

Проверка непараметрических гипотез

Проверка гипотез о виде закона распределения

Иногда можно предположить вид закона распределения генеральной совокупности.

При этом возникает вопрос: на сколько наше предположение соответствует имеющимся экспериментальным данным?

Пусть $g(x; \theta)$, $G(x; \theta)$ – истинный закон распределения генеральной совокупности,
 $f(x; \theta)$, $F(x; \theta)$ – гипотетический закон распределения генеральной совокупности

$$H_0 : g(x; \theta) = f(x; \theta) \quad \text{или} \quad H_0 : G(x; \theta) = F(x; \theta)$$

H_1 : нулевая гипотеза не верна

Используются **выборочные критерии согласия*** (статистики) -- статистические критерии, предназначенные для обнаружения расхождения между гипотетическим законом распределения и реальными статистическими данными (реализацией выборки)

Такие критерии основаны на различных мерах расстояний между анализируемой эмпирической и гипотетической формой закона распределения семейством функций.

Простые, т. е. параметры закона распределения заранее известны

Сложные, когда параметры не известны

- 1. Критерии, основанные на изучении разницы между теоретической плотностью распределения и эмпирической гистограммой. *Например, критерий Хи-квадрат Пирсона*
- 2. Критерии, основанные на расстоянии между теоретической и эмпирической функциями распределения вероятностей. *Например, критерий Колмогорова*
- 3. Корреляционно-регрессионные критерии, основанные на изучении корреляционных и регрессионных связей между эмпирическими и теоретическими порядковыми статистиками.

* Так называют, чтобы выделить данные критерии из множества остальных (параметрических)

Данный критерий
основывается на этой теореме

1. Критерий Хи-квадрат Пирсона

Теорема Пирсона : Если при проведении n испытаний, в которых случайная величина может принимать m различных значений A_1, A_2, \dots, A_m с истинными вероятностями p_1, p_2, \dots, p_m , частота появления данных событий соответствует n_1, n_2, \dots, n_m , то т.е. испытания Бернульи, только с конечным числом исходов выборочная функция:

$$\chi^2 = \sum_{j=1}^m \frac{(n_j - np_j(\theta))^2}{np_j(\theta)} = \sum_{j=1}^m c_j \left(\frac{n_j}{n} - p_j(\theta) \right)^2, \quad c_j = \frac{n}{p_j(\theta)}$$

сходится при $(n \rightarrow \infty)$ к $H^2(m-1)$ – распределению.

Поскольку в силу центральной предельной теоремы (следствие, теорема Бернульи)

$$W_A = \frac{n_A}{n} \xrightarrow{\frac{P}{n}} p_A,$$

то значение случайной величины χ^2 будет близко к нулю. Поэтому значение χ^2 можно рассматривать как критерий верности гипотезы о законе распределения. Если гипотеза верна и мы верно предположили истинный закон распределения, то частоты будут стремиться к истинным вероятностям, а значение критерия к нулю. Если гипотеза не верна, то критерий будет равен какому-то ненулевому значению.

Выборочную функцию χ^2 называют **выборочным критерием согласия хи-квадрат Пирсона**.

А саму гипотезу иногда переписывают в виде:

$$H_0 : g(x; \theta) = f(x; \theta) \quad \begin{matrix} \text{Можно заменить} \\ \Rightarrow \end{matrix} \quad H_0 : p_i = p_i^h, \quad i = 1 \dots m$$

1. Критерий Хи-квадрат Пирсона. Геометрическая интерпретация и свойства

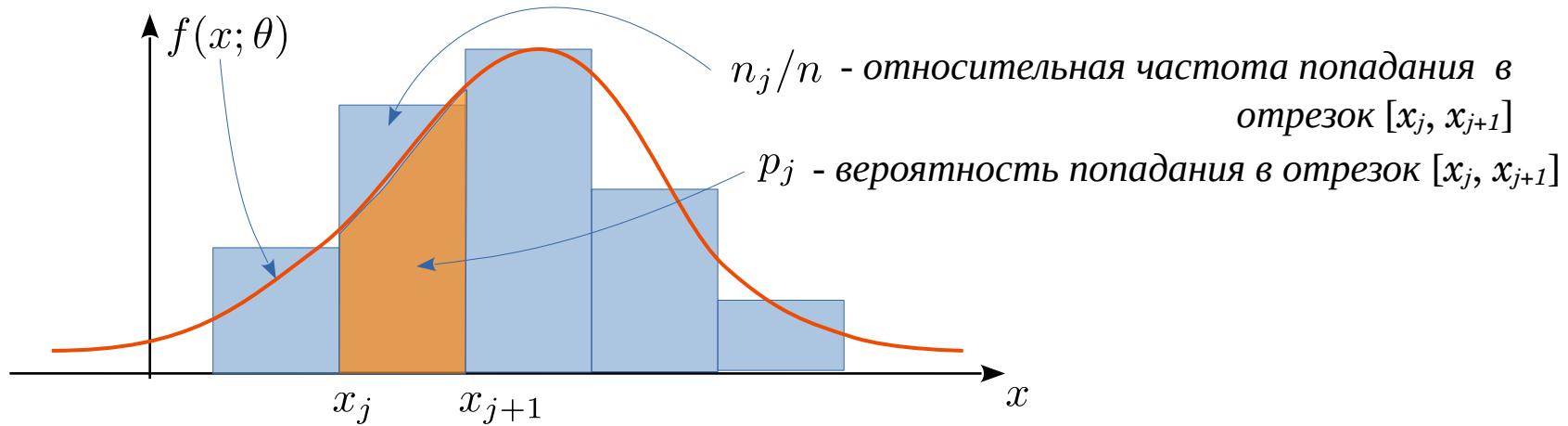
2

Критерий Хи-квадрат Пирсона

$$\chi^2 = \sum_{j=1}^m \frac{(n_j - np_j(\theta))^2}{np_j(\theta)} = \sum_{j=1}^m c_j \left(\frac{n_j}{n} - p_j(\theta) \right)^2, \quad c_j = \frac{n}{p_j(\theta)}$$

где $\chi^2 \in H^2(m-1)$ при $(n \rightarrow \infty)$

Чем меньше вероятность p_j , тем больше коэффициент c_j . Таким образом отсекаются гистограммы сильно отличающиеся на концах (где p_j мало).



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

Рекомендуется при $n > 200$ выбирать m из условия: $m = 4(0,75(n-1)^2)^{1/5} \approx 3,78(n-1)^{2/5}$

Алгоритм проверки гипотезы о виде закона распределения на основе критерия Хи-квадрат (сложной и простой)

Для сложной гипотезы, т. е. когда не известны значения параметров

- 1 * По выборочным данным x_1, x_2, \dots, x_n строим оценки $\hat{\theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_k\}$ параметров выбранного закона распределения $F(x; \theta_1, \dots, \theta_k)$ (используя методы моментов или максимального правдоподобия).

- 2 Диапазон изменения экспериментальных данных разбивается на m интервалов. Определяется относительная частота $n_j/n, j = 1..m$ попадания данных в каждый интервал.

- 3 Находится вероятность попадания случайной величины в j -й интервал:

$$p_j = F(x_{j+1}; \theta) - F(x_j; \theta) \quad \text{или} \quad p_j = F(x_{j+1}; \hat{\theta}) - F(x_j; \hat{\theta})$$

- 4 Вычисляем значение выборочной статистики χ^2

Для сложной гипотезы этот критерий вычисляется аналогично, только подставляются вычисленные в пункте 1 параметры

- 5 По заданному уровню значимости критерия α и числу степеней свободы $m - 1$ (или $m - k - 1$ при неизвестных параметрах θ) из таблиц процентных точек χ^2 - распределения находим критическое значение статистического критерия $\chi^2_{kp,\alpha}$.
k - количество параметров закона распределения

- 6 H_0 принимается, если $\chi^2 < \chi^2_{kp,\alpha}$ и отвергается в противном случае $\chi^2 \geq \chi^2_{kp,\alpha}$

* Для случая, когда параметры закона распределения не известны

П р и м е р : Производился ремонт однотипных агрегатов шасси 100 самолетов.

Фиксировались значения времени на эту работу – значения случайной величины T .

Требуется проверить гипотезу H_0 : СВ T распределена по нормальному закону при уровне значимости $\alpha = 0.10$.

Разряды $\{t_j \dots t_{j+1}\}$, ч	[50...65]	(65...80]	(80...95]	(95...110]	(110...125]	(125...140]	(140...155]
Частоты $P_j = n_j/n$	0,08	0,16	0,18	0,25	0,17	0,11	0,05

1. По результатам наблюдений определяем параметры $\hat{m}_t = \bar{t} = 98.37$, $\hat{\sigma}_t = s = 23.59$

2. Вычисляем значение критерия

Вычисления	Разряды							
	[50...65]	(65...80]	(80...95]	(95...110]	(110...125]	(125...140]	(140...155]	
1	n_j	8	16	18	25	17	11	5
2	p_j	0,06	0,14	0,23	0,33	0,18	0,15	0,03
3	np_j	6	14	23	33	18	15	3
4	$n_j - np_j$	2	2	-5	-8	-1	-4	2
5	$(n_j - np_j)^2$	4	4	25	64	1	16	4
6	$\frac{(n_j - np_j)^2}{np_j}$	0,67	0,29	1,09	1,94	0,06	1,07	1,33
7	$\tilde{\chi}^2 = \sum_{j=1}^m \frac{(n_j - np_j)^2}{np_j}$,							
								$\tilde{\chi}^2 = 6,43$

$$p_j = \frac{1}{s\sqrt{2\pi}} \int_{\hat{t}_{j-1}}^{\hat{t}_j} \exp \left[-\frac{(t - \bar{t})^2}{2s^2} \right] dt$$


3. По таблице определяем критическое значение для уровня значимости $\alpha = 0.10$ и $l = m - k - 1 = 7 - 2 - 1 = 4$

$$\chi^2_{4,0.1} = 7.779$$

4. Поскольку $\hat{\chi}^2 < \chi^2_{4,0.1}$ гипотеза принимается!

2. Критерий Колмогорова

Критерий основан на теореме Гливенко:

$\hat{F}_n(x; \theta)$ эмпирическая функция распределения состоятельная оценка $F(x; \theta)$

Таким образом, если гипотеза верна, то должно проявляться сходство между эмпирической и теоретической функциями распределения и различия между ними должно убывать с увеличением n . Поэтому в качестве критерия можно использовать различные количественные нормы расстояния между ними.

Одна из них: Критерий Колмогорова

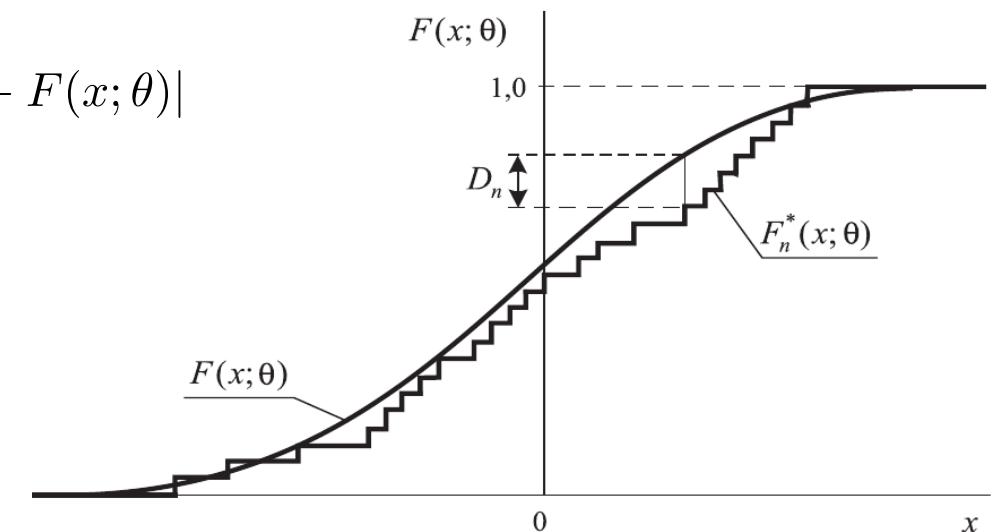
$$D_n(\hat{F}_n(x; \theta), F(x; \theta)) = \max_{|x| < \infty} |\hat{F}_n(x; \theta) - F(x; \theta)|$$

Замечательное свойство D_n состоит в том, что если гипотеза верна, то закон распределения критерия D_n оказывается одним и тем же для всех непрерывных G (истинный закон). Он зависит только от объема выборки n .

Для малых n (примерно $n < 100$) составлены таблицы процентных точек.

Для больших n значение критерия становится очень маленьким, поэтому было решено заменить его на:

$$\Lambda = D_n \sqrt{n}, \quad \Lambda \in K \quad \text{при } n \rightarrow \infty$$



K = распределение Колмогорова (табулировано)

Таблица 1

Уровень значимости α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	1,073	1,224	1,358	1,520	1,627	1,950

При больших n используют асимптотическое приближение

$$\lim_{n \rightarrow \infty} P(\Lambda < z) = F(z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2}$$

Критерий Колмогорова применим, когда выборка достаточно большого размера (примерно $n > 50$)

Алгоритм проверки гипотезы о виде закона распределения на основе критерия Колмогорова-Смирнова

6

Для сложной гипотезы, т. е. когда не известны значения параметров

- 1 * По выборочным данным x_1, x_2, \dots, x_n строим оценки $\hat{\theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_k\}$ параметров выбранного закона распределения $F(x; \theta_1, \dots, \theta_k)$ (используя методы моментов или максимального правдоподобия).

- 2 Строится вариационный ряд, по которому определяется эмпирическая функция распределения:

$$\hat{F}_n(x) = \begin{cases} 0, & x < x_1; \\ \frac{i}{n}, & x_i \leq x < x_{i+1}, i = 1..(n-1); \\ 1, & x \geq x_n \end{cases}$$

- 3 Вычисляется значение критерия:

Свойства значение критерия \hat{D}_n для сложной гипотезы во многом совпадают с D_n для простой гипотезы.

$$D_n = \max_{i \in [1..n]} |\hat{F}_n(x_i) - F(x_i; \theta)| \text{ или } \hat{D}_n = \max_{i \in [1..n]} |\hat{F}_n(x_i) - F(x_i; \hat{\theta})|$$

Однако \hat{D}_n распределена иначе и для нее необходима новая таблица + нет свойства независимости от закона распределения. Для практически важных распределений (нормальное, показательное) составлены таблицы.

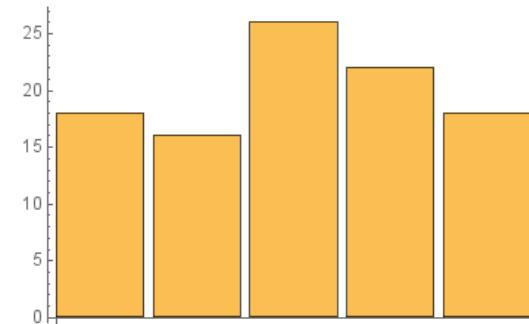
- 4 При больших n находим $\Lambda = \sqrt{n}D_n$ или $\hat{\Lambda} = \sqrt{n}\hat{D}_n$
Значение критерия сравнивается с критическим (см. табл.1) для соответствующего уровня значимости.

- 5 Делаются выводы.

H_0 принимается, если $\Lambda < \lambda_\alpha$ и отвергается в противном случае $\Lambda \geq \lambda_\alpha$

П р и м е р : Имеются данные $n=100$ измерений.

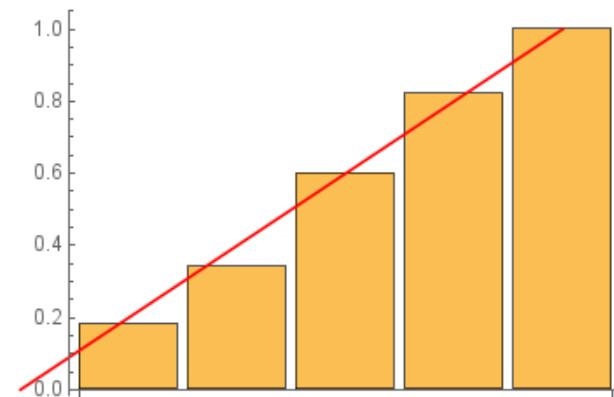
Количество предметов	1	2	3	4	5
Частота	18	16	26	22	18



На уровне значимости $\alpha=0.2$ с помощью критерия Колмогорова определите подчиняются ли данные выборки на интервале $[0,5]$ равномерному закону распределения случайной величины.

x_i	$F(x_i)=0,2x_i$	x_{ni}	$F_n(x_i)$	$ F(x_i) - F_n(x_i) $
1	0,2	18	0,18	0,02
2	0,4	16	0,34	0,06
3	0,6	26	0,6	0
4	0,8	22	0,82	0,02
5	1	18	1	0
$\max=0,06$				

$$F(x) = \begin{cases} 0, & x < 0 \\ x/5, & 0 \leq x \leq 5 \\ 1, & x > 5 \end{cases}$$



$$\Lambda = 0.06 \cdot \sqrt{100} = 0.6$$

При $\alpha = 0.2$, $\lambda_{kp} = \lambda_\alpha = 1.073$

$\Lambda < \lambda_{kp}$ гипотеза принимается !