Rozhledy matematicko-fyzikální

Ondřej Vencálek

Příklad do hodiny věnované statistice

Rozhledy matematicko-fyzikální, Vol. 94 (2019), No. 3, 1–8

Persistent URL: http://dml.cz/dmlcz/147890

Příklad do hodiny věnované statistice

Ondřej Vencálek, PřF UP, Olomouc

1. Úvod

Dal jsem si následující úkol: prostřednictvím krátké (maximálně 45minutové) přednášky motivovat studenty středních škol ke studiu statistiky. Motivovat někoho ke studiu určitého předmětu podle mého znamená nejprve odpovědět na tradiční otázku „k čemu je, či může být, tento předmět dobrý?“, a poté ukázat, že daný předmět je nejen „užitečný“, ale také „zábavný“ – že úsilí vložené do studia je vyváženo radostnými pocity v situacích, kdy člověk přijde na něco, co se zprvu zdálo nejasné. Studenti, kteří mají v oblibě matematiku (především pro ně je přednáška určena), ten radostný, snad až objevitelský pocit dobře znají. Nicméně i ti, kteří neocení krásy statistiky jako matematické disciplíny, mohou ve statistice objevit poměrně univerzálního pomocníka při hledání odpovědí na otázky z nejrůznějších oborů lidské činnosti.

2. Otázky – data – analýzy – odpovědi

První část motivační přednášky má studenty přesvědčit o „užitečnosti“ statistiky.

Lidé pracující v různých oborech si kladou velice rozmanité otázky:

- „Který výrobek má být umístěn na první straně reklamního letáku, aby přilákal co nejvíce zákazníků?“ ptá se zaměstnanec firmy vyrábějící kosmetiku zodpovědný za tvorbu reklamních letáků.
- „Jak stanovit kritéria, která musí splňovat žadatel o hypotéku, tak, aby hypotéka byla poskytována pokud možno jen těm klientům, kteří ji budou schopni bezproblémově splácet?“ ptá se ředitel banky.
- „Které úseky silnic jsou nejnebezpečnější a bylo by dobré na to řidiče upozornit výstražnou cedulí?“ ptají se dopravní policisté.
- „Který ze tří dostupných typů léčby je nejvhodnější pro pacienty trpící určitou nemocí?“ ptají se výzkumníci–lékaři.
- „Která ze dvou různých výukových metod matematiky je vhodnější pro talentované žáky na prvním stupni ZŠ?“ ptá se ředitel školy.

Každá z těchto otázek s sebou většinou nese ještě mnoho dalších otázek a odpovědi nebývají jednoduché. Všechny výše uvedené příklady mají jedno společné – k jejich zodpovězení je třeba hodně zkušeností. Například školní inspektor, který měl možnost pozorovat a porovnávat výsledky žáků učených různými metodami, bude mít o vhodnosti metod pro určitou skupinu žáků dobrou představu. Zkušenost vzniká na základě mnoha pozorování, která si člověk shrnuje do obecných pravidel, např. „Metoda H je lepší pro všechny žáky bez rozdílu.“ nebo „Metoda H je lepší pouze pro žáky s vyšším IQ, pro ostatní je lepší druhá metoda.“ Proces získávání zkušeností můžeme podpořit systematickým zaznamenáváním svých pozorování – sběrem dat – a následnou analýzou těchto záznamů (dat) pomocí vhodných statistických metod. Snažíme se vlastně ve velké spoustě nasbíraných dat najít jednoduchá pravidla, která by poskytla odpovědi na námi kladené otázky.

Zápis na tabuli z této části je vyjádřen schématem:

otázky – data – analýzy – odpovědi

3. Příklad

Ukážeme si, jak taková statistická analýza může vypadat. Z časových důvodů se omezíme jen na opravdu jednoduché typy analýz. Pokud je to jen trochu možné, je dobré, aby studenti mohli vše sami vyzkoušet, což je možné, pokud se ukázková hodina odehrává v počítačové učebně, kde je k dispozici program Microsoft Excel (případně podobný software). Excel sice nepředstavuje špičku mezi softwarovými nástroji pro analýzu dat, ale zato je pro většinu studentů dostupný a mnozí jsou schopni ho velmi dobře používat. Data společně s níže popsanými analýzami jsou k dispozici na `https://rozhledy.jcmf.cz/wp-content/uploads/fev1.xlsx`. Studenti by v ideálním případě měli dostat k dispozici verzi obsahující pouze data.

3.1. Otázka, kterou si klademe

Nyní si ukážeme použití statistiky na příkladu z oblasti lékařského výzkumu. Dnes již víme mnoho o vlivu kouření na lidské zdraví. Je téměř neuvěřitelné, že první přesvědčivé důkazy o tom, že kouření způsobuje rakovinu plic, přinesli až v 50. letech 20. století Richard Doll a Tony Bradford Hill [1, 2]. Diskuse o vlivu tzv. pasivního kouření na lidské zdraví je aktuální stále. Na konci 70. let 20. století probíhal jeden z prvních výzkumů na toto téma. Část dat z této studie je k dispozici na webu

časopisu Journal of Statistics Education[1)]. Do studie byly zahrnuty děti a mládež z rodin, v nichž alespoň jeden z rodičů byl kuřák. Část z těchto mladistvých už sama měla zkušenost s kouřením (uvedli o sobě, že kouří, v dotazníku vyplněném v nepřítomnosti rodičů). Otázka, kterou si tehdy vědci kladli, zní:

Mají děti/mladiství, jejichž otec či matka kouří, horší funkci plic, pokud sami kouří, než jejich vrstevníci ve stejné rodinné situaci, kteří však sami nekouří?

Při snaze nalézt odpověď na tuto otázku pomocí analýzy výše uvedených dat se setkáváme s celou řadou problémů popsaných v článku [3]. Nejprve je třeba vysvětlit, jak vlastně změřit „fungování plic“. Jedním ze způsobů, jak to můžeme udělat, je pomocí tzv. spirometru. Tento přístroj umožní například změřit objem vzduchu (vyjádřený v litrech) vydechnutého během první sekundy „usilovného“ výdechu. Tato veličina se označuje FEV1 (Forced Expiratory Volume). Ukázka dat je uvedena v tabulce 1.

id	věk	výška (cm)	pohlaví	kouření	FEV1 (litry)
1	9	145	*Ž*	*0*	1,708
2	8	171	*Ž*	*0*	1,724
3	7	138	*Ž*	*0*	1,720
4	9	135	*M*	*0*	1,558
...					
653	16	160	*Ž*	*1*	2,795
654	15	169	*Ž*	*0*	3,211

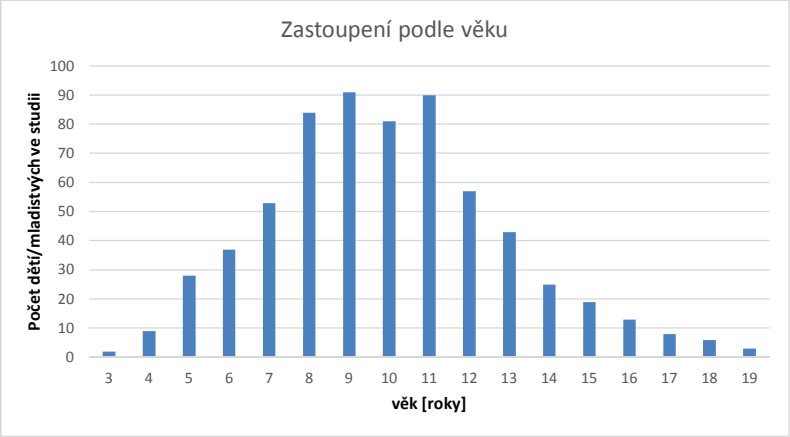
Tabulka 1: ukázka analyzovaných dat

Sloupec *id* je identifikačním číslem jedince v této studii (celkem jich je 654), sloupec *věk* obsahuje údaj o věku vyjádřený v letech, *výška* je udávaná v centimetrech, sloupec *pohlaví* obsahuje buďto symbol „Ž“, pokud jde o ženu, nebo symbol „M“, pokud jde o muže. Ve sloupci *kouření* se vyskytuje jedna z hodnot 1 (pro kuřáky) a 0 (pro nekuřáky). Veličina v posledním sloupci – *FEV1* – nás zajímá především, je měřená v litrech.

[1)] `http://jse.amstat.org/jse_data_archive.htm`

3.2. Seznámení se s daty

Analýza téměř každé datové sady by měla začínat popisnou statistikou – jednoduchými souhrny jednotlivých veličin v datech. U *kvantitativních veličin*, jako je věk, výška či FEV1, můžeme vypočítat průměrné hodnoty (v excelu pomocí příkazu `PRŮMĚR`), u *kvalitativních veličin* je vhodným souhrnem četnostní tabulka, kterou v excelu můžeme získat pomocí příkazu `COUNTIF`. Zjistíme tak například, že z 654 dětí a mladistvých jich kouřilo 65, tedy necelých 10 %. Zastavme se ještě na okamžik u veličiny věk, jejíž povaha je kvantitativní (číselná), avšak udávaných hodnot je relativně málo (věk je uváděn celočíselně). Věk účastníků studie se pohyboval v rozmezí 3 až 19 let (s průměrnou hodnotou přibližně 10 let). Detailnější představu o věkovém složení skupiny účastníků studie si můžeme udělat opět pomocí zjištění četností jednotlivých věkových skupin (zjistíme, že nejvíce zastoupeny jsou věkové skupiny v rozmezí 8–11 let), jejichž hodnoty pak můžeme graficky znázornit pomocí sloupcového grafu, viz obr. 1.



Obr. 1: Zastoupení jednotlivých věkových skupin mezi účastníky studie

Je třeba přiznat, že tvorba této četnostní tabulky a následného grafu by studentům zřejmě zabrala více času, než umožňuje předem stanovený časový rámec, a je proto dobré je pouze seznámit s předem připraveným výsledným grafem. Pokud bychom analýze věnovali třeba dvě vyučovací hodiny, mohli bychom tvorbu grafu nechat na studentech.

3.3. První pokus o zodpovězení otázky

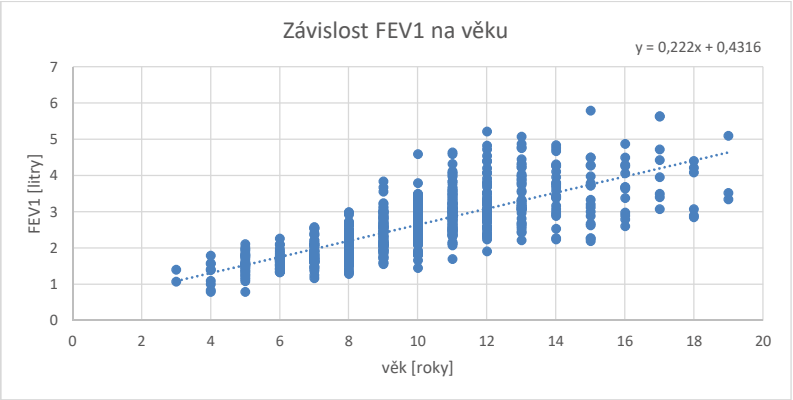
Dali jsme si za cíl zodpovědět otázku, která se týká vztahu veličin kouření a FEV1. Zajímá nás vlastně, jestli hodnoty veličiny FEV1 jsou jiné pro kuřáky než pro nekuřáky. Pokud mají pravdu ti, kteří tvrdí, že kouření škodí zdraví (a to je dnes převládající názor), mohlo by se to projevit třeba zhoršením fungování plic, tedy očekávali bychom nižší hodnoty veličiny FEV1 u kuřáků než u nekuřáků. Je tomu opravdu tak? Jak to zjistíme? Můžeme třeba spočítat průměrnou hodnotu FEV1 pro kuřáky a průměrnou hodnotu FEV1 pro nekuřáky. Je několik způsobů, jak to udělat. Uživatelé excelu většinou umí z dat vyfiltrovat jen určitou část, pro niž pak mohou spočítat průměr. Elegantnějším způsobem je však použití funkce `AVERAGEIF`. Pomocí ní snadno zjistíme, že průměrná hodnota FEV1 pro kuřáky je 3,28 litru, zatímco pro nekuřáky jen 2,57 litru. To je ovšem výsledek opačný, než bychom čekali! A zde je prostor zeptat se překvapených studentů, jak si takový výsledek vysvětlují. Možná se začnou pouštět do spekulací, jestli tedy kouřením vlastně plíce netrénujeme a nezvětšujeme tak jejich „kapacitu“. Možná si také uvědomí, že větší či menší hodnoty FEV1 v důsledku kouření nutně neznamenají, že kouření způsobuje nějakou nemoc. Pro zjištěný výsledek však existuje jednoduché plausibilní vysvětlení – kapacita plic úzce souvisí s fyziologickými rozměry dítěte, a tedy i s jeho věkem. Čím větší (starší) dítě/mladistvý je, tím větší hodnotu FEV1 můžeme očekávat. Kouření se přitom týká spíše starších účastníků studie.

Nyní je taky patrné, jak opatrní musíme být při formulaci otázky, na kterou se ptáme. Vraťme se k ní ještě jednou. Zajímá nás, zda *mají děti/mladiství, jejichž otec či matka kouří, horší fungování plic, pokud sami kouří, než jejich vrstevníci ve stejné rodinné situaci, kteří však sami nekouří?* Kdybychom ve formulaci otázky vynechali vyjádření specifikující, že jde o *vrstevníky*, tedy že chceme porovnávat kapacitu plic u stejně starých kuřáků a nekuřáků, dostali bychom sice správnou, ale zcela zavádějící odpověď, že kuřáci mají větší hodnoty FEV1 (větší kapacitu plic). Poznamenejme ještě, že z formulace otázky je také patrné to, že námi učiněné závěry se budou týkat pouze dětí/mladistvých, jejichž otec či matka kouří (o dětech nekuřáků nemáme v datech žádnou informaci).

3.4. Jak dojít k správné odpovědi

Zatím jsme si ukázali, jak ošidné mohou být závěry, pokud si analýzu příliš zjednodušíme. Jak ale udělat analýzu správně? Především se vraťme k tvrzení, že hodnoty FEV1 rostou s věkem dítěte/mladistvého.

Dá se tato skutečnost nějak doložit údaji v námi analyzovaných datech? Pomocí bodového grafu, v němž x-ové souřadnice jednotlivých bodů jsou určeny věkem a y-ové souřadnice hodnotami FEV1 jednotlivých dětí, můžeme ukázat, že u starších dětí jsou hodnoty FEV1 vskutku vyšší, viz obr. 2.



Obr. 2: Závislost hodnot FEV1 na věku

Otázkou je, jak tuto závislost popsat. Omezíme se na velmi jednoduchý model lineární závislosti hodnot FEV1 na věku. Graficky si to můžeme představit tak, že body proložíme přímkou (viz obr. 2). Tuto přímku popíšeme rovnicí

$$y = ax + b,$$

kde y jsou očekávané hodnoty FEV1 odpovídající věku x. Parametry přímky a a b jsou pro nás neznámé, můžeme je však odhadnout z dat (a to takzvanou metodou nejmenších čtverců, jejíž princip však není nutno nyní objasňovat). Odhady parametrů a a b zjistíme v excelu tak, že klikneme na body v bodovém grafu pravým tlačítkem a zvolíme možnost `Přidat spojnici trendu`. Ponecháme lineární tvar trendu a zaškrtneme možnost `Zobrazit rovnici v grafu`. Tím získáme rovnici přímky s již konkrétními odhady parametrů a a b, která má v našem případě podobu

$$y = 0{,}22x + 0{,}43.$$

Vidíme tedy, že jeden rok věku navíc znamená v průměru přibližně o 0,2 litru větší hodnoty FEV1.

Je nanejvýš důležité nespokojit se jen s výpočtem odhadu parametrů, ale vypočtené hodnoty také interpretovat tak, jako jsme to právě nyní udělali s hodnotou odhadu parametru a (0,22). Poznamenejme, že parametr b, jehož hodnotu jsme odhadli na přibližně 0,43 (litru), má význam očekávané hodnoty FEV1 při nulové hodnotě x, tedy „při narození“. K této hodnotě bychom opravdu dospěli, kdybychom graf přímky „protáhli“ tak, abychom našli průsečík se svislou osou ($x = 0$). Excel vykreslil jen část přímky odpovídající hodnotám x v rozmezí, v němž máme data v našem souboru (3 až 19 let). Možná je to dobrá připomínka toho, že bychom měli být velmi opatrní při „extrapolaci“ našich závěrů mimo toto rozmezí, resp. že bychom se takovýchto extrapolací měli raději vyvarovat – ve skutečnosti o hodnotách FEV1 při narození nevíme z námi analyzovaných dat vůbec nic.

Parametry přímky můžeme získat také pomocí funkce `LINREGRESE` z nabídky excelu `Vzorce` – `Vložit funkci`. Protože jsou výstupem této funkce dvě čísla – odhady parametrů a a b, musíme nejprve označit dvě buňky, do kterých má být výsledek zapsán, poté vložit výše uvedenou funkci `LINREGRESE`, kde za y dosazujeme hodnoty ze sloupce FEV1 a za x hodnoty ze sloupce věk. Vložení funkce však potvrdíme nikoliv obvyklým zmáčknutím klávesy Enter, ale zmáčknutím kláves Ctrl+Shift+Enter, které se používá v situaci, kdy vkládáme vzorec do vícero buněk najednou (v našem případě jsou dvě).

Z předchozích úvah je patrné, že pro správné zodpovězení námi položené otázky musíme ve svých analýzách zohledňovat nejen to, zda daný jedinec kouří, ale také, jakého je věku. Nejjednodušší model předpokládá, že hodnoty FEV1 se u kuřáků a nekuřáků liší o konstantu, která se s věkem nemění, tedy že závislost FEV1 na věku je stejná u kuřáků jako u nekuřáků. Tento model zapíšeme následovně:

$$y = ax + b, \quad \text{pro nekuřáky},$$

$$y = ax + c, \quad \text{pro kuřáky}.$$

Rozdíl mezi stejně starými nekuřáky a kuřáky je v tomto případě

$$(ax + b) - (ax + c) = b - c,$$

nezávisí tedy na x (na věku). Směrnice přímky udávající závislost FEV1 na věku je pro kuřáky stejná jako pro nekuřáky (je rovna a).

Výše uvedený model můžeme zapsat také následovně:

$$y = ax + (c - b)z + b,$$

kde z je veličina nabývající hodnoty 1 pro kuřáky a 0 pro nekuřáky. Hodnoty veličiny z máme v datech ve sloupci `kouření`. Odhady parametrů v tomto modelu (a, b a $c - b$) získáme snadno opět použitím funkce `LINREGRESE`. Protože však máme parametry tři, musíme označit tři buňky, kam budeme vkládat výsledek. Jako `pole_x` tentokrát označíme hodnoty ve sloupcích věk a kouření. Nakonec opět nezapomeneme potvrdit zadání kombinací kláves Ctrl+Shift+Enter.

Odhady parametrů, které takto získáme, jsou 0,23 pro parametr a (výsledek je tedy obdobný tomu, který jsme zjistili v předchozí analýze), 0,37 pro parametr b a $-0{,}21$ pro rozdíl $c - b$.

Odhadli jsme tedy, že při porovnání dětí/mladistvých stejného věku, jejichž rodiče kouří, mají kuřáci přibližně o 0,21 litru nižší hodnoty FEV1 než nekuřáci.

Tento výsledek odpovídá našemu očekávání negativního vlivu kouření na fungování plic. Ukazuje, že i v prostředí, kdy jsou děti/mladiství vystaveni pasivnímu kouření, je funkce jejich plic zhoršena jejich vlastním kouřením.

Zde naši krátkou exkurzi do světa statistiky ukončíme. Možná, že v průběhu hodiny vyvstanou další otázky, kterými by se bylo potřeba zabývat. Statistika se zabývá tím, nakolik můžeme jevy pozorované na daném vzorku zobecňovat. K tomu je však zapotřebí poněkud delšího výkladu...

Literatura

[1] Doll, R., Hill, A. B.: The mortality of doctors in relation to their smoking habits: a preliminary report. *British medical journal*, roč. 228 (1954), s. 1451–1455.

[2] Hutchinson, E.: Smoking gun. *Nature Milestones Cancer*, roč. 1 (2006), č. 1, s. S12–S12.

[3] Kahn, M.: An exhalent problem for teaching statistics. *Journal of Statistics Education*, roč. 13 (2005), č. 2, doi: 10.1080/10691898.2005.11910559.