

Náhled do tajemství dat

12.5.2023 - | Grantová agentura ČR

Když Stanislav Nagy vysvětluje podstatu své práce laikům, začíná jednoduchým příkladem. Chceme zjistit, jaká je obvyklá výška dětí v určitém věku.

„Základní postup je, že změřím výšku dejme tomu dvaceti stejně starých dětí, z naměřených hodnot vypočítám průměr a mám představu, jak jsou tyto děti obvykle vysoké,“ říká doktor Nagy. „Jenomže co když jsem udělal při jednom měření chybu a zapsal jsem místo 140 centimetrů 1400 centimetrů? Pak mi z počítače vyskočí průměr nesmyslně vysoký a pozorování znehodnotí.“

Statistikové proto používají veličinu zvanou medián. Získané hodnoty seřadí od nejnižší po nejvyšší a mediánem je pak číslo, které leží uprostřed. Neboli před mediánem i za ním je stejné množství pozorování. Medián v tomto případě bude zřejmě kolem 140 cm a „ulétnutý“ záznam výsledek neovlivní.

Mnoharozměrný datový prostor

To je jednoduché. Ale co když jedno pozorování zahrnuje dva údaje? *„Dejme tomu u lékaře máme děti stejného věku a u každého změříme výšku a hmotnost,“ říká doktor Nagy. „Získáme ke každému pozorování dvě hodnoty, můžeme je zanést do dvourozměrného grafu a tady už je větší problém najít medián, tedy bod, kolem něhož jsou pozorování rozptýlená, abychom měli představu, jaká kombinace výšky a hmotnosti je v této věkové skupině obvyklá.“*

Když se přidají další měření (třeba ještě krevní tlak, srdeční tep, hladina cukru v krvi...), získají pozorovatelé hodnoty, které se dají zanést jako body zobrazitelné v mnoharozměrných prostorech. Ty si sice dokáže představit málokdo, ale matematika s nimi běžně pracuje.

Až na „drobnost“. Neumí v nich dobře najít právě medián. A to je docela problém. V našem hypotetickém případě jej potřebujeme proto, abychom úplně jednoduše určili, že dítě, jehož naměřené hodnoty se nacházejí poblíž mediánu, prochází zřejmě obvyklým vývojem. Zato dítě, jehož hodnoty se zobrazí daleko od mediánu, může mít poruchu vývoje nebo nějakou chorobu, a je proto nutné věnovat mu lékařskou pozornost.

Nečekané souvislosti

Velké vědecké experimenty přinášejí obrovské množství údajů, při jejichž analýze by se medián hodil. Proto matematictí statistikové navrhli řadu postupů, jak medián v mnoharozměrné sadě dat určit.

Pro potřeby tohoto textu stačí napsat, že statistika při stanovení blízkosti k mediánu používá od roku 1975 termín „hloubka“. Je to funkce, která určuje, jestli má naměřený bod k mediánu blízko, nebo je vzdálený. Bohužel se však postupy určení hloubky od sebe liší, a tak statistici docházejí k různým výsledkům.

Existuje však i jiný obor matematiky, zvaný konvexní geometrie. *„Někdy před šesti lety jsem si všiml, že v tomto oboru existuje obdobný problém, jako je ve statistice hledání mediánu a příslušné hloubky. Ale geometrie jej řeší už od roku 1822,“* popisuje doktor Nagy. V tomto případě se studovaný problém nazývá „popis plovoucího tělesa“ a týká se vlastností objektu volně plovoucího v kapalině.

To, čemu statistici říkají „hloubka“, se v konvexní geometrii nazývá „plovoucí těleso“, je jinak definováno, má úplně jiné použití, ale doktor Nagy si všiml, že obě pojetí se sobě velice podobají. Spolu s kolegy se tedy v rámci projektu financovaného Grantovou agenturou ČR pustil do propojování poznatků z obou matematických oborů. To, že si všiml jejich podobnosti, se ukázalo jako výborný tah.

„Mohli jsme použít poznatky osvědčené v konvexní geometrii k tomu, abychom vyřešili některé těžké problémy ze statistiky,“ raduje se Stanislav Nagy. *„Vyvrátili jsme při tom spoustu chybných závěrů z minulosti.“* Nové výsledky zaznamenali v šestnácti odborných publikacích.

Happyend jen částečný

Čekal je však ještě největší úkol. Zjistit, jestli se postup používaný v konvexní geometrii při popisu plovoucího tělesa dá využít při řešení úkolů v matematické statistice úplně pokaždé.

„Pokud by se ukázalo, že postupy jsou zcela rovnocenné, bylo by to vynikající. Mohli bychom neomezeně propojit obě disciplíny a přineslo by to užitek matematikům i spoustě odborníků z jiných věd,“ vysvětluje doktor Nagy, proč byl jeho další výzkum tak důležitý.

Ověřování věnoval několik let. Výsledek přinesl určité zklamání. *„Obě funkce jsou zcela stejné jen za určitých podmínek. V těch jsme oba matematické obory určitě posunuli dál a z toho mám opravdu radost,“* konstatuje a hned dodává: *„Ale pořád je tam co objevovat. Matematika má hodně záhad, tahle je mezi nimi, ale my ji jednou vysvětlíme.“*

Název projektu nominovaného na Cenu předsedy Grantové agentury ČR: **Geometrické aspekty matematické statistiky**

<https://gacr.cz/nahled-do-tajemstvi-dat>