

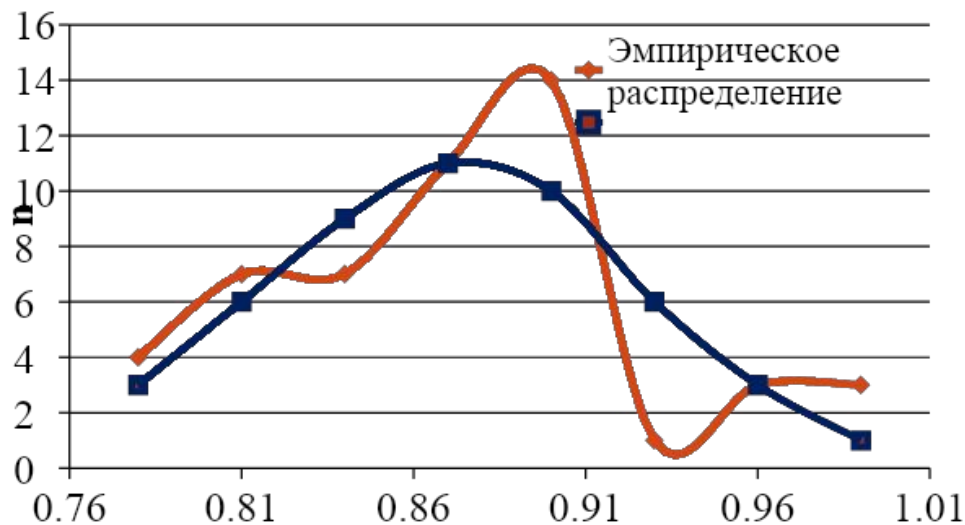
Технологии обработки информации

Лекция 2: Анализ природы данных. Проверка нормальности

Преподаватель: Тазиева Рамиля Фаридовна

Построение эмпирической и теоретической кривых

x_i	n_i	$x_i - \bar{x}$	u_i	$\varphi(u_i)$	y_i	n_i'
0,78	4	-0,0924	-1,70228	0,093684	2,588918	3
0,81	7	-0,0624	-1,14959	0,206032	5,693596	6
0,84	7	-0,0324	-0,5969	0,333842	9,225563	9
0,87	11	-0,0024	-0,04422	0,398553	11,01379	11
0,9	14	0,0276	0,508475	0,350564	9,687659	10
0,93	1	0,0576	1,061164	0,227189	6,278249	6
0,96	3	0,0876	1,613854	0,108479	2,997754	3
0,99	3	0,1176	2,166544	0,038163	1,054609	1



$$u_i = \frac{x_i - \bar{x}}{s}$$

$$y_i = \frac{n \cdot h}{s} \cdot \varphi(u_i)$$

$$\varphi(u_i) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u_i^2}{2}}$$

Гипотеза

Пусть X – наблюдаемая дискретная или непрерывная случайная величина (СВ).

Статистической гипотезой H называется предположение относительно параметров или вида распределения СВ X .

Проверяемая гипотеза называется нулевой гипотезой и обозначается H_0 .

Наряду с гипотезой H_0 рассматривают одну из альтернативных гипотез H_1 .

Проверка статистической гипотезы основывается на принципе, согласно которому маловероятные события считаются невозможными, а события, имеющие большую вероятность, – достоверными.

Принимая или отвергая гипотезу H_0 , можно допустить ошибки двух видов:

Ошибка первого рода состоит в том, что гипотеза H_0 отвергается в случае, когда она верна. Вероятность этой ошибки равна α (уровень значимости).

Ошибка второго рода состоит в том, что H_0 принимается, когда она не верна. ($1-\beta$ – мощность критерия).

Критерии согласия

Проверка гипотезы о виде распределения случайной величины X по результатам выборки:

$$H_0 : F_n(x) = F(x),$$

где $F_n(x)$ - эмпирическая, а $F(x)$ - теоретическая функция распределения вероятностей.

Все известные критерии согласия условно можно разбить на три основные группы :

- критерии, построенные на изучении разности между теоретической плотностью распределения вероятностей и эмпирической гистограммой;
- критерии, основанные на изучении разности между теоретической и эмпирической функциями распределения вероятностей;
- корреляционно-регрессионные критерии, основанные на изучении связей между эмпирическими и теоретическими порядковыми статистиками.

Критерий согласия Пирсона χ^2

Пусть $(x_1, x_2, \dots, x_n)$ - выборка наблюдений СВХ.

Проверяется гипотеза H_0 : X имеет функцию распределения $F_x(x)$.

1. По выборке наблюдений находят оценки параметров предполагаемого закона распределения СВХ.
2. Область возможных значений СВХ разбивается на k множеств $\Delta_1, \dots, \Delta_k$ (например, k интервалов, если X – непрерывная СВ, или k групп, состоящих из отдельных значений, если X – дискретная СВ).
3. Используя предполагаемый закон распределения СВХ, вычисляют вероятность попадания СВХ в интервал Δ_i ($i = \overline{1, k}$) - для НСВХ или вероятность того, что СВХ примет конкретное значение – для ДСВХ.
4. Полученные результаты представляют в таблице:

	Число наблюдений				Вс ег о
	Δ_1	Δ_2	...	Δ_k	
Наблюдаемое	n_1	n_2	...	n_k	n
Ожидаемое	np_1	np_2	...	np_k	n

Критерий согласия Пирсона χ^2

Выборочное значение статистики критерия χ^2 вычисляется по формуле:

$$\chi^2_{\text{в}} = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$$

5. Гипотеза H_0 согласуется с результатами наблюдений на уровне значимости α если $\chi^2_{\text{в}} < \chi^2_{1-\alpha}(k-l-1)$ $\chi^2_{1-\alpha}(k-l-1)$ квантиль порядка $1-\alpha$ распределения χ^2 с $k-l-1$ степенями свободы, l — число неизвестных параметров распределения, оцениваемых по выборке.

6. Если же $\chi^2_{\text{в}} \geq \chi^2_{1-\alpha}(k-l-1)$, то гипотеза H_0 отклоняется.

Критерий Пирсона можно применять для проверки гипотезы о том, что данная выборка взята из генеральной совокупности, распределенной по нормальному, биномиальному закону, по закону Пуассона, по экспоненциальному закону.

Правило применения критерия χ^2 .

Выдвигается нулевая гипотеза:

$$H_0 : F(t) = \frac{1}{S\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{(x-\bar{x})^2}{2S^2}} dx.$$

В предположении, что случайная величина распределена нормально с параметрами μ и σ , вычисляют вероятности p_i ее попадания в интервалы $(-\infty, x_1]$, $(x_1, x_2]$, $\dots$, (x_k, ∞) .

$$p_{i+1} = P(x_i \leq x \leq x_{i+1}) = F(x_{i+1}) - F(x_i),$$

Умножив каждую вероятность на объем выборки n , получаем теоретическую частоту попадания в каждый из интервалов.

Выборочное значение статистики критерия χ^2 вычисляется по формуле

$$\chi^2_{\text{в}} = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$$

Гипотеза H_0 согласуется с результатами наблюдений на уровне значимости α , если

$$\chi^2_{\text{в}} < \chi^2_{1-\alpha}(k-2-1)$$

Если $\chi^2_{\text{в}} \geq \chi^2_{1-\alpha}(k-3)$, то гипотезу отклоняем.

Проверка гипотезы по критерию Пирсона χ^2

левая граница интервала	правая граница интервала	середина x_i	n_i
-5	0.795	0.78	4
0.795	0.825	0.81	7
0.825	0.855	0.84	7
0.855	0.885	0.87	11
0.885	0.915	0.9	14
0.915	0.945	0.93	1
0.945	0.975	0.96	3
0.975	5	0.99	3

=НОРМРАСП(RB; $\bar{x}$; s; ИСТИНА)-НОРМРАСП(LB; $\bar{x}$; s; ИСТИНА)

n_i	Левая (LB)	Правая (RB)	p	np_k	$\frac{(n_k - np_k)^2}{np_k}$
11	-5	0.825	0.191	9.563	0.216
7	0.825	0.855	0.183	9.150	0.505
11	0.855	0.885	0.218	10.876	0.001
14	0.885	0.915	0.192	9.597	2.020
7	0.915	5	0.216	10.814	1.345
					4.088

По заданному уровню значимости $\alpha=0.05$ и $f=5-2-1=2$ степеням свободы с помощью статистической функции ХИ2ОБР(α ;f), $\chi_{0.95}^2(2) = 5.99$

Гипотезу принимаем, т.к. $\chi_{\epsilon}^2 < \chi_{0.95}^2$

Критерий Пирсона χ^2

$$\chi^2_{\epsilon} = \sum_{i=1}^k \frac{(n_i - n_i')^2}{n_i'} = 10,71.$$

Для уровня значимости $\alpha=0.05$ и числе степеней свободы $l=k-r=8-3=5$ по Приложению 4, вводим к которой является $p=1-\alpha=0.95$ и $K=l=5$, находим критическое значение $\chi^2_{\text{крит}} = 11,1$.

Гипотезу принимаем, т.к. $\chi^2_{\text{выб}} < \chi^2_{\text{крит}}$

Критерий Романовского

Если $\frac{|\chi^2_{\epsilon} - f|}{\sqrt{2f}} \geq 3,$

то расхождения можно считать существенными и гипотеза H_0 отвергается.

Если $\frac{|\chi^2_{\epsilon} - f|}{\sqrt{2f}} < 3,$

то расхождения можно считать несущественными, поэтому нет оснований отвергнуть гипотезу H_0 .

Критерий Ястремского

Для проверки соответствия данной выборочной совокупности признака X нормальному распределению по данному критерию, нужно проверить неравенство:

$$J \leq 3\sqrt{2k + 4\Theta},$$

где $J = |c - k|$, $c = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i(1-p_i)}$ n_i — эмпирические частоты, np_i — теоретические частоты, k — число интервалов дискретного вариационного, n — объем выборки. Если $k < 20$, то $\Theta = 0,6$.

Критерий Романовского

Воспользуемся результатами предыдущих расчетов.

а) для случая без объединения интервалов:

$$\frac{|\chi^2_{\epsilon} - f|}{\sqrt{2f}} = \frac{|10.71 - 5|}{\sqrt{10}} = 1.8.$$

б) для случая с объединением интервалов, в которые попало менее 5 значений:

$$\frac{|\chi^2_{\epsilon} - f|}{\sqrt{2f}} = \frac{|4.088 - 2|}{\sqrt{4}} = 1.044.$$

Гипотезу о соответствии данного распределения нормальному закону распределения **принимает**.

$$\frac{|\chi^2_{\epsilon} - f|}{\sqrt{2f}} < 3$$

Критерий Ястремского

Без объединения интервалов

x_i	n_i	n_i'	p_i'	q_i'	$(n_i - n_i')^2$	$n_i' q_i'$	$(n_i - n_i')^2 / (n_i' q_i')$
0,78	4	3	0,06	0,94	1	2,82	0,354609929
0,81	7	6	0,12	0,88	1	5,28	0,189393939
0,84	7	9	0,18	0,82	4	7,38	0,54200542
0,87	11	11	0,22	0,78	0	8,58	0
0,9	14	10	0,2	0,8	16	8	2
0,93	1	6	0,12	0,88	25	5,28	4,734848485
0,96	3	3	0,06	0,94	0	2,82	0
0,99	3	1	0,02	0,98	4	0,98	4,081632653

$$J \leq 3\sqrt{2k + 4\Theta},$$

$$c = \sum_{i=1}^k \frac{(n_i - n_i')^2}{n_i' p_i'} = 11.902 \quad J = |c - k| = |11.902 - 8| = 3.902$$

Условие принятия гипотезы о нормальном распределении (т.е. $3.902 \leq 12.8686$) выполняется.

Критерий Ястремского

Для случая с объединением интервалов, в которые попало менее 5 значений

Эмпирические частоты n_i	p_i	Теоретические частоты np_i	$\frac{(n_i - np_i)^2}{np_i(1 - p_i)}$
11	0.19126030	9.563014902	0.26699365
7	0.18300999	9.150499527	0.61861031
11	0.21751227	10.87561356	0.001818087
14	0.19194079	9.597039292	2.499822427
7	0.21627665	10.81383272	1.716251184
Объем $n=50$		$\sum_{i=1}^n \frac{(n_i - np_i)^2}{np_i(1 - p_i)}$	5.103495659
Число интервалов k	5	$\frac{\left \sum_{i=1}^n \frac{(n_i - np_i)^2}{np_i(1 - p_i)} - k \right }{\sqrt{2k + 4\theta}}$	0.029390792

Поскольку $\frac{J}{\sqrt{2k + 4\theta}} = 0.02939 < 3$, гипотеза о нормальном распределении экспериментальных данных принимается.

Приближенные критерии нормальности распределения

Используются выборочные статистики: асимметрию, эксцесс и их средние квадратические отклонения:

$$S_{A_s} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}$$

$$S_{E_x} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}$$

Если $|A_s| < S_{A_s}$ и $|E_x| < S_{E_x}$, то считают, что выборочная совокупность подчиняется нормальному закону распределения.

Проверку выборочной совокупности на близость ее к нормальному распределению можно производить, используя совокупность статистик χ^2 , A_s и E_x .

$$\chi^2 = \frac{A_s^2}{S_{A_s}^2} + \frac{E_x^2}{S_{E_x}^2}$$

Если вычисленное значение χ^2 меньше $\chi_{кр}^2$, то гипотезу о нормальном распределении выборочной совокупности принимают.

Приближенные критерии

Средние квадратические отклонения асимметрии и эксцесса вычисляют по формулам:

$$S_{As} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}} = \sqrt{\frac{6 \cdot 49}{51 \cdot 53}} = 0.32979 \quad \chi_{выб}^2 < \chi_{крит}^2$$

$$S_{Ex} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}} = \sqrt{\frac{24 \cdot 50 \cdot 48 \cdot 47}{49^2 \cdot 53 \cdot 55}} = 0.62193$$

Условием принятия гипотезы о нормальном распределении являются условия $0.279 < 0.32979$ ($|As| \leq S_{As}$) и $0.293 < 0.62193$ ($|Ex| \leq S_{Ex}$). Оба условия выполняются.

$$\chi^2 = \frac{0.279^2}{0.32979^2} + \frac{0.293^2}{0.62193^2} = 0.9377$$

$$\chi_{выб}^2(0,05;2) = 5,99$$

Гипотезу о нормальном распределении выборочной совокупности принимаем, т. к.

$$\chi_{выб}^2 < \chi_{крит}^2$$

Критерий Колмогорова-Смирнова

Пусть $F_n(x)$ - эмпирическая функция распределения случайной величины X , представленной выборкой $(x_1, x_2, \dots, x_n)$.

Гипотеза $H_0 : F_n(x) = F(x)$

Статистика критерия Колмогорова-Смирнова для несгруппированных данных:

$$D_n^H = D_n(\sqrt{n} - 0,01 + \frac{0,85}{\sqrt{n}})$$

$$D_n^+ = \max_{1 \leq i \leq n} |F_n(x_{i+1}) - F(x_i)|, \quad D_n^- = \max_{1 \leq i \leq n} |F_n(x_i) - F(x_i)|, \quad D_n = \max(D_n^+, D_n^-).$$

Если $D_n^H > D_n(\alpha)$ то гипотеза согласия H_0 отклоняется на уровне значимости α .

α	0,15	0,10	0,05	0,03	0,01
$D_n^H(\alpha)$	0,775	0,819	0,895	0,955	1,035

Статистика критерия Колмогорова-Смирнова для сгруппированных данных:

$$\lambda = \frac{D}{\sqrt{n}} \quad \text{где } D = \max |M - M^*| \text{ — максимум абсолютного значения разности между}$$

накопленными эмпирическими частотами M и накопленными теоретическими частотами M^* , n — объем выборки.

Если $K(\lambda) < 0.05$, то имеет место существенное расхождение между эмпирическим и теоретическим распределениями, которое нельзя считать случайным.

$K(\lambda)$ - вероятность того, что СВ X примет значение большее или равное λ .

Значение функции $K(\lambda)$

λ	$K(\lambda)$	λ	$K(\lambda)$
0,30	1,0000	1,10	0,1777
0,35	0,9997	1,20	0,1122
0,40	0,9972	1,30	0,681
0,45	0,9874	1,40	0,397
0,50	0,9639	1,50	0,222
0,55	0,9228	1,60	0,120
0,60	0,8643	1,70	0,052
0,70	0,7112	1,90	0,015
0,75	0,6272	2,00	0,007
0,80	0,5441	2,10	0,0003
0,85	0,4653	2,20	0,0001
0,90	0,3927	2,30	0,0001
0,95	0,3275	2,40	0,0000
1,00	0,2700	2,50	0,0000

Критерий Колмогорова-Смирнова

x_i	Эмпирически частоты n_i	Накопленные частоты ω_i	$F(x_{cp}, s)$	$F_n(x_i)$	$F_n(x_{i+1})$	D_n^+	D_n^-
0.78	2	2	0.038911	0	0.04	0.038911	0.001089
0.79	2	4	0.057483	0.04	0.08	0.017483	0.022517
0.80	2	6	0.082408	0.08	0.12	0.002408	0.037592
0.81	2	8	0.114710	0.12	0.16	0.005290	0.045290
0.82	3	11	0.155137	0.16	0.22	0.004863	0.064863
0.84	2	13	0.261008	0.22	0.26	0.041008	0.001008
0.85	5	18	0.325262	0.26	0.36	0.065262	0.034738
0.86	4	22	0.395185	0.36	0.44	0.035185	0.044815
0.87	3	25	0.468665	0.44	0.50	0.028665	0.031335
0.88	4	29	0.543232	0.50	0.58	0.043232	0.036768
0.89	3	32	0.616302	0.58	0.64	0.036302	0.023698
0.90	6	38	0.685448	0.64	0.76	0.045448	0.074552
0.91	5	43	0.748632	0.76	0.86	0.011368	0.111368
0.94	1	44	0.890989	0.86	0.88	0.030989	0.010989
0.95	1	45	0.922052	0.88	0.90	0.042052	0.022052
0.96	1	46	0.945887	0.90	0.92	0.045887	0.025887
0.97	1	47	0.963547	0.92	0.94	0.043547	0.023547
0.98	2	49	0.976183	0.94	0.98	0.036183	0.003817
0.99	1	50	0.984914	0.98	1	0.004914	0.015086
n	50				макс	0.065262	0.111368

Критерий Колмогорова-Смирнова

Для несгруппированных данных:

$$F(x_i) = P(X < x_i) = \text{НОРМ.РАСП}(x_i; \bar{x}; s; \text{ИСТИНА})$$

В качестве статистики критерия используется величина:

$$D_n^H = D_n \left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right) = 0.111368 \cdot \left(\sqrt{50} - 0.01 + \frac{0.85}{\sqrt{50}} \right) \\ = 0.799761.$$

Так, для $\alpha = 0.05$ $D_n^H(0.05) = 0.895$ Так как $D_n^H = 0.799761 < D_n^H(0.05) = 0.895$, гипотеза о нормальном распределении не отклоняется.

Критерий Колмогорова для сгруппированной выборки

x_i	n_i	n_i'	w_i	w_i'	$ w_i - w_i' $
0,78	4	3	4	3	1
0,81	7	6	11	9	2
0,84	7	9	18	18	0
0,87	11	11	29	29	0
0,9	14	10	43	39	4
0,93	1	6	44	45	1
0,96	3	3	47	48	1
0,99	3	1	50	49	1

$$\lambda = \frac{\max(|w_i - w_i'|)}{\sqrt{n}} = \frac{4}{\sqrt{50}} = 0,566,$$

$$K(\lambda) = 0,9228$$

Гипотеза о нормальном распределении принимается, если $K(\lambda) > 0.05$. Это условие выполняется. Гипотезу о соответствии данного распределения нормальному закону распределения **принимает**.