## §8. Анализ на категорийни величини

**1. Увод.** Анализът на категорийни величини заема важно място в психологията, понеже много от величините имащи съществен значение за изследванията се отразяват в номинални скали. Такива величини са например полът, религиозната идентификация, етническата идентификация, принадлежност към определен регион, резултати от анкетни проучвания и др. Тези величини не са метрични и изискват специални техники на статистически анализ, различни от изложените дотук.

Примерните данни в тази лекция са получени в резултат от вторична преработка на съпровождащите данни ***GSS93*** на пакета SPSS. Тези данни имат на практика само илюстративен характер, понеже извадките не са подложени на систематичен контрол за представителност. По тази причина направените изводи не могат да бъдат отнесени строго към популацията, която в дадения случай представлява населението на САЩ.

**2. Анализ на пропорцията в една извадка.** Да разгледаме една извадка с обем $n$, в която $f$ на брой индивида (статистически единици) притежават даден признак. В този случай се наблюдава една дихотомна величина – признак, имаща две категории, които могат да бъдат означени условно чрез "ДА(YES)" и "НЕ(NO)" в зависимост от това дали наблюдаваният индивид притежава или не въпросния признак. В числов вид тези категории се отбелязват обикновено с 1 и 0 (понятието дихотомна означава наличие само на две нива на измерване).

Отношението $p = \frac{f}{n}$ представлява ***относителният дял*** – пропорцията, в която признакът е представен в извадката. Когато извадката е представителна, този емпиричен относителен дял носи информация за ***теоретичния дял*** $\pi$ на признака в цялата популация. Тук въз основа на данните от опита може да се проверява нулева хипотеза $H_0 : \pi = \pi_0$, която гласи, че относителният дял в популацията има някаква фиксирана стойност $\pi_0$ срещу двустранната алтернатива $H_{alt} : \pi \neq \pi_0$. Проверяващата статистика на $H_0$ има вида

$$z_{emp} = \sqrt{n}\,\frac{p - \pi_0}{\sqrt{p(1-p)}},$$

която при валидна нулева хипотеза $H_0$ има приблизително нормално стандартно разпределение. При зададено ниво на значимост $\alpha$, нулевата хипотеза $H_0$ се отхвърля когато $\left|z_{emp}\right| \geq z_{1-\frac{\alpha}{2}}$ и се приема в противен случай, като на практика обикновено се работи с оцененото ниво на значимост.

За популационния дял $\pi$ могат да се пресмятат доверителни интервали. При зададено ниво на значимост $\alpha$, $(1-\alpha)100\%$ доверителен интервал за $\pi$ има вида

$$\left( p + z_{\frac{\alpha}{2}}\sqrt{\frac{p(p-1)}{n}},\, p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(p-1)}{n}} \right).$$

Технически отхвърлянето на нулевата хипотеза $H_0$ при ниво на значимост $\alpha$ се извършва, когато фиксираната стойност $\pi_0$ лежи извън този доверителен интервал и съответно $H_0$ се приема, когато $\pi_0$ лежи вътре в интервала.

Да разгледаме пример за проучване мнението в населението на САЩ относно разрешението за носене и притежаване на оръжие от частни лица в извадка от $n = 1480$ индивида. Тук $f = 811$ души са посочили твърдо съгласие, а останалите $n - f = 669$ са

изразили несъгласие или липса на мнение. Пресмятаме $p = \frac{f}{n} = \frac{811}{1480} = 0.548$. При ниво на значимост $\alpha = 0.05$, 95% доверителен интервал за популационния относителен дял има вида

$$\left(0.548 - 0.96\sqrt{\frac{0.548(1-0.548)}{1480}}, 0.548 + 0.96\sqrt{\frac{0.548(1-0.548)}{1480}}\right) = (0.523, 0.573).$$

Този резултат дава основание да предположим, че повече от половината от населението има положително отношение към притежаването и носенето на оръжие. В частност намереният 95% доверителен интервал дава основание за отхвърляне на нулевата хипотеза $H_0 : \pi = 0.5$ при ниво на значимост $\alpha = 0.05$, понеже стойността $\pi_0 = 0.5$ не се съдържа в интервала.

Да проверим сега самостоятелно тази нулева хипотеза $H_0 : \pi = 0.5$, която се интерпретира в предположението, че населението се разделя в две равни групи "ЗА" и "ПРОТИВ", срещу двустранната алтернатива $H_{alt} : \pi \neq 0.5$. За проверяващата статистика намираме

$$z_{emp} = \sqrt{1480}\,\frac{0.548 - 0.5}{\sqrt{0.548(1-0.548)}} = 3.708$$

с оценено ниво на значимост $p_{level} = 0.0002$, което указва отново за отхвърляне на нулевата хипотеза, при което вероятността за грешка (от първи род) е от порядъка на $0.0002$. В този пример добре се вижда приликата между техниката на доверителни интервали и техниката на проверка на статистически хипотези, както и предимството което ни дава последната въз основа на пресмятане оценените ниво на значимост.

**3. Сравняване на пропорции в две извадки.** Да разгледаме сега две ***независими*** извадки с обеми $n_1$ и $n_2$, в които съответно $f_1$ и $f_2$ на брой индивида притежават наблюдавания признак. Тук имаме два относителни дяла $p_1 = \frac{f_1}{n_1}$ и $p_1 = \frac{f_2}{n_2}$. Задачата е да направим статистическо сравнение между $p_1$ и $p_2$. За тази цел проверяваме нулевата хипотеза $H_0 : \pi_1 = \pi_2$ срещу двустранната алтернатива $H_{alt} : \pi_1 \neq \pi_2$. Проверяващата статистика за $H_0$ има вида

$$z_{emp} = \frac{p_1 - p_2}{\sqrt{\left(\frac{f_1 + f_2}{n_1 + n_2}\right)\left(1 - \frac{f_1 + f_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

която при валидна нулева хипотеза $H_0$ има приблизително нормално стандартно разпределение. При зададено ниво на значимост $\alpha$, нулевата хипотеза $H_0$ се отхвърля когато $\left|z_{emp}\right| \geq z_{1-\frac{\alpha}{2}}$ и се приема в противен случай, като на практика се работи с оценените ниво на значимост.

Да разгледаме отново примера с анкетата за одобрение на притежаване и носене на оръжие, като този път обособим две групи за мъжете и за жените. Данните показват $n_{male} = 637$, $f_{male} = 314$ и $n_{female} = 843$, $f_{female} = 497$, откъдето пресмятаме

$$p_{male} = \frac{314}{637} = 0.493 \text{ и } p_{female} = \frac{497}{843} = 0.589.$$

За проверяващата статистика намираме

$$z_{emp} = \frac{0.493 - 0.589}{\sqrt{\left(\frac{314+497}{637+843}\right)\left(1-\frac{314+497}{637+843}\right)\left(\frac{1}{637}+\frac{1}{843}\right)}} = -3.698,$$

с оценено ниво на значимост $p_{level} = 0.0002$, което води до сигурно отхвърляне на нулевата хипотеза. В този случай жените проявяват съществено по-високо одобрение към притежаването и носенето на оръжие.

"Жените показват съществено по-високо одобрение към притежаването и носенето на оръжие, [ $z = -3.698$ ; $p = 0.000$ ]."

Съществува и тест за сравняване на два относителни дяла от зависими извадки. Той може да бъде прилаган например ако желаем да проследим промяната на относителния дял в резултат от някакво въздействие.

**4. Корелация на дихотомни величини.** Ако наблюдаваме едновременно два признака $X$ и $Y$, то резултатите могат да бъдат подредени е една таблица

Таблица 8.1.

| | | $Y$ | |
|---|---|---|---|
| | | ДА | НЕ |
| $X$ | ДА | $a$ | $b$ |
| | НЕ | $c$ | $d$ |

която се нарича ***таблица на спрегнатост***. Тук $a$, $b$, $c$ и $d$ са цели числа, които показват бройките индивиди, характеризирани по съответния начин спрямо двата признака. Например $a$ означава броят на индивидите, притежаващи едновременно и двата признака и т.н. Очевидно за обема на извадката имаме $n = a+b+c+d$. Величината

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

се интерпретира като коефициент на корелация (***коефициент на Пирсън-Браве***) и дава посоката и силата на връзката между признаците.

Може да се проверява нулева хипотеза $H_0$ : *величините са независми* срещу алтернативата $H_{alt}$ : *величините са зависми*. Проверяващата статистика на $H_0$ има вида

$$\chi^2_{emp}(1) = n\frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

която при валидна нулева хипотеза $H_0$ има $\chi^2(1)$ разпределение. При зададено ниво на значимост $\alpha$, нулевата хипотеза $H_0$ се отхвърля, когато $\chi^2_{emp}(1) \geq \chi^2_{1-\alpha;1}$ и се приема в противен случай, като на практика се работи с оцененото ниво на значимост.

В качеството на пример ще разгледаме два признака, първият от които представлява участие в президентските избори 1992г. в САЩ, а вторият представлява одобрение за смъртното наказание. Данните показват следната таблица на спрегнатост, получена върху извадка от $n = 1348$ индивида.

Таблица 8.2.

| | | одобрение на смъртното наказание | |
|---|---|---|---|
| | | ДА | НЕ |
| участие в президентските избори | ДА | 738 | 224 |
| | НЕ | 303 | 83 |

Пресмятанията показват $r=-0.019$ и $\chi^2_{emp}(1)=0.497$ с оценено ниво на значимост $p=0.481$, което не дава основания за отхвърляне на нулевата хипотеза за независимост на признаците.

"Одобрението на смъртното наказание е независимо от електоралната активност на населението, [$\chi^2(1)=0.497$ ; $p=0.481$]."

**5.** $\chi^2$**-тест за независимост.** Да разгледаме две номинални величини $X$ и $Y$, като за $X$ се предполага, че притежава $J_x$ на брой категории (нива), а $Y$ притежава $J_y$ на брой категории и нека разполагаме с данни от $n$ на брой наблюдения. Всеки индивид се отнася точно към една от категориите на всяка от величините $X$ и $Y$, следователно резултатите след изброяване формират следната таблица на спрегнатост.

Таблица 8.3.

| | | $Y$ | | | |
|---|---|---|---|---|---|
| | | категория 1 | категория 2 | ... | категория $J_y$ |
| $X$ | категория 1 | $f_{11}$ | $f_{12}$ | ... | $f_{1J_y}$ |
| | категория 2 | $f_{21}$ | $f_{22}$ | ... | $f_{2J_y}$ |
| | ... | ... | ... | ... | ... |
| | категория $J_x$ | $f_{J_x1}$ | $f_{J_x2}$ | ... | $f_{J_xJ_y}$ |

Тук $f_{ij}$ представлява ***наблюдаваната честота*** на случаите, за които индивидът се характеризира от категория $i$ на величината $X$ и от категория $j$ на величината $Y$. В този случай съществува сравнително лесен тест за проверка на нулева хипотеза $H_0$ : *величините са независими* срещу алтернативата $H_{alt}$ : *величините са зависими*. За тази цел отначало за всяка клетка трябва да се пресметне ***очакваната честота*** $E_{ij}$ въз основа предположението за независимост. Тази очаквана честота се дават по формулата

$$E_{ij}=\frac{(\textit{сума на наблюденията от ред } i)(\textit{сума на наблюденията от ред } j)}{\textit{обем на извадката}}.$$

Проверяващата статистика за нулевата хипотеза $H_0$ има вида

$$\chi^2_{emp}(k)=\sum_{ij}\frac{(f_{ij}-E_{ij})^2}{E_{ij}},\ k=(J_x-1)(J_y-1),$$

която при валидна нулева хипотеза $H_0$ има $\chi^2(k)$ разпределение. При зададено ниво на значимост $\alpha$, нулевата хипотеза $H_0$ се отхвърля, когато $\chi^2_{emp}(k)\geq\chi^2_{1-\alpha;k}$ и се приема в противен случай.

Независимостта означава, че категориите на едната величина са разпределени по един и същ начин във всяка от категориите на другата величина, което фактически показва, че от разпределението на едната величина не можем да правим изводи за категориите на другата величина. При чисто номинални величини статистическата зависимост, ако е налице такава, има сила но няма посока, понеже посоката е свързана с някаква наредба, а за категориите на номиналните величини не се предполага наличие на наредба. Ако се стигне до приемане на нулевата хипотеза $H_0$, то резултатът се оказва лесен за тълкуване в рамките на току що даденото определение за независимост. Ако обаче се стигне до отхвърляне на нулевата хипотеза, то обикновено тълкуването на резултата в типичния случай е свързано със сериозни затруднения, които няма да разискваме подробно.

Да разгледаме например извадка състояща се от $n=1425$ лица граждани на САЩ, за които се отчитат две величини. Първата величина има две категории и показва дали са семейни или не, а втората показва религиозна идентификация в три категории – протестанти, католици и атеисти. Таблицата на спрегнатост има следния вид.

Таблица 8.4.

| | | религиозна идентификация | | |
|---|---|---|---|---|
| | | протестанти | католици | атеисти |
| семейно положение | семейни | $f_{11}=498$<br>$E_{11}=499.716$ | $f_{12}=193$<br>$E_{12}=174.796$ | $f_{13}=57$<br>$E_{13}=74.488$ |
| | несемейни | $f_{21}=454$<br>$E_{21}=452.284$ | $f_{22}=140$<br>$E_{22}=158.204$ | $f_{23}=83$<br>$E_{23}=66.512$ |

Тук заедно с наблюдаваните честоти са приведени и стойностите на очакваните честоти. За степените на свобода имаме $k=(2-1)(3-1)=2$. Пресмятанията показват $\chi^2_{emp}(2)=11.789$ с оценено ниво на значимост $p=0.003$, което води до отхвърляне на хипотезата за независимост $H_0$. Сравнявайки наблюдаваните и очакваните честоти можем да направим извода, че при католиците се наблюдава по-висок дял на семейните отколкото при атеистите.

"Между семейния статус и религиозната идентификация съществува значима зависимост, $[\chi^2(2)=11.789\,;\,p=0.003]$. При католиците преобладават семейните индивиди, докато при атеистите се наблюдава обратната тенденция."

Тълкуването на $\chi^2$ зависимостта представлява сложна задача.

**6. Рангова корелация.** За всеки две рангови или метрични величини $X$ и $Y$ могат да се пресмятат коефициенти на рангова корелация, които могат да се разглеждат като ***непараметрични*** анализи на линейния коефициент на корелация. Ранговите величини могат да бъдат зададени непосредствено чрез ранговете на отделните наблюдения или чрез някакви натурални стойности, които се преобразуват към рангове. За метричните величини натуралните стойности се преобразуват към рангове, при което най-малката стойност получава ранг 1, следващата получава ранг 2 и т.н. Съвпадащите стойности получават среден ранг.

Коефициентът на рангова корелация $r_s$ на ***Спирмън*** се получава от формулата за коефициента на линейна корелация, когато за стойности се вземат съответните рангове. Коефициентът на рангова корелация $\tau$ на ***Кендал*** има по-сложна изчислителна формула, която няма да привеждаме. Тези два коефициента притежават общите характеристики на коефициента на линейна корелация $r$ на Пирсън и имат идентично познавателно значение. В типичния случай на метрична величина, стойностите на $r$, $r_s$ и $\tau$ са близки, като обикновено $\tau$ има най-малка по абсолютна стойност.

Коефициентите на рангова корелация се предпочитат в случая на извадка с малък обем, $n<30$.

В качеството на илюстративен пример ще приведем стойностите на тези корелационни коефициенти между средния успех от зимния и летния семестър за група от $n=12$ девойки. Резултатите показват

Таблица 8.5.

| | |
|---|---|
| коефициент на Пирсън | $r=0.825$ |
| коефициент на Спирмън | $r_s=0.887$ |
| коефициент на Кендал | $\tau=0.751$ |