Лекция 14

Критерий χ^2 (Хи-квадрат) и его применение при проверке гипотез

Распределение χ^2 (Хи-квадрат) Пирсона уже использовалось нами при построении интервальных оценок (доверительных интервалов) для дисперсии. В этой лекции представлено несколько примеров использования этого распределения при проверке гипотез.

Гипотеза о значении дисперсии

Дисперсия является величиной, характеризующей отклонение наблюдаемого ряда значений от средней величины. Проверка гипотез о дисперсии довольно часто связана с анализом качества продукции. Примером может служить дисперсия размера деталей (например, диаметр болтов), дисперсия времени службы аккумуляторов, дисперсия расхода топлива у автомобилей одной модели. Чем меньше дисперсия, тем более устойчивым является производственный процесс, тем более однородную продукцию выпускает предприятие, тем меньше вероятность недопустимых отклонений. Как правило, ограничения на дисперсию закладываются в спецификацию продукции, а выборочное наблюдение организуется с целью выяснить (проверить), укладываются ли реальные изделия в эти рамки или нет.

Формулировки гипотез. Основная гипотеза утверждает, что генеральная дисперсия D равна некоторому предполагаемому значению $H_0 : D = D_0$, альтернативная - что дисперсия не равна этому значению, больше или меньше его, $H_1 : D \neq D_0;\ D > D_0; D < D_0$.

Выборочная информация. Выборочная совокупность при выполнении данного типа исследований характеризуется величинами n - длина выборки s^2 - выборочная дисперсия.

Критерий и условия применимости. Критерием для принятия решения является величина

$$\chi^2 = \frac{(n-1)\cdot s^2}{D_0},$$

подчиняющаяся распределению χ^2 (Хи-квадрат) с числом степеней свободы $k = n-1$ при условии, что генеральная совокупность приближенно описывается нормальным распределением.

Пример. Производитель сигарет хочет проверить, действительно ли среднее квадратическое отклонение содержания никотина в сигаретах составляет 0.8 миллиграмма. Отклонение, рассчитанное для выборки из 20 сигарет, оказалось равным 1г. Проверить гипотезу производителя сигарет при уровне значимости 0.05.

Решение. При решении задачи следует учитывать, что в формуле критерия фигурируют генеральная и выборочная *дисперсии,* тогда как в условиях задачи указаны соответствующие средние квадратические отклонения.

1) Формулируем основную и альтернативную гипотезы:

$$H_0 : D = D_0 = 0.8^2 = 0.64 :\ H_1 : D \neq 0.64 .$$

2) Для проверки гипотезы будет использоваться критерий χ^2 с числом степеней свободы $k = n-1 = 20-1 = 19$. Так как проверяется двусторонняя альтернативная гипотеза, то для заданного уровня значимости $\alpha = 0.05$ нужно по таблице распределения найти два критических значения: $\chi^2_{\alpha/2,\,н} = 8.9$, $\chi^2_{\alpha/2,\,в} = 32.9$. Двусторонняя критическая область, при попадании в которую основная гипотеза отклоняется, имеет вид $\chi^2 < 8.9 \cup \chi^2 > 32.9$, рис. 1.

3) Находим расчетное значение критерия

$$\chi^2 = \frac{(n-1)\cdot s^2}{D_0} = \frac{(20-1)\cdot 1^2}{0.64} \approx 29.7 .$$

4) Принятие решения. Полученное значение находится в некритической области, рис. 1, основная гипотеза принимается.

5) Вывод: Данные выборочного наблюдения подтверждают предположение

производителя о величине среднего квадратического отклонения содержания никотина в сигаретах.

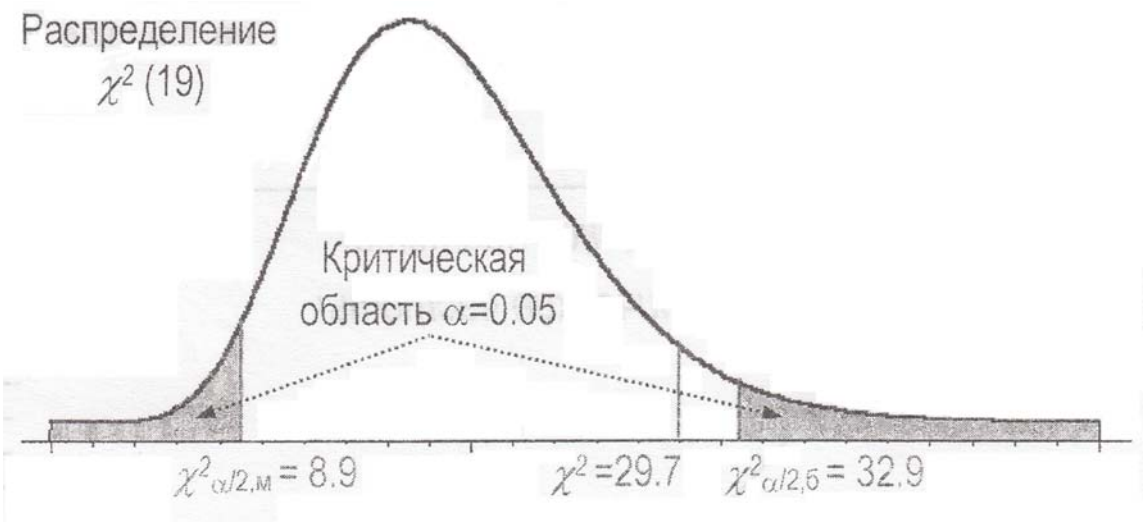

Рис. 1. *Проверка гипотезы о дисперсии содержание никотина*

Пример. Для контроля над расходом топлива отобраны 16 автомобилей. Среднее квадратическое отклонение расхода топлива составило 0.6 литров/100 км. Проверить при уровне значимости 0.01 гипотезу, что генеральное среднее квадратическое отклонение превышает 0.4 литра.

Решение:

1) Основная и альтернативная гипотеза формулируются следующим образом:

$$H_0 : D = D_0 = 0.4^2 = 0.16 : \; H_1 : D > 0.16 .$$

2) Так как альтернативная гипотеза является односторонней (левой), то ищется только правое (большее) критическое значение по распределению χ^2 (Хи-квадрат) с числом степеней свободы 15: $\chi^2_\alpha = 30.58$, а критическая область $\chi^2_\alpha > 30.58$, рис. 2.

3) Используя формулу для расчета критерия по выборочным значениям, находим:

$$\chi^2 = \frac{(n-1)\cdot s^2}{D_0} = \frac{(16-1)\cdot 0.6^2}{0.16} \approx 33.75 .$$

4) Принятие решения. Расчетное значение критерия попадает в

критическую область, рис. 2, поэтому основная гипотеза отклоняется.

5) Вывод: Выборочные данные согласуются с утверждением, что отклонение по расходу топлива у различных автомобилей превышает 0.4 литра на 100 км.

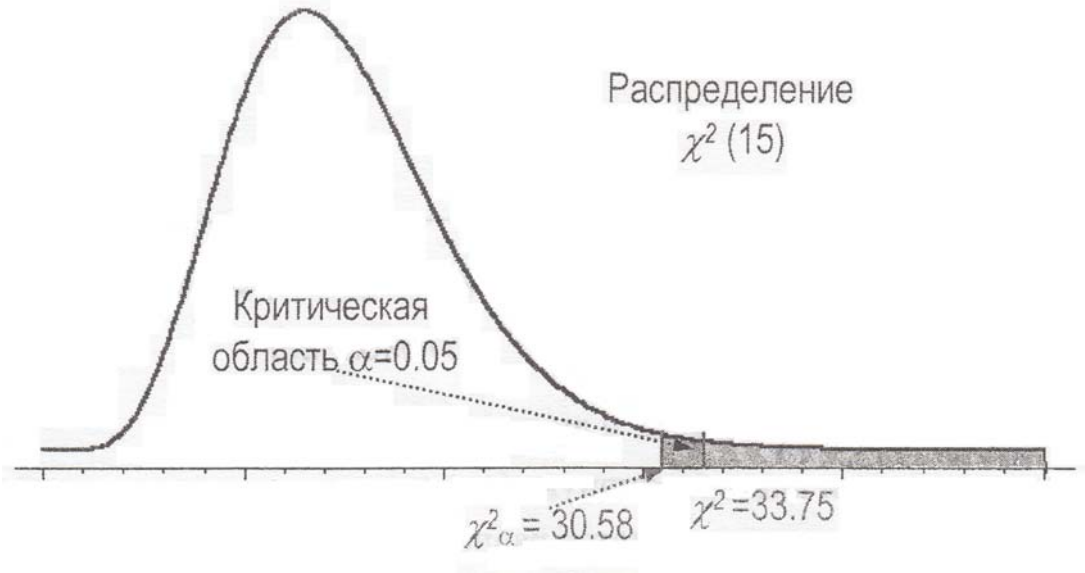

Рис. 2. *Проверка гипотезы о дисперсии расхода топлива*

Гипотеза о равномерном распределении.

Типичной задачей статистики является проверка соответствия полученного в статистическом исследовании (эмпирического) распределения одному из стандартных распределений. Решение этих вопросов также основано на аппарате проверки гипотез и использовании критерия χ^2. Вопрос о соответствии эмпирического частотного распределения ***равномерному*** возникает, например, в следующих ситуациях:

- Производитель прохладительных напитков хочет установить, одинаковой ли популярностью пользуются выпускаемые виды напитков, с целью оптимизации объемов производства.
- Инженер по безопасности движения хочет установить, одинаковое ли количество аварий происходит на дороге в различные дни, или в какие-то дни контроль за движением транспорта должен быть усилен.
- Управляющий станцией скорой помощи хочет знать, равномерно ли

распределяются вызовы в течение дня, с целью оптимизации количества дежурного персонала.

Рассмотрим схему исследования на примере первой из упомянутых ситуаций. Предположим, что компания производит напитки 5 типов. Выборочный опрос 100 покупателей показал, что предпочтения распределились следующим образом, табл. 1, первая и вторая колонка:

Таблица 1. *Проверка гипотезы о равномерном распределении*

Напиток	***Частота*** f_i ***(эмпирическая)***	***Теоретическая частота*** f_i^T
Вишневый	32	20
Клубничный	28	20
Апельсиновый	16	20
Лимонный	14	20
Грейпфрутовый	10	20
Всего	100	100

Теоретические частоты, находящиеся в третей колонке, можно проинтерпретировать следующим образом: они показывают, как распределились бы по группам 100 покупателей, если бы все напитки пользовались равным спросом. Очевидно, что это число равняется общему количеству покупателей, деленному на число различных напитков (число групп) K: $f_i^T = \frac{n}{K} = \frac{100}{5} = 20$.

Полигоны эмпирических и теоретических частот представлены на рис. 3. Проверка гипотез о виде распределения и в этом, и в более сложных случаях основана на анализе суммы квадратов отклонений эмпирических и теоретических частот.

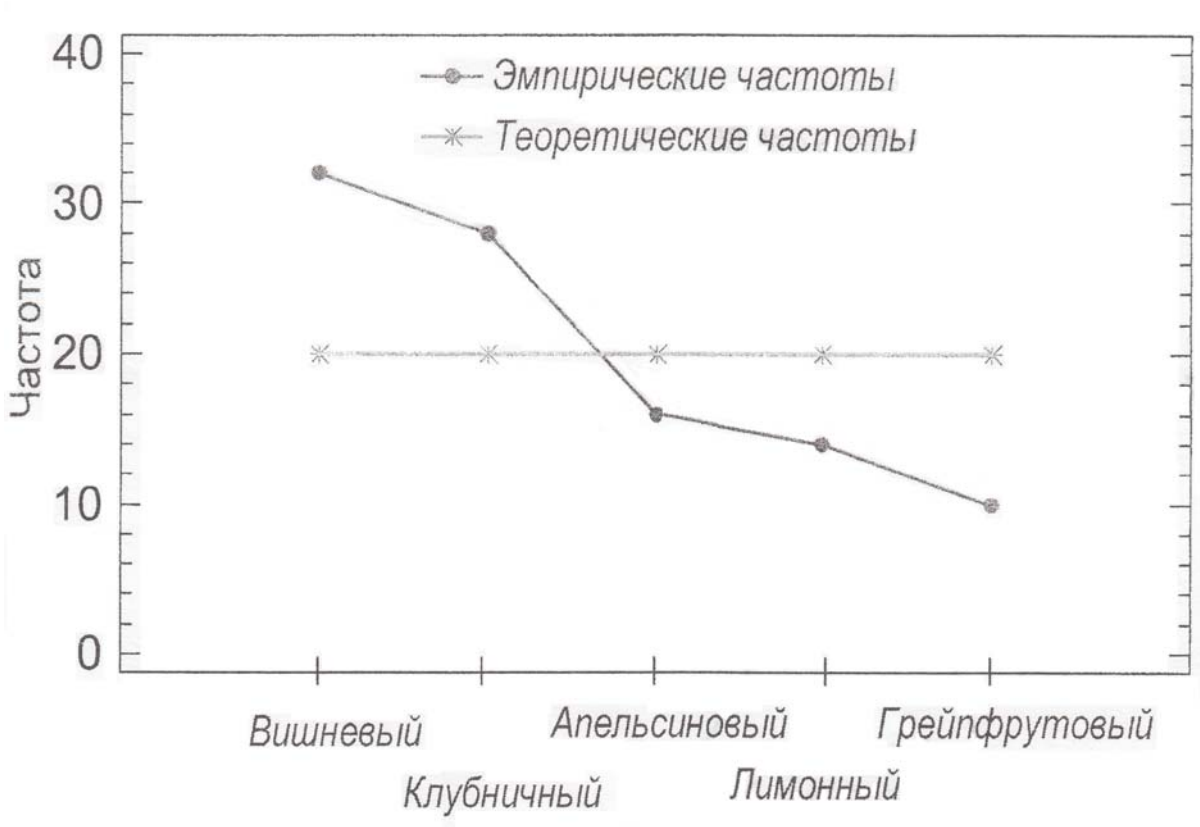

Рис. 3 *Соотношение эмпирических и теоретических частот*

Сформулируем основную и альтернативную гипотезы:

H_0 : Частотное распределение является равномерным;

H_1 : Частотное распределение не является равномерным.

Важно отметить, что даже если вкусы по генеральной совокупности в целом распределены равномерно, для каждой конкретной выборки частоты будут отклоняться от теоретических в силу случайных факторов. Для принятия решения о том, являются ли отклонения частот столь большими, чтобы при заданном уровне значимости можно отклонить гипотезу о равномерности, используется критерий:

$$\chi^2 = \sum_{i=1}^{K} \frac{\left(f_i - f_i^T\right)^2}{f_i^T}$$

имеющий распределение χ^2 с числом степеней свободы k, равным количество групп минус один, $k = K - 1$. Выполняется проверка правой односторонней гипотезы.

Пример. Проверить с помощью критерия χ^2 при уровне значимости 0.05 гипотезу о равномерной распределенности пристрастий покупателей фруктовых напитков.

Решение:

1) H_0 : Вкусы покупателей распределены равномерно;

H_1: Вкусы покупателей распределены неравномерно.

2) Так как проверяемая гипотеза является односторонней правой, то находим критическое значение χ^2_α для распределения с числом степеней свободы 5-1=4; $\chi^2_\alpha = 9.488$.

Критической областью (основная гипотеза отклоняется) является $\chi^2_\alpha > 9.488$, рис. 4.

3) Находим расчетное значение критерия:

$$\chi^2 = \sum_{i=1}^{K} \frac{\left(f_i - f_i^T\right)^2}{f_i^T} = \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(10-20)^2}{20} = 18$$

4) Принятие решения. Расчетное значение попадает в критическую область, 18 > 9.488, рис. 4, основная гипотеза отклоняется.

5) Вывод. Результаты наблюдений подтверждают предположение, что вкусы покупателей распределены неравномерно.

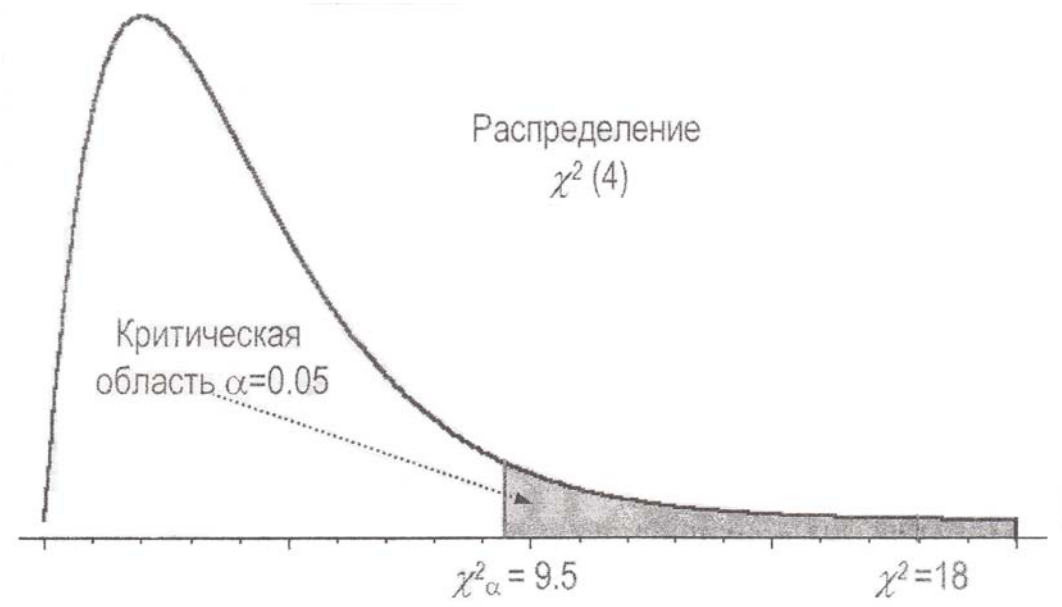

Рис. 4. *Проверка гипотезы о равномерном распределении*

Гипотеза о нормальном распределении

Приближенно нормальное распределение генеральной совокупности является необходимым условием применимости многих статистических методов. Проверить, насколько обоснованно это предположение, можно с помощью гипотезы о виде распределения и критерия χ^2.

Расчет теоретических частот. Аналогично уже рассмотренным гипотезам о равномерном распределении, проверка гипотезы о нормальном распределении основана на сопоставлении эмпирических и теоретических частот. Рассмотрим основные этапы решения этой задачи на примере. В табл. 2., колонки 1 и 2, представлены данные о распределении по росту 300 студентов.

В основе расчета теоретических частот лежит следующий принцип - нужно определить, сколько из общего числа n значений попадет на каждый из интервалов в случае, если генеральная совокупность строго соответствовала бы нормальному распределению с параметрами $\overline{x}$, s. Расчет частот удобно разбить на следующие этапы:

1) По выборочным данным рассчитать среднюю арифметическую $\overline{x}$ и среднее квадратическое отклонение s.

2) Вычислить нормированное отклонение от средней для середин интервалов

$$z_i = \frac{x_i - \overline{x}}{s}.$$

3) Найти значения плотности стандартного нормального распределения $\rho(z_i)$.

4) Вычислить теоретические частоты f_i^T, умножив количество наблюдений n на вероятность попадания значения на интервал в случае нормальной распределенности $\frac{H}{s} \cdot \rho(z_i)$, H - ширина интервала.

Для рассматриваемого примера $\overline{x} = 176.6$, $s^2 = 66.74$, $s = 8.17$. Дальнейшие результаты расчетов представлены в столбцах 4-6, табл. 2. Из-за ошибок округления сумма теоретических частот (297) получилась несколько меньшей истинного значения $n = 300$.

Таблица 2. *Проверка гипотезы о нормальном распределении*

Интервал $[x_i, x_{i+1}]$	Частота f_i	Середина интервала $x_{i,m}$	Нормированное отклонение $z_i = \frac{x_{i,m} - \overline{x}}{s}$	Плотности распределения $\rho(z_i)$	Теоретическая частота $f_i^T = n \cdot \frac{H}{s} \cdot \rho(z_i)$
155-160	8	157.5	-2.34	0.0258	5
160-165	17	162.5	-1.73	0.0893	16
165-170	42	167.5	-1.11	0.2155	40
170-175	54	172.5	-0.5	0.3521	65
175-180	73	177.5	0.11	0.3965	73
180-185	57	182.5	0.72	0.3079	57
185-190	38	187.5	1.33	0.1647	30
190-195	11	192.5	1.95	0.0596	11
Сумма	**300**				297

Проверка гипотезы. Дальнейшая схема проверки гипотез о нормальном

распределении схожа со случаем равномерного распределения. Формулируются две гипотезы:

H_0:Частотное распределение является приближенно нормальным;

H_1:Частотное распределение значительно отклоняется от нормального,

и вычисляется значение критерия $\chi^2 = \sum_{i=1}^{K} \frac{\left(f_i - f_i^T\right)^2}{f_i^T}$, K - число интервалов. Критерий описывается распределением χ^2 с числом степеней свободы $k = K - 3$ и по нему выполняется проверка правосторонней гипотезы.

Пример. Проверить предположение о нормальном распределении роста студентов при уровне значимости 0.05.

1) Формулируем гипотезы:

H_0: Распределение по росту описывается нормальным распределением;

H_1: Распределение по росту значительно отклоняется от нормального.

2) Правое критическое значение χ^2_α определятся по таблице распределения χ^2 с числом степеней свободы $k = K - 3 = 8 - 3 = 5$ при $\alpha = 0.05$, $\chi^2_\alpha = 11.07$.

Критическая область $\chi^2 > 11.07$.

3) Рассчитываем значение критерия:

$$\chi^2 = \sum_{i=1}^{K} \frac{\left(f_i - f_i^T\right)^2}{f_i^T} = \frac{(8-5)^2}{5} + \frac{(17-16)^2}{5} + \frac{(42-40)^2}{40} + \frac{(54-65)^2}{65} +$$
$$+ \frac{(73-73)^2}{73} + \frac{(57-57)^2}{57} + \frac{(38-30)^2}{30} + \frac{(11-1)^2}{11} = 6$$

4) Принятие решения. Расчетное значение критерия не попадает в критическую область 6.0 < 11.07, следовательно, нулевая гипотеза принимается.

5) Вывод: Выборочные данные подтверждают предположение, что распределение студентов по росту является нормальным.

Исследование взаимосвязи атрибутивных признаков

Исследование взаимосвязи между *количественными* признаками проводится в

рамках отдельного раздела статистики - корреляционного и регрессионного анализа, однако перенести применяемые в нем средства на исследование связи между атрибутивными признаками не удается. Эта задача решается с помощью проверки *гипотез о независимости,* основанной на χ^2 - распределении. Рассмотрим схему решения на примере.

Таблицы сопряженности. В крупной клинике предполагают внедрить новую процедуру послеоперационного обслуживания и хотят выяснить, одинаково ли относятся к предполагаемому нововведению врачи и средний медперсонал. Отметим, что вопрос направлен не на выяснение того, нравится им нововведение или нет, а на то, есть ли отличие в восприятии нововведения между двумя группами. Для выявления мнений проводится выборочное исследование, результаты которого представлены в табл. 3.

Таблица 3. *Таблица сопряженности*

Группа	Относится отрицательно	Относятся положительно	Нет определенного мнения
Врачи	10	8	2
Средний медперсонал	5	12	3

Таблица такого типа называется *таблицей сопряженности* и представляет собой средство группировки одновременно по двум атрибутивным признакам. Число строк в таблице равно числу вариантов первого признака, число столбцов - числу вариантов второго, а на пересечении указывается частота - число единиц выборки, у которых признаки совпадают со значением данной строки и столбца. Например, в таблице во второй строке первой колонки указано 5 - столько опрошенных из числа среднего медперсонала отрицательно относятся к нововведению.

Расчет теоретических частот. Проверка гипотез о независимости будет основана на анализе отклонений реальных (эмпирических) значений в таблице от теоретических - значений, которые должны были бы получиться, если бы отношение к процедуре было независимым.

Введем следующие обозначения:

$f_{i,j}$ - частота в i-й строке j-го столбца таблицы; $i = 1,2,\ldots,K_r$, $j = 1,2,\ldots,K_c$.

K_r, K_c - число строк и столбцов (в примере $K_r = 2$, $K_c = 3$);

$n_{r,i}, n_{c,j}$ - суммы значений по строкам и колонкам;

n - общее количество наблюдений.

Тогда алгоритм расчета теоретических значений имеет вид:

1) Рассчитать величину суммы $n_{r,i}, n_{c,j}$

$$n_{r,1} = 10 + 8 + 2 = 20; \quad n_{r,2} = 5 + 12 + 3 = 20;$$

$$n_{c,1} = 10 + 5 = 15; \quad n_{c,2} = 8 + 12 = 20; \quad n_{c,3} = 2 + 3 = 5;$$

$$n = n_{r,1} + n_{r,2} = n_{c,1} + n_{c,2} + n_{c,3} = 40.$$

2) Рассчитать теоретические значения по формуле:

$f_{i,j}^T = \textit{сумма по столбцу} \times \textit{сумма по строке} / \textit{число наблюдений} =$

$$= \frac{n_{c,j} \times n_{r,i}}{n}.$$

$$f_{1,1}^T = \frac{15 \times 20}{40} = 7.5; \; f_{2,1}^T = \frac{15 \times 20}{40} = 7.5$$

$$f_{1,2}^T = \frac{20 \times 20}{40} = 10; \; f_{2,2}^T = \frac{20 \times 20}{40} = 10$$

$$f_{1,3}^T = \frac{5 \times 20}{40} = 2.5; \; f_{2,3}^T = \frac{5 \times 20}{40} = 2.5$$

Используемые соотношения основаны на простом факте: если общее количество мнений, например, в поддержку новой процедуры составляет 15 (сумма по первому столбцу) и мнения эти не зависят от квалификации, то распределение 15 ответов между группами врачи/средний медперсонал должно быть пропорционально численности этих групп в выборке, которые в данном случае равны между собой 20/40 = 0.5. Результаты расчетов представлены в табл. 4. (значения в скобках являются теоретическими частотами).

Таблица 4. *Расчет теоретических частот в таблице сопряженности*

Группа	Относится отрицательно	Относятся положительно	Нет определенного мнения	Сумма $n_{r,i}$
Врачи	10 (7.5)	8 (10)	2 (2.5)	20
Средний медперсонал	5 (7.5)	12 (10)	3 (2.5)	20
Сумма $n_{c,j}$	15	20	5	40

Проверка гипотез. Для того чтобы применить аппарат проверки гипотез, сформулируем следующую пару утверждений:

H_0: Признаки являются независимыми;

H_1:Признаки взаимосвязаны

Принятие решения осуществляется с помощью критерия $\chi^2 = \sum_{i,j} \frac{\left(f_{ij} - f_{ij}^T\right)^2}{f_{ij}^T}$

имеющего распределение χ^2 с числом степеней свободы

$$k = (\textit{число строк} - 1) \times (\textit{число столбцов} - 1) = (K_r - 1) \times (K_c - 1).$$

Пример. Выполнить проверку гипотез о независимости в задачи о клинике при уровне значимости 0.05

1) H_0:Мнение о вводимой процедуре не зависит от квалификации;

H_1:Мнение о вводимой процедуре зависит от квалификации, врачи и средний медперсонал считают по-разному.

2) Критическое значение для правой односторонней гипотезы длярапределения с числом степеней свободы $k = (2-1) \times (3-1) = 2$ составляет $\chi_\alpha^2 = 5.991$.

3) Находим расчетное значение критерия:

$$\chi^2 = \sum_{i,j} \frac{\left(f_{ij} - f_{ij}^T\right)^2}{f_{ij}^T} = \frac{(10-7.5)^2}{7.5} + \frac{(8-10)^2}{10} + \frac{(2-2.5)^2}{2.5} + \frac{(5-7.5)^2}{7.5} +$$

$$+ \frac{(12-10)^2}{10} + \frac{(3-2.5)^2}{2.5} = 2.67 .$$

4) Принятие решения. Расчетное значение меньше критического, 2.67 < 5.991, основная гипотеза принимается.

5) Вывод. Результаты проверки согласуются с предположением, что отношение к новой процедуре не зависит от квалификации.

С помощью проверки гипотезы о независимости можно получить ответ и на вопрос несколько другого типа - влияет ли один количественный признак на другой.

Пример. Департамент здравоохранения провел исследование 500 детей с целью выяснить, влияет ли прививка от гриппа на заболеваемость. Результаты исследования представлены в табл. 5. На уровне значимости 0.05 проверить предположение, что прививки влияют на заболеваемость (уменьшают ее).

Таблица. 5. *Таблица сопряженности для исследования заболевания гриппом*

Группы	*Болел*	*Не болел*
Делал прививку	30	270
Не делал прививку	120	80

Решение. Рассчитаем по представленным данным теоретические частоты, основанные на предположении независимости признаков, табл. 6.

1) Вычислить суммы по строкам и столбцам:

$$n_{r,1} = 30 + 270 = 300; \;\; n_{r,2} = 120 + 80 = 200;$$

$$n_{c,1} = 30 + 120 = 150; \;\; n_{c,2} = 270 + 80 = 350;$$

$$n = n_{r,1} + n_{r,2} = n_{c,1} + n_{c,2} = 500 .$$

2) Вычислить теоретические частоты $f_{i,j}^T = \dfrac{n_{c,j} \times n_{r,i}}{n}$.

$$f_{1,1}^T = \frac{150 \times 300}{500} = 90;\ f_{2,1}^T = \frac{150 \times 200}{500} = 60$$

$$f_{1,2}^T = \frac{350 \times 300}{500} = 210;\ f_{2,2}^T = \frac{350 \times 200}{500} = 140$$

Таблица 6. *Расчет теоретических частот*

Группы	*Болел*	*Не болел*	*Сумма* $n_{r,i}$
Делал прививку	30 (90)	270 (60)	300
Не делал прививку	120 (210)	80 (140)	200
Сумма $n_{c,j}$	150	350	500

Переходим непосредственно к проверке гипотез.

1) Формулировка гипотез:

H_0: Заболеваемость гриппом не зависит от прививок;

H_1: Прививки влияют на заболеваемость.

2) Критическое значение для правой односторонней гипотезы и распределения χ^2 ($k = 1$) составляет $\chi^2_\alpha = 3.84$, критическая область $\chi^2 > 3.84$.

3) Находим расчетное значение критерия:

$$\chi^2 = \sum_{i,j} \frac{\left(f_{ij} - f_{ij}^T\right)^2}{f_{ij}^T} = \frac{(30-90)^2}{90} + \frac{(270-210)^2}{210} + \frac{(120-60)^2}{60} + \frac{(80-140)^2}{140} = 142.85$$

4) Принятие решения. Расчетное значение намного превосходит критическое, 142.85 > 3.84, основная гипотеза отклоняется.

5) Вывод: Данные исследования позволяют утверждать, что прививки снижают заболеваемость гриппом.