SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

4/2024
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: 2

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

Strany/Pages: 21 – 41

Dátum vydania/Publication date: 15. október 2024/October 15, 2024

Roman PAVELKA
Štatistický úrad Slovenskej republiky

IMPUTACE CHYBĚJÍCÍCH DAT POMOCÍ BAYESOVSKÉHO MODELOVÁNÍ

MISSING DATA IMPUTATION USING BAYESIAN DATA MODELLING

ABSTRAKT
Proč jsou chybějící údaje problém? Protože běžné statistické metody a software předpokládají, že všechny hodnoty u všech proměnných v matici dat jsou pozorovány za všechny jednotky účastné ve statistickém zjišťování. Výchozí metodou řešení neodpovědí prakticky u všech statistických softwarů je prosté vymazání případů s chybějícími údaji ukazatelů, které jsou předmětem zájmu. Nejzřetelnější nevýhodou vymazání seznamu (jednotek) je to, že se často vymaže velká část vzorku sesbíraných statistických dat. Odstranění sesbíraných dat, které nejsou vhodné k dalšímu statistickému zpracování, může vést k vážné ztrátě statistické síly analýz. Výzkumníci pochopitelně neradi vyřazují údaje, jejichž sběru věnovali mnoho času, peněz a úsilí, a proto se staly populárními různé metod „záchrany“ případů s chybějícími údaji. Moderní metodou pro doplnění neúplných dat se v posledních několika desetiletích stává bayesovská inference. Bayesovská pravděpodobnost a statistika je mnohem více než všeobecně známý Bayesův vzorec a jeho občasné použití v ukázkových či ilustrativně orientovaných příkladech při výkladu operací s pravděpodobnostmi náhodných jevů. Bayesův pravděpodobnostní vzorec (nazývaný také zákon o inverzní pravděpodobnosti) se především používá v souvislosti s úsudky o neznámém modelu na základě známých dat. Toto dává možnosti pro použití Bayesova vzorce při imputování nepozorovaných dat (neznámý model) na základě dat zjištěných.

ABSTRACT
Why are missing data a problem? Because common statistical methods and software assume that all values for all variables in the data matrix are observed for all units participating in the statistical survey. The default method of dealing with nonresponse in all statistical software is to simply delete cases with missing data for the indicators of interest. The most obvious disadvantage of list (unit) removal is that it often deletes a large portion of the sample of collected statistical data. Removal of the collected data not suitable for further statistical processing can lead to a serious loss of statistical power of the analyses. Researchers are understandably reluctant to discard data that they have spent a lot of time, money and effort collecting, so various methods of 'rescuing' cases with missing data have become popular. The Bayesian inference has become a modern method for completing incomplete data over the last few decades. The Bayesian probability and statistics is far more than the well-known Bayesian formula and its occasional use in demonstrative or illustrative examples in explaining operations with probabilities of random phenomena. The Bayesian' probability formula (also called the law of inverse probability) is primarily used in the context of making judgments about an unknown model based on known data. This provides opportunities for using the Bayesian' formula in imputing unobserved data (unknown model) based on observed data.

KLÍČOVÁ SLOVA
bayesovská inference, inverzní pravděpodobnost, mechanismus chybění, imputace dat

KEY WORDS
Bayesian inference, inverse probability, mechanism of missingness, data imputation

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, periodicitě a složitosti 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?

Neúplné, resp. chybějící informace také narušují zjišťované údaje, vnášejí do nich prvek zkreslení, znehodnocují výsledky a závěry, činí je nevhodnými pro analýzy standardními statistickými metodami, a způsobují, že taková data úřady mohou kvůli odchylkám odmítnout. Náprava takto zjišťovaných dat nespočívá pouze v zanedbání jednotek s neúplnými či chybějícími informacemi. Pokud se tyto jednotky s chybějícími údaji jednoduše vyloučí, ovlivní to zejména statistickou sílu realizovaného zjišťování. Zároveň je pravděpodobné, že zpravodajské jednotky s nevyplněnými hodnotami mohou být jednotkami s extrémními anebo odlehlými hodnotami. Vyloučení těchto jednotek povede k podhodnocení variability, a tedy k zúžení intervalu spolehlivosti. Chybějící údaje samy o sobě představují různé problémy, například:

- absence údajů snižuje statistickou sílu, která se vztahuje k pravděpodobnosti, že test zamítne nulovou hypotézu, pokud je nepravdivá,
- ztracené údaje mohou způsobit zkreslení odhadu parametrů,
- chybějící údaje mohou snížit reprezentativnost vzorku a
- může zkomplikovat analýzu zjišťování nebo studie.

Jako chybějící údaje jsou považovány údaje, které v datové matici sledovaných hodnot chybí u některých (případně u všech) proměnných a u některých (případně u všech) respondentů (případů, resp. zpravodajských jednotek). Tyto chybějící údaje z datové matice jsou problematické především z toho důvodu, že většina současných analytických programů a nástrojů předpokládá datovou matici úplnou. Proto může být tento postup zpracování dat nevhodný a výsledné statistické úsudky mohou být více či méně chybné.

2. POUŽITÉ NÁHODNÉ VELIČINY A JEJICH ZNAČENÍ

Na rozdíl od většiny úloh matematické statistiky, která se soustředí na odhad hypotetických neznámých parametrů (např. průměrné hodnoty hypotetického normálního rozdělení, které vygenerovala data), metodika výběrových šetření se soustředí na odhad neznámých pozorovatelných veličin, jako je např. průměrný příjem rodin v určité konečné populaci. Výraz parametr slouží k označení hypotetických neznámých a nepoznatelných údajů. Nepozorované hodnoty v konečné populaci se považují za chybějící údaje, tj. chybějící hodnoty pozorovatelných veličin. Pozorovatelné veličiny, které jsou předmětem zájmu v konečné populaci, budou pro

účely statistického šetření uspořádány do jednotek (řádky) × proměnné (sloupce). I když snahou v každém statistickém šetření je sledovat všechny hodnoty v této matici, zpravidla v každém šetření populace nejsou všechny hodnoty matice pozorovány.

V konečné populaci složené z $\boldsymbol{N}$ jednotek jsou zpravidla definovány dva druhy proměnných. První druh zahrnuje proměnné, které popisují charakteristiky jednotek zajímavé pro účely statistického zjišťování: a to plně pozorovatelné vysvětlující proměnné (kovariáty) $\boldsymbol{X}$ a vysvětlované (závislé na vysvětlujících proměnných) proměnné $\boldsymbol{Y}$, které jsou předmětem výzkumného zájmu. Druhý druh proměnných zahrnuje indikátorové proměnné – indikátory výběru $\boldsymbol{I}$ a indikátory neodpovědí $\boldsymbol{R}$, které jsou potřebné k popisu, které hodnoty jsou pozorované a které nepozorované. Indikátorové proměnné $\boldsymbol{I}$ a $\boldsymbol{R}$ také pomáhají vyvodit správné induktivní závěry o parametrech populace. Matici proměnných v konečné populaci zobrazuje tabulka č. 1.

Tabulka č. 1: Matice proměnných v konečné populaci o N jednotkách

		Kovariáty *X*			**Závisle proměnné Y**			**Indikátory výběru *I***			**Indikátory odpovědí *R***		
		1	$\cdots$	q	1	$\cdots$	p	1	$\cdots$	p	1	$\cdots$	p
Statistické jednotky v konečné populaci	1	X_{11}	$\cdots$	X_{1q}	Y_{11}	$\cdots$	Y_{1p}	I_{11}	$\cdots$	I_{1p}	R_{11}	$\cdots$	R_{1p}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
	i	X_{i1}	$\cdots$	X_{iq}	Y_{i1}	$\cdots$	Y_{ip}	I_{i1}	$\cdots$	I_{ip}	R_{i1}	$\cdots$	R_{ip}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
	N	X_{N1}	$\cdots$	X_{Nq}	Y_{N1}	$\cdots$	Y_{Np}	I_{N1}	$\cdots$	I_{Np}	R_{N1}	$\cdots$	R_{Np}

Zdroj: [1, s. 29]

$\boldsymbol{X}$ jsou plně pozorované kovariáty, jako jsou indikátor straty nebo velikost jednotky, a jsou zaznamenané pro všechny jednotky $\boldsymbol{N}$ v populaci. Existuje-li více než 1 plně pozorovatelná kovariáta, bude $\boldsymbol{X}_i$ řádkový vektor. $\boldsymbol{Y}$ vyjadřují proměnné, jejichž hodnoty nejsou známy pro všechny jednotky v populaci. Existuje-li více než 1 závislá proměnná, bude $\boldsymbol{Y}_i$ řádkový vektor. Obvykle bude $\boldsymbol{Y}$ zahrnovat pouze ty proměnné, které jsou primárně předmětem zájmu šetření, například jako je příjem jednotlivců, objemy tržeb, obratu nebo přidané hodnoty. Indikátor $\boldsymbol{I}$ bude ukazatelem pro zahrnutí/nezahrnutí do šetření. V případě pouze jedné závislé proměnné $\boldsymbol{Y}_i$ je ukazatel $\boldsymbol{I}$ binární, přičemž $\boldsymbol{I}_i = 1$ značí že *i*-tá jednotka je do šetření zahrnuta (tj. byl učiněn pokus o zaznamenání proměnné $\boldsymbol{Y}_i$) a $\boldsymbol{I}_i = 0$ je *i*-tá jednotka ze šetření vyloučena (tj. nebyl učiněn žádný pokus o zaznamenání $\boldsymbol{Y}_i$). Reprezentuje-li $\boldsymbol{Y}_i$ více než 1 proměnnou, např. p proměnných, $\boldsymbol{I}_i$ je také vektor s p prvky; prvek I_{ij} indikuje, zda byl učiněn pokus o zaznamenání hodnoty Y_{ij} ($I_{ij} = 1$) nebo nebyl ($I_{ij} = 0$). Předpokládá se, že matice indikátorů je plně pozorovaná. Indikátor odpovědí $\boldsymbol{R}$ je indikátorem pro odpověď či neodpověď jednotky. V případě pouze *1* závislé proměnné, je indikátorová proměnná $\boldsymbol{R}_i$ binární, přičemž $\boldsymbol{R}_i = 1$ indikuje že *i*-tá jednotka zaznamenala hodnotu závislé proměnné (tj. hodnota $\boldsymbol{Y}_i$ je pozorována, pokud je učiněn pokus o zaznamenání odpovědi $\boldsymbol{Y}_i$ v případě, že *i*-tá jednotka bude zahrnuta do šetření) a $\boldsymbol{R}_i = 0$ indikuje, že hodnota závislé proměnné *i*-té jednotky pozorována není (samozřejmě v případě zahrnutí *i*-té jednotky do šetření). Pokud $\boldsymbol{Y}_i$ vektor závislých proměnných obsahuje p proměnných, $\boldsymbol{R}_i$ je také vektorem s p proměnnými, kde prvek R_{ij} indikuje odpověď či neodpověď proměnné Y_{ij}. Předpokládá se, že indikátor R_{ij} je znám, je-li indikátor zahrnutí $I_{ij} = 1$. Je-li indikátor zahrnutí *i*-té jednotky

$\boldsymbol{I}_i = 0$, je indikátor odpovědi $\boldsymbol{R}_i$ u *i*-té jednotky nedefinován. Ačkoliv výběrová šetření mohou obecně být vícestupňová se složitějšími návrhy výběru, bude pro účely tohoto článku prezentováno výběrové šetření jednostupňové s jednoduchým náhodným výběrem. V případě ostatních typů statistických šetření se následující matematické vztahy musí příslušně upravit.

3. PRAVDĚPODOBNOSTNÍ SPECIFIKACE POUŽITÝCH PROMĚNNÝCH

Pravděpodobnostní specifikace pro indikátorové proměnné $\boldsymbol{I}$ a $\boldsymbol{R}$ jsou potřebné k tomu, aby bylo možné vyvození správných závěrů pro neznámé hodnoty v konečné populaci. Tyto pravděpodobnostní modely se nazývají „mechanismy“, a nikoliv „modely“, aby mohly být odlišeny od obvyklejších statistických modelů týkajících se rozdělení proměnných $\boldsymbol{X}$, $\boldsymbol{Y}$ [12].

Mechanismus výběru je podle [12] specifikován jako podmíněná pravděpodobnost daná výrazem $Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R})$. Všechny formální statistické metody, které se používají k vyvozování závěrů o populaci, vyžadují explicitní nebo implicitní popis mechanismu výběru. Mechanismus výběru je považován za pravděpodobnostní, pokud každé hodnotě závislé proměnné Y_{ij} přiřazuje kladnou hodnotu pravděpodobnosti, že bude zahrnuta do výběrového souboru, tj.

$$Pr\left(I_{ij} = 1 \middle| \boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}\right) > 0 \qquad \text{pro všechny } I_{ij}, \tag{1}$$

Je-li pro *j*-tou hodnotu závislé proměnné $\boldsymbol{Y}$ *i*-té jednotky pravděpodobnost zahrnutí do výběrového souboru rovná 0, tj. $Pr\left(I_{ij} = 1 \middle| \boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}\right) = 0$, je mechanismus výběru považován jako nepravděpodobnostní. Standardní výběrové techniky jsou navrženy tak, aby umožňovaly jednu obecně přijatelnou specifikaci pro $Pr\left(I_{ij} = 1 \middle| \boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}\right)$. Často se jedná o pravděpodobnostní mechanismus výběru, jak je definován v (1) a (2).

$$Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}) = Pr(\boldsymbol{I}|\boldsymbol{X}) \qquad \text{pro všechny kombinace } (\boldsymbol{I},\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}). \tag{2}$$

Z formulací (1) a (2) jednoduše vyplývá, že:

- mechanismus výběru musí zajistit, aby každá hodnota $\boldsymbol{Y}$ z populace měla šanci být zahrnuta do výběrového souboru (viz 1) a
- mechanismus výběru vzorků může k výběru použít pozorované proměnné $\boldsymbol{X}$, jako jsou ukazatele strat, velikosti, apod. (viz 2).

Pro techniky výběru ale nemohou být použity hodnoty $\boldsymbol{Y}$ nebo $\boldsymbol{R}$, protože nejsou v okamžiku výběru známé.

Příkladem mechanismu výběru nepravděpodobnostního šetření je telefonický průzkum v populaci, která však zahrnuje i některé domácnosti bez telefonů. Pokud proměnná $\boldsymbol{X}_i$ je počet telefonů v *i*-té domácnosti a vyšetřeny mají být všechny jednotky, mechanismus výběru je nepravděpodobnostní a je daný rovnicí

$$Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}) = \begin{cases} 1 & if\ I = 1\ když\ \boldsymbol{X}_i > \boldsymbol{0} \text{ a } I = 0\ když\ \boldsymbol{X}_i = \boldsymbol{0} \text{ a} \\ 0 & jinak \end{cases}. \tag{3}$$

Mechanismus response dat lze podobně jako v předešlém případě specifikovat jako podmíněnou pravděpodobnost. Na rozdíl od mechanismu výběru, mechanismus odpovědí nezávisí na způsobu výběru respondentů. Proto lze tento mechanismus vyjádřit jako pravděpodobnost podmíněnou vysvětlovanými anebo vysvětlujícími proměnnými, tj. $Pr(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{Y})$. Pokud každé hodnotě závislé proměnné Y_{ij} přiřazuje mechanismus odpovědí kladnou hodnotu pravděpodobnosti, že bude hodnota proměnné Y_{ij} pozorovaná, jedná se o pravděpodobnostní mechanismus odpovědí, tj.

$$Pr\big(R_{ij}=1\big|\boldsymbol{X},\boldsymbol{Y}\big)>0 \qquad \text{pro všechny } i,j. \tag{4}$$

Předpoklady o specifických formách mechanismu odpovědí mohou mít zásadní význam pro vhodné úpravy při vyvozování závěrů o neodpovědích.

Jelikož dvojice proměnných $(\boldsymbol{X},\boldsymbol{Y})$ je matice náhodných proměnných (viz tabulka č. 1) o rozměrech $N\times(p+q)$, je nutné specifikovat sdružené rozdělení $Pr(\boldsymbol{X},\boldsymbol{Y})$. Nechť pro $(\boldsymbol{X},\boldsymbol{Y})$ existuje rozdělení $f(\boldsymbol{X},\boldsymbol{Y})$, jehož předpokladem je vzájemná zaměnitelnost $(\boldsymbol{X}_i,\boldsymbol{Y}_i),\ i=1,\dots,N$ tj. řádků matice z tabulky č. 1[1]. Potom podle de Finettiho věty [3] vyplývá, že $Pr(\boldsymbol{X},\boldsymbol{Y})$ lze zapsat ve tvaru, kde při daném vektoru parametrů $\boldsymbol{\theta}$ s marginální (apriorní) hustotou $Pr(\boldsymbol{\theta})$, jsou $(\boldsymbol{X}_i,\boldsymbol{Y}_i), i=1,\dots,N$, nezávisle a identicky rozděleny se společným rozdělením $f_{YX}(\boldsymbol{X}_i,\boldsymbol{Y}_i|\boldsymbol{\theta})$. Potom podle de Finettiho věty lze pravděpodobnost $Pr(\boldsymbol{X},\boldsymbol{Y})$ zapsat ve tvaru:

$$Pr(\boldsymbol{X},\boldsymbol{Y})=\int_{\boldsymbol{\theta}}[\textstyle\prod_{i=1}^{N} f_{YX}(\boldsymbol{X}_i,\boldsymbol{Y}_i|\boldsymbol{\theta})]Pr(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{5}$$

Konkrétním příkladem použití (5) je jednoduchý normální model, ve kterém vystupuje závislá proměnná $\boldsymbol{Y}$ jako skalár (vektor s jedinou složkou, tj. $\boldsymbol{Y}=(Y_1,\dots,Y_N)^T$) bez vektoru vysvětlovaných proměnných $\boldsymbol{X}$. Pro každé $i=1,\dots,N$, je rozdělení vysvětlované proměnné identické a navzájem nezávislé rozdělení odpovídající normálnímu rozdělení $N(\mu,\sigma^2)$, každé s vektorem parametrů $\boldsymbol{\theta}=(\mu,\sigma^2)$ a rozdělením $\boldsymbol{\theta}$ úměrným σ^{-2}, tj.

$$Pr((Y_1,\dots,Y_N)^T|\mu,\sigma^2)=\textstyle\prod_{i=1}^{N}\frac{1}{\sqrt{(2\pi\sigma^2)}}exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right] \tag{6}$$

a

$$Pr(\mu,\sigma^2)\propto\sigma^{-2}. \tag{7}$$

Potom podmíněné rozdělení parametru μ za podmínky realizace parametru σ^2 a proměnné $\boldsymbol{Y}$, tj. $(\sigma^2,y_1,\dots,y_n)$, je $N(\bar{y},\sigma^2/n)$, kde $\bar{y}=\sum_{i=1}^{n}y_i/n$, tj.

$$Pr(\mu|\sigma^2,(y_1,\dots,y_n)^T)=\sqrt{\frac{n}{2\pi\sigma^2}}exp\left[-n\frac{(\mu-\bar{y})^2}{2\sigma^2}\right]. \tag{8}$$

Podmíněné rozdělení parametru σ^2 za podmínky realizace proměnné $(Y_1,\dots,Y_n)^T$ je $(n-1)s^2\chi_{n-1}^{-2}$, kde $s^2=\sum_1^n(y-\bar{y})^2/(n-1)$ a χ_{n-1}^{-2} je inverze rozdělení χ^2 s $(n-1)$ stupni volnosti.

---

[1] Předpokladem je náhodné přiřazení indexů jednotek, které musí být v (řádkových) indexech jednotek zaměnitelné a se stejnou pravděpodobností: $Pr(\boldsymbol{X},\boldsymbol{Y})=Pr\big(row-perm(\boldsymbol{X},\boldsymbol{Y})\big)$, kde $row-perm(\boldsymbol{X},\boldsymbol{Y})$ je libovolná permutace řádků $(\boldsymbol{X},\boldsymbol{Y})$.

Pro podmíněné rozdělení vektoru parametrů (μ, σ^2) za podmínky realizace proměnných $(Y_1, \dots, Y_n)^T$ potom platí

$$Pr(\mu, \sigma^2 | (y_1, \dots, y_n)^T) \propto Pr((Y_1, \dots, Y_n)^T | \mu, \sigma^2) Pr(\mu, \sigma^2). \quad (9)$$

4. BAYESOVSKÁ INFERENCE V PODMÍNKÁCH CHYBĚJÍCÍCH HODNOT

Nechť platí pro $inc = \{(i,j) | I_{ij} = 1\}$ tak, že $\boldsymbol{Y}_{inc}$ označuje složky proměnné $\boldsymbol{Y}$ zahrnuté do výběrového souboru a $\boldsymbol{R}_{inc}$ označuje složky vektoru $\boldsymbol{R}$ zahrnuté do stejného souboru; analogicky je definován výraz $exc = \{(i,j) | I_{ij} = 0\}$. Pro $\boldsymbol{Y}$ a $\boldsymbol{R}$ potom platí $\boldsymbol{Y} = (\boldsymbol{Y}_{inc}, \boldsymbol{Y}_{exc})^T$ a $\boldsymbol{R} = (\boldsymbol{R}_{inc}, \boldsymbol{R}_{exc})^T$. Jak $\boldsymbol{Y}_{exc}$ a $\boldsymbol{R}_{exc}$ jsou nepozorované, protože nejsou do výběrového souboru zahrnuty; $\boldsymbol{R}_{inc}$ je plně pozorována, ale $\boldsymbol{Y}_{inc}$ je plně pozorována pouze za podmínky neexistence neodpovědí.

Nechť platí pro $obs = \{(i,j) | I_{ij} = 1 \text{ a } R_{ij} = 1\}$ tak, že $\boldsymbol{Y}_{obs}$ označuje složky proměnné $\boldsymbol{Y}$ pozorované ve výběrovém souboru; každá složka má přiřazen indikátor $R_{ij} = 1$. Analogicky výraz $mis = \{(i,j) | I_{ij} = 1 \text{ a } R_{ij} = 0\}$ bude použit pro komponenty proměnné $\boldsymbol{Y}$, které jsou do výběrového souboru vybrané, ale nepozorované (chybějící) s přiřazenou hodnotou $R_{ij} = 0$, tj. $\boldsymbol{Y}_{mis}$ a platí, že $\boldsymbol{Y}_{inc} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})^T$. Dále nechť $nob = \{(i,j) | I_{ij} = 0 \text{ anebo } R_{ij} = 0\}$, kde $\boldsymbol{Y}_{nob} = (\boldsymbol{Y}_{exc}, \boldsymbol{Y}_{mis})^T$ vyjadřuje nepozorované složky vysvětlované proměnné $\boldsymbol{Y}$. Na základě tohoto značení lze složky proměnné $\boldsymbol{Y}$ rozdělit na složky pozorované $\boldsymbol{Y}_{obs}$ a složky nepozorované $\boldsymbol{Y}_{nob}$, tj. $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{nob})^T$.

Bayesovská inference pro populační parametr $Q(\boldsymbol{X}, \boldsymbol{Y})$ se odvozuje z jeho aposteriorního rozdělení, tj. jeho podmíněného rozdělení vzhledem k pozorovaným hodnotám $\boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{R}_{inc}, \boldsymbol{I}$ vypočteným podle zadaných modelů. Použije-li se výše zavedená notace, lze toto aposteriorní rozdělení zapsat jako $Pr(Q | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{R}_{inc}, \boldsymbol{I})$. Je-li například aposteriorní rozdělení skaláru Q normální o průměru $\hat{Q} = \hat{Q}(\boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{R}_{inc}, \boldsymbol{I})$ a rozptylem $\hat{U} = \hat{U}(\boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{R}_{inc}, \boldsymbol{I})$, potom interval, který pokrývá populační parametr Q s pravděpodobností 95%, je $\hat{Q} \pm 1{,}96\hat{U}^{1/2}$.

Bayesovská inference také umožňuje vypočítat aposteriorní rozdělení pro nepozorované (chybějící) hodnoty $\boldsymbol{Y}_{nob}$. Jelikož $Q(\boldsymbol{X}, \boldsymbol{Y})$ je funkcí pozorovaných hodnot v $\boldsymbol{X}, \boldsymbol{Y}_{obs}$ a nepozorovaných (chybějících) hodnotách $\boldsymbol{Y}_{nob}$, lze aposteriorní rozdělení nepozorovaných (chybějících) hodnot $\boldsymbol{Y}_{nob}$ vyjádřit jako:

$$Pr(Q | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{R}_{inc}, \boldsymbol{I}) = \int_{\omega(Q)} Pr(\boldsymbol{Y}_{nob} | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{R}_{inc}, \boldsymbol{I}) d\boldsymbol{Y}_{nob}, \quad (10)$$

kde

$$\omega(Q) = \{\boldsymbol{Y}_{nob} | Q(\boldsymbol{X}, \boldsymbol{Y})\}. \quad (11)$$

Aposteriorní rozdělení nepozorovaných (chybějících) hodnot $\boldsymbol{Y}_{nob}$ lze přímo vyjádřit pomocí specifikace pro proměnné $(\boldsymbol{X}, \boldsymbol{Y})$ za pomoci mechanismů výběru a odpovědí jako

$$Pr(\boldsymbol{Y}_{nob}|\boldsymbol{X},\boldsymbol{Y}_{obs},\boldsymbol{R}_{inc},\boldsymbol{I}) = \frac{\int_{\boldsymbol{R}_{exc}} Pr(\boldsymbol{X},\boldsymbol{Y})Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R})Pr(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{Y})d\boldsymbol{R}_{exc}}{\int_{\boldsymbol{Y}_{nob}}\int_{\boldsymbol{R}_{exc}} Pr(\boldsymbol{X},\boldsymbol{Y})Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R})Pr(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{Y})d\boldsymbol{R}_{exc}d\boldsymbol{Y}_{nob}} \quad (12)$$

kde pravděpodobnosti jsou vyhodnocovány v pozorovaných hodnotách (realizaci) $(\boldsymbol{Y}_{nob}|\boldsymbol{X},\boldsymbol{Y}_{obs},\boldsymbol{R}_{inc},\boldsymbol{I})$.

5. MECHANISMUS TVORBY CHYBĚJÍCÍCH DAT A JEHO IGNOROVATELNOST

Jako příčinu vzniku procesu, který generuje chybějící údaje ve statistickém šetření, lze podle [12] považovat způsob výběru vzorků, tj. mechanismus výběru $Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R})$ a míra dosažených pozorovaných hodnot, tj. mechanismus odpovědí $Pr(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{Y})$. V rámci statistické inference za podmínek chybějících dat se k odhadům o populaci používá statistický model pro zjišťované proměnné $(\boldsymbol{X},\boldsymbol{Y})$ vzhledem k danému parametru $\boldsymbol{\theta}$, tj. $f(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})$, doplněný o statistický model pro indikátory chybějících dat podmíněný konkrétní realizací sledovaných náhodných proměnných $(\boldsymbol{X},\boldsymbol{Y})$ a parametrem mechanismu $\boldsymbol{\varphi}$, tj. $g(\boldsymbol{I},\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\varphi})$. Symbol $\boldsymbol{M}$ označuje plně pozorovanou matici binárních indikátorů chybějících dat, přičemž $\boldsymbol{M} = \{M_{ij}\} = \{I_{ij}R_{ij}\}$ s navzájem jednoznačnými (prostými) funkcemi $(\boldsymbol{I},\boldsymbol{M})$ a $(\boldsymbol{I},\boldsymbol{R}_{inc})$.

Je-li model chybějících dat $g(\boldsymbol{I},\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\varphi})$ po realizaci pozorovaných hodnot náhodných proměnných $\boldsymbol{I},\boldsymbol{M},\boldsymbol{X}$ a hodnot pozorovaných složek proměnné $\boldsymbol{Y}$ vyhodnocen jako nezávislý na nepozorovaných (chybějících) složkách proměnné $\boldsymbol{Y}$, tj. na $\boldsymbol{Y}_{nob}$, potom podle [11] chybějící data mohou být považována za náhodně chybějící, označované výrazem *missing-at-random* (MAR), tj.

$$g(\boldsymbol{I},\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\varphi}) = g(\boldsymbol{I},\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y}_{obs},\boldsymbol{\varphi}). \quad (13)$$

Pokud jsou navíc parametry $\boldsymbol{\theta}$ a $\boldsymbol{\varphi}$ odlišné, tj. nezávislé, potom *Věta* 8.1 v [13] dokazuje, že bayesovská inference může zanedbat proces, který generuje chybějící data, a umožňuje odhadnout parametr $\boldsymbol{\theta}$ na základě $f(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})$, apriorního rozdělení $Pr(\boldsymbol{\theta})$ a pozorovaných hodnot $(\boldsymbol{X},\boldsymbol{Y}_{obs})$.

V kontextu bayesovské inference pro náhodnou proměnnou $\boldsymbol{Y}$ je kombinace náhodně chybějících údajů a rozdílných hodnot parametrů $\boldsymbol{\theta}$ a $\boldsymbol{\varphi}$ prakticky ekvivalentní tomu, že mechanismy výběru a odpovědí jsou ignorovatelné. Podle [12] platí, že pokud jsou $\boldsymbol{\theta}$ a $\boldsymbol{\varphi}$ apriorně nezávislé a chybějící data jsou náhodně chybějící (typu MAR), pak pro mechanismus výběru a mechanismus odpovědí platí, že jsou shodné, tj.

$$Pr(\boldsymbol{Y}_{nob}|\boldsymbol{X},\boldsymbol{Y}_{obs},\boldsymbol{R}_{inc},\boldsymbol{I}) = Pr(\boldsymbol{Y}_{nob}|\boldsymbol{X},\boldsymbol{Y}_{obs},\boldsymbol{M},\boldsymbol{I}). \quad (14)$$

Praktický význam toho, že mechanismus tvorby chybějících dat je ignorovatelný, spočívá ve skutečnosti, že aposteriorní rozdělení nepozorované části proměnné $\boldsymbol{Y}$, tedy $\boldsymbol{Y}_{nob}$, může být zjištěno na základě pozorovaných hodnot $\boldsymbol{X},\boldsymbol{Y}_{obs}$ a specifikaci konkrétního rozdělení pro náhodné proměnné $\boldsymbol{X},\boldsymbol{Y}$, tj.:

$$\begin{aligned} Pr(\boldsymbol{Y}_{nob}|\boldsymbol{X},\boldsymbol{Y}_{obs},\boldsymbol{R}_{inc},\boldsymbol{I}) &= Pr(\boldsymbol{Y}_{nob}|\boldsymbol{X},\boldsymbol{Y}_{obs}) \\ &= Pr(\boldsymbol{X},\boldsymbol{Y})/\int Pr(\boldsymbol{X},\boldsymbol{Y})d\boldsymbol{Y}_{nob} \\ &= \int_{\boldsymbol{\theta}} f(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})Pr(\boldsymbol{\theta})d\boldsymbol{\theta}/\int_{\boldsymbol{Y}_{nob}}\int_{\boldsymbol{\theta}} f(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})Pr(\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{Y}_{nob} \end{aligned}, \quad (15)$$

kde specifikované konkrétní rozdělení $Pr(\boldsymbol{X},\boldsymbol{Y}) = \int f(\boldsymbol{X},\boldsymbol{Y})Pr(\boldsymbol{\theta})d\boldsymbol{\theta}$ je využíváno pro realizaci statistické indukce o sledované proměnné $\boldsymbol{Y}$.

V praxi převažuje mechanismus *missing-at-random* (MAR), tj. chybějící data chybí náhodně nezávisle na nepozorovaných (chybějících) složkách proměnné $\boldsymbol{Y}$ V případě, že nepozorovaná data chybí nenáhodně, tj. mechanismus chybění není náhodný, vztahy (13 až 15) neplatí a je nutné použít jeden z následujících postupů:

- **První přístup představují tzv. modely výběru (selection models)**: 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 [13] specifikuje model [5], podle kterého závisí výskyt chybějících hodnot na datové matici $\boldsymbol{Y}$.
- **Druhý přístup tvoří tzv. modely smíšených vzorů (pattern-mixture models)**: Jednotky jsou rozdělené do skupin podle jedinečných vzorů chybějících hodnot. Následně se v každé ze skupin provádí statistická analýza [4]. Použitý pojem „smíšené vzory“ má indikovat, že výsledné marginální rozdělení datové matice $\boldsymbol{Y}$ je směsí několika rozdělení [5].

Podrobnější analýza nenáhodného mechanismu chybění v neúplných datech přesahuje rozsah tohoto článku. Detailnější informace o mechanismu NMAR lze dohledat například v [7] nebo v [1].

6. NÁZORNÉ PŘÍKLADY BAYESOVSKÉ ANALÝZY V SYSTÉMU SAS

Programový systém SAS nabízí 2 způsoby analýzy dat bayesovskou metodou. Ve vybraných procedurách pro statistické modelování se bayesovská analýza volá příkazem BAYES a univerzální proceduru MCMC pro obecné statistické modelování pomocí Bayesovy věty [8]. Statistická inference (viz předcházející kapitoly) – bodové i intervalové odhady - vychází z aposteriorního rozdělení. Bayesovské modelování také nabízí alternativní modelové řešení v analýze chybějících hodnot procedurou MI, ve kterém se s chybějícími hodnotami zachází jako s neznámými parametry a podle toho se odhadují [10]. Tyto chybějící hodnoty jsou jednoduše imputovány přijatelnými hodnotami prostřednictvím simulací aposteriorního rozdělení neúplných údajů.

Jádro procedur pro generování pseudonáhodných čísel z pravděpodobnostních rozdělení tvoří metody založené na simulaci pomocí Markovských řetězců. Podle [14] je Markovský řetězec sekvence náhodných veličin, u nichž rozdělení každého prvku závisí výlučně na hodnotě prvku předchozím. Je-li generované rozdělení konjugované, vzorky jsou generovány přímo z podmíněných aposteriorních rozdělení využitím standardních číselných generátorů[2]. Pro účely simulace metodou Markovských řetězců jsou nejpopulárnějšími algoritmy Gibbsovské vzorkování (Gibbs Sampling) a Metropolis-Hastings algoritmus. Metropolis-Hastings algoritmus simuluje výběr z aposteriorního pravděpodobnostního rozdělení za účelem aproximace požadovaného rozdělení a provádí rozhodování o přijetí nebo o odmítnutí vygenerovaného náhodného vzorku. Pokud je nově generovaný vzorek akceptován, algoritmus generuje vzorek nový, pokud je náhodně generovaný vzorek odmítnut,

[2] Je-li aposteriorní rozdělení členem stejné rodiny rozdělení jako rozdělení apriorní, pak obě tato rozdělení jsou konjugovaná neboli obecně – jsou stejného typu – pozn. aut.

simulace stávajícího výběru se opakuje. Generování náhodných vzorků končí v okamžiku dosažení konvergence k cílovému aposteriornímu rozdělení. Po dosažení konvergence se vybere náhodně z několika nasimulovaných řad a vybírá se řada podle spolehlivosti konvergence k cílovému rozdělení. Gibbsovo vzorkování (jako speciální případ předchozího algoritmu) rozděluje simultánní aposteriorní rozdělení vícerozměrné náhodné veličiny na podmíněná rozdělení pro každý parametr, ze kterých náhodně generuje řetězce a sleduje konvergenci k požadovanému aposteriornímu rozdělení.

Uvedený příklad se poprvé objevil v práci [6]. Datový soubor představuje náhodně vybraný vzorek s dvojrozměrným normálním rozdělením s chybějícími daty. Jedná se o uměle vytvořený problém, kterým však později zabývali i mnozí další následovníci, například v [17] a [18].

Tabulka č. 2: Struktura souboru s normálním rozdělením s chybějícími daty

Sledované proměnné		*Indikátorové (pomocné) proměnné*					
X_1	X_2	I		R		M	
1	1	1	1	1	1	1	1
1	−1	1	1	1	1	1	1
−1	1	1	1	1	1	1	1
−1	−1	1	1	1	1	1	1
2	·	1	1	1	0	1	0
2	·	1	1	1	0	1	0
−2	·	1	1	1	0	1	0
−2	·	1	1	1	0	1	0
·	2	1	1	0	1	0	1
·	2	1	1	0	1	0	1
·	−2	1	1	0	1	0	1
·	−2	1	1	0	1	0	1

Zdroj: [4, s. 27]

Soubor dat (nazvaný BINORM) se po provedeném náhodném výběru skládá z 12 pozorování z dvojrozměrného normálního rozdělení. Strukturu souboru s chybějícími daty zobrazuje tabulka č. 2. U obou komponent tohoto vícerozměrného rozdělení proměnných chybí hodnoty. Dvojrozměrné normální rozdělení má obě komponenty se stejnými nulovými průměry $\boldsymbol{\mu}$, s různými rozptyly σ_1^2 a σ_2^2 a korelačním koeficientem ρ. Tyto předpoklady se promítají do následující kovarianční matice $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{16}$$

Apriorní rozdělení kovarianční matice $\boldsymbol{\Sigma}$ je $Pr(\boldsymbol{\Sigma})$ a odpovídá podle [14] inverznímu Wishartovu rozdělení s determinantem $\left(\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-(p+1)}$, tj.

$$Pr(\boldsymbol{\Sigma}) \propto [\boldsymbol{\Sigma}]^{-\frac{p+1}{2}} = \left(\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-(p+1)}, \tag{17}$$

kde symbol p odpovídá počtu sledovaných proměnných, tedy $p = 2$.

Jelikož se jedná o výběrový soubor o velikosti n náhodně vybraný jednoduchým náhodným výběrem, indikátory zahrnutí pro všechny jednotky jsou rovné 1. Mechanismus výběru je proto dán výrazem

$$Pr(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}) = Pr(\boldsymbol{I}) = 1/\binom{N}{n} \text{ pro } \sum_{i=1}^{N} I_i \text{ a } Pr(\boldsymbol{I}) = 0 \text{ jinak.} \quad (18)$$

Při analýze kompletních pozorování, kdy se do úvahy berou pouze úplná pozorování, by se bayesovským modelováním vytvořil poměrně jednoduchý aposteriorní odhad: 2 datové dvojice ($(1,1)$ a $(-1,-1)$) mají korelaci rovnou 1 a 2 dvojice dat ($(1,-1)$ a $(-1,1)$) mají korelaci -1. V případě, že se budou chybějící hodnoty považovat za zcela náhodně chybějící, lze podle [10] považovat nepozorované hodnoty buď (téměř) dokonale pozitivně nebo dokonale negativně korelující s pozorovanými hodnotami 2 a -2 a že chybějící hodnoty mohou nabývat pouze hodnot, které se od nich příliš neliší. Navíc vyřazením částečně pozorovaných hodnot se ztrácí podstatná část při odhadu rozptylu obou komponent dvojrozměrného normálního rozdělení. Bayesovské modelování chybějících hodnot dvojrozměrného normálního rozdělení vykonávají následující příkazy programového systému SAS:

```
proc mcmc data=BiNorm nmc=20000 seed=17 outpost=postout
     diag=none plots=none monitor=(rho sig1 sig2);
     array x[2] x1 x2;
     array mu[2] (0 0);
     array sigma[2, 2];
     parms rho 0.5 / slice;
     parms sig1 1 sig2 1;
     if (sig1 > 0 and sig2 > 0) then
          lprior = -3 * log(sig1 * sig2 * sqrt(1-rho*rho));
          else lprior = .;
     prior rho sig1 sig2 ~ general(lprior);
     sigma[1,1] = sig1*sig1;
     sigma[1,2] = rho * sig1 * sig2;
     sigma[2,1] = sigma[1,2];
     sigma[2,2] = sig2*sig2;
     model x ~ mvn(mu, sigma);
  run;
```

Příkazy PARMS specifikují tři modelované parametry modelu. Volba SLICE v prvním příkazu PARMS vybere vzorkovací algoritmus, kterým se generuje parametr ρ. I když výpočetně náročnější [9], tento typ algoritmu je velmi vhodný pro simulaci dat z nenormálních rozdělení (kam náleží i rozdělení korelačního koeficientu ρ). Příkazy IF-ELSE zajišťují, aby simulované hodnoty parametrů směrodatné odchylky σ_1 a σ_2 byly zahrnovány do aposteriorních odhadů kovarianční matice pouze při své kladné hodnotě. Příkazem LPRIOR= a příkazem PRIOR se do procedury zadávají simultánní apriorní rozdělení parametrů σ_1, σ_2 a ρ na logaritmické stupnici[3]. Prvky kovarianční matice Σ jsou konstruovány použitím dalších programových příkazů jako funkcí parametrů modelu (v syntaxi příkazů označené jako $sigma[1,1]$ až $sigma[2,2]$). Příkazem MODEL se nastavuje specifikace požadované funkce věrohodnosti nutná pro simulaci a následnou aposteriorní inferenci.

[3] *Použitím výpočtů na logaritmické stupnici se dosahuje větší stabilita a rychlejší konvergence v číselném výpočtu.*

Po provedené simulaci na základě výše uvedené syntaxe je výstupem činnosti procedury MCMC soubor POSTOUT obsahující 20000 vygenerovaných chybějících hodnot v obou proměnných $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$ ve dvojrozměrném normálním rozdělení a vygenerované rozdělení parametrů σ_1, σ_2 a ρ podle zadání. Analýza chybějících hodnot je ilustrována v části *Missing Data Information Table* tabulky č. 3. Proměnná $\boldsymbol{X}_1$ obsahuje 4 chybějící hodnoty, proměnná $\boldsymbol{X}_2$ obsahuje také 4 chybějící hodnoty. Tabulka zaznamenává indexy chybějících hodnot za jednotlivé proměnné. Aposteriorní inference o odhadovaných parametrech σ_1, σ_2 a ρ je zobrazena v části *Posterior Summaries and Intervals* tabulky č. 3.

Tabulka č. 3: Sumární statistiky a inference výsledného rozdělení

Missing Data Information Table

Variable	Number of Missing Obs	Observation Indices	Sampling Method
X1	4	9 10 11 12	Direct
X2	4	5 6 7 8	Direct

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
rho	20000	-0.00184	0.5582	-0.8591	0.8519
sig1	20000	1.6555	0.4458	0.9828	2.6014
sig2	20000	1.6306	0.4444	0.9478	2.4860

Zdroj: vlastní zpracování

Aposteriorní inference o chybějících hodnotách proměnných $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$ je zobrazena ve výstupních tabulkách č. 4.

Tabulka č. 4: Sumární statistiky a inference chybějících hodnot

Parameter	N	Mean	StdDev	P25	P50	P75
X2_5	20000	-0.000219	1.84723	-1.25392	0.003381	1.25390
X2_6	20000	0.001579	1.83949	-1.23536	-0.003435	1.24945
X2_7	20000	0.004489	1.83727	-1.23108	0.036230	1.22791
X2_8	20000	-0.008475	1.84616	-1.25881	0.005462	1.23491
X1_9	20000	-0.023454	1.88411	-1.31367	-0.025875	1.25902
X1_10	20000	-0.003563	1.88864	-1.27118	0.007136	1.29767
X1_11	20000	0.003396	1.89574	-1.30166	0.004094	1.27681
X1_12	20000	0.002021	1.86815	-1.26153	0.018829	1.26093

Parameter	Alpha	CredibleLower	CredibleUpper	HPDLower	HPDUpper
X2_5	0.05	-3.55760	3.61006	-3.61197	3.54433
X2_6	0.05	-3.50325	3.53977	-3.48336	3.54871
X2_7	0.05	-3.60965	3.57439	-3.62454	3.55160
X2_8	0.05	-3.58898	3.55509	-3.57001	3.56999
X1_9	0.05	-3.69331	3.62257	-3.75078	3.55437
X1_10	0.05	-3.66322	3.64678	-3.63647	3.66321
X1_11	0.05	-3.69566	3.69085	-3.73594	3.63509
X1_12	0.05	-3.63658	3.65181	-3.70957	3.55111

Zdroj: vlastní zpracování

Podobně jako v předchozím případě jsou apriorní i podmíněné aposteriorní rozdělení chybějících údajů konjugované. Podmíněná aposteriorní rozdělení chybějících hodnot $Pr(\boldsymbol{X}_1|\cdot)$ a $Pr(\boldsymbol{X}_2|\cdot)$ pro jednotlivé proměnné $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$ jsou jednorozměrná normální rozdělení a jsou dána vzorci:

$$Pr(\boldsymbol{X}_1|\rho,\sigma_1,\sigma_2,x_2) \sim N\left(\rho\frac{\sigma_1}{\sigma_2}x_2, \sigma_1^2(1-\rho^2)\right) \text{ a} \qquad (19)$$

$$Pr(\boldsymbol{X}_2|\rho,\sigma_1,\sigma_2,x_2) \sim N\left(\rho\frac{\sigma_2}{\sigma_1}x_2, \sigma_2^2(1-\rho^2)\right). \qquad (20)$$

Následující příklad ilustruje možnosti procedur z programového systému SAS, které jsou přímo určeny k nahrazování chybějících hodnot způsobem zajišťujícím adekvátní statistickou inferenci. Jedná se o procedury MI [15] a MIANALYZE [16], které byly zavedeny jako standardní součásti programového modulu SAS/STAT počínaje verzí 9.0. Vícenásobná imputace se nesnaží odhadnout každou chybějící hodnotu pomocí simulovaných hodnot, ale spíše modeluje náhodný vzorek chybějících hodnot. Výsledek tohoto procesu dává možnost platných statistických závěrů, které řádně

odrážejí nejistotu způsobenou chybějícími hodnotami; například platné intervaly spolehlivosti pro parametry.

Mnohonásobná imputační inference zahrnuje tři různé fáze:

- chybějící údaje se doplní m krát, aby se vytvořilo m úplných souborů dat,
- m úplných datových souborů se analyzuje pomocí standardních postupů a
- výsledky z m úplných datových souborů se zkombinují pro účely inference.

Procedura mnohonásobné imputace v softwaru SAS/STAT je procedura MI a vytváří soubory mnohonásobně imputovaných dat pro neúplná p-rozměrná vícerozměrná data. Procedura používá metody, které zahrnují vhodnou variabilitu napříč m imputacemi. Jakmile je m úplných datových souborů analyzováno pomocí standardních postupů, lze použít proceduru MIANALYZE k vytvoření platných statistických závěrů o těchto parametrech kombinací výsledků z m úplných datových souborů. Tento postup syntetizuje výsledky tím, že vytvoří průměry bodových odhadů, které jsou předmětem zájmu (průměry, odhady parametrů atd.), napříč imputovanými datovými soubory spolu s upravenými rozptyly a směrodatnými chybami, které zohledňují nejistotu vnesenou do procesu imputace.

Výsledný bodový odhad populačního průměru je kombinovaný bodový odhad populačního průměru, což je průměrná hodnota bodových odhadů přes všech m nezávislých opakování imputací, tj.

$$\bar{\hat{\theta}} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i, \tag{21}$$

kde $\hat{\theta}_i$ je odhad parametru θ z i -tého souboru dat (po imputaci) pro $i = 1, \dots, m$.

Vnitroimputační rozptyl (v rámci imputací) je počítán jako průměr odhadovaných rozptylů $var(\hat{\theta}_i)$ bodových odhadů $\hat{\theta}_i$ populačního průměru přes všech m imputací, tj.

$$\bar{\widehat{W}} = \frac{1}{m}\sum_{i=1}^{m} \widehat{W}_i = \frac{1}{m}\sum_{i=1}^{m} var(\hat{\theta}_i), \tag{22}$$

kde $var(\hat{\theta}_i)$ je odhad rozptylu odhadu populačního průměru pro i -tou imputaci.

Mezimputační rozptyl (mezi imputacemi) je odhadován pomocí vzorce

$$\bar{\hat{B}} = \frac{1}{m}\sum_{i=1}^{m} \left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2, \tag{23}$$

Celkový rozptyl odhadu parametru θ po mnohonásobné imputaci je odhadnut pomocí Rubinovy kombinační formule [12], tj.

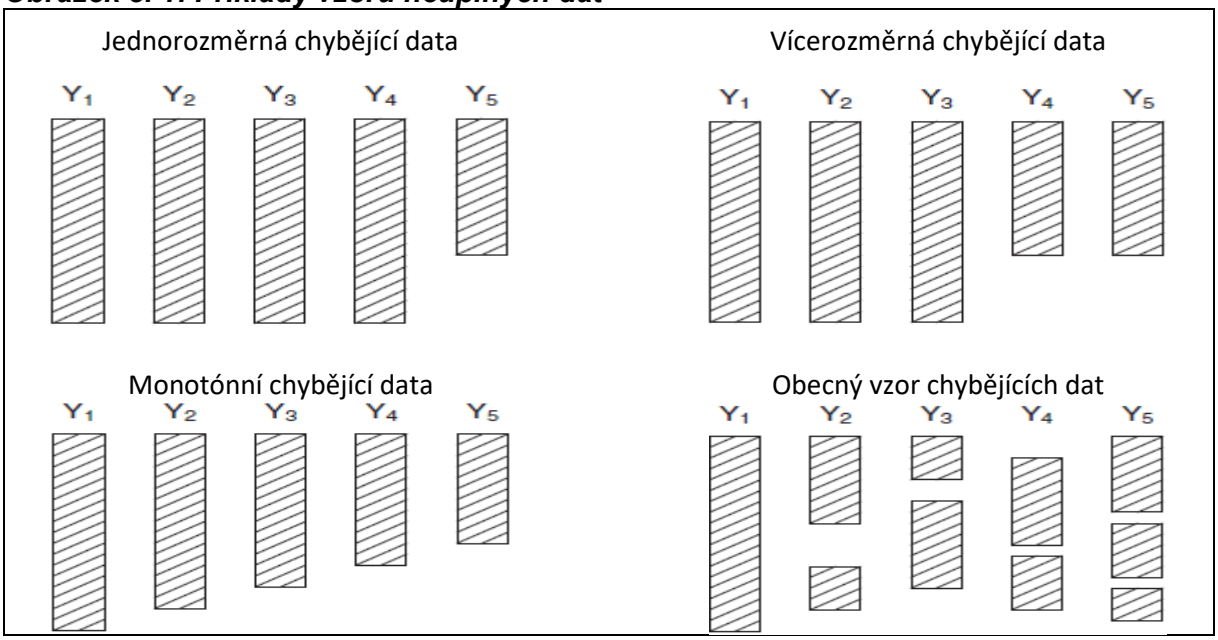
$$var\left(\bar{\hat{\theta}}\right) = \bar{\widehat{W}} + \frac{m+1}{m}\bar{\hat{B}}. \tag{24}$$

Klíčem ke zvolení imputační metody je statistický model, který definuje pravděpodobnostní vztahy všech uvažovaných proměnných a typ imputované proměnné. Vzor chybějících údajů popisuje umístění chybějících hodnot v datové matici. Za tímto účelem lze řádky a sloupce datové matice seřadit tak, aby vznikly

konkrétní vzory chybějících údajů. Zobrazení 4 speciálních vzorů je ilustrováno na obrázku č. 1.

Například monotónní vzor chybějících údajů znamená, že proměnné datové matice se dají uspořádat tak, aby pro každou další proměnnou chyběly hodnoty u všech jednotek, a mohou chybět i nějaké hodnoty navíc. Tento vzor je automaticky splněn, pokud chybějící hodnoty byly sledovány u jediné proměnné.

Obrázek č. 1: Příklady vzorů neúplných dat



Zdroj: [2, str. 132]

Tabulka č. 5: Shrnutí možných metod imputace pomocí procedury MI podle [15]

Vzor chybění	Typ imputované proměnné	Doporučená metoda imputace
Monotónní	Spojitá	Monotónní regrese Monotónní predikovaná shoda v průměru Monotónní propensitní[1] skóre
	Klasifikační (ordinální)	Monotónní logistická regrese
	Klasifikační (nominální)	Monotónní diskriminační funkce
Libovolný (obecný)	Spojitá	Imputace obecným MCMC Imputace monotónní MCMC

1 Odhad pravděpodobnosti přiřazené každému pozorování u proměnné obsahují chybějící hodnoty, že pozorování je chybějící

Zdroj: vlastní zpracování podle [15]

Shrnutí metod mnohonásobné imputace procedurou MI je uvedeno v tabulce č. 5. Podrobnější informace ke každé z výše uvedených imputačních metod jsou uvedeny především v originální dokumentaci programového systému SAS [15]. Jako příklad využití procedury MI při mnohonásobné imputaci bude prakticky předvedena imputace metodou monotónní regrese pro spojité proměnné s chybějícími pozorováními za podmínek ignorovatelného mechanismu chybění MAR. Příklady mnohonásobných imputací, ve kterých jsou aplikovány ostatní imputační metody, jsou uvedeny v originální dokumentaci pro proceduru MI.

Při regresní metodě se pro každou proměnnou s chybějícími hodnotami sestaví regresní model. Na základě tohoto modelu pro proměnnou $\boldsymbol{Y}_j$ se dále sestaví nový regresní model pro proměnnou $\boldsymbol{Y}_{j+1}$, který se použije k imputování chybějících hodnot proměnné [12]. Vzhledem k tomu, že soubor dat má monotónní vzor chybějících dat, proces se opakuje postupně pro ostatní proměnné s chybějícími hodnotami.

Pro spojitou proměnnou $\boldsymbol{Y}_j$ s chybějícími hodnotami platí následující regresní model

$$\boldsymbol{Y}_j = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \cdots + \beta_k \boldsymbol{X}_k, \tag{25}$$

který je vyrovnáván na základě pozorovaných hodnot proměnné $\boldsymbol{Y}_j$ a jejích kovariát $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots \boldsymbol{X}_k$.

Na základě předchozího textu podle [12] pro proměnnou $\boldsymbol{Y}_j$ s chybějícími hodnotami jsou chybějící hodnoty imputovány z rozdělení

$$\boldsymbol{Y}_j \sim Pr(\boldsymbol{Y}_{nob} | \boldsymbol{X}, \boldsymbol{Y}_{obs}) = Pr\left(\boldsymbol{Y}_j \middle| \boldsymbol{X}, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots \boldsymbol{Y}_{j-1}\right). \tag{26}$$

K imputování chybějících hodnot pro proměnnou $\boldsymbol{Y}_j$ jsou v každé imputaci uplatněny následující kroky:

- Vyrovnaný model zahrnuje odhady regresních parametrů $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_k\right)$ a příslušné kovarianční matice $\hat{\sigma}_j^2 \boldsymbol{V}_j$, kde kovarianční matice odpovídá inverzní matici $\boldsymbol{X}'\boldsymbol{X}$ odvozené z absolutního členu a kovariát $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots \boldsymbol{X}_k$.

- Nové parametry $\boldsymbol{\beta} = (\beta_{*0}, \beta_{*1}, \beta_{*2}, \ldots \beta_{*k})$ a σ_{*j}^2 jsou vybírány aposteriorního predikčního rozdělení parametrů [12], což znamená, že parametry jsou simulovány z $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k\right)$, $\hat{\sigma}_j^2$ a $\boldsymbol{V}_j$. Rozptyl je generován jako

$$\sigma_{*j}^2 = \hat{\sigma}_j^2 \left(n_j - k - 1\right)/g, \tag{27}$$

  kde g je $X_{n_j-k-1}^2$ náhodná proměnná a n_j je počet pozorovaných hodnot v proměnné $\boldsymbol{Y}_j$. Regresní koeficienty jsou simulovány podle rovnice

$$\boldsymbol{\beta}_* = \widehat{\boldsymbol{\beta}} + \sigma_{*j} \boldsymbol{V}_{hj}' \boldsymbol{Z}, \tag{28}$$

kde $\boldsymbol{V}_{hj}'$, $\boldsymbol{V}_j = \boldsymbol{V}_{hj}' \boldsymbol{V}_{hj}$, je horní trojúhelníková matice daná jako Choleskyho dekompozice a $\boldsymbol{Z}$ je vektor $k+1$ nezávislých náhodných proměnných.

Chybějící hodnoty jsou nahrazovány $\beta_{*0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + z_i \sigma_{*j}$, kde $x_1, x_2, \ldots x_k$, jsou hodnoty kovariát a z_i je i-tá simulovaná náhodná odchylka.

Regresní metodě imputace je velmi podobná metoda shody predikčních průměrů. Rozdíl od regresní metody imputace spočívá v tom, že pro každou chybějící hodnotu imputuje pozorovanou hodnotu, která je nejblíže předpovězené hodnotě ze simulovaného regresního modelu [12].

Pro nominální klasifikační proměnnou je na základě vyrovnaného logistického regresního modelu je podle [15] simulován nový logistický regresní model z posteriorního predikčního rozdělení parametrů a je použit k imputování chybějících hodnot postupně pro každou proměnnou. Logistický regresní model nominální klasifikační proměnné o K úrovních pro imputování chybějících hodnot proměnné $\boldsymbol{Y}_j$ je vyrovnáván pomocí pozorovaných hodnot proměnných zařazených do imputačního modelu. Postupy imputace jsou analogické jako u monotónní lineární regrese u spojité proměnné.

Příklad byl převzat ze zdroje [15][4] a ilustruje ukázku mnohonásobné imputace regresní metodou (logistická regrese pro klasifikační proměnnou Druhy (Species) a lineární regrese pro spojité proměnné *výška* (Height) a *šířka* (Width)) nad datovým souborem nazývaném Fish, jehož údaje mají monotónní vzor chybění. Datový soubor obsahuje 2 druhy potočních ryb Cejn (Bream) a Štika (Pike), přičemž každá sledovaná ryba obsahuje informace o délce (Length), výšce (Height) a šířce (Width). Některé hodnoty z těchto 3 měření byly nastaveny jako chybějící a výsledný vzor chybění u proměnných délky (Length), výška (Height), Width (šířka) a Druhy (Species) má monotónní charakter. Prvních 10 pozorování tohoto souboru dat ilustruje tabulka č. 6.

Tabulka č. 6: Soubor dat pro imputaci regresní metodou (prvních 10 pozorování)

Obs	Species	Length	Height	Width
1	Bream	30.0	11.520	4.020
2		31.2	12.480	4.306
3	Bream	31.1	12.378	4.696
4	Bream	33.5	12.730	4.456
5		34.0	12.444	.
6	Bream	34.7	13.602	4.927
7	Bream	34.5	14.180	5.279
8	Bream	35.0	12.670	4.690
9	Bream	35.1	14.005	4.844
10	Bream	36.2	14.227	4.959

Zdroj: vlastní zpracování

Mnohonásobná imputace začíná tvorbou m imputovaných souborů dat pomocí procedury MI. Výsledný soubor dat (kombinace m imputovaných souborů) se nazývá *OutFish*. Sekvence příkazů, které vyvolávají proceduru MI a požadují regresní metodu imputace pro proměnné *výška* (Height) a *šířka* (Width) a metodu logistické regrese pro proměnnou Druhy (Species), je následující:

```
proc mi data=Fish seed=1305417 out=OutFish;
      class Species;
      monotone reg(Height Width/ details)
               logistic(Species= Length Height Width Height*Width/
               details);
      var Length Height Width Species;
run;
```

V lineárním regresním modelu bude jako závislá proměnná vystupovat proměnná *výška* (Height), v modelu logistické regrese je na místě závislé nominální proměnná Druhy (Species). Seznam proměnných zahrnutých do modelu je určen výčtem

[4] https://documentation.sas.com/doc/en/statcdc/14.3/statug/statug_mi_examples04.htm

proměnných ze souboru dat v příkazu VAR. Pokud příkaz VAR se seznamem proměnných uveden není, do modelu pro imputaci jsou zahrnuty všechny proměnné. Klasifikační proměnná, pokud má být součástí imputačního modelu, musí být uvedena v příkazu CLASS procedury. Po vykonání výše uvedené příkazové sekvence se na výstupu procedury MI objeví výstupní tabulky a další informace uvedené v následujícím textu.

Metodu a možnosti použité v procesu vícenásobné imputace popisuje tabulka *Model Information*. Ve výchozím nastavení se pro chybějící data se vytvoří 25 imputací (počet imputací procedury MI lze změnit příkazem NIMPUTE=). Tabulka *Monotone Model Specification* zobrazuje monotónní metody použité při imputaci (viz tabulka č. 7).

Tabulka č. 7: Informace o imputačním modelu a použitých metodách imputace

Model Information	
Data Set	WORK.FISH
Method	Monotone
Number of Imputations	25
Seed for random number generator	1305417

Monotone Model Specification	
Method	Imputed Variables
Regression	Height Width
Logistic Regression	Species

Zdroj: vlastní zpracování

Jednotlivé vzory chybějících dat znázorňuje tabulka *Missing Data Patterns* v tabulce č. 8 (včetně odpovídacích absolutních a relativních četností). Přítomnost pozorovaných hodnot proměnné v příslušné skupině je indikována znakem „X“. Naopak nevyplněné hodnoty proměnné ilustruje znak „.“. Pro spojité proměnné je v každé zobrazované skupině zobrazen také průměr proměnné.

Tabulka č. 8: Vzor neúplných dat generovaný procedurou MI

Missing Data Patterns									
							Group Means		
Group	Length	Height	Width	Species	Freq	Percent	Length	Height	Width
1	X	X	X	X	43	82.69	41.997674	12.819512	5.359860
2	X	X	X	.	3	5.77	38.433333	11.797667	4.587667
3	X	X	.	.	4	7.69	42.275000	13.346750	.
4	X	.	.	.	2	3.85	40.150000	.	.

Zdroj: vlastní zpracování

Zobrazit regresní koeficienty v modelu REG, které jsou odhadovány z pozorovaných dat, a regresní koeficienty, které jsou použity v každé imputaci, umožňuje připojení volitelného parametru DETAILS do skupiny příkazů. Ukázka regresních koeficientů simulovaných z aposteriorního rozdělení je v tabulce č. 9 pro prvních 7 iterací.

Tabulka č. 9: Odhady parametrů lineárního modelu pro prvních 7 iterací

Imputed Variable	Effect	Obs-Data	1	2	3	4	5	6	7
Height	Intercept	0.00173	-0.152270	-0.136544	-0.064801	0.036585	0.088415	0.125572	0.019478
Height	Length	-0.22453	-0.133455	-0.155687	-0.319043	-0.108935	-0.215399	-0.080855	0.100551

Zdroj: vlastní zpracování

Použitím stejného volitelného parametru DETAILS u logistického imputačního modelu (model LOGISTIC) lze zobrazit hodnoty regresní koeficienty, které jsou odhadovány z pozorovaných dat, a regresní koeficienty, které jsou nasimulovány v každé imputaci. Ukázka koeficientů logistického imputačního modelu je ilustrovaná tabulkou č. 10.

Tabulka č. 10: Odhady parametrů logistického modelu pro prvních 7 imputací

Imputed Variable	Effect	Obs-Data	1	2	3	4	5	6	7
Species	Intercept	22.80713	22.807129	22.807129	22.807129	22.807129	22.807129	22.807129	22.807129
Species	Length	-14.44698	-14.446980	-14.446980	-14.446980	-14.446980	-14.446980	-14.446980	-14.446980
Species	Height	43.11236	43.112363	43.112363	43.112363	43.112363	43.112363	43.112363	43.112363
Species	Width	-9.64352	-9.643524	-9.643524	-9.643524	-9.643524	-9.643524	-9.643524	-9.643524
Species	Height*Width	-9.73015	-9.730154	-9.730154	-9.730154	-9.730154	-9.730154	-9.730154	-9.730154

Zdroj: vlastní zpracování

Výstupem procedury MI je soubor dat nazvaný *OutFish*, jenž obsahuje kombinaci m imputovaných souborů. Prvních 10 pozorování tohoto souboru dat ilustruje tabulka č. 11. Z obrázku je patrné, že procedura MI nahradila chybějící hodnoty imputovanými, například do druhého pozorování byla doplněna hodnota proměnné Species, a to již v rámci prvního imputovaného souboru.

Tabulka č. 11: Soubor dat OutFish na výstupu procedury MI (prvních 10 pozorování)

Obs	_Imputation_	Species	Length	Height	Width
1	1	Bream	30.0	11.520	4.02000
2	1	Bream	31.2	12.480	4.30600
3	1	Bream	31.1	12.378	4.69600
4	1	Bream	33.5	12.730	4.45600
5	1	Bream	34.0	12.444	4.62964
6	1	Bream	34.7	13.602	4.92700
7	1	Bream	34.5	14.180	5.27900
8	1	Bream	35.0	12.670	4.69000
9	1	Bream	35.1	14.005	4.84400
10	1	Bream	36.2	14.227	4.95900

Zdroj: vlastní zpracování

Výstupem procedury MI je také informace o rozptylu, a to jak mezi imputovanými soubory, tak i v rámci imputovaných souborů, který je potřebný k následné inferenci procedurou MIANALYZE. Informace o rozptylech je zaznamenána v tabulce č. 12.

Tabulka č. 12: Odhad rozptylu v rámci imputovaných souborů a mezi soubory imputace

Variance Information (25 Imputations)							
	Variance				Relative	Fraction	
Variable	Between	Within	Total	DF	Increase in Variance	Missing Information	Relative Efficiency
Height	0.010482	0.296800	0.307701	47.254	0.036730	0.035529	0.998581
Width	0.000534	0.014490	0.015045	47.171	0.038339	0.037032	0.998521

Zdroj: vlastní zpracování

Po vytvoření m imputovaných souborů pomocí procedury MI se přechází lze přejít ke standardní statistické analýze uvedených m imputovaných souborů. Následující

příkazová sekvence vytváří datové soubory, které obsahují odhady parametrů a odpovídající kovarianční matice odhadovaných pomocí logistické regrese. Pomocí standardní procedury LOGISTIC se pro každý imputovaný soubor generují odhady parametrů logistického imputačního modelu (v proceduře MI je tento model zadaný příkazem LOGISTIC). Všech m imputovaných souborů je uloženo v souboru *OutFish* z výstupu procedury MI:

```
proc logistic data=OutFish;
     class Species;
     model Species= Length Height Width Height*Width / covb;
     by _Imputation_;
     ods output ParameterEstimates=lgparms CovB=lgcovb;
run;
```

Výstupem procedury LOGISTIC je soubor s odhady parametrů a odhady jejich standardních chyb nazvaný LGPARMS pro všech m imputovaných souborů. Odhady kovariančních maticí všech m imputačních modelů jsou uloženy ve výstupním souboru LGCOVB. Oba tyto soubory odhadů parametrů a kovariančních matic z výstupu procedury LOGISTIC jsou vstupními soubory do procedury MIANALYZE v poslední fázi procesu imputačního procesu - kombinace výsledků imputace m imputovaných souborů a celkové statistické inference. Podobným způsobem lze analyzovat i lineární regresní model z procedury MI (v proceduře MI je tento model zadaný příkazem REG).

Závěrečnou fází imputačního procesu je použití procedury MIANALYZE, jejíž pomocí lze výsledky z m imputovaných (úplných) datových souborů se zkombinovat pro účely inference. K tomuto účelu lze použít následující sekvenci příkazů:

```
proc mianalyze parms=lgparms covb=lgcovb;
     modeleffects Intercept Length Height Width Height*Width;
run;
```

Vstupními soubory pro proceduru MIANALYZE jsou soubory s odhady parametrů (LGPARMS) a odhadů kovariančních matic (LGCOVB) pro všech m imputovaných (úplných) datových souborů z výstupu analytické procedury LOGISTIC, tj. soubory se vloží do procedury MIANALYZE pomocí příkazů PARMS= a COVB=. Příkaz MODELEFFECTS umožňuje analyzovat vliv jednotlivých efektů (nezávislých proměnných) v imputačním modelu.

Výstup procedury MIANALYZE představují informace, které slouží jako výsledná statistická inference po realizovaném procesu imputace. Procedura MIANALYZE kombinuje vstupní informace všech ze všech m imputovaných souborů do výsledných odhadů a zajišťuje tak statistickou inferenci celého procesu imputace. Tabulka *Model Information* obsahuje informace o datových souborech vstupujících do analýzy procedurou MIANALYZE a údaje o počtu imputací. Odhady celkového rozptylu včetně složek rozptylu jsou zaznamenány v tabulce *Variance Information*. Přehled odhadů parametrů imputačního modelu včetně odhadů standardních chyb regresních koeficientů obsahuje tabulka *Parameter Estimates*. Ukázku statistické inference procedurou MIANALYZE ilustruje tabulka č. 13.

Tabulka č. 13: Ukázka výstupů procedury MIANALYZE po procesu imputace

Model Information	
PARMS Data Set	WORK.LGPARMS
COVB Data Set	WORK.LGCOVB
Number of Imputations	25

Variance Information (25 Imputations)

Parameter	Variance Between	Variance Within	Variance Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Intercept	400.704089	180035	180452	4.5E6	0.002315	0.002310	0.999908
Length	0.080968	38.419197	38.503404	5.02E6	0.002192	0.002187	0.999913
Height	2.408527	1136.358955	1138.863824	4.96E6	0.002204	0.002200	0.999912
Width	32.413980	9735.783085	9769.493624	2.02E6	0.003463	0.003452	0.999862
Height*Width	0.077975	33.699632	33.780726	4.16E6	0.002406	0.002401	0.999904

Parameter Estimates (25 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	-39.590964	424.796213	-872.176	792.9945	4.5E6	-125.766370	-22.874243	0	-0.09	0.9257
Length	-0.542283	6.205111	-12.704	11.6195	5.02E6	-1.379808	-0.139264	0	-0.09	0.9304
Height	5.688145	33.747057	-60.455	71.8312	4.96E6	4.215492	12.157530	0	0.17	0.8661
Width	7.243556	98.840749	-186.481	200.9680	2.02E6	-0.014259	24.732078	0	0.07	0.9416
Height*Width	-0.650649	5.812119	-12.042	10.7409	4.16E6	-1.767419	-0.392503	0	-0.11	0.9109

Zdroj: vlastní zpracování

7. ZÁVĚR

V současné době jsou neúplné (chybějící) údaje velmi častou součástí souborů dat, přičemž 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).

Při řešení problémů imputace chybějících dat by výzkumníci měli zvažovat moderní statistické přístupy vycházející z bayesovského modelování chybějících dat. I když bayesovské metody ve srovnání s klasickými postupy vyžadují hlubší znalosti teorie současně s určitými výpočetními dovednostmi, jejich správná aplikace ve výzkumné činnosti se rozhodně vyplatí. Bayesovská statistika poskytuje obecnější přístup pro statistickou inferenci. Na rozdíl od četnostní statistiky bayesovský přístup považuje odhadované parametry jako náhodné, jejichž rozdělení je možné upřesňovat informacemi získanými z dat. Navíc v kontextu bayesovského přístupu jsou znalosti zkoumaného fenoménu, které lze efektivně využít, zpravidla vždy dostupné. V současné době je aplikace bayesovských metod mnohem přístupnější, protože už existují vhodné volně přístupné programy i placené programové systémy, například právě SAS.

Zvolené příklady do tohoto příspěvku patří k těm jednodušším imputačním problémům, které se však ve statistické praxi vyskytují nejčastěji. Je jasné, že složitější typy imputací vyžadují sofistikovanější statistické přístupy. Cílem zvolených příkladů

bylo přiblížit podstatu nahrazování chybějících hodnot s důrazem na úlohu bayesovského modelování. Dalším cílem článku je ilustrovat možnosti programového systému SAS při řešení náhrady imputacemi chybějících dat. Bližší popis a možnosti procedur SAS pro aplikaci imputačních metod bychom chtěli uveřejňovat v následujících vydáních, neboť problematika imputací a jejich řešení v programovém systému SAS je tak rozsáhlá, že není možné vše obsáhnout jediným příspěvkem.

LITERATURA

[1] ALLISON, P. D.: Missing Data. Thousand Oaks, CA: Sage. Sage University Papers Series. Quantitative Applications in the Social Sciences. No. 07-136. 2001. ISBN 0-7619-1672-5.
[2] BOX, G., E. P. – TIAO, G. C.: Bayesian Inference in Statistical Analysis. 1973. New York: John Wiley & Sons.
[3] GAVALAKIS L. – KONTOYIANNIS I.: Information-theoretic de Finetti-style theorems. 2022 IEEE Information Theory Workshop (ITW).
[4] ENDERS, C., K.: Applied Missing Data Analysis, Second Edition. New York: Guilford Press, 2022. ISBN 978-1-462-54986-3.
[5] GRAHAM, J., W.: Missing data: Analysis and design. 2010. New York: Springer. ISBN 978-1-4614-4017-8.
[6] MURRAY, G. D.: Discussion of Paper by Dempster, Laird, and Rubin. In: Journal of the Royal Statistical Society, Series B, 1977. 39, pp. 27 – 28.
[7] LITTLE, R., J.A.: Pattern-Mixture Models for Multivariate Incomplete Data. In: Journal of the American Statistical Association, 1993. No. 421. pp. 125 – 134.
[8] NAKAGAWA, Sh.: Missing data: Mechanisms, methods, and messages. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), Ecological statistics: contemporary theory and application. 2015. pp. 81 – 105.
[9] NEAL, R. M.:Slice Sampling. 2000. In: Technical Report No. 2005. Department of Statistics and Department of Computer Science. University of Toronto, Toronto, Ontario, Canada. Dostupné na: https://www.cs.toronto.edu/~radford/ftp/slc-samp.pdf.
[10] CHEN, F.: Practical Bayesian Computation using SAS. In: ASA Conference on Statistical Practices. 2014. February 20, s.3.
[11] PAVELKA, R.: Statistická analýza chybějících dat. In: Slovenská štatistika a demografia. 2024. č. 2. pp. 3 – 25.
[12] RUBIN, D. B.: Multiple imputation for nonresponse in surveys. New York: Springer. 1987. ISBN 0-471-08705-X.
[13] RUBIN, D. B.: Inference and Missing Data. In: Biometrika, 1976. Vol. 63, No. 3 pp. 581 – 592.
[14] SAS Institute Inc.: SAS/STAT® 15.3 User's Guide. 2023. Chapter 8 Introduction to Bayesian Analysis Procedures. Cary, NC: SAS Institute Inc.
[15] SAS Institute Inc. SAS/STAT® 15.3 User's Guide. 2023. Chapter 82 The MI Procedure. Cary, NC: SAS Institute Inc.
[16] SAS Institute Inc.: SAS/STAT® 15.3 User's Guide. 2023. Chapter 83 The MIANALYZE Procedure, Cary, NC: SAS Institute Inc.
[17] TAN, M., T. – TIAN, G-L. – NG, K., W., eds.: Bayesian Missing Data Problems: EM, Data Augmentation, and Non-iterative Computation, 2010. New York: Chapman & Hall/CRC.
[18] TANNER, M. A. – WONG, W., H.: The Calculation of Posterior Distributions by Data Augmentation. In: Journal of the American Statistical Association, 82, pp. 528 – 540.

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.
Bayesovské modelování – náhodné výběry z aposteriorního rozdělení, simulace pomocí metody Monte Carlo a Markovských řetězců-představují pevný základ téměř celé kategorie parametrických metod imputace. I když nejdůležitější faktor v podmínkách neúplných dat představuje mechanismus chybějících hodnot, který má nejvýraznější vliv na úspěšnost rozmanitých metod práce s chybějícími hodnotami, znalost a správné použití metod statistického modelování vycházející z daného konkrétního výběru (vzorku) v kombinaci s apriorními znalostmi či informacemi je významným předpokladem pro korektní nahrazování chybějících údajů imputovanými daty.

RESUME

Missing data is especially a problem for analyses using multivariate statistical methods. Although the incomplete data in the random sample may be represented by a relatively small percentage, this situation may result in a relatively very small set of complete data in the observed 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. Removing units from analyses can have significant negative effects - reducing the accuracy of estimates and the power of performed statistical tests - and can lead to distorted results unsuitable for generalization to the target population.
Bayesian modeling-random sampling from the posterior distribution, Monte Carlo simulation, and Markov chains-forms a solid foundation for almost the entire category of parametric imputation methods. Although the most important factor under incomplete data conditions is the mechanism of missing value, which has the most significant impact on the success of the various methods for working with the missing values, knowledge and correct application of statistical modelling methods based on a given specific sample, combined with a priori knowledge or information, is an important prerequisite for a correct replacement of missing data with the imputed data.

PROFESIJNÝ ŽIVOTOPIS

***Ing. Roman Pavelka, PhD**., v letech 1995 – 2010 pracoval v poradenské společnosti Trexima, s. r. o. Na pozici statistik – analytik se zabýval analýzami zejména mzdových a personálních dat. Podílel se na tvorbě pravidelných statistických přehledů a reportů. Spolupracoval s akademickými pracovišti, agenturami i soukromými subjekty na realizaci a vyhodnocování ad hoc statistických výzkumů. Oblast jeho vědeckého zájmu představují výběrová šetření, odhady a statistické modely. V letech 2012 až 2013 se zúčastnil zahraniční stáže ve Velké Británii. Od roku 2013 působil v Národnom ústave certifikovaných meraní vzdelávania (NÚCEM), kde zajišťoval statistické vyhodnocování výsledků testování žáků a studentů. Od roku 2015 pracuje v odboru metod statistických zjišťování Štatistického úradu SR.*

KONTAKT

roman.pavelka@statistics.sk