On Bayes Factors for Success Rate A/B Testing

Maciej Skorski
DELL, Austria

Keywords: Hypothesis Testing, Bayesian Statistics, AB Testing, Information Geometry.

Abstract: This paper discusses Bayes factors, an alternative to classical frequentionist hypothesis testing, within the standard A/B proportion testing setup - observing outcomes of independent trails (which finds applications in industrial conversion testing). It is shown that the Bayes factor is controlled by the *Jensen-Shannon divergence* of success ratios in two tested groups, and the latter one is bounded (under mild conditions) by *Welch's t-statistic*. The result implies an optimal bound on the *necessary sample size* for Bayesian testing, and demonstrates the relation to its frequentionist counterpart (effectively bridging Bayes factors and p-values).

1 INTRODUCTION

1.1 Background and Motivation

A/B Proportions Testing. A/B testing is the methodology of collecting data from two parallel experiments, and deciding which group performs better by means of statistical inference (to account for effects that may be due to change). The most frequent use case concerns two success-counting experiments, for example how many customers convert (purchase, subscribe etc) on two versions of a web page (for example the old page vs the optimized one).

In order to make the decision statistic-driven the business question is formulated as the question about *unknown* conversion rates p_1, p_2 in the compared groups, that are to be estimated from collected (observed) data $\mathcal{D}$. Usually one states the problem as choosing one of the two possibilities

(a) The *null hypothesis* states that conversion rates are equal (zero-effect), written as $H_0 = \{p_1 = p_2\}$

(b) The *alternative hypothesis* claims a difference, for example $H_a = \{p_1 \neq p_2\}$ or $H_a = \{p_1 = p_2 + \delta\}$

A/B Model Assumptions. We assume that in each of the two groups we observe r independent Bernoulli variables (each one is success or failure). Success rates p_i for group i are to be estimated from ths data (we allow unknown rates p_i to depend on hypotheses via prior distributions). The data set $\mathcal{D}$ contains two binary sequences describing outcomes (success or failure for each trial) for each of the two groups.

Statistical Testing and Bayes Factors. The frequentionist approach falsifies the null hypothesis based on the *two-sample t-test* (Welch, 1938), so that it is rejected when the test value is sufficiently unlikely for given data (probability of which, falsely rejecting the null, is p-value).

The Bayesian approach is more coherent and flexible as it directly compares the likelihoods of two hypotheses (for given data). By Bayes' theorem

$$\underbrace{\frac{\Pr[H_0|\mathcal{D}]}{\Pr[H_a|\mathcal{D}]}}_{\text{posterior odds}} = \underbrace{\frac{\Pr[\mathcal{D}|H_0]}{\Pr[\mathcal{D}|H_a]}}_{\text{data likelihood ratio}} \cdot \underbrace{\frac{\Pr[H_0]}{\Pr[H_a]}}_{\text{prior odds}} \tag{1}$$

where the likelihood ratio

$$K = \frac{\Pr[\mathcal{D}|H_0]}{\Pr[\mathcal{D}|H_a]} \tag{2}$$

is also called the *Bayes factor*. Prior odds usually equal 1, when one gives no prior preference to H_0 or H_a. Then the posterior odds equal K and the decision depends on its magnitude: the bigger K from 1, the more it supports H_0. This may be also seen as deciding upon the expected (posterior) cost of a certain risk function (Lavine and Schervish, 1999).

Confidence scales depending on the magnitude have been developed (Kass and Raftery, 1995; Jeffreys, 1998; Lee and Wagenmakers, 2014).

Hypotheses are arbitrary "prior" probabilities on rates p_1, p_2 which can be formally written as $H = \{(p_1, p_2) \to \mathbb{P}_H(p_1, p_2)\}$. This includes point statements of the form $H = \{p_1 = p_2 = 0.01\}$ as well us uncertainty distributions such as beta priors $H = \{p_1 = p_2 = p, p \sim \mathsf{Beta}(0.5, 0.5)\}$. Priors usually are weakly informative, that is they give less preference to "unrealistic" values like those near 0 or 1.

Skorski, M.
On Bayes Factors for Success Rate A/B Testing.
DOI: 10.5220/0007954503320339
In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 332-339
ISBN: 978-989-758-377-3

Under prior distributions specified the data likelihood equals

$$\begin{aligned}\Pr[\mathcal{D}|H] &= \int \Pr[\mathcal{D}|p_1,p_2]\cdot \mathrm{d}\mathbb{P}_H(p_1,p_2)\\ &= \int \prod_{i=1}^{2} p_i^{\bar{p}_i r}(1-p_i)^{(1-\bar{p})_i r}\cdot \mathrm{d}\mathbb{P}_H(p_1,p_2)\end{aligned} \tag{3}$$

where r is the number of events and $\bar{p}_i$ is the success rate *observed* for each group i in the collected data $\mathcal{D}$. Computation of the corresponding factor K can be done in statistical software such as in R package `BayesFactor` (Morey and Rouder, 2018).

1.2 Problem: Bayesian A/B testing

Estimates, neither frequentionist nor bayesian, will not be conclusive without sufficiently many samples. Frequentionists widely use rules of thumbs that are derived based on t-tests. Under the bayesian methodology this is little more complicated because hypotheses can be arbitrary priors over parameters (hence composed out of infinitely many choices). Under the described A/B model, we answer the following questions

- When, given data, a *bayesian* hypothesis on zero effect may be rejected (we want to guarantee that for any H_0 it holds $K \ll 1$ for some H_a)?
- What is the relation to the classical t-test?
- What are most plausible null and alternative hypothesis, for a given dataset?

This will allow us to understand *data limitations* when doing bayesian inference, and relate them to widely-spread frequentionists rules of thumb. It is important to note that Bayesian modelling has recently become very popular in industrial conversion optimization (Keser, 2017), so that these questions are also of considarable practical interest.

1.3 Related Works and Contribution

Our problem, as stated, is a question about *maximizing minimal Bayes factor*; this is because we compare *any null* against its *most favorable* alternative. For certain simple problems, particularly for testing normality, minimal Bayes factors are known to be related to frequentionists p-values (Edwards et al., 1963; Kass and Raftery, 1995; Goodman, 1999), which bridges the Bayesian and frequentionists worlds. This should be contrasted with a wide-spread belief that both methods are very incompatible (Kruschke and Liddell, 2018).

The novel contributions of this paper are (a) determining the Bayes factor - for testing success ratios (b) demonstrating the relation to the frequentionist approach and (c) discussion of Bayesian sample bounds. To the best author's knowledge, this is the first result of this sort in the context of testing success rates in independent trials.

1.3.1 Main Result: Bayes Factor and Welch's Statistic

The following theorem shows that no "zero-effect" hypothesis can be falsified, unless the number of samples is big in relation to a certain *data statistic*. This statistic turns out to be the Jensen-Shannon divergence, well-known in information theory; we further show how to relate it to the Welch's t-statistic. Doing so we connect the classical frequentionist analysis and the Bayes factors analysis.

Theorem 1 (Bayes Factors for Success Rate Testing)**.** *Consider two experiments, each with r independent trials with unknown success probabilities p_1 and p_2 respectively. If observed data $\mathcal{D}$ has $r\cdot\bar{p}_i$ successes for group $i=1,2$, then*

$$\max_{H_0:\{p_1=p_2\}} \min_{H_a} \frac{\Pr[\mathcal{D}|H_0]}{\Pr[\mathcal{D}|H_a]} = \mathrm{e}^{-2r\cdot\mathbf{JS}(\bar{p}_1,\bar{p}_2)} \tag{4}$$

where the maximum is over null hypotheses H_0 such that $p_1=p_2$, the minimum is over all valid alternative hypotheses (priors) H_a over p_1,p_2, and **JS** *denotes the Jensen-Shannon divergence.*

When comparing the Jensen-Shannon divergence with the Welch's t-statistic one should note that the second one is unbounded, as illustrated in Figure 1. Specifically, the Welch's t is unbounded where one rate is close to zero but the second one is closed to one. It is however possible to have a bound of the form $\mathbf{JS}=\Omega(\mathbf{t}^2_{\mathrm{Welch}})$, under mild additional assumptions for example when both rates are smaller than 0.5 (which in practice don't limit the usability). The result is formally stated in the theorem below, the proof appears in Section 4. We note that the current proof fails to achieve the optimal constant (see Section 5).

Theorem 2 (Comparison with T-Test)**.** *Under the condition $0\leqslant\bar{p}_1,\bar{p}_2\leqslant\frac{1}{2}$ it holds that the Jensen-Shannnon divergence is bounded from below by the Welch's t-statistic*

$$\mathbf{JS}(\bar{p}_1,\bar{p}_2) \geqslant \frac{\mathbf{t}_{\mathrm{Welch}}(r,\bar{p}_1,\bar{p}_2)^2}{32r} \tag{5}$$

so that the Bayes factor in Equation (4) *can be bounded by*

$$\max_{H_0:\{p_1=p_2\}} \min_{H_a} \frac{\Pr[H_0|\mathcal{D}]}{\Pr[H_a|\mathcal{D}]} \leqslant \mathrm{e}^{-\mathbf{t}_{\mathrm{Welch}}(r,\bar{p}_1,\bar{p})^2/16}. \tag{6}$$

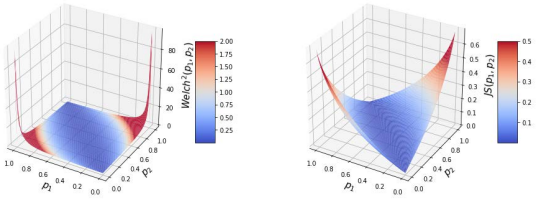

Figure 1: Surface plots of the squared t-statistic $\mathbf{t}^2_{\text{Welch}}$ (left) and Jensen-Shannon divergence **JS** which controls the Bayes factor (right), as functions of success rates p_1, p_2. Note that $\mathbf{t}_{\text{Welch}}$ is unbounded (around the corners).

Remark 1 (Min-max Game Intepretation)**.** *The min-max formula in Equation* (6) *comes from the fact that we evaluate every null against its most plausible alternative (so that the bound holds regardless of the null prior). This can be seen as a two-player zero-sum game where one player chooses the null hypothesis, the second player chooses the alternative and the payoff is the Bayes factor. Theorem 1 describes the saddle point in this game.*

The following corollary shows what are most "plausible" null and alternatives for a given data set (they realize equality in Equation (4))

Corollary 1 (Characterization of Most Favorable Hypotheses)**.** *Note that*

- *Maximally favorable alternative (H_a which maximizes $\Pr[\mathcal{D}|H_a]$) is $p_1 = \bar{p}_1$ and $p_2 = \bar{p}_2$*
- *Maximally favorable null H_0 on zero-effect, that is of the form $p_1 = p_2$, is given by $p_1 = p_2 = \frac{\bar{p}_1 + \bar{p}_2}{2}$*

If null is of the form $p_1 = p_2 = p$ for some constant p, then the bound becomes $\mathrm{e}^{-r\cdot\mathbf{KL}(\bar{p}_1,p)-r\cdot\mathbf{KL}(\bar{p}_2,p)}$.

1.3.2 Application: Sample Bounds

The main result implies the following *sample size rule*

Corollary 2 (Bayesian Sample Bound)**.** *To confirm the non-zero effect ($p_1 \neq p_2$) the number of samples r for the bayesian method should be*

$$r_{bayes} \geqslant \log K_{critical} \cdot \frac{1}{2\mathbf{JS}(\bar{p}_1, \bar{p}_2)} \tag{7}$$

where usually $K_{critical} \approx 10$[1]*. Under the frequenionist method the rule of thumb is $\mathbf{t}_{\text{Welch}} \gg t_{critical}$, which gives (see Section 2)*

$$r_{freq} \geqslant t_{critical} \cdot \frac{\bar{p}_1(1-\bar{p}_1) + \bar{p}_2(1-\bar{p}_2)}{(\bar{p}_1 - \bar{p}_2)^2} \tag{8}$$

where usually $t_{critical} \approx 1.9$[2]

[1]This corresponds to the Bayes factor of 10^{-1} against null, interpreted as strong evidence in common scales (Lee and Wagenmakers, 2014).

[2]This roughly holds for the significance level of 0.95, the exact value depends also on degrees of freedom.

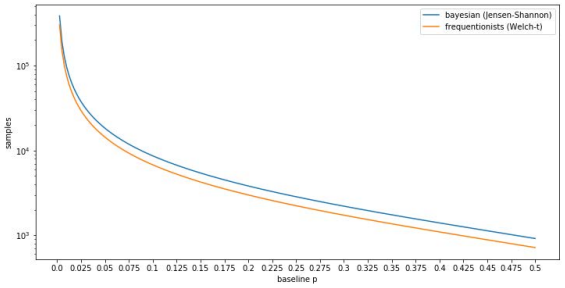

Figure 2: Comparison of the bayesian (7) and the frequentionist (8) sample lower bounds, where observed data are $\bar{p}_1 = p$ and $\bar{p}_2 = p \cdot (1+\delta)$ for $\delta = 0.1$ (relative uplift by 10%), and the zero-effect hypothesis $p_1 = p_2$ is to be rejected. Bounds are away by a constant factor, here are very close for constants calibrated under typical rejection rules: t-statistic of 1.9 and Bayes factor of 0.1.

Note that both formulas needs assumptions on locations of expected rates; testing smaller effects or smaller conversion rates require more samples.

Exact constants, hidden under $K_{critcal}$ and $t_{critcal}$, depend on the significance one wants to achieve: p-value for the extreme t-statistic, respectively the magnitude of the Bayes factor. Apart from constants (or when constants are callibrated for "typical" tests strength) bounds in Equation (7) and Equation (8) are close to each other. The difference is illustrated on Figure 2, for the case when one wants to prove the difference (reject the zero-effect hypothesis) in presence of an observed lift of 10%. The code is attached in Section 5.

It is important to stress that our lower bounds hold with respect to zero-effect hypotheses and regardless of priors, testing effect of a fixed size or using more diffuse priors may require more samples.

1.3.3 Application: Bayes Factors vs P-Values

Since high values of $\mathbf{t}_{\text{Welch}}$ mean small p-values, we conclude that the frequentionist p-values bound the Bayes factor and indeed are evidence against a null-hypothesis in the well-defined bayesian sense.

However, because of the scaling $\mathbf{t}_{\text{Welch}} \to \mathrm{e}^{-\Omega(\mathbf{t}^2_{\text{Welch}})}$ in the minimal Bayes factor in Equation (6), Bayesian rejection corresponds to p-values much lower than the standard frequentionist threshold of 0.05. In a way, the bayesian approach is more conservative and reluctant to reject than frequentionist tests; this conclusion is shared with (Goodman, 1999).

2 PRELIMINARIES

Entropy, Divergence. The binary cross-entropy of p and q is defined by

$$\mathbf{H}(p,q) = -p\log q - (1-p)\log(1-q) \tag{9}$$

which becomes the standard (Shannon) binary entropy when $p = q$, denoted as $\mathbf{H}(p) = \mathbf{H}(p,p)$. The Kullback-Leibler divergence is defined as

$$\mathbf{KL}(p,q) = \mathbf{H}(p,q) - \mathbf{H}(p) \tag{10}$$

and is non-negative. The Jensen-Shannon divergence (Lin, 1991) is defined as

$$\mathbf{JS}(p,q) = \frac{1}{2}\mathbf{KL}\left(p, \frac{p+q}{2}\right) + \frac{1}{2}\mathbf{KL}\left(q, \frac{p+q}{2}\right) \tag{11}$$

being symmetric and non-negative (because KL divergence is non-negative). Alternatively, using Equation (10) and Equation (9) we can write

$$\mathbf{JS}(p,q) = \mathbf{H}\left(\frac{p+q}{2}\right) - \frac{1}{2}\mathbf{H}(p) - \frac{1}{2}\mathbf{H}(q) \tag{12}$$

which shows that the Jensen-Shannon divergence is *bounded*.

The following lemma shows that the cross-entropy function is *convex* in the second argument. This should be contrasted with the fact that the entropy function (of one argument) is concave.

Lemma 1 (Convexity of Cross-entropy)**.** *For any p the mapping $x \to \mathbf{H}(p,x)$ is convex in x.*

Proof. Since $-p\log(\cdot)$ for fixed $p \in [0,1]$ is convex we obtain

$$-\gamma_1 p\log x_1 - \gamma_2 p\log x_2 \geqslant -p\log(\gamma_1 x_1 + \gamma_2 x_2)$$

for any x_1, x_2 and any $\gamma_1, \gamma_2 \geqslant 0$, $\gamma_1 + \gamma_2 = 1$. Replacing x_i by $1 - x_i$ and p by $1 - p$ in the above inequality gives us also

$$\begin{aligned} -\gamma_1(1-p)\log(1-x_1) - \gamma_2(1-p)\log(1-x_2) \\ \geqslant -(1-p)\log(\gamma_1(1-x_1) + \gamma_2(1-x_2)) \\ = -(1-p)\log(1-\gamma_1 x_1) - \gamma_2 x_2) \end{aligned}$$

Adding side by side yields

$$\gamma_1\mathbf{H}(p,x_1) + \gamma_2\mathbf{H}(p,x_2) \geqslant \gamma_1\mathbf{H}(p,x_1) + \gamma_2\mathbf{H}(p,x_2)$$

which finishes the proof. This argument works for multivariate case, when p, x are probability vectors. □

Convexity Properties of Jensen-Shannon Divergence.

Lemma 2. *Let $\delta = p - q$, then for every fixed q we have*

$$\frac{\partial^2}{\partial\delta^2}\mathbf{JS}(p,q) = \frac{\frac{1}{2}\left(p^2 - 2pq - q^2 + 2q\right)}{p\,(p-1)\,(p+q)\,(p+q-2)} \tag{13}$$

which is strictly positive for $0 < p, q < 1$.

Proof. The derivative is calculated in Section 5, with the Python package `SymPy` (Meurer et al., 2017). The numerator (skiping the constant $\frac{1}{2}$) can be written as $2q(1-q) + (p-q)^2$ which is non-negative. The denumerator is non-negative because p, q are between 0 and 1. □

Bernoulii Variables. For a Bernoulli variable with success probability p we denote by $\mathbb{V}\text{ↄ}\backslash(p) = p(1-p)$ the variance. We have the following identity

$$\mathbb{V}\text{ↄ}\backslash(p) + \mathbb{V}\text{ↄ}\backslash(q) = 2\mathbb{V}\text{ↄ}\backslash\left(\frac{p+q}{2}\right) - \frac{(p-q)^2}{2} \tag{14}$$

which in particular demonstrates that the variance is concave.

2-Sample Test. To decide whether means in two groups are equal, under the assumption of unequal variances, one performs the Welch's t-test with the statistic given by (Derrick et al., 2016)

$$\mathbf{t}_{\text{Welch}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{r_1} + \frac{s_2^2}{r_2}}} \tag{15}$$

where s_i are sample variances and μ_i are sample means for group $i = 1, 2$. The null hypothesis is rejected unless the statistic is sufficiently high (in absolute terms). In our case the formula simplifies to

Claim 1. *If $r\theta_1$ and $r\theta_2$ success out of r trials have been observed, respectively in the first and the second group, then*

$$\mathbf{t}_{\text{Welch}}(r,\theta_1,\theta_2) = r^{\frac{1}{2}} \cdot \frac{\theta_1 - \theta_2}{\sqrt{\theta_1(1-\theta_1) + \theta_2(1-\theta_2)}} \tag{16}$$

3 PROOF OF THEOREM 1

Notation. We change the notation slightly, unknown success rates will be p and q, and the number of observed successes $r \cdot \theta_1, r \cdot \theta_2$.

Best Alternative. Maximizing over *alll* possible priors $\mathbb{P}_a$ over pairs (p,q) and using Equation (3) we get

$$\max_{\mathbb{P}_a} \Pr[\mathcal{D}|H_a] = \max_{\mathbb{P}_a} \int \mathrm{e}^{-r\mathbf{H}(\theta_1,p)-r\mathbf{H}(\theta_1,q)} \mathbb{P}_a(p,q)\mathrm{d}(p,q) \quad (17)$$

Since $\mathbf{H}(\theta_1,p) = \mathbf{H}(\theta_1,p) + \mathbf{KL}(\theta_1,p)$, $\mathbf{H}(\theta_1,q) = \mathbf{H}(\theta_1,q) + \mathbf{KL}(\theta_1,q)$ and $\mathbf{KL}$ is non-negative we conclude that the maximum over the integrals equals

$$\max_{\mathbb{P}_a} \Pr[\mathcal{D}|H_a] = \mathrm{e}^{-r\mathbf{H}(\theta_1)-r\mathbf{H}(\theta_2)} \quad (18)$$

achieved for $\mathbb{P}_a$ being a unit mass at $(p,q) = (\theta_1,\theta_2)$, that is at observed rates.

Best Null. Let H_0 states that the effect is 0. This is equivalent to a prior $\mathbb{P}_0(p)$ on the baseline conversion p with the condition $q = p$. We obain

$$\Pr[\mathcal{D}|H_0] = \int \mathrm{e}^{-r\mathbf{H}(\theta_1,p)-r\mathbf{H}(\theta_2,p)} \mathrm{d}\mathbb{P}_0(p) \quad (19)$$

The integrand, due to Lemma 1, has a global maximum at some p. Thus the integral is maximized for $\mathrm{d}\mathbb{P}_0(p)$ being a point mass.

Bayes Factor. Using the above observations on most plausible hypotheses we can write the max-min Bayes factor (in favor of H_0) as

$$\max_{H_0}\min_{H_a} \frac{\Pr[\mathcal{D}|H_0]}{\Pr[\mathcal{D}|H_a]} = \frac{\max_{H_0}\Pr[\mathcal{D}|H_0]}{\max_{H_a}\Pr[\mathcal{D}|H_a]} = \mathrm{e}^{-r\cdot(\mathbf{H}(\theta_1,p)+\mathbf{H}(\theta_1,p)-\mathbf{H}(\theta_1)-\mathbf{H}(\theta_2))} \quad (20)$$

Using the relation between the KL divergence and cross-entropy we obtain

$$\max_{H_0}\min_{H_a} \frac{\Pr[\mathcal{D}|H_0]}{\Pr[\mathcal{D}|H_a]} = \mathrm{e}^{-r\mathbf{KL}(\theta_1,p)-r\mathbf{KL}(\theta_2,p)} \quad (21)$$

We will use the following observation

Claim 2. *The expression* $\mathbf{KL}(\theta_1,p)+\mathbf{KL}(\theta_2,p)$ *is minimized under* $p = \theta^* = \frac{\theta_1+\theta_2}{2}$*, and achieves value* $2\mathbf{JS}(\theta_1,\theta_2)$.

Proof. We have

$$\mathbf{KL}(\theta_1,p)+\mathbf{KL}(\theta_2,p) = \mathbf{H}(\theta_1,p)+\mathbf{H}(\theta_2,p)-\mathbf{H}(\theta_1)-\mathbf{H}(\theta_2)$$

Now the existence of the minimum at $p = \theta^*$ follows by convexity of $p \to \mathbf{H}(\theta_1,p)+\mathbf{H}(\theta_2,p)$, proved in Lemma 1. We note that $\mathbf{H}(\theta_1,p) + \mathbf{H}(\theta_2,p) = 2\mathbf{H}\left(\frac{\theta_1+\theta_2}{2},p\right)$ for any p (by definition), and thus for $p = \frac{\theta_1+\theta_2}{2} = \theta^*$ we obtain $\mathbf{H}(\theta_1,p) + \mathbf{H}(\theta_2,p) = 2H(\theta^*)$ and $\mathbf{KL}(\theta_1,p)+\mathbf{KL}(\theta_2,p) = 2\mathbf{H}(\theta^*) - \mathbf{H}(\theta_1) - \mathbf{H}(\theta_2)$. This combined with the definition of the Jensen-Shannon divergence finishes the proof. □

We can now bound Equation (21) as

$$\min_{H_a} \frac{\Pr[\mathcal{D}|H_0]}{\Pr[\mathcal{D}|H_a]} \leqslant \mathrm{e}^{-2r\cdot\mathbf{JS}(\theta_1,\theta_2)} \quad (22)$$

This ends the proof of Theorem 1, from the proof we also conclude Corollary 1.

4 PROOF OF THEOREM 2

In order to compare **JS** and $\mathbf{t}_{\text{Welch}}$ we use the parametrization $q = p + \delta$. For convenience we slightly abuse the notation writing $\mathbf{t}_{\text{Welch}}(p,q) = \mathbf{t}_{\text{Welch}}(r,p,q)/\sqrt{r} = \mathbf{t}_{\text{Welch}}(1,p,q)$. The result reduces to the following lemma

Lemma 3 (Convexity of the Gap between Jensen-Shannon Divergence and Welch's t)**.** *In the region*

$$0 \leqslant q \leqslant p \leqslant \frac{1}{2} \quad (23)$$

we have that $\mathbf{JS}(p,q) - \frac{1}{32}\cdot\mathbf{t}_{\text{Welch}}(p,q)^2$ *is convex in* $\delta = p-q$ *for any fixed* q*. In particular for every* q *it achieves the minimal value at* $\delta = 0$*, which is equal to* 0.

Proof. It suffices to consider $q \leqslant p$ as both **JS** and $\mathbf{t}_{\text{Welch}}$ are symmetric. Because of continuity we consider the strict inequalities $0 < q < p < \frac{1}{2}$. The proof involves symbolic differentiation and factoring which we do in the Python package `SymPy` (Meurer et al., 2017).

We start with the ratio of the second derivatives

Claim 3 (Second Derivatives Ratio)**.** *We have that*

$$\frac{\frac{\partial^2}{\partial\delta^2}\mathbf{JS}(p,q)}{\frac{\partial^2}{\partial\delta^2}\mathbf{t}_{\text{Welch}}(p,q)} = \frac{1}{4}\cdot\psi\cdot\phi \quad (24)$$

where

$$\psi(p,q) = \frac{\left(p-p^2+q-q^2\right)^3}{p\,(1-p)\,(p+q)\,(2-p-q)}$$

$$\phi(p,q) = \frac{p^2-2pq-q^2+2q}{-2p^3q+p^3+3p^2q+6pq^3-9pq^2-3q^3+4q^2}$$

Proof of Claim. The code deriving formulas is included in Section 5. □

Claim 4 (Denumerator of ϕ is non-negative)**.** *Let* $V = -2p^3q + p^3 + 3p^2q + 6pq^3 - 9pq^2 - 3q^3 + 4q^2$ *be the denumerator of* ϕ*. Then* $V \geqslant 4p^2(1-p^2)$.

Proof of Claim. It holds that $\frac{\partial^2 V}{\partial p^2} = 6 \cdot (p+q-2pq) = 6(p(1-q)+q(1-p))$ which shows that V is convex in p for all $(p,q) \in [0,1]^2$. Next, we have $\frac{\partial V}{\partial p} = -3\,(p-q)\left(2pq - p + 2q^2 - 3q\right)$. Looking for global minimas we note that the second factor $2pq - p + 2q^2 - 3q = -q(1-p) - p(1-q) - 2q(1-q)$ is always non-positive and zero if and only if $p = q = 0$ or $p = q = 1$. The first factor gives us $q = p$. Summing up, the global minimum for each p is given by $q = p$. Substituting this we obtain $V(p,q) \geqslant V(q,q) = 4p^2(1-p)^2$ which finishes the proof. □

Claim 5 (Bound on ϕ)**.** *Let* U, V *be as before. Under the condition* $0 < q < p < \frac{1}{2}$ *we have* $U \cdot p(1-p) - \frac{1}{2}V \geqslant 0$*, and the claim before implies* $\phi = \frac{U}{V} \geqslant \frac{1}{2p(1-p)}$.

Proof of Claim. We have

$$\begin{aligned} Up(1-p) - \frac{V}{2} &= \frac{p-q}{2}\left(\frac{1}{2} - p\right)\left(\frac{1}{2}p^2 - pq - \frac{3}{2}q^2 + 2q\right) \\ &= \frac{p-q}{2}\left(\frac{1}{2} - p\right)\left(\frac{(p-q)^2}{2} + 2q(1-q)\right) \end{aligned}$$

which is non-negative when $\frac{1}{2} \geqslant p \geqslant q \geqslant 0$. By the previous claim U is non-negative so that we can divide both sides. For completeness we include the code used for deriving the formulas in Section 5. □

Summing up, from the claims we obtain

Claim 6 (Bounding the Second Derivative Ratio)**.** *Under the condition* $0 < q < p < \frac{1}{2}$ *we have*

$$\frac{\partial^2}{\partial \delta^2} \frac{\mathbf{JS}(p,q)}{\mathbf{t}_{\text{Welch}}(p,q)} \geqslant \frac{(\mathbb{V}\text{ↄ}\diagdown(p) + \mathbb{V}\text{ↄ}\diagdown(q))^3}{32\mathbb{V}\text{ↄ}\diagdown^2(p)\mathbb{V}\text{ↄ}\diagdown\left(\frac{p+q}{2}\right)} \tag{25}$$

which is bigger than $\frac{1}{32}$.

Proof of Claim. The previous claim implies

$$\begin{aligned} \frac{\partial^2}{\partial \delta^2} \frac{\mathbf{JS}(p,q)}{\mathbf{t}_{\text{Welch}}(p,q)} &= \frac{1}{4} \cdot \psi \cdot \phi \\ &\geqslant \frac{1}{4} \cdot \psi \cdot \frac{1}{2p(1-p)} \end{aligned} \tag{26}$$

which is equivalent to Equation (25) if we consider the explicit form of ψ and use variance expressions. The lower bound of $\frac{1}{32}$ follows as under the assumption $0 < q < p < \frac{1}{2}$ we have

$$\mathbb{V}\text{ↄ}\diagdown(q) < \mathbb{V}\text{ↄ}\diagdown\left(\frac{p+q}{2}\right) < \mathbb{V}\text{ↄ}\diagdown(p) \tag{27}$$

□

The convexity part in the lemma follows directly from the last claim. The minimum for each q exists because of convexity and is achieved for $\delta = 0$, as at this point the first derivative vanishes (see calculations in Section 5). □

5 CONCLUSION

We have studied Bayes factors in the context of A/B testing of event rates, which is relevant to the important problem of conversion optimization.

The result can be easily extended to cover the case of unequal group sizes. Also it is possible to derive bounds for testing a fixed-size effect δ instead of zero-effect as the null hypothesis.

Finally, regarding the problem of comparing the Welch's t and Jensen-Shannon divergence we conjecture that the inequality in Theorem 2 can be improved, namely

Proposition 1 (Open problem)**.** *Let* $0 \leqslant q \leqslant p \leqslant \frac{1}{2}$*. Find the biggest constant* C *such that*

$$\mathbf{JS}(p,q) \geqslant C \cdot \frac{\mathbf{t}_{\text{Welch}}(r,p,q)^2}{r} \tag{28}$$

The current proof is based on global convexity which works under a subptimal constant $C = \frac{1}{32}$.

ACKNOWLEDGMENTS

The author thanks to Evan Miller for inspiring discussions.

REFERENCES

Derrick, B., Toher, D., and White, P. (2016). Why welch's test is type i error robust. *The Quantitative Methods in Psychology*, 12(1):30–38.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The bayes factor. *Annals of internal medicine*, 130 12:1005–13.

Jeffreys, H. (1998). *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. OUP Oxford.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Keser, A. (2017). Goodbye, t-test: new stats models for a/b testing boost accuracy, effectiveness. https://www.widerfunnel.com/goodbye-t-test-new-stats-models-for-ab-testing-boost-accuracy-effectiveness/.

Kruschke, J. K. and Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206.

Lavine, M. and Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, 53(2):119–122.

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., Rathnayake, T., Vig, S., Granger, B. E., Muller, R. P., Bonazzi, F., Gupta, H., Vats, S., Johansson, F., Pedregosa, F., Curry, M. J., Terrel, A. R., Roučka, v., Saboo, A., Fernando, I., Kulal, S., Cimrman, R., and Scopatz, A. (2017). Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103.

Morey, R. D. and Rouder, J. N. (2018). BAYESFACTOR: computation of bayes factors for common designs. r package version 0.9.12-4.2. http://CRAN.R-project.org/package=BayesFactor.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362.

APPENDIX

Code

The following Python code implements the comparison sketched on Figure 2.

```
import numpy as np
from matplotlib import pyplot as plt

H=lambda p,q:-p*np.log(q)-(1-p)*np.log(1-q)
KL=lambda p,q:H(p,q)-H(p,p)
JS=lambda p,q:(KL(p,(p+q)/2)+KL(q,(p+q)/2))/2
Welch=lambda p,q:(p-q)/np.sqrt(p*(1-p)+q*(1-q))

r_bayes= lambda p,q: np.log(10)/(2*JS(p,q))
r_freq= lambda p,q: 1.9**2*Welch(p,q)**(-2)

p= np.linspace(0,0.5,201)
delta= 0.1
q= p*(1+delta)

plt.figure(figsize=(12,6))
plt.plot(r_bayes(p,q),\
        label='bayesian (JS)')
plt.plot(r_freq(p,q),\
        label='frequentionists (Welch-t)')
plt.yscale('log')
plt.xticks(np.arange(0,201,10),\
                np.linspace(0,0.5,21).round(3))
plt.legend()
plt.ylabel('samples')
plt.xlabel('baseline p')
plt.show()
```

Proof of Lemma 3

Below we compute the formula in Claim 3

```
import sympy

p,q,d = sympy.symbols('p q d')

# define JS and Welch's t
welch = (p-q)**2/(p*(1-p)+q*(1-q))
H = -p*sympy.log(p)-(1-p)*sympy.log(1-p)
JS = H.subs(p,(p+q)/2) - 1/2*H - 1/2*H.subs(p,q)
# second derivatives
out1 = JS.subs(p,q+d).diff(d,2).subs(d,p-q)
out2 = welch.subs(p,q+d).diff(d,2).subs(d,p-q)
# ratio of second derivatives
(out1/out2).factor()
out
```

Below we include the code used to derive formulas in the proof of Claim 5

```
import sympy

p,q = sympy.symbols('p q')
U = (p**2 - 2*p*q - q**2 + 2*q)
V = (-2*p**3*q + p**3 + 3*p**2*q \
        +6*p*q**3 - 9*p*q**2 - 3*q**3 + 4*q**2)
(U*p*(1-p)-1/2*V).factor()
```

The following piece of code is used to validate the claim about the global minimum at $\delta = 0$, the end of the proof of Lemma 3

```
import sympy

welch = (p-q)**2/(p*(1-p)+q*(1-q))
H = -p*sympy.log(p)-(1-p)*sympy.log(1-p)
JS = H.subs(p,(p+q)/2) - 1/2*H - 1/2*H.subs(p,q)
#out1 = JS.subs(p,q+d).diff(d,2).subs(d,p-q)
#out2 = welch.subs(p,q+d).diff(d,2).subs(d,p-q)
#(out1/out2).factor()

derivative=(JS-1/32*welch).subs(p,q+d).diff(d,1)
derivative.subs(d,0)
```

The next piece of code evaluates the second derivative in Lemma 2

```
H = -p*sympy.log(p)-(1-p)*sympy.log(1-p)
JS = H.subs(p,(p+q)/2) - 1/2*H - 1/2*H.subs(p,q)

JS.subs(p,q+d).diff(d,2).subs(d,p-q).factor()
```

Plots

```
import numpy as np
from matplotlib import pyplot as plt
from scipy import special as sc

H = lambda p: -sc.xlogy(p,p)-sc.xlogy(1-p,1-p)
JS = lambda p,q: H((p+q)/2)-0.5*H(p)-0.5*H(q)
Welch = lambda p,q: np.divide((p-q),np.sqrt(p*(1-p)+q*(1-q)),
                out=np.zeros_like(p),where=p*(1-p)+q*(1-q)>0)

p = np.linspace(0,1,100+1)
q = np.linspace(0,1,100+1)
xx,yy=np.meshgrid(p,q)
zz_js=JS(xx,yy)
zz_welch = Welch(xx,yy)
zz = zz_welch


labels = {'Welch':'$Welch^2(p_1,p_2)$','JS':'$JS(p_1,p_2)$'}
vmax = {'Welch':2,'JS':0.5}

fig = plt.figure(figsize=(16,6))
for i,(zz,label) in enumerate(zip([zz_welch**2,zz_js],
                labels.keys())):
    ax = fig.add_subplot(1,2,1+i,projection='3d')
    surf=ax.plot_surface(xx, yy, zz,
    cmap=plt.cm.coolwarm, rstride=1, cstride=1,
    linewidth=0, vmax=vmax[label])
    ax.set_xlabel(r"$p_1$")
    ax.set_ylabel(r"$p_2$")
    ax.set_zlabel(labels[label])
    ax.zaxis.labelpad=10
    ax.invert_xaxis()
    fig.colorbar(surf, shrink=0.5, aspect=5)

plt.rc('axes', labelsize=15)
plt.show()
```