ЛАБОРАТОРНАЯ РАБОТА № 1

КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА (ХИ-КВАДРАТ) И КРИТЕРИЙ КОЛМОГОРОВА-СМИРНОВА

Методические указания

Широко используемыми на практике **критериями проверки статистических гипотез** выступают следующие:

- критерий согласия Хи-квадрат
- критерий Крамера-фон Мизеса
- критерий Колмогорова-Смирнова

Критерий Хи-квадрат предпочтителен, когда исследуются большие объемы выборок. При малых объемах выборок этот критерий практически не пригоден.

Нулевая гипотеза при применении общих критериев согласия записывается в форме

$$H0: F_n(x) = F(x),$$

где $F_n(x)$ – эмпирическая функция распределения вероятностей; $F(x)$ – гипотетическая функция распределения вероятностей.

Критерий Пирсона X^2 основан на сравнении эмпирической гистограммы распределения случайной величины с ее теоретической плотностью. Диапазон изменения экспериментальных данных разбивается на k интервалов, и подсчитывается статистика:

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

где n_i – количество значений случайной величины, попавших в i-й интервал; n – объем выборки; $F(x)$ – гипотетический теоретический закон распределения вероятностей случайной величины; $p_i = F(x_{i+1})$ — $\mathrm{F}(x_i)$ – теоретическая вероятность попадания случайной величины в i-й интервал.

Статистика X^2 имеет распределение Хи-квадрат с $f = n$ — 1 степенями свободы в том случае, когда проверяется простая нулевая гипотеза H0, т.е., когда гипотетическое распределение, на соответствие которому проверяется эмпирический ряд данных, известно с точностью до значения своих параметров.

Правило проверки гипотезы:

$$\text{если } X^2 > X^2_{\text{alpha}}(f)$$

то на уровне значимости *alpha*, т. е. с достоверностью (1 — alpha) гипотеза

H0 отклоняется.

На мощность статистического критерия X^2 сильное влияние оказывает *число интервалов* разбиения гистограммы (k) и порядок ее разбиения (т. е. выбор длин интервалов внутри диапазона изменения значений случайной величины). На практике принято считать, что статистику X^2 можно использовать, когда np_i >= 5.

Такое приближение допустимо и тогда, когда не более, чем в 20% интервалов имеет место 1 <= np_i <= 5.

Одна из рекомендаций по расчету k сводится к вычислению:

$$k = 1 + 3{,}32 \cdot \lg n$$

При n >= 200 можно выбирать k из условия

$$k = 4 \cdot \{0{,}75 \cdot (n — 1)^2\}^{1/5} \approx 3{,}78 \cdot (n — 1)^{2/5}.$$

Еще одно простое правило: выбрать как можно большее k, но не превышающее $n/5$:

$$k <= n / 5$$

Критерий Колмогорова-Смирнова также целесообразно использовать для выборки указанных объемов в тех случаях, когда проверяемое распределение непрерывно и известны среднее значение и дисперсия проверяемой совокупности.

Алгоритм реализации критерия Колмогорова-Смирнова предполагает использование критического значения D_{extr} для проверки принятой гипотезы. Для этого используется таблица приведенная в Приложении 1.

Пример выполнения задания

В таблице приведены данные по ежедневному числу дорожно-транспортных происшествий в городе:

Таблица 1

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
	2	3	4	6	4	3
5	4	2	1	4	5	3
4	5	3	5	8	2	2
3	1	3	6	2	1	3
2	7	1				

Приняв уровень значимости α=0,05, ***проверить согласие*** этих данных обычного месяца ***с распределением Пуассона***, пользуясь критерием Хи-квадрат. Перепроверить данные с помощью критерия Колмогорова-Смирнова, по прежнему принимая α =0,05.

Решение

1. Вычислим среднее значение (***x_{cp}***), дисперсию (***D***) и ожидаемое среднее (**λ**) количества дорожно-транспортных происшествий в городе за представленный месяц.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Данные по числу дорожно-транспортных происшествий в городе за месяц**								**Количество дорожных проишествий по дням (x_i)**	**$x=x_i$ - $x_{ср}$**	**x^2**	
2	**ПН**	**ВТ**	**СР**	**ЧТ**	**ПТ**	**СБ**	**ВС**		5	1,5333	2,3511	
3		2	3	4	6	4	3		4	0,5333	0,2844	
4	5	4	2	1	4	5	3		3	-0,4667	0,2178	
5	4	5	3	5	8	2	2		2	-1,4667	2,1511	
6	3	1	3	6	2	1	3		2	-1,4667	2,1511	
7	2	7	1						4	0,5333	0,2844	
8									5	1,5333	2,3511	
9									1	-2,4667	6,0844	
10	**n**	30			**Σ**	95,4667			7	3,5333	12,4844	
11	**$x_{ср}$**	3,4667							3	-0,4667	0,2178	
12	**D**	3,291954							2	-1,4667	2,1511	
13	**λ**	3,3793103							3	-0,4667	0,2178	
14									3	-0,4667	0,2178	

Рисунок 1 – Результат вычисления среднего значения (***x_{cp}***), дисперсии (***D***) и ожидаемого среднего (**λ**)

Формулы ячеек представлены в табл. 2.

Таблица 2

Формулы ячеек

Ячейка	Характеристика	Формула
B11	– среднее значение (***x_{cp}***)	=СРЗНАЧ(A3:G7)
B12	– дисперсия (***D***)	=F10/(30-1)
B13	– ожидаемое среднее (**λ**)	=(B12+B11)/2
F10	– $\Sigma x=(x_i - x_{ср})^2$	=СУММ(K2:K31)

2. Вычислим число случаев исхода (ob), вероятность наступления N происшествий (P), ожидаемое число случаев исхода (ex), количество значений случайной величины, попавших в интервал (obs), вероятность попадания случайной величины в интервал (exp), статистику χ^2

	A	B	C	D	E	F	G	H
35								
36	**N**	**ob**	**P**	**ex**	**Интервал**	**obs**	**exp**	**(obs-exp)2/exp**
37	0	0	0,0340709	1	≤2	10	10	0
38	1	4	0,1151363	3	3	7	7	0
39	2	6	0,1945406	6	4	5	6	0,1667
40	3	7	0,2191377	7	≥5	8	7	0,1429
41	4	5	0,1851336	6			**χ2**	0,3095
42	5	4	0,1251248	4				
43	6	2	0,0704726	2				
44	7	1	0,0340212	1				
45	8	1	0,014371	0				
46	**Σ**	30	0,9920088	30				
47								

Рисунок 2 – Результаты вычисления п.2

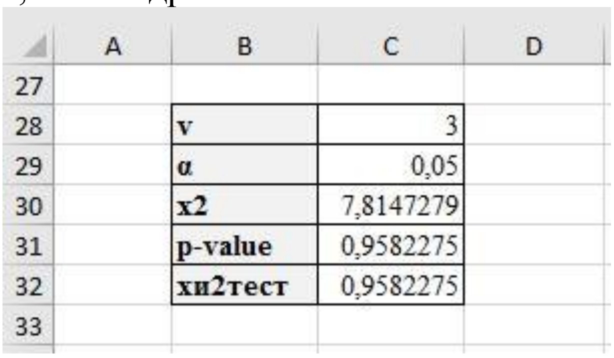
Формулы ячеек представлены в табл. 3.

Таблица 3

Формулы ячеек

Ячейка	Характеристика	Формула
B37	– число случаев исхода	=СЧЁТЕСЛИ($A$3:$G$7;A37)
C37	– вероятность наступления	=ПУАССОН.РАСП(A37;$B$13;ЛОЖЬ)
D37	– ожидаемое число случаев исхода	=ОКРУГЛ(C37*$B$10;0)
H41	– статистика Хи-квадрат	=СУММ(H37:H40)
F37	– количество значений в указанном интервале (obs)	=СЧЁТЕСЛИ(A3:G7;"<=2")
G37	– сумма ожидаемых исходов в указанном интервале (exp)	=СУММ(D37:D39)

3. Вычислим критическое значение Хи-квадрата (максимальное значение для заданного уровня значимости), вероятность получить расчетное значение Хи-квадрат, Хи-квадрат тест



	A	B	C	D
27				
28		**v**	3	
29		**α**	0,05	
30		**x2**	7,8147279	
31		**p-value**	0,9582275	
32		**хи2тест**	0,9582275	
33				

Рисунок 3 – Результаты вычисления п.3

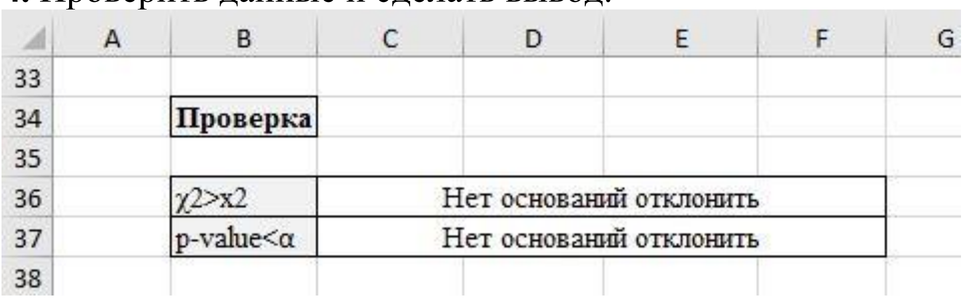
Формулы ячеек представлены в табл. 4.

Таблица 4

Формулы ячеек

Ячейка	Характеристика	Формула
C28	– степень свободы (v)	=кол-во интервалов-1
C30	– критическое значение Хи-квадрата (максимальное значение для заданного уровня значимости)	=ХИ2.ОБР(1- C29; C28)
C31	– p-value (вероятность получить расчетное значение Хи-квадрата)	=ХИ2.РАСП.ПХ(H41; C28)
C32	– Хи-квадрат тест	=ХИ2.ТЕСТ(F37:F40;G37:G40)

4. Проверить данные и сделать вывод.



	A	B	C	D	E	F	G
33							
34		**Проверка**					
35							
36		χ2>x2	Нет оснований отклонить				
37		p-value<α	Нет оснований отклонить				
38							

Рисунок 4 – Результаты проверки гипотез

5. Перепроверить данные с помощью критерия Колмогорова-Смирнова, по-прежнему принимая α =0,05, с помощью таблицы, приведенной в ПРИЛОЖЕНИИ 1.

Рассчитать наблюдаемую вероятность **(PN)**, теоретическую вероятность **(P)**, интегральные вероятности **(PNI и PI)**, абсолютную разность **(AR)**.

	A	B	C	D	E	F
30						
31	**N**	**PN**	**P**	**PNI**	**PI**	**AR**
32	0	0	0,034071	0	0,034071	0,034071
33	1	0,133333	0,115136	0,133333	0,149207	0,015874
34	2	0,2	0,194541	0,333333	0,343748	0,010415
35	3	0,233333	0,219138	0,566667	0,562886	0,003781
36	4	0,166667	0,185134	0,733333	0,748019	0,014686
37	5	0,133333	0,125125	0,866667	0,873144	0,006477
38	6	0,066667	0,070473	0,933333	0,943617	0,010283
39	7	0,033333	0,034021	0,966667	0,977638	0,010971
40	8	0,033333	0,014371	1	0,992009	0,007991
41					**max**	0,034071
42						
43					**Dkp**	0,248301

Рисунок 5 – Результаты вычисления п.5

Формулы ячеек представлены в табл. 5.

Таблица 5

Формулы ячеек

Ячейка	Характеристика	Формула
B32	– число случаев исхода	=СЧЁТЕСЛИ($A$3:$G$7;A32)/30
C32	– вероятность наступления	=ПУАССОН.РАСП(A32;$B$13;ЛОЖЬ)
D33	– интегральная вероятность (PNI)	=D32+B33
E33	– интегральная вероятность (PI)	=E32+C33
F32	– абсолютную разность (AR)	=ABS(D32-E32)
F41	– максимальная разность	=МАКС(F32:F40)
F43	– табличное значение	=1,36/КОРЕНЬ(30)

Выполнить проверку:

	A	B	C	D	E	F
45						
46			*Гипотеза, о том что*			
47	**max (AR)<Dkp**		*экспериментальное*			
48			*распределение*			
49						
50						

Рисунок 6 – Результаты проверки гипотезы по критерию Колмогорова-Смирнова

Задача для самостоятельного выполнения

В таблице приведены данные по ежедневному числу инфекционных заболеваний в городе. Приняв уровень значимости $\alpha=0{,}05$, ***проверить согласие*** этих данных обычного месяца ***с распределением Пуассона***, пользуясь критерием Хи-квадрат. Перепроверить данные с помощью критерия Колмогорова-Смирнова, по прежнему принимая $\alpha=0{,}05$.

Вариант 1

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
			8	5	6	5
3	2	7	2	1	5	4
1	3	3	4	7	3	1
3	3	5	3	4	2	2
4	1	1	2	3	5	

Вариант 2

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
			8	5	6	5
4	1	1	2	1	5	4
3	3	3	4	7	3	1
1	3	5	3	4	2	2
3	2	7	2	3	5	

Вариант 3

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		6	8	5	6	5
4	1	1	2	1	5	4
3	3	3	4	7	3	1
1	3	5	3	4	2	2
3	2	7	2			

Вариант 4

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		6	8	5	6	5
4	1	1	2	1	5	4
3	3	3	4	7	3	1
1	3	5	3	4	2	2
3	2	7	2	5		

Вариант 5

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
6	7	6	8	5	6	5
4	1	1	2	1	5	4
3	3	3	4	7	3	1
1	3	5	3	4	2	2
3	2	7				

Вариант 6

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
	8	5	7	6	6	5
1	4	5	2	1	1	4
3	3	3	4	7	3	1
1	5	3	3	2	4	2
3	2	7	6			

Вариант 7

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		3	7	6	6	5
1	2	5	2	1	4	4
3	3	3	4	5	3	3
1	5	5	3	2	4	2
3	4	7	6			

Вариант 8

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		1	7	8	4	5
4	2	5	2	1	4	4
2	3	3	4	2	3	3
1	2	4	3	2	5	2
3	4	7	8	3		

Вариант 9

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
			7	6	4	5
3	2	5	2	1	4	4
2	3	4	4	2	3	3
4	2	4	3	4	5	2
3	4	6	5	3	2	

Вариант 10

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		8	7	6	4	5
1	3	6	3	2	5	4
2	3	4	4	2	3	3
4	2	4	3	4	5	2
3	4	6	5			

Вариант 11

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
	5	6	7	6	4	5
1	3	6	5	2	5	1
2	3	7	4	5	3	3
3	2	4	3	4	6	2
4	4	8				

Вариант 12

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
		1	6	8	4	5
6	3	6	5	2	5	4
2	4	5	7	5	3	3
3	2	4	3	4	6	2
4	4	8	6	1		

ПРИЛОЖЕНИЕ 1

Критические числа Колмогорова–Смирнова

Степень свободы	Проверка единичной выборки *			Проверка двух выборок **	
N	$D_{0,10}$	$D_{0,05}$	$D_{0,01}$	$D_{0,05}$	$D_{0,01}$
1	0,950	0,975	0,995	–	–
2	0,776	0,842	0,929	–	–
3	0,642	0,708	0,828	–	–
4	0,564	0,624	0,733	1,000	1,000
5	0,510	0,565	0,669	1,000	1,000
6	0,470	0,521	0,618	0,833	1,000
7	0,438	0,486	0,577	0,857	0,857
8	0,411	0,457	0,543	0,750	0,875
9	0,388	0,432	0,514	0,668	0,778
10	0,368	0,410	0,490	0,700	0,800
11	0,352	0,391	0,468	0,636	0,727
12	0,338	0,375	0,450	0,583	0,667
13	0,325	0,361	0,433	0,538	0,692
14	0,314	0,349	0,418	0,571	0,643
15	0,304	0,338	0,404	0,533	0,600
16	0,295	0,328	0,392	0,500	0,625
17	0,286	0,318	0,381	0,471	0,588
18	0,278	0,309	0,371	0,500	0,556
19	0,272	0,301	0,363	0,474	0,526
20	0,264	0,294	0,356	0,450	0,550
25	0,240	0,270	0,320	0,400	0,480
30	0,220	0,240	0,290	0,370	0,430
35	0,210	0,230	0,270	0,340	0,390
Более 35	$\frac{1,22}{\sqrt{N}}$	$\frac{1,36}{\sqrt{N}}$	$\frac{1,63}{\sqrt{N}}$	$1,36\sqrt{\frac{N_1+N_2}{N_1N_2}}$	$1,63\sqrt{\frac{N_1+N_2}{N_1N_2}}$

* Применяется для оценки степени близости выборочных значений к теоретическому распределению.

N – объем выборки.

** Применяется для определения принадлежности двух выборок объемами N_1 и N_2 одному и тому же распределению. При малых размерах выборки $N = N_1 = N_2$.