1 Подготовка и ввод данных

В качестве тестового примера рассмотрим данные о 1600 служащих одной уважаемой фирмы. Целью исследования являлось выявление факторов метаболического синдрома. Состояние каждого из индивидов оценивалось по ряду признаков, расшифровка которых представлена в следующей таблице.

	Кодировка	расшифровка признака
1	No	номер по порядку
2	gen	пол
3	ht	рост
4	wt	вес
5	imb	имт
6	age	возраст
7	cirw	ОТ
8	cirth	ОБ
9	cirn	ОШ
10	hadm	ожирение
11	smk	курение
12	alc	алкоголь
13	sport	спорт
14	iad	информированность об АД
15	iob	физический труд
16	snore	храп
17	apnoe	задержка дыхания во сне
18	sap	СД
19	dap	ДД
20	hr	ЧСС
21	chol	холестерин
22	hdlp	ЛПВП
23	tg	триглицериды
24	AC	коэф.атерогенности
25	glc1	глюкоза в крови
26	dSD	сахарный диабет
27	cidf0	MC1
28	catp0	MC2
29	cwho0	MC3
30	beer	пиво
31	wine	вино
32	spirits	водка
33	money	зарплата

Данные для статистического анализа обычно представляют в виде матрицы наблюдений. Столбцы этой матрицы обозначают признаки, а строки – индивиды.

Если матрица наблюдений подготовлена в Электронных Таблицах и сохранена в текстовом формате под именем „MC.txt“, то для импорта данных в программу

BOA Statistique

необходимо в выбрать команду "Файл"-"Загрузить", затем выбрать из каталога нужный файл.

2 Фильтр

Если в базе данных нужно найти статистические характеристики какой-то категории индивидов, например, курящих мужчин, то для этого в меню "Данные"–"Фильтры" следует указать значения признаков gen=0 и smk=2 и нажать кнопку "Применить".

3 Описательные статистики

Если в базе данных нужно оценить статистическую характеристику определенной категории индивидов, например, вычислить средний уровень ЛПВП у индивидов, занимающихся спортом и употребляющих алкоголь, то для этого в меню "Данные"–"Фильтры" следует указать значения признаков sport=2 и alc=2. Кнопка "Применить".

В меню "Анализ"-"Описательные статистики" выбираем признак hdlp. Находим значения среднего и ошибки среднего: 1.53 ± 0.04 Действительно, эти значения выше нормы, которая близка к 1.

4 Таблицы сопряженности

Для исследования взаимосвязи, например, между абдомиальным ожирением и храпом у мужчин, находим в меню раздел "Анализ – Таблицы сопряженности".

В одном из двух списков признаков выделяем признак hadm, в другом snore, щелкнув мышкой по названиям. Отмечаем птичками "хи-квадрат"и "коэффициент неопределенности". Кнопка "Вычислить". Результаты вычислений представлены в виде таблиц:

Табл.1.1

hadm	*snore*0	*snore*1	*RowTotals*
0	74	53	127
1	64	85	149
	138	138	276

В таблице сопряженности (Табл.1.1) столбцы и строки подписаны в соответствии с градациями выбранных признаков.

Табл.1.2

Хи-квадрат

Статистика	Степени свободы	Значимость
6.43	1	0.01

В табл.1.2 значимость критерия равна $p = 0.01$, то есть с уровнем значимости $\alpha = 0.05$ гипотеза о независимости этих признаков отвергается.

Кнопка "Частоты для hadm" превращает таблицу сопряженности в таблицу пропорций (Табл.1.3) по признаку snore при разных уровнях образования hadm.

Табл.1.3

hadm	*snore*0	*snore*1
0(%)	58,27%	41,73%
1(%)	42,95%	57,05%

Дополнительно используя информацию через "Частоты для snore", получаем, что индивиды с абдомиальным ожирением храпят в 57% случаев, а среди всех храпящих мужчин 62% имеют абдомиальное ожирение. Доверительный уровень вероятности $p = 0.01$ указывет на то, что эти частоты значимы.

Табл.1.4

Коэффициент неопределенности

hadm	snore	Двусторонний
1.7	1.69	1.69

Для измерения степени этой зависимости используем значения коэффициентов неопределенности (Табл.1.4). По данным о храпе мы получаем 1.7% информации об ожирении. По данным об ожирении мы получаем 1.69% информации о храпе. Двусторнний коэффициент неопределенности равен 1.96% и означает усредненную долю совместной информации.

5 Проверка гипотез однородности

Выясним, что является более значимым для повышения ЛПВП: занятия спортом на свежем воздухе или употребление алкоголя. Отдельно будем рассматривать мужчин и жешщин. Если выбираем мужчин, то в меню "Данные"–"Фильтры" указывая

$$gen = 0.$$

Используем сначала параметрические критерии.

"Анализ"

– "Проверка гипотез однородности"

– "для двух независимых выборок"

– "Критерий Стьюдента".

Исследуемый признак hdlp, группирующая переменная sport, группы 0 (не занимаются спортом) и 2 (занимаются интенсивно спортом).

Табл.2.1.

Сравнение групп

Группа	Среднее	Ст.откл.	Кол-во набл.
$sport = 0$	1.30	0.38	147
$sport = 2$	1.31	0.38	91

Табл.2.2.

Тест Фишера на равенство дисперсий

Статистика F	Ст.св.	Значимость
1.0153	146; 90	0.9481

В табл.2.2 представлены результаты проверки гипотезы равенства дисперсии. Значимость $p = 0.9481 > 0.05$, 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 T	Ст.св.	Значимость
−0.0692	236	0.9449

Значимость $p = 0.9449 > 0.05$, поэтому нет оснований отвергнуть гипотезу о равенстве средних, поэтому различие в средних уровнях ЛПВП этих групп отличаются незначимо.

В случае. когда значимость статистики Фишера равенства дисперсий отвергается, то для проверки гипотезы о равенстве средних используется критерий Стьюдента **без предположения равенства дисперсий**.

При небольших объемах выборок рекомендуется использовать непараметрический критерий Манна-Уитни. Например, при проверке значимости влияния употребления алкоголя на ЛПВП у мужчин, которые спортом не занимаются, оказывается небольшой группа вообще непьющих мужчин, которые не занимаются спортом. В фильтре вносим изменения: gen=0, sport=0.

"Анализ"

– "Проверка гипотез однородности"

– "для двух независимых выборок"

– "Критерий Манна-Уитни".

Табл.2.4.

Сравнение групп

Группа	Минимум	Медиана	Максимум	Кол-во набл.
$alk = 0$	0.5	1.2	1.41	6
$alk = 2$	0.79	1.26	2.33	52

Табл.2.5.

Критерий Манна-Уитни

Сумма рангов $alk = 0$	Сумма рангов $alk = 2$	Статистика U	Значимость p
114	1539	93	0.1187

Значимость критерия Манна-Уитни равна $p = 0.1187 > 0.05$, $p = 0.1187 < 0.2$. При такой значимости говорят, что ситуация не определенная, которая может измениться при увеличении объема выборки как в ту, так и другую сторону. Поэтому зачастую используют более слабое выражение, говоря, что имеет место тенденция повышения ЛПВП при употреблении алкоголя у не занимающихся спортом мужчин.

Несложно убедиться, что занятия спортом способствуют повышению ЛПВР у женщин: у не занимающихся спортом ЛПВП в среднем равны 1.56 ± 0.47, у интенсивно занимающихся спортом женщин 1.67 ± 0.46 ($p = 0.0028$ по критерию Стьюдента). Влияние алкоголя на ЛПВП у женщин оказывается незначимым: 1.59 ± 0.48 у совсем не пьющих женщин и 1.66 ± 0.45 у употребляющих часто алкоголь женщин ($p = 0.15$ по критерию Стьюдента).

6 Построение гистограммы

Для категориальных признаков строятся столбиковые гистограммы, в которых для каждой градации признака вычисляется количество соответствующее индивидов. Например, из столбиковой гистограммы для признака sport получаем, что половина индивидов спортом не занимается, а оставшаяся половина делится примерно пополам на тех, кто занимается редко, и тех, кто занимаетсяспортом систематически.

Гистограммы для количественных признаков строятся из расчета заданного количества интервалов. По умолчанию их задается 10. Например, из гистограммы признака cirw (обхват талии) видно, что мода (наиболее часто встречающееся значение, 425 из 1600) находится в пределах от 72 до 79 см.

7 Корреляционный анализ

Для построения корреляционной матрицы для количественных признаков очищаем фильтр, находим в меню раздел

„Анализ“ – „Коэффициенты корреляции“.

Ставим птички в полях „Исследуемые признаки“ и нажимаем кнопку „Вычислить“. Значимыми оказываются корреляции, для которых доверительный уровень вероятности меньше 0.05. Например, у женщин корреляция между индексом массы тела (ИМТ) и ОТ, ОБ и ОШ равны соответственно 0.9, 0.87 и 0.62, а у мужчин 0.9, 0.73 и 0.67. Все корреляции значимо больше нуля, и, естественно, у женщин с ИМТ больше корреляция ОБ, а у мужчин ОШ.