PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

GRZEGORZ KRZYKOWSKI

BAYESIAN TECHNIQUES IN RANDOMIZED RESPONSE

1 INTRODUCTION

Opinion polls are regularly done for nearly 70 years. The first professional reports from an opinion poll were compiled by the American Institute of Public Opinion founded by George Gallup. There were earlier attempts to recognize views of particular groups of people tied by common profession, domicile or religion. However, they were usually opinions of whole groups, or sometimes representatives of these groups, but it was rare to randomely choose these representatives.

Preparing and realizing a contemporary opinion poll takes few major stages, which are carried out using advanced research techniques [9].

According to this procedure an opinion poll is finished with the moment of compiling a final report and releasing the results. At this stage the results and techniques of research are judged by the recipients of the report and a wide group of people interested in the results. This group is usually quite roughly acquainted with the methods of designing and realizing the research and of the statistical analysis of empirical data. Nevertheless, it is the orderer of the research who has a voice in making a decision about the shape and techniques of research. Often, but not always, they go by economical reasons and override content-related problems. As a consequence from the very beginning of a research project the researchers fight stereotypes that are hard to overpass and concern the nature of arising problems. In particular, the problems concern the interpretation of random issues. In one of the first polls concerning the analysis of voting preferences a sample of 10 million people[1] was taken arguing that accurate prognosis needs a huge set of data. It is only the spectacular mistakes in the results of the poll that allow researchers to argue for changing the research techniques. Another argument for new research methods is the expansion of the information technology, infrastructure and the rise of social awareness. In the times of a different view on the research results very important is the transformation of the informative media market. Society expects the media to provide accurate and reliable information substantiated by thorough research. Additionally, the research has become releatively cheap and this results with a huge group of individual research orderers. Banks, economic and social entities have researches done, the results of which are used on their own hook for substantiating argumentation concerning current economic or

[1] A 1936 poll concerning presidential election which was won by Franklin D. Roosevelt, whereas the sample favored his opponent, Alf Landon

political decisions. At the end of XXth century we observed an unimaginably dynamic growth of computer techniques. This growth is reflected by the change of approach to the analysis of statistical data. The main modification of the hitherto existing way of analyzing data is introducing the Bayesian approach in a big way. This approach not only changed the purely technical side of the form of analysis of the survey data, but the main change is the introduction of a new philosophy of giving answers to posed questions in the domain of the observation carried on.

The aim of the work is presenting and solving the issue of estimating the result in the randomized response technique. The techniques used so far of analysing this issue were based on the classical rules of statistical inferring. These methods were analytically correct and easy to compute, but as it sometimes happens, the results obtained in certain cases were unreasonable[2]. In the presented work we deeply analyse the issue of randomized response using of Bayesian methods. We present various choices of prior distributions and we show why some of them do not give the expected results. We present results and Bayesian procedures with the use of a modern V@R technology [7].

2. RANDOMIZED RESPONSE – RESEARCH EXPERIMENT DESCRIPTION

When constructing a survey research we are faced with a variety of partial problems which require detailed techniques. One of classic issues connected with formulating survey questions is balancing possible answers in the sense of positive and negative answer. This issue lies in the domain of psychological investigation and concerns aversion to or acceptance of a selected answer due to a too suggestive question. After the September 11th attack on World Trade Center it was justified to ask a question like the following one: Do you think that the United States should take a military action in retaliation for the terrorist attacks on New York and Washington? This question would suggest the answer too much and thus the actual question asked by The Gallup Organization was: Do you think that the United States should or should not take a military action in retaliation for the terrorist attacks on New York and Washington? Balance of the answers can also be achieved by proposing several answers with a growing strength of the answer. It is worse when we have to deal with sensitive questions.

The issue of truthfulness of the answers to a given questions appeared on the very beginning of the construction of the surveys. We usually assume, that the respondents answer the survey questions honestly and verdically. On the other hand we are aware, that if the question is socially or religiously or professionally sensitive, the truthfulness of the answers is up in the air [5]. The respondents try to avoid answering a sensitive question and most of them either chooses not to answer at all or chooses the answer which is socially accepted – they simply do not tell the truth. In such cases the result of the survey is erroneous and does not reflect the public opinion about the question

[2] Estimators obtained by the method of moments or plug-in method not always give an unambiguous result and not always fall in the allowable range of parameters (compare eg. [1, Chapter 2.1])

posed. The Randomized Response technique aims at discharging the respondent from publishing his standpoint about given issue. This kind of freeing from remorses was used from way back. The biblical penalty of stoning and later the execution by shooting by a firing squad had similar features. The firing squad chosen to shoot the convict was given rounds of ammunition, some of which were blank. They were distributed randomely and in fact no one knew who was the one who shot the convict.

In the classical approach [10] the research technique of randomized response is answering one of two connected questions randomely chosen by the respondent. Balance plays a significant role in the construction of the questions. Suppose the socially sensitive question (denoted Q_0) is as follows:

Q_0: *Did you ever drive in the state of significant alcohol intoxication?*

We add second, complementary question of opposite nature balancing the research issue (denoting Q_1):

Q_1: *Did you always qive up driving when in the state of significant alcohol intoxication?*

The questions may have quite simple form.

Q_0: *Will you vote for candidate x in the next presidential election?*

Q_1: *Will you not vote for candidate x in the next presidential election?*

Questions Q_0 and Q_1 are constructed in such a way that answering "Yes" to one of them is equivalent to answering "No" to the other. We propose the respondent to toss a die (in private) and answering question Q_0 when he/she gets the outcome of 1, 2, 3, 4 or 5, and answering Q_1 when he/she gets 6, but we expect the question to be answered honestly. The result is an answer of "Yes" or "No", but the researcher does not know which question was answered. As a result we don't know the opinion of the person questioned. The assessment of the respondent's standpoint is influenced by his/her opinion about the presented issue together with the frequency of choosing question Q_0. In the simpliest approach the respondent tosses a die and then gives the answer. This schema implies a computational process after the diagram in the figure (1).

From this, the probabilites of getting a positive and negative answer in a single sample are, accordingly:

$$\Pr(\text{Yes}) = \mu\theta + (1-\mu)(1-\theta) = \frac{2}{3}\theta + \frac{1}{6},\ \Pr(\text{No}) = \frac{5}{6} - \frac{2}{3}\theta,$$

where $1 - \mu$ is the probability of getting „6" and θ is the probability of the support for the candidate.

If the answers come from n respondents, we can try to assess the parameter θ, denoting „real" frequency of a positive answer to the sensitive question posed. Using the method of maximum likelihood (method of moments, plug-in principle) we get estimator for θ of the following form

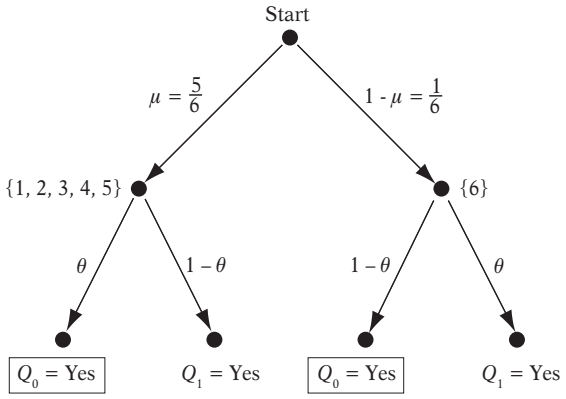


Figure 1. Engel's diagram

$$\hat{\theta}_{ML} = \frac{y_T + \mu - 1}{2\mu - 1} = \frac{3}{2} y_T - \frac{1}{4}, \tag{1}$$

where $y_T = \frac{n_T}{n}$ is the fraction of the answer „Yes" in the n element simple random sample from a binomial distribution with parameters $\left(n, \frac{2}{3}\theta + \frac{1}{6}\right)$.

Right at the first sight we notice that the estimator (1) has significant shortcomings. One of them is inappropriate range. For $n = 20$ respondents and $n_T = 3$ „Yes" answers from (1) we get a negative value of

$$\hat{\theta}_{ML} = -\frac{11}{80}.$$

We get similarly inappropriate (greater than 1) assesment of θ for $n_T = 17$. In such cases we assume values of 0 and 1 to be the estimator of greatest reliabilty, but this is not a favourable solution.

The literature of the subject gives examples of attempts to analyse the survey results using the randomized response technique. These attempts usually enrich the form of posing questions with the cases of not connected questions or systems of questions with multiple answers. The estimation optimalisation in those cases uses classical estimation techniques with a drift towards minimizing the variance [8]. We also find few suggestions of a different approach. One of them concerned using logit transformation [5], [2], another was based on the Bayesian approach [6] and was orientated towards considering a case where in place of the answer „Yes" there were several different answers considered. However, until now there were no studies that would transgress the standards found in monographic studies.

3. BAYESIAN APPROACH – MOTIVATION

In the Bayesian approach we assume that the assesed parameter is a random variable (that is, a random variable of a known distribution). At first we assume that

the distribution of the estimated parameter is known. We call it prior distribution. As a result of observation obtained from a survey we get a modification of prior distribution called posterior distribution. The posterior distribution is obtained by applying Bayes formula to a known conditional distribution of obtained empirical data an the prior distribution. The posterior distribution is the final result of the estimation.

Earlier works the Bayesian estimator was understood as the expected value of the posterior distribution. This value was derived from the conditions of minimum risk with a square loss function. This procedure shows some similarity to classical methods of estimation and justifies the name of this value. Nowadays authors consequently avoid the name „Bayesian estimator" using more often the „prior expected value" or, less accurately, „prior average". For example, such rule is used in monograph [4]. In another monograph [1] the term „Bayesian procedure" is used, and usage of „Bayesian estimator" is retired.

In practice the posterior distribution is presented by density or frequency graphs and by giving quantile characteristics, the average and the standard deviation of the distribution.

In theory the Bayesian approach can be used in almost every model. In practice we use it when we are aiming less at estimating the parameters of the distribution and more on modyfing the hitherto existing knowledge about the parameter. In the presented issue (the randomized response) we usually have very much initial information. The very approach to the question as a sensitive one is introducing an imprecise, but entrenched knowledge about the answers. It evinces in a supposition that the answers will be dishonest, so the classical empirical frequencies will be underrated (or overrated). The common usage of the maximum likelihood estimator according to formula (1) is supported by the prior information. We do assume, that in this approach it is unlikely for the values of the estimator to exceed the range and we treat it as the proper estimate of the frequency of an answer to a sensitive question. However, it turns out that the probability of a negative estimate of the frequency parameter θ may be quite big. We have equalities:

$$\Pr_\theta\left(\hat{\theta}_{ML} < 0\right) = \Pr_\theta\left(n_T < (1-\mu)n\right) = F_\theta\left((1-\mu)n\right)$$

where $F_\theta(\cdot)$ is the cumulative distribution function of a random variable of binomial distribution with parameters (n, θ). As for any $\theta \in (0, 1)$ and $\mu \in \left(\frac{1}{2}, 1\right)$

$$\lim_{n \to \infty} F_\theta\left((1-\mu)n\right) = 1,$$

then $\Pr_\theta\left(\hat{\theta}_{ML} < 0\right)$ may be arbitrarily close to one, so it does not have to be as small as it is commonly taken. Likewise for fixed n and $\mu \in \left(\frac{1}{2}, 1\right)$

$$\lim_{\theta \to 1} \Pr_\theta\left(\hat{\theta}_{ML} > 1\right) = 1 - \lim_{\theta \to 1} \Pr_\theta\left(n_T \leq n\mu\right) = 1,$$

so for every size of the sample and every $\mu \in \left(\frac{1}{2}, 1\right)$ we can select θ, such that $\Pr_\theta\left(\hat{\theta}_{ML} > 1\right)$ is arbitrarily close to one. From this exceeding value of one by the esti-

mator given by the formula (1) is quite a common effect when the value of estimated parameter θ is big.

The Bayesian approach is always connected with some limitations in the choice of parameters and interpretation of the results, but these are limitations we can influence and modify as needed. When using the maximum likelihood estimators we can at most interpret the results.

4. CONSTRUCTION OF A MODEL

Usually a survey of sensitive issues is preceded by partial or preparatory surveys. The results obtained from these preliminary information constitute the initial knowledge which is grouped to form an prior distribution. The issue of constructing an prior distribution is complicated and demands a great experience in analysing empirical data.

The researcher has to be very well acquainted to the content-related side of the analysed phenomenon. In the randomized response technique setting the conditional distribution from which the random sample is taken is quite natural and does not cause too much discrepancy between researchers. It is assumed that it is a binomial distribution with proper parameters, which usually reflects well the nature of the examined phenomenon. It is more difficult to set an prior distribution. We will discuss this issue in the next sections.

Let us assume the simpliest Warner model [10]. In this model we denote by μ a constant value corresponding to the frequency at which we draw question Q_0: We assume that $\mu \in \left(\frac{1}{2}, 1\right)$ which means that as a result of drawing a question by the respondent we obtain Q_0 more often. Then $1 - \mu$ is the probability of obtaining question Q_1. By $\theta \in (0, 1)$ we denote the value of a random variable Θ reflecting the frequency of a positive answer to a sensitive question Q_0. Let X be a random variable corresponding to the answer, which was given by the respondent. The random variable takes values „Yes" and „No" and its conditional distribution (on the condition Θ) has the following form:

$$\begin{aligned} \Pr(X = \text{Yes} \,|\, \Theta = \theta) &= \mu\theta + (1 - \mu)(1 - \theta), \\ \Pr(X = \text{No} \,|\, \Theta = \theta) &= 1 - (\mu\theta + (1 - \mu)(1 - \theta)). \end{aligned} \tag{2}$$

We will use values one and zero interchangeably instead of „Yes" and „No" accordingly, and for simplicity of notation we denote

$$\psi_\mu(\theta) = \mu\theta + (1 - \mu)(1 - \theta). \tag{3}$$

Let $Y_1, Y_2, \ldots, Y_n$ be a random n size simple sample from a distribution given by formula (2). By n_T we denote the number of „Yes" answers in this sample, and by n_N the number of „No" answers. With the above assumptions the combined conditional distribution of the random vector $Y = (Y_1, Y_2, \ldots, Y_n)$ has the following form:

$$\Pr(Y = y \,|\, \Theta = \theta) = \big(\psi_\mu(\theta)\big)^{n_T}\big(1 - \psi_\mu(\theta)\big)^{n_N} \tag{4}$$

where $\psi_\mu(\theta)$ is given by formula (3) and

$$\frac{1}{2} < \mu < 1,\ 0 < \theta < 1,\ y = (y_1,\ y_2,\ \ldots,\ y_n) \in \{Yes,\ No\}^n. \tag{5}$$

The obtained conditional distribution is a binomial distribution with parameters $(n,\ \psi_\mu(\theta))$.

4.1 THE CHOICE OF PRIOR DISTRIBUTION – NONINFORMATIVE PRIORS

In the prior distribution we intend to gather the hitherto existing information about the assessed parameter represented by the random variable Θ. If there are no clues concerning this parameter, we can assume that the prior distribution of variable Θ is one of classical noninformative distributions. For a binomial distribution noninformative distributions are well known [11]. For a randomized response procedure where the parameter of success is given by formula (3), that is as a linear function of an unknown conditioning parameter Θ, there are no such data. Below we bring up noninformative priors for the analysed model deriving from classical assumptions concerning the construction of these distributions and we will present the consequences of these choices. As the first noninformative distribution we assume the uniform distribution on the interval (0, 1), which means the density of variable Θ has the following form:

$$f_\Theta(\theta) = 1,\ \theta \in (0,1).$$

Then we get the posterior distribution

$$f_{(\Theta\,|\,Y)}(\theta \,|\, y) = \psi_\mu^{n_T}(\theta)\big(1 - \psi_\mu(\theta)\big)^{n_N},\ \theta \in (0,1)$$

where notation is given in formula (4), and the range of parameters is described by inequalities (5).

For $\mu = \frac{5}{6}$, $n = 100$ the conditional expected value of parameter Θ is given by the formula

$$\mathbb{E}\big(\Theta \,|\, Y = y\big) = C_\mu \int_0^1 \theta\Big(\frac{2}{3}\theta + \frac{1}{6}\Big)^{n_T}\Big(\frac{5}{6} - \frac{2}{3}\theta\Big)^{n_N} d\theta,$$

where $C_\mu = C_{\frac{5}{6}}$ is a normalisation constant, and n_T and n_N are the number of successes and failures in the sample, accordingly. The graph of this conditional expected value as a function of the number of „Yes" answers (that is as a function of n_T) if presented in figure 2. As it shows for the frequency of successes up to $\frac{1}{2}$ the expected value of posterior distribution is smaller than the empirical result, that is smaller than $\frac{n_T}{n}$, and for the frequency of successes above $\frac{1}{2}$ the expected value of osterior distribution is

above the empirical result. It is worthwile to consider this dependance for a changing parameter μ. For $\mu = 1$ we get a classical result, ie. the expected value of posterior distribution is a linear function of the number of successes (See Section (5.3)) and for $\frac{n_T}{n} < \frac{1}{2}$ is above the empirical result $\frac{n_T}{n}$, and for $\frac{n_T}{n} > \frac{1}{2}$ is below the empirical result (conversely to the figure 2). Together with the decrease of the value of parameter μ we observe the empirical results and the expected value of posterior distribution diverging to the form as in figure 2, which means the expected value of posterior distribution for empirical result $\frac{n_T}{n} < \frac{1}{2}$ falls below this result and for empirical data such that $\frac{n_T}{n} > \frac{1}{2}$ it is greater than the empirical value.

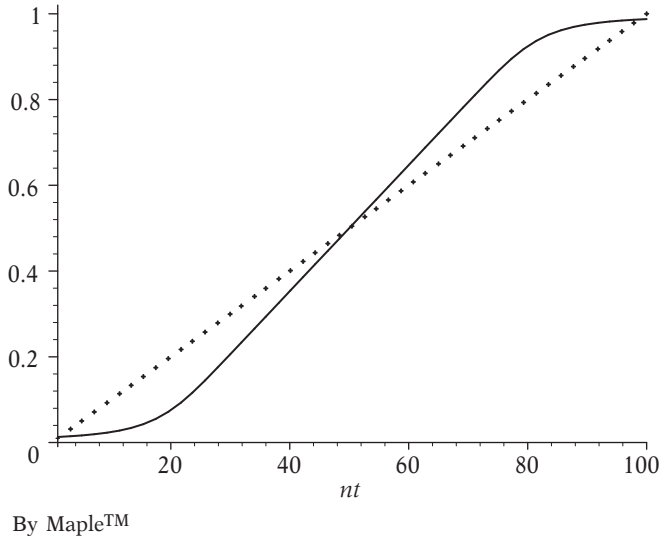


Figure 2. Expected value of the posterior distribution as a function of the number of successes. Dotted line denotes fraction $n_T/100$; solid line denotes the expected value

Another classical noninformative distribution in the described randomized response procedure which we consider is the Jaffreys distribution basing on the Fisher amount of information. In this case the density of the parameter Θ has the form (see Section 5.1)

$$f_\Theta(\theta) = \frac{2\mu - 1}{2\,arcsin(2\mu - 1)} \frac{1}{\sqrt{\psi_\mu(\theta)\big(1 - \psi_\mu(\theta)\big)}},\ \theta \in (0,1),\ \mu \in \Big(\frac{1}{2}, 1\Big),$$

and therefore the density of posterior distribution fulfills the proportion

$$f_{(\Theta|Y)}(\theta \,|\, y) \overset{\theta}{\propto} \psi_\mu(\theta)^{n_T - \frac{1}{2}} \big(1 - \psi_\mu(\theta)\big)^{n_N - \frac{1}{2}}.$$

This distribution has the same feature as the uniform distribution. For the frequency of successes up to $\frac{1}{2}$ the conditional expected value is below the empirical frequency $\frac{n_T}{n}$, and for the frequency of successes above $\frac{1}{2}$, it is above. The graph of conditional expected value as a function of the number of successes is similar to the graph in figure 2.

The third classical noninformative distribution, prior MDIP distribution, also has similar features. We obtain this distribution by maximizing entropy (see section 5.2). In this case the density of prior distribution is given by formula

$$f_\Theta(\theta) = C_\mu \psi_\mu(\theta)^{\psi_\mu(\theta)} \left(1 - \psi_\mu(\theta)\right)^{\left(1 - \psi_\mu(\theta)\right)},\ \theta \in (0,1)$$

where $C_\mu = C_{\frac{5}{6}}$ is the normalization constant, and $\psi_\mu(\theta)$ is given by the formula (4) with limitations to the range of parameters (5).

4.2 CHOOSING BETA DISTRIBUTION AS THE PRIOR DISTRIBUTION

Another group of prior distributions are the distributions that bear some information obtained from hitherto research. In case of a sample taken from a binomial distribution we take as a standard prior distribution Beta distribution with parameters α, β with different way of selecting the α and β parameters. The main reason after choosing Beta distribution is the fact, that Beta distribution is conjugate for binomial distribution [3, §9.4 Tw.2]. In the case of randomized response model Beta distribution is not conjugate (except the case of $\mu = 1$). However the group of Beta distributions is fairly rich and we can expect that taking an prior distribution from this group may be successful used. Let's assume that the „knowledge" about the nature of frequency expressed by random parameter Θ resolves to the fact that we can set the expected value and variance

$$\mathbb{E}\Theta = m_0,\ Var(\Theta) = \sigma_0^2,$$

where m_0 and σ_0^2 are known values. We then assume that the random variable Θ has a Beta distribution with parameters α and β connected to values m_0 and σ_0^2. We can compute the parameters of Beta distribution starting from a system of equations (eg. [4, Sec. A.2])

$$\frac{\alpha}{\alpha + \beta} = m_0,\ \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = \sigma_0^2. \tag{6}$$

This system has the following resolution

$$\alpha = m_0(1 - m_0)\left(\frac{m_0}{\sigma_0^2} - \frac{1}{1 - m_0}\right),\ \beta = (1 - m_0)^2\left(\frac{m_0}{\sigma_0^2} - \frac{1}{1 - m_0}\right). \tag{7}$$

We notice that if some random variable ξ has Beta distribution, and parameters are such that $\alpha \neq 1$ or $\beta \neq 1$ and $m = \mathbb{E}\xi$, $\sigma^2 = \text{Var}(\xi)$, then from inequality $\mathbb{E}\xi > \mathbb{E}\xi^2$ we get the inequality

$$\frac{m}{\sigma^2} - \frac{1}{1-m} = \frac{\mathbb{E}\xi - \mathbb{E}\xi^2}{(1-m)\sigma^2} > 0.$$

Then the parameters α and β in the formula (7) are well described (are non-negative). In the same time we have to notice that the expected value m_0 and variance σ_0^2 in the Beta distribution fulfill the conditions $0 < m_0 < 1$ and $\sigma_0^2 \leq \frac{1}{4}$.

Additionally, a condition resulting from the relationship between variance and expected value has to be fulfilled, which implies the inequality

$$\sigma_0^2 < m_0(1-m_0),\ m_0 \in (0,1). \tag{8}$$

Solving this inequality we have that for a given $\sigma_0^2 \in \left(0, \frac{1}{4}\right)$ value m_0 fulfills the inequality

$$\frac{1}{2} - \sqrt{\frac{1}{4} - \sigma_0^2} < m <_0 \frac{1}{2} + \sqrt{\frac{1}{4} - \sigma_0^2} \tag{9}$$

Let us denote for simplicity

$$m_1 = \frac{1}{2} - \sqrt{\frac{1}{4} - \sigma_0^2},\ m_2 = \frac{1}{2} + \sqrt{\frac{1}{4} - \sigma_0^2} \tag{10}$$

Then $m_0 \in (m_1, m_2)$, $\sigma_0^2 \in \left(0, \frac{1}{4}\right)$.

The Beta distribution taken as the prior distribution has density

$$f_\Theta(\theta) = \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \theta \in (0,1) \tag{11}$$

where α and β are described by the formulas (7) with the condition (8).

The subtle analysis of the choice of α and β parameters in a classical case has been presented in monograph [4, Chapter 5]. The analysis taken out by Gelman and coauthors concerns parameters α and β of Beta distribution, when the conditional distribution of a simple random sample is a binomial distribution. The authors assume as well the conditions resulting from setting prior expected value and variance in the Beta distribution. With the randomized response techniqe, however, we have quite a different model and completely dissimilar expectations about the results of the Bayesian estimation.

The conditional distribution of the random simple sample is given by the formula (4). From this, from the Bayes formula the posterior distribution fulfills the condition

$$f_{(\Theta|Y)}(\theta \mid y) \overset{\theta}{\propto} \psi_\mu(\theta)^{n_T}\left(1-\psi_\mu(\theta)\right)^{n_N}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{12}$$

where $\psi_\mu(\theta)$ is given by formula (3), $\theta \in (0, 1)$ and n_T is equal to the number of „Yes" answers and n_N is equal to the number of „No" answers for the result of the random sample $(y_1, y_2, \ldots, y_n)$. The randomized response research technique assumes in its essence underrating (or overrating) the frequency of classical answers. The respondent of a survey answers „No" (or „Yes") more frequently than it is owed to his attitude. Let us assume that the question is phrased in such a way, that respondents in a classical survey underrate the frequency of positive answers, thus hiding their preferences.

As a consequence, if a classically driven preliminary or partial survey suggests, that the expected value of the answer „Yes" is about $m_0 = 15\%$, then the survey using randomized response technique should give more „Yes" answers than 15%.

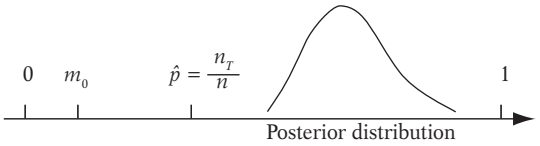


Figure 3. Location of the expected value of the prior distribution (m_0), the empirical value $\hat{p}$ and the posterior distribution

This number of „Yes" answers should imply higher estimate of the frequency parameter θ. The location of these values is schematically presented in figure 3. The schema given in figure 3 does not have to be fulfilled in every survey, but it reflects a natural situation basing on prior premises. This location is usually fulfilled when the success parameter has a value below $\frac{1}{2}$.

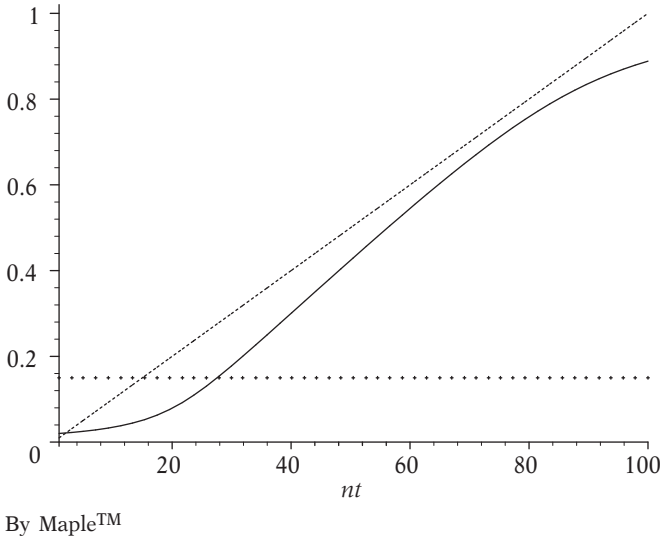


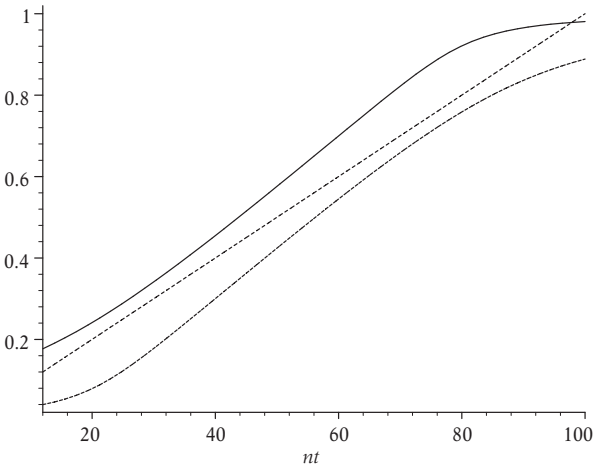
Figure 4. Expected value of posterior distribution as a function of the number of successes. Dotted line denotes fracton $\frac{n_T}{100}$, solid line denotes the expected value, and the crossed line denotes the limit of 15%

The approaches we presented, where the priors is noninformative or is a Beta distribution with parameters α and β derived from assumptions about expected value and variance of Beta distribution given prior, do not thoroughly reflect the nature of the answer to a sensitive question. They do not take into consideration the main characteristics of these answers, that is the underration of the frequency of positive answers to the sensitive question in the classical technique. We solve this problem by using some sort of symetrization by changing the meaning of parameters α and β in the prior assumed Beta distribution.

4.3 COMPUTATIONAL TECHNIQUES AND EXAMPLES

We start illustrating the application of Beta distribution as prior distribution from assuming an prior Beta distribution of expected value $m_0 = 0.15$ and variance $\sigma_0^2 = 0.01$. Then from the formula (7) we get that $\alpha = 1.7625$ and $\beta = 9.9875$. The graph of conditional expected value $\mathbb{E}(\Theta|Y = y)$ as a function of n_T that is the number of "Yes" answers (for $n = 100$ tries) is presented in figure (4).

The solid line, denoting the expected value in posterior distribution, is located below the dotted line, which proves that the conditional expected value for every number of successes $n_T > 2$ underrates the empirical value which comes from the number of obtained successes. The underration is quite substantial. As it can be seen in figure 4 only at $n_T = 27$ to 100 successes in the sample the expected value of posterior distribution is 15%. A similar phonomenon of underration of the frequency of successes by the expected value of posterior distribution is observed at different values of m_0 and σ_0^2. This phenomenon should not occur in the analysed issue, we should rather expect the location of expected value of posterior distribution such as shown in figure 3.



By Maple™

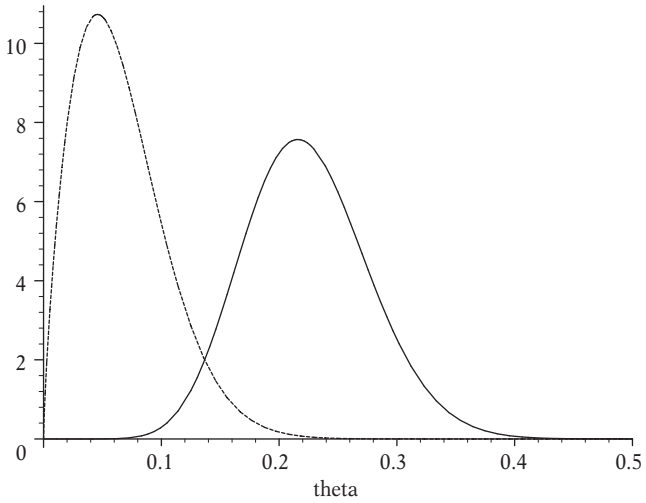
Figure 5. The expected value of posterior distribution as a function of the number of successes. Dotted line denotes the fraction $\frac{n_T}{100}$, solid line below the dotted one denotes the expected value of posterior distribution without modification, and the solid line above the dotted one – with modification.

This underration is caused by a characteristic of Beta distribution signaled in section (5.3), namely the fact that the expected value of posterior distribution is, as a result of assumed prior Beta distribution and binomial likelihood, a linear combination of the initial expected value and the empirical value. This fact is even deepened in case of introducing that parameter of constant frequency μ. The solution to this problem is reversing analogical relationship above the empirical value, in some sense, a symmetrical reflection. We achieve this solution by assuming as the prior distribution the Beta distribution, but with exchanging parameters α and β (compare section (5.3)).

After such symmetrization we get the expected result. The effect of this symmetrization taking place is presented in figure (5). The choice of parameters α and β of the prior distribution is conditioned by the intial information. We may assume, that from earlier analysis basing on classical techniques we have some approximation of the expected value m_0 and variance σ_0^2 of the prior distribution. Then from the formula (7) we find parameters α and β, and then we assume the prior distribution to be $\beta(\beta, \alpha)$. With this choice of parameters the posterior distribution is given by the formula

$$f_{(\Theta|\underline{X})}(\theta \mid \underline{x}) \overset{\theta}{\propto} \psi_{\mu}(\theta)^{n_T}\left(1 - \psi_{\mu}(\theta)\right)^{n_N}\theta^{\beta-1}(1-\theta)^{\alpha-1}, \tag{13}$$

where $\psi_{\mu}(\theta)$ is given by formula (3), $\theta \in (0, 1)$ and n_T is equal to the number of „Yes" answers, and n_N to the number of „No" answers for a random sample $\underline{x} = (x_1, x_2, \ldots, x_n)$. As m_0 and σ_0^2 unambiguously determine the Beta distribution



By MapleTM

Figure 6. Density of posterior distribution before symmetrization (dotted line) and after symmetrization (solid line) for $m_0 = 0.15$, $\sigma_0^2 = 0.01$, $n = 100$, $n_T = 18$.

as the prior distribution, let $\alpha = k\ n\ m_0$ and $\beta = k\ n - \alpha$. Then the new parameter k fulfills the equation

$$k = \frac{1}{n}\left(\frac{m_0(1 - m_0)}{\sigma_0^2} - 1\right). \tag{14}$$

Value k is well chosen (non-negative) if conditions described by inequalities (9) are fulfilled. For example for $n = 100$, $m_0 = 0.15$, $\sigma_0^2 = 0.01$ value of k is 0.1175. Several more values of parameter k are given in table 1.

Table 1

Some values of the proportion parameter k in Beta distribution

μ_0	0.05	0.10	0.15	0.20
$\sigma_0^2 = 0.01$	0.0375	0.0800	0.1175	0.0150
$\sigma_0^2 = 0.04$	0.0019	0.0125	0.0219	0.0300

Eventually we assume as the prior distribution the Beta distribution with parameters $\alpha = k\ n\ (1 - m_0)$ and $\beta = k\ n\ m_0$, where k is given by formula (14).

Let us consider some examples. Let $m_0 = 0.15$ and $\sigma_0^2 = 0.01$ and as a result of simple random sample we obtained $n_T = 18$ successes. Then the posterior distribution has density presented in figure 6. The expected value of the posterior distribution is $\mathbb{E}\ (\Theta|\underline{X}) = 0.2219$, and values of quantiles are given in table 2. This table contains also the quantiles for the greater number of successes $n_T = 20$. Let us note that in this case the estimators of maximum likelihood given by formula (1) for $n = 100$, $n_T = 18$ and $n_T = 20$ are $\Theta_{ML} = 0.02$ and $\Theta_{ML} = 0.05$ accordingly and are of no sense whatsoever in this issue. In the second example we assume $m_0 = 0.10$, $\sigma_0^2 = 0.005$, $n = 200$, $n_T = 30$.

Table 2

Quantiles of posterior distribution ($n = 100$, $m_0 = 0.15$, $\sigma_0^2 = 0.01$)

n_T	$Q_{0.05}$	$Q_{0.10}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.90}$	$Q_{0.95}$
$n_T = 18$	0.1417	0.1579	0.1867	0.2212	0.2580	0.2930	0.3146
$n_T = 20$	0.1558	0.1729	0.2030	0.2389	0.2770	0.3129	0.3351

The graph of the posterior density is presented in figure 7. The expected value

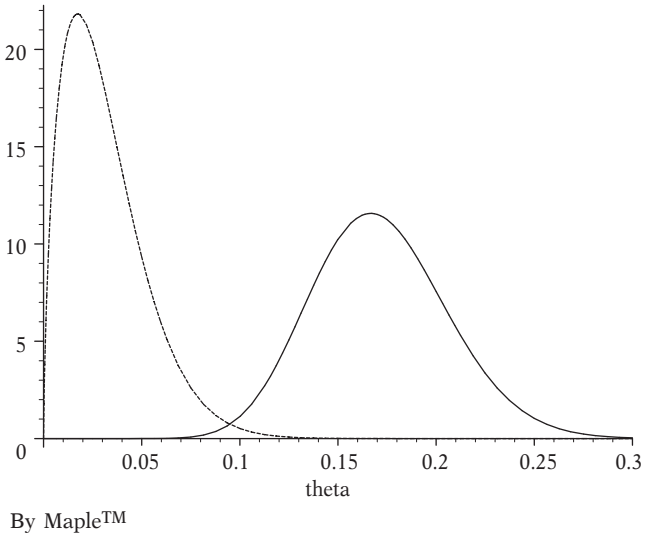


Figure 7. Density of posterior distribution before symmetrization (dotted line) and after symmetrization (solid line) for $m_0 = 0.10$, $\sigma_0^2 = 0.005$, $n = 200$, $n_t = 30$

of posterior distribution is $\mathbb{E}(\Theta|\underline{X}) = 0.1715$, and the values of quantiles are given in table 3. The estimator of greates reliability given by formula (1) for $n = 200$, $n_T = 30$ is negative and we have to assume $\Theta_{ML} = 0$, which does not conform to the conditions of the issue. Lastly, let us note that the Bayesian

Table 3

Quantiles of posterior distribution ($n = 200$, $m_0 = 0.10$, $\sigma_0^2 = 0.005$, $n_T = 30$)

$Q_{0.05}$	$Q_{0.10}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.90}$	$Q_{0.95}$
0.1175	0.1283	0.1474	0.1700	0.1940	0.2169	0.2310

estimatian is as always very sensitive to the size of the sample. For a greater size we have to increase the precision of prior distribution relatively decreasing the value of variance σ_0^2 in the Beta distribution.

5. APPENDIX

5.1 NONINFORMATIVE JEFFREYS' PRIOR

Lemma 1 *Let X be a random variable of a distribution belonging to a family of distributions fulfilling the condition*

$$\Pr(X = x) = \psi_\mu^x(\theta)\big(1 - \psi_\mu(\theta)\big)^{1-x},$$
$$\psi_\mu(\theta) = \theta\mu + (1-\theta)(1-\mu),\ x \in 0,1,\ \theta \in (0,1),\ \mu \in \Big(\frac{1}{2},1\Big).$$

Then the Fisher's amount of information of the parameter θ is given by the formula

$$I(\theta) = \frac{(2\mu-1)^2}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}.$$

Proof. For $x \in \{0,\ 1\}$, $\theta \in (0,\ 1)$, $\mu \in \Big(\frac{1}{2},1\Big)$ we denote

$$f(\theta;x) = \psi_\mu^x(\theta)\big(1-\psi_\mu(\theta)\big)^{1-x}.$$

Then from the definition of the Fisher's amount of information for a family of distributions indexed by parameter θ we have

$$I(\theta) = \mathbb{E}\Big(\frac{\partial \ln f}{\partial \theta}\Big)^2 = \mathbb{E}\Big[\frac{\partial}{\partial \theta}\big(X\ln\psi_\mu(\theta) - (1-X)\ln\big(1-\psi_\mu(\theta)\big)\big)\Big]^2$$
$$= E\Big[X\frac{\psi_\mu'(\theta)}{\psi_\mu(\theta)} + \frac{1-X}{1-\psi_\mu(\theta)}\psi_\mu'(\theta)\Big]^2 = \big[\psi_\mu'(\theta)\big]^2\frac{\mathbb{E}\big[X-\psi_\mu(\theta)\big]^2}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}$$
$$= \frac{2\mu-1}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}Var(X) = \frac{(2\mu-1)^2}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}.$$

The noninformative Jeffreys prior distribution is given by the relationship

$$f_\Theta(\theta) \overset{\theta}{\propto} \frac{1}{\sqrt{I(\theta)}},\ \theta \in (0,1).$$

As for $\mu \in \Big(\frac{1}{2},1\Big)$

$$\int_0^1 \frac{2\mu-1}{\sqrt{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}}d\theta = \frac{2\,arcsin(2\mu-1)}{2\mu-1}$$

then

$$f_\Theta(\theta) = \frac{2\mu-1}{2\,arcsin(2\mu-1)}\frac{1}{\sqrt{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}},\ \theta \in (0,1)$$

is a noninformative prior Jeffreys' distribution.

For $\mu = 1$ we get the noninformative Jeffreys' distribution of a binomial distribution with parameter θ [11].

5.2 NONINFORMATIVE MDPI PRIOR

Let X be a random variable taking a finite number of values m with positive probabilities $p_1, p_2, \ldots, p_m$. Then by the entropy of the distribution $p_1, p_2, \ldots, p_m$ we mean the value

$$H = \sum_{i=1}^{m} p_i \ln p_i.$$

It is easy to show that the entropy is in this case a value limited by $\ln m$ and the maximum value is reached for a uniform distribution on the set $1, 2, \ldots, m$. We define the prior MDIP (Maximum Data Information Prior) distribution by the releationship

$$f_\Theta(\theta) = C_\mu \exp\{H(\theta)\}.$$

If a random variable X has a distribution given by the formula

$$\Pr(X = x) = \psi_\mu^x(\theta)\big(1 - \psi_\mu(\theta)\big)^{1-x},\ x \in (0,1),$$

then the entropy of the distribution of the random variable X fulfills the condition

$$H(\theta) = -\big(\psi_\mu(\theta)\ln\psi_\mu(\theta) + \big(1 - \psi_\mu(\theta)\big)\ln\big(1 - \psi_\mu(\theta)\big)\big).$$

From this

$$f_\Theta(\theta) = C_\mu \psi_\mu(\theta)^{\psi_\mu(\theta)}\big(1 - \psi_\mu(\theta)\big)^{(1-\psi_\mu(\theta))},\ \theta \in (0,1)$$

is the prior MDIP distribution. Because entropy is limited, this distribution is a proper noninformative distribution. The constant C_μ can be computed for a given value of μ. In table 4 we computed come values of the normalization constant

Table 4

Normalization constant in the MDIP distribution

μ	$\mu = 1$ (binomial distribution)	$\mu - \frac{5}{6}$	$\mu - \frac{4}{5}$	$\mu - \frac{3}{4}$	$\mu - \frac{2}{3}$
C_μ	1.6184	1.8454	1.8759	1.9184	1.9626

5.3 BETA PRIOR

Lemma 2 *Let the conditional distribution of a random simple sample* $\underline{X} = (X_1, \ldots, X_n)$ *on the condition of parameter* Θ *be given by the formula*

$$\Pr(\underline{X} = \underline{x} \,|\, \Theta = \theta) = \theta^{\left(\sum_{i=1}^{n} x_i\right)}(1 - \theta)^{\left(n - \sum_{i=1}^{n} x_i\right)}$$

where $\theta \in (0, 1)$, $\underline{x} \in \{0, 1\}^n$ *(binomial distribution). Let the prior distribution of parameter* Θ *be Beta distribution with parameters* (α, β). *Then the expected value of posterior distribution is given by the formula*

$$\mathbb{E}\big(\Theta \,|\, \underline{X} = \underline{x}\big) = tm + (1 - t)\hat{p},$$

where $m = \mathbb{E}\,\Theta$, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $t = \frac{\alpha + \beta}{n + \alpha + \beta}$.

Proof. As parameter Θ has a distribution of $\beta(\alpha, \beta)$, then the prior density is given by formula

$$f_\Theta(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1},\ \theta \in (0,1).$$

From this, the density of posterior distribution fulfills the proportionality condition

$$f_{(\Theta \,|\, \underline{X})}(\theta \,|\, \underline{x}) \overset{\theta}{\propto} \theta^{x_T + \alpha - 1}(1 - \theta)^{x_N + \beta - 1},\ \theta \in (0,1),\ \underline{x} \in \{0,1\}^n,$$

where $x_T = \sum_{i=1}^{n} x_i$, $x_N = n - x_T$. Then the posterior distribution is $\beta(\alpha + n_T, \beta + n_N)$.

From this its expected value may be denoted as follows

$$\mathbb{E}\big(\Theta \,|\, \underline{X} = \underline{x}\big) = \frac{\alpha + n_T}{\alpha + \beta + n} = \frac{\alpha}{\alpha + \beta}\left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) + \frac{n_T}{n}\left(\frac{n}{n + \alpha + \beta}\right),$$

which is the equality from the problem's thesis. Let us note, that if we set $\frac{\alpha}{\alpha + \beta} = m$, then

$$\lim_{\alpha + \beta \to \infty} \mathbb{E}\big(\Theta \,|\, \underline{X} = \underline{x}\big) = m,\ \lim_{n \to \infty}\Big(\mathbb{E}\big(\Theta \,|\, \underline{X} = \underline{x}\big) - \hat{p}\Big) = 0.$$

From this, when choosing parameters α and β of the prior distribution, we decide about the location of the expected value in such a way, that the greater $\alpha + \beta$ the smaller is the influence of the direct data in the posterior distribution. Howsoever, with the increase of the size of the sample, the influence of the prior expected value on the posterior expected value is smaller.

Lemma 3 *If a random variable* ξ *has a distribution* $\beta(\alpha, \beta)$, *then the random variable* $1 - \xi$ *has distribution* $\beta(\beta, \alpha)$:

Proof. Let us denote $\eta = 1 - \xi$. Then

$$f_\eta(\theta) = f_\xi(1 - \theta),\ \theta \in (0,1).$$

From this

$$f_\eta(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}(1-\theta)^{\alpha-1}(\theta)^{\beta-1},\ \theta \in (0,1),$$

which gives the lemma's thesis.

Instytut Informatyki Uniwersytetu Gdańskiego

REFERENCES

[1] Bickel P.J., Doksum K.A., [2001], *Mathematical Statistics. Basic Ideas and Selected Topics. Vol. I*, Prentice Hall, Upper Saddle River, New Jersey, 07458, ii edition.
[2] Corstange D., [Sep. 2-5 2004], Sensitive Questions, Truthful Responses? Randomized Response and Hidden Logit as a Procedure to Estimate It. Annual Meeting of the American Political Science Association.
[3] DeGroot M.H., [1981], *Optymalne decyzje statystyczne*, Państwowe Wydawnictwo Naukowe, Warszawa.
[4] Gelman A., Carlin J.B., Stern H.S., Rubin D.B., **[2000],** *Bayesyan Data Analysis*, CHAPMAN & HALL=CRC, Boca Raton, Lndyn, New York, Washington D.C., II edition.
[5] van der Heijden P., van Gils G., Bouts J., Hox J., [2000], A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning, Sociological Methods and Research, 28(4):505-537.
[6] Kim J.-M., Heo T.-Y., [2005], A Bayesian Analysis of the Multinomial Randomized Response Model using Dirichlet Prior Distribution. Proceedings for the Spring Conference 2005 Korean Statistical Society, pages 239-244.
[7] RMG (RiskMetrics Group), [1999], *Risk Management: A Practical Guide*, Risk-Metrics Group, First edition.
[8] Sing S., Horn S., Singh R., Mangat N.S., [2003], *On the use of modified randomization device for estimating the prevalence of a sensitive attribute*, Statistics in Transition, 6(4):515-522.
[9] Szreder M., [2002], *Badania opinii*, Wydawnictwo Wyższej Szkoły Zarządzania, Gdańsk.
[10] Warner S., **[1965],** *Randomized response: a survey technique for eliminating evasive answer bias*, „Journal of American Statistical Association", 60:63-69.
[11] Yang R., Berger R., [1997], *A Catalog of Noninformative Priors*, Plann. Infer., 79:223-235.

Praca wpłynęła do redakcji w kwietniu 2009 r.

TECHNIKI BAYESOWSKIE W PROCEDURACH LOSOWANIA ODPOWIEDZI

Streszczenie

Zagadnienie prawdziwości odpowiedzi na drażliwe społecznie lub osobiście pytania ankieterów pojawiło się wraz z początkiem badań ankietowych. Często respondenci próbują uniknąć odpowiedzi lub odpowiadają niezgodnie z prawdą. Powstało wiele technik badawczych i analitycznych korygujących wyniki badań ankietowych pozwalających na ocenę rzeczywistej opinii respondentów. W pracy przedstawiono bayesowską analizę techniki Randomize Response. Zaprezentowano argumentację wyboru różnych rozkładów *a priori* zarówno nieinformujących, jak i rozkładów kompensujących informacje wstępne opisane rozkładami parametrycznymi.

Słowa kluczowe: Techniki bayesowskie, randomizowane odpowiedzi, model Warnera

BAYESIAN TECHNIQUES IN RANDOMIZED RESPONSE

Summary

The issue of truthfulness of the answers given to socially or personally sensitive questions appeared at the very beginning of the construction of the surveys. The respondents often try to avoid answering a sensitive question or they do not tell the truth. Various research and analysis techniques have been developed to adjust the surveys results in order to assess the real opinion of the respondents. The work introduces a bayesian analysis of the Randomized Response technique. It presents arguments for choosing various a priori distributions, both non-informative and compensating for the preliminary information described by parametrized distributions.

Key words: Bayesian techniques, randomized response, Warner model