統計関連の関数について

ブライアン・ガウ
(とみながだいすけ訳)

November 2008
(translation: Feb 23, 2010)

1 重み付きの平均と分散

正規分布 $G(\mu, \sigma)$ (または一次および二次のモーメントが有限値をとる他の分布) から得られた N 個の標本 x_i があったとする。各標本には w_i という重みが付いている。この重みの値は、重み同士での相対的な値にのみ意味がある。重み付き平均値の推定量が以下のように与えられる時、

$$\bar{x} = \frac{1}{W}\sum_i w_i x_i \tag{1}$$

$$W = \sum_i w_i \tag{2}$$

母集団の分散 σ^2 の不偏推定量を計算することを考える。

まず標本分散 V を定義し、偏りの修正量を計算する。

$$\begin{aligned}
V &= \frac{1}{W}\sum_i w_i (x_i - \bar{x})^2 &(3)\\
&= \frac{1}{W}\sum_i w_i \left(x_i - \frac{1}{W}\sum_j w_j x_j\right)^2 &(4)\\
&= \frac{1}{W}\sum_i w_i \left(x_i^2 - \frac{2}{W} x_i \sum_j w_j x_j + \frac{1}{W^2}(\sum_j w_j x_j)^2\right) &(5)\\
&= \frac{1}{W}\left(\sum_i w_i x_i^2 - \frac{2}{W}\sum_i w_i x_i \sum_j w_j x_j + \frac{1}{W}(\sum_j w_j x_j)^2\right) &(6)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{W}\left(\sum_i w_i x_i^2 - \frac{1}{W}\sum_i w_i x_i \sum_j w_j x_j\right) && (7)\\
&= \frac{1}{W}\left(\sum_i w_i x_i^2 - \frac{1}{W}\sum_{ij} w_i w_j x_i x_j\right) && (8)
\end{aligned}
$$

期待値を $\langle V\rangle$ とすると、正規分布の場合 $\langle x_i\rangle = \mu$、$\langle x_i x_j\rangle = \mu^2 + \delta_{ij}\sigma^2$ であることからその値が得られる。ここで重みの値そのものにはランダムな要素はない、もしくはあっても σ に比べて無視できるほど小さいとする。

$$
\begin{aligned}
\langle V\rangle &= \frac{1}{W}\left(\sum_i w_i\langle x_i^2\rangle - \frac{1}{W}\sum_{ij} w_i w_j \langle x_i x_j\rangle\right) && (9)\\
&= \frac{1}{W}\left(\sum_i w_i(\mu^2+\sigma^2) - \frac{1}{W}\sum_{ij} w_i w_j(\mu^2+\delta_{ij}\sigma^2)\right) && (10)\\
&= \frac{1}{W}\left(W(\mu^2+\sigma^2) - \frac{1}{W}(W^2\mu^2 + (\sum_i w_i^2)\sigma^2)\right) && (11)\\
&= \frac{1}{W}\left(W\sigma^2 - \frac{1}{W}(\sum_i w_i^2)\sigma^2\right) && (12)\\
&= \left(\frac{W^2 - \sum_i w_i^2}{W^2}\right)\sigma^2 && (13)
\end{aligned}
$$

したがって σ^2 の不偏推定量 U は以下のようになる。

$$
\begin{aligned}
U &= \frac{W^2}{(W^2 - \sum_i w_i^2)}\langle V\rangle && (14)\\
&= \frac{W^2}{(W^2 - \sum_i w_i^2)}\frac{1}{W}\sum_i w_i(x_i-\bar{x})^2 && (15)\\
&= \frac{W}{(W^2 - \sum_i w_i^2)}\sum_i w_i(x_i-\bar{x})^2 && (16)
\end{aligned}
$$

これが GSL で用いられている計算式である。

1.1 備考

これには以下のような特徴がある。

- 重みのスケーリングを変えても式は変わらない。
- 重みの値が全て同じ、つまり $w_i = w$ のように書ける場合、$N/(N^2 - N) = 1/(N-1)$ となり、重みがない場合の分散の不偏推定量と同じになる。
- $V = (1/W)\sum_i w_i(x_i - \bar{x})^2$ が重み付き分散の推定値としてよく使われているが、$\sum_i (w_i/W)^2 \ll 1$ のときは、この式がいい近似値を与える。

重み自体に何らかの「実験誤差」があると考えられるような場合は、ここで示した式による σ^2 の推定値は大きくなりすぎる恐れがある (たとえば母集団の真の分散が $\sigma = 0$ だったとすると、w_i による値が標本のばらつきであるかのように扱われて分散が計算される)。この場合は $\langle x_i x_j \rangle = \mu^2 + \delta_{ij}(\sigma^2 + 1/w_i)$ がいい推定値となる。