SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

2/2024
ročník/volume 34

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 1

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 3 – 25

Dátum vydania/Publication date: 15. apríl 2024/April 15, 2024

Roman PAVELKA
Štatistický úrad Slovenskej republiky

STATISTICKÁ ANALÝZA CHYBĚJÍCÍCH DAT

STATISTICAL ANALYSIS OF MISSING DATA

ABSTRAKT
Standardní statistické metody byly vyvinuty pro analýzu souborů dat v maticovém uspořádání. Řádky datové matice tradičně reprezentují jednotky, označované také jako případy, pozorování nebo předměty – v závislosti na kontextu. Měřené, resp. zjišťované proměnné nebo také charakteristiky pro každou sledovanou jednotku představují sloupce matice dat. Údaje v datech matice jsou téměř vždy reálná čísla u spojitých proměnných, jako je věk anebo příjem, tržby, nebo představují kategoriální odpovědi, které mohou být uspořádané (např. velikostní kategorie, úroveň vzdělání) nebo neuspořádané (nominální), jako je například odvětví ekonomické činnosti, pohlaví, rasa apod. V praxi výběrových zjišťování se však často objevují datové matice pozorovaných hodnot, ve kterých hodnoty některých charakteristik nejsou zaznamenány a jsou chybějící. Například chybějící hodnoty tržeb, obratu anebo jiných ekonomických ukazatelů v podnikových zjišťováních nebo odmítnutí poskytnutí hodnoty příjmu u respondentů v šetření domácností. Příspěvek se zabývá statistickou analýzou takových datových matic, ve kterých hodnoty jedné nebo více proměnných nejsou kompletně vyplněny.

ABSTRACT
Standard statistical methods have been developed for the analysis of data sets in a matrix arrangement. The rows of a data matrix traditionally represent units, also referred to as cases, observations, or subjects - depending on the context. The measured or the surveyed variables or characteristics for each monitored unit represent the columns of the data matrix. Data in matrix data are almost always real numbers for continuous variables such as age or income, turnover, or represent categorical responses that can be ordered (e.g. size category, level of education) or unordered (nominal) such as a sector of economic activity, gender, race, etc. In the practice of sample surveys, however, data matrices of the observed values often appear in which the values of some characteristics are not recorded and are missing. For example, the turnover´s missing values, turnover and/or other economic indicators in business surveys or refusal to provide income values for respondents in household surveys. The paper deals with the statistical analysis of such data matrices in which the values of one or more variables are not completed in full.

KLÍČOVÁ SLOVA
mechanismus chybění, neúplná data, statistická inference, statistické modelování

KEY WORDS
missing data mechanism, incomplete data, statistical inference, statistical modelling

1. ÚVOD DO PROBLEMATIKY CHYBĚJÍCÍCH ÚDAJŮ

Problém neúplných dat se v praxi vyskytuje velmi často. Například míra návratnosti vyplněných statistických formulářů v podnikových šetřeních se pohybuje v závislosti na typu a periodicitě dotazování od 50 % do 70 %, přičemž odpovědi vybraných

statistických jednotek mnohdy nebývají kompletní a také ani bezchybné. Vyšší mírou návratnosti se vyznačují sociální statistiky, například šetření domácností, kde vybrané domácnosti jsou dotazovány pomocí speciálně školených dotazovatelů. Významným faktorem při zjišťování potřebných údajů je neochota respondentů, kteří často odpověď neznají, odpovědi si nepamatují anebo odpovědět jednoduše nechtějí. Následná analýza takto pořízených dat často vychází z předpokladu, že proces, který zapříčinil chybějící data, lze při statistické inferenci více či méně ignorovat. Zde je však potřebné si odpovědět na otázku: jedná se o správné statistické postupy?

Za chybějící údaje jsou považovány takové údaje, které v datové matici chybí u některých (nikoliv u všech) proměnných a u některých (nikoliv u všech) respondentů (případů, resp. zpravodajských jednotek). Tyto chybějící údaje v datové matici jsou problematické především z toho důvodu, že většina současných analytických programů a nástrojů předpokládá datovou matici úplnou. Pokud tomu tak není, dochází při analýzách nejčastěji k ignorování případů, jimž některé sledované hodnoty chybí. Proto může být tento postup zpracování dat nevhodný a výsledné statistické úsudky mohou být více či méně chybné.

Jedním z problémů neúplných dat je skutečnost, že mnohdy není znám důvod chybějících pozorování. Není jasné, zda mechanismus vzniku chybějících dat[1] je zcela náhodný, nebo zda tento mechanismus závisí na pozorovaných či dokonce na nepozorovaných hodnotách [11]. Cílem tohoto příspěvku je provést statistickou analýzu chybějících dat, a to zejména s důrazem na popis a modelování mechanismu chybějících dat. Po nezbytném teoretickém úvodu bude na několika příkladech umělých dat realizováno modelování tohoto mechanismu a načrtnuty metody správného postupu při statistických analýzách neúplných dat.

2. DEFINICE MECHANISMU CHYBĚNÍ DAT A PŘIDRUŽENÝCH VELIČIN

Mechanismus chybějících dat je obvykle definován jako statistický vztah mezi subjekty, proměnnými a pravděpodobností chybějících dat [3]. Tento model umožňuje vypočítat pravděpodobnost chybějícího údaje pro každý zpravodajský subjekt a analyzovanou proměnnou. Příkladem takového modelu pro chybějící data je pravidlo, že „*každá zpravodajská jednotka s pravděpodobností 20 % bude mít ve sledované hodnotě proměnné* $\boldsymbol{Y}$ *údaj chybějící*“. Statistické vyjádření modelu spočívá ve stanovení pravděpodobnosti pro chybějící údaje $P(\boldsymbol{M} = 1) = 0{,}2$, kde $\boldsymbol{M}$ je náhodná indikátorová proměnná $\boldsymbol{M} = 1$ indikující chybějící hodnotu v proměnné $\boldsymbol{Y}$ a $\boldsymbol{M} = 0$ indikující pozorovanou hodnotu. Statistický model o chybějících datech předpokládá, že pravděpodobnost, že chybí údaj u jednoho subjektu (pozorování), je nezávislá na pravděpodobnosti, že bude chybět údaj v proměnné u subjektu (pozorování) jiného. Předpoklad nezávislosti je společný předpoklad při generování chybějících dat [5]. Zjištěný výběrový soubor může obsahovat i více mechanismů generování chybějících dat, a to jak pro každou proměnnou, tak i pro více proměnných.

Stejně jako jiné druhy statistických modelů má mechanismus chybějících dat jeden nebo více parametrů. Tyto parametry jsou spojeny s rozdělením indikátoru chybějících dat $\boldsymbol{M}$. Pro úspěšné modelování chybějících dat je potřebné specifikovat tyto parametry spojené s modelem neúplných dat. Jelikož se tyto parametry modelu neúplných dat týkají populace, parametry je možné pouze odhadovat. Ačkoli se střední

[1] Mechanismus chybějících dat jako proces způsobující chybění v datech byl pro další analýzy dat poprvé formálně definován v práci [11].

hodnota odhadovaných hodnot parametru během opakovaných výběrů z populace rovná skutečné hodnotě parametru, v konkrétních výběrových souborech se odhadovaný parametr obvykle liší od skutečné hodnoty parametru. Znalost pravidla statistického modelu pro chybějící data usnadňuje datovým analytikům zjistit mnoho vlastností vytvářených chybějících dat.

V reálném výzkumu existuje množství důvodů, proč mohou údaje chybět (zabývají se jimi například v práci [7]). Tyto důvody zároveň ovlivňují rozložení chybějících dat v datové matici. Aby bylo možné ze získaných dat vyvodit správné statistické závěry, i za podmínek neodpovědí, je nutné v rámci standardního statistického usuzování zahrnout do statistické inference i vliv zjištěných neodpovědí. Rozdělení chybějících údajů je přitom jedním ze zásadních předpokladů při rozhodování o způsobu práce s těmito daty. Pro zkoumání charakteru rozdělení chybějících dat v datovém souboru proto byly podle [11] zavedeny přidružené indikátorové proměnné, které umožňují přiblížit důvody chybění dat.

3. KLASIFIKACE MECHANISMU CHYBĚNÍ DAT

Mechanismus vzniku chybějících údajů, kterým se pro účely analýzy rozumí jejich vztah k hodnotám dalších proměnných v datovém souboru, je pro nalezení adekvátního postupu v procesu statistické inference podstatný. Vysvětlení vzniku chybějících údajů a argumenty pro nakládání s datovými soubory obsahující chybějící údaje je třeba hledat už v procesu sběru dat. Na základě klasifikace uvedené v práci [11] nebo [2] podle mechanismu vzniku mohou pak chybějící údaje být [10]:

- **zcela náhodné** (Missing Completely at Random − MCAR), kdy rozdělení náhodné veličiny $\boldsymbol{M}$ nezávisí na $\boldsymbol{Y}$, neboli pravděpodobnost, že údaj bude u jednotky chybět, nezávisí na hodnotách pozorované či chybějící náhodné veličiny $\boldsymbol{Y}$ (například chybějící údaje o příjmu nezávisejí na tom, zda jde o příjmy malé či velké, příjmy mužů či žen atd.). Jedná se o nejvíce žádoucí případ, který nezpůsobuje zkreslení prováděných odhadů;
- **náhodné** (Missing At Random − MAR), kdy rozdělení náhodné veličiny $\boldsymbol{M}$ závisí pouze na zjištěných (pozorovaných) hodnotách náhodné veličiny $\boldsymbol{Y}$ (například chybějí spíše údaje o příjmu mužů, nezávisle na tom, zda jde o příjmy menší či větší). Jedná se o případ chybění, které je také žádoucí. MCAR je zřejmě zvláštním případem MAR;
- **nenáhodné** (Not Missing At Random − NMAR), kdy rozdělení náhodné veličiny $\boldsymbol{M}$ závisí na chybějících hodnotách sledované $\boldsymbol{Y}$. Například pokud chybějí převážně údaje o příjmu mužů, které jsou spíše vyšší. Jedná se nejproblematičtější případ.

Formálně se podle [9] výše uvedená klasifikace mechanismu chybění dat dá vyjádřit pomocí podmíněného rozdělení pravděpodobnosti $P(\boldsymbol{M}|y)$ indikátoru chybějících dat (náhodné veličiny) $\boldsymbol{M}$ za podmínky realizace sledované náhodné veličiny $\boldsymbol{Y}$, kde $\boldsymbol{Y} = \left(\boldsymbol{Y}_1, \cdots \boldsymbol{Y}_p\right)^T$ reprezentuje náhodný vektor $\boldsymbol{p}$ sledovaných proměnných a $\boldsymbol{y} = \left(y_1, \cdots y_p\right)^T$ představuje realizaci $\boldsymbol{Y}$. Pro definici jednotlivých typů mechanismů neúplných dat se dále bude předpokládat, že chybějící hodnoty obsahuje pouze náhodná proměnná $\boldsymbol{Y}_1$. Náhodný indikátor neúplných dat $\boldsymbol{M}$ s hodnotou 1, tj. $\boldsymbol{M} = 1$, bude proto indikovat chybějící hodnoty v proměnné $\boldsymbol{Y}_1$.

Zcela náhodný mechanismus chybění dat (anglická zkratka MCAR) vzniká tehdy, kdy rozdělení pravděpodobnosti $P(\boldsymbol{M}|y)$ indikátorové veličiny $\boldsymbol{M}$ nezávisí na realizovaných hodnotách pozorované či chybějící veličině $\boldsymbol{Y}$, tj.

$$P(\boldsymbol{M}=1|y)=P(\boldsymbol{M}=1) \text{ a } P(\boldsymbol{M}=0|y)=P(\boldsymbol{M}=0) \quad (1)$$

Předpoklad toho, že data chybějí zcela náhodně (MCAR), patří k těm nejpřísnějším. K definování zbývajících typů mechanismu chybění je nutné rozdělit realizaci sledované náhodné proměnné $\boldsymbol{Y}_1$ na hodnoty pozorované $(y_{obs})^{\boldsymbol{T}}$ a hodnoty nepozorované (chybějící) $(y_{miss})^{\boldsymbol{T}}$, tj. $y=(y_{obs},y_{miss})^{\boldsymbol{T}}$. Jelikož podle předpokladu chybějící hodnoty obsahuje pouze náhodná proměnná $\boldsymbol{Y}_1$, pak platí, že $y_{miss}=y_1$ a $y_{obs}=\left(y_2,\cdots y_p\right)^T$.

Náhodné rozložení chybějících dat (MAR) je méně striktním předpokladem. Rozdělení chybějících dat reprezentovaném rozdělením indikátorové veličiny $\boldsymbol{M}$ – a i když jsou data rozdělena náhodně – již není nezávislé na sledovaných datech. Pro rozdělení indikátorové veličiny $\boldsymbol{M}$ platí, že závisí na pozorovaných hodnotách $(y_{obs})^{\boldsymbol{T}}$ náhodné proměnné $\boldsymbol{Y}$, ale nezávisí na chybějících hodnotách $(y_{miss})^{\boldsymbol{T}}$ této náhodné proměnné tj. pro náhodné rozložení chybějících dat (MAR) platí:

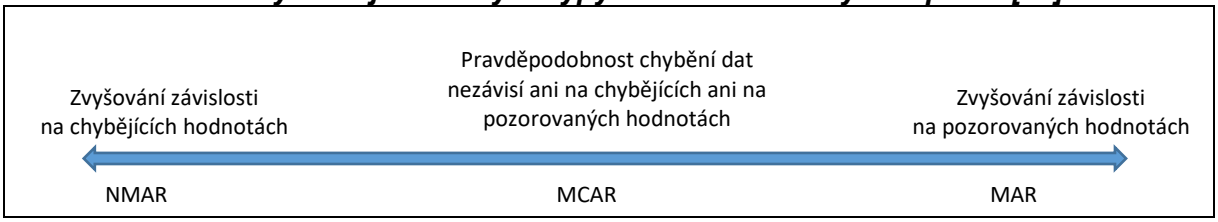
$$\begin{aligned} &P(\boldsymbol{M}=1|(y_{obs},y_{miss})^{\boldsymbol{T}})=P(\boldsymbol{M}=1|(y_{obs})^{\boldsymbol{T}}) \text{ a} \\ &P(\boldsymbol{M}=0|(y_{obs},y_{miss})^{\boldsymbol{T}})=P(\boldsymbol{M}=0|(y_{obs})^{\boldsymbol{T}}) \end{aligned} \quad (2)$$

Nenáhodné rozdělení chybějících dat (NMAR) vzniká tehdy, když rozdělení indikátorové proměnné závisí nejen na pozorovaných hodnotách $\boldsymbol{Y}$, ale i na hodnotách nepozorovaných (chybějících). Pro náhodné rozdělení indikátoru $\boldsymbol{M}$ proto platí:

$$\begin{aligned} &P(\boldsymbol{M}=1|(y_{obs},y_{miss})^{\boldsymbol{T}}) \text{ a} \\ &P(\boldsymbol{M}=0|(y_{obs},y_{miss})^{\boldsymbol{T}}). \end{aligned} \quad (3)$$

Z výše uvedených definic jednotlivých typů mechanismu chybění vyplývá, že zcela náhodný mechanismus chybění dat lze být považován za speciální případ mechanismů chybění náhodného a nenáhodného. Jsou-li náhodně neúplná data nezávislá na pozorovaných hodnotách $(y_{obs})^{\boldsymbol{T}}$, tj. závislost mechanismu chybění na pozorovaných hodnotách je nulová, jsou data zcela náhodně chybějící (MCAR). Jsou-li neúplná data závislá na chybějících hodnotách $(y_{miss})^{\boldsymbol{T}}$, tj. závislost mechanismu chybění na nepozorovaných hodnotách je nenulová, jsou data náhodně chybějící (NMAR). Vztahy mezi jednotlivými typy mechanismy chybění pro spojitá rozdělení ilustruje obrázek č. 1.

Obrázek č. 1: Vztahy mezi jednotlivými typy mechanismu chybění podle [14]



Zdroj: vlastní zpracování podle [12]

4. IGNOROVATELNÝ A NEIGNOROVATELNÝ MECHANISMUS CHYBĚNÍ DAT

Ignorovatelná data jsou typy chybějících dat, které lze efektivně zpracovat moderními technikami pro odhady za podmínek chybějících dat, jako jsou expektační-maximalizační algoritmy, metody maximálně-věrohodné odhadů a metody mnohonásobné imputace. Chybějící data musí splňovat dvě podmínky, aby se stala ignorovatelnými chybějícími daty:

- chybějící údaje jsou buď chybějící podle mechanismu MCAR nebo MAR a
- parametry spojené s konkrétním pravidlem pro chybějící data se liší od parametrů spojených s distribucí proměnných v datovém souboru [11].

Druhá podmínka znamená, že parametry související s rozdělením indikátoru chybění $\boldsymbol{M}$ se liší od parametrů spojených s pravděpodobnostním rozdělením $\boldsymbol{Y}$. Pro pochopení důvodu, proč jsou tyto podmínky potřebné, nechť $\boldsymbol{\theta}$ resp. $\boldsymbol{\varphi}$ označují parametry spojené s rozdělením $\boldsymbol{Y}$, resp. $\boldsymbol{M}$, a nechť $f(y, m; \boldsymbol{\theta}, \boldsymbol{\varphi})$ označuje sdruženou hustotu pravděpodobnosti rozdělení $\boldsymbol{M}$ a $\boldsymbol{Y}$. Vzhledem k tomu, že $\boldsymbol{\theta}$ resp. $\boldsymbol{\varphi}$ jsou parametry vzájemně odlišné, za podmínek neúplných dat, pravděpodobnost pozorovaných dat lze získat prostřednictvím okrajové (marginální) hustoty $(y_{obs})^{\boldsymbol{T}}$ takto:

$$f(y_{obs}, m, \theta, \varphi) = \int f(y_{obs}, y_{miss}; \theta) f(m | y_{obs}, y_{miss}; \varphi) dy_{miss}. \quad (4)$$

Jsou-li neúplná data zcela náhodně chybějící (MCAR), potom:

$$f(m | y_{obs}, y_{miss}; \varphi) = f(m; \varphi). \quad (5)$$

Jsou-li neúplná data náhodně chybějící (MAR), potom:

$$f(m | y_{obs}, y_{miss}; \varphi) = f(m | y_{obs}; \varphi). \quad (6)$$

Jelikož ani $f(m; \varphi)$, ani $f(m | y_{obs}; \varphi)$ nezahrnují y_{miss}, lze hustoty $f(m; \varphi)$ i $f(m | y_{obs}; \varphi)$ vyjmout před integrál (vzorec 4) a pro maximálně-věrohodné odhady je postačující integrál $\int f(y_{obs}, y_{miss}; \theta) dy_{miss}$ k odhadu parametru θ maximalizovat jen a pouze vzhledem k parametru θ. I v případě, že neúplná data typu MAR nesplňují druhou podmínku o různých parametrech θ a φ, statistické metody předpokládají ignorovatelnost mechanismu chybění (i když ne optimální, ale stále vyhovující pro nevychýlené odhady). V praxi jsou proto mechanismy chybění zcela náhodné (MCAR) a náhodné (MAR) považovány za ignorovatelné a mechanismus nenáhodných chybění (NMAR) implikuje mechanismus chybění neignorovatelný.

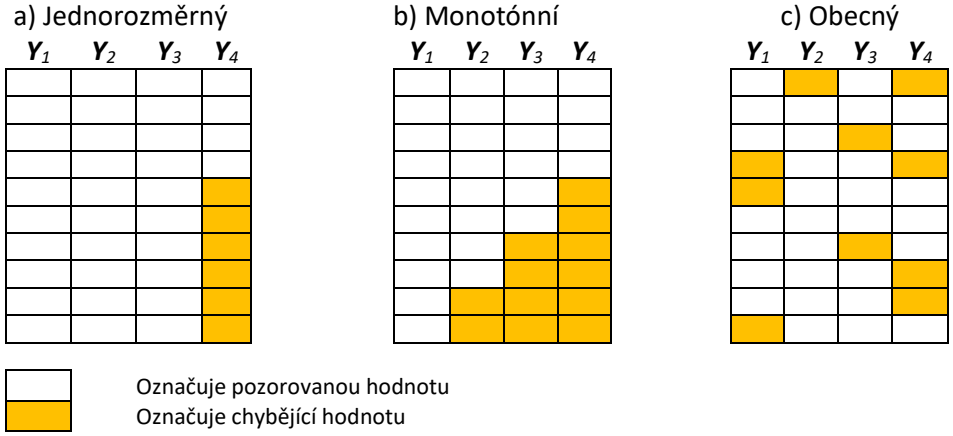
5. VZORY CHYBĚJÍCÍCH DAT

Vzor chybějících dat[2] se týká uspořádání pozorovaných a chybějících hodnot v souboru dat [4]. Často se zaměňuje s mechanismem neúplných dat (např. [8]). Rozdíl je v tom, že specifický mechanismus chybění dat je datové pravidlo v neúplných datech, které popisuje vztah mezi subjekty (zpravodajskými jednotkami) a pravděpodobností chybějících dat. Na druhou stranu specifický chybějící datový vzor je datová konfigurace, která popisuje umístění chybějících hodnot v datech.

[2] *Vzor chybějících dat je překlad anglického originálu Missing Data Pattern.*

Obecně existují tři druhy chybějících datových vzorů. Jednorozměrný vzor vzniká, když chybí hodnoty u jedné proměnné nebo skupiny proměnných, která je buď zcela pozorována, nebo zcela chybí pro každý případ (zpravodajskou jednotku), ale všechny ostatní proměnné jsou zcela pozorovány [12] (viz obrázek č. 2a). Jednorozměrný vzor má nejnižší počet chybějících datových vzorů; jinými slovy, má dva vzory chybějících dat; jeden vzor, kde subjekty mají úplná data, a druhý vzor, kde subjektům více či méně chybí data. Dalším typem chybějícího vzoru je monotónní vzor neúplných dat (např. [12]). V monotónním chybějícím vzoru lze skupinu proměnných $(\boldsymbol{Y}_1, \cdots \boldsymbol{Y}_p)^T$ seřadit tak, že pokud chybí hodnota $\boldsymbol{Y}_j$ u j-té zpravodajské jednotky, pak hodnoty $(\boldsymbol{Y}_{j+1}, \cdots \boldsymbol{Y}_p)^T$ také chybí (viz obrázek č. 2b). Na jednorozměrný vzor lze pohlížet jako na zvláštní případ monotónního vzoru. A konečně, obecný vzor chybějících dat se vytvoří, když může chybět skupina proměnných pro jakýkoli zpravodajský subjekt. Toto vytváří datový soubor s chybějícími hodnotami rozptýlenými v datové matici náhodným způsobem [4] (viz obrázek č. 2c). Uvedené vzory chybění dat jsou ilustrovány na obrázku č. 2.

Obrázek č. 2: Ilustrace základních typů vzorů neúplných dat



Zdroj: vlastní zpracování podle [12]

Přestože datový vzor neúplných dat a mechanismus neúplných dat mají odlišný význam, vzájemně se ovlivňují. Vzhledem ke konkrétnímu pravidlu pro chybějící data s určitým typem mechanismu chybějících dat bude určen počet a typ vzoru neúplných dat. Má-li například datová množina proměnné $(\boldsymbol{Y}_1, \cdots \boldsymbol{Y}_p)^T$ a pokud pravidlem chybějících dat je, že každý subjekt má pravděpodobnost 20 % chybějících hodnot v proměnné $\boldsymbol{Y}_1$, pak je vzor neúplných dat jednorozměrný, což se rovná dvěma vzorům chybění.

I když vzory neúplných dat nemají přímý vliv na mechanismus neúplných dat, je znalost vzoru neúplných dat pro volbu efektivní metody odhadů parametrů velmi důležitá [11].

6. STATISTICKÉ MODELY CHYBĚJÍCÍCH DAT

Statistické modely zcela náhodného mechanismu chybění (MCAR)

Zcela náhodný mechanismus chybění je charakterizován pravděpodobností, že u každé zpravodajské jednotky budou data neúplná v jedné anebo více proměnných. Pravděpodobnost chybějícího údaje je hodnota parametru označovaná jako π.

Hodnota parametru má vliv na očekávané procento chybějících údajů a očekávaný počet vzorů chybění výběrového souboru.

U dat s mechanismem chybění typu MCAR s jednorozměrným vzorem neúplných dat platí pro chybějící údaje pravidlo, že každému zpravodajskému subjektu je přiřazena pravděpodobnost neúplných údajů v proměnné (proměnných). Pro neúplná data tedy platí, že pravděpodobnost chybění je $P(\boldsymbol{M}=1)=\pi$, kde $\boldsymbol{M}$ je indikátor chybějících dat a π je parametr vyjadřující hodnotu pravděpodobnosti chybějících údajů. Na základě znalosti tohoto pravděpodobnostního pravidla mohou výzkumníci určit různé vlastnosti spojené s údaji MCAR, včetně očekávaného procenta chybějících hodnot a očekávaného počtu chybějících vzorů neúplných dat.

K pochopení, jak zcela náhodný mechanismus neúplných dat ovlivňuje vlastnosti dat ve výběrovém souboru, je zavedena náhodná proměnná K udávající počet zpravodajských jednotek (případů) s chybějícími údaji. Za předpokladu nezávislosti šance chybění údajů v proměnné jedné zpravodajské jednotky na šanci chybění u ostatních jednotek, náhodná proměnná K podléhá binomickému rozdělení $K \sim Bin(n,\pi)$, kde n je celkový počet zpravodajských jednotek ve výběrovém souboru a $0 \leq \pi \leq 1$. Jelikož $E(K)=n\pi$ a $Var(K)=n\pi(1-\pi)$, potom očekávané procento chybějících hodnot je:

$$E(\prod)=E\left(\frac{K}{n}\right)=\frac{1}{n}E(K)=\pi, \tag{7}$$

kde $\prod=\frac{K}{n}$ je náhodná proměnná popisující odhadované procento chybějících hodnot ve výběrovém souboru. Rozptyl odhadu procenta chybějících údajů je:

$$Var(\prod)=Var\left(\frac{K}{n}\right)=\frac{1}{n^2}Var(K)=\frac{\pi(1-\pi)}{n}. \tag{8}$$

Rozptyl odhadu procenta podle vzorce (8) vyjadřuje skutečnost, že se ve výběrovém souboru nemusí objevit přesná hodnota očekávaného procenta chybějících údajů.

Na základě pravděpodobnosti neúplných dat π je možné ve výběrovém souboru také odvodit očekávaný počet vzorů chybění. U dat s mechanismem chybění typu MCAR s jednorozměrným vzorem neúplných dat platí, že jeden vzor zahrnuje zpravodajské jednotky s úplnými daty; vzor druhý zahrnuje zpravodajské jednotky s chybějícími hodnotami. Nechť $\boldsymbol{I}_j$, pro $j \in \{1,2\}$, je indikační proměnná události, že j–tý vzor neúplných dat platí alespoň u jedné zpravodajské jednotky z výběrového souboru. Pravděpodobnost, že vzor neúplných dat pro $j=1$ nastal alespoň u jedné zpravodajské jednotky, je $P(I_1=1)=E(I_1)=1-\pi^n$. Pravděpodobnost, že vzor neúplných dat pro $j=2$ nastal alespoň u jedné zpravodajské jednotky, je $P(I_2=1)=E(I_2)=1-(1-\pi)^n$. Nechť D označuje počet odlišných vzorů neúplných dat, tj. $D=\sum_{j=1}^{2} I_j$. Očekávaný počet odlišných chybějících datových vzorů potom je:

$$E(D)=E\left(\sum_{j=1}^{2} I_j\right)=\sum_{j=1}^{2} E\left(I_j\right)=1-\pi^n+1-(1-\pi)^n=2-\pi^n-(1-\pi)^n, \tag{9}$$

Se zvyšováním velikosti výběrového souboru n očekávaný počet vzorů neúplných dat konverguje hodnotě 2 (a to je maximální počet vzorů neúplných dat pro daný mechanismus chybění).

Objevují-li se chybějící hodnoty ve více než 1 proměnné $\boldsymbol{Y}_1, \cdots \boldsymbol{Y}_l$, pro $i \in \{1, \cdots, l\}$, potom lze očekávat při stávajícím mechanismu chybění celkem l parametrů, $\pi_1, \cdots \pi_l$, popisující mechanismů neúplných dat pro l proměnných. Pro každou proměnnou s neúplnými daty $\boldsymbol{Y}_i$, je očekávané procento chybějících hodnot ve výběrovém souboru a jeho rozptyl dané rovnicemi (7) a (8). Daný mechanismus chybění, kde vzor neúplných dat obsahuje až l proměnných s chybnými údaji, může generovat až $m = 2^l$ vzorů neúplných dat. Očekávaný počet různých vzorů neúplných dat ve výběrovém souboru s l proměnnými s chybějícími údaji je definován rovnicí:

$$E(D) = m - \sum_{j=1}^{m} (1 - \eta_j), \tag{10}$$

kde $\eta_1, \cdots, \eta_m$ jsou odpovídající pravděpodobnosti pro m vzorů neúplných dat. U zcela náhodného mechanismu chybění (MCAR) pravděpodobnosti pro vytvoření vzorů neúplných dat, tj. $\eta_1, \cdots, \eta_m$, závisí jen a pouze na pravděpodobnostech neúplných údajů, které se objeví v proměnných $\boldsymbol{Y}_1, \cdots \boldsymbol{Y}_l$, pravděpodobnosti $\pi_1, \cdots \pi_l$.

Statistické modely náhodného mechanismu chybění (MAR)

U výběrových souborů s daty generovanými náhodným mechanismem chybění (MAR) závisí pravděpodobnost, že daná jednotka bude mít chybějící hodnotu, na pozorovaných hodnotách jiných proměnných. To znamená, že pravděpodobnost chybějících hodnot lze předpovídat z pozorovaných hodnot jiných proměnných. Proměnnou, která může předpovídat pravděpodobnost chybějících hodnot, je proto možné nazývat jako prediktor chybějících dat. Prediktorem chybějících dat může být jedna proměnná datový soubor nebo to může být nová proměnná, která je lineární kombinací několika proměnných v datovém souboru.

Modelování neúplných dat za podmínek náhodného mechanismu chybění MAR je složitější než modelování dat MCAR ze dvou důvodů:

- Pravděpodobnostní pravidla pro chybějící data jsou složitější než pravidla pro data zcela náhodná a dají se rozdělit do několika kategorií:
  - s jednoduchou rozhodovací (diskriminační) hodnotou,
  - s vícenásobnými rozhodovacími (diskriminačními) hodnotami,
  - percentilové rozhodovací hodnoty,
  - založené na logistické regresi.
- Mezi indikátorem chybějících dat a prediktorem chybějících dat se může měnit síla a tvar závislosti. Síla závislosti může být slabá nebo silná; tvar závislosti může být lineární nebo nelineární.

Náhodný mechanismus chybění MAR s jednoduchou rozhodovací hodnotou

Pravděpodobnostní pravidla pro chybějící data spojená s jedinou mezní hodnotou zahrnují určení této mezní hodnoty pro každý prediktor chybějících dat. Pro neúplná data s náhodným mechanismem chybění MAR je $\boldsymbol{Y}_1$ proměnná s chybějícími daty a $\boldsymbol{Y}_2$ je prediktor chybějících dat s hraničním bodem a. V tomto případě mají data jednorozměrný vzor chybění. Pravděpodobnostní pravidlo pro neúplná data je: pokud má jednotka hodnotu proměnné $Y_2 \geq a$, pak pravděpodobnost, že chybí hodnota $\boldsymbol{Y}_1$, je π_1; pokud má jednotka hodnotu proměnné $Y_2 < a$, pak jeho pravděpodobnost, že chybí hodnota $\boldsymbol{Y}_1$, je π_2. Formálně lze toto pravděpodobnostní pravidlo definovat takto: nechť $\boldsymbol{M}$ je indikátor chybějících dat pro chybějící data v proměnné $\boldsymbol{Y}_1$ a $\boldsymbol{U}$ je indikátor označující, zda proměnná $\boldsymbol{Y}_2$ nabývá hodnoty a a vyšší (tj. $\boldsymbol{U} = 1$, když $Y_2 \geq a$ a $\boldsymbol{U} = 0$,

když $Y_2 < a$). Indikátor $\boldsymbol{U}$ je náhodná proměnná, také prediktor chybějících údajů, a je lineární funkcí náhodné proměnné $\boldsymbol{Y}_2$. Rovnice pravděpodobnostního pravidla pro mechanismus chybění MAR lze zapsat ve tvaru $P(\boldsymbol{M} = 1|\boldsymbol{U} = 1) = \pi_1$ a $P(\boldsymbol{M} = 1|\boldsymbol{U} = 0) = \pi_2$. Pravděpodobnostní pravidlo chybějících údajů zahrnuje pouze dva indikátory $\boldsymbol{M}$ a $\boldsymbol{U}$, a proto jej lze nejlépe ilustrovat pomocí kontingenční tabulky pro tyto dva indikátory. Tato kontingenční tabulka je uvedena na obrázku č. 3.

Obrázek č. 3: Kontingenční tabulka pro náhodný mechanismus neúplných dat s jednoduchou diskriminační hodnotou

	M	
U	1	0
1	$P(\boldsymbol{M} = 1\|\boldsymbol{U} = 1)P(\boldsymbol{U} = 1) = \pi_1\pi_0$	$P(\boldsymbol{M} = 0\|\boldsymbol{U} = 1)P(\boldsymbol{U} = 1) = (1 - \pi_1)\pi_0$
0	$P(\boldsymbol{M} = 1\|\boldsymbol{U} = 0)P(\boldsymbol{U} = 0) = \pi_2(1 - \pi_0)$	$P(\boldsymbol{M} = 0\|\boldsymbol{U} = 0)P(\boldsymbol{U} = 0) = (1 - \pi_2)(1 - \pi_0)$

$\boldsymbol{M}$ je indikátorová proměnná označující, kdy je hodnota proměnné $\boldsymbol{Y}_1$ chybějící: $\boldsymbol{M} = 1$, když je hodnota proměnné $\boldsymbol{Y}_1$ chybějící, a $\boldsymbol{M} = 0$, když není hodnota proměnné $\boldsymbol{Y}_1$ chybějící

$\boldsymbol{U}$ je indikátorová proměnná označující, kdy je hodnota proměnné $\boldsymbol{Y}_2$ větší nebo rovna diskriminační hodnotě a: $\boldsymbol{U} = 1$, když je hodnota proměnné $\boldsymbol{Y}_2 \geq a$, a $\boldsymbol{U} = 0$, když je hodnota proměnné $\boldsymbol{Y}_2 < a$.

Zdroj**: **vlastní zpracování podle [9]

Symboly π_1 a π_2 označují parametry vyjadřující podmíněné pravděpodobnosti chybějících údajů v proměnné $\boldsymbol{Y}_1$. Parametr π_0 je pravděpodobnost, že hodnota v proměnné $\boldsymbol{Y}_2$ je rovná nebo větší než a, tj. $P(\boldsymbol{Y}_2 \geq a) = P(\boldsymbol{U} = 1) = \pi_0$. Pravděpodobnost π_0 přímo souvisí s mezním bodem a. K nastavení hodnoty pro tento parametr π_0 je potřebné pouze specifikovat mezní hodnotu a. Podobně jako u zcela náhodného mechanismu chybění lze pro každý pravděpodobnostní parametr vypočítat rozptyl spojený s odhadovanou hodnotou [9]. Jestliže n je celkový počet jednotek a $n_1 = n\pi_0$ je počet jednotek s hodnotami proměnné $\boldsymbol{Y}_2 \geq a$, potom příslušné rozptyly pro odhadované π_0, π_1 a π_2 jsou:

$$Var(\prod_0) = \frac{\pi_0(1-\pi_0)}{n}, \tag{11a}$$

$$Var(\prod_1) = \frac{\pi_1(1-\pi_1)}{n} \text{ a} \tag{11b}$$

$$Var(\prod_2) = \frac{\pi_2(1-\pi_2)}{n}. \tag{11c}$$

K určení očekávaného procenta chybějících hodnot v proměnné $\boldsymbol{Y}_1$, je nutné nejprve vypočítat nepodmíněnou pravděpodobnost chybějících údajů v proměnné $\boldsymbol{Y}_1$:

$$\begin{aligned} \pi_{miss} &= P(\boldsymbol{M} = 1) \\ &= P(\boldsymbol{M} = 1|\boldsymbol{U} = 1)P(\boldsymbol{U} = 1) + P(\boldsymbol{M} = 1|\boldsymbol{U} = 0)\big(1 - P(\boldsymbol{U} = 1)\big) \\ &= \pi_1\pi_0 + \pi_2(1 - \pi_0) \end{aligned} \tag{12}$$

Počet zpravodajských jednotek (případů) s chybějícími údaji je náhodná proměnná K podléhající binomickému rozdělení (předpokládá se nezávislost mezi vznikem chybějících údajů), tj. $K \sim Bin(n, \pi_{miss})$. Očekávané procento chybějících údajů v proměnné $\boldsymbol{Y}_1$ ve výběrovém souboru je:

$$E\left(\frac{K}{n}\right) = \pi_1\pi_0 + \pi_2(1-\pi_0), \tag{13}$$

a rozptyl pro odhadované procento (13) chybějících údajů v proměnné $\boldsymbol{Y}_1$ je:

$$Var\left(\frac{K}{n}\right) = \frac{\pi_{miss}(1-\pi_{miss})}{n}. \tag{14}$$

Očekávaný počet vzorů chybějících dat pro náhodný mechanismus chybění se podobně jako v (9) určí rovnicí:

$$E(D) = 2 - \pi_{miss}^n - (1-\pi_{miss})^n, \tag{15}$$

K posouzení vlastností náhodného mechanismu chybění patří také i zjištění síly závislosti mezi chybějícími údaji v proměnné $\boldsymbol{Y}_1$ na pozorovaných hodnotách proměnné $\boldsymbol{Y}_2$ představované mírou asociace mezi indikátorem chybějících údajů $\boldsymbol{M}$ a indikátorem $\boldsymbol{U}$ hodnoty proměnné $\boldsymbol{Y}_2$. Pokud bude tato síla závislosti nulová, pozorované hodnoty proměnné $\boldsymbol{Y}_2$ nemají vliv na chybějící hodnoty proměnné $\boldsymbol{Y}_1$ a z náhodného mechanismu chybění MAR se stává zcela náhodný mechanismus chybění MCAR (viz obrázek č. 1).

Podobně jako u ostatních binárních proměnných, sílu závislosti mezi indikátorem chybějících údajů $\boldsymbol{M}$ a indikátorem $\boldsymbol{U}$ hodnoty proměnné $\boldsymbol{Y}_2$ lze měřit pomocí absolutního rozdílu pravděpodobností (ARD) anebo použitím poměru šancí (OR), tj.

$$ARD = \pi_1 - \pi_2 \text{ a} \tag{16}$$

$$OR = \frac{P(\boldsymbol{M}=1|\boldsymbol{U}=1)/\big(1-P(\boldsymbol{M}=1|\boldsymbol{U}=1)\big)}{P(\boldsymbol{M}=1|\boldsymbol{U}=0)/\big(1-P(\boldsymbol{M}=1|\boldsymbol{U}=0)\big)} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{17}$$

Symbol ARD [1] pro absolutní rozdíl pravděpodobností vychází z anglického názvu *absolute risk difference* a OR (z anglického názvu *Odds Ratio*) znamená poměr šancí. Velká hodnota ARD, resp. hodnota OR větší než 1 (čím dále od 1, tím více) indikuje silnou závislost; je-li $1-\pi_1 = 0$ anebo $1-\pi_2 = 0$ poměr šancí OR není definován. Proto k měření síly závislosti se spíše používá ukazatel ARD.

Rovnice (16) a (17) měří sílu závislosti mezi indikátory (binárními proměnnými) $\boldsymbol{M}$ a $\boldsymbol{U}$ na úrovni populace. Na úrovni výběrového souboru se odhadovaná síla závislosti může lišit vzorek od vzorku. Odchylky spojené s odhadovanou ARD a odhadovaným logaritmem poměru šancí OR jsou následující (odvození viz [14]):

$$Var(\prod_1 - \prod_2) = Var(\prod_1) + Var(\prod_2) = \frac{\pi_1(1-\pi_1)}{n} + \frac{\pi_2(1-\pi_2)}{n}, \tag{18}$$

$$Var\big(log\ (OR)\big) = \frac{1}{\pi_1\pi_0} + \frac{1}{(1-\pi_1)\pi_0} + \frac{1}{\pi_2(1-\pi_0)} + \frac{1}{(1-\pi_2)(1-\pi_0)} \tag{19}$$

kde výrazy ve jmenovatelích jsou definovány v kontingenční tabulce na obrázku č. 3.

Jelikož poměr šancí lze vyjádřit také pomocí logistického regresního modelu [1], vztah mezi indikátorem chybějících údajů $\boldsymbol{M}$ a $\boldsymbol{U}$ indikátorem hodnoty proměnné $\boldsymbol{Y}_2$ je

možné také vyjádřit na základě logistické regrese. Logaritmus poměru šancí může OR být predikován pomocí indikátoru $\boldsymbol{U}$ takto:

$$log\left(\frac{P(\boldsymbol{M}=1)}{1-P(\boldsymbol{M}=1)}\right) = \beta_0 + \beta_1 U, \tag{20}$$

kde koeficient β_0 je hodnota logaritmu[3] poměru šancí OR za platnosti $\boldsymbol{U} = 0$:

$$\beta_0 = log\left(\frac{P(\boldsymbol{M}=1/\boldsymbol{U}=0)}{1-P(\boldsymbol{M}=1)/\boldsymbol{U}=0}\right) = log\left(\frac{\pi_2}{1-\pi_2}\right) \text{ a} \tag{21}$$

kde koeficient β_1 je hodnota logaritmu poměru šancí OR:

$$\beta_1 = log\,(OR) = log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) \tag{22}$$

Z rovnice (20) vyplývá, že čím je vyšší hodnota koeficientu β_1, tím je silnější závislost mezi indikátory $\boldsymbol{M}$ a $\boldsymbol{U}$. Rovnice (20) také dokazuje, že pravděpodobnostní pravidlo s jednou diskriminační hodnotou je současně ekvivalentní logistickému regresnímu modelu, který může popisovat náhodný mechanismus chybění MAR pomocí logistické regrese.

Náhodný mechanismus chybění MAR s vícenásobnou rozhodovací hodnotou

U náhodného mechanismu chybění MAR s vícenásobnou rozhodovací hodnotou je třeba zadat více hraničních bodů pro prediktor chybějících dat. Jednou z výhod použití vícenásobné rozhodovací hodnoty je, že model lze použít k vytvoření lineárního i nelineárního vztahu mezi indikátorem chybějících dat $\boldsymbol{M}$ a prediktorem chybějících dat $\boldsymbol{Y}_2$. K nelineárnímu vztahu dochází, když subjekty s extrémními hodnotami na prediktoru chybějících dat mají vyšší nebo nižší pravděpodobnost, že nebudou chybět, než subjekty s hodnotami středního rozsahu na prediktoru. Naproti tomu k lineárnímu vztahu dochází, když se pravděpodobnost, že bude chybět, postupně zvyšuje nebo snižuje s tím, jak se zvyšuje hodnota prediktoru chybějících dat.

Pro vytvoření nelineárního vztahu předpokládejme, že pravděpodobnost chybějících hodnot u proměnné $\boldsymbol{Y}_1$ závisí na dvou hraničních bodech, a a $-a$, hodnot proměnné $\boldsymbol{Y}_2$. Podobně jako u modelu s jednoduchou rozhodovací hodnotou nechť $\boldsymbol{M}$ je indikátor chybějících dat a $\boldsymbol{U}$ je indikátor označující, zda je hodnota $\boldsymbol{Y}_2$ mezi dvěma diskriminačními body, tj. $\boldsymbol{U} = 1$, když $Y_2 \geq a$ anebo $Y_2 \leq -a$ a $\boldsymbol{U} = 0$, když $-a < Y_2 < a$. Toto pravidlo chybějících dat je stejné jako pravidlo pro model s jedinou rozhodovací hodnotou. Jinými slovy, v případě nelineárního vztahu může být na chybějící datové pravidlo spojené s vícenásobnými rozhodovacími hodnotami nahlíženo tak, jako na pravidlo chybějících dat spojené s jednoduchou rozhodovací hodnotou. Důsledkem je, že v tomto případě lze všechny rovnice pro model s jednoduchou rozhodovací hodnotou použít i pro model s vícenásobnými rozhodovacími hodnotami [14].

K vytvoření lineárního vztahu mezi indikátorem chybějících dat $\boldsymbol{M}$ a prediktorem chybějících dat $\boldsymbol{Y}_2$ je nutné specifikovat alespoň dva hraniční body v prediktoru chybějících dat $\boldsymbol{Y}_2$. Ve většině případů se specifikují tři nebo čtyři hraniční body, obvykle kvartilové nebo kvantilové hodnoty prediktoru chybějících dat $\boldsymbol{Y}_2$. Proměnná

[3] Symbol log označuje logaritmus o základu e, tj. přirozený logaritmus.

$\boldsymbol{Y}_2$ se rozdělí do čtyř nebo pěti skupin, přičemž hodnoty prediktoru chybějících dat $\boldsymbol{Y}_2$ u každé zpravodajské jednotky mají stejnou šanci být v kterékoli ze skupin. Přechází-li se ze skupiny s nejnižšími hodnotami prediktoru $\boldsymbol{Y}_2$ do skupiny s nejvyššími hodnotami, pravděpodobnost chybění obvykle roste nebo klesá konstantní rychlostí.

Nechť $\boldsymbol{M}$ je indikátor chybějících dat a $\boldsymbol{V}$ je diskrétní uniformní náhodná proměnná vytvořená na základě hodnot $\boldsymbol{Y}_2$ (tj. $\boldsymbol{V}$ lze považovat za prediktor chybějících dat), potom náhodná proměnná $\boldsymbol{V}$ vystupující jako prediktor chybějících dat $\boldsymbol{Y}_2$ nabývá následujících hodnot:

$$\boldsymbol{V} = \begin{cases} 1 & \text{if } \boldsymbol{Y}_2 < Q_1 \\ 2 & \text{if } Q_1 \leq \boldsymbol{Y}_2 < Q_2 \\ 3 & \text{if } Q_2 \leq \boldsymbol{Y}_2 < Q_3 \\ 4 & \text{if } \boldsymbol{Y}_2 \geq Q_3 \end{cases} \tag{23}$$

kde symboly Q_1, Q_2, a Q_3 jsou kvartilové hodnoty prediktoru chybějících hodnot $\boldsymbol{Y}_2$. Pravděpodobnostní pravidlo chybějících dat pro proměnnou $\boldsymbol{Y}_1$ je potom:

$$\begin{gathered} P(\boldsymbol{M} = 1|\boldsymbol{V} = 1) = \pi_1 \\ P(\boldsymbol{M} = 1|\boldsymbol{V} = 2) = \pi_2 \\ P(\boldsymbol{M} = 1|\boldsymbol{V} = 3) = \pi_3 \\ a \\ P(\boldsymbol{M} = 1|\boldsymbol{V} = 4) = \pi_4 \end{gathered} \tag{24}$$

Pravděpodobnostní pravidlo chybějících údajů s více diskriminačními hodnotami lze opět nejlépe ilustrovat pomocí kontingenční tabulky pro dva indikátory. Tato kontingenční tabulka je uvedena na obrázku č. 4.

Obrázek č. 4: Kontingenční tabulka pro náhodný mechanismus neúplných dat s více diskriminačními hodnotami

	M	
V	1	0
1	$P(\boldsymbol{M} = 1\|\boldsymbol{V} = 1)P(\boldsymbol{V} = 1) = \pi_1\pi_0$	$P(\boldsymbol{M} = 0\|\boldsymbol{V} = 1)P(\boldsymbol{V} = 1) = (1 - \pi_1)\pi_0$
2	$P(\boldsymbol{M} = 1\|\boldsymbol{V} = 2)P(\boldsymbol{V} = 2) = \pi_2\pi_0$	$P(\boldsymbol{M} = 0\|\boldsymbol{V} = 2)P(\boldsymbol{V} = 2) = (1 - \pi_2)\pi_0$
3	$P(\boldsymbol{M} = 1\|\boldsymbol{V} = 3)P(\boldsymbol{V} = 3) = \pi_3\pi_0$	$P(\boldsymbol{M} = 0\|\boldsymbol{V} = 3)P(\boldsymbol{V} = 3) = (1 - \pi_3)\pi_0$
4	$P(\boldsymbol{M} = 1\|\boldsymbol{V} = 4)P(\boldsymbol{V} = 4) = \pi_4\pi_0$	$P(\boldsymbol{M} = 0\|\boldsymbol{V} = 4)P(\boldsymbol{V} = 4) = (1 - \pi_4)\pi_0$

$\boldsymbol{M}$ je indikátorová proměnná označující, kdy je hodnota proměnné $\boldsymbol{Y}_1$ chybějící: $\boldsymbol{M} = 1$, když je hodnota proměnné $\boldsymbol{Y}_1$ chybějící, a $\boldsymbol{M} = 0$, když není hodnota proměnné $\boldsymbol{Y}_1$ chybějící

$\boldsymbol{V}$ je diskrétní indikátorová proměnná rovnoměrně rozdělená označující, ve kterém kvartilu hodnota proměnné $\boldsymbol{Y}_2$ je: $\boldsymbol{V} = 1$ když $\boldsymbol{Y}_2 < Q_1$, $\boldsymbol{V} = 2$ když $Q_1 \leq \boldsymbol{Y}_2 < Q_2$, $\boldsymbol{V} = 3$ když $Q_2 \leq \boldsymbol{Y}_2 < Q_3$ a $\boldsymbol{V} = 4$ když $\boldsymbol{Y}_2 \geq Q_3$, kde Q_1, Q_2 a Q_3 jsou kvartilové hodnoty proměnné $\boldsymbol{Y}_2$.

***Zdroj**: vlastní zpracování podle [9]*

V pravděpodobnostním pravidlu pro chybějící data vystupuje 5 parametrů. Z nich jsou 4 pravděpodobnostní parametry π_1, π_2, π_3 and π_4. Poslední parametr označený jako π_0 představuje pravděpodobnost indikátoru $\boldsymbol{V}$, tj. $P(\boldsymbol{V} = i) = \pi_0 = 0{,}25$, kde $i \in \{1{,}2{,}3{,}4\}$. Hodnota parametru π_0 je rovna 0,25, protože prediktor chybějících hodnot proměnná $\boldsymbol{Y}_2$ je kvartilovými mezními hodnota rozdělena na 4 stejné intervaly. Pro každý parametr je možné odhadnout rozptyl asociovaný s odhadem počtu procent chybějících údajů. Označí-li se n jako celkový počet hodnot v prediktoru chybějících hodnot, tj. v proměnné $\boldsymbol{Y}_2$, potom počet hodnot v každém kvartilovém intervalu je $n_0 = \pi_0 n = 0{,}25n$. Rozptyl odhadovaného parametru π_0 je definován rovnicí:

$$Var(\prod_0) = \frac{\pi_0(1-\pi_0)}{n}, \tag{25}$$

Rozptyly odhadů parametrů π_j pro $j \in \{1,2,3,4\}$ jsou dány jako:

$$Var(\prod_j) = \frac{\pi_j(1-\pi_j)}{n}, \tag{26}$$

Odhad pravděpodobnosti chybění hodnoty u každé zpravodajské jednotky lze provést pomocí marginální pravděpodobnosti, že indikátor chybění bude roven 1, tj.:

$$\begin{aligned} \pi_{miss} &= P(\boldsymbol{M} = 1) \\ &= P(\boldsymbol{M} = 1|\boldsymbol{V} = 1)P(\boldsymbol{V} = 1) + P(\boldsymbol{M} = 1|\boldsymbol{V} = 2)P(\boldsymbol{V} = 2) \\ &+ P(\boldsymbol{M} = 1|\boldsymbol{V} = 3)P(\boldsymbol{V} = 3) + P(\boldsymbol{M} = 1|\boldsymbol{V} = 4)P(\boldsymbol{V} = 4) \\ &= \pi_1\pi_0 + \pi_2\pi_0 + \pi_3\pi_0 + \pi_4\pi_0 \end{aligned} \tag{27}$$

Počet zpravodajských jednotek (případů) s chybějícími údaji v proměnné $\boldsymbol{Y}_1$ je náhodná proměnná K podléhající binomickému rozdělení (předpokládá se nezávislost vzniků chybějících údajů), tj. $K \sim Bin(n, \pi_{miss})$, kde n je celkový počet hodnot. Očekávané procento chybějících údajů v proměnné $\boldsymbol{Y}_1$ ve výběrovém souboru je:

$$E\left(\frac{K}{n}\right) = \pi_1\pi_0 + \pi_2\pi_0 + \pi_3\pi_0 + \pi_4\pi_0, \tag{28}$$

a rozptyl pro odhadované procento (28) chybějících údajů v proměnné $\boldsymbol{Y}_1$ je:

$$Var\left(\frac{K}{n}\right) = \frac{\pi_{miss}(1-\pi_{miss})}{n}. \tag{29}$$

Očekávaný počet jedinečných vzorů chybějících dat pro náhodný mechanismus chybění MAR s kvantilovými hodnotami se podobně jako v (9) určí rovnicí:

$$E(D) = 2 - \pi_{miss}^n - (1 - \pi_{miss})^n, \tag{30}$$

Sílu závislosti mezi indikátorem chybějících údajů $\boldsymbol{M}$ a indikátorem $\boldsymbol{V}$ hodnoty prediktoru proměnné $\boldsymbol{Y}_2$ lze měřit pomocí průměrného absolutního rozdílu pravděpodobností $AARD$ anebo použitím poměru šancí OR. Pro pravděpodobnostní pravidlo využívající kvartilové hodnoty jako hodnoty rozhodovací je $AARD$ definována ve tvaru:

$$AARD = \frac{\pi_4 - \pi_1}{3}, \tag{31}$$

V poměru šancí OR vystupují marginální pravděpodobnosti pro indikátor chybění $\boldsymbol{M}$ podle (27) jako pravděpodobnosti $\pi_{miss} = P(\boldsymbol{M} = 1)$, že údaj bude chybět, tj.:

$$log\ (OR) = log\ \left(\frac{P(\boldsymbol{M}=1)}{1-P(\boldsymbol{M}=1)}\right) \tag{32}$$

Náhodný mechanismus chybění s vícenásobnými diskriminačními hodnotami je přímým rozšířením modelu s jednoduchou rozhodovací hodnotou.

Náhodný mechanismus chybění MAR s rozhodovacími percentily

Použití percentilů k rozhodování je rozšířením modelu s vícenásobnou rozhodovací hodnotou. V percentilové metodě závisí pravděpodobnost, že u zpravodajské jednotky bude hodnota chybět, na percentilovém pořadí v prediktoru chybějících dat $\boldsymbol{Y}_2$. Proto na percentilovou metodu lze pohlížet jako na metodu s vícenásobnými rozhodovacími hodnotami, kde každá zpravodajská jednotka má svůj vlastní hraniční bod na základě percentilového pořadí hodnot prediktorové proměnné.

Nechť pravděpodobnost, že zpravodajská jednotka bude mít chybějící hodnotu v proměnné $\boldsymbol{Y}_1$, bude závislá na percentilovém pořadí hodnot prediktoru chybějících dat, tj. proměnné $\boldsymbol{Y}_2$. Nechť pravděpodobnost, že zpravodajská jednotka bude mít chybějící hodnotu v proměnné $\boldsymbol{Y}_1$ bude závislá na percentilovém pořadí hodnot prediktoru chybějících dat, tj. proměnné $\boldsymbol{Y}_2$. Nechť náhodná proměnná $\boldsymbol{M}$ bude v úloze indikátoru chybějících dat. Bude-li přímý vztah mezi indikátorem chybějících údajů a prediktorem chybějících údajů $\boldsymbol{Y}_2$, potom pravděpodobnostní pravidlo pro chybějící údaj bude: jestliže hodnota v závislé proměnné bude odpovídat k-tému percentilu proměnné $\boldsymbol{Y}_2$, potom pravděpodobnost, že údaj bude nevyplněn v proměnné $\boldsymbol{Y}_1$, odpovídá $k\%$ pravděpodobnosti chybění údaje v proměnné $\boldsymbol{Y}_1$, tj. $P(\boldsymbol{M} = 1|\boldsymbol{Y}_2 = q_k) = k/100$, kde q_k je hodnota proměnné $\boldsymbol{Y}_2$ korespondující k-tému percentilu. Bude-li nepřímý vztah mezi indikátorem chybějících údajů a prediktorem chybějících údajů, potom pravděpodobnostní pravidlo pro chybějící údaj bude: jestliže hodnota v závislé proměnné bude odpovídat k-tému percentilu proměnné $\boldsymbol{Y}_2$, potom pravděpodobnost, že údaj bude nevyplněn v proměnné $\boldsymbol{Y}_1$, odpovídá $(100 - k)\%$ pravděpodobnosti chybění údaje v proměnné $\boldsymbol{Y}_1$, tj. $P(\boldsymbol{M} = 1|\boldsymbol{Y}_2 = q_k) = 1 - k/100$, kde q_k je hodnota proměnné $\boldsymbol{Y}_2$ korespondující k-tému percentilu. Uvedená pravděpodobnostní pravidla chybění jsou jedinými možnými pravděpodobnostními pravidly spojenými s percentilovými hodnotami. Jiné parametry než výše uvedené, se při aplikaci náhodného mechanismu chybění s percentilovými hodnotami nepoužívají.

Aby bylo možné vypočítat pravděpodobnost chybějících dat, je potřebné stanovit pravděpodobnostní rozdělení percentilové řady hodnot proměnné $\boldsymbol{Y}_2$. K tomu lze využít empirickou distribuční funkci $\boldsymbol{F}(\cdot)$ proměnné $\boldsymbol{Y}_2$, která mapuje hodnoty proměnné $\boldsymbol{Y}_2$. do řady percentilů. Percentilové hodnoty proměnné $\boldsymbol{Y}_2$ proto mají standardní rovnoměrné rozdělení, tj.:

$$F(Y_2) \sim Unif(0,1). \tag{33}$$

Toho důsledkem je, že očekávaná percentilové pořadí zpravodajské jednotky je $50.$ percentil, a tedy pravděpodobnost chybějících dat je vždy 50 %, tj. $P(\boldsymbol{M} = 1) = 0{,}5$. Podobně jako u předešlých modelů počet zpravodajských jednotek (případů) s chybějícími údaji v proměnné $\boldsymbol{Y}_1$ je náhodná proměnná K podléhající binomickému rozdělení (předpokládá se nezávislost mezi vznikem chybějících údajů), tj. $K \sim Bin(n, 0{,}5)$, kde n je celkový počet hodnot. Očekávané procento chybějících údajů v proměnné $\boldsymbol{Y}_1$ ve výběrovém souboru j:

$$E\left(\frac{K}{n}\right) = 0{,}5, \tag{34}$$

a rozptyl pro odhadované procento (34) chybějících údajů v proměnné $\boldsymbol{Y}_1$ je:

$$Var\left(\frac{K}{n}\right) = \frac{0{,}5(1-0{,}5)}{n} = \frac{0{,}25}{n}. \tag{35}$$

Očekávaný počet jedinečných vzorů chybějících dat pro náhodný mechanismus chybění MAR s percentilovými hodnotami se podobně jako v (9) určí rovnicí

$$E(D) = 2 - \pi_{miss}^{n} - (1 - \pi_{miss})^{n} = 2 - 0{,}5^{n} - (1 - 0{,}5)^{n}. \quad (36)$$

U percentilové metody se nedá měnit síla závislosti mezi indikátorem chybějících dat $\boldsymbol{M}$ a prediktorovou proměnou $\boldsymbol{Y}_2$. Způsobem, jak kvantifikovat sílu závislosti mezi indikátorem chybějících dat a proměnnou $\boldsymbol{Y}_2$ jako prediktorovou proměnnou vytvořené percentilovými hodnotami, je najít model logistické regrese, který při dostatečně velkém n aproximuje pravděpodobnostní pravidlo chybějících dat [14], tj.:

$$log\left(\frac{P(\boldsymbol{M}=1)}{1-P(\boldsymbol{M}=1)}\right) = 1{,}70 \cdot \boldsymbol{Y}_2 \quad (37)$$

Náhodný mechanismus chybění MAR jako model logistické regrese

Jak je patrné z rovnic (22), (32) a (37), pravděpodobnostní pravidla chybějících údajů uvedená v předešlých podkapitolách se dají přeformulovat do tvaru modelů logistické regrese. Modeluje-li se náhodný mechanismus chybění přímo pomocí logistické regrese, lze považovat logistický regresní model přímo jako pravděpodobnostní pravidlo chybějících údajů a populační regresní koeficienty asociované s modelem jako parametry tohoto pravděpodobnostního pravidla.

Jestliže pravděpodobnost chybějící hodnoty v proměnné $\boldsymbol{Y}_1$ i-té zpravodajské jednotky je vztažena k prediktoru chybějících hodnot (proměnné $\boldsymbol{Y}_2$), potom regresní logistický model pro i-tou zpravodajskou jednotku je definován jako:

$$log\left(\frac{P(M_i = 1|y_{2,i})}{1-P(M_i = 1|y_{2,i})}\right) = \beta_0 + \beta_1 \cdot y_{2,i}, \quad (38)$$

kde M_i je indikátor chybějící hodnoty a $y_{2,i}$ je hodnota proměnné $\boldsymbol{Y}_2$ u i-té zpravodajské jednotky. Parametry asociované s pravděpodobnostním pravidlem chybějících údajů jsou β_0 a β_1. V závislosti na hodnotě proměnné $\boldsymbol{Y}_2$ je pravděpodobnost, že údaj v proměnné $\boldsymbol{Y}_2$ u i -té zpravodajské jednotky chybí, definována rovnicí:

$$P\left(M_i = 1\middle|y_{2,i}\right) = \frac{1}{1+e^{-\beta_0-\beta_1 y_{2,i}}} \quad (39)$$

Pro velké velikosti výběrového souboru (cca 10 000 zpravodajských jednotek) lze odhadnout očekávané procento chybějících hodnot průměrem pravděpodobností ve tvaru:

$$\pi_{miss} = \frac{1}{n}\sum_{i=1}^{n} P\left(M_i = 1\middle|y_{2,i}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+e^{-\beta_0-\beta_1 y_{2,i}}}. \quad (40)$$

Z hlediska síly závislosti mezi indikátorem chybějících údajů $\boldsymbol{M}$ a prediktorovou proměnou $\boldsymbol{Y}_2$ vyšší hodnoty β_1 indikují silnější závislost. Logistický regresní model však nelze odhadnout, když $P\left(M_i = 1\middle|y_{2,i}\right) = 1$ nebo $P\left(M_i = 0\middle|y_{2,i}\right) = 1$, protože logaritmus šance indikátoru $\boldsymbol{M}$ není v těchto případech definován nebo se rovná nekonečnu. Výhodou logistické regrese je, že pravděpodobnost chybějících hodnot se postupně mění s tím, jak se mění hodnota prediktoru chybějících dat, což vytváří

realističtější situaci vzhledem k metodám popsaným v předešlých podkapitolách. Nevýhodou metody logistické regrese je však to, že neumožňuje nastavit příliš silnou závislost mezi indikátorem chybějících dat $\boldsymbol{M}$ a prediktorovou proměnnou $\boldsymbol{Y}_2$.

Statistické modely nenáhodného mechanismu chybění (NMAR)

Nenáhodný mechanismus chybějících dat (NMAR) se vyznačuje tím, že výskyt chybějících hodnot závisí na hodnotách y_{miss}, které nejsou pozorované. Nenáhodný mechanismus chybění dat představuje podmíněné rozdělení pravděpodobnosti indikátorové proměnné $\boldsymbol{M}$ vzhledem k pozorované proměnné $\boldsymbol{Y}$ určené neznámými parametry φ, tj. $f(m|y_{obs}, y_{miss}; \varphi)$.

Modelovat mechanismus chybějících hodnot umožňují dva rozdílné přístupy určené právě pro data, která nechybí náhodně.

- **První přístup představují tzv. modely výběru (selection models).** U těchto modelů se v prvním kroku specifikuje rozdělení potenciálně kompletních dat (tzn. datové matice $\boldsymbol{Y}$, která je složená z pozorovaných a chybějících hodnot). V dalším kroku se podle [11] specifikuje model [4], podle kterého závisí výskyt chybějících hodnot na datové matici $\boldsymbol{Y}$.
- **Druhý přístup tvoří tzv. modely smíšených vzorů (pattern-mixture models**). U tohoto přístupu jsou jednotky rozdělené do skupin podle jedinečných vzorů chybějících hodnot. Následně se v každé ze skupin provádí statistická analýza [2]. Použitý pojem „smíšené vzory“ má indikovat, že výsledné marginální rozdělení datové matice $\boldsymbol{Y}$ je směsí několika rozdělení [4].

Podrobnější analýza nenáhodného mechanismu chybění v neúplných datech přesahuje rozsah tohoto článku. Detailnější informace o mechanismu NMAR lze dohledat například v [5] nebo v [2].

7. UKÁZKA STATISTICKÝCH MODELŮ CHYBĚJÍCÍCH DAT

Data použitá k ukázce statistických modelů chybějících údajů

Pro účely statistického modelování chybějících údajů byla vygenerována umělá data. Jedná se o datový vzorek se 100 hodnotami v rozsahu $-10{,}4$ až $14{,}1$ u proměnné $\boldsymbol{Y}_1$, pro které jsou jednotlivé modely chybění dat modelovány, a v rozsahu 1 až 100 u prediktorové proměnné $\boldsymbol{Y}_2$. Data jsou generována z normálního rozdělení $N(0, 5)$. Rozdělení pravděpodobnosti π_{miss} chybějících hodnot v proměnné $\boldsymbol{Y}_1$ je modelováno pomocí generátorů náhodné proměnné s rozdělením pravděpodobnosti pro jednotlivé typy mechanismu chybění. V rámci modelování mechanismu náhodného chybění MAR bude pro prediktorovou proměnnou $\boldsymbol{Y}_2$ použito pravděpodobnostní pravidlo s rovnoměrným rozdělením pravděpodobnosti. Pro hodnocení statistických modelů se budou posuzovat zejména momentové charakteristiky a statistické míry sloužící pro měření závislosti náhodných proměnných. Statistické modely chybějících dat a jejich měření (včetně statistických grafů) bylo realizováno pomocí funkcí a procedur statistického systému SAS. Nejdůležitější charakteristiky generovaného vzorku dat pro statistickou analýzu chybějících údajů ilustruje obrázek č. 5.

Obrázek č. 5: Rozdělení generovaných dat (steam-and-leaf graf) bez chybějících údajů a jeho nejdůležitější statistické charakteristiky

-10	4
-9	2
-8	2
-7	9840
-6	874
-5	985
-4	88875411
-3	974400
-2	766211
-1	764310
0	8873301379
1	014556678899
2	112568
3	069
4	1112466899
5	08999
6	0112348
7	5
9	01
10	589
11	0
14	0

Úplná data

Moments			
N	100	Sum Weights	100
Mean	4,64	Sum Observations	464
Std Deviation	3,17031688	Variance	10,0509091
Skewness	-0,0914414	Kurtosis	-1,4556078
Uncorrected SS	3148	Corrected SS	995,04
Coeff Variation	68,3257947	Std Error Mean	0,31703169

Basic Statistical Measures			
Location		Variability	
Mean	4,640000	Std Deviation	3,17032
Median	5,000000	Variance	10,05091
Mode	1,000000	Range	9,00000
		Interquartile Range	7,00000

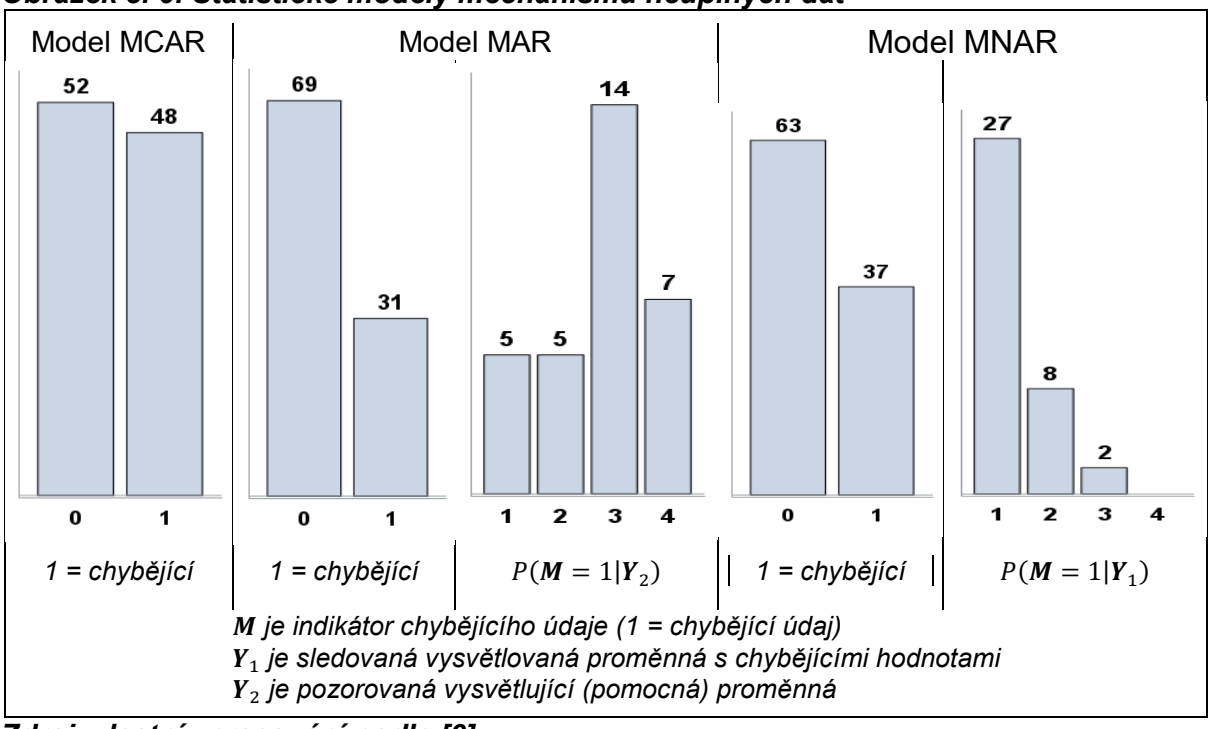
Note: s the smallest of 2 modes with a count of 15.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	14,63576	Pr > |t|	<.0001
Sign	M	44	Pr >= |M|	<.0001
Signed Rank	S	1958	Pr >= |S|	<.0001

Zdroj: vlastní zpracování podle [6]

Na levé straně obrázku č. 5 je tzv. stem-and-leaf graf[4], který vizualizuje rozdělení generovaných hodnot. Na pravé straně obrázku č. 5 jsou zaznamenány nejdůležitější charakteristiky vzorku generovaných dat.

Obrázek č. 6: Statistické modely mechanismů neúplných dat



$\boldsymbol{M}$ *je indikátor chybějícího údaje (1 = chybějící údaj)*
$\boldsymbol{Y}_1$ *je sledovaná vysvětlovaná proměnná s chybějícími hodnotami*
$\boldsymbol{Y}_2$ *je pozorovaná vysvětlující (pomocná) proměnná*

Zdroj: vlastní zpracování podle [6]

[4] *Steam-and-leaf graf je statistický graf podobný histogramu, data jsou uspořádána do tzv. stonků (angl. stems) a listů (angl. leafs). Stonky mohou například být celé části reálných čísel a stonky obsahují číslice, které patří do desetinného rozvoje odpovídajících reálných čísel.*

Princip činnosti jednotlivých modelů neúplných dat reprezentuje obrázek č. 6. Zcela náhodný model neúplných dat (označený MCAR) vykázal ze 100 generovaných hodnot 48 hodnot chybějících a 52 pozorovaných hodnot. U tohoto modelu chybějící hodnoty jsou zcela nezávislé jak od ostatních zjišťovaných proměnných, tak i od rozdělení hodnot v samotné sledované proměnné s chybějícími daty. Jedná se proto o tzv. ignorovatelný mechanismus chybění, který má nejmenší vliv na odhadované hodnoty. U modelu neúplných dat MAR byl zjištěn menší počet chybějících hodnot. Ze 100 generovaných hodnot bylo 69 pozorovaných a 31 chybějících (indikátor chybějících hodnot v grafu má hodnotu 1). U modelu MAR jsou chybějící hodnoty závislé na prediktorové proměnné $\boldsymbol{Y}_2$, na jejímž rozdělení výskyt chyb záleží. V případě použitého statistického modelu chybění MAR jako prediktorová proměnná byla použita proměnná $\boldsymbol{Y}_2$ s rovnoměrným rozdělením pravděpodobnosti. Zjištěné neúplné hodnoty u náhodného modelu chybění MAR byly rozděleny v souladu s předpokládaným rozdělením prediktorové (pomocné) proměnné $\boldsymbol{Y}_2$. Toto je patrné na obrázku č. 6 v pravém grafu statistického modelu neúplnosti dat MAR, kdy prediktorová proměnná $\boldsymbol{Y}_2$ byla rozdělena na 4 stejné intervaly (označené indikátory o hodnotách 1, 2, 3 a 4), každý se stejnou pravděpodobnosti vzniku chybějící hodnoty (ve skutečnosti v prvním intervalu hodnot proměnné $\boldsymbol{Y}_2$ se objevilo 5 chybějících hodnot proměnné $\boldsymbol{Y}_1$, ve druhém také 5, ve třetím intervalu 14 a v intervalu nejvyšších hodnot prediktorové proměnné $\boldsymbol{Y}_2$ bylo 7 chybějících hodnot $\boldsymbol{Y}_2$). Mechanismus chybění MAR patří také k tzv. ignorovatelným mechanismům, které lze při odhadech parametrů zanedbat. Při odhadech parametrů s tímto typem chybějících dat je však nutné brát do úvahy i význam prediktorové (pomocné) proměnné, na jejímž rozdělení výskyt chybějících údajů také záleží, tj. rozdělení chybějících dat je podmíněno rozdělením prediktorové (pomocné) proměnné. Nenáhodný mechanismus chybění (na obrázku č. 6 označený jako MNAR) již při statistické inferenci zanedbat nelze. Nenáhodný mechanismus chybění MNAR z obrázku č. 6 vykazuje 37 chybějících hodnot ve sledované proměnné $\boldsymbol{Y}_1$ v generovaném vzorku (viz pravý graf modelu MNAR). Uvedených 37 chybějících hodnot bylo rozdělených v intervalech nižších hodnot (v grafu označených jako interval 1, 2 a 3) sledované proměnné $\boldsymbol{Y}_1$ (v prvním intervalu hodnot bylo 27 chybějících hodnot, ve druhém intervalu bylo zaznamenáno 8 chybějících hodnot a ve třetím chyběly 2 hodnoty). V praxi je podle [12] mnohdy složité určit rozdělení proměnné $\boldsymbol{Y}_1$, jejíž některé hodnoty jsou chybějící. Chybějící data ve sledované proměnné $\boldsymbol{Y}_1$ závisí na neznámém rozdělení samotné sledované proměnné $\boldsymbol{Y}_1$, která je předmětem statistické inference (viz pravý graf modelu MNAR). Odhady z takto pořízených dat jsou proto nejvíce zatíženy chybami.

Obrázek č. 7: Momentové charakteristiky dat s různými mechanismy chybění

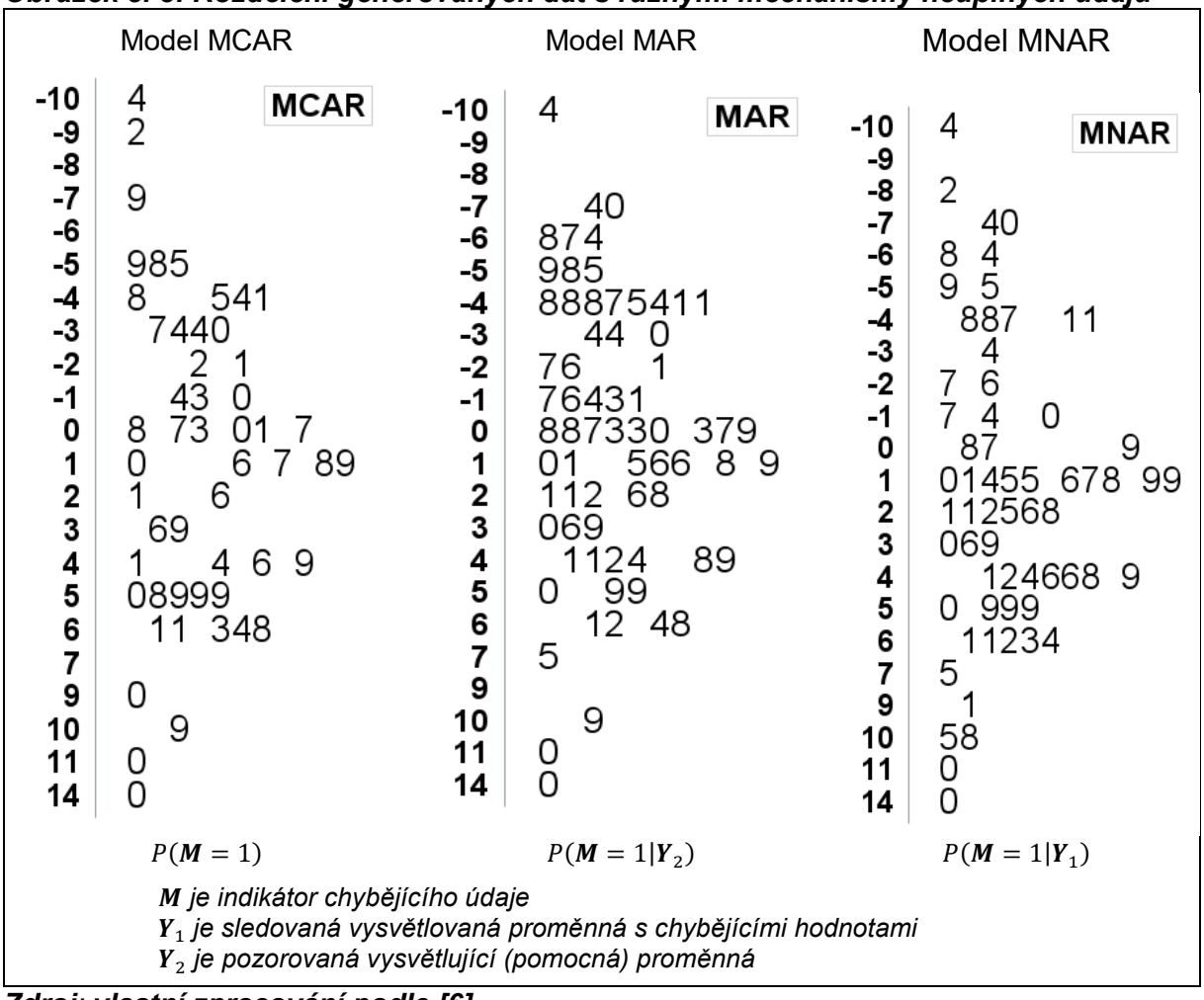
Moments		Mechanismus chybění		
	Úplná data	MCAR	MAR	MNAR
N	100	52	69	63
Mean	**4,640**	**4,596**	**4,666**	**4,730**
Std Deviation	3,170	3,309	3,118	3,122
Skewness	- 0,091	- 0,044	- 0,117	- 0,134
Uncorrected SS	3 148	1 657	2 164	2 014
Coeff Variation	68,326	72,001	66,826	66,007
Sum Weights	100	52	69	63
Sum Observations	464	239	322	298
Variance	10,051	10,951	9,725	9,784
Kurtosis	-1,456	-1,494	- 1,411	- 1,368
Corrected SS	995,04	558,52	661,33	604,41
Std Error Mean	**0,317**	**0,459**	**0,375**	**0,393**

Zdroj: vlastní zpracování podle [6]

Možné odchylky od skutečných hodnot odhadovaných parametrů ilustruje obrázek č. 7. Největší odchylkou od průměrné hodnoty zjištěné z úplných dat (bez chybějících údajů) vykazují data, v nichž byl modelován nenáhodný mechanismus chybění MNAR. Výběrový průměr úplných dat byl zjištěn ve výši 4,640, výběrový průměr dat s mechanismem chybění MNAR byl ve výši 4,730. Nejvyšší variabilita odhadů byla zjištěna pro data se zcela náhodným mechanismem chybění MCAR. Z hlediska statistické inference nad daty s chybějícími údaji vedlo použití znalosti o rozdělení prediktorové (pomocné) proměnné k nejlepším výsledkům, a to jak v přesnosti, tak i variabilitě odhadů. K průměru z úplných dat se nejvíce blíží výběrový průměr z dat s neúplností typu MAR, přičemž variabilita tohoto průměru MAR je nižší než u dat se zcela náhodným mechanismem chybění MCAR.

Jednotlivé mechanismy chybějících údajů v datech mají také vliv na rozdělení hodnot sledované proměnné. Jak působí jednotlivé typy chybových mechanismů na generovaná data, reprezentuje obrázek č. 8 pomocí tzv. stem-and-leaf grafů.

Obrázek č. 8: Rozdělení generovaných dat s různými mechanismy neúplných údajů



Zdroj*: *vlastní zpracování podle [6]

Pro reálnou praxi je také důležité zjištění, zda opravdu existuje souvislost mezi pravděpodobností $\pi_{miss} = P(\boldsymbol{M} = 1)$, že hodnota ve sledované (neúplné) proměnné $\boldsymbol{Y}_1$ bude chybět, a prediktorovou (vysvětlující) proměnnou $\boldsymbol{Y}_2$ v náhodném mechanismu chybění MAR, resp. že hodnota ve sledované (neúplné) proměnné $\boldsymbol{Y}_1$ bude chybět a současně závisí na neznámém rozdělení samotné sledované proměnné

Y_1, která je předmětem statistické inference, jak tomu je v nenáhodném mechanismu chybění MNAR. Za účelem zjištění míry závislosti lze využít standardní statistické testy, a to χ^2-kvadrát test dobré shody, resp. Fisherův exaktní test (při očekávaných hodnotách v buňkách kontingenční tabulky < 5) a další míry asociace. Výsledky testování hypotézy o nezávislosti náhodných proměnných a další koeficienty jsou uvedeny na obrázku č. 9. Na základě výsledků uvedených na obrázku č. 9 lze považovat pravděpodobnost, že hodnota ve sledované proměnné bude chybějící, za pravděpodobnost podmíněnou:

- u statistického modelu MAR na rozdělení hodnot prediktorové proměnné Y_2,
- u statistického modelu MNAR na neznámém rozdělení samotné sledované proměnné Y_1.

Obrázek č. 9: Míry statistické závislosti statistických modelů MAR a MNAR

	MAR			MNAR		
Statistic	**DF**	**Value**	**Prob**	**DF**	**Value**	**Prob**
Chi-Square	3	10,0808	0,0179	2	11,5802	0,0031
Likelihood Ratio Chi-Square	3	9,6504	0,0218	2	11,9903	0,0025
Mantel-Haenszel Chi-Square	1	1,8462	0,2060	1	8,3577	0,0038
Phi Coefficient		0,3175			0,3403	
Contingency Coefficient		0,3026			0,3222	
Cramer's V		0,3175			0,3403	
WARNING: 70% of the cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test.						
Fisher's Exact Test						
Table Probability (P)		<.0001			0,0002	
Pr <= P		0,0246			0,0022	

***Zdroj**: vlastní zpracování podle [6]*

8. ZÁVĚR

V současné době jsou neúplné (chybějící) údaje frekventovanou součástí souborů dat a jejich výskytu není přikládán dostatečný význam. Většina výzkumníků při obvyklých analýzách dat zaznamenaných ve statistickém šetření vypouští zpravodajské jednotky s chybějícími údaji z analýz bez ohledu na mechanismus, který chybějící data generuje. Nepříznivé důsledky takového postupu v odhadování parametrů populace je možné do určité míry kompenzovat využitím informací o populaci (pokud jsou k dispozici). Problematické je také i stanovení směrodatných chyb takto určených odhadů [9].

Metody analýzy neúplných dat používané k zajištění adekvátní statistické inference musí vycházet z konkrétní situace v zaznamenaných datech a jejich uplatnění je závislé na více faktorech. Obecně jsou metody analýzy neúplných dat podle [5] seskupeny do následujících kategorií, jejichž použití se vzájemně nevylučuje:

- **Postupy založené na jednotkách s kompletně vyplněnými odpověďmi**: Zpravodajské jednotky s neúplně vyplněnými údaji se ze statistických analýz vypouští a k dalšímu zpracování se použijí pouze jednotky s kompletně zjištěnými údaji. Tuto strategii, která se nazývá analýzou úplných případů, je snadné implementovat a může být uspokojivá s malým množstvím chybějících údajů. Může však vést k vážným zkreslením a obvykle není příliš efektivní, zejména při odhadování závěrů pro subpopulace.
- **Postupy založené na modifikaci vah pro jednotlivé funkce odhadu**: Do funkcí pro odhad populačních parametrů jsou použity nejen pravděpodobnostní váhy

(pravděpodobnosti zahrnutí jednotek), ale také i váhy odvozené z neodpovědí zpravodajských jednotek.

- **Imputace chybějících hodnot**: Chybějící hodnoty jsou vyplněny a výsledná doplněná data jsou analyzována standardními metodami. Mezi běžně používané postupy pro imputaci patří imputace hot deck, kdy se k imputaci používají jednotky s kompletními údaji; imputace průměrnou hodnotou, které vycházejí z průměrných hodnot kompletně vyplněných jednotek; a regresní imputace, kde se chybějící proměnné pro jednotku odhadují předpokládanými hodnotami z regrese na dostupných proměnných pro tuto jednotku.
- **Metody založené na statistických modelech**: Široká třída metod založených na definici statistického modelu pro úplná data vycházející z funkce věrohodnosti nebo aposteriorního rozdělení. Parametry modelu jsou odhadovány takovými metodami, jako je metoda maximální věrohodnosti anebo expektační-maximalizační algoritmy, apod.

LITERATURA

[1] AGRESTI, A. – KATERI, M.: Categorical data analysis. In: International Encyclopedia of Statistical Science. Springer, 2011, s. 206 – 208. ISBN 978-3-642-04897-5.
[2] ALLISON, P. D.: Missing Data. Thousand Oaks, CA: Sage. Sage University Papers Series. Quantitative Applications in the Social Sciences, 2001, č. 07-136. ISBN 0-7619-1672-5.
[3] ENDERS, C. K.: Applied Missing Data Analysis, Second Edition. New York: Guilford Press, 2022. 546 s. ISBN 978-1-462-54986-3.
[4] GRAHAM, J. W.: Missing data: Analysis and design. New York: Springer, 2012. ISBN 978-1-4614-4017-8.
[5] LITTLE, R. J. A.: Pattern-Mixture Models for Multivariate Incomplete Data. In: Journal of the American Statistical Association, 1993, č. 421, s. 125 – 134.
[6] LITTLE, R. J.A. – RUBIN, D., B.: Statistical Analysis with Missing Data, 3rd Edition. New York: John Wiley & Sons, Inc. 2019. ISBN 978-0-470-52679-8.
[7] McKNIGHT, P. E. – McKNIGHT, K. M. – SIDANI, S. – FIGUEREDO, A. J.: Missing data: a gentle introduction. New York: Guilford Press, 2007. 251 s. ISBN 978-1-59385-394-5.
[8] McLAWHORN, J. – GRIGSBY, T. J.: Missing Data Techniques and the Statistical Conclusion Validity of Survey-Based Alcohol and Drug Use Research Studies: A Review and Comment on Reproducibility. In: Journal of Drug Issues, 2019, č. 1, s. 44 – 56.
[9] NAKAGAWA, S.: Missing data: Mechanisms, methods, and messages. In: G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.). Ecological statistics: contemporary theory and application, 2015, s. 81 – 105.
[10] PECÁKOVÁ, I.: Problém chybějících dat v dotazníkových šetřeních. In: Acta Oeconomica Pragensia. 2014, č. 6, s. 66 – 78.
[11] RUBIN, D. A.: Inference and Missing Data. In: Biometrika, 1976, č. 3, s. 581 – 592.
[12] SCHAFER, J. L. – GRAHAM, J. W.: Missing data: Our view of the state of the art. In: Psychological Methods, 2002, č. 2, s. 147 – 177.
[13] van BUUREN, S.: Flexible Imputation of Missing Data, 2nd Edition. New York: CRC Press, 2018. 444 s. ISBN 978-1-138-58831-8.
[14] ZHANG, X.: How to generate missing data for simulation studies. In: The Quantitative Methods for Psychology, 2023, č. 2, s. 100 – 122.

RESUMÉ

Zejména pro analýzy metodami vícerozměrných statistik představují chybějící údaje problém. Ačkoliv neúplná data ve výběrovém souboru mohou být zastoupena v relativně malém procentu, může tato situace v zjištěných datech vyústit v relativně velmi malý soubor s kompletními údaji; zejména v případě, kdy u různých jednotek chybí hodnoty různých veličin. V běžné praxi výběrových zjišťování jednotky, u kterých byly zaznamenány nevyplněné hodnoty zjišťovaných ukazatelů, jsou převážně z dalších analýz vyloučeny. Vynechání jednotek z analýz může mít značné negativní dopady – snížení přesnosti odhadů a síly vykonávaných statistických testů a může vést až ke zkresleným výsledkům nevhodných k zobecňování na cílovou populaci.

I když mechanismus chybějících hodnot představuje nejdůležitější faktor, který má nejvýraznější vliv na úspěšnost rozmanitých metod práce s chybějícími hodnotami, není mnohdy jednoduché pochopit příčiny vzniku a dynamiku neúplných dat. Datoví analytici zpravidla nikdy nemohou s jistotou znát mechanismus, podle kterého analyzovaná data chybí. V rámci vykonávání odpovídající statistické analýzy dat za podmínek chybějících údajů se vždy vyplatí podrobně promyslet, který z výše uvedených mechanismů chybění je v dané situaci nejrealističtější. Pro adekvátní analýzu dat je dále vhodné hledat ve výběrovém souboru proměnné, které korelují s výskytem nevyplněných hodnot u proměnných vstupujících do procesu analýz. Při realizaci výběrových šetření je užitečné zavést takové postupy a mechanismy, které by alespoň z části vyloučily situace potenciálně vedoucí ke vzniku chybějících hodnot u klíčových proměnných, například včasnými kontrolami sbíraných dat, řešení otázek respondentů ke sběru dat, vhodnou organizací výběrového šetření atd. Je třeba zdůraznit, že právě předcházení vzniku chybějících hodnot může být tím nejlepším řešením problémů plynoucích z chybějících hodnot [13].

RESUME

Missing data is especially a problem for the analyses using multivariate statistical methods. Although the incomplete data in the random sample may be represented by a relatively small percentage, this situation may lead to a relatively very small complete data set in the surveyed data; especially when the values of different quantities are missing for different units. In the normal practice of sample surveys of these units, for which unfilled values of the surveyed indicators were recorded, they are mostly excluded from further analyses. Omission of units from the analyses can have significant negative impacts - reducing the accuracy of estimates and the power of the performed statistical tests - and can lead to biased results inappropriate for generalization to the target population.

Although the mechanism of missing values is the most important factor that affect the most significantly the success of various methods of working with missing values, it is often not easy to understand the causes and dynamics of incomplete data. As a rule, data analysts can never know with certainty the mechanism by which the analysed data is missing. As part of conducting an appropriate statistical analysis of data under conditions of missing data, it is always worth considering in detail which of the above missing data mechanisms is the most realistic in a given situation. For an adequate data analysis, it is also advisable to look for variables in

the sample set that correlate with the occurrence of unfilled values for the variables entering the analysis process. When conducting sample surveys, it is useful to establish such procedures and mechanisms that would at least partially exclude situations potentially leading to the emergence of the missing values for key variables, for example by timely checks of the collected data, solving respondent's questions about data collection, suitable organization of sample surveys, etc. It should be emphasized that the best solution to the issues arising from missing values seems to be the prevention of missing values.

PROFESIJNÝ ŽIVOTOPIS

***Ing. Roman Pavelka, PhD.**, v rokoch 1995 – 2010 pracoval v poradenskej spoločnosti Trexima, s. r. o. Na pozícii štatistik – analytik sa zaoberal najmä analýzami mzdových a personálnych údajov. Podieľal sa na tvorbe pravidelných štatistických prehľadov a reportov. Spolupracoval s akademickými pracoviskami, agentúrami i súkromnými subjektami na realizácii a vyhodnocovaní ad hoc štatistických výskumov. Oblasť jeho vedeckého záujmu predstavujú výberové zisťovania, odhady a štatistické modely. V rokoch 2012 až 2013 sa zúčastnil zahraničnej stáže vo Veľkej Británii. Od roku 2013 pôsobil v Národnom ústave certifikovaných meraní vzdelávania (NÚCEM), kde zaisťoval štatistické vyhodnocovanie výsledkov testovania žiakov a študentov. Od roku 2015 pracuje v odbore metód štatistických zisťovaní Štatistického úradu SR*

KONTAKT

roman.pavelka@statistics.sk