ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Dorota Pekasiewicz**

SEQUENTIAL PROBABILITY RATIO TEST FOR MEAN BASED ON PSEUDO-LIKELIHOOD FUNCTION

Abstract

Hypotheses about expected value of random variable can be verified by means of the parametric sequential probability ratio test in case of the known class of this variable's distribution. The problem with verification of such hypotheses occurs when we have no information about random variable distribution. Then, we have to apply non-parametric methods.

The author of the paper proposes the application of pseudo-likelihood function instead of likelihood one in the statistic of sequential probability ratio test. Examples of application of the test based on the likelihood function ratio in selected kinds of distributions are presented together with the results of Monte Carlo analysis concerning properties of these tests.

Key words: sequential probability ratio test, likelihood function, normal distribution.

1. Introduction

Sequential probability ratio tests presented in the subject literature are used to verify hypotheses about random variable parameters such as expected value or variance. They can be applied if probability function or density function of investigated random variable are known. Information about the random variable distribution enables us to determine likelihood functions and their ratios in case of n-element sample. Calculation of the likelihood function ratio, under the assumption that alternative and null hypotheses are true, allows us to make a decision about acceptance of one of the verified hypotheses or about extra sampling of another unit for the sample, taking into consideration that probabilities of type 1 and type 2 errors are predetermined.

* Ph.D., Chair of Statistical Methods, University of Łódź.

In conducted statistical investigations, however, we do not always have information about the class of distribution of investigated random variable. In these circumstances, we can apply modified sequential probability ratio test. The modification consists in replacing, in the formula for statistic being the test gauge, likelihood functions with pseudo-likelihood ones. Such a transformation can cause a change in the test's properties and in particular, probabilities of errors related to making wrong decisions can be stronger than estimated values. Monte Carlo analysis enables us to detect the change in these properties in the particular examples of applications of sequential test based on the pseudo-likelihood function ratio.

2. Pseudo-likelihood function and its application in sequential tests

Let X be the random variable and θ investigated parameter of this variable. Let us consider the following simple hypotheses about the value of the parameter θ:

$$H_0: \theta = \theta_0$$

$$H_1: \theta = \theta_1, \qquad \text{where } \theta_1 > \theta_0.$$

The above hypotheses can be verified by means of the sequential probability ratio test, whose statistic has the following form:

$$I_n = \ln \frac{L(X_1, X_2, ..., X_n; \theta_1)}{L(X_1, X_2, ..., X_n; \theta_0)} = \ln L(X_1, X_2, ..., X_n; \theta_1) - \ln L(X_1, X_2, ..., X_n; \theta_0) \tag{1}$$

where $X_1, ..., X_n$ is the random sample in n-th step of the sequential procedure, while $L(X_1, ..., X_n, \theta)$ is the likelihood function determined on the basis of probability or density functions of discrete or continuous random variable X, respectively.

If the class of distribution of random variable X is unknown, we can replace likelihood function with pseudo-likelihood one.

The pseudo-likelihood function is defined by the formula:

$$\tilde{L}(X_1, ..., X_n, \theta) = \prod_{i=1}^{n} l(X_i, \theta) \tag{2}$$

where l is probability function or density function of the particular distribution belonging to the linear-exponential distributions family.

Linear-exponential distributions family includes, among others, the following distributions: binomial, Poissson, gamma, univariate normal distribution.

In this case, the statistic of the sequential probability ratio test, which verifies hypotheses about the parameter θ, has the following form:

$$I_n = \ln \frac{\tilde{L}(X_1, X_2, ..., X_n; \theta_1)}{\tilde{L}(X_1, X_2, ..., X_n; \theta_0)} = \ln \tilde{L}(X_1, X_2, ..., X_n; \theta_1) - \ln \tilde{L}(X_1, X_2, ..., X_n; \theta_0) \quad (3)$$

where $\tilde{L}(X_1, ..., X_n, \theta_i)$, for $i = 0, 1$, are determined on the basis of n-element sample.

Selection of linear-exponential distribution depends on values that are taken by the investigated variable X. If X is univariate random variable of real values, we can use normal distribution to construct pseudo-likelihood function. If we have more precise information about values taken by random variable, we can use other distributions from linear-exponential family. And for example, if X is a discrete variable taking values from finite set of natural numbers, then we can apply binomial distribution, and if X takes non-negative values, then we use Poisson distribution. For variable taking only positive values, we can use gamma distribution (see D o m a ń s k i, P r u s k a, 2000).

If we use density function of normal distribution having an independent sample sampling scheme, pseudo-likelihood function is expressed by the following formula:

$$\ln \tilde{L}(X_1, ..., X_n, \theta, \sigma) = \sum_{i=1}^{n} \left\{ \ln \left(\sqrt{2\pi\sigma} \right)^{-1} - \frac{(X_i - \theta)^2}{2\sigma^2} \right\} \quad (4)$$

If an average value is a verified parameter of random variable X, then the statistic of sequential test for the average is of the following form:

$$I_n = \frac{n}{2\sigma^2} \left(\theta_0^2 - \theta_1^2 \right) + \frac{(\theta_1 - \theta_0)}{\sigma^2} \sum_{i=1}^{n} X_i \quad (5)$$

If probabilities of type 1 and type 2 errors α and β are determined, we verify formulated hypotheses. At every stage of sequential procedure, we calculate the value of the statistic I_n and compare it with constants $A = ln \frac{1-\beta}{\alpha}$ and $B = ln \frac{\beta}{1-\alpha}$. At the same time, we make a decision about acceptance of one of the verified hypotheses or about increasing the sample by re-sampling another unit (see P e k a s i e w i c z, 1997).

For sequential tests based on the pseudo-likelihood function ratio, it is impossible to calculate the expected value of the sample size necessary to make a decision about accepting one of the verified hypotheses.

3. Examples of applications of sequential test for the average based on pseudo-likelihood function

In order to investigate possibilities of applying sequential probability ratio tests based on pseudo-likelihood function of verification of hypotheses of the population's average value, simulation experiments were carried out.

Experiments can be divided into three groups. The first group includes these experiments in which the parent population has the distribution χ^2 of k degrees of freedom. In the second one, the population has the distribution being a mixture of two distributions χ^2 of k and l degrees of freedom, respectively. Whereas, in the third group the population's distribution is a mixture of two normal distributions. Distributions of population (of a large number of degrees of freedom, except for χ^2) differ from the normal distribution significantly.

Experiments revealed that verifying hypotheses, probabilities of type 1 and type 2 errors will not exceed 0.05. Formulated hypotheses about the average population were verified 10 000 times by means of sequential test based on the pseudo-likelihood function ratio, while the function used was density function of normal distribution. In each of the groups of experiments, in most cases, hypotheses were formulated so that the null hypothesis was true. However, such pairs of hypotheses in which the alternative hypothesis was true were also verified. The parameter σ occurring in density function of normal distribution was determined at various levels in order to study its impact on the test's result and the sample size. Determining numbers of wrong decisions, probabilities of type 1 and type 2 errors were estimated. The analysis of these results enabled us to answer the question whether the application of pseudo-likelihood function caused significant increase in the number of wrong decisions. Having $\alpha = \beta = 0.05$ and the number of the experiment's repetitions 10 000 times assumed, the number of wrong decisions should not exceed 500.

Table 1 represents results concerning verification of hypotheses about the average value of population of distribution χ^2 of k degrees of freedom. The number of degrees of freedom was determined at various levels (k=2, 3, 10, 20). Only small quantities of degrees of freedom were considered because then population's distribution differs from normal distribution significantly. Table 1 shows real population's parameters and values taken by the parameter θ in zero and alternative hypotheses. Values of the parameter σ were determined at various levels but on average, the value close to the standard deviation of the population was assumed. Table 1 comprises numbers of correct and wrong decisions while applying the considered test. In Table 2 numerical characteristics of random variable, i.e. the sample size was presented for selected sequential tests for the population of distribution χ^2.

Table 1

Number of decisions about acceptance of hypotheses H_0 and H_1 for population of distribution χ^2

Population's parameters	θ_0	θ_1	σ	Acceptance of H_0	Acceptance of H_1
$k=2$ $\mu=1.9699$ $\sigma=1.9486$	2	3	2	9 487	513
			3	9 968	32
			4	10 000	0
	2	2,5	2	9 618	382
			3	9 991	9
			4	10 000	0
	1	2	2	120	9 880
			3	3	9 997
			4	0	10 000
$k=3$ $\mu=2.9718$ $\sigma=2.4051$	3	5	2	8 988	1 012
			3	9 824	176
			4	9 980	20
	3	4	2	8 960	1 040
			3	9 858	142
			4	9 994	6
	2	3	2	681	9 319
			3	37	9 963
			4	0	10 000
$k=10$ $\mu=9.9337$ $\sigma=4.4091$	10	15	4	9 397	603
			5	9 779	221
			6	9 940	60
	10	12	4	9 304	696
			5	9 802	198
			6	9 954	46
	8	10	4	595	9 405
			5	134	9 866
			6	21	9 979
$k=20$ $\mu=19.9643$ $\sigma=6.2967$	20	25	6	9 479	521
			7	9 788	212
			8	9 998	2
	20	23	6	9 454	546
			7	9 768	232
			8	9 910	90
	17	20	6	422	9 578
			7	166	9 834
			8	54	9 946

Source: own calculations.

Table 2

Numerical characteristics of the sample size in sequential tests in case of population of distribution χ^2 of k-degrees of freedom

k	θ_0	θ_1	σ	$\overline{n}$	s_n	min.	max.
2	2	3	2	21.3790	13.8571	1	153
			3	51.4366	26.2041	11	287
			4	90.6125	34.5542	27	317
	2	2.5	2	80.4241	53.7094	3	525
			3	190.8621	96.4866	42	1 139
			4	339.1554	129.9826	104	1 375
	1	2	2	28.4466	20.7431	1	181
			3	60.3618	32.0633	2	288
			4	104.2913	42.1275	21	389
3	3	5	2	5.5734	3.6387	1	37
			3	13.7683	8.0751	1	88
			4	24.2960	11.4974	7	119
	3	4	2	19.3220	13.8300	1	134
			3	51.1260	30.0424	4	296
			4	91.8114	43.2059	21	383
	2	3	2	25.2976	19.7475	1	157
			3	60.0095	37.5562	5	313
			4	103.7714	50.8703	14	471
10	10	15	4	4.1664	2.6173	1	24
			5	6.4983	3.8179	1	36
			6	9.2514	4.9612	1	54
	10	12	4	20.9641	14.6109	2	149
			5	35.3388	22.4286	6	222
			6	51.4042	28.6858	7	252
	8	10	4	25.5359	19.1692	1	171
			5	41.9539	28.4566	4	304
			6	60.3726	36.2686	6	331
20	20	25	6	8.7706	5.8792	1	50
			7	12.2373	7.7372	1	82
			8	18.2445	10.0345	2	113
	20	23	6	22.3284	15.6231	2	152
			7	31.8516	20.8084	5	212
			8	42.0141	25.3918	5	244
	17	20	6	25.1335	18.4165	1	172
			7	35.1477	23.7212	3	203
			8	46.0725	29.0025	4	326

Source: own calculations.

On the basis of the obtained results, we can note that in analysed cases, the application of pseudo-likelihood function in sequential test for the average value brought satisfactory results, i.e. the number of wrong decisions, in most cases, did not exceed the set quantity. Only if the assumed parameter σ was too little, the number of wrong decisions exceeded the set value twice. The selection of larger value σ brought better results but the sample size necessary to make a decision about acceptance of one of the verified hypotheses increased significantly (Table 2). The results presented in Table 2 confirm an obvious fact that smaller differences between values θ_0 and θ_1 bring about increase in the sequential sample size. In some of the analysed cases, the average value of the sample size increased over four times.

Tables 3 and 4 present results concerning verification of hypotheses about the average value of population of distribution being a mixture of two distributions χ^2 of $k = 3$ and $l = 15$ degrees of freedom, whereas two cases were considered. In the first one, a half of population's elements had distribution χ^2 of $k = 3$, and the other half had distribution χ^2 of $k = 15$. In the second case $\frac{1}{4}$ of population had distribution χ^2 of $k = 3$, and $\frac{3}{4}$ had distribution χ^2 of $k = 15$. Populations constructed like that had bimodal distribution of parameters given in tables, which differed much from normal distribution.

Table 3

Number of decisions about acceptance of hypotheses H_0 and H_1 for population of distribution being a mixture of two distributions χ^2

Version	Population's parameters	θ_0	θ_1	σ	Acceptance of H_0	Acceptance of H_1
Mixture of distributions in proportion 1:1	$\mu = 8.9156$ $\sigma = 7.2740$	9	12	7	9 456	544
				8	9 744	256
				9	9 900	100
		9	10	7	9 605	395
				8	9 844	156
				9	9 964	54
		7	9	7	541	9 459
				8	250	9 750
				9	98	9 902
Mixture of distributions in proportion 1:3	$\mu = 11.9315$ $\sigma = 7.0866$	12	15	7	9 623	377
				8	9 831	169
				9	9 942	58
		12	13	7	9 673	327
				8	9 874	126
				9	9 961	39
		10	12	7	507	9 493
				8	237	9 763
				9	83	9 917

Source: own calculations.

Table 4

Numerical characteristics of the sample size in sequential tests in case of population of distribution being a mixture of two distributions χ^2

Version	Population's parameters	θ_0	θ_1	σ	$\overline{n}$	s_n	min.	max.
Mixture of distributions in proportion 1:1	$\mu = 8.9156$ $\sigma = 7.2740$	9	12	7	29.2126	20.1725	4	173
				8	39.9631	26.0689	5	200
				9	51.7331	31.5447	4	279
		9	10	7	235.9860	165.5565	26	1 714
				8	322.5097	208.6677	35	2 043
				9	414.8981	249.4172	59	2 853
		7	9	7	74.7990	56.5367	5	557
				8	101.9941	71.3244	7	600
				9	131.2250	85.3465	11	806
Mixture of distributions in proportion 1:3	$\mu = 11.9315$ $\sigma = 7.0866$	12	15	7	31.1498	21.9002	4	203
				8	41.5367	26.7694	4	257
				9	52.8716	31.3086	8	324
		12	13	7	243.7281	171.4463	28	1 388
				8	327.9857	214.0067	39	2 115
				9	420.3586	250.0663	67	2 241
		10	12	7	73.7039	53.6201	5	491
				8	100.3717	68.4758	7	615
				9	129.8055	82.1655	15	868

Source: own calculations.

Table 5

Number of decisions about acceptance of hypotheses H_0 and H_1 for population of distribution being a mixture of two normal distributions $N(2; 2)$ and $N(10; 3)$

Version	Population's parameters	θ_0	θ_1	σ	Acceptance of H_0	Acceptance of H_1
Mixture of distributions in proportion 1:1	$\mu = 6.0020$ $\sigma = 4.7274$	6	10	4	9 242	758
				5	9 764	236
				6	9 941	59
		6	8	4	9 037	963
				5	9 691	309
				6	9 920	80
		4	6	4	804	9 196
				5	236	9 764
				6	56	9 944
Mixture of distributions in proportion 1:3	$\mu = 8.0014$ $\sigma = 4.4208$	8	12	4	9 633	367
				5	9 933	67
				6	9 987	13
		8	10	4	9 444	556
				5	9 862	138
				6	9 971	29
		10	8	4	31	9 969
				5	1	9 999
				6	0	10 000

Source: own calculations.

Table 6

Numerical characteristics of the sample size in sequential tests in case of population of distribution being a mixture of two normal distributions $N(2; 2)$ and $N(10; 3)$.

Version	Population's parameters	θ_0	θ_1	σ	$\overline{n}$	s_n	min.	max.
Mixture of distributions in proportion 1:1	$\mu = 6.0020$ $\sigma = 4.7274$	6	10	4	6.3291	4.4395	1	50
				5	10.2641	6.7689	2	64
				6	14.5978	8.7425	3	86
		6	8	4	21.1680	15.4753	3	157
				5	37.0820	25.0224	5	250
				6	54.5660	33.0363	8	344
		4	6	4	22.4346	16.5688	2	159
				5	37.7785	25.8705	4	229
				6	55.4512	34.8644	7	452
Mixture of distributions in proportion 1:3	$\mu = 8.0014$ $\sigma = 4.4208$	8	12	4	7.0787	5.0423	1	57
				5	10.7927	6.9813	2	64
				6	14.8552	8.3302	3	77
		8	10	4	23.6228	17.5173	3	223
				5	38.4115	25.2709	4	248
				6	55.2982	31.8338	6	321
		10	8	4	6.3817	4.8452	1	41
				5	10.7105	7.9781	2	76
				6	15.6416	10.0046	2	115

Source: own calculations.

Analogous investigations were carried out for population of distribution being a mixture of two normal distributions $N(2; 2)$ and $N(10; 3)$. Similarly to a mixture of distributions χ^2, two versions of mixtures were considered. In the first one, a half of elements of population had distribution $N(2; 2)$, and the other half had distribution $N(10; 3)$, while in case of the second one, $\frac{1}{4}$ of population had distribution $N(2; 2)$, whereas $\frac{3}{4}$ – distribution $N(10; 3)$. These populations had bimodal distribution differing much from normal distribution.

In groups of experiments in which population has the distribution being a mixture of either two distributions χ^2 or two normal distributions, the application of sequential test based on pseudo-likelihood function to verification of hypotheses about the average value of population resulted in acceptance of true hypotheses of probabilities of type 1 and type 2 errors not stronger than the set one. Similarly to the first group of experiments, the value of assumed parameter σ was of great importance. The value of this parameter, together with the value of the parameter occurring in null and alternative hypotheses affected the sequential sample size significantly.

4. Final remarks

The obtained results of simulations, enable us to claim that the sequential probability ratio test based on the pseudo-likelihood function can be applied to verify hypotheses about the average value of population in case of unknown class of the population's distribution. The application of the test of density function of normal distribution in statistics, despite the fact that the population's distribution differs from the normal one considerably, contributed to making a decisions about acceptance of true hypotheses of probabilities of type 1 and type 2 errors not exceeding the set value 0.05.

In sequential test based on pseudo-likelihood ratio with density function of normal distribution, the assumed quantity of the parameter σ is of great importance. The larger value of this parameter reduces the number of wrong decisions but, at the same time, increases the sample size necessary to make a decision about acceptance of one of the verified hypotheses. It turned out that the best results were obtained when the parameter σ was a little bigger than the real value of the population's standard deviation.

Despite satisfactory results of the conducted simulations, the sequential test based on pseudo-likelihood function requires further analyses. They should concern other classes of populations' distributions and application of other density functions or functions of probabilities of distributions of the family of linear-exponential distributions.

References

D o m a ń s k i Cz., P r u s k a K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.

P e k a s i e w i c z D. (1997), *Application of the sequential probability ratio test to verification of statistical hypotheses*, „Acta Universitatis Lodziensis", Folia Oeconomica **141**, 111–119.

Dorota Pekasiewicz

Ilorazowy test sekwencyjny dla średniej oparty na funkcji pseudowiarygodności

Hipotezy o wartości oczekiwanej zmiennej losowej możemy zweryfikować parametrycznym ilorazowym testem sekwencyjnym, w przypadku znanej klasy rozkładu tej zmiennej. Problem z weryfikacją takich hipotez pojawia się, gdy nie posiadamy informacji o rozkładzie zmiennej losowej i musimy zastosować metody nieparametryczne.

W pracy proponowane jest wykorzystanie funkcji pseudowiarygodności, zamiast funkcji wiarygodności, w statystyce ilorazowego testu sekwencyjnego. Przykłady zastosowania testu opartego na ilorazie funkcji pseudowiarygodności dla wybranych rodzajów rozkładów są zaprezentowane w pracy wraz z wynikami analizy Monte Carlo dotyczącymi własności tych testów.