Statistical Inferences with Categorized Variables and Its Application to a Trimmed Mean

By Dr. Lin-An Chen and Tzu-Yun Huang

Institute of Statistics, National Chiao Tung University,
Hsinchu, Taiwan.

Abstract

Considerable energy has been devoted to the arguments of possible undesired statistical inference properties resulted from categorization of continuous variables that does not stop its popularity in association research of epidemiology for its appealing of convenience in presentation and interpretation of analyzed results. For correction of popularly used untrustworthy statistical methods, we initiate a theoretical study of statistical effect of categorization with parametric and nonparametric estimations for unknown means of categorized variables. We show that the parametric sample mean is very efficient that explains undesired statistical property of classical statistical methods. In nonparametric estimation of the population mean of a (noncategorized) variable, we prove that categorization creates auxiliary information to improve the efficiency of parameter estimation. This shows that the statistical society is far from knowing the statistical properties of categorization and the supplementary population information of an extra variable created by categorization for statistical inferences deserves to receive more attention in literature.

Key words: Auxiliary information; auxiliary variable; categorization of continuous variable; estimation; trimmed mean.

1. Introduction

It is very common that the researchers are interested in assessing the relationship between continuous outcome and explanatory (covariates) variables. In contemporary epidemiologic practice, it is appealing to epidemiologist to modify continuous variables into categorical variables to facilitate data presentation such as low, medium and high risk group. Prock et al. (2004) reported that 84% of epidemiological papers from leading journals made categorization of continuous variables. Categorization of continuous variables is widespread to other areas, for examples, psychology application (MacCallum, et al. (2002)) and marketing application (Irwin and McClelland (2003)). The investigators often then use categorized samples to fit a regression model or to analyze whether subsequently higher categories are associated with increased risk of an outcome by multiple comparison method.

Categorization of continuous variables has been overwhelmingly criticized for problems of undesired statistical properties such as bias in estimation and power loss in hypothesis testing caused by loss information (see, for examples, Royston and Aauerbrei (2008), Taylor and Yu (2002), Walraven and Hart (2008) and Zhao and Kolonel (1992)). As observed by Han (2008), the assumptions of normality, independence and constant variance behind the multiple comparison for these categorized variables are not true (see also Bennette and Vickers (2012)), and these undesired statistical properties, in our opinion, are not prevented when theoretically untrustworthy statistical inference methods are applied.

With categorization still playing important role, it requires theoretically trustworthy statistical methods to deal with categorized samples. We initiate this study by developing distributional theory for parametric and nonparametric categorized sample means. With novel idea of parametrization, the parametric estimation outperforms the classical sample means with very high efficiency. We also observed a surprising and exceptional new statistical theory that the debated categorization creates the desired auxiliary information.

In the statistical inference for unknown parameters of a variable's distribution, any extra variable measured in association with this variable that is used to increase the accuracy of this inference is called an auxiliary variable. By showing that a categorized trimmed mean for estimating population mean of uncategorized variable has asymptotic variance not only smaller than that of the classical trimmed mean but also, more interestingly, smaller than the Cramer-Rao lower bound for this population mean, an evidence of categorization creating auxiliary information is discovered. The knowledge in literature is slim in terms of how much categorization contributes the accuracies in statistical parameter inferences, especially when categorization's auxiliary information is implemented. This approach has taken the first step to recognize the theory of categorization but there is much more waiting for further investigation.

2. Nonparametric Categorized Sample Means

Let Y and X be continuous response and explanatory variables with a joint probability density function (pdf) $f_{XY}(x, y)$. Consider cutoffs $-\infty < a_1 < a_2 < ... < a_{k-1}$ such that intervals $A_1 = (-\infty, a_1]\,, A_2 = (a_1, a_2], ..., A_k = (a_{k-1}, \infty)$ forms a partition of the space of variable X. Cutoffs a_j's are seen as known constants and unknown quantiles in practice. Suppose that we have a random sample $\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \begin{pmatrix} Y_2 \\ X_2 \end{pmatrix}, ..., \begin{pmatrix} Y_n \\ X_n \end{pmatrix}$ from this underlying distribution. The epidemiologists often categorize the sample of variable Y into the following categorized samples, as

$$\{Y_i : X_i \in A_1, i = 1, ..., n\}, \quad \ldots, \quad \{Y_i : X_i \in A_k, i = 1, ..., n\}\,. \tag{2.1}$$

Then classical statistical methods such as t-test and F-test based on these categorized samples for inference of unknown population means $\theta_1 = E[Y|X \in A_1], ..., \theta_k = E[Y|X \in A_k]$, called the categorized means, are applied to verify the relationship between categorized variables. As observed from Han (2008), these categorized variables in k groups are no longer normal, independent and constant variance and then these classical tests are theoretically incorrect.

Theoretically trustworthy inference methods may be developed from the distributional theory of their parametric and nonparametric estimators. We consider the nonparametric estimation in this section. For constant cutoffs, the following averages

$$\hat{\theta}_{cj} = \frac{\sum_{i=1}^n Y_i I(X_i \in A_j)}{\sum_{i=1}^n I(X_i \in A_j)}, j = 1, ..., k, \tag{2.2}$$

called the categorized sample means, are applied without correct distributional theory, in classical ANOVA approach. Here c stands for constant cutoff. Denoting $\hat{\theta}_C = (\hat{\theta}_{c1}, ..., \hat{\theta}_{ck})'$, a nonparametric estimator of vector categorized group means $\theta = (\theta_1, ..., \theta_k)'$, the following theorem states its distributional theory.

Theorem 2.1. $n^{1/2}(\hat{\theta}_C - \theta)$ converges in distribution to k-dimensional multivariate normal distribution $N_k(0_k, \Sigma_C)$ where

$$\Sigma_C = \mathrm{Cov}\begin{pmatrix} p_1^{-1}(Y-\theta_1)I(X \le a_1) \\ p_2^{-1}(Y-\theta_2)I(X \in (a_1, a_2)) \\ \vdots \\ p_k^{-1}(Y-\theta_k)I(X \ge a_k) \end{pmatrix} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_k^2)$$

and where $\sigma_j^2 = \frac{1}{P(X \in A_j)} Var(Y|X \in A_j)$ is the categorized variance.

If choices in nonparametric estimation are available, we recommend the constant cutoffs since estimation of asymptotic covariance matrix is simpler to establish. We define estimator of categorized variance σ_j^2 by

$$S_j^2 = \frac{1}{\sum_{i=1}^n I(X_i \in A_j)} \sum_{i=1}^n (Y_i - \hat{\theta}_j)^2 I(X_i \in A_j), j = 1, ..., k,$$

calling them the categorized sample variances. Hence the following sample matrix

$$\hat{\Sigma}_C = \mathrm{diag}(S_1^2, S_2^2, ..., S_k^2)$$

consitutes a consistent estimator of unknown covariance matrix Σ_C. Its efficiency will be verified later.

In many applications (Shankar et al. (2007), Luo et al. (2007) and Letenneur et al. (2007)), categorization is done on quantile partition as

$$A_1 = (-\infty, F_X^{-1}(\alpha_1)], A_2 = (F_X^{-1}(\alpha_1), F_X^{-1}(\alpha_2)], ..., A_k = (F_X^{-1}(\alpha_{k-1}), \infty)$$

where $F_X^{-1}(\alpha)$ represents the α-th population quantile of random variable X. Frequently the quantile functions $F_X^{-1}(\alpha_j)'s$ are unknown and are estimated with empirical quantiles $\hat{F}_X^{-1}(\alpha_j), j = 1, ..., k-1$, of random sample $X_1, ..., X_n$. This leads to the sample partition

$$\hat{A}_1 = (-\infty, \hat{F}_X^{-1}(\alpha_1)], \hat{A}_2 = (\hat{F}_X^{-1}(\alpha_1), \hat{F}_X^{-1}(\alpha_2)], ..., \hat{A}_k = (\hat{F}_X^{-1}(\alpha_{k-1}), \infty),$$

and the quantiles based categorized sample means as

$$\hat{\theta}_{qj} = \frac{\sum_{i=1}^{n} Y_i I(X_i \in \hat{A}_j)}{\sum_{i=1}^{n} I(X_i \in \hat{A}_j)}, j = 1, ...k \tag{2.3}$$

where q stands for quantile cutoff. We denote $\hat{\theta}_q = (\hat{\theta}_{q1}, ..., \hat{\theta}_{qk})'$ and $\lambda_j = E(Y - \mu_y | X = F_X^{-1}(\alpha_j))$. A representation and asymptotic distribution for this categorized sample mean vector $\hat{\theta}_q$ are introduced below.

Theorem 2.2. (a) The quantile cutoffs based categorized sample mean has the following Bahadur representation:

$$\sqrt{n}(\hat{\theta}_q - \theta) = n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} \alpha_1^{-1}(\psi_1(Y_i, X_i) - E(\psi_1(Y, X))) \\ (\alpha_2 - \alpha_1)^{-1}(\psi_2(Y_i, X_i) - E(\psi_2(Y, X))) \\ \vdots \\ (1 - \alpha_{k-1})^{-1}(\psi_k(Y_i, X_i) - E(\psi_k(Y, X))) \end{pmatrix} + o_p(1)$$

where

$$\psi_1(Y, X) = \begin{cases} Y - \mu_y & \text{if } X \leq F_X^{-1}(\alpha_1) \\ \lambda_1 & \text{if } X > F_X^{-1}(\alpha_1) \end{cases}, \psi_k(Y, X) = \begin{cases} \lambda_{k-1} & \text{if } X \leq F_X^{-1}(\alpha_{k-1}) \\ Y - \mu_y & \text{if } X > F_X^{-1}(\alpha_{k-1}) \end{cases}$$

and, for $j = 2, ..., k - 1$,

$$\psi_j(Y, X) = \begin{cases} \lambda_{j-1} & \text{if } X \leq F_X^{-1}(\alpha_{j-1}) \\ Y - \mu_y & \text{if } F_X^{-1}(\alpha_{j-1}) < X < F_X^{-1}(\alpha_j) \\ \lambda_j & \text{if } X \geq F_X^{-1}(\alpha_j) \end{cases}.$$

(b) We have that $\sqrt{n}(\hat{\theta}_q - \theta)$ is asymptotically normal with distribution $N_k(0_k, \Sigma_q)$ where $k \times k$ matrix $\Sigma_q = (\sigma_{jm}), j, m = 1, ..., k$ with

$$\begin{aligned} \sigma_{11} =& \frac{1}{\alpha_1^2}\{M_1 + (1 - \alpha_1)\lambda_1^2 - (m_1 + (1 - \alpha_1)\lambda_1)^2\}, \\ \sigma_{1j} =& \frac{1}{\alpha_1(\alpha_j - \alpha_{j-1})}\{\lambda_1\lambda_{j-1}(\alpha_{j-1} - \alpha_1) + \lambda_1 m_j + \lambda_1\lambda_j(1 - \alpha_j) \\ & + \lambda_{j-1}m_1 - (m_1 + (1 - \alpha_1)\lambda_1)(\alpha_{j-1}\lambda_{j-1} + (1 - \alpha_j)\lambda_j + m_j)\} \\ \sigma_{1k} =& \frac{1}{\alpha_1(1 - \alpha_{k-1})}\{\lambda_1\lambda_{k-1}(\alpha_{k-1} - \alpha_1) + \lambda_1 m_k + \lambda_{k-1}m_1 \\ & - (m_1 + (1 - \alpha_1)\lambda_1)(\alpha_{k-1}\lambda_{k-1} + m_k)\} \end{aligned}$$

for $j = 2, ..., k-1$,

$$\begin{aligned}
\sigma_{jj} =& \frac{1}{(\alpha_j - \alpha_{j-1})^2}\{\alpha_{j-1}\lambda_{j-1}^2 + (1-\alpha_j)\lambda_j^2 + M_j - (\alpha_{j-1}\lambda_{j-1} \\
&+ (1-\alpha_j)\lambda_j + m_j)^2\}, \\
\sigma_{jj+1} =& \frac{1}{(\alpha_{j+1} - \alpha_j)(\alpha_j - \alpha_{j-1})}\{\lambda_j(\lambda_{j-1}\alpha_{j-1} + m_j + m_{j+1} + \lambda_{j+1}(1-\alpha_{j+1})) \\
&- (\alpha_{j-1}\lambda_{j-1} + (1-\alpha_j)\lambda_j + m_j)(\alpha_j\lambda_j + (1-\alpha_{j+1})\lambda_{j+1} + m_{j+1})\}, \\
\sigma_{jm} =& \frac{1}{(\alpha_j - \alpha_{j-1})(\alpha_m - \alpha_{m-1})}\{\lambda_{m-1}(\lambda_{j-1}\alpha_{j-1} + m_j + \lambda_j(\alpha_{m-1} - \alpha_j)) \\
&+ \lambda_j(m_m + \lambda_m(1-\alpha_m)) - (\alpha_{j-1}\lambda_{j-1} + (1-\alpha_j)\lambda_j + m_j) \\
&(\alpha_{m-1}\lambda_{m-1} + (1-\alpha_m)\lambda_m + m_m)\}, m = j+2, ..., k-1, \\
\sigma_{jk} =& \frac{1}{(\alpha_j - \alpha_{j-1})(1-\alpha_{k-1})}\{\lambda_{k-1}(\lambda_{j-1}\alpha_{j-1} + m_j + \lambda_j(\alpha_{k-1} - \alpha_j)) \\
&+ \lambda_j m_k - (\alpha_{j-1}\lambda_{j-1} + (1-\alpha_j)\lambda_j + m_j)(\alpha_{k-1}\lambda_{k-1} + m_k)\},
\end{aligned}$$

and

$$\sigma_{kk} = \frac{1}{(1-\alpha_{k-1})^2}\{\alpha_{k-1}\lambda_{k-1}^2 + M_k - (\alpha_{k-1}\lambda_{k-1} + m_k)^2\}.$$

where $m_j = E[(Y-\mu_y)I(X \in A_j)]$ and $M_j = E[(Y-\mu_y)^2 I(X \in A_j)]$, for $j = 1, ..., k$ denote the first and second central group moment at jth group for categorized variable Y's.

This theorem generalizes the theory of univariate robust trimmed mean and the outlier mean of Chen, Chen and Chan (2010) to vector case.

Theorems 2.1 and 2.2 provide a basis for theoretically correct nonparametric inferences for unknown parameters θ.

3. Parametric Categorized Sample Means

We consider the parametric approach with normality assumption as

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}). \tag{3.1}$$

In this normal setting, we fix a permutation of distributional parameters as $\Lambda = (\mu_x, \mu_y, \sigma_y^2, \sigma_x^2, \sigma_{yx})$ and consider for simplicity of presentation only quantile cutoffs. Given $0 < \alpha_1 < \alpha_2 < ... < \alpha_{k-1} < 1$, the unknown quantiles under normality assumption are $a_j = F_x^{-1}(\alpha_j) = \mu_x + z_{\alpha_j}\sigma_x$, $j = 1, ..., k-1$. The following theorem is a simplified result from Han (2005).

Theorem 3.1. Consider the quantiles cutoffs and the normal assumption (3.1). The population categorized means forms a vector $\theta_p = (\theta_{1p}, ..., \theta_{kp})$ with

$$\begin{aligned}
&\theta_{1p} = \mu_y - \frac{\sigma_{yx}}{\alpha_1\sigma_x}\phi(z_{\alpha_1}) \\
&\vdots \\
&\theta_{jp} = \mu_y - \frac{\sigma_{yx}}{\sigma_x(\alpha_j - \alpha_{j-1})}(\phi(z_{\alpha_j}) - \phi(z_{\alpha_{j-1}})), j = 2, ..., k-1 \\
&\vdots \\
&\theta_{kp} = \mu_y + \frac{\sigma_{yx}}{\sigma_x(1-\alpha_{k-1})}\phi(z_{\alpha_{k-1}})
\end{aligned}$$

where ϕ is the probability density function (pdf) of standard normal $N(0,1)$.

Suppose that we have a random sample $\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \begin{pmatrix} Y_2 \\ X_2 \end{pmatrix}, ..., \begin{pmatrix} Y_n \\ X_n \end{pmatrix}$. Denoting $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i, S_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2, S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ and $S_{YX} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$, the mle of parameter vector Λ is

$$\hat{\Lambda}_{mle} = (\bar{X}, \bar{Y}, S_Y^2, S_X^2, S_{YX}) \tag{3.2}$$

and the maximum likelihood estimators (mle) of categorized means are

$$\begin{aligned}
&\hat{\theta}_{1p} = \bar{Y} - \frac{s_{yx}}{\alpha_1 s_x}\phi(z_{\alpha_1}), \hat{\theta}_{jp} = \bar{Y} - \frac{s_{yx}}{s_x(\alpha_j - \alpha_{j-1})}\phi(z_{\alpha_j}) - \phi(z_{\alpha_{j-1}})), \\
&j = 2, ..., k-1, \hat{\theta}_{kp} = \bar{Y} + \frac{s_{yx}}{s_x(1-\alpha_{k-1})}\phi(z_{\alpha_{k-1}})
\end{aligned}$$

and its vector $\hat{\theta}_p = (\hat{\theta}_{1p}, ..., \hat{\theta}_{kp})'$.

Theorem 3.2. (a) We have that $n^{1/2}(\hat{\theta}_p - \theta_p)$ converges in distribution to k-dimensional normal distribution $N_k(0_k, \Sigma_p)$ with asymptotic covariance

1896

matrix $\Sigma_p = \Gamma(\Lambda)V_p(\Lambda)\Gamma(\Lambda)'$ where $\Gamma(\Lambda) = \frac{\partial\theta_p(\Lambda)}{\partial\Lambda}$ is the partial derivative of $\theta_p(\Lambda)$ with respect to Λ and $V_p(\Lambda) = -[E\frac{\partial^2 ln\phi_N(X,Y)}{\partial\Lambda\partial\Lambda'}]^{-1}$ is the Crammer-Rao's lower bound for Λ with $\phi_N(X,Y)$ the pdf of normal distribution in (3.1).

(b) The quantiles based cutoffs based partial derivative matrix under the normal distribution is $\Gamma(\Lambda) = (\gamma_{ij})_{i=1,...,k,j=1,...,5}$ with

$$\gamma_{11} = -\frac{\sigma_{xy}}{\sigma_x^2}\frac{\phi(z_{\alpha_1})}{\alpha_1}(\frac{\phi(z_{\alpha_1})}{\alpha_1} + z_{\alpha_1}),\ \gamma_{12} = 1,$$

$$\gamma_{13} = -\frac{\sigma_{xy}}{2\sigma_x^3}\frac{\phi(z_{\alpha_1})}{\alpha_1}(\frac{z_{\alpha_1}\phi(z_{\alpha_1})}{\alpha_1} + z_{\alpha_1}^2 - 1),\ \gamma_{14} = 0,\ \gamma_{15} = -\frac{1}{\sigma_x}\frac{\phi(z_{\alpha_1})}{\alpha_1}$$

for $j = 2, ..., k-1$,

$$\gamma_{j1} = \frac{-\sigma_{xy}}{\sigma_x^2}\frac{1}{\alpha_j - \alpha_{j-1}}(\frac{(\phi(z_{\alpha_j}) - \phi(z_{\alpha_{j-1}}))^2}{\alpha_j - \alpha_{j-1}} + (z_{\alpha_j}\phi(z_{\alpha_j}) - z_{\alpha_{j-1}}\phi(z_{\alpha_{j-1}}))),$$

$$\gamma_{j2} = 1,\ \gamma_{j3} = \frac{-\sigma_{xy}}{2\sigma_x^3}\frac{1}{\alpha_j - \alpha_{j-1}}[(\frac{z_{\alpha_j}\phi(z_{\alpha_j}) - z_{\alpha_{j-1}}\phi(z_{\alpha_{j-1}})}{\alpha_j - \alpha_{j-1}} - 1)$$

$$(\phi(z_{\alpha_j}) - \phi(z_{\alpha_{j-1}})) + (z_{\alpha_j}^2\phi(z_{\alpha_j}) - z_{\alpha_{j-1}}^2\phi(z_{\alpha_{j-1}}))],\ \gamma_{j4} = 0,$$

$$\gamma_{j5} = \frac{-1}{\sigma_x}\frac{\phi(z_{\alpha_j}) - \phi(z_{\alpha_{j-1}})}{\alpha_j - \alpha_{j-1}},\ \gamma_{k1} = \frac{\sigma_{xy}}{\sigma_x^2}\frac{\phi(z_{\alpha_{k-1}})}{1 - \alpha_{k-1}}(z_{\alpha_{k-1}} - \frac{\phi(z_{\alpha_{k-1}})}{1 - \alpha_{k-1}}),$$

$$\gamma_{k2} = 1,\ \gamma_{k3} = \frac{\sigma_{xy}}{2\sigma_x^3}\frac{\phi(z_{\alpha_{k-1}})}{(1 - \alpha_{k-1}}(z_{\alpha_{k-1}}^2 - \frac{z_{\alpha_{k-1}}\phi(z_{\alpha_{k-1}})}{1 - \alpha_{k-1}} - 1)$$

$$\gamma_{k4} = 0,\ \gamma_{k5} = \frac{1}{\sigma_x}\frac{\phi(z_{\alpha_{k-1}})}{1 - \alpha_{k-1}}.$$

(c) Defining 2×2 matrix $A = \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix}$ and 3×3 matrix $B = \begin{pmatrix} 2\sigma_x^4 & 2\sigma_{xy}^2 & 2\sigma_x^2\sigma_{xy} \\ 2\sigma_{xy}^2 & 2\sigma_y^4 & 2\sigma_x^2\sigma_{xy} \\ 2\sigma_x^2\sigma_{xy} & 2\sigma_y^2\sigma_{xy} & \sigma_x^2\sigma_y^2 + \sigma_{xy}^2 \end{pmatrix}$, the Cramer-Rao lower bound for bivariate normal parameter vector Λ is

$$V_p(\Lambda) = -[E\frac{\partial^2 ln\phi_N(X,Y)}{\partial\Lambda\partial\Lambda'}]^{-1} = \begin{pmatrix} A & 0_{2\times3} \\ 0_{3\times2} & B \end{pmatrix}.$$

Further parametric statistical inferences for categorized means θ_p can be constructed with the mle of asymptotic covariance matrix as $\hat{\Sigma}_p = \Gamma(\hat{\Lambda}_{mle})V_p(\Lambda_{mle})\Gamma(\hat{\Lambda}_{mle})'$.

4. Comparison of Parametric and Nonparametric Estimators

We would not investigate the accuracies of theoretically correct inference methods constructed by the parametric and nonparametric categorized sample means but would desire at this moment to compare the accuracies of these two estimation methods. We first compare their asymptotic covariance matrices by evaluating the traces of covariance matrices $\Gamma_{pq}(\Lambda)V_p(\Lambda)\Gamma'_{pq}(\Lambda)$ and Σ_c to compute the relative efficiencies of the nonparametric estimator of categorized group means as

$$eff_N = \frac{\min\{tr(\Gamma_{pq}(\Lambda)V_p(\Lambda)\Gamma'_{pq}(\Lambda)),\ tr(\Sigma_c)\}}{tr(\Sigma_c)}.$$

Considering $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix})$, the efficiency values eff_N under several values of parameters are displayed in Table 1.

Table 1. Efficiencies of nonparametric estimator of categorized group means

	$\sigma_{yx} = 0.2$	0.3	0.5	0.7	0.8
$(\sigma_y^2, \sigma_x^2) = (1, 1)$	0.556	0.599	0.672	0.701	0.675
$(\sigma_y^2, \sigma_x^2) = (2, 1)$	0.530	0.561	0.621	0.671	0.689
$(\sigma_y^2, \sigma_x^2) = (1, 2)$	0.530	0.561	0.621	0.670	0.689

Lower values of eff_N supports the parametric estimation of unknown categorized means when the underlying distribution is known. Since method of nonparametric categorized sample mean is applied for classical ANOVA analysis, this parametric estimation from the new parametrized unknown categorized means in Theorem 3.1 deserves attention in application and study with construction of new ANOVA approach of multiple comparison of categorized means.

Setting the normal distribution $N_2(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix})$, we compare $\hat{\Sigma}_p(\Lambda)$ and $\hat{\Sigma}_C$ through simulation for their efficiencies of estimating the common matrix Σ. Suppose that the categorized sample variance at jth replication be denoted as $S^j = (s^j_{i\ell})_{i,\ell=1,\ldots,k}$ and true covariance matrix is $\Sigma = (\sigma_{i\ell})_{i,\ell=1,\ldots,k}$. We define mean squares error (MSE) by

$$MSE = \frac{1}{mk^2}\sum_{j=1}^{m}\sum_{\ell=1}^{k}\sum_{i=1}^{k}(S^j_{i\ell} - \sigma_{i\ell})^2$$

We re-denote the MSE's for nonparametric and parametric estimators, respectively, by MSE_{np} and MSE_p. With replications $m = 10,000$, sample size $n = 50$ and 100 and some values of variances σ_y^2 and σ_x^2, the results of two MSE's are displayed in Table 2.

Table 2. MSE's for parametric estimator of asymptotic covariance matrix $\Gamma V_p \Gamma'$ (4 groups)

Sample size	$\sigma_{yx} = 0.2$	0.3	0.5	0.7	0.8
$(\sigma_x^2, \sigma_y^2) = (1, 1)$					
$n = 50, MSE_{np}$	4.441	4.422	3.647	2.194	1.338
MSE_p	0.079	0.074	0.049	0.023	0.012
$n = 100, MSE_{np}$	2.208	2.322	2.148	1.323	0.800
MSE_p	0.040	0.036	0.024	0.011	0.006
$(\sigma_x^2, \sigma_y^2) = (2, 1)$					
$n = 50, MSE_{np}$	4.725	4.559	4.528	3.673	3.572
MSE_p	0.083	0.081	0.067	0.050	0.041
$n = 100, MSE_{np}$	2.153	2.240	2.376	2.350	2.293
MSE_p	0.042	0.040	0.034	0.025	0.020
$(\sigma_x^2, \sigma_y^2) = (1, 2)$					
$n = 50, MSE_{np}$	18.61	18.41	17.41	14.86	13.18
MSE_p	0.339	0.312	0.268	0.201	0.164
$n = 100, MSE_{np}$	8.649	8.990	9.321	8.627	7.800
MSE_p	0.165	0.157	0.134	0.103	0.081

The simulated results show that estimation of asymptotic covariance matrix of parametric categorized sample means is much more efficient than that of nonparametric version.

We next consider a simulation study to verify the finite sample efficiencies of parametric and nonparametric estimators of parameter vector θ_p when Y and X have a joint normal distribution. Denoting $\hat{\theta}_N^j$ and $\hat{\theta}_p^j$ as, respectively, nonparametric and parametric estimates of θ at jth replication, we compute the following MSE's

$$MSE_N = \frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_N^j - \theta_p)'(\hat{\theta}_N^j - \theta_p), \;\; MSE_p = \frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_p^j - \theta_p)'(\hat{\theta}_p^j - \theta_p)$$

and the simulated results are displayed in Table 3 where categorization number is 4.

Table 3. MSE's for parametric and nonparametric estimations

Sample size	$\sigma_{yx} = 0.2$	0.3	0.5	0.7	0.8
$n = 30$					
MSE_N	0.145	0.137	0.118	0.087	0.067
MSE_p	0.063	0.059	0.051	0.036	0.027
$n = 50$					
MSE_N	0.082	0.079	0.066	0.049	0.038
MSE_p	0.036	0.034	0.029	0.021	0.016
$n = 100$					
MSE_N	0.039	0.038	0.032	0.023	0.018
MSE_p	0.018	0.017	0.014	0.010	0.007

We see that MSE_p's are all relatively smaller than corresponding MSE_N's that supports our previous observation of superiority of parametric estimation for population categorized means.

Let $\hat{\theta}$ be an estimator of categorized group mean θ that satisfies $n^{1/2}(\hat{\theta} - \theta)$ converging in distribution to normal vector $N_k(0_k, \Sigma)$. Suppose that consistent estimator $\hat{\Sigma}$ for Σ is available. Then following quantity

$$T = n(\hat{\theta} - \theta)'\hat{\Sigma}^{-1}(\hat{\theta} - \theta) \tag{4.1}$$

converges, in distribution, to chi-squares distribution $\chi^2(k)$ of degrees of freedom k when sample size n goes to infinity. Theoretically correct inference methods such as confidence band and test for general linear hypothesis $H_0 : A\theta = 0$ vs $H_1 : A\theta \neq 0$ may be constructed through a chi-square quantity. One way to investigate these inference methods is to compare T like quantity in some way for different approaches. Let the areas of estimated region $(\hat{\theta} - \theta)'\hat{\Sigma}^{-1}(\hat{\theta} - \theta)$ for parametric and nonparametric versions be denoted by $\hat{A}_p$ and $\hat{A}_N$. We denote their MSE's as

$$MSE_p = \frac{1}{m}\sum_{j=1}^{m}(\hat{A}_p^{(j)} - A_p)^2, \; MSE_N = \frac{1}{m}\sum_{j=1}^{m}(\hat{A}_N^{(j)} - A_N)^2$$

where (j) refers to jth replication. With $m = 10,000$, categorization number is 2, we display the simulated results in Table 4.

Table 4. MSE's for region areas

MSE	$\sigma_{yx} = 0.2$	0.3	0.5	0.7	0.8
$(\sigma_y^2, \sigma_x^2) = (1,1), n = 50$					
MSE_N	10.35	7.016	3.462	2.181	2.096
MSE_p	1.798	1.643	1.185	0.628	0.387
$n = 100$					
MSE_N	9.213	6.103	2.793	1.778	1.746
MSE_p	0.876	0.822	0.582	0.314	0.198
$(\sigma_y^2, \sigma_x^2) = (1,2), n = 50$					
MSE_N	12.97	9.929	5.654	3.295	2.562
MSE_p	1.836	1.757	1.539	1.162	0.976
$n = 100$					
MSE_N	11.70	8.800	2.560	2.675	2.052
MSE_p	0.923	0.851	0.993	0.578	0.493
$(\sigma_y^2, \sigma_x^2) = (2,1), n = 50$					
MSE_N	51.96	39.67	22.69	14.13	11.34
MSE_p	7.375	7.084	5.977	4.724	4.048
$n = 100$					
MSE_N	46.82	35.00	19.49	11.31	9.184
MSE_p	3.795	3.577	2.998	2.378	2.010

Accuracy in estimation of unknown parameters gives the parametric sample categorized means the advantage of smaller area of interest. This is another desired property for parametric estimation of unknown categorized means.

In brief summary, attractive properties shown above for parametric estimation is benefited from the new and novel parametrization in Theorem 3.1.

5. Categorization Creating Auxiliary information

In this section, we show that categorization is linked to a theory very important in efficient estimation. Statistician has long been interested in looking for inference method with possible improvement of accuracy. Let $Y_1, ..., Y_n$ be a random sample from a density function $f(y, \theta_y)$ with θ_y being the interest of parameter. We know that Cramer-Rao's theory gives us no chance in improving an uniformly minimum variance unbiased estimator when regularity conditions are assumed. Researchers then turned to find estimator sequence $\{\hat{\theta}_y\}$ asymptotically normal as

$$\sqrt{n}(\hat{\theta}_y - \theta) \rightarrow N(0, v_{\theta_y}) \qquad (5.1)$$

in distribution that has superefficient point θ_y in parameter space as

$$v_{\theta_y} < I(\theta_y)^{-1} \tag{5.2}$$

where $I(\theta_y)$ is the Fisher information at θ_y. In 1951, Hodges produced an estimator (Bickel, et al. (1998)) with one superefficient point. Later, Le Cam (1953) showed that for any sequence of estimators satisfying (5.1), the set of superefficient points has Lebesgue measure zero. This also tells us that estimators with superefficiency is only interesting theoretically but not in practice.

An interest in theory and practice is then looking for a statistic containing auxiliary information so that it improves inference's accuracy. Verifying existence of auxiliary information has received some attention in literature, see, for examples, Kuk and Mak (1989), Rao, Kovar and Mantel (1990) and Martinez-Miranda, Rueda and Arcos (2007) for quantile estimation and Srivastava (1971) for mean estimation. We prove that categorization contributes this improvement. In robust estimation of population mean μ_y, the classical trimmed mean based on random sample $Y_1, ..., Y_n$ is defined as

$$\hat{\mu}_t(\alpha_1, \alpha_2) = \frac{\sum_{i=1}^n Y_i I(\hat{F}_Y^{-1}(\alpha_1) \leq Y_i \leq \hat{F}_Y^{-1}(\alpha_2)}{\sum_{i=1}^n I(\hat{F}_Y^{-1}(\alpha_1) \leq Y_i \leq \hat{F}_Y^{-1}(\alpha_2))}. \tag{5.3}$$

Now, suppose that as in our design for categorization we also have an extra random sample $X_1, ..., X_n$ with Y_i and X_i correlated. For $0 < \alpha_1 < \alpha_2 < 1$, we call the following categorized sample mean

$$\hat{\mu}_{y,cat}(\alpha_1, \alpha_2) = \frac{\sum_{i=1}^n Y_i I(\hat{F}_X^{-1}(\alpha_1) \leq X_i \leq \hat{F}_X^{-1}(\alpha_2)}{\sum_{i=1}^n I(\hat{F}_X^{-1}(\alpha_1) \leq X_i \leq \hat{F}_X^{-1}(\alpha_2))} \tag{5.4}$$

the categorized trimmed mean. This is first example of estimating a distributional parameter of uncategorized variable with estimator based on categorized sample. We prove that categorization creates auxiliary information for robust estimation.

The following theorem with $\alpha_1 = 1 - \alpha_2 = \alpha$ is a direct result from Theorem 2.2.

Theorem 5.1. Suppose that the joint distribution of Y and X is spherically symmetric.

(a) The Barhadur representation for the categorized trimmed mean is

$$\sqrt{n}(\hat{\mu}_{y,cat}(\alpha, 1-\alpha) - \mu_y) = \frac{1}{1-2\alpha} n^{-1/2} \sum_{i=1}^{n} \psi_0(Y_i, X_i) + o_p(1)$$

where

$$\psi_0(Y, X) = \begin{cases} -\lambda_{1-\alpha} & \text{if } X \leq F_X^{-1}(\alpha) \\ Y - \mu_y & \text{if } F_X^{-1}(\alpha) < X < F_X^{-1}(1-\alpha) \\ \lambda_{1-\alpha} & \text{if } X \geq F_X^{-1}(1-\alpha) \end{cases} .$$

(b) Then $\sqrt{n}(\hat{\mu}_{y,cat}(\alpha, 1-\alpha) - \mu_y)$ is asymptotically normal with distribution $N(0, \sigma_{cat}^2)$ where

$$\sigma_{cat}^2 = \frac{1}{(1-2\alpha)^2}[2\alpha\lambda_{1-\alpha}^2 + M_\alpha]$$

where $M_\alpha = E[(Y - \mu_y)^2 I(F_X^{-1}(\alpha) < X < F_X^{-1}(1-\alpha))]$.

We also denote the asymptotic variance of the classical trimmed mean of (5.1) by σ_t^2 (Ruppert and Carroll (1980)).

For verification of our assertion, we design the following setting of mixed distribution:

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim (1-\delta)N_2(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}) + \delta N_2(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{12}^* \\ \sigma_{12}^* & 1 \end{pmatrix})$$

indicating the interest of parameter is $\mu_y = E(Y)$. For each setting of the distributional parameters, we compute the minimum σ_t^2 and σ_{cat}^2 and display them in Tables 5 and 6. Note that the values in parentheses in Table 5 are the trimming proportions achieving smallest asymptotic variances and $I(\mu_y)^{-1}$ represents Y-variable based inverse of Fisher information and, for theoretical interest, we list σ_{cat}^2's when $\sigma_{cat}^2 < I(\mu_y)^{-1}$ holds.

Table 5. Comparison of asymptotic variances ($\delta = 0.1$)

	σ^2_{cat} $(\sigma_{12} = -0.9)$	σ^2_{cat} (-0.8)	σ^2_t	$I(\mu_y)^{-1}$
$\sigma^2_y = 2$				
$\sigma^*_{12} = 1$	1.042(0.06)	1.067(0.05)	1.107(0.05)	1.093
1.4	0.951(0.13)	1.003(0.08)		
$\sigma^2_y = 5$				
$\sigma^*_{12} = 2$	1.131(0.15)	1.202(0.11)	1.230(0.06)	1.220
2.2	1.019(0.2)	1.118(0.14)		
$\sigma^2_y = 9$				
$\sigma^*_{12} = 2.9$	1.150(0.23)		1.296(0.09)	1.256
2.99	1.053(0.26)	1.215(0.2)		

Table 6. Comparison of asymptotic variances ($\delta = 0.2$)

	$\sigma_{12} = 0.5$	0.7	0.9	0.99	$I(\mu_y)^{-1}$
$\sigma^2_y = 2$					
$\sigma^*_{12} = -1$	1.179	1.128	1.042	0.974	1.185
-1.4	1.084	0.985	0.788	0.571	
$\sigma^2_y = 5$					
$\sigma^*_{12} = -2.2$	1.416	1.228	0.892	0.576	1.448
-2.22	1.401	1.205	0.853	0.494	
$\sigma^2_y = 9$					
$\sigma^*_{12} = -2.9$			1.205	0.896	1.532
-2.99		1.460	0.949	0.460	

We have several comments for the results in Tables 5 and 6:

(a) Without extra information from other variables, the classical trimmed mean gains no benefit in outperforming the lower bound $I(\mu_y)$ in any case while, in the designed distributions in terms of variances and covariances, the categorized trimmed means outperform the corresponding lower bounds.

(b) We see the power of auxiliary information that σ^2_{cat} can be as small as 0.46 when the lower bound is 1.532. Auxiliary information greatly improves in reduction of asymptotic variance of trimming estimation.

(c) The fact that the set $\{\mu_y : \sigma^2_{cat} < I(\mu_y)^{-1}, \mu_y \in R\}$ is Lebesgure measure greater than zero supports the consideration of using auxiliary information to modify statistical inference methods.

(d) The auxiliary information exists in this robust estimation when the extra variable has relatively smaller variance and is highly correlated with the response variable. This meets the general understanding in literature.

1896

(e) Not every extra variable X provides auxiliary information. Suppose that we have the following normality assumption:

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{12} \\ \sigma_{12} & \sigma_y^2 \end{pmatrix}).$$

Let $I_{yx}(\mu_y)^{-1}$ be the inverse of Fisher information for μ_y that is derived from the above bivariate distribution. We may see that $I_{yx}^{-1}(\mu_y) = I^{-1}(\mu_y) = \sigma_y^2$ indicating that auxiliary information does not exists when bivariate normal is true.

6. Concluding Remarks

For predicting the unknown population means of categorized variables, we have derived distributional theory for parametric and nonparametric estimators that allows us to construct "theoretically correct" and "advanced" inference methods. Both approaches are shown to be valuable in statistical theory and application. The novel parametrization results in the parametric estimators being much more efficient than the classical ANOVA used sample means. On the other hand the nonparametric categorized sample mean is found involving an auxiliary information that greatly improves the efficiency of nonparametric robust estimation. Ignorance to theory investigation for categorization not only blindly face the use of untrustworthy inference methods but also forfeit the chance to discover interesting information created by categorization for inference improvement. For long being criticized, we have finally taken a big step in knowing it but it deserves to receive more attention in statistical society.

We have several further remarks on this research:

(a) Idea of parametrization provides efficient parametric inference techniques for unknown distributional parameters of categorized variables. Extension of this parametrized parametric approach is desired to non-normal distribution.

(b) The accuracy properties of confidence interval and hypothesis testing method formulated by quantity (4.1) requires for further investigation where testing the usual interest of equal means can be done by setting general linear hypothesis with

$$A=\begin{pmatrix}1 & -1 & 0 & 0 & \dots & 0 & 0\\ 0 & 1 & -1 & 0 & \dots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & \dots & 1 & -1\end{pmatrix}.$$

(c) Intuitively we can expect that some other robust estimation methods such as Winsorized mean in L-estimation and Huber's M-estimation can be benefited with efficiency improvement when the categorization created auxiliary information is applied.

(d) It is expected that asymptotic variance σ^2_{cat} can be even reduced more when multiple auxiliary information is used. The Rao-Blackwell theorem like theory as for how much capacity of improvement could be reached is an interesting theoretical problem.

7. Appendix

We only give the proof for Theorem 2.2 while Theorem 3.1 are induced from Han (2008) and the proofs for Theorems 2.1 and 3.2 being straightforward are skipped.

Proof of Theorem 2.2. With quantile cutoffs, the sample group means may be represented as

$$\hat{\theta}_{qj}-\mu_y=\frac{\sum_{i=1}^n (Y_i-\mu_y)I(\hat{F}_X^{-1}(\alpha_{j-1})\le X_i\le \hat{F}_X^{-1}(\alpha_j))}{\sum_{i=1}^n I(\hat{F}_X^{-1}(\alpha_{j-1})\le X_i\le \hat{F}_X^{-1}(\alpha_j))}. \tag{7.1}$$

Following the approaches of Ruppert and Carroll (1980) and Chen and Chiang (1996), we may see that

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^n (Y_i-\mu_y)[I(X_i\le F_X^{-1}(\alpha)+n^{-1/2}T_n)-I(X_i\le F_X^{-1}(\alpha)] \\ &= E(Y-\mu_y|F_X^{-1}(\alpha))f_X(F_X^{-1}(\alpha))T_n+o_p(1) \end{aligned} \tag{7.2}$$

for any sequence $T_n=O_p(1)$. The fact that $I(X_i\le \hat{F}_X^{-1}(\alpha))=I(X_i\le$

$F_X^{-1}(\alpha)+n^{-1/2}T_X)$ with $T_X=\sqrt{n}(\hat{F}_X^{-1}(\alpha)-F_X^{-1}(\alpha))$ and (7.2) gives

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}(Y_i-\mu_y)I(\hat{F}_X^{-1}(\alpha_{j-1})\le X_i\le \hat{F}_X^{-1}(\alpha_j)) \qquad (7.3)\\
&=[E(Y-\mu_y|X=F_X^{-1}(\alpha_j))f_X(F_X^{-1}(\alpha_j)n^{1/2}(\hat{F}_X^{-1}(\alpha_j)-F_X^{-1}(\alpha_j))\\
&-E(Y-\mu_y|X=F_X^{-1}(\alpha_{j-1}))f_X(F_X^{-1}(\alpha_{j-1})n^{1/2}(\hat{F}_X^{-1}(\alpha_{j-1})\\
&-F_X^{-1}(\alpha_{j-1}))+n^{-1/2}\sum_{i=1}^{n}(Y_i-\mu_y)I(F_X^{-1}(\alpha_{j-1})\le X_i\le F_X^{-1}(\alpha_j))]+o_p(1).
\end{aligned}$$

A representation for regression quantile $\hat{F}_x^{-1}(\alpha)$ as

$$\sqrt{n}(\hat{F}_X^{-1}(\alpha)-F_X^{-1}(\alpha))=f_X^{-1}(F_X^{-1}(\alpha))n^{-1/2}\sum_{i=1}^{n}(\alpha-I(X_i\le F_X^{-1}(\alpha))+o_p(1). \qquad (7.4)$$

may be seen in Ruppert and Carroll (1980). Moreover, we also have

$$n^{-1}\sum_{i=1}^{n}I(\hat{F}_X^{-1}(\alpha_{j-1})\le X_i\le \hat{F}_X^{-1}(\alpha_j))=\alpha_j-\alpha_{j-1}+o_p(1). \qquad (7.5)$$

By plugging (7.4) into (7.3) and with careful re-arrangement, the theorem is followed from (7.1)-(7.3) and (7.5).

Acknowledgement. We would like to acknowledge the help of Professor Stephen Portnoy for his comments and suggestions concerning auxiliary information.

References

Bennette, C. and Vickers, A. (2012). Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, **12**: 21.

Bickel, P. J., Klaassen, C. A.J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer: New York.

Chen, L.-A. and Chiang, Y. C. (1996). Symmetric type quantile and trimmed means for location and linear regression model. *Journal of Nonparametric Statistics.* 7, 171-185.

Chen, L.-A., Chen, Dung-Tsa and Chan, Wenyaw. (2010). The p Value for the Outlier Sum in Differential Gene Expression Analysis. *Biometrika*, 97, 246-253.

Han, Y. (2008). Mathematical and empirical examinations of some epidemiological procedures. Ph.D. Dissertation, School of Public Health, University of Texas-Health Science Center at Houston.

Irwin, J. R. and McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, **40**, 366-371.

Kuk, A. and Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society B*, **1**, 261-269.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, **1**, 277-330.

Letenneur L., Proust-Lima, C. et al. (2007). Flavonoid intake and cognitive decline over a 10-year period. *American Journal of Epidemiology*, **165**: 1364-1371.

Luo J., Margolis K. L. et al. (2007). Body size, weight cycling, and risk of renal cell carcinoma among postmenopausal women: the women's health initiative (United States). *American Journal of Epidemiology*, **166**: 752-759.

MacCallum, R. C., Zhang, S., Preacher, K. J. and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, **7**, 19-40.

Martinez-Miranda, M. D., Rueda, M. and Arcos, A. (2007). Looking for optimal auxiliary variables in sample survey quantile estimation. *Statistics*, **41**, 241-252.

MacCallum, R. C., Zhang, S., Preacher, K. J. and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, **7**, 19-40.

Morgan, T. M. and Elashoff, R. M. (1986). Effect of categorizing a continuous covariate on the comparison of survival time.

Pocock, S. J., Collier, T. J. et al. (2004). Issues in reporting of epidemiological studies: a survey of recent practice. *British Medical Journal*, 329: 883.

Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, **77**, 365-375.

Royston, P. and Sauerbrei, W. (2008). *Multivariate Model-Building: A parametric approach to regression analysis based on fractional polynomials for modeling continuous variables.* Wiley.

Ruppert, D. and Carroll, R.J. (1980). Trimmed least squares estimation in the linear model. *Journal of American Statistical Association* **75**, 828-838.

Shankar A., Klein R. et al. (2007). Association between glycosylated hemoglobin level and cardiovascular and all-cause mortality in type 1 diabets. *American Journal of Epidemiology*, **166**: 393- 402.

Srivastava, S. K. (1971). A generalized estimator for the mean of a finite population using multiauziliary information. *Journal of the American Statistical Association*, **66**, 404-407.

Taylor, J. M. G. and Yu, M. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*, **83**, 2001-2045.

Walraven, C. and Hart, R. G. (2008). Leave'em alone-why continuous variables should be analyzed as such. *Neuroepidemiology*, **30**, 138-139.

Zhao, L. P. and Kolonel, L. N. (1992). Efficiency loss from categorizing quantitative exposures in case-control studies. *American Journal of Epidemiology*, **136**, 464-474.