В.В. АФАНАСЬЕВ

Применение методов математической статистики в научных исследованиях

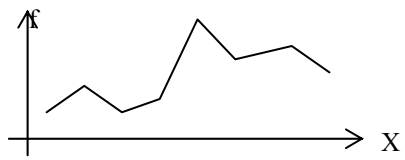
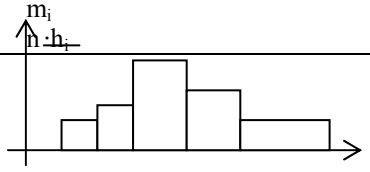
Классика – это то, что все хотели бы прочитать, но никто читать не хочет
Марк Твен

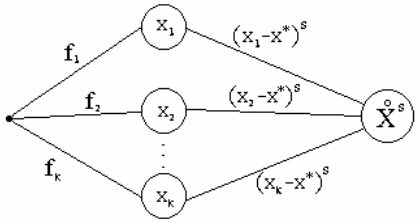
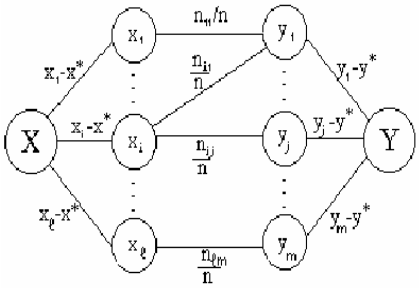
В статье изложены основные идеи доклада автора на традиционном университетском Дне науки 2006 года. Актуальность постановки задачи об использовании методов математической статистики обусловлена, с одной стороны, возросшими прикладными исследованиями, которые по своей сущности являются стохастическими, а с другой – недостаточной подготовленностью исследователей (особенно педагогов) к научно обоснованному проведению и анализу эксперимента, к грамотному выбору средств и критериев математической статистики. В настоящее время в достаточно большом объеме издается научная литература по математической статистике, которая для большинства читателей малодоступна, а изложение материала во многих случаях сухое и скучное. В работе предлагаются наиболее простые и удобные вероятностно-статистические приемы и методы исследования.

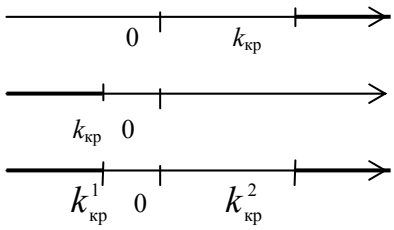
Задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов. Термин «статистика» произошел от латинских слов stato (государство) и status (положение вещей). В XVII веке под статистикой подразумевали «государствоведение» и «политическую арифметику». Карл Федорович Герман, первый руководитель статистического комитета, созданного при Министерстве полиции России, являлся автором работы «Статистическое описание Ярославской губернии» (1805 г.).

Следуя этому историческому посылу, в работе [1] мы проводим исследования реальной жизни Ярославской области, отраженной в статистических данных социально-экономического развития Ярославской области, качественного и количественного состава учительских кадров, урожайности коллективных хозяйств, экономической рентабельности предприятий, спортивных достижений команд и т.д. Основные знания, умения и навыки по курсу «Математическая статистика» сведены автором в следующую опорную таблицу.

Предложенная таблица поможет читателю ориентироваться в большом потоке статистических рассуждений и выводов. На следующих примерах покажем наиболее распространенные и простые вычисления основных параметров реальных статистических данных.

ОСНОВНЫЕ		
ЗНАНИЯ		**УМЕНИЯ НАВЫКИ**
ПОНЯТИЯ	**ТЕОРЕМЫ**	
Выборка Выборка объема n Вариационный, статистический ряд Размах варьирования	$\sum_i f_i = 1$	Находить полигон частот и относительных частот.  Строить гистограмму выборки.
Эмпирическая функция распределения F_n (x)	$F_n(x) \to F(x) \quad n \to \infty$	

Закон распределения		X
Показатели положения: выборочная средняя, мода, медиана Показатели разброса: выборочная дисперсия, неисправленная дисперсия, статистическое и исправленное среднее квадратическое отклонение Асимметрия Выборочные начальные и центральные моменты	$(x+y)^* = x^* + y^*$ $(cx)^* = cx^*$ $D^*(C) = 0;$ $D^*(C \cdot X) = C^2 D^*(X);$ $D^*(X \pm Y) = D^*(X) + D^*(Y);$ $S = \sqrt{\sum \frac{m_i (x_i - x^*)^2}{n-1}}$ $S^2 = \frac{n}{n-1} D^*;$ $M[S^2] = D^*;$ $D_{общ.} = D_{внгр.} + D_{межгр.};$ $D_{межгр.} / D_{общ.} \le 1;$	Вычислять показатели положения. Находить центральные моменты μ_S как полный вес графа распределения статистического ряда: 
Доверительный интервал Надежность	$x^* - t\frac{\sigma}{\sqrt{n}} < a < x^* + t\frac{\sigma}{\sqrt{n}}$, где $F(t) = \frac{\alpha}{2};$ $x^* - t_\alpha \frac{S}{\sqrt{n}} < a < x^* + t_\alpha \frac{S}{\sqrt{n}},$ где $t_\alpha = \gamma(n, \alpha);$ $S(1-q) < \sigma < S(1+q)$	Находить доверительный интервал, который с заданной надежностью α покрывает оцениваемый параметр (математическое ожидание или среднее квадратическое отклонение)
Выборочная ковариация $k(X,Y)$	$k(X,Y) = \frac{1}{n}\sum (x_i - x^*) \times (y_j - y^*) n_{ij}$	Находить выборочную ковариацию как полный вес ковариационного графа: 
Выборочный коэффициент корреляции $r(x,y)$ Выборочное корреляционное отношение Уравнение регрессии *У* на *X* (*X* на *У*) Ранговая корреляция Выборочный коэффициент ранговой корреляции Спирмена Коэффициент ранговой корреляции Кендалла	$r(X,Y) = \frac{\sum n_{xy} xy - x^* y^*}{n \sigma_x^* \cdot \sigma_y^*},$ $\lvert r(X,Y) \rvert \le 1$ $\eta_{yx} = \frac{\sigma_{y_x^*}}{\sigma_y^*},$ $0 \le \eta \le 1,\ \eta \ge \lvert r(X,Y) \rvert,$ $D_{внгр} = D_{общ.}(1 - \eta^2) \cdot$ $y_x^* - y^* = \frac{k(X,Y)}{D_x^*}(x - x^*)$ $r_s^* = 1 - \frac{6\sum d_i^2}{n^3 - n}$ $T_{кр} = t_{кр}(\alpha, k)\sqrt{\frac{1-(r_s^*)^2}{n-2}};$ $r_k^* = \frac{4R}{n(n-1)} - 1;$ $T_{кр} = Z_{кр}\sqrt{\frac{2(2n+5)}{9n(n-1)}}$	Строить графики регрессии $y^*_x = f(x)$ или $x^*_y = \varphi(y)$, пользуясь методом наименьших квадратов. Находить коэффициенты ранговой корреляции Спирмена, Кендалла и область принятия нулевой гипотезы. Обосновать область принятия нулевой гипотезы.

Статистические гипотезы. t – распределение Стьюдента	$t_{\text{эмп}} = \dfrac{x^* - y^*}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \cdot \sqrt{\dfrac{nm(n+m-2)}{n+m}}$;	Отыскивать области принятия нулевой гипотезы(или конкурирующей) 
χ^2 – критерий Пирсона	$x^2_{\text{эмп}} = \sum \dfrac{(n_i - n'_i)^2}{n'_i}$	$T_{\text{эмп}} \le t_{\text{кр}}$
λ – критерий Колмогорова – Смирнова	$\lambda = d_{\max} \cdot \sqrt{\dfrac{n_1 \cdot n_2}{n_1 + n_2}}$	$x^2_{\text{эмп}} \le x^2_{\text{кр}}$
Критерий x^2_r Фридмана Критерий *L* Пейджа Критерий *Q* Розенбаума	$x^2_{r\,\text{эмп}} = \dfrac{12}{n \cdot c \cdot (c+1)} \cdot \sum T_j^2 - 3n(c+1)$ $L_{\text{эмп}} = \sum T_j \cdot j$ $Q_{\text{эмп}} = S_1 + S_2$	$\lambda_{\text{эмп}} > \lambda_{\text{кр}}$ $x^2_{r\,\text{эмп}} < x^2_{r\,\text{кр}}$ $L_{\text{эмп}} \ge L_{\text{кр}}$ $Q_{\text{эмп}} < Q_{\text{кр}}$
Критерий *T* Вилкоксона Критерий *U* Манна-Уитни Критерий *H* Крускала-Уоллиса Критерий Бартлетта	$T_{\text{эмп}} = \sum R_r$ $U = n_1 \cdot n_2 + \dfrac{n_x \cdot (n_x + 1)}{2} - \Phi_x$ $H = \dfrac{12}{N(N+1)} \cdot \sum \dfrac{T_j^2}{n} - 3(N+1)$ $B = 2{,}3\left[\left(\sum n_i - m\right)\lg D^* - \sum (n_j - 1)\lg D_j^*\right]$, $C = 1 + \dfrac{1}{3(m-1)}\left[\sum \dfrac{1}{n_j - 1} - \dfrac{1}{\sum n_j - m}\right]$	$T_{\text{эмп}} \le T_{\text{кр}}$ $U_{\text{эмп}} > U_{\text{кр}}$ $H_{\text{эмп}} < H_{\text{кр}}$ $B/C \le \chi^2(\alpha, m-1)$

Пример 1. Найти, на сколько отличаются требования студентов и преподавателей к личности преподавателя, определенные в ходе социологического опроса.

Требования к личности преподавателя

№	Качества личности	Студенты		Преподаватели		d_i^2
		%	ранг	%	ранг	
		X		Y		
1	Глубокие знания	54	2,5	67	2	0,25
2	Умение объяснять	87	1	70	1	0
3	Увлеченность наукой	10	10	22	6	16
4	Знание практики	34	5,5	46	3	6,25
5	Общительность	39	4	18	7	9
6	Отзывчивость	34	5,5	16	8	6,25
7	Чувство юмора	54	2,5	12	9	42,25
8	Интеллигентность	16	8	32	5	9
9	Требовательность	13	9	42	4	25
10	Демократичность	21	7	8	10	9

	$\sum = 123$

Решение. В качестве измерителей тесноты парных связей между количественными переменными будем использовать коэффициент ранговой корреляции. Пусть объекты генеральной совокупности обладают двумя качественными признаками, которые проранжируем в порядке ухудшения качества. Рассматривая ранги x_1, x_2,..., x_n как возможные значения случайной величины X, а y_1, y_2,..., y_n – как возможные значения величины Y, можно вычислить выборочный **коэффициент корреляции Спирмена** r_s:

$$r_s = 1 - \frac{6 \cdot \Sigma d_i^{\,2}}{(n-1) \cdot n \cdot (n+1)},$$

где $d_i = x_i - y_i$ (разность соответствующих рангов). Заметим, что при равных показателях им присваивается один общий ранг, равный среднему арифметическому соответствующих возможных мест.

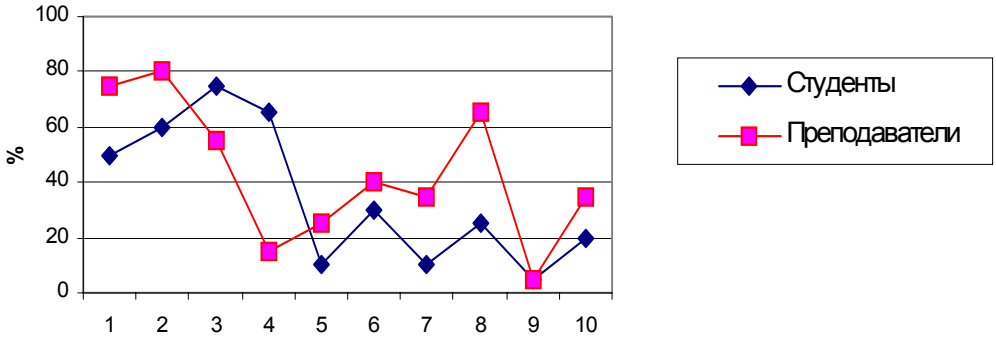
Найдем в нашем случае ранговый коэффициент Спирмена:

$$r_s = 1 - \frac{6 \cdot 123}{9 \cdot 10 \cdot 11} = 1 - \frac{41}{55} = \frac{55 - 41}{55} = \frac{16}{55} \approx 0{,}25.$$

Отсюда следует, что требования студентов и преподавателей к личности преподавателя значимо отличаются.

Пример 2. Жизненные ценности студентов и преподавателей представлены на полигоне частот. Определите, на сколько схожи выборы ценностей у студентов и преподавателей.

Выбор ценностей



1 –

Высокий заработок

6 – Душевное спокойствие

2 – Интересная работа

7 – Профессиональные достижения

3 – Любимый человек

8 – Уважение окружающих

4 – Хорошие друзья

9 – Высокое социальное положение

5 – Собственность, капитал

10 – Чистая совесть

Решение. Проранжируем перечисленные ценности у студентов и преподавателей и занесем их в следующую таблицу:

Ценность	1	2	3	4	5	6	7	8	9	10
Ранг у студентов	4	3	1	2	8,5	5	8,5	6	10	7
Ранг у преподавателей	2	1	4	9	8	5	6	3	10	7
d_i^2	4	4	9	49	0,25	0	2,25	9	0	0

Поскольку $\Sigma d_i^{\,2} = 77{,}5$, то коэффициент ранговой корреляции Спирмена $r_s = 1 - \dfrac{6 \cdot 77{,}5}{9 \cdot 10 \cdot 11} \approx 0{,}53$ и можно утверждать, что между выборами ценностей у студентов и преподавателей существует прямая и средней силы связь.

Пример 3. Найти степень соответствия лучшей жизни и достатка (по данным британского журнала «Economics», ноябрь 2004 г.).

Место	Страна	Место по достатку	Ранг по достатку	d_i^2
1	Ирландия	4	4	9
2	Швейцария	7	5	9
3	Норвегия	3	3	0
4	Люксембург	1	1	9
5	Швеция	19	9	16
6	Австралия	14	8	4
7	Исландия	8	6	1
8	Италия	23	10	4
9	Дания	10	7	4
10	Испания	24	11	1
11	США	2	2	81
				$\Sigma d_i^{\,2} = 138$

Проанализировав места по достатку и найдя разность рангов по лучшей жизни и по достатку, находим коэффициент ранговой корреляции

$$r_s = 1 - \frac{6 \cdot 138}{10 \cdot 11 \cdot 12} \approx 0{,}373 .$$

По найденному коэффициенту видно, что существует прямая и средняя связь между лучшей жизнью и достатком («счастье не в деньгах»).

В следующем примере рассмотрим результаты Единого государственного экзамена в Ярославской области в 2006 году по трем показателям (средний балл, справляемость и успешность). В таком случае, когда исследуется связь между несколькими признаками, корреляцию называют множественной, и она задается всеми коэффициентами r_{ij} парных корреляций, которые записывают в **корреляционную матрицу**:

$$(r) = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ & 1 & r_{23} & \dots & r_{2n} \\ & & 1 & \dots & r_{3n} \\ & & & & \vdots \\ & & & & 1 \end{pmatrix}$$

Заметим, что матрица (r) является треугольной, поскольку $r_{ij} = r_{ji}$, и нет смысла их повторять дважды.

Пример 4. В [3] приводятся основные результаты ЕГЭ в 2006 году по районам Ярославской области. Найти корреляционную матрицу между средним баллом, справляемостью (количество учащихся, получивших отметки “3”, “4”, “5”) и успешностью (количество учащихся, получивших отметки “4”, “5”) по русскому языку, например.

Решение. Дополним традиционные показатели ЕГЭ по русскому языку ([3. С. 16]) их найденными рангами по районам и по трем признакам, запишем все эти данные в следующую таблицу:

Муниципальный район	Кол-во писав-ших	Средний балл (1)	Ранг	Справляе-мость (2)	Ранг	Успеш-ность (3)	Ранг
г. Ярославль	3759	54,90	3	97,05	7	66,27	3
Большесельский МР	69	54,49	4	94,20	17	55,07	14
Борисоглебский МР	80	51,89	11	97,50	5	57,50	9
Брейтовский МР	84	48,32	19	89,29	19	45,24	19
Гаврилов-Ямский МР	168	52,11	10	95,83	10	57,14	10
Даниловский МР	199	53,92	5	99,50	1	60,80	5
Любимский МР	99	51,47	13	93,94	18	56,57	11
Мышкинский МР	70	57,17	1	95,71	12	67,14	2
Некоузский МР	114	52,99	8	97,37	6	56,14	12
Некрасовский МР	124	49,23	18	95,97	9	47,58	17
Первомайский МР	83	51,54	12	95,18	14	59,04	6
Переславский МР	62	53,27	6	98,39	2	58,06	7
Переславль-Залесский	231	54,93	2	98,28	3	69,83	1
Пошехонский МР	95	50,47	15	94,74	15	49,47	18
Ростовский МР	360	52,38	9	94,44	16	57,78	8
Рыбинский МР	1289	53,06	7	97,83	4	62,53	4
Тутаевский МР	364	50,20	16	95,60	13	50,27	15
Угличский МР	241	50,80	14	95,85	10	55,19	13
Ярославский МР	172	49,34	17	96,51	8	47,67	16

Вычисляя суммы квадратов соответствующих рангов, находим и соответствующие коэффициенты ранговых корреляций:

$$r_{12}=1-\frac{6\cdot 653}{18\cdot 19\cdot 20}\approx 0{,}427;\quad r_{13}=1-\frac{6\cdot 187}{18\cdot 19\cdot 20}\approx 0{,}836;\quad r_{23}=1-\frac{6\cdot 490}{18\cdot 19\cdot 20}\approx 0{,}570.$$

Полученные коэффициенты занесем в корреляционную матрицу:

$$(r)=\begin{pmatrix} 1 & 0{,}427 & 0{,}836 \\ & 1 & 0{,}570 \\ & & 1 \end{pmatrix}$$

По матрице (r) видно, что в нашем случае самая сильная связь существует между средним баллом и успешностью, а самая слабая – между средним баллом и справляемостью.

Для определения степени зависимости трех и более показателей используют еще и множественный коэффициент ранговой корреляции, или, иначе говоря, коэффициент конкордации:

$$W=\frac{12S}{m^2\cdot(n-1)\cdot n\cdot(n+1)},$$

где S – сумма квадратов отклонений суммы m рангов от их средней величины,

$$S=\sum_1^n\left(\sum_1^m R_{ij}\right)^2-\frac{\left(\sum_1^n\sum_1^m R_{ij}\right)^2}{n};$$

m – число ранжируемых признаков;

n – число наблюдений.

В заключение рассмотрим результаты финалистов последнего чемпионата мира по четырем показателям (ср. [2. С. 51]).

Пример 5. Найти корреляционную матрицу и коэффициент конкордации для мест на ЧМ–2006, стоимости команд, рейтинга ФИФА и количества клубов мировых футбольных держав.

Группа		Команда	Место на ЧМ-2006	Стоимость команд		Рейтинг ФИФА		Кол-во клубов		$\sum R_{ij}$	$\left(\sum R_{ij}\right)^2$
				млн. евро	ранг		ранг	тыс.	ранг		
A	1.	Германия	3	256	5	16	14	26,7	3	25	625
	2.	Коста-Рика	31	28	31	21	17	0,128	31	110	12100
	3.	Польша	21	44	26	23	18	7,76	8	73	5329
	4.	Эквадор	12	31	29	37	24	0,17	28	93	8649
B	1.	Англия	7	325	2	9	9	42,0	1	19	361
	2.	Парагвай	20	54	22	30	21	1,5	17	80	6400
	3.	Тринидад и Тобаго	29	30	30	51	30	0,135	30	119	14161
	4.	Швеция	14	147	10	14	12	3,23	11	47	2209
C	1.	Аргентина	6	217	6	4	4	3,06	12	28	784
	2.	Кот-д´Ивуар	18	71	21	41	26	0,2	27	92	8464
	3.	Сербия и Черногория	32	92	17	47	27	2,82	13	89	7921
	4.	Голландия	11	174	9	3	3	4,05	10	33	1089
D	1.	Мексика	15	41	27	7	7	1,49	18	67	4489
	2.	Иран	25	48	25	19	15	2,54	14	79	6241
	3.	Ангола	23	27	32	93	32	0,1	32	119	14161
	4.	Португалия	4	198	8	10	10	2,53	15	37	1369
E	1.	Италия	1	304	3	12	11	16,13	6	21	441
	2.	Гана	13	101	12	50	29	0,25	25	79	6241
	3.	США	26	81	18	8	8	1,69	16	68	4624
	4.	Чехия	19	131	11	2	2	4,17	9	41	1681
F	1.	Бразилия	5	410	1	1	1	12,9	7	14	196
	2.	Хорватия	22	95	13,5	20	15	1,19	20	70,5	4970,25
	3.	Австралия	16	95	13,5	49	28	0,25	25	82,5	6808,25
	4.	Япония	27	93	15,5	15	13	19,1	5	60,5	3660,25
G	1.	Франция	2	302	4	5	5	19,8	4	15	225
	2.	Швейцария	10	75	19	36	23	1,45	19	71	5041
	3.	Южная Корея	17	73	20	29	20	0,66	22	79	6241
	4.	Того	30	53	23,5	56	31	0,25	25	109,5	11990,25
H	1.	Испания	9	202	7	6	6	33,6	2	24	576
	2.	Украина	8	93	15,5	40	25	1,09	21	69,5	4830,25
	3.	Тунис	24	37	28	28	19	0,55	23	94	8836
	4.	Саудовская Аравия	28	53	23,5	32	22	0,153	29	102,5	10506,25
										2111	**171219**

$$r_s = \begin{pmatrix} 1 & 0,74 & 0,61 & 0,46 \\ & 1 & 0,76 & 0,52 \\ & & 1 & 0,74 \\ & & & 1 \end{pmatrix}$$

$$S = 171219 - \frac{(2111)^2}{32} = 31959$$

$$W = \frac{12 \cdot 31959}{4^2 \cdot 31 \cdot 32 \cdot 33} \approx 0{,}732$$

По корреляционной матрице видно, что самая сильная связь существует между суммарной стоимостью игроков команд и рейтингом ФИФА ($r_{23} = 0{,}76$), а самая слабая связь – между стоимостью сборных команд и количеством клубов в этих странах ($r_{24} = 0{,}52$). Найденный коэффициент конкордации $W = 0{,}732$ свидетельствует о сильной связи всех четырех рассматриваемых показателей 32(!) мировых футбольных стран, к которым, к сожалению, не относится Россия.

Библиографический список

1. Афанасьев В. В. Теория вероятностей в вопросах и задачах // Учебное пособие. Ярославль: Изд-во ЯГПУ, 2004. 250 с.
2. Афанасьев В. В., Непряев И. И. Математическая статистика в командных видах спорта. Ярославль: Изд-во ЯГПУ, 2006. 120 с.
3. Единый государственный экзамен в Ярославской области в 2006 году / Под ред. М. В. Груздева. Ярославль, 2006. 50 с.