



Факультет гуманитарных наук

Школа филологии

Москва
2025

Про Хи-квадрат

Екатерина Дмитриева



Вступление. Для чего используется хи-квадрат?

- Когда нам нужно проверить, соответствуют ли наблюдения ожидаемому распределению
- Когда нужно проверить, «одинаковые» ли у нас выборки



Задача: проверить, принадлежит ли наша выборка к заданному распределению (гипотеза: принадлежит)

- Пусть у нас есть N наблюдений X_1, X_2, \dots, X_n
- Каждое из наблюдений – число от 1 до k
- Пусть количество наблюдений типа i равно O_i
- У нас есть гипотеза: вероятность выпадения i равна p_i
- Посчитаем вот такое число

$$\chi = \sum_{i=1}^k \frac{(O_i - N \cdot p_i)^2}{N \cdot p_i}$$



Интуиция

$$\chi = \sum_{i=1}^k \frac{(O_i - N \cdot p_i)^2}{N \cdot p_i}$$

- Если наша гипотеза верна, то наиболее ожидаемое количество наблюдений типа i равно $N \cdot p_i$
- Чем больше получилось число χ , тем страннее наш набор наблюдений и тем маловероятнее гипотеза
- При *достаточно больших* значениях χ разумно предположить, что гипотеза неверна



Что такое достаточно большое значение?

- Если вероятность получить столь большое значение χ случайно меньше α (уровень значимости), будем считать, что такого не бывает, следовательно, гипотеза отклоняется.
- NB! Уровень значимости α определяется ДО проведения эксперимента
- Мы можем взять вероятность пронаблюдать столь большой χ (критическое значение) из таблицы распределения хи-квадрат для $(k - 1)$ степени свободы
- В филологии считается допустимой даже $\alpha = 30\%$



Какие можно сделать выводы?

- Если χ получилась больше критического значения, мы можем отклонить нулевую гипотезу
- Если χ получилась меньше, мы не можем отвергнуть нулевую гипотезу
- Второе само по себе не подтверждение гипотезы, но аргумент в её пользу
- Подтвердить гипотезу статистическими методами вообще невозможно



Когда это можно использовать?

- У вас должно быть n наблюдений, каждое наблюдение – одного из фиксированного набора типов
- Хорошо, если число наблюдений $n \geq 50$
- Хорошо, если каждое из наблюдений не очень редкое, $N \cdot p_i \geq 5$
- Если совсем уж никак, можно использовать, если пункты выше не соблюдаются, но не стоит удивляться странным результатам
- NB! Хи-квадрат работает с ЧИСЛАМИ, не с долями, не с процентами

Задача: одинаковы ли две выборки?

I	II	III	IV	V	VI	VII	VIII	Итого
a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	N_a
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	N_b

- Пусть у нас есть два набора наблюдений, и мы хотим проверить, одинаковые ли они. (Гипотеза: они одинаковые)
- Если они одинаковые, то стоит ожидать вероятность i -того класса примерно $p_i = \frac{a_i + b_i}{N_a + N_b}$ (это частота «склеенной» выборки, можно думать, что мы принимаем за эталон ее)

$$\chi = \sum_{i=1}^k \frac{(a_i - N_a \cdot p_i)^2}{N_a \cdot p_i} + \sum_{i=1}^k \frac{(b_i - N_b \cdot p_i)^2}{N_b \cdot p_i}$$

- Критическое значение берем из таблицы для $(k - 1) \cdot (N_{\text{строчек}} - 1) = (8 - 1)(2 - 1) = 7$ степеней свободы



Практические пример: сравнить использования форм ямба Гандлевским и Мандельштамом

форма	Гандлевский	Мандельштам	числа М	Числа Г	Гипотеза о вероятностях	Chi-квадрат М	Chi-квадрат Г	
I	0,09268	0,06666	14	19	0,0795180 7229	0,4361689 181	0,44680718 44	4,9878662 76
II	0,07317	0,03809	8	15	0,0554216 8675	1,1375190 2	1,16526338 7	
III	0,19512	0,18095	38	40	0,1879518 072	0,0547391 0293	0,05607420 3	
IV	0,22439	0,27142	57	46	0,2481927 711	0,4568203 466	0,46796230 62	
V	0,00975	0,00476	1	2	0,0072289 15663	0,1768024 479	0,18111470 27	
VI	0,24878	0,25714	54	51	0,2530120 482	0,0141627 7354	0,01450820 704	
VII	0,15609	0,18095	38	32	0,1686746 988	0,1876731 415	0,19225053 52	
VIII	0	0	0	0	0	0	0	
			210	205	415	2,4638857 51	2,52398052 5	



Выводы

- Вы теперь умеете применять критерий хи-квадрат
- При помощи него можно проверять гипотезы вида «это и это – одно и то же»
- Хи-квадрат можно использовать для сравнения использования форм ямба, сопоставления выборки с заранее заданным эталоном, при вопросе об одинаковости ритмики, проверке статистических гипотез (см. исследования М. Красноперовой), etc
- Применять хи-квадрат в стиховедческих исследованиях – хорошая идея, потому что есть большая традиция его использования, а значит, хорошая сопоставимость результатов и уверенность, что работает



Таблица критических значений

	a = 0,3	a=0,05	a=0,01	a=0,005		a = 0,3	a=0,05	a=0,01	a=0,005
k = 1	1,074	3,841	6,635	7,879	k = 12	14,011	21,026	26,217	28,300
k = 2	2,408	5,991	9,210	10,597	k = 13	15,119	22,362	27,688	29,819
k = 3	3,665	7,815	11,345	12,838	k = 14	16,222	23,685	29,141	31,319
k = 4	4,878	9,488	13,277	14,860	k = 15	17,322	24,996	30,578	32,801
k = 5	6,064	11,070	15,086	16,750	k = 16	18,418	26,296	32,000	34,267
k = 6	7,231	12,592	16,812	18,548	k = 17	19,511	27,587	33,409	35,718
k = 7	8,383	14,067	18,475	20,278	k = 18	20,601	28,869	34,805	37,156
k = 8	9,524	15,507	20,090	21,955	k = 19	21,689	30,144	36,191	38,582
k = 9	10,656	16,919	21,666	23,589	k = 20	22,775	31,410	37,566	39,997
k = 10	11,781	18,307	23,209	25,188	k = 21	23,858	32,671	38,932	41,401
k = 11	12,899	19,675	24,725	26,757	k = 22	24,939	33,924	40,289	42,796

