

Dariusz Biskup

FORMUŁOWANIE SUBIEKTYWNYCH ROZKŁADÓW A PRIORI – PODEJŚCIE PARAMETRYCZNE I NIEPARAMETRYCZNE

1. Wstęp

Statystyka może być określona jako nauka zajmująca się badaniem sytuacji, w których występuje zjawisko niepewności. We wnioskowaniu statystycznym można wyróżnić dwa rodzaje niepewności: niepewność dotyczącą danych statystycznych oraz niepewność dotyczącą parametrów. W ujęciu bayesowskim rozkład prawdopodobieństwa przypisywany jest zarówno danym, jak i parametrom, od których te dane zależą. Jeżeli mamy więc do dyspozycji zbiór obserwacji $x_1, x_2, \dots, x_n$, które zależą od parametru θ (względnie od wektora parametrów), posługujemy się następującymi rozkładami prawdopodobieństwa:

- rozkładem *a priori* $p(\theta)$, który to rozkład opisuje naszą niepewność dotyczącą parametru przed zebraniem danych statystycznych,
- funkcją wiarygodności $P(x_1, x_2, \dots, x_n | \theta)$, która opisuje rozkład obserwacji względem ustalonej wartości parametru θ ,
- rozkładem *a posteriori* $P(\theta | x_1, x_2, \dots, x_n)$, który wyznaczany na podstawie wzoru Bayesa łączy dwa poprzednie rozkłady, a więc informacje *a priori* z informacjami z próby.

Podstawowa różnica pomiędzy podejściem bayesowskim a klasycznym przejawia się w sposobie traktowania parametru θ . W ujęciu bayesowskim parametr θ , mimo że jego wartość jest ustalona i nie zmienia się losowo, ma przypisany rozkład prawdopodobieństwa, który opisuje niepewność badacza dotyczącą jego wartości. Taki sposób traktowania parametru wynika z przekonania (będącego rezultatem przyjęcia odpowiedniego zbioru aksjomatów), że niepewność musi być

mierzona poprzez prawdopodobieństwo (por. [10]). Prawdopodobieństwo jest więc definiowane jako warunkowa miara niepewności, zależna od stanu naszej wiedzy. Wszystkim wielkościom niepewnym występującym w analizie przypisywany jest rozkład prawdopodobieństwa, a wnioskowanie statystyczne jest przeprowadzane wyłącznie przy użyciu reguł rachunku prawdopodobieństwa.

Klasyczna statystyka matematyczna jest jedną z niewielu dziedzin matematyki, które nie mają ujęcia aksjomatycznego. Prowadzi to w wielu sytuacjach do przyjmowania różnych, pozornie intuicyjnie atrakcyjnych technik, które często prowadzą do różnych wniosków. Przegląd paradoksalnych bądź niepożądanych wniosków wynikających z założeń statystyki klasycznej można znaleźć w [1; 7].

Wnioskowanie bayesowskie z kolei ma dobrze ugruntowany, intuicyjnie przekonujący zbiór aksjomatów, który zapewnia wewnętrzną spójność formułowanych wniosków. Ujęcia aksjomatyczne sformułowane zostały m.in. przez Savage'a i deGroota (por. [13; 4]).

2. Subiektywne i obiektywne rozkłady *a priori*

Sformułowanie rozkładu *a priori* parametrów modelu jest podstawowym wymogiem bayesowskiego wnioskowania statystycznego. Rozkład *a priori* wyraża wiedzę badacza o parametrach modelu statystycznego przed przeprowadzeniem eksperymentu. Gdy wiedza ta jest bardzo uboga, stosowane są tzw. nieinformacyjne (obiektywne) rozkłady *a priori*. Nieinformacyjne rozkłady *a priori* stosowane są często nawet wówczas, gdy istnieją pewne informacje *a priori*, jednak sformułowanie informacyjnego *a priori* byłoby bardzo utrudnione. Ma to miejsce zwłaszcza w sytuacji, gdy model zawiera wiele parametrów i trudno jest przełożyć wiedzę badacza (eksperta) na postać rozkładu prawdopodobieństwa. Należy jednak podkreślić, że stosowanie nieinformacyjnych rozkładów *a priori* może prowadzić do błędnych wniosków (por. np. [14]); również sama definicja nieinformacyjnego rozkładu *a priori* jest dosyć nieprecyzyjna. Obecnie przeważa wręcz pogląd, iż niemożliwe jest zbudowanie nieinformacyjnych (obiektywnych) rozkładów *a priori* (por. [1]). Poszukuje się zatem raczej takich rozkładów *a priori*, które będą miały jak najmniejszy wpływ na otrzymany na ich podstawie rozkład *a posteriori*.

Zdaniem wielu, prawidłowa analiza bayesowska powinna zawsze opierać się na informacyjnych (subiektywnych) rozkładach *a priori* (por. [10]). Mają one także inne zalety. Oto kilka z nich (por. [12]):

- 1) subiektywne rozkłady *a priori* są rozkładami właściwymi,
- 2) pozwalają one na uwzględnienie w analizie dodatkowej wiedzy i doświadczenia badacza,
- 3) niekiedy istnieje zbyt mało informacji, aby rozwiązać problem, korzystając z rozkładów obiektywnych i użycie rozkładu subiektywnego jest jedynym sposobem jego rozwiązania.

Podstawową trudnością w stosowaniu informacyjnych rozkładów *a priori* jest trudność w przetłumaczeniu wiedzy eksperta (zwłaszcza jeśli jest to wiedza subiektywna, niewynikająca z badań z przeszłości) na postać rozkładu prawdopodobieństwa. Wśród sposobów formułowania informacyjnych rozkładów *a priori* można wyróżnić dwie zasadnicze metody: parametryczną oraz nieparametryczną. W metodzie parametrycznej zakłada się określoną klasę rozkładu prawdopodobieństwa dla estymowanego parametru (często będzie to tzw. rozkład sprzężony) i wówczas zadanie polega tylko na określeniu parametrów tego rozkładu. Metoda nieparametryczna formułuje rozkład *a priori* o dowolnej postaci – wybierana jest na tyle bogata klasa rozkładów prawdopodobieństwa, iż aproksymuje ona każdy możliwy rozkład prawdopodobieństwa. Metoda parametryczna jest niewątpliwie prostsza, jednak nie zawsze możliwe jest właściwe przedstawianie posiadanej wiedzy poprzez wąsko określoną klasę rozkładów prawdopodobieństwa. Z kolei metoda nieparametryczna wymaga większego wysiłku, a ponadto otrzymywane rozkłady *a priori* niemal zawsze będą niesprężone i wymagać będą stosowania w dalszej analizie metod numerycznych (co jednak wraz z rozwojem metod MCMC nie jest poważnym ograniczeniem).

3. Aspekty psychologiczne

W przypadku stosowania metod parametrycznych od eksperta wymaga się podania przez niego wartości parametrów rozkładu *a priori*. Parametry te mogą mieć bardzo różne interpretacje, jednak często określają one wartość oczekiwaną, odchylenie standardowe lub odsetek (prawdopodobieństwo sukcesu). Przeprowadzono wiele eksperymentów, które miały na celu zbadanie, na ile trafnie potrafimy opisywać tymi parametrami naszą wiedzę (por. [6]).

Testy dotyczące odsetka przeprowadzano poprzez zaprezentowanie grupie osób tablicy o wymiarach 20 na 20 kwadratów. Każdy z 400 kwadratów miał kolor czerwony lub zielony. Zadanie polegało na odgadnięciu odsetka kwadratów czerwonych, gdy tablica była pokazywana przez krótki okres. Okazało się, że odpowiedzi były dość dokładne. Na ogół różnica pomiędzy oszacowanym a prawdziwym odsetkiem nie przekraczała 0,05.

Testy dotyczące parametrów położenia (wartości oczekiwanej, mediany oraz modalnej) przeprowadzone zostały poprzez pokazywanie grupie osób ciągu liczb, a następnie odgadnięcie wartości wymienionych parametrów. Również w tym przypadku odpowiedzi były stosunkowo dokładne. Jedynie w przypadku, gdy generowany ciąg liczb cechowała silna asymetria, pojawiały się niedokładności w oszacowaniu wartości oczekiwanej – jej oszacowanie było zbliżone do mediany.

Znacznie większe trudności sprawia natomiast szacowanie wariancji (odchylenia standardowego). Można dostrzec tutaj kilka niekorzystnych zjawisk. Po pierwsze ludzie mają tendencję do różnego szacowania wariancji w zależności od war-

tości oczekiwanej – można powiedzieć, że zamiast wariancji szacujemy współczynnik zmienności. Ponadto w sytuacji, gdy typowe są duże odchylenia od wartości oczekiwanej (na przykład w przypadku rozkładów dwumodalnych), mamy tendencję do przeszacowywania wariancji. Z kolei gdy dominują małe odchylenia (jak w przypadku rozkładu normalnego), wówczas wariancja jest często niedoszacowana.

Rozwiązaniem trudności związanych z szacowaniem wariancji jest zadawanie pytań o inne parametry – przede wszystkim kwantyle. Na przykład w przypadku rozkładu normalnego zamiast pytać o wartość oczekiwaną i odchylenie standardowe, można pytać o wartości dwóch kwartyli, co jednoznacznie pozwala wyznaczyć wartość oczekiwaną i odchylenie standardowe.

4. Metody parametryczne

4.1. Metody wyznaczania rozkładu *a priori* dla prawdopodobieństwa sukcesu w rozkładzie dwumianowym

Istnieją cztery podstawowe metody parametrycznego określania rozkładu *a priori* dla parametru rozkładu dwumianowego (por. [6]):

1. Metoda pierwsza polega na określeniu przez eksperta mediany dla prawdopodobieństwa sukcesu i/lub kwantyli. Następnie wybierany jest rozkład beta (który jest rozkładem sprzężonym) o parametrach, dla których jego kwantyle są zbliżone do podanych przez eksperta.

2. Metoda *hipotetycznej przyszłej próbki* (*hypothetical future sample*). Ekspert ocenia najpierw odsetek populacji, który ma rozpatrywaną cechę. Następnie stawiane mu są pytania postaci: „Przypuśćmy, że wylosowano próbkę 50 jednostek z populacji i stwierdzono, że 20 z nich ma interesującą nas cechę. Jakie jest teraz prawdopodobieństwo, że wylosowana jednostka będzie miała tę cechę?”.

3. Metoda *ekwiwalentnej próbki a priori* (*equivalent prior sample*). Metoda przedstawiona zostanie na przykładzie. Załóżmy, że celem jest oszacowanie odsetka studentów płci męskiej w całej populacji studentów Akademii Ekonomicznej we Wrocławiu. Osoba formułująca rozkład *a priori* musi podać dwie liczby r i n takie, iż jej wiedza jest w przybliżeniu równoważna (ekwiwalentna) zaobserwowaniu w losowej próbce n -elementowej r studentów płci męskiej (jeżeli rozkład *a priori* formułuje prowadzący ćwiczenia ze studentami, który pamięta, że w prowadzonej przez niego 30-osobowej grupie było 10 studentów, wówczas może on podać takie właśnie wartości). Rozkładem *a priori* jest wówczas rozkład beta z parametrami r oraz $n - r$.

4. Metoda *funkcji gęstości prawdopodobieństwa*. Metoda wymaga podania przez eksperta najbardziej prawdopodobnej wartości dla prawdopodobieństwa sukcesu oraz np. dwóch wartości, które uważa za dwukrotnie mniej prawdopodobne po jednej i po drugiej stronie wartości najbardziej prawdopodobnej.

4.2. Metody wyznaczania rozkładu *a priori* dla populacji o rozkładzie normalnym

Założmy, że celem jest estymacja parametrów m i σ dla cechy X w populacji o rozkładzie normalnym ($X \sim N(m, \sigma)$). W takim przypadku rozkładem sprzężonym jest tzw. rozkład gamma-normalny (por. [3]). Rozkład ten jest jednak dosyć niewygodny w zastosowaniu, ponieważ uniemożliwia niezależne określenie rozkładów *a priori* oddzielnie dla parametru m i dla parametru σ . W związku z tym wygodniej jest zastosować normalny rozkład *a priori* dla wartości oczekiwanej oraz rozkład gamma dla precyzji r ($r = \sigma^2$). Rozkład taki nie jest rozkładem sprzężonym, w związku z czym do estymacji wymagane jest zastosowanie metod numerycznych. Jest on jednak sprzężony warunkowo, co oznacza, że można analitycznie wyznaczyć rozkłady $m|\sigma^{-2}$ i $\sigma^{-2}|m$, co z kolei bardzo upraszcza implementację algorytmu Gibbsa stosowanego w analizie numerycznej.

Wyznaczenia normalnego rozkładu *a priori* dla parametru m najłatwiej dokonać poprzez podanie dwóch kwantyli (najwygodniej kwartyła 1 i kwartyła 3) – zgodnie z uwagami w punkcie 3. Podobnie należy uczynić dla parametru σ .

Przykład 1

Założmy, że należy zbadać iloraz inteligencji X uczniów w pewnej szkole. Przyjęte zostało założenie, że cecha ma rozkład normalny: $X \sim N(m, \sigma)$. Należy zatem wyznaczyć rozkład *a priori* dla wartości oczekiwanej m oraz precyzji r . Założona zostanie parametryczna postać rozkładu *a priori* o postaci:

$$m \sim N(m_0, \sigma_0),$$

$$\sigma^{-2} \sim \text{Gamma}(\alpha, \beta).$$

Przypuśćmy, że ekspert podał następujące informacje *a priori* o badanej cesze:

1) $p(m < 90) = 0,25$, co oznacza, że ocenia prawdopodobieństwo, że wartość oczekiwana ilorazu inteligencji jest mniejsza niż 90, wynosi 0,25,

2) $p(m < 110) = 0,75$, co oznacza, że ocenia prawdopodobieństwo, że wartość oczekiwana ilorazu inteligencji jest mniejsza niż 110, wynosi 0,75,

3) $p(\sigma < 5) = 0,25$ i $p(\sigma < 12) = 0,75$, co jest równoważne warunkom:
 $p\left(\sigma^{-2} < \frac{1}{\sqrt{5}}\right) = 0,75$ i $p\left(\sigma^{-2} < \frac{1}{\sqrt{12}}\right) = 0,25$.

Korzystając z tablic rozkładu normalnego oraz rozwiązując odpowiedni układ równań, łatwo dowiadujemy się, iż $m_0 = 100$ i $\sigma_0 = 14,82$.

Niech $f(x)$ oznacza gęstość rozkładu gamma o postaci:

$$f(x|\alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{dla } x > 0, \\ 0 & \text{dla } x \leq 0. \end{cases}$$

Wyznaczenie wartości parametrów α i β na podstawie kwantyli wymaga numerycznego rozwiązania następującego układu równań:

$$\begin{cases} \int_0^{1/\sqrt{12}} f(x|\alpha, \beta) dx = 0,25, \\ \int_0^{1/\sqrt{5}} f(x|\alpha, \beta) dx = 0,75. \end{cases}$$

W opisanym przykładzie $\alpha = 0,85027$ i $\beta = 29,43306$.

4.3. Rozkłady wielowymiarowe

Wyznaczanie rozkładów *a priori* jednocześnie dla wielu parametrów jest zadaniem znacznie trudniejszym niż w przypadku pojedynczych parametrów. Opracowane zostały m.in. metody dla przypadków, gdy populacja pochodzi z wielowymiarowego rozkładu normalnego (por. [5]) oraz dla liniowego modelu regresji wielu zmiennych (por. [9]).

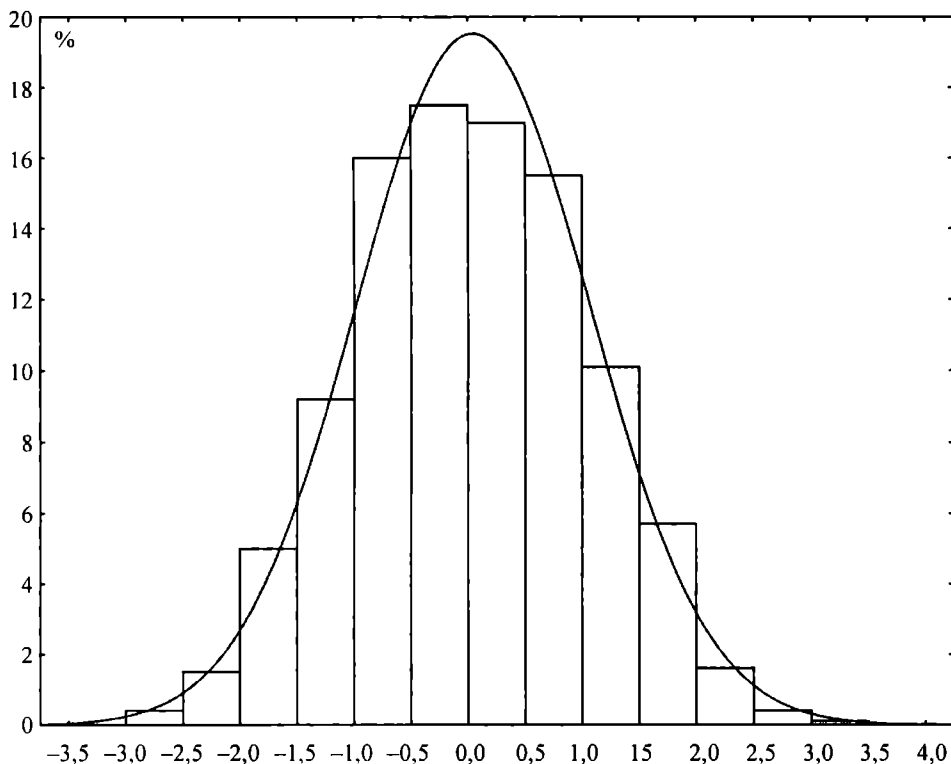
Trudności związane z rozkładami wielowymiarowymi wynikają z konieczności określania macierzy kowariancji rozkładu *a priori*, która opisuje zależności pomiędzy rozpatrywanymi zmiennymi. Oznacza to, że niezbędne jest podanie wartości dla dużej liczby parametrów. Należy również zwracać uwagę, aby wyznaczona macierz kowariancji była dodatnio określona. Mogą również pojawić się trudności w uzyskiwaniu odpowiedzi od ekspertów na nierzadko dosyć zawiłe pytania, jak np. (por. [5]):

- „Przypuśćmy, że zbadano dużą liczbę ptaków i stwierdzono, że ich średnia waga wynosi 170 g, natomiast ich średnia długość równa jest 240 mm. Jakie będzie wówczas twoje oszacowanie mediany ich rozpiętości skrzydeł?” lub
- „Przypuśćmy, że wybrano losowo jednego ptaka. Jeśli jego długość jest większa o 50 mm od przeciętnej, a waga o 30 g większa, o ile twoim zdaniem będzie większa jego rozpiętość skrzydeł? Podaj oszacowanie mediany oraz kwartyla dolnego i górnego”.

5. Metody nieparametryczne

Opracowanych zostało wiele technik nieparametrycznego ustalania rozkładu *a priori*. Jedną z najprostszych jest metoda histogramu. Ekspert ma za zadanie

określenie prawdopodobieństwa, że parametr zawiera się w pewnych z góry określonych przedziałach. W ten sposób otrzymujemy histogram, do którego następnie dopasowywana jest pewna krzywa, która stanowić będzie rozkład *a priori*. Metodę tę zilustrowano na rys. 1.

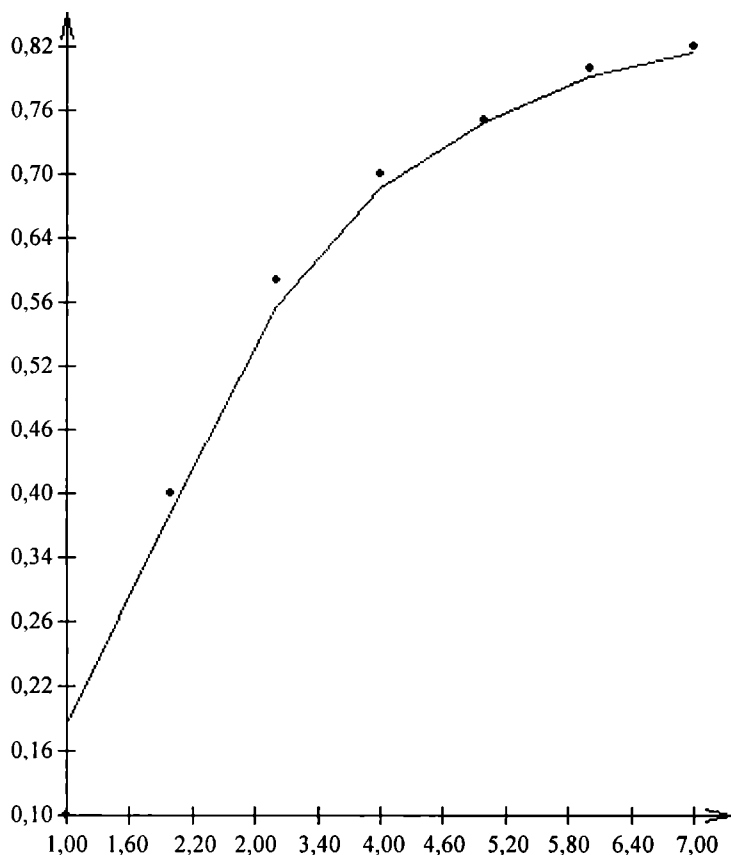


Rys. 1. Nieparametrycznie wyznaczona funkcja gęstości

Źródło: opracowanie własne.

Inna metoda polega na tym, że ekspert podaje pewną liczbę kwantyli rozkładu prawdopodobieństwa, do których następnie w sposób nieparametryczny dopasowywana jest dystrybucja. Zilustrowano to na rys. 2. W przykładzie zastosowano metodę wygładzania jądrowego.

Nieco bardziej skomplikowana metoda została zaproponowana w pracy [11]. Zakłada się w niej, że wiedza eksperta o parametrze może być wyrażona za pomocą funkcji gęstości f , która może mieć dowolną postać. Zakładany jest pewien rozkład *a priori* dla funkcji f , który jest uaktualniany po „zaobserwowaniu danych”, którymi są informacje podane przez eksperta w postaci różnych charakterystyk rozkładu (wartości oczekiwanej, kwantyli itd.). Aby umożliwić praktycznie dowolną



Rys. 2. Nieparametrycznie wyznaczona dystrybuanta

Źródło: opracowanie własne.

postać rozkładu *a priori* parametru, przyjmuje się, że funkcja f może być przedstawiona jako proces gaussowski. Proces ten wyznaczany jest poprzez określenie jego funkcji wartości oczekiwanej $g(x, \theta)$ oraz funkcji wariancji-kowariancji.

Zakłada się, że funkcja wartości oczekiwanej ma rozkład normalny, natomiast funkcja wariancji równa jest $\sigma^2 g(x, \theta)^2$. Z kolei korelacja pomiędzy $f(x)$ a $f(x')$ jest malejącą funkcją odległości pomiędzy punktami x i x' . Parametry procesu gaussowskiego są aktualizowane na podstawie informacji dostarczonych przez eksperta. Będą to najczęściej kwantyle bądź momenty rozkładu.

6. Grupowe wyznaczanie rozkładu *a priori*

W niektórych sytuacjach możemy oczekiwać, iż rozkład *a priori* parametrów wyznaczony zostanie na podstawie wiedzy pewnej grupy osób. Istnieją dwa zasadnicze sposoby grupowego wyznaczania rozkładu *a priori* (por. [8]):

- wyznaczanie rozkładu *a priori* indywidualnie przez każdego eksperta, a następnie wiązanie ich za pomocą odpowiednich procedur,
- wyznaczanie jednego rozkładu *a priori* na zasadzie konsensusu ekspertów.

Pierwsze w wymienionych podejść nazywamy matematycznym, drugie behawioralnym.

Dwie najprostsze metody polegają na arytmetycznym lub geometrycznym uśrednianiu indywidualnych rozkładów *a priori* przy użyciu wag. Jeżeli oznaczymy przez $p_i(\theta)$ rozkład *a priori* parametru θ wyznaczony przez i -tego eksperta oraz przez ω_i wagę przypisaną temu ekspertowi, wówczas łączny rozkład *a priori* wyznacza się jako średnią arytmetyczną rozkładów indywidualnych według wzoru:

$$p(\theta) = \sum_{i=1}^n \omega_i p_i(\theta). \quad (1)$$

Wagi przypisane n ekspertom powinny oczywiście być dodatnie i sumować się do jedności.

Drugi sposób polega na wyznaczeniu średniej geometrycznej rozkładów według wzoru:

$$p(\theta) = k \prod_{i=1}^n p_i(\theta)^{\omega_i}, \quad (2)$$

gdzie k oznacza stałą normującą, która zapewnia, że $p(\theta)$ jest rozkładem prawdopodobieństwa.

Przypisanie wag poszczególnym ekspertom może wynikać z ich kompetencji – w najprostszym wariacie wagi będą jednakowe.

Podejście behawioralne wymaga od ekspertów przeprowadzenia dyskusji, negocjacji i w końcu kompromisu. Więcej informacji o szczegółowych zaleceniach dotyczących tego podejścia można znaleźć w [8].

Literatura

- [1] Berger J.O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 1985.
- [2] Bernardo J.M., *Non-informative Priors Do Not Exist*, „J. Statist. Planning and Inference” 1997, nr 65, 159-189.

-
- [3] Biskup D., *Rozkłady sprzężone w statystyce bayesowskiej*, [w:] *Pomiar statystyczny*, red. W. Ostasiewicz, AE, Wrocław, 2003.
 - [4] DeGroot M.H., *Optymalne decyzje statystyczne*, PWN, Warszawa 1981.
 - [5] Garthwaite P.H., Al-Awadhi S.A., *Non-conjugate Prior Distribution Assessment for Multivariate Normal Sampling*, „Journal of the Royal Statistical Society: Series B” 2001, nr 63.
 - [6] Garthwaite P.H., Kadane J.B., O'Hagan A., *Statistical Methods for Eliciting Probability Distributions*, „Journal of the American Statistical Association” 2005, nr 100, 680-701.
 - [7] Jaynes E.T., *Confidence Intervals vs. Bayesian Intervals. Foundations of Probability Theory. Statistical Inference and Statistical Theories of Science 2*, red. W.L. Harper, C.A. Hooker, Dordrecht: Reidel 1976, 175-257.
 - [8] Jenkinson D., *The Elicitation of Probabilities – A Review of the Statistical Literature*, BEEP Working Paper, 2005.
 - [9] Kadane J.B., Dickey J., Winkler R., Smith W., Peters S., *Interactive Elicitation of Opinion for a Normal Linear Model*, „Journal of the American Statistical Association” 1980, nr 75.
 - [10] Lindley D.V., *The Philosophy of Statistics*, „The Statistician” 2000, nr 49, 293-337.
 - [11] Oakley J.E., O'Hagan A., *Uncertainty in Prior Elicitation: A Nonparametric Approach*, Research Report No. 521/02, University of Sheffield, 2002.
 - [12] Press S.J., *Subjective and Objective Bayesian Statistics*, Wiley 2003.
 - [13] Savage L.J., *The Foundations of Statistics*, Dover Publications, New York 1971.
 - [14] Stone M., *Strong Inconsistency from Uniform Priors*, „Journal of the American Statistical Association” 1976, nr 71, 114-116.

SUBJECTIVE PRIOR ELICITATION – PARAMETRIC AND NON-PARAMETRIC APPROACH

Summary

Prior distribution elicitation is the basic requirement of the Bayesian statistical inference. If there is prior knowledge about a particular problem, then it is necessary to „translate” it into the form of probabilistic distribution. The paper describes the most important ways of prior elicitation basing on the expert's knowledge. The parametric approach was presented, in which it is assumed that the prior distribution can be described using some predetermined class of distributions. The non-parametric approach in turn does not assume any restrictions on the form of the prior. The topic of group prior elicitation was also mentioned.