

INFORMAČNÍ BULLETIN



České statistické společnosti

Ročník 27, číslo 3, září 2016

METODY ANALÝZY EXTRÉMNÍCH HODNOT A JEJICH SOFTWAREVÁ IMPLEMENTACE

METHODS FOR EXTREME VALUE ANALYSIS AND THEIR SOFTWARE IMPLEMENTATION

Jan Holešovský¹, Michal Fusek²

Adresa: ¹Fakulta stavební, VUT v Brně, Veverí 95, 602 00 Brno,

²Fakulta elektrotechniky a komunikačních technologií, VUT v Brně,
Technická 8, 616 00 Brno

E-mail: ¹holesovsky.j@fce.vutbr.cz, ²fusekmi@feec.vutbr.cz

Abstrakt: Teorie extrémních hodnot hraje důležitou roli v celé řadě aplikačních oblastí. Cílem tohoto příspěvku je seznámit čtenáře se základy teorie extrémních hodnot a popsat některé přístupy k jejich analýze. Nejprve bude pozornost soustředěna na modely použitelné v případě nezávislých pozorování, a to na model blokových maxim a prahový model. Dále budou popsány vybrané metody pro analýzu stacionárních řad s krátkodobou závislostí, které v současné době nabývají na významu. V závěru bude představen autory vytvořený software v prostředí Matlab, který lze použít při řešení typických aplikačních problémů teorie extrémních hodnot. Implementovány jsou zejména vybrané odhady parametrů a parametrických funkcí, které jsou v praktických situacích nejčastěji používány. Ačkoliv řešení složitějších úloh obvykle vyžaduje užití řady dalších metod, tento software nabízí jednoduchý nástroj pro základní analýzu extrémních hodnot.

Klíčová slova: Rozdělení extrémních hodnot, prahový model, extrémální závislost, software EVDest.

Abstract: The theory of extreme value distributions plays an important role in many application areas. This paper recapitulates basics of the extreme value theory and introduces some approaches to the analysis of extremes. Two models for analysis of independent observations will be described, specifically the block maxima model and the threshold model. Moreover, selected methods for analysis of stationary series with short-time dependence which have become quite popular recently will be introduced. Finally, procedures for estimation of parameters and typical parametric functions were implemented in Matlab environment in the form of standalone software that can be used for an analysis of common extreme value problems. Although other advanced methods may need to be used for an analysis of more complex problems, this software can be used for a basic analysis of extreme values.

Keywords: Extreme value distribution, threshold model, extremal dependence, software EVDest.

1. Úvod

Teorie extrémních hodnot patří k významně se rozvíjejícím odvětvím matematické statistiky a používá se k odhadům různých řídce se vyskytujících událostí. Tradičně jsou přístupy teorie extrémních hodnot uplatňovány např. v hydrologii či ve finančnictví, ale v současné době se stále častěji vyskytují nová odvětví, kde lze s úspěchem tuto teorii použít. Zmíníme například studie analýzy pevnosti materiálu [21], předpovědi délky života [6], modelování výšky vln [5] či síly větru [11], analýzy termodynamiky zemětřesení [13] nebo práci zabývající se selháváním paměťových buněk [16].

V první části tohoto textu budou nejprve stručně představeny teoretické základy teorie extrémních hodnot (EV z angl. Extreme Value). Zájemce o detailnější studium celé teorie odkazujeme např. na knihy [4] či [6]. Dále budou v odstavci 3. shrnuty nejčastěji používané přístupy k EV odhadům pro případ nezávislých pozorování. Praktické požadavky jednotlivých aplikačních odvětví vedly k mohutnému rozvoji příslušných metod odhadů, většina z nich je ovšem zatížena jistou mírou subjektivního rozhodování. Převážně se jedná právě o problematiku spojenou s aproximativní nezávislostí sledovaných jevů, která se jeví jako velmi limitující. Z tohoto důvodu budou v odstavci 4. představeny metody EV odhadů pro závislá pozorování. Některé základní výsledky byly odvozeny již v knize [14], ale až v poslední době je tato tematika intenzivněji studována, čemuž nasvědčuje řada prací [2, 8, 10, 20].

Veškeré zde prezentované metody byly autory implementovány v prostředí Matlab do volně dostupného softwaru EVDest [7], který nabízí jednoduché použití široké škály přístupů k odhadům tzv. návratových úrovní, které budou popsány dále.

2. Rozdělení extrémních hodnot

Nechť je dán náhodný výběr $X_1, \dots, X_n$ z rozdělení s distribuční funkcí $F(x)$ a označme výběrové maximum $M_n = \max\{X_1, \dots, X_n\}$. Snadno se nahlédne, že limitním rozdělením M_n je degenerované rozdělení v $x^* = \sup\{x : F(x) < 1\}$, tj. pro $n \rightarrow \infty$ platí

$$F_n(x) := P(M_n \leq x) = [F(x)]^n \rightarrow G_{\text{deg}}(x), \quad (1)$$

kde $G_{\text{deg}}(x) = 0$ pro $x < x^*$ a $G_{\text{deg}}(x) = 1$ pro $x \geq x^*$. Stejně jako v případě centrální limitní věty pro výběrové průměry, kterou můžeme chápat jako

analogii základní věty EV teorie odvozené Fisherem a Tippettem, je nutné výběrová maxima řádně normalizovat. Fisher a Tippet [9] ukázali, že pokud existují takové konstanty $a_n > 0$, b_n , že $F_n(a_n x + b_n) \rightarrow G_\xi(x)$ pro $n \rightarrow \infty$, kde $G_\xi(x)$ je nedegenerovaná distribuční funkce, pak $G_\xi(x)$ je tvaru

$$G_\xi(x) = \exp \left\{ - [1 + \xi x]_+^{-1/\xi} \right\}, \quad (2)$$

kde $\xi \in \mathbb{R}$ a $x_+ = \max(x, 0)$. Rozdělení s distribuční funkcí $G_\xi(x)$ se nazývá *rozdělení extrémních hodnot* (zkráceně EV rozdělení). V případě $\xi = 0$ lze příslušnou distribuční funkci získat limitním přechodem ve tvaru $G_0(x) = \exp\{\exp^{-x}\}$. Častěji se však pracuje s tzv. *zobecněným rozdělením extrémních hodnot* (zkráceně GEV rozdělením z angl. Generalized Extreme Value) s distribuční funkcí ve tvaru

$$G(x) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\}, \quad (3)$$

kde $\mu \in \mathbb{R}$ je parametr polohy, $\sigma > 0$ je parametr měřítka a $\xi \in \mathbb{R}$ je parametr tvaru, který hraje klíčovou roli vzhledem k vlastnostem chvostů GEV rozdělení a na přesnost jeho odhadu jsou proto kladeny značné požadavky.

3. Modely pro odhady parametrických funkcí pro nezávislá pozorování

Nyní se zaměříme na dva přístupy užívané v analýze extrémních hodnot, mezi jejichž předpoklady patří nezávislost jednotlivých pozorování. Nejprve popíšeme metodu blokových maxim, která je založena na aproximaci pomocí GEV rozdělení. Tato metoda, jak dále ukážeme, má jisté nevýhody, které se snaží eliminovat druhý přístup, a to prahový model založený na zobecněném Paretově rozdělení. Souvislost mezi oběma přístupy a vhodnost jejich užití bude diskutována v následujícím odstavci. Předpoklad nezávislosti jednotlivých pozorování, který je klíčový pro oba přístupy, bývá v praktických situacích porušen, neboť pozorování nejčastěji získáme ve formě časových řad. Na úrovni extrémních hodnot bývá přítomna jistá „krátkodobá závislost“, tedy závislost významná pro časově (či jinak) blízká měření a často zanedbatelná pro vzdálená pozorování. Této tématice je blíže věnován odstavec 4.

3.1. Model blokových maxim

V předchozí části byla zavedena třída GEV rozdělení pomocí distribuční funkce $G(x)$. Uvažujme náhodný výběr $X_1, \dots, X_k$ dostatečně velkého roz-

sahu $k = mn$ z rozdělení s neznámou distribuční funkcí $F(x)$. Při užití modelu blokových maxim je daný datový soubor rozdělen na m disjunktních bloků rozsahu n , přičemž označme maxima vybraná z jednotlivých bloků jako $M_{n,i} = \max\{X_{(i-1)n+1}, \dots, X_{in}\}$, $i = 1, \dots, m$. Pro dostatečně velké n můžeme náhodný výběr $M_{n,1}, \dots, M_{n,m}$ aproximovat GEV rozdělením, které lze využít ke stanovení odhadů vybraných parametrických funkcí. V environmentálních vědách je pak n voleno převážně tak, aby bloky odpovídaly skupině měření provedených během jednoho roku. Jsou tak jednak eliminovány případné sezónní vlivy s roční periodicitou, které mohou být stále, i když v menší míře, přítomny. Především však tento postup také velmi často vede k redukci krátkodobé závislosti, předpokládáme-li, že tato nenastává na rozhraní jednotlivých bloků. V konkrétních aplikacích pak bývá tato skutečnost obvykle známa. Typicky při vyhodnocení extrémních teplot či srážek lze extrémní hodnoty očekávat dostatečně „uvnitř“ ročních bloků tak, aby příslušná bloková maxima byla dostatečně časově vzdálena. Odhady získané metodou blokových maxim jsou pak obecně považovány za dostatečně robustní vzhledem k porušení předpokladu nezávislosti výběru $X_1, \dots, X_k$.

Obvyklým požadavkem v aplikacích bývá získání odhadu tzv. *návratové úrovně* z_r , tj. hodnoty, která je v průměru překročena jednou za r pozorování. V bankovníctví tak bývá stanovován ekonomický ukazatel VaR (Value at Risk), který udává odhad nejhorší možné ztráty v daném období. V hydrologii se lze s návratovou úrovní běžně setkat např. jako s N -letou vodou, tedy průtokem vody, který se průměrně vyskytuje jednou za N let. Z pohledu statistiky je návratová úroveň z_r vlastně $(1 - r^{-1})$ kvantil modelového rozdělení, kde r je nějaké dostatečně velké číslo (obvykle $r \gg k$).

Ze vztahu (3) plyne, že úroveň návratu z_r je tvaru

$$z_r = \begin{cases} \mu - \frac{\sigma}{\xi} \left\{ 1 - [-\ln(1 - r^{-1})]^{-\xi} \right\} & \text{pro } \xi \neq 0, \\ \mu - \sigma \ln [-\ln(1 - r^{-1})] & \text{pro } \xi = 0, \end{cases} \quad (4)$$

přičemž její odhad $\widehat{z}_r$ lze získat nahrazením parametrů μ, σ, ξ ve vztahu (4) jejich (typicky maximálně věrohodnými) odhady $\widehat{\mu}, \widehat{\sigma}, \widehat{\xi}$. Na základě vlastností maximálně věrohodných (MV) odhadů lze tzv. delta metodou získat odhad rozptylu odhadu návratové úrovně (detaily viz [4]).

V situacích, kdy volba bloků není dána empirickými požadavky, je nutné nalézt kompromis mezi počtem bloků a jejich rozsahem. Malý počet bloků m může způsobit značnou variabilitu odhadnutých parametrů, zatímco velké m (při konstantním mn) vede ke špatné aproximaci rozdělení blokových maxim pomocí limitního GEV rozdělení a tím i k možnému vychýlení odhadů jeho parametrů. Hledání kompromisu mezi malým vychýlením a malým rozptylem

odhadů je tak základní charakteristikou metod EV teorie. Jedním z nejzávažnějších nedostatků modelu blokových maxim je ovšem značná redukce původního výběru. V daném bloku jsou totiž opomíjeny další extrémní hodnoty, které jsou ale menší než příslušná bloková maxima. Ačkoliv jsou obecně odhady získané metodou blokových maxim považovány za málo vychýlené, její použití nedovoluje identifikovat nerovnoměrný výskyt extrémních jevů. Tento problém se snaží eliminovat prahový model popsáný dále.

3.2. Prahový model

Při použití prahového modelu jsou uvažována pozorování, která překračují nějaký předem stanovený práh. Mějme náhodnou veličinu X s distribuční funkcí $F(x)$. Pickands ve své práci [18] ukázal, že je-li dána posloupnost prahů u_n rostoucí s n k bodu x^* , pak podmíněné rozdělení veličiny $Y = X - u_n$ za podmínky $X > u_n$ konverguje pro $n \rightarrow \infty$ k zobecněnému Paretovu rozdělení (GP rozdělení z angl. Generalized Pareto), tj.

$$\lim_{n \rightarrow \infty} P(X - u_n \leq x \mid X > u_n) = H(x), \quad (5)$$

kde $H(x)$ je distribuční funkce GP rozdělení ve tvaru

$$H(x) = 1 - \left(1 + \frac{\xi x}{\sigma_u}\right)_+^{-1/\xi}. \quad (6)$$

V praxi je práh u volen pevně jako nějaká dostatečně vysoká hodnota. Ve vztahu k limitnímu rozdělení výběrových maxim (3) má GP rozdělení stejný parametr tvaru $\xi \in \mathbb{R}$ a pro parametr měřítka $\sigma_u > 0$ GP rozdělení lze odvodit, že $\sigma_u = \sigma + \xi(u - \mu)$, tj. σ_u lineárně závisí na prahu u .

Kritickým bodem uvedeného přístupu je stanovení prahové hodnoty u . Příliš nízká volba prahu má za následek špatné asymptotické vlastnosti prahového modelu a může vést ke značnému vychýlení odhadů parametrů. Naopak vysoké hodnoty u zaručují lepší aproximaci pomocí limitního GP rozdělení, ale vedou k malým počtům nadprahových hodnot a tím k výraznému nárůstu variability odhadů. Ačkoli je této problematice věnována celá řada publikací [1, 3, 15, 17, 19], volba vhodného prahu není dána jednoznačně a nelze tak obecně upřednostnit některou z publikovaných technik. Zde zmíníme dvě často používané grafické metody, které jsou taktéž implementovány v softwaru EVDest. Označme dále $X_{(1)}, \dots, X_{(n)}$ uspořádaný náhodný výběr takový, že $X_{(1)} \leq \dots \leq X_{(n)}$.

První metoda je založena na tzv. Mean Residual Life (MRL) plotu a spočívá ve vynesení závislosti

$$\left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(n-i+1)} - u) \right) \quad \text{pro } x_{(n-i+1)} > u, \quad (7)$$

kde $x_{(i)}$ označuje realizaci náhodné veličiny $X_{(i)}$, $i = 1, \dots, n$, a n_u je počet pozorování, která překročí práh u . Z vlastností limitního GP rozdělení lze odvodit, že $E(X - u \mid X > u) = \sigma_u / (1 - \xi)$, tedy střední hodnota veličiny $Y = X - u$ za podmínky $X > u$ je lineární funkcí prahu u . Vhodná prahová hodnota u_0 se proto v MRL plotu zvolí tak, aby znázorněná závislost byla přibližně lineární pro $u > u_0$. Uvedená střední hodnota existuje pro $\xi < 1$.

Druhá metoda je založena na posouzení stability odhadů $\hat{\xi} = \hat{\xi}(u)$ a $\hat{\sigma}_u = \hat{\sigma}_u(u)$ parametrů GP rozdělení v závislosti na volbě prahu u . Jelikož parametr tvaru ξ GP rozdělení odpovídá parametru tvaru GEV rozdělení, hledáme takovou hodnotu u_0 , pro kterou je $\hat{\xi}$ přibližně konstantní pro $u > u_0$. Parametr měřítka σ_u GP rozdělení je lineární funkcí u při konstantních parametrech GEV rozdělení, a proto hledáme u_0 jako takovou hodnotu, pro niž je $\hat{\sigma}_u$ přibližně lineární pro $u > u_0$.

Jakmile je vybrána vhodná prahová hodnota, lze přistoupit ke statistické inferenci pro prahový model a použít GP rozdělení ke stanovení parametru tvaru ξ a dalších parametrických funkcí. Zde bude opět pozornost věnována odhadu návratové úrovně z_r , která je průměrně překročena jednou za r pozorování a je tvaru

$$z_r = \begin{cases} u + \frac{\sigma_u}{\xi} [(r\lambda_u)^\xi - 1] & \text{pro } \xi \neq 0, \\ u + \sigma_u \ln(r\lambda_u) & \text{pro } \xi = 0. \end{cases} \quad (8)$$

Odhad $\hat{z}_r$ dostaneme nahrazením parametrů σ_u , ξ , λ_u ve vztahu (8) jejich odhady $\hat{\sigma}_u$, $\hat{\xi}$, $\hat{\lambda}_u$. Odhady parametrů σ_u a ξ GP rozdělení lze určit opět metodou maximální věrohodnosti. Parametr λ_u se odhadne jako relativní četnost pozorování nad prahem u , tj. $\hat{\lambda}_u = n_u/n$. Jelikož počet překročení prahu u má binomické rozdělení $\text{Bi}(n, \lambda_u)$, je $\hat{\lambda}_u$ MV odhadem parametru λ_u . Odhad rozptylu odhadu úrovně návratu je opět možné získat delta metodou (viz [4]).

V praxi se častěji používá tzv. N -letá úroveň návratu z_N , což je hodnota měřené veličiny, která je ve výběru překročena v průměru jednou za N let. Při jejím výpočtu se vychází ze vztahu (8) s tím, že $r = Nn_y$, kde n_y je počet pozorování za rok.

Prahová metoda překonává některé nedostatky metody blokových maxim, zejména zpravidla nedochází k tak výrazné redukci původní pozorované řady. Z tohoto důvodu bývá prahový model upřednostňován před modelem blokových maxim. Na druhou stranu, pokud je porušen předpoklad nezávislosti původního náhodného výběru, je nutné tuto metodu kombinovat s dalšími technikami tak, aby byly získány alespoň přibližně nezávislé náhodné veličiny. Konkrétním případem je dříve diskutovaná krátkodobá závislost, jejíž přítomnost vede ke *shlukování* extrémních hodnot nad prahem u . Uvažujeme-li pak v prahovém modelu pozorování ve formě časové řady s krátkodobou závislostí, bývá buď nutné nejdříve z původního datového souboru vyvzorkovat řadu přibližně nezávislých pozorování, nebo tuto závislost jistým způsobem odhadnout. Oba tyto postupy budou popsány v následujícím odstavci.

4. Metody pro odhad parametrických funkcí pro stacionární řady

Použití výše uvedených metod analýzy extrémních hodnot je limitováno předpokladem nezávislosti jednotlivých pozorování ve výběru. Ve většině aplikačních úloh EV teorie představuje tento předpoklad nutnost použití dodatečných aproximačních technik. Praktické úlohy se však často vyznačují jistým společným specifikem, kterým je pouze jistá krátkodobá závislost jednotlivých náhodných veličin. Stacionární řady představují přirozené zobecnění EV teorie, jehož smyslem je popsat náhodné veličiny, které mohou být vzájemně závislé, avšak jejichž vlastnosti jsou homogenní v čase, zejména mají tožná marginální rozdělení. Řada teoretických výsledků pro stacionární řady je známa (viz [4, 14]), ovšem tyto dosud nebyly v plné míře aplikačně rozvinuty. Krátkodobou závislost lze popsat různými způsoby, nejčastěji je však uvažována v podobě následující Leadbetterovy [14] podmínky:

Řekneme, že stacionární řada $X_1, X_2, \dots$ splňuje $D(u_n)$ podmínku, jestliže pro všechny indexy $i_1 < \dots < i_p < j_1 < \dots < j_q$ takové, že $j_1 - i_p > l$, platí

$$|P(X_{i_1} \leq u_n, \dots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \dots, X_{j_q} \leq u_n) - P(X_{i_1} \leq u_n, \dots, X_{i_p} \leq u_n)P(X_{j_1} \leq u_n, \dots, X_{j_q} \leq u_n)| \leq \alpha(n, l),$$

kde $\alpha(n, l) \rightarrow 0$ pro nějakou posloupnost $\{l_n\}$ takovou, že $l_n/n \rightarrow 0$ pro $n \rightarrow \infty$. Obecně požadujeme, aby $D(u_n)$ podmínka byla splněna jen pro nějakou specifickou posloupnost prahů u_n rostoucí s n . Pro posloupnost nezávislých náhodných veličin je zřejmě $\alpha(n, l)$ identicky nulová funkce.

Jeden ze základních výsledků EV teorie říká, že je-li $X_1, X_2, \dots$ stacionární řada splňující $D(u_n)$ podmínku, pak rozdělení řádně normalizovaných

výběrových maxim $M_n = \max\{X_1, \dots, X_n\}$ opět konverguje ke GEV rozdělení.

Označme dále $X_1^*, X_2^*, \dots$ posloupnost nezávislých náhodných veličin se stejným marginálním rozdělením jako $X_1, X_2, \dots$ a $M_n^* = \max\{X_1^*, \dots, X_n^*\}$. Pak za určitých podmínek regularity [14] platí, že existují posloupnosti $a_n > 0$, b_n takové, že

$$P((M_n^* - b_n)/a_n \leq x) \rightarrow G(x) \quad (9)$$

právě tehdy, když

$$P((M_n - b_n)/a_n \leq x) \rightarrow G_\theta(x) = [G(x)]^\theta, \quad (10)$$

kde $G(x)$ je nede degenerovaná distribuční funkce a $0 < \theta \leq 1$. Parametr θ se nazývá extrémální index a představuje míru extrémální závislosti. Pro $\theta \rightarrow 0$ dochází ke shlukování extrémních hodnot.

Základní charakterizace extrémálního indexu vychází z následujícího vyjádření [12]:

$$\theta^{-1} = \lim_{n \rightarrow \infty} E \left[\sum_{i=1}^{r_n} I(X_i > u_n) \mid M_{r_n} > u_n \right], \quad (11)$$

kde $I(A)$ je charakteristická funkce jevu A a pro $n \rightarrow \infty$ je $r_n \rightarrow \infty$ a $r_n/n \rightarrow 0$. Tedy θ^{-1} představuje limitní střední hodnotu velikosti shluků nad prahem u_n . Tato interpretace iniciovala řadu odhadů extrémálního indexu, z nichž některé představíme dále. Jiné možné charakterizace θ lze nalézt např. v [2].

Metody analýzy extrémních hodnot stacionárních řad jsou v principu založeny buď na odhadu extrémálního indexu, nebo na technikách, které odstraňují vliv závislosti a tím i nutnost odhadu θ . Poslední zmiňované budou popsány v následujícím odstavci. Správný odhad θ však dovoluje do statistického zpracování zahrnout všechna pozorování nad daným prahem, což umožňuje daleko přesnější analýzu pro stacionární řady než přístupy založené na předpokladu nezávislosti jednotlivých pozorování.

4.1. Declusterování a jednoduchý odhad extrémálního indexu

Mezi přístupy, které se snaží eliminovat vliv závislosti (tzv. *declusterování* z angl. *declustering*), patří metoda blokových maxim popsaná v odstavci 3.1. Nahradíme-li při zavedení blokových maxim náhodný výběr $X_1, \dots, X_n$ stacionární řadou splňující $D(u_n)$ podmínku, pak pro dostatečně velký rozsah

jednotlivých bloků jsou bloková maxima $M_{n,1}, \dots, M_{n,m}$ přibližně nezávislé náhodné veličiny a tedy GEV rozdělení zůstává vhodnou aproximací jejich rozdělení. Takový způsob eliminace závislosti se nazývá *block declustering*.

Při použití prahového modelu je potřeba použít dodatečné techniky declusterování, kdy jsou pozorování nad prahem u rozdělena na přibližně nezávislé extrémní události, tj. disjunktní podmnožiny nadprahových hodnot. Z těchto událostí dále uvažujeme jen jejich maximální hodnoty. Tento způsob declusterování je v literatuře označován jako tzv. *runs declustering*, kde velikost runs parametru („podtečení“) udává počet po sobě jdoucích podprahových hodnot, kterými mají být dvě události odděleny. Volba tohoto parametru vychází z empirických znalostí a může mít značný vliv na odhady parametrů GP rozdělení. Spolu s volbou prahové hodnoty u je tak častým důvodem kritiky tohoto modelu.

Použití declusterovacích technik vždy vede k redukci původní pozorované řady, což je často příliš omezující. Na základě interpretace extrémálního indexu dle (11) lze zavést odhad následujícího typu

$$\hat{\theta} = \frac{C(u_n)}{N(u_n)}, \quad (12)$$

kde $N(u_n)$ je počet hodnot nad prahem u_n a $C(u_n)$ je počet shluků. Pak je možné definovat jednoduchý odhad θ [2] ve tvaru

$$\hat{\theta}_1(u) := \frac{\sum_{i=1}^{n-1} I(X_{i+1} \leq u < X_i)}{\sum_{i=1}^n I(X_i > u)}. \quad (13)$$

Počet shluků je tak určen jako počet překročení prahu „směrem dolů“, tedy dva shluky jsou od sebe separovány pomocí alespoň jedné podprahové hodnoty. Toto odpovídá situaci, kdy je declusterovací runs parametr v modelu nezávislých pozorování roven jedné. Odhad (13) lze snadno zobecnit i pro separaci shluků pomocí $l \in \mathbb{N}$ podprahových hodnot, dostaneme tak odhad

$$\hat{\theta}_2(u, l) := \frac{\sum_{i=1}^{n-l} I(X_i > u, X_{i+1} \leq u, \dots, X_{i+l} \leq u)}{\sum_{i=1}^n I(X_i > u)}. \quad (14)$$

4.2. Maximové odhady extrémálního indexu

Jiné odhady extrémálního indexu využívají vztahů mezi limitními rozděleními (9) s parametry (μ, σ, ξ) a (10) s parametry $(\mu_\theta, \sigma_\theta, \xi_\theta)$. Snadno lze ukázat, že platí

$$\mu_\theta = \mu + \frac{\sigma}{\xi} (\theta^\xi - 1), \quad \sigma_\theta = \sigma \theta^\xi, \quad \xi_\theta = \xi. \quad (15)$$

Gomes [10] tak navrhla následující metodu odhadu extrémálního indexu. Pro danou stacionární řadu $X_1, \dots, X_{mn}$ označme $X_1^*, \dots, X_{mn}^*$ řadu, která vznikne z původní řady nějakým jejím náhodným přerovnáním. Tato nově vzniklá řada slouží jako aproximace nezávislé řady se stejným marginálním rozdělením. Dále použijeme metodu blokových maxim, tj. máme $M_{n,i} = \max\{X_{(i-1)n+1}, \dots, X_{in}\}$ a $M_{n,i}^* = \max\{X_{(i-1)n+1}^*, \dots, X_{in}^*\}$, $i = 1, \dots, m$, kde m je počet bloků a n je jejich rozsah. Výběr $M_{n,1}, \dots, M_{n,m}$ aproximujeme GEV rozdělením s parametry $(\mu_\theta, \sigma_\theta, \xi_\theta)$ a výběr $M_{n,1}^*, \dots, M_{n,m}^*$ rozdělením s parametry (μ, σ, ξ) .

S využitím vztahů (15) dostáváme odhad extrémálního indexu ve tvaru

$$\hat{\theta}_G(m, n) = (\hat{\sigma}/\hat{\sigma}_\theta)^{-1/\bar{\xi}}, \quad (16)$$

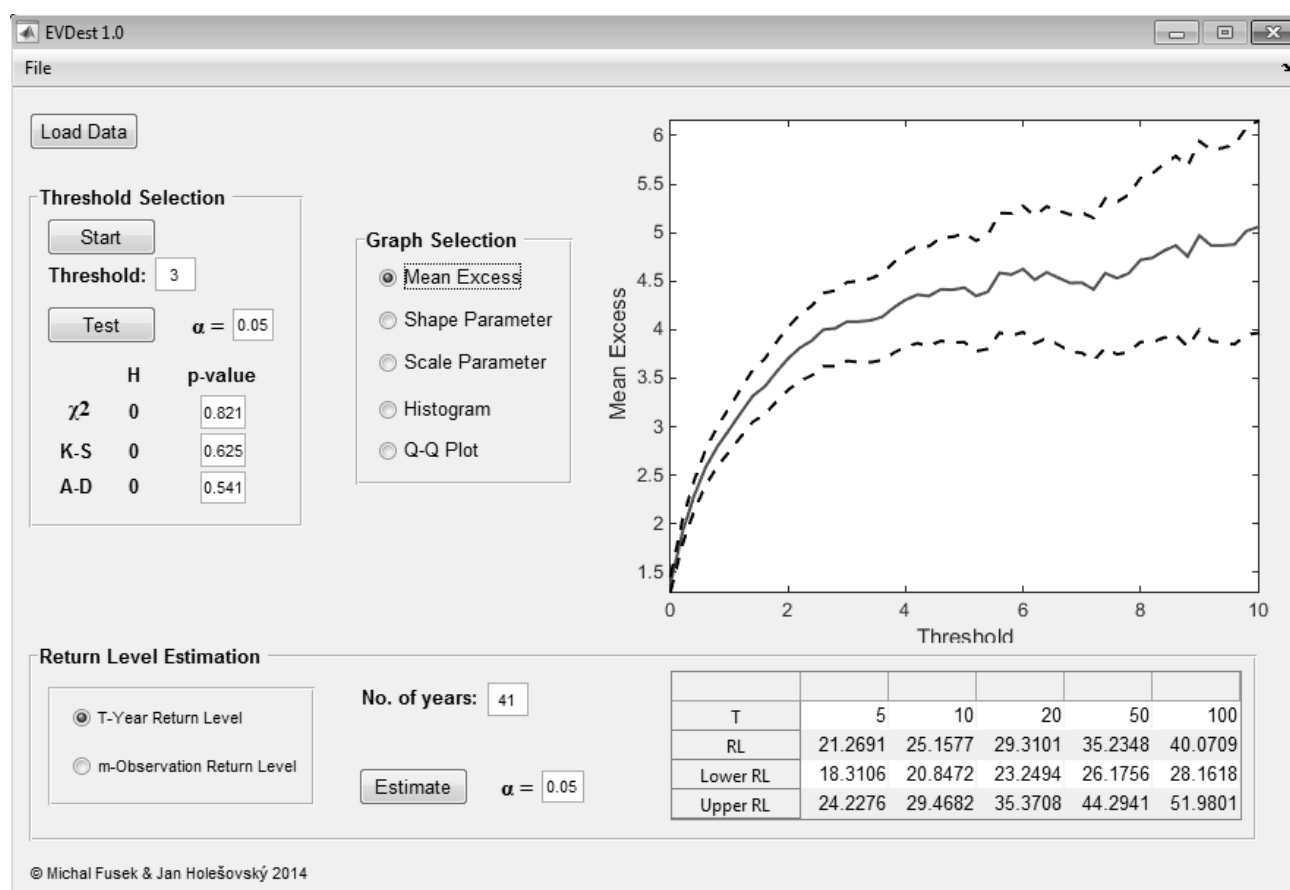
kde $\bar{\xi} = (\hat{\sigma} - \hat{\sigma}_\theta)/(\hat{\mu} - \hat{\mu}_\theta)$. Jakmile je odhadnut extrémální index, je možné stanovit hledané odhady parametrických funkcí s využitím všech nadprahových hodnot. Návratovou úroveň z_r lze získat jako $(1 - r^{-1})$ kvantil limitního rozdělení (10) ve tvaru

$$z_r = u + \frac{\sigma}{\xi} \left[\left(\lambda_u^{-1} \left\{ 1 - [1 - r^{-1}]^{\theta^{-1}} \right\} \right)^{-\xi} - 1 \right]. \quad (17)$$

Odhad $\hat{z}_r$ lze určit nahrazením parametrů $\theta, \lambda_u, \sigma_u, \xi$ ve vztahu (17) jejich MV odhady $\hat{\theta}, \hat{\lambda}_u, \hat{\sigma}_u, \hat{\xi}$. Odhad rozptylu odhadu úrovně návratu je opět možné získat delta metodou.

5. Softwarová implementace

Výše popsané metody byly implementovány do softwaru EVDest (viz obr. 1 a [7]). Program je rozdělen na analýzu nezávislých a závislých výběrů. V případě nezávislých pozorování je implementována metoda blokových maxim a prahový model včetně nástrojů pro volbu optimálního prahu. V případě závislých pozorování je implementován prahový model s možností volby parametru declusterování či odhadu extrémálního indexu pomocí Gomes [10]. U všech modelů je možné vykreslit Q-Q plot a histogram proložený hustotou příslušného rozdělení (GEV nebo GP). Aby bylo možné lépe posoudit vhodnost daného modelu, grafické metody jsou doplněny testy dobré shody (Pearsonův χ^2 test, Kolmogorovův-Smirnovův test, Andersonův-Darlingův test). Představený software tak nabízí základní nástroje pro analýzu extrémních hodnot včetně jednoduché možnosti validace výsledků. Přínosem EVDest je pak především jednoduchost jeho použití bez nutnosti hlubších znalostí v oblasti EV teorie. V neposlední řadě program obsahuje metody odhadu návratových úrovní, které patří k nejžádanějším výstupům v inženýrské praxi.



Obrázek 1: Software EVDest

6. Závěr

Tento příspěvek byl věnován problematice teorie extrémních hodnot a metodám odhadu kvantilů příslušných extrémálních rozdělí. Byly popsány nejčastěji užívané metody pro analýzu nezávislých výběrů, jejich výhody a nevýhody, avšak také jejich robustnost vzhledem k v praxi se často vyskytujícím porušením předpokladu nezávislosti. Právě vyhodnocení stacionárních řad s krátkodobou závislostí patří k častým úlohám EV teorie a zároveň ke stále se rozvíjejícím oblastem statistiky, přičemž její uplatnění v praktických situacích může vést k významnému zpřesnění odhadů extrémních jevů. Závěrem byl představen autory vytvořený software v prostředí Matlab, který obsahuje výše popsané metody. Význam tohoto softwaru spočívá především v jednoduchosti jeho použití pro řešení základních inženýrských úloh, zejména pak pro získání odhadů návratových úrovní tolik potřebných v praxi. Pokročilá analýza extrémních hodnot často vyžaduje použití dalších specifických technik, které tento software sice nenabízí, na druhou stranu však užití pokročilých postupů obvykle předpokládá hlubší studium EV teorie.

Poděkování

Příspěvek vznikl jako součást řešení projektu č. FAST-S-16-3385 realizovaného na VUT v Brně. Autoři by také chtěli poděkovat editorovi a recenzentům za náměty a komentáře, které vedly ke zkvalitnění tohoto příspěvku.

Literatura

- [1] Adamowski, K. (2000): Regional Analysis of Annual Maximum and Partial Duration Flood Data by Nonparametric and L -Moments. In *Journal of Hydrology* 229(3–4), pp. 219–231.
- [2] Ancona-Navarrete, M., Tawn, J. (2000): A Comparison of Methods for Estimating the Extremal Index. In *Extremes* 3(1), pp. 5–38.
- [3] Ben-Zvi, A. (1994): Fit of probability distributions to upper sub-samples of partial duration series. In *Stochastic and Statistical Methods in Hydrology and Environmental Engineering, Extreme Values: Floods and Droughts* 1 (K. W. Hipel, ed.). Kluwer, Dordrecht, NL, pp. 95–107.
- [4] Coles, S. (2001): *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, Berlin, Heidelberg.
- [5] Dawson, T. H. (2000): Maximum wave crest in heavy seas. In *Journal of Offshore Mechanics and Arctic Engineering* 122(3), pp. 222–224.
- [6] De Haan, L., Ferreira, A. (2006): *Extreme Value Theory: An Introduction*. Springer, New York.
- [7] Fusek, M., Holešovský, J. (2014): EVDest. Available online at: <http://www.umat.feec.vutbr.cz/software/evdest.html>.
- [8] Ferro, C. A. T., Segers, J. (2003): Inference for Clusters of Extreme Values. In *J. R. Statist. Soc.* 65, část 2, pp. 545–556.
- [9] Fisher, R. A., Tippett, L. H. C. (1928): Limiting Forms of the Frequency Distribution of the Largest or Smallest Mmembers of a Sample. In *Proceedings of Cambridge Philosophical Society* 24, pp. 180–190.
- [10] Gomes, M. I. (1993): On the Estimation of Parameters of Rare Events in Environmental Time Series. In *Statistics for the Environment 2: Water Related Issues*, Barnett and Turkman.
- [11] Harris, R. I. (2001): The accuracy of design values predicted from extreme value analysis. In *Journal of Wing Engineering and Industrial Aerodynamics* 89(2), pp. 153–164.
- [12] Hsing, T., Husler, J., Leadbetter, M. R. (1988): On the exceedance point process for a stationary sequence. In *Probability Theory and Related Fields* 78, pp. 97–112.

- [13] Lavenda, B., Cipollone, E. (2000): Extreme value statistics and thermodynamics of earthquakes: aftershock sequences. In *Annals of Geophysics* 43(5), pp. 967–982.
- [14] Leadbetter, M. R., Lindgren, G., Rootzén, H. (1983): *Extremes and Related Properties of Random Sequences and Series*, Springer-Verlag, London.
- [15] Madsen, H., Mikkelsen, P. S., Rosbjerg, D., Harremoës, P. (2002): Regional Estimation of Rainfall Intensity-Duration-Frequency Curves Using Generalized Least Squares Regressions of Partial Duration Series. In *Water Resources Research* 38(11), pp. 21-1 – 21-11.
- [16] McNulty, P. J., Scheick, L. Z., Roth, D. R., Davis, M. G., Tortora, M. R. S. (2000): First failure predictions for EPROMs of the type flown on the MPTB satellite. In *IEEE Transactions on Nuclear Science* 47(6), pp. 2237–2243.
- [17] Northrop, P. J., Coleman, C. L. (2014): Improved threshold diagnostic plots for extreme value analyses. In *Extremes* 17(2), pp. 289–303.
- [18] Pickands, J. (1975): Statistical inference using extreme order statistics. In *Ann. Statist.* 3, pp. 119–131.
- [19] Scarrott, C., MacDonald, A. (2012): A review of extreme value threshold estimation and uncertainty quantification. In *Revstat Statistical Journal* 10(1), pp. 33–60.
- [20] Süveges, M. (2007): Likelihood Estimation of the Extremal Index. In *Extremes* 10, pp. 41–55.
- [21] Tryon, R. G., Cruse, T. A. (2000): Probabilistic mesomechanics for high cycle fatigue life prediction. In *Journal of Engineering Materials and Technology – Transactions of the AMSE*, 122, pp. 209–214.

ÚROVEŇ STATISTICKÉ GRAMOTNOSTI STUDENTŮ UČITELSTVÍ BIOLOGIE

LEVEL OF STATISTICAL LITERACY OF STUDENTS OF TEACHING BIOLOGY

Aneta Hybšová

Adresa: Katedra biologie a environmentálních studií, Pedagogická fakulta Univerzity Karlovy v Praze, Magdalény Rettigové 47/4, 116 39 Praha 1 – Nové Město

E-mail: anetahybsova@gmail.com

Abstrakt: S rozvojem vědy a výzkumu roste význam statistické gramotnosti. Především učitelé biologie, neboť právě přesahy mezi biologií a matematikou ukazují žákům praktické využití matematiky a zároveň vědeckost jinak dogmaticky vykládaných teorií a pojmů v biologii. Míru statistické gramotnosti vysokoškolských studentů oboru učitelství biologie určuje výuka statistiky na základních a středních školách a dále pak samotná výuka na vysokých školách. Cílem tohoto příspěvku je popsat míru statistické gramotnosti studentů učitelství biologie a porovnat ji se studenty dalších přírodovědných oborů. Hlavním zjištěním je skutečnost, že studenti učitelství biologie neznají a neumí použít většinu statistických pojmů, které jsou uvedeny v rámcových vzdělávacích programech pro základní a střední školy.

Klíčová slova: statistická gramotnost, RVP, učitel biologie.

Abstract: With the spreading of science and research, the also grows the meaning of statistical literacy. Teachers of biology, or the intersection between biology and mathematics show their students the practical applications of mathematics while also focusing on scientific facts and theories in biology. The level of statistical literacy of students of teaching biology determines the way of teaching statistics at primary and secondary schools. The goal of this work is to compare and contrast the statistical literacy of students studying biology and other subject areas. The main finding is that student studying biology are not familiar with, and therefore unable to utilize, much statistical terminology, event those which are part of the learnings of programs in elementary schools and high schools.

Keywords: statistical literacy, framework education programme, biology teacher.

1. Statistická gramotnost a její měření

Pojem statistická gramotnost je v zahraničí rozvíjen už po více než dvě desetiletí. Příkladem rozvoje statistické gramotnosti je například projekt ISPL¹ nebo konference Turning data into knowledge² zaměřená zejména na statistickou gramotnost a její rozvoj. Výstižně tento pojem definovala Wallman [12] jako schopnost vytvořit, porozumět a kriticky zhodnotit statistické výsledky. Tato schopnost je v dnešní době velice významná [2, 3] v nejrozličnějších odvětvích. V současné době začíná být výuka statistiky implementována nikoliv pouze do matematicky orientovaných vysokoškolských studijních oborů, ale vyučuje se v modifikované formě i v oborech nematematicky zaměřených [3, 9]. Hybšová [5] popsala tři druhy statistické gramotnosti pro učitele biologie:

1.1. Běžná občanská znalost

Každý občan by měl rozumět socioekonomické situaci ve své zemi. Informace o politice, ekonomice a sociálních záležitostech jsou prezentovány v médiích pomocí čísel, například průměrná mzda, odhad míry inflace, atd. Porozumění těmto údajům by se měl žák a student naučit v průběhu základní a střední školy, neboť se jedná zejména o jednoduché popisné statistiky jako je průměr, modus a medián, popřípadě míry variability.

1.2. Didaktická znalost

Učitelé biologie skrze nejrozličnější pokusy vysvětlují studentům přírodní zákony. V průběhu tohoto ověřování učitel získává data, která by měla být analyzována a interpretována. Tento proces by měl být žákům a studentům ukazován zejména proto, aby si uvědomili možnosti využití matematiky na praktických příkladech, a také proto, aby věděli, že věda má nějaký reálný základ a nevzniká na základě dogmat. Spojení výzkumných metod a praxe učitelů popisuje Garfield a Ben-Zvi [2].

1.3. Výzkumná znalost

Moderní učitel biologie by měl být také schopný porozumět nejnovějším trendům v oboru didaktiky biologie i biologie samotné. Ve vědeckých článcích a odborných publikacích se používá základní statistická terminologie.

¹Podrobnosti o projektu lze nalézt zde: <http://iase-web.org/islp/>

²Podrobnosti o konferenci zabývající se statistickou gramotností lze nalézt zde: <http://www.statisteduc.ie.ulisboa.pt/>

Zejména z toho důvodu je nezbytné vzdělávat studenty jako budoucí vědce a podpořit je v používání výzkumných metod. Studenti učitelství biologie by měli být schopni se ptát a pomocí sběru a analýzy dat hledat odpovědi na položené otázky. Tento postup využívají studenti učitelství biologie také při psaní závěrečných prací. Tyto znalosti a schopnosti by měly být vyučovány na vysoké škole, neboť kvalita závěrečných prací nevypovídá pouze o studentech, ale i o kvalitě celé vysoké školy. [5]

Měření statistické gramotnosti se velice různí. Např. Watson a Kelly [13] z University of Tasmania ve své studii zkoumaly u žáků základních škol vývoj porozumění třem základním pojmům: vzorek, náhodný a variabilita. V průzkumu bylo dotázáno celkem 738 žáků, aby definovali vzorek a uvedli jeho příklad. Žáci 7. a 9. tříd (379 žáků) byli navíc dotazováni také na slova náhodný a variabilita. Z odpovědí na otázky vyvodily vědkyně závěry o úrovni porozumění těmito pojmům obecně a zdokumentovaly rozdíly mezi ročníky. Na tento výzkum navázala Kvaszová [7], která se na stejné pojmy zeptala 107 studentů soukromé vysoké školy s ekonomickým zaměřením. Jak sama uvádí „... tito studenti tedy prošli kurzem statistiky na základní škole a v jisté, ovšem hodně odlišné podobě, absolvovali také kurz kombinatoriky, pravděpodobnosti a statistiky na škole střední.“ [7] Ve své studii dochází k závěru, že 20–50 % (podle typu otázky) studentů se s těmito pojmy nikdy důkladně neseznámili a nevytvořili si vlastní představu. Tyto pojmy pak používají čistě formálně, bez konkrétního obsahu.

Statistická gramotnost se stejně jako ostatní gramotnosti rozvíjí po celou dobu života. Právě na základní a střední škole, kdy se myšlení žáků formuje, je významné učit žáky pracovat s informacemi a věrohodností jejich zdroje. Hybšová [6] zkoumala statistický obsah v rámcových vzdělávacích programech pro základní školy [8] a gymnázia [1], za účelem zjistit, co by měli studenti nastupující na vysoké školy umět a jaké kompetence v oblasti statistické gramotnosti by měli mít. Podle Rámcových vzdělávacích programů by studenti po absolvování základní a střední školy měli znát a umět použít pojmy shrnuté v tab. 1.

Nabízí se však otázka, zda při nástupu na vysoké školy tyto pojmy znají a zda je umějí použít, neboť při konstrukci sylabu základů statistiky je nutné vědět, jaká je vstupní úroveň znalostí studentů. Cílem tohoto článku je analyzovat a popsat, zda studenti učitelství biologie tyto pojmy znají a umějí je použít.

2. Výzkumné otázky

V rámci studované problematiky byly formulovány tyto výzkumné otázky:

Tabulka 1: Pojmy, které si podle RVP měli studenti osvojit na základních a středních školách ve vzdělávací oblasti Matematika a její aplikace

Na základní škole	Na gymnáziu
Graf	Náhodný jev
Diagram	Pravděpodobnost
Tabulka	Statistický soubor
Statistický soubor	Charakteristiky statistického souboru
Kvantitativní znak	Aritmetický průměr
Kvalitativní znak	Vážený aritmetický průměr
Četnost	Modus
Modus	Medián
Medián	Percentil
Aritmetický průměr	Kvartil
	Směrodatná odchylka
	Mezikvartilová odchylka

1. Znájí studenti učitelství biologie statistické pojmy, které jsou obsažené v RVP pro základní školy a gymnázia?
2. Umí studenti učitelství biologie na vysoké školy použít pojmy obsažené v RVP ZV a RVP G?
3. Nadhodnocují studenti učitelství biologie své vlastní znalosti statistických pojmů oproti výsledkům jejich znalostí zjištěných pomocí statistických úloh?
4. Cítí studenti učitelství biologie potřebu se ve statistice dále vzdělávat?
5. Existují odlišnosti mezi studenty učitelství biologie a jinými přírodovědnými obory?

3. Metodika

Za účelem zodpovězení výzkumných otázek bylo využito kvantitativního výzkumu, konkrétně dotazníkového šetření. Dotazník obsahuje pět baterií otázek: (1) identifikační otázky, (2) otázky týkající se zaměření žáka, (3) otázky

Tabulka 3: Ukázka záznamového archu – identifikační otázky a zaměření žáka

VŠ	Obor studia	Ročník	Pohlaví	SŠ
UK	lékařství	5	1	gymnázium

Maturita M	Bavila M	Statistika	Kombinatorika
0	0	0	1

týkající se sebehodnocení vlastních znalostí z oblasti statistiky, (4) otázky týkající se další potřeby se ve statistice dále vzdělávat, (5) testové otázky.

3.1. Identifikační otázky a otázky týkající se zaměření žáka

Tato baterie otázek bude sloužit jako třídící kritérium. Studenti byli dotazováni na následující otázky:

- vysoká škola (UK, JČ, atd.);
- název studovaného oboru (lékařství, učitelství biologie, učitelství matematiky, demografie, biologie);
- ročník studia (rozsah 1 až 5);
- pohlaví (0 – muž, 1 – žena);
- typ vystudované střední školy (gymnázium, pedagogická, průmyslová, ekologická, atd.);
- maturita z matematiky (0 – ne, 1 – ano);
- bavila studenta matematika na střední škole (0 – ne, 1 – ano);
- probíral učivo statistika (0 – ne, 1 – ano);
- probíral učivo kombinatorika (0 – ne, 1 – ano).

K zaznamenávání odpovědí na tyto otázky byl vytvořen záznamový arch (viz tab. 2).

3.2. Otázky týkající se sebehodnocení vlastních znalostí z oblasti statistiky

Rámcové vzdělávací programy pro základní školy a gymnázia obsahují v předmětu statistiky následující pojmy ze statistiky: pravděpodobnost, průměr, aritmetický průměr, modus, medián, směrodatná odchylka, mezikvartilová

odchylka, rozptyl, náhodná veličina, relativní četnost, korelační koeficient, kvantil, kvartil a hypotéza [6].

Z účelem zjištění, zda studenti tyto statistické pojmy znají, měli ohodnotit vlastní znalost těchto 14 statistických pojmů (pravděpodobnost, průměr, aritmetický průměr, modus, medián, směrodatná odchylka, mezikvartilová odchylka, rozptyl, náhodná veličina, relativní četnost, korelační koeficient, kvantil, kvartil, testování hypotéz) pomocí následující škály

- 0 – pojem jsem nikdy neslyšel/a;
- 1 – pojem jsem slyšel/a, ale nevím, co přesně znamená;
- 2 – pojem znám a uměl/a bych jej použít.

3.3. Otázky týkající se další potřeby se ve statistice dále vzdělávat

V této části dotazníku se studenti vyjadřovali k pěti otázkám:

- zda v průběhu studia na vysoké škole použili statistické metody (0 – ne, 1 – ano);
- zda by využili možnost konzultace se statistikem (0 – ne, 1 – ano);
- zda v průběhu studia na vysoké škole využili nějaký statistický software (0 – ne, 1 – ano) a pokud ano, tak jaký (otevřená otázka);
- zda budou pomocí statistiky zpracovávat data v rámci bakalářské (0 – ne, 1 – ano), tato otázka byla určena pouze pro studenty 1., 2. a 3. ročníků;
- zda budou pomocí statistiky zpracovávat data v rámci diplomové práce (0 – ne, 1 – ano), tato otázka byla určena pouze pro studenty 4. a 5. ročníků.

3.4. Testové otázky

V rámci statistického testu byly studentům předloženy 4 úlohy, které obsahovaly celkem dvanáct otázek, jejichž potenciální odpovědi byly 4 možnosti – (a) až (d). Studenti měli zakroužkovat u každé otázky právě jednu odpověď. V případě, že studenti odpověděli správně, do záznamového archu byla zapsána hodnota 1, pokud odpověděli špatně, nebo neodpověděli, byla zapsána hodnota 0.

Tyto otázky testovaly schopnost použít pojmy:

- průměr;
- korelační koeficient;

- směrodatná odchylka;
- relativní četnost;
- mezikvartilová odchylka;
- medián;
- kvartil;
- modus.

U každého z těchto pojmů byly vypočítány míry úspěšnosti, vždy počet správných odpovědí ku celkovému počtu odpovědí, a ty byly pomocí testu srovnání dvou relativních četností porovnány s počtem odpovědí „Pojem znám a uměl/a bych jej použít“ (viz bod (2) otázky týkající se sebehodnocení vlastních znalostí z oblasti statistiky). Výsledky testů pro jednotlivé pojmy byly zaznamenány do obr. 1. K zjištění rozdílů míry úspěšnosti v testových otázkách a relativním počtem odpovědí „Pojem znám a uměl/a bych jej použít“ napříč jednotlivými obory studia bylo využito intervalů spolehlivosti (viz obr. 7).

4. Výsledky výzkumu

Výzkumu se zúčastnilo 504 respondentů. Z celkového počtu respondentů bylo 76,2 % žen a 23,8 % mužů. Největší počet respondentů (52,8 %) studuje Univerzitu Karlovu v Praze, 22,8 % respondentů studuje Univerzitu Palackého v Olomouci a 20 % studuje Jihočeskou univerzitu v Českých Budějovicích. Další vysoké školy jsou zastoupeny pouze malým počtem respondentů. Zajímavé je, že 79 % všech respondentů absolvovalo gymnázium, ostatní typy středních škol se vyskytovaly pouze u malého počtu respondentů (5 % pedagogická střední škola, 4,5 % obchodní akademie, 2,3 % přírodovědná/ekologická střední škola).

Do výzkumu bylo zahrnuto 295 (58,5 %) studentů učitelství biologie, 97 (19,2 %) studentů učitelství matematiky, 72 (14,3 %) studentů medicíny, 22 (4,4 %) studentů demografie a 18 (3,6 %) studentů biologie. Výzkumu se zúčastnili studenti napříč jednotlivými ročníky, viz tab. 3. Více než dvě třetiny (81,2 %) respondentů studují bakalářské studium. Zbýlých 18,8 % studuje navazující magisterské studium.

Zajímavé je, že ačkoliv 60 % respondentů uvedlo, že je bavila matematika na střední škole, pouze 40 % respondentů maturovalo z matematiky.

Respondentům byla položena otázka, zda probírali na střední škole kombinatoriku a statistiku. Ačkoliv 84,5 % respondentů uvedlo, že probírali kombinatoriku, pouze 65,7 % uvádí, že probírali statistiku.

Tabulka 4: Počet studentů dle studovaného ročníku vysoké školy

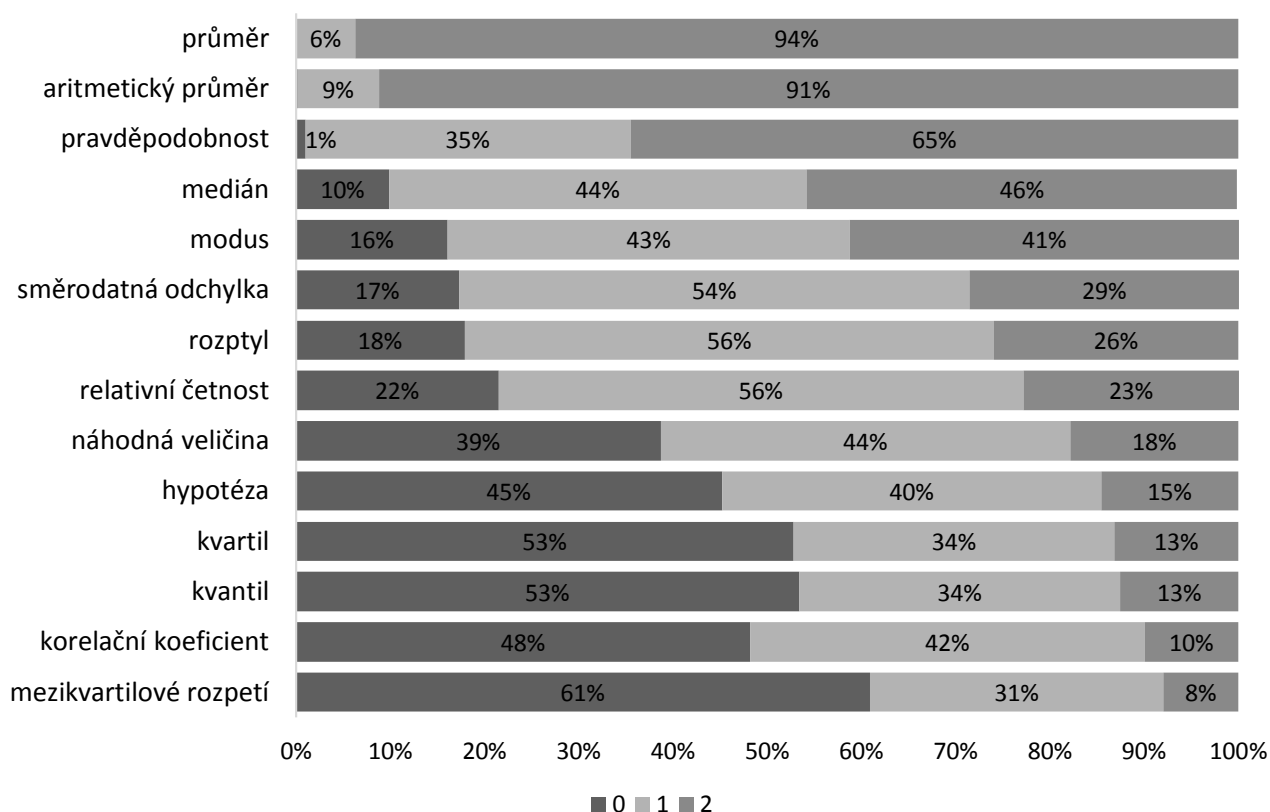
Ročníky	Počet	Relativní počet
1	145	28,8 %
2	180	35,7 %
3	84	16,7 %
4	42	8,3 %
5	53	10,5 %
Celkem	504	100,0 %

Tento zajímavý fakt může být vysvětlen tím, že se kombinatorika a následně po ní statistika vyučuje ve čtvrtém ročníku středních škol [6]. Je tedy pravděpodobné, že se mnoho studentů věnuje maturitním předmětům. Nižší procento kladných odpovědí u statistiky si lze vysvětlit tak, že je statistika úplně poslední probíranou látkou ve čtvrtém ročníku a tudíž se ne vždy stihne probrat.

Jak již bylo řečeno výše, z celkového souboru 504 respondentů, bylo 58,5 % tedy 295 učitelů biologie. Tento skupina respondentů je hlavním předmětem tohoto článku a první čtyři výzkumné otázky se týkají pouze této skupiny, a proto bude nyní blíže popsána. V tomto podsouboru je 20 % mužů a 80 % žen. Zajímavé je, že 72,5 % studentů učitelství biologie studovalo gymnázium a pouze 6,4 % střední školu s pedagogickým zaměřením. Více než 25 % studentů učitelství maturovalo z matematiky a 50 % uvádí, že je na střední škole matematika bavila. Téměř 77 % si pamatuje, že na střední škole probírali kombinatoriku, a pouze 64 % si pamatuje, že na střední škole probírali statistiku. Nejvíce respondentů (36,3 %) studuje 1. ročník a nejméně 4. a 5. ročník (oba po 13,2 %). Z druhého ročníku je 17,6 % a z 3. ročníku 29,7 % respondentů.

4.1. Znají studenti učitelství biologie statistické pojmy, které jsou obsažené v RVP pro základní školy a gymnázia?

Z obr. 1 vyplývá, že se 94 % studentů domnívá, že zná a umí použít průměr. Zajímavé je, že již menší počet studentů, tj. 91 %, uvádí, že zná a umí použít aritmetický průměr. Další pojmy už jsou výrazně méně známé. Pojem pravděpodobnost zná téměř 99 % studentů učitelství biologie, umí jej použít

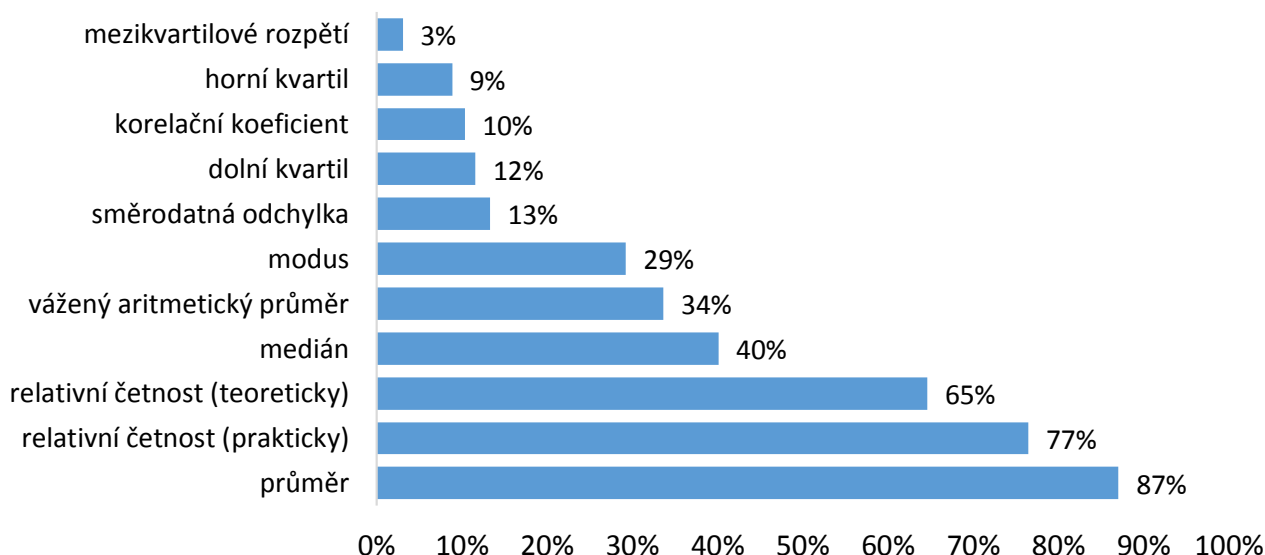


Obrázek 1: Relativní četnosti hodnocení vlastních znalostí jednotlivých pojmů (0 – pojem jsem nikdy neslyšel/a; 1 – pojem jsem slyšel/a, ale nevím, co přesně znamená; 2 – pojem znám a uměl/a bych jej použít) studentů učitelství biologie; $n = 295$

pouze 65 %. Medián a modus zná téměř 90 % dotázaných studentů učitelství biologie, ale méně než polovina je umí použít. Nejméně známými a osvojenými pojmy jsou pojmy mezikvartilová odchylka, korelační koeficient, kvantil a kvartil, které zná a umí použít méně než 13 % studentů učitelství biologie.

4.2. Umí studenti učitelství biologie na vysoké školy použít pojmy obsažené v RVP ZV a RVP G?

Studentům učitelství biologie byl předložen test, zjišťující jejich znalosti v oblasti základních statistických pojmů. Výsledky shrnuje obr. 2. Nejlépe dopadl výpočet průměru, který správně spočítalo 87 % studentů učitelství biologie. Dále byli studenti úspěšní i u praktického výpočtu četností (77 %). Zajímavé je, že již nižší úspěšnost měl výpočet mediánu (40 %) a modu (29 %). Velmi nízkou míru úspěšnosti (34 %) měl i příklad ověřující schopnost vypočítat



Obrázek 2: Počet správných odpovědí u jednotlivých otázek testu u studentů učitelství biologie; $n = 295$

vážený aritmetický průměr. Méně než deset procent studentů správně vypočítalo korelační koeficient, horní kvartil a mezikvartilová odchylka.

4.3. Nadhodnocují studenti učitelství biologie své vlastní znalosti statistických pojmů oproti výsledkům jejich znalostí zjištěných pomocí statistických úloh?

Korelační koeficient mezi počtem správných odpovědí na testové otázky a počtem studentů, že pojmy znají a umí je použít, je 0,40 a na hladině významnosti 5 % je statisticky významný.

Studenti učitelství biologie se statisticky významně nadhodnocují ve schopnosti vypočítat průměr, směrodatnou odchylku, mezikvartilová odchylka, medián a modus, viz tab. 4. Naopak se statisticky významně podhodnocují u výpočtu relativní četnosti. Statisticky nevýznamný rozdíl mezi očekáváním studentů a výsledkem testu byl zjištěn pouze u korelačního koeficientu a kvartilu.

Zajímavé je i srovnání očekávaných znalostí a reálných schopností podle jednotlivých ročníků studia, viz obr. 3. Zatím, co v prvním ročníku je úroveň obou měr nízká a studenti se ani nepodhodnocují, ani nenadhodnocují, v 2., 3. a 4. ročníku se výrazně podhodnocují. Zatím, co míra znalostí roste mezi prvním a druhým rokem studia, míra očekávaných znalostí roste až ve čtvrtém a pátém ročníku. V pátém ročníku se pak opět obě míry vyrovnají a studenti se opět ani nepodhodnocují ani nenadhodnocují.

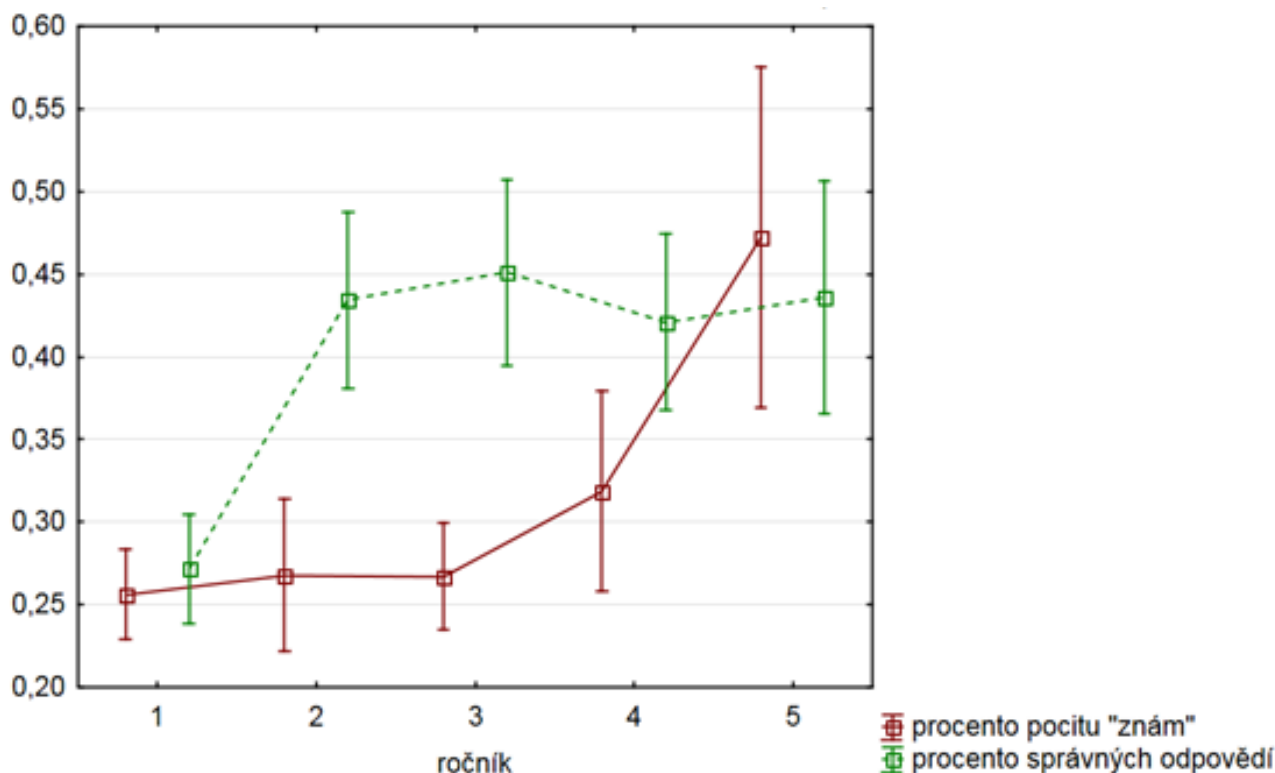
Tabulka 5: Srovnání relativních četností odpovědí Pojem znám a uměl/a bych jej použít a správných odpovědí v testu studentů učitelství biologie dle jednotlivých pojmů včetně výsledku testu o shodě dvou relativních četností na 5 % hladině významnosti; $n = 295$

Pojem	Relativní četnost počtu odpovědí Pojem znám a uměl/a bych jej použít.	Relativní četnost počtu správných odpovědí v testu	Výsledek testu (test o shodě dvou relativních četností)
Průměr	0,91	0,61 ³	Nadhodnocení znalostí
Korelační koeficient	0,10	0,10	Stat. nevýznamný rozdíl
Směrodatná odchylka	0,29	0,13	Nadhodnocení znalostí
Relativní četnost	0,23	0,73 ⁴	Podhodnocení znalostí
Mezikvartilová odchylka	0,08	0,03	Nadhodnocení znalostí
Medián	0,46	0,40	Nadhodnocení znalostí
Kvartil	0,13	0,11 ⁵	Stat. nevýznamný rozdíl
Modus	0,41	0,29	Nadhodnocení znalostí

³ Spočítána jako průměr dvou otázek týkajících se průměru.

⁴ Spočítána jako průměr tří otázek týkajících se relativních četností.

⁵ Spočítána jako průměr dvou otázek týkajících se kvartilů – horní a dolní kvartil.



Obrázek 3: Srovnání průměrného počtu správných odpovědí (v procentech) a průměrného počtu odpovědí „Pojem znám a uměl/a bych jej použít“ (v procentech) studentů učitelství biologie; $n = 295$

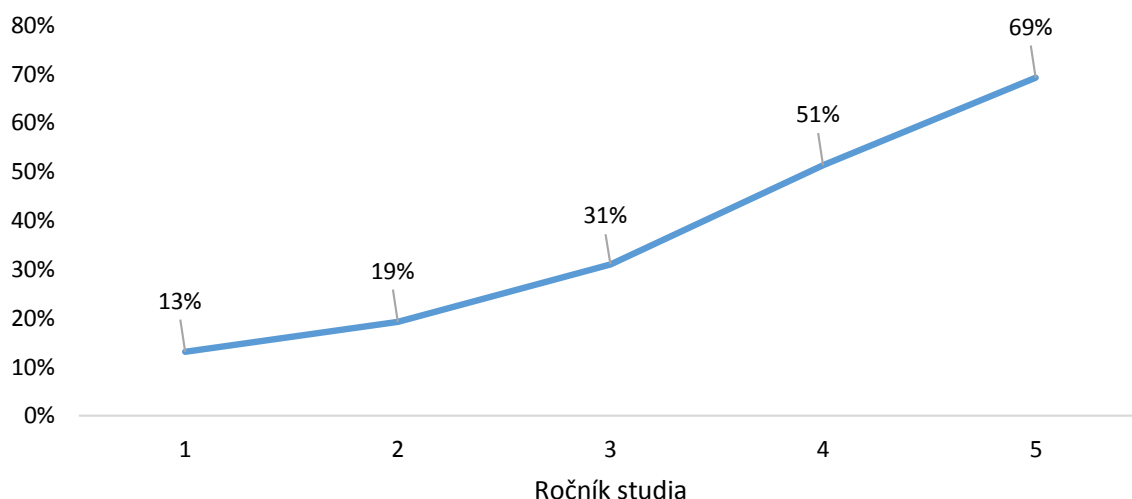
4.4. Cítí studenti učitelství biologie potřebu se ve statistice dále vzdělávat?

V průběhu studia již využilo statistické metody pouze 30 % studentů učitelství biologie. Toto souhrnné číslo rozvádí obr. 4, který ukazuje rozložení využívání statistických metod podle ročníku studia.

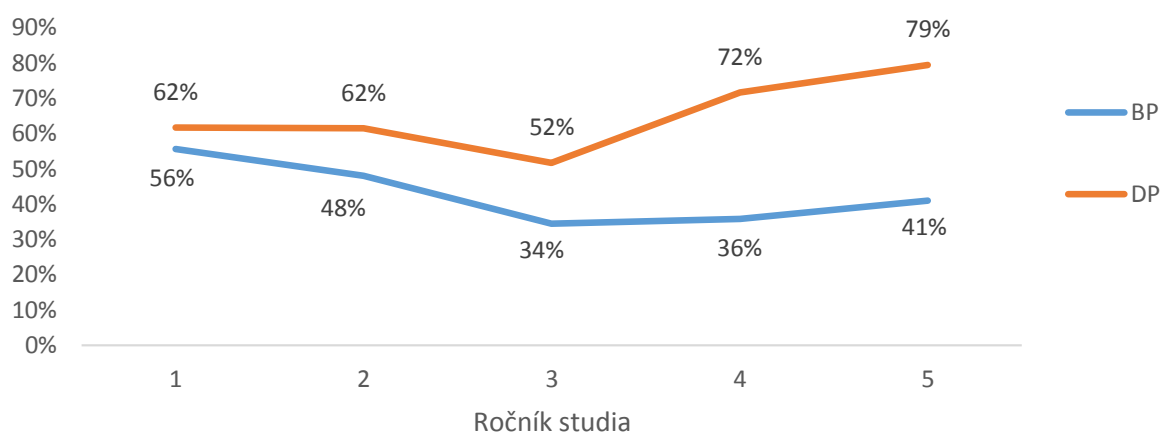
S tím souvisí i fakt, že 45,5 % studentů plánuje využít nebo využilo statistické metody při psaní bakalářské práce a 63 % studentů při psaní práce diplomové. Rozdělení podle studovaného ročníku je uvedeno v obr. 5.

Software při statistických výpočtech využívá pouze 55 % studentů. Z nich 87 % využívá MS Excel, v menší míře je využíváno i CANOCO (7 %), SPSS (3 %), R (2 %) a další.

Studenti učitelství biologie byli dále dotazováni na to, zda by využili možnost konzultace se statistikem. Téměř 42 % uvedlo, že by této možnosti využili. Obr. 6 znázorňuje rozdělení studentů, kteří v dotazníku uvedli, že by využili konzultace se statistikem, podle studovaného ročníku.



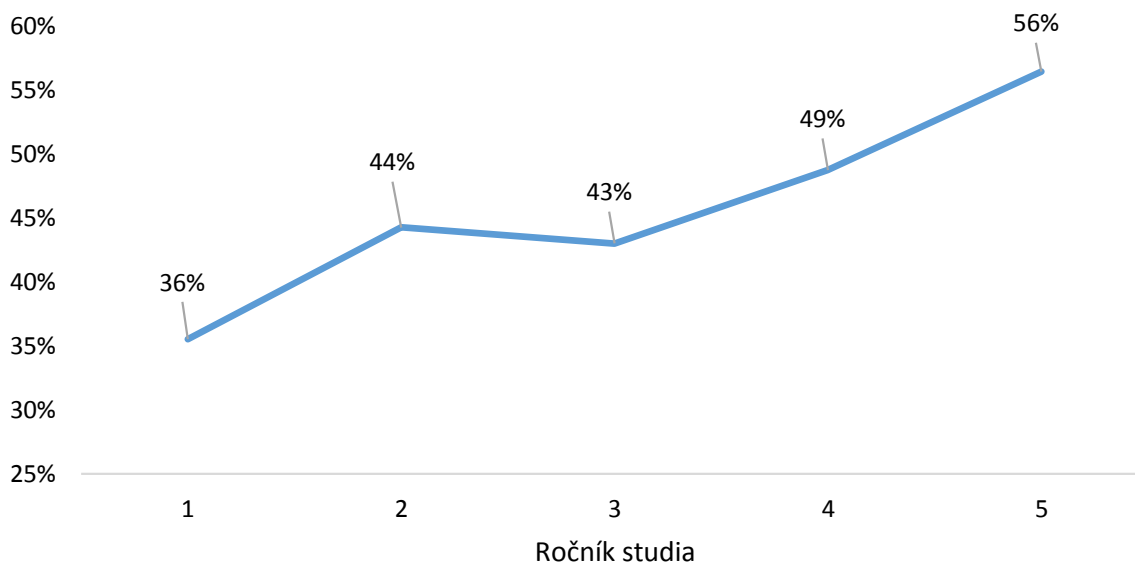
Obrázek 4: Počet studentů učitelství biologie využívajících statistické metody dle ročníku studia; $n = 295$



Obrázek 5: Počet studentů učitelství biologie, kteří využijí statistiku v závěrečné práci; $n = 295$

4.5. Existují odlišnosti mezi studenty učitelství biologie a jinými přírodovědnými obory?

Nejnižší sebevědomí týkající se znalostí pojmů statistiky mají studenti učitelství biologie a matematiky. Naopak vysoké sebevědomí mají studenti biologie a demografie, viz obr. 7. V testu znalostí dopadli nejhůře studenti učitelství biologie a medicíny. Největší procento správných odpovědí pak měli studenti biologie a demografie. Studenti medicíny, biologie a demografie své schopnosti ani nepřecenili, ani nepodcenili. Studenti učitelství biologie a matematiky své znalosti statisticky významně podhodnotili.

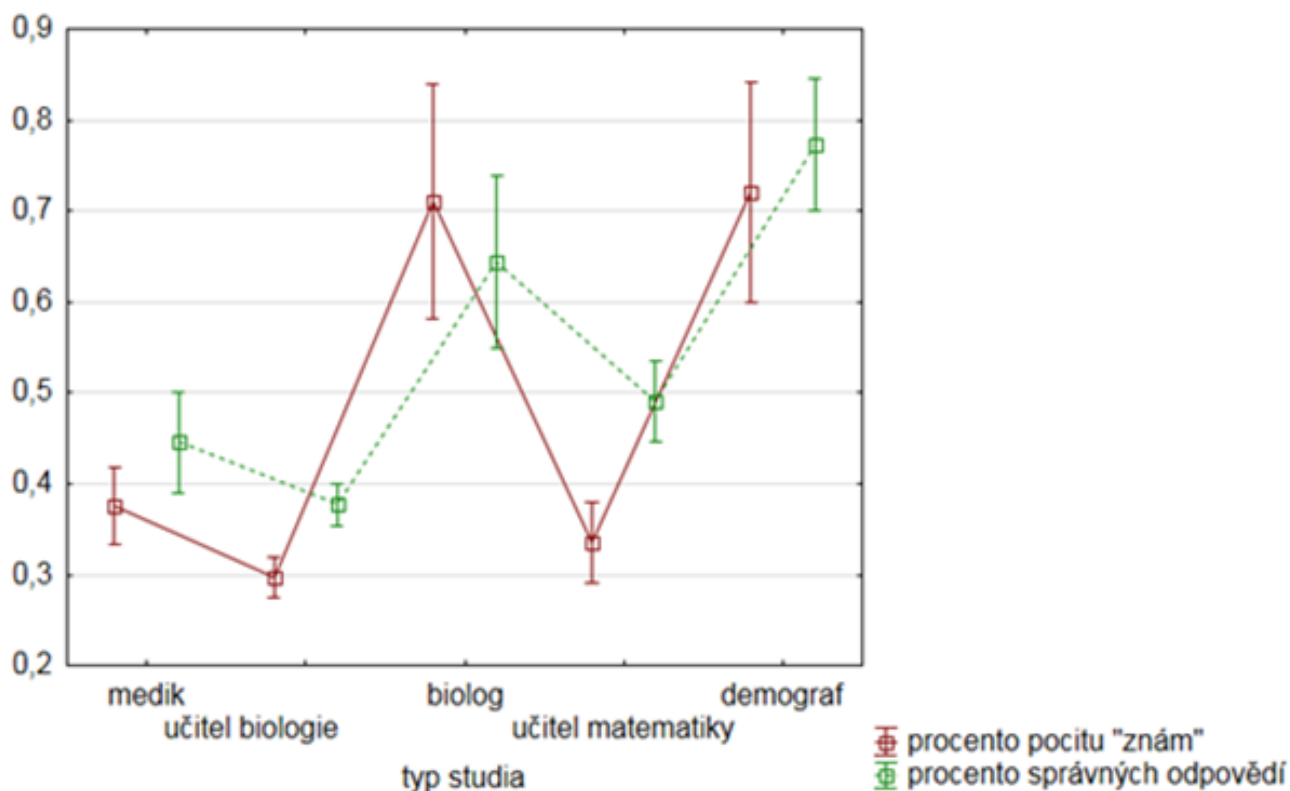


Obrázek 6: Počet studentů učitelství biologie, kteří by využili možnost konzultace se statistikem podle ročníku studia; $n = 295$

Při bližším porovnání studentů jednotlivých typů oborů je zřejmé, že učitelé dopadli v testu hůře než ostatní. Statisticky významně se liší průměr studentů učitelství biologie od studentů učitelství matematiky, biologie a demografie. Odlišnost od mediků je statisticky nevýznamná.

5. Diskuze

Zajímavá je i skutečnost, že pouze 84,5 % studentů si pamatuje, že probírali kombinatoriku na střední škole a pouze 65,7 % si pamatuje, že probírali statistiku na střední škole. Tento fakt může být vysvětlen tím, že je kombinatorika a následně po ní statistika vyučována až ve čtvrtém ročníku středních škol [1]. Je tedy pravděpodobné, že se mnoho studentů věnuje povinným maturitním předmětům a nevěnují pozornost matematice. Ale 40 % respondentů také uvedlo, že z matematiky maturovalo. Je tedy možné, že kombinatoriku a statistiku nepovažují za důležitou a nevěnují jim tolik pozornosti a brzy pojmy, které si osvojili, zapomenou. Nižší procento kladných odpovědí u statistiky si lze vysvětlit tak, že je statistika úplně poslední probíranou látkou ve čtvrtém ročníku, a tudíž se ne vždy stihne probrat. Pokud se žáci a studenti v průběhu vzdělávání na základních a středních školách neseznámí se základními statistickými pojmy, je pak velice těžké odhadnout vstupní míru znalostí pro úvodní kurz statistiky na vysoké škole. A to hlavně pokud je tato úroveň variabilní, tedy od studentů, kteří si neosvojili žádné pojmy až po ty, kteří si osvojili všechny pojmy podle RVP.



Obrázek 7: Srovnání průměrného počtu správných odpovědí v testu a průměrného počtu správných odpovědí Pojem znám a uměl/a bych jej použít (v %); $n = 504$

5.1. Znají studenti učitelství biologie statistické pojmy, které jsou obsažené v RVP pro základní školy a gymnázia?

Hodnocení vlastních znalostí je velice těžká záležitost. Takové sebehodnocení však vypovídá zejména o sebedůvěře v daném oboru. Studenti učitelství biologie si evidentně ve statistice příliš nevěří. Obr. 4 ukazuje, že studenti hodnotili velice kladně pouze svou znalost průměrů. U dalších pojmů si studenti už tak jistí nejsou. Pojmy relativní četnost, náhodná veličina, hypotéza, kvantil, kvartil, korelační koeficient a mezikvartilová odchylka zná a umí vypočítat méně než čtvrtina dotázaných studentů učitelství biologie.

Je zajímavé, že existuje rozdíl v hodnocení vlastních znalostí u pojmů průměr a aritmetický průměr, ačkoliv jde fakticky o tentýž pojem. Někteří studenti se zřejmě zalekli slova „aritmetický“ a ohodnotili svoje znalosti nižším bodem škály. Dále je velice zajímavé, že pojmy, které spolu úzce souvisí, jsou studentům přibližně stejně známé a studenti je umí vypočítat. Například medián a modus, jsou dvě míry polohy, které jsou zpravidla vy-

světlovány zároveň, a zná a umí je vypočítat přibližně 40 % respondentů. Směrodatnou odchylku a rozptyl zná a umí vypočítat 29 % a 26 % respondentům, kvantil a kvartil shodně 13 % respondentů. Pokud tedy probírali základní míry polohy (průměr, modus, medián) nebo základní míry variability (směrodatná odchylka a rozptyl), pojmy si pamatují a umí je použít.

5.2. Umí studenti učitelství biologie na vysoké školy použít pojmy obsažené v RVP ZV a RVP G?

Ačkoliv rámcové vzdělávací programy pro základní školy a gymnázia, jak píše Hybšová [6], obsahují tyto statistické pojmy a studenti na vysokých školách by s nimi měli být již dávno seznámeni, podle výsledku testu je zřejmé, že tomu tak není (viz obr. 5). Studenti učitelství biologie téměř neznají pojmy směrodatná odchylka, dolní a horní kvartil, korelační koeficient a mezikvartilová odchylka, ty zvládlo vypočítat méně než 15 % studentů učitelství biologie. Ani výpočet modu a mediánu, kde byla úspěšnost nižší než 50 %, nevypovídá o znalosti a schopnosti oba pojmy použít. Je zajímavé, že prostý aritmetický průměr vypočítalo 87 %, ale vážený aritmetický průměr pouze 34 % studentů učitelství biologie. Nízká úspěšnost u výpočtu váženého aritmetického průměru je velice překvapující.

5.3. Nadhodnocují studenti učitelství biologie své vlastní znalosti statistických pojmů oproti výsledkům jejich znalostí zjištěných pomocí statistických úloh?

Mezi výsledky testu a hodnocením vlastních znalostí existuje statisticky významná korelace. Znamená to, že studenti poměrně reálně odhadovali svoje znalosti a schopnosti. Z tab. 4 lze vyčíst rozdíly mezi výsledkem testu a očekáváními.

Studenti učitelství biologie neznají pojmy korelační koeficient a kvartil, tomu i odpovídá výsledek testu, kde se prokázalo, že tyto hodnoty neumějí vypočítat. U těchto dvou pojmů byly rozdíly mezi pocitem studentů, že pojem znají a umějí vypočítat, a výsledkem testu, statisticky nevýznamné, přesto jsou obě míry velice nízké. Lze předpokládat, že se s těmito pojmy v průběhu studia setkala pouze malé množství studentů, a z toho důvodu ani nepředpokládali, že by mohli jejich hodnoty spočítat. Absolutním opakem jsou pojmy průměr, směrodatná odchylka, mezikvartilová odchylka, medián a modus. Studenti u těchto pojmů ve velké míře uvádí, že pojmy znají a umějí je použít. V testových příkladech je však míra úspěšnosti nižší než míra předpokladu studentů, že pojem znají a umějí jej použít. U těchto pojmů studenti

statisticky významně nadhodnotili své schopnosti. Lze předpokládat, že studenti tyto pojmy znají, slyšeli je, ale nejsou schopni je vypočítat, popřípadě nerozumí těmto pojmům zcela dobře. Jediný pojem, u kterého studenti podhodnotili své schopnosti, je pojem relativní četnost. V testu byl pojem relativní četnost zakomponován do tří otázek (dvou praktických úloh a jedné teoretické), úspěšnost teoretické byla významně nižší než dvou praktických. Hlavním důvodem bude zřejmě jednoduchost tohoto konceptu a neznalost jeho názvu. Je pravděpodobné, že si studenti nevybavili, co přesně pojem relativní četnost znamená, a zhodnotili, že ji neumí vypočítat. U praktických příkladů, kde šlo o výpočet relativních četností, ale nebylo použito sousloví relativní četnost, studenti neměli problém ji vypočítat.

5.4. Cítí studenti učitelství biologie potřebu se ve statistice dále vzdělávat?

Pouze 30 % studentů učitelství biologie uvádí, že využívá statistické metody. Zajímavé je, jak je těchto 30 % rozloženo podle studovaného ročníku, viz obr. 6. Z něj je zřejmé, že v prvních dvou ročnících je využívání statistických metod málo časté. Ve třetím ročníku, tedy v období psaní bakalářských prací, je již vyšší, 31 %. V pátém ročníku, tedy při psaní diplomové práce, již 69 % studentů využívá statistické metody. Je tedy zřejmé, že potřeba statistiky se zvyšuje s blížícím se koncem studia a psaní závěrečných prací.

Je zajímavé, že možnosti konzultace se statistikem by využilo 41 % studentů učitelství biologie. To značí poměrně veliký zájem. I zde, stejně jako u využívání statistických metod, je vidět rostoucí trend v závislosti na studovaném ročníku (viz obr. 8). V posledním roce studia by tuto možnost využila více než polovina dotázaných studentů učitelství biologie.

5.5. Existují odlišnosti mezi studenty učitelství biologie a jinými přírodovědnými obory?

Je zajímavé, že studenti učitelství (biologie i matematiky) podhodnotili své znalosti a v testu dopadli lépe, než očekávali. Lze jen spekulovat o tom, proč si budoucí učitelé tak málo věří a jakým způsobem by bylo vhodné jejich sebedůvěru zvýšit. Bohužel i v testu znalostí je patrné, že studenti učitelství biologie zaostávají oproti ostatním oborům. Rozdíl mezi výsledky učitelů biologie a biologie lze vysvětlit rozdílnými kurikuly. Hybšová a Leppink [4] porovnali kurikula z hlediska výuky statistiky oborů Biologie a Učitelství biologie na UK V Praze a zjistili značné rozdíly v rozsahu výuky statistiky. Jedním z uvedených příkladů je fakt, že studenti biologie mají nejen přednášky

ale i praktika na počítačích. Dalším důvodem je i větší množství volitelných kurzů týkajících se výzkumu a statistických metod.

Závěr

Analýza znalostí pojmů u studentů učitelství biologie ukázala, že mnoho pojmů, které si studenti měli osvojit na základních a středních školách, studenti neumějí vypočítat a některé vůbec neznají. Navzdory tomuto poměrně negativnímu zjištění se poměrně velká část studentů chce ve statistice dále vzdělávat a dokonce by rádi využili i konzultace se statistikem (viz obr. 6).

Vysoká škola jako vzdělavatel budoucích učitelů biologie vytváří kurzy, které pomáhají studentům lépe se připravit na budoucí povolání. Příprava úvodního kurzu statistiky pro budoucí učitele biologie vzhledem k variabilitě jejich znalostí musí zahrnovat i velice základní pojmy jako je průměr, modus a medián. Ačkoliv by tyto pojmy již měli mít studenti osvojeny, což vyplývá z analýzy RVP [6], z testu vyplývá pravý opak. Znamená to tedy, že je třeba důkladně zopakovat učivo, které by je obsahem rámcových vzdělávacích programů na základních a středních školách, aby se sjednotila vstupní úroveň studentů. Je třeba ale také zvážit i důvody, proč studenti po absolvování základní a střední školy neumí použít základní statistické pojmy, ačkoliv v RVP jsou uvedeny. Důvodů může být několik: Jedním z nich je možnost, že se statistika se v rámci výuky nestihne probrat nebo se probírá pouze povrchně a žáci a studenti si pojmy dostatečně neosvojí. Nejzávažnějším důvodem však bude skutečnost, že látka je probrána v předmětu Matematika a dále se k ní vyučující již nevrací. Fakticky pak studenti vůbec nevyužijí nabyté znalosti v jiných předmětech a osvojené učivo brzy zapomenou. Přesahy z matematiky do jiných předmětů by měly být patrné i v RVP, jak ale Hybšová [6] píše, statistické pojmy jsou probírány pouze v matematice a přesahy do jiných předmětů nejsou v RVP explicitně napsány. Ze závěrů České školní inspekce, která provedla v květnu 2015 šetření Zjišťování výsledků žáků 2015,⁶ jednoznačně vyplývá, že vzhledem k nízké úspěšnosti žáků a studentů v oblasti přírodních věd by se měli změnit postupy výuky matematiky. Bohužel se však nikde neuvádí, že by se měli změnit také postupy výuky přírodních věd. Žákům a studentům by neměli být předkládány pouze holá fakta, ale také by jim měli být ukázány nejružnější výzkumy, postupy a metody ověřování, aby si plně uvědomovali, jakým postupem vznikají poznatky o přírodě, a aby si ověřili, že znalosti nabyté v matematice, lze uplatnit v praxi. Vyučující přírodovědných předmětů a vyučující matematiky by měli mezi sebou

⁶Výsledky lze najít online zde:

<http://www.csicr.cz/getattachment/349e4bab-48e7-44bd-b873-67fbfe346e81>

kooperovat a vytvářet jak v matematice, tak v přírodovědných předmětech prostor pro ukázky praktického využití matematiky a statistiky. Bez tohoto propojení jen těžko žáci a studenti pochopí jinak naprosto zřejmou spojitost mezi matematikou a přírodními vědami.

Nasnadě je hned několik řešení tohoto problému. Prvním je zavádění kurzů statistiky přímo určené pro budoucí učitele přírodních věd. Do sylabů by byla zahrnuta i velice základní témata jako je výpočet váženého aritmetického průměru nebo modu a mediánu, která ačkoliv byla obsažena v RVP pro základní školy i gymnázia, nejsou dostatečně studenty osvojena. Dalším řešením je vytvoření e-learningových kurzů (ve smyslu [10]) přímo zaměřených na výuku přírodních věd s využitím statistiky pro žáky základních a středních škol. Ty by pak sloužily jako didaktický materiál pro žáky a studenty, ale také jako ukázka mezioborových přesahů studentům učitelství biologie. Ačkoliv je na pedagogických fakultách je značný inovativní potenciál [11], implementace takovýchto změn zabere ještě mnoho času.

Poděkování

Výzkum byl podpořen Grantovou agenturou Univerzity Karlovy v Praze (projekt č. 341115).

Literatura

- [1] Balada, J. (2007): *Rámcový vzdělávací program pro gymnázia: RVP G*. Praha: Výzkumný ústav pedagogický v Praze.
- [2] Ben-Zvi, D., Garfield, J. B. (2004): *The challenge of developing statistical literacy, reasoning and thinking*, Dordrecht: Kluwer Academic Publishers.
- [3] Hybšová, A. (2014): Statistická gramotnost a její význam v pedagogice a oborových didaktikách. In: *Aktuální problémy pedagogiky ve výzkumech studentů doktorských studijních programů*, UPOL, Olomouc.
- [4] Hybšová, A., Leppink, J. (2015): The subject of statistics in natural science curricula: a case study. *Journal on Efficiency and Responsibility in Education and Science* 8(1), pp. 8–14, online ISSN 1803-1617, printed ISSN 2336-2375, doi: 10.7160/eriesj.2015.080102.
- [5] Hybšová, A. (2015a): *Statistical literacy in teaching of natural science. Turning data into the knowledge*. University of Lisboa, Lisboa.
- [6] Hybšová, A. (2015b): Statistická gramotnost v RVP. *Informační bulletin České statistické společnosti*, 26(4), pp. 18–35.

- [7] Kvaszová, M. (2012): *Didaktika statistiky*. Dizertační práce. Praha: Univerzita Karlova v Praze.
- [8] *Rámcový vzdělávací program pro základní vzdělávání* (2013). Upravené vydání se změnami účinnými od 1.9.2013. Praha: Národní ústav pro vzdělávání. Dostupné online z <http://www.nuv.cz/cinnosti/kurikulum-vseobecne-a-odborne-vzdelavani-a-evaluace/ramcove-vzdelavaci-programy/upraveny-rvp-zv>
- [9] Markechová, D., Stehlíková, B., Tirpáková, A. (2011): *Štatistické metódy a ich aplikácie*. Nitra: UKF.
- [10] Nocar, D. (2004): *E-learning v distančním vzdělávání*. 1. vyd. Olomouc: Univerzita Palackého.
- [11] Spilková, V., Wildová, R. (2014): Potřebujeme kvalitní nebo kvalifikované učitele? *Pedagogická orientace* 24(3). doi: <http://dx.doi.org/10.5817/PedOr2014-3-423>
- [12] Wallman, K. K. (1993): Enhancing statistical literacy: Enriching our society, *Journal of the American Statistical Association* 88(421), pp. 1–8.
- [13] Watson, J. M., Kelly, B. A., Callingham, R. A., Shaughnessy, J. M. (2003): The measurement of school students [201F?] understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*.

VZPOMÍNKY NA JIŘÍHO ŽVÁČKA MEMORIES ON JIŘÍ ŽVÁČEK

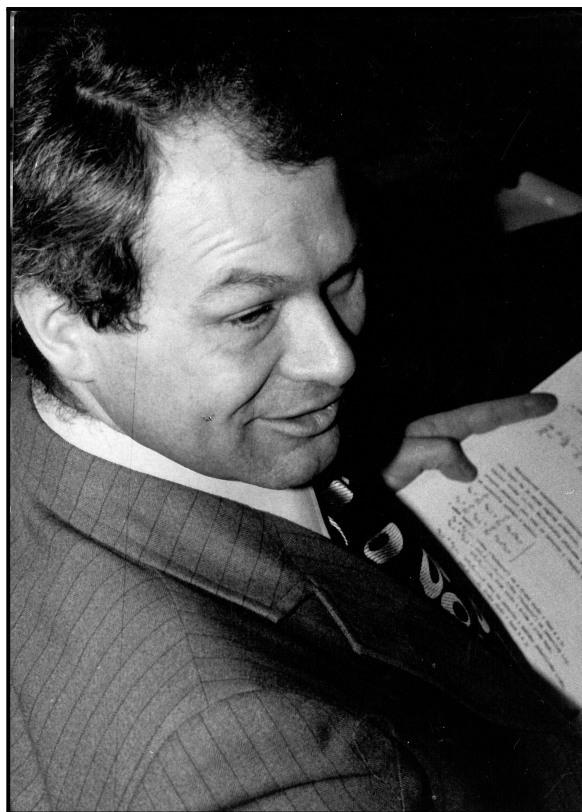
Hana Řezanková, Prokop Závodský

24. listopadu 2015 zemřel ve věku 72 let jeden ze zakládajících členů České statistické společnosti, doc. Ing. Jiří Žváček, CSc., významný odborník, který se zasloužil o rozvoj spolupráce statistiků z různých oblastí vědy a praxe. Rádi bychom touto cestou připomněli alespoň některé jeho zásluhy.

Jiří Žváček se narodil 10. ledna 1943 v Praze v rodině armádního důstojníka, za okupace vězněného. V letech 1960–1965 vystudoval Vysokou školu ekonomickou v Praze. Své znalosti si prohloubil mimořádným studiem na Matematicko-fyzikální fakultě Univerzity Karlovy, kde pak v letech 1980–1982 absolvoval i postgraduální studium. Studiu statistiky a matematiky se věnoval i na studijním pobytu na universitě a technice v Zürichu (1969/70).

Svou profesní dráhu zahájil jako programátor-analytik na Státní plánovací komisi, po dvou letech byl v r. 1967 přijat na katedru statistiky VŠE. V roce 1982 obhájil svou disertační práci ze statistické prognostiky a získal titul kandidáta věd v oboru ekonomická statistika. U posluchačů i kolegů měl Jiří Žváček pověst náročného pedagoga. Pro svou přímou a kritickou povahu nebyl u předlistopadového vedení VŠE v oblibě, docentem byl jmenován teprve v březnu 1990.

Po listopadu 1989 se J. Žváček s odhodláním a pracovitostí sobě vlastní zapojil do reformování a modernizace Vysoké školy ekonomické. V červnu 1990 se stal pedagogickým proděkanem tehdejší Fakulty národohospodářské a koncem roku kandidoval na funkci rektora VŠE. Na žádost vítěze voleb, doc. Štěpána Müllera, přijal pak funkci prorektora VŠE pro pedagogickou činnost a informatizaci (od 1. března 1991). Doc. Žváček byl hlavním autorem návrhu na rychlé zavedení kreditního systému studia, na našich vysokých školách dosud nevyzkoušeného. Využil zde svých zkušeností ze studijního pobytu na amerických univerzitách (v letech 1987–1988). Významně se též



podílel na celkové reorganizaci VŠE do pěti nových fakult k 1. srpnu 1991. Obdivuhodnou činnost doc. Žváčka přerušil naneštěstí již v říjnu 1991 těžký úraz, který utrpěl při horolezeckém výcviku a jehož následkem ochrnul.

Jiří Žváček se ve svých odborných publikacích zaměřoval především na metody a aplikace analýzy a extrapolace časových řad. V aplikacích se soustřeďoval na ekonomický a demografického vývoj, z konkrétních příkladů lze uvést analýzu finančních ukazatelů. K dalším oblastem jeho profesního zájmu patřily metody vícerozměrné statistické analýzy, například regrese a shluková analýza. V souvislosti s rozvojem výpočetní techniky se soustředil na postupy výpočetní statistiky a intenzivně sledoval vývoj v oblasti programového vybavení pro statistické analýzy. Pozornost věnoval také rozvoji metod umělé inteligence, např. v roce 1987 spolu s Jiřím Ivánkem publikoval v časopise *Statistika* (č. 1, str. 4–21) článek s názvem *K použití expertních systémů ve statistice*.

Jako podklad pro jmenovací řízení docentem předložil Jiří Žváček v květnu 1989 seznam 47 článků v odborných časopisech a sbornících (z toho 8 v té době v tisku), 60 výzkumných prací, 4 skripta, 16 recenzí a 13 jiných „drobných prací“. K uvedenému datu publikoval zejména v časopisech *Statistika* (jako autor či spoluautor 15 článků) a *Finance a úvěr* (6 článků), také v časopisech *Acta demographica* a *Wiadomosci statystyczne*. Byl autorem či spoluautorem několika článků ve sbornících vycházejících v polské Wroclawi v sérii *Prace Naukowe AE* a na VŠE v sérii *Acta Oeconomica Pragensia*. Od roku 1986 publikoval též v časopise *Informácie* vydávaném Slovenskou demografickou a statistickou a společností. Byl aktivním účastníkem na mnoha různých konferencích, např. *ROBUST* a *COMPSTAT*. Kromě českého jazyka publikoval v angličtině, němčině a ruštině.

Jeho velká zásluha spočívá v pedagogické činnosti, jednak na VŠE, jednak v rámci školení pracovníků ČSÚ, příp. v rámci jiných kurzů. Jiří Žváček připravil jako autor či spoluautor řadu učebních textů k používání programových systémů jak statistických (SYSTAT, Statgraphics), tak databázových (dBASE) a tabulkových (Lotus 1-2-3), nezbytných pro přípravu datových souborů, a jiných, např. k ovládání programovacího prostředí Turbo Pascal. Nešlo mu o to, vydat publikaci, ale připravit publikaci, která by byla uživateli co nejnázornější. Do svých textů často zařazoval výstřižky vtipů, které vhodně doplňovaly probíranou problematiku.

Po založení České statistické společnosti v roce 1990 byl přispěvatelem nově vzniklého časopisu *Informační bulletin České statistické společnosti*. Kromě odborných článků publikoval i různé informace (např. o vzniku Slovenské statistické a demografické společnosti nebo o novinkách statistického softwaru) a úvahy. Například v příspěvku *Klub nepřátel statistiky* (č. 3 z roku

1990) navrhoval každoroční vyhlašování cen pro „největší statistickou pítomost“.

Činnost doc. Žváčka v oblasti statistiky byla rozsáhlá. Byl i členem redakční rady časopisu *Statistika* a členem několika vědeckých společností – kromě českých (Česká statistická společnost, Jednota čs. matematiků a fyziků, Československá kybernetická společnost) také slovenské (Slovenská demografická a statistická společnost, později SŠDS) a mezinárodní (Mezinárodní společnost pro výpočetní statistiku – IASC). Po založení České statistické společnosti (ČStS) se stal jejím místopředsedou. K jeho dlouholeté činnosti patřila správa webových stránek ČStS, které vytvořil i přes své zdravotní komplikace a které průběžně aktualizoval. V rámci těchto stránek zajišťoval on-line publikování časopisu *Informační bulletin České statistické společnosti*. Kromě působení na VŠE lze zmínit spolupráci doc. Žváčka se statistickým úřadem a z pozdější doby spoluautorství učebnice statistiky s internetovými odkazy a interaktivními programy pro Metropolitní univerzitu Praha.

Celou poslední třetinu svého života strávil doc. Žváček ve zdravotnických zařízeních a trpěl mnoha zdravotními potížemi. Díky své neuvěřitelné houževnatosti, pílí a optimismu pokračoval po celou dobu podle možností v odborné práci. Naučil se ovládat počítač tyčinkou, kterou držel v ústech, a každou chvíli, kdy mu to zdravotní stav umožnil, pracoval. Jen v databázi publikační činnosti VŠE je od roku 1993 evidováno 13 publikací, jichž je spoluautorem. Je uvedeno 7 článků v časopisech (*Statistika*, *Informační bulletin České statistické společnosti* a *Chip*), 3 příspěvky ve sbornících z konferencí a dvojce skripta. Jiří Žváček se stále zajímal o dění ve statistické komunitě. Pokud to bylo možné z organizačního a zdravotního hlediska, účastnil se statistických akcí pořádaných Českou statistickou společností, např. konferencí *ROBUST* a *STAKAN*.

Jiří Žváček byl znám svou houževnatostí, s jakou se snažil splnit cíle, které si předsevzal, ať už jako horolezec, nebo jako prorektor, později jako těžce zdravotně postižený. Jeho pracovitost a píle by měla být pro mnohé příkladem. Uměl však také vnímat krásy přírody a umění, rád cestoval, miloval hory. I v jeho poslední nelehké třetině života mu díky obětavým lidem bylo umožněno zúčastnit se výstav, zájezdů, konferencí a dalších akcí. V červnu 2015 se zúčastnil přátelského setkání pracovníků Fakulty informatiky a statistiky v areálu VŠE na Točné. Byla to jeho poslední návštěva na VŠE. Nádherné počasí, skvělá atmosféra, báječná živá hudba. Doc. Žváček odjížděl při písni „C'est la vie“.

Obsah

Vědecké a odborné články

Jan Holešovský, Michal Fusek

Metody analýzy extrémních hodnot a jejich softwarová implementace 1

Aneta Hybšová

Úroveň statistické gramotnosti studentů učitelství biologie 14

Zprávy a informace

Hana Řezanková, Prokop Závodský

Vzpomínky na Jiřího Žváčka 34

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server <http://www.vyzkum.cz/>.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

Předsedkyně společnosti: prof. Ing. Hana ŘEZANKOVÁ, CSc., KSTP FIS VŠE v Praze, nám. W. Churchilla 4, 130 67 Praha 3, e-mail: hana.rezankova@vse.cz.

Redakce: prof. RNDr. Gejza DOHNAL, CSc. (šéfredaktor), prof. RNDr. Jaromír ANTOCH, CSc., prof. Ing. Václav ČERMÁK, DrSc., doc. Ing. Jozef CHAJDIK, CSc., doc. RNDr. Zdeněk KARPÍŠEK, CSc., RNDr. Marek MALÝ, CSc., doc. RNDr. Jiří MICHÁLEK, CSc., prof. Ing. Jiří MILITKÝ, CSc., doc. Ing. Iveta STANKOVIČOVÁ, PhD., doc. Ing. Josef TVRDÍK, CSc., Mgr. Ondřej VENCÁLEK, Ph.D.

Redaktor časopisu: Mgr. Ondřej VENCÁLEK, Ph.D., ondrej.vencalek@upol.cz.
Informace pro autory jsou na stránkách společnosti, <http://www.statspol.cz/>.

DOI: 10.5300/IB, <http://dx.doi.org/10.5300/IB>
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.