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”)

Калоян Харалампиев
СУ „Св. Климент Охридски“
Имейл: k_haralampiev@phls.uni-sofia.bg

Абстракт: Много често при работата на терен се срещат трудности, които водят до това, че една извадка, която е планирана като представителна, всъщност се оказва непредставителна. Най-често срещаният проблем е високият дял на неоткритите и/или неотговорили лица. В този случай не е коректно да се използват класическите статистически методи за построяване на доверителни интервали и/или за проверка на статистически хипотези. Това налага да се използва бейсовска статистика, която позволява да се построяват доверителни интервали и да се проверяват статически хипотези на базата на данни от непредставителни извадки. Точно такъв беше случаят при изследването на отпадащите студенти в бакалавърска степен на обучение във Философски факултет на Софийски университет „Св. Климент Охридски“. Изследването на отпадналите студенти беше планирано като изчерпателно, но поради ниския процент на попълнени анкетни карти то беше реализирано като непредставителна извадка. Изследването на контролната група от студенти, продължаващи своето образование, беше планирано като представителна извадка, но поради високия процент на неоткритите и/или неотговорили студенти, то също беше реализирано като непредставителна извадка. Това наложи сравнението между двете групи да бъде направено с помощта на бейсовската статистика.

Ключови думи: представителни извадки, непредставителни извадки, статистически изводи и заключения, доверителни интервали, проверка на статистически хипотези, бейсовска статистика.

Using of Bayesian statistics for statistical inferences with non-representative samples (based on an example of the research of the dropout bachelor students in the Faculty of Philosophy of the Sofia University “St. Kliment Ohridski”)

Kaloyan Haralampiev
Sofia University “St. Kliment Ohrdski”
E-mail: k_haralampiev@phls.uni-sofia.bg

Abstract: Very often in the course of fieldwork, there are difficulties which lead to a situation in which a sample that has been planned to be representative is actually non-representative. The most common problem is the high proportion of undiscovered respondents and/or non-respondents. In this case, it is not correct to use the classic statistical methods for confidence intervals and/or hypotheses testing. There is a need to use Bayesian statistics, which allows confidence intervals to be constructed and statistical hypotheses to be tested based on non-representative sampling data. This is exactly the case in the study of the dropout Bachelor students in the Faculty of Philosophy of Sofia University "St. Kliment Ohridski". The dropout study had been planned to be exhaustive, but due to the low percentage of filled-in questionnaires, it was implemented as a non-representative sample. The study of the control group of students continuing their education had been planned as a representative sample, but due to a high

percentage of undiscovered and/or non-responded students, it was again implemented as a non-representative sample. This required the comparison between the two groups to be made by use of Bayesian statistics.

Keywords: representative samples, non-representative samples, statistical inferences, confidence intervals, hypotheses testing, Bayesian statistics.

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

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.

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:

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

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.

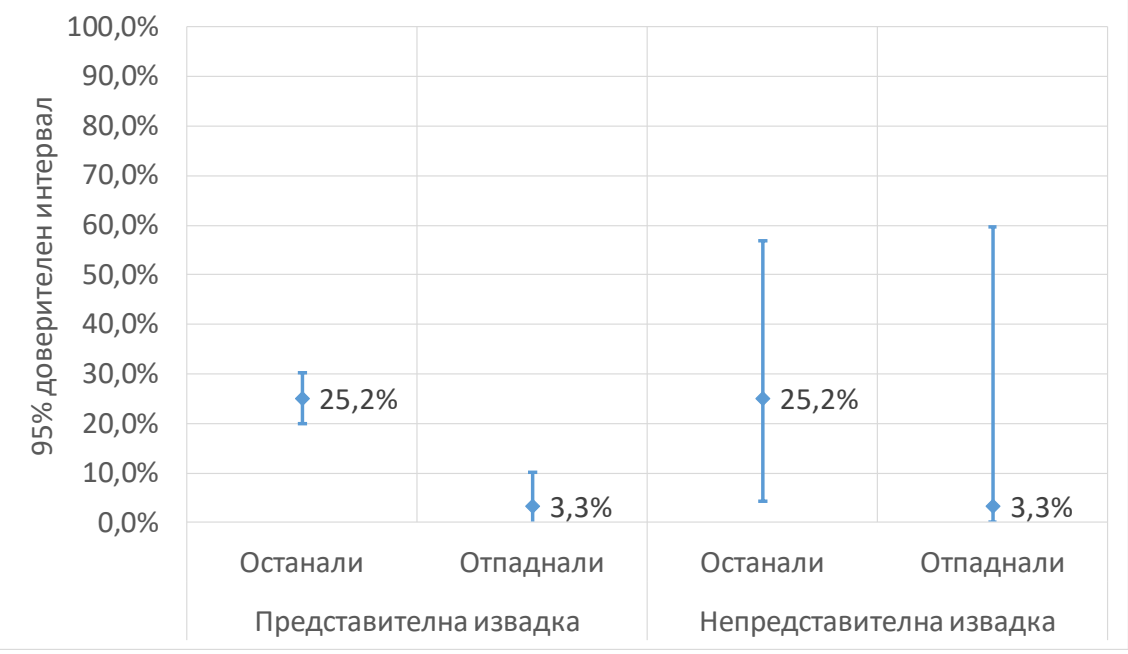
Ето един характерен пример за такова сравнение:

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

		Група		Общо
		Останали	Отпаднали	
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%

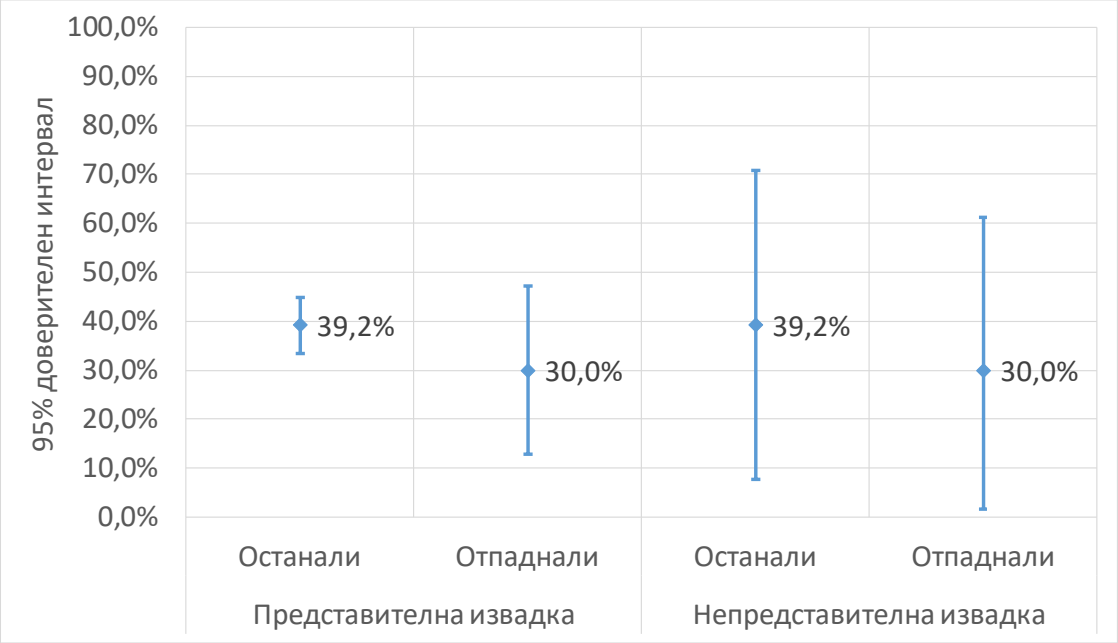
За да се провери дали има статистически значимо различие между двете групи, са построени доверителните интервали за относителните дялове на всеки отделен отговор във всяка от двете групи. Тези доверителни интервали са представени на следващите графики, като всяка графика се състои от две части. Лявата 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“

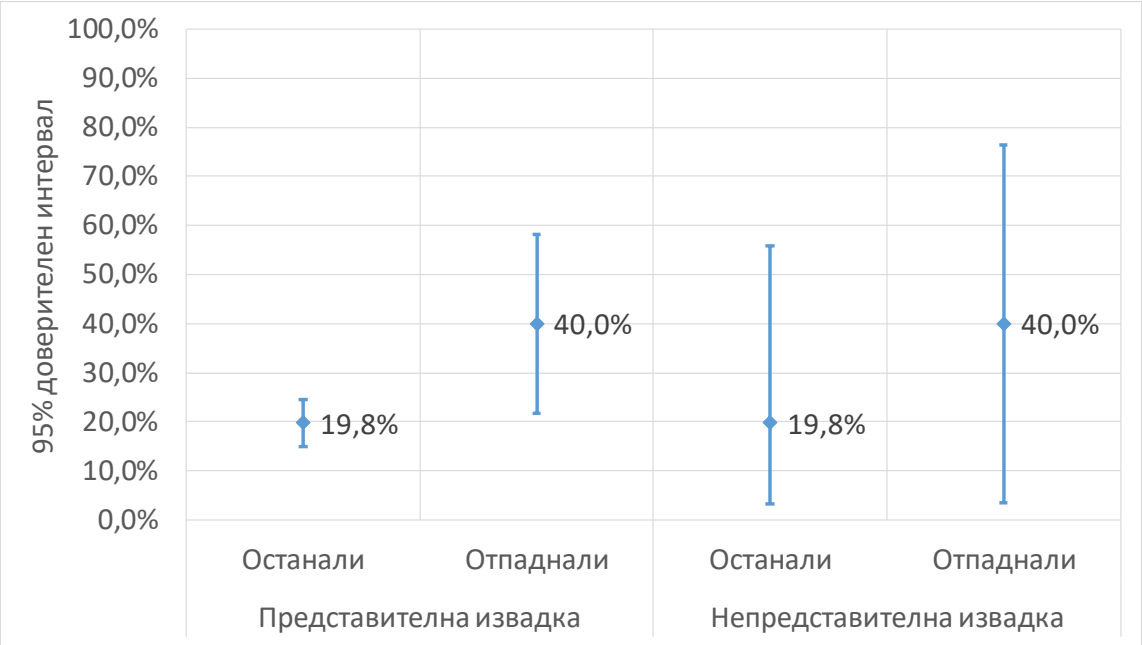
Фиг. 1 показва, че ако двете извадки биха били представителни, би имало статистически значимо различие между двете групи по отношение на относителния дял на отговора „Много важно“. Но тъй като двете извадки не са представителни, този извод може да бъде потвърден само частично, тъй като оценката на относителния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



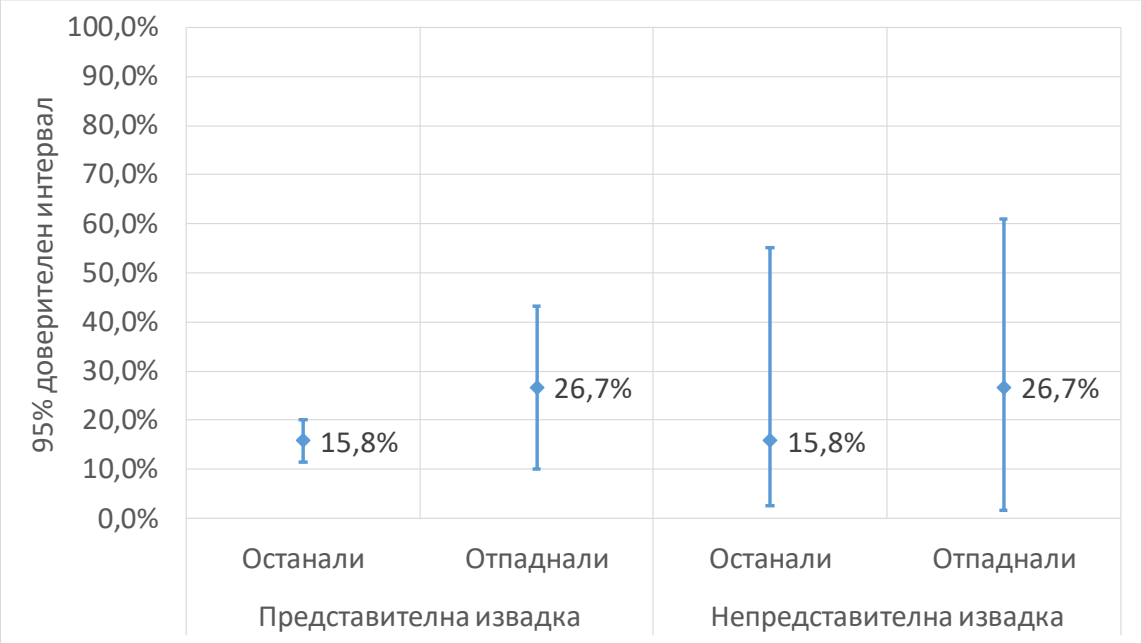
Фиг. 2. Доверителни интервали на относителните дялове на отговора „Важно“

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) $\Delta_{p,max} = \min[(p - \pi_{min}); (\pi_{max} - p)]$,

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) $P(p - \Delta_{p,max} \leq \pi \leq p + \Delta_{p,max}) = F(p + \Delta_{p,max}) - F(p - \Delta_{p,max})$,

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

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.

Сп. „Реторика и комуникации“, брой 36, септември 2018 г.

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”).