Komentář k prezentaci „Statistické metody - nástroj poznání a rozhodování anebo zdroj omylů a lží“

Doc. RNDr. Zdeněk Karpíšek, CSc.
Odbor statistiky a optimalizace
Fakulta strojního inženýrství
Vysoké učení technické v Brně

Motto: „ Jsou tři druhy lží: lži, odsouzeníhodné lži a statistiky.“

Od vědy o státu a hazardních her k matematické statistice

V současném světě má statistika velmi významné a nezastupitelné místo. Moderní řízení výroby, financí a obchodu v zájmu maximalizace jejich efektivnosti je nerealizovatelné bez kvalitních dat a je založeno na neustálém vyhodnocování informací o objektu i jeho okolí za použití exaktních metod. Stejně tak základní i aplikovaný výzkum v takřka všech disciplinách potřebuje rigorózně a co nejpřesněji zpracovat data o studovaných jevech a procesech. Mimořádně významná role v těchto činnostech přísluší statistickým metodám, které prostřednictvím počítačů poskytují soustavu číselných a grafických informací o zkoumaném či řízeném celku, o jeho subsystémech a prvcích, včetně nás samých.

Slovo statistika bývá převětšinou chápáno ve třech pojetích: jako číselné údaje o hromadných jevech, dále jako praktickou činnost spočívající ve sběru, zpracování a vyhodnocování statistických údajů, a jako teoretickou disciplínu, která se zabývá metodami pro popis a odhalování zákonitostí při působení podstatných, relativně stálých činitelů na pozorované jevy a procesy.

Statistické zkoumání lze rozdělit do tří etap. V etapě statistického zjišťování (šetření) získáváme statistické údaje, což jsou číselné nebo slovní hodnoty (obměny) sledovaných statistických znaků. Ve druhé etapě získané soubory údajů po kontrole a verifikaci třídíme a shrnujeme. V závěrečné etapě provádíme vyhodnocování a rozbor získaných statistických údajů pomocí vhodných statistických metod.

Na statistické zkoumání vybraných jevů a procesů pak navazuje aplikace výsledků v daném oboru (např. změna řízení finančních toků, zásah do výroby, vyvození závěrů z vědeckého experimentu aj.). Efektivnost aplikace je však nutno znovu podrobit novému statistickému zkoumání ve smyslu tzv. zpětné vazby.

Statistické metody zpracování a analýzy získaných dat vycházejí z popisné statistiky (získávání údajů, číselné a grafické zpracování datových souborů), teorie pravděpodobnosti (rozdělení náhodných veličin a procesů a jejich charakteristiky), matematické statistiky (odhady parametrů, testy hypotéz, regresní analýza aj.).

Historie výše uvedených disciplin je zajímavá a dosti rozmanitá. Počátky popisné statistiky souvisejí s existencí státních útvarů (jak se tvrdí v úvodech učebnic statistiky), tedy asi do období 7 tisíc let nazpět. Oprávněně se lze domnívat, že sahají ještě hlouběji, neboť první číselné záznamy (vyjadřující spíše množství něčeho než pouhý artefakt) vytvořil pravěký člověk již před 30 tisíci lety. Základy teorie pravděpodobnosti byly položeny zhruba v XVII. století na společenskou objednávku. Konkrétně pro řešení nepříliš ušlechtilých, ale zajímavých otázek spojených s požadavkem hráčů na dosažení úspěchu při provozování víceméně hazardních her. Matematická statistika se vyvíjela pozvolně, často s mírným opožděním oproti teorii pravděpodobnosti, a to v souvislosti s nástupem a rozvojem exaktních metod ve vědeckém bádání při studiu a měření reálných jevů i dějů a následnými aplikacemi těchto metod. Její rychlý a expandující vývoj od konce 19. století byl vyvolán především technickým, průmyslovým a ekonomickým rozvojem, tedy potřebami praxe.

V zjednodušeném pohledu představují metody matematické statistiky spojení metod popisné statistiky s teorií pravděpodobnosti v tom smyslu, že popisované jevy studujeme s ohledem na jejich nahodilé chování. Přitom náhoda nemusí být obsažena jen v jejich postatě (to je otázka nazírání a tedy spíše filozofická), ale také v samotném pozorování celku pouze prostřednictvím jeho části, která jej sice dostatečně reprezentuje, avšak je víceméně náhodně z celku vybrána.

Zmíněné disciplíny nejsou v žádném případě uzavřenými matematickými disciplínami. V současné době rostou možnosti jejich užití velmi výrazně s nasazením počítačů a to spolu s rozvojem lidského poznání a konání vyvolává potřebu vývoje nových metod v těchto disciplinách.

Populace, výběr, náhoda a neurčitost patří k sobě

Při statistickém zkoumání se zabýváme jevy a procesy, které mají hromadný charakter a vyskytují se u rozsáhlého souboru individuálních objektů (výrobky, osoby apod.), nazývaného základní soubor nebo populace. Zkoumané objekty jsou tzv. statistické jednotky a sledujeme u nich vytypované vlastnosti - statistické znaky (veličiny, parametry atd.), které nabývají pozorovatelných hodnot (úrovní).

Podle druhu hodnot dělíme statistické znaky na kvantitativní, které nabývají číselných hodnot (hmotnost, délka, pevnost, cena, životnost,...) a kvalitativní, které nemají číselný charakter a lze je vyjádřit slovně (barva, jakostní třída, podmínky provozu, tvar,...). Sledujeme-li jen jeden znak, hovoříme o jednorozměrném znaku, naopak o vícerozměrném znaku.

Kvantitativní znaky dělíme na diskrétní, jestliže nabývají pouze oddělených číselných hodnot (počet zmetků, počet vad, kusová produkce apod.) a spojité, které nabývají všech hodnot z nějakého intervalu reálných čísel (rozměr výrobku, doba do poruchy, cenový index apod.). Kvalitativní znaky dělíme na ordinální, jejichž slovní hodnoty má smysl uspořádat (jakostní třídy, klasifikace apod.) a nominální, jejichž slovní hodnoty postrádají význam pořadí (barva, tvar, dodavatelé apod.).

Podstatou statistických metod je, že informace o základním souboru nezjišťujeme u všech jeho jednotek, ale jen u některých, které získáme tzv. výběrem. Vedou nás k tomu různá omezení, např. dosažitelnost všech jednotek, velký rozsah základního souboru, způsob získávání informací (zkoušky životnosti, ověření opotřebení atd.), náklady na statistické sledování a další. Počet vybraných jednotek je rozsah výběru. Dle rozsahu dělíme výběry na malé (obvykle do 30 až 50) a velké (řádově stovky, tisíce i více). Toto dělení je relativní a závisí na okolnostech statistického sledování. Výběr by měl být reprezentativní (poskytovat informace bez omezení) a homogenní (bez vlivu dalších různých faktorů). To však často nelze v plné míře verifikovatelně zajistit, a proto obvykle vybíráme statistické jednotky do výběru náhodně. Výběr provází riziko neurčitosti v tom smyslu, že výběr může poskytnout více či méně zkreslené informace o základním souboru. Podle způsobu provedení rozlišujeme různé druhy výběrů: bez opakování, s opakováním, záměrný, oblastní, mechanický aj.

Hodnoty znaku, pozorované či zjištěné na vybraných statistických jednotkách tvoří statistický soubor s týmž rozsahem jako má výběr.

Průměry a jejich ctnosti i nectnosti

Získaný statistický soubor před vlastním zpracováním dle potřeby (např. pro získání grafů nebo pozdější použití metod matematické statistiky) třídíme a to tak, že jej rozdělíme do skupin nazývaných třídy. Každou třídu pak reprezentuje její typická hodnota nazývaná střed a četnost skupiny prvků, které do ní patří. Zpracování souboru spočívá v jeho grafickém znázornění a výpočtu číselných charakteristik.

Grafy poskytují vizuální informace o poloze, variabilitě (rozptylu), symetrii, modalitě (místech soustředění) a jiných vlastnostech souboru, užitečných pro jeho posouzení. Číselné (empirické) charakteristiky statistického souboru jsou čísla, která poskytují důležité a koncentrované informace o výše uvedených atributech souboru. Nejčastěji užívanou charakteristikou polohy jednorozměrného statistického souboru s kvantitativním znakem (např. příjem osob, hmotnost balíčku kávy apod.) je nesporně aritmetický průměr. Předpokládejme např., že průměrný plat 10 pracovníků činí 10 000,- Kč měsíčně. Jestliže se zvýší plat toho, kdo z nich bere nejvíce, konkrétně z 50 000,- Kč na 100 000,- Kč měsíčně a ostatním 9 pracovníkům se platy nezmění (to je ještě příznivá varianta změny), zvýší průměrný plat na 15 000,- Kč. Případný komentář, že se každému zvýšil plat v průměru o 5 000,- Kč, se již snad nedá nazvat ani jako zavádějící. Použijeme-li "robustnější" charakteristiku polohy zvanou medián, dostaneme jiný obraz téže skutečnosti. Mediánem rozumíme prostřední hodnotu souboru uspořádaného od nejmenší do největší hodnoty souboru. Je zřejmé, že medián souboru se výše uvedenou změnou nezmění. To je nectnost průměru, která je navíc znásobena skutečností, že u tzv. kladně asymetrických souborů (převládají co do počtu malé hodnoty nad velkými) je průměr vždy větší než medián.

Průměr má však také významné ctnosti. Jestliže naopak např. ve výrobě sledujeme změny průměrů souborů sledovaných rozměrů výrobků, reaguje průměr dosti citlivě na nevhodné změny ve výrobě a můžeme pomocí něho potřebnými zásahy pozitivně ovlivnit jakost výroby. To je jedním ze základních principů pro statistické sledování a řízení jakosti výroby. Další a významnou ctností průměru je jeho schopnost se blížit se s rostoucím rozsahem souboru k průměru celé populace a také jeho chování při opakovaných výběrech z populace, jak bude popsáno v jednom z dalších odstavců.

Měříme proměnlivost veličin a vztahy mezi nimi

Pro přesnější obraz o celé populaci je však nutno použít další charakteristiky, zejména rozptyl, resp. směrodatnou odchylku, které vyjadřují, jak jsou pozorované hodnoty rozptýleny. Je jasné, že i když např. dva soubory hmotností stejných balíčků kávy získané od dvou balících automatů mají takřka stejné průměrné hmotnosti (a blízké deklarovanému údaji na obalu), je pro zákazníka horší ten automat, který vykazuje větší rozptyl hmotnosti. Zákazník v obchodě může těžko ovlivnit volbu balícího automatu a navíc nadneseně řečeno, platí-li tzv. Murphyho zákon o změně jízdního pruhu v podobné verzi i zde, dostávají se na něj balíčky s menší hmotností i při změně druhu kávy. Na druhé straně nulový vypočtený rozptyl souboru může znamenat chybu ve výpočtu, nevhodný výběr, špatně zvolenou metodu, přístroj nebo způsob měření, vážení apod., ne-li upozornění na příchod soudného dne.

Podobná situace je také při popisu dvourozměrného souboru, kdy pozorujeme na statistických jednotkách dva znaky současně. Můžeme se zabývat každým znakem zvlášť, ale navíc očekáváme jejich byť nepřesnou závislost. Známe z osobní zkušenosti, že např. dle tzv. Parkinsonova zákona obvykle s rostoucími příjmy rostou vydání, s větší výškou postavy člověka bývá často spojena její větší hmotnost apod. Tato závislost však není "přesná", neboť opačných případů zejména u postav je tolik, že nejde o výjimky. Závislost dvou kvantitativních znaků nejčastěji vyjadřujeme tzv. koeficientem korelace. Jde o velmi účinnou charakteristiku závislosti nabývající hodnot od -1 do 1, ale ne vždy správně interpretovanou. Čím je jeho hodnota bližší 1 anebo -1, je závislost znaků těsnější a navíc blízká lineární, jinak řečeno vyjadřuje ji graficky dobře přímka (dokonce pro tyto mezní hodnoty přesně přímka). Jeho nulová hodnota však nemusí vždy znamenat (s výjimkou tzv. normálního rozdělení pravděpodobnosti), že mezi sledovanými znaky není závislost - může být dokonce zcela přesná. Pro "jemnější" vyjádření závislosti se proto dle možností často používá tzv. regresní analýza, příp. další vícerozměrné statistické metody, které umožňují serioznější závěry a předpověď hodnot jednoho znaku pomocí hodnot druhého znaku nebo více znaků.

Rozdělení pravděpodobnosti a zákon velkých čísel jsou obrazem nás i okolního světa

Teoretickou míru možnosti nastoupení náhodného jevu vyjadřuje číselně jeho pravděpodobnost. Od počátků teorie pravděpodobnosti do dneška počítáme v jednoduchých případech tuto míru pomocí poměru "počtu příznivých případů" k "počtu všech možných případů". To však předpokládá, že možné případy jsou stejně pravděpodobné a že je jich konečně mnoho. Proto tato definice selhává např. u falešné hrací kostky anebo doby do poruchy nějakého zařízení. Složitá cesta vývoje tohoto základního pojmu byla po zhruba dvou stoletích zakončena úspěšně ve 30. letech 20. století axiomatickou definicí založenou na teorii množin. Přesto byly již v předcházejícím období nalezeny zákony rozdělení pravděpodobnosti pro modelování reálných jevů, např. rozdělení binomické, hypergeometrické, normální (Gaussovo) aj. Pomocí Bernoulliova zákona velkých čísel bylo odhaleno asymptotické chování relativní četnosti nastoupení náhodného jevu v tom smyslu, že relativní četnost při rostoucím počtu nezávislých pokusů konverguje takřka jistě k pravděpodobnosti tohoto jevu. Odtud pak vychází základní metody teorie odhadu v matematické statistice. Mimořádně významné postavení při modelování reálného světa má právě normální rozdělení, neboť dle limitních vět konverguje za dosti obecných podmínek rozdělení pravděpodobnosti normovaného průměru náhodných veličin s v jistém smyslu libovolnými rozděleními pravděpodobnosti k rozdělení normálnímu. To však neznamená, že jsou jiná rozdělení jen matematickou konstrukcí a že vystačíme pro popis reality jenom s tímto rozdělením. Poznamenejme ještě, že rozdělení pravděpodobnosti je popsáno jeho funkčními charakteristikami (distribuční funkcí, hustotou apod.) a číselnými charakteristikami - parametry (střední hodnotou, rozptylem aj.).

Výběrové charakteristiky a kletba statistikova

Metody matematické statistiky jsou v zásadě založeny na dvou principech:

1. Hodnoty pozorovaného statistického znaku získané výběrem ze základního souboru jsou náhodné. Přepokládáme tedy, že existuje nějaké jejich (třeba i neznámé) rozdělení pravděpodobnosti.
2. Získaný statistický soubor je hodnotou tzv. náhodného výběru, jímž je vícerozměrný náhodný vektor s nezávislými složkami odpovídajícími jednotlivým

pozorováním a tyto složky mají stejné rozdělení pravděpodobnosti jako pozorovaný znak.

Místo znak se obvykle říká náhodná veličina, neboť většinou pozorujeme kvantitativní znaky. Získaný statistický soubor popisují jeho empirické (číselné) charakteristiky, avšak při opakování pozorování náhodné veličiny získáme nové statistické soubory s obecně jinými hodnotami empirických charakteristik. Z uvedených principů vyplývá, že tyto charakteristiky jsou vlastně hodnotami náhodných veličin, které jsou funkcemi pozorované náhodné veličiny. Tyto náhodné veličiny se nazývají výběrové charakteristiky nebo také statistiky. Tak získáme např. výběrový průměr, výběrový rozptyl aj.

Pomocí výběrových charakteristik řešíme dvě základní úlohy matematické statistiky: (1) odhady parametrů a rozdělení, (2) testování statistických hypotéz o parametrech a rozděleních. Přitom využíváme skutečnost, že výběrové charakteristiky nabývají náhodných hodnot (jimiž jsou empirické charakteristiky) blízkých teoretickým charakteristikám a navíc s rozumnými vlastnostmi. Velmi důležitou vlastností je např., že při splnění nepříliš silných podmínek střední hodnota výběrového průměru je rovna střední hodnotě pozorované veličiny ("průměru" populace) a rozptyl výběrového průměru se s rostoucím rozsahem výběru blíží (konverguje) k 0. To znamená, že při dostatečně velkém rozsahu výběru je takřka jistě průměr souboru blízký neznámé střední hodnotě. Není však vše zadarmo, neboť tento rozptyl konverguje k 0 pomaleji, než roste rozsah výběru, konkrétně s druhou odmocninou z rozsahu. Hovoříme proto o tzv. kletbě statistikově, neboť chceme-li zvýšit přesnost svých závěrů v odhadu nebo testu hypotézy o střední hodnotě dvakrát, musíme rozsah souboru zvýšit čtyřikrát, což stojí také zhruba čtyřikrát více času, peněz atd. Podobná situace je i u dalších výběrových charakteristik. Tuto skutečnost bychom proto měli respektovat při závěrech z výsledků různých průzkumů apod.

Je lepší odhad statistika anebo experta?

Odhady neznámých parametrů realizujeme pomocí vhodných výběrových charakteristik a dělíme je na dva druhy: bodové a intervalové. V prvním případě odhadujeme hodnotu parametru jedním číslem a ve druhém případě ji odhadujeme pomocí tzv. konfidenčního intervalu s danou spolehlivostí. Spolehlivost intervalového odhadu obvykle volíme 95% anebo 99%. Spolehlivost např. 95% znamená, že při mnohokrát opakovaných výběrech s týmž rozsahem, obsahuje zhruba 95% těchto intervalových odhadů neznámý parametr a zbývající intervalové odhady jej neobsahují. Riziko chyby 5% je nepříjemné a navíc jeho snížení vede při zachování rozsahu výběru ke zvětšení velikosti intervalu a tím ke zvýšení nepřesnosti odhadu. Pokud se s tím nesmíříme, je nutno zvýšit rozsah výběru, avšak výsledek je poplatný "kletbě statistikově" nebo použít jinak konstruovaný intervalový odhad tohoto parametru (pokud je znám). Intervalový odhad experta proto může být lepší než odhad statistika, avšak vyžaduje těžko ověřitelnou oprávněnost důvěry v to, že se expert nemýlí. Pokud expert i statistik předloží pouze bodové odhady (viz např. odhady získané z průzkumů, které prezentují sdělovací prostředky), je zřejmě spolehlivost těchto odhadů nulová anebo zanedbatelná.

Nezamítnutí statistické hypotézy ještě není potvrzení její správnosti

Při testování statistické hypotézy, kterou nazýváme nulová hypotéza (např. hypotéza o hodnotě parametru rozdělení pravděpodobnosti, které popisuje zkoumaný základní soubor), stavíme proti ní tzv. alternativní hypotézu spočívající v tom, že tento parametr má jinou hodnotu než uvažovanou. Při testování nejprve vypočteme ze získaného statistického souboru hodnotu tzv. testového kritéria (vytvořeného z vhodné výběrové charakteristiky). K tomuto testovému kritériu je pro alternativní hypotézu zkonstruován tzv. kritický obor, do něhož padne hodnota kritéria při platné nulové hypotéze s předem danou pravděpodobností, tzv. hladinou významnosti. Tuto hladinu významnosti volíme obvykle 5% anebo 1%. Jestliže hodnota testového kritéria padne do kritického oboru, nulovou hypotézu zamítáme a alternativní hypotézu nezamítáme a naopak. Přitom nastane vždy jeden ze čtyř případů:

1. Nulová hypotéza platí, avšak my ji zamítáme, takže se dopouštíme tzv. chyby prvního druhu.
2. Nulová hypotéza neplatí, avšak my ji nezamítáme, takže se dopouštíme tzv. chyby druhého druhu.
3. Nulová hypotéza platí a my ji nezamítáme.
4. Nulová hypotéza neplatí a my ji zamítáme.

Případy 3 a 4 jsou vítané, ale případům 1 a 2 se nelze vyhnout, pokud výběrem nevyčerpáme celý základní soubor. Pravděpodobnost chyby prvního druhu (hladinu významnosti) obvykle neměníme a snižujeme pravděpodobnost chyby druhého druhu zvýšením velikosti výběru.

Nezamítnutí hypotézy však ještě neznamená její přijetí a je-li to možné, zvýšíme rozsah výběru a znovu hypotézu testujeme. Je zřejmé, že nezamítnutí nebo přijetí hypotézy není ještě její potvrzení její správnosti. Z hladiny významnosti např. 5% však plyne, že při opakovaných výběrech se chyby prvního druhu dopouštíme jenom ve zhruba 5% testů.

Používání statistických metod při řízení jakosti technologických procesů není samoúčelné ani módou

Ve 30. letech 20. století se statistika stává nástrojem pro hodnocení a zejména řízení jakosti hromadné výroby. Jde především o vytvoření tzv. Shewhartových regulačních diagramů měřením a porovnáváním a jejich nasazení ve strojírenském průmyslu. V zásadě jde o aplikaci a rozpracování odhadů parametrů a testů hypotéz o normálním, binomickém a Poissonově rozdělení pro relativně malé skupiny sledovaných výrobků odebíraných ke kontrole během výroby ve stanovených časových intervalech. Tyto odhady, resp. testy byly převedeny do názorné a manuálně snadno zpracovatelné grafické podoby pro získání rychlých a solidních informací o případných negativních vlivech na kvalitu výroby. Z toho pak při vybočení sledované charakteristiky mimo statisticky určené meze vyplývá možnost přímého zásahu do výrobního procesu nebo vyhledání jeho slabých míst při další přípravě výroby a následný finanční přínos pro výrobce a zvýšení důvěry odběratele.

V dalším období byly statistické metody pro řízení jakosti výroby dále rozpracovány s ohledem na možnosti nasazení počítačů a jsou obsahem norem pro řízení jakosti na státní a nadstátní úrovni (např. Evropská unie) i obsahem příruček jakosti u jednotlivých firem, zejména při jejich certifikaci. Staly se nezbytnou součástí komplexního řízení jakosti v souvislosti s rozvojem obecných metod a zásad. Nejde jen o regulační diagramy, ale také o vyhodnocování způsobilosti výrobních procesů vzhledem k technologickým a konstrukčním požadavkům, různé druhy statistické přejímky, optimalizaci nákladů aj.

Spolehlivost výrobku se dá měřit a úspěšně využívat

Jedním ze základních atributů výrobku je mimo jeho technických a finančních parametrů také jeho schopnost plnit požadované činnosti - tzv. spolehlivost. Existuje rozsáhlá skupina charakteristik spolehlivosti, z níž část respektuje stochastické chování doby bezporuchového stavu sledovaného objektu. Jde jednak o funkční charakteristiky (funkce spolehlivosti, intenzita poruch, hustota obnov aj.), jednak o číselné charakteristiky (střední doba do poruchy, koeficient pohotovosti apod.). Tyto charakteristiky přitom odhadujeme ze statistických souborů pomocí metod matematické statistiky často speciálně orientovaných s ohledem na realizaci zkoušek. Jde zejména o tzv. cenzorované výběry, kdy sledujeme skupiny výrobků např. během omezené doby do poruchy a všechny výrobky ještě nemusí být v poruchovém stavu.

Do oblasti spolehlivosti také spadají pravděpodobnostní modely celých systémů (např. výrobní linky, energetické bloky). Je zřejmé, že výrobky a celky s vyšší spolehlivostí mohou být dražší než výrobky a celky méně spolehlivé, avšak prokazují vyšší užitnou hodnotu pro uživatele či zákazníka a mnohdy také vyšší bezpečnost.

Proč a jak někdy statistiky lžou, a co dál…

Předložené téma o statistických metodách je natolik obsáhlé, že je není možno vtěsnat vyčerpávajícím způsobem do prostoru pro text a do jedné přednášky. Snahou autora je aspoň připomenout nebo naznačit obsah a význam základních pojmů, metod a postupů matematické statistiky s ohledem na možnosti jejich seriozních, ale bohužel také neseriozních aplikací. Statistiky samy o sobě nelžou, horší to ale někdy bývá s jejich realizátory a vykladači. Racionální a zodpovědné používání statistických metod naopak přináší pozitivní výsledky tím, že rozšiřuje naše poznání okolního světa i sebe sama a umožňuje nám účelně rozhodovat o našem dalším konání. A v tom lze vidět i jejich možný přínos v budoucnosti.