Goodness of fit tests for weighted histograms

Nikolai Gagunashvili nikolai@unak.is

University of Akureyri, Iceland

ABSTRACT

Weighted histograms in Monte-Carlo simulations are often used for the estimation of probability density functions. They are obtained as a result of random experiment with random events that have weights. In this paper the bin contents of weighted histogram are considered as a sum of random variables with random number of terms. Goodness of fit tests for weighted histograms and for weighted histograms with unknown normalization are proposed. Sizes and powers of the tests are investigated numerically.

Introduction

A histogram with m bins for a given probability density function $p(x)$ is used to estimate the probabilities

$$p_i = \int_{S_i} p(x)dx,\ i = 1, \ldots, m \tag{1}$$

that a random event belongs to bin i.
The problem of goodness of fit is to test the hypothesis

$$H_0 : p_1 = p_{10}, \ldots, p_{m-1} = p_{m-1,0} \text{ vs. } H_a : p_i \neq p_{i0} \text{ for some } i, \tag{2}$$

where p_{i0} are specified probabilities, and $\sum_{i=1}^{m} p_{i0} = 1$. The test is used in a data analyses for comparison theoretical frequencies np_{i0} with the observed frequencies n_i. The test statistic

$$X^2 = \sum_{i=1}^{m} \frac{(n_i - np_{i0})^2}{np_{i0}} \tag{3}$$

was suggested by Pearson. Pearson showed that the statistic (3) has approximately a χ^2_{m-1} distribution if the hypothesis H_0 is true.

To define a weighted histogram let us write the probability p_i (1) for a given probability density function $p(x)$ in the form

$$p_i = \int_{S_i} p(x)dx = \int_{S_i} w(x)g(x)dx \tag{4}$$

where $w(x) = p(x)/g(x)$ is the weight function and $g(x)$ is some other probability density function.
For weighted histograms again the problem of goodness of fit is to test the hypothesis

$$H_0 : p_1 = p_{10}, \ldots, p_{m-1} = p_{m-1,0} \text{ vs. } H_a : p_i \neq p_{i0} \text{ for some } i, \tag{5}$$

where p_{i0} are specified probabilities, and $\sum_{i=1}^{m} p_{i0} = 1$. In practice the heuristic "chi-square" test statistic is used for this purpose

$$X_h^2 = \sum_{i=1}^{m} \frac{(W_i - np_{i0})^2}{W_{2i}}, \tag{6}$$

where $W_i = \sum_{k=1}^{n_i} w_i(k)$ and $W_{2i} = \sum_{k=1}^{n_i} w_i(k)^2$. It is expected that if hypothesis H_0 is true then statistic X_h^2 has χ^2_{m-1} distribution. The recommended minimal number of events in a bin is equal to 25 for application this test.

The test

The total sum of weights of events in ith bin W_i, $i = 1, \ldots, m$ can be considered as a sum of random variables

$$W_i = \sum_{k=1}^{n_i} w_i(k), \tag{7}$$

where also the number of events n_i is a random value and the weights $w_i(k), k = 1, ..., n_i$ are independent random variables with the same probability distribution function. If hypothesis H_0 is true then

$$\mathrm{E}\, W_i = np_{i0},\ i = 1, \ldots, m. \tag{8}$$

Elements of covariance matrix equal to $\gamma_{ii} = n(p_{i0}/r_i - p_{i0}^2)$ and $\gamma_{ij} = -np_{i0}p_{j0}$ for $i \neq j$ where $r_i = \mu_i/\alpha_{2i}$

Let us now introduce the multivariate T^2 Hotelling statistic

$$(\mathbf{W} - n\mathbf{p}_0)'\Gamma_k^{-1}(\mathbf{W} - n\mathbf{p}_0), \tag{9}$$

where $\mathbf{W} = (W_1, \ldots, W_{k-1}, W_{k+1}, \ldots, W_m)'$, $\mathbf{p}_0 = (p_{10}, \ldots, p_{k-1,0}, p_{k+1,0}, \ldots, p_{m0})'$ and $\Gamma_k = (\gamma_{ij})_{(m-1)\times(m-1)}$ is the covariance matrix for a histogram without bin k. The matrix Γ_k has the form

$$\Gamma_k = \mathrm{diag}\,(n\frac{p_{10}}{r_1}, \ldots, n\frac{p_{k-1,0}}{r_{k-1}}, n\frac{p_{k+1,0}}{r_{k+1}}, \ldots, n\frac{p_{m0}}{r_m}) - n\mathbf{p}_0\mathbf{p}_0', \tag{10}$$

and the Woodbury theorem can be applied to find Γ_k^{-1}. After that the Hotelling statistic can be written as

$$X_k^2 = \sum_{i\neq k} r_i \frac{(W_i - np_{i0})^2}{np_{i0}} + \frac{(\sum_{i\neq k} r_i(W_i - np_{i0}))^2}{n - \sum_{i\neq k} r_i np_{i0}}. \tag{11}$$

and can be transformed to

$$X_k^2 = \frac{1}{n}\sum_{i\neq k} \frac{r_i W_i^2}{p_{i0}} + \frac{1}{n}\frac{(n - \sum_{i\neq k} r_i W_i)^2}{1 - \sum_{i\neq k} r_i p_{i0}} - n \tag{12}$$

that is convenient for numerical calculations. Asymptotically the vector $\mathbf{W}$ has a normal distribution $\mathcal{N}(n\mathbf{p}_0, \Gamma_k^{1/2})$ and therefore the test statistic (11) has χ^2_{m-1} distribution if hypothesis H_0 is true. Notice that for usual histograms when $r_i = 1$, $i = 1, \ldots, m$ the statistic (11) is Pearson's chi-square statistic.

Let us now replace r_i with the estimate $\hat{r}_i = W_i/W_{2i}$ and denote the estimator of matrix Γ_k as $\hat{\Gamma}_k$. Then for positive definite matrices $\hat{\Gamma}_k$, $k = 1, \ldots, m$ the test statistic is given as

$$\hat{X}_k^2 = \frac{1}{n}\sum_{i\neq k} \frac{\hat{r}_i W_i^2}{p_{i0}} + \frac{1}{n}\frac{(n - \sum_{i\neq k} \hat{r}_i W_i)^2}{1 - \sum_{i\neq k} \hat{r}_i p_{i0}} - n. \tag{13}$$

Formula (13) for usual histograms does not depend on the choice of the excluded bin, but for weighted histograms there can be a dependence. A test statistic that is invariant to the choice of the excluded bin and at the same time is Pearson's chi square statistics for the usual histograms can be obtained as the median value of (13) with positive definite matrix $\hat{\Gamma}_k$ for a different choice of excluded bin

$$\hat{X}^2 = \mathrm{Med}\ \{\hat{X}_1^2, \hat{X}_2^2, \ldots, \hat{X}_m^2\}. \tag{14}$$

Usage of $\hat{X}^2$ to test the hypothesis H_0 with a given significance level is equivalent to making a decision by voting. Use of the chi-square tests is inappropriate if any expected frequency is below 1 or if the expected frequency is less than 5 in more than 20% of bins. This restriction known for usual chi-square test is quite reasonable for weighted histograms also and helps to avoid cases when matrix $\hat{\Gamma}_k$ is not positive definite.

The test for histograms with unknown normalization

In practice one is often faced the case that a histogram is defined up to an unknown normalization constant C. Let us denote a bin content of histograms without normalization as $\check{W}_i$, then $W_i = \check{W}_i C$, and the test statistic (12) can be written as

$$X_k^2 = \frac{C}{n}\sum_{i\neq k} \frac{\check{r}_i \check{W}_i^2}{p_{i0}} + \frac{1}{n}\frac{(n - \sum_{i\neq k} \check{r}_i \check{W}_i)^2}{1 - C^{-1}\sum_{i\neq k} \check{r}_i p_{i0}} - n, \tag{15}$$

with $\check{r}_i = Cr_i$. An estimator for the constant C can be found by minimization of (15). The normal equation for (15) has the form

$$\sum_{i\neq k} \frac{\check{r}_i \check{W}_i^2}{p_{i0}} - \frac{(n - \sum_{i\neq k} \check{r}_i \check{W}_i)^2}{(C - \sum_{i\neq k} \check{r}_i p_{i0})^2}\sum_{i\neq k} \check{r}_i p_{i0} = 0 \tag{16}$$

with two solutions

$$\hat{C}_k = \sum_{i\neq k} \check{r}_i p_{i0} \pm \sqrt{\frac{\sum_{i\neq k} \check{r}_i p_{i0}}{\sum_{i\neq k} \check{r}_i \check{W}_i^2/p_{i0}}}(n - \sum_{i\neq k} \check{r}_i \check{W}_i), \tag{17}$$

where $\hat{C}_k$ is an estimator of C. We choose the solution with the positive sign because it converges to a constant $C = 1$ for the case of a usual histogram, while the solution with negative sign does not. Substituting (17) to the (15) we get the test statistic

$$\check{X}_k^2 = \frac{s^2}{n} + 2s, \text{ where } s = \sqrt{\sum_{i\neq k} \check{r}_i p_{i0} \sum_{i\neq k} \check{r}_i \check{W}_i^2/p_{i0}} - \sum_{i\neq k} \check{r}_i \check{W}_i \tag{18}$$

that has a χ^2_{m-2} distribution if hypothesis H_0 is valid. The final statistic $\hat{\check{X}}_k^2$ is obtained by replacing $\check{r}_i$ in (18) with the estimate $\hat{\check{r}}_i = \check{W}_i/\check{W}_{2i}$. As in chapter 2, a test statistic that is "invariant" to choice of the excluded bin can be obtained as the median value of (18) for all possible choices of the excluded bin

$$\hat{\check{X}}^2 = \mathrm{Med}\ \{\hat{\check{X}}_1^2, \hat{\check{X}}_2^2, \ldots, \hat{\check{X}}_m^2\}. \tag{19}$$

Evaluation of the tests' sizes and power

Sizes and power of tests is now evaluated with a numerical example. We take a distribution

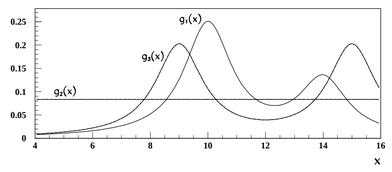
$$p(x) \propto \frac{2}{(x-10)^2+1} + \frac{1}{(x-14)^2+1} \tag{20}$$

defined on the interval $[4, 16]$. Three cases of the probability density function $g(x)$ are considered

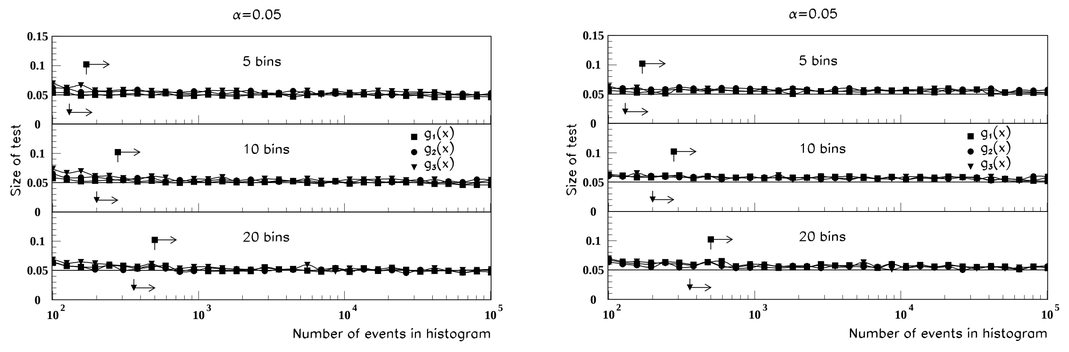
$g_1(x) = p(x)$

$g_2(x) = 1/12$

$g_3(x) \propto \frac{2}{(x-9)^2+1} + \frac{2}{(x-15)^2+1}$

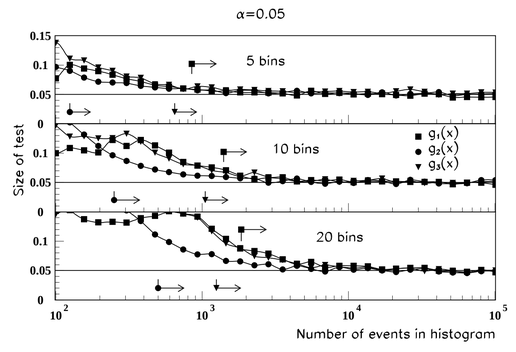


Sizes of tests for histograms with different numbers of bins were calculated for nominal values of size equal to $\alpha = 0.05$ and for a nominal value of size equal to $\alpha = 0.01$. Calculations of test sizes α_s are done using the Monte-Carlo method based on 10000 runs. The same study was done for the chi-square test for histograms with unknown normalization. The results of these calculations are presented in. All cases show that tests sizes are close to nominal values for large numbers n of events and reasonably close to nominal values for low numbers of events.



Sizes of the chi-square tests for histograms (with normalization and without) for different weight functions and different numbers of bins as a function of the number of events n in the histogram. Arrows show regions with appropriate number of events in histogram for test application.

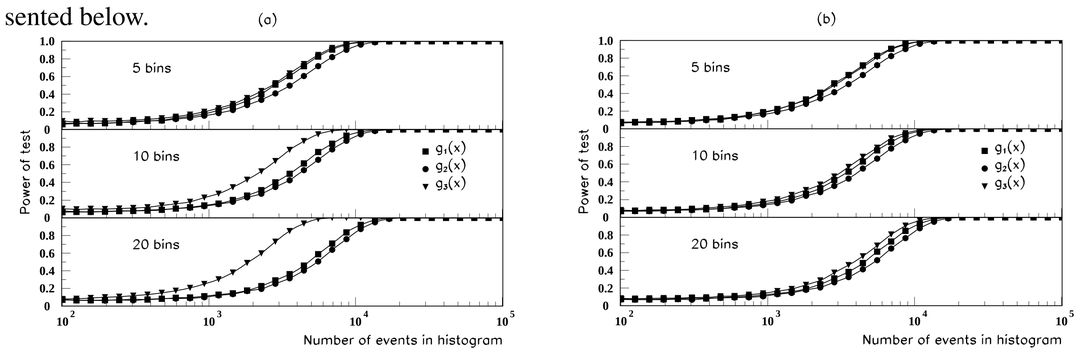
The same computation was done for the size of the heuristic test. It can be noticed that for large number n of events the sizes of tests tend to the nominal value of the test. For small numbers n of events in the histograms the sizes of the tests are generally greater then the nominal values of tests, although some values of sizes are not shown on the figures because they are too big. Comparison of the two tests bring out clearly the superiority of the generalization of Pearson's test over the heuristic test.



At the plot arrows show regions with minimal number of events in bins of histograms equal to 25. The powers of the new chi-square test and the test with unknown normalization were investigated for slightly different values of the amplitude of the second peak of the specified probability distribution function (see Fig. 5):

$$p_0(x) \propto \frac{2}{(x-10)^2+1} + \frac{1.15}{(x-14)^2+1}. \tag{21}$$

The results of these calculations for tests with known normalization(a) and with unknown(b) are presented below.



This paper can be found on http://dx.doi.org/10.1016/j.nima.2008.08.144