ГОДИШНИК НА СОФИЙСКИ УНИВЕРСИТЕТ "СВ. КЛИМЕНТ ОХРИДСКИ"

ФИЛОСОФСКИ ФАКУЛТЕТ

Книга Социология

Том 99, 2006

ANNUAIRE DE L'UNIVERSITE DE SOFIA "ST. KLIMENT OCHRIDSKI"

FAKULTE DE PHILOSOPHIE

Livre Sociologie

Tome 99, 2006

BAYESIAN INFERENCE OF RELATIVE FREQUENCY (IN THE CASE OF ELECTORAL SURVEYS)

KALOYAN HARALAMPIEV

Kaloyan Haralampiev. BAYESIAN INFERENCE OF RELATIVE FREQUENCY (IN THE CASE OF ELECTORAL RESEARCHES).

This article has two major aims. The first one is to introduce Bayesian paradigm in statistics to scientists and researchers in the field of social sciences. The second one is to show how to solve a particular problem in this field. The problem is how to estimate relative frequency (percentage). In the article is shown that if there is no prior information then the Bayesian approach gives the same results as the well known frequentist approach. In all other cases the Bayesian approach provides an opportunity of taking into account whatever prior information we have. On the other hand, if the prior information consists of data from previous surveys then two different possibilities may exist. First, if there are no significant changes in the relative frequencies over time then we can put all data together and we can use them all for estimation of the relative frequencies. Second, if there are significant changes in the relative frequencies over time then the more harmless way is to use only data from the current research with equal prior probabilities. Also, in the article is shown how Bayesian approach deals with two or more parameters which are not independent. In the electoral surveys this allows us to estimate relative frequencies only among the actual voters.

Introduction

This article has two major aims. The first one is to introduce Bayesian paradigm in statistics to scientists and researchers in the field of social science. The second one is to show how to solve a particular problem in this field. The problem is how to estimate relative frequency (percentage). The solution of this problem is demonstrated in two cases – the general one and the special one – concerning the relative frequency of people who choose certain party only among the actual voters.

1. The Bayesian approach

The central position in the Bayesian paradigm in statistics is occupied by Bayes' theorem:

$$(1) \qquad P(H_k \mid DI) = \frac{P(D \mid H_k I).P(H_k \mid I)}{P(D \mid I)}$$

where H_k "stands for some hypothesis whose truth we want to judge, *D* for a set of data, and *I* for whatever "prior information" we have in addition to the data" (Jaynes 1988: 25).

$P(D \mid H_k I)$ is called *sampling distribution* and it represent the probability to get exactly the data we have if the hypothesis H_k is true.

$P(H_k \mid I)$ is called *prior probability* of H_k and "it specifies expert[1] knowledge of H_k before an experiment designed to provided data *D* is performed" (Dose 2002: 1).

$P(D \mid I)$ is called *marginal probability* and it represent the probability to get exactly the data we have, if some of all hypotheses H_i is true, although we don't know which. The marginal probability is received by *marginalization*:

$$(2) \qquad P(D \mid I) = \sum_i P(D \mid H_i I) P(H_i \mid I)$$

When there are an infinite number of hypotheses and they form continuum, we have to replace the sum with the definite integral as following:

$$(3) \qquad P(D \mid I) = \int_R P(D \mid H_i I) P(H_i \mid I) dH_i$$

where *R* is the region in which all hypotheses H_i are defined.

[1] Or theoretical.

The result obtained by Equation (1) - $P(H_k \mid DI)$ - is called *posterior probability* of H_k.

As we can see from Equation (1), in the posterior probability two kinds of information are combined: the theoretical one represented by the prior probability and the empirical one represented by the sampling distribution.

The posterior probability is a basic tool for statistical inference. Using it we can test hypotheses and make confidence intervals.

2. The Bayesian inference of relative frequency

We examine categorical variable with *m* possible categories. In particular, this could be a question with *m* possible preliminary defined answers. We draw a sample from the studied population. As a result we obtain the sample frequency (number) of each category. Let us denote this number by f_i. Therefore data set consist of observed frequencies:

$$D = \{f_1; f_2; ...; f_m\}$$

The relative frequency of each category in the population is denoted by π_i. Then the hypothesis to be verified is that particular relative frequency π_i will receive exactly the value π_{ik}:

$$H_k = \{(\pi_1 = \pi_{1k})(\pi_2 = \pi_{2k})...(\pi_m = \pi_{mk})\}$$

Since all relative frequencies are non-negative and their sum is unity (or 100%) π_{ik} must satisfy the following conditions:

$$(4) \qquad \begin{cases} \sum_{i=1}^{m} \pi_{ik} = 1 \\ \pi_{ik} \geq 0, \; i = 1,2,...,m \end{cases}$$

When the sample is representative and the population is large enough[2] the relative frequencies are continuous and the sampling distribution is *multinomial* (Jaynes 1993: 315, 317-318):

$$(5) \qquad P(D \mid H_k I) = \frac{n!}{f_1! f_2! ... f_m!} \pi_{1k}^{f_1} \pi_{2k}^{f_2} ... \pi_{mk}^{f_m}$$

[2] Usually both conditions are satisfied in social surveys.

where n is the sample size and $f_i!$ is called f_i factorial which means $f_i!=1.2.3...f_i$.

Now we need the prior probabilities for the calculation of the posterior probability.

As Jaynes remarks: "In 'real live' we usually have excellent grounds based on previous experience and theoretical analysis" (Jaynes 1976: 190). I would like to correct this statement by changing only one word: in 'real live' we usually have excellent grounds based on previous experience *or* theoretical analysis. However, thus initial problem is divided into two new problems. The first one is how to obtain prior probabilities with respect to previous experience. The second one is how to obtain prior probabilities with respect to theoretical analysis.

The solution of the second problem is a typical case of application of the *Method of Maximum Entropy*. As a result of theoretical analysis, we obtain some constrains of the prior probability distribution. Then we must maximize Shannon's entropy $\left(-\sum_k P(H_k \mid I)\log P(H_k \mid I) \text{ or } -\int_R P(H_k \mid I)\log P(H_k \mid I)\right)$ with respect to these constrains. Thus we obtain prior probability distribution which is "as uninformative as possible to prevent us from "seeing" things in the data which are not there" (Bretthorst 1988: 14).

The solution of the first problem is much easier. "One simply uses the posterior probability derived in analyzing the previous measurement as the prior probability for the current measurement" (Bretthorst 1990: 11).

However, let us turn back to the very beginning when neither previous experience nor results of theoretical analysis are available. According to the Method of Maximum Entropy if we have no prior information we must assign equal prior probability (Bretthorst 1990: 4-5):

$$P(H_k \mid I) = const = C$$

Thus we already have specified both sampling distribution (Equation (5)) and prior probabilities and we can calculate the marginal probability:

$$P(D \mid I) = \int_{R_1}\int\ldots\int P(D \mid H_k I)P(H_k \mid I)d\pi_{1k}d\pi_{2k}\ldots d\pi_{mk} =$$

(6)

$$= \int_{R_1}\int\ldots\int \frac{n!}{f_1!f_2!\ldots f_m!}\pi_{1k}^{f_1}\pi_{2k}^{f_2}\ldots\pi_{mk}^{f_m}Cd\pi_{1k}d\pi_{2k}\ldots d\pi_{mk} = \ldots = \frac{n!C}{(n+m-1)!}$$

$$R_1 : \begin{cases} \sum_{i=1}^{m} \pi_{ik} = 1 \\ \pi_{ik} \geq 0,\ i = 1,2,...,m \end{cases}$$

Therefore the posterior probability is:

$$(7) \qquad P(H_k \mid DI) = \frac{\frac{n!}{f_1! f_2! ... f_m!} \pi_{1k}^{f_1} \pi_{2k}^{f_2} ... \pi_{mk}^{f_m} C}{\frac{n! C}{(n+m-1)!}} = ... = \frac{(n+m-1)!}{f_1! f_2! ... f_m!} \pi_{1k}^{f_1} \pi_{2k}^{f_2} ... \pi_{mk}^{f_m}$$

Let me note that when there is no prior information, the prior probabilities are canceled out both in the nominator and in the denominator. Thus only data remain that do matter.

Equation (7) concerns all relative frequencies. However, when we have to estimate exactly the relative frequency π_i the other relative frequencies π_j $(j \neq i)$ are *nuisance parameters*, "*i.e.*, parameters which are physically present in the phenomenon and so cannot be safety disregarded in the model, although we are not interested in estimating them" (Jaynes 1993: 2101). However, "in Bayesian methods, nuisance parameters cause very little trouble – any uninteresting parameters are removed by integrating out with respect to their prior probabilities" (Jaynes 1993: 2101). This procedure is also a marginalization:

$$(8) \qquad \begin{aligned} & P(\pi_i = x \mid DI) = \int_{R_2} \int ... \int P(H_k \mid DI) d\pi_{1k} d\pi_{2k} ... d\pi_{i-1,k} d\pi_{i+1,k} ... d\pi_{mk} = \\ & = ... = \frac{(n+m-1)!}{f_i!(n-f_i+m-2)!} x^{f_i} (1-x)^{n-f_i+m-2} \end{aligned}$$

$$R_2 : \begin{cases} \pi_{1k} + \pi_{2k} + ... + \pi_{i-1,k} + \pi_{i+1,k} + ... + \pi_{mk} = 1 - x \\ \pi_{jk} \geq 0,\ j = 1,2,...,i-1,i+1,...,m \end{cases}$$

The obtained result[3] is called *Beta distribution*. Using it we can calculate the following probability:

- The probability that the relative frequency π_i is less than certain number *a*:

$$(9) \qquad P(\pi_i < a \mid DI) = \int_0^a P(\pi_i = x \mid DI) = F(a)$$

$F(a)$ is called *cumulative posterior probability density function*.

We can also calculate the probabilities:

[3] This result is identical with Jaynes' Equation (17-5) (Jaynes 1974: 17-4).

- The probability that the relative frequency π_i is greater than certain number b:

$$(10) \qquad P(\pi_i > b \mid DI) = \int_b^1 P(\pi_i = x \mid DI) = 1 - F(b)$$

- The probability that the relative frequency π_i is lying between two certain numbers a and b:

$$(11) \qquad P(a < \pi_i < b \mid DI) = \int_a^b P(\pi_i = x \mid DI) = F(b) - F(a)$$

Equations (9) and (10) are applied for hypotheses testing. Equation (11) is applied for making confidence intervals.

So, let us apply this method for solving a real problem.

3. Electoral surveys

3.1. An example of Bayesian method of inference

In table 1 data from one electoral survey is presented:

Table 1

Vote intentions, June 2005

Vote intentions	Relative frequency in the sample (%)	Frequency (number)
Coalition for Bulgaria	27,5	277
National Movement "Simeon II"	14,6	147
Coalition Union of Democratic Forces – Democratic Party – Gergiovden	7,4	75
Movement for Rights and Freedom	5,2	52
Coalition Bulgarian National Alliance	3,6	36
Party "Democrats for A Strong Bulgaria"	3,4	34
Other	7,0	71
I have not decided yet	10,0	101
I would not vote	21,3	214
Total	100,0	1007

Source: http://www.aresearch.org/doc.php?en=1&arch=1&id=581

Applying Equation (8) to the data in Table 1 and then Equation (11) with $F(a) = 0{,}025$ and $F(b) = 0{,}975$ we could receive the following results[4]:

[4] All integrals are calculated numerically.

Table 2

Confidence intervals

Vote intentions	Confidence intervals
Coalition for Bulgaria	$P(0{,}247 < \pi_1 < 0{,}301 \mid DI) = 0{,}95$
National Movement "Simeon II"	$P(0{,}125 < \pi_2 < 0{,}168 \mid DI) = 0{,}95$
Coalition Union of Democratic Forces – Democratic Party – Gergiovden	$P(0{,}060 < \pi_3 < 0{,}091 \mid DI) = 0{,}95$
Movement for Rights and Freedom	$P(0{,}039 < \pi_4 < 0{,}066 \mid DI) = 0{,}95$
Coalition Bulgarian National Alliance	$P(0{,}026 < \pi_5 < 0{,}048 \mid DI) = 0{,}95$
Party "Democrats for A Strong Bulgaria"	$P(0{,}025 < \pi_6 < 0{,}046 \mid DI) = 0{,}95$
Other	$P(0{,}056 < \pi_7 < 0{,}087 \mid DI) = 0{,}95$
I have not decided yet	$P(0{,}083 < \pi_8 < 0{,}119 \mid DI) = 0{,}95$
I would not vote	$P(0{,}187 < \pi_9 < 0{,}237 \mid DI) = 0{,}95$

These results are trivial. It is due to the fact that there exists another paradigm in statistics which is best known and widely used. This paradigm is called either frequentist or orthodox. In the case of estimation of relative frequencies with no prior information the results obtained by both paradigms are practically the same.

3.2. Specific problem – relative frequency only among the actual voters

Unfortunately, there is little practical use of the results in Table 2, because in the electoral surveys the most important relative frequencies are the relative frequencies among the actual voters.

There arise two new problems. The first one is how to proceed with the number of people who "have not decided yet" (how or for whom to vote). The second one is how to proceed with the number of people who would not vote.

Since we don't know for which party this person – who had not decided yet – will vote, this information is equivalent to the information we have about other people in the entire population who do not belong to the sample. This means that we may ignore information about the number of people who have not decided yet and thus we reduce the sample[5].

However, the information for the number of people who "would not vote" is crucial. In order to understand its importance, let us denote the size of entire population by N, the frequencies in the population by f_i^* and the number of people in the

[5] This is current practice in poll-survey agencies in Bulgaria. When the results about the actual voters are presented, the number of people who not decided yet is ignored. In this case there is no contradiction between the theory and the current practice.

population who would not vote by f_m^*. Thus the number of actual voters is $N - f_m^*$ and the relative frequency of people who choose certain party only among the actual voters is:

$$\frac{f_i^*}{N - f_m^*} = \frac{N\pi_i}{N - N\pi_m} = \frac{\pi_i}{1 - \pi_m} \tag{12}$$

since

$$\pi_i = \frac{f_i^*}{N} \tag{13}$$

As we can see, the relative frequency of people who would not vote takes part in the denominator of Equation (12). That means that both the numerator and the denominator are unknown parameters[6]. However, the main problem in estimating the relative frequency – among the actual voters only – of people who choose certain party is that π_i and π_m are not independent[7].

The first step solving the aforementioned problem is calculation of joint marginal probability of π_i and π_m. Again, the marginalization is the way to do this:

$$\begin{aligned} &P[(\pi_i = x)(\pi_m = y) \mid DI] = \int\int_{R_3} \ldots \int P(H_k \mid DI) d\pi_{1k} d\pi_{2k} \ldots d\pi_{i-1,k} d\pi_{i+1,k} \ldots d\pi_{m-1,k} = \\ &= \ldots = \frac{(n+m-1)!}{f_i! f_m! (n - f_i - f_m + m - 3)!} x^{f_i} y^{f_m} (1 - x - y)^{n - f_i - f_m + m - 3} \end{aligned} \tag{14}$$

$$R_3 : \begin{vmatrix} \pi_{1k} + \pi_{2k} + \ldots + \pi_{i-1,k} + \pi_{i+1,k} + \ldots + \pi_{m-1,k} = 1 - x - y \\ \pi_{jk} \geq 0,\ j = 1,2,\ldots,i-1,i+1,\ldots,m-1 \end{vmatrix}$$

Then, the easier way to receive the posterior probability is – firstly, to calculate cumulative posterior probability density function $F(u) = P\left(\frac{x}{1-y} < u \middle| DI \right)$ and then – to calculate *posterior probability density function* $P\left(\frac{x}{1-y} = u \middle| DI \right) = f(u) = F'(u)$:

[6] The current practice in poll-survey agencies in Bulgaria is to ignore the number of people who would not vote. Thus we reduce the sample again. However, this is equivalent to considering the denominator as a known parameter and only the numerator remains unknown. In this case there is a contradiction between the theory and the current practice, although the frequentist theory says nothing about this problem.

[7] This is a consequence of Equation (4).

$$F(u) = P\left(\frac{x}{1-y} < u \middle| DI\right) = \int_0^1 \left\{ \int_0^{u(1-y)} P[(\pi_i = x)(\pi_m = y) \mid DI]dx \right\} dy = \ldots =$$

(15)

$$= \frac{(n - f_m + m - 2)!}{f_i!(n - f_i - f_m + m - 3)!} \sum_{j=0}^{n-f_i-f_m+m-3} \frac{C^j_{n-f_i-f_m+m-3}(-1)^{n-f_i-f_m+m-3-j} u^{n-f_m+m-2-j}}{n - f_m + m - 2 - j}$$

$$P\left(\frac{x}{1-y} = u \middle| DI\right) = f(u) = F'(u) =$$

(16)

$$= \ldots = \frac{(n - f_m + m - 2)!}{f_i!(n - f_i - f_m + m - 3)!} u^{f_i}(1-u)^{n-f_i-f_m+m-3}$$

The obtained result is just another Beta distribution. Using it we can test the hypothesis that certain party will pass 4% limit[8]. Also, we can make the confidence interval of relative frequency of people who choose certain party[9] only among the actual voters. Such results are presented in Table 3:

Table 3

Probability of passing 4% limit and confidence intervals for relative frequency of people who choose certain party among the actual voters

Vote intentions	Probability of passing 4% limit (%)	Confidence intervals
Coalition for Bulgaria	100,0	$P\left(0{,}362 < \frac{\pi_1}{1-\pi_m} < 0{,}434 \middle| DI\right) = 0{,}95$
National Movement "Simeon II"	100,0	$P\left(0{,}183 < \frac{\pi_2}{1-\pi_m} < 0{,}242 \middle| DI\right) = 0{,}95$
Coalition Union of Democratic Forces – Democratic Party – Gergiovden	100,0	$P\left(0{,}087 < \frac{\pi_3}{1-\pi_m} < 0{,}132 \middle| DI\right) = 0{,}95$
Movement for Rights and Freedom	100,0	$P\left(0{,}058 < \frac{\pi_4}{1-\pi_m} < 0{,}096 \middle| DI\right) = 0{,}95$
Coalition Bulgarian National Alliance	95,4	$P\left(0{,}038 < \frac{\pi_5}{1-\pi_m} < 0{,}070 \middle| DI\right) = 0{,}95$
Party "Democrats for A Strong Bulgaria"	90,8	$P\left(0{,}036 < \frac{\pi_6}{1-\pi_m} < 0{,}067 \middle| DI\right) = 0{,}95$
Other		$P\left(0{,}082 < \frac{\pi_7}{1-\pi_m} < 0{,}126 \middle| DI\right) = 0{,}95$

[8] Or the hypothesis that certain candidate will win the presidential election.
[9] Or certain candidate for president.

These results are non-trivial. They are not entirely impossible from the frequentist point of view. However, it is quite difficult to obtain them by the frequentist methods.

4. Two important additional points

4.1. Essential one

In section 2 we examined a situation of having no prior information. Let us now examine the situation of having prior information obtained from previous surveys. Then we can use the data from all previous surveys as prior information and the data from the current research as data set.

Suppose that there was no prior information before the first survey. Therefore posterior probability obtained from it is:

$$P(H_k \mid D_1 I) = \frac{(n_1 + m - 1)!}{f_{1,1}! f_{2,1}! \ldots f_{m,1!}} \pi_{1k}^{f_{1,1}} \pi_{2k}^{f_{2,1}} \ldots \pi_{mk}^{f_{m,1}} \tag{17}$$

After that we have carried out the second survey. We could put the posterior probability of the first survey as the prior probability in the second. Therefore the new posterior probability is:

$$P(H_k \mid D_1 D_2 I) = \frac{P(D_2 \mid H_k D_1 I) P(H_k \mid D_1 I)}{P(D_2 \mid D_1 I)} \tag{18}$$

If samples of the two surveys are independent[10] then sampling distribution is:

$$P(D_2 \mid H_k D_1 I) = P(D_2 \mid H_k I) = \frac{n_2!}{f_{1,2}! f_{2,2}! \ldots f_{m,2}!} \pi_{1k}^{f_{2,1}} \pi_{2k}^{f_{2,2}} \ldots \pi_{mk}^{f_{2,m}} \tag{19}$$

Marginalizing the numerator of Equation (18) we obtain:

$$\begin{aligned} P(D_2 \mid D_1 I) = P(D_2 \mid I) &= \int_{R_1} \int \ldots \int P(D_2 \mid H_k D_1 I) P(H_k \mid D_1 I) d\pi_{1k} d\pi_{2k} \ldots d\pi_{mk} = \\ &= \ldots = \frac{n_2!}{f_{1,2}! f_{2,2}! \ldots f_{m,2}!} \cdot \frac{(n_1 + m - 1)!}{f_{1,1}! f_{2,1}! \ldots f_{m,1}!} \cdot \frac{g_{1,2}! g_{2,2}! \ldots g_{m,2}!}{(n_1 + n_2 + m - 1)!} \end{aligned} \tag{20}$$

where

$$g_{i,2} = f_{i,1} + f_{i,2} \tag{21}$$

Thus the posterior probability is:

$$P(H_k \mid D_1 D_2 I) = \frac{(n_1 + n_2 + m - 1)!}{g_{1,2}! g_{2,2}! \ldots g_{m,2}!} \pi_{1k}^{g_{1,2}} \pi_{2k}^{g_{2,2}} \ldots \pi_{mk}^{g_{m,2}} \tag{22}$$

[10] Usually this is the case in social surveys.

It is easy to generalize Equation (22) to *t* surveys:

(23) $$P(H_k \mid D_1 D_2 ... D_t I) = \frac{(n+m-1)!}{g_{1,t}!\, g_{2,t}! ... g_{m,t}!} \pi_{1k}^{g_{1,t}} \pi_{2k}^{g_{2,t}} ... \pi_{mk}^{g_{m,t}}$$

where

(24) $$n = \sum_{j=1}^{t} n_j$$

(25) $$g_{i,t} = \sum_{j=1}^{t} f_{i,j}$$

However, Equations (23) and (7) are practically the same. That means that using the data of all previous researches as prior information and the data of current survey as data set is equivalent to using the data of all surveys as data set with no prior information.

However, does that mean that we can merely put the results of all researches together? Or may be this hides some possible dangers? To answer these questions let us prepare some algebra:

(26) $p_{i,j} = \frac{f_{i,j}}{n_j}$ is sample relative frequency,

(27) $$\bar{p}_i = \frac{\sum_{j=1}^{t} p_{i,j} n_j}{\sum_{j=1}^{t} n_j} = \frac{\sum_{j=1}^{t} f_{i,j}}{n} = \frac{g_{i,t}}{n}$$ is average sample relative frequency

Therefore:

(28) $g_{i,t} = n\bar{p}_i$

(29) $$P(H_k \mid D_1 D_2 ... D_t I) = \frac{(n+m-1)!}{(n\bar{p}_1)!(n\bar{p}_2)!...(n\bar{p}_m)!} \pi_{1k}^{n\bar{p}_1} \pi_{2k}^{n\bar{p}_2} ... \pi_{mk}^{n\bar{p}_m}$$

Equation (29) shows that individual values of sample relative frequencies don't matter but their averages do. Thus, if the relative frequencies are stable over time, then every new piece of information will improve our estimation. However, if the relative frequencies are unstable over time or if there is a trend, then it is very dangerous to put all data together. In the last case it is preferable to use only data from the current survey with equal prior probabilities.

Let us look at an extreme example. In Table 4 data from three electoral researches is presented:

Table 4

Vote intentions (number)

Vote intentions	March 2005	May 2005	June 2005	Total
⋮	⋮	⋮	⋮	⋮
I have not decide yet	287	216	101	604
⋮	⋮	⋮	⋮	⋮
Total	1213	1000	1007	3220

Source: http://www.aresearch.org/doc.php?en=1&arch=1&id=581

If we estimate the relative frequency of people who have not decided yet using only the data obtained in June 2005 we will find that:

$$P(0{,}083 < \pi < 0{,}119 \mid DI) = 0{,}95$$

However, if we put all data together and then estimate the same relative frequency, we will find that:

$$P(0{,}174 < \pi < 0{,}200 \mid DI) = 0{,}95$$

The obtained difference is significant. It is due to changes in the vote intentions during the period from March to June 2005. Obviously, the more correct result concerns only the data obtained in June 2005 without any prior information.

4.2. Technical one

When the sample is large enough some computational efforts can be avoided by passing into limit:

$$\lim_{n\to\infty} P(\pi_i = x \mid DI) = \frac{1}{\sqrt{2\pi\sigma_{1,i}^2}} e^{-\frac{(x-\mu_{1,i})^2}{2\sigma_{1,i}^2}} \tag{30}$$

where

$$\mu_{1,i} = \frac{f_i}{n+m-2} \tag{31}$$

$$\sigma_{1,i}^2 = \frac{\mu_{1,i}(1-\mu_{1,i})}{n+m-2} \tag{32}$$

and

$$\lim_{n\to\infty} P\left(\frac{x}{1-y} = u \middle| DI\right) = \frac{1}{\sqrt{2\pi\sigma_{2,i}^2}} e^{-\frac{(u-\mu_{2,i})^2}{2\sigma_{2,i}^2}} \tag{33}$$

where

(34) $$\mu_{2,i} = \frac{f_i}{n - f_m + m - 3}$$

(35) $$\sigma^2_{2,i} = \frac{\mu_{2,i}(1-\mu_{2,i})}{n - f_m + m - 3}$$

Equations (30) and (33) represent so-called *Gaussian distributions*[11]. Using them we can calculate directly the confidence intervals and the probabilities that certain party will pass 4% limit:

(36) $$P(\mu_{1,i} - 1{,}96.\sigma_{1,i} < \pi_i < \mu_{1,i} + 1{,}96.\sigma_{1,i} \mid DI) = 0{,}95$$

(37) $$P\left(\mu_{2,i} - 1{,}96.\sigma_{2,i} < \frac{x}{1-y} < \mu_{2,i} + 1{,}96.\sigma_{2,i} \middle| DI\right) = 0{,}95$$

(38) $$P\left(\frac{x}{1-y} < 0{,}04 \middle| DI\right) = P\left(z_i < \frac{0{,}04 - \mu_{2,i}}{\sigma_{2,i}} \middle| DI\right)$$

where z_i is z-score of *standardized Gaussian distribution*.

Calculated confidence intervals are presented in Table 5 and Table 6.

Table 5

Confidence intervals for relative frequencies

Vote intentions	Confidence intervals
Coalition for Bulgaria	$P(0{,}246 < \pi_1 < 0{,}301 \mid DI) = 0{,}95$
National Movement "Simeon II"	$P(0{,}123 < \pi_2 < 0{,}167 \mid DI) = 0{,}95$
Coalition Union of Democratic Forces – Democratic Party – Gergiovden	$P(0{,}058 < \pi_3 < 0{,}090 \mid DI) = 0{,}95$
Movement for Rights and Freedom	$P(0{,}038 < \pi_4 < 0{,}065 \mid DI) = 0{,}95$
Coalition Bulgarian National Alliance	$P(0{,}024 < \pi_5 < 0{,}047 \mid DI) = 0{,}95$
Party "Democrats for A Strong Bulgaria"	$P(0{,}022 < \pi_6 < 0{,}045 \mid DI) = 0{,}95$
Other	$P(0{,}054 < \pi_7 < 0{,}086 \mid DI) = 0{,}95$
I have not decided yet	$P(0{,}081 < \pi_8 < 0{,}118 \mid DI) = 0{,}95$
I would not vote	$P(0{,}186 < \pi_9 < 0{,}236 \mid DI) = 0{,}95$

[11] It is also called *normal distribution*.

Table 6

Probability of passing 4% limit and confidence intervals for relative frequency of people who choose certain party among the actual voters

Vote intentions	Probability of passing 4% limit (%)	Confidence intervals
Coalition for Bulgaria	100,0	$P\left(0{,}361 < \frac{\pi_1}{1-\pi_m} < 0{,}434 \middle\| DI\right) = 0{,}95$
National Movement "Simeon II"	100,0	$P\left(0{,}181 < \frac{\pi_2}{1-\pi_m} < 0{,}241 \middle\| DI\right) = 0{,}95$
Coalition Union of Democratic Forces – Democratic Party – Gergiovden	100,0	$P\left(0{,}085 < \frac{\pi_3}{1-\pi_m} < 0{,}131 \middle\| DI\right) = 0{,}95$
Movement for Rights and Freedom	100,0	$P\left(0{,}055 < \frac{\pi_4}{1-\pi_m} < 0{,}094 \middle\| DI\right) = 0{,}95$
Coalition Bulgarian National Alliance	93,4	$P\left(0{,}035 < \frac{\pi_5}{1-\pi_m} < 0{,}068 \middle\| DI\right) = 0{,}95$
Party "Democrats for A Strong Bulgaria"	88,5	$P\left(0{,}033 < \frac{\pi_6}{1-\pi_m} < 0{,}065 \middle\| DI\right) = 0{,}95$
Other		$P\left(0{,}079 < \frac{\pi_7}{1-\pi_m} < 0{,}124 \middle\| DI\right) = 0{,}95$

Comparing Table 5 with Table 2 and Table 6 with Table 3 we can see that the differences between exact and approximate results are negligible small. This loss of accuracy is acceptable because the computational efforts are significantly small when we use the Gaussian approximation.

Conclusions:

1. When we estimate relative frequencies:

1.1. If there is no prior information then both the frequentist and the Bayesian approaches are equivalent. In all other cases the Bayesian approach provides an opportunity of taking into account whatever prior information we have.

1.2. If the prior information consists of the data from previous surveys then two different possibilities may exist:

a) There are no significant changes in the relative frequencies over time. Then we can put all data together and we can use them all for estimation of the relative frequencies.

b) There are significant changes in the relative frequencies over time. Then it is dangerous to put all data together. The more harmless way is to use only data from the current survey with equal prior probabilities

2. The Bayesian approach deals easily with two or more parameters which are not independent. In the electoral surveys this allows us to estimate relative frequencies only among the actual voters. This problem is insoluble (or at least the solution is very difficult) from the frequentist point of view. This is "the real test" of the Bayesian approach according to Jaynes' remark: "the real test of any new principle in science is not its ability to re-derive known results, but its ability to give new results, which could not be (or at least, had not been) derived without it" (Jaynes 1974: 24-1).

REFERENCES

Bretthorst, L. 1988. Bayesian Spectrum Analysis and Parameter Estimation, in Lecture Notes in Statistics, 48**,** Springer-Verlag, New York

Bretthorst, L. 1990, An Introduction of Parameter Estimation Using Bayesian Probability, in Maximum Entropy and Bayesian Methods, P. Fougere (ed.), Kluwer Academic Publishers, Dordrecht the Netherlands

Dose, V. 2002. Bayes in Five Days, Lecture notes from a ten hour tutorial on Bayesian analysis given at the International Max-Planck Research School on bounded plasma, http://www.ipp.mpg.de/OP/Datenanalyse/Publications/bib/node1.html

Jaynes, E. 1974. Probability Theory with Applications in Science and Engineering. http://bayes.wustl.edu/etj/science.pdf.html

Jaynes, E. 1976. Confidence Intervals vs Bayesian Intervals, in Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, W. L. Harper and C. A. Hooker (eds.), D. Reidel, Dordrecht

Jaynes, E. 1988. The Relation of Bayesian and Maximum Entropy Methods, in Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1**,** G. J. Erickson and C. R. Smith (eds.), Kluwer, Dordrecht

Jaynes, E. 1993. Probability Theory: the Logic of Science, http://omega.albany.edu:8008/JaynesBook.html