Statistika ja modelleerimise baaskoolitus õppejõududele ja juhendajatele

Loeng 4

Eesti Maaülikool
27.-30. august 2012
Tanel Kaart

Kahe grupi võrdlus

Kahe grupi keskmiste võrdlus – kolm *t*-testi

Kahe populatsiooni keskmiste võrdlus – *t*-test

Eeldus: $X \sim N(\mu;\sigma^2)$ [või suur (>60) n]

Sõltuvad vaatlused (paariviisiline võrdlus)

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$
$\Leftrightarrow$
$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

Teststatistik: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_{1-2}}\sqrt{n} \underset{H_0}{\sim} t_{n-1}$

Sõltumatud vaatlused

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

Kahe populatsiooni dispersioonide võrdlus – *F*-test

$H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

Teststatistik: $F = \dfrac{s_1^2}{s_2^2} \underset{H_0}{\sim} F_{n_1-1,n_2-1}$

Võrdsed dispersioonid

Teststatistik:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}} \underset{H_0}{\sim} t_{n_1+n_2-2},$$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Mittevõrdsed dispersioonid

Teststatistik: $t = \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \underset{H_0}{\sim} t_\nu,$

$$\nu = \left\{(s_1^2/n_1 + s_2^2/n_2)^2 \Big/ \left[\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}\right]\right\} - 2$$

Sõltuvad vaatlused (paariviisiline võrdlus)

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$
$\Leftrightarrow$
$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

Teststatistik: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_{1-2}}\sqrt{n} \underset{H_0}{\sim} t_{n-1}$

Näide. Soovitakse uurida, kas lehmade ööpäevane piimatoodang langes pärast seda, kui neile lõpetati juurvilja söötmine (olulisuse nivool $\alpha = 0{,}05$).

Lehm	Piim (kg/ööpäevas)		Piimatoodangu
	juurviljaga (x_1)	juurviljata (x_2)	muutused (d)
1	10	11	-1
2	8	7	1
3	11	10	1
4	10	10	0
5	7	6	1
6	8	5	3
7	10	6	4
8	9	4	5
9	8	6	2
10	10	8	2

Kontrollime hüpoteesi piimatoodangu languse kohta, s.t.

$H_0: \mu_d \leq 0$
$H_1: \mu_d > 0$

$n = 10;\ \bar{d} = 1{,}8;\ s_d = 1{,}81$

Andmetest arvutatud teststatistik: $t = \bar{d} \Big/ s_d \sqrt{n} \approx 3{,}14$

Teststatistiku kriitiline väärtus (ühepoolne hüpotees):

$t_{1-\alpha,\,n-1} = t_{0{,}95;9} = 1{,}83$

Järeldus: $t = 3{,}14 > 1{,}83 = t_{1-\alpha,\,n-1}$

$\Rightarrow$ $H_1: \mu_d > 0$ $(p = 0{,}006)$

Sõltumatud vaatlused

Näide. Ettevõttes võrreldi ametiühingusse kuuluvate ja sinna mittekuuluvate töötajate puudumisi aasta jooksul. Viiskümmend vaadeldud ametiühinguliiget puudusid keskmiselt 9,3 päeva, kusjuures standardhälve oli 3,1 päeva. Ametiühingusse mittekuulujad, keda oli 45, puudusid igaüks keskmiselt 8,7 päeva standardhälbega 2,3 päeva. Kontrollida hüpoteesi ettevõtte töötajate keskmiselt puudutud päevade arvu sõltuvusest ametiühingusse kuulumisest olulisuse nivool $\alpha = 0{,}05$.

$n_1 = 50;\ n_2 = 45$
$\bar{x}_1 = 9{,}1;\ \bar{x}_2 = 8{,}7$
$s_1 = 3{,}1;\ s_2 = 2{,}3$
$\alpha = 0{,}05$

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

(1) $H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$

Teststatistik: $F = s_1^2 / s_2^2 = 1{,}817$

Teststatistiku kriitiline väärtus:
$F_{1-\alpha/2;n_1-1;n_2-1} = F_{0{,}975;\,49;\,44} = 1{,}799$

Järeldus: $F = 1{,}817 > 1{,}799 = F_{0{,}975;\,49;\,44} = F_{1-\alpha/2;n_1-1;n_2-1}$

$\Rightarrow H_1 : \sigma_1^2 \neq \sigma_2^2 \quad (p = 0{,}023)$

(2) Teststatistik: $t = |\bar{x}_1 - \bar{x}_2| \Big/ \sqrt{s_1^2/n_1 + s_2^2/n_2} = 0{,}969$

Teststatistiku kriitiline väärtus: $t_{1-\alpha/2,\,\nu} = t_{0{,}975;95} = 1{,}985$

Järeldus: $t = 0{,}969 < 1{,}985 = t_{0{,}975;45} \Rightarrow H_0 : \mu_1 = \mu_2 \quad (p = 0{,}335)$

Märkus: eeldanuks me siiski, et $\sigma_1^2 = \sigma_2^2$, jõudnuks me peale ühise dispersiooni $s^2 = 7{,}566$ ja teststatistiku $t = 0{,}708$ arvutamist samale järeldusele: $\mu_1 = \mu_2$, aga olulisustõenäosus olnuks pisut suurem ($p = 0{,}481$).

Kolm *t*-testi – mis seal vahet on?

Üldine eeldus: $X \sim N(\mu;\sigma^2)$

Andmete olemus		Lisaeeldused	*p*: $\mu_1 > \mu_2$	*p*: $\mu_1 \neq \mu_2$
Sõltumatud vaatlused	mittevõrdne varieeruvus	–	0,026	0,053
	võrdne varieeruvus	$\sigma_1^2 = \sigma_2^2$	0,024	0,048
Sõltuvad vaatlused		$H_0 : \mu_1 = \mu_2$ $\Updownarrow$ $H_0 : \mu_1 - \mu_2 = 0$	0,006	0,012

Piim (kg/ööpäevas)	
juurviljaga (x_1)	juurviljata (x_2)
10	11
8	7
11	10
10	10
7	6
8	5
10	6
9	4
8	6
10	8

Mida enam on lihtsustavaid eelduseid e mida kitsamalt on situatsioon (andmed) enne juhuslikkuse mängu toomist piiritletud,
seda suurem on statistilise testi võimsus
(seda väiksem erinevus on vajalik populatsioonide erinevuse tõestamiseks e seda väiksem on uurija eksimistõenäosus kummutades nullhüpoteesi)!

Mitteparameetrilised testid

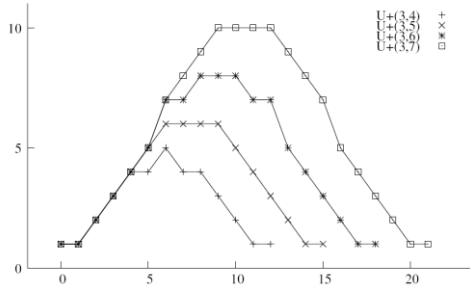
Enam levinud testid sõltumatute vaatluste korral

Mann-Whitney U-test, **Wilcoxoni test**

Eeldused: uuritavad tunnused on vähemalt järjestatavad;
uuritavad tunnused omandavad küllalt palju erinevaid väärtusi.

Idee: kui võrreldavate valimite keskväärtused (jaotused) on võrdsed, peaks nendest moodustatud ühine variatsioonrida olema nö hästi segunenud, st et mõlema valimi elemendid paiknevad enam-vähem vaheldumisi ega ole koondunud variatsioonrea algusesse või lõppu.

Teststatistiku teoreetiline jaotus H_0 korral.



Mitteparameetrilised testid

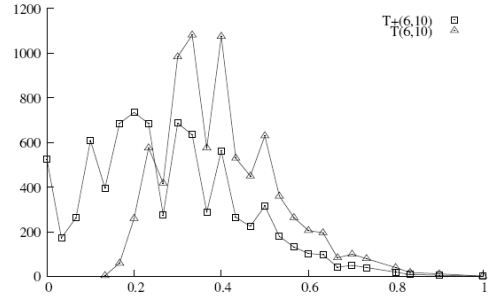
Enam levinud testid sõltumatute vaatluste korral

Kolmogorov-Smirnovi test

Eeldused: uuritavad tunnused on vähemalt järjestustunnused.

Idee: test võrdleb kahe valimi jaotust (mitte üksnes keskmist taset!), leides selleks võrreldavate valimite empiiriliste jaotusfunktsioonide maksimaalse erinevuse – kui see erinevus on piisavalt suur, siis on jaotused järelikult erinevad.

Teststatistiku teoreetiline jaotus H_0 korral.



Mitteparameetrilised testid

Enam levinud testid sõltuvate vaatluste korral

Märgitest

Eeldused: uuritavad tunnused on vähemalt järjestustunnused.

Idee : võrdse keskväärtuse korral peaks paariviisiliste vaatluste vahede hulgas positiivseid ja negatiivseid (tähistatuna vastavalt "+" ja "–", siit ka testi nimi) olema enamvähem võrdselt.

Wilcoxoni astakmärgitest

Eeldused: uuritavad tunnused on vähemalt järjestustunnused.

Idee : võrdse keskväärtuse korral peaks vaatluste vahede hulgas positiivseid ja negatiivseid olema enamvähem võrdselt ning, täiendusena märgitestile, peaksid mõlemad muutuma samades piirides.

Parameetrilised *versus* mitteparameetrilised testid

		p	
	Test	$\mu_1 > \mu_2$	$\mu_1 \neq \mu_2$
Sõltumatud vaatlused	*t*-test, mittevõrdne varieeruvus	0,026	0,053
	t-test, võrdne varieeruvus	0,024	0,048
	Mann-Whitney U-test, Wilcoxoni test	0,038	0,075
	Kolmogorov-Smirnovi test	0,055	0,164
Sõltuvad vaatlused	*t*-test	0,006	0,012
	Märgitest	0,020	0,039

Piim (kg/ööpäevas)	
juurviljaga (x_1)	juurviljata (x_2)
10	11
8	7
11	10
10	10
7	6
8	5
10	6
9	4
8	6
10	8

Permutatsioonitestid [*permutation tests*]

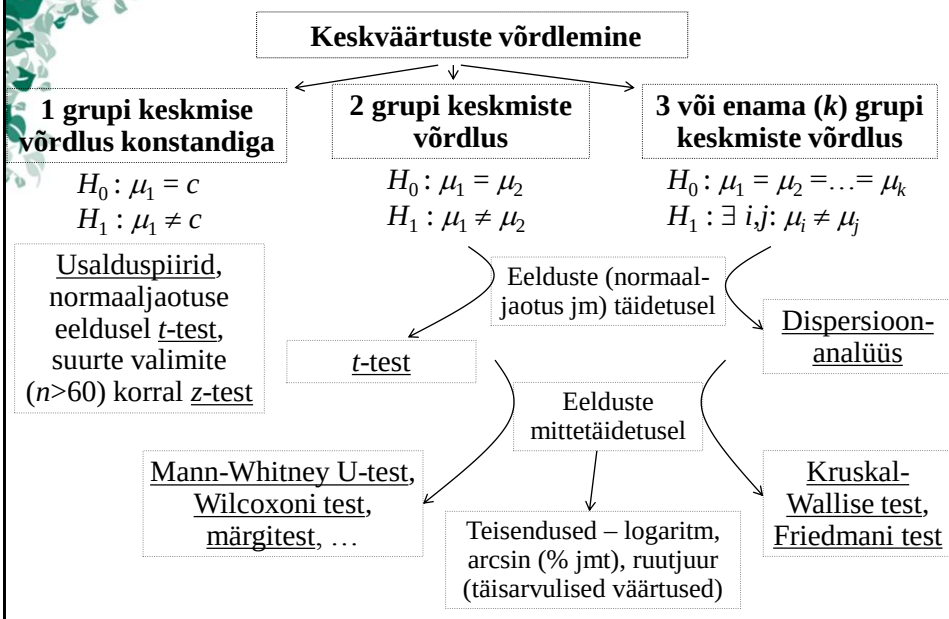
Permutatsioonitest (e **täpne test** [*exact test*] või randomiseerimistest [*randomization test*]) kujutab enesest teststatistiku nullhüpoteesile vastava jaotuse leidmist arvutades teststatistiku väärtused andmete kõikvõimalike ümberpaigutuste korral.

Näiteks kahe grupi keskmiste võrdlemisel, kus gruppide suurused on n_1 ja n_2, arvutatakse esmalt välja andmetele vastav teststatistiku väärtus,

- seejärel moodustatakse ühine andmestik suurusega n_1+n_2,
- millest moodustatakse kõikvõimalikud grupid suurustega n_1 ja n_2 ning arvutatakse kõigil juhtudel teststatistiku väärtus;
- tulemuseks saadud teststatistiku jaotuse alusel leitakse, kui suure sagedusega (tõenäosusega) tulid teststatistiku väärtused võrdsed või suuremad originaalandmeist leitud väärtusest – saadud tõenäosus on täpne 2-poolsele hüpoteesile vastava olulisuse tõenäosuse väärtus.

Juhul, kui kõikvõimalike permutatsioonide teostamine on liiga töömahukas, valitakse neist juhuslikult üksnes teatud hulk ja arvutatakse asümptootiliselt täpne *p*-väärtus – selliseid teste tuntakse **Monte Carlo testidena**.

Keskväärtuste võrdlemine

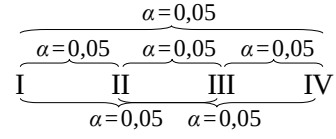


Dispersioonanalüüs

Eesti Maaülikool
Estonian University of Life Sciences

Mitmene võrdlus

Võrdleme näiteks 4 gruppi, lubades iga üksikvõrdluse puhul eksimist 5% tõenäosusega.



Tõenäosus, et üksikvõrdlusel viga ei tehta, on $1-\alpha=0,95$.

Tõenäosus, et kuuel üksikvõrdlusel kokku ei eksita, on $(1-\alpha)^6=0,95^6\approx0,735$.

Mistõttu tõenäosus teha üks (või mitu) vale otsus(t) 4 grupi paarikaupa võrdlemisel on $1-0,735=0,265$ (eksimise tõenäosus on üle 25%!).

Otsite näiteks põhjust, mis võiks soodustada taimedel haiguse tekkimist. Viite läbi uuringu ja fikseerite haigetel ja tervetel taimedel 100 potentsiaalselt haigestumist mõjutava tunnuse väärtused (a'la mulla toitaineterikkus, eelmisel kuul sadanud vihma kogus, taimi kasvata taluniku pikkus jne).

Iga potentsiaalse haigusega seotud tunnuse osas võrdlete haigeid ja terveid taimi kasutades olulisuse nivood 0,05.

Kui nüüd eeldada, et tegelikult ei mõjuta ükski valitud 100-st tunnusest haigestumist, siis sellest hoolimata võiksite antud uuringu puhul lugeda tõestatuks umbes 5 haiguse tekkimist soodustavat tegurit.

Mitmene võrdlus

Bonferroni meetod: piiramaks k üksikvõrdluse puhul ühe või enama vea tegemise tõenäosust olulisuse nivooga α, tuleb kõigil üksikvõrdlustel võtta olulisuse nivooks α/k.

Näiteks 4 grupi võrdlemisel, garanteerimaks kuue võrdluse peale kokku eksimist mitte üle 5%-lise tõenäosusega, tuleb üksikvõrdlustel võtta olulisuse nivooks $\alpha^* = \alpha/k = 0{,}05/6 \approx 0{,}0083$.

Bonferroni-Holmi meetod: teostatakse kõik testid ja järjestatakse saadud olulisuse tõenäosused, $p_1 \le p_2 \le \ldots \le p_k$; otsused nullhüpoteesi kasuks või kahjuks tehakse kasutades olulisuse nivoosid α/k, $\alpha/(k-1)$, $\alpha/(k-2)$, …, $\alpha/2$, α.

Kui teostatavate testide arv kasvab, väheneb kasutatav olulisuse nivoo kiiresti ja alternatiivse hüpoteesi tõestamine osutub sageli äärmiselt raskeks (nõuab tohutu hulga vaatluste olemasolu).

Seetõttu ei ole statistilised meetodid mitte eriti sobivad katse/eksituse-meetodil teaduse tegemiseks (proovime, kas midagi õnnestub)!

False Discovery Rate ("valeavastuste määr")

Idee: piirata (kontrollida) ekslikult vastuvõetud alternatiivsete hüpoteeside osakaalu kõigi vastuvõetud alternatiivsete hüpoteeside seas (piiriks näiteks 5%).

Üks lihtne moodus antud lähenemist ise kasutada on järgmine:

1. Järjesta testide poolt raporteeritavad olulisuse tõenäosused kasvavalt, $p_1 \le p_2 \le \ldots \le p_k$.
2. Kirjuta välja kriitilised suurused α/k, $2*\alpha/k$, $3*\alpha/k$, …, α, kus α näitab maksimaalset nn *False Discovery Rate*'i.
3. Võrdle iga olulisuse tõenäosust temale vastava (sama jrk numbriga) kriitilise väärtusega, kuni jõuad paarini, kus olulisuse tõenäosus on suurem kriitilisest väärtusest. Loe kõigi eelnenud hüpoteeside jaoks alternatiivne hüpotees tõestatuks ja temale järgnevate hüpoteeside puhul jää nullhüpoteesi juurde.

Mõnikord nimetet ka kui **Benjamini-Hochbergi korrektsioon**.

Keskmiste mitmene võrdlus

On *k* gruppi, mille keskmist taset tahame võrrelda.

Sellisel juhul on sageli otstarbekas $k(k–1)/2$ paariviisilist võrdlust (*t*-testi) asendada üheainsa hüpoteeside paari kontrollimisega.

Viimase võib sõnastada kujul:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1: leiduvad sellised grupid *i*,*j*, et $\mu_i \neq \mu_j$

Eeldustel, et

¤ vaatlused on sõltumatud,

¤ uuritav (sõltuv) tunnus on normaaljaotusega ja

¤ uuritava tunnuse varieeruvus võrreldavais gruppides on ühesugune,

on taolise hüpoteeside paari korral rakendatavaks analüüsimeetodiks **dispersioonanalüüs**.

Dispersioonanalüüs

Dispersioonanalüüsil jagatakse tunnused vastavalt nende rollile kaheks:

¤ tunnus, mille keskmisi võrrelda soovitakse, on **uuritav tunnus** e **funktsioontunnus** (lehma piimatoodang, forelli kasvukiirus, talle mass, sea pekipaksus, jne);

¤ (diskreetne või mittearvuline) tunnus, mille väärtuste alusel võrreldavad grupid moodustatakse, on **faktortunnus** (tõug, lüpsiseade, laudatüüp jne).

Dispersioonanalüüsi tulemuste tõlgendamisel räägitaksegi enamasti faktor-tunnuse **mõjust** uuritavale tunnusele.

Näiteks, tõu või lüpsiseadme või laudatüübi vm mõju piimatoodangule, kasvanduse mõju forellide kasvukiirusele, omaniku mõju talle massile, genotüübi (teatud geenikombinatsioonide) mõju sigade pekipaksusele jne.

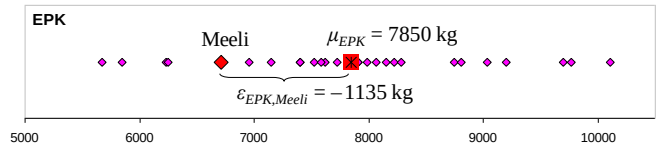
Dispersioonanalüüsi mudel

¤ Iga rühma i (kus $i=1,\ldots,k$) iseloomustab keskmine uuritava tunnuse väärtus μ_i, mistõttu mõõtmistulemused saab esitada mudeliga

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

kus y_{ij} on uuritava tunnuse väärtus i. rühma kuuluval j. objektil ja ε_{ij} on juhuslik mõju (objekti omapära).

Näiteks EPK-tõugu lehm Meeli 1. laktatsiooni piimatoodang 6715 kg on väljendatav kui uuritud EPK-tõugu lehmade 1. laktatsiooni keskmise toodangu $\mu_{EPK}=7850$ kg ja Meeli tõusisese erinevuse $\varepsilon_{EPK,Meeli}=-1135$ kg summa.



Dispersioonanalüüsi mudel

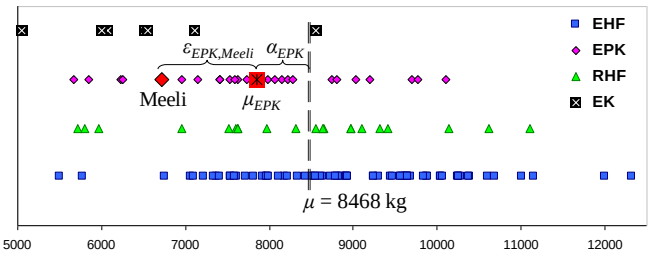
¤ Faktortunnuse mõju uurimiseks esitatakse mudel kujul

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

kus μ tähistab üldkeskmist ja α_i on faktori i. taseme poolt põhjustatud kõrvalekalle üldkeskmisest (i. taseme mõju), $\mu_i = \mu + \alpha_i$.

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ ⇔ H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
H_1: leiduvad grupid i,j, et $\mu_i \neq \mu_j$ H_1: leidub grupp i, et $\alpha_i \neq 0$

Näiteks EPK-tõugu lehm Meeli 1. laktatsiooni piimatoodang 6715 kg on väljendatav kui kõigi uuritud lehmade keskmise 1. laktatsiooni piimatoodangu $\mu=8468$ kg, EPK-tõu mõju (EPK-tõugu lehmade 1. laktatsiooni keskmise toodangu erinevus üld-keskmisest) $\alpha_{EPK}=-618$ kg ja Meeli tõusisese erinevuse $\varepsilon_{EPK,Meeli}=-1135$ kg summa.

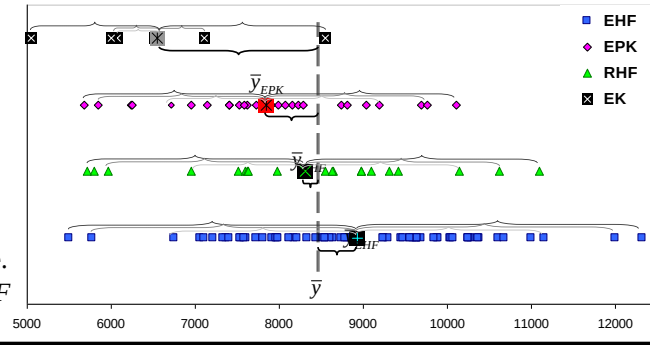


Dispersioonanalüüsi tööpõhimõte

Dispersioonanalüüsi tööpõhimõte seisneb uuritava tunnuse rühmadesisese (nn juhusliku) varieeruvuse $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$
ja rühmadevahelise (faktori mõjust tingitud) varieeruvuse $SSA = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$
võrdlemises – kui rühmadevaheline erinevus on suurem kui rühmadesisene erinevus, on tegu ilmse tõendiga faktortunnuse mõju olemasolu kohta.

Siit ka analüüsi nimetus – dispersioonanalüüs [*analysis of variance*, ANOVA].

Näide.
$i = EK, EPK, RHF, EHF$



Dispersioonanalüüsi tabel

Dispersioonanalüüsiga seotud arvutused koondatakse tavaliselt alljärgnevasse nn. dispersioonanalüüsi tabelisse.

Varieeruvuse allikas	Hälvete ruutude summa	Vabadusastmeid	Keskruut	*F*-suhe	Olulisustõenäosus
Faktor	$SSA = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	p
Viga	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n-k$	$MSE = \frac{SSE}{n-k}$		
Kokku	$SS = SSA + SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n-1$			

Juhul, kui faktortunnuse mõjule vastav keskmine **gruppide vaheline varieeruvus** *MSA* on suurem, kui uuritavate objektide omapärale vastav keskmine **gruppide sisene varieeruvus** *MSE*, on *F*-statistiku väärtus ühest suurem.

Piisavalt suure *F*-suhte väärtuse korral võib lugeda tõestatuks sisuka hüpoteesi – leiduvad vähemalt kaks teineteisest eristuvat gruppi.

Dispersioonanalüüs

Näide. Uuritakse ühes katsefarmis peetava 121 lehma 1. laktatsiooni piimatoodangu sõltuvust tõust (EHF, RHF, EPK, EK). Kontrollitav hüpoteeside paar on kujul:

$H_0: \mu_{EHF} = \mu_{RHF} = \mu_{EPK} = \mu_{EK}$ ⇔ H_0: tõul ei ole mõju
H_1 : leiduvad tõugrupid i,j, et $\mu_i \neq \mu_j$ H_1 : tõul on mõju

Dispersioonanalüüsi mudel on kujul $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,
kus μ on kõigi farmi lehmade keskmine 1. laktatsiooni piimatoodang, α_i on i. tõu keskmine erinevus sellest (i. tõu mõju, i = *EHF*, *RHF*, *EPK*, *EK*) ning y_{ij} ja ε_{ij} on vastavalt i. tõugu j. lehma mõõdetud piimatoodang ja selle erinevus tõu keskmisest (lehma "omapära", $j = 1, \ldots, n_i$, n_i on lehmade arv i. tõus).

Tõug	n_i	$\bar{y}_i = \hat{\mu}_i$	$s_i^2 = \hat{\sigma}_i^2$
EHF	68	8929,9	1776652
RHF	20	8311,5	2265768
EPK	27	7850,1	1335053
EK	6	6549,3	1419571
Kokku	121	8468,1	2093541

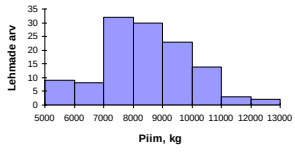
Tõugude mõjud on kõrvaloleva keskmiste toodangute tabeli alusel leitavad kujul

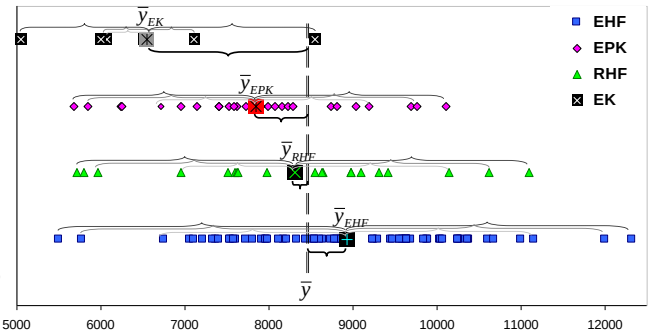
α_{EHF} = 460,8 kg; α_{RHF} = −156,6 kg;
α_{EPK} = −618,0 kg ja α_{EK} = −1918,8 kg.

Nende mõjude erinevuse kontrollimiseks tuleb läbi viia dispersioonanalüüs [viimase eeldused dispersioonide võrdsuse ja normaaljaotuse (vt ka järgmine lk) osas on enamvähem täidetud].

Dispersioonanalüüs

Näide. Uuritakse ühes katsefarmis peetava 121 lehma 1. laktatsiooni piimatoodangu sõltuvust tõust.





Varieeruvuse allikas	Hälvete ruutude summa	Vabadusastmete arv	Keskruut	F-suhe	p
Tõug	47330434	3	15776811	9,053	**0,000019**
Viga	203894493	117	1742688		
Kokku	251224927	120			

< 0,05 => H_1: tõul on mõju

Dispersioonanalüüs

Juhul, kui võrreldavaid gruppe on vaid kaks, on dispersioonanalüüsi tulemused identsed võrdsete dispersioonide eeldusel läbi viidud *t*-testiga.

Näide. Võrreldakse kahest erinevast tõust sigade ööpäevast juurdekasvu.

Andmed:

	Ööpäevane juurdekasv (g)			
Tõug 1	520	550	560	530
Tõug 2	630	690	700	680

MS Exceli protseduuride *t-Test: ...Equal Variances* ja *Anova: Single Factor* väljatrükid.

t-Test: Two-Sample Assuming Equal Variances

	Tõug 1	*Tõug 2*
Mean	540	675
Variance	333,33	966,67
Observations	4	4
Pooled Variance	650	
Hypothesized Mean Difference		
df	6	
t Stat	-7,4885	
P(T<=t) one-tail	0,00015	
t Critical one-tail	1,9432	
P(T<=t) two-tail	0,00029	
t Critical two-tail	2,4469	

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Tõug 1	4	2160	540	333,33
Tõug 2	4	2700	675	966,67

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	36450	1	36450	56,0769	0,00029	5,9874
Within Groups	3900	6	650			
Total	40350	7				

$p < 0{,}05$
=>
erinevat tõugu sead kasvavad erineva kiirusega

Mitmefaktoriline dispersioonanalüüs

Kui vaatlusobjekte saab rühmitada mitme tunnuse (faktortunnuse) järgi, võib osutuda mõttekaks analüüsidagi korraga mitme faktortunnuse mõju (näiteks igal lehmal võib olla fikseeritud tema tõug ja farm, igal kalal tema sugu ja püügikoht).

Dispersioonanalüüsi mudel, mis hõlmab kahe faktortunnuse mõjusid, on kujul:

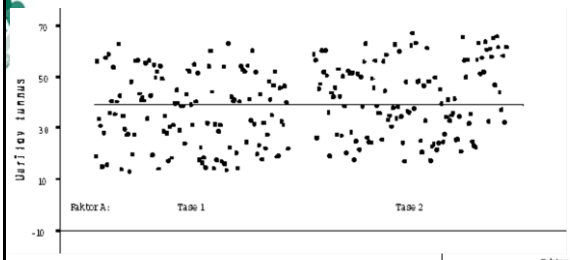
$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

kus μ tähistab üldkeskmist,
α_i-d ja β_j-d märgivad uuritava tunnuse keskmise muutust vastavalt esimese ja teise faktori väärtuste muutumisele (α_i on esimese faktori i. taseme mõju ja β_j on teise faktori j. taseme mõju),
y_{ijk} ning ε_{ijk} on vastavalt esimese faktori i. tasemel ja teise faktori j. tasemel sooritatud k. mõõtmise väärtus ning selle erinevus sama väärtuste kombinatsiooni keskmisest (vaatluse omapära, mudeli viga).

Mitmefaktoriline dispersioonanalüüs

Miks seda vaja on?

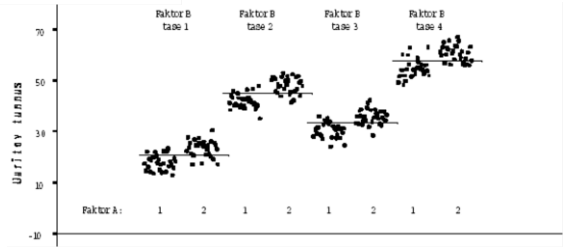
1) Hinnangute ja otsustuste täpsus võib paraneda.



Illustratsiooniks kaks samu andmeid illustreerivat hajuvusdiagrammi.

(joonised M. Mõlsi konspektist)

Osutub, et vaadeldes uuritava tunnuse väärtusi homogeensete gruppide kaupa (faktori B järgi), võib huvipakkuva faktori (A) mõju selgemalt esile tõusta.

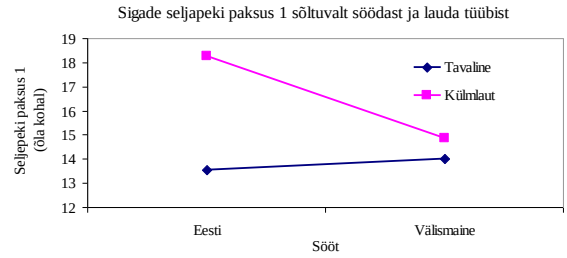


Mitmefaktoriline dispersioonanalüüs

Miks seda vaja on?

2) Võimalik selgemalt väljendada uuritava tunnuse ja faktorite vahelisi seoseid.

3) Interaktsioonid e koosmõjud – uuritava tunnuse väärtused muutuvad ühe faktori tasemete vahel erinevalt, sõltuvalt teise faktori väärtustest.



4) Ilma ei pruugi mudel olla korrektne (jääkliige ei puugi olla normaal-jaotusega).

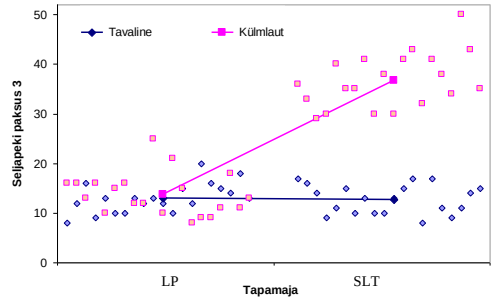
Kahefaktoriline faktoritevahelist interaktsiooni arvestav mudel esitatakse kujul

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

kus γ_{ij} märgib esimese faktori i. taseme ja teise faktori j. taseme koosmõju.

Mitmefaktoriline dispersioonanalüüs

Näide. 80-st seast 40 peeti tavalistes ja 40 välitingimustes. Mõlemast grupist pooled tapeti kohalikus tapamajas (LP) ja pooltel eelnes "paarematele jahimaadele siirdamisele" stressirohke üle 200 kilomeetrine transport auto ja praami abil (SLT).



The SAS System

The GLM Procedure

Dependent Variable: BackFat3 BackFat3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8295.450000	2765.150000	166.77	<.0001
Error	76	1260.100000	16.580263		
Corrected Total	79	9555.550000			

R-Square	Coeff Var	Root MSE	BackFat3 Mean
0.868129	21.34672	4.071887	19.07500

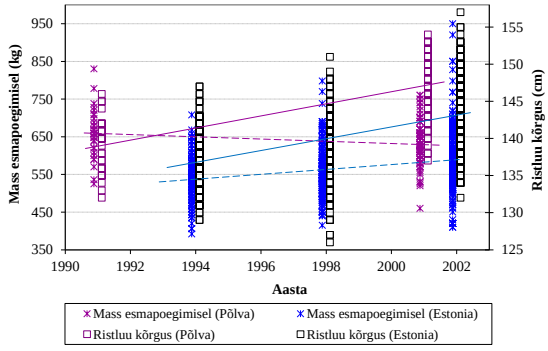
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Kasvukoht	1	3050.450000	3050.450000	183.98	<.0001
Tapamaja	1	2553.800000	2553.800000	154.03	<.0001
Kasvukoht*Tapamaja	1	2691.200000	2691.200000	162.31	<.0001

H_0 : mudel ei ole parem võrreldes konstantse mudeliga
H_1 : mudel on parem võrreldes konstantse mudeliga

Hüpoteeside kontroll mudeli iga faktori (ja nende kombinatsiooni) kohta
$H_0: a_1 = \alpha_2 = \ldots = 0$,
H_1: leidub i, et $a_i \neq 0$;
$H_0: \beta_1 = \beta_2 = \ldots = 0$,
H_1: leidub j, et $\beta_j \neq 0$;
$H_0: \gamma_{11} = \gamma_{12} = \ldots = 0$,
H_1: leiduvad i,j, et $\gamma_{ij} \neq 0$.

Kovariatsioonanalüüs

Näide. Sooviti uurida, kas lehmade suurus on aastate jooksul muutunud.
Kasutada olid kahe aasta andmed ühest ja kolme aasta andmed teisest majandist.
Uuritavad tunnused: mass esmapoegimisel, ristluu kõrgus.
Faktorid: farm (diskreetne, 2 taset), aasta (pidev).
Mudel: $y_{ij} = \mu + \alpha_i + b * x_{ij} + b_i * x_{ij} + \varepsilon_{ij}$



Faktor	Sõltuv tunnus	
	Esmapg. mass	Ristluu kõrgus
Aasta	0,013	<0,001
Farm	<0,001	0,84
Aasta*Farm	<0,001	0,81

Post-hoc testid

Eeldame, et dispersioonanalüüsi (F-testi, mida nimetatakse ka ***omnibus*-testiks**) tulemusena on faktori mõju statistiliselt oluline.

Aga milles see olulisus seisneb?
Millised faktori tasemed on erinevad?

Põhimõtteliselt võib viia läbi nii palju t-teste, kui on huvipakkuvaid võrdlusi, aga siis tekkib mitmese testimise probleem.

Lahenduseks on rakendada t-testide tulemustele mõnd juba kirjeldatud mitmese testimise suhtes korrigeerimise meetodit (näiteks Bonferroni korrektsiooni) või rakendada mõnda nn ***post-hoc* testi**.

Post-hoc testid (*post-hoc* – mõte üksikute tasemete omavahelisest võrdlusest tekkis alles peale dispersioonanalüüsi tulemuste selgumist …) püüavad lisaks faktorite tasemete võrdlemisele kontrolli all hoida ka mitmesel võrdlemisel tekkivat nn katseviisilist viga (*experiment-wise error rate*).

Post-hoc testid

Dunnett test – kui soovitakse teha võrdlusi vaid ühe tasemega (näiteks võrrelda 5 alternatiivi praegu kasutatava standardiga, alternatiivide omavaheline järjestus huvi ei paku);

Tukey test – kui soovitakse võrrelda kõikide gruppide keskmisi omavahel;

Scheffe test – kui soovitakse teha ka keerukamaid võrdluseid (näiteks kas sortide A ja B keskmine saagikus on parem kui sortide C, D ja E keskmine saagikus?);

Fisher LSD (*least significant difference*) **test** – pole päris korrektne mitmese testimise protseduur, sest I-liiki vea tegemise tõenäosus võib olla märgatavalt suurem kui 0,05; põhimõte: teeme esmalt F-testi abil selgeks, et faktori mõju on statistiliselt oluline, kui on, siis unustame, et tegemist on mitmese testimisega, ja teeme läbi kõik üksiktestid.

Üldine lineaarne mudel

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

Katsepõhine *vs* mudelipõhine uuring

- Katsepõhine uuring

  katsetingimused range kontrolli all,
  suhteliselt vähe ja enamasti tasakaalus [*balanced*] andmed,
  analüüsiks standardne regressioon- või dispersioonanalüüs (t-test).

- Mudelipõhine uuring

  juhuslikud ja enamasti mittetasakaalus [*unbalanced*] andmed,
  mittekontrollitud katsetingimused,
  peamine analüüsi alus on uurija intuitsioon/analüüsitava materjali tundmine,
  meetodeiks mitmefaktorilised, sageli mittestandardse vaatluste kovariatsioonistruktuuriga mudelid.

- Mudel mõlemal juhul

  mõõdetud väärtus = sobitatud väärtus + viga.

Sõltuvad ja sõltumatud tunnused

- **Uuritavad** e **sõltuvad tunnused** [*dependent variables*] – tunnused, mille käitumine huvi pakub.
- **Argument-** e **sõltumatud tunnused** [*idependent variables*] e **faktorid** – tunnused, mille mõju uuritavatele tunnustele soovitakse selgitada.
- Faktortunnuse erinevaid väärtusi nimetatakse **tasemeteks** e **nivoodeks** [*levels*].
- Iga faktor jaguneb vastavalt oma tasemete iseloomule **diskreetseks** või **pidevaks**, **arvuliseks** (kvantitatiivne) või **klassifitseerivaks** (kvalitatiivne).

  Näiteks on lehma sünniaasta, laktatsioon jne diskreetsed arvulised faktorid;
  farm tasemetega (väärtustega) 'Vorbuse', 'Ülenurme' jne on diskreetne klassifitseeriv faktor;
  laktatsiooni pikkus, piimatoodang, pekipaksus jne (mõõdetud tunnused) on aga pidevad faktorid.

Faktorite vahekord mudelis

- Faktorid on **lihtsad** ja **tuletatud**. Lihtsate faktorite väärtused on vahetult mõõdetud või registreeritud, tuletatud faktorid moodustatakse lihtsatest. Tüüpilised tuletatud faktorid on **interaktsioonid** e. koosmõjud ning arvuliste faktorite korrutised.

  Näiteks on farm, isa ja laktatsiooni pikkus lihtsad faktorid; farm*isa (koosmõju) ja laktatsioon*laktatsioon (arvulise faktori kõrgem järk) aga tuletatud faktorid.
- Praktikas on faktorite vahel sageli ka **alluvusseosed**. Faktor *A* allub faktorile *B*, kui *A* iga nivoo (tase) esineb koos vaid ühe B nivooga.

  Näiteks võime me tavaliselt lugeda farmi allutatuks maakonnale; kui iga ema on ristatud kindla isaga, on ema allutatud isale.

- Faktorid *A* ja *B* on **ristseoses**, kui *A* iga nivoo kombineerub (saab põhimõtteliselt kombineeruda) *B* kõigi nivoodega.

  Näiteks kui viiel aastal on uuritud pullide tütarde jõudlusandmeid ja igal pullil on igal aastal tütreid, on pull ja aasta ristseoses; kui aga igal aastal on valitud uued pullid, allub pull aastale.

Lineaarsed mudelid

- Lineaarne mudel sisaldab komplekti faktoreid, mis mõjutavad vaatlusi aditiivselt, kusjuures mingi muutuja faktori siseselt võib olla näiteks ruutu võetud.
- Lineaarseid mudeleid sobib rakendada enamustes bioloogilistes uuringutes.
- Mittelineaarsed seosed on tihti lähendatavad lineaarse mudeliga.

- Traditsiooniline lineaarne mudel koosneb kolmest osast:
  - võrrand – mudeli esitus faktorite mõjude summana;
  - juhuslike muutujate keskväärtused ja dispersioonistruktuur;
  - eeldused, kitsendused ja piirangud.

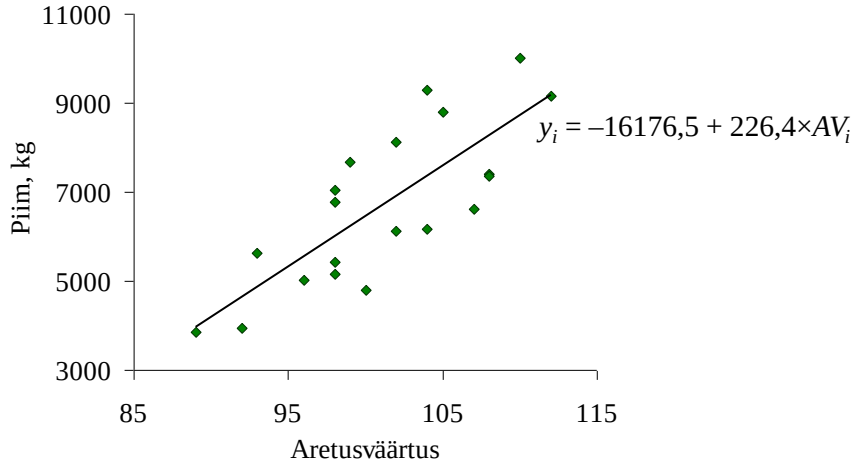
Näiteandmestik

Lehm	Tõug	Farm	Aretusväärtus	Piim. kg
1	EHF	F1	105	8804.256
2	EHF	F3	112	9152.284
3	EHF	F4	98	7055.046
4	EHF	F2	89	3856.88
5	EHF	F1	98	6768.067
6	EHF	F4	99	7676.258
7	EHF	F4	104	9282.086
8	EHF	F1	102	8126.694
9	EHF	F1	110	10017.95
10	EHF	F1	93	5622.356
11	EPK	F2	98	5431.155
12	EPK	F2	108	7406.513
13	EPK	F4	98	5152.659
14	EPK	F3	100	4797.637
15	EPK	F3	96	5011.426
16	EPK	F4	108	7369.143
17	EPK	F2	107	6626.611
18	EPK	F2	104	6170.835
19	EPK	F2	92	3948.281
20	EPK	F3	102	6113.998

Diskreetsed faktorid — Pidevad faktorid — Sõltuv muutuja

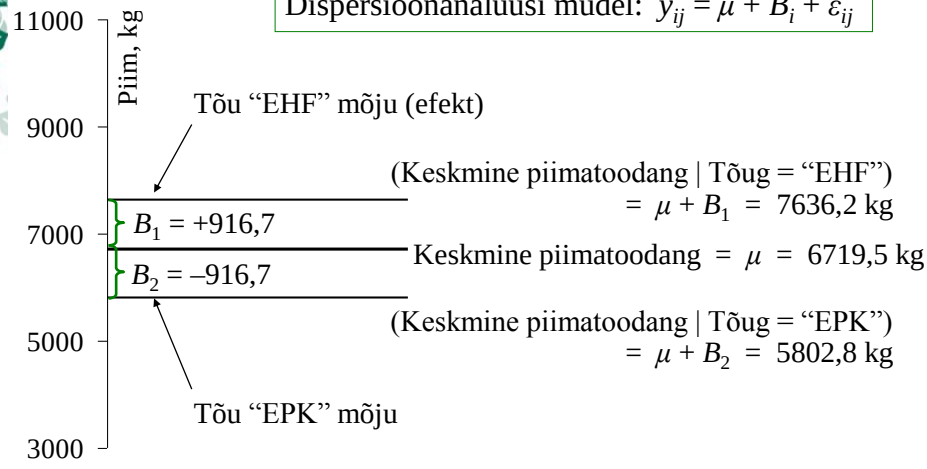
Mudeli esitus

Regressioonanalüüsi mudel: $y_i = \mu + b\ AV_i + \varepsilon_i$



Mudeli esitus

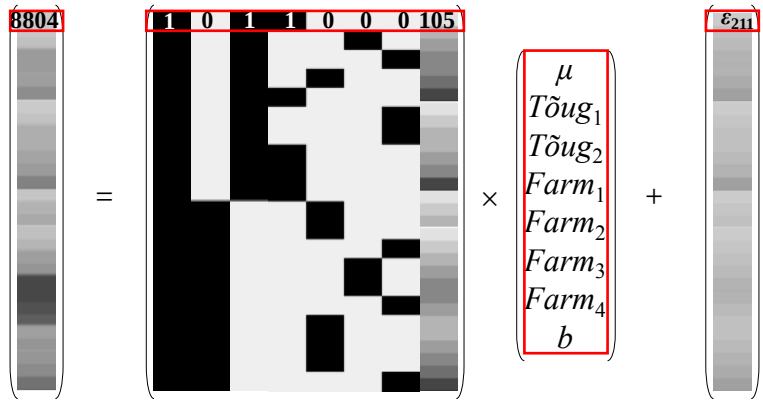
Dispersioonanalüüsi mudel: $y_{ij} = \mu + B_i + \varepsilon_{ij}$



Mudeli esitus

$$y_{ijk} = \mu + T_i + F_j + b \times AV_{ijk} + \varepsilon_{ijk}$$

$$y_{211} = \mu + 0\times T_1 + 1\times T_2 + 1\times F_1 + 0\times F_2 + 0\times F_3 + 0\times F_4 + b \times AV_{211} + \varepsilon_{211}$$



$$\mathbf{y} = \mathbf{X} \times \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} 8804{,}3 \\ 9152{,}3 \\ 7055{,}0 \\ 3856{,}9 \\ 6768{,}1 \\ 7676{,}3 \\ 9282{,}1 \\ 8126{,}7 \\ 10018{,}0 \\ 5622{,}4 \\ 5431{,}2 \\ 7406{,}5 \\ 5152{,}7 \\ 4797{,}6 \\ 5011{,}4 \\ 7369{,}1 \\ 6626{,}6 \\ 6170{,}8 \\ 3948{,}3 \\ 6114{,}0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 105 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 112 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 98 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 89 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 98 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 99 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 104 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 102 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 110 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 93 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 98 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 108 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 98 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 100 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 96 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 108 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 107 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 104 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 92 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 102 \end{pmatrix} \times \begin{pmatrix} \mu \\ Tõug_1 \\ Tõug_2 \\ Farm_1 \\ Farm_2 \\ Farm_3 \\ Farm_4 \\ b \end{pmatrix}$$

Hinnatavad efektid, reparametriseerimine

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Probleem on, et **β** ei ole üheselt hinnatav.

Vaatame näiteks ANOVA-mudelit: $y_{ij} = \mu + B_i + \varepsilon_{ij}$
Keskmine piimatoodang (Tõug = "EHF") = $\mu_1 = \mu + B_1 = 7636{,}2$ kg
Keskmine piimatoodang (Tõug = "EPK") = $\mu_2 = \mu + B_2 = 5802{,}8$ kg
Meil on 2 võrrandit ja 3 tundmatut parameetrit.

Lahendus? Reparametrisatsioon = lisakitsendused

- Klassikaline reparametrisatsioon: $B_1 + B_2 = 0$
($\mu = 6719{,}5$; $B_1 = 916{,}7$; $B_2 = -916{,}7$)
- *SAS*-i reparametrisatsioon: $B_2 = 0$ ($B_1 = 1833{,}4$; $\mu = 5802{,}8$)
- *R*-i reparametrisatsioon : $B_1 = 0$ ($B_2 = -1833{,}4$; $\mu = 7636{,}2$)

Hinnatavad efektid, reparametriseerimine

SAS

```
proc glm data=dk0007;
 class breed farm;
 model milk = breed farm bv / solution;
run;
```

```
                                The SAS System          14:16 Wednesday, February 13, 2008   8

                                The GLM Procedure

Dependent Variable: milk   milk

                                                        Standard
Parameter                 Estimate                         Error    t Value    Pr > |t|

Intercept              -16637.09733 B               1893.708817      -8.79      <.0001
breed      EHF          1527.26447 B                291.258313       5.24      0.0001
breed      EPK             0.00000 B                        .          .           .
farm       F1            -95.49905 B                327.411139      -0.29      0.7748
farm       F2           -678.20733 B                321.650812      -2.11      0.0535
farm       F3           -753.46783 B                340.016142      -2.22      0.0438
farm       F4              0.00000 B                        .          .           .
bv                       227.09839                  18.304106      12.41      <.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal
      equations.  Terms whose estimates are followed by the letter 'B' are not uniquely estimable.
```

Hinnatavad efektid, reparametriseerimine

RGui

```
> cow.model1 <- lm(milk  ~ breed  + farm  + bv , data=dk0007_glm)

> summary(cow.model1)

Call:
lm(formula = milk ~ breed + farm + bv, data = dk0007_glm)

Residuals:
    Min      1Q  Median      3Q     Max
-566.84 -373.07   36.75  312.64  773.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15205.3     1872.2  -8.122 1.15e-06 ***
breed[T.EPK]  -1527.3      291.3  -5.244 0.000124 ***
farm[T.F2]     -582.7      384.2  -1.517 0.151575
farm[T.F3]     -658.0      390.5  -1.685 0.114160
farm[T.F4]       95.5      327.4   0.292 0.774809
bv              227.1       18.3  12.407 6.09e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 483.5 on 14 degrees of freedom
Multiple R-Squared: 0.9455,     Adjusted R-squared: 0.9261
F-statistic:  48.6 on 5 and 14 DF,  p-value: 2.354e-08
```

Hinnatavad funktsioonid, kontrastid

Kontrast on mudeli parameetrite hinnatav lineaarkombinatsioon.

Kontrastide esitamiseks sobib kasutada maatrikskorrutist kujul **lβ**.

Näiteks kontrast, hindamaks tõugudevahelist erinevust, on esitatav kujul

$$\mathbf{l}\times\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \mu \\ Tõug_1 \\ Tõug_2 \\ Farm_1 \\ Farm_2 \\ Farm_3 \\ Farm_4 \\ b \end{pmatrix} = 1\times Tõug_1 - 1\times Tõug_2$$

Milline efekt (erinevus) on hinnatav reavektori **l** abil:

l = (0 0 0 0,5 0,5 –0,5 –0,5 0)?

Vähimruutkeskmised [*least square means, marginal means*]

Vähimruutkeskmine [VRK] kujutab enesest mingi faktori mingile tasemele vastavate väärtuste keskmist, mis on hinnatud mudelist sobivalt defineeritud kontrasti kujul.

Näiteks 2. farmi lehmade piimatoodangu vähimruutkeskmine hinnatakse kujul

$$\text{LSM}(Farm_2) = \left(1 \;\; \tfrac{1}{2} \;\; \tfrac{1}{2} \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; \overline{av}\right) \times \begin{pmatrix} \mu \\ Tõug_1 \\ Tõug_2 \\ Farm_1 \\ Farm_2 \\ Farm_3 \\ Farm_4 \\ b \end{pmatrix}$$

$$= 1 \times \mu + \frac{Tõug_1 + Tõug_2}{2} + 1 \times Farm_2 + b \times \overline{av}$$

Vähimruutkeskmised [*least square means*]

SAS

```
proc glm data=dk0007;
 class breed farm;
 model milk = breed farm bv / solution;
 lsmeans breed / stderr pdiff;
run;
```

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-16637.09733	B	1893.708817	-8.79	<.0001
breed	EHF	1527.26447	B	291.258313	5.24	0.0001
breed	EPK	0.00000	B	.	.	.
farm	F1	-95.43905	B	327.411139	-0.29	0.7748
farm	F2	-678.20733	B	321.650812	-2.11	0.0535
farm	F3	-753.46783	B	340.016142	-2.22	0.0438
farm	F4	0.00000	B	.	.	.
bv		227.09839		18.304106	12.41	<.0001

The GLM Procedure
Least Squares Means

breed	milk LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 Pr > |t|
EHF	7479.37593	180.89567	<.0001	0.0001
EPK	5952.11146	183.24601	<.0001	

keskmine aretusväärtus

$$LSM_{EHF} = Intercept + 1 \times EHF + 0 \times EPK + \frac{(F1 + F2 + F3 + F4)}{4} + coef \times \overline{BV}$$

$$= -16637{,}1 + 1527{,}3 + \frac{(-95{,}5 - 678{,}2 - 753{,}5 + 0)}{4} + 227{,}1 \times 101{,}15 \approx 7479{,}5$$

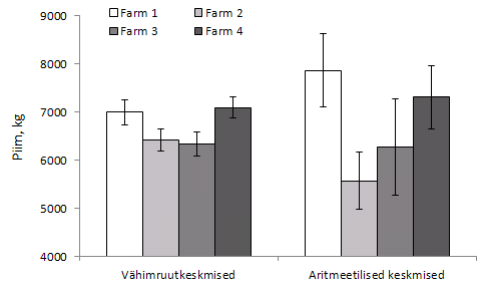
Vähimruutkeskmised *vs* aritmeetilised keskmised

$$y_{ijk} = \mu + T_i + F_j + b \times AV_{ijk} + \varepsilon_{ijk}$$

Least Squares Means

farm	milk LSMEAN	Standard Error	Pr > |t|
F1	7002.03820	261.29676	<.0001
F2	6419.32992	222.82838	<.0001
F3	6344.06942	253.00687	<.0001
F4	7097.53725	218.31010	<.0001

Level of farm	N	milk Mean	milk Std Dev
F1	5	7867.86471	1718.90621
F2	6	5573.37915	1444.58775
F3	4	6268.83613	2006.96723
F4	5	7307.03830	1478.66858



Keskmiste võrdlemine

Variant 1: *post-hoc* testid

Võrreldavad keskmised ei ole teiste faktorite mõjude ja andmebaasi mittetasakaalulisuse suhtes korrigeeritud, st et võrreldakse aritmeetilisi keskmisi; küll sõltuvad mudelist keskmiste hinnangute varieeruvus ja seeläbi ka erinevuste statistiline olulisus.

$$y_{ij} = \mu + F_i + \varepsilon_{ij}$$

Tukey's Studentized Range (HSD) Test for milk

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	16
Error Mean Square	2692645
Critical Value of Studentized Range	4.04609

Comparisons significant at the 0.05 level are indicated by ***.

farm Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
F1 - F4	560.8	-2408.4 3530.0	
F1 - F3	1599.0	-1550.3 4748.3	
F1 - F2	2294.5	-548.3 5137.3	
F4 - F1	-560.8	-3530.0 2408.4	
F4 - F3	1038.2	-2111.1 4187.5	
F4 - F2	1733.7	-1109.1 4576.5	
F3 - F1	-1599.0	-4748.3 1550.3	
F3 - F4	-1038.2	-4187.5 2111.1	
F3 - F2	695.5	-2335.0 3725.9	
F2 - F1	-2294.5	-5137.3 548.3	
F2 - F4	-1733.7	-4576.5 1109.1	
F2 - F3	-695.5	-3725.9 2335.0	

$$y_{ijk} = \mu + T_i + F_j + b \times AV_{ijk} + \varepsilon_{ijk}$$

Tukey's Studentized Range (HSD) Test for milk

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	14
Error Mean Square	233818.8
Critical Value of Studentized Range	4.11051

Comparisons significant at the 0.05 level are indicated by ***.

farm Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
F1 - F4	560.8	-328.1 1449.7	
F1 - F3	1599.0	656.2 2541.8	***
F1 - F2	2294.5	1443.4 3145.5	***
F4 - F1	-560.8	-1449.7 328.1	
F4 - F3	1038.2	95.4 1981.0	***
F4 - F2	1733.7	882.6 2584.7	***
F3 - F1	-1599.0	-2541.8 -656.2	***
F3 - F4	-1038.2	-1981.0 -95.4	***
F3 - F2	695.5	-211.8 1602.7	
F2 - F1	-2294.5	-3145.5 -1443.4	***
F2 - F4	-1733.7	-2584.7 -882.6	***
F2 - F3	-695.5	-1602.7 211.8	

Keskmiste võrdlemine

Variant 2: kontrastid (vähimruutkeskmiste võrdlus)

$$y_{ijk} = \mu + T_i + F_j + b \times AV_{ijk} + \varepsilon_{ijk}$$

Least Squares Means

farm	milk LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
F1	7002.03820	261.29676	<.0001	1
F2	6419.32992	222.82838	<.0001	2
F3	6344.06942	253.00687	<.0001	3
F4	7097.53725	218.31010	<.0001	4

Least Squares Means for effect farm
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: milk

i/j	1	2	3	4
1		0.1516	0.1142	0.7748
2	0.1516		0.8162	0.0535
3	0.1142	0.8162		0.0438
4	0.7748	0.0535	0.0438	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

I ja III tüüpi ruutude summad

SAS

```
proc glm data=dk0007;
 class breed farm;
 model milk = breed farm bv / solution;
run;
```

The SAS System 14:16 Wednesday, February 13, 2008 8

The GLM Procedure

Dependent Variable: milk milk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	56822518.79	11364503.76	48.60	<.0001
Error	14	3273463.43	233818.82		
Corrected Total	19	60095982.22			

R-Square	Coeff Var	Root MSE	milk Mean
0.945529	7.196185	483.5482	6719.507

Source	DF	Type I SS	Mean Square	F Value	Pr > F
breed	1	16806079.82	16806079.82	71.88	<.0001
farm	3	4024063.57	1341354.52	5.74	0.0089
bv	1	35992375.40	35992375.40	153.93	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
breed	1	6429116.53	6429116.53	27.50	0.0001
farm	3	1510502.69	503500.90	2.15	0.1392
bv	1	35992375.40	35992375.40	153.93	<.0001

I ja III tüüpi ruutude summad

RGui

```
> anova(cow.model1)
Analysis of Variance Table

Response: milk
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
breed      1 16806080 16806080  71.8765 6.925e-07 ***
farm       3  4024064  1341355   5.7367  0.008943 **
bv         1 35992375 35992375 153.9328 6.089e-09 ***
Residuals 14  3273463   233819
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(cow.model1, test="F")
Single term deletions

Model:
milk ~ breed + farm + bv
       Df Sum of Sq      RSS      AIC  F value     Pr(F)
<none>               3273463      252
breed   1   6429117  9702580      272  27.4961 0.0001243 ***
farm    3   1510503  4783966      254   2.1534 0.1392264
bv      1  35992375 39265839      300 153.9328 6.089e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Korduvad mõõtmised

ID	Lakt.	Piim, kg
3396	1	4119
3396	2	5857
3396	3	6260
3990	1	3106
3990	2	3934
3990	3	5171
4390	1	2473
4390	2	3301
4390	3	2958
…		

Vaikimisi:

$$\text{var}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 & \cdots \\ 0 & \sigma_\varepsilon^2 & 0 & \cdots \\ 0 & 0 & \sigma_\varepsilon^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Kompaund-sümmeetriline kovariatsioonistruktuur:

$$\text{var}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma_\varepsilon^2+\sigma^2 & \sigma^2 & \sigma^2 & 0 & 0 & \cdots \\ \sigma^2 & \sigma_\varepsilon^2+\sigma^2 & \sigma^2 & 0 & 0 & \cdots \\ \sigma^2 & \sigma^2 & \sigma_\varepsilon^2+\sigma^2 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \sigma_\varepsilon^2+\sigma^2 & \sigma^2 & \cdots \\ 0 & 0 & 0 & \sigma^2 & \sigma_\varepsilon^2+\sigma^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Esimest järku autoregressiivne kovariatsioonistruktuur:

$$\text{var}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \times \begin{pmatrix} 1 & \rho & \rho^2 & 0 & 0 & 0 & \cdots \\ \rho & 1 & \rho & 0 & 0 & 0 & \cdots \\ \rho^2 & \rho & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \rho & \rho^2 & \cdots \\ 0 & 0 & 0 & \rho & 1 & \rho & \cdots \\ 0 & 0 & 0 & \rho^2 & \rho & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, |\rho| \le 1$$

Mudelite võrdlemine

Hierarhiliste mudelite võrdlemiseks

¤ lihtsamal juhul dispersioonanalüüs (*F*-statistik -> *p*-väärtus)

¤ keerulisemal juhul tõepärasuhte test [*likelihood ratio test*]

$$\Lambda(\mathbf{z}) = -2\ln\left[\frac{\hat{L}_k(\mathbf{z})}{\hat{L}(\mathbf{z})}\right] = -2\left\{\ln\left[\hat{L}_k(\mathbf{z})\right] - \ln\left[\hat{L}(\mathbf{z})\right]\right\} \underset{H_0}{\sim} \chi^2(r)$$

Üldisema võrdlemise tarvis

¤ AIC (Akaike informatsiooni kriteerium)

¤ BIC (Bayesi informatsiooni kriteerium)

NB! Testida ei saa, mida väiksem väärtus, seda parem.