The strong contour line relationship between two-sided p values of simple RCT two sample t-tests and BIC based Bayes factor $P_{BIC}(D|H0)$ outputs for fixed sample sizes

John Martin

3/18/2020

Executive Summary

For simple Randomised Control Trial (RCT) analysis there is a strong non-linear relationship between the two-sided p values of two sample t-tests and the corresponding BIC based Bayes factor output $P_{BIC}(D|H0)$. In particular, the spearman rank correlation coefficient $cor_{spearman,RCT}(p_{value}, P_{BIC}(D|H0)) \approx 1$ and a strong contour line relationship dependent on sample size exists between the two statistics that allows psuedo ROC curve interpretation and cutoff estimation. In practice, as $P(H0)$ varies ($0 \leq P(H0) \leq 1$) the position of the distribution of observed data on a given sample size contour line varies according to the statistical power (ie. effect size) and $P(H0)$ probability. Proceeding to posterior odds calculations and derivation of $P_{BIC}(H0|D)$, $P_{BIC}(H1|D)$ probabilities, the above BIC based results have significant relevance under the common prior odds assumption $P(H0)/P(H1) \sim 1$.

Introduction

There is continuing disagreement about the correlation between

- p_{value} the two sample t-test probability of the observed data given the Null Hypothesis and
- $P(H0|D)$ the Bayesian posterior probability of the Null Hypothesis given the observed data.

In this paper, the correlation under repeated sampling for fixed sample sizes is re-examined empirically between Bayes Factor estimates and p-values using (i) BIC based estimates for Bayes Factor ouputs $P_{BIC}(D|H0)$ in relative posterior probability calculations, and explicitly (ii) the Spearman Rank correlation calculation (to allow for a non-linear relationship).

Firstly, the evidence for [1] and against [2], a weak (0.397) correlation coefficient between p_{value} and $P(H0|D)$ is briefly described. Then an existing BIC based method for Bayes Factor calculation is described and used to produce evidence for a strong non-linear relationship between $P_{BIC}(D|H0)$ and two sample t-test p-values under repeated sampling for a fixed sample size. Finally, estimates of pseudo ROC curve cutoff points are attempted for the $P_{BIC}(D|H0)$ (and normalised $P_{BIC}(D|H0)$) contour line for fixed sample sizes to provoke discussion about optimality or otherwise of $\alpha = 0.05$ p_{values}.

Under the commonly used prior odds assumption $\frac{P(H0)}{P(H1)} \approx 1$, the above finding has significant relevance to the posterior probability $P(H0|D)$, p-value correlation for fixed sample sizes.

Previously, Trafimow and Rice [1] raised the issue that frequentist based Null Hypothesis Significance Test (NHST) p-values are not strongly correlated (ie. 0.397) with Bayesian estimates of $P(H0|D)$ based on Bayes formula updating. The simulation used randomly drawn values of $P(D|H0) \rightarrow p_{value}$, $P(H0)$, and $P(D|H1)$ as input to Bayes Formula calculations.

As a critique of the relevance of the evidence presented in [1], Lakens [2]

1. Replicated (with more accuracy) the results from [1], which Lakens identified as employing Pearson correlation calculations (ie. 0.37) of the relationship between p-values and posterior $P(H0|D)$ estimates using Bayes Rule updating under randomly drawn values of $P(D|H0)$, $P(H0)$, and $P(D|H1)$.

2. Then argued that such an approach did not replicate the situation in Randomised Control Trial (RCT) analysis where the statistical power $P(D|H1)$ for a known H1 value is determined by the sample size and coefficient of variation of the data distribution. This pre-determined power along with the existing $P(H0)$ then strengthens the p-value ($P(D|H0)$) and $P(H0|D)$ relationship under repeated sampling rather than the simulation studied in [1].

3. Used simulations to empirically demonstrate the strong non-linear relationship under repeated sampling for fixed $P(H0|D)$ and $P(D|H1)$ for several values of these parameters. Further, Lakens then argued that the results [1] may be inadequate because it assumes a linear regression relationship (ie. Pearson correlation) for the correlation analysis.

Deriving linked $P_{BIC}(D|H0)$ values and p-values under repeated sampling for fixed sample sizes

Masson [3] demonstrated how to calculate estimates of relative posterior probabilities using

$$\frac{P(H0|D)}{P(H1|D)} = \frac{P(D|H0))}{P(D|H1))} \cdot \frac{P(H0)}{P(H1)} \tag{1}$$

$$= BF \cdot \frac{P(H0)}{P(H1)} \tag{2}$$

where BF is the Bayes Factor of the ratio of the likelihoods of the data given H0 and H1 respectively, by using Bayesian Information Criteria (BIC) model fit calculations

$$BIC = -2ln(L) + kln(n) \tag{3}$$

where L is the maximum likelihood of the fitted model, k is the number of free model parameters and n is the sample size.

Explicitly, the BIC based estimate [3] for the Bayes Factor component in (2), is given by

$$\frac{P(H0|D)}{P(H1|D)} \approx \frac{P_{BIC}(D|H0))}{P_{BIC}(D|H1))} \cdot \frac{P(H0)}{P(H1)} \tag{4}$$

$$\approx e^{\frac{(\Delta BIC)}{2}} \cdot \frac{P(H0)}{P(H1)} \tag{5}$$

$$\rightarrow e^{\frac{(\Delta BIC)}{2}} \qquad \text{as } \frac{P(H0)}{P(H1)} \rightarrow 1 \tag{6}$$

where

$$\Delta BIC = BIC_{H1} - BIC_{H0} \tag{7}$$

In comparison, the p_{value} of the frequentist based NHST method is calculated via the probability of the observed t statistic under a two-sided null hypothesis

$$p_{value} = P(\,|t|\,|\,D\,) \tag{8}$$

where $H_{null} : \mu_T = \mu_C$; $H_{alt} : \mu_T \neq \mu_C$

In practice, in the calculations presented in the paper, the two model populations used were (i) an intercept only model and (ii) an intercept with additive treatment effect model

$$H0 : \mu_T = \mu_C \tag{9}$$
$$H1 : \mu_T = \mu_C + \delta \tag{10}$$

To obtain $P_{BIC}(D|H0)$, as explained in [3] the relationship

$$P_{BIC}(D|H0) = \frac{BF}{(BF+1)} \tag{11}$$

is used and the same (r package) lm r-object output was explicitly employed for p_{value} and BF (BIC based Bayes Factor) calculations using summary(lm) and BIC(lm) commands respectively.

Empirical behaviour of $P_{BIC}(D|H0)$ vs p_{value} under repeated sampling for different fixed sample sizes

Under A/A conditions ie. $P(H1) = 0$

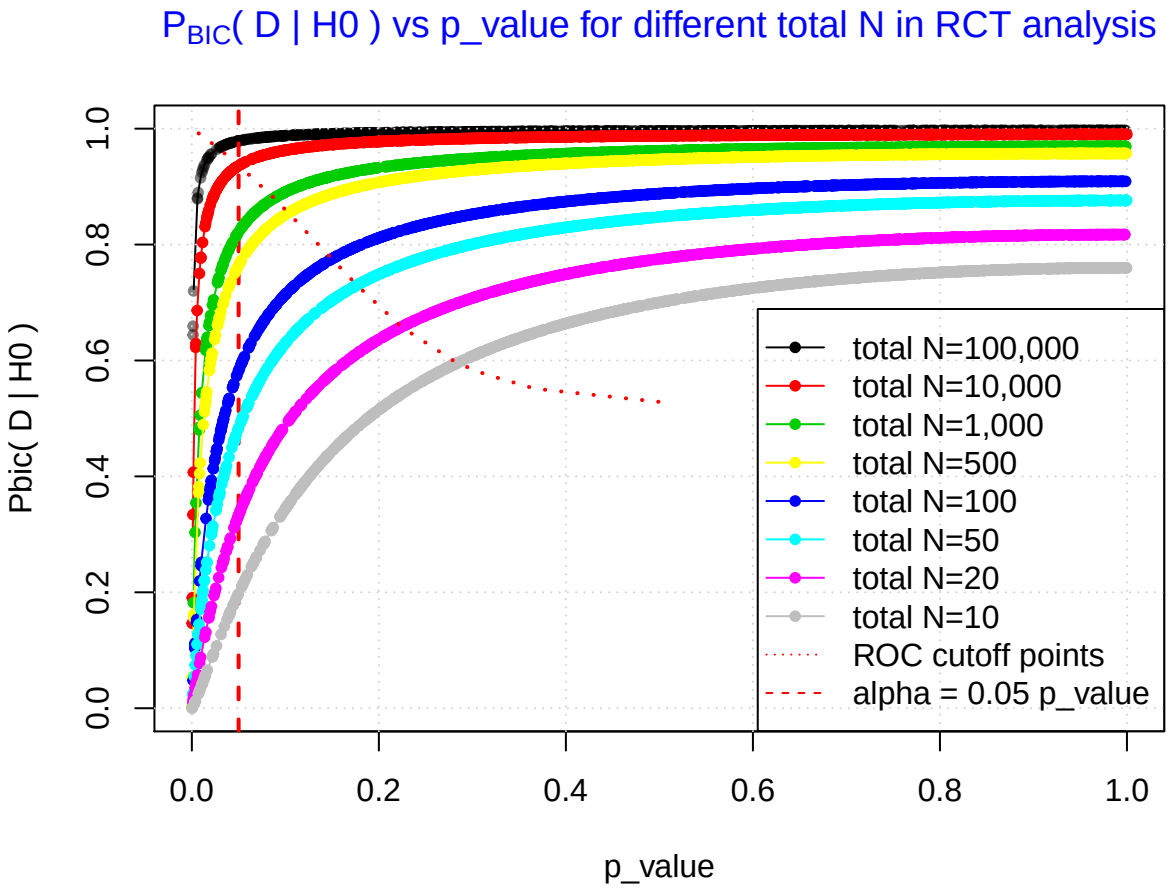

Figure 1. *$P_{BIC}(D|H0)$ as a function of p value for different sample sizes under simple RCT analysis when $P(H0) = 1$.*

Under A/B conditions ie. $P(H1) = 1$

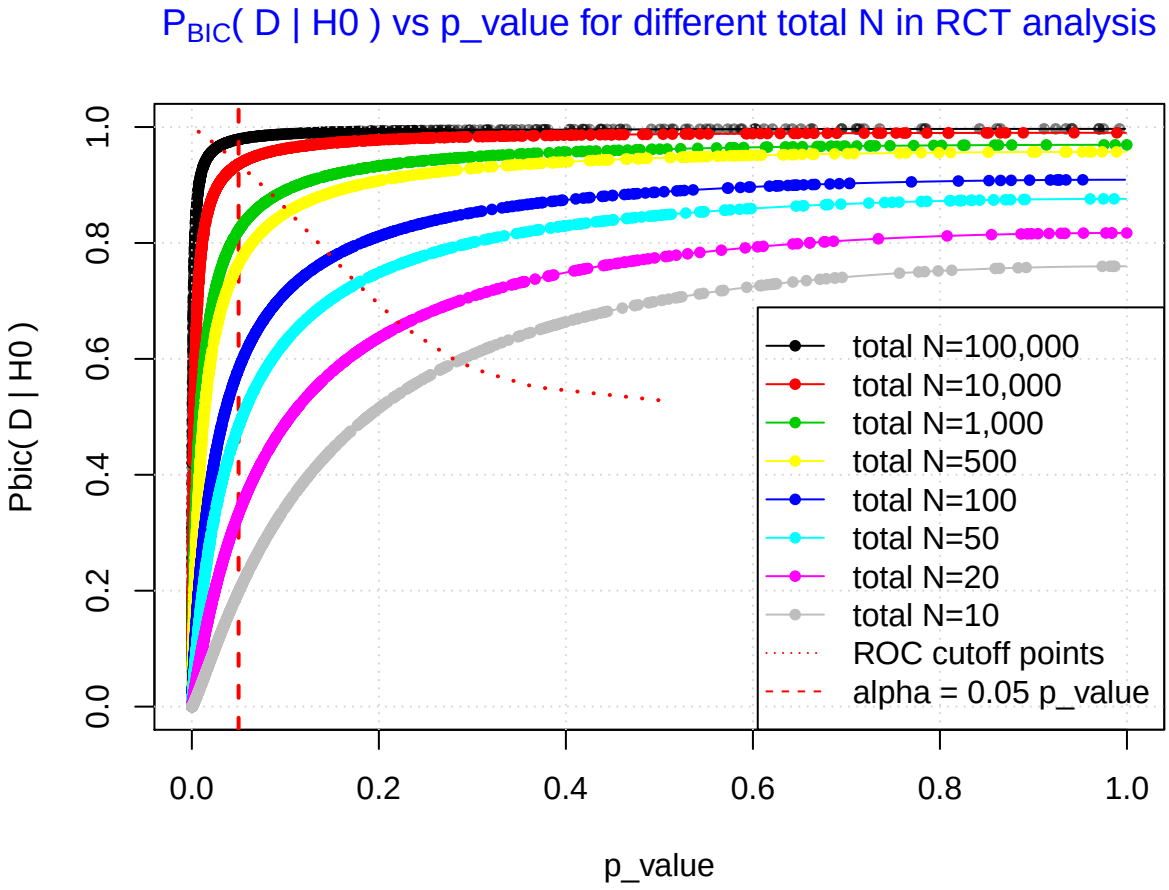

Figure 2. $P_{BIC}(D|H0)$ ***as a function of p value for different sample sizes under simple RCT analysis when*** $P(H0) = 1$***. The contour line barely changes when*** $P(H0) < 1$ ***as*** $cor_{spearman,RCT}(p_{value}, P_{BIC}(D|H0)) \approx 1$ ***remains but the distribution of observed p values and*** $P_{BIC}(D|H0)$ ***values moves to the left. As evidence, of fixed contour line behaviour the fitted lines on the graph are from*** $P(H1) = 0$ ***data and these lines fit the*** $P(H1) = 1$ ***data points very well.***

Looking at Figure 1, with $P(H1) = 0$ under repeated sampling for fixed sample size there is a strong non-linear contour line behaviour between $P_{BIC}(D|H0)$ and p_{value}. The curvature of the contour line depends on the fixed sample size. As the the sample size increases, the density of the observed data pairs of ($P_{BIC}(D|H0)$,p_{value}) with $P_{BIC}(D|H0) < 0.8$ in the region $0 \leq p_{value} \leq 0.05$ decreases.

Using the Spearman Rank correlation calculation $cor_{spearman,RCT}(p_{value}, P_{BIC}(D|H0)) \approx 1$ which is consistent with a strong non-linear relationship between ($P_{BIC}(D|H0)$,p_{value}) observed data in simple RCT analysis when $P(H0) = 1$.

Looking at Figure 2, with $P(H1) = 1$ and using the fitted contour lines from $P(H1) = 0$ data from figure 1, the data using $P(H1) = 1$ lies closely on the same contour line for a fixed sample size but the distribution of the observed data shifts (left and down) to lower $P_{BIC}(D|H0)$ and p_{values} consistent with the measured treatment effect results being determined as statistically significant at a higher rate.

Again using the Spearman Rank correlation calculation $cor_{spearman,RCT}(p_{value}, P_{BIC}(D|H0)) \approx 1$ which is consistent with a strong non-linear relationship between ($P_{BIC}(D|H0)$,p_{value}) observed data in simple RCT

analysis when $P(H0) < 1$. As the treatment effect increases further, more observed data points will shift more dramatically (left and down) to lower $P_{BIC}(D|H0)$ and p_{values} consistent with increasing statistical power.

ROC curve interpretation and analysis of $P_{BIC}(D|H0)$ vs p_{value} behaviour for different fixed sample sizes

It can also be observed from figures 1 & 2 that the $P_{BIC}(D|H0)$ vs p_{value} behaviour for fixed sample sizes has the form of a Receiver Operating Characteristic (ROC) curve which appears often for binary classification problems. In this case the binary classification objective is assigning statistical significance labels to particular observed $P_{BIC}(D|H0)$ and/or p_{value} estimates.

Using the approximation equation (6), such data analysis would also be used to estimate statistical significance for $\hat{P}(H0|D)$.

Using the ROC curve interpretation, the optimal threshold for $P_{BIC}(D|H0)$ and by extrapolation $\hat{P}(H0|D)$ using the BIC based relative posterior probability equations (5) - (6) could be calculated by the elbow point of the ROC curve behaviour. Two common methods are the Youden Index and the maximum height from the diagonal (between minimum and maximum values). In this paper, both methods gives very similar results for the elbow point of $P_{BIC}(D|H0)$ vs p_{value} behaviour.

The elbow point cutoff p_{value} behaviour as a function of sample size model analysis is shown in figure 3 as a log-log plot. The cutoff points are shown in figures 1 & 2 at the intersection of the contour lines and the red dotted line. The alternate cutoff points using the standard $\alpha = .05$ NHST method is shown in figures 1 & 2 at the intersection of the contour lines and the vertical red dashed line. For total sample size of 10,000, the optimal ROC cutoff point is $p_{value} \sim 0.05$ agreeing closely with the standard NHST method.

One immediate comment about the ROC curve cutoff point for N > 10,000 becoming a $p_{value} < .05$ is that [4,5] Bayesian arguments have already been made that the NHST approach ($\alpha = 0.05$) seems too high for large samples based on increasing Bayesian probabilities of false positives and "it must be the case that a large enough sample will produce a significant result" [4]. A possible reply to that argument based on figure 1 data, is that the contour lines for N > 10,000 arises from $P(H0) = 1$.

Two resolutions to the above may be: Firstly, the proper application of the observed data for posterior probability calculations rather than just using the Bayes Factor approach would therefore be that the prior odds $\frac{P(H0)}{P(H1)}$ be included rather than the common usage of $\frac{P(H0)}{P(H1)} \approx 1$. However, using $P(H1) = 0$ is problematic for prior odds estimation. Secondly, the BIC based Bayes Factor approach [3] is an approximation and may not be sensitive enough for the $P(H0) \to 1$ case.

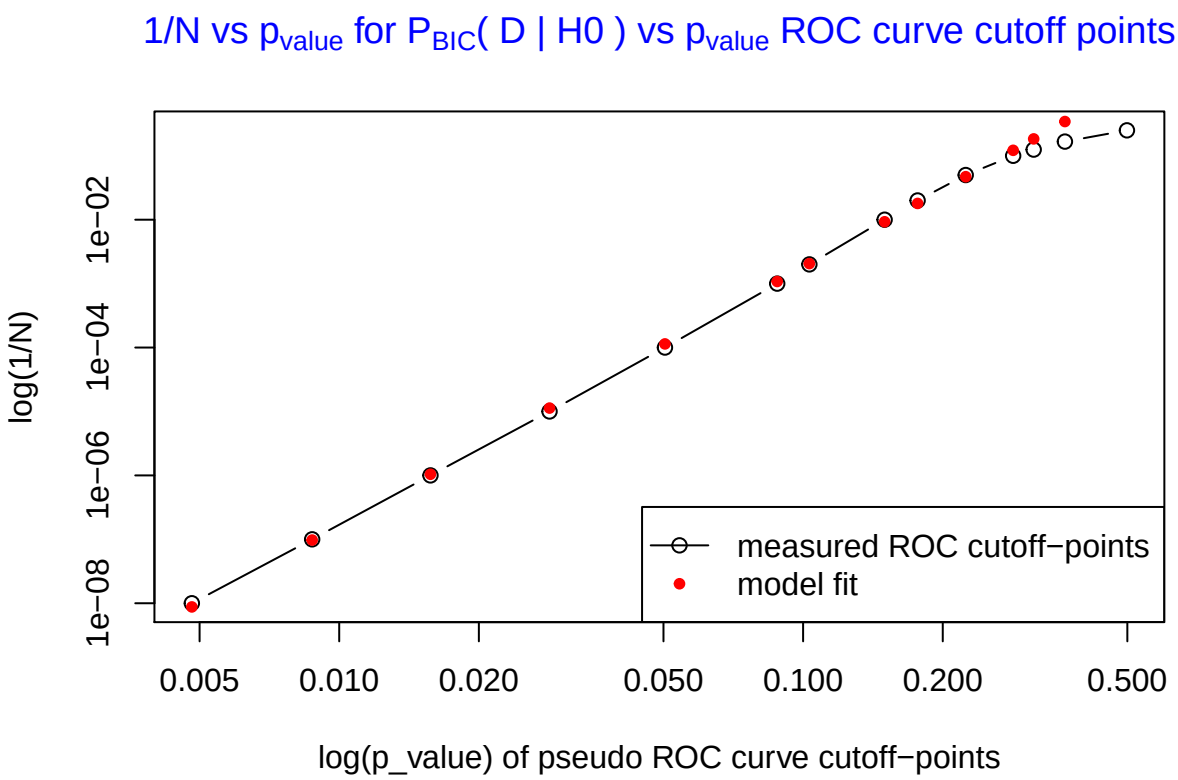

Figure 3. Linear regression fitted log-log scatterplot of 1/(N) vs p_{value} for ROC curve cutoff point analysis.

Normalised $P_{BIC}(D|H0)$ values

It can be seen in figure 1 that $P_{BIC}(D|H0) < 1$ when $P(H0) = 1$. For small sample the evidence for the dominance of H0 or H1 is weaker (consistent with low statistical power) and $P_{BIC}(D|H0) \ll 1$. However, as discussed for Figure 2 the contour line for fixed sample size doesn't move with treatment effect, only the observed data shifts (left and down) along the contour line as statistical power increases.

To improve interpretation of $P_{BIC}(D|H0)$ with respect to P(H0), a simple normalisation factor by inspection can be added to equation (11)

$$normP_{BIC}(D|H0) = \frac{BF}{(BF+1)} \cdot \frac{1}{(1+\frac{1}{\sqrt{N}})} \tag{12}$$

such that $normP_{BIC}(D|H0)$ has the interval $(0, 1]$.

The behaviour of $normP_{BIC}(D|H0)$ is shown in Figure 4 and the same ROC cutoff p_{value} apply as for $P_{BIC}(D|H0)$. To calculate the ROC cutoff point an additional method now becomes available $\frac{dnormP_{BIC}(D|H0)}{dp} = 1$ and is in close agreement with the Youden Index and maximum height from the diagonal (which now has fixed minimum=0 and maximum=1 for $normP_{BIC}(D|H0)$ irrespective of sample size). As $N \to \infty$ the slope of the ROC cutoff line approaches -1.

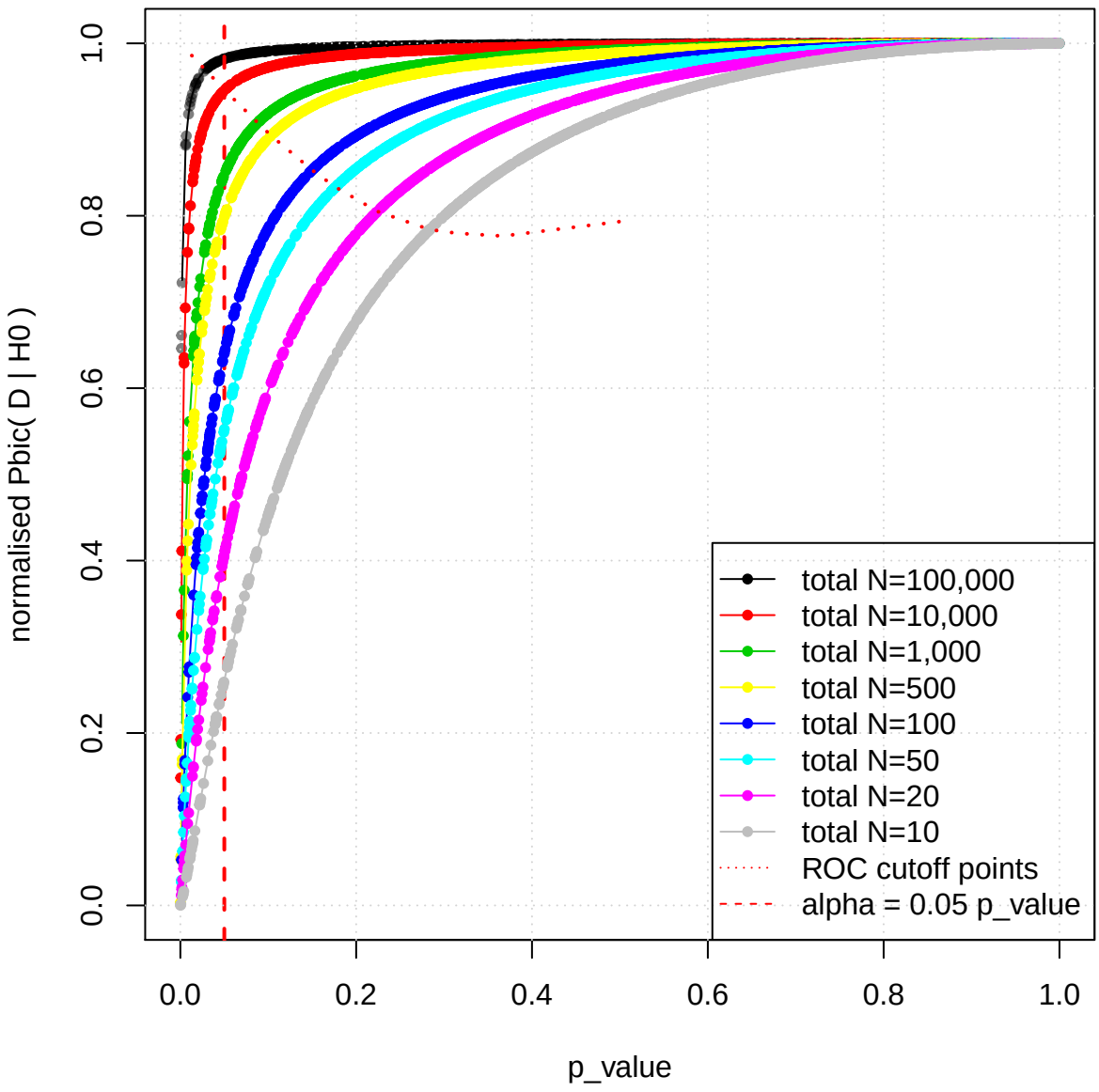

Figure 4. *$normP_{BIC}(D|H0) \equiv \frac{P_{BIC}(D|H0)}{(1+\frac{1}{\sqrt{N}})}$ as a function of p value for different sample sizes under simple RCT analysis when $P(H0) = 1$. Using the $normP_{BIC}(D|H0)$ value, the pseudo ROC cutoff point estimate under Youden Index, elbow point and $\frac{dnormP_{BIC}(D|H0)}{dp} = 1$ methods are in very close agreement.*

Conclusions

Under repeated sampling, for a fixed sample size the Spearman Rank correlation coefficient $cor_{spearman,RCT}(p_{value}, P_{BIC}(D|H0)) \approx 1$ where $P_{BIC}(D|H0)$ is derived from BIC based Bayes Factor calculations [3] indicating a strong relationship between the NHST p_{value} and Bayesian $P_{BIC}(D|H0)$.

This result along with $(p_{value}, P_{BIC}(D|H0))$ contour lines independent of treatment effect and P(H1) gives useful information to interpretation of relative posterior probability based estimates of $P(H0|D)$ (2) assuming $\frac{P(H0)}{P(H1)} \approx constant$.

Finally, a normalised version of P_{BIC}(D|H0) provides easier ROC point cutoff calculation and interpretation.

References

1. Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. The Journal of General Psychology, 136, 261-270.
2. Lakens, D. (2015) https://daniellakens.blogspot.com/2015/11/the-relation-between-p-values-and.html
3. Masson, M. E. J. (2011) A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. Behav Res 43:679–690 DOI 10.3758/s13428-010-0049-5
4. Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.
5. Hagen, R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.