6. prednáška

Štatistika – základné pojmy

"Štatistika je jediná veda, ktorá pri použití rovnakých vzorcov, umožňuje rôznym expertom dostať rôzne výsledky."

" Štatistika je presný súčet nepresných čísel. "

Čo je vlastne štatistika? Slovo **štatistika** má pôvod v minulosti vzdialenej niekoľko storočí. "A v tých dňoch vyšiel rozkaz od cisára Augusta vykonať súpis ľudí na celom svete. Tento prvý súpis sa konal, keď Sýriu spravoval Kvirínius. A všetci šli dať sa zapísať, každý do svojho mesta. Vybral sa aj Jozef z galilejského mesta Nazaret do Judey, do Dávidovho mesta, ktoré sa volá Betlehem, lebo pochádzal z Dávidovho domu a rodu, aby sa dal zapísať s Máriou, svojou manželkou, ktorá bola v požehnanom stave..... "
Už tu, pri prvom sčítaní ľudu v dejinách, sa stretávame s používaním nástrojov vlastných štatistike súčasného obdobia. Ak budeme slovo **štatistika** ďalej rozpytvávať, zacítime v ňom latinský základ – *status* – teda *stav*, ale aj *štát* – stav vecí verejných. Ak siahneme do výkladového slovníka alebo do úvodných kapitol učebníc štatistiky, dozvieme sa, že " štatistika sa zaoberá štúdiom zákonitostí hromadných javov. " Okrem toho sa v učebniciach dočítame, že vo väčšine prípadov je pod pojmom štatistika myslená **matematická štatistika**, čo je odbor matematiky, ktorý aplikuje teóriu pravdepodobnosti pri hľadaní správnych metód usudzovania z neúplných údajov zaťažených navyše aj náhodnou chybou.
Vidíme, že je mnoho významov slova štatistika. Hlavnou ambíciou tejto časti predmetu pravdepodobnosť a štatistika je vybudovanie základov pre správne pochopenie významu slova **štatistika** a pre využitie niektorých štatistických postupov na poznávanie a chápanie sveta, ktorý nás obklopuje. Ak v minulosti sa štatistika považovala za jednoduchý súbor čísel o obyvateľstve, hospodárstve a pod., v súčasnosti sa štatistika považuje za množinu techník používaných na zhromažďovanie, analyzovanie, prezentovanie a interpretáciu dát. Štatistika sa využíva , napr. v ekoómii, na tvorbu podkladov pre rozhodovnie. Umožňuje lepšie porozumieť zdrojom variability a objavovať vzťahy medzi dátami a v dôsledku toho prijímať lepšie rozhodnutia.
Dáta sú fakty, ktoré sa zhromažďujú, spracúvajú a analyzujú na prezntáciu a interpretáciu. Všetky dáta, ktoré sa zhromaždili pre konkrétnu štúdiu, tvoria **súbor dát.** Dáta sú zobrazením istej časti reálneho sveta a forma zobrazenia reálneho sveta môže byť rôzna – fotografia, mapa, kresba, čísla.... My budeme pod dátami rozumieť v prevažnej miere čísla. Také výseky z reálneho sveta, ktoré zahŕňajú viac objektov majúcich nejakú vlastnosť, nazývame **populácia**. Zobrazením buď všetkých alebo len niektorých objektov populácie vznikajú **štatistické dáta**. Ďalšími pojmami, s ktorými sa v štatistike budeme stretávať, sú *štatistická jednotka* a *štatistický znak*.
Štatistická jednotka – základný prvok štatistického súboru – je objekt po-

zorovania, na ktorom skúmame prejav sledovanej vlastnosti. Napr. študent FRI, domácnosť,...
Štatistický znak – vonkajšia merateľná vlastnosť štatistických jednotiek. Napr. výška, váha, VŠP (študenta FRI), spotreba potravín, počet členov, spotreba benzínu (danej domácnosti).
Všeobecne sa rozlišujú dva typy štatistických znakov – *kvalitatívne* a *kvantitatívne*.
Kvalitatívne znaky – nadobúdajú hodnoty, ktoré umožňujú identifikovať znak každej jednotky (napr. pohlavie, národnosť, štátna príslušnosť, ...)
Kvantitatívne znaky – nadobúdajú číselné hodnoty, ktoré reprezentujú množstvo v nejakých merných jednotkách.
Kvantitatívny znak môže byť **diskrétny** – počet detí v domácnosti, počet áut v domácnosti, počet tovarov, ktoré nakúpil zákazník pri jednom nákupe,..., alebo **spojitý** – čas čakania zákazníka v rade pri pokladni, váha, výška, ...
Pri analýze dát používame tak *opisnú* ako aj *induktívnu štatistiku*.
Opisná štatistika – je množina techník (metód), ktoré umožňujú charakterizovať a predstaviť dáta v takej forme, ktorá umožní čo najľahšie ich pochopiť. Napr. vypočítam si priemerný počet bodov z prvej semestrálnej písomky v študijnej skupine, ktorú učím napr. v utorok večer. Charakterizujem úroveň vedomostí študijnej skupiny jedným číslom – aritmetickým priemerom – a nerobím zovšeobecnenia aj pre iné skupiny. Aplikovala som jednu metódu opisnej štatistiky. Ďalšími príkladmi aplikácie opisnej štatistiky, pomocou ktorých sa prezentujú dáta sú *diagramy* a *tabuľky*.
Induktívna štatistika – je množina techník, ktoré umožňujú urobiť úsudky o *základnom súbore* na základe výsledkov analýzy z nich vybratých výberov. Metódy induktívnej štatistiky sú napr. odhadovanie parametrov, testovanie hypotéz, ...
Alfou aj omegou správnych úsudkov o základnom súbore je dobrý výberový súbor. Ako vytvoriť " dobrý výber "? Dobrý výber je taký, keď každá jednotka má rovnakú šancu vo výbere figurovať. Aby sme túto požiadavku splnili, je potrebné urobiť náhodné vyberanie. Jednotky získané náhodným vyberaním vytvárajú *náhodný výber*.
Pričom **základný súbor** môžeme definovať ako množinu všetkých jednotiek uvažovaných v konkrétnej štúdii a **výberový súbor (výber)** je vybratá časť základného súboru. Poznamenajme, že v štatistike sa pod pojmom *súbor dát* nemyslí súbor jednotiek, ktoré doň patria, ale častejšie súbor hodnôt štatistického znaku, ktorý nás zaujíma.

Opisná štatistika – tabuľková a grafická prezentácia dát

V ďalšom sa budeme venvať len výberovým súborom, špeciálne náhodným výberom. Keď nás zaujíma len jedna premenná (jeden štatistický znak) jedná sa o **jednorozmernú opisnú štatistiku**.
Predpokladajme, že máme daný náhodný výber `n` štatistických jednotiek. **n** nazývame **rozsahom náhodného výberu**. Dáta náhodného výberu je vhodné rozumne spracovať tzv. **triedením**.

Najjednoduchším triedením je ich usporiadanie podľa veľkosti, obyčajne vzostupne, čím získame **variačný rad** $x_1 \leq x_2 \leq \cdots \leq x_n$, kde x_i sú namerané hodnoty sledovaného znaku.
Keď je dát veľký počet, môže byť užitočné koncentrovať ich vo forme *rozdelenia početností*. **Rozdelenie početností** je tabuľkové zhrnutie dát, ktoré ukazuje počet (početnosť) jednotiek v každej z tried.
Najskôr sa budeme zaoberať rozdelením početností, v ktorom *každá hodnota premennej tvorí samostatnú triedu*. A venovať sa budeme charakterizácii kvantitatívnych dát. Pre každú nameranú hodnotu x_i variačného radu zistíme, koľkokrát sa vyskytuje medzi nameranými hodnotami, čo označíme $\mathbf{n_i}$ a nazývame **absolútna početnosť i–tej triedy**. Získané hodnoty zapíšeme do tabuľky:

x_i	x_1	x_2	x_3	$\ldots$	x_k
n_i	n_1	n_2	n_3	$\ldots$	n_k

pričom $n_1 + n_2 + n_3 + \cdots + n_k = n$.
Ďalšou možnosťou ako vyjadriť početnosť i–tej triedy je početnosť definovaná vzťahom

$$f_i = \frac{n_i}{n} \quad \text{pre} \quad i = 1, 2, \ldots, k,$$

ktorá predstavuje podiel počtu hodnôt x_i v celkovom počte pozorovaných hodnôt. Nazývame ju **relatívna početnosť i–tej triedy** a často sa uvádza v percentách. **Kumulatívnu absolútnu početnosť i–tej triedy** označujeme N_i a udáva počet jednotiek, ktoré majú hodnotu štatistického znaku menšiu alebo rovnú ako je jej hodnota v i–tej triede. Vypočítame ju podľa vzťahu

$$N_i = \sum_{j=1}^{i} n_j,$$

kde n_j je početnosť j–tej triedy, $j = 1, 2, \ldots, k$.
Kumulatívnu relatívnu početnosť i–tej triedy označujeme F_i a udáva podiel počtu jednotiek, ktoré majú hodnotu premennej menšiu alebo rovnú ako je jej hodnota v i–tej triede. Vypočítame ju podľa vzťahu

$$F_i = \sum_{j=1}^{i} f_j,$$

kde f_j je početnosť j–tej triedy, $j = 1, 2, \ldots, k$.
Na grafické zobrazenie rozdelenia početností sa najčastejšie používa *stĺpcový diagram*, *kruhový diagram* a *diagram kumulatívnych početností*.
Stĺpcový diagram tvoria obdĺžniky, ktorých základne korešpondujú s triedami a ich výšky sa rovnajú početnostiam.
Kruhový diagram – nakreslí sa kruh, ktorý reprezentuje všetky dáta v súbore. Potom sa použijú relatívne početnosti na rozdelenie kruhu na výseky.
Diagram kumulatívnych početností je stĺpcový diagram, v ktorom sa výšky stĺpcov rovnajú kumulovaným početnostiam (absolútnym alebo relatívnym).

Keď je počet rozličných hodnôt štatistického znaku veľký, užitočné je utvoriť triedy tak, že každú z nich tvorí viac ako jedna hodnota. S takýmito situáciami sa stretávame u spojitej premennej (pri spojitom štatistickom znaku). V takomto prípade sa vo všeobecnosti odporúča **počet tried** 5–10, resp. 8–20, pričom sa prihliada na povahu skúmaného znaku a rozsah náhodného výberu. Cieľom je určiť dostatočný počet tried na to, aby sa čo najlepšie ukázala variabilita dát. Zároveň počet tried nemá byť ani príliš veľký, aby niektoré triedy neobsahovali len veľmi malý počet jednotiek.
Ak označíme k počet tried, tak k môžeme určiť jedným z nasledovných spôsobov:

$$k \approx 1 + 3,3 \log n \quad \text{alebo} \quad k \approx 5 \log n \quad \text{alebo} \quad k \approx \sqrt{n},$$

kde n je rozsah náhodného výberu.
Ak k je počet tried, potom **približnú šírku triedy** h vypočítame

$$h \approx \frac{x_{max} - x_{min}}{k},$$

kde x_{max} je maximálna a x_{min} je minimálna hodnota v súbore.
Nech l_i je označenie pre dolnú, u_i označenie pre hornú hranicu triedy a c_i bude označovať stred triedy, platia tieto triviálne vzťahy:

$$h_i = u_i - l_i, \quad c_i = \frac{l_i + u_i}{2} = l_i + \frac{h_i}{2} = u_i - \frac{h_i}{2} \quad \text{pre} \quad i = 1, 2, \ldots, k$$

$$\text{a} \quad l_i = u_{i-1} \quad \text{pre} \quad i = 2, 3, \ldots, k$$

Aby každá nameraná hodnota $x_j, \quad j = 1, 2, \ldots, n$ bola jednoznačne zaradená do jednej z tried, triedy konštruujeme v tvare intervalov $\langle l_i, u_i \rangle$. Pri voľbe dolnej hranice prvého intervalu dbáme, aby bola aspoň trochu menšia ako x_{min} a horná hranica posledného intervalu aspoň o trochu väčšia ako x_{max}. Hranice volíme tak, aby boli čo najzaokrúhlenejšie číselné hodnoty.
Na grafickú reprezentáciu hodnôt v tomto prípade sa najčastejšie používa *histogram*, *polygón* a *ogivná krivka*.
Histogram rozdelenia početností tvoria vedľa seba položené obdĺžniky, ktorých základne sú šírky tried a ich výšky sú absolútne alebo relatívne početnosti. Plochy obdĺžnikov sú úmerné triednym početnostiam.
Polygón rozdelenia početností je alternatívny diagram k histogramu. Konštruuje sa tak, že najskôr sa zakreslia body, z ktorých každý má súradnice: (stred triedy, početnosť) a následne sa spoja úsečkou vždy dva bezprostredne susedné body.
Ogivná krivka je diagram kumulatívnych početností. Zakreslia sa body, z ktorých každý má súradnice (horná hranica triedy, kumulatívna početnosť) a spoja sa úsečkou bezprostredne susedné body.