

χ^2 TEST FOR THE COMPARISON OF WEIGHTED AND UNWEIGHTED HISTOGRAMS

N.D. GAGUNASHVILI

University of Akureyri, Faculty of Information Technology, Borgir, v/Nordurhlóð, IS-600 Akureyri, Iceland
E-mail: nikolai@unak.is

The widely used χ^2 homogeneity test for comparing histograms (unweighted) is modified for cases involving unweighted and weighted histograms. Numerical examples illustrate an application of the method for the case of histograms with a small statistics of events and also for large statistics of events. This method can be used for the comparison of simulated data histograms against experimental data histograms.

1 Introduction

The χ^2 criteria of homogeneity ¹ which is used to compare two or more histograms is well established. Without limiting the general nature of the discussion, we consider two experimental histograms with the same binning and the number of bins equal to r . Let us denote the number of events in the i th bin in the first histogram as n_i and as m_i in the second one. The total number of events in the first histogram is equal to $N = \sum_{i=1}^r n_i$, and $M = \sum_{i=1}^r m_i$ in the second histogram.

The hypothesis of homogeneity is that the two histograms represent random values with identical distributions. This is equivalent to there existing r constants p_1, \dots, p_r , such that $\sum_{i=1}^r p_i = 1$, and the probability of belonging to the i th bin for some measured value in both experiments is equal to p_i . If the hypothesis of homogeneity is valid, then $p_i, i = 1, \dots, r$, can be estimated from the data as

$$\hat{p}_i = \frac{n_i + m_i}{N + M}, \quad (1)$$

and then

$$X^2 = \sum_{i=1}^r \frac{(n_i - N\hat{p}_i)^2}{N\hat{p}_i} + \sum_{i=1}^r \frac{(m_i - M\hat{p}_i)^2}{M\hat{p}_i} \quad (2)$$

has approximately a $\chi^2_{(r-1)}$ distribution ¹.

2 The test

A simple modification of the ideas described above can be used for the comparison of unweighted and weighted histograms. Let us formulate the hypothesis of identity of an unweighted histogram to a weighted histogram so that there exist r constants p_1, \dots, p_r , such that $\sum_{i=1}^r p_i = 1$, and for any i th bin the following equations are valid:

$$n_i = Np_i + \delta(n_i), \quad w_i = Wp_i + \delta(w_i). \quad (3)$$

Here w_i is the weight of the contents of an i th bin, $W = \sum_i w_i$ is the common weight of the weighted histogram; $\delta(n_i), \delta(w_i), i = 1, \dots, r$, are the random residuals with expectation values $E\delta(n_i) = E\delta(w_i) = 0$ and variances $\text{Var}\delta(n_i) = En_i$, $\text{Var}\delta(w_i) = \sigma_i^2$. If we replace the variance $\text{Var}\delta(n_i)$ with the estimate n_i , the variance $\text{Var}\delta(w_i)$ with estimate s_i^2 (sum of squares of weights of events in the i th bin) and the hypothesis of identity is valid, then $p_i, i = 1, \dots, r$, can be estimated from the data by the Least Squares Method ²

$$\hat{p}_i = \frac{N + w_i W / s_i^2}{N^2 / n_i + W^2 / s_i^2}. \quad (4)$$

We may then use the test statistic

$$X^2 = \sum_{i=1}^r \frac{(n_i - N\hat{p}_i)^2}{n_i} + \sum_{i=1}^r \frac{(w_i - W\hat{p}_i)^2}{s_i^2} \quad (5)$$

and it is plausible that this has approximately a $\chi^2_{(r-1)}$ distribution.

This method, as well as the original one ¹, has a restriction on the number of events in a bin. The number of events recommended for the proposed method is more than 25. In the case of a weighted histogram if the number of events is unknown, then we can apply this recommendation for the equivalent number of events as $n_i^{equiv} = w_i^2 / s_i^2$.

The studentised residuals

$$R_i = \frac{n_i - N\hat{p}_i}{\sqrt{n_i} \sqrt{1 - 1/(1 + W^2 n_i / N^2 s_i^2)}} \quad (6)$$

have approximately a normal distribution with mean equal to 0 and standard deviation equal to 1 ². Analysis of the residuals can be useful for the identification of bins that are outliers, or bins that have a big influence on X^2 .

3 Numerical example

The method described herein is now illustrated with an example. We take a distribution

$$\phi(x) = \frac{2}{(x-10)^2+1} + \frac{1}{(x-14)^2+1} \quad (7)$$

defined on the interval $[4, 16]$. Events distributed according to the formula (7) are simulated to create the unweighted histogram. Uniformly distributed events are simulated for the weighted histogram with weights calculated by formula (7). Each histogram has 20 bins. Fig. 1 shows the result of comparison of the unweighted histogram with 2500 events and the weighted one with 500 events.

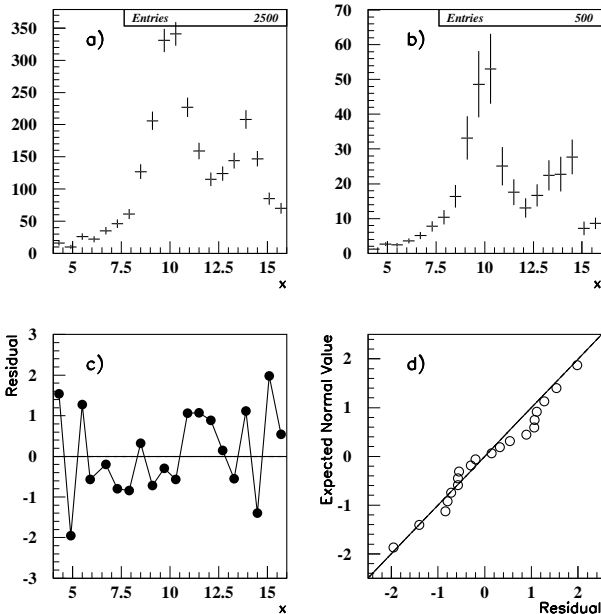


Figure 1. An example of comparison of the unweighted histogram with 2500 events and the weighted histogram with 500 events: a) unweighted histogram; b) weighted histogram; c) studentised residuals plot; d) normal Q-Q plot of residuals.

The value of test statistic X^2 is equal to 21.36 with p -value equal to 0.31, so the hypothesis of identity of the two distributions can be accepted. The behavior of the studentised residuals plot (see Fig. 1c) and the normal Q-Q plot (see Fig. 1d) of residuals are regular and we cannot identify the outliers or bins with a big influence on X^2 .

To investigate the dependence of the distribution of the test statistics on the number of events, three cases were considered. The first case is the unweighted histogram with 1000 events and weighted

with 200 events; the second case is 2500 events in unweighted histogram and 500 events in weighted; and the third case has 10000 and 2000 events respectively. In each case 10000 pairs of histograms were simulated with calculation of X^2 statistics for the each pair. Fig. 2 shows the Chi-square Q-Q plots and the histograms of X^2 statistics. As we can see the real distribution of test statistics obtained for low number of events has a heavier tail than the theoretical χ^2_{19} distribution. It means that the p -value calculated with the theoretical χ^2_{19} distribution is lower than the real p -value and any decision about rejecting the hypothesis of identity of the two distributions is conservative. The distribution of test statistics for the second case is close to the theoretical distribution and confirms that the greater than 25 entries in a bin is reasonable for the application of the method.

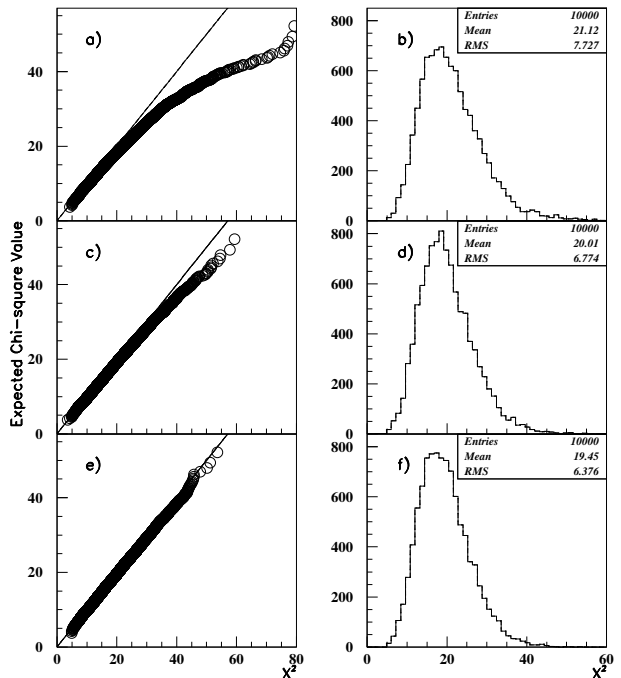


Figure 2. Chi-square Q-Q plots and histograms of X^2 statistics for: a),b) unweighted histograms with 1000 events and weighted with 200; c),d) unweighted histograms with 2500 events and weighted with 500; e),f) unweighted histograms with 10000 events and weighted with 2000.

References

1. H. Cramer, *Mathematical methods of statistics* (Princeton University Press, 1946).
2. G.A.F. Seber, *Linear Regression Analysis* (John Wiley & Sons Inc, 2003).