DUKE NUS
GRADUATE MEDICAL SCHOOL SINGAPORE

transforming medicine, improving lives

SINGAPORE CLINICAL RESEARCH INSTITUTE

Quantifying the strength of evidence for alternative hypotheses – a missing link

Edwin SY Chan PhD

Singapore Clinical Research Institute

&

Duke-NUS Graduate Medical School

Disclosure of Interests (last 3 years)

< **Edwin Chan**>

I certify that, to the best of my knowledge, no aspect of my current personal or professional situation might reasonably be expected to affect significantly my views on the subject on which I am presenting, other than the following*:

< *Employee of the Singapore Clinical Research Institute* >

< *Paid secondment to Duke-NUS Graduate Medical School* >

< *Paid secondment to Singapore Ministry of Health* >

(e.g. employment, consultancy, research grant, honoraria, stock ownership, sponsored education, educational grant, other funding, etc ...)

< *Director, Singapore Branch, Australasian Cochrane Centre* >

(e.g. committee memberships, academic interest, other voluntary engagement, etc...)

** If you have nothing to disclose, delete „other than the following“*

Outline

- Three important but different aims of data analysis – fundamental place of evidence quantification
- Inadequacy of the p-value as a measure of evidence
- Recommendations on how to quantify & report the evidence from quantitative data

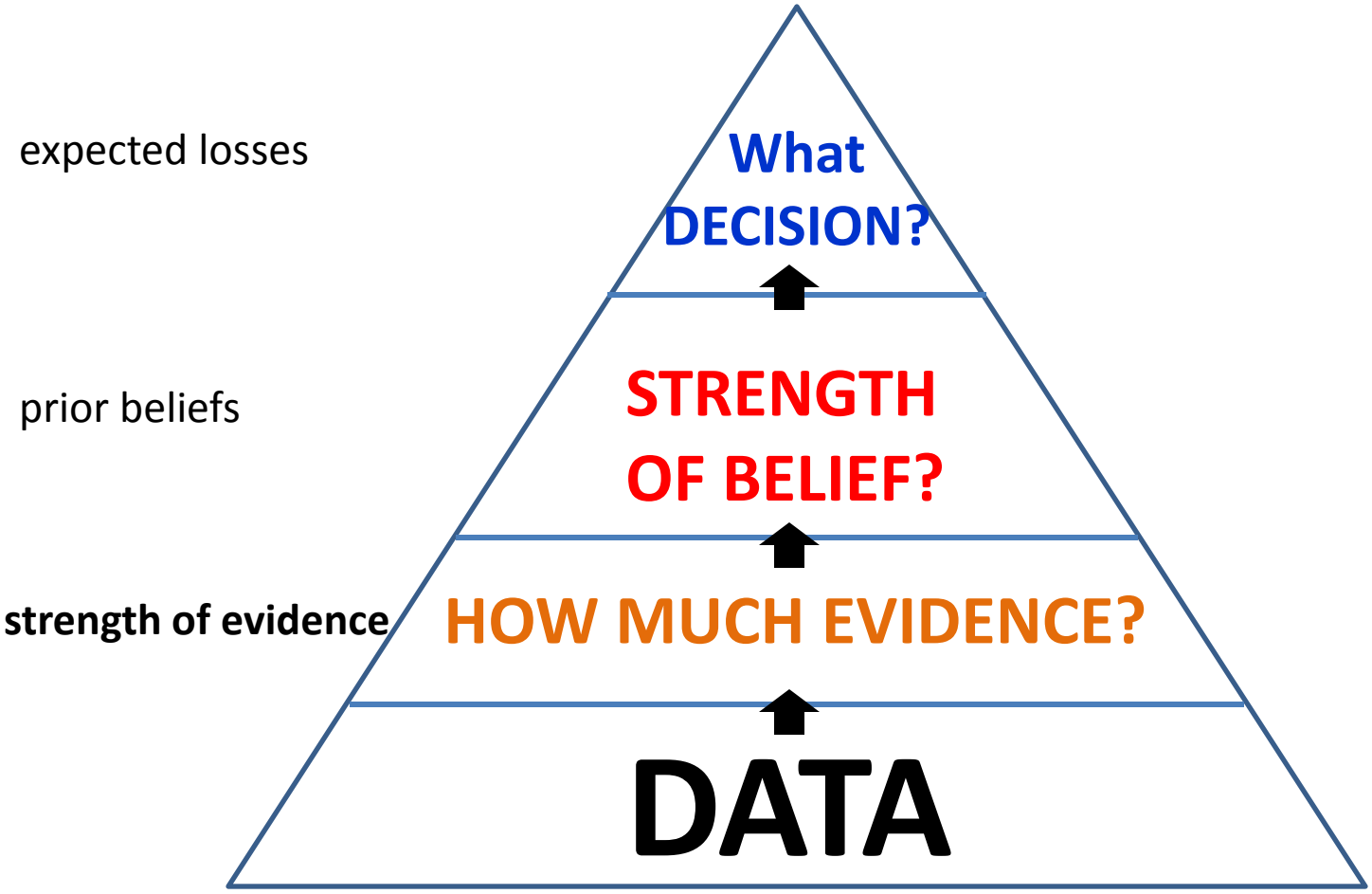
Part 1

- Three important but different aims of data analysis – fundamental place of evidence quantification
- Inadequacy of the p-value as a measure of evidence
- Recommendations on how to quantify & report the evidence from quantitative data

David Cox (1958)

Even in problems where a clear-cut decision is the main object, it very often happens that the assessment of losses and prior information is subjective, so that it will help to *get clear first the relatively objective matter of what the data say* ... it may be argued that one of the main calls for probabilistic statistical methods arises from the need to have agreed *rules for assessing strength of evidence*.

3 related but distinct aims of data analysis



Diagnosing a condition … (Royall 1997)

Consider a diagnostic test for the presence or absence of some disease (**D**) in patient Mr Smith. Suppose that previous research has shown that the Dx test is a "good" one i.e. with **sensitivity = 0.95** & **specificity= 0.98**

The problem – choosing between competing hypotheses

- **Competing Hypotheses**

  A. Mr. Smith **has** the disease (D+)

  B. Mr. Smith **does not have** the disease (D-)

- **Given the test result (DATA), 3 questions could be asked …**

  1. **What is the evidence** supporting hypothesis **D+** over **D-** given the test result?
  2. **What do I believe** Mr Smith has **(D+** or **D-)**, given the evidence & my prior suspicions?
  3. **What should I do** for Mr Smith, given my updated belief and other relevant external considerations?

What is the Evidence?

- What is the evidence supporting hypothesis (D+) versus (D-) given the data a positive test result (T+)?

- Evidence measure: $\frac{\text{Prob of data (T+) assuming D+}}{\text{Prob of data (T+) assuming D-}}$

- **Likelihood ratio (+)** = **47.5** (D+ vs D-)
  - LR+ >>1 $\Rightarrow$ evidence supports D+ much more than D-

What do I believe?

- **Given that the evidence supports hypothesis D+ more than D-**

Should the doctor **believe** Mr. Smith has the disease?

- ✓ not necessarily … it depends on your pre-existing level of suspicion i.e. Pr(**D**)

Scenario	Pr(D)	Pr(D | T+)	Justified Belief in
Common disease	20%	92%	D+
Rare disease	0.1%	4.5%	D-

What should I do/decide?

- Given T+ & that D is rare & belief is in D-, the doctor could still justifiably decide to treat Mr Smith, why?

Basis of the Decision	Justified to Rx	Not Justified to Rx
Evidence: LR+ = 47.5	✓ (Primary studies & Reviews)	
Belief: Post-test probability of D+ = 4.5%		✓ (Primary studies & Reviews)

What should I do/decide?

- Other considerations: expected cost & benefits trade-off

Basis of the Decision	Justified to Rx	Not Justified to Rx
Evidence: LR+ = 47.5	✓ (Primary studies & Reviews)	
Belief: Post-test probability of D+ = 4.5%		✓ (Primary studies & Reviews)
Other considerations: 1. Effective, cheap & harmless Rx 2. Failure to treat is 'disastrous'	✓ Guidelines	

Summary of Part 1

- Within the bounds of a focused research question, investigators need to distinguish between the 3 distinct goals of data analysis & take care to use the appropriate statistical paradigm

- Like Cox, we recommend that we should always start with an objective quantification of the data evidence

Part 2

- Three important but different aims of data analysis – fundamental place of evidence quantification

- Inadequacy of the p-value as a measure of evidence

- Recommendations on how to quantify & report the evidence from quantitative data

P-value as an evidence metric

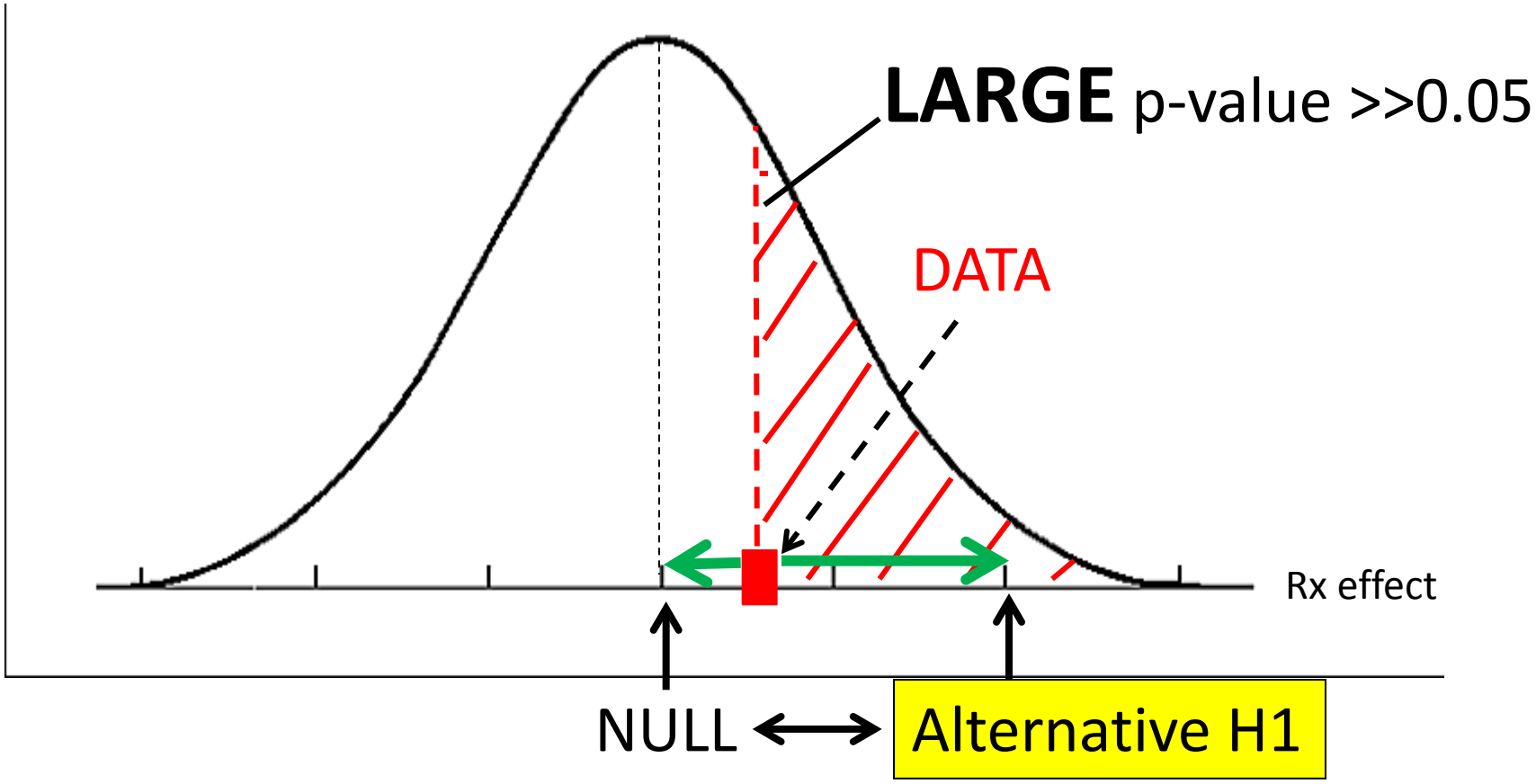
- Definition: the probability of occurrence of the observed result or results more extreme than it, assuming the NULL hypothesis is true

- Key features of this definition:
  - makes use of unobserved data – data "more extreme" than that which was observed
  - attempts to measure evidence for one hypothesis – typically the NULL hypothesis of no difference

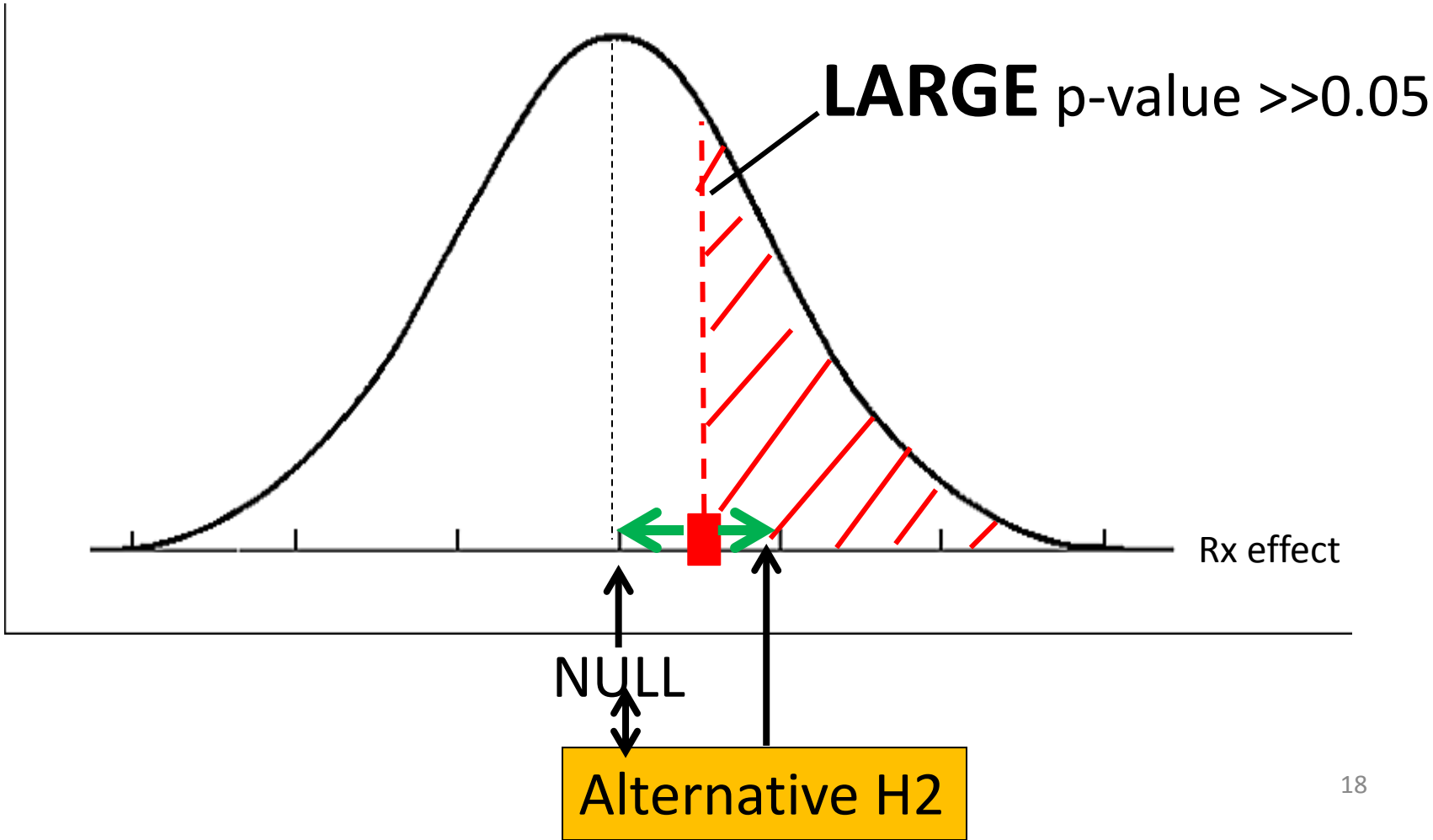
Rationale for the p-value as an evidence metric

- "Large" p-values are "strong" evidence for the NULL hypothesis
- "Small" p-values are "weak" evidence for the NULL hypothesis
- Traditional strong-weak threshold is p = 0.05

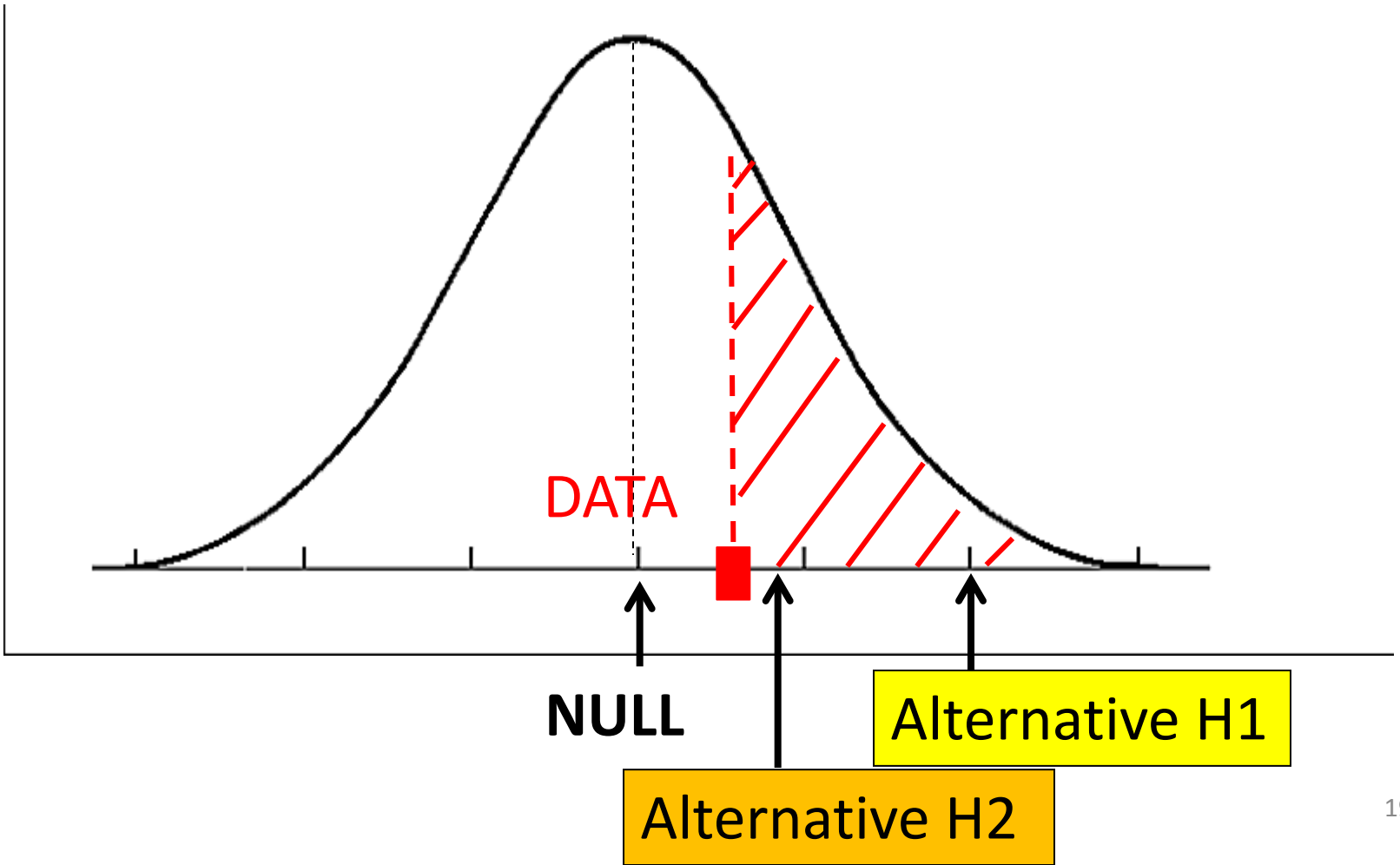
"Large" p-value, data closer to the NULL – more support for the NULL hypothesis (?)



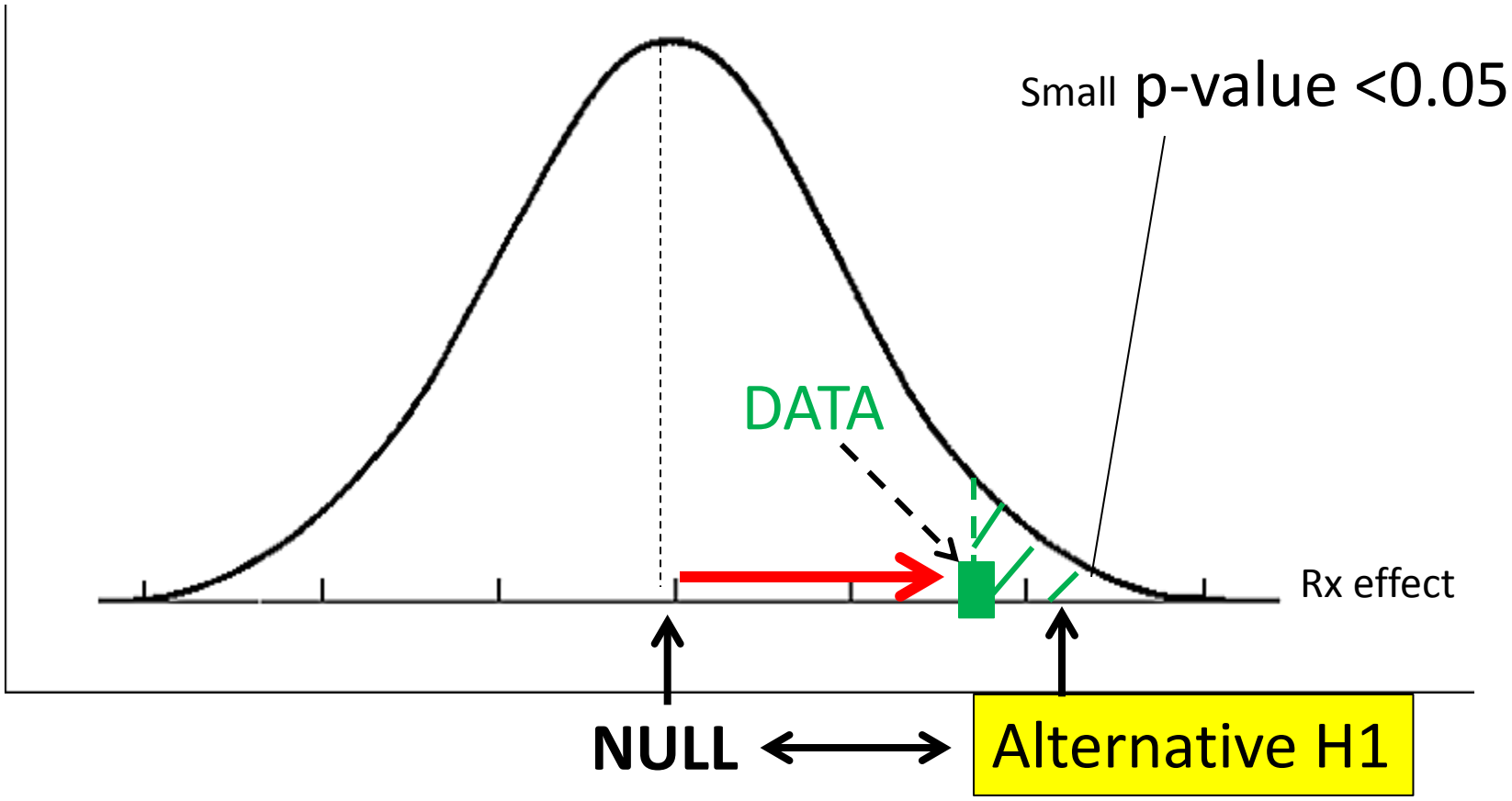
But when the alternative hypothesis changes …



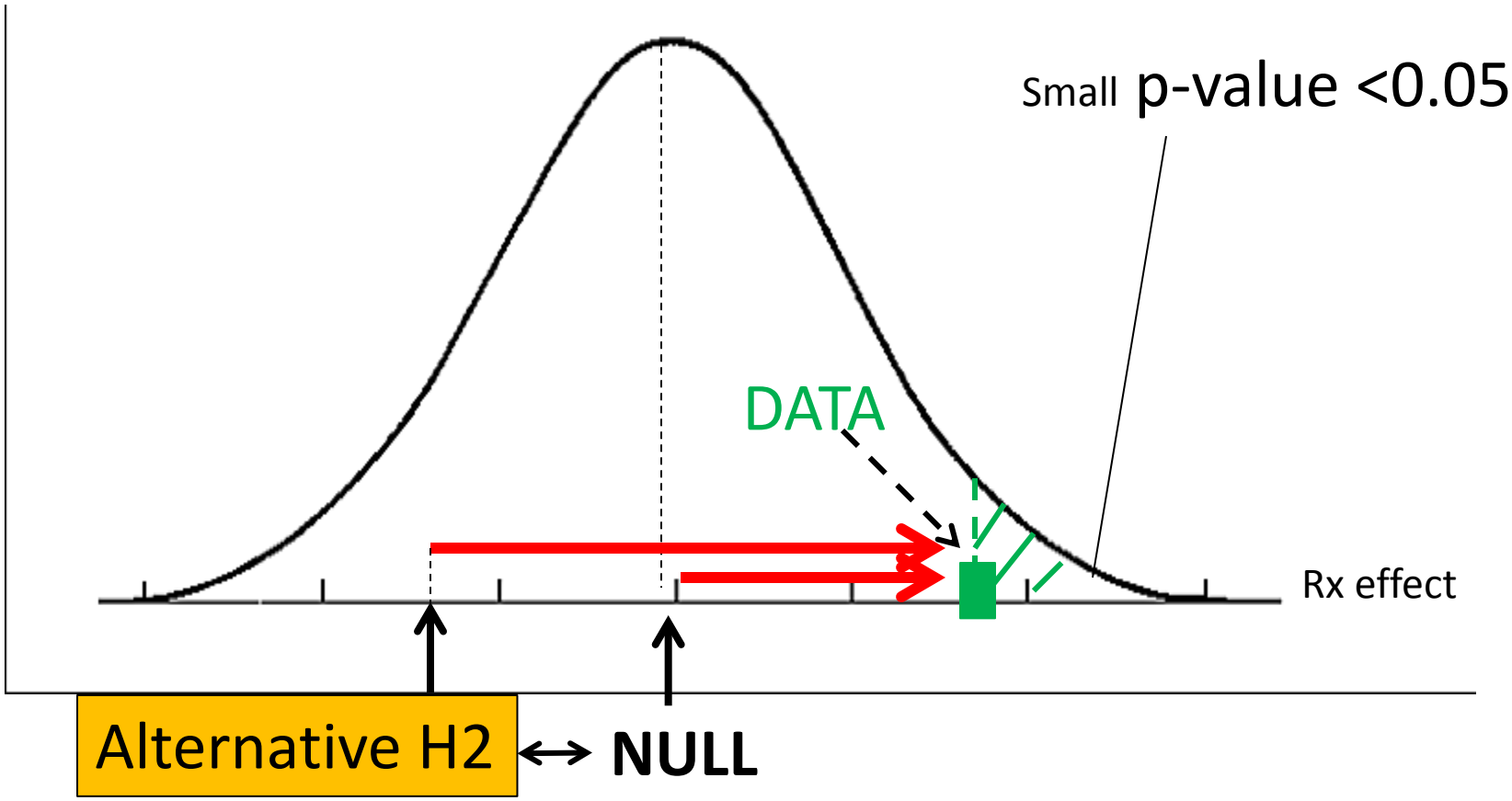
Conclusion: SAME data provides more evidence for NULL than for H1 but less evidence for NULL than for H2, yet the p-value is the same for both cases?!



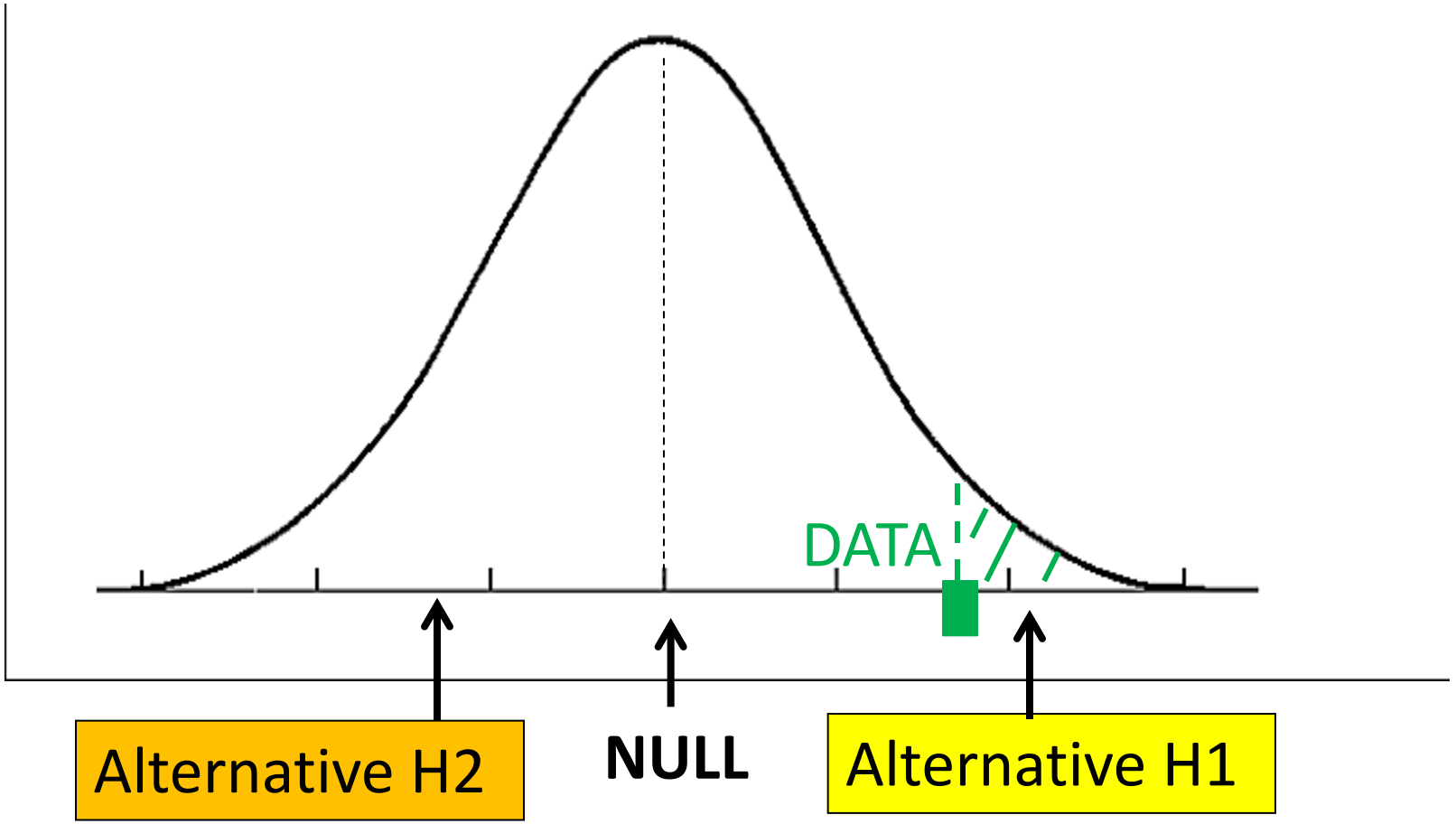
"Small" p-value data further from the NULL – less support for the NULL (?)



But when the alternative hypothesis changes …



Conclusion: SAME data provides less evidence for NULL than for H1 but more evidence for NULL than for H2, yet the p-value is the same for both cases?!

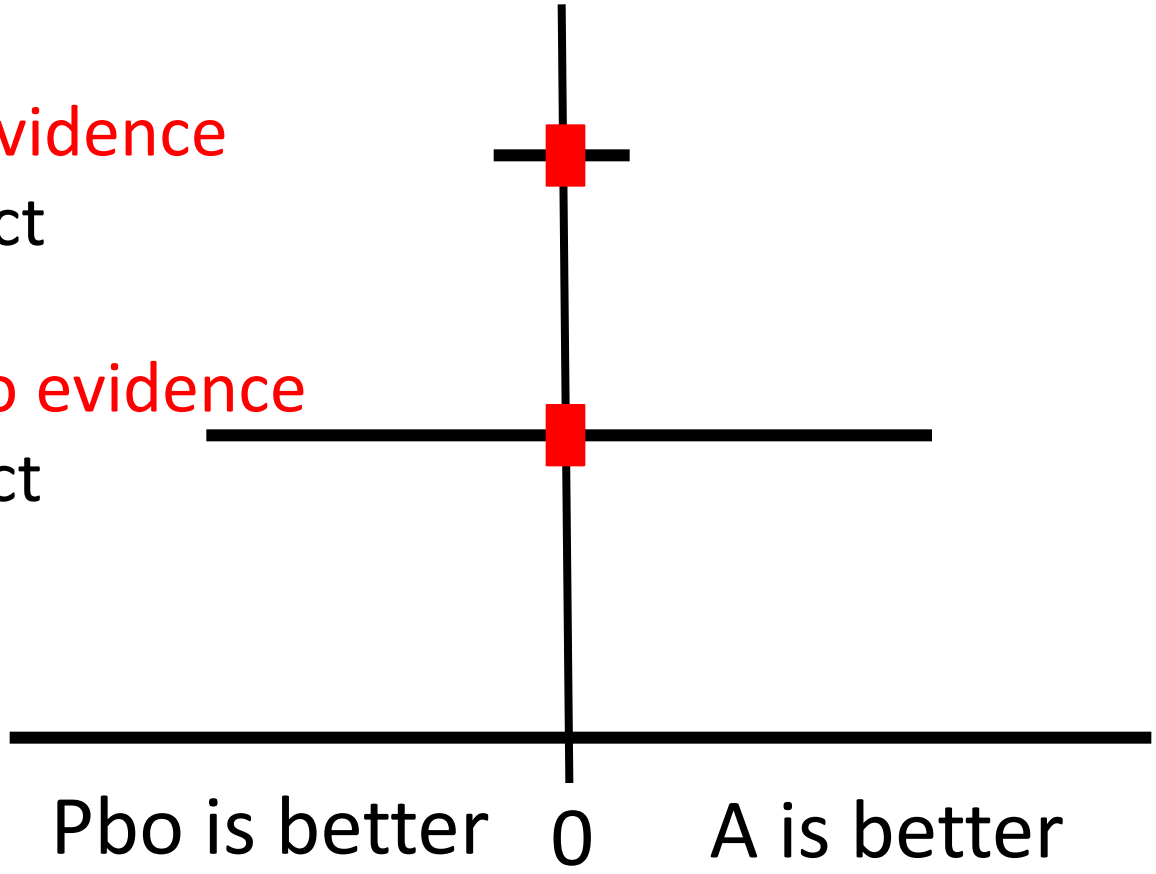


No evidence of effect ≠ Evidence of no effect

What we intuitively know

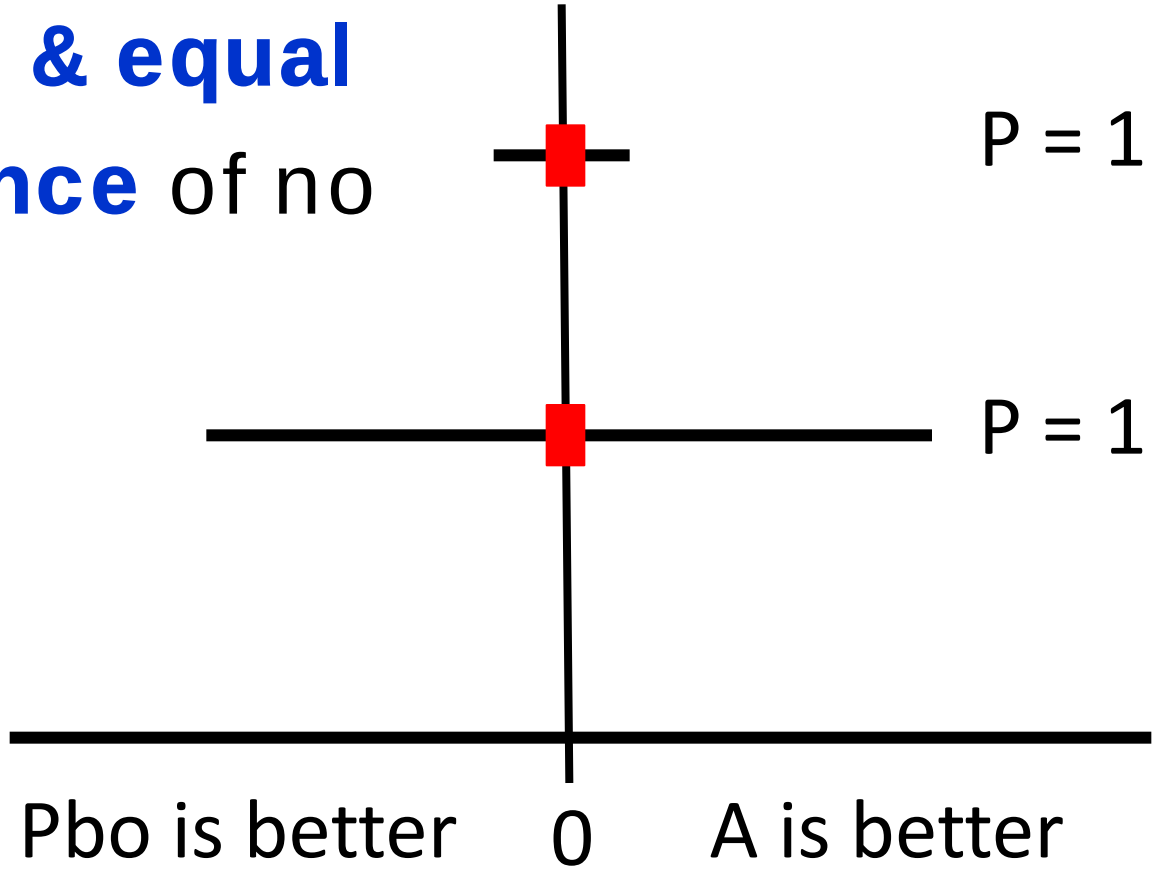
(Strong) Evidence
of no effect

(Weak) No evidence
of an effect



What the p-value concludes

Strong & equal evidence of no effect



Is the confidence interval a solution?

- Short Answer: **NO**

- Points to note:
  - CI's were not proposed as evidence mterics
  - a particular CI is has no useful interpretation (wiithin its own paradigm)therefore frequently misinterpreted as something more meaningful to the researcher by borrowing meaning from other paradigms
    - ✓ Probability interval (Bayesian paradigm)
    - ✓ Evidence support interval (Likelihood paradigm)

Summary of Part 2

- The p-value is a fatally flawed measure of evidence for 2 main reasons:
- (1) it incorporates information from unobserved data
- (2) it does not include specification of the alternative hypotheses

Part 3

- Three important but different aims of data analysis – fundamental place of evidence quantification

- Inadequacy of the p-value as a measure of evidence

- Recommendations on how to quantify & report the evidence from quantitative data

The Likelihood approach to measuring evidence (Hacking 1965)

- Given the study result (the DATA), to quantify the evidence for 2 competing hypotheses A & B, compute:
  - Probability of DATA if Hypothesis A is true = $P_A(DATA)$
  - Probability of DATA, if Hypothesis B is true = $P_B(DATA)$

- If $P_A(DATA) > P_B(DATA)$, we may assert that there is more DATA evidence to support A than B

The Evidence metric – the Likelihood Ratio

- The **ratio of the 2 probabilities** is a measure of the strength of evidence for hypothesis A vs B
  - **Likelihood Ratio** (LR)

$$LR = \frac{P_A(DATA)}{P_B(DATA)} \qquad (0, +\infty)$$

  - ✓ if **LR ≈ 1**: evidence **favors both hypotheses equally**
  - ✓ if **LR > 1**: evidence **favors H_A over H_B**
  - ✓ if **LR < 1**: evidence **favors H_B over H_A**

Dx TEST EXAMPLE revisited

- The 2 possible hypotheses were:
  - Mr Smith has disease (D+) or does not (D-)
- The DATA was the POSITIVE test result (T+)
- $P_{D+}(T+) = 0.95$ and $P_{D-}(T+) = 0.02$
- Since $P_{D+}(T+) > P_{D-}(T+)$, the data support D+ MORE THAN D-, but by how much?
- The Likelihood ratio (LR) is 0.95/0.02 = 47.5
- The evidence is 47.5 times stronger for D+ than for D-

In reality there are many possible alternative hypotheses …

- For example:

Question about …	Hypothesis being estimated is …	Range of possible hypothesis values
INCIDENCE / PREVALENCE	Incidence/prevalence Proportion	0 to 100%
TREATMENT EFFECT	1. Odds ratio 2. Mean difference	0 to +∞ -∞ to +∞
HARM	Relative risk	0 to +∞

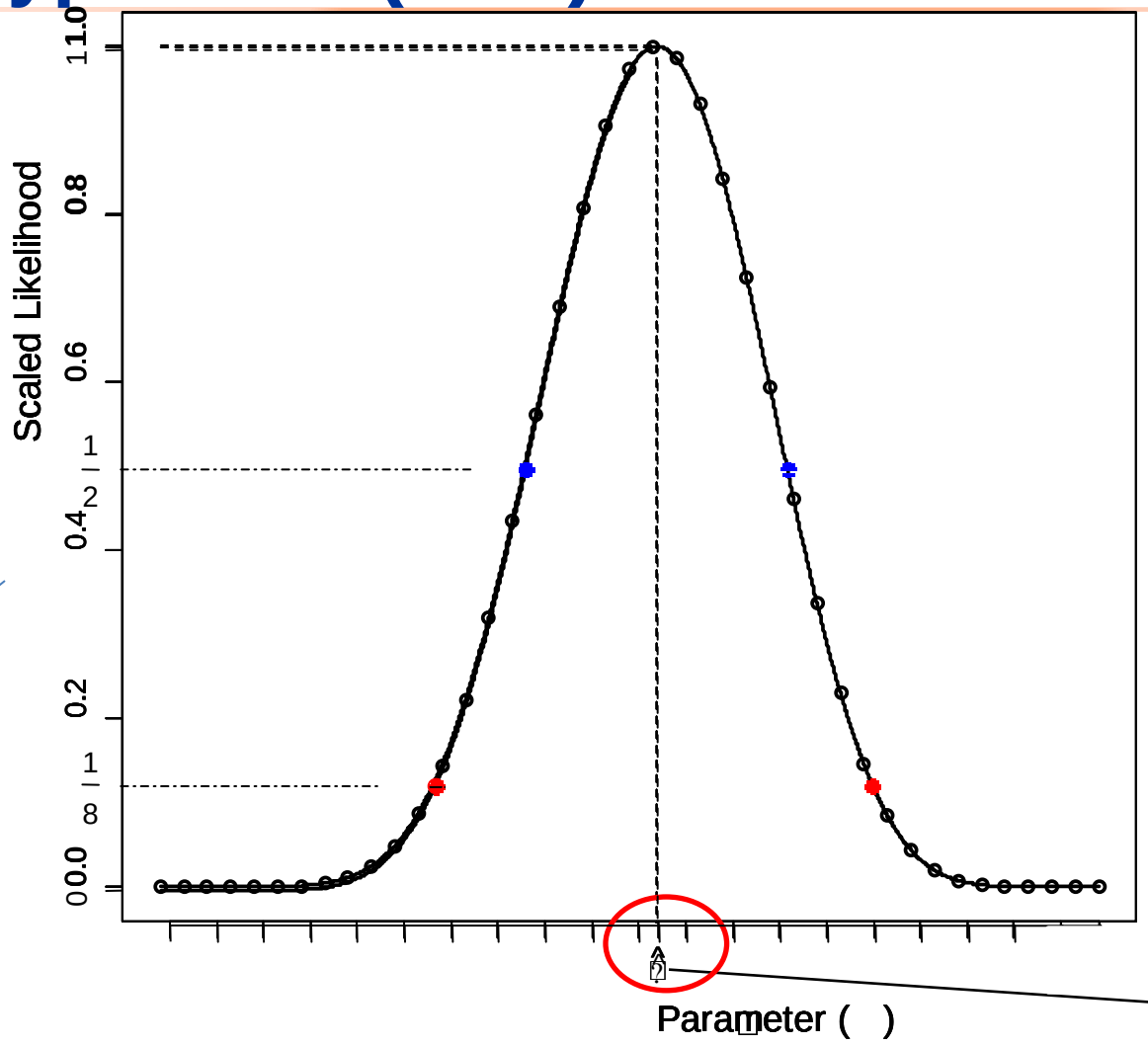
Summarizing the evidence across a continuum of possible hypotheses …

- The likelihood or support function for the given result (data)maps all possible hypotheses to their strength of evidence as compared with the most supported hypothesis - scaled LR(range from 0 -1)

- the most supported hypothesis (MSV) sitself has the maximum value for the scaled LR i.e. 1

The Support Curve & the most supported hypothesis (MSV)

Given the observed DATA

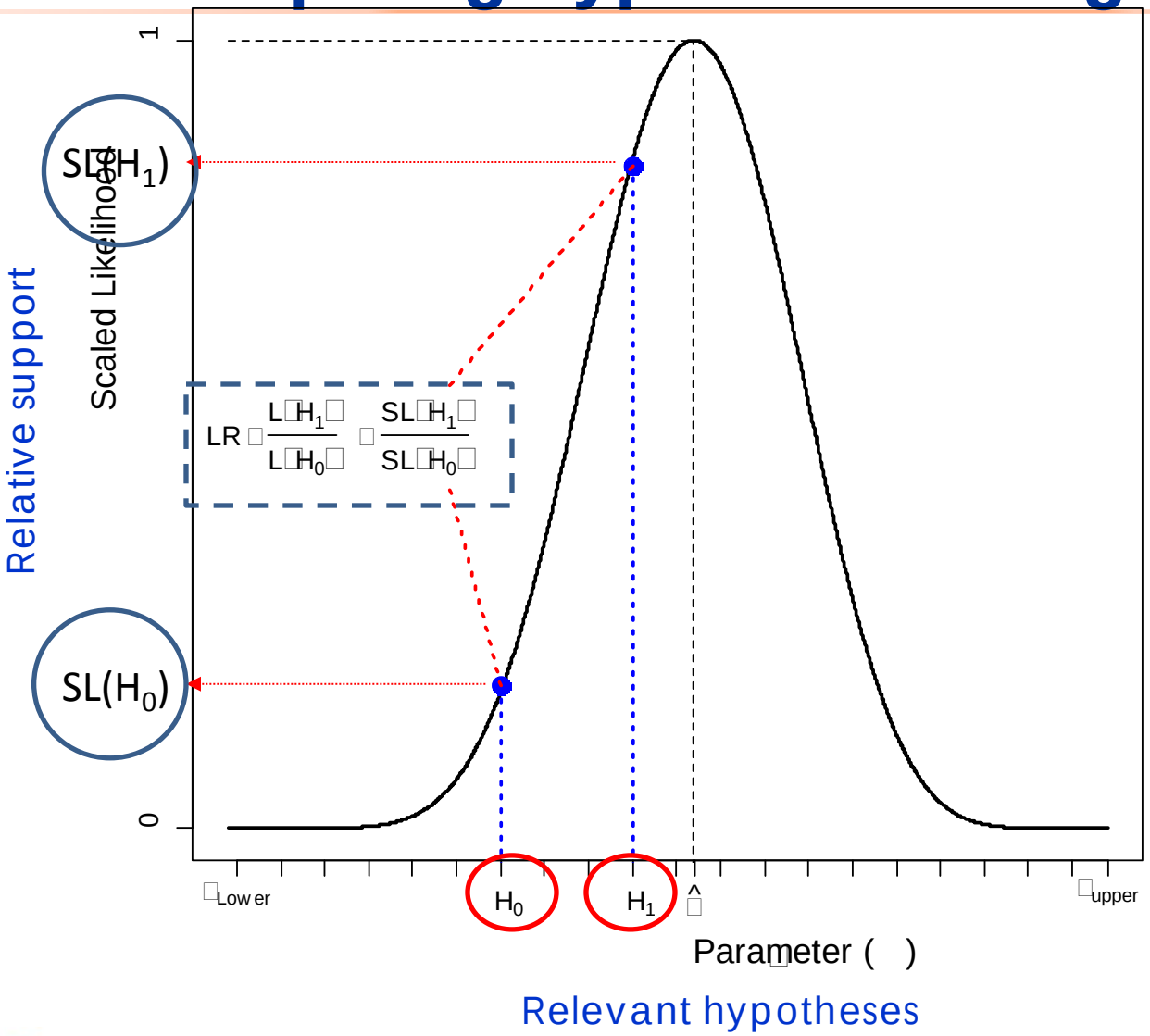


support relative to MSV

Most supported value = MSV

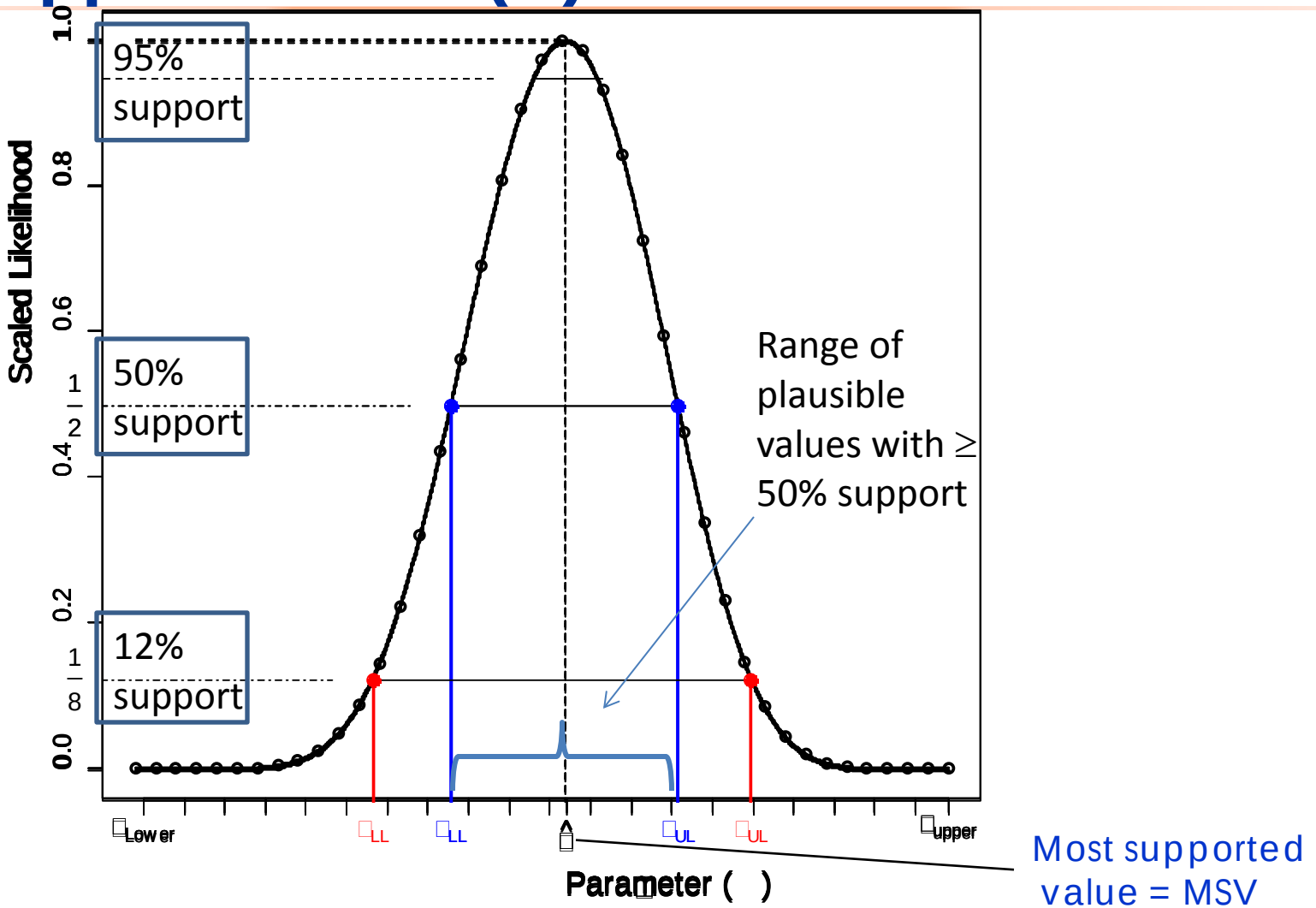
Continuum of possible hypotheses

Quantifying evidential support for any pair of competing hypotheses using the LR



Set of plausibly supported values – the Support Interval (SI)

Given the observed DATA



Meaning of the support interval

- If the 50% SI is (L, U), then hypotheses L & U at the interval limits have only half (50% i.e. scaled LR = 0.5) as much evidence support as the most supported hypothesis (MSV)
- Hypotheses values that are >L but <U have increasingly more than 50% support until the MSV is reached
- The level of support tailored to the objective
  - 90% or more (?) for confirmatory analysis
  - 50% or less (?) for exploratory analysis
  - 12% SI = 95% CI, but this does not mean that the CI can be given an evidentiary interpretation within its own paradigm

Using the Support Interval to test plausibility of specific hypotheses

- RULE: For a pre-specified level of support, if the specific hypothesis of interest is NOT contained within the Support Interval, then it is NOT sufficiently (significantly) supported by the data (compared to MSV)

- Typical hypotheses tested are:
  - NULL hypothesis of no effect
  - Clinically significant treatment effect
  - Clinically non-inferior effect

DATA from an RCT of Budesonide vs Pbo for 21-day asthmatic relapse prevention

- Data from Rowe *et al*

Treatment	Relapse within 21 days		
	Yes	*No*	*Total*
Placebo	**23**	**71**	94
Budesonide	**12**	**82**	94

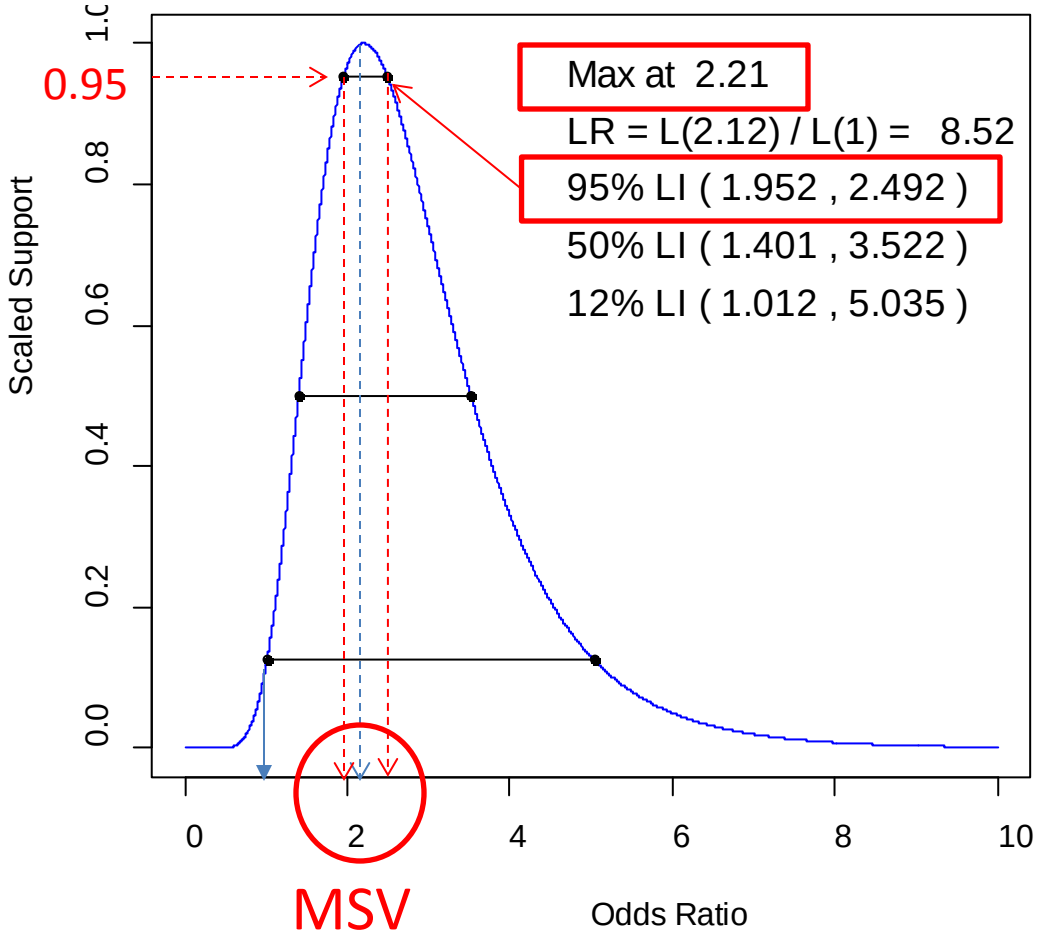
The treatment effect will be quantified as an odds ratio (Pbo vs Budesonide) with possible values for the treatment effect ranging from 0 - ∞

Evidence-based data analysis & reporting

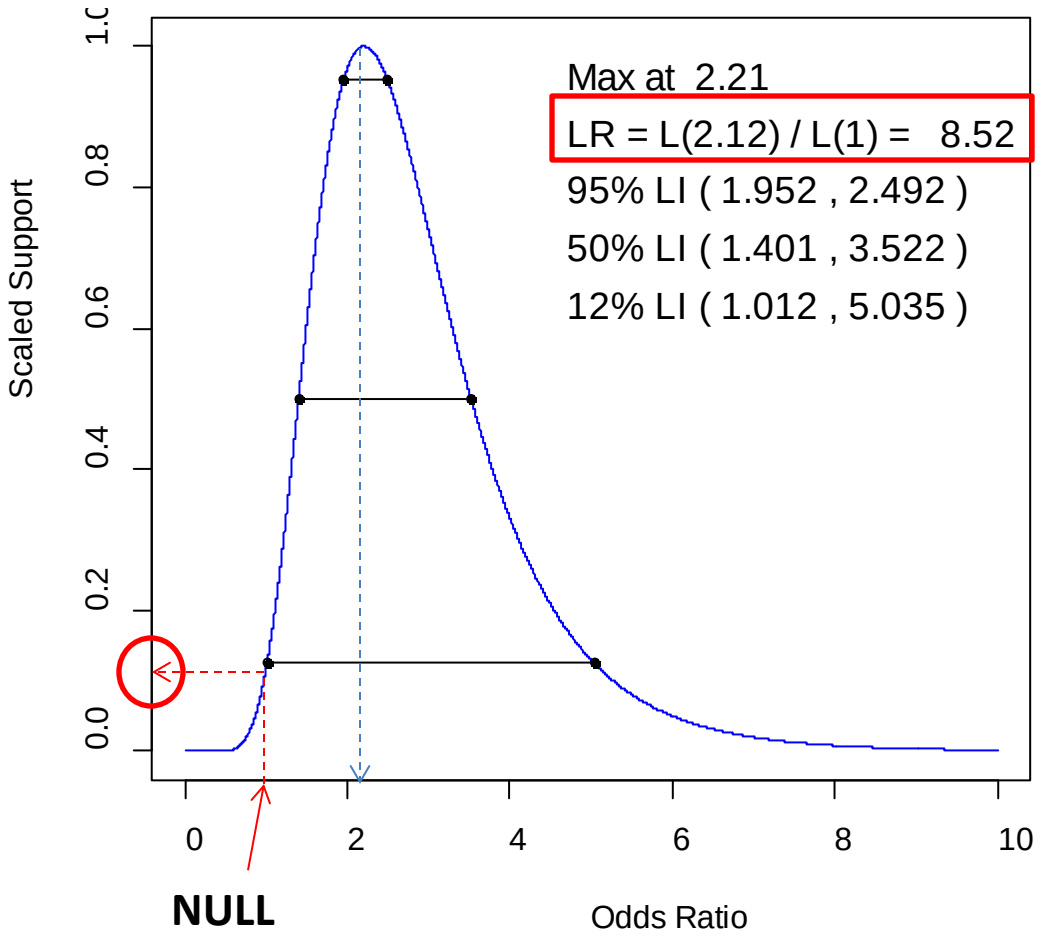
- To report
  - the evidence for all possible values for the treatment effect (OR) compared to the MSV
  - the most supported value for the treatment effect
  - the plausible range of values for the treatment effect at a pre-specified support level, say 95%
  - whether the null hypothesis of no treatment effect (OR = 1) has significant support in comparison to the MSV
  - whether a clinically significant effect (say OR = 2) has significant support in comparison to the MSV

The Support Curve for the possible treatment effect ORs (PBO vs BUD)

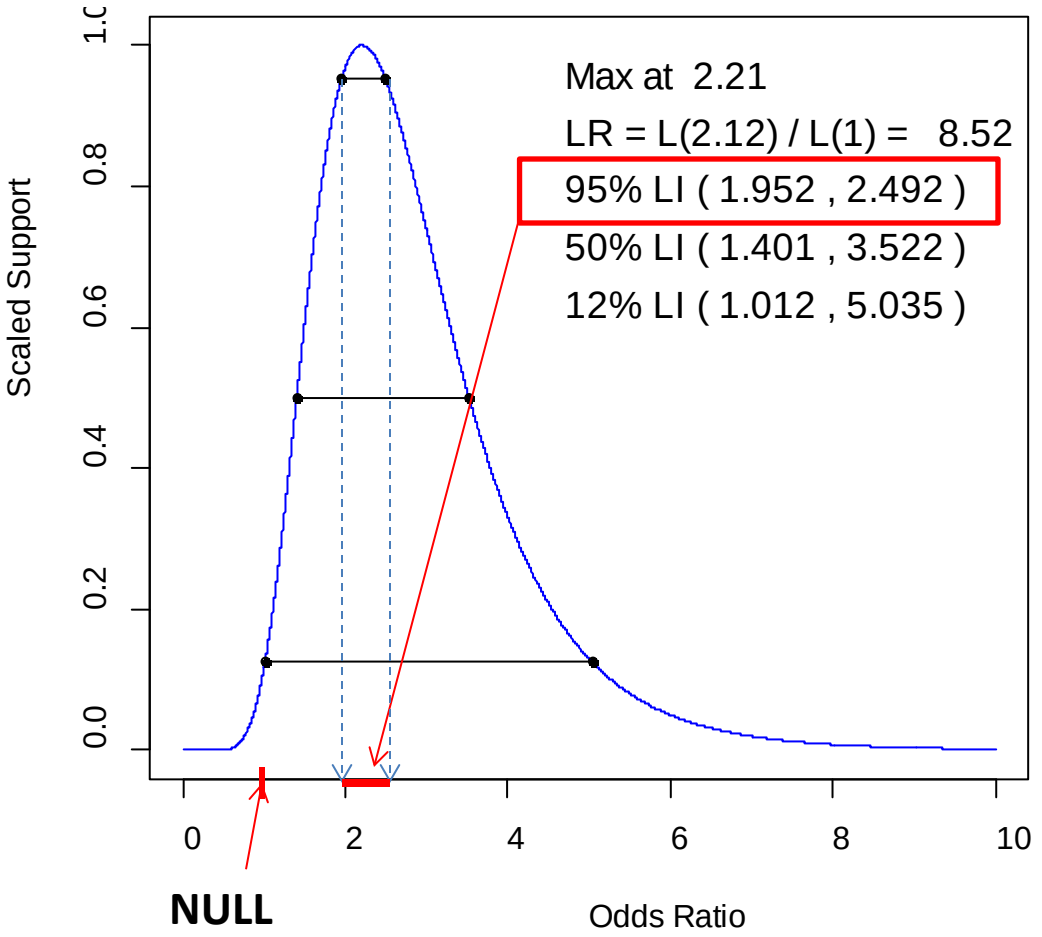
- The MSV and 95% SI estimates are …



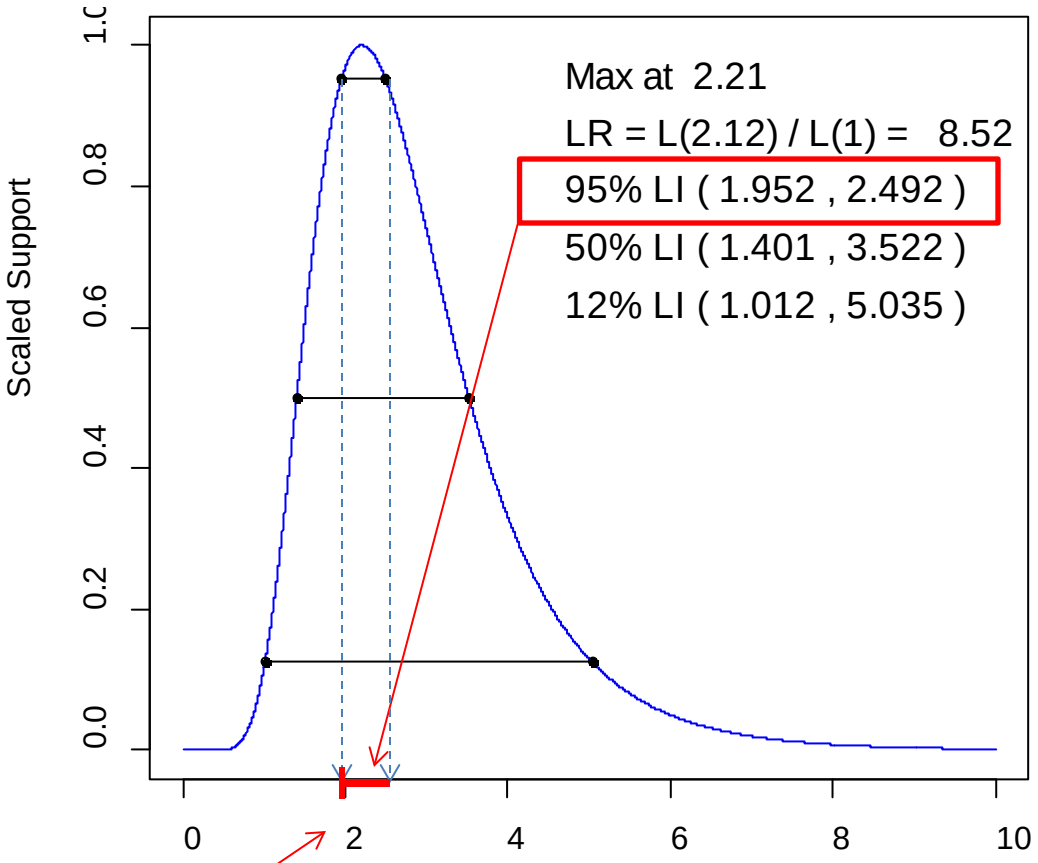
How much evidence is there for the NULL vs the MSV?



Is there significantly strong evidence for no Rx effect?



Is there significantly strong evidence for a clinically important (OR=2) effect?



Comparison with classical approach

Characteristics	Classical Approach	Likelihood Approach
Data model	✓	✓
Estimates of effects	Moments estimates or Maximum likelihood estimate (MLE)	Most supported value (MSV) = MLE
Intervals	Confidence interval (misinterpreted)	Support interval
Strength of evidence	P-value (flawed)	Likelihood ratio
Significance	Statistical	Evidentiary

Comparison with classical approach

Characteristics	Classical Approach	Likelihood Approach
Data model	✓	✓
Estimates of effects	2.21	2.21
Intervals	95% CI (1.03 – 4.77)	95% SI (1.89 – 2.51)
Strength of evidence	0.049	8.52 (MLE vs NULL)
Significance	result is statistically significant	NULL hypothesis has no significant evidential support Clinically significant effect has significant evidential support

The place of evidence measurement within the Ev-B Appraisal & Practice cycle

Stage 3 – Quantify the statistical evidence of the data for competing hypotheses:

LIKELIHOOD approach

Stage 2 – Appraise internal validity: Does **study design, execution minimize** the risk of bias (RoB)

Stage 4 – Reassess your subjective belief: **Bayesian paradigm** - using the evidence from *Stage 3* and your prior belief

Stage 1 – Appraise relevance:

Are the **focused elements** (**PICO**) relevant to my problem?

Stage 5 – Make a decision: **Decision-theoretic methods** – using updated beliefs & external considerations e.g. cost, utilities, patient values, regulations etc

Recommendations for evidence-based reporting of results

- A support curve of the data evidence for all possible hypotheses
- The hypothesis best supported by the data
- Likelihood ratios for important pairs of competing hypotheses
- Support intervals of plausible alternative hypotheses, with support levels appropriate to the analysis objective i.e. confirmatory or exploratory analysis

Acknowledgments

Colleagues

- Pryseley Assam
- Rehena Ganguly
- John Allen
- John Rush

Writings

- Richard Royall
- Kenneth Goodman
- AWF Edwards

DUKE NUS
GRADUATE MEDICAL SCHOOL SINGAPORE

Thank you

Partners in Academic Medicine

SingHealth DUKE NUS GRADUATE MEDICAL SCHOOL SINGAPORE

www.duke-nus.edu.sg