

# Biomeetria

## Enam kui kahe populatsiooni keskväärtuste võrdlemine – dispersioonanalüüs

### Enam kui kahe grupi keskmiste võrdlus



$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1: \text{leiduvad sellised grupid } i, j, \text{ et } \mu_i \neq \mu_j$$

Eeldustel, et

- uuritav (sõltuv) tunnus on normaaljaotusega ja
- uuritava tunnuse varieeruvus võrreldavais gruppides on ühesugune, on rakendatavaks analüüsimeetodiks dispersioonanalüüs.

Dispersioonanalüüsil jagatakse tunnused vastavalt nende rollile kaheks: tunnus, mille keskmisi võrrelda soovitakse, on uuritav tunnus e funktsioon-tunnus (lehma piimatoodang, forelli kasvukiirus, talle mass, sea pekipaksus, jne); (diskreetne või mitteamruline) tunnus, mille väärtuste alusel võrreldavad grupid moodustatakse, on faktortunnus (tõug, lüpsiseade, laudatüüp jne).

Dispersioonanalüüsi tulemuste tõlgendamisel räägitaksegi enamasti faktortunnuse 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

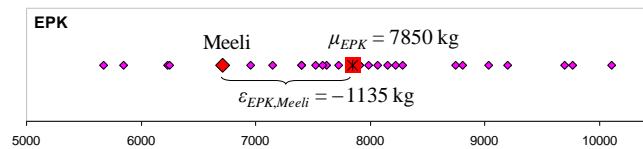


✓ Igat gruppi  $i$  (kus  $i=1, \dots, k$ ) iseloomustab keskmine uuritava tunnuse väärtus  $\mu_i$ , mistõttu mõõtmistulemused saab esitada mudeliga

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

kus  $y_{ij}$  on uuritava tunnuse väärtus  $i$ . gruppi kuuluval  $j$ . objektil ja  $\varepsilon_{ij}$  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 tööpõhimõte



Dispersioonanalüüsi tööpõhimõte seisneb uuritava tunnuse gruppisisese (nn juhusliku) varieeruvuse  $SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$

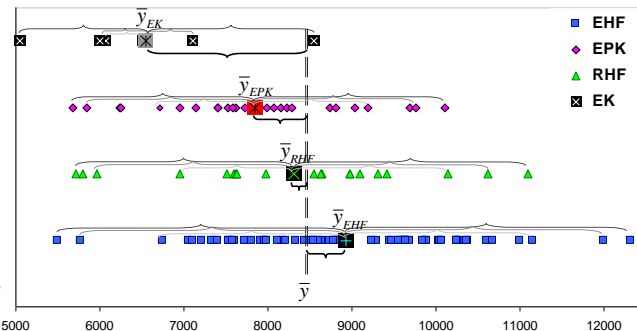
ja gruppide vahelise (faktori mõjust tingitud) varieeruvuse

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

võrdlemises – kui gruppide vaheline erinevus on suurem kui gruppisisene 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	$SS = SSA + SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	$n - 1$			

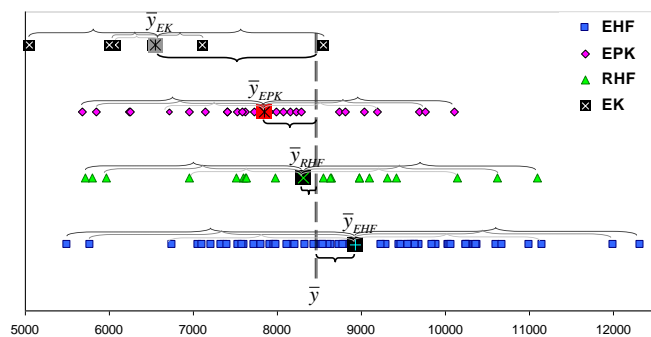
Juhul, kui faktortunnuse mõjule vastav keskmine gruppide vaheline varieeruvus  $MSA$  on suurem, kui uuritavate objektide omapärale vastav keskmine gruppide sisene varieeruvus  $MSE$ , on  $F$ -statistiku väärtus ühest suurem.

Piisavalt suure  $F$ -suhte väärtuse korral võib lugeda tõestatuks sisuka hüpoteesi – leiduvad vähemalt 2 teineteisest selgelt eristuvat gruppi.

## 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	<b>0,000019</b>
Viga	203894493	117	1742688		
Kokku	251224927	120			

$< 0,05 \Rightarrow$   
 $H_1$ : 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 väljatrükkid.*

	Tõug 1	Tõug 2
Mean	540	675
Variance	333,33	966,67
Observations	4	4
Pooled Variance	650	
Hypothesized Mean Difference	6	
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	

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



## Keskväärtuste võrdlemine

Keskväärtuste võrdlemine

**1 grupi keskmise võrdlus konstandiga**

$H_0 : \mu = c$   
 $H_1 : \mu \neq c$

Usalduspiirid;  
normaaljaotuse eeldusel  
t-test;  
suurte valimite ( $n > 60$ ) või teadaoleva dispersiooni  $\sigma^2$  korral  
z-test

**2 grupi keskmiste võrdlus**

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


t-test  
(kolm erinevat!),  
eeldused: uuritav tunnus normaaljaotusega (või suur valim);  
teadaolevate dispersioonide  $\sigma_1^2$  ja  $\sigma_2^2$  korral  
z-test

**3 või enama ( $k$ ) grupi keskmiste võrdlus**

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$   
 $H_1 : \text{leiduvad } i, j: \mu_i \neq \mu_j$

Dispersioonanalüüs,  
eeldused:  
□ uuritav tunnus normaaljaotusega,  
□ uuritava tunnuse varieeruvus võrreldavais gruppides ühesugune

## 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  $(1-\alpha)^6=0,95^6\approx 0,735$ .


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

Bonferroni meetod: piiramaks  $k$  üksikvõrdluse puhul ühe või enama vea tegemise tõenäosust olulisuse nivooaga  $\alpha$ , tuleb kõigil üksikvõrdlustel võtta olulisuse nivooks  $\alpha/k$ .

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

$$\begin{array}{cccc} & & \alpha=0,05 & \\ & \alpha=0,05 & \alpha=0,05 & \alpha=0,05 \\ \text{I} & \text{II} & \text{III} & \text{IV} \\ & \alpha=0,05 & \alpha=0,05 & \end{array}$$

## Dispersioonanalüüsi mudel



✓ Faktortunnuse mõju uurimiseks esitatakse mudel kujul

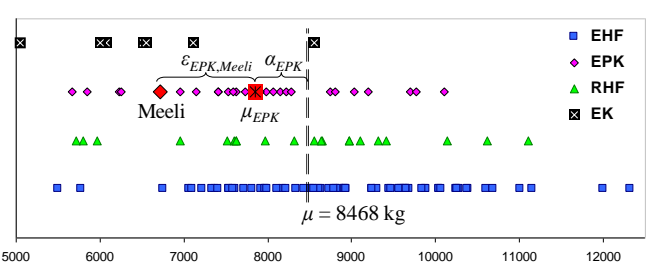
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

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

$H_0: \mu_1 = \mu_2 = \dots = \mu_k \Leftrightarrow H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$   
 $H_1: \text{leiduvad grupid } i, j, \text{ et } \mu_i \neq \mu_j \Leftrightarrow H_1: \text{leidub grupp } i, \text{ et } \alpha_i \neq 0$

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 üldkeskmisest)  $\alpha_{EPK} = -618$  kg ja Meeli tõusisese erinevuse  $\varepsilon_{EPK, Meeli} = -1135$  kg summa.



## Dispersioonanalüüs



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

$$\begin{aligned} H_0: \mu_{EHF} = \mu_{RHF} = \mu_{EPK} = \mu_{EK} & \Leftrightarrow H_0: \text{tõul ei ole mõju} \\ H_1: \text{leiduvad tõugrupid } i, j, \text{ et } \mu_i \neq \mu_j & H_1: \text{tõul on mõju} \end{aligned}$$

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

$$\begin{aligned} \alpha_{EHF} &= 461,8 \text{ kg}; \alpha_{RHF} = -156,6 \text{ kg}; \\ \alpha_{EPK} &= -618,0 \text{ kg ja } \alpha_{EK} = -1918,8 \text{ kg}. \end{aligned}$$

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

## Mitmefaktoriline dispersioonanalüüs



Kui vaatlusobjekte saab rühmitada mitme tunnuse (faktortunnuse) järgi, võib osutada mõttekaks analüüsida korruga mitme faktortunnuse mõju (näiteks igal lehmil 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  $\mu$  tähistab üldkeskmist,

$\alpha_i$ -d ja  $\beta_j$ -d märgivad uuritava tunnuse keskmise muutust vastavalt esimese ja teise faktori väärtuste muutumisele ( $\alpha_i$  on esimese faktori  $i$ . taseme mõju ja  $\beta_j$  on teise faktori  $j$ . taseme mõju),

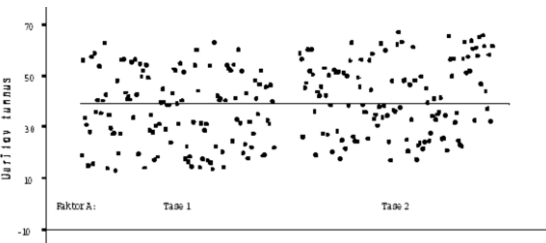
$y_{ijk}$  ning  $\varepsilon_{ijk}$  on vastavalt esimese faktori  $i$ . tasemel ja teise faktori  $j$ . tasemel sooritatud  $k$ . mõõtmise väärtus ning selle erinevus sama väärtuste kombinatsiooni keskmisest (vaatluse omapära, mudeli viga).

## Mitmefaktoriline dispersioonanalüüs



Miks seda vaja on?

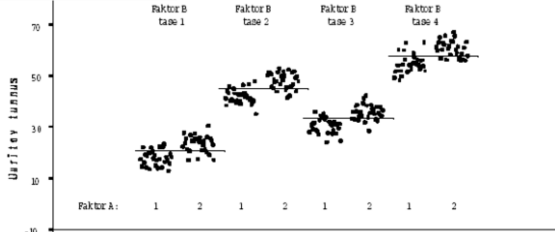
1) Hinnangute ja otsustuste täpsus võib paraneda.




Illustratsiooniks kaks samu andmeid illustreerivat hajuvusdiagrammi.

(joonised M. Mölsi konspektist)

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



## Mitmefaktoriline dispersioonanalüüs

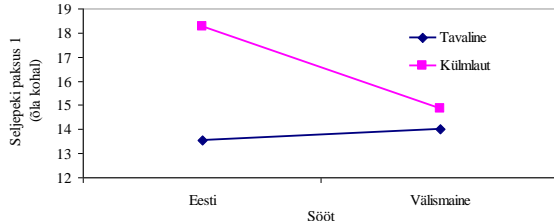


Miks seda vaja on?

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 pruugi olla normaaljaotusega).



Sigade seljapeki paksus 1 sõltuvalt söödast ja lauda tüübist

Kahefaktoriline faktoritevahelist interaktsiooni arvestav mudel esitatakse kujul

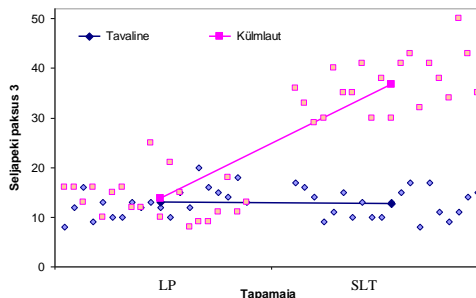
$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

kus  $\gamma_{ij}$  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 “parematele 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

Hüpooteeside kontroll mudeli iga faktori (ja nende kombinatsiooni) kohta

$H_0$  :  $a_1 = a_2 = \dots = 0$ ,

$H_1$  : leidub  $i$ , et  $a_i \neq 0$ ;

$H_0$  :  $\beta_1 = \beta_2 = \dots = 0$ ,

$H_1$  : leidub  $j$ , et  $\beta_j \neq 0$ ;

$H_0$  :  $\gamma_{11} = \gamma_{12} = \dots = 0$ ,

$H_1$  : leiduvad  $i, j$ , et  $\gamma_{ij} \neq 0$ .