MASARYKOVA UNIVERZITA

Přírodovědecká fakulta

Lucie DOUDOVÁ

STATISTICKÁ ANALÝZA POPULACÍ S NEGATIVNĚ BINOMICKÝM ROZDĚLENÍM

Dizertační práce

Školitel: doc. RNDr. Jaroslav Michálek, CSc.

Brno, 2009

Bibliografická identifikace

Jméno a příjmení autora: Lucie Doudová

Název disertační práce: Statistická analýza populací s negativně binomickým rozdělením

Název disertační práce anglicky: Statistical Analysis of Populations with the Negative Binomial Distribution

Studijní program: Matematika

Studijní obor: Pravděpodobnost, statistika a matematické modelování

Školitel: doc. RNDr. Jaroslav Michálek, CSc.

Rok obhajoby: 2010

Klíčová slova v češtině: negativně binomické rozdělení, maximálně věrohodné odhady, korekce odhadů na nestrannost, quasi-likelihood odhady, bayesovský odhad, testy s rušivými parametry, zobecněné lineární modely, síla testů

Klíčová slova v angličtině: negative binomial distribution, maximum likelihood estimators, bias-corrected estimators, quasi-likelihood estimators, bayes estimators, tests with nuisance parameters, generalized linear models, power of the test

Poděkování

Chtěla bych toutou cestou poděkovat svému školiteli doc. RNDr. Jaroslavu Michálkovi, CSc. za jeho rady, trpělivost, ochotu a množství času, který mi věnoval. Ráda bych poděkovala i Mgr. Zuzaně Hübnerové, Ph. D. za poskytnutí programu GLM_NB. Nemenší dík patří všem, kteří mne během tvorby této práce přímo nebo nepřímo podpořili.

Poděkování též patří nadaci Nadání Josefa, Marie a Zdeňky Hlávkových za cestovní stipendium podporující moji účast na zahraniční konferenci TIES 2004.

Abstrakt

Předložená disertační práce se zabývá různými aspekty statistické analýzy populací s negativně binomickým (NB) rozdělením. Práce je rozdělena do osmi kapitol a dodatku. V první kapitole jsou definovány základní pojmy a označení používané v dalším textu. Jsou specifikovány vybrané pojmy z teorie odhadu a testování statistických hypotéz s rušivými parametry a dále potřebné vlastnosti zobecněného lineárního modelu.

V kapitole druhé je zavedeno NB rozdělení a jsou popsány jeho vlastnosti zejména s ohledem na analýzu biologických populací. Detailně jsou diskutovány různé způsoby parametrizace NB rozdělení s ohledem na jejich další použití. V závěru této kapitoly je uvedeno necentrální NB rozdělení. V první části třetí kapitoly jsou popsány momentové a maximálně věrohodné odhady parametrů a algoritmy pro jejich výpočet pro různé typy parametrizací. V další části této kapitoly je řešena otázka korekce vychýlení maximálně věrohodných odhadů a dále pro eliminaci problémů s numerickým hledáním odhadů je popsán a využit quasi-likelihood přístup. Konečně je navržen a popsán bayesovský přístup k nalezení odhadů, který dává dobré výsledky při libovolných hodnotách parametrů. Všechny uvedené odhady jsou na závěr porovnány, je provedeno praktické doporučení pro výběr metody odhadu pro daná data.

Čtvrtá kapitola detailně pojednává o testech hypotéz rovnosti středních hodnot NB rozdělení za různých podmínek na rušivé parametry a dále obsahuje testy hypotéz o rovnosti parametrů κ. Pro porovnání výběrů z NB rozdělení byly odvozeny explicitní tvary Waldovy statistiky, skórové statistiky a věrohodnostního poměru při různých volbách vektoru rušivých parametrů. V navazující páté kapitole jsou pak tyto testy doplněny dalšími testy, které při známé hodnotě parametru κ, vycházejí z teorie zobecněných lineárních modelů. V následující šesté kapitole pak jsou pomocí simulací stanoveny síly dříve zkonstruovaných testů a jsou uvedeny aproximace pro výpočet sil uvedených testů při známém parametru κ. Síly jednotlivých testů jsou graficky porovnány.

V sedmé kapitole jsou uvedené techniky odhadu aplikovány na reálná data. V prvním případě se jedná o studii výskytu spárkaté zvěře v závislosti na typu porostu, v případě druhém je analyzován absolutní počet neutrofilů v závislosti na stupni sepse dětských pacientů Univerzitní nemocnice v Brně. Kapitola je doplněna úvahou o návrhu rozsahu datového souboru s ohledem na požadovanou hodnotu síly

použitých testů.

Programová implementace prakticky všech použitých metod realizovaných ve výpočetním systému MATLAB je obsahem závěrečné kapitoly. V dodatku jsou uvedeny výpočty týkající se korekce maximálně věrohodných odhadů na nestrannost.

Výpočty uvedené v této práci byly provedeny pomocí k tomu účelu sepsaných programů v MATLABu. Tyto programy jsou k dispozici na přiloženém CD.

Literatura

Al-Saleh, M. F., and Al-Batainah, F. K. Estimation of the shape parameter k of the negative binomial distribution. *Appl. Math. and Comput. 143*, 2-3 (2003), 431–441.

Saha, K., and Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics 61*, 1 (2005), 179–185.

White, G. C., and Eberhardt, L. E. Statistical analysis of deer and elk pellet-group data. *J. Wildl. Manage. 44*, 1 (1980), 121-131.

Abstract

The submitted thesis is concerned with various aspects of statistical analysis of the populations with negative binomial distribution (NB). This thesis consists of eight chapters and an appendix. Chapter One contains the definition of basic terms and the notation used in the text as well as the specification of selected terms concerning the theory of estimation and testing of statistic hypothesis with nuisance parameters and necessary characteristics of generalized linear model.

The NB distribution is introduced and its characteristics with respect to the biological population analysis are described in Chapter Two. Various manners of NB parameterization with respect to their possible use are discussed in detail. The noncentral NB distribution is given in the conclusion of the above chapter. Chapter Three deals with moment and maximum likelihood estimators of parameters and algorithms for the calculation of various parameterization types. In the next part of this chapter we solve the bias correction of maximum likelihood estimator and describe and use the quasi-likelihood approach to eliminate imperfections of estimators. Eventually we suggest and describe Bayes approach to finding estimators which approach produces good results under arbitrary parameters. Finally we compare all estimators and offer practical recommendations for selecting the estimation method for the given data.

Chapter Four deals with the hypothesis testing of the equality of NB distribution mean values under various conditions for the nuisance parameters and it also contains κ parameters equality hypothesis tests. To compare the selected data from the NB distribution, explicit values of Wald statistics, score statistics and likelihood ratio have been derived under various vectors of nuisance parameters. Other tests are added in the following chapter which derive from the generalized linear models theory if the κ parameter is known. The powers of formerly constructed tests are determined by means of simulations and approximations for the calculation of power of the given tests if the κ parameter is known are specified in Chapter Six. The powers of individual tests are compared graphically.

We use the above estimation methods in practice in Chapter Seven. In the first case we carry out a study examining the dependence of the number of deer on the vegetation type. In the second case we analyze the absolute neutrophile count depending on the degree of sepsis in child patients in the Teaching Hospital in Brno. This chapter also suggests the experimental design with respect to the required

power of tests used.

Descriptions of procedures implemented in the MATLAB computing system are given in the final chapter. Calculations related to bias correction in MLE are given in the Appendix.

Calculations presented in the dissertation were carried out using MATLAB programs, which are to be found on enclosed CD.

References

Al-Saleh, M. F., and Al-Batainah, F. K. Estimation of the shape parameter k of the negative binomial distribution. *Appl. Math. and Comput. 143*, 2-3 (2003), 431–441.

Saha, K., and Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics 61*, 1 (2005), 179–185.

White, G. C., and Eberhardt, L. E. Statistical analysis of deer and elk pellet-group data. *J. Wildl. Manage. 44*, 1 (1980), 121-131.

Obsah

Přehled značení a základních pojmů

Přehled značení

- $\alpha_{d,n}$ ... Stirlingova čísla druhého druhu
- $D(Y)$ ... rozptyl náhodné veličiny Y
- $E(Y)$ ... střední hodnota náhodné veličiny Y
- η ... lineární prediktor
- $\mathbf{F}$ ... výběrová informační matice
- $f(.)$ ... hustota vzhledem k Lebesgeově míře
- $F(.)$ ... distribuční funkce
- γ_1 ... šikmost
- γ_2 ... špičatost
- $\Gamma(b)$ ... gamma funkce, $\Gamma(b) = \int_0^\infty x^{b-1}e^{-x}dx$, $b > 0$
- $\mathbf{J}$ ... Fisherova informační matice
- ξ_r ... kumulant r-tého řádu
- $\xi_{[r]}$ ... klesající kumulant r-tého řádu
- $l(\theta, y)$ ... logaritmická věrohodnostní funkce
- $L_n^\alpha(y)$ ... Laguerrův polynom

- LR ... věrohodnostní poměr
- μ'_r ... obecný moment r-tého řádu
- μ_r ... centrální moment r-tého řádu
- $\mu_{[r]}$ ... klesající faktoriální moment r-tého řádu
- n ... rozsah výběru
- $\mathbb{N}$ ... množina přirozených
- $o(h)$ ... symbol „malé o“ $\lim_{h\to 0} \frac{o(h)}{h} = 0$
- Ω ... základní prostor
- $\xrightarrow{P}$ ... konvergence podle pravděpodobnosti
- $\boldsymbol{\psi}$ ... vektor rušivých parametrů
- $\psi_Y(.)$ ... charakteristická funkce
- $\Psi(b)$ ... digamma funkce, $\Psi(b) = \frac{\partial}{\partial b} \ln \Gamma(b)$
- $\Psi'(b)$ ... trigamma funkce, $\Psi'(b) = \frac{\partial}{\partial b} \Psi(b)$
- $\mathbb{R}$ ... množina reálných čísel
- S ... skórová statistika
- S_t resp. s_t ... výběrová směrodatná odchylka, resp. její realizace
- σ ... směrodatná odchylka
- σ^2 ... rozptyl
- $\chi^2_\alpha(v)$ ... α-kvantil χ^2 rozdělení s v stupni volnosti
- $\boldsymbol{\tau}$ ... vektor cílových parametrů
- $\boldsymbol{\tau}_0$ ... skutečná hodnota vektoru $\boldsymbol{\tau}$
- Θ ... parametrický prostor
- θ, $\boldsymbol{\theta}$ ... skalární resp. vektorový parametr

- $\mathbf{U}$ ... skórový vektor
- $\mathbf{V}(.)$ ... varianční matice
- W ... Waldova statistika
- $\overline{Y}$, $\overline{y}$ ... výběrový průměr a jeho realizace
- $Y \overset{A}{\sim} N(\mu, \sigma^2)$ ... náhodná veličina Y má asymptoticky normální rozdělení s parametry μ a σ^2
- $\mathbb{Z}$ ... množina celých čísel
- $\binom{n}{k}$ ... binomický koeficient
- ${}_1F_1(a, b, c)$ ... hypergeometrická funkce prvního druhu
- $(\Omega, \mathcal{A}, P)$ ... pravděpodobnostní prostor

Rozdělení použitá v dalším textu

- Binomické rozdělení $Bi(n, \pi)$ s parametry $n \in \mathbb{N}$, $\pi \in (0, 1)$ a dané hustotou

$$\begin{aligned} f(y; n, \pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} && y = 0, 1, \dots \\ &= 0 && \text{jinak.} \end{aligned}$$

- Poissonovo rozdělení $Po(\lambda)$ s parametrem $\lambda > 0$ a dané hustotou

$$\begin{aligned} f(y; \lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} && y = 0, 1, \dots \\ &= 0 && \text{jinak.} \end{aligned}$$

- Negativně binomické rozdělení $NB(\pi, \kappa)$ s parametry $\pi \in (0, 1)$, $\kappa > 0$ a dané hustotou

$$\begin{aligned} f(y; \kappa, \pi) &= \binom{y + \kappa - 1}{y} (1-\pi)^y \pi^\kappa && y = 0, 1, \dots \\ &= 0 && \text{jinak.} \end{aligned}$$

- Exponenciální rozdělení $\mathrm{E}(\lambda; a)$ s parametry $\lambda > 0$ a a dané hustotou

$$f(y; \lambda, a) = \lambda e^{-\lambda(y-a)}, \quad t \geq a$$
$$= 0 \text{ jinak.}$$

- Gamma rozdělení $\mathrm{G}(\lambda; b; a)$ s parametry $\lambda > 0$, $b > 0$ a a dané hustotou

$$f(y; \lambda, b, a) = \frac{\lambda(\lambda(y-a))^{b-1}}{\Gamma(b)} e^{-\lambda(y-a)}, \quad y \geq a$$
$$= 0 \text{ jinak.}$$

- Normální rozdělení $\mathrm{N}(\mu; \sigma^2)$ s parametry μ a $\sigma > 0$ dané hustotou

$$f(y; \mu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}, \quad y \in (-\infty, \infty).$$

Použité zkratky

- BC ... korekce na nestrannost (z anglického bias corrected)
- DEQL ... double extended quasi likelihood
- EQL ... extended quasi likelihood
- GLM ... zobecněný lineární model (z anglického generalised linear model)
- ML ... maximální věrohodnost (z anglického maximum likelihood)
- MM ... metoda momentů
- NB ... negativně binomické rozdělení
- QL ... quasi likelihood

Náhodné veličiny budeme značit velkými písmeny z konce abecedy. Matice budou značeny tučnými velkými písmeny, vektory tučnými malými písmeny. Odhady budeme značit stříškou nebo vlnkou.

Úvod

Negativně binomické (NB) rozdělení patří mezi základní a jedno z nejčastěji používaných diskrétních rozdělení pravděpodobnosti. Jeho použití lze nalézt v ekologických, lékařských, biologických, psychologických aplikacích v pojišťovnictví, kontrole jakosti a v řadě dalších oborů. Tato práce je zaměřena zejména na jeho použití při analýze biologických populací. Ovšem získané výsledky a praktická doporučení mohou být užitečná i v jiných oborech.

Statistické analýze výběrů z NB rozdělení je věnována dlouhá řada teoretických i aplikačních prací, první pocházejí již z první poloviny minulého století např. [24], ty nejnovější jsou ze současnosti např. [1]. Rovněž v řadě softwarových produktů (MATLAB, STATISTICA apod.) jsou k dispozici programy pro statistickou analýzu NB rozdělení, ale mnohé z těchto programů vycházejí ze základních metod odhadu, často se v tomto směru používá momentový a maximálně věrohodný odhad, ale problém je, že pro některé datové soubory tyto metody zcela selhávají. Je to proto, že základní charakteristiky NB rozdělení - střední hodnota μ a rozptyl σ^2 splňují nerovnost $\mu < \sigma^2$, ale při zpracování reálných dat se často stane, že výběrový průměr $\overline{X}$ překročí výběrový rozptyl S^2 i přes to, že data tvoří náhodný výběr z NB rozdělení. Tuto skutečnost lze dobře ilustrovat pomocí simulací. V uživatelské praxi se tato situace popisuje jako „underdispersion“. Jejím důsledkem je pak to, že získané odhady parametrů se špatně interpretují nebo zcela selhávají numerické algoritmy pro jejich výpočet. Uvedené problematice je věnována značná pozornost mnoha autorů (viz např. [8], [9], [13], [61]). Tak je tomu i v této práci.

Vzhledem k tomu, že statistickou analýzu dat z NB rozdělení provázejí výše popsané těžkosti, bylo cílem práce vybrat, případně navrhnout a porovnat statistické postupy, vhodné pro reálnou analýzu populací s NB rozdělením a provést jejich počítačovou implementaci. Protože každý aplikační obor má své specifické problémy, byla práce zaměřena zejména na zpracování biologických populací.

V první kapitole práce jsou uvedeny základní pojmy a označení a jsou stručně

připomenuty potřebné teoretické základy z oblasti teorie odhadu, zobecněných lineárních modelů a testů hypotéz s rušivými parametry, které jsou poté v další části práce využívány.

Druhá kapitola je věnována zavedení NB rozdělení, je uveden standardní přístup pomocí bernouliovské posloupnosti nezávislých pokusů a na něj navazují dvě biologické interpretace NB rozdělení, které naznačují proč mnohé biologické populace často mají NB rozdělení. Kapitola je doplněna o různé způsoby reparametrizace NB rozdělení a pro všechny uvedené reparametrizace byly stanoveny základní funkcionální a číselné charakteristiky (momenty, kumulanty, charakteristická funkce). Závěr kapitoly je věnován zobecnění NB rozdělení na necentrální NB rozdělení.

Další kapitola už je věnována odhadům parametrů NB rozdělení. Nejdříve jsou uvedeny klasické metody odhadu tedy metoda momentů a metoda maximální věrohodnosti. Protože odhady κ získané metodou maximální věrohodnosti nejsou nestranné je uvedena korekce na nestrannost (BC) viz [15]. V článku [63] je tato metoda uvedena pro parametrizaci μ a c a je v něm využita Stirlingova aproximace. V této práci je tato korekce nalezena s využitím gamma, digamma a trigamma funkcí. V [63] je také uvedena korekce maximálně věrohodného odhadu parametru μ. Ten je však nestranný a není třeba ho korigovat. Přesto je tato korekce také přepočítána, protože v [63] je uvedena chybně. Výpočty potřebné k nalezení BC odhadů jsou poněkud zdlouhavé a jsou proto uvedeny v dodatku. V dodatku je také uvedena korekce maximálně věrohodných odhadů pro parametrizaci μ a κ. Metoda maximální věrohodnosti ovšem selhává pro výběry s výběrovým průměrem větším než výběrový rozptyl. Proto je dále uvedena alternativní možnost odhadu - quasi-likelihood přístup ([13], [33], [45], [46], [55]) a nalezeny extended quasi-likelihood (EQL) a double extended quasi-likelihood (DEQL) odhady. Na závěr je uveden bayesovský přístup k hledání odhadů zpracovaný podle článku [2]. Porovnáním se ukazuje, že tato metoda funguje velmi dobře obzvláště v situacích, kdy je κ velké a střední hodnota malá, a to i tam, kde je výběrový rozptyl menší než výběrový průměr. Výpočty zmíněných odhadů byly naprogramovány ve výpočetním systému MATLAB.

Čtvrtá kapitola je zaměřena na analýzu populací, která odpovídá analýze rozptylu pro normální populace. Je zpracována podle [4] pomocí testů s rušivými parametry. Podrobně je rozpracováno pět základních modelů a to test rovnosti středních hodnot μ za předpokladu, že rušivé parametry κ jsou obecně různé neznámé a za předpokladu, že rušivé parametry κ jsou rovny společné neznámé hodnotě. Dále test rovnosti κ za předpokladu, že rušivé parametry μ jsou obecně různé neznámé a za

předpokladu, že rušivé parametry μ jsou rovny společné neznámé hodnotě. A jako poslední je uvažován společný test, že střední hodnoty jsou rovny neznámé společné hodnotě a zároveň, že κ jsou také rovny neznámé společné hodnotě. V tomto případě jsou rušivé parametry pouze ony neznámé společné hodnoty. Obecné tvary věrohodnostního poměru, Waldovy a skórové statistky pro testy s rušivými parametry byly modifikovány a byly odvozeny testovací statistiky pro testy stanovených hypotéz. Odvozený tvar Fisherovy informační matice včetně jejího rozdělení na bloky je uveden, pro všechny uvažované modely, v závěru kapitoly. Algoritmy pro výpočet testovacích statistik byly naprogramovány. V následující páté kapitole je popsán přístup k testům o střední hodnotě NB rozdělení za předpokladu, že parametr κ je známý. Za této podmínky lze užít metod známých z teorie zobecněných lineárních modelů.

Tento přístup k testování hypotézy $H_0 : \mu_1 = \cdots = \mu_n = \mu$ byl společně s testy s rušivými parametry použit v šesté kapitole v simulační studii ke srovnání síly testů rovnosti středních hodnot za situace, kdy parametry κ jsou obecně neznámé a různé, kdy parametry κ jsou si rovny, ale jejich společná hodnota je neznámá a konečně když parametry κ jsou rovny společné známé hodnotě.

Následující sedmá kapitola je věnována aplikacím. Jsou popsány 2 modely a to analýza populací spárkaté zvěře v oblasti Jeseníků a analýza počtu neutrofilů v závislosti na septickém stavu dětských pacientů.

Práce je doplněna souborem programů.

Kapitola 1

Základní pojmy a označení

V této kapitole budou uvedeny základní pojmy a označení, kterých bude dále v práci použito. Budou uvedeny základní vlastnosti zavedených objektů, tvrzení a věty z teorie odhadu a z teorie testování statistických hypotéz, na které se budeme v další části odvolávat. Příslušné věty jsou zpracovány podle [4], [37], [35], [36], [59] a jsou uvedeny bez důkazů. Jejich důkazy je možno nalézt v citované literatuře.

1.1 Rozdělení pravděpodobnosti a jeho charakteristiky

Je-li Y náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$, pak označíme $F(y) = P(Y \leq y)$ její distribuční funkci a příslušnou hustotu vzhledem k nějaké σ-konečné míře μ označíme f. V případě, že míra μ bude Lebesgueova míra bude f hustotou absolutně spojité náhodné veličiny v obvyklém smyslu [4]. V případě, že μ bude čítací míra, bude $f(y) = P(Y = y)$ pravděpodobnostní funkce diskrétní náhodné veličiny Y. Analogické označení pro hustotu a distribuční funkci budeme používat i v případě, že Y bude náhodný vektor.

Dále pokud $X = g(Y)$ je transformovaná náhodná veličina (tj. g je borelovská funkce), označíme

$$EX = Eg(Y) = \int g(y)\ dF(y)$$

její střední hodnotu za předpokladu, že uvedený integrál existuje. Dále pokud budou existovat uvedené střední hodnoty zavedeme obecné momenty μ'_r náhodné veličiny

Y vztahem

$$\mu'_k = EY^k, \qquad k = 0, 1, 2, \ldots$$

a centrální momenty vztahem

$$\mu_k = E(Y - EY)^k, \qquad k = 0, 1, 2, \ldots.$$

Charakteristickou funkci $\psi_Y(t)$ náhodné veličiny Y zavedeme vztahem

$$\psi_Y(t) = \int e^{ity} dF(y). \tag{1.1}$$

Uveďme některé vlastnosti charakteristické funkce. Jestliže existuje $\mu'_r = E(Y^r)$, pak existuje r-tá derivace funkce $\psi_Y(t)$,

$$\psi_Y^{(r)}(t) = (i)^r \int y^r e^{ity} dF(y),$$

a je stejnoměrně spojitá. Dále platí, že

$$\mu'_s = E(Y^s) = (i)^{-s} \psi_Y^{(s)}(0), \quad s \leq r.$$

Jestliže existují momenty μ'_j, $j = 1, \ldots, r$, pak lze $\psi_Y(t)$ rozvinout podle Taylorovy formule

$$\psi_Y(t) = \sum_{j=0}^{r} \mu'_j \frac{(it)^j}{j!} + o(t^r) \quad (\text{pro } t \to 0).$$

Funkci $\varphi(t) = \log \psi_Y(t)$ nazveme vytvořující funkcí kumulantů veličiny Y. Za předpokladu, že existují momenty $\mu'_r = E(Y^r)$, lze $\varphi(t)$ rozvinout podle Taylorovy formule

$$\varphi(t) = \sum_{0}^{r} \xi_j \frac{(it)^j}{j!} + o(t^r) \quad \text{pro } t \to 0,$$

čísla ξ_j se nazývají kumulanty.

Konečně při popisu vlastností NB rozdělení budeme potřebovat klesající faktoriální momenty

$$\mu'_{[k]} = E\left(Y(Y-1)\ldots(Y-k+1)\right), \qquad k = 0, 1, \ldots \tag{1.2}$$

za předpokladu, že uvedené střední hodnoty existují.

Dále připomeňme vztah obecných a faktoriálních momentů [67]. K tomu budeme potřebovat Stirlingova čísla druhého druhu. V této práci je budeme značit $\alpha_{d,n}$ a zavedeme je následujícím rekurentním vzorcem:

$$\begin{aligned} \alpha_{1,n} &= 1 \text{ pro } n = 1, 2, \dots \\ \alpha_{d,n} &= \sum_{j=0}^{n-1} d^j \alpha_{d-1,n-j} \text{ pro } d \geq 2,\ n = 1, 2, \dots . \end{aligned} \tag{1.3}$$

Pak platí (viz [67]), že

$$\mu'_n = E(Y^n) = \sum_{d=1}^{n} \alpha_{d,n-d+1} \mu'_{[d]}. \tag{1.4}$$

Konečně při popisu NB rozdělení a zejména při jeho zobecnění budeme potřebovat hypergeometrickou funkci prvního typu ${}_1F_1(a, b, z)$, $a, b, c \in \mathbb{R}$, $(b)_n \neq 0$ definovanou vztahem

$${}_1F_1(a, b, z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}, \tag{1.5}$$

kde $(a)_n = a(a+1)\dots(a+n-1)$.

Při práci s hypergeometrickou funkcí prvního typu budeme využívat Kumerovu transformaci

$${}_1F_1(\alpha, \beta, y) = e^y \, {}_1F_1(\beta - \alpha, \beta, -y) \tag{1.6}$$

a dále pak vyjádření zobecněných Laguerrových polynomů pomocí hypergeometrické funkce prvního typu ve tvaru

$$L_n^{\alpha}(y) = \frac{(\alpha+1)_n}{n!} \, {}_1F_1(-n, \alpha+1, y). \tag{1.7}$$

1.2 Regulární systém hustot

V tomto odstavci připomeneme zavedení Fisherovy informační matice pro regulární systém hustot a dále skórový vektor a jeho základní vlastnosti. O tyto pojmy a jejich vlastnosti se pak budeme opírat při konstrukci statistických testů pro srovnání negativně binomických populací. Odstavec je vypracován podle [4], [35], [36], [19].

V celém odstavci budeme předpokládat, že náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ má hustotu $f(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^n$ vzhledem k σ-konečné míře μ. Tato hustota závisí na vektorovém parametru $\boldsymbol{\theta}$, tedy $f(\boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta})$. Parametr $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ a množina $\boldsymbol{\Theta}$ je parametrický prostor.

Definice 1.1 *Nechť $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ má hustotu $f(\boldsymbol{y}; \boldsymbol{\theta})$ vzhledem k σ-konečné míře μ. Předpokládejme, že platí:*

1. *$\boldsymbol{\theta} \in \boldsymbol{\Theta}$ a $\boldsymbol{\Theta}$ je neprázdná a otevřená množina v $\mathbb{R}^{\mathrm{m}}$.*
2. *Množina $M = \{\boldsymbol{y} \in \mathbb{R}^n : f(\boldsymbol{y}; \boldsymbol{\theta}) > 0\}$ nezávisí na $\boldsymbol{\theta}$.*
3. *Existují konečné parciální derivace $f_i' = \frac{\partial f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_i}$, pro všechna $i = 1, \ldots, m$ a skoro všechna $\boldsymbol{y} \in M$ vzhledem k μ.*
4. *Pro každé $i = 1, \ldots, m$ a všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí, že*

$$\int_M \frac{\partial \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_i} f(\boldsymbol{y}; \boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{y}) = \mathbf{0} \, .$$

5. *Pro každou dvojici (i, j), $i, j = 1, \ldots, m$ existuje konečný integrál*

$$J_{ij}(\boldsymbol{\theta}) = \int_M \frac{\partial \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_j} f(\boldsymbol{y}; \boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{y}).$$

6. *Matice $\boldsymbol{J}_n(\boldsymbol{\theta}) = (J_{ij}(\boldsymbol{\theta}))_{i,j=1}^m$ je pozitivně definitní pro každé $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Pak se systém hustot $\{f(\boldsymbol{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ nazývá regulární a matice $\boldsymbol{J}_n(\boldsymbol{\theta})$ se nazývá Fisherova informační matice o parametru $\boldsymbol{\theta}$.

Věta 1.2 *Nechť $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ je náhodný výběr z rozdělení o hustotě $f(y; \boldsymbol{\theta})$ (vzhledem k σ-konečné míře μ), $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Nechť systém hustot $\{f(y; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ je regulární a má Fisherovu míru informace $\boldsymbol{J}(\boldsymbol{\theta})$. Pak náhodný vektor $\boldsymbol{Y}$ má hustotu $f_n(\boldsymbol{y}; \boldsymbol{\theta}) = \Pi_{i=1}^n f(y_i; \boldsymbol{\theta})$ vzhledem k součinové míře $\mu_n = \mu \times \cdots \times \mu$. Systém hustot $\{f_n(\boldsymbol{y}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ je regulární a má Fisherovu míru informace $\boldsymbol{J}_n(\boldsymbol{\theta}) = n\boldsymbol{J}(\boldsymbol{\theta})$.*

Důkaz je zřejmý.

Definice 1.3 *Je-li f regulární hustota, pak pro $\mathbf{y} \in M$ definujeme věrohodnostní a logaritmickou věrohodnostní funkci následovně.*

1. *Věrohodnostní funkcí rozumíme funkci* $L(\boldsymbol{\theta}; \boldsymbol{y})$ *vektorového parametru* $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}).$$

2. *Logaritmickou věrohodnostní funkcí rozumíme funkci* $l(\boldsymbol{\theta}; \boldsymbol{y})$

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \ln f(\boldsymbol{y}; \boldsymbol{\theta}).$$

Definice 1.4 *Nechť* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *je náhodný vektor s regulární hustotou* f*. Pak* m*-rozměrný náhodný vektor* $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), \ldots, U_m(\boldsymbol{\theta}))'$ *se složkami*

$$U_i(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \boldsymbol{\theta}_i} \tag{1.8}$$

nazýváme skórový vektor příslušný hustotě f *(též příslušný náhodnému vektoru* $\mathbf{Y}$ *s hustotou* f*).*

Poznámka 1.5 *Systém rovnic*

$$U_i(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \boldsymbol{\theta}_i} = 0, \quad i = 1, \ldots, m \tag{1.9}$$

se nazývá systém věrohodnostních rovnic a slouží ke stanovení maximálně věrohodných odhadů.

Poznámka 1.6 *Pro skórový vektor náhodného vektoru* $\mathbf{Y}$ *s regulární hustotou* $f(\mathbf{y}; \boldsymbol{\theta})$ *platí (s využitím vztahu* (1.8) *a podmínky 4 z definice 1.1)*

$$E\left(U_i(\boldsymbol{\theta})\right) = E\left(\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \boldsymbol{\theta}_i}\right) = \int\limits_M \frac{\partial \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_i} f(\boldsymbol{y}; \boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{y}) = \mathbf{0}$$

a

$$D\left(U_i(\boldsymbol{\theta})\right) = E\left(U_i^2(\boldsymbol{\theta})\right) = -E\left(U_i'(\boldsymbol{\theta})\right), \tag{1.10}$$

když

$$\int\limits_M \frac{\partial^2 \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_i^2} f(\boldsymbol{y}; \boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{y}) = \mathbf{0}. \tag{1.11}$$

Lemma 1.7 *Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ jsou nezávislé náhodné vektory s regulárními hustotami $f_1(\boldsymbol{y}_1; \boldsymbol{\theta}), \ldots, f_n(\boldsymbol{y}_n; \boldsymbol{\theta})$. Pak pro skórový vektor příslušný sdružené hustotě f náhodných vektorů $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ platí*

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}),$$

kde $\boldsymbol{U}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{U}_n(\boldsymbol{\theta})$ jsou nezávislé skórové vektory příslušné hustotám $f_1(\boldsymbol{y}_1; \boldsymbol{\theta}), \ldots, f_n(\boldsymbol{y}_n; \boldsymbol{\theta})$.

Důkaz je zřejmý.

1.3 Vybrané pojmy z teorie zobecněných lineárních modelů

V dalších kapitolách při statistické analýze populací s NB rozdělením bude potřebné provádět srovnání jednotlivých populací, které je analogické analýze rozptylu pro normální populace a dále provádět úvahy analogické úvahám, které se řeší v regresní analýze. V těchto úvahách lze s výhodou použít teorie zobecněných lineárních modelů. Proto v tomto odstavci připomeneme některé pojmy a základní výsledky této teorie. Kapitola je vypracována podle [19], [46] a [48] kde lze rovněž najít další, detailnější výsledky. V tomto odstavci vystačíme se skalárním parametrem $\theta \in \boldsymbol{\Theta} \subset \mathbb{R}$.

Definice 1.8 *Řekneme, že hustota $f(y; \theta)$, $\theta \in \boldsymbol{\Theta} \subset \mathbb{R}$ je exponenciálního typu, pokud*

$$f(y; \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

kde $b(\theta)$ je tzv. přirozený parametr, $a(y)$ a $d(y)$ jsou známé funkce y a $b(\theta)$ a $c(\theta)$ jsou funkce parametru θ.

Je-li $b(\theta) = \theta$, řekneme, že hustota je v kanonickém tvaru a přirozený parametr je θ.

Poznámka 1.9 *Hustota $f(y; \theta)$ může obsahovat další parametr ϕ, který není předmětem našeho zájmu. Tento parametr nazveme rušivým parametrem. Pak regulární hustotu exponenciálního typu s rušivým parametrem ϕ budeme zapisovat ve tvaru*

$$f(y; \theta) = \exp\left\{\frac{y\theta - \gamma(\theta)}{\psi(\phi)} + d(y, \phi)\right\}, \tag{1.12}$$

kde θ a ϕ jsou parametry, $\gamma(\theta)$ je parametrická funkce, $\psi(\theta)$ je nějaká funkce rušivého parametru, d je funkce y a ϕ (viz [48], [19]).

Poznámka 1.10 *Pro regulární hustotu exponenciálního typu v kanonickém tvaru* (1.12) *platí*

$$U = \frac{\partial}{\partial\theta}\left(\frac{Y\theta - \gamma(\theta)}{\psi(\phi)} + d(Y,\phi)\right) = \frac{Y - \gamma'(\theta)}{\psi(\phi)}$$

a odtud tedy

$$\begin{aligned} E(U) &= \frac{E(Y) - \gamma'(\theta)}{\psi(\phi)}, \\ D(U) &= \frac{DY}{\psi^2(\phi)}, \\ U' &= -\frac{\gamma''(\theta)}{\psi(\phi)}. \end{aligned}$$

Protože podle poznámky 1.6 platí $EU = 0$ a $DU = -E(U')$ dostaneme střední hodnotu a rozptyl Y ve tvaru

$$\begin{aligned} EY &= DU - \gamma'(\theta) = \gamma'(\theta), \\ DY &= \psi(\phi)DU = \gamma''(\theta)\psi(\phi). \end{aligned} \tag{1.13}$$

Definice 1.11 *(Zobecněný lineární model) Nechť $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ je náhodný vektor a nechť rozdělení Y_i závisí na pevných vektorech $x_i = (x_{i1}, \ldots, x_{im})' \in \mathbb{R}^m$ prostřednictvím neznámého vektoru parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$. Matice $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$ má rozměr $n \times m$ a hodnost $m < n$ a dále nechť střední hodnoty $\mu_i = EY_i$ existují. Říkáme, že $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$se řídí zobecněným lineárním modelem (GLM z anglického generalized linear model), jestliže dále platí:*

1. *Rozdělení $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ je exponenciálního typu s regulární sdruženou hustotou pravděpodobnosti*

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\theta_i) = \exp\left\{\sum_{i=1}^{n}\left[\frac{y_i\theta_i - \gamma(\theta_i)}{\psi(\phi_i)} + d(y_i,\phi_i)\right]\right\}. \tag{1.14}$$

2. *Parametr θ_i závisí na $\boldsymbol{x}_i$ a $\boldsymbol{\beta}$ prostřednictvím parametru*

$$\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}, \tag{1.15}$$

tzv. lineárního prediktoru, $i = 1, 2, \ldots, n$.

3. *Existuje známá ryze monotónní diferencovatelná funkce g, kterou budeme nazývat linkovací funkce, a platí*

$$\eta_i = g(\mu_i) \qquad \mu_i = g^{-1}(\eta_i). \tag{1.16}$$

Řekneme, že linkovací funkce je kanonická, pokud

$$\theta_i = \eta_i = g(\mu_i).$$

Matici $\boldsymbol{X} = (\boldsymbol{x}_i')_{i=1}^n$ nazýváme maticí plánu.

Poznámka 1.12 *Když pracujeme místo parametru $\theta \in \boldsymbol{\Theta}$ s vektorovým parametrem $\boldsymbol{\beta}$, který je jako proměnná obsažen v hustotě* (1.14) *vzhledem k reparametrizaci* (1.15) *a zavedení linkovací funkce* (1.16)*, označíme složky příslušného skórového vektoru jako $U_j^*(\boldsymbol{\beta})$ a podobně prvky Fisherovy informační matice jako funkce parametru $\boldsymbol{\beta}$ budou značeny $J_{jk}^*(\boldsymbol{\beta})$.*

Věta 1.13 *Mějme náhodný vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ jehož složky mají hustotu* (1.12)*, který se řídí zobecněným lineárním modelem s linkovací funkcí*

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i, \qquad i = 1, \ldots, n.$$

Dále nechť platí (1.10) *a* (1.13)*. Pak*

$$U_j^* = U_j^*(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{x_{ij}(Y_i - \mu_i)}{DY_i} \frac{\partial \mu_i}{\partial \eta_i}$$

a

$$J_{jk}^* = J_{jk}^*(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{x_{ij}x_{ik}}{DY_i} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Poznámka 1.14 *Pro kanonický link (tj. pro $\theta_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$) platí*

$$\begin{aligned}\boldsymbol{U}_n^{*'}(\boldsymbol{\beta}) &= -\boldsymbol{J}_n(\boldsymbol{\beta}),\\ U_j^* &= \sum_{i=1}^{n} U_{j,i} = \sum_{i=1}^{n} \frac{x_{ij}(Y_i - \mu_i)}{\psi_i(\phi)},\\ J_{jk}^* &= \sum_{i=1}^{n} \frac{x_{ij}x_{jk}}{\psi_i(\phi)} V(\mu_i).\end{aligned}$$

Věta 1.15 *Mějme náhodný vektor $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)'$, který se řídí zobecněným lineárním modelem s maticí plánu $\boldsymbol{X}_{n\times m}$. Předpokládejme, že platí* (1.10) *a* (1.13).

Dále mějme matici $\boldsymbol{C}_{m\times q}$ s hodností $h(\boldsymbol{C}) = q < m$. Platí-li obecná lineární hypotéza: $H_0 : \boldsymbol{C}'\boldsymbol{\beta} = \boldsymbol{0}$, pak Waldova statistika

$$W = \widehat{\boldsymbol{\beta}}_{ML}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{J}_n(\boldsymbol{\beta})^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}'\widehat{\boldsymbol{\beta}}_{ML} \overset{A}{\sim} \chi^2(q), \tag{1.17}$$

kde $\widehat{\boldsymbol{\beta}}_{ML}$ je maximálně věrohodný odhad vektorového parametru $\boldsymbol{\beta}$ (viz [3], [35]).

Poznámka 1.16 *Při známé hodnotě parametru ϕ, hypotézu $H_0 : \boldsymbol{C}'\boldsymbol{\beta} = \boldsymbol{0}$ zamítáme na hladině významnosti $\boldsymbol{\alpha}$, pokud platí*

$$W > \chi^2_{1-\alpha}(q).$$

Je-li $\widehat{\boldsymbol{\beta}}_{ML}$ maximálně věrohodný odhad parametru $\boldsymbol{\beta}$, pak $\widehat{\boldsymbol{\beta}}_{ML}$ konverguje za poměrně obecných předpokladů skoro jistě k $\boldsymbol{\beta}$ (viz [3], [35]). Proto při provádění asymptotických testů aproximujeme při provádění statistických analýz Fisherovu informační matici $\boldsymbol{J}_n(\boldsymbol{\beta})$ maticí $\boldsymbol{J}_n(\widehat{\boldsymbol{\beta}}_{ML})$.

Poznámka 1.17 *Speciálním případem obecné lineární hypotézy*

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{0}$$

jsou testy hypotéz

$$H_0 : \beta_j = 0 \text{ pro } j = 1, 2, \ldots, m.$$

Tento test dostaneme pomocí Waldovy statistiky (1.17) *volbou*

$$\boldsymbol{C} = \boldsymbol{c}_{m\times 1} = (0, \ldots, 1, \ldots, 0)'.$$

Při asymptotickém testu této hypotézy lze rovněž vyjít z asymptotické normality maximálně věrohodných odhadů, tedy ze vztahu

$$\widehat{\beta}_{ML,j} \overset{A}{\sim} N\left(\beta_j, s^*_{jj}\right), \text{ kde } s^*_{jj} = (\boldsymbol{J}_n(\boldsymbol{\beta})^{-1})_{jj},$$

přičemž hypotézu $H_0 : \beta_j = 0$ *zamítáme na asymptotické hladině významnosti* α*, pokud*

$$\frac{|\widehat{\beta}_{ML,j}|}{\sqrt{s^*_{jj}}} > u_{1-\frac{\alpha}{2}},$$

kde opět Fisherovu informační matici $\boldsymbol{J}_n(\boldsymbol{\beta})$ *aproximujeme maticí* $\boldsymbol{J}_n(\widehat{\beta}_{ML})$.

O dalších přístupech k testování hypotéz o parametru $\boldsymbol{\beta}$ se zmíníme v následujícím odstavci.

1.4 Testy hypotéz v modelech s rušivými parametry

V tomto odstavci budeme předpokládat, že je dán náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ z rozdělení s hustotou $f(\mathbf{y}, \boldsymbol{\theta})$ a že neznámý parametrický vektor $\boldsymbol{\theta}$ je tvaru $\boldsymbol{\theta} = (\boldsymbol{\tau}', \boldsymbol{\psi}')'$, kde $\boldsymbol{\tau} = (\theta_1, \ldots, \theta_k)'$ je parametr, který je předmětem našeho zájmu. Budeme mu říkat cílový parametrický vektor a jeho složky nazveme cílové parametry. Vektor $\boldsymbol{\psi} = (\theta_{k+1}, \ldots, \theta_m)'$ není předmětem prováděné statistické inference, budeme mu říkat rušivý parametr. Cílem je testovat hypotézu $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ proti alternativě $\boldsymbol{\tau} \neq \boldsymbol{\tau}_0$ bez ohledu na hodnotu rušivého parametru. V této kapitole uvedeme podle [4], [36] příslušné testovací statistiky pro modely s rušivými parametry. Tyto statistiky pak budou využity při analýze populací s NB rozdělením.

Zaveďme označení

- $\boldsymbol{\theta}_0 = (\boldsymbol{\tau}'_0, \boldsymbol{\psi}'_0)'$ skutečná hodnota parametru $\boldsymbol{\theta}$;
- $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{ML} = (\widehat{\boldsymbol{\tau}}'_{ML}, \widehat{\boldsymbol{\psi}}'_{ML})'$ maximálně věrohodný odhad vektorového parametru $\boldsymbol{\theta}$;
- $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{ML} = (\boldsymbol{\tau}'_0, \widetilde{\boldsymbol{\psi}}'_{ML})'$ maximálně věrohodný odhad parametru $\boldsymbol{\theta}$ za platnosti nulové hypotézy.

Test H_0 proti alternativě, že H_0 neplatí lze provést na základě skórové statistiky S, Waldovy statistiky W nebo věrohodnostního poměru LR (viz [4], [36]).

$$S = \frac{1}{n}\left[\mathbf{U}_1(\widetilde{\boldsymbol{\theta}})'\mathbf{J}_{11.2}^{-1}(\widetilde{\boldsymbol{\theta}})\mathbf{U}_1(\widetilde{\boldsymbol{\theta}})\right], \tag{1.18}$$

$$W = n(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0)'\mathbf{J}_{11.2}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0), \tag{1.19}$$

$$LR = 2\left[\ln f(\mathbf{y};\widehat{\boldsymbol{\theta}}) - \ln f(\mathbf{y};\widetilde{\boldsymbol{\theta}})\right], \tag{1.20}$$

Přičemž v uvedených vzorcích bylo užito označení

$\mathbf{U}_1(\boldsymbol{\theta}) = \left(\frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_k}\right)'$ je skórový vektor cílových parametrů

$\mathbf{J}_{11.2}(\boldsymbol{\theta}) = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{22}$, kde

$\mathbf{J}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{pmatrix}$ je Fisherova informační matice rozdělená do bloků tak,

že matice $\mathbf{J}_{11}$ je typu $k \times k$

$\mathbf{F}_i(\boldsymbol{\theta}) = \left(-\frac{\partial^2 \ln f(Y_i;\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}\right)_{r,s=1}^{m}$

$\mathbf{F}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i(\boldsymbol{\theta})$ je výběrová Fisherova informační matice, pro tuto matici

za obecných předpokladů regularity (viz [36]) platí

$E\mathbf{F}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$ a $\mathbf{F}(\boldsymbol{\theta}_0) \xrightarrow{P} \mathbf{J}(\boldsymbol{\theta}_0)$.

Tedy matice $\mathbf{F}(\boldsymbol{\theta}_0)$ je konsistentním odhadem matice $\mathbf{J}(\boldsymbol{\theta}_0)$. Pro výběrovou Fisherovu matici platí $E\mathbf{F}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$ a $\mathbf{F}(\boldsymbol{\theta}_0) \xrightarrow{P} \mathbf{J}(\boldsymbol{\theta}_0)$.

Tyto statistiky mají za platnosti H_0 asymptoticky χ^2 rozdělení [4], [35]. Nulová hypotéza se zamítá (na asymptotické hladině významnosti α) pokud jsou tyto statistiky větší než kvantil $\chi^2_{1-\alpha}(k)$, kde k je počet parametrů obsažených ve vektoru $\boldsymbol{\tau}$.

Kapitola 2

Negativně binomické rozdělení

Negativně binomické rozdělení patří mezi základní rozdělení pravděpodobnosti diskrétního typu, používá se v řadě aplikací. Velmi časté jsou jeho aplikace v teorii spolehlivosti [16], [17], kontrole jakosti [62], [54], psychologii [38] a zejména v biologii [5], [10], [12], [24], [25], [11], [63], [61], [58], [20]. Protože je tato práce zaměřena především na statistickou analýzu biologických populací, budeme při jeho zavedení používat zejména biologické motivace.

V první části kapitoly nejdříve zmíníme tři přístupy k zavedení negativně binomického rozdělení a dále uvedeme jeho základní charakteristiky (tj. pravděpodobnostní funkci, charakteristickou funkci, momenty, kumulanty), abychom s ohledem na jeho další využití vyčerpávajícím způsobem toto rozdělení popsali.

Druhá část této kapitoly bude věnována různým způsobům reparametrizace negativně binomického rozdělení, které pak budou potřebné při odhadu parametrů tohoto rozdělení. Vhodné reparametrizace umožní najít elegantní a numericky nejschůdnější přístupy k získání odhadů tohoto rozdělení.

Poslední odstavec této kapitoly bude věnován necentrálnímu negativně binomickému rozdělení.

2.1 Zavedení negativně binomického rozdělení a jeho základní charakteristiky

Klasický a nejčastější způsob zavedení NB rozdělení vychází z Bernoulliovské posloupnosti nezávislých alternativních pokusů, kdy pravděpodobnost úspěchu v každém pokusu je $\pi \in (0, 1)$.

Rozdělení náhodné veličiny Y, která udává počet neúspěchů předcházejících κ-tému úspěchu $\kappa \in \{1, 2, \dots\}$ se nazývá negativně binomické rozdělení, π a κ jsou jeho parametry. Snadno nahlédneme, že pro jeho pravděpodobnostní funkci (hustotu vzhledem k čítací míře) platí

$$\begin{aligned} f(y;\kappa,\pi) &= \binom{y+\kappa-1}{y}(1-\pi)^y\pi^\kappa && y = 0, 1, \dots \\ &= 0 && \text{jinak.} \end{aligned} \tag{2.1}$$

V některé literatuře (viz [38], [68]) se toto rozdělení nazývá Pascalovo a NB rozdělení je pak zobecněním tohoto rozdělení pro $\kappa > 0$. Lze snadno ukázat, že při $\kappa > 0$ je funkce $f(y;\kappa,\pi)$ daná vzorcem (2.1) stále hustotou (pravděpodobnostní funkcí) vzhledem k čítací míře. V této práci budeme předpokládat, že pro parametry NB rozdělení platí $0 < \pi < 1$ a $\kappa > 0$.

Snadno nahlédneme, že střední hodnota μ a rozptyl σ^2 jsou tvaru

$$\mu = EY = \kappa\frac{1-\pi}{\pi} \qquad \text{a} \qquad \sigma^2 = DY = \kappa\frac{1-\pi}{\pi^2}. \tag{2.2}$$

Pomocí hustoty (2.1) s ohledem na vzorec (1.1) snadno nahlédneme, že charakteristická funkce $\psi_Y(t)$ NB rozdělení je tvaru

$$\psi_Y(t) = \sum_{y=0}^{\infty} e^{ity}P(Y = y) = \pi^\kappa\left[1-(1-\pi)e^{it}\right]^{-\kappa}. \tag{2.3}$$

Hustotu (2.1) lze snadno přepsat do tvaru

$$f(y;\pi,\kappa) = \binom{-\kappa}{y}\pi^\kappa(\pi-1)^y, \tag{2.4}$$

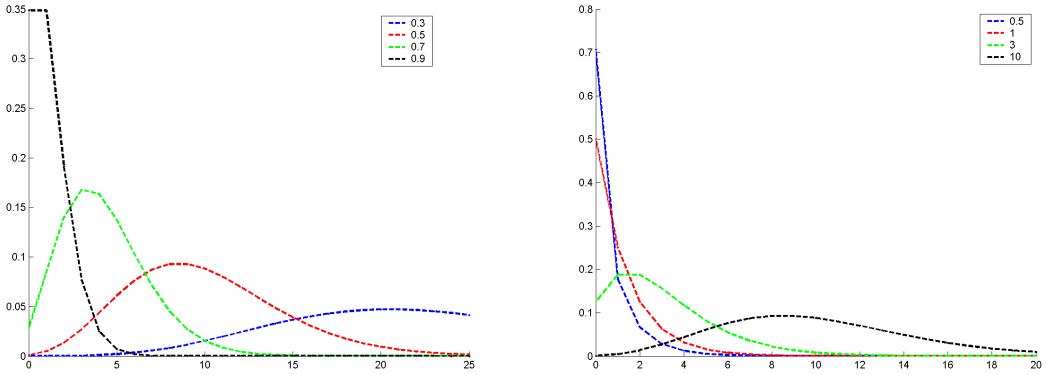
který připomíná pravděpodobnostní funkci binomického rozdělení a podobně charakteristická funkce (2.3) připomíná charakteristickou funkci binomického rozdělení se záporným parametrem $-\kappa$, který odpovídá počtu pokusů. Zavedeme-li navíc nový parametr p vztahem $\pi = \frac{1}{1-p}$, $p \in (0, 1)$ pak pro celočíselné hodnoty parametru κ z intervalu $(-\infty, -\mu)$, kde $\mu = E(Y)$, přímo dostáváme (viz [9]) tvar pravděpodobnostní funkce binomického rozdělení

$$f(y;p,\kappa) = \binom{-\kappa}{y}p^y(1-p)^{-\kappa-y}. \tag{2.5}$$

Proto dostalo toto rozdělení název negativně binomické rozdělení. Tato vlastnost se často využívá ve statistické praxi, když odhad κ vyjde záporný.

Pokud Y bude mít negativně binomické rozdělení tvaru (2.1), budeme v dalším textu používat označení $Y \sim NB(\pi, \kappa)$.

Příklady hustot NB rozdělení jsou pro parametry π a κ na obr. 2.1.



(a) $\kappa = 10$, π postupně 0,3; 0,5; 0,7; 0,9 (b) $\pi = 0{,}5$, κ postupně 0,5; 1; 3; 10

Obrázek 2.1: Pravděpodobnostní funkce NB rozdělení s parametry κ a π. Přestože se jedná o diskrétní rozdělení, je v obrázcích hustota zakreslena spojitě aby více vynikla změna tvaru rozdělení.

Obecné momenty (1.4) hustoty NB rozdělení lze stanovit pomocí Stirlingových čísel 2. druhu (viz odstavec 1.1, vzorec (1.3)) a klesajících faktoriálních momentů (viz odstavec 1.1, vzorec (1.2)). Po jednoduchém výpočtu lze odvodit tvar klesajících faktoriálních momentů pro NB rozdělení

$$\mu_{[d]} = \frac{(\kappa + d - 1)!}{(\kappa - 1)!} \left(\frac{1-\pi}{\pi}\right)^d.$$

Obecné momenty tedy určíme ze vztahu (viz [67])

$$\mu'_n = \sum_{d=1}^{n} \alpha_{d,n-d+1} \mu_{[d]}.$$

Kumulanty pro NB rozdělení můžeme získat ze vztahu (viz [67])

$$\xi_n = \sum_{d=1}^{n} \alpha_{d,n-d+1} (d-1)! \kappa \left(\frac{1-\pi}{\pi}\right)^d. \tag{2.6}$$

Stirlingova čísla 2. druhu jsou pro hodnoty $n = 1, \ldots, 10$ a $d = 1, \ldots, 10$ uvedena v tabulce 2.1. Byla spočtena programem stirling, viz seznam vytvořených programů v kapitole 8.

$d\backslash n$	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	0	1	3	7	15	31	63	127	255	511
3	0	0	1	6	25	90	301	966	3025	9330
4	0	0	0	1	10	65	350	1701	7770	34105
5	0	0	0	0	1	15	140	1050	6951	42525
6	0	0	0	0	0	1	21	266	2646	22827
7	0	0	0	0	0	0	1	28	462	5880
8	0	0	0	0	0	0	0	1	36	750
9	0	0	0	0	0	0	0	0	1	45
10	0	0	0	0	0	0	0	0	0	1

Tabulka 2.1: Stirlingova čísla druhého druhu

Pomocí vzorců (1.4), (2.6) a tabulky 2.1 lze snadno nalézt první čtyři centrální a obecné momenty NB rozdělení, první čtyři kumulanty a rovněž šikmost a špičatost NB rozdělení. Jejich výpočty zde nebudeme provádět, uvedeme pouze výsledné vzorce.

Obecné momenty:

$$\mu'_1 = EY = \frac{\kappa(1-\pi)}{\pi}$$

$$\mu'_2 = EY^2 = \frac{\kappa(1-\pi)}{\pi^2}\left[\kappa(1-\pi)+1\right]$$

$$\mu'_3 = EY^3 = \frac{\kappa(1-\pi)}{\pi^3}\left[\kappa^2(1-\pi)^2 + 3\kappa(1-\pi) + 2 - \pi\right]$$

$$\mu'_4 = EY^4 = \frac{\kappa(1-\pi)}{\pi^4}\left[\kappa^3(1-\pi)^3 + 6\kappa^2(1-\pi)^2 + \kappa(1-\pi)(11-4\pi) + \right.$$
$$\left. + \pi^2 - 6\pi + 6\right]$$

Kumulanty a centrální momenty:

$$\xi_1 = EY = \frac{\kappa(1-\pi)}{\pi}$$
$$\xi_2 = \mu_2 = DY = \frac{\kappa(1-\pi)}{\pi^2}$$
$$\xi_3 = \mu_3 = \frac{\kappa(1-\pi)(2-\pi)}{\pi^3}$$
$$\xi_4 = \mu_4 - 3\xi_2^2 = \frac{\kappa(1-\pi)(\pi^2-6\pi+6)}{\pi^4}$$
$$\mu_4 = \xi_4 + 3\xi_2^2 = \frac{\kappa(1-\pi)\left[3\kappa(1-\pi)+\pi^2-6\pi+6\right]}{\pi^4}$$

Šikmost:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{2-\pi}{\sqrt{\kappa(1-\pi)}}$$

Špičatost:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\pi^2-6\pi+6}{\kappa(1-\pi)}$$

Na zápis hustoty NB rozdělení pomocí parametrů π a κ se budeme odkazovat jako na parametrizaci 1.

Hustota NB rozdělení přechází, pro $\kappa \to \infty$, limitně v hustotu Poissonova rozdělení. Této vlastnosti se využívá při hledání odhadů κ. Vychází-li odhad κ vysoký, často se doporučuje k modelování dat využít, místo NB rozdělení, rozdělení Poissonovo (viz [61]).

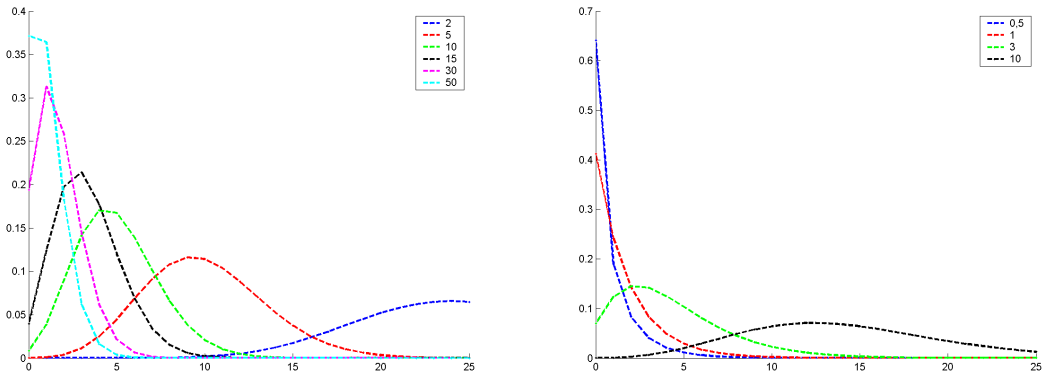
2.2 Negativně binomické rozdělení jako Gamma-Poissonův smíšený model

Při zpracování biologických dat ([44], [66], [21]) se často setkáváme se situacemi, kdy zkoumané veličiny mají Poissonovo rozdělení s parametrem Z, který se náhodně na dílčích subpopulacích mění. Je-li Z náhodná veličina s rozdělením gamma $G(\varphi, \kappa)$ a podmíněné rozdělení veličiny Y za podmínky $Z = z$ je Poissonovo $Po(z)$ lze snadno stanovit nepodmíněnou hustotu veličiny Y, kterou označíme $f(y; \kappa, \varphi)$, jako marginální hustotu ke sdružené hustotě $f(y, z)$ náhodných veličin Y a Z. Postupně pomocí podmíněné hustoty $f_{Y|Z}$ náhodné veličiny Y za podmínky $Z = z$ dostaneme (viz [40])

$$f(y;\kappa,\varphi) = \int_0^\infty f(y,z)dz = \int_0^\infty f_{Y|Z}(y;z)f_Z(z;\varphi,\kappa)dz \qquad (2.7)$$
$$= \binom{y+\kappa-1}{y}\left(\frac{1}{1+\varphi}\right)^y\left(\frac{\varphi}{1+\varphi}\right)^\kappa.$$

Uvedený model pro hustotu f se v liteatuře nazývá Gamma-Poissonův smíšený model (anglicky Gamma-Poisson mixture model), viz monografie [47], která je těmto modelům věnována.

Když se v hustotě (2.7) položí $\pi = \frac{\varphi}{1+\varphi}$, snadno nahlédneme, že se jedná o hustotu NB rozdělení s parametry π a κ.



(a) $\kappa = 50$, φ postupně 2, 5, 10, 15, 30, 50 (b) $\varphi = 0,7$, κ postupně 0,5; 1; 3; 10

Obrázek 2.2: Pravděpodobnostní funkce NB rozdělení s parametry κ a φ. Přestože se jedná o diskrétní rozdělení je v obrázcích hustota zakreslena spojitě aby více vynikla změna tvaru rozdělení.

Příklady hustot NB rozdělení jsou pro parametry κ a φ na obr. 2.2.

V této parametrizaci lze získat vyjádření centrálních a necentrálních momentů a

kumulantů ve tvaru

$$\begin{aligned}
\mu'_1 &= \frac{\kappa}{\varphi},\\
\mu'_2 &= \frac{\kappa}{\varphi} + \frac{\kappa^2}{\varphi^2} + \frac{\kappa}{\varphi^2},\\
\mu'_3 &= \frac{\kappa}{\varphi} + 3\,\frac{\kappa^2}{\varphi^2} + 3\,\frac{\kappa}{\varphi^2} + \frac{\kappa^3}{\varphi^3} + 3\,\frac{\kappa^2}{\varphi^3} + 2\,\frac{\kappa}{\varphi^3},\\
\mu'_4 &= \frac{\kappa}{\varphi} + 7\,\frac{\kappa^2}{\varphi^2} + 7\,\frac{\kappa}{\varphi^2} + 6\,\frac{\kappa^3}{\varphi^3} + 18\,\frac{\kappa^2}{\varphi^3} + 12\,\frac{\kappa}{\varphi^3} + \frac{\kappa^4}{\varphi^4} +\\
&\quad + 6\,\frac{\kappa^3}{\varphi^4} + 11\,\frac{\kappa^2}{\varphi^4} + 6\,\frac{\kappa}{\varphi^4}.
\end{aligned}$$

Kumulanty a centrální momenty

$$\begin{aligned}
\xi_1 &= EY = \frac{\kappa}{\varphi},\\
\xi_2 &= \mu_2 = \frac{\kappa(1+\varphi)}{\varphi^2},\\
\xi_3 &= \mu_3 = \frac{\kappa(1+\varphi)(2+\varphi)}{\varphi^3},\\
\xi_4 &= \mu_4 - 3\xi_2^2 = \frac{\kappa(1+\varphi)(\varphi^2+6\varphi+6)}{\varphi^4}.
\end{aligned}$$

Pro šikmost dostameme

$$\gamma_1 = \frac{2+\varphi}{\sqrt{\kappa(1+\varphi)}}$$

a špičatost je tvaru

$$\gamma_2 = \frac{\varphi^2 + 6\,\varphi + 6}{\kappa\ (1+\varphi)}.$$

Využití tohoto přístupu k modelování biologických populací bude dále využito při analýze populace kopytníků v odstavci 7.1.

2.3 NB rozdělení v populační dynamice

V populační dynamice při modelování růstu populací (viz [53]) se často vychází z lineárního procesu růstu (Yuleova procesu [42]). Stručně připomeňme zavedení

tohoto procesu podle [6], [60], [53].

Předpokládejme, že počáteční velikost populace je a jedinců. Každý jedinec dá za časový interval délky Δt vzniknout novému s pravděpodobností $\lambda\Delta t + o(\Delta t)$ (pro $\Delta t \to 0$ a $\lambda > 0$ je parametr). Pravděpodobnost, že jedinec dá za časový interval délky Δt vznik více než jednomu jedinci, je $o(\Delta t)$. $X(t)$ bude značit velikost populace v čase t, $t \geq 0$. Protože počáteční velikost populace je a, platí $X(0) = a$. Náhodný proces $X(t)$, který vyhovuje uvedeným předpokladům, se nazývá lineární proces růstu nebo též Yuleův proces.

Pravděpodobnost, že velikost populace v čase $t > 0$ bude k jedinců, označíme $p_k(t)$. Pravděpodobnost, že v čase $t + \Delta t$ bude mít populace a jedinců je rovna pravděpodobnosti, že v čase t bylo v populaci a jedinců a za časový okamžik Δt žádný nový jedinec nevznikl.

Celkově

$$p_a(t + \Delta t) = p_a(t)\,(1 - a\lambda\Delta t - o(\Delta t))$$

a odtud

$$\frac{1}{\Delta t}\,(p_a(t + \Delta t) - p_a(t)) = -a\lambda p_a(t) - \frac{o(\Delta t)}{\Delta t} p_a(t).$$

Pro $\Delta t \to 0$ dostáváme lineární diferenciální rovnici pro $p_a(t)$

$$\frac{dp_a(t)}{dt} = a\lambda p_a(t).$$

Podobně pro velikost populace $k > a$ dostaneme soustavu diferenciálních rovnic

$$\frac{dp_k(t)}{dt} = \lambda(k-1)p_{k-1}(t) - \lambda k p_k(t), \qquad k = a+1, a+2, \ldots.$$

Tuto soustavu diferenciálních rovnic lze řešit pomocí vytvořujících funkcí (viz [6]). Při počáteční podmínce $p_a(0) = 1$ dostáváme řešení

$$p_k(t) = \binom{k-1}{a-1} e^{-a\lambda t}(1 - e^{-\lambda t})^{k-a} \text{ pro } k \geq a.$$

Označme $N(t) = X(t) - a$ nárůst populace v čase t. Pak

$$P(N(t) = n) = p_{n+a}(t) = \binom{n+a-1}{a-1} e^{-a\lambda t}(1 - e^{-\lambda t})^n \text{ pro } n = 0, 1, 2, \ldots.$$

Nárůst populace $N(t)$ má tedy negativně binomické rozdělení s parametry $\kappa = a$ a $\pi = e^{-\lambda t}$.

Uvedené zavedení NB rozdělení opravňuje v mnohých experimentálních situacích domněnku, že počty jedinců v sledované biologické populaci mají NB rozdělení (viz [44], [66], [21]).

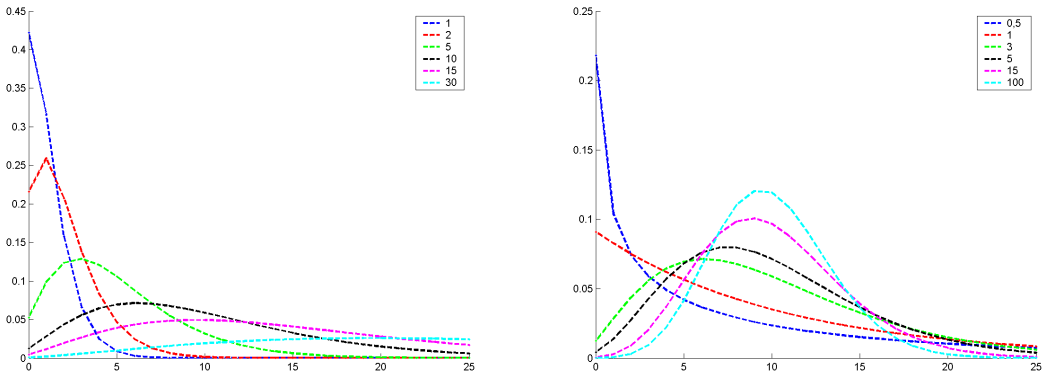
2.4 Další způsoby reparametrizace NB rozdělení

V dalších analýzách bude často třeba testovat hypotézy týkající se střední hodnoty. Proto se často v literatuře používá reparametrizace s parametry $\mu = EY$ a κ (viz např. [5], [2], [13], [28], [44], [43], [66]).

Pro střední hodnotu NB rozdělení v parametrech π a κ podle (2.2) platí $\mu = \frac{\kappa(1-\pi)}{\pi}$ a odtud můžeme vyjádřit $\pi = \frac{\kappa}{\kappa+\mu}$. Po dosazení do (2.1) dostaneme vyjádření hustoty v nových parametrech.

Pomocí těchto nových parametrů pak píšeme $Y \sim NB(\mu, \kappa)$. Zapíšeme-li navíc binomický koeficient $\binom{y+\kappa-1}{y}$ ve vzorci (2.1) pomocí funkce gamma, dostaneme pravděpodobnostní funkci negativně binomického rozdělení ve tvaru

$$\begin{aligned} f(y;\mu,\kappa) &= \frac{\Gamma(y+\kappa)}{\Gamma(y+1)\Gamma(\kappa)} \left(\frac{\mu}{\kappa+\mu}\right)^y \left(\frac{\kappa}{\kappa+\mu}\right)^\kappa && y = 0, 1, \dots \\ &= 0 && \text{jinak.} \end{aligned}$$



(a) $\kappa = 3$, μ postupně 1, 2, 5, 10, 15, 30 (b) $\mu = 0.7$, κ postupně 0,5; 1; 3; 5; 15; 100

Obrázek 2.3: Pravděpodobnostní funkce NB rozdělení s parametry κ a μ. Přestože se jedná o diskrétní rozdělení je v obrázcích hustota zakreslena spojitě aby více vynikla změna tvaru rozdělení.

Příklady hustot NB rozdělení jsou pro parametry μ a κ na obr. 2.3. Z obrázků je dále vidět role parametrů: s rostoucím μ roste rozptyl a rozdělení se stává „plošší“, s rostoucím κ se mění tvar rozdělení, rozdělení se stává „symetričtější“.

V této parametrizaci dostaneme obecné momenty ve tvaru

$$\begin{aligned}
\mu'_1 &= \mu, \\
\mu'_2 &= \mu + \mu^2 + \frac{\mu^2}{\kappa}, \\
\mu'_3 &= \mu + 3\,\mu^2 + 3\,\frac{\mu^2}{\kappa} + \mu^3 + 3\,\frac{\mu^3}{\kappa} + 2\,\frac{\mu^3}{\kappa^2}, \\
\mu'_4 &= \mu + 7\,\mu^2 + 7\,\frac{\mu^2}{\kappa} + 6\,\mu^3 + 18\,\frac{\mu^3}{\kappa} + 12\,\frac{\mu^3}{\kappa^2} + \mu^4 + 6\,\frac{\mu^4}{\kappa} + \\
&\quad + 11\,\frac{\mu^4}{\kappa^2} + 6\,\frac{\mu^4}{\kappa^3}
\end{aligned}$$

a kumulanty a centrální momenty ve tvaru

$$\begin{aligned}
\xi_1 &= EY = \mu, \\
\xi_2 &= \mu_2 = \mu + \frac{\mu^2}{\kappa} = \mu\left(1 + \frac{\mu}{\kappa}\right) = \frac{\mu(\kappa+\mu)}{\kappa}, \\
\xi_3 &= \mu_3 = \frac{\mu(\kappa+\mu)(1+2\mu)}{\kappa^2}, \\
\xi_4 &= \mu_4 - 3\xi_2^2 = \frac{\mu(\kappa+\mu)(\kappa^2+6\kappa\mu+6\mu^2)}{\kappa^3}.
\end{aligned}$$

Dále pro šikmost dostaneme vztah

$$\gamma_1 = \frac{\kappa + 2\mu}{\sqrt{\kappa\mu(\kappa+\mu)}}$$

a pro špičatost platí

$$\gamma_2 = \frac{\kappa^2 + 6\kappa\mu + 6\mu^2}{\kappa\,\mu\,(\kappa+\mu)\,.}$$

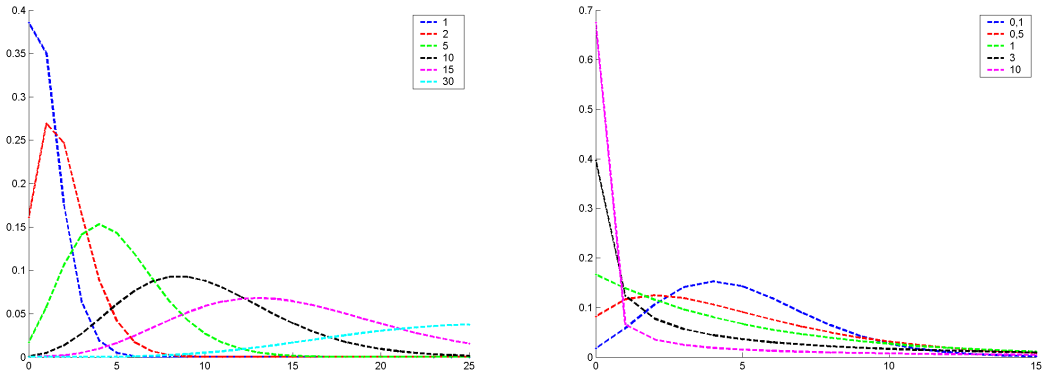
V tomto zápisu, který bude dále často využíván, je $EY = \mu$ a $DY = \mu + \frac{\mu^2}{\kappa}$. Ze vzorce pro rozptyl je dobře patrné, že při daném κ je rozptyl kvadratickou funkcí μ.

Na zápis hustoty NB rozdělení pomocí parametrů μ a κ se budeme odkazovat jako na parametrizaci 2.

Protože nalézt odhad parametru κ je v některých situacích obtížné (zvláště pro

velké κ a malé μ), dávají někteří autoři (viz [7], [13], [46], [57], [31], [18]) přednost parametrizaci $\mu,\ c$, kde $c = \frac{1}{\kappa}$. Pravděpodobnostní funkci lze poté zapsat ve tvaru

$$f(y;\mu,c) = Pr(Y = y_i;\mu,c) = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{y}\left(\frac{1}{1+c\mu}\right)^{c^{-1}}.$$



(a) $c = 3$, μ postupně 1, 2, 5, 10, 15, 30 (b) $\mu = 5$, c postupně 0,1; 0,5; 1; 3; 10

Obrázek 2.4: Pravděpodobnostní funkce NB rozdělení s parametry κ a π. Přestože se jedná o diskrétní rozdělení je v obrázcích hustota zakreslena spojitě aby více vynikla změna tvaru rozdělení.

Příklady hustot NB rozdělení jsou pro parametry μ a c na obr. 2.4.

Příslušné momenty jsou tvaru

$$\begin{aligned}
\mu'_1 &= \mu,\\
\mu'_2 &= \mu + \mu^2 + \mu^2 c,\\
\mu'_3 &= \mu + 3\,\mu^2 + 3\,\mu^2 c + \mu^3 + 3\,\mu^3 c + 2\,\mu^3 c^2,\\
\mu'_4 &= \mu + 7\,\mu^2 + 7\,\mu^2 c + 6\,\mu^3 + 18\,\mu^3 c + 12\,\mu^3 c^2 + \mu^4 +\\
&\quad + 6\,\mu^4 c + 11\,\mu^4 c^2 + 6\,\mu^4 c^3.
\end{aligned}$$

Kumulanty a centrální momenty dostáváme ve tvaru

$$\begin{aligned}
\xi_1 &= EY = \mu \\
\xi_2 &= \mu_2 = \mu\,(1 + c\mu)\,, \\
\xi_3 &= \mu_3 = (1 + c\mu)(1 + 2\,c\mu), \\
\xi_4 &= \mu_4 - 3\xi_2^2 = \mu(1 + c\mu)(1 + 6\,c\mu + 6\,c^2\mu^2).
\end{aligned}$$

Šikmost je dána jako

$$\gamma_1 = \frac{1 + 2c\mu}{\sqrt{\mu(1 + c\mu)}}$$

a špičatost je

$$\gamma_2 = \frac{1 + 6\,c\mu + 6\,c^2\mu^2}{m(1 + c\mu)}.$$

Na zápis hustoty NB rozdělení pomocí parametrů μ a c se budeme odkazovat jako na parametrizaci 3.

2.5 Necentrální negativně binomické rozdělení

Pro úplnost v této kapitole ještě zavedeme podle [56] necentrální negativně binomické rozdělení, které lze získat jako smíšený model negativně binomického rozdělení a Poissonova rozdělení.

Nechť náhodná veličina Y má negativně binomické rozdělení s parametry k a π. Nechť parametr k je také náhodnou veličinou a platí pro něj $K = N + v$, kde N je náhodná veličina s Poissonovým rozdělením ($N \sim Po(\lambda)$) a v je konstanta.

Tedy hustota N je tvaru

$$f_N(n; \lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

a podmíněná hustota Y při daném N je tvaru

$$f_{Y|N}(y; n + v, \pi|n) = \binom{y + n + v - 1}{y}(1 - \pi)^y\pi^{n+v}.$$

Podle Bayesovy věty vyjádříme nepodmíněnou hustotu f_Y náhodné veličiny Y pomocí sdružené hustoty f_{YN} a po jednoduché úpravě dostaneme

$$\begin{aligned} f_Y(y; v, \lambda, \pi) &= \sum_{n=0}^{\infty} f_{YN}(y, n) = \sum_{n=0}^{\infty} f_N(n; \lambda) f_{Y|N}(y; n + v, \pi|n) = \\ &= \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \binom{y + n + v - 1}{y} (1 - \pi)^y \pi^{n+v} = \\ &= e^{-\lambda} (1 - \pi)^y \pi^v \frac{1}{y!} \frac{\Gamma(v + y)}{\Gamma(v)} \left[\frac{\Gamma(v)}{\Gamma(v + y)} \sum_{n=0}^{\infty} \frac{\Gamma(y + n + v)}{n! \Gamma(n + v)} (\lambda\pi)^n \right] = \\ &= e^{-\lambda} (1 - \pi)^y \pi^v \frac{(v)_y}{y!} \sum_{n=0}^{\infty} \frac{(y + v)_n}{n! (v)_n} (\lambda\pi)^n. \end{aligned}$$

Pomocí hypergeometrické funkce (1.5) lze tuto hustotu zapsat ve tvaru

$$f_Y(y; v, \lambda, \pi) = e^{-\lambda} (1 - \pi)^y \pi^v {}_1F_1(y + v, v, \lambda\pi).$$

S využitím Kumerovy transformace (1.6) a zobecněných Laguerrových polynomů (1.7) lze hustotu upravit na tvar

$$f_Y(y; v, \lambda, \pi) = e^{-\lambda(1-\pi)} \pi^v (1 - \pi)^y L_y^{v-1}(-\lambda\pi).$$

Rozdělení náhodné veličiny Y s touto hustotou se nazývá necentrální negativně binomické rozdělení s parametry v, λ, π.

Jeho charakteristiky jsou dány následovně

$$EY = \frac{1 - \pi}{\pi}(v + \lambda), \qquad DY = \frac{1 - \pi}{\pi}(v + \lambda) + \left(\frac{1 - \pi}{\pi} \right)^2 (v + 2\lambda).$$

Poznámka 2.1 *Pro $\lambda \to 0$ hustota necentrálního NB rozdělení konverguje k hustotě NB rozdělení.*

Kapitola 3

Odhady parametrů NB rozdělení

Cílem této kapitoly je popsat metody vhodné pro stanovaní odhadů parametrů NB rozdělení, uvést algoritmy pro jejich výpočet, provést srovnání jednotlivých metod odhadu s ohledem na biologické populace a doporučit pro daný typ populace nejvhodnější metodu odhadu.

Klasické metody odhadu, tedy metoda momentů (MM), metoda maximální věrohodnosti (ML) dávají odhady, které mohou být značně vychýlené. Navíc momentové odhady parametru κ mohu být záporné, což komplikuje jeho interpretaci. Konečně pro NB rozdělení je typické, že jeho střední hodnota μ je menší než rozptyl, ovšem na reálných datech, i pro výběry z NB rozdělení, může dojít k situaci, že výběrový průměr $\overline{Y}$ je větší než výběrový rozptyl S^2. Potvrzují to simulace. Tato situace pak komplikuje další výpočty, zejména výpočet ML odhadu. Uvedená situace bude dále demonstrována na příkladu simulovaných dat a je dobře známa z biologických analýz. Proto se v biologické praxi při provádění statistické inference o parametrech NB rozdělení rozlišují tři situace (viz [13], [57], [61]).

První a nejčastější situace nastává je-li výběrový rozptyl větší než výběrový průměr (tzv. „overdispersion“). Tato situace nastává jsou-li organismy ve shlucích v prostoru či čase. Momentový odhad parametru κ nabývá hodnot větších než 0.

Pokud je výběrový průměr roven výběrovému rozptylu je možno NB rozdělení nahradit Poissonovým, což je limitní případ NB rozdělení pro $\kappa \to \infty$.

V situacích, kdy je pro dané parametry střední hodnota blízká rozptylu, naopak výběrový průměr často překročí výběrový rozptyl (tzv. „underdispersion“) a tato data pak vedou k záporným momentovým odhadům κ. Tato situace nastává, je-li rozmístění organismů pravidelnější než předpokládá Poissonovo rozdělení. V článcích

[8], [9], [13], [61] je ukázána souvislost NB rozdělení (pro záporné hodnoty κ) s binomickým rozdělením (za předpokladu $\kappa \in (-\infty, -\mu)$, viz (2.5)). V této souvislosti Clapham v [12] říká, že „overdispersion“ znamená, že pro jedince je jednodušší usadit se poblíže jiného jedince a pro „underdispersion“ platí opak. Na základě těchto úvah pak Piegorsch [57] doporučuje dát přednost parametrizaci μ, c, která shrnuje předešlé tři situace pro c z intervalu $c \in (-\frac{1}{\mu}, \infty)$.

S ohledem na biologické interpretace populací s negativně binomickým rozdělením bude tato kapitola uspořádána následujícím způsobem. Nejdříve budou uvedeny klasické metody odhadu parametrů NB rozdělení a to metoda momentů (MM) a metoda maximální věrohodnosti (ML). Dále bude uvedena korekce vychýlení maximálně věrohodných odhadů. Konečně, abychom se vyhnuli numerickým nestabilitám při výpočtu maximálně věrohodných odhadů, bude uvedena quasi-likelihood (QL) metoda a její modifikace. Na závěr bude uveden bayesovský přístup k odhadu parametru κ.

3.1 Metoda momentů

Předpokládejme, že je dán náhodný výběr $Y_1, \ldots, Y_n$ z NB rozdělení. Odhady MM snadno získáme řešením momentových rovnic $\mu'_k = M'_k$, kde M'_k jsou obecné výběrové momenty viz [4], [41]. S ohledem na tři nejčastější parametrizace dostáváme odhady ve tvaru:

1. Nechť $Y_i \sim NB(\pi, \kappa)$

$$\widehat{\pi} = \frac{\overline{Y}}{M_2}, \qquad \widehat{\kappa} = \frac{\overline{Y}^2}{M_2 - \overline{Y}}.$$

2. Nechť $Y_i \sim NB(\mu, \kappa)$

$$\widehat{\mu} = \overline{Y}, \qquad \widehat{\kappa} = \frac{\overline{Y}^2}{M_2 - \overline{Y}}.$$

3. Nechť $Y_i \sim NB(\mu, c)$

$$\widehat{\mu} = \overline{Y}, \qquad \widehat{c} = \frac{M_2 - \overline{Y}}{\overline{Y}^2}.$$

$M_2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ je druhý centrální výběrový moment.

Odhady získané touto metodou se často používají jako počáteční odhady při iterativním řešení nelineárních rovnic, které často vycházejí pří hledání maximálně

věrohodných odhadů.

Z výrazů pro $\widehat{\kappa}$ a $\widehat{c}$ je ihned zřejmé, že pro $M_2 < \overline{Y}$ tyto odhady vychází záporné, což může působit výše zmíněné nejen interpretační, ale i numerické problémy a to zejména při automatickém použití těchto odhadů jako počátečních odhadů při iterativním hledání řešení věrohodnostních rovnic.

Odhady získané touto metodou budou dále označeny indexem MM, tedy $\widehat{\kappa}_{MM}$, $\widehat{\pi}_{MM}$, $\widehat{\mu}_{MM}$, $\widehat{c}_{MM}$.

3.2 Maximálně věrohodné odhady

Stanovení ML odhadů pro NB rozdělení provedeme pro všechny tři parametrizace. I když je možné stanovit maximálně věrohodné odhady při různých parametrizacích pomocí Zehnovy věty, která popisuje princip invariance pro maximálně věrohodné odhady (viz [69]), uvádíme dále věrohodnostní rovnice pro různé parametrizace. Jejich numerické řešení totiž dává při vhodné reparametrizaci s ohledem na hodnoty parametrů kvalitnější numerické řešení. Tato situace bude v závěru odstavce ilustrována numerickým příkladem.

- Parametrizace 1

  Jestliže $Y_i \sim NB(\pi, \kappa)$, pak příslušná logaritmická pravděpodobnostní funkce výběru je tvaru

  $$l(\pi, \kappa; \mathbf{y}) = \sum_{i=1}^{n} \left(y_i \ln(1-\pi) + \kappa \ln \pi + \ln \Gamma(y_i + \kappa) - \ln \Gamma(\kappa) - \ln y_i! \right).$$

  Odtud dostaneme systém věrohodnostních rovnic

  $$\begin{aligned} \frac{\partial l}{\partial \pi} &= -\frac{\sum_{i=1}^{n} y_i}{1-\pi} + \frac{\kappa}{\pi} = 0 \\ \frac{\partial l}{\partial \kappa} &= n \ln \pi + \sum_{i=1}^{n} \Psi(y_i + \kappa) - n\Psi(\kappa) = 0, \end{aligned}$$

  kde $\Psi(z)$ je digamma funkce tj. derivace logaritmu gamma funkce. Z první rovnice dostaneme

  $$\widehat{\pi} = \frac{\kappa}{\kappa + \overline{Y}},$$

  druhá se řeší iterativně.

Tyto odhady jsou implementovány v MATLABu - funkce nbinfit, pro další výpočty byla použita tato funkce.

- Parametrizace 2

  Jestliže $Y_i \sim NB(\mu, \kappa)$, pak příslušná logaritmická věrohodnostní funkce výběru je tvaru

$$l(\mu, \kappa; \mathbf{y}) = \sum_{i=1}^{n} \Big[\ln \Gamma(y_i + \kappa) - \ln \Gamma(y_i + 1) - \ln \Gamma(\kappa) + \kappa \ln \frac{\kappa}{\kappa + \mu} + y_i \ln \frac{\mu}{\kappa + \mu} \Big].$$

  Odtud dostamene věrohodnostní rovnice

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= -n \frac{\kappa}{\kappa + \mu} + \frac{\kappa}{\mu(\kappa + \mu)} \sum_{i=1}^{n} y_i = 0 \\ \frac{\partial l}{\partial \kappa} &= \sum_{i=1}^{n} \left[\Psi(y_i + \kappa) - \Psi(\kappa) + \ln \frac{\kappa}{\kappa + \mu} + \frac{\mu}{\kappa + \mu} - y_i \frac{1}{\kappa + \mu} \right] = 0. \end{aligned}$$

  Je vidět, že maximálně věrohodný odhad μ lze jednoduše stanovit pomocí výběrového průměru. Odhad κ je třeba hledat numericky. Pro jeho výpočet je dále použita procedura založená na Newton-Raphsonově iterativní metodě. Jako počáteční odhad κ je použit odhad získaný metodou momentů.

  Program NB_k_mle viz seznam použitých programů v kapitole 8.

- Parametrizace 3

  Jestliže $Y_i \sim NB(\mu, c)$, pak logaritmická věrohodnostní funkce výběru je tvaru

$$l(\mu, c; \mathbf{y}) = \sum_{i=1}^{n} \Big[\ln \Gamma(y_i + c^{-1}) - \ln y! - \ln \Gamma(c^{-1}) + c^{-1} \ln \frac{1}{1 + c\mu} + y_i \ln \frac{c\mu}{1 + c\mu} \Big]. \tag{3.1}$$

  Odtud dostaneme věrohodnostní rovnice

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= -\frac{n}{1 + c\mu} + \frac{1}{\mu(1 + c\mu)} \sum_{i=1}^{n} y_i = 0 \\ \frac{\partial l}{\partial c} &= \sum_{i=1}^{n} \left[-\frac{1}{c^2} \Psi(y_i + c^{-1}) + \frac{1}{c^2} \Psi(c^{-1}) + \frac{1}{c^2} \ln(1 + c\mu) + \frac{y_i - \mu}{c(1 + c\mu)} \right] = 0. \end{aligned}$$

Řešením první rovnice je opět výběrový průměr, druhá se řeší numerickou iterací.

Program NB_c_mle viz seznam použitých programů v kapitole 8.

Odhady získané touto metodou označíme indexem ML. Tedy např. $\widehat{\kappa}_{ML}$.

Tato metoda dává vychýlený odhad parametru κ (popř. c). V programech je použita digamma a trigamma funkce, které jsou v MATLABu definovány pouze v reálném oboru. Ovšem pokud se stane, že odhad κ (c) během výpočtu vyjde záporný vychází digamma a trigamma funkce pro záporný argument. Tato situace je v programech NB_k_mle a MN_c_mle ošetřena tak, že je odhad posunut na hodnotu *eps* a znovu se vstoupí do iteračního cyklu. Pokud se tato situace opakuje vícekrát je výpočet ukončen.

Podle zkušeností autorky práce jsou ML odhady parametru κ nebo c většinou rovnocenné. V běžných výběrech dávají (po přepočtu) stejné výsledky. Problém nastává u výběrů, kdy je κ velké a obzvlášť je-li zároveň μ malé. Pro tuto situaci je výhodnější parametrizace μ a c. V těchto případech program NB_k_mle často selhává (řešení buď není nalezeno, nebo je nalezen odhad $\kappa = \infty$), program NB_c_mle většinou dá odhad c velmi blízký 0. V těchto situacích je výhodnější použít jiný postup založený na aproximaci NB rozdělení rozdělením Poissonovým či binomickým (viz odstavec 2.1) nebo použít jiné metody odhadu jako jsou quasi-likelihood metody a zejména bayesovský přístup k hledání odhadů.

Uvedené situace budou nyní demonstrovány na simulovaných datech.

Příklad 3.1 *Pro výběr z NB rozdělení s parametry $\mu = 2$ a $\kappa = 100$ rozsahu 30 byla simulací získána realizace uvedená v tabulce 3.1. Ze získaných dat byly spočteny*

varianta	0	1	2	3	4
četnost	6	6	2	10	6

Tabulka 3.1: Realizace výběru rozsahu 30 z NB rozdělení s parametry $\mu = 2$ a $\kappa = 100$.

momentové odhady pro parametrizace 2 a 3: $\widehat{\kappa}_{MM} = 82{,}4889$ a $\widehat{c}_{MM} = 1/\widehat{\kappa}_{MM} = 0{,}0121$.

Maximálně věrohodné odhady pro uvedené parametrizace vycházejí odlišně (z důvodů numerické nestability). Maximálně věrohodný odhad v parametrizaci 1 vychází $\widehat{\kappa}_{ML_1} = 2{,}8551 \cdot 10^6$, pro parametrizaci 2 $\widehat{\kappa}_{ML_2} = \infty$, pro parametrizaci 3 $\widehat{\kappa}_{ML_3} = 1/\widehat{c}_{ML_3} = 1/(8{,}9589 \cdot 10^6) = 1{,}1162 \cdot 10^{14}$.

Příklad 3.2 *Pro výběr z NB rozdělení s parametry $\mu = 2$ a $\kappa = 100$ rozsahu 30 byla získána nová realizace uvedená v tabulce 3.2. Ze získaných dat byly spočteny*

varianta	0	1	2	3	4	5	7	9
četnost	2	3	2	8	10	1	3	1

Tabulka 3.2: Realizace výběru rozsahu 30 z NB rozdělení s parametry $\mu = 2$ a $\kappa = 100$.

momentové odhady pro parametrizace 2 a 3 $\widehat{\kappa}_{MM} = -39{,}4091$ a $\widehat{c}_{MM} = -0{,}0254$. V MATLABu je pro takovéto situace doporučeno místo funkce nbinfit a NB rozdělení použít Poissonovo rozdělení. Funkce NB_k_mle dává odhad $\widehat{\kappa}_{ML_2} = \infty$ a funkce NB_c_mle dává odhad $\widehat{\kappa}_{ML_3} = 1/c_{ML_3} = 1/(7.356 \cdot 10^{-16} = 1.3593 \cdot 10^{15})$.

Tedy v této situaci je dobré místo hledání ML odhadů NB rozdělení použít aproximaci Poissonovým rozdělením nebo zvolit jiné odhady, např. quasi-likelihood nebo bayesovské. O těchto odhadech bude pojednáno v dalším odstavci.

3.3 Korekce vychýlení maximálně věrohodného odhadu

Vzhledem k výše uvedeným numerickým nestabilitám budeme dále pracovat s parametrizací 3, tedy s parametry μ a c.

Maximálně věrohodný odhad parametru c nebývá nestranný. V práci [15] je popsána aproximace vychýlení $b_c(\boldsymbol{\theta})$ (viz [63]) odhadu parametru pomocí třetích derivací věrohodnostní funkce, která je řádu n^{-1}. Pro NB rozdělení je aproximace vychýlení odhadu parametru c a μ stanovena v práci [63]. V práci [63] nejsou odhady počítány přímo pomocí Γ funkce, ale je použita Stirlingova aproximace. V této práci byly odvozeny odhady s využitím Γ funkce.

Odhady získané touto metodou označíme indexem BC (z anglického Bias Corrected).

Dříve, nežli popíšeme vychýlení odhadu c, zavedeme označení $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (m, c)'$ pro vektor parametrů a dále U, V a W pro první, druhou a třetí derivaci logaritmické věrohodnostní funkce (3.1). Tedy

$$U_r^{(i)} = \frac{\partial l_i}{\partial \theta_r},\ V_{rt}^{(i)} = \frac{\partial^2 l_i}{\partial \theta_r \partial \theta_t},\ W_{rtu}^{(i)} = \frac{\partial^3 l_i}{\partial \theta_r \partial \theta_t \partial \theta_u},\ r, t, u = 1, 2,$$

a dále položme

$$J_{rt} = E\left(-\sum_{i=1}^{n} V_{rt}^{(i)}\right),\ I_{rtu} = E\left(\sum_{i=1}^{n} W_{rtu}^{(i)}\right),\ K_{r,tu} = E\left(\sum_{i=1}^{n} U_{r}^{(i)} V_{tu}^{(i)}\right)$$

a $\mathbf{M}$ budeme značit matici inverzní k Fisherově informační matici, tedy

$$\mathbf{M} = \left(M^{ij}\right)_{i,j=1,2} = \mathbf{J}^{-1}.$$

Po technickém výpočtu, který je poněkud zdlouhavý a proto je uveden v dodatku A.1, dostaneme

$$\begin{aligned}
M^{11} &= \frac{\mu(1+c\mu)}{n} & M^{22} &= -\frac{1}{n}\left[\frac{\mu}{c^2(1+c\mu)} + \frac{\Delta_1 - \Psi'(c^{-1})}{c^4}\right]^{-1} \\
I_{111} &= 2\frac{1+2c\mu}{\mu^2(1+c\mu)^2} & K_{1,11} &= -\frac{1}{2} I_{111} \\
I_{112} &= \frac{n}{(1+c\mu)^2} & K_{1,12} &= -I_{112}.
\end{aligned}$$

Dále

$$\begin{aligned}
I_{122} &= 0 \qquad K_{2,12} = -\frac{\mu}{c(1+c\mu)^2} + \frac{1}{c^2(1+c\mu)^2}(\Delta_{y0} - \mu\gamma) \\
I_{222} &= n\left\{-\frac{\mu(4+5c\mu)}{c^3(1+c\mu)^2} - \frac{6}{c^5}\left[\Delta_1 - \Psi'(c^{-1})\right] - \frac{1}{c^6}\left[\Delta_2 - \Psi''(c^{-1})\right]\right\} \\
K_{2,22} &= n\left\{\frac{3+4c\mu}{c^4(1+c\mu)^2}\left(\Delta_{y0} - \mu\gamma\right) - \frac{\mu(1+2c\mu)}{c^3(1+c\mu)^2} + \frac{2}{c^5}(\gamma^2 - \Delta_{00}) + \right. \\
&\quad \left. + \frac{1}{c^6}(\gamma\Delta_1 - \Delta_{01}) + \frac{1}{c^5(1+c\mu)}(\Delta_{y1} - \mu\Delta_1)\right\},
\end{aligned}$$

kde

$$\begin{aligned}
\gamma &= \ln(1+c\mu) + \Psi(c^{-1}) & \Delta_0 &= E(\Psi(y+c^{-1})) \\
\Delta_{00} &= E(\Psi^2(y+c^{-1})) & \Delta_1 &= E(\Psi'(y+c^{-1})) \\
\Delta_2 &= E(\Psi''(y+c^{-1})) & \Delta_{y0} &= E(y\Psi(y+c^{-1})) \\
\Delta_{y1} &= E(y\Psi'(y+c^{-1})) & \Delta_{01} &= E(\Psi(y+c^{-1})\Psi'(y+c^{-1})).
\end{aligned}$$

Označíme-li dále $b_c(\widehat{\mu}_{ML}, \widehat{c}_{ML})$ aproximaci vychýlení ML odhadu parametru c, kde $b_c(\mu_{ML}, c_{ML}) = E(\widehat{c}_{ML} - c)$ je vychýlení odhadu $\widehat{c}_{ML}$ a dále $\widehat{c}_{BC}$ odhad $\widehat{c}_{ML}$

korigovaný na nestrannost, pak (viz [63]) platí

$$b_c(\mu, c) = \frac{1}{2}[(M^{22})^2(I_{222} + 2K_{2,22}) + M^{11}M^{22}(I_{112} + 2K_{1,12})]$$

a tedy

$$\widehat{c}_{BC} = \widehat{c}_{ML} - b_c(\widehat{\mu}_{ML}, \widehat{c}_{ML}).$$

Protože ML odhad parametru μ je roven výběrovému průměru a je tedy nestranný, není třeba ho korigovat. I přesto je zde korekce uvedena, neboť v [63] je uvedena chybně. Podobně jako pro c lze nalézt i odhad vychýlení pro parametr μ:

$$b_\mu(\mu_{ML}, c_{ML}) = \frac{1}{2}[(M^{11})^2(I_{111} + 2K_{1,11}) + M^{11}M^{22}(I_{122} + 2K_{1,12})]$$

a pro korigovaný odhad poté platí $\widehat{\mu}_{BC} = \widehat{\mu}_{ML} - b_\mu(\widehat{\mu}_{ML}, \widehat{c}_{ML})$.

Analogicky lze odvodit korekci vychýlení ML odhadů pro parametrizaci μ a κ. Tato korekce je uvedena v dodatku A.2.

Ve většině případů vede tato korekce ke zlepšení odhadu.

Tato korekce byla naprogramována, viz funkce bias_corr_c v kapitole 8.

3.4 Quasi-likelihood přístupy

Jak bylo uvedeno v předchozích odstavcích, je výpočet ML odhadů numericky značně náročný. Pro takové situace se doporučuje použít quasi-likelihood (QL) přístup. QL odhady vycházejí z vlastností logaritmické věrohodnostní funkce pro hustoty exponenciálního typu v kanonickém tvaru (1.12). Jsou konstruovány tak, že se logaritmická věrohodnostní funkce nahradí tzv. quasi-likelihood funkcí, která je pro další použití jednodušší a ponechá si základní vlastnosti logaritmické věrohodnostní funkce. Vlastnosti tzv. quasi-likelihood funkce jsou podrobně popsány v [45]. Dále jsou tyto odhady analyzovány např. v [46], [13], [55], [32], [34], [33]. Přístup ke konstrukci QL odhadů na základě uvedené literatury dále stručně popíšeme.

Budeme uvažovat náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, jehož složky jsou nezávislé a mají hustotu exponenciálního typu v kanonickém tvaru (1.12). Dále předpokládejme, že $\mathbf{Y}$ má střední hodnotu $\boldsymbol{\mu}$ a varianční matici $\sigma^2\mathbf{V}(\boldsymbol{\mu})$, kde σ^2 je konstanta a

varianční matice $\mathbf{V}(\boldsymbol{\mu})$ je diagonální matice známých funkcí, tedy

$$\mathbf{V}(\boldsymbol{\mu}) = diag\{V_1(\mu_1), \ldots, V_n(\mu_n)\}.$$

Dále ze vztahu skórového vektoru a Fisherovy matice $\mathbf{J}$ dostáváme:

$$U_i = \frac{\partial l_i}{\partial \mu_i} = \frac{Y_i - \mu_i}{D(Y_i)} = \frac{Y_i - \mu_i}{\sigma^2 V(\mu_i)}$$

a odtud zřejmě (viz (1.10))

$$\begin{aligned} E(U_i) &= 0, \\ D(U_i) &= E(U_i^2) = 1/[\sigma^2 V(\mu_i)], \\ E\left(\frac{\partial U_i}{\partial \mu_i}\right) &= E\left(\frac{\partial^2 l_i}{\partial \mu_i^2}\right) = -E\left(\frac{\partial l_i}{\partial \mu_i}\right)^2 = -D(U_i) = -J_{ii} = -1/[\sigma^2 V(\mu_i)]. \end{aligned}$$

Nyní zavedeme quasi-likelihood (QL) funkci (podle [46]) jako integrál

$$l_{Q_i}(\mu; y_i) = \int_{y_i}^{\mu} \frac{y_i - t}{\sigma^2 V(t)} dt. \tag{3.2}$$

Pro derivaci l_{Q_i} dostaneme

$$\frac{\partial l_{Q_i}(\mu; y_i)}{\partial \mu} = \frac{y_i - \mu_i}{\sigma^2 V_i(\mu_i)}.$$

Dále lze analogicky jako byla zavedena statistika LR (1.20) zavést quasi-devianční funkci (viz [55])

$$D_i(y_i; \mu) = -2\sigma^2[l_{Q_i}(\mu; y_i) - l_{Q_i}(y_i; y_i)] = 2\int_{\mu}^{y_i} \frac{y_i - t}{V(t)} dt. \tag{3.3}$$

Tato funkce závisí pouze na y_i a μ a nezávisí na σ^2.

Všechny výše zmíněné úvahy byly vztaženy pouze k μ, přičemž rušivý parametr σ^2 se předpokládal konstantní. Další úvahy lze rozšířit i na rušivý parametr σ^2.

Zavedeme místo QL funkce funkci $l_{Q^+} = l_{Q^+}(\mu, \sigma^2; y_i)$ tzv. extended quasi-likelihood funkci (EQL), která bude analogií QL funkce (3.2), ale bude mít dříve popsané vlastnosti logaritmické věrohodnostní funkce i vzhledem k parametru σ^2.

Funkci $l_{Q^+}(\mu, \sigma^2; y_i)$ zavedeme (podle [46]) vztahem

$$l_{Q^+}(\mu, \sigma^2; y_i) = l_Q(\mu; y_i) + h(\sigma^2; y_i), \tag{3.4}$$

kde $h(\sigma^2; y_i)$ je vhodná funkce σ^2 a y. Pomocí (3.3) lze l_{Q^+} upravit na tvar

$$l_{Q^+}(\mu, \sigma^2; y_i) = -\frac{D(y_i; \mu)}{2\sigma^2} + h(\sigma^2; y_i),$$

kde funkci $h(\sigma^2; y_i)$ uvažujeme ve formě

$$h(\sigma^2; y_i) = -\frac{1}{2}h_1(\sigma^2) - h_2(y_i)$$

pro vhodně zvolené h_1 a h_2.

Funkce l_{Q^+} má vzhledem k μ stejné vlastnosti jako funkce l_Q a podobně pro σ^2 musí platit $E(\partial l_{Q^+}/\partial\sigma^2) = 0$. Tedy po úpravě

$$\sigma^4 h_1'(\sigma^2) = E[D(Y_i; \mu)].$$

Jako první aproximaci lze použít $E[D(Y_i; \mu)] = \sigma^2$ a odtud dostaneme $h_1(\sigma^2) = \log(\sigma^2) + const$, tedy

$$l_{Q^+}(\mu, \sigma^2; y_i) = -\frac{D(y_i; \mu)}{2\sigma^2} - \frac{1}{2}\log\sigma^2.$$

Zlepšení se dosáhne (viz [46]), pokud se k funkci $h_1(\sigma^2)$ přičte $\log(2\pi V(y_i))$ (což je konstanta vzhledem k σ^2).

Výslednou EQL funkci lze poté zapsat jako (viz [46])

$$l_{Q^+}(\mu, \sigma^2; y_i) = -\frac{D(y_i; \mu)}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2 V(y_i)). \tag{3.5}$$

Nelder a Pregibon [55] ve svém článku na str. 226 navrhují, z numerických důvodů, modifikaci empirické varianční funkce pro binomické, Poissonovo a negativně binomické rozdělení. O této modifikaci pro NB rozdělení bude pojednáno dále.

Double extended quasi-likelihood (DEQL) funkce vychází z dalšího zpřesnění aproximace $E[D(Y_i; \mu)]$. Pomocí kumulantů vyšších řádů, lze tuto střední hodnotu

aproximovat následovně (viz [46])

$$E[D(Y_i;\mu)] \cong \sigma^2\{1+\sigma^2(2V'^2/V - 3V'')/12\}, \tag{3.6}$$

kde V je varianční funkce.

Quasi-likelihood (respektive EQL, DEQL) funkce pro náhodný výběr $Y_1,\ldots,Y_n$ se získá jako součet jednotlivých QL (resp. EQL, DEQL) funkcí. Jejich maximalizací se pak získají QL (EQL, DEQL) odhady.

Quasi-likelihood (EQL a DEQL) metody je vhodné použít k odhadu parametrů NB rozdělení obzvláště v situacích, kdy je numericky náročné stanovit ML odhady. Lze je použít i v situaci, kdy vychází odhady parametrů záporné.

3.5 Quasi-likelihood odhady pro negativně binomické rozdělení

Budeme uvažovat parametrizaci 3 (tedy parametry μ a c). Interval přípustných hodnot pro parametr c je totiž narozdíl od intervalu pro parametr κ souvislý ($c \in (-\frac{1}{\mu},\infty)$).

- Quasi-likelihood (QL) přístup k odhadu parametru μ

  Pro $NB(\mu,c)$ lze varianční funkce zapsat ve tvaru $V(\mu)=\mu(1+c\mu)$. Potom po dosazení do (3.2) a úpravě dostaneme Quasi-likelihood funkci ve tvaru

  $$l_Q = \sum_{i=1}^{n} l_{Q_i} = \sum_{i=1}^{n}\left[y_i \ln(\mu) + \left(y_i + c^{-1}\right)\ln\left(\frac{1+cy_i}{1+c\mu}\right) - y_i \ln(y_i)\right].$$

  Pro její derivaci platí

  $$\frac{\partial l_Q}{\partial \mu} = \sum_{i=1}^{n}\left[\frac{y_i}{\mu} - \frac{1+cy_i}{(1+c\mu)}\right] = U.$$

- Extended quasi-likelihood (EQL) funkci dostaneme po dosazení do (3.5) s využitím (3.3) jako

  $$l_{EQ_i} = l_{Q_i} - \frac{1}{2}\ln(2\pi\sigma^2 V(y_i)). \tag{3.7}$$

Pro binomické, Poissonovo a negativně binomické rozdělení se z numerických důvodů doporučuje použití modifikované varianční funkce (viz [55], [13]), kterou lze pro NB rozdělení zapsat ve tvaru

$$V(y, \frac{1}{6}) = \frac{(1+cy)^2(y+\frac{1}{6})(c^{-1}+\frac{1}{6})}{(y+c^{-1}+\frac{1}{6})}.$$

Dosazením do (3.7) a úpravou se dostane výsledný tvar extended quasi-likelihood funkce

$$l_{EQ} = \sum_{i=1}^{n} \left[y_i \ln \mu + (y_i + c^{-1}) \ln \frac{1+cy_i}{1+c\mu} - y_i \ln y_i - \ln(1+cy_i) - \frac{1}{2}\ln(y_i + \frac{1}{6}) \right.$$
$$\left. -\frac{1}{2}\ln(c^{-1}+\frac{1}{6}) + \frac{1}{2}\ln(y_i + c^{-1} + \frac{1}{6}) - \frac{1}{2}\ln 2\pi \right].$$

Derivace l_{EQ} podle μ je totožná s derivací funkce l_Q.

Pro derivaci podle c dostáváme

$$\frac{\partial l_{EQ}}{\partial c} = \sum_{i=1}^{n} \left[-\frac{1}{c^2}\ln\left(\frac{1+cy_i}{1+c\mu}\right) + \frac{y_i - \mu}{c(1+c\mu)} - \frac{y_i}{1+cy_i} + \frac{3}{c(6+c)} + \right.$$
$$+\frac{1+6y_i}{2(6cy_i+6+c)} - \frac{1}{2c} = -\frac{1}{c^2}\ln\left(\frac{1+cy_i}{1+c\mu}\right) + \frac{y_i-\mu}{c(1+c\mu)} - \frac{y_i}{1+cy_i} +$$
$$\left. +\frac{1+6y_i}{2(6cy_i+6+c)} - \frac{1}{2(6+c)} \right]$$

a pro druhou derivaci

$$\frac{\partial^2 l_{EQ}}{\partial c^2} = \sum_{i=1}^{n} \left[\frac{2}{c^3}\ln\left(\frac{1+cy_i}{1+c\mu}\right) - \frac{y_i-\mu}{c^2(1+c\mu)(1+cy_i)} - \frac{(y_i-\mu)(1+2c\mu)}{c^2(1+c\mu)^2} + \right.$$
$$\left. +\frac{y_i^2}{(1+cy_i)^2} - \frac{(1+6y_i)^2}{2(6cy_i+6+c)^2} + \frac{1}{2(6+c)^2} \right].$$

- Double extended quasi-likelihood (DEQL) funkci získáme dosazením do (3.4) s využitím (3.6) a (3.7) jako

$$l_{DEQ_i} = l_{EQ_i} - \frac{1}{2}(2V'^2/V - 3V'')/12,$$

kde $V = y_i(1+cy_i)$.

Po úpravě se získá DEQL funkce ve tvaru

$$l_{DEQ} = \sum_{i=1}^{n} \left[y_i \ln \mu + (y_i + c^{-1}) \ln \frac{1+cy_i}{1+c\mu} - (y_i + \frac{1}{2}) \ln y_i - \frac{1}{2} \ln (1+cy_i) + \right.$$
$$\left. + \frac{c}{12(1+cy_i)} - \frac{c}{12} - \frac{1}{12y_i} - \frac{1}{2} \ln 2\pi \right].$$

Derivace l_{DEQ} podle μ je opět totožná s derivací funkce l_Q.

Pro její první a druhou derivaci podle c dostáváme

$$\frac{\partial l_{DEQ}}{\partial c} = \sum_{i=1}^{n} \left[-\frac{1}{c^2} \ln \left(\frac{1+cy_i}{1+c\mu} \right) + \frac{y_i - \mu}{c(1+c\mu)} - \frac{y_i}{2(1+cy_i)} + \frac{1}{12(1+cy_i)^2} - \frac{1}{12} \right]$$

$$\frac{\partial^2 l_{DEQ}}{\partial c^2} = \sum_{i=1}^{n} \left[\frac{2}{c^3} \ln \left(\frac{1+cy_i}{1+c\mu} \right) - \frac{y_i - \mu}{c^2(1+c\mu)(1+cy_i)} - \frac{(y_i - \mu)(1+2c\mu)}{c^2(1+c\mu)^2} + \right.$$
$$\left. + \frac{y_i^2}{2(1+cy_i)^2} - \frac{y_i}{6(1+cy_i)^3} \right].$$

Odhady parametrů se opět získají maximalizací EQL případně DEQL funkce. Rovnice pro odhad parametru c jsou nelineární, odhad parametru μ je stejně jako při metodě maximální věrohodnosti roven výběrovému průměru.

Metoda byla algoritmizována a výsledný program QL_c viz kapitola 8.

3.6 Bayesovský odhad

Protože ML odhady parametrů selhávají pro případ, že $\overline{Y} > M_2'$, lze použít bayesovský přístup k hledání těchto odhadů (viz [4]). Pro NB rozdělení s parametry μ a κ je tento přístup navržen v práci [2].

V této práci je zavedena podmíněná apriorní hustota parametru μ při daném $\sigma^2 = s^2$ vztahem $f_{\mu|\sigma^2}(m|s^2) = 1/s^2$ pro $0 \leq m < s^2$ a rovna 0 jinak. Dále apriorní hustota σ^2 je $f_{\sigma^2}(s^2) = 1/s^2$ pro $s^2 > 0$ (tzv. Jeffery apriorní hustota). Pak sdružená apriorní hustota (μ, σ^2) je $f_{\mu,\sigma^2}(m, s^2) = 1/s^4$ pro $0 \leq m < s^2 < \infty$.

Pro stanoveni aposteriorní hustoty $f_{\mu,\sigma^2|\mathbf{Y}}(m, s^2|\mathbf{y})$ je v [2] použita aproximace hustoty náhodného výběru $Y_1, \ldots, Y_n$ normálním rozdělením na základě centrální limitní věty (CLV).

Nechť $Y_i \sim NB(\pi, \kappa)$, pak pro přirozené κ můžeme psát $Y_i = \sum_{j=1}^{\kappa} X_j$, kde X_j jsou nezávislé stejně rozdělené veličiny z $NB(\pi, 1)$. Tedy pro velké hodnoty κ má Y_i (podle CLV) asymptoticky normální rozdělení $N(\mu, \sigma^2)$, $0 < \mu < \sigma^2$.

Pak lze podle Bayesovy věty zapsat aposteriorní hustotu $f_{\mu,\sigma^2|\mathbf{Y}}(m, s^2|\mathbf{y})$ ve tvaru

$$f_{\mu,\sigma^2|\mathbf{Y}}(m, s^2|\mathbf{y}) = \frac{f_{\mu,\sigma^2}(m, s^2) f_{Y|\mu,\sigma^2}(y|m, s^2)}{\int_0^{\infty} \int_0^{s^2} f_{\mu,\sigma^2}(m, s^2) f_{Y|\mu,\sigma^2}(y|m, s^2)\ dm\ ds^2}. \tag{3.8}$$

Pro hustotu $f_{\mathbf{Y}|\mu,\sigma^2}(\mathbf{y}|m, s^2)$ dále na základě uvedené aproximace můžeme psát

$$f_{\mathbf{Y}|\mu,\sigma^2}(\mathbf{y}|m, s^2) = (2\pi)^{-\frac{n}{2}} (s^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - m)^2}{s^2}\right\}$$

a po dosazení do (3.8) lze aproximovanou aposteriorní hustotu vyjádřit ve tvaru

$$f_{\mu,\sigma^2|\mathbf{Y}}(m, s^2|\mathbf{y}) = \frac{(s^2)^{-\frac{n}{2}-2} e^{-\frac{1}{2s^2}\sum (y_i - m)^2}}{\int_0^{\infty} \int_0^{s^2} s^{-\frac{n}{2}-2} e^{-\frac{1}{2s^2}\sum (y_i - m)^2}\ dm\ ds^2}.$$

Pomocí této hustoty lze stanovit bayesovský odhad μ

$$\widehat{\mu}_B = E(\mu|\mathbf{y}) = \int_0^{\infty} \int_0^{s^2} m f(m, s^2|\mathbf{y})\ dm\ ds^2,$$

který lze podle [2] aproximovat výrazem

$$E(\mu|\mathbf{y}) \cong \overline{y} - \frac{E\left[\frac{\sigma}{\sqrt{n}}\left(\phi\left(\frac{(\sigma^2 - \overline{y})\sqrt{n}}{\sigma}\right) - \phi\left(-\frac{\overline{y}\sqrt{n}}{\sigma}\right)\right)\right]}{E\left[\Phi\left(\frac{(\sigma^2 - \overline{y})\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\overline{y}\sqrt{n}}{\sigma}\right)\right]},$$

kde $\phi(z)$ je hustota standardizovaného normálního rozdělení a Φ je její distribuční funkce. Střední hodnoty na pravé straně uvedeného výrazu jsou definovány vzhledem k modifikovanému apriornímu rozdělení parametru σ^2 při daném $\mathbf{y}$. Detaily viz [2].

Podobně pro σ^2

$$\widehat{\sigma}_B^2 = E(\sigma^2|\mathbf{y}) \cong \frac{E\left[\sigma^2\left(\Phi\left(\frac{(\sigma^2 - \overline{y})\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\overline{y}\sqrt{n}}{\sigma}\right)\right)\right]}{E\left[\Phi\left(\frac{(\sigma^2 - \overline{y})\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\overline{y}\sqrt{n}}{\sigma}\right)\right]}$$

a konečně bayesovský odhad parametru κ je tvaru

$$\widehat{\kappa}_B = \frac{\widehat{\mu}_B^2}{\widehat{\sigma}_B^2 - \widehat{\mu}_B}.$$

Protože $\widehat{\mu}_B \leq \widehat{\sigma}_B^2$ je vždy $\widehat{\kappa}_B > 0$.

Uvedený typ odhadu je naprogramován v programu Bayes_k (viz seznam programů v kapitole 8).

Přestože odvození tohoto odhadu je založeno na aproximaci NB rozdělení normálním rozdělením, tedy pro situaci kdy je κ velké přirozené, fungují tyto odhady dobře i pro ostatní situace (viz [2] a dále odstavec 3.7).

3.7 Porovnání odhadů

Porovnání jednotlivých metod odhadu parametrů NB rozdělení bylo provedeno pomocí simulací. Pozornost byla zaměřena na situace, kdy ML metoda selhává, tedy pro případ, že je „malé“ μ a „velké“ κ. Pro tuto situaci byly porovnány metody MM, EQL, DEQL a Bayesovské. Metoda ML nebyla uvažována, protože (jak je vidět z předchozích příkladů) v těchto situacích selhává. Dále pak bylo použito aproximace výběru pomocí binomického (na základě vzorce (2.5), na tuto metodu aproximace se dále budeme odvolávat zkratkou BiM) a Poissonova rozdělení podle doporučení z odstavce 2.1.

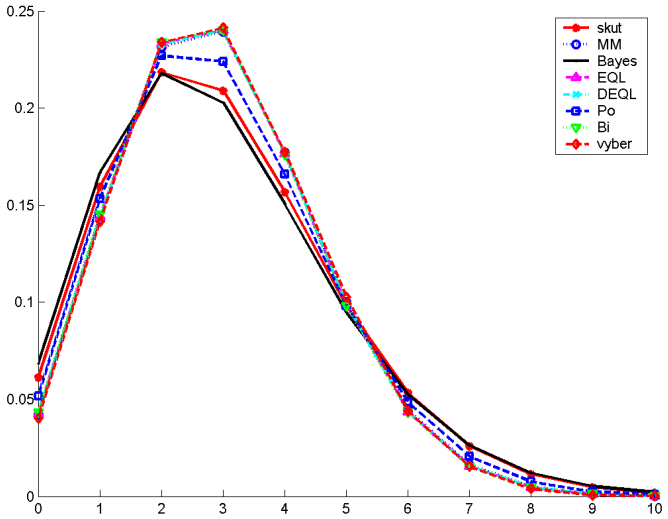
Průběh simulace: 1000 krát se nasimuloval výběr z $NB(\mu, \kappa)$ pro $\mu = 3$ a $\kappa = 20$ rozsahu 30 a 500 s výběrovým rozptylem menším než výběrový průměr, pro každý se nalezly odhady μ a κ. Jako výsledný odhad se vzal průměr z 1000 nalezených odhadů.

	MM	Bayes	EQL	DEQL	BiM
$n = 30$	-23.945	13.742	-23.093	-23.057	23.169
$n = 500$	-267.23	60.93	-265.92	-265.89	265.89

Tabulka 3.3: V tabulce jsou uvedeny průměrné odhady parametru $\kappa = 20$ pro simulace rozsahu 30 a 500 pro uvedené metody odhadu.

Odhady parametrů získáváme z výběrů a proto pro výběry s rozptylem menším než průměr tyto odhady budou většinou značně vzdálené od skutečné hodnoty (mohou vycházet i záporné) viz tabulka 3.3, ale rozdělení, které získáme dosazením odhadnutých parametrů, může přesto být dosti blízké původnímu rozdělení.

V následujících obrázcích (3.1 a 3.2) jsou zachyceny hustoty pro parametry odhadnuté MM, Baysovské odhady, EQL a DEQL. Dále je použita aproximace Poissonovým (PoM) a Binomickým (BiM) rozdělením. Odhady MM a oba QL vyšly záporně (viz tabulka 3.3) a pro vykreslení aproximujeme NB rozdělení binomickým rozdělením. Přestože se jedná o diskrétní rozdělení jsou pro větší přehlednost jednotlivé body spojeny.



Obrázek 3.1: Porovnání odhadnutých hustot s hustotou teoretickou a výběrovou pro rozsahy výběru $n = 30$.

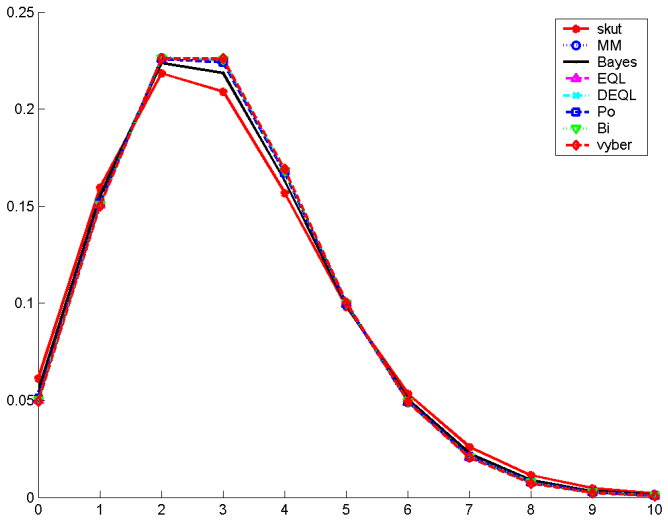
Z obrázků 3.1 a 3.2 se dá usoudit, že v těchto situacích Bayesův odhad funguje nejlépe. Lépe je možné tento fakt demonstrovat pomocí „vzdáleností“ srovnávaných rozdělení.

Pro porovnání odchylky dvou diskrétních rozdělení s hustotami p a q lze použít řadu charakteristik založených na f-divergencích (viz [30], [39], [65]). Zde použijeme 3 statistiky:

$$I(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \qquad \text{I divergence}$$

$$D_{1/2}(p,q) = 2\left(1 - \sum_x \left(p(x)q(x)\right)^{1/2}\right) \qquad \text{Hellingerova vzdálenost}$$

$$\chi^2(p,q) = \sum_x \frac{\left(p(x) - q(x)\right)^2}{q(x)} \qquad \chi^2 \text{ divergence}$$



Obrázek 3.2: Porovnání odhadnutých hustot s hustotou teoretickou a výběrovou pro rozsahy výběru $n = 500$.

Ještě uveďme vzdálenosti daných funkcí:

	MM	BiM	EQL	DEQL	Bayes	PoM	výb
I div.	0,01844	0,01984	0,01966	0,01959	*0,00126*	0,00410	0,02683
Hell. vzdál.	0,00969	0,01020	0,01015	0,01013	*0,00264*	0,00345	0,01252
χ^2 div.	0,05163	0,05788	0,05676	0,05627	*0,00185*	0,01070	0,11114

Tabulka 3.4: f-divergence uvedených odhadnutých hustot k teoretické pro $n = 30$

	MM	BiM	EQL	DEQL	Bayes	PoM
I divergence	0,00057	0,00055	0,00049	*0,00047*	0,02888	0,00693
Hellinger distance	0,00034	0,00031	0,00028	*0,00027*	0,01689	0,00393
χ^2 divergence	0,00090	0,00093	0,00079	*0,00075*	0,04682	0,01160

Tabulka 3.5: f-divergence uvedených odhadnutých hustot k výběrové pro $n = 30$

Z tabulek je patrné, že hustota odhadnutá s využitím Bayesovských odhadů je nejblíže teoretické hustotě. Naopak hustoty získané pomocí MM a QL odhadů jsou blízké výběrové hustotě (četnostem).

Další simulace byla provedena pro výběry s rozptylem větším než průměr. Tato simulace byla použita pro srovnání kvality odhadu parametru c. Pro $\mu = 2$ a $c \in \{0{,}5; 0{,}1; \ldots; 2\}$ bylo postupně vygenerováno 1000 výběrů (nevyhovující byly

	MM	BiM	EQL	DEQL	Bayes	PoM	výb
I div.	0,00595	0,00592	0,00593	0,00592	*0,00214*	0,00505	0,00712
Hell. vzdál.	0,00283	0,00282	0,00282	0,00282	*0,00108*	0,00242	0,00332
χ^2 div.	0,01414	0,01403	0,01404	0,01403	*0,00474*	0,01178	0,01794

Tabulka 3.6: f-divergence uvedených odhadnutých hustot k teoretické pro $n = 500$

	MM	BiM	EQL	DEQL	Bayes	PoM
I divergence	5,99	5,88	*5,87*	5,88	124,72	15,82
Hellinger distance	3,18	3,14	*3,13*	3,14	65,01	8,26
χ^2 divergence	11,03	10,73	*10,72*	10,73	232,76	29,67

Tabulka 3.7: f-divergence uvedených odhadnutých hustot k výběrové pro $n = 500$, všechny hodnoty jsou $\cdot 10^{-5}$

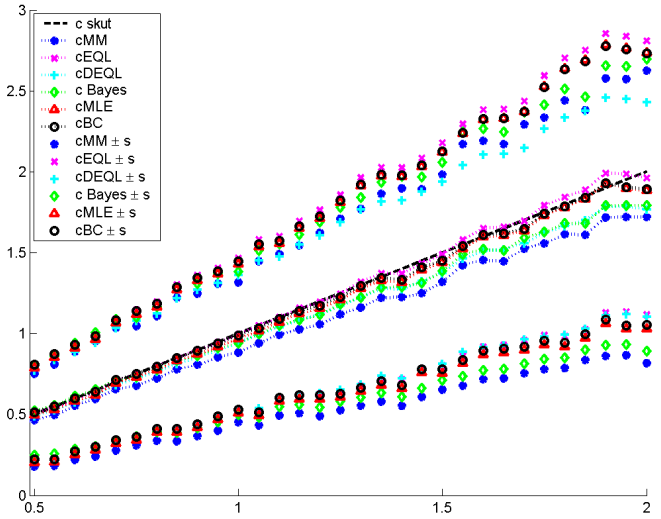
odstraněny) a spočteny odhady (MM, EQL, DEQL, Bayes, ML, BC). Byl spočten průměr příslušného odhadu a jeho výběrová směrodatná odchylka. Tyto hodnoty byly vyneseny do grafu a pro přehlednost byly průměry odhadů spojeny čarou. Simulace byla opakována pro rozsahy $n = 30$ a 50.

V obrázcích 3.3 a 3.4 je na ose x vynesen parametr $c \in \{0{,}5; 0{,}1; \ldots; 2\}$. Skutečná hodnota parametru c je pro přehlednost vynesena černou přerušovanou čarou. Modrou hvězdičkou spojenou tečkovanou čarou je vynesen momentový odhad parametru c a modrou hvězdičkou bez spojení je vynesena $\pm$ jeho výběrová směrodatná odchylka. Růžovým x je vynesen průměrný extended quasi-likelihood odhad $\pm$ jeho směrodatná odchylka, světle modrým + je vynesen průměrný DEQL odhad ($\pm s$), zelený kosočtverec znázorňuje průměrný bayesovský odhad ($\pm s$), červený trojúhelník znázorňuje ML odhad ($\pm s$) a černý kroužek ML odhad opravený na nestrannost ($\pm s$).

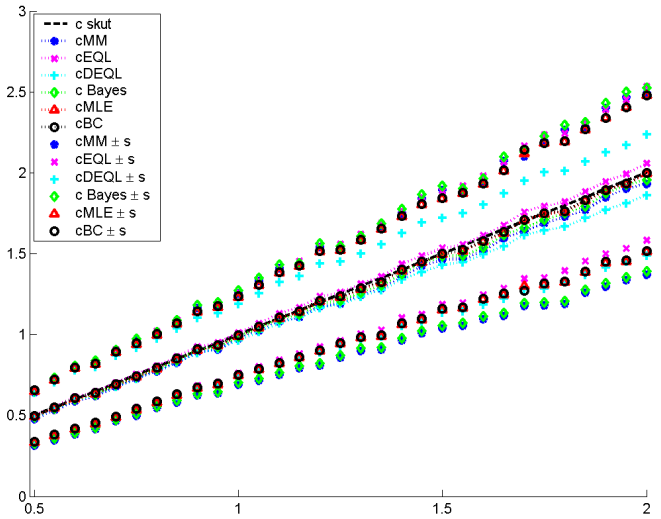
Jako nejlepší odhady (pro situace kdy je výběrový rozptyl větší než výběrový průměr) vychází ML odhady a ML odhady korigované na nestrannost. Velmi kvalitní se zdají být i EQL odhady, které lze navíc použít i pro výběry kdy je výběrový rozptyl menší než výběrový průměr. Také bayesovské odhady dávají dobré výsledky zejména pro menší hodnoty c (tedy velké κ). Obzvláště výhodné pro bayesovské odhady je to, že fungují velmi dobře i pro výběry s výběrovým rozptylem menším než výběrový průměr.

Závěrem lze pro automatické stanovení odhadů parametrů NB rozdělení doporučit následující postup:

Nejdříve spočíst výběrový průměr $\overline{y}$ a výběrový rozptyl s^2.



Obrázek 3.3: Srovnání odhadů parametru c (pro rozsah výběru $n = 30$) získaných metodou momentů, metodou maximální věrohodnosti a její korekce, quasi-likelihood metodou (EQL, DEQL) a bayesovským přístupem. Pro názornost jsou průměrné odhady spojeny čarou, skutečná hodnota parametru c je vynesena černou přerušovanou čarou.



Obrázek 3.4: Srovnání odhadů parametru c (pro rozsah výběru $n = 100$) získaných metodou momentů, metodou maximální věrohodnosti a její korekce, quasi-likelihood metodou (EQL, DEQL) a bayesovským přístupem. Pro názornost jsou průměrné odhady spojeny čarou, skutečná hodnota parametru c je vynesena černou přerušovanou čarou.

Je-li $\overline{y} < s^2$ dává nejkvalitnější odhady metoda maximální věrohodnosti a korekce ML odhadů na nestrannost. Numericky jednodušší, ale přesto kvalitní odhady dávají i quasi-likelihood metody.

Pro případ $\overline{y} > s^2$ se jako nejkvalitnější ukazují Bayesovské metody, které vždy dávají kladný odhad κ. Alternativně lze použít i quasi-likelihood metody a v případě, že daný odhad vyjde záporný pracovat dále s binomickým rozdělením podle vztahu (2.4) uvedeném v odstavci 2.1.

V případě, že $\overline{y}$ je blízký, nebo roven s^2, je vhodné přejít k aproximaci Poissonovým rozdělením.

Kapitola 4

Srovnání populací s NB rozdělením

Budeme se věnovat situaci, kdy je dáno p nezávislých náhodných výběrů z NB rozdělení, tedy předpokládáme, že $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ je náhodný výběr z $NB(\mu_i, \kappa_i)$ rozsahu n_i pro $i = 1, \ldots, p$. Celkem je tedy náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$ n-rozměrný, $n = \sum_{i=1}^{p} n_i$ a jeho rozdělení záleží na vektorovém parametru $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\kappa}')' = (\mu_1, \ldots, \mu_p, \kappa_1, \ldots, \kappa_p)'$. Cílem dalších úvah budou testy hypotéz o parametrech μ a κ. Speciálně při testování hypotéz o parametru μ budeme považovat parametr κ za rušivý a naopak.

Testy hypotéz v modelech s rušivými parametry uvedené v odstavci 1.4 nyní použijeme k testování hypotéz o rovnosti parametrů, tedy k testu obdobných hypotéz, které se vyskytují v analýze rozptylu. V těchto případech je vhodné rozdělení vhodně reparametrizovat.

Hypotézu o rovnosti parametrů μ lze zapsat ve tvaru $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$. Tento zápis však neumožňuje bezprostřední použití statistik z odstavce 1.4. Proto provedeme reparametrizaci. Označme $\mu_i = \mu + \alpha_i$, s podmínkou $\alpha_1 = 0$. Tato reparametrizace je dobře známá z analýzy rozptylu. Hypotézu H_0 lze nyní pomocí parametru $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_p)'$ přepsat ve tvaru $H_0 : \boldsymbol{\alpha} = \mathbf{0}$. Tento zápis nulové hypotézy je pro test rovnosti parametrů μ vhodnější. Podobně pro test rovnosti parametrů κ se nejprve vyjádří $\kappa_i = \kappa + \beta_i$ za podmínky $\beta_1 = 0$ a H_0 se vektorově zapíše ve tvaru $H_0 : \boldsymbol{\beta} = \mathbf{0}$, kde $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_p)'$.

Specifikujme dále jednotlivé nulové hypotézy, jejichž testováním se budeme zabývat:

1. Uvažme hypotézu $H_{01} : \boldsymbol{\alpha} = \mathbf{0}$ oproti alternativě, že H_{01} neplatí. S ohledem na rušivé parametry lze rozlišit 2 experimentální situace:

   (a) $S_{01.1}$ - Parametry κ jsou obecně různé tj. existuje $i, j = 1, \ldots, p$ tak, že $\kappa_i \neq \kappa_j$, pak pracujeme s vektorem parametrů
   $\boldsymbol{\theta} = (\alpha_2, \ldots, \alpha_p, \mu, \kappa_1, \ldots, \kappa_p)'$.

   (b) $S_{01.2}$ - Parametry κ jsou totožné tj. pro všechna $i, j = 1, \ldots, p$ platí $\kappa_i = \kappa_j$, pak pracujeme s vektorem parametrů
   $\boldsymbol{\theta} = (\alpha_2, \ldots, \alpha_p, \mu, \kappa)'$, kde κ je společná hodnota parametrů $\kappa_1, \ldots, \kappa_p$.

2. Uvažme hypotézu $H_{02} : \boldsymbol{\beta} = \mathbf{0}$ oproti alternativě, že H_{02} neplatí. S ohledem na rušivé parametry lze opět rozlišit 2 experimentální situace:

   (a) $S_{02.1}$ - Parametry μ jsou obecně různé tj. existuje $i, j = 1, \ldots, p$ tak, že $\mu_i \neq \mu_j$, pak pracujeme s vektorem parametrů
   $\boldsymbol{\theta} = (\beta_2, \ldots, \beta_p, \mu_1, \ldots, \mu_p, \kappa)'$.

   (b) $S_{02.2}$ - Parametry μ_i jsou totožné a rovné společné hodnotě μ, pak pracujeme s vektorem parametrů
   $\boldsymbol{\theta} = (\beta_2, \ldots, \beta_p, \mu, \kappa)'$.

3. Uvažme hypotézu $H_{03} : \boldsymbol{\alpha} = \mathbf{0}$ a zároveň $\boldsymbol{\beta} = \mathbf{0}$ oproti alternativě, že H_{03} neplatí, pak pracujeme s vektorem parametrů
   $\boldsymbol{\theta} = (\alpha_2, \ldots, \alpha_p, \beta_2, \ldots, \beta_p, \mu, \kappa)'$ (S_{03}).

V tabulce 4.1 jsou pro názornost symbolicky uvedeny tvary nulové hypotézy a alternativy a označení jednotlivých testů.

Experimentální situace	H_0	H_1
$S_{01.1}$	μ, κ_i	μ_i, κ_i
$S_{01.2}$	μ, κ	μ_i, κ
$S_{02.1}$	μ_i, κ	μ_i, κ_i
$S_{02.2}$	μ, κ	μ, κ_i
S_{03}	μ, κ	μ_i, κ_i

Tabulka 4.1: Přehled uvažovaných nulových hypotéz a odpovídajících alternativ.

V dalším budou odvozeny konkrétní tvary Fisherovy informační matice potřebné pro konstrukci výše zmíněných testů. Zaveďme nejprve označení a_i, b_i a c_i pro střední hodnoty druhých derivací logaritmické věrohodnostní funkce podle parametrů μ_i

a κ_i. Jejich výpočet je uveden v A.2. Programová implementace získaných algoritmů pro výpočet potřebných statistik je v programu statistiky_E (viz kapitola 8). Tedy

$$\begin{aligned}
a_i = \quad E\left(\frac{\partial^2 l}{\partial \mu_i^2}\right) \quad &= E\left\{\sum_{j=1}^{n_i}\left[\frac{\kappa_i}{(\kappa_i+\mu_i)^2} - y_{ij}\frac{\kappa_i(\kappa_i+2\mu_i)}{\mu_i^2(\kappa_i+\mu_i)^2}\right]\right\} \\
&= \sum_{j=1}^{n_i}\frac{-1}{\mu_i(\kappa_i+\mu_i)} \quad \text{pro } i=1,\ldots,p \\
c_i = E\left(\frac{\partial^2 l}{\partial \mu_i \partial \kappa_i}\right) &= E\left\{\sum_{j=1}^{n_i}\left[\frac{-\mu_i}{(\kappa_i+\mu_i)^2} + y_{ij}\frac{1}{(\kappa_i+\mu_i)^2}\right]\right\} = 0 \\
b_i = \quad E\left(\frac{\partial^2 l}{\partial \kappa_i^2}\right) \quad &= E\left\{\sum_{j=1}^{n_i}\left[\Psi'(y_{ij}+\kappa_i) - \Psi'(\kappa_i) + \frac{\mu_i}{\kappa_i(\kappa_i+\mu_i)} - \frac{\mu_i}{(\kappa_i+\mu_i)^2} + \right.\right. \\
&\quad \left.\left. + y_{ij}\frac{1}{(\kappa_i+\mu_i)^2}\right]\right\} = \sum_{j=1}^{n_i}\left[\Delta_1^{ij} - \Psi'(\kappa_i) + \frac{\mu_i}{\kappa_i(\kappa_i+\mu_i)}\right]
\end{aligned}$$

pro $i=1,\ldots,p$, kde $\Delta_1^{ij} = E[\Psi'(y_{ij}+\kappa_i)]$ a připomeňme, že Ψ' je trigamma funkce.

Podle typu hypotézy je třeba hledat druhé derivace logaritmické věrohodnostní funkce i podle parametrů α_i popř. β_i, které lze nyní jednoduše získat z výše uvedených vztahů užitím vlastností derivací složených funkcí. Protože pro reparamerizaci $\mu_i = \mu + \alpha_i$ platí $\frac{\partial \mu_i}{\partial \alpha_i} = 1$ (pod. pro κ), platí $\frac{\partial l_i}{\partial \alpha_i} = \frac{\partial l_i}{\partial \mu_i}$ apod. Dle typu nulové hypotézy platí $\mu_i = \mu + \alpha_i$ nebo $\kappa_i = \kappa + \beta_i$.

4.1 Test rovnosti středních hodnot při různých κ

Uvažujme náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$ rozsahu $n = \sum_{i=1}^p n_i$ z negativně binomického rozdělení s vektorem parametrů $\boldsymbol{\theta} = (\alpha_2 \ldots, \alpha_p, \mu, \kappa_1, \ldots, \kappa_p)'$, kde konkrétně

$$\begin{aligned}
\mathbf{Y}_1 &\sim NB(\mu, \kappa_1) \\
\mathbf{Y}_2 &\sim NB(\mu+\alpha_2, \kappa_2) \\
&\vdots \\
\mathbf{Y}_p &\sim NB(\mu+\alpha_p, \kappa_p).
\end{aligned}$$

Test rovnosti středních hodnot proto lze formulovat jako test nulové hypotézy H_{01} : $\boldsymbol{\alpha} = \mathbf{0}$ s vektorem rušivých parametrů $\boldsymbol{\psi} = (\mu, \kappa_1, \ldots, \kappa_p)'$.

Za platnosti nulové hypotézy by tedy složky vektoru $\mathbf{Y}$ měly NB rozdělení se společnou střední hodnotou μ a obecně odlišnými hodnotami $\kappa_1, \ldots, \kappa_n$. V alternativě se předpokládají jak různé střední hodnoty tak i různá κ.

Pro testovací statistiky je třeba znát odhad Fisherovy informační matice, pro kterou je třeba vyjádřit střední hodnoty druhých derivací logaritmické věrohodnostní funkce. Ke stanovení testovací statistiky při testu hypotézy $H_{01.1}$ v situaci $S_{01.1}$ je potřebné stanovit následující hodnoty

$$
\begin{aligned}
E\left(\frac{\partial^2 l}{\partial \alpha_i^2}\right) &= a_i & E\left(\frac{\partial^2 l}{\partial \alpha_i \partial \mu}\right) &= a_i & &\text{pro } i = 2, \ldots, p \\
E\left(\frac{\partial^2 l}{\partial \alpha_i \partial \kappa_1}\right) &= 0 & E\left(\frac{\partial^2 l}{\partial \alpha_i \partial \kappa_i}\right) &= c_i = 0 & &\text{pro } i = 2, \ldots, p \\
E\left(\frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j}\right) &= 0 & E\left(\frac{\partial^2 l}{\partial \kappa_i \partial \kappa_j}\right) &= 0 & &\text{pro } i \neq j \\
E\left(\frac{\partial^2 l}{\partial \mu \partial \kappa_i}\right) &= c_i = 0 & E\left(\frac{\partial^2 l}{\partial \kappa_i^2}\right) &= b_i & &\text{pro } i = 1, \ldots, p \\
& & E\left(\frac{\partial^2 l}{\partial \mu^2}\right) &= \sum_{i=1}^{p} a_i. & &
\end{aligned}
$$

Příslušné veličiny a_i, b_i, c_i byly zavedeny v předchozím odstavci. Pomocí nich lze schematicky zapsat Fisherovu informační matici. Toto vyjádření je z typografických důvodů na konci této kapitoly. Po dosazení získané Fisherovy informační matice do statistik z odstavce 1.4 dostaneme příslušné testovací statistiky.

4.2 Test rovnosti středních hodnot při stejných κ

Jedná se o náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$ rozsahu $n = \sum_{i=1}^{p} n_i$ z negativně binomického rozdělení s vektorem parametrů $\boldsymbol{\theta} = (\alpha_2 \ldots, \alpha_p, \mu, \kappa)'$, kde konkrétně

$$
\begin{aligned}
\mathbf{Y}_1 &\sim NB(\mu, \kappa) \\
\mathbf{Y}_2 &\sim NB(\mu + \alpha_2, \kappa) \\
&\vdots \\
\mathbf{Y}_p &\sim NB(\mu + \alpha_p, \kappa).
\end{aligned}
$$

Nulová hypotéza je tedy $H_{01} : \boldsymbol{\alpha} = \mathbf{0}$ a vektor rušivých parametrů $\boldsymbol{\psi} = (\mu, \kappa)'$.

Za platnosti nulové hypotézy by složky vektoru $\mathbf{Y}$ měly NB rozdělení se stejnou střední hodnotou a stejným parametrem κ. V alternativě se předpokládají různé střední hodnoty a stejná κ.

Schématický zápis Fisherovy informační matice je uveden na konci této kapitoly. Po dosazení získané Fisherovy informační matice do statistik z odstavce 1.4 dostaneme příslušné testovací statistiky.

4.3 Test rovnosti κ při různých středních hodnotách

Jedná se o náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$ rozsahu $n = \sum_{i=1}^{p} n_i$ z negativně binomického rozdělení s vektorem parametrů $\boldsymbol{\theta} = (\beta_2, \ldots, \beta_p, \mu_1, \ldots, \mu_p, \kappa)'$, kde konkrétně

$$
\begin{aligned}
\mathbf{Y}_1 &\sim NB(\mu_1, \kappa) \\
\mathbf{Y}_2 &\sim NB(\mu_2, \kappa + \beta_2) \\
&\vdots \\
\mathbf{Y}_p &\sim NB(\mu_p, \kappa + \beta_p).
\end{aligned}
$$

Nulová hypotéza je tedy $H_{01} : \boldsymbol{\beta} = \mathbf{0}$ a vektor rušivých parametrů
$\boldsymbol{\psi} = (\mu_1 \ldots, \mu_p, \kappa)'$.

Za platnosti nulové hypotézy by složky vektoru $\mathbf{Y}$ měly NB rozdělení se stejným parametrem κ a obecně různou střední hodnotou. V alternativě se předpokládají jak různá κ tak i různé střední hodnoty.

Schématický zápis Fisherovy informační matice je uveden na konci této kapi-

toly. Po dosazení získané Fisherovy informační matice do statistik z odstavce 1.4 dostaneme příslušné testovací statistiky.

4.4 Test rovnosti κ při stejných středních hodnotách

Jedná se o náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$ rozsahu $n = \sum_{i=1}^{p} n_i$ z negativně binomického rozdělení s vektorem parametrů $\boldsymbol{\theta} = (\beta_2, \ldots, \beta_p, \mu, \ldots, \mu, \kappa)'$, kde konkrétně

$$\begin{aligned}
\mathbf{Y}_1 &\sim NB(\mu, \kappa) \\
\mathbf{Y}_2 &\sim NB(\mu, \kappa + \beta_2) \\
&\vdots \\
\mathbf{Y}_p &\sim NB(\mu, \kappa + \beta_p).
\end{aligned}$$

Nulová hypotéza je tedy $H_{01} : \boldsymbol{\beta} = \mathbf{0}$ a vektor rušivých parametrů $\boldsymbol{\psi} = (\mu, \ldots, \mu, \kappa)'$.

Za platnosti nulové hypotézy by složky vektoru $\mathbf{Y}$ měly NB rozdělení se stejným parametrem κ i střední hodnotou. V alternativě se předpokládají různá κ a stejné střední hodnoty.

Schématický zápis Fisherovy informační matice je uveden na konci této kapitoly. Po dosazení získané Fisherovy informační matice do statistik z odstavce 1.4 dostaneme příslušné testovací statistiky.

4.5 Test rovnosti středních hodnot a zároveň rovnosti κ

Jedná se o náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$ rozsahu $n = \sum_{i=1}^{p} n_i$ z negativně binomického rozdělení s vektorem parametrů $\boldsymbol{\theta} = (\alpha_2 \ldots, \alpha_p, \mu, \kappa_1, \ldots, \kappa_p)'$, kde

konkrétně

$$\begin{aligned}\mathbf{Y}_1 \sim & NB(\mu, \kappa)\\ \mathbf{Y}_2 \sim & NB(\mu + \alpha_2, \kappa + \beta_2)\\ & \vdots\\ \mathbf{Y}_p \sim & NB(\mu + \alpha_p, \kappa + \beta_p).\end{aligned}$$

Nulová hypotéza je tedy $H_{01} : \boldsymbol{\alpha} = \mathbf{0}$ a vektor rušivých parametrů $\boldsymbol{\psi} = (\mu, \kappa_1, \ldots, \kappa_p)'$.

Za platnosti nulové hypotézy by složky vektoru $\mathbf{Y}$ měly NB rozdělení se stejným parametrem κ i stejnou střední hodnotou. V alternativě se předpokládají jak různá κ tak i různé střední hodnoty.

Schématický zápis Fisherovy informační matice je uveden na konci této kapitoly. Po dosazení získané Fisherovy informační matice do statistik z odstavce 1.4 dostaneme příslušné testovací statistiky.

4.6 Fisherova informační matice pro jednotlivé testy

Fisherova informační matice pro test rovnosti středních hodnot při různých κ (viz odstavec 4.1)

$$\mathbf{J} = \frac{1}{n}\left(\begin{array}{cccc|ccccc} a_2 & 0 & \ldots & 0 & a_2 & 0 & \ldots & \ldots & 0\\ 0 & a_3 & & \vdots & a_3 & \vdots & \ddots & & \vdots\\ \vdots & & \ddots & 0 & \vdots & \vdots & & \ddots & \vdots\\ 0 & \ldots & 0 & a_p & a_p & 0 & \ldots & \ldots & 0\\ \hline a_2 & a_3 & \ldots & a_p & \sum_{i=1}^{p} a_i & 0 & \ldots & \ldots & 0\\ 0 & \ldots & \ldots & 0 & 0 & b_1 & 0 & \ldots & 0\\ \vdots & \ddots & & \vdots & \vdots & 0 & \ddots & & \vdots\\ \vdots & & \ddots & \vdots & \vdots & \vdots & & \ddots & 0\\ 0 & \ldots & \ldots & 0 & 0 & 0 & & 0 & b_p \end{array}\right)$$

Fisherova informační matice pro test rovnosti středních hodnot při stejných κ

(viz odstavec 4.2)

$$\mathbf{J} = \frac{1}{n}\left(\begin{array}{cccc|cc} a_2 & 0 & \dots & 0 & a_2 & 0 \\ 0 & a_3 & & \vdots & a_3 & 0 \\ \vdots & & \ddots & 0 & \vdots & \vdots \\ 0 & \dots & 0 & a_p & a_p & 0 \\ \hline a_2 & a_3 & \dots & a_p & \sum_{i=1}^{p} a_i & 0 \\ 0 & \dots & \dots & 0 & 0 & \sum_{i=1}^{p} b_i \end{array}\right)$$

Fisherova informační matice pro test rovnosti κ při různých středních hodnotách (viz odstavec 4.3)

$$\mathbf{J} = \frac{1}{n}\left(\begin{array}{cccc|ccccc} b_2 & 0 & \dots & 0 & b_2 & 0 & \dots & \dots & 0 \\ 0 & b_3 & & \vdots & b_3 & \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \vdots & \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & b_p & b_p & 0 & \dots & \dots & 0 \\ \hline b_2 & b_3 & \dots & b_p & \sum_{i=1}^{p} b_i & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 & a_1 & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & \vdots & & \ddots & 0 \\ 0 & \dots & \dots & 0 & 0 & 0 & & 0 & a_p \end{array}\right)$$

Fisherova informační matice pro test rovnosti κ při stejných středních hodnotách (viz odstavec 4.4)

$$\mathbf{J} = \frac{1}{n}\left(\begin{array}{cccc|cc} a_2 & 0 & \dots & 0 & a_2 & 0 \\ 0 & a_3 & & \vdots & a_3 & 0 \\ \vdots & & \ddots & 0 & \vdots & \vdots \\ 0 & \dots & 0 & a_p & a_p & 0 \\ \hline a_2 & a_3 & \dots & a_p & \sum_{i=1}^{p} a_i & 0 \\ 0 & \dots & \dots & 0 & 0 & \sum_{i=1}^{p} b_i \end{array}\right)$$

Fisherova informační matice pro test rovnosti středních hodnot při různých κ

(viz odstavec 4.5)

$$\mathbf{J} = \frac{1}{n}\left(\begin{array}{cccccccc|cc} a_2 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & a_2 & 0 \\ 0 & a_3 & & \vdots & \vdots & \ddots & & \vdots & a_3 & \vdots \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & a_p & 0 & \dots & \dots & 0 & a_p & 0 \\ 0 & \dots & \dots & 0 & b_2 & 0 & \dots & 0 & 0 & b_2 \\ \vdots & \ddots & & \vdots & 0 & b_3 & & \vdots & \vdots & b_3 \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & 0 & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 0 & \dots & 0 & b_p & 0 & b_p \\ \hline a_2 & a_3 & \dots & a_p & 0 & \dots & \dots & 0 & \sum_{i=1}^{p} a_i & 0 \\ 0 & \dots & \dots & 0 & b_2 & b_3 & \dots & b_p & 0 & \sum_{i=1}^{p} b_i \end{array}\right)$$

Připomeňme, že a_i a b_i byly definovány v úvodu této kapitoly a že závisí na parametrech μ_i a κ_i. Zároveň je v uvedených maticích naznačeno jejich rozdělení na bloky potřebné k výpočtu matice $\mathbf{J}_{11.2}$ potřebné ve statistikách S (viz (1.18)) a W (viz (1.19)) uvedených v odstavci 1.4. Tyto statistiky společně se statistikou LR (viz (1.20)) lze využít k testu uvedených hypotéz, přičemž parametry μ_i a κ_i nahradíme jejich ML odhady.

Testy o střední hodnotě byly dále využity v simulační studii v kapitole 6.

Ukázka využití těchto testů na reálných datech je v kapitole 7, která je věnována aplikacím.

Pro výpočet uvedených testovacích statistik byl vytvořen program statistiky_E, který je uveden v kapitole 8.

Kapitola 5

Statistická inference o středních hodnotách NB rozdělení při známém parametru κ

V této kapitole budeme předpokládat, že parametr κ je známý. Tento předpoklad dále umožňuje využít při statistické inferenci teorie zobecněných lineárních modelů. Speciálně bude možné střední hodnotu μ popsat v závislosti na doprovodných proměnných a provádět analýzy populací s NB rozdělením, které jsou obdobné analýze rozptylu v regresní analýze.

Hustotu NB rozdělení můžeme zapsat ve tvaru

$$f(y;\mu,\kappa) = \exp\left\{y \ln \frac{\mu}{\kappa+\mu} + \kappa \ln \frac{\kappa}{\kappa+\mu} + \ln \binom{y+\kappa-1}{y}\right\} \tag{5.1}$$

a za předpokladu, že parametr κ je pevný a známý se jedná o hustotu exponenciálního typu.

Zavedením nového parametru $\theta = \ln \frac{\mu}{\kappa+\mu}$ lze hustotu (5.1) převést na kanonický tvar. Dostaneme

$$f(y;\mu,\kappa) = \exp\left\{y\theta - \kappa \ln \frac{1}{1-e^{\theta}} + \ln \binom{y+\kappa-1}{y}\right\}. \tag{5.2}$$

Pro statistickou analýzu náhodného výběru s hustotou (5.2) lze tedy použít teorie zobecněných lineárních modelů.

Protože hustota (5.2) je v kanonickém tvaru, je v dalším zvolen kanonický link

$\theta = \eta$.

Porovnáním s hustotou (1.12) dostaneme

$$\gamma(\theta) = \kappa \ln \frac{1}{1 - e^{\theta}},$$

a dále pro NB rozdělení platí:

- přirozený parametr: $\theta = \ln \frac{\mu}{\kappa+\mu}$,
- rušivý parametr: $\phi = 1$,
- střední hodnota: $\mu = \mu(\theta) = EY(y;\theta) = \gamma'(\theta) = \kappa \frac{e^{\theta}}{1-e^{\theta}}$,
- rozptyl: $DY(y;\theta) = -\gamma''(\theta)\psi(\phi) = \kappa \frac{1}{(1-e^{\theta})^2}$,
- kanonická linkovací funkce: $g(\mu) = \ln \frac{\mu}{\kappa+\mu}$.

Pro náhodný vektor $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_p')'$, kde $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ a $Y_{ij} \sim NB(\mu_i, \kappa)$ lze sdruženou hustotu vektoru $\mathbf{Y}$ zapsat ve tvaru

$$f(\mathbf{y}; \boldsymbol{\mu}, \kappa) = \exp\left\{ \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left[y_{ij} \ln \frac{\mu_i}{\kappa + \mu_i} + \kappa \ln \frac{\kappa}{\kappa + \mu_i} + \ln \binom{y_{ij} + \kappa - 1}{y_{ij}} \right] \right\}.$$

Zavedeme-li nový parametr $\theta_i = \ln \frac{\mu_i}{\kappa+\mu_i}$ můžeme sdruženou hustotu vektoru $\mathbf{Y}$ zapsat ve tvaru

$$f(\mathbf{y}; \boldsymbol{\theta}, \kappa) = \exp\left\{ \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left[y_{ij}\theta_i - \kappa \ln \frac{1}{1 - e^{\theta_i}} + \ln \binom{y_{ij} + \kappa - 1}{y_{ij}} \right] \right\},$$

což odpovídá hustotě exponenciálního typu v kanonické formě bez rušivých parametrů.

Pro test hypotézy $H_0 : \mu_1 = \cdots = \mu_p = \mu$ je použit kanonický link

$$\theta_i \;=\; \ln \frac{\mu_i}{\kappa + \mu_i} \;=\; \delta + \beta_i$$

s podmínkou $\beta_1 = 0$, kde δ je nový parametr.

Parametry, na které se zaměřujeme, jsou tedy $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_p)'$, pro test rovnosti středních hodnot nás parametr δ nezajímá.

Odhady hledaných parametrů nalezneme řešením věrohodnostních rovnic (1.9).

Pro $\mu_i = \kappa \frac{e^{\eta_i}}{1-e^{\eta_i}}$ lze věrohodnostní rovnice přepsat do tvaru

$$\sum_{i=1}^{n} x_{ij}(Y_i - \mu_i) = \sum_{i=1}^{n} x_{ij}(Y_i - \kappa \frac{e^{\eta_i}}{1 - e^{\eta_i}}) = 0.$$

K testování hypotéz jsou k dispozici známé statistiky - věrohodnostní poměr (deviance) LR, skórová S a Waldova W statistika, které jsou analogické statistikám zavedeným v odstavci 1.4.

Pro výpočet zmíněných statistik pro test hypotézy byla použita funkce GLM_NB [29], kterou vytvořila Zuzna Hübnerová.

Kapitola 6

Síly testů o středních hodnotách NB rozdělení

V této kapitole budeme vycházet z předpokladů popsaných v předchozích kapitolách. Tato kapitola je zaměřena na srovnání sil testů tří následujících hypotéz.

H_{01}**:** $\mu_1 = \cdots = \mu_p$ za předpokladu, že $\kappa_1, \ldots, \kappa_p$ jsou neznámé a obecně různé

H_{02}**:** $\mu_1 = \cdots = \mu_p$ za předpokladu, že $\kappa_1 = \cdots = \kappa_p$ $(= \kappa)$ a κ je neznámé

H_{03}**:** $\mu_1 = \cdots = \mu_p$ za předpokladu, že $\kappa_1, \ldots, \kappa_p$ $(= \kappa)$ a κ je známé

Testy uvedených hypotéz byly provedeny pomocí statistik - věrohodnostní poměr (deviance) LR (viz (1.20)), Waldova W (viz (1.19)) a skórová S statistika (viz (1.18)). Síly těchto testů byly nalezeny pomocí simulací. Pro test hypotézy H_{03} byla provedena i aproximace teoretické síly (viz [14], [22], [23], [28]).

Pro test rovnosti středních hodnot (hypotézy H_{01} a H_{02}) byla použita reparametrizace

$$\mu_i = \mu + \alpha_i, \ \alpha_1 = 0. \tag{6.1}$$

Nulová hypotéza byla pro tyto dva testy zapsána ve tvaru $\boldsymbol{\alpha} = \mathbf{0}$ a rušivý parametr pro test hypotézy H_{01} je $\boldsymbol{\psi} = (\mu, \kappa_1, \ldots, \kappa_p)'$ a pro test hypotézy H_{02} je $\boldsymbol{\psi} = (\mu, \kappa)'$.

Jak bylo uvedeno v předchozí kapitole, v situaci, kdy je κ známé, lze k testování hypotézy H_{03} využít metod GLM. Hustotu NB rozdělení lze zapsat v kanonickém

tvaru a je tedy vhodné zvolit kanonický link

$$\ln \frac{\mu_i}{\kappa + \mu_i} = \delta + \beta_i. \tag{6.2}$$

Protože matice plánu není plné hodnosti, byla zvolena doplňující podmínka $\beta_1 = 0$. Rušivým parametrem je tedy v této situaci pouze parametr δ.

6.1 Aproximace síly testu v případě známého κ

V případě konstantního a známého parametru κ jsou možné dva přístupy k aproximaci síly testu.

První, Pitmanův přístup, kde se hypotéza H_0 testuje proti posloupnosti alternativ

$$\mathrm{A}_n : \boldsymbol{\beta} = \boldsymbol{\beta}_H + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n = o(n^{-1/2}), \quad n = \sum_{i=1}^{p} n_i,$$

který poskytuje aproximaci distribučních funkcí testových statistik LR, W a S řádu $o_p(1)$ a $o_p(n^{-1/2})$ (viz [14]).

Druhý, vyvážená ANOVA (pro $n_1 = n_2 = \cdots = n_p = N$) umožní aproximaci testových statistik LR a S řádu $o_p(1)$ (viz [28]).

V [22] je uvedeno, že pro posloupnost alternativ A_n lze distribuční funkci testovacích statistik LR, S a W aproximovat necentrálním χ^2 rozdělením s $p-1$ stupni volnosti a parametrem necentrality λ, který závisí na $\boldsymbol{\varepsilon}_n$ a δ.

Aby bylo možno rozlišit mezi aproximacemi distribučních funkcí statistik LR, S a W byly odvozeny aproximace (viz [14])

$$\begin{aligned}
F_{\mathrm{LR}}(t) &= G_{p-1,\lambda}(t) + \sum_{j=0}^{2} b_j^{\mathrm{LR}} G_{p-1+2j,\lambda}(t) + o(n^{-1/2}), \\
F_W(t) &= G_{p-1,\lambda}(t) + \sum_{j=0}^{3} b_j^{W} G_{p-1+2j,\lambda}(t) + o(n^{-1/2}), \\
F_S(t) &= G_{p-1,\lambda}(t) + \sum_{j=0}^{3} b_j^{S} G_{p-1+2j,\lambda}(t) + o(n^{-1/2}),
\end{aligned}$$

kde $G_{p-1,\lambda}(t)$ je distribuční funkce necentrálního $\chi^2(p-1,\lambda)$ rozdělení. Koeficienty v uvedených aproximací lze nalézt v [14].

V práci [28] je odvozeno, že pro vyvážené třídění lze statistiku LR (popř. S) apro-

ximovat lineární kombinací nezávislých veličin majících necentrálních χ^2 rozdělení.

Těchto výsledků bude dále využito při porovnání sil jednotlivých testů.

6.2 Simulované síly

V simulační studii byly generovány výběry z $NB(\mu_i, \kappa)$, $i = 1, 2, 3$ a testovány hypotézy H_{01}, H_{02} a H_{03}. Pro nulovou hypotézu bylo zvoleno $\mu_i = \mu = 10$. V alternativě pro μ_i platí $\mu_i = \mu + (1 - i)h$ a h bylo postupně zvoleno (0,2; 1,0; ...; 4). Pro nulovou hypotézu a každou z alternativ byl výběr 1000× opakován. Z další studie byly vyřazeny výběry, u nichž vyšel výběrový rozptyl menší než výběrový průměr (vzhledem k problémům s nalezením ML odhadů κ). Rozsahy výběrů byly voleny postupně 30, 50 a 100. Toto bylo opakováno pro $\kappa = 1, \ldots, 11$.

Výsledky jsou graficky znázorněny na obrázcích 6.1, 6.3 a 6.4. V obrázcích 6.1 a 6.2 jsou hodnoty statistiky LR označeny $*$, hodnoty statistiky S jsou označeny kroužkem o a hodnoty statistiky W trojúhelníkem, modrou čerchovanou čarou je značen test hypotézy H_{01}, tedy test rovnosti μ za předpokladu, že κ se mohou lišit a jsou neznámé, červenou přerušovanou čarou je značena síla testu hypotézy H_{02}, tedy předpoklad, že κ_i jsou si rovny, ale společná hodnota κ je neznámá a černou plnou čarou je značena síla testu hypotézy H_{03}, když je κ známé.

Na obrázku 6.1 je názorně vidět jak rychle s rostoucím rozsahem výběru roste síla. Na obrázku 6.2 je ukázáno chování statistik v okolí nulové hypotézy. Jedná se o výřezy z grafů uvedených na předchozím obrázku 6.1 pro n 30, 50, 100 a κ 3, 11. Z těchto grafů je vidět, že pro malé rozsahy (viz obr. 6.2(a) a 6.2(b)) není dodržena hladina 0,5. Soustředíme-li se například na test hypotézy H_{01} (modrá čerchovaná čára), je z uvedených obrázků patrné, že Waldova statistika má tendenci hladinu překračovat, zatímco skórová statistika má tendenci hladiny nedosáhnout. S rostoucím rozsahem výběru dochází k ustálení kolem hladiny 0,5 (viz obr. 6.2(f)). Z obrázku 6.2(d) je patrné, že k tomuto ustálení dochází nejdříve u statistik pro test hypotézy H_{03}, tedy za předpokladu, že parametr κ je známý.

Na obrázku 6.3 je ukázáno porovnání aproximovaných a simulovaných sil testu pro hypotézu H_{03}. Na tomto obrázku je použito následující značení: aproximovaná síla pro vyvážené třídění je označena modře, aproximovaná síla pro Pitmanův přístup červeně se znakem $\Diamond$ a simulovaná síla černě se znakem $*$. Statistika LR je označena plnou čarou, statistika S čarou přerušovanou a statistika W čerchovanou čarou.

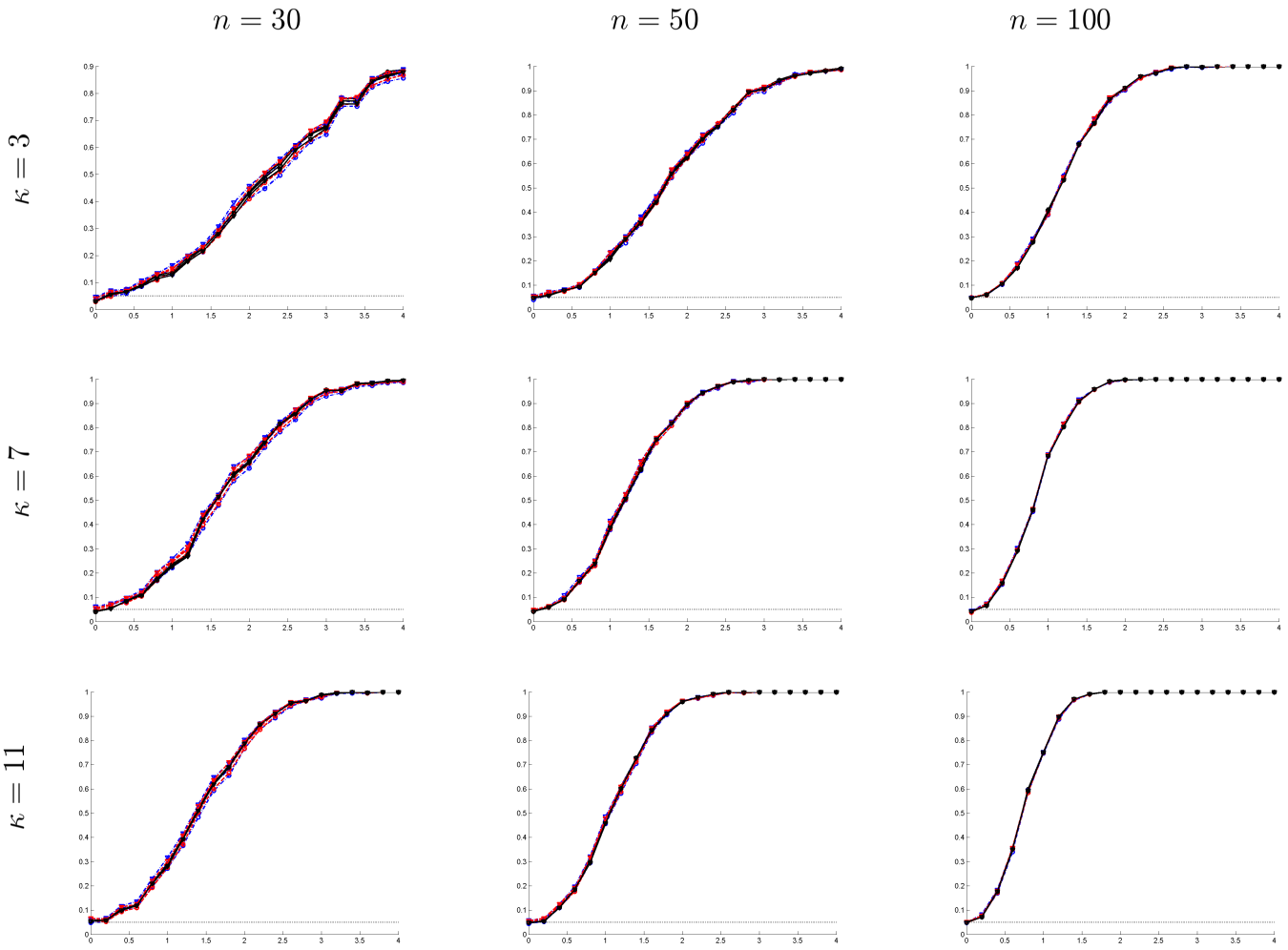
Tyto síly vychází srovnatelné, pouze aproximované síly pro Pitmanův přístup mají nepříjemnou vlastnost a to, že nabývají i hodnot větších než jedna.

Na obrázku 6.4 je pro jednotlivé statistiky ukázán vliv κ na sílu testu. Je zde použito označení: modrou čerchovanou čarou je značen test hypotézy H_{01}, tedy test rovnosti μ za předpokladu, že κ se mohou lišit a jsou neznámé, červenou přerušovanou čarou je značena síla testu hypotézy H_{02}, tedy předpoklad, že κ_i jsou si rovny, ale společná hodnota κ je neznámá a černou plnou čarou je značena síla testu hypotézy H_{03}, když je κ známé. Pro $\kappa = 1$ je použit znak $\triangle$, pro $\kappa = 3$ je použit znak $\circ$, pro $\kappa = 5$ je použit znak $\square$, pro $\kappa = 7$ je použit znak $\star$, pro $\kappa = 9$ je použit znak $\bigtriangledown$, pro $\kappa = 11$ je použit znak $\bullet$. Z těchto grafů je názorně vidět, jak s rostoucím κ roste síla testu.

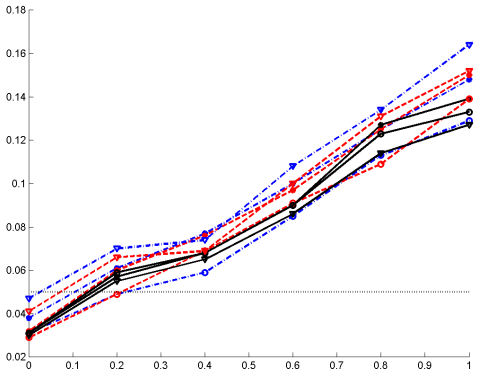
Na závěr lze tedy říci, že při velkých rozsazích vycházejí testy ekvivalentní.

Rozdíly mezi silami jednotlivých testů jsou minimální. Většinou vychází jako slabě silnější test rovnosti μ za předpokladu, že κ jsou známé, následován testem rovnosti μ za předpokladu, že κ jsou rovny neznámé společné hodnotě. Toto ovšem neplatí pokaždé. Výjimku tvoří malé rozsahy a malé κ.

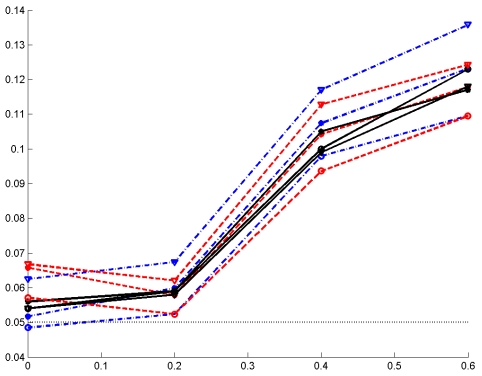
Trochu jiné chování má Waldova statistika, která se s ostatními srovná až pro větší výběry.



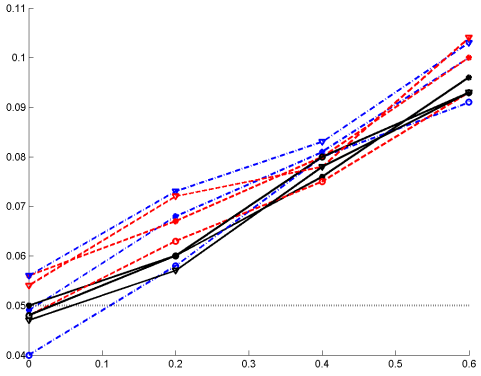
Obrázek 6.1: Srovnání simulovaných sil pro statistiky LR, S a W pro $\kappa = 3$, 7 a 11 pro výběry rozsahu $n = 30$, 50 a 100.



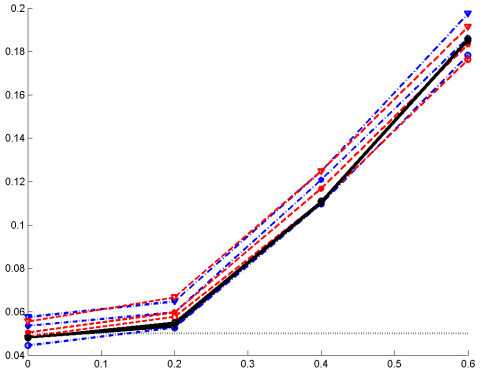
(a) $\kappa = 3$, $n = 30$



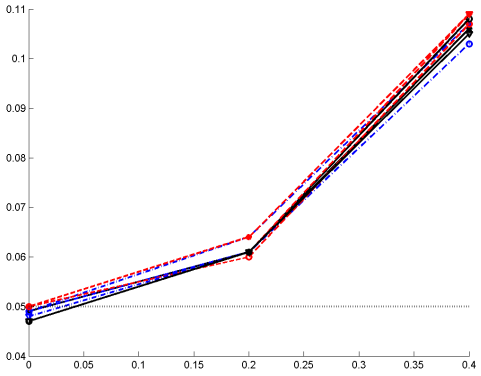
(b) $\kappa = 11$, $n = 30$



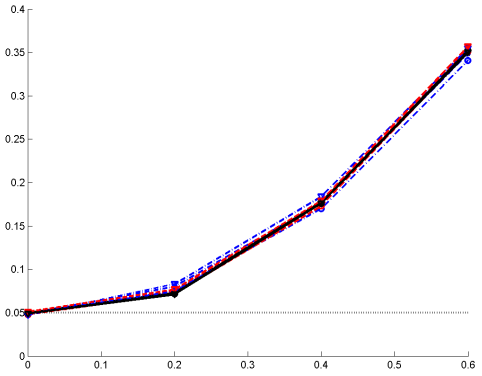
(c) $\kappa = 3$, $n = 50$



(d) $\kappa = 11$, $n = 50$

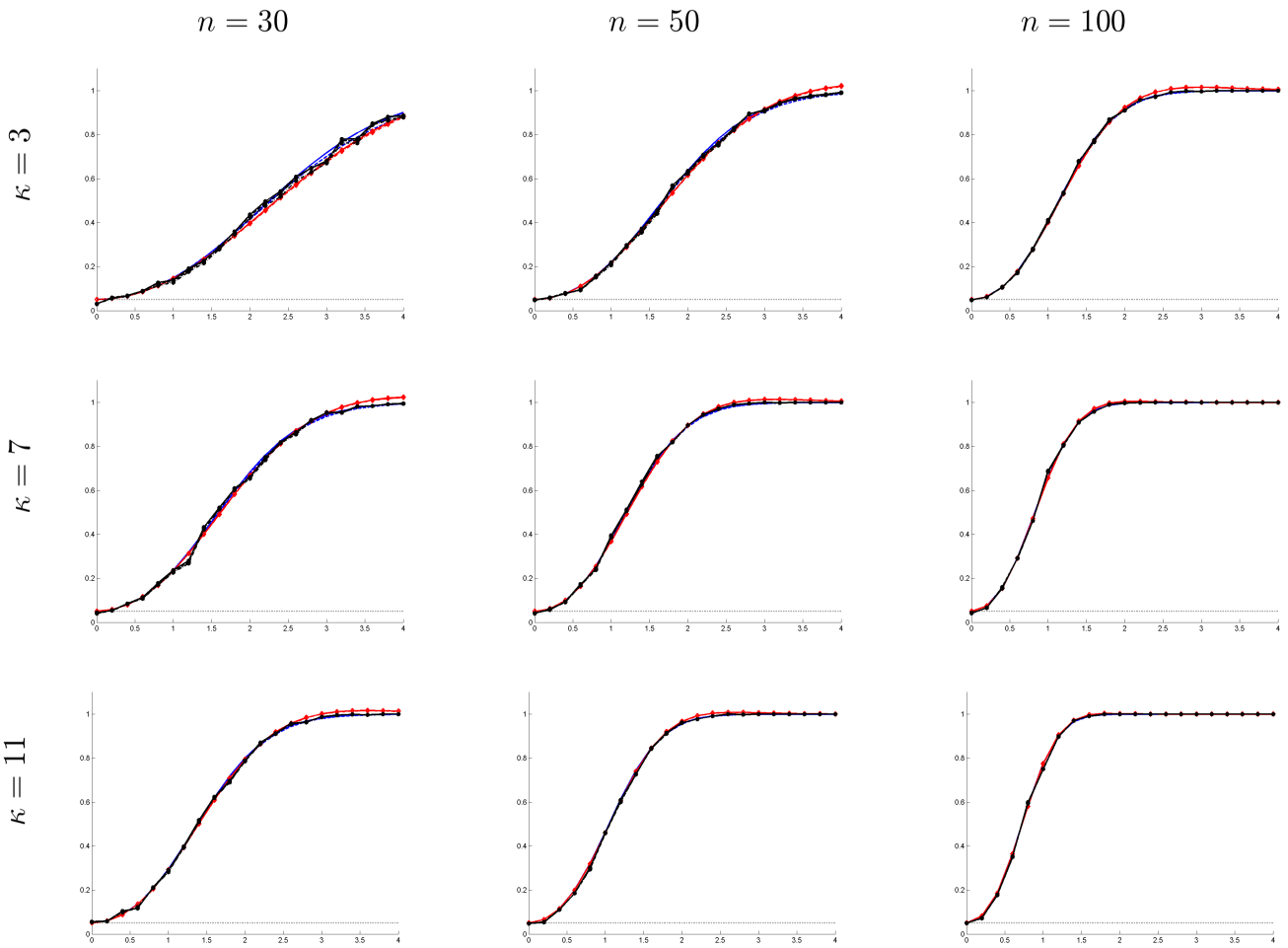


(e) $\kappa = 3$, $n = 100$

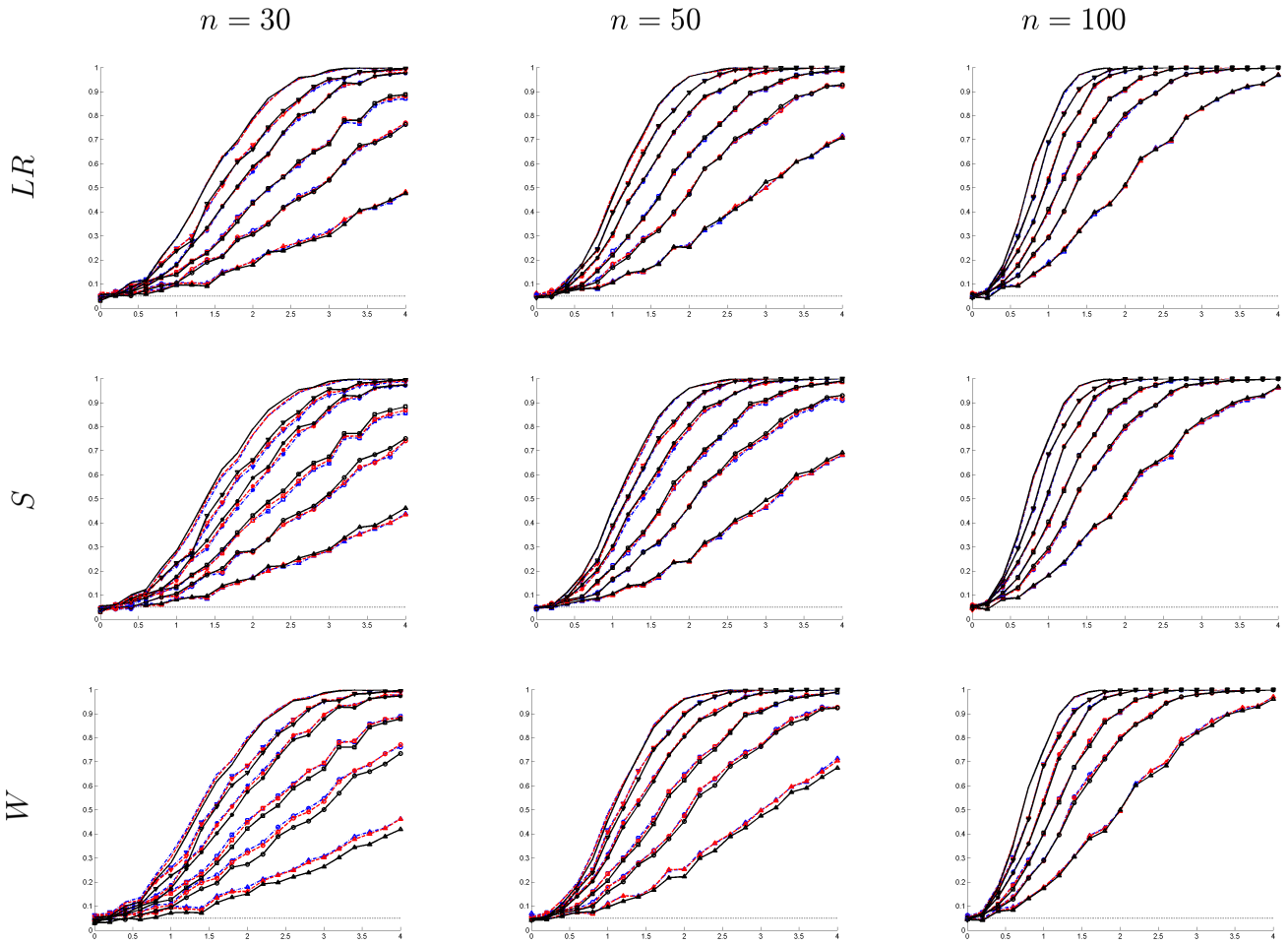


(f) $\kappa = 11$, $n = 100$

Obrázek 6.2: Zvětšená část z obrázku 6.1 pro κ rovno 3 a 11 a rozsah 30, 50 100.



Obrázek 6.3: Srovnání simulovaných sil pro H_0 a asimptotických sil pro vyvážené třídění pro $\kappa = 3$, 7 a 11 pro výběry rozsahu n =30, 50 a 100.



Obrázek 6.4: Srovnání simulovaných sil pro $\kappa = 1$, 3, 5, 7, 9 a 11, pro statistiky LR, S a W pro výběry rozsahu n =30, 50 a 100.

Kapitola 7

Aplikace

7.1 Analýza populací spárkaté zvěře v oblasti Jeseníků

V rámci ekologického průzkumu bylo sledováno, kde a v jakém porostu se zdržuje spárkatá zvěř v oblasti Jeseníků (600–1400 m n. m.). Proto byly v rámci výzkumu vytýčeny dílce o rozměrech 50×4 m v různých částech sledované oblasti a na těchto dílcích byla pracovníky ÚBO AV ČR sledována relativní densita trusu spolu s přítomností okusových a neokusových dřevin a pokryvností trav. Podrobný popis sledované oblasti je v publikacích [27] a [26].

Z literatury je známo (viz [44] nebo [66]), že distribuce zimního trusu na dané oblasti má negativně binomické rozdělení. To odpovídá přístupu k zavedení NB rozdělení pomocí gamma a Poissonova rozdělení v odstavci 2.7. V tomto pojetí je hustota náhodné veličiny Z úměrná času, po který se kopytníci zdrží na dané ploše a o počtu hromádek trusu na dané ploše pak lze předpokládat, že má Poissonovo rozdělení s parametrem rovným pozorované hodnotě Z pro danou plochu. Z těchto předpokladů pak dospějeme k NB rozdělení počtu hromádek trusu na daném dílci. Tato hypotéza byla také potvrzena pomocí testů dobré shody.

Ve sledovaném průzkumu byly studované plochy podle přítomnosti okusových a neokusových dřevin a podle pokryvnosti trav rozděleny do čtyř skupin. První skupina je charakterizována nízkou pokryvností trav i nízkým počtem okusových a neokusových dřevin, ve druhé skupině je vyšší pokryvnost trav, počet okusových a neokusových dřevin zůstává nižší, do třetí skupiny patří oblasti s vyšší pokryvností trav a vyšší přítomností neokusových dřevin, přítomnost okusových dřevin je nízká a

čtvrtá skupina je charakterizována vyšší přítomností okusových dřevin, přítomnost neokusových dřevin a pokryvnost trav je nižší. Budeme předpokládat, že hodnoty density trusu na jednotlivých dílcích dané oblasti tvoří náhodný výběr z NB rozdělení. Protože byly studovány čtyři skupiny pokusných ploch, předpokládáme, že jsou dány čtyři nezávislé náhodné výběry z $NB(\mu_i, \kappa_i)$. Rozsahy těchto výběrů (počty dílců v jednotlivých skupinách studovaných ploch) byly postupně 203, 84, 107 a 21.

Nejdříve byla testována hypotéza, že parametr κ je pro všechny čtyři skupiny dílců stejný, tedy hypotéza $\kappa_1 = \cdots = \kappa_4 = \kappa$ při libovolných středních hodnotách μ_i v jednotlivých skupinách dílců. Pro testování byla použita LR statistika (viz (1.20)) uvedená v závěru odstavce 1.4 (μ je vektor rušivých parametrů a κ je vektor testovaných parametrů). Odhady parametrů v základním modelu byly $\widehat{\kappa} = (1{,}28; 0{,}82; 1{,}20; 1{,}51)'$ a v redukovaném modelu za platnosti hypotézy $\kappa_1 = \cdots = \kappa_4 = \kappa$ byl nalezen odhad $\widetilde{\kappa} = 1{,}16$. Odhady parametrů μ_1, μ_2, μ_3, μ_4 byly $\widehat{\mu} = \widetilde{\mu} = (3{,}25; 2{,}94; 6{,}22; 3{,}67)'$. Hodnota testovací statistiky LR vyšla 2,997. Kvantil rozdělení$\chi^2_{1-\alpha}(3)$ má pro $\alpha = 0{,}05$ hodnotu 7,81. Tedy nulovou hypotézu na hladině 5 % nezamítáme a budeme dále předpokládat, že hodnoty parametru κ jsou pro všechny čtyři skupiny pokusných dílců stejné.

Dále budeme testovat hypotézu $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (za předpokladu, že parametr κ je stejný pro dané výběry). Metodou maximální věrohodnosti byly získány následující odhady $\widehat{\mu} = (3{,}25; 2{,}94; 6{,}22; 3{,}67)'$, $\widetilde{\mu} = 3{,}97$ a pro κ vyšly odhady $\widehat{\kappa} = 1{,}16$, $\widetilde{\kappa} = 1{,}03$. Po dosazení dostaneme hodnotu LR statistiky $LR = 34{,}419$. Protože platí $LR > \chi^2_{1-\alpha}(3) = 7{,}81$, zamítáme na hladině významnosti 5 % nulovou hypotézu, že střední hodnoty μ_i, $i = 1, \ldots, 4$ jsou stejné.

Na závěr ještě uveďme, že test hypotézy $\mu_1 = \mu_2 = \mu_3 = \mu_4$ za předpokladu, že rušivé parametry κ_i, $i = 1, \ldots, 4$ nejsou stejné by vedl k maximálně věrohodným odhadům $\widehat{\mu} = (3{,}25; 2{,}94; 6{,}22; 3{,}67)'$, $\widetilde{\mu} = 3{,}97$ a $\widehat{\kappa} = (1{,}28; 0{,}82; 1{,}20; 1{,}51)'$, $\widetilde{\kappa} = (1{,}21; 0{,}75; 0{,}96; 1{,}49)'$. Po dosazení je testovací statistika $LR = 33{,}512$ a protože platí $LR > \chi^2_{1-\alpha}(3) = 7{,}851$ opět na hladině 5 % zamítáme nulovou hypotézu rovnosti středních hodnot.

Závěrem lze konstatovat, že se dané čtyři skupiny oblastí neliší v parametru κ, ale liší se ve středních hodnotách μ_i, $i = 1, \ldots, 4$. Nejvyšší střední hodnota je ve třetí skupině, která odpovídá oblastem s vyšší pokryvností trávou a vyšším podílem neokusových dřevin. Naopak nejnižší je ve druhé skupině, která odpovídá oblastem s vyšší pokryvností trav a nižším počtem okusových a neokusových dřevin.

7.2 Statistická analýza počtu neutrofilů v závislosti na septickém stavu dětských pacientů

V období od září 2003 do prosince 2006 byla ve fakultní nemocnici v Brně provedena studie do níž bylo zahrnuto 1231 osob a to 579 dětských pacientů ve věku 0–19 let a 641 zdravých osob. U všech těchto osob bylo sledováno 12 genů, u nemocných byla navíc sledována imunitní odezva, která byla měřena stupněm sepse na škále 1 až 6, kde hodnota 1 odpovídá horečnatým stavům (tělesná teplota nad 39° C nebo nad 38,5° C naměřená následně ve dvou šestihodinových intervalech), hodnota 2 syndromu systémové zánětlivé odpovědi, hodnota 3 septickým stavům, hodnota 4 těžkým septickým stavům, hodnota 5 septickému šoku a hodnota 6 mnohočetnému selhání orgánů (viz [49]). Předmětem analýz publikovaných v [51], [64], [52] a [50] je statistická analýza závislostí mezi imunitní odezvou a variantou jednotlivých genů.

Mimo výše uvedené charakteristiky byly ve skupině nemocných měřeny i další medicínské charakteristiky. V této práci bude pozornost zaměřena na jednu z nich a to na absolutní počet neutrofilů (ANC z anglického absolute neutrophil count). Tato charakteristika je k dispozici pro 533 pacientů a 5 stupňů sepse (pro stav 6 hodnoty ACN nebyly k dispozici). Po poradě s odborníky bylo těchto pět skupin pacientů sdruženo do tří o rozsazích 413, 37 a 83 pacientů. Provedený χ^2 test dobré shody nezamítl hypotézu o NB rozdělení ANC v zmíněných třech skupinách septických stavů.

Nejdříve byl proveden test rovnosti parametrů κ (viz odstavec 4.3). Hodnoty tohoto parametru v základním modelu byly $\hat{\kappa} = (3{,}41; 3{,}20; 2{,}42)'$ a za platnosti nulové hypotézy $\tilde{\kappa} = 3{,}18$. Odhady pro parametry μ_1, μ_2, μ_3 vyšly stejné za platnosti alternativy i za platnosti nulové hypotézy $\hat{\mu} = \tilde{\mu} = (10{,}62; 15{,}62; 12{,}72)$. Hodnota testovacích statistik LR, S, W byla postupně 2,62, 2,76 a 3,24. Tyto statistiky se srovnaly s kvantilem $\chi^2_{0,95}(2) = 5{,}99$ a hypotéza o rovnosti parametrů κ se tedy nezamítla.

Poté byl proveden test rovnosti středních hodnot za předpokladu rovnosti parametru κ (viz sekce 4.2). Za platnosti nulové hypotézy byl nalezen odhad $\tilde{\mu} = 11{,}31$, odhad $\hat{\mu}$ je stejný jako v předešlém testu. Pro parametr κ vyšly odhady $\hat{\kappa} = 3{,}18$ a $\tilde{\kappa} = 3{,}05$. Hodnoty testovacích statistik LR, S, W pro test shody středních hodnot za předpokladu rovnosti κ byly postupně 17,35, 19,47 a 13,51. Po srovnání s kvantilem

$\chi^2_{0,95}(2)$ byla hypotéza o shodnosti středních hodnot zamítnuta.

Tedy dané tři skupiny septických stavů se neliší v parametru κ, ale ve střední hodnotě je významný rozdíl.

7.3 Plánování rozsahu experimentu

Pro plánování experimentu je jednou ze základních otázek jaký musí být minimální rozsah datového souboru, aby bylo možno rozlišit konkrétní alternativu od nulové hypotézy.

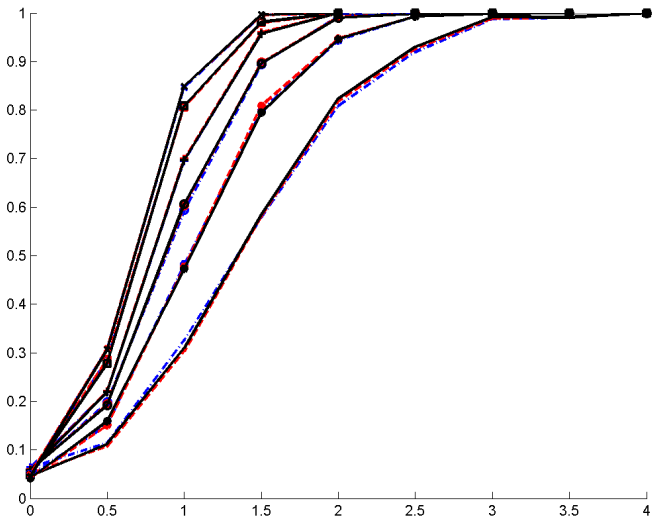
Na základě dat z předchozího odstavce byla provedena simulační studie, která se pokouší dát odpověď na otázku „Jaký je minimální rozsah dat nutný k tomu, aby se dal rozlišit posun ve střední hodnotě větší nebo roven h?“

Simulace byla provedena následovně: Pro nulovou hypotézu bylo zvoleno $\mu_i = \mu = 10$. V alternativě pro μ_i platí $\mu_i = \mu + (1 - i)h$ a h bylo postupně zvoleno (0,5; 1; ...; 4). κ bylo pro všechny výběry rovno 3,18. Pro nulovou hypotézu a každou z alternativ byl výběr 1000× opakován. Z další studie byly vyřazeny výběry u nichž vyšel výběrový rozptyl menší než výběrový průměr (vzhledem k problémům s nalezením ML odhadů κ). Rozsahy výběrů byly voleny postupně 245, 389, 533, 677, 821 a 965. tento celkový rozsah byl pokaždé rozdělen ve stejném poměru v jakém byla rozdělena reálná data (tedy 413:37:83).

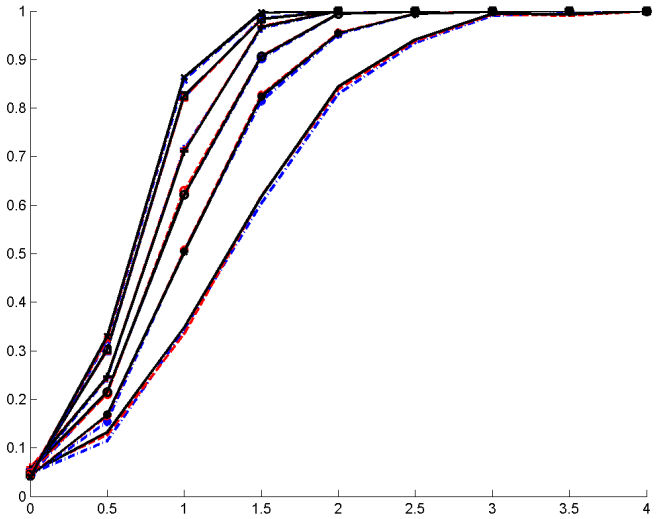
Síla testu rovnosti μ byla, podobně jako v kapitole 6, počítána pro 3 situace: V první předpokládáme κ obecně různá, neznámá (v obrázcích značeno modrou čerchovanou čarou). V druhé je předpoklad, že κ jsou rovny neznámé společné hodnotě (značeno červenou přerušovanou čarou) a v poslední se předpokládají κ rovny známé společné hodnotě (značeno černou plnou čarou).

Na obrázcích 7.1 a 7.2 je na ose x vynesena hodnota h, na ose y je v 7.1 vynesena síla zmíněných testů pro věrohodnostní poměr LR a v 7.2 pro skórovou statistiku S. Pro rozsah 245 je použit znak •, pro rozsah 389 je použit znak ⋆, pro rozsah 533 je použit znak ∘, pro rozsah 677 je použit znak +, pro rozsah 821 je použit znak □, pro rozsah 965 je použit znak ×.

Z uvedených obrázků je např. vidět, že přejeme-li si rozlišit posun v střední hodnotě o $h = 1$ (tedy alternativu $\boldsymbol{\mu} = (10, 11, 12)'$) a chybu druhého druhu 0,7 potřebujeme pro statistiku LR (obr. 7.1) rozsah výběru alespoň 821, rozdělený mezi tři skupiny, tedy (636; 57; 128). Pro statistiku S (obr. 7.2) by stačil rozsah výběru 677, tedy (525; 47; 105).



Obrázek 7.1: Srovnání simulovaných sil pro rozsah postupně 245, 389, 533, 677, 821 a 965 pro statistiku LR.



Obrázek 7.2: Srovnání simulovaných sil pro rozsah postupně 245, 389, 533, 677, 821 a 965 pro statistiku S.

Kapitola 8

Programy

funkce `mom_o(Y,ON)`
Tato funkce dává momentové odhady parametrů μ a κ pro tři základní situace. Tedy momentový odhad κ z celého vektoru `Y`, momentový odhad μ z celého vektoru `Y`, momentový odhad κ z částí vektoru `Y` určených maticí `ON`, momentový odhad μ z částí vektoru `Y` určených maticí `ON`.

```
[kappa_cmc,mi_c,kappa_i,mi_i]=mom_o(Y,ON)
```

Vstupy:
`Y` je vektor vstupních dat,
`ON` je matice 0 a 1 říkající, která část vektoru `Y` tvoří jeden výběr, není-li `ON` zadáno bere se automaticky, že `Y` tvoří jeden výběr.

Výstupy:
`kappa_cmc` - momentový odhad κ z celého vektoru `Y`,
`mi_c` - momentový odhad μ z celého vektoru `Y`,
`kappa_i` - momentový odhad κ z částí vektoru `Y` určených maticí `ON`,
`mi_i` - momentový odhad μ z částí vektoru `Y` určených maticí `ON`,
`kappa_cmr` - momentový odhad κ získaný jako průměr odhadů `kappa_i`.

funkce `NB_k_mle(Y,kappa_mom,ON,rm,rk,mez,rozdil)`
Tato funkce dává maximálně věrohodné odhady parametrů μ a κ pro různé vstupní situace.

```
[mi_mle,kappa_mle,citac2]=NB_k_mle(Y,kappa_mom,ON,rm,rk,mez,rozdil)
```

Vstupy:
`Y` - vektor vstupních dat,
`kappa_mom` - momentový odhad parametru κ (sloupcový vektor, jehož počet řádků odpovídá počtu sloupců následující matice `ON`),
`ON` - matice 0 a 1 říkající, která část vektoru `Y` tvoří jeden výběr, není-li `ON` zadáno bere se automaticky, že `Y` tvoří jeden výběr (podmínkou je, ze `kappa_mom` je číslo),
`rm` a `rk` - určují situaci pro niž hledáme ML odhady:
různá μ, různá κ: `rk`=1, `rm`=1,
různá μ, stejná κ: `rk`=0, `rm`=1,
stejná μ, různá κ: `rk`=1, `rm`=0,
stejná μ, stejná κ: `rk`=0, `rm`=0,
tj.:
`rk` – různé κ: 1 – ano; 0 – ne,
`rm` – různé μ: 1 – ano; 0 – ne,
`mez` - omezující podmínka udávající maximální počet iterací,
`rozdil` - omezující podmínka udávající při jaké přesnosti se má iterační proces zastavit.
Iterační proces se zastaví, jakmile je splněna alespoň jedna z podmínek `mez` a `rozdil`.

Výstupy:
`mi_mle` - ML odhad μ za daných podmínek `rk`, `rm`,
`kappa_mle` - ML odhad κ za daných podmínek `rk`, `rm`,
`citac2` - kontrolní proměnná, ukazuje, kolikrát během výpočtu vyšlo κ záporné.

funkce `NB_c_mle(Y,c_mom)`
Tato funkce dává maximálně věrohodný odhad parametrů μ a c.

```
[c_mle,mi_mle,citac2,krok]=NB_c_mle(Y,c_mom)
```

Vstupy:
`Y` - vektor vstupních dat,
`c_mom` - momentový odhad parametru c.

Výstupy:
`mi_mle` - ML odhad μ,
`c_mle` - ML odhad c,
`citac2` - kontrolní proměnná, ukazuje, kolikrát během výpočtu vyšlo c záporné.

funkce `bias_corr_c(Y,mi_mle,c_mle)`
Výpočet odhadů korigovaných na nestrannost pro parametrizaci μ a c.

```
[cBC,miBC,bc,bm]=bias_corr_c(Y,mi_mle,c_mle)
```

Vstupy:
`Y` - je vektor vstupních dat,
`mi_mle` - je ML odhad parametru μ,
`c_mle` - ML odhad parametru c.

Výstupy:
`cBC` - ML odhad c opravený na nestrannost,
`miBC` - ML odhad μ opravený na nestrannost,
`bc` - odhad vychýlení parametru c,
`bm` - odhad vychýlení parametru μ.

funkce `QL_c(Y,c_mom,typ,iter)`
Vypočte quasi-likelihod odhady (EQL nebo DEQL) pro parametrizaci μ, c.

```
[c_QL,m]=QL_c(Y,c_mom,typ,iter)
```

Vstupy
`Y` - vektor dat,
`c_mom` - momentový odhad c,
`typ`
`typ`=1 spočte EQL odhad c,
`typ`=2 spočte DEQL odhad c,
`iter` udává počet kroků iterace.

Výstupy
`c_QL` - QL odhad dle zvoleného typu (EQL nebo DEQL),
`m` - odhad μ získaný jako průměr `Y`.

funkce `bayes_k(x,L)`
vypočte bayesovské odhady pro parametrizaci μ a κ.

```
[k_b,mu_b]=bayes_k(x,L)
```

Vstupy:
`x` - data,
`L` - rozsah výběru z inverzního gamma rozdělení, který se bude generovat (určuje přesnost), není-li zadáno, nastaví se na 1000.

Výstupy:
`k_b` - bayesovský odhad κ,
`m_b` - bayesovský odhad μ.

funkce `statistiky_E(Y,ON,kappa_mle_H,mi_mle_H,kappa_mle_A1,mi_mle_A1,i)` Vypočte statistiky LR, W a S pro test hypotéz.

```
[LR,W,S]=statistiky_E(Y,ON,kappa_mle_H,mi_mle_H,kappa_mle_A1,mi_mle_A1,i)
```

Vstupy
`Y` - vstupní vektor dat,
`ON` - je matice 0 a 1 říkající, která část vektoru `Y` tvoří jeden výběr,
`kappa_mle_H` - ML odhad κ za platnosti nulové hypotézy,
`mi_mle_H` - ML odhad μ za platnosti nulové hypotézy,
`kappa_mle_A1` - ML odhad κ za platnosti alternativy,
`mi_mle_A1` - ML odhad μ za platnosti alternativy,
`i` - typ testu
`i`=1 - vektor parametru `alpha_i`, `mi`, `kappa_i`,
`i`=2 - vektor parametru `alpha_i`, `mi`, `kappa`,

i=3 - vektor parametru beta_i, kappa, alpha_i,
i=4 - vektor parametru beta_i, kappa, alpha,
i=5 - vektor parametru alpha_i, beta_i, mi, kappa.

Výstupy
LR - devianční statistika (likelihood ratio),
W - Waldova statistika,
S - skórová statistika.

funkce odhad_Fim_0(n,kappa,mi,i)
Vypočte bloky odhadnuté Fisherovy informační matice (FIM).

```
[J112,J11,J12,J22,J]=odhad_Fim_0(n,kappa,mi,i)
```

Vstupy
n - rozsah,
kappa - vektor odhadu κ,
mi - vektor odhadu střední hodnoty,
i - typ testu pro který je matice hledána:
i=1 - vektor parametrů alpha_i, mi, kappa_i,
i=2 - vektor parametrů alpha_i, mi, kappa,
i=3 - vektor parametrů beta_i, kappa, alpha_i,
i=4 - vektor parametrů beta_i, kappa, alpha,
i=5 - vektor parametrů alpha_i, beta_i, mi, kappa.

Výstupy
J112=-(J11-J12*inv(J22)*J12') - potřeba do statistik,
J11, J12, J22 - bloky FIM (po přenásobení -1),
J - FIM.

funkce der1_kmi(Y,ON,kappa,mi)
Vypočte hodnoty první derivace logaritmické věrohodnostní funkce pro konkrétní κ a μ.

```
[dam,dekb]=der1_kmi(Y,ON,kappa,mi)
```

Vstupy:
`Y` - vstupní vektor dat,
`ON` - je matice 0 a 1 říkající, která část vektoru `Y` tvoří jeden výběr,
`kappa` - vektor κ,
`mi` - vektor μ.

Výstupy:
`dam` - první derivace podle μ,
`dekb` - první derivace podle κ.

funkce `Eder2_kmi_1(n,kappa,mi)`
Vypočte odhad střední hodnoty druhé derivace logaritmické věrohodnostní funkce pro konkrétní κ a μ.

```
[a,b,c]=Eder2_kmi_1(n,kappa,mi)
```

Vstupy:
`n` - vektor rozsahu jednotlivých výběrů,
`kappa` - vektor κ,
`mi` - vektor μ.

Výstupy:
`a` - střední hodnota druhé derivace podle μ^2,
`b` - střední hodnota druhé derivace podle κ^2,
`c` - střední hodnota druhé derivace podle μ a κ (tedy vektor 0).

funkce `logvf(Y,ON,kb,am)`
Sdružená logaritmická věrohodnostní funkce pro NB rozdělení.

```
[lvf] = logvf(Y,ON,kb,am)
```

Vstupy:
`Y` - vstupní vektor dat,

`ON` - je matice 0 a 1 říkající, která část vektoru `Y` tvoří jeden výběr,
`kb` - vektor κ,
`am` - vektor μ.

Výstup:
`lvf` - sdružená logaritmická věrohodnostní funkce pro NB.

funkce `stirling(nn,nd)`
Vypočte Stirlingova čísla druhého druhu pro všechna $n \in (1, \texttt{nn}]$ a $d \in (0, \texttt{dn}]$.

```
[alpha]=stirling(nn,dn)
```

Vstupy:
`an`, `dn` - parametry pro Stirlingova čísla druhého druhu, `an` ≥ 1, `dn` ≥ 2.
Výstup:
`alpha` - matice rozměrů `dn`$\times$`nn`, v níž jsou Stirlingova čísla.

Příloha A

Stanovení korekce ML odhadů na nestrannost

A.1 Korekce pro NB rozdělení s parametry μ a c

Nechť $Y_i \sim NB(\mu, c)$. Nejdříve připomeňme hustotu v této parametrizaci

$$f(\mu, c; y_i) = \frac{\Gamma(y_i + c^{-1})}{y_i! \Gamma(c^{-1})} \left(\frac{c\mu}{1+c\mu}\right)^{y_i} \left(\frac{1}{1+c\mu}\right)^{c^{-1}}$$

a sdruženou logaritmickou věrohodnostní funkci

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \big[y_i \ln(\mu) - (y_i + c^{-1}) \ln(1 + c\mu) + y_i \ln c + \\ + \ln \Gamma(y_i + c^{-1}) - \ln \Gamma(c^{-1}) - \ln y_i! \big] .$$

Připomeňme ještě značení zavedené v odstavci 3.3. $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (m, c)'$ pro vektor parametrů a dále U, V a W pro první, druhou a třetí derivaci logaritmické věrohodnostní funkce (3.1). Tedy

$$U_r^{(i)} = \frac{\partial l_i}{\partial \theta_r},\ V_{rt}^{(i)} = \frac{\partial^2 l_i}{\partial \theta_r \partial \theta_t},\ W_{rtu}^{(i)} = \frac{\partial^3 l_i}{\partial \theta_r \partial \theta_t \partial \theta_u},\ r, t, u = 1, 2$$

a dále položme

$$J_{rt} = E\left(-\sum_{i=1}^{n} V_{rt}^{(i)}\right),\ I_{rtu} = E\left(\sum_{i=1}^{n} W_{rtu}^{(i)}\right),\ K_{r,tu} = E\left(\sum_{i=1}^{n} U_r^{(i)} V_{tu}^{(i)}\right).$$

Dále označme

$$\gamma \ln(1+c\mu) + \Psi(c^{-1}),\ \gamma_1 = \Psi'(y_i + c^{-1}) - \Psi'(c^{-1}),\ \gamma_2 = \Psi''(y_i + c^{-1}) - \Psi''(c^{-1}).$$

Pak pro první derivace vypočteme

$$\begin{aligned}
U_m^{(i)} &= \frac{\partial l_i}{\partial \mu} = \frac{y_i}{\mu} - \frac{1+cy_i}{1+c\mu} \\
U_c^{(i)} &= \frac{\partial l_i}{\partial c} = \\
&= c^{-2}\ln(1+c\mu) + \frac{y_i - \mu}{c(1+c\mu)} - \frac{1}{c^2}\big[\Psi(y_i + c^{-1}) - \Psi(c^{-1})\big] = \\
&= \frac{1}{c^2}\left[\ln(1+c\mu) + \Psi(c^{-1})\right] - \frac{1}{c^2}\Psi(y_i + c^{-1}) + \frac{y-\mu}{c(1+c\mu)}
\end{aligned}$$

Pro druhé derivace dostaneme

$$\begin{aligned}
V_{\mu\mu}^{(i)} &= \frac{\partial^2 l_i}{\partial \mu^2} = -\frac{y_i}{\mu^2} + \frac{c(cy_i+1)}{(1+c\mu)^2} \\
&= -\frac{1+2c\mu}{\mu^2(1+c\mu)^2} y_i + \frac{c}{(1+c\mu)^2} \\
V_{\mu c}^{(i)} &= \frac{\partial^2 l_i}{\partial \mu \partial c} = -\frac{y_i(1+c\mu) - (1+cy_i)\mu}{(1+c\mu)^2} = -\frac{y_i - \mu}{(1+c\mu)^2} \\
V_{c\mu}^{(i)} &= V_{\mu c}^{(i)} \\
V_{cc}^{(i)} &= \frac{\partial^2 l_i}{\partial c^2} \\
&= -\frac{2}{c^3}\ln(1+c\mu) + \frac{1}{c^2}\frac{\mu}{1+c\mu} - \frac{y_i - \mu}{c^2(1+c\mu)^2}(1+2c\mu) + \\
&+ \frac{2}{c^3}[\gamma] - \frac{1}{c^2}\left[\Psi'(y_i + c^{-1})(-c^{-2}) - \Psi'(c^{-1})(-c^{-2})\right] \\
&= -\frac{2}{c^3}\ln(1+c\mu) + \frac{\mu}{c^2(1+c\mu)} - \frac{y_i - \mu}{c^2(1+c\mu)^2}(1+2c\mu) + \\
&+ \frac{2}{c^3}[\gamma] + \frac{1}{c^4}\big[\underbrace{\Psi'(y_i + c^{-1}) - \Psi'(c^{-1})}_{\gamma_1}\big]
\end{aligned}$$

a pro třetí derivace pak dostaneme

$$
\begin{aligned}
W^{(i)}_{\mu\mu\mu} &= \frac{\partial^3 l_i}{\partial \mu^3} = 2\left[\frac{y_i}{\mu^3} - \frac{c^2(1+cy_i)}{(1+c\mu)^3}\right] \\
W^{(i)}_{\mu\mu c} &= \frac{\partial^3 l_i}{\partial \mu^2 \partial c} \\
&= \frac{(cy_i + 1 + cy_i)(1+c\mu)^2 - c(cy_i+1)2(1+c\mu)\mu}{(1+c\mu)^4} = \\
&= \frac{1 + 2cy_i - c\mu}{(1+c\mu)^3} = \frac{1}{(1+c\mu)^2} + 2\frac{c(y_i-\mu)}{(1+c\mu)^3} \\
W^{(i)}_{\mu c\mu} &= W^{(i)}_{c\mu\mu} = W^{(i)}_{\mu\mu c} \\
W^{(i)}_{\mu cc} &= \frac{\partial^3 l_i}{\partial \mu \partial c^2} = -(-2)\frac{y_i - \mu}{(1+c\mu)^3}\mu = 2\frac{\mu(y_i-\mu)}{(1+c\mu)^3} \\
W^{(i)}_{c\mu c} &= W^{(i)}_{cc\mu} = W^{(i)}_{\mu cc} \\
W^{(i)}_{ccc} &= \frac{\partial^3 l_i}{\partial c^3} = \\
&= \frac{6}{c^4}\ln(1+c\mu) - \frac{2}{c^3}\frac{\mu}{1+c\mu} - \frac{2}{c^3}\frac{\mu}{1+c\mu} - \frac{1}{c^2}\frac{\mu^2}{(1+c\mu)^2} + \\
&+ (y_i-\mu)\frac{2\mu c^2(1+c\mu)^2 - (1+2c\mu)[2c(1+c\mu)^2 + 2c^2(1+c\mu)\mu]}{c^4(1+c\mu)^4} \\
&- \frac{6}{c^4}[\gamma] + \frac{2}{c^3}\left[\Psi'(y_i + c^{-1})(-c^{-2}) - \Psi'(c^{-1})(-c^{-2})\right] - \\
&- \frac{4}{c^5}[\gamma_1] + \frac{1}{c^4}\left[\Psi''(y_i + c^{-1})(-c^{-2}) - \Psi''(c^{-1})(-c^{-2})\right] = \\
&= \frac{6}{c^4}\ln(1+c\mu) - \frac{\mu(4+5c\mu)}{c^3(1+c\mu)^2} + \\
&- (y_i-\mu)2\frac{\mu c(1+c\mu) - (1+2c\mu)^2}{c^3(1+c\mu)^3} - \\
&- \frac{6}{c^4}[\gamma] - \frac{6}{c^5}[\gamma_1] - \frac{1}{c^6}\left[\Psi''(y_i + c^{-1}) - \Psi''(c^{-1})\right] \\
&= \frac{6}{c^4}\ln(1+c\mu) - \frac{\mu(4+5c\mu)}{c^3(1+c\mu)^2} + (y_i-\mu)\frac{2(1+3c\mu+3c^2\mu^2)}{c^3(1+c\mu)^3} \\
&- \frac{6}{c^4}[\gamma] - \frac{6}{c^5}[\gamma_1] - \frac{1}{c^6}[\gamma_2] \\
&= \frac{6}{c^4}\ln(1+c\mu) + \frac{6\mu y_i}{c^2(1+c\mu)^2} - \frac{4\mu}{c^3(1+c\mu)^2} - \frac{11\mu^2}{c^2(1+c\mu)^2} + \\
&+ \frac{2(y_i-\mu)}{c^3(1+c\mu)^3} - \frac{6}{c^4}[\gamma] - \frac{6}{c^5}[\gamma_1] - \frac{1}{c^6}[\gamma_2]
\end{aligned}
$$

Zaveďme označení:

$$\begin{aligned}
\gamma &= \ln(1+c\mu) + \Psi(c^{-1}) & \Delta_0 &= E(\Psi(y_i + c^{-1})) \\
\Delta_{00} &= E(\Psi^2(y + c^{-1})) & \Delta_1 &= E(\Psi'(y_i + c^{-1})) \\
\Delta_2 &= E(\Psi''(y_i + c^{-1})) & \Delta_{y0} &= E(y\Psi(y_i + c^{-1})) \\
\Delta_{y1} &= E(y\Psi'(y_i + c^{-1})) & \Delta_{01} &= E(\Psi(y_i + c^{-1})\Psi'(y_i + c^{-1}))
\end{aligned}$$

Pro střední hodnoty U_i platí

$$\begin{aligned}
E(U_\mu^{(i)}) &= 0 \\
E(U_c^{(i)}) &= 0
\end{aligned}$$

odtud

$$\begin{aligned}
E(\Psi(y_i + c^{-1})) &= \ln(1+c\mu) + \Psi(c^{-1}) \\
\Delta_0 &= \gamma
\end{aligned}$$

Pro střední hodnoty V_{ij} platí

$$\begin{aligned}
E(V_{\mu\mu}^{(i)}) &= -\frac{1}{\mu(1+c\mu)} \\
E(V_{\mu c}^{(i)}) &= 0 \\
E(V_{cc}^{(i)}) &= -\frac{2}{c^3}\ln(1+c\mu) + \frac{\mu}{c^2(1+c\mu)} + \frac{2}{c^3}[\Delta_0 - \Psi(c^{-1})] + \frac{1}{c^4}[\Delta_1 - \Psi'(c^{-1})] = \\
&= -\frac{2}{c^3}\gamma + \frac{2}{c^3}\Delta_0 + \frac{\mu}{c^2(1+c\mu)} + \frac{1}{c^4}\Delta_1 - \frac{1}{c^4}\Psi'(c^{-1}) = \\
&= \left(\frac{\mu}{c^2(1+c\mu)} + \frac{1}{c^4}\Delta_1 - \frac{1}{c^4}\Psi'(c^{-1})\right)
\end{aligned}$$

Pro střední hodnoty W_{ijk} platí

$$\begin{aligned}
E(W_{\mu\mu\mu}^{(i)}) &= 2\frac{1+2c\mu}{\mu^2(1+c\mu)^2} \\
E(W_{\mu\mu c}^{(i)}) &= \frac{1}{(1+c\mu)^2}
\end{aligned}$$

$$
\begin{aligned}
E(W^{(i)}_{\mu cc}) &= 0 \\
E(W^{(i)}_{ccc}) &= \frac{6}{c^4}\ln(1+c\mu) - \frac{\mu(4+5c\mu)}{c^3(1+c\mu)^2} - \\
&\quad - \frac{6}{c^4}[\Delta_0 - \Psi(c^{-1})] - \frac{6}{c^5}[\Delta_1 - \Psi'(c^{-1})] - \frac{1}{c^6}[\Delta_2 - \Psi''(c^{-1})] = \\
&= -\frac{\mu(4+5c\mu)}{c^3(1+c\mu)^2} - \frac{6}{c^5}[\Delta_1 - \Psi'(c^{-1})] - \frac{1}{c^6}[\Delta_2 - \Psi''(c^{-1})]
\end{aligned}
$$

Než přistoupíme k výpočtu střední hodnoty součinu $U_i V_{jk}$ uveďme nejdříve pomocné výsledky

$$
\begin{aligned}
E(y_i U^{(i)}_{\mu}) &= \frac{Ey^2}{\mu} + \frac{Ey + cEy^2}{1+c\mu} = 1 + c\mu + \mu - \frac{(1+c\mu)(\mu+c\mu)}{1+c\mu} = 1 \\
E(y_i U^{(i)}_{c}) &= \frac{1}{c^2}(\mu\gamma - \Delta_{y0}) + \frac{\mu(1+c\mu)}{c(1+c\mu)} \\
E(\Psi(y_i + c^{-1})U^{(i)}_{c}) &= \frac{1}{c^2}(\gamma\Delta_0 - \Delta_{00}) + \frac{1}{c(1+c\mu)}(\Delta_{y0} - \mu\Delta_0) \\
E(\Psi'(y_i + c^{-1})U^{(i)}_{c}) &= \frac{1}{c^2}(\gamma\Delta_1 - \Delta_{01}) + \frac{1}{c(1+c\mu)}(\Delta_{y1} - \mu\Delta_1)
\end{aligned}
$$

Konečně pro střední hodnoty součinu $U_i V_{jk}$ dostaneme

$$
\begin{aligned}
E[U^{(i)}_{\mu} V^{(i)}_{\mu\mu}] &= -\frac{1+2c\mu}{\mu^2(1+c\mu)^2} E(y_i U^{(i)}_{\mu}) + \frac{c}{(1+c\mu)^2} E(U^{(i)}_{\mu}) = \\
&= -\frac{1+2c\mu}{\mu^2(1+c\mu)^2}
\end{aligned}
$$

$$
\begin{aligned}
E[U^{(i)}_{\mu} V^{(i)}_{\mu c}] &= -\frac{1}{(1+c\mu)^2} E(y U^{(i)}_{\mu}) + \frac{\mu}{(1+c\mu)^2} E(U^{(i)}_{\mu}) = \\
&= -\frac{1}{(1+c\mu)^2}
\end{aligned}
$$

$$
\begin{aligned}
E[U^{(i)}_{c} V^{(i)}_{\mu c}] &= -\frac{1}{(1+c\mu)^2} E(y U^{(i)}_{c}) + \frac{\mu}{(1+c\mu)^2} E(U^{(i)}_{c}) = \\
&= -\frac{1}{c^2(1+c\mu)^2}(\mu\gamma - \Delta_{y0}) - \frac{\mu}{c(1+c\mu)^2} = \\
&= -\frac{\mu}{c(1+c\mu)^2} + \frac{1}{c^2(1+c\mu)^2}(\Delta_{y0} - \mu\Delta_0)
\end{aligned}
$$

$$
\begin{aligned}
E[U_c^{(i)}V_{cc}^{(i)}] = &\left(-\frac{2}{c^3}\ln(1+c\mu) + \frac{\mu}{c^2(1+c\mu)} + \frac{\mu}{c^2(1+c\mu)^2}(1+2c\mu) - \right. \\
&\left. -\frac{2}{c^3}\Psi(c^{-1}) - \frac{1}{c^4}\Psi'(c^{-1})\right) E(U_c^{(i)}) - \frac{1+2c\mu}{c^2(1+c\mu)^2}E(yU_c^{(i)}) + \\
&+ \frac{2}{c^3}E(\Psi(y_i + c^{-1})U_c^{(i)}) + \frac{1}{c^4}E(\Psi'(y_i + c^{-1})U_c^{(i)}) = \\
&= -\frac{1+2c\mu}{c^2(1+c\mu)^2}\left(\frac{1}{c^2}(\mu\gamma - \Delta_{y0}) + \frac{\mu}{c}\right) + \\
&+ \frac{2}{c^3}\left(\frac{1}{c^2}(\gamma\Delta_0 - \Delta_{00}) + \frac{1}{c(1+c\mu)}(\Delta_{y0} - \mu\Delta_0)\right) + \\
&+ \frac{1}{c^4}\left(\frac{1}{c^2}(\gamma\Delta_1 - \Delta_{01}) + \frac{1}{c(1+c\mu)}(\Delta_{y1} - \mu\Delta_1)\right) = \\
&= \frac{3+4c\mu}{c^4(1+c\mu)^2}(\Delta_{y0} - \mu\gamma) - \frac{\mu(1+2c\mu)}{c^3(1+c\mu)^2} + \\
&+ \frac{2}{c^5}(\gamma\Delta_0 - \Delta_{00}) + \\
&+ \frac{1}{c^6}(\gamma\Delta_1 - \Delta_{01}) + \frac{1}{c^5(1+c\mu)}(\Delta_{y1} - \mu\Delta_1)
\end{aligned}
$$

Uvedené výsledky byly použity v odstavci 3.3.

A.2 Korekce pro NB rozdělení s parametry μ a κ

Nechť $Y_i \sim NB(\mu, \kappa)$. Nejdříve připomeňme hustotu v této parametrizaci

$$
f(y;\mu,\kappa) = \frac{\Gamma(y+\kappa)}{\Gamma(y+1)\Gamma(\kappa)}\left(\frac{\mu}{\kappa+\mu}\right)^y\left(\frac{\kappa}{\kappa+\mu}\right)^\kappa
$$

a sdruženou logaritmickou věrohodnostní funkci

$$
\begin{aligned}
l = \sum_{i=1}^{n}\Big[&\ln\Gamma(y_i+\kappa) - \ln\Gamma(y_i+1) - \ln\Gamma(\kappa) \\
&+ \kappa\ln\frac{\kappa}{\kappa+\mu} + y_i\ln\frac{\mu}{\kappa+\mu}\Big].
\end{aligned}
$$

Připomeňme ještě značení zavedené v odstavci 3.3. $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (m, \kappa)'$ pro vektor parametrů a dále U, V a W pro první, druhou a třetí derivaci logaritmické

věrohodnostní funkce (3.1). Tedy

$$U_r^{(i)} = \frac{\partial l_i}{\partial \theta_r},\ V_{rt}^{(i)} = \frac{\partial^2 l_i}{\partial \theta_r \partial \theta_t},\ W_{rtu}^{(i)} = \frac{\partial^3 l_i}{\partial \theta_r \partial \theta_t \partial \theta_u},\ r,t,u = 1,2$$

a dále položme

$$J_{rt} = E\left(-\sum_{i=1}^{n} V_{rt}^{(i)}\right),\ I_{rtu} = E\left(\sum_{i=1}^{n} W_{rtu}^{(i)}\right),\ K_{r,tu} = E\left(\sum_{i=1}^{n} U_r^{(i)} V_{tu}^{(i)}\right).$$

Pak pro první derivace vypočteme

$$U_\mu = \frac{\partial l}{\partial \mu} = \sum_{i=1}^{n}\left[-\frac{\kappa}{\kappa+\mu} + y_i \frac{\kappa}{\mu(\kappa+\mu)}\right] = \sum_{i=1}^{n}\left[-\frac{\kappa+y_i}{\kappa+\mu} + \frac{y_i}{\mu}\right]$$
$$U_\kappa = \frac{\partial l}{\partial \kappa} = \sum_{i=1}^{n}\left[\Psi(y_i+\kappa) - \Psi(\kappa) + \ln\frac{\kappa}{\kappa+\mu} + \frac{\mu}{\kappa+\mu} - y_i\frac{1}{\kappa+\mu}\right]$$

Pro druhé derivace dostaneme

$$V_{\mu\mu} = \frac{\partial^2 l}{\partial \mu^2} = \sum_{i=1}^{n}\left[\frac{\kappa}{(\kappa+\mu)^2} - \frac{\kappa(\kappa+2\mu)}{\mu^2(\kappa+\mu)^2} y_i\right] = \sum_{i=1}^{n}\left[\frac{\kappa+y_i}{(\kappa+\mu)^2} - \frac{y_i}{\mu^2}\right]$$
$$V_{\mu\kappa} = \frac{\partial^2 l}{\partial \mu \partial \kappa} = \sum_{i=1}^{n}\left[\frac{y_i-\mu}{(\kappa+\mu)^2}\right]$$
$$V_{\kappa\kappa} = \frac{\partial^2 l}{\partial \kappa^2} = \sum_{i=1}^{n}\left[\Psi'(y_i+\kappa) - \Psi'(\kappa) + \frac{\mu}{\kappa(\kappa+\mu)} - \frac{\mu}{(\kappa+\mu)^2} + y_i\frac{1}{(\kappa+\mu)^2}\right]$$

a pro třetí derivace pak dostaneme

$$W_{\mu\mu\mu} = \frac{\partial^3 l}{\partial \mu^3} = \sum_{i=1}^{n}\left[-2\frac{\kappa+y_i}{(\kappa+\mu)^3} + 2\frac{y_i}{\mu^3}\right]$$
$$W_{\mu\mu\kappa} = \frac{\partial^3 l}{\partial \mu^2 \partial \kappa} = \sum_{i=1}^{n}\left[-\frac{1}{(\kappa+\mu)^2} - 2\frac{y_i-\mu}{(\kappa+\mu)^3}\right]$$
$$W_{\mu\kappa\kappa} = \frac{\partial^3 l}{\partial \mu \partial \kappa^2} = \sum_{i=1}^{n}\left[-2\frac{y_i-\mu}{(\kappa+\mu)^3}\right]$$
$$W_{\kappa\kappa\kappa} = \frac{\partial^3 l}{\partial \kappa^3} = \sum_{i=1}^{n}\left[\Psi''(y_i+\kappa) - \Psi''(\kappa) + \frac{\mu(2\kappa+\mu)}{\kappa^2(\kappa+\mu)^2} + \right.$$
$$\left. + 2\frac{\mu}{(\kappa+\mu)^3} - 2y_i\frac{1}{(\kappa+\mu)^3}\right]$$

Zaveďme označení:

$$\begin{aligned}
\Delta_1 &= E(\Psi'(y_i+\kappa)) \\
\Delta_2 &= E(\Psi''(y_i+\kappa)) \qquad & \Delta_{y0} &= E(y\Psi(y_i+\kappa)) \\
\Delta_{y1} &= E(y\Psi'(y_i+\kappa)) \qquad & \Delta_{01} &= E(\Psi(y_i+\kappa)\Psi'(y_i+\kappa))
\end{aligned}$$

Pro střední hodnoty U_i platí

$$E(U_\mu) = 0, \qquad E(U_\kappa) = 0$$

Pro střední hodnoty V_{ij} platí

$$\begin{aligned}
E(V_{\mu\mu}) &= -n\frac{\kappa}{\mu(\kappa+\mu)} \\
E(V_{\mu\kappa}) &= 0 \\
E(V_{\kappa\kappa}) &= n\left[\Delta_1 - \Psi'(\kappa) + \frac{\mu}{\kappa(\kappa+\mu)}\right]
\end{aligned}$$

Pro střední hodnoty W_{ijk} platí

$$\begin{aligned}
E(W_{\mu\mu\mu}) &= 2n\frac{\kappa(\kappa+2\mu)}{\mu^2(\kappa+\mu)^2} \\
E(W_{\mu\mu\kappa}) &= -n\frac{1}{(\kappa+\mu)^2} \\
E(W_{\mu\kappa\kappa}) &= 0 \\
E(W_{\kappa\kappa\kappa}) &= n\left[\Delta_2 - \Psi''(\kappa) - \frac{\mu(2\kappa+\mu)}{\kappa^2(\kappa+\mu)^2}\right]
\end{aligned}$$

Než přistoupíme k výpočtu střední hodnoty součinu U_iV_{jk} uveďme nejdříve pomocné výsledky

$$\begin{aligned}
E(yU_\mu) &= \frac{\kappa}{\mu(\kappa+\mu)}\left[-\mu\sum_{i=1}^{p}E(y_i) + \sum_{i=1}^{p}E(y_i^2)\right] = n \\
E(yU_\kappa) &= n\left[\Delta_{y0} - \mu\psi(\kappa) + \mu\ln\frac{\kappa}{\kappa+\mu} - \frac{\mu}{\kappa}\right]
\end{aligned}$$

Konečně pro střední hodnoty součinu $U_i V_{jk}$ dostaneme

$$
\begin{aligned}
E(U_\mu V_{\mu\mu}) &= -n\frac{\kappa(\kappa+2\mu)}{\mu^2(\kappa+\mu)^2} \\
E(U_\mu V_{\mu\kappa}) &= n\frac{1}{(\kappa+\mu)^2} \\
E(U_\mu V_{\kappa\kappa}) &= n\left[\frac{1}{(\kappa+\mu)^2} + \Delta_1\frac{-\kappa}{\kappa+\mu} + \Delta_{y1}\frac{\kappa}{\mu(\kappa+\mu)}\right] \\
E(U_\kappa V_{\mu\mu}) &= -n\frac{\kappa(\kappa+2\mu)}{\mu^2(\kappa+\mu)^2}\left[\Delta_{y0} - \mu\Psi(\kappa) + \mu\ln\frac{\kappa}{\kappa+\mu} - \frac{\mu}{\kappa}\right] \\
E(U_\kappa V_{\mu\kappa}) &= n\frac{1}{(\kappa+\mu)^2}\left[\Delta_{y0} - \mu\Psi(\kappa) + \mu\ln\frac{\kappa}{\kappa+\mu} - \frac{\mu}{\kappa}\right] \\
E(U_\kappa V_{\kappa\kappa}) &= n\left\{\Delta_{01} + \Delta_1\left[-\Psi(\kappa) + \ln\frac{\kappa}{\kappa+\mu} + \frac{\mu}{\kappa+\mu}\right] - \frac{\Delta_{y1}}{\kappa+\mu} + \right. \\
&\quad \left. + \frac{1}{(\kappa+\mu)^2}\left[\Delta_{y0} - \mu\Psi(\kappa) + \mu\ln\frac{\kappa}{\kappa+\mu} - \frac{\mu}{\kappa}\right]\right\}
\end{aligned}
$$

Uvedené výsledky jsou analogií pro použití v odstavci 3.3 pro parametry μ a κ.

Literatura

[1] Aban, I. B., Cutter, G. R., and Mavinga, N. Inferences and power analysis concerning two negative binomial distributions with an application to mri lesion counts data. *Comput. Stat. Data Anal. 53*, 3 (2009), 820–833.

[2] Al-Saleh, M. F., and Al-Batainah, F. K. Estimation of the shape parameter k of the negative binomial distribution. *Appl. Math. Comput. 143*, 2-3 (2003), 431–441.

[3] Anděl, J. *Matematická statistika*. SNTL, Praha, 1978.

[4] Anděl, J. *Základy matematické statistiky*. MATFYZPRESS, Praha, 2005.

[5] Anscombe, F. J. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics 5*, 2 (1949), 165–173.

[6] Bailey, N. T. J. *The elements of stochastic processes with applications to the natural sciences*. John Wiley & Sons Inc., New York, 1964.

[7] Barnwal, R. K., and Paul, S. R. Analysis of one-way layout of count data with negative binomial variation. *Biometrika 75*, 2 (1988), 215–222.

[8] Binet, F. E. Fitting the negative binomial distribution. *Biometrics 42* (1986), 989–992.

[9] Binns, M. Sequential estimation of the mean of a negative binomial distribution. *Biometrika 62*, 2 (1975), 433–440.

[10] Bliss, C. I. Fitting the negative binomial distribution to biological data (with "Note on the efficient fitting of the negative binomial" by R. A. Fisher. *Biometrics 9* (1953), 176–200.

[11] Breslow, N. E. Extra-poisson variation in log-linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 33*, 1 (1984), 38–44.

[12] Clapham, A. R. Over-dispersion in grassland communities and the use of statistical methods in plant ecology. *Journal of Ecology 24* (1936), 232–251.

[13] Clark, S.J.and Perry, J. N. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics 45*, 1 (1989), 309–316.

[14] Cordeiro, G. M., Botter, D. A., and Ferrari, S. L. d. P. Nonnull asymptotic distributions of three classic criteria in generalised linear models. *Biometrika 81*, 4 (1994), 709–720.

[15] Cox, D. R., and Snell, E. J. A general definition of residuals. *J. Roy. Statist. Soc. Ser. B 30* (1968), 248–275.

[16] Das, N. Study of implementing control charts assuming negative binomial distribution with varying sample size in a software industry. *Software Quality Professional 6*, 1 (2003), 38–39.

[17] Daykin, C. D., Pentikäinen, T., and Pesonen, M. *Practical risk theory for actuaries*, vol. 53 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1994.

[18] Dean, C., and Lawless, J. F. Tests for detecting overdispersion in Poisson regression models. *J. Amer. Statist. Assoc. 84*, 406 (1989), 467–472.

[19] Dobson, Annete, J. *An Introduction to Generalized Linear Models*. Chapman & Hall, New York, 1990.

[20] Doudová, L. Srovnání výběrů z NB-rozdělení. In *XXV. mezinárodní kolokvium o řízení osvojovacího procesu: sborník abstraktů a elektronických verzí recenzovaných příspěvků na CD-ROMu*. UO, Brno, 2007, pp. –. Adresář: 6clanky/1doudovl.pdf.

[21] Doudová, L., Heroldová, M., Homolka, M., and Michálek, J. Statistická analýza dat s negativně binomickým rozdělením. In *Biometrické metody a modely v současné vědě a výzkumu – Sborník referátů ze XVII. letní školy biometriky*. ÚKZÚZ, Brno, 2006, pp. 73–79.

[22] Fahrmeir, L. Asymptotic testing theory for generalized linear models. *Statistics 18*, 1 (1987), 65–76.

[23] Fahrmeir, L., and Kaufmann, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist. 13*, 1 (1985), 342–368.

[24] Fisher, R. A. The negative binomial distribution. *Ann. Eugenics 11* (1941), 182–187.

[25] Fisher, R. A., Corbet, A. S., and Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology 12*, 1 (1943), 42–58.

[26] Homolka, M., and Heroldová, M. Native red deer and introduced chamois: foraging habits in a subalpine meadow-spruce forest area. *Folia Zoologica 50* (2001), 89–98.

[27] Homolka, M., and Matouš, J. Density and distribution of red deer and chamois in subalpine meadow habitats in the jeseníky mountains (czech republic). *Folia Zoologica 48* (1999), 1–10.

[28] Hrdličková, Z. Approximation of powers of some tests in one-way MANOVA type multivariate generalized linear model. *Comput. Statist. Data Anal. 52*, 8 (2008), 4059–4075.

[29] Hübnerová, Z., and Doudová, L. One-way anova type model with negative binomial distribution. In *International Environmetrics Society North American Regional Meeting 2007*.

[30] Karpíšek, Z., Jurák, P., and Neradová, V. Divergence a kvazinormy diskrétních rozdělení pravděpodobnosti. In *6th International Conference APLIMAT 2007 (Part I)*. STU, Bratislava, 2007, pp. 387–395.

[31] Lawless, J. F. Negative binomial and mixed Poisson regression. *Canad. J. Statist. 15*, 3 (1987), 209–225.

[32] Lee, Y., and Nelder, J. A. Hierarchical generalized linear models.

[33] Lee, Y., and Nelder, J. A. The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ratio data. *J. Roy. Statist. Soc. Ser. C 49*, 3 (2000), 413–419.

[34] Lee, Y., and Nelder, J. A. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika 88*, 4 (2001), 987–1006.

[35] Lehmann, E. L. *Testing Statistical Hypothesis*, second ed. Springer-Verlag, New York, 1997.

[36] Lehmann, E. L. *Elements of large-sample theory*, second ed. Springer-Verlag, New York, 1998.

[37] Lehmann, E. L., and Casella, G. *Theory of Point Estimation*, second ed. Springer-Verlag, New York, 1998.

[38] Lehmann, G. *Statistik: Einführung in die matematischen Grundlagen für Psychologen, Wirtschafts- und Sozialwissenschaftler.* Spektrum Akademischer Verlag, Heidelberg, Berlin, 2002.

[39] Liese, F., and Vajda, I. f-divergences: sufficiency, deficiency and testing of hypotheses. In *Advances in inequalities from probability theory and statistics*, Adv. Math. Inequal. Ser. Nova Sci. Publ., New York, 2008, pp. 113–149.

[40] Likeš, J., and Machek, J. *Počet pravděpodobnosti.* SNTL, Praha, 1981.

[41] Likeš, J., and Machek, J. *Matematická statistika.* SNTL, Praha, 1988.

[42] Mandl, P. *Pravděpodobnostní dynamické modely.* Academia, Praha, 1985.

[43] Maul, A., and El-Shaarawi, A. H. Analysis of two-way layout of count data with negative binomial variation. *Environmental Monitoring and Assessment 17*, 2–3 (1991), 315–322.

[44] McConnell, B. R., and Smith, J. G. Frequency distributions of deer and elk pellet groups. *J. Wildl. Manage. 34*, 1 (1970), 29–36.

[45] McCullagh, P. Quasilikelihood functions. *Ann. Statist. 11*, 1 (1983), 59–67.

[46] McCullagh, P., and Nelder, J. A. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983.

[47] McLachlan, G., and Peel, D. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

[48] Michálek, J. Zobecněný lineární model - aplikace v biometrice. In *Sborník prací letní školy Biometrické metody a modely v současné vědě a výzkumu*. ÚKZÚZ, Brno, 2002, pp. 13–21.

[49] Michálek, J., and Michálek, J. j. Hierarchické modely pro mnohorozměrné kontingenční tabulky s aplikacemi v medicíně. In *XXVII. mezinárodní kolokvium o řízení osvojovacího procesu: sborník abstraktů a elektronických verzí recenzovaných příspěvků na CD-ROMu*. FEM UO, Brno, 2008. Adresář: 6clanky/2michalej.pdf.

[50] Michálek, J., and Michálek, J. j. Statistická analýza genů ovlivňujících průběh sepse. In *Biometrické metody a modely v současné vědě a výzkumu – Sborník referátů ze XVIII. letní školy biometriky*. Agentura SAPV, Nitra, 2008, pp. 79–93.

[51] Michálek, J., Světlíková, P., Fedora, P., Klimovič, M., Klapačová, L., Bartošová, D., Hrstková, H., and Hubáček, J. Interleukine - 6 gene variants and the risk of sepsis development in children. *Human Immunology 68* (2007), 756–760.

[52] Michálek, J., Světlíková, P., Fedora, P., Klimovič, M., Klapačová, L., Bartošová, D., Hrstková, H., and Hubáček, J. A. Bactericidal permeability increasing protein gene variants in children with sepsis. *Intensive Care Medicine 33* (2007), 2158–2164.

[53] Michálek, J., Šmerek, M., and Šotová, J. *Matematické modelování rizik*. FEM UO, Brno, 2007.

[54] Montgomery, D. C. *Introduction to Statistical Quality Control*. John Wiley & Sons, New York, 1996.

[55] Nelder, J. A., and Pregibon, D. An extended quasilikelihood function. *Biometrika 74*, 2 (1987), 221–232.

[56] Ong, S. H., and Lee, P. A. The noncentral negative binomial distribution. *Biometrical J. 21*, 7 (1979), 611–627.

[57] Piegorsch, W. W. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics 46*, 3 (1990), 863–867.

[58] Power, J. H., and Moser, E. B. Linear model analysis of net catch data using the negative binomial distribution. *Can. J. Fish. Aquat. Sci. 56* (1999), 191–200.

[59] Rao, R. C. *Lineární metody statistické indukce a jejich aplikace.* Academia, Praha, 1978.

[60] Resnick, S. I. *Adventures in Statistic Processes.* Birkhäuser, Boston, 2002.

[61] Ross, G. J. S., and Preece, D. A. The negative binomial distribution. *The Statistician 34* (1985), 323–336.

[62] Ryan, T. P. *Statistical Methods for Quality Improvement.* Wiley series in probability and statistic. Wiley, New York, 2000. Second Edition.

[63] Saha, K., and Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics 61*, 1 (2005), 179–185.

[64] Světlíková, P., Fornůsek, M., Fedora., M., Klapačová, L., Bartošová, D., Hrstková, H., Klimovič, M., Novotná, E., Hubáček, J., and Michálek, J. Sepsis characteristics in children with sepsis. In *Česko-Slovenská Pediatrie*, 59. 2004, pp. 632–636.

[65] Vajda, I. *Theory of statistical Inference and Information.* Kluwer Academic Publishers, Dordrecht, Boston, 1989.

[66] White, G. C., and Eberhardt, L. E. Statistical analysis of deer and elk pellet-group data. *J. Wildl. Manage. 44*, 1 (1980), 121–131.

[67] Wimmer, G. *Diskrétne jednorozmerné rozdelenia pravdepodobnosti.* MATFYZPRESS, Praha, 2000.

[68] Wimmer, G., and Altmann, G. *Thesaurus of univariate discrete probability distributions*. STAMM, Essen, 1999.

[69] Zacks, S. *The theory of statistical inference*. John Wiley & Sons Inc., New York, 1971. Wiley Series in Probability and Mathematical Statistics.