**Základní zpracování kategoriálních dat, třídění a popisné statistiky, frekvenční analýza, nezávislost**

**Zpracování kategoriálních dat:**

### **Typy kategoriálních proměnných**

Kategoriální proměnné lze dále rozdělit na:

* **Nominální (jmenné)** – nemají přirozené pořadí (např. barva očí, náboženské vyznání).
* **Ordinální (pořadové)** – lze je uspořádat podle velikosti nebo intenzity, ale rozdíly mezi kategoriemi nelze kvantifikovat (např. spokojenost: nízká, střední, vysoká).

Pro ordinální proměnné lze v některých případech využít specifické testy, které zohledňují jejich uspořádání (např. **Spearmanův korelační koeficient** nebo **testy na trend**).

Kategoriální data, též označována jako znaky kvalitativní vyjadřujeme slovně – jedná se např. o druh vlastnictví bytu, místo trvalého bydliště, typ vystudované střední školy,…).

**Kategoriální** znaky můžeme **rozdělit pomocí četností.**

Na rozdíl od znaků kvalitativních se ale u kategoriálních znaků neprovádí uspořádání údajů do rostoucí posloupnosti, ale jedná se pouze o výčet vyskytujících se variant znaku.

Vše se uspořádá do tabulky a vznikne tabulka rozdělení četností, která zahrnuje:

a) **absolutní četnosti** (kolikrát se znak vyskytl)

b) **relativní četnosti (**poměr k celkovému počtu pozorování)

c) **kumulativní absolutní četnosti**

**d) kumulativní relativní četnosti**

**Sestavení dat do tabulky četností je vhodné zejména pro malé soubory.**

Konstruujeme **intervalové rozdělení četností**, kdy rozdělíme data do stejně širokých intervalů a následně zjišťujeme počet jednotek v každém úseku.

**Třídění a popisné statistiky:**

Z popisné statistiky se nejčastěji využívá pojem **modus a míra variability**.

* **Modus** představuje nejčetnější hodnotu znaku.
* **Míra variability** se u kvantitativních znaků využívá zejména při posuzování vypovídací schopnosti aritmetického průměru.

Vypovídací schopnost aritmetického průměru je pak tím větší, čím je variabilita sledovaného znaku menší a naopak.

**Frekvenční analýza:**

Frekvenční (četnostní) analýza představuje základní krok při zpracování kategoriálních dat. Slouží ke kvantitativnímu vyjádření výskytu jednotlivých kategorií a jejich vztahů. Při analýze více proměnných se využívají tzv. **vícerozměrné kontingenční tabulky**, které umožňují studium interakcí mezi více kategoriálními znaky najednou.

Kromě četností lze sledovat i tzv. **podmíněné relativní četnosti**, tedy pravděpodobnosti výskytu jedné hodnoty proměnné za podmínky, že druhá proměnná má určitou hodnotu. To je užitečné např. při zkoumání, jak se změnila preference zákazníků v závislosti na věkové kategorii.

**Kontingenční tabulka:**

Díky kontingenční analýze lze analyzovat kategoriální data a porovnávat je.

Nejrozšířenější formou je tabulka 2x2, která obsahuje dvě proměnné (zkoumáme např. závislost mezi pohlavím a typem preferovaného dopravního prostředku). Existují ale také i tabulky 3xk nebo obecně r x s (řádky x sloupce).

V tabulce je uveden počet výskytů (četností) jednotlivých kombinací a suma sloupců a řádků.

**Test nezávislosti (Pearsonův chí-kvadrát test):**

Jedná se o metodu, která slouží k porovnávání dvou kategorických proměnných.

Jedná se o test, kde se stanovuje nulová a alternativní hypotéza.

**Nulová hypotéza:** mezi kategorickými proměnnými není významný vztah

**Alternativní hypotéza:** mezi kategorickými proměnnými existuje významný vztah

Při testu nezávislosti se srovnávají pozorované (empirické) a očekávané (teoretické) četnosti pomocí kontingenční tabulky.

Pokud budou naše pozorované hodnoty hodně odlišné od těch očekávaných, pak nulovou hypotézu zamítáme.

Tím prosadíme, že spolu proměnné souvisí – že mají mezi sebou významný vztah.

Pro použití testu jsou dvě podmínky, a to:

1. Pozorování v tabulce jsou nezávislá, každý prvek je tak zahrnutý v tabulce pouze jednou.

2. Buňky by měli mít očekávanou četnost větší než 5.

Pokud druhou podmínku test nesplňuje, jedná se většinou o kategorie jako ,,velmi spokojen“, ,,spíše spokojen“, které pak mohou být sloučeny do kategorie ,,spokojen“.

Test nezávislosti je citlivý na velikost výběrového souboru.

Pokud vybereme příliš mnoho pozorování, pak i malé rozdíly mezi pozorovanými a očekávanými četnostmi pak mohou být statisticky významné.

**Pearsonova rezidua:**

Hodnoty, které jsou blízké k nule představují dobrou shodu mezi pozorovanými a očekávanými četnostmi.

* Rezidua, která je větší jak + nebo – 2 představuje velký rozdíl.

**Test homogenity multinomických rozdělení:**

Pomocí tohoto testu testujeme shodnost jednoho ze sledovaných znaků za různých podmínek, které vyjadřují kategorie druhého znaku.

Nezjišťujeme tedy pouze vztah mezi dvěma proměnnými, ale porovnáváme rozložení jedné proměnné ve více skupinách.

I test homogenity má podmínky stejné jako test nezávislosti – dostatečně velké očekávané četnosti.

Můžeme například zkoumat, jestli je věková struktura hospitalizovaných pacientů ve dvou nemocnicích stejná.

**Nulová hypotéza**: pravděpodobnosti qi1,…..qic nezávisí na řádkovém indexu i (všechny řádky matice jsou stejné)

**Alternativní hypotéza:** pravděpodobnosti qi1,……qic závisí na řádkovém indexu i (všechny řádky matice nejsou stejné)

Postupujeme téměř totožně jako u nezávislosti – srovnáváme pozorované hodnoty s hodnotami očekávanými pomocí kontingenční tabulky

**Vizualizace kategoriálních dat:**

Kvantitativní data může graficky znázornit polygonem četností.

Sestavení: V pravoúhlém souřadném systému spojíme úsečkami body o souřadnicích (xi, ni).

Pro grafické znázornění intervalových četností můžeme využít histogram, což je sloupcový graf, který je tvořen obdélníky podle zvolených intervalů.

Pro kvantitativní znaky můžeme také využít bodové grafy, kdy vodorovná osa označuje hodnoty znaku a svislá osa pak hodnoty druhého znaku.

Oblíbený je také výsečový graf, kdy jsou relativní hodnoty znaku uspořádány do výsečí kruhu.

Lze také využít i mozaikový graf, kde se do obdélníkových oblastí znázorňuje počet a podíl výskytů kombinací kategorií. Velikost ploch obdélníků odpovídá velikosti četnosti.