

Teorija ir praktika

## Statistikos taikymas mokslinių tyrimų analizėje

V. Kasiulevičius, G. Denapienė

Vilniaus universitetas, Medicinos fakultetas

### Aprašomoji statistika

Aprašomoji statistika nagrinėja kintamųjų grupavimo požymius, grupavimo intervalus, grupavimų rūšis, įvertina duomenų variaciją ir koncentraciją, pavaizduoja statistikos duomenis grafikais ir lentelėmis. Aprašomosios statistikos objektas yra vidurkia, moda, mediana, standartinis nuokrypis, variacijos koeficientas. **Mediana** (angl. *median*) – požymio reikšmė, kuri variacinę eilutę dalija į dvi lygias dalis. **Moda** (angl. *mode*) – tai dažniausiai pasikartojanti požymio reikšmė variacinėje eilutėje. **Vidurkis** (angl. *mean, average*) – tai vidutinė požymio reikšmė, nustatyta tiriant skirtingus objektus. Jis apskaičiuojamas sudedant reikšmes ir sumą padalijant iš tų reikšmių skaičiaus. **Standartinis nuokrypis** (SD, *standard deviation*), vidutinis kvadratinis nuokrypis – tai dydis, rodantis, kiek kiekviena reikšmė yra vidutiniškai nukrypusi nuo vidurkio. Tai tiriamojo požymio reikšmių sklaidos apibūdinimas, apibrėžiamas kaip požymio įgyjamų reikšmių ir vidurkio skirtumų kvadratų sumos vidurkis. Šis dydis yra žymimas įvairiai:  $\sigma$ ,  $s$ , SD. **Variacijos koeficientas** – standartinio nuokrypio santykis su vidurkiu. Imtis, kurios kintamieji duomenų apdorojimo

programoje išdėstyti didėjimo arba mažėjimo tvarka, vadinama **variacione eilute**. Kai duomenų daug, sudaromos vienodų ar artimų reikšmių grupės bei surašomi variantų pasikartojimo dažniai. Taip sudaroma **intervalinė (pasiskirstymo) variacinė eilutė** [1, 2].

1 lentelė. Sistolinio kraujospūdžio intervalinė variacinė eilutė tyrime

Sistolinio AKS dydis	Dažnis
Iki 120 mmHg	0
Nuo 120 iki 130 mmHg	34
Nuo 130 iki 140 mmHg	43
Nuo 140 iki 150 mmHg	161
Nuo 150 iki 160 mmHg	123

**Diagramos** – vaizdus duomenų pateikimo būdas. Stulpelinės diagramos būna įvairių rūšių. Dažnai naudojama dažnių histograma, nuokrypių nuo vidurkių stulpelinė diagrama. Sudėtis (%) dažniausiai vaizduojama skrituline diagrama [1, 2].

### Analitinė statistika

Analitinė statistika nagrinėja taikymą statistinių kriterijų, kuriais remiantis priimamos arba atmestos hipotezės pasirinktu reikšmingumo lygmeniu. **Hipotezė** – teiginys apie kokį nors reiškinį, kurio teisingumas iš anksto

Adresas: V. Kasiulevičius  
Santariškių g. 2, Vilnius  
Tel. (8-682) 21009  
El. paštas: vytautas.kasiulevicius@santa.lt

nežinomas. Statistikoje svarbios dvi hipotezių rūšys – nulinė ir alternatyvioji. **Nulinė hipotezė** (angl. null hypothesis) teigia, kad lyginamų imčių skirstiniai nesiskiria. Ji žymima  $H_0$ . **Alternatyvioji hipotezė** (angl. alternative hypothesis) teigia priešingai – lyginamų imčių skirstiniai skiriasi. Ji žymima  $H_1$ . Priimant sprendimą dėl hipotezės galimos dviejų rūšių klaidos (2 lentelė).

2 lentelė. Hipotezių priėmimo klaidos

	$H_0$ teisinga	$H_0$ neteisinga
Paneigti $H_0$	I rūšies klaida: Paneigti $H_0$ , kai ji yra teisinga	teisingas sprendimas
Priimti $H_0$	teisingas sprendimas	II rūšies klaida: Priimti $H_0$ , kai ji yra neteisinga

**Hipotezės tikrinimas** (angl. hypothesis testing) yra procedūra, kai imties (ar imčių) duomenys tikrinami naudojant statistinius kriterijus. Tuo tarpu daug statistinių kriterijų yra skirta nustatyti, ar skiriasi dviejų arba daugiau populiacijų požymiai – vidurkiai arba medianos.

**Statistinių kriterijų taikymo principai**

Prieš taikant apibrėžtą statistikos testą, dažniausiai reikia atsakyti į tris pagrindinius klausimus: 1) Kokiai matavimo skalei priklauso tiriamas kintamasis? 2) Ar kintamojo rezultatai intervalų skalėje pasiskirstę pagal normalųjį skirstinį? 3) Ar lyginamos imtys yra priklausomos ar nepriklausomos?

**Kokiai matavimo skalei priklauso tiriamas kintamasis?**

Atlikę tyrimą, gautus duomenis arba kintamuosius suvedame į statistinę duomenų apdorojimo programą (EXCEL, Statistica, SPSS, EpiInfo ar kt.). Tai atlikdami turime pasirinkti matavimo skalę. Yra šios tyrimo duomenų matavimo skalės:

1) Vardinė, arba nominalinė skalė. Būdingiausi vardinės skalės pavyzdžiai yra lytis, medikamentų sąrašas, medicinos specialybių sąrašas ir t. t. Kintamasis, įgyjantis tik dvi reikšmes (kategorijas), dar vadinamas binariniu. Skaičiai, kuriais koduojami atskiri objektai ar jų savybės, neturi jokios empirinės reikšmės, tik rodo, kokia čia ypatybė ar objektas. Kintamųjų, priklausančių vardinėi skalei (nominaliųjų kintamųjų), apdorojimo galimybės gana ribotos – galima tikrai įvertinti, kurių objektų (savybių) yra daugiau ar mažiau, koks bendras visų sąrašė esančių objektų kiekis. Pagal nominaliuosius kintamuosius

dažnai vykdoma kokybinė duomenų klasifikacija, arba grupavimas – imtis suskaidoma pagal šių kintamųjų kategorijas. Gautoms dalinėms imtims taikomi vienodi statistikos testai, jų rezultatai palyginami tarpusavyje.

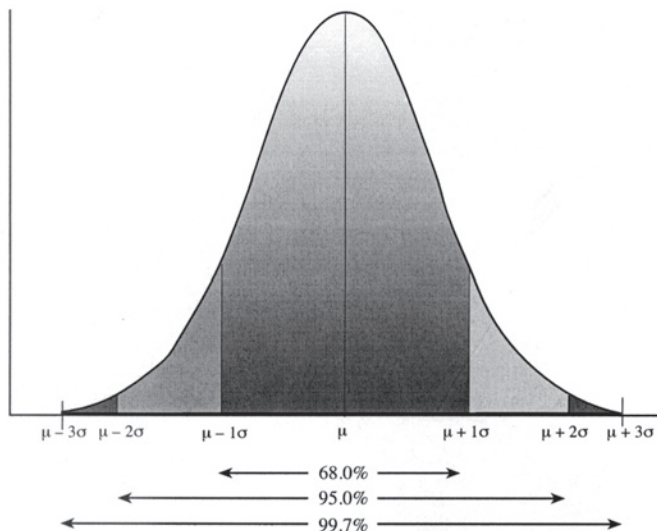
2) Rangų skalė. Joje nustatoma objekto (reiškinių) vieta pagal pasirinktą kiekybinį arba kokybinį požymį vienos rūšies objektų (reiškinių) grupėje. Pavyzdžiui, hipertenzijos laipsnis, ligos stadija, pacientų fizinis aktyvumas (pvz.: 1 = jokio fizinio aktyvumo, 2 = fiziškai aktyvūs retkarčiais, 3 = pakankamas fizinis aktyvumas) ir t. t. Su matuojamais ranginiais kintamaisiais galima atlikti daugiau statistinių operacijų negu su nominaliaisiais kintamaisiais. Be dažnių įvertinimo, galima apskaičiuoti medianą, rangų koreliacijos koeficientą, palyginti atskiras imtis naudojant neparimetrinius testus.

3) Intervalų skalė. Šioje skalėje nurodomi kiekybiniai kintamųjų reikšmių skirtumai, išreikšti matavimo vienetais (milimetrais, sekundėmis, laipsniais ir pan.). Šie skirtumai gali būti tarp atskirų intervalų arba nuo kurio nors pasirinkto atskaitos taško, t. y. nulinė reikšmė dar nereiškia, kad tiriamasis požymis visai nepasireiškia, o tiktai, kad jis nesiskiria nuo sąlyginio atskaitos nulio. Duomenis intervalų skalėje galima apdoroti visais be apribojimų statistikos metodais.

4) Santykių skalė. Ši skalė skiriasi nuo intervalų skalės tik tuo, kad joje nulinis taškas yra griežtai apibrėžtas ir visiškai atitinka dydžio nebuvimą.

**Ar kintamojo rezultatai intervalų skalėje pasiskirstę pagal normalųjį skirstinį?**

Skirstinys (tikimybinis pasiskirstymas ar pasiskirstymo dėsnis; angl. probability distribution) – tai požymio reikšmių, arba atsitiktinių dydžių, ir jų tikimybių tarpusavio ryšys. Normalusis skirstinys (angl. normal distribution) – tai tolydžiųjų požymių reikšmių skirstinys (pasiskirstymo dėsnis), atitinkantis tokias sąlygas: vidurkio ( $\mu$ ), modos ir medianos reikšmės sutampa, skirstinio kreivė yra simetriška, o simetrijos ašis yra ties vidurkiu, skirstinio kreivės forma priklauso nuo vidurkio ir standartinio nuokrypio ( $\sigma$ ), normalųjį skirstinį turinčių atsitiktinių dydžių suma taip pat turi normalųjį skirstinį. Normalusis skirstinys siejamas su vokiečių matematiko Karlo Friedricho Gauso (vok. Gauss) (1777–1855) vardu ir vadinamas Gauso pasiskirstymu arba Gauso skirstiniu (angl. Gaussian distribution). Normaliajam skirstiniui taikoma trijų sigmų taisyklė: 1) patekimo į intervalą  $\mu - \sigma$  ir  $\mu + \sigma$  tikimybė yra 68%, 2) patekimo į intervalą



1 pav. Normalusis skirstinys

$\mu - 2\sigma$  ir  $\mu + 2\sigma$  tikimybė yra 95%, 3) patekimo į intervalą  $\mu - 3\sigma$  ir  $\mu + 3\sigma$  tikimybė yra 99%.

Kaip matyti 1 paveiksle, beveik visas plotas po normaliąja kreive yra trijų kvadratinų nuokrypių nuo centro ribose. Taigi, jei kintamojo skirstinys normalus, tai praktiškai visos kintamojo reikšmės yra ne daugiau kaip  $3\sigma$  atstumu nutolusios nuo centro.

Kolmogorovo–Smirnovu testu galima patikrinti, ar realus skirstinys atitinka normalųjį skirstinį, kadangi nuo gautų rezultatų priklauso, kokie analizės metodai – parametriniai ar neparametriniai bus taikomi. Tyrime analizuojamas skirstinys nuo normaliojo skiriasi reikšmingai, jeigu gauta  $p$  – reikšmė mažesnė už nustatytą reikšmingumo lygmenį (paprastai – 0,05).

### Ar lyginamos imtys yra priklausomos, ar nepriklausomos?

Nepriklausomos imtys (angl. *independent samples*) – dvi ar daugiau imčių, kurių kiekvienos tiriama objektai niekaip nesusiję su kitų imčių tiriama objektais. Nepriklausomos imtys, t. y. imtys, kurioms negalima nustatyti dėsningo ir vienareikšmio atitikimo, jos gali turėti skirtingus stebėjimus, kuriuos paprastai skiria kategorinis vardinės skalės kintamasis.

Priklausomos imtys (angl. *dependent samples*) – dvi ar daugiau imčių, kurių kiekvienos tiriama objektai kaip nors susiję su kitų imčių tiriama objektais. Pvz., kelis kartus per metus atlikti tų pačių pacientų kraujospūdžio matavimai. Tokiu atveju priklausomos imtys sudaro tiriamo vyksmo parametrų reikšmes skirtingais laiko momentais.

### Statistinio kriterijaus pasirinkimas

Atsakę į šiuos tris klausimus galime nesunkiai pasirinkti statistinį kriterijų mūsų tyrimo rezultatams patikrinti.

3 lentelė. Statistinio kriterijaus pasirinkimas, kai yra normalusis kintamųjų pasiskirstymas

Imtys	Statistinis kriterijus
Dvi nepriklausomos imtys	Stjudento t-testas
Trys ir daugiau nepriklausomų imčių	Paprasta dispersinė analizė ANOVA
Dvi priklausomos imtys	Porinis studento t-testas
Trys ir daugiau priklausomų imčių	Blokuotųjų duomenų dispersinė analizė ANOVA

4 lentelė. Statistinio kriterijaus pasirinkimas, kai nėra normaliojo kintamųjų pasiskirstymo

Imtys	Statistinis kriterijus
Dvi nepriklausomos imtys	Waldo–Wolfowitzo runs testas Mano–Witnio U testas
Trys ir daugiau nepriklausomų imčių	Kruskalo ir Wallis H-testas Medianos testas
Dvi priklausomos imtys	Wilcoxon testas Ženklo testas
Trys ir daugiau priklausomų imčių	Friedmano testas

### Kas nulemia statistinio kriterijaus galią?

Tikimybė pagrįstai atmesti neteisingą  $H_0$  hipotezę vadinama testo galia. Testo galia yra priešinga II rūšies klaidos tikimybei. Mažinant I klaidos tikimybę  $\alpha$  didėja II klaidos tikimybė. Iš to seka, kad mažėja ir testo galia. Statistinis kriterijus yra tuo geresnis, kuo yra mažesnės abiejų rūšių klaidos. Paprastai kriterijai yra sudaromi taip, kad fiksuotai I rūšies klaidai II rūšies klaida būtų minimali. Todėl dažniausiai pasirenkama  $\alpha$  reikšmė tyrime yra 0,05. Paprastai yra skaičiuojama ne II rūšies klaidos tikimybė, o jai priešingo įvykio tikimybė  $1 - \beta$  – *kriterijaus galia*. Kriterijaus galia  $\beta$  – tai tikimybė atmesti hipotezę  $H_0$ , kai ji klaidinga. Kriterijaus galia leidžia palyginti du kriterijus, turinčius tą patį  $\alpha$  ir taikomas tokio pat dydžio imtims. Galingesnis kriterijus yra tas, kuriam  $1 - \beta$  yra didesnis. Didinant imtį kriterijaus galia paprastai didėja [3, 4].

### Koreliacija

Terminą **koreliacija** pirmas pavartojo Kiuvje biologijoje (1806). Tuo tarpu matematiškai koreliaciją pirmasis aprašė prancūzas O. Brave (1846). K. Galtonas 1886 m. koreliaciją panaudojo biometrijoje. Taigi **koreliacijos analizė** – statistikos metodas, tiriantis požymių tarpusavio ryšių stiprumą. Pasiskirsčiusiems pagal normalųjį

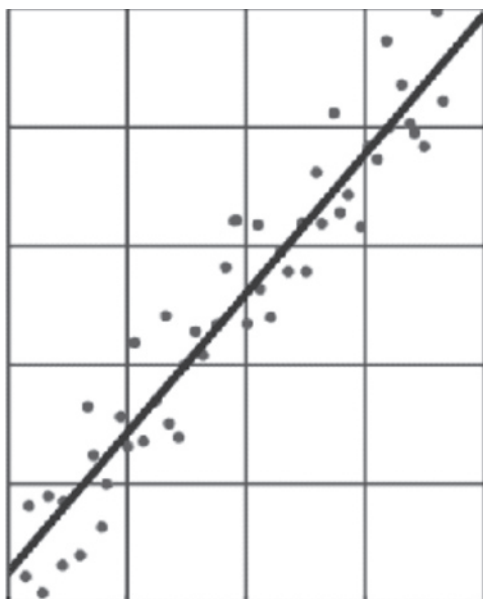
dėsnį intervaliniams kintamiesiems yra skaičiuojamas Pirsono (angl. *Pearson*) koreliacijos koeficientas. Intervaliniams kintamiesiems, kuriems normalumo prielaida nėra tenkinama, ir ranginiams kintamiesiems yra skaičiuojamas Spirmeno (angl. *Spearman*) arba Kendallo  $\tau$ -b koreliacijos koeficientas [5, 6].

5 lentelė. Koreliacijos koeficiento (r reikšmės) vertinimas

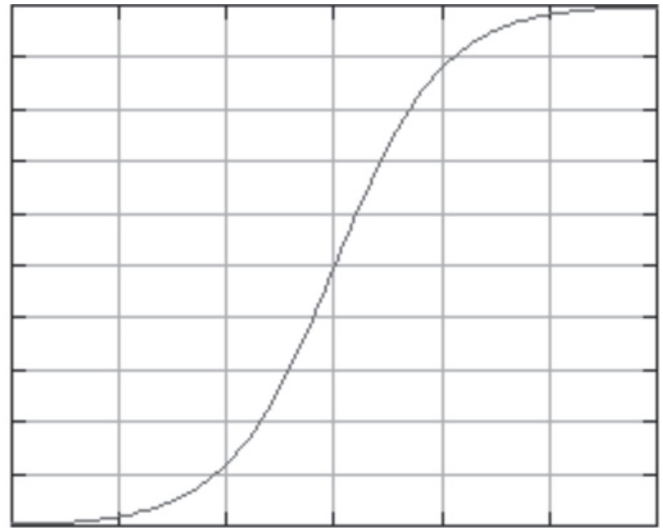
r reikšmė	Vertinimas
0,00–0,19	Labai silpnas tarpusavio ryšys
0,20–0,39	Silpnas ryšys
0,40–0,69	Vidutinis ryšys
0,70–0,89	Stiprus ryšys
0,90–1,00	Labai stiprus tarpusavio ryšys

**Regresinė analizė**

Regresinė analizė nustato statistinio ryšio pobūdį ir aprašo priklausomojo (pasekmės) kintamojo vidutinių reikšmių priklausomybę nuo vieno ar kelių nepriklausomųjų (priežasties) kintamųjų reikšmių matematine formule ir kartu prognozuoja šio kintamojo reikšmes. Regresinė analizė skirstoma į tiesinę ir logistinę regresiją. Tiesinė regresija (2 pav.) dalijama į paprastą tiesinę regresiją, kai egzistuoja vienas nepriklausomas kintamasis, ir daugelio faktorių tiesinę regresiją, kai egzistuoja keletas nepriklausomų kintamųjų. Savo ruožtu logistinė regresija (3 pav.) skirstoma į binarinę logistinę regresiją ir daugiareikšmę logistinę regresiją [5, 6].



2 pav. Grafinis tiesinės regresijos modelis



3 pav. Grafinis logistinės regresijos modelis

**Dispersinė analizė**

Dispersinė analizė – dviejų ar daugiau populiacijų vidurkių lygybės tikrinimas. Metodas pagrįstas tarpgruopinės ir grupių vidutinių dispersijų palyginimu, kur visos grupės lyginamos vienu metu. Dispersinei analizei naudojami kintamieji turi būti pasiskirstę pagal normalųjį dėsnį, atsitiktiniai ir nepriklausomi. Dispersinė analizė leidžia įvertinti tiriamojo faktoriaus įtaką esant mažam duomenų ar kartotinių bandymų kiekiui. Tai efektyvus metodo skirtumo tarp trijų ir daugiau grupių reikšmingumui patikrinti. Yra trys dispersinės analizės metodai: vieno faktoriaus, dviejų faktorių ir daugelio faktorių dispersinė analizė.

**Literatūra:**

1. Armitage P, Matthews JNS, Berry G. Statistical Methods in Medical Research. Blackwell Science: 2001.
2. Hogg R. Probability and Statistical Inference. 7th ed. Pearson: 2006.
3. Hill T, Lewicki P. Statistics methods and applications. StatSoft, Tulsa, OK. 2007.
4. Douglas G. Altman Practical Statistics for Medical Research. Chapman & Hall, CRC: 1999.
5. Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates. 2003.
6. Pampale FC. Logistic regression: a primer. Sage Publications: 2000.

*Straipsnis įteiktas redakcijai 2008 m. gegužės 21 d., parengtas spaudai 2008 m. birželio 28 d.*

**STATISTICS IN SCIENTIFIC RESEARCH ANALYSIS****V. Kasiulevičius, G. Denapienė**

Vilnius University, Faculty of Medicine

**Abstract**

The paper deals with descriptive and inferential statistics. Descriptive statistics (DS) is used to describe the basic features of data under study. DS provides simple summaries about a sample and the measures. DS helps researchers to simplify

large amounts of data in a sensible way. Each descriptive statistics reduces lots of data to a simpler summary. Inferential statistics (IS) helps to reach conclusions that extend beyond the immediate data alone. IS is used to infer from a sample data what the population might think. IS means the use of statistics to make inferences concerning some unknown aspect of the population from a sample. A common method used in inferential statistics is estimation. Examples of inferential statistics methods include hypothesis testing, correlations, linear and logistic regression, and dispersion analysis.