LÉKAŘSKÁ STATISTIKA

# Jaká data statistika potěší?

Statistické studie jsou v oblasti medicíny nutností, a proto jsou také stále častěji vyžadovány. Jen velmi malá část lékařů zpracovává svoje data bez pomoci odborníka – statistika. Vyhledáním statistika ovšem lékaři práce nekončí. Zdravotník a statistik mluví každý naprosto odlišným jazykem, proto je pro správnost závěrů analýzy nezbytné, aby se domluvili a vzájemně pochopili. Tento příspěvek si klade za cíl ulehčení spolupráce obou stran.

Mgr. Lenka Blažková
analytička a konzultantka, StatSoft CR, s. r. o.

## Jak začít?

Jestliže máte šanci před samotným sběrem dat vyhledat statistika, udělejte to. Poradí vám, jak velký je **minimální nutný rozsah souboru** pro ověření vašich hypotéz, jaký **formát pro data** zvolit, řekne vám, které **metody** jsou pro řešení konkrétního problému vhodné, zvládne poradit i několik **triků**, jak si ušetříte práci, a zodpoví vaše dotazy.

## Zkuste data učesat, než je předáte

Nejčastěji bývají data uložena v tabulce nebo několika tabulkách aplikace Excel. Jak má taková tabulka v ideálním případě vypadat? Předpokládejme, že data tvoří záznamy o léčených pacientech.

• **První řádek tabulky obsahuje záhlaví:** Jde o stručné označení jednotlivých sledovaných veličin a parametrů, ideální je užití krátkých výstižných názvů bez diakritiky (čeština může některým statistickým programům dělat potíže). Některé statistické programy také zakazují mezery, tečky nebo další určité znaky v názvu, mohou omezovat jeho délku a například zakazovat číslici jako první znak.

| Nr | gender | age (years) | body stature (m) | BMI (kg/m2) | body weight pre (kg) | post 1.stage( kg) |
|---|---|---|---|---|---|---|
| 1 | M | 33 | 1,73 | 23,9 | 71,5 | 70,1 |
| 2 | M | 41 | 1,8 | 22,5 | 73,0 | 71,2 |
| 3 | M | 52 | 1,86 | 23,1 | 80,0 | 76,2 |
| 4 | M | 40 | 1,78 | 25,3 | 80,1 | 77,1 |
| 5 | M | 40 | 1,83 | 27,2 | 91,1 | 89,9 |
| 6 | M | 35 | 1,75 | 23 | 70,3 | 69,9 |
| 7 | M | 42 | 1,77 | 23,6 | 73,9 | 71,7 |

• **Ve zbylých řádcích** jsou záznamy hodnot sledovaných veličin a parametrů pro jednotlivé pacienty. Co řádek, to jeden pacient.

• **Identifikační číslo (ID):** Každý pacient by měl mít své (unikátní) identifikační číslo. Raději se vyhněte rodným číslům (podle zákona o ochraně osobních údajů byste rodná čísla nikdy neměli předat statistikovi, totéž platí pro jména pacientů a další osobní údaje jako adresa nebo telefon; přesvědčte se, že v souboru ke statistickému zpracování tyto citlivé informace nejsou!) a použijte jednoduše pořadová čísla. V minulosti existují případy, kdy se rodná čísla dvou lidí shodovala. Problémů z toho plynoucích je lepší se vyvarovat. **Každá tabulka by měla obsahovat sloupec s ID**. Pokud jsou data uložena ve více tabulkách, dají se za předpokladu, že každý pacient má svoje unikátní číslo, na základě ID spárovat údaje z různých tabulek, které patří téže osobě. Použití ID zaručuje dostatečnou anonymitu pacientů, pokud je ovšem pro lékaře nutné později zjistit jméno konkrétního pacienta reprezentovaného pouze jeho ID číslem, původní záznamy lékařů (obsahující jak údaje o pacientech, tak přiřazené ID) to umožní.

• Pokud některý údaj **chybí**, nechejte v příslušné buňce tabulky **prázdné místo**. Vepisování nul či jiných znaků vede k znepřehlednění souboru. Navíc nelze rozlišit, zda byla 0 naměřená, či zda je v buňce proto, že údaj nebyl zjištěn.

• Tabulka ke zpracování by neměla obsahovat **žádné souhrnné řádky** (**součty, průměry** atp.), neboť statistické programy načtou jako proměnnou vždy všechny hodnoty sloupce. Jestliže souhrny potřebujete, mohou být na několika posledních řádcích tabulky, ale neměly by se volně střídat s původními daty. Jakmile je jejich výskyt v souboru náhodný, musí statistik taková data odmazávat manuálně a stojí to hodně času i energie. Navíc je u všech ručních úprav větší pravděpodobnost výskytu chyb.

| | A | B | C |
|---|---|---|---|
| 1 | **Název** | **Celý název** | **Význam** |
| 2 | HIE | Hypoxicko ischemická encefalopatie | stav, kdy mozek není dostatečně okysličován, podle závažnosti rozdělujeme na skupiny I. , II. a III. (nejhorší stav) |
| 3 | S-urea | Krevní sérum - močovina (mmol/l) | normální hodnoty: 0 - 6 týdnů **1.7-5.0**, 6 týdnů - 1 rok **1.4-5.4** |
| 4 | SDate | Start date | datum, kdy začala léčba |
| 5 | Age | Gestační věk | stáří novorozence včetně prenatálního vývoje (jendotka: týdny) |

• Zachovávejte **konzistenci v záznamech:** Například **„M“** a **„m“** a **„muž“** ve sloupci udávající pohlaví může statistický program vyhodnotit pokaždé jako různou hodnotu. Použité výrazy by měly být vždy identické, zejména je zapotřebí ohlídat, že jste omylem nevložili mezeru před samotné slovo. Zkontrolujte i malá a velká písmena, některé programy je nerozlišují, ale jiné ano.

• **Pro oddělení desetinných míst používejte vždy stejný znak** (buď tečku, nebo čárku).

• **Opakovaná měření:** Pokud opakovaně zaznamenáváte nějakou charakteristiku o pacientovi v průběhu léčby, v názvu proměnné by měl být obsažen (vždy stejný) název parametru a příslušný časový kód (1D, 3M, 1Y) oddělený nejlépe podtržítkem (příklad: **krea_1D**).

• **Datum:** V praxi nezáleží, jaký formát pro datum použijete, ověřte si ale, že je stejný v celém souboru, popřípadě ve všech tabulkách.

S takto upravenými daty již statistika nevyděsíte. Analytici jsou sice po očištění a úpravě dat schopni vyhodnotit leccos, ale uvědomte si, že vhodným zpracováním do přehledných tabulek v první řadě děláte něco užitečného, co po vás už nikdo nebude muset opakovat, a také snižujete finanční náklady na komerční zpracování statistikem-externistou.

## Slovníček

Na paměti ovšem mějte, že statistik je jenom obyčejný člověk. Nejsou mu vlastní zkratky, jež lékaři užívají denně. Uvítá proto, když v souboru s daty bude i list, v němž budou k jednotlivým proměnným (nejlépe seřazeným dle abecedy anebo podle pořadí výskytu v tabulce) následující údaje:

1. ***Použitý název* (*zkratka*)** v záhlaví tabulky a celý název, stručný popis, co proměnná znamená.
2. ***U číselných údajů*** vždy teoreticky možné **minimum, maximum, rozsah** pro normální a abnormální hodnoty (statistik neví, co je vysoký cholesterol, je potřeba mu explicitně říci, že si má všímat hodnot HDL pod 1,4 a LDL nad 3,4). Pro správný přístup k měřenému parametru též potřebuje vědět, zda jsou hodnoty **spojité** (proměnná může nabývat jakékoli hodnoty z nějakého intervalu), nebo **diskrétní** (možné jsou jen některé hodnoty, např. 1, 2, 3, 4, 5, a žádná další – tj. nelze naměřit 2,5). Dobré je uvést i případné jednotky, v nichž je veličina měřena.
3. ***U kategorických veličin*** uvést vždy všechny teoreticky možné hodnoty. Pokud existuje nějaké přirozené pořadí kategorií (jež chcete zachovat v grafech atp.), uveďte jej.
4. ***Kódované údaje*** – pokud jsou jednotlivé kategorie kódovány (např. hodnoty 0, 1), uveďte, co který kód reprezentuje (tedy 0 = muž, 1 = žena apod.).

## Zadání

Zadání analýzy by mělo obsahovat stručný úvod do problematiky, které se věnujete. **Nejdůležitější je ovšem seznam otázek, na něž hledáte odpověď, nebo hypotéz, které potřebujete ověřit.** Nic nezapomeňte a buďte dostatečně detailní. Jestliže se vám nedaří jasně zformulovat, co potřebujete znát, obraťte se na statistika.

Někdy je spíše na škodu, když se snažíte vyjmenovat statistické metody, jež jsou z vašeho pohledu vhodné nebo obvyklé. Vše totiž závisí na datech a je možné, že ve vašem případě bude vhodný jiný postup než ten, který jste měli možnost zahlédnout v článku na podobné téma. Někdy je taková specifikace také zbytečně široká, například pojem multivariantní analýza zahrnuje všechny metody, kde se vyskytuje více proměnných, od regresních úloh až po exploratorní vícerozměrné techniky. Statistikovi tak lépe poslouží jiná studie na podobné téma, kde se pracuje s pro medicínu obvyklými statistickými metodami.

Je důležité, abyste uvedli, které proměnné jsou pro vás závislé (vysvětlované pomocí ostatních) a které nezávislé (slouží tedy jako vysvětlující parametry). Někdy je vztah symetrický, není podstatné rozdělení na závisle a nezávisle proměnné, tuto skutečnost případně uveďte.

Telefonicky nebo při osobním setkání si vzájemně vše vyjasníte. Ověřujte, že si vzájemně rozumíte, například parafrázováním požadavků druhé strany.

## Účel

• Nezapomeňte uvést, pro jaké účely výstup potřebujete – je pro vás vhodnější **dokument Word**, nebo **prezentace v PowerPointu**? Pokud již máte k dispozici šablonu pro požadovanou prezentaci, zašlete ji statistikovi – budete tak mít méně práce s upravováním formátu.

• Budete grafy **tisknout pouze černobíle**, a je pro vás tedy dobré, aby již byly **generovány ve stupních šedi**?

• **Potřebujete editovat tabulky a grafy**, nebo mohou být ve formě obrázků?

• Některý *software* poskytuje anglický i český výstup. Informujte se, v jakém programu budou vaše data zpracovávána a zda je dostupná **jazyková verze**, která vám nejlépe vyhovuje.

Všechny tyto podrobnosti jsou důležité; pokud si je uvědomíte až v době, kdy už je statistik v polovině práce, může to znamenat nutnost zpracovávat vše od začátku. Proto i zde platí heslo: dvakrát měř a jednou řež.