**Применение χ2-критерия для проверки гипотезы о равенстве двух или нескольких долей**

В предыдущих заметках были описаны процедуры проверки гипотез о числовых и категорийных данных: [одновыборочные критери](http://baguzin.ru/wp/?p=5790), несколько [двухвыборочных](http://baguzin.ru/wp/?p=5832), а также [основы дисперсионного анализа](http://baguzin.ru/wp/?p=5884), позволяющего изучать один или [два интересующих нас фактора](http://baguzin.ru/wp/?p=5919). В настоящей заметке мы рассмотрим методы проверки гипотез о различиях между долями признака в генеральных совокупностях на основе нескольких независимых выборок, а также изучим критерий «хи-квадрат» (χ2-критерий) для проверки независимости двух категорийных переменных.[[1]](#footnote-1)

Для иллюстрации применяемых методов используется сценарий, в котором оценивается степень удовлетворенности постояльцев отелей, принадлежащих компании Т. С. Resort Properties. Представьте себе, что вы — менеджер компании, владеющей пятью отелями, расположенными на двух курортных островах. Если гости удовлетворены обслуживанием, велика вероятность, что они вернутся на следующий год и порекомендуют своим друзьям остановиться именно в вашем отеле. Чтобы оценить качество обслуживания, постояльцев просят заполнить анкету и указать, довольны ли они гостеприимством. Вам необходимо проанализировать данные опроса, определить общую степень удовлетворенности запросов постояльцев, оценить вероятность того, что гости приедут вновь в следующем году, а также установить причины возможного недовольства некоторых клиентов. Например, на одном из островов компании принадлежат отели Beachcomber и Windsurfer. Одинаково ли обслуживание в этих отелях? Если нет, как эту информацию можно использовать для улучшения качества работы компании? Более того, если некоторые постояльцы заявили, что больше к вам не приедут, какие причины они указывают чаще других? Можно ли утверждать, что эти причины касаются лишь конкретной гостиницы и не относятся ко всей компании в целом?

[Ранее](http://baguzin.ru/wp/?p=5832) был описан *Z*-критерий для сравнения двух долей. Сейчас мы рассмотрим способ, основанный на сравнении количества успехов в двух группах, а не их долей в генеральных совокупностях. В этой процедуре тестовая χ2-статистика аппроксимируется χ2-распределением с одной степенью свободы. Результат, полученный с помощью χ2-критерия эквивалентен результату применения *Z*-критерия.

Для сравнения количества успехов в двух независимых группах необходимо заполнить таблицу перекрестной классификации с двумя входами, содержащую количество успехов и неудач в каждой из групп (рис. 1). Такую таблицу часто называют таблицей сопряженности признаков или факторной (ранее мы уже приводили примеры таких таблиц; см. [Представление категорийных данных в виде таблиц и диаграмм](http://baguzin.ru/wp/?p=5342), а в заметке [Основные понятия теории вероятностей](http://baguzin.ru/wp/?p=5467) мы использовали ее для определения понятия вероятности).

|  |  |  |  |
| --- | --- | --- | --- |
| **Столбцы (группы)** | | | |
| **Строки** | **1** | **2** | **Итого** |
| Успехи | *X1* | *X2* | *X* |
| Неудачи | *n1 – X1* | *n2 – X2* | *n – X* |
| Всего | *n1* | *n2* | *n* |

Рис. 1. Шаблон факторной таблицы 2х2

Здесь использованы следующие обозначения: *X1* — количество успехов в первой группе, *X2* — количество успехов во второй группе, *n1 – X1* — количество неудач в первой группе, *n2 – X2* — количество неудач во второй группе, *X = X1 + X2* — общее количество успехов, *n – X* = (*n1 – X1*) + (*n2 – X2*) — общее количество неудач, *n1* — объем первой выборки, *n2* — объем второй выборки, *n* = *n1* + *n2* — суммарный объем выборок. Представленная таблица имеет две строки и два столбца, поэтому она называется факторной таблицей 2x2. Ячейки, образованные пересечением каждой строки и столбца, содержат количество успехов или неудач.

Проиллюстрируем применение таблицы сопряженности признаков на примере сценария, описанного выше. Предположим, что на вопрос «Вернетесь ли вы в следующем году?» утвердительно ответили 163 из 227 постояльцев отеля Beachcomber, и 154 из 262 постояльцев отеля Windsurfer. Существует ли статистически значимая разность между степенью удовлетворенности постояльцев отелей (представляющая собой вероятность того, что постояльцы вернутся в следующем году), если уровень значимости равен 0,05?

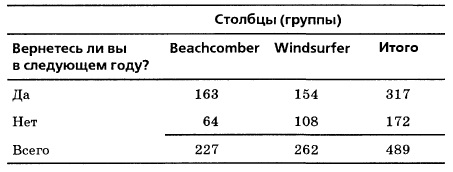
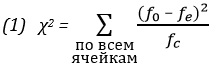


Рис. 2. Факторная таблица 2х2 для оценки качества обслуживания постояльцев

В первой строке указывается количество постояльцев каждого отеля, заявивших о своем желании вернуться в следующем году (успех); во второй строке – количество постояльцев, выразивших недовольство (неудача). Ячейки, расположенные в столбце «Итого», содержат общее количество гостей, планирующих вернуться в отель в следующем году, а также общее количество гостей, недовольных обслуживанием. Ячейки, расположенные в строке «Всего», содержат общее количество опрошенных постояльцев каждого отеля. Доля постояльцев, планирующих вернуться, вычисляется путем деления количества постояльцев, заявивших об этом, на общее количество опрошенных гостей данного отеля. Затем для сравнения вычисленных долей применяется χ2-критерий.

Чтобы проверить нулевую и альтернативные гипотезы *Н0: р1 = р2*; *Н1: р1 ≠ р2* используем тестовую χ2-статистику.

*Критерий «хи-квадрат» для сравнения двух долей.* Тестовая χ2-статистика равна сумме квадратов разностей между наблюдаемым и ожидаемым количеством успехов, деленных на ожидаемое количество успехов в каждой ячейке таблицы:



где *f0* — наблюдаемое количество успехов или неудач в конкретной ячейке таблицы сопряженности признаков, *fe* — теоретическое, или ожидаемое, количество успехов или неудач в конкретной ячейке таблицы сопряженности признаков при условии, что нулевая гипотеза является истинной.

Тестовая χ2-статистика аппроксимируется χ2-распределением с одной степенью свободы.

Чтобы вычислить ожидаемое количество успехов или неудач в каждой ячейке таблицы сопряженности признаков, необходимо понимать их смысл. Если нулевая гипотеза является истинной, т.е. доли успехов в двух генеральных совокупностях равны, выборочные доли, вычисленные для каждой из двух групп, могут отличаться друг от друга лишь по случайным причинам, причем обе доли являются оценкой общего параметра генеральной совокупности *р*. В этой ситуации статистика, объединяющая обе доли в одной общей (средней) оценке параметра *р*, содержит больше информации, чем каждая из них в отдельности. Эта статистика, обозначаемая символом *р̅*, представляет собой общую долю успехов в объединенных группах (т.е. равна общему количеству успехов, деленному на суммарный объем выборок). Ее дополнение, *1 – р̅*, представляет собой общую долю неудач в объединенных группах. Используя обозначения, смысл которых описан в таблице на рис. 1. можно вывести формулу (2) для вычисления параметра *р̅*:

*(2) p̅ =*

где *р̅* – средняя доля признака.

Чтобы вычислить ожидаемое количество успехов *fe* (т.е. содержимое первой строки таблицы сопряженности признаков), необходимо умножить объем выборки на параметр *р̅*. Чтобы вычислить ожидаемое количество неудач *fe* (т.е. содержимое второй строки таблицы сопряженности признаков), необходимо умножить объем выборки на параметр *1 – р̅*.

Тестовая статистика, вычисленная по формуле (1), аппроксимируется χ2-распределением с одной степенью свободы. При заданном уровне значимости α нулевая гипотеза отклоняется, если вычисленная χ2-статистика больше χU2, верхнего критического значения χ2-распределения с одной степенью свободы. Таким образом, решающее правило выглядит следующим образом: гипотеза *H0* отклоняется, если χ2 > χU2, в противном случае гипотеза *Н0* не отклоняется (рис. 3).



Рис. 3. Критическая область χ2-критерия для сравнения долей при уровне значимости α

Если нулевая гипотеза является истинной, вычисленная χ2-статистика близка к нулю, поскольку квадрат разности между наблюдаемой *f0* и ожидаемой *fе* величинами в каждой ячейке очень мал. С другой стороны, если нулевая гипотеза *Н0* является ложной и между долями успехов в генеральных совокупностях существует значимая разница, вычисленная χ2-статистика должна быть большой. Это объясняется разностью между наблюдаемым и ожидаемым количеством успехов или неудач в каждой ячейке, которая увеличивается при возведении в квадрат. Однако вклады разностей между ожидаемыми и наблюдаемыми величинами в общую χ2-статистику могут быть неодинаковыми. Одна и та же фактическая разность между *f0* и *fe* может оказать большее влияние на χ2-статистику, если в ячейке содержатся результаты небольшого количества наблюдений, чем разность, соответствующая большему количеству наблюдений.

Для того чтобы проиллюстрировать χ2-критерий для проверки гипотезы о равенстве двух долей, вернемся к сценарию, описанному в ранее, результаты которого приведены на рис. 2. Нулевая гипотеза (Н0: р1 = р2) утверждает, что при сравнении качества обслуживания в двух отелях доли постояльцев, планирующих вернуться в следующем году, практически одинаковы. Для оценки параметра *р*, представляющего собой долю гостей, планирующих вернуться в отель, если нулевая гипотеза является истинной, используется величина *р̅*, которая вычисляется по формуле

*p̅ = = 0,6483*

Доля гостей, оставшихся недовольными обслуживанием = 1 – 0,6483 = 0,3517. Умножая эти две доли на количество опрошенных постояльцев отеля Beachcomber, получаем ожидаемое количество гостей, планирующих вернуться в следующем сезоне, а также число отдыхающих, которые больше не остановятся в этом отеле. Аналогично вычисляются ожидаемые доли постояльцев отеля Windsurfer:

Да — Beachcomber: *р̅* = 0,6483, *n1* = 227, следовательно, *fe* = 147,16.   
Да — Windsurfer: *р̅* = 0,6483, *n2* = 262, следовательно, *fe* = 169,84.  
Нет — Beachcomber: 1 – *р̅* = 0,3517, *n1* = 227, следовательно, *fe* = 79,84.   
Нет — Windsurfer: 1 – *р̅* = 0,3517, *n2* = 262, следовательно, *fe* = 92,16.

Расчеты представлены на рис. 4.

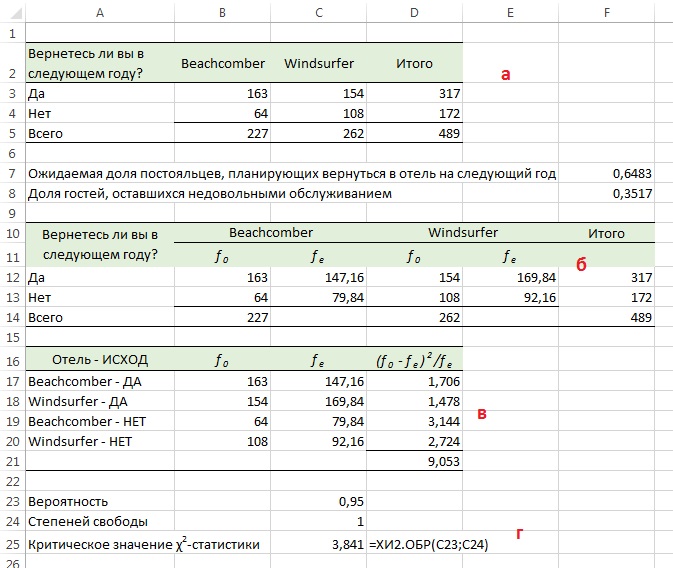


Рис. 4. χ2-статистика для отелей: (а) исходные данные; (б) факторная таблица 2х2 для сравнения наблюдаемого (*f0*) и ожидаемого (*fe*) количества постояльцев, удовлетворенных и не удовлетворенных обслуживанием; (в) вычисление χ2-статистики при сравнении доли постояльцев, удовлетворенных обслуживанием; (г) расчет критического значения тестовой χ2-статистики

Для расчета критического значения тестовой χ2-статистики применяется функция Excel =ХИ2.ОБР(). Если уровень значимости α = 0,05 (вероятность, подставляемая в функцию ХИ2.ОБР есть 1 –α), а χ2-распределение для факторной таблицы 2x2 имеет одну степень свободы, критическое значение χ2-статистики равно 3,841. Поскольку вычисленное значение χ2-статистики, равное 9,053 (рис. 4в), превышает число 3,841, нулевая гипотеза отклоняется (рис. 5).



Рис. 5. Определение критического значения тестовой χ2-статистики с одной степенью свободы при уровне значимости α = 0,05

Вероятность *р* того, что нулевая гипотеза верна при χ2-статистикие равной 9,053 (и одной степени свободы) рассчитывается в Excel с помощью функции =1 – ХИ2.РАСП(9,053;1;ИСТИНА) = 0,0026. *р*-значение, равное 0,0026, — это вероятность того, что разность между выборочными долями постояльцев, удовлетворенных обслуживанием в отелях Beachcomber и Windsurfer, равна или больше 0,718 – 0,588 = 0,13, если на самом деле их доли в обеих генеральных совокупностях одинаковы. Таким образом, существуют веские основания утверждать, что между двумя отелями есть статистически значимая разница в обслуживании постояльцев. Исследования показывают, что количество гостей, удовлетворенных обслуживанием в отеле Beachcomber, больше количества постояльцев, планирующих снова остановиться в гостинице Windsurfer.

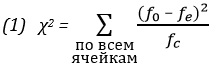
*Проверка предположений, касающихся факторной таблицы 2x2.* Для получения точных результатов на основе данных, приведенных в таблице 2x2, необходимо, чтобы количество успехов или неудач превышало число 5. Если это условие не выполняется, следует применять точный *критерий Фишера*.

При сравнении процента клиентов, удовлетворенных качеством обслуживания в двух отелях, критерии Z и χ2 приводят к одинаковым результатам. Это можно объяснить существованием тесной связи между стандартизованным нормальным распределением и χ2-распределением с одной степенью свободы. В этом случае χ2-статистика всегда является квадратом Z-статистики. Например, при оценке степени удовлетворенности гостей мы обнаружили, что *Z*-статистика равна +3,01, а χ2-статистика — 9,05. Пренебрегая ошибками округления, легко убедиться, что вторая величина является квадратом первой (т.е. 3,012 = 9,05). Кроме того, сравнивая критические значения обеих статистик при уровне значимости α = 0,05, можно обнаружить, что величина χ12 равная 3,841, является квадратом верхнего критического значения Z-статистики, равного +1,96 (т.е. χ12 = Z2). Более того, *р*-значения обоих критериев одинаковы.

Таким образом, можно утверждать, что при проверке нулевой и альтернативной гипотез *Н0: р1 = р2*; *Н1: р1 ≠ р2* критерии Z и χ2 являются эквивалентными. Однако, если необходимо не просто обнаружить различия, но и определить, какая доля больше (р1 > р2), *следует* применять Z-критерий с одной критической областью, ограниченной хвостом стандартизованного нормального распределения. Далее будет описано применение критерия χ2 для сравнения долей признака в нескольких группах. Необходимо отметить, что Z-критерий в этой ситуации применять невозможно.

**Применение χ2-критерия для проверки гипотезы о равенстве нескольких долей**

Критерий «хи-квадрат» можно распространить на более общий случай и применять для проверки гипотезы о равенстве нескольких долей признака. Обозначим количество анализируемых независимых генеральных совокупностей буквой *с*. Теперь таблица сопряженности признаков состоит из двух строк и *с* столбцов. Чтобы проверить нулевую и альтернативные гипотезы *Н0: р1 = р2* = … = *р2*, *Н1:* не все *рj* равны между собой *(j = 1, 2, …, c),* используется тестовая χ2-статистика:



где *f0* — наблюдаемое количество успехов или неудач в конкретной ячейке факторной таблицы 2х*с*, *fe* — теоретическое, или ожидаемое, количество успехов или неудач в конкретной ячейке таблицы сопряженности признаков при условии, что нулевая гипотеза является истинной.

Чтобы вычислить ожидаемое количество успехов или неудач в каждой ячейке таблицы сопряженности признаков, необходимо иметь в виду следующее. Если нулевая гипотеза является истинной и доли успехов во всех с генеральных совокупностях равны, соответствующие выборочные доли могут отличаться друг от друга лишь по случайным причинам, поскольку все доли представляют собой оценки доли признака *р* в общей генеральной совокупности. В этой ситуации статистика, объединяющая все доли в одной общей (или средней) оценке параметра *р*, содержит больше информации, чем каждая из них в отдельности. Эта статистика, обозначаемая символом *р̅*, представляет собой общую (или среднюю) долю успехов в объединенной выборке.

Вычисление средней доли:

*(3) p̅ =*

Чтобы вычислить ожидаемое количество успехов *fe* в первой строке таблицы сопряженности признаков, необходимо умножить объем каждой выборки на параметр *р̅*. Чтобы вычислить ожидаемое количество неудач *fe* во второй строке таблицы сопряженности признаков, необходимо умножить объем каждой выборки на параметр *1 – р̅*. Тестовая статистика, вычисленная по формуле (1), аппроксимируется χ2-распределением. Количество степеней свободы этого распределения задается величиной *(r – 1)(c – 1)*, где *r*— количество строк в факторной таблице, *с* — количество столбцов в таблице. Для факторной таблицы *2\*с* количество степеней свободы равно *(2 – 1)(с – 1) = с – 1*. При заданном уровне значимости α нулевая гипотеза отклоняется, если вычисленная χ2-статистика больше верхнего критического значения χU2, присущего χ2-распределению с *с – 1* степенями свободы. Таким образом, решающее правило выглядит следующим образом: гипотеза *Н0* отклоняется, если χ2 > χU2 (рис. 6), в противном случае гипотеза отклоняется.



Рис. 6. Критическая область χ2-критерия для сравнения с долей при уровне значимости α

*Проверка предположений, касающихся факторной таблицы 2\*с.* Для получения точных результатов на основе данных, приведенных в факторной таблице 2\**с*, необходимо, чтобы количество успехов или неудач было достаточно большим. Некоторые статистики полагают, что критерий дает точные результаты, если ожидаемые частоты превышают 0,5. Более консервативные исследователи требуют, чтобы не более 20% ячеек таблицы сопряженности признаков содержали ожидаемые величины, которые меньше 5, причем ни одна ячейка не должна содержать ожидаемую величину меньше единицы. Последнее условие нам представляется разумным компромиссом между этими крайностями. Чтобы удовлетворить это условие, категории, содержащие небольшие ожидаемые величины, следует объединить в одну. После этого критерий становится более точным. Если по каким-либо причинам объединение нескольких категорий невозможно, следует применять альтернативные процедуры.

Для того чтобы проиллюстрировать χ2-критерий для проверки гипотезы о равенстве долей в нескольких группах, вернемся к сценарию, описанному в начале главы. Рассмотрим аналогичный опрос, в котором принимают участие постояльцы трех отелей, принадлежащих компании Т. С. Resort Resources (рис. 7а).

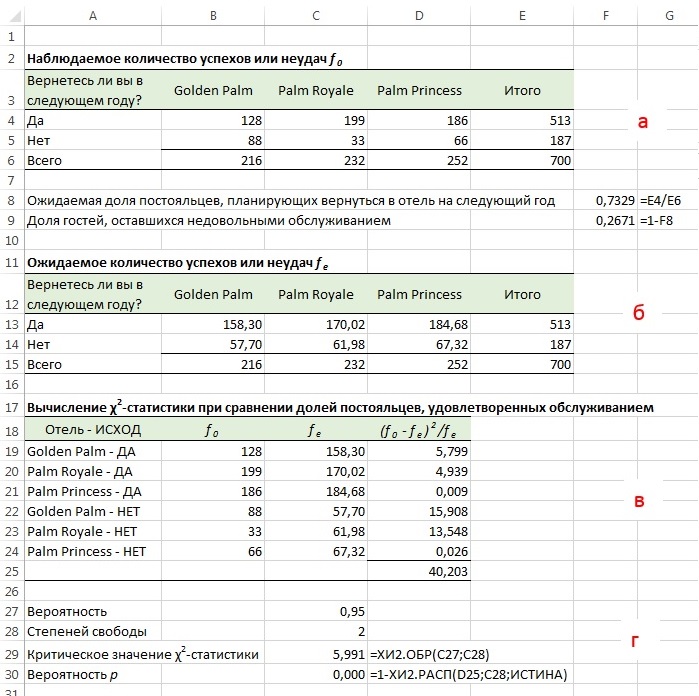


Рис. 7. Факторная таблица 2x3 для сравнения количества постояльцев, удовлетворенных и не удовлетворенных обслуживанием: (а) наблюдаемое количество успехов или неудач – *f0*; (б) ожидаемое количество успехов или неудач – *fe*; (в) вычисление χ2-статистики при сравнении долей постояльцев, удовлетворенных обслуживанием

Нулевая гипотеза утверждает, что доли клиентов, планирующих вернуться в следующем году, во всех отелях практически одинаковы. Для оценки параметра *р*, представляющего собой долю гостей, планирующих вернуться в отель, используется величина *р̅* = *Х / n* = 513 / 700 = 0,733. Доля гостей, оставшихся недовольными обслуживанием, равна 1 – 0,733 = 0,267. Умножая три доли на количество опрошенных постояльцев в каждом из отелей, получаем ожидаемое количество гостей, планирующих вернуться в следующем сезоне, а также число клиентов, которые больше не остановятся в этом отеле (рис. 7б).

Чтобы проверить нулевую и альтернативные гипотезы используют тестовую χ2-статистику, вычисленную с помощью ожидаемых и наблюдаемых величин по формуле (1) (рис. 7в).

Критическое значение тестовой χ2-статистики определяется по формуле =ХИ2.ОБР(). Поскольку в опросе принимают участие постояльцы трех отелей, χ2-статистика имеет (2 – 1)(3 – 1) = 2 степени свободы. При уровне значимости α = 0,05 критическое значение χ2-статистики равно 5,991 (рис. 7г). Так как вычисленная χ2-статистика, равная 40,236, превышает критическое значение, нулевая гипотеза отклоняется (рис. 8). С другой стороны, вероятность *р* того, что нулевая гипотеза верна при χ2-статистикие равной 40,236 (и двух степенях свободы) рассчитывается в Excel с помощью функции =1-ХИ2.РАСП() = 0,000 (рис. 7г). *р*-значение равно 0,000 и меньше уровня значимости α = 0,05. Следовательно, нулевая гипотеза отклоняется.

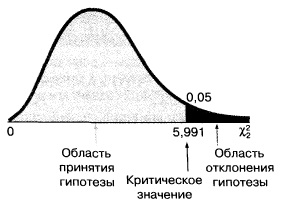


Рис. 8. Области принятия и отклонения гипотезы о равенстве трех долей при уровне значимости, равном 0,05, и двух степенях свободы

Отклоняя нулевую гипотезу при сравнении долей, указанных в факторной таблице 2\**с*, мы можем утверждать лишь, что доли постояльцев, удовлетворенных обслуживанием в трех отелях, не совпадают. Для того чтобы выяснить, какие доли отличаются от других, необходимо применять иные методы, например процедуру Мараскуило.

*Процедура Мараскуило* позволяет сравнивать все группы попарно. На первом этапе процедуры вычисляются разности psj – psj’ (где *j* ≠ *j’*) между *с(с – 1)/2* парами долей. Соответствующие критические размахи вычисляются по формуле:

*(4) Критический размах =*

При общем уровне значимости α, величина представляет собой квадратный корень из верхнего критического значения распределения "хи-квадрат", имеющего *с – 1* степеней свободы. Для каждой пары выборочных долей необходимо вычислить отдельный критический размах. На последнем этапе каждая из *с(с – 1)/2* пар долей сравнивается с соответствующим критическим размахом. Доли, образующие конкретную пару, считаются статистически значимо разными, если абсолютная разность выборочных долей |psj – psj| превышает критический размах.

Проиллюстрируем процедуру Мараскуило на примере опроса постояльцев трех отелей (рис 9а). Применяя критерий «хи-квадрат», мы убедились, что между долями постояльцев разных отелей, собирающихся вернуться в следующем году, существует статистически значимая разница. Поскольку в опросе участвуют постояльцы трех отелей, необходимо выполнить 3(3 – 1)/2 = 3 попарных сравнений и вычислить три критических размаха. Для начала вычислим три выборочных доли (рис. 9б). При общем уровне значимости, равном 0,05, верхнее критическое значение тестовой χ2-статистики для распределения «хи-квадрат», имеющего (с – 1) = 2 степени свободы определяется по формуле =ХИ2.ОБР(0,95;2) = 5,991. Итак, = 2,448 (рис. 9в). Далее, вычислим три пары абсолютных разностей и соответствующие критические размахи. Если абсолютная разность больше ее критического размаха, то соответствующие доли считаются значимо разными (рис. 9г).

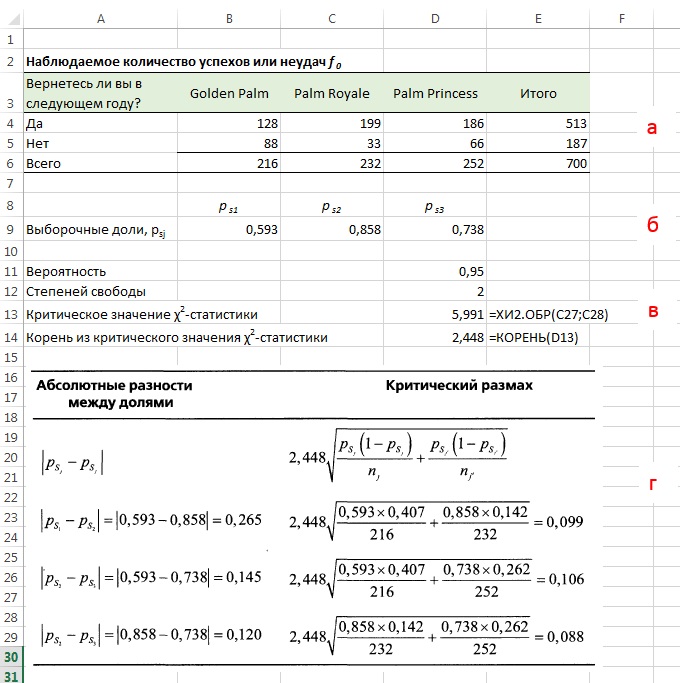


Рис. 9. Результаты выполнения процедуры Мараскуило для проверки гипотезы о равенстве долей удовлетворенных постояльцев трех отелей: (а) данные опроса; (б) выборочных доли; (в) верхнее критическое значение тестовой χ2-статистики для распределения «хи-квадрат»; (г) три пары абсолютных разностей и соответствующие критические размахи

Как видим, при уровне значимости, равном 0,05, степень удовлетворенности постояльцев отеля Palm Royal (ps2 = 0,858) выше, чем у постояльцев отелей Golden Palm (ps1 = 0,593) и Palm Princess (ps3 =0,738). Кроме того, степень удовлетворенности постояльцев отеля Palm Princess выше, чем у постояльцев отеля Golden Palm. Эти результаты должны заставить руководство проанализировать причины таких различий и попытаться определить, почему степень удовлетворенности постояльцев отеля Golden Palm значительно ниже, чем у постояльцев других отелей.

Предыдущая заметка [Блочный рандомизированный эксперимент](http://baguzin.ru/wp/?p=5948)

Следующая заметка

К оглавлению [Статистика для менеджеров с использованием Microsoft Excel](http://baguzin.ru/wp/?p=5285)

1. Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 708–730 [↑](#footnote-ref-1)