**Katseandmete analüüs**

**Loeng 3. Ühefaktoriline dispersioonanalüüs. Dispersioonanalüüsi eeldused.**

**Kirjutas Toomas Tammaru veebruaris 2002, viimati sätistatud septembris 2020; konsultant Ants Kaasik**

**Ühefaktoriline dispersioonanalüüs**

 on t-testi üldistus või ehk täpsemalt on t-test ühefaktorilise **dispersioonanalüüsi** (***analysis of variance*,** siit siis üldtuntud lühend **ANOVA**) erijuht. Seda siis sedamoodi, et dispersioonanalüüsi korral võib võrreldavaid rühmi (või siis manipulatsioonide nn. tasemeid, kui ise loome need võrreldavad rühmad) olla rohkem kui kaks.

 Miks mitte palju t-teste? Esimene mõte võiks ju olla teha rohkem kui kahe võrreldava rühma korral mitmeid t-teste võrdlemaks rühmi paarikaupa ja kui üks selline võrdlus viitab statistilisele olulisuse, et siis olekski nagu tõestatud, et rühmade vahel erinevus on. Siin on aga põhimõtteline probleem: kui me teeme suure hulga teste, siis on parasjagu tõenäoline, et mõned testid ületavad valitud olulisuse nivoo (tavaliselt siis =0.05) juhuslikult. Seega - mida rohkem paarikaupa t-teste teeme, seda suurem on oht, et teeme vale järelduse uuritava erinevuse/ mõju olemasolu kohta olukorras, kus seda mõju tegelikult ei ole (vt hilisemas loengus juttu *post hoc* võrdlustest). Niimoodi siis järelikult ei tohi ja tuleb teistmoodi teha, st ANOVAt teha.

 ANOVA põhineb **dispersiooni komponentideks lahutamisel** (vt seletust 1. loengus). Nii võib dispersiooni jagada rühmade keskmiste (ehk **tasemete**, nagu siinpuhul öeldakse, vt pilti allpool tasemete visualiseerimiseks) dispersiooniks ja **jääkide** (*residual*) dispersiooniks ümber nende keskmiste (**tasemete vaheline ja tasemesisene varieeruvus**, nagu öeldakse), need kaks komponenti annavad kogudispersiooni kokku. Jääk on siis iga üksikvaatluse erinevus oma rühma keskmisest. Vt pilti (siin siis on kaks rühma ja jäägid +1, -2 jne; üks must mumm on üks vaatlus):

 

sõltuva

muutuja

väärtus

 

Loogika siis selline, et uurimaks seda, kas rühmade keskväärtuste vahel on erinevusi, küsitakse “kas rühmade keskmiste dispersioon on seletatav üksikvaatluste dispersiooniga?”. On ju igati tõenäone, et kui vaatlusi kasvõi suvaliselt rühmadeks jagada, siis puhtjuhuslikult on rühmade keskmised natuke erinevad - seda erinevamad, mida rohkem on seda va dispersiooni seal üldkogumis. Idee on siis selles, et kui rühmade keskmiste erinevus on suurem, kui üksikvaatluste dispersiooniga seletada võib, siis loetakse statistika mõttes tõestatuks, et rühmade keskmiste (tasemete) vahel see erinevus tõesti on.

 Seda kõike formaliseeritakse nn F-statistiku arvutamisega:



kus MS=SS/df, ehk siis vabadusastmete arvuga jagatud ruutude summa, nimi on *mean square* ehk keskruut.

 See SS on *sum of squares* ehk jääkide dispersiooni ümber oma rühma keskmise (jääkhajuvust) iseloomustav ruutude summa ehk erinevused keskmisest on ruutu tõstetud ja kokku liidetud (ühel juhul siis iga vaatluse erinevus vastava rühma keskmisest ruutu tõstetud ja teisel juhul igale objektile vastava rühma keskmise erinevus üldkeskmisest ruutu tõstetud, mõlemal juhul on ruudud summeeritud). SS erinevus dispersioonist on seega see, et see on N-ga (valimi suurusega) läbi jagamata. See on tehniline aspekt ja me võime rahulikult mõelda SS-dest kui dispersiooni komponentidest.

Neid MS’e ja SS’e on siis kaks tükki - *model* ehk *between* ja *error* ehk *within*, need siis kirjeldavad neid kahte ülalseletatud dispersiooni komponenti - ehk siis erinevusi rühmade keskmiste (tasemete) vahel ja üksikväärtuste hajuvust ümber rühma keskmise (jääkhajuvus), ehk ametlikuma terminina ka tasemesisene ja tasemetevaheline hajuvus. Ehk siis vaatamata teisendustele on küsimus ikkagi rühmasisese ja rühmadevahelise dispersiooni võrdlemises, küsitakse ikka, kas rühmadevaheline dispersioon on piisavalt suur võrreldes rühmasisese dispersiooniga, st piisavalt suur järeldamaks seda, et rühmade keskmised ka üldkogumis erinevad on.

 Selle F põhjal siis otsustatakse, kas selline pilt on juhusega seletatav või mitte - ehk siis F põhjal leitakse p, nii nagu t järgi t-testi puhul. F ja p on omavahel seotud F-jaotuse jaotusfunktsiooni kaudu (*et keda selline termin huvitab*) neid jaotusfunktsioone on jällegi erinevaid vastavalt vabadusastmete arvule. **Seega on F-st p arvutamisel on oluline vabadusastmete arv, need kirjutatakse tekstis sageli (*siin kursuses kindlasti*) F-i alaindeksitena.** Nagu ka neid MS-e, nii on ka vastavaid vabadusastmeid kaks tükki - mudeli oma ja hälvete oma (vt üldist juttu vabadusastmete kohta loengus 2). Vastavat ülaltoodud valemile (F = …) räägitakse ANOVA kontekstis lugeja ja nimetaja vabadusastmetest (*numerator degrees of freedom* ehk *ndf* ja *denominator degrees of freedom* ehk *ddf*). Mudeli vabadusastmed (mis siis iseloomustavad mudeli keerukust) on määratud võrreldavate rühmade (tasemete) arvu (k) poolt: k-1 ja hälvete vabadusastmete arv vaatluste arvu (n) poolt (n-k). Esimene neist df-dest iseloomustab siis iseloomustab mudeli keerukust, teine andmete hulka, nagu eespool esitatud üldine jutt vabadusastmetest räägibki. Artiklis esitatud F-i indeksite vaatamisest saame seega kiire ülevaate sellest, et kuimitut rühma võrreldi ja palju andmeid oli.

 **Ühefaktoriline** dispersioonanalüüs (*one-way ANOVA*) on selline, kus andmestik on rühmadeks jagatud vaid ühe tunnuse alusel, sellisest siis siiani juttu olnud. Muist võimalusist hiljem. Selle ühe sõltumatu muutuja väärtusi (ehk siis faktori tasemeid, ehk võrreldavaid rühmi) võib olla kuitahes palju, see ei mõjuta seda, et ta jääb ühefaktoriliseks.

 Öeldakse, et ANOVA on **tasakaaluline** (valim on tasakaalus), kui kõikides võrreldavates rühmades on samapalju objekte. Hiljem sellest, kui asi nii pole.

Alljärgnevalt näide programmi (SAS) poolt väljastatud dispersioonanalüüsi tulemustest.



Uuri seda ja saa aru, kus on siin

 - mudeli vabadusastmed (3)

 - jääkhajuvuse vabadusastmed (16)

 - mudeli ruutude summa (25)

 - jääkhajuvuse ruutude summa (40)

 - F’i väärtus (3,33)

 - statistiline olulisus p (0,0461)

 - determinatsioonikordaja R2 (38,5%).

Näeme, et peale SS ja F ja p on tulemustes ära toodud ka selline statistik nagu determinatsioonikordaja R-ruut. Selline näitaja iseloomustab nn mudeli sobivust, ehk ta mõõdab seda, kui suure osa andmetes olevast dispersioonist mudel ära seletab - ehk siis siin tuleb jälle mängu esimeses loengus üleskiidetud dispersiooni aditiivsuse omadus, dispersiooni saab jagada mudeli poolt äraseletatud ja seletamata jäävaks osaks (jääkhajuvuseks). Manipulatiivse katse puhul (ise tekitame need rühmad mingit manipulatsiooni rakendades) on R-ruut tõlgendatav ka manipuleeritava faktori (nt toidutaim, temperatuur vms) mõju tugevuse iseloomustajana.

 R-ruut arvutatakse SSmodel/ SStotal, see viimane (SS-total) on siis kogudispersioon ehk SS-modeli ja SS-errori summa (ehk siis tasemetevaheline ja tasemetesisene dispersioon annab kokku kogudispersiooni). R-ruutu võib tõlgendada ka nii, et see näitab, mitu protsenti jääkide dispersioon väheneb, kui jääke mõõdetakse rühmade keskmiste suhtes võrreldes sellega, kui neid arvutatakse kogu valimi keskmise suhtes. Vaata tahvlile, kuidas erinevad R-ruudu väärtused joonisel välja näevad. R-ruut võib saada väärtusi nullist üheni (e. 100%), mida kõrgem väärtus, seda paremini sobiv mudel.

 Sageli väljastatakse programmide poolt ka kohandatud R-ruudu väärtus (*adjusted R-square*), mis on valimi põhjal antud hinnanguks R-ruudule üldkogumis: valimi R-ruut hindab üldkogumi R-ruutu süstemaatiliselt üle, nii nagu dispersioon valimis hindab alla dispersiooni üldkogumis (vt eespool).

 Ühefaktorilise dispersioonanalüüsi tulemusi esitame tavaliselt “uuritavad järved erinesid ahvenate pikkuse poolest (F 7, 33=7,36, p=0,006)” (kui võrdleme meist sõltumatult olemas olevaid rühmi) või “toidutaime liigil oli mõju nukukaalule (F 3,16=3,33, p=0,046)” (kui võrreldavate rühmade vaheline erinevus on tõlgendatav manipulatsiooni mõju uurimisena). Pane tähele **tõlgendust**: „järved erinesid“ tähendab siin seda, et „kõik kaheksa järve ei ole kala pikkuse poolest ühesugused“, me ei väida, et misiganes kaks järve erinevad üksteisest ahvenate pikkuste poolest: kaheksa hulgas võib olla ka järvi, kus ahvenate pikkused ei erine. Selle uurimisest edaspidi, et milliste konkreetsete järvede vahel on ja kus ei ole erinevusi.

R-ruudu võime lisada, kui see on sisuliselt huvitav - kui huvitab kirjeldada, kui palju mudel varieeruvusest seletab ehk siis kui suur on uuritava (võrreldavaid rühmi moodustava) faktori mõju võrreldes juhuslike hälvete ehk taustamüraga.

 Natuke terminoloogiast veel. ANOVA (ja regressiooni jms selliste analüüside puhul, kus räägime mudelist) puhul saab rääkida mudeli poolt **ennustatud** (*predicted*) **väärtustest**. Ehk siis võib rääkida nii, et mudel ennustab, et sõltumatute muutujate nii- ja naasuguste väärtuste korral on sõltuva muutuja väärtus kõige tõenäosemalt just selline. Ühefaktorilise ANOVA korral on asi lihtne - ennustatud väärtuseks on rühma keskmine. Ehk siis kui võrdleme kalaliike ja sõltuvaks muutujaks on mass, siis massi ennustatud väärtuseks on iga liigi keskmine mass - mingit paremat ennustust ei oska me ju tehagi, kui sõltumatuid muutujaid rohkem pole. Erinevust reaalse (vaadeldud, *observed*) väärtuse ja ennustatud väärtuse vahel nimetatakse siis jäägiks, nagu eespool juba juttu oli. Seda asja pole tegelikult ühefaktorilise analüüsi puhul mingit mõtet nii keeruliseks ajada, aga keerulisemates olukordades on sellel (st ennustatud väärtuse mõistel) juba asjalik sisu.

Siiani läksime andmetele oma analüüsidega julmalt peale pikemalt mõtlemata. Nii tegelikult ei tohi, enne analüüsima hakkamist tuleb **kindlasti** kontrollida, kas analüüsi **eeldused** on täidetud. Lugu on paraku nii, et kogu see eespool seletatud matemaatika annab õige tulemuse vaid siis, kui muutujate jaotused vastavad teatud tingimustele. T-testi (st kindlasti ka t-testi puhul, kuigi seal me seda ei maininud!) ja ANOVA puhul eeldame, et võrreldavate rühmade sees on vaatlused **normaaljaotusega** ja lisaks veel, et nende jaotuste **dispersioonid** on võrdsed.

 Teine eeldus pole väga range - st väike kõrvalekalle ei põhjusta suuri probleeme (dispersioonid peavad suisa kordades erinema, et see tulemust tuntavalt mõjutaks), esimesest saab üle arvutuslikult. Veidi allpool sellest, mis saab siis, kui eeldused on rikutud. Pane aga hoolega tähele, et jutt käib jaotustest rühmade sees, mitte jaotusest mitme rühma peale kokku. St suurte rühmadevaheliste erinevuste korral on üldjaotus muidugi mitme tipuga, aga see ei sega teste tegemast.

Normaaljaotusele vastavust on võimalik statistiliselt testida (st et kas erineb oluliselt või mitte), kuid tavaliselt pole seda ehk niiväga mõtet teha, sest:

 - ANOVA on nö robustne normaaljaotuse eelduste rikutuse suhtes, st mõõdukad hälbed „normaalsusest“ ei mõjuta tulemusi märkimisväärselt - mittemõõdukad hälbed on aga silmaga näha;

 - ega looduses pole ükski asi täpselt normaaljaotusega, seega mõõdab see normaalsuse test (st selle *p*) pigem valimi suurust kui jaotuse hälbimist normaaljaotusest;

 - kui valimi suurus on nii umbes 10, siis peaks see hälbimine normaaljaotusest küll väga suur olema, et erinevust oluliseks saada. Ka üldisemalt, jaotuse kuju puudutavad testid pole eriti tundlikud, asjalikeks testideks on vaja vähemalt hulka kümneid vaatlusi.

Aga – seisukohad ses osas kipuvad erinema ja kui ajakirja toimetaja vms nõuab, siis tuleb muidugi tegeleda jaotuste normaalsuse testimisega. Selleks on omad võimalused, praksis räägitakse.

Seega - vaata pigem silmaga ja võta midagi ette a) ilmse; b) süstemaatilise jama korral. Ette saab võtta a) muutuja teisendamist; b) mitteparameetrilise meetodi rakendamist, c) permutatsioonipõhiseid analüüse. Kohe varsti asjust a) ja b), c) ehk kursuse lõpus.

 Ilmne jama on eelkõige selline, kus mõned üksikud väärtused on väga kaugel. See on küll tõsine probleem - üksikud tugevasti hälbivad punktid võivad kergesti testi tulemused pea peale keerata.

 Süstemaatiline jama on näiteks selline, kui rühmades on jaotused süstemaatiliselt ühele poole asümmeetrilised (*skewed*). Siis võiks näiteks logaritmida (vt kohe allpool) ja rahule jääda võiks siis, kui peale teisendust on mõlemale poole viltuseid umbes samapalju. Teine süstemaatilise jama variant on selline, kus valimite keskmised ja hajuvused on omavahel seoses, enamasti siis nii, et suurema keskmisega kaasneb suurem hajuvus, ka selle vastu võib aidata logaritmimine.

 Aga igal juhul: enne misiganes analüüside tegemist vaata oma andmeid, st tee enda jaoks pilte, millelt näed jaotused ära. Muuhulgas ka sellel põhjusel, et trükiviga (nt koma vales kohas) on kerge sisse jääma ja võib palju pahandust teha.

**Muutujate teisendamine (normaliseerimisteisendused)**

 Muutujate teisendamine on üks võte saavutamaks analüüside eelduste täitumist. Asi käib siis sedamoodi, et sõltuva muutuja mõõdetud väärtused asendatakse miski matemaatilise funktsiooniga neist väärtusist. Selle mõttega siis, et kui sõltuv muutuja ise ei ole normaaljaotusega (st kui just seda tahame, st normaalseks saada, nagu see enamasti on), siis sobivalt valitud funktsioon muutujast võib olla.

 Kõige levinum teisendus on ehk logaritmimine, logaritmimine aitab negatiivse asümmeetriaga jaotuste puhul, ehk siis selliste puhul, kellel on pikk saba paremale. Ökoloogias tundub selline olukord just kõige sagedasem olevat. Peale asümmeetria vähendab logaritmimine ka dispersioonide erinevust. Näiteks sellise jaotuse (ilmne jama):



 saab logaritmimisega selliseks:



millel pole enam häda midagi.

Pane tähele, et logaritm on defineeritud vaid positiivse arvu puhul. Seega kui väärtusi näiteks alates nullist, siis tuleb enne logaritmimist liita väärtustele arv üks.

Logaritminime on efektiivne siis, kui väärtused on nulli lähedal, kui väärtused on vahemikus näiteks 100 kuni 120, on logaritmimise mõju jaotuse kujule väike (vt tahvlile). Siis võib väärtused nullile lähemale tuua, lahutades näiteks kõigist väärtustest 98.

Kui logaritmimise mõju on liiga “tugev” - jaotus hakkab teisele poole viltu olema, võib proovida “pehmemat” ruutjuurfunktsiooni ja muidugi ei pea see astendaja olema just 0,5 (nagu ruutjuure puhul).

Kui pikk saba on hoopis vasakule, võib proovida ruutu tõstmist.

Teisendamine pole siiski päris probleemivaba tegevus, seda sellepärast, et peale teisendamist käib testitav hüpotees rangelt võttes enam mitte keskväärtuste (et kas nood on rühmiti võrdsed), vaid logaritmide keskväärtuste kohta. Võib ilmselt ette kujutada olukorda, kus keskväärtused ja logaritmide keskväärtused erinevad eri suundades. Siiski pole ökoloogia praktikas nii peened probleemid kuigi sageli olulised.

 Olulisem asjaolu on see, et logaritmteisendus muudab interaktsioonide ehk koosmõjude tõlgendust, aga sellest siis edaspidi.

**Mitteparameetrilised meetodid**

 Siiski on olukordi, kus jaotusi ei saa sümmeetriliseks miski teisendusega, näiteks kahe tipuga (bimodaalsete) jaotuste puhul. Sellisel juhul tuleb kasutada nn. **mitteparameetrilisi** meetodeid. Need (vastandina **parameetrilistele**, millest eelpool juttu oli – nimetus tuleb sellest, et testitavad hüpoteesid käivad pidevate jaotuse parameetrite – nt keskväärtus – kohta) on siis sellised, mis ei eelda midagi jaotuse kuju kohta.

 Miks siis mitte kasutada mitteparameetrilisi meetodeid alati ja igal pool? No tollepärast, et

 - nende võimsus (*power, tuleb lähemalt juttu kursuse lõpus*) on sageli (kuid mitte päris alati, sõltub olukorrast) madalam – see tähendab, et parameetriliste meetodite puhul saame enamasti väiksemad erinevused statistiliselt olulisteks; ehk parameetrilised meetodid leiavad tegelikkuses eksisteeriva seose paremini üles

 - parameetrilised meetodid võimaldavad peale olulisuse testimise igasuguseid lisaoperatsioone, näiteks leida usaldusintervalle keskmiste erinevustele jne., mitte-parameetriliste puhul on selliseid võimalusi vähem, eelkõige ei saa me vastust küsimusele, et *kui suur* on see erinevus üldkogumite keskväärtuste vahel;

 ANOVA’t asendavatest mitteparameetrilistest meetoditest esituvad siin kaks:

 - Mann-Whitney U-test (ja sarnane Kruskal-Wallise testi - viimane nimi juhuks, kui võrreldavaid rühmi on rohkem kui kaks) idee on selline, eri rühmadest pärit objektid pannakse ühte ritta ja järjestatakse vastavalt uuritava tunnuse väärtustele. Kui uuritavate üldkogumite vahel erinevust ei ole, on eri üldkogumitest pärit objektid omavahel hästi segunenud. Mida suurem see erinevus on, seda rohkem on ühest populatsioonist pärit väärtused rea ühes otsast. Seda segunemust on võimalik matemaatiliselt iseloomustada ja otsustada, kas “mittesegunemus” on suurem kui seda juhuslikult eeldada võiks. Igal juhul on kogu analüüsi ideoloogia põhimõtteliselt teistsugune.

 - Mediaanitest küsib lihtsalt, kas võrreldavates rühmades on rohkem või vähem vaatlusi, mis jäävad ühele või teisele poole ühise jaotuse (rühmad kokku pandud) mediaani. Saadud tulemust analüüsitakse nagu sagedustabelit.

Tulemuse võime kirjutada “.... leiti erinevus (Kruskal-Wallise test: Z=...., p= .... )”.

Nii ja lisaks sellele olukorrale, kus kasutame mitteparameetrilisi meetodeid seetõttu, et jaotuste normaalsus tingimus on rikutud, võib ja tuleb neid kasutada ka olukordades, kus sõltuv muutuja on **järjestustunnus,** st selline diskreetne muutuja, mille tasemetel (klassidel) on sisuline järjekord (suurem/väiksem, parem/halvem), näiteks on selline muutuja (koolipoisi) hinne. Me saame võrrelda kooli klasse ja küsida, kas a-klassis õpitakse paremini kui b-klassis ja sellisele küsimusele vastamiseks kasutada ANOVA mitteparameetrilisi analooge. Küll ei tohi me järjestustunnuste puhul kasutada tavalist parameetrilist ANOVAt, sest see eeldab ju sõltuva muutuja normaaljaotust, mis on ju pidev jaotus ja mida järjestustunnustel põhimõtteliseltki olla ei saa!

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* jutu lõpp \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*