ヒストグラムの適合度の尤度比検定*

吉田 正俊†

2008 年 11 月 13 日

以下の問題を解こうとして、ネットを調べてたんですが、よくわからないので自作してみました。これでいいんでしょうか? こういう問題です：

Observed data が N 個あってそれのヒストグラムを作りました。このヒストグラムを fitting してやるために二種類のモデルによるヒストグラムを作りました。モデル $H0$ とモデル $H1$ があって、$H1$ は $H0$ のモデルにパラメータをひとつ付加したもの。それぞれのモデルからデータを generate してやって、ヒストグラムを作る。

モデル $H0$ による fitting の適合度とモデル $H1$ による fitting の適合度とを比べて有意差があるか知りたい。そこで、ふたつの適合度の尤度比検定をしてやろうというわけです。

それぞれの fitting の尤度関数さえ計算できれば、モデル間で差がないとする帰無仮説において「-2*loglikelihood の差」が自由度 1($H0$ と $H1$ での自由度の差) のカイ二乗分布に従うことを使って検定できます。

(なお、全部ヒストグラムではなくて確率密度関数にしても成り立つはず。あと、要はモデル選択なので、AIC とか BIC でやるかという話もあるけど、とりあえず簡単のため尤度比検定で。)

Obserbed data は N 個：

$$data = \{data(1), data(2), data(3), ...data(N)\} \tag{1}$$

ヒストグラムの bin は n 個：

$$\{x(1), x(2), x(3), ...x(n)\} \tag{2}$$

Observed data、モデル H0 での best-fitted data、モデル H1 での best-fitted data のそれぞれのヒストグラムでの各 bin の頻度：

$$N_O(x(i)) = \{N_O(x(1)), N_O(x(2)), ..., N_O(x(n))\} \tag{3}$$
$$N_{E0}(x(i)) = \{N_{E0}(x(1)), N_{E0}(x(2)), ..., N_{E0}(x(n))\} \tag{4}$$
$$N_{E1}(x(i)) = \{N_{E1}(x(1)), N_{E1}(x(2)), ..., N_{E1}(x(n))\} \tag{5}$$

それぞれのヒストグラムのデータの総数は N になるように揃えてあります。(データ数と同じになるように fitting のときに使った simulation のヒストグラムを normalize してある。)

---

* 元の記事：http://pooneil.sakura.ne.jp/archives/permalink/001153.php

† 所属：生理学研究所・発達生理学研究系・認知行動発達研究部門 連絡先：pooneil68@gmail.com

$$\sum_{i=1}^{n} N_O(x(i)) = \sum_{i=1}^{n} N_{E0}(x(i)) = \sum_{i=1}^{n} N_{E1}(x(i)) = N \tag{6}$$

まずそれぞれの fitting の尤度関数を作ってみる。まず Observed data のある 1 個 $data(j)$ がヒストグラムの bin x(i) に落ちるとすると、$data(j)$ がモデル $H0$ の分布の bin $x(i)$ に落ちる確率は

$$\begin{aligned} P(data(j)|H0) &= \frac{N_{E0}(x(i))}{\sum_{i=1}^{n} N_{E0}(x(i))} \\ &= \frac{N_{E0}(x(i))}{N} \end{aligned} \tag{7}$$

と書けます。だから、尤度関数 $L(H0)$ は全 observed data の数だけこれを掛け合わせたものです。Observed data の

$$data = \{data(1), data(2), data(3), ...data(N)\} \tag{8}$$

がそれぞれ落ちる bin がたとえば

$$\{x(m), x(n), x(q), x(r)\} \tag{9}$$

とかだとすると、

$$\begin{aligned} L(H0) &= P(data(1)|H0) * P(data(2)|H0) * ... * P(data(n)|H0) \\ &= \frac{N_{E0}(x(m))}{N} * \frac{N_{E0}(x(n))}{N} * \frac{N_{E0}(x(q))}{N} * ... * \frac{N_{E0}(x(r))}{N} \\ &= N_{E0}(x(m)) * N_{E0}(x(n)) * N_{E0}(x(q)) * ... * N_{E0}(x(r)/(N^N) \end{aligned} \tag{10}$$

といったかんじになります。これを log-likelihood に変換すると、

$$\begin{aligned} LL(H0) = ln(N_{E0}(x(m))) + ln(N_{E0}(x(n))) + ln(N_{E0}(x(q))) + \\ ... + ln(N_{E0}(x(r)) - N * ln(N) \end{aligned} \tag{11}$$

です。全 observed data は $x(i)$ の bin のどこかに落ちるから、たとえば、$x(1)$ の bin に落ちる observed data は $N_O(x(1))$ 個ある、というふうに整理すると、

$$\begin{aligned} LL(H0) = N_O(x(1)) * ln(N_{E0}(x(1))) + N_O(x(2)) * ln(N_{E0}(x(2))) + \\ ... + N_O(x(n)) * ln(N_{E0}(x(n))) - N * ln(N) \end{aligned} \tag{12}$$

となります。これを bin ごとに足し合わせると

$$LL(H0) = \sum_{i=1}^{n} (N_O(x(i)) * ln(N_{E0}(x(i)))) - N * ln(N) \tag{13}$$

と書けます。同様に H1 のモデルのときは

$$LL(H1) = \sum_{i=1}^{n} (N_O(x(i)) * ln(N_{E1}(x(i)))) - N * ln(N) \tag{14}$$

となります。あとは 2LL を作るだけ。二番目の項が消えます。

$$\begin{aligned} 2LL &= -2 * (LL(H1) - LL(H0)) \\ &= 2 * \sum_{i=1}^{n} (N_O(x(i)) * ln(N_{E0}(x(i)))) - 2 * \sum_{i=1}^{n} (N_O(x(i)) * ln(N_{E1}(x(i)))) \end{aligned} \tag{15}$$

$\sum$ はどちらも共通の bin $x(i)$ で足し合わせているから合体できます。

$$2LL = 2 * \sum_{i=1}^{n} \left(N_O(x(i)) * ln \frac{N_{E0}(x(i))}{N_{E1}(x(i))} \right) \tag{16}$$

ということで計算できました。これで合ってるのか答えがネットを探していても見つからないのだけれど、G-test や KL diveregence の値と似ているから、間違った方向には行ってないでしょう。ということで検算のために G-test の式に近づけてみましょう。上記の 2LL 式を変形してやると、

$$\begin{aligned} 2LL &= 2 * \sum_{i=1}^{n} \left(N_O(x(i)) * ln\{ \frac{N_{E0}(x(i))}{N_{E1}(x(i))} * \frac{N_O(x(i))}{N_O(x(i))} \} \right) \\ &= 2 * \sum_{i=1}^{n} \left(N_O(x(i)) * \{ ln \frac{N_{E0}(x(i))}{N_O(x(i))} - ln \frac{N_{E1}(x(i))}{N_O(x(i))} \} \right) \\ &= -2 * \sum_{i=1}^{n} \left(N_O(x(i)) * \{ ln \frac{N_O(x(i))}{N_{E0}(x(i))} - ln \frac{N_O(x(i))}{N_{E1}(x(i))} \} \right) \\ &= 2 * \sum_{i=1}^{n} \left(N_O(x(i)) * ln \frac{N_O(x(i))}{N_{E1}(x(i))} \right) - 2 * \sum_{i=1}^{n} \left(N_O(x(i)) * ln \frac{N_O(x(i))}{N_{E0}(x(i))} \right) \\ &= G(H1) - G(H0) \end{aligned} \tag{17}$$

となって、二つの G 検定値の差となっています。なんてこった orz ちなみにモデル H0 で observed data を fitting したときの G 検定の G-statistics は、

$$G\,(H0) = 2 * \sum_{i=1}^{n} \left(N_O(x(i)) * ln \frac{N_O(x(i))}{N_{E0}(x(i))} \right) \tag{18}$$

となります。というわけでたぶんこんな長々と計算しなくても、二つの fitting をして、G-statistics を計算して、その差をカイ二乗検定すれば良かったというオチ、のようです。じつははじめに計算をしたときは、G statistics ではなくてカイ二乗の方を使ったのだけれど、カイ二乗の差をまたカイ二乗検定、ってなんかへんではないか?と思って上のような計算をしてみた次第。あと、G-test 自体が observed data と fitted data とのあいだでの尤度比検定なのは知っていたのですが、二つのモデルの比較で単純にそれらをさし引いて良いかどうかがわからなかったわけです。以上の計算からすると、最尤推定の考えからしてもさし引いて良い、ということになりそうですが。

ま、いつも通り自分で疑問出して自分で納得してるってかんじなのですが、まだ納得しているわけではないんです。というのも、計算される値が大きすぎるし、ググっても、関係してくる項目が見つからない。まだなんか間違えている気がします。

また、G 検定はいわば KL divergence の離散バージョンと式の上では同じですから (ところでこのことを web で探しても明確に書かれているものを見つけられない)、情報幾何で使われるイメージ化の方法が使えます。ここでしている-2LL のカイ二乗検定というのは、[Observed data -¿ model H0 での best fitted data] の距離と [Observed data -¿ model H1 での best fitted data]) の距離との差を取っているということになります。たとえば「神経回路網と EM アルゴリズム」とかにあるようなとり扱いをすると、observed data の集団を多様体の中のある一点として、H0 のモデルに属した曲面と H1 のモデルに属した曲面とがあって、observed data の点からそれぞれの曲面に垂線を下ろした当たったところがそれぞれのモデルでの best-fitted data の集団の位置で、その垂線 (正確には m 測地線) の距離が KL-diveregence です。

さて、そうするとわたしがわからないのは、そういったまっすぐでない空間での二つの KL-divergence をたんにさっ引いてよいのかっていう問題になるのかも知れません。ていうかそこまでいくと、2LL の原理を調べろ、ってことでネイマンーピアソンまで戻って勉強しないといけない。こういうことやってると JSTOR にある昔の統計学の論文とか読み出してどんどんはまることになるので、このへんまでにしておきます。