EXPERIMENTATION

Bayesian vs. Frequentist Inference in Conversion Testing

A practitioner's guide to Bayesian and frequentist A/B testing: what p-values and posteriors really mean, the role of priors, and when each fits CRO.

Bhaskar Roy Sarkar, cazyweb · June 28, 2026 ·

Bayesian · frequentist · A/B testing · priors · conversion rate optimization

Download PDF

ABSTRACT

Frequentist and Bayesian inference offer distinct but complementary answers to the question 'what should this experiment tell me?' Frequentist methods control long-run error rates, producing p-values and confidence intervals whose guarantees are statements about procedure rather than about any single result. Bayesian methods start from a prior belief about effect sizes, update it with data, and produce posterior distributions that directly quantify probability over parameter values - enabling intuitive decision rules such as 'the probability that B beats A.' Neither framework is universally superior: frequentist guarantees are indispensable when regulator-level Type I error control is required, while Bayesian posteriors fit naturally with business decision theory and early-stopping. This paper examines both inferential philosophies in depth, relates them to sequential testing, and offers concrete guidance on which fits a given CRO program.

1. Two Questions, Two Philosophies

Every A/B test ends with a decision under uncertainty. Should we ship variant B, or hold? Two statistical traditions answer this question differently, and confusing them is one of the most common sources of misinterpretation in conversion optimisation practice.

The **frequentist** tradition - formalised by Neyman and Pearson in 1933 [1] and distinct from Fisher's original significance-testing conception [2] - treats probability as a long-run frequency. A procedure's quality is measured by what would happen if you repeated it across many experiments: what fraction of truly null effects would be falsely flagged, what fraction of true effects would be detected. Specific test outcomes are evaluated against this procedural benchmark.

The **Bayesian** tradition treats probability as a degree of belief over propositions [3]. Given data, you update a prior distribution over the unknown quantity (here, the difference in conversion rates) to a posterior distribution. The posterior is a complete probabilistic description of what you know after the experiment. Inference is a matter of reading off credible intervals, posterior probabilities, or expected losses from that distribution.

Both traditions are internally coherent. Understanding them precisely - rather than accepting caricatures - is what allows a CRO practitioner to choose the right tool.

---

2. What a Frequentist p-Value Actually Means

The p-value is the probability, computed under the null hypothesis (no difference between A and B), of observing a test statistic at least as extreme as the one you obtained. Formally: p = P(T >= t_obs | H_0). This is not the probability that the null hypothesis is true. It is not the probability that you made an error. It is a statement about data, conditional on the null being true, not a statement about the null conditional on the data.

The American Statistical Association, in a formal 2016 position statement co-authored by Ronald Wasserstein and Nicole Lazar, enumerated six principles clarifying this [4]. Among them: 'P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.' The statement was motivated by pervasive misreading of p-values across scientific fields - the same misreadings that plague CRO dashboards.

The complementary concept is the **confidence interval (CI)**. A 95% frequentist CI is constructed by a procedure that, if repeated across many independent experiments, would contain the true parameter value 95% of the time. This is a statement about the long-run coverage of the procedure, not about any particular interval. Once you have computed a specific interval - say, [+0.3%, +2.1%] - it is not correct to say there is a 95% probability the true lift lies in that range. Either it does or it does not; the probability is not defined in the frequentist sense post-experiment.

This is not a gotcha against frequentism - it is just what the guarantee actually is. When the guarantee is understood correctly, it is useful: if you always stop tests at the pre-planned sample size and act on p < 0.05, you will falsely declare a winner at most 5% of the time across your programme, under the assumptions of the model.

---

3. What a Bayesian Posterior and Credible Interval Mean

In the Bayesian framework, you begin with a **prior distribution** p(theta) over the parameter of interest - here, the difference in conversion rates delta = p_B - p_A. You observe data D and apply Bayes' theorem: p(theta | D) proportional to p(D | theta) * p(theta). The resulting **posterior distribution** p(theta | D) is the object of inference [3].

A **credible interval** is simply the central or highest-density region of this posterior. A 95% credible interval contains 95% of the posterior probability

mass. Unlike the frequentist CI, this is a direct probability statement: given the prior and the data, there is a 95% posterior probability that the true parameter falls in this interval. This is the interpretation practitioners usually want, which is why Bayesian credible intervals are often easier to communicate to non-statisticians.

The posterior also yields the metric most directly useful to business decisions: the **posterior probability that B beats A**, i.e., P(delta > 0 | D). This can be read straight from the posterior. No p-value inversion or effect-direction inference is needed.

3.1 The Role of Priors

Priors are the most contested part of Bayesian practice. The classical objection is that they are subjective: two analysts with different priors will reach different posteriors from the same data. This objection is real but often overstated in practice.

First, **weakly informative priors** - priors that encode only broad constraints, such as 'conversion rate lifts are unlikely to exceed 50%' - have negligible influence on posteriors once sample sizes reach the hundreds or thousands typical in CRO [3]. For large experiments, the data dominate.

Second, **empirical Bayes** or **hierarchical** priors can be calibrated to the organisation's historical test results, making the prior data-driven rather than arbitrary. A prior built from 200 past tests is not a guess; it is regularised information.

Third, and most importantly, priors make assumptions explicit. Frequentist analyses also embed assumptions - about the null effect size, the sampling distribution, the stopping rule - but these are often invisible. Bayesian priors surface the analyst's beliefs and make them auditable [2].

Kruschke's 2013 paper on Bayesian estimation as a replacement for the t-test [5] demonstrates that specifying a weakly informative prior on means and

standard deviations yields complete posterior distributions that answer the practical question ('how large is the effect, and how sure are we?') more directly than a dichotomous significance test.



Bayesian updating: a prior belief, combined with the data (the likelihood), yields a posterior - a full distribution over the effect, not a single number. More data makes the posterior narrower and more confident.

4. Bayesian A/B Testing in Practice

4.1 The Posterior Probability and Expected Loss

In a standard two-variant Bernoulli conversion test, the conjugate prior for a conversion rate p is the Beta distribution, Beta(alpha, beta). After observing k conversions in n visits, the posterior is Beta(alpha + k, beta + n - k). The posterior probability that B beats A is then the integral P(p_B > p_A | D), which is computed analytically or by Monte Carlo simulation [6].

VWO's SmartStats system, described in a technical whitepaper by Chris Stucchio, operationalises this as the core decision metric [6]. The whitepaper, which accompanies the public launch of VWO's Bayesian engine, explains both the Beta-Bernoulli model and the role of the loss function in stopping decisions.

The complementary decision metric is **expected loss**: if you stop the test now and declare B the winner, what is the expected conversion you give up if A is actually better? Formally, expected_loss(B) = E[max(delta, 0) | D, delta < 0], integrated over the posterior. You ship B when its expected loss falls below a business-defined threshold - for example, 0.1% of conversion rate [6]. Expected loss encodes both the probability of being wrong and the magnitude of the error if you are, which makes it a principled business decision rule. Its use in CRO was popularised by Stucchio's work at VWO and has since spread across experimentation platforms.

A rigorous academic treatment of Bayesian inference for the A/B test, including Bayes factor formulations and informative priors, is provided by Gronau, Raj K. N., and Wagenmakers (2021) [7].

4.2 When Can You Stop?

A key practical advantage of the Bayesian approach is that the posterior is always a valid summary of current knowledge. You can compute P(B beats A) at any time, and the number is what it is - your current state of belief given the data seen. There is no multiple-testing penalty for checking the posterior repeatedly in the way that there is for frequentist p-values (see Section 5).

However, this does not mean early stopping is free of consequences. If you use a threshold rule ('stop when P(B beats A) > 0.95') and apply it repeatedly, the decision rule has implicit frequentist properties - specifically, a Type I error rate that depends on the stopping threshold and the prior [8]. Bayesian inference tells you what you believe now; it does not automatically guarantee a particular long-run false-positive rate for a stopping procedure. Practitioners who want both - interpretable posteriors and bounded false-positive rates - can use sequential Bayesian methods or hybrid approaches.

Thompson Sampling, a Bayesian adaptive method in which traffic is dynamically reallocated toward the arm with the higher posterior probability of being best, extends this logic to active traffic optimisation [9]. It is more applicable to bandit problems (exploit now while learning) than to inference

problems (reach a decision about a fixed set of variants), and the distinction matters: Thompson Sampling minimises regret during the test but may produce biased effect estimates at the end.

---

5. Frequentist Guarantees: Type I Error and the Peeking Problem

The frequentist framework's central strength is the **Type I error rate guarantee**. If you set alpha = 0.05 and run a fixed-sample test to its pre-planned size, you will falsely reject a true null at most 5% of the time across your programme, regardless of the actual distribution of true effects. This guarantee requires no assumption about the prior distribution of effects - it holds for any mixture of null and non-null experiments.

The Neyman-Pearson framework [1] establishes the trade-off between Type I error (false positive, alpha) and Type II error (false negative, 1 - power). Power analysis before the test determines the sample size required to detect a minimum detectable effect at the desired power (typically 80%) given alpha.

The guarantee breaks when experimenters **peek** - that is, check results before the pre-planned sample size is reached and stop early if $p < 0.05$. Armitage, McPherson, and Rowe (1969) [10] showed formally that repeated significance testing on accumulating data inflates the actual Type I error rate far above the nominal level. With unrestricted continuous monitoring at alpha = 0.05, the true false-positive rate can exceed 30%. In practice, teams who watch their dashboards daily and stop tests the moment they cross significance are not running 5%-alpha experiments; they may be running 20-30%-alpha experiments without knowing it.

Simmons, Nelson, and Simonsohn (2011) [11] documented how this and related 'researcher degrees of freedom' - stopping when significant, adding covariates, excluding outliers after looking at results - compound into false-positive rates that make many published findings unreliable. Their analysis, conducted in the

context of academic psychology, is directly applicable to CRO: a team that peeks, extends or stops tests based on visual inspection of dashboards, and selectively reports variants is performing exactly this category of practices.

---

6. Sequential Testing: Bridging the Gap

Sequential testing is the frequentist answer to the peeking problem: it provides valid inference at any point during data collection by using adjusted thresholds that account for multiple looks. The sequential probability ratio test (SPRT) is the classical approach, but it requires specifying the alternative effect size in advance.

Johari, Pekelis, and Walsh - along with Pete Koomen in the journal version - developed a variant called 'always valid inference,' which produces p-values and confidence intervals that remain valid under continuous monitoring without specifying a fixed alternative [8]. This work forms the statistical backbone of Optimizely's Stats Engine, which implements these methods to allow experimenters to check results at any time while maintaining their nominal Type I error rates [8].

This is important context for the Bayesian vs. frequentist debate: sequential frequentist methods largely close the 'you cannot peek' objection against frequentism. A frequentist experimenter using always-valid inference can stop at any time, read valid p-values, and maintain error-rate control. The remaining differences between sequential frequentist and Bayesian approaches are philosophical (what probability means) and practical (how prior information is incorporated and how decision rules are framed).

The Microsoft CUPED method (Controlled-experiment Using Pre-Experiment Data), introduced by Deng, Xu, Kohavi, and Walker (2013) [12], addresses a related but different problem: variance reduction. By using pre-experiment covariates to reduce noise in the outcome metric, CUPED increases the sensitivity of both

frequentist and Bayesian tests. It is orthogonal to the stopping-rule debate but highly practical for CRO teams with access to historical user data.

---

7. When Each Framework Fits a CRO Programme

Neither framework is correct in the abstract; the question is which fits the practical constraints and objectives of a given programme.

Frequentist testing with fixed-horizon design fits when: (a) regulatory, legal, or audit requirements demand Type I error guarantees that are framework-agnostic and easy to communicate; (b) the team cannot resist peeking and cannot implement sequential correction; (c) effect sizes are similar across tests and power analysis is straightforward; (d) the organisation needs a bright-line decision rule that is hard to game.

Sequential frequentist testing (always-valid inference, group sequential methods) fits when: (a) tests need to stop early to reduce exposure to losing variants; (b) teams want to monitor experiments continuously; (c) Type I error guarantees are still required. This is the best of both worlds for many commercial teams, and platforms like Optimizely already implement it.

Bayesian testing fits when: (a) the team has meaningful prior information from past tests (same product, similar audiences) that should inform the decision; (b) the decision is better framed as expected-loss minimisation than hypothesis rejection; (c) stakeholders want probability statements ('there is a 92% chance B is better') rather than reject/fail-to-reject outcomes; (d) the programme runs many small tests where accumulating priors across tests is valuable.

A practical note: Berger (2003) [2] argues that Fisher, Jeffreys, and Neyman might have agreed more than their polemics suggested, precisely because conditional frequentist testing and Bayesian testing can produce similar numerical answers under well-calibrated priors. Berger and Sellke (1987) [13] showed the opposite in a provocative sense: a p-value of 0.05 corresponds to a Bayesian posterior probability of the null as high as 0.30 under reasonable

prior assumptions, suggesting that standard frequentist thresholds are more lenient than commonly believed.

For most mid-size ecommerce teams (hundreds of tests per year, limited statistical staffing), the practical recommendation is: use a sequential frequentist engine to eliminate the peeking problem, adopt minimum detectable effect calculations before launching tests, and reserve Bayesian decision rules for situations where expected-loss framing is genuinely superior to binary decisions - such as multi-armed bandit traffic allocation or portfolio-level experiment prioritisation.

---

8. Conclusion

The frequentist/Bayesian debate is sometimes presented as a contest to be settled. It is better understood as two solutions to different formulations of the inference problem. Frequentism offers guaranteed long-run error control without requiring prior beliefs; Bayesianism offers direct probability statements about hypotheses but requires specifying (and defending) a prior. Sequential methods, developed partly in response to real-world CRO and tech-industry needs, extend frequentist validity to continuous monitoring, reducing the practical gap.

For CRO practitioners, the takeaways are concrete: understand what a p-value actually guarantees (and does not), implement sequential testing or strict fixed-horizon discipline to prevent inflated false-positive rates, consider Bayesian framing when expected-loss decision rules fit the business problem, and never treat statistical significance as the sole arbiter of shipping decisions. Effect size, business impact, and implementation cost all belong in the decision - statistics provides the evidentiary input, not the full answer.

References

1. Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*. https://doi.org/10.1098/rsta.1933.0009

2. Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing?. *Statistical Science*. https://doi.org/10.1214/ss/1056397485

3. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd). *Chapman & Hall/CRC*. https://sites.stat.columbia.edu/gelman/book/

4. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*. https://doi.org/10.1080/00031305.2016.1154108

5. Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/a0029146

6. Stucchio, C. (2015). Bayesian A/B testing at VWO. *Visual Website Optimizer (technical whitepaper)*. https://vwo.com/downloads/VWO_SmartStats_technical_whitepaper.pdf

7. Gronau, Q. F., Raj K. N., A., & Wagenmakers, E.-J. (2021). Informed Bayesian inference for the A/B test. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v100.i17

8. Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*. https://doi.org/10.1287/opre.2021.2135

9. Russo, D., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*. https://doi.org/10.1561/2200000070

10. Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*. https://doi.org/10.2307/2343787

11. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*. https://doi.org/10.1177/0956797611417632

12. Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM 2013)*. https://doi.org/10.1145/2433396.2433413

13. Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1987.10478397

Cite this

APA Copy

```
Bhaskar Roy Sarkar (2026). Bayesian vs. Frequentist Inference in
Conversion Testing. cazyweb Research.
https://cazyweb.com/research/bayesian-vs-frequentist-conversion-
testing
```

BIBTEX Copy

```
@techreport{bayesian-vs-frequentist-conversion-testing,
  author = {Bhaskar Roy Sarkar},
  title = {Bayesian vs. Frequentist Inference in Conversion Testing},
  institution = {cazyweb Research},
  year = {2026},
  url = {https://cazyweb.com/research/bayesian-vs-frequentist-
conversion-testing}
}
```

RELATED READING

Statistical significance in A/B testing, explained

Always-valid inference in conversion testing

Want a team to run experiments like this for you? See how we help

Let's build something that converts.

Start a project with us, or join the bench of specialists and agencies we deliver with.

Start a project ↗

cazyweb

The engineering and growth team behind leading brands and companies across the US and Europe.

CRO and ecommerce insights

Occasional, genuinely useful notes. No fluff.

you@company.com

Subscribe →

Work email only. No spam. Unsubscribe any time. Privacy.

Solutions

- Product Optimization
- Experimentation Program
- cazyweb Platform

For Ecommerce

- CRO Program
- Email & Retention
- Paid Ads
- Landing Pages
- Revenue Calculator FREE
- A/B Test Calculator FREE
- CRO Agency
- Shopify CRO
- Ecommerce Growth
- Website Performance

Company

Explore

CazyWeb IT Solutions
GSTIN: 19JTLPS8476P2ZZ
hello@cazyweb.com

© 2026 cazyweb. All rights reserved.