Review

Statistical Significance Reconsidered: The Role of Bayesian Methods in the Life Sciences

Reconsidération de la signification statistique : le rôle des méthodes bayésiennes dans les sciences de la vie

Ishaan S. Goswami[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: igosw085@uottawa.ca

Abstract | Résumé

Statistical reasoning underpins the interpretation of experimental results in the life sciences. For decades, frequentist hypothesis testing has dominated research practice, with the p-value serving as a benchmark of significance. Yet, p-values are often misunderstood and limited in what they reveal, and issues of statistical power further complicate inference. Bayesian methods offer an alternative framework that addresses several of these limitations. By incorporating prior knowledge and comparing the probabilities of competing hypotheses, Bayesian inference yields richer measures such as Bayes Factors, posterior probabilities, and credible intervals. However, Bayesian approaches also present challenges, including the subjectivity of priors and the computational intensity of complex models. This review synthesizes the conceptual and practical differences between frequentist and Bayesian approaches, with particular attention to their implications for experimental design, ethical considerations, and interpretation in modern biological research. It further examines how disciplinary norms, data complexity, and regulatory constraints shape methodological preferences across fields. Rather than positioning the two paradigms as competing frameworks, this work argues for statistical bilingualism as a necessary foundation for transparent, context-aware, and reproducible scientific inference.

———

Le raisonnement statistique sous-tend l'interprétation des résultats expérimentaux dans les sciences de la vie. Pendant des décennies, le test d'hypothèses fréquentistes a dominé la pratique de la recherche, la valeur p servant de référence significative. Pourtant, les valeurs p sont souvent mal comprises et limitées dans ce qu'elles révèlent, et les questions de puissance statistique compliquent encore davantage l'inférence. Les méthodes bayésiennes offrent un cadre alternatif qui répond à plusieurs de ces limitations. En incorporant des connaissances préalables et en comparant les probabilités d'hypothèses concurrentes, l'inférence bayésienne permet de produire des mesures plus riches telles que les facteurs bayésiens, les probabilités postérieures et les intervalles crédibles. Cependant, les approches bayésiennes présentent également des défis, notamment la subjectivité des a priori et l'intensité computationnelle des modèles complexes. Cette revue synthétise les différences conceptuelles et pratiques entre les approches fréquentistes et bayésiennes, en portant une attention particulière à leurs implications pour la conception expérimentale, les considérations éthiques et l'interprétation dans la recherche biologique moderne. Il examine également comment les normes disciplinaires, la complexité des données et les contraintes normatives influencent les préférences méthodologiques selon les domaines. Plutôt que de présenter les deux paradigmes comme des cadres concurrents, ce travail plaide en faveur du bilinguisme statistique comme fondation nécessaire pour une inférence scientifique transparente, consciente du contexte et reproductible.

Keywords: Bayesian inference; frequentist statistics; p-values; Bayes Factors; statistical power; experimental design; reproducibility; model comparison; hypothesis testing

Introduction

The importance of statistical significance in the life sciences

Interpreting whether experimental results are meaningful, rather than merely interesting, is a question of statistics. R.A. Fisher demonstrated the vitality of statistics through his "lady tasting tea" experiment. He outlined that "[e]very experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (1). The significance of data can be assessed through statistical tests, treating experimental outcomes as classical statistical events.

Frequentist Hypothesis Testing

In frequentist statistics, the p-value quantifies how compatible the observed data are with the null hypothesis, which is the assumption that there is no true effect. Specifically, it represents the probability of obtaining results as extreme as those observed, if the null hypothesis were true. A small p-value indicates that the observed results would be unlikely under the null. Conventionally, $p < 0.05$ is taken to indicate statistical significance, meaning that assuming the null hypothesis were true, results at least as extreme

as those observed would occur with probability less than 5%.

If one were to simulate experiments repeatedly under the null hypothesis, p-values would follow a uniform distribution between 0 and 1. In practice, however, this uniformity depends on correct model specification and data independence. When the null hypothesis is false, p-values cluster toward smaller values, and their distribution shape depends on factors such as effect size, sample size, and statistical power.

Power and sample size

In classical statistics, there are two types of errors: Type I (α) and Type II (β) errors. Type I error, or the false positive, occurs when the null hypothesis is rejected, even though it is actually true. It is typically set at 0.05, which means that if the null is true, it will be wrongly rejected 5% of the time. Type II error, or the false negative, occurs when one fails to reject the null hypothesis, even though the alternative is true. Type 2 error relies on the sample size, effect size, and variability of the data. A smaller sample size inherently increases the standard error, which makes the test less sensitive. Statistical power (1- β) is the probability that a test correctly detects an effect when one exists. It is often set at 80% (2). This means that high power means there is a lower chance of a Type 2 error.

Power analysis helps determine the sample size required (n) to detect a specified effect size with a desired significance level (α) and power (1-β). Power analysis should be conducted a priori to determine an adequate sample size, rather than after the fact. This ensures the study is designed to detect the effect of interest with sufficient sensitivity. An underpowered study is ethically questionable, as it may entail wasted reagents, time, and potentially unnecessary harm to animal and human participants without yielding interpretable results. If an underpowered study is not presented transparently, it could drive further research and policy towards a direction that is fundamentally inconclusive or misleading. As Gaskill and Garner (3) emphasize, underpowered studies not only waste resources but risk ethical violations in animal research. However, this does not mean that an excessively large sample should be collected. This can lead to the detection of statistically significant effects which may have no real biological or clinical significance. Both extremes undermine the ethics and efficiency of scientific and medical research.

Limitations of frequentist hypothesis testing

A small p-value doesn't prove that the null hypothesis is incorrect, and a large p-value does not prove the null hypothesis is correct. Rather, they indicate how comparable the observed data are with the assumption that there is no true effect between the variables. Gigerenzer (4) argues that statistical significance often substitutes for substantive reasoning. A large p-value may simply reflect insufficient data, insufficient statistical power, a poorly specified model, such as variance misspecification, or high variability in the data.

It is important to note that one can never accept the null hypothesis. Rather, you can only fail to reject it. As summarized by the American Statistical Association's official statement, a p-value near 0.05 alone provides only weak evidence against the null hypothesis and should not be treated as a bright-line criterion for discovery (5). That may not seem like a major semantic difference, but it has massive implications in how one can analyze data. The absence of evidence for the null hypothesis at high p-values is not evidence of absence. Thus, while p-values can hint at inconsistencies between the data and the null, it cannot confirm a competing hypothesis. This is a limitation that Bayesian methods are designed to address.

Bayesian Hypothesis Testing

In Bayesian statistics, the likelihood of data being obtained under both the null and alternative can be compared. Bayesian methods allow for the quantification of the degree of support for one hypothesis over another, rather than rejecting based on thresholds. This requires an assignment of prior probabilities and likelihoods under both hypotheses. These can be informed by how a scientist thinks the data would work or can be uniform priors if there are no strong assumptions. Probability is interpreted as a degree of belief, not just frequency. This allows one to update their beliefs about a hypothesis using the observed data. This is governed by Bayes' theorem:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \quad \text{(Eq. 1)}$$

$P(H|D) = Posterior: probability\ of\ hypothesis\ H\ given\ data\ D$

$P(D|H) = Likelihood: probability\ of\ data\ given\ H$

$P(H) = Prior: belief\ about\ H\ prior\ to\ data$

$P(D) = Marginal\ Likelihood: hard\ to\ compute, normalizing\ factor$

After defining two competing hypotheses, probability models must also be defined under both hypotheses. For example, under the null hypothesis, (H_0), it is assumed there is no true effect. In this case, the observed data X are modeled as arising from a normal distribution centered at zero:

$$X \sim \mathcal{N}(0, \sigma^2) \quad \text{(Eq. 2)}$$

Where σ^2 represents the variance of the observation. This formulation reflects the assumption that any observed deviation from zero is due solely to random noise, and that there is no mean difference.

The alternative hypothesis (H_1), on the contrary, allows for a non-zero effect size wherein the data are modeled as:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{(Eq. 3)}$$

Where μ represents the true, unknown, effect. In Bayesian statistics, μ is treated as a random variable and assigned a prior distribution:

$$\mu \sim \mathcal{N}(0, \tau^2) \quad \text{(Eq. 4)}$$

This prior reflects the uncertainty about the effect size before observing the data, with τ^2 controlling the expected magnitude of plausible effects. The choice of prior thus influences the resulting inference, particularly in cases with limited data.

Now that the hypotheses' probability models have been defined, likelihoods for the data appearing under both hypotheses can be generated. P(Data | Null) finds the likelihood of the data under the null. This is a simple density evaluation. For P(Data | Alternative), one calculates the marginal likelihood, which is the evidence for the alternative. This is defined as:

$$P(D|H_1) = \int P(D|\theta) \cdot P(\theta|H_1) d\theta \quad \text{(Eq. 5)}$$

Under the alternative hypothesis, the marginal likelihood of the observed data is given by Equation (5). This quantity represents the probability of the data after integrating over all possible parameter values. In this expression, denotes the marginal likelihood of the data under the alternative hypothesis. The term is the likelihood, representing the probability of observing the data given a specific parameter value . The term is the prior distribution over the parameter , reflecting our uncertainty about its value under . The integral sums (or averages) the likelihood over all possible values of , weighted by the prior distribution.

This integrates the likelihood over the prior distribution of parameters under the alternative. The more parameters or flexibility a model has, the more "room" it has to explain data, even when the effect is not real. Hence, this integration favours simpler models such as those with fewer parameters or more constrained prior structures, in line with Ockham's Razor, which advises against unnecessary complexity unless the data demands it. However, the life sciences have ample systems that are complex, necessitating balance, not oversimplification.

The Bayes Factor (BF_{10}) compares how well two competing hypotheses predict the observed data. It is defined as the likelihood of the data under the alternative hypothesis to the data under the null. The Bayes Factor is not the posterior probability of a hypothesis; rather, it is a relative measure of the strength of the evidence for one hypothesis over the other. The Bayes Factor is defined as:

$$BF_{10} = \frac{P(Data|Alternative)}{P(Data|Null)} \quad \text{(Eq. 6)}$$

The Bayes Factor can be combined with prior odds to yield the posterior odds, which quantifies how much more likely one hypothesis is than another after also accounting the prior beliefs regarding data distribution.

There are certain heuristic guidelines to interpret the Bayes Factor. These are not hard statistical laws and are meant to act as tools to facilitate analysis. Jeffreys (6) introduced his scale to facilitate comparative analyses. Kass and Raftery (7) introduced stricter cutoffs than Jeffreys and their scale is widely cited in applied works. Lee and Wagenmakers (8) introduced their scale for applications in psychology. These are summarized by Taboga (9). According to these conventions, a Bayes Factor between 1 and 3 is considered anecdotal or barely worth mentioning, 3 to 10 indicates moderate or substantial evidence, 10 to 30 reflects strong evidence, 30 to 100 suggests very strong evidence, and values above 100 represent extreme or decisive support for the alternative hypothesis. It is necessary to emphasize that conclusions should always integrate experimental design, data quality, and prior plausibility

Once the Bayes Factor has been computed, it can be combined with prior odds to yield posterior probabilities, which quantifies the updated degree of belief in each hypothesis given the data. This requires specification of prior probabilities for both the null and alternative hypotheses. When interpreting results, thus, the Bayes Factor and the assumptions encoded in the prior must be explicitly stated. Unlike p-values, which measure compatibility with the null hypothesis, Bayes factors directly quantify the relative evidence that the data provides for one hypothesis compared with another.

For simple models that have known distributions and linear relationships, this process can be done in an analytical manner using closed-form solutions. This, however, is only the case when the prior and the likelihood distributions are conjugate or belong to the same distributional family as each other. One classic conjugate prior-likelihood pair is the normal likelihood with a normal prior which results in a normal posterior. This is used to model continuous variables with unknown mean but known variance. Another commonly used pair is the binomial likelihood with a beta prior which yields a beta posterior. However, the life sciences rarely follow these straightforward models. Oftentimes, Bayesian inference must rely on numerical approximation methods. The most widely used method is the Markov Chain Monte Carlo algorithm, which allows for sampling from the posterior distribution, even if it is computationally difficult to compute exactly. Markov Chain Monte Carlo algorithms, such as Metropolis-Hastings and Hamiltonian, generate a sequence of samples that approximate the posterior distribution via different means. By generating enough samples, it will converge to the true posterior.

In Bayesian statistics, the corollary to confidence intervals is the credible interval. A 95% credible interval means that given the observed data and the specified prior, there is a 95% probability that the parameter lies within that interval.

Posterior Predictive Distributions are a standard tool in the modern Bayesian workflow used to estimate the probability distribution of future or replicated data conditional on what has already been observed, as outlined by Gelman et al. (10). They are obtained by averaging predictions across all possible parameter

values, weighted by their posterior probability. Posterior predictive checks allow scientists not just to test a hypothesis, but to simulate new experiments. This is valuable in the life sciences, where replication is costly. This is also a part of the model criticism workflow, where experiments can be simulated a priori to evaluate expected performance.

The pitfalls of Bayesian priors

Bayesian inference is powerful as it can quantify updated belief in a hypothesis given the model and prior assumptions. The double-edged sword of Bayesian statistics is the prior model. The Bayesian inference about the alternative hypothesis can be biased and misleading based on the selection of the prior. This is one of the central critiques in Bayesian statistics. The Bayesian updating equation, which explains the reliance of the posterior upon the prior is defined as:

$$Posterior \propto Likelihood \times Prior \quad \text{(Eq. 7)}$$

The prior encodes what one already believes about a parameter or a hypothesis before seeing the data. There are three major ways a prior can be biased: I) it is too narrow, II) it is too wide, and III) it is too biased. If a prior is too narrow, the probability distribution is tightly centred around 0, which may fail to detect real effects that occur farther away in the tails. The data must show very strong deviation from zero to show that the alternative is more probable than the null. If the prior is too wide, like a uniform distribution, evidence for an effect may be inflated, even if the data generated is weak. This dilutes any evidence, as probability mass is spread across many implausible values, contrary to Ockham's Razor. If a prior is too biased, the posterior can be biased towards an expected effect due to the proportional relation between them. This may make weak data appear more supportive of a hypothesis than it really is. These three scenarios pose an important ethical dilemma to Bayesian statisticians. A model may be too conservative to accurately model complicated data, or too general to ascertain significant results. If one thinks the data will fit a certain distribution, they could fit the prior to make that hypothesis valid.

This can be overcome, though. A sensitivity analysis can be run, where the Bayesian model is run under multiple different priors. If the conclusion changes dramatically, then it shows the result is fragile. But it is important that these are run at the same time to avoid selective reporting of results that depend on a particular prior specification. Priors should exclude implausible values while still remaining sufficiently flexible to allow the data to inform the posterior.

Comparing the Two Schools & Conclusions

Having reviewed both the frequentist and Bayesian frameworks, it is valuable to directly compare the two. While both approaches rely on probability theory, they interpret probability very differently. This leads to different manners of designing experiments, analyzing the data, and drawing conclusions about the underlying biological systems.

Fundamentally, the two schools diverge on what probability means. Frequentists view probability as the long-run frequency of events. Here, randomness lies only in the data that is observed. Inference focuses on error rates across hypothetical repeated experiments. Bayesians, on the other hand, view probability as a degree of belief about uncertainty. Parameters are treated as random variables with their own probability distributions, and inference comes from updating prior beliefs about the distribution of results with new data to obtain a posterior distribution. With respect to reporting information, the frequentist would say with their confidence intervals, that if an experiment was conducted 1000 times, 95% of confidence intervals would contain the true success rate. A Bayesian would instead say that given their selected prior and the observed data, there is a 95% probability that the true success rate lies within a certain range.

The differences between the two schools are more evident in hypothesis testing. Classical frequentist statistics focuses on null hypothesis significance testing. As outlined above, this framework cannot directly provide information on the validity of the alternative hypothesis. The Bayesian school instead uses the Bayes Factor and posterior odds to directly compare two hypotheses given collected data with the incorporation of prior biological knowledge. This makes Bayesian statistics more robust, but also more open to being abused.

Uncertainty is treated in different ways in the two schools. Frequentists use the theoretical confidence interval, wherein if the experiment was repeated an infinite number of times, 95% of the intervals would contain the true parameter. This is often misunderstood to be a probability statement about the true value of the parameter itself. Bayesians instead use the credible interval, which directly represents the probability of the parameter; given the observed data and the prior, there is an X% chance that the parameter lies within the interval.

Experimental design also looks quite different under the two schools. In the frequentist tradition, researchers rely on a power analysis to ensure that an experiment is sensitive enough to detect a true effect. The emphasis is on controlling error probabilities over the long run, so that conclusions drawn across many repetitions of the same study are statistically reliable.

The Bayesian perspective allows experimental design to incorporate prior distributions and simulate posterior predictive outcomes before data collection begins. However, simulation-based planning can also be conducted within frequentist frameworks. This approach helps researchers anticipate how informative a study will likely be, given both prior knowledge and the planned sampling. Bayesian adaptive designs thus have the potential to stop trials early when strong evidence emerges, saving resources and reducing unnecessary exposure of animals or patients to ineffective treatments. However, regulatory adoption is

still ongoing. Bayesian adaptive trials in clinical research exemplify how probabilistic updating can improve both ethical and statistical efficiency (11).

Frequentist statistics is often praised for its simplicity and objectivity, since it relies on minimal prior assumptions. The use of standardized thresholds, such as the familiar $p < 0.05$, makes results easy to compare across studies and facilitates regulatory acceptance. In contrast, the strength of Bayesian methods lies in their flexibility, as they allow researchers to incorporate prior biological knowledge, update inferences as new data arrives, and generate richer outputs such as probabilities of hypotheses, model comparisons, and predictive distributions.

Yet these strengths come with trade-offs. Frequentist methods, while simple, cannot assign probabilities directly to hypotheses and can foster rigid reliance on significance thresholds, hence obscuring biological nuance. Bayesian approaches, on the other hand, are criticized for being sensitive to the choice of prior, which can introduce bias if not carefully justified, and for the computational intensity of techniques like Markov Chain Monte Carlo when applied to large or high-dimensional datasets. In practice, the perceived advantages and drawbacks of each framework often depend on the goals of the study, the complexity of the data, and the broader scientific or regulatory context.

Neither framework is universally superior. Frequentist methods remain dominant in the life sciences for historical, institutional, and regulatory reasons. At the same time, Bayesian methods are steadily gaining traction in areas such as genomics, neuroscience, epidemiology, and adaptive clinical trials, where richer probabilistic statements and the incorporation of prior knowledge offer clear advantages.

A pragmatic stance is to view the two schools as complementary tools rather than competitors. Frequentist methods provide standardized inference and well-defined error control, which support comparability and regulatory acceptance. Bayesian methods, on the other hand, allow deeper integration of prior information, adaptive study designs, and more intuitive probability statements. For modern life scientists, fluency in both schools offers a richer and more flexible toolkit for interpreting complex, noisy, and high-dimensional data.

One consideration is that the preference for frequentist or Bayesian methods across scientific disciplines is not purely statistical, but reflects the structure of the questions being asked, as well as institutional and practical constraints. Frequentist approaches remain dominant in fields such as clinical research, where regulatory frameworks demand standardized thresholds, controlled error rates, and comparability across studies (12). In contrast, Bayesian methods have gained traction in domains such as genomics, neuroscience, and systems biology, where data are high-dimensional, prior knowledge is informative, and adaptive or iterative experimental designs are advantageous (13-15). These differences are further reinforced by disciplinary training, computational accessibility, and historical precedent. As a result, methodological preference often emerges not from theoretical superiority, but from alignment with the epistemological and logistical demands of a field.

Statistical reasoning sits at the heart of modern life sciences. Regardless of the question being investigated, researchers must parse out what the data shows and determine how confident they are in those conclusions. Frequentist and Bayesian statistics offer different answers and methodologies to these challenges. However, they both provide unique insights and pose limitations that must be acknowledged, not disregarded. For life scientists, the key lesson is that statistics is not just a set of mechanical tests but a framework for reasoning under uncertainty. As biological research increasingly grapples with big data, complex models, and ethical constraints on experimentation, fluency in both frequentist and Bayesian methods will be essential to navigate the increasing complexity, uncertainty, and scale of modern life sciences research.

Editorial Conflict of Interest Statement

Ishaan S. Goswami is Co-Editor-in-Chief of the *University of Ottawa Science Undergraduate Research Journal*. He was fully recused from all aspects of the editorial process for this manuscript, including reviewer selection, peer review, and final decision-making. The manuscript was handled independently by other members of the editorial board.

References

1. R. A. Fisher, The Design of Experiments (Oliver & Boyd, 1935).
2. C. C. Serdar, M. Cihan, D. Yücel, M. A. Serdar, Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia Medica 31, 010502 (2021).
3. B. N. Gaskill, J. P. Garner, Power to the people: Power, negative results, and sample size. J. Am. Assoc. Lab. Anim. Sci. 59, 9–16 (2020).
4. G. Gigerenzer, Mindless statistics. J. Socio-Econ. 33, 587–606 (2004).
5. R. L. Wasserstein, N. A. Lazar, The ASA's statement on p-values: Context, process, and purpose. Am. Stat. 70, 129–133 (2016).
6. H. Jeffreys, Theory of Probability (Oxford Univ. Press, ed. 3, 1939).
7. R. E. Kass, A. E. Raftery, Bayes factors. J. Am. Stat. Assoc. 90, 773–795 (1995).
8. M. D. Lee, E.-J. Wagenmakers, Bayesian Cognitive Modeling: A Practical Course (Cambridge Univ. Press, 2014).
9. M. Taboga, Lectures on Probability Theory and Mathematical Statistics (Kindle Direct Publishing, 2021).
10. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, Bayesian Data Analysis (CRC Press, ed. 3, 2013).
11. D. A. Berry, Bayesian clinical trials. Nat. Rev. Drug Discov. 5, 27–36 (2006).

12. Center for Drug Evaluation and Research, “Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry” (U.S. Food and Drug Administration, 2020); https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.
13. E. C. Goligher, A. Heath, M. O. Harhay, Bayesian statistics for clinical research. The Lancet 404, 1067–1076 (2024).
14. H. P. Kärkkäinen, M. J. Sillanpää, Back to basics for Bayesian model building in genomic selection. Genetics 191, 969–987 (2012).
15. K. P. Kording, Bayesian statistics: Relevant for the brain? Curr. Opin. Neurobiol. 25, 130–133 (2014).