ŘEHÁK, J., ŘEHÁKOVÁ, B. 1986. *Analýza kategorizovaných dat v sociologii*, Academia: Praha.

# Kapitola 4 Rozložení četností

## 4.1 Výskyt jevu, četnost, procento

Statistická analýza kategorizovaných dat je založena na studiu výskytů jednotlivých jevů a vztahů mezi nimi. *Jev* ve statistice chápeme obecně jako souhrn určitých projevů, vlastností, vztahů, podmínek, který je empiricky identifikovatelný, o němž můžeme vždy jednoznačně prohlásit, že buď nastal, nebo nenastal. Takový jednoznačný výsledek bývá však v praxi zatížen chybou a my se můžeme dopouštět dvou omylů s různými pravděpodobnostmi: jev nastal, ale prohlásíme, že nenastal, nebo jev nenastal, ale prohlásíme, že nastal. Proto je empirická identifikovatelnost jevu odstupňována podle pravděpodobnosti správného určení. Statistický jev je vždy vztažen k určitému komplexu podmínek, za nichž má odlišení „nastal-nenastal" smysl. Statistická analýza dat vychází z obou uvedených aspektů a jejich rozbor je jedním ze základních východisek postupů i konečné interpretace výsledků.

Jako příklad může sloužit často sledovaný jev **a** ≡ „spokojenost v práci". Místo něho však zjišťujeme jev **b** ≡ „respondent prohlásil, že je v práci spokojen". Vztah mezi oběma je sociologicky i metodologicky velmi složitý. Zatímco výskyty druhého jevu, **b**, zjišťujeme u respondentů vybraného souboru přesně (až na technické chyby v záznamu a přenosu dat), údaje o prvním jevu (**a**) jsou jím reprezentovány jen do jistého (většinou neurčeného) stupně spolehlivosti a platnosti. Respondenti nemusí vyjádřit skutečnou spokojenost, ať už záměrně či neúmyslně, pod vlivem nevhodně voleného sběru dat nebo v důsledku nějaké okamžité situace na pracovišti apod. Negací druhého jevu je jev opačný, „respondent neprohlásil, že je spokojen", který však může znamenat: respondent není spokojen, nechce svou spokojenost vyjádřit, nebo to dokonce vůbec prohlásit nemohl, protože nepracuje. Základní podmínkou sledování výskytu uvedeného jevu je tedy pracovní aktivita respondenta, která je např. u výzkumů pracovních kolektivů automaticky splněna, jindy však musí být zjišťována.

V praxi sociologicko-statistické analýzy je určení kontextu, v němž má smysl o jevu mluvit, totožné s určením souboru, k němuž má význam vztáhnout výskytovost jevu. Takový soubor lze vymezit pomocí vhodně zvoleného doplňkového jevu (**ā**) k jevu zkoumanému tak, že sjednocení obou charakterizuje kontext.





V uvedeném příkladě je možno doplněk k jevu „respondent prohlásil, že je v práci spokojen" volit jako jev „respondent prohlásil, že je v práci nespokojen". Opačné oběma jevům je „spokojen — nespokojen", společné (a tím i definicí souboru, k němuž vztahujeme výskytovost) je „prohlásil", „v práci" a všechna obecně platná omezení v daném šetření. Logicky tedy z analýzy vynecháme ekonomicky neaktivní respondenty a ty, kteří o spokojenosti nevypověděli.

Kontext, ke kterému vztahujeme výskytovost jevu, může být pro různé cíle definován různými způsoby, v rozmanité šíři a podmíněnosti. Každá statistická analýza je podmíněná: zvolené omezující podmínky jsou základním kvalitativním východiskem pro interpretaci statistických výsledků. Prakticky je podmíněná analýza prováděna vhodnou redukcí souboru dat. Obzvlášť silně se nutnost určení významových souvislostí projeví u komparace souborů, která je možná jen při srovnatelném základu. V běžné praxi se setkáváme s celou řadou rušivých vlivů, s nimiž není vždy lehké se vyrovnat: mateřská dovolená a vojenská služba jako důvody absence v podniku pro ženy a muže a různé věkové skupiny; otázka po důvodech změny zaměstnání pro osoby, které nejsou zaměstnány, nebo u nichž ke změně nedošlo; zjišťování postojů a názorů u osob, které si je nevytvořily, či si je dokonce ani vytvořit nemohly; typy školního vzdělání pro různé věkové kategorie apod.

Statistické jevy, jejich identifikovatelnost i způsob identifikace a komplex podmínek, za kterých má smysl o nich mluvit, se určují nejen u každé analýzy, ale už při přípravě sběru dat, při jejich záznamu a přenosu, při tvorbě dotazníků a záznamových listů, instrukcí pro pozorování, tazatele apod. Interpretace výsledků se opírá nejen o číselné závěry, ale i o rozbor empirické situace, metodologie sběru i teorii vztahů mezi podstatovými jevy, jež nás zajímají, a zjišťovanými empirickými jevy, které jsou vlastním předmětem statistické analýzy.

Statistický jev se váže ke statistické jednotce, u níž nastává, k jejímu místu, času, kontextu. Při sběru dat zjišťujeme u každé statistické jednotky, zda u ní jev nastal, nenastal, či nastat nemůže. U $n^*$ jednotek souboru tak máme empirický údaj: $m$ = počet jednotek, u nichž jev nastal, $\bar{m}$ = počet jednotek, u nichž jev nenastal, $m^*$ = počet jednotek, u nichž jev nemá smysl, $m'$ = počet jednotek, u nichž chybí informace, nebo je zjevně chybná. Z analýzy vynecháme $M = m^* + m'$ jednotek (tzv. vynechávaná data) a pro daný jev pracujeme se souborem o velikosti $n = m + \bar{m} = n^* - m^* - m'$.

Rozdíl mezi absolutním a poměrovým ukazatelem výskytu je dán otázkami: „kolik?" a „jaká část? (jaký podíl?)". V sociologické analýze ve většině případů pracujeme s poměrovými údaji, které jsou charakterizovány *relativními četnostmi* jevů $f = \frac{m}{n}$, tj. podílem souboru, u něhož jev nastal. Doplňková relativní četnost opačného jevu je $g = 1 - f = \frac{\bar{m}}{n}$. Někdy určujeme také podíl vynechávaných dat pro daný jev: $v = \frac{m' + m^*}{n^*}$. V praxi většinou uvádíme stonásobky relativních četností,

kterým říkáme *procenta*. Vlastnosti relativních četností jsou velmi jednoduché:

a) Relativní četnost můžeme určovat vždy, existuje-li neprázdný ($n > 0$) soubor jednotek, pro něž má jev smysl.

b) $f = 0$ právě když jev vůbec nenastal.

c) $f = 1$ právě když jev nastal u všech jednotek.

d) Čím vyšší je $f$, tím častější je jev.

Při analýze více jevů **a**, **b**, **c**, ... značíme obvykle četnosti $n_a$, $n_b$, $n_c$, ... resp. $f_a$, $f_b$, $f_c$, ...

Absolutní četnost $m$ má někdy sama o sobě praktický význam (např. počet osob, které odešly z pracovního kolektivu, musí být nahrazen bez ohledu na velikost skupiny), většinou však nás zajímá výskytovost jako podíl počtu výskytů v souboru (onemocní-li pět osob v třicetičlenném kolektivu, je to méně závažné, než onemocní-li stejný počet osob v patnáctičlenném kolektivu).

Každý jev **a** určuje jednoznačně dichotomickou proměnnou $\boldsymbol{A} = (a, \bar{a})$ = („jev **a** nastal", „jev **a** nenastal"); proto analýzu výskytovosti jevu provádíme také pomocí metod dalších paragrafů. Rozložení proměnné $\boldsymbol{A}$ je $(f, 1-f)$ resp. $(f_a, f_{\bar{a}})$.

## 4.2 Rozložení četností

*Kategorizovanou proměnnou* můžeme statisticky chápat jako soubor jevů, pro který platí:

— každé dva jevy jsou neslučitelné (žádné dva nemohou nastat současně);
— soubor jevů je úplný (alespoň jeden z jevů musí nastat);
— každý z jevů má smysl (každý z jevů může nastat);
— každý z jevů je identifikovatelný (v určitém stupni spolehlivosti);
— při identifikaci určujeme jednoznačně, který z jevů nastal.

Každá kategorie pak odpovídá jednomu z jevů; určení toho z jevů, který u statistické jednotky nastal, je totožné s určením kategorie, do které ji zařadíme. Proto kategorie znaku $\boldsymbol{A} = \{a_1, a_2, \ldots, a_K\}$ považujeme za soubor možných jevů, které lze zjistit.

Statistická analýza vychází ze vztahů všech $K$ četností $\{n_1, n_2, \ldots, n_K\}$ resp. $\{f_1, f_2, \ldots, f_K\}$, a navíc z typu znaku, tj. z relací, které platí mezi $\{a_k\}$ tak, jak byly určeny vnějším sociologickometodologickým kritériem. Jde-li o prostý seznam jevů, hovoříme o nominálním znaku, jsou-li jevy uspořádány, jde o ordinální znak, přiřazujeme-li jevům čísla, dostáváme kardinální kategorizovaný znak. Zvláštní roli hraje znak dichotomický, jehož dvě hodnoty se vzájemně vylučují a k jehož statistickému popisu postačuje údaj o jedné kategorii, tj. $f_1$ nebo $f_2$ (druhý údaj plyne automaticky, $f_2 = 1 - f_1$).

Tabulka četností v třídění 1. stupně zahrnuje $K$ nezávislých parametrů. Buď je to rozložení $\{n_k\}_K$, z něhož plyne výběrový rozsah $n = \Sigma n_k$, nebo $\{n, f_k\}_K$, kde jedna z relativních četností je odvoditelná z ostatních ($\Sigma f_k = 1$). V praxi analýzy je





vhodné využít grafická zobrazení rozložení četností, která mají celou řadu tvarů. Nejvhodnější je *histogram* (*sloupkový graf*) a pro nominální znaky také *kruhový graf*. Existuje celá řada dalších vhodných i méně vhodných ilustrativních metod, které lze vidět v publikacích statistické služby, v odborných článcích a knihách.

Příklad 4.1. Důvody změny zaměstnání. Ve výzkumu „Životní dráhy mládeže" byla položena otázka: „Změnil jste zaměstnání? Jestliže ano, jaký jste k tomu měl důvod?" Při záznamu odpovědí byly kódovány statistické jevy, které odpovídaly předem určeným kategoriím znaku, „důvody změn", a doplňkové jevy: „absence změny", „chybějící informace". Výsledky třídění 1. stupně jsou uvedeny v tab. 4.1a.

Tabulka 4.1. *Změna zaměstnání*

a) *Rozložení četností pro změnu zaměstnání a její důvody* (soubor mládeže ČSSR, 18—29 let)

| Kód | Kategorie | Absolutní četnost | Relativní četnost | Procento |
|---|---|---|---|---|
| 1 | neměnil zaměstnání | 1 628 | 0.8555 | 86 |
| 2 | rodinné důvody | 74 | 0.0389 | 4 |
| 3 | finanční důvody | 57 | 0.0300 | 3 |
| 4 | zlepšení podmínek resp. výhodnější dojíždění | 48 | 0.0252 | 3 |
| 5 | nová práce lépe odpovídá zájmům a schopnostem | 22 | 0.0116 | 1 |
| 6 | lepší možnost růstu a postupu | 8 | 0.0042 | 0 |
| 7 | zdravotní důvody | 17 | 0.0089 | 1 |
| 8 | reorganizace | 6 | 0.0032 | 0 |
| 9 | ostatní | 12 | 0.0063 | 1 |
| 0 | chybí informace | 31 | 0.0163 | 2 |
| | Celkem | 1 903 | 1.0001 | 101 |

(Zdroj: V. Dubský, Životní dráhy mládeže, výzkumný soubor, ÚFS ČSAV, Praha 1978).

b) *Důvody změny zaměstnání* (Výzkum „Životní dráhy mládeže", soubor mládeže 15—29 let, $n = 232$).

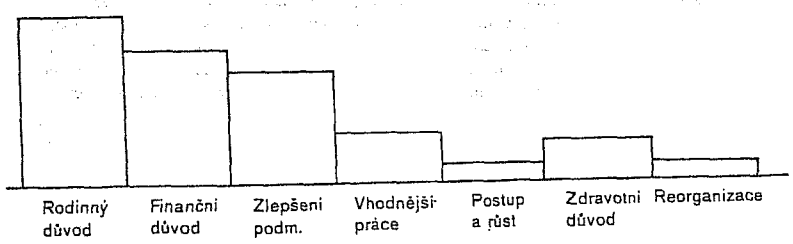
| Důvod | Rodinný důvod | Finanční důvod | Zlepšení podmínek | Vhodnější práce | Postup a růst | Zdravotní důvody | Reorganizace | Celkem |
|---|---|---|---|---|---|---|---|---|
| Procentní zastoupení | 32% | 25% | 21% | 9% | 3% | 7% | 3% | 100% |

Proměnná „důvody změny zaměstnání" obsahuje však jen kategorie, které mají význam za podmínky, že respondent změnil zaměstnání. Proto z analýzy vynecháme kategorii 1 a 0 (= kód pro chybějící informaci). Nakonec vynecháme i málo obsazenou kategorii „ostatní důvody", která nemá interpretační význam (z důvodů obsahové heterogenity i nízkého procenta zastoupení). Po redukci

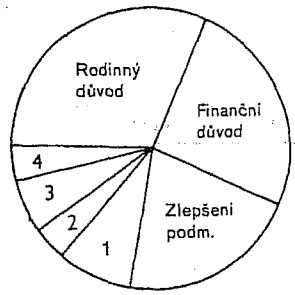
dostaneme tab. 4.1b, která charakterizuje výskytovost důvodů změny na redukovaném souboru, a která je vhodným východiskem pro analýzu dat.

Rozložení z tab. 4.1b můžeme zobrazit histogramem nebo kruhovým grafem (viz. obr. 4.1).

a)




b)




Obr. 4.1. Důvody změny zaměstnání
a) Histogram
b) Kruhový graf: 1 = vhodnější práce; 2 = postup a růst; 3 = zdravotní důvod; 4 = reorganizace

Pro publikaci tabulek rozložení četností platí obvyklé zásady:

1. Každá tabulka je plně informativní a vypovídá sama o sobě. Obsahuje název nebo přesnou charakteristiku proměnné, charakteristiku souboru, místa, času, kontextu, případně i metodu, která je využita, a výsledky statistické analýzy.

2. Řádky a sloupce jsou jasně označeny slovním popisem (především jde o význam kategorií proměnné, význam charakteristik a čísel v tabulce), pouze obecně přijaté a dobře definované statistické symboly mohou být výjimkou.

3. Hlavní informace se umisťuje do záhlaví tabulky, doplňková informace do poznámek k tabulce (nikoliv pod čáru).

4. V poznámkách pod tabulkou (případně v záhlaví) je uveden zdroj dat, pokud nejde o data, která patří autorům, o data určená jinde (např. v rejstříku použitých dat) nebo společná pro celou publikaci.

5. Absolutní četnosti se uvádějí pouze tehdy, mají-li vlastní informativní hodnotu. Relativní čísla jsou většinou vyjádřena v procentech, a to zaokrouhleně



ŘEHÁK, J., ŘEHÁKOVÁ, B. 1986. *Analýza kategorizovaných dat v sociologii*. Academia: Praha.



na celá čísla (méně často na jedno desetinné místo), vždy k nim uvádíme velikost souboru $n$.

6. Zaokrouhlování při dělení $\frac{n_i}{n}$ vede k tomu, že součet procent nemusí být přesně 100, ale může dávat 99, 101, či 100,1, 99,9 apod. Dříve se procenta v jednotlivých kategoriích upravovala tak, aby součet dával 100%, v současné době se od takových úprav upouští.

7. V poznámkách u tabulky (nebo i přímo v tabulce) zpravidla uvádíme procento vynechávaných hodnot.

Obdobná pravidla platí pro přípravu grafů: plná informativnost, vhodné měřítko, které zajišťuje přehlednost, slovní popis, případná slovní informace přímo v grafu nesmí rušit vjem, uvedení zdroje.

V tabulce rozložení četností pro *nominální* znak můžeme kategorie řadit sestupně podle četností jejich obsazení. Tím získáváme větší přehled a rychlejší informaci. Někdy uvádíme jen ty kategorie, které hrají v rozložení výraznou a interpretovatelnou roli. Takovou formu volíme, především jde-li o tzv. dlouhé znaky (velké $K$), a u znaků s předem neomezeným počtem hodnot. Typickými příklady proměnných, které většinou tabelujeme tímto způsobem, jsou: respondentův *nejoblíbenější* zpěvák (sportovec, kniha, film, opera), příčina pracovní neschopnosti, záměr trávení dovolené. Uvedená forma se však nehodí pro ordinální a kardinální znaky, neboť by porušila vztahy mezi kategoriemi.

## 4.3 Kumulativní četnosti (distribuční funkce)

U ordinálních a kardinálních znaků jsou kategorie seřazeny podle vnějšího kritéria určeného obsahem. Z tohoto jednoznačného řazení vychází řada analytických metod založených na kumulativních četnostech, vyjadřujících postupné přibývání výskytů podél stupnice uvažované proměnné. Používáme *absolutní* i *relativní kumulativní četnosti*

$$M_k = \sum_{j=1}^{k} n_j = \text{počet jednotek v kategoriích } 1, 2, \ldots, k,$$

(4.1)
$$F_k = \frac{M_k}{n} = \sum_{j=1}^{k} f_j = \text{podíl jednotek v kateg. } 1, 2, \ldots, k,$$

$$P_k = \sum_{j=1}^{k} p_j = \text{podíl jednotek v kateg. } 1, 2, \ldots, k \text{ v základním souboru.}$$

Pro popis vzorců v dalších částech zavedeme úmluvu

(4.2)
$$M_0 = F_0 = P_0 = 0.$$

Poznamenejme, že $M_K = n$, $F_K = P_K = 1$. Souboru relativních čísel $\{F_k\}_K$ resp. $\{P_k\}_K$ říkáme *distribuční funkce*. Smysl a využití kumulativních četností ilustruje příklad 4.2.

Příklad 4.2. Příchody do zoo.

Při sociologickém šetření struktury návštěvníků Pražské zoo a délky jejich pobytu bylo zjišťováno rozložení příchodů (metodou načítání příchozích u vchodu). Četnosti získané během jednoho výzkumného dne uvádí tab. 4.2.

**Tabulka 4.2.** *Příchody do ZOO Praha v sobotu 12. 8. 1978*
(charakteristika dne: skoro zataženo, chladno)

| | Příchody v hodinách | | | | | | |
|---|---|---|---|---|---|---|---|
| | do 08.00 | do 09.00 | do 10.00 | do 11.00 | do 12.00 | do 13.00 | do 14.00 |
| Počet příchozích | 20 | 219 | 956 | 1 034 | 971 | 547 | 759 |
| Procento | 0.3 | 3.7 | 16.1 | 17.4 | 16.3 | 9.2 | 12.7 |
| Kumulativní četnost | 20 | 239 | 1 195 | 2 229 | 3 200 | 3 747 | 4 506 |
| Kumulativní procento | 0.3 | 4.0 | 20.1 | 37.5 | 53.8 | 63.0 | 75.7 |

| | Příchody v hodinách | | | | | |
|---|---|---|---|---|---|---|
| | do 15.00 | do 16.00 | do 17.00 | do 18.00 | do 19.00 | Celkem |
| Počet příchozích | 924 | 438 | 75 | 11 | 0 | 5 954 |
| Procento | 15.5 | 7.3 | 1.3 | 0.2 | 0.0 | 100.0 |
| Kumulativní četnost | 5 430 | 5 868 | 5 943 | 5 954 | 5 954 | — |
| Kumulativní procento | 91.2 | 98.5 | 99.8 | 100.0 | 100.0 | — |

Absolutní četnosti jsou důležité pro zhodnocení náporu na pokladnu, pro služby uvnitř zahrady, pro požadavky na městskou dopravu. Relativní četnosti dobře ukazují rozložení náporu během dne a umožňují porovnání podobných údajů z jiných dnů. Kumulativní četnosti skýtají okamžitou informaci o tom, kolik lidí již do zoo přišlo, ale také podíl, kolik jich do určité hodiny přišlo, a tudíž jaká část jich ještě přijde. Znak „hodina příchodu" může být chápán nominálně (charakteristika určité části dne), ordinálně (průběžný posun během dne) i kardinálně (kvantifikovat můžeme např. časovým odstupem od otevírací nebo zavírací hodiny, od vrcholného zatížení restaurace apod.). Kumulativní četnosti umožňují také rychlý výpočet četností výskytů v určitém intervalu složeném ze sousedních kategorií: relativní četnost kategorií

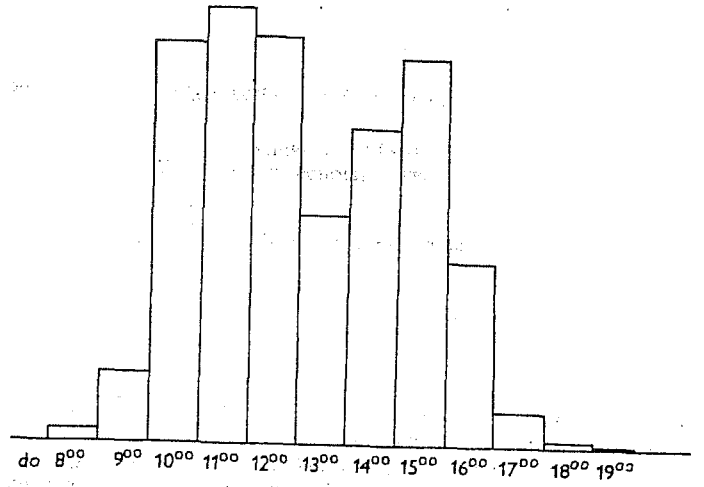
$$(i, i+1, \ldots, j) = f_i + f_{i+1} + \ldots + f_j = F_j - F_{i-1}. \tag{4.3}$$



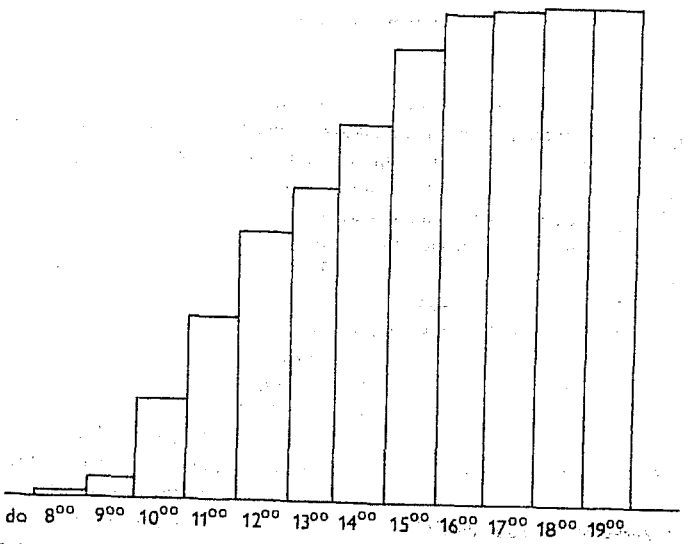


Tak např. mezi 10. a 14. hod. přišlo 75.7% − 20.1% = 55.6% návštěvníků. Četnosti lze graficky zobrazit pomocí obr. 4.2.

a)




b)




*Obr. 4.2.* Přehled příchodů do Zoo Praha, 12. 8. 1978
a) Četnosti příchodů v hodinových intervalech
b) Kumulativní četnosti příchodů

## 4.5 Zvláštní případ tabelací: vícenásobné výběrové otázky

Tabelace vícenásobných výběrových otázek není ve striktním slova smyslu tříděním 1. stupně. Vzhledem k častému výskytu v sociologických šetřeních se o ní však zmíníme. Vícenásobné výběrové otázky jsou instrukce typu: „Z přiloženého seznamu vyberte dvě položky, které považujete za nejdůležitější", „Jmenujte





**Tabulka 4.3.** *Názor na důležitost cílů v zaměstnání*
(Pokyn: „Vyberte dva z předložených cílů, které považujete za nejdůležitější", $n = 1903$)

| Cíl | Počet voleb | Procento z počtu voleb | Procento z počtu respondentů |
|---|---|---|---|
| Řídit lidi, být vedoucím | 164 | 5 | 9 |
| Materiální zajištění | 949 | 26 | 50 |
| Společenská úcta, vážnost, prestiž | 198 | 5 | 10 |
| Možnost přinést maximální užitek lidem, společnosti | 552 | 15 | 29 |
| Tvůrčí činnost, možnost vytvářet nové | 313 | 9 | 16 |
| Každodenní svědomité plnění svých povinností | 427 | 12 | 22 |
| Radost z vykonané práce | 828 | 23 | 44 |
| Možnost rozšiřovat obzor | 200 | 6 | 11 |
| Celkem | 3 631 | 101% | 191% |

(Zdroj: V. Dubský, Životní dráhy mládeže, výzkumný soubor, ÚFS ČSAV, Praha 1978)



nejvýše tři oblíbené zpěváky", „Uveďte tři nejpodstatnější příčiny jevu". Přitom instrukce neobsahuje pokyn k seřazení položek. Analýza těchto dat je složitá, neboť jevy jsou specifickým způsobem závislé. Třídění se provádí tak, že zjišťujeme

(4.8) $m_j$ = počet voleb, které dostala položka „$j$",

a odhad četnosti pro každou z $J$ položek ($J$ je počet buď předložených, nebo jmenovaných možností):

$$r_j = \frac{m_j}{n}. \tag{4.9}$$

Jiným způsobem je tabelace jednorozměrné tabulky relativních četností vzhledem k počtu realizovaných voleb

$$g_j = \frac{m_j}{M}, \quad M = \sum_{j=1}^{J} m_j, \tag{4.10}$$

tj. podílu voleb kategorie „$j$" na všech realizovaných volbách.

Můžeme též uvést index $R$:

$$R = \frac{M}{Ln} = \frac{\sum_{j=1}^{J} r_j}{L}, \tag{4.11}$$

který vyjadřuje, do jaké míry respondenti využili povolených $L$ voleb.

Častou chybou při zpracování odpovědí na vícenásobné výběrové otázky je to, že děláme třídění 1. stupně pomocných a jen formálně zavedených znaků, které vzniknou tak, že např. 3 možné volby kódujeme: 1. znak = první zatržená hodnota v seznamu, 2. znak = druhá zatržená hodnota v seznamu, 3. znak = třetí zatržená hodnota v seznamu. Rozložení těchto pomocných znaků nemá smysl a žádný interpretační význam. Např. lze snadno ověřit, že kód první položky se nemůže vůbec vyskytnout u 2. a 3. znaku, kód druhé položky se nemůže vyskytnout u 3. znaku, kód třetí položky se může vyskytnout u 1. znaku jen tehdy, využil-li respondent pouze jednu volbu. Hodnoty $\{m_j\}$ vznikají součtem rozložení uvedených pomocných znaků.

Příklad 4.3. Cíle v povolání.

Mladým lidem ve věku *18—29 let byl dán* tazatelem pokyn: „Ve svém povolání se lidé snaží dosáhnout nejrůznějších cílů. Vyberte dva z nich, které jsou podle Vašeho názoru nejdůležitější." Osm důvodů bylo předloženo na kartě, uvedené volby byly zakroužkovány v záznamovém listě. Kódování bylo provedeno pomocí dvou pomocných znaků ***A*** = kód první zakroužkované kategorie, ***B*** = kód druhé zakroužkované kategorie. Tabulka 4.3 vznikla jako součet absolutních četností znaku ***A*** a ***B*** (dílčí tabulky nemají smysl).

(Může překvapit nízké procento kategorie, „materiální zajištění", neboť při samostatném dotazu bychom očekávali téměř stoprocentní odpověď „*ano*, je to důležité".) Součet 191% ukazuje, že využití

# Counting Responses

# 3

*How can you summarize the various responses people give to a question?*

- What is a frequency table, and what can you learn from it?
- How can you tell from a frequency table if there have been errors in coding or entering data?
- What are percentages and cumulative percentages?
- What are pie charts and bar charts, and when do you use them?
- When do you use a histogram?
- What are the mode and the median?
- What do percentiles tell you?

Whenever you ask a number of people to answer the same questions, or when you measure the same characteristics for several people or objects, you want to know how frequently the possible responses occur. This can be as simple as just counting up the number of yes or no responses to a question. Or it can be considerably more complicated if, for example, you've asked people to report their annual income to the nearest penny. In this case, simply counting the number of times each unique income occurs may not be a useful summary of the data. In this chapter, you'll use the Frequencies procedure to summarize and display values for a single variable. You'll also learn to select appropriate statistics and charts for different types of data.

▶ The data analyzed in this chapter are in the *gss.sav* data file. For instructions on how to obtain the Frequencies output shown in the chapter, see "How to Obtain a Frequency Table" on p. 48.





## Describing Variables

To see what's actually involved in examining and summarizing data, you'll use the nine variables from the General Social Survey described in Table 3.1. (You will use data from only 1500 respondents, since the SPSS student system is restricted in the number of cases in a data file.)

*What's the General Social Survey?* The General Social Survey is administered yearly by the National Opinion Research Center to a sample of about 1500 persons 18 years of age and older. The sample represents the population of non-institutionalized adults living in the United States. (College dormitories are excluded from the survey!) Questions on many different topics—from how often you pray to where you were living at age 16—are included. Data from the General Social Survey are distributed at a nominal cost and are widely used by researchers and students (Davis & Smith, 1993). ■■■

**Table 3.1 Variables from the General Social Survey**

| Variable Name | Description |
|---|---|
| *age* | Age of respondent in years |
| *sex* | 1=Male, 2=Female |
| *educ* | Years of education |
| *income91* | Total family income in 1993 (classified into one of 21 income categories) |
| *wrkstat* | Work status (1=Full-time work, 2=Part-time work, 3=Temporarily not working, 4=Unemployed (laid off), 5=Retired, 6=In school, 7=Keeping house, 8=Other) |
| *richwork* | "Would you continue or stop working if you became rich?" (1=Continue, 2=Stop) |
| *satjob* | Job satisfaction (1=Very satisfied, 2=Moderately satisfied, 3=A little dissatisfied, 4=Very dissatisfied) |
| *life* | "Do you find life exciting, pretty routine, or dull?" (1=Dull, 2=Routine, 3=Exciting) |
| *impjob* | "How important to your life is having a fulfilling job?" (1=One of the most important, 2=Very important, 3=Somewhat important, 4=Not too important, 5=Not at all important) |

*All of these variables are defined as numeric in SPSS, but in most cases the numbers are just codes for non-numeric information. Value labels for each variable specify what the codes really mean.*

*In the SPSS Data Editor, to display (or hide) value labels, from the menus choose:*

*View*
*Value Labels*

Start by looking at the variable *impjob*, which tells you how important a fulfilling job is to the respondent. Since there are only five possible responses, you can easily count how many people gave each of them.

## A Simple Frequency Table

In Figure 3.1, you see the frequency table for the job importance variable.

**Figure 3.1 Frequency table of job importance**

*To obtain this frequency table, from the menus choose:*

*Statistics*
*Summarize*
*Frequencies...*

*In the Frequencies dialog box, select the variables impjob, as shown in Figure 3.11.*

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | One of most important | 316 | 21.1 | 21.4 | 21.4 |
| | Very important | 833 | 55.5 | 56.3 | 77.7 |
| | Somewhat important | 238 | 15.9 | 16.1 | 93.8 |
| | Not too important | 62 | 4.1 | 4.2 | 98.0 |
| | Not at all important | 30 | 2.0 | 2.0 | 100.0 |
| | Total | 1479 | 98.6 | 100.0 | |
| Missing | Don't know | 7 | .5 | | |
| | No answer | 14 | .9 | | |
| | Total | 21 | 1.4 | | |
| Total | | 1500 | 100.0 | | |

The response "very important" was chosen by 833 people. This response is coded in the data file as the number 2.

From a frequency table, you can tell how frequently people gave each response. The first row is for the response *one of the most important* (coded in the data with the value 1). The second row is for the response *very important* (coded in the data with the number 2). To determine how many people gave each response, look at the column labeled *Frequency*. For example, you find that 316 people find a fulfilling job to be *one of the most important* things to them, and 238 find it to be *somewhat important*. Only 30 people find having a fulfilling job *not at all important*. In the row labeled *Total*, you see that 1479 people selected one of the five possible valid responses.

The second part of the table tells you how many people did not select one of the five choices. There are two rows in the frequency table for the responses *don't know* and *no answer*. *Don't know* is used for people unwilling to commit themselves to a response. *No answer* is used when the response is illegible, lost, or not recorded by the interviewer. When the data file was defined, both *don't know* and *no answer* were identified as missing-value codes. That is, you don't have a valid answer for people whose responses are coded as *don't know* or *no answer*. In the *Frequency* column, you see that the response *don't know* was selected by 7 people and that the response was not available for 14 people. A total of 21 failed to select a valid response; that is, their response was identified as missing.

In the last row of the frequency table, you see that a total of 1500 people participated in the survey. Of these, 21 failed to select one of the five available responses; that is, their response was identified as *missing*. The other 1479 provided a valid response.

*Why do you use different codes for* don't know *and* no answer? It's important to pinpoint why data values are missing. A response of *don't know* tells you that a person probably doesn't have strong feelings about the topic. It's unlikely that they find a job to be very important. A response of *no answer* doesn't tell you anything about a person's opinion of the importance of a job. The number of *no answer* responses tells you whether the survey was carefully conducted. You'll see later that if there are many cases with missing values, you may have serious problems in drawing conclusions from your data. ■■■

In the frequency table, value labels, which are descriptions of the codes assigned when you define a variable, are used to identify rows. If you don't assign these descriptions, the actual codes are shown. If your codes are not inherently meaningful, you should assign value labels to them so that the output is easier to understand. Assigning a value label once is much easier than repeatedly having to look up the meanings of codes.

Only responses actually selected by the participants are included in the frequency table. If no one selected the response *not at all important*, it would not be included in the table. Similarly, if you accidentally enter a code that does not correspond to a valid response—say a code of 0, 6, or 7 for the job importance variable—you will find it as a row in the frequency table. That's why frequency tables are useful for detecting mistakes in the data file. If you find wrong codes in your data values, you must correct the data file before proceeding.

## Percentages

A frequency count alone is not a very good summary of the data. For example, if you want to compare your results to those of another survey, it won't do you much good to know simply that 762 people in that survey chose the response *very important*. From the count alone, you can't tell if the other survey's results are similar to yours. To compare the two surveys, you must convert the observed counts to percentages.

From a percentage, you can tell what proportion of people in the survey gave each of the responses. Unlike counts, you can compare percentages across surveys with different numbers of cases. You compute a percentage by dividing the number of cases that gave a particular response by the total number of cases. Then you multiply the result by 100.

*To change the number of decimal places shown in the output, double-click the pivot table to activate it, select the cell or column of interest, then choose:*

*Format*
*Cell Properties...*

*and change the Decimals specification.*

In Figure 3.1, you find percentages in the column labeled *Percent*. Note that the 316 people who gave the response *one of the most important* are 21.1% of the 1500 people in your survey. Similarly, the 238 people who gave the response *somewhat important* are 15.9% of your sample. The 7 people who *don't know* are 0.5% of the total sample. (The actual percentage is 0.47%, but by default only one decimal place is shown.) The sum of the percentages over all the possible responses, including *don't know* and *no answer*, is 100%.

## Percentages Based on Valid Responses

To get the numbers in the column labeled *Percent*, you divide the observed frequency by the total number of cases in the sample and multiply by 100. Cases with codes identified as *missing* are included in the denominator. That can be a problem. For example, the General Social Survey does not ask all questions of all people. The question "Would you continue or stop working if you became rich?" was asked of only two-thirds



of people who were working or temporarily unemployed. Figure 3.2 shows the responses of people to this question.

*To obtain this table, in the Frequencies dialog box select the variable richwork, as shown in Figure 3.11.*

**Figure 3.2 Frequency table of continue working**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Continue working | 448 | 29.9 | 69.8 | 69.8 |
| | Stop working | 194 | 12.9 | 30.2 | 100.0 |
| | Total | 642 | 42.8 | 100.0 | |
| Missing | Not applicable | 842 | 56.1 | | |
| | Don't know | 11 | .7 | | |
| | No answer | 5 | .3 | | |
| | Total | 858 | 57.2 | | |
| Total | | 1500 | 100.0 | | |

The percentage of people giving the response *continue working* is 29.9. What does that mean? Does it mean that about 30% of people in the survey would continue working if they became rich? No. It means that about 30% of the people in the sample, regardless of whether they were asked the question or volunteered an answer, gave the response *continue working*. Of the 1500 people in the survey, 56.1% weren't even asked the question (recorded in the table as *Not applicable*). An additional 1% were asked and either gave the response *don't know* or their response was lost (*no answer*). All of these missing people are included in the denominator of the *Percent* calculation.

If you want to know what percentage of people who gave an acceptable answer selected *continue working*, look at the *Valid Percent* column. Almost 70% of people who answered the question claim that they would continue working if they struck it rich. (It's up to you whether you believe that percentage!) That's quite different from 30%. To calculate the entries in the *Valid Percent* column, you must exclude all people who gave an answer identified as *missing*. Valid percentages sum to 100 over all possible answers that are not missing. In this example, there are only two valid answers: *continue working* and *stop working*. Of the people who gave one of these answers, 69.8% selected the first and 30.2% selected the second. These two percentages sum to 100.

### Problems with Missing Data

Removing people who aren't asked a question from the calculation of percentages is not troublesome. They don't make interpretation of the results difficult. However, if a lot of people who are asked the question refuse to answer, that can be a problem. In Figure 3.2, you see that only 11 people gave an answer of *don't know*. They represent fewer than 2% of the 653 people who were actually asked the question. So, you don't have to worry much about their impact on any conclusions you draw.

In contrast, however, consider the following situation. You conduct an employee satisfaction survey among 100 employees and find that 55 of them rate themselves as satisfied, 4 rate themselves as unsatisfied, and the remaining 41 decline to answer your question. That means that 55% of the polled employees consider themselves satisfied. However, if you exclude those who refused to answer from the denominator, 93% of the employees who answered the question consider themselves satisfied.

Which is the correct conclusion? Unfortunately, you don't know. It's possible that you have a company full of satisfied employees, many of whom don't like to answer questions. It's also possible that almost half of your employees are unhappy but are wary of voicing their dissatisfaction. When your data have many missing values because of people refusing to answer questions, it may be difficult, if not impossible, to draw correct conclusions. When you report percentages based on cases with nonmissing values, you should also report the percentage of cases that refused to give an answer.

### Cumulative Percentages

There's one more percentage of interest in the frequency table. It's called the cumulative percentage. For each row of the frequency table, the **cumulative percentage** tells you the percentage of people who gave that response and any response that precedes it in the frequency table. It is the sum of the valid percentages for that row and all rows before it. Since there are only two possible valid answers for the continue working variable, the cumulative percentages in Figure 3.2 are of little interest. Instead, consider Figure 3.1 again. The cumulative percentage for *somewhat important* is 93.8. This means that over 93% of the people who answered the question said that a fulfilling job was at least somewhat important to their lives. Only 6.2% of the people rated the importance of a fulfilling job as less than *somewhat important*. Cumulative percentages are most useful when there is an underlying order to the codes assigned to a variable.



## Sorting Frequency Tables

Unless you specify otherwise, SPSS produces a frequency table in which the order of the rows corresponds to the values of the codes you assign to the responses. The first row is for the smallest number found in the data values, and the last is for the largest. Codes that have been declared missing are at the end of the table. For example, if you had assigned the code 1 to *stop working*, it would have appeared first in the frequency table in Figure 3.2.

When you have several possible responses and the codes are not arranged in a meaningful order, you may want to rearrange the frequency table so that it's easier to use. You can determine the order of the rows in the table based on the frequency of values in the data. For example, Figure 3.3 shows a frequency table for the work status variable when the table is sorted in descending order of frequencies. Look at the column labeled *Frequency*. The frequencies go from largest to smallest.

**Figure 3.3 Frequency table sorted by counts**

*To obtain this output, select Format in the Frequencies dialog box. Then select Descending counts, as shown in Figure 3.12.*

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Working fulltime | 747 | 49.8 | 49.8 | 49.8 |
| | Retired | 231 | 15.4 | 15.4 | 65.2 |
| | Keeping house | 200 | 13.3 | 13.3 | 78.5 |
| | Working parttime | 161 | 10.7 | 10.7 | 89.3 |
| | Unempl, laid off | 51 | 3.4 | 3.4 | 92.7 |
| | School | 42 | 2.8 | 2.8 | 95.5 |
| | Other | 36 | 2.4 | 2.4 | 97.9 |
| | Temp not working | 32 | 2.1 | 2.1 | 100.0 |
| | Total | 1500 | 100.0 | 100.0 | |

Table is sorted by the counts in the Frequency column.

Sorting a frequency table will usually change the values in the *Cumulative Percent* column, since the cumulative percentages depend on the order of the rows in the table. When the work status table is sorted by decreasing frequency, the cumulative percentage for *retired* is the percentage of people retired or working full time. In the default frequency table, however, in which the rows are sorted by the values of the codes,

the cumulative percentage for *retired* is the sum of the valid percentages for codes 1 through 5.

## Pie Charts

The information in a frequency table is easier to see if you turn it into a visual display, such as a bar chart or a pie chart. In Figure 3.4, you see a pie chart of the frequency table in Figure 3.3. There is a "slice" for each row of the frequency table. From the pie chart, you can easily see that almost half of your sample is *working full time*. It's also easy to see that the number of people who are *retired, keeping house,* and *working part time* are roughly equal. If you have many small slices in a pie chart, you can combine them into an *other* category. For example, Figure 3.5 is the pie chart for the same frequency table, except that all slices that have fewer than 5% of the cases (*in school, temporarily not working, unemployed,* and *other*) are combined into a single slice.

**Figure 3.4 Pie chart of work status**

*To obtain a pie chart, select Pie charts in the Frequencies Charts dialog box, as shown in Figure 3.14.*

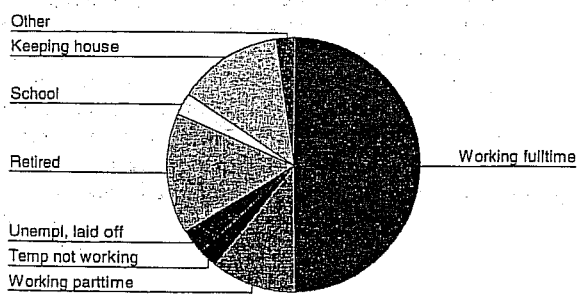
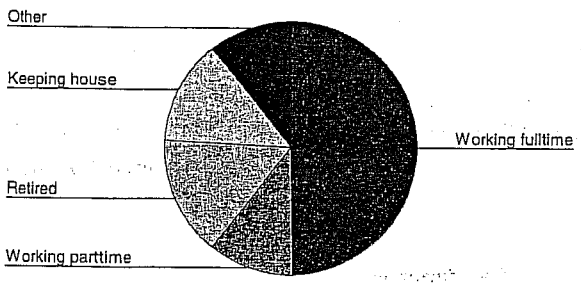



**Figure 3.5 Work status with categories collapsed**

*You can collapse categories in a pie chart after it has been created. See "Modifying Chart Options" on p. 518 in Appendix A.*








would expect, the tallest bar is for the *working full time* category. It's about three times as tall as the next largest bar, which represents *retired*.

**Figure 3.6 Bar chart of work status**

*To obtain this output, select Bar charts in the Frequencies Charts dialog box, as shown in Figure 3.14.*

*You can also obtain bar charts using the Graphs menu, as discussed in Appendix A.*