

# Лог-линейный анализ

---

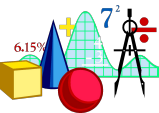
Стат. методы в  
психологии  
(Радчикова Н.П.)



---

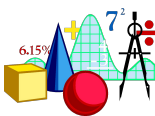
# Цели

- **Что делать, если таблица сопряженности не двухмерная, а трехмерная или еще хуже?**





**Применять  
лог-линейный  
анализ!**

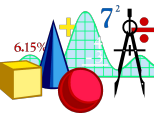




# МОДЕЛИ

**Математики любят модели.**

**Каждая модель соответствует  
определенной гипотезе о  
переменных, входящих в таблицу  
сопряженности.**

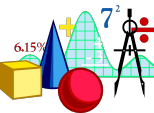




# МОДЕЛИ

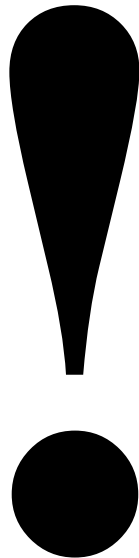
**Идея состоит в том, чтобы взять модель и проверить, совпадают ли эмпирические данные с предсказанными моделью результатами.**

**Та модель, где совпадение наибольшее, признается лучшей, т.е. наиболее адекватно описывающей полученные данные.**

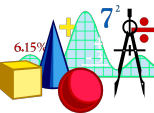




# МОДЕЛИ



**В модели лог-линейного  
анализа переменные  
НЕ ДЕЛЯТСЯ  
на независимые и  
зависимые переменные**

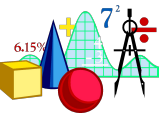




# ДВУХМЕРНАЯ МОДЕЛЬ

Рассмотрим сначала лог-линейную модель для двухмерной таблицы сопряженности с  $r$  строками и  $c$  столбцами

Наблюдаемое значение =  
ожидаемое значение + ошибка





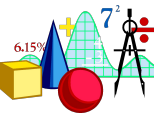
# ДВУХМЕРНАЯ МОДЕЛЬ

★ Наблюдаемое значение – это эмпирическая частота  $n_{ij}$  в каждой клетке таблицы

★ Ожидаемое значение – это теоретическая частота  $F_{ij}$

Поэтому можно написать:

$$n_{ij} = F_{ij} + \text{ошибка}$$







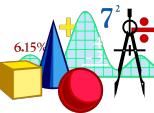
# ДВУХМЕРНАЯ МОДЕЛЬ

★ Наблюдаемое значение – это эмпирическая частота  $n_{ij}$  в каждой клетке таблицы

★ Ожидаемое значение – это теоретическая частота  $F_{ij}$

Поэтому можно написать:

$$n_{ij} = F_{ij} + \text{ошибка}$$





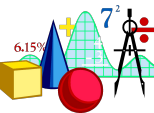
# ДВУХМЕРНАЯ МОДЕЛЬ


Предположив, что наблюдения  
независимы, получаем:

$$F_{ij} = N p_{i.} p_{.j} = N \frac{F_{i.}}{N} \frac{F_{.j}}{N} = \frac{F_{i.} F_{.j}}{N}$$

$p_{i.}$  – это вероятность попасть в категорию  $i$  переменной 1,

$p_{.j}$  – это вероятность попасть в категорию  $j$  переменной 2.





# Помните, как мы определяли теоретическую частоту?

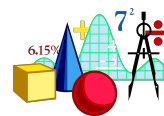
Для выделенной ячейки:

Подставив все это в формулу

$$F_{ij} = N p_{i.} p_{.j} = N \frac{F_{i.}}{N} \frac{F_{.j}}{N} = \frac{F_{i.} F_{.j}}{N}$$

получим теоретическую частоту для выделенной клетки:

$$F_{ij} = (200/550) * (350/550) * 550 = 127,3.$$

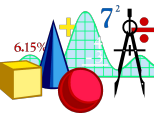




# ДВУХМЕРНАЯ МОДЕЛЬ

Возьмем натуральный логарифм и  
получим:

$$\ln F_{ij} = \ln F_{i.} + \ln F_{.j} - \ln N$$





# ДВУХМЕРНАЯ МОДЕЛЬ

А это выражение можно представить в виде:

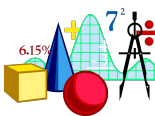
$$\ln F_{ij} = u + u_{1(i)} + u_{2(j)}$$

где

$$u = \frac{\sum_{i=1}^r \sum_{j=1}^c \ln F_{ij}}{rc},$$

$$u_{1(i)} = \frac{\sum_{j=1}^c \ln F_{ij}}{c} - u$$

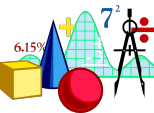
$$u_{2(j)} = \frac{\sum_{i=1}^r \ln F_{ij}}{r} - u$$





# ДВУХМЕРНАЯ МОДЕЛЬ

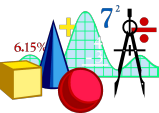
- ★ говорят, что  $\mu$  представляет собой «общий средний эффект»
- ★  $\mu_{1(i)}$  - «главный эффект» уровня  $i$  переменной, расположенной по строкам
- ★  $\mu_{2(j)}$  - «главный эффект» уровня  $j$  переменной, расположенной по столбцам





# ДВУХМЕРНАЯ МОДЕЛЬ

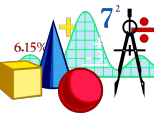
**Значения, представленные как главные эффекты в этой модели, просто отражают разницу между маргинальными частотами по строкам или столбцам и мало нас интересуют**





# ДВУХМЕРНАЯ МОДЕЛЬ

**Лог-линейная модель может быть проверена посредством оценки параметров (т.е. теоретических частот) и сравнением этих оценок с наблюдаемыми (эмпирическими) частотами. Это можно сделать с помощью известной нам процедуры  $\chi^2$  Пирсона**



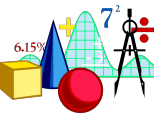




# ДВУХМЕРНАЯ МОДЕЛЬ

**Если модель с независимыми переменными плохо подходит для оценки исходной таблицы (т.е.  $\chi^2$  получился значимый), то в модель следует ввести дополнительной слагаемое, которое будет представлять собой связь между переменными**

$$\ln F_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

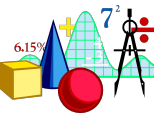




# ДВУХМЕРНАЯ МОДЕЛЬ

$$\ln F_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

**Эта модель всегда полностью  
описывает  
таблицу сопряженности размером 2\*2.**





# ТРЕХМЕРНАЯ МОДЕЛЬ

$$\ln F_{ij} = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$$

$u$  – общий «средний» эффект

$u_1$  – главный эффект переменной 1

$u_2$  – главный эффект переменной 2

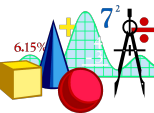
$u_3$  – главный эффект переменной 3

$u_{12}$  – взаимодействие между переменными 1 и 2

$u_{13}$  – взаимодействие между переменными 1 и 3

$u_{23}$  – взаимодействие между переменными 3 и 2

$u_{123}$  – взаимодействие между тремя переменными  
(взаимодействие второго порядка)

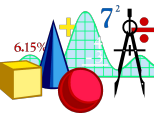




# ТРЕХМЕРНАЯ МОДЕЛЬ

**ЦЕЛЬ:**

**найти модель с минимальным  
количеством параметров,  
которая бы адекватно  
предсказывала эмпирические  
частоты**

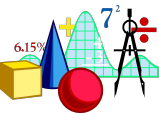




# ТРЕХМЕРНАЯ МОДЕЛЬ

**Следует помнить,  
что данная модель – иерархическая.**

**Это значит, что если в модель  
включены эффекты более высоких  
порядков, то автоматически  
включаются и эффекты более низких  
порядков.**





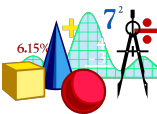
# ТРЕХМЕРНАЯ МОДЕЛЬ

Например, если слагаемое  $u_{123}$  включено, то будут включены и слагаемые  $u_1$ ,  $u_2$ ,  $u_3$ ,  $u_{12}$ ,  $u_{13}$  и  $u_{23}$ .

Например, модель

~~$$\ln F_{ij} = u + u_1 + u_2 + u_3 + u_{123}$$~~

недопустима.

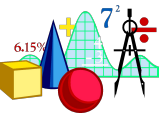




# ТРЕХМЕРНАЯ МОДЕЛЬ

**Каждая модель, которую можно придумать для трехмерной таблицы сопряженности, соответствует определенной гипотезе о переменных, входящих в таблицу.**

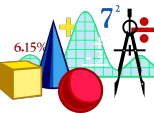
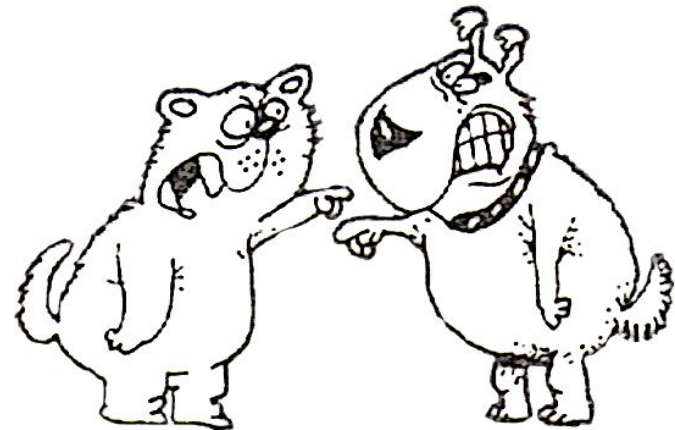
**Рассмотрим каждую модель подробнее.**





# Любимый пример

**Усложним любимый пример: пусть теперь мы хотим проверить, правда ли, что мужчины больше любят собак, а женщины – кошек, и не зависит ли это отношение от возраста**







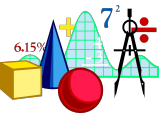
# Модель (1)

$$(1) \ln F_{ij} = u$$

Все частоты в таблице одинаковы

	мужчины	
	собака	кошка
Ребенок	40	40
Взрослый	40	40

	женщины	
	собака	кошка
Ребенок	40	40
Взрослый	40	40





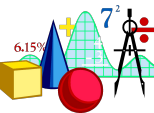
# Модель (2)

[1]

$$(2) \ln F_{ij} = u + u_1$$

Маргинальные частоты для переменных 2 и 3  
равны

	мужчины		женщины	
	собака	кошка	собака	кошка
Ребенок	20	20	20	20
Взрослый	10	10	10	10





# Модель (3)

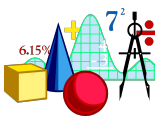
[1] [2]

(3)  $\ln F_{ij} = u + u_1 + u_2$

Маргинальные частоты для переменной  
3 равны

	мужчины	
	собака	кошка
Ребенок	10	10
Взрослый	30	10

	женщины	
	собака	кошка
Ребенок	10	10
Взрослый	30	10

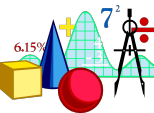




---

Эти модели являются неинтересными, так как не позволяют эмпирическим частотам отражать эмпирическую разницу в маргинальных частотах каждой переменной. Фактически они сводятся к двумерному случаю.

И, видимо, могут быть проинтерпретированы как случай, когда все три переменные независимы.





# Модель (4)

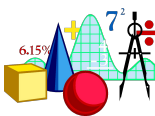
[1] [2] [3]

$$(4) \ln F_{ij} = u + u_1 + u_2 + u_3$$

Все переменные независимы (?)

	мужчины	
	собака	кошка
Ребенок	20	20
Взрослый	40	20

	женщины	
	собака	кошка
Ребенок	10	10
Взрослый	30	10



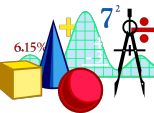


# Модель (5)

[12] [3]

$$(5) \ln F_{ij} = u + u_1 + u_2 + u_3 + u_{12}$$

Переменные 1 и 2 зависимы и обе независимы от переменной 3.





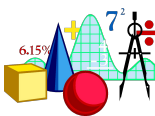
# Модель (5)

[12] [3]

Все дети любят кошек, а взрослые – собак.  
Переменные «возраст» и «домашнее животное»  
связаны, и обе они не зависят от пола.

	мужчины	
	собака	кошка
Ребенок	5	40
Взрослый	40	5

	женщины	
	собака	кошка
Ребенок	5	40
Взрослый	40	5



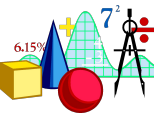


## Модель (6)

[12] [13]

$$(6) \ln F_{ij} = u + u_1 + u_2 + u_3 + u_{12} + u_{13}$$

Переменные 2 и 3 независимы на каждом уровне переменной 1, но каждая зависит от переменной 1.







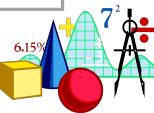
# Модель (6)

[12] [13]

Возраст и предпочтение домашнего животного связаны с полом, но возраст и предпочтение домашнего животного не связаны.

	мужчины	
	собака	кошка
Ребенок	40	20
Взрослый	80	40

	женщины	
	собака	кошка
Ребенок	40	80
Взрослый	10	20



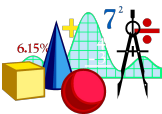


## Модель (7)

[12] [13] [23]

$$(7) \ln F_{ij} = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23}$$

Каждая пара переменных связана, но направление связи одинаково для каждого уровня третьей переменной.





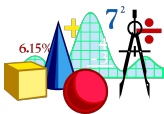
# Модель (7)

[12] [13] [23]

Женщины любят собак, а мужчины кошек.  
Дети любят кошек, а взрослые собак.  
Женщины взрослые, а мужчины – дети.

	мужчины	
	собака	кошка
Ребенок	20	80
Взрослый	20	20

	женщины	
	собака	кошка
Ребенок	20	20
Взрослый	80	20



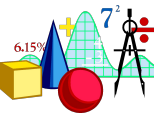


# Модель (8)

[123]

$$(8) \ln F_{ij} = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$$

Взаимодействие второго порядка.  
Все переменные связаны.





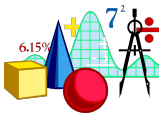
# Модель (8)

[123]

Маленькие мальчики любят кошек, а взрослые мужчины – собак. Маленькие девочки любят собак, а взрослые женщины – кошек.

	мужчины	
	собака	кошка
Ребенок	5	40
Взрослый	40	5

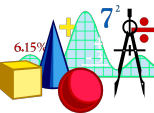
	женщины	
	собака	кошка
Ребенок	40	5
Взрослый	5	40





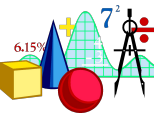
---

Больше для трехмерного случая никаких  
моделей придумать нельзя.





**Лог-линейные  
модели можно  
подбирать для  
четырех и более  
переменных  
аналогичным  
образом**



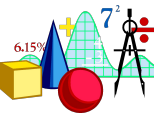


**★ Главная идея метода:**

**Подбираем последовательно модели от самых простых до самых сложных и проверяем, насколько предсказанные моделью частоты совпадают с эмпирическими частотами.**

**★ Если совпадают, процесс подбора модели закончен.**

**★ Поэтому удачной будет та модель, для которой хи-квадрат незначимый!**





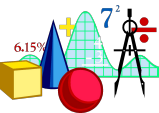


**Эти ценные сведения о лог-линейном  
анализе можно почерпнуть в**

**Everitt B.S.  
Making Sense of Statistics  
in Psychology. –**

**Oxford University Press, 1996. – 350 p.**

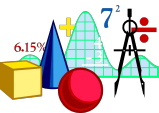
**(перевод – в папке «Дополнительная  
литература»)**





---

**А нам теперь интересно, как найти  
подходящую модель, если у нас есть  
ТОЛЬКО данные.**





# Это можно сделать в программе STATISTICA, в специальном модуле Statistics - Advanced Linear/Nonlinear Models -Log-Linear Analysis of Frequency Tables

The screenshot shows the STATISTICA software interface with the following menu structure:

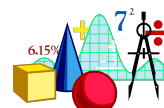
- File
- Edit
- View
- Insert
- Format
- Statistics
- Data Mining
- Graphs
- Tools
- Data
- Window
- Help

The 'Statistics' menu is open, showing the following options:

- Basic Statistics/Tables
- Multiple Regression
- ANOVA
- Nonparametrics
- Distribution Fitting
- Advanced Linear/Nonlinear Models** (highlighted with a red circle)
  - General Linear Models
  - Generalized Linear/Nonlinear Models
  - General Regression Models
  - General Partial Least Squares Models
  - NIPALS Algorithm (PCA/PLS)
  - Variance Components
  - Survival Analysis
  - Nonlinear Estimation
  - Fixed Nonlinear Regression
  - Log-Linear Analysis of Frequency Tables** (highlighted with a red circle)
  - Time Series/Forecasting
  - Structural Equation Modeling
- Multivariate Exploratory Techniques
- Industrial Statistics & Six Sigma
- Power Analysis
- Automated Neural Networks
- PLS, PCA, Multivariate/Batch SPC
- Variance Estimation and Precision (VEPAC)
- Statistics of Block Data
- STATISTICA Visual Basic
- Batch (ByGroup) Analysis
- Probability Calculator

The background shows a data table with the following content:

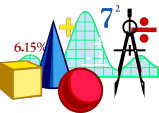
	Результаты тест			
	1			4
	Белорусский			
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13	32	37		23
14	28	14		10
15	30	34		63
16	14	13		15
17	18	24		11
18	39	11		8
19	22	22		13





Стандартное обозначение модели	Обозначение в программе STATISTICA
[1]	1
[1][2]	1 2
[1][2][3]	1 2 3
[12][3]	12
[12][13]	12 13
[12][13][23]	12 13 23
[123]	123

**Иногда в программе STATISTICA вместо пробела используется запятая**





# Выбор переменных

**Log-Linear Analysis**

Table to be analyzed:

AGE	SEX	METHOD
3	x 2	x 6

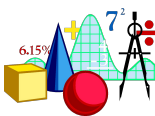
Input file: Raw Data

**Variables:** AGE-METHOD

Variable containing frequencies:

Select codes Selected

OK Cancel Open Data SELECT CASES \$ TO W





Тут можно выбрать коды

Log-Linear Analysis

Table to be analyzed:

AGE	SEX	METHOD
3	x 2	x 6

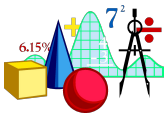
Input file: Raw Data

Variables: AGE-METHOD

Variable containing frequencies:

Select codes Selected

OK Cancel Open Data SELECT CASES \$ TO W





# Окно выбора модели

Тут можно проверить все простые модели

Log-Linear Model Specification: S

Table to be analyzed:

(1)					
AGE					
3	x	2	x	6	

Minimum cell frequency: 5, Maximum: 797, Sum: 5305,

Quick | **Advanced** | Review/Save

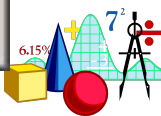
Specify model to be tested | Structural zeros:

**Test all marginal & partial association models** | Delta: .50

Automatic selection of best model | Maximum number of iterations: 50

Convergence criterion: .010

OK | Cancel | Options





# Окно выбора модели

Тут можно задать модель, которую хотим проверить

Table to be analyzed:

(1)					
AGE					
3	x	2	x	6	

Minimum cell frequency: 5, Maximum: 797, Sum: 5305,

Quick | Advanced | Review/Save

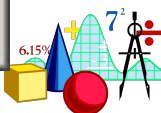
**Specify model to be tested** | Structural zeros:

Test all marginal & partial association models | Delta: .50

Automatic selection of best model | Maximum number of iterations: 50

Convergence criterion: .010

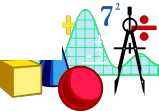
OK | Cancel | Options







**Какой ужас!  
А если я забыл, как  
обозначаются  
модели?!!  
Или совсем не  
помню, какие модели  
бывают?!!**





# Окно выбора модели

Тогда надо жать на эту кнопку!  
«Автоматический выбор лучшей модели»

Log-Linear Model Specification: 9

Table to be analyzed:

(1)
AGE
3

Minimum cell frequency:

Quick | Advanced | Review/Save

Specify model to be tested

Structural zeros:

Test all marginal & partial association models

**Automatic selection of best model**

Delta: .50

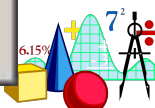
Maximum number of iterations: 50

Convergence criterion: .010

OK

Cancel

Options





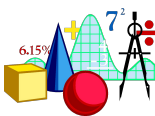
## Automatic Selection of Best Model

Best initial model: Chi-Square = 15,14346      df = 10      p = ,1270  
21,31,32

Best Model: Chi-Square = 15,14339      df = 10      p = ,1270  
21,31,32

**Осталось только  
проинтерпретировать!**

 Further evaluation





## Automatic Selection of Best Model



Best initial model: Chi-Square = 15,14346      df = 10      p = ,1270  
21,31,32

Best Model:  
21,31,32

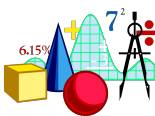
,1270

**А тут можно оценить  
выбранную модель более  
подробно**



Further evaluate the best model

Cancel





**Ура!**  
**Я могу посчитать**  
**лог-линейный**  
**анализ!**

