I Chí-kvadrát test v kontingenčních tabulkách

I.I Úvod

Pokud máme dvě kategoriální proměnné, můžeme prozkoumat jejich vztah tak, že obě proměnné vložíme do tabulky.

Motivační příklad. Z produkce firmy vybereme vzorek 200 součástek vyrobených třemi různými stroji (označme je A,B,C). U každé součástky ve vzorku bylo prozkoumáno, jestli je vadná. Je také zaznamenán stroj, který vytvořit tuto součástku . Výsledky jsou následující:

	Stoj			
	A	B	C	celkem
vadných	8 (12,9%)	6 (8,8%)	12 (17,1%)	26 (13%)
v pořádku	54	62	58	174
celkem	62	68	70	200

Manažer si přeje určit, jestli existuje nějaký vtah mezi procentem vadných součástek a strojem, který je vytvořil. Musí tedy stanovit nulovou a alternativní hypotézu tak, aby pomocí ní mohl odpovědět na svoji výzkumnou otázku. Obecně nulová hypotéza říká, že mezi řádkovými a sloupcovými proměnnými neexistuje žádná souvislost, a alternativní hypotéza, že tam nějaký vztah je. V tomto případě je ekvivalentním tvrzením, že:

H_0: Není žádný rozdíl mezi stroji v procentním poměru vadných součástek.
H_1: Je rozdíl mezi stroji v procentním poměru vadných součástek.

Abychom otestovali nulovou hypotézu, porovnáme pozorované hodnoty buněk s očekávanými hodnotami spočítanými za předpokladu, že nulová hypotéza platí. Pokud by nulová hypotéza platila, řádkové (nebo sloupcové) procentní poměry vadných komponent by byly stejné. Očekávanou četnost spočítáme jako součin řádkové marginální četnosti a sloupcové marginální četnosti podělený rozsahem souboru.

V našem vzorku jsou očekávané četnosti buněk následující:

	Stoj		
	A	B	C
vadných	$\frac{26 \cdot 62}{200} = 8,06$	$\frac{26 \cdot 68}{200} = 8,84$	$\frac{26 \cdot 70}{200} = 9,10$
v pořádku	$\frac{174 \cdot 62}{200} = 53,94$	$\frac{174 \cdot 68}{200} = 59,16$	$\frac{174 \cdot 70}{200} = 60,90$

Všimněte si, že (v souladu s nulovou hypotézou) očekávané řádkové procentní podíly jsou stejné ($\frac{8,06}{62} = \frac{8,84}{68} = \frac{9,10}{70} = 13\%$ = celkový procentní podíl vadných součástek).

I.II Provedení Chí kvadrát testu

Abychom otestovali nulovou hypotézu, vypočítáme statistiku, která porovnává celý soubor pozorovaných četností se souborem očekávaných četností. Tato statistika se nazývá Chí kvadrát a je definována jako:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

kde O_{ij} = pozorované (**o**bserved) četnosti buněk a E_{ij} = očekávané (**e**xpected) četnosti buněk a suma jde přes všech s x r buněk v tabulce, kde r = počet řádků a s = počet sloupců v tabulce. V našem případě:

$$\chi^2 = \frac{(8-8,06)^2}{8,06} + \frac{(6-8,84)^2}{8,84} + \cdots + \frac{(58-60,90)^2}{60,90} = 2,11$$

Počet stupňů volnosti je (r - 1) x (s - 1). V našem případě je to (3 – 1) x (2 - 1) = 2. P-hodnota Chí kvadrát testu je v tomto případě P(χ^2_2>2,11). V tabulkách Chí kvadrát rozložení si dohledáme p-hodnotu 0,25 <p <0,5. V tomto případě jsme tedy neprokázali (nemáme tedy žádný důkaz), že by se procentní podíly vadných komponent lišily mezi stroji.

I.III Předpoklady použití chí-kvadrát testu ve kontingenčních tabulkách

Chí kvadrát testy má smysl použít pouze, pokud je vzorek přiměřené velikosti. Je možné použít následující návody:

1. Pro 2 x 2 tabulky:
   - Pokud je celková velikost větší než 40, můžeme Chí kvadrát test použít.
   - Pokud je celková velikost vzorku mezi 20 a 40 a nejmenší očekávaná četnost je alespoň 5, Chí kvadrát test lze použít.
   - Jinak je nutné provést Fisherův přesný test.
2. Pro jiné tabulky:
   - Chí kvadrát lze použít, pokud maximálně 20% očekávaných četností je menších než 5 a žádná není menší než 2.