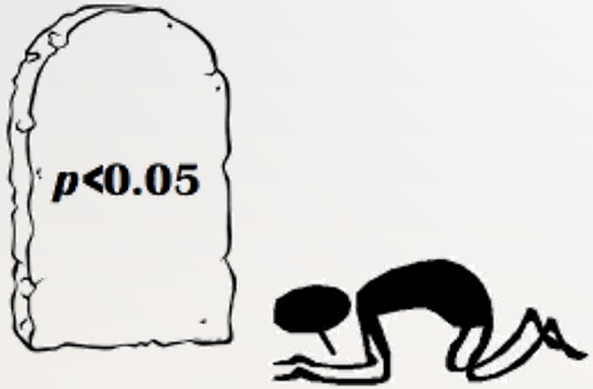
The Reproducibility Crisis: *p*-value Misuse or Lack of an Evidence Measure?

Yes!



Jeffrey D. Blume, PhD
School of Data Science
University of Virginia

Reproducibility

- Reproduce what?
  - Convincing results ?
  - Study conclusions ?
  - Subsequent decisions ?
  - Statistical Evidence ? (how to define?)

- Complications
  - Lack of consensus study goals
  - Lack of consensus of what should be reproducible
  - What does it mean for a random variable to be reproducible?

- Simpler approach may be to be less granular
  - Data support alternative, null or were inconclusive

Evidential metrics

Example:
Diagnostic Test

1. Measure of the strength evidence
   - → Axiomatic and intuitive justification
   - → Summary statistic, yardstick

   Positive Test
   Negative Test

2. Propensity to collect data that will yield a misleading #1
   - → Error rates
   - → Properties of the study design (!)

   Sensitivity
   Specificity

3. Probability that an observed #1 is misleading
   - → False Discovery rate, False Confirmation rate
   - → Chance that an observed result is mistaken
   - → Properties of the observed data (!)

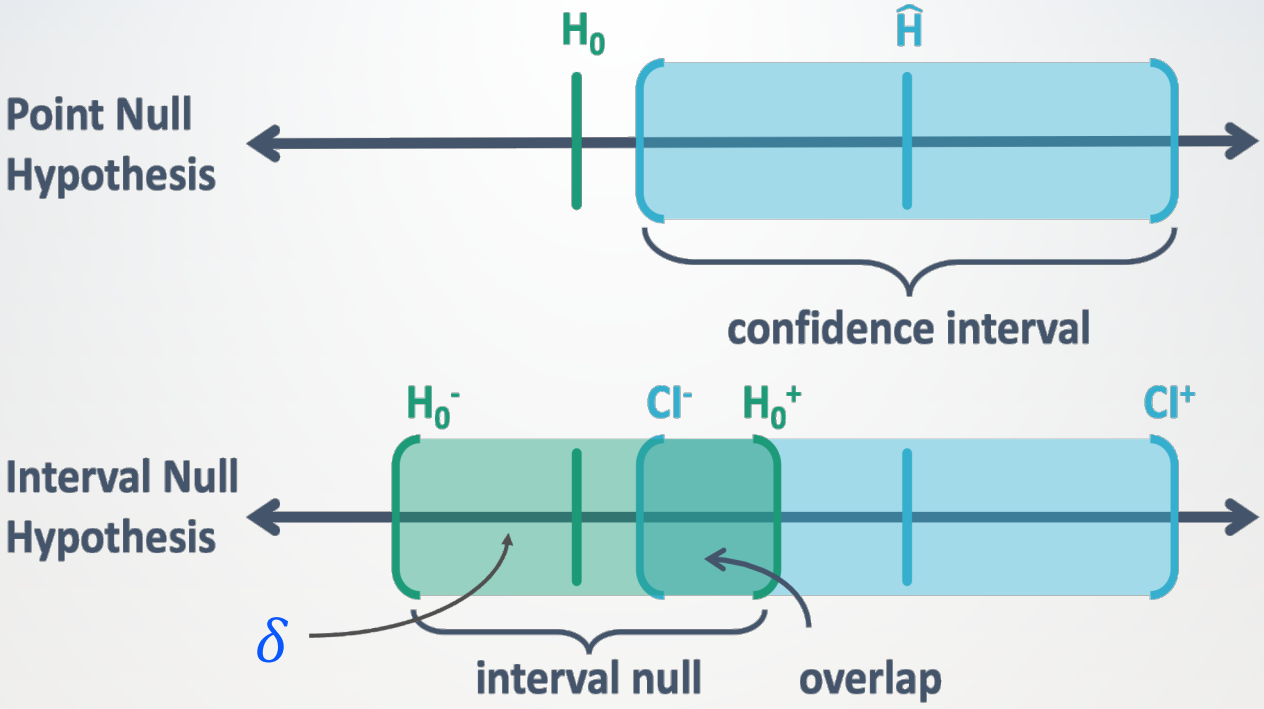
   PPV
   NPV

This is now

Evidential Metric	What it measures	Hypothesis Testing	Significance Testing
1	strength of the evidence	Absent	Tail-area probability (p-value)
2	propensity for study to yield misleading evidence	Tail-area probability (error rates)	Absent
3	propensity for observed results to be misleading	misinterpret #2	misinterpret #1

- **Confusion:** the tail-area probability is used to measure *three* distinct quantities.
- **Reproducibility:** Depends on the intended metric

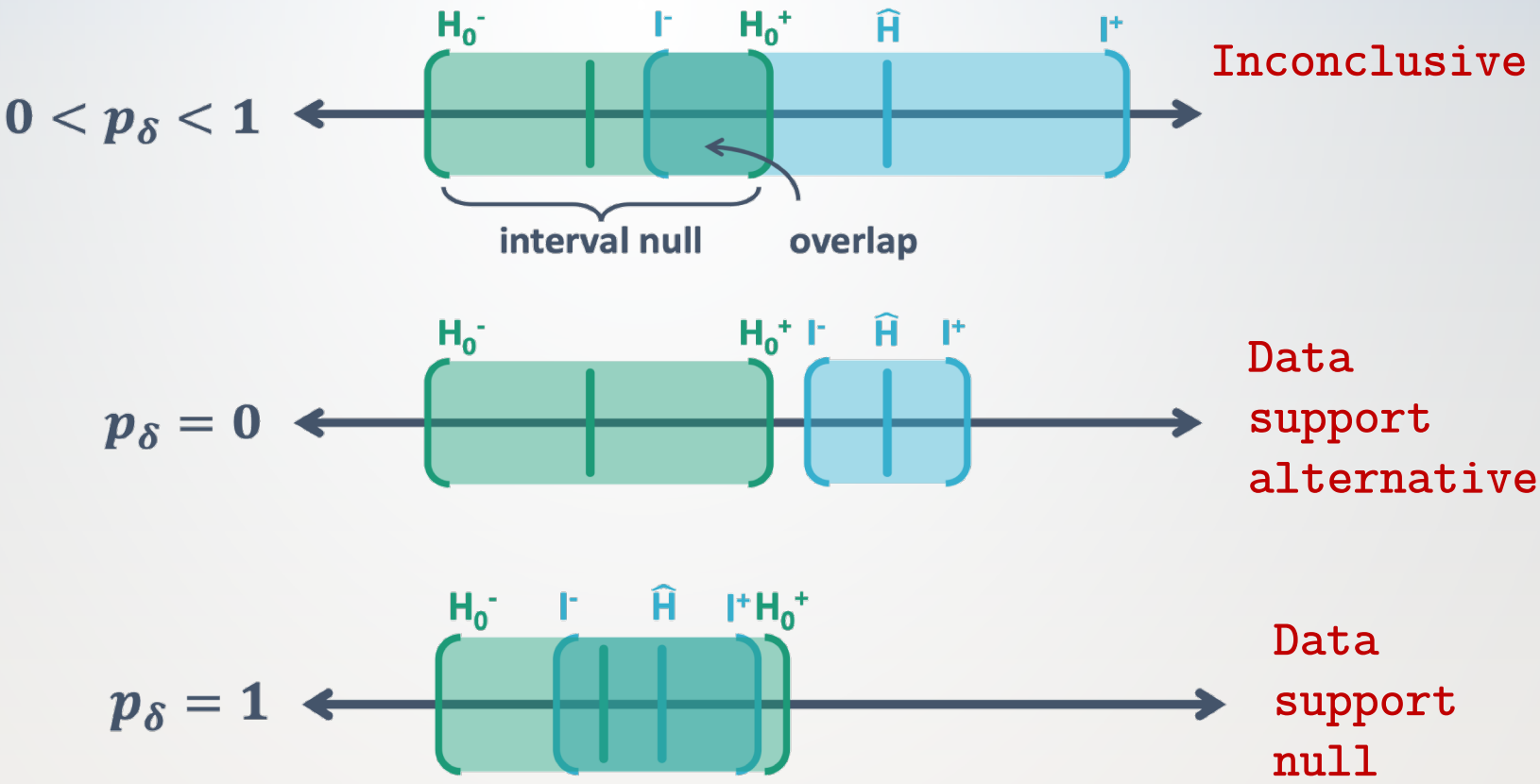
Move to interval nulls



Point null hypothesis H_0 and interval null hypothesis $[H_0^-, H_0^+]$

Data-supported hypothesis $\hat{H}$ and confidence interval $[CI^-, CI^+]$

Reproduce this



works with confidence, credible, and support intervals

Blume, BFF7 2022

Definition

Second-generation *p*-value (SGPV)

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2|H_0|}, 1\right\}$$

Proportion of data-supported hypotheses that are also null hypotheses

Small-sample correction factor

shrinks proportion to ½ when $|I|$ wide

when $|I| > 2|H_0|$

Second-generation p-value

- Statistical properties detailed in recent pubs
- Retains strict error control
- StatisticalEvidence.com

Evidential Metric	What it measures	Likelihood
1	Summary measure	SGPV (p_δ)
2	Operating characteristics	$P(p_\delta = 0 \mid H_0)$ $P(p_\delta = 1 \mid H_1)$ $P(0 < p_\delta < 1 \mid H)$
3	False discovery rates	$P(H_0 \mid p_\delta = 0)$ $P(H_1 \mid p_\delta = 1)$

Blume, BFF7 2022

The p-value (what it is)

- Number between 0 and 1
- Smaller ⇒ support for an alternative hypothesis
- Larger ⇒ data are inconclusive
- Clinical significance is ignored
- Sample size confounds comparisons
- Interpretation
  - awkward
  - assumes null hypothesis true
  - rooted in inductive reasoning
- Not clear if/when ‘adjustments’ are necessary

The ^p-value (what ~~it is~~)

2nd-generation

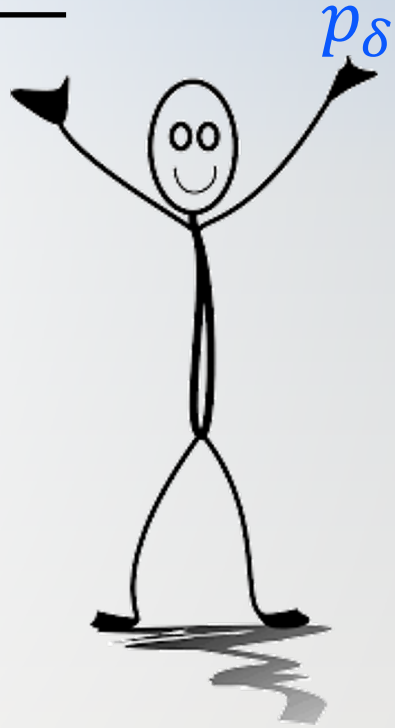
we want

Version 2.0

- ✓ Number between 0 and 1 → near 0 supports alt; near 1 supports null; near $\frac{1}{2}$ inconclusive
- ✓ Smaller ⇒ support for an alternative hypothesis
- Larger ⇒ data ~~are inconclusive~~ support null
- Clinical significance is ~~ignored~~ incorporated
- ✗ ~~Sample size confounds comparisons~~
- Interpretation → Fraction of data-supported hypotheses that are null
  - ~~awkward~~ straightforward
  - assumes ~~null hypothesis true~~ conditions on observed data
  - ~~rooted in inductive reasoning~~ descriptive, summarizes
- ~~Not~~ clear if/when 'adjustments' are necessary

Ideally, never

Second-generation p-value

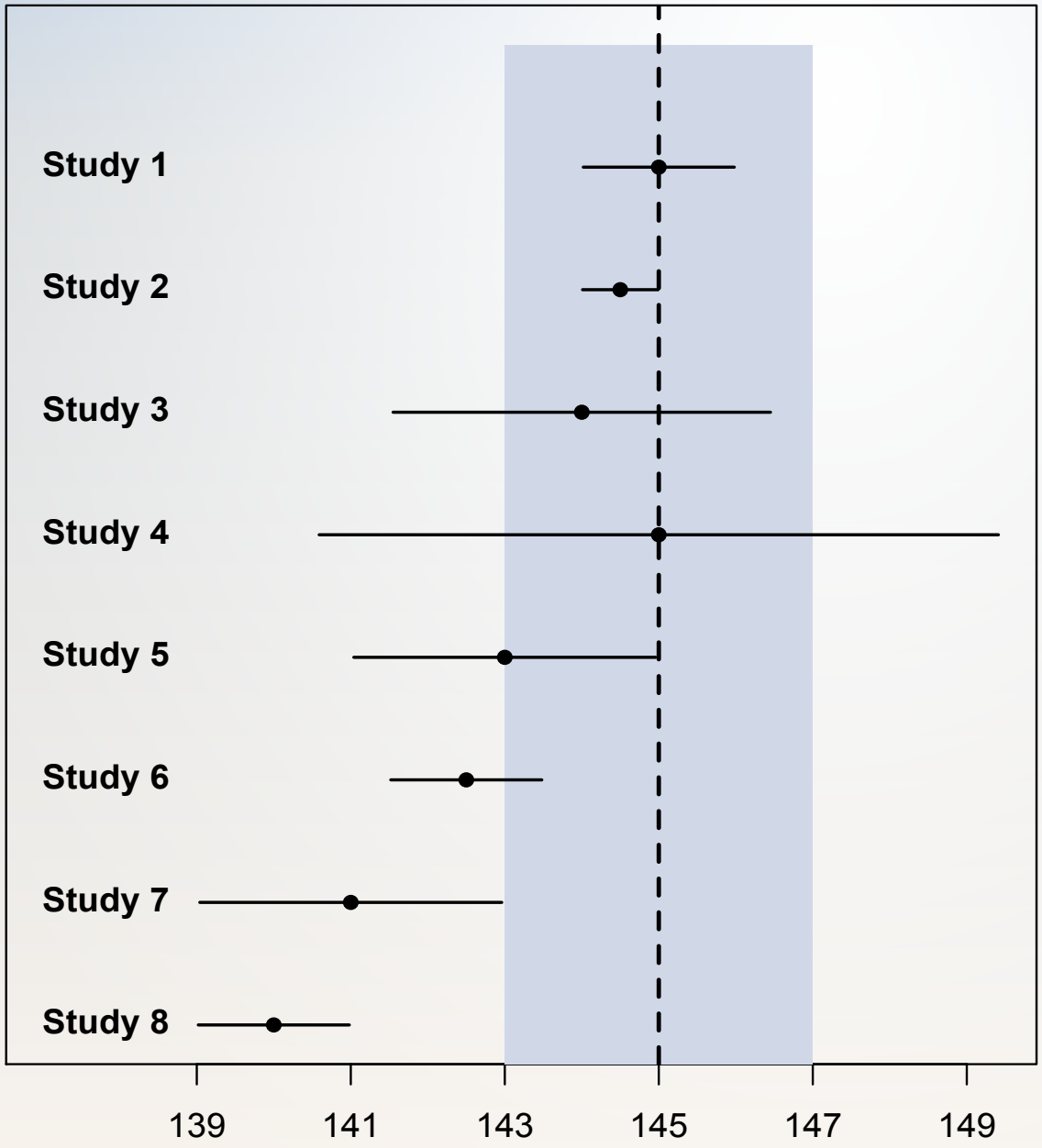


- SGPV is in [0,1] and denoted by p_δ
- δ for scientific significance
  1. $p_\delta = 0 \Rightarrow$ null incompatible with data
  2. $p_\delta = 1 \Rightarrow$ null compatible with data
  3. $0 < p_\delta < 1 \Rightarrow$ data are inconclusive

- Fraction of data-supported hypotheses that are null
- Retains strict error control, all rates $\to 0$

Systolic Blood Pressure

- SBP is reported to the nearest 2 mmHg
- Null Hypothesis: mean SPB is 145 mmHg
- Interval Null hypothesis: mean is 143 to 147 mmHg
- Results from 8 mock studies



p-value	2nd Gen P
1	1
0.0455	1
0.4237	0.7041
1	0.5
0.0455	0.5
<0.0001	0.2499
<0.0001	0
<0.0001	0

Blume, BFF7 2022

Statistical Properties

Suppose interval I has coverage probability 1-α, then

Three 'Error' Rates

1. $P(p_\delta = 0 | H_0) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
2. $P(p_\delta = 1 | H_1) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
3. $P(0 < p_\delta < 1 | H)$ controlled through sample size

Will not examine today

Two False Discovery Rates

1. $P(H_0 \mid p_\delta = 0)$
2. $P(H_1 \mid p_\delta = 1)$

Will graph to illustrate, if time allows

False discovery rates

- Impact of α=0.05 vs α=0.05/7128 (7128 comparisons)

- False Discovery Rate (FDR)

Error rates

$$P(H_0|p < \alpha) = \left[1 + \frac{(1-\beta)}{\alpha} r\right]^{-1}$$

- False Confirmation Rate (FCR)

$$P(H_1|p > \alpha) = \left[1 + \frac{(1-\alpha)}{\beta} \frac{1}{r}\right]^{-1}$$

Error rates

$r = P(H_1)/P(H_0)$

False discovery rates

- Second-generation p-values
- False Discovery Rate (FDR)

Error Rates

$$P(H_0|p_\delta = 0) = \left[1 + \frac{P(p_\delta = 0|H_1)}{P(p_\delta = 0|H_0)} r\right]^{-1}$$

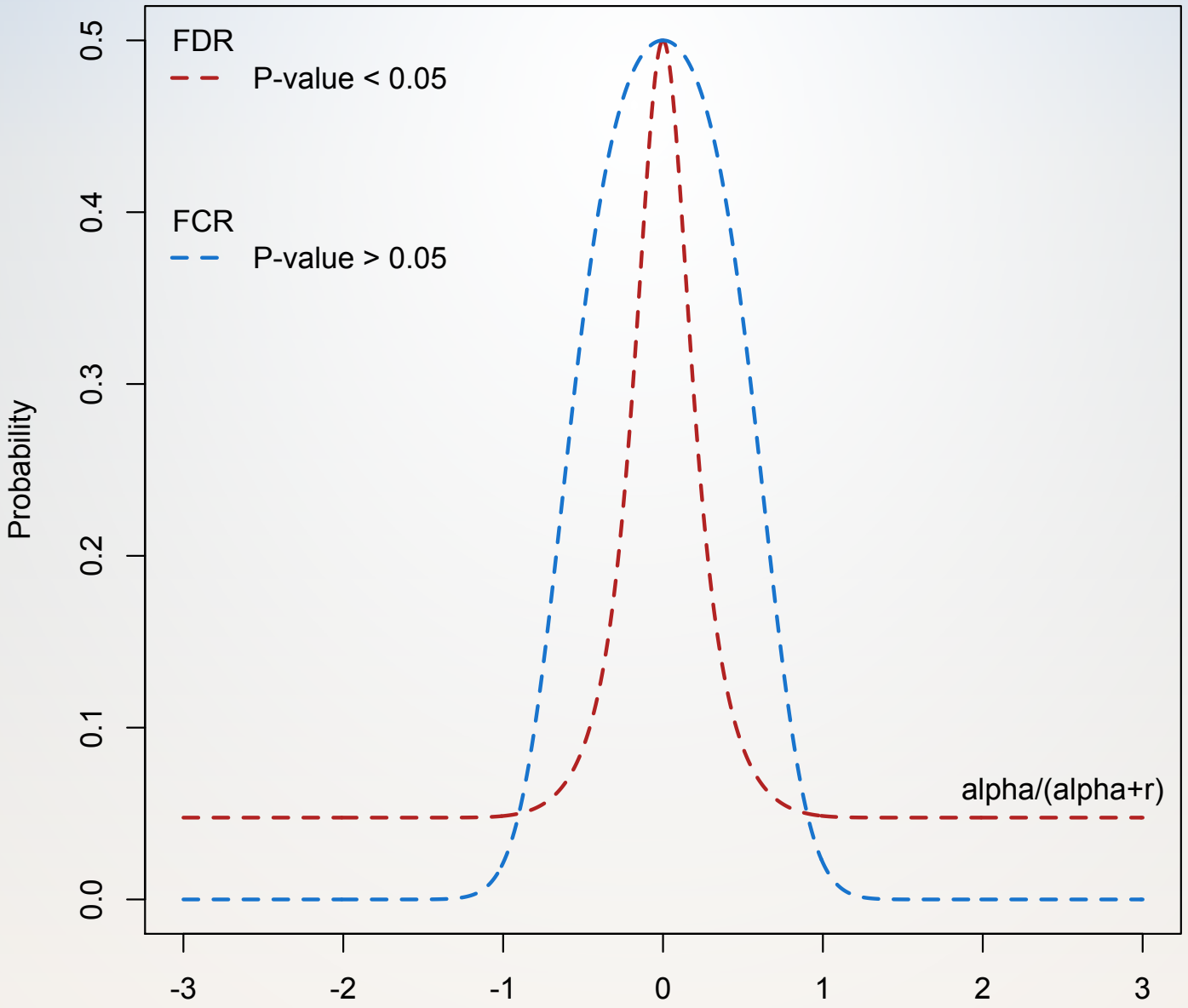
- False Confirmation Rate (FCR)

$$P(H_1|p_\delta = 1) = \left[1 + \frac{P(p_\delta = 1|H_0)}{P(p_\delta = 1|H_1)} \frac{1}{r}\right]^{-1}$$

Error Rates

$r = P(H_1)/P(H_0)$

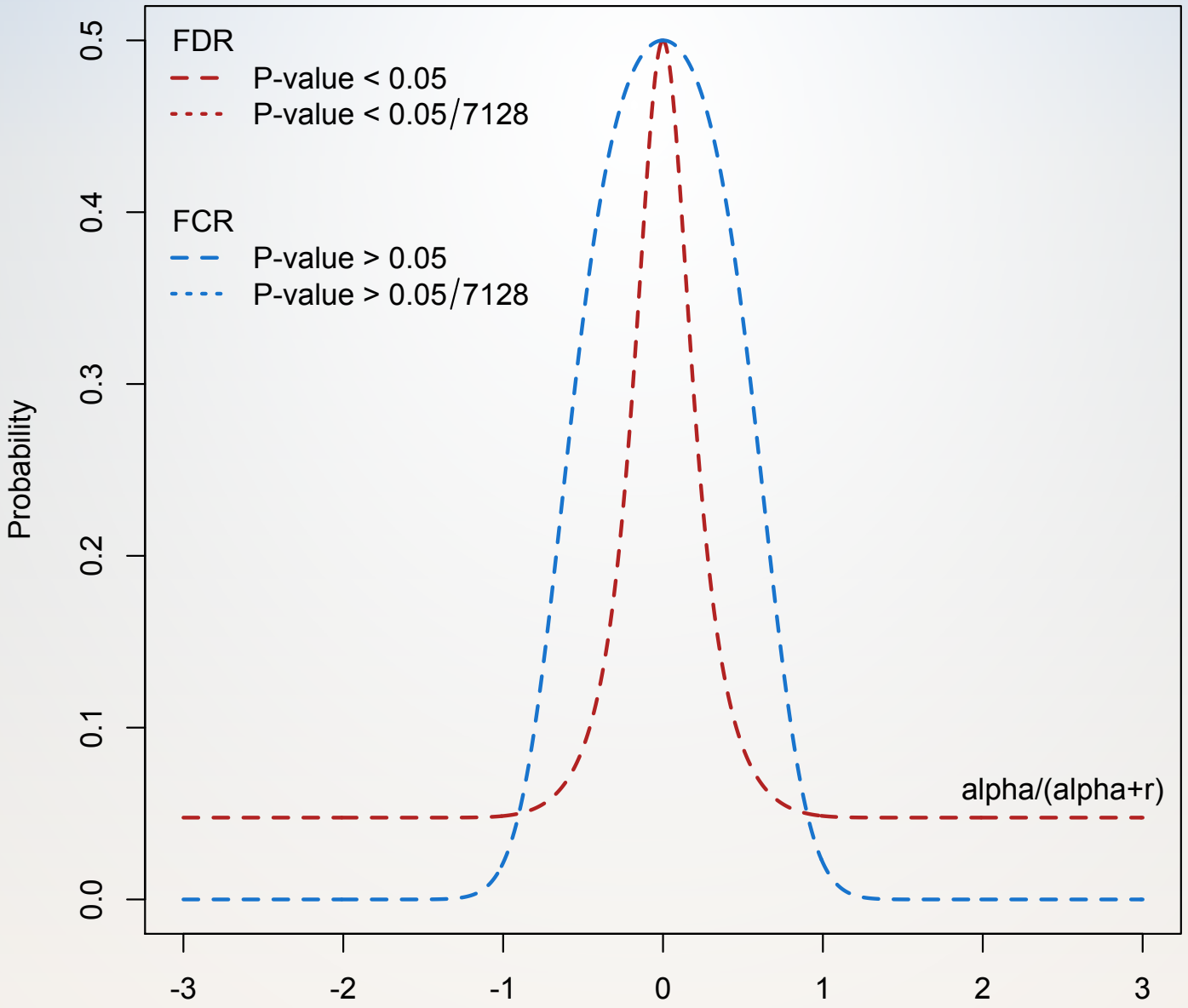
False discovery and confirmation rates



Blume, BFF7 2022

r=1

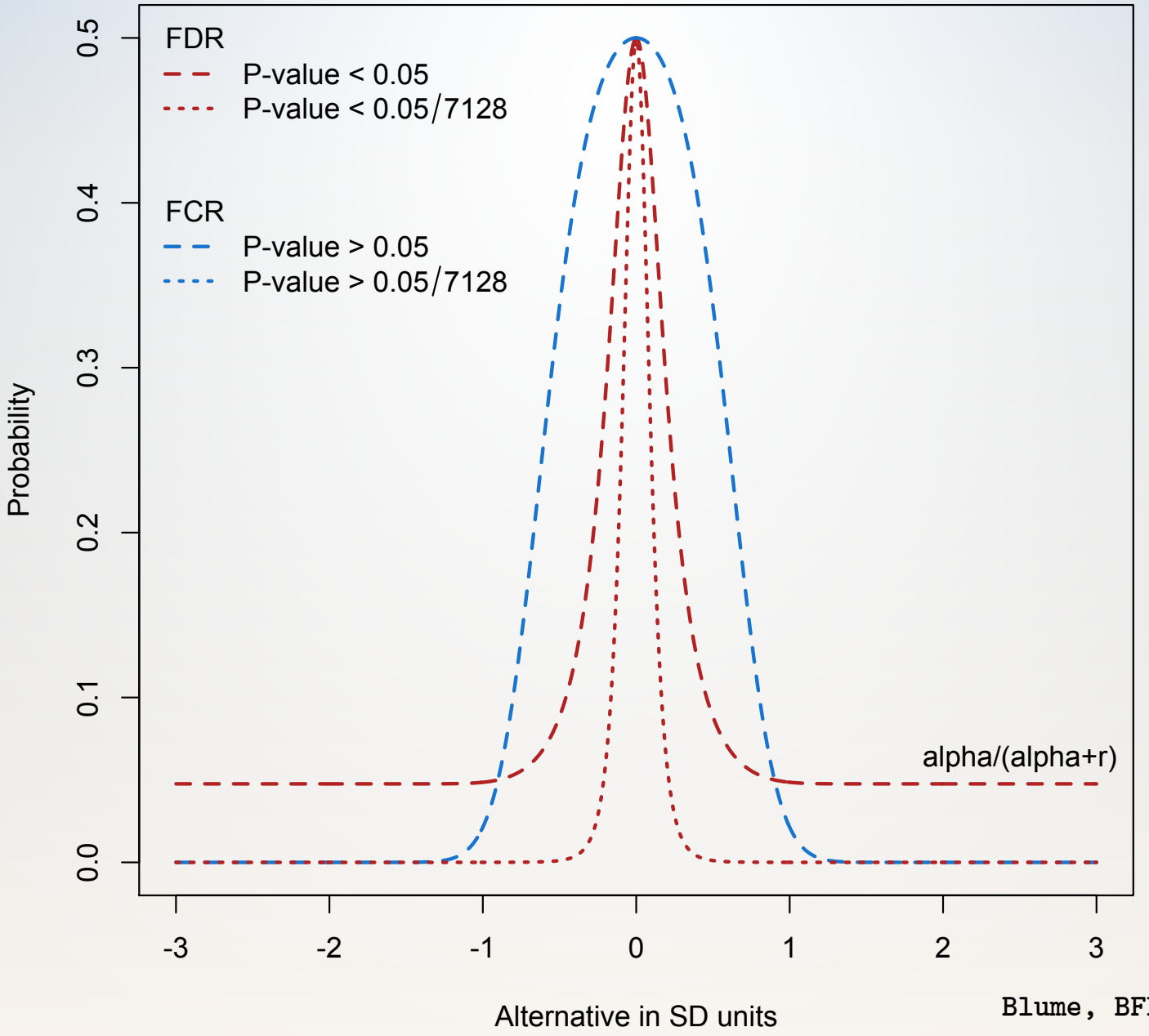
False discovery and confirmation rates



Blume, BFF7 2022

r=1

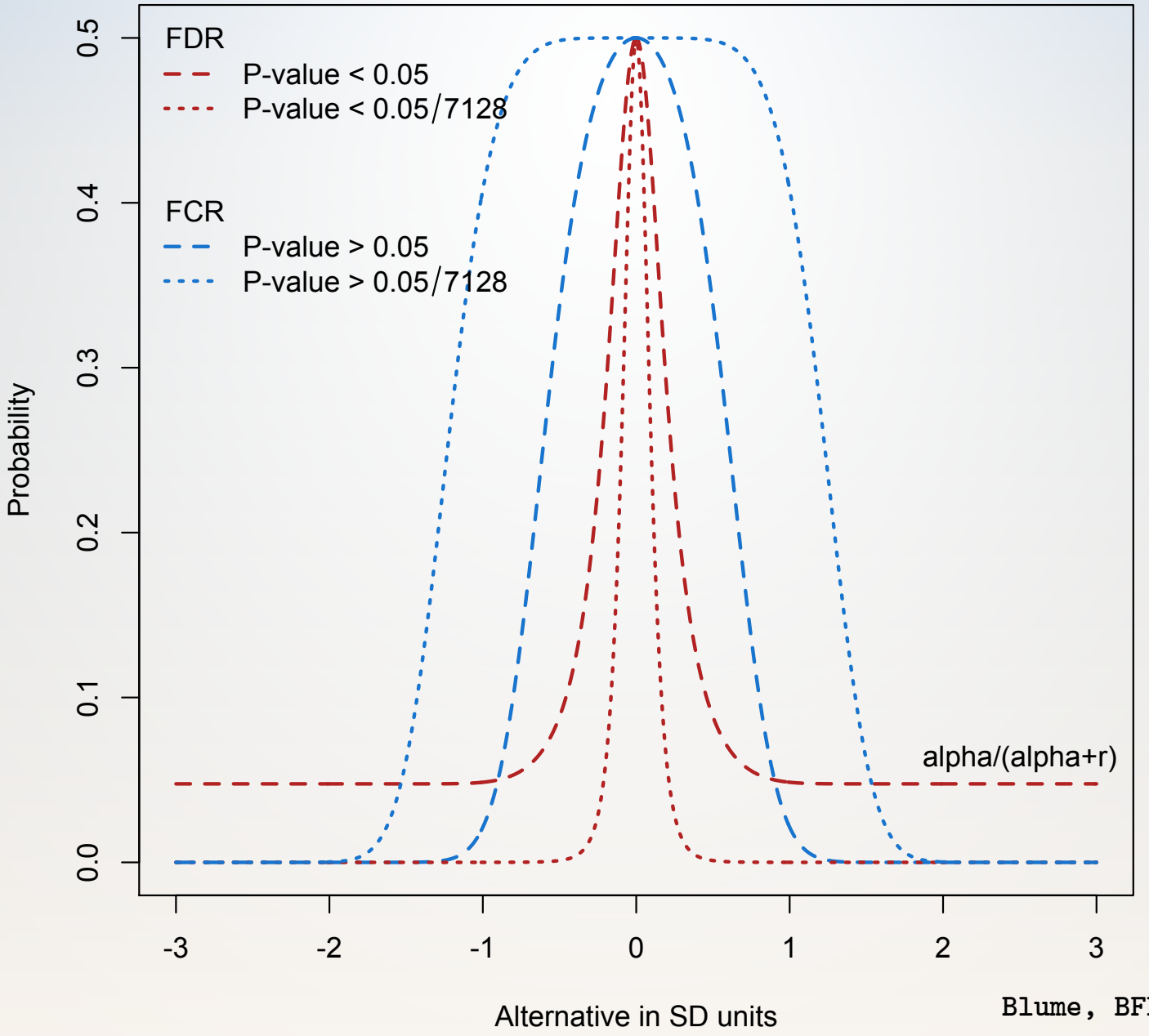
False discovery and confirmation rates



Blume, BFF7 2022

r=1

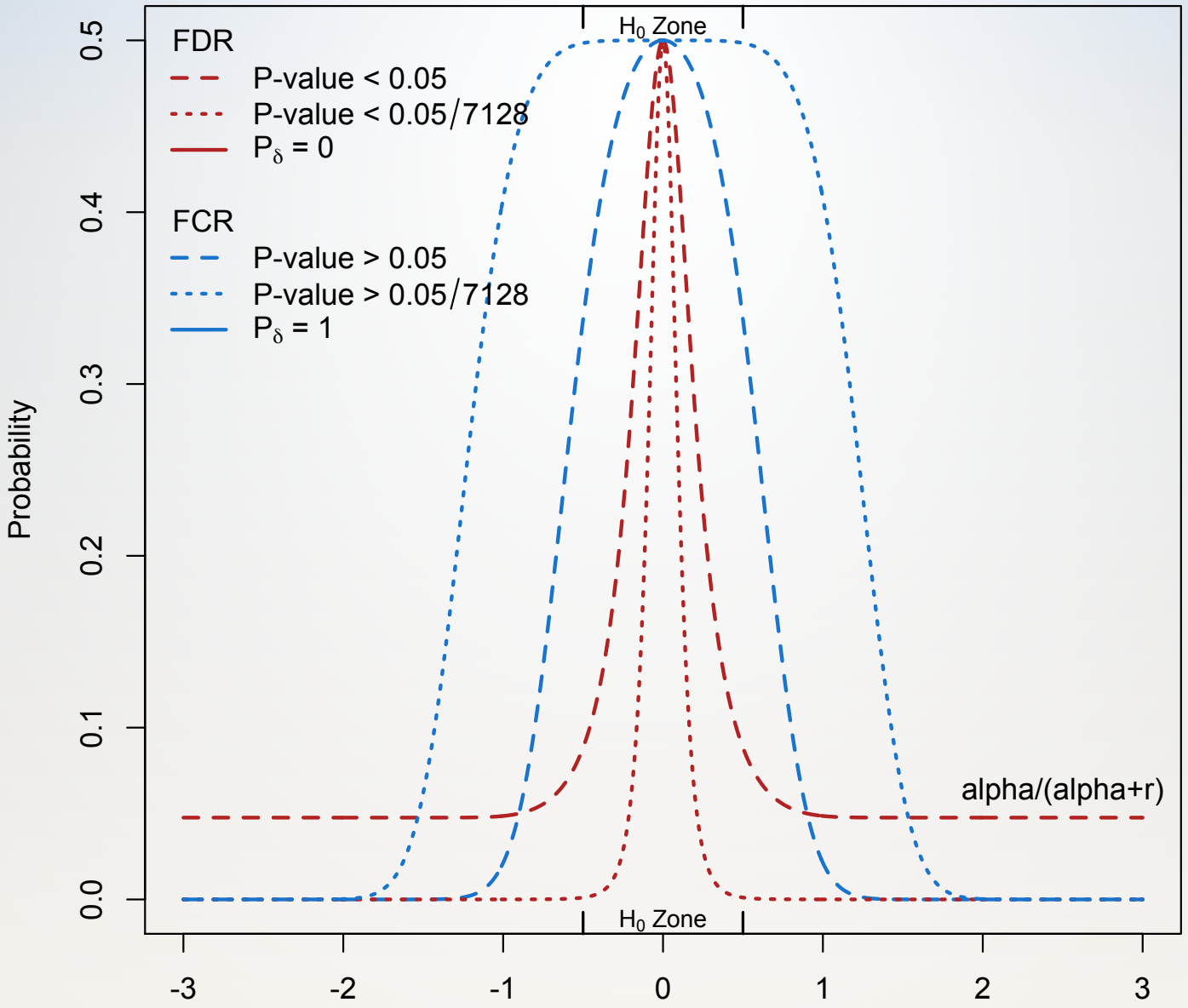
False discovery and confirmation rates



Blume, BFF7 2022

r=1

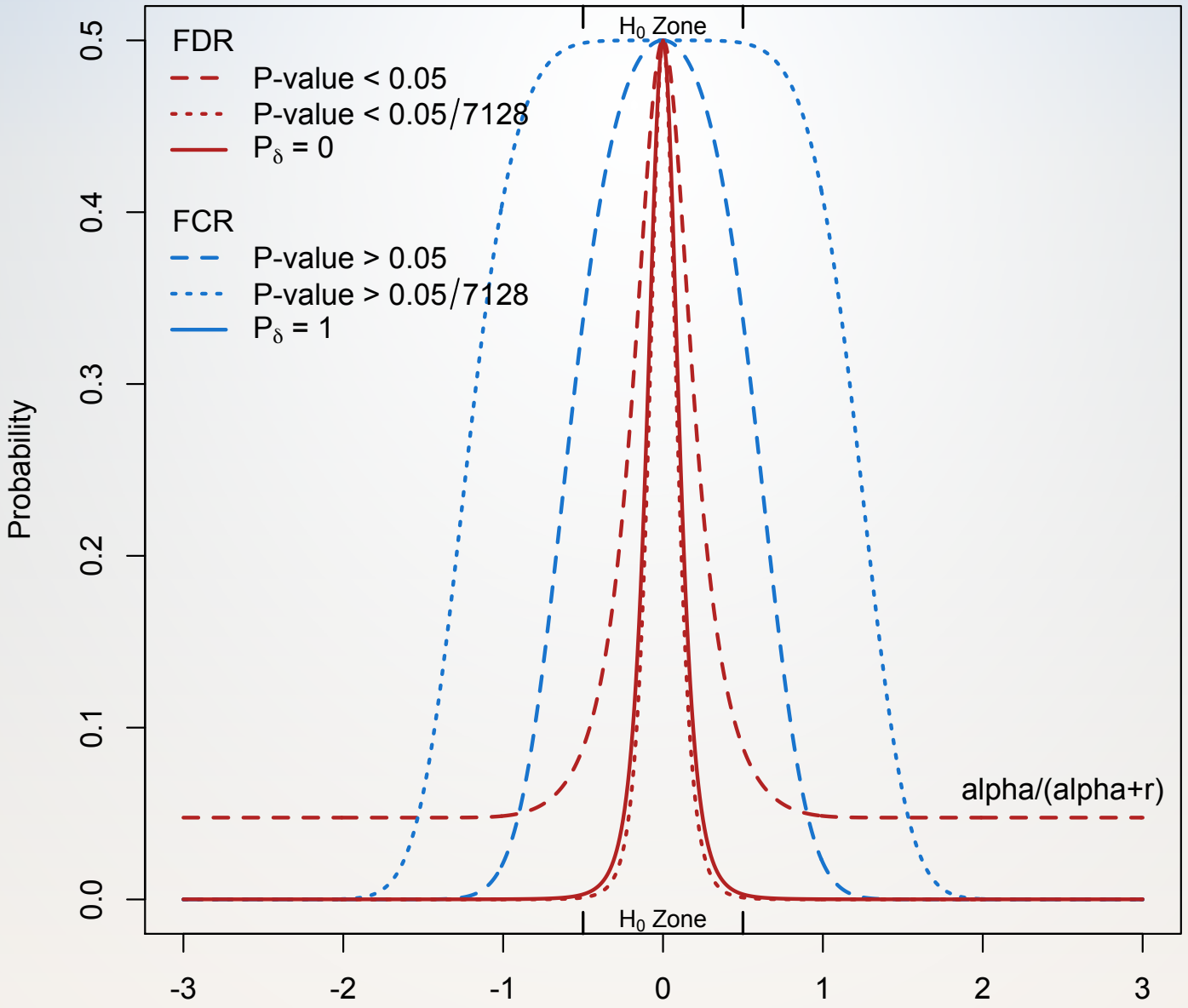
False discovery and confirmation rates



Blume, BFF7 2022

r=1

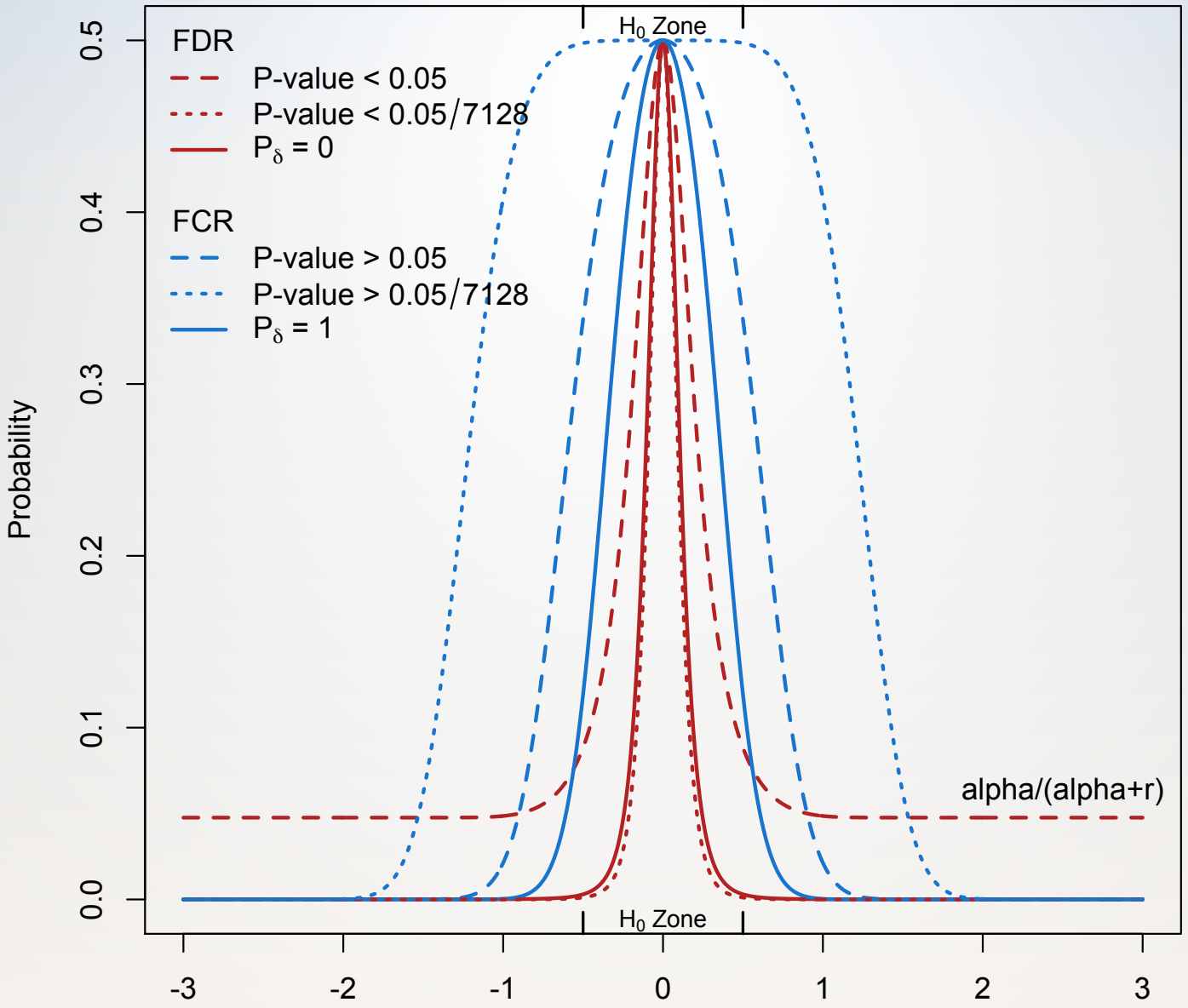
False discovery and confirmation rates



Blume, BFF7 2022

r=1

False discovery and confirmation rates



Blume, BFF7 2022

r=1

Remarks

- Critical to establish evidential metrics & role
- Second-generation p-values…
  - Indicate compatibility with null or alternative
  - Indicate when the data are inconclusive (!)
  - Straightforward to compute and interpret
  - Controls error rate using *science*
  - Reduces the false discovery rate
- Anchoring the scale of the effect size…
  - Eliminates most Type I Errors
  - Improves scientific translation of statistical model

Acknowledgements

- Students
  - Yi Zuo (Variable Selection with SGPVs; TAS 2021)
  - Valerie Welty (FDR and SGPVs)
  - Megan Murray (SGPVs & Equivalence Tests)

- Website / Papers / Code
  - statisticalevidence.com
  - Google "Second-Generation p-value"
  - Cran packages (CRAN.R-project.org w/ vignettes)
    - SGPV
    - ProSgpv
    - FDRestimation

Thank you for your attention.

Questions?