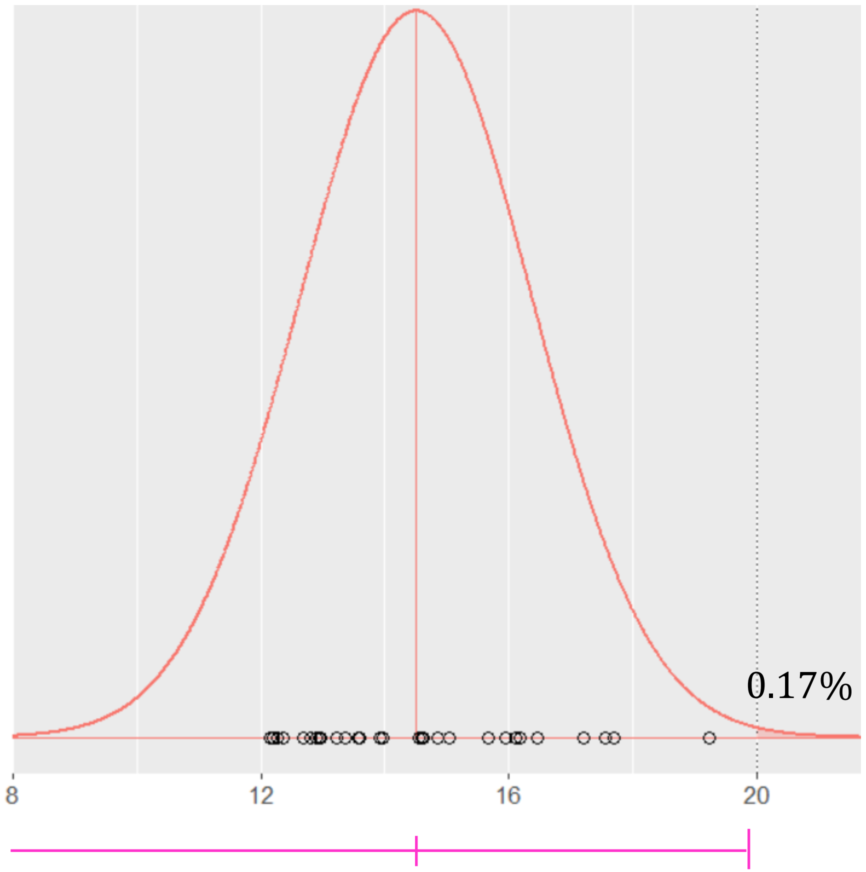


→ TI < USL
→ The POOS is lower than 2%
→ Smart Risk decision: Go ☺

Can we calculate the POOS?

GSK

Smart Risk: Probability Out Of Specification (POOS)



At least 98.7% *of future products will be lower than 20 (95% confidence)*

At most 1.3% *will be greater than 20 (95% confidence)*

$$1-\phi\left(z=\frac{USL-\bar{X}}{S}\right)=0.17\%$$

Interpretation

We expect 0.17% products OOS

95% Upper Bound by 'reversing' the TI

$$14.51+t_{0.95,29,z_{pred}\sqrt{30}}\frac{1.87}{\sqrt{30}}=20$$

```
uniroot(function(k) mean(x) + qt(0.95, 29, qnorm(k)*sqrt(n)) * sd(x) / sqrt(n) - 20, …….)
```

```
[1] 0.9867714
```

The POOS is 0.17% with a 95% upper bound equal to 1.3%

This « worst » scenario is lower than the maximal OOS (2%)

→ Smart Risk Decision: Go ☺

Smart Risk: Summary

Calculate the TI with the desired confidence level, the prediction (coverage) is related to the maximal tolerated risk

If the TI does not overlap the ‘threshold’

→ Go ☺

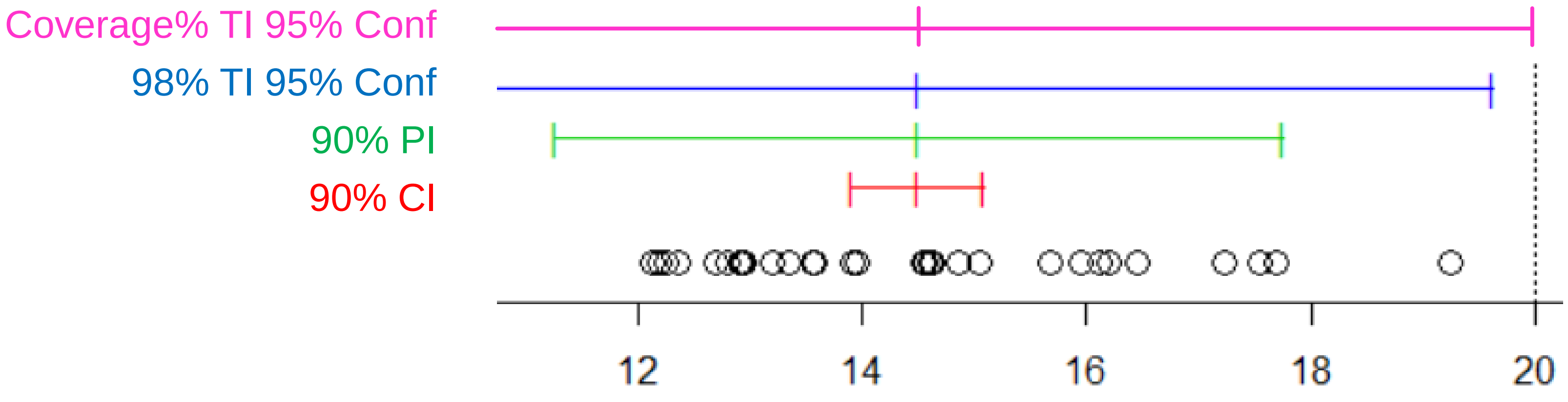
→ Otherwise, stop

Or, calculate the POOS and its upper bound

If the upper bound exceeds the maximal tolerated POOS

→ Stop

→ Otherwise, Go ☺



Overview of presentation

Smart Risk

Medical Research: misinterpretations

1 sample t-test

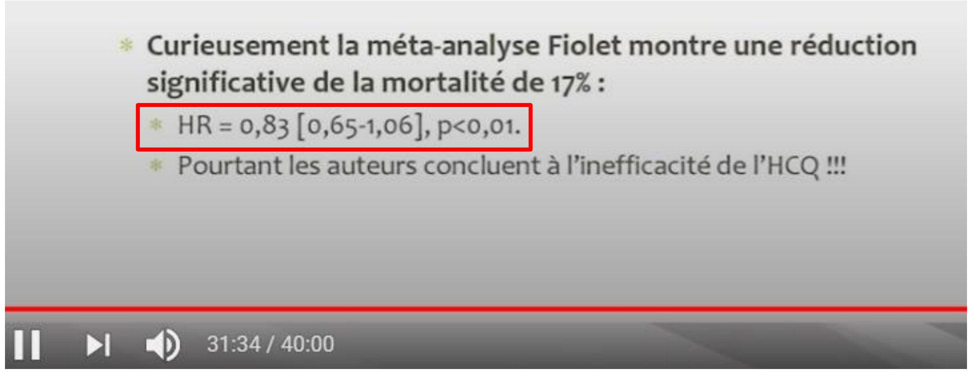
- Significance tests: Confidence Interval (CI), p-value, s-value
- Prediction Interval (PI)
- Tolerance Interval (TI)
- Individual Success Probability (ISP)
- p-value or ISP

Measurement error

2-arms clinical trials (B-value)

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



https://www.youtube.com/watch?v=FIhp7PtCEtY
Critical analysis of treatments for COVID-19
(Analyse critique des traitements de la COVID-19)

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 ($p = 0.000000000000000036$).

- 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

1-sample t-test
Toy Example: systolic blood pressure (SBP) (mmHg)

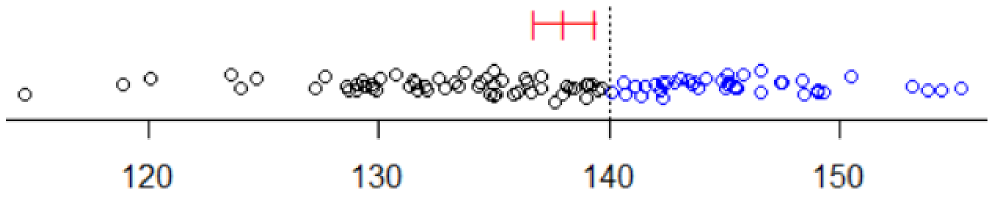
$H_0: \mu = 140$ $\quad\quad\quad \alpha = 0.05$

$H_1: \mu < 140$ $\quad\quad\quad n = 100$

The mean is estimated by $\bar{X} = 138.11$ and the standard deviation by $S = 7.97$

The 90% CI for the mean is therefore: [136.79, 139.43]

The 1-sided 95% CI for the mean is: $]-\infty, 139.43]$



The estimated mean is < 140

Its uncertainty as well: 140 ∉ CI

→ *H_0 is rejected*

→ *The mean is significantly lower than 140 mmHg*

1-sample t-test

p-value, s-value

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

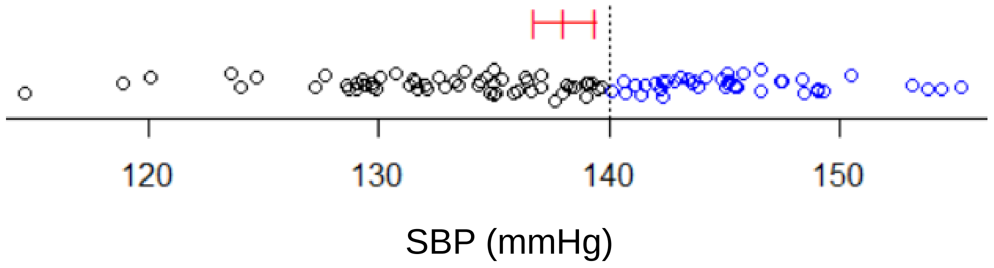
What about the p-value?

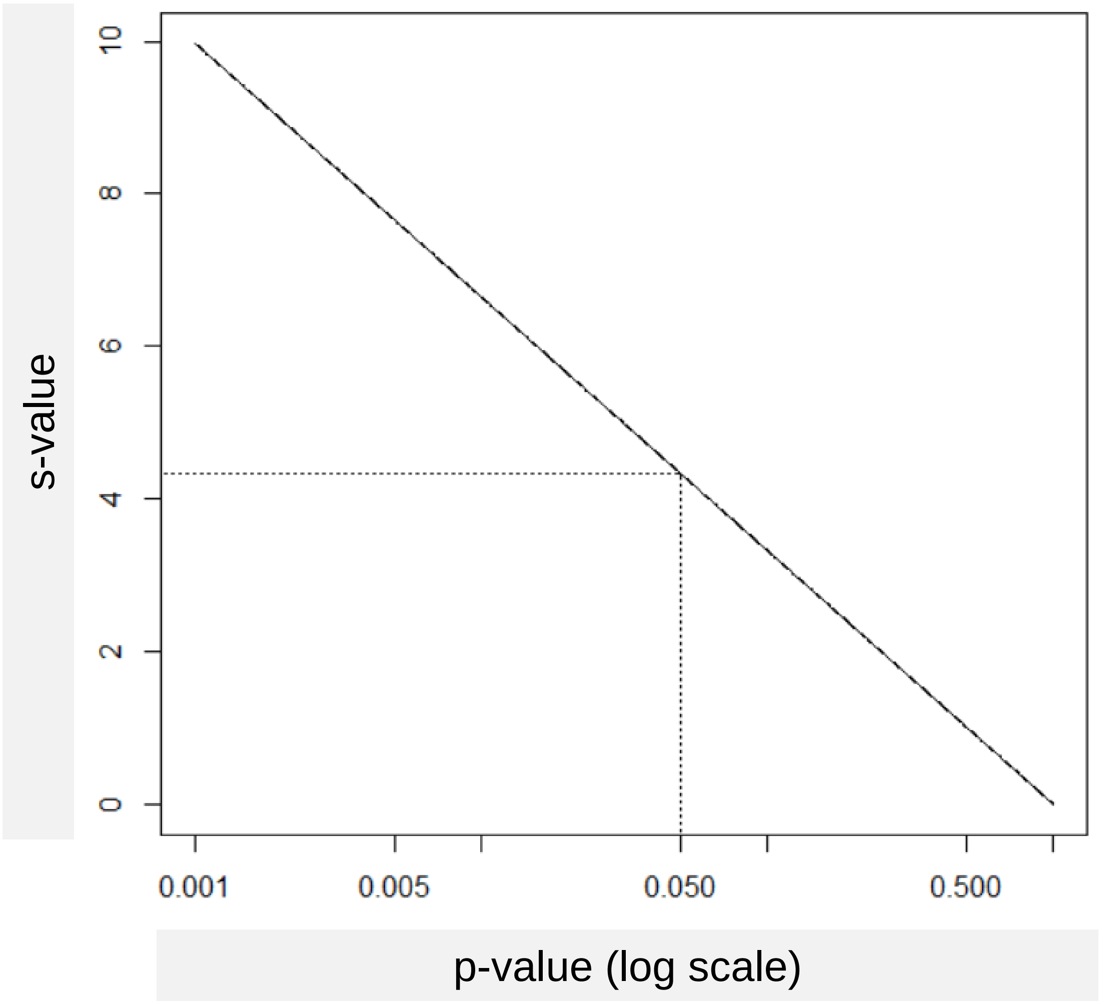
p-value = 0.0098 (significant at $\alpha = 0.05$)

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





GSK

1-sample t-test
Individual Success Probability

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

What about the proportion of patients with a SBP > 140 ?

$$\phi\left(\frac{\bar{X} - 140}{S}\right) = 41\%$$

This will be called, here, the Individual Success Probability (ISP)

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

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

				$H_1: \mu < 140$		ISP (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

The p-values collapse (the s-values soar) while the ISP remains constant

How to quantify the uncertainty on the ISP?

1-sample t-test

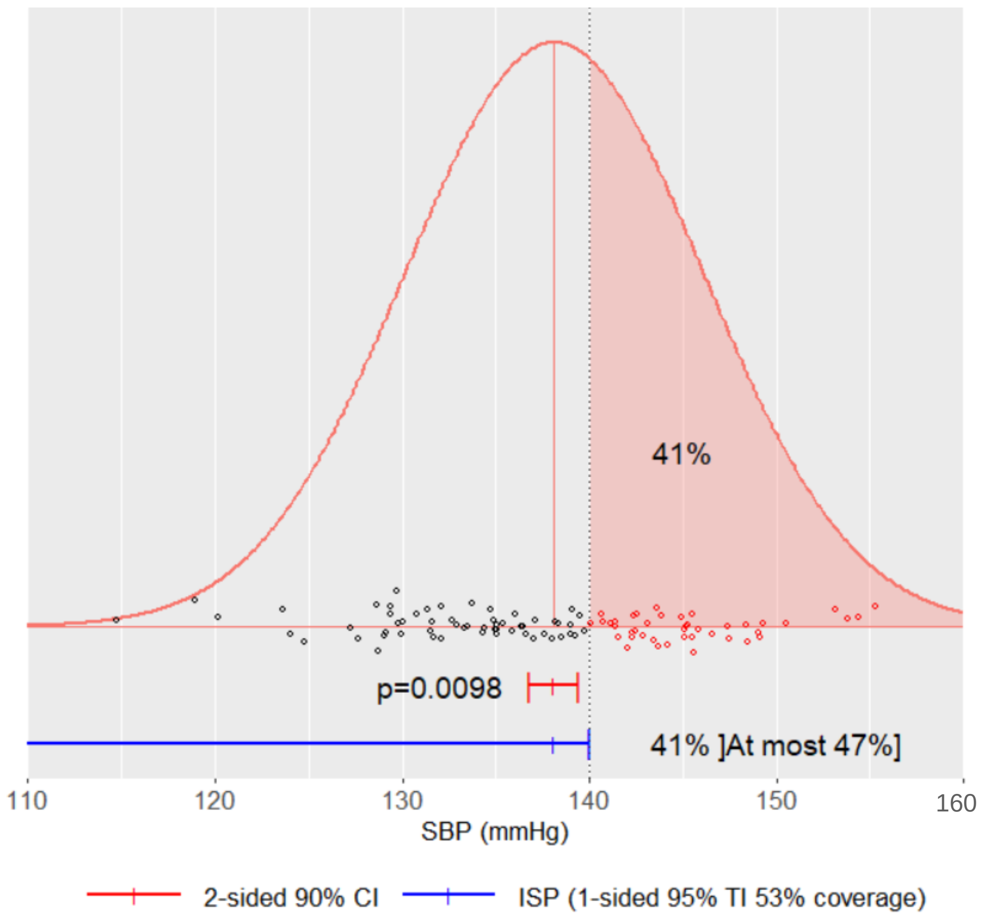
How to take into account the uncertainty on the ISP?

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 ISP



1-sample t-test

Toy Example: systolic blood pressure (SBP) (mmHg)

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

				$H_1: \mu < 140$		ISP (95% CI)	
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 [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**]%

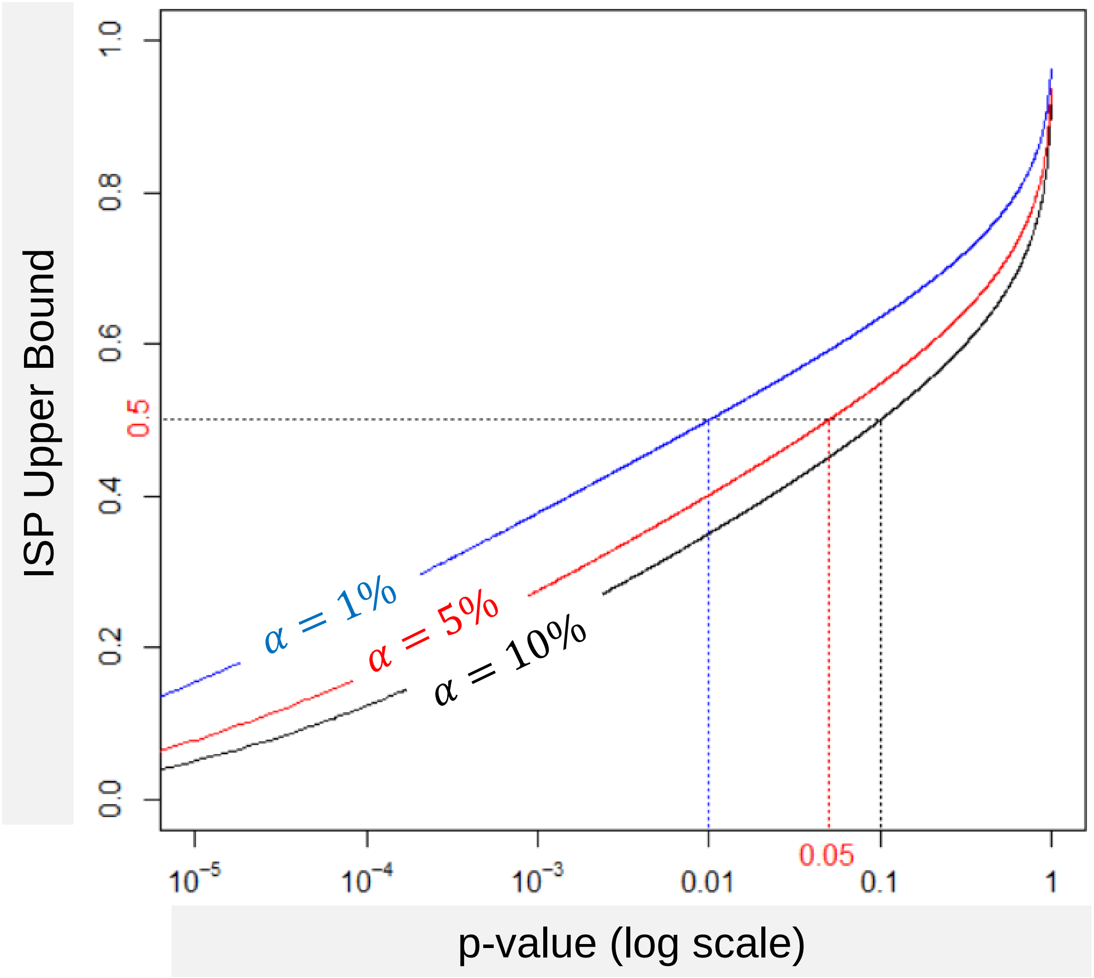
CI and p-value might be confusing

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

- ✓ ***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

One-to-one function ISP & p-value



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

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

Main advantages of the ISP 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

Overview of presentation

Smart Risk

Medical Research: misinterpretations

1 sample t-test

- Significance tests: Confidence Interval (CI), p-value, s-value
- Prediction Interval (PI)
- Tolerance Interval (TI)
- Individual Success Probability (ISP)
- p-value or ISP

Measurement error

2-arms clinical trials (B-value)

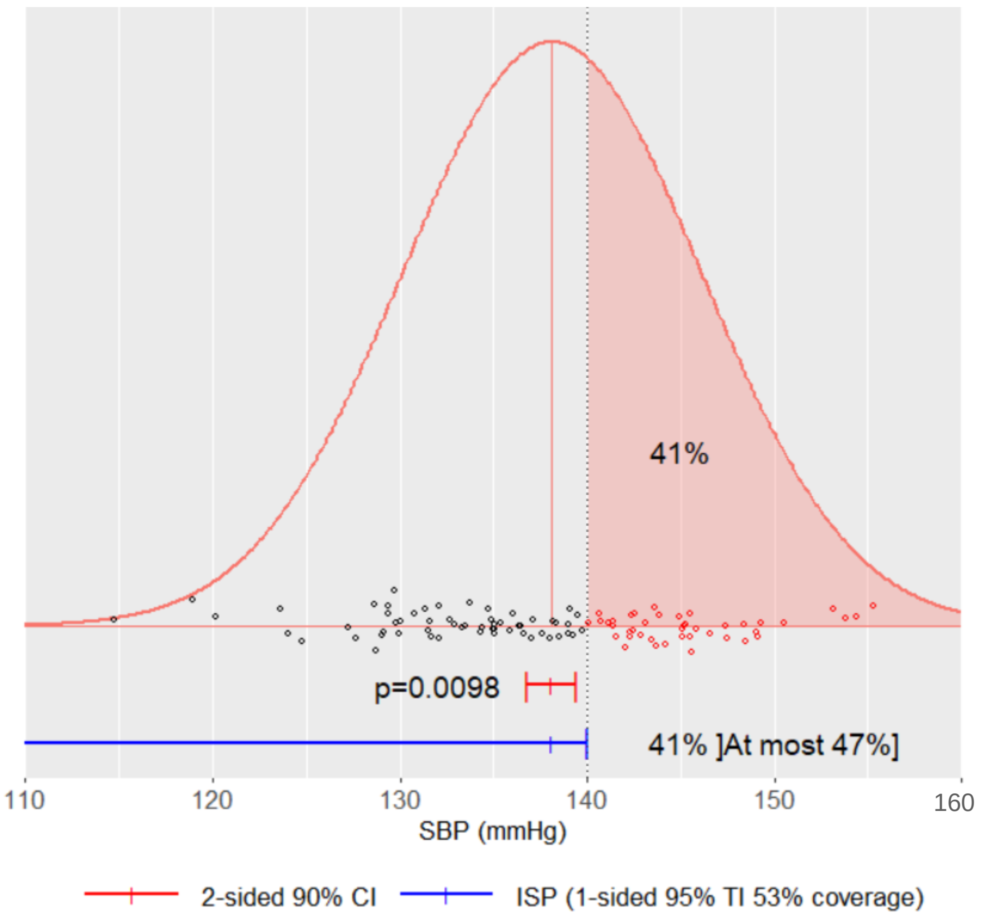
ISP and measurement error

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
(and not its measured value)

Overview of presentation

Smart Risk

Medical Research: misinterpretations

1 sample t-test

- Significance tests: Confidence Interval (CI), p-value, s-value
- Prediction Interval (PI)
- Tolerance Interval (TI)
- Individual Success Probability (ISP)
- p-value or ISP

Measurement error

2-arms clinical trials (B-value)

2-arms clinical trials: Parallel treatment groups

What if you undergo treatment A,
and your friend treatment B?

- A^{n+1} is the outcome of a new patient under treatment A
- B^{n+1} is the outcome of a new patient under treatment B

A^{n+1} and B^{n+1} are independent

→ $P(A^{n+1} > B^{n+1})$ is the ISP (for A over B)
→ $P(B^{n+1} > A^{n+1})$ is the ISP (for B over A)

Remark
$P(A^{n+1} > B^{n+1})$ is here simplified as $P(A > B)$,
and also named the « B-value » in the literature (or D-value when reversed) *

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

2-arms clinical trials: Summary

	Mean	Pooled	Mean Difference	$H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$		$P(D < 0)$	$P(D > 0)$
n	Diff.	SD	95% CI	p-value	# Head	b-value	d-value
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 constant

d-value constant

How to add the 95% CI ?

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

✓ Well-established methodology in non-clinical statistics

Example: comparability,…

GSK

2-arms clinical trials: Summary

n	Mean Diff.	Pooled SD	Mean Difference 95% CI	$H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ p-value	# Head	b-value	d-value	Success Probability $P(D < 0)$ 95% CI	$P(D > 0)$ 95% CI
50	0.12	1.41	[-0.27, 0.52]	p=0.54	0.9	53.5	46.5	[42.5, 64.2]%	[35.8, 57.5]%
100	0.12	1.41	[-0.15, 0.40]	p=0.38	1.4	53.5	46.5	[45.7, 61.2]%	[38.8, 54.3]%
500	0.12	1.41	[0, 0.25]	p=0.05	4.3	53.5	46.5	[50.0, 57.0]%	[43.0, 50.0]%
1000	0.12	1.41	[0.04, 0.21]	p=0.006	7.5	53.5	46.5	[51.0, 56.0]%	[44.0, 49.0]%
5000	0.12	1.41	[0.08, 0.16]	p<.001	31	53.5	46.5	[**52.4**, 54.6]%	[45.4, **47.6**]%

p-value collapse *s-value soar* *b-value constant* *d-value constant*

ISP interpretation ($n = 5000$) *

× ~~*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 their friends under A*

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

Borderline test ($n = 500$)

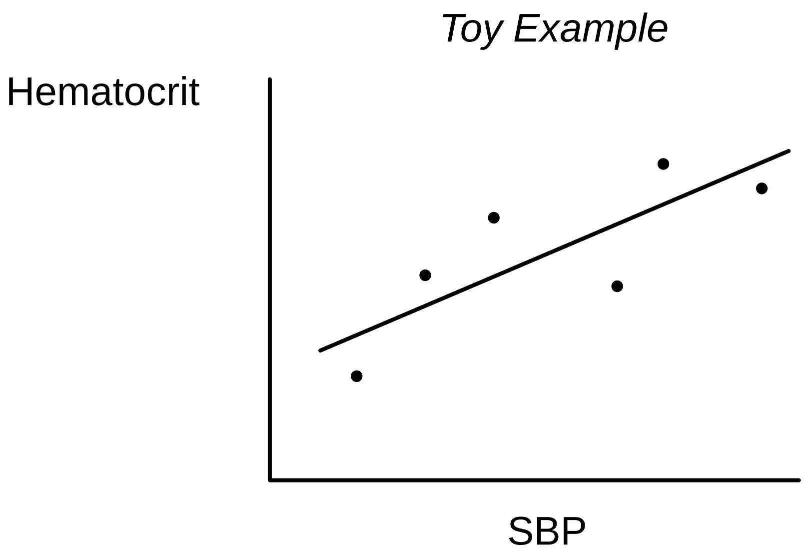
- p-value = 5%
- CI bound = 0
- ISP bound = 50%

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* 2019.

Demystify a (statistical) urban legend

How do you interpret a slope ?



Slope $= \hat{\beta} = 1.5$

Interpretation ? (OLS assumptions fulfilled)

No !

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

*→ Interpret coefficients as comparisons, not effects **

→ Like the Tolerance Interval for Differences

* 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

Prediction
Tolerance
ISP
Bayesian,…

Rear brake

CI Means
p-value

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

While CI and p-value can be confusing or controversial,
Smart Risk, Tolerance Intervals and Success Probabilities
have straightforward interpretations

Last but not least

References

- Francq, Lin, Hoyer, Cartiaux, Kenett: Individual Success Probability: Beyond the t-test and p-values. (2022) (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

Dan Lin
Walter Hoyer
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. R Kenett is an employee of the KPA group and the Samuel Neaman Institute.

GSK

2-arms clinical trials: Summary

n	Mean Diff.	Pooled SD	Mean Difference 95% CI	$H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ p-value	# Head	b-value	d-value	Success Probability $P(D < 0)$ 95% CI	$P(D > 0)$ 95% CI
50	0.12	1.41	[-0.27, 0.52]	p=0.54	0.9	53.5	46.5	[42.5, 64.2]%	[35.8, 57.5]%
100	0.12	1.41	[-0.15, 0.40]	p=0.38	1.4	53.5	46.5	[45.7, 61.2]%	[38.8, 54.3]%
500	0.12	1.41	[0, 0.25]	p=0.05	4.3	53.5	46.5	[50.0, 57.0]%	[43.0, 50.0]%
1000	0.12	1.41	[0.04, 0.21]	p=0.006	7.5	53.5	46.5	[51.0, 56.0]%	[44.0, 49.0]%
5000	0.12	1.41	[0.08, 0.16]	p<.001	31	53.5	46.5	[**52.4**, 54.6]%	[45.4, **47.6**]%

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

Our reply

The d-value you can't buy…

"Individual Success Probability: Beyond the t-test and p-values" (2022, under review)

GSK