ESTIMATING THE BINOMIAL PROPORTION AND
THE RISK DIFFERENCE IN MULTI-CENTER
STUDIES WITH ADJUSTING SPARSITY

By

Chukiat Viwatwongkasem

A Dissertation Submitted in Partial
Fulfillment of The Requirements for The Degree of
Doctor of Philosophy (Statistics)
School of Applied Statistics
National Institute of Development Administration
ISBN 974-231-669-4
2005
ESTIMATING THE BINOMIAL PROPORTION AND THE RISK DIFFERENCE IN MULTI-CENTER STUDIES
WITH ADJUSTING SPARSITY
Chukiat Viwatwongkasem
School of Applied Statistics

The Examining Committee Approved This Dissertation Submitted in Partial Fulfillment of The Requirements for The Degree of Doctor of Philosophy (Statistics).

Associate Professor...........................................Committee Chairman

(Jirawan Jitthavech, Ph.D.)

Professor.........................................................Committee

(Dankmar Böhning, Ph.D.)

Associate Professor...........................................Committee

(Vichit Lorchirachoonkul, Ph.D.)

Professor.........................................................Committee

(Prachoom Suwattee, Ph.D.)

Associate Professor...........................................Committee

(Samruam Chongcharoen, Ph.D.)

..............................................................Dean

(Lersan Bosuwan, Ph.D.)
Date: 30-09-48
The conventional proportion estimator \( \hat{p} = X/n \) (number of events divided by sample sizes) for estimating the binomial parameter \( p \) encounters a number of problems when data are sparse. It is suggested that \( p \) be estimated using the class \( \hat{p}_c \), where \( \hat{p}_c = (X + c)/(n + 2c) \). Choosing optimal point \( c \) is investigated in a center study from various perspectives. For a multi-center study of size \( k \), the optimal weights \( \hat{f}_{ej} \) that minimize the MSE of \( \hat{p}_{cw} = \sum_{j=1}^{k} \hat{f}_{ej} \hat{p}_{ej} \) are derived, subject to \( \sum_{j=1}^{k} \hat{f}_{ej} = 1 \), where \( \hat{p}_{ej} = (X_j + c)/(n_j + 2c) \). The performance in terms of the smallest MSE of \( \hat{p}_{cw} \) is compared with the well-known summary proportion estimators.

The results show that the optimal value of \( c \) for minimizing the MSE of \( \hat{p}_c \) is dependent on \( p \) and equals \( c = 2p(1-p)/(1-2p)^2 \). To eliminate \( p \), Bayes risk with respect to the uniform prior and Euclidean loss are used, leading to the minimum point \( c = 1 \). This result is true and can be extended for application in a multi-center study of size \( k \). The best performance in terms of the smallest Bayes risk of \( \hat{p}_{cw} \) is provided at the point \( c = 1 \) for moderate to large sample sizes \( n_j \geq 16 \), followed by \( \hat{p}_{cw} \) defined by \( c = 0.5 \) (at least for smaller sample sizes \( n_j \leq 8 \)).
For estimating the risk difference $\theta = p_1 - p_2$, the optimal point $(c_1, c_2)$ with the smallest MSE of the adjusted risk difference estimator 

$$\hat{\theta}_c = \hat{p}_{c_1} - \hat{p}_{c_2} = \frac{(X_1 + c_1)}{(n_1 + 2c_1)} - \frac{(X_2 + c_2)}{(n_2 + 2c_2)}$$

found in a particular case of $c = c_1 = c_2$ and $n = n_1 = n_2$ as $c = \frac{p_1(1 - p_1) + p_2(1 - p_2)}{2\theta^2}$, $\theta \neq 0$. Another idea from a Bayesian perspective found the optimal point $(c_1, c_2)$ for minimizing the average MSE of $\hat{\theta}_c$ is $(c_1, c_2) = (1, 1)$. In a multi-center study of size $k$, the optimal weights $f_{c_j}$ that minimize the MSE of $\hat{\theta}_{cw} = \frac{1}{\sum_{j=1}^{k} f_{c_j}} \frac{\hat{\theta}_{c_j}}{\hat{\theta}_{c_j}}$ are derived subject to $\sum_{j=1}^{k} \hat{f}_{c_j} = 1$, where

$$\hat{\theta}_{c_j} = \hat{p}_{c_1j} - \hat{p}_{c_2j} = \frac{(X_{1j} + c_1)}{(n_{1j} + 2c_1)} - \frac{(X_{2j} + c_2)}{(n_{2j} + 2c_2)}.$$  

Under a common risk difference $\theta$ over all centers, the performance in the sense of smallest MSE of $\hat{\theta}_{cw}$ is that, if $\theta = 0$, $\theta = 0.1$, $\theta = 0.2$, and $\theta = 0.3$, the proposed summary estimator $\hat{\theta}_{cw}$ adjusted by $c = c_1 = c_2 = 2$ is the best choice, with the smallest MSE. For $\theta = 0.4$, the proposed estimator $\hat{\theta}_{cw}$ adjusted by $c = c_1 = c_2 = 1$ performs best. However, since the true value of $\theta$ is usually not available; in practice, we suggest choosing the estimator $\hat{\theta}_{cw}$ adjusted by $c = c_1 = c_2 = 1$, which can minimize the Bayes risk with respect to uniform prior and Euclidean loss function.

Alternatively, using the ratio estimation method, the optimal point $c$ that reduces the bias of $\hat{p}_c = (\bar{X} + c)/(\bar{n} + 2c)$ where $\bar{X} = \frac{1}{k} \sum_{j=1}^{k} X_j$ and $\bar{n} = \frac{1}{k} \sum_{j=1}^{k} n_j$ are the sample means over all $k$ centers, i.e. 

$$\hat{c} = \frac{\bar{n}}{k(1 - 2\hat{p})}\left[\frac{s_{xn}}{4c^2 + 4c\bar{n} + \left(\frac{1}{k(1 - 2\hat{p})}(s_x^2 - 2s_{xn})\right)}\right]$$

where $\hat{p} = \bar{X}/\bar{n}$, $s_x^2 = \frac{1}{k - 1} \sum_{j=1}^{k} (n_j - \bar{n})^2$, and $s_{xn} = \frac{1}{k - 1} \sum_{j=1}^{k} (X_j - \bar{X})(n_j - \bar{n})$. We suggest selecting $\hat{p}_c = (\bar{X} + c)/(\bar{n} + 2c)$ when center size $k$ or sample size $n_j$ are moderate to large.
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and deepest appreciate to Assoc. Prof. Jirawan Jitthavech, my major advisor, Prof. Dankmar Böhning, Assoc. Prof. Vichit Lorchirachoonkul, Prof. Prachoom Suwattee, and Assoc. Prof. Samruam Chongcharoen, my co-advisors, for their constructive comments, valuable suggestions, constant support, and warm encouragement. They have sacrificed their precious time to give advice and have the grace to correct mistakes and make improvements so that this study come out wholly complete. I also acknowledge the opportunity they give me to pursue this study and to express my ideas. I am truly grateful for their help and thank them profusely.

I am greatly indebted to the teachers of the School of Applied Statistics, National Institute of Development Administration, who impart their knowledge and experiences, which really guide me through this study.

I would like to thank deeply the National Research Council of the Kingdom of Thailand (NRCT) and the Deutsche Forschungsgemeinschaft (DFG: German Research Foundation), Federal Republic of Germany. Most of my work has been supported by the NRCT in co-operation with the DFG.

I am sure that I forgot some names of contributors to this study (at least, for all my friends and colleagues). I would like to express my deep thanks to all persons—if they have been mentioned above or not—for their energy and help, spirit, patience, and co-operation.

Finally, I am also grateful and owe my deepest appreciation to my parents, my wife, my daughters, my brothers, and my sister for their continuous support, great willingness, encouragement and love throughout my entire life.

Chukiat Viwatwongkasem

September 2005