2012年6月15日　統計数理研究所　オープンハウス

Generalization of t-statistic

Osamu Komori　School of Statistical Thinking,　Project Assistant Professor

1 Generalized t statistic

For a binary class label $y \in \{0,1\}$, let $\{x_{0i} : i = 1,\dots,n_0\}$ be a sample with $y = 0$ and $\{x_{1j} : j = 1,\dots,n_1\}$ be a sample with $y = 1$, where $n = n_0 + n_1$. Then we propose a generalized t-statistic defined by

$$L_U(\beta) = \frac{1}{n_1}\sum_{j=1}^{n_1} U\left\{\frac{\beta^{\mathrm{T}}(x_{1j} - \bar{x}_0)}{(\beta^{\mathrm{T}} S_0 \beta)^{1/2}}\right\}, \tag{1}$$

where U is an arbitrary real-valued function: $\mathbb{R} \to \mathbb{R}$; $\bar{x}_y$ and S_y are the sample mean and the sample variance given y, respectively. The expectation of $L_U(\beta)$ is defined by

$$\mathbb{L}_U(\beta) = E_1\left[U\left\{\frac{\beta^{\mathrm{T}}(x - \mu_0)}{\beta^{\mathrm{T}}\Sigma_0\beta}\right\}\right], \tag{2}$$

where E_y, μ_y and Σ_y denote the conditional expectation, mean and variance, respectively, given y. For the distribution of the control group ($y = 0$), we assume normality such as

$$x_0 \sim N(\mu_0, \Sigma_0). \tag{3}$$

That is, the information of 0-group population is assumed to be simply reduced to the statistics $\bar{x}_0$ and S_0; while we carefully have to choose U to extract the information of 1-group population. In the cancer data analysis based on the gene expression data, a small part observations of disease group ($y = 1$) is usually over- or down-expressed. To treat this heterogeneity, several types of t-statistics are proposed to individually detect genes that are useful in cancer studies (Tibshirani and Hastie, 2007; Wu, 2007; Lian, 2008).

If we adopt a linear function $U(w) = w$, then the generalized t-statistic becomes the simple t-statistic standardized by S_0:

$$L_{\mathrm{I}}(\beta) = \frac{\beta^{\mathrm{T}}(\bar{x}_1 - \bar{x}_0)}{(\beta^{\mathrm{T}} S_0 \beta)^{1/2}}. \tag{4}$$

When U is the cumulative function of the standard normal distribution: $U(w) = \Phi(w)$, the generalized t-statistic is viewed as c-statistic (area under the ROC curve) because of the normality assumption of 0-group population in (3):

$$L_\Phi(\beta) = \frac{1}{n_1}\sum_{j=1}^{n_1} \Phi\left\{\frac{\beta^{\mathrm{T}}(x_{1j} - \bar{x}_0)}{(\beta^{\mathrm{T}} S_0 \beta)^{1/2}}\right\}, \tag{5}$$

which converges to $\mathrm{pr}(\beta^{\mathrm{T}} x_0 < \beta^{\mathrm{T}} x_1)$ as n_0 and n_1 go to infinity by a conditional expectation argument (Su and Liu, 1993). Hence, the generalized t-statistic is a natural extension of the common statistics such as t-statistic and c-statistic. Moreover, there is some relationship with Fisher linear discriminant function if we choose a specific quadratic function as U, which is discussed in detail later.

2 Asymptotic consistency and normality

Let us consider the estimator associated with the generalized t-statistic as

$$\widehat{\beta}_U = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}}\, L_U(\beta). \tag{6}$$

Then we consider the following assumption:

(A) $\qquad E_1(g \mid w = a) = 0 \quad$ for all $a \in \mathbb{R}$,

where $w = \beta_0^{\mathrm{T}}(x - \mu_0)$, $g = (I - P_0)(x - \mu_0)$ with I being the $p \times p$ unit matrix and $P_0 = \Sigma_0 \beta_0 \beta_0^{\mathrm{T}}$, where

$$\beta_0 = \frac{\Sigma_0^{-1}(\mu_1 - \mu_0)}{\{(\mu_1 - \mu_0)^{\mathrm{T}}\Sigma_0^{-1}(\mu_1 - \mu_0)\}^{1/2}}. \tag{7}$$

Theorem 2.1 *Under Assumption (A), $\widehat{\beta}_U$ is asymptotically consistent with β_0 for any U.*

Next we consider the following assumption in addition to (A):

(B) $\qquad \mathrm{var}_1(g \mid w = a) = \Sigma_0^* \quad$ for all $a \in \mathbb{R}$,

where var_y denotes the conditional variance of x given y and $\Sigma_0^* = (I - P_0)\Sigma_0(I - P_0^{\mathrm{T}})$.

Theorem 2.2 *Under Assumptions (A) and (B), $n_1^{1/2}(\widehat{\beta}_U - \beta_0)$ is asymptotically distributed as $N(0, \Sigma_U)$, where*

$$\Sigma_U = c_U \Sigma_0^*, \tag{8}$$

$$c_U = \frac{E_1\{U'(w)^2\} + \pi_1/\pi_0\left[E_1\{U'(w)w\}\right]^2 + \pi_1/\pi_0\left[E_1\{U'(w)\}\right]^2}{\left[E_1\{U'(w)S(w)\} + E_1\{U'(w)w\}\right]^2}, \tag{9}$$

in which $\pi_0 = \mathrm{pr}(y = 0)$, $\pi_1 = \mathrm{pr}(y = 1)$, $S(w) = \partial \log f_1(w)/\partial w$ and U' denotes the first derivative of U.

Theorem 2.3 *The optimal U function under Assumptions (A) and (B) has the following form:*

$$U_{\mathrm{opt}}(w) = \log \frac{f_1(w)}{\phi(w, \mu_w, \sigma_w^2)}, \tag{10}$$

where $\mu_w = E(w)$ and $\sigma_w^2 = \mathrm{var}(w)$. Moreover, the minimum of c_U is given by

$$\min_U c_U = \frac{\sigma_w^2}{\mu_{1,S^2} - 1 + (\pi_0\mu_{1,w}^2 + \sigma_{1,w}^2 - 1)(\pi_0 + \pi_1\mu_{1,S^2})}, \tag{11}$$

where $\mu_{1,w} = E_1(w)$, $\sigma_{1,w}^2 = E_1\{(w - \mu_{1,w})^2\}$ and $\mu_{1,S^2} = E_1\{S(w)^2\}$.

Remark 2.1 *The expectation of generalized t-statistic based on U_{opt} is equivalent to the Kullback-Leibler divergence given as:*

$$\mathbb{L}_{U_{\mathrm{opt}}}(\beta) = \int f_1(w) \log \frac{f_1(w)}{\phi(w, \mu_w, \sigma_w^2)} dw. \tag{12}$$

That is, the maximization of the generalized t-statistic is considered as the maximization of the Kullback-Leibler divergence.

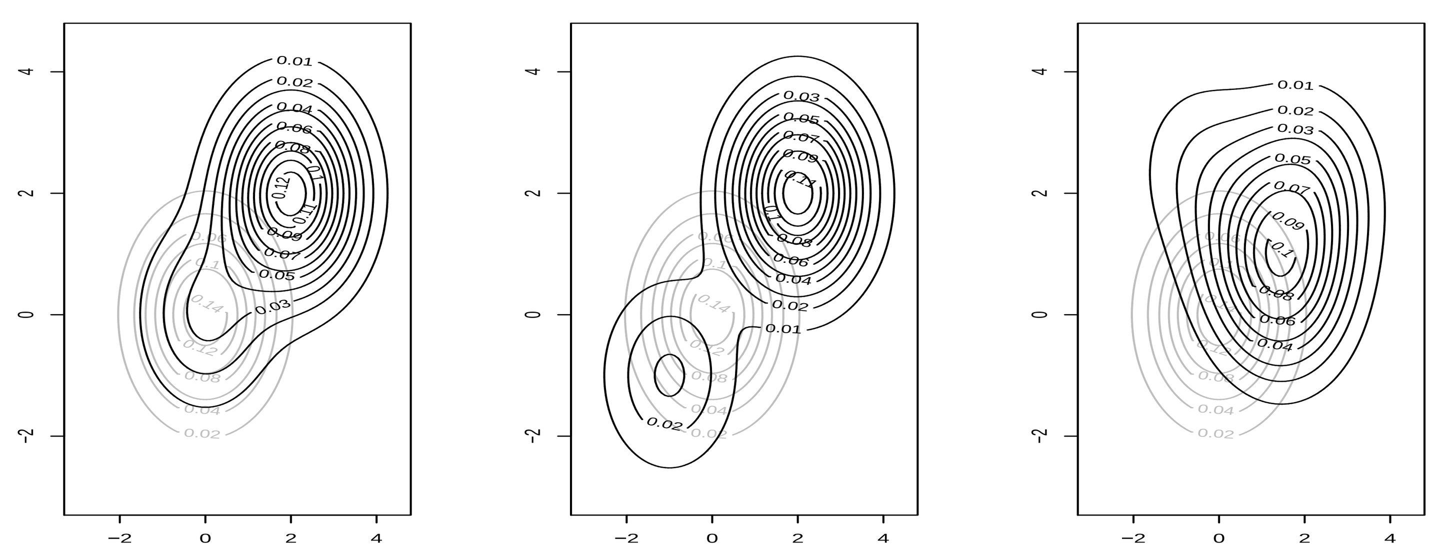

Fig1. Contour plots of probability densities of $y = 0$ in gray and $y = 1$ in black, which satisfy Assumptions (A) and (B).

References

Lian, H. (2008). MOST: detecting cancer differential gene expression. *Biostatistics* **9**, 411–418.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

Tibshirani, R. and Hastie, T. (2007). Outlier sums for differential gene expression analysis. *Biostatistics* **8**, 2–8.

Wu, B. (2007). Cancer outlier differential gene expression detection. *Biostatistics* **8**, 566–575.