

Chí-kvadrát test *Chi square test*

Chí-kvadrát test je statistická metoda používaná k určení, zda existuje významná asociace mezi dvěma kategorickými proměnnými, často v tabulce 2 krát 2. Test je vhodný pro větší soubory dat a udává přibližnou hodnotu p , pro menší soubory je doplňován o Yatesovu korekci.

Chí-kvadrát test

Chí-kvadrát test je statistická metoda používaná k určení, zda existuje významná asociace mezi dvěma kategorickými proměnnými. Jednoduše řečeno, pomáhá nám pochopit, zda existuje vztah mezi dvěma veličinami, které jsou kategorie nebo skupiny.

Představte si, že máte soubor dat s veličinami "kouření" a "diagnóza rakoviny plic." Chcete vědět, zda existuje vztah mezi těmito dvěma proměnnými. Chí-kvadrát test vám může pomoci zjistit, zda existuje významná asociace mezi kouřením (kategorie jako "kuřák" a "nekuřák") a diagnózou rakoviny plic (kategorie jako "diagnostikován rakovinou plic" a "nediagnostikován rakovinou plic").

Jak to funguje?

Chí-kvadrát test porovnává pozorovaná data s daty, která bychom očekávali, kdyby mezi proměnnými neexistoval žádný vztah. Vypočítá číslo nazývané chí-kvadrát statistika, které nám říká, jak moc se pozorovaná data liší od dat, která bychom očekávali, kdyby mezi proměnnými neexistovala žádná asociace.

Kroky k provedení chí-kvadrát testu:

1. Formulace hypotéz

- Nulová hypotéza (H_0): Neexistuje žádná asociace mezi proměnnými.
- Alternativní hypotéza (H_1): Existuje asociace mezi proměnnými.

2. Sbírání a organizace dat

- Shromážděte data pro obě kategorické proměnné.
- Organizujte data do kontingenční tabulky, také známé jako křížová tabulka nebo dvoucestná tabulka. Pro tabulky s 2x2 políčky se užívá název čtyřpolní tabulka.

3. Výpočet očekávaných frekvencí a ověření předpokladů testu

- Vypočítejte očekávané frekvence pro každou buňku v kontingenční tabulce za předpokladu, že mezi proměnnými neexistuje žádná asociace. Očekávané frekvence se vypočtou podle vzorce:

$$\text{očekávaná frekvence} = \frac{\text{součet řádku} \times \text{součet sloupce}}{\text{celkový součet}}$$



- Ověřte předpoklady testu, že všechny očekávané frekvence jsou větší než 5. Pokud tato podmínka není splněna, není možno test použít.

4. Výpočet chí-kvadrát statistiky

- Vypočítejte chí-kvadrát statistiku pomocí vzorce, kde:
 - χ^2 = Chí-kvadrát statistika
 - O = Pozorovaná frekvence
 - E = Očekávaná frekvence

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- Sčítáme přes všechna políčka kontingenční tabulky. Yatesova korekce pro malé soubory má vzorec trochu upravený.

5. Stanovení významnosti

- Určete počet stupňů volnosti (df): $df = (\text{počet řádků} - 1) \times (\text{počet sloupců} - 1)$.
- Porovnejte vypočítanou chí-kvadrát hodnotu s kritickou hodnotou chí-kvadrát rozdělení se stupni volnosti (df) ze statistické tabulky nebo pomocí softwaru.
- Pokud je vypočítaná chí-kvadrát hodnota větší než kritická hodnota, zamítnete nulovou hypotézu a usoudíte, že existuje významná asociace mezi proměnnými.

Jak výsledek Interpretovat?

Při použití softwaru dostaneme jako výsledek testu p-hodnotu (pravděpodobnostní hodnota). Pokud je p-hodnota menší než předem stanovená úroveň významnosti (obvykle 0,05), zamítnete nulovou hypotézu a usoudíte, že existuje významná asociace mezi proměnnými.

Pokud je p-hodnota větší než úroveň významnosti, nezamítáte nulovou hypotézu, což naznačuje, že neexistuje dostatek důkazů k závěru, že existuje asociace mezi proměnnými.

Jak test prakticky provést?

Chí-kvadrát test patří mezi nejčastěji používané statistické testy – dle analýz je využíván přibližně v každém pátém publikovaném vědeckém článku. Jeho praktická znalost a správná interpretace jsou proto nezbytné pro každého, kdo se zabývá analýzou kategoriálních dat.

Cílem chí-kvadrát testu je posoudit, zda existuje statisticky významný rozdíl mezi skutečnými (pozorovanými) a očekávanými četnostmi v jednotlivých kategoriích. Očekávané četnosti představují hodnoty, které bychom pozorovali, pokud by nulová hypotéza byla pravdivá – tedy pokud by mezi sledovanými proměnnými neexistovala žádná asociace. Pokud by byly pozorované a očekávané četnosti shodné, hodnota testového kritéria χ^2 by byla nulová. V praxi se hodnota χ^2 interpretuje obtížně, a proto vhodné mít p-hodnotu, která udává pravděpodobnost, že rozdíl mezi pozorovanými a očekávanými četnostmi vznikl náhodou.

P-hodnotu nelze jednoduše vypočítat, je třeba požit vhodný software nebo online kalkulačtor. Je však důležité si uvědomit, že výpočet χ^2 je sice technicky jednoduchý, ale test je pouze asymptotický (dobré výsledky dává pouze při dostatku dat). Pro zpřesnění se v některých



případech, zejména pokud máme málo pozorování (<30) u kontingenčních tabulek 2×2, využívá Yatesova korekce, která zohledňuje odchylky způsobené malým počtem pozorování.

Chí-kvadrát test má však své limity. Není vhodný pro malé výběrové soubory, u nichž nejsou splněny předpoklady testu – tedy tehdy, když některá očekávaná četnost klesne pod hodnotu 5. V těchto situacích je vhodnější použít Fisherův exaktní test, který poskytuje přesnější výsledek.

V praxi tak pro výpočet nejčastěji využijeme statistické programy či online nástroje. V poslední době lze rovněž využít i umělou inteligenci v podobě jazykových modelů, které jsou schopné jednoduše připravit např. skript pro statistický program R, Excel nebo Python.

Příklad použití chí-kvadrát testu

Řekněme, že jste provedli chí-kvadrát test, abyste zjistili vztah mezi pohlavím (muž/žena) a preferencí volby (kandidát A/kandidát B). Po provedení testu jste zjistili chí-kvadrát statistiku 5,36 a p-hodnotu 0,021. Protože p-hodnota je menší než 0,05, zamítli byste nulovou hypotézu a usoudili, že existuje významná asociace mezi pohlavím a preferencí volby.

Jak jsme k výpočtu dospěli?

Nejprve uspořádejte svá data do kontingenční tabulky. Tato tabulka by měla mít dva řádky představující jednu kategorickou proměnnou a dva sloupce představující druhou kategorickou proměnnou. Každá buňka v tabulce představuje kombinaci kategorií z obou proměnných. Pro každý řádek a sloupec v kontingenční tabulce vypočítejte celkový počet pozorování. Tyto součty vám pomohou vypočítat očekávané frekvence.

Představme si, že máme kontingenční tabulku reprezentující vztah mezi pohlavím a preferencí volby:

	Kandidát A	Kandidát B	Celkem
Muž	30	20	50
Žena	40	60	100
Celkem	70	80	150

Pro výpočet očekávané frekvence pro buňku představující "Muž" a "Kandidát A," použijeme vzorec:

$$\text{očekávaná frekvence } E_{MA} = \frac{(50 \times 70)}{150} = 23,33$$



Podobně budeme postupovat u všech dalších možných kombinací:

$$E_{MB} = \frac{(50 \times 80)}{150} = 26,67$$

$$E_{\bar{z}A} = \frac{(100 \times 70)}{150} = 46,67$$

$$E_{\bar{z}B} = \frac{(100 \times 80)}{150} = 53,33$$

Všechny očekávané frekvence jsou větší než 5, můžeme tedy použít chí-kvadrát test.

Výsledky si pro přehlednost můžeme uvést do tabulky s pozorovanými (O) a očekávanými (E) hodnotami:

	Kandidát A O (E)	Kandidát B O (E)	Celkem
Muž	30 (23,33)	20 (26,67)	50
Žena	40 (46,67)	60 (53,33)	100
Celkem	70	80	150

Dalším krokem je výpočet chí-kvadrát statistiky dle vzorce uvedeného v bodě 4 a následně suma výsledků pro každou možnost.

Muž – kandidát A

$$\frac{(30 - 23,33)^2}{23,33} = \frac{6,67^2}{23,33} = \frac{44,49}{23,33} = 1,91$$

Muž – kandidát B

$$\frac{(20 - 26,67)^2}{26,67} = \frac{6,67^2}{26,67} = \frac{44,49}{26,67} = 1,67$$

Žena – kandidát A

$$\frac{(40 - 46,67)^2}{46,67} = \frac{6,67^2}{46,67} = \frac{44,49}{46,67} = 0,95$$

Žena – kandidát B

$$\frac{(60 - 53,33)^2}{53,33} = \frac{6,67^2}{53,33} = \frac{44,49}{53,33} = 0,83$$

Souhrnná suma: $\chi^2 = 1,91 + 1,67 + 0,95 + 0,83 = 5,36$.

Dále zvolíme hladinu významnosti (obvykle je to 0,05).

Vypočítáme stupně volnosti (df) a v tabulkách nalezneme příslušnou kritickou hodnotu.



$df = (\text{počet řádků} - 1) \times (\text{počet sloupců} - 1) = (2 - 1) \times (2 - 1) = 1$.

Kritická hodnota z tabulek (při hladině významnosti 0,05 a $df = 1$) je: 3,841.

Výsledek

$\chi^2 = 5,36 > 3,841 \rightarrow$ zamítáme nulovou hypotézu a můžeme vypočítat hodnotu p .

P-hodnotu lze určit z tabulek pouze přibližně. Přesně ji spočteme pomocí statistického nástroje či online kalkulatoru. Pro $\chi^2 = 5,36$ a $df = 1$ je $p = 0,021$.

Protože vypočítaná hodnota chí-kvadrátu je větší než kritická hodnota a hodnota $p < 0,05$, existuje statisticky významná asociace mezi pohlavím a volební preferencí. Jinými slovy, to, zda člověk volil kandidáta A nebo B, pravděpodobně souvisí s tím, zda volil muž či žena.

Závěr

Opět platí, že není potřeba umět pomocí vzorců vypočítat chí-kvadrát test, protože lze využít off-line i online software pro jeho výpočet. Použití software je vhodnější, protože dostaneme p-hodnotu, kterou nelze jednoduše vypočítat. Podstatné je uvědomit si, zda test mohu použít (jsou splněny předpoklady testu), jak s údaji nakládat v praxi a jak je správně interpretovat.

Body k zapamatování

- Chí-kvadrát test je použit jako statistický test u každého pátého publikovaného článku, znalost jeho použití a interpretace je tak velmi důležitá.
- Chí-kvadrát test hodnotí míru rozdílu mezi skutečnými a očekávanými četnostmi naměřených hodnot. Očekávaná četnost představuje četnost, kterou bychom pozorovali, pokud by nulová hypotéza byla pravdivá (pokud by skutečné a očekávané četnosti byly stejné, hodnota χ^2 by byla 0).
- Hodnotit přímo hodnotu χ^2 je velmi obtížné, v praxi proto hledáme p-hodnotu.
- Je potřeba si uvědomit, že chí-kvadrát test je jednoduché vypočítat, ale výsledek je pouze přibližná p-hodnota. U malých souborů je používána Yatesova korekce, která hodnotu p zpřesňuje.
- Test není vhodný při nesplnění předpokladů, pak se používá Fisherův exaktní test.

Zdroje

- [1] HARRIS, Michael a TAYLOR, Gordon. Medical statistics made easy. Fourth edition. Banbury, England: Scion, 2021. ISBN 978-1-911510-63-5.

