# Katseandmete analüüs

**Loeng 5. Kahefaktoriline dispersioonanalüüs**

**Kirjutas Toomas Tammaru märtsis 2002, sätistatud 2003-2016, konsultant Ants Kaasik**

Eelseletatud mitme üldkogumi võrdlus on siis mõnikord tõlgendatav nii, et uurime neid rühmi tekitava **faktori** mõju sõltumatule muutujale, see teine variant oli siis selline, kus uurime meist sõltumatult olemas olnud rühmade vahelisi erinevusi. Tõlgenduslikult on need variandid erinevad, statistika poolest mitte.

 Nii võime faktori mõju uurimise näitena uurida vastse toidutaime mõju putuka nukukaalule. Samas võib meid huvitada ka miski **teise sõltumatu muutuja samaaegne mõju** sõltuvale muutujale või siis on see teine muutuja meie tahte vastaselt olemas ja mõjutamas ja asja segasemaks tegemas. Nii pole parata, et mainitud toidutaime mõju uurimisel nukukaalule on kasvatatavad putukad kahest soost ja sugu mõjutab samuti nukukaalusid, faktoreid on seega kaks.

 Olgu olukord selline

|  |  |  |
| --- | --- | --- |
|  | sugu isane | sugu emane |
| taim a | viis arvu | viis arvu |
| taim b | viis arvu | viis arvu |

Kui tahame analüüsida kahe faktori mõju korraga, on tegemist on siis **kahefaktorilise dispersioonanalüüsiga** (*two-way ANOVA*), mille väljund võib välja näha järgmisel viisil (SAS programmi näitel).

Dependent Variable: kaal

 Sum of

 Source DF Squares Mean Square F Value Pr > F

 Model 2 41.63200000 20.81600000 33.68 <.0001

 Error 17 10.50600000 0.61800000

 Corrected Total 19 52.13800000

 R-Square Coeff Var Root MSE kaal Mean

 0.798496 13.12404 0.786130 5.990000

 Source DF Type III SS Mean Square F Value Pr > F

 taim 1 20.00000000 20.00000000 32.36 <.0001

 sugu 1 21.63200000 21.63200000 35.00 <.0001

Tulemus on kombeks kirja panna **ANOVA tabelina**, mis siin eelpool toodud näitest saadud numbritega tehtud on:



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

kuna MS=SS/df, siis neid pole ehk mõtet kirjutada. Sellise tabeli võib muidugi tekitada ka ühefaktorilise ANOVA puhul, aga kuna tol puhul on infot vähem, võib olla libedam tulemused lihtsalt teksti sees esitada, nagu seal juttu oli.

 Pane hoolega tähele, et sõltuvat muutujat siin tabelis ei ole, see tuleb tabeli tekstis ära seletada, et mis muutuja see oli, milles esinevat varieeruvust me uurisime. Teine asi, mida siit tabelist ei näe, on mõju suund – et siis kas isased või emased suuremad olid, see tuleb artikli tekstis jutuks võtta või graafiliselt esitada.

Pane siis veel tähele, et nüüd on meil **kummagi faktori kohta omad** statistikud: p-väärtus muidugi ja ka need muud, nt SS’d. Seega saab iga faktori kohta saab arvutada ka oma R-ruuduga analoogse suuruse SS(effect)/SS(total), mis on siis faktori mõju tugevuse mõõt - väljendab, kui mitu % faktor koguhajuvusest kirjeldab, ehk siis koguhajuvus jagatakse kolmeks osaks – ühe faktori poolt seletatav, teise faktori poolt seletatav ja jääkhajuvus. Miskipärast on seda suhet kombeks tähistada η2 (eeta-ruut), kui see käib ühe faktori ja mitte kogu mudeli kohta (st kogu mudeli oma on siis ikka R-ruut).

 Kui meid kogu tabel ei huvita, võime ikkagi esitada tulemustes vaid ühe faktori mõju, näiteks nii “*the effect of host plant on pupal weight was statistically confirmed F1,17 =32.3, p<0.0001; two-way ANOVA with sex as an additional factor”* - kuid need muud faktorid mis seal mudelis olid tuleb siis ikka kindlasti ära mainida.

 Siin on näitena kummalgi faktoril kaks **taset** (st kahte rühma on jagatud nii sugu kui toidutaim), neid võib aga olla kuitahes palju, sellest ei sõltu see, et mitmefaktoriline see ANOVA on.

Pane kindlasti tähele, et kahefaktorilise ANOVA puhul kummagi faktori kohta saadavad statistikud pole **sugugi samad** kui mõlema faktori kohta eraldi tehtud **ühefaktorilistes** ANOVA-des, st kahefaktoriline ANOVA annab olulist lisainformatsiooni võrreldes olukorraga, kus teeme kaks erinevat ANOVAt, kummagi faktori jaoks eraldi. Seda siis nii, et **kahefaktorilises ANOVAs me uurime ühe faktori mõju olukorras, kus teise faktori mõju on juba arvesse võetud.** Ehk teisisõnu, igasuguse mitmefaktorilise mudeli puhul sõltub iga üksiku faktori kohta käivad näitajad (sh statistiline olulisus) sellest, millised muud faktorid seal mudelis on. Seejuures võib olulisustõenäosus uute faktorite lisamisel nii kasvada kui ka kahaneda.

Sellega aga kahefaktorilise ANOVA võimalused ei piirdu, siin tuleb sisse **interaktsiooni** ehk **koosmõju** mõiste. Interaktsioon väljendab seda, et **ühe faktori mõju sõltub teise faktori tasemest.** St interaktsiooni uurides küsime, kas toidutaime mõju sõltub putuka soost. Võime küsida ka vastupidi, st kas soo mõju sõltub toidutaimest - vahet pole, interaktsioon on *sümmeetriline* asi. Kumbapidi me tulemuse oma töös esitame, see sõltub ainuüksi sellest, et kumb interpretatsioon on meile sisuliselt huvitavam.

 *Pane hästi tähele, et kahe faktori koosmõju olemasoluks* ***ei ole*** *sugugi piisav, et lihtsalt mõlemad faktorid samaaegselt mõjuvad, see ei ole veel kaugeltki koosmõju. Koosmõju* ***ei ole*** *lihtsalt see, et kaks faktorit mõjuvad “koos” ehk “üks faktor mõjub ja teine mõjub ka”!!! Sesosas on eestikeelne termin natuke eksitav.*

Sellistes andmetes on koosmõju:

|  |  |  |
| --- | --- | --- |
|  | sugu isane | sugu emane |
| taim a | keskmine 3 | keskmine 6 |
| taim b | keskmine 6 | keskmine 8 |

sest on ju nii, et toidutaime mõju sõltub soost (b on a-st parem kas kahe või kolme ühiku võrra ja mitte samapalju mõlema soo jaoks) ehk siis teistpidi, soo mõju kaalule on erinev eri taimedel (emased isastest suuremad kas kahe või kolme ühiku võrra vastavalt ja mitte samapalju).

Vastavad ANOVA tulemused oleksid sellised, siin oleme koosmõju “mudelisse sisse võtnud” nagu seda öeldakse, kahefaktorilist analüüsi võib seega teha “koosmõju sisse võttes” või seda “mitte sisse võttes”:

 Source DF Type III SS Mean Square F Value Pr > F

 taim 1 0.05000000 0.05000000 0.08 0.7772

 sugu 1 79.20200000 79.20200000 131.24 <.0001

 taim\*sugu 1 20.00000000 20.00000000 33.14 <.0001

ja nende põhjal tehtud artiklisse sobiv tabel järgmine:



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Pane tähele, et **interaktsiooni puudumine** tähendab siis seda, et faktorite mõjud on **aditiivsed** ehk liituvad, st kui emane nukk on isasest 10 mg **võrra** raskem ühel taimel siis on ta seda samapalju ka teisel taimel kasvanuna. St siis võrra, aga mitte korda!

 Interaktsiooni on asjalik kujutada nii, et kanname graafikule sõltuva muutuja väärtused (lahtrite keskmised) sõltuvana ühe faktori tasemetest ja ühendame joonega punktid, mis vastavad teise sõltumatu muutuja võrdsetele väärtustele. Vaata tahvlile! Kui sedaviisi saadud jooned on paralleelsed, pole andmetes koosmõju. Koosmõju on eriti ilmne, kui jooned lõikuvad, st ühe faktori mõju on vastupidine teise faktorite eri tasemetel (“suuna muutusega koosmõju”). **Lõikumine pole siiski vajalik koosmõju olemasoluks, piisab sellest, et jooned pole paralleelsed** (“suuna muutuseta koosmõju”).

 Ja veel pane tähele, et logaritmteisendus muudab aditiivse mudeli multiplikatiivseks, ehk siis kui logaritmitud sõltuva muutuja korral tähendab interaktsiooni puudumine, et iga emane on isasest raskem samapalju **kordi** toidutaimest sõltumatult. Seda siis sellepärast, et logaritm korrutisest on tegurite logaritmide summa. Tuleb endale seega aru anda, et interaktsiooni olemasolu või puudumine võib sõltuda teisendusest. Seda siiski eelkõige siis, kui vaid ühe faktori mõjud ei muuda teise faktori tasemest sõltuvalt suunda, mõju suuna muutmisega interaktsiooni on teisendusega raske mõjutada.

 Olulise koosmõju korral võivad faktorite **peamõjud** (*main effects*, vastandina interaktsioonile; st soo ja taime mõjud ülalnäites) olla ise olla olemas (ja statistiliselt olulised), aga võivad seda ka mitte olla. Ülaltoodud näites tähendaks siis soo peamõju puudumine seda, et üle kahe taime kokku arvutatuna on emased ja isased keskmiselt sama kaaluga. Kui samaaegselt on olemas sugu\*taim koosmõju, tähendab peamõju puudumine seda, et isased on ühel taimel täpselt samapalju emastest suuremad kui palju nad on teisel taimel emastest väiksemad. Taime peamõju puudumine tähendaks seda, et üle kahe soo arvutatud keskmine kaal on eri taimedel kasvanud putukatel täpselt sama.

 Pane tähele, et miski faktori peamõju olulisus ei tähenda tingimata seda, et tema mõju suund on kõigil teise faktori tasemetel sama. St siis ülalesitatud näites ei pruugi faktori “sugu” olulisus tähendada seda, et soo mõju on mõlemal taimel samapidine - nt kui ühel taimel on emased palju suuremad kui isased, teisel taimel aga natuke väiksemad, siis võib faktoril “sugu” ikka ilmneda peamõju: st emased on küll üldiselt suuremad, kuid sellest veel ei järeldu, et nad seda igal taimel on.

Vaata loengu slaide ja ole kindel, et oskad tekitada olukordi, kus on ja ei ole koosmõju ja on ja ei ole peamõju!

 **Koosmõju** olemasolul tuleb seega olla väga ettevaatlik **peamõjude tõlgendamisega** ja tõlgendus sõltub sellest, mis meid sisuliselt huvitab. Olgu lugu nii, et toidutaimel B on tugev positiivne mõju emaste nukukaalule (võrreldes siis standardiks võetud miski A-ga) ja nõrk negatiivne mõju isaste nukukaalule. Olgu siis nii, et saame toidutaime mõju peamõju oluliseks. Kui meid sisuliselt huvitab, kas keskmine nukukaal on suurem B-metsas kui A-metsas (näiteks miski linnu toidubaasi seisukohalt), siis võime seda olulisest peamõjust rahulikult järeldada ja koosmõjule tähelepanu mitte pöörata. Kui meid aga huvitab toidutaime mõju füsioloogiline tagamaa, siis ei saa me kuidagi koosmõjust üle ega ümber oma tulemuste esitusel ja arutamisel. Ehk siis peame asja arutama sugude kaupa eraldi.

 Ehk siis üldiselt igal juhul: koosmõju olemasolul uuri ühe koosmõjus osaleva faktori mõjusid teise faktori tasemetel eraldi. St soo mõju uurimiseks tee ühefaktorilisi ANOVAsid iga taime jaoks eraldi, vaid nii võid kindel olla, et ei tee sisulist viga oma kahefaktorilise ANOVA tulemuste interpreteerimisel.

 Rohkem kui kahefaktoriline ANOVA on muidugi ka väga võimalik, faktoreid võib põhimõtteliselt olla kuitahes palju **(*multi-way ANOVA*).** Nii võime ühe katse tulemuste põhjal uurida, kas nukukaal sõltub peale toidutaime ja soo ka rööviku värvist ja augu olemasolust purgi kaane sees. Põhimõtted ja tõlgendused on põhimõtteliselt samad, oluliselt keerulisemaks läheb hulgafaktorilise analüüsi puhul aga koosmõjude interpreteerimine. Neljafaktorilise puhul tuleb meil ju võimalikud üks nelja faktori koosmõju, neli kolme faktori oma, kuus kahe faktori oma. Vt näidet:

**4-way ANOVA kõik koosmõjud:**

 Source DF Type III SS Mean Square F Value Pr > F

 taim 1 5.0041161 5.0041161 4.83 0.0293

 sugu 1 609.9239388 609.9239388 588.16 <.0001

 taim\*sugu 1 159.7450216 159.7450216 154.05 <.0001

 varv 1 6.6006350 6.6006350 6.37 0.0125

 taim\*varv 1 12.8614007 12.8614007 12.40 0.0005

 sugu\*varv 1 0.1566191 0.1566191 0.15 0.6980

 taim\*sugu\*varv 1 0.0470915 0.0470915 0.05 0.8315

 auk 1 0.3428916 0.3428916 0.33 0.5660

 taim\*auk 1 0.5017713 0.5017713 0.48 0.4876

 sugu\*auk 1 2.2543296 2.2543296 2.17 0.1421

 taim\*sugu\*auk 1 0.0663268 0.0663268 0.06 0.8006

 varv\*auk 1 0.7646613 0.7646613 0.74 0.3916

 taim\*varv\*auk 1 3.4534288 3.4534288 3.33 0.0696

 sugu\*varv\*auk 1 3.4380134 3.4380134 3.32 0.0703

 taim\*sugu\*varv\*auk 1 0.0000439 0.0000439 0.00 0.9948

Ehk siis palju ja asi keeruline. Kolme faktori interaktsioonidest edaspidi, kõrgemateni siin ei jõua.

 Nii, regressioonanalüüsi puhul oli juttu, et selle matemaatiline põhimõte on dispersioonanalüüsi omale väga sarnane. See sarnasus on suisa selline, et **diskreetseid** ja **pidevaid** muutujaid võib kasutada läbisegi sama ANOVA mudeli sõltumatute muutujatena, pidevaid nimetatakse sel puhul **kovariaatideks** (ehk kovariantideks) ja kogu analüüsi **kovariatsioonanalüüsiks** (ANCOVA). Kõigepealt näide: huvitagu tigudega toitmise mõju linnu muna suurusele. Samas on aga teada, et muna suurust mõjutab oluliselt ka vanemlinnu suurus. Olen teinud mänguandmestiku, kus me tigudega toitmise mõju oluliseks, kui analüüsime asja ühefaktorilise ANOVA-ga (toitmine – ei või jah – ainus kahetasemeline sõltumatu muutuja):

 Sum of

 Source DF Squares Mean Square F Value Pr > F

 Model 1 10.22727273 10.22727273 2.96 0.1008

 Error 20 69.09090909 3.45454545

 Corrected Total 21 79.31818182

 R-Square Coeff Var Root MSE pikkus Mean

 0.128940 40.48524 1.858641 4.590909

 Source DF Type III SS Mean Square F Value Pr > F

 trea 1 10.22727273 10.22727273 2.96 0.1008

 kovariaadi (st kehakaalu) kaasamisel ja samade andmetega aga küll: /algus puudu/

 Source DF Type III SS Mean Square F Value Pr > F

 kaal 1 52.04090909 52.04090909 57.99 <.0001

 trea 1 10.22727273 10.22727273 11.40 0.0032

Vaata loengus esitatud illustratsiooni asja kohta:



Tulemuse esitame samasuguse ANOVA tabeli kujul nagu eespool näidatud, kovariaadi df on alati 1.

 Miks siis kovariaadi lisamine tõstis manipulatsiooni mõju statistilist olulisust? Tuletame meelde, et F-statistik arvutatakse F=MS(model)/MS(error), ehk siis mida suurem on jääkhajuvus (mudelis poolt seletamata jääv dispersioon), seda väiksem on F ja seda vähem lootust on mudeli statistilisele olulisusele. Kovariaadi kaasamine vähendas seletamata jäävat dispersiooni ja seega suurendas F-statistiku väärtust - teisisõnu, kovariaat võttis suure hulga juhuslikku hajuvust enda arvele ja tegi pildi oluliselt selgemaks: enne jagasime dispersiooni kaheks komponendiks, nüüd aga kolmeks, aga manipulatsiooni mõju testimisel kasutame ikka neid kahte, mis ennegi. Ehk võib-olla kõige intuitiivsemalt väljendudes - eemaldas kehakaalu segava mõju. **Tüüpilisel juhul kaasamegi kovariaate olukorras, kus nende endi mõju meid ei huvitagi, st kaasame vaid muude, sisuliselt huvitavate faktorite mõju selgemaks tegemiseks** (see ei pruugi muidugi nii olla – kovariaadi enda mõju võib ka huvitada).

 Selliseid **kovariaate** (ja ka diskreetseid lisafaktoreid, mida sel puhul ei nimetata kovariaatideks vaid pigem lisafaktoriteks - *aga mingit põhimõttelist vahet pole*, v.a. mis interaktsioonidesse puutub, aga sellest hiljem) **võib** pildi selgemaks tegemise eesmärgil **kaasata** **palju kulub**, kuid siiski on siin oma lõivsuhe:

 - mudeli keerulisemaks ajamine vähendab analüüsi võimsust iga üksiku faktori suhtes (vähendab hälvete vabadusastmeid). Näitlikustamaks: mida keerulisem on mudel, seda suurem on oht saavutada sobivus puhtjuhuslikult - statistiline analüüs võtab seda arvesse. Hälvete vabadusastmete vähenemine uute faktorite kaasamise tõttu on seda suurema tähtsusega, mida väiksem on valim. Seega tuleb kovariaatide ja lisafaktorite kaasamisel kaaluda, kas kasud või kahjud olulisemaks osutuvad.

 **Üldjuhul ei kaasata mudelisse kovariaati, mis ise ei osutu oluliseks.** Siiski, muarust võib kaasata sellise muutuja, mille mõju uuritavale muutujale on teada muudest katsetest ja mis lihtsalt näiteks ebapiisava valimi suuruse tõttu antud juhul ei saavuta olulisust.

 Suvaliste mitteoluliste faktorite kaasamisega on ka teatud eetiline probleem - kui meid huvitava efekti olulisus on veidi alla vajaliku piiri, siis hulka juhuslikke kovariaate läbi proovides võime kergesti saavutada selle, et mõnes mudelis saame p-väärtuse juhuslikult teisele poole piiri. See pole aga aus võte enam mõistagi. Aga muideks - kui sisuliselt tähtis tulemus hakkab sõltuma mitteoluliste kovariaatide kaasamisest või mittekaasamisest, on see muarust selgeks viiteks selle kohta, et valim on liiga väike ja soovitaksin statistilise akrobaatika asemel pigem uuesti metsa minna ja andmeid juurde koguda.

 Üks viis veidi süstemaatilisemalt otsustada eri faktorite mudelisse kaasamise või väljajätmise üle on (kovariaatide või lisafaktorite) *backward elimination* (“tagurpidi eemaldamine”). Sellise protseduuri käigus kaasame mudelisse alguses kõik asjassepuutuvad faktorid ja hakkame neid **ühekaupa** välja jätma alates kõige vähem olulisest: ehk siis teeme **uue analüüsi** nii, et selles on nüüd ühe võrra vähem sõltumatuid muutujaid. Sellise tegevuse (**mudeli lihtsustamise**) lõpetame siis, kui mudelisse pole jäänud enam ühtegi statistiliselt mitteolulist efekti. Miks ühekaupa - aga sellepärast, et ühe faktori olulisus võib sõltuda teise (ka mitteolulise) efekti olemasolust ja kaks korraga elimineerides võtame selle riski, et elimineerime sel viisil kogemata efekti, mis ühekaupa lähenedes oleks oluliseks jäänud ja seega lõpliku mudeli koosseisu kaasatud.

 Mudeli lihtsustamisele on tänapäeval kasutada ka teistsugune lähenemisviis, informatsioonikriteeriumidele põhinev selline (Akaike), sellest lähemalt viimases loengus, siin lihtsalt mainitud saamise tarbeks.

 ANCOVA puhul saab arvutada diskreetsete sõltumatute muutujate eri tasemetele vastavad sõltuva muutuja sellised keskmised, milles kovariaadi (segav - kui asjakohane) mõju on eemaldatud. Ehk siis ülaltoodud näites - millised oleksid tigudega söödetud ja söötmata lindude muna pikkuse keskmised, kui kovariaadi - antud juhul siis kehasuuruse - väärtused oleksid neil kõigil samad. Nendele ***least square means***idele võib arvutada ka standardvead, selliseid asju sobib ka pildi peale panna.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*