ЗА ПАРАДИГМИТЕ В СТАТИСТИКАТА – БЕЙСОВСКА СТАТИСТИКА[1]

ГЛ. АС. Д-Р КАЛОЯН ХАРАЛАМПИЕВ[2]

There are two paradigms in contemporary statistics. The first one, which is best known and widely used, is called frequentist. The second one, which is practically unknown (especially in Bulgaria) and it is used only by narrow circle of scientists and researchers (mainly in the rest of the world), is called Bayesian. Major aim of this paper is to introduce Bayesian statistics. The presentation is made in the following order – presenting of basic principles of Bayesian statistics; brief historical review of basic ideas; description of the technique for practical application. In the first part of the paper the Bayes theorem and related probabilities are presented. In the second part the development of Bayesian statistics during the centuries is shown. Special attention is paid on the solving of basic theoretical problems related to obtaining of the required probabilities. In the third part the Bayesian algorithm of inference based on sample information is presented.

Key words: prior probability; posterior probability; entropy; Bayesian inference

Въведение

В съвременната статистика има две парадигми. Едната, която е най-позната и най-използвана, се нарича честотна (frequentist). Другата, която е почти непозната (особено в България) и е използвана само от тесен кръг учени и изследователи (главно по света), се нарича бейсовска (Bayesian). Но дори и там, където бейсовската статистика е позната, тя не се осъзнава като отделна парадигма. Това е валидно в еднаква степен както за представителите на честотната парадигма, така и за представителите на бейсовската статистика.

Представителите на честотната статистика възприемат бейсовската статистика просто като още няколко метода в общия корпус от методи. При това на тези методи в най-добрия случай се гледа като на допълващи, а в най-лошия – като на безсмислена, дори погрешна екзотика.

Представителите на бейсовската статистика осъзнават обособеността и относителната самостоятелност на своята „теория”, но понеже не я възприемат като отделна парадигма, не могат да си обяснят защо остават неразбрани, въпреки големите усилия, които полагат да представят ясно и логично своите идеи. Показателно е, че в текстовете по бейсовска статистика почти не се среща термина „парадигма” (paradigm). Обикновено се говори за „подход” (approach), „метод” (method) или „теория” (theory)[3].

Целта на този доклад е да представи бейсовската статистика. Това ще стане в следната последователност – (1) представяне на основните принципи на бейсовската статистика, (2) кратък исторически преглед на основните идеи, (3) описание на техниката за практическо приложение.

---

[1] Текстът е резултат от изследване, финансирано със средства, отпуснати целево от държавния бюджет на СУ „Св. Климент Охридски” за научни изследвания през 2007 година, договор 033/2007.

[2] *Софийски университет „Св. Кл. Охридски”, Философски факултет, катедра „Социология”, главен асистент*

[3] Единственото изключение, което съм срещнал в литературата, е Loredo (1990). Но и тук има особеност – думата „парадигма” се среща само веднъж и то в резюмето на статията.

1. Бейсовска статистика

Ще започна своето представяне на бейсовската статистика с една опростена до крайност постановка – при всяко изследване разполагаме само с данните от изследването и искаме да проверим някакви хипотези. В такъв случай, вероятността конкретна хипотеза да е вярна и едновременно с това да разполагаме точно с тези данни, с които разполагаме, може да се представи по два начина. Първо, това е вероятността хипотезата да е вярна умножена по вероятността да разполагаме точно с тези данни, при условие че хипотезата е вярна, или второ, вероятността да разполагаме точно с тези данни умножена по вероятността хипотезата да е вярна, при условие че това са данните. Символично това се записва така:

$$P(HD) = P(H).P(D \mid H) = P(D).P(H \mid D)$$

където H е хипотезата, D са данните, а с вертикална черта ($\mid$) се означава „при условие".

От тези вероятности интерес представлява вероятността хипотезата да е вярна в светлината на разполагаемите данни. Тя може да се изрази от горното равенство:

$$P(H \mid D) = \frac{P(H).P(D \mid H)}{P(D)}$$

Полученият резултат е познатата теорема на Бейс, на чието име всъщност е наречена бейсовската статистика.

На практика, освен данните от изследването и проверяваните хипотези, при всяко изследване разполагаме още и с т.нар. априорна информация (prior information). Това може да е теоретично знание за изследвания обект, а може да е информация от някакви минали изследвания на същия обект. Априорната информация се отразява по следния начин в теоремата на Бейс:

$$P(H \mid DI) = \frac{P(H \mid I).P(D \mid HI)}{P(D \mid I)}$$

$P(H \mid I)$ се нарича априорна вероятност (prior probability), $P(D \mid HI)$ се нарича извадково разпределение (sampling distribution), извадкова вероятност (sampling probability) или функция на правдоподобие (likelihood function), $P(D \mid I)$ се нарича пълна вероятност (marginal probability, marginal likelihood или global likelihood), а $P(H \mid DI)$ се нарича апостериорна вероятност (posterior probability).

Априорната вероятност е вероятността хипотезата да е вярна само в светлината на априорната информация за изследвания обект. Извадковото разпределение е вероятността да се получат точно тези данни, които са получени, ако хипотезата е вярна. Пълната вероятност е вероятността да се получат точно тези данни, които са получени, независимо дали хипотезата е вярна или не.

От теоремата на Бейс е видно, че за получаването на апостериорната вероятност е необходимо преди това да се определят априорната вероятност, извадковото разпределение и пълната вероятност. Начините за тяхното определяне са възниквали исторически по различно време, затова ще ги опиша на съответните места в следващия раздел.

2. Малко история

Представителите на бейсовската статистика поставят на първо място в своята история швейцарския математик Якоб Бернули (Jakob Bernoulli, известен и като James Bernoulli) (1654-1705). В своята книга „Ars Conjectandi” (1713) Бернули разисква основни понятия от теорията на вероятностите като „възможно събитие”, „невъзможно събитие”, „сигурно събитие”, „вероятно събитие”. Също така, той развива идеята за извадковото разпределение (макар да не го нарича по този начин) и извежда конкретно извадково разпределение – т.нар. биномно разпределение (binomial distribution). Макар че Бернули не достига до изчисляване на апостериорни вероятности, той все пак формулира един основен принцип за определяне на априорните вероятности[4], който се използва и до днес, а именно принципа на безпристрастността (principle of indifference) (Bernoulli 2005: 28), според който, когато няма никаква априорна информация, априорните вероятности трябват да бъдат равни.

За първи път апостериорни вероятности се появяват в книгата на английския протестантски пастор Томас Бейс (Thomas Bayes) (1702-1761) „Есе относно решаване на проблем в доктрината на шанса” (1763). Бейс не формулира теоремата на Бейс, нито посочва какви априорни вероятности и извадкови разпределения е използвал, а по-скоро предлага начини за изчисляване на някакви вероятности. От текста му обаче личи, че той има предвид именно апостериорните вероятности и търси начини за тяхното изчисляване.

Теоремата на Бейс е формулирана за първи път от френския математик и астроном Пиер-Симон Лаплас (Pierre-Simon Laplace) (1749-1827) (Jaynes 2003: 112; Loredo 1990: 86). Той е и човекът, който разработва бейсовската статистика почти във вида, в който е позната и днес (Jaynes 1986: 5 и 6; Loredo 1990: 83 и 87). Но Лаплас не може да се справи с един основен проблем: докато преди него, по негово време и след това са предложени различни вероятностни разпределения, отнасящи се до различни конкретни ситуации, то определянето на априорните вероятности се оказало изключително трудно. Поради невъзможност за намиране на универсално решение на този проблем, Лаплас широко използва принципа на безпристрастността, въведен от Бернули. Но такъв избор на априорните вероятности е изглеждал равностоен на всички останали и в такъв смисъл – субективен.

Именно обвиненията в субективност са били в основата на последвалите критики срещу Лаплас, започнали в средата на XIX век. Най-ожесточеният критик на Лаплас е бил английският философ логик Джон Вен (John Venn) (1834-1923). Критиката на Вен е била толкова силна, че оказва изключително влияние на неговите съвременници и последователи (Jaynes 2003: 315).

[4] Макар че в забележка е казано, че това е „много стар принцип”, използван още от времето на Кеплер (Bernoulli 2005: 46).

От хората, попаднали под влиянието на Вен, най-значима роля има английският генетик и еволюционен биолог Роналд Фишер (Ronald Fisher) (1890-1962). Всъщност в основата на честотната статистика стоят работите на Фишер, Нейман (Jerzy Neyman) (1894-1981) и Пирсън (Karl Pearson) (1857-1936) (Jaynes 2003: 492). Фишер разработва метод, основан единствено върху извадковото разпределение, наречен метод на максималното правдоподобие (maximum likelihood principle), който напълно отхвърля идеята за априорната информация, а от тук и за априорните вероятности.

Критиките срещу Лаплас и утвърждаването на честотната парадигма довеждат до почти пълен отказ от използването на априорни вероятности и на теоремата на Бейс за решаване на практически проблеми. По времето на Фишер едва ли не единственият негов опонент е английският математик, геофизик и астроном Харолд Джефрис (Harold Jeffreys) (1891-1989). Джефрис възражда идеите на Лаплас като дава бейсовско решение на всички познавателни задачи, формулирани от честотната статистика. Джефрис също така въвежда нов принцип за определяне на априорната вероятност (наречена днес априорна вероятност на Джефрис). Но като цяло Джефрис не предлага обосновано универсално решение на проблема за определянето на априорните вероятности, така че проблемът с евентуалната субективност продължава да стои.

Следващата важна стъпка в развитието на бейсовската статистика е направена от американския физик Ричард Кокс (Richard Cox) (1898-1991) в неговата статия „Вероятност, честота и рационално очакване” (1946).

Преди представянето на работата на Кокс обаче е важно да се направи едно важно уточнение. Едно от фундаменталните различия между честотната и бейсовската статистика е в разбирането на това какво е вероятност. Според честотната статистика вероятността е обективна характеристика на изследвания обект, която се проявява при безкраен брой опити (Loredo 1990: 84). Според бейсовската статистика вероятността е измерител на знанието (state of knowledge) за изследвания обект. В този смисъл, от гледна точка на честотната статистика вероятността е вътрешно присъща характеристика на изследвания обект, а от гледна точка на бейсовската статистика вероятността не характеризира обекта, а знанието за него.

Кокс използва бейсовската дефиниция на вероятността и извежда теоремата на Кокс, която гласи, че ако вероятността отговаря на следните три условия (desiderata):

1) вероятността се описва с реално число;
2) има съответствие между вероятностите и здравия разум (common sense);
3) вероятността е консистентна, т.е. ако една вероятност може да бъде получена по няколко различни начина, то крайният резултат трябва да бъде едно и също число,

то тя удовлетворява две теореми – теоремата за умножение на вероятности (product rule) и теоремата за събиране на вероятности (sum rule) (Jaynes 2003: 17-19 и 24-33; Loredo 1990: 95-98).

Теоремата за умножение на вероятности гласи, че, ако A и B са две събития, вероятността те да се случат едновременно е:

$$P(AB) = P(A).P(B \mid A) = P(B).P(A \mid B)$$

Теоремата за събиране на вероятности гласи, че сумата от вероятността на събитието А и неговото противоположно ($\overline{A}$) е единица:

$$P(A) + P(\overline{A}) = 1$$

Ценността на теоремата на Кокс е, че тя показва, че всяка формула за работа с вероятности или е следствие от теоремите за умножение и събиране на вероятности, или е неконсистентна (Jaynes 1986: 9). Това дава възможност да се изгради цялостна теория, базирана само на тези две теореми и техните следствия.

Най-важното следствие от теоремата за умножение на вероятностите всъщност е самата теорема на Бейс.

Теоремата за събиране на вероятности също има важни следствия, използвани в бейсовската статистика, които образуват своеобразна верига:

Първо следствие: ако А и В са две събития, вероятността да се случи едното или другото е:

$$P(A + B) = P(A) + P(B) - P(AB)$$

Второ следствие: ако А и В са две събития, които не могат да се случат едновременно, вероятността да се случи едното или другото е:

$$P(A + B) = P(A) + P(B)$$

Трето следствие: ако A_k са няколко събития, които не могат да се случат едновременно, вероятността да се случи някое от тях е:

$$P\left(\sum_k A_k\right) = \sum_k P(A_k)$$

Чрез третото следствие се намира пълната вероятност, тъй като проверяваната хипотеза никога не е само една[5] и всички хипотези образуват пълна група. Пълна група означава, че хипотезите не се припокриват взаимно и че взети заедно покриват всички възможности (mutually exclusive and exhaustive).

Тогава:

$$\sum_k P(H_k \mid I) = 1$$

$$\sum_k P(H_k \mid DI) = 1$$

$$\sum_k P(H_k \mid I).P(D \mid H_k I) = \sum_k P(DH_k \mid I) = P\left(D \sum_k H_k \mid I\right) = P(D \mid I)$$

Полученият резултат показва, че пълната вероятност всъщност е сума от произведенията на извадковото разпределение и априорната вероятност. Това означава, че теоремата на Бейс може да се запише по следния начин:

[5] Най-малкият брой хипотези е две – изследваният феномен да е налице или да липсва.

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

Както е видно от формулата, за получаването на апостериорната вероятност е необходимо да са известни единствено априорната вероятност и извадковото разпределение.

Вече посочих, че има различни извадкови разпределения, отнасящи се до различни конкретни ситуации. Но по времето, когато Кокс е публикувал своята теорема, все още не е имало универсално решение за определяне на априорната вероятност. Това решение е предложено от американския електроинженер и математик Клод Шенън[6] (Claude Shannon) (1916-2001). В неговата статия „Математическа теория на комуникациите" (1948) Шенън предлага формула за измерване на ентропията (entropy):

$$-\sum_k P(H_k \mid I).\log P(H_k \mid I)$$

Ентропията на Шанън стои в основата на метода на максималната ентропия, според който във всяка конкретна ситуация априорните вероятности трябва да се изберат така, че ентропията да е максимална. По този начин, първо, априорната информация се оползотворява напълно, и второ, се гарантира, че изследователят не приписва на изследвания обект свойства, каквито той може и да не притежава (Jaynes 1986: 10).

Методът на максималната ентропия решава последния, останал нерешен фундаментален проблем в бейсовската статистика – определянето на априорните вероятности – и след неговото решаване бейсовската парадигма може да бъде окончателно оформена. Това е направено от американския физик Едуин Джейнс (Edwin Jaynes) (1922-1998). Той обобщава приносите на Бейс, Лаплас, Джефрис, Кокс и Шанън, прехвърля мостове към логиката, комуникационната теория и теорията за вземане на решения (decision theory) и разработва общата теория на бейсовската статистика. Също така, Джейнс добавя две нови условия за консистентност в теоремата на Кокс:

3.2) деидеологизация: при всяко изследване трябва да се използва цялата налична информация;

3.3) консистентност по Джейнс: ако двама изследователи разполагат с една и съща априорна информация и едни и същи данни, те трябва да получат едни и същи апостериорни вероятности. Това всъщност е бейсовската дефиниция за обективност на вероятностите.

С работите на Джейнс приключва теоретичното разработване на бейсовската парадигма. Авторите, пишещи след него, имат приноси основно в техническите детайли (например избор и обосноваване на различни извадкови разпределения и априорни вероятности), но основните принципи, инструменти и техники се възприемат като даденост и директно се използват.

[6] Тук е важно да се отбележи, че Шенън изобщо не се е занимавал с проблема за определянето на априорните вероятности. Той работи в съвсем друга област, а неговото откритие е интегрирано в бейсовската статистика по-късно от Джейнс.

3. Практическо приложение на бейсовската статистика

Единственият дял от статистиката, където има различие между честотната и бейсовската статистика, са статистическите изводи, основани на информация от извадки (inference). Задачата тук е да се направи извод (доверителен интервал или проверка на хипотези) за стойността на някакъв параметър на генералната съвкупност (Θ), разполагайки с данни от извадка. Бейсовската статистика решава този проблем по следния начин:

1) Дефиниране на хипотезите H_k:

$$H_k : \Theta = x_k \ , \ x_1 \le x_2 \le x_3 \le \ldots$$

2) Определяне на априорните вероятности $P(\Theta = x_k \mid I)$. Това става с помощта на метода на максималната ентропия.

3) Определяне на извадковото разпределение. Тук по-скоро става въпрос не за определяне, а за избор на някакво теоретично вероятностно разпределение. Този избор е съобразен с конкретния оценяван параметър и с типа на подбора. Така например, ако се оценява относителен дял от извадка, получена чрез безвъзвратен подбор, извадковото разпределение ще е хипергеометрично, а ако подборът е възвратен, тогава извадковото разпределение ще е биномно (полиномно).

4) Изчисляване на пълната вероятност. Това става чрез третото следствие на теоремата за събиране на вероятности.

5) Изчисляване на апостериорните вероятности $P(\Theta = x_k \mid DI)$. Това става чрез теоремата на Бейс. По този начин на практика се получава плътността на разпределението (probability density function)[7]:

$$f(x_k) = P(\Theta = x_k \mid DI)$$

6) От плътността на разпределение се получава функцията на разпределение (cumulative probability density function) $F(x_k)$:

$$F(x_k) = P(\Theta \le x_k \mid DI)$$

7) Чрез функцията на разпределение се проверяват статистически хипотези и се построяват доверителни интервали.

$$P(\Theta \le a \mid DI) = F(a)$$

$$P(\Theta > b \mid DI) = 1 - F(b)$$

$$P(b < \Theta \le a \mid DI) = F(a) - F(b)$$

Литература:

1. Bayes, T. 1958. Essay towards Solving a Problem in the Doctrine of Chances. In: Biometrika, 45.

2. Bernoulli, J. 2005. On the Law of Large Numbers. NG Verlag, Berlin.

[7] Впрочем и при определянето на априорните вероятности се получава плътността на разпределението, но априорните вероятности имат самостоятелно приложение, само когато по някаква причина не може да се определи извадковото разпределение.

3. Jaynes, E. 1986. Bayesian Methods: General Background. In: Maximum-Entropy and Bayesian Methods in Applied Statistics, J. H. Justice (ed.), Cambridge Univercity Press, Cambridge. Статията е достъпна в Интернет на адрес: http://bayes.wustl.edu/etj/articles/general.background.pdf

4. Jaynes, E. 2003. Probability Theory: the Logic of Science. Cambridge University Press, Cambridge. Непълно издание на тази книга е достъпно в Интернет на адрес: http://omega.albany.edu:8008/JaynesBook.html

5. Jeffreys, H. 1948. Theory of Probability (Second edition). Oxford University Press, London.

6. Loredo, T. 1990. From Laplace to SN 1987A: Bayesian Inference in Astrophysics. In: Maximum Entropy and Bayesian Methods, P. F. Fougere (ed), Kluwer Academic Publishers, Dordrecht. Статията е достъпна в Интернет на адрес: http://bayes.wustl.edu/gregory/articles.pdf

7. Shannon, C. 1948. A Mathematical Theory of Communication. In: The Bell System Technical Journal, 27. Статията е достъпна в Интернет на адрес: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf