SIZES AND POWERS OF TESTS FOR MODELS WITH SAMPLE SELECTION BIASES -A MONTE CARLO COMPARISON-

Kazumitsu Nawata
Department of Advanced Social and International Studies
University of Tokyo
and
Michael McAleer
Department of Economics
University of Western Australia

Abstract

The t-test is widely used in hypothesis testing in models of qualitative choice. However, the t-test can sometimes perform poorly and yield misleading results, especially when the sample size is small. This paper analyses the finite sample size and power properties of the t-test in models with sample selection biases. Three versions of the t-test are compared with the likelihood ratio and Lagrange multiplier tests, which are asymptotically equivalent to the t-test. The finite sample problems with the t-test are shown to be much more serious than in models such as binary choice models, and it is recommended that the t-test not be used in such models. Powers of the tests are also presented.

1. Introduction

For the single equation hypothesis given by $H_0: h(\beta) = 0$, the asymptotically equivalent t-test and Wald test are the most widely used tests in hypothesis testing. The test statistic is calculated as $h(\hat{\beta})$/ (estimated standard error). However, since the standard errors are usually calculated from the asymptotic covariance matrix, the t-test does not behave well in some models and can sometimes yield misleading results. One such example is the case of a nonlinear hypothesis. Gregory and Veall (1985) compared the hypotheses $H_0: \beta_1 \beta_2 = 1$ and $H_0^*: \beta_1 = 1/\beta_2$. Although the two nonlinear hypotheses are mathematically identical, the t-test based on the second hypothesis was shown to perform quite poorly. Lafontaine and White (1986) considered $H_0: \beta_1^k = 1$ and showed that the test statistic can take arbitrary values by appropriate selection of k.

When the model considered is not the standard linear regression model, the t-test may not perform well even for a simple linear hypothesis, such as the test of significance of an individual coefficient. Griffiths et al. (1987) analysed hypothesis testing in binary choice models and showed that the t-test for the simple coefficient does not yield the correct size and rejects the true null hypothesis too frequently in finite samples.

This paper examines the finite sample properties of the t-test in models with sample selection biases (for details of the model, see Amemiya (1985)), which are widely used in various fields of economics such as labour economics. Problems with the t-test are shown to be much more serious than in models such as binary choice models, and it is recommended that the t-test should not be used in such models. First, using the Monte Carlo technique, the t-test is compared with the likelihood ratio and Lagrange multiplier tests, which are asymptotically equivalent to the t-test. The performance of the likelihood ratio test is shown to be far superior to the t-test under the true null hypothesis. The powers of the tests are also compared.

2. The Model and its Maximum Likelihood Estimator

The model considered in this paper is

$$(2.1)\ y_{1i} = x_{1i}'\alpha + u_{1i}$$
$$d_i = 1(y_{1i} > 0)$$
$$y_{2i} = x_{2i}'\beta + u_{2i},\ i = 1,2,\ldots, N,$$

where $1(\bullet)$ is an indicator function such that $1(\bullet) = 1$ if $\bullet$ is true and 0 otherwise. y_{1i} is not observable and only the sign of y_{1i} (i.e. d_i) is observable. y_{2i} is observable if and only if $y_{1i} > 0$ (i.e. $d_i = 1$). u_{1i} and u_{2i} are jointly normal with zero means, variances 1 and σ_2^2, respectively, and covariance σ_{12}. N is the number of observations. Heckman's (1976, 1979) two step estimator is widely used to estimate the model. However, since Heckman's two step estimator often performs poorly (see, e.g., Nawata (1993,1994), and Nawata and Nagase (1996)), it is sensible to estimate the model by the maximum likelihood method. Setting $\rho = \sigma_{12}/\sigma_2$, the log-likelihood function is given by

$$(2.2)\ \ln L(\theta) = \sum_{i=1}^{N} \ln f_i(\theta)$$
$$\ln f_i(\theta) = (1 - d_i)\log[1 - \Phi(x_{1i}'\alpha)]$$
$$+ d_i[\log\Phi[\{x_{1i}'\alpha + \rho/\sigma_2(y_{2i} - x_{2i}'\beta)\}(1 - \rho^2)^{-1/2}]$$
$$- \log\sigma_2 + \log\phi[\sigma_2^{-1}(y_{2i} - x_{2i}'\beta)]]\ ,$$

where $\theta' = (\alpha', \beta', \sigma_2, \rho)$, and ϕ and Φ are the density and distribution functions of the standard normal distribution, respectively.

Standard methods such as those used in the LIMDEP, STAT, and TSP computer software packages are incomplete. The problems are:
1. The procedures sometimes do not converge.
2. Owing to the existence of local maxima, the results may not be correct even if the procedures do converge (see Olsen (1982)).
For further details, see Nawata (1995). In this paper, the maximum likelihood estimator (MLE) $\hat{\theta}$ is obtained by the procedure suggested in Nawata (1994, 1995), which modifies the method given in Olsen (1982). The procedure is:

i) Choose M_1 equidistant points from (-1,1). Let δ be the distance between any two points. Let $\rho = 0$ and calculate $\hat{\alpha}_0$, $\hat{\beta}_0$, and $\hat{\sigma}_0$ which maximise the conditional likelihood function. Note that these estimators are the Probit MLE and the least squares estimator using the $y_{1i} > 0$ observations in the first and second equations of (2.1), respectively.

ii) Let $\hat{\alpha}_j$, $\hat{\beta}_j$, and $\hat{\sigma}_j$ be the j-th round estimators. Increase ρ by δ and choose the initial values of the iteration as the j-th round estimators $\hat{\alpha}_j$, $\hat{\beta}_j$, and $\hat{\sigma}_j$ and calculate the (j+1)-th round estimators. Since the likelihood function is a continuous function, the previous estimators are considered to be in the neighbourhood of the maximum values.

iii) Continue (ii) and calculate the estimators up to the largest values of ρ determined in (i).

iv) In the same way, calculate the estimators from 0 to the smallest value of ρ .

v) Choose the values which maximise the likelihood function.

vi) Choose M_2 points in the neighbourhood of the value of ρ determined in (v), and repeat the procedure.

vii) Determine the final estimators.

3. Monte Carlo Experiments

3.1 The Basic Model and the Null Hypothesis

In this section, the performance of the t-test is evaluated by Monte Carlo experiments. The basic model for the Monte Carlo experiment is:

$$(3.1)\ y_{1i} = \alpha_0 + \alpha_1 x_{1i} + u_{1i}$$
$$d_i = 1(y_{1i} > 0)$$
$$y_{2i} = \beta_0 + \beta_1 x_{2i} + u_{2i},\ i = 1,2,\ldots, N.$$

x_{1i} and x_{2i} are independent variables, and u_{1i} and u_{2i} are jointly normal random variables. They are determined as:

$$(3.2)\ x_{1i} = \xi_{1i}$$
$$x_{2i} = [\eta\,\xi_{1i} + (1 - \eta)\,\xi_{2i}]/\sqrt{\eta^2 + (1 - \eta)^2},$$

and

$$(3.3)\ u_{1i} = \epsilon_{1i}$$
$$u_{2i} = 10\bullet[\pi\epsilon_{1i} + (1 - \pi)\epsilon_{2i}]/\sqrt{\pi^2 + (1 - \pi)^2}$$
$$\rho_0 = \pi/\sqrt{\pi^2 + (1 - \pi)^2}\ .$$

ξ_{1i} and ξ_{2i} are independent and uniformly distributed on (0,20], and ϵ_{1i} and ϵ_{2i} are independent standard normal random variables. ρ_0 is the true correlation coefficient of u_{1i} and u_{2i}.

The null hypothesis is given by:

$$(3.4)\ H_0\colon \rho = 0.$$

This hypothesis tests the existence of sample selection biases and is always important in this model (i.e. if there are no sample selection biases, the model can be estimated by the Probit MLE and least squares estimator).

The following items are considered.

i) Sizes and powers of the tests, with π =0.0, 0.1, 0.2, 0.3, 0.4 ,0.5.

ii) Effects of the correlation of x_{1i} and x_{2i}, with η =0.0, 0.8, 1.0.

iii) Effects of sample sizes, with N =200 and 400.

The true parameter values of α's and β'$s$are:

$$(3.5)\ \alpha_0 = -1.0, \quad \alpha_1 = 0.1, \\ \beta_0 = -10.0, \quad \beta_1 = 1.0.$$

Since the degree of censoring is close to 50% in many empirical examples , the degree of censoring is chosen to be 50%. The MLE is calculated by the scanning method described in Section 2, and ρ is chosen from [-0.99, 0.99] with an interval of 0.01 in Step (i), after which it is chosen in the neighbourhood of the maximum value with an interval of 0.001. The number of repetitions is 5000 for each case.

3.2 Test Statistics

The t-test, Likelihood Ratio (LR) test, and Lagrange Multiplier (LM) test are evaluated in the experiments. To compare these tests directly, the t-test statistic is squared, so that the asymptotic distributions of the three tests have the χ^2 distribution with one degree of freedom. Since the number of observations is at least 200 and the number of unknown parameters is just 6, differences in the t- and standard normal distributions are ignored. The test statistics for the three tests are calculated as follows.

1. <u>t-test</u>

$$(3.6)\ t^2 = \hat{\rho}^2 \,/\, \hat{V}(\hat{\rho}) \,.$$

$\hat{V}(\hat{\rho})$ is the estimated variance of $\hat{\rho}$, and is obtained from using the following three different methods.

i) The inverse of the Hessian matrix , $-I(\theta)^{-1}$, evaluated at $\theta = \hat{\theta}$, where $I(\theta) = \partial^2 \ln L / \partial\theta\partial\theta'$.

ii) The inverse of the sum of the outer products of the first derivatives of $\ln f_i$, $J(\theta)^{-1}$,evaluated at $\theta = \hat{\theta}$, where $J(\theta) = \sum \partial \ln f_i / \partial\theta \bullet \partial \ln f_i / \partial\theta'$.

iii) $I(\theta)^{-1} J(\theta) I(\theta)^{-1}$ evaluated at $\theta = \hat{\theta}$.

The test statistics based on (i), (ii) and (iii) are denoted in this paper as t_1^2, t_2^2 and t_3^2. An alternative method of evaluating the asymptotic variances is the inverse of the information matrix. However, since the likelihood function is a complicated form, calculation of the information matrix is quite difficult and is not practical for this model.

2. <u>LR Test</u>

$$(3.7)\ LR = 2[\ln L(\hat{\theta}) - \ln L(\bar{\theta})]$$

where $\bar{\theta}$ is the constrained MLE under H_0.

3. <u>LM Test</u>

$$(3.8)\ LM = -\frac{\partial \ln L}{\partial\theta'}\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right]^{-1}\frac{\partial \ln L}{\partial\theta}$$

evaluated at $\theta = \bar{\theta}$, the constrained MLE under H_0.

4. Results of the Monte Carlo Experiments

4.1 Sizes of the tests

The asymptotic distributions of the test statistics follow the χ^2distribution with one degree of freedom, $\chi^2(1)$. The results of the Monte Carlo experiments for the five test statistics (including t_1^2, t_2^2 and t_3^2 for the t-test) are given in Tables 1 - 6. 90% - 99% denote the percentiles of the estimated test statistics and the corresponding $\chi^2(1)$ values. Max. denotes the maximum calculated test statistic.

Table 1. Estimates of the Test Statistics when ρ_0 = 0, η = 0 , N = 200

Test	90%	95%	99%	Max.
t_1^2	4.71	9.15	38.46	4925.6
t_2^2	4.40	9.31	44.92	4810.3
t_3^2	5.21	10.12	38.44	47.03
LR	2.85	4.25	7.47	18.29
LM	3.01	4.64	9.33	31.41
$\chi^2(1)$	2.70	3.84	5.02	-

Table 2. Estimates of the Test Statistics when ρ_0 = 0, η = 0 , N = 400

Test	90%	95%	99%	Max.
t_1^2	3.53	4.95	11.58	49.85
t_2^2	3.24	4.88	12.16	72.56
t_3^2	3.63	5.45	14.01	55.03
LR	2.85	3.88	6.60	14.83
LM	2.99	4.00	7.23	20.07

Table 3. Estimates of the Test Statistics when $\rho_0 = 0$, $\eta = 0.8$, $N = 200$

Test	90%	95%	99%	Max.
t_1^2	58.43	123.6	424.6	4835
t_2^2	53.37	115.8	454.3	4315
t_3^2	70.98	154.8	543.2	7458
LR	3.10	4.36	7.15	16.65
LM	1.69	3.41	16.48	172.6

Table 4. Estimates of the Test Statistics when $\rho_0 = 0$, $\eta = 0.8$, $N = 400$

Test	90%	95%	99%	Max.
t_1^2	27.78	53.69	164.9	740.3
t_2^2	26.03	48.93	163.8	681.9
t_3^2	32.36	60.59	192.3	959.8
LR	2.91	3.97	6.85	11.40
LM	2.23	4.76	19.19	239.2

Table 5. Estimates of the Test Statistics when $\rho_0 = 0$, $\eta = 1.0$, $N = 200$

Test	90%	95%	99%	Max.
t_1^2	86.71	187.8	765.5	5866
t_2^2	78.70	168.4	838.1	4004
t_3^2	50.75	91.24	245.1	1620
LR	3.05	4.37	7.45	16.13
LM	0.24	0.51	2.78	326.2

Table 6. Estimates of the Test Statistics when $\rho_0 = 0$, $\eta = 1.0$, $N = 400$

Test	90%	95%	99%	Max.
t_1^2	47.39	81.57	216.8	1291
t_2^2	44.82	76.46	241.6	1023
t_3^2	107.0	226.5	1035.0	3583
LR	2.73	3.94	6.36	13.25
LM	0.31	0.65	2.83	168.8

The results of the three different t-tests ($t_1^2 \sim t_3^2$) are quite similar. However, the distributions of $t_1^2 \sim t_3^2$ are surprisingly different from $\chi^2(1)$, especially when N is small and η is large. They have much heavier tails than $\chi^2(1)$, the absolute t-values are quite large and even exceed 10 in some trials, despite the fact that the null hypothesis is correct. Although the *LR* test statistic is excessively large, its distributions are much closer to $\chi^2(1)$ than are $t_1^2 \sim t_3^2$ for all cases. The distributions of the *LM* test statistic are close to $\chi^2(1)$ for cases where $\eta = 0$; however, the *LM* test statistic takes negative values in many trials for cases where $\eta = 0.8$ and 1.0. The percentage of trials in which the *LM* test statistic becomes negative is given in the following table. It can be seen that the situation is particularly bad when $\eta = 0.8$ and $\eta = 1.0$, even when $N = 400$.

Table 7. Percentage of Trials in which the *LM* Test Statistic is Negative

N	$\eta = 0$	$\eta = 0.8$	$\eta = 1.0$
200	0.0%	32.9%	45.4%
400	0.0%	26.8%	44.7%

Tables 8 and 9 give the sizes of the tests for the significance levels 5% and 1%, respectively. (Although the *LM* test statistic becomes negative for many trials, the tests are conducted using the simple rule: "reject the null hypothesis if the test statistic is larger than the critical value obtained from the $\chi^2(1)$ distribution". Therefore, the results of the *LM* test may not be reliable in these cases.)

The performances of the three t-tests are extremely poor. All of $t_1^2 \sim t_3^2$ reject the correct null hypothesis far too frequently, especially for cases where $\eta = 0.8$ and 1.0. When the significance level is 5%, t_1^2 rejects the null hypothesis 12.1%, 8.2%, 40.3%, 34.2%, 46.3% and 40.7% of the time for ($\eta = 0$, $N = 200$), ($\eta = 0$, $N = 400$), ($\eta = 0.8$, $N = 200$), ($\eta = 0.8$, $N = 400$), ($\eta = 1.0$, $N = 200$) and ($\eta = 1.0$, $N = 400$); when the significance level is 1%, the rejection frequencies are 7.2%, 2.9%, 33.7%, 26.7%, 41.0% and 34.1%.

Although t_2^2 seems slightly better than t_1^2, it still rejects the null hypothesis 11.6%, 7.7%, 37.0%, 31.8%, 44.4% and 38.2% at the 5% significance level, and 6.9%, 2.8%, 31.3%, 24.7%, 38.8% and 32.6% at the 1% significance level. t_3^2 rejects the null hypothesis 13.4%, 9.4%, 49.6%, 41.4%, 51.3% and 44.7% at the 5% significance level, and 8.1%, 3.3%, 40.6%, 31.7%, 43.7% and 37.1% at the 1% significance level. Except for the case where ($\eta = 0$, $N = 400$), t_3^2 is the worst among the three t-tests. These results suggest the three versions of the t-test are unreliable.

Although the LR test rejects the null too frequently, the performance of the *LR* test is much better than the three t-tests in all cases. The *LR* test rejects the null hypothesis 6.0%, 5.2%, 6.6%, 5.4%, 6.5% and 5.2% at the 5% significance level, and 1.5%, 1.0%, 1.3%, 1.2%, 1.5% and 0.9% at the 1% significance level, which is far superior to the t-tests.

Table 8. Sizes of the Tests at the 5% Significance Level

N	η = 0		η = 0.8		η = 1.0	
	200	400	200	400	200	400
t_1^2	12.1%	8.2%	40.3%	34.2%	46.3%	40.7%
t_2^2	11.6%	7.7%	37.0%	31.8%	44.4%	38.2%
t_3^2	13.4%	9.4%	49.6%	41.4%	51.3%	44.7%
LR	6.0%	5.2%	6.6%	5.4%	6.5%	5.2%
LM	6.9%	5.7%	4.6%	6.2%	0.6%	0.8%

Table 9. Sizes of the Tests at the 1% Significance Level

N	η = 0		η = 0.8		η = 1.0	
	200	400	200	400	200	400
t_1^2	7.2%	2.9%	33.7%	26.7%	41.0%	34.1%
t_2^2	6.9%	2.8%	31.3%	24.7%	38.8%	32.6%
t_3^2	8.1%	3.3%	40.6%	31.7%	43.7%	32.1%
LR	1.5%	1.0%	1.3%	1.2%	1.5%	0.9%
LM	2.3%	1.2%	2.6%	3.6%	0.3%	0.5%

Although the performance of the *LM* test may not be reliable when η = 0.8 and 1.0, it rejects the null hypothesis 6.9%, 5.7%, 4.6%, 6.2%, 0.6% and 0.8% at the 5% significance level, and 2.3%, 1.2%, 2.6%, 3.6%, 0.3% and 0.5% at the 1% significance level.

4.2 Powers of the Tests

In order to evaluate the powers of the tests, the cases where $\pi = 0.1, 0.2, 0.3, 0.4$ and 0.5 are considered in addition to $\pi = 0$. The powers of the tests at the 5% significance level are given in Figures 1-6, in which T1 ~ T3 represent the powers of $t_1^2 \sim t_3^2$. As before, the *LM* test statistic takes on negative values in many trials.

When $\eta = 0$, the powers increase rapidly as π increases, at almost the same rate for all tests. The powers of t_1^2 and t_2^2 are approximately 5% higher than for the *LR* and *LM* tests, for all values of π, and the powers of the *LR* and *LM* tests are very similar. The power of t_3^2 is similar to t_1^2 and t_2^2 for $N = 200$, and it is about 10% higher than t_1^2 and t_2^2 for $N = 400$ and $\pi \geq 0.2$. (Since the size of t_3^2 is also closest to 5%, t_3^2 seems best among the three t-tests in this case.)

When $\eta = 0.8$ and 1.0, the powers of the tests do not increase considerably for small values of π. The powers of the *LR* test are about 30~45% lower than for the t-tests for all values of π. The powers of the *LM* test are the lowest and do not increase for these cases; indeed, the *LM* test is almost powerless where $\eta = 1.0$. Although the powers of t_3^2 are highest and t_2^2 are lowest among the three t-tests, the differences in the powers of the three tests are not very large.

5. Conclusion

This paper has examined three versions of the t-test, *LR* and *LM* tests for testing hypotheses in models with sample selection biases. The t-test rejects the true null hypothesis too frequently, especially when the sample size is small, whereas the *LR* test performs much better than the t-test for all sample sizes under the true null hypothesis. These results of the paper suggest that, although it is the most widely used test, the t-test in models with sample selection biases should be interpreted cautiously.

References

Amemiya, T., (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, MA.

Gregory, A., and M. R. Veall, (1985), "On Formulating Wald Tests of Nonlinear Restrictions", *Econometrica*, 53, 1465-68.

Griffiths, W. E., R. C. Hill, and P. J. Pope, (1987) "Small Sample Properties of Probit ModelEstimators", *Journal of the American Statistical Association*, 82, 929-937.

Heckman, J., (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection Bias and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement*, 5, 475-492.

________, (1979), "Sample Selection Bias as a Specification Error", *Econometrica* ,47,153-161.

Lafontaine, F., and K. J. White, (1987), "Obtaining Any Wald Statistic You Want",*Economics Letters*, 21, 35-40.

Nawata, K., (1993), "A Note on Estimation of Models with Sample Selection Biases", *Economics Letters*, 42, 15-24.

______, (1994), "Estimation of the Sample-Selection Biases Models by the Maximum Likelihood Estimator and Heckman's Two-Step Estimator in Models with Sample-Selection Biases", *Economics Letters*, 45, 33-40.

______, (1995), "Estimation of Sample-Selection Models by the Maximum Likelihood Method", *Mathematics and Computers in Simulation*, 39, 299-303.

Nawata, K., and N. Nagase, (1996), "Estimation of Sample Selection Bias Models," *Econometric Reviews*, 15, 387-400.

Olsen, R. J., (1982), "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator", *International Economic Review*, 23, 223-240.

Figure 1. Power of the Tests (Significance Level = 5%, η = 0, N = 200)

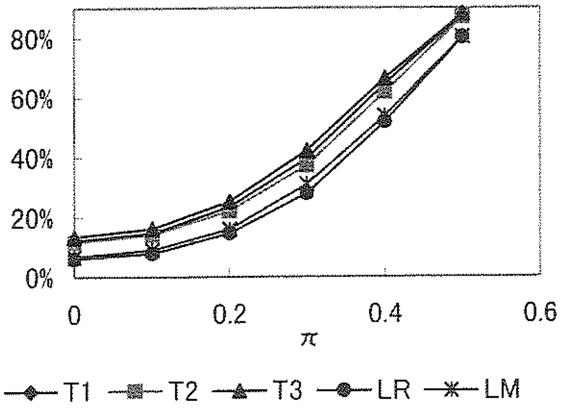


Figure 2. Power of the Tests (Significance level = 5%, η = 0, N = 400)

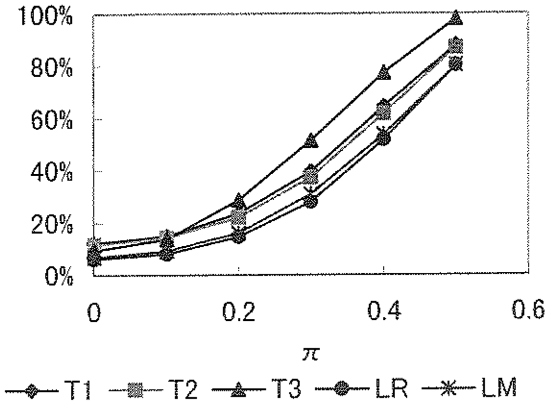


Figure 3. Power of the Tests (Significance Level = 5%, η = 0.8, N = 200)

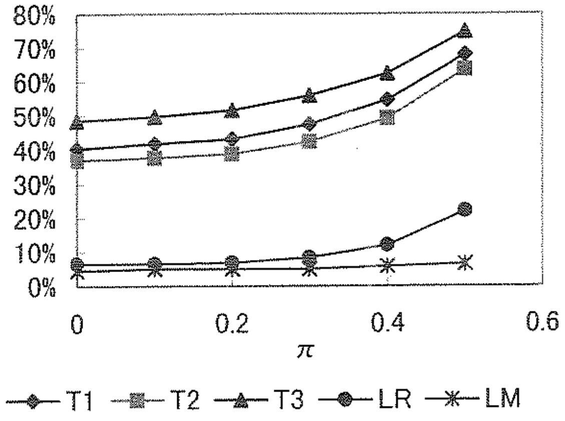


Figure 4. Power of the Tests (Significance Level = 5%, η = 0.8, N = 400)

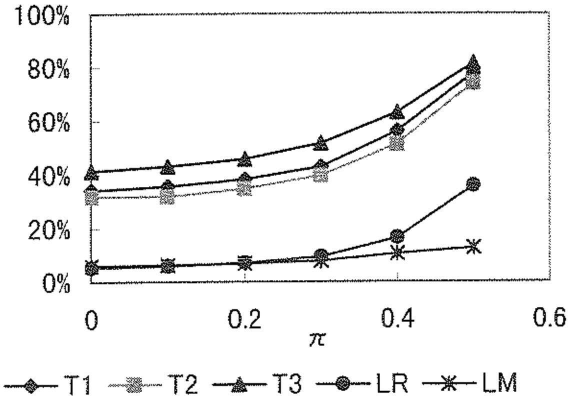


Figure 5. Power of the Tests (Significance Level = 5%, η = 1.0, N = 200)

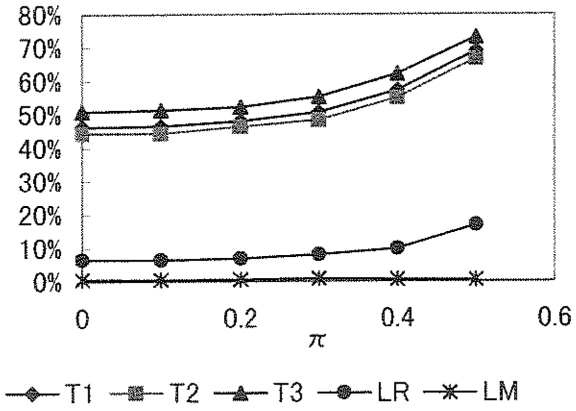
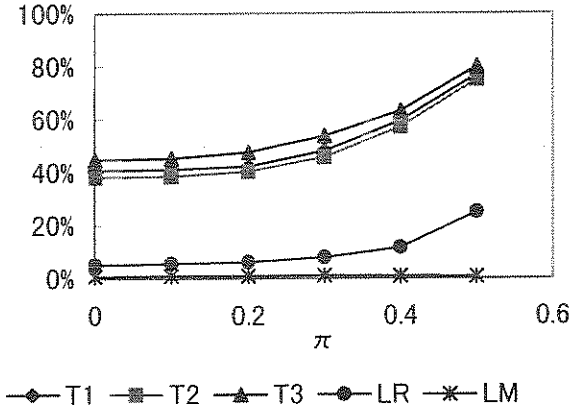


Figure 6. Power of the Tests (Significance Level = 5%, η = 1.0, N = 400)


π
—◆—T1 —■—T2 —▲—T3 —●—LR —✳—LM