7. prednáška

Opisné charakteristiky

Kvantitatívne dáta možno vyjadriť aj v koncentrovanej podobe, a to pomocou *opisných charakteristík*. Charakteristiky, ktoré sa počítajú z dát výberu, sa nazývajú *výberové*. Budeme pracovať s dátami, ktoré získame z náhodného výberu. Prvky náhodného výberu (reprezentatívnej vzorky) X_i sú nositeľmi hodnôt x_i sledovaného znaku, možno ich považovať za náhodné premenné. Matematicky definujeme náhodný výber takto:
Náhodný výber je n–tica náhodných premenných $X_1, X_2, \ldots, X_n$, ktoré sú nezávislé a majú rovnaké rozdelenie pravdepodobnosti ako náhodná premenná X, ktorej hodnoty v základnom súbore pozorujeme.
Napr. ak vyšetrujeme výšku desaťročných chlapcov, môžeme výšku považovať za náhodnú premennú X. Náhodným výberom vybraní chlapci z populácie desaťročných chlapcov sú nositeľmi hodnôt sledovaného znaku (výšky), ozn. $X_1, X_2, \ldots, X_n$ – vybraní chlapci tvoria náhodný výber. Akonáhle odmeriame výšku každého chlapca z náhodného výberu, dostávame n–ticu reálnych čísel $x_1, x_2, \ldots, x_n$, ktorú nazývame *realizácia náhodného výberu*.
Tak, ako sme v teórii pravdepodobnosti číselné charakteristiky náhodných premenných definovali pomocou teoretických momentov, v štatistike budeme postupovať rovnako, ale použijeme *výberové momenty*.
Nech $X_1, X_2, \ldots, X_n$ je náhodný výber zo základného súboru. Pod **počiatočným výberovým momentom k–teho rádu** rozumieme výraz:

$$v_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Pod **centrálnym výberovým momentom k–teho rádu** rozumieme výraz:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^k,$$

kde

$$\overline{X} = v_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$$

nazývame **aritmetický priemer**.
Výberové charakteristiky v zásade delíme na *charakteristiky polohy* a *charakteristiky variability*.

Charakteristiky polohy

Hádam najznámejšou a najpoužívanejšou charakterristikou polohy je **aritmetický priemer**:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

kde n je rozsah náhodného výberu a x_i je hodnota sledovaného štatistického znaku i–tej jednotky.
V prípade, že hodnoty sledovaného štatistického znaku sú triedené, tak výberový aritmetický priemer počítame podľa vzťahu

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m} n_i x_i,$$

kde n je rozsah náhodného výberu, m je počet tried, n_i je absolútna početnosť i–tej triedy a x_i je hodnota sledovaného štatistického znaku i–tej triede.
Aritmetický priemer vyjadruje, aký objem hodnôt znaku pripadá v priemere na jednotku súboru.
Ďalšou charakteristikou polohy je **modus** M_o– je to najčastejšie sa vyskytujúca hodnota štatistického znaku. Pokiaľ sa v neklesajúcej postupnosti hodnôt znaku budú vyskytovať hodnoty s maximálnou početnosťou vedľa seba, *modus bude ich priemer*. Pokiaľ sa tieto vedľa seba nevyskytujú, každú z nich uvádzame ako modus.
V prípade, že hodnoty znaku sú utriedené do intervalov, možno priamo nájsť len *modálny interval*. Hodnota modusu v modálnej triede sa dá vypočítať podľa komplikovaného vzťahu – nebudeme uvádzať.
Hoci je modus najľahšie pochopiteľnou mierou polohy, jeho používanie ako miery polohy je najmenšie. Je to v dôsledku veľkej štandardnej chyby a skutočnosti, že často nie je určený jednoznačne.
Medián M_e je hodnota, ktorá súbor nameraných hodnôt delí na dve rovnako početné skupiny. Ak zoradíme všetky hodnoty znaku do neklesajúcej postupnosti, tak medián je hodnota, ktorá

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & \text{ak n je nepárne} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{ak n je párne} \end{cases}$$

Medián predstavuje najtypickejšiu charakteristiku úrovne hodnôt znaku tak, ako ju chápe väčšina ľudí. Okrem toho je stabilný, neovplyvňujú ho odľahlé hodnoty. Ale aj napriek tomu, že aritmetický priemer je ovplyvniteľný extrémnymi hodnotami, má mnohé zaujímavé vlastnosti, preto sa v praxi používa najčastejšie.

Charakteristiky variability

V mnohých praktických situáciách je okrem charakteristiky polohy užitočné poznať aj *charakteristiky variability*. Predstavme si situáciu, že máme firmu a hľadáme dodávateľov bližšie nešpecifikovaného materálu. Do úvahy pripadajú dvaja dodávatelia. K dispozícii máme dáta o lehotách ich dodávok pre iné firmy. Priemerná dodacia lehota bola u oboch dodávateľov rovnaká. V tom prípade pri výbere zaváži variabilita – vyberieme si firmu s menšou variabilitou.
V nasledujúcom si všimneme niektoré charakteristiky variability podrobnejšie. Najjednoduchšou charakteristikou variability je **variačné rozpätie R**

(*Range*). Je to charakteristika, ktorá je ovplynená extrémnymi hodnotami, používa sa menej. Vypočíta sa podľa vzťahu:

$$R = x_{max} - x_{min}.$$

Výberový rozptyl (*Sample Variance*) je najpoužívanejšia miera variability. Definujeme ho dvomi spôsobmi (prečo, povieme neskôr):

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2 = \overline{X^2} - \overline{X}^2$$

V prípade triedeného súboru:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{m} n_i\left(X_i - \overline{X}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{m} n_i X_i^2\right) - \overline{X}^2$$

Druhý spôsob, ako definovať výberový rozptyl je:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \left(\frac{1}{n-1}\sum_{i=1}^{n} X_i^2\right) - \frac{n}{n-1}\overline{X}^2$$

Zrejme platí:

$$S_{n-1}^2 = \frac{n}{n-1}S_n^2.$$

Vidíme, že výberový rozptyl je hodnota vždy väčšia alebo rovná nule. Nule je rovná len v prípade, že všetky namerané hodnoty sú rovnaké ($x_i = \overline{X}$ pre všetky $i = 1, 2, \ldots, n$). Väčšiu vypovedaciu hodnotu ako výberový rozptyl pre laickú verejnosť má **smerodajná odchýlka** (*Standard Deviation*), ktorá je odmocninou z výberového rozptylu.

$$S_n = \sqrt{S_n^2}, \text{ resp.} \quad S_{n-1} = \sqrt{S_{n-1}^2}.$$

Jej výhodou oproti rozptylu je to, že má rovnaký rozmer ako namerané hodnoty a aritmetický priemer. Užitočnou charakteristikou využívajúcou smerodajnú odchýlku je **štandardná chyba** (*Standard Error*) $\frac{S_n}{\sqrt{n}}$, resp. $\frac{S_{n-1}}{\sqrt{n}}$. Dokumentuje, s akou presnosťou aritmetický priemer charakterizuje sledovaný štatistický znak v základnom súbore. Čím je štandardná chyba menšia, tým presnejšie aritmetický priemer charakterizuje celý základný súbor.

Ďalšie charakteristiky rozdelenia pozorovaných hodnôt

Okrem polohy a variability je možné jediným číslom vyjadriť i ďalšie charakteristiky vystihujúce tvar rozdelenia (rozloženia) dát. Týmito charakteristikami sú **koeficient šikmosti g_1** (*Skewness*) a **koeficient špicatosti g_2** (*Kurtosis*).

Šikmosť je definovaná vzťahom $g_1 = \frac{m_3}{S_{n-1}^3}$

Nulová šikmosť znamená, že rozdelenie je symetrické okolo aritmetického priemeru. Kladná šikmosť znamená, že dáta sú koncentrované v ľavej časti variačného rozpätia, záporná šikmosť znamená opak – dáta sú koncentrované v pravej časti variačného rozpatia.
Špicatosť je definovaná vzťahom $g_2 = \frac{m_4}{S_{n-1}^4} - 3$
Špicatosť vzťahujeme k najčastejšie sa vyskytujúcemu rozdeleniu – k normálnemu rozdeleniu, ktoré má pomer $\frac{m_4}{S_{n-1}^4}$ rovný 3. Kladná špicatosť znamená, že rozdelenie dát je špicatejšie ako normálne rozdelenie a záporná špicatosť znamená, že rozdelenie dát je plochejšie ako normálne. Štvorica čísel *aritmetický priemer*, *smerodajná odchýlka*, *šikmosť* a *špicatosť* nám umožňuje urobiť si predstavu o tvare rozdelenia dát a porovnávať rôzne dáta.

Rozdelenia výberových charakteristík

Ak poznáme rozdelenie náhodného výberu $X_1, X_2, \ldots, X_n$, ktoré je dané distribučnou funkciou $F(x, \theta)$, bude nás zaujímať aj rozdelenie náhodných premenných, ktoré sú funkciami náhodného výberu a nezávisia od parametrov rozdelenia náhodného výberu. Takéto funkcie nazývame **štatistikami** a označujeme $T_n = T(X_1, X_2, \ldots, X_n)$. Poznamenajme, že všetky doteraz definované výberové charakteristiky sú štatistiky a považujeme ich za náhodné premenné. Keď do vzťahu, ktorý definuje príslušnú výberovú charakteristiku, dosadíme realizácie náhodného výberu, dostaneme číslo. Napríklad vzťah $\frac{1}{n}\sum_{i=1}^{n} X_i$ aj $\frac{1}{n}\sum_{i=1}^{n} x_i$ predstavuje aritmetický priemer, v prvom prípade pozeráme na aritmetický preiemer ako na náhodnú premennú, v druhom prípade aritmetický priemer je číslo. Budeme si to pripomínať a budeme medzi nimi rozlišovať.
Ukážeme si, že ak $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia so známou strednou hodnotou $E(X)$ a disperziou $D(X)$, potom $E(\overline{X}) = E(X)$
a $D(\overline{X}) = \frac{1}{n}D(X)$:

Špeciálne, ak náhodný výber pochádza z normálneho rozdelenia $N(m, \sigma^2)$, potom $\overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$.

Ak náhodný výber pochádza z ľubovoľného rozdelenia so strednou hodnotou $E(X)$ a disperziou $D(X)$, potom pre veľké náhodné výbery centrálna limitná veta umožňuje aproximovať rozdelenie výberového priemeru normálnym rozdelením: $\overline{X} \sim N\left(E(X), \frac{D(X)}{n}\right)$. V praxi sa táto aproximácia používa pre $n \geq 30$.

Vypočítať $E(S_n^2),\ D(S_n^2),\ E(S_{n-1}^2),\ D(S_{n-1}^2)$ nie je celkom jednoduché, presahuje rámec nášho kurzu štatistiky. Praktický význam pre nás bude mať poznatok, že ak náhodný výber pochádza z normálneho rozdelenia, potom náhodná premenná

$$\frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2(n),$$

resp.

$$\frac{(n-1)S_{n-1}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2(n-1).$$