Статистически методи в социологията – втора част
(Извадкови изследвания)

Уважаеми колеги,

В този текст са представени анотациите на темите за лекции по "Статистически методи в социологията – втора част (Извадкови изследвания)". Няма отделни теми за упражненията, тъй като по време на упражненията Вие ще използвате емпирични данни, за да прилагате наученото на лекции.

Всяка тема е представена заедно със списък на основната и допълнителната литература. Посочените литературни източници в графата "Основна литература" са алтернативни и можете да ползвате който и да е от тях. Литературните източници, посочени в графата "Допълнителна литература" хвърлят допълнителна светлина върху някои детайли на разглеждания проблем и можете да ги ползвате при проявен по-голям интерес към съответната тема. За някои теми няма достатъчно информация в посочената основна и допълнителна литература, затова в края на текста са дадени допълнителни текстове по тях.

1. Статистическо извадково изучаване

В практиката на емпиричните изследвания не винаги, когато трябва да се изучи една съвкупност, има възможност да се наблюдават всички единици. Това много често е неизгодно или невъзможно. Изходът е да се наблюдават част от единиците на съвкупността и въз основа на тях да се правят изводи за цялата съвкупност.

От тази тема ще научите:

- Какво е генерална съвкупност и какво е извадка.
- Какво е параметър и какво е оценка.
- Какви грешки се допускат при всяко извадково изследване.
- Какви са условията, за да бъде една извадка представителна (репрезентативна).

Основна литература: Венедиков 1992: 7-10 и 41-43; Гатев, Гатева 2008: 141-143; Съйкова и колектив 2002: 42-49

Допълнителна литература: Брогли, Петкова 1988: 13-15; Венедиков 1993: 38-40; Енциклопедичен речник по социология 1996: 77, 86-87 и 343-344;

Калинов 2001: 8-9; Петров, Велева-Стефанова 2009: 163-164; Трифонов, Цонкова 2007: 114-116; Харалампиев 2003: 35 и 37

2. Модели на представителни извадки

За да може извадката да отговаря на условията за представителност, разгледани в първа тема, тя трябва да бъде формирана по определена технология.

От тази тема ще научите как се формират и какви особености имат:

- Простата случайна извадка.
- Систематичната извадка.
- Районираната (стратифицирана) извадка.
- Гнездовата извадка (едностепенна, двустепенна и многостепенна).

Допълнителна литература: Брогли, Петкова 1988: 15; Венедиков 1992: 54-56; Венедиков 1993: 200-202; Гатев, Гатева 2008: 144-145; Енциклопедичен речник по социология 1996: 154-156 и 256-257; Калинов 2001: 113-116; Парчев 1998: 111-112 и 117-120; Петров, Велева-Стефанова 2009: 165-167, Трифонов, Цонкова 2007: 116-118

3. Модели на непредставителни извадки

В практическата работа не винаги може да се формира представителна извадка. Понякога е по-лесно да се формира непредставителна извадка, а друг път просто са налице данни, които са получени по начин, който не може да гарантира представителност.

От тази тема ще научите как се формират и какви особености имат:

- Анкетата по пощата.
- Телефонното интервю.
- Квотната извадка.
- Типологичната извадка.
- Извадката по метода на отзовалите се.

Допълнителна литература: Енциклопедичен речник по социология 1996: 153-154, 157, 268-269 и 502; Парчев 1998: 113-115

4. Статистическо оценяване

При извадково изследване параметрите на генералната съвкупност са неизвестни, а могат да се получат само техните оценки. Задачата е на базата на получените оценки да се направят изводи за интересуващите ни параметри.

От тази тема ще научите:

- Какви видове оценки има.
- Какво е емпирично разпределение и какво е теоретично разпределение.
- Кои са най-често използваните теоретични разпределения.
- Какво е стохастично разпределение.

Основна литература: Съйкова и колектив 2002: 49-59; Харалампиев 2003: 35-37

Допълнителна литература: Венедиков 1992: 11-15; Енциклопедичен речник по социология 1996: 478; Калинов 2001: 57-65, 112-113, 121-124 и 133-134; Манов 2001: 123-138

5. Доверителен интервал на средна аритметична и на относителен дял

В емпиричните социологически изследвания най-често се работи с качествени признаци, затова е важно да можем да построяваме доверителни интервали на относителни дялове. От друга страна относителният дял може да се разглежда като частен случай на средната аритметична, затова е важно да можем да построяваме доверителни интервали и на средни аритметични.

От тази тема ще научите:

- Какви важни свойства има стохастичното разпределение.
- Какво е стандартна грешка и какво е максимална грешка.
- Какво е гаранционен множител и какво е гаранционна вероятност.
- Как се избира гаранционната вероятност и как се получават гаранционният множител, стандартната и максималната грешка.
- Как се построява доверителен интервал на средна аритметична и на относителен дял.
- Как се интерпретира доверителният интервал на средната аритметична и на относителния дял.

Основна литература: Брогли, Петкова 1988: 76-81; Гатев, Гатева 2008: 152-158; Манов 2001: 144-151; Петров, Велева-Стефанова 2009: 167-173; Харалампиев 2003: 37-44

Допълнителна литература: Венедиков 1992: 22-23; Калинов 2001: 134-137; Парчев 1998: 115-117; Трифонов, Цонкова 2007: 120-123

6. Статистическа проверка на хипотези (тестове за значимост)

Има случаи, когато точните стойности на параметрите в генералната съвкупност не са от съществено значение, а е по-важно сравняването им с определени числа или помежду им. Това налага да се формулират хипотези.

От тази тема ще научите:

- Какво е статистическа хипотезата.
- Какви статистически хипотези се дефинират и проверяват.
- Какви грешки се допускат при статистическата проверка на хипотези.
- Как се избира рискът за грешка.
- Как се прави окончателният извод при статистическата проверка на хипотези.
- Какви методи за проверка на хипотези има.

Основна литература: Брогли, Петкова 1988: 6 и 83-87; Гатев, Гатева 2008: 167-173; Калинов 2001: 124, 129 и 151-167; Манов 2001: 154-155 и 166-168; Съйкова и колектив 2002: 79-95; Трифонов, Цонкова 2007: 129-133; Харалампиев 2003: 45-48

Допълнителна литература: Венедиков 1992: 16; Енциклопедичен речник по социология 1996: 478-479 и 538; Петров, Велева-Стефанова 2009: 181-186

7. Проверка за равенство на средна аритметична и на относителен дял с дадено число

Има случаи, в които ние знаем предварително или колко е била в миналото, или колко трябва да бъде стойността на даден параметър. Тъй като не сме в състояние да проведем изчерпателно наблюдение, задачата е, по информация от представителна извадка, да проверим дали стойността на

параметъра в момента на изследването съвпада с тази, която е била в миналото или с тази, която трябва да бъде.

От тази тема ще научите:

- Как се избира „даденото число”.
- Как се проверява хипотезата за равенство на средната аритметична или на относителният дял с дадено число.

Основна литература: Калинов 2001: 124-128, 171-173 и 179-182; Манов 2001: 155-166; Харалампиев 2003: 48-50 и 55-57

Допълнителна литература: Венедиков 1992: 24-25; Гатев, Гатева 2008: 175-176; Трифонов, Цонкова 2007: 133-138

8. Проверка за равенство на две средни аритметични и на два относителни дяла при независими извадки

Когато разполагаме с две извадки те могат да са независими или да са зависими. Независими извадки означава, че всяка извадка е формирана отделно, без връзка с другата.

От тази тема ще научите как се прави проверка за равенство на две средни аритметични и на два относителни дяла при независими извадки.

Основна литература: Брогли, Петкова 1988: 90-92 и 103-104; Гатев, Гатева 2008: 176-180 и 182-183; Калинов 2001: 197-204 и 208-210; Манов 2001: 178-186 и 195-197; Петров, Велева-Стефанова 2009: 191-198; Харалампиев 2003: 50-55 и 57-60

Допълнителна литература: Венедиков 1992: 25-27; Парчев 1998: 55-56

9. Проверка за равенство на две средни аритметични и на два относителни дяла при зависими извадки

Зависими извадки означава, че е формирана само едната извадка, а това кои единици попадат в другата се определя от техните връзки с единиците от първата извадка.

От тази тема ще научите как се прави проверка за равенство на две средни аритметични и на два относителни дяла при зависими извадки.

Основна литература: Калинов 2001: 204-208 и 327-328

Допълнителна литература: Брогли, Петкова 1988: 92-94; Манов 2001: 186-189

10. Проверка за съгласуваност на емпирично и теоретично разпределение и на две емпирични разпределения

Освен проверка на хипотези относно параметрите на генералната съвкупност, можем да проверяваме хипотези и относно разпределенията на единиците. При това можем да сравняваме емпиричното разпределение с някакво познато теоретично разпределение, използвано за еталон, или да сравняваме две емпирични разпределения едно с друго.

От тази тема ще научите:

- Как се прави проверката за съгласуваност на емпирично и теоретично разпределение и на две емпирични разпределения помежду им.
- Как се измерва големината на различието между две разпределения.

Основна литература: Манов 2001: 212-218, 280-283 и 436-437

Допълнителна литература: Венедиков 1992: 31-34; Гатев, Гатева 2008: 183-188; Енциклопедичен речник по социология 1996: 535-537; Калинов 2001: 313-327

11. Нарушаване на условията за приложение на параметричните методи – решение първо: непараметрични методи за проверка на хипотези

За да можем да прилагаме методите за проверка на хипотези, разгледани в предходните теми, трябва разпределението на единиците в генералната съвкупност да отговаря на определени условия. Ако условията са нарушени, тогава не бихме могли да твърдим, че емпиричната характеристика на критерия има съответното теоретично разпределение. В такъв случай можем да използваме непараметрични методи за проверка на хипотези, които не зависят от свойствата на разпределението на единиците в генералната съвкупност.

От тази тема ще научите:

- Кои непараметрични тестове се използват за сравняване на две медиани въз основа на информация от две независими извадки.
- Кои непараметрични тестове се използват за сравняване на две медиани въз основа на информация от две зависими извадки.

Основна литература: Манов 2001: 432-435, 436 и 437-440

Допълнителна литература: Брогли, Петкова 1988: 97-99; Калинов 2001: 330-332 и 334-335

Уважаеми колеги,

След като преминете тази тема и на упражнения Вие ще направите първия тест. Той е с 8 въпроса. Всеки правилен отговор носи една точка. Междинната оценка от първия тест се определя по следната скала:

От 0 до 4 точки – слаб (2)

5 точки – среден (3)

6 точки – добър (4)

7 точки – много добър (5)

8 точки – отличен (6)

12. Нарушаване на условията за приложение на параметричните методи – решение второ: бейсовски подход

Друго възможно решение на проблема, поставен в единадесета тема, е т.нар. бейсовски подход.

От тази тема ще научите:

- Каква е разликата между честотния и бейсовския подход.
- Как се проверяват на хипотези и как се построяват доверителни интервали при бейсовския подход.

Основна литература: Харалампиев 2007

13. Анализ на връзки между качествен признак фактор и качествен признак резултат

Тъй като при провеждане на емпирични социологически изследвания се получава информация най-вече за качествени признаци, то най-често срещаният в социологическата практика случай за анализ на връзки е именно изследване на връзки между качествен признак фактор и качествен признак резултат.

От тази тема ще научите как се установява наличието на връзка между два качествени признака, когато информацията е получена от извадково изследване.

Основна литература: Венедиков 1993: 147-150; Енциклопедичен речник по социология 1996: 536; Манов 2001: 298-304 и 311-312; Съйкова и колектив 2002: 112-126; Харалампиев 2003: 71-76

Допълнителна литература: Брогли, Петкова 1988: 63-65 и 102-103; Венедиков 1992: 83-84; Парчев 1998: 37-39 и 51-53

14. Анализ на връзки между качествен признак фактор и количествен признак резултат

Макар в социологическите изследвания сравнително рядко да се срещат количествени признаци, доста често има признаци, чиито значения са разположени на бална скала. На практика балната скала е своеобразен заместител на интервалната скала и с нея се работи по същия начин, както с интервална. Това налага да можем да изследваме връзката между качествен признак фактор и количествен признак резултат.

От тази тема ще научите как се установява наличието на връзка между качествен признак фактор и количествен признак резултат, когато информацията е получена от извадково изследване.

Основна литература: Брогли, Петкова 1988: 99-102; Гатев, Гатева 2008: 191-198; Калинов 2001: 103-105 и 232-250; Манов 2001: 224-229 и 333-334; Петров, Велева-Стефанова 2009: 198-203; Съйкова и колектив 2002: 362-365 и 369-370; Трифонов, Цонкова 2007: 144-149; Харалампиев 2003: 82-88

15. Анализ на връзки между количествен признак фактор и количествен признак резултат

Освен случаите на използване на бална скала, за които стана дума в предходната тема, много често в социологическите изследвания ролята на фактори се изпълнява от някакви обективни характеристики, които са количествени признаци. Това налага да можем да изследваме връзката между два количествени признака.

От тази тема ще научите:

- Как се установява наличието на връзка между два количествени признака, когато информацията е получена от извадково изследване.

- Как се построяват доверителни интервали и как се проверяват статистически хипотези относно регресионните и корелационните коефициенти.

Основна литература: Брогли, Петкова 1988: 40-57 и 60-62; Калинов 2001: 68-81 и 342-357; Манов 2001: 237-248, 250-254 и 268-276; Петров, Велева-Стефанова 2009: 228-229 и 235-239; Съйкова и колектив 2002: 165-176, 264-268 и 289-291; Харалампиев 2003: 94-98

Допълнителна литература: Гатев, Гатева 2008: 229-234

Уважаеми колеги,

След като преминете тази тема и на упражнения Вие ще направите втория тест. Той е с 16 въпроса. Всеки правилен отговор носи една точка. Крайната оценка от двата теста се определя по следната скала:

От 0 до 12 точки – слаб (2)

От 13 до 15 точки – среден (3)

От 16 до 18 точки – добър (4)

От 19 до 21 точки – много добър (5)

От 22 до 24 точки – отличен (6)

Тези от вас, които не са се явили на някой от тестовете през годината, задължително се явяват на краен тест, който съдържа 24 въпроса. Всеки верен отговор носи една точка. Крайната оценка се получава по горната скала:

Тези от вас, които са се явили и на двата теста, но са на мнение, че получената оценката не отразява знания им, имат право да поискат анулиране на оценката. В такъв случай се явяват на тест, който съдържа 24 въпроса. Оценката се получава по горната скала. Получената оценка е окончателна.

Литература, налична в библиотека „Социални науки”:

Брогли, Я., Л. Петкова. 1988. *Статистически методи в спорта*. София: „Медицина и физкултура”

Венедиков, Й. 1992. *Статистика, социология и още нещо...* София: Информационно обслужване

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

Гатев, К., Н. Гатева. 2008. *Статистика. Статистически методи в емпиричните изследвания и бизнеса*. София: „Парадигма”

Енциклопедичен речник по социология. 1996. София: „М-8-М”

Калинов, К. 2001. *Статистически методи в поведенческите и социалните науки*. София: Нов български университет

Манов, А. 2001. *Статистика със SPSS*. София: „Тракия-М”

Парчев, И. 1988. *Избор на партия, избор на президент*. София: Статистическо издателство и печатница при НСИ

Петров, С., С. Велева-Стефанова. 2009. *Обща теория на статистиката*. София: „Парадигма”

Съйкова, И., А. Стойкова-Къналиева, С. Съйкова. 2002. *Статистическо изследване на зависимости*. София: Университетско издателство „Стопанство”

Трифонов, Т., В. Цонкова. 2007. *Статистика в икономиката и управлението*. Велико Търново: „Астарта”

Харалампиев, К. 2003. *Въведение в основните статистически методи за анализ*. София: „Балкани”

Харалампиев, К. 2007. „За парадигмите в статистиката – бейсовска статистика”. В: *Актуални проблеми на статистическата теория и практика*. Сборник с доклади от научна конференция. София: университетско издателство „Стопанство”
http://kaloyan-haralampiev.info/Publications%20by%20Type/Papers%20on%20Conferences/(2007)%20About%20the%20Paradigms%20of%20Statistics%20-%20Bayesian%20statistics/

ДОПЪЛНИТЕЛНИ ТЕКСТОВЕ

Към темите, за които няма достатъчно информация в посочената към тях основна и допълнителна литература

Допълнителен текст към теми №2 и №3

„Модели на представителни извадки" и „Модели на непредставителни извадки"

Съдържание

Въведение

Ако винаги, когато трябва да се изучи една статистическа съвкупност, има възможност да се наблюдават всички статистически единици, няма проблем да се реализира изчерпателно наблюдение. Много често обаче наблюдението на всички единици в една съвкупност е неизгодно или невъзможно.

Първо, ако съвкупността е прекалено голяма, това води до огромен разход на *ресурси – хора* (*екип*), *време* (*срокове*) и *пари*, за да бъде наблюдавана изчерпателно.

Второ, ако съвкупността е динамично променяща се, това води до необходимостта от много кратки срокове за провеждане на емпиричното изследване, което от своя страна води до необходимостта от голям екип и много пари.

Изходът е да се наблюдава само част от единиците на съвкупността и въз основа на тях да се правят изводи за цялата съвкупност.

Основни понятия

Изучаваната съвкупност се нарича *генерална съвкупност* (*population*), а наблюдаваната част от нея – *извадка* (*sample*). Описателните числови характеристики на отделните признаци в генералната съвкупност се наричат *параметри* (*parameters*), а изчислените от извадката – *оценки* на параметрите (*estimations*).

При извадково изследване параметрите на генералната съвкупност са неизвестни, а могат да се получат само техните оценки.

Ако получените оценки са конкретни числа, те се наричат точкови оценки. Ясно е, че точковите оценки могат да съвпадат с неизвестния параметър в генералната съвкупност, но могат и да не съвпадат с него, като е много по-вероятно да не съвпадат, отколкото да съвпадат. За да се избегне този недостатък на точковите оценки, се изчислява един интервал около точковата оценка, в който с определена вероятност се намира неизвестният параметър на генералната съвкупност. Този интервал се нарича *интервална* оценка (*доверителен интервал*) (*confidence interval*).

Условия за представителност на извадката

Изчисляването на доверителните интервали изисква извадката да бъде *представителна* (*representative*), а за да бъде една извадка представителна, е необходимо да отговаря на три условия:

Първо, генералната съвкупност трябва да бъде точно и ясно дефинирана. Голяма заблуда е, че една извадка може да бъде представителна „по принцип“. Всяка извадка е представителна само спрямо генералната съвкупност, от която е излъчена. Много често в специализираната литература това условие се формулира като необходимост от *изчерпателен* (*адресен*) *списък* на статистическите единици в генералната съвкупност.

Второ, единиците, които попадат в извадката, трябва да бъдат определени чрез *случаен подбор* (*random choice*). Под случаен подбор се разбира такъв подбор, който осигурява *равен шанс* на всяка статистическа единица да попадне в извадката. Изискването за случаен подбор е изключително важно, тъй като само при случаен подбор може да се използва теорията на вероятностите, с чиято

помощ се доказва, че структурата на извадката е приблизително еднаква на структурата на генералната съвкупност. А това от своя страна означава, че количествените съотношения, пропорции и коефициенти, получени от извадката, са приблизително равни на съотношенията, пропорциите и коефициентите в генералната съвкупност.

Трето, забранява се заместването на статистическите единици, попаднали в извадката, с други, в случаи на проблеми или трудности при откриването им. Тази забрана се налага, тъй като всяко заместване води до нарушаване на пропорциите в извадката, отдалечаването им от пропорциите в генералната съвкупност и по този начин получаване на съотношения, пропорции и коефициенти, които надценяват или подценяват съответните съотношения, пропорции и коефициенти в генералната съвкупност.

Забраната за заместване обаче може да доведе до ниска *възвращаемост* (*response rate*), а това от своя страна също може да наруши пропорциите в реализираната извадка и да доведе до надценяване или подценяване на съотношенията, пропорциите и коефициентите на генералната съвкупност. Затова анкетьорите имат две много важни задачи – да открият лицата, попаднали в извадката, и да ги убедят да се включат в проучването.

За изпълнението на първата задача анкетьорите получават специални инструкции. Например, първото посещение да бъде в работен ден в работно време. В случай на неуспех второто посещение да бъде в работен ден след работно време. В случай на повторен неуспех третото посещение да бъде в събота или в неделя. И ако и третото посещение е неуспешно, тогава изследването лице се отписва от извадката и не се търси повече.

За изпълнението на втората задача анкетьорите минават специален инструктаж, на който се обучават, първо, да различават категоричните откази от меките откази, и второ, да превръщат меките откази в съгласие за участие. С категоричните откази не се работи, защото дори и да се получи съгласие за участие в анкетното проучване, остават съмнения доколко изследваното лице ще отговаря искрено и точно на въпросите.

Само ако и трите условия за представителност са изпълнени, само тогава извадката е представителна. Нарушаването дори само на едно от тези условия прави извадката непредставителна.

За да може извадката да отговаря на условията за представителност, тя трябва да бъде формирана по определена *технология*. Различните начини на формиране на извадка се наричат *модели* на извадката.

Определяне на обема на извадката

При всеки модел, най-напред е необходимо да се определи *обемът* (*size*) на извадката, който би осигурил желана точност.

Обемът на извадката няма връзка с нейната представителност. Възможно е една извадка да е малка, но представителна, както и да е голяма и непредставителна. Разбира се, възможни са и другите две комбинации – малка и непредставителна, както и голяма и представителна.

Обемът на извадката обаче е важен за нейната *точност*. По-голяма извадка означава по-голяма точност, и обратно. Така че, най-добрата комбинация е една извадка да е голяма и представителна. Но голямата извадка изисква повече ресурси – хора, време и пари. Затова при планиране на обема на извадката са вземат предвид и двете страни – от едната страна е точността, а от другата страна са необходимите ресурси. При това са възможни два подхода:

Първо, определя се желаната точност. След това се изчислява обемът на извадката, който ще осигури тази точност. След това се правят разчети дали този обем може да бъде реализиран с наличните ресурси. Ако не може, тогава се търсят допълнителни ресурси.

Второ, определя се обемът на извадката, който би могъл да бъде реализиран с наличните ресурси. След това се правят разчети каква точност ще осигури този обем на извадката. Ако точността е по-ниска от желаната, тогава също се търсят допълнителни ресурси.

Модели на извадката

След като е определен обемът, се пристъпва към формирането на самата извадка.

Има различни модели, които осигуряват представителни извадки. Основните са три – *прост случаен подбор*, *райониран* (*стратифициран*) *подбор* и *гнездови подбор*.

Прост случаен подбор

При простия случаен подбор се прилага някаква форма на лотария с номерата на единиците от техния изчерпателния списък. При това са възможни две различни ситуации:

- след изтеглянето на номера на единицата, тя се връща в генералната съвкупност и има шанс отново да попадне в извадката при следващо изтегляне. Този начин на подбор се нарича „*схема с връщане*" или *възвратен* подбор;
- след изтеглянето на номера на единицата, тя не се връща в генералната съвкупност. Този начин на подбор се нарича „*схема без връщане*" или *безвъзвратен подбор*.

Възвратният подбор е основният теоретичен модел на формиране на представителна извадка. На практика, обаче, подборът почти винаги е безвъзвратен.

Простият случаен подбор е скъп, защото може да се окаже, че трябва да се изпрати анкетьор и да му са платят пътни, дневни и квартирни пари, за да отиде в населено място, където да анкетира само едно лице.

На простия случаен подбор съответстват два аналогични модела, които обаче не са представителни, тъй като при тях е нарушено някое (но не всички) от условията за представителност. Това са *анкетата по пощата* и *телефонното интервю*.

Анкета по пощата

При анкетата по пощата подготовката е както при простия случаен подбор: от изчерпателния списък на статистическите единици чрез някаква форма на лотария се избират тези единици, които попадат в извадката, и им се изпращат анкетните карти. Когато се използва класическата поща и се изпращат анкетни карти на хартия, те се изпращат заедно със самоадресиран плик за връщане на попълнената карта. Но може да се изработи онлайн въпросник и да се използва електронна поща за изпращане на линка към анкетата. Тогава не се очаква респондентите да връщат попълнените анкетни карти, а платформите за онлайн анкетиране автоматично събират отговорите в електронни бази данни.

Анкетата по пощата има два проблема, които се отразяват на нейната

представителност. Първо, по никакъв начин не може да се контролира дали лицето, попаднало в извадката, е попълнило анкетната карта само или някой друг е направил това вместо него. Второ, на практика се връщат около една четвърт от изпратените анкетни карти. Това е сериозен проблем, тъй като мненията на наблюдаваната една четвърт може да съвпадат, но може и да не съвпадат с мненията на останалите три четвърти. Ако съвпадат, това е добре, но ако не съвпадат, тогава информацията, с която разполагаме е силно изкривена. Най-лошото е, че няма как знаем дали има съвпадение в мненията или не. Тези обстоятелства не позволяват да се оценява точността на информацията и следователно не може да се правят генерализации за цялата изследвана съвкупност. Могат да се правят изводи само за съвкупността на отзовалите се.

Телефонно интервю

При телефонното интервю подготовката е сходна с тази при анкетата по пощата: от изчерпателния списък на статистическите единици чрез някаква форма на лотария се избират тези единици, които попадат в извадката и след това те се интервюират по телефона. Проблемите, обаче, са малко по-различни. Първо, може не всички лица в генералната съвкупност да имат телефон, което прави част от статистическите единици недостъпни за изследване. Както и по-горе, мненията на достъпните единици може да съвпадат, но може и да не съвпадат с мненията на недостъпните, което евентуално може да доведе до изкривяване на информацията. Затова този метод е използван обикновено в общества с почти сто процентова телефонизация на лицата и домакинствата или когато може да се допусне, че лицата без телефон не представляват интерес от гледна точка на целта на изследването. Второ, много по-трудно се убеждават анкетираните лица, че анонимността им ще бъде гарантирана. Това по същество не е проблем на представителността, но също може да доведе до големи нестохастични грешки, ако интервюираните лица премълчават или преиначават информацията.

Систематичен подбор

Разновидност на простия случаен подбор е т.нар. *систематичен подбор*. При него броят на статистическите единици в генералната съвкупност се разделя на броя на единиците в извадката. По този начин се получава т.нар. *стъпка на*

подбора. След това чрез някаква форма на лотария се избира *стартов номер*. Номерата на следващите единици се получават като към стартовия номер се прибавя стъпката на подбора. Очевидно систематичният подбор винаги е безвъзвратен и осигурява представителна извадка.

Недостатък на систематичния подбор е, че при някои изследвания стъпката на подбора може да съвпадне с цикъл във вътрешната структура на изучаваната съвкупност. Това води до попадане в извадката на статистически единици, които имат систематично по-високи или систематично по-ниски значения на изучавания признак, а това от своя страна означава, че стойността на интересуващия ни параметър ще бъде надценена или подценена.

Райониран (стратифициран) подбор

Простият случаен подбор е теоретичната основа на извадковите изучавания, но изследователската практика е наложила да се търсят такива модели, които да водят до подобряване на точността и до намаляване на ресурсите (хора, време, пари), необходими за изследването (в сравнение с простата случайна извадка). За съжаление не е създаден такъв модел, който да решава и двете задачи.

Моделът, създаден за подобряване на точността на оценките, се нарича райониран (стратифициран) подбор. При него първо се избират няколко признака, които се намират във връзка с изучаваните. След това се осигурява информация за разпределенията на статистическите единици в генералната съвкупност по тези признаци. Обикновено се избират такива признаци, за които има общодостъпна публична статистическа информация. На базата на разпределенията се прави многомерна групировка по значенията на избраните признаци. По този начин се получават т.нар. *райони* (*страти*). Единиците, попадащи в извадката, се избират чрез някаква форма на лотария във всеки район (страта).

Тъй като единиците във всеки район (страта) са много по-хомогенни, отколкото в цялата съвкупност, това води до намаляване на разсейването, а оттам и до подобряването на точността. От друга страна предварителната работа по осигуряването на информация за районiращите признаци и по формирането на районите (стратите) води до увеличаване на ресурсите на изследването (хора,

време, пари).

Квотна извадка

Непредставителният аналог на районираната (стратифицирана) извадка се нарича *квотна извадка*. При нея:

- избират се няколко признака, които се намират във връзка с изучаваните;
- осигурява се предварителна информация за разпределенията на статистическите единици в генералната съвкупност по тези признаци;
- прави се многомерна групировка по значенията на избраните признаци;
- след това обаче, във всяка група не се прилага някаква форма на лотария, а се определя *квота*. Това е броят на единиците, които трябва да бъдат наблюдавани. Обикновено квотата се определя пропорционално на дела на съответната група в генералната съвкупност;
- анкетьорите търсят и анкетират лица, отговарящи на условията за попадане в съответната група, докато запълнят квотата.

Очевидно е нарушено условието за равния шанс на единиците да попаднат в извадката. Също така, при този модел заместването на една единица с друга, не само, че се допуска, ами без него моделът не би могъл да се реализира. На практика обаче, квотната извадка е много по-широко използвана в сравнение с представителния си аналог – районираната (стратифицирана) извадка. Това се дължи основно на две причини: първо, нейното осъществяване е по-лесно, и второ, квотната извадка е единствената извадка със стопроцентово изпълнение на предварително планирания обем. От друга страна е ясно, че колкото връзката между квотообразуващите и изучаваните признаци е по-силна, толкова точността на изследването е по-голяма. Проблемът е в това, че теоретично е невъзможно да бъде измерена силата на посочената връзка, докато при районираната (стратифицираната) извадка това е възможно.

Гнездови подбор

Моделът, създаден за намаляване на ресурсите, се нарича гнездови подбор. При него най-напред се формират т.нар. *гнезда*. Това са подсъвкупности от статистически единици, които са разположени териториално близо помежду си и могат да бъдат наблюдавани едновременно. След това се прави извадка от гнезда, като изборът на гнездата, попаднали в извадката, отново става чрез някаква форма на лотария. В зависимост от следващата стъпка извадката бива *едностепенна* гнездова извадка, *двустепенна* гнездова извадка или *многостепенна* гнездова извадка.

Ако в избраните гнезда се наблюдават всички статистически единици, то извадката е едностепенна гнездова извадка.

Ако в избраните гнезда се прилага някаква форма на лотария за избиране на наблюдаваните единици, то извадката е двустепенна гнездова извадка. При това са възможни два варианта – извадката във всяко гнездо е фиксиран процент от размера на гнездото или извадката във всяко гнездо е с равен обем. От практически съображения по-често се избира вторият вариант, защото по-лесно се планира, управлява и контролира работата на терен.

Ако в рамките на гнездата се формират под-гнезда, в рамките на под-гнездата – под-под-гнезда и т.н. (възможни са няколко йерархични нива на под-гнезда), то извадката е многостепенна гнездова извадка. Многостепенни гнездови извадки обикновено се използват в държави с голяма територия и/или голямо население или в институции със сложна йерархична структура.

Тъй като статистическите единици могат да бъдат наблюдавани едновременно в рамките на много кратък времеви интервал, това означава, че с едни и същи разходи за пътни, дневни и квартирни пари могат да се наблюдават няколко единици, а това води до намаляване на ресурсите на изследването (хора, време, пари). От друга страна, близкото териториално разположение на статистическите единици води до взаимно влияние между тях, което води до т.нар. *вътрешногнездова корелация* и влошава точността на изследването.

Типологична извадка

Накрая ще разгледаме два модела на типични непредставителни извадки (при тях са нарушени и трите условия за представителност). Това са

типологичната извадка и *методът на отзовалите си*.

При типологичната извадка се наблюдават само няколко единици, за които се счита, че са „*типични*" за изследваната съвкупност. Очевидно тук не е необходимо да се прави изчерпателен списък на единиците в генералната съвкупност. Също така, подборът не е случаен, а е субективно решение на изследователя. Често, за да се намали субективността, се прибягва до използването на метода на *експертните оценки* за определяне на „типичните" единици. Намаляването на субективността обаче, не премахва неслучайността на подбора, т.е. не се осигурява равен шанс на единиците да попаднат в извадката, и следователно не осигурява представителност на извадката.

Метод на отзовалите се

При метода на отзовалите се се изготвят голям брой анкетни карти, които се поставят на място, до което имат достъп голям брой хора. Лицата сами решават дали да попълнят анкетните карти или не. Този модел е най-проблематичен по отношение на дефиниране на генералната съвкупност. На практика изобщо не е ясно кои лица биха попълнили анкетната карта и дали те формират подходяща съвкупност от гледна точка на целите и задачите на изследването. Също така, тук изобщо липсва подбор от страна на изследователя (дори и субективен като в предходния модел). Подборът е дело на самите изследвани лица. След като няма подбор, е невъзможно да се говори за случайност и за равен шанс на единиците да попаднат в извадката. Също като при анкетата по пощата, тези обстоятелства не позволяват да оценяваме точността на информацията и следователно не можем да правим генерализации за цялата изследвана съвкупност. Могат да се правят изводи само за съвкупността на отзовалите се.

* * *

На практика най-често в реалните емпирични изследвания моделите на представителни извадки се комбинират. На първата стъпка обикновено се прави стратифициран подбор по статистически райони или по области. Това гарантира, че всеки статистически район (респ. всяка област) ще бъде представен/а в извадката и извадката ще има добро териториално покритие. На втората стъпка

се прави гнездови подбор, обикновено двустепенен. Това гарантира поевтиняване на изследването. На третата стъпка се прилага прост случаен или систематичен подбор за избирането на конкретните изследвани лица.

Препоръчителна литература

На български език

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

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

На английски език

Alvi, M. 2016. „A Manual for Selecting Sampling Techniques in Research“. Munich Personal RePEc Archive, Paper No. 70218, https://mpra.ub.uni-muenchen.de/70218/

Perumal, T. 2014. „Research Methodology. Topic 10: Sampling“. Open University Malaysia, https://www.tankonyvtar.hu/en/tartalom/tamop412A/2011-0021_22_research_methodology/adatok.html

Допълнителен текст към тема №9

„Проверка за равенство на две средни аритметични и на два относителни дяла при зависими извадки”

При зависими извадки се формират двойки от единици – едната единица е от първата извадка, а другата – нейната съответна от втората извадка.

При сравняването на две средни аритметични, за всяка двойка единици се изчислява разликата между стойностите на интересуващия ни признак. По този начин се формира извадка от разлики. Въз основа на нея може да се направи проверка дали средната разлика в генералната съвкупност е равна на нула. Тази проверка е аналогична на проверката за равенство на средна аритметична с дадено число.

При сравняване на два относителни дяла се разглежда дихотомен признак (или признак, изкуствено приведен към дихотомен чрез противопоставяне на интересуващото ни значение на признака на всички останали значения) и се приписва стойност 1 на интересуващото ни значение и 0 на другото. След това двойките единици се разпределят по начина, показан в следната таблица:

Разпределение на двойките единици

Значения на признака в първата извадка	Значения на признака във втората извадка		Общо
	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
Общо	$a+c$	$b+d$	n

В таблицата a е броят на двойките, за които и двете единици притежават интересуващото ни значение на признака, а d е броят на двойките, за които и двете единици не притежават интересуващото ни значение на признака. От друга страна, b и c е броят на двойките, за които едната единица притежава интересуващото ни значение на признака, а другата не го притежава.

Относителният дял на интересуващото ни значение в първата извадка е:

$$p_1 = \frac{a+b}{n} = \frac{a}{n} + \frac{b}{n}$$

Относителният дял на интересуващото ни значение във втората извадка е:

$$p_2 = \frac{a+c}{n} = \frac{a}{n} + \frac{c}{n}$$

Очевидно тези относителни дялове имат общ компонент $\frac{a}{n}$ и различни компоненти – съответно $\frac{b}{n}$ и $\frac{c}{n}$. Следователно разликата между двата относителни дяла се определя от разликата между *b* и *c*. Критерият на проверката има χ^2-разпределение, чиято емпирична стойност се получава по формулата:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Допълнителен текст към тема №10

„Проверка за съгласуваност на емпирично и теоретично разпределение и на две емпирични разпределения”

Освен проверка на хипотези относно параметрите на генералната съвкупност, можем да проверяваме хипотези и относно разпределенията на единиците. При това можем да сравняваме емпиричното разпределение с някакво познато теоретично разпределение, използвано за еталон или да сравняваме две емпирични разпределения едно с друго. Проверката в двата случая се извършва с едни и същи критерий – χ^2 и критерий на Колмогоров-Смирнов.

Емпиричната стойност на χ^2 при проверката за съгласуваност на емпирично и теоретично разпределение се изчислява по формулата:

$$\chi^2 = \sum \frac{(f - \hat{f})^2}{\hat{f}},$$

където f са честотите на емпиричното разпределение, а $\hat{f}$ – на теоретичното.

При сравняването на две разпределения първо се построява следната таблица:

Емпирични разпределения на единиците в двете извадки

Извадка	Значения на признака				Общо
	x_1	x_2	…	x_c	
Първа	f_{11}	f_{12}	…	f_{1c}	$f_{1\bullet}$
Втора	f_{21}	f_{22}	…	f_{2c}	$f_{2\bullet}$
Общо	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet c}$	n

Емпиричната стойност на χ^2 се изчислява по формулата:

$$\chi^2 = n\left(\sum_i \sum_j \frac{f_{ij}^2}{f_{i\bullet} \cdot f_{\bullet j}} - 1\right)$$

При приложението на критерия на Колмогоров-Смирнов, първо, значенията на признака се подреждат във възходящ ред и се изчисляват

прогресивно-кумулативните относителни дялове за всяко от двете разпределения. След това се изчисляват разликите (d) между прогресивно-кумулативните относителни дялове за всяко значение на признака и се намира най-голямата разлика (по абсолютна стойност). Емпиричната характеристика има разпределение на Колмогоров и при сравняване на емпирично и теоретично разпределение се изчислява по формулата:

$$K = d_{\max}.\sqrt{n},$$

а при сравняването на две емпирични разпределения – по формулата:

$$K = d_{\max}.\sqrt{\frac{n_1.n_2}{n_1 + n_2}}$$

Вместо разпределението на Колмогоров може да се използва отново χ^2–разпределение, тъй като величината $(2.K)^2$ има χ^2–разпределение.

Ако се установи, че между две разпределения има различие, може да се измери големината на различието. За целта най-напред се изчисляват относителните дялове за всяко от двете разпределения. След това се намират разликите между относителните дялове за всяко значение на признака. Ако се изчисли средната аритметична от разликите, резултатът винаги ще е нула, независимо от големината на различието между двете разпределения. По тази причина се изчислява средната квадратична разлика по формулата:

$$\bar{d} = \sqrt{\frac{\sum (p' - p'')^2}{m}},$$

където p' са относителните дялове за едното разпределение, p'' – за другото, а m е броят на значенията на признака.

Средната квадратична разлика измерва средното различие между относителните дялове на двете разпределения. Тя обаче зависи от броя на значенията на признака и при различни стойности на m има различна горна граница, която е $\sqrt{\frac{2}{m}}$. Ако разделим средната квадратична разлика на горната й граница, ще получим измерител, който е нормиран в границите от 0 до 1. Този измерител се нарича нормирано евклидово разстояние и се изчислява по формулата:

$$d_n = \sqrt{\frac{\sum (p' - p'')^2}{2}}$$

Колкото стойността на d_n е по близка до 0, толкова различието между сравняваните разпределения е по-малко, а колкото е по-близка до 1, толкова различието е по-голямо.

Допълнителен текст към тема №11

„Нарушаване на условията за приложение на параметричните методи – решение първо: непараметрични методи за проверка на хипотези”

При проверките на хипотези, разгледани до сега, изчислявахме някакви емпирични характеристики, за които твърдяхме, че имат някакво познато теоретично разпределение. За да имат емпиричните характеристики съответното теоретично разпределение, трябва разпределението на единиците в генералната съвкупност да отговаря на определени условия. Най-често се поставя изискването разпределението на единиците в генералната съвкупност да е нормално или близко до нормалното. Ако условията са нарушени, не бихме могли да твърдим, че емпиричната характеристика има съответното теоретично разпределение. Едно възможно решение в тази ситуация е да използваме непараметрични критерии за проверка на хипотези, които не зависят от разпределението на единиците в генералната съвкупност.

Ще разгледаме следните непараметрични критерий – сериен тест, серийно-рангов тест (критерий на Ман-Уитни), знаков тест и знаково-рангов тест (критерий на Уилкоксън). Първите два се използват за сравняване на центровете на две разпределения въз основа на информация от две независими извадки, а последните два – за сравняване на центровете на две разпределения въз основа на информация от две зависими извадки.

При серийния тест двете извадки се обединяват и единиците се подреждат във възходящ ред. След това за всяка единица се отбелязва принадлежността ѝ към едната или към другата извадка. По този начин се образуват т.нар. серии. Серията е последователност от съседни единици, които принадлежат на една и съща извадка. За критерий на проверката служи броят на сериите. За него са възможни всички стойности, намиращи се между две крайности.

Първата крайна ситуация е налице, когато всички единици от едната извадка имат значения на признака по-малки от единиците в другата извадка. Тогава сериите са само две и очевидно центърът на едното разпределение е по-малък от центъра на другото разпределение.

Втората крайна ситуация е налице, когато единиците от двете извадки са разположени последователно по следния начин: единица от първата извадка,

единица от втората извадка, единица от първата извадка, единица от втората извадка и т.н. Тогава сериите са $2.\min(n_1;n_2)+1$ и очевидно центровете на двете разпределения са приблизително еднакви.

На практика броят на сериите е между двете крайни стойности като, когато е по-близо до 2, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности се различават, а когато е по-близо до $2.\min(n_1;n_2)+1$, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности са еднакви.

При серийно-ранговия тест (критерия на Ман-Уитни) отново двете извадки се обединяват и единиците се подреждат във възходящ ред. След това на всяка единица се приписва ранг, отговарящ на мястото й в подредбата. Събират се ранговете на единиците от едната извадка и на единиците от другата извадка. За критерий на проверката служи по-голямата от двете суми. Отново крайните ситуации са същите:

Когато всички единици от едната извадка имат значения на признака по-малки от единиците в другата извадка, по-голяма е сумата на ранговете във втората извадка, която е $n_1.n_2+\frac{n_2.(n_2+1)}{2}$.

Когато единиците от двете извадки са разположени последователно, двете суми на ранговете ще са приблизително еднакви. Тъй като общата сума на ранговете е $\frac{(n_1+n_2).(n_1+n_2+1)}{2}$, то двете суми ще са приблизително по $\frac{(n_1+n_2).(n_1+n_2+1)}{4}$.

На практика по-голямата сума се намира между двете крайни стойности като, когато е по-близо до $n_1.n_2+\frac{n_2.(n_2+1)}{2}$, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности се различават, а когато е по-близо до $\frac{(n_1+n_2).(n_1+n_2+1)}{4}$, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности са еднакви.

При знаковия тест се формират двойки от единици по начина, показан в девета тема и се изчисляват разликите между стойностите на интересуващия ни

признак за всяка двойка. Тук обаче се използва само знакът на разликата. За критерий на проверката служи броят на този знак („плюс” или „минус”), който се среща по-често. За него са възможни всички стойности, намиращи се между две крайности.

Първата крайна ситуация е налице, когато всички единици от едната извадка имат значения на признака по-малки от единиците в другата извадка. Тогава всички знаци ще бъдат еднакви, техният брой ще бъде n и очевидно центърът на едното разпределение ще бъде по-малък от центъра на другото разпределение.

Втората крайна ситуация е налице, когато при половината единици от едната извадка, значенията на признака са по-малки, а при другата половина са по-големи в сравнение с другата извадка. Тогава броят на знаците „плюс” и „минус” ще бъде приблизително равен на $\frac{n}{2}$ и очевидно центровете на двете разпределения ще бъдат приблизително еднакви.

На практика по-големият брой на знаците е между двете крайни стойности, като когато е по-близо до n, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности се различават, а когато е по-близо до $\frac{n}{2}$, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности са еднакви.

При знаково-ранговия тест (критерия на Уилкоксън) отново се изчисляват разликите между стойностите на интересуващия ни признак за всяка двойка. След това разликите се подреждат във възходящ ред по абсолютна стойност и им се приписва ранг, отговарящ на мястото им в подредбата. Събират се ранговете на положителните и на отрицателните разлики. За критерий на проверката служи по-голямата от двете суми. Отново крайните ситуации са същите:

Когато всички единици от едната извадка имат значения на признака по-малки от единиците в другата извадка, всички знаци са еднакви и сумата от техните рангове е $\frac{n.(n+1)}{2}$.

Когато при половината единици от едната извадка, значенията на признака са по-малки, а при другата половина са по-големи в сравнение с другата

извадка, двете суми на ранговете ще са приблизително равни на $\frac{n.(n+1)}{4}$.

На практика по-голямата сума се намира между двете крайни стойности като, когато е по-близо до $\frac{n.(n+1)}{2}$, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности се различават, а когато е по-близо до $\frac{n.(n+1)}{4}$, изводът ще бъде, че центровете на разпределенията в двете генерални съвкупности са еднакви.