Bayesian *a posteriori* performance estimation for speech recognition and psychophysical tasks

Stacy L. Tantum, Leslie M. Collins,† and Chandra S. Throckmorton

Department of Electrical and Computer Engineering, Duke University
129 Hudson Hall, Box 90291, Durham, North Carolina 27708-0291
stacy.tantum@duke.edu, lcollins@ee.duke.edu, chandra.throckmorton@duke.edu

†*Author to whom correspondence should be addressed.*

Abstract: Null hypothesis significance testing remains the *de facto* method to quantitatively compare experimental data measured under various conditions, despite the fact that it has been questioned across a wide variety of disciplines over the past half-century. This letter presents a Bayesian approach for performance estimation from which Bayesian methods for performance comparisons then naturally follow. The Bayesian approach offers several advantages, including robustness to data partitioning effects that could complicate the accurate reporting of research results for experiments in which the number of data samples that can be collected is limited.

PACS numbers: 43.10.Pr,43.66.Yw

1. Introduction

Many speech recognition and psychophysical experiments are administered with the goal of comparing subject performance under two or more conditions. Tests of statistical significance, such as t-Tests or analysis of variance (ANOVA), typically are applied after the experimental data are collected to determine if the differences in performance measured by the experiment are meaningful (i.e., a null hypothesis test that the means are equal). The use of null hypothesis significance testing (NHST) remains the *de facto* method to quantitatively compare experimental data measured under various conditions, despite the fact that it has been questioned across a wide variety of disciplines over the past half-century (e.g. Selvin, 1957; Rozeboom, 1960; Lecoutre *et al.*, 2001). The principle concerns regarding NHST revolve around the interpretation (or mis-interpretation) of p-values, the bias toward not rejecting the null hypothesis, the common assumption that not rejecting the null hypothesis is the same as accepting the null hyothesis, and the potential for an insignificant difference to be deemed significant due to a large sample (effect size). To this list of concerns, we add the potential dependence of the outcome of the significance test on the order in which the data are collected.

Speech recognition tests typically compare signal processing algorithms such as speech processing for cochlear implants or adaptive noise cancellation. Similarly, psychophysical tests typically compare acoustic stimuli (or electric stimuli in the case of cochlear implants) generated under conditions designed to understand and/or explain the physiology of the auditory system or identify mechanisms that affect subject performance on listening tests. This process typically entails presenting some number of trials (K) of a forced choice test to arrive at a single estimate of the probability

of a correct response, defined as the number of correct responses divided by the total number of trials. This procedure is then repeated a number of times (T) to generate a total of T estimates of the probability of correct response. The T estimates of the probability of correct response for each of the conditions to be compared then become the samples of probability of correct response required for the chosen test of statistical significance.

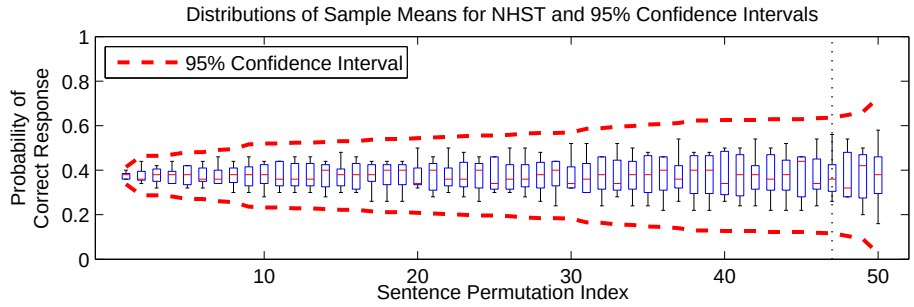
Figure 1a illustrates the variability in $T = 5$ samples of probability of correct response depending on the order in which 50 sentences, each with 5 key words, are presented in groups of 10 ($K = 50$ words per group), assuming the order of presentation of the sentences does not affect whether or not each word is correctly identified. The order of words within each sentence is maintained; it is only the order of the sentences that is randomly permuted. The box plots for 50 random permutations of the sentence order are presented in ascending order of the standard deviation of the T samples of probability correct. The actual sentence order under which this data was collected is permutation 47, denoted by the vertical dashed line. In contrast, Fig. 1b demonstrates the stability of a Bayesian *a posteriori* probability estimate of percent discrimination for the same 50 random permutations of sentence order. (The distinction and relationship between probability of correct response and probability of discrimination will be discussed in Sec. 2.) Intuitively, the order in which the sentences are presented should not affect the estimate of the probability of correct response. A standard NHST paradigm fails to meet this intuitive expectation, yet the Bayesian technique is consistent with this expectation. Similarly, applying NHST (a t-Test) to the measured data indicates the difference in the means is not statistically significant ($p = 0.10$), which is a surprising result given the apparent separation between the data measured under the two conditions shown by the box plots in Fig. 1c. The Bayesian approach, however, results in the *a posteriori* probability density functions (pdfs) shown in Fig. 1d. The probability that percent discrimination for condition B is smaller than percent discrimination for condition A can be calculated from the *a posteriori* pdfs and the result, $P(p_B < p_A) = 0.9994$, is more intuitively consistent with the measured data.

These results from an actual psychophysical experiment are consistent with previously published work demonstrating the potential benefits of a Bayesian approach over NHST (e.g. Selvin, 1957; Rozeboom, 1960; Lecoutre *et al.*, 2001). The Bayesian approach presented here is an alternative to NHST which mitigates the potential for the outcome of the data analysis to depend on the random order in which the data were collected. This benefit may be particularly useful for situations in which the number of data points that can be collected is limited, such as when the eligible subject pool is small or the individual experiments are time-consuming, which is not uncommon for psychophysical experiments. The remainder of this letter presents a Bayesian *a posteriori* distribution for subject performance on M-ary speech recognition and psychophysical tasks, including a description of the relationship of the Bayesian approach to statistical significance tests. A Matlab toolbox to assist others who would like to apply this Bayesian approach for performance estimation is available at: `URL provided upon publication`.

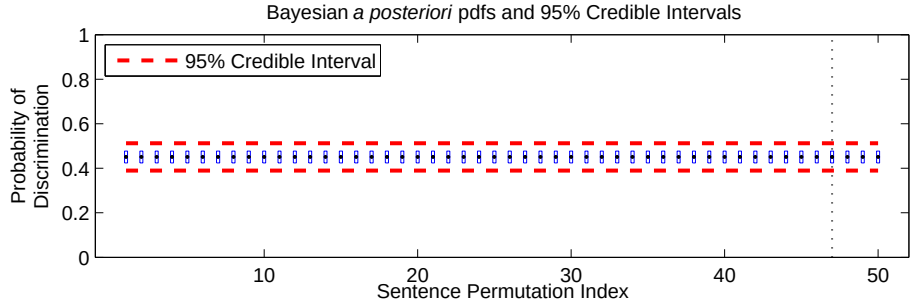
2. Bayesian performance estimation

The Bayesian *a posteriori* approach for performance estimation applied here was first developed by Morrison for forced binary choice tests (Morrison, 1978) and later extended to forced M-ary choice tests (Morrison and Brockway, 1979). This work is extended here to perform Bayesian analysis and comparisons of experimental data using the previously developed Bayesian performance estimates.

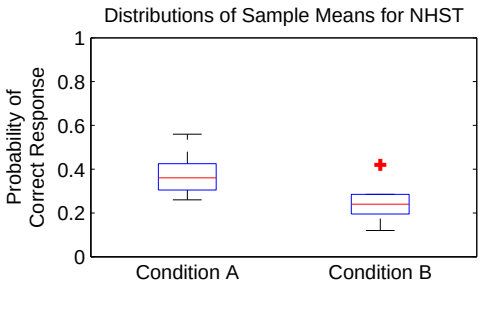
A key notion central to the development of this approach is the distinction between the subject *knowing* the correct response and the subject *providing* the correct response, as a correct response can be given by chance. For an M-ary forced choice



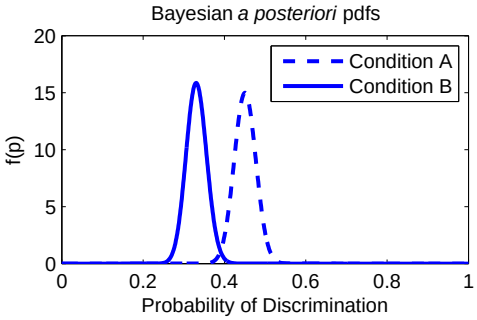
(a) NHST probability of correct response.



(b) Bayesian probability of discrimination.



(c) Measured experimental data.



(d) Bayesian *a posteriori* pdfs for the measured data.

Fig. 1. a) Box plots resulting from 50 random permutations of the speech recognition test sentence presentation. b) Mean and standard deviation of the Bayesian *a posteriori* pdfs for the same 50 random permutations of the speech recognition test sentence presentation. c) Box plots of the measured experimental data. d) Bayesian *a posteriori* pdfs for the measured experimental data.

test, the probability of providing a correct response, c, and the probability of knowing the correct response due to the ability to discriminate among the M choices, p, are related by (Morrison and Brockway, 1979)

$$c = p + (1-p)\frac{1}{M}. \tag{1}$$

Theoretically, the probability of a correct response, c, is bounded between $\frac{1}{M}$ and 1, while the probability of discrimination, p, is bounded between 0 and 1. In the limit as $M \to \infty$, such as for open set speech recognition tests, $c \to p$. The goal of performance estimation within a Bayesian framework is then finding the posterior for the probability that the subject *knows* the correct response, p, given the number of correct responses

provided, r, out of k trials, $f(p|r,k)$.

Assuming the distribution of p is $f(p)$, it follows that the distribution of c is (Morrison and Brockway, 1979)

$$g(c) = \frac{M}{M-1} f\left(\frac{Mc-1}{M}\right). \tag{2}$$

The distribution for the number of correct responses provided given the number of trials, $p(r|k)$, is then given by a mixed binomial distribution (Morrison and Brockway, 1979)

$$\begin{aligned} p(r|k) &= \binom{k}{r} \int_{1/M}^{1} c^r (1-c)^{k-r} g(c) dc \\ &= \binom{k}{r} \int_{1/M}^{1} c^r (1-c)^{k-r} \frac{M}{M-1} f\left(\frac{Mc-1}{M}\right) dc. \end{aligned} \tag{3}$$

If it is assumed that p follows a beta distribution with parameters x and y,

$$f(p) = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)} p^{x-1}(1-p)^{y-1} = B(x,y), \tag{4}$$

then Morrison and Brockway (Morrison and Brockway, 1979) give the conditional probability of the number of correct responses, r, given the number of trials, k, when it is assumed that $f(p)$ follows a beta distribution with parameters x and y,

$$\begin{aligned} p_{MBB}(r|k,x,y) = \\ \binom{k}{r} \frac{(M-1)^{k-r}}{M^k B(x,y)} \int_0^1 p^{x-1}(1-p)^{k-r+y-1}[1+(M-1)p]^r dp, \end{aligned}$$

$$= \binom{k}{r} \frac{1}{M^k B(x,y)} \sum_{j=0}^{r} \binom{r}{j} (M-1)^{k-r+j} B(x+j, k-r+y),$$

$$r = 0, 1, \ldots, k, \tag{5}$$

where the term $[1+(M-1)p]^r$ was replaced by its binomial expansion.

The modified beta binomial (MBB) distribution (5) models the distribution of the number of correct responses, r, but the desired quantity is the probability of discrimination, or the probability that the subject *knows* the correct response, p, given the number of correct responses provided, r, out of k trials, $f(p|r,k)$. Applying Bayes' Rule,

$$f(p|r,k) = \frac{f(r|k,p)f(p)}{\int_p f(r|k,p)f(p)}, \tag{6}$$

where $f(p)$ is the prior on the probability that the subject knows the correct response. Recognizing the (assumed) beta distribution for the prior $f(p) = B(x,y)$ is characterized by the parameters x and y, the above relationship can be re-written as

$$f(p|r,k) = \frac{f(r|k,x,y)B(x,y)}{\int_{x,y} f(r|k,x,y)B(x,y)}, \tag{7}$$

where the likelihood $f(r|k,x,y)$ can be found via the MBB distribution (5) and the prior $f(p) = B(x,y)$ is either assumed (e.g., uniform prior, Jeffreys prior) or estimated (e.g., evidence approximation). In the example presented in Fig. 1, $f(p)$ was estimated using the evidence approximation.

The posterior distribution for the probability that the subject *knows* the correct response (Eq. 5) is key to applying the Bayesian performance estimates to assess statistical significance. Bayesian credible intervals can be calculated from this distribution. In addition, these distributions allow the relative values of the probability of discrimination under different conditions to be compared in a probabilistic sense. Both of these approaches to assessing statistical significance from the Bayesian posterior distributions are discussed in the following section.

An additional benefit of estimating the probability of discrimination p, rather than the probability or correct response c, is the probability of discrimination is independent of the number of possible choices, M, whereas the probability of correct response is affected by the number of possible choices because a correct response can be obtained purely by chance. Since the probability of discrimination is independent of the number of possible choices, performance can be compared across experiments in a meaningful way even if the number of possible choices is not consistent across experiments. This allows, for example, a meaningful comparison of vowel and consonant recognition performance which typically consist of differing numbers of tokens.

3. Applying Bayesian performance estimates

Given the posterior for the probability that the subject knows the correct response, $f(p|r,k)$, a Bayesian credible interval, an interval within which the true parameter value falls with a stated probability, can be calculated. Given posteriors for the probability that the subject knows the correct response under varying experimental conditions, Bayesian tests to compare the two conditions can be performed to assess the statistical significance of performance differences.

3.1. Bayesian credible interval

A Bayesian credible interval, the conceptual analog to a confidence interval often calculated within the context of null hypothesis significance tests, can be calculated from the posterior for the probability that the subject knows the correct response, $f(p|r,k)$. Although discussed as conceptual analogs of each other, Bayesian credible intervals and confidence intervals are very different. An n% confidence interval is one random interval identified by a method that will produce an interval that contains the true parameter value n% of the time. In contrast, an n% Bayesian credible interval is an interval for which the true parameter value falls within the stated interval with probability $n/100$.

There are multiple approaches to calculating an n% Bayesian credible interval. The approach taken here is to find the central interval that captures n% of the probability, so probability of $\frac{1}{2}(n/100)$ lies both above and below the interval. Other approaches include determining the smallest interval over which the desired probability is captured or choosing the interval such that the mean of the distribution is the central point of the interval.

3.2. Bayesian comparisons

The approach of probabilistically comparing posterior distributions to assess the similarity, or lack thereof, of two parameter estimates is the preferred approach within the context of Bayesian probability and statistics, (e.g. Jaynes, 1976). This may be considered to be a conceptual analog to standard NHSTs but, again, the Bayesian probabilistic comparison is very different from the NHST. A p-value calculated as part of a NHST provides the probability that if the experiment were (hypothetically) completed numerous times, p% of the experiments would result in a declaration that the difference in the parameters is significant when the difference truly is not significant. In contrast, the Bayesian probabilistic comparison provides the probability that one parameter is greater than (or less than) the other parameter.

The probability that the probability of discrimination under condition A, p_A, is smaller than the probability of discrimination under condition B, p_B, can be calculated from their respective posteriors, $f(p_A|r,k)$ and $f(p_B|r,k)$:

$$P(p_A < p_B) = \int_{-\infty}^{\infty} \int_{p_A}^{\infty} f(p_A|r,k) f(p_B|r,k) dp_B dp_A. \quad (8)$$

4. Conclusions

Numerous researchers have noted potential issues with null hypothesis significance tests (NHSTs) (e.g. Selvin, 1957; Rozeboom, 1960; Lecoutre *et al.*, 2001). We affirm these concerns with the observation that results of NHSTs may depend on the random ordering of the experimental trials, although the trial order should be irrelevant with respect to the significance test. Ideally, the significance test results would be independent of the order in which the responses are obtained when ordering of the trials is not a relevant component of the experiment design. The Bayesian approach to *a posteriori* performance estimation presented here achieves this ideal. In addition, the variance of the *a posteriori* distribution naturally reflects the confidence in the estimate given the measured data, and a Bayesian credible interval can be computed from the posterior distribution. Testing the significance of the difference in performance between two tests performed under different conditions also naturally follows from the posterior distributions; the probability that performance under one condition is greater (or less) than performance under the other condition can be calculated. Finally, estimating the probability that the subject *knows*, rather than *provides*, the correct response allows for performance comparisons across experimental conditions whose number of choices, M, are different because the effect of obtaining a correct response by chance is removed from the parameter estimate. Within the context of speech recognition tasks, this allows, for example, comparing in a meaningful way performance on vowel and consonant recognition tasks which typically consist of differing numbers of tokens.

A Matlab toolbox to assist others who would like to apply this Bayesian approach for performance estimation is available at: `URL provided upon publication`.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number 1-R01-DC007994-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors gratefully acknowledge Sara I. Duran, who provided the psychophysical data that formed the basis of the simulations presented in Figure 1.

References

Jaynes, E. T. (**1976**). *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume II, chapter Confidence Intervals Vs Bayesian Intervals, 175–257 (D. Reidel Publishing Company).

Lecoutre, B., Lecoutre, M.-P., and Poitevineau, J. (**2001**). "Uses, abuses, and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable?", International Statistical Review **69**, 399–417.

Morrison, D. G. (**1978**). "A probability model for forced binary choices", The American Statistician **32**, 23–35.

Morrison, D. G. and Brockway, G. (**1979**). "A modified beta binomial model with applications to multiple choice and taste tests", Psychometrika **44**, 427–442.

Rozeboom, W. (**1960**). "The fallacy of the null-hypothesis significance test", Psychological Bulletin **57**, 416–428.

Selvin, H. (**1957**). "A critique of tests of significance in survey research", American Sociological Review **22**, 519–527.