Úvod

(Proč se zabývat statistikou?)

Statistika je metoda analýzy dat, která nachází široké uplatnění v celé řadě ekonomických, technických, přírodovědných a humanitních disciplín. Její význam v poslední době neustále roste, což úzce souvisí s rozvojem výpočetní techniky, která je používána jak při sběru a přenosu dat, tak při jejich zpracování a ukládání informací.

Role statistiky je zcela nezastupitelná, neboť moderní řízení je založeno na nepřetržitém vyhodnocování informací o celku i jeho subsystémech, a tyto informace poskytuje a následně zpracovává právě statistika.

Přiměřená znalost základních statistických pojmů je důležitá také proto, že pomáhá porozumět odborné literatuře, jejíž některé části statistiku v hojné míře využívají.

Aplikovat statistiku znamená shromažďovat data o studovaných jevech a zpracovávat je, tj. třídit, numericky vyhodnocovat a interpretovat. Statistika se tak ocitá v těsném sousedství informatiky a výpočetní techniky a je připravena řešit problémy pomocí kvantitativní analýzy dat.

1. Základní, výběrový a datový soubor

(Jaké jsou základní pojmy popisné statistiky?)

Po prostudování této kapitoly budete umět:

- vymezit základní soubor a jeho objekty
- stanovit výběrový soubor
- spočítat absolutní a relativní četnosti množin ve výběrovém souboru a znát vlastnosti relativní četnosti a podmíněné relativní četnosti
- ověřit četnostní nezávislost dvou množin ve výběrovém souboru
- vytvořit datový soubor
- uspořádat jednorozměrný datový soubor a stanovit vektor variant
- vypočítat absolutní a relativní četnost jevu ve výběrovém souboru

Pro zvládnutí této kapitoly budete potřebovat 4 - 5 hodin studia.

Nejprve se seznámíme s definicí základního a výběrového souboru a pojmem absolutní a relativní četnosti množiny v daném výběrovém souboru. Uvedeme příklad, s jehož různými variantami se budeme setkávat ve všech kapitolách věnovaných popisné statistice. Rovněž shrneme vlastnosti relativní četnosti.

1.1. Definice

Základním souborem rozumíme libovolnou neprázdnou množinu Ε. Její prvky značíme ε a nazýváme je *objekty*. Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, ..., \varepsilon_n\}$ základního souboru Ε nazýváme *výběrový soubor rozsahu n*. Je-li $G \subset \Epsilon$, pak symbolem N(G) rozumíme *absolutní četnost* množiny G ve výběrovém souboru, tj. počet těch objektů množiny G, které patří do výběrového souboru. *Relativní četnost* množiny G ve výběrovém souboru zavedeme vztahem $p(G) = \frac{N(G)}{n}$.

1.2. Příklad

Základním souborem Ε je množina všech ekonomicky zaměřených studentů 1. ročníku českých vysokých škol. Množina G_1 je tvořena těmi studenty, kteří uspěli v prvním zkušebním termínu z matematiky a množina G_2 obsahuje ty studenty, kteří uspěli v prvním zkušebním termínu z angličtiny. Ze základního souboru bylo náhodně vybráno 20 studentů, kteří tvoří výběrový soubor $\{\varepsilon_1, ..., \varepsilon_{20}\}$. Z těchto 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech. Zapište absolutní a relativní četnosti úspěšných matematiků, angličtinářů a oboustranně úspěšných studentů.

Řešení:

$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, p(G_1) = \frac{12}{20} = 0{,}6, p(G_2) = \frac{15}{20} = 0{,}75,$

$p(G_1 \cap G_2) = \frac{11}{20} = 0{,}55$

Vidíme, že úspěšných matematiků je 60%, angličtinářů 75% a oboustranně úspěšných studentů jen 55%.

1.4. Definice

Nechť je dán výběrový soubor $\{\varepsilon_1, ..., \varepsilon_n\} \subset \Ε$. Hodnoty znaků X, Y, ..., Z pro i-tý objekt označíme $x_i = X(\varepsilon_i)$, $y_i = Y(\varepsilon_i)$, ..., $z_i = Z(\varepsilon_i)$, i = 1, ..., n. Matice

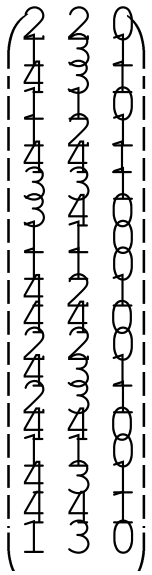
$$\begin{pmatrix} x_1 & y_1 & \cdots & z_1 \\ x_2 & y_2 & \cdots & z_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_m & y_n & \cdots & z_n \end{pmatrix}$$

typu n × p se nazývá *datový soubor*. Její řádky odpovídají jednotlivým objektům, sloupce znakům.

Libovolný sloupec této matice nazýváme *jednorozměrným datovým souborem*. Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle velikosti, dostaneme *uspořádaný datový soubor* $\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$, kde $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$. Vektor $\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix}$, kde $x_{[1]} < ... < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá *vektor variant*.

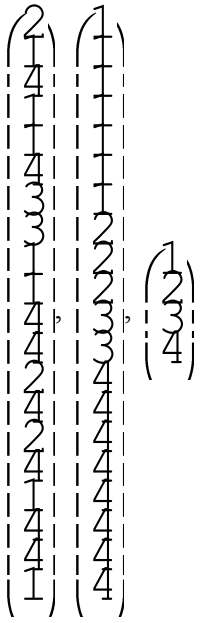
1.5. Příklad

Pro studenty z výběrového souboru u vedeného v příkladu 1.2 byly zjišťovány hodnoty znaků X – známka z matematiky v prvním zkušebním termínu, Y – známka z angličtiny v prvním zkušebním termínu, Z – pohlaví studenta (0 ... žena, 1 ... muž). Byl získán datový soubor



Utvořte jednorozměrný uspořádaný i neuspořádaný datový soubor pro známky z matematiky a vektory variant pro známky z matematiky.

Řešení:



Definice

Podle stupně kvantifikace znaky třídíme takto:

a) (n) *Nominální znaky* připouštějí obsahovou interpretaci jedině relace rovnosti $x_1 = x_2$ (popřípadě $x_1 \neq x_2$), tj. hodnoty znaku představují jen číselné kódy kvalitativních pojmenování. Např. městské tramvaje jsou očíslovány, ale např. č. 4 a 12 říkají jen to, že jde o různé tratě: nic jiného se z nich o vztahu obou tratí nedá vyčíst.
b) *Ordinální znaky* připouštějí obsahovou interpretaci kromě relace rovnosti i v případě relace uspořádání $x_1 < x_2$ (popřípadě $x_1 > x_2$), tj. jejich uspořádání vyjadřuje větší nebo menší intenzitu zkoumané vlastnosti. Např. školní klasifikace vyjadřuje menší nebo

větší znalosti zkoušených (jedničkář je lepší než dvojkař), ale intervaly mezi známkami nemají obsahové interpretace (netvrdíme, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem. Podobný charakter mají různá bodování ve sportovních, uměleckých a jiných soutěžích.

c) *Intervalové znaky* připouštějí obsahovou interpretaci kromě relace rovnosti a uspořádání též u operace rozdílu $x_1 - x_2$ (popřípadě součtu $x_1 + x_2$), tj. stejný interval mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný rozdíl v extenzitě zkoumané vlastnosti. Např. teplota měřená ve stupních Celsia představuje intervalový znak. Naměříme-li ve čtyřech dnech polední teploty 0, 2, 4, 6, znamená to, že každým dnem stoupla teplota o 2 stupně Celsia. Bylo by však chybou interpretovat tyto údaje tvrzením, že ze druhého na třetí den vzrostla teplota dvakrát, kdežto ze třetího na čtvrtý pouze jedenapůlkrát.

d) *Poměrové znaky* umožňují obsahovou interpretaci kromě relace rovnosti a uspořádání a operace rozdílu ještě u operace podílu x_1/x_2 (popřípadě součinu $x_1 \times x_2$*),* tj. stejný poměr mezi jednou dvojicí hodnot a druhou dvojicí hodnot znamená i stejný podíl v extenzitě zkoumané vlastnosti. Např. má-li jedna osoba hmotnost 150 kg a druhá 75 kg, má smysl prohlásit, že první je dvakrát hmotnější než druhá.

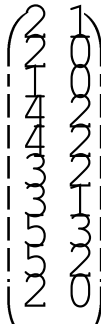
Zvláštní postavení mají:

e) *Alternativní znaky-dichotomické,* které nabývají jen dvou hodnot, např. 0,1, což znamená absenci a prezenci nějakého jevu. Například 0 bude znamenat neúspěch, l úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.

Proměnné závislé a nezávislé

Kontrolní otázky a úkoly

1. Uveďte příklad základního souboru z praxe pedagogů
2. Je dán dvourozměrný datový soubor



Znak X znamená počet členů domácnosti a znak Y počet dětí do 15 let v této domácnosti.

a) Utvořte uspořádané datové soubory pro znaky X a Y.
b) Najděte vektory variant znaků X a Y.
c) Vypočtěte relativní četnost tříčlenných domácností.
d) Vypočtěte relativní četnost nejvýše tříčlenných domácností.
e) Vypočtěte relativní četnost bezdětných domácností.
f) Vypočtěte relativní četnost dvoučlenných bezdětných domácností.
g) Vypočtěte podmíněnou relativní četnost dvoučlenných bezdětných domácností.