MRU111

Notes on comparing two histograms. Bill Menke, October 18, 2013

Suppose that one has two observed histograms, $h_1(x)$ and $h_2(x)$. Are they significantly different? If not, then there is a theoretical histogram $h_0(x)$ for which there is less than, say, a 95% probability that both "$h_1(x)$ compared with $h_0(x)$" and "$h_2(x)$ compared with $h_0(x)$" fails a Pierson's Chi Square Test. Let $p[\chi^2_{10}]$ be the p-val for chi-squared, with χ^2_{10} the Pierson's statistic for $h_1(x)$ is compared with $h_0(x)$, and similarly with $p[\chi^2_{20}]$. Then define a parameter,

$$c = \max \{ p[\chi^2_{10}], p[\chi^2_{20}] \}$$

that is the maximum of the two p-vals. Now find the function a theoretical histogram $h_0(x)$ that minimized c:

minimize c with respect to $h_0(x)$

if the minimal c is less than 0.95, then the two observed histograms are not significantly different.

I would suppose that the minimal $h_0(x)$ would be close to the average of $h_1(x)$ and $h_2(x)$:

$$h_0(x) \sim [h_1(x) + h_2(x)]/2$$

The attached code tests these ideas for a histogram with four bins, using a grid search. The minimal $h_0(x)$ is close to the average of $h_1(x)$ and $h_2(x)$ in this test case.

A grid search will be impractical for histograms with many bins. In that case, it might be possible to define

$$c' = \{ p[\chi^2_{10}]^N + p[\chi^2_{20}]^N \}^{1/N}$$

which satisfies $c' \to c$ in the limit of $N \to \infty$, and then minimize c' with respect to h_0 by a gradient method, choosing some large N. One could use the average as the starting h_0.

```
clear all;
x = [1, 2, 3, 4]';
Dx = 1;
N=4;

% first histogram
h1 = [34, 6, 2, 0]';
N1 = sum(h1);

% second histogram, but normalized to the same number of total counts as
% the first
h2 = [6, 9, 10, 13]';
N2 = sum(h2);
h2 = h2 * N1 / N2;

% Now answer the question:
% Is there a theoretical histogram such that
% the null hypothesis that the difference between the theoretical
% histogram and an observed histogram arises due to random variation
% CANNOT be rejected to greater than 95% certainty
% for both observed histograms

% or more simply put
% Could both observed historgams be "the same" in these sense
% of arising out of random variation from some common theoretical
% histogram?

% loop over all possible theoretical histograms (but with integer counts)
% calculate Pval for null hypothesis for both observed histograms
% minimize the larger Pval of the pair over all possible theoretical
% histograms

% I would expect that "minimal" theoretical histogram is about halfway
% between the two observed histograms ... which seems to be the case

nu = N-1;
pvmin = 1.0;
h0min = [1, 1, 1, 1];
for i1 = [1:N1-1]
for i2 = [1:N1-1]
for i3 = [1:N1-1]
    k=N1-(i1+i2+i3);
    if( k > 0 )
        h0 = [ i1, i2, i3, k ]';
        c1 = sum((( h1 - h0) .^ 2) ./ h0);
        c2 = sum((( h2 - h0) .^ 2) ./ h0);
        pv1 = 1 - chi2pdf(c1,nu);
        pv2 = 1 - chi2pdf(c2,nu);
        pv = max( [pv1, pv2]' );
        if( pv < pvmin )
            pvmin = pv;
            h0min = h0;
        end
    end
end
```

```
end
end
fprintf('H1: %d %d %d %d\n', h1(1), h1(2), h1(3), h1(4) );
fprintf('H2: %d %d %d %d\n', floor(h2(1)), floor(h2(2)), floor(h2(3)),
floor(h2(4)) );
fprintf('HA: %d %d %d %d\n', floor((h1(1)+h2(1))/2), floor((h1(2)+h2(2))/2),
floor((h1(3)+h2(3))/2), floor((h1(4)+h2(4))/2) );
fprintf('HT: %d %d %d %d\n', h0min(1), h0min(2), h0min(3), h0min(4) );

fprintf('Piersons p-value %f\n', pvmin);
```

```
>> rfhist2

H1 (observed): 34 6 2 0

H2 (observed): 6 9 11 14

HA (average) 20 7 6 7

HT (minimal theoretical): 21 6 7 8

Piersons p-value 0.999913 (so two observed histograms are significantly different)
```