

# Теория вероятностей и математическая статистика

## Лекция 18

Проверка непараметрических  
гипотез, корреляционный и  
регрессионный анализ

## Проверка гипотезы о виде статистической модели

**Статистической моделью** называется выборочное пространство, на котором задано некоторое семейство распределений.

Пусть дана случайная выборка  $\vec{X}_n$ .

**Статистической гипотезой** называется всякое предположение о распределении вероятностей случайной выборки.

**Критерием согласия** называется статистический критерий для проверки гипотезы о виде функции распределения следующего вида

$$H_0 : F(t) = F_0(t), \quad \forall t \in \mathbb{R}, \quad H_1 : \exists t \in \mathbb{R}, \quad F(t) \neq F_0(t).$$

Критерий согласия проверяет согласие с нулевой гипотезой.

Для непрерывной статистической модели существуют критерии согласия Колмогорова и  $\omega^2$ , для дискретной статистической модели – критерий согласия  $\chi^2$  (Пирсона).

## Критерий согласия Колмогорова

$$W : D(\vec{x}_n) > D_{1-\alpha}(n),$$

где  $D_{1-\alpha}(n)$  – квантили случайной величины  $D(\vec{X}_n)$  при условии истинности  $H_0$  и значение величины  $D(\vec{X}_n)$  на выборке  $\vec{x}_n$  определяется как

$$D(\vec{x}_n) = \sup_t |F_n(t) - F_0(t)| = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F_0(x_{(i)}), F_0(x_{(i)}) - \frac{i-1}{n} \right\},$$

где  $F_n(t)$  – эмпирическая функция распределения для  $\vec{x}_n$ .

**Теорема:** функция распределения случайной величины  $D(\vec{X}_n)$  при истинности  $H_0$  не зависит от  $F_0(t)$  и, в частности, совпадает с функцией распределения случайной величины  $\sup_{0 \leq t \leq 1} |\hat{F}(t, \vec{Y}_n) - t|$ , где  $Y$  распределено равномерно на  $[0; 1]$ .

Поэтому таблицу квантилей  $D_{1-\alpha}(n)$  можно строить по равномерному распределению, а использовать для любых  $\vec{X}_n$ .

## Пример

Для выборки  $\vec{x}_{10}$  с вариационным рядом

−1,20   −1,15   −0,91   −0,29   −0,12   0,16   1,06   1,09   1,22   1,29

проверим гипотезу  $H_0$  о том, что  $X$  имеет стандартное нормальное распределение на уровне значимости  $\alpha = 0,1$ .

Используем критерий согласия Колмогорова. В точках вариационного ряда найдём значения функции стандартного нормального распределения  $F_0(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-s^2/2} ds$ :

0,115   0,125   0,181   0,386   0,452   0,564   0,855   0,862   0,899   0,901

Находим значения выражения  $\frac{i}{n} - F_0(x_{(i)})$ :

−0,015   0,075   0,119   0,014   0,048   0,036   −0,155   −0,062   0,001   0,099

Находим значения выражения  $F_0(x_{(i)}) - \frac{i-1}{n}$ :

0,115   0,025   −0,019   0,086   0,052   0,064   0,255   0,162   0,099   0,001

Находим максимум по двум последним спискам  $D(\vec{x}_{10}) = 0,255$

и по таблице квантилей  $D_{1-\alpha}(n) = D_{0,9}(10) = 0,369$ .

Так как  $D(\vec{x}_{10}) < D_{0,9}$ , то оснований отклонить гипотезу  $H_0$  нет.

Таблица 6.2. Критические значения для наибольшего отклонения эмпирического распределения от теоретического (критерий Колмогорова)

n \ Q	Q					n \ Q	Q				
	20%	10%	5%	2%	1%		20%	10%	5%	2%	1%
1	0,90000	0,95000	0,97500	0,99000	0,99500	51	0,14697	0,16796	0,18659	0,20864	0,22386
2	68377	77639	84189	90000	92929	52	14558	16637	18482	20667	22174
3	56481	63604	70760	78456	82900	53	14423	16483	18311	20475	21968
4	49265	56522	62394	68887	73424	54	14292	16332	18144	20289	21768
5	44698	50945	56328	62718	66853	55	14164	16186	17981	20107	21574
6	0,41037	0,46799	0,51926	0,57741	0,61661	56	0,14040	0,16044	0,17823	0,19930	0,21384
7	38148	43607	48342	53844	57581	57	13919	15906	17669	19758	21199
8	35831	40962	45427	50654	54179	58	13801	15771	17519	19590	21019
9	33910	38746	43001	47960	51332	59	13686	15639	17373	19427	20844
10	32260	36866	40925	45662	48893	60	13573	15511	17231	19267	20673
11	0,30829	0,35242	0,39122	0,43670	0,46770	61	0,13464	0,15385	0,17091	0,19112	0,20506
12	29577	33815	37543	41918	44905	62	13357	15263	16956	18960	20343
13	28470	32549	36143	40362	43247	63	13253	15144	16823	18812	20184
14	27481	31417	34890	38970	41762	64	13151	15027	16693	18667	20029
15	26588	30397	33760	37713	40420	65	13052	14913	16567	18525	19877
16	0,25778	0,29472	0,32733	0,36571	0,39201	66	0,12954	0,14802	0,16443	0,18387	0,19729
17	25039	28627	31796	35528	38086	67	12859	14693	16322	18252	19584
18	24360	27851	30936	34569	37062	68	12766	14587	16204	18119	19442
19	23735	27136	30143	33685	36117	69	12675	14483	16088	17990	19303
20	23156	26473	29408	32866	35241	70	12586	14381	15975	17863	19167
21	0,22617	0,25858	0,28724	0,32104	0,34427	71	0,12499	0,14281	0,15864	0,17739	0,19034
22	22115	25283	28087	31394	33666	72	12413	14183	15755	17618	18903
23	21645	24746	27490	30728	32954	73	12329	14087	15649	17498	18776
24	21205	24242	26931	30104	32286	74	12247	13993	15544	17382	18650
25	20790	23768	26404	29516	31657	75	12167	13901	15442	17268	18528

## Критерий согласия $\omega^2$ (омега-квадрат)

$$W : \omega^2(\vec{X}_n) > \omega_{1-\alpha}^2(n),$$

где  $\omega_{1-\alpha}^2(n)$  – квантили статистики

$$\omega^2(\vec{X}_n) = \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^n \left( F_0(X_{(i)}) - \frac{2i-1}{2n} \right)^2$$

при условии истинности  $H_0$ .

**Теорема:** распределение статистики  $\omega^2(\vec{X}_n)$  при истинности  $H_0$  не зависит от  $F_0$ .

## Пример

Для выборки  $\vec{x}_{10}$  с вариационным рядом

−1,20   −1,15   −0,91   −0,29   −0,12   0,16   1,06   1,09   1,22   1,29

проверим гипотезу  $H_0$  о том, что  $X$  имеет стандартное нормальное распределение на уровне значимости  $\alpha = 0,05$ .

Используем критерий согласия  $\omega^2$ . В точках вариационного ряда найдём значения функции стандартного нормального распределения  $F_0(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-s^2/2} ds$ :

0,115   0,125   0,181   0,386   0,452   0,564   0,855   0,862   0,899   0,901

$$\begin{aligned}\omega^2(\vec{x}_{10}) &= \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^n \left( F_0(x_{(i)}) - \frac{2i-1}{2n} \right)^2 = \frac{1}{12 \cdot 10^2} + \frac{1}{10} (0,065^2 + (-0,025)^2 + \\ &+ (-0,069)^2 + 0,036^2 + 0,002^2 + 0,014^2 + 0,205^2 + 0,112^2 + 0,049^2 + (-0,049)^2) \approx \\ &\approx 0,007641 < \omega_{0,95}^2(10) \approx 0,0461.\end{aligned}$$

Гипотеза  $H_0$  не отклоняется.

Таблица 6.4а. Критерий  $\omega^2$ . Функция распределения  $a_1(x)$

$x$	0	1	2	3	4	5	6	7	8	9
0,0	0,00000	00001	00300	02568	06685	12372	18602	24844	30815	36386
1	41513	46196	50457	54329	57846	61042	63951	66600	69019	71229
2	73253	75109	76814	78383	79829	81163	82396	83536	84593	85573
3	86483	87329	88115	88848	89531	90167	90762	91317	91836	92321
4	92775	93201	93599	93972	94323	94651	94960	95249	95521	95777
0,5	0,96017	96242	96455	96655	96843	97020	97186	97343	97491	97630
6	97762	97886	98002	98112	98216	98314	98406	98493	98575	98653
7	98726	98795	98861	98922	98981	99036	99083	99137	99183	99227
8	99268	99308	99345	99380	99413	99444	99474	99502	99528	99553
9	99577	99599	99621	99641	99660	99678	99695	99711	99726	99740
1,0	0,99754	99764	99776	99787	99799	99812	99820	99828	99837	99847
1	99856	99862	99869	99876	99883	99890	99895	99900	99905	99910
2	99916	99919	99923	99927	99931	99935	99938	99941	99944	99947
3	99950	99953	99955	99957	99959	99962	99964	99965	99967	99969
4	99971	99972	99973	99975	99976	99978	99978	99979	99980	99980

$$\omega_{\gamma}^2(n) = \frac{x}{n}, \quad a_1(x) = \gamma$$

## Критерий согласия $\chi^2$ (Пирсона)

Пусть дискретная СВ  $X$  принимает  $r$  различных значений с частотами  $n_k(\vec{x}_n)$ .

$$H_0 : p_k = p_{k0}, \quad \forall k = \overline{1, r}, \quad H_1 : \exists k = \overline{1, r}, \quad p_k \neq p_{k0}.$$

Критерий согласия  $\chi^2$  (Пирсона)

$$W : \chi^2(\vec{x}_n) > \chi_{1-\alpha}^2(r-1),$$

где  $\chi_{1-\alpha}^2(r-1)$  – квантили  $\chi^2$ -распределения с  $r-1$  степенями свободы:

$$\chi^2(\vec{X}_n) = \sum_{k=1}^r \frac{(n_k(\vec{X}_n) - np_{k0})^2}{np_{k0}} = n \sum_{k=1}^r \frac{\left( \frac{n_k(\vec{X}_n)}{n} - p_{k0} \right)^2}{p_{k0}}.$$

**Теорема (Пирсона):** распределение случайной величины  $\chi^2(\vec{X}_n)$  при  $n \rightarrow +\infty$  слабо сходится к  $\chi^2$ -распределению с  $r-1$  степенями свободы.

Критерий  $\chi^2$  используется только, если  $np_k \geq 10$  или если  $r \geq 20$  и  $np_k \geq 5$ .

Критерий  $\chi^2$  можно использовать и для непрерывных случайных величин, и для дискретных со счётным множеством различных значений. Для этого множество всех значений разбивают на конечное множество промежутков.

## Пример

При 4040 бросаниях монеты Ж. Бюффон (1707–1788) получил 2048 гербов и 1992 решки.

Проверим на уровне значимости 0,05 совместимость этих данных с гипотезой о том, что вероятность герба равна 0,5.

$$\begin{aligned}\chi^2(\vec{x}_n) &= \sum_{k=1}^r \frac{(n_k(\vec{X}_n) - np_{k0})^2}{np_{k0}} = \\ &= \frac{(2048 - 4040 \cdot 0,5)^2}{4040 \cdot 0,5} + \frac{(1992 - 4040 \cdot 0,5)^2}{4040 \cdot 0,5} = 0,776 < \\ &< \chi_{0,95}^2(1) = 3,841.\end{aligned}$$

Статистические данные не противоречат гипотезе  $H_0$ .

Квантили распределения  $\chi^2$

	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999
1	3,9E-05	0,0002	0,001	0,004	0,02	0,1	0,1	1,1	1,6	2,7	3,8	5,0	6,6	7,9	10,8
2	0,010	0,02	0,051	0,10	0,21	0,45	0,71	2,41	3,22	4,61	5,99	7,38	9,21	10,60	13,82
3	0,072	0,11	0,22	0,35	0,58	1,01	1,42	3,66	4,64	6,25	7,81	9,35	11,34	12,84	16,27
4	0,21	0,30	0,48	0,71	1,06	1,65	2,19	4,88	5,99	7,78	9,49	11,14	13,28	14,86	18,47
5	0,41	0,55	0,83	1,15	1,61	2,34	3,00	6,06	7,29	9,24	11,07	12,83	15,09	16,75	20,52
6	0,68	0,87	1,24	1,64	2,20	3,07	3,83	7,23	8,56	10,64	12,59	14,45	16,81	18,55	22,46
7	0,99	1,24	1,69	2,17	2,83	3,82	4,67	8,38	9,80	12,02	14,07	16,01	18,48	20,28	24,32
8	1,34	1,65	2,18	2,73	3,49	4,59	5,53	9,52	11,03	13,36	15,51	17,53	20,09	21,95	26,12
9	1,73	2,09	2,70	3,33	4,17	5,38	6,39	10,66	12,24	14,68	16,92	19,02	21,67	23,59	27,88
10	2,16	2,56	3,25	3,94	4,87	6,18	7,27	11,78	13,44	15,99	18,31	20,48	23,21	25,19	29,59
11	2,60	3,05	3,82	4,57	5,58	6,99	8,15	12,90	14,63	17,28	19,68	21,92	24,72	26,76	31,26
12	3,07	3,57	4,40	5,23	6,30	7,81	9,03	14,01	15,81	18,55	21,03	23,34	26,22	28,30	32,91
13	3,57	4,11	5,01	5,89	7,04	8,63	9,93	15,12	16,98	19,81	22,36	24,74	27,69	29,82	34,53
14	4,07	4,66	5,63	6,57	7,79	9,47	10,82	16,22	18,15	21,06	23,68	26,12	29,14	31,32	36,12
15	4,60	5,23	6,26	7,26	8,55	10,31	11,72	17,32	19,31	22,31	25,00	27,49	30,58	32,80	37,70
16	5,14	5,81	6,91	7,96	9,31	11,15	12,62	18,42	20,47	23,54	26,30	28,85	32,00	34,27	39,25
17	5,70	6,41	7,56	8,67	10,09	12,00	13,53	19,51	21,61	24,77	27,59	30,19	33,41	35,72	40,79
18	6,26	7,01	8,23	9,39	10,86	12,86	14,44	20,60	22,76	25,99	28,87	31,53	34,81	37,16	42,31
19	6,84	7,63	8,91	10,12	11,65	13,72	15,35	21,69	23,90	27,20	30,14	32,85	36,19	38,58	43,82
20	7,43	8,26	9,59	10,85	12,44	14,58	16,27	22,77	25,04	28,41	31,41	34,17	37,57	40,00	45,31
21	8,03	8,90	10,28	11,59	13,24	15,44	17,18	23,86	26,17	29,62	32,67	35,48	38,93	41,40	46,80
22	8,64	9,54	10,98	12,34	14,04	16,31	18,10	24,94	27,30	30,81	33,92	36,78	40,29	42,80	48,27
23	9,26	10,20	11,69	13,09	14,85	17,19	19,02	26,02	28,43	32,01	35,17	38,08	41,64	44,18	49,73
24	9,89	10,86	12,40	13,85	15,66	18,06	19,94	27,10	29,55	33,20	36,42	39,36	42,98	45,56	51,18
25	10,52	11,52	13,12	14,61	16,47	18,94	20,87	28,17	30,68	34,38	37,65	40,65	44,31	46,93	52,62
26	11,16	12,20	13,84	15,38	17,29	19,82	21,79	29,25	31,79	35,56	38,89	41,92	45,64	48,29	54,05
27	11,81	12,88	14,57	16,15	18,11	20,70	22,72	30,32	32,91	36,74	40,11	43,19	46,96	49,64	55,48
28	12,46	13,56	15,31	16,93	18,94	21,59	23,65	31,39	34,03	37,92	41,34	44,46	48,28	50,99	56,89
29	13,12	14,26	16,05	17,71	19,77	22,48	24,58	32,46	35,14	39,09	42,56	45,72	49,59	52,34	58,30
30	13,79	14,95	16,79	18,49	20,60	23,36	25,51	33,53	36,25	40,26	43,77	46,98	50,89	53,67	59,70
35	17,19	18,51	20,57	22,47	24,80	27,84	30,18	38,86	41,78	46,06	49,80	53,20	57,34	60,27	66,62
40	20,71	22,16	24,43	26,51	29,05	32,34	34,87	44,16	47,27	51,81	55,76	59,34	63,69	66,77	73,40
45	24,31	25,90	28,37	30,61	33,35	36,88	39,58	49,45	52,73	57,51	61,66	65,41	69,96	73,17	80,08
50	27,99	29,71	32,36	34,76	37,69	41,45	44,31	54,72	58,16	63,17	67,50	71,42	76,15	79,49	86,66
75	47,21	49,48	52,94	56,05	59,79	64,55	68,13	80,91	85,07	91,06	96,22	100,84	106,39	110,29	118,60
100	67,33	70,06	74,22	77,93	82,36	87,95	92,13	106,91	111,67	118,50	124,34	129,56	135,81	140,17	149,45

## Корреляционный анализ

**Корреляционным анализом** называется раздел математической статистики, исследующий зависимость между случайными величинами с помощью выборочных коэффициентов корреляции.

Выборочный коэффициент линейной корреляции

$$\hat{\rho}(\vec{X}_n, \vec{Y}_n) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

При проверке гипотезы  $H_0 : \rho = 0$  при  $H_1 : \rho \neq 0$  используют статистику

$$\frac{\hat{\rho}(\vec{X}_n, \vec{Y}_n) \sqrt{n-2}}{\sqrt{1 - \hat{\rho}^2(\vec{X}_n, \vec{Y}_n)}} \sim T(n-2),$$

и тогда

$$W : \frac{|\hat{\rho}| \sqrt{n-2}}{\sqrt{1 - \hat{\rho}^2}} \geq t_{1-\alpha/2}(n-2).$$

## Пример

**Пример 6.5.** Вычислим значение  $\hat{\rho}$  для пары случайных величин  $(\xi, \eta)$ , где  $\xi$  — рост (в см), а  $\eta$  — масса тела (в кг) наугад выбранного студента-первокурсника. Выборка объема  $n = 15$  представлена в табл. 6.2.

Чтобы оценить показатель  $\rho$  связи двух случайных величин, сначала найдем выборочные средние этих величин:

$$\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = \frac{2620}{15} = 173,3; \quad \bar{y} = \frac{1}{15} \sum_{i=1}^{15} y_i = \frac{945}{15} = 63,1.$$

Затем определяем суммы

$$\sum_{i=1}^{15} (x_i - \bar{x})^2 = 747,33; \quad \sum_{i=1}^{15} (y_i - \bar{y})^2 = 1171,4;$$

$$\sum_{i=1}^{15} (x_i - \bar{x})(y_i - \bar{y}) = 293,3.$$

Таким образом,  $\hat{\rho} = \frac{293,3}{\sqrt{747,33 \cdot 1171,4}} = 0,313.$

Таблица 6.2

Номер наблюдения	Рост, см		Масса тела, кг	
	$x_i$	$x_i - \bar{x}$	$y_i$	$y_i - \bar{y}$
1	165	-8,3	72,9	9,8
2	171	-2,3	48,4	-14,7
3	182	8,7	66,3	3,2
4	165	-8,3	64,1	1,0
5	183	9,7	62,7	-0,4
6	180	6,7	76,0	12,9
7	183	9,7	73,8	10,7
8	166	-7,3	50,6	-12,5
9	173	-0,3	52,3	-10,8
10	172	-1,3	56,5	-6,6
11	174	0,7	66,8	3,7
12	170	-3,3	61,6	-1,5
13	164	-9,3	72,8	9,7
14	168	-5,3	52,6	-10,5
15	184	10,7	68,6	5,5
$\Sigma$	2600		945	

# Регрессионный анализ

**Регрессионным анализом** называется раздел математической статистики, исследующий зависимость между случайными величинами с помощью уравнений регрессии.

**Регрессией** называется функциональная связь в среднем любых случайных величин.

**Теоретическим уравнением регрессии** называется уравнение вида

$$y = M_x Y$$

где  $M_x Y$  – условное математическое ожидание СВ  $Y$  при заданном  $x$ .

**Эмпирическое уравнение регрессии** можно построить в виде многочлена

$$y = B_0 + B_1 x + \dots + B_k x^k$$

**методом наименьших квадратов:** по  $n$  экспериментальным точкам  $(x_1, y_1)$ , ...,  $(x_n, y_n)$  найти многочлен  $y(x)$ , для которого сумма квадратов отклонений в точках  $x_i$  от  $y_i$  минимальная, т. е.

$$Q = \sum_{i=1}^n (y_i - y(x_i))^2 = \sum_{i=1}^n \varepsilon^2 \rightarrow \min.$$

# Линейная регрессионная модель

$$y = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p$$

При некоторых значениях факторов  $x_{i1}, \dots, x_{ip}$ , где  $i \in \overline{1, n}$ , произведены измерения отклика  $y_i$  со случайной ошибкой  $\varepsilon_i$ :

$$Y = X\theta + \varepsilon,$$

где  $X$  – матрица размера  $n \times p$  и  $Q = \|Y - X\theta\|^2 = (Y - X\theta)^\top (Y - X\theta)$ .

**Теорема:** если ранг  $X = p$ , то оценка параметра  $\theta = (\theta_1, \dots, \theta_p)$  по методу наименьших квадратов имеет вид  $\hat{\theta} = (X^\top X)^{-1} X^\top Y$ .

Формулу в этой теореме нельзя упростить, так как в общем случае матрица  $X$  не является квадратной.

**Теорема:** если случайные величины  $\varepsilon_i$  независимы и одинаково распределены с  $M\varepsilon_i = 0$  и конечной дисперсией  $D\varepsilon_i = \sigma^2$ , то оценка  $\hat{\theta} = (X^\top X)^{-1} X^\top Y$  является несмещённой и состоятельной.

Для оценивания параметров нелинейной зависимости

$$y = \theta_1 \varphi_1(t) + \theta_2 \varphi_2(t) + \dots + \theta_p \varphi_p(t)$$

обозначим  $x_{ij} = \varphi_j(t_i)$  и получим линейную модель.

## Пример

В «Основах химии» Д.И. Менделеев приводит следующие данные о количестве азотнатриевой соли  $NaNO_3$ , которое можно растворить в 100 г воды в зависимости от температуры  $t$ .

$t_i$	0	4	10	15	21	29	36	51	68
$y_i$	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Построим по этим данным приближённую эмпирическую формулу

$$y = \theta_1 + \theta_2 t + \theta_3 t^2.$$

$$X^T = \begin{pmatrix} 1 \\ t_i \\ t_i^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 4 & 10 & 15 & 21 & 29 & 36 & 51 & 68 \\ 0 & 16 & 100 & 225 & 441 & 841 & 1296 & 2601 & 4624 \end{pmatrix},$$

$$X^T X = \begin{pmatrix} 9 & 234 & 10\ 144 \\ 234 & 10\ 144 & 531\ 828 \\ 10\ 144 & 531\ 828 & 30\ 788\ 836 \end{pmatrix}, \quad \hat{\theta} = (X^T X)^{-1} X^T Y,$$

$$y \approx 66,71 + 0,9604 t - 0,001359 t^2.$$