

Условия задач заимствованы из учебника:

Agresti, A. (1996) An introduction to categorical data analysis. John Wiley & Sons Inc., NY

### Вариант 1.

Описание эксперимента:

3.10. Table 3.6 shows results of a three-center clinical trial designed to compare a drug to placebo for treating severe migraine headaches. At each center, subjects were randomly assigned to treatments.

Table 3.6

Center	Group	Response	
		Success	Failure
1	Drug	6	4
	Placebo	2	8
2	Drug	4	3
	Placebo	1	5
3	Drug	5	3
	Placebo	3	6

Данные вводятся вручную

Задания:

- 1) Изобразить графически число успешных и не успешных применений препарата и имитатора препарата в трех клиниках (barplots). Сравнить полученные результаты.
- 2) Проверить наличие статистической зависимости между использованием препарата и головной болью (точный критерий Фишера) и гипотезу условной независимости (СМН-test)
- 3) Перевести данные в формат, пригодный для анализа с использованием GLM.
- 4) С использованием модели логистической регрессии описать зависимость наличия головной боли от использования препарата и от клиники. С использованием AIC выбрать оптимальную модель логистической регрессии для описания этих данных.
- 5) С использованием логарифмической модели (Пуассона) проверить гипотезы однородности зависимости наличия головной боли от применения препарата и условной независимости наличия головной боли и факта использования препарата.
- 6) Интерпретировать результаты анализа.

## Вариант 2.

Описание эксперимента:

**8.2.** Table 8.8 displays primary food choice for a sample of alligators, classified by length ( $< 2.3$  meters,  $> 2.3$  meters) and by the lake in Florida in which they were caught.

**Table 8.8**

Lake	Size	Primary Food Choice				
		Fish	Invertebrate	Reptile	Bird	Other
Hancock	$\leq 2.3$	23	4	2	2	8
	$> 2.3$	7	0	1	3	5
Olkawaha	$\leq 2.3$	5	11	1	0	3
	$> 2.3$	13	8	6	1	0
Trafford	$\leq 2.3$	5	11	2	1	5
	$> 2.3$	8	7	6	3	5
George	$\leq 2.3$	16	19	1	2	3
	$> 2.3$	17	1	0	1	3

*Source:* Wildlife Research Laboratory, Florida Game and Fresh Water Fish Commission.

Данные вводятся вручную

Задания:

- 1) Изобразить графически процент аллигаторов, предпочтевающих различные типы пищи, с классификацией по размеру и месту обитания (круговые диаграммы). Сравнить полученные результаты. Построить объединенную диаграмму.
- 2) Проверить наличие статистической зависимости между предпочтением определенных типов пищи и размером, для каждого из мест обитания (Хи-квадрат) и гипотезу условной независимости при условии места обитания (CMH-test)
- 3) Перевести данные в формат, пригодный для анализа с использованием GLM.
- 4) С использованием модели логистической регрессии описать зависимость размера аллигатора от предпочитаемого типа пищи и места обитания. С использованием AIC выбрать оптимальную модель логистической регрессии для описания этих данных.
- 5) Используя две группы предпочтений “Рыба” vs. “Другое” описать зависимость выбора предпочтения от размера и места обитания с использованием модели логистической регрессии.
- 6) С использованием логарифмической модели (Пуассона) описать зависимости факторов друг на друга (без группировки и с группировкой из п. 5). В случае без дополнительной группировки выбрать наилучшую модель с использованием AIC. В случае группировки из п.5 проверить однородность влияния размера на предпочтение рыбного типа пищи и размера, при условии места обитания, и проверить условную независимость предпочтительного типа пищи от размера.
- 7) Интерпретировать результаты анализа.

### Вариант 3.

Описание эксперимента:

**3.8.** Table 3.5 refers to the effect of passive smoking on lung cancer. It summarizes results of case-control studies from three countries among nonsmoking women married to smokers. Test the hypothesis that having lung cancer is independent of passive smoking, controlling for country. Report the P-value, and interpret. (Note: Weak associations in observational studies are suspect. With relatively small changes in the data, perhaps representing effects of misclassification or other bias, the association could disappear. See, for instance, R. L. Tweedie et al., Garbage in, garbage out, *Chance*, 7: no. 2, 20–27 (1994)).

**Table 3.5**

Country	Spouse Smoked	Cases	Controls
Japan	No	21	82
	Yes	73	188
Great Britain	No	5	16
	Yes	19	38
United States	No	71	249
	Yes	137	363

Source: Blot and Fraumeni, *J. Nat. Cancer Inst.*, 77: 993–1000 (1986).

Данные вводятся вручную

Задания:

- 1) Изобразить графически число заболевших и здоровых при наличии факта пассивного курения по объединенной выборке и для каждой из трех стран (barplots). Сравнить полученные результаты.
- 2) Проверить наличие статистической зависимости между пассивным курением и заболеванием для каждой из стран (Хи-квадрат) и гипотезу условной независимости (CMH-test)
- 3) Перевести данные в формат, пригодный для анализа с использованием GLM.
- 4) С использованием модели логистической регрессии описать зависимость заболеваемости от курения и от страны проживания. С использованием AIC выбрать оптимальную модель логистической регрессии для описания этих данных.
- 5) С использованием логарифмической модели (Пуассона) проверить гипотезы однородности зависимости заболеваемости от пассивного курения и условной независимости заболеваемости от пассивного курения.
- 6) Интерпретировать результаты анализа.

#### Вариант 4.

Описание эксперимента:

**6.6.** Table 6.13 refers to applicants to graduate school at the University of California at Berkeley for the fall 1973 session. Admissions decisions are presented by gender of applicant, for the six largest graduate departments. Denote the three variables by  $A$  = whether admitted,  $G$  = gender, and  $D$  = department.

**Table 6.13**

Department	Whether admitted, male		Whether admitted, female	
	Yes	No	Yes	No
1	512	313	89	19
2	353	207	17	8
3	120	205	202	391
4	138	279	131	244
5	53	138	94	299
6	22	351	24	317

Source: P. Bickel et al., *Science*, 187: 398–403 (1975).

Данные вводятся вручную

Задания:

- 1) Изобразить графически частоты для двух групп абитуриентов, принятых и не принятых в высшие школы с классификацией по полу (barplot). Сравнить полученные результаты.
- 2) Проверить наличие статистической зависимости между поступлением в школу и полом (Хи-квадрат) и гипотезу условной независимости (CMH-test)
- 3) Перевести данные в формат, пригодный для анализа с использованием GLM.
- 4) С использованием модели логистической регрессии описать зависимость шансов поступления в школу от пола и факультета. С использованием AIC выбрать оптимальную модель логистической регрессии для описания этих данных.
- 5) С использованием логарифмической модели (Пуассона) проверить гипотезы однородности зависимости шансов поступления в школу от пола и условной независимости шансов поступления от пола.
- 6) Интерпретировать результаты анализа.

## Вариант 5.

Описание эксперимента:

- 2.21. Table 2.15 refers to a study that assessed factors associated with women's attitudes toward mammography (Hosmer and Lemeshow, 1989, p. 220). The columns refer to their response to the question, "How likely is it that a mammogram could find a new case of breast cancer?" Analyze these data.

Table 2.15

Mammography Experience	Detection of Breast Cancer		
	Not Likely	Somewhat Likely	Very Likely
Never	13	77	144
Over one year ago	4	16	54
Within the past year	1	12	91

Source: Hosmer and Lemeshow (1989), p. 224. Reprinted with permission of John Wiley & Sons, Inc.

Данные вводятся вручную

Задания:

- 1) Изобразить графически частоты для трех групп пациентов, убежденных в полной неэффективности, частичной эффективности и эффективности метода (barplot).
- 2) Проверить наличие статистической зависимости между фактом прохождения обследования и срока с момента последнего обследования с ответом на вопрос об его эффективности методами классического категориального анализа (Хи-квадрат)
- 3) Перевести данные в формат, пригодный для анализа с использованием GLM.
- 4) Проверить наличие влияния мнения об эффективности метода обследования на проведение хотя бы одного обследования с использованием модели логистической регрессии.
- 5) Проверить наличие влияния мнения об эффективности метода обследования на сроки проведения последнего обследования с использованием логарифмической модели (Пуасона).
- 6) Интерпретировать результаты анализа.

## **Вариант 6.**

Описание эксперимента:

**5.5.** Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, severe forward flexion of the spine following corrective spinal surgery. The age in months at the time of the operation for the 18 subjects for whom kyphosis was present were 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157 and for 22 of the subjects for whom kyphosis was absent were 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206.

Файл данных: "Kypnosis.csv"

Задания:

- 1) Сгруппировать пациентов по возрасту на момент операции в 3 группы – до 5 лет; от 5 до 10 лет и более 10 лет. Представить графически полученные данные (barplot)
- 2) С использованием группировки из п. 1 проверить зависимость шансов возникновения осложнений от возраста методами классической статистики (Хи-квадрат).
- 3) Провести анализ зависимости наблюдаемой переменной от возраста по исходным данным (без группировки) с использованием обобщенных линейной и полиномиальной 2-го порядка регрессионных моделей (GLM модель логистической регрессии). Проверить значимость включения в модель члена 2-го порядка. Проверить значимость влияния возраста на возникновение осложнений. Выбрать лучшую модель с использованием AIC.
- 4) Провести анализ зависимости наблюдаемой переменной от возраста по группированным данным с использованием однофакторного дисперсионного анализа и модели логистической регрессии. Проверить значимость влияния фактора на возникновение осложнений.
- 5) Графически изобразить логистическую регрессию, полученную методами регрессионного анализа и дисперсионного анализа по группированным данным (для дисперсионного анализа значение ассоциировать с серединой интервала группировки).
- 6) Интерпретировать результаты анализа.

## Вариант 7.

Описание эксперимента:

**5.22.** Table 5.15 refers to results of a case-control study about effects of cigarette smoking and coffee drinking on myocardial infarction (MI) for a sample of men under 55 years of age.

**Table 5.15**

Cups Coffee per Day	Cigarettes per Day							
	0		1–24		25–34		≥ 35	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
0	66	123	30	52	15	12	36	13
1–2	141	179	59	45	53	22	69	25
3–4	113	106	63	65	55	16	119	30
≥ 5	129	80	102	58	118	44	373	85

*Source:* L. Rosenberg et al., *Am. J. Epidemiol.*, 128: 570–578 (1988).

Файл данных: "Coffee\_sigar.csv"

Задания:

- 1) Представить графически данные (barplot)
- 2) Проверить зависимость наблюдаемой переменной с интенсивностью курения методами классической статистики (Хи-квадрат).
- 3) Провести анализ зависимости наблюдаемой переменной с интенсивностью курения данных с использованием GLM (Пуассоновская модель).
- 4) Проверить зависимость наблюдаемой переменной от двух факторов – интенсивности курения и интенсивности употребления кофе (Пуассоновская модель). Является ли значимым взаимодействие факторов. Выбрать оптимальную модель с использованием AIC и BIC.
- 5) Интерпретировать результаты анализа.

## Вариант 8.

Описание эксперимента:

- 5.1.** For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, Table 5.10 shows the temperature ( $^{\circ}\text{F}$ ) at the time of the flight and whether at least one primary O-ring suffered thermal distress.

**Table 5.10**

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			

*Note:* Ft = flight no., Temp = temperature, TD = thermal distress (1 = yes, 0 = no).

*Source:* Data based on Table 1 in S. R. Dalal, E. B. Fowlkes, and B. Hoadley. *J. Amer. Statist. Assoc.*, 84: 945–957 (1989). Reprinted with permission of the American Statistical Association.

Файл данных: "Distress.csv"

Задания:

- 1) Разбить значения температур на 3 группы и изобразить графически наличие TD при различных температурах ( bargplot ).
- 2) Проверить зависимость наличия TD от температуры (группировка из п. (1)) методами классического категориального анализа (Хи-квадрат )
- 3) Проверить зависимость наличия TD от температуры без использования группировки с использованием GLM (модель логистической регрессии). Изобразить графически полученную зависимость вероятности TD от температуры.
- 4) Проверить зависимость наличия TD от температуры с группировкой из п. (1) с использованием GLM. Сделать это двумя способами: с использованием модели логистической регрессии и с использованием пуассоновской модели.
- 5) Интерпретировать результаты анализа

## Вариант 9.

Описание эксперимента:

4.5. Table 4.6 refers to a sample of subjects randomly selected for an Italian study on the relation between income and whether one possesses a travel credit card (such as American Express or Diners Club). At each level of annual income in millions of lira, the table indicates the number of subjects sampled and the number of them possessing at least one travel credit card. Analyze these data.

Table 4.6

Income	Number Cases	Credit Cards	Income	Number Cases	Credit Cards
24	1	0	48	1	0
27	1	0	49	1	0
28	5	2	50	10	2
29	3	0	52	1	0
30	9	1	59	1	0
31	5	1	60	5	2
32	8	0	65	6	6
33	1	0	68	3	3
34	7	1	70	5	3
35	1	1	79	1	0
38	3	1	80	1	0
39	2	0	84	1	0
40	5	0	94	1	0
41	2	0	120	6	6
42	2	0	130	1	1
45	1	1			

Source: *Categorical Data Analysis*, Quaderni del Corso Estivo di Statistica e Calcolo delle Probabilità, n. 4., Istituto di Metodi Quantitativi, Università Luigi Bocconi, a cura di R. Piccarreta.

Файл данных: "Credit.csv"

Задания:

- 1) Разбить уровень доходов на 3 группы и изобразить графически частоты наличия хоть одной карты при различных уровнях дохода (barplot).
- 2) Проверить зависимость наличия хоть одной карты от уровня дохода (группировка из п. (1)) методами классического категориального анализа (Хи-квадрат)
- 3) Проверить зависимость наличия хоть одной карты от уровня дохода без использования группировки с использованием GLM (модель логистической регрессии). Изобразить графически полученную зависимость вероятности наличия хоть одной карты от уровня дохода.
- 4) Проверить зависимость наличия хоть одной карты от уровня дохода с группировкой из п. (1) с использованием GLM. Сделать это двумя способами: с использованием модели логистической регрессии и с использованием пуассоновской модели.
- 5) Интерпретировать результаты анализа.

## Вариант 10.

**8.18.** Table 8.13 refers to a study in which subjects were randomly assigned to a control group or a treatment group. Daily during the study, treatment subjects ate cereal containing psyllium. The purpose of the study was to analyze whether this had a desirable effect in lowering LDL cholesterol.

**Table 8.13**

Beginning	Ending LDL Cholesterol Level							
	Control				Treatment			
	$\leq 3.4$	3.4–4.1	4.1–4.9	> 4.9	$\leq 3.4$	3.4–4.1	4.1–4.9	> 4.9
$\leq 3.4$	18	8	0	0	21	4	2	0
3.4–4.1	16	30	13	2	17	25	6	0
4.1–4.9	0	14	28	7	11	35	36	6
> 4.9	0	2	15	22	1	5	14	12

Source: Dr. Sallee Anderson, Kellogg Co.

Задания:

- 1) Изобразить графически процент наблюдений, имеющих различные уровни холестерина, с классификацией по начальному уровню холестерина и проведению лечения (круговые диаграммы). Сравнить полученные результаты. Построить объединенную диаграмму. Интерпретировать полученные результаты.
- 2) Построить график зависимости уровней холестерина в начальный и конечный моменты времени с классификацией по проведению лечения. Визуально оценить изменение уровня холестерина при наличии и при отсутствии лечения.
- 3) Сформировать данные с группировкой “невысокий уровень холестерина” ( $\leq 4.1$ ) и “высокий уровень холестерина” ( $> 4.1$ ). Проверить зависимость уровня холестерина у пациентов, проходивших, и не проходивших лечение, для групп с невысоким и высоким начальным уровнем холестерина. Для каждой из групп проверить гипотезу независимости уровня холестерина в конечный момент времени от проведения лечения (точный критерий Фишера). Проверить гипотезу условной независимости конечного уровня холестерина от лечения при условии начального уровня холестерина (CMH-test)
- 4) Перевести данные в формат, пригодный для анализа с использованием GLM.
- 5) С использованием группировки п.2 и модели логистической регрессии описать зависимость уровня холестерина в конечный момент времени от начального уровня холестерина и лечения. С использованием AIC/BIC выбрать оптимальную модель логистической регрессии для описания этих данных.
- 6) С использованием логарифмической модели (Пуассона) описать зависимости факторов друг на друга (без группировки и с группировкой из п. 2). Сформулировать и проверить гипотезы однородности зависимости и условной независимости. При наличии группировки сравнить результаты с п.2. В случае без дополнительной группировки выбрать наилучшую модель с использованием AIC/BIC. Интерпретировать результаты анализа.