DOE/EV/02874-67

MASTER

SAMPLE SIZE FOR LOGISTIC REGRESSION WITH SMALL RESPONSE PROBABILITY

ALICE S. WHITTEMORE

TECHNICAL REPORT NO. 33

MARCH 1980

PREPARED UNDER THE AUSPICES OF

SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

SIMS

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Prepared For
THE U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT NO. DE-AS02-76EV02874

DISCLAIMER

SAMPLE SIZE FOR LOGISTIC REGRESSION WITH SMALL RESPONSE PROBABILITY

Alice S. Whittemore
Stanford University

TECHNICAL REPORT NO. 33

MARCH 1980

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

Department of Energy (DOE)
Rockefeller Foundation
Sloan Foundation
Environmental Protection Agency (EPA)
National Science Foundation (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

SAMPLE SIZE FOR LOGISTIC REGRESSION WITH SMALL RESPONSE PROBABILITY

Alice S. Whittemore*
Stanford University

SUMMARY

The Fisher information matrix for the estimated parameters in a multiple logistic regression can be approximated by the augmented Hessian matrix of the moment generating function for the covariates. The approximation is valid when the probability of response is small. With its use one can obtain a simple closed form estimate of the asymptotic covariance matrix of the maximum likelihood parameter estimates, and thus approximate sample sizes needed to test hypotheses about the parameters. The method is developed for selected distributions of a single covariate, and for a class of exponential-type distributions of several covariates. It is illustrated with an example concerning risk factors for coronary heart disease.

Key Words: Fisher information matrix, Hessian matrix, maximum likelihood estimates, moment generating function.

*Alice S. Whittemore is Adjunct Professor, Family, Community, and Preventive Medicine, Stanford University Medical School, Stanford, California 94305. Research was supported by NIH YES Grant No. 1 R01 CA23214-01.

1. INTRODUCTION

Much of the recent biostatistical and epidemiological literature has been concerned with the association between a binary response R, such as disease or death, and a vector $\underset{\sim}{X}' = (X_1,\ldots,X_s)$ of covariates. Here we deal with studies in which a random sample is drawn from the joint distribution of $(R,\underset{\sim}{X})$. In addition, the conditional probability $p(\underset{\sim}{x})$ of response given $\underset{\sim}{X} = \underset{\sim}{x}$ is specified by the model logit $p(\underset{\sim}{x}) = \theta_0 + \underset{\sim}{\theta}'\underset{\sim}{x}$, where logit $p = \log[p/(1 - p)]$, and the unknown parameters $\theta_0,\theta_1,\ldots,\theta_s$ are estimated by maximum likelihood. This paper gives approximate sample sizes needed to test hypotheses about θ_1 with specified significance and power against given alternatives in the case when the probability of response is small. This is done using a simple closed form approximation to the asymptotic covariance matrix of the maximum likelihood estimates.

2. THE APPROXIMATION

Set R = 1 if response occurs, with R = 0 otherwise. For a given sample size N the likelihood of the observations $(r_\nu,\underset{\sim}{x}^{(\nu)})$, $\nu = 1,\ldots,N$ can then be written

$$L(\theta_0,\underset{\sim}{\theta}) = \prod_{\nu=1}^{N} f(\underset{\sim}{x}^{(\nu)})p(\underset{\sim}{x}^{(\nu)})^{r_\nu}\,[1 - p(\underset{\sim}{x}^{(\nu)})]^{1-r_\nu}\,.$$

Here $f(\underset{\sim}{x})$, the joint p.d.f. for $\underset{\sim}{X}$, is assumed to depend upon none of the unknown parameters θ_0, $\underset{\sim}{\theta}$. If the logistic model is valid, then the maximum likelihood estimates $\hat{\theta}_0$, $\hat{\underset{\sim}{\theta}}$ are asymptotically normally distributed with mean θ_0, $\underset{\sim}{\theta}$ and with covariance matrix given by the inverse of the

$(s+1) \times (s+1)$ Fisher information matrix $I(\theta_o, \underset{\sim}{\theta})$. The $(i,j)^{th}$ entry of I is

$$I_{ij} \equiv - E \frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j} = NE\, X_i X_j e^{\theta_o + \underset{\sim}{\theta}'\underset{\sim}{X}} \left(1 + e^{\theta_o + \underset{\sim}{\theta}'\underset{\sim}{X}}\right)^{-2} , \qquad (1)$$

$$i,j = 0,1,\ldots,s ,$$

where $X_o \equiv 1$ and $\underset{\sim}{X}' = (X_1,\ldots,X_s)$. (See for example Cox, 1970.) Expanding the right hand side of (1) in powers of $\exp(\theta_o + \underset{\sim}{\theta}'\underset{\sim}{X})$ yields

$$I_{ij} = N \sum_{k=1}^{\infty} c_k e^{k\theta_o} E\, X_i X_j e^{k\underset{\sim}{\theta}'\underset{\sim}{X}} , \qquad (2)$$

where c_k is $(k!)^{-1}$ times the k^{th} derivative of $z(1+z)^{-2}$ evaluated at $z = 0$.

Let $m(\underset{\sim}{t}) = E \exp(\underset{\sim}{t}'\underset{\sim}{X})$ denote the moment generating function of $\underset{\sim}{X}$, with $m_i \equiv \partial m/\partial t_i$, $i = 1,\ldots,s$ and $m_{ij} = \partial^2 m/\partial t_i \partial t_j$, $i,j = 1,\ldots,s$. We extend this notation by defining $m_o = m_{o,o} \equiv m$, and $m_{o,i} = m_{i,o} \equiv m_i$, $i = 1,\ldots,s$. Then (2) may be rewritten

$$I_{ij} = N \sum_{k=1}^{\infty} c_k e^{k\theta_o} m_{ij}(k\underset{\sim}{\theta}) , \quad i,j = 0,1,\ldots,s . \qquad (3)$$

To express (3) in matrix form, let $\underset{\sim}{m}^{(1)}$ denote the s dimensional column vector of first partials of m, and let $m^{(2)}$ be the $s \times s$ Hessian matrix of second partials of m. We define the augmented Hessian of m to be the $(s+1) \times (s+1)$ matrix H defined by

$$H(\underset{\sim}{\theta}) \equiv \begin{pmatrix} m & \underset{\sim}{m}^{(1)\prime} \\ \underset{\sim}{m}^{(1)} & \underset{\sim}{m}^{(2)} \end{pmatrix} .$$

Then (3) becomes

$$I(\theta_o,\underset{\sim}{\theta}) = N \sum_{k=1}^{\infty} c_k e^{k\theta_o} H(k\underset{\sim}{\theta}) \ . \tag{4}$$

We assume that $\exp(\theta_o + \underset{\sim}{\theta}'\underset{\sim}{x})$ is sufficiently small on the support of f so that I can be approximated by the first term of (4). Since $c_1 = 1$, we rewrite (4) accordingly:

$$I(\theta^o,\underset{\sim}{\theta}) = Ne^{\theta_o} H(\underset{\sim}{\theta}) + O\left(e^{2\theta_o}\right) \ .$$

Thus the asymptotic covariance matrix of the estimates $\hat{\theta}_o$, $\hat{\underset{\sim}{\theta}}$ is approximately $\left[Ne^{\theta_o} H(\underset{\sim}{\theta})\right]^{-1}$. In particular, the asymptotic variance of $\hat{\theta}_1$ is

$$A \text{ var } \hat{\theta}_1 \simeq \left(Ne^{\theta_o}\right)^{-1} v(\underset{\sim}{\theta}) \ , \tag{5}$$

where $v(\underset{\sim}{\theta})$ is the second diagonal entry of $H^{-1}(\underset{\sim}{\theta})$.

In what follows we shall use the approximation (5) to estimate the sample size needed to test at level α and with power $1 - \beta$ the hypothesis $\theta_1 = 0$ against alternatives $\theta_1 = \tilde{\theta}_1$. To do so, we shall treat the distribution of $\hat{\theta}_1$ as normal with mean θ_1 and approximate variance (5). Then according to normal theory the required sample size N must satisfy

$$Ne^{\theta_o} \geq \left[v^{1/2}(\underset{\sim}{\theta}^o)z_\alpha + v^{1/2}(\underset{\sim}{\tilde{\theta}})z_\beta\right]^2 / \tilde{\theta}_1^2 \ , \tag{6}$$

where $\underset{\sim}{\theta}^o = (0,\theta_2,\ldots,\theta_s)'$, $\underset{\sim}{\tilde{\theta}} = (\tilde{\theta}_1,\theta_2,\ldots,\theta_s)'$ and z_c is the $100(1-c)^{th}$ percentile of the standard normal distribution.

In Section 3 we present sample sizes needed for inferences about the parameter $\theta_1 \equiv \theta$ relating response to a single covariate X. Then in Section 4 we consider the case of more than one covariate.

3. THE UNIVARIATE CASE

It is evident from the definition of H that when s = 1 the second diagonal entry of H^{-1} is simply

$$v(\theta) = [m/(mm_{11} - m_1^2)](\theta) \ . \tag{7}$$

Having specified the error probabilities α and β, the alternative $\tilde{\theta}$, the distribution for X and the approximate response probability e^{θ_o} corresponding to x = 0, the statistician can readily estimate N from (6) and (7). As an illustration, Table 1 shows estimates of Ne^{θ_o} for selected values of α, β and $\tilde{\theta}$ and for the normal, exponential, Poisson and Bernoulli distributions for X. To facilitate comparison of the sample size estimates given by the first three distributions, in these cases X has been normalized to have mean 0 and variance 1. Thus e^{θ_o} approximates the response probability at the mean X value, and $e^{\tilde{\theta}}$ is the odds ratio of response corresponding to an increase in X of one standard deviation.

Estimates of Ne^{θ_o} for alternatives $\tilde{\theta} < 0$ are given for the exponential and Poisson distributions in Table 2. Similar values for the normal and Bernoulli distributions can be obtained from Table 1 by using the symmetry of these distributions.

Table 1 shows that the sample size needed to achieve a given power against a given alternative is quite sensitive to the distribution of the covariate. In particular, relatively small samples are needed in the Bernoulli case when $\pi = 0.5$, and fewer observations are needed to detect a positive association when the covariate is rare ($\pi = 0.1$) than when it is prevalent ($\pi = 0.9$). The latter has been noted for the case of retrospective sampling by Chase and Klauber (1965).

The power of a test about θ, conditional on the observed values $x^{(1)},\ldots,x^{(N)}$, is related to the sample size N by (6) and (7) with the population distribution for X replaced by its empirical distribution (Cox, 1970, p. 87). Table 1 shows that an investigator can achieve substantial reduction in sample size by controlling the values of X in his sample. For example, when X is dichotomous the Bernoulli parameter minimizing the right-hand side of (6) for given $\alpha = \beta$ and $\tilde{\theta} << 1$ is approximately $\pi = 0.5$. Hence for alternatives close to the null hypothesis, maximum conditional power is achieved by choosing the sample so that $x = 1$ for half of the individuals.

4. THE MULTIVARIATE CASE

We now examine the sample size needed to achieve a given power and significance level for tests about θ_1 when the distribution for $\underset{\sim}{X}$ is of a general multivariate exponential type, as described by Bildikar and Patil (1968). In this case the moment generating function for $\underset{\sim}{X}$ is of the form

$$m(\underset{\sim}{t}) = e^{q(\underset{\sim}{\gamma}+\underset{\sim}{t})-q(\underset{\sim}{\gamma})} , \qquad \underset{\sim}{\gamma} + \underset{\sim}{t} \in (\underset{\sim}{a},\underset{\sim}{b}) . \tag{8}$$

Here $(\underset{\sim}{a},\underset{\sim}{b})$ is an interval in Euclidean s-space which may be finite or infinite, $\underset{\sim}{\gamma}$ is a vector of parameters, and q is a real valued function of s variables whose Hessian matrix of second derivatives exists and is positive definite.

This family of multivariate distributions includes among others the multivariate normal distribution, the multinomial distribution, and multivariate Poisson and negative exponential distributions. The mean of $\underset{\sim}{X}$ is given by the vector $\underset{\sim}{q}^{(1)}(\underset{\sim}{\gamma})$ of first partials of q, evaluated at $\underset{\sim}{\gamma}$, and the variance of $\underset{\sim}{X}$ is given by the Hessian $\underset{\sim}{q}^{(2)}(\underset{\sim}{\gamma})$.

It is shown in Theorem 1 of the Appendix that $v(\underset{\sim}{\theta})$ is $e^{q(\underset{\sim}{\gamma})-q(\underset{\sim}{\gamma}+\underset{\sim}{\theta})}$ times the first diagnonal entry of the inverse of $\underset{\sim}{q}^{(2)}$, evaluated at $\underset{\sim}{\gamma} + \underset{\sim}{\theta}$:

$$v(\underset{\sim}{\theta}) = e^{q(\underset{\sim}{\gamma})-q(\underset{\sim}{\gamma}+\underset{\sim}{\theta})}[\underset{\sim}{q}^{(2)}(\underset{\sim}{\gamma} + \underset{\sim}{\theta})]^{-1}_{11} . \tag{9}$$

Before using (9) in (6) to estimate sample sizes we must specifiy the function q. Depending on the form of q, we will also need some functions of $\underset{\sim}{\gamma}$, and functions of $\underset{\sim}{\theta}$ under both null and alternate hypotheses.

To illustrate such a calculation we consider further the special case in which $\underset{\sim}{X}$ has a multivariate normal distribution with mean $\underset{\sim}{\mu}$ and positive definite covariance $\sum$. Then $m(\underset{\sim}{t})$ is given by (8) with $\underset{\sim}{\gamma} = \sum^{-1}\underset{\sim}{\mu}$ and $q(\underset{\sim}{\gamma}) = \underset{\sim}{\gamma}'\sum\underset{\sim}{\gamma}/2$. As verified in Corollary 1 of the Appendix, equation (9) reduces to

$$v(\underset{\sim}{\theta}) = [\text{var } X_1 \exp(\underset{\sim}{\theta}'\underset{\sim}{\mu} + \underset{\sim}{\theta}'\sum\underset{\sim}{\theta}/2)(1 - \rho^2_{1\cdot 2\ldots s})]^{-1} . \tag{10}$$

Here $\rho^2_{1\cdot 2\cdots s}$ is the multiple correlation coefficient relating X_1 to $X_2,\ldots,X_s$, with ρ^2 set to zero when $s = 1$. If each of the covariates X_i has been normalized to have mean 0 and variance 1, then (10) becomes

$$v(\underset{\sim}{\theta}) = \exp(\underset{\sim}{\theta}'\underset{\sim}{\sum}\underset{\sim}{\theta}/2)(1 - \rho^2_{1\cdot 2\cdots s})^{-1} , \qquad (11)$$

where $\sum$ is the correlation matrix of $\underset{\sim}{X}$. Hence in this case sample size estimation via (6) and (11) requires that we specify the correlation coefficient $\rho^2_{1\cdot 2\cdots s}$, the value of $\underset{\sim}{\theta}'\underset{\sim}{\sum}\underset{\sim}{\theta}$ under null and alternative hypotheses, and the approximate response probability e^{θ_o} at the mean covariate levels.

It is evident from (10) that for $s > 1$ the asymptotic variance of $\hat{\theta}_1$ is inflated by the factor $(1 - \rho^2_{1\cdot 2\cdots s})^{-1}$, which achieves its lower bound of one if and only if X_1 is independent of $X_2,\ldots,X_s$. Hence this well-known feature of the least squares estimators in classical multiple regression pertains also to these estimators, provided that the distribution of $\underset{\sim}{X}$ is multivariate normal and the response probability is small. In particular, inclusion of covariates which are correlated with X_1 but independent of response R leads to loss of power.

As a second specialization of (8) we consider a covariate $\underset{\sim}{X}$ having a bivariate Poisson distribution, for which

$$q(\underset{\sim}{\gamma}) = e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_1+\gamma_2} .$$

Thus $EX_i = \text{var } X_i = e^{\gamma_i} + e^{\gamma_1+\gamma_2}$, $i = 1,2$, and the covariance of X_1 and X_2 is $e^{\gamma_1+\gamma_2}$. The parameter of interest θ_1 represents the log-odds-ratio associated with unit increase in X_1 at any level of X_2. In this example (9) becomes

$$v(\underset{\sim}{\theta}) = e^{q(\underset{\sim}{\gamma})-q(\underset{\sim}{\gamma}+\underset{\sim}{\theta})}\left[1 - e^{-\gamma_1-\theta_1}\right]\Big/[1 + q(\underset{\sim}{\gamma} + \underset{\sim}{\theta})] \ . \qquad (12)$$

To determine sample sizes using (6) and (12), we must specify $\underset{\sim}{\gamma}$, e^{θ_o} (which is the approximate response probability at $\underset{\sim}{X} = \underset{\sim}{0}$), and the null and alternate vectors $\underset{\sim}{\theta}^o$ and $\underset{\sim}{\tilde{\theta}}$.

Before presenting an example of the calculations described in the last two sections, we first address the issue of how small the response probability must be in order to achieve reasonable accuracy in our approximations.

5. APPROXIMATION ERROR AND CORRECTION

In this section we examine the error in the sample size approximation (6). We also present a correction for situations when the approximation is not good. Recall that the approximate response probability e^{θ_o} at $x = 0$ is assumed small. Let N and N_1 denote the sample size estimates based upon use of the true asymptotic variance, and of our approximate variance, in (6).

We first treat the univariate case. It is shown in Theorems 2 and 4 of the Appendix that for $\alpha = \beta$ the fractional error $(N-N_1)N_1$ is given by

$$(N-N_1)/N_1 = 2e^{\theta_o}\delta(\tilde{\theta}) + O\left(e^{2\theta_o}\right) \ , \qquad (13)$$

where

$$\delta(\theta) = [v^{1/2}(0) + v^{1/2}(\theta)R(\theta)][v^{1/2}(0) + v^{1/2}(\theta)]^{-1} \ , \qquad (14)$$

and

$$R(\theta) = v(\theta)[m_{11}(2\theta) - 2m^{-1}(\theta)m_1(\theta)m_1(2\theta) + m^{-2}(\theta)m(2\theta)m_1^2(\theta)] \ . \qquad (15)$$

It is clear from (15) and (14) that $R(0) = \delta(0) = 1$. Thus for alternatives $\tilde{\theta}$ close to zero, the fractional error in the suggested sample size estimate is roughly $2e^{\theta_o}$. Also, we see from (13) that when $2e^{\theta_o}\delta(\tilde{\theta})$ is large the correction

$$\tilde{N} = N_1\left[1 + 2e^{\theta_o}\delta(\tilde{\theta})\right] \qquad (16)$$

can be used to obtain more accurate sample size estimates $\tilde{N}$.

When the covariate X has a Bernoulli distribution the exact asymptotic variance of $\hat{\theta}$ and the corresponding sample size N can readily be calculated. Figure 1 shows the curves $(N-N_1)/N_1$ and $(N-\tilde{N})/\tilde{N}$ as functions of e^{θ_o} for an odds ratio $e^{\tilde{\theta}} = 2$ and for $\pi = 0.1$, 0.5 and 0.9. We see that the fractional error in using N_1 is less than 0.10 provided $e^{\theta_o} < 0.03$, and that the corrected estimate $\tilde{N}$ is quite good even for e^{θ_o} as large as 0.10.

Values of $(\tilde{N}-N_1)/N_1$ for the case when X has a normal distribution are shown in Table 2. It is evident from this table that the correction is needed when $e^{\theta_o} \geq 0.05$ and $e^{\tilde{\theta}} = 2$, i.e. when the odds ratio relating the sixteenth and eighty-fourth percentiles of the distribution is four.

For the multivariate case, the correction (16) is too complicated to be useful and is therefore not included in this paper. However for alternatives $\tilde{\underset{\sim}{\theta}} = (\tilde{\theta}_1, \theta_2, \ldots, \theta_s)'$ close to the zero vector, the fractional error in N_1 is again close to $2e^{\theta_o}$ (Appendix, Corollary 2). Thus when necessary, sample size estimates for multivariate regressors can be corrected by the approximate factor $1 + 2e^{\theta_o}$.

6. EXAMPLE

We illustrate the preceding sample size calculations with the problem of testing the null hypothesis that risk of coronary heart disease (CHD) among white males aged 39-59 is unaffected by serum cholesterol levels. According to the data of Hulley et al. (1980), the probability that such an individual will develop CHD during an 18 month study period is approximately 0.07. Under the null hypothesis then, the probability of CHD at the mean serum cholesterol level is 0.07. The cholesterol levels in this population are well represented by a Gaussian distribution. Hence approximate sample sizes needed to detect a given odds-ratio corresponding to a level one standard deviation above the mean are obtained by multiplying appropriate entries from the first column of Table 1 by $(0.07)^{-1}$. For example, approximately N = 582 observations are needed to detect an odds ratio of $e^{0.5} = 1.65$ with $\alpha = 0.05$ significance and $1 - \beta = 0.95$ power, while N = 15,425 observations are needed to detect an odds ratio $e^{0.1} = 1.11$ at these values of α and β.

We see from Figure 2 that for a mean response probability as high as 0.07, correction factors of approximately 1.195 and 1.14 for odds ratios of 1.65 and 1.11 are required. Application of these correction factors to the above sample sizes yields the more accurate estimates N = 695 and N = 17,584 respectively.

We now compare these estimates with the sample size needed to test the above null hypothesis while controlling for the effects of triglyceride. Previous studies indicate that the joint distribution of cholesterol and log triglyceride is bivariate normal with correlation coefficient $\rho = 0.4$, and that the odds ratio of CHD among those with

log triglyceride levels that are one standard deviation above the mean is approximately 1.25 = exp(0.22). Thus to test the hypothesis $\underset{\sim}{\theta}' = (0, 0.22)$ against $\underset{\sim}{\theta}' = (0.5, 0.22)$ with $\alpha = \beta = 0.05$, we use this information in (11) to obtain $v(0, 0.22) = 1.16$ and $v(0.5, 0.22) = 0.98$. Using 0.07 as an estimate of CHD risk at the mean levels of cholesterol and log triglyceride, we find from (6) that N must be at least 661. The correction factor of 1.14 increases the required size to 754 observations. Similar calculations using this approximate correction factor show that if the log odds ratio corresponding to log triglyceride were $\theta_2 = 0.5$ with the remaining parameters unchanged, then N = 665 observations would be needed, while if $\theta_2 = 1$, approximately 435 observations would be required.

REFERENCES

Anderson, T. W. (1958), An Introduction to Linear Statistical Models, New York: John Wiley & Sons.

Bildikar, S. and Patil, G. B. (1968), "Multivariate Exponential-Type Distributions," Annals of Mathematical Statistics, 39, 1316-1326.

Chase, G. and Klauber, M. R. (1965), "A Graph of Sample Sizes for Retrospective Studies," American Journal of Public Health, 55, 1993-1996.

Cox, D. R. (1970), Analysis of Binary Data, London: Methuen & Co., Ltd.

Hulley, S. B., Rosenman, R. A., Bawol, R., and Brandt, D., "Is Serum Cholesterol a Cause of Coronary Heart Disease?" (in press).

APPENDIX

A. MULTIVARIATE EXPONENTIAL-TYPE DISTRIBUTIONS FOR $\underset{\sim}{X}$

When the covariates in the regression can be assumed to have a multivariate exponential-type distribution as defined in Section 5, the function v of formula (6) can be determined from the following result.

Theorem 1. Let $\underset{\sim}{X}$ have a multivariate exponential-type distribution with moment-generating function $m(\underset{\sim}{\theta}) = \exp[q(\underset{\sim}{\gamma} + \underset{\sim}{\theta}) - q(\underset{\sim}{\gamma})]$. Let $H = \begin{pmatrix} m & \underset{\sim}{m}^{(1)'} \\ \underset{\sim}{m}^{(1)} & \underset{\sim}{m}^{(2)} \end{pmatrix}$ be the augmented Hessian of m where, as usual, the superscripts (1) and (2) denote the vector of first partials and Hessian of second partials, respectively. Then

$$H^{-1}(\underset{\sim}{\theta}) = e^{q(\underset{\sim}{\gamma})-q(\underset{\sim}{\gamma}+\underset{\sim}{\theta})} \begin{pmatrix} 1 + \underset{\sim}{q}^{(1)'}\underset{\sim}{q}^{(2)}\underset{\sim}{q}^{(1)}(\underset{\sim}{\gamma}+\underset{\sim}{\theta}) & -\underset{\sim}{q}^{(1)'}\underset{\sim}{q}^{(2)^{-1}}(\underset{\sim}{\gamma}+\underset{\sim}{\theta}) \\ -q^{(2)^{-1}}q^{(1)}(\underset{\sim}{\gamma}+\underset{\sim}{\theta}) & \underset{\sim}{q}^{(2)^{-1}}(\underset{\sim}{\gamma}+\underset{\sim}{\theta}) \end{pmatrix}.$$

In particular, the second diagonal entry v of H^{-1} is given by

$$v(\underset{\sim}{\theta}) = e^{q(\underset{\sim}{\gamma})-q(\underset{\sim}{\gamma}+\underset{\sim}{\theta})}\left[\underset{\sim}{q}^{(2)^{-1}}(\underset{\sim}{\gamma}+\underset{\sim}{\theta})\right]_{11}, \tag{A.1}$$

where the subscript ij denotes the (ij)th entry of a matrix.

Proof. Since $\underset{\sim}{q}^{(2)}$ is a positive definite matrix, there exists a non-singular $s \times s$ matrix $B = B(\underset{\sim}{\theta})$ such that $BB' = \underset{\sim}{q}^{(2)}$. Define the $(s+1) \times (s+1)$ matrix $[\underset{\sim}{q}^{(1)};B]$ by

$$[\underset{\sim}{q}^{(1)};B](\underset{\sim}{\theta}) = \begin{pmatrix} 1 & \underset{\sim}{0}' \\ \underset{\sim}{q}^{(1)}(\underset{\sim}{\theta}) & B(\underset{\sim}{\theta}) \end{pmatrix}.$$

Straightforward calculation shows that H can be written

$$H(\underset{\sim}{\theta}) = e^{q(\underset{\sim}{\gamma}+\underset{\sim}{\theta})-q(\underset{\sim}{\gamma})}[\underset{\sim}{q}^{(1)};B][\underset{\sim}{q}^{(1)};B]' ,$$

where $[\underset{\sim}{q}^{(1)};B]$ is evaluated at $\underset{\sim}{\gamma} + \underset{\sim}{\theta}$. The desired result can now be obtained using $[\underset{\sim}{q}^{(1)};B]^{-1} = [-B^{-1}\underset{\sim}{q}^{(1)};B^{-1}]$.

Corollary 1. Let $\underset{\sim}{X}$ be normally distributed with mean $\underset{\sim}{\mu}$ and positive definite covariance matrix $\sum$. Then

$$v(\underset{\sim}{\theta}) = [\text{var } X_1 \exp(\underset{\sim}{\theta}'\underset{\sim}{\mu} + \underset{\sim}{\theta}'\sum \underset{\sim}{\theta}/2)(1 - \rho^2_{1\cdot 2\ldots})]^{-1} ,$$

where $\rho_{1\cdot 2\ldots s}$ is the multiple correlation coefficient relating X_1 to $X_2,\ldots,X_s$, with $\rho_{1\cdot 2\ldots s} = 0$ when $s = 1$.

Proof. The distribution for $\underset{\sim}{X}$ is of the exponential type, with $q(\underset{\sim}{\gamma}) = \underset{\sim}{\gamma}'\sum \underset{\sim}{\gamma}/2$, where $\underset{\sim}{\gamma} = \sum^{-1}\underset{\sim}{\mu}$. Thus $\underset{\sim}{q}^{(1)}(\gamma) = \underset{\sim}{\mu}$, $\underset{\sim}{q}^{(2)}(\underset{\sim}{\gamma}) = \sum$, and (A.1) becomes

$$v(\underset{\sim}{\theta}) = \exp(-\underset{\sim}{\theta}'\underset{\sim}{\mu} - \underset{\sim}{\theta}'\sum\theta/2)\sum^{-1}_{11} ,$$

where $\sum^{-1}_{11} = [\text{var } X_1(1 - \rho^2_{1\cdot 2\ldots s})]^{-1}$ (Anderson 1958, pp. 32, 344).

B. APPROXIMATION ERROR

Let $\varepsilon = e^{\theta_0} << 1$, and let $\underset{\sim}{X}$ have moment generating function $m(\underset{\sim}{\theta})$. Denote by H_ℓ the matrix given by $(N\varepsilon)^{-1}$ times the sum of the first ℓ terms in the right hand side of equation (4). Thus

$$H_\ell(\underset{\sim}{\theta}) = \sum_{k=1}^{\ell} c_k \varepsilon^{k-1} H(k\underset{\sim}{\theta}) \ , \tag{A.2}$$

where $H(\underset{\sim}{\theta})$ is the augmented Hessian of m. We define $H_\infty(\underset{\sim}{\theta})$ to be $(N\varepsilon)^{-1}$ times the Fisher information $I(\varepsilon,\underset{\sim}{\theta})$. Let $v_\ell(\underset{\sim}{\theta})$ denote the second diagonal entry of $H_\ell^{-1}(\underset{\sim}{\theta})$. Note that $H_1 = H$, $v_1 = v$ of (5), and $(N\varepsilon)^{-1}v_\infty$ is the asymptotic variance of $\hat{\theta}_1$. Finally we set N_ℓ equal to the sample size estimate obtained using v_ℓ in formula (6). Thus $N_\infty = N$ is the estimate obtained using the asymptotic variance of $\hat{\theta}_1$. We first prove the following result.

Theorem 2. For $\ell \geq 2$ and $\alpha = \beta$, N_ℓ and N_1 are related by

$$(N_\ell - N_1)/N_1 = 2\varepsilon[v^{1/2}(\underset{\sim}{\theta}^0)R(\underset{\sim}{\theta}^0) + v^{1/2}(\tilde{\underset{\sim}{\theta}})R(\tilde{\underset{\sim}{\theta}})] \Big/ [v^{1/2}(\underset{\sim}{\theta}^0) + v^{1/2}(\tilde{\underset{\sim}{\theta}})] + O(\varepsilon^2) \ .$$

Here $\underset{\sim}{\theta}^0 = (0,\theta_2,\ldots,\theta_s)'$, $\tilde{\underset{\sim}{\theta}} = (\tilde{\theta}_1,\theta_2,\ldots,\theta_s)'$, and $2\varepsilon R(\underset{\sim}{\theta})$ is, within terms of order ε^2, the fractional difference $(v_\ell - v)/v$ evaluated at $\underset{\sim}{\theta}$:

$$v_\ell(\underset{\sim}{\theta}) = v(\underset{\sim}{\theta})[1 + 2\varepsilon R(\underset{\sim}{\theta})] + O(\varepsilon^2) \ . \tag{A.3}$$

Proof. Rewriting (6) as an equality gives

$$cN_\ell = [v_\ell^{1/2}(\underset{\sim}{\theta}^0) + v_\ell^{1/2}(\tilde{\underset{\sim}{\theta}})]^2 \ , \tag{A.4}$$

where $c = \varepsilon(\tilde{\theta}_1/z_\alpha)^2$. Substitution of (A.3) into (A.4) yields

$$cN_\ell = v(\underset{\sim}{\theta}^0)[1 + 2\varepsilon R(\underset{\sim}{\theta}^0)] + v(\tilde{\underset{\sim}{\theta}})[1 + 2\varepsilon R(\tilde{\underset{\sim}{\theta}})] + 2[v(\underset{\sim}{\theta}^0)v(\tilde{\underset{\sim}{\theta}})]^{1/2} \cdot \{1 + 2\varepsilon[R(\underset{\sim}{\theta}^0) + R(\tilde{\underset{\sim}{\theta}})]\}^{1/2} + O(\varepsilon^2) \ . \tag{A.5}$$

Use of the binomial expansion $(1 + x)^{1/2} = 1 + \frac{1}{2}x + O(x^2)$ in (A.5) gives, after rearrangement of terms,

$$cN_\ell = [v^{1/2}(\underset{\sim}{\theta}^o) + v^{1/2}(\tilde{\underset{\sim}{\theta}})]^2 + 2\varepsilon[v^{1/2}(\underset{\sim}{\theta}^o)R(\underset{\sim}{\theta}^o) + v^{1/2}(\tilde{\underset{\sim}{\theta}})R(\tilde{\underset{\sim}{\theta}})] \cdot [v^{1/2}(\underset{\sim}{\theta}^o) + v^{1/2}(\tilde{\underset{\sim}{\theta}})] + O(\varepsilon^2)$$

$$= cN_1\{1 + 2\varepsilon[v^{1/2}(\underset{\sim}{\theta}^o)R(\underset{\sim}{\theta}^o) + v^{1/2}(\tilde{\underset{\sim}{\theta}})R(\tilde{\underset{\sim}{\theta}})]/[v^{1/2}(\underset{\sim}{\theta}^o) + v^{1/2}(\tilde{\underset{\sim}{\theta}})]\} + O(\varepsilon^2) ,$$

as required.

Theorem 3. $R(\underset{\sim}{0}) = 1$.

Proof. Note in (A.2) that $c_1 = 1$ and $c_2 = -2$. Thus for $\ell \geq 2$,

$$H_\ell(\underset{\sim}{0}) = (1 - 2\varepsilon)H(\underset{\sim}{0}) + O(\varepsilon^2) . \tag{A.6}$$

Inverting both sides of (A.6) and expanding $(1 - 2\varepsilon)^{-1}$ yields

$$H_\ell^{-1}(\underset{\sim}{0}) = (1 + 2\varepsilon)H^{-1}(\underset{\sim}{0}) + O(\varepsilon^2) .$$

In particular, $v_\ell(\underset{\sim}{0}) = v(\underset{\sim}{0})(1 + 2\varepsilon) + O(\varepsilon^2)$. Comparison of this expression with (A.3) yields the desired result.

Theorems 2 and 3 give the following corollary.

Corollary 2. For null and alternative parameters $\underset{\sim}{\theta}$ near the zero vector, the fractional error $(N-N_1)/N_1$ is approximately 2ε plus terms of order ε^2.

Explicit expressions for R in terms of the moment generating function m are unwieldy for $s > 1$. For the univariate case we have the following result.

Theorem 4. For $s = 1$,

$$R(\theta) = v(\theta)[m_{11}(2\theta) - 2m^{-1}(\theta)m_1(\theta)m_1(2\theta) + m^{-2}(\theta)m(2\theta)m_1^2(\theta)] .$$

Proof. For $s = 1$ we have

$$v_\ell(\theta) = [\sum c_k\varepsilon^{k-1}m(k\theta)]/\{[\sum c_k\varepsilon^{k-1}m(k\theta)][\sum c_k\varepsilon^{k-1}m_{11}(k\theta)] - [\sum c_k\varepsilon^{k-1}m_1(k\theta)]^2\} , \qquad (A.7)$$

the summations being taken from $k = 1$ to $k = \ell$. Expanding the denominator in (A.7) and noting that $c_1 = 1$, $c_2 = -2$ yields, after simplification,

$$v_\ell(\theta) = v(\theta)[1 - 2\varepsilon m(2\theta)^{-1}m(\theta) + O(\varepsilon^2)] \cdot \{1 - 2\varepsilon v(\theta)[m_{11}(2\theta) + m^{-1}(\theta)m(2\theta)m_{11}(\theta) - 2m^{-1}(\theta)m_1(\theta)m_1(2\theta)] + O(\varepsilon^2)\}^{-1} . \qquad (A.8)$$

Use of the binomial expansion $(1 - x)^{-1} = 1 + x + O(x^2)$ in (A.8) gives

$$v_\ell(\theta) = v(\theta)[1 - 2\varepsilon m(2\theta)m^{-1}(\theta)][1 + 2\varepsilon v(\theta)\{m_{11}(2\theta) + m^{-1}(\theta)m(2\theta)m_{11}(\theta) - 2m^{-1}(\theta)m_1(\theta)m_1(2\theta)\}] + O(\varepsilon^2) ,$$

or

$$v_\ell(\theta) = v(\theta)[1 + 2\varepsilon R(\theta)] + O(\varepsilon^2) ,$$

where $R(\theta)$ is of the required form.

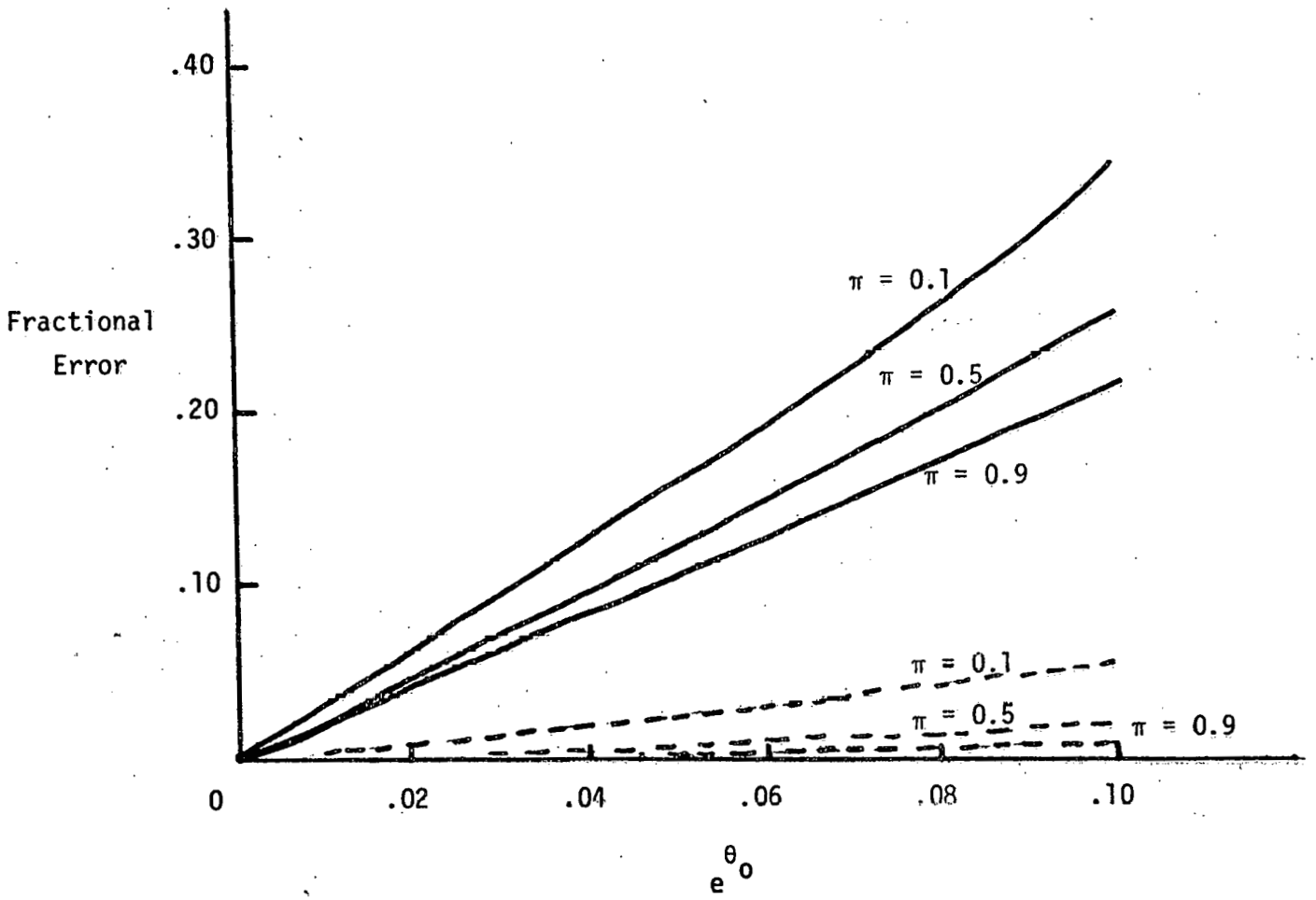

Figure 1. Fractional error $(N-N_1)/N_1$ (———) and $(N-\tilde{N})/\tilde{N}$ (------) for a Bernoulli covariate with parameter π, and with odds ratio $e^{\theta} = 2$. N is the sample size based on use of the asymptotic variance of the maximum likelihood estimate $\hat{\theta}$ in (6). N_1 is the sample size estimate based on use of (7) in (6), and $\tilde{N}$ is the corrected estimate given by (16).

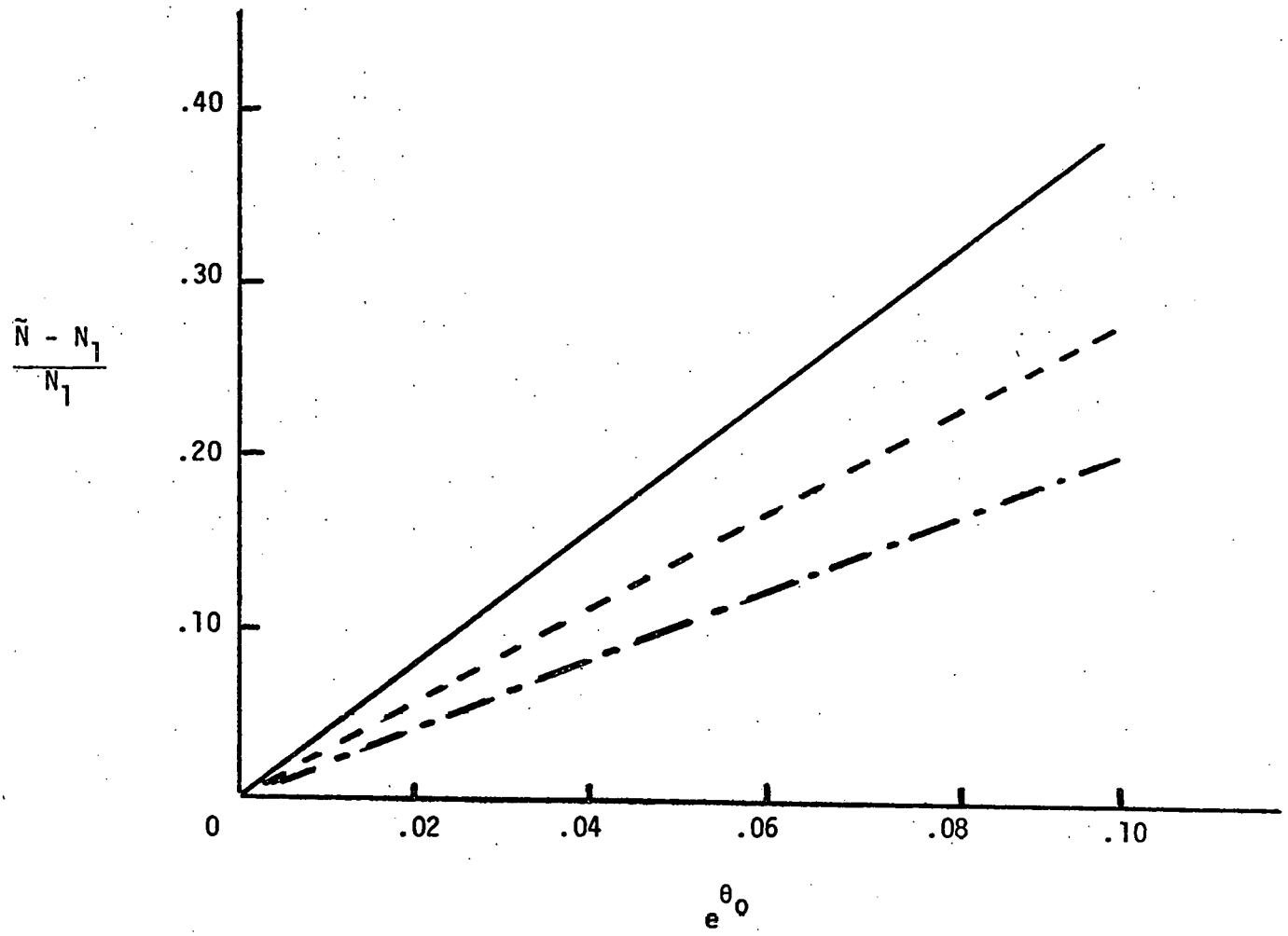

Figure 2. Fractional error $(\tilde{N}-N_1)/N_1$ for a standard normal covariate with odds ratio corresponding to one standard deviation given by $e^{\theta} = 2$ (———), $e^{\theta} = 1.65$ (------) and $e^{\theta} = 1.11$ (—·—·—). N_1 and $\tilde{N}$ are as described in Figure 1.

TABLE 1

Values of Ne^{θ_o}, where N is Sample Size for One-Tailed Test of $\theta = 0$ vs $\vartheta = \tilde{\theta} > 0$

			Distribution of X					
			normal	exponential[a]	Poisson[a]	Bernoulli with parameter π		
	$v(\theta)$		$e^{-\theta^2/2}$	$e^{\theta}(1-\theta)^3$, $\epsilon < 1$	$\exp(1 - e^{\theta})$	$(1-\pi)^{-1} + (\pi e^{\theta})^{-1}$		
α	β	$\tilde{\theta}$				$\pi = .1$	$\pi = .5$	$\pi = .9$
0.01	0.01	0.10	2158.709	1948.758	2054.669	23004.410	8449.242	23931.110
		0.50	81.399	45.750	64.246	782.255	311.265	942.809
		1.00	17.119		10.964	164.968	72.237	232.795
		2.00	2.531		1.466	32.518	16.634	57.486
0.01	0.05	0.10	1573.624	1445.928	1510.669	16891.140	6182.324	17451.760
		0.50	59.949	37.768	49.430	591.472	231.080	689.347
		1.00	13.011		9.137	128.903	54.359	170.578
		2.00	2.148		1.432	26.707	12.712	42.213
0.01	0.10	0.10	1299.457	1208.750	1254.808	14017.530	5118.336	14415.070
		0.50	49.853	33.826	42.325	500.655	193.209	570.421
		1.00	11.052		8.231	111.495	45.866	141.358
		2.00	1.957		1.414	23.843	10.837	35.033
0.05	0.01	0.10	1572.275	1393.376	1483.673	16533.810	5130.852	17423.170
		0.50	58.678	29.380	44.267	548.910	222.568	684.618
		1.00	11.947		6.918	111.809	50.940	168.678
		2.00	1.563		0.757	20.861	11.543	41.563
0.05	0.05	0.10	1079.709	974.398	1027.672	11505.980	4226.008	11969.480
		0.50	40.713	22.882	32.133	391.256	155.684	471.559
		1.00	8.562		5.484	82.511	36.130	116.436
		2.00	1.266		0.733	16.264	8.320	28.752
0.05	0.10	0.10	854.859	781.599	818.719	9157.668	3355.025	9479.520
		0.50	32.475	19.838	26.458	318.099	124.902	374.171
		1.00	6.988		4.787	68.709	29.273	92.533
		2.00	1.120		0.720	14.047	6.817	22.885
0.10	0.01	0.10	1279.578	1135.551	1217.205	13659.080	5046.645	14375.250
		0.50	48.082	21.864	35.133	441.371	181.352	563.834
		1.00	9.570		5.140	87.684	41.104	138.712
		2.00	1.142		0.474	15.701	9.208	34.128
0.10	0.05	0.10	854.330	760.951	808.113	9056.555	3334.802	9468.281
		0.50	31.975	16.464	24.430	301.376	121.558	372.313
		1.00	6.570		3.915	61.993	27.330	91.787
		2.00	0.890		0.455	11.750	6.357	22.630
0.10	0.10	0.10	655.768	591.807	624.163	6988.234	2566.595	7269.742
		0.50	24.727	13.898	19.516	237.632	94.556	286.405
		1.00	5.200		3.330	50.114	21.944	70.718
		2.00	0.769		0.445	9.878	5.053	17.463

a) With parameter $\lambda = 1$, and normalized to have mean zero.

TABLE 2

Values of Ne^{θ_o}, where N is Sample Size for One-Tailed Test of $\theta = 0$ vs $\theta = \tilde{\theta} < 0$

			Distribution of X	
			exponential[a]	Poisson[a]
	$v(\theta)$		$e^{\theta}(1-\theta)^3$, $\theta < 1$	$\exp(1 - e^{\theta})$
α	β	$\tilde{\theta}$		
0.01	0.01	-1.00	39.896	30.433
		-0.50	127.867	106.408
		-0.10	2380.076	2270.852
0.01	0.05	-1.00	26.502	20.999
		-0.50	87.594	74.949
		-0.10	1706.730	1641.191
0.01	0.10	-1.00	20.478	16.683
		-0.50	69.230	60.427
		-0.10	1393.449	1347.236
0.05	0.01	-1.00	31.757	23.383
		-0.50	98.920	80.164
		-0.10	1761.987	1668.188
0.05	0.05	-1.00	19.954	15.221
		-0.50	63.955	53.222
		-0.10	1190.429	1135.800
0.05	0.10	-1.00	14.779	11.584
		-0.50	48.420	41.107
		-0.10	931.406	893.694
0.10	0.01	-1.00	27.797	20.004
		-0.50	85.005	67.691
		-0.10	1470.418	1384.842
0.10	0.05	-1.00	16.843	12.521
		-0.50	52.870	43.156
		-0.10	953.117	904.302
0.10	0.10	-1.00	12.119	9.245
		-0.50	38.843	32.324
		-0.10	723.014	689.835

a) With parameter $\lambda = 1$, and normalized to have mean zero.