Bendrai finansuojama pagal
Europos Sąjungos programą
„Erasmus+“

Matematikos išlyginamasis kursas
4b skyrius – Statistika

TIKIMYBĖ IR STATISTIKA

Tikimybė ir statistika yra glaudžiai susijusios, kadangi visi statistiniai teiginiai yra pagrindinai teiginiai apie tikimybę. Nežiūrint į tai, kartais jie abu atrodo kaip labai skirtingi dalykai. Tikimybė yra logiškai savarankiška; yra keletas taisyklių, o atsakymai logiškai išplaukia iš taisyklių. Statistikoje taikome tikimybę, kad iš duomenų padarytume išvadas.

Tikimybės pavyzdys:
Įsivaizduokime dabar teisingą monetą, kuri išmesta turi vienodas galimybes atsiversti skaičiumi ar herbu. Dabar išmetame ją 100 kartų. Mūsų klausimas šiame pavyzdyje būtų toks: koks yra panašumas 60-ties ar daugiau skaičiumi į viršų atsivertusių monetų. Į šį klausimą yra tik vienas atsakymas ir mes išmoksime jį apskaičiuoti.

Statistikos pavyzdys:
Šiame pavyzdyje vėl turime monetą, tačiau mes nežinome ar ji yra teisinga ar sukonstruota, kad atsiverstų daugiau skaičiumi į viršų (ar daugiau herbu į viršų). Mes vėl išmetame monetą 100 kartų ir suskaičiuojame 60 skaičiumi į viršų. Dabar kyla klausimas, kokią išvadą mes galime padaryti iš šių duomenų. Yra daug būdų, kaip galime atsakyti į šį klausimą atsižvelgiant į išvados formą ir į tikimybių skaičiavimus naudojamus išvadai pagrįsti.

Ko mes išmokome iš pavyzdžių?
Pirmame pavyzdyje atsitiktinis procesas yra visiškai žinomas. Žinome, kad kiekvieno išmetimo metu yra 50% tikimybė gauti skaičiumi į viršų ar herbu į viršų. Ką mes siekiame išsiaiškinti yra tikimybė konkretaus scenarijaus (mažiausiai 60 atsivertusių skaičiumi į viršų), kuris atsirastų iš šio ***žinomo*** atsitiktinio proceso. Antrame pavyzdyje dirbame su konkrečiu scenarijumi (gavome 60 atsivertusių herbu į viršų) ir mūsų tikslas yra panaudoti tai apibūdinti **nežinomą** atsitiktinį procesą.

2. Statistika

Oksfordo besimokančiųjų žodynas apibrėžia **statistiką** kaip skaičiais rodomos informacijos rinkinį. Statistika taip pat yra statistikos rinkimo ir analizės mokslas, t.y. skaičiais rodomos informacijos rinkimas ir analizavimas. Kaip elementariausią statistikos pavyzdį, mes galime tiesiog atsiversti laikraštį ir eiti į sporto skyrių. Tarkime, dabar yra laikotarpis, kai žaidžiamas yra Vimbldonas ir mes sporto skiltyje randame tokią informaciją:

Final · Center Court

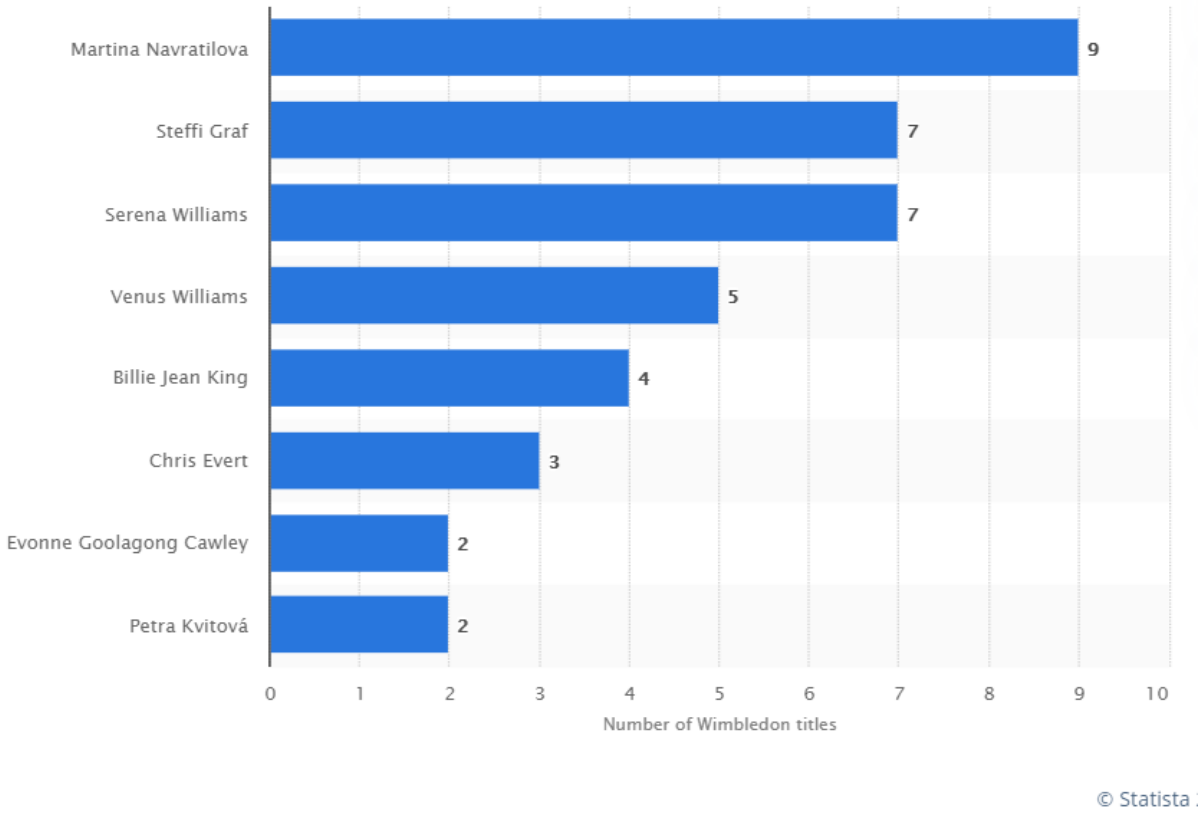
		Final	
11	S. Williams	2	2
7	S. Halep	6	6

Serena Williams		Simona Halep
2	Aces	1
1	Double faults	0
68%	First serve %	76%
59%	Win % on 1st serve	83%
50%	Win % on 2nd serve	45%
0/1	Break points	4/5
0	Tiebreaks won	0
12	Receiving points won	21
38	Points won	55
4	Games won	12
1	Max games won in a row	5
5	Max points won in a row	7

Matome 2019m. sužaistų finalinių rungtynių tarp Simonos Halep ir Selenos Williams statistiką. Kiekvienas numeris turi savo“istoriją” . Viršutiniame dešiniajame kampe matome mažą matricą su pirmuoju ir antruoju stulpeliu rodančiais 2:6. Kiekvienas stulpelis žymi atitinkamą sužaistą kėlinį, o kiekvienoje eilutėje nurodoma, kiek partijų kiekvienas iš žaidėjų laimėjo. Žemiau šešėlyje matome išsamesnę informaciją apie rungtynes. Šiuos skaičius galime interpretuoti kaip atsakymus į konkrečius klausimus apie rungtynes:

- Kiek neatmušamų padavimų Serena Williams atliko?
  - Atsakymas: 2
- Kokias dvigubas Simonos Halep klaidas galėjome pastebėti?
  - Atsakymas: 0
- Koks buvo Simonos sėkmingų padavimų procentas?
  - Atsakymas: 76%
- Kiek rungčių laimėjo Serena Williams?
  - Atsakymas: 4 (kuris atitinka 2 pirmame kėlinyje ir 2 antrame kėlinyje, kuriuos mes matėme matricos dešiniajame kampe)

Aukščiau pateiktame pavyzdyje mes iliustravome surinktus duomenis per vienas konkrečias rungtynes. Mums taip pat gali būti įdomu sužinoti atsakymus į skirtingus klausimus tokius kaip, kas laimėjo daugiausiai moterų finalinių varžybų titulų laikotarpyje nuo 1968 iki 2019. Kad atsakytume į šį klausimą, naudosime kitokio tipo iliustraciją:



Mūsų antrasis pavyzdys iliustruoja duomenų rinkimą per tam tikrą laiką. Mums įdomu sužinoti, kas laimėjo moterų finalines rungtynes 1968 -2019m., grupuojant metus pagal vardus ir rūšiuojant grupes pagal metų skaičių jose.

Tipiški žingsniai, kuriuos atliktų statistė savo darbe:

1. Duomenų rinkimas:
    - Moterų finalo laimėtojų 1968- 2019 rezultatai
2. Duomenų suskirstymas į lenteles, diagramas, grafikus
    - Grupavimas pagal žaidėjų vardus, nurodant kiek metų kiekvienas žaidėjas laimėjo
3. Išvadų darymas iš duomenų analizės
    - Viršuje pateikta diagrama rodo, kad daugiausiai laimėjimų pelnė Martina Navratilova

Kai šie trys žingsniai, kurie aprašo ir apibendrina duomenų rinkinį, įtraukiami į statistinį tyrimą, tyrimą vadinsime ***aprašomąja statistika***. Neapdoroti duomenys dažnai pateikiami kaip sąrašas, masyvas arba pavadinimų ir vardų duomenų bazė.

<u>Kaip mes renkame neapdorotus duomenis?</u>

Kaip matėme pavyzdžiuose, duomenų, su kuriais dirbame, pobūdis gali būti įvairių formų. Mes galime žiūrėti į duomenis kaip į rezultatą:

1. ***Klausimynas***
    - Klausimų sąrašas, kuriame asmuo gali pažymėti vieną iš kelių galimų atsakymų ar užpildyti atsakymą raštu
    - Pavyzdys: surašymas, rinkimų rezultatų skaičiavimas

2. **Registracijos žurnalas**
    - Ar dienoraštis, kuriame žmogus reguliariai surašo informaciją

- Pavyzdys: ligoninių diagrama (vaistai skiriami kas 8 val.), lauko temperatūra (7val. Iš ryto kiekvieną dieną), pergalės Vimbldone (moterų finalinių rungtynių nugalėtoja kartą per metus)

Surašymas yra geras pavyzdys, kaip mes surenkame informaciją iš stebimų gyventojų. Remdamiesi surašymo duomenimis, šalies vadovybė gali priimti tokius sprendimus kaip didesnės biudžeto dalies paskirstymas labiau apgyvendintoms vietovėms ir t.t

Tačiau daugeliu atvejų mes negalime pasiekti visos mūsų tyrimo auditorijos. Panagrinėkime reklamos įmonę, kuri nori sužinoti, kurios laidos yra labiausiai žiūrimos. Turėdama šią informaciją, įmonė gali įsigyti reklamos laiko per šių laidų pertraukas ir padidinti pasiekiamą auditoriją. Iš tikrųjų tai būtų labai brangu ir atimtų daug laiko imituoti surašymą ir bandyti pasiekti kiekvieną žiūrovą. Šiam tyrimui reikės tik to, ką vadiname ***imtis***. Imtis yra svarstomos "populiacijos" dalis.

Imčių parinkimas arba ***imties atranka*** turi būti atlikta rūpestingai ir vadovautis tam tikra logiška intuicija:

1. Imtis turi būti ***teisinga*** ir ***atstovauti visai populiacijai,*** kuri yra tiriama
2. Imtyje turi būti ***pagrįstas skaičius*** testuojamų ar skaičiuojamų dalykų
3. ***Imčių pavyzdžiai ar atsitiktinė atranka*** turi būti naudojami tyrime

RASTI KOKĮ PRASMINGĄ IMTIES ATRANKOS IR APKLAUSOS PAVYZDĮ (YOUGOV GALI VYKTI) IR PAAIŠKINTI JĮ PAPRASTAIS ŽODŽIAIS !!!!!

Reklamos įmonė samdo YouGov, kad padėtų jiems suprasti, koks laikas būtų optimalus konkrečiam produkto rodymui JK.

Tada YouGov parengia internetinę anketą, kuri yra siunčiama jų registruotų vartotojų grupei, atstovaujančiai įvairaus amžiaus socialines ir ekonomines grupes ir kitus demografinius tipus (šiuo metu JK yra 1 mln. vartotojų)

Kaip mes pateikiame neapdorotus duomenis?

Šio skyriaus pradžioje apžvelgėme duomenų organizavimą. Naudojome du duomenų aprašymo pavyzdžius, pagrįstus Vimbldono moterų finalinių rungtynių rezultatais.

Dabar panagrinėkime tradiciškesnį pavyzdį. 32 mokinių klasė surašė egzaminą su šiais rezultatais:

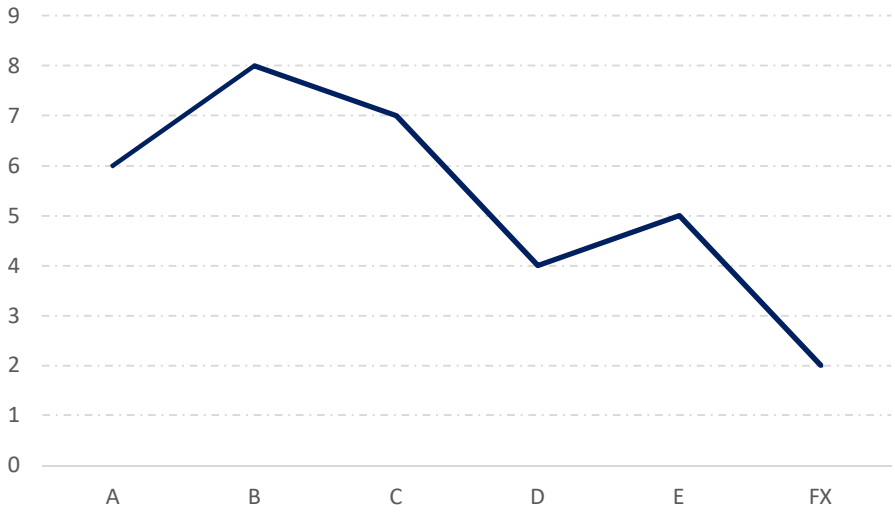
82, 90, 61, 80, 88, 96, 58, 74, 69, 82, 53, 78, 72, 58, 54, 95, 99, 86, 80, 95, 99, 88, 83, 51, 65, 77, 28, 82, 70, 73, 100, 19

Išlaikymo slenkstis yra 50 taškų. Kiekvienas mokinys, surinkęs daugiau negu 50 taškų (bet ne 50) išlaikys dalyką ir jam bus priskirtas įvertinimas pagal vertinimo intervalus. Žemiau mes naudojame įvairius įrankius, kurie padeda mums sutvarkyti ir suskirstyti mūsų duomenis:

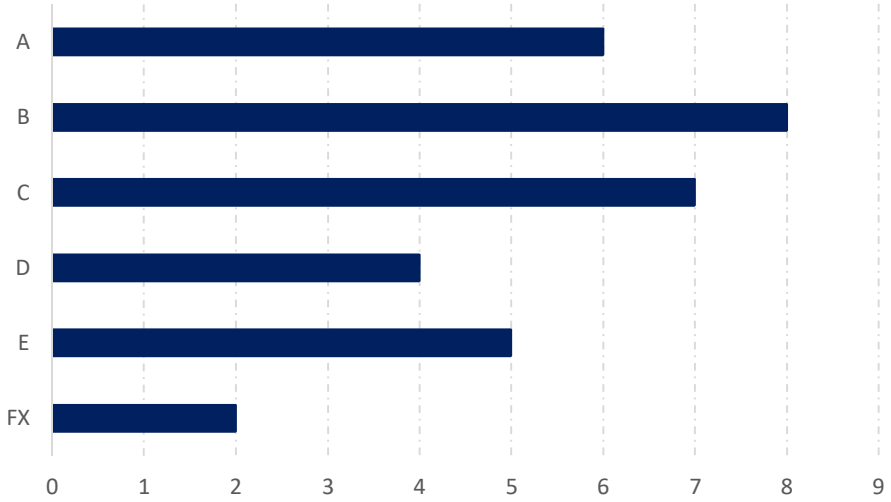
1. *Duomenų sutvarkymas lentelėje:*

Lygis	Rezultatas	# Mokinių	% Mokinių
A	100 - 91	6	19%
B	90 – 81	8	25%
C	80 – 71	7	22%
D	70 – 61	4	13%
E	60 – 51	5	16%
FX	50 – 0	2	6%

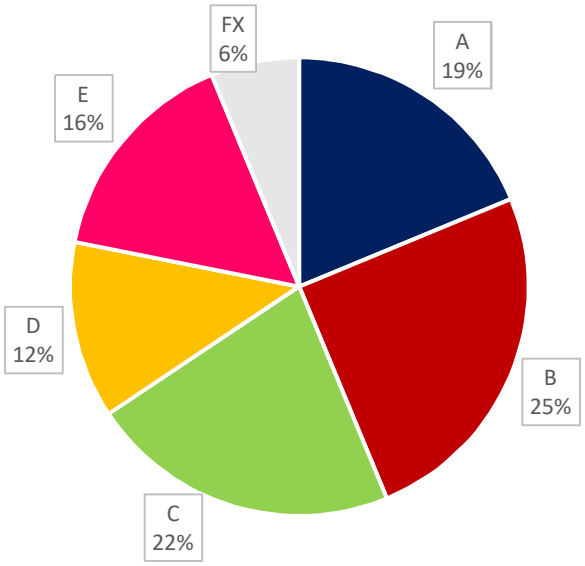
2. *Suskirstymas naudojant Linijos diagramą:*



3. *Suskirstymas naudojant Juostinę diagramą:*



4. *Suskirstymas naudojant skritulinę diagramą*



Ką mums sako vidurkis, mediana, moda?

Mes jau parodėme, kaip rinkti, tvarkyti ir pateikti savo duomenis. Žvelgiant į diagramas, jau galime perskaityti šiek tiek informacijos apie tai, kurie požymiai buvo dažniausiai. Norėdami įvertinti kitų tipų informaciją, paaiškinsime, ką reiškia, vidurkis, mediana ir modusas.

1. **Vidurkis:**
    - Statistikoje tai ką anksčiau apibūdinome kaip aritmetinį vidurkį, vadinsime ***vidurkiu***
    - Šiame pavyzdyje mums gali būti įdomu sužinoti koks buvo bendras testo rezultatų vidurkis ar rezultatų vidurkis kiekvienam lygiui.
    - ***Norėdami apskaičiuoti N skaičių aibės vidurkį, sudėsime visus šiuos skaičius ir padalysime šią sumą iš N***
    - Todėl pirmajam vidurkiui reikės tik susumuoti visus balus (= 2,385) ir padalyti tai iš laikiusių testą mokinių skaičiaus (= 32); Gautas bendrų balų vidurkis bus lygus **74.53**
    - Rezultatų vidurkis kiekvienam lygiui pareikalaus papildomo žingsnio
    - Šiame žingsnyje mes pažvelgsime į individualius rezultatus, kuriuos mokiniai pasiekė kiekvienam lygiui (A lygiui yra: 96, 95, 99, 95, 99 ir 100)
    - Kai turėsime šią informaciją, mes taikysime tą patį skaičiavimą, kaip ir anksčiau; susumuokite balus ir padalykite šią sumą iš balų skaičiaus (t.y. (96+95+99+95+99+100)/6 = 97.33)

Lygis	Įvertinimų vidurkis
A	97.33
B	85.13
C	76.29
D	66.25
E	54.80
FX	23.50
Bendras	74.53

2. **Mediana:**
    - Užuot žiūrėję į vidurkį, galime domėtis, kokia yra skaičiaus reikšmė mūsų duomenų viduryje
    - ***Mediana yra vidurinis balas skaitine tvarka išdėstytų duomenų rinkinyje; Sekos su lyginiu elementų skaičiumi mediana bus lygi dviejų vidurinių skaičių aritmetiniam vidurkiui***
    - Kodėl mums būtų įdomu žinoti šį skaičių kartu su įprastu vidurkiu?
    - Įsivaizduokime penkias lėkštes su sausainiais; ant keturių jų yra penki sausainiai, o paskutinė lėkštė tuščia
    - Į virtuvę ateina penki vaikai ir kiekvienas čiumpa po vieną lėkštę ir suvalgo lėkštės turinį
    - Ką vidurkis rodo šiuo atveju yra, kad mes turėjome (5+5+5+5+0)/5 = 4 sausainius vienam vaikui
    - Tačiau mes žinome, kad vienas iš vaikų neturėjo jokio sausainio ir ji tikrai nesutiktų su teiginiu, kad vidutiniškai ji suvalgė keturis sausainius
    - Šių lėkščių mediana bus vidurinė šios sekos reikšmė: 0, 5, 5, 5, 5 => mediana = 5
    - Vidurkis žemiau vidutinio skaičiaus rodo, kad viena ar daugiau lėkščių turėjo mažiau nei penkis sausainius ir todėl kai kurie vaikai dar galėjo būti alkani
    - Kitą vertus, mes galėtume turėti keturias tuščias lėkštes ir vieną su penkiais sausainiais; Vidurkis mums rodo, kad kiekvienas vaikas turėjo po vieną sausainį (0+0+0+0+5)/5 = 1, bet mediana suteikia mums realesnį vaizdą

- Šiuo atveju vidurkis būtų lygus 0, todėl mums gali tekti atidžiai pasižiūrėti, kaip sausainiai buvo išdalinti ir kiek vaikų vis dar yra alkani
- Žemiau mes matome kiekvieno lygio medianą ir bendrą visų balų medianą; ką mes galime pastebėti iš rezultatų?

Lygis	Medianos balai
A	97.50
B	84.50
C	77.00
D	67.00
E	54.00
FX	23.50
Bendras	79.00

3. **Moda:**
   - Analizuodami savo duomenis mes taip pat galime paklausti savęs, kuris rezultatas pasirodo dažniausiai; tas rezultatas yra vadinamas ***moda***
   - Norėdami rasti moda, vėl sutvarkome duomenis skaitine tvarka ir ieškome skaičiaus, kuris kartojasi dažniausiai
   - Mūsų pavyzdyje šis skaičius buvo **82**; buvo trys mokiniai, surinkę 82 balus, buvo du mokiniai, kurie pasiekė rezultatą 58, 80, 88, 95 ir 99
   - Kas nutiktų scenarijuje, kur trys mokiniai surinktų 82, bet taip pat trys mokiniai surinktų 58?
   - Pagal šį scenarijų sakytume, kad skaičių rinkinys turi ***dvi modas*** arba, kad yra ***bimodalinis***
   - Ką daryti, jeigu kiekvienas mokinys pasiekia skirtingą rezultatą? (t.y. kiekvienas skaičius sekoje pasirodo tik vieną kartą)
   - Jei kiekvienas skaičius duomenų rinkinyje pasirodo tiek pat kartų, ***modos nėra***

Šaltiniai:

https://www.statista.com/statistics/280393/womens-tennis-players-with-the-most-victories-at-wimbledon/
https://yougov.co.uk/about/panel-methodology/
Dressler and Keenan: Integrated Mathematics (Second Edition) (https://books.google.sk/books/about/Integrated_Mathematics_Course_1.html?id=h4oiXcjGT3oC&redir_esc=y)