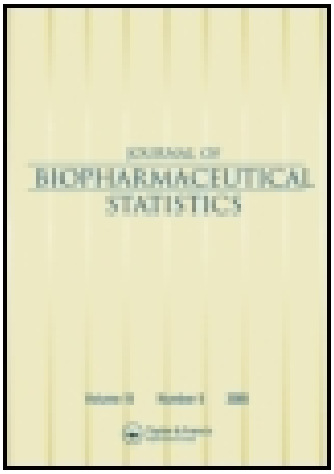

Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

A Note on Sample Size Calculation Based on Propensity Analysis in Nonrandomized Trials

Sin-Ho Jung [a], Shein-Chung Chow [a b] & Eric M. Chi [c]
[a] Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA
[b] Department of Statistics, National Cheng-Kung University, Tainan, Taiwan
[c] US Medical Affairs Biostatistics, Amgen, Inc, Thousand Oaks, California, USA
Published online: 02 Feb 2007.

To cite this article: Sin-Ho Jung, Shein-Chung Chow & Eric M. Chi (2007) A Note on Sample Size Calculation Based on Propensity Analysis in Nonrandomized Trials, Journal of Biopharmaceutical Statistics, 17:1, 35-41, DOI: 10.1080/10543400601044790

To link to this article: http://dx.doi.org/10.1080/10543400601044790

Journal of Biopharmaceutical Statistics, 17: 35–41, 2007

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400601044790

Taylor & Francis
Taylor & Francis Group

A NOTE ON SAMPLE SIZE CALCULATION BASED ON PROPENSITY ANALYSIS IN NONRANDOMIZED TRIALS

Sin-Ho Jung
Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA

Shein-Chung Chow
Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA and Department of Statistics, National Cheng-Kung University, Tainan, Taiwan

Eric M. Chi
US Medical Affairs Biostatistics, Amgen, Inc. Thousand Oaks, California, USA

In nonrandomized trials, patients are not randomly assigned to treatment groups with equal probability. Instead, the probability of assignment varies from patient to patient depending on patients baseline covariates. This often results in a non-comparable treatment groups due to treatment imbalance. As a result, the United States Food and Drug Administration (FDA) recommended that the method of propensity score analysis be employed to overcome this problem. In this note, a formula for sample size calculation is developed based on a proposed weighted Mantel–Haenszel test on the strata defined by the propensity score analysis. It was shown that the sample size formula derived by Nam (1998) based on the test statistic proposed by Gart (1985) is a special case of the sample size formula derived in this note.

Key Words: Nonrandomized trials; Propensity score analysis; Weighted Mantel–Haenszel test.

1. INTRODUCTION

As indicated in Yue (2006), the use of propensity analysis in nonrandomized trials has received much attention, especially in the area of medical device clinical studies. In a nonrandomized study, patients are not randomly assigned to treatment groups with equal probability. Instead, the probability of assignment varies from patient to patient depending on patient's baseline covariates. This often results in a non-comparable treatment groups due to imbalance of the baseline covariates and consequently invalid the standard methods commonly employed in data analysis. To overcome this problem, Yue (2006) recommends the method of propensity score developed by Rosenbaum and Rubin (1983, 1984) be used.

Received September 15, 2006
Address correspondence to Sin-Ho Jung, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA; E-mail: sinho.jung@duke.edu

In her review article, Yue (2006) described some limitations for use of propensity score. For example, propensity score method can only adjust or observed covariates and not for unobserved ones. As a result, it is suggested that a sensitivity analysis be conducted for possible hidden bias. In addition, Yue (2006) also posted several statistical and regulatory issues for propensity analysis in nonrandomized trials including sample size calculation. In this discussion, our emphasis will be placed on the issue of sample size calculation in the context of propensity scores. We propose a procedure for sample size calculation based on weighted Mantel–Haenszel test with different weights across score subclasses.

In the next section, a proposed weighted Mantel–Haenszel test, the corresponding formula for sample size calculation, and a formula for sample size calculation ignoring strata are briefly described. Section 2 summarizes the results of several simulation studies conducted for evaluation of the performance of the proposed test in the context of propensity analysis. A brief concluding remark is given in the last section.

1.1. Weighted Mantel–Haenszel Test

Suppose that the propensity score analysis defines J strata. Let n denote the total sample size, and n_j the sample size in stratum j ($\sum_{j=1}^{J} n_j = n$). The data on each subject comprise of the response variable $x = 1$ for response and 0 for no response; j and k for the stratum and treatment group, respectively, to which the subject is assigned ($1 \le j \le J$; $k = 1, 2$). We assume that group 1 is the control. Frequency data in stratum j can be described as follows:

	Group		
Response	1	2	Total
Yes	x_{j11}	x_{j12}	x_{j1}
No	x_{j21}	x_{j22}	x_{j2}
Total	n_{j1}	n_{j2}	n_j

Let $O_j = x_{j11}$, $E_j = n_{j1}x_{j1}/n_j$, and

$$V_j = \frac{n_{j1}n_{j2}x_{j1}x_{j2}}{n_j^2(n_j - 1)}.$$

Then, the weighted Mantel–Haenszel (WMH) test is given by

$$T = \frac{\sum_{j=1}^{J} \hat{w}_j(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} \hat{w}_j^2 V_j}},$$

where the weights $\hat{w}_j$ converges to a constant w_j as $n \to \infty$. The weights are $\hat{w}_j = 1$ for the original Mantel–Haenszel (MH) test and $\hat{w}_j = \hat{q}_j = x_{j2}/n_j$ for the statistic proposed by Gart (1985).

Let $a_j = n_j/n$ denote the allocation proportion for stratum j ($\sum_{j=1}^{J} a_j = 1$), and $b_{jk} = n_{jk}/n_j$ denote the allocation proportion for group k within stratum j ($b_{j1} + b_{j2} = 1$). Let p_{jk} denote the response probability for group k in stratum

j and $q_{jk} = 1 - p_{jk}$. Under $H_0 : p_{j1} = p_{j2},\ 1 \le j \le J,\ T$ is approximately $N(0, 1)$. The optimal weights maximizing the power depend on the allocation proportions $\{(a_j, b_{1j}, b_{2j}),\ j = 1, \ldots, J\}$ and effect sizes $(p_{j1} - p_{j2}, 1, \ldots, J)$ under H_1.

1.2. Power and Sample Size

In order to calculate the power of WMH, we have to derive the asymptotic distribution of $\sum_{j=1}^{J} \hat{w}_j(O_j - E_j)$ and the limit of $\sum_{j=1}^{J} \hat{w}_j^2 V_j$ under H_1. We assume that the success probabilities $(p_{jk},\ 1 \le j \le J,\ j = 1, 2)$ satisfy $p_{j2}q_{j1}/(p_{j1}q_{j2}) = \phi$ for $\phi \ne 1$ under H_1. Note that a constant odds ratio across strata holds if there exists no interaction between treatment and the propensity score when the binary response is regressed on the treatment indicator and the propensity score using a logistic regression. Following derivations are based on H_1. It can be verified that

$$\begin{aligned} O_j - E_j &= \frac{n_{j1}n_{j2}}{n_j}(\hat{p}_{j1} - \hat{p}_{j2}) \\ &= \frac{n_{j1}n_{j2}}{n_j}(\hat{p}_{j1} - p_{j1} - \hat{p}_{j2} + p_{j2}) + \frac{n_{j1}n_{j2}}{n_j}(p_{j1} - p_{j2}) \\ &= na_jb_{j1}b_{j2}(\hat{p}_{j1} - p_{j1} - \hat{p}_{j2} + p_{j2}) + na_jb_{j1}b_{j2}(p_{j1} - p_{j2}) \end{aligned}$$

where $\hat{p}_{jk} = x_{j1k}/n_{jk}$. Thus, under H_1, $\sum_{j=1}^{J} \hat{w}_j(O_j - E_j)$ is approximately normal with mean $n\delta$ and variance $n\sigma_1^2$, where

$$\begin{aligned} \delta &= \sum_{j=1}^{J} w_j a_j b_{j1} b_{j2}(p_{j1} - p_{j2}) \\ &= (1 - \phi)\sum_{j=1}^{J} w_j a_j b_{j1} b_{j2} \frac{p_{j1}q_{j1}}{q_{j1} + \phi p_{j1}} \end{aligned}$$

and

$$\begin{aligned} \sigma_1^2 &= n^{-1}\sum_{j=1}^{J} w_j^2 \frac{n_{j1}^2 n_{j2}^2}{n_j^2}\left(\frac{p_{j1}q_{j1}}{n_{j1}} + \frac{p_{j2}q_{j2}}{n_{j2}}\right) \\ &= \sum_{j=1}^{J} w_j^2 a_j b_{j1} b_{j2}(b_{j2}p_{j1}q_{j1} + b_{j1}p_{j2}q_{j2}). \end{aligned}$$

Also under H_1, we have

$$\sum_{j=1}^{J} w_j^2 V_j = n\sigma_0^2 + o_p(n),$$

where

$$\sigma_0^2 = \sum_{j=1}^{J} w_j^2 a_j b_{j1} b_{j2}(b_{j1}p_{j1} + b_{j2}p_{j2})(b_{j1}q_{j1} + b_{j2}q_{j2}).$$

Hence, the power of WMH is given as

$$\begin{aligned} 1-\beta &= P(|T| > z_{1-\alpha/2} \mid H_1) \\ &= P\left(\frac{\sigma_1}{\sigma_0} Z + \sqrt{n}\frac{|\delta|}{\sigma_0} > z_{1-\alpha/2}\right) \\ &= \overline{\Phi}\left(\frac{\sigma_0}{\sigma_1} z_{1-\alpha/2} - \sqrt{n}\frac{|\delta|}{\sigma_1}\right), \end{aligned}$$

where Z is a standard normal random variable and $\overline{\Phi}(z) = P(Z > z)$. Thus, sample size required for achieving a desired power of $1-\beta$ can be obtained as

$$n = \frac{(\sigma_0 z_{1-\alpha/2} + \sigma_1 z_{1-\beta})^2}{\delta^2}. \tag{1}$$

Following the steps as described in Chow et al. (2003), the sample size calculation for the weighted Mantel–Haenszel test can be carried out as follows:

(1) Specify the input variables

- Type I and II error probabilities, (α, β).
- Success probabilities for group 1, $p_{11}, \ldots, p_{J1}$, and the odds ratio ϕ under H_1. Note that $p_{j2} = \phi p_{j1}/(q_{j1} + \phi p_{j1})$.
- Incidence rates for the strata, $(a_j, j = 1, \ldots, J)$. (Yue, in her article, proposes to use $a_j \approx 1/J$.)
- Allocation probability for group 1 within each stratum, $(b_{j1}, j = 1, \ldots, J)$.

(2) Calculate n by

$$n = \frac{(\sigma_0 z_{1-\alpha/2} + \sigma_1 z_{1-\beta})^2}{\delta^2},$$

where

$$\delta = \sum_{j=1}^{J} a_j b_{j1} b_{j2} (p_{j1} - p_{j2})$$

$$\sigma_1^2 = \sum_{j=1}^{J} a_j b_{j1} b_{j2} (b_{j2} p_{j1} q_{j1} + b_{j1} p_{j2} q_{j2})$$

$$\sigma_0^2 = \sum_{j=1}^{J} a_j b_{j1} b_{j2} (b_{j1} p_{j1} + b_{j2} p_{j2})(b_{j1} q_{j1} + b_{j2} q_{j2}).$$

1.3. Sample Size Calculation Ignoring Strata

We consider the case where ignoring strata and combining data across J strata. Note that, in the presence of unbalanced allocation in baseline covariates, the analysis ignoring strata will be biased. We want to investigate the impact of ignoring strata in sample size and parameter estimators.

Let $n_k = \sum_{j=1}^{J} n_{jk}$ denote the sample size in group k. Ignoring strata, we may estimate the response probabilities by $\hat{p}_k = n_k^{-1} \sum_{j=1}^{J} x_{j1k}$ for group k and $\hat{p} = n^{-1} \sum_{j=1}^{J} \sum_{k=1}^{2} x_{j1k}$ for the pooled data. The WHM ignoring strata reduces to

$$\tilde{T} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(n_1^{-1} + n_2^{-1})}}$$

where $\hat{q} = 1 - \hat{p}$.

Noting that $n_{jk} = na_j b_{jk}$, we have

$$E(\hat{p}_k) \equiv p_k = \sum_{j=1}^{J} a_j b_{jk} p_{jk} \Big/ \sum_{j=1}^{J} a_j b_{jk} \tag{2}$$

and $E(\hat{p}) \equiv p = \sum_{j=1}^{J} a_j (b_{j1} p_{j1} + b_{j2} p_{j2})$. So, under H_1, $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean $\tilde{\delta} = p_1 - p_2$ and variance $n^{-1}\tilde{\sigma}_1^2$, where $\tilde{\sigma}_1^2 = p_1 q_1 / b_1 + p_2 q_2 / b_2$, $q_k = 1 - p_k$ and $b_k = \sum_{j=1}^{J} a_j b_{jk}$. The odds ratio in this case is $\tilde{\phi} = q_1 p_2 / q_1 / p_2$ which equals ϕ when $b_{11} = \cdots = b_{J1}$ (balanced allocation) and $a_1 = \cdots = a_J$. Also under H_1, $\hat{p}\hat{q}(n_1^{-1} + n_2^{-1}) = n^{-1}\tilde{\sigma}_0^2 + o_p(n^{-1})$, where $\tilde{\sigma}_0^2 = pq(b_1^{-1} + b_2^{-1})$. Hence, the sample size ignoring strata is given as

$$\tilde{n} = \frac{(\tilde{\sigma}_0 z_{1-\alpha/2} + \tilde{\sigma}_1 z_{1-\beta})^2}{(p_1 - p_2)^2}. \tag{3}$$

Analysis based on propensity score is to adjust for possible unbalanced baseline covariates between groups. Under balanced baseline covariates (or propensity score), we have $b_{11} = \cdots = b_{J1}$, and, from Eq. (2), $p_1 = p_2$ when H_0: $p_{j1} = p_{j2}$, $1 \le j \le J$ is true. Hence, under balanced covariates, the test statistic $\tilde{T}$ ignoring strata will be valid too. However, by not adjusting for the covariates (or, propensity), it will have a lower power than the stratified test statistic T, see Nam (1998) for Gart's test statistic. On the other hand, if the distribution of covariates is unbalanced, we have $p_1 \neq p_2$ even under H_0, and the test statistic $\tilde{T}$ ignoring strata will not be valid.

1.4. Remarks

For nonrandomized trials, the sponsors usually estimate sample size in the same way as they do in randomized trials. As a result, the United States Food and Drug Administration (FDA) usually gives a warning and request sample size justification (increase) based on the consideration of the degree of overlap in the propensity score distribution.

When there exists an unbalance in covariate distribution between arms, a sample size calculation ignoring strata is definitely biased. The use of different weights will have an impact on statistical power but will not affect the consistency of the proposed weighted Mantel–Haenszel test. Note that the sample size formula by Nam (1998) based on the test statistic proposed by Gart (1985) is a special case of the sample size given in Eq. (1) where $\hat{w}_j = \hat{q}_j$ and $w_j = 1 - b_{j1} p_{j1} - b_{j2} p_{j2}$.

2. SIMULATIONS

Suppose that we want to compare the probability of treatment success between control ($k = 1$) and new ($k = 2$) devices. We consider partitioning the combined data into $J = 5$ strata based on propensity score, and the allocation proportions are projected as $(a_1, a_2, a_3, a_4, a_5) = (0.15, 0.15, 0.2, 0.25, 0.25)$ and $(b_{11}, b_{21}, b_{31}, b_{41}, b_{51}) = (0.4, 0.4, 0.5, 0.6, 0.6)$. Also, suppose that the response probabilities for control device are given as $(P_{11}, P_{21}, P_{31}, P_{41}, P_{51}) = (0.5, 0.6, 0.7, 0.8, 0.9)$, and we want to calculate the sample size required for a power of $1 - \beta = 0.8$ to detect an odds ratio of $\phi = 2$ using two-sided $\alpha = 0.05$. For $\phi = 2$, the response probabilities for the new device are given as $(P_{12}, P_{22}, P_{32}, P_{42}, P_{52}) = (0.6667, 0.7500, 0.8235, 0.8889, 0.9474)$. Under these settings, we need $n = 447$ for MH.

In order to evaluate the performance of the sample size formula, we conduct simulations. In each simulation, $n = 447$ binary observations are generate under the parameters (allocation proportions and the response probabilities) specified for sample size calculation. MH test with $\alpha = 0.05$ is applied to each simulation sample. Empirical power is calculated as the proportion of the simulation samples that reject H_0 out of $N = 10{,}000$ simulations. The empirical power is obtained as 0.7978, which is very close to the nominal $1 - \beta = 0.8$.

If we ignore the strata, we have $p_1 = 0.7519$ and $p_2 = 0.8197$ by (2) and the odds ratio is only $\tilde{\phi} = 1.5004$ which is much smaller than $\phi = 2$. For $(\alpha, 1 - \beta) = (0.05, 0.8)$, we need $\tilde{n} = 1151$ by Eq. (3). With $n = 422$, $\widetilde{T}$ with $\alpha = 0.05$ rejected H_0 for only 41.4% of simulation samples.

The performance of the test statistics, T and $\widetilde{T}$, are evaluated by generating simulation samples of size $n = 447$ under

$$H_0 : (p_{11}, p_{21}, p_{31}, p_{41}, p_{51}) = (p_{12}, p_{22}, p_{32}, p_{42}, p_{52}) = (0.1, 0.3, 0.5, 0.7, 0.9).$$

Other parameters are specified at the same values as above. For $\alpha = 0.05$, the empirical type I error is obtained as 0.0481 for T with MH scores and 0.1852 for $\widetilde{T}$. While the empirical type I error of the MH stratified test is close to the nominal $\alpha = 0.05$, the unstratified test is severely inflated. Under this H_0, we have $p_1 = 0.7953$ and $p_2 = 0.7606$ ($\tilde{\phi} = 0.8181$), which are unequal due to the unbalanced covariate distribution between groups.

Now, let's consider a balanced allocation case, $b_{11} = \cdots = b_{J1} = 0.3$, with all the other parameter values the same as in the above simulations. Under above $H_1 : \phi = 2$, we need $n = 499$ for T and $\tilde{n} = 542$ for $\widetilde{T}$. Note that the unstratified test $\widetilde{T}$ requires a slightly larger sample size due to loss of efficiency. From $N = 10{,}000$ simulation samples of size $n = 499$, we obtained an empirical power of 0.799 for T and only 0.770 for $\widetilde{T}$. The unstratified analysis by $\widetilde{T}$ is slightly underpowered. From similar simulations under H_0, we obtained an empirical type I error of 0.0470 for T with MH scores and 0.0494 for $\widetilde{T}$. Note that both tests control the type I error very accurately in this case. Under H_0, we have $p_1 = p_2 = 0.78$ ($\tilde{\phi} = 1$).

3. CONCLUSION

Sample size calculation plays an important role in clinical research when designing a new medical study. Inadequate sample size could have a significant

impact on the accuracy and reliability of the evaluation of treatment effect especially in medical device clinical studies, which are often conducted in a nonrandomized fashion (although the method of propensity score may have been employed to achieve balance in baseline covariates). We propose a unified sample size formula for weighted Mantel–Haenszel tests based on large-sample assumption. We found through simulations that our sample size formula accurately maintains the power. When the distribution of the covariates is unbalanced between groups, an analysis ignoring the strata could be severely biased.

REFERENCES

Chow, S. C., Shao, J., Wang, H. (2003). *Sample Size Calculation in Clinical Research*. New York: Marcel Dekker, Inc.

Gart, J. J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2×2 tables. *Biometrika* 72:673–677.

Nam, J. M. (1998). Power and sample size for stratified prospective studies using the score method for testing relative risk. *Biometrics* 54:331–336.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Rosenbaum, P. R., Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association* 95:749–759.

Yue, L. (2006). Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *Journal of Biopharmaceutical Statistics* 17(1):1–13.