Page 1: Thank you Tianhao for introducing the balanced accuracy to me and encouraging me to construct confidence interval for balanced accuracy using frequentist method. Without Tianhao’s introduction and guidance, I would not make research on this topic and find the interesting result. I am very excited to have this chance today to present the topic about balanced accuracy, its variance and its confidence interval.

Page 2: What is the balanced accuracy? I firstly use a binary classification case to introduce the concept and then generalize the balanced accuracy to multiclass classification case. Consider a 2x2 table that gives the number of subjects with positive and negative test predictions as shown in predict class, and with actual positive or negative numbers as shown in true class. As we all know, the sensitivity equals to true positive number divided by the sum of true positive number and false negative number; the specificity equals to true negative number divided by the sum of true negative number and false positive number; overall accuracy equals to true positive number plus true negative number and then divided by total number. When the data is balanced, that is, a plus c equals to b plus d, we can use sensitivity, specificity and overall accuracy to evaluate the performance of a diagnostic tool, for example, MRI or CT. However, when the data is unbalanced, overall accuracy would not work well and produce the biased result. A better approach is to use balanced accuracy. In the binary case, balanced accuracy equals to sen plus spe and then divided by 2. Overall accuracy equals to balanced accuracy only when data is balanced. When data is unbalanced, we need to use balanced accuracy.

Page 3: In this slide, I use a 3x3 cross table to demonstrate how to calculate balanced accuracy for multiclass classification case. Balanced accuracy is average accuracy over all the classes. It derives all its information from the diagonal elements and the column sums. The mathematical formula is shown in the left. For the three-class classification case, we can get the balanced accuracy easily as shown in the right formula.

Page 4: Balanced accuracy is a popular measurement in pattern classification and machine learning but not well discussed in the field of statistics. Statisticians always try to estimate the variance of a statistic and then construct confidence interval to measure the uncertainty. In the field of machine learning, variance and confidence interval for balanced accuracy are rarely discussed. Only Bayesian credible interval is proposed for balanced accuracy and implemented in Matlab and Python platform. R function is written to estimate balanced accuracy but there is no R function to calculate confidence interval.

Page 5: Is the Bayesian posterior method acceptable to measure the uncertainty for balanced accuracy? The answer is obvious. From the perspective of science and statistics, the Bayesian method is a great solution to measure uncertainty for balanced accuracy using the probabilistic logic. But it may face the challenges from regulatory on the prior choice because the prior choice is always subjective using Bayesian method. Can we construct the confidence interval for balanced accuracy using frequentist method? Before construct the confidence interval, can we estimate the variance of balanced accuracy?

Page 6: In the past several days, with Tianhao’s help and encouragement, I focused on calculating the interval for binary balanced accuracy. For the multiclass classification, I will discuss it in the end. The variance of balanced accuracy is difficult to estimate because the correlation between sensitivity and specificity should be considered. Thus, bootstrap method seems to be a feasible solution within frequentist school.

Page 7: Bootstrap is a great statistical idea using the sample data as a population from which repeated sample are drawn. Thus, the population is to the sample as the sample is to the bootstrap samples. We can use this general approach to construct interval based on building a sampling distribution for balanced accuracy by resampling from the data at hand.

Page 8: I write a R function to estimate interval for balanced accuracy using boot R function. Biased-corrected, accelerated percentile interval is preferable among the several bootstrapping methods as it does not need to meet the requirements of other methods.

Page 9: The bootstrap method is very easy to use, but it is a compromised method and produces different results between the different simulations. We still need to construct frequentist confidence interval. As we know, the variance of balanced accuracy is difficult to estimate and then becomes an obstacle for constructing confidence interval. Is there any index similar with balanced accuracy that is well discussed in the field of statistics?

Page 10: Balanced accuracy has the very similar mathematical form with both balanced error rate and Youden index. Using very simple calculation, we can know that the variance of balanced accuracy equals to the variance of BER and equals to the variance of Youden index divided by 4. So if we know the variance of Youden index or balanced error rate, we can estimate the variance of balanced accuracy easily.

Page 11: There is no literature on variance of balanced error rate, and based on literatures we can know the variance of Youden index. As the correlation between sensitivity and specificity is not zero and should be considered in estimating variance of Youden index. So we use the variance of Youden index estimated by Chen Fangyao in 2013 to estimate the variance of balanced accuracy, and use the Wald method to construct confidence interval.

Page 12: Based on the formula of variance of balanced accuracy, I write a R function to estimate the frequentist confidence interval for balanced accuracy.

Page 13: I use a hypothetical data example to demonstrate the result of bootstrap interval, confidence interval and credible interval. Suppose that there is a diagnostic tool used to predict whether patient has disease or not. Among the 160 patients who actually have disease, this diagnostic tool predicts 80 patients as positive cases and 80 cases as negative cases; among the 120 who actually have no disease, this diagnostic tool predicts 40 patients as positive cases and 80 cases as negative cases. The balanced accuracy is 0.58333. The BCa bootstrap interval is from 0.5237 to 0.6402, and the frequentist confidence interval is from 0.516654 to 0.6500127.

Page 14: Henry Carrillo uses Fourier transforms of class-specific performance distributions to obtain a posterior distribution that can be used to evaluate and compare competing classifiers. To facilitate its use, they provide an open-source MATLAB implementation. His colleague Cheng Soon Ong re-write matlab codes in Python language and Kay H Brodersen write a R package “micp” but it is not available in the website. I wrote a email to contact with Brodersen but have not received the responses. I re-write matlab codes in R language but it produce the wrong results.

Page 15: Bayesian method works for both binary and multiple classification cases. We provide frequentist confidence interval of balanced accuracy for using the variance of youden index. But our method only works for binary classification and do not consider the multiple classification case. When estimating variance of balanced accuracy for multiple classes, the youden index can not be used because youden index only fits for binary class. Can we use Cleophas method mentioned in Chen’s paper to estimate variance of balanced accuracy for multiple classes? We prove that the following formula 1 is not equal to formula 2 and formula 3 by using R programming. So Cleophas method can not be used to estimate variance of balanced accuracy for multiple classes. In future research, we will focus on estimating variance of balanced accuracy for multiple classes.

Page 16: As Canadian Philosopher – Ian Hacking said, [There is not just one way to build a house, or even to grow tomatoes. We should not expect something as motley as the growth of knowledge ***(statistics)*** to be strapped to one methodology.](https://www.azquotes.com/quote/905210) There is no unified method to estimate parameter and construct interval in the field of statistics. In fact, there are several competing schools to do the same thing. Bayesian method provides probabilistic interval for statistics parameter but the prior choice is always subjective. Frequentist confidence interval is acceptable by the regulatory but the confidence is not probabilistic. Fisherian fiducial interval aims to estimate interval in frequentist school in the style of a Bayesian posterior, that is, enjoy the benefit of Bayesian probabilistic taste without choosing prior. For further information on fisherian fiducial interval and related topics, we can read the papers on “confidence distribution” and “inferential model” raised by Chinese statisticians.