ЛАБОРАТОРНАЯ РАБОТА № 6. СТАТИСТИКА КСИ-КВАДРАТ ПИРСОНА

Хи-квадрат Пирсона один из самых популярных статистических критериев для анализа качественных данных (номинальных, порядковых, ранговых), анализа частот. Однако, как и у каждого статистического критерия у хи-квадрата есть свои собственные правила применения метода, его интерпретации.

Хи-квадрат используется прежде всего для анализа таблиц сопряженности (вид таблицы, которая учитывает совместное влияние фактора на исход, данные в таблице сопряженности должны быть представлены в виде частоты номинальных данных или интервалами, но не непрерывными количественными величинами). Стоит отметить, что при работе с сопряженными таблицами хи-квадрат часто является поддержкой для анализа влияния факторов риска с помощью расчета рисков (абсолютный и относительный риски) и отношение шансов.

Таблицы сопряженности могут принимать различные формы, простейшая таблица сопряженности выглядит следующим образом:

	Исход есть	Исхода нет	Всего
Фактор риска есть	A	B	A+B
Фактора риска нет	C	D	C+D
Всего	A+C	B+D	A+B+C+D

Как заполнить таблицу сопряженности? Обратимся к простому примеру:

Например, Вы хотите с помощью таблицы сопряженности и как следствия хи-квадрата Пирсона выяснить есть ли различия в частоте артериальной гипертонии в группах курящего и некурящего населения. Предполагается, что по остальным параметрам Ваши группы равномерны и превалирующим фактором риска развития артериальной гипертензии будет именно курение.

Для проведения исследования на основании ретроспективных данных (дизайн: случай-контроль) были отобраны две группы исследуемых — в первую вошли 70 человек, ежедневно выкуривающих не менее 1 пачки сигарет, во вторую группу вошли 80 некурящих такого же возраста, пола, и социального уровня (прочие систематически ошибки случайны).

В первой группе у 40 человек отмечалась артериальная гипертензия. Во второй — у 32 человек. Соответственно, референсное (нормальное) артериальное давление в группе «курильщиков» наблюдалось у 30 человек (70 — 40 = 30), а в группе «некурящих» нормальное АД наблюдалось у 48 (80 — 32 = 48).

Имея эти данные мы можем заполнить простейшую таблицу сопряженности:

	Повышенное АД	АД в пределах норма	Всего
«Курильщики»	40	30	70
«Не курят»	32	48	80

Всего	72	78	150

АД- артериальное давление

Как видно из таблицы: каждая строка соответствует группе пациентов, которая подвергается влиянию фактора, каждый столбец, в свою очередь, обозначает частоту исходов в группе (к примеру: произошло/ не произошло, как в нашем примере).

Таблицы сопряженности служат удобным средством визуализации комбинации частот «фактор- исход» и субстратом для расчета хи-квадрата Пирсона, который в нашем случае сможет дать статистически точный ответ о случайности или не случайности наших находок.

Условия применения статистического критерия хи-квадрата Пирсона

1. Тип данных: параметры должны быть качественными цельночисленными частотами, измеренными в номинальной шкале (Например, тип диагноза)

   бинарными (пол: мужской/женский, наличие или отсутствие заболевания)

   порядковыми (степень артериальной гипертензии),

2. Желательно, чтобы общее количество наблюдений было более 20,
3. Ожидаемая частота, соответствующая нулевой гипотезе должна быть более 5, если ожидаемое явление принимает значение менее 5, то необходимо использовать точный Критерий Фишера.
4. Для четырехпольных таблиц (2x2): Если ожидаемое значение принимает значение менее 10 (а именно 5<x<10), необходим расчет поправки Йетса таблиц сопряженности
5. Сравниваемые частоты должны быть примерно одного размера
6. Сопоставляемые группы должны быть независимыми (то есть единицы наблюдения в них разные, в отличие от связанных групп, анализирующих изменения «до-после» у одних и тех единиц наблюдений до и после вмешательства. Для таких ситуаций существует отдельный тест МакНемара (McNemar)

   Запрещается: использовать хи-квадрат для анализа непрерывных абсолютных данных, процентов и долей

Как рассчитать критерий хи-квадрат Пирсона?

Для оценки достоверности различий по методу хи-квадрата Пирсона (критерий соответствия, коэффициент согласия) анализируется различия между реальной существующими частотами в группах (Observed) и рассчитываемыми по формуле ожидаемыми «гипотетическими» частотами, которые соответствуют распределению хи-квадрат. При малом различии ожидаемых и наблюдаемых частот (хи-квадрат не достиг своего критического значения) мы принимаем нулевую гипотезу об отсутствии различий. Если же различия оказываются существенными (критическое значение хи-квадрата достигаются для заданного числа степеней свободы) мы отвергаем нулевую гипотезу и говорим о наличии статистически значимых различий.

Чем больше теоретические числа, рассчитанные на основе Но-гипотезы, **будут отличаться от фактических**, тем более «хи -квадрат» будет отличаться от 0, **тем с большей**

вероятностью можно отклонить Но-гипотезу и говорить о статистической достоверности имеющихся различий в сравниваемых совокупностях.

Основная формула для расчета хи-квадрата Пирсона:

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Зачем учитывать количество степеней свободы при расчете хи-квадрата?

Для того, чтобы не утомлять читателя пространными разъяснениями «о сумме квадратом нормально распределенных случайных величин» скажем лишь, что оценка критического значения хи-квадрата зависит от степени свободы изменения частот, что это значит на практике для пользователя хи-квадрата? То, что чем более многопольная таблица перед Вами, тем больше степеней свободы, чем она меньше, тем меньше. Формула расчета хи-квадрата следующая:

Degree of freedom (d.f.) = (c-1)(r-1)

Column (c) – количество столбцов частотами, r- количество строк с частотами.

Таким образом, количество степеней свободы для стандартной 2x2 таблицы сопряженности составит:

d.f. = (2-1)*(2-1)=1

и так далее.

Примеры расчета хи-квадрата Пирсона

Пример 1:

Необходимо определить наличие влияния предшествующей степени нарушения кровообращения на исход комиссуротомии (хирургическое разделение спаек при стенозе клапанного отверстия сердца). Пациенты поступали на комиссуротомию с различными исходными уровнями нарушения кровообращения. После комиссуротомии пациенты были выписаны с различными исходами операции.

Фактор: Степень нарушения кровообращения

Исход: Результативность операции

Таблица: наблюдаемые (Observed) частоты распределения влияния степени нарушения кровообращения на результаты операции комиссуротомии

Степень нарушения кровообращения	Всего больных	Выписан с хорошим	Выписан с удовлетворительным	Выписан с ухудшением

		результатом операции	результатом операции	
II	30	20	8	2
III	80	43	20	17
IV	60	10	40	10
Всего	170	73	68	29
H0-гипотеза	100%	43%	40%	17%

Первый этап

Расчет ожидаемых (Expected) величин (на основании групповых частот)

Степень нарушения кровообращения	Всего больных	"Ожидаемые" данные (P_1)		
		Хорошие	Удовлетвори-тельные	Ухудшение
II	30	13	12	5
III	80	34	32	14
IV	60	26	24	10
Всего	170	73	68	29
H_0-гипотеза	100%	43%	40%	17%

Второй этап

Сопоставление наблюдаемых и ожидаемых частот с нахождением их разницы (O-E)

Степень нарушения кровообращения	Выписан с хорошим результатом операции	Выписан с удовлетворительным результатом операции	Выписан с ухудшением
II	+7	-4	-3
III	+9	-12	+3
IV	-16	+16	0
Всего	0	0	0

Третий этап

Рассчитываем сумму отношений квадрата разности значений и делим ожидаемые данные (хи-квадрат) $(O-E)^2/E$

Степень нарушения кровообращения	Выписан с хорошим результатом операции	Выписан с удовлетворительным результатом операции	Выписан с ухудшением
II	49/13=3,77	16/12=1,33	9/5=1,80
III	81/34=2,38	144/32=4,50	9/14=0,64

IV	256/26=9,85	256/24=10,66	0/10*=0,10
Всего	16	16,49	2,54

как видно из данной таблицы одно из ожидаемых значений равно 0, в данном случае будет подставлена 1, корректнее применить точный критерий Фишера (см. Условия применения хи-квадрата Пирсона)

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 35{,}03$$

Четвертый этап

Необходимо соотнести полученное значение хи-квадрата с критическим значением хи-квадрата.Возникает вопрос, откуда брать критическое значение? Критическое значение хи-квадрата, как и для большинства, статистических критериев зависит от степени свободы и уровня достоверности (alpha), который Вы выбираете.В нашем случае, наше количество степеней свободы равно (3-1)*(3-1)=4, уровень значимости, который мы хотим соблюсти равен 0,05Обратимся к таблице критических значение хи-квадрата:

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.84	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.99	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314

- Хи-квадрат (для d.f.=4 p=0.05) = 9.488
- Хи-квадрат (для d.f.=4 p=0.01) = 13.27735,03 > 13,277;
- p<0,01

Пример корректной интерпретации: Предшествующая степень нарушения кровообращения влияет на исход комиссуротомии (однако! Мы не можем говорить о направленности связи, то есть: улучшает-ухудшает сказать не можем), оптимально указать степень свободы, точное

значение хи-квадрата, если есть возможность рассчитать точное значение достоверности, то так же стоит указать и его или остановиться на критическом значении достоверности (p<0,05 или p<0,01 и так далее).В нашем случае:d.f.=4, x^2=35,03, p< 0.01
Пример 2: Вернемся к нашему примеру с влиянием курения на развитие артериальной гипертензии:Исходная четырехпольная таблица:

	Повышенное АД	АД в пределах норма	Всего
«Курильщики»	40	30	70
«Не курят»	32	48	80
Всего	72	78	150

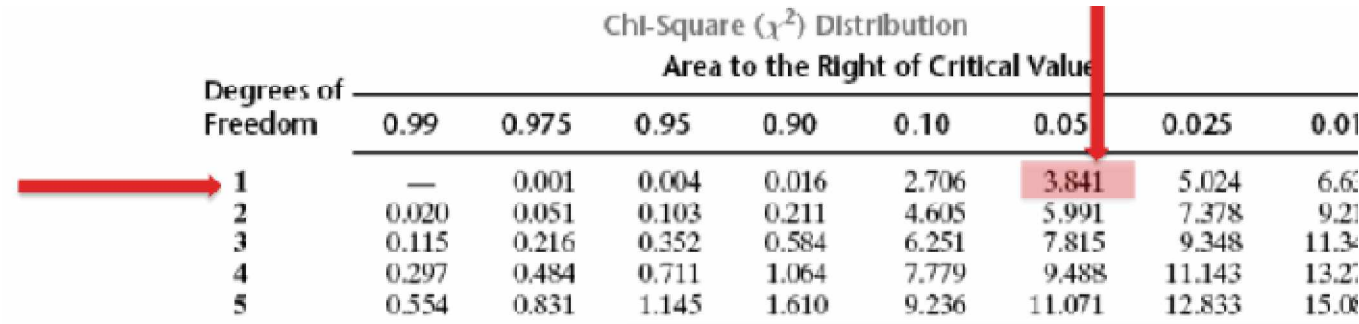
Для четырехпольных таблиц существует упрощенная формула расчета значения хи-квадрата:

	Исход +	Исход 0	Всего
Фактор +	a	b	a+b
Фактор 0	c	d	c+d
Всего	a+c	b+d	N

$$\chi^2 = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

- x^2= (40x48 – 32x30)x150 / (70)(80)(72)(78) = $(1920 – 960)^2$x150/31449600 = 138240000/31449600 = 4,395
- Сравним полученное значение хи-квадрата с критическим значением (для степени свободы 1, и уровнем значимости 3,841)

Правильная интерпретация: Курение оказывает влияние на формирование повышенного артериального давления *df*=1, x^2= 4,395, p<0,05



Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.6
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.2
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.3
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.2
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.0

Заключение по хи-квадрату Пирсона

хи-квадрат Пирсона является удобным статистическим методом для анализа изменения частот, оформленными в таблицы сопряженности для несвязанных групп. Как и все статистически

инструменты хи-квадрат Пирсона имеет свои правила, преимущества и ограничения применения. Будьте внимательны и хи-квадрат Пирсона Вас не разочарует.