Emilija Vasiliūnaitė
Sintetinės biologijos projektas
Lietuvoje sugautų žiurkių duomenų analizė

I dalis. Įvadas. Duomenų kilmė, tikslas, galimos klaidos ir jų likvidavimas, analizės kryptys

Turimi duomenys apie Lietuvoje 2013.09.21 - 2017.10.01 sugautas žiurkes (**1 pav.**). Tai mano kolegės surinkti duomenys, jų pakankamai nemažai, todėl galima mėginti bandyti atlikti šiokią tokią statistinę analizę (nors reikėtų daugiau duomenų su „teigiamomis“ žiurkėmis, kuriose rasti poliomos virusai, nes jie pasitaiko sąlyginai retai, todėl reikia ištirti labai daug mėginių). Turiu savo analogiškus duomenis su pelėmis, tačiau jų imtis mažesnė (n=34) ir virusas buvo rastas tik dviejose pelėse, tad nieko statistiškai neišanalizuosi. Žiurkės buvo gaudytos siekiant ištirti žiurkių poliomos virusų 1 ir 2 (RatPyV1 ir 2) genomų pokyčius bei jų paplitimą tarp *Rattus norvegicus* ir *Rattus rattus* žiurkių Lietuvoje. Anksčiau lietuviškos žiurkės dėl PyV nebuvo tirtos. Abu poliomos virusai iki šiol buvo rasti tik *Rattus norvegicus* žiurkėse, todėl įdomu ištirti, ar jie infekuoja ir *Rattus rattus*.

```
In [100]: data_all.head()
```

Out[100]:

	Nr	Rūšis	Vieta	Pagavimo_data	Lytis	Q(g.)	L(mm)	Luo(mm)	P(mm)	A(mm)	Amžius	Pastabos	RatPyV1	RatPyV2
0	R121	R.norvegicus	Joniškio raj., Veršių km.	2013.09.21	vyr.	162.0	187.0	152.0	38.0	15.0	ad.	Tvartas, laikomos karvės, kiaulės	NaN	NaN
1	R122	R.norvegicus	Joniškio raj., Šarūnų km.	2013.09.27	vyr.	184.0	202.0	156.0	39.0	17.0	ad.	Tvartas, laikomos karvės, kiaulės	NaN	NaN
2	R123	R.norvegicus	Panevėžio raj., Krekenava	2014.06.10	vyr.	248.0	233.0	171.0	39.0	17.0	ad.	Stumbrynas	NaN	NaN
3	R124	R.norvegicus	Gedimino pr., Vilnius	2014.10.20	mot.	38.0	120.0	90.0	28.0	14.0	juv.	NaN	NaN	NaN
4	R125	R.norvegicus	Joniškio raj., Šarūnų km.	2015.02.15.	vyr.	139.0	185.0	153.0	38.0	15.5	sub.	Tvartas, laikomos karvės, kiaulės	NaN	NaN

1 pav. Turimi žiurkių duomenys.
Turimi duomenys apie Lietuvoje sugautų žiurkių rūšį, pagavimo vietą ir datą, lytį, svorį, kūno ilgį, uodegos ilgi, galinės pėdos ilgį, ausies ilgį, amžių, pagavimo vietos tipą bei žiurkėse rastus poliomos virusus.

Nėra informacijos, kaip žiurkės buvo gaudytos, kaip buvo nustatytas svoris, uodegos, kūno, pėdos, ausies ilgiai. Greičiausiai esama tam tikrų paklaidų, tačiau jos neturėtų turėti labai didelės reikšmės duomenų prasmei, t. y. Iš duomenų galimos padaryti išvados neturėtų reikšmingai keistis. Rūšies, lyties nustatyme taip pat galėjo įsivelti klaidų. Jei žiurkės rūšis buvo nustatyta blogai ir joje buvo aptiktas poliomos virusas, galimos klaidingos išvados. Dėl šios priežasties daugelis žiurkių, kuriose buvo aptikti poliomos virusai, buvo papildomai tiriamos, nustatant jų rūšį molekuliniais metodais, pagal mitochondrijų *Citochoromo B* geno seką. Lyties nustatymas greičiausiai nėra labai svarbus, nes bent jau iki šiol nebuvo pastebėta jokių koreliacijų tarp gyvūno lyties ir poliomos viruso infekcijos dažnio/būdingumo, tad dabar nesitikima jų nustatyti, bet, žinoma, būtų labai įdomu tokią koreliaciją nustatyti.

Pagavimo vieta ir data yra svarbūs kintamieji, nes jie turėtų atspindėti poliomos virusų paplitimą tarp žiurkių. Panašiu metu toje pačioje vietoje pagautos žiurkės turi didelę tikimybę būti užsikrėtusios tuo pačiu virusu. Bet kuriuo atveju, čia neturėtų būti įvelta daug klaidų. Dar vienas įdomus kintamasis – žiurkių pagavimo vietos pobūdis. Ar tam tikroje vietoje (pvz.: tvarte, kur laikomos vištos) žiurkės dažniau būna užsikrėtusios, nei sugautos stumbryne? Taip pat galima įdomiai pasiaiškinti ir žiurkės amžiaus/užsikrėtimo sąveikas (galbūt vyresnės žiurkės dažniau būna užsikrėtusios?).

Sudėtingiausia dalis, kurioje pasitaiko daugiausiai klaidų, yra poliomos viruso buvimo/nebuvimo mėginyje nustatymas. Šio tyrimo metu buvo analizuoti keli žiurkių mėginiai – krūtinės ertmės skystis, plaučiai, inkstai, blužnis (lentelėje duomenys sudėtiniai, t. y. Jei bent viename iš šių mėginių buvo rastas PyV, žiurkė „teigiama“. Daugelyje žiurkių PyV buvo rasti bent keliuose skirtinguose mėginiuose, tai patikina, kad žiurkė tikrai buvo užsikrėtusi virusu). Šiuose mėginiuose atskirai su specifiniais pradmenimis PGR metodu buvo ieškota RatPyV1 ir 2. Kadangi pagrindinis paieškos metodas – PGR, yra didelė tikimybė mėginius užkrėsti vienus nuo kitų. Be to, laboratorijoje daug dirbama su kitų rūšių poliomos virusais, kurie jau yra plazmidėse. Kai kurių poliomos virusų sekos panašios pakankamai, kad pradmenys ant jų sėstų ir padaugintų tam tikrus fragmentus. Dėl šių priežasčių imtasi atsargumo priemonių: žiurkių mėginiai laikomi ir su jais dirbama atskirame kambaryje. Į šį kambarį nenešama jokia plazmidinė DNA, taip pat čia nepatenka mėginiai po PGR (reakcija leidžiama kitame kambaryje ir po jos su mėginiais toliau dirbama kitame kambaryje), į jį ribotas pašalinių žmonių patekimas. Taip pat laikomasi kitų švaraus darbo taisyklių – pirštinės, darbo vietos švara, vienkartiniai pipečių antgaliai su oro filtrais, darbo priemonių dezinfekija ir t. t.

Poliomos viruso buvimas nustatomas po PGR paleistame agarozės gelyje esant tam tikro ilgio DNR fragmentui. Deja, kartais matomi nespecifiniai panašaus dydžio fragmentai, tokiu atveju vienintelis būdas įsitikinti, ar tikrai mėginyje yra poliomos virusas – fragmentą sekvenuoti. Žinoma, sekoskaita bet kuriuo atveju yra patikimiausias būdas nustatyti, ar žiurkė tikrai užsikrėtusi poliomos virusu. Daugelyje pateikiamų „teigiamų“ žiurkių vieno ar kito poliomos viruso buvimas buvo patvirtintas sekoskaita.

Reikia paminėti, kad kai kurių duomenų trūksta. Ne visose žiurkės buvo ištirtos dėl poliomos virusų, trūksta kai kurių duomenų apie žiurkių morfologiją, lytį, pagavimo vietą. Dėl PyV kai kurios žiurkės nebuvo tirtos, nes nebuvo gauti jų mėginiai, o kodėl trūksta kitų duomenų – sunku pasakyti.

II dalis. Duomenų redagavimas ir aprašymas

Kiekviena žiurkė turi savo numerį nuo R121 iki R237, iš viso turimi duomenys apie 117 žiurkių. Vienas svarbus šių duomenų aspektas – trūkstamos reikšmės (**1 lentelė**). Didžiausia problema – žiurkės R121-154 (išskyrus R132 ir R133) bei R170 nebuvo tirtos dėl poliomos virusų. Kadangi trūksta aktualiausių duomenų – šias žiurkes reiks pašalinti iš tyrimo. Jų pašalinimas neturėtų sudaryti didelių problemų, nes šių duomenų trūksta atsitiktinai, be to, jų vis tiek negalima panaudoti analizei.

Kintamasis	**Kiek vnt. duomenų trūksta**
Nr	0
Rūšis	0
Vieta	0
Pagavimo_data	0
Lytis	5
Q(g.)	1
L(mm)	18
Luo(mm)	18
P(mm)	16
A(mm)	19
Amžius	0
Pastabos	2

RatPyV1	33
RatPyV2	33

1 lentelė. Trūkstami duomenys.

Duomenų trūksta ir keliems kitiems kintamiesiems. Kadangi visų žiurkių su trūkstamais duomenimis šalinti iš analizės nenoriu, nes tektų atmesti ir keletą žiurkių, kuriose buvo rasti poliomos virusai, o aptiktų virusų ir taip nedaug, teks priskirti trūkstamus duomenis. Trūkstamiems tolydiesiems duomenims apie žiurkių svorį (Q(g.)), kūno ilgį (L(mm)), uodegos ilgį (Luo(mm)), galinės pėdos ilgį (P(mm)), ausies ilgį (A(mm)) sugeneruoti naudojau Fancyimpute IterativeImputer algoritmą. **2 lentelėje** aprašytos pagrindinės šių kintamųjų statistikos, po nežinomųjų elementų užpildymo ir visų likusių žiurkių su nežinomaisiais pašalinimo. Jos skiriasi nuo statistikų, sudarytų pagal pirminį duomenų variantą (ar reikšmingai, netikrinau. R193 žiurkei pagal svorį ir kūno ilgį priskyriau vyrišką lytį. Likusias žiurkes, turinčias nežinomųjų, pašalinau. Iš 117 liko 80 tinkamų analizei žiurkių. (visa informacija apie duomenų pašalinimą/užpildymą ir statistikas failuose p*ython-3-notebook-remove-missing-data.ipynb* ir *Python3-notebook-rats-descriptive-after-missing.ipynb*).

	Q(g.)	L(mm)	Luo(mm)	P(mm)	A(mm)
count	80.000000	80.000000	80.000000	80.000000	80.000000
mean	151.744211	165.380657	164.272802	34.918336	18.809298
std	117.636167	41.535664	29.110053	5.145127	2.752830
min	21.000000	88.000000	101.000000	23.000000	11.000000
25%	65.750000	131.338998	146.646865	31.984427	18.000000
50%	109.000000	163.500000	163.748109	34.382182	18.468760
75%	231.500000	197.750000	187.500000	39.250000	20.000000
max	435.000000	249.000000	231.000000	47.000000	27.000000

2 lentelė. Tolydieji žiurkių duomenys.
Reikšmės – Q(g.) svoris gramais, L(mm) kūno ilgis milimetrais (be uodegos), Luo(mm) uodegos ilgis milimetrais, P(mm) galinės pėdos ilgis, A(mm) ausies ilgis milimetrais.

Žiurkių svorio ir ausies ilgio matavimai pasiskirstę nenormaliai pagal scipy.stats.normaltest ir Šapiro testus (scipy.stats.shapiro) ($p < 0,05$). Kūno ilgio duomenys pagal normaltest testą pasiskirstę normaliai ($p = 0.062$), pagal Šapiro – nenormaliai ($p = 0.044$). Likę duomenys – uodegos ir pėdos ilgio – pasiskirstę normaliai pagal abu testus ($p >0,05$).

Surinkti nominalieji duomenys apie žiurkės

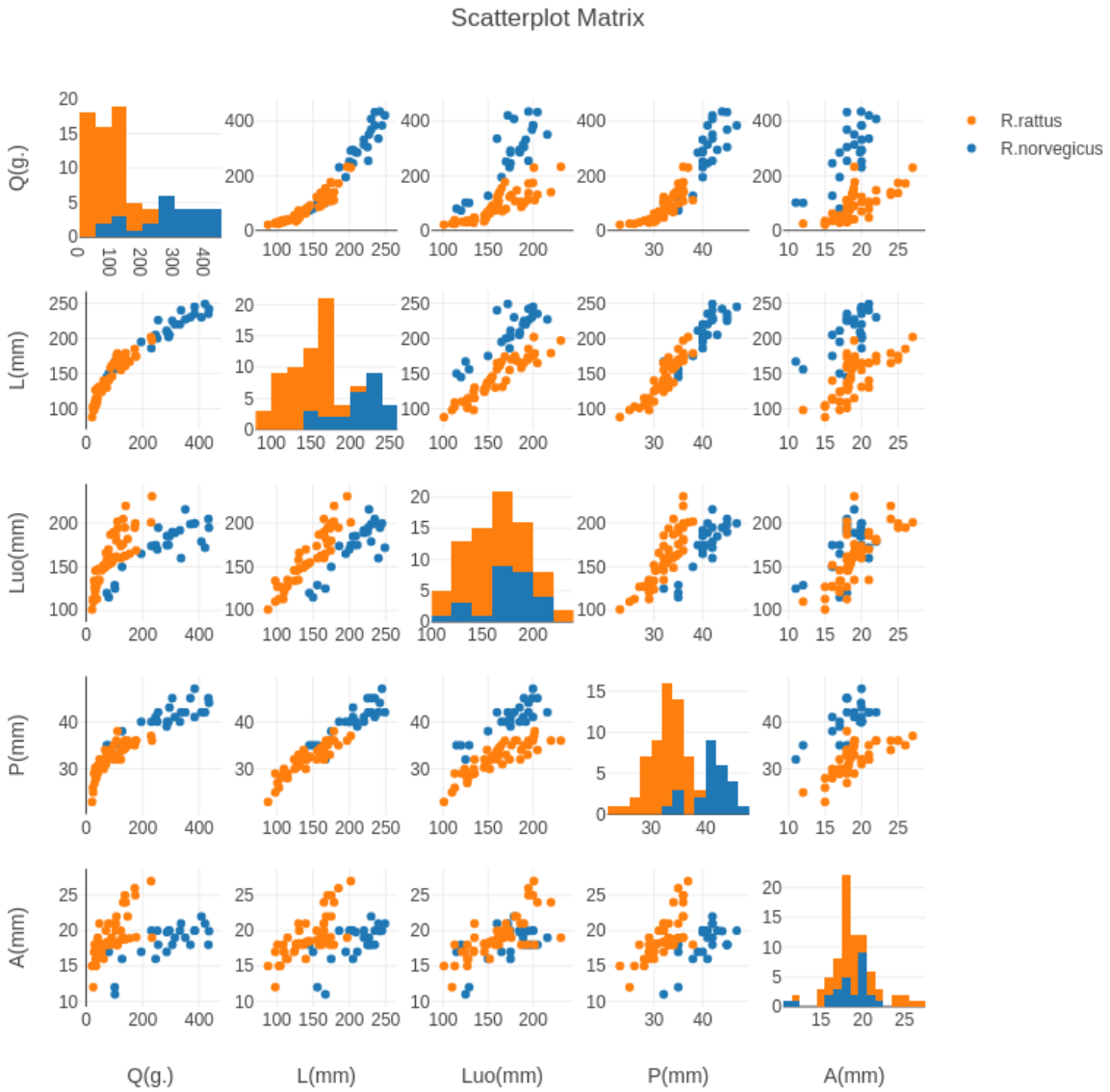
- rūšį (trys rūšys, 'R.norvegicus' - 26 , 'R.rattus' - 54);
- pagavimo vietą (10 vietų, Zarasų raj., Pakniškių km.', 'Vilniaus raj.', 'Grubupių km., Šilutės raj. ', 'Zarasų raj., Antazavės gyv.', 'Vilnius, senamiestis', 'Vilnius, Gedimino pr.', 'Joniškio raj., Šarūnų km.', 'Rokiškio raj., Kovelių km.', 'Rokiškio raj., Laibgalių km.', 'Zarasų raj., Dusetų gyv.' Laibgalių km.', dažniausia sugavimo vieta - Zarasų raj., Pakniškių km. (46 kartus pasikartoja);
- pagavimo datą (36 datos nuo 2014.01.03 iki 2017.10.01, dažniausia data 2017.09.07, pasikartoja 10 kartų);
- lytį (2 rūšys, 'vyr.', 'mot.', mot. - 38, vyr. - 42);
- amžių ('ad.', 'juv.', 'sub.', ad. 36, juv. 27, sub. 17);
- tikslią sugavimo vietą (8 reikšmės, 'Tvartas, laikomos karvės, kiaulės', 'Tvartas, laikomos vištos', 'Tvartas, laikomos karvės', 'Ūkinis pastatas, gyvūnų nelaikoma', 'Gyvenamasis

namas', 'Tvartas, laikomos kiaulės', 'Kiaulių auginimo kompleksas', 'Paukštynas', dažniausia reikšmė - Tvartas, laikomos vištos, 39);

- žiurkėse aptikti poliomos virusai – žiurkių poliomos virusas 1 (RatPyV1) ir žiurkių poliomos virusas 2 (RatPyV2); RatPyV1 aptiktas 10, RatPyV2 – 5 žiurkėse.

III dalis. Duomenų analizė.

Pradėjau nuo žiurkių morfologinių duomenų analizės. Pirmiausia nusibraižiau Scatterplotus (**2 paveikslas.**) pagal žiurkės rūšį ir svorį (Q(g.)), kūno ilgį (L(mm)), uodegos ilgį (Luo(mm)), galinės pėdos ilgį (P(mm)), ausies ilgį (A(mm)), kad pamatyčiau bendras duomenų tendencijas. R. norvegicus ir R. rattus iš esmės pagal visus matavimus išsiskiria į dvi grupes. Tai patvirtino ir atlikti Wilcoxon rank-sum testai - pagal svorio ($p = 2{,}325e^{-09}$), kūno ($p = 5{,}672e^{-10}$), uodegos ($p = 0{,}015$) ir pėdos ($p = 6{,}665e^{-11}$)) ilgio duomenis dviejų rūšių žiurkės reikšmingai skyrėsi.



2 pav. *R.norvegicus* ir *R.rattus* morfologinių duomenų sklaidos grafikai.

Taip pat iš grafiko atrodo, kad daugelis parametrų koreliuoja tarpusavyje, todėl atlikau neparametrinį Spearman’o koreliacijos testą. Rezultatai pateikti **3 lentelėje**. Su kitais parametrais

silpai-vidutiniškai teigiamai koreliuoja ausies ilgis, kiti parametrai pasižymi stipriomis teigiamomis koreliacijomis. Išanalizavus *R.rattus* ir *R. norvegicus* duomenis atskirai, matyti, kad *R. rattus* ausų ilgis stipriai koreliuoja su kitais parametrais (r>0,7), o *R. norvegicus* – silpnai arba vidutiniškai (0,29 < r 0,62). Kaip ir galima būtų tikėtis, pagal amžių duomenys taip pat susiklasifikuoja į tris amžiaus grupes. Tai galima matyti ir iš scatter grafiko. Susigrupavimą pagal amžių patvirtino ir Wilcoxon rank-sum testai. Visi parametrai reikšmingai skiriasi tarp amžiaus grupių, išskyrus ausies ilgį, kuris tarp sub ir ad pelių reikšmingai nesiskyrė (p = 0.099).

	Q(g.)	**L(mm)**	**Luo(mm)**	**P(mm)**	**A(mm)**
Q(g.)	1.000000	0.975300	0.807508	0.944059	0.507445
L(mm)	0.975300	1.000000	0.814310	0.950974	0.478751
Luo(mm)	0.807508	0.814310	1.000000	0.760459	0.629610
P(mm)	0.944059	0.950974	0.760459	1.000000	0.426929
A(mm)	0.507445	0.478751	0.629610	0.426929	1.000000

3 lentelė. Sprearman koreliacijos koeficientai tarp žiurkių morfologinių duomenų

Atlikau Chi_square testus su lygindama RatPyV1 ir RatPyV2 su kitais nominaliaisiais kintamaisiais. Testas parodė RatPyV1 priklausomybę nuo žiurkės rūšies (p = 0,002), RatPyV2 tokia priklausomybe nepasižymėjo. RatPyV1 taip pat priklausė nuo žiurkės pagavimo vietovės (p = 0,0007) bei jos ypatybių (t. y. Kokiame pastate pagauta, „Pastabos“) (p = 0,031). Taip pat atlikau Wilcoxon rank-sum testus, lyginant RatPyV1 ir RatPyV2 užsikrėtusioms žiurkėms būdingą kiekvieną morfologinį parametrą. Žiurkių svoris (p = 0,0199), kūno ilgis (p = 0,0199), uodegos ilgis (p = 0,0274) ir galinės pėdos ilgis (p = 0,0169) reikšmingai skyrėsi tarp lygintų grupių. Ausies ilgis reikšmingai nesiskyrė (p = 0,1589).

Atliktus veiksmus (ir dar visokių nesąmonių) galima rasti faile *Python3-notebook-rats-analysis.ipynb.*

III dalis. Išvados

Iš tiesų nustebino, kad visai pavyko įgyvendinti pradinius tikslus. Pavyko patikrinti ne tik tolydžius duomenis, bet ir nominalius, ir netgi rasti koreliacijų/sąryšių, atskirti duomenų grupes. Gauti rezultatai atitinka randamus literatūroje. *R. novegicus* morfologija nuo *R. rattus* išties gan smarkiai skiriasi, mano gauti duomenys su tuo sutinka. Taip pat nebuvo netikėta, kad morfologiniai parametrai (teigiamai) koreliuoja tarpusavyje, bet smagu, kad statistinė analizė tai patvirtino. Pavyko parodyti, kad žiurkių morfologiniai parametrai reikšmingai skiriasi priklausomai nuo žiurkės amžiaus.

Tiesa, buvo netikėta, kad duomenys nėra pasiskirstę normaliai, juk tai atsitiktiniai matavimai, tikėjausi normalaus pasiskirstymo. Matyt, vis dėlto įsivėlusi tam tikra paklaida, gal dėl mano missing data likvidavimo/dalies duoemnų pašalinimo, gal tiesiog žiurkės su tam tikrais požymiais lengviau ir dažniau pagaunamos, o galbūt įsivėlė matavimo paklaidos. Dėl duomenų nenormalumo teko taikyti neparametrinius testus.

Džiugiausia, kad pavyko išanalizuoti ir apsikrėtimo RatPyV1 ir RatPyV2 santykį su kitais kintamaisiais. Vyliausi, jog bus PyV priklausomybė nuo žiurkės rūšies, ir iš tiesų RatPyV1 yra priklausomas nuo rūšies. RatPyV2 analizė neparodė priklausomybės nuo rūšies, bet to ir buvo galima tikėtis, nes užsikrėtusių RatPyV2 žiurkių kiekis labai mažas, todėl sunku tikėtis reikšmingų rezultatų. Vis dėlto, išanalizavus Įdomu, kad buvo rasti skirtumai tarp RatPyV1 ir RatPyV2 užsikrėtusių žiurkių morfologinių duomenų. Panašu, kad šie skirtumai egzistuoja būtent dėl apkrėstų žiurkių rūšims būdingų morfologinių skirtumų. Taip pat, kaip ir vyliausi, buvo surastos RatPyV1 priklausomybės nuo sugavimo vietos ir jos pobūdžio.

Akivaizdu, kad duomenų mažai, norint padaryti tikslias išvadas, bet jie gali nurodyti tam tikras tolimesnių tyrimų kryptis. Svarbiausia, kad stebimos geros tendencijos – rezultatai atitinka tikėtinus pagal iki šiol turimą informaciją.