

INTERAKTIVNÍ STATISTICKÁ ANALÝZA V OBORU PĚSTOVÁNÍ LESA - I. VYHODNOCENÍ JEDNOROZMĚRNÝCH DAT

COMPUTER-ASSISTED STATISTICAL ANALYSIS IN SILVICULTURE - I. UNIVARIATE DATA TREATMENT

DAVID DUŠEK¹⁾ - MILAN MELOUN²⁾ - JIŘÍ NOVÁK¹⁾

¹⁾Výzkumný ústav lesního hospodářství a myslivosti, v. v. i., VS Opočno

²⁾Univerzita Pardubice, Katedra analytické chemie, Pardubice

ABSTRACT

Forest research should be based on the computer-assisted statistical data analysis. At present, a progress of both new methods and software provides full service for researcher. This paper is focused on a detailed description of the computerized interactive statistical analysis of an one-dimensional data. Application of these methods is demonstrated on two examples (large and small sample) from a silvicultural discipline. A special attention is paid to the exploratory data analysis. From the results it can be concluded, that Box-Cox transformation and following calculation of re-transformed mean was suitable procedure for asymmetrically distributed data in a case of large sample. Concerning small sample, the Horn procedure leads to correct results in comparison with classical methods. Therefore, we suppose wide scope of activity of both techniques in a silvicultural discipline.

Klíčová slova: průzkumová analýza dat, analýza jednorozměrných dat, Hornův postup, pěstování lesa

Key words: exploratory data analysis, analysis of one-dimensional data, Horn procedure, silviculture

ÚVOD

S rozvojem výpočetní techniky a dostupného softwaru pro osobní počítače vzrůstá uplatnění matematicko-statistických metod v lesnickém výzkumu i lesnické praxi. Statistické metody v lesnictví začal ve 20. letech minulého století jako první ve střední Evropě systematicky používat prof. Alexandr Leporský (LEPORSKÝ 1953). V dalším období vznikla řada učebních textů pro tento obor jak tuzemských (např. STONE 1963, ŠMELKO, WOLF 1977, DRÁPELA, ZACH 1996, 1999, DRÁPELA 2000), tak i zahraničních (např. PRODAN 1961, KOZAK et al. 2008). V dobách před masovým uplatněním počítačů byl výzkumník zpravidla odkázán na jednoduché aproximativní postupy, často bez ověřování základních předpokladů použitých metod. Dnes umožňuje moderní výpočetní technika interaktivní statistickou analýzu dat za pomoci rigorózních matematických postupů.

Článek je zaměřen na popis a příkladné použití moderních metod počítačově orientované interaktivní statistické analýzy jednorozměrných dat, použitelné především v lesnickém výzkumu, konkrétně v oboru pěstování lesa. Zvláštní pozornost je zde věnována průzkumové analýze dat a nutnému ověření základních předpokladů o výběru dat.

MATERIÁL A METODIKA

Popis a použití statistických metod jsou demonstrovány na dvou příkladech dat získaných z experimentů Výzkumného ústavu lesního hospodářství a myslivosti, v. v. i., s porostní výchovou. Pro příklad analýzy velkého výběru byl analyzován soubor 232 výčetních tloušťek naměřených kalibrovanou milimetrovou průměrkou na nevychováva-

né kontrolní ploše experimentu s výchovou smrku ztepilého (DUŠEK, SLODIČÁK 2009, tab. 1). Pro příklad analýzy malého výběru (Hornův postup) byla použita data laboratorního stanovení (při 80 °C) hmotnosti sušiny holorganických horizontů (L + F + H) odebraných ze sedmi půdních sond v porostech smrku pichlavého (ŠPULÁK, DUŠEK 2009, tab. 4). Analýzy byly provedeny podle následujících postupů:

Analýza velkých výběrů

Obecný postup statistické analýzy jednorozměrných dat

1. V průzkumové analýze dat (EDA) se vyšetřují statistické zvláštnosti dat. Jedná se především o detekci lokální koncentrace dat, určení parametrů tvaru rozdělení výběru (šikmost, špičatost), identifikace odlehklých hodnot a odhalení odchylek od předpokládaného normálního rozdělení (MELOUN, MILITKÝ 2004). Tato fáze analýzy umožňuje odhalit nečekané a na první pohled obtížně zjištěitelné vlastnosti dat, čímž se získá hlubší vhled do zkoumané problematiky (HENDL 2004).
2. Ověření základních předpokladů kladených na výběr se provádí vždy i v případě rutinních měření. Jedná se především o rozdělení výběru a jeho homogenitu, nezávislost prvků a dostatečný rozsah výběru. Pouze při splnění základních předpokladů je možné užít klasických odhadů polohy a rozptýlení a jejich intervalových odhadů.
3. V případě porušení některého z předpokladů o výběru je nutné použít transformaci dat, a to mocninnou, exponenciální nebo Boxo-Coxovu.

Při vyčíslení nejlepších odhadů parametrů polohy, rozptýlení a tvaru se vedle klasických odhadů, jako je aritmetický průměr, rozptyl a směrodatná odchylka, nabízí také robustní odhady (medián, uřezaný průměr, winsorizovaný rozptyl) a adaptivní M-odhady.

Techniky EDA

Při průzkumové analýze dat se vychází z pořádkových statistik $x_{(i)}$, tj. ze vzestupně uspořádaných prvků výběru. Symbol $P_i = i/(n+1)$ označuje pořadovou pravděpodobnost. Hodnoty P_i jsou závislé na předpokládaném rozdělení výběru. Pro normální rozdělení se doporučuje $P_i = (i - 3/8)/(n + 1/4)$ a pro účely průzkumové analýzy se často volí $P_i = (i - 1/3)/(n + 1/3)$. Vynesením hodnot $x_{(i)}$ proti P_i se získá hrubý odhad kvantilové funkce $Q(P)$, jenž jednoznačně charakterizuje rozdělení výběru. Pro grafická znázornění exploratorní analýzy je možno použít následující diagnostické grafy:

- Diagram rozptýlení představuje jednorozměrnou projekci kvantilového grafu na osu x . V rozmitnutém diagramu rozptýlení jsou body pro lepší přehlednost náhodně rozmitnuté na ose y . Oba diagramy odhalí lokální koncentraci dat a také odlehla či podezřelá měření.
- Kvantilový graf umožňuje rozlišit tvar rozdělení a jeho případné zešikmení k vyšším nebo nižším hodnotám. Na osu x se vynášejí pořadová pravděpodobnost $P_i = i/(n+1)$, na osu y pořadová statistika $x_{(i)}$. Pro snadnější porovnání s normálním rozdělením se do grafu obvykle zakresluje i kvantilová funkce normálního rozdělení, založená na klasických a robustních odhadech. Podle toho, která z křivek lépe vystihuje data, lze jako odhad střední hodnoty volit aritmetický průměr nebo medián.
- V grafu rozptýlení s kvantily se na osu x vynášejí $P_i = (i - 1/3)/(n + 1/3)$ a na osu y pořadová statistika $x_{(i)}$. Navíc se zde zakreslují tři pomocné kvantilové obdélníky: kvartilový obdélník F, oktilový obdélník E a sedecilový obdélník D. Vzájemná poloha obdélníků indikuje symetrii rozdělení. V případě symetrického unimodálního rozdělení leží obdélníky symetricky uvnitř sebe. Odlehla pozorování se projeví náhlým vzrůstem kvantilové funkce mimo sedecilový obdélník. Vícemodální rozdělení se projeví úseky s téměř nulovou směrnici uvnitř obdélníku F.
- Diagram percentilů slouží k posouzení symetrie a tvaru rozdělení. Hodnoty percentilů jsou na osu y vynášeny ve vhodném intervalu. Obvykle se volí intervaly 0 - 2, 2 - 5, 5 - 10, 10 - 15, 15 - 25, 25 - 35, 35 - 45, 45 - 55, 55 - 65, 65 - 75, 75 - 85, 85 - 90, 90 - 95, 95 - 99, 99 - 100.
- Graf jádrového odhadu hustoty pravděpodobnosti pomáhá porovnat empirický jádrový odhad rozdělení s aproximační křivkou zpravidla normálního rozdělení. Slouží k posouzení rozložení dat, jejich případné asymetrie nebo k identifikaci nehomogenity v datech. Pro tvar jádra se obvykle volí normální rozdělení. Podobu grafu zásadně ovlivní zvolená šířka jádra. Podrobnosti ke konstrukci grafu jádrového odhadu hustoty pravděpodobnosti lze nalézt např. v MELOUN, MILITKÝ (2004) nebo VENABLES, RIPLEY (1999).
- Histogram četností dat v jednotlivých třídách s konstantní šířkou patří k nejstarším technikám. Kvalita histogramu je ovlivněna počtem tříd L . V širokém rozmezí velikosti výběrů n lze použít vztah $L = \text{int}(2.46(n - 1)^{0.4})$.
- Graf pohoří vzniká modifikací grafu empirické distribuční funkce. Vynášejí se hodnoty $y_i = 100 P_i$ pro $P_i \leq 0.5$ a $y_i = 100 - 100 P_i$ pro $P_i > 0.5$ v závislosti na $x_i = x_{(i)}$. Tento graf má vrchol při hodnotě mediánu a umožňuje identifikovat asymetrii, odlehla měření nebo směs více rozdělení.
- Graf polosum slouží jako citlivý indikátor asymetrie rozdělení. Na osu x se vynášejí pořadkové statistiky $x_{(i)}$, na osu y hodnoty polosum $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$. V případě symetrického rozdělení body oscilují kolem horizontální přímky představující medián

a měřítko osy y je detailní. Pro snadnější posouzení náhodnosti či nenáhodnosti trendu lze do grafu vykreslit i pomocný intervalový odhad mediánu.

- Graf symetrie má podobný význam jako graf polosum. Na osu x se vynášejí $0.5 \times u^2 P_i$ pro $P_i = i/(n+1)$, na osu y hodnoty polosum $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$, kde u_{P_i} jsou kvantily normalizovaného normálního rozdělení. V případě asymetrického rozdělení vykazují body výrazný trend. Směrnice odhaleného trendu je úměrná šikmosti rozdělení.
- Diferenční kvantilový graf je pomůckou, která slouží k posouzení rozdělení se špičatostí odpovídající normálnímu rozdělení. V případě normální špičatosti leží data na horizontální přímce. Na osu x se vynášejí kvantily normalizovaného normálního rozdělení u_{P_i} , na osu y : $d_{(i)} = x_{(i)} - s u_{P_i}$ kde $s = 0.74128 R_F$ je robustní odhad směrodatné odchylky a R_F je interkvartilové rozpětí.
- Krabicový graf je tvořen krabičkou představující na začátku 25% dolní kvantil F_D a na konci krabice 75% percentil čili horní kvantil F_H . Čarou uvnitř krabičky se zobrazuje medián jako robustní odhad parametru polohy. V případě vrubového krabicového grafu je navíc vykreslen zářez v krabici představující robustní interval spolehlivosti mediánu. Pro meze tohoto intervalu platí

$$I_D = M - \frac{1.57 R_F}{\sqrt{n}} \quad \text{a} \quad I_H = M + \frac{1.57 R_F}{\sqrt{n}}$$

kde R_F je interkvartilové rozpětí $\tilde{x}_{0.75} - \tilde{x}_{0.25}$. Dvě úsečky vně krabice, označované jako „vousy“, představují nejbližší data uvnitř tzv. vnitřních hradeb. Data vně vnitřních hradeb jsou považována za odlehle hodnoty.

- Rankitový Q-Q graf poslouží k odhalení diagnostik normality a odlehklých hodnot. Na osu x se vynášejí $P_i = (i - 0.5)/n$ pro $n \geq 11$ nebo $P_i = (i - 3/8)/(n + 1/4)$ pro $n \leq 10$ (VENABLES, RIPLEY 1999) a na osu y pořadkové hodnoty $x_{(i)}$. Pro data s normálním rozdělením má přibližně tvar přímky. V případě odlehklých měření leží koncové body mimo přímku. Graf umožňuje určit, je-li odchylka od normality způsobena jen několika měřeními nebo všemi daty.
- Pravděpodobnostní P-P graf je alternativou ke Q-Q grafům. V případě shody výběrového rozdělení se zvoleným teoretickým (nejčastěji normálním) rozdělením leží body na přímce s jednotkovou směrnici a nulovým úsekem. Odchylky od teoretického rozdělení výběru se projeví především ve střední části grafu.

Ověření základních předpokladů o výběru

Na předpokladu normality výběrového rozdělení je založena celá klasická analýza dat. Z mnoha testů normality uvedeme test podle Jarque-Bera (JARQUE, BERA 1987)

$$LM = n \left(\frac{(b_1^{1/2})^2}{6} + \frac{(b_2 - 3)^2}{24} \right),$$

kde $b_1^{1/2} = m_3 / m_2^{3/2}$, $b_2 = m_4 / m_2^2$, m_i je i -tý centrální moment

$$m_i = \sum (x_j - \bar{x})^i / n$$

a n je velikost výběru.

Často je používána modifikace (URZUA 1996)

$$ALM = \left(\frac{(b_1^{1/2})^2}{c_1} + \frac{(b_2 - c_2)^2}{c_3} \right),$$

$$\text{kde } c_1 = \frac{6(N-2)}{(N+1)(N+3)}, \quad c_2 = \frac{3(N-1)}{(N+1)}, \quad c_3 = \frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}$$

Za předpokladu normality má statistika asymptoticky $\chi^2_{1-\alpha}(2)$ rozdělení. Přesněji lze kritické kvantily určit např. z Monte Carlo simulací, což je vhodné především pro menší výběry.

Testy normality bývají však méně citlivé v porovnání s diagnostickými grafy a jejich závěry jsou obvykle vágní. U malých výběrů často zcela selhávají. Je proto výhodné využít výše uvedených diagnostických grafů průzkumové analýzy, které navíc dokáží podat informaci, jedná-li se o systematické odchýlení, či zdali je odchylka od normality způsobena jen několika odlehlými body.

K nehomogenitě výběru dochází tam, kde se vyskytuje nerovnoměrnost ve vlastnostech vzorku nebo tam, kde se výrazně mění podmínky v průběhu experimentu. Je možné se pokusit rozdělit daný výběr do podskupin a poté analyzovat každou podskupinu zvlášť. V případě vybočujících měření, která silně ovlivňují odhady parametrů polohy a rozptýlení, je možné tato odlehlá měření identifikovat a případně je vyloučit z další analýzy. Jednoduchou technikou identifikace vybočujících bodů za předpokladu normálního rozdělení je například modifikace vnitřních hradeb B_D a B_H podle vztahů

$$B_D = \tilde{x}_{0.25} - K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \text{ a } B_H = \tilde{x}_{0.75} + K(\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

kde $\tilde{x}_{0.25}$ a $\tilde{x}_{0.75}$ je dolní a horní kvartil.

Hodnotu parametru K pro pravděpodobnost 95 %, že žádný prvek z normálního rozdělení o rozsahu n nebude mimo vnitřní hradby, lze určit v rozmezí $8 \leq n \leq 100$ podle aproximace

$$K \approx 2,25 - \frac{3,6}{n}.$$

Body ležící mimo tyto hradby se považují za vybočující.

Při nekonstantních podmínkách měření, změnách stavu měřicího zařízení nebo při nesprávném, nenáhodném výběru vzorků dochází k porušení předpokladu nezávislosti prvků výběru. To se projeví zvýšeným rozptylem oproti rozptylu homogenního výběru a rozptyl výběrového průměru pro závislá data nemusí klesat s růstem velikosti výběru. K identifikaci časové závislosti měření nebo k identifikaci závislosti, jenž souvisí s pořadím jednotlivých, měření se testuje významnost autokorelačního koeficientu prvního řádu ρ_1 podle von Neumannova kritéria

$$t_n = \frac{T_1 \sqrt{n+1}}{\sqrt{1-T_1}}, \text{ kde } T_1 = 1 - \frac{T}{2} \sqrt{\frac{n^2-1}{n^2-4}}$$

a T je von Neumannův poměr

$$T = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Pokud platí, že $|t_n| > t_{1-\alpha/2}(n+1)$, zamítáme hypotézu o nezávislosti prvků výběru na zvolené hladině významnosti α . Mnoho testů hypotéz (např. t -test, ANOVA aj.) je založeno na předpokladu nezávislosti dat. Není-li podmínka nezávislosti splněna, je nutno prověřit celý experimentální design. V některých případech není možné zajistit nezávislý výběr. Takové závislé výběry se označují jako selekční výběry a modely, které zahrnují omezení, za nichž byly výsledky získány, se nazývají selekční modely.

Dostatečný rozsah výběru pozitivně ovlivní přesnost odhadů parametru polohy a rozptýlení. S rostoucím rozsahem výběru dochází ke zúžení intervalů spolehlivosti odhadů, klesá riziko chyby II. druhu a roste tak síla testů. Pro výběry pocházející z normálního rozdělení lze jejich minimální velikost určit např. ze vztahu

$$n_{\min} = \left[\frac{t_{1-\alpha/2}(n_1-1)}{d} \right]^2 s_0^2(x),$$

kde d je zvolené číslo představující maximální přípustnou chybu, $t_{1-\alpha/2}(n_1-1)$ je kvantil Studentova rozdělení s (n_1-1) stupni volnosti, n_1 je počet předběžných hodnot, z nichž se určí odhad výběrového rozptylu a $s_0^2(x)$ je odhad výběrového rozptylu pokusně zvoleného výběru. Minimální velikost výběru může být volena také s ohledem na relativní chybu směrodatné odchylky $\delta(s)$ podle vztahu

$$n_{\min} = \frac{g_2(x)-1}{4\delta^2(s)} + 1,$$

kde $g_2(x)$ je špičatost rozdělení výběru.

Transformace dat

K nelineární transformaci dat je vhodné přistoupit v případech, kdy je na základě předchozí analýzy identifikováno rozdělení zásadně odlišné od normálního rozdělení. Uvedené transformace jsou vhodné pro asymetrická unimodální rozdělení a vedou ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení.

Mocninná transformace vede k zesymetričtění rozdělení výběru, ale nezachovává měřítka a je vhodná pouze pro kladná data. Mocninná transformace má tvar

$$y = g(x) = \begin{cases} x^\lambda & (\lambda > 0) \\ \ln x & \text{pro } (\lambda = 0) \\ -x^{-\lambda} & (\lambda < 0) \end{cases}$$

Optimální hodnota parametru λ se volí taková, která vede k minimální šikmosti a k hodnotě špičatosti nejbližší normálnímu rozdělení.

Exponenciální transformace je použitelná i pro záporné hodnoty a má tvar

$$y = g(x) = \begin{cases} \frac{e^{\lambda x} - 1}{\lambda} & (\lambda \neq 0) \\ x & \text{pro } (\lambda = 0) \end{cases}$$

Boxova-Coxova transformace má tvar

$$y = g(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \ln x & (\lambda = 0) \end{cases} \text{ pro}$$

Pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti. Logaritmus věrohodnostní funkce má tvar

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^n \ln x_i,$$

kde $s^2(y)$ je rozptyl transformovaných dat. Do grafu logaritmu věrohodnostní funkce lze také zakreslit 95% intervaly spolehlivosti. Transformace je tím výhodnější, čím jsou intervaly spolehlivosti užší. Pokud ovšem tento interval obsahuje hodnotu +1, není transformace ze statistického hlediska přínosná a lze užít aritmetický průměr výběru.

Jednoduchý postup zpětné transformace v Boxově-Coxově transformaci pro $\lambda \neq 0$ vede na retransformovaný průměr definovaný vztahem

$$\begin{aligned} \bar{x}_R &= g^{-1}(\bar{y}) = \exp(\bar{y}) & \text{pro } \lambda = 0 \\ \bar{x}_R &= g^{-1}(\bar{y}) = (1 + \bar{y}\lambda)^{\frac{1}{\lambda}} & \text{pro } \lambda \neq 0, \end{aligned}$$

kde $\bar{y}$ je aritmetický průměr transformovaných dat. Tyto odhady jsou však poněkud vychýlené.

Korektnějším postupem zpětné transformace je vyčíslení střední hodnoty a rozptylu transformovaných dat a následná retransformace s využitím Taylorova rozvoje v okolí transformovaného průměru, jenž vede k odhadům nových retransformovaných parametrů polohy a rozptýlení. Pro Boxovu-Coxovu transformaci, kde $\lambda \neq 0$, bude $\bar{x}_R$ jedním z kořenů kvadratické rovnice, pro které platí

$$\bar{x}_{R,1,2} = \left[0,5(1 + \bar{y}\lambda) \pm 0,5\sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2s^2(y))} \right]^{\frac{1}{\lambda}},$$

kde $\bar{y}$ je aritmetický průměr transformovaných dat a $s^2(y)$ je rozptyl těchto dat. Odhadem $\bar{x}_R$ bude kořen $\bar{x}_{Ri}$, který je blíže mediánu $\tilde{x}_{0,5} = g^{-1}(\tilde{y}_{0,5})$. Pro případ, kdy $\lambda = 0$, bude odhadem retransformovaného průměru

$$\bar{x}_R = \exp(\bar{y} + 0,5s^2(y)).$$

Analýza malých výběrů

Malých výběrů užíváme pouze tam, kde z nějakých důvodů experimentu (biologických, technických, ekonomických) není možné dosáhnout vyššího počtu měření, neboť závěry učiněné na základě analýzy malých výběrů jsou vždy zatíženy větší mírou nejistoty. Pro stanovení odhadů polohy a rozptýlení výběrů o velikosti od 4 do 20 měření lze využít Hornův postup pivotů. Výpočet vychází z hloubky pivotů, která je definována vztahy $h = \text{int}((n+1)/2)/2$ nebo $h = \text{int}(((n+1)/2) + 1)/2$ podle toho, které h vyjde jako celé číslo. Dolní pivot je $x_D = x_{(h)}$ a horní pivot $x_H = x_{(n+1-h)}$. Odhadem parametru polohy je pivotová polوسума $P_L = (x_D + x_H)/2$ a odhadem parametru rozptýlení pivotové rozpětí $R_L = x_H - x_D$. Náhodná veličina $T_L = P_L/R_L$ má přibližně symetrické rozdělení, jejíž kvantily jsou tabelovány v např. MELOUN, MILITKÝ (2004). Pro 95% intervaly spolehlivosti střední hodnoty potom platí nerovnost $P_L - T_{L,0.975}(n) \leq \mu \leq P_L + T_{L,0.975}(n)$.

Zpracování dat bylo provedeno v softwarech ADSTAT 2.0, QC-Expert 2.5 a R 2.8.1. Grafy v článku byly vytvořeny v programu R 2.8.1, což je nekomerční obdoba programu S-Plus. Skripty jazyka R pro jednorozměrnou analýzu dat včetně grafických výstupů použitých v tomto článku jsou k dispozici u autorů.

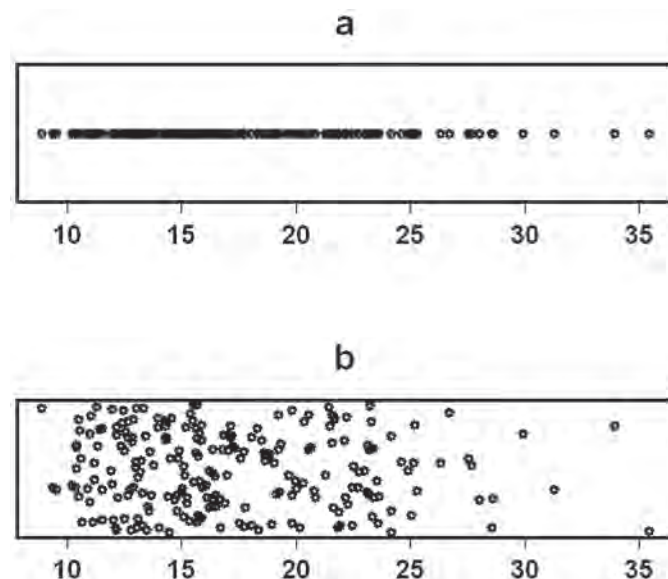
VÝSLEDKY A DISKUSE

Analýza velkého výběru

Exploratorní analýza jednoznačně ukázala na levostranně asymetrické rozdělení. Již z diagramu rozptýlení (obr. 1) je patrná větší koncentrace bodů v jeho levé části. Lepší pohled na data poskytují grafy na obrázku 2 a i zde je na všech grafech identifikována zřejmá asymetrie. Jako zvláště citlivý na porušení symetrie se jeví graf polوسum nebo graf symetrie. Diferenční kvantilový graf indikuje špičatost výrazně odlišnou od normální. Dva body za vnitřními hradbami krabicového grafu jsou zřejmě důsledkem levostranného rozdělení a neměly by být zde považovány za odlehlá měření. Také výrazná nelinearita v rankitovém Q-Q grafu jasně ukazuje na neshodu s normálním rozdělením a stejně lze interpretovat i P-P graf. Empirické rozdělení bylo kromě normálního porovnáno také s rozděleními Weibullovým, gamma a logaritmicko-normálním. Parametry Weibullova a gamma rozdělení byly vypočítány metodou maximální věrohodnosti. Z P-P grafu se jako nejvhodnější jeví aproximace logaritmicko-normálním rozdělením (obr. 3).

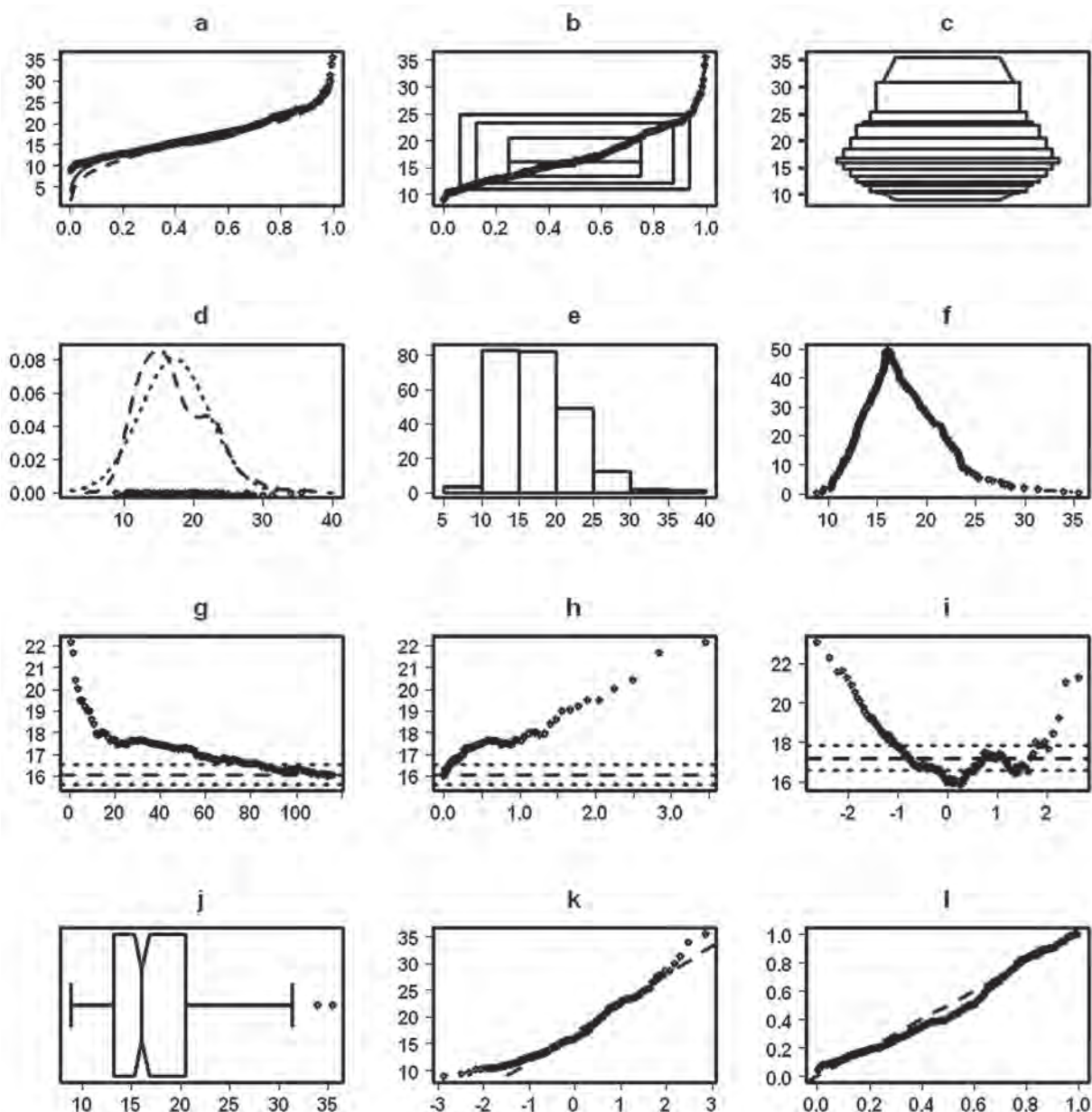
Závěry učiněné z diagnostických grafů jsou také v souladu s výsledky provedených statistických testů. Testem normality podle Jarque-Bera byla zamítnuta nulová hypotéza o shodě s normálním rozdělením ($p < 0,001$). Test homogenity založený na modifikovaných vnitřních hradbách neidentifikoval žádná odlehlá měření a výběr je možno považovat za homogenní. Tento test je nicméně založen na předpokladu normality, která byla zamítnuta, a proto není v tomto případě jeho použití zcela korektní. Ani test nezávislosti neprokázal závislost v datech ($p = 0,46$) a vzhledem k charakteru dat byl tento výsledek rovněž vysoce pravděpodobný. Rozsah výběru je dostatečný k tomu, aby relativní chyba směrodatné odchylky nebyla větší než 5,2 % (tab. 2).

Byly také vyčísleny klasické i robustní parametry polohy, parametry rozptýlení a tvaru (tab. 3). Vzhledem ke zjištěné asymetrii výběrového rozdělení lze předpokládat, že klasický odhad střední hodnoty bude zkreslený. Proto byl vyčíslen retransformovaný průměr po Boxově-Coxově transformaci. Maximalizací logaritmu věro-



Obr. 1.

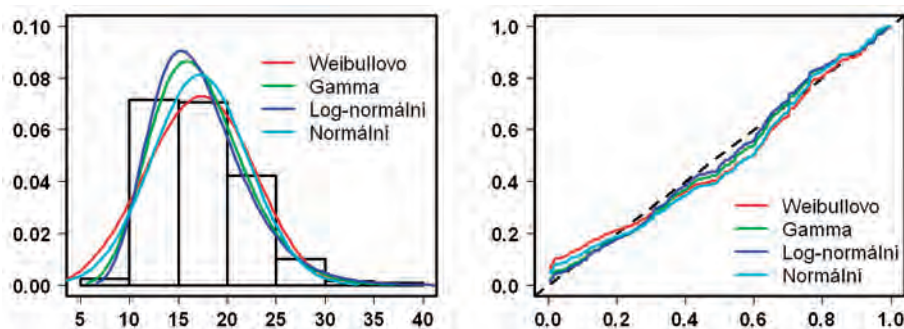
Diagram rozptýlení (a) a rozmitnutý diagram rozptýlení (b)
Dot plot (a) and jittered dot plot (b)

**Obr. 2.**

Kvantilový graf (a), graf rozptýlení s kvantily (b), diagram percentilů (c), graf jádrového odhadu hustoty pravděpodobnosti (d), histogram (e), graf pohorí (f), graf polosum (g), graf symetrie (h), diferenční kvantilový graf (i), vrubový krabicový graf (j), Q-Q graf (k) a P-P graf (l)
 Quantile plot (a), quantile-box plot (b), percentile plot (c), kernel estimation of probability density plot (d), histogram (e), mountain plot (f), half-sum plot (g), symmetry plot (h), differential quantile plot (i), notched box-and-whisker plot (j), Q-Q plot (k), P-P plot (l)

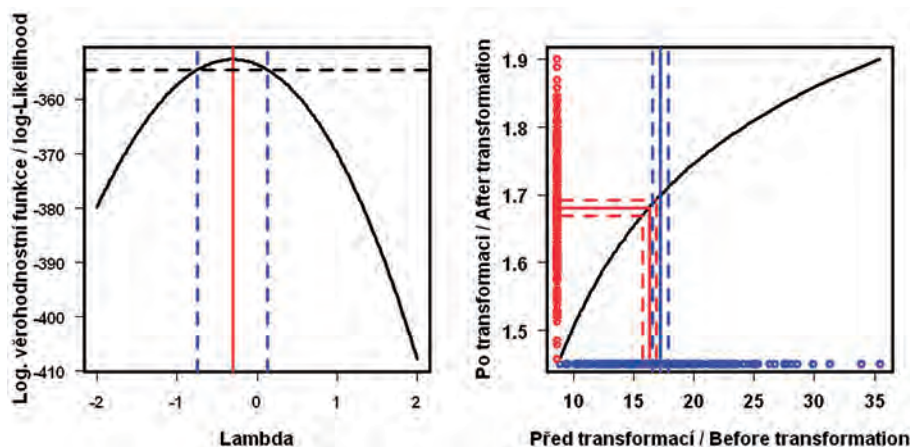
hodnostní funkce (obr. 4) rezultoval odhad exponentu $\lambda = -0,3052$, program ADSTAT použil optimalizovaný odhad exponentu z pěti rozličných numerických přístupů, který vedl k hodnotě $\lambda = -0,40$, směřující především k minimální šikmosti. Intervalový odhad parametru λ neobsahuje hodnotu +1 a transformaci lze proto považovat za oprávněnou. To, že konfidenční interval pokrývá hodnotu 0, by opravňovalo k provedení výpočetně podstatně jednodušší logaritmické transformace. Správnější hodnota odhadu polohy čili retransformovaného průměru 16,29 cm (dolní interval spolehlivosti 15,72 cm a horní interval 16,88 cm) je zde nižší než hodnota klasického odhadu čili aritmetického průměru 17,19 cm s konfidenčním intervalem 16,56 cm a 17,83 cm. Interval spolehlivosti retrans-

formovaného průměru je asymetrický, což by bylo zřetelnější při analýze menšího výběru, při velikosti našeho výběru ($n = 232$) leží bodový odhad téměř ve středu konfidenčního intervalu (obr. 4). Z obrázku 5 je patrné, že provedená transformace vedla ke zlepšení symetrie rozdělení. Odchylky od linearitu na obou koncích Q-Q grafu po transformaci dat nejsou způsobeny asymetrií, ale tzv. krátkými konci.



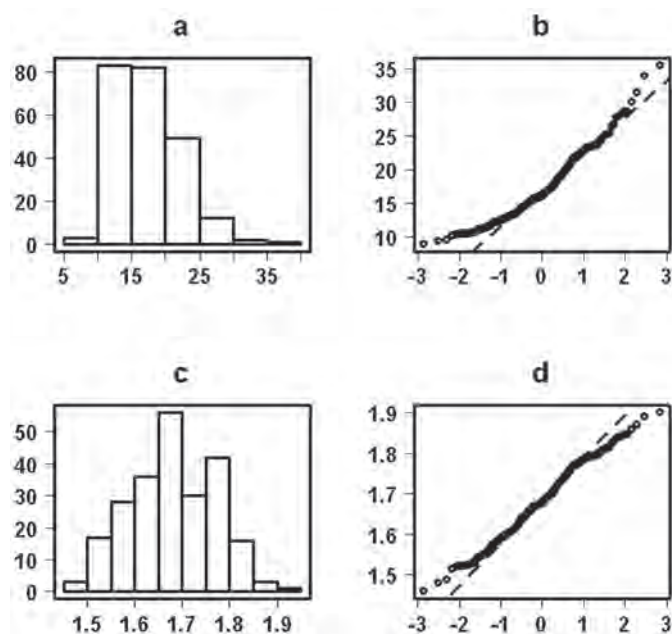
Obr. 3.

Histogram a P-P graf pro porovnání empirického rozdělení s rozdělením Weibullovým, gamma, logaritmicko-normálním a normálním
Histogram and P-P plot where the empirical distribution is compared to Weibull, gamma, log-normal and normal distribution



Obr. 4.

Graf logaritmu maximální věrohodnosti v Boxově-Coxově transformaci (vlevo) a graf transformace (vpravo) s původními (modré) a transformovanými daty (červené). Jsou zobrazeny bodové odhady původního a retransformovaného průměru (plné čáry) a 95% konfidenční intervaly (přerušované čáry)
The plot of logarithm of likelihood function for Box-Cox transformation (left) and plot with original (blue) and transformed data (red, right)
Mean of original data and mean of transformed data (solid lines) and 95% confidence intervals (dashed lines) are showed.



Obr. 5.

Histogram (a) a Q-Q graf (b) před transformací a histogram (c) a Q-Q graf (d) po Boxově-Coxově transformaci
Histogram (a) and Q-Q plot (b) before transformation and histogram (c) and Q-Q plot (d) after transformation

Analýza malého výběru

Klasický odhad střední hodnoty dat malého výběru z tabulky 4 vedl k hodnotě 771 gramů se spodní mezí 653 g a horní mezí 888 g a směrodatná odchylka byla 127,05. Medián rezultoval v hodnotu 760 g se spodní mezí 505 g a horní mezí 1 015 g a mediánová směrodatná odchylka byla 104,08. Hornovým postupem byl však nalezen rigoróznější odhad střední hodnoty 797 g se spodní mezí 709 g a horní mezí 885 g a s pivotovým rozpětím 122. Pro srovnání byla provedena i exponenciální transformace, která vedla k hodnotě retransformovaného průměru 788 g se spodní mezí 660 g a horní mezí 882 g.

ZÁVĚR

Počítačově orientovaná průzkumová analýza dat představuje užitečný nástroj k vyšetření statistických zvláštností v datech. Výsledky průzkumové analýzy jsou klíčové pro volbu další strategie vyhodnocení dat. Jak bylo demonstrováno na příkladu analýzy velkého výběru, je hodnota průměru i jeho intervalových odhadů pro nesymetrická rozdělení neko-rektní. Vhodným řešením byla Boxova-Coxova transformace, vedoucí ke zesymetričtění rozdělení výběru a následné vyčíslení hodnoty retrans-formovaného průměru.

Na příkladu analýzy malého výběru bylo ukázáno, že aplikace Hor-nova postupu poskytuje korektnější odhady parametrů polohy a rozpětí než klasické postupy požadující větší výběry. Pivotová polosuma a pivoto-vé rozpětí byly poměrně robustní i vůči odlehlým hodnotám ve výběru.

Tab. 1.

Hodnoty výčetních tlouštěk (v cm) naměřených kalibrovanou milimetrovou průměrkou na nevychované kontrolní ploše experimentu s výcho-vou smrku ztepilého (DUŠEK, SLODIČÁK 2009)

Values of diameter at breast height (in cm) which were measured by the calibrated calliper on the unthinned control plot of thinning experiment in Norway spruce stand (DUŠEK, SLODIČÁK 2009)

14,85	12,65	23,20	15,25	23,65	15,40	16,70	15,05
11,15	18,05	15,80	19,80	15,20	19,95	12,80	14,25
18,35	23,15	13,55	18,80	15,85	26,70	24,15	14,45
16,90	17,15	18,75	15,75	14,30	18,65	22,45	11,45
11,55	10,40	21,70	14,35	23,35	13,80	18,60	13,00
22,45	21,60	16,10	15,75	18,50	16,90	35,45	15,70
16,05	10,50	14,90	11,30	23,60	20,80	19,80	25,05
11,35	18,05	10,55	21,65	10,90	11,90	19,15	13,00
23,40	15,80	24,90	13,45	15,25	23,35	12,85	15,90
24,15	13,35	26,30	15,75	15,80	10,40	9,40	21,85
28,55	15,65	19,35	12,40	13,10	14,50	8,90	11,95
17,15	16,50	18,30	19,05	10,20	11,55	11,05	12,45
14,80	25,15	16,20	12,75	9,55	27,70	12,65	19,70
20,30	22,70	12,90	12,55	10,65	13,95	31,30	22,50
23,10	12,15	13,05	12,90	13,20	21,90	21,70	16,50
11,55	20,35	16,80	15,85	20,55	10,60	11,95	11,20
20,55	23,55	23,40	15,65	13,60	15,10	33,90	10,35
15,10	27,55	10,50	17,75	23,00	12,85	15,55	16,35
28,00	16,85	11,00	21,80	17,55	21,60	16,45	17,05
13,05	24,60	14,00	12,50	12,15	11,20	15,15	19,80
13,65	12,10	17,20	20,15	15,70	14,30	16,35	15,35
22,75	13,30	14,80	18,85	17,75	20,85	20,00	21,30
11,00	17,00	19,25	14,70	23,55	21,50	16,35	23,20
25,20	14,40	15,70	13,30	23,25	10,40	14,55	23,30
22,10	17,35	13,25	13,95	15,00	13,50	21,45	22,20
25,30	15,55	16,35	12,65	14,35	12,00	14,30	22,25
16,20	29,90	18,95	22,15	17,60	15,85	16,65	17,80
17,25	24,15	12,20	12,15	17,15	20,65	15,85	28,60
12,35	15,50	19,30	19,25	12,50	13,10	17,05	13,70

Tab. 2.

Testy základních předpokladů o výběru
Tests of basic assumptions about sample

Jarque Beraův test normality/Jarque-Bera test for normality	
ALM statistika/ALM statistics	28,51
Tabulkový kvantil pro $\alpha = 0,05$ /Critical value $\alpha = 0,05$	5,99
Pravděpodobnost/P-value	6,44E-7
Test homogeneity/Test for homogeneity	
Dolní mez/Lower bound	-3,09
Horní mez/Upper bound	36,97
Počet vybočujících dat/Number of outliers	0
von Neumannův test nezávislosti/Von Neumann test for randomness	
von Neumannova statistika/Von Neumann statistics	0,74
Tabulkový kvantil pro $\alpha = 0,05$ /Critical value $\alpha = 0,05$	1,97
Pravděpodobnost/P-value	0,46
Minimální velikost výběru/Minimal size of sample	
Pro 25% relativní chybu směrodatné odchylky/For 25% relative error of standard deviation	11
Pro 10% relativní chybu směrodatné odchylky/For 10% relative error of standard deviation	63
Pro 5% relativní chybu směrodatné odchylky/For 5% relative error of standard deviation	248
Pro 1% relativní chybu směrodatné odchylky/For 1% relative error of standard deviation	6187

Tab. 3.

Klasické a robustní odhady parametrů (vše v cm) pro příklad analýzy velkého výběru
Classical and robust estimation of parameters (all values in cm) for example of large sample

Klasické odhady parametrů/Classical parameters	
Průměr/Mean	17,19
95% LCL průměru/95% CI lower bound	16,56
95% UCL průměru/95% CI upper bound	17,83
Směrodatná odchylka/Standard deviation	4,93
Rozptyl/Variance	24,29
Šikmost/Skewness	0,81
Špičatost/Kurtosis	3,47
Robustní odhady parametrů/Robust parameters	
Medián/Median	16,08
95% LCL mediánu/95% CI lower bound	15,45
95% UCL mediánu/95% CI upper bound	16,70
Retransformovaný průměr/Retransformed mean	16,29
95% LCL/95% CI lower bound	15,72
95% UCL/95% CI upper bound	16,88

Tab. 4.

Hodnoty laboratorně stanovené (při 80 °C) hmotnosti sušiny (v gramech) holorganických horizontů (L + F + H) odebraných ze sedmi půdních sond v porostech smrku pichlavého (ŠPULÁK, DUŠEK 2009)

Values of dry mass weight (in grams, analyzed in laboratory at 80 °C) of holorganic horizons (L + F + H) which were sampled from seven soil pits in blue spruce stands (ŠPULÁK, DUŠEK 2009)

530	736	760	760	814	858	938
-----	-----	-----	-----	-----	-----	-----

Poděkování:

Publikace vznikla v rámci řešení výzkumného záměru Mze 000-2070203 „Stabilizace funkcí lesa v antropogenně narušených a měnících se podmínkách prostředí“ a MSM 0021627502.

LITERATURA

- DRÁPELA K., ZACH J. 1996. Biometrika, biostatistika – vybrané části. Skriptum MZLU Brno, 150 s.
- DRÁPELA K., ZACH J. 1999. Statistické metody I. Skriptum MZLU Brno, 135 s.
- DRÁPELA K. 2000. Statistické metody II. Skriptum MZLU Brno, 144 s.
- DUŠEK D., SLODIČÁK M. 2009. Struktura a statická stabilita porostů pod různým režimem výchovy na bývalé zemědělské půdě. Zprávy lesnického výzkumu, 54: 12-16.
- HENDL J. 2004. Přehled statistických metod zpracování dat. Praha, Portál: 583 s.
- JARQUE C. M., BERA A. K. 1987. A test for normality of observation and regression residuals. International Statistical Review, 55: 163-167.
- KOZAK A., KOZAK R., STAUDHAMMER CH., WATTS S. 2008. Introductory Probability and Statistics - Application for Forestry and Natural Sciences. Cambridge, Cambridge University Press: 408 s.
- LEPORSKÝ A. 1953. Statistické metody. Praha, SPN.
- MELOUN M., MILITKÝ J. 2004. Statistická analýza experimentálních dat. Praha, Academia: 953 s.
- PRODAN M. 1961. Forstliche Biometrie. München, BVL: 432 s.
- STONE B. 1963. Statistické metody v lesnictví. Praha, Státní zemědělské nakladatelství: 251 s.
- ŠMELKO Š., WOLF J. 1977. Štatistické metódy v lesníctve. Bratislava, Príroda: 330 s.
- ŠPULÁK O., DUŠEK D. 2009. Comparison of the impact of blue spruce and reed *Calamagrostis villosa* on forest soil chemical properties. Journal of Forest Science, 2009, 54 - v tisku
- VENABLES W., RIPLEY B. 1999. Modern Applied Statistics with S-PLUS. New York, Springer-Verlag: 501 s.
- URZUA M. 1996. On the correct use of omnibus tests for normality. Economics Letters, 53, s. 247-251.

COMPUTER-ASSISTED STATISTICAL ANALYSIS IN SILVICULTURE - I. UNIVARIATE DATA TREATMENT**SUMMARY**

Statistical data processing is a traditional discipline of forest research. Contemporary modern statistical data analyses undergo a rapid development due to the new potential of computer technology. Presented paper is focused on a detailed description of the modern interactive statistical analysis on PC in the frame of silvicultural data processing. Firstly, we paid attention to analysis of one-dimensional data. Application of these methods is demonstrated on two examples. One example is presented by a large sample ($n = 232$, values of diameter at breast height in cm) from the unthinned control plot of a thinning experiment in Norway spruce stand, tab. 1. As an example of small sample, data from a blue spruce experiment were used ($n = 7$, values of dry mass weight of holorganic horizons - surface layers - from seven soil pits, tab. 4).

Exploratory data analysis indicated an asymmetrical left-sided distribution of the large sample data. It was supported by a collection of plots and diagrams (figs. 1 – 3). Therefore, we expected incorrect values of a classical and robust estimation of parameters (tab. 3). Consequently, the Box-Cox transformation method was used (fig. 4). Transformation of the data from the large sample resulted to better parameters of distribution (fig. 5).

In case of small sample, we found, that results from the Horn procedure are correct in comparison with classical methods. Both, the pivot half-sum and pivot range were robust against outliers in a small sample.

In conclusion, we recommend both presented techniques to be applied in a silvicultural research.

Recenzováno

ADRESA AUTORA/CORRESPONDING AUTHOR:

Ing. David Dušek, Výzkumný ústav lesního hospodářství a myslivosti, v. v. i., Opočno
Na Olivě 550, 571 73, Opočno, Česká republika
tel.: 494 668 391 - 2; e-mail: dusek@vulhmop.cz