СОФИЙСКИ УНИВЕРСИТЕТ „СВЕТИ КЛИМЕНТ ОХРИДСКИ”
ФИЛОСОФСКИ ФАКУЛТЕТ, КАТЕДРА „СОЦИОЛОГИЯ”

УЧЕБНО ПОМАГАЛО

ПО

„ПРИЛОЖЕНИЕ НА БЕЙСОВСКАТА СТАТИСТИКА В СОЦИАЛНИТЕ НАУКИ”

ЛЕКЦИИ: ДОЦ. Д-Р КАЛОЯН ХАРАЛАМПИЕВ
УПРАЖНЕНИЯ: ДОЦ. Д-Р КАЛОЯН ХАРАЛАМПИЕВ

Подготовката и издаването на това учебно помагало е подпомогнато и финансирано от High Education Support Program, Open Society Institute, Budapest

СОФИЯ
2007

СЪДЪРЖАНИЕ

АНОТАЦИЯ НА КУРСА ... 2
ТЕМИ ... 3
ЛИТЕРАТУРА ... 5
АНОТАЦИИ НА ТЕМИТЕ И ТЕКСТОВЕ ... 7
ПЪРВИ РАЗДЕЛ. ВЪВЕДЕНИЕ В БЕЙСОВСКАТА СТАТИСТИКА ... 7
ТЕКСТ КЪМ ПЪРВИ РАЗДЕЛ ... 8
За парадигмите в статистиката – бейсовска статистика ... 8
ВТОРИ РАЗДЕЛ. ИЗВОДИ, ОСНОВАВАЩИ СЕ НА НЕПРЕДСТАВИТЕЛНИ ИЗВАДКИ ... 14
ТЕКСТОВЕ КЪМ ВТОРИ РАЗДЕЛ ... 16
Използване на бейсовска статистика за статистически изводи при непредставителни извадки (през примера на изследването на отпадащи студенти в бакалавърска степен на обучение във Философския факултет на Софийския университет "Св. Климент Охридски") ... 16
Студентската оценка на преподаването – проблемът за точността на изводите ... 25
Анкетите в Интернет: възможност за статистически изводи и интерпретиране на резултатите ... 33
Телевизионните гласувания по телефона – проблемът за „победителя" ... 39
Още една гледна точка към проблема за отказите при социологически изследвания ... 42
ТРЕТИ РАЗДЕЛ. ИЗВОДИ, ОСНОВАВАЩИ СЕ НА ПРЕДСТАВИТЕЛНИ ИЗВАДКИ ... 47
Текст към трети раздел ... 49
Бейсовско оценяване на относителни дялове (В случая на електоралните изследвания) ... 49
ЧЕТВЪРТИ РАЗДЕЛ. ВЗЕМАНЕ НА РЕШЕНИЯ В УСЛОВИЯ НА РИСК ... 60
Текст към четвърти раздел ... 62
Лапласово правило за приемственост – интерпретации и приложения ... 62

АНОТАЦИЯ НА КУРСА

Изследователските проблеми, изискващи приложението на статистически методи, могат да се обособят в три големи групи:

В първата група попадат тези проблеми, които имат теоретично решение. Наистина, като правило, теоретичните решения на един и същ проблем са повече от едно и всяко от тях се основава на различна логика и на различни предположения. Това от своя страна поставя изискването да се познават всички решения и във всяка конкретна ситуация да се прави избор.

Във втората група попадат тези проблеми, които нямат теоретично решение, но в своята практическа дейност изследователите ги свеждат, не съвсем коректно, към аналогични проблеми, които имат теоретично решение.

В третата група попадат тези проблеми, които нито имат теоретично решение, нито могат да се сведат до вече решени проблеми.

Целта на настоящият курс е да покаже как от позициите на една малко позната парадигма в статистиката, наречена бейсовска, може да се даде теоретично решение на всеки изследователски проблем и в този смисъл трите групи проблеми да се сведат към първата. Разбира се, тук по-скоро става въпрос за потенциал, понеже е практически невъзможно да бъдат описани изчерпателно всички проблеми, изискващи приложението на статистически методи. По тази причина, половината от курса е посветена на теоретично въведение, а в другата половина са разгледани избрани изследователски проблеми. В теоретичното въведение са изложени принципите, инструментите и техниките на бейсовската статистика. Това означава, че в тези теми по необходимост има малко повече математика, но тя е необходима основа за разбирането на различните приложения. А в този курс ще се разглеждат приложения, отнасящи се и до трите групи изследователски проблеми:

Към третата група проблеми се отнасят изводите, основани на информация от непредставителни извадки. Примери за непредставителни извадки напоследък има достатъчно много – анкетите в Интернет, гласуванията по телефона или чрез SMS за различни телевизионни предавания (Песен на Евровизия, Тест на нацията, Big Brother, Star Academy, Музикална ку-ку академия, Великолепната шесторка и т.н., и т.н.).

Към втората група проблеми се отнасят изводите, основани на информация от представителни извадки, но отнасящи се за няколко взаимно свързани параметъра на генералната съвкупност. Такъв, например, е проблемът за оценяването на процента на гласувалите за някоя партия (кандидат), но само сред действителните гласоподаватели.

Към първата група проблеми се отнася определянето на вероятността нова единица да попадне в конкретна група, формирана по значението на даден признак. Това успешно може се съчетае с техниката на многомерните групировки и по този начин се получава лесна за употреба алтернатива на методите, носещи общото наименование „вземане на решения в условия на риск”.

ТЕМИ

1. Малко история

Теория на вероятностите.

Честотна статистика.

Бейсовска статистика.

[Основна литература: Харалампиев 2007]

[Допълнителна литература: Jaynes 1986; Jaynes 2003; Loredo 1990]

2. Малко математика

Вероятност. Условна вероятност.

Теореми за събиране и умножаване на вероятности. Теорема за пълната вероятност и теорема на Бейс. Априорни и апостериорни вероятности. Метод на максималната ентропия.

Разпределение, плътност на разпределение и функция на разпределение.

[Основна литература: Харалампиев 2007]

[Допълнителна литература: Bretthorst 1990; Jaynes 1988; Jaynes 2003; Loredo 1990]

3. Статистически изводи и заключения относно относителни дялове в генералната съвкупност на базата на априорни вероятности

Безвъзвратен подбор, възвратен подбор, големи генерални съвкупности.

[Основна литература: Харалампиев 2004б; Харалампиев 2018]

[Допълнителна литература: Харалампиев 2004а]

4. Статистически изводи и заключения относно важни параметри на разпределенията (средни аритметични, стандартни отклонения и т.н.) в генералната съвкупност на базата на априорни вероятности

Безвъзвратен подбор, възвратен подбор, големи генерални съвкупности.

[Основна литература: Харалампиев 2004б]

[Допълнителна литература: Харалампиев 2004а]

5. Приложения в областта на социалните науки – непредставителни извадки

Анкети в Интернет.

Телевизионни гласувания – „Песен на Евровизия", „Вот на доверие", „Сблъсък", „Дуел", „Big Brother", „Star Academy", „Музикална ку-ку академия", „Тест на нацията" и т.н. Проблемът за „победителя".

Студентска оценка за качеството на преподаване.

Определяне на допустимия дял на отказите при социологическите изследвания.

[Основна литература: Харалампиев 2004а; Харалампиев 2004б; Харалампиев 2005; Харалампиев 2006; Харалампиев 2012]

6. Статистически изводи и заключения относно относителни дялове в генералната съвкупност на базата на апостериорни вероятности

Безвъзвратен подбор, възвратен подбор, големи генерални съвкупности.

Най-вероятно разпределение (най-вероятни относителни дялове) в генералната съвкупност.

[Основна литература: Харалампиев 2004б; Haralampiev 2006]

7. Статистически изводи и заключения относно важни параметри на разпределенията (средни аритметични, стандартни отклонения и т.н.) в генералната съвкупност на базата на апостериорни вероятности

Безвъзвратен подбор, възвратен подбор, големи генерални съвкупности.

[Основна литература: Харалампиев 2004б]

8. Приложения в областта на социалните науки – представителни извадки

Предизборни проучвания. Проблемът за процента спрямо действителните гласове.

[Основна литература: Haralampiev 2006]

9. Определяне на вероятността нова единица да попадне в конкретна група, формирана по значенията на един или няколко признака.

Определяне на вероятността нова единица да попадне в конкретна група, формирана по значенията на един или няколко признака при работа с априорни и с апостериорни вероятности.

Банкови правила за отпускане на кредити. Правила за определяне на застрахователни премии. Възможности за приложение в областта на маркетинговите и политическите изследвания.

[Основна литература: Харалампиев 2008]

[Допълнителна литература: Jaynes 2003]

10. Приложение в областта на прогнозирането – доверителни интервали на екстраполационни прогнози.

[Основна литература: Харалампиев 2004б]

ЛИТЕРАТУРА

Харалампиев, К. 2004а. Анкетите в интернет: възможност за статистически изводи и интерпретиране на резултатите. Социологически проблеми, брой 3-4. Достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/statija9.pdf

Харалампиев, К. 2004б. Нетрадиционен поглед върху традиционни статистически проблеми. Балкани, София

Харалампиев, К. 2005. Телевизионните гласувания по телефона – проблемът за „победителя". Социологически проблеми, брой 3-4. Достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/statija13.pdf

Харалампиев, К. 2007. За парадигмите в статистиката – бейсовска статистика. Международна научна конференция „Актуални проблеми на статистическата теория и практика", Равда. Достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/doklad-1.pdf

Харалампиев, К. 2008. Лапласово правило за приемственост – интерпретации и приложения. Научна конференция с международно участие „Авангардни научни инструменти в управлението", Равда. Достъпна в Интернет на адрес: http://vsim-journal.info/static_cont/vsim_e-journal_vol_01-2010-1_screen.pdf

Харалампиев, К. 2009. Студентската оценка на преподаването – проблемът за точността на изводите. В: Социологията пред предизвикателството на различията. Юбилеен сборник, посветен на 30-годишнината на катедра „Социология". Университетско издателство „Св. Климент Охридски", София. Достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/statija12.pdf

Харалампиев, К. 2012. Още една гледна точка към проблема за отказите при социологически изследвания. Социологически проблеми, брой 1-2. Достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/statiya19.pdf

Харалампиев, К. 2018. Използване на бейсовска статистика за статистически изводи при непредварителни извадки (през примера на изследването на отпадащи студенти в бакалавърска степен на обучение във Философския факултет на Софийския университет "Св. Климент Охридски"). Реторика и комуникации, брой 36. Достъпна в Интернет на адрес: http://rhetoric.bg/%D0%B8%D0%B7%D0%BF%D0%BE%D0%BB%D0%B7%D0%B2%D0%B0%D0%BD%D0%B5-%D0%BD%D0%B0-%D0%B1%D0%B5%D0%B9%D1%81%D0%BE%D0%B2%D1%81%D0%BA%D0%B0-%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0-%D0%B7

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. Достъпна в Интернет на адрес: http://bayes.wustl.edu/glb/intro.pdf

Haralampiev, K. 2006. Bayesian Inference of Relative Frequency (In the Case of Electoral surveys). Annuaire de l'universite de Sofia "St. Kliment Ohridski", Fakulte de philosophie, Livre – Sociologie, Tome 99. Достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/article1.pdf
Превод на статията на български език е достъпен в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/prevod-article1.pdf

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

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. Достъпна в Интернет на адрес: http://bayes.wustl.edu/etj/articles/relationship.pdf

Jaynes, E. 2003. Probability Theory: the Logic of Science. Cambridge Univercity Press, Cambridge. В този курс е цитирано непълно издание на тази книга, достъпно в Интернет на адрес: http://omega.albany.edu:8008/JaynesBook.html

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. Достъпна в Интернет на адрес: http://bayes.wustl.edu/gregory/articles.pdf

АНОТАЦИИ НА ТЕМИТЕ И ТЕКСТОВЕ

ПЪРВИ РАЗДЕЛ.
ВЪВЕДЕНИЕ В БЕЙСОВСКАТА СТАТИСТИКА

В съвременната статистика има две парадигми. Едната, която е най-позната и най-използвана, се нарича честотна. Другата, която е почти непозната (особено в България) и е използвана само от тесен кръг учени и изследователи (главно по света) се нарича бейсовска. Нещо повече, различието между двете парадигми е по-дълбоко и се открива още в теорията на вероятностите. Това различие е исторически обусловено. В зората на развитието на теорията на вероятностите тя е била единна, но с появата на по-сериозни проблеми се появяват различни теоретични схеми, от чиито позиции се решават тези проблеми. Постепенно в рамките на теорията на вероятностите се обособяват двете парадигми, които се пренасят в математическата статистика, а оттам и в приложната статистика.

В този раздел ще бъде разгледана накратко историята на теорията на вероятностите и статистиката, основните принципи, инструменти и техники на бейсовската статистика и основните й полета на приложение в момента.

<u>Първа тема:</u> Малко история

Теорията на вероятностите е сравнително нов дял на математиката. Нейната история започва преди около четири века. Статистиката е още по-млада наука. Въпреки тази кратка история и двете науки са претърпели изключително бурно развитие. Появата на сериозни проблеми е довело до възникването на различни теоретични постановки, от чиито позиции се решават тези проблеми. По този начин са се обособили две парадигми – честотна и бейсовска, които в момента се намират в ожесточен сблъсък помежду си. „Новата" парадигма се стреми към утвърждаване, а старата – към запазване на позициите си.

От тази тема ще научите:

- Кои са основните етапи в развитието на теорията на вероятностите и статистиката.
- Решаването на какви практически проблеми способства за получаването на превес на едната или другата парадигма.
- Какви са полетата на приложение на бейсовската статистика.
- Какви са възможностите и какво е бъдещето на приложението на бейсовската статистика в социалните науки.

<u>Втора тема:</u> Малко математика

За разлика от честотната статистика, която е набор от различни методи, приложими в различни конкретни ситуации, бейсовската статистика е принцип, който се прилага по един и същи начин независимо от конкретния проблем.

От тази тема ще научите:

- Как се дефинира понятието вероятност от гледна точка на двете парадигми в статистиката.
- На какви условия трябва да отговаря вероятността от гледна точка на бейсовската парадигма.
- Какви са основните принципи, инструменти и техники на бейсовската статистика.
- Как се дефинират и за какво се използват разпределението, плътността на разпределение и функцията на разпределение.
- Какъв е алгоритъмът за решаване на конкретен познавателен проблем.

ТЕКСТ КЪМ ПЪРВИ РАЗДЕЛ

За парадигмите в статистиката – бейсовска статистика*

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)[1].

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. 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 (|) 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:

---

* Доклад, представен на конференцията „Актуални проблеми на статистическата теория и практика", Равда, 2007

[1] 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.

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

$P(H \mid I)$ 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[2], които 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).

---

[2] 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 A 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 A и B са две събития, вероятността да се случи едното или другото е:

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

Второ следствие: ако *A* и *B* 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[3] 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:

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

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[4] (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

---

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

[4] 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.

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.

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)[5]:

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

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.

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

[5] 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.

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.

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://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

ВТОРИ РАЗДЕЛ.
ИЗВОДИ, ОСНОВАВАЩИ СЕ НА НЕПРЕДСТАВИТЕЛНИ ИЗВАДКИ

В последните години се наблюдава бум на непредставителните извадки. Разбира се, би било пресилено всяко използване на непредставителна извадка да бъде наречено изследване, но така или иначе, някаква информация се трупа и при това е ясно, че по никакъв начин не може да се гарантира нейната представителност. Нещо повече, може да се твърди, че вече едва ли не, само изрично планираните изследвания (и то не винаги) гарантират представителност на извадката, а всички останали „изследвания" са непредставителни. Получава се така, че голям обем информация се натрупва сякаш „от само себе си". Тази информация би било добре също да бъде анализирана. От позициите на честотната статистиката обаче това е невъзможно, тъй като данните не могат да се разглеждат като случайна величина. Това изискване на честотната статистика е прекалено ограничаващо, защото бидейки неслучайна величина, данните все пак носят някаква информация, която би могла да бъде използвана.

В този раздел първо ще бъде показано как би могла да бъде извлечена цялата релевантна информацията от непредставителните извадки по принцип. След това ще бъдат разгледани примери, в които информацията е резултат от натрупване „от само себе си" (анкети в Интернет и гласувания по телефона) и изследвания, планирани като изчерпателни, но на практика реализирали се като непредставителна извадка (студентска оценка на качеството на преподаване).

<u>Трета тема:</u> Статистически изводи и заключения относно относителни дялове в генералната съвкупност на базата на априорни вероятности

В масовия случай на използване на непредставителни извадки (анкетите по Интернет и гласуванията по телефона) отговорите на задаваните въпроси формират качествен признак. При изследване на качествени признаци единствено възможната числова характеристика е относителният дял. Приложението на теоремата на Бейс в този случай изисква да се познават априорната вероятност и извадковото разпределение на относителния дял. Извадковото разпределение от своя страна изисква данните да са случайна величина. Това е валидно, само когато извадката е представителна. Следователно, когато извадката е непредставителна, извадковото разпределение не може да използва. В този случай единствено възможните вероятности са априорните вероятности.

От тази тема ще научите:

- Кои данни от извадката са релевантни при непредставителни извадки.
- Какви стойности може да приема относителният дял в генералната съвкупност.
- Как се получават плътността на разпределение и функцията на разпределение на вероятностите на относителния дял в генералната съвкупност.
- Как с помощта на функцията на разпределение се построяват доверителни интервали и се проверяват статистически хипотези относно относителният дял в генералната съвкупност.
- Как видът на подбора и големината на генералната съвкупност влияе върху крайните изводи.

<u>Четвърта тема:</u> Статистически изводи и заключения относно важни параметри на разпределенията (средни аритметични, стандартни отклонения и т.н.) в генералната съвкупност на базата на априорни вероятности

Ако в изследването са включени и количествени признаци, освен относителните дялове, могат да се изчисляват и обобщаващи числови характеристики за център на разпределението, за разсейването, асиметрията, връхната изтеченост и др. Когато извадката е непредставителна, изводите относно тези числови характеристики отново се базират на априорни вероятности.

От тази тема ще научите:

- Кои данни от извадката са релевантни при непредставителни извадки.

- Какви стойности може да приема средната аритметична в генералната съвкупност.
- Какъв е видът на разпределението на вероятностите на средната аритметична в генералната съвкупност.
- Как с помощта на функцията на разпределение се построяват доверителни интервали и се проверяват статистически хипотези относно средната аритметична в генералната съвкупност.
- Какъв е центърът и какво е разсейването на разпределението на вероятностите на средната аритметична в генералната съвкупност.
- Как видът на подбора и големината на генералната съвкупност влияят върху крайните изводи.

<u>Пета тема:</u> Приложения в областта на социалните науки – непредставителни извадки

Приложенията в областта на социалните науки се отнасят до анализирането на данни от:

- Непредставителни извадки, излъчени чрез безвъзвратен подбор от сравнително малка генерална съвкупност. Като пример ще бъде разгледана студентската оценка на качеството на преподаване. Допълнително ще бъде разгледан проблемът за влиянието на дела на извадката в генералната съвкупност върху точността на изводите.
- Непредставителни извадки, излъчени чрез възвратен подбор от сравнително малка генерална съвкупност.
- Непредставителни извадки, излъчени от достатъчно голяма генерална съвкупност, независимо от вида на подбора. Като пример ще бъдат разгледани анкетите в Интернет и гласуванията по телефона, свързани с различни телевизионни предавания. Допълнително ще бъде разгледан проблемът за определянето на победителя от гласуването в генералната съвкупност.

ТЕКСТОВЕ КЪМ ВТОРИ РАЗДЕЛ

Използване на бейсовска статистика за статистически изводи при непредставителни извадки (през примера на изследването на отпадащи студенти в бакалавърска степен на обучение във Философския факултет на Софийския университет "Св. Климент Охридски")*

Уводни думи

Много често при емпиричните изследвания, при работа на терен, се срещат трудности, които водят до това, че една извадка, която е планирана като представителна, всъщност се оказва непредставителна. Най-често такъв проблем е високият дял на неоткритите и/или неотговорилите лица. В такъв случай не е коректно да се използват класическите статистически методи за построяване на доверителни интервали и/или за проверка на хипотези, тъй като те са създадени при изричното допускане за представителност на извадката. Налага се да се прибегне до средствата на бейсовската статистика, която позволява да се построяват доверителни интервали и да се проверяват статистически хипотези на базата на данни от непредставителни извадки.

В предишни публикации [1], [2] съм показал как бейсовската статистика се използва за построяване на доверителни интервали на относителни дялове (проценти) при непредставителни извадки.

Най-общо доверителният интервал на относителен дял се получава по формулата:

(1) $$P(a \le \pi \le b) = F(b) - F(a),$$

където:

π е неизвестният относителен дял (процент) в генералната съвкупност;

a и b са границите на доверителния интервал;

P е означение за „вероятност";

$F(x)$ е т.нар. функция на разпределение.

Както се вижда, за построяването на доверителния интервал функцията на разпределение е ключова. При непредставителни извадки функцията на разпределение има следния вид [3]:

(2) $$F(x) = 1 - \frac{C_{N(1-x)-n(1-p)+m-1}^{m-1}}{C_{N-n+m-1}^{m-1}},$$

където:

N е обемът на генералната съвкупност;

n е обемът на извадката;

p е оценката на относителния дял, получена от извадката;

m е броят на разновидностите на признака;

C е означение за „комбинация".

Тази формула може да се прилага директно само при сравнително малки генерални съвкупности. При големи генерални съвкупности е по-удобно формулата да се преработи като се извърши граничен преход. В горецитираните публикации [4], [5] е извършен следният граничен преход:

(3) $$F(x) = \lim_{N\to\infty;\frac{n}{N}\to 0}\left[1 - \frac{C_{N(1-x)-n(1-p)+m-1}^{m-1}}{C_{N-n+m-1}^{m-1}}\right] = 1 - (1-x)^{m-1},$$

Тази формула показва, че когато делът на извадката спрямо генералната съвкупност е пренебрежимо малък, данните, получени от извадката, на практика не участват при изчисляването на функцията на разпределение. А това поражда парадокс – в такъв случай данните не участват и при построяването на доверителните интервали, което означава, че тези доверителни интервали могат да се построяват и без данни, а това поставя въпроса защо тогава изобщо ни е нужна извадката. Всъщност, когато извадката е непредставителна и нейният дял спрямо генералната съвкупност е пренебрежимо малък, това означава, че информацията, получена от нея, също е пренебрежимо малка, т.е. няма никакъв смисъл от такава извадка.

---

* Статия, публикувана в списание „Реторика и комуникации", брой 36/2018

За съжаление, когато планираме национално представителни извадки, генералната съвкупност се състои от няколко милиона, а извадката обикновено се състои от (няколко) хиляди, а това означава, че делът на извадката е пренебрежимо малък. И когато, поради проблеми на теренното изследване, извадката се окаже непредставителна, тогава събраните данни са абсолютно неизползваеми.

Но има и друга ситуация. Тя се получава, когато делът на извадката спрямо генералната съвкупност не е пренебрежимо малък. Тогава граничният преход ще изглежда по следния начин:

$$(4) \qquad F(x) = \lim_{N\to\infty}\left[1 - \frac{C^{m-1}_{N(1-x)-n(1-p)+m-1}}{C^{m-1}_{N-n+m-1}}\right] = 1 - \left[\frac{1-x-\frac{n}{N}(1-p)}{1-\frac{n}{N}}\right]^{m-1}$$

В тази ситуация данните, получени от извадката, вече участват при построяването на доверителните интервали [6], и нещо повече – колкото делът на извадката спрямо генералната съвкупност е по-голям, толкова точността на доверителните интервали е по-висока. Но въпреки това точността остава по-ниска спрямо доверителните интервали, които бихме получили, ако извадката е представителна.

Резултати от изследването

Точно такава ситуация се получи в изследването на отпадащите студенти в бакалавърска степен на обучение във Философския факултет на Софийския университет „Св. Климент Охридски“.

Изследването с отпадналите студенти беше планирано като изчерпателно. Генералната съвкупност включваше 511 отпаднали студенти. Те са се обучавали в редовна форма в бакалавърска степен и са от четири випуска – от 2013/2014 до 2016/2017 година.

За изследването им беше разработена онлайн анкета. В главните книги обаче намерихме информация за имейлите само на 235 отпаднали студенти. До всички тях беше изпратено електронно писмо с линк към анкетата. Периодично бяха изпращани напомнящи писма. В резултат:

- 47 мейла (20,0%) бяха върнати като грешни;
- 5 отпаднали студента (2,1%) отказаха да попълнят анкетата;
- 131 отпаднали студента (55,7%) изобщо не са отворили анкетата;
- 4 отпаднали студента (1,7%) са отворили анкетата, но не са я попълнили;
- 48 отпаднали студента (20,4%) са попълнили анкетата (29 изцяло и 19 частично).

Ние сме анализирали отговорите на тези 48 отпаднали студенти. Макар че възвръщаемостта е около 20%, все пак трябва да се има предвид, че ние разполагаме с имейлите само на 46% от всички отпаднали студенти, така че тези 48 бивши студенти, попълнили анкетата, всъщност са само 9,4% от цялата генерална съвкупност.

Изследването на контролната група от студенти, продължаващи своето образование, беше планирано като представителна извадка. Генералната съвкупност включваше 1661 студенти бакалавърска степен редовно обучение, продължаващи своето образование. Планираният обем на извадката беше 500 студенти. Моделът на извадката беше комбинация от стратифициран и прост случаен подбор, като стратификацията беше направена по специалност и курс. Във всяка страта чрез прост случаен подбор бяха избрани конкретните студенти, които да бъдат анкетирани.

В резултат от работата на терен се получиха следните резултати:

- отсъства поради мобилност по програма „Еразъм“ – 1 студент (0,2%);
- на студентска бригада – 13 студенти (2,6%);
- колегите му не го познават – 3 студенти (0,6%);
- отказали да попълнят анкетата – 16 студенти (3,2%);
- отказали се от следването (по думите на колегите им) – 12 студенти (2,4%);
- прекъснал – 1 студент (0,2%);
- сменили специалността – 3 студенти (0,6%);
- не са открити от анкетьорите – 141 студенти (28,4%);
- попълнили анкетата – 307 студенти (61,8%).

Трябва да се има предвид, че тази ниска възвръщаемост не е поради ниско качество на работата на анкетьорите. Напротив, тъй като теренната работа се провеждаше в края на летния семестър и по време на сесията, анкетьорите направиха всичко възможно да открият студентите, попаднали в извадката, като ги издирваха преди всеки изпит. Неоткриването на студент най-често означава, че този студент не се е явил на нито един изпит по време на сесията. Ето характерна извадка от кореспонденция с един от анкетьорите:

„Ходила съм да ги търся преди всичките им задължителни изпити, като на устните изпити, които се провеждат в рамките на два дена, съм ходила и двата дни и съм стояла от сутрин до следобед – до самия край на изпита, за да видя дали някой от списъка ще се появи.

Равносметка – общо събраните анкети дотук са 14 от 43 (8 от първи курс и 6 от втори). Това прави около една трета от извадката, която получих. Направих всичко възможно, но тези хора са като мухи без глави – нито си знаят изпитите, нито знаят нещо особено за изпитите. Въобще направо съм възмутена – аз ставах в 6:30 ч., за да ходя да ги търся по изпитите и стоях по цял ден, а те самите не ходят. Нямам думи!

Остава да отида на последния изпит на второкурсниците и това е – за някои от празните места колегите им казаха, че не са ги чували, но аз все пак ще видя дали няма да стане някое чудо и да се появят на изпита, затова ще ги попълня и тях със съответния статус и приключвам окончателно със събирането на данните.“

И това не беше изолиран случай. Другите анкетьори споделяха сходни проблеми.

Също така, поради проблеми с откриването на четвъртокурсниците, анкетата с тях беше проведена от членовете на екипа по време на държавните изпити и/или на защитите на дипломните работи.

Така че направихме всичко възможно да обхванем всички студенти, попаднали в извадката, и ниската възвръщаемост е по причини, които са извън наш контрол.

В крайна сметка възвръщаемостта по специалност и курс се получи както следва:

Таблица 1. Възвръщаемост по специалност и курс

Специалност	Първи курс	Втори курс	Трети курс	Четвърти курс	Общо
БИН	91,7%	100,0%	66,7%	30,8%	69,6%
Европеистика	40,0%	71,4%	0,0%	91,7%	48,3%
Културология	55,6%	78,6%	78,6%	63,6%	68,4%
Политология	55,0%	80,0%	28,6%	57,1%	55,6%
Психология	75,0%	70,0%	75,0%	100,0%	79,1%
Публична администрация	55,0%	64,3%	58,3%	58,3%	58,6%
Социология	57,1%	90,9%	81,8%	58,8%	68,3%
Философия	29,2%	72,2%	69,2%	0,0%	43,3%
Общо	55,8%	76,5%	57,3%	59,6%	61,8%

Ниската възвръщаемост и неравномерното ѝ разпределение по специалност и курс правят извадката непредставителна. Това наложи сравнението между двете групи да бъде направено с помощта на бейсовската статистика.

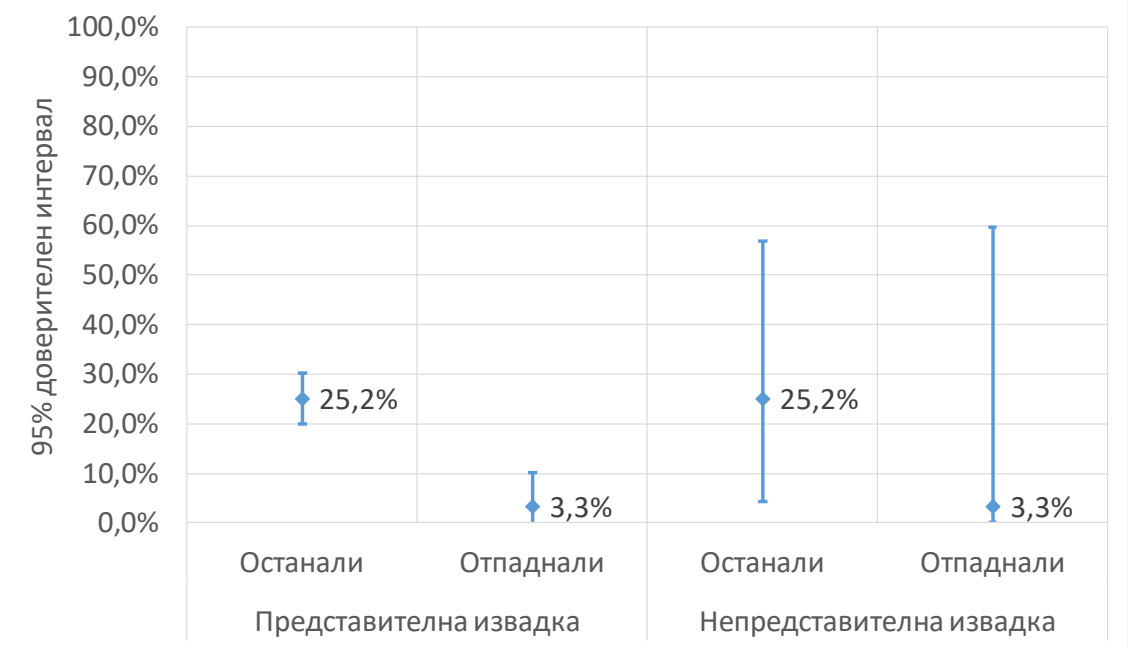
Ето един характерен пример за такова сравнение:

Таблица 2. Разпределение на двете групи студенти по важността на очакването за добър доход като причина за кандидатстване

		Група		Общо
		Останали	**Отпаднали**	**Общо**
Завършването на тази специалност ще увеличи възможностите ми за добър доход	Много важно	25,2%	3,3%	23,1%
	Важно	39,2%	30,0%	38,3%
	Маловажно	19,8%	40,0%	21,8%
	Изобщо не е важно	15,8%	26,7%	16,9%
Общо		**100,0%**	**100,0%**	**100,0%**

За да се провери дали има статистически значимо различие между двете групи, са построени доверителните интервали за относителните дялове на всеки отделен отговор във всяка от двете групи. Тези доверителни интервали са представени на следващите графики, като всяка графика се състои от две части. Лявата част представя доверителните интервали такива, каквито биха били, ако извадките биха били представителни. Тъй като двете извадки не са представителни, дясната част на графиката представя доверителните интервали такива, каквито са, изчислени по алгоритъма от Приложение 1.

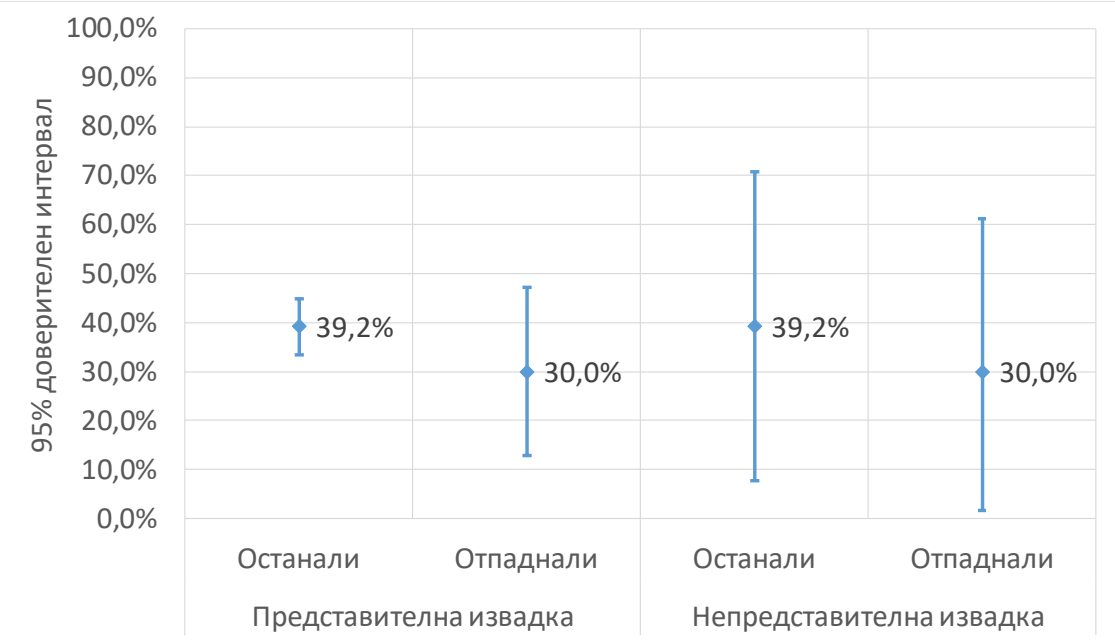
Доверителните интервали задават диапазона на извадковата грешка. Ако разликата между оценките на относителните дялове в двете групи надхвърля грешката, т.е. ако оценката на относителния дял в едната група е извън доверителния интервал на относителния дял в другата група, и обратно, то разликата между двете групи е статистически значима. Ако разликата между оценките на относителните дялове в двете групи не надхвърля грешката, т.е. ако оценката на относителния дял в едната група е в рамките на доверителния интервал на относителния дял в другата група, и обратно, то разликата между двете групи е статистически незначима.



Фиг. 1. Доверителни интервали на относителните дялове на отговора „Много важно“

Фиг. 1 показва, че ако двете извадки биха били представителни, би имало статистически значимо различие между двете групи по отношение на относителния дял на отговора „Много важно“. Но тъй като двете извадки не са представителни, този извод може да бъде потвърден само частично, тъй като оценката на относителния дял в групата на отпадащите студенти (3,3%) е извън

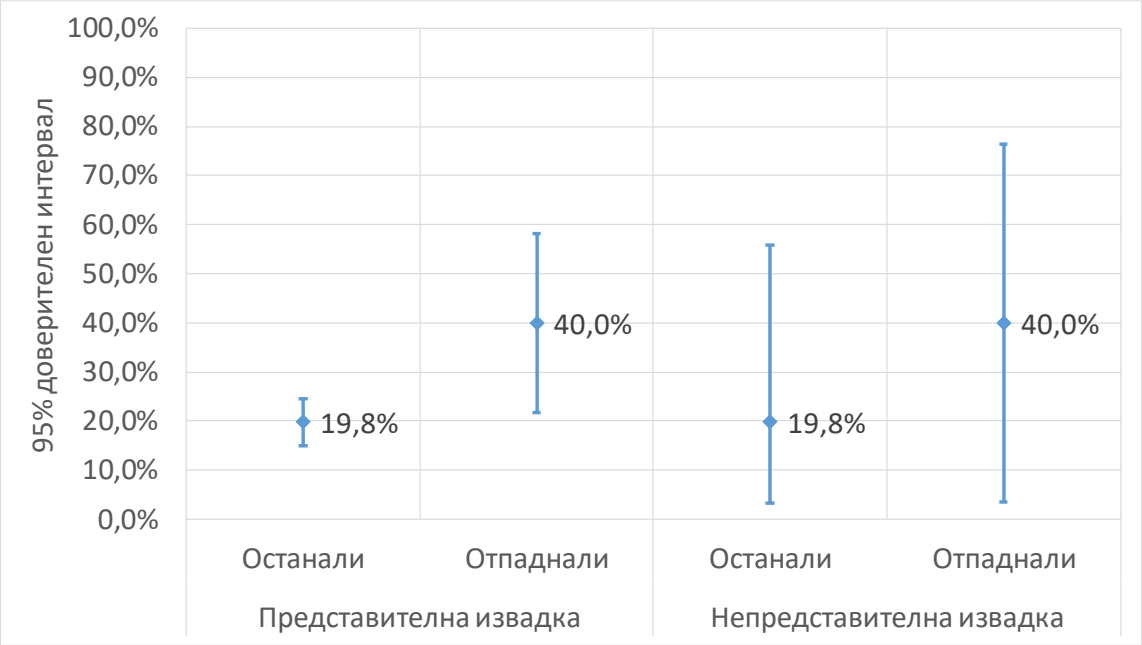
доверителния интервал на относителния дял в групата на студентите, продължаващи своето образование, но обратното не е вярно – оценката на относителния дял в групата на студентите, продължаващи своето образование, (25,2%) е в рамките на доверителния интервал на относителния дял в групата на отпадналите студенти. Иначе казано, разликата между оценките на двата относителни дяла надхвърля извадковата грешка в групата на студентите, продължаващи своето образование, но не надхвърля грешката в групата на отпадащите студенти. Именно затова статистически значимото различие се потвърждава само частично.



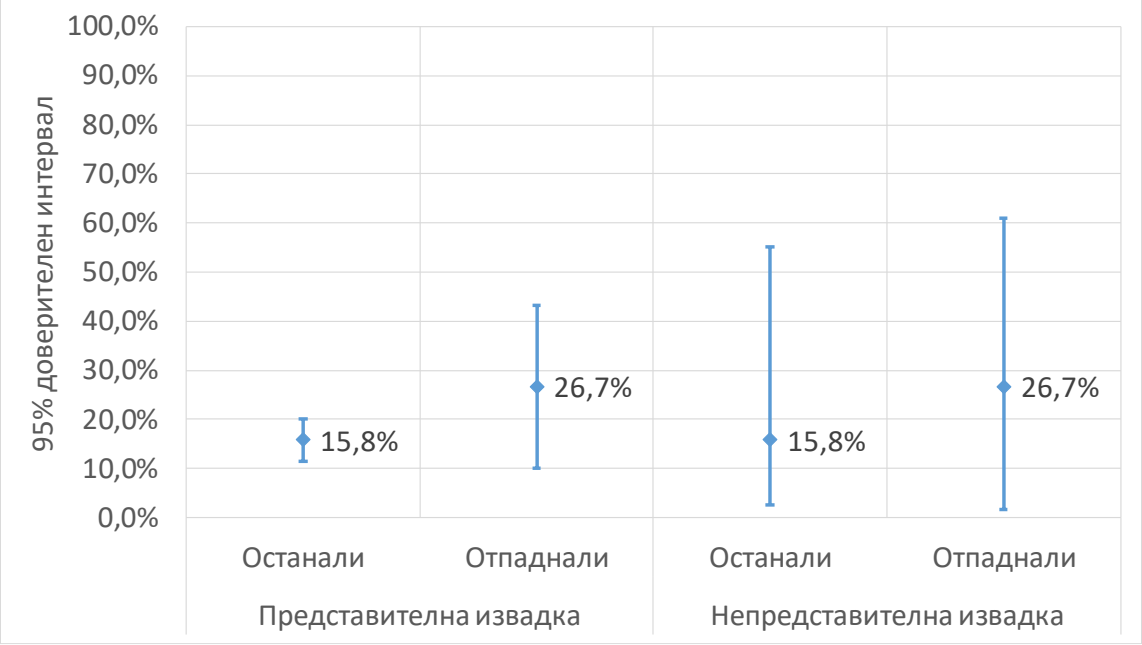
Фиг. 2. Доверителни интервали на относителните дялове на отговора „Важно“

Фиг. 2 показва, че ако двете извадки биха били представителни, различието между двете групи по отношение на относителния дял на отговора „Важно“ нямаше да бъде статистически значимо. Въпреки че двете извадки не са представителни, този извод може да бъде потвърден напълно. Този резултат не е неочакван, тъй като грешката при непредставителните извадки е по-голяма спрямо грешката при представителните извадки. Така че, ако едно различие е по-малко от по-малката грешка, то то със сигурност ще е по-малко и от по-голямата грешка.

Фиг. 3 показва, че ако двете извадки биха били представителни, би имало статистически значимо различие между двете групи по отношение на относителния дял на отговора „Маловажно“. Но тъй като двете извадки не са представителни, то този извод не може да бъде потвърден, тъй като оценката на относителния дял във всяка една от двете групи попада в рамките на доверителния интервал на относителния дял в другата група. Иначе казано, разликата между оценките на двата относителни дяла не надхвърля извадковата грешка. Именно затова статистически значимото различие не се потвърждава.



Фиг. 3. Доверителни интервали на относителните дялове на отговора „Маловажно“



Фиг. 4. Доверителни интервали на относителните дялове на отговора „Изобщо не е важно“

Фиг. 4 показва, че ако двете извадки биха били представителни, различието между двете групи по отношение на относителния дял на отговора „Изобщо не е важно“ нямаше да бъде статистически значимо, което се потвърждава и при непредставителни извадки.

Като равносметка от цялото изследване – от 105 въпроса, по които беше направено сравнение между двете групи студенти, при 37 би имало статистически значимо различие, ако двете извадки биха били представителни. Тъй като извадките не са представителни, всяко едно от тези 37 сравнения беше проверено с помощта на бейсовската статистика, като само при 5 от тях статистически значимото различие се потвърди, и то само частично.

Заключение

В заключение може да се обобщи, че когато поради различни причини, свързани с работата на терен, планираната като представителна извадка се окаже непредставителна, също могат да се построяват доверителни интервали и да се проверяват статистически хипотези, но:

- Делът на извадката спрямо генералната съвкупност не трябва да бъде пренебрежимо малък. Ако делът на извадката спрямо генералната съвкупност е пренебрежимо малък и тя е непредставителна, то тя на практика е абсолютно безполезна.
- Грешките, изчислени от непредставителни извадки, са по-големи от грешките, които биха били изчислени от същите извадки, ако извадките биха били представителни. Това води до по-широки доверителни интервали, а оттам и до по-рядко установяване на статистически значими различия.

Благодарности:

Тази статия е резултат от изследване, финансирано от Фонд „Научни изследвания“ на Софийски университет „Св. Климент Охридски“, проект №80-10-78/20.04.2017.

Приложение 1

Алгоритъм за построяване на доверителни интервали при непредставителни извадки, когато делът на извадката спрямо генералната съвкупност не е пренебрежимо малък

1. Изчислява се максимално възможната максимална грешка по формулата [7]:

(A.1) $\Delta_{p,max} = \min\left[(p - \pi_{min}); (\pi_{max} - p)\right]$,

където:

(A.2) $\pi_{min} = p\frac{n}{N}$ [8];

(A.3) $\pi_{max} = p\frac{n}{N} + 1 - \frac{n}{N}$ [9].

2. Построява се доверителният интервал по формула (1):

(A.4) $P\left(p - \Delta_{p,max} \leq \pi \leq p + \Delta_{p,max}\right) = F\left(p + \Delta_{p,max}\right) - F(p - \Delta_{p,max})$,

където функцията на разпределение се изчислява по формула (4).

3.1. Ако така изчислената вероятност е по-голяма от желаната, тогава максималната грешка се намалява и отново се построява доверителният интервал. Тази стъпка се повтаря, докато се получи желаната вероятност.

3.2. Ако така изчислената вероятност е по-малка от желаната, тогава доверителният интервал трябва да се разшири. Това разширяване обаче може да стане само в едната посока и по този начин доверителният интервал ще стане асиметричен.

3.2.1. Ако $\min[(p - \pi_{min}); (\pi_{max} - p)] = p - \pi_{min}$, тогава доверителният интервал може да се разшири само надясно. Тогава директно може да се изчисли горната граница на доверителния интервал, като се реши следното уравнение:

(A.5) $F(x) = 1 - \left[\frac{1-x-\frac{n}{N}(1-p)}{1-\frac{n}{N}}\right]^{m-1} = P$,

където P е желаната вероятност.

След решаването на това уравнение се получава горната граница на доверителния интервал и самият доверителен интервал:

(A.6) $P\left[\pi_{min} \leq \pi \leq \pi_{min} + \left(1 - \frac{n}{N}\right)\left(1 - \sqrt[m-1]{1-P}\right)\right] = P$

3.2.2. Ако $\min[(p - \pi_{min}); (\pi_{max} - p)] = \pi_{max} - p$, тогава доверителният интервал може да се разшири само наляво. Тогава директно може да се изчисли долната граница на доверителния интервал, като се реши следното уравнение:

(A.7) $F(x) = 1 - \left[\frac{1-x-\frac{n}{N}(1-p)}{1-\frac{n}{N}}\right]^{m-1} = 1 - P$,

След решаването на това уравнение се получава долната граница на доверителния интервал и самият доверителен интервал:

(A.8) $P\left[\pi_{max} - \left(1 - \frac{n}{N}\right)\sqrt[m-1]{P} \leq \pi \leq \pi_{max}\right]$

Цитати и бележки:

[1] Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани, 71–79.

[2] Харалампиев, К. (2004б). Анкетите в Интернет: възможност за статистически изводи и интерпретиране на резултатите, *Социологически проблеми*, бр. 3–4, 207–211.

[3] Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани, 55.

[4] Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани, 66–67.

[5] Харалампиев, К. (2004б). Анкетите в Интернет: възможност за статистически изводи и интерпретиране на резултатите, *Социологически проблеми*, бр. 3–4, 208.

[6] Алгоритъмът за построяване на доверителните интервали е описан в Приложение 1.

[7] Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани, 72.

[8] Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани, 54.

[9] Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани, 55.

Библиография:

Харалампиев, К. (2004а). *Нетрадиционен поглед върху традиционни статистически проблеми*. София: Балкани.

Харалампиев, К. (2004б). Анкетите в Интернет: възможност за статистически изводи и интерпретиране на резултатите, *Социологически проблеми*, бр. 3–4, 203–211.

Статията е по проект „Формиране на компетентности и усъвършенстване на умения за прилагане на съвременни методи и методики за научни изследвания от млади учени” (Договор № ДМ10/2 от 14.12.2016 г. по Фонд „Научни изследвания”).

Студентската оценка на преподаването – проблемът за точността на изводите*

За целите на атестацията на преподавателите и по-общо за целите на управлението на качеството на учебния процес е нужна информация за това как студентите оценяват учебните дисциплини и преподавателите си. Изследванията на студентските мнения обаче се сблъскват с един постоянстващ проблем – отговарят не всички студенти, а само тази част от тях, които са открити по време на провеждане на изследването и са пожелали да попълнят анкетната карта (Илиева 2002: 145-146, Маринова 2006: 9 и 14, Харалампиев, Караджов 2003: 138). Това прави изследването непредставително и то по метода на отзовалите се. Въпреки това, има три съществени разлики от класическия метод на отзовалите се, които значително облекчават работата на изследователя:

Първо, при класическия метод на отзовалите се най-същественият проблем е, че всъщност не е ясно коя е изучаваната генерална съвкупност и следователно не е ясно как трябва да се генерализират изводите от извадката. При изследване на студентските мнения генералната съвкупност е ясно дефинирана, въпреки че са възможни вариации, описани по-долу в текста.

Второ, класическият метод на отзовалите се предполага, че едно лице би могло да отговори няколко пъти на анкетните въпроси и в този смисъл подборът е възвратен. При изследване на студентските мнения един студент оценява само един път конкретна дисциплина и конкретен преподавател и в този смисъл подборът е безвъзвратен.

Трето, от първите две различия следва третото. При класическия метод на отзовалите се, тъй като не е ясна изследваната съвкупност и е възможно едно лице да отговаря многократно, обикновено изводите се генерализират за безкрайно голяма генерална съвкупност[6]. При изследване на студентските мнения е обратното – генералната съвкупност е крайна и сравнително малка[7].

Тези три различия позволяват при изследване на студентските мнения чрез извадки, за които не може да се гарантира представителност, да се правят по-точни изводи в сравнение с изводите при класическия метод на отзовалите се[8].

Основната цел на настоящата статия е да покаже от какво зависи точността на изводите при изследване на студентските мнения. Успоредно с това ще бъде показано как се правят самите изводи, валидни за цялата генерална съвкупност.

И така, изследването започва с дефиниране на изследваната съвкупност. Още тук възниква един етичен проблем – нужно ли е да правим всичко възможно, за да обхванем всички студенти, след като част от тях така или иначе не посещават лекциите и упражненията? Иначе казано – имат ли право студенти, които не са посещавали учебните занимания, да дават оценка на дисциплината и на преподавателя?

Различните възможности за отговор и съответно за формиране на генералната съвкупност и извадката са представени в таблица 1:

---

* Статия, публикувана в „Социологията пред предизвикателството на различията“, Юбилеен сборник, посветен на 30-годишнината на катедра „Социология”. Университетско издателство „Св. Климент Охридски“

[6] Обикновено това са „читателиТЕ”, „зрителиТЕ”, „потребителиТЕ на Интернет”, „българиТЕ”, „хораТА” и т.н.

[7] Обикновено при такива изследвания наблюдаваната съвкупност са студентите от една и съща специалност и един и същи курс.

[8] Правенето на статистически изводи при непредставителни извадки е описано в Харалампиев 2004а.

Таблица 1

Разпределение на всички студенти по посещаемост на учебните занимания

В момента	По принцип		Общо
	Посещават	Не посещават	
Присъстват	f_{11}	f_{12}	$\sum_j f_{1j}$
Отсъстват	f_{21}	f_{22}	$\sum_j f_{2j}$
Общо	$\sum_i f_{i1}$	$\sum_i f_{i2}$	$\sum_i \sum_j f_{ij}$

Първа възможност – генералната съвкупност са студентите, които „по принцип" посещават учебните занимания, а извадката са тези от тях, които в момента на изследването присъстват и желаят да попълнят анкетната карта. В този случай обемът на извадката е $f_{11} = n$, а обемът на генералната съвкупност е $\sum_i f_{i1} = N$.

Втора възможност – генералната съвкупност са всички студенти, независимо дали посещават учебните занимания или не, а извадката са тези от тях, които в момента на изследването присъстват и желаят да попълнят анкетната карта. В този случай обемът на извадката е $\sum_j f_{1j} = n$, а обемът на генералната съвкупност е $\sum_i \sum_j f_{ij} = N$.

Ясно е, че има два начина за дефиниране на генералната съвкупност. За всеки от тях извадката се дефинира еднозначно. Това означава, че може да се работи с универсалните означения за обемите на извадката и на генералната съвкупност (съответно *n* и *N*), като в тях се влага различно съдържание в зависимост от конкретния начин на дефиниране на генералната съвкупност.

Най напред ще разгледам анкетните въпроси, чиито отговори са разположени на бална или интервална скала. Тези скали имат следната особеност – значенията на признаците са **числа**. Това означава, че може да се приложи методът за правене на изводи относно количествени признаци при непредставителни извадки, предложен от Харалампиев (Харалампиев 2004б: 8-20) и по-конкретно методът за оценяване на средната аритметична (Харалампиев 2004б: 31-34).

Най-общо при непредставителни извадки неизвестната средна аритметична в генерална съвкупност има симетрично вероятностно разпределение, което е приблизително нормално, с център

$$\bar{\mu} = \frac{\frac{(N-n)(x_1 + x_m)}{2} + \sum_{i=1}^{m} x_i f_i}{N} \quad (1)$$

и разсейване

$$\sigma_\mu = \frac{x_m - x_1}{2N} \sqrt{\frac{(N-n)(N-n+m)}{3(m-1)}}, \quad (2)$$

където x_i са значенията на признака, f_i са честотите, x_1 е най-малкото значение на признака, x_m е най-голямото значение на признака, а *m* е броят на всички значения.

Освен това, възможните стойности на средната са ограничени в интервала между

$$\mu_{\min} = \frac{x_1(N-n) + \sum_{i=1}^{m} x_i f_i}{N} \quad (3)$$

и

(4) $$\mu_{\max} = \frac{x_m(N-n) + \sum_{i=1}^{m} x_i f_i}{N}$$

От симетричността на разпределението следва, че $\bar{\mu}$ е едновременно и най-вероятна стойност, и следователно σ_μ, което показва средния размер на отклоненията около центъра (т.е. в случая около най-вероятната стойност), е измерител на точността. Затова нека видим от какво зависят тези две стойности.

Първо, формула (1) може да се запише във вида:

(5) $$\bar{\mu} = \left(1 - \frac{n}{N}\right).\frac{x_1 + x_m}{2} + \frac{n}{N}.\bar{x},$$

където $\bar{x}$ е извадковата средна.

От формула (5) се вижда, че най-вероятната стойност на средната аритметична в генералната съвкупност е претеглена средна от извадковата средна и геометричния център на разпределението $\left(\frac{x_1 + x_m}{2}\right)$, с тегла съответно делът на извадката в генералната съвкупност и делът на останалата част от генералната съвкупност. Следователно, колкото делът на извадката е по-голям, толкова най-вероятната стойност на средната аритметична в генералната съвкупност ще се доближава до извадковата средна, и обратно, колкото делът на извадката е по-малък, толкова най-вероятната стойност на средната аритметична в генералната съвкупност ще се доближава до геометричния център на разпределението.

Този резултат е различен от резултата, които бихме получили, ако извадката би била представителна. При представителни извадки най-вероятната стойност на средната аритметична в генералната съвкупност съвпада с извадковата средна, докато тук, тъй като извадката е непредставителна, извадковата средна вече е изместена оценка на средната аритметична в генералната съвкупност, като колкото делът на извадката е по-малък, толкова извадковата средна е по-изместена.

Второ, формула (2) може да се запише във вида:

(6) $$\sigma_\mu = \frac{x_m - x_1}{2}\sqrt{\frac{\left(1 - \frac{n}{N}\right)\left(1 - \frac{n}{N} + \frac{m}{N}\right)}{3(m-1)}}$$

От формула (6) се вижда, че колкото делът на извадката е по-голям, толкова разсейването е по-малко, т.е. точността е по-голяма.

Този резултат също е различен от резултата, които бихме получили, ако извадката би била представителна. При представителни извадки точността на оценката зависи в много по-голяма степен от абсолютния обем на извадката и в много по-малка степен от дела й в генералната съвкупност, докато тук се вижда, че **при непредставителни извадки абсолютният обем на извадката не оказва НИКАКВО влияние върху точността,** а единствено делът й има значение.

Трето, горният извод може да се допълни, ако от формули (3) и (4) се изчисли размахът на разпределението:

(7) $$\mu_{\max} - \mu_{\min} = (x_m - x_1)\left(1 - \frac{n}{N}\right)$$

От формула (7) отново се вижда, че колкото делът на извадката е по-голям, толкова разсейването е по-малко, т.е. точността е по-голяма.

Тези изводи могат да бъдат илюстрирани с пример: 22 студенти са оценили преподавателя си по петобална скала със значения от 1 до 5. Извадковата средна е 3,95. Генералната съвкупност обхваща 66 студента. Делът на извадката е 33%[9]. Следователно най-вероятната средна оценка е 3,32 с разсейване ±0,41 и размах 2,67.

Нека запазим абсолютния обем на извадката, а да променим обема на генералната съвкупност, като по този начин променим дела на извадката. Нека сега генералната съвкупност да обхваща 40 студента[10]. Тогава делът на извадката ще бъде 55% и най-вероятната средна стойност ще бъде 3,52 с разсейване ±0,29 и размах 1,80. Видно е, че в този случай оценката е по-точна и е по-близо до извадковата средна. Сравнението между двата случая е представено графично на фигура 1.

С това целите на статията са изпълнени. Нека обобщим:

Първо, най-вероятната стойност на средната аритметична в генералната съвкупност се намира между геометричния център на разпределението и извадковата средна, като близостта до едната или до другата стойност се определя от дела на извадката в генералната съвкупност.

Второ, точността зависи от диапазона между най-голямото и най-малкото значение на признака и от дела на извадката в генералната съвкупност.

Трето, формули (1) и (2) са достатъчни за правенето на статистически изводи относно средната аритметична в генералната съвкупност при непредставителни извадки и безвъзвратен подбор.

Въпреки че първоначално дефинирани цели са изпълнени, ще продължа с осветляването на някои важни особености на метода, които сега са скрити. Тяхното игнориране е потенциално опасно, защото директното приложение на формули (1) и (2) (без отчитане на тези особености) може да доведе до грешки и неточности.

Първо, ключово важно изискване е числовите стойности, с които се описват значенията на признака да са **последователни** числа. На практика при балните скали това е изпълнено, защото там значенията на признака са предварително дефинирани и анкетираното лице трябва да избере измежду вече зададените стойности. При интервалните скали проблемът е по-сериозен, защото се допуска, че може да няма предварително зададени стойности, а анкетираното лице само да вписва числата, които са отговор на анкетните въпроси. В този случай има различни варианти за справяне с проблема. Ако признакът е прекъснат (дискретен), за негови значения могат да се вземат всички цели числа в диапазона между най-малкото и най-голямото посочено число, макар че някои от тях може да не са избрани от нито едно анкетирано лице. Ако признакът е непрекъснат, трябва да се направи интервална групировка и за значения на признака да се вземат средите на интервалите.

Второ, макар че при балните скали числовите стойности са последователни и са предварително дефинирани, може да възникне следната ситуация: най-малката и/или най-голямата стойност не са посочени от нито едно анкетирано лице. Тогава възниква друг етичен проблем – можем ли да интерпретираме тази числова стойност като значение на признака? Иначе казано – може преподавателят да е толкова добър (лош), че нито един студент в генералната съвкупност да не му постави най-ниската (най-високата) оценка, а може в генералната съвкупност да има студенти, които биха поставили най-ниска (най-висока) оценка, но нито един от тях не е попаднал в извадката. За да се провери дали това наистина е проблем, ще бъде разгледана ситуацията, при която най-ниската оценка не е

---

[9] Тези данни са от реално проведено изследване, макар че името на преподавателя и специалността са премълчани. Нарочно е подбран този преподавател, защото при него делът на извадката е близък до посочения от Маринова дял на извадката при анкетата по пощата – 31,1% (Маринова 2006: 9), до посочения от Илиева като долна граница в изследванията, проведени в ХТМУ-София дял 30% (Илиева 2002: 145) и до посочените дялове от Караджов и Харалампиев в изследвания, проведени в Богословски факултет на Софийски университет (Харалампиев, Караджов 2003: 138).

[10] Изборът на обема на генералната съвкупност е фиктивен, само за целите на илюстрацията. Новият дял на извадката е близък до посочения от Маринова дял на извадката при оценяване на упражненията по Интернет – 50% (Маринова 2006: 14). Не е правена илюстрация на ситуациите с дял на извадката близък до посочения от Маринова дял при оценяване на лекциите по Интернет – 80% (Маринова 2006: 14) и с посочения от Илиева като горна граница в изследванията, проведени в ХТМУ-София дял 95% (Илиева 2002: 145).

посочена от нито един студент, като ситуацията, при която не е посочена най-високата оценка, е огледален образ.

Ако приемем, че в генералната съвкупност има студенти, които биха поставили най-ниската оценка, то най-вероятната стойност на средната аритметична в генералната съвкупност, разсейването, най-малката и най-голямата възможна стойност на средната се получават по описания по-горе начин.

Ако приемем, че в генералната съвкупност няма студенти, които биха поставили най-ниската оценка, то най-вероятната стойност на средната аритметична в генералната съвкупност, разсейването, най-малката и най-голямата възможна стойност на средната се получават по вариации на горните формули:

(8) $$\bar{\mu}' = \left(1 - \frac{n}{N}\right).\frac{x_2 + x_m}{2} + \frac{n}{N}.\bar{x}$$

(9) $$\sigma'_{\mu} = \frac{x_m - x_2}{2}\sqrt{\frac{\left(1 - \frac{n}{N}\right)\left(1 - \frac{n}{N} + \frac{m-1}{N}\right)}{3(m-2)}}$$

(10) $$\mu'_{\min} = \frac{x_2(N-n) + \sum_{i=1}^{m} x_i f_i}{N}$$

(11) $$\mu'_{\max} = \frac{x_m(N-n) + \sum_{i=1}^{m} x_i f_i}{N}$$

Тогава разликите между стойностите, получени при двете различни допускания са:

(12) $$\bar{\mu}' - \bar{\mu} = \left(1 - \frac{n}{N}\right).\frac{x_2 - x_1}{2}$$

(13) $$(\sigma'_{\mu})^2 - \sigma_{\mu}^2 = -\frac{1}{3}\left(\frac{x_2 - x_1}{2}\right)^2\left(1 - \frac{n}{N}\right)\left[1 - \frac{n}{N} + \frac{2(m-1)}{N}\right]$$

(14) $$\mu'_{\min} - \mu_{\min} = \left(1 - \frac{n}{N}\right).(x_2 - x_1)$$

(15) $$\mu'_{\max} - \mu_{\max} = 0$$

Тези резултати дават възможност да се направят няколко извода:

Първо, ако допуснем, че в генералната съвкупност няма студенти, които биха поставили най-ниската оценка, новото разпределение на средната аритметична в генералната съвкупност ще съвпада в горния си край със старото (формула (15)) обаче в долния си край двете разпределения ще се различават с не повече от една единица (формула (14)), а разликата между най-вероятните стойности ще бъде най-много половин единица (формула (12)).

Второ, най-малката възможна стойност и най-вероятната стойност на новото разпределение са по-близо до извадковата средна, отколкото съответните им стойности в старото разпределение.

Трето, ако допуснем, че в генералната съвкупност няма студенти, които биха поставили най-ниската оценка, новото разпределение ще бъде с по-малко разсейване от старото (формула (13)), следователно изводът ще бъде по-точен.

Четвърто, с увеличаване на дела на извадката в генералната съвкупност, разликите между стойностите, получени при двете различни допускания, стават все по-малки.

Нека отново илюстрираме получените изводи с пример. За целта да модифицираме предходния пример: 22 студенти са оценили преподавателя си по петобална скала със значения от 1 до 5, като нито един студент не е посочил оценка 1. Отново извадковата средна е 3,95, генералната съвкупност

обхваща 40 студента и делът на извадката е 55%. Най-вероятната стойност на средната аритметична в генералната съвкупност и разсейването обаче вече са други – съответно 3,75 и ±0,25. Сравнението с предходния пример показва, че новата най-вероятна стойност е още по-близо до извадковата средна и новото разсейване е още по-малко. Сравнението е представено графично на фигура 2.

Получените резултати показват, че това, че най-малката и/или най-голямата стойност не са избрани от нито едно анкетирано лице наистина е проблем. Допускането, че в генералната съвкупност има студенти, които биха поставили най-ниската (най-високата) оценка води до по-консервативни изводи, а допускането, че в генералната съвкупност няма студенти, които биха поставили най-ниската (най-високата) оценка води до по-точни изводи, които обаче биха могли да се окажат прекалено оптимистични. Статистически коректният подход в такива ситуации е да се дадат и двете възможни решения, макар че това не решава проблема[11], а само прехвърля отговорността за решаването му на следващия субект в управленската верига.

До тук разгледах признаците, които позволяват да се изчисляват средни стойности. Това са признаците, разположени на бални и интервални скали. Но освен тях съществуват и признаци, чиито значения са разположени на номинална скална. Тези признаци не позволяват да се изчисляват средни стойности. Единствената тяхна числова характеристика е относителният дял. Обаче за разлика от средната аритметична в генералната съвкупност, която има симетрично вероятностно разпределение, вероятностното разпределение на относителния дял в генералната съвкупност е крайно асиметрично L-разпределение (Харалампиев 2004б: 55). Това означава, че неговата средна и стандартно отклонение не са адекватни измерители съответно на най-вероятната стойност и на разсейването около нея. Все пак и при тези признаци могат да се правят изводи относно най-вероятната стойност на относителния дял в генералната съвкупност и относно точността.

Първо, най-вероятната стойност на относителния дял на *i*–тото значение на признака в генералната съвкупност е най-малката възможна стойност:

(16) $$\pi_{i,\min} = \frac{f_i}{N}$$ (Харалампиев 2004б: 54)

Тази формула може да се запише във вида:

(17) $$\pi_{i,\min} = \left(1 - \frac{n}{N}\right).0 + \frac{n}{N}.p_i ,$$

където p_i е извадковият относителен дял на *i*–тото значение на признака.

От формула (17) се вижда, че най-вероятната стойност на относителния дял в генералната съвкупност е претеглена средна от извадковия относителен дял и нулата, с тегла съответно делът на извадката в генералната съвкупност и делът на останалата част от генералната съвкупност. Следователно, колкото делът на извадката е по-голям, толкова най-вероятната стойност на относителния дял в генералната съвкупност ще се доближава до извадковия относителен дял, и обратно, колкото делът на извадката е по-малък, толкова най-вероятната стойност на относителния дял в генералната съвкупност ще се доближава до нулата.

Второ, след като стандартното отклонение не е адекватен измерител на разсейването, вместо него може да се използва средното отклонение около най-вероятната стойност:

(18) $$\delta = \left(1 - \frac{n}{N}\right).\frac{1}{m}$$

[11] Спорен въпрос е дали този проблем изобщо е решим. От статистическа гледна точка той е нерешим (макар че любовта на статистиците към по-консервативните решения е широко известна), така че за решаването му трябва да се търсят други критерии, извън статистическата теория. Не случайно в началото дефинирах този проблем като етичен, а не като статистически.

От формула (18) се вижда, че колкото делът на извадката е по-голям, толкова средното отклонение около най-вероятната стойност е по-малко, т.е. точността е по-голяма.

Трето, формули (16) и (18) са достатъчни за правенето на статистически изводи относно относителния дял в генералната съвкупност при непредставителни извадки и безвъзвратен подбор.

Четвърто, въпросът „Какво става, ако най-малкото и/или най-голямото значение на признака не са посочени от нито едно анкетирано лице?" вече се формулира като „Какво става, ако някое от значенията на признака не е посочено от нито едно анкетирано лице?". От формули (16) и (18) се вижда, че ако допуснем, че в генералната съвкупност няма студенти, които биха посочили непосоченото значение на признака, това не оказва влияние върху най-вероятната стойност на другите относителни дялове в генералната съвкупност, а само влошава точността.

Накрая, още веднъж да обобщим получените резултати:

При изследване на студентските мнения чрез извадки, за които не може да се гарантира представителност, най-вероятната стойност на изследвания параметър в генералната съвкупност (средна аритметична или относителен дял) е претеглена средна от две осреднявани величини, едната от които е оценката от извадката. Оценката от извадката участва в осредняването с тегло равно на дела на извадката в генералната съвкупност. Следователно, колкото делът на извадката в генералната съвкупност е по-голям, толкова най-вероятната стойност на параметъра е по-близо до оценката от извадката. Точността на изводите също зависи от дела на извадката в генералната съвкупност – по-голям дял на извадката води до по-голяма точност.

ЛИТЕРАТУРА

Илиева, М. 2002. Студентската оценка за преподаването. ХТМУ, С.

Маринова, С. 2006. Доклад относно някои елементи на реформата на висшето образование в Германия: системи за оценка на качеството на преподаване; кредитна система; форми на тюторство. Непубликуван ръкопис.

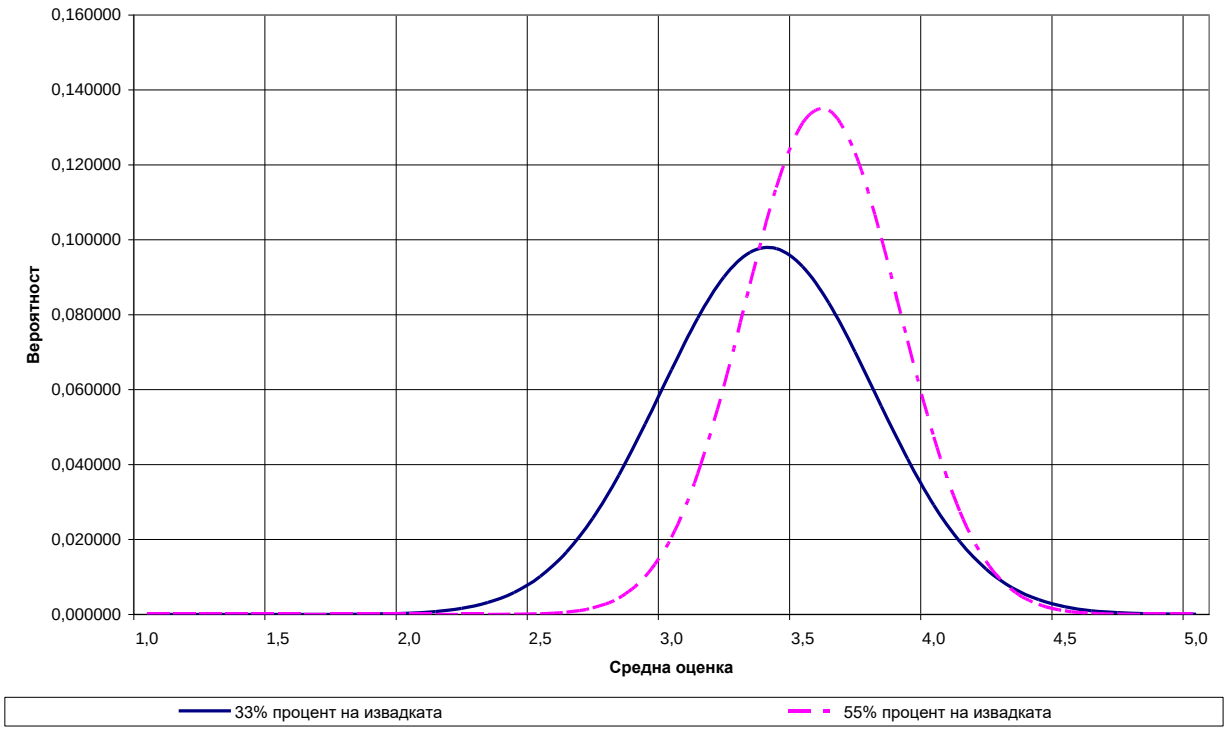
Харалампиев, К., С. Караджов**.** 2003. Доклад със сравнителен анализ на данни, получени от емпирични социологически изследвания, проведени със студенти от Богословски факултет на СУ „Св. Климент Охридски" през декември 2002, май 2003 и декември 2003 година. „Богословска мисъл", 3-4

Харалампиев, К. 2004а. Анкетите в Интернет – възможност за статистически изводи и интерпретиране на резултатите. „Социологически проблеми", 3-4

Харалампиев, К. 2004б. Нетрадиционен поглед върху традиционни статистически проблеми. Балкани, С.

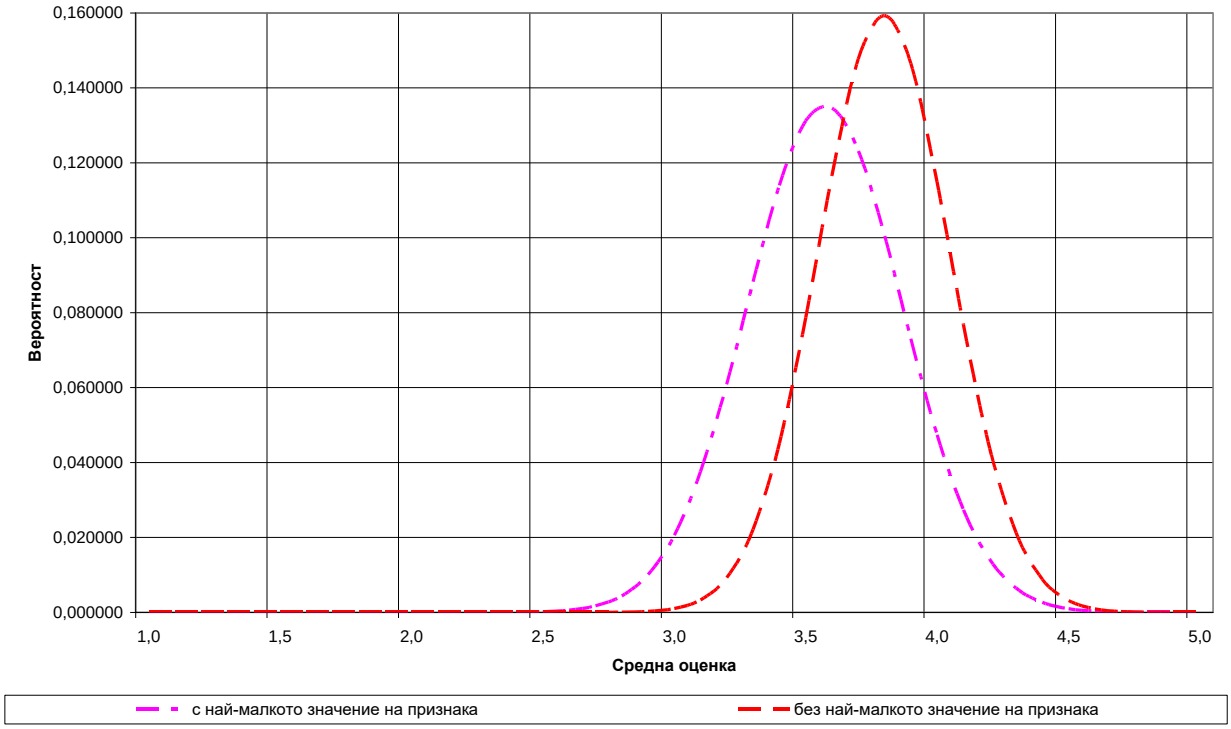
Фигура 1

Вероятностно разпределение на средната аритметична в генералната съвкупност при два различни дяла на извадката



Фигура 2

Вероятностно разпределение на средната аритметична в генералната съвкупност при включване и изключване на най-малкото значение на признака



Анкетите в Интернет: възможност за статистически изводи и интерпретиране на резултатите*

В статията ще бъдат разгледани анкетите, поместваны на различни интернет страници. Става въпрос за ситуацията, при която е зададен въпрос (обикновено само един), посочени са няколко отговора с възможност за алтернативен избор между тях и има два бутона – "Гласувай" и "Виж резултатите". Всеки посетител на интернет страницата сам решава дали да отговори на зададения въпрос или не. Обикновено всеки посетител може да види резултатите от "гласуването" до момента, дори и ако не е отговорил на въпроса.

Основният въпрос е как да интерпретираме видяните от нас резултати, или иначе казано, върху кои съвкупности имаме право да разпростираме нашите изводи. Тук са възможни поне три различни ситуации:

Ситуация първа: най-коректно би било да отнасяме резултатите само към лицата отговорили на въпроса. В този случай не се налага използването на никакви статистически и/или математически методи. Твърденията, че определен процент **от отговорилите на въпроса** са посочили даден отговор, а определен процент – са посочили друг отговор и т.н. са напълно достатъчни без да е необходимо допълнително да им се придава тежест с помощта на формули и изчисления.

Ситуация втора: желанието е резултатите да се отнесат не само към отговорилите на въпроса, но и към всички посетили дадената интернет страница. В този случай отговорилите на въпроса са непредставителна извадка, получена по метода на отзовалите се от съвкупността на посетителите на страницата (Енциклопедичен речник по социология 1997: 158). Специфичното в случая е, че броят на посетителите е крайно число и е известен на администратора на страницата, макар че не винаги е известен на потребителя, гледащ резултатите от гласуването. Това означава, че администраторът (или поръчителят на анкетата) може да направи съответните статистически изводи относно интересуващите го отговори[12], докато за потребителите това не винаги е възможно.

Ситуация трета: желанието е резултатите да се отнесат към една по-широка съвкупност като например всички потребители на Интернет, цялото население на България, а защо не и всички българи, независимо в коя точка на света се намират[13]. В този случай отново имаме непредставителна извадка по метода на отзовалите се, но вече излъчена от някаква неопределена хипотетична съвкупност, която е достатъчно голяма[14].

По нататък в статията ще разглеждам само третата ситуация като ще се опитам да дам отговор на два въпроса:

- "Възможно ли е в тази ситуация да се направят статистически изводи относно хипотетичната съвкупност?"[15] и
- "Как решаването на статистическия проблем решава съдържателния проблем за допустимостта на генерализирането на изводите?"

Статистическият проблем в третата ситуация се решава по същия начин както във втората, като към съответните формули се прилага граничен преход (Харалампиев 2004). Това става по различен начин в зависимост от вида на признака – количествен или качествен.

---

* Статия, публикувана в списание „Социологически проблеми", брой 3-4/2004

[12] По-подробно техниката за правенето на тези статистически изводи е описана в първите две глави на книгата "Нетрадиционен поглед към традиционни статистически проблеми" (Харалампиев 2004).

[13] В случая става въпрос за български интернет страници, но тази ситуация е валидна за каквито и да е страници. Още повече, че за страниците, които са на езици с международна употреба като английски, испански, френски и др. съвкупността може да се разшири до всички англоговорящи, всички испаноговорящи, всички френскоговорящи и т.н.

[14] Терминът "достатъчно голяма" също е неопределен, но тук се използва само за да укаже, че имаме право да използваме математическия инструмент на граничен преход.

[15] Проблемът е в това, че класическата статистическа теория отговаря отрицателно на този въпрос. В статията е направен опит да се покаже, че въпросът има и положителен отговор.

1. Количествени признаци

Когато се изследват количествени признаци най-често се изчисляват средни аритметични.

Тъй като е много трудно да се намери пример за количествен признак в анкетите в Интернет, илюстрацията ще бъде направена на базата на признак, който е на бална скала.

И така, въпросът и неговите отговори, така както са дадени в Интернет (http://anketi.abv.bg, 24.07.2002 година, 10:06 часа) е:

"Каква оценка бихте поставили на едногодишното управление на НДСВ и Симеон Сакскобургготски?

Отличен (5,50-6): 191
Много добър (4,50-5,50): 198
Добър (3,50-4,50): 355
Задоволителен (2,50-3,50): 615
Слаб (2): 1442
1 - преписаха предишните: 454
Общо гласували: 3255"

За целите на по-нататъшната работа да подредим тези резултати в следната таблица:

Таблица 1.

Разпределение на единиците в извадката по отговорите на въпроса "Каква оценка бихте поставили на едногодишното управление на НДСВ и Симеон Сакскобургготски?"[16]

Отговори	Групови интервали	Среди на интервалите (x)	Честоти (f)	$x.f$
Слаб	Над 1,50 до 2,50	2	1442	2884
Задоволителен	Над 2,50 до 3,50	3	615	1845
Добър	Над 3,50 до 4,50	4	355	1420
Много добър	Над 4,50 до 5,50	5	198	990
Отличен	Над 5,50	6	191	1146
Общо			2801	8285

Средната оценка на управлението на правителството е:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i f_i}{\sum_{i=1}^{m} f_i} = \frac{8285}{2801} = 2{,}96\,,$$

където m е броят на различните значения на признака.

И така, можем ли да кажем "Народът писа "Задоволителен"[17] (2,96) на правителството"?.

От статистическа гледна точка нещата стоят така: средната на хипотетичната съвкупност се изменя непрекъснато в диапазона от най-малкото до най-голямото значение на признака, т.е. от 2,00 до 6,00 и има приблизително нормално разпределение с център:

$$\bar{\mu} = \frac{x_{\min} + x_{\max}}{2} = \frac{2+6}{2} = \frac{8}{2} = 4{,}00$$

и разсейване:

[16] Отговорът "1 – преписаха предишните" е изключен, тъй като той по същество е извън скалата за оценяване на управлението на правителството.

[17] Запазена е оригиналната терминология от интернет страницата.

$$\sigma_{\mu} = \frac{x_{\max} - x_{\min}}{2} \cdot \frac{1}{\sqrt{3.(m-1)}} = \frac{6-2}{2} \cdot \frac{1}{\sqrt{3.(5-1)}} = \frac{4}{2} \cdot \frac{1}{\sqrt{3.4}} = \frac{1}{\sqrt{3}} = 0{,}5774$$ (Харалампиев 2004).

Използвайки таблица за нормално разпределение може да се построи доверителеният интервал на средния бал[18]:

(1) $P(\bar{x} - \Delta_{\bar{x}} < \mu < \bar{x} + \Delta_{\bar{x}}) = P,$

където $\Delta_{\bar{x}}$ е максималната грешка, μ е неизвестната средна аритметична в хипотетичната съвкупност, а P е гаранционната вероятност.

Работата започва с определянето на максимално възможната максимална грешка:

$$\Delta_{\bar{x},\max} = \min[(\bar{x} - x_{\min});(x_{\max} - \bar{x})] = \min[(2{,}96 - 2{,}00);(6{,}00 - 2{,}96)] =$$

$$= \min(0{,}96;3{,}04) = 0{,}96$$

Следователно най-широкият възможен доверителен интервал, симетричен относно изчислената средна, е:

$$2{,}96 - 0{,}96 < \mu < 2{,}96 + 0{,}96$$

$$2{,}00 < \mu < 3{,}92$$

Неговата гаранционна вероятност се получава с помощта на таблица за нормално разпределение:

$$P(2{,}00 < \mu < 3{,}92) = P(\mu < 3{,}92) = 0{,}4449$$

Ако искаме гаранционната вероятност да бъде P=0,95, доверителният интервал трябва да се разшири докато се получи доверителен интервал с желаната гаранционна вероятност. Това разширяване може да стане само надясно[19]. Окончателният доверителен интервал е:

$$P(2{,}00 < \mu < 4{,}95) = P(\mu < 4{,}95) = 0{,}9500$$

И така, писа ли народът "Задоволителен" на управлението на правителството или не?

Видно е, че дори първият доверителен интервал е достатъчно широк и включва в себе си значенията "Слаб", "Задоволителен" и "Добър". Можем, разбира се, да го стесним, но това ще доведе до допълнително намаляване на гаранционната вероятност, която и без това е малка. Ако стесним доверителния интервал, например, до границите от 2,50 до 3,42 (така че да се запази симетрията) ще получим:

$$P(2{,}50 < \mu < 3{,}42) = 0{,}1529,$$

т.е. вероятността на твърдението "Народът писа "Задоволителен" на управлението на правителството" е едва 0,1529.

От друга страна доверителният интервал с гаранционна вероятност P=0,95 е пределно широк и включва в себе си значенията "Слаб", "Задоволителен", "Добър" и "Много добър".

Изводът от всичко казано до тук е, че категоричното твърдение е крайно несигурно, а сигурното твърдение е толкова широко, че на практика няма познавателен смисъл.

2. Качествени признаци

При анализа на качествени признаци не могат да се изчисляват средни аритметични, тъй като значенията на признака се описват с някакви категории[20]. Това, което се прави е изчисляването на относителните дялове на всяко конкретно значение на признака.

---

[18] Чете се "Вероятността неизвестната средна аритметична да се намира в границите от … до … е …".

[19] За да не се нарушава симетрията разширяването би трябвало да стане и в двете посоки, но тъй като наляво от 2,00 няма значещи стойности, то разширяването се прави само в едната посока. По този начин се избягват безсмислени записи като $P(0{,}97 < \mu < 4{,}95) = 0{,}9500$, тъй като е ясно, че средната аритметична изобщо не може да попадне в интервала от 0,97 до 2,00.

[20] На практика често значенията на качествените признаци се шифрират с числа. Тези числа обаче, показват единствено различие между значенията на признака, без да измерват големината на това различие. Нещо повече - когато качественият

Нека разгледаме един пример. Въпросът е: "Притежавате ли мобилен телефон?" (http://www.mtel.bg, 07.11.2002 година, 17:43 часа). Отговорите са:

"Да, клиент съм на М-Тел	12032	74,00%
Да, ползвам други оператори	1881	11,57%
Не, не притежавам	2346	14,43%
Общо гласували	16259"	

И така, можем ли да кажем, че 74,00% от всички българи са клиенти на М-Тел?

Нека първо да построим доверителните интервали на относителните дялове на трите отговора при гаранционна вероятност P=0,95.

Относителните дялове на всеки отговор в хипотетичната съвкупност се изменят непрекъснато в диапазона от 0 до 1 и имат следната функция на разпределение[21]:

(2) $$P(\pi_i < \pi_{ik}) = 1-(1-\pi_{ik})^{m-1},$$

където π_i е неизвестният относителен дял на i–тото значение на признака в хипотетичната съвкупност, а π_{ik} е конкретно число в границите от 0 до 1 (Харалампиев 2004: 67).

Самият доверителен интервал отново се получава по формула (1) като $\bar{x}$ се замести с p_i (изчисленият от данните относителен дял на i–тото значение на признака), а μ се замести с π_i.

Максимално възможната максимална грешка се получава по формулата:

(3) $$\Delta_{p_i,\max} = \min[p_i;(1-p_i)]$$ (Харалампиев 2004)

И така:

- за отговора "Да, клиент съм на М-Тел":

$$p_1 = \frac{12032}{16259} = 0{,}7400$$

$$\Delta_{p_1,\max} = \min[0{,}7400;(1-0{,}7400)] = \min(0{,}7400;0{,}2600) = 0{,}2600$$

$$P(0{,}7400-0{,}2600 < \pi_1 < 0{,}7400+0{,}2600) = P(0{,}4800 < \pi_1 < 1{,}0000) =$$

$$= P(0{,}4800 < \pi_1) = (1-0{,}4800)^{3-1} = 0{,}5200^2 = 0{,}2704$$

Така получената гаранционна вероятност е по-малка от предварително избраната от нас. Това означава, че доверителният интервал трябва да се разшири докато се получи доверителен интервал с желаната гаранционна вероятност. Това разширяване може да стане само наляво.

Според формула (2) търсим π_{1k} така, че:

$$P(\pi_1 > \pi_{1k}) = 1-P(\pi_1 < \pi_{1k}) = 1-1+(1-\pi_{1k})^2 = 0{,}95$$

$$1-\pi_{1k} = \sqrt{0{,}95}$$

$$\pi_{1k} = 1-\sqrt{0{,}95} = 0{,}0253$$

В крайна сметка доверителният интервал е:

$$P(0{,}0253 < \pi_1 < 1{,}0000) = 0{,}9500$$

- за отговора "Да, ползвам други оператори":

$$p_2 = \frac{1881}{16259} = 0{,}1157$$

признак е неподредим, е възможно неговите значения да бъдат подреждани по различни начини, което от своя страна означава и различни подредби на съответните шифри. Това от своя страна води до несъстоятелността на средната аритметична най-малкото по две причини - първо, защото е средна от шифрите, а не от значенията на признака, и второ, защото при едни и същи изходни данни, различните подредби на значенията на признака ще доведат до различни подредби на шифрите, а оттам и до различни средни аритметични. От друга страна, когато качественият признак е подредим, е възможна само една единствена подредба на неговите значения, което прави практически възможно (макар и не съвсем теоретически коректно) изчисляването на средна аритметична. Именно такъв пример беше разгледан в точка 1.

[21] Чете се "Вероятността неизвестният относителен дял на i–тото значение на признака да е по-малък от … е …".

$$\Delta_{p_2,\max} = \min[0{,}1157;(1-0{,}1157)] = \min(0{,}1157;0{,}8843) = 0{,}1157$$

$$P(0{,}1157 - 0{,}1157 < \pi_2 < 0{,}1157 + 0{,}1157) = P(0{,}0000 < \pi_2 < 0{,}2314) = 0{,}4093$$

Така получената гаранционна вероятност е по-малка от предварително избраната от нас. Това означава, че доверителният интервал трябва да се разшири докато се получи доверителен интервал с желаната гаранционна вероятност. Това разширяване може да стане само надясно.

Според формула (2) търсим π_{2k} така, че:

$$P(\pi_2 < \pi_{2k}) = 1 - (1 - \pi_{2k})^2 = 0{,}95$$

$$(1 - \pi_{2k})^2 = 1 - 0{,}95 = 0{,}05$$

$$1 - \pi_{2k} = \sqrt{0{,}05}$$

$$\pi_{2k} = 1 - \sqrt{0{,}05} = 0{,}7764$$

В крайна сметка доверителният интервал е:

$$P(0{,}0000 < \pi_2 < 0{,}7764) = 0{,}9500$$

- за отговора “Не, не притежавам”:

$$p_3 = \frac{2346}{16259} = 0{,}1443$$

$$\Delta_{p_3,\max} = \min[0{,}1443;(1-0{,}1443)] = \min(0{,}1443;0{,}8557) = 0{,}1443$$

$$P(0{,}1443 - 0{,}1443 < \pi_3 < 0{,}1443 + 0{,}1443) = P(0{,}0000 < \pi_3 < 0{,}2886) = 0{,}4939$$

Така получената гаранционна вероятност е по-малка от предварително избраната от нас. Аналогично на предходния случай се получава:

$$P(0{,}0000 < \pi_3 < 0{,}7764) = 0{,}9500$$

Да обобщим получените резултати.

Таблица 2.

Доверителни интервали на относителните дялове

Относителни дялове	Доверителен интервал, симетричен относно изчисления от данните относителен дял	Доверителен интервал при гаранционна вероятност P=0,95
π_1	$P(0{,}4800 < \pi_1 < 1{,}0000) = 0{,}2704$	$P(0{,}0253 < \pi_1 < 1{,}0000) = 0{,}9500$
π_2	$P(0{,}0000 < \pi_2 < 0{,}2314) = 0{,}4093$	$P(0{,}0000 < \pi_2 < 0{,}7764) = 0{,}9500$
π_3	$P(0{,}0000 < \pi_3 < 0{,}2886) = 0{,}4939$	$P(0{,}0000 < \pi_3 < 0{,}7764) = 0{,}9500$

Оказва се, че и при качествените признаци положението е същото както при количествените – тесните доверителни интервали са с малки гаранционни вероятности, а доверителните интервали с достатъчно голяма гаранционна вероятност, са толкова широки, че на практика нямат познавателен смисъл.

Нека сега отново да зададем въпроса дали 74,00% от българите са клиенти на М-Тел. За щастие в случая знаем отговора, защото в съобщение за медиите на М-Тел от 30.09.2002 година е казано, че “всеки пети българин общува чрез М-Тел” (http://www.mobiltel.bg, 07.11.2002 година). Това означава, че приблизително $\pi_1 = \frac{1}{5} = 0{,}2000$, т.е. приблизително 20,00% от българите са клиенти на М-Тел. Видно е, че тази стойност не принадлежи на симетричния доверителен интервал, който обаче е с много малка гаранционна вероятност (едва 0,2704), и в същото време принадлежи на доверителния интервал, който е с гаранционна вероятност P=0,95. Освен това е видно и голямото разминаване

между резултата от анкетата ($p_1 = 0{,}7400$) и действителната стойност на параметъра ($\pi_1 = 0{,}2000$).

И така, в случая по-категоричното[22] твърдение се оказва невярно. От друга страна, доверителният интервал, получен при гаранционна вероятност *P*=0,95 е толкова широк, че може да означава практически всичко, включително и разлика от 54 пункта между резултата от анкетата и действителната стойност.

В резултат на всичко казано до тук може да се направят следните изводи:

1) *при непредставителните извадки по принцип и при разглеждания тип анкети в Интернет в частност, могат да се правят статистически изводи отнасящи се за неопределени, достатъчно големи хипотетични съвкупности;*

2) решаването на статистическия проблем не решава съдържателния проблем за допустимостта на генерализирането на изводите, тъй като единствено дава поредното доказателство, че "… по информацията, получена от "отзовалите се", т.е. на пожелалите да вземат участие в анкетата, не могат да се правят обобщени изводи и оценки за мнението на всички лица, към които е била насочена анкетата" (Съйкова, Чакалов 1977: 31).

Накрая бих искал да прехвърля мост към един друг тип анкети. Става въпрос за ситуацията, при която водещ на телевизионно предаване задава въпрос в ефир, например: "Трябва ли треньорът на националния отбор по футбол да подаде оставка?". На зрителите се дават два телефонни номера и всички, които желаят да отговорят с "Да" звънят на единия, а желаещите да отговорят с "Не" звънят на другия. Като изключим скритата цел за установяване на рейтинга на самото предаване, по всичко останало тези анкети са еднакви с разглежданите по-горе в статията. И така, ако въпросният водещ твърди, че определен процент **от обадилите се** са отговорили с "Да", а определен процент – с "Не", в това няма никакъв проблем. Ако обаче твърдението е, че определен процент **от зрителите** или пък определен процент **от българите** смятат, че треньорът на националния отбор по футбол трябва да си подаде оставката, то или верността на това твърдение е гарантирана с прекалено малка вероятност, или обратното – вероятността е голяма, но коректно изказано (като интервал, а не като едно число) твърдението би било толкова широко, че би могло да означава практически всичко. Така или иначе формулировката, отнасяща се до всички зрители или до всички българи просто е заблуждаваща.

ЛИТЕРАТУРА

Енциклопедичен речник по социология. 1997. "M&M" Михаил Мирчев, София.

Съйкова, И., Б. Чакалов. 1977. *Методология и методика на социологическите изследвания*. Наука и изкуство, София

Харалампиев, К. 2004. Нетрадиционен *поглед към традиционни статистически проблеми*. Балкани, София.

[22] Макар че интервал от 48,00% до 100,00% също е достатъчно широк.

Телевизионните гласувания по телефона – проблемът за „победителя”*

Тази статия е своеобразно продължение на статията „Анкетите в Интернет – възможност за статистически изводи и интерпретиране на резултатите” (Харалампиев 2004а). Там се отговаряше на въпроса върху кои съвкупности могат да се разпростират резултатите от анкетите в Интернет.

В настоящата статия ще бъде разгледано отговарянето по телефона на въпросите, задавани в ефира на различни телевизии. На практика се срещат поне две различни ситуации:

В единия случай става дума за зададен (обикновено само един) въпрос с предварително формулирани отговори, за които са дадени различни телефонни номера. За отговарящия е достатъчно само да набере съответния номер или да изпрати SMS, без да е необходимо устно да отговаря на въпроса. Пример за такава ситуация са редовно задаваните въпроси в шоуто „Сблъсък”, където обадилите се подкрепят едната или другата страна в сблъсъка.

Втората ситуация е класирането на участници[23] в различни телевизионни шоу-игри. В този случай за всеки участник има отделен телефонен номер и обадилите се избират „победителя”[24]. Примери за такива телефонни класации са „Big Brother”, „Star Academy”, „Вот на доверие”, определянето на българската песен за конкурса на Евровизия.

Въпреки че тук не става въпрос за анкети в Интернет, проблемът е принципно същият – върху кои съвкупности могат да се разпростират резултатите от гласуването. Общата тенденция във всички гореизброени предавания е резултатите да се разпростират върху съвкупността на **зрителите**. От гледна точка на конкретните предавания това е разбираемо – всяко предаване е или вътрешна продукция на съответната телевизия, или има сключен договор с нея за външна продукция. В този смисъл във взаимен интерес и на предаването, и на телевизията е непрекъснато да се напомня в ефир, че именно **техните** зрители са направили избора. На практика обаче **обадилите се** са непредставителна извадка (получена по метода на отзовалите се) от генералната съвкупност на **зрителите**.

Начинът на правене на статистически изводи в този случай е същият като описания в горецитираната статия (Харалампиев 2004а) и затова тук няма да бъде показван отново. Искам обаче да направя следваща крачка и да разгледам един специфичен проблем, който възниква при телевизионните гласувания по телефона – проблемът за „победителя”.

Във всички гореизброени предавания отговорът или участникът, събрал най-много гласове е обявяван за „победител”[25] на зрителите. В действителност той е „победител” на обадилите се. Дали обаче „победителят” на обадилите се е и победител на зрителите?

Необходимото условие един отговор или един участник да има най-много привърженици в генералната съвкупност се дава със системата[26]:

$$(1)\qquad \begin{cases} \pi_1 > \pi_2 \geq \pi_3 \geq \ldots \geq \pi_m \geq 0 \\ \sum_{i=1}^{m} \pi_i = 1 \end{cases},$$

където π_i е относителният дял на *i*–тия отговор или на *i*–тия участник, а m ($m \geq 2$) е броят на всички отговори или на всички участници.

Решаването на тази система води до следния резултат:

---

* Статия, публикувана в списание „Социологически проблеми”, брой 3-4/2005

[23] Тук и навсякъде по-нататък в текста под „участник” се избира лице, което участва в съответното телевизионно предаване, а не обадило се лице, което дава гласа си по телефона.

[24] „Победител” е в кавички, по две причини. Първо, защото участникът, получил най-много гласове, може да е само финалист, а не победител („Вот на доверие”) или дори да е губещ („Big Brother” и „Star Academy”). Второ, и по-важно, в статията ще бъде показано, че „победителят” в извадката и победителят в генералната съвкупност може и да не е един и същ!

[25] Както стана ясно в шоу-игрите „Big Brother” и „Star Academy” „победителят” всъщност е губещ.

[26] Тук относителните дялове са подредени в низходящ ред, което е често срещана практика при различните класации.

(2) $$\pi_1 > \frac{1}{m},$$

т.е. необходимото условие един отговор или един участник да е „победител" в генералната съвкупност е той да има повече от една *m*–та част от гласовете.

Но необходимо условие още не означава и достатъчно условие[27]. Това просто означава, че отговор или участник, който има по-малко от една *m*–та част от гласовете в генералната съвкупност **не може** да бъде „победител", но ако има повече от една *m*–та част от гласовете в генералната съвкупност това не му гарантира автоматично „победата". В този смисъл необходимото условие е условие за **възможна** победа, а не условие за **сигурна** победа.

Вероятността да е изпълнено условието (2) при непредставителни извадки се дава с формулата:

(3) $$P\left(\pi_1 > \frac{1}{m}\right) = \left(1 - \frac{1}{m}\right)^{m-1} = \left(\frac{m-1}{m}\right)^{m-1}$$ (Харалампиев 2004б: 67)

Трябва да се отбележи, че тази вероятност зависи само от броя на отговорите или на участниците и се изменя в диапазона от 36,8%[28] до 50,0%. При два отговора или двама участника вероятността е точно 50,0% и с нарастването на броя на отговорите или на участниците вероятността се доближава до долната си граница. Всичко това означава, че вероятността конкретен отговор или конкретен участник **да бъде възможен „победител"** в генералната съвкупност е под или най-много равна на 50,0%, следователно вероятността конкретен отговор или конкретен участник **да не бъде „победител"** е най-малко 50%.

Във връзка с всичко казано до тук трябва да се изтъкне още един факт – съществува мнението, че проблемът всъщност е в това, че един зрител може да се обади неограничен брой пъти. Когато резултатите от гласуването са разпростират върху цялата генерална съвкупност обаче, това мнение е погрешно. В (Харалампиев 2004б: 66-67) е показано, че при големи генерални съвкупности (каквито са телевизионните аудитории) се получават едни и същи резултати, независимо от това дали подборът на единиците, попаднали в извадката, е бил възвратен или безвъзвратен. Следователно проблемът не е в това дали едно лице ще гласува само един път или повече пъти. Същинският проблем е в начина, по който е направен подборът на гласуващите и по-точно в пълната липса на подбор[29].

Нека сега да разгледаме поотделно горепосочените примери.

В предаването „Сблъсък" има две споречи страни и обадилите се подкрепят едната или другата. В такъв случай формула (3) придобива вида:

$$P\left(\pi_1 > \frac{1}{2}\right) = \left(\frac{2-1}{2}\right)^{2-1} = \frac{1}{2} = 0{,}500$$

Вече беше посочено, че когато *m*=2 необходимото условие е и достатъчно, т.е. вероятността едната от двете страни да е сигурен победител е 50,0%. Дали вероятност 50,0% е голяма или малка е въпрос на субективна преценка, но само ще отбележа, че вероятността извадката правилно да е определила победителя в генералната съвкупност е равна на вероятността другата страна всъщност да има по-голяма подкрепа.

В шоу-игрите „Big Brother" и „Star Academy" изборът е между двама номинирани за изгонване участници, следователно изводите не се различават от предходния пример. Понякога обаче се случва номинираните да са трима. Тогава формула (3) придобива вида:

---

[27] Необходимото условие е и достатъчно само когато изборът е между два отговора или между двама участника, т.е. *m*=2.

[28] $0{,}3678794 = \frac{1}{e}$, където *e* е т.нар. неперово число.

[29] Всъщност липсва „изследовател" (продуцент, сценарист, режисьор, водещ или някои друг), който да формира представителна извадка от лица, които следва да гласуват, а инициативата е оставена на зрителите, които сами решават дали да се обадят или не.

$$P\left(\pi_1 > \frac{1}{3}\right) = \left(\frac{3-1}{3}\right)^{3-1} = \left(\frac{2}{3}\right)^2 = 0{,}444,$$

т.е. вероятността конкретен участник да е възможен „победител" в генералната съвкупност е 44,4%, следователно вероятността да не е „победител" в генералната съвкупност е 55,6%. Видно е, че в този случай вероятността извадката правилно да е определила „победителя" в генералната съвкупност е по-малка от вероятността някой от другите двама участници всъщност да има по-голям относителен дял.

В шоу-играта „Вот на доверие" изборът е между петима участници, следователно формула (3) придобива вида:

$$P\left(\pi_1 > \frac{1}{5}\right) = \left(\frac{5-1}{5}\right)^{5-1} = \left(\frac{4}{5}\right)^4 = 0{,}410,$$

т.е. вероятността конкретен участник да е възможен победител в генералната съвкупност е 41,0%, следователно вероятността да не е победител в генералната съвкупност е 59,0%. Видно е, че в този случай вероятността извадката правилно да е определила победителя в генералната съвкупност е около един път и половина по-малка от вероятността някой от другите четирима участници всъщност да е победител.

При определянето на българската песен за конкурса на Евровизия изборът е между 12 песни, следователно формула (3) придобива вида:

$$P\left(\pi_1 > \frac{1}{12}\right) = \left(\frac{12-1}{12}\right)^{12-1} = \left(\frac{11}{12}\right)^{11} = 0{,}384,$$

т.е. вероятността конкретна песен да е възможен победител в генералната съвкупност е 38,4%, следователно вероятността да не е победител в генералната съвкупност е 61,6%. Видно е, че в този случай вероятността извадката правилно да е определила победителя в генералната съвкупност е също около един път и половина по-малка от вероятността някоя от другите 11 песни всъщност да е победител.

В текста до тук не стана дума за „Шоуто на Слави" и проведената там "Музикална ку-ку академия". Това е така, защото там обадилите се не избират победител между отделните участници, а ги оценяват по шестобална скала. В този случай определянето на необходимото и достатъчното условие за победа в генералната съвкупност не е по силите на автора. Тази техническа трудност изключи „Музикална ку-ку академия" от изследването, но трябва да се отбележи, че и в този случай обадилите се са непредставителна извадка от генералната съвкупност на зрителите и следователно разпростирането на резултатите върху съвкупността на зрителите е некоректно (Харалампиев 2004а: 210).

И така, да обобщим. При разпростирането на резултатите от телефоните гласувания върху генералната съвкупност на всички зрители, вероятността извадката правилно да определи „победителя" в генералната съвкупност е по-малка или най-много равна на вероятността извадката погрешно да определи „победителя" в генералната съвкупност.

ЛИТЕРАТУРА

Харалампиев, К. 2004а. Анкетите в Интернет – възможност за статистически изводи и интерпретиране на резултатите. *Социологически проблеми*, 3-4: 203-211

Харалампиев, К. 2004б. Нетрадиционен *поглед върху традиционни статистически проблеми*. Балкани, София

Още една гледна точка към проблема за отказите при социологически изследвания*

Напоследък[30], покрай актуалните в момента електорални изследвания, отново придобива актуалност въпросът колко е допустимият дял на отказите при социологическите изследвания[31]. Моят преглед на литературата показа, че няма точен отговор на този въпрос. Публикациите по темата за отказите при социологическите изследвания основно могат да се разделят на две групи – в едната група попадат тези публикации, които разглеждат причините за отказите (Атанасов и кол. 2006: 40-46; Дерменджиева, Цветкова, Милева 2010).

В другата група са публикациите, които разглеждат последиците от отказите, а именно: „ако съществуват систематични отклонения в мненията на отговорилите и неотговорилите, ниският процент на участие очевидно създава значителен риск от *изместване* на резултатите" (Парчев 1998: 113). И още „Възможно е изобщо да няма изместване (ако мненията на отговорилите и неотговорилите по зададените въпроси не се различават), или пък да има значително изместване. Тук статистическата теория не може да ни помогне. Възможни са само позоваванията, при това не напълно убедителни, на миналия опит" (пак там).

Аз обаче смятам, че статистическата теория може да помогне доста, но не „класическата" статистическа теория (която наистина не може да помогне), а една друга парадигма в статистиката, позната като бейсовска статистика.

Тук няма да навлизам в дълбочина в разликите между двете парадигми[32], само ще отбележа, че различията се отнасят единствено до начина на правене на изводи, базирани на извадки. Важно последствие от тези различия обаче е, че бейсовската статистика позволява да се изчисляват грешки и при непредставителни извадки.

А в конкретния случай реализираната извадка може да се разглежда като непредставителна извадка от планираната, получена по метода на отзовалите се. Тогава основен става въпросът доколко реализираната извадка достатъчно добре възпроизвежда планираната извадка. Или иначе казано, каква е грешката, но не грешката на реализираната извадка спрямо цялата генерална съвкупност, а грешката на реализираната спрямо планираната извадка. Основното допускане е, че планираната извадка е формирана по всички правила за правене на представителни извадки. Тогава, ако относителните дялове в реализираната извадка се различават малко (т.е. грешката е малка) спрямо планираната, то реализираната извадка ще е достатъчно добра.

В предишна моя публикация съм показал как се правят изводи за относителни дялове, базирани на непредставителни извадки, в два случая – на малка генерална съвкупност (Харалампиев 2004: 54-58) и на голяма генерална съвкупност (Харалампиев 2004: 66-68 и 71-79). Особеното на втория случай е, че имплицитно се допуска, че делът на извадката спрямо генералната съвкупност е пренебрежимо малък. Това допускане е напълно оправдано при всички социологически изследвания, тъй като обемите на извадката обикновено са от порядъка на (няколко) хиляди, а обемите на генералната съвкупност – от порядъка на (няколко) милиона, така че делът на извадката наистина е пренебрежимо малък.

---

* Статия, публикувана в списание „Социологически проблеми", брой 1-2/2005

[30] Виж например отговорите на Цветозар Томов в съвместното му интервю с Андрей Райчев в „Гласове" (http://www.glasove.com/erata-borisov--intervyu-s-andrey-raychev-i-tsvetozar-tomov--ii-chast-16287) или коментарите на Стойко Тонев от 29 септември 2011 в „Делниците на един луд" (http://frognews.bg/news_39164/CHetvarti_kilometar/ или http://www.reduta.bg/?p=1440).

[31] В тази статия няма да се спирам на проблема със заместването на отказалите. Професор Венедиков винаги е настоявал, че заместването е недопустимо, но в неговите публикации не успях да открия никъде експлицитно заявено това твърдение. Това което открих е, че трябва „да наблюдаваме точно тези единици, които са избрани чрез лотарията" (Венедиков 1992: 43; Венедиков 1993: 40). Разбира се, от гледна точка на езика на математиката твърдението „недопустимо е заместването на единици" е пряко следствие от твърдението „наблюдават се точно тези единици, които са избрани". От своя страна аз, в една публицистична статия, писана по повод на предходните местни избори през 2007 година, съм показал защо заместването е опасно и че е недопустимо да се прави (Харалампиев 2007: 16). Това че заместването на единици е (масова) практика е въпрос на друг разговор.

[32] Основните разлики съм описал в предишна моя публикация (Харалампиев 2008: 146-153).

За конкретния проблем, който се разисква в настоящата статия, очевидно нито първия, нито втория случай е подходящ, защото делът на реализираната спрямо планираната извадка не е пренебрежимо малък. Затова трябва да се разработи трети случай – на голяма генерална съвкупност, при която делът на извадката не е пренебрежимо малък.

Ако означим с n обема на планираната извадка, а с $\hat{n}$ обема на реализираната извадка, то делът на реализираната спрямо планираната извадка (т.нар. response rate) ще бъде $\frac{\hat{n}}{n}$, а $1-\frac{\hat{n}}{n}$ ще е делът на отказите.

Ако се модифицират съответните формули (Харалампиев 2004: 54-55), се получава следното:

Минималната възможна стойност на относителния дял в планираната извадка е:

(1) $$p_{\min} = \hat{p}.\frac{\hat{n}}{n},$$

където:

p е относителният дял на интересуващото ни значение на изследвания признак в планираната извадка;

$\hat{p}$ е относителният дял на интересуващото ни значение на изследвания признак в реализираната извадка.

Максималната възможна стойност на относителния дял в планираната извадка е:

(2) $$p_{\max} = \hat{p}.\frac{\hat{n}}{n}+1-\frac{\hat{n}}{n}$$

Още тук можем да направим първия извод: целият диапазон на възможните стойности на относителния дял в планираната извадка е:

(3) $$p_{\max} - p_{\min} = 1-\frac{\hat{n}}{n}$$

Тоест делът на отказите съвпада с диапазона на възможните стойности. Иначе казано, ако делът на отказите е 5%, тогава диапазонът на възможните стойности ще бъде 5 процентни пункта, а ако делът на отказите е 50%, тогава диапазонът на възможните стойности ще бъде 50 процентни пункта.

Този извод обаче е твърде консервативен. Това е така, защото различните стойности в диапазона не са еднакво вероятни. Затова може да се построи доверителният интервал на относителния дял в планираната извадка и да се търси връзка между неговата ширина и делът на отказите.

Първата стъпка за построяването на доверителен интервал е да се определи т.нар. функция на разпределение. Затова продължаваме с модифицирането на формулите от цитираната публикация (Харалампиев 2004: 55):

(4) $$P(p \le x) = F(x) = \lim_{n\to\infty}\left(1-\frac{C^{m-1}_{n-\hat{n}+m-nx+\hat{n}\hat{p}-1}}{C^{m-1}_{n-\hat{n}+m-1}}\right) = 1-\left[\frac{1-x-\frac{\hat{n}}{n}(1-\hat{p})}{1-\frac{\hat{n}}{n}}\right]^{m-1},$$

където m е броят на значенията на изследвания признак.

Тъй като функцията на разпределение всъщност е вероятността интересуващия ни относителен дял в планираната извадка да не надхвърли дадена стойност, то доверителния интервал се получава като разлика от две функции на разпределение:

(5) $$P(x < p \le y) = F(y) - F(x) = 1-\alpha,$$

където α е рискът за грешка.

В конкретния случай обаче формула (5) не е приложима в този си вид. Това е така, защото формула (4) е функция на разпределение на т.нар. L-разпределение. Това е разпределение, чиято най-вероятна стойност е минималната възможна стойност и всяка следваща стойност е по-малко

вероятна от предходната. В този случай е по-подходящо доверителният интервал да се затвори само отгоре, т.е.:

(6) $$P(p \le x) = F(x) = 1 - \alpha$$

Ако функцията на разпределение от формула (4) се замести във формула (6) и полученото уравнение се реши спрямо x, се получава:

(7) $$x = \left(1 - \frac{\hat{n}}{n}\right)\left(1 - \sqrt[m-1]{\alpha}\right) + \hat{p}.\frac{\hat{n}}{n}$$

Тогава ширината на доверителния интервал ще бъде[33]:

(8) $$2\Delta = x - p_{\min} = \left(1 - \frac{\hat{n}}{n}\right)\left(1 - \sqrt[m-1]{\alpha}\right)$$

Тази формула е достатъчна за изчисляването на ширината на доверителния интервал при всяко конкретно изследване, и на тази основа, за оценка на това с колко относителният дял в реализираната извадка се различава от планираната.

Обаче формула (8) може да се пререши спрямо дела на отказите:

(9) $$1 - \frac{\hat{n}}{n} = \frac{2\Delta}{1 - \sqrt[m-1]{\alpha}}$$

Формула (9) е полезна, защото тя дава теоретичен отговор на въпроса колко е допустимия размер на дела на отказите при предварително зададени от нас изисквания за ширината на доверителния интервал и за риска за грешка.

Тъй като в практиката на социологическите изследвания има общоприети стойности за размера на грешката и за риска за грешка, може да се направи една таблица, в която за различните стойности на m са изчислени допустимите дялове на отказите. Общоприетата стойност за размера на грешката е 3 процентни пункта[34], а за риска за грешка е 5%. Получените резултати са представени в таблица 1.

Таблица 1: *Допустими размери на дела на отказите при различен брой на значенията на изследвания признак*

Брой на значенията на изследвания признак	Допустим размер на дела на отказите
2	6,3
3	7,7
4	9,5
5	11,4
6	13,3
7	15,3
8	17,2
9	19,2
10	21,2

[33] При представителни извадки разпределението на относителния дял в извадката е нормално и доверителният интервал се получава като към оценката на относителния дял се прибавя и се изважда максималната грешка (Δ). Затова ширината на доверителния интервал е 2Δ. В нашия случай разпределението не е нормално и доверителния интервал не се определя като оценката плюс/минус максималната грешка, но означението 2Δ е запазено за удобство.

[34] Тук напълно се солидаризирам със становището на Ивайло Парчев, че тази стойност се е превърнала в „магическа" за практиката на социологическите изследвания (Парчев 1998: 121). Пак там е дадено много добро обяснение за това откъде е дошла и защо се е утвърдила тази „магическа" стойност.

Таблица 1 е направена до $m = 10$, защото рядко при социологически изследвания броя на значенията на изследваните признаци надхвърля 10. Но ако все пак това се случи, тогава заинтересованият читател може сам да изчисли допустимия размер на отказите по формула (9) или ширината на доверителния интервал по формула (8).

Във връзка с таблица 1 трябва да се акцентира на три важни неща:

Първо, ако в консервативната формула (3) заложим отново, че желаем диапазонът на възможните стойности да бъде равен на две максимални грешки, тогава допустимият дял на отказите ще се получи равен на 6,0%. Тоест, когато работим с доверителни интервали, а не с диапазона на възможните стойности, получаваме по-либерални оценки за допустимия дял на отказите.

Второ, обикновено в анкетната карта на всяко социологическо изследване има множество въпроси, всеки с различен брой отговори. Затова, когато определяме допустимия дял на отказите, трябва да използваме признака с най-малко на брой значения. И също така обикновено в анкетните карти се срещат дихотомни въпроси. А както се вижда от таблица 1, допустимият дял на отказите при дихотомните признаци съвсем малко се отличава от допустимия дял на отказите, получен на база на диапазона на възможните стойности.

Трето, числата в таблица 1 все пак трябва да се разглеждат повече като илюстрация, а не като твърд еталон. Това е така, защото те са изчислени при зададено изискване ширината на доверителния интервал да бъде шест процентни пункта. Но при конкретно изследване тази ширина може да се окаже или по-висока, или по-ниска от желаната. А това означава и, че праговете трябва съответно да се намалят или да се увеличат. Така например, при едно електорално изследване, за големите партии, които събират 20-30% подкрепа, доверителен интервал с ширина от шест процентни пункта може да е подходящ, но за малките партии, които събират 3-4% подкрепа, този доверителен интервал е напълно неприложим. За малките партии може би по подходящ е доверителен интервал с ширина от един процентен пункт. Но намаляването на ширината на доверителния интервал шест пъти автоматично намалява и допустимия дял на отказите също шест пъти!

При всяко социологическо изследване винаги ще има откази. Но резултатите, получени досега, дават възможност на изследователя да направи две неща. Първо, преди започване на теренното изследване може да изчисли по формула (9) допустимия дял на отказите и по време на теренната работа да удържа отказите в тези граници, и второ, след приключването на терена да изчисли по формула (8) диапазона, в който (най-вероятно) попада разликата в интересуващите ни относителни дялове между реализираната и планираната извадка.

ЛИТЕРАТУРА

Атанасов, А. и кол. 2006. *Електоралните изследвания. Изследователски проблеми и прогностични възможности*. Издателство „FABER", Велико Търново.

Венедиков, Й. 1992. *Статистика, социология и още нещо...* Издателство „Информационно обслужване", София.

Венедиков, Й. 1993. *Общественото мнение. Епистемологични проблеми*. Университетско издателство „Св. Климент Охридски", София

Дерменджиева, Б., В. Цветкова, Н. Милева. 2010. „Отказите от участия в изследвания: проблем на обществото?". В: *Благополучие и доверие: България в Европа?*, състав. Н. Тилкиджиев и Л. Димова. Издателство „Изток-Запад", София.

Парчев, И. 1998. *Избор на партия, избор на президент. Осем етюда върху една таблица*. Статистическо издателство и печатница при НСИ, София.

Харалампиев, К. 2004. *Нетрадиционен поглед върху традиционни статистически проблеми*. Издателство „Балкани", София.

Харалампиев, К. 2007. Агенциите рядко грешат умишлено. *Вестник „Кеш", 39: 16*

Харалампиев, К. 2008. „За парадигмите в статистиката – бейсовска статистика”. В: *Актуални проблеми на статистическата теория и практика*. Университетско издателство „Стопанство”, София.

ТРЕТИ РАЗДЕЛ.
ИЗВОДИ, ОСНОВАВАЩИ СЕ НА ПРЕДСТАВИТЕЛНИ ИЗВАДКИ

Представителните извадки са на практика основният източник на информация при изследванията в социалните науки. Както беше посочено, те са основният източник на информация и при изследванията в природните науки, макар че там това не се посочва изрично. По тази причина, в честотната статистика методите за правене на изводи, основаващи се на информация от представителни извадки са силно развити. При тривиалните, най-често срещани и най-често описвани в учебниците изследователски проблеми практически няма разлика между решенията на честотната и бейсовската статистика. Предимствата на бейсовската статистика са във възможността за включване на наличната априорна информация, в по-доброто разбиране на предимствата и рисковете от това включване и в решаването на нетривиални изследователски проблеми.

В този раздел първо ще бъде показано как от позициите на бейсовската статистика се решават тривиалните статистически проблеми за оценяване на относителен дял и на средна аритметична по информация от представителни извадки. След това ще бъде разгледан един нетривиален проблем, свързан с оценяването на относителен дял, който е отношение на два други взаимно зависими неизвестни относителни дяла.

<u>Шеста тема:</u> Статистически изводи и заключения относно относителни дялове в генералната съвкупност на базата на апостериорни вероятности

При представителните извадки данните са случайна величина, т.е. няма пречки за използването на извадковото разпределение. Следователно, може да се приложи теоремата на Бейс и да се получат апостериорните вероятности.

От тази тема ще научите:

- Какви стойности може да приема относителният дял в генералната съвкупност.
- Какъв е видът на извадковото разпределение при безвъзвратен и при възвратен подбор.
- Какъв е видът на разпределението на апостериорните вероятности на относителния дял в генералната съвкупност.
- Как се построяват доверителни интервали и се проверяват статистически хипотези относно относителният дял в генералната съвкупност.
- Как големината на генералната съвкупност влияе върху крайните изводи.
- Как се определя най-вероятното разпределение на единиците в генералната съвкупност.

<u>Седма тема:</u> Статистически изводи и заключения относно важни параметри на разпределенията (средни аритметични, стандартни отклонения и т.н.) в генералната съвкупност на базата на апостериорни вероятности

При представителни извадки изводите относно средната аритметична и другите числови характеристики на количествените признаци се базират на апостериорни вероятности.

От тази тема ще научите:

- Какви стойности може да приема средната аритметична в генералната съвкупност.
- Какъв е видът на извадковото разпределение при безвъзвратен и при възвратен подбор.
- Какъв е видът на разпределението на апостериорните вероятности на средната аритметична в генералната съвкупност.
- Как се построяват доверителни интервали и се проверяват статистически хипотези относно средната аритметична в генералната съвкупност.
- Как големината на генералната съвкупност влияе върху крайните изводи.

<u>Осма тема:</u> Приложения в областта на социалните науки – представителни извадки

Приложенията в областта на социалните науки се отнасят до анализирането на данни от:

- Представителни извадки, излъчени чрез безвъзвратен подбор от сравнително малка генерална съвкупност.
- Представителни извадки, излъчени чрез възвратен подбор от сравнително малка генерална съвкупност.
- Представителни извадки, излъчени от достатъчно голяма генерална съвкупност, независимо от вида на подбора. Като пример ще бъдат разгледани предизборните проучвания. Допълнително ще бъде разгледан проблемът за процента спрямо действителните гласове.

ТЕКСТ КЪМ ТРЕТИ РАЗДЕЛ

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

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

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

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

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

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

---

* Превод на статия, публикувана в Годишник на Софийски университет „Св. Климент Охридски", Философски факултет, книга „Социология", том 99, 2008

[35] Или теоретичното.

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:

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

Когато извадката е представителна и генералната съвкупност е достатъчно голяма[36], относителните дялове са непрекъснати и извадковото разпределение е *полиномно* (Jaynes 1993: 315, 317-318):

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

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 $\left(-\sum_k P(H_k \mid I) \log P(H_k \mid I) \text{ или } -\int_R P(H_k \mid I) \log P(H_k \mid I)\right)$, 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).

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:

---

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

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

(6)

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

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

Следователно апостериорната вероятност е:

(7)

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

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:

$$P(\pi_i = x \mid DI) = \int\int \ldots \int_{R_2} P(H_k \mid DI) d\pi_{1k} d\pi_{2k} \ldots d\pi_{i-1,k} d\pi_{i+1,k} \ldots d\pi_{mk} =$$

(8)

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

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

Полученият резултат[37] 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*:

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

(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. 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

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[38]:

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

[38] Всички интеграли са изчислени числово.

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[39].

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:

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

тъй като

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

Както се вижда, относителният дял на хората, които няма да гласуват участва в знаменателя на формула (12). Това означава, че и числителят и знаменателят са неизвестни параметри[40]. 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[41].

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:

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

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

---

[39] 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.

[40] 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.

[41] Това е следствие от формула (4).

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

$$F(u) = P\left(\frac{x}{1-y} < u \middle| DI\right) = \int_0^1 \left\{ \int_0^{u(1-y)} P[(\pi_i = x)(\pi_m = y) \mid DI]dx \right\} dy = \ldots = \tag{15}$$

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

$$P\left(\frac{x}{1-y} = u \middle| DI\right) = f(u) = F'(u) = \tag{16}$$

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

Полученият резултат е още едно бета разпределение. Чрез него можем да проверим хипотезата, че дадена партия ще премине четирипроцентовата бариера[42]. Също така можем да построим доверителните интервали на относителните дялове на хората, които биха избрали дадена партия[43], 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	$P\left(0{,}036 < \frac{\pi_6}{1-\pi_m} < 0{,}067 \middle| DI\right) = 0{,}95$

[42] Или хипотезата, че даден кандидат-президент ще спечели президентските избори.
[43] Или даден кандидат за президент.

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 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:

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

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

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

Ако извадките на двете изследвания са независими[44], извадковото разпределение е:

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

Маргинализирайки числителя на формула (18) получаваме:

$$P(D_2 \mid D_1 I) = P(D_2 \mid I) = \int_{R_1} \int \ldots \int P(D_2 \mid H_k D_1 I) P(H_k \mid D_1 I) d\pi_{1k} d\pi_{2k} \ldots d\pi_{mk} =$$

$$= \ldots = \frac{n_2!}{f_{1,2}!\, f_{2,2}! \ldots f_{m,2}!} \cdot \frac{(n_1 + m - 1)!}{f_{1,1}!\, f_{2,1}! \ldots f_{m,1}!} \cdot \frac{g_{1,2}!\, g_{2,2}! \ldots g_{m,2}!}{(n_1 + n_2 + m - 1)!} \tag{20}$$

където

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

Така апостериорната вероятност е:

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

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

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

където

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

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

[44] Обикновено такъв е случаят в социалните изследвания.

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

Следователно:

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

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

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

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

Формули (30) и (33) представляват така нареченото *гаусово разпределение*[45]. 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	$P(0{,}186 < \pi_9 < 0{,}236 \mid DI) = 0{,}95$

[45] 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).

ЛИТЕРАТУРА

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, https://www.researchgate.net/publication/245581841_Bayes_in_five_days

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

ЧЕТВЪРТИ РАЗДЕЛ.
ВЗЕМАНЕ НА РЕШЕНИЯ В УСЛОВИЯ НА РИСК

Вземането на решения в условия на риск се свързва с изследването на извадка от лица и на тази основа с определянето на поведението на лице (лица), които не принадлежат към нея. Например, банка трябва да определи надеждността на нов кредитоискател на базата на информацията за всичките си предишни клиенти, а застрахователна компания трябва да определи риска от сключване на застраховка с потенциален клиент също на базата на информацията за всичките си предишни клиенти.

Методите за вземане на решения в условия на риск се основават на идеята, че за новата единица разполагаме с цялата налична информация освен значението на признака, отразяващ риска. Така например, от клиентите си банката (застрахователната компания) може да изиска всякаква информация – възраст, образование, семейно положение, доходи, имущество, професия, стаж и т.н., и т.н. Но банката (застрахователната компания) няма как да знае дали клиентът ще си върне заема (ще претърпи застрахователно събитие) или не. Това което банката (застрахователната компания) може да направи е да види каква част от миналите клиенти със същите характеристики (възраст, образование, семейно положение, доходи, имущество, професия, стаж и т.н., и т.н.) са си върнали заема (са претърпели застрахователно събитие) и каква част не са. Имайки тази информация може да се определи рискът новата единица да попадне в едната или другата група.

В този раздел първо ще бъде показано как се изчислява вероятността нова единица да попадне в конкретна група, формирана по значенията на признака, отразяващ риска. След това полето на приложение ще бъде разширено с построяването на доверителни интервали на екстраполационни прогнози.

<u>Девета тема:</u> Определяне на вероятността нова единица да попадне в конкретна група, формирана по значенията на един или няколко признака.

Всички честотни методи за вземане на решения в условия на риск се основават на определянето на вероятността новата единица да попадне в една или друга група, формирана по значенията на признака, отразяващ риска. Основната идея е да се моделира връзката между тази вероятност и всички останали признаци. Но тъй като всички останали признаци обикновено са много, това води до получаването на детайлно дефинирани подсъвкупности, които обаче са с малък обем. Това от своя страна налага признаците да се подреждат по важност и само най-важните да се включат в анализа.

Бейсовският анализ дава възможност във всяка подсъвкупност директно да се изчислява вероятността новата единица да попадне в една или друга група, формирана по значенията на признака, отразяващ риска. При това размерът на конкретната подсъвкупност няма значение.

От тази тема ще научите:

- По какво информацията при вземането на решения в условия на риск се различава от класическите извадки
- Как се изчислява вероятността нова единица да попадне в една или друга група, формирана по значенията на признака, отразяващ риска
- Какви са скритите предпоставки на лапласовото правило за приемственост
- Как се интерпретира числовата стойност, получена от лапласовото правило за приемственост

<u>Десета тема:</u> Приложение в областта на прогнозирането – доверителни интервали на екстраполационни прогнози.

Проблемът при прогнозирането е сходен с разгледания в предходната тема – действителната стойност на прогнозираната числова характеристика се различава от прогнозната стойност и би било добре предварително да се изчисли вероятността разликата между двете стойности да попадне в определен интервал. По този начин може предварително да се определи интервалът, в който е най-

вероятно да попадне действителната стойност. В същото време, интервалите се формират на базата на анализ на данни за вече минали периоди, т.е. отново се изчислява вероятността нови единици да попаднат в една или друга група.

От тази тема ще научите:

- Как се правят екстраполационни прогнози
- Как се изчислява вероятността разликата между действителната и прогнозната стойност да попадне в даден интервал
- Как се построяват доверителни интервали на екстраполационните прогнози

ТЕКСТ КЪМ ЧЕТВЪРТИ РАЗДЕЛ

Лапласово правило за приемственост – интерпретации и приложения*

Вземането на решения в условия на риск (decision making under risk) най-общо се свързва с избор между няколко различни алтернативи, когато предварително не е ясно коя от тях ще се осъществи на практика. Т.е. всичките са еднакво възможни, но не са еднакво вероятни. Освен вероятността (probability) за осъществяването на всяка алтернатива, при вземането на решение се отчита още и потенциалната полза (utility) от нея.

В настоящия доклад ще бъде разгледана само една част от задачата, а именно, определянето на вероятността за попадане в предварително дефинирана група. По-конкретно: изследване на извадка (sample) от лица и на тази основа определяне на принадлежността на лице (лица), което (които) не принадлежи (принадлежат) на извадката, към предварително дефинираните групи. Този анализ има съществени разлики спрямо класическите извадкови изследвания, които са следните:

Първо, извадката не се формира по правилата за формиране на представителни извадки, а представлява масив от данни, които се натрупват сякаш „от само себе си" за сравнително продължителен период от време;

Второ, изводите не се правят за цялата генерална съвкупност, а само за единиците от генералната съвкупност, които не са попаднали в извадката;

Трето, обикновено изводите не се правят за всички останали единици, а само за сравнително малка част от тях, най-често само за една единица.

Например, банка трябва да определи надеждността на нов кредитоискател на базата на информацията за всичките си предишни клиенти, а застрахователна компания трябва да определи риска от сключване на застраховка с потенциален клиент също на базата на информацията за всичките си предишни клиенти.

Всички методи за изчисляване на вероятностите[46] се основават на идеята, че за новата единица разполагаме с цялата налична информация, освен за признака, отразяващ риска. Така например, банката (застрахователната компания) може да изиска от клиентите си всякаква информация – възраст, образование, семейно положение, доходи, имущество, професия, стаж и т.н., и т.н. Но банката (застрахователната компания) няма как да знае дали клиентът ще си върне заема (ще претърпи застрахователно събитие) или не. Това което банката (застрахователната компания) може да направи е да види каква част от миналите клиенти със същите характеристики (възраст, образование, семейно положение, доходи, имущество, професия, стаж и т.н., и т.н.) са си върнали заема (са претърпели застрахователно събитие) и каква част не са. Имайки тази информация, може да се изчисли вероятността новата единица да попадне в едната или другата група.

Вероятността $n+1$-вата единица да попадне в *i*–тата група се получава по лапласовото правило за приемственост (Laplace rule of succession):

$$P(i \mid DI) = \frac{f_i + 1}{n + m}$$

където:

f_i е броят на единиците в извадката, попадащи в *i*–тата група,

n е обемът на извадката,

m е броят на групите (Jaynes 2003: 571; Харалампиев 2004: 82).

Лапласовото правило за приемственост всъщност е апостериорна вероятност (posterior probability), което означава, че тя е получена от някакви априорни вероятности (prior probability) и някакво

---

* Доклад, представен на научната конференция с международно участие „Авангардни научни инструменти в управлението", Равда, 2008

[46] Тук се визират както вече познатите и използвани в практиката методи логистична регресия (logistic regression) и класификационни дървета (classification trees), така и предлагания в настоящия доклад метод.

извадково разпределение (sampling distribution). От така записаната формулата обаче не става ясно какви са те и това трябва да бъде уточнено специално:

- априорните вероятности са равни (Харалампиев 2004: 20). Това означава, че липсва каквато и да била априорна информация;

- извадковото разпределение е хипергеометрично (Харалампиев 2004: 21). Това означава, първо, че подборът е безвъзвратен, и второ, че извадката е представителна.

Обаче извадката по никакъв начин не може да бъде приета за представителна. Следователно приложението на лапласовото правило за приемственост не е съвсем теоретически коректно. Неговото приложение се основава на един компромис – макар че извадката не е представителна, нейният дял в генералната съвкупност е достатъчно голям. Това е така, защото за генерална съвкупност се приема извадката плюс новата единица (Jaynes 2003: 570). В такъв случай генералната съвкупност ще има обем $n+1$, а делът на извадката ще е $\frac{n}{n+1}$. С нарастването на обема на извадката, нейният дял ще се доближава до 100%, следователно точността ще нараства, макар че извадката е непредставителна.

Има и още една причина да се пренебрегне фактът, че извадката е непредставителна. Основната причина за използването на представителни извадки е, че те възпроизвеждат сравнително точно съотношенията между подсъвкупностите в генералната съвкупност. Ако извадката не е представителна, някои подсъвкупности ще бъдат свръхпредставени, а други ще бъдат недопредставени. Като резултат от това числовите характеристики в цялата съвкупност ще бъдат или надценени, или подценени. Но, при вземането на решения в условия на риск, интерес представлява не цялата съвкупност, а конкретните подсъвкупности и то такива подсъвкупности, които са формирани по значенията на доста голям брой признаци. Това, дали те са свръхпредставени или недопредставени в извадката, няма значение, защото новата единица не е единица изобщо, а единица с определени характеристики и в този смисъл попада в конкретна подсъвкупност. Разбира се съществува възможността единиците в тази подсъвкупност да се диференцират по някакви други признаци, които не са били взети предвид. В този случай нарушаването на представителността би могло да доведе до изкривяване на крайните резултати, разбира се, в рамките на грешката на непредставителната извадка.

Интересен въпрос с важни практически последствия е каква е вероятността $n+k+1$-вата единица да попадне в *i*–тата група, ако е имало *k* предходни нови единици, които принадлежат на конкретната подсъвкупност, но за тях още не е налична необходимата информация за признака, отразяващ риска. Например, банката има *k* нови клиенти със съответните характеристики, но те са клиенти по-малко от месец и още не е ясно дали ще си връщат заема без проблеми или не. Но въпреки това трябва да се определи вероятността $n+k+1$-вият клиент със същите характеристики да попадне в рисковата група.

Може да се докаже, че вероятността $n+k+1$-вата единица да попадне в *i*–тата група, ако липсва информация за това в коя групи попадат *k*–те предходни нови единици, също се получава по лапласовото правило за приемственост. В този смисъл лапласовото правило за приемственост дава вероятността нова единица изобщо (не непременно първата) да попадне в *i*–тата група.

Този извод обаче не е безпроблемен. Вече беше казано, че лапласовото правило за приемственост предполага представителна извадка, но може да се прилага и при непредставителни извадки, поради високия дял на извадката. Но в случая делът на извадката е $\frac{n}{n+k+1}$ и намалява с увеличаването на *k*. Това означава, че когато се изчислява вероятността на коя да е от всичките останали единици от генералната съвкупност, които не са попаднали в извадката, да попадне в една или друга група, е най-добре да се използват данни от представителни извадки.

Представителна извадка задължително трябва да се използва и когато вместо да определяме вероятността нова единица да попадне в *i*–тата група, искаме да оценим (inference) относителния дял

(relative frequency) на единиците в генералната съвкупност, попадащи в *i*–тата група. Този относителен дял има бета разпределение (Haralampiev 2006: 4), а средната аритметична на бета разпределението се получава по същата формула като лапласовото правило за приемственост.

Трябва обаче да се прави разлика между двете. Макар числовият резултат да е един и същ, неговите интерпретации са различни. В случая на лапласовото правило за приемственост резултатът е вероятност, а в случая на бета разпределение резултатът е най-вероятната стойност на относителния дял в генералната съвкупност[47].

И така, една и съща формула може да се интерпретира по три различни начина:

Първо, като вероятност първата нова единица да попадне в *i*–тата група;

Второ, като вероятност нова единица изобщо (не непременно първата) да попадне в *i*–тата група;

Трето, като най-вероятна стойност на относителния дял на единиците в генералната съвкупност, попадащи в *i*–тата група.

Първите две интерпретации могат да се прилагат за вземане на решения в условия на риск и при това имат следните предимства:

Първо, тяхното приложение е изключително лесно;

Второ, чрез тях се дава директен отговор на въпроса каква е вероятността новата единица да попадне в една или друга група;

Трето, не може да се получи вероятност за попадане в група равна на нула, както и вероятност за попадане в група равна на единица, т.е. нито една алтернатива не се приема предварително нито за невъзможна, нито за сигурна;

Четвърто, лапласовото правило за приемственост може да се прилага, дори когато извадката е малка. Нещо повече, когато в масива липсва каквато и да е информация, лапласовото правило за приемственост дава равна вероятност за попадане във всяка група, т.е. липсата на каквато и да е информация (както на данни, така и на априорна информация) води до невъзможност за избор между различните алтернативи;

Пето, това правило е самонастройващо се, защото в момента, в който новата единица влезе в масива, тя увеличава извадката с единица и изводът за следващите нови единици се базира на по-голямата извадка. Нещо повече, с увеличаването на обема на извадката точността се увеличава.

От своя страна, третата интерпретация дава възможност полето на приложение на лапласовото правило за приемственост да се разшири с използването му в социологическите, политическите и маркетинговите изследвания.

Литература:

1. Харалампиев, К. 2004. Нетрадиционен поглед върху традиционни статистически проблеми. Балкани, С.

2. Haralampiev, K. 2008. Bayesian Inference of Relative Frequency (In the Case of Electoral Surveys). Annuaire de l'universite de Sofia "St. Kliment Ohridski", Fakulte de Philosophie, Livre – Sociologie, Tome 99. Статията е достъпна в Интернет на адрес: http://kaloyan-haralampiev.info/wp-content/uploads/2010/03/article1.pdf (При цитирането номерата на страниците са давани по варианта в Интернет.)

3. Jaynes, E. 2003. Probability Theory: the Logic of Science. Cambridge Univercity Press, Cambridge. Непълно електронно издание на тази книга е достъпно в Интернет на адрес: http://omega.albany.edu:8008/JaynesBook.html (При цитирането номерата на страниците са давани по печатното издание.)

[47] Всъщност средната аритметична би била най-вероятната стойност, само ако при бета разпределението тя съвпадаше с модата. Но, поради асиметрията на бета разпределението, това не е така. Все пак, при големи стойности на *n* разликата между модата и средната аритметична е доста по-малка от желаната точност, т.е. те на практика съвпадат. Риск от неправилно отъждествяване на двете съществува единствено при малки стойности на *n*.