The Strength of Evidence from Statistical Significance and *P*-values

Daniel J. Benjamin
Center for Economic and Social Research,
Behavioral and Health Genomics Center, and Economics Department
University of Southern California

BITSS Panel on Transparency and Reproducibility • 2 August 2017

Related Papers

Bayarri, M.J., Daniel J. Benjamin, James O. Berger, and Thomas M. Sellke (2016). "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses." *Journal of Mathematical Psychology*, 72: 90-103. Invited paper for special issue on "Bayesian hypothesis testing."

Benjamin, Daniel J., and James O. Berger (2016). "Comment: A simple alternative to p-values." *The American Statistician*. Invited comment on "The American Statistical Association Statement on Statistical Significance and p-values."

Benjamin, Daniel J., et al. (2017). "Redefine Statistical Significance." Forthcoming, *Nature Human Behaviour*.

Common Practice: Heuristic-Based

- Reject H_0 if $P < \alpha \equiv 0.05$.
  - Treat such findings as providing strong evidence for a true effect.
- Often, ignore power (except for when required for grant proposals).
- When do power calculations, aim for sample size N that gives power of 0.80.
- (In talk, will remain within paradigm of null hypothesis significance testing.)

Setup

- Test $H_0: \theta = 0$ versus $H_1: \theta = \theta_1$.
  - For simplicity, consider one-sided test.
- Test statistic t.
- P-value at t_{obs} is $P \equiv \Pr(t > t_{obs} | H_0)$.
- Significance threshold $\alpha \equiv \Pr(t > t_{crit} | H_0)$.
  - Implicitly defines t_{crit}.
  - Type I error rate = $\Pr(P < \alpha | H_0) = \alpha$.
- Power $\equiv \Pr(t > t_{crit} | H_1) = \Pr(P < \alpha | H_1)$.
  - Determined by α and sample size *N*.

Pre-Experimental Odds

(based on Wacholder et al., 2004; Ioannidis, 2005; Benjamin et al., 2012; Maniadis, Tufano, and List, 2014; Bayarri et al., 2016)

Fix α. If result is statistically significant, what are the odds of H_1 relative to H_0 ?

$$\Pr(H_1 \mid P<\alpha)$$

$$=\Pr(P<\alpha|H_1)\Pr(H_1)/\Pr(P<\alpha|H_1)\Pr(H_1)+\Pr(P<\alpha|H_0)\Pr(H_0).$$

$$\Pr(H_0 \mid P<\alpha)$$

$$=\Pr(P<\alpha|H_0)\Pr(H_0)/\Pr(P<\alpha|H_1)\Pr(H_1)+\Pr(P<\alpha|H_0)\Pr(H_0).$$

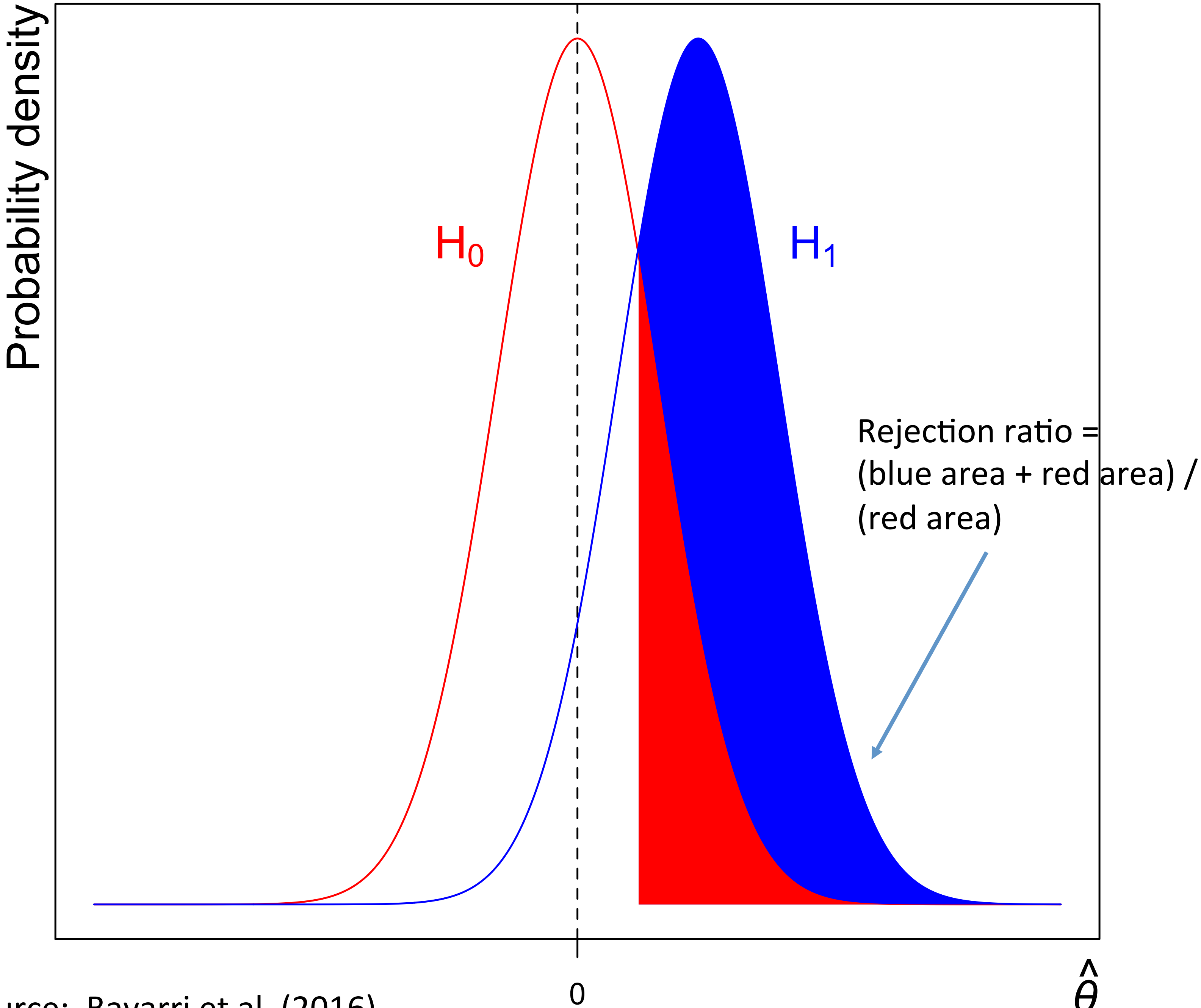
Pre-Experimental Odds

(based on Wacholder et al., 2004; Ioannidis, 2005; Benjamin et al., 2012; Maniadis, Tufano, and List, 2014; Bayarri et al., 2016)

Fix α. If result is statistically significant, what are the odds of H_1 relative to H_0 ?

$$\underbrace{\Pr(H_1 \mid P<\alpha)/\Pr(H_0 \mid P<\alpha)}_{\text{Posterior ratio}} = \underbrace{\Pr(P<\alpha \mid H_1)/\Pr(P<\alpha \mid H_0)}_{\text{“Rejection ratio”}} \times \underbrace{\Pr(H_1)/\Pr(H_0)}_{\text{Prior ratio}}.$$

Rejection ratio $\equiv power/\alpha$ is strength of evidence from statistical significance.



Source: Bayarri et al. (2016).

What's the Prior Odds?

- Of course, varies by context.
- Some evidence indicates ~1:10 (on average) for psychology:
  - Analysis of results from OSC (2015) replication project. (Johnson et al., 2016)
  - Prediction market about outcomes of the OSC replication project. (Dreber et al., 2015)
- Results from experimental economics replication project suggest more like ~1:5 (on average) for experimental economics. (Camerer et al., 2016)

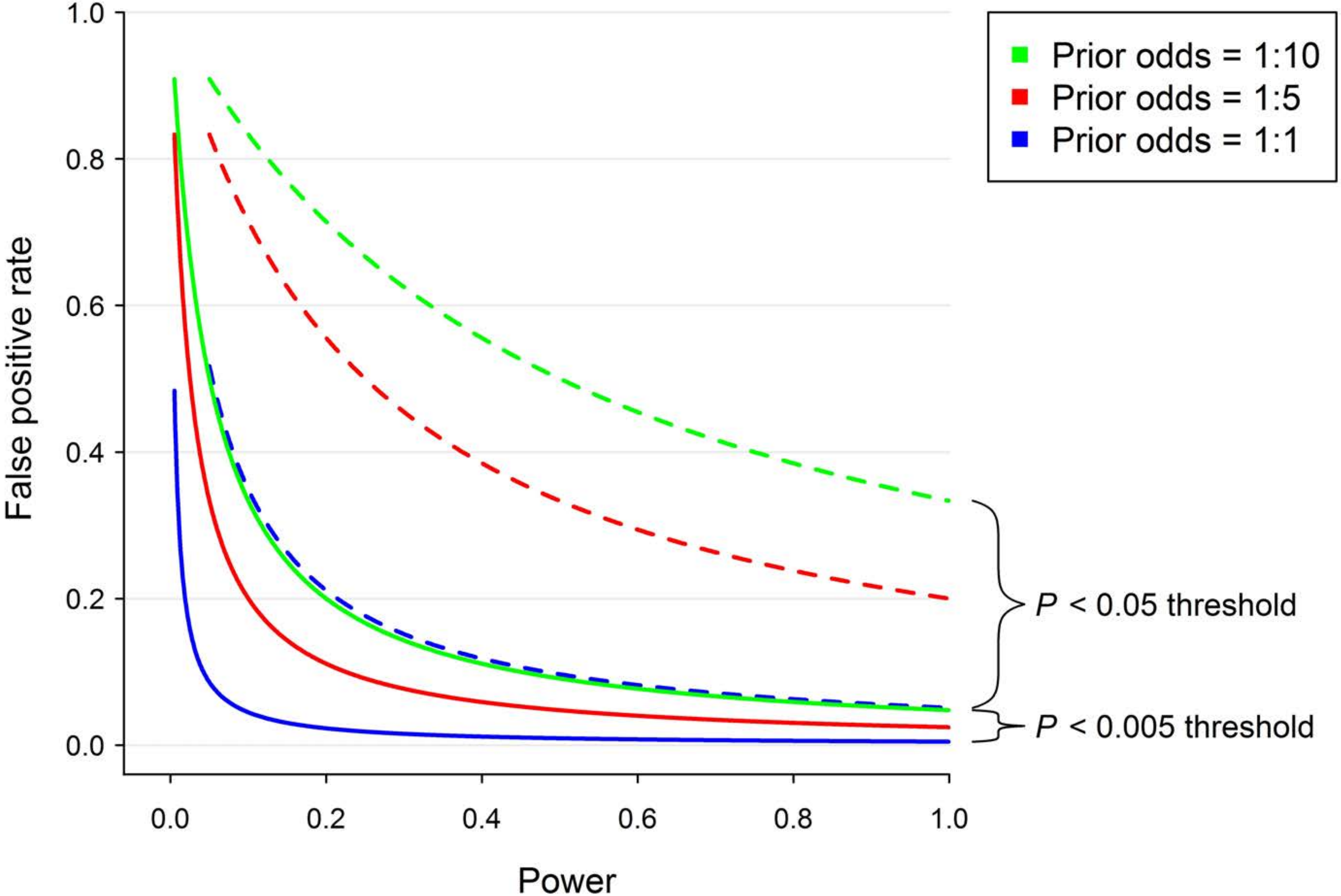
Application: Simple Experiment

- Treatment and control group, each with sample size *N*.
- Effect size *r* = 0.21, “typical” according to meta-analysis of studies in social psychology. (Richard, Bond, and Stoke-Zoota, 2003)

Per-condition *N*	**10**	**20**	**30**	**40**	**50**
Power	0.12	0.16	0.20	0.24	0.28
Rejection ratio	2.4	3.3	4.1	4.8	5.5

Per-condition *N*	**100**	**150**	**200**	**250**	**280**
Power	0.44	0.57	0.68	0.76	0.80
Rejection ratio	8.7	11.4	13.5	15.2	16.0

Source: Bayarri et al. (2016).



Source: Adapted from Benjamin et al. (2017). False positive rate $\equiv \Pr(P<\alpha \; \& \; H_0) / \Pr(P<\alpha)$.

Some Implications

1. Power matters for strength of evidence implied by statistical significance.
   - Common fallacies:
     - Power no longer matters once you've run the experiment.
     - If significant despite low power, even more convincing(!).
   - Other problems with low power: (Gelman and Carlin, 2014)
     - increases probability of wrong sign.
     - increases expected exaggeration of estimated effect size.
2. If prior odds are low, need lower α.
   - Rejection ratio is bounded above by $1/\alpha$ (since power is bounded above by 1).

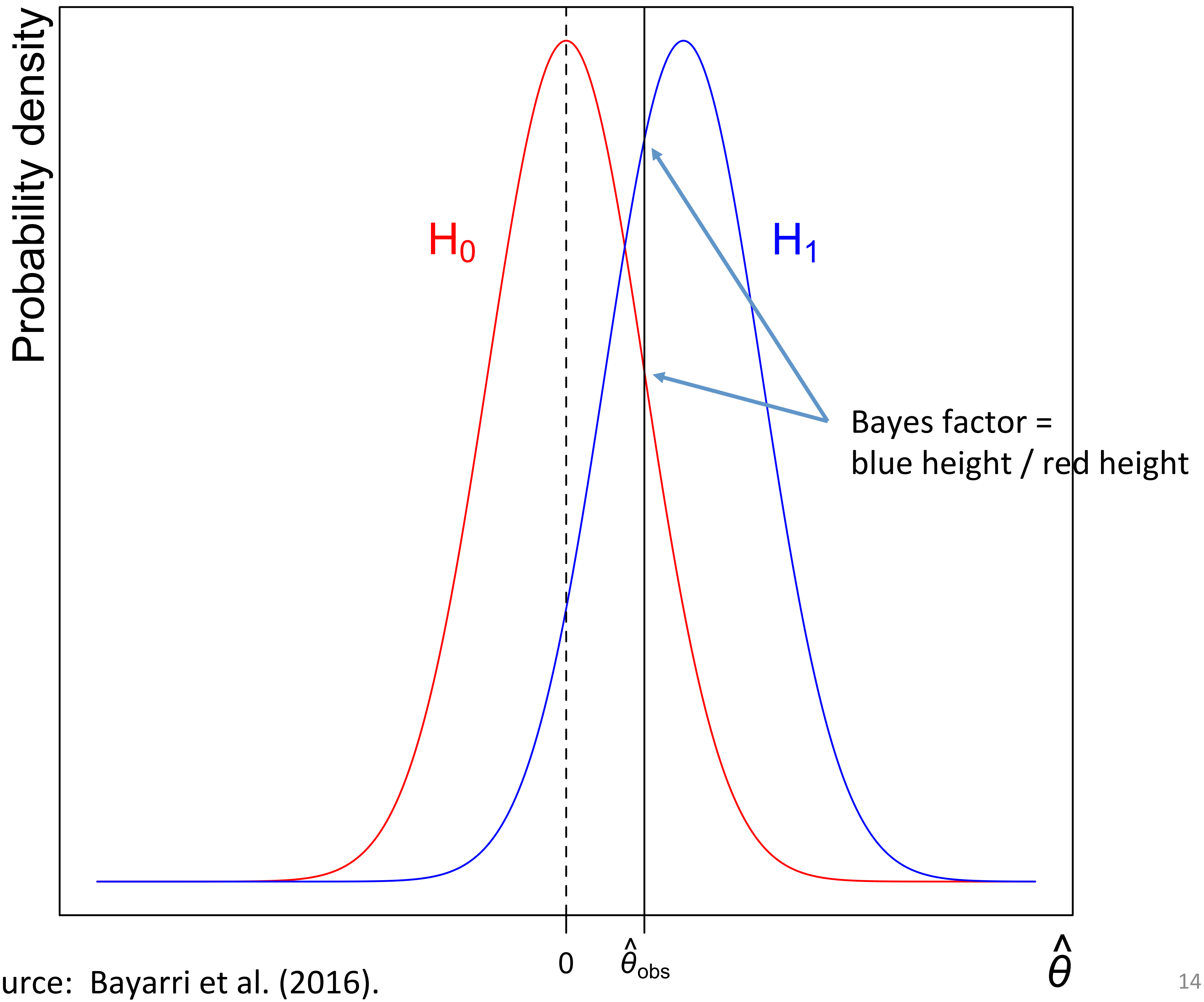
Post-Experimental Odds (Bayes Factors)

If result has *P*-value P_{obs}, what are the odds of H_1 relative to H_0?

$$\Pr(H_1 \mid P = P_{obs}) / \Pr(H_0 \mid P = P_{obs}) = f(P = P_{obs} \mid H_1) / f(P = P_{obs} \mid H_0) \Pr(H_1) / \Pr(H_0).$$

Posterior ratio = Bayes factor × Prior ratio

Bayes factor is the strength of evidence from the observed data.

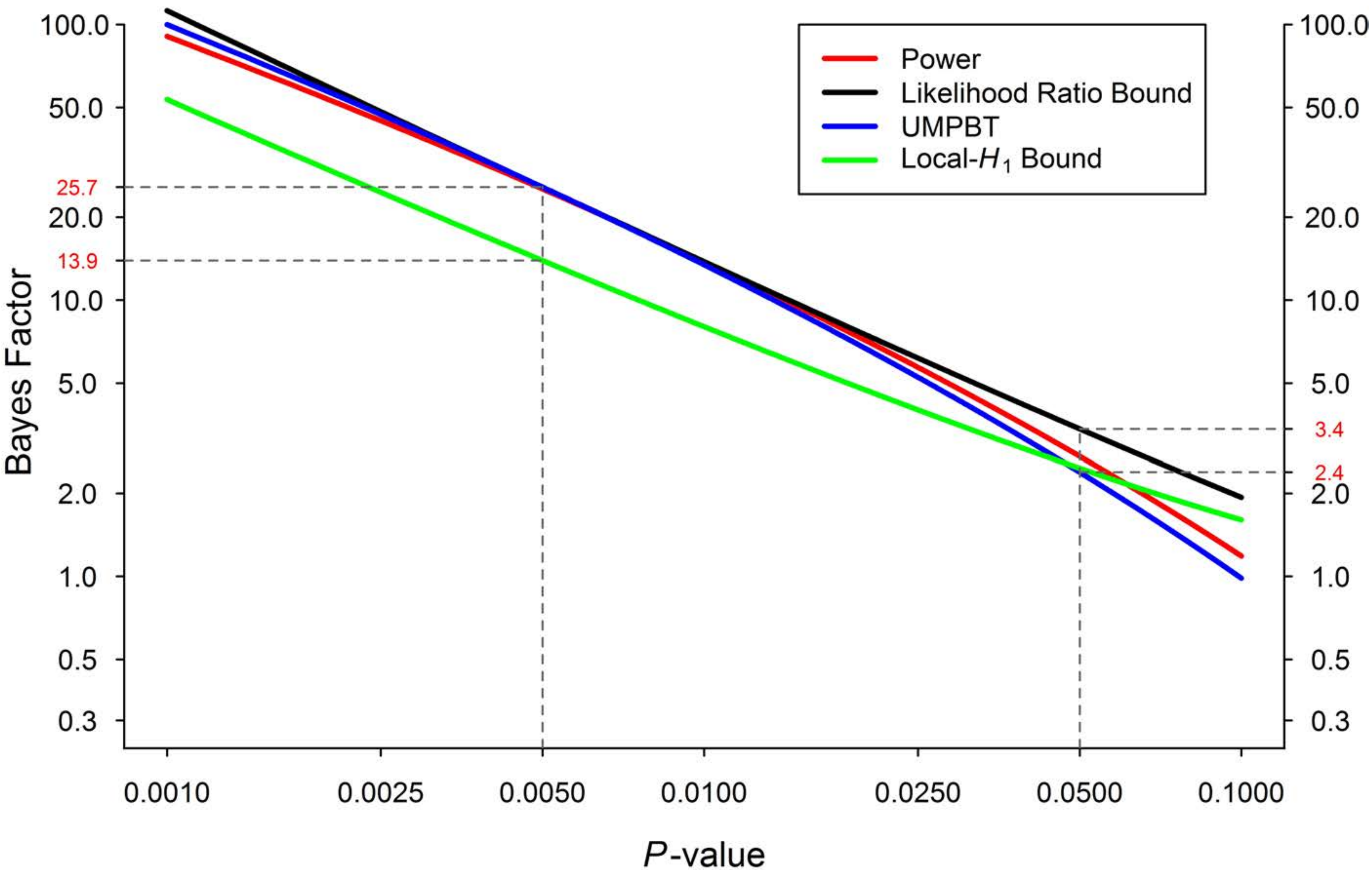


Source: Bayarri et al. (2016).

P-value ↔ Bayes factor ?

- Calculating P-value only requires specifying H_0, but BF requires specifying H_0 and H_1.
- But often, H_1 is not specified.
- Can obtain a correspondence (or bound) under some generic assumptions about H_1.

- For example, consider a draw of a sample mean, $x \sim N(\theta, 1)$, with $H_0: \theta = 0$.
- Every $P = P_{obs} \rightarrow x = x_{obs}$.
- Setting $H_1: \theta = x_{obs}$ gives an upper bound for BF. (Edwards, Lindman, and Savage, 1963)



Source: Benjamin et al. (2017).

Some Implications

1. Calculations illustrate the fact that knowing that $P=0.05$ is *much* weaker evidence than knowing that $P<0.05$.
   - In general, Bayes factor for $P=\alpha$ is smaller than rejection ratio for $P<\alpha$ (for any level of power). (Proved in Bayarri et al., 2016)
   - Intuitively, $P<0.05$ includes many (much more convincing!) *P*-values smaller than 0.05.
   - Report $P=P_{obs}$, not $P<\alpha$ and definitely not $P<P_{obs}+\varepsilon$.

2. $P=0.05$ is actually pretty weak evidence: roughly 3:1 odds of H_1 versus H_0.

Suggestions For Reproducible Research

- Pre-experimental design:
  - Consider whether prior odds warrant lower significance threshold.
  - Under realistic anticipated effect size, calculate power (really!) and report it.
- Post-experimental evaluation of evidence:
  - Using (ex ante) anticipated effect size for H_1, calculate Bayes factor.
  - If can't, then calculate Bayes factor implied by the evidence under a range of assumptions about H_1.
  - Evaluate H_1 in light of Bayes factor and plausible prior odds.
- Pre-register prior odds, significance threshold, anticipated effect size, and power calculations.