

Štatistické pojmy

Last update:08.07.2019

Pre jednoduchšie pochopenie štatistiky je potrebné vysvetliť niektoré základné štatistické pojmy.

Štatistický súbor

Štatistika sa zaoberá javmi, ktoré nazývame **hromadné javy**. Rozlišujeme dva druhy hromadných javov. Jeden druh je taký, čo sa vyskytuje vo veľkom počte objektov - výška, hmotnosť, farba očí, pohlavie, vek, profesia, počet detí a pod. Druhý druh je taký, ktorý je výsledkom veľkého počtu opakovaní, napr. opakované váženie nejakého predmetu, opakované hody kockou, mincou a pod. Objekty, ktoré skúma štatistika voláme **štatistický súbor** - napr. množina osôb, ktoré žijú na nejakom území.

Základný štatistický súbor rozsahu **N** predstavuje množinu všetkých štatistických jednotiek. V prípade, že nemožno skúmať základný súbor (z časových, finančných alebo iných dôvodov), vytvárame z neho **výberový súbor** podľa vopred stanovených pravidiel. V prieskume sa údaje zbierajú len za časť populácie, tzv. vzorku. Tieto údaje sa potom použijú na odhad charakteristík celej populácie. V tomto prípade sa musí zabezpečiť, aby vzorka reprezentovala príslušnú populáciu. Napríklad podiel osôb vo veku do 18 rokov alebo podiel žien a mužov vo vybranej vzorke domácností musí odrážať realitu v celkovom počte obyvateľov.

Výber musíme urobiť tak, aby vybrané objekty mali vlastnosť, ktorú má aj celý základný súbor. Túto vlastnosť nazývame **štatistický znak** (môže to byť pohlavie, profesia, počet detí, vek, vzdelanie a pod.). Rôzne hodnoty štatistického znaku voláme obmeny, varianty.

(napríklad kvalitatívny znak pohlavie: muži = 1, ženy = 2).

Znak, ktorý má iba jednu nemennú hodnotu sa volá **konštanta**. V štatistike sa sledujú hlavne znaky, ktoré majú aspoň dve alebo viac obmien. Takýto znak má názov **premenná**.

Štatistické triedenie

Údaje získané štatistickým zisťovaním najlepšie zjednodušíme a sprehľadníme tým, že ich usporiadame do určitých skupín – tried.

Metóda usporiadania štatistických údajov do určitých skupín (tried) podľa určitého znaku sa nazýva triedenie. triedenie = rozdelenie štatistických jednotiek do takých skupín (tried), aby čo najlepšie vynikli charakteristické vlastnosti skúmaných javov;

triediaci znak = štatistický znak, ktorý je kritériom triedenia štatistického súboru;

trieda = skupina štatistických jednotiek s rovnakou hodnotou (variantom) znaku.

Príklady triedenia štatistických súborov:

- pri známovaní v škole rozlišujeme triedu jednotiek, triedu dvojok, trieda trojok, ...

- triedenie osôb podľa dosiahnutej úrovne vzdelania - skupina so základným vzdelaním, skupina so stredným vzdelaním, bez maturity, skupina so stredným vzdelaním s maturitou.

Keď má štatistický znak malý počet obmien, tak každá jeho obmena predstavuje samostatnú triedu. Na prezentáciu takýchto údajov sa používa tabuľka rozdelenia početností, ktorá sa tiež volá **frekvenčná tabuľka**. Obvykle ide o **prvýkrok pri spracovaní štatistických údajov**.

Rozdelenie početnosti – frekvenčná tabuľka

Tabuľky početnosti slúžia na zobrazenie údajov, je to **prehľadný spôsob usporiadania údajov**. Štatistické údaje sa zaraďujú do tried s určitým počtom a vytvorí sa rozdelenie početnosti.

Tabuľky početnosti obsahujú informácie:

- o sledovaných hodnotách znaku a o frekvencii jeho výskytu v štatistickom súbore (absolútna početnosť), číslo n_j , udáva, koľko jednotiek štatistického súboru patrí do triedy s istou vlastnosťou alebo koľko jednotiek súboru má hodnotu znaku x_j

- o pomernom zastúpení hodnôt znaku v súbore (relatívna početnosť), = číslo v_j , udáva, aká časť súboru patrí do triedy s danou vlastnosťou; je určená podielom absolútnej početnosti n_j a rozsahu súboru n .

- o zaradení individuální hodnoty znaku v celkovom rozdelení početnosti daného znaku (kumulatívne početnosti)

Príklad

V domácnostiach sa zisťoval počet obytných miestností s možnými výsledkami: 1, 2, 3, 4, 5+ (znamená 5 a viac miestností). Pri tomto zisťovaní bolo náhodne vybratých 25 respondentov s nasledujúcimi výsledkami, ktorí žili v bytoch s nasledovným počtom miestností: 1, 3, 2, 4, 5+, 2, 3, 2, 5+, 1, 4, 2, 3, 3, 5+, 3, 3, 2, 4, 4, 2, 3, 4, 3, 2

Vytvorte tabuľku rozdelenia početností a relatívnych početností. Pokúste sa o stručnú interpretáciu získaných výsledkov, ktorá bude obsahovať najpočetnejšiu a najmenej početnú hodnotu sledovaného znaku.

Riešenie

Sledovaný znak označíme symbolom x (počet obytných miestností), ktorý môže nadobudnúť hodnoty x_j ($j = 1, 2, 3, 4, 5+$). Absolútne početnosti označíme n_j , relatívne početnosti v_j .

Tabuľka: Absolútne a relatívne početnosti podľa počtu obytných miestností **charakteristiky polohy (stredú)** - vyjadrujú určitú úroveň znaku, okolo ktorej sú ostatné hodnoty znaku viac alebo menej koncentrované; **charakteristiky variability (premenlivosti)** - vyjadrujú, ako sa hodnoty znaku prvkov súboru líšia od zvolenej charakteristiky polohy, respektíve od seba navzájom

Charakteristiky polohy (úrovne, stredú) - čísla, ktoré určitým spôsobom charakterizujú tzv. "priemernú hodnotu" sledovaného štatistického znaku. Delíme ich na dve základné skupiny:

- **priemery** - môžu byť jednoduché alebo vážené

- a) aritmetický priemer
- b) geometrický priemer
- c) harmonický priemer

- ostatné stredné hodnoty - medián, modus

Aritmetický priemer

Aritmetický priemer, v bežnom jazyku sa nazýva **priemer**, je súčtom hodnôt skupiny čísel vydelených počtom čísel v skupine. Priemer možno chápať ako zhrnutie skupiny čísel do jedného čísla.

Aritmetický priemer \bar{x} vyjadruje, aký objem hodnôt znaku X pripadá v priemere na jednu jednotku súboru. Je definovaný vzťahom:

N - rozsah súboru,

x_i - hodnota znaku X u i -tej jednotky.

Príklad

Máme 9 čísel v skupine: 10, 12, 11, 15, 13, 35, 41, 23, 20. Súčet týchto 9 čísel je 180. Potom sa suma 180 delí číslom 9, aby sa dosiahol priemer. Priemer je $180/9 = 20$.

Príklad

Na základe štatistických hlásení o narodeniach a úmrtiach a súboroch o sťahovaní spracúva Štatistický úrad SR každoročnú bilanciu počtu obyvateľov Slovenskej republiky za všetky obce.

K dispozícii máme stredný (priemerný) počet trvale bývajúceho obyvateľstva v jednotlivých krajoch SR k 31. 12. 2018 (pre lepšiu prácu s údajmi bol počet obyvateľov zaokrúhlený a uvádza sa v tisícoch): Bratislavský kraj 646, Trnavský kraj 562, Trenčiansky kraj 588, Nitriansky kraj 679, Žilinský kraj 691, Banskobystrický kraj 650, Prešovský kraj 823, Košický kraj 798.

Vypočítajte aritmetický priemer počtu obyvateľov žijúcich v jednom kraji Slovenskej republiky.

Riešenie

Uvedené údaje dosadíme do vzorca pre aritmetický priemer. Nesmieme zabudnúť, že do vzorca budeme dosadzovať čísla uvedené v tisícoch, preto aj výsledný počet obyvateľov bude v tisícoch.

$$\bar{x} = (646+562+588+679+691+650+823+798)/8 = 679,625 \text{ (približne) } 680$$

Môžeme konštatovať, že v každom kraji Slovenskej republiky žije v priemere 680 tisíc obyvateľov.

V oficiálnych štatistikách je najčastejším typom priemeru **vážený priemer**, pretože málokedy majú všetky položky

rovnaký význam. Vo váženom priemere sa každá položka, ktorá sa berie do úvahy, vynásobí číslom (váhou), ktoré odráža relatívny význam položky, potom sa výsledok spočíta a následne sa vydolí počtom položiek.

Vážený aritmetický priemer \bar{x} používame, ak pracujeme s triedeným súborom hodnôt znaku X . Na jeho výpočet použijeme vzťah :

N - rozsah súboru,

x_i - hodnota znaku X u i -tej jednotky.

Modus

Mo je najčastejšie sa vyskytujúca hodnota znaku X , resp., v prípade triedeného súboru hodnota s najväčšou absolútnou početnosťou.

Príklad

V domácnostiach sa zisťoval počet obytných miestností s možnými výsledkami: 1, 2, 3, 4, 5+ (znamená 5 a viac miestností). Pri tomto zisťovaní bolo náhodne vybraných 25 respondentov, pričom každý býval v byte s určitým počtom miestností (prvý býval v jednoizbovom byte, druhý v trojizbovom ...: 1, 3, 2, 4, 5+, 2, 3, 2, 5+, 1, 4, 2, 3, 3, 5+, 3, 3, 2, 4, 4, 2, 3, 4, 3, 2

Hodnota 3 sa v tomto príklade vyskytuje najčastejšie, to znamená, že modus je 3.

Mod (x) = 3, hodnota 3 je najčastejšia, vyskytuje sa až 8 krát.

Medián

Medián M_e je hodnota, ktorá súbor zistených hodnôt delí na 2 rovnako početné skupiny, t. j. skupiny, z ktorých prvá obsahuje 50 % štatistických jednotiek, ktoré majú hodnotu znaku X menšiu ako medián, druhá obsahuje 50 % zvyšných štatistických jednotiek, ktoré majú hodnotu väčšiu ako medián. Ak zoradíme všetky hodnoty znaku podľa veľkosti do neklesajúcej (resp. nerastúcej) postupnosti, tak mediánom bude hodnota, ktorá je v strede uvažovanej postupnosti.

$$M_e = x_{k+1}, \text{ ak } N = 2k + 1,$$

$$M_e = (x_k + x_{k+1})/2, \text{ ak } N = 2k$$

V prípade triedeného súboru:

a - horná hranica triedy, ktorá predchádza mediálny interval,

N - rozsah súboru,

n_1 - počet všetkých prvkov pod mediálnym intervalom,

n_2 - počet prvkov mediálneho intervalu,

h - šírka triedy.

Príklad

Veľkosti hrán piatich kociek sú 6, 8, 9, 10 a 11 (v cm). Stanovte hodnotu mediánu:

a) pre ich povrch,

b) pre objem týchto kociek.

Riešenie

a) Najprv si vypočítame povrchy (v cm^2) jednotlivých kociek. Pre dané rozmery hrán sa postupne rovnajú: 216, 384, 486, 600, 726.

Teraz je možné určiť medián:

$$\text{Med } (x) = x_{(n+1)/2} = x_{6/2} = x_3 = 486, \text{ pretože } n \text{ je nepárne.}$$

b) Najprv si vypočítame objemy (v cm^3) jednotlivých kociek. Pre dané rozmery hrán sa postupne rovnajú: 216, 512, 729, 1000, 1331. Teraz je možné určiť medián:

$$\text{Med } (x) = x_{(n+1)/2} = x_{6/2} = x_3 = 729, \text{ pretože } n \text{ je nepárne.}$$

Variabilita

Charakteristiky variability (premenlivosti, rozptýlenia) = čísla, ktoré udávajú, do akej miery sa hodnoty znaku odchyľujú od zvolenej charakteristickej polohy, respektíve od seba navzájom. Patria k nim:

- variačné rozpätie
- rozptyl
- smerodajná odchýlka

Variačné rozpätie v_r je iba približnou charakteristikou variability hodnôt sledovaného znaku. Je definovaný ako rozdiel najväčšej a najmenej hodnoty kvantitatívneho znaku, t. j. $v_r = X_{\max} - X_{\min}$.

Rozptyl

Rozptyl σ^2 predstavuje aritmetický priemer druhých mocnín (štvorcov) odchýlok od priemeru \bar{x} . Je definovaný vzorcom:

V prípade triedeného súboru ho vypočítame podľa vzťahu:

N - rozsah súboru,

m - počet tried v súbore,

n_j - absolútna početnosť j -tej triedy ($j = 1, 2, \dots, m$),

x_j - hodnota znaku X , ktorá reprezentuje j -tu triedu.

K dôležitým vlastnostiam rozptylu patrí:

a) Rozptyl konštanty je rovný nule.

b) Ak pripočítame ku všetkým hodnotám znaku konštantu, rozptyl sa nezmení.

c) Ak všetky hodnoty znaku vynásobíme konštantou a , potom rozptyl takto vzniknutých hodnôt je rovný súčinu rozptylu pôvodného súboru a druhej mocniny konštanty a .

Príklad

V predajni hospodárskych potrieb dostali zásielku 18 metiel, pričom metly mali mať predpísanú dĺžku 120 cm. Meraním sa však zistilo, že skutočné dĺžky sú nasledujúce:

117, 122, 125, 118, 119, 120, 122, 123, 116, 117, 121, 120, 123, 120, 119, 121, 124, 115.

Určite rozptyl metiel.

Riešenie

Najskôr vypočítame priemer:

$$\bar{x} = (117+122+125+118+\dots+121+124+115) / 18 = 2162 / 18 = \mathbf{120,1}$$

Dosadením do vzorca vypočítame rozptyl:

Výsledný rozptyl sa rovná 6,69.

Štandardná (smerodajná) odchýlka δ je definovaná ako $\delta = \sqrt{\sigma^2}$ a udáva, ako sa v priemere v danom súbore odchyľujú hodnoty znaku od aritmetického priemeru.