5.2 Charakteristiky rozložení nominální proměnné

Rozložení nominálních znaků se od sebe odlišují v několika aspektech. Jednak je to obsazení každé jednotlivé kategorie, jednak to jsou vlastnosti rozložení $\{f_k\}$ jakožto celku. Každou z kategorií můžeme hodnotit jak samu o sobě, tak ve vztahu k ostatním. Z tohoto hlediska jsou pro analýzu dat zajímavé ty, které jsou málo četné vzhledem k ostatním (většinou je vynecháváme nebo slučujeme), a pak ty, které jsou proti ostatním obsazeny výrazněji, jsou četnější.

Zvláštní místo mezi hodnotami znaku zaujímá *modální kategorie* (*modus, modální hodnota*), tj. nejčetněji obsazená hodnota znaku. Označuje to, co je v daném souboru typické, charakteristické, nejčastější. Rozložení může mít dvě nebo i více modálních kategorií. Podle toho mluvíme o *unimodálním, bimodálním, trimodálním*, resp. *k-modálním* rozložení.

(5.13) modus = *Mo* = *nejčetnější hodnota znaku*.

Spolu s modální kategorií určujeme také *modální četnost*

(5.14)
$$n_{Mo} = \max_{1 \leq k \leq K} n_k,$$

$$f_{Mo} = \max_{1 \leq k \leq K} f_k .$$

Modus patří k charakteristikám polohy, které označují typické hodnoty rozložení.

Základní vlastností všech typů dat je jejich *variabilita, rozptýlenost, heterogenita* vzhledem k dané proměnné. Vlastnost heterogenity resp. vnitřní diferenciace souboru je sociologickým aspektem, který je nutno měřit: velká nebo naopak malá rozrůzněnost souboru může mít své příčiny, které sociologicky objasňujeme. Statistická analýza, která vede k odhalování příčin a o níž bude řeč později, je celá založena na statistickém vysvětlení variability rozložení. Charakteristiky variability mají ještě druhý interpretační význam: ukazují do jaké míry jsou data koncentrována kolem své charakteristické hodnoty určené použitou mírou polohy (v případě nominálního znaku je to modální kategorie), jak dalece je tato hodnota typická pro celý soubor.

Míry variability:

1. *Variační poměr*

(5.15)
$$v = v(\mathbf{A}) = 1 - \frac{n_{Mo}}{n} = 1 - f_{Mo} =$$

$$= 1 - \frac{\text{četnost modální kategorie}}{\text{velikost souboru}}$$

(pokud je více modálních kategorií bereme četnost jen jednou).

2. *Nominální variance (nomvar)*

$$\text{nomvar} = \text{nomvar}(\boldsymbol{A}) = 1 - \sum_{k=1}^{K} f_k^2 = \tag{5.16}$$

$$= 1 - \sum_{k=1}^{K} \left(\frac{n_k}{n}\right)^2 = \frac{n^2 - \Sigma n_k^2}{n^2} =$$

$$= \sum_{k=1}^{K} f_k(1 - f_k) =$$

= *podíl všech dvojic jednotek, které nemají stejnou hodnotu znaku.*

3. *Entropie*

$$H = H(\boldsymbol{A}) = -\sum_{k=1}^{K} f_k \ln f_k \tag{5.17}$$

(ln je přirozený logaritmus a $0 \times \ln 0 = 0$).

4. *Normalizovaná nominální variance*

$$\text{norm. nomvar} = \text{norm. nomvar}(\boldsymbol{A}) = \frac{K\left(1 - \sum_{k=1}^{K} f_k^2\right)}{K-1}. \tag{5.18}$$

5. *Normalizovaná entropie*

$$H^* = H^*(\boldsymbol{A}) = \frac{H}{\ln K} = -\frac{\sum_{k=1}^{K} f_k \ln f_k}{\ln K}. \tag{5.19}$$

. Vlastnosti těchto měr můžeme shrnout:

1. Jsou rovny nule, právě když celý soubor je soustředěn do jedné kategorie, tj. $n_{Mo} = n$, $f_{Mo} = 1$ (případ nulového rozptýlení, úplné homogenity).

2. Maximální hodnoty nabývají pro rovnoměrné rozložení dat ($f_k = \frac{1}{k}$ pro $k = 1$, ..., K, případ maximálního rozptýlení).

3. Mezi nulou a maximální hodnotou charakterizují stupeň vnitřní heterogenity dat — čím vyšší hodnota, tím vyšší heterogenita souboru.

4. Obor hodnot, je-li počet kategorií K předem dán, je pro

$$v \in \left\langle 0, \frac{K-1}{K} \right\rangle,$$

$$\text{nomvar} \in \left\langle 0, \frac{K-1}{K} \right\rangle,$$

$$H \in \langle 0, \ln K \rangle,$$

$$\text{norm. nomvar} \in \langle 0, 1 \rangle,$$

$$H^* \in \langle 0, 1 \rangle.$$

Pro výzkumnou praxi lze doporučit:

— v pro rychlé výpočty a charakteristiku typičnosti modální hodnoty;

— nomvar pro běžnou analýzu dat a programování (i když je výpočet pracnější, odráží celou strukturu tabulky, zatímco v vychází pouze z modální četnosti);

— norm. nomvar pro analýzy, v nichž chceme komparovat heterogenitu vzhledem k proměnným o různém, ale předem daném počtu hodnot, tj. chceme-li komparovat variability a abstrahovat od délky znaku, chceme-li vyjádřit podíl heterogenity k dosaženému maximu.

Obě míry založené na entropii (H, H^*) mají velmi podobné vlastnosti jako nominální variance. Té dáváme přednost ze dvou důvodů: má jednoduchý operacionální význam (podíl nepodobných dvojic), vychází z obecnějšího modelu, z něhož plynou též analogické a srovnatelné míry pro jiné typy znaků. H a H^* mají zase vztah k technikám testování hypotéz založeným na věrohodnostním poměru, na loglineárním modelu a teorii informace.

Míry variability uvedené v (5.15)—(5.19) se týkaly souboru dat z něhož byly odvozeny. Při odhadu populačních měr používáme v ve stejném tvaru, zatímco bodový odhad nomvar můžeme zlepšit použitím faktoru $\frac{n}{(n-1)}$, a dostat tak nevychýlený odhad

(5.20) $$\textit{odhad}\ \text{nomvar}\ \textit{pro populaci} = \frac{n}{n-1}\ [\text{nomvar}\ \textit{výběrového souboru}].$$

Pro velká n je tato úprava nepodstatná. Nepříjemné vlastnosti může mít při normalizaci (5.18), neboť při rovnoměrném rozložení překročí hranici 1 a má hodnotu $\frac{n}{(n-1)}$.

Příklad 5.3. Způsob získávání periodik.

Ve výzkumu čtenářů periodického tisku byla položena otázka na způsob získávání jednotlivých titulů. V tab. 5.2 je uvedeno 11 vybraných typických rozložení, u nichž můžeme dobře vidět význam jednotlivých měr. Modální kategorie i rozptýlenost (nepodobnost čtenářů vzhledem ke způsobu získávání) přinášejí kvalitativní údaje o vztahu čtenáře a každého periodika.

Modální kategorie tu určují, zda jde o periodikum výrazně předplatitelské, zda je jednotlivě kupováno či získáváno některou z dalších možností. Předplácením se modálně odlišují *I*, *J*, zatímco ostatní jsou kupována jednotlivě. Kvalitativní pohled (a z něho odvozená klasifikace titulů) může v tomto speciálním případě jít ještě dále: *A* má vysoké zastoupení kategorie „k dispozici v práci", čímž se odlišuje od všech ostatních; některé tituly jsou zajímavé vysokým procentem půjčování, rovněž je zajímavý vztah předplácení a kupování. Interpretaci ovlivňuje tedy nejen nejčetněji obsazená kategorie, ale i nejrůznější vztahy mezi kategoriemi. Heterogenita distribucí (dále měřená pomocí nominální variance) ukazuje, do jaké míry existuje typický způsob získávání. V tabulce 5.3 je uveden přehled hodnot základních měr polohy a variability a dále 95%-ní interval spolehlivosti pro procento pravidelných čtenářů každého titulu.

Tabulka 5.2. *Způsob získávání denního tisku a týdeníků u pravidelných čtenářů* (výběrový soubor o velikosti 1298 osob)

Periodikum	Předplácí[2]	Kupuje[3]	K dispozici v práci	Půjčuje si	Získává jinak	Celkem	n[1]	Procento z celku
A	20.2	49.1	20.2	9.8	0.6	99.9	173	13.3
B	24.1	50.0	0.0	20.4	5.6	100.1	54	4.2
C	16.2	64.7	2.7	14.3	2.2	100.1	414	31.9
D	2.6	82.9	0.7	12.1	1.7	100.0	414	31.9
E	22.5	55.9	9.2	10.8	1.5	99.9	195	15.2
F	37.4	52.7	3.2	5.8	1.0	100.1	313	24.1
G	4.8	67.4	3.5	21.7	2.6	100.0	230	17.7
H	11.7	77.6	2.4	7.1	1.2	100.0	170	13.1
I	47.8	37.4	6.2	7.1	1.5	100.0	337	26.0
J	48.3	24.1	6.9	16.4	4.3	100.0	116	8.9
K	8.6	68.1	3.2	18.8	1.3	100.0	559	43.1

Poznámky:

[1] Pravidelné čtení bylo zjišťováno samostatnou otázkou; n pro každý titul označuje počet osob, které prohlásily, že čtou všechna čísla nebo vynechají čtení jen výjimečně.

[2] Kategorie *„předplácí"* vznikla ze dvou hodnot: *„předplácí sám"*, *„předplácí někdo jiný z rodiny"*.

[3] Kategorie *„kupuje"* vznikla ze dvou hodnot: *„kupuje sám"*, *„kupuje někdo jiný z rodiny"*.

Z hodnot nominální variance je hned vidět, že ***D*** se odlišuje výraznější homogenitou (30% všech dvojic je různých, jen 17% osob je mimo modální kategorii „kupuje", která je tak velmi výrazným znakem čtenářů). Nejdiferencovanější jsou ***J***, ***A***, ***B***, ***I***. U ***A*** je to 67% nepodobných dvojic, přičemž maximálně jich může být nepodobných 80% (to plyne z max nomvar $= {}^{(K_K-1)}/_{K} = {}^{4}/_{5} = 0.8$). Normalizovaný nomvar ukazuje na stupeň, v jakém jsou v souboru vyčerpány všechny možné nepodobnosti, které lze realizovat (u ***A*** je to 83%, tedy velmi vysoká heterogenita). Konkrétní interpretace sociologická a případné závěry pak vyplynou podle konkrétních titulů, možností i záměrů vydavatelů.

Numerický výpočet jednotlivých měr ukážeme pro rozložení časopisu ***J***:

a) Nejčetnější hodnota — modus — odpovídá kategorii „předplácí", $f_{Mo} = 0.483$ (viz (5.13));

b) $v = 1 - 0.483 = 0.517$ (viz (5.15));

c) $\text{nomvar} = 1 - (0.483)^2 - (0.241)^2 - (0.069)^2 - (0.164)^2 - (0.043)^2 = 0.675$ (viz (5.16));

d) $H = -0.483 \ln 0.483 - 0.241 \ln 0.241 - 0.069 \ln 0.069 - 0.164 \ln 0.164 - 0.043 \ln 0.043 =$

$= 1.3107$ (viz (5.17));

e) $\text{norm.nomvar} = \dfrac{5 \times 0.675}{4} = 0.844$ (viz (5.18));

Tabulka 5.3. *Statistické charakteristiky pro rozložení dat z tab. 5.2.* ($K = 5$)

Periodikum	Modální kategorie	Modální procento	Variační poměr	nomvar	Norm. nomvar	Odhad nomvar	95% interval spol. pro % prav. čten.	
A	kupuje	49.1	.509	.668	.835	.672	11.5	15.1
B	kupuje	50.0	.500	.647	.809	.659	3.1	5.3
C	kupuje	64.7	.353	.533	.666	.534	29.4	34.4
D	kupuje	82.9	.171	.297	.371	.298	29.4	34.4
E	kupuje	55.9	.441	.617	.771	.620	13.2	17.2
F	kupuje	52.7	.473	.578	.723	.580	21.8	26.4
G	kupuje	67.4	.326	.494	.618	.496	15.6	19.8
H	kupuje	77.6	.224	.378	.473	.380	11.3	14.9
I	předpl.	47.8	.522	.623	.779	.625	23.6	28.4
J	předpl.	48.3	.517	.675	.844	.681	7.4	10.4
K	kupuje	68.1	.319	.492	.615	.493	40.4	45.8

f) $H^* = \frac{1.3107}{\ln 5} = 0.814$ (viz (5.19));

g) odhad nomvar pro základní soubor $= 0.681$ (viz (5.20));

h) 95%-ní interval spolehlivosti pro relativní četnost pravidelných čtenářů (viz **ISD 1**) je

$$0.089 \pm 1.96 \sqrt{\frac{0.089 \times 0.911}{1\,298}} = 0.089 \pm 0.015 = (0.074, 0.104).$$

5.3 Porovnání četností u kategorií nominální proměnné

Sociologické závěry často plynou ze vzájemného porovnání četností u jednotlivých kategorií proměnné. Pro soubor dat, z něhož počítáme rozložení $\{f_k\}_K$, je porovnání jednoduché, při zobecňování na základní soubor je však zajímavý vztah čísel p_j, p_k, nikoliv f_j, f_k. Proto si klademe otázku: je rozdíl čísel f_j, f_k natolik velký, že je nenáhodný, a tudíž způsobený rozdílem populačních hodnot p_j, p_k? Přístup statistické metodologie k tomuto problému spočívá v testování hypotézy, že mezi p_j a p_k není žádný rozdíl a ve zhodnocení toho, zda rozdíl $f_j - f_k$ je natolik velký, že hypotéze odporuje, či zda není tak velký, aby hypotézu vyvracel (a může tedy vzniknout i za její platnosti pouhým působením náhody při výběru). Řešení úlohy obsahuje metoda **THD 1**.

Metoda **THD 1**: Test hypotézy $p_i = p_j$.