SUSIJĘ POŽYMIAI. SĄSAJOS ĮVERTINIMAS

1. Bendrasis supratimas

Iki šiol buvo kalbama apie *pavienius* tiriamųjų objektų (ar reiškinių) požymius. Tačiau tyrinėtojas neretai fiksuoja du ar kelis tų objektų (reiškiniu) požymius, tad ištyrus vieną imtį susidaro faktiškai du ar keli pasiskirstymai – kiekvienam požymiui atskiras pasiskirstymas. Žinoma, galima kiekvieną jų tyrinėti autonomiškai, nė kiek neatsižvelgiant į kitų, "gretutinių" požymių pasiskirstymus, bet taip pat galima ir kelti klausimą, ar vieno požymio pasiskirstymas kaip nors siejasi su kito požymio pasiskirstymu, ar ne. Kitais žodžiais sakant – ar tiriamieji požymiai vienas kitą kaip nors įtakoja, ar yra tarpusavy susiję ir *priklausomi* vienas nuo kito, ar nėra.

Iš karto reikia pasakyti, jog „priklausomybės spektras“ čia, apskritai paėmus, labai platus – nuo visiškos nepriklausomybės (jokių sąsajų tarp požymių nebuvimo) iki absoliučios priklausomybės, kuomet vieno požymio reikšmės pokyčiai neabejotinai sukelia (sąlygoja) ryškius, vienareikšmiškus kito požymio reikšmės pokyčius. Todėl kyla specifinis uždavinys: kaip nors išmatuoti ir įvertinti skirtingų požymių, būdingų imties objektams, tarpusavio sąsajos laipsnį (mastą) bei nustatyti, *kaip* vienas iš jų konkrečiai įtakoja kitą.

Nuosekliai bei detaliai šiuos uždavinius nagrinėja statistikos skyrius, vadinamas **koreliacine** ir **regresine analize**. *Koreliacijos* (ši sąvoka yra artima priklausomybės savokai ir reiškia maždaug tą patį, ką „stochastinė sąsaja“) tyrimas leidžia konstatuoti patį sąsajos buvimą, apčiuopti jos egzistavimą, o *regresijos* – jau ir atsakyti į klausimą, kaip konkrečiai vieno požymio reikšmės pokytis įtakoja kito požymio pasiskirstymą.

Aprašomosios gi statistikos priemonėmis pakanka tik apčiuopti ir parodyti, ar požymiai vienas kito pasiskirstymą kaip nors „veikia“, ar ne. Dėl paprastumo galima apsiriboti ir dviem požymiais, kurie abu yra būdingi kiekvienam imties objektui (kartais jie vadinami poriniais požymiais): jeigu tokių požymių yra daugiau (3, 4 ir t.t.), tai juos iš principo galima irgi sugrupuoti po du ir taip „suvesti“ į porinius, tik porų tuomet susidarys daugiau.

2. Poriniai dažnumai ir jų lentelės

Vienas iš parankių būdų dviejų požymių priklausomybei pasekti yra vadinamieji poriniai dažnumai. Itin paranku juos skaičiuoti tada, kai požymių reikšmės būna diskrečiosios ir negrupuotos; tolydžiąsias bei „smulkiažingsnes“ diskrečiąsias reikšmes, norint nustatyti porinius jų dažnumus, tenka sugrupuoti intervalais.

Visų paprasčiausias atvejis, savaime suprantama, susidaro tuomet, kai abudu stebimieji imties objektų požymiai yra binariniai, t. y. tegali įgyti tik po vieną kurią iš dviejų išvis jiems tebūdingų reikšmių. Pavyzdžiui, binariški būtų tokie skiemenų požymiai kaip **atvirumas** (skiemuo tegali būti arba *atvirasis*, arba *uždarasis*), **ilgumas** (skiemuo būna arba *ilgasis*, arba *trumpasis*), **kirčiuotumas** (skiemuo esa arba kirčiuotas, arba ne). Suprantama, jog kiekvienas imtin patekęs skiemuo turės visus tris šiuos požymius, kurių visi yra kokybiniai ir kiekvienas tegali įgauti tik po vieną iš dviejų paminėtųjų atitinkamų reikšmių. Jeigu būtų fiksuojami ir tiriami ne visi trys, bet tik du kurie iš jų (tarkime, atvirumas ir ilgumas), tai ir susidarytų situacija, kuriai pravarti būtų „minimali“ porinių dažnumų lentelė:

	Atvirasis (1)	Uždarasis (2)	*Iš viso*:
Ilgasis (1)	$m_{1,1}$	$m_{1,2}$	[ilgųjų]
Trumpasis (2)	$m_{2,1}$	$m_{2,2}$	[trumpųjų]
Iš viso:	[atvirųjų]	[uždarųjų]	[**visų**]

Prityręs „dažnumininkas“ iš tokia lentele pateiktų konkrečių duomenų gana patikimai galės spręsti, ar tie požymiai vienas su kitu siejasi, ar ne. Be abejo, yra daug šiai sąsajai įvertinti skirtų ir specialių apskaičiavimų pagal atitinkamas formules.

Kai bent vienas, o tuo labiau – abu iš porą sudarančių požymių gali įgauti ne dvi, bet daugiau reikšmių, porinių dažnumų lentelė tampa didesnė, nes susidaro daugiau galimų „reikšmių porų“ $m_{i,j}$ (čia vieno požymio reikšmių „eilės numeriai“ žymimi i, antrojo – j).

Pavyzdys. Turime imtį-tekstą iš 325 žodžių. Kiekvieną žodį apibūdina tokie du tiriamieji požymiai: ilgis skiemenimis (galimos reikšmės – nuo 1 iki 5) ir ilgis raidėmis (galimos reikšmės – nuo 1 iki 13). Jau a priori aišku, jog tie požymiai vienas su kitu susiję, nors ir nevienareikšmiškai: kuo žodyje daugiau skiemenų, tuo apskritai turėtų būti daugiau ir raidžių, ir atvirkčiai - kuo daugiau raidžių, tuo daugiau galima tikėtis būsiant ir skiemenų. Ištyrus imtį nustatyti tokie šių požymių reikšmių porų dažnumai:

Ilgis	Ilgis skiemenimis					Iš
raidėmis	1	2	3	4	5	viso:
2	57	0	0	0	0	57
3	46	4	0	0	0	50
4	8	33	0	0	0	41
5	1	46	6	0	0	53
6	0	24	20	0	0	44
7	0	11	18	0	0	29
8	0	0	17	6	0	23
9	0	0	9	3	0	12
10	0	0	1	3	3	7
11	0	0	0	2	3	5
12	0	0	0	2	0	2
13	0	0	0	2	0	2
Iš viso:	112	118	71	18	6	325

Kaip matome, iš teoriškai galimų 13*5=65 reikšmių porų daugelio nepasitaiko iš viso, o tos, kurios pasitaiko, pasiskirsčiusios irgi netolygiai. Matoma ir bendra sąsajos tendencija: didėjant vieno požymio reikšmėms, didėja ir kito reikšmės.

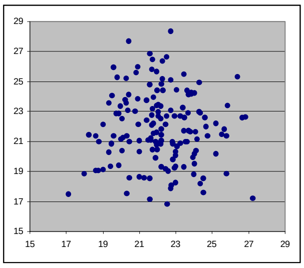
Į porinių dažnumų lentelę galima žiūrėti kaip į dviejų skirtingų pasiskirstymų (ilgio raidėmis ir ilgio skiemenimis) „bendrą plotą“ (sandaugą), atskleidžiantį jų tarpusavio sąveikos struktūrą. Joje vieno požymio, sakysim, *ilgio raidėmis* pasiskirstymas (jį atitinka vertikalioji skiltis "Iš viso") vaizduojamas ir „diferencijuotai“ – išskaidytas į tiek „dalinių“ pasiskirstymų, kiek yra kito požymio – ilgio skiemenimis – reikšmių, tad lygindami tuos „dalinius“ pasiskirstymus tarpusavyje ir galime pamatyti konkrečių antrojo požymio reikšmių „įtaką“ jiems. Kiekvieną tokį dalinį pasiskirstymą „reprezentuoja“ fonu paženklintas lentelės stulpelis. Kitaip gi žiūrint, *ilgio skiemenimis* pasiskirstymas, kurį „visą“ perteikia horizontalioji skiltis "Iš viso", toje pat lentelėje yra pateiktas „išskaidytas“ į dvylika „dalinių“ pasiskirstymų, atitinkančių 12 skirtingų kito požymio – ilgis raidėmis – reikšmių. Kiekvieną iš šių dalinių pasiskirstymų „reprezentuoja“ fonu paženklintos lentelės eilutės.

Tiesiogiai vizualizuoti (t.y. pavaizduoti grafiškai) porinių dažnumų lentelės duomenis paranku yra *trimačiais* grafikais.

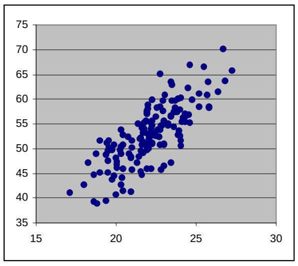
3. Reikšmių porų grafikai

Jie pravartūs ir parankūs tada, kai abu tiriamieji požymiai yra tolydieji ir imtyje užfiksuotos jų reikšmės praktiškai nesikartoja. Vadinasi, norint sudaryti jų porinių dažnumų lentelę, tektų juos grupuoti intervalais (nes negrupuojant porų susidarytų praktiškai tiek, kiek yra bandymų imtyje, ir visų jų abs. dažumas būtų praktiškai vienodas bei lygus 1), o ta procedūra duomenis neišvengiamai „sugrubina“ ir taip yra prarandama dalis informacijos apie vieno požymio reikšmių atitikimą kito reikšmėms. Tokiu atveju racionalu yra tiriamųjų požymių tarpusavio priklausomybę pavaizduoti dvimačiu (plokštuminiu) grafiku (mat, „trečiasis matavimas“, t.y. požymių reikšmių porų dažnis, čia nebetenka prasmės, nes praktiškai tampa pastovus, lygus 1). Toks grafikas čia ir pavadintas reikšmių porų grafiku.

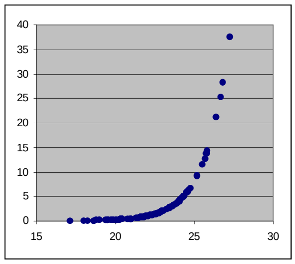
Daroma taip: vieno iš tiriamųjų požymių reikšmių skalė sutapatinama su x-sų ašimi, o antrojo – su y-ų ašimi. Tuomet kiekvieną imties objektą atitinkanti konkreti tų požymių reikšmių pora vaizduojama **tašku**, kurio *abscisė* (x) lygi vieno, o *ordinatė* (y) – kito požymio reikšmėms (savaime suprantama, jog vietoj taško galimi ir kiti nestambūs ženkliukai: kvadratėliai, apskritimai, trikampiai ir pan.). Taškų grafike iš viso susidaro tiek, kiek imtyje yra objektų (o tuo pačiu – ir tiriamųjų požymių reikšmių porų). Tų taškų *pasklidima*s bei *išsidėstymas* grafiko plote, aprėptame x ir y ašių, labai vaizdžiai parodo požymių „įtaką“ vienas kitam: kai požymiai vienas su kitu visiškai nesusiję, taškų sklaida primena pasklidą jų „spiečių“ ar „debesį“, o kuo požymių sąsaja darosi glaudesnė, tuo taškai vis glaužiau „buriasi“ į vienokios ar kitokios formos (pavidalo) liniją, kol galiausiai, kai pasiekiama maksimali (funkcinė) požymių priklausomybė, į tą liniją „susirikiuoja“ visi. Pavyzdžiui:



1 pav.



2 pav.



3 pav.

1 pav. pavaizduotas visiškai nesusijusių požymių reikšmių porų grafikas, 2 pav. – nepilnai (iš dalies) susijusių tiesine priklausomybe, o 3 pav. – absoliučiai (t.y. funkciškai) susijusių netiesiniu ryšiu (konkrečiai čia juos siejanti priklausomybė yra $y = 2^{(x\text{-}22)}$).

Iš linijos, į kurią, stiprėjant požymių sąryšiui, artėja taškai, pavidalo galima spręsti apie sąsajos pobūdį, apie tai, ar ta sąsaja tiesiška, ar ne. Statistikoje priimta skirti *tiesiškas* ir *netiesiškas* požymių sąsajas: pirmosios analitiškai išreiškiamos (aprašomos) tiesės lygtimi, o antrosios – atitinkamomis kreivių (parabolių) lygtimis. Tiesa, apie tikslų požymių priklausomybės reiškimą (aprašymą) tomis lygtimis galima kalbėti tiktai tada, kai priklausomybė yra absoliuti, funkcinė, o kai ji būna nepilna, dalinė, ją tomis lygtimis galima tiktai aproksimuoti, t.y. apibendrinti su tam tikra paklaida.

Pridurtina, kad reikšmių porų grafikuose pravartu statmenimis, išvestais iš x-sų ir y-ų ašių, paženklinti atitinkamų tiriamųjų požymių reikšmių *vidurkius*: tuomet grafikai tampa dar informatyvesni.

4. Požymių sąsajos skaitinis įvertinimas

Jis yra problemiškas dėl to, kad gerokai priklauso nuo sąsajos „formos“, nuo konkretaus tos kreivės, į kurią, stiprėjant sąsajai, išsirikiuoja taškai, pavidalo. Šiaipjau įprasta yra specialiais skaičiavimais tikrinti, kokiu mastu (laipsniu) tiriamųjų požymių sąsaja imtyje artima *tiesinei* priklausomybei. Ypač logiška tai daryti tuomet, kai taškai reikšmių porų grafike išsidėsto į viena kuria nors kryptimi ištęstą elipsę, vadinasi, kai tam tikrą „atstojamąją“ (apibendrinamąją) priklausomybės tiesę jau galima įžiūrėti ir iš akies. Savaime suprantama, kuo labiau taškų sklaidos elipsė ištęsta ir kuo mažiau „vingiuoja“, tuo požymių priklausomybė stipresnė bei tiesiškesnė.

Yra du populiarūs skaitiniai požymių sąsajos artimumo tiesinei sąsajai matai: *kovariacijos* koeficientas (kartais sakoma – kovariacijos rodiklis; jo konkretus dydis priklauso ir nuo mato, kuriuo matuotos požymių reikšmės imtyje, vienetų, todėl jis nėra parankus) bei tiesinės *koreliacijos koeficientas* (kitaisp dar vadinamas K. Pirsono koreliacijos koeficientu; žymimas paprastai $\boldsymbol{r}_{\mathbf{xy}}$) – normuotas „artimumo tiesei" rodiklis, kurio reikšmė gali kisti intervale [-1; 1], todėl labiau parankus ir bene plačiausiai taikomas. Beveik visose statistikos knygose yra pateikiama ir formulių šiems sąsajos artimumo tiesei matams apskaičiuoti; skaičiuoklė *Excel* (o tuo labiau – specializuoti statistinių programų paketai) turi tam skirtas specialias funkcijas.

Tiesinės koreliacijos koeficiento r_{xy} reikšmė, moduliu (absoliutiniu dydžiu) artima vienetui, rodo, kad požymiai vienas su kitu susije ganėtinai *tiesišku* ryšiu, kurį galima išreikšti tiesės lygtimi $y = a + bx$, o to koeficiento ženklas atspindi priklausomybės pobūdį: teigiamas koeficientas rodo, jog požymiai yra susiję pozityviu (teigiamos priklausomybės) būdu, ir, didėjant vieno požymio reikšmėms, *didės* ir antrojo reikšmės bei atvirkščiai. Tuo tarpu neigiamas r_{xy} rodo, kad požymiai yra susiję atvirkštiniu (neigiamos priklausomybės) būdu, ir didėjant vieno iš jų reikšmėms, kito reikšmės mažės bei atvirkščiai.

Dar tikslesnį bei konkretesnį požymių sąsajos pavidalą galima nustatyti, specialiais būdais apskaičiuojant minėtosios tiesės lygties $y = a + bx$ koeficientų a ir b konkrečias reikšmes (tam reikalingas formules irgi galima rasti statistikos knygose, regresinei analizei skirtuose skyriuose). Koeficientas a (laisvasis narys) čia kartais vadinamas *postūmio* koeficientu ir reiškia ieškomiosios „priklausomybės tiesės" poslinkį y-ų ašies atžvilgiu, kitaip sakant, – tašką y-ų ašyje, į kurį ši tiesė atsiremtų tada, kai $x = 0$. Koeficientas b yra krypties koeficientas: jo ženklas reiškia „priklausomybės kryptį" (panašiai kaip ir r_{xy} ženklas – teigiamas rodo, jog didėjant x reikšmėms didės ir y reikšmės, neigiamas gi – kad bus atvirkščiai, kad x reikšmėms didėjant, y reikšmės mažės), o dydis – pokyčio (prieaugio) mastą: kuo didesnė b reikšmė, tuo labiau pakinta y, pakeitus x tuo pačiu žingsniu.

Svarbu yra teisingai suvokti tiesinės koreliacijos koeficiento r_{xy} pramę: jis atspindi *ne* tiek požymių sąsajos *stiprumo* (glaudumo, intensyvumo), kiek tos sąsajos (kitaip sakant, požymių tarpusavio priklausomybės) *panašumo į tiesinę priklausomybę* laipsnį. Todėl palyginti nedidelės, ne kažin ką nuo nulio tenutolusios (į bet kurią pusę!) reikšmės gali reikšti du iš esmės skirtingus dalykus: viena – kad tiriamieji požymiai vienas su kitu nėra arba tėra gana menkai susiję (situacija, panaši į pavaizduotąją 1 pav.), o antra – kad požymių sąsaja, nors ir būdama labai stipri, nėra tiesiška ir jos neįmanoma teisingai aprašyti *tiesės* lygtimi (situacija, panaši pavaizduotąją 3 pav.). Tad šiuo antruoju atveju (kai $|r_{xy}|$ yra palyginti mažas) itin pravartu būna požymių priklausomybę pavaizduoti taip pat ir reikšmių porų grafiku: lengviau bus nustatyti, kuri iš minėtų dviejų priežasčių labiau tikėtina.

Kai tiriamųjų požymių priklausomybė yra iš esmės tiesiška, bet ne absoliuti (ne funkcinė), ir taškai reikšmių porų grafike išsidėsto pabiru, nors ir primenančiu tiesę, lauku, iškyla uždavinys parinkti tokią tiesę (t.y. apskaičiuoti tokias tiesės lygties koeficientų a ir b reikšmes), kuri geriausiai atspindėtų (aproksimuotų) visų tų taškų išsidėstymą. Tam paprastai taikomas vadinamasis *mažiausių kvadratų* metodas: siekiama, kad atstumų tarp dabartinės, faktinės, taškų padėties ir „teorinės" jų padėties, kurią taškai turėtų užimti būdami išrikiuoti į projektuojamą tiesę, kvadratų suma būtų minimali. Tai daugiau teoretikų, o ne praktikų, „taikytojų" veiklos baras: „taikytojams" svarbu teisingai pasirinkti tokius „gatavus" koeficientų a ir b apskaičiavimo būdus (formules), kurie tą kvadratų minimumą savaime užtikrintų.

Iš principo yra galima ir netiesiškų sąsajų (priklausomybių), siejančių tiriamuosius požymius, aproksimacija atitinkamomis kreivėmis, ir specializuotos kompiuterių programos leidžia tai padaryti gana patogiai bei greitai (iš parankių tam uždaviniui programų paminėtina, pvz., *Microcal Origin*). Tačiau didesnės apdairos bei atidos reikalauja netiesiškai susijusių požymių priklausomybės pati interpretacija.