Лекция 12

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Проверка гипотез. Дисперсионный анализ и критерий хи-квадрат

Здравствуйте!

После изучения материала лекции вы сможете определять и решать ситуации, в которых уместен дисперсионный анализ (ANOVA), объяснять логику проверки гипотез применительно к ANOVA, выполнять тест ANOVA, используя пятиступенчатую модель в качестве руководства, и правильно интерпретировать результаты, определять и объяснять понятия дисперсии, общая сумма квадратов, сумма квадратов между, сумма квадратов внутри и среднеквадратичные оценки, знать разницу между статистической значимостью и важностью взаимосвязей между переменными, а также изучить критерий хи-квадрат.

Дисперсионный анализ (от латинского Dispersio – рассеивание / на английском Analysis Of Variance – ANOVA) применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную.

Вычисление ANOVA.

Даже несмотря на то, что мы рассматривали ANOVA как критерий значимости различий между средними значениями выборки, вычислительная процедура фактически включает в себя разработку двух отдельных оценок дисперсии численности, сигма квадрат. Напомним, что такое дисперсия это стандартное отклонение в квадрате. Одна оценка дисперсии населения основана на величине вариации в каждой из категорий независимой переменной, а другая – на величине.

Чтобы решить эту формулу, сначала найдите сумму квадратов баллов (другими словами, возведите в квадрат каждый балл, а затем сложите квадратов баллов). Затем возведите в квадрат среднее значение всех показателей, умножьте это значение на общее количество случаев в выборке (N) и вычтите это количество из суммы квадратов.
Вышеуказанная может показаться смутно знакомой. Аналогичное выражение

$$\sum \left(X_i - \overline{X}\right)^2$$

появляется в формуле для стандартного отклонения и дисперсии. Все три статистики включают в себя информацию о вариации баллов (или, в случае SST, квадратов баллов) вокруг среднего значения (или, в случае SST, квадрата среднего, умноженного на. Другими словами, все три статистики – это мера вариации или разброса баллов.
Чтобы построить две отдельные оценки дисперсии, мы разделим общую вариацию (SST) на две составляющие. Один компонент отражает характер изменение внутри- в каждой из категорий и называется суммой квадратов внутри (SSW). Другой компонент основан на вариации между категориями и называется суммой квадратов между (SSB). SSW и SSB являются компонентами SST, как показано в

Формула 12.1 SST=SSB+SSW.

Давайте начнем с вычисления SSB, нашей меры вариации между категориями. Мы используем средства категории в качестве сводной статистики, чтобы определить размер разницы от категории к категории. Формула для суммы квадратов между (SSB) имеет вид:

Формула 12.2

$$SSB = \sum N_k \left(\overline{X_k} - \overline{X}\right)^2$$

где где SSB - сумма квадратов между категориями, Nк — количество в категории.
Xk- = среднее значение категории. Чтобы найти SSB, вычтите общее среднее значение всех баллов (X) из среднего значения для каждой категории (Xk), возведите в квадрат разницу, умножьте на количество наблюдений в категории (Nk) и сложите результаты по всем категориям.
Вторая оценка дисперсии населения (SSW) основана на количестве вариаций в категориях. Формула показывает, что общая сумма квадратов (SST) равна сумме SSW и SSB. Это соотношение означает, что мы можем найти SSW простым вычитанием.

Формула 12.3 SSW=SST-SSB.

Давайте на секунду остановимся, чтобы вспомнить, что мы здесь делаем. Если нулевая гипотеза верна, то не должно быть большого различия от категории к категории, и SSW и SSB должны быть примерно равны. Если нулевая гипотеза не соответствует действительности, между категориями будут большие различия относительно различий внутри категорий, и SSB должен быть намного больше, чем SSW. SSB будет увеличиваться по мере увеличения разницы между категориями, особенно если между категориями (SSW) нет особых различий. Чем больше SSB по сравнению с SSW, тем больше вероятность, что мы отвергнем нулевую гипотезу.

Следующим шагом в вычислительной программе является построение оценок дисперсии. Для этого мы делим каждую сумму квадратов на соответствующие степени свободы. Чтобы найти степени свободы, связанные с SSV, вычтите число категорий (k) из числа случаев (N). Степени свободы, связанные с SSB, представляют собой число категорий минус 1. В итоге, dfw=N-k, где dfw – степени свободы, N-общее количество случаев, k-количество категорий.

Фактические оценки дисперсии населения, называемые среднеквадратичными оценками, рассчитываются путем деления каждой суммы квадратов на соответствующие степени свободы:

Формула 12.4 Величина вариации внутри категорий $= \frac{SSW}{dfw}$

Формула 12.5 Величина вариации между категориями $= \frac{SSB}{dfb}$

Тестовая статистика, рассчитанная на шаге 4 пятиступенчатой модели, называется коэффициентом F, и ее значение определяется по следующей формуле:

Формула 12.6 $$F = \frac{\text{Величина вариации между категориями}}{\text{Величина вариации внутри категорий}}$$

Как вы можете видеть, значение коэффициента F является функцией величины вариации между категориями (на основе SSB) по сравнению с величиной вариации внутри категорий (на основе SSW). Чем больше вариация между категориями относительно вариации внутри, тем выше значение коэффициента F и тем более вероятно, что мы отвергнем нулевую гипотезу.

Рассмотрим пример:
Была выбрана случайная выборка из 20 стран с четырьмя уровнями дохода. Мы использовали эти экономические категории в главе 4 при обсуждении коэффициентов рождаемости в связи с остроумными коробками. Напомним, что страны с низким уровнем дохода в основном занимаются

сельским хозяйством и имеют самое низкое качество жизни. Ожидаемая продолжительность жизни по уровню дохода страны с высоким доходом - индустриальные, самые амбициозные и современные. Страны с доходом выше и ниже среднего редко встречаются между этими крайностями. Отражены ли эти уровни доходов в разнице в ожидаемой продолжительности жизни (сколько лет средний гражданин может прожить при рождении), данные приведены в таблице 12.1.
Чтобы найти F (получено) и провести тест ANOVA, вычисления будут организованы в виде таблицы:

Таблица 12.1 Уровни доходов в разнице в ожидаемой продолжительности жизни

18-29		30-45		45-64		65+	
X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
7	49	10	100	12	144	17	289
8	64	12	144	15	225	20	400
10	100	13	169	17	289	24	576
15	255	17	289	20	400	27	729
40	438	52	702	64	1058	88	1994
$\overline{X}_k = 10.0$		$\overline{X}_k = 13.0$		$\overline{X}_k = 16.0$		$\overline{X}_k = 22.0$	
$\overline{X} = 15.25$							

Тест ANOVA покажет нам, достаточно ли велики эти различия, чтобы оправдать вывод о том, что они произошли случайно. Следуя обычной вычислительной процедуре, результаты вы сейчас видите. Теперь мы можем провести проверку значимости. Мы нашли полученное соотношение F= 34,44. Данной значение сравним с критическим равным 3,24. Нулевая гипотеза («Значения населения равны») может быть отвергнута. Различия в ожидаемой продолжительности жизни между странами с разными уровнями доходов статистически значимы и отражают различия в популяциях, из которых были отобраны эти данные.

ANOVA подходит для тестирования различий между средними значениями отношения интервалов – уровня, зависящего от переменной, между тремя или более категориями независимой переменной. Это приложение называется односторонним анализом отклонений, поскольку оно включает влияние одной переменной (общая численность населения) на другую (например, экономически активное население). Это простейшее применение ANOVA, и вы должны знать, что метод имеет множество более продвинутых и сложных форм. Например, вы можете столкнуться с исследовательскими проектами, в которых наблюдали влияние двух отдельных переменных (например, численности населения и количества безработных) на какую-то третью переменную. Одним из важных ограничений ANOVA является то, что для него требуется переменная, зависящая от отношения интервалов, и примерно равное число случаев в каждой категории независимой переменной. Первое условие может быть трудно встретить с полной уверенностью для многих переменных, представляющих интерес в социальных науках. Последнее условие может создавать проблемы, когда гипотеза исследования требует сравнения между группами, которые по своей природе неравны по количеству (например, белые против чернокожих американцев). Ни одно из этих ограничений не должно быть особенно вредным, потому что ANOVA допускает некоторое отклонение от своих модельных допущений, но вы должны знать об этих ограничениях при планировании собственных исследований, а также

при оценке адекватности исследований, проводимых другими. Другое ограничение ANOVA, фактически относится ко всем формам проверки значимости. Тесты значимости предназначены для выявления неслучайных различий или различий, настолько больших, что вряд ли они могут быть произведены одним случайным случаем. Проблема в том, что различия, которые являются статистически значимыми, не обязательно важны ни в каком другом смысле. Последнее ограничение ANOVA относится к гипотезе исследования, которая просто утверждает, что, по крайней мере, один из популяционных средств отличается от других. Очевидно, что когда мы отвергаем нулевую гипотезу, нам хотелось бы знать, какие различия между выборочными средними значимы. Иногда мы можем сделать это определение простым интуитивным анализом.

Сделаем некоторые выводы:
1.Односторонний дисперсионный анализ - это мощный критерий значимости, который обычно используется, когда представляют интерес сравнения более чем двух категорий или выборок. Возможно, проще всего представить ANOVA как продолжение теста на разницу в средних значениях выборки.

2. ANOVA сравнивает количество вариаций внутри категорий с количеством вариаций между категориями. Если нулевая гипотеза об отсутствии различий неверна, между категориями должны быть относительно большие различия и относительно небольшие различия между категориями. Чем больше различий от категории к категории относительно различий внутри категорий, тем больше вероятность того, что мы сможем отвергнуть нулевую гипотезу.

3. Вычислительная процедура даже для простых приложений ANOVA может быстро стать довольно сложной (факт, который указывает на ценность компьютеризированных статистических пакетов, таких как SPSS). Основной процесс заключается в построении отдельных оценок дисперсии населения на основе различий в категориях и различий между категориями. Тестовая статистика – это коэффициент F, который основан на сравнении этих двух оценок.

4. Тест ANOVA может быть организован в знакомую пятиступенчатую модель для проверки значимости результатов выборки. Хотя допущения модели (шаг 1) требуют высококачественных данных, тест может допускать некоторое отклонение, если размеры выборки примерно равны. Нулевая гипотеза принимает знакомую форму заявления о том, что нет никаких различий какой-либо важности среди значений совокупности, в то время как альтернативная гипотеза утверждает, что по крайней мере одно среднее значение совокупности отличается. Распределением выборки является F-распределение, а тест всегда односторонний. Решение отклонить или не отклонить нулевую гипотезу основано на сравнении полученного отношения F с критическим отношением F, определенным для данного альфа-уровня и степеней свободы. Решение отклонить нулевую гипотезу указывает только на то, что одно или несколько средств населения отличается от других. Мы часто можем определить, какое из выборочных средств учитывает разницу, проверяя выборочные данные, но этот неформальный метод следует использовать с осторожностью.

Хи-квадрат.

Критерий хи-квадрат является одним из наиболее часто используемых тестов гипотез в социальных науках, популярность которого во многом обусловлена тем, что допущения и требования в шаге 1 пятиступенчатой модели просты удовлетворить. Тест может проводиться с переменными, измеренными на номинальном уровне (самый низкий уровень измерения), и является непараметрическим, что означает, что в нем вообще не требуется никаких предположений о форме популяции или распределении выборки.

Почему выгодно иметь простые для удовлетворения предположения и требования? Решение отклонить нулевую гипотезу (шаг 5) не является конкретным: это означает, что только одно утверждение в модели (шаг 1) или нулевая гипотеза (шаг 2) неверно. Обычно, конечно, мы выделяем нулевую гипотезу отказа. Чем более мы уверены в модели, изложенной в шаге 1, тем выше наша уверенность в том, что нулевая гипотеза является ошибочным предположением. «Слабая» или легко удовлетворяемая модель означает, что наше решение отклонить нулевую гипотезу может быть принято с большей уверенностью.

Квадрат Хи также был популярен благодаря своей гибкости. Его можно использовать с переменными, имеющими много категорий или оценок. Тест с двумя выборками больше не будет применим, но хи-квадрат легко обрабатывает более сложные переменные. Кроме того, в отличие от теста ANOVA, тест хи-квадрат можно проводить с переменными на любом уровне измерения.
Двусторонние таблицы.
Хи-квадрат вычисляется из двумерных таблиц, так называемых, потому что они отображают оценки случаев по двум различным переменным одновременно. Двусторонние таблицы используются для проверки значимых связей и для других целей, которые мы рассмотрим в следующих главах. На самом деле, эти таблицы очень часто используются в исследованиях.
Прежде всего, двумерные таблицы имеют (конечно) два измерения. Мы называем горизонтальное (поперечное) измерение в виде строк, а вертикальное измерение (вверх и вниз) - в виде столбцов. Каждый столбец или строка представляет оценку по переменной, а пересечения строк и столбцов (ячеек) представляют объединенные оценки по обеим переменным.
Давайте использовать пример, чтобы уточнить. Предположим, что исследователь интересуется жизнью пожилых людей и, в частности, интересуется, влияет ли их участие в добровольных группах их семейное положение. Чтобы упростить анализ, исследователь ограничил выборку людьми, которые в настоящее время состоят в браке или не состоят в браке (в том числе одинокие и разведенные), и оценил участие в добровольных ассоциациях как простую дихотомию: люди были классифицированы как высокие или с низким уровнем участия.

По соглашению мы обозначаем независимую переменную (переменную, которая считается причиной) в столбцах и зависимую переменную в строках. В данном примере семейное положение является причинно-следственной переменной (вопрос был «На участие ли влияет семейное положение?»). И каждый столбец будет представлять оценку по этой переменной. Каждый ряд, с другой стороны, будет представлять оценку уровня вовлеченности (высокая или низкая).

Таблица 12.2 Схема двумерной таблицы для выборки из 100 пожилых людей

	Семейное положение		Всего
	состоят в браке	не состоят в браке	
Высокий уровень			50
Низкий уровень			50
Итого	50	50	100

Обратите внимание на некоторые дополнительные детали таблицы. Во-первых, промежуточные итоги включены для каждого столбца n и строки. Они называются маргинальными строчками или столбцами, и в этом случае они говорят нам, что 50 членов выборки состоят в браке, а 50 не состоят в браке (маргинальные столбцы), и что 50 имеют высокий уровень участия и 50 – низкие (маргинальные строки).

Во-вторых, общее количество случаев в выборке (N – 100) сообщается на пересечении маргинальных строк и столбцов. Наконец, внимательно обратите внимание на маркировку на столе. Каждая строка и столбец идентифицированы, и таблица имеет описательный заголовок, который включает в себя имена переменных, с зависимой переменной, перечисленной вожделение. Четкие, полные надписи и краткие заголовки должны быть включены во все таблицы, графики и диаграммы.

Как вы заметили, в таблице 12.2 отсутствует некоторая важная информация: номера в таблице. Чтобы завершить таблицу, нам нужно классифицировать семейное положение и уровень участия каждого члена в выборке, вести учет того, как часто встречается каждая комбинация баллов, и записывать эти числа в соответствующие ячейки таблицы. Поскольку каждая переменная имеет две возможные оценки, возможны четыре комбинации оценок, каждая из которых соответствует ячейке в таблице. Например, женатые люди с высоким уровнем участия будут учитываться в верхней левой ячейке, неженатые люди с низким уровнем участия будут учитываться в нижней правой ячейке и т.д. Когда мы закончим подсчет, в каждой ячейке будет отображаться количество раз, когда произошла каждая комбинация баллов.
Наконец, обратите внимание, что мы могли бы расширить таблицу, чтобы она включала переменные с более чем двумя показателями. Если бы мы хотели включить людей с другими семейным статусом (овдовевшие, разделенные и т. д.), мы просто добавим столбцы. Более сложные зависимые переменные также могут быть легко приспособлены. Если бы мы измерили уровень участия по трем категориям (например, высокий, средний и низкий), мы просто добавили бы строку в таблицу.

Тест хи-квадрат имеет несколько различных применений, но мы рассмотрим только тест хи-квадрат на независимость. Мы столкнулись с термином независимость в связи с требованиями для случая проверки гипотезы с двумя выборками и для теста ANOVA. В контексте хи-квадрат независимость относится к взаимосвязи между переменными, а не между выборками. Две переменные являются независимыми, если классификация случая в определенную категорию одной переменной не влияет на вероятность того, что случай попадет в какую-либо конкретную категорию второй переменной. Например, переменные в таблице 12.2 будут независимы друг от друга, если классификация случая как состоящего в браке или не состоящего в браке не влияет на классификацию случая как высокую или низкую по участию. Другими словами, переменные являются независимыми, если уровень участия и семейное положение полностью не связаны друг с другом.
Рассмотрим таблицу 12.2 еще раз. Если бы переменные были независимыми, то частоты ячеек определялись бы исключительно случайным образом, и мы обнаружили бы, что около половины респондентов, состоящих в браке, имели бы высокий рейтинг по участию, а половина – низко. Такая же схема будет иметь место для 50 не состоящих в браке респондентов, и, следовательно, в каждой из четырех ячеек будет около 25 случаев, как показано в таблице 12.3.

Таблица 12.3 Схема 1 двумерной таблицы для выборки из 100 пожилых людей

	Семейное положение		Всего
	состоят в браке	не состоят в браке	
Высокий уровень	25	25	50
Низкий уровень	25	25	50
Итого	50	50	100

Такая структура частот связи указывает на то, что семейное положение не влияет на уровень участия человека. Вероятность быть классифицированной как высокая или низкая будет 0,50 для обоих семейных статусов, и поэтому переменные будут независимыми.

Нулевая гипотеза для критерия независимости хи-квадрат заключается в том, что переменные независимы. В предположении, что нулевая гипотеза верна, мы вычисляем частоты ячеек, которые мы ожидаем найти, если бы действовал только случайный случай. Эти частоты называются ожидаемыми частотами (обозначены fe), и мы сравниваем их, ячейка за ячейкой, с частотами, фактически наблюдаемыми в таблице (наблюдаемые частоты, обозначены f0). Если нулевая гипотеза верна и переменные независимы, то должно быть мало Разница между ожидаемой и наблюдаемой частотами. Однако если нулевая гипотеза неверна, между ними должны быть большие различия. Чем больше различие между ожидаемой (fe) и наблюдаемой (f0) частотами, тем меньше вероятность того, что переменные являются независимыми, и тем более вероятно, что мы сможем отклонить нулевую гипотезу.

Вычисление хи-квадрат

Для проведения теста хи-квадрат – как и во всех тестах гипотезы – мы вычисляем статистику теста хи, из данных выборки и затем помещаем это значение в распределение выборки всех возможных результатов выборки. В частности фактическое значение хи-квадрат будет сравниваться со значением хи-квадрат критическим, которое будет определено из таблицы распределения хи-квадрат для конкретного альфа-уровня и степеней свободы.
Процедура расчета хи-квадрат приведена в формуле 12.7:

Формула 12.7 $$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Для каждой ячейки вычтите ожидаемую частоту из наблюдаемой частоты, возведите в квадрат результат, а затем разделите на ожидаемую частоту для этой ячейки. Суммируйте результирующие значения для всех ячеек.

Эта формула требует ожидаемой частоты для каждой ячейки в таблице. В таблице 12.3 предельные значения одинаковы для всех строк и столбцов, а ожидаемые частоты очевидны fe = 25 для всех четырех ячеек. В более обычном случае ожидаемые частоты не будут очевидны, предельные значения будут неравными, и мы должны использовать формулу 12.8, чтобы найти ожидаемую частоту для каждой ячейки:

Формула 12.8

$$f_e = \frac{\text{Число наблюдений в строке} \times \text{Число наблюдений в столбце}}{N}$$

То есть ожидаемая частота для любой ячейки равна общему числу наблюдений в строке (предельному значению строки), умноженному на общее число наблюдений в столбце (предельному значению столбца), деленному на общее число наблюдений в таблице (N).

Приведем пример. Случайная выборка из 100 специальностей социальной работы была классифицирована следующим образом:
Образование в области социальной работы аккредитовало их программы бакалавриата (столбец или независимая переменная) и были ли они наняты на должности по социальной работе в течение трех месяцев после окончания обучения (строка или зависимая переменная). Таблица 12.4, 12.5, 12.6 вспомогательные расчетные таблицы, они сейчас перед вами.

Таблица 12.4 Схема двумерной таблицы для выборки из 100 выпускников программы бакалавриата в области социальной работы и их аккредитация.

	Статус аккредитации		Всего
	аккредитованы	не аккредитованы	
Работают на должности по социальной работе	30	10	40
Не работают на должности по социальной работе	25	35	60
Итого	55	45	100

Таблица 12.5 Данные таблицы выборки из 100 выпускников программы бакалавриата в области социальной работы и их аккредитация.

	Статус аккредитации		Всего
	аккредитованы	не аккредитованы	
Работают на должности по социальной работе	22	18	40
Не работают на должности по социальной работе	33	27	60
Итого	55	45	100

Таблица 12.6 Результаты расчетов таблицы выборки из 100 выпускников программы бакалавриата в области социальной работы и их аккредитация.

1	2	3	4	5
f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$\frac{(f_0 - f_e)^2}{f_e}$
30	22	8	64	2,91
10	18	-8	64	3,56
25	33	-8	64	1,94
35	27	8	64	2,37
100	100	0		10,78

Начиная с верхней левой ячейки (выпускники аккредитованных программ, которые работают в качестве социальных работников), ожидаемая частота для этой ячейки с использованием формулы 12.8 составляет (40*55) / 100 или 22. Для другой ячейки в этом ряду (выпускники неаккредитованных программ, работающие в качестве социальных работников) ожидаемая частота составляет (40*45)/100 или 18. Для двух ячеек в нижнем ряду ожидаемые частоты равны (60*55)/100 или 33 и (60*45)/100 или 27 соответственно. Ожидаемые частоты для всех четырех ячеек приведены в таблице 12.5.

Обратите внимание, что маргинальные строки и столбцы, а также общее количество случаев в таблице 12.5 точно такие же, как в таблице 12.4. Границы строк и столбцов для ожидаемых частот должны всегда равняться таковым для наблюдаемых частот, что обеспечивает удобный способ проверки вашей арифметики до этой точки.

Значение для хи-квадрат для этих данных теперь можно найти, решив формулу 12.7. Будет полезно использовать вычислительную таблицу, такую как Таблица 12.6, для организации нескольких шагов, необходимых для вычисления хи-квадрат. В таблице перечислены наблюдаемые частоты

(f0) в столбце 1 в порядке от верхней левой ячейки до нижней правой ячейки, перемещаясь слева направо по всей таблице и сверху вниз. В столбце 2 перечислены ожидаемые частоты (fe) в том же порядке. Дважды проверьте, чтобы убедиться, что вы перечислили частоты ячеек в одном и том же порядке для обоих этих столбцов.
Следующим шагом является вычитание ожидаемой частоты из наблюдаемой частоты для каждой ячейки и перечисление этих значений в столбце 3. Чтобы заполнить столбец 4, возведите в квадрат значение в столбце 3, а затем в столбце 5 разделите значение столбца 4 на ожидаемая частота для этой ячейки. Наконец, сложите столбец 5. Сумма этого столбца получается хи-квадрат. Для таблицы 12.4 хи квадрат фактическое = 10,78.

Обратите внимание, что итоговые значения для столбцов 1 (f0) и 2 (fe) абсолютно одинаковы. Это всегда будет иметь место, и если итоговые значения не совпадают, вы допустили вычислительную ошибку, возможно, при расчете ожидаемых частот. Также обратите внимание, что сумма столбца 3 всегда будет 0, еще один удобный способ проверить свою математику на данный момент. Это значение образца для квадрата хи все еще должно быть проверено на предмет его значимости.

Теперь мы готовы провести тест на хи-квадрат на независимость. Напомним, что если переменные не зависят друг от друга, оценка случая по одной переменной не будет связана с его оценкой по другой переменной. Как всегда, пятиступенчатая модель проверки значимости обеспечит основу для организации процесса принятия решений. Данные, представленные в таблице 12.4 послужат нашим примером.

Шаг 1. Создание предположений и выполнение требований теста. Обратите внимание, что, поскольку тест непараметрический, мы не делаем никаких предположений о форме распределения выборки.
Модель: Независимые случайные выборки.
Уровень измерения номинальный.

Шаг 2. Изложение нулевой гипотезы. Нулевая гипотеза утверждает, что две переменные независимы. Если нулевая гипотеза верна, различия между наблюдаемой и ожидаемой частотами будут небольшими. Как обычно, гипотеза исследования прямо противоречит нулевой гипотезе. Гипотезы имеют вид.

Шаг 3. Выбор распределения выборки и определение критического региона. Распределение выборки хи-квадратов выборки, в отличие от распределений Z и t, имеет положительный сдвиг, причем более высокие значения выборочных хи-квадратов в верхнем хвосте распределения (справа). Таким образом, с помощью критерия хи-квадрат критическая область устанавливается в верхнем хвосте распределения выборки.
Значения для хи-квадрат (критические) приведены в таблице распределений. Эта таблица аналогична таблице t, где альфа-уровни расположены сверху, а степени свободы – сбоку. Тем не менее, с квадратом хи, степени свобода определяем по формуле 12.9:
Формула 12.9

$$df = (r - 1)(c - 1)$$

и хи-квадрат критическое будет равно 3,841.

Шаг 4. Вычисление тестовой статистики. Вычисление хи-квадрат было введено в предыдущем разделе. Как вы помните, у нас было расчетное значение равно 10,78.

Шаг 5. Принятие решения и интерпретация результатов теста. Сравнивая статистику теста с критической областью, мы видим, что тестовая статистика попадает в критическую область, и поэтому мы отвергаем нулевую гипотезу независимости.

Наблюдаемая в таблице 12.4 структура частот ячеек вряд ли произошла случайно. Переменные являются зависимыми. В частности, исходя из этих выборочных данных, вероятность обеспечения занятости в сфере социальной работы зависит от статуса аккредитации программы. Мы должны четко понимать, что именно тест хи-квадрат делает и не говорит нам.
Значительный хи-квадрат означает, что переменные (вероятно) зависят друг от друга в популяции. В нашем примере это означает, что существует связь между аккредитацией и тем, работает ли человек в качестве социального работника. Но какова именно связь между переменными? Какой тип выпускников с большей вероятностью найдет работу по профессии? Чтобы сделать это определение, мы должны произвести дополнительный расчет.

Мы можем выяснить, как независимая переменная (статус аккредитации в нашем примере) влияет на зависимую переменную (занятость в качестве социального работника), вычисляя проценты столбца или вычисляя проценты в каждом столбце двумерной таблицы. Эта процедура аналогична расчету процентов для частотных распределений, как это делать мы уже знаем.
Чтобы вычислить процентное содержание столбца, разделите частоту каждой ячейки на общее количество наблюдений в столбце (маргинальный столбец) и умножьте результат на 100. Для таблицы 12.4, начиная с верхней буквы, мы видим, что имеется 50 случаев в этой камере и 55 случаев в кол. Таким образом, 30 из 55 выпускников аккредитованных программ работают в качестве социальных работников. Поэтому процент столбца для этой ячейки составляет (30/55)*100 = 54,55 %. Для нижней левой ячейки процент столбца (25/55)*100 = 45,45%. Для двух ячеек в правой колонке n (выпускники неаккредитованных программ) проценты монеты составляют (10/45)*100 = 22,22% и (35/45)*100=77,78%. В таблице 12.7 показаны процентные доли всех столбцов для исходной таблицы.

Таблица 12.7 Данные таблицы выборки из 100 выпускников программы бакалавриата в области социальной работы и их аккредитация (%).

	Статус аккредитации		Всего
	аккредитованы	не аккредитованы	
Работают на должности по социальной работе	54.55%	22.22%	40%
Не работают на должности по социальной работе	45.45%	77.78%	60%
Итого	100% (55)	100% (45)	100%

Проценты столбца делают отношения между переменными более очевидными и мы легко видим из таблицы, что именно студенты из аккредитованных программ с большей вероятностью будут работать в качестве социальных работников. Почти 55% этих студентов работают в качестве

социальных работников против примерно 22% студентов из неаккредитованных программ. Мы уже знаем, что это соотношение является значительным (маловероятно, что оно вызвано случайной случайностью), и теперь, с помощью процентов в столбцах, мы знаем, как связаны две переменные. Согласно этим результатам, выпускники аккредитованных программ имеют решающее преимущество в обеспечении социальной работы.

Подведем итоги. Критерий хи-квадрат для проверки гипотезы на независимость подходит для ситуаций, в которых представляющие интерес переменные были организованы в виде таблицы. Нулевая гипотеза состоит в том, что переменные являются независимыми или что классификация случая в определенную категорию по одной переменной не влияет на вероятность того, что случай будет классифицирован в какую-либо конкретную категорию второй переменной. В тесте хи-квадрат мы сначала находим частоты, которые появляются в ячейках, если переменные были независимы (fe), а затем сравниваем эти частоты, ячейка за ячейкой, с частотами, фактически наблюдаемыми в номинально измеренных переменных, его модельные предположения.