Бейсовско оценяване на относителни дялове
(В случая на електоралните изследвания)

Калоян Харалампиев

Въведение

Тази статия има две главни цели. Първата е да представи на учените и изследователите, работещи в областта на социалните науки, бейсовската парадигма в статистиката. Втората е да покаже как се решава конкретен проблем в тази област. Проблемът е как да се оценяват относителни дялове (проценти). Решението на този проблем е показано в два случая – общ и частен, отнасящ се до относителните дялове на хората, които биха избрали дадена партия, но само сред действителните гласоподаватели.

1. Бейсовски подход

Централно място в бейсовската парадигма в статистиката заема теоремата на Бейс:

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

където H_k "е някаква хипотеза, чиято истинност искаме да проверим, D са данните, а I е всяка "априорна информация", с която разполагаме в допълнение към данните" (Jaynes 1988: 25).

$P(D \mid H_k I)$ се нарича *извадково разпределение* и представлява вероятността да получим точно тези данни, които сме получили, ако хипотезата H_k е вярна.

$P(H_k \mid I)$ се нарича *априорна вероятност* на H_k и "специфицира експертното[1] знание за H_k преди експериментът, проектиран за осигуряване на данните D, да е проведен" (Dose 2002: 1).

$P(D \mid I)$ се нарича *пълна вероятност* и представлява вероятността да получим точно тези данни, които сме получи, ако някоя от всичките хипотези H_i

[1] Или теоретичното.

е вярна, макар да не знаем точно коя. Пълната вероятност се получава чрез *маргинализация*:

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

Когато броят на хипотезите е безкрайно голям и те формират континуум, сумирането трябва да се замести с интегриране както следва:

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

където R е областта, в която всички хипотези H_i са дефинирани.

Резултатът получен по формула (1) - $P(H_k \mid DI)$ - се нарича *апостериорна вероятност* на H_k.

Както е видно от формула (1), в апостериорната вероятност са комбинирани два типа информация: теоретична, представена от априорната вероятност и емпирична, представена от извадковото разпределение.

Апостериорната вероятност е основен инструмент за статистическо оценяване. Чрез нея могат да се проверяват хипотези и да се построяват доверителни интервали.

2. Бейсовско оценяване на относителни дялове

Разглеждаме качествен признак с m възможни значения. В частност това може да бъде въпрос с m възможни, предварително дефинирани отговора. Излъчваме извадка от изучаваната генерална съвкупност. В резултат получаваме извадковата честота (броя) на всяко значение на признака. Нека означим този брой с f_i. Следователно данните се състоят от наблюдаваните честоти:

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

Относителният дял на всяко значение на признака в генералната съвкупност е означено с π_i. Тогава проверяваната хипотеза е, че относителният дял π_i ще получи точно стойността π_{ik}:

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

Тъй като всички относителни дялове са неотрицателни и тяхната сума е единица (или 100%) π_{ik} трябва да удовлетворява следните условия:

(4) $$\left|\begin{array}{l}\sum_{i=1}^{m}\pi_{ik}=1\\ \pi_{ik}\geq 0,\ i=1,2,...,m\end{array}\right.$$

Когато извадката е представителна и генералната съвкупност е достатъчно голяма[2], относителните дялове са непрекъснати и извадковото разпределение е *полиномно* (Jaynes 1993: 315, 317-318):

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

където n е обемът на извадката, а $f_i!$ се нарича f_i факториел, което означава $f_i!=1.2.3...f_i$.

Сега за изчисляването на апостериорната вероятност се нуждаем от стойностите на априорната вероятност.

Както отбелязва Jaynes: "В 'истинския живот' обикновено имаме отлична база, основаваща се на минал опит и теоретичен анализ" (Jaynes 1976: 190). Бих искал да коригирам това твърдение, променяйки само една дума – в 'истинския живот' обикновено имаме отлична база, основаваща се на минал опит *или* теоретичен анализ. Обаче така първоначалният проблем се разделя на два нови проблема. Първият е как да получим априорната вероятност, основана на миналия опит. Вторият е как да получим априорната вероятност, основана на теоретичен анализ.

Решението на втория проблем е типичен случай на приложение на *Метода на максималната ентропия*. Като резултат от теоретичния анализ получаваме някакви ограничения относно априорното разпределение. След това трябва да максимизираме ентропията на Shannon $\left(-\sum_k P(H_k\mid I)\log P(H_k\mid I)\right.$ или $\left.-\int_R P(H_k\mid I)\log P(H_k\mid I)\right)$, съобразявайки се с тези ограничения. Така получаваме априорно вероятностно разпределение, което е "толкова неинформативно, колкото е възможно, за да се предпазим от „виждане" в данните на неща, които не съществуват" (Bretthorst 1988: 14).

[2] Обикновено в социалните изследвания и двете условия са изпълнени.

Решението на първия проблем е много по-лесно. “Просто използваме апостериорната вероятност, получена при анализирането на минали данни като априорна вероятност в конкретното изследване” (Bretthorst 1990: 11).

Да се върнем обаче към самото начало, когато няма нито минал опит, нито теоретичен анализ. В съответствие с метода на максималната ентропия, ако нямаме никаква априорна информация, трябва да припишем равни априорни вероятности (Bretthorst 1990: 4-5):

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

Така вече определихме и извадковото разпределение (формула (5)), и априорната вероятност и можем да изчислим пълната вероятност:

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

(6)

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

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

Следователно апостериорната вероятност е:

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

Да отбележим, че когато няма никаква априорна информация, априорните вероятности в числителя и в знаменателя се съкращават. Така само данните остават да имат значение.

Формула (7) се отнася до всички относителни дялове. Когато обаче трябва да оценим точно относителния дял π_i, останалите относителни дялове π_j ($j \neq i$) са *неудобни параметри*, “т.е. параметри, които физически са представени във разглеждания феномен и не могат безопасно да бъдат пренебрегнати в модела, въпреки че не се интересуваме от тяхното оценяване” (Jaynes 1993: 2101). Обаче “в бейсовския метод неудобните параметри причиняват много дребни затруднения – всички параметри, които не ни интересуват, се отстраняват чрез интегриране, съобразявайки се с тяхната априорна вероятност” (Jaynes 1993: 2101). Това също е маргинализация:

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

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

Полученият резултат[3] се нарича *бета разпределение*. Чрез него можем да изчислим следните вероятности:

- Вероятността относителният дял π_i да бъде по-малък от дадено число *a*:

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

$F(a)$ се нарича *функция на разпределение*.

Можем още да изчислим вероятностите:

- Вероятността относителният дял π_i да бъде по-голям от дадено число *b*:

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

- Вероятността относителният дял π_i да се намира в интервала между две дадени числа *a* и *b*:

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

Формули (9) и (10) се използват за проверка на хипотези. Формула (11) се използва за построяване на доверителни интервали.

И така, нека приложим този метод за решаването на реален проблем.

[3] Този резултат е идентичен с формула (17-5) на Jaynes (Jaynes 1974: 17-4).

3. Електорални изследвания

3.1. Пример за бейсовско оценяване

В таблица 1 са представени данни от едно електорално изследване:

Таблица 1

Намерения за гласуване, юни 2005

Намерения за гласуване	Относителен дял в извадката (%)	Честота (брой)
Коалиция за България	27,5	277
НДСВ	14,6	147
Коалиция ОДС-Демократическа партия-Гергьовден	7,4	75
ДПС	5,2	52
Коалиция БНС	3,6	36
ДСБ	3,4	34
Други	7,0	71
Не съм решил	10,0	101
Няма да гласувам	21,3	214
Общо	100,0	1007

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

Прилагайки формула (8) към данните в таблица 1 и след това формула (11) с $F(a) = 0{,}025$ и $F(b) = 0{,}975$ ще получим следните резултати[4]:

Таблица 2

Доверителни интервали

Намерения за гласуване	Доверителни интервали
Коалиция за България	$P(0{,}247 < \pi_1 < 0{,}301 \mid DI) = 0{,}95$
НДСВ	$P(0{,}125 < \pi_2 < 0{,}168 \mid DI) = 0{,}95$
Коалиция ОДС-Демократическа партия-Гергьовден	$P(0{,}060 < \pi_3 < 0{,}091 \mid DI) = 0{,}95$
ДПС	$P(0{,}039 < \pi_4 < 0{,}066 \mid DI) = 0{,}95$
Коалиция БНС	$P(0{,}026 < \pi_5 < 0{,}048 \mid DI) = 0{,}95$
ДСБ	$P(0{,}025 < \pi_6 < 0{,}046 \mid DI) = 0{,}95$
Други	$P(0{,}056 < \pi_7 < 0{,}087 \mid DI) = 0{,}95$
Не съм решил	$P(0{,}083 < \pi_8 < 0{,}119 \mid DI) = 0{,}95$
Няма да гласувам	$P(0{,}187 < \pi_9 < 0{,}237 \mid DI) = 0{,}95$

Тези резултати са тривиални. Това е така, защото съществува и друга парадигма в статистиката, която е по-позната и широко използвана. Тази парадигма се нарича честотна или ортодоксална. Когато се оценяват относителни

[4] Всички интеграли са изчислени числово.

дялове без никаква априорна информация, резултатите, получени чрез двете парадигми, са практически еднакви.

3.2. Специфичен проблем – относителни дялове само сред действителните гласоподаватели

За нещастие, практическата полза от резултатите в таблица 2 е малка, защото в електоралните изследвания най-важните относителни дялове са относителните дялове само сред действителните гласоподаватели.

Тук възникват два нови проблема. Първият е как да постъпим с броя на хората, които още не са решили (как или за кого да гласуват). Вторият е как да постъпим с броя на хората, които няма да гласуват.

Тъй като не знаем за коя партия ще гласуват тези хора, които още не са решили, тази информация е еквивалентна на информацията, която имаме за останалите хора в цялата генерална съвкупност, които не са попаднали в извадката. Това означава че можем да игнорираме информацията за броя на хората, които още не са решили и по този начин да редуцираме извадката[5].

Обаче информацията за броя на хората които няма да гласуват е решаваща. За да разберем нейната важност, нека означим обема на цялата генерална съвкупност с *N*, честотите в генералната съвкупност с f_i^* и броя на хората в генералната съвкупност, които няма да гласуват с f_m^*. Тогава броят на действителните гласоподаватели е $N - f_m^*$ и относителният дял на хората, които са избрали дадена партия, само сред действителните гласоподаватели е:

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

тъй като

$$\pi_i = \frac{f_i^*}{N} \quad (13)$$

Както се вижда, относителният дял на хората, които няма да гласуват участва в знаменателя на формула (12). Това означава, че и числителят и

[5] Това е обичайна практика на социологическите агенции в България. Когато се представят резултатите, отнасящи се за действителните гласоподаватели, броят на хората, които още не са решили, се игнорира. В този случай няма противоречие между теорията и текущата практика.

знаменателят са неизвестни параметри[6]. Обаче главният проблем при оценяването на относителния дял на хората, които биха избрали дадена партия – само сред действителните гласоподаватели – е, че π_i и π_m не са независими[7].

Първата стъпка от решаването на горепосочения проблем е изчисляването на съвместната вероятност на π_i и π_m. Отново начинът да се направи това е маргинализация:

$$P[(\pi_i = x)(\pi_m = y) \mid DI] = \int\int_{R_3} \ldots \int P(H_k \mid DI) d\pi_{1k} d\pi_{2k} \ldots d\pi_{i-1,k} d\pi_{i+1,k} \ldots d\pi_{m-1,k} =$$

$$= \ldots = \frac{(n+m-1)!}{f_i! f_m! (n - f_i - f_m + m - 3)!} x^{f_i} y^{f_m} (1-x-y)^{n-f_i-f_m+m-3} \tag{14}$$

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

След това най-лесният начин да се получи апостериорната вероятност е първо да се изчисли функцията на разпределение $F(u) = P\left(\frac{x}{1-y} < u \middle| DI \right)$ и след това да се изчисли *плътността на разпределение* $P\left(\frac{x}{1-y} = u \middle| DI \right) = f(u) = F'(u)$:

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

$$= \frac{(n - f_m + m - 2)!}{f_i! (n - f_i - f_m + m - 3)!} \sum_{j=0}^{n-f_i-f_m+m-3} \frac{C^j_{n-f_i-f_m+m-3} (-1)^{n-f_i-f_m+m-3-j} u^{n-f_m+m-2-j}}{n - f_m + m - 2 - j} \tag{15}$$

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

$$= \ldots = \frac{(n - f_m + m - 2)!}{f_i! (n - f_i - f_m + m - 3)!} u^{f_i} (1-u)^{n-f_i-f_m+m-3} \tag{16}$$

Полученият резултат е още едно бета разпределение. Чрез него можем да проверим хипотезата, че дадена партия ще премине четирипроцентовата бариера[8]. Също така можем да построим доверителните интервали на

[6] Обичайната практика на социологическите агенции в България е да се игнорира броят на хората, които няма да гласуват. Така извадката се редуцира отново. Това обаче е еквивалентно на разглеждането на знаменателя като известен параметър и оставянето само на числителя като неизвестен параметър. В този случай има противоречие между теорията и текущата практика, макар че честотната теория мълчи по проблема.

[7] Това е следствие от формула (4).

[8] Или хипотезата, че даден кандидат-президент ще спечели президентските избори.

относителните дялове на хората, които биха избрали дадена партия[9], само сред действителните гласоподаватели. Тези резултати са представени в таблица 3:

Таблица 3

Вероятност за преминаване на четирипроцентовата бариера и доверителни интервали на относителните дялове на хората, които биха избрали дадена партия, само сред действителните гласоподаватели

Намерения за гласуване	Вероятност за преминаване на четирипроцентовата бариера (%)	Доверителни интервали
Коалиция за България	100,0	$P\left(0{,}362 < \frac{\pi_1}{1-\pi_m} < 0{,}434 \middle\| DI\right) = 0{,}95$
НДСВ	100,0	$P\left(0{,}183 < \frac{\pi_2}{1-\pi_m} < 0{,}242 \middle\| DI\right) = 0{,}95$
Коалиция ОДС-Демократическа партия-Гергьовден	100,0	$P\left(0{,}087 < \frac{\pi_3}{1-\pi_m} < 0{,}132 \middle\| DI\right) = 0{,}95$
ДПС	100,0	$P\left(0{,}058 < \frac{\pi_4}{1-\pi_m} < 0{,}096 \middle\| DI\right) = 0{,}95$
Коалиция БНС	95,4	$P\left(0{,}038 < \frac{\pi_5}{1-\pi_m} < 0{,}070 \middle\| DI\right) = 0{,}95$
ДСБ	90,8	$P\left(0{,}036 < \frac{\pi_6}{1-\pi_m} < 0{,}067 \middle\| DI\right) = 0{,}95$
Други		$P\left(0{,}082 < \frac{\pi_7}{1-\pi_m} < 0{,}126 \middle\| DI\right) = 0{,}95$

Тези резултати не са тривиални. Те не са напълно невъзможни от гледна точка на честотната статистика, обаче е твърде трудно да бъдат получени чрез честотни методи.

4. Два важни допълнителни момента

4.1. Същественият

Във втора точка разгледахме ситуацията на липса на априорна информация. Нека сега разгледаме ситуацията на налична априорна информация, получена от минали изследвания. Тогава можем да използваме данните от всички

[9] Или даден кандидат за президент.

минали изследвания като априорна информация, а данните от конкретното изследване като данни.

Да предположим, че няма никаква априорна информация преди първото изследване. Следователно получената апостериорната вероятност е:

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

След това провеждаме второто изследване. Можем да използваме апостериорната вероятност от първото изследване като априорна вероятност във второто. Следователно новата апостериорна вероятност е:

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

Ако извадките на двете изследвания са независими[10], извадковото разпределение е:

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

Маргинализирайки числителя на формула (18) получаваме:

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

където

$$(21) \qquad g_{i,2} = f_{i,1} + f_{i,2}$$

Така апостериорната вероятност е:

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

Формула (22) лесно се обобщава за *t* изследвания:

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

където

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

[10] Обикновено такъв е случаят в социалните изследвания.

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

Обаче формули (23) и (7) са практически еднакви. Това означава, че използването на данните от всички минали изследвания като априорна информация и данните от конкретното изследване като данни е еквивалентно на използването на данните от всички изследвания като данни без никаква априорна информация.

Означава ли това обаче, че можем просто да обединим резултатите от всички изследвания? Или може би това крие някакви възможни опасности? За да отговорим на този въпрос, нека да извършим някои преобразувания:

(26) $$p_{i,j} = \frac{f_{i,j}}{n_j}$$ е извадковият относителен дял

(27) $$\overline{p}_i = \frac{\sum_{j=1}^{t} p_{i,j} n_j}{\sum_{j=1}^{t} n_j} = \frac{\sum_{j=1}^{t} f_{i,j}}{n} = \frac{g_{i,t}}{n}$$ е средният извадков относителен дял

Следователно:

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

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

Формула (29) показва, че индивидуалните стойности на извадковите относителни дялове нямат значение, но тяхната средна има. И така, ако относителните дялове са стабилни във времето, тогава всяко ново парченце информация ще подобри оценката. Обаче, ако относителните дялове не са стабилни във времето или ако има тренд, тогава е много опасно всички данни да се обединяват. В този случай е за предпочитане да се използват само данните от конкретното изследване с равни априорни вероятности.

Нека да разгледаме един екстремен пример. В таблица 4 са представени данни от три електорални изследвания:

Таблица 4

Намерения за гласуване (брой)

Намерения за гласуване	Март 2005	Май 2005	Юни 2005	Общо
⋮	⋮	⋮	⋮	⋮
Не съм решил	287	216	101	604
⋮	⋮	⋮	⋮	⋮
Общо	1213	1000	1007	3220

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

Ако оценим относителния дял на хората, които още не са решили, използвайки само данните от юни 2005 година, ще получим:

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

Ако обаче обединим всички данни и тогава оценим същия относителен дял, ще получим:

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

Получената разлика е съществена. Тя се дължи на промените в намеренията за гласуване през периода от март до юни 2005 година. Очевидно по-коректният резултат е този, получен от данните от юни 2005 година без никаква априорна информация.

4.2. Техническият

Когато извадката е достатъчно голяма, някои изчислителни трудности могат да бъдат избегнати чрез граничен преход:

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

където

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

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

и

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

където

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

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

Формули (30) и (33) представляват така нареченото *гаусово разпределение*[11]. Чрез него можем директно да построим доверителните интервали и да изчислим вероятността дадена партия да премине четирипроцентовата бариера:

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

където z_i са стойностите на *стандартизираното гаусово разпределение*.

Построените доверителни интервали са представени в таблици 5 и 6.

Таблици 5

Доверителни интервали на относителните дялове

Намерения за гласуване	Доверителни интервали
Коалиция за България	$P(0{,}246 < \pi_1 < 0{,}301 \mid DI) = 0{,}95$
НДСВ	$P(0{,}123 < \pi_2 < 0{,}167 \mid DI) = 0{,}95$
Коалиция ОДС-Демократическа партия-Гергьовден	$P(0{,}058 < \pi_3 < 0{,}090 \mid DI) = 0{,}95$
ДПС	$P(0{,}038 < \pi_4 < 0{,}065 \mid DI) = 0{,}95$
Коалиция БНС	$P(0{,}024 < \pi_5 < 0{,}047 \mid DI) = 0{,}95$
ДСБ	$P(0{,}022 < \pi_6 < 0{,}045 \mid DI) = 0{,}95$
Други	$P(0{,}054 < \pi_7 < 0{,}086 \mid DI) = 0{,}95$
Не съм решил	$P(0{,}081 < \pi_8 < 0{,}118 \mid DI) = 0{,}95$
Няма да гласувам	$P(0{,}186 < \pi_9 < 0{,}236 \mid DI) = 0{,}95$

[11] Още се нарича *нормално разпределение*.

Таблица 6

Вероятност за преминаване на четирипроцентовата бариера и доверителни интервали на относителните дялове на хората, които биха избрали дадена партия, само сред действителните гласоподаватели

Намерения за гласуване	Вероятност за преминаване на четирипроцентовата бариера (%)	Доверителни интервали
Коалиция за България	100,0	$P\left(0{,}361 < \frac{\pi_1}{1-\pi_m} < 0{,}434 \middle\| DI\right) = 0{,}95$
НДСВ	100,0	$P\left(0{,}181 < \frac{\pi_2}{1-\pi_m} < 0{,}241 \middle\| DI\right) = 0{,}95$
Коалиция ОДС-Демократическа партия-Гергьовден	100,0	$P\left(0{,}085 < \frac{\pi_3}{1-\pi_m} < 0{,}131 \middle\| DI\right) = 0{,}95$
ДПС	100,0	$P\left(0{,}055 < \frac{\pi_4}{1-\pi_m} < 0{,}094 \middle\| DI\right) = 0{,}95$
Коалиция БНС	93,4	$P\left(0{,}035 < \frac{\pi_5}{1-\pi_m} < 0{,}068 \middle\| DI\right) = 0{,}95$
ДСБ	88,5	$P\left(0{,}033 < \frac{\pi_6}{1-\pi_m} < 0{,}065 \middle\| DI\right) = 0{,}95$
Други		$P\left(0{,}079 < \frac{\pi_7}{1-\pi_m} < 0{,}124 \middle\| DI\right) = 0{,}95$

Сравнявайки таблица 5 с таблица 2 и таблица 6 с таблица 3 се вижда, че различията между точните и приблизителните резултати са пренебрежимо малки. Тази загуба на точност е приемлива, защото изчислителните усилия са значително по-малки, когато се използва гаусовото приближение.

Изводи:

1. Когато оценяваме относителни дялове:

1.1. Ако няма никаква априорна информация, тогава честотният и бейсовският подход са еквивалентни. Във всички останали случай бейсовският подход дава възможност да се вземе предвид всяка априорна информация, с която разполагаме.

1.2. Ако априорната информация се състои от данни от минали изследвания, тогава могат да съществуват две различни възможности:

а) Няма съществени изменения на относителните дялове във времето. Тогава можем да обединим всички данни и да ги използваме за оценяване на относителните дялове.

б) Има съществени изменения на относителните дялове във времето. Тогава е опасно да се обединяват всички данни. По-безопасният път е да се използват само данните от конкретното изследване с равни априорни вероятности.

2. Бейсовският подход лесно работи с два или повече параметъра, които не са независими. В електоралните изследвания това позволява да се оценяват относителните дялове само сред действителните гласоподаватели. Този проблем е нерешим (или най-малкото решението е много трудно) от честотна гледна точка. Това е "истинският тест" на бейсовския подход в съответствие с казаното от Jaynes: "истинският тест на всеки нов принцип в науката не е неговата способност да получи отново познати резултати, а неговата способност да даде нови резултати, които не биха могли да се получат (или най-малкото не са получени) без него" (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