Калоян Валентинов Харалампиев

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[1], 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[2]. В този случай отново имаме непредставителна извадка по метода на отзовалите се, но вече излъчена от някаква неопределена хипотетична съвкупност, която е достатъчно голяма[3].

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?"[4] 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 (Хараламписв 2004). Това става по различен начин в зависимост от вида на признака – количествен или качествен.

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.

---

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

[2] 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.

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

[4] 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.

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?”[5]

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.

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

И така, можем ли да кажем “Народът писа “Задоволителен”[6] (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:

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

и разсейване:

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

(Харалампиев 2004).

Използвайки таблица за нормално разпределение може да се построи доверителеният интервал на средния бал[7]:

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

където $\Delta_{\overline{x}}$ 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:

$$\Delta_{\overline{x},\max} = \min[(\overline{x} - x_{\min});(x_{\max} - \overline{x})] = \min[(2{,}96 - 2{,}00);(6{,}00 - 2{,}96)] =$$
$$= \min(0{,}96;3{,}04) = 0{,}96$$

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

---

[6] Запазена е оригиналната терминология от интернет страницата.

[7] Чете се “Вероятността неизвестната средна аритметична да се намира в границите от … до … е …”.

желаната гаранционна вероятност. Това разширяване може да стане само надясно[8]. 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[9].

---

[8] 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.

[9] 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 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

Това, което се прави е изчисляването на относителните дялове на всяко конкретно значение на признака.

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[10]:

(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).

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":

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

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.

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

$$P(0{,}7400-0{,}2600<\pi_1<0{,}7400+0{,}2600) = P(0{,}4800<\pi_1<1{,}0000) =$$
$$= P(0{,}4800<\pi_1) = (1-0{,}4800)^{3-1} = 0{,}5200^2 = 0{,}2704$$

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”:

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

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

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

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”:

---

[10] Чете се “Вероятността неизвестният относителен дял на *i*–тото значение на признака да е по-

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

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

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

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

малък от … е …”.

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[11] 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 –

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

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.