P-VALUE ANO, ČI NE?

Roman Biskup

Katedra aplikované matematiky a informatiky, Zemědělská fakulta,
Jihočeská univerzita v Českých Budějovicích

ABSTRAKT

Tento článek se zabývá možnostmi rozhodování o výsledcích testování hypotéz v závislosti na zvoleném postupu a interpretaci hodnot, jež k tomuto slouží. Jsou zde diskutována úskalí jednotlivých přístupů, jak z hlediska pochopení látky, správného nastavení testu tak časově optimálního rozhodování.

KLÍČOVÁ SLOVA

Výuka statistiky, testování hypotéz, testové kritérium, kritický obor, hladina významnosti, *p-value*, jednostranný test, oboustranný test.

SUMMARY

The paper deals with possibilities of decision making about results of hypothesis testing dependent on the chosen methodology and values interpretation. Possible obstacles in particular approaches regarding the understanding of presented topics by students, the suitability of test application and the time-optimal decision-making regarding time is discussed in this contribution.

KEYWORDS

Teaching of Statistic, Hypothesis testing, Test Criterion, Rejection Area, Level of Significance, *p-value*, One-Tailed Test, Two Tailed Test.

Úvod

V základním kurzu Statistiky na Zemědělské fakultě Jihočeské univerzity v Českých Budějovicích (dále jen ZF) se standardně vyučuje testování hypotéz. Vzhledem k tomu, že se k výpočtům také používá statistický software, konkrétně STATISTICA komplet CZ 6.1, studenti nemusí dodržovat standardní postup testování hypotéz uváděný v klasických učebnicích statistiky. Studenti ale dost dobře nechápou invariatnost klasického postupu, operujícího s kritickým oborem a testovým kritériem, a interpretace hodnoty *p-value*.

Přestože výuka je vedena tak, aby byla zachována návaznost k teorii a pochopení vztahů, hodnocení výsledů testů je směřováno k interpretaci hodnoty *p-value*. Studenům pak připadá nesmyslné stanovování kritického oboru a počítání, či prosté přepisování hodnoty testového kritéria do výstupu práce, když nakonec rozhodují na základě, pro některé magické hodnoty – *p-value*. Hodnoty, jež ale někdy musí modifikovat.

Zhodnocení výsledků testů hypotéz

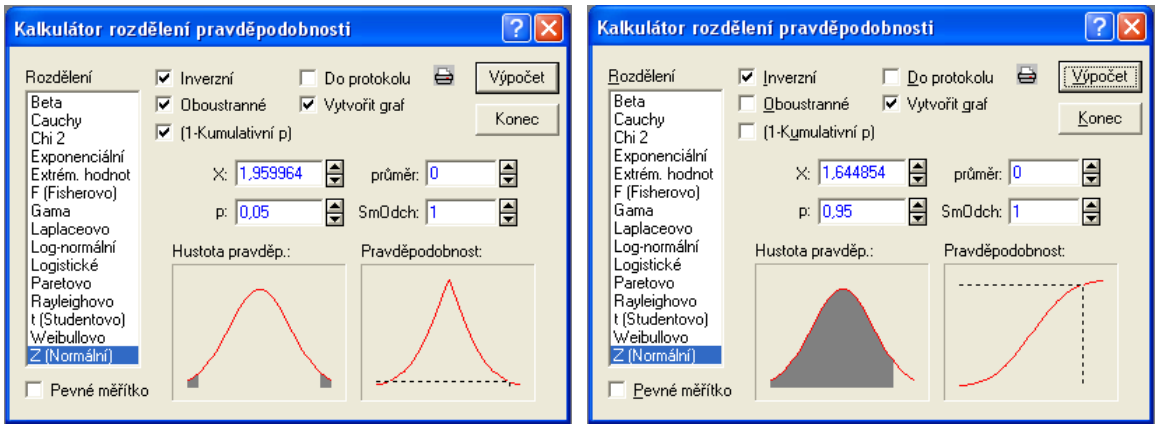
V současné době je bezesporu vhodné použít ve výuce statistiky software, neboť díky němu lze zpracovat reálná data. K vyhodnocení testování hypotéz lze přistoupit ve skrze dvojím způsobem. Zapomenout na plný komfort, jež nám software nabízí a využít jen dílčí výsledky

zahrnuté ve výstupní sestavě toho kterého statistického programu, nebo vytvořit vhodnou metodiku pro rozhodování o výsledku testu na základě interpretace jediné hodnoty, obvykle značené *p-value*. Výhody a nevýhody jednotlivých přístupů diskutují následující řádky.

Klasický postup

Pokud se do jisté míry oprostíme od výstupní tabulky, kterou nám obvykle nabídne statistický software, či vynecháme-li danou funkci/makro „automaticky řešící“ testování úplně, dostáváme se nejblíže výpočtovým prostředkům minulosti. Těmito prostředky, jimiž disponoval student statisticky v nedávné minulosti, jsou „tužka, papír, kalkulačka a statistické tabulky“. Přestože studenti na ZF jistě nepotřebují znát přímo matematickou statistiku, tento přístup jim pomáhá pochopit samotný princip testování, souvislost vzorců testových kritérií s předpisy pro intervaly spolehlivosti a konstrukci testového kritéria vůbec. Statistický software tak zajišťuje jen tu nejzákladnější výpočtovou činnost statistiků a student tak nemusí ztrácet čas výpočtem výběrových popisných charakteristik potřebných k vyčíslení toho, kterého testového kritéria.

Na ZF se tímto způsobem řeší alespoň několik příkladů právě kvůli výše zmíněným důvodům. Posléze je testování hypotéz prováděno s plnou podporou statistického softwaru na základě interpretace hodnoty *p-value*. Při těchto několika málo příkladech řešených „elementárně“, jež dotace na předmět a rozsah látky dovolují, je poslední příležitost jak studentům zafixovat testovací kritérium jakožto náhodnou veličinu. Náhodnou veličinu zkonstruovanou tak, aby za splnění daných předpokladů sledovala určité rozdělení. Je třeba si uvědomit, že studenti v současné době využívají software také k tomu, aby hledali hodnoty pravděpodobností a distribučních funkcí známých rozdělení. Málokterý z nich tudíž vstřebá to, co studentům v minulých dobách neunikalo. V dobách nedávno minulých studenti řešící například jednoduché úlohy z oblasti náhodných veličin museli používat v případě normálního rozdělení transformace na normální normované rozdělení a hodnoty distribučních funkcí respektive kvantilů hledat v tabulkách. Proto jim některé vzorce testových kritérií a srovnávání s kritickou hodnotou mohlo přijít apriori poměrně srozumitelné a logické.



Obrázek 1: Vizualizace vypočtených hodnot Pravděpodobnostním kalkulátorem

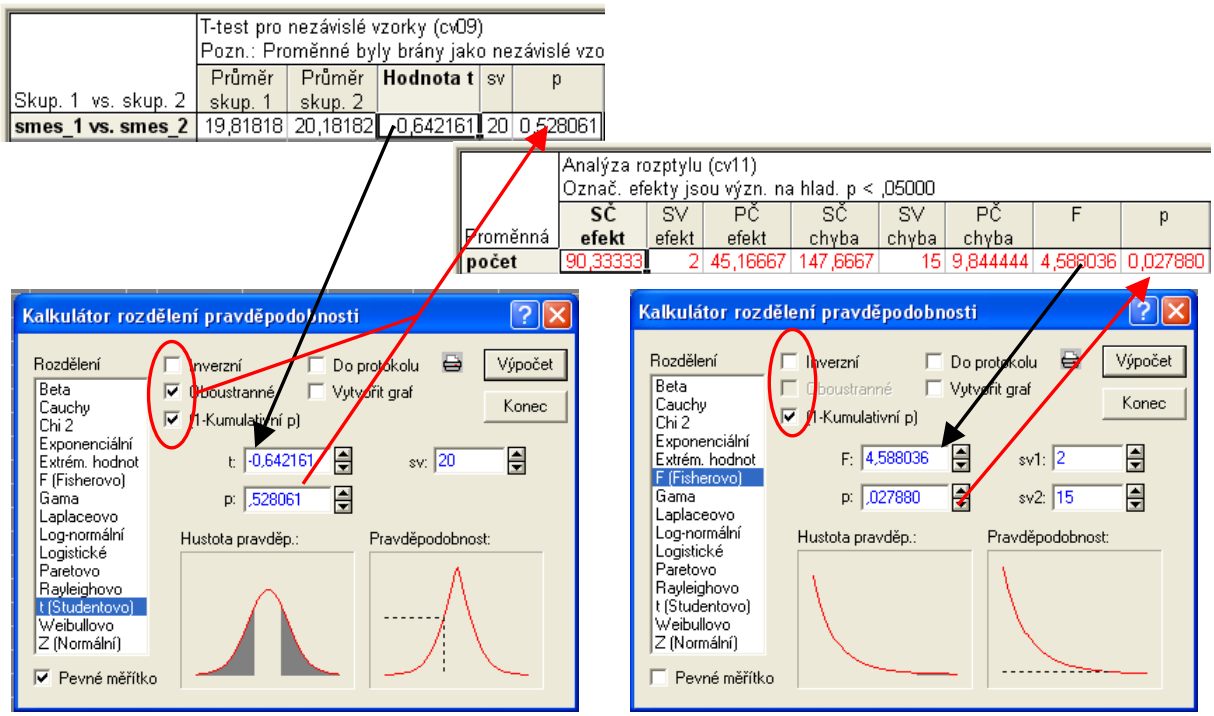
Vyžívají-li nynější studenti době například program STATISTICA komplet Cz 6.1 jedná se na této úrovni jen o ulehčení automatického vypočtu výběrových popisných charakteristik, maximálně spočtení hodnoty testového kritéria. Je třeba zmínit, že studenti na ZF již

nepoužívají statistické tabulky a hledají kritické hodnoty pomocí programu STATISTICA komplet Cz 6.1 funkcí „Pravděpodobnostní kalkulátor“. Hodnota *p-value*, jež je součástí výstupních tabulek, je v této fázi výuky zatím mlčky ignorována. Pravděpodobností kalkulátor umožňuje vedle prostého výpočtu kvantilu či distribuční funkce také vizualizaci těchto hodnot, což je velice názorné (viz Obrázek 1) a pro mnohé dostatečně ilustrativní. Ilustrativní natolik, aby pochopily nerovnosti v přepisech kritických oborů. Pravděpodobnostní kalkulátor navíc umožňuje přecházet od jednostranných k oboustranným kvantilům, atd.

Student při zachování klasického postupu testování hypotéz volí nulovou a alternativní hypotézu, hladinu významnosti, předpis testového kritéria, tvar kritického oboru a dále na základě výběru počítá hodnotu testového kritéria, hledá kritickou hodnotu (pro konstrukci kritického oboru) a na základě zjištěných číselných údajů rozhoduje o zamítnutí respektive nezamítnutí nulové hypotézy ve prospěch hypotézy alternativní (viz např. [1]). Pokud je tento přístup dostatečně vstřebán nebo zažit, nebývá následně problém studenta naučit vyhodnocovat výsledek testu na základě hodnoty *p-value* bez bezmyšlenkovitého používání pomocných pravidel, ale o tom dále.

Vyhodnocování na základě hodnoty *p-value*

Většina statistických softwarů ve svých výstupních formacích zobrazuje hodnotu obecně nazývanou *p-value*. Kdyby všechny hypotézy mohli mít jen jeden tvar (nikoliv jednostranné a oboustranné alternativy zároveň), tak by to studenti (na ZF určitě) ocenili. Program STATISTICA komplet Cz 6.1, jako ostatně většina statistických softwarů, v případě více možností hypotéz nabízí totiž pouze tu oboustrannou a jí příslušné *p-value*. Jak tedy bývá hodnota *p-value* studentům představována a jak je počítána statistickým softwarem?



Obrázek 2: Vizualizace vztahu testového kritéria a hodnoty *p-value*

S odpovědí na položenou otázku začněme u výpočetního backgroundu, jež využívá statistický software, konkrétně STATISTICA komplet Cz 6.1. Pokud existuje, nebo je používána pouze jedna možnost jak stanovit alternativní hypotézu (ANOVA, *F*-test, …) program provede na základě dat výpočet testového kritéria a na základě tvaru alternativní hypotézy určí takovou nejnižší hladinu významnosti, na níž je ještě možné zamítnout nulovou hypotézu. Tuto hodnotu označí za *p-value* (viz Obrázek 2 – vpravo). Pokud existují, nebo jsou používány jak jednostranné tak oboustranné alternativy pro testování hypotéz a tedy i alternativní hypotézu (vesměs varianty *t*-testů, …) program provede na základě dat výpočet testového kritéria a bez ohledu na alternativní hypotézu (jak by ji také mohl znát) určí nejnižší takovou hladinu významnosti, na níž je ještě možné zamítnout nulovou hypotézu pro oboustrannou alternativu testu. To znamená, že software vypočte pravděpodobnost s jakou by náhodná veličina sledující teoretické rozdělení nabyla hodnotu větší, než je absolutní hodnota z hodnoty testového kritéria a tuto hodnotu vynásobí dvěmi. Vynásobená hodnota je následně uživateli předložena jako *p-value* (viz Obrázek 2 - vlevo).

Studentům bývá *p-value* přibližována následovně. *P-value* poskytuje obecněji více informací o výsledku statistického testu. Předpokládejme, že *p-value* vyjde rovna 0,05. Z toho lze usoudit, že nulovou hypotézu lze zamítnout například na hladině významnosti $\alpha = 0{,}1$, ale již ne na hladině významnosti $\alpha = 0{,}01$ a $\alpha = 0{,}001$. Nejnižší možnou hladinu významnosti, na které ještě můžeme nulovou hypotézu zamítnout je právě $\alpha = 0{,}05$. Čím nižší vyjde *p-value*, tím více jsme přesvědčeni, že nulová hypotéza není správná a je třeba jí zamítnout. *P-value* je tedy nejnižší taková hodnota hladiny významnosti, na níž je ještě možné zamítnout nulovou hypotézu. [1]

Jaký je tedy rozdíl mezi hladinou významnosti a *p-value*? Podstata *p-value* a hladiny významnosti je v podstatě stejná. Hladina významnosti je maximální předpokládaná pravděpodobnost chyby zamítnutí nulové hypotézy za předpokladu, že byla správná. Hladinu významnosti určujeme předem. *P-value* je taková nejnižší možná pravděpodobnost chyby pro zamítnutí nulové hypotézy určená na základě hodnoty testového kritéria (tj. na základě výsledků výběru). Z tohoto důvodu obě hodnoty označujeme různými symboly – α a *p-value*.

Nejjednodušším způsobem, jak rozhodovat o výsledku testu spočívá v porovnání *p-value* (vypočte statistický software) a hladiny významnosti (určíme před testem sami). Platí následující pravidlo, které na základě srovnání hladiny významnosti a hodnoty *p-value* rozhoduje o zamítnutí respektive nezamítnutí nulové hypotézy:

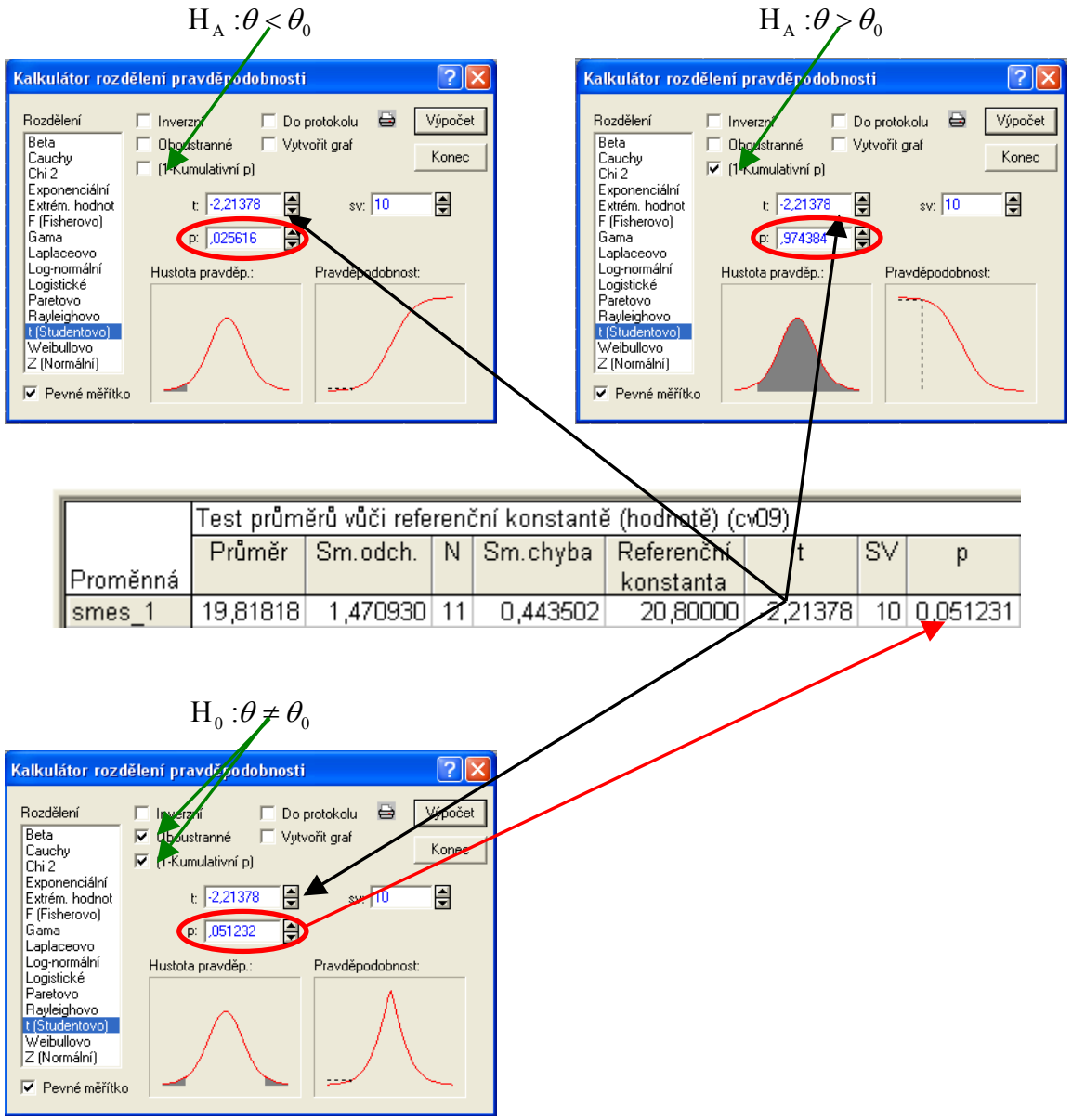
---

[1] Definice *p*-value jak ji uvádí materiály k programu STATISTICA komplet Cz 6.1 [2]:

Statistical Significance (p-level). The statistical significance of a result is an estimated measure of the degree to which it is "true" (in the sense of "representative of the population"). More technically, the value of the *p-level* represents a decreasing index of the reliability of a result. The higher the *p-level*, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the p-level represents the probability of error that is involved in accepting our observed result as valid, that is, as "representative of the population." For example, the *p-level* of .05 (i.e.,1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is a "fluke." In other words, assuming that in the population there was no relation between those variables whatsoever, and we were repeating experiments like ours one after another, we could expect that approximately in every 20 replications of the experiment there would be one in which the relation between the variables in question would be equal or stronger than in ours. In many areas of research, the *p-level* of .05 is customarily treated as a "border-line acceptable" error level.

- Zamítni nulovou hypotézu, když je *p-value* menší nebo rovno zvolené hladině významnosti.
- Nezamítej nulovou hypotézu, když je *p-value* větší než zvolená hladina významnosti.

Vše vypadá sympaticky a jednoduše, ale jak bylo zmíněno výše oni nejsou jen oboustranné respektive přesně určené varianty testů. Nicméně k hodnotě *p-value* příslušné jednostranné alternativě testování hypotéz se lze dostat jednoduchou úvahou. Pro studenta, který úplně nepochopil sestavování testového kritéria a způsobu výpočtu hodnoty *p-value* softwarem, to přináší pouze další pravidla, které nechápe jako alternativy ke klasickému způsobu rozhodování na základě padnutí nebo nepadnutí testového kritéria do kritického oboru. Pro studenta, jež má znát tedy jak teorii i svižně počítat příklady, to pak někdy znamená, že se učí dva postupy v nichž nevidí souvislost.



Obrázek 3: Výpočet hodnoty *p-value* za pomoci Pravděpodobnostního kalkulátoru

Učitelům navíc nezbývá nic jiného než formulovat pravidla pro modifikaci hodnoty *p-value* tak, aby to byla ta požadovaná (příslušná jednostranným alternativám) hodnota. Obsah pravidel se většinou různí podle učitelů. Někteří vytvářejí přesné postupy jak na základě tvaru alternativní hypotézy editovat data do programu a následně modifikovat *p-value* vypočtené softwarem. Jiní předkládají, jak na základě alternativní hypotézy a znaménka hodnoty testového kritéria rozhodnout, co dále a jak hodnotu *p-value* modifikovat, aby se studenti navzdory „oboustranné" interpretaci testu softwarem dostali k tomu co chtějí. V prvním případě je hodnota, o kterou stojíme jako o výslednou, většinou přímo rovna polovině hodnoty *p-value* vypočtené softwarem. V druhém případě se hodnota vypočtená softwarem na základě vztahu mezi alternativní hypotézou a znaménkem testového kritéria buďto dělí dvěmi, nebo dělí dvěmi a odečítá od jedné.

Vhodný mezikrok

Určitým mezikrokem mezi oběma metodami se jeví využití vnitřních poměrně elementárních prostředků daného softwaru k výpočtu přímo hodnoty *p-value*. Konkrétně program STATISTICA komplet Cz 6.1 nabízí možnost využití Pravděpodobnostního kalkulátoru a na základě hodnoty testového kritéria a tvaru alternativní hypotézy přímo určit hodnotu *p-value*. Tento postup „testování" je o něco méně komfortní, nicméně odstraňuje do značné míry z výpočtu hodnoty *p-value* punc černé skříňky. Obrázek 3 tento postup zobrazuje. Vedle vypočtení jednotlivých hodnot *p-value* student jasně vidí princip výpočtu *p-value* softwarem a lépe tak chápe pravidla, která operují s tvary alternativních hypotéz, tu „dělením dvěmi, tu dělením dvěmi a odečítáním od jedné".

Závěr

Z výše napsaného by mělo zřejmé, že učitelé statistiky (a nejen oni) se jistě musí cítit schizofrenně. Na jedné straně vyučují teorii, která neslouží jen k tomu, aby studenti dokázali vyhodnotit například svůj pokus, ale také k tomu, aby získali algoritmický a systémový přístup ke zpracování dat a výzkumu jako takového. Na straně druhé je zřejmým požadavkem na výuku statistiky, předložit studentům prostředek jak efektivně, tedy za využití softwaru, zpracovávat datové soubory. Statistika však není na školách typu ZF profilovým předmětem a proto je bohužel na okraji zájmů nejen studentů, kteří ze zřejmých důvodů nemohou tuto obsažnou látku optimálně vstřebat a statistický software jim v tomto přidává práci.

Pokud se vrátíme zpět k tématu diskutovaném v příspěvku je patrno, že látka může být podána jedním, druhým nebo oběma způsoby. První způsob je vhodnější pro lepší procvičení a pochopení/vstřebání teorie a vytvoření užitečných návyků pro práci s daty obecně. Jeho nevýhodou je zřejmá výpočetní náročnost, nerespektování trendů ve výuce, které směřují k využití výpočetní techniky všude, kde je to jen možné. Druhý způsob přináší studentům možnost jak efektivně (z pohledu metodologie nikoliv nutně správně) zpracovat data přesně – řekněme numericky bez chyby. Na druhé straně vytváří ze studenta člověka, který spoléhá na software, černou skříňku, jejíž princip fungování nezná a tím pádem jí musí bezmezně věřit. Ideálním se proto jeví kombinace obou přístupů, které ale kladou zvýšené nároky na studenta jak po straně technicko-informatické zručnosti tak po straně znalostní.

Mou odpovědí na vyzývavou otázku z názvu tohoto příspěvku je *p-value* ano i ne. Výuka by měla, dle mého názoru, studenty dovést k hodnotě *p-value* jako k hodnotě, kterou si sami

umějí vypočítat (viz „Vhodný mezikrok“) a jež skutečně podává více informací, než pouhé zhodnocení výsledku testu na základě padnutí respektive nepadnutí testového kritéria do kritického oboru.

Literatura

[1] Čermáková, A., Střeleček, F.: Statistika I. České Budějovice, 1995. skripta ZF JU, ISBN 80-7040-126-5.

[2] Elektronická učebnice k programu STATISTICA

Kontakt

Roman Biskup, Mgr., Jihočeská univerzita v Českých Budějovicích, Zemědělská fakulta, Katedra aplikované matematiky a informatiky, Studentská 13, 387 05 České Budějovice, biskup@zf.jcu.cz