

Подготовка данных. В качестве тестового примера рассмотрим данные о 280 наркоманах, проходивших курс лечения амбулаторно и в диспансере. Состояние каждого из индивидов оценивалось по ряду признаков, расшифровка которых представлена в следующей таблице.

	Кодировка	расшифровка признака
1	<i>prcod</i>	лечебный препарат: 1 – налтрексон и плацебо, 2 – плацебо и плацебо, 3 – налтрексон и антидепрессант, 4 – плацебо и антидепрессант
2	<i>intpla</i>	место лечения: 1 – амбулаторно, 2 – в диспансере
3	<i>sex</i>	пол: 1 – женский, 0 – мужской
4	<i>age</i>	возраст
5	<i>educat</i>	образование: 1 – 8 классов, 2 – средняя школа, 3 – неполное высшее, 4 – высшее
6	<i>curwor</i>	занятость: 0 – не работает, 1 – работает
7	<i>asilmed</i>	индексы зависимости тяжести: медицинский статус
8	<i>asi2emp</i>	социальный статус (занятость, заработок)
9	<i>asi3alc</i>	употребление алкоголя
10	<i>asi4dr</i>	употребление наркотиков
11	<i>asi5leg</i>	проблемы с законом
12	<i>asi6fam</i>	семейные отношения
13	<i>asi7psy</i>	психиатрический статус
14	<i>asid3 dyr</i>	длительность героиновой зависимости
15	<i>tlfba2</i>	обычная доза приема алкоголя (гр)
16	<i>tlfbh2</i>	обычная доза употребления героина (мг)
17	<i>st</i>	использование стимулирующих средств
18	<i>ha</i>	использование препаратов, вызывающих галлюцинации
19	<i>se</i>	использование успокоительных средств
20	<i>cravin</i>	собственная оценка необходимой дозы героина
21	<i>rabdru</i>	оценка риска в употреблении наркотиков
22	<i>rubsex</i>	оценка риска в сексуальных связях
23	<i>gaf</i>	оценка общего функционирования
24	<i>bdi</i>	оценка депрессии
25	<i>sstati</i>	оценка уровня тревожности
26	<i>end</i>	итог лечения: 1 – программа лечения выполнена, 0 – программа сорвана

Данные для статистического анализа обычно представляют в виде матрицы наблюдений. Столбцы этой матрицы обозначают признаки, а строки – индивиды. Например, для первых семи признаков и трех индивидов матрица наблюдений имеет вид:

	prcod	intpla	sex	age	educat	curwor	asil med
1	4	1	0	18	1	1	0.19
2	2	2	1	30	4	1	0.44

Индивид, порядковый номер которого равен 2, является женщиной в возрасте 30 лет, которая во время лечения в диспансере получала две таблетки плацебо. У нее высшее образование, есть работа. Медицинский индекс тяжести зависимости равен 0.44 и так далее.

Если матрица наблюдений подготовлена в Электронных Таблицах и сохранена в текстовом формате под именем "наркоманы.txt", то для импорта данных в программу *BOA Statistique* необходимо выбрать команду "Файл" - "Загрузить", затем выбрать из каталога нужный файл под названием "наркоманы".

1. Исследование зависимости между категориальными признаками. Для исследования взаимосвязи, например, между занятостью и образованием наркоманов, находим в меню раздел "Анализ – Таблицы сопряженности".

В одном из двух списков признаков выделяем признак *educat* (образование), в другом *curwor* (занятость), щелкнув мышкой по названиям. Отмечаем птичками "хи-квадрат" и "коэффициент неопределенности". Кнопка "Вычислить". Результатом вычислений представлены в виде таблиц:

Табл.1.1

	<i>curwor</i> 0	<i>curwor</i> 1
<i>educat</i> 1	12	9
<i>educat</i> 2	170	52
<i>educat</i> 3	17	11
<i>educat</i> 4	5	4

В таблице сопряженности (Табл.1.1) столбцы и строки подписаны в соответствии с градациями выбранных признаков. Например, на пересечении второго столбца и третьей строки значение 11 означает, что 11 человек с незаконченным высшим образованием заняты на работе или учебе.

Табл.1.2

Хи-квадрат	Статистика	Степени свободы	Значимость
	7.625	3	0.054

В табл.1.2 значимость критерия равна $p = 0.054$, то есть с уровнем значимости $\alpha = 0.05$ нет оснований отвергнуть гипотезу о независимости этих признаков. Однако близкое расположение к стандартному уровню значимости $\alpha = 0.05$ свидетельствует о том, что распределение по занятости в группах с различным уровнем образования далеко не одинаково.

Кнопка "Частоты для *educat*" превращает таблицу сопряженности в таблицу пропорций (Табл.1.4) по признаку *curwor* при разных уровнях образования *educat*.

Табл.1.4

	<i>curwor</i> 0	<i>curwor</i> 1
<i>educat</i> 1	57%	43%
<i>educat</i> 2	77%	23%
<i>educat</i> 3	61%	39%
<i>educat</i> 4	56%	44%

Из этой таблицы видно, что среди всех наркоманов, кроме тех, кто имеет среднее образование, наблюдается примерно одинаковое количество занятых и безработных, а среди окончивших среднюю школу процент занятых почти в два раза ниже. Вывод можно сформулировать так: *с уровнем значимости большим 0.054 между занятостью и уровнем образования среди наркоманов имеет место зависимость, связанная со значимо меньшим количеством работающих наркоманов с законченным средним образованием.*

Табл.1.3

Коэффициент неопределенности	educat	curwor	Двусторонний
	1.78	2.19	1.96

Для измерения степени этой зависимости используем значения коэффициентов неопределенности (Табл.1.3). По данным о занятости мы получаем 1.78% информации об уровне образования. По данным об уровне образования мы получаем 2.19% информации о занятости. Двусторонний коэффициент неопределенности равен 1.96% и означает усредненную долю совместной информации.

2.1. Фильтр и описательные статистики.

Если в базе данных нужно найти статистические характеристики какой-то категории индивидов, например, уровень тревожности у 170 безработных наркоманов со средним образованием, то для этого в меню "Данные"—"Фильтры" следует указать значения признаков $curwor=0$ и $educat=2$. Кнопка "Применить".

В меню "Анализ"—"Описательные статистики" выбираем признак $sstat1$ (тревожность). Находим значения среднего и ошибки среднего: 49.09 ± 0.71

Для сравнения этой группы с работающими после средней школы наркоманами в меню "Данные"—"Фильтры" следует изменить значения признаков на $curwor=1$ и $educat=2$. Через пункты "Анализ"—"Описательные статистики" признак $sstat1$ (тревожность) в этой группе характеризуется показателями 47.38 ± 1.06 .

2.2. Проверка гипотез однородности. Выясним, насколько значима разница в показателях тревожности у работающих и неработающих наркоманов с законченным средним образованием. В меню "Данные"—"Фильтры" указываем на выбор значения признака $educat=2$ и используем параметрические критерии. Меню: "Анализ"—"Проверка гипотез однородности"—"для двух независимых выборок"—"Критерий Стьюдента". В "Параметрах" в качестве зависимой переменной указываем $DEGREE$, в качестве группирующей $STATE$, отмечаем, что нас интересуют группы 2 и 3. Кнопка "Вычислить".

Табл.2.1.

Сравнение групп			
Группа	Среднее	Ст.откл.	Кол-во набл.
$curwor = 0$	49.09	9.25	170
$curwor = 1$	47.38	7.65	52

Табл.2.2.

Тест Фишера на равенство дисперсий

Статистика F	Ст.св.	Значимость
1.4626	169; 51	0.1142

В табл.2.2 представлены результаты проверки гипотезы равенства дисперсий. Значимость $p = 0.1142 > 0.05$, поэтому нет оснований отвергнуть гипотезу о равенстве дисперсий. В связи с этим для проверки гипотезы о равенстве средних используется критерий Стьюдента в предположении одинаковых дисперсий.

Табл.2.3.

Критерий Стьюдента (в предположении равенства дисперсий)

Статистика F	Ст.св.	Значимость
1.2117	220	0.2269

Значимость $p = 0.2269 > 0.05$, поэтому нет оснований отвергнуть гипотезу о равенстве средних, поэтому различие в средних уровнях тревожности этих групп отличаются незначимо.

Для непараметрического критерия однородности используем путь: "Анализ"—"Проверка гипотез однородности"—"для двух независимых выборок"—"Критерий Манна-Уитни".

Табл.2.4.

Сравнение групп				
Группа	Минимум	Медиана	Максимум	Кол-во набл.
$curwor = 0$	23	48	71?170	
$curwor = 1$	31	48	63?52	

Табл.2.5.

Критерий Манна-Уитни

Сумма рангов $curwor = 0$	Сумма рангов $curwor = 1$	Статистика U	Значимость p
19360	5394	4825	0.3183

Значимость критерия Манна-Уитни равна $p = 0.3183 > 0.05$, предыдущему критерию это не противоречит. **3.1.**

Значимость коэффициента корреляции. Для построения корреляционной матрицы признаков, относящихся к индексам тяжести ($asi1, asi3-asi5$), находим в меню раздел "Анализ"—"Коэффициенты корреляции". В поле "Исследуемые признаки" ставим птички на признаки $asi1, asi3-asi5$. Кнопка "Вычислить".

Наибольшая положительная корреляция наблюдается между индексами тяжести $asi3_alc$ (употребление алкоголя) и $asi4_dr$ (употребление наркотиков), $R = 0.33, P < 10^{-4}$. Наибольшая отрицательная корреляция наблюдается между индексами тяжести $asi1_med$ (медицинский статус) и $asi5_leg$ (проблемы с законом), $R = -0.1955, P = 0.001$.

Особенно эта отрицательная корреляция проявляется в группе неработающих ($curwor = 0$) и принимающих успокоительные средства ($se = 1$). Для неработающих наркоманов, которые принимают успокоительные средства, "проблемы с законом" отрицательно коррелируют с "медицинским статусом": $R = -0.29, P = 0.036$, что означает: чем хуже здоровье, чем меньше проблем с законом.