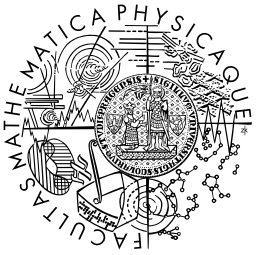


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Viktor Hržič

Odhady parametrov latentného rozdelenia pre ordinálne dáta

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D.

Studijní program: Finančná matematika

Studijní obor: Finančná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

V prvom rade by som chcel vyjadriť veľké poďakovanie mojej vedúcej bakalárskej práce RNDr. Šárke Hudecovej, Ph.D. za jej čas venovaný mne a najmä za výber veľmi zaujímavej témy a následné nespočetné množstvo cenných rád, ktoré mi pomohli túto tému spracovať. Ďalej by som sa chcel poďakovať rodičom, ktorí mi boli oporou vo všetkých možných formách počas celého štúdia. A v neposlednom rade patrí jedno ďakujem aj spolužiakovi Martinovi Romaňákovi, ktorý mi v prípravách na skúšky pomohol natoľko, že aj vďaka nemu mám teraz vôbec možnosť písať bakalársku prácu.

Název práce: Odhady parametrov latentného rozdelenia pre ordinálne dáta

Autor: Viktor Hržič

Katedra: Katedra pravdepodobnosti a matematickej štatistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravdepodobnosti a matematickej štatistiky

Abstrakt: Hlavným cieľom práce je predstavenie problematiky ordinálnych dát a najmä následne vytvorených odhadov latentného rozdelenia hustoty. Takto získané odhady sú porovnávané v simulačnej štúdii s inými metódami tvorby odhadov, ktoré sú v praxi používané častejšie. Čitateľ bude oboznámený s metódou maximálnej vierohodnosti, ktorá bola využitá pri tvorbe jednotlivých odhadov. Jedna časť je venovaná aj alternatívnemu prístupu využitím Bernsteinových polynómov. Bolo zistené, že s prihliadnutím na benefity, ktorými sú bezpochyby jednoduchší zber dát v ordinalizovanej podobe a minimalizácia chýb ktorých sa môže dopustiť respondent, sme dostali veľmi kvalitné a hodnotné výsledky.

Klíčová slova: Odhady parametrov, parametrické modely, metóda maximálnej vierohodnosti, ordinálne dáta, latentné rozdelenie

Title: Estimation of latent distribution for ordinal data

Author: Viktor Hržič

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D.

Abstract: The main goal of the bachelor thesis is to introduce problematics of ordinal data together with estimations of latent density distribution based on ordinal data. The estimations obtained using ordinal data are compared to the ones that are more common and used on daily basis. The reader will be introduced to the maximum likelihood method that is used in the development of each estimate. One chapter is dedicated to alternative approach using Bernstein polynomials. When we take all benefits into account such as easier collecting the data in ordinal form or minimization of possible errors committed by respondent, we obtained very valuable and promising results.

Keywords: Estimation of parameters, parametric models, maximum likelihood estimation, ordinal data, latent distribution

Obsah

Úvod

Pri zbieraní dát sa častokrát uvažujú iba kategórie, do ktorých spadá dané číslo alebo hodnota. Napríklad, pri platoch zamestnancoch je možné získavať údaje v určitých intervaloch, či už z dôvodov obmedzení plynúcich z pracovnej zmluvy alebo z osobnej preferencie ľudí uviesť iba ich platovú triedu. Ak sú tieto kategórie navyše aj zoradené, jedná sa o ordinálne dáta.

Je správne a prirodzené tušiť, že takto získané údaje strácajú časť pôvodnej informácie. Komplikácie navyše prichádzajú, ak dĺžky jednotlivých intervalov nie sú rovnaké. Prípadne, ak hustota skúmanej veličiny podlieha anomálii, ktorá je ťažko zistiteľná, pokiaľ máme dostupné iba ordinálne dáta.

V prvej kapitole sa okrem základných definícií parametrických modelov pozrieme hlbšie na niektoré detaily nášho problému ilustrované na konkrétnom príklade a obrázkoch. A taktiež bude objasnený cieľ práce: hľadanie distribučnej funkcie, prípadne hustoty latentného rozdelenia.

V úvode druhej kapitoly bude v krátkosti vysvetlený princíp metódy maximálnej vierohodnosti a uvedené vlastnosti takto získaných odhadov. Neskôr sa prejde ku konkrétnym odhadom parametrov potrebných k popisu pravdepodobnostného rozloženia latentného rozdelenia. Ďalej budú ukázané čiastočne vyjadrené rovnice pre numerické počítanie využité v nasledujúcej kapitole. Na záver bude vysvetlený ešte jeden odlišný prístup odhadovania latentného rozdelenia a to konkrétne aproximáciou zmesi rôznych beta rozdelení.

Tretia a zároveň posledná kapitola nás prevedie simulačnou štúdiou, kedy sa bude pracovať s nami nasimulovanými dátami, o ktorých budeme vedieť z presne akého rozdelenia pochádzajú. Následne budú ordinalizované a s ich využitím získame potrebné odhady parametrov jednotlivých rozdelení. Výsledky budú porovnávané s metódami, ktoré pracujú priamo so spojitými dátami. Na konci tejto tretej kapitoly bude aplikovaná alternatívna metóda popísaná na konci druhej kapitoly.

Vlastný prínos spočíva v podrobnom spracovaní danej tematiky, detailnom rozpísaní niektorých vzťahov a postupov. Ďalej v doplnení motivačného príkladu a predovšetkým v implementácií predstavených algoritmov v prostredí štatistického programu a následného vyhodnotenia výsledkov simulačnej štúdie.

1. Základné pojmy a oboznámenie sa s problémom

Túžba po stále dokonalejšom poznaní nie je opomenutá ani v kontexte riešenia problému s ordinálnymi dátami. Ľudom častokrát nestačí vedieť iba to, s akou pravdepodobnosťou skončia ich údaje v niektorej z predom danej kategórii. Nástroje a metódy k získaniu ďalších informácií nám poskytnú nasledujúce odstavce a kapitoly.

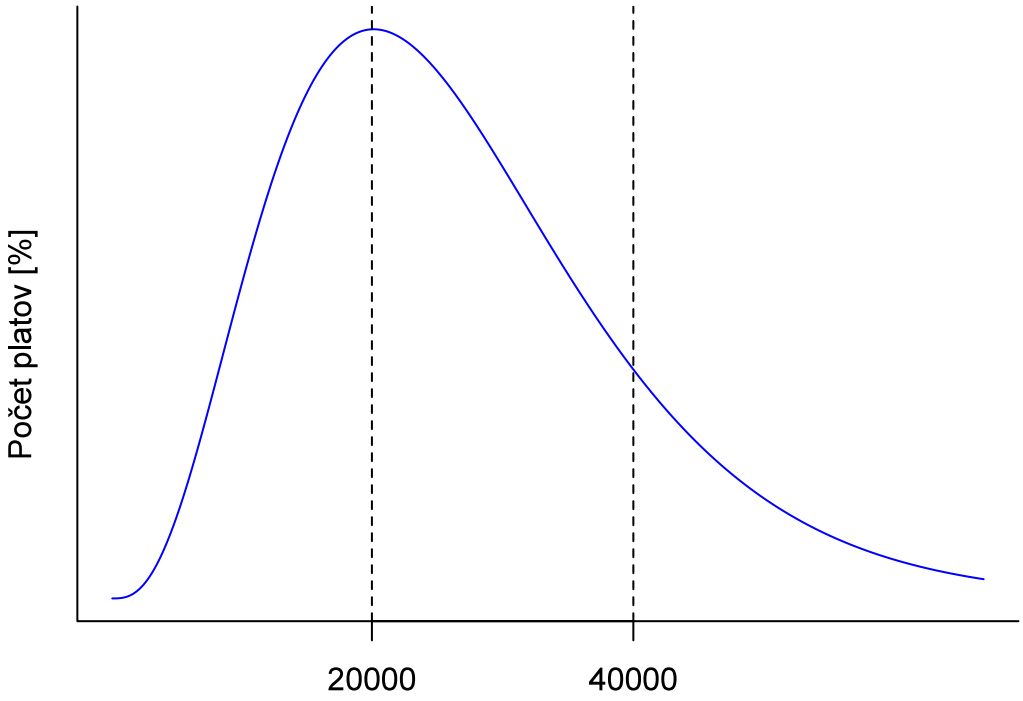
1.1 Podrobnosti problému

V úvode boli naznačené niektoré nedostatky ordinálnych dát. Vrátime sa ešte na chvíľu k situácii s platmi zamestnancov. Presné, avšak nám nie úplne známe, hodnoty platov budú chápané ako náhodný výber $X_1, \ldots, X_n$. To, že nám nie sú úplne presne známe, znamená, že máme k dispozícii iba ich ordinalizované hodnoty. Formálne sa dá vzťah medzi ordinálnou premennou Y a latentnou spojitou premennou X popísať nasledovnou rovnosťou:

$$Y = \sum_{k=1}^{J} k \cdot \mathbb{1}(m_k > X > m_{k-1})$$

a $-\infty \leq m_0 < m_1 < \cdots < m_{J-1} < m_J \leq \infty$ je postupnosť známych medzných bodov, ktoré rozdeľujú definičný obor náhodnej veličiny X do J kategórii. Nech $Y_1, \ldots, Y_n$ je teda n nezávislých, rovnako rozdelených ordinálnych pozorovaní z latentného rozdelenia premennej X, ktorého distribučnú funkciu F_X a hustotu f_X sa snažíme odhadnúť. Na obrázku 1.1 s tromi kategóriami ($J = 3$), predstavuje modrá krivka práve presný popis hustoty latentného rozdelenia náhodnej veličiny X. Krivka prirodzene nespĺňa formálne matematické vlastnosti hustoty, slúži skôr na ilustráciu problému. Pozorovaná veličina Y nám hovorí, že do ktorej kategórie dané pozorovanie spadlo. Má multinomické rozdelenie s J kategóriami a vieme teda pomocou relatívnych početností povedať intuitívne a relatívne presne s akou pravdepodobnosťou bude nejaký plat vyšší ako 40 tisíc korún. Avšak dáta nám neposkytujú takmer žiadnu informáciu o tom, aká bude pravdepodobnosť, že nejaký náhodne vybraný plat bude väčší ako 60 tisíc. Práve k tomuto slúžia odhady parametrov latentného rozdelenia, ktoré nevidíme priamo z dát, ale v istom zmysle nám poskytuje kompletnejšiu informáciu o rozdelení skúmanej veličiny.

Ale na to, aby bolo možné odhadovať konkrétne parametre, je potrebné predpokladať o našich napozorovaných kategorizovaných dátach, že spĺňajú určitý model zo svojej povahy. Pri výbere modelu sa zohľadňujú informácie ako nenulovosť hodnôt náhodnej veličiny, symetria rozdelenia alebo jeho predpokladaná šikmosť, či špicatosť. Pokiaľ tieto informácie nemáme k dispozícii, je vhodné sa obrátiť na odborníkov z danej oblasti alebo použiť na rozhodnutie spracované databázy s podobným charakterom a zameraním z minulosti.



Obr. 1.1: Ilustratívne rozdelenie platov s použitím rozdelenia Γ s preškálovanými hodnotami na x-ovej osi

1.2 Základné definície a modely

Nasledujúce definície predstavujú množinu možných modelov pre popis latentného rozdelenia skúmaného počas celej práce.

Definícia 1. *Ak náhodná veličina X má normálne rozdelenie, tak použijeme označenie $X \sim \mathcal{N}(\mu, \sigma^2)$, kde $\mu \in \mathbb{R}$ je stredná hodnota a $\sigma > 0$ je smerodajná odchýlka. Rozdelenie náhodnej veličiny X je definované pomocou hustoty, ktorá má tento tvar:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \text{ pre } x \in \mathbb{R}.$$

Špeciálnym prípadom je normované normálne rozdelenie, kedy $\mu = 0$ a $\sigma = 1$.

Pod špeciálnym označením normovaného normálneho rozdelenia definovaného v 1 využijeme neskôr nasledujúce označenie pre hustotu a hodnoty distribučnej funkcie:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right],$$
$$\Phi(z) = \int_{-\infty}^{z} \phi(u)\,\mathrm{d}u \quad \text{pre} -\infty < z < \infty,$$

kde hodnoty distribučnej funkcie $\Phi(z)$ je nutné počítať numericky a hodnoty sú tabulované.

Z definícií 1 a špeciálneho prípadu v podobe normovaného normálneho rozdelenia plynú nasledovné vzťahy pre distribučnú funkciu a hustotu normálneho rozdelenia $\mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = P\left[X \leq x\right] = P\left[Z \leq \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right) \text{ pre } x \in \mathbb{R},$$
$$f_X(x) = F'_X(x) = \phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \text{ pre } x \in \mathbb{R}.$$

Definícia 2. *Ak má náhodná veličina X normálne rozdelenie so strednou hodnotou $\mu \in \mathbb{R}$ a rozptylom $\sigma^2 > 0$, potom náhodná veličina $Y = e^X$ má logaritmicko-normálne rozdelenie. Budeme používať označenie $Y \sim \mathcal{LN}(\mu, \sigma^2)$.*

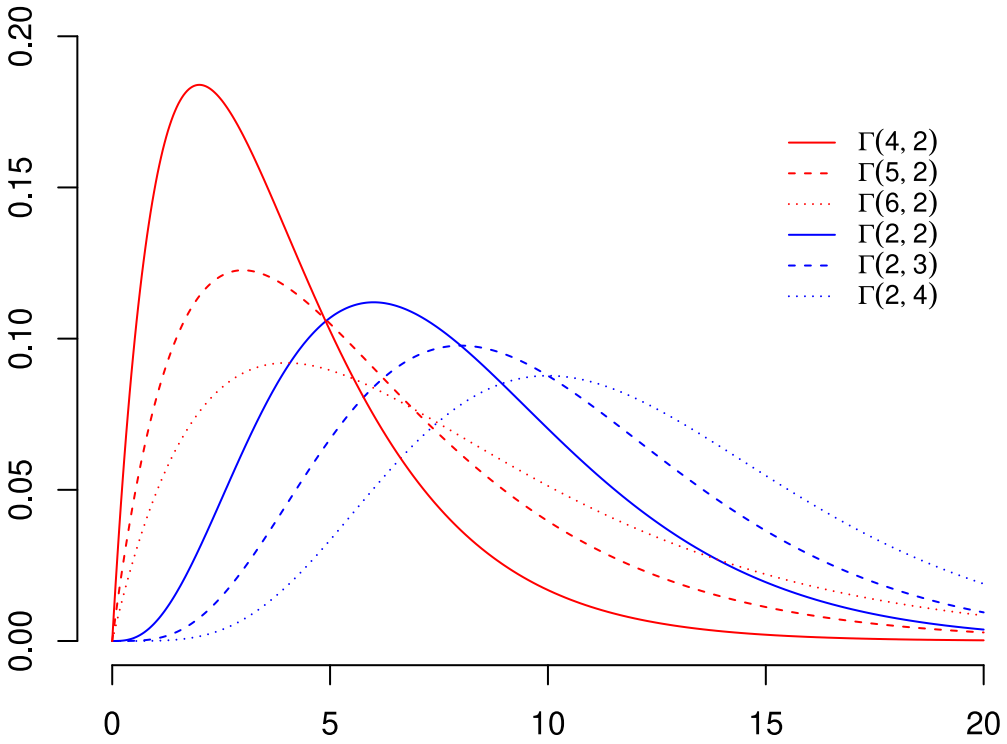
Z definície a vety o transformácií náhodnej veličiny plynie, že hustota náhodnej veličiny Y, ktorá má logaritmicko-normálne rozdelenie, je daná ako:

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma y} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right], & \text{pre } y > 0, \\ 0, & \text{inak.} \end{cases}$$

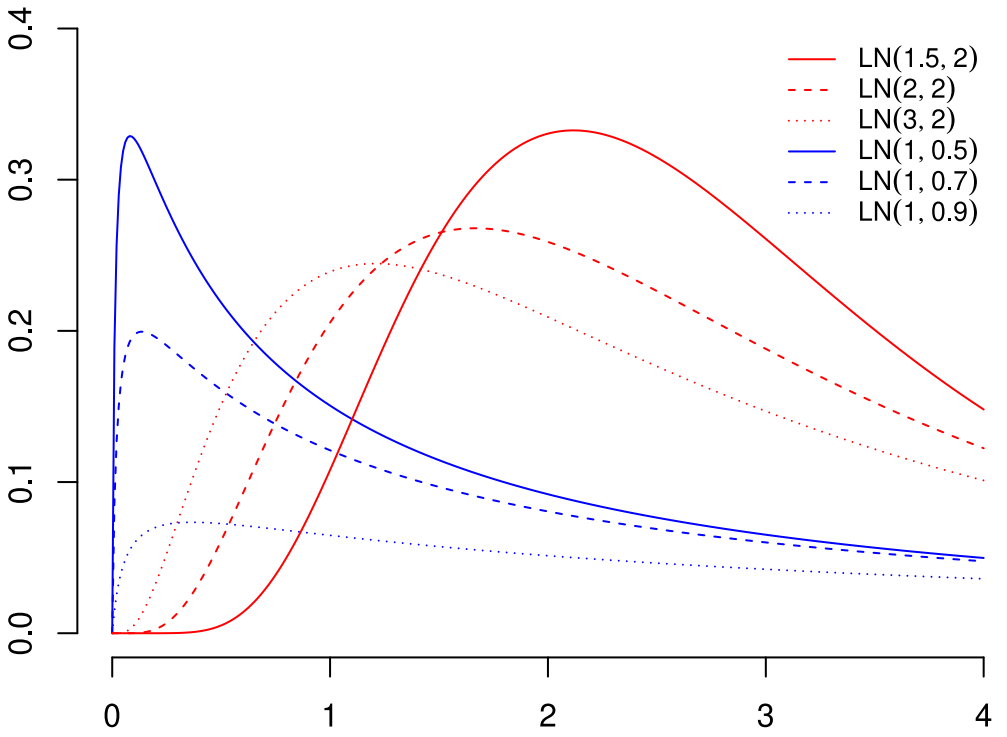
Definícia 3. *Označením $X \sim \Gamma(\alpha, \beta)$ budeme rozumieť rozdelenie náhodnej veličiny X, ktorá má Gama rozdelenie, kde α je parameter tvaru a β je parameter škály. Hustota tohto rozdelenia je daná nasledovne:*

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{pre } x > 0, \\ 0, & \text{inak.} \end{cases}$$

Obrázky 1.2 a 1.3 demonštrujú vplyv jednotlivých parametrov v gama a logaritmicko-normálnom rozdelení. V oboch prípadoch je najprv jeden parameter fixný a mení sa hodnota toho druhého. Pri zmene farby sa vymení aj situácia a fixným sa stáva druhý parameter.



Obr. 1.2: Hustota $\Gamma(\alpha, \beta)$ rozdelenia pre rôzne voľby parametrov α a β



Obr. 1.3: Hustota $\mathcal{LN}(\mu, \sigma^2)$ rozdelenia pre rôzne voľby parametrov μ a σ^2

2. Odhady parametrov latentného rozdelenia

Majme náhodný výber $X_1, \ldots, X_n$, modelom $\mathcal{F}$ budeme rozumieť vopred stanovenú množinu rozdelení, kam patrí aj rozdelenie, z ktorého pochádzajú naše dáta. Toto rozdelenie je charakterizované parametrami. Symbolom Θ budeme ďalej označovať množinu všetkých prípustných hodnôt parametra $\boldsymbol{\theta}$. Ak $F_X \in \mathcal{F}$ je skutočné rozdelenie, z ktorého pochádza náhodný výber a $\boldsymbol{\theta_X}$ je skutočná hodnota hľadaného parametra, tak potom jeho odhadom rozumieme ľubovoľnú merateľnú funkciu dát $\widehat{\boldsymbol{\theta}} \equiv T(X_1, \ldots, X_n)$.

A teraz je namieste si povedať zopár slov všeobecne k maximálne vierohodným odhadom a uviesť niektoré užitočné vlastnosti sformulované v tvrdeniach uvedených neskôr. Pracujeme s náhodným výberom z rozdelenia $f(x, \boldsymbol{\theta}_X)$ vzhľadom k σ-konečnej miere μ a parametrickým modelom:

$$\mathcal{F} = \{\text{rozdelenie s hustotou } f(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\},$$

kde d je dĺžka vektora $\boldsymbol{\theta}$. Náhodný výber $X_1, \ldots, X_n$ má združenú hustotu v tvare $\prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}_X)$. Ďalej máme vierohodnostnú funkciu:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i, \boldsymbol{\theta}).$$

Maximálne vierohodný odhad $\widehat{\boldsymbol{\theta}}$ parametra $\boldsymbol{\theta}$ je taký prvok z Θ, ktorý maximalizuje vierohodnostnú funkciu, napočítanú v pozorovaných hodnotách $X_1, \ldots, X_n$. Hľadáme teda:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

Vyššie uvedenú maximalizačnú úlohu typicky riešime tak, že logaritmickú vierohodnosť parciálne derivujeme podľa zložiek $\boldsymbol{\theta}$ a položíme rovnú nule, a následne overíme, či sa skutočne jedná o lokálny extrém. Rovnaké postupy a vlastnosti budú platiť aj pri ostatných maximálne vierohodných odhadoch spomenutých v práci. Uvedené tvrdenia a definície vychádzajú z Anděl (2007).

Definícia 4. *Nech náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ má hustotu $f(\boldsymbol{x}, \boldsymbol{\theta})$ vzhľadom k nejakej $\sigma-$konečnej miere μ. Predpokladajme, že platí:*

(A) $\boldsymbol{\theta} \in \Theta$, kde Θ je neprázdna otvorená množina v $\mathbb{R}^d$.

(B) Množina (nosič) $M = \{\boldsymbol{x} : f(\boldsymbol{x}, \boldsymbol{\theta}) > 0\}$ nezávisí na $\boldsymbol{\theta}$.

(C) Pre skoro všetky $\boldsymbol{x} \in M$ vzhľadom k μ a pre všetky $i = 1, \ldots, d$ existujú parciálne derivácie $f_i'(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i}$.

(D) Pre každé i a pre všetky $\boldsymbol{\theta} \in \Theta$ platí $\int_M f_i'(\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{x}) = 0$.

(E) Pre každú dvojicu (i, j) existuje konečný integrál

$$J_{ij}(\boldsymbol{\theta}) = \int_M \frac{f_i'(\boldsymbol{x}, \boldsymbol{\theta}) f_j'(\boldsymbol{x}, \boldsymbol{\theta})}{f^2(\boldsymbol{x}, \boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{x}).$$

(F) Matica $\boldsymbol{J}_n(\boldsymbol{\theta}) = \|J_{ij}(\boldsymbol{\theta})\|_{i,j=1}^m$ *je pozitívne definitná pre každé* $\boldsymbol{\theta} \in \Theta$.

Potom sa systém hustôt $\{f(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ *nazýva regulárny a* $\boldsymbol{J}_n(\boldsymbol{\theta})$ *sa nazýva* Fisherova informačná matica.

K predpokladom z predchádzajúcej definície budeme ešte navyše považovať za splnené dve podmienky regularity.

(1) Nech $\Theta \subset \mathbb{R}^d$ je parametrický priestor, ktorý obsahuje taký neprázdny otvorený interval ω, že skutočná hodnota parametra $\boldsymbol{\theta}_0$ patrí do ω.

(2) Nech $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Potom $f(x, \boldsymbol{\theta}_1) = f(x, \boldsymbol{\theta}_2)$ $[\mu]$ (skoro všade vzhľadom k μ) platí práve vtedy, ak je $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Veta 1 (Vlastnosti maximálne vierohodného odhadu)**.** *Nech je systém hustôt* $\{f(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ *regulárny a má Fisherovu maticu* $\boldsymbol{J}(\boldsymbol{\theta})$. *Nech platia podmienky regularity (1), (2) a navyše sú splnené nasledujúce predpoklady.*

(A) Derivácia $\frac{\partial^3 f(x,\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k}$ *existuje pre skoro všetky* x, *pre všetky* $\boldsymbol{\theta} \in \omega$ *a pre všetky* $i, j, k = 1, \ldots, d$.

(B) Pre všetky $\boldsymbol{\theta} \in \omega$ *platí*

$$\int_M f''_{ij}(x, \boldsymbol{\theta})\,\mathrm{d}\mu(x) = 0, \quad i, j = 1, \ldots, d.$$

(C) Pre všetky $i, j, k = 1, \ldots, d$ *existujú funkcie* $M_{ijk}(x) \geq 0$ *tak, že*

$$\mathrm{E}_{\boldsymbol{\theta}_0} M_{ijk}\,(X) < \infty$$

a

$$\left|\frac{\partial^3 \ln f(x, \boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k}\right| \leq M_{ijk}(x) \quad \textit{pre všetky } \boldsymbol{\theta} \in \omega \textit{ a skoro všetky } x \in M.$$

Potom platia nasledujúce tvrdenia.

(i) Pokiaľ $n \to \infty$, *potom ku každému* $\epsilon > 0$ *existuje s pravdepodobnosťou blížiacou sa k jednej také riešenie* $\widehat{\boldsymbol{\theta}}_n$ *systému vierohodnostných rovníc, že* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| < \epsilon$.

(ii) Položme

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial L(\boldsymbol{\theta})}{\partial\theta_1} \\ \ldots\ldots \\ \frac{\partial L(\boldsymbol{\theta})}{\partial\theta_d} \end{pmatrix}.$$

Potom pre $n \to \infty$ *platí*

$$\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}\left[\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0)\right].$$

(iii) Ak existuje pre každé dostatočne veľké n a pre každú hodnotu $\boldsymbol{X}$ taký koreň $\widehat{\boldsymbol{\theta}}_n$ systému vierohodnostných rovníc, že $\widehat{\boldsymbol{\theta}}_n$ je konzistentným odhadom parametra $\boldsymbol{\theta}_0$, potom

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}\right) \xrightarrow{D} \mathrm{N}\left(\mathbf{0},\, [\boldsymbol{J}(\boldsymbol{\theta}_0)]^{-1}\right).$$

Dôkaz. Jednotlivé kroky dôkazu sú podrobne popísané v kapitole 6.4 knihy Lehmann (1983).

□

Toto tvrdenie nám dáva informáciu o konzistentnosti a asymptotickej normalite maximálne vierohodných odhadov. Metódu maximálnej vierohodnosti budeme aplikovať pri hľadaní odhadov parametrov distribučnej funkcie a hustoty latentného rozdelenia.

2.1 Prvotné odhady distribučnej funkcie

Pripomeňme proces ordinalizovania pozorovaní veličiny X na Y:

$$Y = j \Leftrightarrow m_{j-1} < X \leq m_j.$$

Pravdepodobnosť náležania pozorovania j-tej kategórii bude označovaná ako p_j. V zmysle distribúcie budeme používať $p_j = P[Y = j]$ a ak navyše k tomu je $F_X(x) = P[X \leq x]$ distribučná funkcia, potom platí:

$$p_j = F_X(m_j) - F_X(m_{j-1}). \tag{2.1}$$

Pre jednoduchosť zápisu položíme: $F_X(m_0) = 0 = 1 - F_X(m_J)$. Kedy si nultý medzný bod predstavíme na začiatku nosiča skúmanej veličiny a posledný na jeho konci.

Na intuitívnej úrovni budeme tušiť, že pôvodná distribučná funkcia by sa dala bez parametrizácie odhadovať iba v medzných bodoch, pretože z povahy dát nemáme informáciu o tom, akému pravdepodobnostnému rozloženiu podliehajú dáta vnútri jednotlivých kategórii. Ako bolo spomenuté v prvej kapitole, musíme usúdiť, že distribučná funkcia F_X spĺňa nejaký parametrický model, napríklad v zmysle definícii 1 až 3. Tento fakt budeme značiť nulou v dolnom indexe. Budeme teda predpokladať, že $F_X(x)$ patrí do rodiny rozdelení $F_0(x, \boldsymbol{\theta})$ pre $\boldsymbol{\theta} \in \Theta$. V takomto modeli sa dá vyjadriť ľubovoľná charakteristika rozdelenia (stredná hodnota, rozptyl, ...) ako funkcia zložiek $\boldsymbol{\theta}$.

2.2 Maximálne vierohodný odhad p_j

Nech vektor $\boldsymbol{f}$ vznikol z diskretizovanej náhodnej veličiny Y a tvoria ho početnosti jednotlivých kategórií. Pod označením f_j budeme rozumieť počet prvkov v j-tej kategórii. Teda dostávame $f_j = \sum_{i=1}^{n} \mathbb{1}(Y_i = j)$, pričom platí $\sum_{k=1}^{J} f_k = n$. K dispozícií máme práve tento vektor početností jednotlivých kategórií $\boldsymbol{f} = (f_1, f_2, \ldots, f_J)^\top$. Pripomeňme, že pravdepodobnosť toho, že jedno pozorovanie z pôvodného náhodného výberu spadne do j-tej kategórie, sme označili ako p_j,

pre ktorú prirodzene platí $\sum_{k=1}^{J} p_k = 1$. Vektor $\boldsymbol{p}$ je teda vektor pravdepodobností z čoho plynie, že $p_j \in [0,1]$ pre $1, 2, \ldots, J$.

Teraz sa pozrieme na odhady pravdepodobností jednotlivých kategórií, ktoré plynú z toho, že $\boldsymbol{f}$ má multinomické rozdelenie. Matematickým zápisom značené ako $\boldsymbol{f} \sim Mult(n, \boldsymbol{p})$, kde n je počet pozorovaní. Združená pravdepodobnosť náhodného vektora $\boldsymbol{f}$ vzhľadom k súčinovej počítacej miere na $\mathbb{Z}^J$, ktorá nám hovorí o tom, koľko z n pozorovaní spadlo do ktorej kategórie, bude vyzerať nasledovne:

$$\begin{aligned}
&\binom{n}{f_1}\binom{n-f_1}{f_2}\cdots\binom{n-f_1-\ldots-f_{J-1}}{f_J} p_1^{f_1}\cdot p_2^{f_2}\cdot\ldots p_J^{f_J} = \\
&= \frac{n!}{f_1!\ldots f_J!} p_1^{f_1}\cdot p_2^{f_2}\cdot\ldots p_J^{f_J} = n!\prod_{k=1}^{J}\frac{p_k^{f_k}}{f_k!},\ \text{kde}\sum_{k=1}^{J} f_k = n,\ f_k \in \mathbb{N}_0\,\forall\, k.
\end{aligned}$$

Zlogaritmovaním posledného výrazu dostaneme logaritmickú vierohodnosť multinomického rozdelenia pre jednotlivé kategórie postupne v tvare:

$$\begin{aligned}
\ell^*(\boldsymbol{p}) &= \log(n!) + \log\left(\prod_{k=1}^{J}\frac{p_k^{f_k}}{f_j!}\right) \\
&= \log(n!) + \sum_{k=1}^{J}\log\left(\frac{p_k^{f_k}}{f_k!}\right) \\
&= \log(n!) + \sum_{k=1}^{J} f_k \log\left(p_k\right) - \sum_{k=1}^{J}\log\left(f_k!\right). \qquad (2.2)
\end{aligned}$$

Na nájdenie maxima danej funkcie použijeme Lagrangeove multiplikátory. Postup bude spočívať v tom, že si vytvoríme novú funkciu, nazývanú Lagrangeova. Tá sa bude skladať z pôvodnej logaritmickej vierohodnosti a zároveň bude obsahovať obmedzenie týkajúce sa súčtu pravdepodobností. Konkrétne, členy v rovnosti $\sum_{k=1}^{J} p_k = 1$ presunieme na jednu stranu, položíme rovné nule a prenásobíme -1, aby krajšie vychádzali znamienka. Dostávame tak Lagrangeovu funkciu v tvare:

$$\mathscr{L}(\boldsymbol{p}, \lambda) = \ell^*(\boldsymbol{p}) + \lambda\left(1 - \sum_{k=1}^{J} p_k\right).$$

Aj s pomocou parciálnych derivácií p_j pre $j = 1, \ldots, J-1$ budeme hľadať $\arg\max_{\boldsymbol{p}} \mathscr{L}(\boldsymbol{p}, \lambda)$. Po kratšom výpočte:

$$\begin{aligned}
\frac{\partial}{\partial p_j}\mathscr{L}(\boldsymbol{p}, \lambda) = \frac{\partial}{\partial p_j}\ell^*(\boldsymbol{p}) + \frac{\partial}{\partial p_j}\lambda\left(1 - \sum_{k=1}^{J} p_k\right) = 0 \\
\frac{\partial}{\partial p_j}\sum_{k=1}^{J} f_k \cdot \log p_k - \lambda\frac{\partial}{\partial p_j}\sum_{k=1}^{J} p_k = 0.
\end{aligned}$$

Deriváciu súčtu získame ako súčet derivácií, pričom nenulová derivácia je len

pri sčítacom indexe $k = j$, odkiaľ dostávame ďalšie kroky výpočtu:

$$\begin{aligned} \frac{f_j}{p_j} - \lambda &= 0, \\ p_j &= \frac{f_j}{\lambda}, \\ \sum_{k=1}^{J} p_k &= \sum_{k=1}^{J} \frac{f_k}{\lambda}, \\ 1 &= \frac{1}{\lambda} \sum_{k=1}^{J} f_k, \\ \lambda &= n. \end{aligned} \tag{2.3}$$

Dosadením poslednej rovnosti do (2.3) dostávame záver v podobe maximálne vierohodného odhadu $\widehat{p}_j = \frac{f_j}{n}$. Našli sme teda maximálne vierohodný odhad parametra $\boldsymbol{p}$, ktorý spĺňa vlastnosti ako konzistentnosť, či asymptotická normalita, ktoré plynú z vety 1. Avšak, naším pôvodným cieľom bolo nájsť odhad latentného rozdelenia F_X. Práve tomu sa bude venovať nasledujúca kapitola.

2.3 Maximálne vierohodný odhad parametru $\boldsymbol{\theta}$

V tejto kapitole budeme čiastočne vychádzať z Ghosh a kol. (2018). Upravíme logaritmickú vierohodnosť z (2.2) a to tak, že časť neobsahujúca argumenty p_j, ktorá sa dá považovať za konštantu, bude vynechaná z procesu hľadania maximálne vierohodného odhadu. Týmto spôsobom obdržíme logaritmickú vierohodnosť v tvare:

$$\ell^*(\boldsymbol{p}) = \underbrace{\log\left(\frac{n!}{\prod_{k=1}^{J} f_k!}\right)}_{K} + \sum_{k=1}^{J} f_k \log p_k. \tag{2.4}$$

Máme vierohodnostnú funkciu pre parametrickú verziu problému a potom logaritmickú vierohodnosť použitím parametrizovaného vzťahu (2.1) dosadeného do (2.4) v zmysle $\ell(\boldsymbol{\theta}) = \ell^*(\boldsymbol{p}(\boldsymbol{\theta}))$, kde $p_j(\boldsymbol{\theta}) = F_X(m_j, \boldsymbol{\theta}) - F_X(m_{j-1}, \boldsymbol{\theta})$ Potom už uvažujeme vierohodnosť po vynechaní konštanty K:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} \left[F_0(m_j, \boldsymbol{\theta}) - F_0(m_{j-1}, \boldsymbol{\theta})\right]^{f_j}, \text{kde stále } f_j = \sum_{i=1}^{n} \mathbb{1}(Y_i = j),$$

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{j=1}^{J} f_j \log\left[F_0(m_j, \boldsymbol{\theta}) - F_0(m_{j-1}, \boldsymbol{\theta})\right].$$

Je nutné vychádzať z konkrétnych pravdepodobnostných modelov pre veličinu X. Postupne pre normálne, logaritmicko-normálne a gamma rozdelenie máme:

$$\begin{aligned} F_N &= \left\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\right\}; \boldsymbol{\theta} = (\mu, \sigma^2)^\top, \Theta = \mathbb{R} \times \mathbb{R}^+, \\ F_{LN} &= \left\{\mathcal{LN}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\right\}; \boldsymbol{\theta} = (\mu, \sigma^2)^\top, \Theta = \mathbb{R} \times \mathbb{R}^+, \\ F_G &= \left\{\Gamma(\alpha, \beta), \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+\right\}; \boldsymbol{\theta} = (\alpha, \beta)^\top, \Theta = \mathbb{R}^+ \times \mathbb{R}^+. \end{aligned}$$

Naše tri potenciálne modely sú definované práve dvomi parametrami, z čoho plynie, že bude treba riešiť sústavu dvoch rovníc. Rovnaká myšlienka sústavy dvoch rovníc bola použitá v (Tamhane a kol. (2002)). Existencia parciálnych derivácií plynie zo splnenia podmienok regularity spomenutých v úvode druhej kapitoly. Zovšeobecnenie pre viacparametrické modely by sa robilo analogicky.

$$\frac{\partial}{\partial\theta_1}\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} \frac{f_j}{F_0(m_j,\boldsymbol{\theta}) - F_0(m_{j-1},\boldsymbol{\theta})} \frac{\partial\left(F_0(m_j,\boldsymbol{\theta}) - F_0(m_{j-1},\boldsymbol{\theta})\right)}{\partial\theta_1} = 0 \quad (2.5)$$

$$\frac{\partial}{\partial\theta_2}\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} \frac{f_j}{F_0(m_j,\boldsymbol{\theta}) - F_0(m_{j-1},\boldsymbol{\theta})} \frac{\partial\left(F_0(m_j,\boldsymbol{\theta}) - F_0(m_{j-1},\boldsymbol{\theta})\right)}{\partial\theta_2} = 0 \quad (2.6)$$

Pred tým, ako budú vyjadrené rovnice pre jednotlivé parametre rozdelení z prvej kapitoly, je potrebné si definovať pojmy ako chybová funkcia, jej vzťah k distribučnej funkcii normálneho a tým pádom aj logaritmicko-normálneho rozdelenia. Taktiež pribudne alternatívny zápis pre distribúciu gama rozdelenia využívajúci neúplnú gama funkciu.

Definícia 5. *Pod nasledovným označením budeme rozumieť chybovú funkciu:*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,\mathrm{d}t.$$

Nasledujúce tvrdenie je uvedené v literatúre Andrews (1998), jeho samostatne doplnený dôkaz vznikol iba okomentovaním dôkazu uvedeného v iných zdrojoch.

Tvrdenie 2. *Nech* $\Phi(x)$ *značí normované normálne rozdelenie a* erf *je chybová funkcia z definície 5, potom platí nasledujúci vzťah:*

$$\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right].$$

Dôkaz. Z definície 5 pri substitúcií $t^2 = z^2/2$ dostávame $t = z/\sqrt{2}$, pretože t nenadobúda záporné hodnoty, a teda $\mathrm{d}t = \mathrm{d}z/\sqrt{2}$. Z medzí $t = 0$ a $t = x$ sa stanú $z = 0$ a $z = x\sqrt{2}$. Rozdelením integrálu sa budeme snažiť nájsť niečo, čo pripomína distribučnú funkciu normovaného normálneho rozdelenia:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{2\pi}} \int_0^{x\sqrt{2}} e^{-\frac{z^2}{2}}\,\mathrm{d}z = 2\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x\sqrt{2}} e^{-\frac{z^2}{2}}\,\mathrm{d}z - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{z^2}{2}}\,\mathrm{d}z\right).$$

Integrály na pravej strane sú hodnoty distribučnej funkcie normovaného normálneho rozdelenia z definície 1. Dostávame teda:

$$\operatorname{erf}(x) = 2\left(\Phi(x\sqrt{2}) - \Phi(0)\right) = 2\left(\Phi(x\sqrt{2}) - \frac{1}{2}\right) = 2\Phi(x\sqrt{2}) - 1.$$

Algebraickými úpravami sa dostaneme k chcenému záveru:

$$\Phi(x) = \frac{1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)}{2}.$$

□

Alternatívnym vyjadrením distribučnej funkcie pre $\Gamma(\alpha, \beta)$ rozdelenie môže byť nasledujúca rovnosť:

$$F_X(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}, \tag{2.7}$$

kde v čitateli vystupuje neúplná gama funkcia definovaná ako:

$$\gamma(\alpha, \beta x) = \int_0^{\beta x} t^{\alpha-1} e^{-t} \, \mathrm{d}t.$$

Môžeme vidieť, že v rovnostiach (2.5) a (2.6) pre parametre rozdelení, je potrebné parciálne zderivovať distribučné funkcie a potom už len dosadiť postupnosť medzných bodov. Ako prvé sa pozrieme na normálne rozdelenie (model F_N), kde využijeme vlastnosť popísanú v tvrdení 2 a ihneď potom dostávame aj sústavu rovníc:

$$\begin{aligned}
F_X(x) &= \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] = \frac{1}{2} + \frac{1}{2} \cdot \frac{2}{\sqrt{\pi}} \int_0^{\frac{x-\mu}{\sigma\sqrt{2}}} e^{-t^2} \, \mathrm{d}t, \\
\frac{\partial F_X(x)}{\partial \mu} &= \frac{1}{\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(\frac{\partial}{\partial \mu} \frac{x-\mu}{\sigma\sqrt{2}}\right) = \frac{1}{\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(\frac{-1}{\sigma\sqrt{2}}\right) = \frac{-1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \\
\frac{\partial F_X(x)}{\partial \sigma} &= \frac{1}{\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(\frac{\partial}{\partial \sigma} \frac{x-\mu}{\sigma\sqrt{2}}\right) = \frac{1}{\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(\frac{-\sqrt{2}(x-\mu)}{2\sigma^2}\right) = \\
&= \frac{(\mu - x)}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.
\end{aligned}$$

V prípade logaritmicky-normálneho rozdelenia (model F_{LN}) by sa postupovalo takmer identicky. Jediný rozdiel by nastal v hornej medzi integrálu, kde by namiesto x vystupoval $\log x$. Ďalej budeme obdobným spôsobom derivovať aj distribučnú funkciu gama rozdelenia (model F_G) v tvare uvedenom v (2.7). V priebehu výpočtu bude využitý vzťah pre deriváciu gama funkcie, ktorý bol použitý v Li a Qin (2017):

$$\Gamma^{(n)}(\alpha) = \frac{\mathrm{d}^n}{\mathrm{d}\alpha^n} \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} \log^n t \, \mathrm{d}t \qquad \text{pre } n = 1,2,\ldots.$$

Ako bolo avizované, podobným spôsobom dostávame aj sústavu rovníc pre gama rozdelenie:

$$\begin{aligned}
\frac{\partial F_X(x)}{\partial \alpha} &= \frac{\gamma'(\alpha, \beta x) \cdot \Gamma(\alpha) - \gamma(\alpha, \beta x) \cdot \Gamma'(\alpha)}{(\Gamma(\alpha))^2} = \\
&= \frac{\int_0^{\beta x} t^{\alpha-1} e^{-t} \log t \, \mathrm{d}t \cdot \Gamma(\alpha) - \gamma(\alpha, \beta x) \cdot \int_0^{\infty} t^{\alpha-1} e^{-t} \log t \, \mathrm{d}t}{(\Gamma(\alpha))^2}, \\
\frac{\partial F_X(x)}{\partial \beta} &= \frac{1}{\Gamma(\alpha)} (\beta x)^{\alpha-1} e^{-\beta x} x.
\end{aligned}$$

Po dosadení týchto výrazov do (2.5) a (2.6) môžeme numericky riešiť dané sústavy rovníc. Pokiaľ sú splnené všetky predpoklady tvrdení, potom je získaný, maximálne vierohodný odhad parametra $\boldsymbol{\theta}$ konzistentný a asymptoticky normálny.

2.4 Aproximácia Bernsteinovými polynómami

V tejto časti budeme predpokladať, že krajné medze m_0 a m_J sú konečné. Metóda je teda vhodná pre rozdelenia s konečným nosičom. Túto vlastnosť nespĺňa žiadne z rozdelení spomenutých v prvej kapitole. Pritom táto požiadavka je z reálneho pohľadu celkom ľahko splniteľná. V situácií s výškami platov je pomerne náročné nájsť nekonečne kladný alebo záporný plat. Na druhej strane pevné stanovenia hornej platovej hranice môže byť problematické. Podkapitola bude vychádzať predovšetkým z článku Ghosh a kol. (2018), s doplnením detailov v niektorých častiach. Je potrebné si zaviesť všeobecne beta rozdelenie, ktoré bude neskôr používané.

Definícia 6. *Označením $X \sim \mathcal{B}(\alpha, \beta)$, kde $\alpha > 0$ a $\beta > 0$, budeme rozumieť rozdelenie náhodnej veličiny X, ktorá má beta rozdelenie. Hustota tohto rozdelenia je daná nasledovne:*

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{pre } x \in [0,1], \\ 0, & \text{inak.} \end{cases}$$

V menovateli výrazu z definície sa vyskytuje beta funkcia daná predpisom:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x,$$

pre ktorú platí pri prirodzených α a β identita:

$$B(\alpha, \beta) = \frac{\alpha+\beta}{\alpha\beta} \Big/ \binom{\alpha+\beta}{\alpha}. \tag{2.8}$$

Označme ako $\mathscr{B}(x, \alpha, \beta)$ distribučnú funkciu beta rozdelenia $\mathcal{B}(\alpha, \beta)$ v bode x a ako $b(x, \alpha, \beta)$ hustotu tohto rozdelenia. Ďalej budeme uvažovať zmes preškálovaných beta rozdelení:

$$F_0(x, \boldsymbol{\theta}) = \sum_{l=1}^{N} \theta_l \mathscr{B}\left(\frac{x - m_0}{m_J - m_0}, l, N - l + 1\right), \tag{2.9}$$
$$\text{kde } \theta_l \geq 0 \,\forall\, l \text{ a } \sum_{l=1}^{N} \theta_l = 1.$$

Uvedená funkcia $F_0(\,\cdot\,)$ je distribučnou funkciou pre $N \in \{2, 3, \dots\}$, $\boldsymbol{\theta} \in \Theta$, kde

$$\Theta = \left\{(\theta_1, \dots, \theta_N) \in [0,1]^N : \sum_{l=1}^{N} \theta_l = 1\right\}.$$

To, že je naozaj distribučnou funkciou plynie zo skutočnosti, že výsledná funkcia (2.9) vznikla konvexnou kombináciou distribučných funkcií beta rozdelení. Pre zvolené N dostávame ďalší parametrický model obsahujúci všetky možné zmesi beta rozdelení:

$$F_B^N = \left\{\sum_{l=1}^{N} \theta_l \mathscr{B}\left(\frac{x - m_0}{m_J - m_0}, l, N - l + 1\right); \boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^\top \in \Theta\right\}.$$

Vektor $\boldsymbol{\theta}$ v podstate predstavuje koeficienty alebo váhy využité z bázických funkcií daných rôznymi beta rozdeleniami. Distribučnej funkcii z (2.9) odpovedá hustota

$$f_0(x, \boldsymbol{\theta}) = \sum_{l=1}^{N} \theta_l b\left(\frac{x - m_0}{m_J - m_0}, l, N - l + 1\right) \frac{1}{m_J - m_0}. \tag{2.10}$$

S využitím vlastnosti uvedenej v (2.8) sa dá vyjadriť $b(u, l, N - l + 1)$ v tvare

$$b(u, l, N - l + 1) = N\binom{N-1}{l-1} u^{l-1}(1-u)^{N-l}\mathbb{1}(u \in [0, 1]).$$

Definícia 7. *Bernsteinov polynóm rádu n funkcie f(x) definovanej na uzavretom intervale [0, 1] je daný výrazom:*

$$B_n(x) = \sum_{v=0}^{n} f\left(\frac{v}{n}\right)\binom{n}{v} x^v(1-x)^{n-v}.$$

Veta 3. *Pre funkciu f(x) obmedzenú na [0, 1] platí vzťah*

$$\lim_{n\to\infty} B_n(x) = f(x)$$

v každom bode spojitosti x funkcie f. A navyše vzťah platí rovnomerne na [0, 1], ak je f(x) spojitá na tomto intervale.

Dôkaz. Jednotlivé kroky dôkazu je možné nájsť v knihe Lorentz (2012). □

Tvrdenie 4. *Nech f je spojitá hustota na intervale $[m_0, m_J]$, kde m_0 a m_J sú konečné reálne hodnoty. Ak položíme*

$$\widehat{\theta}_l = \frac{f(m_0 + (l-1)(m_J - m_0)/(N-1))}{\sum_{l=1}^{N} f(m_0 + (l-1)(m_J - m_0)/(N-1))},$$

tak platí rovnomerná konvergencia $f_0(x, \widehat{\boldsymbol{\theta}})$ k $f(x)$ pre $N \to \infty$. Vektor $\widehat{\boldsymbol{\theta}}$ je zložený z takto vyššie uvedených $\widehat{\theta}_l$.

Dôkaz. Bez ujmy na všeobecnosti budeme predpokladať $m_0 = 0$, $m_J = 1$. Dosadením $\widehat{\theta}_l$ do (2.10) dostaneme:

$$f_0(x, \widehat{\theta}) = \sum_{l=1}^{N} \frac{f\left(\frac{l-1}{N-1}\right)}{\sum_{i=1}^{N} f\left(\frac{i-1}{N-1}\right)} N\binom{N-1}{l-1} x^{l-1}(1-x)^{N-l}\mathbb{1}(x \in [0, 1]).$$

Označíme $v = l - 1$, $n = N - 1$ a výraz prepíšeme ako:

$$f_0(x, \widehat{\theta}) = \sum_{v=0}^{n} \frac{f\left(\frac{v}{n}\right)}{\sum_{i=0}^{N} f\left(\frac{i}{n}\right)} (n+1)\binom{n}{v} x^v(1-x)^n\mathbb{1}(x \in [0, 1]).$$

Tento výraz rozdelíme na dve časti nasledovne:

$$\underbrace{\frac{n+1}{\sum_{v=0}^{n} f\left(\frac{v}{n}\right)}}_{g_n} \underbrace{\sum_{v=0}^{n} f\left(\frac{v}{n}\right)\binom{n}{v} x^v(1-x)^n\mathbb{1}(x \in [0, 1])}_{Bernsteinov\ polynóm}.$$

Z predpokladu je $f(x)$ spojitá a podľa vety 3 postupnosť funkcií $B_n(x)$ konverguje rovnomerne k $f(x)$ pre $n \to \infty$. Postupnosť g_n je postupnosť konečných konštánt a konverguje k 1. Prepíšeme menovateľ výrazu g_n na:

$$\sum_{v=0}^{n} f\left(\frac{v}{n}\right) \frac{1}{n+1},$$

ktorý pre $n \to \infty$ konverguje k $\int_0^1 f(x)\,\mathrm{d}x$. Dostávame tak v podstate Riemannovský integrál hustoty na celom jej nosiči, ktorý sa z vlastností hustoty rovná 1. Z čoho dostávame požadovaný záver.

□

V našom prípade predstavuje funkcia f hľadanú hustotu latentného rozdelenia na danom konkrétnom nosiči. Neformálne povedané, pre N dostatočne veľké existuje funkcia z modelu F_B^N, ktorá bude hľadané f aproximovať dostatočne dobre.

Dostávame tak veľmi flexibilný nástroj na odhadovanie neznámej spojitej hustoty latentnej premennej X. Otázkou naďalej ostáva aké veľké N zvoliť a taktiež ako odhadnúť $\boldsymbol{\theta}$, keď f nepoznáme. Vo svojej práci Babu a kol. (2002) odporúča voľbu $N \in \{2, 3, \ldots [n/\log(n)]\}$. So zvyšujúcim sa N očakávame lepšiu presnosť aproximácie, ale zároveň klesá výpočetná stabilita.

Samotný výpočet odhadov parametra $\boldsymbol{\theta}$ vychádza z Anderson-Darlingovho štatistického testu. Ten je nástrojom na zistenie toho, či môžu byť dáta aproximované nejakým parametrickým rozdelením. Testová štatistika daného testu dáva indíciu k tomu, prečo budeme minimalizovať práve funkciu uvedenú nižšie. Spomínaný test meria vzdialenosť medzi empirickou distribučnou funkciou a funkciou spĺňajúcou parametrický model nasledovne:

$$n \int_{-\infty}^{\infty} \left(\widehat{F}(x) - F(x)\right)^2 w(x)\,\mathrm{d}F(x).$$

Pod označením $w(x)$ rozumieme nejakú váhovú funkciu a jej konkrétnou voľbou $w(x) = (F(x)(1 - F(x)))^{-1}$ dostávame štatistiku:

$$n \int_{-\infty}^{\infty} \frac{\left(\widehat{F}(x) - F(x)\right)^2}{F(x)(1 - F(x))}\,\mathrm{d}F(x).$$

Cieľom bude teda nájsť podľa Ghosh a kol. (2018):

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min} \sum_{j=1}^{J} \frac{\left(\widehat{F}(m_j) - F_0(m_j, \boldsymbol{\theta})\right)^2}{\left(\widehat{F}(m_j) + \epsilon\right)\left(1 - \widehat{F}(m_j) + \epsilon\right)},$$

kde ϵ zabezpečuje malý posun z dôvodu numerickej stability. V práci Anscombe (1948) je navrhnutá hodnota $\epsilon = 3/(8n)$. V danom výraze vystupuje ešte $\widehat{F}$, ktorá značí empirickú distribučnú funkciu. Empirická funkcia náhodného výberu $X_1, X_2, \ldots, X_n$ je všeobecne daná predpisom:

$$\widehat{F}(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, u]}(X_i).$$

V našej situácii, kedy pracujeme s ordinálnymi dátami, má zmysel počítať empirickú distribučnú funkciu iba v medzných bodoch, a to tak, že sa spravia čiastočné súčty prvkov vektora $\boldsymbol{f}$, ktoré budú podelené celkovým počtom realizácií. Matematicky zapísané ako:

$$\widehat{F}(m_j) = \frac{1}{n} \sum_{i=1}^{j} f_i, \quad \text{pre } j = 1, \ldots, J.$$

Prakticky tento postup ukážeme podrobnejšie v časti 3.3.

3. Simulačná štúdia

Simulačná štúdia je rozdelená do troch podkapitol. V časti 3.1 vychádzame zo správne zvoleného pravdepodobnostného modelu, v 3.2 sa používala rovnaká metóda, ale na chybne zvolený model. Posledná časť 3.3 sa venuje aproximácii využitím Bernsteinových polynómov. Simulácie boli realizované v prostredí matematického softwaru `R` R Core Team (2022) s využitím dodatočného balíčka `truncnorm` od Mersmann a kol. (2023). Pri implementácii algoritmov boli použité rôzne funkcie na báze numerického riešenia optimalizačných úloh.

3.1 Vlastnosti a porovnanie MLE odhadov

Pri samotnom simulovaní sa vychádzalo z rozdelenia, ktorého parametre sme presne poznali, aby bolo možné overiť fungovanie metódy popísanej v práci. Vybrané bolo rozdelenie $\mathcal{N}(25, 9)$. ktoré nám poskytlo nasimulované hodnoty platov v tisícoch so strednou hodnotou 25 a štandardnou odchýlkou 9. Tie boli následne roztriedené do jednotlivých kategórií. Jednotlivé medze boli zvolené v ekvidistantných intervaloch, tak aby do každej kategórie spadli aspoň nejaké pozorovania. Pri piatich kategóriach to bola postupnosť $\{-\infty, 10, 20, 30, 40, +\infty\}$ a pri zahustení intervalov s cieľom obdržania presnejšieho výsledku zase postupnosť $\{-\infty, 10, 15, 20, 25, 30, 35, 40, +\infty\}$. V tabuľke 3.1 je možné nájsť pre každý parameter jeho priemerné vychýlenie a štandardnú odchýlku, ktoré boli vždy počítané z 500 simulácii s postupne so zväčšujúcim sa rozsahom výberu. Konkrétna veľkosť rozsahu výberu je uvedená vždy v zátvorke prvého stĺpca. Riadky MLE v tabuľke prezentujú odhady metódou maximálnej vierohodnosti, vychádzajúcej zo spojitej veličiny. Pre odhad strednej hodnoty μ a smerodajnej odchýlky σ boli teda použité odhady dané predpismi:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n,$$
$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Pri rovnakom rozsahu výberu sa vždy pracovalo s rovnakými dátami, aby bolo možné si všimnúť prípadné zlepšenia v odhadoch. Zlepšenia sme mohli vidieť na tom, že s narastajúcim počtom pozorovaní sa navzájom zmenšovala smerodajná odchýlka jednotlivých odhadov. Zväčšujúci sa počet pozorovaní sme dosiahli zväčšením rozsahu výberu. Efekt poklesu smerodajnej odchýlky podporilo aj navýšenie počtu kategórii. V tabuľke 3.1 si môžeme okrem iného všimnúť, že vo vychýlení sa výsledky líšia až na druhom, prípadne treťom desatinnom mieste. Pridávanie kategórií na okraj definičného oboru nemalo značný prínos, akurát to spôsobovalo problémy s prázdnymi kategóriami pri výpočtoch. Trochu prekvapivo, navýšenie počtu kategórií nemalo vždy pozitívny vplyv na vychýlenie odhadu.

V tomto prípade sme mali správne zvolený pravdepodobnostný model. Inými slovami povedané, snažili sme odhadnúť nejaké normálne rozdelenie na dátach pochádzajúcich pôvodne z normálneho rozdelenia. Nie vždy máme takéto šťastie.

Typ odhadu	**BIAS**(μ)	**SD**(μ)	**BIAS**(σ)	**SD**(σ)
5 kategórií (100)	0,012	0,851	−0,050	0,649
8 kategórií (100)	0,005	0,806	−0,046	0,558
MLE (100)	−0,002	0,773	−0,032	0,452
5 kategórií (200)	−0,016	0,453	0,016	0,287
8 kategórií (200)	−0,019	0,406	0,016	0,253
MLE (200)	−0,018	0,391	0,016	0,205
5 kategórií (300)	−0,041	0,281	−0,001	0,186
8 kategórií (300)	−0,041	0,279	−0,010	0,173
MLE (300)	−0,043	0,276	0,010	0,147

Tabuľka 3.1: Výsledky simulačnej štúdie pre meniaci sa rozsah výberu, ktorý je uvedený v zátvorke

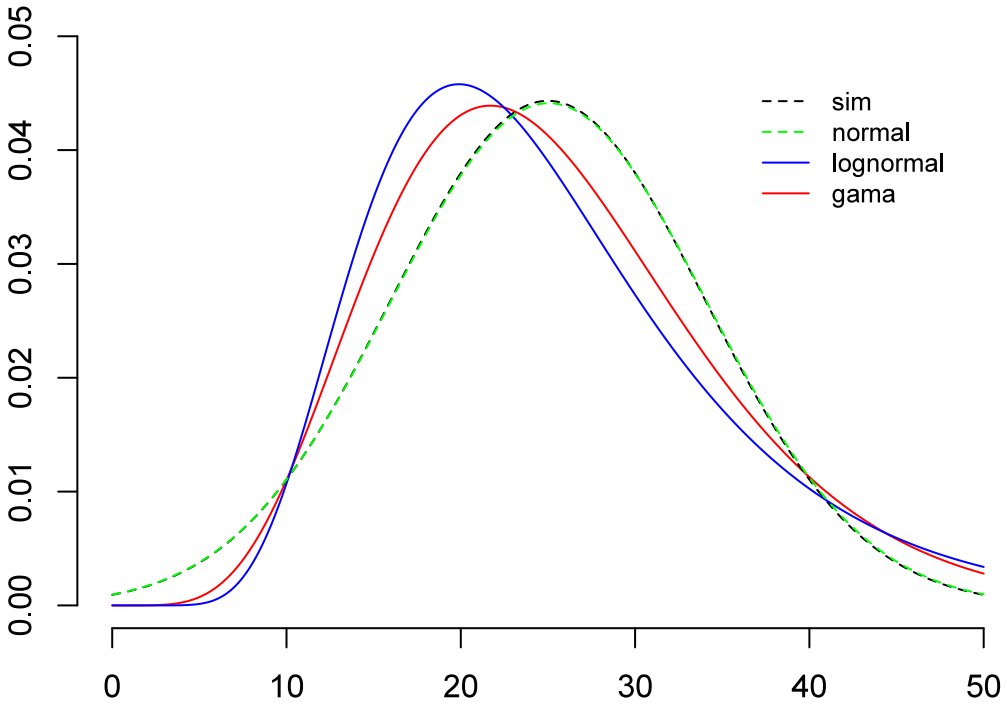
3.2 Odhady pri zlej špecifikácii modelu

Porovnanie použitia iných modelov demonštruje obrázok 3.1, kde je čiernou farbou vyznačené skutočné pravdepodobnostné rozdelenie latentnej veličiny. Pozornosť by sa mala upriamiť predovšetkým na situáciu v oboch takzvaných chvostoch. Keďže žiadne z použitých rozdelení nespĺňa predpoklad symetrie ako normálne rozdelenie, je prirodzené, že práve na nenulových okrajoch nosiča budú vznikať najväčšie odchýlky. Ich konkrétnejšie hodnoty budú popísané v nasledujúcom odstavci.

Ostaneme v scenári, ktorý uvažuje 5 a 8 kategórií s rozsahom výberu 300 pozorovaní. V tabuľke 3.2 môžeme nahliadnuť na jednotlivé odhady parametrov normálneho, log-normálneho a gama rozdelenia a taktiež smerodajné odchýlky týchto odhadov.

Rozdelenie (kat)	**MEAN**(θ_1)	**MEAN**(θ_2)	**SD**(θ_1)	**SD**(θ_2)
normálne (5)	24,959	8,999	0,2807	0,1859
normálne (8)	24,958	8,990	0,2785	0,1727
log-normálne (5)	3,150	0,403	0,0006	0,0005
log-normálne (8)	3,155	0,402	0,0006	0,0004
gama (5)	6,908	0,273	0,5411	0,1440
gama (8)	6,954	0,274	0,4938	0,1278

Tabuľka 3.2: Odhady parametrov rozdelení



Obr. 3.1: Porovnanie odhadnutých latentných pravdepodobnostných rozdelení pre rôzne pravdepodobnostné modely

Väčšiu výpovednú hodnotu však bude mať tabuľka 3.3, v ktorej sú spočítané stredné hodnoty a rozptyly jednotlivých rozdelení pri rozsahu výberu rovnému 300 a spriemerovaní 500 simulácii. To nám dáva prvú silnejšiu indíciu o tom, ako veľmi sa líšia dané charakteristiky od pôvodnej strednej hodnoty 25 a presného rozptylu našich dát s hodnotou 81. Pri počítaní strednej hodnoty a rozptylu logaritmicko-normálneho rozdelenia boli použité predpisy pre strednú hodnotu a rozptyl:

$$\mathrm{E}X = \exp\left(\mu + \frac{1}{2\sigma^2}\right),$$
$$\operatorname{var} X = \exp\left(2\mu + \sigma^2\right)\left(\exp(\sigma^2) - 1\right).$$

A podobne aj pre strednú hodnotu a rozptyl gama rozdelenia:

$$\mathrm{E}X = \frac{\alpha}{\beta}, \quad \operatorname{var} X = \frac{\alpha}{\beta^2}.$$

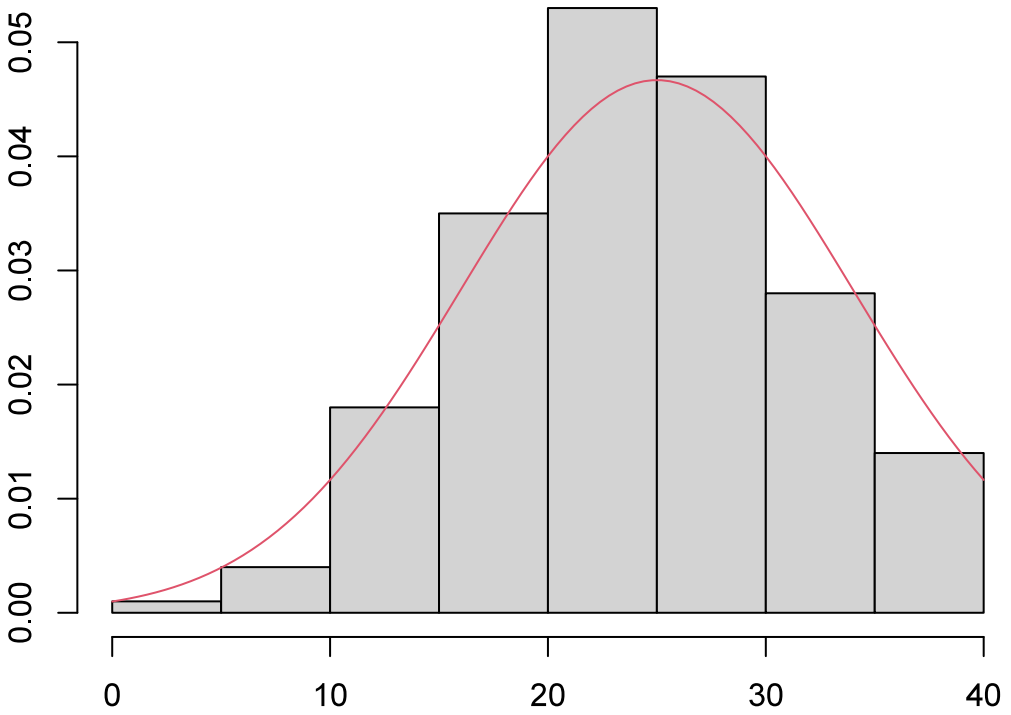
Ostáva nám ešte zodpovedať otázku z motivačnej časti tejto práce týkajúcu sa pravdepodobnosti výskytu platu, ktorý je vyšší alebo rovný nejakej hodnote z posledného intervalu. Bude nás teda zaujímať, s akou pravdepodobnosťou vieme dostať plat vyšší ako 45 tisíc. Konečne, v poslednom stĺpci tabuľky sú pravdepodobnosti toho, že plat vybraný z daného rozdelenia s odhadnutými parametrami bude väčší ako 45 tisíc. Matematicky povedané, zisťujeme pravdepodobnosť $\mathrm{P}(X > 45) = 1 - \mathrm{P}(X \leq 45)$. Typicky počítanú cez doplnkovú pravdepodobnosť a distribučnú funkciu. Presná hodnota tejto pravdepodnosti je 0,0131.

Rozdelenie (kat)	**EX**	**varX**	**P**$(X>45)$
normálne (5)	24,959	80,982	0,0130
normálne (8)	24,958	80,820	0,0129
log-normálne (5)	25,313	112,783	0,0515
log-normálne (8)	25,433	113,403	0,0525
gama (5)	25,351	93,028	0,0369
gama (8)	25,400	92,779	0,0370

Tabuľka 3.3: Stredné hodnoty, rozptyly rozdelení a odhady pravdepodobnosti vytvorené spriemerovaním výsledkov z 500 simulácií

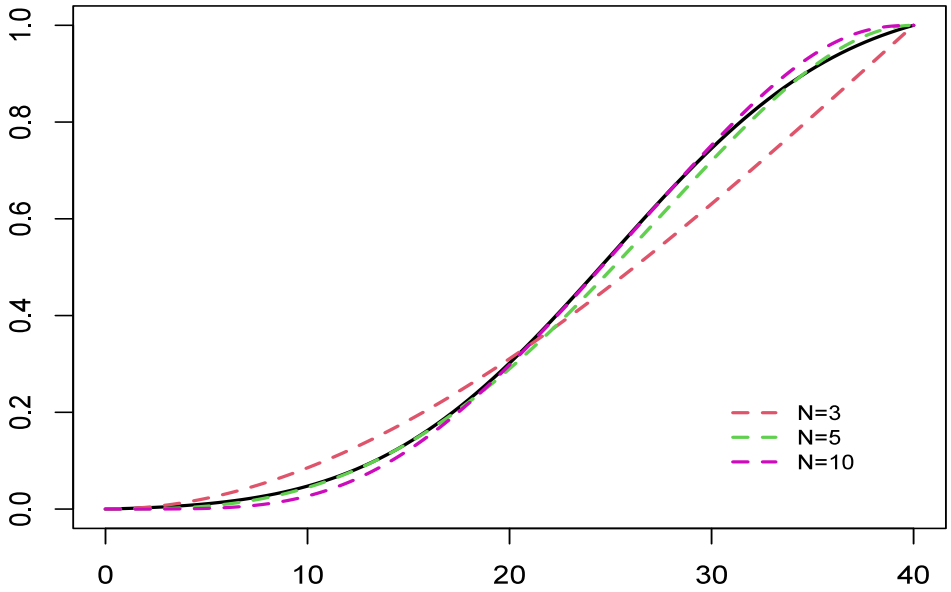
3.3 Odhady pre rozdelenie s konečným nosičom

V poslednej časti simulačnej štúdie vyžadujeme, aby pravdepodobnostné rozdelenie, ktoré budeme odhadovať, malo konečný nosič. Z toho dôvodu si vyberieme useknuté normálne rozdelenie. Dáta budeme generovať podobne ako v minulých častiach z $\mathcal{N}(25, 9)$. Tentokrát zahodíme všetky pozorovania menšie ako 0 a väčšie ako 40. Postup opakujeme, až kým nenazbierame požadovaný rozsah výberu $n = 200$. Konečný nosič je teda interval $(0, 40)$, ktorý sa dobre delí práve na 8 rovnako dlhých podintervalov s dĺžkou 5 (J=8). Toto delenie bude odrazovým mostíkom pre ďalšie pokračovanie. Situácia po ordinalizácii dát je vidno na stĺpcovom grafe relatívnych početností jednotlivých kategórií 3.2, kde je červenou farbou skutočné pravdepodobnostné rozdelenie useknutého normálneho rozdelenia.

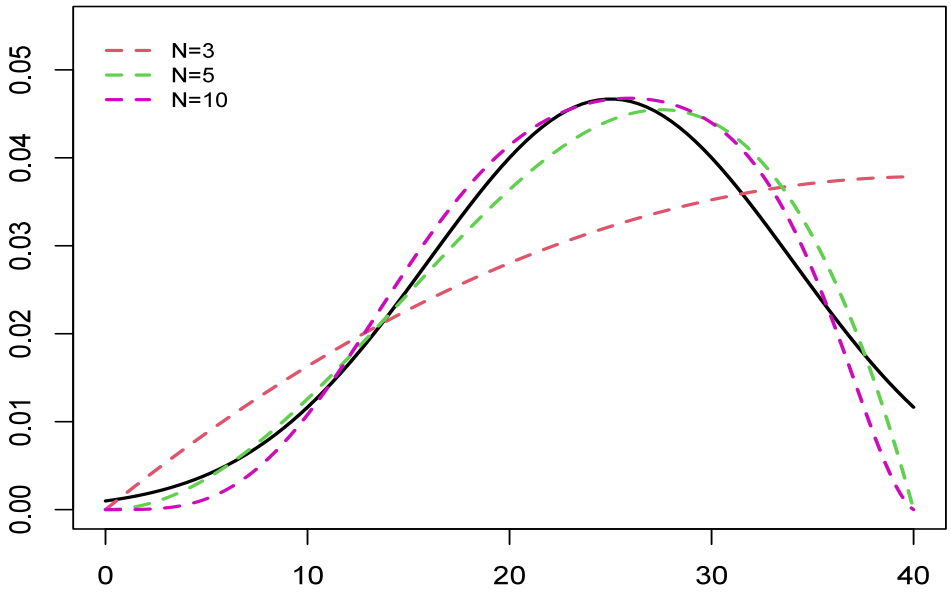


Obr. 3.2: Graf relatívnych početností jednotlivých kategórii pre n=200 a J=8

Okrem počtu kategórií môžeme voliť už spomínané ϵ, alebo počet váh beta rozdelení použitých pri aproximácii. Pre meniace sa ϵ nemožno pozorovať voľným okom výrazné zmeny pri vykreslení jednotlivých odhadov pravdepodobnostných rozdelení. Situácia bude trochu iná pri počítaní pravdepodobností vychádzajúcich z odhadov. Na obrázku 3.3 sú vidno odhady pre distribučnú funkciu useknutého normálneho rozdelenia s postupne sa zvyšujúcim počtom použitých beta rozdelení. Ďalej na obrázku 3.4 je situácia podobná, tento raz sú vykreslené odhady hustoty. Čiernou farbou je vyznačená skutočná distribučná funkcia a hustota. V obidvoch situáciach bolo použité odporúčané $\epsilon = 3/(8n)$ s rozsahom výberu n=200.

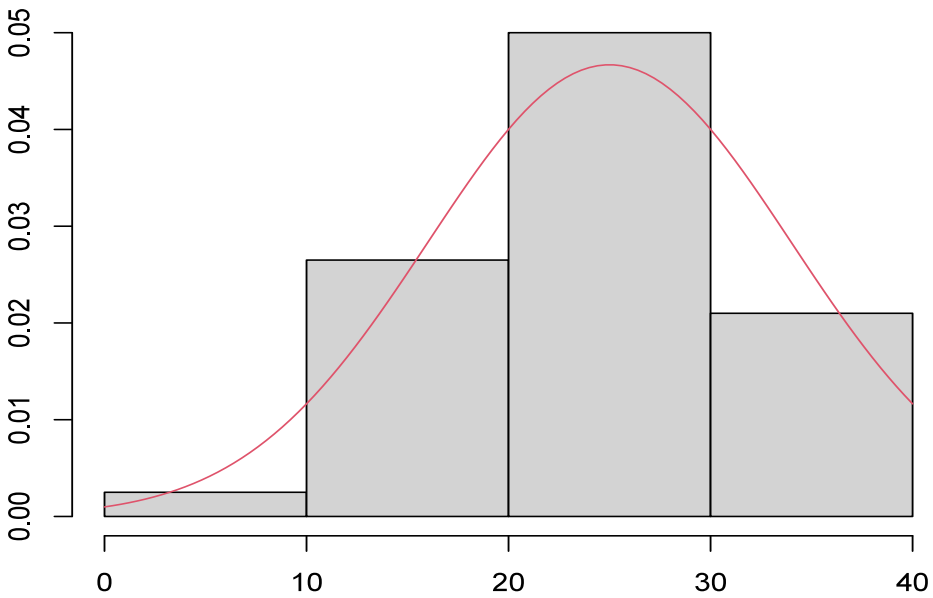


Obr. 3.3: Odhady distribučnej funkcie useknutého normálneho rozdelenia (J=8)

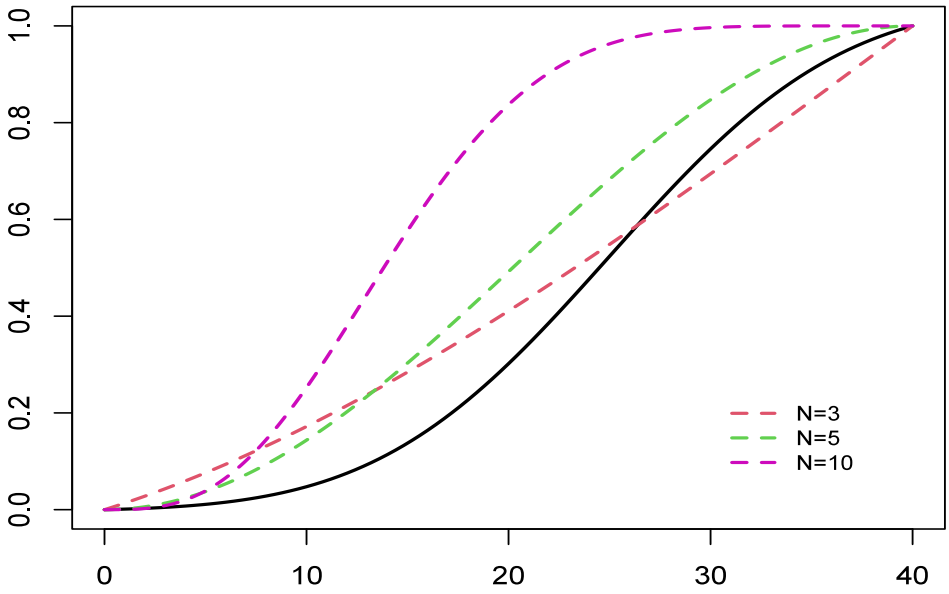


Obr. 3.4: Odhady hustoty useknutého normálneho rozdelenia (J=8)

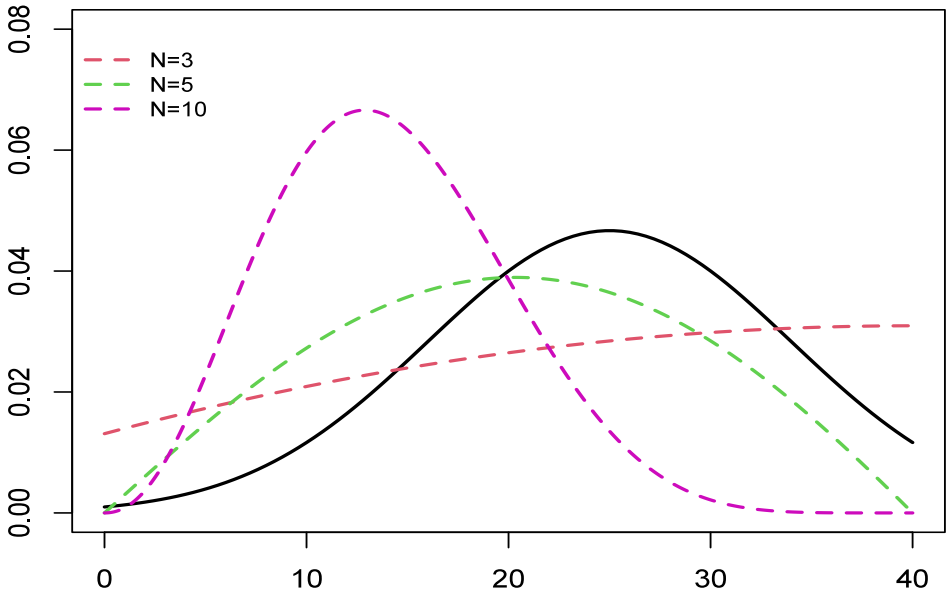
Môžeme si všimnúť, že v tomto prípade sa so zvyšujúcim N výrazne zlepšuje aj celková aproximácia rozdelenia. V situácii na obrázku 3.5, kedy máme k dispozícii dáta rozdelené iba do štyroch kategórií sa odhady budú dopúšťať viditeľne väčších výchyliek. Interval $(0, 40)$ bol v tomto prípade rozdelený na 4 menšie intervaly s dĺžkou 10. Z obrázkov 3.6 a 3.7 je už výrazne ťažšie vyčítať zlepšenie zvýšením počtu použitých beta rozdelení. Stále platí, že čiernou farbou je vyznačená skutočná distribučná funkcia a hustota.



Obr. 3.5: Graf relatívnych početností jednotlivých kategórii pre n=200 a J=8



Obr. 3.6: Odhady distribučnej funkcie useknutého normálneho rozdelenia (J=4)



Obr. 3.7: Odhady hustoty useknutého normálneho rozdelenia (J=4)

V podobnom duchu, ako sme skúmali správanie sa odhadov za hranicou posledného medzného bodu, sa teraz pozrieme na kraj obmedzeného nosiča. Konkrétne nás bude zaujímať P($35 < X < 40$). Výsledné odhady jednotlivých pravdepodobností sme dostali spriemerovaním výsledkov 100 simulácii s rôznymi vstupmi a pevným rozsahom výberu n=200. Presná hodnota tejto pravdepodobnosti je P = 0,090018. Do hry vstupuje tentokrát aj zmena ϵ, okrem odporúčanej hodnoty $\epsilon_1 = 0{,}001875$ vyskúšame aj $\epsilon_2 = 0{,}010000$. V tabuľke 3.4 sú práve odhady spomínanej pravdepodobnosti P. V poslednom riadku je celková suma jednotlivých výchyliek a smerodajnej odchýlky. Môžeme skonštatovať, že si pri voľbe odporúčaného ϵ sme dostali veľmi podobný výsledok. Ďalej sa dá vidieť, že nie v každej situácii si polepšíme zvýšením počtu použitých beta rozdelení.

Počet kategórii (N)	**BIAS(ϵ_1)**	**SD(ϵ_1)**	**BIAS(ϵ_2)**	**SD(ϵ_2)**
8 kategórií (3)	0,0748	0,0230	0,0718	0,0195
8 kategórií (5)	−0,0025	0,0208	−0,0034	0,0197
8 kategórií (10)	−0,0027	0,0186	−0,0025	0,0186
4 kategórie (3)	0,0641	0,0001	0,0647	0,0003
4 kategórie (5)	0,0039	0,0191	0,0025	0,0183
4 kategórie (10)	−0,0792	0,0472	−0,0829	0,0458
Celková suma	0,2272	0,1288	0,2278	0,1222

Tabuľka 3.4: Smerodajné odchýlky a vychýlenia odhadov pravdepodobností v porovnaní so skutočnou hodnotou P = 0,090018

Príliš nízke hodnoty smerodajných odchýlok v niektorých prípadoch mohli byť zapríčinené tým, že viaceré koeficienty pri odhadoch boli rovné nule.

Záver

Práca sa nám snažila nepriamo odpovedať na otázku, či je podstatný rozdiel v práci so spojitými dátami v porovnaní s ich ordinalizovanou verziou. Počas celej práce bola v rôznych podobách využívaná metóda maximálnej vierohodnosti. V simulačnej štúdii sme si mohli všimnúť, že vo vychýlení sa výsledky príliš nelíšia. To, či je takto veľká odchýlka v rámci nejakej normy, záleží asi od konkrétnej situácie. Podstatnejšia je skutočnosť zachovania konzistentnosti takto vzniknutých odhadov. Tieto uvedené fakty nás vedú k možnému využitiu v praxi. Dovolím si tvrdiť, že z užívateľského pohľadu je oveľa jednoduchšie a najmä rýchlejšie vyplniť dotazník s predom prichystanými odpoveďami. Znižuje to možnosť vzniku chyby v dôsledku inak uvedenej fyzikálnej jednotky, nesprávne zapísanej desatinnej čiarky alebo pri ručne písaných odpovediach to minimalizuje aj problém s nečitateľnosťou písma. Firma tak zbytočne neprichádza o cenné údaje a vie, že s ordinalizovanými dátami vie dosiahnuť obdobných výsledkov.

Pri práci s Bernsteinovými polynómami sme mali dodatočnú požiadavku, ktorý nebol splnený pri žiadnom z doterajších prípadov, a to konkrétne konečný nosič odhadovanej hustoty. Výhodou tohto prístupu by mohla byť flexibilita pri aproximácii hustoty, ktorá má nejakým spôsobom netradičný tvar a bežné parametrické modely nie sú v tomto prípade dostačujúce.

Zoznam použitej literatúry

Andrews, L. C. (1998). *Special functions of mathematics for engineers*, volume 49. Spie Press.

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**(3/4), 246–254.

Babu, G. J., Canty, A. J. a Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, **105**(2), 377–392.

Ghosh, S. K., Burns, C. B., Prager, D. L., Zhang, L. a Hui, G. (2018). On nonparametric estimation of the latent distribution for ordinal data. *Computational Statistics & Data Analysis*, **119**, 86–98.

Lehmann, E. (1983). Theory of point estimation, New York: Johnwiley. *Lehmann Theory of Point Estimation 1983*.

Li, A. a Qin, H. (2017). The representations on the partial derivatives of the extended, generalized gamma and incomplete gamma functions and their applications. *IAENG Int. J. Appl. Math*, **47**(3), 312–318.

Lorentz, G. G. (2012). *Bernstein polynomials*. American Mathematical Soc.

Mersmann, O., Trautmann, H., Steuer, D. a Bornkamp, B. (2023). *Truncated normal distribution*. URL `https://github.com/olafmersmann/truncnorm`.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Tamhane, A., Ankenman, B. a Yang, Y. (2002). The beta distribution as a latent response model for ordinal data (i): estimation of location and dispersion parameters. *Journal of Statistical Computation and Simulation*, **72**(6), 473–494.