Základní pojmy a cíle statistiky [1]

[1]Tyto materiály byly vytvořeny za pomoci grantu FRVŠ číslo 1145/2004.

Předmět zkoumání Statistiky

Definice statistiky

Statistika zasahuje do mnoha oblastí našeho moderního života. Lze se s ní setkat jak ve velmi specifických vědních disciplínách, tak i v běžném každodenním životě. Příkladem mohou být výzkumy veřejného mínění, stanovení farmakokinetiky účinné látky v krvi sledované osoby při bioekvivalenčních studiích, vývoj ceny sledovaných akcí, či prostorové modely území využívající různých regresních modelů.

Statistiku lze definovat jako vědní obor, zabývající se hromadnými jevy a procesy.

Existují však i jiné definice.

Statistika zahrnuje jak získávání, tak i analýzu a interpretaci napozorovaných dat. Cílem statistického zpracování dat je podání informace o vlastnostech, povaze a zákonitostech projevujících se na pozorovaných datech.

Počátky statistiky

Slovo statistický je odvozeno od latinského slova "status". To lze přeložit jako "stav" či v přeneseném významu jako "stát". V 16. - 17. století se začínají vyskytovat slova "statistico" a "státística". Význam těchto slov spočíval v souhrnu znalostí o státních záležitostech, či jako označení znalce státních záležitostí. V polovině 18. století se v Německu vyučovala disciplína tzv. státověda "die Staatswissenschaft" ta byla zpočátku spíše slovním popisem.

Významná jména spojená se statistikou

Mezi přední světové statistiky bezesporu patří tato jména: Galileo Galilei, Blaise Pascal, Piere Fermat, Jacob Bernoulli, Thomas Bayes, Karl Friedrich Gauss, Andrej Andrejevič Markov, Andrej Nikolajevič Kolmogorov, Karel Pearson, atd ...

Hromadný jev

Jevy které se vyskytují v masovém měřítku a mohou se neustále opakovat nazýváme hromadnými jevy. V zásadě existují dva typy hromadných jevů:

- Prvním je opakované měření sledované vlastnosti jednoho předmětu
- Druhým typem je pak sledování vlastnosti(í) na předem definované množině sledovaných předmětů.

Hranice mezi individuálním a hromadným pojetím je neostrá.

Statistický soubor

Je definován jako množina sledovaných statistických jednotek. Statistika rozeznává **základní statistický soubor** a **výběrový statistický soubor**. Počet jednotek v analyzovaném statistickém souboru se nazývá **rozsah souboru**. Z hlediska rozsahu souboru se může jednat o **konečný** či **nekonečný** soubor.

Statistická jednotka

Statistické jednotky jsou obecně nositeli sledovaných vlastností. Pokud sledujeme výsledky hospodaření podniků v ČR, pak je statistickou jednotkou právě podnik. Ten je nositelem sledovaných vlastností tj. **statistických znaků** jako jsou zisk, obrat, rentabilita vloženého kapitálu atd...

Statistický znak

Je obrazem určité vlastnosti každé statistické jednotky ze statistického souboru. Hodnota statistického znaku je pak označením stupně dané vlastnosti. Hodnota statistického znaku se často nazývá pozorování.

Lze rozlišovat statistické znaky: **konstantní** jejich hodnota je pro všechny jednotky stejná. Dále **identifikační** a **proměnlivé** neboli variabilní. Právě tyto znaky - proměnné jsou předmětem statistického zkoumání. Zamyslete se nad tím, že vyjádření hodnot nemusí být vždy číselné.

Výsledkem statistického šetření, tj. analýzy dat je množina informací o hodnotách jednotlivých statistických znaků.

Indukce $\times$ dedukce

Indukce znamená usuzování z partikulárního na obecné.

Dedukce znamená usuzování na základě obecných poznatků či axiómů na partikulární.

Typy statistických znaků

V praxi se lze setkat s různými druhy dat. Příkladem mohou být záznamy o pohlaví dítěte, záznamy o postoji voliče k programu politické strany, počet potratů, ceny mléčných výrobků v různých obchodech, nebo soubor dat zachycující hmotnost substance obsažené v produkčním médiu po několikadenní fermentaci.

Nominální znaky

V případě práce s nominálními daty je nutné zavést **disjunktní kategorie**, které budou vyčerpávat všechny možnosti. Jednotlivé hodnoty znaku mohou být vyjádřeny slovně či pomocí čísel. Číselné vyjádření nominálního znaku je v tomto případě kódem, který slouží pouze k odlišení. Pokud existují pouze dvě hodnoty jichž může sledovaný znak nabývat, pak takový znak nazýváme **dichotomickým** či **binárním**. Příkladem takovéhoto znaku může být záznam narozených chlapců, nebo záznam výskytu infekční choroby A u sledovaných pacientů. Veličinám vyjádřeným v nominálním měřítku se také v některých případech říká **faktory**.

Ordinální znaky

Pokud pracujeme s daty které **lze uspořádat**, tj. jsme schopni rozlišit, zda jde o vyšší respektive nižší úroveň sledovaného znaku, pak lze zavést **ordinální stupnici**. Jsme tedy schopni pro každou dvojici hodnot určit, která v daném uspořádání předchází. Ordinální veličina bývá někdy nazývaná uspořádaným faktorem. Příkladem ordinálního znaku může být postoj zákazníka ke koupi určitého produktu, nebo stádium nemoci či stupeň zaplevelení.
Pro úplnost ze uvést, že nominální a ordinální znaky jsou znaky **kvalitativní**. Následující znaky patří naopak mezi znaky **kvantitativní**, neboli mezi znaky **kardinální**.

Intervalové znaky–metrické znaky

Na rozdíl od výše uvedených předpokládá, že mezi sousedními hodnotami jsou **stejné vzdálenosti**. Běžné je přiřazení jednotlivých hodnot stupnice přirozeným nebo reálným číslům. Typickým příkladem intervalového měřítka je Celsiova stupnice. Při zavedení takovéto stupnice **má smysl se ptát o kolik se dvě hodnoty liší**. Ptáme se tedy na rozdíl dvou hodnot. Lze dodat, že nula nemusí znamenat neexistenci sledované vlastnosti.

Poměrové veličiny–kardinální veličiny

Zde údaj znamená násobek přesně definovaného jednotkového množství. Nula v tomto případě znamená neexistenci sledované vlastnosti. Vyjádření pomocí reálných čísel je zde nezbytné. Při porovnávání hodnot v podílovém měřítku **se lze ptát jak na otázku o kolik se dvě hodnoty liší**, **tak i na otázku kolikrát je daná hodnota větší či menší než hodnota druhá**.

Podle počtu hodnot, kterých může daný statistický znak nabývat lze znaky dělit na diskrétní a spojité.

Diskrétní znaky

Jsou takové znaky, které **mohou nabýt konečného či spočetného počtu** hodnot. Příkladem může být počet lidí zaměstnaných ve státním podniku Budvar.

Spojité znaky

Jsou takové statistické znaky, které **mohou nabývat nekonečně mnoha hodnot v konečném či nekonečném intervalu**. Příkladem může být čistý zisk po zdanění.

Třídění

V podstatě jde o uspořádání získaných dat. Volba třídícího znaku je zpravidla dána účelem třídění. Dle počtu třídících znaků rozeznáváme třídění jednostupňové, **dvoustupňové**, či vícestupňové.

Jde-li o třídění kategoriálního znaku, nebo jedná-li se o numerický znak s malým

počtem hodnot, lze provádět tzv. **třídění prosté**.

Je-li třídícím znakem numerická proměnná s velkým počtem hodnot, pak je vhodnější provádět **intervalové třídění**.

Volba počtu intervalů je velmi důležitá, ale neexistuje žádné obecné doporučení pro jejich určení. Pokud je intervalů příliš mnoho, jsou většinou příliš krátké a informace obsažené v nich jsou nepřehledné. Pokud je jich naopak málo, pak jsou do stejného intervalu zařazeny zcela odlišné statistické jednotky. Vodítkem pro určení počtu intervalů může být **Sturgesovo pravidlo**. To je definováno takto:

$$k = 1 + 3,3 \log n \tag{1}$$

V některých případech je vhodné charakterizovat statistický soubor prostřednictvím tzv. četností. Zpravidla rozeznáváme několik druhů četností.

- **Absolutní četnost** zpravidla ji značíme prostřednictvím symbolu n_i a udává kolikrát se hodnota x_i znaku X vyskytuje v souboru.
- **Relativní četnost** p_i udává, v jak velké části souboru je hodnota znaku X rovna x_i
- **Kumulativní absolutní četnost** k_{n_i} udává počet statistických jednotek, u nichž byla hodnota statistického znaku $X \leq x_i$ tj.

$$k_{n_i} = n_1 + n_2 + \ldots + n_i$$

- **Kumulativní relativní četnost** K_{P_i} udává jaká část souboru vykazovala hodnoty $X \leq x_i$ tj.

$$k_{p_i} = p_1 + p_2 + \ldots + p_i$$

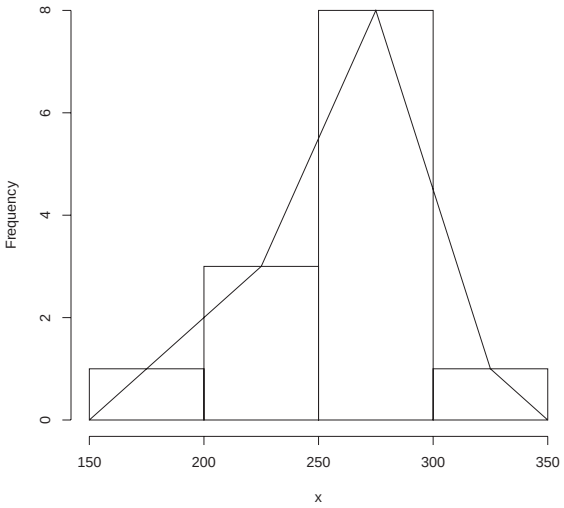
Relativní a kumulativní četnosti se počítají pouze pokud mají smysl. Při určování mezí je třeba volit meze tak, aby nedocházelo k nejasnostem, tj. abychom mohli každou hodnotu jednoznačně zařadit do určitého intervalu.

Grafické znázornění souboru

Znázornění provádíme zpravidla pomocí polygonu četností, nebo pomocí histogramu četností.

Polygon četností

Jde o spojnicový graf rozdělení četností kvantitativního znaku. Vzniká tak, že na x-vé ose zobrazíme hodnoty znaku X a na y-ové ose zobrazíme četnosti. Body pak spojíme lomenou čarou. V případě intervalového třídění bude každý interval zasoupen středem intervalu.



Histogram četností

Jde o grafické znázornění intervalového rozdělení četností. Na ose x zobrazíme meze intervalů hodnot X a nad nimi vytvoříme sloupce. Šířka sloupce odpovídá velikosti intervalu, výška pak četnostnostem znaků v příslušném intervalu.

Tyto dva grafy jsou sice nejužívanější, lze však využít i jiné typy grafického znázornění.

Box–plot

Pro vytvoření prvotní představy o vlastnosti souboru dat zpravidla využíváme několika typů grafů. Jedním z nich je tzv. Box-plot. Ten vhodným způsobem zachycuje minima, maxima, průměr či medián a dále pak horní a dolní kvartily. Nevýhodou tohoto grafu je, že nelze zhodnotit zda studovaný soubor má unimodální či vícemodální rozdělení.

Obrázek 1: Ukázka Box-plotu