DAUGAVPILS UNIVERSITĀTE
Dabaszinātņu un matemātikas fakultāte
Matemātikas katedra

Studiju kurss

Eksperimentu plānošana un analīze

3.lekcija

Docētājs: Dr. P. Daugulis

2008./2009.studiju gads

Saturs

Lekcijas mērķis:

- apgūt viena faktora eksperimentu statistisko analīzi, ieskaitot dispersiju analīzi,
- apgūt blokošanas pamatus.

1. Viena faktora eksperimenti

1.1. Ievads

1.1.1. Pamatjēdzieni

Viena faktora eksperiments:

- ir viens faktors,
- faktoram ir m līmeņi.

Ir iespējami divu gadījumi:

- faktora līmeņi ir fiksēti - *fiksēto efektu eksperiments*,
- faktora līmeņi tiek izvēlēti nejaušā veidā - *nejaušo efektu eksperiments*.

Pagaidām apskatīsim tikai fiksēto efektu eksperimentus. Vispārīgā gadījumā uzskatām, ka efektu līmeņi ir doti kvalitatīvi - neizmantosim to skaitliskos lielumus.

Katram faktora līmenim atbilst savs gadījuma lielums, kuru reprezentē eksperimenta rezultāts - izlase.

1.1. piemērs. Ir jānosaka betona stiprība atkarībā no cementa īpatsvara. Tiek izvēlēti 10 īpatsvara lielumi, katram gadījumam tiek ņemti vairāki paraugi, tiek mērītas visu paraugu stiprības. Faktors - cementa īpatsvars.

Viena faktora eksperimenta gadījumā ir jāsalīdzina m izlases, kas atbilst, iespējams, dažādām normālā sadalījuma funkcijām ar dažādiem parametriem (vidējo vērtību un dispersiju).

Dotas m izlases

$$\left\{ \begin{array}{l} X_1 = \{x_{11}, ..., x_{1n}\} \\ X_2 = \{x_{21}, ..., x_{2n}\} \\ ... \\ X_m = \{x_{k1}, ..., x_{mn}\} \end{array} \right.$$

Apskatīsim tikai vienkāršāko gadījumu, kad izlasēs ir vienāds skaits elementu - n. Vispārīgo gadījumu apskatīt patstāvīgi.

1.1. piezīme. Ievērosim, ka vienas un divu izlašu analīze ir viena faktora eksperimenta speciālgadījumi.

Lietderīgi ir izmantot kastes-ūsu diagrammas.

1.2. piezīme. SPSS → Graphs → Interactive.
SPSS → Analyze → Descriptive Statistics → Explore

1.1.2. Statistiskie modeļi

Vidējo vērtību modelis:

$$x_{ji} = \mu_j + \epsilon_{ji},$$

kur

- x_{ji} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai parametra vērtībai,
- μ_j ir faktora j-tajam līmenim atbilstošās ģenerālkopas apakškopas vidējā vērtība,
- ϵ_{ji} ir kļūdas gadījuma lielums, $E(\epsilon_{ji}) = \overline{\epsilon_j} = 0$, parasti uzskata arī, ka $\epsilon_{ji} \sim N(0, \sigma^2)$.

Faktora efektu modelis:

$$x_{ji} = \mu + \tau_j + \epsilon_{ji},$$

kur

- x_{ji} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai parametra vērtībai,

- μ ir visas ğenerālkopas vidējā vērtība,
- τ_j ir faktora j-tajam līmenim atbilstošās vidējās vērtības un kopējā vidējās vērtības μ starpība, *faktora efekts*, apmierina vienādību
$$\sum_{j=1}^{m} \tau_j = 0,$$
- ϵ_{ji} ir kļūdas gadījuma lielums, $E(\epsilon_{ji}) = \overline{\epsilon_{ji}} = 0$, parasti uzskata arī, ka $\epsilon_{ji} \sim N(0, \sigma^2)$.

Tā kā μ, τ_j ir konstantie gadījuma lielumi, tad pieņemam, ka
$$x_{ji} \sim N(\mu + \tau_j, \sigma^2).$$

1.2. Fiksēto efektu modeļa dispersiju analīze

Viena faktora eksperimentu analīzes svarīgākā metode - *dispersiju analīze (analysis of variance, ANOVA).*

1.2.1. Hipotēzes

Definēsim šādus novērtējumus:

- j-tā līmeņa izlases vidējā vērtība

$$\overline{x_j} = \frac{\sum_{i=1}^{n} x_{ji}}{n},$$

- visas izlases vidējā vērtība

$$\overline{x} = \frac{\sum_{i,j} x_{ji}}{nm}$$

.

Ievērosim, ka

$$E(x_{ji}) = \mu_j = \mu + \tau_j.$$

Vidējo vērtību modeļa hipotēzes:

- H_0: $\mu_1 = \mu_2 = ... = \mu_m$,
- H_1: $\mu_k \neq \mu_l$ vismaz vienam pārim k, l.

Efektu modeļa hipotēzes:

- H_0: $\tau_1 = \tau_2 = ... = \tau_m = 0$,
- H_1: $\tau_k \neq 0$ vismaz vienam k.

1.2.2. Dispersijas sadalīšana

Dispersijas analīzes pamatā ir izlases pilnās dispersijas sadalīšana divu daļu summā un sekojošā statistiskā analīze.

Definēsim

$$DP = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ji} - \overline{x})^2 = \sum_{j,i} (x_{ji} - \overline{x})^2.$$

Ievērosim, ka

$$DP = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \overline{x})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \overline{x_j} + \overline{x_j} - \overline{x})^2 =$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \overline{x_j})^2 + (\overline{x_j} - \overline{x})^2 + 2(x_{ji} - \overline{x_j})(\overline{x_j} - \overline{x}) =$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \overline{x_j})^2 + \sum_{j=1}^{m}\sum_{i=1}^{n}(\overline{x_j} - \overline{x})^2 + \sum_{j=1}^{m}\sum_{i=1}^{n}2(x_{ji} - \overline{x_j})(\overline{x_j} - \overline{x}) =$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \overline{x_j})^2 + n\sum_{j=1}^{m}(\overline{x_j} - \overline{x})^2 + 2\sum_{j=1}^{m}(\overline{x_j} - \overline{x})\underbrace{\left(\sum_{i=1}^{n}(x_{ji} - \overline{x_j})\right)}_{=0}$$

Redzam, ka

$$\underbrace{DP}_{\text{pilnā dispersija}} = \underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \overline{x_j})^2}_{\text{intragrupu dispersija}} + \underbrace{n\sum_{j=1}^{m}(\overline{x_j} - \overline{x})^2}_{\text{starpgrupu dispersija}} \quad .$$

Apzīmējumi: $DP = DI + DS$.

1.3. piezīme. Var pierādīt šādas formulas:

$$DP = \sum_{j=1}^{m}\sum_{i=1}^{n} x_{ji}^2 - \frac{1}{mn}\Big(\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ji}\Big)^2,$$

$$DS = \frac{1}{n}\sum_{j=1}^{m}\Big(\sum_{j=1}^{m} x_{ji}^2\Big) - \frac{1}{mn}\Big(\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ji}\Big)^2.$$

Brīvības pakāpju skaits:

- DP - $nm-1$,
- DI - $nm-m$,
- DS - $m-1$.

Definēsim

- $MI = \frac{DI}{nm-m}$,
- $MS = \frac{DS}{m-1}$.

1.1. teorēma. Ja $x_{ji} \sim N(\mu + \tau_j, \sigma^2)$, tad

$$E(MI) = \sigma^2,$$

$$E(MS) = \sigma^2 + \frac{1}{m-1} \sum_{j=1}^{m} \tau_j^2.$$

1.2.3. Hipotēžu pārbaude

Dotas m normāla sadalījuma ģenerālkopas izlases ar vienādām dispersijām

$$X_1 = \{X_{11}, X_{12}, ..., X_{1n}\} \sim N(\mu_1, \sigma),$$
$$...$$
$$X_m = \{X_{m1}, X_{m2}, ..., X_{mn}\} \sim N(\mu_2, \sigma).$$

Statistiskais modelis:

$$X_{ji} = \mu_j + \epsilon_{ji} = \mu + \tau_j + \epsilon_{ji}, \text{ kur } \epsilon_{ji} \sim N(0, \sigma^2).$$

Uzdevums ir salīdzināt τ_j.

Parametri: τ_j.

Novērtējošās statistikas: $\mu = \overline{X}$, $\mu_j = \overline{X_j}$, DP, DI, DS, MI, MS.

Testa statistika:

$$F = \frac{MS}{MI} \sim F(m-1, nm-m).$$

1.solis
Formulēt hipotēzes:

- H_0: $\tau_1 = ... = \tau_m = 0$.
- H_1: $\tau_i \neq 0$ vismaz vienam i.

2.solis
Noteikt testa statistikas vērtību:

$$F_0 = \frac{MS}{MI} \to f_0.$$

3.solis
Noteikt kritisko vērtību $f_{\alpha,m-1,nm-m}$.

4.solis
Noraidīt H_0, ja $f_0 > f_{\alpha,m-1,nm-m}$.

1.4. piezīme. SPSS izmantošana. Dati ir jāievada divās kolonnās: pirmajā kolonnā ir faktora līmenis, otrajā kolonnā tajā pašā rindā ir atbilstošais mērījums.

SPSS → Analyze → Compare Means → One-Way ANOVA → faktoru kolonnu nosūtīt uz Factor box, rezultātu kolonnu nosūtīt uz Dependent list, atķeksēt Descriptives Statistics box izvēlnē → Continue → OK.

Tiek aprēķināta $F = f_0$ vērtība un tai atbilstošā P-vērtība (minimālā α vērtība, ar kuru var pieņemt hipotēzi).

1.2.4. Modeļa parametru novērtēšana

Faktora j-tā līmeņa grupas vidējās vērtības parametra

$$\mu_j = \mu + \tau_j,$$

$$\widehat{\mu_j} = \overline{x_j} = \frac{\sum_{i=1}^{n} x_{ji}}{n}$$

ticamības intervāls:

$$\overline{x_j} - t_{\alpha/2,nm-m}\sqrt{\frac{MI}{n}} \leq \mu_j \leq \overline{x_j} - t_{\alpha/2,nm-m}\sqrt{\frac{MI}{n}}.$$

1.3. Modeļa pareizības pārbaude

Definēsim *atlikumu:*

$$e_{ji} = x_{ji} - \overline{x_j}.$$

Lai noteiktu, vai modelis ir labs, ir jāpēta atlikumi.

1.3.1. Normalitātes pieņēmums

Lai noteiktu, vai atlikumu sadalījums ir aptuveni normāls, var izmantot normālā sadalījuma pārbaudi ($P - P$, $Q - Q$ plots).

Īpaša vērība ir jāievērš *izņēmumiem (outliers)* - mērījumiem, kas ļoti ievērojami atšķiras no normālā sadalījuma (ja tos izmet, tad sadalījums kļūst ievērojami tuvāks normālajam). Var būt nepieciešams veikt divas analīzes - ar un bez izņēmumiem.

Lai noteiktu izņēmumus var izmantot normētos atlikumus

$$d_{ji} = \frac{e_{ji}}{\sqrt{MI}} \sim N(0, 1).$$

Ja $|d_{ji}| > 3$, tad to var uzskatīt par izņēmumu.

1.3.2. Atlikumu atkarība no laika

Ir vēlams atlikt punktus (i, e_i) vai (t, e_t), kur e_i ir i-tā eksperimenta atlikums, e_t ir eksperimenta atlikums, kas ir veikts laikā t.

Uzmanība ir jāpievērš tad, ja

- vairāki pēc kārtas ejoši atlikumi nemaina zīmes,
- atlikumi samazinās laika gaitā vai ir citas tendences, kas ir atkarīgas no laika.

1.3.3. Atlikumu atkarība no rezultātiem

Ir vēlams atlikt punktus $(\overline{x_j}, e_{ji})$. Jāseko, lai nebūtu redzamas nekādas likumsakarības

Iespējama problēma - nekonstanta dispersija σ^2 (atkarīga no faktora līmeņa). Tā var palielināties, ja palielinās mērāmie lielumi x_{ji}.

Lai noteiktu, vai dispersijas ir dažādas, izmanto *Bartleta testu*. Ja dispersija ir mainīga, tad var būt nepieciešams pielietot *dispersiju stabilizējošus pārveidojumus*.

1.4. Rezultātu praktiskā interpretācija

1.4.1. Regresijas modelis

Ja faktora līmenis ir definēts kvalitatīvi, tad var mēģināt izteikt μ kā funkciju no līmeņa:

$$x = \mu_a = x(a).$$

1.2. piemērs. $x(a) = \beta_0 + \beta_1 a + \beta_2 a^2$.

1.4.2. Dažādu līmeņu vidējo vērtību salīdzināšana ar kontrastu metodi

Ja hipotēze par μ_j vienādību ir noraidīta, tad var būt nepieciešams salīdzināt līmeņu vidējās vērtības.

μ_j var salīdzināt grafiski.

Par *kontrastu* sauc lineāru kombināciju

$$\Gamma = \sum_{j=1}^{m} c_j \mu_j, \text{ kur } \sum_{j=1}^{m} c_j = 0.$$

Kontrastu var domāt kā vektoru $\vec{c} = (c_1, ..., c_m)$.

Bieži ir jāpēta hipotēze

$$H_0 : \sum_{j=1}^{m} c_j \mu_j = 0.$$

Šādas hipotēzes par pārbaudīt ar testiem, kas ir t-testa vai F-testa variācijas:

- $$\frac{\sum_{j=1}^{m} c_j \mu_j}{\sqrt{nMI \sum_{j=1}^{m} c_j^2}} \sim t_{nm-m},$$

- ja $DC = \frac{\left(\sum_{j=1}^{m} c_j \mu_j\right)^2}{n \sum_{j=1}^{m} c_j^2}$, tad

$$\frac{DC}{MI} \sim F(1, nm - m).$$

Var būt nepieciešams atrast *ievērojamos kontrastus (significant contrasts).*

Var pētīt *ortogonālos kontrastus* - kontrastu sistēmu, kas atbilst savstarpēji ortogonāliem vektoriem.

1.5. Eksperimentu skaita noteikšana

1.5.1. Ticamības intervāla metode

Nepieciešamo eksperimentu (mērījumu) skaita katrai faktora līmeņa vēretībai var noteikt, fiksējot vēlamo vidējās vērtības ticamības intervāla lielumu un atrisinot to attiecībā uz mērījumu skaitu n.

Ja ticamības intervāls ir $2l$, tad

$$l = t_{\alpha/2, mn-m} \sqrt{\frac{2MI}{n}}.$$

1.6. Regresijas pieeja dispersiju analīzei

1.6.1. Modeļa parametru novērtēšana ar mazāko kvadrātu metodi

Tika izmantots statistiskais modelis

$$x_{ji} = \mu + \tau_j + \epsilon_{ji}.$$

Ir doti eksperimentāli noteiktie lielumi x_{ji}. Kā atrast tādus μ un τ_j, lai ϵ_{ji} kopumā būtu pēc iespējas mazāki?

Definēsim

$$L = \sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \mu - \tau_j)^2.$$

Meklēsim tādas μ, τ_j vērtības, kas minimizē L.

Ir jāatrisina vienādojumu sistēma

$$\begin{cases} \frac{\partial L}{\partial \mu} = 0, \\ \frac{\partial L}{\partial \tau_j} = 0, \ \forall\ j. \end{cases}$$

Atvasinot L, iegūsim sistēmu

$$\begin{cases} -2\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ji} - \mu - \tau_j) = 0, \\ -2\sum_{i=1}^{n}(x_{ji} - \mu - \tau_j) = 0, \end{cases}$$

kuru ērtāk ir pētīt formā

$$\begin{cases} (nm)\mu + n\tau_1 + ... + n\tau_m = \overline{x}, \\ n\mu + n\tau_1 = \overline{x_1}, \\ n\mu + n\tau_2 = \overline{x_2}, \\ ... \\ n\mu + n\tau_m = \overline{x_m}. \end{cases}$$

Redzam, ka vienādojumi nav lineāri neatkarīgi. Var pievienot vēl vismaz vienu vienādojumu, parasti

$$\sum_{j=1}^{m} \tau_j = 0.$$

Atrisinot šo *mazāko kvadrātu normālo sistēmu* iegūsim modeļa parametru novērtējumus $\widehat{\mu}$, $\widehat{\tau_j}$ - to pašu rezultātu, ko ieguvām agrāk:

$$\widehat{\mu} = \overline{x}, \widehat{\tau_j} = \overline{x_j} - \overline{x}.$$

1.6.2. Dispersijas sadalījuma interpretācija

DS ir vienāda ar dispersiju, ko iegūst salīdzinot faktora dažādu līmeņu grupu vidējās vērtības ar visas mērījumu kopas vidējo, tāpēc to sauc par *izskaidroto dispersiju.* Atlikums DI ir *neizskaidrotā dispersija.*

1.7. Viena faktora eksperimenti ar nejaušiem faktora līmeņiem

Var būt situācija, kurā faktora iespējamo vērtību kopa ir liela. Ja eksperimentā faktora līmeņi tiek izvēlēti nejaušā veidā, tad tiek izmantots *nejaušo faktoru līmeņu eksperiments.*

1.7.1. Statistiskais modelis

Lineārais statistiskais modelis:

$$x_{ji} = \mu + \tau_j + \epsilon_{ji}, \text{ kur}$$

- $\tau_j \sim N(0, \sigma_\tau^2)$,
- $\epsilon_{ji} \sim N(0, \sigma^2)$,
- τ_j, ϵ_{ji} ir neatkarīgi.

No neatkarības seko, ka $D(x_{ji}) = \sigma_\tau^2 + \sigma^2$.

1.7.2. Dispersiju analīze

Tāpat kā fiksētu līmeņu gadījumā, var definēt DP, DS, DI, MP, MS, MI. Ir spēka dispersiju sadalījums

$$DP = DS + DI.$$

Hipotēzes:

- $H_0 : \sigma_\tau^2 = 0$,
- $H_1 : \sigma_\tau^2 \neq 0$.

1.2. teorēma.

1. $E(MS) = \sigma^2 + n\sigma_\tau^2$.
2. $E(MI) = \sigma^2$.

Testa statistika:

$$F = \frac{MS}{MI} \sim F(m-1, nm-m).$$

1.solis
Formulēt hipotēzes:

- $H_0 : \sigma_\tau^2 = 0$.
- $H_1 : \sigma_\tau^2 \neq 0$.

2.solis
Noteikt testa statistikas vērtību:

$$F_0 = \frac{MS}{MI} \rightarrow f_0.$$

3.solis
Noteikt kritisko vērtību $f_{\alpha,m-1,nm-m}$.

4.solis
Noraidīt H_0, ja $f_0 > f_{\alpha,m-1,nm-m}$.

2. Randomizēto bloku dizaini

2.1. Blokošana ar vienu blakusparametru - randomizēto pilno bloku dizains

2.1.1. Ievads

Blakusparametri var būt

- nezināmi un nekontrolējami - pret tiem cīnās ar randomizācijas palīzību,
- zināmi (izmērāmi) un nekontrolējami - pret cīnās ar kovariances analīzes palīzību (netiks apskatīta)
- zināmi un kontrolējami - pret tiem cīnās ar blokošanas palīzību.

Blokošana - vairāku eksperimentālo vienību apvienošana tā, lai katrā eksperimentālajā vienībā būtu vienādas kontrolējamo blakusparametru vērtības.

2.1. piemērs. Vienkāršākais blokošanas piemērs - sapārotais t-tests.

Ir jātestē 6 veidu urbji, ir vairāki metāla gabali. Metāla gabali var būt ar dažādām īpašībām - ar dažādām blakusparametru vērtībām. Lai kontrolētu šo blakusparametru iespaidu, ir vēlams visus urbjus testēt uz katra parauga.

Eksperimenta dizainu, kurā visas faktoru vērtības tiek testētas kopā ar visām kontrolējamo parametru vērtībām, sauc par *randomizēto pilno bloku dizainu (randomized complete block design, RCBD).* Katrai eksperimentālajai vienībai dažādiem faktoru līmeņiem atbilstošie apakšeksperimenti tiek veikti nejaušā veidā.

Blokošanu var uzskatīt par nejaušināšanas ierobežojumu.

2.1.2. Statistiskais modelis un hipotēzes

Ir dotas m faktora vērtības un b bloki. Tādējādi ir iegūti šādi mērījumu rezultāti:

$$\begin{cases} x_{11}, ..., x_{m1} - \text{ 1.bloks,} \\ x_{12}, ..., x_{m2} - \text{ 2.bloks,} \\ ..., \\ x_{1b}, ..., x_{mb} - \text{ b.bloks.} \end{cases}$$

Vidējo vērtību modelis:

$$x_{ji} = \mu_{ji} + \epsilon_{ji},$$

kur

- x_{ji} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai parametra vērtībai i-tajā blokā,
- μ_{ji} ir faktora j-tajam līmenim i-tajā blokā atbilstošās ģenerālkopas apakškopas vidējā vērtība,

- ϵ_{ji} ir kļūdas gadījuma lielums, $E(\epsilon_{ji}) = \overline{\epsilon_j} = 0$, parasti uzskata arī, ka $\epsilon_{ji} \sim N(0, \sigma^2)$.

Faktora efektu modelis:

$$x_{ji} = \mu + \tau_j + \beta_i + \epsilon_{ji},$$

kur

- x_{ji} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai parametra vērtībai i-tajā blokā,
- μ ir visas ģenerālkopas vidējā vērtība,
- τ_j - *j-tā faktora efekts*, apmierina vienādību
$$\sum_{j=1}^{m} \tau_j = 0,$$
- β_i - *i-tā bloka efekts*, apmierina vienādību
$$\sum_{i=1}^{b} \beta_i = 0,$$

- ϵ_{ji} ir kļūdas gadījuma lielums, $E(\epsilon_{ji}) = \overline{\epsilon_j} = 0$, parasti uzskata arī, ka $\epsilon_{ji} \sim N(0, \sigma^2)$.

Vidējo vērtību modeļa hipotēzes:

- H_0: $\mu_1 = \mu_2 = ... = \mu_m$,
- H_1: $\mu_k \neq \mu_l$ vismaz vienam pārim k, l.

Efektu modeļa hipotēzes:

- H_0: $\tau_1 = \tau_2 = ... = \tau_m$,
- H_1: $\tau_k \neq 0$ vismaz vienam k.

2.1.3. Dispersiju analīze

Definēsim šādas vidējās vērtības:

- $\overline{x} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} x_{ji}}{mb}$- visas mērījumu kopas vidējā vērtība;
- $\overline{x_{j*}} = \frac{\sum_{i=1}^{n} x_{ji}}{mb}$- j-tā līmeņa vidējā vērtība;

- $\overline{x_{*i}} = \frac{\sum_{j=1}^{m} x_{ji}}{mb}$- i-tā bloka vidējā vērtība;

Definēsim šādas dispersijas:

- $DP = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ji} - \overline{x})^2$- pilnā dispersija;
- $DS = \sum_{j=1}^{m} \sum_{i=1}^{n} (\overline{x}_{j*} - \overline{x})^2 = b \sum_{j=1}^{m} (\overline{x}_{j*} - \overline{x})^2$- starpgrupu (starplīmeņu) dispersija;
- $DB = \sum_{j=1}^{m} \sum_{i=1}^{n} (\overline{x}_{*i} - \overline{x})^2 = m \sum_{i=1}^{n} (\overline{x}_{*i} - \overline{x})^2$- starpbloku dispersija;
-

$$DI = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ji} - \overline{x}_{j*} - \overline{x}_{*i} + \overline{x})^2 =$$

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \Big(x_{ji} - (\overline{x}_{j*} - \overline{x}) - (\overline{x}_{*i} - \overline{x}) - \overline{x} \Big)^2$$

- intragrupu dispersija;

Var pierādīt, ka

$$DP = DS + DB + DI.$$

Definēsim

- $MS = \frac{DS}{m-1}$,
- $MB = \frac{DB}{b-1}$,
- $MI = \frac{DS}{(m-1)(b-1)}$.

2.1. teorēma. Ja $x_{ji} \sim N(\mu + \tau_j, \sigma^2)$, tad

$$E(MI) = \sigma^2,$$

$$E(MS) = \sigma^2 + \frac{b}{m-1} \sum_{j=1}^{m} \tau_j^2,$$

$$E(MB) = \sigma^2 + \frac{m}{b-1} \sum_{i=1}^{b} \beta_i^2.$$

Var pierādīt, ka $\frac{MS}{MI} \sim F(m-1,(m-1)(b-1))$.

Hipotēze H_0 tiek noraidīta, ja $f_0 > f_{\alpha,m-1,(m-1)(b-1)}$.

2.1.4. Modeļa piemērotības pārbaude

Definēsim

- *līmeņa atlikumu:*

$$e_{ji} = x_{ji} - \overline{x}_{j*},$$

- *bloka atlikumu:*

$$e_{ji} = x_{ji} - \overline{x}_{*i},$$

Lai noteiktu, vai modelis ir labs, ir jāpēta atlikumi.

2.1.5. Modeļa aditivitāte

Var redzēt, ka dotais modelis ir aditīvs - nav mijiedarbības starp faktora un bloka parametriem.

Ja ir mijiedarbība starp faktoru un bloku, tad šāds modelis neder, ir jāizmanto citi modeļi, kas satur nelineārus locekļus.

2.2. piemērs. Bloka parametrs un faktora parametrs ir ķīmiskie sastāvi. Ja i-tā bloka viela reaģē ar j-tā līmeņa vielu, tad modelī jābūt loceklim $c\beta_i\tau_j$.

2.1.6. Iztrūkstošo mērījumu aizvietošana

Var gadīties, ka daži mērījumi netiek veikti.

2.3. piemērs. Ja 6 urbji jātestē uz viena metāla parauga, tad tas var sabrukt.

Iztrūkstošās vērtības var atrast, ar mazāko kvadrātu metodi minimizējot DI.

2.1.7. Modeļa parametru novērtēšana ar mazāko kvadrātu metodi

2.2. Blokošana ar diviem blakusparametriem - latīņu kvadrātu dizains

2.2.1. Ievads

Ja ir divi blakusparametri, tad blokošanu var un vajadzētu veikt divos dažādos veidos.

Ir iespējami divi naivi risinājumi:

- veikt divas neatkarīgas blokošanas, katru ar vienu blakusparametru, ja ir p divu blakusparametru līmeņi un p faktora līmeņi, tad eksperimentu skaits būtu $2p^2$, veicot blokošanu pēc viena blakusparametra, otrais blakusparametrs tiktu ignorēts;
- veikt blokošanu pēc abiem parametriem, bloku skaits būtu p^2, mērījumu skaits būtu p^3.

Netriviāls risinājums - apvienot divas bloku sistēmas vienā - veidot blokus, kas satur tādus blakusparametru pārus, kuros katrs no tiem ir sastopams tikai vienu reizi.

Citiem vārdien sakot, veiksim eksperimentu saskaņā ar šādu kārtību:

- izvēlēsimies p pirmā blakusparametra līmeņus,
- izvēlēsimies p otrā blakusparametra līmeņus,
- izvēlēsimies p faktora līmeņus, parasti apzīmē arī ar lielajiem latīņu burtiem (A,B,...),
- izveidosim tabulu, kurā rindas tiek indeksētas ar pirmā blakusparametra p līmeņiem un kolonnas tiek indeksētas ar p otrā blakusparametra līmeņiem,
- aizpildīsim tabulu ar faktora līmeņu simboliem tā, lai katrā rindā un katrā kolonnā katrs simbols būtu vienu reizi.

Šādu eksperimenta dizainu sauc par *latīņu kvadrāta dizainu.* Šādā dizainā blokiem atbilst tabulas rindas un kolonnas.

2.4. piemērs. $P = 5$:

	1	2	3	4	5
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

Tas ir *ciklískā latīņu kvadrāta* piemērs.

2.5. piemērs. Ir doti p lauki un p kultūras. Blakusfaktori - lauks un kultūras tips. Lauki tiek apstrādāti ar p veidu herbicīdiem. Faktors - herbicīda tips. Lai nebūtu jāpārbauda p^3 kombinācijas, konstruējam latīņu kvadrātu un pārbaudam tikai p^2 kombinācijas.

Ir doti p sagataves un p darbinieki (blakusfaktori). No katras sagataves var izgatavot ne vairāk kā p detaļas, ir p veidu detaļu projekti. Lai nebūtu jāpārbauda p^3 kombinācijas, konstruējam latīņu kvadrātu un pārbaudam tikai p^2 kombinācijas.

2.2.2. Statistiskais modelis

Faktora efektu modelis:

$$x_{jik} = \mu + \tau_j + \beta_i + \alpha_k + \epsilon_j,$$

kur

- x_{jik} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai faktora vērtībai i-tajā kolonnā un k-tajā rindā,
- μ ir visas ģenerālkopas vidējā vērtība,
- τ_j - *j-tā faktora efekts*, apmierina vienādību

$$\sum_{j=1}^{p} \tau_j = 0,$$

- β_i - *i-tās kolonnas efekts*, apmierina vienādību

$$\sum_{i=1}^{p} \beta_i = 0,$$

- α_k - *k-tās rindas efekts*, apmierina vienādību

$$\sum_{k=1}^{p} \alpha_k = 0,$$

- ϵ_{jil} ir kļūdas gadījuma lielums, $E(\epsilon_{jik}) = \overline{\epsilon_{jik}} = 0$, parasti uzskata arī, ka $\epsilon_{jik} \sim N(0, \sigma^2)$.

Modelis ir pilnīgi aditīvs - netiek ņemtas vērā mijiedarbības starp parametriem.

Efektu modeļa hipotēzes:

- H_0: $\tau_1 = \tau_2 = ... = \tau_m$,
- H_1: $\tau_k \neq 0$ vismaz vienam k.

2.2.3. Dispersiju analīze

Definēsim šādas vidējās vērtības:

- $\overline{x} = \frac{\sum_{j,i,k} x_{jik}}{p^3}$- visas mērījumu kopas vidējā vērtība;
- $\overline{x}_{j**} = \frac{\sum_{i,k} x_{jik}}{p^2}$- visas j-tā līmeņa vidējā vērtība;
- $\overline{x}_{*i*} = \frac{\sum_{j,k} x_{jik}}{p^2}$- visas i-tās kolonnas vidējā vērtība;
- $\overline{x}_{**k} = \frac{\sum_{j,i} x_{jik}}{p^2}$- visas k-tās rindas vidējā vērtība;

Definēsim šādas dispersijas:

- $DP = \sum_{j,i,k}(x_{jik} - \overline{x})^2$- pilnā dispersija;
- $DS = \sum_{j,i,k}(\overline{x}_{j**} - \overline{x})^2 = p^2 \sum_j (\overline{x}_{j**} - \overline{x})^2$- starpgrupu (starplīmeņu) dispersija;
- $DC = \sum_{j,i,k}(\overline{x}_{*i*} - \overline{x})^2 = p^2 \sum_i (\overline{x}_{*i*} - \overline{x})^2$- starpkolonnu dispersija;
- $DR = \sum_{j,i,k}(\overline{x}_{**k} - \overline{x})^2 = p^2 \sum_k (\overline{x}_{**k} - \overline{x})^2$- starprindu dispersija;

Var pierādīt, ka

$$DP = DS + DS + DC + DR + DI.$$

Definēsim

- $MS = \frac{DS}{p-1}$,
- $MC = \frac{DB}{p-1}$,
- $MR = \frac{DR}{p-1}$,
- $MI = \frac{DI}{(p-2)(p-1)}$.

Var pierādīt, ka $\frac{MS}{MI} \sim F(p-1, (p-2)(p-1))$.

Hipotēze H_0 tiek noraidīta, ja $f_0 > f_{\alpha,p-1,(p-2)(p-1)}$.

2.3. Grieķu-latīņu kvadrātu dizains

2.3.1. Ievads

Ja ir 3 blakusparametri, tad var būt nepieciešams veikt blokošanu 3 veidos.

Turpinot latīņu kvadrātu ideju, var definēt *grieķu-latīņu kvadrātus*:

- izvēlēsimies p pirmā blakusparametra līmeņus,
- izvēlēsimies p otrā blakusparametra līmeņus,
- izvēlēsimies p trešā blakusparametra līmeņus, parasti apzīmē arī ar mazajiem grieķu burtiem,
- izvēlēsimies p faktora līmeņus, parasti apzīmē arī ar lielajiem latīņu burtiem (A,B,...),
- izveidosim tabulu, kurā rindas tiek indeksētas ar pirmā blakusparametra p līmeņiem un kolonnas tiek indeksētas ar p otrā blakusparametra līmeņiem,

- aizpildīsim tabulu ar trešā blakusparametra simboliem (grieķu burtiem) tā, lai katrā rindā un katrā kolonnā katrs simbols būtu vienu reizi,
- aizpildīsim tabulu ar faktora līmeņu simboliem tā, lai katrā rindā un katrā kolonnā katrs simbols būtu vienu reizi un katra grieķu un latīņu burtu pāris ir tieši vienu reizi.

2.6. piemērs.

	1	2	3	4	5
1	Aα	Bγ	Cϵ	Dβ	Eδ
2	Bβ	Cδ	Dα	Eγ	Aϵ
3	Cγ	Dϵ	Eβ	Aδ	Bα
4	Dδ	Eα	Aγ	Bϵ	Cβ
5	Eϵ	Aβ	Bδ	Cα	Dγ

2.3.2. Statistiskais modelis

Faktora efektu modelis:

$$x_{jikl} = \mu + \tau_j + \beta_i + \alpha_k + \Psi_l + \epsilon_j,$$

kur

- x_{jik} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai faktora vērtībai i-tajā kolonnā un k-tajā rindā,
- μ ir visas ģenerālkopas vidējā vērtība,
- τ_j - *j-tā faktora efekts*, apmierina vienādību

$$\sum_{j=1}^{p} \tau_j = 0,$$

- β_i - *i-tās kolonnas efekts*, apmierina vienādību

$$\sum_{i=1}^{p} \beta_i = 0,$$

- α_k - *k-tās rindas efekts,* apmierina vienādību

$$\sum_{k=1}^{p} \alpha_k = 0,$$

- Ψ_l - *l-tā grieķu parametra efekts,* apmierina vienādību

$$\sum_{l=1}^{p} \Psi_l = 0,$$

- ϵ_{jil} ir kļūdas gadījuma lielums, $E(\epsilon_{jik}) = \overline{\epsilon_{jik}} = 0$, parasti uzskata arī, ka $\epsilon_{jik} \sim N(0, \sigma^2)$.

Modelis ir pilnīgi aditīvs - netiek ņemtas vērā mijiedarbības starp parametriem.

2.4. Balansētie nepilno bloku dizaini

2.4.1. Ievads

Var gadīties, ka randomizētā pilnā bloku dizaina realizācija nav iespēja, jo nav iespējams katrā blokā veikt visus nepieciešamos eksperimentus ar visiem faktora līmeņiem.

2.7. piemērs. Ir jātestē 4 urbji uz metāla gabaliņiem. Uz katra metāla gabaliņa ir iespējams veikt tikai 3 urbumus.

Ir vēlams izstrādāt tādus blokošanas veidus, lai katrā blokā būtu tikai daļa no faktora līmeņiem - *randomizētos nepilno bloku dizainus.*

Par *balansēto nepilno bloku dizainu (balanced incomplete block design, BIBD)* sauc dizainu, kurā katrs faktoru līmeņu pāris vienā blokā ir sastopams vienādu skaitu reižu. Tas ir nepieciešams, tad, ja visu faktoru līmeņu salīdzināšanas ir vienādi svarīgas. Pretējā gadījumā statistiskie modeļi būs mazāk dabiski, statistiskā analīze būs grūtāka, kļūdas nebūs vienādas.

2.4.2. BIBD diskrēti-matemātiskie aspekti

Naivais risinājums - ja ir m faktoru līmeņi un blokā var būt k ($k < m$) faktoru vērtības, tad izveidosim C_m^k blokus, katrā blokā būs viena no iespējamajām m faktoru kopas apakškopām ar k elementiem. Katrs faktoru pāris kopā būs tieši C_{m-2}^{k-2} blokos.

2.8. piemērs. 4 urbji līmeņi un 3 urbumi vienā metāla gabalā.

Eksistē arī netriviāli risinājumi, kad bloku skaits ir mazāks nekā C_m^k.

BIBD sauc par (v, b, r, k, λ)-dizainu, ja

- faktora līmeņu skaits ir v;
- bloku skaits ir b;

- katrs faktora līmenis ir tieši r blokos;
- katrā blokā ir tieši k faktora līmeņi;
- katrs faktora līmeņu pāris kopā ir tieši λ blokos.

"Nepilns" nozīmē, ka $v > k$. "Balansēts" nozīmē, ka λ ir definēts.

Par (v, b, r, k, λ)-dizaina *incidences matricu* A sauc bināru $v \times b$ matricu, kurā

- rindas tiek indeksētas ar faktora līmeņiem;
- kolonnas tiek indeksētas ar blokiem;
- rūtiņā (i, j) ir 1 $\iff$ i-tais līmenis ir j-tajā blokā.

2.9. piemērs. Incidences matricas piemērs.

BIBD parametri nav neatkarīgi, tos saista vismaz šādas sakarības:

- $bk = vr$ (jāuzzīmē divdaļīgs grafs, kuram daļas ir līmeņi un bloki, šķautne nozīmē līmeņa iekļaušanu blokā, jāskaita šķautnes pie katras daļas);
- $\lambda(v-1) = r(k-1)$ (jāuzzīmē divdaļīgs grafs, kuram daļas ir līmeņu pāri un bloki, šķautne nozīmē pāra iekļaušanu blokā, jāskaita šķautnes pie abām daļām, jāizmanto iepriekšējā īpašība);
- $AA^T = (r-\lambda)\mathcal{E}_v + \lambda\mathcal{J}_v$ ($\mathcal{E}_v$ - vienības matrica, $\mathcal{J}_v$ - vieninieku matrica);
- *Fišera nevienādība:* $b \geq v$ (pierāda izmantojot incidences matricas).

BIBD dizainu sauc par *simetrisku*, ja $b = v$ ($\implies r = k$). Apzīmē kā (v, k, λ). Tādējādi simetriskajam dizainam ir mazākais iespējamais bloku skaits ar dotu faktora līmeņu skaitu.

Simetriskie (v, k, λ)-dizaini apmierina papildus nosacījumus (*Bruck-Ryser-Chowla teorēma*):

- ja v ir pāra skaitlis, tad $k - \lambda$ ir vesela skaitļa kvadrāts;

- ja v ir nepāra skaitlis, tad vienādojumam
$$x^2 = (k - \lambda)y^2 + (-1)^{\frac{v-1}{2}} \lambda z^2$$
eksistē netriviāls atrisinājums veselos skaitļos.

Klasiskas BIBD dizainu sērijas:

- *afīnās plaknes* - $(n^2, n, 1)$,
- *projektīvās plaknes* - $(n^2 + n + 1, n + 1, 1)$,
- *Steinera trijnieku sistēmas* - $(v, 3, 1)$, ja $v \in \{1, 3\} (\mathrm{mod}\ 6)$, $v \geq 3$,
- *Adamāra dizaini* - $(4n + 3, 2n + 1, n)$,
- *unitālie dizaini* - $(n^3 + 1, n + 1, 1)$.

2.10. piemērs. Mazākā projektīvā plakne, $n = 2$, *Fano plakne:*

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

2.4.3. Statistiskais modelis

BIBD modelis:

$$x_{ji} = \mu + \tau_j + \beta_i + \epsilon_{ji},$$

kur

- x_{ji} ir eksperimenta rezultāta gadījuma lielums, kas atbilst j-tajai faktora vērtībai i-tajā blokā,
- μ ir visas ģenerālkopas vidējā vērtība,

- τ_j - *j-tā faktora efekts*, apmierina vienādību

$$\sum_{j=1}^{m} \tau_j = 0,$$

- β_i - *i-tā bloka efekts*, apmierina vienādību

$$\sum_{i=1}^{b} \beta_i = 0,$$

- ϵ_{ji} ir kļūdas gadījuma lielums, $E(\epsilon_{ji}) = \overline{\epsilon_j} = 0$, parasti uzskata arī, ka $\epsilon_{ji} \sim N(0, \sigma^2)$.

BIBD modeļa hipotēzes:

- H_0: $\tau_1 = \tau_2 = ... = \tau_m$,
- H_1: $\tau_k \neq 0$ vismaz vienam k.

2.4.4. Dispersiju analīze

BIBD gadījumā dispersiju analīze ir grūtāka.

Ir spēkā formula

$$DP = DS + DB + DI,$$

kur

- $$DB = \frac{1}{k}\sum_{i=1}^{b}\Big(\sum_{i\text{-tais bloks}} x_{*i}\Big)^2 - \frac{1}{bk}\Big(\sum_{i,j} x_{ji}\Big)^2,$$
  brīvības pakāpju skaits - $b-1$,
- $DS = \frac{k}{\lambda m}\sum_{j=1}^{m} Q_j^2$, kur
  $$Q_j = (\sum_i x_{ji}) - \frac{1}{k}\sum_{i=1}^{b} n_{ji}(\sum_j x_{ji}),$$
  n_{ji} ir 1/0, ja j-tais līmenis ir/nav i-tajā blokā, brīvības pakāpju skaits $m-1$.

DP brīvības pakāpju skaits ir $bk - 1$.

Var pierādīt, ka $\frac{MS}{MI} \sim F(m-1, bk-m-b+1)$.

Hipotēze H_0 tiek noraidīta, ja $f_0 > f_{\alpha,m-1,bk-m-b+1}$.

3. 3.mājasdarbs

3.1 Cementa stiprība tiek testēta 4 dažādām maisījuma sagatavošanas tehnoloģijām. Tiek iegūti šādi mērījumi:

$$\text{1.tehnoloģija} - 4519, 4493, 4495, 4512$$

$$\text{2.tehnoloģija} - 4453, 4448, 4460, 4441$$

$$\text{3.tehnoloģija} - 4552, 4545, 4557, 4547$$

$$\text{4.tehnoloģija} - 4398, 4405, 4411, 4402.$$

(a) Pārbaudiet hipotēzi, ka tehnoloģijas ietekmē stiprību, ar $\alpha = 0.05$

(b) veiciet normālā sadalījuma pārbaudi atlikumiem, noskaidrojiet, vai ir izņēmumi.