Bounds on Bayes Factors for Binomial A/B Testing

Maciej Skorski

Abstract: Bayes factors, in many cases, have been proven to bridge the classic -value based significance testing and bayesian analysis of posterior odds. This paper discusses this phenomena within the binomial A/B testing setup (applicable for example to conversion testing). It is shown that the bayes factor is controlled by the Jensen-Shannon divergence of success ratios in two tested groups, which can be further bounded by the Welch statistic. As a result, bayesian sample bounds almost match frequentionist’s sample bounds. The link between Jensen-Shannon divergence and Welch’s test as well as the derivation are an elegant application of tools from information geometry.

1 Introduction

1.1 Motivation

A/B testing A/B testing is the technique of collecting data from two parallel experiments and comparing them by probabilistic inference. A particularly important case is assessing which of two success-counting experiments achieves a higher success rate. This naturally applies to evaluating conversion rates on two naturally applies to evaluating conversion rates on two experiments: A/B testing counts proportions Suppose that empirical data $\mathcal{D}$ has $r_i = r$ runs and $r \cdot \bar{p}_i$ successes in the $i$-th experiment, $i = 1, 2$. Under the binomial counting model, the data likelihood under a hypothesis $H$ equals

$$\Pr(\mathcal{D}|H) = \int_0^1 p_i^{r_i} (1 - p_i)^{(1-r_i)} \, d\mathbb{P}_H(p_1, p_2) \quad (2)$$

where the prior distribution $\mathbb{P}_H(\cdot, \cdot)$ reflects what is assumed prior to seeing data (and what will be tested); one can for example choose $\{p_1 = p_2 = 0.1\%\}$ for $H = H_0$ and $\{p_1 \neq p_2\}$ for $H_a$ uniformly over all valid values of $p_1, p_2$, but in practice more informative priors are used because some configurations of values are unrealistic (e.g. extremely low or high conversion). The corresponding factor $K$ can be computed for example by the R package BayesFactor [Morey and Rouder, 2018].

Problem: Bayesian A/B testing power Estimates, neither frequentist nor bayesian, will not be conclusive without sufficiently many samples. Frequentists 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. Under the binomial A/B model, we will answer the following questions

- when, given data, a bayesian hypothesis on zero effect may be rejected ($K \ll 1$ for some $K_0$)?
- what is the relation to the classical t-test?
This will allow us to understand data limitations when doing bayesian inference, and relate them to widely-spread frequentionist rule of thumbs.

### 1.2 Related Works and Contribution

Our problem, as stated, is a question about maximizing minimal bayes factor. It is known that for certain problems bayes factors can be related to frequentionist’s p-values [Edwards et al., 1963][Kass and Raftery, 1995][Goodman, 1999] and thus bridges the Bayesian and frequentionist world (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) bounding the Bayes factor for binomial distributions (b) discussion of sample bounds for binomial A/B testing in relation to the frequentionist approach.

#### 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 dataset statistic. This statistic turns out to be the Jensen-Shannon divergence, well-known in information theory. It is in turn bounded by the Welch’s t-statistic.

**Theorem 1** (Bayes Factors for Binomial Testing). Consider two independent experiments, each with \( r \) independent trials with unknown success probabilities \( p_1 \) and \( p_2 \) respectively. Let observed data \( D \) has \( r \cdot \overline{p}_i \) successes and \( r \cdot (1 - \overline{p}_i) \) failures for group \( i \). Then

\[
\max_{H_0: \{p_1 = p_2\}} \min_{H_a} \frac{\Pr[H_0 | D]}{\Pr[H_a | D]} = e^{-2r \cdot JS(\overline{p}_1, \overline{p}_2)}
\]

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

Moreover, the Jensen-Shannon divergence is bounded by the Welch’s t-statistic (on \( D \))

\[
JS(\overline{p}_1, \overline{p}_2) \geq \frac{t_{Welch}(\overline{p}_1, \overline{p}_2)^2}{4r}
\]

so that we can bound

\[
\max_{H_0: \{p_1 = p_2\}} \min_{H_a} \frac{\Pr[H_0 | D]}{\Pr[H_a | D]} \leq e^{-t_{Welch}(\overline{p}_1, \overline{p}_2)^2/2}
\]

#### Remark 1 (Most favorable hypotheses). Note that

- Maximally favorable alternative (\( H_a \) which maximizes \( \Pr[D | H_a] \)) is \( p_1 = \theta_1 \) and \( q_1 = \theta_2 \)
- Maximally favorable null of the form \( p_1 = q_1 \) is \( p_1 = q_1 = \frac{\overline{p}_1 + \overline{p}_2}{2} \)

If null is of the form \( p = q = \theta_0 \) then the bound becomes \( e^{-t_{KL}(\overline{p}_1, \theta_0)^{-2}} = -t_{KL}(\overline{p}_2, \theta_0)^{-2} \).

**Corollary 1** (Bayesian Sample Bound). To confirm the non-zero effect (\( p_1 \neq p_2 \)) the number of samples for the bayesian method should be

\[
r \gg \frac{1}{2JS(p_1, p_2)}
\]

Under the frequentionist method the rule of thumb is \( t_{Welch} \gg 1 \), which gives (see Section 2)

\[
r \gg \frac{2(p_1(1 - p_1) + p_2(1 - p_2))}{(p_1 - p_2)^2}
\]

Note that both formulas needs assumptions on locations of the parameters. In particular, testing smaller effects or effects with higher variance require more samples.

Bounds [Equation (6)] and [Equation (7)] are close to each other by a constant factor (a different small factor is necessary to make the bound small in both the bayesian credibility and p-value sense). The difference (under the normalized constant) is illustrated on Figure 1 for the case when one wants to test a relative uplift of 10%.

Since high values of \( t_{Welch} \) means small p-values, we conclude that the frequentionist p-values bounds the bayes factor and indeed, are evidence against a null-hypothesis in the well-defined bayesian sense. However, because of the scaling \( t_{Welch} \rightarrow e^{-t_{Welch}^2/2} \), this is true for p-values much lower than the standard threshold of 0.05. In some sense, the bayesian approach is more conservative and less reluctant to reject than frequentionist tests; this conclusion is shared with other works [Goodman, 1999].
2 Preliminaries

Entropy, Divergence The binary cross-entropy of \( p \) and \( q \) is defined by

\[
H(p, q) = -p \log(1 - p) - (1 - p) \log(1 - q)
\]

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

\[
KL(p, q) = H(p, q) - H(p)
\]

and the Jensen-Shannon divergence \([\text{Lin}, 1991]\) is defined as

\[
JS(p, q) = H(p, q) - \frac{1}{2} H(p) - \frac{1}{2} H(q)
\]

(always positive because the entropy is concave).

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 \( x \to H(p, x) \) is convex in \( x \).

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

\[-y_1 p \log x_1 - y_2 p \log x_2 \geq -p \log(y_1 x_1 + y_2 x_2)\]

for any \( x_1, x_2 \) and any \( y_1, y_2 \geq 0, y_1 + y_2 = 1 \). Replacing \( x_i \) by \( 1 - x_i \) and \( p \) by \( 1 - p \) in the above inequality gives us also

\[-y_1 (1 - p) \log(1 - x_1) - y_2 (1 - p) \log(1 - x_2) \geq - (1 - p) \log(y_1 (1 - x_1) + y_2 (1 - x_2))
\]

\[= - (1 - p) \log(1 - y_1 x_1 - y_2 x_2)\]

Adding side by side yields

\[y_1 H(p, x_1) + y_2 H(p, x_2) \geq y_1 H(p, x_1) + y_2 H(p, x_2)\]

which finishes the proof. This argument works for multivariate case, when \( p, x \) are probability vectors. \( \square \)

**Lemma 2** (Quadratic bounds on KL/cross-entropy). For any \( p \) it holds that

\[KL(p, x) \geq \left( \frac{1}{p} + \frac{1}{1 - p} \right) \cdot (x - p)^2\]

**Proof.** We will prove a general version. Let \( (p_i), \) and \( (x_i) \), be probability vectors of the same length. By the elementary inequality

\[\log(1 + u) \geq u - \frac{1}{2} u^2\]

we obtain

\[-\log(x_i/p_i) = -\log(1 - (p_i - x_i)/p_i) \geq \frac{p_i - x_i}{p_i} + \frac{1}{2} \left( \frac{p_i - x_i}{p_i} \right)^2\]

multiplying both sides by \( p_i \) and adding inequalities side by side we obtain

\[-\sum p_i \log(x_i/p_i) \geq -\sum (x_i - p_i) + \sum \frac{(p_i - x_i)^2}{2p_i}\]

which means \( KL(x, p) \sum \frac{(p_i - x_i)^2}{2p_i} \). Our lemma follows by specializing to the vectors \((p, 1 - p)\) and \((x, 1 - x)\). \( \square \)

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

\[t_{\text{Welch}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\]

where \( s_i \) are sample variances and \( \mu_i \) 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

\[t_{\text{Welch}}(r, \theta_1, \theta_2) = r^{-\frac{1}{2}} \cdot \frac{\theta_1 - \theta_2}{\sqrt{\theta_1 (1 - \theta_1) + \theta_2 (1 - \theta_2)}}\]

3 Proof

We change the notation slightly, unknown success rates will be \( p \) and \( q \), and corresponding successes \( r \cdot \theta_1, r \cdot \theta_2 \).

**Alternatives** Maximizing over all possible priors \( P_a \) over pairs \((p, q)\) we get

\[\max_{\theta_1} \Pr[D/H_a] =
\]

\[c \cdot \max_{\theta_1} \int_{[0, 1]^2} e^{-H(\theta_1, p) - H(\theta_1, q)} \cdot [P_a(p, q)] d(p, q)\]
where \( c = \frac{1}{b(x_{1} + 1, (1-\theta_1)^{-1}) b(x_{2} + 1, (1-\theta_2)^{-1})} \) is a normalizing constant, which equals
\[
\max_{p} \Pr[D|H_0] = c \cdot e^{-rH(\theta_1) - rH(\theta_2)}
\]
achieved for \( P_a \) being a unit mass at \((p,q) = (\theta_1, \theta_2)\).

**Null** Let \( H_0 \) states that the baseline is \( p \) and the effect is 0. Then we obtain
\[
\Pr[D|H_0] = c \cdot e^{-rH(\theta_1) - rH(\theta_2)}
\]
with the same normalizing constant \( c \).

**Bayes factor** If none of two hypothesis is a priori preferred, that is when \( \Pr[H_0] = \Pr[H_a] \), then the Bayes factor equals the likelihood ratio (by Bayes theorem)
\begin{equation}
\Pr[H_0|D] = \frac{\Pr[D|H_0]}{\Pr[D|H_a]} = \frac{\Pr[D|H_0]}{\Pr[D|H_a]}
\end{equation}
In turn the likelihood ratio (in favor of \( H_0 \)) equals
\[
\min_{H_0} \Pr[D|H_0] = e^{-r[H(\theta_1,p) + H(\theta_2,p) - H(\theta_1) - H(\theta_2)]}
\]
(the normalizing constant \( c \) cancels).

Using the relation between the KL divergence and cross-entropy we obtain
\[
\min_{H_0} \Pr[D|H_0] = e^{-rKL(\theta_1,p) - rKL(\theta_2,p)}
\]
We will use the following observation

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

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

We can now bound Equation (24) as
\[
\min_{H_0} \Pr[D|H_0] \leq e^{-2rJS(\theta^*)}
\]
This proves the first part of Theorem 1.

Connecting t-statistic and bayes factor exponent
Recall that by Claim 1 under t-test we have
\[
T \approx r^2 \cdot (\theta_1 - \theta_2) \cdot (\theta_1(1 - \theta_1) + \theta_2(1 - \theta_2))
\]
It remains to connect \( |\theta_1 - \theta_2| \) and \( JS(\theta_1, \theta_2) \). By Lemma 2 we have the following refinement of Pinsker’s inequality

**Claim 3.** We have \( KL(\theta,p) \geq \frac{(\theta-p)^2}{2\theta(1-\theta)} \).

Using \( 2JS(\theta_1, \theta_2) = KL(\theta_1, \theta^*) + KL(\theta_2, \theta^*) \), the inequality from Claim 3 and the Welch’s formula in Equation (26) we obtain

**Claim 4.** We have
\[
JS(\theta_1, \theta_2) \geq \frac{t_{Welch}(r, \theta_1, \theta_2)^2}{4r}
\]

**Proof.** Claim 3 implies
\[
2JS(\theta_1, \theta_2) \geq (\theta_1 - \theta_2)^2 \cdot \left( \frac{1}{2\theta_1(1 - \theta_1)} + \frac{1}{2\theta_1(1 - \theta_1)} \right)
\]
we recognize the Welch’s statistic and write
\[
2JS(\theta_1, \theta_2) \geq \frac{t_{Welch}(r, \theta_1, \theta_2)^2}{2r}
\]

Combining Equation (25) and Equation (27) implies the second part of the theorem.

**Acknowledgments**

The author thanks to Evan Miller for inspiring discussions.

**REFERENCES**

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). Bayesian analysis of stochastic processes. *Journal of the American Statistical Association*, 90(430):773–795.

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

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