STAT7630: Bayesian Statistics
Lecture Slides # 10

Checking Model Quality & Bayesian Hypothesis Testing
Chapter 8 Posterior Inference & Prediction

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: October, 2024

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

Assessing Bayesian Model Adequacy

- Key steps in evaluating the adequacy of a Bayesian model include:
    1. Assessing the sensitivity of the posterior distribution to the choice of prior and likelihood.
    2. Ensuring that the observed data aligns with predictions based on the posterior distribution.
    3. Evaluating the model's robustness to outliers and individual data points.

Sensitivity Analysis

- Regular sensitivity checks on the data model/likelihood are recommended but seldom performed.
- One approach is to evaluate how the posterior changes when selecting alternative models for the data (e.g., Poisson vs. negative binomial for count data).
- More frequently, we focus on assessing the sensitivity of the posterior to the prior specification.
- Key questions include:
  1. How does the posterior change when we modify the functional form of the prior?
  2. What is the impact when we retain the prior form but alter its parameters?
- If the posterior remains robust under these variations, we gain confidence in the reliability of our inferences.

Sensitivity Analysis: Example 1(a)

- Consider $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 known.
- Different prior choices for μ:
  - Conjugate prior: $\mu \sim N(\delta, \tau^2)$
  - Noninformative prior: $p(\mu) = 1$
  - Another prior: $\mu \sim t$-distribution centered at δ
- Key Question: How does the posterior change under these 3 priors?
- Methods for comparison:
  1. Plot the posterior distributions for each prior.
  2. Examine several posterior quantiles for each prior.

Local Sensitivity Analysis

- Evaluating a broad class of prior specifications can be challenging, particularly for multidimensional parameters θ.
- Local sensitivity analysis focuses on the effect of small changes in the hyperparameter values on the posterior distribution.
- Example 1(a): $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 known.
- Conjugate prior for μ: $\mu \sim N(\delta, \tau^2)$
- Compare the resulting posterior (using plots or quantiles) with these alternative priors:
    - $\mu \sim N(\delta - \tau, \tau^2)$
    - $\mu \sim N(\delta + \tau, \tau^2)$
    - $\mu \sim N(\delta, 0.5\tau^2)$
    - $\mu \sim N(\delta, 2\tau^2)$
- See `R` example for implementation.

Local Sensitivity Analysis: Example 1(b)

- Consider $Y_1, \ldots, Y_{200}$ as annual deaths from horse kicks for 10 Prussian cavalry corps over 20 years.
- Model: $Y_i \overset{iid}{\sim} \text{Poisson}(\lambda)$, with prior $\lambda \sim \text{Gamma}(\alpha, \beta)$.
- Compare posterior distributions for λ under the following priors:
    - $\lambda \sim \text{Gamma}(2, 4)$
    - $\lambda \sim \text{Gamma}(4, 8)$
    - $\lambda \sim \text{Gamma}(1, 2)$
    - $\lambda \sim \text{Gamma}(0.1 \times 2, \sqrt{0.1} \times 4)$
    - $\lambda \sim \text{Gamma}(3 \times 2, \sqrt{3} \times 4)$
- See the `R` example with Prussian horse kick data for detailed analysis.
- **Recommendation**: If the posterior is highly sensitive to the prior specification, consider using a more "objective" prior or be ready to justify your choice of prior.

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

Prior Predictive Distribution

- Recall that for a fixed value of θ, the data Y follow the distribution $p(Y|\theta)$.
- Since the true value of θ is uncertain, we should average over all possible values of θ to obtain a more accurate representation of the distribution of Y.
- Prior to observing data, the uncertainty in θ is captured by the prior distribution $p(\theta)$.
- For a new data point y_{new}, the **prior predictive distribution** is given by:

$$p(y_{\text{new}}) = \int_{\Theta} p(y_{\text{new}}, \theta) d\theta = \int_{\Theta} p(y_{\text{new}}|\theta) p(\theta) d\theta$$

Posterior Predictive Distribution (Post-Sample)

- After observing the data, the uncertainty in θ is updated using the posterior $p(\theta|\mathbf{y})$.
- The **posterior predictive distribution** for a new data point y_{new} is given by:

$$p(y_{\text{new}}|\mathbf{y}) = \int_{\Theta} p(y_{\text{new}}|\theta, \mathbf{y}) p(\theta|\mathbf{y}) d\theta = \int_{\Theta} p(y_{\text{new}}|\theta) p(\theta|\mathbf{y}) d\theta$$

  (since given θ, y_{new} is independent of the sample data $\mathbf{y}$)
- This distribution describes how we expect new data to behave.
- If the observed data align well with this pattern, it suggests that our model and prior are well-chosen.

Posterior Predictive Distribution: Example 2

- Recall the model: $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$, with $\lambda \sim \text{Gamma}(\alpha, \beta)$.
- The posterior distribution for λ given $\mathbf{y}$ is $\lambda|\mathbf{y} \sim \text{Gamma}\left(\sum y_i + \alpha, n + \beta\right)$.
- The posterior predictive distribution for a new data point y_{new} is:

$$p(y_{\text{new}}|\mathbf{y}) = \int_0^\infty p(y_{\text{new}}|\lambda)p(\lambda|\mathbf{y})d\lambda$$

- Substituting the Poisson pmf and Gamma posterior:

$$p(y_{\text{new}}|\mathbf{y}) = \int_0^\infty \frac{\lambda^{y_{\text{new}}} e^{-\lambda}}{y_{\text{new}}!} \times \frac{(n+\beta)^{\sum y_i + \alpha}}{\Gamma\left(\sum y_i + \alpha\right)} \lambda^{\sum y_i + \alpha - 1} e^{-(n+\beta)\lambda} d\lambda$$

Posterior Predictive Distribution: Example 2 (continued)

The posterior predictive distribution for y_{new} simplifies to:

$$\begin{aligned}
p(y_{\text{new}}|\mathbf{y}) &= \frac{(n+\beta)^{\sum y_i+\alpha}}{\Gamma\left(\sum y_i+\alpha\right)\Gamma(y_{\text{new}}+1)} \int_0^\infty \lambda^{y_{\text{new}}+\sum y_i+\alpha-1} e^{-(n+\beta+1)\lambda}\, d\lambda \\
&= \frac{(n+\beta)^{\sum y_i+\alpha}}{\Gamma\left(\sum y_i+\alpha\right)\Gamma(y_{\text{new}}+1)} \frac{\Gamma\left(y_{\text{new}}+\sum y_i+\alpha\right)}{(n+\beta+1)^{y_{\text{new}}+\sum y_i+\alpha}} \\
&= \frac{\Gamma\left(y_{\text{new}}+\sum y_i+\alpha\right)}{\Gamma\left(\sum y_i+\alpha\right)\Gamma(y_{\text{new}}+1)} \left(\frac{n+\beta}{n+\beta+1}\right)^{\sum y_i+\alpha} \left(\frac{1}{n+\beta+1}\right)^{y_{\text{new}}}
\end{aligned}$$

This is a (generalized) negative binomial distribution, NB(r, p), with $r = \sum y_i + \alpha$, $p = \frac{1}{n+\beta+1}$ and has mean and variance as

$$\text{Mean} = \frac{\sum y_i + \alpha}{n+\beta}$$

$$\text{Variance} = \frac{\sum y_i + \alpha}{(n+\beta)^2(n+\beta+1)}.$$

Posterior Predictive Distribution: Key Insight

- The posterior predictive distribution retains the same mean as the posterior distribution.
- However, the variance is greater due to additional sampling uncertainty when predicting a new data point.
- This reflects the variability introduced by drawing a new value, in addition to the uncertainty in the parameter λ.
- See the `R` example using the Prussian cavalry data for an illustration of this concept.

Posterior Predictive Distribution: Example 1(a)

Model Setup:

- Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. from $N(\mu, \sigma_0^2)$, where σ_0^2 is known, but μ is unknown.
- Place a normal prior on μ, i.e., $\mu \sim N(\delta, \tau^2)$.

$$p(\mu|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma_0^2}\right)\left(\mu - \left(\frac{\delta}{\tau^2}+\frac{n\bar{y}}{\sigma_0^2}\right)\left(\frac{1}{\tau^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)^2\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma_1^2}(\mu-\mu_1)^2\right)$$

where **posterior mean and variance** are

$$\mu_1 = \left(\frac{\delta}{\tau^2}+\frac{n\bar{y}}{\sigma_0^2}\right)\left(\frac{1}{\tau^2}+\frac{n}{\sigma_0^2}\right)^{-1} \quad \text{and} \quad \sigma_1^2 = \left(\frac{1}{\tau^2}+\frac{n}{\sigma_0^2}\right)^{-1}$$

Posterior Predictive Distribution: Example 1(a)

Predictive Distribution for New Data:

$$p(y_{\text{new}}|\mathbf{y}) \propto \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left[\frac{(y_{\text{new}}-\mu)^2}{\sigma_0^2}+\frac{(\mu-\mu_1)^2}{\sigma_1^2}\right]\right) d\mu$$

Expected Value & Variance of the Posterior Predictive Distribution:

$$\mathbf{E}[Y_{\text{new}}|\mathbf{y}] = \mu_1 \quad \text{and} \quad \text{Var}[Y_{\text{new}}|\mathbf{y}] = \sigma_0^2 + \frac{\sigma_1^2}{n}$$

Final Form of the Predictive Distribution:

$$Y_{\text{new}}|\mathbf{y} \sim N\left(\mu_1, \sigma_0^2 + \frac{1}{n}\sigma_1^2\right)$$

Posterior Predictive Distribution

Model Diagnostics:

- Simulated data from the posterior predictive distribution can be used to assess model fit (Gelman et al., 2003).
- Poor fit occurs when replicated data differ significantly from observed data.
- The posterior predictive distribution enables explicit model comparison (Chen, Dey, Ibrahim, 2000).

Posterior Predictive Distribution: Monte Carlo Sampling

- While the form of $p(y_{\text{new}}|\mathbf{y})$ can sometimes be derived analytically, it is often more practical to sample from $p(y_{\text{new}}|\mathbf{y})$ using Monte Carlo methods.
- The procedure is:
  1. For $j = 1, \ldots, J$, sample $\mu^{[j]}$ from $p(\mu|\mathbf{y})$.
  2. Then, sample $y_{\text{new}}^{[j]}$ from $p(y_{\text{new}}|\mu^{[j]})$.
- The resulting $y_{\text{new}}^{[1]}, \ldots, y_{\text{new}}^{[J]}$ form an independent and identically distributed (iid) sample from $p(y_{\text{new}}|\mathbf{y})$.
- See the `R` example using the lead data for an implementation.

Outline

- Model Quality/Adequacy
    - Sensitivity Analysis
    - Posterior Predictive Distribution
- Hypothesis Testing
    - Classical (Frequentist) Hypothesis Testing
    - Bayesian Hypothesis Testing
    - Bayes Factor
    - BIC

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

Hypothesis Testing: Classical Approach

- Classical hypothesis testing focuses on the ***p*-value**: the probability (under H_0) that a test statistic would take a value as extreme as, or more favorable to, H_a than the observed value.
- Consider iid data $\mathbf{y} = y_1, \ldots, y_n$ from $f(y|\theta)$, where $-\infty < \theta < \infty$.
- We test $H_0 : \theta \leq 0$ vs. $H_a : \theta > 0$ using a test statistic $T(\mathbf{Y})$, a function of the data.
- If the observed test statistic is $T(\mathbf{y}) = T^*$, the *p*-value is:

$$p\text{-value} = P(T(\mathbf{Y}) \geq T^*|\theta = 0) = \int_{T^*}^{\infty} f_T(t|\theta = 0)dt$$

where $f_T(t|\theta)$ is the density function of $T(\mathbf{Y})$.

Issues with Classical Hypothesis Testing

- The *p*-value averages over possible values of T (and thus sample values) that did not occur and are unlikely to occur.
- This approach violates the **Likelihood Principle**, as it relies on hypothetical data rather than solely the observed data.
- The concept of repeated testing, which motivates the probabilities of **Type I** and **Type II** errors, becomes questionable in contexts where the study cannot be replicated.

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

The Bayesian Approach to Hypothesis Testing

- In Bayesian hypothesis testing, we compute the posterior probabilities that θ falls within the null or alternative regions.
- Consider a one-sided hypothesis test of the form:

$$H_0 : \theta \leq c \quad \text{vs.} \quad H_a : \theta > c$$

  for some constant c, where $-\infty < \theta < \infty$.
- We can assign prior probabilities to θ, such that:

$$p_0 = P(-\infty < \theta \leq c) = P(\theta \in \Theta_0)$$

  and

$$p_1 = 1 - p_0 = P(c < \theta < \infty) = P(\theta \notin \Theta_0)$$

  where Θ_0 represents the set of θ-values for which H_0 holds.

The Bayesian Approach (continued)

- The posterior probability that H_0 is true is given by:

$$P(\theta \in \Theta_0|\mathbf{y}) = \int_{-\infty}^{c} p(\theta|\mathbf{y})d\theta$$

- Using Bayes' Law, this can be expressed as:

$$P(\theta \in \Theta_0|\mathbf{y}) = \frac{\int_{-\infty}^{c} p(\mathbf{y}|\theta)p_0 d\theta}{\int_{-\infty}^{c} p(\mathbf{y}|\theta)p_0 d\theta + \int_{c}^{\infty} p(\mathbf{y}|\theta)p_1 d\theta}$$

- The denominator is the marginal likelihood of **Y**, which normalizes the posterior distribution.

The Bayesian Approach: Simplified with Uninformative Priors

- Often, we use an uninformative prior where $p_0 = p_1 = \frac{1}{2}$.
- In this case, the posterior probability $P(\theta \in \Theta_0|\mathbf{y})$ simplifies to:

$$P(\theta \in \Theta_0|\mathbf{y}) = \frac{\int_{-\infty}^{c} p(\mathbf{y}|\theta)d\theta}{\int_{-\infty}^{\infty} p(\mathbf{y}|\theta)d\theta}$$

- This simplifies the calculation, focusing on the ratio of the integral over the null region to the total likelihood of the data.

Hypothesis Testing Example: Coal Mining Strike Data

- **Example 1**: Coal mining strike data
- Let Y represent the number of strikes in a sequence before cessation.
- We have data $y_1, \ldots, y_{11}$ for 11 such sequences in France.
- While a Poisson model is natural, the variance in these data greatly exceeds the mean.
- Thus, we select a geometric model Geometric(θ), where:

$$f(y|\theta) = \theta(1-\theta)^y$$

- Here, θ is the probability of cessation of the strike sequence, and y_i represents the number of strikes before cessation.
- We use a prior for θ such that:

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$$

Hypothesis Testing Example: Coal Mining Strike Data (continued)

- The posterior distribution for θ is:

$$p(\theta|\mathbf{y}) \propto p(\theta)L(\theta|\mathbf{y}) = \theta^{n-1}(1-\theta)^{\sum y_i - 1/2}$$

- This is kernel of the Beta$(n, \sum y_i + 1/2)$ distribution.
- We test the hypothesis:

$$H_0 : \theta \leq 0.05 \quad \text{vs.} \quad H_a : \theta > 0.05$$

- The posterior probability $P(\theta \leq 0.05|\mathbf{y})$ is:

$$P(\theta \leq 0.05|\mathbf{y}) = \int_0^{0.05} p(\theta|\mathbf{y}) d\theta$$

- This is the area to the left of 0.05 in the Beta$(n, \sum y_i + 1/2)$ density, and can be computed directly or via Monte Carlo methods.
- See the `R` example with coal mining strike data.

Two-Sided Hypothesis Tests

- Two-sided hypothesis tests take the form:

$$H_0 : \theta = c \quad \text{vs.} \quad H_a : \theta \neq c$$

  for some constant c.
- A continuous prior on θ is not suitable for this test because it would lead to:

$$P(\theta \in \Theta_0) = 0 \quad \text{and} \quad P(\theta \in \Theta_0 | \mathbf{y}) = 0$$

  for any observed data set $\mathbf{y}$.
- **Solution 1:** One solution is to place a prior probability mass on the point $\theta = c$, but many Bayesians find this approach problematic.
- The difficulty lies in assigning an appropriate value to the point mass, which can significantly influence the posterior results.

Two-Sided Tests: Solutions

- **Solution 2**: Define a small $\varepsilon > 0$ such that if θ is within ε of c, it is considered "practically indistinguishable" from c.
- Set $\Theta_0 = [c - \varepsilon, c + \varepsilon]$ and compute the posterior probability that $\theta \in \Theta_0$.
- **Example 1**: Testing $H_0 : \theta = 0.10$ vs. $H_a : \theta \neq 0.10$ with $\varepsilon = 0.003$. Here, $\Theta_0 = [0.097, 0.103]$ and:

$$P(\theta \in \Theta_0 | \mathbf{y}) = \int_{0.097}^{0.103} p(\theta | \mathbf{y}) d\theta = 0.033$$

  (calculated using *R*).
- **Solution 3** (mimicking the classical approach): Derive a 100(1 - α)% highest posterior density (HPD) credible interval for θ. Reject $H_0 : \theta = c$ at level α if and only if c lies outside the credible interval.

Two-Sided Tests: Bayesian Decision Theory

- In Bayesian decision theory, we incorporate the **cost** of making incorrect decisions regarding H_0 or H_a using a **loss function**.
- The goal is to evaluate the **Bayes risk** of a decision rule, which is the expected loss based on the posterior distribution of θ.
- This approach provides a framework for making decisions that account for both the uncertainty in θ and the consequences of incorrect conclusions.

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

The Bayes Factor

- The **Bayes Factor** provides a way to formally compare two competing models, say M_1 and M_2.
- It is similar to testing a "full model" vs. "reduced model" (e.g., with a likelihood ratio test) in classical statistics.
- However, with the **Bayes Factor**, one model does not have to be nested within the other.
- Given a data set **y**, we compare models:

$$M_1 : f_1(\mathbf{y}|\theta_1) \quad \text{and} \quad M_2 : f_2(\mathbf{y}|\theta_2)$$

- We may specify prior distributions $p_1(\theta_1)$ and $p_2(\theta_2)$, which lead to prior probabilities for each model $p(M_1)$ and $p(M_2)$.

The Bayes Factor (continued)

- By Bayes' Law, the **posterior odds** in favor of Model 1 versus Model 2 is:

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{\frac{\int_{\Theta_1} p(M_1) f_1(\mathbf{y}|\theta_1) p_1(\theta_1) d\theta_1}{p(\mathbf{y})}}{\frac{\int_{\Theta_2} p(M_2) f_2(\mathbf{y}|\theta_2) p_2(\theta_2) d\theta_2}{p(\mathbf{y})}}$$

- Simplifying this, we get:

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(M_1)}{p(M_2)} \cdot \frac{\int_{\Theta_1} f_1(\mathbf{y}|\theta_1) p_1(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(\mathbf{y}|\theta_2) p_2(\theta_2) d\theta_2}$$

- This gives us:

$$[\text{posterior odds}] = [\text{prior odds}] \times [\text{Bayes Factor } BF(\mathbf{y})]$$

The Bayes Factor (continued)

- Rearranging, the **Bayes Factor** is:

$$BF(\mathbf{y}) = \frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} \times \frac{p(M_2)}{p(M_1)}$$

- This simplifies to:

$$BF(\mathbf{y}) = \frac{p(M_1|\mathbf{y})/p(M_2|\mathbf{y})}{p(M_1)/p(M_2)}$$

- The **Bayes Factor** is the ratio of the posterior odds for M_1 to the prior odds for M_1.

The Bayes Factor (continued)

- **Note**: If the prior model probabilities are equal, i.e., $p(M_1) = p(M_2)$, then the **Bayes Factor** equals the posterior odds for M_1.
- **Note**: If $p(M_1) = p(M_2)$ and the parameter spaces Θ_1 and Θ_2 are the same, the **Bayes Factor** reduces to a likelihood ratio.
- Note also that in general:

$$BF(\mathbf{y}) = \frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} \times \frac{p(M_2)}{p(M_1)} = \frac{\frac{p(M_1,\mathbf{y})}{p(\mathbf{y})p(M_1)}}{\frac{p(M_2,\mathbf{y})}{p(\mathbf{y})p(M_2)}}$$

$$= \frac{p(M_1,\mathbf{y})}{p(M_1)} \Big/ \frac{p(M_2,\mathbf{y})}{p(M_2)} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}.$$

- This shows that the **Bayes Factor** is the ratio of the likelihoods under each model. $\neq LRT = \frac{\sup L_{\Theta_1}}{\sup L_{\Theta_1 \cup \Theta_2}}$

Computing the Bayes Factor with Multiple Parameters

- When models M_1 and M_2 specify parameter spaces Θ_1 and Θ_2 (instead of single values for the parameters θ_1 and θ_2), the **Bayes Factor** $BF(\mathbf{y})$ compares the marginal likelihoods of each model:

  $$BF(\mathbf{y}) = \frac{\int_{\Theta_1} \underline{f_1(\mathbf{y}|\theta_1)}\,\underline{p_1(\theta_1)}\, d\theta_1}{\int_{\Theta_2} \underline{f_2(\mathbf{y}|\theta_2)}\,\underline{p_2(\theta_2)}\, d\theta_2}$$

  where

  - $f_1(\mathbf{y}|\theta_1)$ and $f_2(\mathbf{y}|\theta_2)$ are likelihoods under each model
  - $p_1(\theta_1)$ and $p_2(\theta_2)$ are priors over parameter spaces
- This integration accounts for all possible parameter values, weighted by their priors.
- **Approximations**: For complex models, use methods like:
  - Monte Carlo integration
  - Importance sampling
  - Laplace approximations
  - Markov Chain Monte Carlo (MCMC)

The Bayes Factor (continued)

- A **Bayes Factor** much greater than 1 supports Model 1 over Model 2.
- **Jeffreys' Rules** for interpreting the Bayes Factor when Model 1 represents the null model:
  - $BF(\mathbf{y}) \geq 1$: Model 1 supported
  - $0.316 \leq BF(\mathbf{y}) < 1$: Minimal evidence against Model 1 (Note: $0.316 = 10^{-1/2}$)
  - $0.1 \leq BF(\mathbf{y}) < 0.316$: Substantial evidence against Model 1
  - $0.01 \leq BF(\mathbf{y}) < 0.1$: Strong evidence against Model 1
  - $BF(\mathbf{y}) < 0.01$: Decisive evidence against Model 1
- Clearly, these labels are somewhat arbitrary.

The Bayes Factor and Posterior Probability of Model 1

- When comparing two models, M_1 and M_2, the posterior probability of Model 1 can be expressed i.t.o. the Bayes Factor $BF(\mathbf{y})$.
- By Bayes' Rule, $P(M_1|\mathbf{y}) = \dfrac{P(\mathbf{y}|M_1)P(M_1)}{P(\mathbf{y})}$.
- Since $P(\mathbf{y}) = P(\mathbf{y}|M_1)P(M_1) + P(\mathbf{y}|M_2)P(M_2)$, we can express $P(M_1|\mathbf{y})$ as:

$$P(M_1|\mathbf{y}) = \frac{P(\mathbf{y}|M_1)P(M_1)}{P(\mathbf{y}|M_1)P(M_1) + P(\mathbf{y}|M_2)P(M_2)}$$

- Using the definition of the Bayes Factor, $BF(\mathbf{y}) = \frac{P(\mathbf{y}|M_1)}{P(\mathbf{y}|M_2)}$, we can rewrite this as: $P(M_1|\mathbf{y}) = \dfrac{BF(\mathbf{y}) \cdot P(M_1)}{BF(\mathbf{y}) \cdot P(M_1) + P(M_2)}$
- Therefore, the posterior probability of Model 1 is:

$$\boxed{P(M_1|\mathbf{y}) = \frac{1}{1 + \frac{1}{BF(\mathbf{y})} \cdot \frac{P(M_2)}{P(M_1)}}}$$

Example 2(a): Comparing Two Means (Bayes Factor Approach)

- **Data**: Blood pressure reduction was measured for 11 patients who took calcium supplements and 10 patients who took a placebo.
- The data are modeled as normally distributed with a common variance:

  indep

  - **Calcium group**: $Y_{1j} \overset{iid}{\sim} N(\mu_1, \sigma^2)$, $j = 1, \ldots, 11$
  - **Placebo group**: $Y_{2j} \overset{iid}{\sim} N(\mu_2, \sigma^2)$, $j = 1, \ldots, 10$
- Consider the two-sided test for whether the mean blood pressure reduction differs between the two groups:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 \neq \mu_2$$

Example: Comparing Two Means (continued)

- We will place a prior on the difference of standardized means:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$$

  with mean μ_Δ and variance σ^2_Δ.

- Consider the classical two-sample t-statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\underbrace{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}_{s^2_{\text{pooled}}}}} \cdot \sqrt{n^*}$$

effective sample size

where $n^* = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}$.

Example: Comparing Two Means (continued)

- H_0 and H_a define two models for the distribution of the t-statistic T:
  - Under H_0, $T \sim$ t (central) with $n_1 + n_2 - 2$ degrees of freedom.
  - Under H_a, $T \sim$ noncentral t-distribution with noncentrality parameter $\underline{\mu_\Delta \sqrt{n^*}}$.
- Using the prior, the **Bayes Factor** for H_0 over H_a is:

$$BF(\mathbf{y}) = \frac{t_{n_1+n_2-2}(t^* \mid 0, 1)}{t_{n_1+n_2-2}(t^* \mid \mu_\Delta \sqrt{n^*}, 1 + n^* \sigma_\Delta^2)}$$

  where:
  - $t_{n_1+n_2-2}(x \mid \mu, \sigma^2)$: t-distribution density with $n_1 + n_2 - 2$ degrees of freedom, noncentrality μ, and variance σ^2
  - t^*: observed t-statistic
- See the R example to compute $BF(\mathbf{y})$ and $P(H_0 \mid \mathbf{y})$.

Example 2(a): Comparing Two Means (Gibbs Sampling Approach)

μ_1: mean BP red w/ calcium
reduction

μ_2: mean BP red w/ placebo
reduction

- We revisit the same data set, now testing whether calcium yields a better BP reduction than the placebo:

$$H_0 : \mu_1 \le \mu_2 \quad \text{vs.} \quad H_a : \mu_1 > \mu_2$$

makes conditional posteriors easy.

for Gibbs sampling

- We set up the following sampling model:

$$Y_{1j} = \mu + \tau + \varepsilon_{1j}, \quad j = 1, \ldots, 11$$

$$Y_{2j} = \mu - \tau + \varepsilon_{2j}, \quad j = 1, \ldots, 10$$

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

μ, τ, σ^2

- Hence, $\mu_1 = \mu + \tau$ and $\mu_2 = \mu - \tau$.

$H_0: \tau \le 0$ vs $H_a: \tau > 0$

Example: Comparing Two Means (continued)

- We assume independent priors for μ, τ, and σ^2:

$$\mu \sim N(\mu_\mu, \sigma_\mu^2), \quad \tau \sim N(\mu_\tau, \sigma_\tau^2), \quad \sigma^2 \sim IG\left(\frac{\nu_1}{2}, \frac{\nu_1\nu_2}{2}\right)$$

- The Gibbs sampling process iteratively samples from the following full conditional distributions:
  - $\mu | \mathbf{y}_1, \mathbf{y}_2, \tau, \sigma^2 \sim$ Normal
  - $\tau | \mathbf{y}_1, \mathbf{y}_2, \mu, \sigma^2 \sim$ Normal
  - $\sigma^2 | \mathbf{y}_1, \mathbf{y}_2, \mu, \tau \sim$ Inverse Gamma (IG)

In R, $\mu \sim N(2, 50)$

$\tau \sim N(\underline{0}, 50)$

$\sigma^2 \sim IG(1/2, 10)$

- Recall that each conditional distribution leverages the latest sampled values for the other parameters, creating a Markov chain that approximates the joint posterior distribution.
- The specific parameters for these distributions can be found in the accompanying `R` code.

Full Conditional Distributions in Gibbs Sampling

- **Conditional distribution for σ^2:**

$$\sigma^2 \mid \mathbf{y}_1, \mathbf{y}_2, \mu, \tau \sim IG\left(\frac{\nu_1 + n_1 + n_2}{2}, \frac{\nu_1\nu_2 + \sum_{j=1}^{n_1}(y_{1j} - \mu - \tau)^2 + \sum_{j=1}^{n_2}(y_{2j} - \mu + \tau)^2}{2}\right)$$

- **Conditional distribution for μ:**

$$\mu \mid \mathbf{y}_1, \mathbf{y}_2, \tau, \sigma^2 \sim N\left(\frac{\frac{\mu_\mu}{\sigma_\mu^2} + \frac{\sum_{j=1}^{n_1}(y_{1j}-\tau)+\sum_{j=1}^{n_2}(y_{2j}+\tau)}{\sigma^2}}{\frac{1}{\sigma_\mu^2} + \frac{n_1+n_2}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{n_1+n_2}{\sigma^2}}\right)$$

- **Conditional distribution for τ:**

$$\tau \mid \mathbf{y}_1, \mathbf{y}_2, \mu, \sigma^2 \sim N\left(\frac{\frac{\mu_\tau}{\sigma_\tau^2} + \frac{\sum_{j=1}^{n_1}(y_{1j}-\mu)-\sum_{j=1}^{n_2}(y_{2j}-\mu)}{\sigma^2}}{\frac{1}{\sigma_\tau^2} + \frac{n_1+n_2}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_\tau^2} + \frac{n_1+n_2}{\sigma^2}}\right)$$

Derivation of the Posterior Conditionals (Sketch)

Model Setup:

- **Two Groups with Measurements**:

$$Y_{1j} = \mu + \tau + \varepsilon_{1j}, \quad j = 1, \ldots, n_1$$

$$Y_{2j} = \mu - \tau + \varepsilon_{2j}, \quad j = 1, \ldots, n_2$$

  where $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- This implies:

$$Y_{1j} \sim N(\mu + \tau, \sigma^2) \quad \text{and} \quad Y_{2j} \sim N(\mu - \tau, \sigma^2)$$

Prior Distributions:

- **Independent priors for μ, τ, and σ^2**:

$$\mu \sim \underline{N(\mu_\mu, \sigma^2_\mu)}, \quad \tau \sim \underline{N(\mu_\tau, \sigma^2_\tau)}, \quad \sigma^2 \sim \underline{IG\left(\frac{\nu_1}{2}, \frac{\nu_1 \nu_2}{2}\right)}$$

- Recall that *IG* denotes the Inverse Gamma distribution, which is commonly used as a prior for variance parameters.

Derivation of the Posterior Conditionals (Sketch)

- The **joint posterior** distribution of μ, τ, and σ^2 is proportional to the likelihood times the prior:

$$p(\mu, \tau, \sigma^2 \mid \mathbf{y}_1, \mathbf{y}_2) \propto \overbrace{p(\mu) \cdot p(\tau) \cdot p(\sigma^2)}^{\text{prior}} \cdot \overbrace{p(\mathbf{y}_1, \mathbf{y}_2 \mid \mu, \tau, \sigma^2)}^{\text{likelihood}}$$

- Conditional distribution for σ^2 given μ, τ, and the data is derived from combining the likelihood and the prior for σ^2: $p(\sigma^2 \mid \mu, \tau, \mathbf{y}_1, \mathbf{y}_2) \propto p(\mathbf{y}_1, \mathbf{y}_2 \mid \mu, \tau, \sigma^2) \cdot p(\sigma^2)$ and simplifying further.
- Conditional distribution for μ given τ, σ^2, and the data is derived from combining the likelihood and the prior for μ. $p(\mu \mid \tau, \sigma^2, \mathbf{y}_1, \mathbf{y}_2) \propto p(\mathbf{y}_1, \mathbf{y}_2 \mid \mu, \tau, \sigma^2) \cdot p(\mu)$ and simplifying.
- Conditional distribution for τ given μ, σ^2, and the data is derived from combining the likelihood and the prior for τ. $p(\tau \mid \mu, \sigma^2, \mathbf{y}_1, \mathbf{y}_2) \propto p(\mathbf{y}_1, \mathbf{y}_2 \mid \mu, \tau, \sigma^2) \cdot p(\tau)$ and simplifying.

Example: Comparing Two Means (continued)

- **R Example**: A Gibbs Sampler can be used to obtain approximate posterior distributions for μ and, more importantly, for τ.
- Note that:

$$P(H_a|\vec{y}) = \underline{P(\mu_1 > \mu_2|\mathbf{y}) = \boxed{P(\tau > 0|\mathbf{y})}}$$

- Additionally, we can compute the posterior predictive probability:

$$P(Y_1 > Y_2)$$

- These results provide insight into the effectiveness of calcium in reducing blood pressure compared to the placebo.

Advantages of Bayes Factor

- Bayesian hypothesis testing enables researchers to discriminate evidence of absence from absence of evidence.
- Bayesian results are relatively straightforward to interpret and communicate.
- Bayes factor hypothesis testing encourages researchers to quantify evidence on a continuous scale.
- For most statistical scenarios, Bayes factor hypothesis testing is "now relatively easy." software, package
- Bayesian inference allows researchers to monitor the results as the data accumulate. (eg sequential updating
- Bayes factor hypothesis testing allows researchers to include prior knowledge for a more diagnostic test.
- Models compared do not have to be nested.

Issues with Bayes Factors

- **Note**: When an **improper prior** (one that does not integrate to a finite number over its support) is used for θ, the **Bayes Factor** is not well-defined.
- Recall that $BF(\mathbf{y}) = \frac{\text{Posterior Odds for } M_1}{\text{Prior Odds for } M_1}$, and the "prior odds" is meaningless for an improper prior.
- Several methods exist to define types of Bayes Factors with improper priors (e.g., Local Bayes Factors, Intrinsic Bayes Factors, Partial Bayes Factors, Fractional Bayes Factors), but none are ideal.
- One criticism of Bayes Factors is the (implicit) assumption that one of the competing models (M_1 or M_2) is correct.
- Another criticism is that the Bayes Factor depends heavily on the choice of prior.

Outline

Model Quality/Adequacy

- Sensitivity Analysis
- Posterior Predictive Distribution

Hypothesis Testing

- Classical (Frequentist) Hypothesis Testing
- Bayesian Hypothesis Testing
- Bayes Factor
- BIC

The Bayesian Information Criterion (BIC)

- The **BIC** can be used as a substitute for the Bayes factor to compare two (or more) models.

not exactly Bayesian, derivation relies on Bayes estimation

- Conveniently, the BIC does not require specifying priors.
- For parameters θ and data $\mathbf{y}$, the BIC is calculated as:

Similar to "consistent AIC" of Bozdogan

$$\text{BIC} = -2\log L(\hat{\theta}|\mathbf{y}) + p\log(n)$$

# of parameters not prespecified

where p is the number of free parameters in the model, and $L(\hat{\theta}|\mathbf{y})$ is the maximized likelihood given the observed data y.

- Good models have relatively small BIC values:
  - A small value of $-2\log L(\hat{\theta}|\mathbf{y})$ indicates a good fit to the data.
  - A small value of the "overfitting penalty" term $p\log(n)$ indicates a simple, parsimonious model.

The BIC: Comparing Two Models

- To compare two models M_1 and M_2, we calculate:

$$
\begin{aligned}
S &= \underline{-\frac{1}{2}}\left(\underline{\mathrm{BIC}_{M_1} - \mathrm{BIC}_{M_2}}\right) \\
&= \log L(\hat{\theta}_1|\mathbf{y}) - \log L(\hat{\theta}_2|\mathbf{y}) - \frac{1}{2}(p_1 - p_2)\log(n)
\end{aligned}
$$

- A small value of S would favor M_2, while a large S would favor M_1.
- As $n \to \infty$:

$$\underline{\frac{S - \log(BF(\mathbf{y}))}{\log(BF(\mathbf{y}))} \to 0}$$

- For large n, we have the approximation:

$$\underline{\mathrm{BIC}_{M_1} - \mathrm{BIC}_{M_2}} = -2S \approx -2\log(\underline{BF(\mathbf{y})})$$

The BIC: Additional Notes

- Differences in BIC values can be used to compare several non-nested models.
- The BIC should only be trusted as a substitute for Bayes Factors when:
  1. No reliable prior information is available.
  2. The sample size is quite large.
- See R examples:
  - (1) Calcium data example.
  - (2) Regression example on the Oxygen Uptake data set.

BIC vs AIC - Overview

- **AIC (Akaike Information Criterion)** and **BIC (Bayesian Information Criterion)** are model selection tools balancing model fit and complexity.
- **AIC** formula: $\text{AIC} = \underline{-2\log(L) + 2k}$
  - L is the likelihood of the model, k is the number of parameters.
  - AIC penalizes complexity by $2k$, aiming to balance fit and simplicity.
- **BIC** formula: $\text{BIC} = \underline{-2\log(L) + k\log(n)}$
  - n is the sample size.
  - BIC uses $k\log(n)$, imposing a stronger penalty as n grows.
- **AIC** - Based on **information theory**, aims to minimize the Kullback-Leibler (KL) divergence between the true model and estimated model.
- **BIC** - Based on **Bayesian principles**, approximates the posterior probability of a model, penalizing complexity more strictly as n increases.

Complexity Penalties

- **AIC**: Fixed penalty of $2k$, regardless of sample size. Often selects more complex models, especially with small samples.
- **BIC**: Penalty grows with sample size, $k \log(n)$, leading to simpler models in large datasets.
- **Comparison**:
  - **AIC** is generally more flexible, often better for prediction.
  - **BIC** is more conservative, often better for finding the true model structure.

When to Use AIC vs. BIC

- **AIC** - Preferred for prediction-focused applications; minimizes out-of-sample error.
- **BIC** - Suitable for identifying the most likely true model, particularly useful in scientific contexts.
- **Sample Size** - BIC's penalty increases with n, making it more conservative as sample size grows.

Consistency of AIC and BIC

- **BIC is Consistent**:
    - As $n \to \infty$, BIC will identify the true model with high probability, assuming it's among the candidates.
    - Useful for accurate model selection when the goal is interpretability.
- **AIC is Not Consistent**:
    - AIC may not identify the true model, even as n grows, due to its fixed penalty.
    - However, AIC is **asymptotically efficient**, minimizing prediction error.

$H_0: \theta \in \Theta_0$ vs $H_1: \theta \in \Theta_1$

$$P(H_j|\text{data}) = \frac{p(\text{data}|H_j)}{p_0\, P(\text{data}|H_0) + p_1\, P(\text{data}|H_1)} \quad ①$$

$P(\text{data}|H_j)$: marginal density of data under H_j.

Since posterior probabilities are sensitive to priors, p_0 & p_1, it is often suggested to use BF instead.

$BF = P(\text{data}|H_0) / p(\text{data}|H_1)$.

Ex 2(a) $H_0: \mu_1 = \mu_2 = \mu$ vs $H_1: \mu_1 \neq \mu_2$

$\Rightarrow \Theta_0 = (\mu, \sigma^2)$ $\Theta_1 = (\mu_1, \mu_2, \sigma^2)$,

- Such Bayesian 2-sample t-tests are found in literature, but only implicitly as a special case of more complex regression formulations or related to estimation problems as in Bolstad (2004)
- Mathematical Derivation of BF(x) is provided in Gönen et al (2004) (on pg 12)
- Many software packages provide pdf of noncentral t having scale parameter 1.0 and a simple modification is needed in a general case

$$T_\nu(t|a,b) = T_\nu(t/\sqrt{b}\,|\,a/\sqrt{b}, 1)/\sqrt{b}$$

e.g. in R, $BF = \dfrac{dt(t^*, n_1+n_2-2)}{\frac{1}{sqrt(post\nu)}\, dt(t^*/sqrt(\nu post), n_1+n_2-2, nc)}$

where $t = t^*$, $\nu_{post} = 1 + n^* \sigma_\Delta^2$, $nc = \dfrac{\sqrt{n^*}\,\mu_\Delta}{\sqrt{1+n^*\sigma_\Delta^2}}$

- For the case $\mu_\Delta = 0$,

$$BF = \left(\frac{1 + t^2/\nu}{1 + t^2/(\nu(1+n^*\sigma_\Delta^2))}\right)^{-(\nu+1)/2} \cdot \sqrt{1 + n^* \sigma_\Delta^2}$$

$H_0: \delta = \mu_1 - \mu_2 = 0$ vs $H_1: \delta \neq 0$

To obtain the usual two-sample t statistic, prior info is modeled for $\Delta = \delta/\sigma$ rather than δ.

Let $\mu = \frac{\mu_1 + \mu_2}{2}$ and reparametrize $\Theta_1: (\mu_1, \mu_2, \sigma^2)$ as (μ, δ, σ^2).

The prior for $\Delta = \frac{\delta}{\sigma} \mid \mu, \sigma^2 \sim N(\mu_\Delta, \sigma^2_\Delta)$

Here (μ, σ^2): nuisance parameters, so we impose for (μ, σ^2): an uninformative prior regardless of $\delta = 0$ or not, (If (μ, σ^2) has informative prior, BF is not necessarily as in the slides).

Uninformative prior ensures BF depends on data only through t^*.

In summary, $\Delta = \delta/\sigma \mid \mu, \sigma^2 \sim N(\mu_\Delta, \sigma^2_\Delta)$ and $p(\mu, \sigma^2) \propto 1/\sigma^2$ (or c/σ^2 for $c > 0$).

- Gonen et al 2005 says $P_0 = 1/2$, is this correct? (only reasonable b/c $p(\mu, \sigma^2)$ is not proper
- BaM says $D = |\mu_1 - \mu_2|/\sigma$, is this correct? (You'd need folded normal prior, very likely $B(x)$ on pg12 will not be as formed).

BaM: Bayesian Analysis, Jeff Gill

REVIEW ARTICLE
https://doi.org/10.1038/s41593-020-0660-4

nature neuroscience

Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence

Christian Keysers[1,2] ✉, Valeria Gazzola[1,2] and Eric-Jan Wagenmakers[2]

Most neuroscientists would agree that for brain research to progress, we have to know which experimental manipulations have no effect as much as we must identify those that do have an effect. The dominant statistical approaches used in neuroscience rely on *P* values and can establish the latter but not the former. This makes non-significant findings difficult to interpret: do they support the null hypothesis or are they simply not informative? Here we show how Bayesian hypothesis testing can be used in neuroscience studies to establish both whether there is evidence of absence and whether there is absence of evidence. Through simple tutorial-style examples of Bayesian *t*-tests and ANOVA using the open-source project JASP, this article aims to empower neuroscientists to use this approach to provide compelling and rigorous evidence for the absence of an effect.

Neuroscientists would need to know and publish whether a manipulation does not have an effect as much as whether it does. One may use drugs to block a candidate pathway. If the drug has an effect, that pathway is involved; if it doesn't, one would like to conclude the pathway is not involved. Or one may alter activity in a brain region *X* and measure behavior *B*. If de-activating *X* changes *B*, *X* is involved in *B*; if *B* remains unchanged, one would like to conclude that *X* is not involved in *B*.

Neuroscience research is characterized by advanced measurement techniques and sophisticated experimental designs, but the data analyses almost always employ the standard framework of frequentist statistics, featuring *P* value null-hypothesis significance testing (NHST). NHST is arguably appropriate when one wants to quantify evidence against the null hypothesis (H_0: there is no effect) and therefore for the presence of an effect (but see ref. [1]); however, NHST is problematic when one wants to quantify evidence for the null hypothesis. It is notoriously difficult to establish whether non-significant results support the null hypothesis (i.e., yield evidence for absence) or are simply not informative (i.e., show absence of evidence[2–4]). NHST biases us to emphasize positive effects, because those are the effects it equips us to quantify, and to ignore null findings. If we agree that the absence of an effect is important information, isn't this bias unacceptable? Here we aim to highlight how an alternative statistical framework—Bayesian inference—can resolve this problem in neuroscience practice.

We will first illustrate why it is problematic to quantify evidence for the null hypothesis based on the dominant frequentist approaches. We will then show how Bayesian statistics provides a way out of this predicament through simple tutorial-style examples of Bayesian *t*-tests and ANOVA using the open-source project JASP[5].

The *P* value predicament

When we conduct a *t*-test to compare two conditions A and B, a resulting *P* value below a critical threshold α shows that one is unlikely to encounter differences this extreme or more if the experimental manipulation had no effect (H_0: $\mu_A = \mu_B$). For a fixed sample size, the smaller the *P*, the more evidence we have against H_0. Fisher argued that a low *P* value signals that "either the null hypothesis is false, or an exceptionally rare event has occurred."[6] But what if we find no significant effect (for example, $P = 0.3$)? Apart from sampling variability (i.e., 'bad luck'), there are two fundamentally different causal explanations for a non-significant *P* value: the manipulation had a non-zero effect, but the sample size was too small to detect it (i.e., there was insufficient power); or the manipulation had no effect (i.e., the true effect is zero). When sample size is small, either explanation is plausible. As sample size grows, a non-significant *P* value increasingly suggests the manipulation did not have an effect (or an effect so small it is not meaningful). While a power analysis can help disentangle these alternatives, the relationship between sample size, power, *P* value and evidence for H_0 is complex enough that we are rightly reticent to draw strong conclusions from a non-significant *P* value. This has been famously and elegantly phrased in the antimetabole: 'absence of evidence [read: the data are not informative, the design was underpowered] is not evidence of absence [read: the data provide support in favor of the null]'[7].

Intuitively, one may believe that if lower *P* values provide more evidence against H_0, higher *P* values should provide more evidence in favor of H_0. We would thus expect that if we simulate truly random data with no effect, high *P* values should be relatively frequent, especially with large sample sizes. This, however, is not the case. When we draw random samples from two identical distributions (i.e., where H_0 is true; Fig. 1a leftmost column), $P < 0.05$ is rare (as expected), but all *P* values are equally likely. As sample size increases, and we thus intuitively have more evidence that the two distributions have the same mean, high *P* values do not become more frequent (Fig. 1a, leftmost column comparing top and bottom row). Higher *P* values are thus not a reliable metric for more evidence for H_0.

Hence, NHST leaves the neuroscientist in a peculiar predicament: significant *P* values indicate evidence against H_0 (but see refs. [1,8]), but non-significant *P* values do not allow us to conclude that the data support H_0. This inherent limitation of *P* values impedes our ability to draw the important conclusion that a manipulation has no effect and hence that a particular molecular pathway or brain circuitry is not involved or that a particular stimulus dimension does not matter for brain activity.

[1]Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands. [2]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands. ✉e-mail: c.keysers@nin.knaw.nl

A Bayesian solution

In contrast to frequentist NHST, which focuses exclusively on the null hypothesis (H_0), Bayesian hypothesis testing aims to quantify the relative plausibility of alternative hypotheses H_1 and H_0 (Box 1).

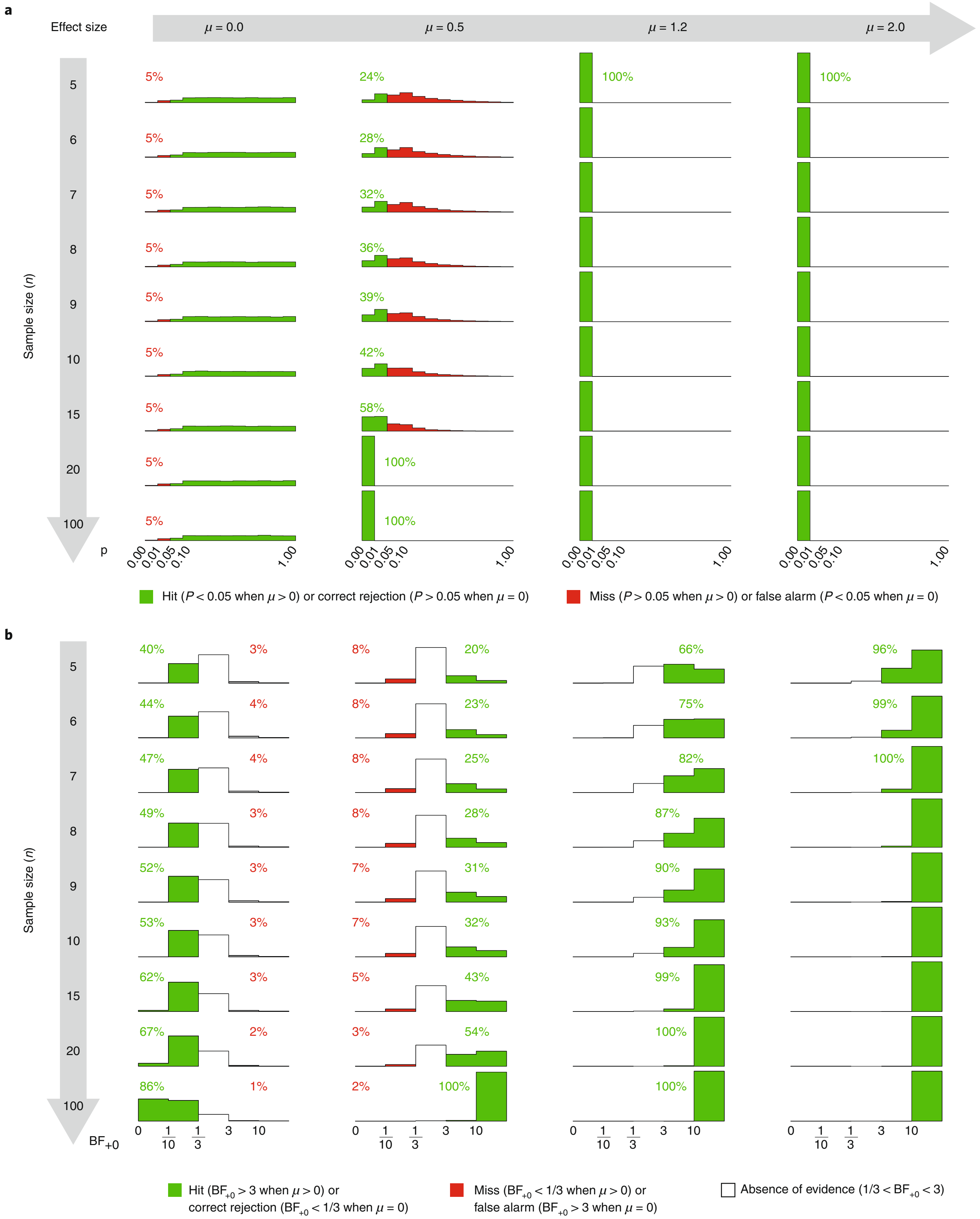
Figure 2 shows an example of how evidence is computed, using a Bayesian approach, for the case of a *t*-test when the question of interest is whether an experimental manipulation has a positive effect. This translates into two rival hypotheses: the manipulation had no effect versus the manipulation increased the dependent variable. Rather than expressing hypotheses in raw values specific to a given experiment, they are expressed using the population standardized effect size δ (with $\delta = (\mu_A - \mu_B)/\sigma$). The sceptic's hypothesis, H_0: $\delta = 0$, states that the effect is absent, whereas the alternative hypothesis, H_+: $\delta > 0$, states that the effect is positive (Fig. 2a). Note that a 'one-tailed' H_1 is denoted as H_+ to indicate the direction of the hypothesized effect. To quantify which hypothesis best predicts the data, we quantify the observed effect size d ($d = (m_A - m_B)/s$) in the data and transform it into a *t*-value ($t = d \times \sqrt{n}$), because the distribution of *t*-values expected for any δ is well known. Next, we transform the qualitative hypotheses H_0 and H_+ into quantitative predictions about the probability of encountering every *t*-value using this *t*-distribution. This is achieved by assigning prior probability distributions to δ (Fig. 2b), and then computing the probability of each observable t based on these δ-value distributions (Fig. 2c). For the sceptic's H_0: $\delta = 0$, the distribution of effect sizes is simply a spike at $\delta = 0$ (red in Fig. 2b), and this makes predictions about the likelihood of each observable *t*-value using the same distribution that is used in a frequentist *t*-test with n participants: the Student's t distribution with $n - 2$ degrees of freedom (red in Fig. 2c). For H_+: $\delta > 0$, we need to be specific about the probability of each possible positive δ to become specific about t. The one-tailed nature of our hypothesis is reflected in a truncated distribution, with negative values having zero probability under H_+ (ref. [9] p. 283; note that two-tailed hypotheses are usually implemented by means of symmetrical distributions, for example, the dotted line in Fig. 3b). We also know that most neuroscience papers report effect sizes of $\delta <$ 1 (ref. [10]), with smaller effect sizes being more common than larger effect sizes; this is reflected in a peak for small positive δ and low probability for $\delta > 1$. Indeed, that we feel that we need to perform a test in the first place corresponds to this presumption that the effect size must be fairly small[9]. These considerations about the plausible direction and magnitudes of the effect under H_+ generate the prior distribution shown in blue in Fig. 2b (see section "Default priors provide an objective anchor" for guidance on how to define this prior distribution). For each of the hypothesized δ values, we can make predictions about t using the non-central t distribution with $\mu = \delta$. The mixture of these non-central *t*-distributions associated with each δ, weighted by the prior plausibility of that δ, predicts the probability of each possible *t*-value under H_+ (blue in Fig. 2c). When the data arrive (Fig. 2d), we first calculate the *t*-value for our data, which we will call t_1, and then see where t_1 falls on the *t*-distribution expected under H_0 (red) and under H_+ (blue). The traditional frequentist P value corresponds to the area to the right of t_1 on the red distribution; note that the predictions from H_+, indicated by the blue distribution, are entirely ignored in the frequentist approach. In contrast, for the Bayesian approach, we take the ordinates $p(t_1 \mid H_0)$ and $p(t_1 \mid H_+)$ and calculate the evidence that the data provide in favor of H_+ over H_0 as $p(t_1 \mid H_+) \div p(t_1 \mid H_0)$ (Fig. 2e). At that specific t_1 value, the ratio equals 4, indicating that our data was predicted four times better by H_+ than H_0; we may conclude that our data supports H_+. The evidence—the relative predictive performance of H_0 versus H_+—is known as the Bayes factor[9,11,12] (Box 1). We abbreviate it as BF and use subscripts to denote which model is in the numerator versus the denominator; thus, $\mathrm{BF}_{+0} = p(t_1 \mid H_+) \div p(t_1 \mid H_0)$ and $\mathrm{BF}_{0+} = p(t_1 \mid H_0) \div p(t_1 \mid H_+)$.

If the *t*-value from our data were to be closer to 0, as exemplified by another hypothetical *t*-value, t_2 (Fig. 2e), the ordinates of the red and blue distributions would be about equally high, indicating that the observed t_2 is about equally likely to occur under H_0 and H_+; hence the predictive performance of H_0 and H_+ is about equal, the Bayes factor is near 1, and consequently we have absence of evidence. If the *t*-value were to fall at t_3 (Fig. 2e), this value would be 4 times more likely to occur under H_0 than under H_+; consequently, $\mathrm{BF}_{+0} = ¼$, that is, $\mathrm{BF}_{0+} = 4$, and we may conclude that our data support H_0—in other words, we have some evidence of absence.

Thus, the P value of a frequentist approach has two logical states, significant versus not significant, which translate into evidence for H_1 ("great, I found the effect") versus a state of suspended disbelief ("I did not find an effect, but it could be because I was unlucky or because the effect does not exist or because my sample size was too small"), whereas the BF has three qualitatively different logical states: $\mathrm{BF}_{10} > x$ ("great, I have compelling evidence for the effect"), $1/x < \mathrm{BF}_{10} < x$ ("oops, my data are not sufficiently diagnostic"), $\mathrm{BF}_{10} < 1/x$ ("great, I have compelling evidence for the absence of the effect"). Here x is the researcher-defined target level of evidence. The BF should primarily be seen as a continuous measure of evidence. However, since larger deviations from 1 provide stronger evidence, Jeffreys proposed reference values to guide the interpretation of the strength of the evidence[9]. These values were spaced out in exponential half steps of 10, $10^{0.5} \approx 3$, $10^1 = 10$, $10^{1.5} \approx 30$, etc., to be equidistant on a log scale. He then compared these values with critical values in frequentist *t*-tests (see Extended Data Fig. 1a for a modern equivalent) and χ^2 tests, and declared, "Users of these tests speak of the 5 per cent point [$p = 0.05$] in much the same way as I

Fig. 1 | *P* value of a *t*-test and BF_{+0} as a function of effect size and sample size. a, Each histogram shows the distribution of P values obtained from 1,000 one-tailed one-sample *t*-tests based on n random numbers drawn from a normal distribution with mean μ and s.d. = 1. To differentiate levels of significance, the first bin was split into multiple bins based on standard critical values. Note how, when there is an effect in the data (i.e., $\mu > 0$, all but leftmost column), increasing sample size (downwards) or effect size (rightwards) leads to a leftwards shift of the distribution: more evidence for an effect leads to lower P values. In this case, P values <0.05 are considered hits and are shown in green, while P values >0.05 are considered misses and shown in red. However, somewhat counterintuitively, the converse does not hold true: in the absence of an effect, ($\mu = 0$, leftwards column), increasing sample size does not lead to a rightward shift (increase) of the P values. Instead the distribution is completely flat, with all P values equally likely (note that the distribution seems to thin out below 0.05, but this is because we subdivided the last leftmost bin into several bins to resolve levels of significance). In this case, $P < 0.05$ represents false alarms, shown in red, and $P > 0.05$ represents correct rejections, shown in green. P values are thus not a symmetrical instrument: cases with much evidence for H_1 (high effect size and sample size) give us quasi-certainty to find a very low P value, whereas cases with much evidence for H_0 (for example, $\mu = 0$ with $n = 100$) do not make P values close to 1 highly likely; instead, any P value remains as likely as any other. **b**, Distribution of BF_{+0} (using $r = \sqrt{2}/2$ for the effect size prior Cauchy width) values obtained from 1,000 *t*-tests based on n random numbers drawn from an $N(\mu,1)$ normal distribution with mean μ and s.d. = 1. Each histogram has the same bounds specified below the graphs, representing conventional limits for moderate and strong evidence. When an effect is absent ($\mu = 0$, leftmost column), evidence of absence (green bars and percentages, $\mathrm{BF}_{+0} < 1/3$) increases with increasing sample size and the false alarm rate is well controlled. When an effect is present ($\mu > 0$), evidence for a positive effect ($\mathrm{BF}_{+0} >$ 3, green bars and green percentages) increases with sample size and effect size, and misses ($\mathrm{BF}_{+0} < 3$, red bars and red percentages) are rare ($\mu = 0.5$) or absent ($\mu = 1.2$ or 2). When percentages are not shown, they are 0% (red) or 100% (green). Data can be found at https://osf.io/md9kp/.

should speak of the K = $10^{1/2}$ [i.e. $BF_{10} = 3$] point, and of the 1 per cent [$p = 0.01$], point as I should speak of the K = 10^1 point [i.e. $BF_{10} = 10$]; and for moderate numbers of observations the points are not very different."[9] These reference values remain in use: BF > 3 is considered moderate evidence for the hypothesis in the numerator (i.e., H_1 if $BF_{10} > 3$), roughly similar to $P < 0.05$; BF > 10 is considered



Box 1 | Bayesian updating

The Bayesian formalism describes how an optimal observer updates beliefs in response to data. In the context of hypothesis testing, at the start, observers entertain a set of two or more rival accounts. In the context of a t-test, they would be called hypotheses H_0 and H_1; in the case of an ANOVA, they would be called models. Each is specified via parameters we can call θ, for example, the effect size δ in a t-test hypothesis or a regression parameter β in an ANOVA. Prior to looking at the data, the rival accounts have prior probabilities, and the parameter values within each account also have prior probabilities. At the level of the accounts, we may assume them to be equally believable a priori (for example, prior hypothesis probabilities $p(H_0) = p(H_1) = 0.5$). At the level of the parameters within each account, they are associated with prior parameter distributions (for example, H_0: $\delta = 0$, H_1: d ~ Cauchy; Fig. 2). When data become available, the probabilities are reallocated: accounts and parameters-within-accounts that predict the data relatively well receive a boost in credibility, whereas those that predict the data poorly suffer a decline[30]. Note the similarity to models of reinforcement learning[31]. Mathematically, this updating is done using Bayes' rule, as we describe below separately for parameters and accounts.

Updating parameter estimates

$$\underbrace{p(\theta|data)}_{\text{posterior beliefs about }\theta} = \underbrace{p(\theta)}_{\text{prior beliefs about }\theta} \times \underbrace{\frac{p(data|\theta)}{p(data)}}_{\text{predictive updating factor}}$$

Here the probability of each possible value of θ within an account after seeing the data (i.e., posterior parameter beliefs) are calculated as the product of the prior probability of that value (i.e. parameter prior beliefs) times the predictive updating factor. The latter reflects how likely the observed data is according to that particular parameter value divided by the average predictive performance across all values of θ weighted by their prior probability, i.e. $p(data) = \int p(data|\theta) \cdot p(\theta) d\theta$. This posterior parameter belief is the basis for the credible intervals (CI) that the Bayesian analysis provides for the parameters conditional on a given model.

Updating the plausibility of the rival accounts

For two rival accounts of the data (for example, H_0 vs H_1), Bayes' rule can best be written in the form of odds[32]:

$$\underbrace{\frac{p(H_0|data)}{p(H_1|data)}}_{\text{posterior odds for }H_0\text{vs }H_1} = \underbrace{\frac{p(H_0)}{p(H_1)}}_{\text{prior odds for }H_0\text{vs }H_1} \times \underbrace{\frac{p(data|H_0)}{p(data|H_1)}}_{\text{predictive updating factor}}$$

This equation shows that the change from prior hypothesis odds to posterior hypothesis odds is brought about by the predictive updating factor—commonly known as the Bayes factor[12]. For instance, assume the rival hypotheses are equally plausible a priori (i.e., $p(H_0) = p(H_1) = 0.5$). The prior hypothesis odds are then equal to one. If the predictive updating factor is 10 (i.e., the observed data is 10 times more likely under H_0 than under H_1), this means that the posterior odds are then also 10. Given that for mutually exclusive hypotheses $p(H_0)+p(H_1) = 1$, these odds mean that the data have increased the probability of H_0 from 0.5 (the prior hypothesis probability) to $10/11 \approx 0.91$ (the posterior H_0 probability).

The Bayes factor quantifies the degree to which the data warrant a change in beliefs, and it therefore represents the strength of evidence that the data provide for H_0 vs H_1. Note that this strength measure is symmetric: evidence may support H_0 just as it may support H_1; neither of the rival hypotheses enjoys a special status.

For a neuroscientist who wants to know whether or not their manipulation had an effect, the posterior odds might seem like the most obvious metric, as they reflect the plausibility of one hypothesis over another after considering the data. However, these posterior odds depend both on the evidence provided by the data (i.e., the Bayes factor) and the prior odds. The prior odds capture subjective beliefs before the experiment and introduce an often-undesirable element of subjectivity that could bias the conclusions drawn from the posterior beliefs. Scientists who embrace a certain theoretical standpoint and those who do not might fiercely disagree on these prior odds while agreeing on the evidence, that is, the extent to which the data should change their beliefs. As beliefs are considered less valuable for scientific reporting than evidence, the data-informed Bayes factor is the less controversial and thus favored metric to report.

There are three broad qualitative categories of Bayes factors. First, the Bayes factor may support H_1; second, the Bayes factor may support H_0; third, the Bayes factor may be near 1 and support neither of the two rival hypotheses. In the second case we have 'evidence of absence', and in the third care we have 'absence of evidence' (see also ref. [2]). More fine-grained classification schemes have been proposed[16].

To develop an intuition for the continuous strength of evidence that a Bayes factor provides, one may use a probability wheel. Examples are shown in Fig. 3b. To construct the wheel, we have assumed that H_0 and H_1 are equally likely; the red part in the wheel is then the posterior probability for H_1, and the blue part is the complementary probability for H_0. Now pretend that the wheel is a pizza, with the red area covered with pepperoni and the blue area covered with mozzarella. Imagine that you poke your finger blindly onto the pizza and that it comes back covered in the non-dominant topping (in this case, pepperoni). How surprised are you? Your level of imagined surprise is an indication for the strength of evidence that a Bayes factor provides. We additionally compare the BF with traditional P values in Extended Data Fig. 1.

strong evidence, roughly similar to $P < 0.01$ (ref. [13]). Because $BF_{10} = 1/BF_{01}$, this also defines the bounds for evidence for the hypothesis in the denominator: BF < 1/3 is moderate and BF < 1/10 is strong evidence. BF values between 1/3 and 3 indicate that there is insufficient evidence to draw a conclusion for or against either hypothesis. While these guidelines enable us to reach somewhat discrete conclusions, the magnitude of the BF should be considered as a continuous quantity, and the strength of the conclusions expressed in the discussion section of a paper should reflect the magnitude of the BF. For new discoveries, Jeffreys suggested that $x = 10$ is more appropriate than $x = 3$; however, each scientist and field will need to decide whether to privilege the sensitivity of the test for small samples or effects by using smaller x values such as 3, or to avoid false conclusions by using higher x values such as 10. Regardless, readers can judge the strength of the evidence directly from the numerical value of BF, with a BF twice as high providing evidence twice as strong. In contrast, it can be difficult to interpret an actual P value as strength of evidence, as $P = 0.01$ does not provide five times as much evidence as $P = 0.05$.

Crucially, the three-state system of the Bayes factor allows us to differentiate between evidence of absence and absence of evidence. This represents a fundamental conceptual step forward in the way we interpret data: instead of one outcome (i.e., $P < \alpha$) that generates knowledge, we now have two (i.e., $BF_{10} > x$ and $BF_{01} > x$).

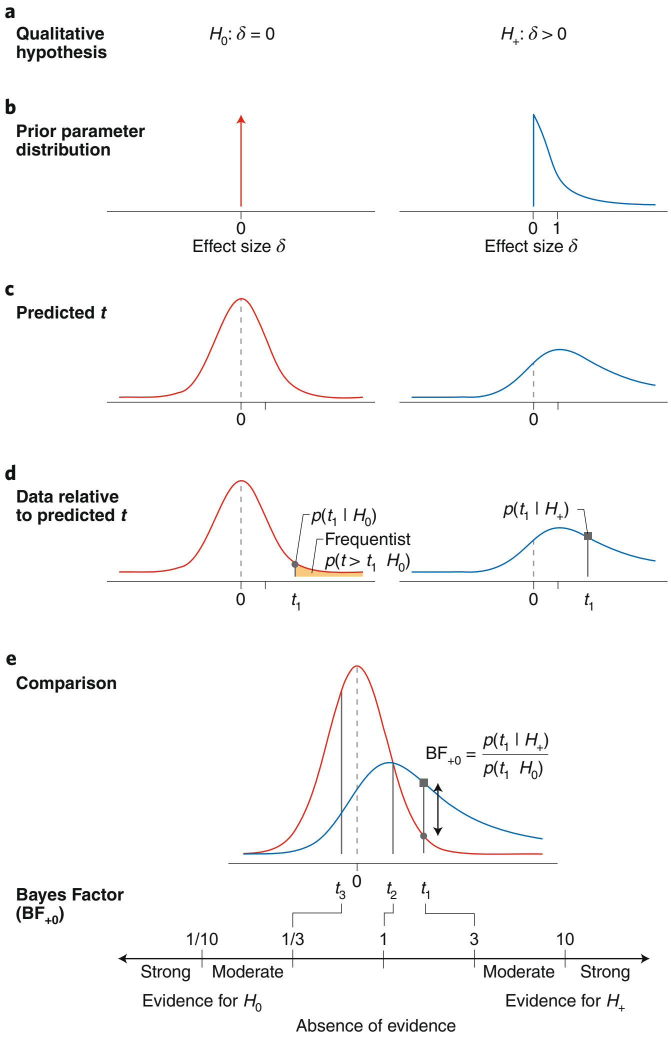


Fig. 2 | Hypothesis testing under the Bayesian framework. a, Two competing qualitative hypotheses are expressed in terms of a test parameter, such as the population effect size δ. H_+ represents a directional hypothesis of a positive effect size. **b**, The two rival hypotheses are formulated in terms of specific probability distributions expressing the plausibility or probability of each effect size value. **c**, Each effect size distribution is transformed into expected *t*-values. For H_0, this is simply the standard *t*-distribution used in frequentist *t*-tests. For H_+, for each hypothesized effect size, a non-central *t*-distribution with that effect size is multiplied with the hypothesized probability of that effect size in **b**. All of these weighted non-central *t*-distributions are then summed together to get the distribution in **c**. **d**, After the data is obtained, the observed *t*-value (t_1) can be interrogated in each distribution. Note that, in frequentist statistics, the *P* value is derived from the H_0 distribution alone, as the area where $t > t_1$. **e**, The likelihood of t_1 under H_0 and H_+ is then compared to calculate the BF_{+0}. Here we illustrate three examples of observed *t*-values. At an observed value of t_1, the blue distribution is 4 times higher than the red; hence $\mathrm{BF}_{+0} = 4$, and we have (moderate) evidence for H_+. At an observed value of t_2, where the two distributions are equal, $\mathrm{BF}_{+0} = 1$ and we have absence of evidence. At an observed value of t_3, the red distribution is 4 times higher than the blue; hence $\mathrm{BF}_{0+} = 4$ and we have moderate evidence for H_0. Here we illustrated one-tailed hypotheses, as these respect the directional nature of the underlying theory and yield more diagnostic predictions. More agnostic two-tailed hypotheses are calculated using the same principles, but the truncated blue distribution in **b** is then replaced with a non-truncated, symmetric distribution, as shown in the dotted line in Fig. 3b. Data can be found at https://osf.io/md9kp/.

Figure 1b shows how a Bayesian *t*-test performs compared to a frequentist *t*-test (Fig. 1a). The target level of evidence was set at $x = 3$, considered similar to the α-level of 0.05 in Fig. 1a (ref. [9]). When an effect is absent ($\mu = 0$), the Bayesian test will seldom come to the erroneous conclusion that an effect is present (less than 4% $\mathrm{BF}_{+0} > 3$), similarly to the frequentist approach. However, unlike the frequentist approach, the Bayesian *t*-test provides increasing evidence for the absence of an effect (see green percentages in Fig. 1b) with increasing sample size. Similarly, evidence for an effect increases as sample size or effect size increases (Extended Data Fig. 1b). Hence, unlike the frequentist *P* value, the BF has a symmetric property of quantifying evidence for the presence or the absence of an effect that scales with evidence in either direction, be it due to increased sample size or effect size. In each case, inconclusive cases (i.e., absence of evidence, defined here as $1/3 < \mathrm{BF} < 3$) become increasingly rare as sample size increases.

Figure 1b also shows the statistical power to provide evidence for or against an effect. When an effect is absent, evidence of absence ($\mathrm{BF}_{+0} < 1/3$) in the presence of noise is limited when sample size is very small (40% at $n = 5$), but reasonable in sample sizes often used in neuroscience ($n = 20$–100). When an effect is present, evidence for the presence of an effect ($\mathrm{BF}_{+0} > 3$) is slightly less frequent than that of the frequentist approach ($P < 0.05$), but not dramatically different. However, as sample sizes become very large, the Bayes factor and *P* values diverge more dramatically: *P* values will become significant even for arguably irrelevantly small effect sizes (for example, at $n = 1{,}000$, $d = 0.05$, $t(999) = d\sqrt{1000}$, $P = 0.05$), whereas the BF continues to require more relevant effect sizes (Extended Data Fig. 1b). It should be noted that for two-tailed tests, evidence for the null hypothesis becomes substantially harder to provide and requires larger sample sizes because the predictions of the null hypothesis are directly flanked by the high likelihood of finding small effect sizes in either direction under H_1.

If Bayesian inference is so simple and informative, why isn't it used more? We speculate that one of the main reasons is pragmatic: until recently it was difficult to conduct Bayesian analyses for standard statistical scenarios. However, a number of packages are now available that make Bayesian hypothesis tests easier to perform. Here we focus on the multiplatform open-source program JASP (Jeffreys's Amazing Statistics Program; https://jasp-stats.org), which uses an accessible graphical user interface; the R-package BayesFactor[14] is a powerful alternative.

JASP, a convenient tool for Bayesian inference

In the JASP graphical user interface, developed to facilitate the adoption of Bayesian inference, analyses are selected from drop-down menus, variables are dragged and dropped into windows, and output is generated on the fly. Increasingly detailed analyses can be executed by ticking checkboxes. As a result, for many statistical scenarios, a comprehensive Bayesian (re)analysis can be performed in a matter of seconds. The examples below showcase the ways in which the output from such Bayesian analyses should be interpreted and how they allow researchers to go beyond the conclusions from the classical frequentist *P* values. On the Open Science Forum (https://osf.io/md9kp/), we provide csv files associated with the examples presented below, as well as R code to replicate the BF values for power users to apply such analyses to a large number of units (for example, to classify hundreds of neurons recorded using calcium imaging into those responding and those not responding to a particular stimulus) and a video illustrating how to use JASP.

Example of a two-sample *t*-test

To illustrate the Bayesian *t*-test, we use an example inspired by ref. [15], in which we hypothesized that the anterior cingulate cortex (ACC) is critical for 'emotional contagion' in rats and that deactivating the ACC by locally injecting muscimol should thus reduce

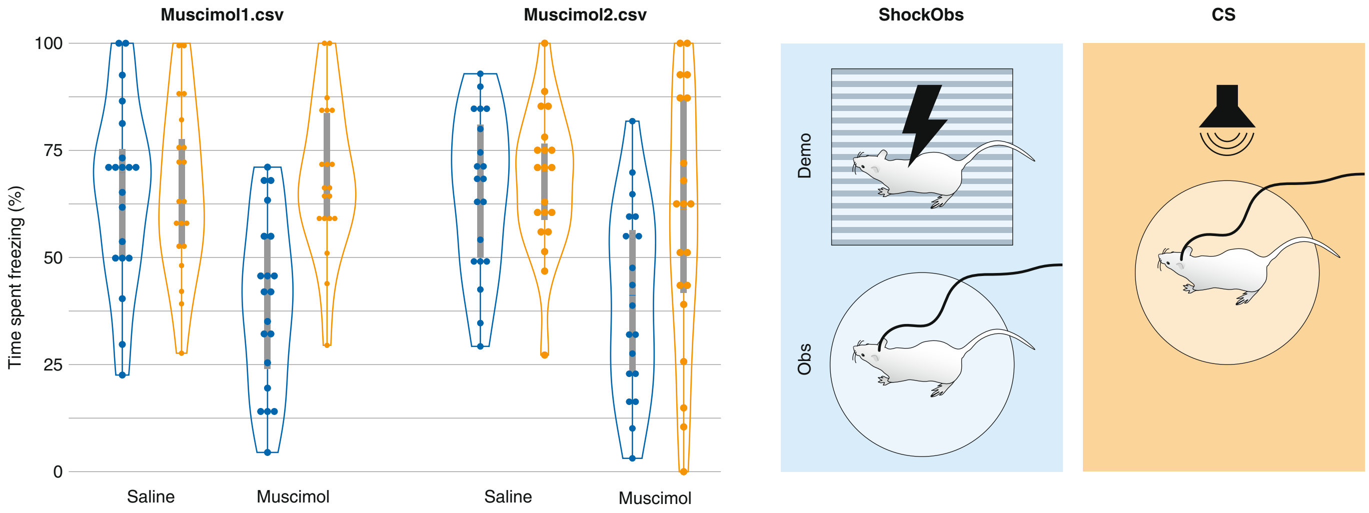


Fig. 3 | Illustration of the data for the two simulated scenarios. Muscimol1 data were simulated using $\mu = 70$ and $\sigma = 20$ for all conditions (imposing a floor of 0 and a ceiling of 100), except ShockObs (in blue) under muscimol, which was simulated using $\mu = 40$. Muscimol2 data were simulated using the same parameters except for CS (in orange) under muscimol, which had $\mu = 65$ and $\sigma = 40$. Based on these data, we should find evidence for H_+: saline > muscimol in all cases for ShockObs. For CS (orange), muscimol1 should reveal evidence for H_0 (evidence of absence) given that data were drawn from the same $\mu = 70$, $\sigma = 20$ distributions. For muscimol2, CS was drawn from different distributions for saline and muscimol, but with $n = 20$, it might be hard to adjudicate the difference, and we might thus expect absence of evidence. Data can be found at https://osf.io/md9kp/. Plots are violin plots, with the gray bar showing the middle two quartiles.

emotional contagion compared to injecting saline. The injected animal observed a conspecific receive electroshocks (ShockObs), and its freezing was measured as an index of emotional contagion. There was a non-social control condition in which the injected animal was exposed to a shock-conditioned tone (CS playback). To illustrate how to analyze this kind of design using Bayesian statistics, we generated two synthetic data sets (see additional materials on OSF (https://osf.io/md9kp/) for muscimol1.csv and muscimol2.csv) that illustrate two slightly different scenarios. We use simulated data rather than the actual data from the paper to guide the reader though alternative scenarios and to allow the reader to modify the data and test the effect this has on the analysis (see additional materials on OSF (https://osf.io/md9kp/) for the script GenerateMuscimolData.R used to generate the data).

Video 1 (see additional materials at https://osf.io/md9kp/) shows how to setup the analyses in JASP to examine the data of Muscimol1.csv. Our main analyses of interest are two independent sample *t*-tests on the freezing measures that compare H_+: saline > muscimol against H_0: saline = muscimol separately for the ShockObs and CS conditions. To assess the specificity of the effect, we will use an ANOVA (see below). We use a one-tailed alternative hypothesis because deactivating the ACC should reduce (not enhance) freezing in the muscimol condition and hence lead to higher freezing in the saline condition. The frequentist approach can also be performed in JASP by selecting 'Independent Samples T-Test'. Thus, this single package enables scientists to combine frequentist and Bayesian approaches on the same data set.

The frequentist approach shows that for ShockObs, muscimol reduced freezing significantly ($t_{(38)} = 3.961$, $P < 0.001$), i.e., the observed difference in freezing is unlikely under H_0. For CS, the result is nonsignificant ($t_{(38)} = -0.519$, $P = 0.7$), which could signal evidence for absence or absence of evidence. To adjudicate between these alternative interpretations, we perform the 'Bayesian Independent Samples T-Test'. Here too we select ShockObs and CS as dependent variables, group as the Grouping Variable, and the one-tailed group1>group2 analysis (after selecting saline as group1 and muscimol as group2 in the data viewer as shown in Video1). The results are shown in Fig. 4.

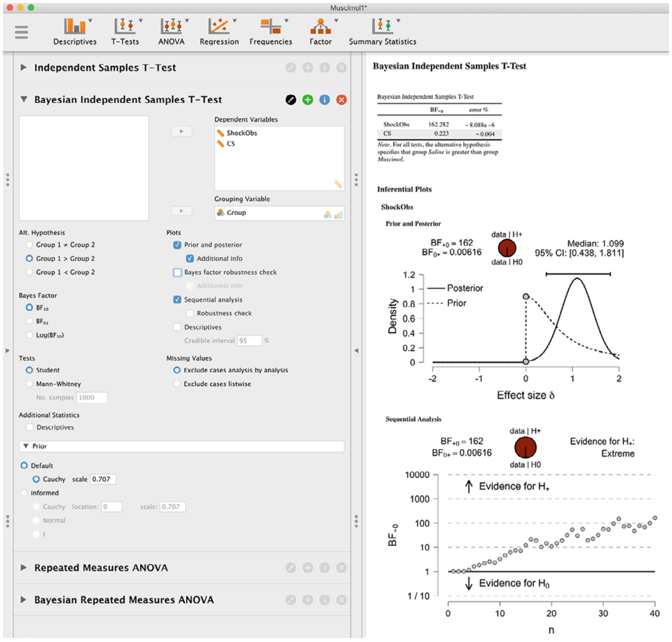
In the input panel on the left, we select BF_{10} as the output, i.e., $p(\text{data} \mid H_+) \div p(\text{data} \mid H_0)$, with a one-tailed hypothesis of group1[saline] > group2[muscimol]. The results table on the right summarizes the main outcomes. For ShockObs, $\text{BF}_{+0} = 162.282$, indicating that the data are 162 times more likely under H_+ than under H_0. The data thus provides what is considered extremely strong evidence for our hypothesized reduction in socially triggered freezing following ACC deactivation. For CS, $\text{BF}_{+0} = 0.223$. This value is below 1/3 and, according to the classification scheme by Jeffreys[9,16], our data thus provide moderate evidence for H_0, i.e., that ACC deactivation does not lead to a reduction of non-socially triggered freezing. Switching to option BF_{01} in the lefthand panel inverts the Bayes factor: now BF_{0+} for CS equals 4.494 (1/0.223), meaning that the data are 4.5 times more likely under H_0 than under H_+.

For the muscimol2 data, the frequentist *t*-test again reveals a significant reduction in ShockObs ($t_{(38)} = 3.8$, $P < 0.001$) and a non-significant result for CS ($t_{(38)} = 1.2$, $P = 0.11$). The Bayesian analysis confirms that the data provide extremely strong evidence for a reduction of freezing for ShockObs ($\text{BF}_{+0} = 120$). However, this time, for CS, $\text{BF}_{+0} = 0.97$. This result indicates an absence of evidence (in contrast to muscimol1, which showed moderate evidence of absence).

Example of an ANOVA

We can also examine whether muscimol had a greater effect on ShockObs than on CS by assessing evidence for an interaction between group (saline vs muscimol) and condition (ShockObs vs CS)[17,18]. In a frequentist approach, we can conduct this analysis using the JASP 'Repeated Measures ANOVA' (rmANOVA) menu option. The results show significant main effects of condition ($F_{(1,38)} = 14.6$, $P < 0.001$) and group ($F_{(1,38)} = 5.4$, $P = 0.026$) and a significant condition × group interaction ($F_{(1,38)} = 14.3$, $P < 0.001$). We can also perform this analysis using the 'Bayesian Repeated Measures ANOVA' menu option (Fig. 5), the functionality of which is based on the BayesFactor R package[19].

The Bayesian approach to the rmANOVA is to compare the predictive performance of models with and without each of the factors



▶ Independent Samples T–Test

▼ Bayesian Independent Samples T–Test

Dependent Variables: ShockObs, CS

Grouping Variable: Group

Alt. Hypothesis
○ Group 1 ≠ Group 2
● Group 1 > Group 2
○ Group 1 < Group 2

Bayes Factor
● BF_{10}
○ BF_{01}
○ $Log(BF_{10})$

Tests
● Student
○ Mann–Whitney
No. samples 1000

Additional Statistics
□ Descriptives

Plots
☑ Prior and posterior
☑ Additional info
□ Bayes factor robustness check
Additional info
☑ Sequential analysis
□ Robustness check
□ Descriptives
Credible interval 95 %

Missing Values
● Exclude cases analysis by analysis
○ Exclude cases listwise

▼ Prior
● Default
● Cauchy scale 0.707
○ Informed
○ Cauchy location: 0 scale: 0.707
○ Normal
○ t

▶ Repeated Measures ANOVA

▶ Bayesian Repeated Measures ANOVA

Bayesian Independent Samples T-Test

Bayesian Independent Samples T-Test

	BF_{+0}	error %
ShockObs	162.282	~ 8.088e −6
CS	0.223	~ 0.004

Note. For all tests, the alternative hypothesis specifies that group *Saline* is greater than group *Muscimol*.

Inferential Plots

ShockObs

Prior and Posterior

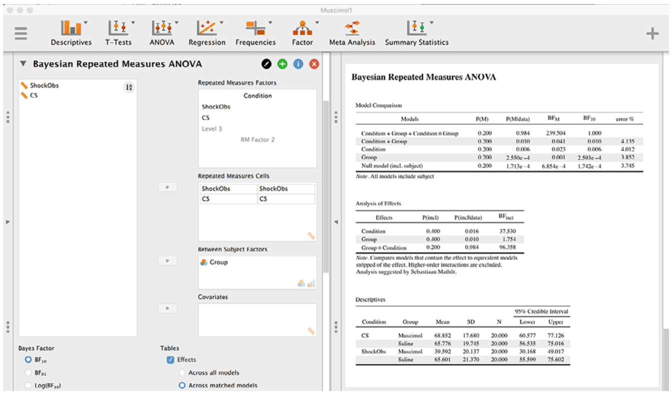
Sequential Analysis

Fig. 4 | Screenshot from the 'Bayesian Independent Samples T-Test' in JASP. Top right: the Bayes factor for the two variables, followed by the inferential plot showing the credible interval of the effect size and the sequential analysis. The inferential plots shown on the right are discussed in sections "Default priors provide an objective anchor" and "Accumulation of evidence." Data can be found at https://osf.io/md9kp/, including a muscimol1.jasp file that can be loaded to replicate the analysis within JASP or to view the results of the analysis within OSF.

and interactions. Conceptually, it starts from a null model that predicts data based on a constant for each subject without considering any experimental factors. It computes the likelihood L_{null} of that null model, i.e., the probability of the observed data D under this null model. It then also calculates the likelihood $\mathrm{L}_{\text{group}}$ of a model additionally including an effect of group. If the Bayes factor calculated as $\mathrm{L}_{\text{group}}/\mathrm{L}_{\text{null}}$ is >1, there is evidence for the effect of group. If BF < 1, i.e., the null model outperforms the more complex group model, there is evidence for the absence of an effect of group. If BF ≈ 1 we have evidence of absence. This Bayes factor can be interpreted using the same bounds discussed in Fig. 2 and Extended Data Fig. 1.

Complex models always fit data at least as well as simpler models. How can a simpler model thus ever outperform a more complex model in the Bayesian sense? The answer is simple: a Bayes factor model comparison does not compare the fit of models for a specific parameter value (i.e., the maximum likelihood) but the predictive performance of models across all plausible parameter values (i.e., average likelihood)[20–22]. If we consider the models $D =$ subject $+ \beta \times$ group (i.e., the group model) and $D =$ subject (i.e., the null model), the average likelihood of the data under the models is the weighted average of the probability of the data D under the full range of plausible values assigned to β in the parameter prior: $\mathcal{L} = \int P(D|\beta)P(\beta)d\beta$. Hence, the null model's $\mathcal{L}$ is calculated entirely at $\beta = 0$, whereas the group model's $\mathcal{L}$ considers $\beta = 0$, but averaged with the predictions from all other plausible β values. The effect of this integration over β can be appreciated in Extended Data Fig. 2. Essentially, because the null model concentrates all its predictions on $\beta = 0$, small differences across the two groups are more likely under this null model, providing evidence for absence.

Figure 5 applies this logic to our data. The top table in the output panel indicates all the models that are being considered and compared. This includes the abovementioned null model with subject constants only, a model that adds the effect of condition, one adding only group, one adding both main effects and one also including the interaction. The $P(M)$ column indicates the prior probabilities of these various rival models, which are set equal so as not to influence the outcome of the test. Note that this model prior probability reflects how likely you are to believe each model to be true and is different from the parameter prior distribution that characterizes each model (Box 1). Next, we see how likely each model is after having seen our data, $P(M|\text{data})$. This shows that the full model with the interaction (condition + croup + condition × group) is by far the most likely ($P(M|\text{data}) = 0.983$). The following columns indicate the relative likelihood of each model compared with the average of all other models (BF_M) or compared with the best or worst model (BF_{10}). For instance, BF_M for the null model is $P(M|\text{data})$ for the null model divided by the average of the $P(M|\text{data})$ over all other models. For BF_{10}, the calculation depends on what is chosen in the menu 'Order'. Selecting 'Compare to null model', as we did in Fig. 5, shows the models with the null model on top, and all other BF_{10} values can be read as describing how much more likely that model is than the null model. If one selects "Compare to best model", the best model is shown first, and all other BF_{10} values express likelihood relative to that best model. Switching to BF_{01} then inverts the BF and expresses how much better the best model is than each of the other models. The error column estimates the margin of error in the BF computation.



▼ Bayesian Repeated Measures ANOVA

ShockObs
CS

Repeated Measures Factors
Condition
ShockObs
CS
Level 3
RM Factor 2

Repeated Measures Cells

ShockObs	ShockObs
CS	CS

Between Subject Factors: Group

Covariates

Bayes Factor
● BF_{10}
○ BF_{01}
○ $Log(BF_{10})$

Tables
☑ Effects
○ Across all models
● Across matched models

Bayesian Repeated Measures ANOVA

Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Condition + Group + Condition ∗ Group	0.200	0.984	239.504	1.000	
Condition + Group	0.200	0.010	0.041	0.010	4.135
Condition	0.200	0.006	0.023	0.006	4.012
Group	0.200	2.550e −4	0.001	2.593e −4	3.852
Null model (incl. subject)	0.200	1.713e −4	6.854e −4	1.742e −4	3.745

Note. All models include subject

Analysis of Effects

Effects	P(incl)	P(incl|data)	BF_{incl}
Condition	0.400	0.016	37.530
Group	0.400	0.010	1.754
Group ∗ Condition	0.200	0.984	96.358

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt.

Descriptives

					95% Credible Interval	
Condition	Group	Mean	SD	N	Lower	Upper
CS	Muscimol	68.852	17.680	20.000	60.577	77.126
	Saline	65.776	19.745	20.000	56.535	75.016
ShockObs	Muscimol	39.592	20.137	20.000	30.168	49.017
	Saline	65.601	21.370	20.000	55.599	75.602

Fig. 5 | Screenshot of the Bayesian repeated measures ANOVA of muscimol1. Data can be found at https://osf.io/md9kp/, including a muscimol1.jasp file that can be loaded to replicate the analysis within JASP or to view the results of the analysis within OSF.

The analysis showed that amongst the tested models, the full model is the most likely in the light of our data, but which of its components improved its predictive performance? To explore this question systematically, select the 'Effects' option, which generates the 'Analysis of Effects' table (Fig. 5). This analysis uses the $P(M|\text{data})$ column of the model comparison above to quantify the contribution of each component. When selecting the default option 'across all models', for each component, the $\mathrm{BF}_{\text{incl}}$ (last column) is calculated as p(models with that factor | data) ÷ p(models without that factor | data). For condition for instance, $\mathrm{BF}_{\text{incl}}$ is the average $P(M|\text{data})$ for all models with condition (i.e., condition, condition + group, and condition + group + condition × group) divided by that of all models without condition (i.e., null model and group). Selecting 'across matched models' restricts the comparison to models that only differ in the presence or absence of a particular component, and for condition, $\mathrm{BF}_{\text{incl}}$ is then the average $P(M|\text{data})$ for condition and condition + group divided by the average $P(M|\text{data})$ of their matched models, i.e., models identical except for the absence of condition, namely the null and group models. In this calculation, the interaction model is not included in the nominator, because it lacks a

Box 2 | Six advantages of a Bayesian analysis for pragmatic neuroscientists

Pragmatic neuroscientists may be convinced to start conducting Bayesian analyses—and Bayes factor hypothesis tests in particular—only when the practical advantages of doing so are sufficiently evident. Below is a select overview of such practical advantages:

1. **Bayesian hypothesis testing enables researchers to discriminate evidence of absence from absence of evidence**.
Non-significant P values are notoriously ambiguous. Indeed, a P value of 0.25 may indicate that the experiment was underpowered ('absence of evidence') or that the data support the null hypothesis ('evidence of absence').

2. **Bayesian results are relatively straightforward to interpret and communicate**.
Compared to frequentist conclusions, Bayesian conclusions are remarkably intuitive. While $P < 0.01$ is not 5 times as convincing as $P < 0.05$, $BF_{10} = 6$ really does mean twice the evidence compared to $BF_{10} = 3$. When neuroscientists make positive claims (for example, that the ACC is necessary for vicarious freezing), reviewers and readers may find it convincing if these claims are accompanied by an assessment of the statistical evidence, that is, an assessment of the extent to which H_1 outpredicted H_0.

3. **Bayes factor hypothesis testing encourages researchers to quantify evidence on a continuous scale**.
The advantage of retaining a continuous representation of evidence was stressed by Rozeboom[33]: "The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true."

4. **For most statistical scenarios, Bayes factor hypothesis testing is now relatively easy**.
Until recently, carrying out a Bayesian analysis for a standard statistical test required mathematical expertise and knowledge of probabilistic programming. This alone would be enough to deter many pragmatic neuroscientists who just wish to conduct a quick Bayesian t-test. However, recent R packages[14], Shiny apps[34] and graphical user interface (GUI)-based software packages such as JASP[35] now provide comprehensive Bayesian analyses that can be conducted with a minimum of effort.

5. **Bayesian inference allows researchers to monitor the results as the data accumulate**.
As illustrated in Box 1 and Supplementary Fig. 1, the Bayesian predict–update cycle of learning continues indefinitely. In an experimental setting, neuroscientists may decide to terminate data collection when the result is deemed compelling or when they have run out of time, money or patience[8,36]. This means that experiments can be flexibly shortened or lengthened according to the evidence that has already been collected. If error control guarantees are put in place, such flexibility can reduce the required sample size by as much as 50%[34,37].

6. **Bayes factor hypothesis testing allows researchers to include prior knowledge for a more diagnostic test**.
Although the default prior parameter distributions allow for a robust reference analysis[38], these distributions can be adjusted in light of relevant background information. This background information acts to sharpen the predictions from the models, making them easier to discriminate. For instance, prior distributions for effect size may respect the direction of the prediction, or even its location[39].

matched model group + condition × group. We recommend the 'matched model option' as it provides a more conservative estimate of each factor's contribution.

This effects table then allows us to draw inferences about the contribution of each factor and interaction in the spirit of a traditional ANOVA. BF_{incl} for condition (similarly to the main effect of condition) is 37.5, indicating that the models including the factor conditions are much (37.5 times) more likely than those not including it. The BF_{incl} for group (main effect of group) is 1.7, showing that models with group are marginally more likely than those without that main effect, but the evidence is too weak to be conclusive. BF_{incl} for the interaction is 96, meaning that the full model with the interaction is 96 times more likely than that without. This effect of interaction provides extremely strong evidence that deactivating the ACC has a much stronger impact on ShockObs than on the CS condition. However, performing the same analysis on muscimol2, where evidence that muscimol reduced freezing in the CS condition was inconclusive ($BF_{+0} = 0.97$), provides no evidence for an interaction ($BF_{incl} = 1.16$, i.e., absence of evidence). Thus, in muscimol2, we remain uncertain whether deactivating the ACC impairs freezing in the CS condition (because the t-test BF_{+0} is inconclusive) and whether deactivating the ACC has a stronger effect on ShockObs than CS. Had we found a $BF_{incl} < 1/3$, we would have had evidence of absence: that muscimol has the same effect on ShockObs and CS.

Default priors provide an objective anchor

As shown in Fig. 2, to calculate a Bayes factor we have to specify H_1 such that its predictive adequacy can be assessed. We are generally uncertain about the true value of the parameters (such as effect size), and most neuroscientists would be reticent to pin down their expectations to a single value. In the Bayesian framework, this uncertainty is reflected in the use of a prior distribution across the parameter values instead of a single value. Defining this prior distribution introduces an element of subjectivity, one that scientists fear jeopardizes the objectivity and generalizability of their inferences (for example, ref. [23], but see ref. [24]). There is however a simple two-step solution: first, use a default prior that is designed to fulfil general statistical desiderata[25]; then, check how robust your inference is against motivated changes in the prior.

For the t-test and ANOVA, there is broad consensus on certain parameter priors being appropriate under most circumstances. We recommend using these default parameter priors to increase the objectivity of the analyses and to provide a common frame of reference that ensures the direct comparability of Bayes factors from different experiments. Indeed, these defaults are implemented in JASP (and in the BayesFactor package in R for those that prefer a command line environment). Above, we performed all our inferences without considering prior distributions. However, it is informative to consider these parameter priors in more detail.

For the t-test, the default prior is the Cauchy distribution with a scale parameter of $r = \sqrt{2}/2 \approx 0.707$ as shown in Figs. 2 and 4. A Cauchy distribution resembles a Gaussian distribution but has fatter tails. The prior specifies the a priori plausibility of each effect size, and the default specifies that half the effect sizes are within the scale parameter, i.e., ±0.707, with smaller effect sizes more likely than larger effect sizes. For ANOVA, the parameters are also assumed to follow a Cauchy prior distribution, but their scale depends on the type of factor one explores (fixed effects $r = 0.5$, random effects $r = 1$, and covariates $r = 0.354$; see ref. [20] for details).

To examine the effect of changing the width of that prior distribution in our t-test example, it suffices to select the option 'Bayes factor robustness check' to generate the plots of Fig. 3a. The default width of the prior distribution for t-tests is the above mentioned Cauchy with scale 0.707 (ref. [19]); the prior that is used can be displayed (and changed) by pulling down the 'Prior' option on the bottom-left (Fig. 4). The robustness graph on the top of Fig. 3a shows how BF_{+0} changes as a function of the prior scale or width, with the scale set in the menu 'Prior' shown as the 'user prior' at the gray circle. Wider priors (wide, black circle; ultrawide, empty circles), assume that larger effects are more likely than the default prior. We consider wider priors to be less informed because if one has no expectation about effect size, all effect sizes should be considered equally likely a priori, and the prior would be infinitely wide. For ShockObs (Fig. 3a, left), evidence for H_+ is extremely high for all but the narrowest prior distributions, and our conclusion that deactivating the ACC reduces freezing is thus robust against reasonable changes in the prior. For CS (Fig. 3a, right), evidence favors H_0, also robustly across all but the narrowest prior distributions. In both cases, such robustness is reassuring and warrants confident conclusions. In contrast, when conclusions vary dramatically across a range of reasonable prior distributions, caution may be in order. Note that when the scale parameter is zero, H_+ reduces to H_0, and the Bayes factor equals 1 regardless of the data; this explains why all robustness lines will converge to 1 for the narrowest prior distributions.

Selecting the option 'Prior and posterior and additional info' outputs the results shown in Fig. 3b for our one-tailed hypothesis. Under H_+, the prior and posterior distributions are shown as dotted and black lines, respectively. This posterior shows the effect size distribution after updating the prior based on the data (Box 1 and Box 2). The posterior median and credible interval summarize the Bayesian estimate of the effect if H_+ holds (median $\delta = 1.109$, 95% credibility interval: [0.406, 1.810]). This effect size estimate is not simply the Cohen's d observed in the sample (which equals 1.24) but a combination of prior distribution and data (Box 1). The Cauchy prior distribution assumes that small effect sizes are more likely than large effect sizes; this knowledge exerts a small pull toward zero on the sample estimates—a reasonable and conservative approach—leading to the Bayesian point estimate of $\delta = 1.1$ (using the median and assuming H_+ is true). For small sample sizes, the estimate will be more influenced by the prior, whereas for larger sample sizes, the estimate will approach the sample value d. This property is desirable in the way it counteracts the systematic overestimation of effect sizes in frequentist approaches with low power[26]. For CS (right), the posterior is folded at zero because of our one-tailed hypothesis, which implies that negative effect sizes are impossible. For parameter estimation of d, we recommend adopting a two-tailed hypothesis by clicking on 'Group1≠Group2'; this leads to estimates that are more suitable to report as effect size estimates (second row). Note that for the muscimol1 column, the posterior distribution for effect size is mostly unaffected by whether a two-sided or a one-sided prior distribution is used; in contrast, the Bayes factor against the null hypothesis is about twice as high for the one-sided analysis as for the two-sided analysis (i.e., $\mathrm{BF}_{+0} = 162$ and $\mathrm{BF}_{10} = 81$).

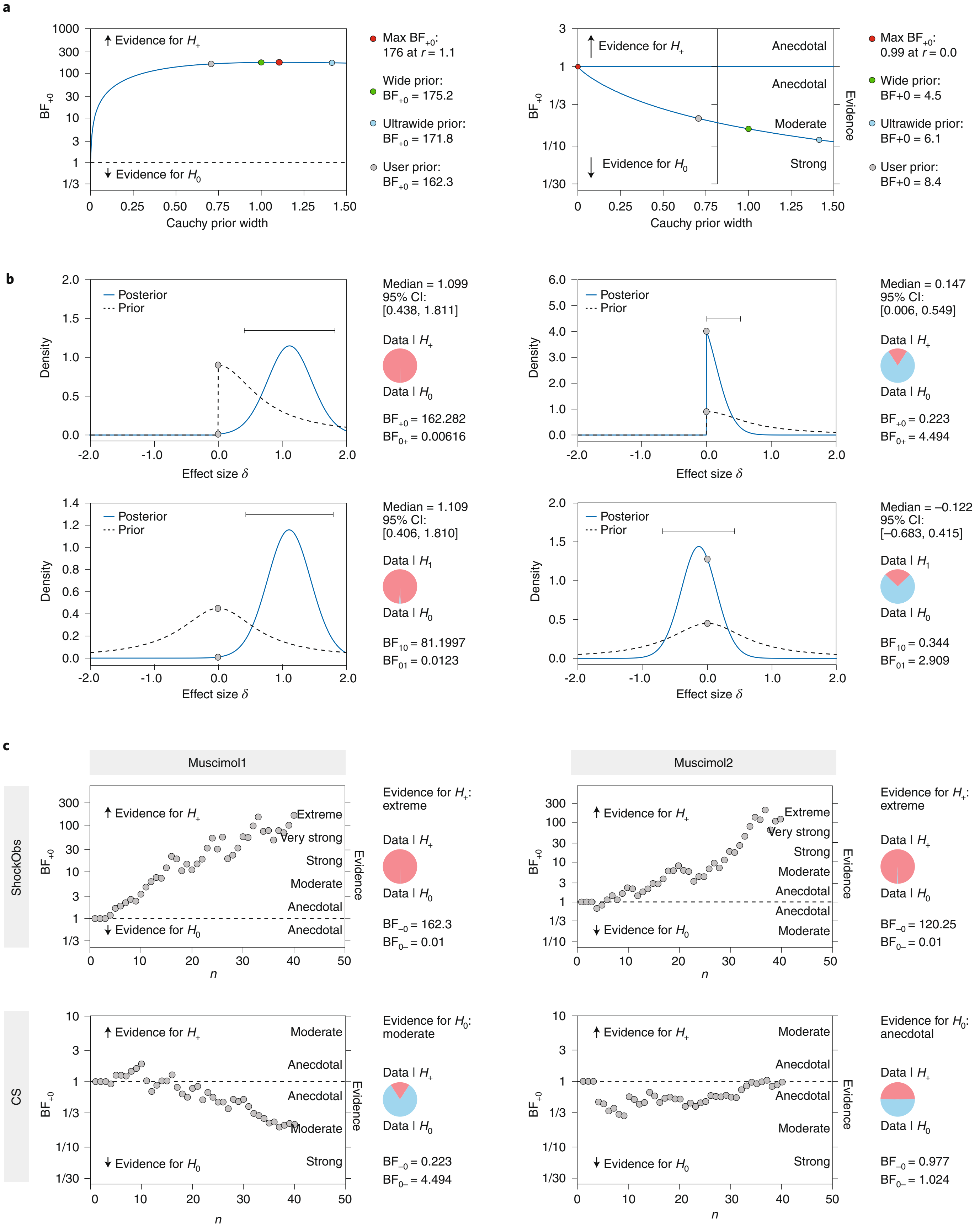
We recommend reporting the median and 95% credible interval (abbreviated as 95% CI; although this Bayesian CI is often numerically close to the frequentist confidence interval, the intervals are conceptually different; see ref. [27]) in addition to the BF to provide complementary information. For instance, for ShockObs, the BF_{+0} reveals strong evidence for the presence of an effect, but it does not indicate the strength of the effect. This is because the same effect size δ will lead to different BF values at different sample sizes (Extended Data Fig. 1b). The 95% CI provides us with information about this effect size, namely that the effect for ShockObs is probably very large (as suggested by the median $\delta = 1.1$) and that we can be quite confident that it exceeds $\delta = 0.4$ (lower bound of the 95% CI). If one looks for effects of clinical relevance, knowing that a manipulation has an effect in a group of 1,000 patients (as revealed by the BF) is often less interesting than knowing how strong the effect is likely to be (as revealed by the CI). A 95% CI that does not include $\delta = 0$ is a further indication for the presence of an effect. For CS, the BF_{+0} provides evidence for the absence of an effect. In such cases, it is perhaps not relevant to consider the 95% CI, because the CI only makes sense under H_1. However, the bounds of the CI specify that even if H_+ were true (despite the observed data being 4 times more likely under H_0), the effect size is unlikely to exceed 0.4 (upper bound of the CI), and is likely to be very small (median = −0.12). This informs the kind of group size that would be needed to systematically study such an effect. A 95% CI that includes $\delta = 0$ is in line with the notion that the data reflect the absence of an effect; however, unlike the BF, the CI alone cannot distinguish absence of evidence from evidence for absence. If scientists prefer to see the CI in the original units of measurement (for example, number of days of illness saved by a medication) the bounds should be multiplied by the population s.d., σ.

For the ANOVA, extracting credible intervals of effect sizes in JASP is a work in progress[28]. In the meantime, post hoc Bayesian t-tests could be performed to obtain Bayesian CI for specific contrasts of interest, or the effect size (for example, η^2) of the corresponding frequentist ANOVA could be reported.

The effect of the directionality of H_1 on the BF and posterior distribution is important. In frequentist statistics, one-tailed hypothesis testing is sometimes frowned upon; if one focuses on the risk of false positives, a more-conservative two-tailed statistics is arguably preferable, and the only difference is typically that P values double. With Bayesian statistics, the focus shifts to giving H_1 and H_0 a more balanced 'chance', and the ability to provide evidence for H_0 becomes an important consideration. In that context, if we hypothesize a specific direction of effect (for example, that injecting muscimol into the ACC should reduce freezing in response to ShockObs but not CS), we strongly recommend testing this directional hypothesis with the appropriate directional H_+ effect size prior distribution. The reason is particularly apparent in small group sizes: with $n = 8$, under a two-tailed Bayesian one-sample t-test, $t > 2.8$ (corresponding to $\delta \sim 0.8$) can provide evidence for H_1 ($\mathrm{BF}_{10} > 3$), but even $t = 0$

Fig. 6 | Further outputs for the Bayesian *t*-test on muscimol1.csv. a, Clicking the option 'Bayes Factor Robustness Check' will plot for each variable (ShockObs on the left and CS on the right) the BF as a function of the effect size prior. The user prior (gray) is by default set at Cauchy scale 0.707 as recommended in ref. [19]. The wide and ultrawide prior are flatter priors that are sometimes used, especially when the goal is parameter estimation. As can be seen, there is extreme evidence for H_1 in ShockObs, across all but the smallest priors (i.e., the gray, black and white dots all have $\mathrm{BF}_{+0} > 160$), and there is moderate evidence for H_0 for all but the smallest priors for CS (most $\mathrm{BF}_{0+} > 4.5$). The interpretation of the data does thus not depend on the choice of prior scale within a reasonable range. **b**, Priors and posteriors for ShockObs and CS together with median and CI of the effect size. Results are shown for a one-tailed prior (top row) often more suited for hypothesis testing and two-tailed prior (bottom row) more suited for parameter estimation. **c**, Accumulation of evidence with increasing sample size using the 'Sequential analysis' option. Data can be found at https://osf.io/md9kp/, including a muscimol1.jasp file that can be loaded to replicate the analysis within JASP or to view the results of the analysis within OSF.

(the datum with the highest evidence for H_0) falls short of providing modest evidence for H_0 ($BF_{01} = 2.97$). Using the theoretically appropriate H_+ resolves this imbalance, as even small negative t-values can provide evidence for H_0 over H_+ (for example, $t = -0.3$, $BF_{0+} = 3.62$). One-tailed testing is thus typically a fairer balance between the ability to provide evidence for H_0 and H_1.



Finally, it is important to consider that some scenarios do call for user-defined priors (see ref. [24] for a more extensive discussion of how to create informed priors). For instance, to test a claim that a candidate drug has an effect size $\delta > 0.8$ one would need to specify custom priors with H_0: $\delta < 0.8$ vs H_1: $\delta > 0.8$ and compare their likelihoods.

Accumulation of evidence

While designing experiments, we are typically uncertain about the effect sizes to expect. Determining the number of subjects or participants we need to provide sufficient power a priori is then difficult. By selecting the option 'Sequential Analysis' we can see how the BF changes as one considers an increasing number of data points in our Bayesian t-test examples (Figs. 3c and 6). For muscimol1, we observe a clear upward trend to ShockObs in favor of H_+ and a downward trend to CS playback in favor of H_0. Such consistent trends provide confidence in the effect a posteriori. Importantly, this analysis can be performed during data collection, effectively replacing a predefined sample size by a principled data collection plan: for example, collect a minimum of $n = 20$ animals (10 per group) at first, and then keep adding new animals to the saline and muscimol group until the BF_{+0} crosses a predetermined critical value (for example, $\mathrm{BF}_{+0} > 6$ or $\mathrm{BF}_{+0} < 1/6$) or until a preset maximum of animals (for example, $n = 40$) has been reached (Supplementary Notes). In our example, we would have stopped at $n = 20$ animals in the ShockObs condition and continued until $n = 40$ animals in the CS condition, thus saving $n = 20$ animals to reach the same conclusions. Such an approach is unacceptable in NHST (Supplementary Note and Supplementary Figure 1). This is because Bayesian statistics can provide evidence for H_0 and H_1, whereas NHST can only provide evidence against H_0. Hence, testing until a significant result is found in NHST will per definition always find evidence against H_0.

For muscimol2, the BF_{+0} values show no steep and consistent trend toward providing evidence in favor of either hypothesis (Fig. 3c, bottom right). This is typical of small effect sizes. For $n > 20$, the BF shows a mild upwards trend, and extending this trend shows that hundreds of animals would probably have to be added for the analysis to provide evidence for the presence of an effect ($\mathrm{BF}_{+0} > 3$). This $n > 100$ projection is in line with the outcome of a traditional power analysis for $\delta = 0.4$, which is the effect size we used to generate the simulated data in muscimol2.

Reporting both frequentist and Bayesian results

One concern for aspiring Bayesian neuroscientists is that reviewers in neuroscience journals may be unfamiliar with Bayes factors and may be more impressed by $P < 0.01$ than by $\mathrm{BF}_{10} = 10.3$. Our pragmatic recommendation is to consistently report both the frequentist and Bayesian statistics (for example, $t_{(38)} = 3.961$, $P < 0.001$, $\mathrm{BF}_{+0} = 162$, with median posterior $\delta = 1.1$, 95% CI = [0.4, 1.8]). Where evidence for H_1 is presented, one can report a P value with a standard frequentist test and add the BF_{10} to provide additional quantification. Where there is no evidence for H_1, reporting BF_{01} is an attractive way to adjudicate between absence of evidence and evidence of absence.

This hybrid approach is a powerful opportunity to reap the best of both statistical approaches. In borderline cases where frequentist and Bayesian approaches do not quite concur (for example, $P < 0.04$ suggesting a significant effect, but $\mathrm{BF}_{10} = 2.3$ suggesting only anecdotal evidence), we still recommend reporting both and discussing the divergence as showing that obtaining more data will be important to strengthen the evidence. Additionally, reporting the CI on the effect size is important. Extended Data Fig. 3 provides examples of wording appropriate to report the kind of analyses we discussed above.

Concluding comments

Bayesian inference offers unique practical advantages for neuroscience (Box 2). Bayes factors provide a continuous and symmetric measure of statistical evidence. The Bayes factor can support H_0 as much as it can support H_1. There is a bias toward publishing significant results, and we have become increasingly aware of the negative impact that the resulting P value hacking has on the progress and replicability of science. Bayesian statistics provide a principled tool for reducing this bias by allowing us to provide equally compelling evidence for the absence and the presence of an effect.

We have presented examples of neuroscience scenarios in which Bayesian statistics are simple to adopt. Some applications will require more development. For example, neuroimaging requires statistical testing over thousands of voxels and, therefore, correction for multiple comparisons, and frameworks for the latter are still in their infancy for the Bayes factor. Also, the Bayesian t-test and ANOVA we leveraged here assume normally distributed data, but neuroscience datasets can have highly non-normal distributions. Non-parametric Bayesian tests so far only exist for certain applications (for example, some t-tests and regressions have a tick-mark for non-parametric approaches in JASP, and R code exists for a number of additional cases[29]), but remain in development for others (for example, ANOVA).

Neuroscientists have been slow to take up Bayesian statistics, presumably out of a perception that Bayesian hypothesis testing is difficult to perform and interpret. With the emergence of new software and accessible packages, performing Bayesian equivalents of the most prevalent tests has become easy. Supplementing frequentist approaches with Bayesian analyses will lead to richer data interpretations that allow more informative conclusions. Null findings become interpretable and more easily publishable. We finally have a principled tool to shed light on the hitherto dark side of our scientific enterprise: evidence of absence.

Data availability

All data and code can be downloaded at https://osf.io/md9kp/.

Received: 11 December 2019; Accepted: 21 May 2020;
Published online: 29 June 2020

References

1. Benjamin, D. J. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
2. Dienes, Z. Using Bayes to get the most out of non-significant results. *Front. Psychol.* **5**, 781 (2014).
3. Gallistel, C. R. The importance of proving the null. *Psychol. Rev.* **116**, 439–453 (2009).
4. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).
5. Love, J. et al. JASP: Graphical statistical software for common statistical designs. *J. Stat. Softw.* **88**, 1–17 (2019).
6. Wagenmakers, E.-J. *et al.* The need for Bayesian hypothesis testing in psychological science. in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions* (eds. Lilienfeld, S. O. & Waldman, I.) 123–138 (Wiley, 2017).
7. Altman, D. G. & Bland, J. M. Absence of evidence is not evidence of absence. *Br. Med. J.* **311**, 485 (1995).
8. Edwards, W., Lindman, H. & Savage, L. J. Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242 (1963).
9. Jeffreys, H. *Theory of Probability* (Oxford University Press, 1961).
10. Szucs, D. & Ioannidis, J. P. A. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15**, e2000797 (2017).
11. Etz, A. & Wagenmakers, E.-J. J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Stat. Sci.* **32**, 313–329 (2017).
12. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
13. Lee, M. D. & Wagenmakers, E.-J. *Bayesian Cognitive Modeling: A Practical Course* (Cambridge University Press, 2013).

14. Morey, R. D. & Rouder, J. N. BayesFactor: computation of Bayes factors for common designs. v. 0.9.12–4.2 https://cran.r-project.org/package=BayesFactor (2018).
15. Carrillo, M. et al. Emotional mirror neurons in the rat's anterior cingulate cortex. *Curr. Biol.* **29**, 1301–1312.e6 (2019).
16. Jeffreys, H. *Theory of Probability* (Oxford University Press, 1939).
17. Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E.-J. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* **14**, 1105–1107 (2011).
18. Gelman, A. & Stern, H. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006).
19. Morey, R. D. & Rouder, J. N. Bayes factor approaches for testing interval null hypotheses. *Psychol. Methods* **16**, 406–419 (2011).
20. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56**, 356–374 (2012).
21. Rouder, J. N., Engelhardt, C. R., McCabe, S. & Morey, R. D. Model comparison in ANOVA. *Psychon. Bull. Rev.* **23**, 1779–1786 (2016).
22. Myung, I. J. & Pitt, M. A. Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon. Bull. Rev.* **4**, 79–95 (1997).
23. Efron, B. Why isn't everyone a Bayesian? *Am. Stat.* **40**, 1–5 (1986).
24. Lee, M. D. & Vanpaemel, W. Determining informative priors for cognitive models. *Psychon. Bull. Rev.* **25**, 114–127 (2018).
25. Bayarri, M. J., Berger, J. O., Forte, A. & Garcia-Donato, G. Criteria for Bayesian model choice with application to variable selection. *Ann. Stat.* **40**, 1550–1577 (2012).
26. Cremers, H. R., Wager, T. D. & Yarkoni, T. The relation between statistical power and inference in fMRI. *PLoS ONE* **12**, e0184923 (2017).
27. Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D. & Wagenmakers, E.-J. The fallacy of placing confidence in confidence intervals. *Psychon. Bull. Rev.* **23**, 103–123 (2016).
28. Marsman, M., Waldorp, L., Dablander, F. & Wagenmakers, E. J. Bayesian estimation of explained variance in ANOVA designs. *Stat. Neerl.* **73**, 351–372 (2019).
29. van Doorn, J., Marsman, M., Ly, A. & Wagenmakers, E.-J. Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *J. Appl. Stat.* https://doi.org/10.1080/02664763.2019.1709053 (2020).
30. Wagenmakers, E.-J., Morey, R. D. & Lee, M. D. Bayesian benefits for the pragmatic researcher. *Curr. Dir. Psychol. Sci.* **25**, 169–176 (2016).
31. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction.* (The MIT Press, 1998).
32. Wrinch, D. & Jeffreys, H. On certain fundamental principles of scientific inquiry. *Philos. Mag.* **42**, 368–374 (1923).
33. Rozeboom, W. W. The fallacy of the null-hypothesis significance test. *Psychol. Bull.* **57**, 416–428 (1960).
34. Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D. & Wagenmakers, E.-J. A tutorial on Bayes factor design analysis using an informed prior. *Behav. Res. Methods* **51**, 1042–1058 (2019).
35. Wagenmakers, E.-J. et al. Bayesian inference for psychology. Part II: example applications with JASP. *Psychon. Bull. Rev.* **25**, 58–76 (2018).
36. Rouder, J. N. Optional stopping: no problem for Bayesians. *Psychon. Bull. Rev.* **21**, 301–308 (2014).
37. Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: planning for compelling evidence. *Psychon. Bull. Rev.* **25**, 128–142 (2018).
38. Consonni, G., Fouskakis, D., Liseo, B. & Ntzoufras, I. Prior distributions for objective Bayesian analysis. *Bayesian Anal.* **13**, 627–679 (2018).
39. Gronau, Q. F., Ly, A. & Wagenmakers, E.-J. Informed Bayesian *t*-tests. *Am. Stat.* **74**, 137–143 (2019).

Acknowledgements

C.K. is funded by NWO VICI grant 453-15-009; V.G. is funded by ERC grant 758703 and NWO VIDI grant 452-14-015; E.J.W. is funded by NWO VICI grant 453-16-003. We thank F. Bartos for help with Fig. 2.

Author contributions

All authors conceived the project together and contributed to the writing of the manuscript. E.J.W. coordinates the development of JASP.

Competing interests

E.J.W. declares that he coordinates the development of the open-source software package JASP (https://jasp-stats.org), a non-commercial, publicly-funded effort to make Bayesian statistics accessible to a broader group of researchers and students. C.K. and V.G. declare no competing interests.

Additional information

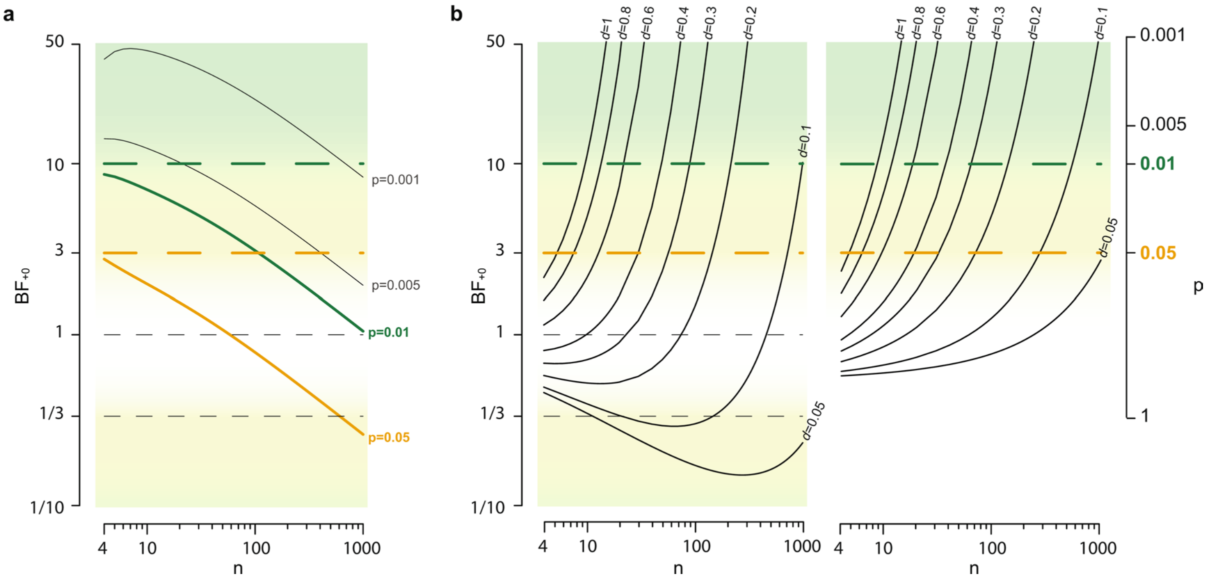
Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-0660-4.

Correspondence should be addressed to C.K.

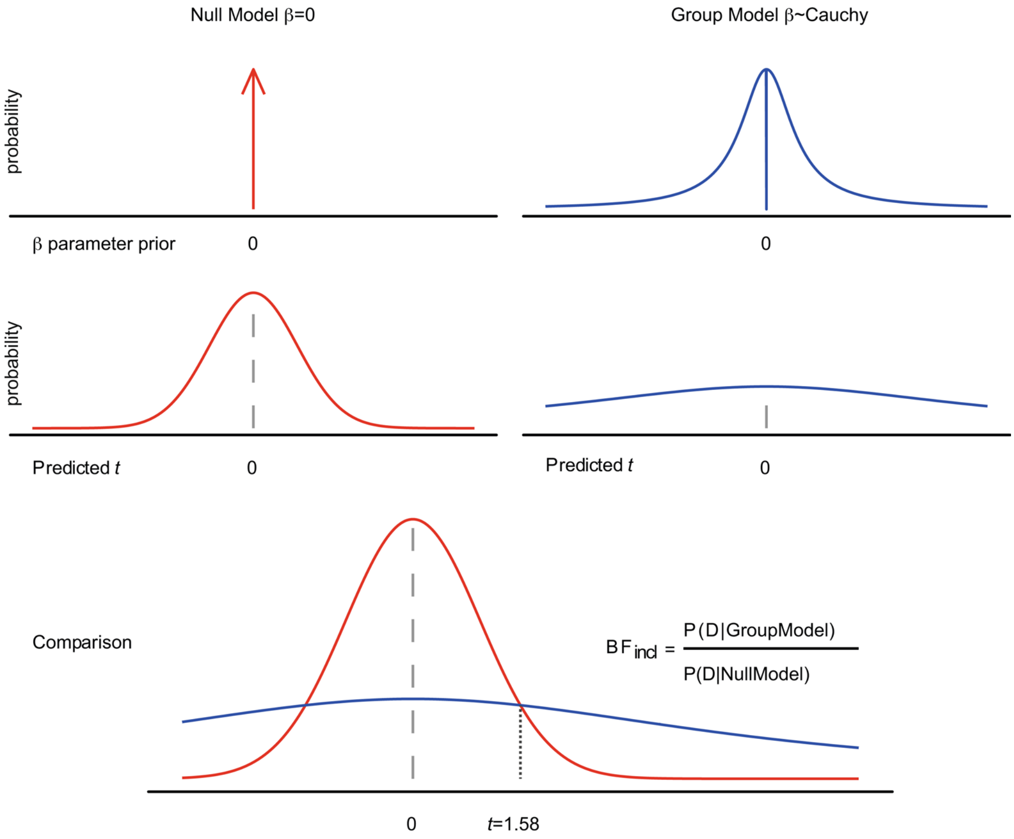
Peer review information *Nature Neuroscience* thanks Denise Cai, Zhe Dong, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

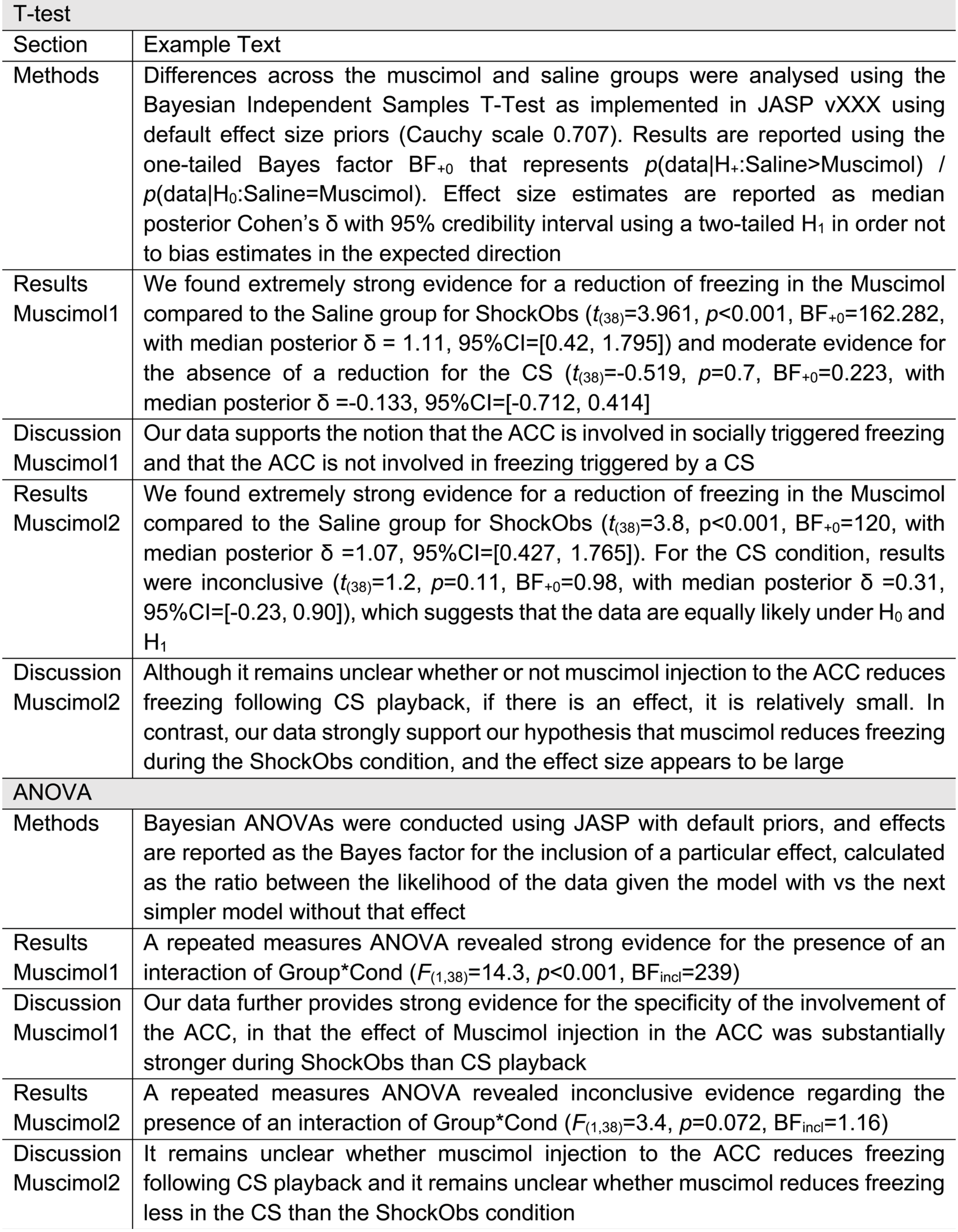
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | The relationship between BF, p, and effect sizes values. a, This log-log plot shows the BF_{+0} values corresponding to familiar critical *p* values for a one-tailed one-sample *t*-test at different sample sizes (*n*). The curves show the BF_{+0} values obtained in a Bayesian *t*-test based on the critical *t*-value that provides *P*=0.05 (yellow), *P*=0.01 (green), *P*=0.005 (black) and *P*=0.001 (black). The yellow dashed horizontal line indicates the BF_{+0}=3 bound for moderate evidence considered by Jeffreys[9] to be similar to *P*=0.05, the green one the BF_{+0}=10 for strong evidence considered similar to *P*=0.01. The two black dashed lines mark BF_{+0}=1, i.e. the line of no evidence, and BF_{+0}=1/3, the bound for moderate evidence of absence. The background gradient reminds the reader that the BF reference values of 3 and 10 should not be considered hard bounds. Instead the BF should be interpreted as a continuous value, with values diverging more from 1 supporting stronger conclusions. This panel makes two points. First, there is no simple equivalence between *p* and BF that holds over all sample sizes. This is because in a frequentist *t*-test, the observed effect size (*d*) sufficient to generate a specific *p* value decreases with $\sqrt{n}$ more rapidly than for the BF. As a result, at large *n*, very small effect sizes generate 'significant' *t*-test: at *n*=1000, the critical *t*-value for a one-tailed *P*=0.05 is 1.65, corresponding to $d=1.65/\sqrt{n}=0.05$. For the BF, such a minuscule effect is 4 times more likely under H_0 than H_+ (BF_{+0}=0.26). Hence, for small sample sizes *p* and BF support similar conclusions (e.g., *P*=0.05 at *n*=4 corresponds to BF_{+0}>3, supporting the same conclusion of evidence for an effect), but for large sample sizes the frequentist and Bayesian conclusions can diverge in the presence of very small effect sizes (e.g., *P*=0.05 at *n*=1000 corresponds to BF_{+0}<1/3, see Jeffreys, H. Some Tests of Significance, Treated by the Theory of Probability. *Proc. Cambridge Philos. Soc.* **31**, 203–222 (1935)). Considering confidence or credible intervals of the effect size in addition to *p* or BF values helps interpret such cases. Second, the fact that the dashed lines are above the curve of the same color for all n>4 shows that BF_{+0}=3 and BF_{+0}=10 indeed protect against Type I errors in a frequentist sense at least at *P*=0.05 or *P*=0.01, respectively. In other words, if BF_{10}>3, p<0.05, and if BF_{10}>10, *p*<0.01, but how much lower than 0.05 or 0.01 the exact *P value* is, depends on *n*. **b**, BF_{+0} (left) and *p* (right) values as a function of measured effect- and sample-sizes. These panels illustrate the measured effect sizes necessary to provide evidence for an effect at different sample sizes in a one-sample one-tailed t-test using the BF vs. traditional p values. Each curve connects the results at different sample sizes for the specified value of *d*. The logarithmic BF and p scales are aligned so as to place BF=3 next to *P*=0.05, and BF=10 next to *P*=0.01.

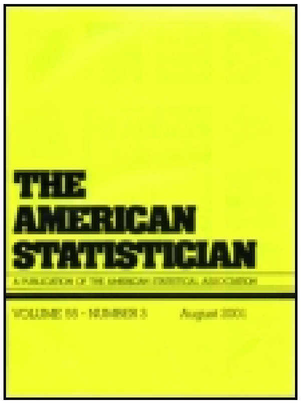


Extended Data Fig. 2 | Evidence for or against a factor in a Bayesian ANOVA. A Bayesian ANOVA is a form of model comparison. This figure illustrates how the Bayes factor can provide evidence for a simpler model by concentrating its predictions on a single parameter value. This example ANOVA determines whether or not the data D depend on the value of the factor Group by comparing the Null Model D=0*Group (left) against the Group Model D=β*Group, with a Cauchy prior on β (right). The top row illustrates the prior probability attributed to the different values of β under the two competing models. Note how both models include β = 0 as a possibility, but given that the probability values must integrate to 1 over the entire β space, for the Null Model *p*(β = 0) = 1 while for the Group Model, the probability is distributed across all plausible alternative values. The middle row shows the predicted t-values based on these priors, where t represents the difference between the data from the two groups as in Fig. 2. Note how these predictions are more peaked for the Null compared to the Group model. The bottom row compares the predicted probability of finding particular *t*-values under the two models, and shows how values close to zero (i.e., small or no difference between the groups) are predicted more often by the Null compared to the Group Model, while the opposite is true for large *t*-values. If conducting the experiment reveals a measured *t*-values close to zero, the Bayes Factor for including the factor Group would be substantially below 1, providing evidence for the absence of an effect of Group, while the inverse would be true for high *t*-values.



T-test	
Section	Example Text
Methods	Differences across the muscimol and saline groups were analysed using the Bayesian Independent Samples T-Test as implemented in JASP vXXX using default effect size priors (Cauchy scale 0.707). Results are reported using the one-tailed Bayes factor BF_{+0} that represents p(data|H_+:Saline>Muscimol) / p(data|H_0:Saline=Muscimol). Effect size estimates are reported as median posterior Cohen's δ with 95% credibility interval using a two-tailed H_1 in order not to bias estimates in the expected direction
Results Muscimol1	We found extremely strong evidence for a reduction of freezing in the Muscimol compared to the Saline group for ShockObs ($t_{(38)}$=3.961, p<0.001, BF_{+0}=162.282, with median posterior δ = 1.11, 95%CI=[0.42, 1.795]) and moderate evidence for the absence of a reduction for the CS ($t_{(38)}$=-0.519, p=0.7, BF_{+0}=0.223, with median posterior δ =-0.133, 95%CI=[-0.712, 0.414]
Discussion Muscimol1	Our data supports the notion that the ACC is involved in socially triggered freezing and that the ACC is not involved in freezing triggered by a CS
Results Muscimol2	We found extremely strong evidence for a reduction of freezing in the Muscimol compared to the Saline group for ShockObs ($t_{(38)}$=3.8, p<0.001, BF_{+0}=120, with median posterior δ =1.07, 95%CI=[0.427, 1.765]). For the CS condition, results were inconclusive ($t_{(38)}$=1.2, p=0.11, BF_{+0}=0.98, with median posterior δ =0.31, 95%CI=[-0.23, 0.90]), which suggests that the data are equally likely under H_0 and H_1
Discussion Muscimol2	Although it remains unclear whether or not muscimol injection to the ACC reduces freezing following CS playback, if there is an effect, it is relatively small. In contrast, our data strongly support our hypothesis that muscimol reduces freezing during the ShockObs condition, and the effect size appears to be large
ANOVA	
Methods	Bayesian ANOVAs were conducted using JASP with default priors, and effects are reported as the Bayes factor for the inclusion of a particular effect, calculated as the ratio between the likelihood of the data given the model with vs the next simpler model without that effect
Results Muscimol1	A repeated measures ANOVA revealed strong evidence for the presence of an interaction of Group*Cond ($F_{(1,38)}$=14.3, p<0.001, BF_{incl}=239)
Discussion Muscimol1	Our data further provides strong evidence for the specificity of the involvement of the ACC, in that the effect of Muscimol injection in the ACC was substantially stronger during ShockObs than CS playback
Results Muscimol2	A repeated measures ANOVA revealed inconclusive evidence regarding the presence of an interaction of Group*Cond ($F_{(1,38)}$=3.4, p=0.072, BF_{incl}=1.16)
Discussion Muscimol2	It remains unclear whether muscimol injection to the ACC reduces freezing following CS playback and it remains unclear whether muscimol reduces freezing less in the CS than the ShockObs condition

Extended Data Fig. 3 \| Examples of how to report results.



The American Statistician

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

The Bayesian Two-Sample t Test

Mithat Gönen[a], Wesley O Johnson[a], Yonggang Lu[a] & Peter H Westfall[a]

[a] Mithat Gönen is Associate Attending Biostatistician, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10021 . Wesley O. Johnson is Professor, Department of Statistics, University of California at Irvine, Irvine, CA 92697 . Yonggang Lu is a Ph.D. student, and Peter H. Westfall is Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409 (E-mail addresses: and . The author order is alphabetical. The authors are grateful to the referees, the associate editor, and the editor for their suggestions that greatly improved the article.

Published online: 01 Jan 2012.

To cite this article: Mithat Gnen, Wesley O Johnson, Yonggang Lu & Peter H Westfall (2005) The Bayesian Two-Sample t Test, The American Statistician, 59:3, 252-257, DOI: 10.1198/000313005X55233

To link to this article: http://dx.doi.org/10.1198/000313005X55233

The Bayesian Two-Sample t Test

Mithat GÖNEN, Wesley O. JOHNSON, Yonggang LU, and Peter H. WESTFALL

This article shows how the pooled-variance two-sample t statistic arises from a Bayesian formulation of the two-sided point null testing problem, with emphasis on teaching. We identify a reasonable and useful prior giving a closed-form Bayes factor that can be written in terms of the distribution of the two-sample t statistic under the null and alternative hypotheses, respectively. This provides a Bayesian motivation for the two-sample t statistic, which has heretofore been buried as a special case of more complex linear models, or given only roughly via analytic or Monte Carlo approximations. The resulting formulation of the Bayesian test is easy to apply in practice, and also easy to teach in an introductory course that emphasizes Bayesian methods. The priors are easy to use and simple to elicit, and the posterior probabilities are easily computed using available software, in some cases using spreadsheets.

KEY WORDS: Bayes factor; Posterior probability; Prior elicitation; Teaching Bayesian statistics.

1. INTRODUCTION AND THE TEST

The two-sample comparison is a staple in elementary statistics courses. A typical course sequence is as follows: one-sample problems (means and proportions, tests and intervals), two-sample comparisons (differences of means and proportions, tests and intervals), then more advanced topics (ANOVA, regression). Single-sample problems involving the selection of a population reference value for the mean, μ_0, are less interesting than their two-sample counterparts. Most designed experiments involve this latter category, where the samples are experimental and control (drug and placebo in most clinical trials), and interesting applications also exist in virtually all areas of scientific inquiry.

Assuming the data y_{ir} $(i = 1, 2; r = 1, \ldots, n_i)$ are independent and normally distributed with means μ_i and common variance σ^2, the pooled-variance two-sample t test is commonly used for testing $H_0 : \mu_1 = \mu_2$ against the two-sided alternative $H_1 : \mu_1 \neq \mu_2$. The test statistic is

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s_p / n_\delta^{1/2}}, \tag{1}$$

where

$$s_p^2 = \left\{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right\} / (n_1 + n_2 - 2)$$

is the pooled variance estimate, $\overline{y}_i$ and s_i^2 are the sample mean and sample variance for group i, and

$$n_\delta = \left(n_1^{-1} + n_2^{-1}\right)^{-1},$$

which may be called the "effective sample size" for the two-sample experiment. Letting $\nu = n_1 + n_2 - 2$ denote the degrees of freedom and $t\{1 - \alpha/2, \nu\}$ denote the $1 - \alpha/2$ quantile of the T_ν distribution, H_0 is rejected in favor of H_1 when $|t| \geq t\{1 - \alpha/2, \nu\}$; the two-sided p value is obtained as $p = 2 \times P(T \geq |t|)$, where T has the T_ν distribution. This test has many optimality properties (Lehmann 1986), it is routinely produced by statistical software, and it is found in most elementary statistics texts.

Although the two-sample t statistic is well understood and widely accepted, it is difficult to find motivation for it in the Bayesian hypothesis testing literature. Recent literature suggesting that we should teach Bayesian methods at the elementary learning stage includes Albert (1997a), Albert and Rossman (2001), Antelman (1997), Berry (1996, 1997), and Bolstad (2004); however, none of these discuss the two-sample t statistic, at least not from the Bayesian formulation of hypothesis testing.

In the general Bayesian formulation of hypothesis testing, one places prior probabilities π_0 and π_1 $(\pi_0 + \pi_1 = 1)$ on hypotheses H_0 and H_1, respectively, then updates these values via Bayes' theorem to obtain the posterior probabilities

$$P(H_j \mid \text{data}) = \frac{\pi_j P(\text{data} \mid H_j)}{\pi_0 P(\text{data} \mid H_0) + \pi_1 P(\text{data} \mid H_1)}, \; j = 0, 1,$$

where $P(\text{data}|H_j)$ denotes the marginal density of the data under hypothesis j. Because the posterior probabilities are sensitive to the priors π_0 and π_1, it is often suggested to use the Bayes factor (BF) instead:

$$\text{BF} = \frac{P(\text{data} \mid H_0)}{P(\text{data} \mid H_1)}.$$

When BF > 1 the data provide evidence for H_0, and when BF < 1 the data provide evidence for H_1 (and against H_0). Jeffreys (1961) suggested BF $< .1$ provides "strong" evidence against H_0 and BF $< .01$ provides "decisive" evidence. The posterior probability is simply related to the Bayes factor as

$$P(H_0 \mid \text{data}) = \left[1 + \frac{\pi_1}{\pi_0} \frac{1}{\text{BF}}\right]^{-1}.$$

Much of the literature on Bayes factors and posterior probabilities is concerned with calculating or approximating (either analytically or via Monte Carlo) the marginal densities

$$P(\text{data} \mid H_j) = \int P(\text{data} \mid \boldsymbol{\theta}_j, H_j) \Pi_j(\boldsymbol{\theta}_j \mid H_j) \, d\boldsymbol{\theta}_j,$$

Mithat Gönen is Associate Attending Biostatistician, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10021 (E-mail: gonenm@mskcc.org). Wesley O. Johnson is Professor, Department of Statistics, University of California at Irvine, Irvine, CA 92697 (E-mail: wjohnson@ics.uci.edu). Yonggang Lu is a Ph.D. student, and Peter H. Westfall is Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409 (E-mail addresses: gentlelu@yahoo.com *and* peter.westfall@ttu.edu). The author order is alphabetical. The authors are grateful to the referees, the associate editor, and the editor for their suggestions that greatly improved the article.

DOI: 10.1198/000313005X55233

where $\boldsymbol{\theta}_j$ is the parameter vector under hypothesis H_j and $\Pi_j(\boldsymbol{\theta}_j \mid H_j)$ is its prior distribution. Relevant references are Jeffreys (1961), Dickey (1971), Zellner and Siow (1980), Berger and Sellke (1987), Bernardo and Smith (1994), Carlin and Chib (1995), Chib (1995), Kass and Raftery (1995), and Albert (1997b).

When considering the two-sample case in particular where the hypotheses are $H_0 : \mu_1 = \mu_2 = \mu$, versus $H_1 : \mu_1 \neq \mu_2$, the parameter vectors are $\boldsymbol{\theta}_0 = (\mu, \sigma^2)$ and $\boldsymbol{\theta}_1 = (\mu_1, \mu_2, \sigma^2)$, and one may consider a variety of priors $\Pi_j(\boldsymbol{\theta}_j \mid H_j)$. Such analyses for the Bayesian two-sample t test are found in the literature, but only implicitly as a special case of more complex regression formulations, or as related to the estimation problem as in Bolstad (2004). The aims of this article are two-fold: first we present the model and a reasonable prior for which the BF depends on the data only through the pooled-variance two-sample t statistic, as well as the associated (central and noncentral) T_ν distributions; and second, we show how one might use the results for prior selection, data analysis, and learning about Bayesian statistics.

For the two-sample problem with normally distributed, homoscedastic, and independent data, with prior distributions as specified in Section 2, the Bayes factor for testing $H_0 : \mu_1 = \mu_2 = \mu$, versus $H_1 : \mu_1 \neq \mu_2$ is

$$\mathrm{BF} = \frac{T_\nu(t \mid 0, 1)}{T_\nu(t \mid n_\delta^{1/2}\lambda, 1 + n_\delta\sigma_\delta^2)}. \tag{2}$$

Here t is the pooled-variance two-sample t statistic (1), λ and σ_δ^2 denote the prior mean and variance of the standardized effect size $(\mu_1 - \mu_2)/\sigma$ under H_1, and $T_\nu(. \mid a, b)$ denotes the noncentral t probability density function (pdf) having location a, scale $b^{1/2}$, and df ν. Specifically, $T_\nu(. \mid a, b)$ is the pdf of the random variable $Y/\sqrt{U/\nu}$, where Y is distributed normally with mean a and variance b, and where U has the chi-square distribution with ν degrees of freedom, independent of Y. The mathematical derivation of (2) and further details are available online (Gönen, Westfall, Johnson, and Lu 2004). The data enter the BF only through the pooled-variance two-sample t statistic (1), providing a Bayesian motivation for its use. Benefits of having the analytic result (2) are: (i) one can explain the Bayesian two-sample t test in terms of unconditional (central and noncentral T) distributions; (ii) it allows simple sensitivity analysis with respect to prior inputs, as we show in Section 4; and (iii) it allows for simple explanations of interesting Bayesian topics such as the noncorrespondence between posterior probabilities and p values (Berger and Sellke 1987), and "Lindley's Paradox" (Lindley 1957), both of which are also illustrated in Section 4.

Calculation of (2) requires evaluation of the noncentral T pdf with general scale parameter. Many software packages provide the pdf of the noncentral t having scale parameter 1.0, and a simple modification is needed for the general case: $T_\nu(t \mid a, b) = T_v(t/b^{1/2} \mid a/b^{1/2}, 1)/b^{1/2}$. Thus, for example, using the statistics freeware package R (http://www.r-project.org/), the Bayes factor can be computed as

```
BF = dt(t,n1+n2-2)/(dt(t/sqrt(postv),
     n1+n2-2,nc)/sqrt(postv))
```

where "t" is the value of the two-sample t statistic, `postv` $= 1 + n_\delta\sigma_\delta^2$ and `nc`$= n_\delta^{1/2}\lambda/(1 + n_\delta\sigma_\delta^2)^{1/2}$. The noncentral t density is also available in commercial packages including SAS, SPSS, and Mathematica, and it may be obtained using specialized programs or add-ins with other packages as well. For the case where the prior mean λ of the effect size is assumed to be zero, the Bayes factor requires only the central T pdf and is calculated more simply (e.g., using a spreadsheet) as

$$\mathrm{BF} = \left[\frac{1 + t^2/\nu}{1 + t^2/\{\nu(1 + n_\delta\sigma_\delta^2)\}}\right]^{-(\nu+1)/2} (1 + n_\delta\sigma_\delta^2)^{1/2}.$$

Assessment of priors is discussed generically in Section 2, and Section 3 discusses prior selection in a specific context involving clinical trials. Section 4 presents an analysis of a dataset comparing blood pressure drop in patients receiving either calcium supplements or placebo, along with a sensitivity analysis, and Section 5 concludes.

2. PRIOR DISTRIBUTION AND ASSESSMENT

Let $N(y \mid a, b)$ denote the pdf of a normally distributed random variable with mean a and variance b, and as usual, $Y \sim N(a, b)$ means that Y has pdf $N(y \mid a, b)$. The assumption for the two-sample t test is that the data are conditionally independent with $Y_{ir}|\{\mu_i, \sigma^2\} \sim N(\mu_i, \sigma^2)$. The goal is to test the null hypothesis $H_0 : \delta = \mu_1 - \mu_2 = 0$ against the two-sided alternative $H_1 : \delta \neq 0$.

To obtain the usual two-sample t statistic, prior knowledge is modeled for δ/σ rather than for δ. Let $\mu = (\mu_1 + \mu_2)/2$, and reparameterize (μ_1, μ_2, σ^2) to (μ, δ, σ^2). The prior for δ/σ is specified as

$$\delta/\sigma \mid \{\mu, \sigma^2, \delta/\sigma \neq 0\} \sim N(\lambda, \sigma_\delta^2).$$

For Jeffreys (1961), dependence of the prior for δ on the value of σ is implicit in his assertion "from conditions of similarity, it [the mean] must depend on σ, since there is nothing in the problem except σ to give a scale for [the mean]." This dependence is also found in Dickey (1971), Zellner and Siow (1980) and Berger, Boukai, and Wang (1997).

The standardized effect size δ/σ is a familiar dimensionless quantity, easily modeled a priori. Cohen (1988) reported that $|\delta/\sigma|$ values of .20, .50, and .80 are "small," "medium," and "large," respectively, based on a survey of studies reported in the social sciences literature. These benchmarks can be used to check whether the specifications of hyperparameters λ and σ_δ^2 are reasonable; a simple check based on $\lambda \pm 3\sigma_\delta$ can determine whether the prior allows unreasonably large effect sizes.

The remaining parameters (μ, σ^2) are assigned a standard noninformative prior, no matter whether $\delta = 0$ or $\delta \neq 0$. Although noninformative priors are attractive in the sense of minimizing prior inputs, they also ensure that the Bayes factor depends on the data only through the two-sample t statistic. One can verify numerically that, when the prior for (μ, σ^2) is informative, two different datasets having identical t statistics and sample sizes can yield different Bayes factors.

To summarize, the prior is as follows:

$$\Pi(\delta\,/\sigma|\,\mu, \sigma^2, \delta \neq 0) = N(\delta/\sigma \mid \lambda, \sigma_\delta^2),$$

with the nuisance parameters assigned the improper prior

$$\Pi(\mu, \sigma^2) \propto 1/\sigma^2.$$

Finally, the prior is completed by specifying the probability that H_0 is true:

$$\pi_0 = P(\delta = 0),$$

where π_0 is often taken to be 1/2 as an "objective" value (Berger and Sellke 1987). However, π_0 can be simply assigned by the experimenter to reflect prior belief in the null; it can be assigned to differentially penalize more complex models (Jeffreys 1961, p. 246); it can be assessed from multiple comparisons considerations (Jeffreys 1961, p. 253; Westfall, Johnson, Utts 1997); and it can be estimated using empirical Bayes methods (Efron, Tibshirani, Storey, and Tusher 2001). The next section provides a case study for prior assessment.

It should be mentioned prominently that Jeffreys, who pioneered the Bayesian testing paradigm, derived a Bayesian test for $H_0 : \mu_1 = \mu_2$ that is also a function of the two-sample t statistic (1). However, his test (Jeffreys 1961, sec. 5.41) uses an unusually complex prior that partitions the simple alternative $H_1 : \mu_1 \neq \mu_2$ into three disjoint events depending upon a hyperparameter μ: $H_{11} : \mu_2 = \mu \neq \mu_1$, $H_{12} : \mu_1 = \mu \neq \mu_2$, and $H_{13} : \{(\mu_1 \neq \mu_2)$ and neither equals $\mu\}$. Jeffreys further suggested prior probabilities in the ratio 1 : 1/4 : 1/4 : 1/8 for H_0, H_{11}, H_{12}, and H_{13}, respectively, adding another level of avoidable complexity. An additional concern with Jeffreys' two-sample t test is that it does not accommodate prior information about the alternative hypothesis.

3. A CASE STUDY: CLINICAL TRIALS

This section provides a case study in clinical trials to suggest how priors can be specified. Prior information to suggest the expected effect size (i.e., the value of λ) is routinely used for sample size calculations. In clinical trials, the outcome is considered positive if it is significant in the correct tail using a standard two-sided test with Type I error probability $\alpha = .05$. The large-sample sample size calculation formula for two-sample tests is given by

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\delta/\sigma)^2},$$

where $n = n_1 = n_2 = 2n_\delta$ is the sample size per group and β is the Type II error probability. The analyst must specify δ/σ. In a study powered at $100(1-\beta)\% = 80\%$, the analyst will have used

$$\delta/\sigma = \frac{z_{1-\alpha/2} + z_{1-\beta}}{n_\delta^{1/2}},$$

or $\delta/\sigma = (1.96{+}.84)/n_\delta^{1/2} = 2.80/n_\delta^{1/2}$ as an anticipated standardized effect size. For example, if $n = 100$, then the analyst anticipated $\delta/\sigma = 2.80/50^{1/2} = .396$ ["small" to "medium" in the terminology of Cohen (1988)].

The value σ_δ can be expressed as a function of the prior probability that the effect is in the wrong direction. For example, if $\lambda = .396$ and one thinks $P(\delta < 0 \mid \delta \neq 0) = .10$, then one obtains $\sigma_\delta = .309$ using normal distribution calculations. More generally, if $\lambda = 2.80/n_\delta^{1/2}$, then $\sigma_\delta = 2.19/n_\delta^{1/2}$, again assuming $P(\delta < 0 \mid \delta \neq 0) = .10$. These calculations involved the choice of zero for the tenth percentile of the prior on δ/σ; other percentiles could have been selected as well. Yet another calibration would involve selection of σ_δ based on a prior assumed value for $P(\delta/\sigma > 2\lambda \mid \delta \neq 0)$. It would be useful to try several such values to ensure consistency.

The remaining parameter to specify is $\pi_0 = P(H_0)$. Observing that it is unethical to randomize patients when the outcome is certain, the quantities $P(\delta \leq 0)$ and $P(\delta > 0)$ should be roughly comparable. One may set $\pi_0 = .5$, which, in conjunction with $P(\delta < 0 \mid \delta \neq 0) = .10$, yields $P(\delta \leq 0) = .5 + .10(.5) = .55$. Alternatively, one may first set $P(\delta \leq 0) = .5$, which, in conjunction with $P(\delta < 0 \mid \delta \neq 0) = .10$, implies $\pi_0 = .444$.

If historical (meta-analysis) data are available on rejection rates, one can check whether the prior specification is consistent with historical data by calculating the proportion of nulls that would be expected to be rejected. Since (for large sample sizes) the t statistic is approximately distributed as $N(0, 1)$ when $\delta = 0$, and approximately (marginally) distributed as $N(n_\delta^{1/2}\lambda, 1 + n_\delta\sigma_\delta^2)$ when $\delta \neq 0$, the proportion of rejected nulls (upper-tailed, $\alpha = .025$) is expected to be

$$\pi_0(.025) + \pi_1 \left[1 - \Phi\left(\frac{1.96 - n_\delta^{1/2}\lambda}{\sqrt{1 + n_\delta\sigma_\delta^2}}\right)\right].$$

Using, as suggested above, $\lambda = 2.80/n_\delta^{1/2}$, and $\sigma_\delta = 2.19/n_\delta^{1/2}$, this expression yields 33.1% rejections when $\pi_0 = .5$ and 36.5% when $\pi_0 = .444$. For comparison, Lee and Zelen (2000) surveyed the oncology literature for a variety of diseases and found that only 28.7% of the randomized trials reported rejection of the null hypothesis. Hence the choice of $\pi_0 = .5$, along with $(\lambda, \sigma_\delta) = (2.80/n_\delta^{1/2}, 2.19/n_\delta^{1/2})$, yields a model that is roughly consistent with results of randomized trials, at least in oncology.

4. AN EXAMPLE

The Data and Story Library (DASL; the Web site is http://lib.stat.cmu.edu/DASL) provides datasets that illustrate the use of basic statistical methods. Under the "Pooled t test" method one finds the "Calcium and Blood Pressure Story," which contains a subset of the data shown by Lyle et al. (1987). As posted on the DASL Web site, the data consist of blood pressure measurements on a subgroup of 21 African-American subjects, 10 who have taken calcium supplements and 11 who have taken placebo. The primary analysis variable is the blood pressure difference ("Begin" minus "End"). Summary statistics are as follows:

Group	n	mean	StdDev
Calcium	10	5.0000	8.7433
Placebo	11	−.2727	5.9007

Here, $s_p = 7.385$, $n_\delta = 5.238$, and $t = 1.634$; the positive t value suggests calcium is beneficial for reducing blood pressure. The two-sided frequentist p value, from the T_{19} distribution, is $p = .1187$.

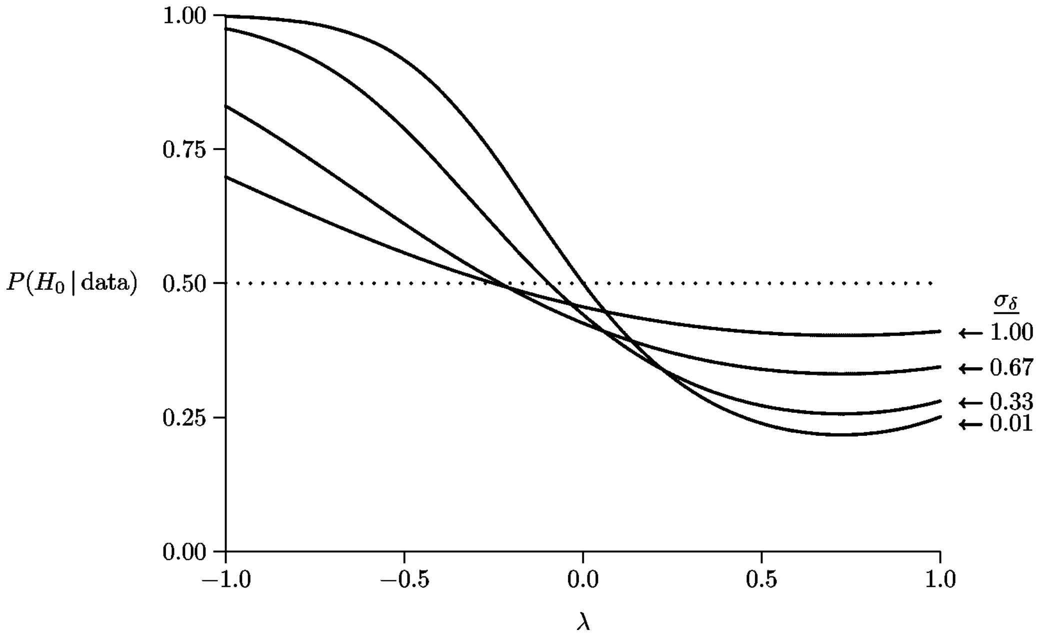


Figure 1. Posterior probabilities of H_0 as a function of λ, when π_0 = .5, and σ_δ = .01, .33, .67, 1.00 (solid lines). The prior probability π_0 = .5 is also shown (dotted line).

To perform the Bayesian test, priors must be specified. The previous section provided a case study to suggest particular values based on frequentist power considerations; however, this particular study was not powered for the African-American subgroup and those results do not apply. For the purposes of discussion, we will be as generic as possible in our initial specification and then provide sensitivity analysis.

Although not experts in the subject matter, if we suppose that the direction of an effect is completely uncertain, then we would set $\lambda = 0$. Further, we might assume that a standardized effect size greater than 1 is unlikely; setting $\sigma_\delta = 1/3$ seems reasonable as this would imply $P(|\delta/\sigma| > 1 \mid H_1) = .003$. We now compute the Bayes factor: $\text{BF} = .791$, suggesting that the data support $H_1 : \mu_1 \neq \mu_2$ better than $H_0 : \mu_1 = \mu_2$. If we wish to calculate posterior probabilities, then we need the prior probabilities as well; generically we may set $\pi_0 = .5$. With these settings we have $P(H_0 \mid \text{data}) = .442$. Although it is true that the null hypothesis that calcium has no effect is less likely after seeing the data, the results are not compelling.

Figure 1 shows a sensitivity analysis of the posterior probability $P(H_0 \mid \text{data})$ with respect to λ, for $\sigma_\delta = .01, .33, .67,$

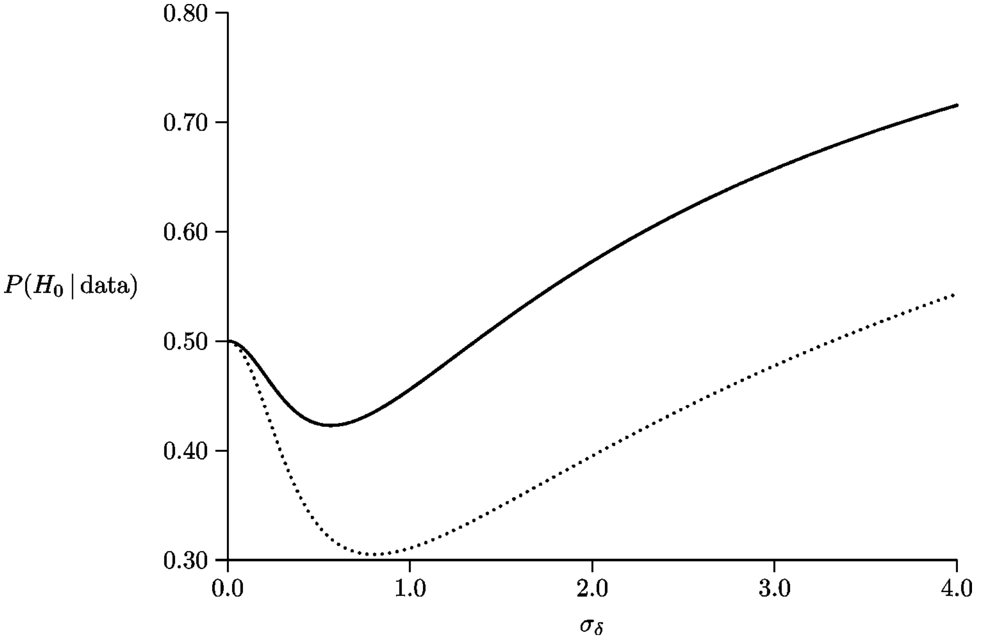


Figure 2. Posterior probability of H_0 as a function of σ_δ, when λ = 0 and π_0 = .5, both for the observed data (solid line) where the p value is p = .1187, and for hypothetical data with p = .05 (dotted line). The minimum posterior probability for the case where p = .05 is $P(H_0 \mid$ data) = .305, illustrating Berger and Sellke's "irreconcilability" of frequentist p values with posterior probabilities in the case of the two-sample t test.

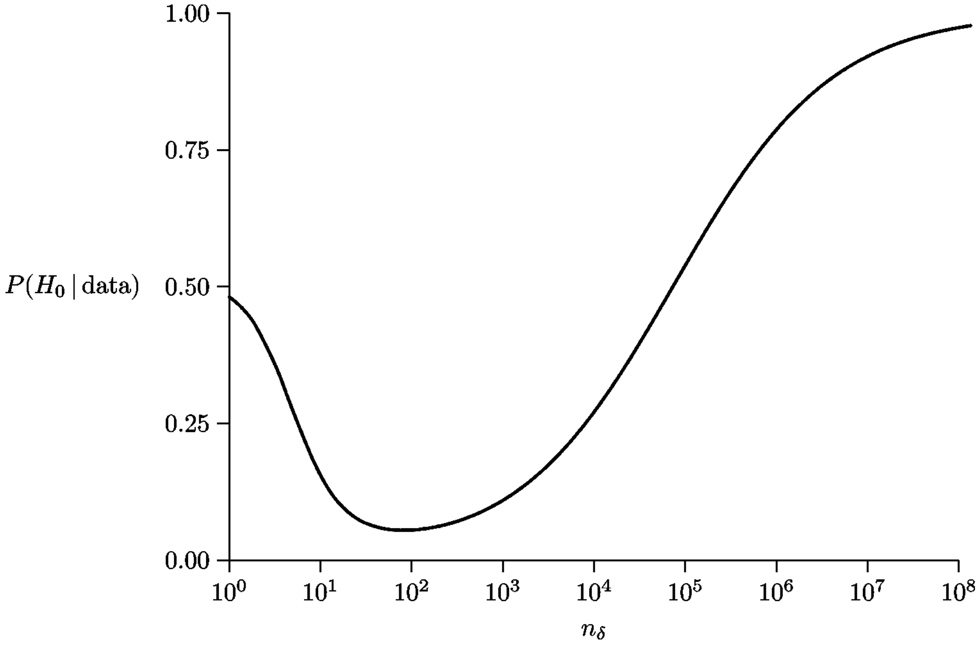


Figure 3. Posterior probability of H_0 as a function of n_δ when π_0 = .5, (λ; σ_δ) = (0, 1/3), and t = 3.00, illustrating Lindley's paradox.

and 1.00, assuming the prior probability is $\pi_0 = .5$. There is not reasonable evidence against H_0 no matter which combinations of the prior values λ and σ_δ are chosen. Smaller posterior probabilities of H_0 occur for λ near the sample estimate $(\overline{y}_1 - \overline{y}_2)/s_p = .714$, but even the smallest value ($P(H_0 \,|\, \text{data}) = .217$, occurring when $\sigma_\delta = .01$) is not small enough to rule out H_0. The graph also shows large differences in the posterior probability for different λ; for example, if λ is near -1 (meaning that, if there is a difference, then calcium is expected to be much worse than placebo for reducing blood pressure), the positive t statistic $t = 1.634$ provides much more evidence for H_0 than for H_1. Although this lack of sensitivity may be troubling, one can question whether such values of λ would have been reasonable choices; after all, presumably the goal of the study was to assess whether calcium causes greater reductions in blood pressure, and therefore nonnegative values of λ might have been more plausible a priori.

Figure 2 shows the special case where $\lambda = 0$ and σ_δ is varied over a wider range. Here the minimum posterior probability is $P(H_0 \,|\, \text{data}) = .423$, much larger than the frequentist p value ($p = .1187$). This graph highlights the central point of Berger and Sellke (1987); namely, that $P(H_0 \,|\, \text{data})$ is typically much higher than the frequentist p value. For comparison, the posterior probability that results when $t = 2.093$, for which the frequentist two-sided p value is exactly .05, is also displayed in the graph as a dotted line. The curve corresponding to $t = 2.093$ ($p = .05$) dramatizes Berger and Sellke's (perhaps surprising) conclusion that H_0 will be true in at least 30% of studies for which the p value is observed to be in a small neighborhood of .05 (assuming that H_0 is true, a priori, in 50% of all studies considered, and assuming that the prior effect sizes for the nonnull studies come from a symmetric unimodal distribution centered at 0).

Although the posterior probability $P(H_0 \,|\, \text{data})$ does not appear to be overly sensitive to the prior inputs λ and σ_δ (provided a sensible range of inputs is considered), it is clearly much more sensitive to the prior probability π_0. For example, when (λ, σ_δ) $= (0, 1/3)$, the posterior probabilities are determined as follows:

Prior Probability π_0 :	.100	.250	.500	.750	.900
Posterior Probability $P(H_0 \,|\, \text{data})$:	.081	.209	.442	.704	.877

The posterior is sensitive to the prior as expected, but what is more interesting is that *these* data barely modify one's prior belief about H_0.

As a concluding note, it is simple to discuss "Lindley's Paradox" (Lindley 1957) using (2). Lindley noticed that data from large sample sizes that are "highly significant" from a frequentist standpoint can support H_0 better than H_1. Imagine, in the case above, that $t = 3.00$, highly significant by any measure. From the frequentist standpoint, the result would be considered even more significant for larger values of n_1 and n_2. On the other hand, $t = 3.00$ becomes less likely under H_1 for extremely large n_δ: the denominator of (2) decreases (since the variance $1 + n_\delta \sigma^2$ increases) while the numerator remains fixed. Figure 3 shows the effect of increasing n_δ (assuming $n_1 = n_2$) on the posterior probability of H_0 when $t = 3.00$, showing a minimum posterior probability of .055 at $n_\delta = 81.5$ ($n_1 = n_2 = 163$), and increasing to 1.0 thereafter for larger n_δ. This seeming "paradox" is not really a paradox at all, since the frequentist statistical significance with large n_δ is a result of a large sample amplification of a very small effect size.

5. CONCLUSION

The two-sample comparison is one of the most important problems in statistics. From the teaching standpoint, two-sample testing problems are usually much more interesting and relevant than single-sample problems. However, it is difficult to find the Bayesian two-sample t test explicitly in the literature. We present a simple, relatively easy-to-elicit prior for which the Bayes factor for the two-sample comparison of means is a function of the usual two-sample t statistic, thus providing a Bayesian motivation for this statistic. Because the analytic result itself is easy to teach and compute, and because it facilitates discussions of Bayesian concepts such as prior selection and Lindley's Paradox, we recommend that this test be incorporated routinely when teaching elementary statistics from a Bayesian perspective.

[Received June 2003. Revised March 2005.]

REFERENCES

Albert, J. (1997a), "Teaching Bayes Rule: A Data-Oriented Approach," *The American Statistician*, 51, 247–253.

——— (1997b), "Bayesian Testing and Estimation of Association in a Two-Way Contingency Table," *Journal of the American Statistical Association*, 92, 685–693.

Albert, J., and Rossman, A. (2001), *Workshop Statistics: Discovery with Data, A Bayesian Approach*, Emeryville, CA: Key College.

Antelman, G. (1997), *Elementary Bayesian Statistics*, Cheltenham: Edward Elgar Publishing.

Berger, J. O., Boukai, B., and Wang, Y. (1997), "Unified Frequentist and Bayesian Testing of a Precise Hypothesis," *Statistical Science*, 12, 133–160.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Berry, D. A. (1996), *Statistics: A Bayesian Perspective*, Belmont, CA: Wadsworth.

——— (1997), "Teaching Elementary Bayesian Statistics with Real Applications in Science," *The American Statistician*, 51, 241–246.

Bolstad, W. M. (2004), *Introduction to Bayesian Statistics*, Hoboken, NJ: Wiley.

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society,* Ser. B, 57, 473–484.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press.

Dickey, J. M. (1971), "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters," *Annals of Mathematical Statistics*, 42, 204–223.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Gönen, M., Westfall, P. H., Johnson, W. O., and Lu, Y. (2004), "The Two-Sample t Test: A Bayesian Perspective," unpublished manuscript, http://www.ba.ttu.edu/isqs/westfall/Bayes2samplet.pdf.

Jeffreys, H. (1961), *Theory of Probability*, Oxford: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley.

Lee, S. J., and Zelen, M. (2000), "Clinical Trials and Sample Size Considerations: Another Perspective," *Statistical Science*, 15, 95–110.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192.

Lyle, R. M., Melby, C. L., Hyner, G. C., Edmondson, J. W., Miller., J. Z., and Weinberger, M. H. (1987), "Blood Pressure and Metabolic Effects of Calcium Supplementation in Normotensive White and Black Men," *Journal of the American Medical Association*, 257, 1772–1776.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian Perspective on the Bonferroni Adjustment," *Biometrika*, 84, 419–427.

Zellner, A., and Siow, A.(1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, Valencia: University Press, pp. 585–603.