Statistické metody a zpracování dat

I. Úvod, základní pojmy

Petr Dobrovolný

Statistika **je vědní obor zabývající se zkoumáním jevů, které mají hromadný charakter.**

Statistika je v určitém smyslu jazykem pro **shromažďování**, **zpracování, rozbor, hodnocení a interpretaci** hromadných jevů

Statistika se těší pochybnému vyznamenání tím, že je nejvíce nepochopeným vědním oborem

(H. Levinson)

Jsou tři druhy lži: lež prostá, lež odsouzeníhodná a statistika. Statistika je zvlášť rafinovanou formou lži.

Významy pojmu STATISTIKA

I. Statistika jako praktická činnost

- Statistická evidence (např. sběr údajů, třídění, sumarizace apod.),
- Instituce, která tuto evidenci provádí (např. ČSÚ, ministerstva aj.)
- Souhrn údajů o nějaké skutečnosti (statistika nezaměstnanosti, ročenka meteorologických pozorování atd.)

Významy pojmu STATISTIKA

II. Statistika jako vědní disciplina

- **Popisná** statistika (výsledky nelze zobecnit)
- **Matematická** (induktivní) statistika - cílem je zobecnit výsledky (odhad a testování hypotéz) - použití **počtu pravděpodobnosti**
- Teorie výběrových zjišťování či měření
- Aplikované vědy („–metrie“ a „-grafie“): biometrie, dendrometrie, ekonometrie, chemometrie atd.
- Vědy se silným statistickým základem: klimatologie, hydrologie, sociologie, psychologie, demografie aj.

Co je typické pro statistiku

- Zkoumá **hromadné** jevy.
- Zabývá se proměnlivými - **variabilními** - vlastnostmi.
- Pracuje s čísly a vyjadřuje se pomocí čísel - zajímá se především o **kvantitativní stránku** reality.
- Používá výpočetní techniku k vytváření a správě statistických **databází**, k provádění hromadného **zpracování** a **analýzy** dat a ke **komunikaci.**

Co statistika „umí“

• **Zjišťování** (počet domácností ČR, počet pracovníků v odvětví XY)

• Popis **struktury** (věková struktura obyvatel ČR, roční chod hodnot meteorologických prvků)

• **Shrnování** dílčích ukazatelů v čase a prostoru (průměrná nezaměstnanost v regionu)

• **Srovnávání** agregovaných ukazatelů v čase nebo prostoru (trend vývoje počtu obyvatelstva, teploty vzduchu dvou lokalit)

• **Předvídání** jejich budoucí úrovně (tržby v maloobchodě v příštím roce)

• Měření **závislosti** (závislost mezd na HDP, závislost met. Prvku na nadmořské výšce).

… a co statistika „neumí“:

Statistika selhává, pokud:

- Nemá k dispozici adekvátní číselné údaje
- Chybí-li představa o velikosti chyb měření a vlivu různých doprovodných činitelů
- Nemá-li k dispozici dostatečně rozsáhlý soubor případů
- Není-li v datech přítomna proměnlivost (variabilita).

Vymezení základních pojmů I

Hromadné jevy: Jevy, které jsou výsledkem působení velkého množství příčin, jejich vlastnosti se neprojevují v jednotlivých jevech, ale jen v souboru a to prostřednictvím řady náhod.

Některé jevy, které v geografii studujeme pomocí statistických metod mají povahu jevů náhodných – tzv. stochastických (hydrologické jevy či meteorologické jevy).

Vymezení základních pojmů II

Statistická jednotka: je to určitý jev či prvek, který je předmětem statistického šetření a pro který se zjišťují údaje

Statistická jednotka musí být přesně vymezena na počátku vlastního šetření a to z hlediska **věcného, časového, prostorového.**

Statistický znak: je to určitá vlastnost statistické jednotky, kterou se snažíme postihnut. Tzv. **shodné (společné) znaky** vymezují příslušnost statistické jednotky k určitému statistickému souboru. Ostatní jsou znaky **proměnlivé (variabilní)**.

Vymezení základních pojmů III

Statistické znaky lze dělit na znaky **prostorové, časové a věcné.**

Věcné znaky se dělí na znaky **kvantitativní** a **kvalitativní**

Kvalitativní znaky mohou být **alternativní** a **možné**

Kvantitativní znaky dělíme nejčastěji na znaky **spojité** a **diskrétní**.

Statistické znaky můžeme získat přímo – (např. **měřením**) a nebo **nepřímo** (výpočtem). Tyto potom nazýváme znaky odvozenými.

Znaky **nominální, ordinální, poměrové, intervalové**

Vymezení základních pojmů IV

Statistický soubor: skupina statistických jednotek stejného druhu (věcně, prostorově a časově vymezených)

Je to množina všech prvků, které jsou předmětem daného statistického zkoumání. Každý z prvků je statistickou jednotkou.

Prvky tvořící statistický soubor mají určité společné vlastnosti - tzv. **identifikační znaky** - umožňující určit, zda prvek do daného statistického souboru patří nebo nepatří. Identifikační znaky tedy statistický soubor vymezují.

Z hlediska cílů statistického zkoumání sledujeme na prvcích statistického souboru jednu nebo více vlastností - **sledované znaky**. Je-li vlastnost měřitelná v nějakých jednotkách, jde o kvantitativní znak, jinak jde o kvalitativní znak.

Vymezení základních pojmů IV

Statistický soubor můžeme podle různých hledisek dále dělit:

Statistický soubor **jednorozměrný**, **vícerozměrný**
Statistický soubor **základní** a **výběrový**

Výběrový soubor je podmnožinou základního souboru. Je vytvořen ze statistických jednotek, vybraných podle určitého hlediska.

Reprezentativní výběr: Pokud zkoumaný výběr dobře odráží strukturu celého zkoumaného souboru, nazýváme jej reprezentativním výběrem.

Rozsah statistického souboru: počet statistických jednotek v souboru: N – rozsah základního souboru n – rozsah výběrového souboru

Popisná (deskriptivní) statistika

Popisná (deskriptivní) statistika se zabývá uspořádáním souborů, jejich popisem a účelnou sumarizací.

Jak mohou být tyto jevy jednoduše popsány (charakterizovány, sumarizovány)?

Existují dvě základní možnosti, které se vzájemně doplňují:

- **Numerické metody** – jedním nebo několika málo čísly lze vystihnout určité vlastnosti jevu. Jsou přesnější a objektivnější
- **Grafické metody** – sestrojení vhodného typu grafu. Jsou názornější a umožňují vystihnout vztahy.

Induktivní (matematická) statistika

Induktivní (matematická) statistika se vyvinula z popisné statistiky a jejím základem je **teorie pravděpodobnosti**.

Matematická statistika zkoumá soubory nepřímo prostřednictvím výběrů

Induktivní statistika se zabývá metodami jak poznatky **přenášet** a umožňuje z pozorovaných dat vytvářet **obecné závěry** s udáním *stupně jejich spolehlivosti.*

Výpočet stupně spolehlivosti závěrů je však objektivní, neboť je založen na poznatcích teorie pravděpodobnosti a nezávisí na subjektivním názoru hodnotitele.

Základní etapy statistického zpracování dat

- **Zjišťování** - shromáždění a zaznamenání údajů, jejich kontrola aj.,
- **Zpracování -** uspořádání, seskupení, shrnování, sumarizace,
- **Analýza** - výpočet charakteristik, měření závislostí, srovnávání, měření dynamiky
- **Prezentace** výsledků - tabulkové či grafické vyjádření a slovní zhodnocení výsledků předcházejících etap.

Základní dělení statistických údajů

- podle zdroje — **primární** a **sekundární**,
- podle reálnosti situace — **skutečné** a **simulované**,
- podle periodicity zjišťování — **průběžné**, **periodické** a **jednorázové**,
- podle časového hlediska — **okamžikové** a **intervalové**.
- podle použité škály měření – **nominální, ordinální, intervalové, poměrové**

Typy geografických dat

- **Nominální (kategorie využití země)**
- **Ordinální (řád vodního toku, stupnice síly větru)**
- **Intervalová (teplota vzduchu) nula = data**
- **Poměrová (množství srážek, délka vodního toku) nula = neexistence jevu**

Typy geografických dat

Nominální data – hodnota představuje konkrétní kategorii či třídu a vyjadřuje její označení (jméno), kategorie se nesmějí překrývat – jsou disjunktní. Každý objekt je zařaditelný alespoň do jedné kategorie, žádný nespadá do více jak jedné. Čísla, která označují kategorie jsou pouze symboly a nelze s nimi provádět aritmetické operace. Např. telefonní čísla, kategorie tříd využití země. V nejjednodušší podobě mají binární charakter (s vegetací či bez vegetace) a lze je pouze porovnávat.

Ordinální data – data, která lze seřadit do uspořádané posloupnosti podle určitého kritéria. Je známé pořadí kategorií, rozdíl však nemá smysl. Např. řád vodního toku, třída silnice, bonita půdy atd.

Typy geografických dat

Intervalová data – umožňují provádět i odečítání mezi kategoriemi definovat rozdíl mezi kategoriemi. Teplota vzduchu. Stupnice většinou nezačíná nulou. Poměr dat závisí na zvolených jednotkách

Poměrová data – vedle rovnosti, uspořádání a odčítání umožňují také dělení. Nula vyjadřuje neexistenci jevu – objem, délka …

1.2 Statistika a výpočetní technika

- Výpočetní technika zasahuje do všech etap statistického zpracování dat.
- Exploze výpočetní techniky umožňuje provádět výpočty, které byly dříve nerealizovatelné (z důvodů velkého objemu dat, pracnosti, …).
- Na druhou stranu však roste nebezpečí výběru nesprávného postupu.

Výhody počítačového zpracování I.

Přesnost a rychlost: Dobré počítačové programy (software) nám dají velmi rychle správné výsledky. Dřívější ruční zpracování dat bylo často zatíženo aritmetickými chybami a bylo časově velmi náročné.

Univerzálnost: Počítače zpřístupňují širokou škálu statistických metod a umožňují provést velmi rychle i rozsáhlé komplexní statistické analýzy.

Grafika: Počítače umožňují snadné grafické zobrazení pozorovaných dat a výsledků statistického zpracování.

Flexibilita: Velkou výhodou počítačů je, že umožňují rychle provést nové zpracování při změnách v datech či transformaci některých veličin.

Výhody počítačového zpracování II.

Nové veličiny: Snadno lze vytvářet nové veličiny pomocí požadovaných transformací.

Velikost datových souborů: Počítače umožňují zpracování velmi rozsáhlých souborů dat pomocí vhodného softwaru, což bylo ještě před deseti lety velmi obtížné.

Snadný přenos dat: Jakmile se jednou data dostala do počítače, lze je snadno přenést elektronicky (například pomocí Internetu) na jiné místo.

…ale

Nevýhody počítačového zpracování I.

Chyby v softwaru.

Ne všechny statistické programy jsou spolehlivé. Je dobré používat programy, které mají dobrou pověst a jsou používány již dostatečně dlouho, takže byla postupně odstraněna většina jejich chyb. K takovým programům patří například BMDP, SAS, SPSS, STATISTICA, S PLUS, STATGRAPHICS a další.

Univerzálnost.

Může vést k výběru nevhodné metody zpracování. Je velmi důležité, aby každý, kdo používá statistický software, si byl vědom úrovně svých statistických znalostí a užíval pouze ty metody, kterým rozumí. Pozor na používání neznámých statistických metod.

Nevýhody počítačového zpracování II

Černá skříňka.

Počítač vzdaluje uživatele od dat i metody zpracování. Statistická analýza se provádí automaticky, nová data se zpracovávají a výsledky se ukládají, aniž by byly posouzeny člověkem. Protože většinou výsledky zachycují jen průměrné efekty, může se zcela ztrácet citlivost k individuálním pozorováním.

Špatná data plodí špatné závěry.

Jestliže data jsou nasbírána či naměřena špatně (například jsou špatně kladené otázky v dotazníku), nelze očekávat, že závěry z takových dat budou správné. Sem náleží i nesprávné zpracování datových souborů, chybějící či ovlivněné (tzv. nehomogenní) údaje.

Statistický software

1. Programové vybavení založené na využití vlastního programovacího jazyka (Splus, SAS)
2. Interaktivní zpracování v „oknech“ MINITAB, SPSS, STATGRAPHICS
3. Programové vybavení s knihovnou statistických, matematických a grafických funkcí (EXCEL)