Srovnání skupin

Výsledkem srovnání skupin je jednak absolutní **velikost rozdílu** mezi skupinami a jednak **statistická významnost** tohoto rozdílu. Oba parametry jsou důležité a nesmí se zaměňovat. Pokud je rozdíl statisticky nevýznamný, znamená to, že může být jen dílem náhody a celkem nezáleží na tom, jaká je jeho velikost. Pokud je ale statisticky významný, je potřeba kriticky zhodnotit jeho velikost. Při dostatečně velkém vzorku jsme totiž schopni odhalit jako statisticky významné i nepatrné rozdíly, jejichž praktický význam může být… nepatrný.

Ke zhodnocení významnosti rozdílu mezi sledovanými skupinami (nebo jednou skupinou a nějakým referenčním údajem-číslem, např. při hodnocení pravdivosti měření) používáme obvykle **statistické testy**. Při srovnání volíme dvě hypotézy (tvrzení), které si konkurují. Nepřítomnost rozdílu/efektu nazveme **nulová hypotéza H_0**, a přítomnost rozdílu/efektu **alternativní hypotéza H_1**. Na základě statistického testu potom rozhodneme, jestli zamítneme H_0 nebo ne (pozor, nezamítnout neznamená přijmout).

Je třeba vybrat správný test a určit **hladinu významnosti α**. To je nejvyšší tolerovaná pravděpodobnost nesprávného zamítnutí nulové hypotézy (tzv. chyby I. druhu). Obvykle se volí α=0,05, tj. 5% pravděpodobnost chyby I. druhu. Pravděpodobnost, že v případě, že jsme H_0 nezamítli, tato skutečně platí, se nazývá **síla testu**.

Výsledkem statistického testu je často tzv. **P hodnota**, tedy přímo pravděpodobnost chyby I. druhu. Pokud je P<α, zamítáme H_0.

Při srovnávání skupiny spojitých dat s referenční hodnotou můžeme také postupovat tak, že spočítáme **95% interval spolehlivosti průměru**. Pokud leží referenční hodnota vně intervalu spolehlivosti, je odlišná od průměru sledované skupiny na hladině významnosti 0,05.

Statistické testy při srovnání skupin u kvalitativních dat

Kvalitativní data nejprve shrneme do tabulky. Pokud sledujeme jednu charakteristiku (veličinu), jde o tzv. **tabulku četností**. Pokud sledujeme dvě veličiny a jejich souvislost, jde o tzv. **kontingenční tabulku**.

barva	n_{barva}	n_{barva} / n
bílá	621	0,904
černá	66	0,096
∑	687	1,00

Ve třetím sloupci tabulky četností máme **pravděpodobnost výskytu** daného znaku. **Střední chyba (analogie SEM)** odhadu pravděpodobnosti se rovná √(p(1-p)/n).

Na rozdíl od spojitých veličin, u **binomických** veličin nelze pro zjištění, zda je výsledná pravděpodobnost rovna zvolené hodnotě, použít interval spolehlivosti. Binomické rozdělení má totiž pro různé hodnoty pravděpodobnosti různý rozptyl. Místo toho používáme **Binomický test** (dostupný v Excelu).

Při srovnávání skupin vlastně analyzujeme shodu struktury (homogenitu) jednotlivých výběrů.

Srovnání 2 skupin **binárních dat**: **Čtyřpolní tabulka**. Rozdíl mezi skupinami testujeme pomocí **Pearsonova (Chi square) testu** (dostupný v Excelu). Předpoklady tohoto testu jsou nezávislost pozorování (každý prvek patří jen do jedné buňky) a četnosti ve většině případů > 5 a ve všech případech > 2. Pokud máme menší četnosti (a v každém případě, pokud je celkové n < 20, použijeme **Fisherův exaktní test**. V případě, že se jedná o párová data, tedy vlastně jednu skupinu vzorků ve dvou

různých stavech (např. před a po léčbě), použijeme **McNemarův test** (pro párová data ovšem vypadá čtyřpolní tabulka jinak).

nemoc	1. skupina	2. skupina	celkem
ano	16 (a)	10 (b)	26
ne	4 (c)	12 (d)	16
celkem	20	22	42

Párová data	Po - ano	Po - ne
Před - ano	10 (A)	6 (B)
Před - ne	0 (C)	4 (D)

Pro **nominální/ordinální data** nebo pro srovnání více než 2 skupin bude tabulka větší. Pokud chceme zjistit, mezi kterými skupinami je rozdíl, musíme tabulku rozdělit do 2×2 tabulek.

Korespondenční analýza patří mezi ordinační analýzy. Pracuje se vzorky popsanými dvěma kategoriálními proměnnými s velkým počtem kategorií, konkrétně umístí jednotlivé kategorie těchto proměnných do tzv. korespondenční mapy. Místo velké kontingenční tabulky s frekvencemi získáme dvourozměrnou mapu, ve které jsou blízko sebe zobrazeny kategorie stejné proměnné, které mají podobný profil, pokud jde o druhou proměnnou. Co se týká srovnání kategorií různých proměnných, záleží na směru a vzdálenosti od počátku soustavy souřadnic. Větší vzdálenost od počátku značí větší vliv. Stejný směr od počátku značí pozitivní korelaci a opačný směr negativní korelaci.