1 Pirmoji paskaita. TYRIMO DUOMENŲ TVARKYMAS. ĮVAIRIŲ ŪKIO RODIKLIŲ SKAITINIŲ CHARAKTERISTIKŲ SKAIČIAVIMAS IR ĮVERTINIMAS

Šioje paskaitoje nagrinėjami klausimai:

1. Duomenų surinkimas ir jų tvarkymas.

2. Duomenų analizė ir grafinis pateikimas.

3. Skaitinės duomenų charakteristikos.

4. Duomenų standartizavimas.

5. Prognozės.

1.1 Tyrimo duomenų tvarkymas, analizavimas, grafinis pateikimas

Statistinio tyrimo metu sukaupti duomenys niekaip neapibūdina visumos. Norint ką nors pasakyti apie visumą, reikia gauti apibendrintas duomenų charakteristikas, atskleisti reiškinių esmę, tarpusavio ryšius, susisteminti duomenis.

Paprasčiausias būdas pateikti surinktus duomenis yra jų grupavimas pagal svarbiausius požymius. Tarkime, kad imtyje, sudarytoje iš n elementų, stebimas tam tikras kintamasis. Tegul kintamojo reikšmės yra: x_1, x_2, ..., x_n. Ši eilutė vadinama statistine eilute.

Apibrėžimas 1.1. Imties kintamojo reikšmių išdėstymas nemažėjančia tvarka vadinamas variacine eilute. T. y. $x_1^* \leqslant x_2^* \leqslant ... \leqslant x_n^*$, čia $x_{min} = x_1^*$ – mažiausias elementas, o $x_{max} = x_n^*$ – didžiausias imties elementas.

Apibrėžimas 1.2. Skirtumą $x_{max} - x_{min} = r$ vadiname imties pločiu.

Statistinėje eilutėje $x_1, x_2, ..., x_n$ kintamojo reikšmės gali kartotis. Galime suskaičiuoti, kiek kartų kiekviena reikšmė sutinkama statistinėje eilutėje. Tegul turime s skirtingų reikšmių: $x_1, x_2, ..., x_s$.

Apibrėžimas 1.3. Kintamojo reikšmės x_i dažnis n_i parodo, kiek kartų reikšmė x_i pasikartojo statistinėje eilutėje.

Apibrėžimas 1.4. Kintamojo reikšmės x_i santykinis dažnis n_i/n parodo, kurią statistinės eilutės dalį sudaro reikšmės x_i.

Praktikoje skaičiuojami kiekybinių ir kokybinių kintamųjų dažniai, santykiniai dažniai, sukauptieji dažniai ir santykiniai sukauptieji dažniai.

Apibrėžimas 1.5. Dažnių lentele vadiname lentelę, kurioje pateikiamos visos skirtingos kintamojo reikšmės (x_i) ir jų dažniai (n_i), santykiniai dažniai (n_i/n), sukauptieji dažniai ($n_1 + n_2 + ... + n_i$) arba santykiniai sukauptieji dažniai ($n_1/n + n_2/n + ... + n_i/n$).

Dažnių lentelę galima grafiškai iliustruoti dažnių daugiakampiu arba histograma.

Apibrėžimas 1.6. Jei grupavimas atliekamas pagal laiko kintamąjį, tai gautos grupuotos eilutės vadinamos dinamikos eilutėmis.

Apibrėžimas 1.7. Jei grupuojame tik pagal vieną požymį, tai toks grupavimas vadinamas paprastuoju, o jei turime 2 ir daugiau požymių, tai sakome, kad grupavimas yra kombinuotasis.

Dažnai turime skirtingai sugrupuotus duomenis ir norėdami palyginti sukauptą informaciją turime juos pergrupuoti. Dalis stebėjimo požymių yra faktoriniai, o kita dalis – rezultatiniai. Faktorniai požymiai yra tokie, kurie įtakoja nagrinėjamo proceso ar reiškinio kaitą, t. y. jie yra faktoriai įtakojantys procesą. Rezultatiniai požymiai kinta priklausomai nuo faktorių.

Kaip sudaryti grupes? Sudarant grupes reikia atsižvelgti į šiuos reikalavimus:

1) nustatomi grupavimo požymiai,

2) apibrėžiamas grupių skaičus (turėtų būti ne mažesnis už 4 ir ne didesnis už 20), nustatomi grupavimo intervalai,

3) atskiros kintamojo reikšmės turi patekti tik į vieną grupę.

Jeigu duomenų aibė yra gana simetriška, tai grupių skaičių patariama rinkti pagal tokią taisyklę:

$$k = 1 + 3,322 \lg n$$

čia $\lg n$ – skaičiaus n dešimtainis logaritmas. Galime imti ir

$$k \approx \sqrt{n}.$$

Paprastai statistiniai rezultatai pateikiami sąsajų lentelėmis. Tai – statistinės lentelės. Jose atsispindi, kokį reiškinį ar procesą nagrinėjame ir tą reiškinį ar procesą apibūdinantys rodikliai. Visos lentelės gali būti skirstomos į pasiskirstymo ir dinamikos lenteles. Pasiskirstymo lentelėse matome ekonominių reiškinių kitimą erdvėje, o dinamikos lentelėse – kitimą laike. Statistinės lentelės turi būti aiškios ir lengvai suprantamos. Jos turi turėti pavadinimą ir paantraštę (jei ji reikalinga – trumpa informacija apie duomenis, duomenų matavimo vienetus ir pan.), pagrindinę informaciją, išnašas (jei reikia) ir nuorodas į duomenų šaltinius, jei informacija surinkta ne autoriaus. Pavadinimai turi būti pilni ir aiškūs, nurodyti matavimo vienetai ir pan. Tuo atveju, kai duomenų nėra rašomas $-$, kai informacija yra, bet nepateikta rašoma ..., kai skiltis nepildoma – $\times$.

Greta statistinų lentelių pateikiami statistiniai grafikai. Grafikas yra vaizdinė priemonė. Grafikus galima pasitelkti ne tik pradiniams duomenims, bet ir analizės rezultatams pateikti. Grafikai gali būti pateikiami kartu su lentelėmis arba be jų. Aišku, kad grafikas bus informatyvus, jei jame atsispindės ryšiai, glūdintys pradiniuose duomenyse. Grafiko tipas turi būti parinktas taip, kad analizuojant grafiką netektų jo papildomai komentuoti ar grįžti prie surinktų duomenų. Visi grafiko elementai turi būti aiškūs ir gerai matomi. Be to, grafikas turi būti sudarytas taip, kad ir nespalvotas jo vaizdas išliktų informatyvus. Visi grafikai turi būti numeruojami. Pagal vaizdavimo formą galima išskirti tokias pagrindines grafikų (diagramų) rūšis:

1) linijinės,

2) stulpelinės,

3) skritulinės,

4) sklaidos.

Iš grafikų galima pastebėti tam tikras skirstinio savybes. Svarbios skirstinio formos charakteristikos yra asimetrijos koeficientas ir ekscesas.

Apibrėžimas 1.8. Skirstinio simetriškumas vertinamas akaičiuojant asimetrijos koeficientą

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma^3}.$$

Jei asimetrijos koeficientas teigiamas, tai turime teigiamą asimetriją. Jei asimetrijos koeficientas neigiamas, tai asimetrija neigiama. Kai asimetrijos koeficientas lygus 0, tai grafikas yra simetriškas.

Apibrėžimas 1.9. Smailiaviršūniškumui vertinti skaičiuojamas ekscesas

$$E = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma^4} - 3.$$

Jei eksceso reikšmė teigiama, tai grafikas yra smailas; jei – neigiama, tai grafikas yra lėkštas, o kai ekscesas lygus 0, tai turime standartinio normaliojo skirstinio atvejį.

1.2 Skaitinės duomenų charakteristikos

Norėdami palyginti statistinius duomenis ar atlikti matematinius skaičiavimus, turime turėti tam tikras skaitines reikšmes. Nagrinėsime dviejų tipų skaitines charakteristikas: padėties ir sklaidos.

Pagrindinės duomenų padėties charakteristikos: vidurkis, moda, mediana, kvartiliai. Visos šios charakteristikos gali būti skaičiuojamos kiekybiniams kintamiesiems, o kokybiniams kintamiesiems galime nustatyti tik modą.

Apibrėžimas 1.10. Vidurkis (Mean) – tai taškas, vidutiniškai artimiausias visiems statistinės eilutės elementams. Vidurkis apibrėžiamas formule:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

čia n – imties didumas, x_1, x_2, ..., x_n – statistinės eilutės elementai.

Vidurkis yra viena iš svarbiausių ir dažniausiai naudojamų padėties charakteristikų, tačiau kartais jis netiksliai atspindi vidutinę duomenų reikšmę. Tokiais atvejais vidurkis nėra informatyvus ir tenka naudoti kitas skaitines charakteristikas.

Apibrėžimas 1.11. Moda (Mode) – tai dažniausiai duomenų aibėje pasikartojanti reikšmė. Ji žymima Mo.

Jei visos reikšmės statistinėje eilutėje pasikartoja vienodai dažnai, tai sakome, kad statistinė eilutė modos neturi. Jei kelių gretimų variacinės eilutės reikšmių dažnis yra vienodas ir didesnis nei likusių statistinės eilutės reikšmių dažniai, tai moda yra šių reikšmių vidurkis. Jei dvi negretimos variacinės eilutės reikšmės turi tą patį dažnį, tai turėsime bimodalinį skirstinį. Jei negretimų variacinės eilutės reikšmių, turinčių tą patį dažnį, yra daugiau nei dvi, tai skirstinys yra multimodalinis. Moda gali būti skaičiuojama tiek kiekybiniams, tiek kokybiniams kintamiesiems.

Apibrėžimas 1.12. Mediana (Median) – vidurinioji variacinės eilutės reikšmė. Jei stebėjimų skaičius n yra nelyginis, tai mediana yra variacinės eilutės $\frac{n+1}{2}$ – asis narys. Jei stebėjimų skaičius n yra lyginis, tai mediana yra variacinės eilutės $\frac{n}{2}$ – ojo ir $\frac{n}{2}+1$ – ojo narių aritmetinis vidurkis.

Mediana dažniausiai naudojama rangų, intervalų bei santykių skalės duomenims. Ją patariama naudoti tuomet, kai duomenų aibėje yra išskirčių – reikšmių, daug kartų didesnių (mažesnių) už kitas reikšmes. Šiuo atveju modos ir medianos reikšmės sutampa ir yra informatyvesnės nei vidurkio reikšmė.

Apibrėžimas 1.13. Skaičiai, dalijantys variacinę eilutę į keturias maždaug lygias dalis, vadinami kvartiliais (Quartiles). Jie žymimi Q_1, Q_2, Q_3.

Apibrėžimas 1.14. Skaičiai, suskirstantys variacinę eilutę į 100 maždaug vienodų dalių, vadinami procentiliais (Percentile).

Pagrindinės sklaidos charakteristikos yra duomenų aibės plotis, kvartilių skirtumas, vidutinis nuokrypis, dispersija, standartinis nuokrypis, imties kitimo ir osciliacijos koeficientai. Šios charakteristikos gali būti taikomos tik kiekybiniams kintamiesiems.

Apibrėžimas 1.15. Duomenų aibės plotis (Range) yra lygus maksimalios ir minimalios variacinės eilutės reikšmių skirtumui.

Duomenų aibės plotis labai jautrus išskirtims. Jei duomenų aibėje yra išskirčių, vertėtų vietoj duomenų aibės pločio skaičiuoti kvartilių skirtumą.

Apibrėžimas 1.16. Kvartilių skirtumas – skirtumas tarp trečiojo ir pirmojo kvartilių, t. y. $Q_3 - Q_1$.

Apibrėžimas 1.17. Vidutinis nuokrypis žymimas raide d ir skaičiuojamas pagal šią formulę:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

Vidutinis nuokrypis nėra toks patogus naudoti kaip dispersija ar standartinis nuokrypis.

Apibrėžimas 1.18. Imties dispersija (Variance) parodo duomenų sklaidą apie vidurkį ir randama pagal formulę:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Dispersija – viena iš populiariausių duomenų sklaidos charakteristikų. Akivaizdu, kad dispersija rodo vidutinę kvadratinę reikšmių sklaidą. Todėl dažniausiai vietoje dispersijos skaičiuojamas imties standartinis nuokrypis, kurį lengviau interpretuoti ir lyginti su duomenimis.

Apibrėžimas 1.19. Imties standartinis nuokrypis (Std. deviation)

$$s = \sqrt{s^2}.$$

Lyginant skirtingų duomenų aibių sklaidas praverčia imties kitimo koeficientas.

Apibrėžimas 1.20. Imties kitimo (variacijos) koeficientas

$$cv = \frac{s}{\bar{x}}.$$

Jis gali būti išreiškiamas ir procentais.
Ar imtis vienarūšė parodo osciliacijos koeficientas.

Apibrėžimas 1.21. Osciliacijos koeficientas

$$v_r = \frac{r}{\bar{x}} \cdot 100\%.$$

Apibendrindami turime pastebėti, kad išvardintos padėties ir sklaidos charakteristikos gali būti skaičiuojamos ir grupuotiems duomenims. Suprantama, kad grupuotos imties skaitinės charakteristikos skirsis nuo negrupuotos imties charakteristikų.

Duomenų padėties ir sklaidos charakteristikos gali būti pateikiamos grafiškai naudojant dėžutės pavidalo ir paklaidų diagramas.

1.3 Duomenų standartizavimas

Normalusis skirstinys yra vienas plačiausiai naudojamų tolydžiųjų skirstinių.

Apibrėžimas 1.22. Sakome, kad atsitiktinis dydis turi normalųjį skirstinį, jei jo tankio funkcija išreiškiama taip:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

čia μ ir σ^2 yra skirstinio parametrai – vidurkis ir dispersija.

Normaliojo skirstinio parametrai lemia tankio funkcijos formą. Kuo didesnė dispersija, tuo lėkštesnė tankio funkcija. Tankio funkcija maksimalią reikšmę įgyja taške $x = \mu$, o maksimali funkcijos reikšmė lygi $\frac{1}{\sigma\sqrt{2\pi}}$. Tiek Puasono, tiek binominis skirstinys, esant dideliam bandymų skaičiui yra aproksimuojami normaliuoju skirstiniu. Kita vertus, normaliojo skirstinio svarba paaiškėja iš centrinės ribinės teoremos. Pateiksime atskirą jos atvejį:

1.1 Teorema. *Jei nepriklausomieji atsitiktiniai dydžiai X_i, $i \geqslant 1$ yra vienodai pasiskirstę su vidurkiais $MX_i = m$ ir baigtinėmis dispersijomis $DX_i = \sigma^2 \neq 0$, tai šių atsitiktinių dydžių sumos $\frac{S_i - nm}{\sigma\sqrt{n}}$ pasiskirstymo funkcija konverguoja į normaliojo skirstinio pasiskirstymo funkciją, t. y. $P(\frac{S_i - nm}{\sigma\sqrt{n}} < x) \to \Phi(x)$, kai $n \to \infty$. Čia $S_i = X_1 + X_2 + ... + X_n$.*

Kadangi atliekant statistinę duomenų analizę svarbi ne tik pati stebimoji reikšmė, bet ir jos padėtis duomenų aibėje. Todėl paprastai reikšmės lyginamos su vidurkiu. Tačiau čia tenka įvertinti ir duomeų sklaidos įtaką. O jei tenka lyginti kelias duomenų aibes, tai labai praverčia duomenų standartizavimas. Labiausiai paplitęs standartizavimo metodas susijęs su z reikšmių skaičiavimu:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Stnadartizuotos reikšmės turi vidurkį, lygų 0 ir dispersiją lygią 1. Teigiamoji z reikšmė nurodo, kad nagrinėjamoji reikšmė yra didesnė už vidurkį, o neigiamoji – kad mažesnė. Normaliajam skirstiniui galioja trijų sigma taisyklė, t. y. beveik visos atsitiktinio dydžio reikšmės yra susitelkusios intervale $[\mu - 3\sigma, \mu + 3\sigma]$. Intervale $[\mu - 2\sigma, \mu + 2\sigma]$ yra apie 96% reikšmių, o intervale $[\mu - \sigma, \mu + \sigma]$ – apie 68%.

1.4 Prognozės

Paprasčiausias būdas prognozuoti, tai įvertinti jau turimus ryšius tarp duomenų ir naudojantis sukaupta informacija numatyti mus dominančio dydžio kitimo tendencijas. Šiuo atveju labai svarbu nustatyti ir aiškiai apibrėžti priežastis ir padarinius, t. y. reikia nustatyti priklausomuosius ir nepriklausomuosius kintamuosius. Nepriklausomieji kintamieji yra priežastis, o pasekmė yra priklausomieji kintamieji. Tačiau kartais mes galime nustyti priklausomybes tarp kintamųjų nesusietų priežastiniu ryšiu. Ryšiai tarp kintamųjų nustatomi remiantis koreliacine ir regresine analize.

Koreliacinė analizė padeda įvertinti statistinio ryšio stiprumą ir nustatyti svarbiausius faktorius. Regresinė analizė įvertina statistinio ryšio formą.

Tarkime, kad stebime intervalinių kintamųjų porą, gautą matuojant dvimatį normalųjį atsitiktinį dydį. Norime nustatyti, ar kintamieji koreliuoja. Ryšius tarp kintamųjų galime nustatyti detaliau panagrinėję sklaidos diagramas. Suprantama, jei tik pagal grafiką imsime tvirtinti, kad kintamuosius sieja tam tikra priklausomybė, neišvengsime klaidų. Koreliacijos koeficientas apibūdina kintamųjų ryšio stiprumą. Jo įgyjamos reikšmės priklauso intervalui $[-1, 1]$. Jei koreliacijos koeficientas lygus 1, tai sakoma, kad tarp tiriamų kintamųjų yra stiprus tiesinis ryšys. Jei koreliacijos koeficientas lygus -1, tai sakoma, kad tarp kintamųjų yra stiprus atvirkštinis ryšys. Jei koreliacijos koeficientas lygus 0, tai sakoma, kad stebimi dydžiai yra tiesiškai nepriklausomi. Literatūroje galima surasti įvairių koreliacijos koeficiento reikšmių interpretacijų. Remsimės viena iš jų: jei koreliacijos koeficiento įgyjama reikšmė yra intervale nuo -0,2 iki 0,2, tai sakome, kad kintamuosius sieja labai silpna koreliacija,
nuo 0,2 iki 0,5 (nuo -0,2 iki -0,5) – silpna teigiama (neigiama) koreliacija,
nuo 0,5 iki 0,7 (nuo -0,5 iki -0,7) – vidutinio stiprumo teigiama (neigiama) koreliacija,
nuo 0,7 iki 0,9 (nuo -0,7 iki -0,9) – stipri teigiama (neigiama) koreliacija,
nuo 0,9 iki 1 (nuo -0,9 iki -1) – labai stipri teigiama (neigiama) koreliacija.

Koks koreliacijos koeficientas gali būti skaičiuojamas dirbant su vienais ar kitais duomenimis, lemia kintamųjų matavimo skalė:

- intervaliniams kintamiesiems skaičiuojamas Pirsono koreliacijos koeficientas,
- jei bent vienas iš kintamųjų nėra pasiskirstęs pagal normalųjį dėsnį, tai skaičiuojamas Spirmeno ranginės koreliacijos koeficientas,
- jei bent vienas kintamasis priklauso rangų skalei, taip pat skaičiuojamas Spirmeno ranginės koreliacijos koeficientas.

Koreliacinė analizė labai glaudžiai susijusi su regresine analize. Tačiau čia vertėtų prisiminti, kad koreliacijos koeficientas yra tiesinio ryšio matas. Taigi tais atvejais, kai koreliacijos koeficiento reikšmės yra mažos, mes galime tvirtinti, kad kintamuosius sieja silpnas tiesinis ryšys, bet tarp šių kintamųjų gali būti labai stiprus netiesinis ryšys. Regresinėje analizėje priklausomasis kintamasis yra tas, kuriuo mes domimės, kurio reikšmes norime prognozuoti, o nepriklausomasis – kuriuo bandome aiškinti priklausomojo kintamojo kitimą. Pažvelgę į duomenų sklaidos diagramą joje galime pastebėti galimo ryšio tarp kintamųjų pobūdį. Tik tuo atveju, jei sklaidos diagramos taškai telkiasi apie tiesę, galime taikyti paprastąją tiesinę regresiją, kurios pagrindą sudaro mažiausiųjų kvadratų metodas. Remiantis šiuo metodu sudaroma tiesinė priklausomybė išreiškiama lygtimi:

$$y = \alpha x + \beta + \varepsilon,$$

čia α, β – nežinomi parametrai, o ε – paklaida. Vienas iš regresinės kreivės tinkamumo rodiklių yra determinacijos koeficientas. Jis parodo regresinio nuokrypio kvadratų sumos ir bendro nuokrypio nuo vidurkio kvadratų sumos santykį. Kuo determinacijos koeficiento reikšmė artimesnė 1, tuo geriau tinka duomenims sudarytoji regresinė kreivė. Paprastosios tiesinės regresijos apibendrinimas yra daugialypė tiesinė regresija. Šiuo atveju vieno (priklausomojo) kintamojo reikšmės nusakomos nagrinėjant kelis nepriklausomuosius kintamuosius. Čia tenka atsakyti į klausimą, kurie nepriklausomieji kintamieji yra tikrai svarbūs, kurie turi didžiausią įtaką priklausomąjam kintamąjam. Kaip ir paprastosios tiesinės regresijos atveju, labai svarbus reikalavimas, kad sudarius prognozę paklaidos būtų kaip galima artimesnės 0. Bendroji daugialypės regresijos lygtis:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + \beta + \varepsilon,$$

čia α_1, ..., α_k, β – nežinomi parametrai, o ε – paklaida. Lygties išraiškoje esantis koeficientas α_i parodo kiek pakinta priklausomojo kintamojo reikšmė vienu vienetu padidėjus nepriklausomojo kintamojo x_i reikšmei, kai kitų nepriklausomųjų kintamųjų reikšmės yra fiksuotos. Norėdami įvertinti kiekvieno iš nepriklausomųjų kintamųjų svarbą vertiname gautų regresinės tiesės koeficientų reikšmes. Kuo didesnis determinacijos koeficientas, tuo daugiau informacijos apie priklausomąjį kintamąjį suteikia visi nepriklausomieji kintamieji. Tuo atveju, kai nepriklausomųjų kintamųjų daug, tai determinacijos koeficientas yra artimas 1 ir juo vertėtų abejoti. Šiuo atveju labiau praverčia koreguotasis determinacijos koeficientas, kuris skaičiuojamas atsižvelgiant ir į imties didumą ir į nepriklausomųjų kintamųjų skaičių:

$$r_k^2 = 1 - \frac{n-1}{n-k-1}(1 - r^2),$$

čia n – imties didumas, k – nepriklausomųjų kintamųjų skaičius, r^2 – determinacijos koeficientas. Kvadratinė šaknis iš determinacijos koeficiento yra vadinama daugialypės koreliacijos koeficientu ir parodo, kaip stipriai priklausomasis kintamasis priklauso nuo visų nepriklausomųjų kintamųjų. Tai jokiu būdu nereiškia, kad visi kintamieji svarbūs. Dalinė koreliacija leidžia nustatyti priklausomojo kintamojo priklausomybę nuo vieno iš nepriklausomųjų kintamųjų eliminavus kitų kintamųjų įtaką. Taip pat reikėtų atkreipti dėmesį į tai, ar dalinės koreliacijos nėra mažesnės už empyrines. Tokiu atveju nepriklausomieji kintamieji iš tikrųjų yra susiję, t. y. koreliuoja tarpusavyje. Tuomet tenka atsisakyti dalies kintamųjų arba didinti imtį.

Nepriklausomųjų kintamųjų svarbą lengviausia įvertinti nagrinėjant ir analizuojant standartizuotąsias regresinės funkcijos koeficientų rekšmes. Absoliutiniu didumu didesnė standartizuotoji reikšmė nurodo didesnę įtaką turintį kintamąjį.