EMPÜ Matemaatika Instituut

E. ARUVEE, U. ENGSTRAND, U. OLSSON

Biomeetria

TARTU 2001

EPMÜ Matemaatika Instituut

E. ARUVEE, U. ENGSTRAND, U. OLSSON

Biomeetria

TARTU 2001

Biomeetria

Toimetaja N. Veske

Käsikirja ettevalmistamist toetas Rootsi Instituut
(grant 2080/1999 (380/98))

ISBN 9985-882-90-3

Eessõna

Käesolev õpik valmis koostöös Rootsi Põllumajandusülikooli kolleegidega Ulf Olssoni ja Ulla Engstrandiga ning Rootsi Instituudi grandi toetusel. Tuginetud on Rootsi Põllumajandusülikooli loengute materjalidele ja pikaajalistele kogemustele matemaatilise statistika õpetamisel.

Esitatud materjal on raamatu "Matemaatiline statistika" loogiliseks jätkuks.

Raamatus käsitletakse katseplaneerimise küsimusi ning andmeanalüüsi erinevaid meetodeid. Näidatakse, kuidas leida statistikute väärtusi käsitsiarvutamisel ja programmipakette kasutades. Teksti on illustreeritud statistikapakettide Minitab ja SAS väljatrükkidega.

Tänan kolleeg Jüri Vardjat käsikirja läbilugemise ja tehtud kasulike märkuste eest. Raamatu eestikeelse käsikirja valmimise eest kuulub eriline tänu toimetaja Nora Veskele, kelle tähelepanelik töö oli raamatu valmimisel hindamatuks abiks.

Tartus, märts 2001. a Eve Aruvee

Sisukord

Peatükk 1

Katseplaneerimise põhimõisted

1.1 Kuidas katset planeerida?

Katsed on kallid. Iga mõõtmine on seotud kulutustega, mis väljendub ajas ja rahas. Paljud bioloogilised katsed on nn võrdlevad katsed. Igas niisuguses katses soovime võrrelda erinevaid töötlusi, näiteks liike, toitumisskeeme jne. Katseplaneerimise eesmärk on koguda niipalju asjakohast informatsiooni kui võimalik ja seda võimalikult madala hinnaga. Katseplaneerimise kahe meetodi vahel valides võib esineda olukord, kus mõlemad meetodid annavad samatasemelisi tulemusi, kuid erineva hinnaga. Sel juhul eelistame odavamat meetodit. Siin on ka eetiline aspekt: korralikult planeeritud katses, kus uuritakse loomi, võib vajaminevate loomade arv olla väiksem.

Paljud inimesed usuvad intuitiivselt, et peamine faktor, mis katse efektiivsust mõjutab, on vaatluste arv. Kuid on teisi faktoreid, mis mõjutavad efektiivsust, näiteks plokk-katsed, mida käsitleme hiljem. Plokkimine tähendab, et võrdleme erinevaid töötlusi katseühikutest moodustatud plokkide vahel, mis on võimalikult lihtsad. Näiteks sigade erinevaid toitumisskeeme võib võrrelda sigade pesakondade vahel. Sel viisil on bioloogilise varieeruvuse mõju nõrgendatud ja saame efektiivsema katse. Teiselt poolt võime saada oskamatu katse, kui raiskame suure hulga ressursse mõõtmistele, kus varieeruvus on väike. Sageli on vähem efektiivsem teha 2 mullaproovi ja igale proovile 10 keemilist analüüsi, kui teha 10 mullaproovi ja igale 2 analüüsi. Seepärast, et sageli on keemiliste analüüside viga väike, aga mullaproovide vahel võib olla varieeruvus suur.

See tähendab, et vaatlustel on erinev väärtus, mis sõltub katseplaanist. Järelikult plaan mõjutab andmete analüüsi.

1.2 Terminid ja definitsioonid

Erinevatele töötlustele (erinevad väetised, temperatuurid jne) valime katsetes objektide hulga. Neid nimetatakse faktoriteks. Faktoreid tähistame tähtedega näiteks A. Faktori, mis enamasti on nominaaltunnus, erinevaid väärtusi nimetatakse tasemeteks (nivoodeks). Tähistame faktori A tasemete arvu a-ga.

Näited.

1. Soovime võrrelda vasikate erinevaid toitumisskeeme. Faktor A on toit, millel on kolm taset ($a = 3$): A_1 - kõrge proteiini tase, A_2 - normaalne tase ja A_3 - madal tase.

2. Soovime võrrelda tomatite väetamiseks kasutatvat nelja ökoloogilist väetist. Väetis on faktor, millel on neli taset ($a = 4$): A_1 - kompost, A_2 - saepuru, A_3 -hein ja A_4 - kontrollväetis.

Katseühik on väikseim ühik, mis esindab individuaalset töötlust.

Näited.

1. Kui vasikaid on söödetud individuaalselt, siis on vasikas kasteühik. Kui kõiki vasikaid ühest latrist on toidetud koos, siis on latter katseühik.

2. Kui igat tomatitaime on väetatud individuaalselt, siis on taim katseühik. Kui on väetatud 10-ne taimega kasti, siis on kast katseühik.

Katsed võivad üheaegselt sisaldada mitmeid faktoreid. Sel juhul nimetatakse neid mitmefaktorilisteks katseteks.

Näide. Vasikate toitumiskatses on üks faktor A=toit 3 tasemega (a=3) ja teine faktor B= laut 10 tasemega (b=10).

Faktor võib olla fikseeritud või juhuslik, sõltuvalt valitud faktori tasemetest. Faktor on fikseeritud, kui katse planeerija valib faktori tasemed ja teostab katse ainult teda huvitavate taseme väärtustega. Faktor on juhuslik, kui tasemed valitakse juhusliku valimina faktori võimalikest tasemetest.

Näide. Mingis põllumajanduslikus katses on vaadeldud lautu. Sageli on kasulik käsitleda lautu juhusliku faktorina, mis on võetud lautade suurest üldkogumist.

Tüüpiline fikseeritud faktor on sugu, liik, toit, väetis jne. Tüüpiline juhuslik faktor on laut, kari, alus. Juhusliku faktorina käsitletakse mõnikord aastat.

1.3 Mõned võtmeküsimused katseplaneerimisel

1.3.1 Replikatsioon (kordus)

Näide. Soovime selgitada, kumb kahest maisisordist annab kõrgema toodangu. Kasvatame sorti A ühel põllul ja sorti B teisel põllul. Osutub, et sort A andis kõrgemat saaki. Kas võib selle informatsiooni põhjal teha üldistusi teise sordi kohta? Kas võib soovitada kõigil teistel põllumeestel julgesti kasvatada sorti A?

Tõenäoliselt mitte. Ühel põllul kasvanud saak ei ole piisav tõendus. Me vajame rohkem andmeid enne, kui võime anda üldiseid soovitusi sordi A kasutamiseks. Niisuguse tõenduse võime saada, kui kordame katset mitmeid kordi. Sellist protsessi nimetatakse replikatsiooniks.

1.3.2 Millal on tegemist replikatsioonidega?

Näide. Soovime võrrelda kahte erinevat sorti maisitoodangut. Katses kasutatakse kahte põldu. Selleks, et saada replikatsioone, jagatakse kumbki põld kaheks osaks. Sorti A kasvatatakse ühe põllu mõlemal osal ja sorti B kasvatatakse teise põllu mõlemal osal. Kas me praegu teeme katset replikatsiooniga?

Näide. Soovime hinnata kindlas piirkonnas keskmist orgaanilise aine sisaldust mullas. Kui palju reaalseid replikatsioone on meil järgmistel juhtudel?

Võtame ühe mullaproovi ühelt alalt. Teeme 200 analüüsi sellele proovile.

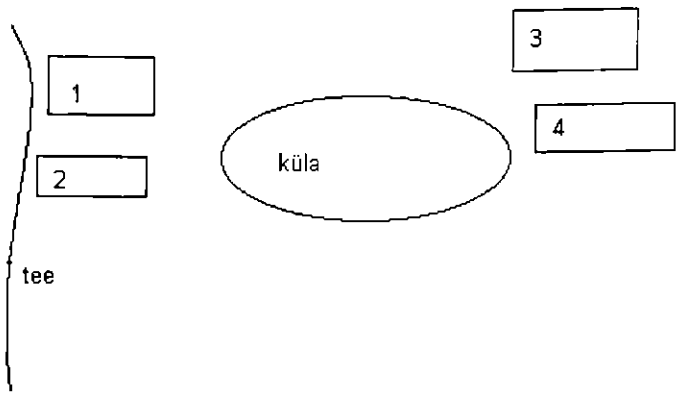
Võtame ühe mullaproovi 10-lt erinevalt alalt. Teeme 20 analüüsi igale proovile.

Valime piirkonnast 10 ala suurusega 10x10 m. Võtame igast ruudust 10 mullaproovi. Teeme 2 analüüsi igale mullaproovile.

Valime piirkonnast 10 ala suurusega 10x10 m. Võtame igast ruudust 10 mullaproovi. Segame iga ruudu 10 mullaproovi. Teeme 2 analüüsi igale segatud mullaproovile.

1.4 Lokaalne kontroll

Näide. Soovime võrrelda kahte sorti maisi. Meil on kasutada katses neli põldu. Kaks põldu paiknevad koos peatee ääres ja kaks põldu paiknevad koos mingil kaugusel külast:



Kuidas valime sortidele A ja B põllud 1, 2, 3 ja 4?

Üks võimalus on kasvatada sorti A kahel põllul, mis paiknevad tee ääres (põllud 1 ja 2) ja sorti B kasvatada kahel ülejäänud põllul. Aga on mõned põhjused, miks see ei ole hea. Kui osutub, et sort A annab suuremat toodangut, siis võib see olla tingitud sellest, et teeäärne muld on parem. Ei tea kindlalt, kas tulemus sõltus mullast või sellest, et sort A oli parem.

Parem alternatiiv on kasvatada sorti A ühel teeäärsel põllul ja sorti B teisel. Sarnaselt kasvatame sorti A põllul, mis paikneb külast eemal (põllud 3 või 4) ja sorti B teisel külatagusel põllul. Sel viisil on sortide A ja B võrdlemine vähem mõjutatud võimalikest mulla erinevustest. See võib anda rahuldava tulemuse kahe sordi võrdlemiseks. Kasutasime lokaalset kontrolli.

1.5 Randomiseerimine

Võime kasvatada sorti A põldudel 1 ja 2. Kuidas otsustame, missugust põldu kasutada?

Paljudele uurijatele ei näi see küsimus tähtsana. Uurija valib vabalt ühe põllu sordile A ja teise põllu sordile B. Ta kirjutab oma aruandes, et põllud valiti juhuslikult. Aga kas see "juhuslik" valik on tegelikult juhuslik?

Meil kõigil on teatud kalduvusi - isegi kui me ei ole alati teadlikud nendest. Agronoom võib tunda, et üks põldudest on rohkem sobilik sordile A ja võib määrata vastava põllu. Aga see võib hävitada erapooletu katse iseloomu. Sortide A ja B vaheline võrdlus võib muutuda mõjutatavaks.

Selle probleemi lahendamise ainus võimalus on määrata sordid A ja B põldudele 1 ja 2 nii, et protsessi jooksul ei saa uurija seda kontrollida – läbi randomiseerimise. Selle kehtestamiseks põllul on palju võimalusi:

- Visata münti. Kui "vapp" on peal, kasutame põldu 1 sordile A. Seda meetodit saab kasutada ainult juhul, kui on kaks töötlust.
- Kirjutada töötluste nimed paberitükkidele, segada paberid ja võtta pimesi üks paber. Seda meetodit võib kasutada hoolimata töötluste arvust.
- Kasutada juhuslike arvude tabelit.
- Kasutada juhuslike arve, mis on genereeritud arvuti poolt.
- Kasutada arvutipakettide Minitabi või SAS-i katseplaneerimise osa.

Üldine nõuanne. Peab jälgima, et võimalikult palju varieeruvust oleks kaasatud lokaalsesse kontrolli (plokkimisse). Näiteks ei ole tark määrata sordid A ja B põldudele 1, 2, 3 ja 4 täiesti juhuslikult, sest siis võib juhtuda, et määrame sordi A põldudele 1 ja 2 ja sordi B põldudele 3 ja 4. Nagu eelpool märkisime, võib see viia ebaõigele sortide A ja B võrdlemise tulemuseni.

Seega on vaja kõigepealt paika panna katse plaani üldine struktuur, seejärel kasutada randomiseerimist, et otsustada, missugust töötlust kus kasutada.

1.5.1 Tasakaal ja täielikkus

Tasakaalus katse tähendab, et katses on igale töötluse kombinatsioonile mõõdetud võrdne arv vaatlusi. Kui see nii ei ole, siis katse on tasakaalustamata.

Kui mingil töötlusel kombinatsioon puudub, siis nimetatakse katset mittetäielikuks.

Näide. A ja B on kaks faktorit, tähistame katseühiku X-ga.

	A_1	A_2	A_1	A_2	A_1	A_2
B_1	XXX	XXX	XXX	XX	XX	XXX
B_2	XXX	XXX	XXX	XXX		X
B_3	XXX	XXX	XX	X	XXX	XX
	Tasakaalus		Tasakaalustamata		Mittetäielik	

Üldine reegel - tasakaalus katset on kergem analüüsida ja interpreteerida.

1.6 Mõned näited katseplaneerimisest

1.6.1 Töötluse struktuur ja plaani struktuur

Katses on kasulik vahet teha töötluse struktuuri ja plaani struktuuri vahel.

Töötluse struktuur näitab, missugune töötlus või töötluse kombinatsioon on katses huvifookuses. Näiteks kui soovime võrrelda nelja sorti maisi, siis sort ongi töötlus. Teistes näidetes võib olla töötlustena kasutatud väetise hulka; piimakarja toitumisskeemide tüüpe; heinasaaki aastas.

Plaani struktuur näitab, mil viisil püüame lokaalset kontrolli saavutada. Täielikult randomiseeritud plaanis ei ole lokaalne kontroll kasutatav. Erinevat tüüpi randomiseeritud plokkplaanis, ladina ruutude plaanis, liigendatud plokkide plaanis jne peab arvestama juhusliku varieeruvuse erinevaid viise. Järgnevalt vaatleme neid plaane lähemalt.

1.6.2 Täielikult randomiseeritud plaan

Üks faktor

Näide. Võrdleme kolme sorti maisi: A, B ja C. Meie käsutuses on põld, mida saab jagada kuueks plokiks. Järelikult saame kasvatada igat sorti maisi kahel plokil (replikatsioon). Kui määrame sordid täiesti juhuslikul viisil põllulappidele, võib tulemus olla näiteks järgmine:

B	A	A
A	C	C
C	B	B
C	A	B

Kaks või enam faktorit

Näide. Soovime piimakarjas kontrollida, kuidas kaks toiduskeemi A_1 ja A_2 mõjutavad piimatoodangut. Lehmad on kahte tõugu: B_1 ja B_2. Järelikult on tegemist nelja töötluse kombinatsiooniga: A_1B_1, A_1B_2, A_2B_1 ja A_2B_2. Kui meil on neli lehma igast tõust, siis üks viis katse planeerimiseks on juhuslikult valida kaks lehma igast tõust toituma toiduskeemiga A_1 ja kaks toiduskeemiga A_2. Moodustub täielikult randomiseeritud kahefaktoriline plaan.

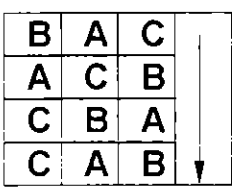
Sellist tüüpi plaanides võib üldiselt olla rohkem kui kaks faktorit. Näiteks põlluvilja väetamise katses võib lisada faktorid N, P ja K. Kui igal faktoril on kolm taset (0, keskmine ja kõrge tase), siis töötluste kombinatsioonide arv on $(3 \quad 3 \quad 3) = 27$ Kui soovime katses igat kombinatsiooni korrata kaks korda, siis vajame kokku 54 plokki. Näeme, et kui suurendame töötluste arvu, siis töötluste kombinatsioonide arv kasvab väga kiiresti. See sunnib lihtsusele: püüda hoida igas katses töötluste arv väike ja kui see pole võimalik, püüda vähendada tasemete arvu. Näiteks väetamise katse võib olla palju väiksem, kui igal faktoril on ainult kaks taset. Sel juhul saame ainult $(2 \quad 2 \quad 2) = 8$ kombinatsiooni.

Võime küsida: "Miks peame ühes katses vaatlema mitmeid faktoreid, selle asemel, et teha katse igale faktorile?" Üks vastus on, et mitmefaktoriline katse on odavam, kui mitmete üksikute katsete tegemine. Tähtsam põhjus on siiski selles, et mitmefaktorilised katsed võimaldavad uurida faktoritevahelisi koosmõjusid.

1.6.3 Randomiseeritud plokkidega katseplaan

Täielikult randomiseeritud plaanis lokaalset kontrolli ei kasutata. See võib põhjustada suurt bioloogilist varieeruvust. Sageli on kasu, sellest kui katse ühikute vahel esinevad erinevused hõlmata plokkidesse.

Näide. Eeldame, et põllul on kallak noole suunas (vt järgnevat joonist). Noole lõpupool olevatel maatükkidel on paremate omadustega pinnas. Seda fakti peab arvestama katseplaani täiustamiseks, et katse avastaks töötlustevahelisi erinevusi selgemalt (ja odavamalt!).



Selles plaanis esineb iga töötlus üks kord igas reas. Oleme randomiseerinud kolm töötlust igale reale. Ja kuna eeldame, et read on ühesugustes tingimustes, siis töötlustevaheliseks võrdluseks oleme kasutanud lokaalset kontrolli, vähendamaks juhuslikku varieeruvust.

Näide. Plokkimist saab kasutada ka loomadega seotud katsetes. Uurija saab grupeerida loomad plokiks, et oleks lihtsam. Informatsioon, mida plokkides kasutada saab, võib olla näiteks: vanus, eelmise laktatsiooniperioodi piimatoodang, kaal jne. Üldiselt on kasulik, kui plokkide moodustamise tunnused on seotud katses uuritavate tunnustega.

Eeldame, et loomad on järgmiste karakteristikutega:

Lehma nr	Tõug	Eelmise laktatsiooni perioodi toodang	Vanus
1	B_1	4,6	4
2	B_1	2,4	8
3	B_1	1,9	12
4	B_1	5,0	6
5	B_2	4,1	5
6	B_2	4,4	4
7	B_2	3,2	8
8	B_2	2,8	7

Lehmatõu B_1 hulgas lehmad 1 ja 4 on üsna sarnased: neil on kõrge piimatoodang ja nad on umbes ühevanused. Lehmad 2 ja 3 on samuti üsna sarnaste tunnustega. Võime moodustada ühe ploki, mis koosneb lehmadest 1 ja 4 ning teise ploki, mis koosneb lehmadest 2 ja 3. Lehmatõu B_2 hulgas lehmad 5 ja 6 peaksid moodustama ühe ploki ja lehmad 7 ja 8 peaksid moodustama teise ploki. Meil on 4 plokki, 2 lehma igas plokis. Igast plokist valime juhuslikult ühe lehma, keda söödame toiduskeemi A_1 järgi, teist lehma söödame toiduskeemi A_2 järgi.

Rõhutame, et selles näites tõug ei ole plokkimise muutuja, sest ta on üks faktor, mida soovime katses uurida.

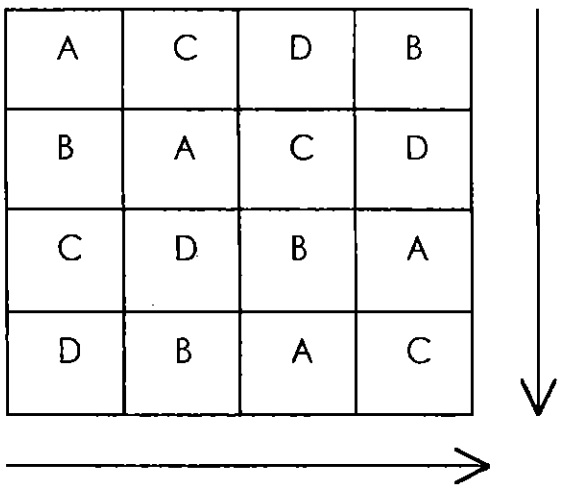
1.6.4 Ladina ruutude plaan

Mõnikord esineb kahte liiki varieeruvust, mida peame jälgima plokkide moodustamise juures. Näiteks katsepõld võib omada kallakut kahes suunas. Katseplaanis tuleb seda arvestada nii, et põllul esineb erinevus mõlema rea ja mõlema veeru vahel.

Näide. Soovime võrrelda nelja sorti nisu A, B, C ja D. Meil on põld, mille võime jaotada 16 maalapiks nii, et igat sorti nisu võime kasvatada neljal maalapil (replikatsioon). Aga põld on küllaltki ebaühtlaste tingimustega: ülemises vasakpoolses nurgas on parem mullastik kui all parempoolses nurgas ning mulla koostis muutub vähehaaval nende piiride vahel.

Katse planeerimise üks võimalus on kasutada ladina ruutude plaani. See

töötab järgmiselt. Jaotame töötlused juhuslikult 16 ala vahel, aga kindlasti nii, et iga töötlus on esindatud ainult üks kord igas reas ja igas veerus. Tulemus võib olla järgmine:



Ladina ruutude plaani saab kasutada ka loomadega seotud katsetes.

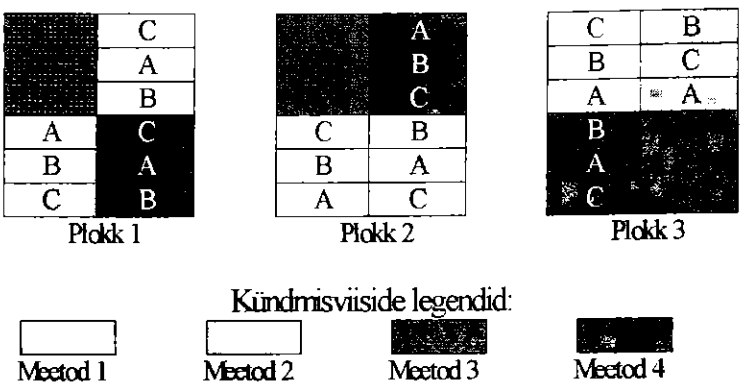
Näide. Soovime uurida piimalehmade nelja toiduskeemi A, B, C ja D. Iga toiduskeemi järgi toideti lehmi kolm nädalat. Iga perioodi kolmandal nädalal mõõdeti piimatoodang, et vältida ülekande mõjusid. Tulemused olid järgmised:

		Lehma number			
		1	2	3	4
Periood	1	A: 33,3	B: 29,5	C: 35,2	D: 35,7
	2	B: 32,9	D: 30,8	A. 34,1	C: 27,6
	3	C: 26,5	A: 28,5	D: 32,5	B: 27,5
	4	D: 30,3	C: 27,8	B: 32,7	A: 27,2

1.6.5 Liigendatud plokkplaan

Näide. Mõnikord on praktiline põhjus kasutada erinevaid plokkide moodustamise skeeme erinevatele tunnustele. Näiteks põlluvilja erinevate sortide katses (A, B ja C) ning kündmise erinevate viiside (1, 2, 3 ja 4) kui faktorite korral võib juhtuda, et erinevaid kündmise viise võib kasutada ainult üsna laiadel maa-aladel, samal ajal kui erinevaid liike on võimalik kasvatada väikesel alal. Niisugusel juhul kündmise viisid jaotatakse peaplokkide vahel, samal ajal kui erinevaid sorte kasvatatakse liigendatud plokkides peaplokkide sees.

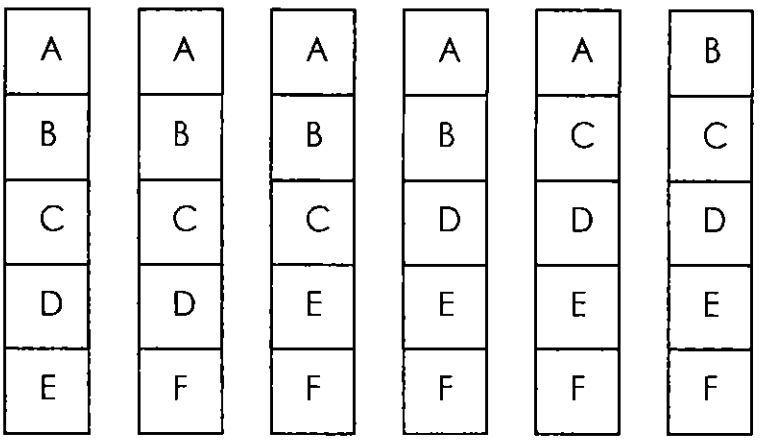
Kündmise erinevad viisid paigutatakse juhuslikult igasse peaplokki ning seejärel paigutatakse erinevad sordid juhuslikult igasse peaplokki. Niisugusne katse, kus kogu katset on korratud kolmes plokis, näeb välja järgmine:



1.6.6 Mittetäielikud plokid

Mõnedes katsetes, kus tundub loomulik kasutada randomiseeritud plokkide plaani, võivad plokid olla piiratud suurusega, nii et ei ole võimalik kõikides plokkides kasutada kõiki töötlusi (või töötluste kombinatsioone).

Näide. Oletame, et soovime võrrelda 6 töötlust: A, B, C, D, E ja F Meie käsutuses on 30 katseühikut ja need on jaotatud 6 ploki vahel, 5 ühikut plokis. Järelikult ei ole võimalik panna kõik kuus töötlust kõikidesse plokkidesse. (Näiteks selline olukord võib esineda, kui katses olevad loomad on üles kasvatatud tallis, kus on piiratud arv kohti, või kui kasvatame kasvuhoones kindlaksmääratud arvuga taimi). Küsimus on, kuidas kasutada 30 ühikut parimal viisil, nii et töötluste vaheline võrdlus oleks võimalikult efektiivne. Kui teeme plaani selliselt, et iga töötlus esineb viis korda, siis tulemusplaani nimetatakse tasakaalustatud lõpetamata plokkide plaaniks. Näiteks sellise olukorra üks plaan on:



Eelnevas plaanis esinevad kõik töötlused A, B,...,F viis korda. Peaaegu igat töötlust (näiteks C) on võrreldud iga teise töötlusega neli korda.

1.6.7 Andmeanalüüs

Enamus raamatuid (eriti "klassikalised") katseplaneerimise kohta sisaldavad osi, kuidas teha arvutusi: leheküljed lehekülgede järel on hälvete ruutude summad. Arvutusskeemid võivad anda lugejale ettekujutuse katseplaneerimise olemusest. Tänapäeval teostatakse andmeanalüüs arvutipakettide abil. Arvutipaketid, nagu Minitab või SAS, saavad kergelt hakkama ühe- või kahefaktorilise dispersioonanalüüsi mudelitega, samuti ka üldistatud lineaarsete mudelitega (GLM). Kõiki plaane, millest oli eelpool juttu, saab analüüsida Minitabi või SAS-i protsetuuriga GLM.

1.7 Ülesanded

1.1

Sojaube kasvatati kasvuhoones, kus igast sordist oli 10 taime. Kasutati täielikult randomiseeritud katseplaani. Iga sordi saagist valiti juhuslikult viis seemet ja mõõdeti iga seemne õlisisalduse protsent. Seega tehti iga sordi kohta 50 mõõtmist. Selleks, et arvutada iga sordi keskväärtuse standardviga, katsetaja arvutas 50-ne vaatluse standardhälbe ja jagas selle $\sqrt{50}$. Selgita, mis võiks sellel arvutusel viga olla.

1.2

Rottide toitumiskäitumise katses jaotati 18 rotti juhuslikult kolme töötluse T1, T2 ja T3 vahel. Loomi hoiti individuaalsetes puurides riiulil. Riiulil oli kolm riiulit, igal kuus puuri. Valgustingimised olid muutuvad; alumistel loomadel oli pimedam. Valmistati ette järgmised kolm plaani:

(i) Valiti 18 rotti juhuslikult18 positsioonile riiulis.

(ii) Pandi kõik T1 rotid esimesle riiulilele, kõik T2 rotid teisele ja kõik T3 rotid kolmandale riiulile.

(iii) Igale riiulile pandi kaks T1 rotti, kaks T2 rotti ja kaks T3 rotti.

Selgita iga plaani headust ning põhjenda, missugust plaani sina eelistaksid.

Peatükk 2

Ühefaktoriline dispersioonanalüüs

2.1 Sissejuhatav näide

Dispersioonanalüüs on statistiline meetod, millega saab võrrelda enam kui kahe üldkogumi keskväärtusi. Andmetes olev varieeruvus jaotatakse osadeks nii, et seda saab interpreteerida. Seejures on tähtis osa ka statistiliste hüpoteeside kontrollimisel. Erinevust kahe grupi vahel, näiteks kahe töötluse vahel, aitavad meil kindlaks määrata t-testid.

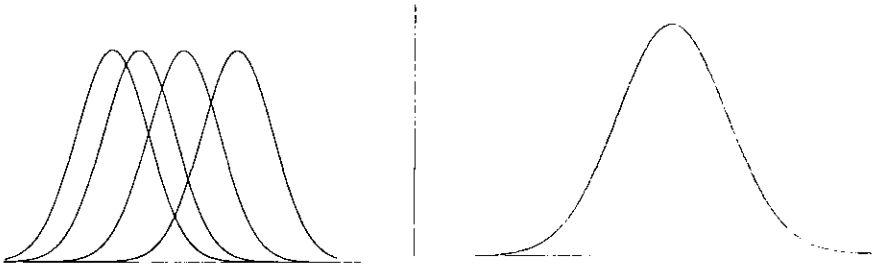
Dispersioonanalüüsis seatakse üles statistiline mudel, mida püütakse lähendada katsele või situatsioonile, mida proovitakse analüüsida. Tegelikult võime formuleerida mitmeid asjatundlikke mudeleid. Analüüsi ülesanne on kindlaks määrata üks mudel, mis nendest mudelitest on "parim" Näiteks esitame järgneva katseandmestiku, kus maisi saagikust mõõdeti 4·6 =24 põllulapil, mida väetati erinevat tüüpi väetistega.

Maisisaak nelja erineva väetise korral.

Väetis		1	2	3	4	5	6	$\sum_{j=1}^{6} y_{ij}$	$\bar{y}_{i\cdot}$
1	Kontroll	99	40	61	72	76	84	432	72
2	K_2O+N	96	84	82	104	99	105	570	95
3	K_2O+P_2O_5	63	57	81	59	64	72	396	66
4	N+P_2O_5	79	92	91	87	78	71	498	83
	$\sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij} =$	1896		$\bar{y}_{\cdot\cdot} =$	79				

Tähistame üksiku vaatluse y_{ij}, kus esimene indeks tähistab töötlust (väetist) ja teine indeks j tähistab töötluses olevat vaatlust. Meie näites $y_{11} = 99$ ja $y_{24} = 104$. Vaatluste keskväärtus ühe töötluse piires on $\bar{y}_{i.}$ ja kõikide vaatluste keskväärtus on $\bar{y}_{..}$ Indeks " " tähendab, et keskväärtus on leitud üle kõigi vaatluste. Töötlusel A on a taset. Vaatluste arv töötluses i on n_i ja kõigi vaatluste arv on N. Meie näites a=4, kõik n_i-d on võrsed 6-ga ja N=24.

Sellist tüüpi andmestiku korral on vaja kindlaks määrata, kas töötluste vahel on erinevust. Peame eristama kahte järgnevat situatsiooni:



Alternatiiv 1. Erinevate töötluste andmetel on jaotused erinevate keskväärtustega.

Alternatiiv 2. Kõigi nelja töötluste andmed omavad sama jaotust st keskväärtuste vahel ei ole erinevust.

2.2 Mudel ja piirangud

Mudeli juhule, kui keskväärtus on erinev, võib kirjutada järgmiselt

$$y_{ij} = \mu + \alpha_i + e_{ij},$$

kus μ on üle kõikide töötluste arvutatud keskväärtus (nn üldkeskmine), α_i on faktortunnuse i-nda taseme keskväärtuse ja üldkeskmise erinevus ning e_{ij} on juhuslik viga (jääk). Sageli eeldatakse, et jääkide dispersioon (hoolimata töötlusest) on võrdne σ_e^2, jäägid on sõltumatud ja normaaljaotusega.

α_i-d nimetatakse sageli ka töötluse i-ndaks efektiks, st ta on faktortunnuse i-nda taseme põhjustatud muutus. Töötluste efektide jaoks võetakse kasutusele lisakitsendus $\sum \alpha_i = 0$, mis tuleneb sellest, et parameeter α_i on sisuliselt kõrvalekalle üldkeskmisest ja see määrab mudeli üheselt.

Alternatiiv 2 korral saame lihtsustatud mudeli $y_{ij} = \mu + e_{ij}$, st kõik α_i on sel juhul nullid. Selleks, et ühte nendest kahest mudelist eelistada, tuleks uurida kas töötlustel on samad (üldkogumi) keskväärtused. Nullhüpoteesi võime püstitada järgmistel viisidel.

H_0: kõik α_i =0.

H_0: $\sum \alpha_i^2 = 0$.

Alternatiivne hüpotees on, et keskmised ei ole võrdsed, mis tähendab et H_1: $\sum \alpha_i^2 \neq 0$.

2.3 Hälvete ruutude summa jaotamine

Üksikust vaatlusest y_{ij} kauguse väljendamiseks üldkeskmiseni $\overline{y}_{..}$ võime kirjutada avaldise $(y_{ij} - \overline{y}_{..}) = (y_{ij} - \overline{y}_{i.}) + (\overline{y}_{i.} - \overline{y}_{..})$. $(y_{ij} - \overline{y}_{i.})$ näitab, kui kaugel on vaatlus selle töötluse keskväärtusest. $(\overline{y}_{i.} - \overline{y}_{..})$ mõõdab i-nda töötluse kaugust üldkeskmisest. Tõstame need ruutu ja summeerime üle kõigi vaatluste.

Andmestiku summaarset varieeruvust esitame avaldisega
$SS_T = \sum (y_{ij} - \overline{y}_{..})^2$

Seda võib jaotada kahte ossa, millest üks osa on tingitud töötlusest ja teine juhuslikust muutlikkusest järgmiselt:

$SS_T = \sum\sum (y_{ij} - \overline{y}_{..})^2 = \sum\sum (y_{ij} - \overline{y}_{i.} + \overline{y}_{i.} - \overline{y}_{..})^2 = \sum\sum (y_{ij} - \overline{y}_{i.})^2 + \sum n_i (\overline{y}_{i.} - \overline{y}_{..})^2$ (+ termid (liidetavad), mis on võrdsed nulliga).

Tähistame $SS_A = \sum n_i (\overline{y}_{i.} - \overline{y}_{..})^2$ ja $SS_E = \sum\sum (y_{ij} - \overline{y}_{i.})^2$ Järelikult oleme jaotanud summaarse varieeruvuse (the total Sum of Squares) kaheks osaks, üks on tingitud töötluste erinevustest (SS_A) ja teine on juhuslikust varieeruvusest (SS_E): $SS_T = SS_A + SS_E$. Üldine printsiip on, et kui SS_A on väike, siis töötluste vahel pole reaalset erinevust ja kui on suur, siis töötluste erinevus on olemas. Teiselt poolt mõõdab SS_E ainult varieeruvust iga töötluse sees ja võib mitte mõjutatud olla töötluste erinevustest.

Iga hälvete ruutude summa assotseerub *vabadusastmega*. Üldiselt on vabadusastmete arv summas võrdne termi (liidetavate) arv miinus lineaarsete kitsenduste arv termis. Sel juhul:

$SS_T = \sum\sum (y_{ij} - \overline{y}_{..})^2$ sisaldab N termi. Selles on üks kitsendus, nimelt $\sum (y_{ij} - \overline{y}_{..}) = 0$. Järelikult sisaldab see hälvete ruutude summa $N - 1$ vabadusastet;

$SS_A = \sum n_i (\overline{y}_{i.} - \overline{y}_{..})^2$ sisaldab t termi (= töötluste arv). Selles on üks kitsendus $\sum n_i (\overline{y}_{i.} - \overline{y}_{..}) = 0$. Ruutude summa sisaldab $t - 1$ vabadusastet;

$SS_E = \sum\sum (y_{ij} - \overline{y}_{i.})^2$ sisaldab N termi. Iga töötluse i jaoks on üks kitsendus, mis on $\sum (y_{ij} - \overline{y}_{i.}) = 0$, seega kokku on a kitsendust. Järelikult omab ruutude summa $N - a$ vabadusastet.

Kui jagame iga hälvete ruutude summa vastavate vabadusastmete arvuga, saame tulemuse, mida nimetatakse ruutkeskmiseks ja tähistatakse:

$MS_A = \frac{SS_A}{a-1}$,

$MS_E = \frac{SS_E}{N-a}$.

Ruutkeskmine on kindla dispersiooni hinnang. Ruutkeskmine viga (MS_E) on jääkide e_{ij} disprsiooni σ_e^2 hinnang, mis mõõdab andmestikus juhusliku varieeruvuse hulka. Seda võib väljendada: $E(MS_E) = \sigma^2$

On võimalik näidata, et $E(MS_A) = \sigma^2 + \frac{N}{(a-1)} \sum \alpha_i^2$.

See annabki meile otsustamise vahendi, kas töötluste keskväärtused on võrdsed või mitte.

Kui kõik töötlused on võrdsed, siis kõikide töötluste efektid on nullid, st kõik α_i on nullid. Seega on samaväärne öelda, et $\sum \alpha_i^2 = 0$. Järelikult, kui töötluste keskväärused on võrdsed, siis MS_A ja MS_E peaksid olema "lähedaselt võrdsed" Nende võrdlemiseks arvutame nende suhte: $F = \frac{MS_A}{MS_E}$ Kui nullhüpotees on õige, siis suhe järgib niinimetatud F-jaotust $(a-1)$ vabadusastmega lugejas ja $(N-a)$ vabadusastmega nimetajas. Illustreerimaks kuidas seda teha, pöördume tagsi meie numbrilise näite juurde ja esitame mõned arvutuste tulemused.

Hinnang $\bar{y}_{..} = 79$.

Hinnang α_1: $\widehat{\alpha}_1 = 72 - 79 = -7$

Hinnang α_2: $\widehat{\alpha}_2 = 95 - 79 = 16$.

Hinnang α_3: $\widehat{\alpha}_3 = 66 - 79 = -13$.

Hinnang α_4: $\widehat{\alpha}_4 = 83 - 79 = 4$.

Näeme, et $\sum \hat{\alpha}_i = -7 + 16 - 13 + 4 = 0$.

$\mathrm{SS}_T = \sum_{i=1}^{4} \sum_{j=1}^{6} (y_{ij} - \bar{y}_{..})^2 = (99-79)^2 + (40-79)^2 + ... + (71-79)^2 = 6212$ (24 termi)

$\mathrm{SS}_A = \sum_{i=1}^{4} \sum_{j=1}^{6} (\bar{y}_{i.} - \bar{y}_{..})^2 = 6(72-79)^2 + 6(95-79)^2 + 6(66-79)^2 + 6(83-79)^2 =$ 2940 (4 termi)

$\sum_{i=1}^{4} \sum_{j=1}^{6} (y_{ij} - \bar{y}_{i.})^2 = 3272$

Näeme, et $\mathrm{SS}_A + SS_E = 2940 + 3272 = 6212 = \mathrm{SS}_T$. Ühtlasi märgime, et see on kiireim viis käsitsi arvutamisel. Vastavad valemid on toodud raamatu lisas.

2.4 Dispersioonanalüüsi tabel

Tulemused on sageli kokku võetud järgmise tabelina, mida nimetatakse dispersioonanalüüsi tabeliks.

Allikas	Vabadusastmed	SS	MS	E(MS)
Mudel, A	a-1	SS_A	MS_A	$\sigma_e^2 + \frac{N}{a-1}\sum \alpha_i^2$
Viga	N-a	SS_E	MS_E	σ_e^2
Kokku	N-1	SS_T		

Meie näite korral on tabel järgmine:

Allikas	SS	Vabadusastmed	MS	F
Mudel	2940	$4-1=3$	980,0	5,99
Viga	3272	$4(6-1)=20$	163,6	
Kokku	6212	$24-1=23$		

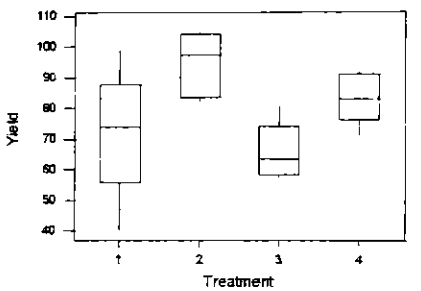
F-statistiku väärtuse arvutame järgmise valemi põhjal: $F = \frac{MS_A}{MS_E} = \frac{980}{163,6} = 5,99$. Et otsustada, kas nullhüpotees kehtib, vaatame F-jaotuse tabelit 5%-lise olulisuse nivoo juures. Leiame, et kriitiline väärtus vabadusastmetega 3 ja 20 on 3,493. Vastav 1%-line piir on 5,849. Meie arvutatud F-statistiku väärtus on 5,99, mis on suurem kui 5,849. Järelikult on tulemus oluline 1% olulisuse nivool.Võtame vastu sisuka hüpoteesi: võime pidada tõestatuks, et erinevatel väetistel on maisisaagile erinev mõju. Risk eksida ei ole suurem kui 1%. Sama tulemuse saame, kui kasutame Minitab'i:

One-way Analysis of Variance

Analysis of Variance for SAAK

Source	DF	SS	MS	F	P
Väetis	3	2940	980	5.99	0.004
Error	20	3272	164		
Total	23	6212			

Et paremini aru saada, mis toimub, on hea kasutada mingit vahendit kirjeldavast statistikast. Näiteks hetkel on ülevaatlikkuse tagamiseks hea teha karpdiagramm töötluste kaupa, kus yield tähistab saaki ja treatment väetist:



Peale dispersioonanalüüsi tegemist on loomulik püstitada küsimus "missugune nendest töötlustest on erinev" Selle küsimuse adresseerime järgmisele peatükile.

2.5 Ülesanded

2.1

Toiduaineteteadlane uurib küpsetatud koogis piimahulga kogust (mõõdetud milliliitrites 100 grammi kohta). Küpsetati 15 kooki. Kasutati ühesugust retsepti, aga piima lisati kolmes erinevas koguses (1: madal; 2: keskmine ja 3: kõrge). 15 kooki jaotati juhuslikult võrdselt iga koostise taseme järgi. Kookide kaalud olid järgmised:

Kogus	Koogi number				
	1	2	3	4	5
Madal	351	369	381	386	370
Keskmine	390	394	406	407	415
Kõrge	398	409	415	399	434

A. Dispersioonanalüüsi kasutades uuri, kas koogi kaal sõltub erinevast piimahulgast. Analüüs peaks sisaldama hüpoteese, dispersiooni tabelit, tulemuste kontrolli ja lõppjäreldusi.

B. Leia keskmise ja kõrge piimahulga erinevus koogi kaalus.

C. Missugusi eeldusi on vaja teha selleks analüüsiks?

2.2

Uuriti, kuidas käituvad kindlat liiki linnud sarnastes loodustingimustes ja ühises keskkonnas. Iga liigi linnulaulul oli teistest eristatav äratundmistunnus. Üks uurimise all olev karakteristik oli laulu pikkus sekundites. Uuriti kolme liiki linde: linavästrikku, rasvatihast ja pruunrästast. Saadi järgmised tulemused. (Ühtlasi on antud iga liigi jaoks summad ja ruutude summad.)

Linavästrik	Rasvatihane	Pruunrästas
1,11	2,17	0,42
1,23	1,85	0,93
0,91	1,99	0,77
0,95	1,74	0,37
0,99	1,54	0,50
1,08	1,86	0,48
1,18	1,87	0,68
1,29	2,04	0,62
1,12	1,69	0,39
0,88		0,67
1,34		1,03
		0,79
$\sum x = 12,08$	$\sum x = 16,75$	$\sum x = 7,65$
$\sum x^2 = 13,5030$	$\sum x^2 = 31,4649$	$\sum x^2 = 5,3843$

A. Kontrolli hüpoteesi, et iga liigi keskmine laulupikkus on võrdne kõigi kolme liigi keskmisega.

B. Püstita analüüsiks vajalikud eeldused.

2.3

Uuriti 57 hiire füsioloogilise funktsiooni erinevust vastandlikes keskkondades (kasutati röntgenikiirte mõju). Üks muutuja, mida uuriti, oli uriini hulk. Järgmised andmed näitavad iga roti uriini hulka enne ja pärast mõjutust. Tabelis veerg "Erinevus" on arvutatud veergude "pärast"-"enne" põhjal. Analüüsi eesmärk oli uurida vastandlike keskkondade mõju uriini hulgale.

Keskkond	Enne	Pärast	Erinevus	Keskkond	Enne	Pärast	Erinevus
Diatrizoate	1.670	34.590	32.920	Omnipaque	1.187	9.700	8.513
Diatrizoate	2.010	27.860	25.850	Omnipaque	2.470	18.580	16.110
Diatrizoate	2.970	23.720	20.750	Omnipaque	1.540	8.760	7.220
Diatrizoate	2.680	23.060	20.380	Omnipaque	1.530	10.560	9.030
Diatrizoate	0.960	8.020	7.060	Omnipaque	2.020	12.130	10.110
Hexabrix	3.320	9.790	6.470	Omnipaque	1.053	7.810	6.757
Hexabrix	1.890	7.520	5.630	Omnipaque	2.090	3.250	1.160
Hexabrix	2.380	5.460	3.080	Omnipaque	1.460	17.570	16.110
Hexabrix	2.033	2.990	0.957	Omnipaque	1.213	5.200	3.987
Hexabrix	4.990	7.360	2.370	Omnipaque	0.960	5.860	4.900
Hexabrix	1.390	8.390	7.000	Ringer	1.450	1.520	0.070
Hexabrix	1.206	6.090	4.884	Ringer	1.247	1.220	-0.027
Hexabrix	2.660	3.770	1.110	Ringer	1.680	2.020	0.340
Hexabrix	2.286	6.430	4.144	Ringer	1.210	1.290	0.080
Isovist	1.480	3.580	2.100	Ringer	1.600	2.110	0.510
Isovist	2.273	3.040	0.767	Ringer	1.673	1.770	0.097
Isovist	1.080	1.040	-0.040	Ringer	2.010	2.410	0.400
Isovist	1.890	6.690	4.800	Ultravist	2.060	15.000	12.940
Isovist	1.550	4.290	2.740	Ultravist	5.200	12.500	7.300
Isovist	1.166	3.610	2.444	Ultravist	1.990	17.340	15.350
Isovist	1.453	2.320	0.867	Ultravist	1.727	8.310	6.583
Isovist	1.550	1.330	-0.220	Ultravist	1.700	17.380	15.680
Isovist	2.130	3.650	1.520	Ultravist	0.980	4.460	3.480
Mannitol	1.680	10.870	9.190	Ultravist	1.120	6.870	5.750
Mannitol	2.240	3.030	0.790	Ultravist	1.580	13.760	12.180
Mannitol	1.946	12.170	10.224				
Mannitol	1.140	5.920	4.780				
Mannitol	2.160	16.800	14.640				
Mannitol	1.906	8.890	6.984				
Mannitol	1.400	8.910	7.510				
Mannitol	1.213	10.760	9.547				
Mannitol	2.717	8.250	5.533				

Küsimused:

A. "Ringer" on kontrolltöötlus. Seega eeldame, et sellel ei ole mõju uriini hulgale. Kontrolli, kas sellist eeldust on mõistlik kasutada eelneva andmestiku kohta.

B. Kas võib eeldada, et uriini hulga muutuse ("Erinevus") dispersioon on sama

"Ringeril" ja "Ultravistil"?

Peatükk 3

Kui F-test on oluline...

Dispersioonanalüüsis interpreteeritakse F-testi oluline väärtus, kui "erinevus töötluste keskväärtuste vahel" Kuid missugune töötlus on erinev, seda F-test üksi ei näita. Näiteks kui teeme ühefaktorilise dispersioonanalüüsi andmestiku mingi muutuja nelja erineva töötluse kohta, siis F-testi oluline väärtus ei anna informatsiooni, missugune töötlustest on "parim" Selle informatsiooni peame kätte saama edaspidisest analüüsist lähtudes F-testi olulisest tulemusest.

Näide. Järgnev andmestik esitab nelja odrasordi katse tulemused.

Sort						Keskmine
A	89	76	101	89	84	87,8
B	95	80	119	104	103	100,2
C	124	113	116	122	113	117,6
D	103	113	101	85	100	100,4

Andmestiku dispersioonanalüüs annab järgmised tulemused (leitud SAS programmiga):

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2249.0000000	749.6666667	7.27	0.0027
Error	16	1650.0000000	103.1250000		
Corrected Total	19	3899.0000000			

R-Square	C.V	Root MSE	YIELD Mean
0.576815	10.00497	10.155048	101.50000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
VARIETY	3	2249.0000000	749.6666667	7.27	0.0027

Näeme, et F-test on oluline (sest Pr > F on väiksem kui 0,0027), aga ikkagi me ei tea, missugune liikidest on teistest oluliselt parem (või halvem). On olemas kindlad meetodid F-testi jätkamiseks. Järgnevas kirjeldame lühidalt mõnda sellist ja illustreerime neid eelneva näite andmete põhjal.

3.1 Töötlustevaheline võrdlemine

3.1.1 Paarisvõrdlus

Kõige otsesem tee sortide võrdlemiseks on kahefaktorilise dispersioonanalüüsi kasutamine. Esiteks võime võrrelda sorte A ja B (kontrollime hüpoteesi H_0 : $\mu_A - \mu_B = 0$), kasutades standardset valemit

$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s^2(\frac{1}{n_A}+\frac{1}{n_B})}}$, kus $s^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}$ on arvutatud valimi põhjal. Valem s^2 põhineb eeldusel, et sortidel A ja B on võrdne dispersioon σ^2 Ühefaktorilise dispersioonanalüüsi korral eeldasime, et kõigil neljal töötlusel on võrdne dispersioon, mitte ainult töötlustel A ja B. Järelikult võime σ^2 hinnanguna kasutada hinnangut disprsioonanalüüsi tabelist, st võime kasutada arvuti väljatrükist väärtust s_e^2=(Mean square for error). Vaadeldava näite korral $s_e^2 = 103{,}125$.

Järelikult, kontrollides hüpoteesi $H_0 : \mu_A = \mu_B$, on vaja arvutada $t = \frac{87{,}8-100{,}2}{\sqrt{103{,}125(\frac{1}{5}+\frac{1}{5})}} = -1{,}93$ vabadusastme arvuga 16. Vastav kriitiline väärtus on t_{kr}=2,12. t-test ei ole oluline. Sellise lähenemise eelis on selles, et kasutame katsest rohkem informatsiooni. Hetkel σ^2 hinnangus kasutame 16 vabadusastet 8 (st 5 + 5 − 2) asemel.

3.1.2 Paarisvõrdlemise tulemuste esitamine

Tulemuste esitamise üks võimalus on esitada SAS väljatrükk. Järgnev väljatrükk iseloomustab eelneva näite andmeid.

T tests (LSD) for variable: SAAK

NOTE: This test controls the type I comparisonwise error rate not not the experimentwise error rate

Alpha= 0.05 df= 16 MSE= 103.125
Critical Value of T= 2.12
Least Significant Difference= 13.615

Means with the same letter are not significantly different

T Grouping	Mean	N	SORT
A	117.600	5	C
B	100.400	5	D
B			
B	100.200	5	B
B			
B	87.800	5	A

Väljatrüki alumisest osast leiame sortidele keskväärtused (veerg "Mean"). Veerus "T Grouping" on täht A sordi C ja täht B on kõigi ülejäänud sortide A, B ja D ees. See tähendab, et ei ole olulist erinevust sortide A, B ja D vahel, aga sordi C keskväärtus on oluliselt erinev ülejäänud sortidest.

Väljatrüki ülemises osas näeme teksti "Least Significant Difference" (LSD), mis näitab, kui suurt erinevust on vaja kahe sordi vahel, et olla järjestikku olulised. $LSD = t\sqrt{MS_e\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 2,12\sqrt{103,125\left(\frac{1}{5}+\frac{1}{5}\right)} = 13,615.$ Vabadusastmete arv t-statistiku leidmiseks on $N - a = 16$. Kui keskväärtuste erinevus kahe sordi vahel on suurem kui 13,615, siis erinevus on oluline. Sordil C on keskväärtus 117,6 ja sordil D on keskväärtus 100,4, seega keskväärtuste erinevus on: $117,6 - 100,4 = 17,2$. See on suurem kui 13,615, järelikult erinevus on oluline 5% olulisuse nivoo juures. LSD väärtus on kasutatav tasakaalustatud katsete korral, teistel juhtudel on vaja arvutatada t-statistiku väärtus igale võrdlusele eraldi.

Analüüsime keerukamat näidet, mille esitasime 2. peatüki sissejuhatuses. Selles võrreldi nelja väetise mõju maisi saagile. Dispersioonanalüüsi tabel

näitas nende olulist erinevust. Esitame SAS väljatrüki paarisvõrdlusest.

```
                General Linear Models Procedure

              T tests (LSD) for variable: Saak

NOTE: This test controls the type I comparisonwise error
      not the experimentwise error rate.

    Alpha= 0.05  df= 20  MSE=
       Critical Value of T=
  Least Significant Difference= 15.404

Means with the same letter are not significantly different.
T Grouping          Mean      N  Väetis

          A        95.000     6  2
          A
   B      A        83.000     6  4
   B
   B      C        72.000     6  1
          C
          C        66.000     6  3
```

Väetised 2 ja 4 ei ole käesoleval juhul oluliselt erinevad (seda näitab T Grouping veeru täht A). Väetistel 4 ja 1 ei ole olulist erinevust (T Grouping B). Ka väetisted 1 ja 3 ei ole olulist erinevust (T Grouping C). Kõik teised paarisvõrdlused , st väetised 2 ja 1; 2 ja 3; 4 ja 3 on oluliselt erinevad.

3.1.3 Kontrastid

Sageli ei ole võrdlused lihtsad paarisvõrdlused, vaid on vaja võrrelda erinevate töötluste kombinatsioone. Näiteks töötlus A on standard ja meid huvitab kontrollida, kas töötluse A keskmine erineb teiste töötluste keskmisest, st peaksime kontrollima hüpoteesi $H_0: \mu_A - \frac{\mu_B+\mu_C+\mu_D}{3} = 0$. Niisugust töötluste kombinatsiooni keskväärtust nimetatakse kontrastiks. Kontrasti üldine valem on $L = \sum h_i\mu_i$, kus $\sum h_i = 0$. Märgime, et paarisvõrdlus on kontrasti leidmise erijuht.

Kontrollides hüpoteesi kontrasti kohta, on vaja hinnata kontrasti ja dispersiooni väärtusi. Meie näites on kontrasti hinnang $\hat{L} = \bar{y}_A - \frac{\bar{y}_B+\bar{y}_C+\bar{y}_D}{3} = 87,8 - \frac{110,2+117,6+100,4}{3} = -18,267$ Kontrasti hinnangu dispersioon, kasutades keskväärtuste ja dispersiooni reegleid, on $D(\bar{y}_A - \frac{1}{3}\bar{y}_B + \frac{1}{3}\bar{y}_C + \frac{1}{3}\bar{y}_D) = D(\bar{y}_A) + \frac{1}{9}D(\bar{y}_B) + \frac{1}{9}D(\bar{y}_C) + \frac{1}{9}D(\bar{y}_D) = \frac{\sigma^2}{n_A} + \frac{1}{9}(\frac{\sigma^2}{n_B} + \frac{\sigma^2}{n_C} + \frac{\sigma^2}{n_D})$. Kasutame σ^2 hinnanguna MS_e-d, järelikult $\hat{D}\left(\hat{L}\right) = 103,125(\frac{1}{5} + \frac{1}{9}(\frac{1}{5} + \frac{1}{5} + \frac{1}{5})) = 27,5$.

Kontrollime hüpoteesi $H_0 : L = 0$, arvutame $t = \frac{\hat{L}}{\sqrt{\hat{D}(\hat{L})}} = \frac{-18,267}{\sqrt{27,5}} = -3,48$. Tabelist leiame t kriitilise väärtuse kui vabadusastmete arv on 16 ja võrdleme seda arvutatud t-ga. 5%-lise olulisuse nivoo juures on t kriitiline väärtus 2,12 ja 1%-ne t kriitiline väärtus on 2,92. Leitud kontrast on oluline juba 1% olulisuse nivoo juures.

3.2 Probleemid, kui teeme mitmeid teste

Kui teeme mitmeid teste, kerkivad ka probleemid. Eelnevas näites oli neli sorti. Võime teha sortide vahel kokku 6 erinevat paarisvõrdlust ($A-B, A-C, A-D, B-C$, $B-D$ ja $C-D$). Üldiselt, kui teeme kuus sõltumatut testi α= 5%-lisel olulisuse nivool, siis on 26% tõenäosus[1], et vähemalt üks test on oluline, isegi kui sortide erinevust ei ole! See on üsna suur risk. Ja kui meil on näiteks 10 sorti, siis risk on 90%, et vähemalt üks test 10·9/2=45 paarisvõrdlusest on oluline "ainult edu korral" See aga tähendab, et me ei tööta enam 5%-lisel olulisuse nivool.

Seda probleemi nimetatakse olulisuse hulga probleemiks. Ta kerkib, kui teeme suure arvu olulisi teste. Siis võime saada ka mingi arvu olulisi tulemusi, isegi kui ei ole "tegelikke" erinevusi liikide vahel.

3.3 Lahendused

Olulisuse hulga probleemi üldine lahendus on suurendada nõudeid olulistele tulemustele. Kui standard t-test 16 vabadusastmega on oluline 5%-lise olulisuse nivoo juures (|t|>2,12), siis juhul kui on olulisuse hulga risk, peaksime määrama piirid veidi kaugemale, võib-olla |t|>3,0. See tähendab, et kasutame igale testile väiksemat olulisuse nivood. Küsimus on: kui palju kaugemale või missugust uut olulisuse nivood peaksime kasutama? Erinevad autorid vastavad sellele erinevalt. Nende mõningaid soovitusi vaadeldakse allpool. Meetodite valik põhineb SAS paketis protseduuridel ANOVA ja GLM.

3.3.1 Bonferroni

Bonferroni t-test (BON funktsioon SAS-is) leiab paarikaupa erinevusi kahe töötluse keskväärtuse vahel, kui olulisuse nivoo on teada. Tavaline t-test

[1] Arvutatakse $1-(1-0,05)^6 = 0,26$

$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s^2(1/n_A + 1/n_B)}}$ on oluline ε nivool, kus $\varepsilon = \alpha/c$ ja c võrdluste arv. Meie näites võime teha $c = 6$ paarikaupa võrdlemist. Kui soovime kontrollida kõiki hüpoteese $H_0 : \mu_A - \mu_B = 0$ 5%-lise olulisuse nivoo juures, siis peame teostama iga individuaalse testi olulisuse nivool $\varepsilon = 0,05/6 = 0,0083$. Selle asemel, et kasutada piiri |t|>2,12, peame |t|>3,01 puhul hüpoteesi H_0 tagasi lükkama.

3.3.2 Sidak

Sidak'i t-test (SIDAK funktsioon SAS-is) leiab olulist erinevust tasemel, kui t-test on oluline ε olulisuse nivool, kus $\varepsilon = 1 - (1 - \alpha)^{\frac{1}{c}}$ Nagu varemgi, tähistame ka siin c-ga võimalike võrdluste arvu. Näites $c = 6$ ja $\alpha = 0,05$, saame $\epsilon = 1 - (1 - 0,05)^{\frac{1}{6}} = 0,0085$. Vastav t piir on 3,00.

3.3.3 Sheffé

Vastavalt Sheffé testile on erinevus kahe töötluse keskväärtuse vahel oluline olulisuse nivool, kui t-suhe on suurem kui $\sqrt{(k-1)\ F(\alpha; k-1, v)}$. $F()$ tähistab F-jaotuse tabelist kriitilist väärtust, kus vabadusastmete arv on sama kui standardses F- testis. Meie näite andmetes F-väärtus 5%-lise olulisuse nivoo juures vabadusastmetega (3 ja 16) on 3,24. Järelikult nullhüpoteesi $H_0 : \mu_A - \mu_B = 0$ peaks tagasi lükkama, kui t suhe on suurem kui $\sqrt{(4-1)\ 3,24} = 3,12$.

3.3.4 Tukey

Eelnevaid meetodeid võib kasutada mitte ainult paarisvõrdlusel, vaid ka üldiste kontrastide korral. Tukey esitab töötluste keskväärtuste paarisvõrdluseks spetsiaalse testi. Tukey meetodi põhjal on kaks keskväärtust oluliselt erinevad mingil nivool, kui t suhe on suurem kui $q(\alpha; k, v)/\sqrt{2}$, kus q on niinimetaud Studenti haare (standardvea abil normeeritud haare). See tabel on toodud raamatu lisas. α on olulisuse nivoo, k on töötluste võrdluste arv ja v on vastavate vabadusastmetega standardse F-testi nimetaja. Meie näite korral k=4, v=16 ja α= 0,05. Saame q=4,05, st peame hüpoteesi $H_0 : \mu_A - \mu_B = 0$ tagasi lükkama, kui t suhe on suurem kui $4,05/\sqrt{2} = 2,86$.

Eelnevas esitasime 2. peatüki sissejuhatava näite paarisvõrdluseks SAS väljatrüki, millelt saime välja lugeda paarikaupa t-testi ja LSD võrdlemise

tulemused. Esitame nüüd Tukey meetodiga leitud SAS väljatrüki:

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: Saak

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 20 MSE= 163.6
Critical Value of Studentized Range= 3.958
Minimum Significant Difference= 20.66952

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	VÄETIS
A	95.000	6	2
A			
B A	83.000	6	4
B			
B	72.000	6	1
B			
B	66.000	6	3

Kui esimeses paarisvõrdluses olid väetised 2 ja 1, 2 ja 3 ning 4 ja 3 oluliselt erinevad, siis nüüd näeme, et väetised 4 ja 3 ei ole enam erinevad. Põhjus on selles, et Tukey testis on suuremad nõudmised olulisusele kui paarikaupa t-testis. Varem oli LSD 15,404, nüüd on 20,669. See arvutati avaldisest $2{,}86\sqrt{MS_e\left(\frac{1}{n}+\frac{1}{n}\right)}$. Tukey testi saab teha ka paketiga Minitab. Selle väljatrükk näeb teistmoodi välja, kuid järeldused on samad.

Tukey Simultaneous Tests
Response Variable Saak
All Pairwise Comparisons among Levels of VÄETIS

VÄETIS = 1 subtracted from:

Level VÄETIS	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
2	23.000	7.385	3.1146	0.0258
3	-6.000	7.385	-0.8125	0.8478
4	11.000	7.385	1.4896	0.4619

VÄETIS = 2 subtracted from:

Level VÄETIS	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
3	-29.00	7.385	-3.927	0.0043
4	-12.00	7.385	-1.625	0.3879

VÄETIS = 3 subtracted from:

Level VÄETIS	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
4	17.00	7.385	2.302	0.1311

3.3.5 Aste-alla meetod (Step-down method)

Aste-alla meetodi korral on protseduur järgmine. Esiteks kontrollime hüpoteesi, kas kõik k keskväärtust on võrdsed, kasutades mingit olulisuse nivood γ_k. Kui see on oluline, kontrollime iga alamhulga (k-1) keskväärtust, kasutades seejuures mingit olulisuse nivood γ_{k-1}; vastasel juhul protseduur seiskub. Olulisele alamhulgale jätkub protseduur. SAS on varustatud võimsa aste-alla meetodiga, mis põhineb Studenti haardel (so standarvea abil normeeritud haare), vastav protseduur on REGWQ. Aste-alla meetodi järgi on keeruline käsitsi arvutada, kuid arvutit kasutades on see lihtne. Minitab ei oma sellist protseduuri.

3.4 Soovitused

Mitmeid lisameetodeid mitmese võrdlemisele jaoks võib leida SAS käsiraamatutest. Paneb imestama, kui palju neid on. Missugune on neist parim?

Sellele küsimusele vastates, peame defineerima, mis on "parim" Hüpoteeside kontrollimise juures töötavad statistikud järgmise kriteeriumi järgi.

Tegelik olulisuse nivoo peab olema võrdne või vähemalt ligikaudu võrdne määratud tasemel. Kontrollides hüpoteesi 0,05 olulisuse nivool on hüpoteesi H_0 tagasilükkamise tõenäosus 5%.

Antud olulisuse nivoo korral peab test omama suuremat tõenäosust, et lükata H_0 tagasi. Test peab sisaldama suuremat jõudu.

Mitmese võrdlemise korral kerkib lisaküsimus: kas soovime kontrollida hüpoteesi, kus α on iga individuaalse võrdlemise korral ("comparisonwise error rate") veasuhte mõõt, või kas tahame, et α oleks mingi hüpoteesi korral terve katse veasuhte mõõt. Esimesel juhul võime kasutada lihtsat t-testi. Teisel juhul kui ei ole vaja usalduspiire, on soovitatav kasutada SAS REGWQ protseduuri. Kui on vaja usalduspiire, siis on soovitav kasutada SAS TUKEY funktsiooni.

3.5 Ülesanded

3.1 Järgnev andmestik pärineb kolme töötlusega katsest

A_1	A_2	A_3
2	12	23
5	10	26
8	18	22
3	16	19
8		24

A. Arvuta ruutude summad ja tee dispersioonanalüüsi tabel.

B. Kui tulemus on oluline, tee töötlustevaheline paarisvõrdlemise test LSD-võrdlemisena.

3.2 Katses moodustati Allimum schoenoprasum sordist 3 gruppi A_1, A_2 ja A_3, et võrrelda selle sordi jõudlust. Iga grupi taimi kasvatati 10 potis. Teatud ajal mõõdeti kõikide taimede kõrgus y. Andmeid võib kokkuvõtvalt esitada järgmiselt:

$n_1 = 10 \quad \bar{y}_1 = 160 \quad s_1 = 20$

$n_2 = 10 \quad \bar{y}_2 = 150 \quad s_2 = 15$

$n_3 = 10 \quad \bar{y}_3 = 180 \quad s_3 = 18$

A. Koosta mudel, tee dispresioonanalüüsi tabel ja kontrolli, kas gruppide keskmine kõrgus on võrdne.

B. Tee kõik paarisvõrdlemised (5% LSD).

3.3 Sööt A_1 baseerub maisil ja sööt A_2 baseerub odral. A_1 ja A_2 on ühesugune kalorsus. Sööt A_3 sisaldab võrdses koguses sööta A_1 ja A_2. Iga söödaga söödeti 10 juhuslikult valitud looma. Loomad kaaluti enne ja pärast katset. Üks loom, kellele söödeti sööta A_3, eemaldati katsest. Katse tulemused olid järgmised:

$n_1 = 10 \quad \bar{y}_1 = 15,5 \quad s_1 = 2,0,$

$n_2 = 10 \quad \bar{y}_2 = 14,5 \quad s_2 = 1,5,$

$n_3 = 9 \quad \bar{y}_3 = 17,5 \quad s_3 = 1,8.$

A. Koosta mudel, tee dispersioonanalüüsi tabel ja kontrolli, kas söödad annavad võrdset keskmist kaalu juurdekasvu.

B. Kontrolli järgmisi kontraste ja selgita nende mõõte: $L_1 = \alpha_1 - \alpha_2$, $L_2 = 0,5\alpha_1 + 0,5\alpha_2 - \alpha_3$.

3.4

Kasuta ülesande 2.3 andmeid.

Selle andmestiku dispersioonanalüüsi tulemused on, Minitab'i kasutades, järgmised. Leia vastused küsimärkide asemele.

One-Way Analysis of Variance

Analysis of Variance for Erinevus

Source	DF	SS	MS	F
Contrast	?	1788.0	?	?
Error	?	905 1	?	
Total	?	2693.1		

Kirjeldava statistika statistikud kontrastsete keskkondade erinevusele on järgmised:

Descriptive Statistics

Variable	Contrast	N	Mean	Median	Tr Mean	StDev	SE Mean
Erinevus	Diatrizo	5	21.39	20.75	21.39	9.48	4.24
	Hexabrix	9	3.961	4.144	3.961	2.227	0 742
	Isovist	9	1.664	1.520	1 664	1.569	0.523
	Mannitol	9	7.69	7.51	7.69	3.90	1.30
	Omnipaqu	10	8.39	7.87	8.33	4.84	1.53
	Ringer	7	0.2100	0.0970	0.2100	0.2035	0.0769
	Ultravis	8	9.91	9.74	9.91	4.69	1.66

A. Uuri, missugusel kontrastsel keskkonnal ("contrast") on oluline mõju uriini hulgale.

B. Eelnevat informatsiooni kasutades tee erinevate keskkondade paarisvõrdlemine.

Peatükk 4

Katsed plokkstruktuuridega

4.1 Randomiseeritud plokkplaan

4.1.1 Sissejuhatav näide

Katses uuriti[1] tomatite saagikust katmikalal. Originaalkatses oli kuus töötlust (A kuni F), aga meie jätame mitmesugustel põhjusetel ära töötlused E ja F Allesjäänute hulgast oli A kontrolltöötlus, viljakuse tõstmiseks kasutati erinevaid liike komposte (töötlused B, C ja D).

Tomateid kasvatati kastides. Igas kastis oli võrdne arv tomatitaimi. Kasvuhoone mahutas kaks rida kaste, üks rida paiknes lõuna pool ja teine oli põhja pool. Katseplaan oli järgmine: iga kasti taimede kogutoodang kaaluti ja kast sai numbri, mis väljendas toodanguperioodi kilogrammide arvu ruutmeetri kohta.

E	D	A	B	F	C	Lõuna
ei kasutatud	23	12	20	ei kasutatud	22	
F	A	C	E	B	D	Põhja
ei kasutatud	7	20	ei kasutatud	15	20	↓

[1]Situatsioon põhineb Lena Gäredal'i (Rootsi Põllumajandusülikool) uurimusel.

Ülesandeks oli uurida, kas erinevad kompostid avaldavad mõju tomatitoodangule.

Katse on randomiseeritud plokkplaan, kus rida vaadeldakse plokina. (Esialgses katses oli plokke rohkem). Andmete põhjal koostasime järgmise tabeli:

Plokk	Töötlus	Toodang
1	A	12
1	B	20
1	C	22
1	D	23
2	A	7
2	B	15
2	C	20
2	D	20

4.1.2 Mudel ja kitsendused

Niisuguste andmete korral on mudel $y_{ij} = \mu + \alpha_i + \beta_i + e_{ij}$, kus $i = 1, ..., a$ ja $j = 1, ..., b$. Vaatluste arv $N = ab$. Eeldame, et $e_{ij} \sim N(0, \sigma_e^2)$. Kitsendused on $\sum \alpha_i = \sum \beta_j = 0$.

4.1.3 Ruutude summad

Varieeruvust põhjustab töötlus (A) ja plokk (B), ülejäänud varieeruvus on juhuslik ja kuulub jääkidele. Võime kirjutada

$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$.

$(y_{ij} - \bar{y}_{..})$ on katses ühe vaatluse erinevus üldkeskmise väärtusest. Vahega $(\bar{y}_{i.} - \bar{y}_{..}) = \hat{\alpha}_i$ võrreldakse üldkeskmist i-nda töötluse keskväärtusega, see on töötluse mõju. Vahega $(\bar{y}_{.j} - \bar{y}_{..}) = \hat{\beta}_i$ võrreldakse j-nda ploki keskväärtust üldkeskmisega, see on ploki mõju. $(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) = e_{ij}$ on ülejääk ehk jääk.

Tõstame iga termi ruutu ja summeerime üle kõigi $a \quad b = N$ vaatluse. Paljud termid ruutu tõstes on võrdsed 0-ga, alles jäävad:

$$\sum\sum(y_{ij} - \bar{y}_{..})^2 = \sum\sum(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum\sum(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum\sum(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$
$$SS_T = \quad SS_A \quad + \quad SS_B \quad + SS_e.$$

Ruutude summa ülc kõigi SS_T on (N-1) vabadusastmega, SS_A on (a-1) vabadusastmega ja arvutatakse täpselt nii nagu ühefaktorilise katse korral. SS_B on (b-1) vabadusastmega ja arvutatakse analoogiliselt SS_A-ga. Seega

$SS_e = SS_T - SS_A - SS_B$ on $(N-1-a+1-b+1) = (a-1)(b-1)$ vabadusastmega. Käsitsi arvutamise valemid on toodud raamatu lisas.

Võib näidata, et $E(MS_A) = \sigma_e^2 + b\sum \alpha_i^2/(a-1)$ ja $E(MS_B) = \sigma_e^2 + a\sum \beta_i^2/(b-1)$. Lisaks $E(MS_e) = E(s_e^2) = \sigma_e^2$. SAS raamatus on kasutusel tähistused $Q(A) = b\sum \alpha_i^2/(a-1)$ ja $Q(B) = a\sum \beta_i^2/(b-1)$, mida sageli kasutatakse.

Dispersioonanalüüsi tabel:

Allikas	Vabadusastmed	SS	MS	E(MS)
Töötlus A	$a-1$	SS_A	MS_A	$\sigma_e^2 + Q(A)$
Plokk B	$b-1$	SS_B	MS_B	$\sigma_e^2 + Q(B)$
Jääk	$(a-1)(b-1)$	SS_e	$MS_e = s_e^2$	σ_e^2
Kokku	$N-1$	SS_T		

4.1.4 Arvuline näide

Meie näites on ruutude summad arvutatud veel mugavamate valemitega, võrreldes raamatu lisas esitatutega.

$$SS_T = \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = (12 - 17,375)^2 + (20 - 17,375)^2 +$$

$$+ (20 - 17,375)^2 = 215,875,$$

$$SS_B = \sum_i n_i (\bar{x}_i - \bar{x}_{..})^2 = 4(19,25 - 17,375)^2 + 4(15,50 - 17,375)^2 =$$

$$28,125,$$

$$SS_A = \sum_j n_j (\bar{x}_{.j} - \bar{x}_{..})^2 = 2(9,5 - 17,375)^2 + 2(17,5 - 17,375)^2 +$$

$$+2(21 - 17,375)^2 + 2(21,5 - 17,375)^2 = 184,375,$$

$$SS_e = SS_T - SS_B - SS_{Tr} = 215,875 - 28,125 - 184,375 = 3,375.$$

Dispersioonanalüüsi tabel:

Allikas	Vabadusastmeid (df)	SS	MS=SS/df	F
Töötlus	$t-1=3$	184,375	61,458	54,629
Plokk	$b-1=1$	28,125	28,125	25,000
Jääk	$(t-1)(b-1)=3$	3,375	1,125	
Kokku	$n-1=7$	215,875		

Kontrollime hüpoteesi H_0, milles väidame, et töötlustel ei ole mõju (st $\sum \alpha_j^2 = 0$). Leiame tabelist $F_{kr} = 54,629$ vabadusastmetega (3; 3). 5% piir on 9,277 1% piir on 29,457 0,1% piir on 141,11. Tulemus on oluline (juba) 1%-lise olulisuse nivoo juures.

Kontrollime hüpoteesi H_0, milles väidame, et plokkidel ei ole mõju (st $\sum b_i^2 = 0$). Nüüd F_{kr}=25,000 vabadusastmetega (1; 3). 5% piir on 10,128.

1% piir on 34,116. 0,1% piir on 167,03. Järelikult tulemus on oluline 5%-lise olulisuse nivoo juures.

Kokkuvõte.

Töötluste keskväärtused: A: 9,5 B: 17,5 C: 21,0 D: 21,5.

Plokkide keskväärtused: 1: 19,25 2: 15,50.

Järeldused.

Plokkide vaheline erinevus on oluline.

Töötluste vaheline erinevus on oluline.

Eeldused olid: vaatlused peavad olema sõltumatud (st jääkide sõltumatus) ja jäägid normaaljaotusega.

Märgime, et plokkide mõju ei ole üldiselt peamine huvi. Katses kasutame plokke sellepärast, et saada efektiivsemat katset. Kui plokid on moodustatud õigesti, võime eeldada ploki olulist mõju, aga isegi kui efekt ei ole formaalselt oluline, jääme oma mudeli juurde. Ploki mõju olemasolu võime kasutada tulevaste katsete planeerimisel. Kui plokidel ei ole mõju, näiteks 15 kuni 25% tasemel, siis plokkide moodustamist tuleb vaadelda kui mitteõnnestunut ja tuleks kasutada mõnda teist plokkimise faktorit.

4.1.5 Analüüs arvutiga

Analüüsiks kasutame SAS paketi protseduuri GLM.

Järgnevat SAS programmi kasutasime meie näite andmetele:

```
DATA tomat;
INPUT plokk kompost $ saak;
CARDS;
1     A     12
1     B     20
1     C     22
1     D     23
2     A     7
2     B     15
2     C     20
2     D     20
;
PROC GLM DATA=tomat;
CLASS plokk kompost;
MODEL saak = plokk kompost;
MEANS kompost/tukey;
RUN;
```

SAS väljatrükk:

General Linear Models Procedure Dependent Variable: SAAK

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	212.50000000	53.12500000	47.22	0.0048
Error	3	3.37500000	1.12500000		
Corrected Total	7	215.87500000			

R-Square	C.V	Root MSE	YIELD Mean
0.984366	6.104519	1.06066017	17.375000008

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLOKK	1	28.12500000	28.12500000	25.00	0.0154
KOMPOST	3	184.37500000	61.45833333	54.63	0.0041

Alpha= 0.05 df= 3 MSE= 1.125
Critical Value of Studentized Range= 6.825
Minimum Significant Difference= 5.11842

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Kompost
A	21.500	2	D
A			
A	21.000	2	C
A			
A	17.500	2	B
B	9.500	2	A

Näeme, et nii töötluse kui ka plokkide mõjud on olulised. Kontrolltöötlusel A on keskväärtus oluliseselt väiksem kui ülejäänud töötlustel.

Analüüsiks kasutame paketti Minitab

Minitabi kasutamisel tuleb ka sõltuvaks muutujaks ette anda muutuja "saak" ja mudel, mis koosneb plokkidest "kompost" Minitabi väljatrükk on järgmine:

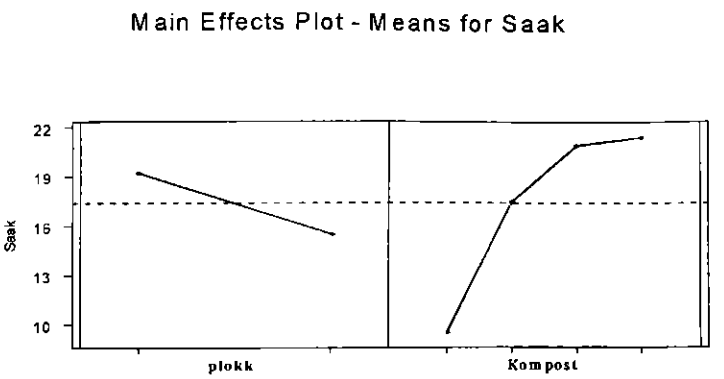
General Linear Model

Factor	Levels	Values
Plokk	2	1 2
Kompost	4	A B C D

Analysis of Variance for SAAK

Source	DF	Seq SS	Adj SS	Adj MS	F	P
PLOKK	1	28.125	28.125	28.125	25.00	0.015
Kompost	3	184.375	184.375	61.458	54.63	0.004
Error	3	3.375	3.375	1.1252		
Total	7	215.875				

Minitab'i abil saab koostada nn peamõjude graafiku, mis näitab peamõju iga taseme keskväärtust. Minitab'i peamõjude graafik:



4.2 Katse, kus kasutame ladina ruute

4.2.1 Sissejuhatav näide

Uuriti tootmisprotsessi, kus võrreldi viit erinevat toote retsepti: A, B, C, D ja E. Protsessi tulemus sõltus ka sellest missugust toormaterjali patakat kasutati ja missugune operaator oli selle masinasse pannud. Kõik operaatorid kasutasid kõiki toormaterjali patakaid ladina ruutude plaanis.

	Operaator				
Patakas	1	2	3	4	5
1	A 24	B 20	C 19	D 24	E 24
2	B 17	C 24	D 30	E 27	A 36
3	C 18	D 38	E 26	A 27	B 21
4	D 26	E 31	A 26	B 23	C 22
5	E 22	A 30	B 20	C 29	D 31

Randomiseerimisel tähistati erinevad retseptid juhuslikult koodidega A, B, C, D ja E.

4.2.2 Mudel ja eeldused

Mudel, mida kasutatakse ladina ruutude katseplaanis, on $y_{ij(k)} = \mu + \alpha_i + \beta_j + \gamma_{(k)} + e_{ij(k)}$. Kolmas indeks (k) on pandud sulgudesse, kuna vaatluse väärtus

on teada niipea, kui on teada indeksid i ja j. Eeldame , et $e_{ij(k)} \sim N(0, \sigma_e^2)$. α_i on töötluse i mõju, β_j on rea j mõju ja $\gamma_{(k)}$ on veeru k mõju, kus $i = 1, ..., a$; $j = 1, ..., b$; $k = 1, ...a$; $N = a^2$ Kitsendused on $\sum \alpha_i = 0$, $\sum \beta_j = 0$ ja $\sum \gamma_{(k)} = 0$.

4.2.3 Ruutude summad ja dispersioonanalüüsi tabel

Varieeruvus andmetes on tingitud töötlusest (A), ridadest (B) ja veergudest (C). Kõik ülejäänud varieeruvused on juhuslikud ja kuuluvad jääkidele. Kirjutame samasuse

$(y_{ij(k)} - \bar{y}_{...}) = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{..(k)} - \bar{y}_{...}) + (y_{ij(k)} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..(k)} + 2\bar{y}_{...})$.

Kui tõstame selle avaldise ruutu ja summeerime üle kõigi N vaatluse, saame

$(y_{ij(k)} - \bar{y}_{...})^2 = \sum(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum(\bar{y}_{.j} - \bar{y}_{...})^2 + \sum(\bar{y}_{..(k)} - \bar{y}_{...})^2 + \sum(y_{ij(k)} - \bar{y}_{i.} - \bar{y}_{.j} + y_{..(k)} + \bar{y}_{...})^2$,

$SS_T = SS_A + SS_B + SS_C + SS_e$.

Dispersioonanalüüsi tabel:

Allikas	Vabadusastmed	SS	MS	E(MS)
Faktor A	$a-1$	SS_A	MS_A	$\sigma_e^2 + Q(A)$
Rida B	$a-1$	SS_B	MS_B	$\sigma_e^2 + Q(B)$
Veerg C	$a-1$	SS_C	MS_C	$\sigma_e^2 + Q(C)$
Jääk	$(a-1)(a-2)$	SS_e	$MS_e = s_e^2$	σ_e^2
Kokku	$N-1$	SS_T		

4.2.4 Analüüs arvuti abil

Sellist tüüpi katse mudel sümbolkujus on Y = Rida Veerg Töötlus

Analüüs SAS paketiga

Meie näite SAS programm on järgmine:

```
DATA ruut;
INPUT rida veerg töötlus $ y;
CARDS
1 1 A 24
1 2 B 20
1 3 C 19
1 4 D 24
1 5 E 24
2 1 B 17
2 2 C 24
2 3 D 30
2 4 E 27
2 5 A 36
3 1 C 18
3 2 D 38
3 3 E 26
3 4 A 27
3 5 B 21
4 1 D 26
4 2 E 31
4 3 A 26
4 4 B 23
4 5 C 22
5 1 E 22
5 2 A 30
5 3 B 20
5 4 C 29
5 5 D 31
PROC GLM;
CLASS rida veerg töötlus;
MODEL y = rida veerg töötlus;
RUN;
```

Tulemus on:

General Linear Models Procedure
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	548.00000	45.66667	4.28	0.0089
Error	12	128.00000	10.66667		
Corrected Total	24	676.00000			
	R-Square	C.V.	Root MSE		Y Mean
	0.810651	12.85821	3.2660		25.400
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rida	4	68.00000	17.00000	1.59	0.2391
Veerg	4	150.00000	37.50000	3.52	0.0404
Töötlus	4	330.00000	82.50000	7.73	0.0025

Olulisi tulemusi võime saada paarisvõrdlemisel. Analüüs Minitab'iga on analoogiline, seepärast ei lisanud me seda siia.

4.3 Ülesanded

4.1

Katses võrreldi kolme väetist: A, B ja C. Katses oli 12 maalappi, mis oli jagatud 4 plokki, igas plokis 3 maalappi. Saadi järgmised tulemused:

	Väetis		
Plokk	A	B	C
1	69	72	60
2	75	74	64
3	70	78	65
4	66	68	55

A. Selgita dispersioonanalüüsi abil, kas väetistel on erinev mõju. Analüüs peab sisaldama hüpoteese, dispersioonanalüüsi tabelit, tulemuste kontrolli ja järeldusi.

B. Väetis A on standardväetis ning B ja C on uued väetised. Võrdle sobival viisil iga uut väetist standardväetisega.

C. Missuguseid eeldusi on vaja teha analüüsiks?

4.2

Uuriti kolme erineva pesulahuse mõju bakterite kasvu pidurdamiseks piimakonteinerites. Analüüs tehti laboratooriumis ning igal päeval tehti ainult kolm katset. Kuna päev võib esitada varieeruvust, jagas eksperimenteerija vaadeldava "päeva" plokiliseks faktoriks. Vaatluseid koguti nelja päeva jooksul. Andmestik on esitatud järgnevas tabelis, kus arvud näitavad bakterite hulka peale pesemist (st madal väärtus näitab, et enamus baktereid hävis).

	Päev			
Lahus	1	2	3	4
A	13	22	18	39
B	16	24	17	44
C	5	4	1	22

A. Selgita dispersioonanalüüsi abil, kas lahustel on olulist erinevust bakterite hävitamisel.

B. Kas on olulist erinevust lahuste A ja B vahel?

C. Missuguseid eeldusi on vaja teha analüüsiks?

4.3

Uuriti, kas jahihooaeg mõjutab hirvede harjumuspärast käitumist. Valiti välja neli liikumisteed, mida hirved kasutasid. Määrati keskmine jälgede arv nädalas iga tee kindlal teelõigul enne jahihooaega, jahihooajal ja peale jahihooaega. Saadi

järgmised tulemused:[2]

Tee	Enne	Jahihooajal	Peale
1	62,5	57,0	49,0
2	46,5	53,3	50,0
3	45,0	59,3	37,0
4	24,0	35,7	50,0

A. Selgita dispersioonanalüüsi abil, kas jahihooaeg mõjutab hirvede harjumusi. Püstita formaalsel viisil hüpoteesid.

B. Kui ülesandes A leiti oluline erinevus, siis kasuta mitmese võrdlemise meetodit, et näidata erinevuse olemasolu.

C. Missuguseid eeldusi on vaja teha analüüsiks?

4.4

Uuriti nelja hübriidteraviljasordi vastupidavust seenhaigustele. Seejuures ei olnud midagi teada nende saagikuse potentsiaali kohta. Igat hübriidi kasvatati sama maa viies erinevas piirkonnas. Saadi järgmised toodangud:

	Piirkond					
Hübriid	NW	NE	C	SE	SW	Keskmine $\bar{Y}_{i\cdot}$
FR-11	62,3	64,0	64,3	65,0	66,4	64,40
BCM	63,3	62,7	66,2	66,8	64,5	64,70
DBC	60,8	64,3	65,2	62,2	65,1	63,52
RC-3	55,4	56,0	59,8	58,0	58,8	57,60
Keskmine $\bar{Y}_{\cdot j}$	60,45	61,75	63,875	63,00	63,70	$\bar{Y}_{\cdot\cdot}$ =62,555

$\sum\sum Y_{ij} = 1251,1; \quad \sum\sum Y_{ij}^2 = 78480,51$

Kirjelda sobivate dispersioonanalüüsi vahenditega seda andmestikku. Kas erinevad hübriidid annavad erinevat toodangut? Kas esineb piirkonna mõju? Missuguseid soovitusi võib anda hübriidide valiku kohta? Missuguseid eeldusi on vaja teha analüüsiks?

4.5

Uuriti nelja väetise (A, B, C ja D) mõju nisusaagile. Kasutati nelja nisusorti ning nelja põldu. Iga põld jaotati neljaks võrdseks plokiks ning kasutati ladina ruutude katseplaani. Katsetulemused on esitatud tabelis.

[2]Andmed on võetud: D. Brown, Dept. of Biology, Ratford University; and Ratford University Consulting Service, October 1990.

	Põld				
Nisusort	1	2	3	4	Keskmine
1	35,5 (A)	24,5 (B)	14,7 (C)	35,5 (D)	27,55
2	14,3 (B)	6,2 (C)	13,7 (D)	24,5 (A)	14,67
3	14,1 (C)	16,2 (D)	34,3 (A)	19,7 (B)	21,07
4	15,0 (D)	64,5 (A)	34,6 (B)	19,0 (C)	33,27
Keskmine	19,72	27,85	24,32	24,67	24,14

Väetiste (A, B, C ja D) keskväärtused olid vastavalt 39,70; 23,27; 13,5 ja 20,10.

A. Tee dispersioonanalüüsi tabel.

B. Koosta mudel ja mudeli kasutamise eeldused.

C. Kontrolli olulisuse nivoo $\alpha = 0,01$ korral, kas väetiste kasutamine põhjustab saagi erinevuse.

D. Kontrolli olulisuse nivoo $\alpha = 0,01$ korral, kas nelja nisusordi saagikus omavahel erines oluliseselt.

E. Leia punkthinnang igale nisusordile.

F Leia iga väetise jaoks keskmise saagi hinnang.

4.6

Mercer ja Hall (1911) ja Fisher (1925) analüüsisid klassikalises katses peedi juure kaalu. Katses oli viis töötlust: A, B, C, D ja E. Põld jaotati 25 tükiks ja tükid olid paigutatud ruutudena suurusega 5·5. Igat töötlust kasutiti üks kord igas reas ja veerus. Tulemused olid järgmised, kus iga töötluse lapile on sulgudes antud juure kaal põllulapil.

	Veerg					
Rida	1	2	3	4	5	Keskmine
1	D (376)	E (371)	C (355)	B (356)	A (335)	358,6
2	B (316)	D (338)	E (336)	A (356)	C (332)	335,6
3	C (326)	A (326)	B (335)	D (343)	E (330)	332,2
4	E (317)	B (343)	A (330)	C (327)	D (336)	330,0
5	A (321)	C (332)	D (317)	E (318)	B (306)	318,8
Keskmine	331,1	342,0	334,4	340,0	327,8	

Analüüs on tehtud paketiga Minitab.

General Linear Model

Factor	Levels	Values				
Row	5	1	2	3	4	5
Column	5	1	2	3	4	5
Treat	5	1	2	3	4	5

Analysis of Variance for KAAL

Source	DF	Seq SS	Adj SS	Adj MS	F
Row	4	4240.2	4240.2	1060.1	
Column	4	701.8	701.8	175.5	
Treat	4	330.2	330.2	82.6	
Error	12	1754.3	1754.3	146.2	
Total	24	7026.6			

Küsimused.

A. Formuleeri ja kontrolli sobivaid hüpoteese. Leia p väärtus täpsusega, mida võimaldab tabel.

B. Püstita eeldused ja arutle milliseid neist on vaja analüüsiks.

C. Kontrolli hüpoteesi, et töötluse A keskväärtus on erinev teiste töötluste keskmisest.

Peatükk 5

Mitmefaktoriline dispersioonanalüüs

Mitmefaktorilises katses on mitmed faktorid kombineeritud samas katses. Ideaalne oleks, kui ühe faktori iga tase oleks kombineeritud teise faktori iga tasemega. Öeldakse, et faktorid ristuvad.

5.1 Kaks fikseeritud faktorit

5.1.1 Sissejuhatav näide

Soovime võrrelda kahte sorti seemneid (1 ja 2) kolme erineva väetamise taseme juures. Seega meil on 2 3 = 6 töötluse kombinatsiooni. Jagame põllu 24 tükiks. Paigutame juhuslikult 6 töötluse kombinatsiooni põllutükile nii, et iga töötluse kombinatsioon kordub neli korda. Näiteks võib tulemus välja näha põllul järgmiselt (tähistame vastavalt L, M või H madala, keskmise ja kõrge väetise taseme jaoks ning 1 või 2 vastava seemnesordi jaoks):

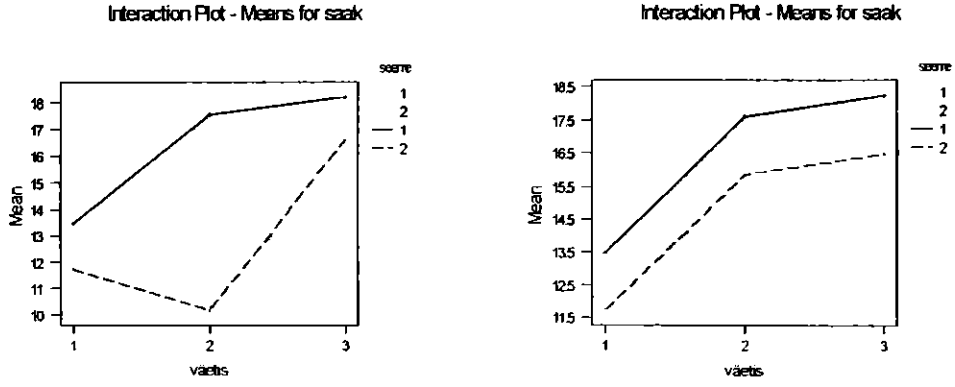
L1 14,3	M1 18,1	H1 17,6	M1 17,6	L1 14,5	M1 17,1
H1 18,2	H1 18,9	H2 15,7	H2 17,5	H2 16,7	H2 16,6
M2 10,5	M1 17,6	L1 11,5	M2 12,8	L2 12,6	L2 11,2
L2 11,0	M2 8,3	H1 18,2	M2 9,1	L2 12,1	L1 13,6

Moodustame andmestikust järgmise tabeli:

	1=madal	2=keskmine	3=kõrge	kokku
seeme 1	14,3	18,1	17,6	
	14,5	17,6	18,2	
	11,5	17,1	18,9	
	13,6	17,6	18,2	
keskmine	13,475	17,6	18,225	16,433
seeme 2	12,6	10,5	15,7	
	11,2	12,8	17,5	
	11,0	8,3	16,7	
	12,1	9,1	16,6	
keskmine	11,725	10,175	16,625	12,842
üldkeskmine	12,600	13,888	17,425	14,638

Lõppkokkuvõtteks kontrollime selle andmestiku kohta hüpoteese: kas erinevad seemnesordi saagid on võrdsed? Kas erineva väetisekoguse mõjul saame erineva seemnesaagi?

Aga enne, kui seda teeme, vaatame järgnevaid graafikuid, kus on esitatud erinevate töötluste kombinatsioonide keskväärtused:



Vasakpoolsel graafikul näeme (seda kinnitame formaalselt hiljem), et seemnesordi 1 korral suureneb saak väetise hulga kasvades ja seemnesordil 2 on madalaim keskmine saagikus. Näeme ka, et sellel on väikseim saak väetisehulga 2 korral ja väetisehulga 3 korral on saak peaaegu sama, mis seemnesordi 1 korral. See võib olla koosmõju näiteks kahe faktori vahel.

Parempoolsel graafikul on keskväärtuste erinevus ühesugune kõigi väetisehulkade korral. Kui selline graafik kirjeldab reaalsust, siis faktoritevaheline koosmõju puudub.

Selle analüüsi tähtis osa on hinnata, kas faktorite vahel leidub mõnda olulist koosmõju.

5.1.2 Mudel ja kitsendused

Faktoril A on a taset ja faktoril B on b taset. Tähistame ühe vaatluse y_{ijk}, kus $i = 1, ..., a$, $j = 1, ..., b$ ja $k = 1, ..., n$. Seega on vaaatluste üldarv $N = abn$. Eeldame, et katse on tasakaalus.

Iga töötluse kombinatsiooni keskväärtus $\bar{y}_{ij.} = \left(\sum_{k=1}^{n} y_{ijk}\right)/n$. Selle avaldise järgi arvutatud väärtused on esitatud eelpool vaadeldavas näites andmete tabelis reas "keskmine" Faktori B kindla taseme keskväärtus $\bar{y}_{.j.} = \left(\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}\right)/na$. Faktori A kindla taseme keskväärtus $\bar{y}_{i..} = \left(\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}\right)/nb$. Üldkeskmine $\bar{y}_{...} = \left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}\right)/nba$.

Mudel on $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$. Eeldame , et $e_{ijk} \sim N(0, \sigma_e^2)$. Kitsendused on $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=0}^{b} \beta_j = 0$ ja $\sum_{i=1}^{a}(\alpha\beta)_{ij} = \sum_{j=0}^{b}(\alpha\beta)_{ij} = 0$.

5.1.3 Hälvete ruutude summad ja dispersioonanalüüsi tabel

Mingi vaatluse ja üldkeskmise erinevuse võib kirjutada $(y_{ijk} - \bar{y}_{...}) = (\bar{y}_{ij.} - \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{ij.}) = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (y_{ij.} - \bar{y}_{.j.} - \bar{y}_{.j} + \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{ij.})$.

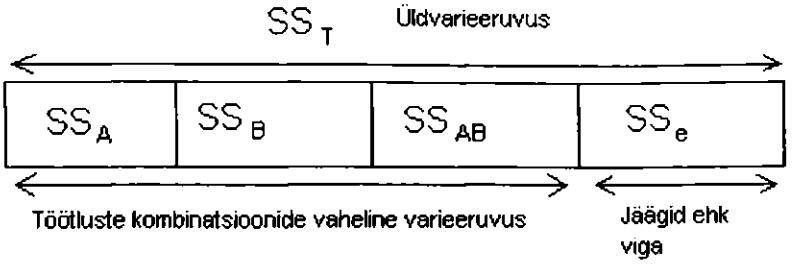
Eelmises avaldises jaotasime vaatluse ja üldkeskmise erinevuse esiteks töötluse kombinatsiooni (Tööt) $(\bar{y}_{ij.} - \bar{y}_{...})$ ja jääkide $e_{ijk} = (\bar{y}_{ijk} - \bar{y}_{ij.})$ järgi. Seejärel lahutasime töötluste kombinatsiooni mõju faktori A $(\bar{y}_{i..} - \bar{y}_{...})$ mõjuks, faktori B $(y_{.j.} - \bar{y}_{...})$ mõjuks ja faktorite A ja B koosmõjuks $(y_{ij.} - \bar{y}_{.j.} - \bar{y}_{.j} + \bar{y}_{...})$.

Peale ruutu tõstmist ja summeerimist saame:
$\sum_{ijk}(y_{ijk} - \bar{y}_{...})^2 = \sum_{ijk}(\bar{y}_{ij.} - \bar{y}_{...})^2 + \sum_{ijk}(\bar{y}_{ijk} - \bar{y}_{ij.})^2 = \sum_{ijk}(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_{ijk}(\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_{ijk}(y_{ij.} - \bar{y}_{.j.} - \bar{y}_{.j} + \bar{y}_{...})^2 + \sum_{ijk}(\bar{y}_{ijk} - \bar{y}_{ij.})^2$

Kokkuvõtvalt võib selle esitada (mudeli all on iga osa vabadusastmete arv)

SS_T	=	$SS_{TÖÖT}$	+	SS_e	=	SS_A	+	SS_B	+	SS_{AB}	+	SS_e
N-1		(ab-1)		(N-ab)		(a-1)		(b-1)		(a-1)(b-1)		(N-ab)

Järelikult oleme üldvarieeruvuse jaotanud järgmisteks osadeks:



$SS_{Tööt}$ ei ole tavaliselt dispersioonanalüüsi tabeli osa, seda kasutatakse faktorite koosmõju SS_{AB} arvutuste kergendamiseks.

Dispersioonanalüüsi tabel.

Allikas	Vabadusastmed	SS	MS	E(MS)
Faktor A	$a-1$	SS_A	MS_A	σ_e^2+Q(A, A*B)
Faktor B	$b-1$	SS_B	MS_B	σ_e^2+Q(B, A*B)
Koosmõju A*B	$(a-1)(b-1)$	SS_{AB}	MS_{AB}	σ_e^2+Q(A*B)
Jäägid	$N-ab$	SS_e	MS_e=s_e^2	σ_e^2
Kokku	$N-1$	SS_T		

Märgime, et $Q(A*B) = n\sum\sum(\alpha\beta)_{ij}^2/(a-1)(b-1)$. See on koosmõju mõõt. Faktori A mõju ja võimaliku koosmõju mõõt on $Q(A, A*B)$. Kui $Q(A*B) = 0$, siis $Q(A, A*B) = Q(A) = nb\sum\alpha_i^2/(a-1)$. $Q(B, A*B)$ on faktori B mõju ja võimaliku koosmõju mõõt. Kui $Q(A*B) = 0$, siis $Q(B, A*B) = Q(B) = na\sum\beta_i^2/(b-1)$.

Kui faktoritel (A ja B) võib esineda koosmõju, siis ongi vaja alustada selle olemasolu kontrollist. Kui selgub, et koosmõju ei ole oluline, alles siis kontrollime peamõjusid.

Seega hüpoteesi

$H_0 : Q(A*B) = 0$ (koosmõju ei ole),

$H_1 : Q(A*B) > 0$ (koosmõju esineb)

kontrollimiseks kasutame F-testi $F_{(a-1)(b-1),(N-ab)} = MS_{AB}/s_e^2$.

Kui koosmõju ei ole oluline, siis kontrollime peamõjusid:

H_0 $Q(A) = 0$ (faktoril A ei ole),

$H_1 : Q(A) > 0$ (faktoril A on mõju).

Selleks kasutame F-testi $F_{(a-1),(N-ab)} = MS_A/s_e^2$. Analoogiliselt kontrollime faktori B mõju.

H_0 $Q(B) = 0$ (faktoril B ei ole),

$H_1 : Q(B) > 0$ (faktoril B on mõju).
Kasutame F-testi $F_{(b-1),(N-ab)} = MS_B/s_e^2$.

5.1.4 Analüüs arvutil

Mudel sümbolkujul

Kahefaktorilise dispersioonanalüüsi mudeli kirjutame sümbolites $Y = A \ B \ A * B$, kus A ja B on faktorid. Meie näites kasutame mudelit

$$SAAK = Seeme \ V\ddot{a}etis \ Seeme * V\ddot{a}etis.$$

Analüüs paketiga SAS

SAS-i programm koos andmetega on järgmine:

```
OPTIONS LS=60;
OPTIONS PS=75;
DATA kaksfaktorit;
INPUT Seeme Väetis $ SAAK;
CARDS;
1 L 14.3
1 L 14.5
1 L 11.5
1 L 13.6
1 M 18.1
1 M 17.6
1 M 17.1
1 M 17.6
1 H 17.6
1 H 18.2
1 H 18.9
1 H 18.2
2 L 12.6
2 L 11.2
2 L 11.0
2 L 12.1
2 M 10.5
2 M 12.8
2 M  8.3
2 M  9.1
2 H 15.7
2 H 17.5
2 H 16.7
2 H 16.6
;
PROC GLM;
CLASS Seeme Väetis;
MODEL SAAK = Seeme Väetis
  Seeme*Väetis;
MEANS Seeme Väetis Seeme*Väetis;
RUN;
QUIT;
```

Analüüsi tulemused on:

General Linear Models Procedure

Dependent Variable: SAAK

Source	DF	Sum of Squares	F Value	Pr > F
Model	5	221.37875000	36.23	0.0001
Error	18	21.99750000		
Corrected Total	23	243.37625000		

R-Square	C.V	SAAK Mean
0.909615	7.552374	14.6375000

Source	DF	Type I SS	F Value	Pr > F
SEEME	1	77.40041667	63.33	0.0001
VÄETIS	2	99.87250000	40.86	0.0001
SEEME*VÄETIS	2	44.10583333	18.05	0.0001

Source	DF	Type III SS	F Value	Pr > F
SEEME	1	77.40041667	63.33	0.0001
VÄETIS	2	99.87250000	40.86	0.0001
SEEME*VÄETIS	2	44.10583333	18.05	0.0001

General Linear Models Procedure

Level of SEEME	N	SAAK Mean	SAAK SD
1	12	16.4333333	2.34106943
2	12	12.8416667	3.09969451

Level of VÄETIS	N	SAAK Mean	SAAK SD
1	8	12.6000000	1.38770108
2	8	13.8875000	4.18208338
3	8	17.4250000	1.04163333

Level of SEEME	Level of VÄETIS	N	SAAK Mean	SAAK SD
1	1	4	13.4750000	1.37204227
1	2	4	17.6000000	0.40824829
1	3	4	18.2250000	0.53150729
2	1	4	11.7250000	0.75443135
2	2	4	10.1750000	1.97209702
2	3	4	16.6250000	0.73654599

Analüüs näitab, et koosmõju on oluline. Sel juhul peamõjude kontrollimine ei oma mõtet.

Analüüs paketiga Minitab

Minitab'i analüüs annab selle andmestiku põhjal järgmise tulemuse.

General Linear Model

Factor	Type	Levels	Values
Seeme	fixed	2	1 2
Väetis	fixed	3	1 2 3

Analysis of Variance for SAAK, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Seeme	1	77.400	77.400	77.400	63.33	0.0002
Väetis	2	99.873	99.873	49.936	40.86	0.0002
Seeme*Väetis	2	44.106	44.106	22.053	18.05	0.0002
Error	18	21.997	21.997	1.222		
Total	23	243.376				

Unusual Observations for SAAK

Obs	SAAK	Fit	StDev Fit	Residual	St Resid
3	11.5000	13.4750	0.5527	-1.9750	-2.06R
14	12.8000	10.1750	0.5527	2.6250	2.74R

R denotes an observation with a large standardized residual.

5.1.5 Eelnevale analüüsile järgnev analüüs

Analüüs juhuks, kui esineb koosmõju

Kui koosmõju on oluline, ei ole mõtekas kõnelda faktori A üldisest mõjust. Siis on faktori A mõju faktori B erinevatele tasemetele erinev Järgneva analüüsi käigus on vaja võrrelda töötluste kombinatsioonide keskväärtusi μ_{ij}. Võime leida kontrastid $L = \sum h_{ij}\mu_{ij}$, kus $\sum h_{ij} = 0$. Kontrasti hinnang avaldub $\hat{L} = \sum h_{ij}\bar{y}_{ij}$. Dispersioon kontrasti hinnangust on $D(\hat{L}) = \sigma_e^2 \sum \frac{h_{ij}^2}{n}$, mille hinnang omakorda on $\hat{D}(\hat{L}) = MS_e \sum \frac{h_{ij}^2}{n}$.

Näide. Oletame, et soovime kontrollida, kas esineb erinevust tugevalt väetatud põldudel kasvanud taimede seemnesaakide vahel. Andmete tabelist saame, et keskväärtus seemnesordi 1 korral on 18,225 ja seemnesordi 2 korral 16,625. Leiame kontrasti $L = 1 \quad \mu_{13} - 1 \quad \mu_{23}$. Kontrasti hinnang on $\hat{L} = 18,225 - 16,625 = 1,6$. Dispersiooni hinnang kontrastile on $\hat{D}(\hat{L}) = 1^2 \frac{s_e^2}{n} + (-1)^2 \frac{s_e^2}{n} = 1,222 \quad \frac{2}{4} = 0,611$. Et selgitada, kas kontrast on oluline, on vaja arvutada $t = \frac{\bar{y}_{13} - \bar{y}_{23}}{\sqrt{s_e^2 \frac{2}{n}}} = \frac{18,225 - 16,625}{\sqrt{1,222 \frac{2}{4}}} = \frac{1,6}{\sqrt{0,611}} = 2,05$. Hetkel on vabadusastmete arv 18. 5%-lise olulisuse nivoo korral t kriitiline väärtus on 2,101. Järelikult erinevus ei ole oluline 5%-lise olulisuse nivoo juures.

Analüüs juhuks, kui koosmõju ei ole

Kui koosmõju ei ole, siis jätkame analüüsi faktorile A ja faktorile B eraldi. Keskväärtuse $\mu_{i.}$ kontrast leitakse $L = \sum h_i \mu_i$, kus $\sum h_i = 0$. Kontrasti hinnang avaldub $\hat{L} = \sum h_i \bar{y}_i$. Dispersioon kontrasti hinnangust on $D(\hat{L}) = \sigma_e^2 \sum \frac{h_i^2}{nb}$. Standardviga arvutatakse $\widehat{SE}_{\hat{L}} = \sqrt{s_e^2 \sum \frac{h_i^2}{nb}}$, kus vabadusastmete arv on $(N - ab)$. Faktori B jaoks on arvutused analoogilised. Kontrasti usaldusvahemik arvutatakse $\hat{L} \pm t_{N-ab} \widehat{SE}_{\hat{L}}$.

5.1.6 Kaks fikseeritud faktorit plokk-katses

Praktikas planeeritakse paljud katsed randomiseeritud plokk-katsena. Taimearetuses kasutatakse sageli mõistet "replikatsioon", kui kogu katset korratakse iseseisvates plokkides.

Näide. Porrulaugu ökoloogilise toodangu katses kasvatati taimi ridade erinevatel kaugustel: 50 cm või 70 cm. Taimi väetati (kompost) erinevatel aegadel (0 nädal, 2 nädalat või 4 nädalat peale istutamist). Neid $2 \cdot 3 = 6$ töötluse kombinatsiooni korrati kolmes plokis, mille vaatluste üldarv on $N = 2 \ 3 \ 3 = 18$. 16 nädala pärast mõõdeti taimedes mingi kindla aine hulk. Katse tulemused on järgnevas tabelis.

Plokk	Väetamine	Kaugus	Aine
1	0 nädalat	50 cm	15,78
1	2 nädalat	70 cm	19,01
1	4 nädalat	70 cm	11,93
1	0 nädalat	70 cm	13,79
1	4 nädalat	50 cm	11,69
1	2 nädalat	50 cm	17,04
2	4 nädalat	70 cm	10,35
2	2 nädalat	50 cm	11,86
2	0 nädalat	50 cm	7,71
2	0 nädalat	70 cm	11,41
2	4 nädalat	50 cm	9,73
2	2 nädalat	70 cm	14,21
3	2 nädalat	70 cm	12,59
3	2 nädalat	50 cm	13,93
3	4 nädalat	70 cm	11,32
3	4 nädalat	50 cm	11,71
3	0 nädalat	70 cm	13,44
3	0 nädalat	50 cm	14,06

Võib arvata, et sellist tüüpi katsete korral ei ole koosmõju plokkide ja peafaktorite vahel. Mudel peab sisaldama peamõjusid aga ka koosmõjusid faktorite vahel. Katses, kus faktoril A on a taset, b plokki ja faktoril C on c taset (nii et N=abc), on dispersioonanalüüsi tabel:

Allikas	Vabadusastmed	SS	MS	E(MS)
plokk	b-1	SS_B	MS_B	$\sigma_e^2 + Q(Plokk)$
A	a-1	SS_A	MS_A	$\sigma_e^2 + Q(A, A*C)$
C	c-1	SS_C	MS_C	$\sigma_e^2 + Q(C, A*C)$
AC	(a-1)(c-1)	SS_{AC}	MS_{AC}	$\sigma_e^2 + Q(A*C)$
jäägid	(ac-1)(b-1)	SS_e	MS_e	σ_e^2
kokku	N-1	SS_T		

Mudel sümbolkujus on

$$AINE = Plokk\ Väetamine\ Kaugus\ Väetamine * Kaugus.$$

SAS väljatrükk põhineb eelneval mudelil.

General Linear Models Procedure

Dependent Variable: AINE AINE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	89.839638	12.834234	4.10	0.0221
Error	10	31.279566	3.127957		
Corrected Total	17	121.119204			

R-Square	C.V.	Root MSE	AINE Mean
0.741746	13.74822	1.7686	12.864

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLOKK	2	47.805226	23.902613	7.64	0.0097
VÄETAMINE	2	40.304653	20.152326	6.44	0.0159
KAUGUS	1	1.153301	1.153301	0.37	0.5572
VÄETAMINE*KAUGUS	2	0.576458	0.288229	0.09	0.9127

Näeme, et kolm plokki annavad $3 - 1 = 2$ vabadusastet. Töötluse kombinatsioonid $a{\cdot}b = 6$ anavad 5 vabadusastet, mille jaotame järgmiselt: väetamisele $a - 1 = 2$, kaugusele $b - 1 = 1$ ja koosmõjule $(a - 1)(b - 1) = 2 \quad 1 = 2$ vabadusastet. Väljatrükist näeme, et olulised on ploki ja väetamise mõjud. Analüüsi tuleks jätkata nagu eelpool paarisvõrdlusega.

5.1.7 Rohkem kui kaks faktorit

Kahefaktorilise katse analüüsi meetodeid võib üldistada ka katse jaoks, kus on rohkem kui kaks faktorit. Vaatleme tasakaalus katset kolme faktoriga, kus faktorite A, B ja C tasemete arv on vastavalt a, b ja c. Kõiki töötluste kombinatsioone korrati n korda, seega kõikide vaatluste arv $N = a \quad b \quad c \quad n$. Katses esineva varieeruvuse võib jagada vastavalt dispersioonanalüüsi tabelile.

Allikas	Vabadusastmed	SS
A	a-1	SS_A
B	b-1	SS_B
C	c-1	SS_B
AB	(a-1)(b-1)	SS_{AB}
AC	(a-1)(c-1)	SS_{AC}
BC	(b-1)(c-1)	SS_{BC}
ABC	(a-1)(b-1)(c-1)	SS_{ABC}
Jääk	abc(n-1)	SS_e
Kokku	N-1	

Kui faktorid on fikseeritud, siis kõikide mõjude kontrollimiseks kasutatakse F-testi murru nimetajas MS_e-d. Võib esineda mitut tüüpi koosmõjusid, mis teevad interpreteerimise keeruliseks. Hetke situatsioonis võib esineda nii koosmõju kahe faktori vahel, kui ka koosmõju kõigi kolme faktori vahel. Niisuguste koosmõjude interpreteerimiseks on vaja omada head ettekujutusvõimet. Kui teostame katse (1 replikatsioon) ainult üks kord (n=1), siis jäägi vabadusastmete arv on 0. Sel juhul proovime mudeli koostada ilma kolme faktori koosmõjuta, siis käitub see koosmõju kui mudeli viga.

Kui katse on tasakaalustamata, siis kõik faktorid ja koosmõjud säilitavad oma vabadusastmed vähemasti seni, kui katses ei ole tühje lahtreid. Vabadusastmete vähenemine võib toimuda mudeli vea (jäägi) hinnaga.

5.1.8 Tasakaalustamata katsed

Hälvete ruutude summade erinevad tüübid

Üldvarieeruvuse jaotamine osadeks on korduvates katsetes ühene ainult tasakaalus katsete korral st kõikide töötluse kombinatsioonide jaoks on replikatsioonide arv ühesugune. Kui katse on tasakaalustamata kas katseplaani põhjal või kogemata, siis hälvete ruutude summad sõltuvad faktorite mudelisse toomise järjekorrast. Faktori B hälvete ruutude summa SS sõltub sellest, kas faktor A on juba mudelis või mitte. See fakt võib tasakaalustamata katses segada andmete interpreteerimist. SAS väljatrükkides võime näha nelja erinevat tüüpi hälvete ruutude summasid.

I tüüp (I type SS) tähendab, et hälvete ruutude summa arvutatakse iga faktori jaoks kui muutus SS_e-s kui faktor lisatakse mudelisse esinemise järjekorras vastavalt etteantud mudelile. Kui mudel on Y = A B A*B, siis SS_A arvutatakse nagu oleks tegemist ühefaktorilise katsega (mudel: Y=A). Siis

arvutatakse SS_B, mis tähendab SS_e ümberarvutamist. Mudel on sel juhul Y= A B ja lõpuks leiame koosmõju SS, kui lisame mudelisse ka faktorite koosmõju. Mõnikord nimetatakse I tüüpi hälvete ruutude summasid ka järjestikusteks hälvete ruutude summadeks.

II tüüp (II type SS) tähendab, et igale faktorile arvutatakse SS siis, kui faktor on lisatud mudelisse viimasena välja arvatud koosmõjudele. Kõik peamõjud, mis on koosmõjude osad, peavad olema kaasa arvatud. Mudeli Y = A B A*B korral SS arvutatakse järgmiselt: SS(A|B), SS(B|A) ja SS(AB|A,B).

III tüüp (III type SS) on katse arvutada SS tasakaalus katsena. Selle arvutamise põhimõtet me siin ei selgita. Üldiselt kasutatakse III tüüpi SS-i, kui katse on tasakaalustamata. Sellega kaasnev probleem on, et kõikide faktorite koosmõjude hälvete ruutude summad ei ole üldiselt samad kui üld SS. Minitab'i väljatrükis on III tüübi asemel "Adjusted Sum of Squares"

IV tüüp (IV type SS) erineb III tüübist tühjade lahtrite käsitlemise meetodite poolest st nad on lõpetamata katsed.

Kui katse on tasakaalus, siis kõik SS-id on võrdsed. Praktikas tehakse tasakaalustamata katse analüüs kasutades III tüüpi SS-i (või "Adjusted Sum of Squares" Minitab'is). Kahjuks ei ole see meetod eksimatu.

Vähimruutkeskmine

Tasakaalustamata katses ei ole küllalt selge, mida tähendab keskväärtus. Arutleme selle üle väikese näite põhjal. Katses on faktorid A ja B, kumbki kahe tasemega.

	B1	B2	Keskmine
A1	10, 12	13, 15, 14, 16, 15, 17	14
A2	14, 16	19, 22, 21, 24, 23, 26	20,625

Kui arvutame keskväärtused igale neljale lahtrile, saame

	B1	B2	Keskmine
A1	11	15	13
A2	15	22,5	18,75

Kõigi A1 vaatluste keskväärtus on 14 ja kõigi A2 vaatluste keskväärtus 20,625. Teist tüüpi keskväärtusi saame teisest tabelist kui "keskväärtus keskmisest" 11 ja 15 keskväärtus on 13, mis on ka A1 keskväärtus. Vastav A2 keskväärtus on 18,75. Keskväärtuse arvutamise see viis püüab anda keskväärtuse, mis saadakse tasakaalus katse korral. SAS-is on selliste keskväärtuste arvutamiseks funktsioon LSMEANS. Selle kohta on samuti võimalik

kontrollida hüpoteese ja leida usaldusvahemikke. Järgnev SAS programm arvutab meie näite jaoks mõlemad keskväärtused: keskväärtuse tavalises mõttes ja LSMEANS-i.

```
PROC GLM;
CLASS a b;
MODEL y = a b a*b;
MEANS a /t;
LSMEANS a / t;
RUN;
```

Programmis funktsioon MEANS arvutab tavalise keskväärtuse ja LSMEANS vähimruutkeskmise ("least squares means"). Mõlemal juhul anname a väärtuse t-testi teostamiseks, et võrrelda andmestikus A1 ja A2.

```
General Linear Models Procedure

                T tests (LSD) for variable: Y(

NOTE: This test controls the type I comparisonwise error rate not
      the experimentwise error rate.

              Alpha= 0.05  df= 12  MSE= 3.625
                 Critical Value of T= 2.18
            Least Significant Difference= 2.0742

Means with the same letter are not significantly different.

          T Grouping          Mean     N  A
                  A       20.6250     8  2
                  B       14.0000     8  1


General Linear Models Procedure
                      Least Squares Means

          A              Y   T / Pr > |T| H0:
                    LSMEAN   LSMEAN1=LSMEAN2
          1   13.0000000       -5.23088
                                0.0002
          2   18.7500000
```

Üldine kommentaar. Tasakaalustamata andmete korral on töötluste vahelise võrdlemise jaoks sobiv arvutada vähimruutkeskmist, aga samade piirangutega, kui hälvete ruutude summa III tüüpi korral. See ei lahenda veel kõiki probleeme. Kui võimalik püüa teha tasakaalus katset.

5.2 Ülesanded

5.1

Uuriti lindude pliisisaldust veres. Kolmest erinevast liigist (harakas, kotkas, öökull) püüti neli lindu ja analüüsiti nende verd. Igale linnule anti number (1 kuni 4) ja märgiti üles linnu liik ning pliisisaldus veres. Andmete analüüsimisel kasutati järgnevat SAS programmi:

```
DATA Linnud;
Input Plii Liik $ Linnunr;
CARDS;
49 Harakas 1
44 Harakas 2
45 Harakas 3
45 Harakas 4
48 Öökull  1
47 Öökull  2
53 Öökull  3
48 Öökull  4
50 Kotkas  1
54 Kotkas  2
52 Kotkas  3
57 Kotkas  4
;
PROC GLM;
CLASS Liik Linnunr;
MODEL Plii = Liik Linnunr;
MEANS Liik;
RUN;
QUIT;
```

Üks osa väljatrükist on järgmine:

```
Dependent Variable: Plii
Source           DF Sum of Squares  Mean Square  F Value  Pr > F
Model             5  119.16666667  23.83333333  2.49     0.1491
Error             6   57.50000000   9.58333333
Corrected Total  11  176.66666667

          R-Square      C.V          Root MSE      Plii Mean
          0.674528    6.275059   3.09569594      49.33333333

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

Liik              2  113.16666667  56.58333333    5.90    0.0382
Linnunr           3    6.00000000   2.00000000    0.21    0.8869
```

Statistikud väidavad, et ei ole õige lisada analüüsi linnu number (LINNUNR).

A. Tee midagi statistikute väite vastu! Selgita, miks või miks mitte.

B. Tee uus analüüs andmestikule, kus LINNUNR ei ole lisatud mudelisse. See tähendab, et sinu analüüs peab andma sama tulemuse kui SAS programm

```
PROC GLM;
CLASS Liik;
MODEL Plii = Liik;
RUN;
```

C. Jätka oma analüüsi t-testiga, kus võrdle kotkast öökulliga.

5.2

Järgnevad andmed on saadud rohttaimede uurimusest. Katses vaadeldi kahte liiki rohttaimi. Neid niideti kaks, kolm ja neli korda aastas. Katse teostati neljas plokis. Tulemused on järgmised:

	Elevandirohi			Guatemala rohi		
Plokk	2	3	4	2	3	4
1	109	222	187	277	246	252
2	97	125	163	293	263	181
3	133	134	143	260	194	224
4	113	173	179	325	190	248

Andmestiku analüüsimisel kasutati SAS-i. Üks osa väljatrükist on järgmine:

Allikas	Vabadus-astmed	Hälvete ruutude summa	Keskruudud
G: rohu tüüp		57526,04	
H: Niitmine/aasta		225,00	
G * H		18176,33	
Plokk		4478,46	
Viga		12734,79	
Kokku		93140,63	

A. Täienda eelnevat tabelit, täites vabadusastmete ja keskruutude veerud.

B. Püstita asjakohane hüpotees ja kontrolli seda.

C. Interpreteeri mudeli koosmõjud. Missuguse rohttaime tüübi ja niitmise kombinatsioon osutub parimaks?

5.3

Uuriti fungitsiidi (Captan) ja kahe pestitsiidi (Dildren ja Diazinon) mõju faasanite munatoodangule. Märgiti üles iga katses oleva faasani munade arv. Tehti kaheksa vaatlust igas kombinatsioonis. Tulemused:

	Pestitsiid		
Fungitsiid	Ei ole	Dieldren	Diazinon
Ei ole	15 29 18 30 25 21 26 24	22 32 7 24 21 5 9 18	30 25 28 21 17 19 16 30
Captan	37 12 5 38 24 21 6 10	1 13 1 10 6 18 20 25	9 7 0 1 0 4 4 9

Dispesioonanalüüsi tulemuste üks osa on esitatud:

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
FUNG		1026.75000			
PEST		595 29167			
FUNG*PEST		570 37500			
Error		2968.25000			
Total		5160 66667			

A. Täienda eelnevat tabelit, täites vabadusastmete ja keskruutude veerud.

B. Püstita asjakohane hüpotees ja kontrolli seda.

C. Interpreteeri mudeli koosmõjud. Kas mõni fungitsiidi ja pestitsiidi kombinatsioon on eriti ohtlik munade toodangule?

D. Kas on mõnda tunnust, et analüüsis olevad eeldused võivad olla rikutud? Soovita, mida teha, et parandada analüüsi.

5.4

Katses uuriti paberi tugevust mõjutavaid faktoreid. Paberimassi keedeti 3 tundi. Uuriti järgmisi muutujaid:

lehtpuupuit – lehtpuupuidu protsent massis (2%, 4% või 8%);

rõhk – rõhk keetmise ajal (400, 500 või 650);

tugevus – paberi tugevus.

Tehti kolm katset igale kombinatsioonile lehtpuupuidu massi protsendi ja rõhu vahel. Järgneva SAS programmi abil saab andmestiku analüüsida:

```
DATA Paber;
INPUT Puit Rõhk Tugevus;
CARDS;
2       400     196.6
2       400     196.9
2       400     195.8
2       500     197.7
2       500     196.0
2       500     197.0
2       650     199.8
2       650     205.4
2       650     200.1
4       400     198.5
4       400     197.2
4       400     197.7
4       500     195.6
4       500     196.0
4       500     196.9
4       650     198.4
4       650     198.0
4       650     197.6
8       400     197.5
8       400     196.6
8       400     197.0
8       500     195.6
8       500     196.2
8       500     195.5
8       650     197 4
8       650     198.1
8       650     197 7
;
PROC GLM DATA=Paber;
CLASS Puit Rõhk;
MODEL Tugevus = Puit Rõhk
                Puit*Rõhk;
RUN;
```

Järgnev on üks osa tulemuste väljatrükist. Mõne tulemuse kohale on asendatud küsimärk.

```
Dependent Variable: Tugevus
Source      DF    Sum of Squares   Mean Square   F Value   Pr>F
Model       ??     77.46666667          ??
Error       ??     24.66000000          ??
Total       ??    102 12666667
     R-Square       C.V          Root MSE          Tugevus Mean
         ?          0.592710     1 17046999        197.47777778
Source      DF     Type I SS     Mean Square   F Value   Pr>F
Puit         ?      9.42888889          ??
Rohk         ?     40 78222222          ??
Puit*Rohk    ?     27.25555556          ??
```

A. Täienda eelnevat tabelit, arvutades vabadusastmete arv ja keskruudud.

B. Kasutades F-testi leia, kas mõnel faktoril on oluline mõju.

C. Arvuta R-square.

D. Intrerpreteeri koosmõjud.

5.5

Uurimuses püüti määrata vee hulga ja taimeliigi mõju hernetaime varre pikkusele. Kasutati kahte liiki taimi ja kolme erinevat veehulka. Uuriti kaheksateist lehtedeta taime, mis jaotati juhuslikult kolme alagruppi ja igale grupile määrati juhuslik veehulk. Sarnast ptotsetuuri korrati ka 18 tavalise taime korral. Koostati järgnev SAS programm andmete analüüsimiseks:

```
OPTIONS PS=55;
DATA Hernes;
INPUT Tüüp  $ Vesi  $ Pikkus;
CARDS;
lehtedeta      madal     69.0
lehtedeta      madal     71.3
lehtedeta      madal     73.2
lehtedeta      madal     75.1
lehtedeta      madal     74.4
lehtedeta      madal     75.0
lehtedeta     keskmine   96.1
lehtedeta     keskmine  102.3
lehtedeta     keskmine  107.5
lehtedeta     keskmine  103.6
lehtedeta     keskmine  100.7
lehtedeta     keskmine  101.8
lehtedeta      kõrge    121.0
lehtedeta      kõrge    122.9
lehtedeta      kõrge    123.1
lehtedeta      kõrge    125.7
lehtedeta      kõrge    125.2
lehtedeta      kõrge    120.1
tavaline       madal     71.1
tavaline       madal     69.2
tavaline       madal     70.4
tavaline       madal     73.2
tavaline       madal     71.2
tavaline       madal     70.9
tavaline      keskmine   81.0
tavaline      keskmine   85.8
tavaline      keskmine   86.0
tavaline      keskmine   87.5
tavaline      keskmine   88.1
tavaline      keskmine   87.6
tavaline       kõrge    101.1
tavaline       kõrge    103.2
tavaline       kõrge    106.1
tavaiine       kõrge    109.7
tavaline       kõrge    109.0
tavaline       kõrge    106.9
;
PROC GLM DATA=Hernes;
  CLASS Tüüp Vesi;
  MODEL Pikkus = Tüüp Vesi
                 Tüüp*Vesi;
RUN;
```

Analüüsi tulemused saadeti faksi teel. Kahjuks osa faksist oli loetamatu. Järgnev on ainult osa väljatrükist, mida oli võimalik lugeda:

Dependent Variable: PIKKUS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	12489.00000000	2497.800000	338.42	0.0001
Error	30	221.42000000	7.380667		
Corrected Total	35	12710.42000000			

Koosta täielik dispersioonanalüüsi tabel, mis põhineb andmestikul ja sinu dispersioonanalüüsi tabelil:

A. Kontrolli, kas taimeliigil on mõju varre pikkusele ja interpreteeri võimalikud olulised mõjud.

B. Kontrolli, kas vee hulgal on mõju varre pikkusele ja interpreteeri võimalikud olulised mõjud.

C. Kontrolli, kas esineb koosmõju vee hulga ja taimeliigi vahel. Kui on see nii, siis interpreteeri koosmõjud.

5.6

Uuriti, kas kuivaine sisaldus (DM) taimedes sõltub taimeliigist ja lõikamise ajast. Taimi kasvatati kasvuhoones 10-s potis, igas oli kahte liiki taimi; potid paigutati juhuslikult kasvuhoone lavadele. Iga viie lõikuse ajal võeti igast liigist kaks potti ja mõõdeti kuivaine sisaldus. Järgnev SAS programmi kasutati andmete analüüsimiseks:

```
DATA Q5;
INPUT Liik Aeg DM;
CARDS;
1      1      92
1      1      92
1      2      90
1      2      85
1      3      87
1      3      102
1      4      99
1      4      99
1      5      99
1      5      110
2      1      107
2      1      98
2      2      97
2      2      103
2      3      98
2      3      100
2      4      106
2      4      110
2      5      110
2      5      116
;
PROC GLM DATA=Q5;
  CLASS Liik Aeg;
MODEL DM=Liik Aeg Liik*Aeg;
RUN;
```

Saadi järgnev väljatrükk :

```
General Linear Models Procedure
Dependent Variable: DM
                          Sum of          Mean
Source          DF       Squares        Square     F Value      Pr>F
Model           ?     1024.00000          ??          ??          ??
Error           ?      272.00000          ??
Total           ?     1296.00000
            R-Square         C.V          Root MSE        DM Mean
            0 790123        5.215362      5.21536         100.000
Source     DF      Type III SS     Mean Square    F Value   Pr>F
Liik        ?      405.000000          ??            ??
Aeg         ?      584.000000          ??            ??
Liik*Aeg    ?       35.000000          ??            ??
```

Küsimused:

A. Koosta dispersioonanalüüsi tabel. Samuti kontrolli sobivat hüpoteesi ja sõnasta lõppjäreldused.

B. Missugune ajahetk andis erineva tulemuse? Küsimusele vastamiseks on vaja teha paarisvõrdlus, kasutades mõnda meetodit ühesuguse olulisuse nivooga.

C. Kas esineb olulist erinevust kahe taimeliigi vahel viiendal lõikusel?

5.7

Kakskümmend neli meest, kellel igal oli liigkaalu ligikaudu 10 kg, jaotati 12 töötluse vahel. Tekkisid kombinatsioonid nelja toiduratsiooni ja kolme erineva pikkusega sõrkjooksu vahel. Iga mees tarbis päeva jooksul võrdse kalorihulga, kuid toiduratsioonid erinesid proteiini, rasva ja süsivesinike proportsiooni poolest. Andmed on lisatud järgnevasse SAS programmi:

```
DATA Ulekaal;
INPUT Toit $Jooks Kaal_kadu;
Cards;
Normaalne  0 8.5
Normaalne  0 11.5
Normaalne  1 14.0
Normaalne  1 16.0
Normaalne  2 24.5
Normaalne  2 19.5
K_prot  0 15 5
K_prot  0 16.5
K_prot  1 20.0
K_rasv  1 13.0
K_rasv  1 11.0
K_rasv  2 22.0
K_rasv  2 27.0
K_karbo  0 15.5
K_karbo  0 13.5
K_karbo  1 21.0
K_karbo  1 18.0
K_karbo  2 24.5
K_karbo  2 27 5
;
PROC GLM DATA=Ulekaal;
```

```
K_prot  1 23.0
K_prot  2 27 0
K_prot  2 24.0
K_rasv  0 8.5
K_rasv  0 7.5
```

```
CLASS Toit Jooks;
MODEL Kaal_kadu = Toit Jooks
                  Toit*Jooks;
MEANS Toit Jooks /Tykey;
RUN;
```

SAS väljatrükk on:

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: WT_LOSS

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 12 MSE= 4.541667
Critical Value of Studentized Range= 4.199
Minimum Significant Difference= 3.6529

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	DIET
A	21.000	6	K_prot
A			
A	20.000	6	K_karbo
B	15.667	6	Normaalne
B			
B	14.833	6	K_rasv

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: KAAL_KADU

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 12 MSE= 4.541667
Critical Value of Studentized Range= 3.773
Minimum Significant Difference= 2.8427

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	JOoks
A	24.500	8	2
B	17.000	8	1
C	12.125	8	0

Küsimused:

A. Kas toiduratsioonide vahel on olulist erinevust? Kui on, siis interpreteeri erinevused.

B. Kas sõrkjooksu tasemete vahel on olulist erinevust? Kui on, siis interpreteeri erinevused.

C. Kas esineb mõnda olulist koosmõju toiduratsiooni ja sõrkjooksu vahel? Kui on, siis interpreteeri erinevused.

D. Mees, kes ei jooksnud ja sai "normaalset" toiduratsiooni, näib olevat kaalus maha võtnud. Kas see mahavõtmine on oluline?

E. Väljatrükk sisaldab järgnevat märkust Tukey meetodi juures:

```
NOTE: This test controls the type I experimentwise error rate,
but generally has a higher type II error rate than REGWQ.
```

Selgita, mida praktiliselt mõeldakse selle märkusega. Mida tähendab "type II error"?

5.8

Taimefüsioloog uurib üleujutuse mõju kahe puuliigi juure ainevahetusele: üleujutust taluvat jõekaske ja üleujutust mittetaluvat euroopa kaske. Kummagi liigi neli puud ujutati üheks päevaks üle ja nelja vaadeldi kontrollpuudena. Mõõdeti iga puu juures leiduv adenosiin triosphadi(ATP) kontsentratsioon. Andmestik (ATP molekulide arv mg koemahlas) on toodud järgnevas tabelis:

Jõekask		Euroopa kask	
Üleujutatud	Kontroll	Üleujutatud	Kontroll
1,45	1,70	0,21	1,34
1,19	2,04	0,58	0,99
1,05	1,49	0,11	1,17
1,07	1,91	0,27	1,30

Küsimused:

A. Püstita ja kontrolli sobivat hüpoteesi.

B. Arvuta 95%-lised usalduspiirid üleujutatud euroopa kase ATP nivoole.

C. Kas on olulist erinevust jõekase ja euroopa kase vahel võrreldes kontrollgrupiga?

5.9

Uuriti õhu reostusainete (osooni ja vaskdioksiidi) mõju. Teatud liiki uba kasvatati avamaal lavades. Mõned lavad suitsutati korduvalt vaskdioksiidiga. Mõne lava õhk filtreeriti süsinikuga, et eemaldada ümbritsevat osooni. Juhuslikult määrati

kolm lava ühe töötluse kombinatsiooni kohta. Kui töötlusest möödus üks kuu, mõõdeti iga lava ubade saak (kg).Tulemused on esitatud tabelis[1]

	OSOON PUUDUB		OSOON ESINEB	
	Vaskdioksiid		Vaskdioksiid	
	Puudub	Esineb	Puudub	Esineb
	1,52	1,49	1,15	0,65
	1,85	1,55	1,30	0,76
	1,39	1,21	1,57	0,69
Keskmine	1,587	1,417	1,340	0,700
Standardhälve	0,237	0,181	0,213	0,056

Arvutuste abistamiseks: $SS_{Viga} = 0{,}27513$; $SS_{kogu} = 1{,}62889$

Küsimused. Esita andmestiku täielik analüüs, et kontrollida, kas järgnevate faktorite mõju on oluline:

A. Osoon.

B. Vaskdioksiid.

C. Osooni ja vaskdioksiidi koosmõju.

D. Püstita analüüsiks vajalikud eeldused.

E. Kirjuta katse kohta väga lühidalt järeldused, mida võiks avaldada teaduslikus artiklis.

5.10

Uuriti kanga vastupidavust hõõrdumisel. Andmed väljendavad kanga kaalukaotust hõõrdumisel. Võrreldi erinevaid kanga tüüpe. Nad erinesid ka pinnatöötluste (pind), kasutatava täidisetüübi (täidis) ja kasutatava täidise protsendi (prop) poolest. Kontrolliti iga tüübi kahte kanga tükki. Järgnevat SAS programmi kasutati andmete analüüsiks:

[1] Andmet on võetud: Heggestad ja Bennet, Science (1981) pp 1507-1514.

```
DATA Vabrik;
INPUT Pind $ Täidis $ Prop Vastup;
CARDS;
YES A 25 192
YES A 25 188
YES A 50 217
YES A 50 222
YES A 75 252
YES A 75 263
NO  A 25 169
NO  A 25 152
NO  A 50 187
NO  A 50 196
NO  A 75 225
NO  A 75 270
YES B 25 127
YES B 25 105
YES B 50 123
YES B 75 117
YES B 75 125
NO  B 25  82
NO  B 25  82
NO  B 50  94
NO  B 50  89
NO  B 75  76
NO  B 75 105
;
PROC GLM DATA=Vabrik;
CLASS Pind Täidis Prop;
MODEL Vastup = Pind Täidis Prop
    Pind*Täidis        Pind*Prop
                Täidis*Prop
  Pind*Täidis*Prop;
RUN;
```

Saadi järgnev väljatrükk (mõned tulemused on kustutatud):

Dependent Variable: Vastup

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	11	90092.1250	8190.1932	40.86	0.0001
Error	12	2405.5000	200.4583		
Total	23	92497.6250			

R-Square	C.V	Root MSE	Vastup Mean
0.973994	8.939751	14.1583	158.375

Source	DF	Type I SS	Mean Square	F Value	Pr>F
Pind		5017.0417	5017.0417		
Täidis		70959.3750	70959.3750		
Prop		7969.0000	3984.5000		
Pind*Täidis		57.0417	57.0417		
Pind*Prop		44.3333	22.1667		
Täidis*Prop		6031.0000	3015.5000		
Pind*Täidis*Prop		14.3333	7.1667		

Küsimused:

A. Täienda dispersioonanalüüsi tabelit vabadusastmete arvu ja F väärtuste lisamisega. Missugune mõju on oluline? Missuguse olulisuse nivoo juures?

B. Iga olulise koosmõju jaoks valmista koosmõju joonis ja interpreteeri koosmõju.

C. Täidise tüübi B jaoks kontrolli, kas on olulist erinevust 25%-lise täidise ja 75%-lise täidise vahel.

Peatükk 6

Juhuslikud ja hierarhilised mudelid

6.1 Juhuslike faktoritega mudelid

Vaatleme katset, kus ühe või enama faktori tasemed esindavad juhuslikult valitud üldkogumi tasemeid. Näiteks loomade uurimise katse võib teha erinevates karjades, põlluvilja katse võib teha "ühe talu uuringuna" erinevates taludes, metsa katse võib teha erinevates metsades ja loomade toitumiskatses võib kasutada pullide juhuslikku valimit. Juhusliku faktori kasutamine katses mõjutab katse lõppjäreldusi. Vaatleme lühidalt ühe ja mitme juhusliku faktoriga andmeanalüüsi.

6.1.1 Ühefaktoriline dispersioonanalüüs

Näide

Eesti musta-valgekirju tõugu lehmade 40-st karjast valiti igast karjast juhuslikult 5 looma (n = 5). Mõõdeti iga lehma piima rasvasus. Järelikult on faktor "kari" 40 tasemel ($\alpha = 40$). Niisuguses situatsioonis ei huvita meid eriti karjade vahelise olulise erinevuse otsimine. Leides erinevuse karja 1 ja karja 2 vahel ei saa tulemust üldistada teiste karjade valimitele, kuna tõenäoliselt ei ole need kaks karja järgmiste valimite osad. Peamine huvi on hinnata populatsiooni keskväärtus μ, hinnata karjadevaheline dispersioon σ_A^2 ja lehmadevaheline dispersioon karjades σ_e^2. Niisuguseid dispersiooni hinnanguid nimetatakse *dispersiooni komponentideks.* Tähistame muutujad:

mõõdetud rasvasus on rasv, kari on kari ja lehma number on nr ning SAS -i programm on järgmine:

```
DATA piimarasv;
INPUT rasv kari nr;
CARDS;
3.80      1     1
3.55      1     2
3.59      1     3
3.38      1     4
3.71      1     5
3.52      2     1
3.28      2     2
3.52      2     3
2.95      2     4
.....jne
3.73     39     4
3.26     39     5
3.56     40     1
3.13     40     2
3.32     40     3
3.82     40     4
3.43     40     5
;
```

Mudel ja eeldused

Selle katse mudel on $y_{ij} = \mu + a_i + e_{ij}$, kus a_i on i-nda karja mõju ning i = 1,...,a, j = 1,...,n ja N = an. Märgime, et kasutame sümbolit a_i sümboli α_i asemel, sest on kujunenud tavaks kasutada kreeka tähti fikseeritud faktorite tähistamiseks ja ladina tähti juhuslike faktorite korral. Soovime hinnata keskväärtust μ ja dispersiooni komponente $D(a_i) = \sigma_A^2$ ning $D(e_{ij}) = \sigma_e^2$.

Dispersioonanalüüsi tabel

Formaalselt on hälvete ruutude summaks jaotamine sama kui fikseeritud faktori korral. Kuid oodatav keskruut E(MS) on erinev, nagu näeme järgmisest dispersioonanalüüsi tabelist:

Allikas	Vabadusastmed	SS	MS	E(MS)
A	a-1	SS_A	MS_A	$\sigma_e^2 + n\sigma_A^2$
Jäägid	N-1	SS_e	$MS_e = s_e^2$	σ_e^2
Kokku	N-1	SS_T		

Dispersiooni komponentide hindamine

On võimalik kontrollida hüpoteesi $H_0 : \sigma_A^2 = 0$, mille vastandhüpotees on $H_1 : \sigma_A^2 > 0$, kasutades F-jaotust $F(a-1, N-1) = MS_A/s_e^2$. Meid huvitavad dispersiooni komponendid. Näeme dispersioonanalüüsi tabeli veerust E(MS), et võime hinnata σ_e^2, kui $\hat{\sigma}_e^2 = MS_e$, millest saame $\hat{\sigma}_A^2 = (MS_A - MS_e)/n$.

Keskväärtuse μ usalduspiirid

Populatsiooni keskväärtuse μ hinnang on $\hat{\mu} = \bar{y}_{..}$ Siis kehtib $D(\bar{y}_{..}) = [n\sigma_A^2 + \sigma_e^2]/N$, mille hinnang on MS_A/N (a-1) vabadusastmega. Järelikult usalduspiirid määratakse avaldisest $\bar{y}_{..} \pm t_{(1-\alpha/2,a-1)}\sqrt{\frac{MS_A}{N}}$.

Analüüsimine arvutiga

Ka juhusliku faktori korral arvutused ei erine sellest kas faktor on fikseeritud. Arvutuste jaoks kasutame ikka protseduuri **GLM**, kuid nüüd anname ette, et faktor on juhuslik.

```
PROC GLM DATA=piimarasv;
CLASS kari;
MODEL rasv = kari;
RANDOM kari;
RUN;
```

Väljatrükist leiame dispersioonanalüüsi tabeli ja selle lõpust oodatava keskruudu. See lihtsustab dispersiooni komponentide arvutamise.

```
Dependent Variable: RASV

Source                    DF     Sum of Squares    F Value     Pr > F

Model                     39       4 00587000       1.87     0.0038
Error                    160       8.79128000
Corrected Total          199      12.79715000

                    R-Square             C.V            RASV Mean
                    0.313028        6.807160          3.44350000

Source                    DF       Type III SS  F Value     Pr > F
KARI                      39       4.00587000       1.87     0 0038

                  General Linear Models Procedure

Source       Type III Expected Mean Square
KARI      Var(Error) + 5 Var(KARI)
```

SAS-i teise protseduuriga MIXED võib dispersiooni komponendid arvutada otse. Seda protseduuri kasutatakse nagu protseduuri GLM, aga juhuslik faktor ei tohi siis olla mudeli osa. Ta antakse eraldi kui RANDOM. Sel juhul ei sisalda mudel ühtegi fikseeritud faktorit, seega mudel on tühi. Ning programm on:

```
PROC MIXED DATA=piimarasv;
  CLASS kari;
  MODEL rasv =  /SOLUTION;
  RANDOM kari;
RUN;
```

Võtmesõna SOLUTION kasutamise tulemusena hinnatakse üks fikseeitud parameeter μ ja programmi väljatrükk on:

```
Covariance Parameter Estimates (REML)

             Cov Parm       Estimate

             Kari         0.00955382
             Residual     0.05494550

              Model Fitting Information for RASV

            Description                          Value
            Observations                       200.0000
            Res Log Likelihood                  -8.5267
            Akaike's Information Criterion     -10.5267
            Schwarz's Bayesian Criterion       -13.8200
            -2 Res Log Likelihood               17.0535

                  Solution for Fixed Effects

Effect         Estimate     Std Error     DF       t  Pr > |t|

INTERCEPT    3.44350000    0.02266215     39  151.95    0.0001
```

Saame hinnangud $\sigma_A^2 = 0,00956$ ja $\sigma_e^2 = 0,055$. Samuti saame (väljatrükis intercept) $\hat{\mu} = 3,4435$, mille standardviga on 0,2266 ja vabadusastmete arv on 39. Seda võib kasutada μ usaldusvahemiku arvutamiseks 3,4455±2,02 0,02266; 3,4435±0,0458.

6.1.2 Kahefaktoriline dispersioonanalüüs

Ristuvate faktoritega mudelid võivad sisaldada kahte, ühte või mitte ühtegi juhuslikku faktorit. Katse näide, kus mõlemad faktorid on juhuslikud, on

niisugune, kus igast kaheteistkümnest juhuslikult valitud karjast mõlemad lehmad on ristatud iga viieteistkümne juhuslikult valitud pulliga. Katse tulemusena mõõdame näiteks järglaste kaalud. See võib olla juhuslike mõjudega mudel. Mõlemad karjad ja pullid on juhuslikud faktorid. Sellise andmestiku struktuur on:

	Kari j			
	1	2	...	12
Pull i				
1	XX	XX		XX
2	XX	XX		XX
...				
15	XX	XX		XX

Juhuslike mõjudega mudel, kus mõlemad faktorid on juhuslikud, on $y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{ijk}$, kus i=1, ..., a; j=1, ..., b; k=1, ..., n; N=abn. $a_i \sim N(0, \sigma_A^2)$; $b_i \sim N(0, \sigma_B^2)$; $ab_{ij} \sim N(0, \sigma_{AB}^2)$; $e_{ijk} \sim N(0, \sigma_e^2)$. Selles mudelis on peamised huviobjektid dispersiooni komponentide hinnangud ja üldkeskmine.

Juhul, kui üks faktoritest on fikseeritud ja teine on juhuslik, nimetatakse mudelit segamudeliks. Näiteks tahame võrrelda kolme erinevat sorti kanatoitu. Valime juhuslikult kaksteist kanatoitjat. Igaüks neist kasutab kahes puuris igat toitu. Siis kaalume kanad. Andmete struktuur on:

	Toitja j			
	1	2	...	12
Sööt i				
1	XX	XX		XX
2	XX	XX		XX
3	XX	XX		XX

Segamudel, kui faktor A on fikseeritud ja üks faktor B on juhuslik, on järgmine $y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{ijk}$, kus $i = 1, ..., a$; $j = 1, ..., b$; $k = 1, ..., n$; $N = abn$. $b_i \sim N(0, \sigma_B^2)$; $ab_{ij} \sim N(0, \sigma_{AB}^2)$; $e_{ijk} \sim N(0, \sigma_e^2)$.

Märgime, et koosmõju on juhuslik seni, kuni üks osadest on juhuslik. Selles mudelis võime kontrollida fikseeritud faktorite olulisust ja hinnata juhuslike faktorite dispersioonikomponete.

Hälvete ruutude summad ja dispersioonanalüüsi tabel

Hälvete ruutude summad arvutatakse ühtemoodi hoolimata sellest, kas faktorid on juhuslikud või fikseeritud. Keskruudu keskväärtus on ikkagi mõjuta-

tud. Võime võrrelda erinevaid mudeleid summeerides E(MS) dispersioonanalüüsi tabelis

Allikas	Vabadusastmed	SS	MS
A	a-1	SS_A	MS_A
B	b-1	SS_B	MS_B
AB	(a-1)(b-1)	SS_{AB}	MS_{AB}
Viga	(N-ab)	SS_e	MS_e
Kokku	N-1	SS_T	

Keskruudu keskväärtused erinevatele mudelitele on:

Allikas	Fikseeritud E(MS)	Sega E(MS)	Juhuslik E(MS)
A	$\sigma_e^2 + Q(A, AB)$	$\sigma_e^2 + n\sigma_{AB}^2 + Q(A)$	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$
B	$\sigma_e^2 + Q(B, AB)$	$\sigma_e^2 + n\sigma_{AB}^2 + na\sigma_B^2$	$\sigma_e^2 + n\sigma_{AB}^2 + na\sigma_B^2$
AB	$\sigma_e^2 + Q(AB)$	$\sigma_e^2 + n\sigma_{AB}^2$	$\sigma_e^2 + n\sigma_{AB}^2$
Viga	σ_e^2	σ_e^2	σ_e^2
Kokku			

Keskruudu keskväärtuste tabelit kasutame kahel juhul. Esiteks näitab E(MS), kuidas erinevaid dispersioonikomponente peaks arvutama. Teiseks annab E(MS) hüpoteeside kontrollimiseks sobivad teststatistikud.

Fikseeritud faktoritega mudelis kontrollitakse kõik faktorid ja koosmõjud keskruudu vea põhjal nagu näitasime 5. peatükis.

Märgime, et segamudelis faktori A ja koosmõju AB jaoks keskruudu keskmine E(MS) ainult eristab neid termis Q(A). Kontrollides hüpoteesi $H_0 : Q(A) = 0$, arvutame $F = MS(A)/MS(AB)$. Järelikult kontrollitakse fikseeritud faktori A koosmõju termi nimetajas, mitte vea termis. Dispersioonikomponentide on hinnangud on $\hat{\sigma}_e^2 = MS_e$, $\hat{\sigma}_{AB}^2 = (MS_{AB} - MS_e)/n$ ja $\hat{\sigma}_A^2 = (MS_B - MS_{AB})/a$.

Juhuslike faktoritega mudelis kontrollitakse peamõjud A ja B kasutades F-testi murru nimetajas MS(AB). Dispersioonikomponentide hinnangud on $\hat{\sigma}_e^2 = MS_e$, $\hat{\sigma}_{AB}^2 = (MS_{AB} - MS_e)/n$, $\hat{\sigma}_B^2 = (MS_B - MS_{AB})/a$ ja $\hat{\sigma}_A^2 = (MS_A - MS_{AB})/b$.

Nagu võib näha, on sedalaadi mudelite kontrollimine ja hindamine natuke petlik. Õnneks on abi käepärast. SAS protseduur Mixed võib käsitleda kõiki mudeli tüüpe ja isegi Minitab'i GLM osa on küllalt arukas selle töö jaoks. Vaatame mõningaid näiteid. Esiteks juhuslike faktoritega mudel. SAS programm on:

```
PROC MIXED;
CLASS A B;
MODEL Y=  / S DDFM=SATTERTH ;
RANDOM  A  B  A*B;
RUN;
```

Väljatrükk on:

Covariance Parameter Estimates (REML)

Cov Parm	Estimate
A	4.04846825
B	1.03990476
A*B	0.61820635
Residual	0.22777778

Model Fitting Information for

Description	Value
Observations	360.0000
Res Log Likelihood	-459.871
Akaike's Information Criterion	-463.871
Schwarz's Bayesian Criterion	-471.638
-2 Res Log Likelihood	919.7423

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT	30.19333333	0.60051962	21.3	50.28	0.0001

Programm annab meile dispersioonikomponentide hinnangu ja üldise keskväärtuse hinnangu (intercept), sisaldades standardviga ja selle vabadusastet.

Minitab'i analoogilises analüüsis antakse faktorid A ja B juhuslikena, küsitakse lisaväljatrükke E(MS) ja dispersioonikomponentide jaoks. Väljatrükk on:

General Linear Model

```
Factor     Type Levels Values
A          random     15  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
B          random     12  1  2  3  4  5  6  7  8  9 10 11 12

Analysis of Variance for y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
A          14   1380.784   1380.784     98.627   67.36  0.000
B          11    359.275    359.275     32.661   22.31  0.000
A*B       154    225.485    225.485      1.464    6.43  0.000
Error     180     41.000     41.000      0.228
Total     359   2006.544

Expected Mean Squares, using Adjusted SS

Source       Expected Mean Square for Each Term
1 A          (4) +  2.0000(3) + 24.0000(1)
2 B          (4) +  2.0000(3) + 30.0000(2)
3 A*B        (4) +  2.0000(3)
4 Error      (4)

Error Terms for Tests, using Adjusted SS

Source       Error DF  Error MS  Synthesis of Error MS
1 A            154.00     1.464  (3)
2 B            154.00     1.464  (3)
3 A*B          180.00     0.228  (4)

Variance Components, using Adjusted SS

Source    Estimated Value
A                  4.0485
B                  1.0399
A*B                0.6182
Error              0.2278
```

Analoogilise analüüsi segamudeli jaoks saab teha kasutades protseduuri MIXED järgmiselt:

```
PROC MIXED;
CLASS A B ;
MODEL Y = A;
RANDOM B A*B ;
 LSMEANS A / PDIFF;
RUN;
```

Väljatrükk on:

```
The MIXED Procedure

Class Level Information

Class     Levels  Values

A              3  1 2 3
B             12  1 2 3 4 5 6 11 12 13 14 15 16

Covariance Parameter Estimates (REML)

Cov Parm       Estimates

B              0.19676768
A*B            0.14242424
Residual       0.27555556

Tests of Fixed Effects

Source      NDF   DDF  Type III F  Pr > F

A             2    22       77.64  0.0001
```

Analoogiline analüüs Minitab'iga on:

```
General Linear Model

Factor     Type Levels Values
A         fixed      3 1 2 3
B        random     12  1  2  3  4  5  6 11 12 13 14 15 16

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
A           2    87.0178    87.0178    43.5089   77.64  0.000
B          11    19.1511    19.1511     1.7410    3.11  0.011
A*B        22    12.3289    12.3289     0.5604    2.03  0.028
Error      36     9.9200     9.9200     0.2756
Total      71   128.4178

Expected Mean Squares, using Adjusted SS

Source       Expected Mean Square for Each Term
1 A         (4) +  2.0000(3) + Q[1]
2 B         (4) +  2.0000(3) +  6.0000(2)
3 A*B       (4) +  2.0000(3)
4 Error     (4)

Error Terms for Tests, using Adjusted SS

Source       Error DF  Error MS  Synthesis of Error MS
1 A             22.00    0.5604  (3)
2 B             22.00    0.5604  (3)
3 A*B           36.00    0.2756  (4)

Variance Components, using Adjusted SS

Source    Estimated Value
B                  0.1968
A*B                0.14240
Error              0.27562
```

6.2 Hierarhilised mudelid

6.2.1 Rist- ja hierarhilised faktorid

Mõnikord on katsed tehtud nii, et samal katseühikul mõõdetakse mitu korda tulemus. Võib mõõta saagi kogust (või keemilist muutujat) põllul mitmetel väikestel maa-aladel või mõõta puu tihedust samast puust mitmete proovidega. Eesmärgiks võib olla täpsema väärtuse saamine kui seda saaks ühe mõõtmisega. Oletame näiteks, et teeme katse mingil alal, et võrrelda a töötlust (faktor A). Igal alal mõõdame b väiksel alal tulemuse.

Sellist tüüpi katse on looduses hierarhiline. Faktor B (ala) on hierarhiline faktorile A. See tähendab, et iga mõõtmise jaoks on erinevad alad. Võrdleme andmestruktuuri hierarhilises katses ristplaaniga:

A_1				A_a			
B_{11}	...	B_{1b}	...	B_{a1}	...	B_{ab}	
X		X		X		X	Hierarhiline plaan
X		X		X		X	
X		X		X		X	

	A_1			A_2			
B_1	X	X	X	X	X	X	Ristplaan
B_2	X	X	X	X	X	X	
B_3	X	X	X	X	X	X	

Ristplaanis faktori B iga tase esineb koos faktori A iga tasemega.

Hierarhilises plaanis on faktoril B uusi tasemeid (näiteks uus ala) faktori A iga taseme jaoks.

Hierarhilise plaani mudel on $y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$, kus $i = 1, ..., a$; $j = 1, ..., b$; $k = 1, ..., n$.

Antud juhul me ei käsitle käsitsi arvutamist. Vaatleme arvulist näidet, kus esimese astme faktor on fikseeritud.

Näide

Tahame võrrelda peedi kasvatamiseks kolme ökoloogilist väetist. Iga kolme väetise jaoks randomiseerime neli ala. Igal platsil on juhulsikult valitud kaks taime. Mõõdetakse raua sisaldus igas taimes. Tulemus võib olla:

Väetis	Ala	Taim	Raud
1	1	1	102,4
1	1	2	98,3
1	2	1	99,7
1	2	2	99,3
1	3	1	100,1
1	2	2	100,4
1	4	1	97,0
1	4	2	99,2
2	1	1	96,2
2	1	2	98,8
2	2	1	100,7
2	2	2	98,1
2	3	1	101,2
2	3	2	101,5
2	4	1	97,5
2	4	2	97,6
3	1	1	103,8
3	1	2	104,1
3	2	1	105,6
3	2	2	104,7
3	3	1	109,1
3	3	2	108,4
3	4	1	101,4
3	4	2	102,6

Alade kõikvõimalikud väärtused on siin hierarhilised ühele töötlusele, mille tähistame B(A). See tähendab: "B on hierarhiline A-le" Muidugi on see faktor juhuslik. Nende andmete analüüsimiseks kasutame protseduuri MIXED järgmises programmis:

```
PROC MIXED DATA=peet
CLASS väetis ala
MODEL raud = väetis
RANDOM ala(väetis);
RUN;
```

Väljatrükk on:

```
Covariance Parameter Estimates (REML)

Cov Parm          Estimate

ALA(Väetis)     3.35069444
Residual        1.55625000

  Model Fitting Information for RAUD

Description                          Value
Observations                       24.0000
Res Log Likelihood                -45.0707
Akaike's Information Criterion    -47.0707
Schwarz's Bayesian Criterion      -48.1152
-2 Res Log Likelihood              90.1413

        Tests of Fixed Effects

Source      NDF   DDF  Type III F  Pr > F
Väetis        2     9       10.57  0.0043
```

Väetiste mõjud on olulised ($p = 0{,}0043$). Märgime, et selles testis on murru lugejal 2 vabadusastet ja 9 vabadusastet nimetajal. Tegelikult on test sama kui see, mille saaksime, kui arvutaksime igal alal kahe taime keskväärtuse ja teeksime siis nendele keskväärtustele ühefaktorilise dispersioonanalüüsi. Illustreerime seda punkti, lisades ala keskväärtuse analüüsi paketiga Minitab.

General Linear Model

```
Factor     Type Levels Values
Väetis    fixed      3 1 2 3

Analysis of Variance for Mean, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
Väetis      2     87.301     87.301     43.651   10.57  0.004
Error       9     37.159     37.159      4.1292
Total      11    124.4612

Unusual Observations for Mean

Obs      Mean       Fit   StDev Fit  Residual   St Resid
21   108.750   104.963       1.016     3.788      2.15R

R denotes an observation with a large standardized residual.
```

Miks teha mitu mõõtmist

Kui praktikas on andmeanalüüs sama keskväärtuse analüüsiga iga ala jaoks, siis võib küsida, kas on vaja teha mõõtmisi rohkem kui üks. Vastus on, et see oleneb varieeruvuse hulgast ja uute katsete tegemise maksumusest. Katses arvutatakse täpsus töötluse keskväärtuse dispersioonina $D(\bar{y}_{i..})$. See tähendab, et $D(\bar{y}_{i..}) = D(\bar{b}_{i..}) + D(\bar{e}_{i..}) = \frac{\sigma^2_{B(A)}}{b} + \frac{\sigma^2_e}{bn}$. Kui suurendame n-i, siis avaldise viimane osa muutub väiksemaks, aga esimene osa ei muutu. Järelikult tõestab täpsus tegelikult midagi, kui suurendame n-i. Näite andmete korral on $\hat{D}(\bar{y}_{i..}) = 1,23$, kui $n = 1$ ja $\hat{D}(\bar{y}_{i..}) = 1,03$, kui $n = 2$ ja $\hat{D}(\bar{y}_{i..}) = 0,88$, kui $n = 10$. Kunagi ei saa 0,84-st väiksemat väärtust, isegi kui n on lõpmatu.

6.3 Ülesanded

6.1

Uuriti loomatoidu proteiini sisaldust. Moodustati kaks valimit 8-st juhuslikult valitud kotist. Mõõdeti igas kotis oleva toidu proteiini sisaldus. Kuid iga koti lipikule oli kirjutatud: "Proteiini sisaldus 16%"

	Koti nr								
	1	2	3	4	5	6	7	8	
	10	14	8	14	13	10	8	16	
	12	15	6	18	11	12	9	18	
$\sum$	22	29	14	32	24	22	17	34	194

Hinda mudeli parameetrid ja arvuta proteiini sisalduse 95%-line usaldusvahemik. Kas lipiku tekst on õige?

6.2

Analüüsiti 20 juhuslikult valitud kana 8-t muna. Hinda keskväärtus ja dispersioonikomponendid.

Andmed: $\sum\sum y_{ij} = 1200$, $\sum\sum y_{ij}^2 = 16790$, $\mathrm{SS}_A = 4990$.

6.3

Katsesse valiti juhuslikult 20 seafarmi (B). Kasutati 15 juhuslikult valitud kulti (A). Iga kult ristati igas farmis kahe emisega. Registreeriti sündinud põrsaste arv. Tulemused:

$\mathrm{SS}_A = 2296$, $\mathrm{SS}_B = 931$, $\mathrm{SS}_{AB} = 1064$, $\mathrm{SS}_T = 4591$, $\sum\sum\sum y_{ijk} = 8760$.

Hinda keskväärtus ja arvuta 95%-line usaldusvahemik. Hinda dispersioonikomponendid.

6.4

Toitumiskatsesesse valiti juhuslikult 20 lindlat (B). Võrreldi kolme erinevat sööta (A). Igas lindlas kasutati ühe sööda jaoks kümmet puuri. Söödad olid: A_1– kõrge proteiini tase + vitamiinide lisa; A_2– kõrge proteiini tase + ilma vitamiinideta; A_3– normaalne proteiini tase + vitamiinide lisa. Tulemused:

$\mathrm{SS}_A = 1596$, $\mathrm{SS}_B = 4750$, $\mathrm{SS}_e = 5400$, $\mathrm{SS}_T = 13266$;

Keskväärtused: A_1:15,3; A_2:12,0; A_3:11,7

Analüüsi katset.

Peatükk 7

Liigendatud plokk-katsed

7.1 Sissejuhatav näide

Liigendatud plokkplaan on mitmefaktorilise katseplaani erijuht. Seda on sobiv kasutada, kui katses oleva ühe faktori olemus on niisugune, et on kergem või odavam kasutada seda suure katse ühikuna, mida nimetatakse peaplokiks. Katse teisi faktoreid kasutame väiksemate ühikutena, mida nimetame alamplokkideks, ja nad paiknevad peaploki sees. Peaploki faktorite näited on: niisutus, pinnase töötluse jaoks kasutatavad tugevad masinad jne.

Liigendatud plokkplaanis hinnatakse peaploki faktorit väiksema täpsusega kui alamplokkide faktoreid. Järelikult võib mõningal juhul kasutada liigendatud plokkplaani ka juhul, kui uurija huvitub oma katses rohkem ühest faktorist. Faktor, millele on vaja täpsemat informatsiooni määratakse, alamplokile ja faktor, millele on piisavalt töötlemata hinnanguid, määratakse peaploki faktorina.

Näide. Katses uuriti heina kaalu. Peaploki faktor oli heinamaa tasemetega A ja B. Alamploki töötlus oli bakterite inokulatsioon kolmel tasemel: kontroll, elus ja surnud. Plokkides tehti neli replikatsiooni. Põldudel saadi järgmised tulemused:

Repl. 1	Repl.2	Repl. 3	Repl.4
K 29,4	S 28,7	S 29,7	K 26,7
E 34,4	E 33,4	K 28,6	E 31,8
S 32.5	K 28,9	E 32,9	S 28,9
K 27,4	E 36,4	K 27,2	S 28,6
E 34,5	S 32,4	E 32,6	E 30,7
S 29,7	K 28,7	S 29,1	K 26,8

Legend K=Kontroll E=Elus S=Surnud

Põld A Põld B

7.2 Mudel ja dispersioonanalüüsi tabel

Tähistame peaploki faktori A-ga, ploki B-ga ja alamploki faktori C-ga. Liigendatud plokk-katses näeb mudel veidi teisiti välja. Kui vaatleme plokki (ja järelikult ka AB koosmõju) juhuslikuna, siis mudeli võib kirjutada: $y_{ijk} = \mu + \alpha_i + b_j + \alpha b_{ij} + \gamma_k + \alpha\gamma_{ik} + e_{ijk}$. Siin on α_i peaploki faktori mõju, b_j ploki mõju ja γ_k alamploki faktori mõju. Kasutame kitsendusi $\sum \alpha_i = \sum \gamma_k = \sum \alpha\gamma_{ik} = 0$. Eeldame veel, et kõik juhuslikud faktorid on sõltumatud.

Üldvarieeruvuse võib jaotada järgmiseks dispersioonanalüüsi tabeliks:

Allikas	Vabadus-astmed	SS	MS	E(MS)
A (peaplokk)	a-1	SS_A	MS_A	$\sigma_e^2 + c\sigma_{ab}^2 + Q(A, AB)$
B (plokk)	b-1	SS_B	MS_B	$\sigma_e^2 + c\sigma_{ab}^2 + ac\sigma_b^2$
AB (peaploki viga)	(a-1)(b-1)	SS_{AB}	MS_{AB}	$\sigma_e^2 + c\sigma_{ab}^2$
C (alamplokk)	c-1	SS_C	MS_C	$\sigma_e^2 + Q(C, AC)$
AC	(a-1)(c-1)	SS_{AC}	MS_{AC}	$\sigma_e^2 + Q(AC)$
Alamploki viga	a(b-1)(c-1)	SS_e	MS_e	σ_e^2+
Kokku	N-1	SS_T		

Katsel on kahte tüüpi ühikuid: peaplokid ja alamplokid. Peaploki analüüs võib olla põhimõtteliselt ühefaktoriline dispersioonanalüüs peaploki keskväär-

tusele. Sama F-test kehtib, kui kontrollime peaploki faktori A sõltuvust koosmõjust AB. Seda võib näha ka keskruudu keskväärtuse avaldistes. Alamploki faktorit C kontrollitakse alamploki vea suhtes. Sellist tüüpi katsete korral kontrollitakse erinevaid faktoreid erinevalt: peaploki faktorit vastandatakse peaploki veaga ja alamploki faktorit alamploki veaga. Arvuti programmid võivad selle korrektseks teostamiseks vajada abi.

7.3 Analüüs arvutiga

7.3.1 Andmeanalüüs, kasutades protseduuri GLM

Eelpool esitatud liigendatud plokk-katse jaoks SAS programm võib olla järgmine: peaploki faktor on heinamaa, alamploki faktor on Inokul ja plokk on Replik.

```
DATA Splitpl;
INPUT Repl Cult $ Inoc $ Drywt;
CARDS;
1 A C 27.4
1 A D 29.7
    [data lines omitted here]
4 B D 28.6
4 B L 30.7
;
PROC GLM DATA=Splitpl;
  CLASS Repl Cult Inoc;
  MODEL Drywt = Repl Cult Repl * Cult Inoc Cult * Inoc;
  TEST H=Cult E=Repl * Cult;
RUN;
```

Pöörame tähelepanu TEST avaldisele. See avaldis käsib SAS-is kontrollida heinamaa faktorit, kasutades koosmõju Replik*Heinamaa kui termi viga, nagu eelpool öeldud. Samuti märgime, et SAS trükib Heinamaa kontrolli kasutades automaatselt alamploki vea mudeli veana. See kontroll ei ole korrektne.

Programmi SAS väljatrüki osad on esitatud järgnevas:

General Linear Models Procedure

Dependent Variable: DRYWT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	157.20833333	14.29166667	20.26	0.0001
Error	12	8.46500000	0.70541667		
Corrected Total	23	165.67333333			

R-Square	C.V.	Root MSE	DRYWT Mean
0.948905	2.761285	0.83989087	30.41666667

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
REPL	3	25.32000000	8.44000000	11.96	0.0006	
CULT	1	2.40666667	2.40666667	3.41	0.0895	#
REPL*CULT	3	9.48000000	3.16000000	4.48	0.0249	
INOC	2	118.17583333	59.08791667	83.76	0.0001	
CULT*INOC	2	1.82583333	0.91291667	1.29	0.3098	

Tests of Hypotheses using the Type III MS for REPL*CULT as an error term

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
CULT	1	2.40666667	2.40666667	0.76	0.4471	##

Märkus: (#) on heinamaa kontroll vale mudeli veaga. Ära kasuta seda! Selle asemel kasuta (##).

7.3.2 Analüüs protseduuriga Mixed

Vaadeldes teisiti liigenadatud plokk-katset, võib teda esitada segatud mudelite idee kaudu. Võime vaadelda plokke (replik) juhusliku faktorina ja seega ka koosmõju plokkidega (replik*heinamaa) kui juhuslikke faktoreid. Kasutame analüüsimiseks SAS protseduuri Mixed järgmises programmis:

```
PROC MIXED DATA=Splitpl;
  CLASS Repl Cult Inoc;
  MODEL Drywt = Cult Inoc Cult*Inoc;
  RANDOM Repl Repl*Cult;
RUN;
```

Väljatrükk on:

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
REPL	1.24748966	0.88000000	1.22640094	0.72	0.4730
REPL*CULT	1.15987399	0.81819444	0.86538380	0.95	0.3444
Residual	1.00000000	0.70541667	0.28798515	2.45	0.0143

Model Fitting Information for DRYWT

Description	Value
Observations	24.0000
Variance Estimate	0.7054
Standard Deviation Estimate	0.8399
REML Log Likelihood	-32.5313
Akaike's Information Criterion	-35.5313
Schwarz's Bayesian Criterion	-36.8669
-2 REML Log Likelihood	65.0626

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
CULT	1	3	0.76	0.4471
INOC	2	12	83.76	0.0001
CULT*INOC	2	12	1.29	0.3098

Analüüsimine protseduuriga Mixed on korrektne. Protsetuuriga Mixed analüüsi võib kasutada koos mitmese võrdlusega jne, seda ka protsetuuri GLM korral.

7.3.3 Analüüs, kasutades Minitab'i

Samasuguse analüüsi võime saada, kui kasutame Minitab'is mudelit
Heinamaa Replik Heinamaa*Replik Inoc Heinamaa*Inoc,
kus plokid (st Replik) on antud kui juhuslikud faktorid. Dispersioonanalüüsi tabel on:

```
General Linear Model
Factor        Type Levels Values
Heinamaa      fixed      2 A1         A2
Replik        random     4 1 2 3 4
Inokul        fixed      3 C          D          L
Analysis of Variance for Hein_kaal, using Adjusted SS for Tests
Source           DF     Seq SS     Adj SS     Adj MS       F      P
Heinamaa          1      2.407      2.407      2.407    0.76  0.447
Replik            3     25.320     25.320      8.440    2.67  0.221
Heinamaa*Replik   3      9.480      9.480      3.160    4.48  0.025
Inokul            2    118.176    118.176     59.088   83.76  0.000
Heinamaa*Inokul   2      1.826      1.826      0.913    1.29  0.310
Error            12      8.465      8.465      0.705
Total            23    165.6732
```

Kui soovime, võime saada ka keskruudu keskväärtuse E(MS) ja dispersiooni hinnangud:

```
Expected Mean Squares, using Adjusted SS
Source          Expected Mean Square for Each Term
1 Heinamaa         (6) +  3.0000(3) + Q[1; 5]
2 Replik           (6) +  3.0000(3) +  6.0000
3 Heinamaa*Replik  (6) +  3.0000
4 Inokul           (6) + Q[4; 5]
5 Heinamaa*Inokul  (6) + Q[5]
6 Error            (6)
Error Terms for Tests, using Adjusted SS
Source        Error DF  Error MS  Synthesis of Error MS
1 Heinamaa        3.00     3.160  (3)
2 Replik          3.00     3.160  (3)
3 Heinamaa*Replik 12.00    0.705  (6)
4 Inokul          12.00    0.705  (6)
5 Heinamaa*Inokul 12.00    0.705  (6)
Variance Components, using Adjusted SS
Source         Estimated Value
Replik               0.8800
Heinamaa*Replik      0.8182
Error                0.7054
```

Tulemused on samad, kui kasutasime protseduuri Mixed.

7.4 Ülesanded

7.1

Katses võrreldi kahe maisisordi toodangut. Katse tehti kahel põllul, kasutati lahutatud plokkide plaani. Iga põld jagati neljaks peaplokiks, neid kasteti erineva veehulgaga. Iga peaplokk jagati kaheks alamplokiks ning alamplokkidele külvati juhuslikult kahte sorti seemneid. Varieeruvuse erinevate allikate hälvete ruutude summad on esitatud järgnevas tabelis.

Allikas	Vabadusastmed	Hälvete ruutude summa	Keskruut	F
Plokk		133,9806		
Kastmine		63,7869		
Seeme		28,8906		
Kastmine*Seeme		1,0719		
Kastmine*Plokk		7,2619		
Viga		9,1425		
Kokku		244,1344		

A. Täienda tabelit: täida vabadusastmete ja keskruudu veerud.

B. Kontrolli, kas kastmise mõju on oluline.

C. Kontrolli, kas seemne mõju on oluline.

7.2

Liigendatud plokk-katse kavandati järgmisel viisil.

Katse tehti neljas plokis. Iga plokk jagati kuueks peaplokiks, mida töödeldi erinevate putukamürkidega. Iga peaplokk jagati viieks alamplokiks. Alamploki töötlus oli "kaugus põllu äärest" (Arvati, et putukamürgi mõju võib olla põllu keskel tugevam kui äärel.) Igas alamplokis märgiti üles kindlat liiki ellujäänud putukate arv.

A. Koosta selle katse kohta dispersioonanalüüsi tabel ja leia iga mõju vabadusastmete arv

B. Kuidas kontrollida hüpoteese:

i) et erinevatel mürkidel on sama mõju,

ii) et kaugusel põllu äärest ei ole mõju.

C. Selles katses uuritav tunnus on "peale mürgitamist ellujäänud putukate arv" Kas andmed vajavad enne analüüsi muutmist? Kui jah, siis missugused muutused võivad arvesse tulla?

7.3

Järgnev tabel iseloomustab Alfalfa katset. Katse kavandati kuues plokis. Iga plokk jagati kolmeks peaplokiks. Peaplokis kasvatati Alfalfa erinevaid sorte. Neid koristati esiteks kaks korda. Peaplokk jagati seejärel liigendatud plokkideks. Nende töötlus oli "kolmanda lõikuse kuupäev" Andmestik kõigi kolme lõikuse saagi kohta on esitatud järgnevas tabelis:

		Plokk					
Liik	Kuupäev	1	2	3	4	5	6
Ladak	ei ole	2,17	1,88	1,62	2,34	1,58	1,66
	1.sept	1,58	1,26	1,22	1,59	1,25	0,94
	20.sept	2,29	1,60	1,67	1,91	1,39	1,12
	7.okt	2,23	2,01	1,82	2,10	1,66	1,10
Cossack	ei ole	2,33	2,01	1,70	1,78	1,42	1,35
	1.sept	1,38	1,30	1,85	1,09	1,13	1,06
	20.sept	1,86	1,70	1,81	1,54	1,67	0,88
	7.okt	2,27	1,81	2,01	1,40	1,31	1,06
Ranger	ei ole	1,75	1,95	2,13	1,78	1,31	1,30
	1.sept	1,52	1,47	1,80	1,37	1,01	1,31
	20.sept	1,55	1,61	1,82	1,56	1,23	1,13
	7.okt	1,56	1,72	1,99	1,55	1,51	1,33

Järgnevat SAS programmi kasutati andmestiku analüüsimisel:

```
PROC GLM DATA=Alfalfa;
CLASS Liik Plokk Kuupäev;
MODEL SAAK = Liik Plokk Kuupäev Liik*Plokk Liik*Kuupäev;
RUN;
```

SAS väljatrükk on järgmine:

```
Dependent Variable: SAAK
Source      DF Sum of Squares Mean Square F Value Pr>F
Model       26      7.86321944  0.30243152   10.81 0.0001
Error       45      1 25854583  0 02796769
Total       71      9.12176528

R-Square         C.V       Root MSE      SAAK Mean
```

```
0.862028      10.47312      0 16723542    1.59680556
Source         DF     Type I SS
LIIK                  0.17801944
PLOKK                 4.14982361
Kuupaev               1.96247083
LIIK*PLOKK            1.36234722
LIIK*KUUPaEV          0.21055833
```

A. Kontrolli, kas liikide toodangus on olulist erinevust.

B. Kontrolli, kas erinevate kuupäevade toodang on erinev

C. Kontrolli ja interpreteeri koosmõju Liik*Kuupäev.

Peatükk 8

Dispresioonanalüüsi eelduste kontroll

8.1 Jääkide analüüs

Dispersioonanalüüsis tehakse statistiliste otsustuste langetamiseks järgmised eeldused:

1) juhuslikud vead on normaaljaotusega ja sõltumatud,

2) juhuslike vigade keskväärtus on null ja dispersioonid on võrdsed kõikidel faktori tasemetel,

3) kasutame põhjendatud mudelit.

Kaks esimest eeldust on seotud jääkide analüüsiga. Kontrollimaks, kas need eeldused on täidetud, peame kõigepealt arvutama jäägid.

Näide. Katses uuriti maisisaaki, mida kasvatati 6-l erineval põllul. Iga põld oli jaotatud neljaks põllulapiks, mida väetati erinevat tüüpi väetistega. Esitame järgneva katseandmestiku, kus maisisaak mõõdeti 4·6 =24 põllulapil.

Väetis		1	2	3	4	5	6	$\sum_{j=1}^{6} y_{ij}$	$\bar{y}_{i\cdot}$
1	Control	99	40	61	72	76	84	432	72
2	K_2O+N	96	84	82	104	99	105	570	95
3	K_2O+P_2O_5	63	57	81	59	64	72	396	66
4	N+P_2O_5	79	92	91	87	78	71	498	83
	$\sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij} =$	1896		$\bar{y}_{\cdot\cdot} =$	79				

Mudel on $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. Keskväärtuse μ hinnang on $\hat{\mu} = \bar{y}_{\cdot\cdot} = 79$.

Samuti saame hinnata töötluse α_i mõjusid $\hat{\alpha}_i = \bar{y}_i - \bar{y}_{..}$, mis on vastavalt $\hat{\alpha}_1 = -7$, $\hat{\alpha}_2 = 16$, $\hat{\alpha}_3 = -13$ ja $\hat{\alpha}_4 = 4$. Oleme hinnanud kõik mudeli parameetrid. Saame arvutatda y-i väärtuse hinnangu, arvestamata juhuslikku varieeruvust $\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i$. Kontrollgrupi jaoks saame $\hat{y}_{1j} = 79 + (-7) = 72$. Kõik $\hat{y}$ väärtused on selles lihtsas näites vastavate töötluste keskväärtused.

Kui $\hat{y}$ väärtused on arvutatud, saame arvutada jäägid $\hat{e}_{ij} = y_{ij} - \hat{y}_{ij}$. Kontrollgrupi esimese vaatluse korral $\hat{e}_{11} = 99 - 72 = 27$ Järgnevas tabelis on esitatud kõikide vaatluste jäägid.

Töötlus		1	2	3	4	5	6	$\sum$
1	Kontroll	27	-32	-11	0	4	12	0
2	$K_2O + N$	1	-11	-13	9	4	10	0
3	$K_2O + P_2O_5$	-3	-9	15	-7	-2	6	0
4	$N + P_2O_5$	-4	9	8	4	-5	-12	0

Oleme arvutanud jäägid, mida kasutame dispersioonanalüüsi eelduste kontrollimiseks.

8.2 Normaalsus

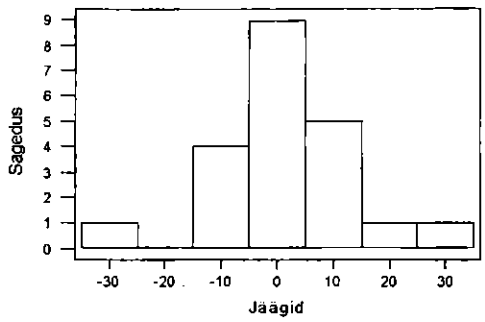
Hüpoteeside kontrollimine dispersioonanalüüsis on teostatav vaid siis, kui võib eeldada, et vaatlused igas grupis (töötluse kombinatsioonis jne) järgivad normaaljaotust. Teisiti väljendades võib öelda, et selle analüüsi jäägid on normaaljaotusega.

8.2.1 Kuidas leida kõrvalekaldeid normaaljaotusest

Lihtsaim võimalus kõrvalekallete avastamiseks normaaljaotusest on teha iga töötluse kombinatsioonile graafik. Sealt võime näha suuri hälbeid normaaljaotusest.

Mõnikord on randomiseeritud plokki-katses ainult üks vaatlus töötluse kombinatsiooni kohta. Sel juhul pole mõttekas kujutada iga kombinatsiooni eraldi graafikul. Sama kehtib katses, kus faktoritel on vähe replikatsioone. Kui valim koosneb kolmest vaatlusest, siis on raske määrata, kas andmestik järgib normaaljaotust. Palju tõhusam on uurida kogu katse jääke. Tuleb arvutada jäägid ja kujutada need graafikul, et näha, kas nad järgivad nor-

maaljaotust. Meie näite jääkide histogramm on järgmine:



Graafik ei näita tõsist kõrvalekaldumist normaaljaotusest.
SAS kasutamise näide.

```
PROC GLM;
   CLASS väetis;
   MODEL SAAK = Väetis;
   OUTPUT OUT=uus P=pred R=res;
RUN;
```

SAS määrangus OUTPUT luuakse uus andmestik (nimi on "uus"), et arvutada disprsioonanalüüsi jäägid ja prognooside väärtused ning need salvestatakse uude andmestikku vastavalt nimedega Res ja Pred. Seejärel koostatakse graafik.

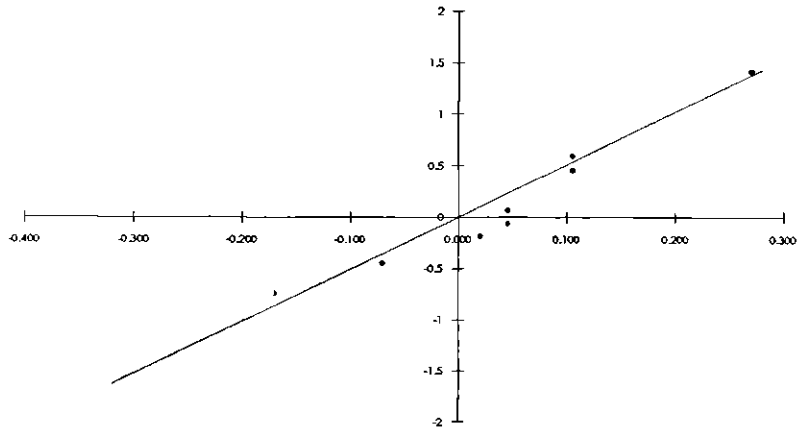
8.2.2 Normaalsuse kontroll

Histogrammi põhjal võib olla raske otsustada, kas jaotus on normaalne. Normaalsuse kontrolliks on olemas kindlat tüüpi graafik (normal probability plot). Vaatleme kuidas teha normaalsuse kontrolli.

1. Arvutame jäägid.
2. Sorteerime need arvulisse järjekorda.
3. Arvutame täpsustatud kumuleeritud suhtelise sageduse p, kasutame valemit (r-3/8)/(n+1/4) (täpsustamata alternatiiv oleks p=r/n).
4. Kasutame normaaljaotuse tabelit, arvutame normaaljaotuse jaoks Z, mis vastab täpsustatud suhtelisele sagedusele p. Näiteks p=0,031 annab $Z = -1,8663$ ja p=0,87 annab $Z = 1,1264$. Selles näites n=20.

(2) Jäägid (sorteeritud)	Järjekorra nr	(3) Täpsustatud kumuleeritud suhteline sagedus p	(4) Normaaljaotuse Z
-0,320	1	0,031	-1,8663
-0,320	2	0,080	-1,4051
-0,255	3	0,130	-1,1264
-0,255	4	0,179	-0,9192
-0,170	5	0,228	-0,7455
-0,080	6	0,278	-0,5888
-0,070	7	0,327	-0,4482
-0,030	8	0,377	-0,3134
0,020	9	0,426	-0,1866
0,045	10	0,475	-0,0627
0,045	11	0,525	0,0627
0,105	12	0,574	0,1866
0,105	13	0,623	0,3134
0,105	14	0,673	0,4482
0,105	15	0,722	0,5888
0,120	16	0,772	0,7455
0,120	17	0,821	0,9192
0,180	18	0,870	1,1264
0,270	19	0,920	1,4051
0,280	20	0,969	1,8663

Kui jäägid on normaaljaotusega, siis graafik peab olema (ligilähedaselt) lineaarne, st graafikul on jäägid horisontaalteljel ja normaaljaotuse Z vertikaalteljel. Eelpool olevate andmete graafik on järgmine (joon aitab jälgida lineaarsust):

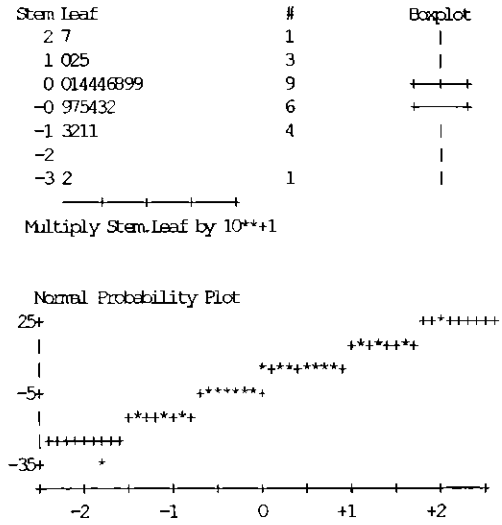


Näeme, et andmed paiknevad ligilähedaselt sirgel, järelikult on jäägid normaaljaotusega.

SAS protseduur UNIVARIATE annab (üsna elementaarse) normaalsuse kontrolli graafiku. SAS programm, mis leiab meie näitele jäägid, on järgmine (eeldame, et jäägid on esitatud andmestikus LISA nime all Res).

```
PROC UNIVARIATE PLOT DATA=LISA;
   VAR Res;
RUN;
```

SAS protseduur PLOT abil saame terve hulga teisi graafikuid: tüvi-leht-diagrammi, karpdiagrammi ja nn tõenäosuspaberi.



8.2.3 Normaalsuse kontroll

Kas andmed on normaaljaotusega? Selle kontrollimiseks on võimalik teha hüpoteesi kontroll. SAS protseduur Univariate sisaldab niisuguseid parameetreid, mis on esitatud järgnevas:

```
PROC UNIVARIATE NORMAL DATA=LISA;
   VAR Res;
RUN;
```

Valimi väikese mahu ja võtmesõna NORMAL korral teostatakse Shapiro-Wilks normaalsuse test. Kui valimi maht on suur (>2000), siis kontrollitakse hüpoteesi normaaljaotuse kohta, kasutades Kolmogorov-Smirnovi testi. Kui vastav p-väärtus on väike, siis hüpotees lükatakse tagasi, st et jaotus ei ole normaalne.

Esitame ühe osa protseduuri UNIVARIATE näiteandmestiku jääkide väljatrükist:

Univariate Procedure

Variable=RES

Moments

N	24	Sum Wgts	24
Mean	0	Sum	0
Std Dev	11.92732	Variance	142.2609
Skewness	-0.34902	Kurtosis	1.501019
USS	3272	CSS	3272
CV		Std Mean	2.434653
T:Mean=0	0	Pr>|T|	1.0000
Num ^= 0	23	Num > 0	12
M(Sign)	0.5	Pr>=|M|	1.0000
Sgn Rank	2	Pr>=|S|	0.9530
W:Normal	**0.96923**	**Pr<W**	**0.6483**

Quantiles(Def=5)

100% Max	27	99%	27
75% Q3	8.5	95%	15
50% Med	0.5	90%	12
25% Q1	-8	10%	-12
0% Min	-32	5%	-13
		1%	-32
Range	59		
Q3-Q1	16.5		
Mode	4		

Viimane rasvases kirjas rida näitab meile, et Shapiro-Wilks statistik on 0,94 ja p väärtus on 0,6483. Seega on ta suurem kui 0,05 ning me ei saa tagasi lükata hüpoteesi normaalsuse kohta.

8.2.4 Mis juhtub, kui andmed ei ole normaaljaotusega?

Kui normaaljaotuse eeldus on rikutud, siis tõenäosusavaldis, testid jne võivad olla valed. Kui valed nad on, see sõltub normaaljaotusesest kõrvalekaldumise astmest. Paljudel juhtudel, normaaljaotusest mõõduka kõrvalekaldumise korral, ei ole tulemuste õigsusele mingit tõsist mõju.

8.2.5 Parandused

Kontrolli, kas üksik "kauge" vaatlus põhjustab mittenormaalsust. Kaalutle, kas kustutada niisugune vaatlus andmestikust (niisugused 'erindid' on illustreeritud edaspidi).

Mõnikord võib andmeid teisendada, et muuta nad normaaljaotusele lähedasemaks.

8.3 Homoskedastilisus

Kahe valimi lihtsas t-testis eeldasime, et valimid pärinevad võrdsete dispersioonidega üldkogumist. Seda kasutatakse kui argumenti, et ühendada dispersioonid lihtsasse hinnangusse. Sama eeldust kasutatakse ka dispersioonanalüüsis: eeldatakse, et dispersioonid on kõigi töötluste kombinatsioonide korral ühesugused.

Mõned terminid. Juhul, kus kõik dispersioonid on võrdsed, nimetatakse seda homoskedastilisuseks. Kui dispersioonid ei ole võrdsed, on meil tegemist heteroskedastilisusega.

8.3.1 Kuidas kõrvaldada heteroskedastilisust?

1. Arvuta kõigi töötluste kombinatsioonide dispersioonid ja võrdle neid. Sel viisil saab avastada ilmse kõrvalekalde homoskedastilisusest.

2. Kahe valimi t-testi jaoks on võimalik kontrollida hüpoteesi, dispersioonide võrdsusest, st $H_0: \quad \sigma_1^2 = \sigma_2^2$, mille vastandhüpotees on, et nad ei ole võrdsed. Arvuta $F = \frac{s_1^2}{s_2^2}$, s.o kahe valimi dispersioonide suhe. Kasuta lugejas suurema dispersiooniga valimit! Võrdle seda väärtust F-jaotuse tabelis oleva väärtusega, kui vabadusastmed on (n_1 - 1; n_2 - 1). Kui olulisuse nivoo on 0,05, peame vaatama F-jaotuse tabelit, kus $\alpha = 0{,}025$.

Märgime, et SAS-i protseduur TTEST teeb selle kontrolli.

3. Eelpool kontrollisime testi, kui kaks dispersiooni olid võrdsed. Kui erinevate valimite dispersioonid pärinevad ühise dispersiooniga üldkogumist, siis on võimalik kontrollida hüpoteesi, kasutades Bartlett'i testi. SAS programm sel juhul on järgmine:

```
PROC GLM;
CLASS Väetis;
MODEL SAAK = Väetis;
MEANS Väetis / HOVTEST=BARTLETT;
RUN;
```

Millest saame järgmise väljatrüki:

Bartlett's Test for Equality of SAAK Variance

Source	DF	Chisq Value	Prob>Chisq
Väetis	3	5.4070	0.1443

Näeme, et test ei ole oluline (p=0,1443), seega ei saa hüpoteesi, et dispersioonid on võrsed, tagasi lükata.

4. Heteroskedastilisust saab avastada graafikul, kui kujutame graafikul jäägid ja y väärtuste hinnangud ($\hat{y}$ väärtused).

8.3.2 Mis juhtub, kui andmed on heteroskedastilised?

Sel juhul tõenäosusavaldis võib muutuda vigaseks, kuna MS ei hinda ühtegi "õiget" dispersiooni.

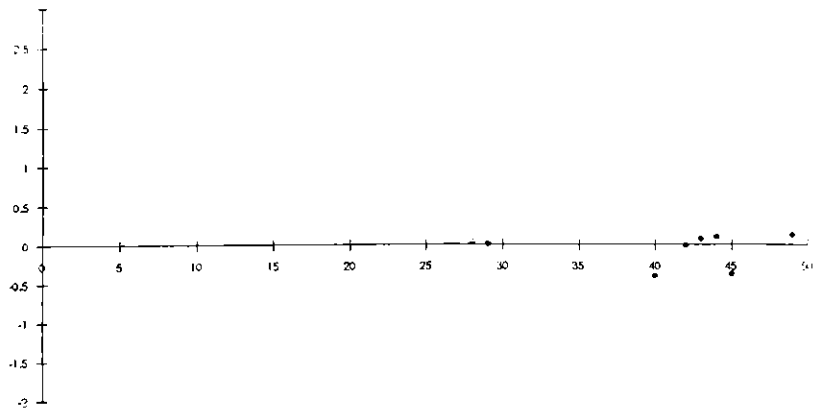
8.3.3 Parandused

Kasutame andmete teisendamist, et "stabiliseerida" dispersioone.

8.4 Sõltumatus

Teine oluline eeldus oli, et jäägid oleksid sõltumatud. See tähendab, et üks jääk, mis on suur, ei tohi mõjutada järgmist jääki. Üks näide, kus jäägid võivad olla sõltuvad, on pestitsiidide uurimuses. Pestitsiidide kõrge tase ühel maa-alal võib levida kõrvalolevale alale ning järelikult muutuvad naaberala tingimused. Sõltumatute vaatluste saamiseks peab korraldama katse nii, et töötlus ühes ühikus ei tohi mõjutada kõrvalolevaid ühikuid.

Kui eksisteerib loomulik järjekord katse ühikute hulgas, siis võib järgmisel viisil avastada sõltuvust jääkide hulgas. Kujutame jäägid graafikul loomulikus järjestuses. Jääkide pikk sama märgiga järgnevus võib osutada andmete sõltuvusele. On olemas formaalsed testid (Durbin-Watsons test; "Runs" test), mis kontrollivad järjestuses esinevat sõltuvust. Järgnev graafik on jääkidevahelise sõltuvuse näide. Sarnane Minitabi graafik on esitatud hiljem.



8.4.1 Mis juhtub, kui jäägid ei ole sõltumatud?

Sõltumatuse puudus võib tõsiselt mõjutada analüüsi. Peab püüdma selgitada, miks jäägid on sõltuvad.

8.5 Kehtiv mudel

Viimane eeldus oli, et analüüsiks kasutatav mudel peab olema kehtiv. Näiteks kahefaktorilises katses eeldame mudelit

y = μ + {A töötluse mõju} + {B töötluse mõju} + {AB koosmõju} + Viga.

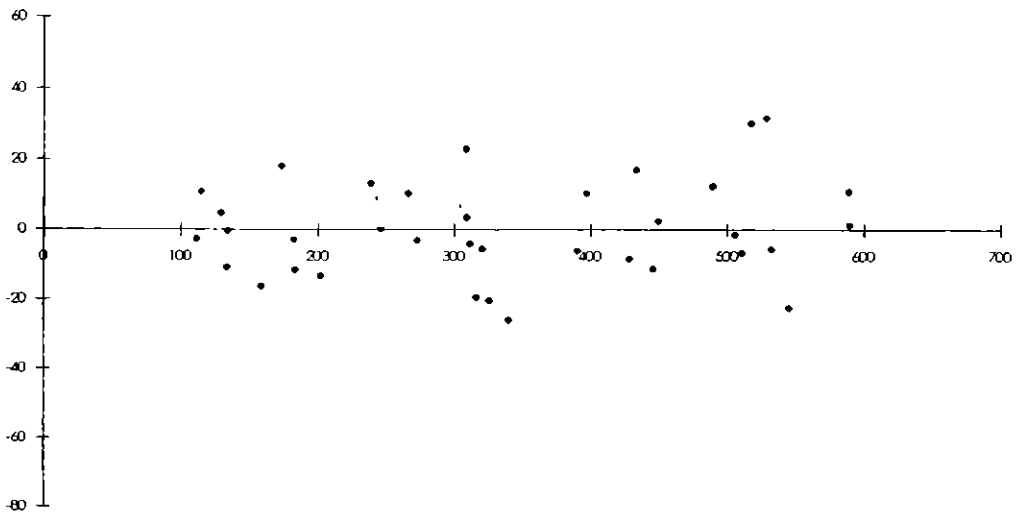
Kui mudel on vale, siis tuleb analüüs ka vigane.

Märgime, et mudel tähendab liidetavaid mõjusid. Näiteks faktori A mõju on liidetud keskväärtusele. Üks ilmne mudeli kõrvalekaldumise võimalus on juhul, kui mõjud ei ole liidetavad, aga on korrutatavad. ("Kasutatav väetis A tõstab saaki 10%".) Selllisel juhul võib analüüsi parandada log-teisendus.

8.5.1 Erindid

Erindid on vaatlused, mis mingil põhjusel hälbivad ülejäänud andmete hulgast. Mõnikord võivad erindid osutada andmete huvitavatele tunnustele. Järelikult on halb heita kõrvale kõiki "imelikke" vaatlusi. Erindit võib andmestikust välja jätta, kui see on põhjendatud või nad on vähemalt tugevasti kaheldavad ning on põhjustatud andmete kogumise veast, on valesti mõistetavad jne.

Näiteks on järgnev graafik, kus võib arvata, et ühte vaatlust võib vaadelda erindina.

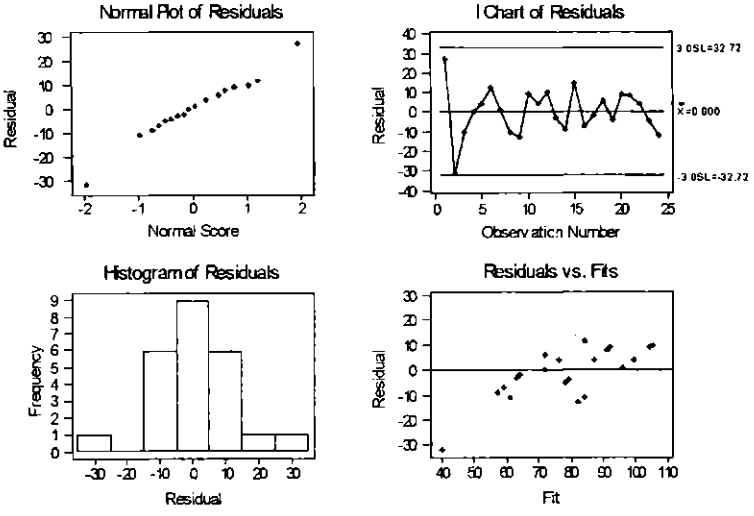


8.6 Jääkide graafikud paketiga MINITAB

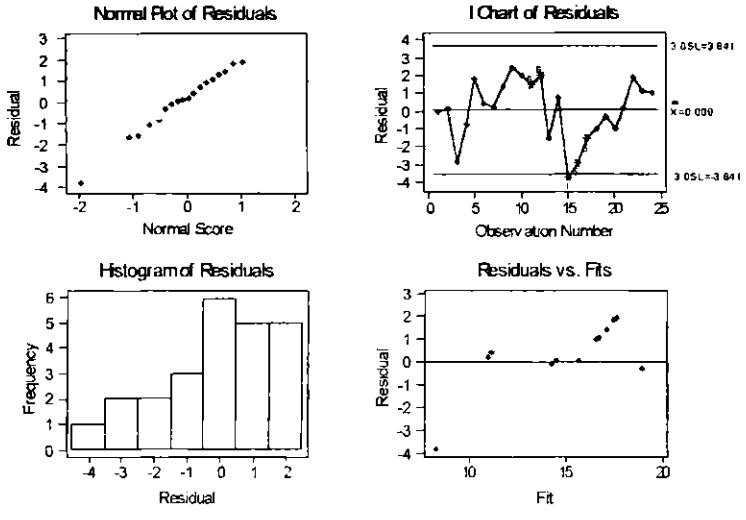
Programmipaketil Minitab on mõned suurepärased vahendid eespool olevate erinevat tüüpi jääkide graafikute tegemiseks. Kasutades seda, peab Minitabiga moodustama analüüsist uue veeru jääkide ja oodatava väärtuse ("fits") jaoks. See töötab Minitab mitmetes erinevates dispersioonanalüüsi, regressiooni või üldiste lineaarsete mudelite protseduurides. Ette tuleb anda veergude nimed, Minitab arvutab jäägid vastavalt RESI1 ja oodatava väärtuse ($\hat{y}$ väärtuse) kui FITS1. (Kui teha teisi erinevaid analüüse, võivad nimed olla RESI2, RESI3 jne).

Kui soovime, et Minitab valmistaks jääkide graafiku, siis väljastatakse järgnevat tüüpi graafikud.

Ühefaktoriline dispersioonanalüüsi näide



Kahefaktorilise dispersioonanalüüsi näide



AINEREGISTER

Normaaljaotusel põhinevad teststatistikud ja usaldusintervallid

Parameeter	Punkthinnang	Usaldusvahemik	Tast statistik
Üldine Θ	$\hat{\Theta}$	$\hat{\Theta} \pm \begin{Bmatrix} t \\ või \\ z \end{Bmatrix} \cdot \begin{Bmatrix} \sqrt{\hat{V}ar(\hat{\Theta})} \\ või \\ \sqrt{Var(\hat{\Theta})} \end{Bmatrix}$	$\begin{Bmatrix} t \\ või \\ z \end{Bmatrix} = \frac{\hat{\Theta} - \Theta_0}{\sqrt{Var(\hat{\Theta})}}$
Eeldused	z kui $Var(\hat{\Theta})$ on üldkogumis teada (praktikas ebatavaline) t kui $Var(\hat{\Theta})$ ei ole teada ja peame hindama valimi põhjal ($t \approx z$ kui vabadusastmete arv on suur)		
μ	$\overline{X}$	$\overline{X} \pm z\sqrt{\frac{\sigma^2}{n}}$	$z = \frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$
Eeldused	Eeldame, et teame σ^2 (sageli ebareaalne).		
μ	$\overline{X}$	$\overline{X} \pm t_{n-1}\sqrt{\frac{s^2}{n}}$	$t_{n-1} = \frac{\overline{X} - \mu_0}{\sqrt{\frac{s^2}{n}}}$
Eeldused	σ^2 ei ole teada, kasutame s^2 kui σ^2 hinnangut. Kui valim on väike, siis X jaotus peab olema ligilähedane normaaljaotusele.		

Parameeter	Punkthinnang	Usaldusvahemik	Tast statistik
p	$\hat{p}$	$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Eeldused	Valim on suur. Rusikareegel: mõlemad $n\hat{p}$ ja $n(1-\hat{p})$ on suuremad 10-st.		
$\mu_d = \mu_x - \mu_y$	$\bar{d}$	$\bar{d} \pm t_{n-1}\sqrt{\frac{s_d^2}{n}}$	$t_{n-1} = \frac{\bar{d} - 0}{\sqrt{\frac{s_d^2}{n}}}$
Eeldused	Ühildatud valimid, kus n= paaride arvuga. Kui valimid on väiksed, siis d jaotus peab olema ligilähedane normaaljaotusele.		
$\mu_x - \mu_y$	$\bar{X} - \bar{Y}$	$\bar{X} - \bar{Y} \pm z\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$	$z = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$
Eeldused	Eeldame, et kaks sõltumatut valimit on suured (mõlemad > 30).		
$\mu_x - \mu_y$	$\bar{X} - \bar{Y}$	$\bar{X} - \bar{Y} \pm t_{n_X+n-2_Y}\sqrt{s^2(\frac{1}{n_X} + \frac{1}{n_Y})}$, kus $s^2 = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X + n_Y - 2}$	$t_{n_X+n-2_Y} = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{s^2(\frac{1}{n_X} + \frac{1}{n_Y})}}$, kus $s^2 = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X + n_Y - 2}$
Eeldused	Kaks sõltumatut valimit. Eeldame, et $\sigma_X^2 = \sigma_Y^2$ X ja Y jaotus peab olema ligilähedane normaaljaotusele, eriti kui valimid on väikesed.		

Parameeter	Punkthinnang	Usaldusvahemik	Tast statistik
$\mu_x - \mu_y$	$\overline{X} - \overline{Y}$	$\overline{X} - \overline{Y} \pm t_\gamma \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$	$t_\gamma = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$
Eeldused	Kaks sõltumatut valimit. Ei vaja eeldust, et $\sigma_X^2 = \sigma_Y^2$. X ja Y jaotus peab olema ligilähedane normaaljaotusele, eriti kui valimid on väikesed. γ on vabadusastmete arvu ligilähedane arv: $\gamma = \frac{\left[s_X^2/n_X + s_Y^2/n_Y \right]^2}{\frac{\left[s_X^2/n_X \right]^2}{n_X - 1} + \frac{\left[s_Y^2/n_Y \right]^2}{n_Y - 1}}$		
$p_X - p_Y$	$\hat{p}_X - \hat{p}_Y$	$\hat{p}_X - \hat{p}_Y \pm z \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$	$z = \frac{(\hat{p}_X - \hat{p}_Y) - 0}{\sqrt{\hat{p}_0(1-\hat{p}_0)(\frac{1}{n_X} + \frac{1}{n_Y})}}$
Eeldused	Kaks sõltumatut valimit on suured. Rusikareegel: mõlemad $n\hat{p}$ ja $n(1-\hat{p})$ on suuremad 10-st.		

Kõikide testide ja intervallide korral: valim(id) peavad olema moodustatud nii, et vaatlused on sõltumatud. Kahe valimi juhu korral vajame samuti eeldust, et kaks keskväärtust või proportsiooni oleksid sõltumatud. Välja arvatud ühildatud valimite korral.

Dispersioonanalüüsi jaoks käsitsi arvutamise valemid

Tähistused:

N = vaatluste üldarv
a = faktori A tasemete (nivoode) arv
b = faktori B tasemete arv (või plokkide arv plokkplaanis).
T = vaatluste summa; kui on antud indeksiga, siis summa on arvutatud üle kõigi vaatluste vastava indeksi järgi.
Punkt: indeksi kohal punkt tähendab vastava indeksi summat (või keskmist) üle kõigi väärtuste.

Ühefaktoriline dispersioonanalüüs

n_i = i-na töötluse vaatluste arv.

N = $\sum n_i$

Faktori korrektsioon $K = \frac{\left(\sum y_{ij}\right)^2}{N} = \frac{T_{..}^2}{N}$

$$SS_T = \sum (y_{ij} - \bar{y}_{..})^2 = \sum y_{ij}^2 - K$$

$$SS_A = \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum \frac{T_{i.}^2}{n_i} - K$$

$$SS_e = SS_T - SS_A$$

Tasakaalus plokkidega ühefaktoriline dispersioonanalüüs

a on faktori A tasemed, b plokid, igas plokis igale töötlusele tehti üks replikatsioon.

$N = a \cdot b$

Faktori korrektsioon $K = \frac{\left(\sum y_{ij}\right)^2}{N} = \frac{T^2}{N}$

$$SS_T = \sum y_{ij}^2 - K$$

$$SS_A = \frac{1}{b} \sum T_{i.}^2 - K$$

$$SS_B = \frac{1}{a} \sum T_{.j}^2 - K$$

$$SS_e = SS_T - SS_A - SS_B$$

Tasakaalus ladina ruutude plaan

a on faktori A tasemed, a read, a veerud. $N = a \cdot a$

Faktori korrektsioon $K = \frac{\left(\sum y_{ij}\right)^2}{N} = \frac{T_{..}^2}{N}$

$$SS_T = \sum y_{ij}^2 - K$$

$$SS_A = \frac{1}{a}\sum T_{i.}^2 - K$$

$$SS_{Rida} = \frac{1}{a}\sum T_{j}^2 - K$$

$$SS_{Veerg} = \frac{1}{a}\sum T_{.k}^2 - K$$

$$SS_e = SS_T - SS_A - SS_{Rida} - SS_{Veerg}$$

Tasakaalus kahefaktoriline dispersioonanalüüs

a on faktori A tasemed, b faktori Btasemed, n on replikatsioonid igale kombinatsioonile A-st ja B-st.

$N = a \cdot b \ \ n$

Faktori korrektsioon $K = \frac{\left(\sum y_{ijk}\right)^2}{N}$

$$SS_T = \sum y_{ijk}^2 - K$$

$$SS_{Töötl} = \frac{1}{n}\sum T_{ij.}^2 - K$$

$$SS_A = \frac{1}{nb}\sum T_{i..}^2 - K$$

$$SS_B = \frac{1}{na}\sum T_{.j.}^2 - K$$

$$SS_{AB} = SS_{Töötl} - SS_A - SS_B$$

$$SS_e = SS_T - SS_{Töötl}$$

Liigendatud plokkplaan

(A = peaploki faktor, B = plokk, C = liigendatud ploki faktor)
a faktori A tasemed, b plokid, c faktori C tasemed

$N = a \cdot b \ \ n$

Faktori korrektsioon $K = \dfrac{\left(\sum y_{ijk}\right)^2}{N}$

$$SS_T = \sum y_{ijk}^2 - K$$

$$SS_{Pea} = \frac{1}{c}\sum T_{ij.}^2 - K$$

$$SS_A = \frac{1}{bc}\sum T_{i..}^2 - K$$

$$SS_B = \frac{1}{ac}\sum T_{.j.}^2 - K$$

$$SS_{AB} = SS_{Pea} - SS_A - SS_B$$

$$SS_C = \frac{1}{bc}\sum T_{..k}^2 - K$$

$$SS_{AC} = \frac{1}{b}\sum T_{i.k}^2 - SS_A - SS_C$$

$$SS_e = SS_T - SS_A - SS_B - SS_{AB} - SS_C - SS_{AC}$$