TARTU RIIKLIK ÜLIKOOL

V. Tšervjakov

MATEMAATILISE STATISTIKA ALUSED

TARTU 1970

TARTU RIIKLIK ÜLIKOOL

V. Tšervjakov

MATEMAATILISE STATISTIKA ALUSED

Õppevahend geograafidele

Tartu 1970

Originaali tiitel:

Червяков В.А. Основы математической статистики в географии. Издательство Дальневосточного госуд. ун-та. Владивосток, 1966 (ротапринт).

Tõlkinud U. Pragi

Tõlkija ei peatu käesoleva õpiku eesmärkidel, mida autoritekstis on küllalt valgustatud, vaid püüab motiveerida tõlkimisel tehtud muutusi.

Õpiku venekeelse väljaande ilmumisele järgnes selle elav arutelu geograafide hulgas. Ühel sellisel "Matemaatika ja geograafia" arutelul suvekoolis väljendati soovi, et õpik sisaldaks ka "ettehaaravaid" osi, millel küll ei ole kitsalt utilitaarset tähtsust, mis aga stimuleerivad üliõpilasi matemaatilise statistika sügavamale omandamisele. Peale selle peaks õpik andma üliõpilastele kas või üldise ettekujutuse neist statistilistest meetoditest, mida täpsemalt käsitleda ei ole võimalik piiratud mahu tõttu, millega aga uuema geograafilise kirjanduse lugemisel tuleb sageli kokku puutuda. Õpiku autor töötab praegu vastavate täienduste juures.

On otstarbekas lülitada kõnesolnud täiendused eestikeelsesse väljaandesse. Et aga venekeelse väljaande parandatud trüki ilmumine võtab veel aega, tuli need osad kirjutada tõlkijal, kes täienduste sisu suhtes eelnevalt autoriga põhimõtteliselt kokku leppis.

Autori nõusolekul on osa näiteid asendatud Eesti NSV oludes sobivamatega. Samuti on tõlkes muudetud valemite numeratsiooni, varustatud paragrahvid järjekorranumbritega ja tehtud muid vormilisi muudatusi.

Alljärgnevalt esitame olulisemate muudatuste loetelu.

1. Pikemalt on peatutud statistilise kogumi, väljavõtte ja kogumi elemendi mõistetel.
2. Välja on jäetud paragrahv "Arvulise materjali saamise viisid geograafilistes uurimustes", mis kordab mujal avaldatud materjali.
3. On selgitatud statistilise jaotuse mõistet.
4. Rangemalt on käsitletud standardhälbe ja keskmise ruuthälbe erinevust ja aritmeetilise keskmise vea leidmist. Seejuures on sisse toodud tõenäosustaseme mõiste.
5. Lisatud on VII peatükk, mis käsitleb statistiliste kriteeriumide kasutamist.

T õ l k i j a .

S i s s e j u h a t u s .

Meie ajale on iseloomulik huvi tõus matemaatiliste meetodite vastu mitmesuguste teaduslike ja praktiliste ülesannete lahendamisel.

Numbriliste meetodite juurutamine ka geograafilistesse uurimustesse on aktuaalne ülesanne. Geograafil tuleb tihti tegemist teha suure hulga kvantitatiivse informatsiooniga. Selle asjatundlik töötlemine tõstab järelduste objektiivsust. Hindamatut abi võib siin anda matemaatiline statistika, mis uurib statistiliste andmete süstematiseerimist, töötlemist ja kasutamist teaduslikel ja praktilistel eesmärkidel. Kahjuks alahindavad geograafid seda distsipliini, millest annab tunnistust ka vastava õpiku puudumine.

Pakutav õppevahend on esimeseks näiteks matemaatilise statistika elementide esitamisel geograafiliseks uurimistööks.

Õpik on ette nähtud eelkõige geograafia üliõpilastele. Arvestades nende matemaatilise ettevalmistuse taset, on autor püüdnud esitada materjali võimalikult lihtsas vormis. Siin kirjeldatud meetodite omandamiseks piisab keskkooli matemaatikast. Valgustamist on leidnud vaid need matemaatilise statistika osad, mida on kasutatud või mida võidakse kasutada geograafias. Tõestused on ära jäetud. Selle asemel on toodud palju geograafilise sisuga näiteid, üksikasjalikult tutvustatakse arvutusskeemi ja lihtsustatud arvutusviise. Arvuline materjal on valitud nii, et aritmeetilised raskused ei segaks asja sisuga tutvumist.

Õppevahend taotleb 1) õpetada üliõpilasi kasutama statistiliste andmete kogumise ja ümbertöötamise lihtsamaid viise, 2) näidata statistiliste meetodite rakendamise iseärasusi geograafias, eriti seoses kaardimaterjali kasutamisega.

Geograaf, kes kasutab matemaatilise statistika meetodeid, peab küllalt põhjalikult tundma uuritavate nähtuste olemust ja oskama teaduslikul tasemel interpreteerida saadud tulemusi. Praktika näitab, et ilma küllaldase eruditsioonita, oskuseta püstitada hüpoteese, leida ja selgitada nähtuste põhjuslikke sidemeid, kindlaks määrata kasutatud mõistete täpne sisu ja maht igal konkreetsel juhul - ilma selleta viivad matemaatilised arvutused tihtigi vale järeldustele. On selge, et süüdi ei ole seejuures matemaatika.

I. NUMBRILISE INFORMATSIOONI SAAMINE JA ANDMETE GRUPEERIMINE.

1. Statistilised kogumid ja väljavõttemeetod.

Enne kui asuda arvutuste juurde, peab geograafil olema küllaldane kvantitatiivne informatsioon nähtuste ja protsesside kohta uuritaval alal.

Matemaatiline statistika vaatleb ühetaoliste nähtuste hulka, mida nimetatakse s t a t i s t i l i s e k s k o g u m i k s . Kogumiks võivad näiteks olla territoriaalsed ühikud: maastikud, geograafilised vööndid, rajoonid jne.

Üksikuid nähtusi, ühikuid, mis kuuluvad kogumisse, nimetatakse kogumi e l e m e n t i d e k s . Kõigi elementide arvu nimetatakse kogumi m a h u k s ja tähistatakse N .

Geograafilise keskkonna nähtusi uuritakse mitte ainult ruumis, vaid ka ajas. Seega võivad kogumi elementideks olla ka ajalised ühikud: aasta, kuu, päev jne.

Kogumi elementidele on omane vähemalt üks tunnus, millel on teatud arvuline väärtus või mille alusel võib elemente järjestada.

Ühel ja samal elemendil võib muidugi olla mitu arvudes väljendatavat tunnust. Näiteks administratiivrajoonid ei erine mitte üksnes looduslike, vaid ka majanduslike näitajate poolest (kultuuride saagikus, külvipinnad jms.). Linnad võivad üksteisest erineda elanike arvult, tööstuse kogutoodangu rahaliselt väärtuselt, geograafilistelt koordinaatidelt jne. Aastad kogumi elementidena nähtuste dünaamika uurimisel erinevad saagikuse, temperatuuride summa, sademete hulga jne. poolest.

Niisiis, statistilise kogumi elementideks on uuritavad objektid. Seejuures uurib matemaatiline statistika objektide kogumit, mitte aga üksikuid elemente: saadud tulemused iseloomustavad kogumit tervikuna.

Illustreerime toodud mõisteid tabeliga 1.

T a b e l 1.

Eesti NSV territoorium ja rahvaarv.

Majandusrajoon	Pindala tuh.km^2	Rahvaarv	Linnarahvastiku arv
Loode-Eesti	12,7	526 669	393 865
Kirde-Eesti	7,0	244 785	173 055
Kagu-Eesti	12,6	308 102	143 208
Edela-Eesti	8,7	156 839	80 848
Eesti saared	4,0	48 801	13 375

Kogumiks on siinkohal Eesti NSV. Kogumisse kuulub 5 elementi - majandusrajooni. Kogumi elemente on iseloomustatud kolme tunnusega.

Objekte ja nähtusi kirjeldavate arvuliste näitajate saamise protsessi nimetatakse statistiliseks vaatluseks.

Kogumi maht võib kõikuda mõnest elemendist lõpmatuseni. Näiteks klimaatiliste nähtuste uurimisel moodustab elemendi iga territooriumi punkt, iga punkti iseloomustab näiteks aastase sademete hulga kindel arvuline väärtus. Kogum on sel juhul lõpmata suure mahuga.

Et uurida ülaltoodud kogumit, oleks tarvis rajada vaatlusjaam igasse territooriumi punkti, mis on loomulikult võimatu. Andmeid kõigi 2000 NSV Liidu linna kohta on küll võimalik, kuid väga raske koguda ja töödelda. Statistilise materjali suure mahu juures kasutatakse tavaliselt v ä l j a v õ t t e i d (mahuga n). Väljavõte moodustab teatud viisil valitud osast kogumi elementidest, tema abil otsustatakse kogu kogumi üle. Väljavõtteks on näiteks praegune vaatlusjaamade võrk. Nähtavasti saab NSV Liidu linnade iseloomulikke omadusi leida, kui võtta vaatluse alla iga kümnes või isegi kahekümnes neist. Küllalt suur väljavõtte maht ja õige teostamine garanteerivad, et väljavõtte uurimisel saadakse peaaegu samad tulemused kui kogumi uurimisel. Väljavõtte r e p r e s e n t a t i i v s u s võimaldab tal asendada kogumit.

Objektide valik väljavõttesse peab rahuldama järgmist hädavajalikku tingimust: igal kogumi elemendil on võrdsed šansid sattuda väljavõttesse. Niisugune nõudmine välistab eelarvamusliku, subjektiivse lähenemise (näiteks teraviljade saagikuse määramisel koostada väljavõte ainult paremat majanditest).

Levinumateks väljavõtete tegemise viisideks on järgmised.

1. Juhuslik kordamisega valik. Oletame, et mingil alal on 2000 majandit (N = 2000). Uurija otsustas piirduda 30-st majandist koosneva väljavõttega (n = 30). Igale

majandile omistatakse kindel number, tema nimi kirjutatakse eraldi kaardile. Kõik 2000 kaarti segatakse hoolega. Valitakse huupi üks kaart, sellel märgitud majand läheb väljavõttesse. Nüüd pannakse kaart pakki tagasi ja segatakse uuesti. Seda protsessi korratakse 30 korda. Nii saadakse 30 majandit, mis kuuluvad väljavõttesse. Võib juhtuda, et üks majand tuleb välja korduvalt - siis ka tema tunnuste väärtused kirjutatakse üles sama arv kordi.

2. Juhuslik kordamiseta valik, mis erineb eelmisest vaid selle poolest, et kord väljatõmmatud kaarti ei panda pakki tagasi. Seetõttu ei saa ükski element esineda väljavõttes kaks korda.

3. Mehhaaniline valik seisneb selles, et kogumi elemendid nummerdatakse ja väljavõttesse võetakse näiteks 3-ga, 5-ga, 10-ga vms. jagunevate numbritega elemendid. Valitakse näiteks kõigist antud ala majanditest väljavõttesse iga kümnes.

4. Valik juhuslike arvude tabeli abil. Väljavõttesse lähevad need elemendid, mille järjekorranumber langeb kokku järjekordse juhuslike arvude tabelist võetud arvuga.

5. Seerialine valik. Uuritav kogum jagatakse osadeks (seeriateks) ja väljavõte tehakse igast seeriast või osast seeriatest mingi eeltoodud meetodiga. See viis on sobiv, kui uuritav kogum on väga ebaühtlane. Sel juhul püütakse seeriad koostada nii, et seeria piires kogumi elemendid ei oleks väga erinevad.

Praktikas kasutatakse ka kombineeritud väljavõtte tegemise meetodeid. Geograaf peab sõltuvalt konkreetsetest tingimustest leidma meetodi, mis kindlustab väljavõtte suurima representatiivsuse.

Kuna osa ei suuda ikkagi absoluutselt täpselt kirjeldada tervikut, siis erinevad kogumi omadused alati mingil määral väljavõtte omadustest. See erinevus, väljavõtte representatiivsusviga, sõltub eelkõige väljavõtte mahust n ,

mille suurendamine toob aga kaasa arvutus- ja mõõtmistöö-
de mahu suurenemise. Tuleb leida mõistlik tasakaal, mille
puhul väljavõtte viga ei oleks kuigi suur, väljavõtte maht
aga võimalikult väike. Niisuguse tasakaalu määramine kuu-
lub samuti statistika ülesannete hulka.

On tähtis täpselt määrata, mis on kogum, väljavõte,
element.

Tartu linna sidemete uurimisel oli vaja selgita-
da, milliste rajoonide haigeid teenindavad Tartu raviasu-
tused ja millises proportsioonis . Kõigi registreerimis-
kaartide läbivaatamine kõigis haiglates on praktiliselt
võimatu. Seepärast otsustati kasutada väljavõtete meeto-
dit.

Me vajame patsientide arvu teatud rajoonist. Kuna
statistilised meetodid annavad vaid kogumit iseloomusta-
vaid tulemusi, siis peab iga rajoon moodustama omaette ko-
gumi. Arvutused tuleb teha kõigi kogumite jaoks eraldi.

Mis on kogumi element? Selleks ei saa olla üksik pat-
sient. Kui palju me teda ka uuriksime, me ei saa teada se-
da, kui palju on patsiente. Elemendil ei oleks lihtsalt
tunnust, mis meid huvitab. Seega peab elemendiks olema
teatud patsientide grupp, mis koosneb muutlikust arvust
patsientidest. Osa neid gruppe tuleb valida väljavõttes-
se, teine osa mitte. Me taotleme sellist elemendi määrat-
lust, et meil ei tekiks raskusi gruppide piiritlemisega,
et ei jääks üle patsiente, keda ei saa paigutada kuhugi
gruppi (muidu oleks kogum suurem kui tema elementide sum-
ma, kogum koosneks elementidest ja veel millestki), et me
saaksime lihtsalt eraldada väljavõttesse kuuluvad elemen-
did ülejäänutest, kasutades alfabeetilist haigete regist-
reerimise kartoteeki ja et töötlemisele tulev kartoteegi-
kaartide arv oleks võimalikult väike.

Ei kõlba järgmine definitsioon:
element on patsientide grupp, kes viibis haiglas teatud
kindlal perioodil, näiteks jaanuarikuul. Võimalik, et pat-
sient saabus haiglasse näiteks 30. jaanuaril ja kirjutati

välja 5. veebruaril. Millisesse elementi peaks kuuluma see patsient?

Parem olgu elemendiks teatud külanõukogust, linnast või alevist pärit patsiendid. Piiritlemisraskusi siin ei teki. Kõikvõimalikud kaardid on jaotatud elementide vahel. Võiksime igas rajoonis osa külanõukogusid valida väljavõttesse ja töötada läbi ainult nende haigete kaardid, kes elunevad neis külanõukogudes. Ent me ei leia kõiki vajalikke kaarte kartoteegist muidu, kui läbi vaadates kõik vajalikud ja mittevajalikud kaardid.

Määrame, et elemendi moodustab patsientide grupp, kelle perekonnanimi algab kindla tähega ja kes loomulikult elavad momendil meid huvitavas rajoonis. Elemendi piiritlemine on jällegi selge, elementide eraldamine kartoteegis ei ole raske. Kuid selleks, et väljavõttesse satuks küllaldane arv elemente (umbes kümmekond), peame läbi töötama umbes poole kartoteegist. Määratlus on ebaefektiivne.

Sobivaks elemendi definitsiooniks osutub järgmine. Kogumi elemendiks on mingis külanõukogus elavate patsientide grupp, kelle perekonnanimi algab teatava tähega. Me näeme, et ka siin on elemendid rangelt piiritletud, samal ajal kuulub igasse elementi vähem kaarte kui eelmisel juhul ja küllaldase mahuga väljavõtte saamiseks on vaja läbi töötada vähem kaarte. Väljavõtte teostamise printsiip võib endiselt olla alfabeetiline: väljavõttesse kuuluvad kõik elemendid, mida moodustavad patsientide grupid, kelle perekonnanimi algab kindla tähega, sõltumata nende elukohast.

Elementi iseloomustavaks arvuliseks tunnuseks on temasse kuuluvate kaartide arv. Elementide arv väljavõttes võrdub valitud tähtede arvu ja antud rajooni külanõukogude, linnade ja alevite arvu korrutisega.

Praktiliselt on meil seega tarvis valida kartoteegist teatava tähega algava perekonnanimega patsientide kaardid. See ei valmista raskusi. Iga kaardi järgi saab määrata patsiendi elukoha. Esmalt sorteerime kaardid rajoonide järgi -

eraldame kogumid. Võtame siis näiteks Põlva rajooni kaardid ja sorteerime need nüüd külanõukogude järgi hunnikutesse. Saame patsientide grupid, kes elavad antud külanõukogus ja kelle perekonnanimi algab valitud tähega,s.t. elemendid. Sama protseduuri võib korrata kõigi tähtede jaoks, alustades iga tähega uusi hunnikuid. Siis me oleksime jaganud kõik kaardid esiteks kogumite vahel ja teiseks kogumi elementide vahel. Tegelikult me lõpetame kaartide sorteerimise siis, kui elementide arv igas kogumis on küllalt suur. Selleks momendiks sorteeritud kaartide grupid moodustavad väljavõtted, millest igaüks koosneb elementidest. Antud töös osutus küllaldaseks ühe tähe läbivaatamine. Seejärel loeme üle, mitu kaarti sattus igasse hunnikusse, mis kuulub näiteks Põlva rajooni kogumisse. Saadud arvud on väljavõtte elemente iseloomustavad suurused. Leiame nende alusel elemendi keskmise kaartide arvu, mille korrutame kogumi elementide arvuga, s. o. Põlva rajooni külanõukogude ja alevite arvu korrutisega alfabeedi tähtede arvuga. Saame kogumi mahu hinnangu, mille täpsuse me muidugi ka peame arvutama. Edasi kordame protseduuri Jõgeva, Tartu jne. rajoonide jaoks.

2. Rühmituste olemus ja liigid.

Andmete statistilisele töötlemisele eelneb sageli nende rühmitamine, s. t. statistilise kogumi jaotamine rühmadesse, mis mingi tunnuse poolest on homogeensed. Selleks on tarvis kogumi elementide hulgas leida homogeensed rühmad ja alles seejärel anda elementide üldistatud käsitlus. Näiteks uurides uhteorge, võime need rühmitada asendi alusel reljeefil: nõgudes, nõlvadel ja kõrgendike lagedel paiknevad uhteorud. Asustatud punktid võime rühmitada linnadeks ja maa-asulateks.

Geograafias on tähtis rühmitamine territoriaalse tunnuse järgi. Uhteorge võib rühmitada maastike kaupa, looduslike vööndite kaupa jne. Majanduslikud andmed rühmitatakse

tavaliselt administratiivsete ühikute lõikes.

Statistikas tarvitatakse siiski kõige sagedamini rühmitamist hulgalise tunnuse järgi. Vaatleme sellise rühmituse näidet. Tabelis 2 on antud 25 uhteoru pikkus meetrites.

Tabel 2.

Nr.	L	Nr.	L
1	35	14	21
2	42	15	37
3	38	16	28
4	24	17	17
5	31	18	59
6	52	19	43
7	43	20	36
8	26	21	45
9	39	22	22
10	54	23	36
11	36	24	41
12	48	25	38
13	33		

Leiame tabeli järgi pikima ja lühima uhteoru pikkused. Nendeks on vastavalt 59 ja 17 m. Nende arvudega on piiratud tunnuse muutumispiirkond. Jagame muutumispiirkonna võrdseiks vahemikeks, olgu neid vahemikke näiteks 5. Loeme ära, mitme uhteoru pikkused langevad mingisse vahemikku. Saadud arve nimetatakse vahemike sagedusteks (vt. tabel 3).

Tabelit, kus on loetaldud tunnuse väärtuste vahemikud ja antud nende vahemike sagedused, nimetatakse jaotuse intervallreaks e. variatsioonireaks.

Vahemike arv sõltub kogumi või väljavõtte mahust. See tuleb määrata nii, et ühte vahemikku ei jääks ei liiga palju ega liiga vähe elemente. Esimesel juhul lähevad kaotsi olulised erinevused vahemike piires, teisel juhul on aga saadud sagedused mitterepresentatiivsed. 100 - 500 elementi on soovitav jagada 8 - 16 vahemikku.

Mõnikord on otstarbekas jagada muutumispiirkond mittevõrdseiks vahemikeks, eriti kui mõnes vahemikus on vähe vaatlusi. Ent sel juhul raskeneb mitme arvutusskeemi kasutamine.

Tunnuse väärtuse jaotuse kirjeldamiseks kasutatakse laialt ka näitlikku graafilist meetodit (histogramm, jaotus-

polügoon, jaotuskõver. Neid graafikuid tohib joonistada ainult võrdsete vahemike korral.

Tõmbame koordinaatteljed. Abstsisseteljele kantakse vahemike piirid (meie näites uhteorgude pikkus), ordinaatteljele vastavad sagedused. Iga vahemiku jaoks ehitatakse ristkülik, mille alus võrdub vahemiku ulatusega, kõrgus aga sagedusega. Saadud kujundit nimetatakse h i s t o g r a m m i k s (vt. joon. 1).

T a b e l 3 .

Uhteorgude pikkuse vahemik	Sagedus
10 - 20	1
21 - 30	5
31 - 40	10
41 - 50	6
51 - 60	3

Samas teljestikus võime me iga vahemiku keskelt tõmmata abstsisseteljele ristsirge, millele kanname sagedusega võrdse lõigu. Saadud punktid ühendame sirglõikudega. Saame j a o t u s p o l ü g o o n i (polügoon = hulknurk) (vt. joon. 2, murdjoon).

Täpsema ettekujutuse statistilise materjali seaduspärasusest annab j a o t u s k õ v e r a t e koostamine. Nende täpsem joonistamine nõuab keerulisi arvestusi, ent geograafilises uurimistöös on tihti küllaldane tõmmata jaotuskõver silma järgi, polügooni teravaid nurki siludes.[1]

Jaotusgraafikute analüüs (nende maksimumide ja miinimumide otsimine, laugete ja järskude osade paiknemine) pakub uurijale suurt huvi. Joonistel 1 ja 2 kujutatud histogrammist ja polügoonist võib järeldada, et antud rajoonis on valdavad uhteorud pikkusega 30 - 40 m, lühikesi ja pikki uhteorgusid on üpris vähe. Sageli võrdleb geograaf jaotusgraafikuid ühe ja sama nähtuse kohta, kuid eri aladel. Kui jaotusgraafikud osutuvad sarnasteks

[1] Sel juhul saame diferentsiaalse jaotuskõvera. Allpool ongi juttu ainult diferentsiaalsetest jaotuskõveratest.

(mitte ilmtingimata kokkulangevateks), siis võib järeldada, et antud nähtus, meie näites erosioon, toimib mõlemal alal ühesugustes tingimustes.

3. Juhusliku suuruse mõiste. Tähtsamad statistilised jaotused ja nende tähendus.

Tavaliselt on statistilise kogumi elemente iseloomustavad arvulised väärtused erinevad. Selle erinevuse põhjused võivad olla mitmesugused. Näiteks on linnade erinev rahvaarv tingitud nende asendi, ajaloolise arengu jne. iseärasustest. Sageli osutub kasulikuks mitte süveneda erinevuste kõigisse põhjustesse, vaid pidada neid erinevusi juhuslikeks. Juhuslikkus tähendab siis neid põhjusi, mida me ei pea otstarbekaks täpsemalt identifitseerida. Kogumi elementide väärtused moodustavad siis *juhusliku suuruse* muutumispiirkonna.

Juhuslikuks suuruseks nimetatakse suurust (näiteks linnade elanike arv), mis mingi tõenäosusega p võib omandada kõik või teatud väärtused[1] muutumispiirkonnas.

Seega siis on eelmises paragrahvis kirjeldatud graafikud juhusliku suuruse graafiliseks esitusviisiks.

Kujutame ette praktiliselt piiramatu mahuga kogumit (näiteks Maakera täiskasvanud elanikkond), kus juhuslikuks suuruseks on inimese pikkus. Jagame juhusliku suuruse muutumispiirkonna mingil viisil vahemikkudeks, näiteks iga 10 cm järel. Ehitame jaotuspolügooni. Olgu kogumi maht N , aga vahemikku i langevate elementide arv m_i . Siis on vahemikku iseloomustav *suhteline sagedus*

i

[1] Aasta keskmine temperatuur mingis punktis võib olla 3°, aga ka 3,141° ja isegi π ° (iseasi, kas me nii täpselt mõõta suudame). Linna elanike arv aga ei saa olla 62 541,25 inimest. Esimest tüüpi suurust nimetatakse pidevaks, teist aga diskreetseks. Käesolevas õppevahendis ei ole tähtis, kas suurus on pidev või diskreetne.

$$\mu_i = \frac{m_i}{N} . \qquad (I.1)$$

Jagame iga vahemiku kaheks, siis jääb vahemiku pikkuseks 5 cm. Arvutame uuesti suhtelised sagedused ja ehitame jaotuspolügooni. Polügoon erineb eelmisest üsna vähe, olles ainult umbes 2 korda madalam. Muudame sellepärast mastaapi nii, et polügoon oleks sama kõrge kui eelmine.

Kordame vahemike poolitamist ja polügoonide ehitamist senikaua, kuni vahemiku pikkus muutub praktiliselt punktiks abstsissteljel. Siis saab polügoonist jaotuskõver, suhtelisest sagedusest aga juhusliku suuruse antud väärtuse t õ e n ä o s u s p .

Analüütilisest geomeetriast teame, et igale kõverale saab vastavusse seada funktsiooni. Sarnase kujuga, kuid erineva paigutusega kõverad erinevad selle funktsiooni konstantsete kordajate - parameetrite poolest. Näiteks igale sirgele vastab funktsioon kujuga $y = ax + b$. Sirged $y = 3x + 5$, $y = 5x + 3$ ja $y = 4x + (-6)$ erinevad vaid asendi poolest graafikul. Suurused a ja b on siinkohal parameetrid.

Jaotuskõverat kirjeldab analüütilises vormis selliste või teistsuguste parameetritega t õ e n ä o s u s t i h e d u s e f u n k t s i o o n .

Mõned tõenäosustiheduse funktsioonid on statistikas eriti tuntud. Juhuslike suuruste suhtes, mis jaotuvad vastavalt neile funktsioonidele, on tõestatud rida tähtsaid tulemusi. Kui uuritava juhusliku suuruse tõenäosustiheduse funktsioon on lähedane mõnele neist funktsioonidest, saab neid tulemusi vahetult kasutada. Teiselt poolt sõltub mitmete statistiliste valemite kuju sellest, kuidas on jaotunud juhuslik suurus, mille kohta neid valemeid rakendatakse. Sellepärast on otstarbekas teada tähtsamaid statistilisi jaotusi.

Kõige lihtsam on ühtlane jaotus. Sel juhul võtab juhuslik suurus muutumispiirkonnas võrdse tõenäosusega üks-

kõik millise võimaliku väärtuse. Ühtlase jaotuse graafikuks on abstsissteljega paralleelne sirglõik (vt. joon.3). Ühtlane jaotus tekib üksikult mõjuva juhusliku põhjuse toimel. Võtame näiteks juhusliku suuruse - täringul visatud silmade arvu.

Muutumispiirkond koosneb siin kuuest väärtusest: 1, 2, 3, 4, 5, 6. Kui täring on kvaliteetne (s. t. ei toimi teisi juhuslikke põhjusi), on iga silmade arvu väljatuleku tõenäosus 1/6.

Ühtlase jaotuse parameetriks on muutumispiirkonna ulatus.

Praktiliselt ei esine aga peaaegu kunagi olukorda, kus juhuslik suurus toimiks üksi. Väga tihti on tegemist suure arvu enam-vähem võrdse tähtsusega põhjustega. Igaüks neist tekitab ühtlase jaotuse, kõigi jaotuste summeerimisel saadakse n o r m a a l n e e. G a u s s i j a o t u s. Nimetus viitab asjaolule, et see jaotusetüüp on looduse ja ühiskonna nähtuste juures kõige enam levinud. Seepärast, kui ei ole teada midagi lähemat juhusliku suuruse jaotuse kohta, aga katsetingimused ei räägi vastu normaalse jaotuse tekkimise tingimustele, oletatakse, et jaotus on tõesti normaalsele lähedane, ja kasutatakse vastavaid valemeid.

Normaalne jaotusfunktsioon on liiga keeruline, et seda siin esitada. Märgime ainult jaotuskõvera iseloomulikke omadusi: 1) jaotuskõver on sümmeetriline; 2) mingi väärtuse (keskväärtuse) lähedased väärtused esinevad suurema tõenäosusega kui keskväärtusest kaugemad väärtused; 3) muutumispiirkond on lõpmatu, s. t. võib esineda ükskõik milliseid väärtusi, kuid kaugete väärtuste tõenäosus on väga väike (vt. joon. 4).

Mõnikord on tegemist ka kahe või enama jaotuse summaga. Toodud näites on Maakera täisealise elanikkonna pikkus juhuslik suurus, mille jaotus on kahe erinevate parameetritega normaalse jaotuse summa. Füüsiliselt vastab neile jaotustele kaks erinevat kogumit: mehed ja naised.

Ka mitmed geograafe huvitavad suurused jaotuvad konk-

reetse statistilise tõenäosustiheduse funktsiooni järgi. Mõõtmisvead jaotuvad peaaegu alati normaalselt. Klimaatilised ja hüdroloogilised näitajad jaotuvad aastate lõikes normaalsele lähedaselt. Tihedas asulate võrgus jaotub kaugus lähima asulani üsna keerulise gammajaotuse järgi.

II. NÄHTUSE KESKVÄÄRTUS JA MUUTLIKKUSE NÄITAJA.

1. Keskväärtus.

Statistilise kogumi lühikese, üldistatud kirjeldusena kasutatakse kõige enam keskväärtust. Geograafil tuleb sageli arvutada ja kasutada keskmist õhutemperatuuri, keskmist kõrgust, põllukultuuride keskmist saagikust jms. Keskväärtuses hävivad, kustutavad üksteist vastastikku arvuliste tunnuste juhuslikud kõikumised, nähtuste ja objektide individuaalsed erinevused tasanduvad. Teiste sõnadega, keskväärtuse arvutamine on juhuslikkusest loobumine ja üldise, olulise, kogumit tervikuna iseloomustava väljatoomine.

Kõige levinumaks keskväärtuseks on aritmeetiline keskmine, mida arvutatakse valemiga:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} . \qquad (II.1)$$

$\bar{x}$ on aritmeetilise keskmise, x_i - tunnuse i-nda väärtuse, n - vaatluste arvu standardtähistus. (Märk $\sum_{i=1}^{n}$ tähendab, et tuleb liita kõik järgneva avaldise väärtused, mille

indeks on vahemikus 1 kuni n . Kui indeksi muutumise vahemik on silmanähtav, nagu käesolevas valemis, võib tarvitada lihtsustatud tähistust $\sum_i$ või isegi $\sum$.)

Arvutame aritmeetilise keskmise lihtsast näitest. Jõe vooluhulk on mõõdetud iga 2 kuu tagant. Tulemuseks on arvud: 130, 280, 180, 140, 190, 150 m/sek. Leiame aasta keskmise vooluhulga.

$$\bar{x} = \frac{130+280+140+180+190+150}{6} = \frac{1070}{6} \approx 178{,}7 \text{ m/sek.}$$

Arvutuste õigsuse kontrolliks leiame summa:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \frac{n \sum x_i}{n} - n\bar{x} = n\bar{x} - n\bar{x} = 0 .$$

Ja tõesti:

$$\begin{array}{r} 178{,}7 - 130 \\ 178{,}7 - 280 \\ 178{,}7 - 180 \\ + \; 178{,}7 - 140 \\ 178{,}7 - 190 \\ 178{,}7 - 150 \\ \hline 1070 \quad - 1070 = 0 . \end{array}$$

Arvutused valemi (II.1) järgi muutuvad mahukaks, kui on tegemist suure arvu vaatlustega. Võtame kas või toodud näite uhteorgude pikkusest (tabel 2). Kas ei ole võimalust arvutusi lihtsustada?

Meil on olemas jaotuse intervallrida, s. t. vahemike keskpunktid ja sagedused. Kui asendame kõik samasse vahemikku langevad uhteorgude pikkused vahemiku keskmise pikkusega, ei tee me suurt viga. Kui vahemikku k langeb m_k uhteoru pikkus, aga vahemiku keskpunkt on x_k meetrit, on kõigi vahemikku k langenud uhteorgude pikkuste summa $m_k \cdot x_k$ meetrit. Liidame nüüd "vahemiku summad". On selge, et uhteorgude üldarv võrdub sageduste summaga. Saame

$$\bar{x} = \frac{\sum_k m_k x_k}{\sum_k m_k} . \qquad \text{(II.2)}$$

See on kaalutud aritmeetilise keskmise valem. Kaaludeks on siinkohal sagedused m_k . Arvutame nüüd tabeli 2 andmete järgi uhteorgude keskmise pikkuse. Sagedused võtame tabelist 3. Saame

$$\bar{x} = \frac{1.15 + 5.25 + 10.35 + 6.45 + 3.55}{1 + 5 + 10 + 6 + 3} = 37 \text{ m}.$$

Võrdluseks leiame $\bar{x}$ ka valemi (II.1) järgi. Tulemuseks on $\bar{x} = 36{,}96$ m, seega viga on vaid $0{,}04 : 37 \approx 0{,}1$ %.

2. Keskväärtuste tähendus.

Keskväärtused lubavad:

1) määrata nähtuste arengutendentsi,

2) hinnata üksiku väärtuse suurust võrreldes keskväärtusega,

3) kindlaks teha seose olemasolu kahe nähtuse vahel nende keskväärtuste võrdlemisel eri kogumites.

1) Vaatleme tabelit 4, kus on toodud juuli keskmised temperatuurid Leningradis. Nende andmete põhjal ei saa mingit järeldust teha. Kui aga heita pilk tabelisse 5, kus juuli keskmised on võetud 50 aasta kohta, selgub kohe seaduspärasus - aasta-aastalt muutub kliima soojemaks umbes 0,05° aastas.

T a b e l 4.

Aasta	Juuli keskmine temperatuur
1900	16,3
1901	19,3
1902	14,9
1903	16,5
1904	14,3
1905	16,9

T a b e l 5.

Aastad	Juuli keskmine temperatuur
1876-1925	17,4
1877-1926	17,4
1878-1927	17,5
1879-1928	17,5
1880-1929	17,6
1881-1930	17,6

Tabel 6.

Rajooni nr.	Huumusesisaldus %	Saagikus ts/ha
1	3	6
2	4	8
3	5	9
4	7	10

2) Nelja uhteoru kasvukiiruseks karjatamiseks kasutataval kõrbealal on vastavalt 5, 6, 8 ja 9 m/aastas. Kuivõrd karjatamine soodustab erosiooni, näitab asjaolu, et uhteorgude kasvu keskmine kiirus kogu selles piirkonnas on 2 m aastas.

3) Tabelis 6 on toodud mulla keskmine huumusesisaldus $\bar{x}$ ja teraviljade saagikus $\bar{y}$ neljas rajoonis. Mida rohkem huumust, seda suurem saagikus. Kui me ei oleks võtnud keskmisi, vaid võrrelnud näitajaid x ja y iga põllutüki kohta, vaevalt oleks see seaduspärasus kuigi selgesti ilmnenud. Sõltub ju saagikus mitte üksnes mulla huumusesisaldusest, vaid ka kliimast, veerežiimist, harimisest, väetiste kasutamisest jne.

Keskmisi näitajaid kasutatakse laialdaselt geograafias. Kliima iseärasused on tabatavad vaid ilmastikunähtuste paljuaastaste keskmiste kaudu. Ka põllumajanduskultuuride saagikus ja tööstustoodangu maht kapitalistlikes maades tuleb objektiivsem, kui see arvutada mitme aasta keskmise näitajana.

Mitte just harva leiavad rakendamist ka keskmiste keskmised. Näiteks võib summeerida paljude aastate temperatuurikeskmisi kõigis meteojaamades ja jagada vaatluspunktide arvuga. Nii võib arvutada näiteks Eesti NSV jaanuarikuu keskmise temperatuuri. Keskväärtuse võtmine üle mingi ajalõigu ja seejärel veel üle mingi territooriumi ongi levinuim viis sellise "teist järku" keskmise leidmiseks. Muidugi on võimalik arvutada ka kolmandat, neljandat jne. järku keskmisi.

3. Limiidid ja muutumisulatus.

Aritmeetiline keskmine annab meile vaid tunnuse tüüpilisema väärtuse või tema väärtuste raskuskeskme. Teiseks oluliseks probleemiks on muutlikkuse (variaabluse) leidmine.

Koosnegu üks statistiline kogum elementidest väärtusega 1, 3, 5, 7, 9, teine elementidest väärtusega 3, 4, 5, 6, 7. Aritmeetiline keskmine võrdub mõlemal juhul viiega. On aga selge, et need kogumid on erinevad. Erinev on just nende kõikumise ulatus, v a r i a a b l u s . Püüame leida variaabluse mõõtu.

Kõige lihtsamaks mõõduks on l i m i i d i d - suurim ja vähim väärtus. Geograafias on see üsna levinud viis. Tsiteerigem Tselinnõi krai atlast (1964): "Aasta päikesepaiste kestus on siin 2000 - 2500 tundi. Ööpäeva päikese radiatsioon juulis on krai põhjaosas 550, lõunas 600 cal/cm^2... Huumuse ja lämmastiku tagavarad poolemeetrises mullakihis kõiguvad vastavalt 350 ning 400 t ja 23 ning 25 t vahel hektari kohta". Limiidid võimaldavad hinnata umbkaudu ka keskväärtust, mis tõenäoliselt on kuskil pooles vahemikus. Ka meie kahe kogumi erinevuse toovad limiidid välja. Limiitide vahe - suuruse m u u t u m i s e u l a t u s - lisatakse samuti juurde.

Limiidid arvestavad ainult kogumi kahte väärtust, mis pealegi võivad olla juhuslikud. Ülejäänud kogumi elemendid ei mõju kuidagi sellele variaablusenäitajale. Toome jällegi näite kahest kogumist:

1) 1, 9, 10, 11, 12, 13, 60. $\bar{x} = 18.$
2) 1, 2, 3, 5, 10, 45, 60. $\bar{x} = 18.$

On silmanähtav, et nende kogumite elementide variaablus on erinev, kuid ei limiidid ega aritmeetiline keskmine ei näita seda erinevust.

4. Muutlikkuse näitajad.

Arvestades, et keskväärtus väljendab seda üldist, mis on omane kogumile ja et me soovime leida hajuvuse näitajat, mis võtaks arvesse kõiki kogumi elemente, võiks selliseks näitajaks olla elementide keskmine kaugus aritmeetilisest keskmisest:

$$\frac{\sum (x_i - \bar{x})}{n} .$$

Nagu me aga esimeses paragrahvis nägime, võrdub see avaldis nulliga igasugustes kogumites ja väljavõtetes. Positiivsed ja negatiivsed vahed kustutavad üksteist vastastikku. Et sellest raskusest üle saada, võib kasutada kahte teed: kas võtta vahede absoluutväärtused või tõsta nad ruutu.

K e s k m i n e a b s o l u u t n e h ä l v e on aritmeetiline keskmine kogumi (väljavõtte) elementide x_i ja kogumi (väljavõtte) aritmeetilise keskmise $\bar{x}$ vahede absoluutväärtustest:

$$\theta = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)}{n} . \qquad \text{(II.3)}$$

Eelmises paragrahvis näitena toodud kogumites $\theta_1 = 13{,}4$ ja $\theta_2 = 21{,}1$.

Sagedamini kasutatakse aga teist viisi - vahede ruutu tõstmist, kuna nii saadud näitajad sisalduvad ka paljudes teistes statistilistes tulemustes, kuuluvad mitme jaotuse parameetrite hulka ja võimaldavad valemite edasiarendamist.

Juhusliku suuruse k e s k m i n e r u u t h ä l v e avaldub valemiga

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}} \qquad \text{(II.4)}$$

Siin N , nagu ikka, tähistab kogumi mahtu. Väljavõtte põhjal ei saa keskmist ruuthälvet nii arvutada, sest väljavõttest saadud aritmeetiline keskmine ei ole täiesti täpne. Sellepärast räägitakse väljavõtte puhul s t a n d a r d h ä l b e s t , mille valemis juurealuse avaldise nimetaja on vea kompenseerimiseks muudetud. Standardhälve

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} . \qquad (II.5)$$

Küllalt suure väljavõtte mahu n juures erinevad suurused 1/n ja 1 : (n-1) väga vähe, näiteks kui n = 20, siis 1/n ja $\frac{1}{n-1}$ vahe on kõigest $\frac{1}{380}$. Niisiis erinevus standardhälbe ja keskmise ruuthälbe vahel muutub järjest väiksemaks. Seepärast, kui n > 20, ei tehta nende kahe termini vahel enam vahet ja standardhälve arvutatakse keskmise ruuthälbe valemi järgi. Geograafias aga esineb vahel ka väikese mahuga väljavõtteid ja sellistel juhtudel on kasulik meeles pidada valemite II.4 ja II.5 erinevust.

Suurust σ^2 nimetatakse juhusliku suuruse d i s p e r s i o o n i k s .

Meie näites

$$\sigma_1 = \frac{4216}{7} = \pm 24{,}5 ;$$

$$\sigma_2 = \frac{5764}{7} = \pm 28{,}7 .$$

Nagu nägime, on keskmine ruuthälve tavaliselt veidi suurem kui keskmine absoluutne hälve.

Rühmitatud andmete korral omandab keskmise ruuthälbe valem kuju:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} m_i (x_i - \bar{x})^2}{\sum m_i}} . \qquad (II.6)$$

Soovitame lugejal tuletada see valem iseseisvalt.

Arvutame σ tabeli 3 andmete põhjal. Arvutused koondame tabelisse 7.

Klimaatiliste, mullastikuliste, majanduslike näitajate keskmised ruuthälbed on kindlate territooriumide ja ajalõikude jaoks rangelt seaduspärased. Kahjuks kasutavad geograafid tunnuste variaabluse uurimisel liiga harva näitajat σ (või s), piirdudes limiitidega.

Tabel 7.

Uhteorgude pikkusvahemik m	Vahemiku keskpunkt x_i	Uhteorgude arv m_i	$m_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$m_i (x_i - \bar{x})^2$
10–20	15	1	15	-22	484	484
20–30	25	5	125	-12	144	720
30–40	35	10	350	- 2	4	40
40–50	45	6	270	8	64	384
50–60	55	3	165	18	324	972
K o k k u		25	925			2600

$$\bar{x} = \frac{925}{25} = 37 \text{ m} ; \qquad \sigma = \sqrt{\frac{2600}{25}} = \sqrt{104} = 10{,}2 \text{ m} .$$

$$\sigma^2 = 104 .$$

5. Variatsioonikordaja.

Keskmine ruuthälve on tunnuse muutlikkuse nimega mõõt. Seega ei saa teda kasutada variaabluse võrdlemiseks, kui tunnused on mõõdetud erinevates ühikutes. Iseloomustagu looduslike tingimuste variaablust ühes ja samas rajoonis tabelis 8 toodud arvud.

Ei ole võimalik kindlaks teha, milline tunnus on antud alal kõige muutlikum, sest ei saa võrrelda kraade meetrite-

ga. Erisuguste tunnuste variaabluse võrdlemiseks kasutatakse uut näitajat - v a r i a t s i o o n i k o r d a j a t .

T a b e l 8 .

T u n n u s e d	
1. Kõrgus merepinnast	100 m
2. Aasta keskmine õhutemperatuur	2 kraadi
3. Aasta sademete hulk	50 mm

Variatsioonikordaja V on keskmise ruuthälbe suhe aritmeetilisse keskmisse:

$$V = \frac{\sigma}{\bar{x}} . \qquad (II.7)$$

Tavaliselt väljendatakse variatsioonikordaja protsentides.

T a b e l 9 .

T u n n u s e d	$\bar{x}$	σ	V
1. Kõrgus merepinnast	200 m	100 m	50 %
2. Aasta keskmine õhutemperatuur	8	2	25 %
3. Aasta sademete hulk	500 mm	50 mm	10 %

Tabeli 9 andmete põhjal võib järeldada, et kõige muutlikum on kõrgus merepinnast, kõige vähem muutlik aga sademete hulk.[1]

---

[1] Variatsioonikordaja kasutamisega peab olema ettevaatlik. Toodud näites ei tohi variatsioonikordajat kasutada, kui aasta keskmine temperatuur on 0° või negatiivne. Negatiivne variatsioonikordaja on võrreldamatu positiivsega.

6. $\bar{x}$ ja σ arvutamise lihtsustatud meetodid.

Aritmeetiliste keskmiste ja keskmiste ruuthälvete arvutamine lihtsustub, kui me kasutame nende näitajate järgmisi omadusi (soovitame lugejal iseseisvalt tõestada omadused 1), 2) ja 3)):

1) Kui kõik sagedused korrutada (jagada) ühe ja sama arvuga b (1/b), siis $\bar{x}$ ja σ ei muutu.

2) Kui tunnuse kõik väärtused korrutada (jagada) ühe ja sama arvuga b (1/b), siis tuleb $\bar{x}$ korrutada (jagada) sama arvuga, σ aga tema absoluutväärtusega:

$$\frac{\sum bx_i m_i}{\sum m_i} = b\bar{x} ;$$

$$\sqrt{\frac{\sum (bx_i - b\bar{x})^2 m_i}{\sum m_i}} = \sigma |b| .$$

3) Kui kõigile tunnuse väärtustele liita (lahutada) üks ja sama arv b (-b), siis $\bar{x}$ suureneb (väheneb) selle arvu võrra, aga σ jääb muutmata:

$$\frac{\sum m_i (x_i + b)}{\sum m_i} = \bar{x} + b$$

$$\sqrt{\frac{\sum (x_i + b - (b+\bar{x})^2 m_i}{\sum m_i}} = \sigma .$$

4) Dispersioon võrdub ruudu keskmisega miinus keskmise ruut:

$$\sigma^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2 . \qquad \text{(II.8)}$$

Tõestame viimase omaduse.

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \frac{\sum 2\bar{x}x_i}{n} + \frac{\sum \bar{x}^2}{n} .$$

Kirjutame kaks viimast summat teisiti:

$$\sum 2\bar{x}x_i = 2\bar{x}x_1 + 2\bar{x}x_2 + \ldots + 2\bar{x}x_n = 2\bar{x}(x_1 + x_2 + \ldots + x_n) =$$

$$= 2\bar{x} \sum x_i$$

$$\sum_{i=1}^{n} \bar{x}^2 = \underbrace{\bar{x}^2 + \bar{x}^2 + \ldots + \bar{x}^2}_{n \text{ korda}} = n\bar{x}^2 .$$

Seega siis

$$\sigma^2 = \frac{\sum x_i^2}{n} - \frac{2\bar{x} \sum x_i}{n} + \frac{n(\bar{x})^2}{n} = \frac{\sum x_i^2}{n} - 2(\bar{x})^2 + (\bar{x})^2 =$$

$$= \frac{\sum x_i^2}{n} - (\bar{x})^2 .$$

Rühmitatud näitajate jaoks saame:

$$\sigma^2 = \frac{\sum m_i x_i^2}{\sum m_i} - (\bar{x})^2 . \qquad \text{(II.9)}$$

Valemite II.8 ja II.9 puhul ei ole tarvis ruutu tõsta vahesid $x_i - \bar{x}$, mis tihti tulevad murdarvud. Toome näite (tabelid 10 ja 11).

T a b e l 10.

Esimene viis.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-0,75	0,5625
5	2,25	5,0625
1	-1,75	3,0625
3	0,25	0,0625
11	0	8,7500

T a b e l 11.

Teine viis.

x_i	x_i^2
2	4
5	25
1	1
3	9
11	39

$$\bar{x} = \frac{11}{4} = 2{,}75 \ ;$$

$$\sigma = \sqrt{\frac{8{,}75}{4}} = \sqrt{2{,}19} \approx 1{,}5; \qquad \sigma = \sqrt{\frac{39}{4} - \left(\frac{11}{4}\right)^2} = \sqrt{9{,}75 - 7{,}56} =$$

$$= \sqrt{2{,}19} \approx 1{,}5 \ .$$

Kasutame nüüd keskmise ja variaabluse leidmisel kõiki nelja omadust. Andmed on toodud tabelis 12; seal on teostatud ka arvutused.

T a b e l 12 .

Sademete hulga vahemik mm	Vaatlusjaamade arv m_i	Vahemiku keskpunkt x_i	m_i'	x_i'	$m_i'x_i'$	$m_i'x_i'^2$
400 - 500	5	450	1	-2	-2	4
501 - 600	10	550	2	-1	-2	2
601 - 700	20	650	4	0	0	0
701 - 800	15	750	3	1	3	3
K o k k u	50		10		-1	9

Kuna sagedused m_i jagunevad 5-ga, siis, kasutades omadust 1), asendame nad sagedustega $m_i' = m_i : 5$. Lihtsustame ka tunnuse väärtused x_i, kasutades valemit

$$x_i' = \frac{x_i - c}{t} \ , \qquad \text{(II.10)}$$

kus t on vahemiku ulatus, c aga vabalt valitud punkt. Meie näites on sobiv võtta $c = 650$; $t = 100$.

Kui võtta c väärtuseks mingi vahemiku keskpunkt, siis saab uued väärtused x_i' leida isegi arvutamata. Nimelt vastab sellele vahemikule $x_i' = 0$, tunnuse väiksema-

te väärtuste suunas tulevad järjekorras -1, -2 jne., suuremate väärtuste suunas 1, 2, ...

Nüüd arvutame vajalikud näitajad uute, lihtsustatud andmete põhjal. Kuuendasse tulpa kirjutame korrutised $m_i'x_i'$, seitsmendasse tulpa $-m_i'x_i'^2$. Arvutame vajalikud summad. Edasi saame:

$$\bar{x}' = \frac{-1}{10} = -0,1 ;$$

$$\sigma' = \sqrt{\frac{9}{10} - (-0,1)^2} = 0,94 .$$

Nüüd, kasutades omadusi 2) ja 3), leiame:

$$\bar{x} = \bar{x}' \cdot t + c \qquad \text{(II.11)}$$

$$\sigma = \sigma' \cdot t \qquad \text{(II.12)}$$

$$\bar{x} = -0,1 \cdot 100 + 650 = 640 \text{ m}$$

$$\sigma = 0,94 \cdot 100 = 94 \text{ m} .$$

III. KORRELATSIOONITEOORIA ALUSED.

1. Korrelatsiooni mõiste.

Dialektiline lähenemine looduse ja ühiskonna probleemidele nõuab protsesside ja nähtuste uurimist nende vastastikustes seostes.

Geograafilise keskkonna nähtused sõltuvad paljudest, sageli muutuvatest põhjustest, mis on tihti teadmata. Neid seoseid aitab leida ja uurida korrelatsiooniteooria - uuri-

ja jaoks äärmiselt oluline matemaatilise statistika osa.

Kõik sidemed jagunevad funktsionaalseteks ja korrelatiivseteks.

F u n k t s i o n a a l n e s õ l t u v u s eeldab suuruste ranget vastavust: ühe suuruse - argumendi mingile väärtusele vastab teise suuruse - funktsiooni kindlalt määratud väärtus (harva mitu väärtust). Funktsionaalse sõltuvuse graafilisel kujutamisel, kui paigutada ühele teljele argument, teisele funktsiooni väärtus, saame mingi joone, sirge või kõvera. Funktsionaalsed sõltuvused esinevad matemaatilistes abstraktsioonides, näiteks ringi pindala sõltuvus ringi raadiusest: $S = \pi R^2$.

Katsete puhul on meil enamasti tegemist k o r r e l a t i i v s e s e o s e g a , kus ühe suuruse mingile väärtusele võib vastata mitu teise suuruse väärtust, mis pealegi ei ole kindlalt määratavad. Põhjuseks on siin kõrvaliste faktorite olemasolu. Näiteks sõltub kultuuride saagikus mitte üksnes mulla viljakusest, vaid ka ilmast, majanduslikest faktoritest ja isegi majandi juhtimisest. Kui saagikuse seost mingi ühe faktoriga kujutada graafiliselt, ei saa me joont, vaid laialivalguva punktide hulga (vt. joonis 5), nn. korrelatsiooniellipsi.

2. Seose tiheduse mõõtmine.

Korrelatsiooniteooria aluseks on seose tiheduse mõiste. Kuna see mõiste geograafilises kirjanduses harva esineb, selgitame seda graafiliselt, ehitades k o r r e l a t s i o o n i v ä l j a . Kujutame iga vaatluste paari punktiga koordinaatsüsteemis. Konkreetsuse mõttes paigutame x-teljele rajooni hüdrotermilise koefitsiendi, y-teljele nisu keskmise saagikuse selles rajoonis. Kui me niiviisi kõik vaatlused üles märgime, saame rea punkte, millele võib ümber tõmmata ellipsikujulise joone. Mida ümaram on see ellips, seda nõrgem on seos nähtuste vahel ja vastupidi. Seose täieliku puu-

dumise korral saame ellipsi asemel ringi, funktsionaalse sõltuvuse korral tõmbub ellips kokku jooneks.

Joonisel 6 on kujutatud kaks korrelatsioonivälja. Väli a näitab uhteorgude kasvukiiruse y sõltuvust basseini suurusest x_1, väli b - kaldenurgast x_2. Kuna välja b punktid koonduvad ümaraks ellipsiks, võime öelda, et kaldenurga mõju kasvukiirusele on vähem tähtis. Ellipsi pikitelje suund NO-SW näitab, et seos on positiivne, s. t. kaldenurga suurenemine toob kaasa ka uhteorgude kasvukiiruse suurenemise. Negatiivse seose korral on ellipsi pikitelg suunatud „loodest kagusse".

Graafiline seose tiheduse määramine on küll näitlik, ent ebatäpne ja aeganõudev. Matemaatiline statistika lubab arvutada seose tiheduse arvulise näitaja - k o r r e l a t s i o o n i k o r d a j a .

L i n e a a r s e k o r r e l a t s i o o n i k o r d a j a r_{xy} mõõdab lineaarse seose tihedust kahe tunnuse x ja y vahel. Lineaarne tähendab siin, et otsitavat seost võib väga ligikaudselt väljendada lineaarse funktsiooniga $y = ax + b$. Lineaarse seose puudumine ei tarvitse kaugeltki tähendada mingi keerulisemat laadi seose puudumist.

Alati $-1 \leqslant r \leqslant +1$. Kui $r = 0$, puudub täielikult lineaarne sõltuvus nähtuste vahel. Kui $r = 1$ või $r = -1$, on sõltuvus funktsionaalne ja seost võib täpselt väljendada funktsiooniga $y = ax + b$. Korrelatsioonikordaja positiivsed väärtused tähendavad, et x suurenemisel reeglina y suureneb, negatiivsed väärtused - et x suurenemisel y väheneb.

Geograafid teevad tihti mitte just absoluutselt täpse järelduse: kui tunnuste x ja y vaheline korrelatsioonikordaja on r, siis tunnuse x muutlikkus selgitab $r^2 \cdot 100$ % tunnuse y muutlikkusest.

Korrelatsioonikordaja arvutatakse valemist

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} . \qquad \text{(III.1)}$$

Arvutame näiteks korrelatsiooni jaanuarikuu sademete hulga vahel Taškendis (x) ja 12 km kaugusel linnast asuvas vaatlusjaamas (y). Et vaatluspunktid on lähedal, võib arvata, et korrelatsioonikordaja tuleb suur. Kasutada on 9 aasta andmed (vt. tabel 13).

Tabel 13.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
87	86	+34,2	+29,1	995,22	1169,64	846,81
47	56	- 5,8	- 0,9	5,22	33,64	0,81
74	84	+21,2	+27,1	574,52	449,44	734,41
86	72	+33,2	+15,1	501,32	1102,24	228,01
38	47	-14,8	- 9,9	146,52	219,04	98,01
15	17	-37,8	-39,9	1508,22	1428,84	1592,01
41	43	-11,8	-13,9	164,02	139,24	193,21
8	19	-44,8	-37,9	1697,92	2007,04	1436,41
79	88	+26,2	+31,1	814,82	686,44	967,21
475	512	- 0,2	- 0,1	6407,78	7235,56	6096,89

$$\bar{x} = \frac{475}{9} = 52{,}8 \; ; \qquad \bar{y} = \frac{512}{9} = 56{,}9$$

$$r_{xy} = \frac{\frac{1}{9} \cdot 6407{,}78}{\sqrt{\frac{7235{,}56}{9}} \cdot \sqrt{\frac{6096{,}89}{9}}} = 0{,}96 \; .$$

$$r^2 = 0{,}92 \; .$$

Seega on sademete hulga vahel neis kahes punktis tõepoolest tihe seos. Võib öelda, et 92 % ulatuses on sademete hulka mõjutavad põhjused mõlemas punktis samad.

3. Korrelatsioonikordaja lihtsustatud arvutusviisid.

Tabeli 13 lahtrite täitmine on väga tülikas. Püüame teisendada korrelatsioonikordaja valemi lihtsamale kujule. Selleks teisendame valemi III.1 lugejat:

$$\frac{1}{n}\sum(x_i - \bar{x})(y_i-\bar{y}) = \frac{\sum x_i y_i}{n} - \frac{\sum \bar{x}y_i}{n} - \frac{\sum \bar{y}x_i}{n} + \frac{\sum \bar{x}\bar{y}}{n} =$$

$$= \frac{\sum x_i y_i}{n} - \bar{x}\frac{\sum y_i}{n} - \bar{y}\frac{\sum x_i}{n} + \frac{n\bar{x}\bar{y}}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y} =$$

$$= \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} .$$

Seega

$$r_{xy} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} . \qquad \text{(III.2)}$$

Nagu näitab tabel 14, tuleb selle skeemi järgi arvutada vähem suurusi (andmed on samad).

T a b e l 14.

x^2	x	xy	y	y^2
7569	87	7482	86	7396
2209	47	2632	56	3136
5476	74	6216	84	7056
7396	86	6192	72	5184
1444	38	1785	47	2209
225	15	255	17	289
1681	41	1763	43	1849
64	8	152	19	361
6241	79	6952	88	7744
32 305	475	33 430	512	35 224

$$\bar{x} = \frac{475}{9} = 52{,}8 \; ; \qquad \bar{y} = \frac{512}{9} = 56{,}9 \; ;$$

$$\sigma_x = \sqrt{\frac{32305}{9} - 52{,}8^2} = 28{,}3 \; ;$$

$$\sigma_y = \sqrt{\frac{35224}{9} - 56{,}9^2} = 26{,}0 \; ;$$

$$r_{xy} = \frac{\frac{33430}{9} - 52{,}8 \cdot 56{,}9}{28{,}3 \cdot 26{,}0} = 0{,}96 \; .$$

On lihtne tõestada, et korrelatsioonikordaja suurus ei muutu, kui ühe tunnuse kõiki väärtusi korrutada ühe ja sama arvuga või liita neile üks ja sama arv. Samuti võib algandmeid ümardada, võttes tingimuseks, et muutumisulatus peab olema vähemalt 7 - 8 säilitatava viimase koha ühikut. Meie näites võime ümardada algandmed täiskümneteks, jagada saadud arvud 10-ga (s. t. korrutada 0,1-ga) ja liita nüüd kõigile väärtustele -5, et saada umbes võrdne arv positiivseid ja negatiivseid sissekandeid (vt. tabeleid 15, 16, 17). Saame $r = 0{,}95$. Tühine erinevus on tekkinud ümardamise arvelt.

T a b e l 15 .

x teisendamine.

x	$\frac{x}{10}$	$x' = \frac{x}{10} - 5$
87	9	4
47	5	0
74	7	2
86	9	4
38	4	-1
15	1	-4
41	4	-1
8	1	-4
79	8	3

T a b e l 16.

y teisendamine.

y	$\frac{y}{10}$	$y' = \frac{y}{10} - 5$
86	9	4
56	6	1
84	8	3
72	7	2
47	5	0
17	2	-3
43	4	-1
19	2	-3
88	9	4

Tabel 17.

x'^2	x'	$x'y'$	y'	y'^2
16	4	16	4	16
0	0	0	1	1
4	2	6	3	9
16	4	8	2	4
1	-1	0	0	0
16	-4	12	-3	9
1	-1	1	-1	1
16	-4	12	-3	9
9	3	12	4	16
79	3	67	7	65

$$\bar{x}' = \frac{3}{9}\ ; \quad \bar{y}' = \frac{7}{9}\ ; \quad \sigma'_x = \frac{\sqrt{702}}{9}\ ; \quad \sigma'_y = \frac{\sqrt{536}}{9}\ ;$$

$$r_{xy} = \frac{\frac{67}{9} - \frac{3}{9}\cdot\frac{7}{9}}{\frac{\sqrt{702}}{9}\cdot\frac{\sqrt{536}}{9}} = 0{,}95\ .$$

4. Spearmani astakkorrelatsioon.

Mõnikord ei saa geograafi huvitavaid tunnuseid mõõta arvudes või on see mõõtmine liiga töömahukas. Seose tihedust saab aga määrata ka sellistel juhtudel, on ainult tarvis, et tunnuste väärtusi oleks võimalik reastada kasvavas või kahanevas järjekorras. Olgu üheks tunnuseks maahindepall x, teiseks kartuli saagikus y. Tunnused on määratud maatükkidel A, B, C, D, E. Maahindepallide kahanemise järjekord on: 1. A; 2. B; 3. C; 4. D; 5. E. Kartuli saagikuse kahanemise järjekord on: 1. A; 2. C; 3. D; 4. B; 5. E. Kirjutame ühele maatükile vastavad järjekorranumbrid tabelisse 18 kohakuti.

Tabel 18.

Maatükk	Järjekorranumber x järgi	Järjekorranumber y järgi	x'-y'	$(x'-y')^2$
	x'	y'		
A	1	1	0	0
B	2	4	-2	4
C	3	2	1	1
D	4	3	1	1
E	5	5	0	0
			Kokku	6

Tabeli teises ja kolmandas veerus on vastavalt tunnuste x ja y astakud, mida me võime vaadelda kui tunnuste väärtusi ja arvutada korrelatsiooni tunnuste vahel, kasutades Spearmani astakkorrelatsiooni valemit:

$$\varrho = 1 - \frac{6 \cdot \sum (x'-y')^2}{n^3 - n} . \qquad \text{(III.3)}$$

Meie näites

$$\varrho = 1 - \frac{6 \cdot 6}{125 - 5} = 1 - \frac{36}{120} = 0{,}7 .$$

Astakkorrelatsiooni võib kasutada ka siis, kui me vajame vaid seose tiheduse ligikaudset suurust, ehkki arvulised andmed on meil olemas.

Tuleb märkida, et ϱ väärtused alla 0,6 ei ole statistiliselt olulised, s. t. nad ei peegelda enam korralikult erinevusi seose tiheduses.

5. Nähtuste sõltuvuse empiiriliste valemite saamine.

Me juba märkisime, et lineaarse korrelatsiooni kordaja väärtus 1 või -1 tähendab, et tunnuste väärtuste x ja y vahel eksisteerib funktsionaalne sõltuvus $y = ax+b$. Kui $r \neq 1$, siis ei ole selline valem enam täpne, vaid ligikaudne. Ometigi on ka ligikaudne valem kasulik, kui ta ei ole väga vigane, s. t. kui korrelatsioon on küllalt tihe (tavaliselt, kui $r > 0{,}8$). Seepärast on mõtet püüda katseandmete põhjal määrata parameetrid a ja b, siis võib valemit kasutada tunnuse y arvutamiseks tunnuse x järgi või vastupidi väljaspool tehtud vaatlusi. Parameetrid avalduvad valemitega:

$$a = r \cdot \frac{\sigma_y}{\sigma_x}; \qquad b = \bar{y} - a\bar{x}. \qquad \text{(III.4)}$$

Olgu meil tarvis leida valem, mis kirjeldaks saagikuse y sõltuvust huumuse protsendist mullas x. Katseandmetest on teada, et $r = 0{,}90$ (seega valemit on mõtet leida), $\bar{x}=5\%$, $y = 20$ ts/ha, $\sigma_x = 3\%$, $\sigma_y = 2$ ts/ha. Saame:

$$a = 0{,}9 \cdot 2/3 = 0{,}6; \qquad b = 20 - 0{,}6.5 = 17.$$

Seega siis $y = 0{,}6x + 17$.

Me mõõtsime huumusesisalduse uudismaatükil, see oli 10%. Millist saaki võiks sellelt tükilt loota, kui me ta üles harime? Et $x = 10$, siis

$$y = 0{,}6.10 + 17 = 23 \text{ ts/ha}.$$

Mida suurem on korrelatsioonikordaja absoluutväärtus, seda usaldusväärsem on selline valem.

6. Mitmene korrelatsioon.

Kui nähtust mõjutavad mitu faktorit, kerkib küsimus nende faktorite koosmõju tugevuse mõõtmisest.

Korrelatsioonanalüüs algab esmaste korrelatsioonikordajate r_{x_1y} leidmisest, mis kirjeldavad sideme tugevust ühe faktori ja nähtuse vahel. Võib leida mitu sellist kordajat r_{x_1y}, r_{x_2y}, ... , kus y tähistagu näiteks saagikust, x_1 - sademete hulka, x_2 - efektiivsete temperatuuride summat, x_3 - vegetatsiooniperioodi pikkust jne. Järgmiseks astmeks oleks m i t m e s e k o r r e l a t s i o o n i k o r d a j a R leidmine, mis määrab, kui tugevasti sõltub saagikus kõigist neist faktoreist kokku. R arvutamine on väga töömahukas, seepärast vaadelgem siin ainult lihtsaimat juhtu - kahe faktori koosmõju. Tähistagu y nagu ennegi kultuuride saagikust, x_1 - hüdrotermilist koefitsienti, x_2 - tootmispõhifondide maksumust. Kõigepealt leiame korrelatsiooni nende nähtuste vahel paarikaupa. Olgu need korrelatsioonid:

$$r_{x_1y} = 0{,}80 ,$$

$$r_{x_2y} = 0{,}67 ,$$

$$r_{x_1x_2} = 0{,}31 .$$

(Siinkohal kaks märkust. Esiteks, valemist III.1 on silmanähtav, et $r_{xy} = r_{yx}$. Teiseks ei ole põhjust arvata, et hüdrotermilise koefitsiendi ja tootmispõhifondide maksumuse vahel oleks mingi seos, ehkki korrelatsioonikordaja tuleb nullist erinev. Asi on selles, et väikesed korrelatsioonikordaja väärtused (alla 0,5 - 0,6) ei ole statistiliselt olulised, s. t. nad näitavad vaid kahe arvude rea juhuslikke ühiseid jooni. Peaaegu võimatu on leida kaht rida, millel ei oleks mingeid ühiseid jooni.)

Mitmese korrelatsiooni kordaja väljendab siin seda määra, kuivõrd saagikus sõltub kliimaolude ja tehnilise varustatuse koosmõjust. Selle arvutamiseks kasutatakse valemit:

$$R = \sqrt{\frac{r^2_{x_1y} + r^2_{x_2y} - 2r_{x_1y}r_{x_2y}r_{x_1x_2}}{1 - r^2_{x_1x_2}}} \quad . \qquad \text{(III.5)}$$

Meie näites

$$R = \sqrt{\frac{0{,}80^2 + 0{,}67^2 - 2.0{,}80.0{,}67.0{,}31}{1 - 0{,}31^2}} = 0{,}92 \; .$$

See tähendab, et $0{,}92^2 = 84\ \%$ teravilja saagikuse erinevustest on seletatav hüdrotermilise koefitsiendi ja tootmispõhifondide maksumuse erinevustega, kuna ülejäänud 16 % seletamiseks on vaja teisi põhjusi.

Lineaarset sõltuvust saab kahe mõjuva faktori puhul väljendada valemiga:

$$y = a + bx_1 + cx_2 \; . \qquad \text{(III.6)}$$

Siin

$$b = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{x_2y} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}} \; ,$$

$$c = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{x_2y} - r_{x_1y}\, r_{x_1x_2}}{1 - r^2_{x_1x_2}} \; , \qquad \text{(III.7)}$$

$$a = \bar{y} - b\bar{x}_1 - c\bar{x}_2 \; .$$

IV. VÄLJAVÕTTE STATISTILISTE PARAMEETRITE VEAD.

1. Uurimisvigade liigid.

Mõõtmiste tagajärjel saadud arvulised näitajad on alati vigased. Võib eristada kolm vigade allikat:

1. Metoodilist laadi vead, mis tekivad ebaõige katse- või vaatlusmetoodika rakendamise tagajärjel.

2. Mõõtmisriistade ja arvutuste ebatäpsus.

3. Representatiivsusvead, kui väljavõtte põhjal otsustatakse üldkogumi üle. Neid vigu ei ole võimalik vältida, küll aga võib neid õige ja otstarbeka väljavõttega miinimumini viia. Peale selle on välja töötatud meetodid, mis võimaldavad väljavõtte alusel hinnata nende vigade suurust.

Tuleb märkida, et kuna vead on jaotunud enamasti normaalselt, on teoreetiliselt võimalik kuitahes suurt representatiivsusviga teha. Ent väga suure vea tegemise tõenäosus on kaduvalt väike, praktiliselt null. Statistiline veahinnang tähendabki piiride leidmist, millest viga võib suurem olla vaid teatud tõenäosusega p. Kui võtta p küllalt väike, siis võib oletada, et nii väikese tõenäosusega sündmust (vea ulatumine piiridest välja) ei esine. Kõige sagedamini kasutatakse väärtusi $p = 0{,}05$ või $p = 0{,}01$. Sel juhul räägitakse vea maksimaalsest suurusest 95 % või 99 % tasemel või olulisusnivoo juures. Eriti vastutusrikka uurimise juures tuleb leida vea suurus 99,99 % tasemel, s. t. suurus, mida viga võib ületada vaid 1 juhul 10000-st. Märgime veel, et tegelik viga pole tavaliselt niigi suur. Näiteks 60 % tasemel, s. t. igal 3 juhul 5-st on vea maksimaalne suurus 2 korda väiksem kui 95% tasemel.

Käesolevas peatükis vaadeldakse keskväärtuse, standard-

hälbe ja korrelatsioonikordaja representatiivsusvigu. Representatiivsusvea definitsioonist on ilmne, et kui on uuritud terve kogum, siis on selline viga võrdne nulliga.

2. Väljavõtte aritmeetilise keskmise viga.

Ühelt poolt sõltub representatiivsusviga tunnuse variaablusest üldkogumis, teiselt poolt väljavõtte mahust. Mida suurem on tunnuse variaablus, seda "halvematest" elementidest representatiivsuse mõttes koosneb kogum. Kui aga kogumi variaablus võrdub nulliga, s. t. kõik elemendid on võrdse väärtusega, siis on ükskõik, millise elemendi me võtame, saame alati täpse väärtuse - representatiivsusviga võrdub nulliga. Samuti: mida suurem on väljavõte, seda vähem on neid kogumi elemente, mis väljavõttesse ei sattunud ja mille arvestamata jätmisest viga tekibki.

Kui väljavõtte maht on vähemalt 30 % kogumi mahust, kasutatakse aritmeetilise keskmise vea arvutamiseks järgmist valemit:

$$m = \frac{ds}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}, \qquad \text{(IV.1)}$$

kus m on aritmeetilise keskmise viga;
s - väljavõtte standardhälve;
n - väljavõtte maht ja
N - kogumi maht.

Suurusest d tuleb juttu allpool.

Kui aga väljavõtte maht on väike võrreldes kogumi mahuga, s. t. suhe n:N on lähedane nullile, siis omandavad parempoolne tegur $\sqrt{1 - \frac{n}{N}} \approx 1$ ja valem lihtsama kuju

$$m = \frac{ds}{\sqrt{n}} . \qquad \text{(IV.2)}$$

Suuruse d väärtus sõltub tasemest, millel viga arvutatakse. Kui viga on jaotunud normaalselt, siis saab d väärtuse leida tabelitest. Kasutatavamate tasemete jaoks esitame osa d väärtusi:

Tõenäosuse tase	d
68 %	1,00
95 %	1,96
99 %	2,58
99,99 %	3,80

Toome näiteid valemite (IV.1) ja (IV.2) rakendamisest.

1) Uuriti teraviljade saagikust territooriumil, kus asub 800 majandit. Teostati väljavõte mahuga 400 majandit. Väljavõtte standardhälve s = 2 ts/ha, keskmine saagikus $\bar{x}$ = 18,2 ts/ha. On vaja määrata aritmeetilise keskmise võimalik viga tasemetel 68 % ja 95 %.

Kuna väljavõte on suhteliselt suur, kasutame valemit (IV.1).

$$m_{68} = \frac{1.2}{\sqrt{400}} \cdot \sqrt{1 - \frac{400}{800}} = 0,07 \text{ ts/ha} ,$$

$$m_{95} = 1,96\, m_{68} = 0,14 \text{ ts/ha} .$$

Seega tõenäosusega 68 % on aritmeetilise keskmise tõeline väärtus 18,2 $\pm$ 0,07 ts/ha, s. t. tõeline keskmine saagikus on 18,13 ja 18,27 ts/ha vahel; 95 % tõenäosusega on tõeline keskmine 18,06 ja 18,34 ts/ha vahel.

On selline viga suur või väike? Sellele vastamiseks arvutame s u h t e l i s e v e a

$$\mu_{95} = \frac{m_{95}}{\bar{x}} , \qquad \text{(IV.3)}$$

$$\mu_{95} = \frac{0,14}{18,2} \cdot 100\ \% = 0,5\ \% .$$

Seega on viga praktiliselt tühine ja me võime ütelda, et kõigi majandite keskmine saagikus on 18,2 ts/ha.

2) 800 majandist võeti väljavõttesse ainult 25. Oletame, et me saime väljavõtte standardhälbe väärtuseks nagu ennegi 2 ts/ha, väljavõtte keskmiseks aga 18,4 ts/ha.

Kuna väljavõtte maht on umbes 3 % kogumi mahust, on lihtsam kasutada valemit (IV.2).

$$m_{68} = \frac{1.2}{\sqrt{25}} = 0{,}4 \text{ ts/ha} ,$$

$$m_{95} = \frac{1{,}96.2}{\sqrt{25}} = 0{,}78 \text{ ts/ha} ,$$

$$\mu_{95} = \frac{0{,}78}{18{,}4} \cdot 100\,\% \approx 4{,}2\,\% .$$

Seega on suhteline viga üsna suur. Väljavõtte maht osutus liiga väikeseks.

3) Olgu meil tehtud väljavõte mahuga 25 majandit, tulemustega, mis me esitasime eespool. Kuna suhteline viga oli meie jaoks liiga suur, huvitab meid küsimus, kui suur peaks olema väljavõte, et suhteline viga 95 % tasemel oleks alla 2 %?

$$\bar{x} = 18{,}4 \text{ ts/ha} ,$$
$$s = 2 \text{ ts/ha} ,$$
$$\mu_{95} < 0{,}02 .$$

Siis

$$m_{95} = \mu_{95} \cdot \bar{x} < 0{,}02 \cdot 18{,}4 = 0{,}368 \text{ ts/ha} ,$$

$$m_{95} = \frac{1{,}96s}{\sqrt{n}} = \frac{1{,}96.2}{\sqrt{n}} = \frac{3{,}92}{\sqrt{n}} ,$$

$$0{,}368 > \frac{3{,}92}{\sqrt{n}} ,$$

$$\sqrt{n} > \frac{3{,}92}{0{,}368} \approx 10{,}7 ,$$

$$n > 114{,}49 .$$

Seega peab väljavõtte maht olema vähemalt 115 majandit.

3. Standardhälbe viga.

Küllalt suure väljavõtte mahu (vähemalt 30) puhul määrab standardhälbe representatiivsusvea valem

$$m(s) = \frac{s}{\sqrt{2n}} \quad . \qquad \text{(IV.4)}$$

Seega sõltub ka standardhälbe viga standardhälbest endast ja väljavõtte suurusest. Arvutame ka standardhälbe suhtelise vea

$$\mu(s) = \frac{m(s)}{s} = \frac{s}{s\sqrt{2n}} = \frac{1}{\sqrt{2n}} \quad . \qquad \text{(IV.5)}$$

Siit näeme, et standardhälbe suhteline viga sõltub vaid väljavõtte mahust.

Arvutame standardhälbe representatiivsusvea näite 1) jaoks.

$$m(s) = \frac{2}{\sqrt{2.400}} = 0{,}07 \text{ ts/ha} \ ,$$

$$\mu(s) = \frac{100\ \%}{\sqrt{800}} = 3{,}5\ \% \ .$$

4. Väljavõtte korrelatsioonikordaja representatiivsusviga.

Ka korrelatsioonikordaja määratakse sageli väljavõtte alusel. Küllalt suure väljavõtte ($n > 50$) puhul on korrelatsioonikordaja representatiivsusviga

$$m(r) = \frac{1 - r^2}{\sqrt{n}} \quad . \qquad \text{(IV.6)}$$

Näeme, et viga sõltub siingi pöördvõrdeliselt väljavõtte mahu ruutjuurest. Mida tihedam on seos nähtuste vahel, seda väiksem on viga.

Olgu 64 punkti andmete alusel arvutatud mingi kliima-

faktori ja düsenteeriasse haigestumuse korrelatsioon. Olgu $r = 0{,}82$. On selge, et me ei suuda arvutada korrelatsiooni kõigis maakera punktides tehtud vaatluste alusel. Seepärast tekib representatiivsusviga

$$m(r) = \frac{1 - 0{,}82^2}{\sqrt{64}} = \frac{0{,}33}{8} \approx 0{,}04$$

ja

$$r = 0{,}82 \pm 0{,}04 .$$

Tekib küsimus - kui me korrelatsiooni hindame teatud veaga, kas siis seos nähtuste vahel üldse eksisteeribki, s. t. kas r nullist erinev väärtus on statistiliselt oluline, kas see pole saadud mitte juhuslikult, vea tulemusena. Et sellele vastata, arvutame vea ülemise piiri praktiliselt täiesti kindlalt, tõenäosusega 99 %. See võrdub, nagu nägime punktis 2, suurusega $2{,}6\, m(r)$. Kui nüüd $|r| : 2{,}6m(r) \geq 1$, siis võib lugeda korrelatsiooni oluliseks ja seost reaalselt eksisteerivaks, vastasel juhul tuleb seos lugeda mittetõestatuks (ta võib olla reaalne, kuid võib ka mitte olla).

Meie näites $0{,}82 > 2{,}6.0{,}04$, seega seos on reaalne.

Olgu mingis teises näites $r = -0{,}30$ ja $m(r) = 0{,}15$. Siis ei saa läbitöötatud andmete alusel midagi kindlat öelda seose olemasolu kohta kahe vaadeldud nähtuse vahel.

(Isegi kui korrelatsioonikordaja selles näites oleks leitud veatult, on ta liialt väike, et olla oluline. Tõepoolest, kuna $(-0{,}30)^2 = 0{,}09$, siis selgitab esimese näitaja muutlikkus vaid 9 % teise nähtuse muutlikkusest.)

5. Aritmeetilise keskmise vea täpsustamine.

Nägime, et aritmeetilise keskmise viga sõltub standardhälbest. Standardhälbe omakorda leiame me teatud representatiivsusveaga. Selletõttu ei ole valemite IV.1 või IV.2 järgi määratud aritmeetiline keskmise viga täiesti õige.

Täpsema tulemuse annab valem

$$m(\bar{x}) = \frac{ts}{\sqrt{n}} . \qquad (IV.7)$$

See valem erineb valemist IV.2 ainult ühes suhtes - koefitsiendi d asemel seisab koefitsient t . t väärtused sõltuvad mitte üksnes tõenäosustasemest, vaid ka väljavõtte mahust. Nende saamiseks kasutatakse spetsiaalset statistilist tabelit (vt. tabel 27). Selles tabelis on veergude pealkirjadeks tõenäosustaseme väärtused, ridade pealkirjadeks ühe võrra vähendatud väljavõtte maht.

Arvutame uuesti aritmeetilise keskmise vea punkt 2 teises näites:

$$\bar{x} = 18,4 \text{ ts/ha} ,$$
$$s = 2 \text{ ts/ha} ,$$
$$n = 25 .$$

n - 1 on siis 24. Leiame vea tasemel 95 %. Tabelist 27 leiame, et t = 2,06

$$m(\bar{x}) = \frac{2,06 . 2}{\sqrt{25}} = \frac{4,12}{5} = 0,82 \text{ ts/ha} .$$

Seega on erinevus väike võrreldes valemi IV.2 abil saadud tulemusega.

Tabeli 27 uurimisel selgub: mida suurem on n , seda vähem erinevad t ja d . Kui $n = \infty$, siis $t = d$. Üldiselt, kui $n > 30$, loetakse t ja d erinevus tühiseks ja siis on muidugi ükskõik, kas tarvitada valemit IV.2 või IV.7. Ainult väikeste väljavõtete korral annab valem IV.7 tunduvalt usaldusväärsemaid tulemusi.

V. KARTOGRAAFILISE MATERJALI STATISTILINE TÖÖTLEMINE.

1. Statistilise pinna ja selle kaardi mõiste.

Geograafiline kaart annab võimaluse uurida loodust, majandust, kultuurilisi nähtusi suurtel aladel (riikides, looduslikes vööndites, isegi kogu maakeral). Kaasaegsete kaartide kujutamisvahendid võimaldavad paljusid ruumilisi seaduspärasusi leida visuaalselt, heites vaid pilgu kaardile. Kuid kaart võimaldab ka mõõtmisi teha, saada nähtuste arvulisi väärtusi. Hea kaart annab uurija käsutusse tohutu hulga arvulist informatsiooni reljeefi, kliima, taimestiku, mullastiku, rahvastiku, majandusliku tegevuse jne. kohta. Selle arvulise informatsiooni töötlemiseks on vajalikud statistilised meetodid.

S t a t i s t i l i s e k s p i n n a k s nimetatakse abstraktset reljeefi, mille moodustavad antud tunnuse väärtused mingi territooriumi lõpmatus arvus punktides. Sellise pinna lihtsaimaks näiteks on absoluutsete kõrguste pind - see, mida geograafid tavaliselt nimetavad reljeefiks. Tihti räägitakse ka baarilisest reljeefist - õhurõhu pinnast. Kuid statistilisi pindu saab joonistada peaaegu igale nähtusele: temperatuuri, magneetilise deklinatsiooni, rahvastiku tiheduse, keskmise saagikuse jne. pinnad. Kõige loomulikum on selliste pindade koostamine pidevate geograafiliste nähtuste jaoks, selliste nähtuste jaoks, mis esinevad uuritava ala kõigis punktides. Need nähtused varieeruvad ruumiliselt sujuvalt, ilma märkimisväärsete hüpeteta, mistõttu moodustuvad sujuvad pinnad; neid kujutatakse kõige sagedamini samasuurusjoontega (isohüpsid, isotermid, isodeemid - rahvastiku samatihedusjooned, isokroonid jne.). Muidugi on enamiku nähtuste statistiline pind üldistus, võimatu on tegelikkuses näha baarilist reljeefi või rahvastiku tiheduse pinda. Samasuurusjoonte kõrval kasutatakse statis-

tiliste pindade kujutamiseks kaardil ka teisi meetodeid, näiteks kartogramme ja punktide meetodit.

Punktide meetodit kasutatakse massiliste hajutatud nähtuste kujutamiseks, eriti sageli põllumajanduskaartide juures (maarahvastiku, karja suuruse, külvipindade kaardid). Punktid paigutatakse kaardile vastavalt nähtuse tegelikule paiknemisele või ühtlaselt territoriaalse ühiku piires. Loomulik, et mida väiksemad on sellised ühikud, seda õigemini on kujutatud nähtuste paiknemine. Statistilise pinna kõrgusest loob ettekujutuse punktide tihedus: kus punkte on tihedamini, seal on statistiline pind kõrgem. Lohkudes on punkte hõredalt või üldse mitte.

Pidevate suuruste kujutamiseks kartogrammina muudame me pinna kunstlikult astmeliseks, kuna diskreetsete nähtuste statistilise pinna kujutamiseks on kartogrammide meetod eriti sobiv. Viirutuste või värvi tumedusaste võimaldab kindlaks teha statistilise pinna kõrgendikud ja lohud.

Niisiis, samasuurusjoonte kaardid, kartogrammid ja punktkaardid loovad statistilise pinna visuaalse efekti, peegeldavad näitlikult looduslike ja ühiskondlike nähtuste arvuliste väärtuste ruumilist jaotust. Seega võib neid kaarte nimetada statistiliste pindade kaartideks.

2. Arvuliste andmete saamine kaardilt.

Statistiliste pindade kaardid võimaldavad määrata nähtuse arvulist väärtust igas kaardi punktis. Seega on statistiline kogum lõpmata suur. Andmete töötlemiseks tuleb teha väljavõte. Kuid kaardi andmete alusel saadud statistiliste näitajate täpsus ei sõltu ainult väljavõtte suurusest, vaid ka selle tegemise viisist: tuleb õigesti valida k o n t r o l l p u n k t i d , kus mõõdetakse uuritava nähtuse suurust. Väljavõte on representaabel üldiselt siis, kui kontrollpunktid on jaotatud ühtlaselt üle territooriumi, võrdeliselt üksikute alade pindalaga. Seda

võib saavutada mitmesuguste geograafiliste võrkude kaardile asetamisega.

Kõige levinum kontrollpunktide valimise viis on ruutude võrgu asetamine kaardile ja nähtuse mõõtmine võrgu sõlmedes. Sellise võrgu võib spetsiaalselt valmistada trafaretina, tihti aga saab kasutada kaardivõrku (meridiaanide ja paralleelide lõikepunktid, kilomeetrivõrk topograafilisel kaardil).

Korrelatsioonarvutustes tuleb ühes kontrollpunktis mõõta mitme nähtuse väärtusi. Tavaliselt on need nähtused kujutatud eri kaartidel, mis pealegi on tihti koostatud erisugustes mõõtkavades ja projektsioonides. Tekib küsimus, kuidas saavutada kontrollpunktide ühtimist mõlemal kaardil. Kontrollpunktidena võib kasutada kartograafilise võrgu sõlmi, tuleb ainult arvestada, et põhja-lõuna suunas ulatusliku territooriumi kaartidel ei ole see võrk enam ühtlane - meridiaanid koonduvad pooluste suunas. See kartograafilise võrgu omadus toob kaasa märkimisväärse vea siiski ainult aladel, mille ulatus on 1500 - 2000 km ja rohkem põhja-lõuna suunas.

Arvulise väärtuse saamine kontrollpunktis tähendab statistilise pinna kõrguse määramist. Samasuurusjoonte kaardil võib neid suurusi määrata silma järgi, interpoleerides, samuti nagu leitakse absoluutne kõrgus topograafiliselt kaardilt. Kartogrammidel on reljeef astmelise kujuga, seepärast ei ole interpoleerimisel mõtet. Kõige raskem on mõõta nähtuse väärtust punktkaardilt. Kontrollpunkti ümber joonistatakse sobiva, kõigi kontrollpunktide jaoks sama raadiusega ring ja loetakse ringi jäänud punktid. Siis nähtuse väärtus kontrollpunktis

$$I = \frac{W \cdot m}{S} , \qquad (V.1)$$

kus W on punkti mõõtkava, m - ringi jäänud punktide arv, S - ringi tegelik pindala.

Olgu meil antud näiteks kaart mõõdus 1:10 000 000,

kuhu on punktidega kantud rahvastiku paiknemine. Üks punkt tähistab 1000 inimest; W = 1000. Kontrollringi raadius kaardil on 2 mm, tegelikkuses seega 20 km. Ringi sisse jäi 8 punkti. Ringi tegelik pindala on

$$\pi R^2 = 3{,}14 \cdot 400 = 1256 \text{ km}^2 .$$

Siis on rahvastiku tihedus kontrollpunktis

$$\frac{1000 \cdot 8}{1256} \approx 6{,}4 \ \frac{\text{inimest}}{\text{km}^2} .$$

Kui ringi raadius võtta liiga väike, jääb ringi sisse liiga vähe punkte ja ühe punkti juhuslik sisse- või väljajäämine muudab liiga palju tulemust. Sel meetodil arvutame me aga rahvastiku tiheduse kogu ringis, mitte üksnes kontrollpunktis. Mida suurem ring, seda enam võib nii saadud tihedus erineda tõelisest.

Lihtsam on kasutada teist teed. Kartogrammidel ja samasuurusjoonte kaartidel on statistilised suurused antud rühmitatuna. Vahemike piirideks on samasuurusjooned või kartogrammi astmed, vahemike sagedusteks aga sellesse vahemikku langevate kontrollpunktide arv. Vaatleme Valgevene NSV aasta keskmiste temperatuuride kaarti (Valgevene NSV atlas, 1958, lk. 31), millele on peale asetatud ruutvõrk (joon.7). Vahemike arv (isotermide vahede arv) võrdub siin seitsmega. Vahemikud ja igasse vahemikku langevate kontrollpunktide arv (sagedus) on toodud tabelis 19.

T a b e l 19 .

Vahemikud	Sagedused
4,0–4,5°	6
4,5–5,0°	18
5,0–5,5°	100
5,5–6,0°	64
6,0–6,5°	71
6,5–7,0°	56
7,0–7,5°	11
Summa	326

Nagu varemgi, võime nende andmete põhjal joonistada histogrammi (joon. 7), arvutada keskväärtuse, limiidid, standardhälbe.

Samuti töödeldakse kartogramme ka statistiliselt. Mõnevõrra keerulisemaks osutub punktkaardi kasutamine, seepärast peatume sellel pikemalt.

Punktkaardi puhul tuleb vahemike piirid alles määrata, aluseks võttes kontrollringi jäänud punktide arvu. Tabel 20 on koostatud kartuli külvipinna kaardi järgi Valgevene NSV-s. Iga kontrollpunkti ümber tõmmati 1,5 mm raadiusega ring ja loeti punktide arv ringis, mis varieerus 0-st 50-ni. See muutumispiirkond jagati 10 vahemikuks. Iga vahemiku sageduseks on kontrollpunktide arv, mille ümber tõmmatud ringidesse sattus vastav arv punkte.

Tabel 20.

Vahemikud	Sagedused
1 - 5	2
6 - 10	15
11 - 15	24
16 - 20	41
21 - 25	26
26 - 30	34
31 - 35	20
36 - 40	11
41 - 45	6
46 - 50	3
Kokku	182

Saadud tabeli võib muidugi valemi (V.1) abil teisendada külvipinna hektariteks ja seejärel arvutada statistilised näitajad, arvutuslikult on aga mugavam vastupidine protseduur: tähistades punktide arvu ringis n, leida $\bar{n}$ ja $\sigma(n)$, kasutades lihtsustatud viise ja alles seejärel teisendada tulemused hektariteks. Teeme vastava arvutuse tabeli 20 andmete jaoks. Seejuures teostame teisenduse

$$n_i' = \frac{n_i - 23}{5} .$$

Vt. tabel 21.

Tabel 21.

Vahemik	Selle keskpunkt n	n'	m	n'm	n'^2m
1 - 5	3	-4	2	-8	32
6 - 10	8	-3	15	-45	135
11 - 15	13	-2	24	-48	96
16 - 20	18	-1	41	-41	41
21 - 25	23	0	26	0	0
26 - 30	28	1	34	34	34
31 - 35	33	2	20	40	80
36 - 40	38	3	11	33	99
41 - 45	43	4	6	24	144
46 - 50	48	5	3	15	75
	Kokku		182	4	736

$$\bar{n}' = \frac{4}{182} \approx 0{,}02 \; ; \quad (\bar{n}')^2 = 0{,}0004 \approx 0 \; ;$$

$$s(n') = \sqrt{\frac{736}{182} - 0{,}0004} \approx 2 \; ;$$

$$\bar{n} = 0{,}02.5 + 23 = 23{,}1 \; ;$$

$$s(n) = 2.5 = 10 \; .$$

Et leida kartuli keskmine osatähtsus külvipinnas Valgevene NSV-s ja selle osatähtsuse standardhälve, on meil tarvis veel järgmisi suurusi: V - punkti mõõtkava, s. t. mitmele hektarile kartuli külvipinnale vastab üks punkt, k - kaardi mõõtkava ja külvipinna osatähtsus Valgevene NSV territooriumis. Arvutusskeemi lõpetamiseks võtame need suurused suvaliselt (seetõttu ei ole lõppresultaadid õiged). Olgu kaardi mõõt 1:1 000 000, V = 50 ha ja k = 40 %. Siis oleks kontrollringi tegelik pindala $S = \pi . 1{,}5^2$ km$^2 \approx 700$ ha.

$$\bar{x} = \frac{V.\bar{n}}{S.k} = \frac{5.23{,}1}{700.0{,}4} = 41 \; \% \; .$$

$$s = \frac{V.s(n)}{S.k} = \frac{5.10}{700.0{,}4} = 18 \; \% \; .$$

Arvutame ka võimaliku vea 95-% tasemel. Kuna vaatluste - kontrollpunktide arv oli 182, siis aritmeetilise keskmise viga

$$m(\bar{x}) = \frac{1{,}96.18}{\sqrt{182}} \approx 2{,}6 \; \%$$

ja standardhälbe viga

$$m(s) = \frac{18}{\sqrt{364}} \approx 1 \; \% \; .$$

Seega on kartulit Valgevene NSV-s $41 \pm 2{,}6$ % külvipinnast, kuna tema osatähtsuse standardhälve moodustab 18 ± 1 %.

Võib kasutada aga ka teist katseskeemi, mis ei ole seotud väljavõtete tegemisega. Ehkki meil ei ole võimalik loendada kõiki territooriumi punkte, kus tunnuse väärtus langeb mingisse vahemikku, on ometi selge, et see punktide

arv on võrdeline pindala suurusega, mille võtavad enda alla sellesse vahemikku kuuluvad tunnuse väärtused kaardil. Vastav pindala ongi siis vahemiku sageduseks m_i. (Meenutame, et $\bar{x}$ ja σ väärtus ei muutu, kui sagedusi korrutada ühe ja sama arvuga.)

Toome näiteks rahvastiku keskmise tiheduse arvutamise liiduvabariigis, mis koosneb 5 oblastist. Kartogrammilt on näha, et iga oblast on värvitud erineva tooniga, seega on erineva rahvastiku tihedusega. Planimeetri või paleti abil mõõdame oblastite pindala, ükskõik, kas tegeliku või kaardi mõõtkavas. Andmed koondame tabelisse 22.

T a b e l 22.

Rahvastiku tihedus x	Oblasti pindala S	S'	x'	S'x'	S'x'^{2}
12	250	50	-2	-100	200
14	85	17	-1	- 17	17
16	100	20	0	0	0
18	70	14	1	14	14
20	75	15	2	30	60
Kokku	580	116		- 73	291

Kasutame jälle lihtsustatud arvutusskeemi, taandades pindalad (sagedused) 5-ga, lahutades x_i-st 16 ja jagades vahed 2-ga.

Saame:

$$\bar{x}' = \frac{\sum S'x'}{\sum S'} = \frac{-73}{116} = -0{,}63 ;$$

$$\sigma' = \sqrt{\frac{\sum S'x'^2}{\sum S'} - \bar{x}'^2} = \sqrt{\frac{291}{116} - 0{,}63^2} = 1{,}45 .$$

Lähme üle tõelistele $\bar{x}$ ja σ väärtustele:

$$\bar{X} = -0{,}63 \cdot 2 + 16 = 14{,}7 \text{ in/km}^2;$$

$$\sigma = 1{,}45 \cdot 2 = 2{,}9 \text{ in/km}^2 .$$

Kuna me võime sagedusi korrutada või jagada ükskõik millise arvuga, ilma et keskmine või keskmine ruuthälve muutuksid, ei ole meil tarvis üle viia kaardi mõõtkavas pindala tegelikuks. Veel enam, pindalade asemel võime sagedustena kasutada isegi paleti ruutude või planimeetri jaotuste arvu.

Kui meie käsutuses on rahvastiku paigutuse punktkaart, siis vabariigi keskmise rahvastiku tiheduse saame leida valemist:

$$\bar{x} = \frac{nV}{S} . \qquad (V.2)$$

Siin n on punktide üldarv, S - tegelik pindala. Oletame, et vabariigi pindala on 45 000 km^2, punkte on 1280, üks punkt vastab 1000 inimesele. Siis on rahvastiku keskmine tihedus:

$$\bar{x} = \frac{1280 \cdot 1000}{45000} = 28{,}4 \frac{\text{inimest}}{\text{km}^2} .$$

Keskmise tiheduse võib arvutada rajoonide kaupa, siis on võimalik leida ka keskmine ruuthälve, mis iseloomustab rahvastiku paigutuse ruumilist variaablust vabariigis.

Seega võib statistilise pinna kaardilt leida $\bar{x}$ ja σ, kasutamata väljavõtteid ja seetõttu suurema täpsusega. Punktkaardil on see meetod kasutatav, kui punkte ei ole väga palju, muidu tehakse vigu juba punktide kokkulugemisel.

3. Statistiliste näitajate määramise visuaalkartograafiline meetod.

Statistiliste pindade kaardid lubavad hinnata põhiliste statistiliste näitajate suurust ligikaudu, silma järgi. Nii saab kiiresti otsustada, kas nende näitajate täpsem arvutamine tasub ennast ära. Mõnigi kord piisab väga ligikaudsetest hinnangutest.

Statistilise pinna kaardi reljeef võimaldab näha tema põhilisi seaduspärasusi, lokaliseerida reljeefi ebatasasusi. Suurte alade kaartidel moodustab reljeef tihti lained - kõrgemate ja madalamate osade reeglipärase vaheldumise. Näiteks võivad olla tsonaalsete nähtuste pinnad (temperatuur, õhurõhk, sademed jms.). Laineharjad ja -nõod vahelduvad põhja-lõuna suunas nähtuse olemusest tingituna. Tsonaalsetele "lainetele" võivad olla "peale asetatud" teist, kolmandat jne. järku kohalikud "lained", mille kutsuvad esile atsonaalsed faktorid (maapinna reljeef, kaugus merest vms.). Väikeste alade kaartidel näeme aga sageli vaid laine üksikuid osi, enamikus nõlva.

Me räägime, et statistiline reljeef on lihtne, kui tegemist on ühe kaldpinnaga, ja keeruline, kui ta on laineline. Sõltuvalt reljeefi kujust on ka statistiliste näitajate määramine kas lihtsam või keerulisem.

Limiitide leidmine ei valmista mingeid raskusi samasuurusjoonte kaardil või kartogrammil. Teame ka, et keskväärtus asub kusagil limiitide vahel, keskmine ruuthälve aga moodustab alla poole muutumisulatusest.

1. Kui territooriumil suured ja väikesed tunnuse väärtused hõlmavad ligikaudu võrdse ala, on keskväärtuse hinnanguks limiitide vahe keskpaik.

2. Kui suured (väikesed) väärtused on ülekaalus, siis on keskväärtus lähemal maksimumile (miinimumile).

Esimesel juhul on reljeefi kallakus ühesugune, samasuurusjoonte vahe püsiv. Kui samasuurusjooned on tihedamalt väikeste väärtuste osas, on reljeef kumer ja keskmine nihutatud maksimumi suunas ja vastupidi.

Raskem on hinnata keskmise ruuthälbe suurust. Empiiriliselt on kindlaks tehtud, et lihtsa reljeefi ja tasase kallaku korral on keskmine ruuthälve umbes 2/7 limiitide vahest. Suurema või väiksema täpsusega jääb see hinnang kehtima ka keerulisema reljeefi puhul.

Võtame Usbeki NSV atlasest kaardi leheküljel 33: sooja perioodi sademete hulk. Kaart on koostatud samasuurusjoonte meetodil. Minimaalne sademete hulk on vahemikus 0-30 mm, seega umbes 15 mm. Maksimaalne sademete hulk ulatub üle 400 mm; kahjuks ei ole antud selle vahemiku ülapiiri. Võtame maksimaalseks väärtuseks umbes 600 mm. Samasuurusjooned on tihedasti koos suurte väärtuste osas, seega keskväärtus on nihutatud miinimumi suunas. Hindame seda 50 mm-ks. Keskmise ruuthälbe leiame, korrutades limiitide vahe 2/7-ga. Seejuures ei ole mõtet lahutada 600-st 15, sest meie limiitide hinnangu viga on ilmselt suurem kui 15. Maksimaalne väärtus võib samahästi olla ka 615 või 590 mm. Keskmise ruuthälbe ligikaudseks suuruseks saame 600,2:7 150-200 mm.

Palju täpsema tulemuse saame Valgevene NSV aasta keskmise temperatuuri määramisel, kuna see statistiline pind on tasane. Limiidid on 4,0° ja 7,5°. Keskväärtus on siis (4,0 + 7,5):2 = 5,7°. Keskmise ruuthälbe suuruseks on (7,5 - 4,0) . 2:7 = 1° .

Korrelatsioonikordajat võib samuti määrata silma järgi, kuid äärmiselt ebatäpselt. Umbkaudseks hindamiseks on kasulik tarvitada graafilist meetodit.

Kummagi tunnuse statistilise reljeefi kaardile tõmmatakse mingis punktis joon kallaku suurima järskuse suunas. Nende joonte vahelise nurga α koosinus ongi korrelatsioonikordaja ligikaudne väärtus selles punktis:

$$r = \cos \alpha . \qquad (V.3)$$

Joonisel 8 on näidatud nurga α konstrueerimine punktis O . Sellest punktist tõmmatakse samasuurusjoontega risti olevad sirged. Sirgetevahelise nurga võib mõõta ja leida koosinuse väärtuse tabelist või arvutuslükatil. Võib aga ka ehitada täisnurkse kolmnurga, mille lähiskaateti suhe hüpotenuusisse on teatavasti nurga koosinus:

$$r = \frac{OD}{OC} = 0,88$$

meie joonisel. Seejuures on oluline ainult, et nurk ODC

oleks täisnurk; kummal sirgel ja kus me valime punkti D , on ükskõik.

Et kaks risti asuvate haaradega nurka on võrdsed, võib nurgaks α kasutada lihtsalt samasuurusjoontevahelist nurka, pidades meeles, et kahe kõvera vaheline nurk on nendele kõveratele lõikepunktis joonistatud puutujate vaheline nurk.

Samasuurusjooned lõikuvad väga harva kõigis kaardi osades ühesuguse nurga all. Mõõtes lõikenurka mitmes piirkonnas, saame kindlaks teha korrelatsiooni territoriaalse varieeruvuse, mis peaks pakkuma geograafile suurt huvi.

Vaatleme NSV Liidu jaanuarikuu isotermide kaarti. Püüame määrata sõltuvuse koha geograafilise laiuse (sellest sõltub päikesekiirguse summa) ja temperatuuri vahel. Esimese näitaja samasuurusjoonteks on loomulikult paralleelid. Et koha geograafiline laius ei määra kaugeltki täielikult jaanuarikuu temperatuuri, on tegemist korrelatiivse seosega. Korrelatsioonikordaja on negatiivne: geograafilise laiuse suurenemine tähendab reeglina temperatuuri alanemist. Määrates samasuurusjoontevahelise nurga mitmes kaardi osas, näeme, et Euroopa osas lõikuvad isotermid paralleelidega peaaegu täisnurga all, seega korrelatsioon praktiliselt puudub (cos 90° = 0). Lääne-Siberis on lõikumisnurgad väikesed, korrelatsioonikordaja järelikult suur (cos 0° = 1). See on loomulikult seletatav tsüklonaalse tegevuse mõju nõrgenemisega.

4. Väljavõtte mahu määramine ligikaudse keskmise ruuthälbe järgi.

Keskmise ruuthälbe ligikaudne hinnang võimaldab meil määrata väljavõtte nõutavat mahtu, kui täpsem arvutus osutub vajalikuks. Teatavasti oli aritmeetilise keskmise vea hinnang väljavõttes (vt. valem IV.2):

$$m = \frac{ds}{\sqrt{n}} .$$

Arvutame siit n :

$$n = \frac{d^2 s^2}{m^2} . \qquad (V.4)$$

Olgu meil vaja leida Valgevene NSV aasta keskmine temperatuur nii, et suhteline viga tõenäosusega 95 % ei ületaks 5 %. Vastav d väärtus on teatavasti 1,96. Eelmises paragrahvis leidsime, et ligikaudne keskmine aastatemperatuur oli 5,7°, millest 5 % on umbes 0,3°; m = 0,3. Keskmise ruuthälbe ja seega ka standardhälbe hinnanguks oli 1° . Asetades need suurused valemisse V.4 saame:

$$n = \frac{1{,}96^2 \cdot 1^2}{0{,}3^2} \approx 45 .$$

Kindluse mõttes suurendame seda arvu veel veidi ja võtame kontrollpunktide arvuks 50.

Nüüd määrame ruudustiku suuruse. Paletiga mõõdame ära kaardipinna suuruse Valgevene NSV all, see on 31,3 cm^2. Siis peab ühe ruudu pindala olema 31,3 : 50 = 0,6 cm^2, ruudu külje pikkus aga $\sqrt{0{,}6} \approx 0{,}8$ cm. Sellisel juhul saame umbes 50 kontrollpunkti.

Võrdluseks toome 100 kontrollpunkti andmete läbitöötamisel saadud parameetrite väärtused: $\bar{x} = 5{,}8°$, s = 0,7° .

Täiesti analoogiliselt, kasutades korrelatsioonikordaja vea valemit, saame arvutada kontrollpunktide arvu, mis on vajalik korrelatsioonikordaja täpseks määramiseks. Siingi vajame korrelatsioonikordaja ligikaudset väärtust.

Eriti oluline on korrelatsioonikordaja ligikaudne määramine mitmese korrelatsiooni arvutamisel. Me saame silma järgi eraldada faktorid, mille omavaheline korrelatsioon on väike ega võta neid faktoreid arvesse täpsemate arvutuste juures. Sellega me vähendame tunduvalt arvutuste mahtu.

VI. MITMESUGUSED ARVUTUSMEETODID.

1. r arvutamine rühmitatud andmete järgi.

Väike väljavõte ei kindlusta alati statistiliste näitajate küllaldast täpsust. Nende näitajate usaldusväärsus kasvab vaatluste arvu suurenemisel. Kuid väljavõtte suure mahu puhul on eeltoodud korrelatsioonikordaja arvutusskeemid ebamugavad. Otstarbekam on kasutada teist skeemi, mille aluseks on andmete eelnev grupeerimine ja tunnuste tõeliste väärtuste asendamine lihtsustatud väärtustega.

Statistiliste pindade kaardid võimaldavad teha üksköik kui palju vaatlusi ja rühmitada andmeid suvalisel viisil. Tutvume üksikasjaliselt korrelatsioonikordaja arvutusskeemiga kaardi andmetest.

Tselinnõi krai kaardile kanti teraviljakultuuride saagikuse kartogramm ja hüdrotermilise koefitsiendi samasuurusjooned. Vaja on arvutada nende kahe nähtuse vahelise korrelatsiooni kordaja. Kartogrammi astmed järgnevad 3 ts/ha intervalliga, samasuurusjooned on tõmmatud iga 0,3 mm/kraadi järel.

Paneme kaardile ruudustiku. Kontrollpunktid ruutude tippudes nummerdame mingil viisil ära. Näiteks alustame vasakust ülemisest nurgast ridade kaupa, lõpetades paremas alumises nurgas.

Joonistame korrelatsioonivälja tabeli kujul (tabel 23). Veergude pealkirjadeks (teljele x) märgime hüdrotermilise koefitsiendi vahemikud 0,1 - 0,4, 0,4 - 0,7, 0,7 - 1,0; ridade pealkirjadeks (teljele y) kirjutame saagikuse intervallid 0-3, 3-6, 6-9, 9-12, 12-15.

Saame ruudustiku, mille täidame suuruste x ja y väärtustega igas kontrollpunktis. Näiteks kontrollpunktis 1 kuulub saagikus intervalli 12,1 - 15,0, hüdrotermiline koefitsient intervalli 0,7 - 1,0. Vastavasse tabeli laht-

risse paneme punkti. Nii töötame läbi kõik kontrollpunktid (üldse oli neid 340). Seejärel loeme kokku punktide arvu igas tabeli lahtris ja kirjutame selle punktide asemele; saame tabeli 23 sisemise osa.

T a b e l 23 .

	x'	-1	0	1	Veergude number 1	2	3
y'	y \ x	0,1–0,4	0,4–0,7	0,7–1,0	1	1y'	1y'^{2}
2	12,1–15,0			18 9	9	18	36
1	9,1–12,0		18 18	25 25	43	43	43
0	6,1–9,0	0 5	0 130	0 41	176	0	0
-1	3,1–6,0	-11 11	-62 62	-11 11	84	-84	84
-2	0,0–3,0	-44 22	-12 6		28	-56	112
Ridade number 1	h	38	216	86	340	-79	275
2	hx'	-38	0	86	48		
3	hx'^{2}	38	0	86	124		
4	$\sum$ my'	-55	-56	32	-79		
5	x'$\sum$ my'	55	0	32	87		

Täiendame tabelit ülalt ja vasakult ühe rea ja veeruga, kuhu me, nagu varemgi, paigutame suurused x' ja y'; vastavalt keskmises veerus ja reas x' = 0 ja y' = 0 . Nii saame tunnuste lihtsustatud väärtused (vt. ptk. 2 p. 6 ja ptk. 3 p. 3).

Joonistame paremale 3 täiendavat veergu: 1, kuhu kirjutame tabeli vastava rea summa, ly' ja ly'^2, mille sisu selgub pealkirjast.

Alla joonistame ka 3 täiendavat rida: h, hx' ja hx'^2, kus h on jällegi vastava veeru sissekannete summa. Kontrolliks: veeru 1 ja rea h summad peavad võrduma vaatluste üldarvuga 340.

Nüüd korrutame kõik põhitabeli sissekanded sissekande rea ees seisva y' väärtusega; tulemused kirjutame põhitabeli ruutudesse üles parempoolsesse nurka. Summeerime uued sissekanded veergude kaupa ja kirjutame summad täiendavasse ritta 4: $\sum my'$. Rea 4 summa peab võrduma veeru 2 summaga. Selle summa (-79) kirjutame tabeli parempoolsesse alumisse nurka veeru ja rea kohale.

Seejärel arvutame ka rea 5, korrutades rea 4 x' vastava väärtusega, ja leiame rea 5 ja veeru 3 summad. Tähistades nüüd veeru i sissekannete summa V_i ja rea j sissekannete summa R_j , võime arvutada korrelatsioonikordaja valemist:

$$r = \frac{n.R_5 - R_2V_2}{\sqrt{nR_3-R_2^2} \cdot \sqrt{nV_3-V_2^2}} . \qquad (VI.1)$$

Meie näites:

$$r = \frac{340.87 - 48.(-79)}{\sqrt{340.124 - 48^2} \cdot \sqrt{340.275 - (-79)^2}} = 0,57 .$$

2. σ ja r parandamine.

Andmete rühmitamine põhjustab vea, kuna tunnuste tõelised väärtused asendatakse vastava vahemiku keskmise väärtusega. Selletõttu osutub tunnuse variaablus suurendatuks. Dispersiooni parandamist selle vea allika arvel teostatakse valemiga:

$$\hat{\sigma}^2 = \sigma^2 - \frac{t^2}{12} . \qquad (VI.2)$$

Siin $\hat{\sigma}^2$ on parandatud dispersiooni väärtus, σ^2 - tema esialgne väärtus, t - vahemiku ulatus. Seda parandust nimetatakse Sheppardi paranduseks.

Et saada valemit keskmise ruuthälbe (või standardhälbe) parandamiseks, võtame ruutjuure valemi VI.2 mõlemast poolest:

$$\hat{\sigma} = \sqrt{\sigma^2 - \frac{t^2}{12}} . \qquad (VI.3)$$

Näiteks määrati rühmitatud andmetest saagikuse standardhälbeks Tselinnõi krais 2,6 ts/ha. Andmed olid rühmitatud vahemikkudesse ulatusega 3 ts/ha. Parandatud standardhälve võrdub:

$$\hat{s} = \sqrt{2,6^2 - \frac{9}{12}} = \sqrt{6,76 - 0,75} = 2,45 \text{ ts/ha} .$$

Ebavõrdsete vahemikkude puhul tuleb **korrektsioonivalem** väga keeruline.

Rühmitatud andmete järgi arvutatud korrelatsioonikordaja on reeglina vähendatud. Parandus arvutatakse valemiga:

$$\hat{r} = r \frac{\sigma_x \sigma_y}{\hat{\sigma}_x \hat{\sigma}_y} . \qquad (VI.4)$$

Soovitav on suuruste σ_x, σ_y, $\hat{\sigma}_x$, $\hat{\sigma}_y$ asemel kasutada lihtsustatud, kergemini leitavaid suurusi σ'_x, σ'_y, $\hat{\sigma}'_x$, $\hat{\sigma}'_y$.

Teeme paranduse eelmises paragrahvis arvutatud näite juurde. Säilitades seal sisseviidud tähistused, saame:

$$s'_x = \sqrt{\frac{R_3}{n} - \left(\frac{R_2}{n}\right)^2} = 0,59 .$$

$$s'_y = \sqrt{\frac{V_3}{n} - \left(\frac{V_2}{n}\right)^2} = 0,87 .$$

Vahemike ulatus lihtsustatud väärtustes on kõikjal 1. Seega Sheppardi parandus võrdub alati 1/12 = 0,08333. Parandatud standardhälbed:

$$\hat{s}'_x = 0,51 \; ; \qquad \hat{s}'_y = 0,82 \; .$$

Et vaatluste arv n on suur, ei ole standardhälbe s ja keskmise ruuthälbe δ vahe oluline ja me võime valemis VI.4 tarvitada standardhälbeid.

$$\hat{r} = 0,57 \cdot \frac{0,59 \cdot 0,87}{0,51 \cdot 0,82} = 0,70 \; .$$

Korrelatsioonikordaja viga arvutatakse valemiga IV.6:

$$m = \frac{1 - r^2}{\sqrt{n}} = 0,03 \; .$$

Seega on tegelik korrelatsioonikordaja:

$$r = 0,70 \pm 0,03 \; .$$

Korrelatsioonikordaja parandamine on otstarbekas, kui kasutatud vahemike arv on väiksem kui 6.

VII. STATISTILISED KRITEERIUMID.

1. Statistilise kriteeriumi mõiste.

Nii mõnigi kord tarvitame väljendusi: ligikaudu võrdne, erinevused on tingitud juhuslikkusest, erinevused ei ole olulised. Selliseid otsustusi teevad geograafid veel "tunde järgi". Matemaatiline statistika aga lubab seesugu-

seid järeldusi teha ranges, tõestatavas vormis. Selleks tuleb kasutada sobivat s t a t i s t i l i s t k r i t e e r i u m i .

Esimeseks sammuks on nn. n u l l h ü p o t e e s i püstitamine. Huvitagu meid näiteks kahe rajooni asulastik. Me arvutasime asula keskmise suuruse mõlemas rajoonis. See osutus mõnevõrra erinevaks. Ent kas me võime väita, et rajoonis A on asulad suuremad kui rajoonis B ? On ju meil tegemist siiski ainult väljavõttega. Isegi kui me uuriksime kõiki asulaid, ei ole meil võimalik määrata nende rahvaarvu kõigil ajamomentidel. Võib-olla on asulate suuruse erinevus tekkinud juhuslikult praegusel momendil, üldiselt seda aga ei esine? Võib-olla on see erinevus tingitud meie andmete ebatäpsusest - isegi rahvaloenduse andmed ei ole täiesti vabad vigadest? Püstitame nullhüpoteesi: asulate keskmine suurus mõlemas rajoonis ei ole oluliselt erinev. Selle hüpoteesi õigsust kontrollime S t u d e n t ' i e. t - k r i t e e r i u m i g a .

Samas näites arvutasime ka asula suuruse standardhälbe, mis rajoonis A osutus suuremaks. Me võiksime teha järelduse, et rajoonis A on asulate suurus ebaühtlasem, rahvastik on rohkem kontsentreeritud kui rajoonis B. Kuid jällegi - kas standardhälbe erinevus ongi oluline või on tegemist lihtsalt juhuslikkusega? Siin on nullhüpoteesiks: mõlema rajooni asulate suuruse standardhälbed on tegelikult võrdsed. Hüpoteesi õigsust kontrollime F i s c h e r i e. F - k r i t e e r i u m i abil.

Toome veel ühe näite. Meil õnnestus suure hulga vaatlusandmete läbitöötamisel saada mingi sõltuvus kahe geograafilise nähtuse vahel, näiteks leida, et linna rahvastiku tihedus D muutub sõltuvalt kaugusest r linna tsentrist vastavalt Clark-Sauškini valemile: $D(r) = D_0 e^{-br}$. Loomulikult ei saa selline valem täpselt kehtida. Võib ju mõnes linnas keskuse lähedal olla kasvõi suur järv või lai jõgi (Riia), kus rahvastiku tihedus on null. Kas aga see valem väljendab tõesti mingit reaalset tendentsi, s. t.

tema vead ei ole olulised? Või oleme surunud lihtsalt faktidele peale skeemi, mis on täiesti vale? Viimasel juhul peavad vead olema olulised. Võtame teatud hulga punkte linnas või linnarajoone, arvutame nende rahvastiku tiheduse ja rahvastiku tiheduse Clark-Sauškini valemi järgi. Saame kaks rida arve. Püstitame hüpoteesi: need read ei erine oluliselt. Hüpoteesi kontrollime χ^2 - k r i t e e r i u m i g a .

On olemas rida statistilisi jaotusi - χ^2-jaotus, Fischeri jaotus, Student'i jaotus. Kõigi kriteeriumide korral on tõestatav järgmine väide: kui nullhüpotees on vale, siis on katseandmetest teatud viisil arvutatav suurus mingi tõenäosusega p suurem kui vastava jaotuse väärtus. Nende jaotuste valemid on väga keerulised, seetõttu on koostatud statistilised tabelid, mida kasutamegi. Tabelist saame arvu, mida χ^2 , F või t antud tõenäosusega ei ületa. Kui meie arvutatud väärtus on suurem tabeli väärtusest, siis sama tõenäosusega on nullhüpotees vale. Vastupidisel juhul võib (kuid ei pea) nullhüpotees olla õige. Sel juhul me ütleme: erinevused ei ole tasemel p olulised.

Vaatleme lähemalt χ^2-kriteeriumi.

2. χ^2 - kriteerium.

Seda kriteeriumi tohib kasutada ainult rühmitatud andmete jaoks. Tema mehhanism seisneb umbes järgnevas: kui kaks rida on sarnased, siis ei tohi ühtede ja samade vahemike sagedused mõlemas reas kuigi palju erineda. Selle juures peab vaatluste arv mõlemas reas olema võrdne. Kui see tingimus ei ole täidetud, tuleb arvutada suhtelised sagedused, mille summa on alati 100 %. Ühegi vahemiku sagedus ei tohi olla liiga väike, vastasel juhul ei ole arvutatud tulemus jaotunud χ^2-jaotuse järgi. Igasse vahemikku peab langema vähemalt 4 - 5 vaatlust,

selleks võime me neid liita, kuna vahemikud ei pea olema võrdsed. χ^2-jaotusel on üks parameeter - vabadusastmete arv f :

$$f = n - r , \qquad (VII.1)$$

kus n on vahemike arv, r aga täiendavate tingimuste arv. Alati on olemas vähemalt üks täiendav tingimus - kahe rea sageduste summa peab olema võrdne. On võimalik nõuda andmetelt teiste täiendavate tingimuste täitmist. Rakendame χ^2-kriteeriumi järgmise näite juures. Meil on olemas rühmitatud andmed kahe rajooni A ja B asulate suuruse kohta (vt. tabel 24). Tekib küsimus: kas asulastik mõlemas rajoonis on sarnane? Teatud erinevusi muidugi on, kuid leida kahte rajooni, kus asulad on täiesti ühesugused, on praktiliselt võimatu. Võib-olla on need erinevused juhuslikud, tingitud sellest, et meie käsutuses ei ole mitte kõige parem väljavõte? Sest isegi kui me arvestasime kõiki rajoonide A ja B asulaid, on meil ikkagi tegemist väljavõttega; me ei ole arvestanud ega saagi arvestada kõiki ajamomente.

T a b e l 24 .

Elanike arv asulas	Asulate arv A. rajoonis	Asulate arv B. rajoonis
1 - 20	15	19
21 - 50	28	31
51 - 100	41	40
101 - 500	24	16
501 - 1000	5	3
2000	-	1
2600	1	-
Kokku	114	110

Et vastata sellele küsimusele, kasutame χ^2-kriteeriumi. Nullhüpoteesiks on meil: kahe rajooni asulastiku jaotus

suurusjärkude vahel ei erine oluliselt, erinevused on tingitud juhuslikkusest.

Kuna andmed on rühmitatud, siis on üks χ^2-kriteeriumi kasutamise tingimus juba täidetud. Kuid sageduste summad kummaski rajoonis on erinevad, peale selle esineb vahemikke, kuhu langeb liiga vähe jaotusi. Liidame kolm viimast vahemikku ja arvutame suhtelised sagedused (vt. tabel 25).

Tabel 25.

Vahemik	Sagedus m_A %	Sagedus m_B %	$m_A - m_B$	$(m_A-m_B)^2$	$\frac{(m_A-m_B)^2}{m_B}$
0-20	13,15	17,27	-4,1	16,81	0,97
21-50	24,56	28,18	-3,6	12,56	0,44
51-100	36,00	36,36	-0,3	0,09	0,00
101-500	21,05	14,55	6,5	42,25	2,90
üle 500	5,26	3,63	1,6	2,56	0,70
Summa	100 %	100 %	0	χ^2 =	5,01

Samas tabelis on arvutatud ka suurus

$$\chi^2 = \sum_{i=1}^{n} \frac{(m_i^A - m_i^B)^2}{m_i^B} . \qquad \text{(VII.2)}$$

Vahemike arv n=5 . Vabadusastmete arv on seega 4, kuna meil on vaid üks täiendav tingimus: sageduste summad peavad olema võrdsed.

Nüüd võime pöörduda χ^2-jaotuse tabeli poole (vt. tabel 28). Selle veergude pealkirjadeks on suurused 1 - p , ridade pealkirjadeks aga vabadusastmete arv f . Seega siis, kui me võtame tõenäosustasemeks 95 % = 0,95, vajame veergu 0,05 ja rida 4. Nende lõikepunktist leiame, et χ^2 väärtus võib olla kuni 9,5. Meie saime χ^2 väärtuseks 5,0. Seega ei saa nullhüpoteesi tagasi lükata. Nähtavasti

on tõesti kahe rajooni asulastik üsna sarnane.

Teise näite laename J. V. Medvedkovi tööst "Экономгеографическая изученность районов капиталистического мира". Selle töö 2. osast on võetud tabel 26. Andmete grupeerimise alus on siin teistsugune, vaatluspunktid on rühmitatud looduslike piirkondade järgi. Sellest hoolimata on tegemist ikkagi vahemikega ning teises ja kolmandas lahtris toodud arvud ei ole midagi muud kui suhtelised sagedused (lihtsuse mõttes on tähistuse m_i^A asemel tarvitatud A_i ja m_i^B asemel B_i).

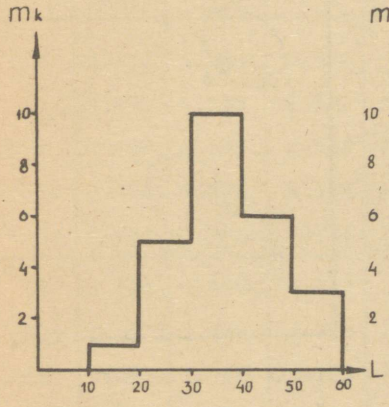
Tabeli pealiskaudne vaatlus jätab mulje, et rahvastik on jaotunud üksikute looduslike piirkondade vahel üsnagi ühtlaselt, sõltuvalt nende piirkondade suurusest. Korrelatsioonisse kahe rea vahel tuleb suhtuda ettevaatlikult, sest tegemist on ainult nelja vahemikuga. Siiski võime arvutada korrelatsioonikoefitsiendi. Saame, et $r_{AB} = 0,83$. See õigustab nullhüpoteesi püstitamist: rahvastiku ja pindala jaotusread ei erine oluliselt. Sisuliselt tähendab see, et rahvastik on ühtlaselt jaotunud üle kogu ala ja looduslike tingimuste erinevusel on üsna väike tähtsus rahvastiku paiknemisel (moodustab juhusliku faktori).

T a b e l 26.

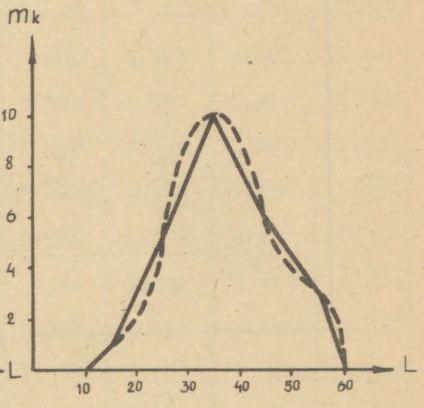
Looduslikud piirkonnad	Elanike arvu % üldarvust (A_i)	Pindala % üldpindalast (B_i)	$A_i - B_i$	$(A_i - B_i)^2$	$\frac{(A_i - B_i)^2}{B_i}$
Rannik	30	20	10	100	5,00
Eelmäestikud	15	10	5	25	2,50
Alumine platoo	40	50	-10	100	2,00
Kõrgplatoo	15	20	- 5	25	1,25
Summa	100 %	100 %	0	$\chi^2 = 10,75$	

Kontrollime nullhüpoteesi jällegi χ^2-kriteeriumi abiga. Kõik selle kriteeriumi rakendamise tingimused on täidetud, sellepärast ei ole ümberarvutusi vaja teha. χ^2 leidmine on kirjeldatud tabeli 26 parempoolses osas. Vabadusastmete arv on ilmselt 3. Pöördume χ^2-tabeli poole. Leiame, et tõenäosusega 95 % ei tohi χ^2 väärtus ületada 7,82. Kuna $10,75 > 7,82$, siis tuleb nullhüpotees tagasi lükata. Loodusliku piirkonna pindala ei ole määrav faktor rahvastiku paigutuse kujunemisel. Otsustavat faktorit tuleb otsida mujalt.

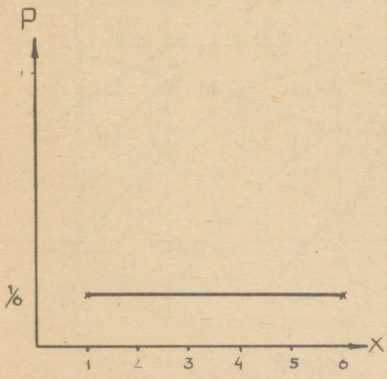
χ^2-kriteeriumi kasutavad geograafid kõige rohkem. Selle üheks põhjuseks on asjaolu, et kriteerium ei esita mingeid nõudeid võrreldavate ridade jaotusfunktsiooni suhtes. F- ja t-kriteeriumide puhul on aga nõudeks, et võrreldavad juhuslikud suurused oleksid jaotunud normaalselt. Seepärast kasutatakse F-kriteeriumi kõige rohkem saadud empiiriliste valemite suhtelise efektiivsuse kontrollimisel. Nullhüpoteesiks on sel juhul väide: ühe valemi abiga saadud tulemused ei ole täpsemad teise valemi abiga saadud tulemustest. See võimaldab valida kasutamiseks lihtsama valemi. Seejuures lähtutakse eeldusest, et valemite vead on jaotunud normaalselt, mis peaaegu alati on täidetud. Muidugi võib F-kriteeriumi tarvitada ka teistel juhtudel, kuid siis peab eelnevalt kasutama veel üht kriteeriumi, mis kontrollib hüpoteesi: antud jaotus on normaalne. Kõigi nende teemade käsitlus käesolevas õppevahendis on aga selle piiratud mahu tõttu võimatu.



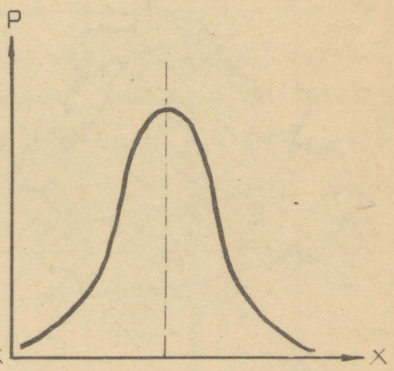
Joon. 1



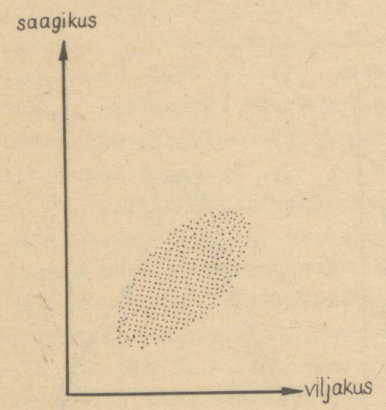
Joon. 2



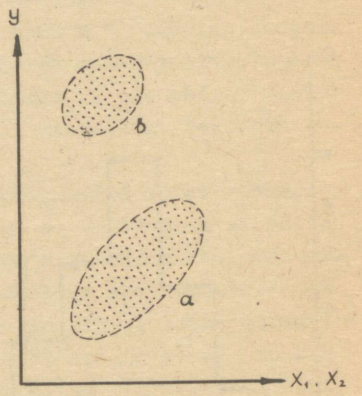
Joon. 3



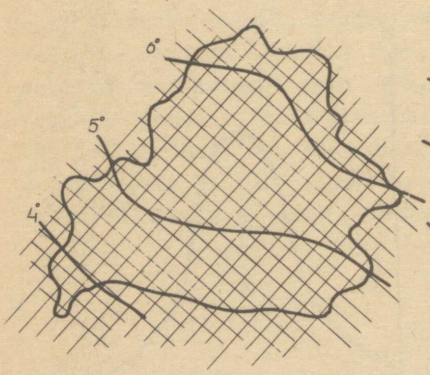
Joon. 4



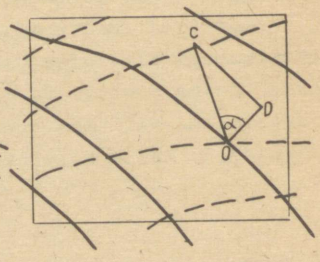
Joon. 5



Joon. 6



Joon. 7



Joon. 8

Tabel 27.

Koefitsiendi t väärtused.

n - 1	Tõenäosustase						
	70 %	80 %	90 %	95 %	98 %	99 %	99,9 %
2	1.336	1.886	2.920	4.303	6.965	9.925	31.60
3	1.250	1.638	2.353	3.189	4.541	5.844	12.94
4	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	1.134	1.440	1.943	2.447	3.143	3.707	5.959
8	1.108	1.397	1.860	2.306	2.896	3.355	5.041
10	1.093	1.372	1.812	2.228	2.764	3.169	4.587
12	1.083	1.356	1.782	2.179	2.681	3.055	4.318
14	1.076	1.345	1.761	2.145	2.624	2.977	4.140
16	1.071	1.337	1.746	2.120	2.583	2.921	4.015
18	1.067	1.330	1.734	2.103	2.552	2.878	3.922
20	1.064	1.325	1.725	2.086	2.528	2.845	3.850
22	1.061	1.321	1.717	2.074	2.508	2.819	3.792
24	1.059	1.318	1.711	2.064	2.492	2.797	3.745
26	1.058	1.315	1.706	2.056	2.479	2.779	3.707
28	1.056	1.313	1.701	2.048	2.467	2.763	3.674
30	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Tabel 28.

χ^2-jaotuse tabel.

f	95 %	99 %	99,9 %	f	95 %	99 %	99,9 %
1	3,84	6,63	10,8	16	26,3	32,0	39,3
2	5,99	9,21	13,8	17	27,6	33,4	40,8
3	7,81	11,3	16,3	18	28,9	34,8	42,3
4	9,49	13,3	18,5	19	30,1	36,2	43,8
5	11,1	15,1	20,5	20	31,4	37,6	45,3
6	12,6	16,8	22,5	21	32,7	38,9	46,8
7	14,1	18,5	24,3	22	33,9	40,3	48,3
8	15,5	20,1	26,1	23	35,2	41,6	49,7
9	16,9	21,7	27,9	24	36,4	43,0	51,2
10	18,3	23,2	29,6	25	37,7	44,3	52,6
11	19,7	24,7	31,3	26	38,9	45,6	54,1
12	21,0	26,2	32,9	27	40,1	47,0	55,5
13	22,4	27,7	34,5	28	41,3	48,3	56,9
14	23,7	29,1	36,1	29	42,6	49,6	58,3
15	25,0	30,6	37,7	30	43,8	50,9	59,7

S i s u k o r d .

Eessõna 3

Sissejuhatus 5

I. Numbrilise informatsiooni saamine ja andmete grupeerimine 6

1. Statistilised kogumid ja väljavõttemeetod. 6

2. Rühmituste olemus ja liigid 12

3. Juhusliku suuruse mõiste. Tähtsamad statistilised jaotused ja nende tähendus 15

II. Nähtuse keskväärtus ja muutlikkuse näitaja .. 18

1. Keskväärtus 18

2. Keskväärtuste tähendus 20

3. Limiidid ja muutumisulatus 22

4. Muutlikkuse näitajad 23

5. Variatsioonikordaja 25

6. $\bar{x}$ ja σ arvutamise lihtsustatud meetodid. 27

III. Korrelatsiooniteooria alused 30

1. Korrelatsiooni mõiste 30

2. Seose tiheduse mõõtmine 31

3. Korrelatsioonikordaja lihtsustatud arvutusviisid 34

4. Spearmani astakkorrelatsioon 36

5. Nähtuste sõltuvuse empiiriliste valemite saamine 38

6. Mitmene korrelatsioon 39

IV. Väljavõtte statistiliste parameetrite vead 41

1. Uurimisvigade liigid 41

2. Väljavõtte aritmeetilise keskmise viga 42

3. Standardhälbe viga 45
4. Väljavõtte korrelatsioonikordaja representatiivsusviga 45
5. Aritmeetilise keskmise vea täpsustamine... 46

V. Kartograafilise materjali statistiline töötlemine 48

1. Statistilise pinna ja selle kaardi mõiste. 48
2. Arvuliste andmete saamine kaardilt 49
3. Statistiliste näitajate määramise visuaal-kartograafiline meetod 55
4. Väljavõtte mahu määramine ligikaudse keskmise ruuthälbe järgi 58

VI. Mitmesugused arvutusmeetodid 60

1. r arvutamine rühmitatud andmete järgi .. 60
2. σ ja r parandamine 62

VII. Statistilised kriteeriumid 64

1. Statistilise kriteeriumi mõiste 64
2. χ^2-kriteerium 66

B. A. Червяков

ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

На эстонском языке

Тартуский государственный университет
ЭССР, г. Тарту, ул. Юликооли, I8

Vastutav toimetaja T. Raitviir
Korrektor E. Oja

TRÜ rotaprint 1970. Paljundamisele antud 18. II 1970.
Trükipoognaid 4,75. Tingtrükipoognaid 4,4. Arvestuspoognaid 3,6. Trükiarv 400. Paber 30 x 42. 1/4.
MB 00441. Tell. nr. 114.

Hind 20 kop.

Hind 20 kop.