Tina Memo No. 2004-010
To Appear in; Towards a Quantitative mehtodology for the Quantitative Assessment
of Cerebral Blood Flow in Magnetic Resonance Imaging. PhD Thesis, M.L.J.Scott, Manchester, 2004.

Critical Values for the Test of Flatness of a Histogram Using the Bhattacharyya Measure.

M.L.J.Scott.

Last updated
26 / 11 / 2004

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Critical Values for the Test of Flatness of a Histogram Using the Bhattacharyya Measure.

M. Scott

Introduction

The distance between the observed (h_i) and expected (t_i) bin values of a histogram (for each bin i) can be calculated using several measures. The χ^2 test for the histogram is:

$$\chi^2 = \sum_{i=1}^{m} \frac{(h_i - t_i)^2}{t_i} \tag{1}$$

This measure is unsuitable for expected bin values of less than 5, where the Fisher Exact test is normally recommended.

However, calculation of Fisher's test is not straight forward. To compute the hypothesis probability, we must first compute the probability of the current configuration and then integrate all possible permutations over this and all other less frequent distributions, in order to compute the fraction of samples which would be less like that observed. While this is directly available in standard statistical packages it is not simple to implement should a test be required in a new piece of software.

The failure of the standard χ^2 test is due to lack of agreement between the Poisson and Gaussian distributions for small values. This agreement can be improved substantially using the square-root transform [1]. The Bhattacharyya distance measure (B) can then be used [3] instead as the basis for the statistical test:

$$B = \sum_{i=1}^{m} (\sqrt{h_i} \times \sqrt{t_i}) \tag{2}$$

Such measures are typically used to determine whether the data from which the histogram is constructed originates from a uniform distribution; in which case the value of t is constant for all bins.

Having determined the Bhattacharyya measure for the histogram, it is necessary to be able to say whether the value is consistent with the histogram being flat, ie, what is the probability (p) that the B value would not have been obtained by chance alone. If this p-value is below some cut-off value (typically 5% or 1%, and corresponding to a given B value) then the observed histogram is taken as being statistically significantly different from the expected (flat) histogram. Tables of the critical values of B (at p=0.05 etc) for given numbers of entries in the histograms and numbers of bins of the histograms are not commonly available (as they are for the χ^2 distribution for example), so this document details the methods used to produce the Bhattacharyya distributions for given histograms and the corresponding critical values at p=0.05, p=0.01 and p=0.001.

Methods

The distribution of possible values of B for a given histogram of n entries and m bins extending from 0.0 to 1.0 was estimated using a Monte-Carlo simulation of 5000 iterations. n values were randomly generated from a uniform distribution from 0.0 to 1.0, and inserted into a histogram of m bins. The Bhattacharyya measure (B) was calculated for each histogram, with an expected value in each bin of n/m. The 5000 values for B were histogrammed, using 40 bins, for visual inspection.

Figure 1 shows an example B value distribution. Note that the greatest possible value for B is when the histogram is flat. For all i bins, the expected and observed bin values are equal, and B equals the number of bins multiplied by the expected value, ie, the number of entries (n) in the histogram. Therefore, lower the B value, the less flat the histogram. The result of this is that in terms of determining flatness significance values, the Bhattacharyya distribution is one-sided.

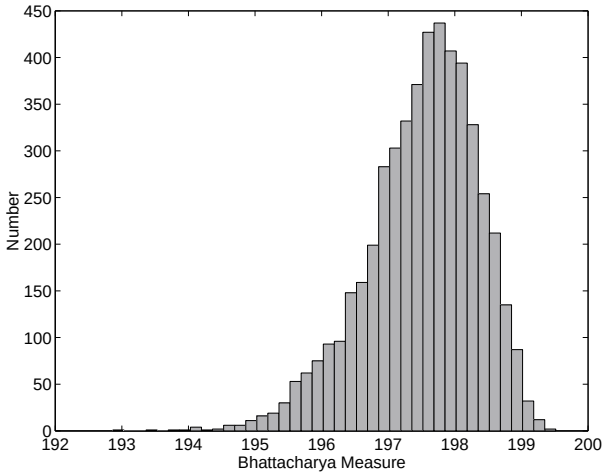


Figure 1: Example distribution of Bhattacharyya values, for 200 entries and 20 histogram bins.

Critical B values at p=0.05, p=0.01 and p=0.001 (below which 5%, 1% and 0.1% of the Bhattacharyya value distribution occurs) were estimated as the 250th, 50th and 5th values of the monotonically increasing sorted values of the B distribution.

The errors on the estimation of the critical values can be given by the error estimation for a binomial distribution, given a probability P and number of entries N:

$$\sigma = \frac{1}{\sqrt{N}} \times \sqrt{P - P^2} \tag{3}$$

For 5000 entries and P values of 0.05, 0.01 and 0.001, the errors on the estimation of P are 0.0031, 0.0014 and 0.00045 respectively. This is sufficient for the use of these variables in statistical tests. If more precision were required it could be achieved by increasing the Monte-Carlo sanple size.

Tables

	Number of Bins									
No. Entries	5	10	15	20	25	30	35	40	45	50
10	0.773	0.683	0.605	0.535	0.515	0.470	0.441	0.412	0.412	0.391
20	0.892	0.815	0.749	0.705	0.663	0.617	0.594	0.559	0.543	0.518
30	0.954	0.883	0.832	0.789	0.748	0.712	0.682	0.653	0.633	0.615
40	0.967	0.920	0.881	0.842	0.807	0.776	0.747	0.722	0.698	0.677
50	0.975	0.941	0.908	0.878	0.848	0.820	0.795	0.770	0.748	0.728
60	0.978	0.961	0.931	0.904	0.877	0.854	0.831	0.807	0.787	0.766
70	0.983	0.965	0.941	0.920	0.899	0.876	0.855	0.836	0.815	0.799
80	0.985	0.971	0.951	0.934	0.914	0.896	0.877	0.858	0.839	0.823
90	0.986	0.975	0.961	0.943	0.927	0.910	0.894	0.877	0.861	0.845
100	0.988	0.977	0.967	0.950	0.937	0.922	0.906	0.891	0.876	0.861
150	0.992	0.985	0.979	0.972	0.963	0.955	0.946	0.937	0.926	0.916
200	0.994	0.989	0.985	0.980	0.975	0.969	0.963	0.957	0.950	0.943
250	0.995	0.991	0.988	0.984	0.980	0.977	0.972	0.968	0.963	0.958
500	0.998	0.996	0.994	0.992	0.991	0.989	0.987	0.986	0.984	0.982
1000	0.999	0.998	0.997	0.996	0.995	0.995	0.994	0.993	0.992	0.992

Table 1: Critical B values (normalised to the number of entries) for a probability of 0.05.

Tables 1, 2 and 3 give the critical Bhattacharyya values (for histograms of a given number of entries and number of bins) for probability values of 0.05, 0.01 and 0.001 respectively. Figures 2(a), 2(b) and 2(c) show corresponding 3D plots of the values against numbers of bins and entries. Note that for comparative purposes, the values in the tables have been normalised to the number of entries in the histogram. From the tables and figures it can be

No. Entries	Number of Bins									
	5	10	15	20	25	30	35	40	45	50
10	0.740	0.615	0.558	0.524	0.472	0.437	0.409	0.407	0.384	0.364
20	0.863	0.792	0.711	0.674	0.636	0.589	0.566	0.538	0.522	0.497
30	0.934	0.854	0.801	0.759	0.722	0.686	0.659	0.631	0.611	0.588
40	0.951	0.902	0.854	0.815	0.778	0.751	0.727	0.700	0.677	0.656
50	0.963	0.922	0.890	0.854	0.826	0.796	0.772	0.747	0.727	0.706
60	0.971	0.938	0.909	0.882	0.857	0.830	0.811	0.783	0.762	0.744
70	0.974	0.950	0.927	0.899	0.877	0.856	0.837	0.817	0.796	0.779
80	0.978	0.962	0.936	0.916	0.898	0.879	0.859	0.839	0.819	0.806
90	0.981	0.967	0.946	0.927	0.910	0.892	0.876	0.860	0.841	0.828
100	0.983	0.969	0.954	0.937	0.923	0.908	0.891	0.875	0.860	0.845
150	0.988	0.980	0.973	0.964	0.953	0.946	0.934	0.925	0.914	0.904
200	0.992	0.986	0.980	0.975	0.968	0.961	0.955	0.949	0.942	0.933
250	0.993	0.989	0.985	0.981	0.977	0.971	0.966	0.961	0.957	0.952
500	0.997	0.994	0.992	0.990	0.989	0.987	0.985	0.984	0.981	0.980
1000	0.998	0.997	0.996	0.995	0.994	0.994	0.993	0.992	0.991	0.990

Table 2: Critical B values (normalised to the number of entries) for a probability of 0.01.

No of Entries	Number of Bins									
	5	10	15	20	25	30	35	40	45	50
10	0.716	0.597	0.502	0.476	0.421	0.397	0.396	0.371	0.350	0.338
20	0.838	0.731	0.657	0.613	0.600	0.550	0.527	0.509	0.480	0.478
30	0.873	0.805	0.772	0.717	0.676	0.657	0.633	0.609	0.572	0.559
40	0.922	0.863	0.812	0.776	0.746	0.720	0.691	0.675	0.647	0.634
50	0.945	0.886	0.855	0.824	0.799	0.768	0.742	0.713	0.701	0.676
60	0.958	0.915	0.890	0.862	0.824	0.806	0.789	0.759	0.735	0.721
70	0.962	0.924	0.899	0.879	0.856	0.831	0.810	0.788	0.774	0.752
80	0.969	0.935	0.916	0.894	0.879	0.857	0.833	0.818	0.792	0.784
90	0.972	0.943	0.935	0.905	0.892	0.873	0.855	0.844	0.821	0.810
100	0.978	0.959	0.938	0.921	0.902	0.891	0.870	0.854	0.836	0.830
150	0.986	0.975	0.953	0.953	0.939	0.934	0.919	0.911	0.904	0.888
200	0.987	0.981	0.974	0.969	0.962	0.948	0.947	0.940	0.929	0.916
250	0.991	0.986	0.981	0.976	0.971	0.963	0.959	0.952	0.947	0.940
500	0.996	0.993	0.991	0.989	0.987	0.985	0.983	0.980	0.977	0.976
1000	0.998	0.996	0.995	0.995	0.993	0.992	0.991	0.991	0.990	0.989

Table 3: Critical B values (normalised to the number of entries) for a probability of 0.001.

seen that the critical B values increase with increasing numbers of entries and decreasing number of bins. This is indicative of the fact that the greater the number of entries and the fewer the number of bins, the tighter the distribution of frequency of occurrence of possible histogram shapes.

Notice that the typical number of entries per bin for these tables is less than 5. For values greater than 5 there is no need not to use the conventional χ^2 test. However, it should be noted that for large values the test statistic given by;

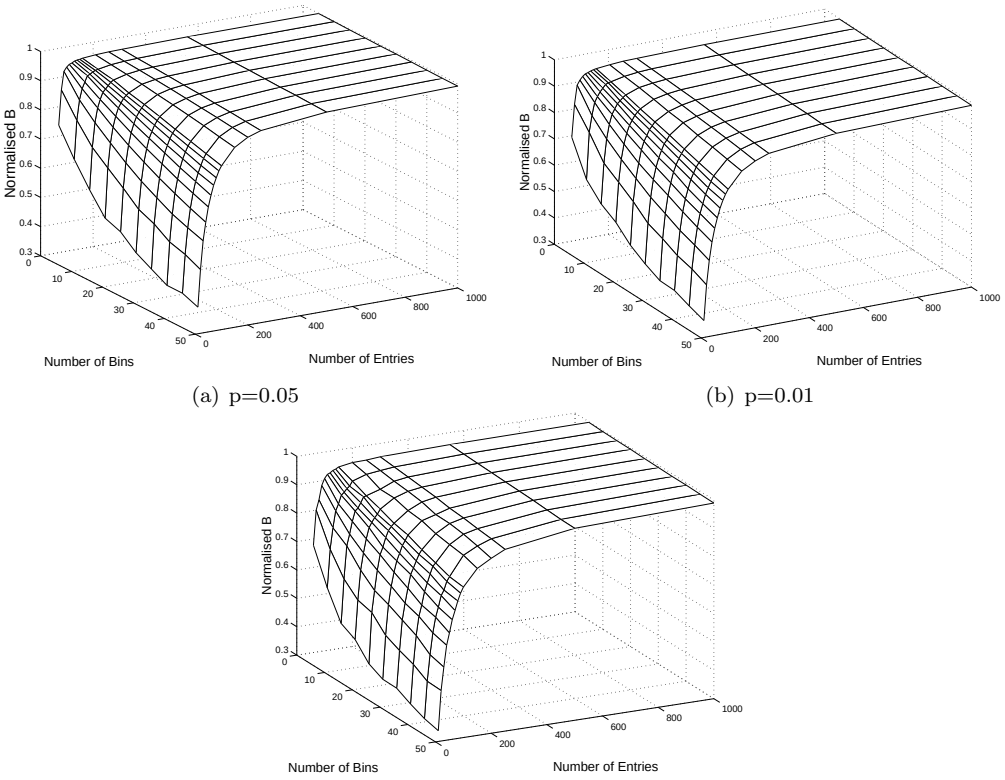
$$T = (N - B)/2$$

is a effectively a χ^2 variable with m degrees of freeedom.

Histograms

Figures 3(a)-3(l) show the Bhattacharyya distributions for histograms of various numbers of bins (left-right) and numbers of entries (top-bottom). Note that as the number of bins increases beyond the number of entries, the histograms become increasingly quantised. As the number of entries in the histogram increases, the distribution of Bhattacharyya values becomes increasingly Gaussian-like.

If the data is naturally quantised into the bins of a histogram then the application of this measure is strightforward.



(a) p=0.05

(b) p=0.01

(c) p=0.001

Figure 2: 3D plots of normalised Bhattacharyya critical values against numbers of bins and numbers of entries given the probability.

If however, we wish to compare distributions of continuous variables in this way one might be tempted to think that the number of bins m is a free parameter, and to scan all possible values in order to get the most significant result. This would be incorrect at several levels. The key point however, is probably that the ambiguity between data introduced by the histogram binning process should reflect knowledge of the genuine ambiguity of data, ie: measurement error. This observation is consistent with the ideas outlined in [2].

Matlab Code

Below is listed the Matlab code for obtaining the distributions and critical values for the Bhattacharyya measure.

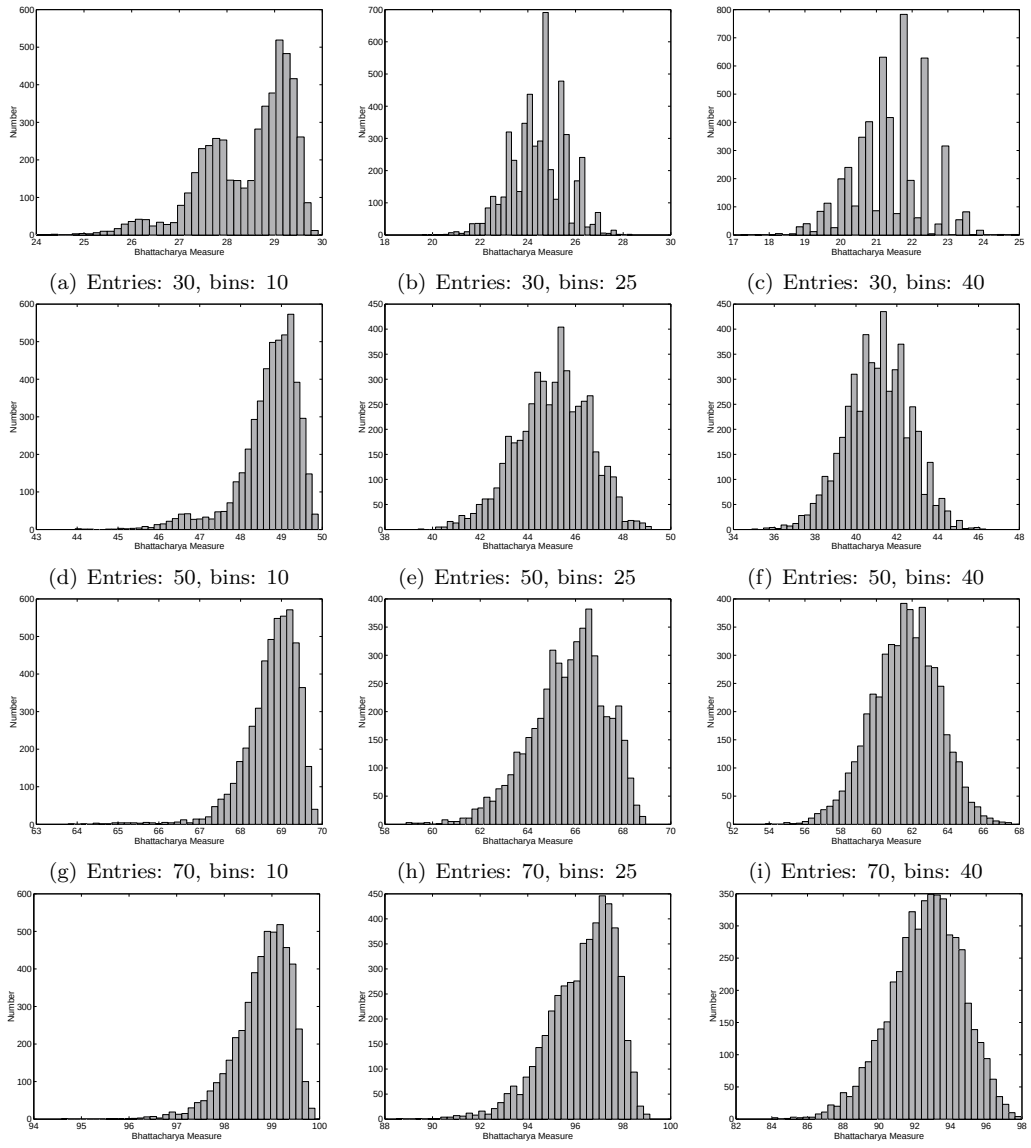
```
function [crit_val] = bhat_tables(no_entries, no_bins)

     y = [1/(2*no_bins):1/no_bins:1-(1/(2*no_bins))];  % finds bin centres

     unif_data = rand(no_entries, 5000);               % creates random uniform data
     unif_data_hist = hist(unif_data, y);              % histograms data
     bval = sum(sqrt(unif_data_hist)*sqrt(no_entries/no_bins)); % calculates B values
     hist(bval,40)                                     % displays B value distribution

     sorted_bvals = sort(bval);                        % sorts B values into ascending order

     crit_val(:,1) = sorted_bvals(:,250);              % gets B values at 5, 1 and 0.1%
     crit_val(:,2) = sorted_bvals(:,50);
     crit_val(:,3) = sorted_bvals(:,5);
```



(a) Entries: 30, bins: 10 (b) Entries: 30, bins: 25 (c) Entries: 30, bins: 40

(d) Entries: 50, bins: 10 (e) Entries: 50, bins: 25 (f) Entries: 50, bins: 40

(g) Entries: 70, bins: 10 (h) Entries: 70, bins: 25 (i) Entries: 70, bins: 40

(j) Entries: 100, bins: 10 (k) Entries: 100, bins: 25 (l) Entries: 100, bins: 40

Figure 3: Bhattacharyya Distributions of histograms with given numbers of entries and bins.

References

1. N.A. Thacker and P.A. Bromiley, The Effects of a Square Root Transform on a Poisson Distributed Quantity. Tina memo, 2001-010, 2001.
2. N.A. Thacker and P.A. Bromiley, The Equal Variance Domain: Issues Surrounding the Use of Continuous Probability Densities in Algorithm Design., Tina memo, 2004-005, 2004.
3. N. A. Thacker, F. J. Aherne and P. I. Rockett, *The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data*, Kybernetika, Vol. 32 No. 4, pp. 1-7, 1997