

Měření závislosti kvalitativních znaků

- Kvalitativní znaky mají slovní charakter a získáváme je v sociologických průzkumech, při terénním šetření apod.
- Slovní charakter mají odpovědi na otázky týkající se např. pohlaví, vzdělání nebo povolání respondenta atd.
- K popsání vztahu závislosti spojených kvantitativních veličin slouží **korelační koeficient**.
- K charakterizování závislosti kvalitativních znaků slouží tzv. **kontingenční tabulky**

Klasifikace kvalitativních znaků:

- Podle počtu možných obměn dělíme znaky na **alternativní** (také dvojné) nabývající pouze dvou obměn a znaky **množné**, nabývající více než dvou obměn,
- Podle možnosti určit objektivní pořadí obměn na znaky, které **mají pořadový charakter** (např. vzdělání, stupeň souhlasu či nesouhlasu apod.) a znaky, které tento charakter nemají (např. povolání, typ absolvovaného vzdělání, značka výrobku) a u nichž tedy objektivní uspořádání není možné,
- Podle toho zda lze jednoznačně vymezit kde „začíná“ a „končí“ každá obměna znaku nebo nelze (např. u barevných odstínů) dělíme znaky na **nespojité a spojitě**.

Statistická analýza kvalitativních znaků:

- Statistické zpracování jednoho slovního znaku spočívá jednak v jeho třídění
- Nejčastěji se jedná o prosté třídění podle jednotlivých obměn slovního znaku a o stanovení absolutních nebo relativních četností.
- V omezené míře lze určovat charakteristiky úrovně (modus, u pořadových znaků medián, nikdy aritmetický průměr).
- Existují i speciální charakteristiky proměnlivosti.
- O měření závislosti má smysl uvažovat, je-li k dispozici dvojice slovních znaků.

Měření závislosti kvalitativních znaků

Spočívá v sestavení tzv. **kontingenční tabulky**

Z kontingenční tabulky lze určit intenzitu závislosti ve dvojici slovních znaků.

Nelze z ní však určit průběh závislosti. O směru závislosti má smysl se vyslovit pouze v případě pořadových slovních znaků.

Máme-li dva alternativní znaky dostaneme tzv. čtyřpolní tabulku.

Měření závislosti kvalitativních znaků

Obecně může mít každý kvalitativní znak A r tříd a znak B s tříd. Výsledky šetření potom sestavujeme do kontingenční tabulky $r \times s$.

Pozorované četnosti v jednotlivých buňkách označujeme dvěma indexy – obecně n_{ij} .

Také marginální četnosti mají dva indexy.

Ten, přes který je sčítáno je označen hvězdičkou – tedy n_{2*} značí součet četností v druhé řádce, n_{*1} značí součet četností v prvním sloupci.

Tabulka bývá doplněna hodnotami procentuálních (relativních) četností. Častým požadavkem je konstantní délka intervalů tvořících třídy.

Stejně jako v případě kvantitativních znaků ověřujeme i zde existenci vztahu testy významnosti a hodnotíme ho vhodnou mírou závislosti.

Kontingenční tabulka typu $r \times s$

Tříděný znak		Znak B					Součet	
		b_1	b_2	...	b_j	...		b_s
Znak A	a_1	n_{11}	n_{12}				n_{1s}	n_{1*}
	a_2	n_{21}						n_{2*}
	\vdots							\vdots
	a_i				n_{ij}			n_{i*}
	\vdots							\vdots
	a_r	n_{r1}					n_{rs}	n_{r*}
	Součet	n_{*1}	n_{*2}	...	n_{*j}	...	n_{*s}	$n_{**} = n$

Posuzování závislosti v kontingenčních tabulkách

Podmíněné četnosti uvnitř kontingenční tabulky mají podobný význam jako body korelačního diagramu — jejich rozmístění umožňuje usuzovat na charakter závislosti tříděných znaků.

Pro posouzení nezávislosti obou znaků můžeme vedle pozorovaných četností stanovit pro jednotlivá pole také očekávané (teoretické) četnosti :

$$n'_{ij} = \frac{n_{i*}n_{*j}}{n}$$

tedy jako součin okrajových četností příslušného řádku a sloupce dělený rozsahem souboru.

Pro každé pole kontingenční tabulky existuje dvojice četností - četnost pozorovaná a četnost vypočtená.

Hypotéza nezávislosti

Ukazatel, který pro tabulku jako celek měří rozdílnost pozorovaných a vypočtených četností v jednotlivých polích tabulky se nazývá čtvercová kontingence χ^2

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Je to bezrozměrná hodnota a platí: $\chi^2 \geq 0$

Hodnoty nula nabývá pouze v případě, že znaky v kontingenční tabulce jsou nezávislé.

Vypočtená hodnota χ^2 se porovnává na zvolené hladině významnosti p s kritickou hodnotou χ^2 rozdělení pro $(r-1)(s-1)$ stupňů volnosti.

Hypotézu zamítáme, jestliže vypočtená hodnota je větší než tabulková, případně, když jí příslušující p -hodnota je menší než zvolená hladina významnosti.

Koeficienty kontingence

Maximální možná hodnota čtvercové kontingence závisí na rozměrech kontingenční tabulky a rozsahu souboru - z toho důvodu není nejvhodnějším ukazatelem intenzity závislosti.

Na bázi čtvercové kontingence jsou konstruovány vhodnější ukazatele - **koeficienty kontingence**.

Jsou konstruovány tak, aby jejich hodnota závisela pouze na intenzitě závislosti.

Koeficienty kontingence měří intenzitu závislosti pro dvojici slovních znaků.

Pearsonův koeficient kontingence:

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

nabývá hodnot $0 \leq P < 1$

Příklad analýzy závislosti v tabulce r x s

Pro výběr 234 studentů zjišťujeme, zda existuje vztah mezi sportem, který provozují a sportovními pořady, které sledují v televizi.

Sestavíme tabulku typu 4 x 4:

Obľíbenost při sledování televize	Obľíbenost při sportování				Řádkové součty
	hry	atletika	gymnastika	plavání	
hry	133	6	2	4	145
atletika	15	10	4	3	32
gymnastika	4	1	25	0	30
plavání	9	0	1	17	27
Sloupcové součty	161	17	32	24	234

Hypotéza nezávislosti H_0 : Neexistuje vztah mezi provozovaným sportem a sportem sledovaným v TV.

Vypočtená hodnota testovacího kritéria $\chi^2 = 273,3$

Kritická hodnota z tabulek pro $p=0,05$ a $(4-1) \times (4-1) = 9$ stupňů volnosti:

Závěr: H_0 zamítáme, existuje významný vztah. $\chi^2 = 16,9$

Sílu tohoto vztahu lze posoudit Pearsonovým koeficientem

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{273,3}{273,3 + 234}} = 0,71$$

kontingence

Testování nezávislosti v tabulce 2 x 2

	Zájem o statistiku		řádkové součty
	ano	ne	
chlapci	a	b	a + b
dívky	c	d	c + d
sloupcové součty	a + c	b + d	n

Pro výpočet testovacího kritéria χ^2 v tabulce 2 x 2 můžeme využít zjednodušený vzorec:

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

Protože v 2x2 tabulce můžeme uvažovat i směr poruchy nulové hypotézy – proto musíme rozhodnout, zda použijeme test jednostranný či dvoustranný.

Kritické hodnoty jsou uvedeny v tabulce χ^2 -rozdělení o jednom stupni volnosti.

Příklad analýzy závislosti v tabulce 2 x 2

	Zájem o statistiku		řádkové součty
	ano	ne	
chlapci	30	36	66
dívky	11	63	74
sloupcové součty	41	99	140

Hypotéza nezávislosti H_0 : Relativní četnost studentů se zájmem o statistiku je nezávislá na pohlaví.

Vypočtená hodnota testovacího kritéria: $\chi^2 = \frac{140(30 \times 63 - 11 \times 36)^2}{41 \times 99 \times 66 \times 74} = 15,8$

Kritická hodnota χ^2 -rozdělení z tabulek pro $p=0,05$: 3,84

Závěr: H_0 zamítáme, existuje významný rozdíl.

Zájem u chlapců: $30/66 = 0,45$

Zájem u dívek: $11/74 = 0,14$

Chlapci mají zhruba 3x větší zájem o statistiku než dívky.