A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 285, 2013

Dorota Pekasiewicz[*]

BAYESIAN STATISTICAL TESTS FOR PROPORTION FOR INDEPENDENT AND DEPENDENT SAMPLING[1]

Abstract. As a result of the use of the Bayesian statistical tests, the decision of the acceptance of the hypothesis for which the posterior risk is lower, is made. The risk depends on the prior parameter's distribution, the loss function and the sampling scheme.

In the paper, the Bayesian statistical tests for the proportion, for different prior distributions and independent and dependent sampling, are considered. Apart from theoretical considerations, the results of simulation studies on the properties of these tests are presented.

Key words: Bayesian test, risk function, independent sample, dependent sample, prior distribution.

I. IDEA OF THE BAYESIAN TESTS

Let Θ be the set of admissible parameter values θ and Θ_0, Θ_1 – nonempty subsets of Θ which satisfy the condition: $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

Let us formulate the null hypothesis:

$$H_0 : \theta \in \Theta_0 , \tag{1.1}$$

against the alternative:

$$H_1 : \theta \in \Theta_1 . \tag{1.2}$$

When we have the information about prior distribution $g(\theta)$ of parameter θ, we can verify the above hypotheses using the Bayesian statistical tests (see Domański Cz., Pruska K., (2000), French S., Rios Insua D., (2000), Krzyśko M., (2004)). On the basis of random sample $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ we make one of the two decisions:

[*] Ph. D., Department of Statistical Methods, University of Łódź.

[1] The research was supported by the project number DEC-2011/01/B/HS4/02746 from the National Science Centre.

- d_0 – decision about acceptance of hypothesis H_0,
- d_1 – decision about acceptance of hypothesis H_1.

We accept the hypothesis whose posterior risk is smaller.

The posterior distribution depends on prior distribution (see Szreder M., (1994) and on the loss function.

We define the loss function:

$$L(\theta, d_i(\mathbf{x})) = \begin{cases} 0 & \text{dla} \quad \theta \in \Theta_i \\ c_i & \text{dla} \quad \theta \in \Theta - \Theta_i \end{cases} \quad \text{for} \quad i=0,1, \tag{1.3}$$

where $\mathbf{x}$ is a realization of random sample $\boldsymbol{X}$ and c_0, c_1 are fixed values.

The risk function is defined as $r(d, \mathbf{x}) = EL(\theta, d(\mathbf{x}))$.

For discrete prior distribution of parameter θ, the risk function is expressed by the formula:

$$\begin{aligned} r(d_i, \mathbf{x}) &= \sum_{\theta_k} L(\theta_k, d(\mathbf{x})) \cdot g(\theta_k \mid \mathbf{x}) = \\ &= \sum_{\theta_k \in \Theta_i} L(\theta_k, d(\mathbf{x})) \cdot g(\theta_k \mid \mathbf{x}) + \\ &+ \sum_{\theta_k \in \Theta - \Theta_i} L(\theta_k, d(\mathbf{x})) \cdot g(\theta_k \mid \mathbf{x}) = \\ &= 0 + c_i \sum_{\theta_k \in \Theta - \Theta_i} g(\theta_k \mid \mathbf{x}) = c_i P(\theta \in \Theta - \Theta_i \mid \mathbf{x}), \end{aligned} \tag{1.4}$$

where $g(\theta_k \mid \mathbf{x})$ is the posterior distribution of θ.

For continuous prior distribution of parameter θ, the risk function has the following form:

$$\begin{aligned} r(d_i, \mathbf{x}) &= \int_{\Theta} L(\theta, d_i) g(\theta \mid \mathbf{x}) d\theta = \int_{\Theta_i} L(\theta, d_i) g(\theta \mid \mathbf{x}) d\theta + \\ &+ \int_{\Theta - \Theta_i} L(\theta, d_i) g(\theta \mid \mathbf{x}) d\theta = 0 + c_i \int_{\Theta - \Theta_i} g(\theta \mid \mathbf{x}) d\theta = c_i P(\theta \in \Theta - \Theta_i \mid \mathbf{x}), \end{aligned} \tag{1.5}$$

where $g(\theta \mid \mathbf{x})$ is the posterior distribution of θ.

The acceptance of the null hypothesis is connected with the inequality (see Domański Cz., Pruska K., (2000)):

$$c_0 P\left(\theta \in \Theta_1 \middle| \mathbf{x}\right) < c_1 P\left(\theta \in \Theta_0 \middle| \mathbf{x}\right), \tag{1.6}$$

being true. This is equivalent to

$$\frac{P\left(\theta \in \Theta_0 \middle| \mathbf{x}\right)}{P\left(\theta \in \Theta_1 \middle| \mathbf{x}\right)} > \frac{c_0}{c_1}. \tag{1.7}$$

For $\Theta_0 \cup \Theta_1 = \Theta$, the region of acceptance of the null hypothesis is the following:

$$\left\{ \mathbf{x}: \quad P\left(\theta \in \Theta_0 \middle| \mathbf{x}\right) > \frac{c_0}{c_0 + c_1} \right\}. \tag{1.8}$$

II. BAYESIAN TESTS FOR PROPORTION FOR INDEPENDENT SAMPLING

Let X be the two point random variable with probability function $P(X = x) = xp + (1-x)(1-p)$ for $x = 0, 1$.

Let us formulate the null hypothesis:

$$H_0 : p \le p_0, \tag{2.1}$$

against the alternative:

$$H_1 : p > p_0, \tag{2.2}$$

where p_0 is a fixed value.

We verify these hypotheses on the basis of independent random sample $\boldsymbol{X} = \left(X_1, X_2, ..., X_n\right)$.

We consider two cases of prior distribution of parameter p:

- discrete uniform distribution,
- uniform distribution on the interval $[a, b]$.

Firstly, let us assume that the probability function of p has the following form: $g\left(p_k\right) = P\left(p = p_k\right) = \frac{1}{l}$ for $k=1,2,\ldots,l$ and the loss function is expressed by formula (1.3).

The posterior distribution is of the following form:

$$g(p_k|\mathbf{x}) = \frac{f(\mathbf{x}|p_k)\cdot g(p_k)}{\sum_{j=1}^{l} f(\mathbf{x}|p_j)\cdot g(p_j)} = \frac{{p_k}^m(1-p_k)^{n-m}}{\sum_{j=1}^{l} {p_j}^m(1-p_j)^{n-m}} \text{ ,for } k=1,2,\ldots,l, \quad (2.3)$$

where m is the number of "ones" in the sample and

$$P(p \le p_0|\mathbf{x}) = \sum_{p_k \le p_0} g(p_k|\mathbf{x}) = \frac{\sum_{p_k \le p_0} {p_k}^m(1-p_k)^{n-m}}{\sum_{j=1}^{l} {p_j}^m(1-p_j)^{n-m}}. \quad (2.4)$$

If $P(p \le p_0|\mathbf{x}) > \frac{c_0}{c_0+c_1}$ we accept the hypothesis H_0.

Next, we consider the case, when p has the prior uniform distribution on the interval $[a, b]$, where $a \ge 0$ and $b \le 1$.

The posterior distribution of parameter p has the following form:

$$\begin{aligned} g(p|\mathbf{x}) &= \frac{f(\mathbf{x}|p)\cdot g(p)}{\int f(\mathbf{x}|p)\cdot g(p)dp} = \frac{p^m(1-p)^{n-m}}{\int_a^b p^m(1-p)^{n-m}dp} = \\ &= \frac{p^m(1-p)^{n-m}}{B(m+1,n-m+1)(F(b)-F(a))}, \end{aligned} \quad (2.5)$$

where F is the cumulative distribution function of the beta distribution with parameters $m+1$ and $n-m+1$.

Therefore,

$$\begin{aligned} P(p \le p_0|\mathbf{x}) &= \int_a^{p_0} \frac{p^m(1-p)^{n-m}}{B(m+1,n-m+1)(F(b)-F(a))}dp = = \\ &= \frac{1}{B(m+1,n-m+1)(F(b)-F(a))}\int_a^{p_0} p^m(1-p)^{n-m}dp = \frac{F(p_0)-F(a)}{F(b)-F(a)}. \end{aligned} \quad (2.6)$$

We compute probability $P\left(p \le p_0 \middle| \mathbf{x}\right)$, compare it with $\frac{c_0}{c_0 + c_1}$ and accept either H_0 or H_1 .

III. BAYESIAN TESTS FOR PROPORTION FOR DEPENDENT SAMPLING

Let $\boldsymbol{X} = \left(X_1, X_2, ..., X_n\right)$ be random sample drawn without replacement from the population containing N –elements.

In this case, similarly as for independent sampling scheme, we consider discrete and continuous uniform prior distributions of parameter *p*.

For discrete prior distribution of parameter *p* with probability function $g\left(p_k\right) = P\left(p = p_k\right) = \frac{1}{l}$ for k=1,2,…,l, the posterior probability function has the following form:

$$g\left(p_k \middle| \mathbf{x}\right) = \frac{f\left(\mathbf{x} \middle| p_k\right) \cdot g\left(p_k\right)}{\sum_{j=1}^{l} f\left(\mathbf{x} \middle| p_j\right) \cdot g\left(p_j\right)} = \frac{\prod_{i=0}^{m-1}\left(Np_k - i\right) \cdot \prod_{s=0}^{n-m-1}\left(N\left(1-p_k\right) - s\right) \cdot \frac{1}{l}}{\sum_{j=1}^{l}\left(\prod_{i=0}^{m-1}\left(Np_j - i\right) \cdot \prod_{s=0}^{n-m-1}\left(N\left(1-p_j\right) - s\right)\right) \cdot \frac{1}{l}} =$$

$$= \frac{\prod_{i=0}^{m-1}\left(Np_k - i\right) \cdot \prod_{s=0}^{n-m-1}\left(N\left(1-p_k\right) - s\right)}{\sum_{j=1}^{l}\left(\prod_{i=0}^{m-1}\left(Np_j - i\right) \cdot \prod_{s=0}^{n-m-1}\left(N\left(1-p_j\right) - s\right)\right)}, \tag{3.1}$$

when $m \neq 0$ and $m \neq n$.

For $m = 0$ we obtain:

$$g\left(p_k \middle| \mathbf{x}\right) = \frac{\prod_{s=0}^{n-1}\left(N\left(1-p_k\right) - s\right)}{\sum_{j=0}^{l}\prod_{s=0}^{n-1}\left(N\left(1-p_j\right) - s\right)} \tag{3.2}$$

and for $m = n$: $g(p_k|\mathbf{x}) = \dfrac{\prod_{i=0}^{n-1}(Np_k - i)}{\sum_{j=1}^{l}\prod_{i=0}^{n-1}(Np_j - i)}$. (3.3)

In this case

$$P(p \le p_0|\mathbf{x}) = \begin{cases} \dfrac{\sum_{p_k \le p_0}\left(\prod_{i=0}^{m-1}(Np_k - i)\cdot\prod_{s=0}^{n-m-1}(N(1-p_k)-s)\right)}{\sum_{j=1}^{l}\left(\prod_{i=0}^{m-1}(Np_j - i)\cdot\prod_{s=0}^{n-m-1}(N(1-p_j)-s)\right)} & \text{for } m \ne 0 \wedge m \ne n, \\ \dfrac{\sum_{p_k \le p_0}\prod_{s=0}^{n-1}(N(1-p_k)-s)}{\sum_{j=1}^{l}\prod_{s=0}^{n-1}(N(1-p_j)-s)} & \text{for } m = 0, \\ \dfrac{\sum_{p_k \le p_0}\prod_{i=0}^{n-1}(Np_k - i)}{\sum_{j=1}^{l}\prod_{i=0}^{n-1}(Np_j - i)} & \text{for } m = n. \end{cases} \quad (3.4)$$

For the uniform distribution of the interval [a, b], where $a \ge 0$ and $b \le 1$, the posterior distribution function has the form:

$$g(p|\mathbf{x}) = \frac{f(\mathbf{x}|p)\cdot g(p)}{\int_a^b f(\mathbf{x}|p)\cdot g(p)dp} = \frac{\prod_{i=0}^{m-1}(Np - i)\cdot\prod_{s=0}^{n-m-1}(N(1-p)-s)\cdot\frac{1}{b-a}}{\int_a^b\left(\prod_{i=0}^{m-1}(Np - i)\cdot\prod_{s=0}^{n-m-1}(N(1-p)-s)\frac{1}{b-a}\right)dp} = \frac{\prod_{i=0}^{m-1}(Np - i)\cdot\prod_{s=0}^{n-m-1}(N(1-p)-s)}{\int_a^b\left(\prod_{i=0}^{m-1}(Np - i)\prod_{s=0}^{n-m-1}(N(1-p)-s)\right)dp}, \quad (3.5)$$

when $m \ne 0$ and $m \ne n$.

$$\text{For } m=0:\ g(p|\mathbf{x})=\frac{\prod_{s=0}^{n-1}(N(1-p)-s)}{\int_a^b\prod_{s=0}^{n-1}(N(1-p)-s)dp}, \tag{3.6}$$

$$\text{and for } m=n:\ g(p|\mathbf{x})=\frac{\prod_{i=0}^{n-1}(Np-i)}{\int_a^b\prod_{i=0}^{n-1}(Np-i)dp}. \tag{3.7}$$

The posterior distribution function is expressed by formula:

$$P(p\le p_0|\mathbf{x})=\begin{cases}\dfrac{\int_a^{p_0}\left(\prod_{i=0}^{m-1}(Np-i)\cdot\prod_{s=0}^{n-m-1}(N(1-p)-s)\right)dp}{\int_a^b\left(\prod_{i=0}^{m-1}(Np-i)\cdot\prod_{s=0}^{n-m-1}(N(1-p)-s)\right)dp} & \text{for } m\neq 0\wedge m\neq n,\\ \dfrac{\int_a^{p_0}\prod_{s=0}^{n-1}(N(1-p)-s)dp}{\int_a^b\prod_{s=0}^{n-1}(N(1-p)-s)dp} & \text{for } m=0,\\ \dfrac{\int_a^{p_0}\prod_{i=0}^{n-1}(Np-i)dp}{\int_a^b\prod_{i=0}^{n-1}(Np-i)dp} & \text{for } m=n.\end{cases} \tag{3.8}$$

IV. ANALYSIS OF THE PROPERTIES OF THE BAYESIAN TESTS FOR PROPORTION

In order to analyze the properties of the Bayesian tests for proportion, the populations were generated with the two point distribution with parameter *p*.

We considered the following prior distribution of parameter *p*:

a) discrete uniform distribution with probability function $P\left(p=\frac{k}{10}\right)=\frac{1}{10}$, for *k*=1, …, 9, 10 (distribution D1),

b) discrete uniform distribution with probability function $P(p = p_k) = \frac{1}{10}$ for $p_k = 0{,}2 + 0{,}05k$, k=1, 2, …, 10 (distribution D2),

c) continuous uniform distribution on the interval [0, 1] (distribution D3),

d) continuous uniform distribution on the interval [0.2, 0.7] (distribution D4).

We assume that the loss function is of the form:

$$L(p, d_0) = \begin{cases} 0 & \text{dla } p \le p_0 \\ c_0 & \text{dla } p > p_0 \end{cases}, \; L(p, d_1) = \begin{cases} 0 & \text{dla } p > p_0 \\ c_1 & \text{dla } p \le p_0 \end{cases} \text{ and } c_0 = c_1.$$

For the generated populations and different fixed values p_0 and different sample sizes n, we made R=1000 repetitions of the Bayesian procedure of hypotheses verification. Both dependent and independent samples were drawn. For the dependent sampling a finite population of N=1000 elements was generated.

The results of the Monte Carlo analysis for selected values of p_0 and sample sizes are presented in tables 4.1–4.2.

In these tables the following notations are used:

LF_0 – the number of acceptance decisions of false H_0,

LF_1 – the number of acceptance decisions of false H_1,

n – sample size.

The figures 4.1–4.4 present percentage of false decisions $(LF_1/R, LF_0/R)$ for independent and dependent sampling.

Table 4.1. The percentage of acceptance decision of false H_0 and H_1 for independent random sampling and fixed prior distribution of parameter p

p_0	n	distribution D1		distribution D2		distribution D3		distribution D4	
		LF_1/R	LF_0/R	LF_1/R	LF_0/R	LF_1/R	LF_0/R	LF_1/R	LF_0/R
1	2	3	4	5	6	7	8	9	10
0.30	50	0.027	0.028	0.052	0.043	0.032	0.020	0.063	0.046
	100	0.011	0.016	0.030	0.043	0.017	0.016	0.035	0.030
0.35	50	0.027	0.028	0.048	0.056	0.032	0.032	0.061	0.050
	100	0.011	0.016	0.042	0.042	0.020	0.017	0.038	0.031
0.40	50	0.031	0.027	0.042	0.058	0.035	0.027	0.070	0.055
	100	0.015	0.017	0.041	0.029	0.014	0.023	0.040	0.036
0.45	50	0.031	0.027	0.051	0.048	0.037	0.027	0.055	0.058
	100	0.015	0.017	0.049	0.046	0.029	0.012	0.050	0.028

Table 4.1 (cont.)

1	2	3	4	5	6	7	8	9	10
0.50	50	0.032	0.032	0.044	0.060	0.019	0.035	0.046	0.064
	100	0.015	0.022	0.027	0.055	0.020	0.022	0.039	0.040
0.55	50	0.032	0.032	0.048	0.045	0.027	0,030	0.047	0.049
	100	0.015	0.022	0.030	0.034	0.016	0.017	0.046	0.045
0.60	50	0.015	0.021	0.047	0.046	0.016	0.029	0.040	0.051
	100	0.011	0.010	0.031	0.033	0.019	0.016	0.037	0.040
0.65	50	0.015	0.021	0.025	0.053	0.027	0.026	0.023	0.063
	100	0.011	0.010	0.030	0.028	0.015	0.018	0.017	0.045

Source: own calculations.

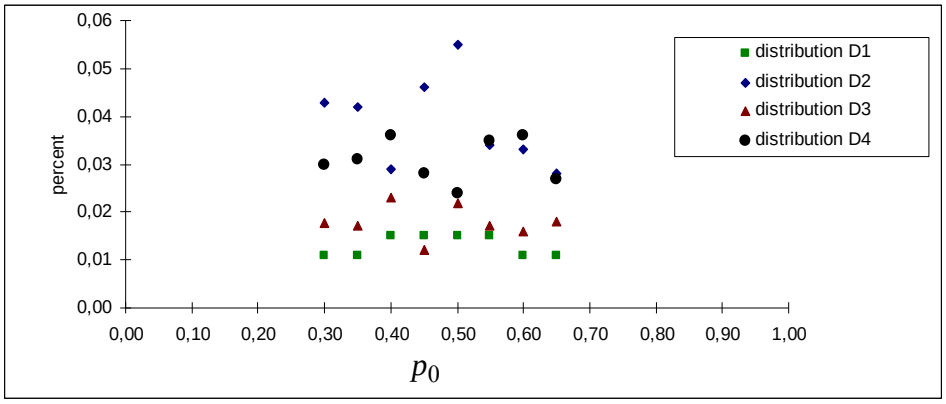

Figure 4.1. Percentage of acceptance decision of false H_0 for independent sampling of n=100 elements

Source: own calculations.

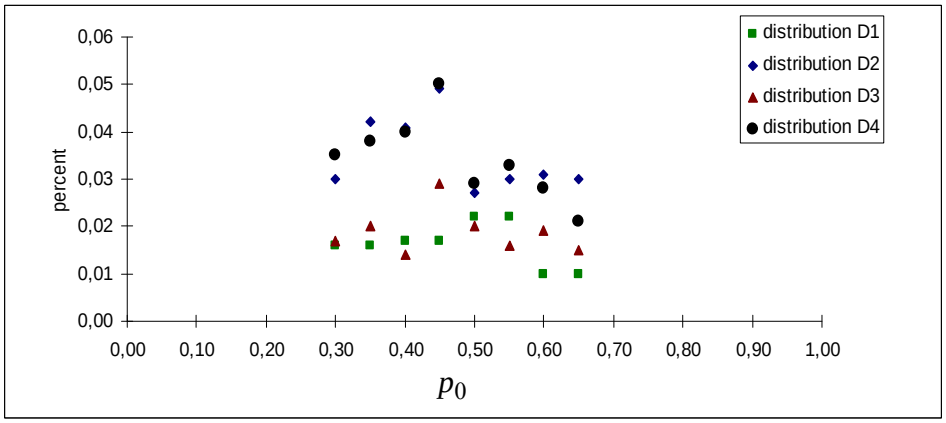

Figure 4.2. Percentage of acceptance decision of false H_1 for independent sampling of n=100 elements

Source: own calculations.

Table 4.2. The percentage of acceptance decision of false H_0 and H_1 for dependent random sampling and fixed prior distribution of parameter p

p_0	n	distribution D1		distribution D2		distribution D3		distribution D4	
		LF_1/R	LF_0/R	LF_1/R	LF_0/R	LF_1/R	LF_0/R	LF_1/R	LF_0/R
0.30	50	0.020	0.025	0.062	0.038	0.022	0.023	0.041	0.053
	100	0.013	0.010	0.035	0.036	0.015	0.016	0.037	0.036
0.35	50	0.020	0.025	0.052	0.049	0.020	0.022	0.045	0.055
	100	0.013	0.010	0.034	0.030	0.030	0.019	0.036	0.028
0.40	50	0.018	0.031	0.037	0.067	0.022	0.027	0.052	0.055
	100	0.023	0.014	0.047	0.035	0.023	0.020	0.054	0.035
0.45	50	0.018	0.031	0.053	0.041	0.024	0.034	0.048	0.053
	100	0.023	0.014	0.031	0.041	0.023	0.013	0.036	0.033
0.50	50	0.026	0.023	0.042	0.046	0.021	0.038	0.047	0.057
	100	0.011	0.016	0.029	0.030	0.017	0.018	0.039	0.036
0.55	50	0.026	0.023	0.067	0.046	0.031	0.022	0.060	0.058
	100	0.011	0.016	0.028	0.031	0.018	0.024	0.036	0.039
0.60	50	0.017	0.023	0.067	0.056	0.023	0.020	0.054	0.065
	100	0.012	0.022	0.035	0.041	0.012	0.024	0.032	0.055
0.65	50	0.017	0.023	0.028	0.061	0.022	0.029	0.031	0.060
	100	0.012	0.022	0.033	0.044	0.017	0.017	0.031	0.050

Source: own calculations.

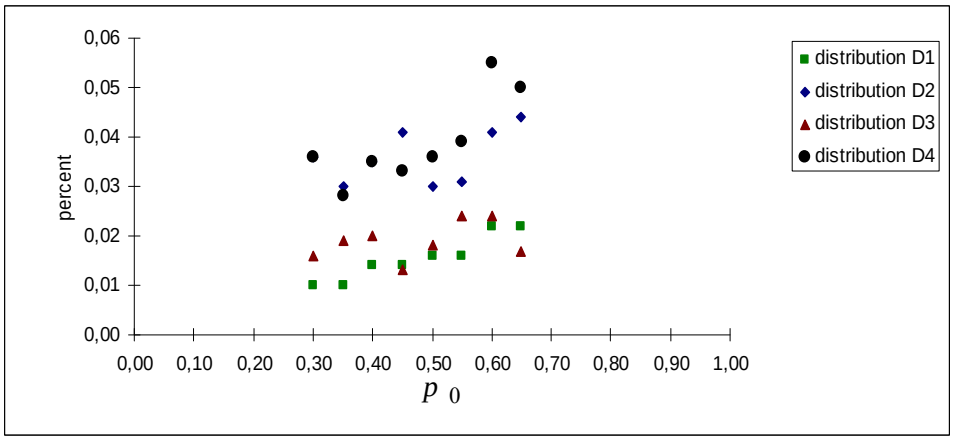

Figure 4.3. Percentage of acceptance decisions of false H_0 for dependent sampling of n=100 elements

Source: own calculations.

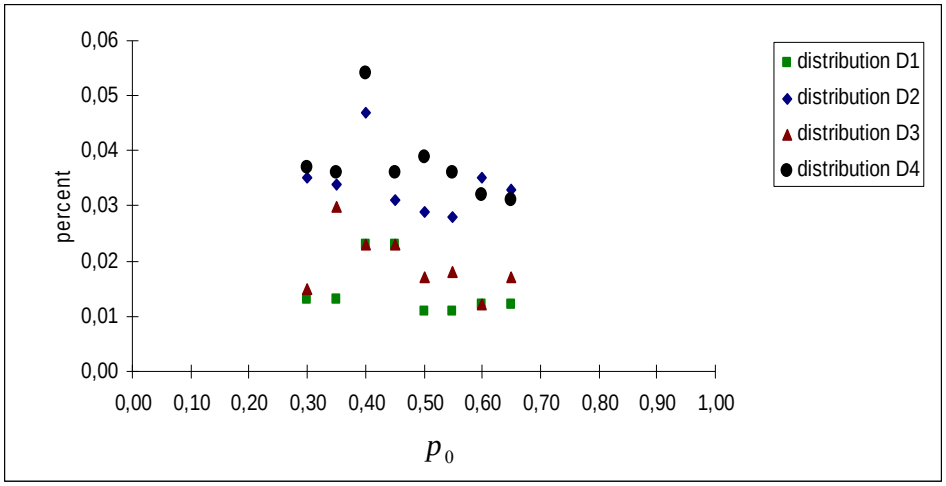

Figure 4.4. Percentage of acceptance decision of false H_1 for dependent sampling of n=100 elements

Source: own calculations.

V. CONCLUSIONS

The Bayesian tests for proportion can be applied to verify hypotheses about parameter p, for independent and dependent sampling. Therefore, they can be applied to both finite and infinite populations.

The frequencies of false decisions of the acceptance of H_0 and H_1 were smaller than 0.05 for the considered prior distributions, D1 and D2, of parameter p and independent sampling. For distributions D3 i D4 and sample sizes smaller than 100 there were few cases of the percentage of false decisions being slighty greater than 0.05.

For the dependent sampling similar results were obtained. For the assumed prior distributions of parameter p , the use of the Bayesian test with dependent sampling allowed to accept one of the hypotheses with probabilities of errors not exceeding 0.05 for sample size $n \geq 100$. For smaller samples the frequencies of the false decisions of the acceptance of false hypothesis were sometimes slighty greater than 0.05.

The prior distributions considered are instances of possible types of distribution of parameter p . The results obtained encourage to further research involving possibly the normal truncated distribution as the prior distribution of parameter p.

REFERENCES

Domański Cz., Pruska K. (2000), *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa.
French S., Rios Insua D. (2000), *Statistical Decision Theory*, Arnold, London.
Krzyśko M. (2004), *Statystyka matematyczna*, t. II, Wydawnictwo Naukowe UAM, Poznań.
Szreder M. (1984), *Informacje a priori w klasycznej i bayesowskiej estymacji modeli regresji*, Wydawnictwo Uniwersytetu Gdańskiego.

Dorota Pekasiewicz

BAYESOWSKIE TESTY STATYSTYCZNE DLA WSKAŹNIKA STRUKTURY DLA NIEZALEŻNEGO I ZALEŻNEGO SCHEMATU LOSOWANIA PRÓBY

W wyniku zastosowania bayesowskich testów statystycznych podejmujemy decyzję o akceptacji hipotezy, dla której ryzyko a posteriori jest mniejsze. Ryzyko a posteriori zależy od rozkładu a priori rozważanego parametru, funkcji straty i schematu losowania próby.

W pracy rozpatrywane są bayesowskie testy statystyczne dla wskaźnika struktury, w przypadku różnych rozkładów a priori, przy niezależnym i zależnym schemacie losowania próby. Oprócz rozważań teoretycznych, zaprezentowane są wyniki analiz symulacyjnych dotyczących własności tych testów.