Dichotomization and Estimation of Interaction through a Boolean Framework

Julia Nagrebetskaya, Vladimir Panov

Ural Federal University
Department of Mathematics, Mechanics and Computer Science
Institute of Industrial Ecology of the Ural Branch of RAS

Yekaterinburg, Russia

7.05.2023

In biomedical research, it is a quite common situation when the acting factors have a finite number of levels.

The simplest case of this kind is binary acting factors and a binary response [MacMahon and Pugh, 1967, Rothman, 1976]. Such a binary model considered in the sufficient causes theory which is one of the basic models of causality in epidemiology and evidence based medicine [Rothman, 1976, Greenland and Poole, 1988, VanderWeele and Richardson, 2012].

It was shown by the authors that efficient mathematical formalization of the binary theory of sufficient causes can be performed within the Boolean framework. In particular, this formalization allows us to introduce new concepts for mathematical analysis of the binary factors interaction. One such a new concept is proposed in [Authors, 2015-2019, USBEREIT, 2022] under the name "degree of interaction in a given response".

Although the natural application of the Boolean model of sufficient causes theory is the analysis of binary experiments, it can be also applied to continuous or multilevel factors and response.

Let random variables $X_1, X_2, \ldots, X_n$ are the independent factors and a variable Y be a response. We can build their binary versions $x_1, x_2, \ldots, x_n, \ y$ as follows.

Let's dichotomize every variable X_i, $i \in \{1, 2, \ldots, n\}$, by its threshold value.

As a result, the factors $x_1, x_2, \ldots, x_n$ can be considered as Boolean variables, and the response y as a Boolean function $y = f(\boldsymbol{x})$ of Boolean variables $\boldsymbol{x} = x_1, x_2, \ldots, x_n$.

Thus, for each $i \in \mathbb{N}_n = \{1, 2, \ldots, n\}$ we obtain realizations $\boldsymbol{x_i} = \{x_i^s\}_{s=1}^N$ and $\boldsymbol{y} = \{y^s\}_{s=1}^N$ of Boolean variables x_i and y respectively.

Let a tuple $\boldsymbol{\alpha} \in \mathbb{B}^n$. We denote by $m_{\boldsymbol{\alpha}}$ the number of tuples $\boldsymbol{x}^s = (x_1^s, x_2^s, \ldots, x_n^s)$, $s \in \mathbb{N}_N$, such that $\boldsymbol{x}^s = \boldsymbol{\alpha}$ and $y^s = 1$.

Denote by $N_{\boldsymbol{\alpha}}$ the number of all tuples $\boldsymbol{x}^s$ such that $\boldsymbol{x}^s = \boldsymbol{\alpha}$. Then the ratio

$$w_{\boldsymbol{\alpha}} = \frac{m_{\alpha}}{N_{\boldsymbol{\alpha}}}$$

is the relative frequency of the event $\{y = 1 \mid \boldsymbol{x} = \boldsymbol{\alpha}\}$. We consider a full factorial experiment, so $N_{\boldsymbol{\alpha}} > 0$.

It is clear that $w_{\boldsymbol{\alpha}}$ is a statistical estimate of the probability

$$p_{\boldsymbol{\alpha}} = P(y = 1 \mid \boldsymbol{x} = \boldsymbol{\alpha}).$$

Consider the support $C_f = \{\boldsymbol{\alpha} \in \mathbb{B}^n \mid f(\boldsymbol{\alpha}) = 1\}$ of a Boolean function $f \in \mathbb{B}(\boldsymbol{x})$ as a fuzzy set with membership function $A : \mathbb{B}^n \to [0; 1]$

$$A(\boldsymbol{\alpha}) = p_{\boldsymbol{\alpha}}^{f(\boldsymbol{\alpha})}$$

Here notations $z^0 = 1 - z$ and $z^1 = z$ are used.

Thus, we obtain a probability distribution of the random Boolean function f for the given response y:

$$\{(f, p_f) \mid f \in \mathbb{B}(\boldsymbol{x})\},$$

where

$$p_f = \prod_{\boldsymbol{\alpha} \in \mathbb{B}^n} A(\boldsymbol{\alpha})$$

For each Boolean function $f(\boldsymbol{x})$ and any $k \in \mathbb{N}_n$ one can calculate the degree $\mu_{f,k}$ of the joint action of k variables, which allows one to find the probability distribution of the random variable $\mu_{f,k}$:

$$\{(\mu_{f,k}, p_{ki}) \mid i \in \{0, 1, \ldots, n\}\},$$

where

$$p_{ki} = P(\mu_{f,k} = i) = \sum_{\substack{\mu_{f,k}=i \\ f \in \mathbb{B}(x)}} p_f.$$

This distribution allows us to calculate the mean of the degree of joint action

$$\overline{\mu}_{f,k} = \sum_{i=0}^{n} i \cdot p_{ki},$$

which is some characteristic of the **degree (strength) of interaction of the factors in the original data**.

As the simplest case, consider a deterministic model in which all factors X_i for $i \in \mathbb{N}_n$ and response Y take values 0 or 1, and all combinations of $\boldsymbol{\alpha} \in \mathbb{B}^n$ levels of factors appear once.

The response's values for all possible sets of $\boldsymbol{\alpha} \in \mathbb{B}^n$ define the response as a Boolean function $g(\boldsymbol{X}), \boldsymbol{X} = X_1, X_2, \ldots, X_n$.

Then for any Boolean function $f \in \mathbb{B}(\boldsymbol{x})$ we have $p_f = 1$ if and only if $f = g$.

Hence, for the considered deterministic model the mean $\overline{\mu}_{f,k}$ coincides with the degree $\mu_{g,k}$ of the joint action of k factors in the corresponding Boolean function.

All calculations have been performed in CAS Wolfram *Mathematica* v. 13.0.

Throughout the following, only two factors will be considered ($n = 2$). Since $\mu_{f,1} = 1$ for all non-constant Boolean functions $f \in \mathbb{B}(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2)$, we are only interested in the degree $\mu_f = \mu_{f,2}$ of the joint action of the both factors x_1, x_2.

Let us denote $\boldsymbol{X}_i = \{X_i^s\}_{s=1}^N$, $i \in \{1, 2\}$, $\boldsymbol{Y} = \{Y^s\}_{s=1}^N$ measured numerical values of the factors X_i, and the response Y.

For the Boolean function $f_1(\boldsymbol{x}) = x_1 \wedge x_2$, and fixed $l \in \mathbb{N}$ consider simulation data for $N = 2^2 l$, whose dichotomization leads to a case very close to that of the deterministic model.

Let for $\boldsymbol{X}^s = (X_1^s, X_2^s)$, $\boldsymbol{\xi}^s = (\xi_1^s, \xi_2^s)$, $s \in \mathbb{N}_N$,

$$\left.\begin{array}{l} \boldsymbol{X}^1 = (0,0) + (\xi_1^1, \xi_2^1) \\ \boldsymbol{X}^2 = (0,0) + (\xi_1^2, \xi_2^2) \\ \dots \\ \boldsymbol{X}^l = (0,0) + (\xi_1^l, \xi_2^l) \end{array}\right\} l \qquad \left.\begin{array}{l} \boldsymbol{X}^{2l+1} = (1,0) + (\xi_1^{2l+1}, \xi_2^{2l+1}) \\ \boldsymbol{X}^{2l+2} = (1,0) + (\xi_1^{2l+2}, \xi_2^{2l+2}) \\ \dots \\ \boldsymbol{X}^{3l} = (1,0) + (\xi_1^{3l}, \xi_2^{3l}) \end{array}\right\} l$$

$$\left.\begin{array}{l} \boldsymbol{X}^{l+1} = (0,1) + (\xi_1^{l+1}, \xi_2^{l+1}) \\ \boldsymbol{X}^{l+2} = (0,1) + (\xi_1^{l+2}, \xi_2^{l+2}) \\ \dots \\ \boldsymbol{X}^{2l} = (0,1) + (\xi_1^{2l}, \xi_2^{2l}) \end{array}\right\} l \qquad \left.\begin{array}{l} \boldsymbol{X}^{3l+1} = (1,1) + (\xi_1^{3l+1}, \xi_2^{3l+1}) \\ \boldsymbol{X}^{3l+2} = (1,1) + (\xi_1^{3l+2}, \xi_2^{3l+2}) \\ \dots \\ \boldsymbol{X}^{4l} = (1,1) + (\xi_1^{4l}, \xi_2^{4l}) \end{array}\right\} l$$

$$\left.\begin{array}{l} Y^1 = f_1(0,0) + \xi_3^1 \\ Y^2 = f_1(0,0) + \xi_3^2 \\ \dots \\ Y^l = f_1(0,0) + \xi_3^l \end{array}\right\} l \qquad \left.\begin{array}{l} Y^{2l+1} = f_1(1,0) + \xi_3^{2l+1} \\ Y^{2l+2} = f_1(1,0) + \xi_3^{2l+2} \\ \dots \\ Y^{3l} = f_1(1,0) + \xi_3^{3l} \end{array}\right\} l$$

$$\left.\begin{array}{l} Y^{l+1} = f_1(0,1) + \xi_3^{l+1} \\ Y^{l+2} = f_1(0,1) + \xi_3^{l+2} \\ \dots \\ Y^{2l} = f_1(0,1) + \xi_3^{2l} \end{array}\right\} l \qquad \left.\begin{array}{l} Y^{3l+1} = f_1(1,1) + \xi_3^{3l+1} \\ Y^{3l+2} = f_1(1,1) + \xi_3^{3l+2} \\ \dots \\ Y^{4l} = f_1(1,1) + \xi_3^{4l} \end{array}\right\} l$$

where $\xi_1^s, \xi_2^s, \xi_3^s$ are realizations of a uniform distribution ξ on an interval $[-\varepsilon, \varepsilon], \varepsilon > 0$.

The random variable ξ can be interpreted as the measurement error of each factor and response which now takes values $0, 1$.

It is clear that the data set $\{(x_1^s, x_2^s, y^s) \mid s \in \mathbb{N}_N\}$ obtained after dichotomization consists of l copies of triples $(0,0,0)$, $(0,1,0)$, $(1,0,0)$ and $(1,1,1)$.

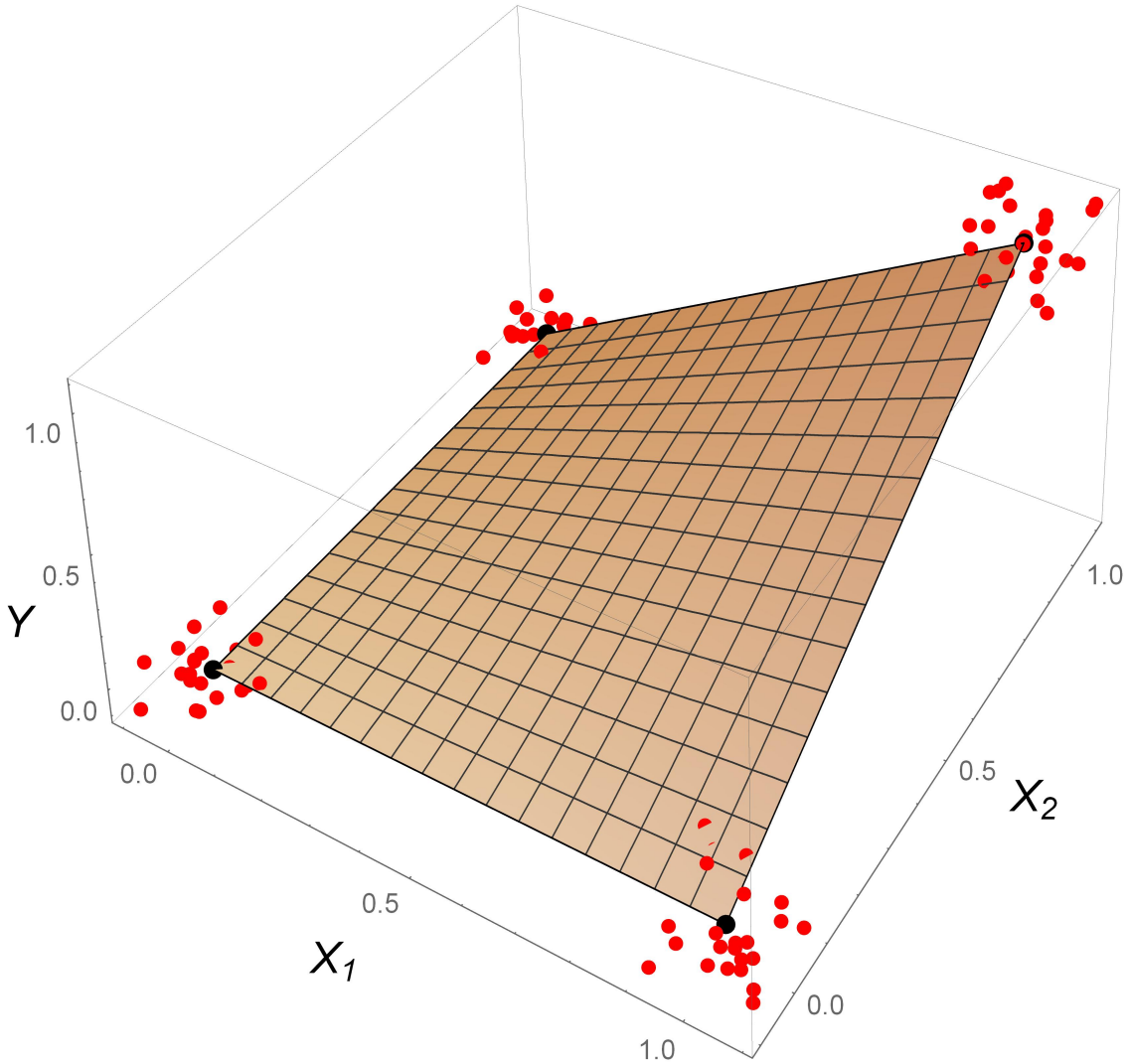
Thus, we find ourselves within the deterministic model.

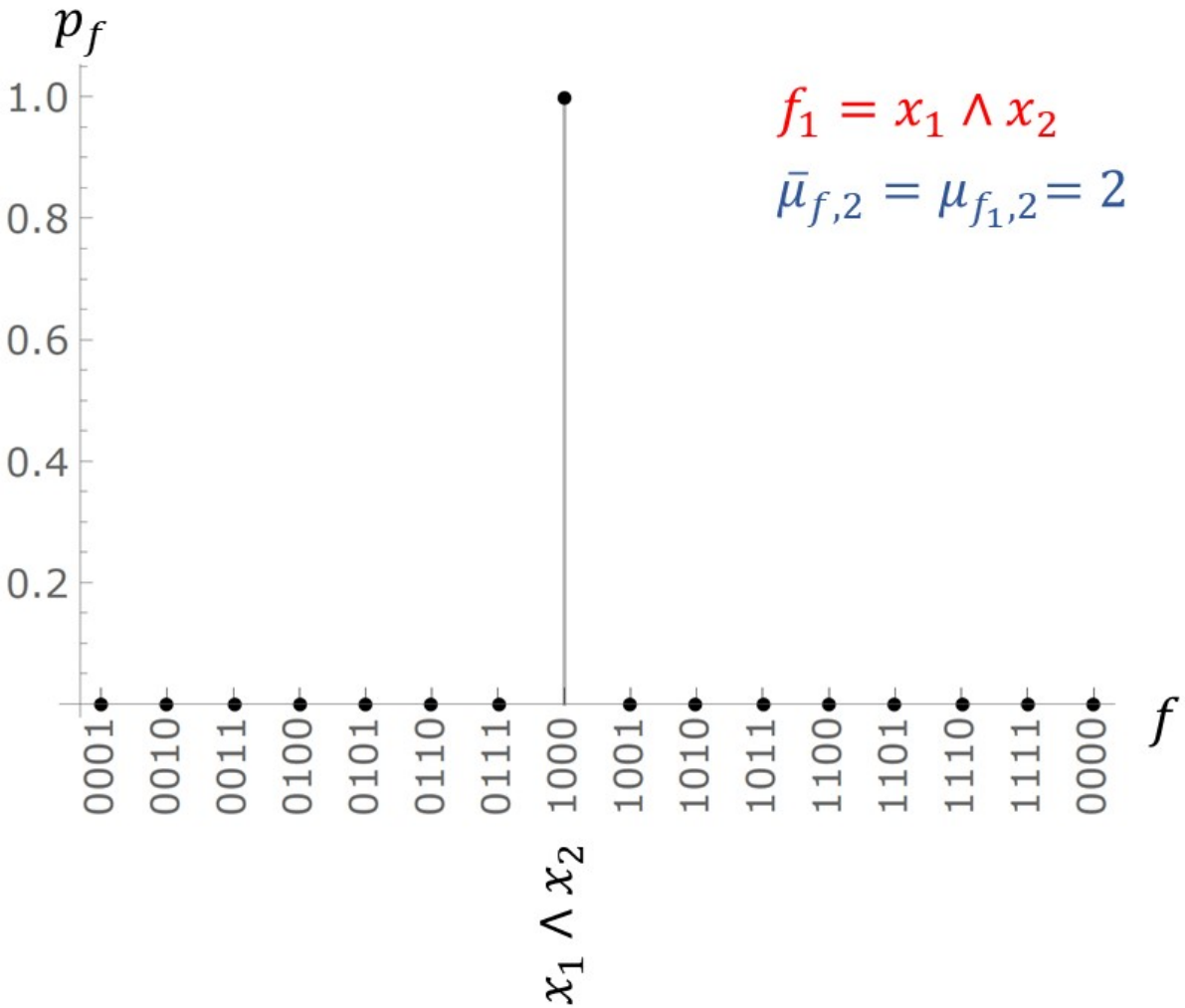
For this example, a numerical experiment was carried out at $N = 100$ and $\varepsilon = 0.1$.

Then for a Boolean function $f \in \mathbb{B}(\boldsymbol{x})$ the equality $p_f = 1$ holds if and only if $f = f_1$. Besides, $\overline{\mu}_f = \mu_{f_1} = 2$, i. e. the proportion of the degree of interaction of the factors X_1, X_2 relative to the maximal possible one is 100%.

The same results can be obtained for other Boolean functions, for example, $f_2 = x_1 \vee x_2$ or $f_3 = \neg x_1$.

Slide 12. Data for the almost deterministic model and the surface $Y = X_1 X_2$





Each Boolean function $f = \beta_{11} x_1 x_2 + \beta_{10} x_1 \bar{x}_2 + \beta_{01} \bar{x}_1 x_2 + \beta_{00} \bar{x}_1 \bar{x}_2$ is encoded by $\beta_{11}\beta_{10}\beta_{01}\beta_{00}$.

Let the factors X_1, X_2 are continuous random variables taking values on the interval $[0;1]$ which distributions will be specified below. We consider two types of these random variables' distributions and generate their realizations $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ for each type.

Consider Boolean functions $f_1 = x_1 \wedge x_2$, $f_2 = x_1 \vee x_2$, and $f_3 = \neg x_1$ which are matched to product $g_1(X_1, X_2) = X_1 X_2$, probabilistic sum $g_2(X_1, X_2) = X_1 + X_2 - X_1 X_2$ and involutive negation $g_3(X_1) = 1 - X_1$, respectively, in fuzzy logic.

We define responses Y_1, Y_2, Y_3 as follows $Y_j = g_j(X_1, X_2) + \xi$, $j \in \mathbb{N}_3$, where ξ is uniformly distributed on the interval $[-\varepsilon; \varepsilon]$. The random variable ξ simulates the measurement error of the response values.

Using these functions, we construct data arrays such that after their dichotomization the corresponding Boolean functions f_1, f_2, f_3 will have the **highest probability** in the distribution of all Boolean functions (see slide 5).

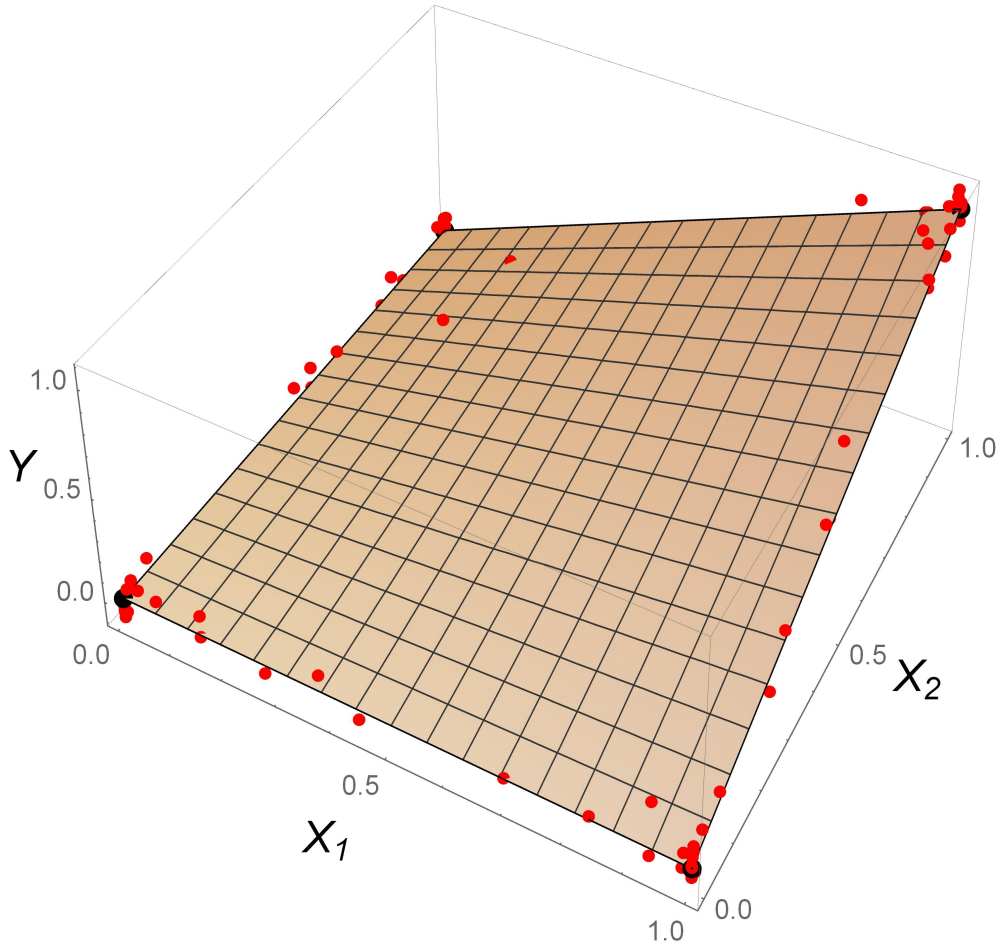
For most of the values of factors X_1, X_2 to be "close" to 0 or 1, we set $X_i = F(Z), i \in \{1, 2\}$, where a random variable Z is uniformly distributed on the interval $[0; 1]$, and F is the Bates' cumulative distribution function.

Then, by generating for $N = 100$ and $\varepsilon = 0.1$ realizations of $\boldsymbol{X}_1$, and $\boldsymbol{X}_2$ of the factors X_1, X_2, we can calculate the corresponding values for the responses Y_j, $j \in \{1, 2, 3\}$, and obtain the desired results.

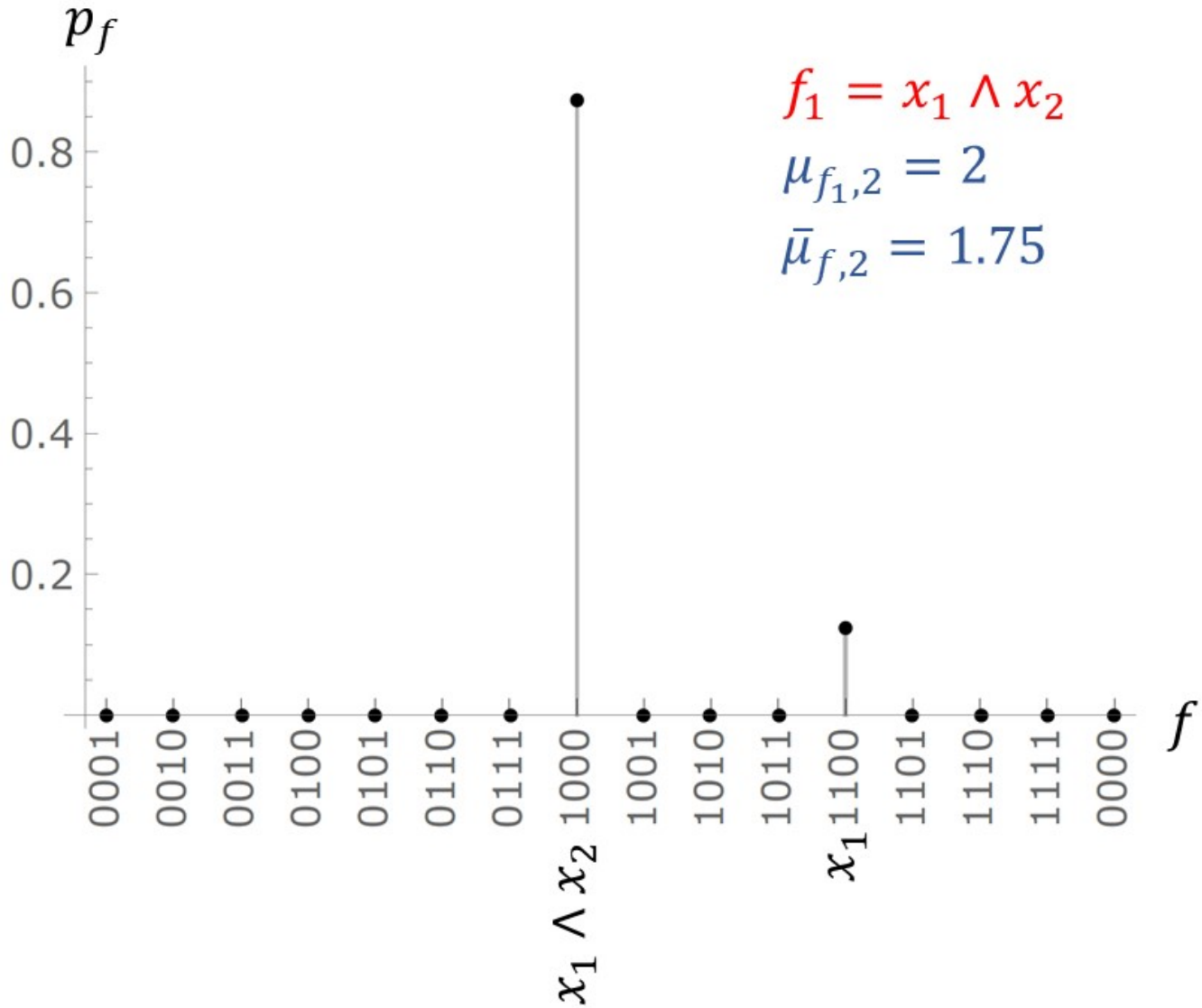
The generated data and the surface $Y = X_1 X_2$, $X_1, X_2 \in [0; 1]$, are presented in the next slide for $j = 1$.

The mean of the degree value is $\overline{\mu}_f = 1.75$, hence, **the proportion of the degree of joint action of the factors** X_1, X_2 **relative to the maximal value is** 87.5%.

Slide 16. Data for the Example 2 and the surface $Y = X_1 X_2$ for $j = 1$.



Slide 17. **Example 1**. Probability Distribution of Boolean Function and the mean of the degree of joint action

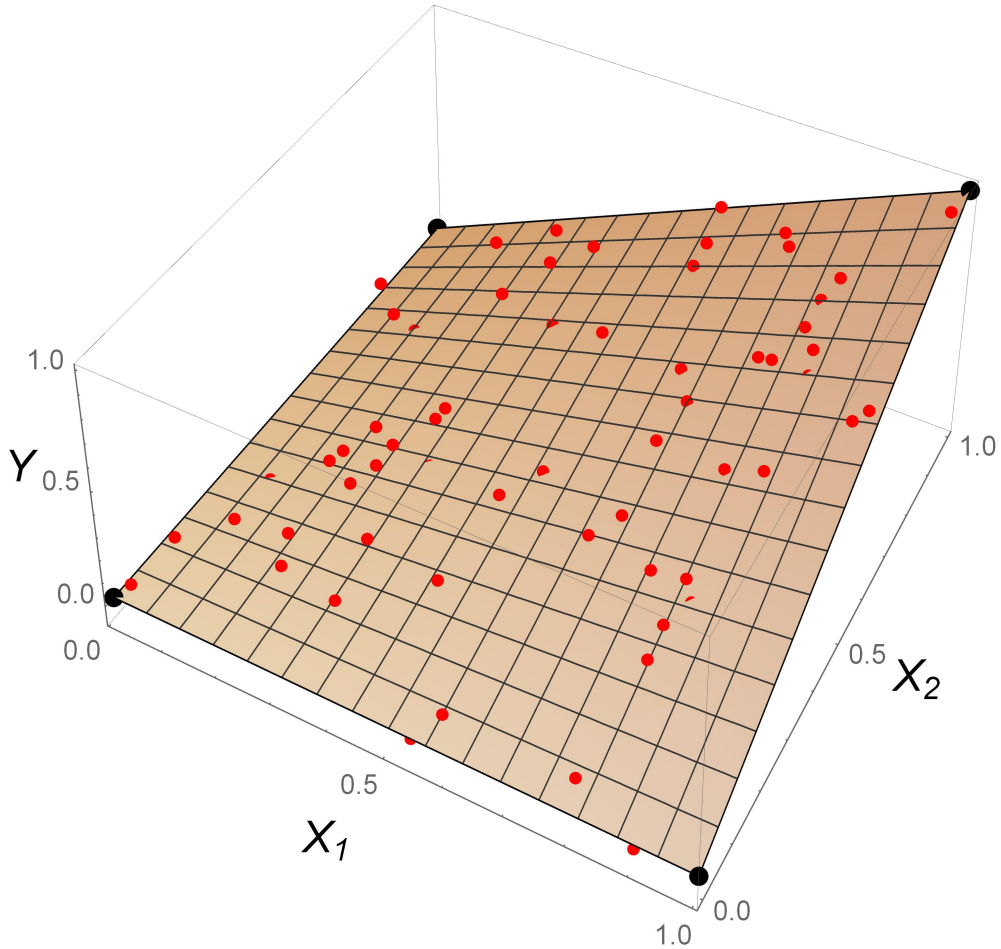


Let's consider data simulation with the factors X_1, X_2 which are uniformly distributed on the segment $[0; 1]$.

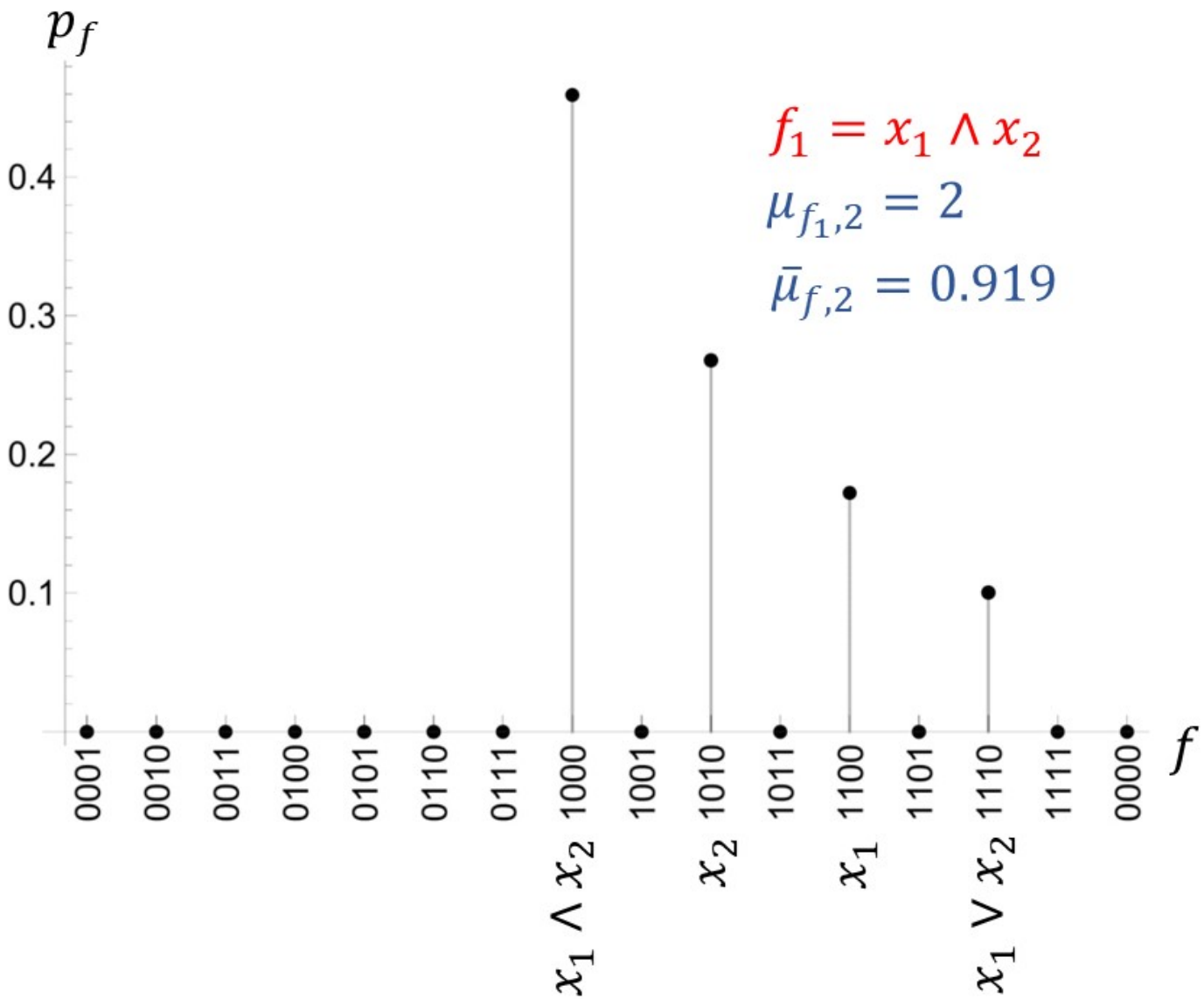
We generate for $N = 100$ and $\varepsilon = 0.1$ realizations $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ of the factors X_1, X_2 and corresponding values of the responses Y_j, $j \in \{1, 2, 3\}$
(see, next slide for $j = 1$).

We obtain the following results which as well as the previous ones confirm the constructed mathematical model.

Slide 19. Data for the Example 2 and the surface $Y = X_1 X_2$ for $j = 1$.



Slide 20. **Example 2**. Probability Distribution of Boolean Function and the mean of the degree of joint action



1. B. MacMahon and T. F. Pugh, *Causes and entities of disease*. 1em plus 0.5em minus 0.4emBoston: Little Brown, 1967, pp. 11–18.

2. K. J. Rothman, "Causes," *Am. J. of Epidemiology*, vol. 104, no. 6, pp. 587–592, 1976.

3. S. Greenland and C. Poole, "Invariants and noninvariants in the concept of interdependent effects," *Scand. J. Work Environ. Health*, vol. 14, pp. 125–129, 1988.

4. T. J. VanderWeele and T. S. Richardson, "General theory for interactions in sufficient cause models with dichotomous exposures," *Ann. Statistics*, vol. 40, pp. 2128–2161, 2012.

5. V. Panov and J. V. Nagrebetskaya, "Boolean algebras and classification of interactions in sufficient-component cause model," *Int. J. Pure Appl. Math.*, vol. 98, no. 2, pp. 239–259., 2015.

6. V. Panov and J. V. Nagrebetskaya, "Classification of combined action of binary factors and Coxeter groups," *J. Discr. Math. Sci. and Cryptography*, vol. 21, no. 3, pp. 661–667, 2018

7. J. Nagrebetskaya and V. Panov, "Joint action of binary factors in the sufficient causes theory and its classification," *Int. J. Innovative Technology and Exploring Engineering*, vol. 9, no. 1, pp. 2146–2152, 2019.

8. J. Nagrebetskaya and V. Panov, "Spectrum of Joint Action of Factors in the Binary Theory of Sufficient Causes," in *2022 Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 2022, pp. 208–211.

Thank you for your attention!