КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ МАТЕМАТИКИ И МЕХАНИКИ им. Н.И. ЛОБАЧЕВСКОГО
Кафедра математической статистики

Е.К. Каштанова

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Казань – 2025

УДК 519.2(07)
ББК 22.172я7

Принято на заседании методической комиссии ИММ
Протокол № 7 от 10 июля 2025 года.

Рецензент:
Кандидат физико-математических наук,
доцент кафедры математической статистики КФУ **С.В. Симушкин**

Каштанова Е.К.
Практикум по математической статистике: учебное пособие/ Е.К. Каштанова – Казань: Изд-во Казан. ун-та, 2025. – 100 с.

Учебное пособие «Практикум по математической статистике» предназначено для обучения основам математической статистики студентов направления подготовки: 38.03.01 «Экономика» очной, очно-заочной форм обучения.

При разработке курса большое внимание было уделено практической направленности курса. Учебное пособие «Практикум по математической статистике» предназначено для развития общепрофессиональной компетенции «Способен осуществлять сбор, обработку и статистический анализ данных, необходимых для решения поставленных экономических задач». Учебное пособие содержит теоретический материал, разбор типовых задач, задачи, таблицы, терминологический словарь, рекомендуемую литературу. Учебное пособие может быть рекомендовано к использованию преподавателями данной дисциплины и студентами в течение семестра и при подготовке к зачету и экзамену.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Целью любого статистического исследования является принятие решения относительно обследуемого объекта. Такое решение принимается с помощью специальных статистических правил на основе наблюдения некоторых характеристик объекта (обычно числовых или векторных). Результаты наблюдений трактуются как реализации случайных величин, и поэтому возможны ошибки в принятии решений. Математическая статистика занимается построением статистических правил, минимизирующих частоту принятия неправильных решений, а также вычислением этой частоты для традиционных (известных) статистических правил.

Все выводы математической статистики основываются на статистических данных. Приведем наиболее простой и распространенный способ получения данных.

Пусть X - наблюдаемая характеристика (случайная величина – с.в.) исследуемого объекта. Проведем серию опытов, состоящую из n раз независимых наблюдений X. В результате мы получаем n наблюдений, которые обозначим $x_1, x_2, .., x_n$. Обычно в математической статистике они называются выборкой из распределения случайной величины X (устаревшее – “выборка из генеральной совокупности значений X”) и представляют собой, строго говоря, результат наблюдения n независимых копий $X_1,.., X_n$ с.в. X.

Следующие задачи являются типичными для математической статистики.

1. Оценка неизвестных параметров в распределении наблюдаемых характеристик (случайных величин),

2. Проверка гипотез (высказываний, предположений) о распределении наблюдаемых случайных величин.

1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

1.1 Вариационный ряд

Ключевые слова: вариационный ряд, дискретный вариационный ряд, интервальный вариационный ряд, частота, частость, полигон, гистограмма.

Случайные величины выборки x_1, x_2 … x_n, расположенные в порядке возрастания их значений: $x_{(1)} \le x_{(2)} \ldots \le x_{(n)}$,, образуют *вариационный ряд*. В частности, $x_{(1)} = \min\{x_1,..,x_n\}$, $x_{(n)} = \max\{x_1,..,x_n\}$. Элементы вариационного ряда $x_{(i)}$ называются *вариантами*, i=1..n.

Если в выборке x_1, x_2,.., x_n значения x_i встречается n_i раз, то число n_i называется *частотой* элемента выборки x_i. Общая сумма частот будет равна объему выборки, т.е. $\sum n_i = n$.

Относительной частотой (частостью) называется отношение частоты к объему выборки

$$w_i = \frac{n_i}{n}.$$

Относительная частота показывает, какую долю занимает данный признак (например, значение x_2) в выборке.

Дискретным вариационным рядом называется перечень вариант и соответствующих им частот или относительных частот (таблица 1).

Таблица 1

X	x_1	x_2	…	x_r
n_i	n_1	n_2	…	n_r

Если признаки варьируют в широком диапазоне, т.е. число различающихся вариант велико, то данные целесообразно группировать в интервалы.

Интервальным вариационным рядом называется соответствие между интервалами и частотами, которые равны сумме частот вариант, относящихся к интервалу (таблица 2, с.6).

Таблица 2

Границы интервалов	$(a_0 - a_1)$	$(a_1 - a_2)$	…	$(a_{m-1} - a_m)$
n_i	n_1	n_2	…	n_m

В интервальных вариационных рядах вместо частот могут быть использованы и относительные частоты.

Для повышения наглядности эмпирических распределений используется их графическое представление. Наиболее распространенным способом графического представления является гистограмма и полигон.

Гистограмма представляет собой фигуру, состоящую из прямоугольников, высота которых пропорциональна числу наблюдений, попавших в данный интервал.

Для гистограммы частот высота равна $h_i = \frac{n_i}{\Delta}$, где Δ – ширина интервала, а для гистограммы относительных частот – $h_i = \frac{w_i}{\Delta} = \frac{n_i}{n\Delta}$.

Алгоритм построения гистограммы.

1. Число интервалов разбиения вычисляется по формуле Стерджесса: r=1+3,322lgn

2. Ширина интервала определяется по следующей формуле:

$$\Delta \approx \frac{x_{\max} - x_{\min}}{r - 1}.$$

3. Область значений наблюдаемой с.в. X разбивается на интервалы одинаковой длины Δ: (a_{i-1},a_i) $a_i=a_{i-1}+\Delta$, i =1..r. Причем $a_0 = x_{\min} - \frac{\Delta}{2}$.

4. Вычисляются значения n_i – количество данных, попавших в i-й интервал (a_{i-1},a_i), $i=1..r$.

5. Над каждым i-м интервалом строится прямоугольник высоты $h_i = \frac{n_i}{n\Delta}$.

Площадь каждого i-го прямоугольника будет равна $\frac{n_i}{n\Delta} * \Delta = \frac{n_i}{n}$, т.е. относительной частоте. В этом случае площадь (при больших значениях n) будет оценкой вероятности попадания с.в. X в i-й интервал. Площадь ступенчатой фигуры под графиком гистограммы будет равна единице. Таким образом, гистограмма является статистическим аналогом функции плотности для непрерывных распределений.

Полигоном называется ломаная линия, соединяющая точки (x_1,h_1), (x_2,h_2),.., (x_r, h_r). Как и в случае гистограммы, для полигона частот $h_i = \frac{n_i}{\Delta}$, а для полигона относительных частот $h_i = \frac{w_i}{\Delta} = \frac{n_i}{n\Delta}$.

Полигон можно получить из готовой гистограммы следующим образом: середины верхних сторон прямоугольников, образующих гистограмму, нужно соединить отрезками прямых.

Изображая на одном рисунке несколько полигонов одновременно, мы можем сделать анализ процесса или явления в разных условиях.

► **Пример.** В одном микрорайоне города был проведен мониторинг цен на морковь. По результатам обследования 27 магазинов, которые относятся к разным торговым сетям, были получены следующие данные:

31,6	32,6	35,6	30,6	34,9	33,9	35,3	36,9	34,1
33,7	32,8	32,6	35,6	34,2	35,5	34,2	34,6	33,1
35,8	31,8	36,1	33,4	35,9	37,4	37,9	32,9	36,5

Представьте данные в виде таблицы частот. Постройте гистограмму и полигон относительных частот.

Решение.

1. Всего наблюдений $n = 27$.

Найдем число интервалов разбиения: $r = 1+3{,}322 \bullet \lg(27) = 5{,}76 \approx 6$

2. Определим ширину интервала:

$$\Delta = \frac{x_{\max} - x_{\min}}{r-1} = \frac{37,9-30,6}{6-1} = 1,46 \approx 1,5.$$

3. Найдем границы интервалов.

$a_0 = x_{\min} - \frac{\Delta}{2} = 30,6 - \frac{1,5}{2} = 29,85,$

$a_1 = a_0 + \Delta = 29,85 + 1,5 = 31,35;$

$a_2 = a_1 + \Delta = 31,35 + 1,5 = 32,85$ и т.д.

4. Найдем частоту n_i - число наблюдений, попавших в i-й интервал (a_{i-1}, a_i), $i=1..7$. Например, 2-му интервалу (31,35-32,85) принадлежат 5 значений: 31,6; 31,8; 32,6; 32,6; 32,8. Следовательно, $n_2=5$. Полученные значения занесем в таблицу 3 (с. 9).

5. Построим гистограмму. Каждый i-й интервал «достроим» до прямоугольника: основанием прямоугольника будет сам интервал (a_{i-1}, a_i), а высота равна $\frac{n_i}{n\Delta}$. Например, над интервалом (31,35-32,85) построим прямоугольник, параллельный оси OY, высотой $h_2 = \frac{n_2}{n\Delta} = \frac{5}{27 \cdot 1,5} = 0.123$ (рис 1, с. 9)

6. Построим полигон. Найдем середины интервалов по формуле

$$u_i = \frac{a_{i-1} + a_i}{2}.$$

$u_1 = \frac{a_0 + a_1}{2} = \frac{29,85 + 31,35}{2} = 30,5$ и т.д.

7. Для построения полигона соединим отрезками точки (u_i, h_i), $i=1..6$: (30,6; 0,025), (32,1; 0,123), (33,6; 0,198),.., (38,1; 0,025) (рис. 2, с. 9).

Проанализируем полученные результаты. Самая низкая цена составила 30,6 д.е., а самая высокая – 37,4 д.е. Наиболее часто цены на морковь наблюдаются в интервале (31,35-34,35) д.е. Указанный диапазон цен наблюдался в 15 магазинах, что соответствует 55,56% от всех обследованных торговых точек.

Таблица 3

Номер интервала	Границы интервала (a_{i-1},a_i)	Середина интервала, u_i	Частота n_i	Высота $h_i = \frac{n_i}{n\Delta}$
1	29,85-31,35	30,6	1	0,025
2	31,35-32,85	32,1	5	0,123
3	32,85-34,35	33,6	8	0,198
4	34,35-35,85	35,1	7	0,173
5	35,85-37,35	36,6	5	0,123
6	37,35-38,85	38,1	1	0,025

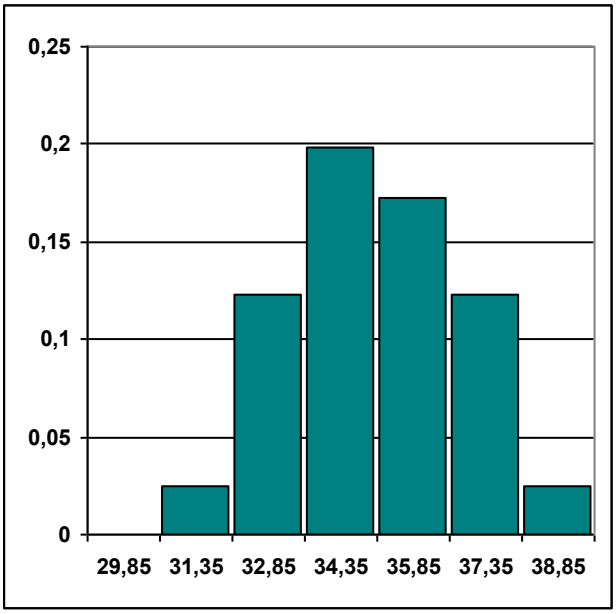


Рис.1. Гистограмма

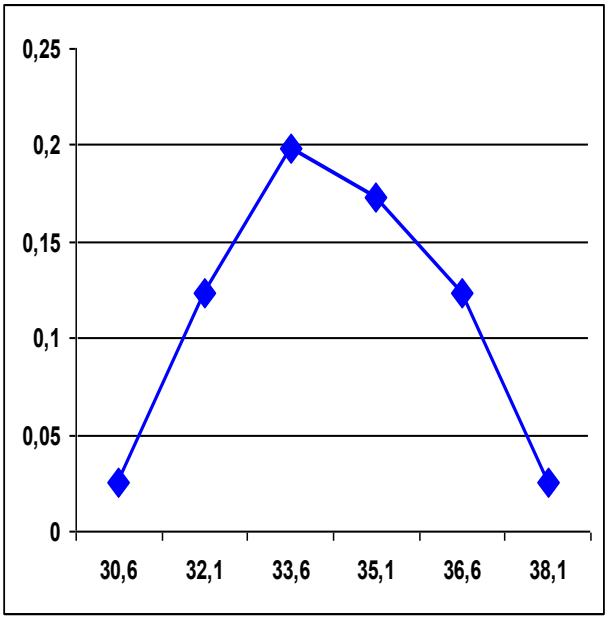


Рис. 2. Полигон

Задачи к §1.1

1. Составьте таблицу частот и постройте полигон частот для следующего распределения размеров женской одежды, проданных магазином за день: 44, 48, 52, 54, 46, 44, 52, 48, 48, 56, 52, 46, 44, 42, 50, 58, 54, 46, 44, 48, 50, 48, 52, 54, 48, 46, 54, 52, 46, 48, 48, 52, 54, 56, 44, 46.

2. В таблице приведены данные о дневном удое на одной молочной ферме в летний период. Постройте гистограмму относительных частот.

Дневной удой (л)	Менее 22	22-24	24-26	26-28	28-30	30-32
Число коров, голов	5	25	30	60	30	10

3. Распределение численности мужчин и женщин по возрастным группам (на 1 января 2021 г., тысяч человек)

	Все население	
	мужчины	женщины
Всего		
в том числе в возрасте, лет:		
0 – 4	4 155,3	3 924,3
5 – 9	4 907,9	4 643,5
10 – 14	4 231,2	4 025,4
15 – 19	3 712,3	3 559,7
20 – 24	3 446,2	3 329,8
25 – 29	4 401,0	4 181,3
30 – 34	6 272,1	6 151,4
35 – 39	6 016,9	6 113,9
40 – 44	5 296,2	5 619,0
45 – 49	4 752,4	5 227,5
50 – 54	4 163,5	4 729,0
55 – 59	4 413,1	5 443,3
60 – 64	4 347,1	5 953,4
65 – 69	3 253,3	5 215,6
70 и более	4 479,4	10 206,1

(Источник: Российский статистический ежегодник 2024)

Постройте полигоны частот для возрастных групп а) мужчин; б) женщин; в) всего населения.

Постройте гистограмму относительных частот для возрастных групп всего населения .

4. Распределение населения по возрастным группам

(на 1 января, тысяч человек)

	2001	2011	2019	2020	2021
Все население	**146 304**	**142 865**	**146 781**	**146 749**	**146 171**
в том числе в возрасте, лет:					
0 – 4	6 367	8 051	9 032	8 579	8 080
5 – 9	7 762	7 117	9 085	9 309	9 552
10 – 14	11 789	6 601	7 825	8 048	8 257
15 – 19	12 322	8 237	6 947	7 161	7 272
20 – 24	11 106	12 122	7 114	6 889	6 776
25 – 29	10 451	12 012	10 222	9 427	8 582
30 – 34	9 620	11 016	12 718	12 633	12 423
35 – 39	11 333	10 211	11 735	12 004	12 131
40 – 44	12 651	9 251	10 594	10 707	10 915
45 – 49	11 434	10 561	9 674	9 896	9 980
50 – 54	9 409	11 509	9 021	8 846	8 892
55 – 59	4 995	10 063	10 811	10 443	9 856
60 – 64	8 906	7 982	10 027	10 107	10 300
65 – 69	5 903	3 913	8 179	8 339	8 469
70 и более	12 256	14 219	13 797	14 361	14 686
Городское население	**107 072**	**105 421**	**109 454**	**109 562**	**109 252**
в том числе в возрасте, лет:					
0 – 4	4 347	5 654	6 760	6 487	6 110
5 – 9	5 226	5 011	6 546	6 730	6 982
10 – 14	8 189	4 526	5 562	5 739	5 904
15 – 19	9 370	6 064	5 056	5 221	5 300
20 – 24	8 493	9 160	5 140	4 942	4 862
25 – 29	7 999	9 193	8 035	7 357	6 588
30 – 34	7 179	8 474	9 926	9 893	9 773
35 – 39	8 341	7 777	9 148	9 382	9 497
40 – 44	9 370	6 840	8 111	8 227	8 432

45 – 49	8 667	7 676	7 280	7 481	7 568
50 – 54	7 299	8 381	6 536	6 438	6 532
55 – 59	3 761	7 470	7 778	7 521	7 103
60 – 64	6 384	6 085	7 278	7 305	7 433
65 – 69	4 041	2 939	6 086	6 178	6 244
70 и более	8 406	10 171	10 212	10 661	10 924
Сельское население	**39 232**	**37 444**	**37 327**	**37 186**	**36 919**
в том числе в возрасте, лет:					
0 – 4	2 020	2 397	2 272	2 092	1 970
5 – 9	2 536	2 106	2 539	2 580	2 569
10 – 14	3 600	2 075	2 263	2 310	2 353
15 – 19	2 952	2 173	1 891	1 940	1 972
20 – 24	2 613	2 962	1 974	1 947	1 914
25 – 29	2 452	2 819	2 187	2 070	1 994
30 – 34	2 441	2 542	2 792	2 740	2 650
35 – 39	2 992	2 434	2 587	2 621	2 634
40 – 44	3 281	2 411	2 483	2 480	2 483
45 – 49	2 767	2 885	2 394	2 414	2 412
50 – 54	2 110	3 128	2 485	2 407	2 361
55 – 59	1 234	2 593	3 033	2 922	2 753
60 – 64	2 522	1 897	2 749	2 801	2 867
65 – 69	1 862	974	2 093	2 161	2 225
70 и более	3 850	4 048	3 585	3 701	3 762

(Источник: Российский статистический ежегодник 2024)

1. По данным 2021 г. постройте гистограммы относительных частот для возрастных групп а) все население; б) городское население; в) сельское население.

2. Сравните 2 полигона частот для возрастной группы «сельское население», построенные по данным 2011 и 2021 гг.

3. Сравните 2 полигона частот для возрастных групп «городское население» и «сельское население», построенные по данным 2019 г.

5. Распределение населения по возрастным группам
(тысяч человек)

	Годы	Все население	в том числе в возрасте, лет						
			0–4	5–9	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34
Страны БРИКС	2023	146151	6864	9074	9185	7843	7439	7383	10113
Россия	2023	146151	6864	9074	9185	7843	7439	7383	10113
Бразилия	2022	214 829	14 676	14 696	14 576	15 319	16 727	17 092	17 044
Индия	2021	1 363 006	114 273	117 666	118 051	124 282	127 244	119 900	109 575
Китай	2011	1 347 305	76 271	72 093	73 507	94 574	127 726	105 011	96 645
Южно-Африканская Республика	2022	60 605	5 695	5 604	5 714	5 102	4 679	5 204	5 596

(продолжение таблицы)

	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69	70 и более
Россия	12786	11744	10483	9390	8700	10152	9234	15761
Бразилия	12786	11744	10483	9390	8700	10152	9234	15761
Индия	17 162	16 220	14 269	12 862	11 693	9 957	7 903	14 633
Китай	98 863	88 765	80 107	69 566	57 144	44 544	34 406	58 620
Южно-Африканская Республика	113 994	126 342	118 651	73 181	84 314	62 009	42 416	80 573
	5 130	4 033	3 307	2 682	2 260	1 846	1 437	2 316

(Источник: Российский статистический ежегодник 2024)

1. Постройте 5 полигонов относительных частот на одной плоскости для стран БРИКС.
2. Постройте 2 полигона частот на одной плоскости для Индии и Китая. Сравните их.

6. В результате исследования затрат времени на обработку одной детали (мин) получены следующие результаты:

2,1	5,4	3,2	3,3	0,9	4,1	2,8	1,8	1,7	2,4
3,6	1,5	2,7	2,9	3,5	1,8	0,9	1,8	3,3	3,8
4,1	5,2	3,7	4,1	2,9	1,8	3,4	1,8	3,7	3,8
3,4	3,8	1,5	5,1	2,9	3,1	3,0	5,0	4,7	3,8

Постройте гистограмму частот по выборке, предварительно проведя группировку. В качестве длины интервала взять следующие значения: Δ=0,3; б) Δ=0,5; в) Δ=1,2.

7. Постройте полигон относительных частот по данным о количестве работников в магазине: 5, 7, 1, 6, 4, 5, 6, 4, 5, 12, 17, 9, 5, 6, 12 , 13, 9, 18, 5, 6, 4, 8, 11, 13, 15, 8, 19, 12, 15, 6, 4, 8, 19, 21, 5, 4, 6, 18 (чел.).

8. Время решения контрольной задачи учениками 4-го класса (в секундах):

60 41 51 33 42 45 21 53 60 43
52 47 46 49 49 14 57 54 59 56
47 28 48 58 32 42 58 61 30 23
35 47 72 41 45 44 55 30 40 38
65 39 48 43 60 54 42 59 50 42

Построить гистограмму относительных частот. В качестве ширины интервала предлагается взять Δ=5.

6. Постройте гистограмму частот продажи акций (млн д.е.) по результатам работы биржи:

Продано на сумму (млн д.е.)	0,5-1,0	1,0-1,5	1,5-2,0	2,0-2,5	2,5-3,0	3,0-3,5
Количество продаж	46	123	525	228	35	28

1.2. Числовые характеристики выборки

Ключевые слова: среднее арифметическое, мода, медиана, размах вариации, дисперсия, стандартное отклонение, коэффициент вариации.

Вариационные ряды и графики эмпирических распределений дают наглядное представление о том, как изменяется признак в выборочной совокупности. Но они не являются достаточными для полной характеристики выборки. Числовые характеристики выборки дают количественное представление об эмпирических данных и позволяют их сравнивать между собой. Наибольшее практическое применение имеют характеристики положения и характеристики вариации эмпирических распределений.

1.2.1. Характеристики положения

В этом разделе мы рассмотрим характеристики положения, определяющие положение центра эмпирического распределения.

1. *Выборочное среднее* $\bar{x}$ *(среднее арифметическое)*.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ , где} \qquad (1)$$

n – объем выборки; x_i – варианты выборки.

Для сгруппированных данных формулы выборочного среднего приведены в таблице 1 (с. 18).

Среднее показывает, какие значения признака наиболее характерны для данной совокупности в целом.

2. *Медианой (*$\hat{Me}$) называется значение признака, которое приходится на центральный член вариационного ряда.

Таким образом, одна половина вариационного ряда меньше (или равна) медианы, а вторая – больше (или равна).

$$\hat{Me} = \begin{cases} x_{k+1}, & n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & n = 2k \end{cases}, \qquad (2)$$

где n – объем выборки

Для интервального ряда сначала определяется медианный интервал, потом по формуле вычисляется медиана.

а) *Медианным* будет тот интервал, в котором сумма частот n_i впервые окажется больше $\frac{\sum n}{2}$ **:**

$$\begin{cases} n_1 + n_2 + ... + n_{k-1} < \dfrac{\sum n}{2} \\ n_1 + n_2 + ... + n_k > \dfrac{\sum n}{2} \end{cases}$$, т.е. k-й интервал будет медианным.

б) Медиана для интервального ряда:

$$\hat{Me} = x_{Me} + \Delta \frac{\frac{\sum n}{2} - S_{Me-1}}{n_{Me}}, \qquad (3)$$

где x_{Me} – нижняя граница медианного интервала,

Δ – величина медианного интервала,

S_{Me-1} – сумма накопленных частот, предшествующих медианному интервалу,

n_{Me} – частота медианного интервала.

Медиана применяется при статистическом контроле качества продукции и технологического процесса на предприятии; при изучении распределения семей по величине дохода и т. п.

3. *Модой* ($\hat{Mo}$) называется наиболее часто встречающееся в выборке значение признака.

Для дискретного ряда мода определяется визуально по максимальной частоте или частости.

Для интервального ряда сначала определяется модальный интервал, потом по формуле вычисляется мода.

Модальным называется интервал, имеющий наибольшую частоту (частость).

$$\hat{Mo} = x_{Mo} + \Delta \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})}, \text{ где} \quad (4)$$

x_{Mo} – нижняя граница модального интервала,

Δ – величина модального интервала,

n_{Mo} – частота модального интервала,

n_{Mo-1} – частота интервала, предшествующего модальному,

n_{Mo+1} – частота интервала, следующего за модальным.

Совокупность может иметь одну моду (унимодальное распределение) или несколько мод (мультимодальное распределение).

Мода широко используется при изучении покупательского спроса, ходовых размеров одежды и обуви, регистрации цен и т.д.

Мода и медиана, в отличие от среднего, мало чувствительны к колебаниям крайних вариант. Поэтому, если эмпирическое распределение сильно асимметрично, то предпочтительнее использовать моду и медиану. Вообще, среднее, мода и медиана совпадают только в случае симметричного и унимодального (с одним максимумом) распределения.

В статистических исследованиях применяются различные виды средних величин. Выбор средней должен осуществляться с учетом содержания исследуемого показателя и формы представления данных.

1.2.2. Характеристики вариации

Средние величины не дают полной информации о варьирующем признаке. При одинаковых средних признаки могут отличаться по величине и разбросу значений. Поэтому наряду со средними значениями вычисляют и характеристики вариации (рассеяния) выборки.

1. *Размах вариации R* показывает разность между наибольшими и наименьшими значениями выборки:

$$R = x_{max} - x_{min}\,. \quad (5)$$

2. Выборочная дисперсия и выборочное стандартное отклонение являются важнейшими характеристиками вариации.

Выборочной дисперсией S^2 называется средний квадрат отклонения значений вариант от их средней арифметической:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (6)$$

Дисперсия после преобразования принимает вид:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2. \qquad (7)$$

Если массив данных был подвергнут группировке, то выборочная дисперсия вычисляется по формулам таблицы 1.

Таблица 1

Данные	Выборочное среднее	Выборочная дисперсия
$x_1, x_2, .., x_n$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$	$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2$
Дискретный вариационный ряд	$\bar{x} = \frac{1}{n}\sum_{j=1}^{r}x_j n_j$	$S^2 = \frac{1}{n}\sum_{j=1}^{r}(x_j - \bar{x})^2 n_j = \frac{1}{n}\sum_{j=1}^{r}x_j^2 n_j - (\bar{x})^2$
Интервальный вариационный ряд	$\bar{x} = \frac{1}{n}\sum_{j=1}^{r}v_j n_j$	$S^2 = \frac{1}{n}\sum_{j=1}^{r}(v_j - \bar{x})^2 n_j = \frac{1}{n}\sum_{j=1}^{r}v_j^2 n_j - (\bar{x})^2$

$n = \sum_{j=1}^{r} n_j$ – объем выборки; r – число интервалов группировки;

x_j – варианта дискретного ряда; v_j – середина j-го интервала интервального вариационного ряда;

n_j – частота варианты x_j или j-го интервала.

Выборочное стандартное отклонение (или *среднее квадратическое отклонение*) определяется как корень квадратный из дисперсии:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ . \qquad (8)$$

Выборочное стандартное отклонение имеет ту же размерность, что и сам признак. Поэтому на практике обычно используется стандартное отклонение, а не дисперсия.

3. *Коэффициент вариации*

Выборочное стандартное отклонение выражается в тех же единицах измерения, что и характеризуемый им признак. В практике же бывают ситуации, когда сравнивают изменчивость признаков, выраженных разными единицами. В таких случаях используют не абсолютные, а относительные показатели вариаций. Одним из таких показателей является *коэффициент вариации*:

$$C_v = \frac{S}{\bar{x}}100\%\ . \qquad (9)$$

Если $C_v \le 10$ %, то варьирование считается слабым, при 11 % $\le C_v \le$ 25 % – средним, и значительным при $C_v > 25$ %. Если коэффициент вариации не превышает 10 %, то выборку можно считать однородной, т.е. полученной из одной генеральной совокупности.

Следует заметить, что к понятиям среднее, дисперсия, стандартное отклонение часто добавляются слова «эмпирическое». Тем самым подчеркивается, что указанные характеристики вычислены по экспериментальным, опытным данным.

► **Пример 1.** Для активизации работы сотрудников руководство компании внедрило систему поощрительных баллов. За 1 месяц сотрудники отдела продаж заработали следующие баллы: 5, 17, 14, 7, 3. Найти средний балл, медиану, дисперсию и стандартное отклонение.

Решение. Число сотрудников равно $n = 5$.

1) Среднее количество поощрительных баллов находим по формуле среднего арифметического (1):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{5+17+14+7+3}{5} = \frac{46}{5} = 9{,}2 \text{ балла}.$$

2) Для определения медианы сначала построим вариационный ряд:

3, 5, 7, 14, 17.

Найдем медиану двумя способами.

Способ 1 – визуальный.

Медиану, как середину вариационного ряда, можно определить зрительно – это значение «7».

$\hat{Me}$

3, 5, **7**, 14, 17.

Способ 2 – по формуле (2).

$n = 5$ – число нечетное, т.е. $n = 2k + 1 = 5$. Следовательно, $k = 2$.

$\hat{Me} = x_{k+1} = x_{2+1} = x_3 = 7$ *балла* – это поощрительные баллы сотрудника с условным номером 34. Из оставшихся у половины рабочих (№№ 1–2) баллов меньше медианного, а у другой половины (№№4–5) – больше.

3) Дисперсию можно найти двумя способами.

Способ 1. Используем определение дисперсии (6):

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{7}((5-9{,}2)^2 + (17-9{,}2)^2 + (14-9{,}2)^2 + (7-9{,}2)^2 +$$

$$+ (3-9{,}2)^2 = 28{,}96$$

Стандартное отклонение равно $S = \sqrt{S^2} = \sqrt{28{,}96} = 5{,}38$.

Способ 2. Используем упрощенную формулу выборочной дисперсии (7):

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{7}(5^2 + 17^2 + 14^2 + 7^2 + 3^2) - (9{,}2)^2 = 28{,}96.$$

Способ 2 требует меньше вычислений. Но, если мы оперируем большими числами, то предпочтительнее будет использовать способ 1 или перейти к условным вариантам (§2.3). ■

► Пример 2 (продолжение Примера § 1.1).

В таблице представлены результаты мониторинга цен на морковь, продажи которого осуществлялись в одном микрорайоне города. Найти выборочное среднее, выборочную дисперсию и выборочное стандартное отклонение.

Цена (X), д.е.	29,85-31,35	31,35-32,8	32,85-34,35	34,35-35,85	35,85-37,35	35,85-37,35
Частота (n_i), ед.	1	5	8	7	5	1

Решение. Общее количество обследованных магазинов составило: $n = \sum n_i = 1 + 5 + 8 + 7 + 5 + 1 = 27$.

Преобразуем данный интервальный вариационный ряд в безынтервальный. Для этого в качестве вариант возьмем середины интервалов. Например, для 4-го интервала середина будет $v_4 = \frac{34,35+35,85}{2} = 35,1.$ В результате таких преобразований мы получаем следующую таблицу.

Цена (X), д.е.	30,6	32,1	33,6	35,1	36,6	38,1
Частота (n_i), ед.	1	5	8	7	5	1

1) Вычислим выборочное среднее.

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} v_j n_j = \frac{1}{27}(30,6\cdot 1 + 32,1\cdot 5 + 33,6\cdot 8 + 35,1\cdot 7 + 36,6\cdot 5 + 38,1\cdot 1) = 34,32 \text{ д.е.}$$

2) Вычислим выборочную дисперсию.

$$S^2 = \frac{1}{n}\sum_{j=1}^{n} v_j^2 n_j - \left(\bar{v}\right)^2 = \frac{1}{27}(30,6^2\cdot 1 + 32,1^2\cdot 5 + 33,6^2\cdot 8 + 35,1^2\cdot 7 + 36,6^2\cdot 5 + 38,1^2\cdot 1) - (34,32)^2 = 3,23$$

3) Выборочное стандартное отклонение равно

$$S = \sqrt{S^2} = \sqrt{3,23} = 1,797 \text{ д.е.}$$

Итак, мы нашли, что среднее цена на морковь в указанном микрорайоне составляет 34,32 д.е., а выборочное стандартное отклонение – 1,797 д.е. ■

Задачи к § 1.2

В задачах 1 – 4 вычислить выборочное среднее, медиану, моду, выборочную дисперсию.

1. 6, 4, 3, 3, 6, 4, 5, 1, 2, 1, 3;

2. 1, 2, 3, 4, 5, 5, 12;

3. 1, 2, 3, 4, 5, 5, 9;

4. 3.1, 3.0, 1.5, 1.8, 2.5, 3.1, 2.4, 2.8, 1.3.

5. В откормочном цехе фермерского хозяйства имеется бычки, вес которых распределяется следующим образом:

Вес бычков, кг	598-603	603-608	608-613	613-618
Количество, голов	7	5	14	3

Найти средний вес бычка, коэффициент вариации.

7. С целью изучения мнения покупателей о качестве продукции, выпускаемой фирмой, в магазине проводили опрос. Покупателям предлагалось дать оценку качества по десятибалльной шкале. Результаты представлены в следующей таблице.

Оценка качества, балл	1-2	3-4	5-6	7-8	9-10
Количество покупателей	4	6	42	46	32

Определите средний балл качества продукции, коэффициент вариации.

8. Найдите выборочную дисперсию, если коэффициент вариации равен 15,7%, а среднее арифметическое – 7,62.

9. Для характеристики производственного стажа работников одной из отраслей промышленности произведено обследование различных категорий работников. В результате обследования были получены следующие данные:

Стаж работы, лет	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16
Число рабочих	8	15	20	25	10	8	2	8
Число мастеров	6	10	22	20	26	7	6	11
Число технологов	0	5	20	10	32	20	10	5

Определить средний стаж работников по каждой категории и выборочное стандартное отклонение.

10. Распределение автомобилей автотранспортного предприятия по величине суточного пробега за 1 июня следующее.

Суточный пробег автомобиля, км	До 160	160-180	180-200	Свыше 200
Количество автомобилей	6	8	12	4

Найдите: а) средний суточный пробег одного автомобиля; б) выборочную дисперсию.

11. В бригаде сборщиков заработная плата за смену 420 д.е. у четырех рабочих, 510 д.е. – у трех рабочих и по 365 д.е. – у семи рабочих. Найдите среднюю заработную плату по бригаде за смену.

12. Найдите выборочное стандартное отклонение, если для 75 наблюдений сумма квадратов равна 2679, выборочное среднее – 6,45.

13. Браки по возрастам жениха и невесты

	2000	2010	2018	2019	2020
Всего браков	**897 327**	**1 215 066**	**893 039**	**950 167**	**770 857**
По возрасту жениха, лет					
до 18	3 703	1 131	454	564	598
18 – 24	403 851	372 782	170 440	177 912	143 938
25 – 34	303 216	564 776	456 639	469 220	372 864
35 и более	186 133	276 219	265 506	302 389	253 457
не указан	424	158	–	82	–
По возрасту невесты, лет					
до 18	29 889	11 698	4 593	5 141	4 569
18 – 24	511 446	554 772	285 580	296 984	239 507
25 – 34	212 528	451 318	386 652	395 684	314 172
35 и более	143 193	197 162	216 214	252 276	212 609
не указан	271	116	–	82	–

(Источник: Российский статистический ежегодник 2024)

1. По данным 2010 и 2020 гг. найдите средний возраст а) жениха; б) невесты.

2. Найдите средний возраст жениха по данным 2000, 2010, 2018, 2019 гг. Сравните значения.

14. Дисперсия признака равна 3600, коэффициент вариации равен 50%. Найти среднюю величину признака.

15. Структура безработных по продолжительности поиска работы (в процентах к итогу).

	Безработные – всего 100%	В том числе ищут работу			
		до 3 месяцев	от 3 до 6 месяцев	от 6 до 12 месяцев	12 месяцев и более
Всего					
2021	100	33,5	21,8	22,2	22,4
2022	100	39,6	20,4	22,1	18,0
2023	100	41,9	21,6	20,0	16,4
Мужчины					
2021	100	34,5	21,5	22,2	21,8
2022	100	40,6	20,2	21,3	17,9
2023	100	42,7	21,7	19,2	16,4
Женщины					
2021	100	32,6	22,1	22,3	23,0
2022	100	38,7	20,5	22,8	18,0
2023	100	41,2	21,5	20,9	16,4

(Источник: Российский статистический ежегодник 2024)

1. По данным 2023 г. вычислите среднее время поиска работы а) для всего населения; б) для мужчин; в) для женщин.

2. Вычислите среднее время поиска работы для женщин по данным 2021, 2022, 2023 гг.

16. Имеются данные о сроках функционирования коммерческих банков на начало года.

Срок функционирования, лет	1-2	2-3	3-4	4-5	5-6	Свыше 6
Удельный вес банков, %	16	20	28	14	14	8

Определить средний срок функционирования банков, моду и медиану, стандартное отклонение.

Контрольные вопросы

1. Как строится гистограмма?
2. Каким образом следует выбирать интервалы группировки при построении гистограммы?
3. Сколько интервалов нужно выбрать для построения гистограммы?
4. В каких случаях предпочтительнее строить гистограмму, а в каких случаях − полигон?
5. Какие виды средних величин применяются в математической статистике?
6. Почему выборочное среднее интервального ряда является приближенной средней, от чего зависит степень ее приближения?
7. Дайте определение выборочной дисперсии.
8. Что называется стандартным отклонением? По каким формулам оно вычисляется?

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

Ключевые слова: точечная оценка, статистика, состоятельность, несмещенность, эффективность, доверительный интервал, надежность, уровень значимости, предельная ошибка выборки.

Одна из основных задач математической статистики заключается в нахождении распределения случайной величины X (с.в.) по данным выборки. Очень часто вид распределения с.в. X известен с точностью до неизвестных параметров. В этом случае нам остается найти приближенные значения этих параметров. Допустим, $F(x;\theta)$ – функция распределения с.в. X, содержащая неизвестный параметр θ, и пусть $x_1,x_2,..,x_n$ – выборка из генеральной совокупности. Требуется определить, используя эти наблюдения, число, которое можно было бы взять в качестве значения θ, или интервал, о котором можно было бы утверждать, что он с вероятностью, близкой к 1, накрывает это значение. В первом случае будем говорить о точечной оценке, во втором – о доверительном интервале (рис. 1).

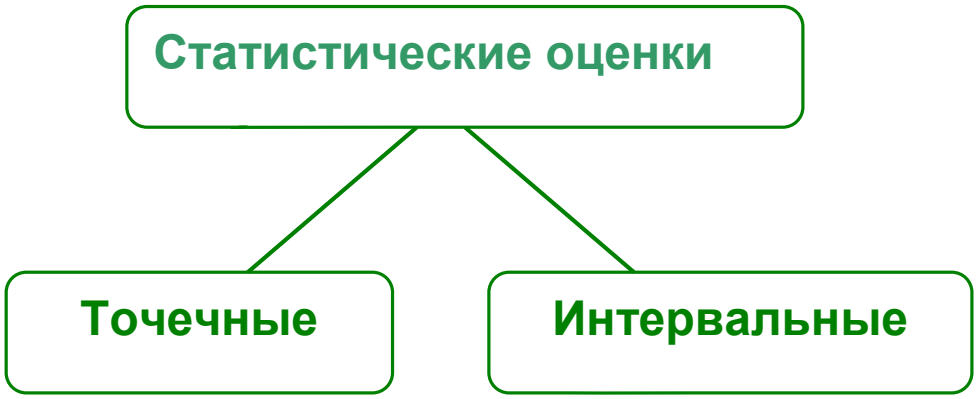


Рис. 1

Задача оценки параметров состоит в получении наилучших в определенном смысле оценок параметров распределения на основе выборочных данных.

1. Точечные оценки.

Для нахождения оценки неизвестного параметра θ мы должны построить некоторую статистику $T(x_1, x_2, .., x_n)$, которая является приближением к θ. Такая статистика T называется *оценкой* и обозначается, обычно, $\theta = \theta(x_1, x_2, .., x_n)$. В дальнейшем термин «статистика» будет часто использоваться, поэтому уточним его: *статистикой* будем называть любую функцию от наблюдений $T = T(x_1, x_2, .., x_n)$.

Пусть $x_1, x_2, .., x_n$ – реализация случайной выборки из распределения со средним значением μ и дисперсией σ^2. Припишем каждому значению вес $1/n$. Тогда пара $(x_k, \frac{1}{n})$, $k=1..n$, образует выборочное распределение.

Все характеристики, подсчитанные по выборочному распределению, называются выборочными характеристиками. В частности, в качестве оценки для истинного среднего μ возьмем *выборочное среднее* $\overline{x}$ (§ 1.2). Мода и медиана также являются оценками для μ. Оценкой дисперсии σ^2 служит выборочная дисперсия S^2 (или S_x^2) (§ 1.2).

Важнейшими свойствами оценки, по которым мы определяем ее близость к истинному значению, являются свойства состоятельности, несмещенности, эффективности.

Точечная оценка $\hat{\theta}$ называется *состоятельной*, если при неограниченном увеличении объема выборки $(n \to \infty)$ она стремится к истинному значению параметра θ.

Так, выборочное среднее $\bar{x}$ является состоятельной оценкой истинного среднего μ, а выборочная дисперсия S^2 – состоятельная оценка дисперсии σ^2.

Статистика θ называется *несмещенной* оценкой параметра θ, если ее математическое ожидание равно самому оцениваемому параметру, т.е. $E\theta = \theta$.

Следовательно, в случае несмещенной оценки отклонение оценки от истинного значения носит несистематический, случайный характер, что означает отсутствие систематической ошибки.

Из приведенных выше оценок только выборочное среднее является несмещенной оценкой среднего μ. Несмещенной оценкой дисперсии σ^2 является исправленная выборочная дисперсия, вычисляемая по формуле:

$$S^2_{несм} = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где S^2 вычисляется по формуле (2), § 1.2.

Оценка является *эффективной,* если она имеет наименьшую дисперсию по сравнению с другими несмещенными оценками того же параметра.

Выборочное среднее $\bar{x}$ и выборочная дисперсия S^2 являются эффективными оценками соответствующих параметров.

2. Интервальные оценки.

В предыдущем разделе мы рассмотрели точечные оценки. Но если мы имеем дело с выборкой малого объема, то случайный характер величины θ может привести к значительному расхождению между θ и $\hat{\theta}$, т.е. к ошибкам. В этом случае параметр θ удобно оценивать некоторым интервалом.

Интервальной называют оценку, которая определяется двумя числами – концами интервалов.

Доверительным интервалом для параметра θ называют интервал $(\hat{\theta}_1, \hat{\theta}_2)$ со случайными концами, накрывающий истинное значение θ с вероятностью не меньшей 1 – α, т.е.

$$P\ (\hat{\theta}_1 < \theta < \hat{\theta}_2) \geq 1 - \alpha.$$

Число $\gamma = 1 - \alpha$ называется *доверительным уровнем (надежностью, доверительная вероятность*).

Число α называется *уровнем значимости.*

Интерпретация доверительного уровня: в среднем в (1−α)100% случаев фиксированное, но неизвестное значение θ будет накрыто доверительным интервалом. Обычно задают надежность, равную 0,95; 0,99; 0,999 (т.е. α = 0,05; 0,01; 0,001) и риск ошибиться в первом случае будет 1 раз на 20 испытаний, во втором – один раз на 100 испытаний и в третьем – один раз на 1000 испытаний.

Наибольшее отклонение оценки $\hat{\theta}$ от оцениваемого параметра θ, которое возможно с заданным доверительным уровнем γ, называется предельной ошибкой выборки Δ.

Очень часто доверительные интервалы параметра θ представлены в виде $(\hat{\theta}-\Delta;\ \hat{\theta}+\Delta)$.

Предельная ошибка выборки характеризует точность оценки: чем больше значение Δ, тем шире интервал и, следовательно, тем меньше точность.

1. Доверительный интервал для среднего значения нормального распределения при известной дисперсии:

$$\bar{x}-u_\alpha\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{x}+u_\alpha\frac{\sigma}{\sqrt{n}}, \text{ где} \qquad (1)$$

u_α – квантиль нормального распределения, который находится из уравнения $\Phi(u_\alpha)=1-\frac{\alpha}{2}$.

Для случая повторной выборки $\Delta=u_\alpha\frac{\sigma}{\sqrt{n}}$. В случае бесповторной выборки в эту формулу включается поправка. Но для простоты вычислений (с точки зрения учебного процесса) задачи данного раздела будут сформулированы только для случая повторной выборки.

2. Доверительный интервал для среднего значения нормального распределения при неизвестной дисперсии:

$$\bar{x}-t_\alpha\frac{s}{\sqrt{n-1}}\le\mu\le\bar{x}+t_\alpha\frac{s}{\sqrt{n-1}}, \text{ где} \qquad (2)$$

S^2 – выборочная дисперсия, t_α – квантиль распределения Стьюдента.

3. Доверительные интервалы для оценки дисперсии нормального распределения:

$$\frac{nS^2}{\chi^2_{\alpha}} \le \sigma^2 \le \frac{nS^2}{\chi^2_{1-\alpha}}, \quad \text{где} \qquad (3)$$

$\chi^2_{\alpha}, \chi^2_{1-\alpha}$ – квантили распределения χ^2 (хи-квадрат) с $\nu = n - 1$ степенями свободы.

4. Доверительный интервал для доли ($n > 100$)**:**

$$w - u_{\alpha}\sqrt{\frac{w(1-w)}{n}} \le p \le w + u_{\alpha}\sqrt{\frac{w(1-w)}{n}}, \text{ где} \qquad (4)$$

w – относительная частота, u_{α} – квантиль нормального распределения, который находится из уравнения $\Phi(u_{\alpha}) = 1 - \frac{\alpha}{2}$.

Замечание. При достаточно больших объемах выборки ($n > 30$) распределение Стьюдента приближается к нормальному. Поэтому, если σ^2 неизвестна и $n > 30$, то доверительный интервал для среднего можно вычислить по формуле (1). При этом величина стандартного отклонения σ заменяется на $S_{несм}$.

▶ **Пример 1**. С целью исследования качества выпускаемых автомобильных шин случайным образом были выбраны 156 легковых автомобилей. Средний пробег шин в городских условиях составил 48,3 тыс км. Постройте с надежностью 0,95 доверительный интервал для среднего пробега шин, если стандартное отклонение известно и равно 5,2 тыс км.

Решение. Т.к. стандартное отклонение известно, то доверительный интервал строится по формуле (1). По условию $n = 156$; $\sigma = 5{,}2$; $\bar{x} = 48{,}3$. Поскольку $\gamma = 0{,}95$, то $\alpha = 1 - \gamma = 1 - 0{,}95 = 0{,}05$. Найдем $u_{0{,}05}$ из уравнения $\Phi(u_{0{,}05}) = 1 - \frac{0{,}05}{2} =$ **0,975**. По таблице «Нормальное распределение» (приложение, таблица I) находим $u_{0{,}05} =$**1,96**, т.е. Ф(1,96) = 0,975.

Нормальное распределение $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$

x	Сотые доли				
	…	4	5	**6**	7
…		…	…	…	…
1,8		0,9671	0,9678	0,9686	0,9693
1,9		0,9738	0,9744	**0,9750**	0,9756
2,0		0,9793	0,9798	0,9803	0,9808

Рис. 2. Фрагмент таблицы «Нормальное распределение»

Подставляя эти данные в формулу, получаем:

$$48{,}3-1{,}96\frac{5{,}2}{\sqrt{156}}\le\mu\le 48{,}3+1{,}96\frac{5{,}2}{\sqrt{156}},$$

$$47{,}483\le\mu\le 49{,}116.$$

То есть, с надежностью 0,95 искомый доверительный интервал для среднего времени пробега шин равен (47,483; 49,1) тыс км.

Найдем точность оценки: $\Delta=u_\alpha\frac{\sigma}{\sqrt{n}}=1{,}96\frac{5{,}2}{\sqrt{156}}=0{,}816.$ ■

▶ **Пример 2.** В рекрутинговом агентстве для определения среднего размера заработной платы по позиции «Программист языка NN» случайным образом выбраны 12 предложений о вакансиях. Средняя заработная плата по указанной позиции составила 150 тыс. д.е., а выборочное стандартное отклонение – 68,5 тыс. д.е. Определите границы 95 %-х интервалов для среднего, дисперсии, стандартного отклонения размера заработной платы.

Решение. По условию n = 12,

$\bar{x}=150$ *тыс д.е.*, $S^2=68{,}5$, $S=8{,}276$ *тыс д.е.*

Доверительная вероятность равна γ = 0,95. Тогда уровень значимости $\alpha = 1-\gamma = 1-0{,}95 =$ **0,05.**

а) Найдем доверительный интервал для среднего размера заработной платы.

По таблице «Критические точки распределения Стьюдента» (таблица III приложения) при заданном уровне значимости $\alpha = 0{,}05$, помещенному в верхней строке таблицы (для двусторонней критической области), и числе степеней свободы $\nu = n - 1 = 12 - 1 = 11$ находим $t_\alpha = 2{,}2$ (рис. 3).

Критические точки распределения Стьюдента

Число степеней свободы ν	Уровень значимости α (двусторонняя критическая область)					
	0, 10	0,05	0,02	0,01	0,002	0,001
…	…	…	….	…	…	…
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,2	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32

Рис. 3. Фрагмент таблицы «Критические точки распределения Стьюдента»

По формуле (2) определяем границы доверительного интервала:

$$7{,}93 - 2{,}2\frac{1{,}14}{\sqrt{12-1}} \le \mu \le 7{,}93 + 2{,}2\frac{1{,}14}{\sqrt{12-1}},$$

$$7{,}93 - 0{,}76 \le \mu \le 7{,}93 + 0{,}76$$

$$7{,}17 \le \mu \le 8{,}69.$$

Таким образом, можно утверждать, что доверительный интервал (7,17; 8,69) (тыс.д.е.) накрывает среднее значение размера заработной платы с вероятностью 0,95.

б) Найдем доверительный интервал для дисперсии.

По таблице «Критические точки распределения χ^2» (приложение, таблица II) найдем квантили распределения χ^2 с $\nu = n - 1 = 12 - 1 = 11$ степенями свободы (рис. 3, с. 34):

$\chi^2_\alpha = \chi^2_{0{,}05} = 19{,}675$ и $\chi^2_{1-\alpha} = \chi^2_{0{,}95} = 4{,}575$.

Критические точки распределения χ^2

ν \ α	0,99	**0,95**	0,90	0,10	**0,05**	0,02	0,01	0,001
…								
10	2,558	3,940	4,865	15,987	18,307	21,161	23,209	29,588
11	3,053	**4,575**	5,578	17,275	**19,675**	22,618	24,725	31,264
12	3,571	5,226	6,304	18,549	21,026	24,054	26,217	32,909

Рис. 4. Фрагмент таблицы «Критические точки распределения χ^2»

Вычислим доверительный интервал для дисперсии по формуле (4):

$$\frac{12\cdot 1{,}3}{19{,}675} \le \sigma^2 \le \frac{12\cdot 1{,}3}{4{,}575},$$

$$0{,}79 \le \sigma^2 \le 3{,}4.$$

Доверительный интервал для стандартного отклонения:

$$\sqrt{0{,}79} \le \sigma \le \sqrt{3{,}4} \quad \text{или } 0{,}889 \le \sigma \le 1{,}844.$$

С надежностью 0,95 доверительный интервал для дисперсии равен (0,79; 3,4) (млн руб.)^2, а для стандартного отклонения – (0,889; 1,844) (млн руб.). ■

▶ **Пример 3.** Опрос 150 случайно отобранных жителей города показал, что 94 опрошенных довольны работой нового мэра. С вероятностью 0,98 оцените границы доли жителей города, которые также положительно оценивают работу мэра.

Решение. Из условия известно, что $n = 150$, $m = 94$, $\gamma = 0{,}98$. Тогда уровень значимости равен $\alpha = 1 - \gamma = 0{,}02$. Границы доверительного интервала будем искать по формуле (5).

Вычислим относительную частоту: $w = \frac{m}{n} = \frac{94}{150} = 0{,}627$.

Найдем $u_{0,01}$ из уравнения $\Phi(u_{0,01}) = 1 - \frac{\alpha}{2} = 1 - \frac{0,02}{2} =$ **0,99**. В таблице «Нормальное распределение» (рис. 5) нет значения 0,995, а есть ближайшие значения 0,9949 и 0,9951. Поскольку 0,99 есть среднее арифметическое из 0,9949 и 0,9951, то $u_{0,01} = \frac{2,32 + 2,33}{2} =$ **2,325**.

Нормальное распределение $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$

x	Сотые доли				
	…	1	**2**	**3**	4
…		…	…	…	…
2,2		0,9864	0,9868	0,9871	0,9875
2,3		0,9896	**0,9898**	**0,9901**	0,9904
2,4		0,9920	0,9922	0,9925	0,9927

Рис. 5. Фрагмент таблицы «Нормальное распределение»

Следует заметить, что квантили с любой степенью точности можно найти с помощью статистических пакетов или, хотя бы, Excel.

Вычисляем доверительный интервал по формуле (4):

$$0,627 - 2,325\sqrt{\frac{0,627(1-0,627)}{150}} \le p \le 0,627 + 2,325\sqrt{\frac{0,627(1-0,627)}{150}},$$

$$0,0531 \le p \le 0,1735.$$

С вероятностью 0,98 можно утверждать, что доля жителей города, которые положительно оценивают работу мэра, заключена от 6,31 % до 17,35 %. ■

Задачи к § 3

В задачах этого раздела предполагается, что все случайные повторные выборки имеют нормальное распределение или достаточно близкое к нормальному.

1. При взвешивании 12 контейнеров было установлено, что средний вес груза равен 135 кг, а выборочное стандартное отклонение равно 6 кг. Построить доверительные интервалы для оценки средней массы груза с надежностями 0,95 и 0,99. Сравнить их.

2. По выборке объема $n = 23$, извлеченной из нормальной совокупности, были найдены выборочное среднее $\overline{x} = 17,5$ и выборочное стандартное отклонение $S = 4,2$. С надежностью 0,95 построить доверительные интервалы для среднего, дисперсии и стандартного отклонения.

3. Анализ расхода воды в 15 населенных пунктах, имеющих примерно одинаковое количество населения, показал, что среднее значение расхода воды составляет 60 000 л в день, выборочная дисперсия – 3450 $л^2$. Определить с надежностью 0,9 возможные пределы ежедневного среднего расхода воды.

4. В результате обследования 120 коров установлено, что у данной группы коров средний годовой удой равен 3200 л. Найти доверительный интервал для оценки среднего годового удоя с надежностью 0,94, если известно, что дисперсия равна 62500 $кг^2$.

5. Для выявления среднего времени простоя одного ткацкого станка было взято на выборку 87 станков. Оказалось, что в среднем станок простаивает 24 мин со стандартным отклонением 5 мин. С надежностью 0,97 оценить доверительный интервал для среднего времени простоя.

6. Для определения соответствия качества продуктов указанным нормам было проверено 16 образцов. По результатам обработки данных было определено, что среднее содержание углеводов в единице продукции равно 18,8 г, а выборочная дисперсия – 4,7 $г^2$. Постройте 95%-е доверительные интервалы для среднего и дисперсии содержания углеводов в единице продукции.

7. Исследование возраста участников некоторой компьютерной игры дало следующие результаты (лет): 16, 19, 24, 19, 43, 30, 28, 19, 24,

22, 20, 17, 24, 26, 14, 19, 21, 20. Определите границы для среднего и дисперсии возраста с надежностью 0,99.

8. Для стада из 26 голов был установлен средний годовой привес скота – 196 кг. Определить с надежностью 0,98 возможные пределы вариации годового привеса скота, если выборочное стандартное отклонение равно 50 кг.

9. Среднее время пребывания в очереди на кассу супермаркета составляет 56 сек., а выборочная дисперсия – 114. Определить доверительный интервал с надежностью 0,999 для среднего значения, если было произведено 18 замеров.

10. В продуктовом магазине проводился опрос покупателей (127 человек). На вопрос «Удобно ли расположены товары?» 85 покупателей ответили утвердительно. Найдите границы для доли покупателей, которые ответили утвердительно. Надежность принять 0,93.

11. На большом участке леса случайно отобрана сосны и измерены их диаметры.

Диаметр, см	21-23	23-25	25-28	28-31	31-36
Число деревьев, ед.	2	6	9	4	2

Найти надежность того, что средний диаметр деревьев на всем участке заключен в границах от 25,44 до 27,20 см. Построить с надежностью 0,99 доверительный интервал для среднего диаметра всех деревьев.

12. В универмаге планируется провести очередное изучение структуры и стоимости покупок. Сколько следует опросить покупателей, чтобы с вероятностью 0,99 определить стоимость покупки с точностью не менее 75 д.е. Дисперсию принять равной 36 500 (по прошлому обследованию).

13. Для предварительного контроля качества было выбрано 132 изделия. Оказалось, что бракованных изделий 15 штук. Определить с

надежностью 0,91 возможные пределы для доли бракованных изделий.

14. Проводились исследования по определению среднего веса яблока нового сорта. Каким должен быть объем выборки, чтобы с вероятностью 0,99 можно было утверждать, что средний вес яблока в ней отличается от среднего веса яблока по всей совокупности не более чем на 5г.? Считать дисперсию равной 576 $г^2$.

15. Для установления среднего размера вклада определенной категории вкладчиков в банках города необходимо провести случайную повторную выборку лицевых счетов. Считая стандартное отклонение равным 150 д.е., найти минимальный объем выборки при условии, что с вероятностью 0,9 точность оценки не превысит 10 д.е.

16. Вариация (стандартное отклонение) ежесуточного дохода случайно выбранных 12 киосков оказалась равной 95 д.е. Найдите точность, которая с надежностью 0,98 определяет среднесуточный доход.

17. Найти точность, с которой определен средний годовой привес скота (200кг) для стада из 81 головы, если стандартное отклонение равно 48 кг. Оценку произвести, приняв доверительные вероятности $\gamma_1 = 0,9$; $\gamma_2 = 0,95$; $\gamma_3 = 0,99$. Как отличаются оценки?

18. Результаты исследования длительности оборота оборотных средств торговых фирм города (в днях) представлены в группированном виде.

Длительность оборота (дни)	14-23	23-32	32-41	41-50	50-59	59-68
Число фирм	2	3	9	17	10	6

Построить доверительный интервал с надежностью 0,95 для средней длительности оборота оборотных средств в торговых фирмах города при условии, что стандартное отклонение известно и равно 10 дням.

Контрольные вопросы

1. В чем состоит задача оценки параметров?
2. Какие требования предъявляются к оценкам параметров?
3. Дайте определение состоятельности оценки и проинтерпретируйте смысл этого определения.
4. Можно ли сказать, что состоятельная оценка лучше не состоятельной оценки?
5. Определите разницу в понятиях: «статистика» (оценка) и «параметр» распределения.
6. В каких случаях строятся доверительные интервалы?
7. Можно ли утверждать, что чем выше надежность, тем «лучше» доверительный интервал?
8. Какие значения уровня значимости являются наиболее употребительными?

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Ключевые слова: статистические гипотезы, нулевая и конкурирующая гипотезы, ошибки I и II рода, уровень значимости, критическая область.

На практике часто приходится по результатам обследований, испытаний, опытов проверять различные предположения о характеристиках исследуемого объекта. В этом случае решение об истинности предположения принимается путем проверки статистической гипотезы.

Статистической гипотезой, обозначаемой H, называется любое предположение относительно вида или параметра распределения случайной величины X.

Нулевой гипотезой называют проверяемую гипотезу H_0.

Альтернативной (конкурирующей) называют гипотезу H_1, которая противоречит нулевой.

Правило, по которому мы решаем принять или отклонить гипотезу H_0, называется *критерием К*.

Так как решение мы будем принимать на основе наблюдений с.в. X, то нам необходимо выбрать подходящую статистику, которая называется *статистикой Т критерия К*. Критерий, как статистическое правило, определяется заданием критической области – описанием тех данных, при которых мы отклоняем гипотезу H_0.

Критической областью называют множество всех значений статистики критерия, при которых нулевая гипотеза отклоняется. Критическая область изображена на рис. 1–3 (с. 41-42).

Эта область обычно задается с помощью статистики T и имеет вид неравенства $T(c)>c$, где c выбирается по заданному уровню значимости.

Положение критической области зависит от формулировки альтернативной гипотезы H_1. Предположим, что проверяется гипотеза H_0: $\theta = \theta_0$, а альтернативная гипотеза – H_1: $\theta \neq \theta_0$. Т.е. допускается, что

различие может быть в обе стороны. Поэтому такие критерии называются *двусторонними.* На рис. 1 изображена двусторонняя критическая область

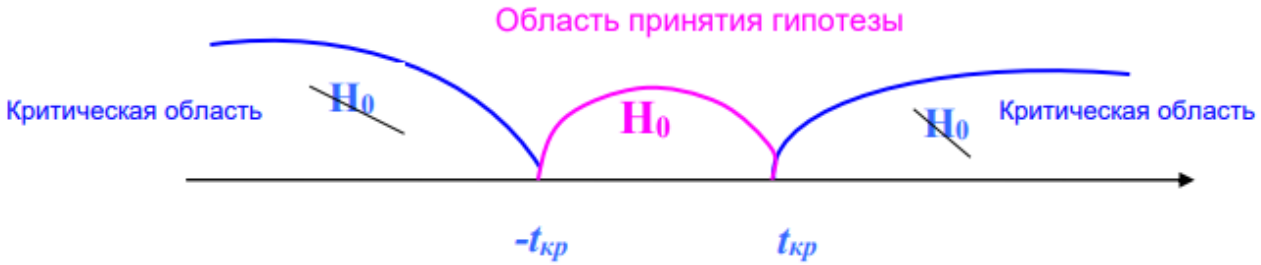


Рис. 1

В случае альтернативной гипотезы H_1: $\theta > \theta_0$ предполагается, что значение θ может быть расположено на числовой оси правее значения θ_0. Поэтому критерий называется односторонним, а точнее – правосторонним. Его критическая область изображена на рис. 2.

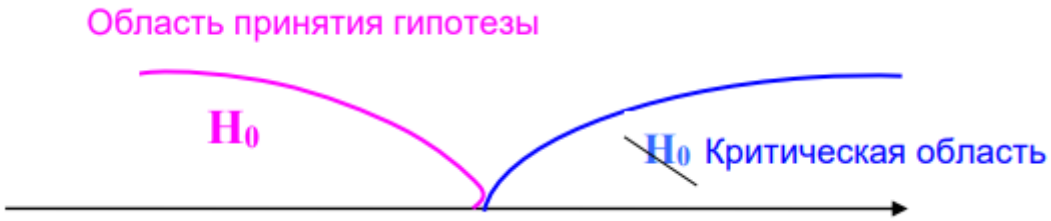


Рис. 2

Другим вариантом одностороннего критерия является левосторонний критерий: H_1: $\theta < \theta_0$. На рис. 3 (с. 42) изображена левосторонняя критическая область.

При проверке гипотез возможны два типа ошибок:

1) отклонение гипотезы H_0, когда она верна – *ошибка первого рода*;

2) принятие гипотезы H_0, когда она ложна – *ошибка второго рода*.

Заданные ограничения на вероятность ошибки первого рода обозначаются α и называются *уровнем значимости критерия*.

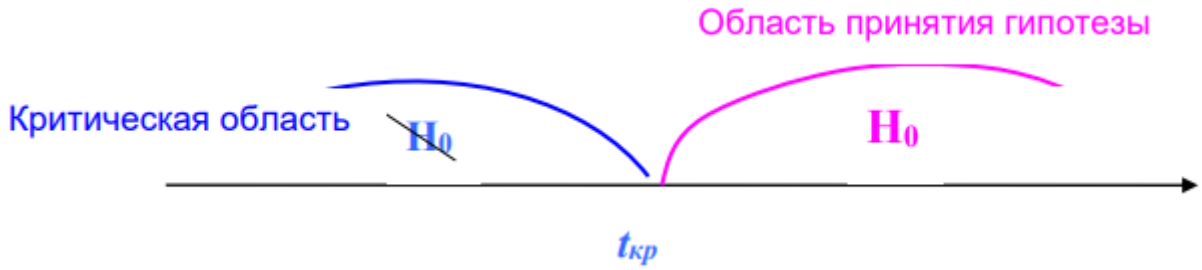


Рис. 3

Вероятность ошибки второго рода обозначается β. Вероятности α и β удобно представить, как это сделано в следующей таблице.

Ошибки при проверке гипотез

	Решение	
	Принять H_0	Принять H_1
Справедлива H_0	Правильное с вероятностью 1 – α	Ошибочное с вероятностью α
Справедлива H_1	Ошибочное с вероятностью β	Правильное с вероятностью 1 – β

Вероятность (1 – β) не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0, когда она неверна, называется *мощностью критерия*.

Следует особо подчеркнуть, что любая гипотеза должна формулироваться, уровень значимости α задаваться, а критерий выбираться исследователем всегда до получения экспериментальных данных, по которым эта гипотеза будет проверяться.

Проверка статистических гипотез может быть разбита на ряд этапов:

1. Сформулировать нулевую и альтернативную гипотезы.
   Назначить уровень значимости α.
   Выбрать критерий.
2. Получить экспериментальные данные.

3. Вычислить статистику (например, T) данного критерия.

4. Принять статистическое решение. В зависимости от имеющихся таблиц мы будем использовать один из следующих способов.

Способ 1. По таблице находим границы критической области.

1) Если значение статистики T попадает в область принятия гипотезы, то принимается гипотеза H_0. То есть считается, что гипотеза H_0 не противоречит результатам наблюдений.

2) Если значение статистики попадает в критическую область, то гипотеза H_0 отклоняется как не согласующаяся с результатами наблюдений (рис. 4).

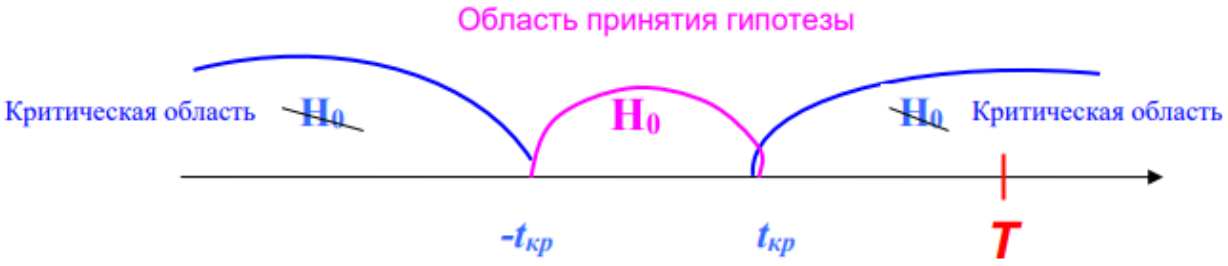


Рис. 4

Способ 2. По таблице находится критический уровень значимости $\alpha_{кр}$. Гипотеза отвергается, если критический уровень значимости $\alpha_{кр}$ меньше заданного α. Способ 2 обладает большей информативностью.

3.1. Проверка гипотезы о величине среднего значения нормального распределения (одновыборочный критерий Стьюдента)

Условия применения t-критерия: случайная выборка $X_1,..,X_n$ получена из нормального распределения с неизвестным средним μ и неизвестной дисперсией σ^2.

Гипотеза H_0 : $\mu = \mu_0$

Альтернатива H_1 : $\mu \neq \mu_0$ или H_1 : $\mu > \mu_0$ (H_1 : $\mu < \mu_0$)

Уровень значимости: α.

Порядок применения:

1. Принимается предположение о нормальности, формулируется гипотеза H_0 и альтернатива H_1, задается уровень значимости α .

2. Получают выборку $X_1,..,X_n$ объемом n.

3. Вычисляется статистика Стьюдента:

$$T = \frac{\bar{x} - \mu_0}{S}\sqrt{n-1},$$

где $\bar{x}, S_x^2$ – выборочные среднее и дисперсия выборки $X_1,..,X_n$,.

Величина T при справедливости нулевой гипотезы имеет t-распределение Стьюдента с $\nu = n - 1$ степенями свободы.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) Если H_1 : $\mu \neq \mu_0$, то $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Значение $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α, помещенному в верхней строке таблицы (для двусторонней критической области). Если $|T| > t_{кр}$, то гипотеза H_0 отвергается.

б) Если H_1 : $\mu > \mu_0$, то $t_{кр} = t_{прав\ кр}(\alpha; \nu)$. Если $T > t_{кр}$, то гипотеза H_0 отвергается.

в) Если H_1 : $\mu < \mu_0$, то $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$. Если $T < t_{лев\ кр}$, то гипотеза H_0 отвергается.

Значения $t_{кр} = t_{прав\ кр}(\alpha; \nu)$, $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости α, помещенному в нижней строке таблицы (для односторонней критической области).

Вывод: если гипотеза H_0 отвергается, то выборочное среднее значимо отличается от μ_0 на уровне значимости α. В противном случае различие не является статистически значимым.

▶ **Пример.** По результатам работы последних месяцев средний чек в кофейне составлял 150 д.е. С целью повышения среднего чека был использован метод upselling (предложение улучшенной или более дорогой версии заказа). В частности, посетителям предлагался кофе с добавками (сироп, взбитые сливки, кокосовое молоко и т.п.). Случайным образом выбраны 18 чеков, среднее значение чека которых составило 158,2 д.е., а выборочное стандартное отклонение – 18,4 д.е. На уровне значимости 0,01 проверьте предположение об эффективности предпринятых мер, направленных на увеличение среднего чека в указанной кофейне.

Решение. Используем *t*-критерий и действуем в указанном порядке.

1. Предполагаем, что распределение расхода материала приближенно нормальное.

Гипотеза H_0 : μ = 150.

Альтернатива H_1 : μ > 150.

Уровень значимости: α = 0,01.

2.-3. Из условия задачи известно: n = 18, $\bar{x} = 158,2$; $S_x = 18,4$.

$$t = \frac{\bar{x} - \mu_0}{S}\sqrt{n-1} = \frac{158,2-150}{18,4}\sqrt{18-1} = 1,837$$

и ν = n – 1 = 18 – 1 = 17.

4. Находим по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости α = 0,01, помещенному в нижней строке таблицы (для односторонней критической области), и ν =17 значение $t_{кр} = t_{прав\ кр}\,(\alpha; \nu) = t_{прав\ кр}\,(0,01; 17) = 2,57$. Т.е. критическая область есть (2,57; +∞). На рис. 1 (с. 46) изображена критическая область.

Так как | t | < $t_{кр}$ (1,837 < 2,57), то гипотеза H_0 принимается.

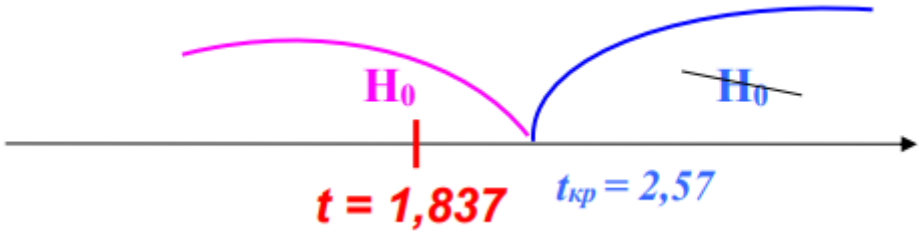


Рис. 1

Вывод: имеющиеся данные не подтверждают эффективность предпринятых мер по увеличению значения среднего чека. ■

Задачи к §3.1

В задачах данного раздела предполагается, что выборки имеют нормальное распределение, либо достаточно близкое к нормальному.

1. 22 сентября – Международный день без автомобиля. Ежедневная норма выручки для автобуса некоторого маршрута составляет 79 тыс д.е. Средняя выручка 22 сентября 20** г. составила 82 тыс д.е. при выборочной дисперсии – 17. На маршруте работало 8 автобусов. Можно ли утверждать, что 22 сентября 20** г. выручка повысилась? Уровень значимости принять 0,05.

2. По выборке объема $n = 24$, извлеченной из нормальной совокупности, были найдены выборочное среднее $\bar{x} = 17{,}5$ и выборочное стандартное отклонение $S = 4{,}2$. Необходимо при уровне значимости 0,05 проверить гипотезу $H_0 : \mu = 16$ при альтернативе а) $H_1 : \mu \neq 16$; б) $H_1 : \mu > 16$.

3. В откормочном цехе фермерского хозяйства имеется 20 бычков, вес которых распределяется следующим образом:

Вес бычков, кг	598-603	603-608	608-613	613-618
Количество	5	4	10	1

Проверьте гипотезу о том, что средний вес бычка соответствует установленной для этого возраста норме − 610 кг. Уровень значимости предлагается взять 0,05.

4. Согласно ГОСТу «Шоколад. Общие технические требования» горький (черный) шоколад должен содержать 33% и более какао-масло. Для проверки качества были взяты 11 образцов одного производителя. В приведенной ниже таблице указано содержание какао-масло (гр) на 100 гр продукции.

	Номер образца										
	1	2	3	4	5	6	7	8	9	10	11
Содержание какао-масло (гр)	30	32	34	33	36	34	29	31	37	32	36

Соответствует ли ГОСТу указанная марка шоколада?

5. Затраты времени на сборку прибора у 70 работников цеха представлены в следующей таблице.

Время, мин	49-52	52-55	55-58	58-61	61-64
Число сборщиков, чел.	12	13	25	11	9

Проверьте гипотезу о том, что средние затраты времени на сборку прибора не превышают норматив − 1 час. Уровень значимости принять 0,05.

6. Норма расхода дизельного топлива трактором Т-180 за час работы составляет 13 кг. Фактический расход соответствовал норме. Условия строительства несколько усложнились. Для проверки обоснованности нормы в новых условиях было отобрано в случайном порядке 8 тракторов. Выборочная средняя составила 14,6 кг при $S^2 = 5{,}12$ кг^2. Проверьте при уровне значимости 0,01 необходимость пересмотра нормы расхода дизельного топлива.

7. Из продукции автомата, обрабатывающего болты с номинальным значением контролируемого размера 40 мм, была взята выборка объема 36. Выборочное среднее контролируемого размера составило 40,2 мм с выборочной дисперсией 1,2 мм^2. Можно ли по результатам проведенного выборочного обследования утверждать, что контролируемый размер в продукции автомата не имеет положительного смещения по отношению к номинальному размеру? Принять $\alpha = 0{,}01$.

8. Для выявления среднего времени простоя одного ткацкого станка было взято на выборку 27 станков. Оказалось, что в среднем станок простаивает 25,7 мин со стандартным отклонением 5 мин. Свидетельствуют ли эти данные о том, что простои в указанный период увеличились, если за норму принять среднее значение простоя ткацкого станка за предыдущий год – 24,1 мин. Проверьте предположение при уровне значимости 0,05.

9. В период подготовки к экзамену студенты (группа девушек) были протестированы по тесту Спилберга на уровень реактивной тревожности. С целью снижения тревожности в группе был проведен тренинг по новой методике и студентки еще раз были протестированы по тесту Спилберга (в период подготовки к экзамену). Результаты представлены в следующей таблице.

Уровень реактивной тревожности	Испытуемые (код имени)							
	И.А.	К.Л.	В.Д.	К.О.	Н.В.	В.З.	Ф.О.	Ю.Ю.
До тренинга	32	34	28	43	35	26	41	32
После тренинга	26	29	31	35	29	26	31	24

Подтверждают ли результаты тестирования эффективность новой методики? Проверьте это предположение на уровне значимости 0,05.

(*Указание.* Предлагается создать новую выборку $z_i = x_i - y_i$, где x_i и y_i – уровни тревожности у девушки до и после тренинга соответственно. Далее проверить гипотезу $H_0 : \mu_z = 0$, которая означает, что никаких статистически значимых изменений не произошло.)

3.2. Проверка гипотезы о равенстве средних значений двух нормальных распределений (двухвыборочный критерий Стьюдента)

Условия применения t-критерия: случайные выборки $X_1,..,X_n$ и $Y_1,..,Y_m$ имеют нормальное распределение с неизвестными средними μ_x и μ_y, общей неизвестной дисперсией σ^2; n, $m < 30$.

Гипотеза $H_0 : \mu_x = \mu_y$.

Альтернатива $H_1 : \mu_x \neq \mu_y$ или $H_1 : \mu_x > \mu_y$ $(\mu_x < \mu_y)$ в зависимости от того, что требуется доказать: простое различие средних или то, что одно из них больше другого.

Уровень значимости: α.

Порядок применения:

1. Принимается предположение о нормальности, формулируется гипотеза H_0 и альтернатива H_1, задается уровень значимости α .

2. Получают две независимые выборки $X_1,..,X_n$ и $Y_1,..,Y_m$ объемами n и m соответственно.

3. Вычисляется статистика Стьюдента:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где $\bar{x}, S_x^2$ – выборочные среднее и дисперсия выборки $X_1,..,X_n$,

а $\bar{y}, S_y^2$ – второй выборки $Y_1,..,Y_m$.

Величина T при справедливости нулевой гипотезы имеет t-распределение Стьюдента с $\nu = n + m - 2$ степенями свободы.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) $H_1 : \mu_x \neq \mu_y$ $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Значение $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распре-

деления Стьюдента» по заданному уровню значимости α, помещенному в верхней строке таблицы (для двусторонней критической области). Если $|T| > t_{кр}$, то гипотеза H_0 отвергается.

б) $H_1 : \mu_x > \mu_y$ $t_{кр} = t_{прав\ кр}(\alpha; \nu)$. Если $T > t_{кр}$, то гипотеза H_0 отвергается.

в) $H_1 : \mu_x < \mu_y$ $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$. Если $T < t_{лев\ кр}$, то гипотеза H_0 отвергается.

Значения $t_{кр} = t_{прав\ кр}(\alpha; \nu)$, $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости α, помещенному в нижней строке таблицы (для односторонней критической области).

Вывод: если гипотеза H_0 отвергается, то выборочные средние значимо различаются на уровне значимости α. В противном случае различие не является статистически значимым.

Пример. В магазинах сети «AAA» средняя цена на яблоки сорта «Гала» (1 кг) составила 76,4 д.е., а в магазинах сети «BBB» – 85,1 д.е. В исследовании участвовало 15 магазинов «AAA» и 14 магазинов «BBB» одного района города. Выборочные стандартные отклонения равны 15,2 д.е. («AAA») и 11,6 д.е. («BBB»). Можно ли утверждать, что цены на яблоки в указанных сетевых магазинах значительно отличаются при уровне значимости 0,05?

Решение.

Обозначим с.в. X – цена на яблоки (1 кг) в магазинах сети «AAA»;

с.в. Y – цена на яблоки (1 кг) в магазинах сети «BBB».

Используем *t*-критерий и действуем в указанном порядке.

1. Предполагаем, что цена на яблоки имеет приближенно нормальное распределение.

Гипотеза $H_0 : \mu_x = \mu_y$.

Альтернатива $H_1 : \mu_x \neq \mu_y$.

Уровень значимости: $\alpha = 0{,}05$.

2. Из условия задачи известно:

$$\bar{x} = 76,4; \quad S_x = 15,2; \quad n = 15,$$

$$\bar{y} = 85,1; \quad S_y = 11,6; \quad m = 14.$$

3. Вычисляем статистику T:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} =$$

$$= \frac{76,4 - 85,1}{\sqrt{15 \cdot 15,2^2 + 14 \cdot 11,6^2}} \sqrt{\frac{15 \cdot 14(15+14-2)}{15+14}} = -1,66,$$

и $\nu = 15 + 14 - 2 = 27$.

4. Находим $t_{кр}$ по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости $\alpha = 0,01$, помещенному в верхней строке таблицы (для двусторонней критической области): $t_{кр} = t_{двуст\ кр}(\alpha; \nu) = t_{двуст\ кр}(0,05; 27) = 2,05$.

Критическая область имеет следующий вид.

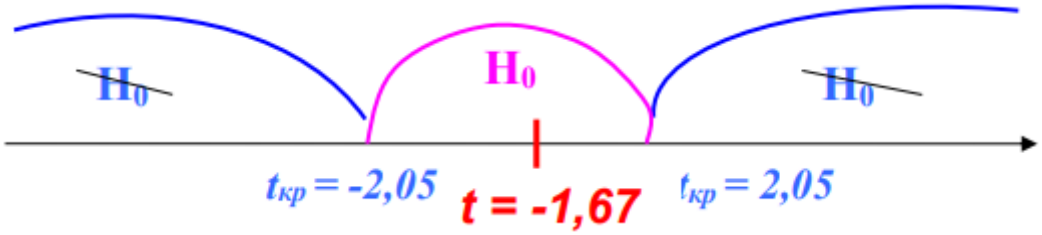


Рис. 1

Поскольку $|t| < t_{кр}$ ($|-1,67| < 2,05$), то гипотеза H_0 принимается.

Вывод: на основе имеющихся данных не выявлено статистически значимых различий между ценами на яблоки в указанных сетевых магазинах при уровне значимости 0,05. ■

Задачи к §3.2

В задачах данного раздела предполагается, что выборки имеют нормальное распределение, либо достаточно близкое к нормальному.

1. С целью повышения производительности труда для бригады (16 человек) были организованы курсы повышения квалификации. Через месяц после окончания курсов были произведены замеры: среднее время обработки одной детали составило 12,4 мин., выборочное стандартное отклонение – 5,4 мин. Аналогичные замеры для работников (19 человек), не проходивших повышение квалификации: среднее время обработки одной детали составило 16,8 мин., выборочное стандартное отклонение – 7,2 мин. Какой вывод можно сделать об эффективности курсов повышения квалификации?

2. По двум независимым выборкам объема $n = 11$ и $m = 18$, извлеченных из нормальной совокупности, были найдены выборочные средние $\overline{x} = 42$, $\overline{y} = 45$ и выборочные стандартные отклонения $S_x = 3{,}78$ и $S_y = 2{,}91$. Необходимо проверить гипотезу $H_0 : \mu_x = \mu_y$ при альтернативе $H_1 : \mu_x \neq \mu_y$ на уровне значимости а) 0,05; б) 0,01.

3. По результатам опроса на тему «Какую сумму Вы планируете потратить на новогодние подарки?» получены следующие данные.

	Возраст респондентов			
	18-24	26-46	47-60	60+
Среднее значение (д.е.)	875	1200	1500	937
Выборочное стандартное отклонение (д.е.)	87	116	94	62
Количество опрошенных, чел.	25	19	14	27

Можно ли утверждать, что существуют статистически значимые различия в расходах на подарки у разных возрастных групп?
(*Указание*. Проверьте это предположение, сравнивая попарно данные по разным возрастным группам. Сколько статистических гипотез необходимо проверить?)

4. В цехе работают 2 линии по выпуску керамической плитки. У сотрудника отдела контроля качества возникло предположение, что

линии выпускают неодинаковую по качеству продукцию. Для проверки случайным образом с линии А было взято 8 плиток, а с линии В – 6 плиток и сделаны замеры по толщине. Средняя толщина для линии А оказалась 7,42 мм и 0,057 мм^2, а для линии В – 7,63 мм и 0,0457 мм^2. Допуская риск 1%, можно ли утверждать, что линии выпускают одинаковую продукцию?

5. В сети кофеен решили повысить средний чек с помощью предложений дополнительных продаж. Была использована методика Cross selling (предложение сопутствующих товаров). Для эксперимента были выбраны 2 кофейни с одинаковым средним чеком – 300,4 д.е. В кофейне №1, в которой к каждому напитку предлагали десерт со скидкой 10%, средний чек составил 336,5 д.е., а в другой кофейне №2 со скидкой 15% средний чек составил 345,8 д.е. В кофейне №1 на проверку случайным образом взяли 24 чека, а в кофейне №2 – 19 чеков. Выборочные дисперсии для кофеен №1и №2 составили соответственно 186 (д.е.)^2 и 165 (д.е.)^2.

Свидетельствуют ли полученные данные о лучшем результате при скидке 15%? Действительно ли предпринятые меры способствовали повышению среднего чека для каждой из кофеен?

6. Средний недельный объем продаж для 15 торговых точек района А составил 3100 тыс. д.е. при стандартном отклонении 400 тыс.д.е., а для 10 торговых точек района В – 2600 тыс. д.е. при стандартном отклонении 500 тыс. д.е. Значимо ли различие средних недельных объемов продаж в районах А и В при 5 %-м уровне значимости? Значимо ли превышение среднего недельного объема продаж в районе А по сравнению с районом В? Чем отличается этот вопрос от предыдущего?

7. Для испытания шерстяной материи на прочность произведены две выборки объемов в 10 и 12 образцов. Средняя прочность оказалась равной 135 г и 136 г при выборочных дисперсиях 4 и 6. Определить при уровне значимости 0,01 существенность расхождения в обеих выборках.

8. Для сравнения предлагаемого нового сорта пшеницы В со старым А под каждый сорт было отведено по 3 участка. В результате был получен урожай, в пересчете в центнерах на гектар, равный:

Сорт А – 15,6; 18,2; 13,3.

Сорт В – 17,1; 16,3; 18,8.

Определить, можно ли, допуская риск 5%, считать, что сорт В урожайнее сорта А?

9. Взято на выборку несколько рыб (севрюга) в осенние и весенние уловы и измерена их длина. Результаты даны в таблице:

Длина рыбы (севрюги) в см.

Весенний улов	81,5	129,4	115,3	148,1	173,6	76,8	102,3	-
Осенний улов	79,1	105,3	181,2	120,3	143,4	165,8	153,8	178,3

Выявить при уровне значимости 0,01, является ли расхождение между длиной самцов севрюги существенным, т.е. влияет ли сезон улова на длину?

10. На двух лесных участках взяты на выборку по 30 деревьев для определения их среднего диаметра. Они оказались равными 31 см и 32 см, их выборочные дисперсии равны 40 и 60 соответственно. С вероятностью 0,95 оценить расхождение между средними диаметрами деревьев**.**

3.3. Проверка гипотезы о равенстве дисперсий двух нормальных распределений (*F*–критерий Фишера)

Условия применения F–критерия: $X_1,..,X_n$ и $Y_1,..,Y_m$ – две независимые случайные выборки (n, $m < 100$) из нормальных распределений с дисперсиями σ_x^2 и σ_y^2 и неизвестными средними μ_x и μ_y .

Гипотеза $H_0 : \sigma_x^2 = \sigma_y^2$.

Альтернатива $H_1 : \sigma_x^2 \neq \sigma_y^2$. Если предположить, что одна из выборок имеет большую дисперсию (обозначим ее σ_1^2), чем другая (σ_2^2), то можно сформулировать одностороннюю гипотезу $H_1 : \sigma_1^2 > \sigma_2^2$.

Уровень значимости: α.

Порядок применения:

1. Принимается предположение о нормальности распределения выборок, формулируется гипотеза H_0 и альтернатива H_1, задается уровень значимости α.

2. Получают две независимые выборки $X_1,..,X_n$ и $Y_1,..,Y_m$ объемами n и m соответственно.

3. Рассчитывают значения несмещенных выборочных дисперсий S_x^2 и S_y^2. Большую из дисперсий (S_x^2 или S_y^2) обозначают S_1^2, а меньшую дисперсию – S_2^2.

Вычисляется значение *F*–критерия по формуле

$$F = \frac{S_1^2}{S_2^2},$$

$\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$, где ν_1 – число степеней свободы большей дисперсии S_1^2.

4. По таблице «Процентные точки распределения Фишера-Снедекора (F-распределения)» (приложение, таблица IV) находят критическое

значение F-критерия при заданном уровне значимости α и числе степеней свободы $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$.

При альтернативе $H_1 : \sigma_x^2 \neq \sigma_y^2$ критическую точку $F_{кр}(\frac{\alpha}{2}, \nu_1, \nu_2)$ находят по уровню значимости $\frac{\alpha}{2}$ и числе степеней свободы ν_1, ν_2.

При альтернативе $H_1 : \sigma_x^2 > \sigma_y^2$ находят критическую точку $F_{кр}(\alpha, \nu_1, \nu_2)$.

Если $F < F_{кр}$, то нет оснований отвергнуть нулевую гипотезу. То есть, различие дисперсий не является статистически значимым.

▶ **Пример.** Для повышения урожайности (увеличение веса 1 яблока) была разработана новая схема весенней подкормки деревьев. На контрольном участке подкормка деревьев производилось по традиционной схеме. Случайным образом на контрольном участке 29 яблок, а на экспериментальном – 24. Вариация (выборочная дисперсия) массы 1 яблока составили; на контрольном участке – 85,4 гр^2, а на экспериментальном – 64,7 гр^2 . На уровне значимости 0,05 проверьте предположение о том, что разброс массы 1 яблока (вариация) на экспериментальном участке значительно ниже, чем на контрольном.

Решение.

с. в. X – масса 1 яблока (гр), собранного на контрольном участке,

с. в. Y – масса 1 яблока (гр), собранного на экспериментальном участке.

1. Гипотеза $H_0 : \sigma_x^2 = \sigma_y^2$.

Альтернатива $H_1 : \sigma_x^2 > \sigma_y^2$, т.к., согласно предположению, разброс значений массы 1 яблока в экспериментальной группе (Y) меньше, чем в контрольной (X).

Уровень значимости: $\alpha = 0{,}05$.

Принимаем предположение о нормальности распределения обеих выборок.

2.-3. По формулам (§§1.2.1, 1.2.2) вычисляем выборочные дисперсии. Исправленные (несмещенные) дисперсии найдем по формуле (§2): $S^2_{несм} = \frac{n}{n-1} S^2$.

Контрольный участок: $n_x = 25$, $S^2_x = 85{,}4$, $S^2_{y,несм} = \frac{25}{25-1} 85{,}4 = 88{,}96$,

Экспериментальный участок: $n_y = 29$, $S^2_y = 64{,}7$,

$$S^2_{y,несм} = \frac{29}{29-1} 64{,}7 = 67{,}01.$$

Обозначим

$$S^2_1 = \max(S^2_{x,несм}, S^2_{y,несм}) = \max(88{,}96; 67{,}01) = 88{,}96 = S^2_{x,несм}.$$

Тогда $S^2_2 = S^2_{y,несм} = 67{,}01$.

Вычисляем значение *F*-критерия

$$f = \frac{S^2_1}{S^2_2} = \frac{88{,}96}{67{,}01} = 1{,}327,$$

$\nu_1 = n_1 - 1 = n_x - 1 = 25 - 1 = 24$ и $\nu_2 = n_2 - 1 = n_y - 1 = 29 - 1 = 28$.

4. Найдем границы критической области по таблице «Процентные точки распределения Фишера-Снедекора (F-распределение)» (приложение, таблица IV). Так как α = 0,05, то используем таблицу «F-распределение: верхние 5%-е точки» и находим $F_{кр}(\alpha; \nu_1; \nu_2) = F_{кр}(0{,}05; 24; 28) = 1{,}91$. Критическая область изображена на рис.1.

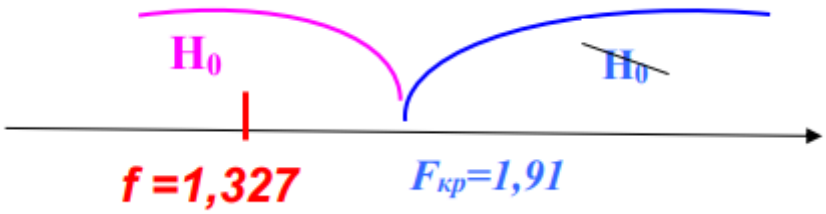


Рис. 1

Поскольку $f < F_{кр}$ (1,327 < 1,91), то гипотеза H_0 принимается. Следовательно, на уровне значимости 0,05 различие дисперсий не является статистически значимым.

Вывод: полученные данные не подтверждают предположения о том, том, что вариация массы 1 яблока на экспериментальном участке значительно ниже, чем на контрольном. ■

Задачи к §3.3

В задачах данного раздела предполагается, что выборки имеют нормальное распределение, либо достаточно близкое к нормальному.

1. До наладки станка была проверена точность изготовления 10 втулок и найдено значение оценки дисперсии диаметра 9,6. После наладки подверглись контролю еще 15 втулок и получено новое значение оценки дисперсии – 5,7. Можно ли считать, что в результате наладки станка точность изготовления деталей увеличится? Принять α = 0,05.

2. По двум независимым выборкам объема n = 13 и m = 16, извлеченных из нормальной совокупности, были найдены несмещенные выборочные дисперсии S_x^2 = 25,6 и S_y^2 = 8,94. Необходимо проверить гипотезу $H_0 : \sigma_x^2 = \sigma_y^2$ при альтернативе $H_1 : \sigma_x^2 > \sigma_y^2$ на уровне значимости а) 0,05; б) 0,01.

3. Для исследования двух типов удобрений на урожайность пшеницы было засеяно по 15 опытных участков. Несмещенные выборочные дисперсии, характеризующие вариацию урожайности, соответственно, равны 0,25 и 0,49. Проверить, зависит ли вариация урожайности пшеницы от типа внесенных удобрений а) при α = 0,1; б) при α = 0,02.

4. Два токарных автомата изготовляют детали по одному чертежу. Из продукции первого станка было отобрано 9 деталей, а из продукции второго – 11 деталей. Выборочные дисперсии контрольного размера, определенные по этим выборкам, равны соответственно

5,9 мкм^2 и 23,3 мкм^2. Проверить гипотезу о равенстве дисперсий при $\alpha = 0{,}10$, если альтернативная гипотеза утверждает, что: а) дисперсии не равны; б) дисперсия размера для второго станка больше, чем для первого.

5. Определить, можно ли считать существенно различными режимы работ трактористов А и В, если у тракториста А при средней глубине вспашки в 21 см выборочное стандартное отклонение, определенное в результате 30 замеров, равнялось 2 см, а у тракториста В при той же средней глубине вспашки, определенное в результате 20 замеров, оказалось 3,9 см. Уровень значимости взять равным 0,10.

6. Для эксперимента были выбраны 2 группы учащихся, одна из которых имеет гуманитарную специализацию, а другая – по химии. Ниже представлены результаты контрольной работы по математике, максимальный балл равен 20.

Направление специализации	Оценка за контрольную работу, балл								
гуманитарное	8	15	18	9	12	17	6	11	17
химическое	15	12	17	9	10	18	-	-	-

Значения средних баллов по обоим направлениям специализации не имеют статистически значимых различий. Можно ли сказать, что и по вариативности результатов контрольной работы учащиеся обоих направлений специализации также не отличаются?

7. Компания решила инвестировать средства в ценные бумаги компаний А и В. Выборка из 21 наблюдения по активу А показала, что выборочная дисперсия ее доходности равна 0,072, а по активу В – 0,068 при 16 наблюдениях. Эксперты предполагают, что бумаги компании А менее устойчивы к колебаниям рынка, поэтому инвестировать средства следует только в бумаги компании В. Проверьте предположение экспертов на уровне значимости 0,05.

8. Одни и те же изделия получают на двух производственных линиях. Качество изделия зависит от продолжительности процесса обработки сырья. На второй линии введены некоторые усовершенствования, сократившие вариацию времени обработки, в связи с чем качество изделий повысилось. Затем были проведены выборочные измерения вариации времени обработки на обеих линиях. В результате были получены следующие значения несмещенных выборочных дисперсий: $S_x^2 = 3.579 мин^2$ при 21 наблюдении и $S_y^2 = 2{,}5963 мин^2$ при 11 наблюдениях. Можно ли считать существенными расхождения между вариациями продолжительности процесса обработки сырья на первой и второй линиях? Проверить гипотезу при уровне значимости 0,10.

Контрольные вопросы

1. Как следует выбирать нулевую гипотезу?
2. Как следует выбирать альтернативную гипотезу?
3. Какие ошибки существуют при проверке гипотез?
4. Какое условие налагается на применение критерия Стьюдента?
5. Можно ли применять критерий Фишера как предварительный тест для проверки условий применимости критериев Стьюдента?
6. В каких случаях применяются односторонние и двусторонние критерии?
7. Какое условие налагается на применение критерия Фишера?
8. Какое распределение имеет статистика Фишера?
9. Какой уровень значимости лучше выбрать – 5% , 10% или 1%?

4. СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ СВЯЗИ

Ключевые слова: функциональная зависимость, статистическая зависимость, корреляционная зависимость, факторный признак, результативный признак

В исследованиях большое место занимает определение связи между признаками. При изучении зависимости между различными явлениями и их признаками различают функциональную и статистическую зависимости.

При *функциональной связи* значение одной величины однозначно определяется значением другой и выражается математической формулой. В реальной жизни такие зависимости чаще наблюдаются в физике, биологии, химии, геометрии и т.д. Например, связи между общим стажем работы и стажем работы на данном предприятии, связи между зарплатой и стажем работы на данном предприятии.

Но более широкое распространение в нашей жизни имеет *статистическая зависимость*. В этом случае при фиксированном значении одной величины другая имеет некоторую свободу и может принимать различные значения. Например, зависимость между стажем и производительностью труда, выпавшими осадками и полученным урожаем, ростом и весом и т.д.

Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой, называется *корреляционной.*

При анализе любого типа связи мы выделяем зависимые и независимые признаки.

Признаки, обуславливающие изменение других, связанных с ними признаков, называются *факторными (независимыми*).

Признаки, зависимые от факторных признаков, называются *результативными (зависимыми).*

Рассмотрим, например, зависимость между доходом и потреблением товаров. Очевидно, что независимым признаком является доход (*X*), а зависимым – потребление товаров (*Y*). То есть здесь связь «причина-следствие», $X \rightarrow Y$. Такой тип связи еще называют «направленной». В нашем примере с увеличением доходов снижается относительное потребление продовольственных товаров и увеличивается потребление непродовольственных товаров. Но следует помнить, что мы рассматриваем статистическую зависимость, и здесь не будет однозначной определенности, скорее мы определим общую тенденцию. В нашем примере возможна ситуация, когда рост доходов приведет к увеличению относительного потребления продовольственных товаров. Это может быть связано с разными привычками потребителей, неодинаковым ассортиментом продовольственных товаров, действием рекламы, новой информацией о здоровом образе жизни и т.д.

В случае *косвенной зависимости* признаки (*X* и *Y*) не имеют причинно-следственной связи, потому что оба являются следствием общей для них причины (признак *Z*, рис. 1).

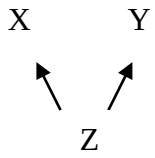


Рис. 1

Так в дореволюционной России была установлена тесная связь (корреляционная) между количеством пожарных команд в городе и числом пожаров. И пожары, и количеством пожарных команд в городе зависят от общего фактора – размеров города, его площади.

Следует упомянуть еще о *ложной связи*. Под ложной связью понимается чисто формальная связь между явлениями, не находящая никакого логического объяснения и основанная лишь на количественном соотношении между ними.

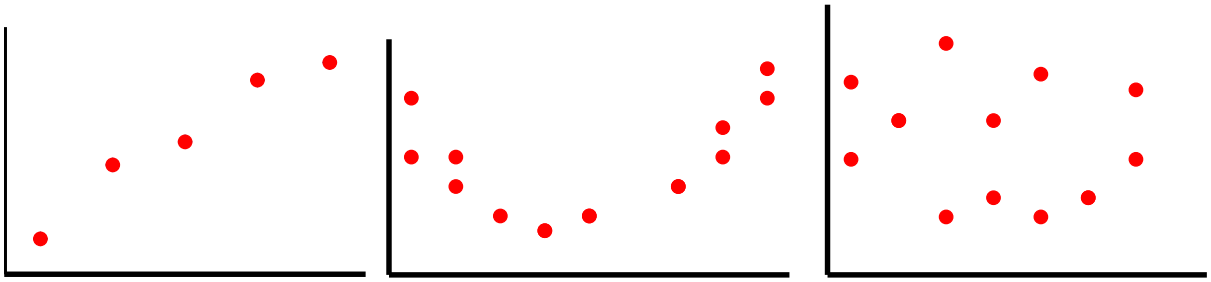
На сегодняшний день разработано большое число коэффициентов, которые определяют самые разнообразные виды связей практически для всех типов данных. Общее название этих коэффициентов – коэффициенты связи или меры связи. В пособии рассмотрена корреляционная связь.

4.1. Линейный коэффициент корреляции. Элементы регрессионного анализа

Ключевые слова: корреляционная зависимость, факторный признак, результативный признак, положительная корреляция, отрицательная корреляция, линия регрессии.

Корреляционная зависимость – это разновидность статистической зависимости, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой.

Для описания связи между случайными признаками X и Y также используется графическое представление данных. Каждую пару чисел $(x_1, y_1),..,(x_n, y_n)$, где x_i – i-ое наблюдение признака X, а y_i – i-ое наблюдение признака Y, которое соответствует значению x_i, можно рассматривать как координатные точки в прямоугольной системе координат. Тогда совокупность наблюдений, представленная таким образом, образует на плоскости скопление точек. Такого рода график называется *корреляционным полем* или *точечной диаграммой*. По конфигурации скопления точек мы можем сделать предварительные выводы о типе зависимости. По форме корреляционная связь бывает линейная и криволинейная. Например, если между точками можно провести прямую



а) б) в)

Рис. 1

линию так, что точки будут достаточно близки к ней, то речь идет о линейной зависимости (рис. 1.а). В случае рисунка 1.б зависимость наблюдается, но точно не линейная, а скорее похожая параболическую (криволинейная зависимость). Рисунок 1.в является примером того, что между признаками точно нет зависимости. В этом случае проводить дальнейшие исследования просто нет смысла.

Основная задача *корреляционного анализа* – определение степени и направления связи между признаками.

В данном разделе мы рассмотрим коэффициент корреляции, который показывает меру линейной зависимости между двумя признаками. Этот коэффициент еще называют коэффициентом корреляции Пирсона или линейным коэффициентом корреляции.

Выборочный коэффициент корреляции определяется по формуле:

$$r = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{S_x \cdot S_y}, \quad -1 \le r \le 1.$$

или

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{S_x \cdot S_y},$$

где $\bar{x} = \frac{1}{n}\sum x_i$ – выборочное среднее случайной величины X;

$S_x^2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$ – выборочная дисперсия случайной величины X.

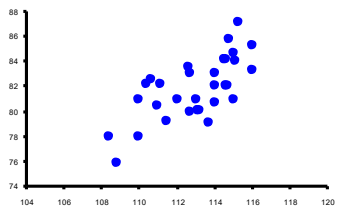
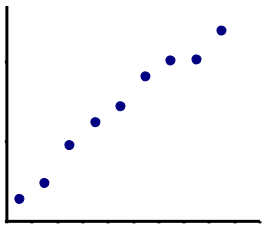
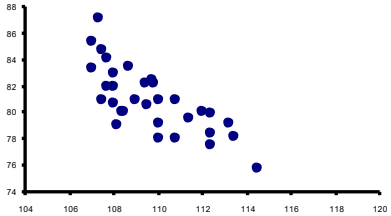
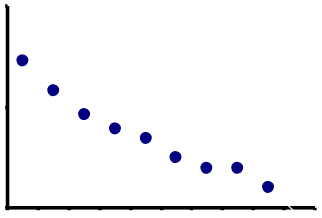
В случае, когда число наблюдений над признаками достаточно велико, для большей наглядности и упрощенности расчетов массив наблюдаемых значений подвергают группировке и представляют в виде корреляционной таблицы. Тогда выборочный коэффициент корреляции имеет следующий вид:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}(z_i - \bar{x})(t_j - \bar{y})}{S_x \cdot S_y},$$

где n_{ij} – частота попадания в i-й класс по признаку X и j-й класс по признаку Y, z_i и t_j – середины классов, k и m – число классов по признакам X и Y ссоответственно.

В таблице 1 представлены варианты корреляционного поля и их интерпретация.

Таблица 1

 $r > 0$ Слабая прямая связь	 $r \to 1$ Строгая прямая связь
 $r < 0$ Слабая обратная связь	 $r \to -1$ Строгая обратная связь

Обычно считают, что если $|r| \geq 0{,}7$, то связь сильная; при $0{,}5 < |r| < 0{,}7$ – средняя связь; при $|r| < 0{,}5$ – слабая связь. Следует отметить, что указанная интерпретация значений коэффициента корреляции достаточно условна. Сила связи определяется сущностью анализируемых явлений. Например, в социологии значения *r* редко достигают границ варьирования, и связь считается сильной, если $|r| > 0{,}3$.

Если коэффициент корреляции больше нуля ($r > 0$), то речь идет о *положительной корреляции (прямая связь)*. Это означает, что с увеличением одного признака значение другого признака имеет тенденцию к возрастанию. Например, зависимость уровня производительности труда от стажа работы.

Отрицательной корреляции (*обратной связи*) соответствует коэффициент, который меньше нуля ($r < 0$). В этом случае увеличение одного признака сопровождается убыванием другого. Например, если цены на потребительские товары растут, то спрос на них падает и наоборот.

Если признаки независимые, то коэффициент корреляции равен нулю. Обратное утверждение неверно. Если коэффициент равен нулю, то мы не можем сказать, что признаки независимы. Мы можем только утверждать, что между признаками нет корреляционной зависимости, но возможна другая зависимость, которую наш коэффициент не определяет.

Если $|r| = 1$, то между случайными величинами *X* и *Y* существует функциональная зависимость ($Y = aX + b$). И, следовательно, по значениям одного признака мы можем предсказывать значение другого.

Коэффициент корреляции Пирсона является симметричным коэффициентом $r_{xy} = r_{yx}$, т.е. он только указывает на степень связи между признаками. А вот что является причиной (независимый признак), а что – следствием (зависимый признак), приходится определять исходя из предыдущих исследований или обычного здравого смысла.

Пример 1. В таблице представлены данные о 5 однотипных предприятиях отрасли: выпуск продукции (X, тыс ед.) и расход условного топлива (Y, т.). Вычислить выборочный коэффициент корреляции, построить корреляционное поле.

X	1	1	2	3	6
Y	1	2	3	4	5

Решение. Чаще всего, чем больше выпуск продукции, тем выше расход условного топлива. Поэтому в нашей задаче выпуск продукции (X) – независимая переменная, а расход условного топлива (Y) – зависимая от X переменная.

Для построения корреляционного поля на координатную плоскость нанесем точки с координатами (1; 1), (1; 2), (2; 3), (3; 4), (6; 5) (рис. 2). По конфигурации скопления точек мы можем предположить, что связь между стажем и разрядом есть, причем, скорее всего эта связь линейная и положительная.

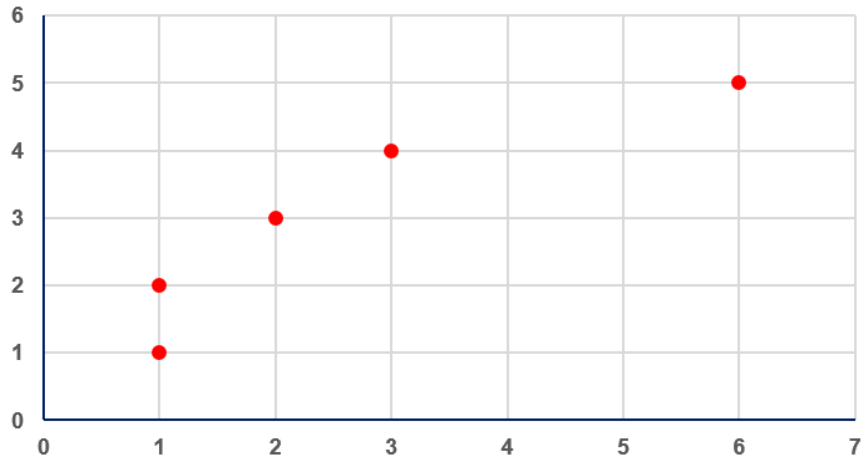


Рис. 2

Найдем выборочный коэффициент корреляции по следующей формуле: $r = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{S_x \cdot S_y}$.

Для удобства представим наши вычисления в виде следующей расчетной таблицы:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
1	2	1	4	2
2	3	4	9	6
3	4	9	16	12
6	5	36	25	30
$\sum = 13$	15	51	55	51

$n = 5$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\cdot 13 = 2{,}6;\ S_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}\right)^2 = \frac{1}{5}\cdot 51 - (2{,}6)^2 = 3{,}4;$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{5}\cdot 15 = 3,\ \ S_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\bar{y}\right)^2 = \frac{1}{5}\cdot 55 - 3^2 = 2,$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{S_x \cdot S_y} = \frac{\frac{1}{5}\cdot 51 - 2{,}6\cdot 3}{\sqrt{3{,}4\cdot 2}} = 0{,}92.$$

Полученное значение коэффициента $r = 0{,}92$ очень близко к 1, что указывает на сильную зависимость между выпуском продукции и расходом условного топлива. ■

Пример 2. Имеется следующая информация по однотипным предприятиям торговли о возрасте (продолжительности эксплуатации) типового оборудования (X, лет) и затратах (Y, тыс. д.е.) на их ремонт. Вычислить выборочный коэффициент корреляции.

Таблица 2

X \ Y	0–1,5	1,5–3,0	3,0–4,5	4,5–6,0	6,0–7,5
1–3	1	1	–	–	–
3–5	–	3	2	1	–
5–7	–	1	2	4	6
7–9	–	–	–	2	2

Решение. Признаки X и Y представлены в таблице 2 в виде интервальных рядов. Поэтому сначала мы находим середины классов $(\tilde{x}_i, \tilde{y}_j)$. Например, середина первого интервала по признаку Y будет $\tilde{y}_1 = \frac{0+1{,}5}{2} = 0{,}75$. В итоге получаем таблицу 3.

Таблица 3

X \ Y	0,75	2,25	3,75	5,25	6,75	n_X
2	1	1	-	-	-	2
4	-	3	2	1	-	6
6	-	1	2	4	6	13
8	-	-	-	2	2	4
n_Y	1	5	4	7	8	n = 25

В столбце с обозначением n_x записывается в i-й строке число наблюдений только признака X, которые принадлежат i-му интервалу без учета другого признака (§4.5). Аналогично и для строки n_Y. Рассмотрим признаки X и Y по отдельности.

1) Распределение признака X:

X	2	4	6	8
n_i	2	6	13	4

В строке n_i мы записываем значения столбца n_X.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \frac{2\cdot 2 + 4\cdot 6 + 6\cdot 13 + 8\cdot 4}{25} = \frac{138}{25} = 5{,}52,$$

$$\frac{1}{n}\sum x_i^2 n_i = \frac{4\cdot 2 + 16\cdot 6 + 36\cdot 13 + 64\cdot 4}{25} = \frac{828}{25} = 33{,}12,$$

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - (\bar{x})^2 = 33{,}12 - (5{,}52)^2 = 2{,}65,\ S_x = \sqrt{S_x^2} = \sqrt{2{,}65} = 1{,}63.$$

2) Распределение признака Y:

Y	0,75	2,25	3,75	5,25	6,75
n_i	1	5	4	7	8

В строке n_j мы записываем значения столбца n_Y.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i n_i = \frac{0{,}75\cdot 1 + 2{,}25\cdot 5 + 3{,}75\cdot 4 + 5{,}25\cdot 7 + 6{,}75\cdot 8}{25} = \frac{117{,}75}{25} = 4{,}71,$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 n_i = \frac{0{,}75^2\cdot 1 + 2{,}25^2\cdot 5 + 3{,}75^2\cdot 4 + 5{,}25^2\cdot 7 + 6{,}75^2\cdot 8}{25} = \frac{639{,}56}{25} = 25.58$$

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n} n_i y_i^2 - (\bar{y})^2 = 25{,}58 - (4{,}71)^2 = 3{,}4,\ \ S_y = \sqrt{S_y^2} = \sqrt{3{,}4} = 1{,}84.$$

3) Вычисляем выборочный коэффициент корреляции.

$$\sum_{i=1}^{4}\sum_{j=1}^{5} n_{ij} x_i y_j = 1\cdot 2\cdot 0{,}75 + 1\cdot 2\cdot 2{,}25 + 3\cdot 4\cdot 2{,}25 + 2\cdot 4\cdot 3{,}75 + 1\cdot 4\cdot 5{,}25 +$$

$$+1\cdot 6\cdot 2{,}25 + 2\cdot 6\cdot 3{,}75 + +4\cdot 6\cdot 5{,}25 + 6\cdot 6\cdot 6{,}75 + 2\cdot 8\cdot 5{,}25 + 2\cdot 8\cdot 6{,}75 = 703{,}5$$

$$r_{xy} = \frac{\frac{1}{n}\sum n_{ij} x_i y_j - \bar{x}\cdot\bar{y}}{S_x S_y} = \frac{\frac{1}{25}\cdot 703{,}5 - 5{,}52\cdot 4{,}71}{1{,}63\cdot 1{,}84} = 0{,}71.$$

Значение выборочного коэффициента корреляции свидетельствует о средней степени зависимости между возрастом (продолжительностью эксплуатации) типового оборудования и затратами на их ремонт. ■

Основываясь только на значении выборочного коэффициента корреляции, особенно если это значение не очень близко к ± 1, нельзя сделать вывод о достоверности корреляции между признаками. Этот вывод может быть сделан с помощью соответствующих критериев значимости корреляции. Мы рассмотрим *t*-критерий (Стьюдента).

Условия применения t-критерия: $(X_1, Y_1),..,(X_n, Y_n)$ – случайная выборка из двумерного нормального распределения со средними μ_x и μ_y, дисперсиями σ_x^2, σ_y^2 и коэффициентом корреляции ρ.

Гипотеза $H_0: \rho = 0$.

Альтернатива H_1: $\rho \neq 0$. Мы можем взять и одностороннюю альтернативу (H_1: $\rho > 0$ или H_1: $\rho < 0$), если знак корреляции может быть определен заранее.

Уровень значимости: α.

Порядок применения:

1. Принимается предположение о нормальности, формулируется гипотеза H_0 и альтернатива H_1, задается уровень значимости α.
2. Получают двумерную выборку объема n.
3. Вычисляют статистику Стьюдента:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

где r – выборочный коэффициент корреляции; n – объем выборки.

Статистика T при справедливости нулевой гипотезы имеет *t*-распределение Стьюдента с $\nu = n - 2$ степенями свободы.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) $H_1: \rho \neq 0$ $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Значение $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α, помещенному в верхней строке таблицы (для двусторонней критической области). Если $|T| > t_{кр}$, то гипотеза H_0 отклоняется.

б) $H_1: \rho > 0$ $t_{кр} = t_{прав\ кр}(\alpha; \nu)$. Если $T > t_{кр}$, то гипотеза H_0 отклоняется.

в) $H_1: \rho < 0$ $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$. Если $T < t_{лев\ кр}$, то гипотеза H_0 отклоняется.

Значения $t_{кр} = t_{прав\ кр}(\alpha; \nu)$, $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α, помещенному в нижней строке таблицы (для односторонней критической области).

Вывод «гипотеза H_0 отвергается» означает, что коэффициент корреляции статистически значимо отличается от нуля на уровне значимости α. То есть, подтверждается наличие связи.

Пример 2 (продолжение). Проверим гипотезу о значимости корреляции, используя коэффициент корреляции из примера 2.

Решение.

1. Принимаем предположение о нормальности двумерного распределения.

Гипотеза $H_0: \rho = 0$.

Альтернатива $H_1: \rho > 0$ (т.к. по нашим предположениям здесь существует прямая связь: повышение продолжительности эксплуатации типового оборудования ведет к увеличению затрат на их ремонт).

Уровень значимости: $\alpha = 0{,}05$.

2.-3. В Примере 2 мы уже вычислили выборочный коэффициент корреляции: $r = 0{,}71$; $n = 25$. Находим статистику Стьюдента

$$t = \frac{0{,}71\sqrt{25-2}}{\sqrt{1-0{,}71^2}} = 4{,}84,$$

$$\nu = n - 2 = 25 - 2 = 23.$$

4. Для односторонней критической области находим по таблице «Критические точки распределения Стьюдента» значение $t_{кр} = t_{прав\ ст\ кр}(\alpha; \nu) = t_{прав\ ст\ кр}(0{,}05; 23) = 1{,}71$. Получаем критическую область $(1{,}71; +\infty)$, которая изображена на рис.1.

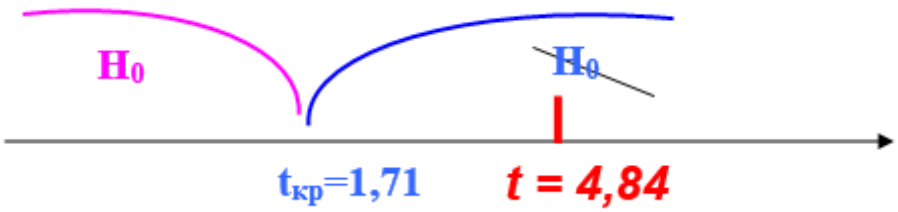


Рис. 1

Поскольку $t > t_{кр}$, т.е. значение статистики принадлежит критической области $(4{,}84 \in (1{,}71; +\infty))$, то гипотеза H_0 отвергается.

Вывод: между продолжительностью эксплуатации типового оборудования и затратах на их ремонт существует статистически значимая положительная связь ■

Как мы уже писали ранее, основной задачей корреляционного анализа является определение степени и направления связи между признаками. В регрессионном анализе также изучается связь между зависимой переменной и одной или несколькими независимыми переменными, но главным является установление формы зависимости. Поэтому основная задача *регрессионного анализа* – определение аналитического выражения связи между признаками.

Пусть переменная Y зависит от одной переменной X. При этом предполагается, что переменная X принимает заданные (фиксированные) значения, а зависимая переменная Y имеет случайный разброс из-

за ошибок измерения, влияния неучтенных факторов или других причин. В этом случае уравнение взаимосвязи (парная регрессионная модель) может быть представлена в виде

$$Y = f(x) + \varepsilon, \qquad (1)$$

где ε – случайная ошибка наблюдений.

Рассмотрим простейшую и наиболее часто используемую форму регрессии – парную линейную регрессию:

$$f(x) = \alpha + \beta x. \qquad (2)$$

Пусть проведено n независимых наблюдений с.в. Y при значениях переменной $X = x_1, x_2, .., x_n$. При этом измерения с.в. Y дали следующие результаты: $y_1, y_2, .., y_n$. Тогда линейная регрессионная модель имеет вид:

$$y_i = \alpha + \beta x_i + \varepsilon_i . \qquad (3)$$

Причем, $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$.

Оценкой модели (3) по выборке является уравнение регрессии:

$$\hat{y}_x = a + bx . \qquad (4)$$

Принято говорить: «регрессия Y на X».

Параметры линейного уравнения парной регрессии:

$$b = b_{Y|X} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x^2}, \qquad (5)$$

$$a = \bar{y} - b_{Y|X} \bar{x} . \qquad (6)$$

Параметр $b_{Y|X}$, т.е. коэффициент при x в уравнении линейной регрессии, называется коэффициентом регрессии (коэффициентом регрессии Y на X). Коэффициент регрессии показывает, на сколько (в абсолютном выражении) изменяется в среднем значение результативного признака Y при изменении факторного признака X на единицу.

Если $b_{Y|X} > 0$, то зависимость прямая,

если $b_{Y|X} < 0$, то зависимость обратная.

До сих пор мы предполагали, что переменная Y зависит от переменной X. Рассмотрим противоположную ситуацию, когда переменная Y – независимая, а переменная X зависит от Y. Тогда уравнение регрессии X на Y:

$$\hat{x}_y = a + by,$$

$$b = b_{X|Y} = \frac{\frac{1}{n}\sum xy - \bar{x}\cdot\bar{y}}{S_y^2},\ a = \bar{x} - b_{X|Y}\bar{y}.$$

Между коэффициентом корреляции r и коэффициентами регрессии $b_{Y|X}$ и $b_{X|Y}$ существует взаимосвязь:

$$r = \sqrt{b_{Y|X}\cdot b_{X|Y}}.$$

Поэтому коэффициенты регрессии могут быть вычислены через коэффициент корреляции:

$$b_{Y|X} = r\frac{S_y}{S_x},\ b_{X|Y} = r\frac{S_x}{S_y}.$$

▶ **Пример (продолжение Примера 1).** В таблице представлены данные о 5 однотипных предприятиях отрасли: выпуск продукции (X, тыс ед.) и расход условного топлива (Y, т.). По приведенным данным а) найти уравнение регрессии Y на X; б) построить линию регрессии Y на X; в) оценить средний расход условного топлива при выпуске продукции 7 тыс ед.

X	1	1	2	3	6
Y	1	2	3	4	5

Решение. Признак X (выпуск продукции) – независимый, а признак Y (расход условного топлива) – зависимый от X признак.

а) Предварительный анализ корреляционного поля (таблица 1, с. 65) позволяет предположить, что зависимость между стажем и разрядом является линейной. Тогда

$$\hat{y}_x = a + b_{Y|X}\, x.$$

Параметры линейного уравнения парной регрессии будем находить по формулам (5), (6). Для расчета параметров воспользуемся вспомогательной таблицей (с. 68), уже рассмотренной нами в Примере 1. Добавим в таблицу еще одну строку – средние значения по каждому столбцу.

$n = 5$.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
1	2	1	4	2
2	3	4	9	6
3	4	9	16	12
6	5	36	25	30
$\sum = 13$	15	51	55	51
$\frac{\Sigma}{n} = 2{,}6$	3	10,2	11	10,2

$$b = b_{Y|X} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{10{,}2 - 2{,}6 \cdot 3}{10{,}2 - (2{,}6)^2} = 0{,}7$$

$$a = \bar{y} - b_{Y|X}\, \bar{x} = 3 - 0{,}7 \cdot 2{,}6 = 1{,}19.$$

Значение $b_{Y|X}$ =0,7 показывает, что с увеличением продукции на 1 тыс ед. условный расход топлива в среднем возрастает на 0,7.

$$\hat{y}_x = a + b_{Y|X}\, x = 1{,}19 + 0{,}7x.$$

Это уравнение характеризует зависимость среднего значения разряда рабочего от его стажа.

б) Построим линию регрессии Y на X:

$\hat{y}_x = 1{,}19 + 0{,}7x.$

Учитывая, что график прямой строится по любым двум точкам, то при $x_1 = 1$ вычислим значение $\hat{y}_1 = 1{,}19 + 0{,}7 \cdot 1 = 1{,}89;$

$x_5 = 6$ значение $\hat{y}_5 = 1{,}19 + 0{,}7 \cdot 6 = 5{,}39.$

Полученный график линии регрессии Y на X изображен на рис. 3.

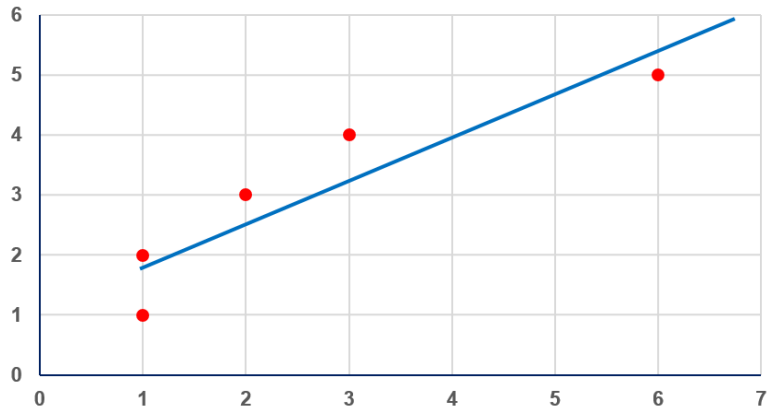


Рис. 3

в) Найдем оценку среднего расхода условного топлива при выпуске продукции 7 тыс ед. В уравнение линии регрессии Y на X подставим значение стажа $x = 7$:

$\hat{y}_x = 1{,}19 + 0{,}7x = 1{,}19 + 0{,}7 \cdot 7 = 6{,}09.$

Таким образом, мы получили, что при выпуске продукции 7 тыс ед. средний расход условного топлива прогнозируется 6,09. ■

Задачи к §4

В задачах 1– 4 по виду корреляционного поля определите тип связи.

1.

X	2	5	6	9	11
Y	3	6	4	6	3

2.

X	9	10	12	5
Y	6	4	7	3

3.

X	4	3	6	7	10
Y	8	12	5	3	2

4.

X	5	8	4	3	9
Y	9	7	9	1	1

5. По данным шести однотипных мясомолочных хозяйств исследуется влияние электровооруженности (X, кВт) на производительность труда (Y, тыс.руб./чел.). Статистические расчеты дали следующие результаты: $\bar{x} = 70; S_x = 21; \bar{y} = 30; \quad S_y = 9; \quad \sum x_i y_i = 13722.66$. Определить коэффициент корреляции между электровооруженностью и производительностью труда. Построить линию регрессии.

6. В задаче представлены данные о стаже работы (X, лет) и производительности труда (Y, деталей в час). Вычислить выборочный коэффициент корреляции, построить корреляционное поле и линию регрессии для следующей выборки:

X	9	10	12	5
Y	6	4	7	3

7. В торговых организациях города исследовалась зависимость между размером дневного товарооборота и числом рабочих мест. Результаты представлены в следующей таблице.

Группа предприятий по размеру товарооборота, тыс. д.е.	Группа предприятий по числу рабочих мест			
	1-3	3-5	5-7	7-9
10-20	2	-	-	-
20-30	-	2	1	-
30-40	-	3	4	-
40-50	-	-	2	1

Построить линию регрессии.

8. В следующей таблице представлены данные, отражающие статистическую связь издержек обращения (*Y*, тыс. р.) и товарооборота (*X*, тыс. р.):

X	5.0	5.2	5.8	6.4	6.6	7.0
Y	17.6	17.5	18.0	18.1	18.2	18.5

Построить уравнение регрессии. Объяснить его. Спрогнозировать издержки обращения при заданном товарообороте в 7,5 тыс. р.

9. На металлургическом заводе исследовалась зависимость предела прочности $(H/мм^2)$ от предела текучести $(H/мм^2)$. Результаты прочности (x_i) и текучести (y_i) стали 50 марок приведены в таблице:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
77	81	81	54	129	100	104	94	94	84
96	77	57	40	145	95	108	84	112	94
86	76	86	61	142	206	93	73	136	162
92	86	80	68	120	118	124	107	104	98
98	53	87	88	95	109	112	94	103	77
53	47	163	145	107	107	113	107	114	88
63	36	133	136	133	120	94	99	123	94
80	40	159	129	140	114	112	100	111	76
64	49	153	126	149	113	116	104	127	84
66	60	134	96	147	123	93	88	129	73

Составить корреляционную таблицу и вычислить выборочный коэффициент корреляции.

10. Для установления зависимости между стажем работы и выработкой рабочих (швей) вычислен коэффициент корреляции *r=0,91, n=12*. Требуется при уровне значимости 0,05 проверить гипотезу о значимости коэффициента корреляции.

11. По данным таблицы изменения веса поросят (*Y*, кг) в зависимости от их возраста (*X*, недели) построить линию регрессии. Какой будет предположительно вес 10-недельного поросенка?

X	0	1	2	3	4	5	6	7	8
Y	1,3	2,5	3,9	5,2	6,3	7,5	9,0	10,8	13,1

12. Выборочный коэффициент корреляции, вычисленный по выборке объема 18, равен 0,25. Проверить значимость этого результата при альтернативных гипотезах: а) $H_1 : \rho \neq 0$; б) $H_1 : \rho < 0$. Принять $\alpha = 0{,}05$.

13. В опытном хозяйстве на протяжении 38 месяцев отмечали расходы на механизацию работ (*X*, тыс. руб.) и получение привеса всего скота (*Y*, ц). Установили, что имеет место прямая корреляционная зависимость между ними: *r=0,8*. Проверить значимость этой связи при *α=0,01*.

14. Даны несколько уровней значимости: α_1=0,01; α_2=0,02; α_3=0,05; α_4=0,001. Укажите наименьший уровень значимости (из приведенных), при котором значим выборочный коэффициент корреляции *r=0,1*, если объем выборки равен 102.

15. При исследовании корреляционной зависимости между стоимостью основных производственных фондов (млн д.е.) и объемом валовой продукции (млн д.е.) для 10 предприятий был получен выборочный коэффициент корреляции r = 0,6. Проверьте значимость этой связи при α = 0,05.

16. Для исследования были выбраны 27 однотипных предприятий одной отрасли промышленности, которые характеризуются следующими данными.

X – балансовая прибыль (тыс. д.е.),

Y – объем реализованной продукции (тыс. д.е.).

По результатам наблюдений были вычислены следующие суммы:

$\sum_{i=1}^{n} x_i = 190{,}3;\ \sum_{i=1}^{n} x_i^2 = 1519{,}708,\ \sum_{i=1}^{n} x_i y_i = 12835{,}22;$

$\sum_{i=1}^{n} y_i = 1679{,}9;\ \sum_{i=1}^{n} y_i^2 = 112949{,}7.$

1) Найдите среднюю балансовую прибыль (тыс. д.е.), средний объем реализованной продукции (тыс. д.е.).

2) Найдите выборочный коэффициент корреляции и составьте уравнение регрессии. Проверьте значимость коэффициента корреляции.

17. Что можно сказать о взаимосвязи между признаками X и Y, если коэффициент корреляции между ними оказался равным: а) 0,05; б) 1; в) 0,93; г) 1,3; д) – 0,93?

18. В таблице представлены данные об обороте розничной торговли и расходах на рекламу по десяти магазинам:

Магазин	Оборот розничной торговли, млн. руб.	Расходы на рекламу, тыс. руб
1	15,2	9,5
2	16,7	10,2
3	18,5	11,5
4	20,4	13,3
5	18,2	11,4
6	24,3	16,0
7	28,2	17,2
8	26,4	15,1
9	20,6	13,0
10	27,5	16,3

Выявите наличие, направление и форму связи между оборотом розничной торговли и расходами на рекламу. Найдите выборочный коэффициент корреляции и составьте уравнение регрессии. Проверьте значимость коэффициента корреляции.

Контрольные вопросы

1. Какая связь называется функциональной, в каких областях науки она наиболее широко распространена?
2. В чем состоит отличие между функциональной и стохастической связью?
3. Какой признак называют факторным?
4. Какой признак называют результативным?
5. Какая связь называется корреляционной и в чем ее сущность? Приведите примеры корреляционной зависимости.
6. Какие бывают виды связи по направлению?
7. Раскройте понятие «корреляционное поле». Какова его роль в корреляционном анализе?
8. Каковы пределы изменения коэффициента корреляции и интерпретация его величины?
9. Что показывает знак линейного коэффициента корреляции?
10. С какой целью используется корреляционный анализ?
11. С какой целью проверяется гипотеза о значимости выборочного коэффициента корреляции?

ПРИЛОЖЕНИЕ

Таблица I

Нормальное распределение $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$

	Сотые доли									
x	0	1	2	3	4	5	6	7	8	9
0,0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767

Окончание таблицы I

x	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,999	0,999
3.1	0,999	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3.4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Таблица II

Критические точки распределения χ^2

ν \ α	0,99	0,95	0,90	0,10	0,05	0,02	0,01	0,001
1	$0,0^3157$	0,00393	0,0158	2,706	3,841	5,412	6,635	10,827
2	0,0201	0,103	0,211	4,605	5,991	7,824	9,210	13,815
3	0,115	0,352	0,584	6,251	7,815	9,837	11,345	16,266
4	0,297	0,711	1,064	7,779	9,488	11,688	13,277	18,467
5	0,544	1,145	1,610	9,236	11,070	13,388	15,086	20,515
6	0,872	1,635	2,204	10,645	12,592	15,033	16,812	22,457
7	1,239	2,167	2,823	12,017	14,067	16,622	18,475	24,322
8	1,646	2,733	3,490	13,362	15,507	18,168	20,090	26,125
9	2,088	3,325	4,168	14,684	16,919	19,679	21,666	27,877
10	2,558	3,940	4,865	15,987	18,307	21,161	23,209	29,588
11	3,053	4,575	5,578	17,275	19,675	22,618	24,725	31,264
12	3,571	5,226	6,304	18,549	21,026	24,054	26,217	32,909
13	4,107	5,892	7,042	19,812	22,362	25,472	27,688	34,528
14	4,660	6,571	7,790	21,064	23,685	26,873	29,141	36,123
15	5,229	7,261	8,547	22,307	24,996	28,259	30,578	37,697
16	5,812	7,962	9,312	23,542	26,296	29,633	32,000	39,252
17	6,408	8,672	10,085	24,769	27,587	30,995	33,409	40,790
18	7,015	9,390	10,865	25,989	28,869	32,346	34,805	42,312
19	7,633	10,117	11,651	27,204	30,144	33,687	36,191	43,820
20	8,260	10,851	12,443	28,412	31,410	35,020	37,566	45,315
21	8,897	11,591	13,240	29,615	32,671	36,343	38,932	46,797
22	9,542	12,338	14,041	30,813	33,924	37,659	40,289	48,268
23	10,196	13,091	14,848	32,007	35,172	38,968	41,638	49,728
24	10,856	13,848	15,659	33,196	36,415	40,270	42,980	51,179
25	11,524	14,611	16,473	34,382	37,652	41,566	44,314	52,62(1
26	12,198	15,379	17,292	35,563	38,885	42,856	45,642	54,052
27	16,151	18,114	20,703	40,113	44,140	46,963	46,963	55,476
28	13,565	16,928	18,939	37,916	41,337	45,419	48,278	56,893
29	14,256	17,708	19,768	39,087	42,557	46,693	49,588	58,302
30	14,953	18,493	20,599	40,256	43,773	47,962	50,892	59,703

Окончание таблицы II

ν \ α	0,99	0,95	0,90	0,10	0,05	0,02	0,01	0,001
36	23,269	25,643	28,735	47,212	50,999	55,489	58,619	67,985
40	26,509	29,051	32,345	55,759	60,436	63,691	63,691	73,402
46	26,657	31,439	34,215	58,641	62,830	67,771	71,201	81,400
50	29,707	34,764	37,689	63,167	67,505	72,613	76,154	86,661
56	34,350	39,801	42,937	69,919	74,468	79,815	83,513	94,461
60	37,485	43,188	46,459	74,397	79,082	84,580	88,379	99,607
66	42,240	48,305	51,770	81,085	85,965	91,681	95,626	107,258
70	45,442	51,739	55,329	85,527	90,531	96,388	100,425	112,317

Таблица III

Критические точки распределения Стьюдента

Число степеней свободы ν	Уровень значимости α (двусторонняя критическая область)					
	0, 10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Таблица IV

Процентные точки распределения Фишера-Снедекора (F-распределения)

Таблица IV.1

F-распределение: верхние 5%-е точки

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	285
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24

Продолжение таблицы IV.1

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

ν_2 \ ν_1	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,1	250,1	251,2	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	29,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73

Окончание таблицы IV.1

ν_2 \ ν_1	12	15	20	24	30	40	60	120	∞
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,87	1,82	1,77	1,65
29	2,01	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Таблица IV.2

F-распределение: верхние 1%-е точки

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	4052	4999,5	5403	5625	5764	5859	1928	5981	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,73	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	1225	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,44
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	529	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	383	3,51	3,29	3 12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47

Продолжение таблицы IV.2

ν_2 \ ν_1	12	15	20	24	30	40	60	120	∞
1	6106	6157	6209	6235	6261	6287	6313	6339	6336
2	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	90,50
3	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	14,37	14,20	14,12	13,43	13,84	13,75	13,65	13,56	13,46
5	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,47	6,31	6,16	6,07	5,99	5,91	5,Я2	5,74	5,65
8	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,16	4,01	3,86	3,78	3,70	3,62	3,54	1,45	3,36
13	3,96	3,82	3,66	3,S9	3,51	3,43	3,14	3,25	3,17
14	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	2,96	2,81	2,66	2,58	2,50	2,42	2,31	2,21	2,13
27	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,01
30	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,18
∞	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Таблица IV.3

F-распределение: верхние 0,1%-е точки

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	4053*	5000*	5404*	5625*	5764*	5859*	5981*	6023*	6056*	6107*
2	998,5	999,0	999,2	999,2	999,3	999,3	999,4	999,4	999,4	999,4
3	167,0	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2
4	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	48,47	48,05
5	47,18	37,12	33,20	31,09	29,75	2884	28,16	27,64	27,24	26,92
6	35,51	27,00	23,70	21,92	20,81	20,03	19,46	19,03	18,69	18,41
7	29,25	21,69	18,77	17,19	16,21	I5,52	15,02	14,63	14,33	14,08
8	25,42	18,49	15,83	14,39	13,44	12,86	12,40	12,04	11,77	11,54
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89
10	21,04	14,91	12,55	11,28	10,48	9,92	9,52	9,20	8,96	8,75
11	19,69	13,81	11,56	10,35	9,58	9,05	8,66	P, 35	8,12	7,92
12	18,64	12,97	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29
13	17,81	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80
14	17,14	11,78	9,73	8,62	7,92	7,43	7,08	6,80	6,58	6,40
15	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08
16	16,12	10,97	9,00	7,94	7,27	6,81	6,46	6,19	5,98	5,81
17	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58
18	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39
19	15,08	10,16	8,28	7,26	6,62	6,18	5,85	5,59	5,39	5,22
20	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08
21	14,59	9,77	7,94	6,95	6,32	5,88	5,31	5,31	5,11	4,95
22	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83
23	14,19	9,47	7,67	6,69	6,08	5,65	5,33	5,09	4,89	4,73
24	14,03	9,34	7,55	6,59	5,98	5,55	5,23	4,99	4,80	4,64
25	13,88	9,22	7,45	6,49	5,88	5,46	5,15	4,91	4,71	4,56
26	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48
27	13,61	9,02	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41
28	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35
29	13,39	8,85	7,12	6,19	5 59	5,18	4,87	4,64	4Л5	4,29
30	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24
40	12,61	8,25	6,60	5,70	5,13	4,73	4,44	421	4,02	3,87
60	11,97	7,76	6,17	5,31	4,76	4,37	4,09	3,87	3,69	3,54
120	11,38	7,32	5,79	4,95	4,42	4,04	3,77	3,55	3,38	3,24
∞	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96

Окончание таблицы IV.3

ν_2 \ ν_1	12	15	20	24	30	40	60	120	∞
1	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
2	999,4	999,4	999,4	999,5	Ф995	999,5	999,5	999,5	999,5
3	128,3	127,4	126,4	125,9	125,4	125,0	124,5	124,0	123,5
4	47,41	46,76	46,10	45,77	45,43	45,09	44,75	44,40	44,05
5	26,42	25,91	25,39	25,14	24,87	2460	24,33	24,06	23,79
6	17,99	17,56	17,12	16,89	16,67	16,44	1621	15,99	15,75
7	13,71	13,32	12,93	12,73	12,53	12,33	12,12	11,91	11,70
8	11,19	10,84	10,48	10,30	10,11	9,92	9,73	9,53	9,33
9	9,57	9,24	8,90	8,72	8,55	8,37	8,19	8,00	7,81
10	8,45	8,13	7,80	7,64	7,47	7,30	7,12	6,94	6,76
11	7,63	7,32	7,01	6,85	6,68	6,52	*6,35*	6,17	6,00
12	7,00	6,71	6,40	6,25	6,09	5,93	5,76	5,59	5,42
13	6,52	6,23	5,93	5,78	5,63	5,47	5,30	5,14	4,97
14	6,13	5,85	5,56	5,41	5,25	5,10	4,94	4,77	4,60
15	5,81	5,54	5,25	5,10	4,95	4,80	4,64	4,47	4,31
16	5,55	5,27	4,99	4,85	4,70	4,54	4,39	4,23	4,06
17	5,32	5,05	4,78	4,63	4,48	4,33	1,18	4,02	3,85
18	5,13	4,87	4,59	4,45	4,30	4,15	4,00	3,84	3,67
19	4,97	4,70	4,43	4,29	4,14	3,99	3,84	3,68	3,51
20	4,82	4,56	4,29	4,15	4,00	3,86	3,70	3,54	3,38
21	4,70	4,44	4,17	4,03	3,88	3,74	3,58	3,42	3,26
22	4,58	4,33	4,06	3,92	3,78	3,63	3,48	3,32	3,15
23	4,48	4,23	3,96	3,82	3,68	3,53	3,38	3,22	3,05
24	4,39	4,14	3,87	3,74	3,59	3,45	3,29	3,14	2,97
25	4,31	4,06	3,79	3,66	3,52	3,37	3,22	3,06	2,89
26	4,24	3,99	3,72	3,59	3,44	3,30	3,15	2,99	282
27	4,17	3,92	3,66	3,52	3,38	3,23	3,08	2,92	2,75
28	4,11	3,86	3,60	3,46	3,32	3,18	3,02	2,86	2,69
29	4,05	3,80	3,54	3,41	3,27	3,12	2,97	2,81	2,64
30	4,00	3,75	3,49	3,36	3,22	3,07	2,92	2,76	2,59
40	3,64	3,40	3,15	3,01	2,87	2,73	2,57	2,41	2,23
60	3,31	3,08	2,83	2,69	2,55	2,41	2,25	2,08	1,89
120	3,02	2,78	2,53	2,40	2,26	2,11	1,95	1,76	1,54
∞	2,74	2,51	2,27	2,13	1,99	1,84	1,66	1,45	1,00

* Эти значения надо умножить на 100,

Терминологический словарь

Альтернативная (конкурирующая гипотеза) – это гипотеза H_1, которая противоречит нулевой гипотезе.

Вариационный ряд – последовательность вариант, записанных в возрастающем порядке.

Выборка – это совокупность случайно отобранных объектов.

Выборочное среднее $\overline{x}$ – это среднее арифметическое всех вариант выборки.

Выборочная дисперсия S^2 – это средний квадрат отклонения значений вариант от их средней арифметической.

Выборочное стандартное отклонение (или *выборочное среднее квадратическое отклонение*) – это корень квадратный из выборочной дисперсии

Гистограмма – фигура, состоящая из прямоугольников, высота которых пропорциональна числу наблюдений, попавших в данные интервалы числовой прямой.

Дискретный вариационный ряд – это перечень вариант и соответствующих им частот или относительных частот.

Доверительный интервал для параметра θ – это интервал $(\hat{\theta}_1, \hat{\theta}_2)$ со случайными концами, накрывающий истинное значение θ с вероятностью не меньшей 1– α, т.е. $P(\left|\theta - \hat{\theta}\right| < \Delta) \geq 1 - \alpha$.

Доверительный уровень (доверительная вероятность, надежность) – это вероятность γ того, что доверительный интервал накроет истинное значение параметра.

Интервальная оценка – это оценка, которая определяется двумя числами – концами интервалов.

Интервальный вариационный ряд – это соответствие между интервалами и частотами, которые равны сумме частот вариант, относящихся к интервалу.

Критерий К – это правило, по которому определяется, принять или отклонить гипотезу H_0.

Критическая область – это множество всех значений статистики критерия, при которых нулевая гипотеза отклоняется

Математическая статистика – раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации результатов наблюдений с целью выявления статистических закономерностей.

Медиана ($\hat{Me}$) – это значение признака, которое приходится на центральный член вариационного ряда.

Мода ($\hat{Mo}$) – это значение варианты, частота или относительная частота которой имеет наибольшее значение.

Мощность критерия – вероятность (1– β) не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0, когда она неверна.

Надежность – см. доверительный уровень.

Область принятия гипотезы – это множество всех значений статистики критерия, при которых нулевая гипотеза принимается

Объем выборки – это число элементов выборки.

Оценка (статистическая оценка) – это функция от наблюдаемых случайных величин, принимающая значения в пространстве возможных значений оцениваемого параметра.

Ошибка второго рода – это принятие ложной гипотезы H_0,

Ошибка первого рода – это отклонение гипотезы H_0, когда она верна.

Полигон – это ломаная кривая, соединяющая точки (x_1,h_1), (x_2,h_2),.., (x_r, h_r), абсцисса которых соответствует варианте или середине интервала, а ордината пропорциональна частоте варианты или соответствующего интервала.

Предельная ошибка выборки Δ – это наибольшее отклонение оценки $\hat{\theta}$ от оцениваемого параметра θ, которое возможно с заданным доверительным уровнем γ.

Признаки зависимые – см. признаки результативные.

Признаки независимые – см. признаки факторные.

Признаки результативные (зависимые) – это признаки, зависимые от факторных признаков.

Признаки факторные (независимые) – это признаки, которые обуславливают изменение других, связанных с ними признаков.

Ранг – это порядковый номер объекта в ранжированной выборке.

Ранжированная выборка – это выборка, в которой объекты (или значения признаков) расположены в порядке возрастания или убывания их свойств.

Связь корреляционная – статистическая связь между двумя переменными, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой.

Связь функциональная – связь, при которой значение одной величины однозначно определяется значением другой и выражается математической формулой.

Статистика – это любая функция от наблюдений.

Статистической гипотезой, обозначаемой *H*, называется любое предположение относительно вида или параметра распределения с.в. *X*.

Частота – это число наблюдений варианты.

Частость (относительная частота) – это отношение частоты к объему выборки.

ЛИТЕРАТУРА

Основная литература

1. Ватутин В.А., Ивченко Г.И., Медведев Ю.И. и др. Теория вероятностей и математическая статистика в задачах: учебное пособие для вузов. – 2-е изд., испр. – М.: Дрофа, 2003. – 328 с.: ил.
2. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel: учебное пособие для вузов. – Изд. 3-е, доп. и перераб. – Ростов-на-Дону: Феникс, 2005. – 480 с.
3. Кремер Н.Ш. Теория вероятностей и математическая статистика. – 3-е изд., перераб. и доп. – М.: ЮНИТИ-ДАНА, 2009. – 551 с.
4. Общий курс высшей математики для экономистов: Учебник под ред. В.И. Ермакова. – М.: ИНФРА-М, 2003. – 575 с.

Дополнительная литература

5. Айвазян С.А., Мешалкин Л.Д., Енюков И.С. Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1985. – 472 с.
6. Гмурман В.Е. Теория вероятностей и математическая статистика. – 9-е изд., стер. – М.: Высш. шк., 2003. – 479 с., ил.
7. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1998. – 400 с., ил.
8. Елисеева И.И. Статистические методы измерения связей. – Л.: ЛГУ, 1982. – 136 с.
9. Калинина В.Н., Панкин В.Н. Математическая статистика. – М.: Высшая школа. – 4-е изд., испр. – М., 2002. – 336 с., ил.
10. Кимбл Г. Как правильно пользоваться статистикой / пер. с англ. Б.И. Клименко. – Финансы и статистика, 1982. – 294 с., с ил.

11. Справочник по прикладной статистике / под ред. Э. Ллойда и У. Ледермана. Том 2. – М.: Финансы и статистика, 1990. – 526 с.
12. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов: пер. с нем. – М.: Финансы и статистика, 1983. – 302 с.

Учебное издание

Каштанова Елена Кирилловна

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Учебное пособие

Подписано к использованию 24.07.2025

Научная библиотека им. Н.И. Лобачевского