Modelowanie obserwacji w doświadczeniach nauk przyrodniczych

Stanisław Mejza

Uniwersytet Przyrodniczy w Poznaniu,
Katedra Metod matematycznych i Statystycznych
Wojska Polskiego 28, PL-60-637 Poznań,
e-mail: smejza@up.poznan.pl

NATURA

(*POPULACJA*)

AKSJOMAT

W NATURZE (PRZYRODZIE – POPULACJI) wszystkie zjawiska, procesy przebiegają według jednej funkcji (reguły, zasady, relacji).

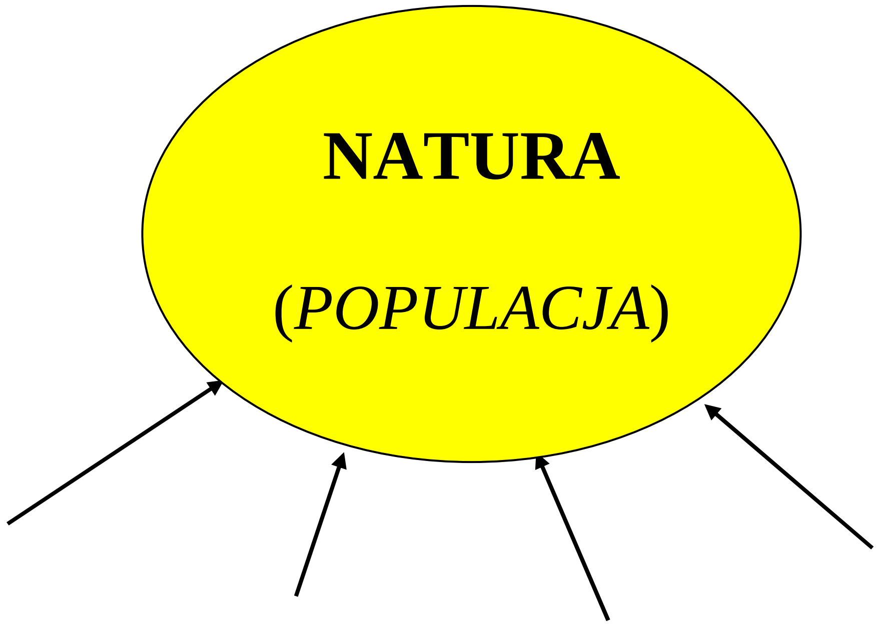
CEL BADAŃ - ROZPOZNAĆ FUNKCJĘ (RELACJĘ)

I. OPIS (Naturalny bieg NATURY)

Fakty: 1) co rozpoznać ?

2) czym dysponujemy ?

3) co obserwujemy ?



Czynniki doświadczalne

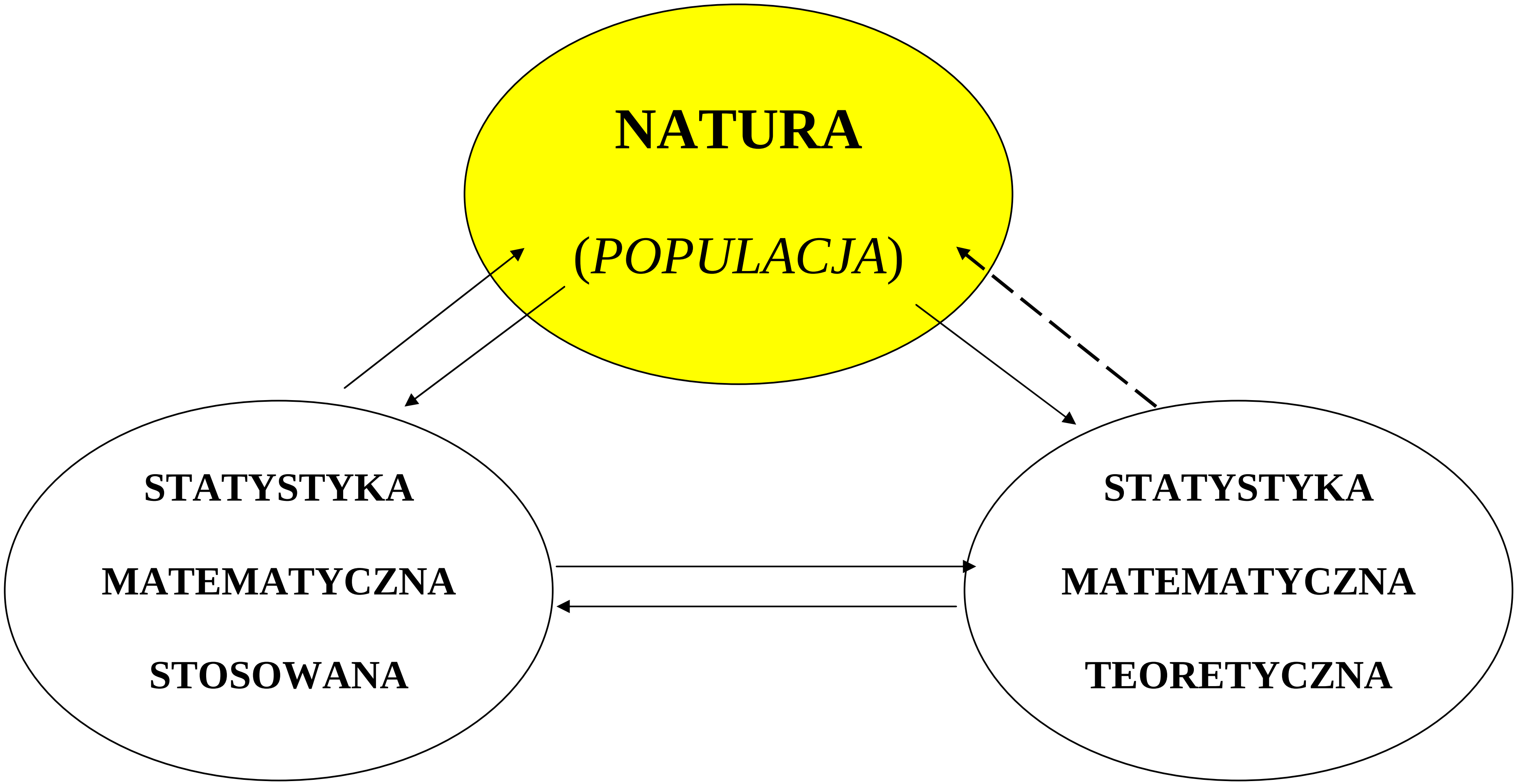
II. **WNIOSKOWANIE - Badanie reakcji NATURY**

COMPARATIVE EXPERIMENTS -
badania (doświadczenia) porównawcze

1. Design of Comparative Experiments: Meaning?

$\Rightarrow$ **NOT** experiments to determine the exact value of G.

$\Rightarrow$ **BUT** experiments to find out if A is better than B, and, if so, by how much.



Nauki rolnicze !!!!!!!!!!

1. Introduction

The experiment is an important tool of research in natural sciences. Hence, planning, modelling and inference problems are of fundamental importance for every experimenter using experiments in his research work.

Approaches to the model building of observations.

- assumption (expected value +dispersion structure) – **sampling ?**

- **derivation – sampling methods**

  model assumed
  derived model

In the first approach we assume a'priori a form of the linear model, usually before performing the experiment.

The linear model and its dispersion structure are assumed to be independent from the type of experiment, on the structure of the experimental material and on the sampling methods.

Sometimes, some additional assumptions concerning dispersion structure (correlation, auto correlation, etc.) are added.

The problem is how to check these assumptions.

Derivation

The model is strictly connected with a given experiment i.e., with the structure of its experimental material and with the method of assigning treatments to the units (sampling).

This is given by the so-called scheme of randomization.

At the beginning let us consider the factors which have an influence on the value of the observed data (called also observed response, observed yield).

Observed response is a sum of three components:

- a conceptual response connected with an experimental unit,
- a pure effect due to treatment (combinations),
- a technical effect connected with measurements.

Additionally: *additivity among these three components is assumed*.

Conceptual response

Note that every unit (plot) possesses some kind of fertility which gives some yield in the case when treatments do not occur on a unit and in the case in which no treatments have an effect on the yield. This yield will be called zero yield (conceptual response).

Pure effect

The increase (or decrease) in zero yield due to the treatment used on the experimental unit will be called pure effect (due to treatment). Usually the sum of zero yield (conceptual response) and pure effect due to treatment is called the pure yield (pure response) and is often the base of the statistical analysis.

Technical error

Let us note that when observing the response of the unit in reality, any observation may be affected by a "technical error", an error due to some technical inaccuracy in performing the experiment and due to some error connected with measurements of the response (data). This error is also called measurement error (cf. Neyman et al., 1935).

The paper deals with the problems connected with model building for popular types of designed experiments. The one- factor experiments carried out in design with one or more blocking systems are taken into account only, i.e. a block design and nested block design.

The several different schemes of randomization’s (sampling) for the designs mentioned above will be presented. Finally, the consequences of different schemes of randomization to the linear model of observation is discussed.

Experiment (R, Ω), R - experimental material structure, Ω - treatment structure.

The starting point of statistical considerations is the theoretical treatment plan of the experiment. In this plan, Ω; we take into account all the experimenter's suggestions concerning the statistical properties of design and the experimental conditions available in R.

It means that plan Ω will not be chosen at random.

Ω - statistical properties:
estimability, testability,
balance (variance balance, efficiency balance),
optimality,

etc……..

The basic problem worked out here, is the way of assigning plan Ω to a given experimental material. This is defined by the scheme of randomization. It describes how to assign the theoretical units of plan Ω (with their treatments), to the experimental plots. In our considerations the treatments will not be randomised.

Suppose that the randomization is performed as described by Nelder (1954) by randomly permuting, for example for block design, blocks within their total area and by randomly permuting units within blocks.

(see: **Caliński & Kageyama, 2000, 2002)**

It will be assumed that the treatments under consideration are homogeneous (or additive) in the sense that the variation of the response among the available experimental units does not depend on the treatment received (cf. Kempthorne, 1952, Nelder, 1965).

Modelling (Fisher's principles)

Observed yield = (block structure effects)
+ (treatment effects)
+ (unit errors)
+ (technical errors)

Block designed experiment

Example

B_1	B_2	B_3	B_4	B_5

B_1	B_4	B_3	B_2	B_4	B_3

Case A.

Population of experimental units (potential population of units):

- b blocks of sizes $K_1, K_2, \ldots, K_b$,

Plan Ω

- b blocks of sizes $k_1, k_2, \ldots, k_b$ units ($k_i \leq K_i$, $i=1,2,\ldots,b$)

Randomizations:

-- blocks are not randomised. It means that because of some reasons the blocks of the plan Ω are assigned to the experimental ones in an arbitrary, non random way;

-- units within each experimental block are assigned at random to the units of plan Ω;

-- all the *b* randomizations are independent;

Formal derivation:

Let d^i_{jt} is equal to 1, if within the *i*-th block the *t*-th unit of experimental material is assigned to the *j*-th unit of plan Ω; otherwise d^i_{jt} is equal to 0.

Distribution of d^i_{jt} (discrete).

Conceptual response of the *j*-th unit of the *i*-th block after randomization: $Y_{ij} = \Sigma_t d^i_{jt} m_{it}$, then using equality $m_{it} = m_{i.} + (m_{it} - m_{i.})$, $\rightarrow$ $Y_{ij} = \alpha_i + \varepsilon_{ij}$, $j= 1,2,\ldots, k_i$, $\mu_i = m_{i.}$, denotes the effect of the *i*-th block,

$\varepsilon_{ij} = \Sigma_t d^i_{jt}(m_{it} - m_{i.})$, denotes the random effect of the *j*-th unit in the *i*-th block, $\sigma_i^2 = K_i^{-1}\Sigma_t (m_{it} - m_{i.})^2$ - denotes the unit variance.

Dispersion structure is given below.

This scheme of randomization and the additivity assumptions generate the following linear model for observed yield y_{it} obtained for the t-th treatment in the i-th block, (t=1,2,…v):

$$y_{it} = \mu + \alpha_i + \tau_t + e_{it} + \varepsilon_{it}, \qquad E(y_{ijt}) = \mu + \alpha_i + \tau_t, \qquad (1)$$

$$Cov(y_{it}, y_{i't'}) = \begin{cases} \sigma_i^2 + \sigma^2, & i = i',\ t = t', \\ -(K_i - 1)^{-1}\sigma_i^2, & i = i',\ t \neq t', \qquad \sigma_i^2 = K^{-1}\Sigma_t (m_{it} - m_{i.})^2 \\ 0, & otherwise. \end{cases}$$

The parameters in the model (1) are as follows: μ - the mean of experiment, α_i - the effect of the i-th block, τ_t - the effect of the t-th treatment, e_{it} - a plot error, ε_{it} - a technical error, σ_i^2 - a plot variance of the i-th block, σ^2 - a technical error variance.

Particular cases:

1) $k_i = K_i, i = 1,2,...,b.$

2) $K_{ij} \to \infty$, the observations can be considered uncorrelated.

3) $K_{ij} \to \infty$, $\sigma_i^2 = \sigma_e^2$ for all i,

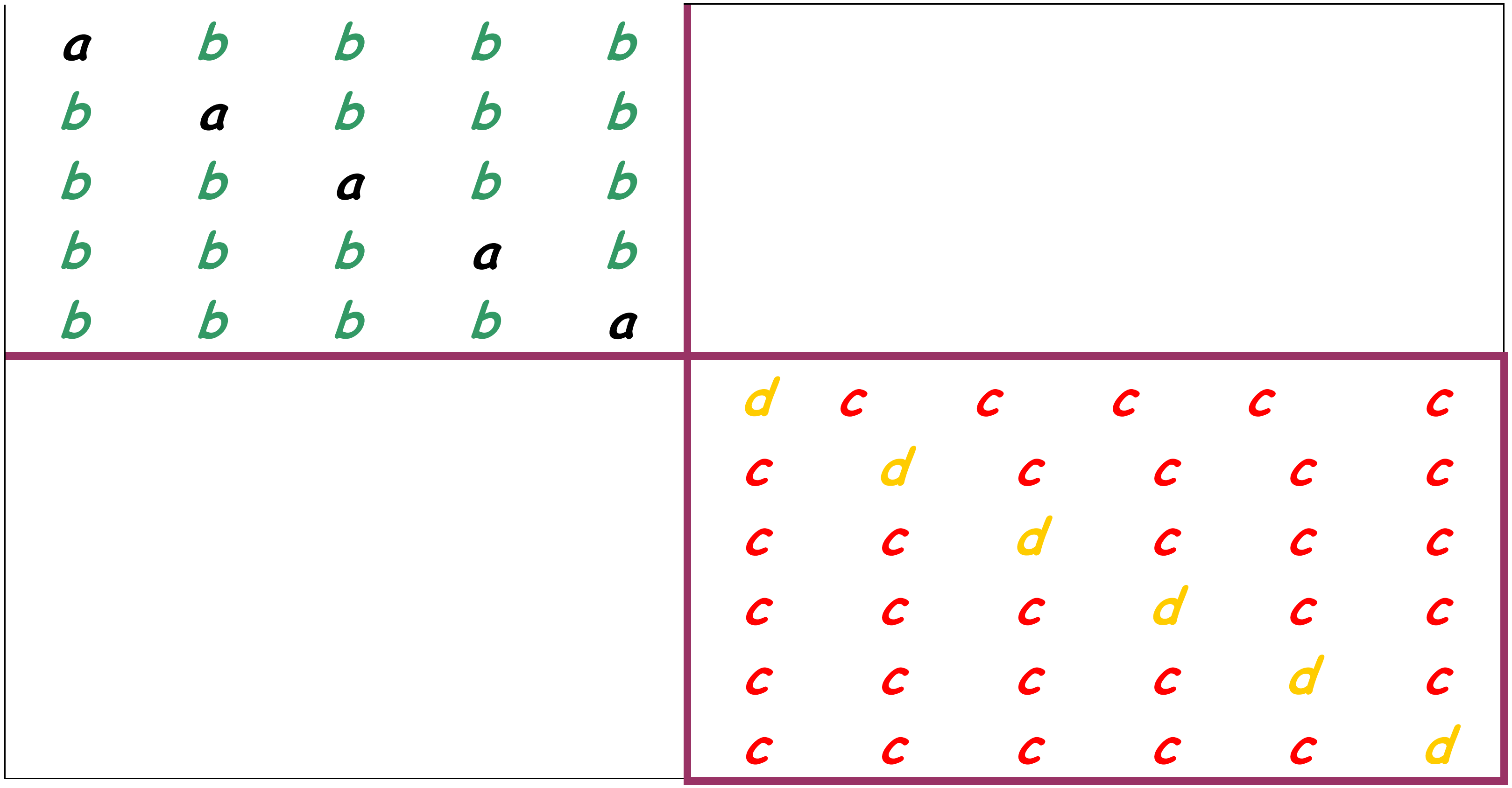
leads to often used fixed linear model for experiment carried out in BD. These assumptions are restrictive and in a practice are seldom fulfilled.

Example

B_1	B_2	B_3	B_4	B_5

B_1	B_4	B_3	B_2	B_4	B_3

Dispersion structure



Fixed ?

Observed yield = (block structure effects)
+ (treatment effects)
+ (unit errors)
+ (technical errors)

Randomization?

The assumptions are restrictive and in a practice are seldom fulfilled.

***Case B*.**

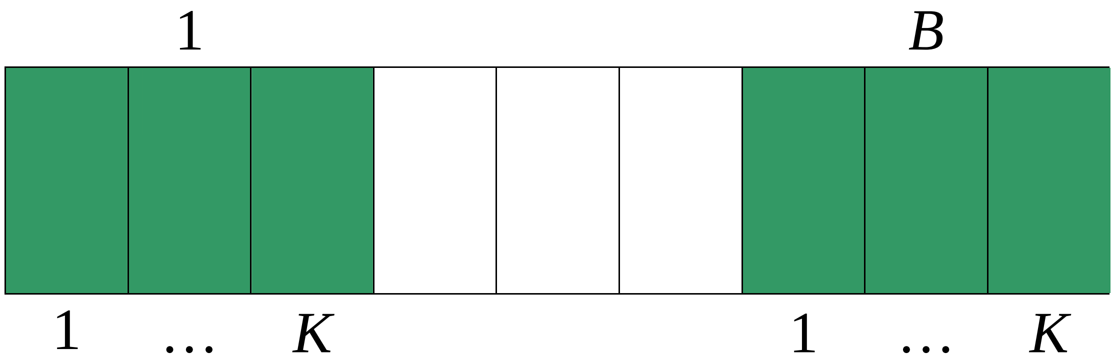
Population of experimental units (potential population of units):

- B blocks of sizes K,

Plan Ω :

b blocks of sizes $k_1, k_2,...,k_b$ units ($k_i \le K$, $i=1,2,\ldots,b$)

Structure of experimental units



Randomization:

1) Randomization of blocks,
2) Independent randomization of units within blocks.

The linear model:

$$E(y_{it}) = \mu + \tau_t, \qquad (2)$$

$$\mathrm{Cov}(y_{it}, y_{i't'}) = \begin{cases} \sigma_B^2 + \sigma_e^2 + \sigma^2, & i = i', t = t',, \\ \sigma_B^2 - (K-1)^{-1}\sigma_e^2, & i = i', t \neq t' \\ -(B-1)^{-1}\sigma_B^2, & \textit{otherwise}, \end{cases},$$

where σ_B^2 denotes the block variance within the *i*-th block, σ_e^2 denotes the plot variance within the *i*-th block.

Particular cases:

1) $b = B$, $k_i = k = K$;
2) $B \to \infty$, $K \to \infty$

Example

B_1	B_2	B_3	B_4	B_5

B_1	B_4	B_3	B_2	B_4	B_3

Dispersion structure

a	b	b	b	b	d	d	d	d	d	d
b	a	b	b	b	d	d	d	d	d	d
b	b	a	b	b	d	d	d	d	d	d
b	b	b	a	b	d	d	d	d	d	d
b	b	b	b	a	d	d	d	d	d	d
d	d	d	d	d	a	b	b	b	b	b
d	d	d	d	d	b	a	b	b	b	b
d	d	d	d	d	b	b	a	b	b	b
d	d	d	d	d	b	b	b	a	b	b
d	d	d	d	d	b	b	b	b	a	b
d	d	d	d	d	b	b	b	b	b	a

Mixed ?

Observed yield = (block structure effects)
+ (treatment effects)
+ (unit errors)
+ (technical errors)

Randomization?

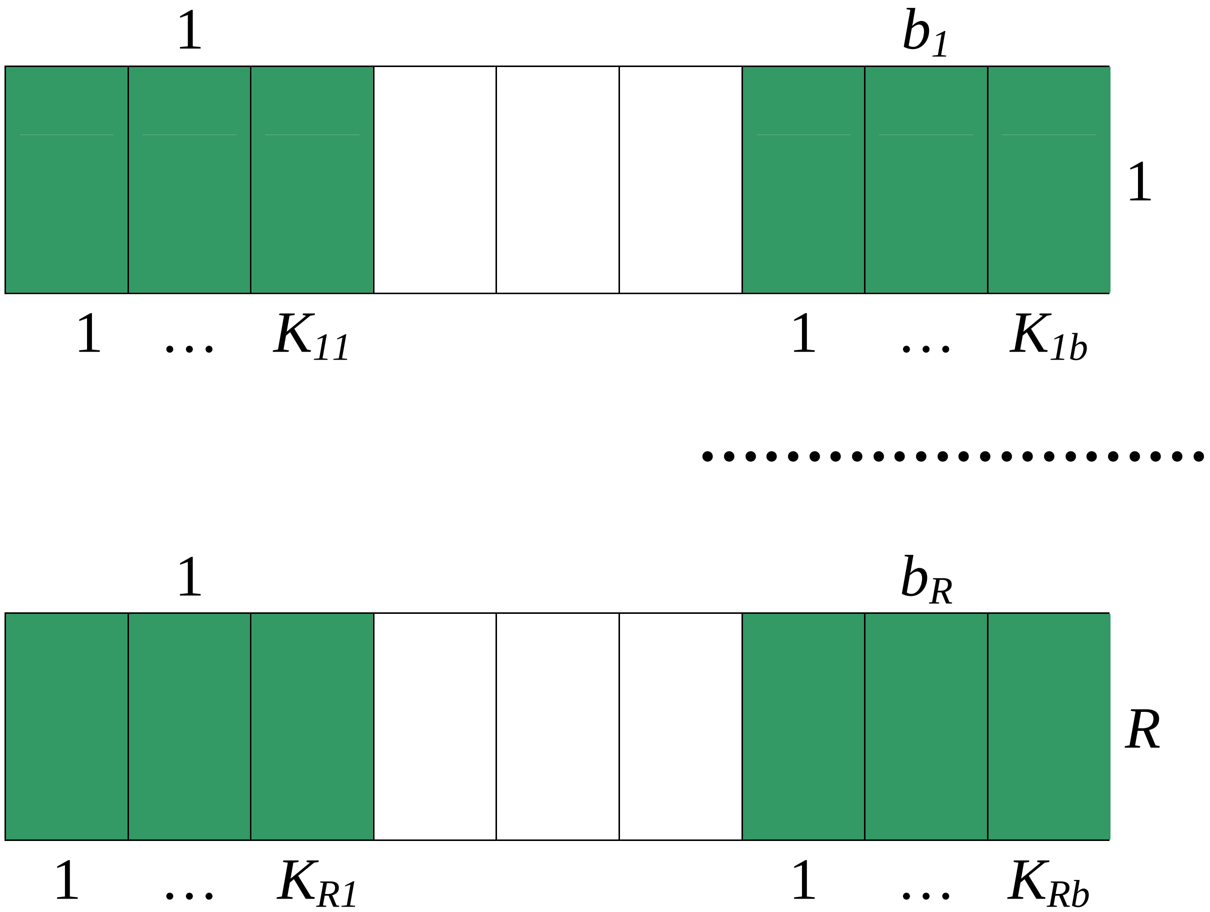
Random?

Observed yield = (block structure effects)
+ (treatment effects)
+ (unit errors)
+ (technical errors)

Randomization?

Nested block design

Structure of potential experimental units - **Case A**



Case A.

Population of experimental units (potential population of units):

- R superblocks,
- superblocks of sizes $b_1,\ b_2,\ldots,\ b_R$ blocks,
- K_{ij} - number of units in the j-th block of the i-th superblock,
- $j = 1,2,\ldots,\ b_i,\ \ i = 1,2\ \ldots,\ R.$

Plan Ω

- R superblocks,
- b_i blocks,
- $k_{ij}\ \ (\leq K_{ij}),\ \ i=1,2,\ldots,b_i$

Randomization:

1) Superblocks – not randomized,
2) Blocks – not randomized
3) Units - independent randomization of units within blocks.

Linear model for observed yield y_{ijt} obtained for the t-th treatment in the j-th block of the i-th superblock:

$$y_{ijt} = \mu + \alpha_i + \beta_{ij} + \tau_t + e_{ijt} + \varepsilon_{ijt}, \quad E(y_{ijt}) = \mu + \alpha_i + \beta_{ij} + \tau_t, \qquad (3)$$

$$\mathrm{Cov}(y_{ijt}, y_{i'j't'}) = \begin{cases} \sigma_{ij}^2 + \sigma^2, & i = i',\ j = j',\ t = t', \\ -(K_{ij} - 1)^{-1}\sigma_{ij}^2, & i = i',\ j = j',\ t \neq t', \\ 0, & \textit{otherwise.} \end{cases}$$

The parameters in the model (3) are as follows:

μ - the mean of experiment,

α_i - the effect of the *i*-th superblock,

β_{ij} - the effect of the *j*-th block of the *i*-th superblock,

τ_t - the effect of the *t*-th treatment,

e_{ijt} - a plot error,

ε_{ijt} - a technical error,

σ_{ij}^2 - a plot variance of the *j*-th block in the *i*-th superblock,

σ^2 - a technical error variance.

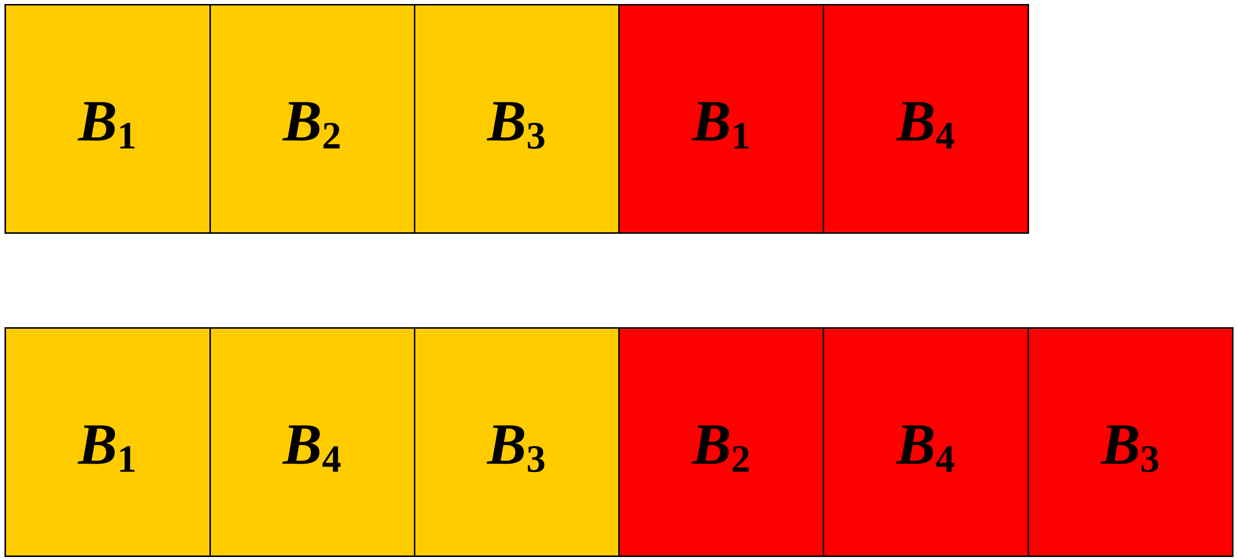
Particular cases:

1) $K_{ij} \to \infty$, the observations can be considered uncorrelated,

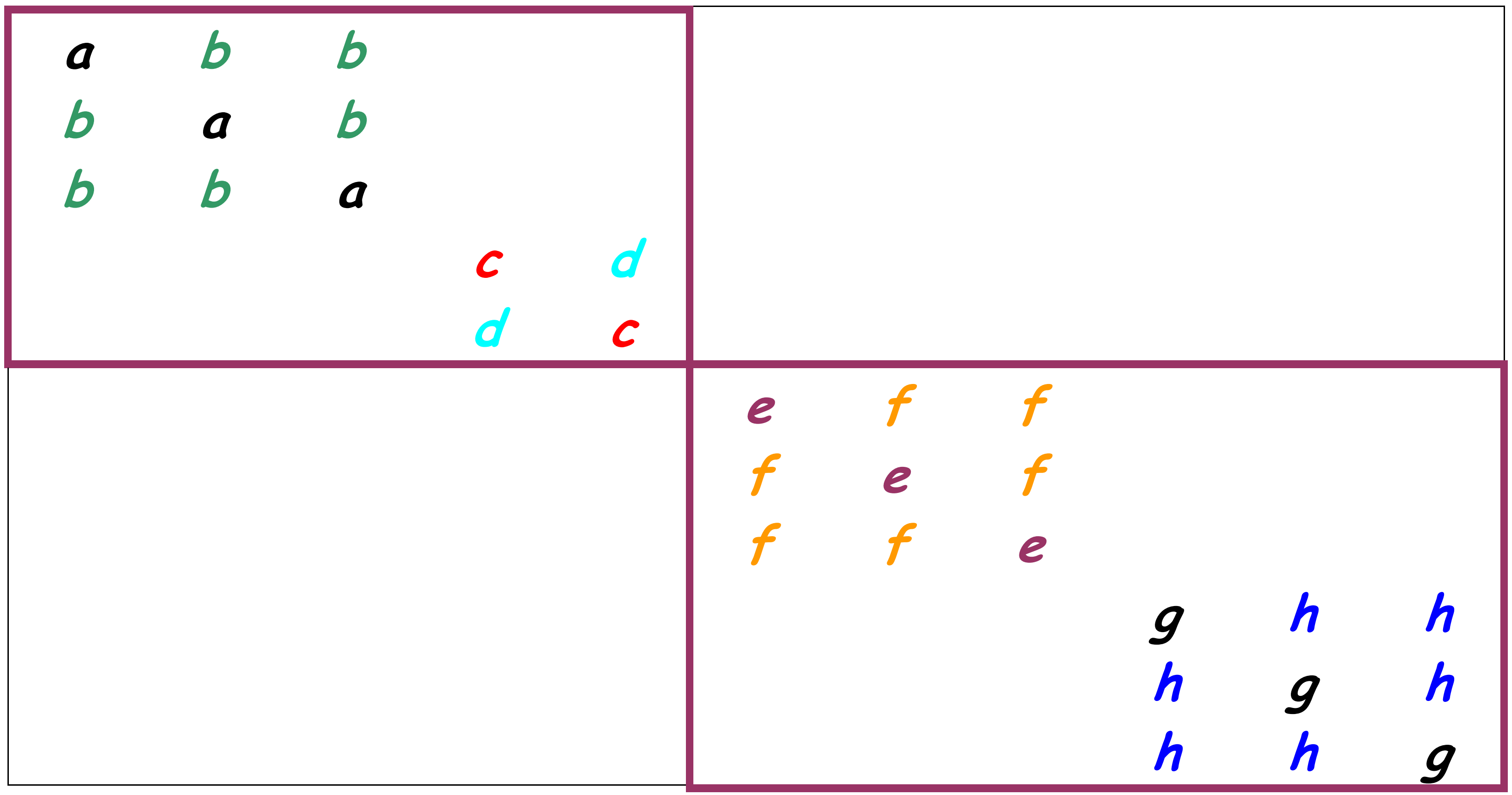
2) $\sigma_{ij}^2 = \sigma_e^2$ for all *i*, *j*, leads to often used fixed linear model for experiment carried out in NBD.

These assumptions are restrictive and in a practice are seldom fulfilled.

Example:



Dispersion structure



Case B.

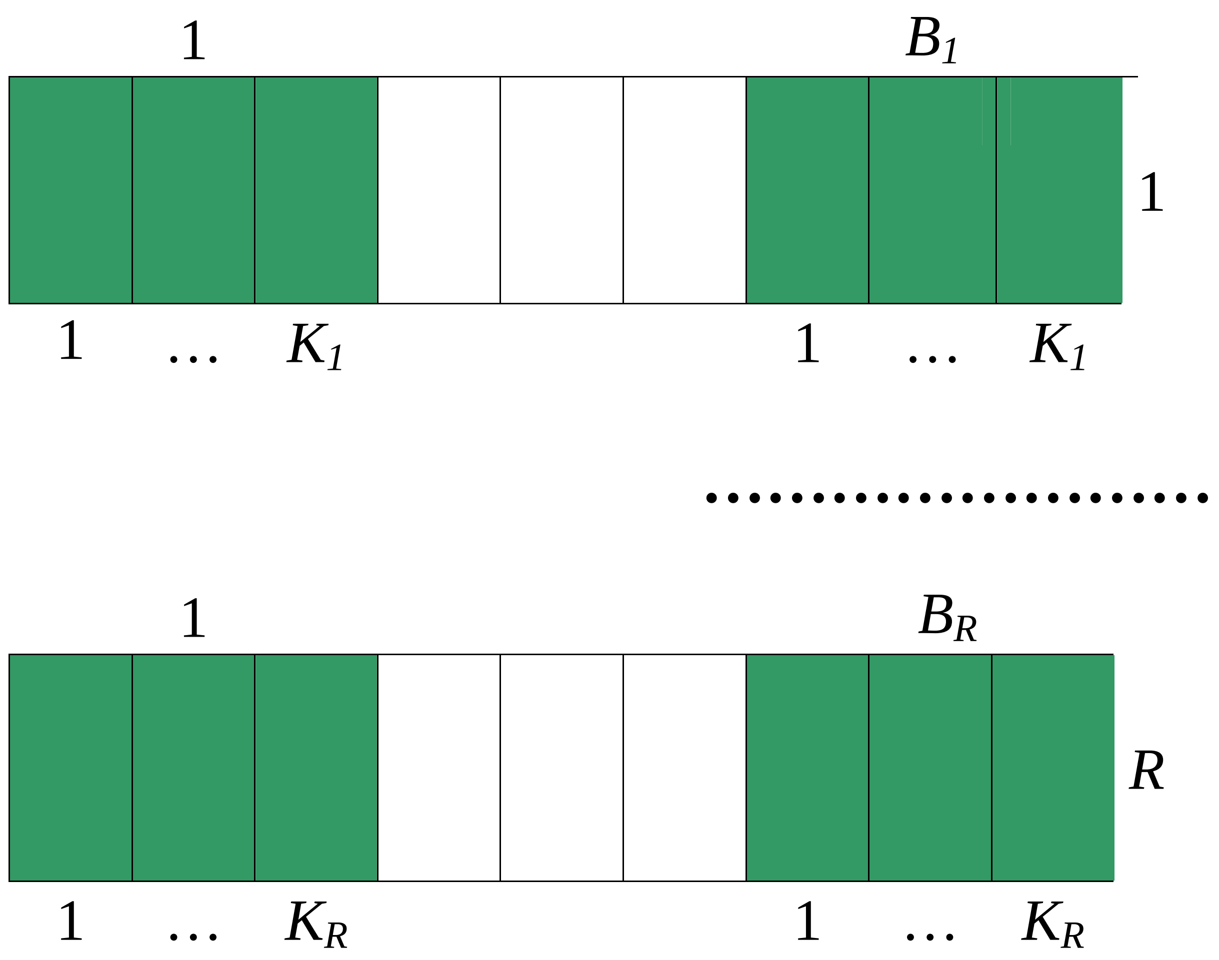
Population of experimental units (potential population of units):

- R superblocks,
- superblocks of sizes $B_1, \; B_2, \ldots, \; B_R$ blocks,
- K_i - number of units in the j-th block of the i-th superblock,
- $j = 1,2,\ldots, K_i, \; i = 1,2 \ldots, R$.

Plan Ω

- R superblocks,
- $b_i \quad (\leq B_i)$, blocks,
- $k_i \; (\leq K_i), \; i=1,2,\ldots,R$

Structure of potential experimental units



Randomization:

1) Superblocks – not randomized,
2) Blocks – randomized
3) Units - independent randomization of units within blocks.

The linear model

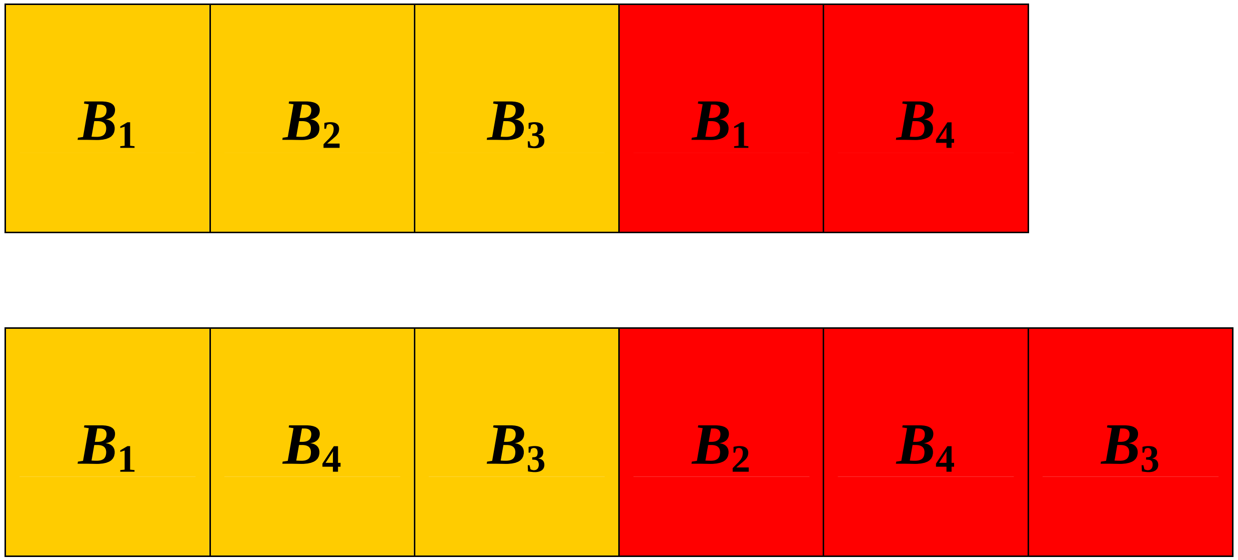
$$E(y_{ijt}) = \mu + \alpha_i + \tau_t, \qquad (4)$$

$$\mathrm{Cov}(y_{ijt}, y_{i'j't'}) = \begin{cases} \sigma_{Bi}^2 + \sigma_{oi}^2 + \sigma^2, & i = i',\ j = j', t = t', \\ \sigma_{Bi}^2 - (K_i - 1)^{-1}\sigma_{oi}^2, & i = i',\ j = j', t \neq t' \\ -(B_i - 1)^{-1}\sigma_{Bi}^2, & i = i',\ j \neq j, \\ 0, & \textit{otherwise}, \end{cases}$$

- σ_{Bi}^2 denotes the block variance within the *i*-th superblock,

- σ_{oi}^2 denotes the plot variance within the *i*-th superblock, $i = 1,2,\ldots, R$.

Example



Dispersion structure

a	b	b	c	c						
b	a	b	c	c						
b	b	a	c	c						
c	c	c	a	b						
c	c	c	b	a						
					d	e	e	f	f	f
					e	d	e	f	f	f
					e	e	d	f	f	f
					f	f	f	d	e	e
					f	f	f	e	d	e
					f	f	f	e	e	d

Particular cases:

1) $B_i \rightarrow \infty$, $K_i \rightarrow \infty$, the observations can be considered uncorrelated,

2) $\sigma^2_{Bi} = \sigma^2_B$ and $\sigma^2_{oi} = \sigma^2_o$. for all i, leads to often used mixed linear model for experiment carried out in NBD.

It is worth noting here that the assumptions about the equality of the variances are extremely restrictive. This is the main disadvantage of this model.

***Case C*.**

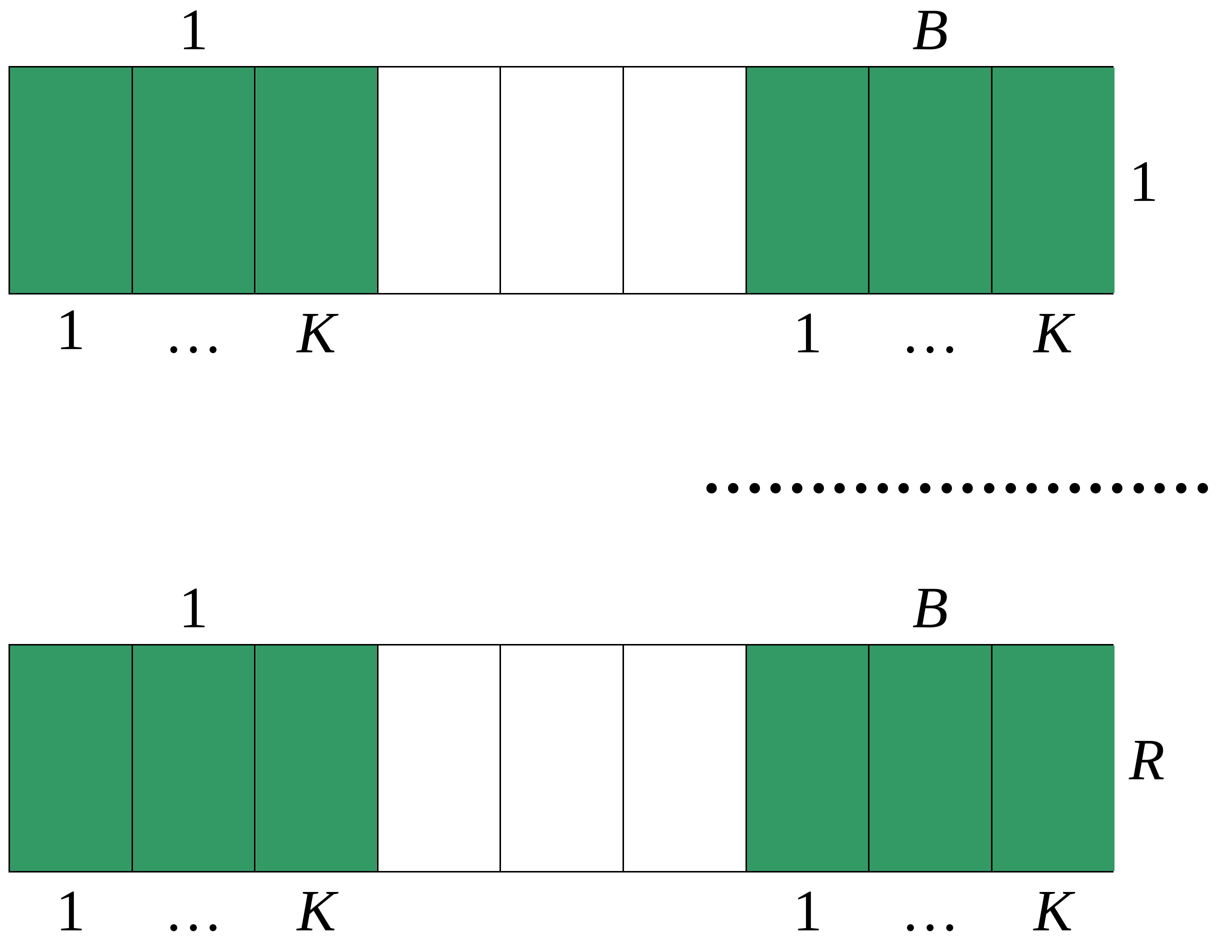
Population of experimental units (potential population of units):

- R superblocks,
- superblocks of sizes B , blocks,
- K - number of units in the j-th block of the i-th superblock,

Plan Ω

- r superblocks,
- b_i $(\leq B)$, blocks,
- $k_{i1}, k_{i2}, \ldots, k_{ib_i}$ $(\leq K)$

Structure of experimental units



Randomization

Now, threefold randomization is performed i.e.

1) the randomization of the superblocks,
2) randomization of the blocks within the superblocks,
3) and randomization of the units (plots) within the blocks.

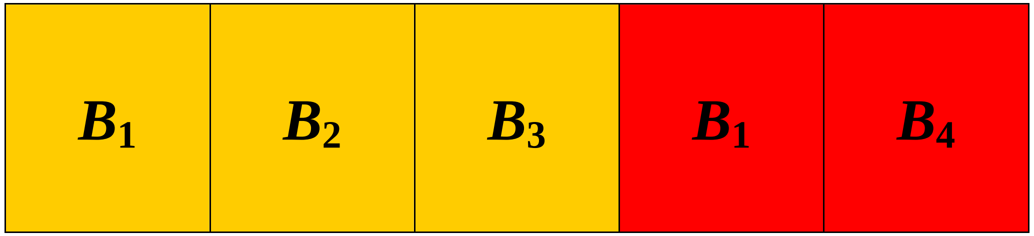
The obtained model has a form:

$$E(y_{ijt}) = \mu + \tau_t , \tag{5}$$

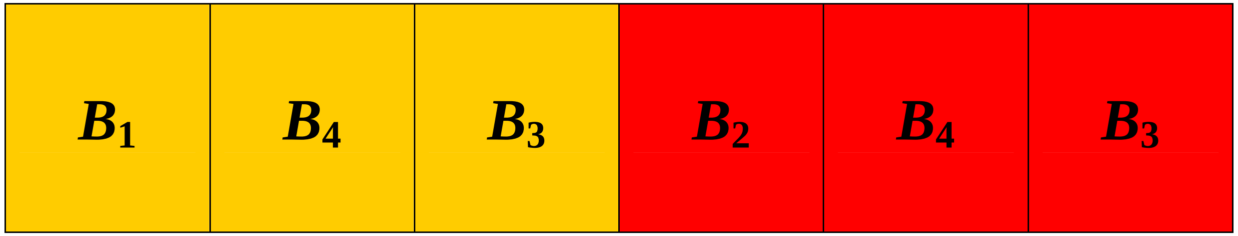
$$\mathrm{Cov}(y_{ijt}, y_{i'j't'}) = \begin{cases} \sigma_a^2 + \sigma_b^2 + \sigma_e^2 + \sigma^2, & i = i',\ j = j',\ t = t', \\ \sigma_a^2 + \sigma_b^2 - (K \ \ -1)^{-1}\sigma_e^2, & i = i',\ j = j',\ t \neq t', \\ \sigma_a^2 - (B \ \ -1)^{-1}\sigma_b^2, & i = i',\ j \neq j, \\ -(R \ \ -1)^{-1}\sigma_a^2, & i \neq i', \end{cases}$$

- σ_a^2 denotes the superblock variance,
- σ_b^2 denotes the block variance,
- σ_e^2 denotes the unit variance.

Example



333



Dispersion structure

a	b	b	c	c	d	d	d	d	d	d
b	a	b	c	c	d	d	d	d	d	d
b	b	a	c	c	d	d	d	d	d	d
c	c	c	a	b	d	d	d	d	d	d
c	c	c	b	a	d	d	d	d	d	d
d	d	d	d	d	a	b	b	c	c	c
d	d	d	d	d	b	a	b	c	c	c
d	d	d	d	d	b	b	a	c	c	c
d	d	d	d	d	c	c	c	a	b	b
d	d	d	d	d	c	c	c	b	a	b
d	d	d	d	d	c	c	c	b	b	a

Particular cases:

1) $R \to \infty$, $B \to \infty$ and $K \to \infty$.

We obtain the classic linear mixed model of the NBDs under this assumptions

2) $r=R$, $b_i=B$, $k_{i1}=k_{i2}=\ldots=k_{ib_i}=k$,

Usually the whole population of units takes part in an experiment.

This makes the model useful to practice.

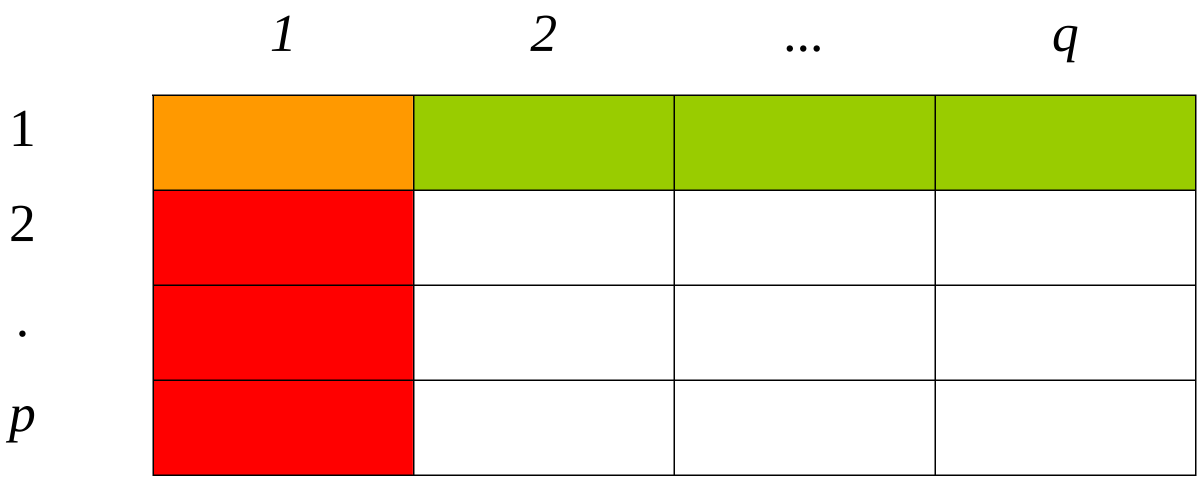
1) $R \to \infty$, $B \to \infty$ and $K \to \infty$.

$$\mathrm{Cov}(y_{ijt}, y_{i'j't'}) = \begin{cases} \sigma_a^2 + \sigma_b^2 + \sigma_e^2 + \sigma^2, & i = i',\ j = j', t = t', \\ \sigma_a^2 + \sigma_b^2, & i = i',\ j = j', t \neq t', \\ \sigma_a^2, & i = i',\ j \neq j, \\ 0, & i \neq i', \end{cases}$$

Dispersion structure

a	b	b	c	c						
b	a	b	c	c						
b	b	a	c	c						
c	c	c	a	b						
c	c	c	b	a						
					a	b	b	c	c	c
					b	a	b	c	c	c
					b	b	a	c	c	c
					c	c	c	a	b	b
					c	c	c	b	a	b
					c	c	c	b	b	a

Row – column design



Randomization:

1) the randomization of the rows (columns),

4) randomization of the columns (rows),

The obtained model has a form:

$$E(y_{ijt}) = \mu + \tau_t, \qquad (6)$$

$$\mathrm{Cov}(y_{ijt}, y_{i'j't'}) = \begin{cases} \sigma_a^2 + \sigma_b^2 + \sigma_e^2 + \sigma^2, & i = i',\ j = j', \\ \sigma_a^2 - (q-1)^{-1}\sigma_b^2 - (p\ \ -1)^{-1}\sigma_e^2, & i = i',\ j \neq j', \\ -(p-1)^{-1}\sigma_a^2 + \sigma_b^2 - (p-1)\sigma_\eta^2, & i \neq i',\ j = j', \\ -(p-1)^{-1}\sigma_a^2 - (q-1)^{-1}\sigma_b^2 + & \\ (p-1)^{-1}(q-1)^{-1}\sigma_\eta, & i \neq i', j \neq j' \end{cases}$$

Let p=2, q=5,

Dispersion structure

a	b	b	b	b	e	d	d	d	d
b	a	b	b	b	d	e	d	d	d
b	b	a	b	b	d	d	e	d	d
b	b	b	a	b	d	d	d	e	d
b	b	b	b	a	d	d	d	d	e
e	d	d	d	d	a	b	b	b	b
d	e	d	d	d	b	a	b	b	b
d	d	e	d	d	b	b	a	b	b
d	d	d	e	d	b	b	b	a	b
d	d	d	d	e	b	b	b	b	a

Analysis

1) **Randomization model**

2) **Multistratum experiments (OBS, Genaral balance)**

3) **Approximation by normal distribution**

   (normal model)

References (selected)

Calinski T., 1994: On the randomization theory of experiments in nested block designs. *Biometrical Letters* **31**, 45-78.

Caliński, T. Kageyama, S. (2000). Block Designs: A Randomization Approach, Vol. I: Analysis. Lecture Notes in Statistics 150. Springer-Verlag, New York.

Caliński, T. Kageyama, S. (2003). Block Designs: A randomization approach, Vol. II: Design. Lecture Notes in Statistics 170. Springer-Verlag, New York

Hinkelmann, K., Kempthorne, O., 1994. Design and analysis of experiments. Vol. I. Introduction to experimental design. Wiley, New York

Houtman A.M., Speed T.P., 1983: Balance in designed experiments with orthogonal block structure. *Ann. Statist.* **11,** 1069-1085**.**

Mejza S., (1992). On some aspects of general balance in designed experiments. Statistica LII, 263-278.

Mejza S., 1994: On modelling of experiments in natural sciences. *Biometrical Letters* **31**, 79-100.

Mejza S., and Kageyama S., 1998: Some Statistical properties of nested block designs. Atkinson A.C., Pronzato L., and Wynn H., (Eds) Proceedings of MODA 5, Physica Verlag, Heidelberg, 231-238.

Nelder J.N., 1965: The analysis of experiment with orthogonal block structure. *Proc. R. Soc. Lond. A.*, **283**, 147-178.

Neyman J., Iwaszkiewicz, K. and Kolodziejczyk, St., 1935: Statistical problems in agricultural experimentation. *J. Roy. Statist. Soc. Suppl.*, **2**, 107-154.