Data a jejich role v období epidemie

„Statistika“ jako alibi pro (špatná) rozhodnutí nebo statistika jako prostředek poznání reality?

Arnošt Komárek

Katedra pravděpodobnosti a matematické statistiky, MFF UK

> Motto (Simon N. Wood): *Statistics is about the honest interpretation of data. It's not always a popular subject: honest interpretation of data is difficult, and muchless appealing than less honest interpretation.*
>
> *Something that the Covid-19 crisis has emphasised is the fact that statisticians have not managed to adequately communicate how fundamental random sampling is to proper measurement of things like infection rates. That data* ***somehow related*** *to the thing we want to measure* ***are not the same*** *as data that actually measure it.*

Data jsou klíčová nejenom v období epidemie. V únoru 2021 si přitom *Česká statistická společnost* ústy své odborné skupiny posteskla, že v České republice jsme v rámci epidemie COVID-19 **zcela selhali v procesu poznání založeném na datech** (tento povzdech platí 100% i nadále, jenom odborná skupina ČStS již rezignovala s tím něco dělat). Jedním z problémů je fakt, že v ČR máme k epidemii k dispozici téměř výhradně toliko *observační data*. Taková data nejsou bezcenná, ale s ohledem na svoji *nereprezentativnost* pro populaci zájmu i *nekonzistenci* (vypovídací schopnost se mění v čase, s místem, ...) jsou extrémně náchylná k **dezinterpretacím** či kreativnímu výkladu, pomocí něhož lze zdůvodnit téměř jakékoliv rozhodnutí. Příkladem je reportovaný denní počet *pozitivně otestovaných*. Tento je Ministerstvem zdravotnictví (MZ) mylně považován za počet nově *nakažených*, je z něho počítána jakási *incidence* a na základě změny tohoto „incidenčního čísla“ je rozhodováno o změnách v epidemických opatřeních, resp. jsou jím zdůvodňována mimořádná opatření MZ. Ministerské incidenční číslo přitom nijak neodráží epidemickou situaci, jako spíš to, *kolik* lidí a z *jakých skupin* aktuálně testujeme. A potřebujeme-li navýšit „incidenci“ v jedné skupině, není nic jednoduššího, než úředně donutit osoby z této skupiny se ve větší míře testovat, že ANO pane ministře. Absurditu ministerské incidence ilustruje např. tento výtah z denního „tisku“ (idnes.cz, 9.9.2021):

> *„Statistiky nově nakažených naopak zcela zaplňují momentálně lidé do 30 až 40 let, přičemž díky plošným testům ve školách jich třetinu tvoří děti. Protože po čtvrtku se pokračovat celorepublikově v testování zatím nemá, jejich poměr patrně klesne. A s tím i celkový počet nově evidovaných jako nakažení.“*

ANO, `idnes.cz` připouští, že s koncem testování ve školách klesne podíl nakažených dětí v populaci. ANO, stačí přestat testovat a epidemie je pryč! V rámci přednášky ukážeme, vesměs na volně dostupných prezentacích *Ústavu zdravotnických informací a statistiky* několik dalších příkladů kreativní (dez)interpretace informace obsažené v observačních datech. Nebo se s ohledem na časové omezení přednášky budeme věnovat spíše něčemu zajímavějšímu?

Dalším horkým tématem, který souvisí s daty, je tzv. *matematické modelování* a jeho využití při „řízení“ epidemie (např. Brauner a kol., 2021). Matematické modelování má svoji nezastupitelnou roli ve fyzice, kde úspěšně slouží k i *dlouhodobému* předpovídání chování jednodušších systémů neživé přírody, jejichž mechanizmům dobře rozumíme (proudění za jasně daných podmínek, ...), též ke *krátkodobému* předpovídání chování i poměrně složitých systémů neživé přírody, jejichž mechanizmům relativně dobře rozumíme

(počasí, . . .). V průběhu epidemie COVID-19 se nicméně mnozí „modeláři“[1] začali pokoušet o modelování chování *značně složitého* systému *živé* přírody (interakce viru a lidské společnosti), kterému navíc ani po roce a půl příliš nerozumíme. A kromě věštění budoucnosti se COVIDoví modeláři pustili též do předvídání *„Co se stane, když ministr Hamáček (ne)uzavře obyvatele v jejich katastrech.“*, tedy vlastně do statistické kauzální analýzy. Zde se hlavním zaklínadlem a téměř všeovlivňující „řídící proměnnou“ stala *intenzita kontaktů*, viz např. Šmíd a kol. (2021). Autoři (i mnohých dalších „modelů“) jaksi pozapomněli na pojmy dobře známé statistikům, jako je *confounding*, *overfitting*, *validace modelu*, rozdíl mezi *korelací* a *kauzalitou*, . . . V přednášce vysvětlíme význam některých z těchto pojmů a s tím spojené záludnosti v kontextu COVIDového modelování. I když není to vlastně jedno? Významná část COVIDových modelářů si totiž nechce připustit ještě základnější předpoklad smyslupnosti svého konání. Pokud chceme modelovat realitu, potřebujeme v první řadě data, jež odrážejí tuto realitu, tedy data *reprezentativní* a *konzistentní* (Kulich, 2021). Pokud taková data nemáme, je možná lepší nemít žádný model (a mlčet) než špatný (a nebezpečný) model a poskytovat politikům vítané alibi pro jejich rozhodnutí. Ale každému co jeho jest.

Sběr relevantních dat o epidemii COVID-19 (či čemkoliv jiném) dá práci, chvíli to trvá, něco to stojí a hlavně to vyžaduje plánování i znalosti. Ani ve světě není situace na tomto poli úplně růžová, i když mnohde zdaleka ne tak tristní jako v České republice. Britský profesor statistiky Simon Wood[2] si proto položil otázku, jak by bylo možné s pouze observačními daty vyhodnotit *validním* způsobem[3] efektivitu tzv. *nefarmakologických intervencí* (≡ vládních opatření) v boji s epidemií. Simon Wood využívá data o počtu úmrtí v souvislosti s COVID-19, jež lze alespoň v rámci jednoho státu považovat za reprezentativní i konzistentní (resp. mezi běžně dostupnými observačními daty nejlepší možné přiblížení tomuto ideálu) a navrhuje stochastický model (Wood, 2021), jehož primárním cílem je odhadnout kolik osob z těch, kdo zemřeli, se ten který den nakazilo, tzv. *fatální infekce*, které by nás měly trápit nejvíce (na rozdíl od infekcí dětí, jež se často vůbec nedozvědí, že byly infikovány). Zpětným srovnáním s chronologií nefarmakologických intervencí, resp. jinými událostmi, u nichž se předpokládal „efekt“ lze následně zjistit, které z těchto infekcí mohly mít skutečný efekt na dynamiku šíření *fatálních* infekcí a které spíše nikoliv. Prakticky jediným předpokladem, který Wood činí, je předpokládané parametrické rozdělení času mezi infekcí a smrtí a dále se musí vypořádat s omezenou dostupností dat pro odhad tohoto rozdělení. Jinak řečeno, Woodův přístup pouze stochasticky posouvá zemřelé do okamžiku, kdy se nakazili, informaci o nefarmakologických opatřeních nevyužívá (a tudíž o jejich efektu nemusí ani nic předpokládat) a až následně kouká, zda v okolí dne, kdy bylo zavedeno to které opatření, nastala změna v dynamice šíření viru či nikoliv. „Kvalitu“ posunu lze přitom ověřit běžnými postupy kontroly modelu. Wood svůj přístup používá k odhadu dynamiky epidemie ve Velké Británii a Švédsku (do ledna 2021) a dochází k závěru, že v obou těchto státech se průběh epidemie, navzdory značně rozdílnému přístupu jednotlivých vlád, příliš nelišil (až na jisté „fázové“ posuny). Na své webové stránce výsledky lapidárně komentuje takto:

> *Sweden, which did not lockdown, saw declining infections only a day or two after the UK (Swedish GDP dropped about 3% in 2020 against the UK's 10% drop).*

K čemu by Wood dospěl v České republice víme díky Robertu Strakovi s drobným přispěním autora tohoto textu (Straka and Komárek, 2021). Co mohlo způsobit jeden z nejvyšších počtů úmrtí v souvislosti s COVID-19? Opravdu jde o náhodu, že počet mrtvých koreluje *kladně* s délkou uzavření škol? Štěkal PES, když měl štěkat a byl naopak zalezlý v boudě, když bylo nebezpečí na ústupu? Nebo jsme tu měli spíše reverzního PESa, který štěká na mírumilovného souseda a před zlodějem zaleze do boudy? Měla smysl blokáda okresních hranic?

[1] Někteří z nich v sobě objevili skryté vlohy pro tuto disciplínu až na jaře 2020 a patrně nebude náhodou, že alespoň v České republice se do tohoto typu „matematického modelování“ pustilo pouze minimum osob, jež se matematickému modelování věnovalo již před rokem 2020.

[2] Kromě jiného editor jednoho z nejvýše hodnocených statistických časopisu *Journal of the Royal Statistical Society, Series B* a takřka duchovní otec (resp. jeden z duchovních otců) aditivních modelů.

[3] To jest jinak než pomocí „modelu“, který má x neověřitelných předpokladů, ignoruje několik zákonitostí statistické inference a je odhadnut na základě observačních dat vystavených v neznámé míře výběrovému vychýlení.

Reference

Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulík, V., Norman, A. J., Teperowski Monrad, J., Besiroglu, T., Ge, H., Hartwick, M. A., Whye Teh, Y., Chindelevitch, L., Galand, Y., Kulveit, J. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science*, **371**(6531), DOI 10.1126/science.abd9338.

Kulich, M. (2021). O datech. *Pojednání, Odborná skupina České statistické společnosti,* únor 2021, http://www.statspol.cz/wp-content/uploads/2021/02/o_datech_publ.pdf.

Straka, R. a Komárek, A. (2021). Odhad průběhu epidemie SARS-CoV-2 v ČR na základě počtu úmrtí: Statistické modelování pomocí reálných dat. *Technická zpráva, Sdružení mikrobiologů, imunologů a statistiků,* červen 2021, https://smis-lab.cz/wp-content/uploads/2021/06/2021-06-07-Straka.pdf.

Šmíd, M., Berec, L., Kuběna, A. A., Levínský, R., Trnka, J., Tuček, V. and Zajíček, M. (2021). SEIR Filter: A Stochastic Model of Epidemics. *medRχiv*, 1–39, DOI 10.1101/2021.02.16.21251834.

Wood, S. N. (2021). Inferring UK COVID-19 fatal infection trajectories from daily mortality data: Were infections already in decline before the UK lockdowns? *Biometrics*, 1–14, DOI 10.1111/biom.13462.

idnes.cz (9.9.2021). V nemocnicích končí s covidem hlavně neočkovaní, obézní a nemocní senioři. https://www.idnes.cz/zpravy/domaci/nemocnice-koronavirus-covid-hospitalizace-testovani.A210909_080136_domaci_remy