ELICITING BETA PRIOR DISTRIBUTIONS FOR BINOMIAL SAMPLING

José Rafael TOVAR CUEVAS[1]

▪ *ABSTRACT: A procedure is introduced to obtain Beta hyperparameter values for a Bayesian analysis of binary data in situations with very little or null prior information about the parameter of interest and when information has been published about the same. The elicitation procedure is illustrated, using three examples within the clinical diagnostic tests environment.*

▪ *KEYWORDS: Binomial sampling; Chebyshev's inequality; elicitation method.*

1 Introduction

The elicitation of a prior distribution is perhaps the most important subject in Bayesian data analyses. According to Garthwaite *et al.* (2005a,b), elicitation is the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a joint probability distribution. In that way the data analyst (or facilitator) obtains information from an expert in the matter or from historical sources about the unknown quantity and then expresses it in terms of a probability distribution known as the prior distribution. From both psychological and statistical perspectives, many authors have studied the heuristics and biases to take into account in the elicitation process. Some good reviews on this topic were done by Hogarth (1975) and Garthwaite *et al.* (2005a,b). Other authors such as Kadane *et al.* (1980), Kadane and Wolfson (1998), Chaloner and Duncan (1983), Nadler Lins and Campello de Souza (2001) and Enøe *et al.* (2000), have focused their efforts toward constructing methods of elicitation for different probability models.

In Bayesian analyses of proportions, the specification of prior distributions is generally based on subjective probabilities about the unknown parameters through

[1]Universidad del Rosario, Escuela de Medicina y Ciencias de la Salud, Carrera 24 No 63C69 Barrio Siete de Agosto, Santa Fe de Bogotá, Colombia. E-mail: *jose.tovarc@urosario.edu.co*

the use of the beta distributions family. This family of distributions is very flexible, contains a wide variety of shapes, and has components that are natural conjugate priors for binomial sampling distributions, which greatly simplifies the subsequent derivations. Using the prior predictive distribution of binomial sampling with beta prior distribution, Chaloner and Duncan (1983) developed an algorithm to obtain the beta prior hyperparameters, approximating them with the predictive mode; method that was later modified by Gavasakar (1988). Bedrick *et al.* (1996, 1997) developed a methodology to obtain prior distributions for the parameters of the logistic regression model, using the expert's information about the probabilities of success, obtaining prior distributions called conditional means priors.

Within a clinical diagnostic environment, some authors have developed procedures to obtain beta prior hyperparameters. Joseph *et al.* (1995) proposed two approaches to reach that goal, using fixed intervals to obtain information from experts. In the approach they determined equally tailed 95% credible intervals (i.e. 2.5% in each tail) derived from a review of the relevant literature and clinical opinions obtained from a panel of experts on the matter. The prior density is selected by matching the center of the interval's range with the mean of a beta distribution and matching the standard deviation of the beta distribution with the quarter of the total range in the chosen interval. The second approach is to match the end points of the given intervals to beta distributions with similar 95% probability intervals, using simulated data from a computer program such as R or S-plus. Enøe *et al.* (2000) proposed to elicit the most probable value or best guess (θ_0) which may be an actual estimate based on previous data. The next step is to get the value (θ_L) for which the expert is certain $(1-\alpha/2)$ that the parameter will be larger and a second value (θ_U) for which the experimenter is $(1-\alpha/2)$ certain that the parameter will be smaller. Then, for a given guess θ_0 and a given value of b, a is found. With the obtained values of a and b, a software as R can be used to determine whether the appropriate percentiles of the specified Beta(a,b) prior distribution are given by (θ_L, θ_U).

In this paper a method similar to that developed by Joseph *et al.* (1995) is introduced to obtain the Beta hyperparameters. Similarly equally tailed $(1-\alpha)$ percent probability intervals are used, taking the center value of the interval as the mean of a Beta distribution. To approximate the prior variance, Chebyshev's inequality is used. The procedure can be extended for situations where it is not possible to have an expert's opinion on the subject in order to construct the intervals; i.e, not having any prior information about the performance of the parameter of interest or for situations where it is not possible to have an expert's opinion of expert on the matter but there is some published information.

This paper is organized as follows: in Section 2, the eliciting situations and a procedure to approximate the variance of the Beta distribution which will be used to obtain the prior hyperparameters are introduced, in Section 3, the procedure is illustrated using three examples and finally in the Section 4, is presented some conclusions.

2 A conceptual framework

2.1 Clinical diagnosis and copula functions

Given that two of the data sets used to illustrate the application of the elicitation procedure are within a clinical diagnostic environment which is possible to use the information obtained from two indexes reported in the literature (Böhning and Patilea´s Indexes), it is important to clarify some concepts that will be used in the examples.

Within a design to evaluate the performance of a clinical diagnostic procedure that includes the measure of two biological traits as screening tests, it is assumed that the test outcomes are realizations of random variables (generally non-observable) V_1 and V_2 measured on a continuous positive scale; i.e., $V_1 > 0$ and $V_2 > 0$. It is also assumed that two cut-off values ξ_1 and ξ_2 were chosen for the tests in order to determine when an individual is classified as positive or negative. Thus it can be assumed that an individual is positive for test ν if $V_\nu > \xi_\nu$ which is expressed through another binary random variable T_ν that takes the value of one ($T_\nu = 1$ if and only if $V_\nu > \xi_\nu$ for $\nu = 1, 2$). To measure the degree of the dependence between the random variables V_1 and V_2, the use of a copula function can be considered.

Copula functions are used to link marginal distributions with a joint distribution using the probability integral transformation. For specified marginal univariate distribution functions $F_1(v_1), \ldots, F_m(v_m)$, the function $C(F_1(v_1), \ldots, F_m(v_m))$ defined using the copula function C, results in a multivariate distribution function such that, $F(v_1, \ldots, v_m) = C(F_1(v_1), \ldots, F_m(v_m))$. In the special case of bivariate distributions ($m = 2$); and the random variables V_1 and V_2, the continuous distribution functions F_1 and F_2 make it possible to define $U = F_1(v_1)$ and $W = F_2(v_2)$ so $U \sim U(0, 1)$ and $W \sim U(0, 1)$ which are usually dependent if V_1 and V_2 are dependent, (thus independent V_1 and V_2 imply that U and W are independent); consequently, the problem is reduced to specifying a bivariate distribution between two uniform variables, it is a copula.

Besides screening tests, there is a random binary variable D that denotes the true health status of an individual. It is established through the "gold standard" procedure which classifies individuals without error: $D = 1$ denotes an disease individual and $D = 0$ denotes a healthy individual, where the unknown probabilities to be estimated; $P(D = 1) = p$, $S_j = P(T_j = 1|D = 1)$ and $E_j = P(T_j = 0|D = 0)$ can be defined as the prevalence of the disease in the population, the sensibility of the jth test and the specificity of the jth test ($j = 1, 2$). In the second example of this paper, it is assumed that the screening test outcomes are interdependent; and the Gumbel copula function dependence parameter (Gumbel, 1960, Nelsen, 2006) is used to model that interdependence.

The Gumbel copula function is defined as follows:

$$\begin{aligned} C(u, w) &= u + w - 1 + (1 - u)(1 - w) \exp\{-\phi \ln(1 - u) \ln(1 - w)\} \\ & \quad 0 \leq u, w \leq 1, \quad \phi \in (0, 1). \end{aligned} \tag{1}$$

When there is a design with two Gumbel structures of dependence diagnostic tests -each with an outcome (measured or not) on a continuous scale and a cut-off point to dichotomization- the contributions to likelihood function are as shown in Table 1.

Table 1 - Contributions of all possible combinations of outcomes of T_1, T_2 and D to likelihood if independence or a Gumbel copula dependence structure is assumed, where f_i is the number of individuals in cell i, $i = 1, 2, ..., 8$. (ϕ_k is the Gumbel dependence parameter ($k = D, ND$))

					Contribution to likelihood	
i	D	T_1	T_2	f_i	Independent tests	Gumbel dependence structure
1	1	1	1	f_1	pS_1S_2	$pS_1S_2Q_1$
2	1	1	0	f_2	$pS_1(1-S_2)$	$pS_1[1-S_2Q_1]$
3	1	0	1	f_3	$p(1-S_1)S_2$	$pS_2[1-S_1Q_1]$
4	1	0	0	f_4	$p(1-S_1)(1-S_2)$	$p[1-S_1-S_2+S_1S_2Q_1]$
5	0	1	1	f_5	$(1-p)(1-E_1)(1-E_2)$	$(1-p)(1-E_1)(1-E_2)Q_2$
6	0	1	0	f_6	$(1-p)(1-E_1)E_2$	$(1-p)(1-E_1)[1-(1-E_2)Q_2]$
7	0	0	1	f_7	$(1-p)E_1(1-E_2)$	$(1-p)(1-E_2)[1-(1-E_1)Q_2]$
8	0	0	0	f_8	$(1-p)E_1E_2$	$(1-p)[E_1+E_2-1+(1-E_1)(1-E_2)Q_2]$
					$Q_1 = \exp(-\phi_D \ln S_1 \ln S_2)$,	$Q_2 = \exp(-\phi_{ND} \ln(1-E_1) \ln(1-E_2))$

2.2 Böhning and Patilea indexes

Böhning and Patilea (2008) developed two Indexes to measure the association between diagnostic test outcomes, assuming that they have the same binary structure. The first of them, the λ index for diseased individuals ($k = D$) and non-diseased individuals ($k = ND$) is defined as,

$$\begin{aligned} \lambda_k &= \frac{P(T_1 = 1|T_2 = 1, D = k)}{P(T_1 = 1|D = k)} \\ &= \frac{P(T_1 = 1, T_2 = 1|D = k)}{P(T_1 = 1|D = k)P(T_2 = 1|D = k)}; \qquad \lambda_k \in (0, \infty) \end{aligned} \tag{2}$$

Therefore, if $\lambda_k = 1$ the tests results are associated; if $\lambda_k < 1$ there is negative association between tests and if $\lambda_k > 1$, the association between tests is positive.

Similarly, the δ index for each population of individuals is defined by,

$$\delta_k = \frac{P(T_1 = 1, T_2 = 1|D = k)P(T_1 = 0, T_2 = 0|D = k)}{P(T_1 = 1, T_2 = 0|D = k)P(T_1 = 0, T_2 = 1|D = k)} \quad \delta_k \in (0, \infty) \tag{3}$$

Thus, δ_k is defined as the odds ratio in the k^{th} diseased state; and when $\delta_k = 1$ there is no association between tests; a negative association is expressed by $\delta_k < 1$ and a positive association by $\delta_k > 1$. The λ_k and δ_k indexes are used to estimate the numbers of truly diseased (n_+) and healthy (n_-) individuals within the group with negative results in both screening tests when a design with verification bias has been used.

3 Methods

3.1 Procedure for Eliciting prior distributions

When eliciting a prior distribution for unknown quantities in binomial sampling, three different situations can occur:

- Situation 1: There is prior information about the unknown quantities from published results of other studies or from a specialist in the subject under study.
- Situation 2: There is no information about the topic of interest and/or there is no specialist in the subject; however information can be constructed from data or results of research conducted to estimate related parameters or functions of the unknown quantities.
- Situation 3: There is no information obtainable about the parameter of interest, or the information that is available is considered too poor to carry out the eliciting process.

Situation 1: When it is possible to obtain results from other studies about the first two distributional moments of prior distribution to be used, the matching hyperparameters θ_0 and σ^2 can be obtained with the arithmetic mean and variance computed from the published results. Then Equation 7 (Section 2.2) is used. This situation is illustrated in Example 1.

Situation 2: When there is no prior information on the parameters of interest, subjective information can be constructed, using published information on functions or other parameters related in some way to the parameter to be estimated. Then fixed intervals are proposed with a given probability that can include the unknown quantity; and the procedure described in Section 2.2 is applied to obtain the Beta hyperparameters of various prior distributions. These are then evaluated, observing changes and trends in the posterior distributions before selecting one of them, using an appropriate selection criterion.

situation 3: When there is prior information on the parameters of interest, but there is no information about their functions or other related parameters, the following alternative can be considered: divide the parametric space into k-bounded, equally tailed $1-\alpha$ percent probability exclusive intervals so that the sum of their ranges is equal to the range of the full parametric space. Then the hyperparameter values are obtained for each of them, using the procedure described in Section 2.2. Next, with each of the k-based prior distributions, the unknown quantity is estimated and the subsequent results are compared, using any of the existing methods for selecting models (e.g., Bayes factor, DIC, Kullback-Leibler distance or maximum entropy).

3.2 Obtaining hyperparameters for the Beta distribution

Let us to define by θ the proportion of interest which is a random amount with an unknown probability distribution $\pi(\theta)$. A natural choice for $\pi(\theta)$ is the

Beta distribution with hyperparameters a and b. Using the available information obtained from some source, it is possible to determine equally tailed $1-\alpha$ percent probability intervals with bounds θ_1 and θ_2 in the same way as Joseph, Gyorkos and Coupal (1995) did within a fixed-intervals approach; thus, $(1-\alpha)\%$ of the values of θ falls within the interval (θ_1, θ_2). In a second step, the center of the interval (θ_0) is matched with the mean of the beta (a,b) distribution; i.e.

$$\theta_0 = \frac{\theta_1 + \theta_2}{2} \cong \frac{a}{a+b} \tag{4}$$

Now there are two percentiles of the prior distribution (e.g., 2.5^{th} and 97.5^{th} percentiles) and its first moment, thus it is possible to use the Chebyshev's inequality to obtain the variance as follows;

$$\begin{aligned} P(|\theta - E(\theta)| \geq k\sigma) &\leq \frac{1}{k^2} = \alpha \\ P([\theta - E(\theta)]^2 \geq k^2\sigma^2) &\leq \alpha \\ P(\alpha[\theta - \theta_0]^2 \geq \sigma^2) &\leq \alpha \end{aligned} \tag{5}$$

In accordance with last line in Equation (5), variance will be a function of the established prior probability to interval that contains θ and the distance between θ_0 and a percentile of the distribution. If θ is replaced by one of the known values θ_1 or θ_2 in Equation (5), an approximate value can be obtained for the variance of the Beta prior distribution, as follows;

$$\sigma^2 \leq \alpha[\theta_1 - \theta_0]^2 \cong \frac{ab}{(a+b)^2(a+b+1)} \tag{6}$$

As the mean $\theta_0 = E(\theta)$ and the variance $\sigma^2 = V(\theta)$ can be written as functions of the Beta prior hyperparameters, then the problem becomes a matter of resolving a system of two equations with two unknowns (using the right sides part of (4) and (6)) to finding the values of a and b i.e:

$$\begin{aligned} \omega &= \frac{\theta_0}{(1-\theta_0)} \\ a &= \omega b \\ b &= \frac{\omega - [(\omega+1)^2\sigma^2]}{(\omega^3 + 3\omega^2 + 3\omega + 1)\sigma^2} \end{aligned} \tag{7}$$

4 Examples

To illustrate the proposed elicitation procedure, two data sets from clinical diagnostic studies that include the use of two screening tests and a gold standard to verify the final health status of each individual, and a third one obtained in a prevalence study. In the first example independence between test outcomes is assumed, and the focus is mainly on estimating test performance parameters,

using the contributions to likelihood that appear in the left column of Table 1. In the second example, it is assumed that both clinical tests respond to underlying (observable or not) biological traits expressed on a continuous scale, dichotomized using a cut-off point. There is a weak nonlinear dependence between outcomes that can be modeled using the Gumbel copula, and the likelihood contributions are those that appear in the right column of Table 1. The third example deals with the problem of estimating the prevalence of three sexually transmissible infections in a Colombian indigenous community.

4.1 Example 1: Urinary tract infection (UTI)

The UTI data introduced by Ali et al. (2007) were used to estimate the performance parameters from two fast tests for screening for UTI in children ranging from 1mo to 11 yr of age. The diagnostic procedure includes two screening tests [the presence of nitrite in the urine (T_1) and Leukocyte Sterase levels in urine (T_2) and urine culture as gold standard. Therefore, we have a vector of five parameters to estimate; $\theta = [S_1,\ S_2,\ E_1,\ E_2,\ p]$. The authors compared their outcomes for performance test estimates and prevalence with those obtained in other four studies. The interest is to estimate the same parameter vector using a Bayesian approach, then it is necessary to eliciting prior distributions for five unknown proportions and under the assumption of prior independence between the parameters that compose the vector θ, it is possible to use a Beta$(\alpha_{\theta_{1,k}}, \beta_{\theta_{1,k}})$, $k = 1, 2, ..., 5$ prior distribution for each of them, where $\alpha_{\theta_{1,k}}$ and $\beta_{\theta_{1,k}}$ denote the prior hyperparameters. To elicit the hiperparameter values it was used the results of the other four studies (Table 2) as follows: First, it was matched the means and variances calculated using the outcomes of the five studies with θ_{0_i} and σ_i^2, next it was used the equation (7) to obtain the $(a_{\theta_i}, b_{\theta_i})$, $\quad i = 1, 2, ..., 5$ hyperparameters of the Beta distributions.

Table 2 - Outcomes obtained by Ali *et al.* (2007) and results of other four published studies

		Nitrite (T_1)		Leucocyte Sterase (T_2)	
study	Incidence %	S_1	E_1	S_2	E_2
Ali *et al.* (2007)	67.4	38.2	88.4	85.4	58.1
Sharief *et al.* (1998)	5.2	54.6	96.8	100	78.1
Weinberg and Gan (1991)	4.01	56.0	98.1	85.4	92.7
Lohr *et al.* (1993)	14.8	37.3	100	79.4	72.7
Cannon *et al.* (1986)	14.4	72.7	99.6	84.6	71.4
Mean θ_0	21.2	51.8	96.6	87.0	74.6
Variance σ^2	963.3	214.4	22.5	59.4	156.5

Thus the informative prior distributions for prevalence and performance test parameters are given by:

$$S_1 \sim Beta(4.15, 4.5), \quad S_2 \sim Beta(15.7, 2.4) \quad E_1 \sim Beta(0.5, 13),$$

$$E_2 \sim Beta(8.3, 2.8), \sim p \sim Beta(2.1, 22.3)$$

4.2 Example 2: Cancer data

In this example a data set introduced by Smith *et al.* (1997) was used, where 19, 476 men were screened for prostate cancer using Prostatic Specific Antigen (PSA) (T_1) and the Digital Rectal exam DRE (T_2). The PSA level was considered suspicious for cancer if it exceeded 4.0 ng/ml. Subjects with positive results on either the DRE or PSA were submitted to an ultrasound-guided needle biopsy test, which was considered as gold standard. Smith, Bullock and Catalona's study data set was used by Böhning and Patilea (2008) to illustrate the use of two indexes by they developed. In this example it was assumed that the test results were dependent and could be modeled using the Gumbel copula structure. Thus the problem was reduced to estimating the vector $\theta = [S_1,\ S_2,\ E_1,\ E_2,\ p,\ \phi_D,\ \phi_{ND}]$. assuming prior independence among the components.

Given that this study had a verification bias (individuals with negative outcomes were not verified by gold standard), the numbers of individuals with negative outcomes in screening tests that were not verified by the gold standard (n_+ and n_-) were calculated in order to obtain some prior information about the prevalence and test performance parameters. Böhning and Patilea's indexes were used so there were two estimates for each parameter: one from the number obtained using the λ estimate and the other obtained using the δ estimate (Table 3). It was assumed that the two estimates of each parameter were the limits of an interval containing 95% of the possible values of the proportion of interest.

Table 3 - Estimated values for quantities of nonverified individuals with prostate cancer using Böhning and Patilea indexes; values in brackets were calculated using δ index and the other one using λ index

	Diseased subjects $\lambda = 2.42$, $\delta = 3.08$			Non-diseased subjects $\lambda = 2.40$, $\delta = 3.03$		
	$DRE+$	$DRE-$	Total	$DRE+$	$DRE-$	Total
$PSA+$	189	292	481	141	755	896
$PSA-$	145	1431[691]	1576[836]	1002	15521[16261]	16523[17263]
Total	334	1723[983]	2057[1317]	1143	16276[17016]	17419[18159]

With the constructed intervals and using the procedure described in Section 2, values for the hyperparameters a_i and b_i were obtained for each Beta prior distribution. (Table 4).

Given the difficulty of obtaining prior information about the copula dependence structure, it was necessary to construct subjective information using the values of δ_k and λ_k estimates. It was found that the indexes had similar values within both groups (diseased and healthy individuals), which implies that the two screening tests have a similar dependence structure in both populations. Both indexes were positive and were close to unity; therefore it is possible to conclude that there is some type of dependence (positive) between the two test

Table 4 - Prior distribution hyperparameters for performance tests, prevalence and Gumbel dependence parameters using the the author's proposed method

θ	Interval	θ_0	a	b
S_1	0.234 - 0.365	0.301	303	704
S_2	0.162 - 0.254	0.208	324	1232
E_1	0.949 - 0.951	0.950	902500	47500
E_2	0.934 - 0.937	0.936	501758	34595
p	0.068 - 0.106	0.087	379	4002
ϕ_1	0.0 - 0.25	0.125	17	122
ϕ_2	0.25 - 0.75	0.500	39.5	39.5
ϕ_3	0.75 - 1.0	0.875	122	17

outcomes, but very weak; this copula function does not model linear positive dependencies (when $\phi = 1$ the Pearson correlation linear coefficient (ρ) takes the value -0.40365) and its dependence parameter belongs to (0.1) interval. The constructed subjective information about the Gumbel dependence parameter does not allow eliciting a prior distribution in terms of a particular distribution of probability; therefore, the eliciting procedure associated with Situation 3 in Section 2 is more appropriate in this case. The parametric space of ϕ was divided into three arbitrary and exclusive intervals; namely: $\phi \in (0, 1/4)$, $\phi \in (1/4, 3/4)$ and $\phi \in (3/4, 1)$; and it was assumed that within each of these intervals, the most probable values of the parameter could be contained with a probability of 0.95. Next, the resulting approximated variance and beta hyperparameters were computed in each interval, obtaining Beta(17, 122), Beta(39.5, 39.5) and Beta(122, 17) as the prior distributions for the Gumbel dependence parameters. Using each prior distribution and the likelihood function that appear in Table 1, the estimators for the parameters were computed. The prior distribution with the best performance was selected, using the DIC as criterion

4.3 Example 3: Prevalences in an indogenous community

An epidemiological survey was carried by Dr. Carlos Alberto Rojas and his research group from the Universidad de Antioquia. The main goal of the study was to estimate the prevalence of HIV infection, syphilis and hepatitis B in a sample of 295 individuals from an indigenous community in eastern Colombia. According to the expert (Dr. Rojas), it is very difficult to obtain prior information about the sexual infections prevalent in indigenous communities because it is a very little studied subject, especially in countries like Colombia, where most of the indigenous communities still have their own regulations and policies. In this section the data set of that study was used to illustrate the elicitation procedure. In the epidemiological survey, 3 HIV infection cases, 8 syphilis infection cases and 0 cases of hepatitis B were found. To elicit the prior distributions, the expert was asked his opinion with respect to the observed proportion of cases (1.02% for VIH and 2.7% for Syphilis). He considered that the value for HIV was higher

than to be expected for an indigenous community (the parameter should be about 1%), while for syphilis the prevalence should range from 1-3%. To improve the information about HIV prevalence, more data were sought in an official document published by the Ministerio de la Protección Social (2010), where exact information about the prevalence of HIV among indigenous people was not found, just the prevalence of infection in different population groups of the country. In agreement with the published source, the estimated prevalence of HIV infection in populations ranging from 14-49 years of age is 0.59%; and in the total Colombian population, the estimated prevalence is about 0.22%. Then an interval was constructed, assuming that the prevalence of HIV infection among the indigenous population must be from 0.0022-0.0059 (probability of 0.95). This was discussed with the expert who agreed. Using the proposed procedure, a Beta(95, 23475) was obtained as the prior distribution for the prevalence of HIV infection.

In the case of the prevalence of syphilis, the same assumption was made as to the interval (0.01, 0.03), which was divided into three (0.01, 0.016667), (0.0166667, 0.023333) and (0.023333, 0.03). Applying the procedure once again, the following prior distributions were obtained: Beta(316, 23364), Beta(705, 34566) and Beta(1244, 45425). For each prior distribution, 100 000 samples of its subsequent distribution were simulated; and the MCMC estimates, the DIC and the Bayes factor (BF) were obtained. The selected prior distribution had the lowest DIC and the highest BF value (Table 5). In accordance with the expert's opinion, hepatitis B is less common than HIV and syphilis infections in indigenous communities so its prevalence must be very close to zero. Therefore, it was assumed that the prevalence would never be greater than 1% so it would be within the interval (0, 0.01) even when the expert expected the value to be closer to zero than 0.01. The proposed interval was divided into three parts as follows: (0, 0025), (0.0025, 0.0075) and (0.0075, 0.01) to obtain the beta prior distributions Beta(20, 15959), Beta(80, 15839) and Beta(971, 110048), with the first one being selected as the prior distribution (Table 5).

Table 5 - Hyperparameters for prior distributions, DIC values and BFs obtained in estimating the prevalences of hepatitis B and siphilys. (y= no. of cases observed among the 295 individuals tested)

	Interval	Hyperparameters (a,b)	DIC	BF
	(0, 0.0025)	20, 15959	0.725	$K_{1,2}$=2.993
Hepatitis B	(0.0025, 0.0075)	80, 15839	2.919	$K_{2,3}$=3.081
y=0	(0.0075, 0.01)	971, 110048	5.170	$K_{1,3}$=9.224
	(0.01, 0.01666)	316, 23364	7.132	$K_{2,1}$=3.481
Siphilys	(0.01666, 0.02333)	705, 34566	7.130	$K_{3,2}$=1.401
y=8	(0.02333, 0.03)	1244, 45425	3.926	$K_{3,1}$=38.56

5 Concluding remarks

The elicitation of prior distribution is perhaps the most important step in a Bayesian data analysis. In that part of the estimation process, the statistician or facilitator takes the subjective information that an expert has about the unknown quantities in the phenomenon under study and translates it into terms of a probability distribution that increases the information contained in the data obtained under experimental conditions. In some cases there may be no expert in the matter so the data analyst has to look for historical or published information about the quantities in different sources in order to propose a prior distribution. In this paper a procedure was constructed from a combination of different strategies published in the literature in order to obtain the prior distribution for binomial sampling schemes. The method proposed herein not only considers situations in which it is easy to obtain published information or there is a specialist on the subject of the study, but also other cases where the data analyst does not have any information about the possible distributional performance of the unknown quantity of interest. To obtain the hyperparameter values, the midpoint of a fixed interval is taken as the mean; and Chebyshev's inequality is used to obtain an approximation of the prior variance. In agreement with Casella and Berger (1990), Chebyshev's inequality is necessarily conservative given that it can be applied to completely arbitrary distributions. Thus the approximation for the variance may be somewhat inaccurate and larger than the true variance of the parameter. In cases where it is not possible to get any information about the second population moment and it is important to obtain the hyperparameter values, this approximation can be a useful tool. The proposed elicitation procedure can be used only when the parameters of the prior distributions have some functional relationship with the first two distributional moments; in other cases it cannot be applied. An advantage of this procedure is that, unlike Joseph's method, it does not assume symmetry with respect to the tail areas in the prior distribution, which is readily observable in the examples where both parameters a and b have very different values. The procedure's performance was illustrated using the beta distribution and proportions as unknown quantities to be estimated; but the method could be used with other families of distributions whose parameters have the aforementioned characteristic.

When it is necessary to evaluate partitions of the parametric space (Situation 3), it is important to observe that in the intervals obtained at the extremes, the negligible tail probability is placed at the endpoint of the interval. This results in a truncated distribution at one end of the constructed interval, which is not generally a problem because there is usually no interest in estimating the extreme value (zero or one) of the proportion. Given that there are k (in the examples, k=3, which is an arbitrarily selected number) prior distributions with their corresponding distributions (after using the observed data), it is necessary to select one of them as the best candidate for carrying out the estimation process. In this case two selection-model criteria from the literature were used to select the prior distribution. The prior distribution obtained in this way and under situations of absolute or great

ignorance about the possible distributional performance of the unknown quantity can be evaluated with data obtained in new experiments. In the examples presented herein, three arbitrary and exclusive intervals were used to divide the parametric space; but there are many possible ways to do this as can be observed in the detailed study by Nadler Lins and Campello de Souza (2001) about constructing intervals within a parametric space.

In the first example, the estimates of the incidences obtained in studies developed in different countries to elicit the author's prior distribution, implies that it is being assumed that the prevalence and incidence of urinary infections in children have the same distribution and are similar in the different countries where the studies were conducted. That is a strong assumption that may be difficult to prove, but it is possible to think that the distribution of the prevalence covers a wide range that contains all the values observed in the studies. That may be a rough approximation; but the same could be argued for a specialist on the subject; or that the results obtained with it can be compared with those obtained using other methods of estimation. In the second example (cancer data), it was necessary to estimate two Gumbel-type dependence parameters, which is very complicated due to the difficulty of interpreting exactly what the copula dependence parameter means. Few people are familiar with the copula functions -models of relatively novel use, whose theory is still evolving- thus it is very complicated to grasp the concept of the performance of a copula parameter. One way to address this problem would be to elicit prior to a known dependence parameter (e.g. Spearman's rho or Kendall's tau) under the assumption that eliciting prior distributions for linear dependence forms is easier than doing so for other types of dependencies. In agreement with Clemen *et al.* (2000), who did a study about assessing linear dependencies, evaluating six different methods of asking the expert on a given subject, the most accurate way to obtain a subjective dependence measure was to ask the expert to estimate the correlation between two variables of interest. Given that the indexes constructed from the agreement among pairs of observations have a functional relationship with the copula parameters, that relationship could be used to elicit the prior distribution; however, for many copula functions, the function link between the indexes of concordance and the dependence parameter involves integrals that do not have an analytic solution, which makes the elicitation process more difficult.

Acknowledgments

The author would like to thank Dr. Carlos Alberto Rojas, for his help with the data and his valuable participation in the eliciting process. I am also grateful to Professor Jorge Alberto Achcar for his insightful observations on the subject of this paper.

TOVAR CUEVAS, J. R. Obtenção de distribuições priori Beta para verossimilhanças Binomiais. *Rev. Bras. Biom.,* São Paulo, v.30, n.1, p.159-172, 2012.

▪ RESUMO: Neste artigo é introducido um procedimento para obter valores dos hiperparametros da distribuição Beta na analise Bayesiana de dados com estrutura binária em situações nas que se tem muito pouca ou nenhuma informação priori sobre o parámetro de interesse e quando é possivel ter alguma informação publicada. O procedimento de elicitação é ilustrado usando três exemplos dentro do ambiênte dos testes para diagnóstico clínico.

▪ PALAVRAS-CHAVE: Distribuição de amostragem binomial; desigualdade de Chebyshev; método de elicitação.

References

ALI, S. H. G.; MOODAMBAIL, A. R.; HAMRAH, E. K. B.; BIN-NAKHI, H. A.; SADEQ, S. A. Reliability of rapid dipstick test in detecting urinary tract infection in symptomatic children. *Kuwait Med. J.*, v.39, p.36-38, 2007.

BEDRICK, E. J.; CHRISTENSEN, R.; JOHNSON, W. A new perspective on priors for generalized linear models. *J. Am. Stat. Assoc.*, v.91, p.1450-1460, 1996.

BEDRICK, E. J.; CHRISTENSEN, R.; JOHNSON, W. Bayesian binomial regression: predicting survival at a trauma center. *Am. Stat.*, v.51, p.211-218, 1997.

BÖHNING, D.; Patilea V. A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the positives only. *J. Am. Stat. Assoc.*, v.103, p.212-221, 2008.

CANNON, H. J.; GEETZ, E. S.; HAMOUDI, A. C.;MARCON, M. J. Rapid screening and microbiologic processing of pediatric urine specimens. *Diagn Microbiol Infect Dis.*, v.4, p.11-17, 1986.

CASELLA, G.; BERGER, R. L. Statistical inference. Belmont, CA: Duxbury Press, 1990.

CHALONER, K. M.; DUNCAN, G. T. Assessment of a Beta prior distribution: PM elicitation. *J. Roy. Stat. Soc. D-Sta.*, v.32, p.174-180, 1983.

CLEMEN, R. T.; FISHER, G. W.; WINKLER, R .L. Assessing dependence: some experimental results. *Manage. Sci.*, v.46, p.1100-1115, 2000.

ENØE, C. GEORGIADIS, M. P.; JOHNSON, W. O. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.*, v.45, p.61-81, 2000.

GARTHWAITE, P. H.; KADANE, J. B.; O'HAGAN, A. Statistical methods for eliciting probability distributions. *J. Am. Stat. Assoc.*, v.100, n.470, p.680-701, 2005a.

GARTHWAITE, P. H.; KADANE, J. B.; O'HAGAN, A. Statistical methods for eliciting probability distributions. Published in http://www.stat.cmu.edu/tr/tr808/tr808.pdf, January 4, 2005b, downloaded in September 2011.

GAVASAKAR, U. A comparison of two elicitation methods for a prior distribution for a binomial parameter. *Manage. Sci.*, v.34, p.784-790, 1988.

GUMBEL, E. J. Bivariate exponential distributions. *J. Am. Stat. Assoc.*, v.55, p.698-707, 1960.

HOGARTH, R. M. Cognitive process and the assessment of subjective probability distributions. *J. Am. Stat. Assoc.*, v.70, p.271-289, 1975.

JOSEPH, L.; GYORKOS, T. W.; COUPAL, L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in absence of a gold standard. *Am. J. Epidemiol.*, v.141, p.263-272, 1995.

KADANE, J. B., WOLFSON, L. J. Experiences in elicitation. *J. Roy. Stat. Soc. D-Sta.*, v.47, p.2-18, 1998.

KADANE, J. B.: DICKEY, J. M.; WINKLER, R. L.; SMITH, W.; PETERS, S. C. Interactive elicitation of opinion for a normal linear model. *J. Am. Stat. Assoc.*, v.75, p.845-854, 1980.

LOHR, J. A., PORTELLA, M. G., GEUDER,T. G. et al. Making a presumptive diagnosis of urinary tract infection by using urinalysis performed in an on-site laboratory. *J. Pediatr.*, v.122, p.22-25, 1993.

MINISTERIO DE LA PROTECCIÓN SOCIAL DE COLOMBIA. Resumen de la epidemia por VIH/SIDA en Colombia. Published in <http://www.minproteccion social.gov.co/salud/Documentos/resumen situación epidemia VIH SIDA 1983-2009.pdf>, 2010, downloaded in October 2011.

NADLER LINS, G.; CAMPELLO DE SOUZA, F. A protocol for the elicitation of prior distributions. In: 2nd International Symposium on Imprecise Probabilities and Their Applications. Ithaca, NY: Cornell University, 26-29 June, 2001. Electronic proceedings (pdf format), 8p.

NELSEN, R. B. An introduction to copulas. 2.ed. New York: Springer Series in Statistics. Springer Science+Business Media, 2006.

SHARIEF, N., HAMEED, M., PETTS, D. Use of rapid disptick tests to exclude urinary tract infections in children. *Br. J. Biomed. Sci.*, v.55, p.242-246, 1998.

SMITH, D. S.; BULLOCK, A. D.; CATALONA; W. J. Racial differences in operating characteristics of prostate cancer screening tests. *J. Urol.*, v.158, p.861-1865, 1997.

WEINBERG, A. G., GAN, V. N. Urine screen for bacteria in symptomatic pediatric outpatiens. *Pediatr. Infect. Dis. J.*, v.10, p.651-654, 1991

Received in 17.02.2012.

Approved after revised in 11.07.2012.