МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«КУЗБАССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ имени Т. Ф. ГОРБАЧЕВА»

Кафедра строительства подземных сооружений шахт
и разработки месторождений полезных ископаемых

Составитель
Г. К. Клюкин

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методические указания к практическим занятиям для студентов очной формы обучения

Рекомендованы учебно-методической комиссией направления подготовки 21.05.04 «Горное дело» в качестве электронного издания для использования в учебном процессе

Кемерово 2016

Рецензенты:

Войтов М. Д. – кандидат технических наук, профессор кафедры строительства подземных сооружений, шахт и разработки месторождений полезных ископаемых

Першин В. В. – доктор технических наук, профессор, заведующий кафедрой строительства подземных сооружений, шахт и разработки месторождений полезных ископаемых, председатель учебно-методической комиссии направления подготовки 21.05.04 «Горное дело», образовательная программа «Шахтное и подземное строительство»

Клюкин Геннадий Константинович

Обработка экспериментальных данных [Электронный ресурс] : методические указания к практическим занятиям по научно-исследовательской работе для студентов направления подготовки 21.05.04 «Горное дело», образовательная программа «Шахтное и подземное строительство» очной формы обучения / сост. Г. К. Клюкин; КузГТУ. – Электрон. дан. – Кемерово, 2016. – Систем. требования: Pentium IV; ОЗУ 256 Мб; Windows XP; мышь. – Загл. с экрана.

Представлен инструментарий для обработки экспериментальных данных – расчет средней арифметической величины в различных значениях, основы теории случайных ошибок и методов оценки случайных погрешностей в измерениях (интервальная оценка точности и надежности с помощью доверительной вероятности, определение минимального количества измерений), методы графической обработки результатов измерений, методы подбора эмпирических формул, метод средних квадратов, метод наименьших квадратов, корреляционный анализ, регрессионный анализ.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Любое научное исследование связано с определением параметров, которые характеризуют исследуемый объект или процесс. Совокупность таких параметров характеризует объект или процесс количественно и качественно.

Параметры получают путем измерения, т.е. путем определения численного значения некоторой величины посредством *единицы* измерения. Измерение предполагает наличие следующих основных элементов: объекта измерения, эталона, измерительных приборов, метода измерения.

Измерение как процедура развилось из операции сравнения. Без эталона получается *качественный* результат (больше – меньше; выше – ниже). Сравнения же объектов с эталоном дают возможность получить *количественные* характеристики. Такие сравнения называются измерением.

Итак, *измерение* – метод научного исследования процесса определения численного значения некоторой величины посредством определенной заранее единицы измерения.

Измерение относится к методам эмпирического уровня. Важнейшим показателем качества измерения, его научной ценности является *точность*.

При совокупности ряда данных полученных при измерениях перед исследователем всегда стоит вопрос: «Какое из данных результатов является истинным?». Одним из ответов на этот вопрос является определение среднего арифметического значения ряда полученных данных.

1. Средние арифметические величины

1.1. Средняя арифметическая простая величина

Простая среднеарифметическая величина представляет собой среднее слагаемое, при определении которого общий объем данного признака в совокупности данных поровну распределяется между всеми единицами, входящими в данную совокупность, т.е. – это *сумма всего набора значений, поделенная на число значений*.

В математике и статистике это – наиболее часто используемое и наиболее полезное измерение центральной тенденции, так как оно использует все данные распределения.

Пример:

1) Чтобы найти среднее арифметическое трех чисел, надо сложить эти числа и результат разделить на 3:

$$(12{,}6 + 14{,}7 + 16{,}5) / 3 = 14{,}6.$$

2) Чтобы найти среднее арифметическое четырех чисел, надо сложить эти числа и результат разделить на 4:

$$(40{,}52 + 44{,}63 + 52{,}34 + 58{,}29) / 4 = 48{,}945.$$

Среднегодовая выработка продукции на одного работающего ***x*** – это такая величина объема продукции, которая приходилась бы на каждого работника, если бы весь объем выпущенной продукции в одинаковой степени распределялся между всеми работниками организации.

Среднеарифметическая простая величина исчисляется по формуле:

$$x = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (1)$$

Пример. Бригада из 6 рабочих получает в месяц 3; 3,2; 3,3; 3,5; 3,8 и 3,1 тыс. руб. Найти среднюю заработную плату. Решение:

$$(3 + 3{,}2 + 3{,}3 + 3{,}5 + 3{,}8 + 3{,}1) / 6 = 3{,}32 \text{ тыс. руб.}$$

1.2. Средняя арифметическая взвешенная

Если объем совокупности данных большой и представляет собой ряд распределения, то исчисляется взвешенная средне-арифметическая величина.

Пример. Определить средневзвешенную производительность породопогрузочных машин на шахте x_{cv}.

$$x_{cv} = \frac{\Sigma x_i w_i}{\Sigma w_i}, \qquad (2)$$

где x_i – производительность i-ой погрузочной машины; w_i – количество отработанных часов в месяце различными по типу i-ми погрузочными машинами;

Здесь общий объем $x_i w_i$ погруженной горной массы в месяц (сумму произведений количества рабочих часов погрузки породы в месяц на производительность погрузочной машины *соответствующего типа*) делят на суммарное количество отработанных часов.

Пример. Автомобиль в течение промежутка времени t_1 движется со скоростью υ_1, затем в течение следующего промежутка времени t_2 – со скоростью υ_2 и так далее до последнего промежутка времени t_n, в течение которого он движется со скоростью υ_n, то средняя скорость движения авто за суммарный промежуток времени ($t_1 + t_2 + \ldots + t_n$) будет равна среднему арифметическому взвешенному скоростей $\upsilon_1, \ldots, \upsilon_n$ с набором весов $t_1, \ldots, t_n$:

$$\upsilon_{\text{ср}} = \frac{\sum_{i=1}^{n} t_i \upsilon_i}{\sum_{i=1}^{n} t_i}. \qquad (3)$$

В общем случае, взвешенная средняя арифметическая – сумма произведений значения признака к частоте повторения данного признака отнесенная к сумме частот повторения всех признаков. Используется, когда варианты исследуемой совокупности встречаются неодинаковое количество раз.

Пример. Найти среднюю заработную плату рабочих цеха за месяц.

Заработная плата одного рабочего, x_i, тыс. руб.	Число рабочих, f_i в бригаде
3,2	20
3,3	35
3,4	14
4,0	6
Итого:	75

Средняя заработная плата может быть получена путем деления общей суммы заработной платы на общее число рабочих:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} + \frac{64{,}0 + 115{,}5 + 47{,}6 + 24{,}0}{20 + 35 + 14 + 6} = \frac{251{,}1}{75}. \quad (4)$$

Ответ: 3,35 тыс. руб.

1.3. Средняя арифметическая для интервального ряда

При расчете средней арифметической для интервального вариационного ряда сначала определяют среднюю *для каждого* интервала, как полу сумму верхней и нижней границ, а затем – среднюю всего ряда. В случае открытых интервалов значение нижнего или верхнего интервала определяется по величине интервалов, примыкающих к ним.

Средняя арифметическая величина для интервального вариационного ряда является *мерой средней тенденции для интервальных переменных.*

Пример. Определить средний возраст студентов вечернего отделения.

Возраст в годах	Число студентов, f	Среднее значение интервала, x'	Произведение середины интервала (возраст) на число студентов, $x' \cdot f$
до 20	65	(18 + 20) / 2 = 19 18 в данном случае граница нижнего интервала. Вычисляется как 20 – (22 – 20)	1235

Возраст в годах	Число студентов, f	Среднее значение интервала, x'	Произведение середины интервала (возраст) на число студентов, $x' \cdot f$
20–22	125	(20 + 22) / 2 = 21	2625
22–26	190	(22 + 26) / 2 = 24	4560
26–30	80	(26 + 30) / 2 = 28	2240
30 и более	40	(30 + 34) / 2 = 32	1280
Итого	500		11940

$$\bar{x} = \frac{\sum x'f}{\sum f} = \frac{11940}{500} = 23{,}9 \text{ года.}$$

Средние арифметические, вычисляемые из интервальных рядов являются приближенными. Степень их приближения зависит от того, в какой мере фактическое распределение единиц совокупности внутри интервала приближается к равномерному.

В том случае, когда все веса равны между собой, среднее арифметическое взвешенное будет равно среднему арифметическому значению.

1.4. Свойства средней арифметической величины

Эти свойства более полно раскрывают ее сущность и упрощают расчет:

1. Произведение средней на сумму частот всегда равно сумме произведений вариантов на частоты, т.е.

$$\bar{x}\sum f = \sum xf. \qquad (5)$$

2. Средняя арифметическая суммы варьирующих величин равна сумме средних арифметических этих величин:

$$\bar{x} = \frac{\sum x}{n} = \frac{\sum (y+z)}{n} = \frac{\sum y}{n} + \frac{\sum z}{n} = \bar{y} + \bar{z}. \qquad (6)$$

3. Алгебраическая сумма отклонений индивидуальных значений признака от средней арифметической равна нулю:

$$\sum (x_i - \bar{x}) f_i = 0. \qquad (7)$$

4. Сумма квадратов отклонений вариантов от средней меньше, чем сумма квадратов отклонений от любой другой произвольной величины a, т.е.:

$$\sum(x_i - \bar{x})^2 f_i \prec \sum(x_i - a)^2 f_i . \quad (8)$$

5. Если все варианты ряда уменьшить или увеличить на одно и то же число a, то средняя арифметическая уменьшится или увеличится на это же число a:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{\sum(x \mp a)f}{\sum f} \pm a . \quad (9)$$

6. Если все варианты ряда уменьшить или *увеличить в A* раз, то средняя арифметическая также уменьшится или увеличится в A раз:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{\sum \frac{x}{A} f}{\sum f} A = \frac{\sum xAf}{\sum f} : A . \quad (10)$$

7. Если все частоты (веса) увеличить или уменьшить в d раз, то средняя арифметическая не изменится:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{\sum x \frac{f}{d}}{\sum \frac{f}{d}} A = \frac{\sum xfd}{\sum fd} . \quad (11)$$

Контрольные вопросы

1. Задача любого научного исследования.
2. Измерение – метод научного исследования процесса.
3. Определение средней арифметической простой величины.
4. Определение средней арифметической взвешенной.
5. Определение средней арифметической для интервального ряда.
6. Назовите основные свойства средней арифметической величины.

2. Основы теории случайных ошибок и методов оценки случайных погрешностей в измерениях.

Одной из основных задач математической обработки результатов эксперимента является оценка истинного значения измеряемой величины по получаемым результатам, т.е. ставится задача вычисления приближенного значения, а с возможно меньшей ошибкой. Для этого надо знать основные свойства ошибок измерений.

Известно, что даже при достаточно точных измерениях одной и той же величины результаты отдельных измерений отличаются друг от друга, и, следовательно, следовательно, содержат ошибки. Ошибкой измерения называется разность $x - a$ между результатом измерения x и истинным значением a измеряемой величины. Ошибка измерения обычно неизвестна, как неизвестно и истинное значение измеряемой величины.

Для этого надо знать основные свойства ошибок измерений. Ошибки могут быть грубыми (промахи), систематическими и случайными.

Грубые ошибки возникают вследствие нарушения основных условий измерения. При обнаружении грубой ошибки результат измерения следует сразу отбросить. Внешний признак грубой ошибки – резкое отличие по величине от результатов остальных измерений.

Систематические ошибки – иногда удается выделить причины ошибок, эффект действия которых может быть рассчитан – неправильная регулировка прибора, все снятые показания будут смещены. Другой пример – изменение внешних условий – например, температуры, если известно ее влияние на результаты измерений.

Принято считать, что каждая из таких причин вызывает *систематическую ошибку.* Выявление систематических ошибок требует специальных исследований. Как только систематические ошибки обнаруживаются и их величины рассчитаны, они могут быть легко устранены путем введения соответствующих поправок.

Грубые и Систематические ошибки должны быть устранены из математической обработки.

Случайные ошибки – это ошибки, остающиеся после устранения всех выявленных систематических ошибок, т.е. ошибки ре-

зультатов измерений, исправленных путем введения соответствующих поправок называются *случайными*.

Случайные ошибки измерения вызываются большим количеством факторов, которые нельзя выделить и учесть в отдельности. Случайные ошибки это суммарный эффект действия таких факторов. Случайные ошибки измерения – неустранимы. Их влияние на оценку истинного значения измеряемой величины и определение ее значения со значительно меньшей ошибкой может быть учтено методами теории вероятностей.

Если повторять много раз измерение некоторой величины в неизменных условиях и подсчитывать число m тех результатов измерения, которые попадают в любой выделенный (отмеченный) интервал: отношение этого числа к общему числу n произведенных измерений (*относительная частота попадания в отмеченный интервал*) при достаточно большом числе измерений оказывается близким к постоянному числу (для своего интервала).

Каждому интервалу (z_1, z_2) соответствует вполне определенное число называемое *вероятностью попадания случайной величины z в этот интервал* и обозначается $P(z_1 < z < z_2)$. Практически именно к этой вероятности близки относительные частоты:

$$\frac{m}{n} \approx P(z_1 < z < z_2). \qquad (12)$$

Случайные ошибки измерения характеризуются определенным законом их распределения.

Правило, позволяющее для любых интервалов (z_1, z_2) находить вероятности $P(z_1 < z < z_2)$, называется *законом распределения вероятностей случайной величины z.*

Задача анализа случайных погрешностей – определить истинное значение измеренной величины и оценить возможные ошибки.

Основой теории случайных ошибок – являются **Предположения:**

1. При большом числе измерений случайные погрешности одинаковой величины, но разного знака встречаются одинаково часто;

2. Большие погрешности встречаются реже, чем малые (вероятность появления погрешности уменьшается с ростом ее величины);

3. При бесконечно большой выборке, истинное значение измеряемой величины равно среднеарифметическому результату.

Совокупность событий, содержащих самые различные варианты массового явления, называют *генеральной совокупностью* или *большой выборкой*.

Обычно изучается обычно лишь часть *генеральной совокупности*, которую называют *выборочной* совокупностью или *малой выборкой*.

Для выборочной совокупности число измерений «*n*» – может быть ограничено. При $n > 30$ среднее значение данной совокупности измерений «*x*» достаточно приближается к истинному значению.

2.1. Интервальная оценка точности и надежности с помощью доверительной вероятности

Для большой выборки и нормального закона распределения общей оценочной характеристики является дисперсия *D* и коэффициент вариации k_B

$$D = \sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - 1), \qquad (13)$$

где σ – *среднеквадратическое отклонение* и вычисляемое по формуле

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}; \qquad (14)$$

D – дисперсия отклонений (квадрат среднего квадратического отклонения), на использовании которой основаны практически все методы математической статистики; $k_B = \sigma / \bar{x}$ – коэффициент вариации, который характеризует изменчивость относительно среднего арифметического $\bar{x}$ – чем больше k_B тем больше изменчивы x_i относительно $\bar{x}$.

Для оценки точности и надежности попадания в заданный интервал измеренных величин *истинного* значения нужно знать вероятность попадания. Эти вероятность и интервал называются *доверительными.*

Доверительной вероятностью измерения (достоверностью) называется *вероятность того*, что *истинное* значение измеряемой величины попадает в *данный* доверительный *интервал*, т.е. в зону $a \le x_Д \le в$, эта величина определяется в долях единицы или процентах.

Доверительная вероятность $P_Д$

$$P_Д = P[a \le x_Д \le в] = 0{,}5[\varphi(в - \bar{x}) / \sigma - \varphi(a - \bar{x}) / \sigma] = \varphi(\mu / \sigma), \quad (15)$$

где $\varphi(t)$ – интегральная функция Лапласа:

$$\varphi(t) = \frac{2}{\sqrt{2\pi}} \int_0^{t_i} e^{\frac{t^2}{2}} dt. \quad (16)$$

Аргументом этой функции является *отношение* μ и среднеквадратичного отклонения σ, т.е.

$$t = \mu / \sigma\mu = \sigma t, \quad (17)$$

где t – гарантийный коэффициент.

$$\mu = (в - \bar{x}); \ \mu = -(a - \bar{x}). \quad (18)$$

Чаще всего *доверительную* вероятность принимают равной P = 0; 0,90; 0,95; 0,9973, которая означает, что только в 10, 5 и 1 случаях из 100 ошибка может выйти за установленные границы. Задавшись конкретным уровнем вероятности, выбирают величину гарантийного коэффициента t (*нормированного отклонения*) по таблице значений интеграла Лапласа (табл. 1).

Когда установлена доверительная вероятность $P_Д$, то устанавливается точность измерений (*доверительный интервал* 2μ) на основе соотношения $P_Д = \varphi(\mu / \sigma)$. Половина доверительного интервала равна

$$\mu = \sigma \arg\varphi(p_Д) = \sigma t, \quad (19)$$

где $\arg\varphi(p_Д)$ – аргумент функции Лапласа, а при $n < 30$ – функции Стьюдента.

Доверительным называется *интервал* значений x_i, в который *попадает* истинное значение $x_Д$ величины с заданной вероятностью.

Доверительный интервал характеризует точность измерения *данной выборки*, а доверительная вероятность – достоверность измерения.

Таблица 1

Интегральная функция Лапласа

t	$P_Д$	t	$P_Д$	t	$P_Д$
0,00	0,0000	0,75	0,5467	1,50	0,8664
0,05	0,0399	0,80	0,5763	1,55	0,8789
0,10	0,0797	0,85	0,6047	1,60	0,8904
0,15	0,1192	0,90	0,6319	1,65	0,9011
0,20	0,1585	0,95	0,6579	1,70	0,9109
0,25	0,1974	1,00	0,6827	1,75	0,9199
0,30	0,2357	1,05	0,7063	1,80	0,9281
0,35	0,2737	1,10	0,7287	1,85	0,9357
0,40	0,3108	1,15	0,7419	1,90	0,9426
0,45	0,3473	1,20	0,7699	1,95	0,9488
0,50	0,3829	1,25	0,7887	2,00	0,9545
0,55	0,4177	1,30	0,8064	2,25	0,9756
0,60	0,4515	1,35	0,8230	2,50	0,9876
0,65	0,4843	1,40	0,8385	3,00	0,9973
0,70	0,5161	1,45	0,8529	4,00	0,9999

Пример. Пусть, например, выполнено 30 измерений прочности дорожной одежды участка автомобильной дороги. При этом средний модуль упругости одежды составил E = 170 МПа, а вычисленное значение среднеквадратического отклонения – σ = 3,1 МПа.

Требуемую точность измерений можно определить для разных уровней доверительной вероятности ($p_Д$ = 0,9; 0,95; 0,9973), приняв значения t по табл. 1. В этом случае соответственно:

μ = ($\pm$ 3,1 · 1,65 = 5,1); ($\pm$ 3,1 · 2,0 = 6,2); ($\pm$ 3,1 · 3,0 = 9,3) МПа.

Следовательно, для данного средства и метода доверительный интервал μ возрастает примерно в два раза, если увеличить $p_Д$ только на 10 %.

Обратная задача – установить достоверность измерений для установленного доверительного интервала при μ = $\pm$ 7,0 МПа:

$$t = \mu / \sigma = 7{,}0 / 3{,}1 = 2{,}26.$$

По табл. 1 определяем $p_Д$ = 0,97. Это означает, что в заданный доверительный интервал из 100 измерений не попадают только три.

Значение $(1 - p_Д)$ называют *уровнем значимости*. Из него следует, что при нормальном законе распределения погрешность, превышающая доверительный интервал будет встречаться **один раз** из $n_И$ измерений, т.е. приходится *браковать* одно из $n_И$ измерений:

$$n_И = p_Д / (1 - p_Д). \qquad (20)$$

По данным приведенного выше примера можно вычислить количество измерений, из которых *одно* измерение *превышает* доверительный интервал. При $p_Д$ = 0,90 – $n_И$ = 9; при $p_Д$ = 0,95 – – $n_И$ = 19; при $p_Д$ = 0,9973 – $n_И$ = 367.

2.2. Определение минимального количества измерений

Для проведения опытов *с заданной* точностью и *достоверностью* необходимо знать то количество измерений, при котором экспериментатор уверен в положительном исходе.

В связи с этим нужно установить *минимальное, но достаточное* количество измерений. Задача сводится к установлению минимального объема выборки (числа измерений) N_{min} при заданных значениях доверительного интервала 2μ и доверительной вероятности.

При выполнении измерений необходимо знать их точность

$$\Delta = \sigma_0 / \bar{x}, \qquad (21)$$

где σ_0 – среднеарифметическое значение среднеквадратичного отклонения $\sigma_0 = \sigma / \sqrt{n}$.

Значение σ_0 часто называют *средней ошибкой*.

Доверительный *интервал ошибки* измерения Δ определяется также как для измерений $\mu = t \cdot \sigma_0$, с помощью t легко определить доверительную вероятность ошибки измерений из табл. 1.

В исследованиях часто по заданной точности Δ и доверительной вероятности измерения определяют минимальное количество измерений, гарантирующих требуемые значения Δ и $p_Д$.

$$\mu = \sigma \arg\varphi(p_Д) = \frac{\sigma_0}{\sqrt{n}} t . \qquad (22)$$

При $N_{min} = n$ получаем

$$N_{min} = \frac{\sigma^2 t^2}{\sigma_0^2} = \frac{k_в^2 t^2}{\Delta^2}, \qquad (23)$$

где $k_в^2$ – коэффициент вариации (переменчивость), %; Δ – точность измерений, %.

Для определения N_{min} может быть принята следующая последовательность вычислений:

1. Проводится предварительный эксперимент с количеством измерений n, которые составляют в зависимости от трудоемкости опыта 25-50 шт.

2. Вычисляется среднеквадратическое отклонение σ по формуле (14).

3. В соответствии с поставленными задачами эксперимента устанавливается требуемая точность измерений Δ, которая не должна превышать точности прибора.

4. Задается доверительная вероятность и устанавливается нормируемое отклонение t, значение которого обычно задается (зависит так же от точности метода).

5. Определяется N_{min} и тогда в дальнейшем в процессе эксперимента число измерений не должно быть меньше N_{min}.

Для нахождения границы доверительного интервала при малых значениях n применяют метод предложенный в 1908 г. английским математиком В. Госсетом (псевд. Стьюдент). Кривые распределения Стьюдента в случае $n \to \infty$ (практически при $n > 20$ переходят в кривые нормального распределения, рис. 1.

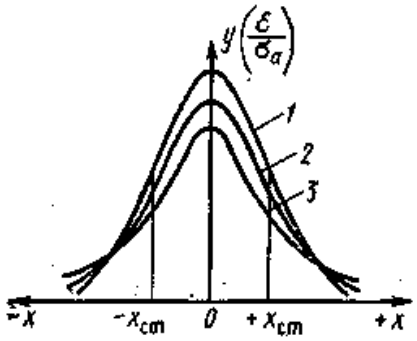


Рис. 1. Кривые распределения Стьюдента для различных значений:
1 – $n \to \infty$; 2 – $n = 10$; 3 – $n = 2$

1) Для малой выборки доверительный интервал

$$\mu_{СТ} = \sigma_0 \cdot \alpha_{СТ}, \qquad (24)$$

где $\alpha_{СТ}$ – коэффициент Стьюдента, принимаемый по табл. 2 в зависимости от значения заданной доверительной вероятности $p_Д$.

Зная $\mu_{СТ}$, можно вычислить действительное значение изучаемой величины для малой выборки

$$x_Д = \bar{x} \pm \mu_{СТ}. \qquad (25)$$

Таблица 2

Коэффициент Стьюдента $\alpha_{СТ}$

n	$p_Д$					
	0,80	0,90	0,95	0,99	0,995	0,999
2	3,080	6,31	12,71	63,70	127,30	637,20
3	1,886	2,92	4,30	9,92	14,10	31,60
4	1,638	2,35	3,188	5,84	7,50	12,94
5	1,533	2,13	2,77	4,60	5,60	8,61
6	1,476	2,02	2,57	4,03	4,77	6,86
7	1,440	1,94	2,45	3,71	4,32	6,13
8	1,415	1,90	2,36	3,50	4,03	5,40
9	1,397	1,86	2,31	3,36	3,83	5,04
10	1,383	1,83	2,26	3,25	3,69	4,78
12	1,363	1,80	2,20	3,11	3,50	4,49
14	1,350	1,77	2,16	3,01	3,37	4,22
16	1,341	1,75	2,13	2,95	3,29	4,07
18	1,333	1,74	2,11	2,90	3,22	3,96
20	1,328	1,73	2,09	2,86	3,17	3,88
30	1,316	1,70	2,04	2,75	3,20	3,65
40	1,306	1,68	2,02	2,70	3,12	3,55
50	1,298	1,68	2,01	2,68	3,09	3,50
60	1,290	1,67	2,00	2,66	3,06	3,46
∞	1,282	1,64	1,96	2,58	2,81	3,29

2) *Возможна и иная постановка задачи*. По n известных измерений малой выборки необходимо определить доверительную вероятность $p_Д$ при условии, что погрешность среднего значения не выйдет за пределы $\pm \mu_{СТ}$.

Задачу решают в такой последовательности:

– в начале вычисляется среднее значение $\bar{x}$, σ_0 и $\alpha_{CT} = \mu_{CT} / \sigma_0$; $\sigma_0 = \sigma / \sqrt{n}$;

– затем с помощью величины α_{CT}, известного n и табл. 2 определяют доверительную вероятность $p_Д$.

В процессе обработки экспериментальных данных следует исключать грубые ошибки ряда. Появление этих ошибок вполне вероятно, а в наличие их ощутимо влияет на результат измерений. Однако прежде чем исключить то или иное измерение, необходимо убедиться, *что это действительно грубая ошибка, а не отклонение вследствие статистического разброса*. Известно несколько методов определения грубых ошибок статистического ряда.

Первый метод. Наиболее простым способом исключения из ряда резко выделяющихся измерений является правило трех сигм: разброс случайных величин от среднего значения не должен превышать:

$$X_{max,\,min} = \bar{x} \pm 3\sigma. \qquad (26)$$

Более достоверными являются методы, которые базируются на использовании доверительного интервала.

Пусть имеется статистический ряд малой выборки, подчиняющийся закону нормального распределения. При наличии грубых ошибок критерии их появления вычисляются по формулам:

$$\beta_1 = (x_{max} - \bar{x}) / \sigma\sqrt{\frac{n-1}{n}}; \qquad (27)$$

$$\beta_2 = (\bar{x} - x_{min}) / \sigma\sqrt{\frac{n-1}{n}}, \qquad (28)$$

где x_{max}, x_{min} – наибольшее и наименьшее значение из n измерений.

Составлена табл. 3 критерия появления грубых ошибок β_{max} в зависимости от доверительной вероятности $p_Д$ и количества измерений (n = 50; 1,41 ≤ β ≤ 2,8 – 3,37 при $p_Д$ = 0,999).

Если $\beta_1 > \beta_{max}$, то значение x_{max} необходимо исключить из статистического ряда, как грубую ошибку. При $\beta_2 < \beta_{max}$ исключается величина x_{min}.

После исключения грубых ошибок определяют новые значения x и σ из (n – 1) или (n – 2) измерений.

Второй метод. При анализе измерений применяется для приближенной оценки по следующей методике:

– вычислить с помощью формулы (14) среднеквадратическое отклонение σ;

– определить по формуле (21) среднеарифметическое значение среднеквадратического отклонения σ_0;

– принять доверительную вероятность $p_Д$ и найти доверительные интервалы $\mu_{СТ}$;

– окончательно установить действительные значения измеряемой величины $x_Д$ по формуле (9) – $x_Д = \bar{x} \pm \mu_{СТ}$.

Третий метод. В случае более глубокого анализа экспериментальных данных рекомендуется такая последовательность:

1) После получения экспериментальных данных в виде статистического ряда его анализируют и исключают систематические ошибки.

Таблица 3

Критерий появления грубых ошибок

n	β_{max} при $p_Д$			n	β_{max} при $p_Д$		
	0,90	0,95	0,99		0,90	0,95	0,99
3	1,41	1,41	1,41	15	2,33	2,49	2,80
4	1,64	1,69	1,72	16	2,35	2,52	2,84
5	1,79	1,87	1,96	17	2,38	2,55	2,87
6	1,89	2,00	2,13	18	2,40	2,58	2,90
7	1,97	2,09	2,26	19	2,43	2,60	2,93
8	2,04	2,17	2,37	20	2,45	2,62	2,96
9	2,10	2,24	2,46	25	2,54	2,72	3,07
10	2,15	2,29	2,54	30	2,61	2,79	3,16
11	2,19	2,34	2,61	35	2,67	2,85	3,22
12	2,23	2,39	2,66	40	2,72	2,90	3,28
13	2,26	2,43	2,71	45	2,76	2,95	3,33
14	2,30	2,46	2,76	50	2,80	2,99	3,37

2) Анализируют ряд в целях обнаружения грубых промахов и ошибок: устанавливают подозрительные значения x_{max} и x_{min}; определяют среднеквадратическое отклонение σ, вычисляют критерии β_1 и β_2 и сопоставляют с β_{max} и β_{min} исключают при необходимости из статистического ряда x_{max} или x_{min} и получают новый ряд из новых членов.

3) Вычисляют среднеарифметическое $\bar{x}$, погрешность отдельных измерений ($\bar{x} - x_i$) и среднеквадратическое очищенного ряда σ.

4) Находят среднеквадратическое σ_0 серии измерений, коэффициент вариации $k_В$.

5) При большой выборке задаются доверительной вероятностью $p_Д = \varphi(t)$ или уравнением значимости $(1 - p_Д)$ и по табл. 1 определяется (t).

6) При малой выборке ($n \leq 30$) в зависимости от принятой $p_Д$ и числа членов ряда принимают коэффициент Стьюдента $\alpha_{СТ}$; далее определяют доверительный интервал.

7) Устанавливают действительное значение исследуемой величины – $x_Д = \bar{x} \pm \mu_{СТ.}$

8) Оценивают относительную погрешность, % результатов серии измерений при заданной доверительной вероятности $p_Д$:

$$\delta = 100\ \delta_0\ \alpha_{СТ} / \bar{x}, \%. \qquad (29)$$

Во многих случаях в процессе экспериментальных исследований приходится иметь дело с косвенными измерениями. При этом в расчетах применяют те или иные функциональные зависимости типа:

$$y = f(x_1, x_2, \ldots, x_n). \qquad (30)$$

Так как в данную функцию подставляют не истинные, а приближенные значения, то и окончательный результат также будет приближенным. В связи с этим одной из основных задач теории случайных ошибок является определение ошибки функции, если известны ошибки их аргументов.

При исследовании функции одного переменного предельные абсолютные $\varepsilon_{ПР}$ и относительные $\delta_{ПР}$ ошибки (погрешности) вычисляют так:

$$\varepsilon_{ПР} = \pm\ \varepsilon_X\ f'(x); \qquad (31)$$

$$\delta_{ПР} = \pm\ d \cdot \ell n(x), \qquad (32)$$

где $f'(x)$ – производная функции $f(x)$; $d \cdot \ell n(x)$ – дифференциал натурального логарифма функции

Если исследуется функция многих переменных, то

$$\varepsilon_{ПР} = \pm \sum_{1}^{n} \left| \frac{\partial f(x_1, x_2, ..., x_i)}{\partial x_i} \mathrm{d}x_i \right|; \qquad (33)$$

$$\delta_{ПР} = \pm d \mid \ell \mathrm{n}\,(x_1, x_2, \ldots, x_n)|. \qquad (34)$$

В эти выражения под знаком суммы и дифференциала принимаются абсолютные значения.

Методика определения ошибок с помощью этих уравнений включает следующее:

– вначале определяют абсолютные и относительные ошибки аргументов (независимых переменных). Обычно величина $x_Д \pm \varepsilon$ каждого переменного измерена, следовательно, абсолютные ошибки для аргументов известны, т.е. $\varepsilon_{X1}, \varepsilon_{X2}, \ldots, \varepsilon_{Xn}$. Затем вычисляют относительные ошибки независимых переменных.

$$\delta_{X1}, = \varepsilon_{X1} / x_Д, \delta_{X2}, = \varepsilon_{X2} / x_Д, \ldots, \delta_{Xn}, = \varepsilon_{Xn} / x_Д. \qquad (35)$$

Находят частные дифференциальные функции и по формуле (33) вычисляют $\varepsilon_{ПР}$ в размерностях функции $f(y)$ и с помощью (34) вычисляют $\delta_{ПР}$ (%).

В исследованиях часто возникает вопрос о достоверности данных, полученных в опытах.

Решение такой задачи можно проиллюстрировать примером.

Пусть установлена прочность контрольных образцов бетона до виброперемешивания $R_1 = \bar{R}_1 \pm \sigma_0 = 20 \pm 0{,}5$ МПа и прочность бетонных образцов после виброперемешивания $R_2 = \bar{R}_2 \pm \sigma_0 = 23 \pm \pm 0{,}6$ МПа. Прирост прочности составляет 15 %. Это упрочнение относительно небольшое, его можно отнести за счет разброса опытных данных. В этом случае следует провести проверку на достоверность экспериментальных данных по правилу 3-х сигм:

Отклонение истинного значения измеряемой величины от среднего арифметического значения результатов измерений не превосходит утроенной средней квадратической ошибки этого среднего значения.

$$\frac{\bar{x}}{\sigma_1} \geq 3 \qquad (36)$$

В данном случае проверяется разница $\bar{x} = R_1 - R_2 = 3$ МПа. Ошибка измерения равна $\sigma_0 = \sqrt{\sigma_1^2 + \sigma_2^2}$, поэтому

$$(R_1 - R_2) / \sqrt{\sigma_1^2 + \sigma_2^2} = 3{,}0 / (0{,}25 + 0{,}36) \cdot 3{,}84 > 3.$$

Следовательно, полученный прирост прочности является достоверным.

Контрольные вопросы

1. Основная задача математической обработки результатов эксперимента.
2. Основные свойства ошибок измерений.
3. Основа теории случайных ошибок. Назовите основные предположения.
4. Типы ошибок, характеристика.
5. Определение средней арифметической для интервального ряда.
6. Вероятность попадания случайной величины z в интервал.
7. Доверительная вероятность.
8. Интервальная оценка точности и надежности с помощью доверительной вероятности.
9. Определение минимального количества измерений.

3. Методы графической обработки результатов измерений

Для графического изображения результатов измерений (наблюдений), как правило, применяют систему прямоугольных координат. Если анализируется методом функция $y = f(x)$, то наносят в системе прямоугольных координат значения x_1, y_1, x_2, y_2, … , x_n, y_n (рис. 2).

Прежде чем строить график необходимо знать ход (течение) исследуемого явления. Как правило, качественные закономерности и форма графика экспериментатору ориентировочно известны из теоретических исследований.

Точки на графике необходимо соединять плавной кривой так, чтобы она по возможности проходила ближе ко всем экспериментальным точкам. Если соединить точки прямыми отрезка-

ми, то получим ломаную кривую. Она характеризует изменение функции по данным эксперимента.

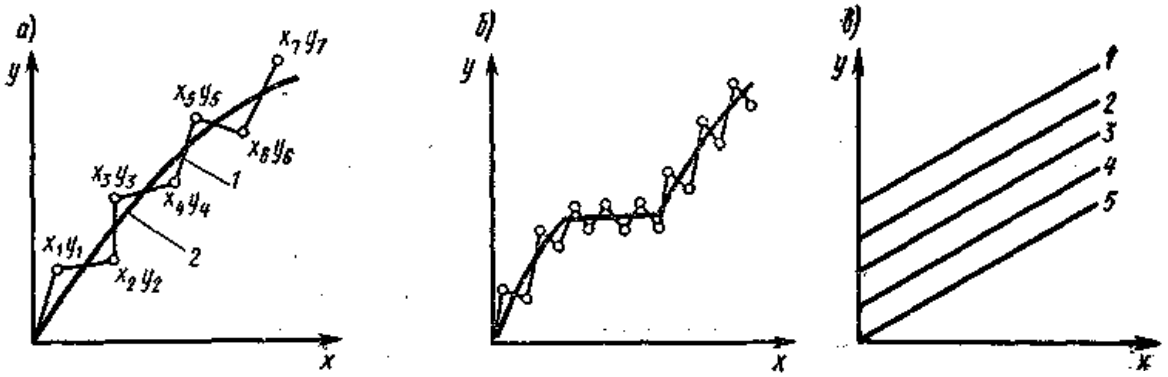


Рис. 2. Графическое изображение функции $y = f(x)$:
а) плавная зависимость: 1 – кривая по результатам непосредственных измерений; 2 – плавная кривая;
б) при наличии скачка; в) при трех переменных:
z_5 = const; z_4 = const; z_3 = const; z_2 = const; z_1 = const

Обычно функции имеют плавный характер. Поэтому при графическом изображении результатов измерений следует проводить между точками плавные кривые. Резкое искривление графика объясняется погрешностями измерений. Если бы эксперимент повторили с применением средств измерений более высокой точности, то получили бы меньшие погрешности, а ломаная кривая больше бы соответствовала плавной кривой.

Однако могут быть и исключения, так как иногда исследуются явления, для которых в определенных интервалах наблюдается быстрое скачкообразное изменение одной из координат (см. рис 2, б). Это объясняется сущностью физико-химических процессов, например, фазовыми превращения влаги, радиоактивным распадом атомов в процессе исследования радиоактивности и т.п.

В таких случаях необходимо особо тщательно соединять точки кривой. Общее «осреднение» всех точек плавной кривой может привести к тому, что скачек функции подменится якобы погрешностями измерений.

Иногда при построении графика одна – две точки резко удаляются от кривой. В таких случаях вначале следует проанализировать физическую сущность явления, и, если нет основания по-

лагать наличие скачка функции, то такое рез кое отклонение можно объяснить грубой ошибкой или промахом. Это может возникнуть тогда, когда данные измерений предварительно не исследовались на наличие грубых ошибок измерений. Нужно повторить опыт.

Часто при графическом изображении результатов экспериментов приходится иметь дело с тремя переменными $в = f(x, y, z)$. В этом случае применяют метод разделения переменных. Одной из величин z в пределах интервала измерений ($z_1 – z_n$) задают несколько последовательных значений. Для двух остальных переменных x и y строят графики $x, y = f_1(x)$ при z = const. В результате на одном графике (см. рис 2, в) получают семейство кривых , $y = f_1(x)$ для различных значений z. Если необходимо графически изобразить функцию с четырьмя переменными и более $\alpha = f(в, x, y, z)$, строят серию типа предыдущей, по каждой из них при $в_1, в_2, \ldots , в_n$ = const или принимается из N переменных $(N – 1)$ постоянными и строят графики: вначале $(N – 1) = f_1(x)$, далее $(N – 2) = f_2(x)$, $(N – 3) = f_3(x)$ и т.д. Таким образом, можно проследить изменения любой переменной величины в функции от другой при постоянных значениях остальных. Этот метод графического анализа требует тщательности, большого внимания к результатам измерений. Однако он в большинстве случаев является наиболее простым и наглядным.

При графическом изображении результатов эксперимента большую роль играет выбор системы координат или координатной сетки. Координатные сетки (рис. 3) бывают равномерными и неравномерными. У равномерных координатных сеток ординаты и абсциссы имеют равномерную шкалу.

Например, в системе прямоугольных координат длина, откладываемых единичных отрезков на обеих осях одинаковая. Из неравномерных координатных сеток наиболее распространены полулогарифмические, логарифмические, вероятностные. Полулогарифмическая сетка имеет равномерную ординату и логарифмическую абсциссу.

Логарифмическая координатная сетка имеет обе координатные оси логарифмические. Вероятностная сетка имеет обычно ординату равномерную, а по абсциссе вероятностную шкалу.

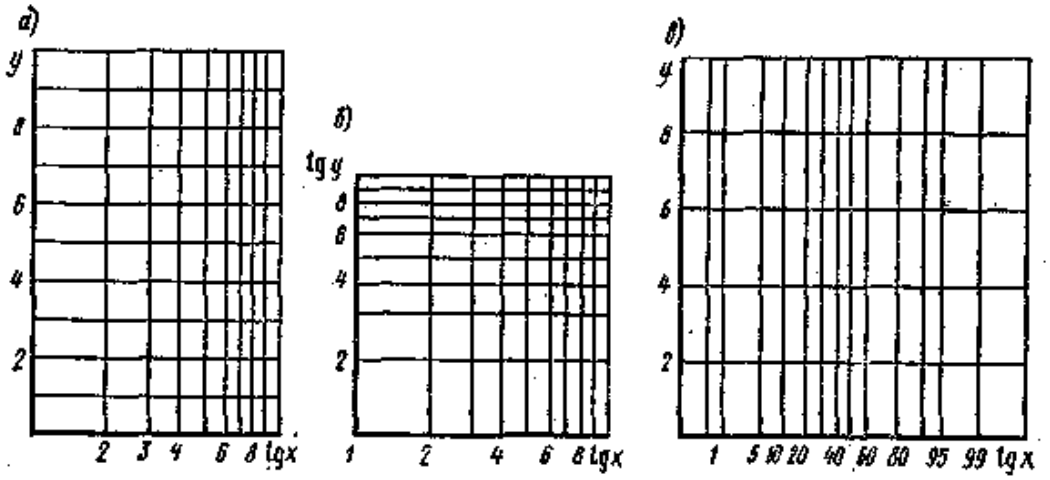


Рис. 3. Координатная сетка: а) полулогарифмическая;
б) логарифмическая; в) вероятностная

Назначение неравномерных сеток различное – в большинстве случаев их применяют для более наглядного изображения функции (и *если нужна контрастность*).

Функция $y = f(x)$ имеет различную форму при различных сетках. Так, многие криволинейные функции спрямляются на логарифмических сетках.

Большое значение в практике графического изображения экспериментальных данных вероятностная сетка, применяется в различных случаях – при обработке измерений для оценки точности, при определении расчетных характеристик (расчетной влажности, расчетных значений модуля упругости, межремонтных сроков службы).

Масштаб по координатным осям обычно применяют различный. От выбора его зависит форма графика – он может быть плоским (узким) или вытянутым (широким) вдоль оси (рис. 4).

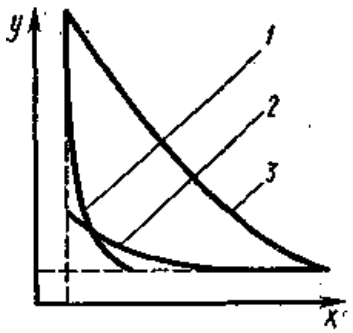


Рис. 4. Форма графика в зависимости от масштаба:
1 – плоская; 2 – уширенная; 3 – нормальная

В некоторых случаях строят номограммы, существенно облегчающие применение для систематических расчетов сложных теоретических или эмпирических формул в определенных пределах измерения величин.

Контрольные вопросы

1. Какая система координат используется для графического изображения результатов наблюдений?
2. Как следует относиться, когда в определенных интервалах наблюдается быстрое скачкообразное изменение одной из координат?
3. Какой метод используется когда, при графическом изображении результатов экспериментов участвуют три и более переменных?
4. Какую координатную сетку для достижения контрастности при графическом изображении результатов эксперимента?

4. Методы подбора эмпирических формул

В процессе экспериментальных исследований получается статистический ряд измерений двух величин, когда каждому значению функции $y_1, y_2, \ldots, y_n$, соответствует определенное значение аргумента $x_1, x_2, \ldots, x_n$.

На основе экспериментальных данных можно подобрать алгебраическое выражение функции $y = f(x)$, которые называют эмпирическими формулами. Такие формулы подбираются лишь в пределах измеренных значений аргумента $x_1 - x_n$. И имеют тем большую ценность, чем больше соответствуют результатам эксперимента.

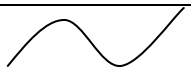
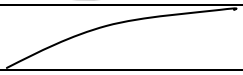
Замену точных аналитических выражений приближенным, более простым называют аппроксимацией, а функции – аппроксимирующими.

Процесс подбора эмпирических формул состоит из двух этапов:

I этап. Данные измерений наносят на сетку прямоугольных координат, соединяют экспериментальные точки плавной кривой и выбирают ориентировочно вид формулы (табл. 4).

Таблица 4

Виды математических функций, используемые при выравнивании

Название функции	Вид функции	Формула
Прямая линия		$\hat{y}_t = a_0 + a_1 t$
Парабола 2-го порядка	 или	$\hat{y}_t = a_0 + a_1 t + a_2 t^2$
Парабола 3-го порядка		$\hat{y}_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3$
Гипербола		$\hat{y}_t = a_0 + \frac{a_1}{t}$
Показательная		$\hat{y}_t = a_0 a_1^t$
Степенная		$\hat{y}_t = a_0 t^{a_1}$
Ряд Фурье		$\hat{y}_t = a_0 + \sum_{k=1}^{m} (a_k \cos kt + b_k \sin kt)$

II этап. Вычисляют параметры формул, которые наилучшим образом соответствовали бы принятой формуле. Подбор эмпирических формул необходимо начинать с самых простых выражений. Так, например, результаты измерений многих явлений и процессов аппроксимируют простейшими эмпирическими уравнениями типа

$$y = a + вx, \qquad (37)$$

где *а, в* – постоянные коэффициенты.

Поэтому при анализе графического материала необходимо по возможности стремиться к использованию линейной функции. Для этого применяют метод выравнивания, заключающийся в том, что кривую, построенную по экспериментальным точкам, представляют линейной функцией.

Для преобразования некоторой кривой в прямую линию вводят новые переменные

$$X = f_1(x, y);\ Y = f_2(x, y). \qquad (38)$$

В искомом уравнении они должны быть связаны линейной зависимостью

$$Y = a + вx. \qquad (39)$$

Значения X и Y можно вычислить на основе решения системы уравнений (38). Далее строят прямую (рис. 5), по которой легко графически вычислить параметры a (ордината точки пересечения прямой с осью Y) и $в$ (тангенс угла наклона прямой с осью X).

$$в = \text{tg}\alpha = (Y_i - a) / x_i. \qquad (40)$$

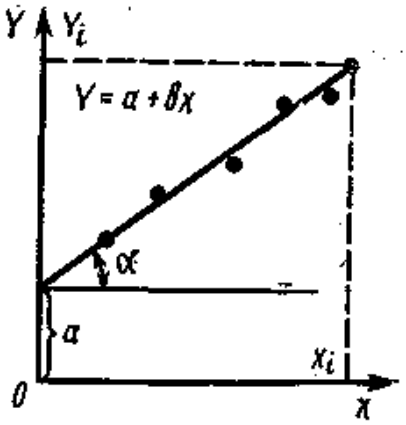


Рис. 5. Графическое определение параметров x и y

При графическом определении параметров a и $в$ обязательно, чтобы прямая строилась на координатной сетке, у которой началом является точка $X = 0$ и $Y = 0$. Для расчета необходимо точки X_i и Y_i принимать на крайних участках прямой.

Пример. Подобрать эмпирическую формулу для следующих измерений:

x	12,1	19,2	25,9	33,3	40,5	46,4	54,0
y	1	2	3	4	5	6	7

Графический анализ этих измерений показывает, что в прямоугольных координатах точки хорошо ложатся на прямую линию и их можно выразить зависимостью $Y = a + вx$.

Выбираем координаты крайних точек и подставляем в два уравнения и решаем их вычитанием, т.е. $A_0 + 7A_1 = 54{,}0$; и $A_0 + A_1 = 12{,}1$. Отсюда $A_1 = 41{,}9 / 6 = 6{,}98$ (= $в$); A_0 (a) = 12,1 – – 6,98 = = 5,12 (= a).

Эмпирическая формула примет вид $Y = 5{,}12 + 6{,}98X_1$.

Графический метод выравнивания может быть применен в тех случаях, когда экспериментальная кривая на сетке прямоугольных координат имеет вид плавной кривой.

а) Если экспериментальный график имеет вид (рис. 6), то необходимо применить формулу

$$y = ax^{в}. \qquad (41)$$

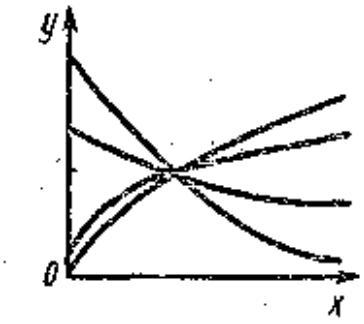


Рис. 6. Вид графика эмпирической формулы $y = ax^{в}$

Заменяя $X = \lg x$ и $Y = \lg y$ получим

$$Y = \lg a + вx. \qquad (42)$$

При этом экспериментальная кривая превращается в прямую на логарифмической сетке.

б) Если экспериментальный график имеет вид, рис. 7, то целесообразно использовать выражение

$$y = ae^{вx}. \qquad (43)$$

При замене $Y = \lg y$ получим

$$Y = \lg a + вx. \qquad (44)$$

Здесь экспериментальная кривая превращается в прямую на полулогарифмической сетке.

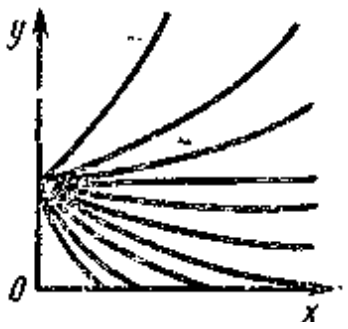


Рис. 7. Вид графика эмпирической формулы $y = ae^{вx}$

в) Если экспериментальный график имеет вид, рис. 8, то целесообразна эмпирическая формула

$$y = c + ax^{в}. \qquad (45)$$

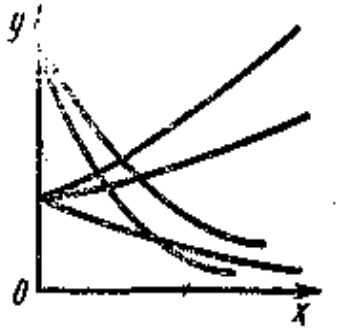


Рис. 8. Вид графика эмпирической формулы $y = c + ax^{в}$

1) Если $в$ задано, то надо принять $X = x^{в}$ и тогда получим прямую линию на сетке прямоугольных координат

$$y = c + aX. \qquad (46)$$

2) Если же $в$ неизвестно, то надо принять $X = \lg x$ и $Y = \lg(y - c)$. В этом случае будет тоже прямая, но на логарифмической сетке.

В последнем случае необходимо предварительно вычислить c. Для этого на экспериментальной кривой принимают три произвольные точки x_1, y_1; x_2, y_2; $x_3 = \sqrt{x_1 x_2}$, y_3 и вычисляют c в виде отношения

$$c = \frac{y_1 y_2 - y_3^2}{y_1 + y_2 - 2y}. \qquad (47)$$

$г$) Если экспериментальный график имеет вид (рис. 9), то нужно пользоваться формулой

$$Y = c + ae^{вx}. \qquad (48)$$

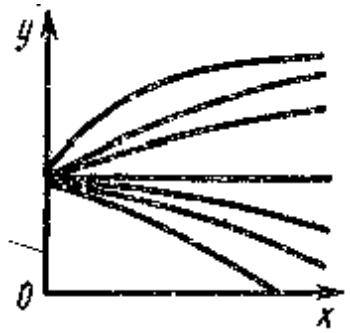


Рис. 9. Вид графика эмпирической формулы $y = c + ae^{вx}$

Путем замены $Y = \lg(y - c)$ можно построить прямую на полулогарифмической сетке

$$Y = \lg a + вx \cdot \lg c, \qquad (49)$$

где c – предварительно определено по формуле (47). В этом случае

$$x_3 = (x_1 + x_2) / 2. \qquad (50)$$

д) Если экспериментальный график имеет вид (рис. 10), применяют выражение

$$Y = a + в / x. \qquad (51)$$

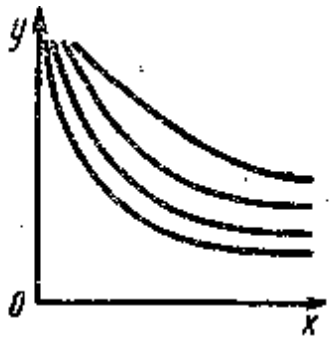


Рис. 10. Вид графика эмпирической формулы $y = a + b / xae^{вx}$

Путем замены $x = 1 / z$ можно получить прямую линию на сетке прямоугольных координат

$$Y = a + вz. \qquad (52)$$

е) Если экспериментальный график имеет вид, рис. 11, то используют формулу

$$y = 1 / (a + вx). \qquad (53)$$

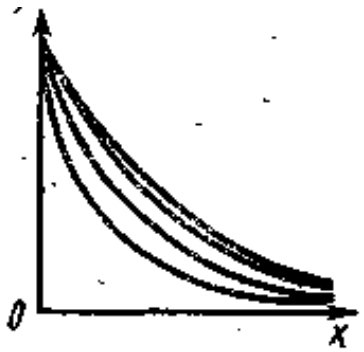


Рис. 11. Вид графика эмпирической формулы $y = c + ae^{nx + m2}$

Если принять $y = 1 / z$, то $z = a + вx$ – прямая на сетке прямоугольных координат.

Аналогично уравнению

$$y = \frac{1}{a+вx+cx^2}. \quad (54)$$

Путем замены $y = 1 / z$ можно придать вид

$$z = a + вx + c^2. \quad (55)$$

ж) Сложную степенную функцию

$$Y = c + ae^{nx + mx2} \quad (56)$$

можно преобразовать в простую, при $\lg y = z$, $\lg a = p$ и $\lg e = q$; $m \cdot \lg e = r$ получается зависимость вид $z = p + qx + rx^2$.

С помощью графиков и выражений можно практически всегда подобрать уравнение эмпирической формулы.

Пример. Подобрать эмпирическую формулу по следующим результатам измерений.

1	1,5	2,0	2,5	3,0	3,5	4,0	4,5
15,2	20,6	27,4	36,7	49,2	66,0	87,4	117,5

График на основе этих данных имеет вид, что. соответствует кривым (рис. 12, а).

$$y = ae^{вx}. \quad (57)$$

После логарифмирования выражения $\lg y = \lg a + вx \cdot \lg e$, если обозначить $\lg y = Y$, то

$$Y = \lg a + вx \cdot \lg e, \quad (58)$$

т.е. в полулогарифмических координатах выражение для Y представляет собой прямую линию (рис. 12, б).

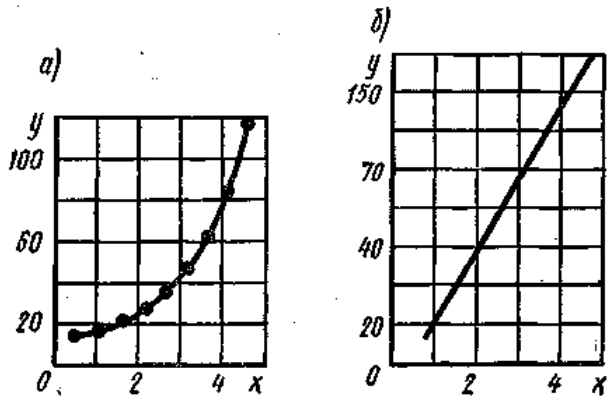


Рис. 12. Подбор эмпирической характеристики:
а) эмпирическая; б) спрямленная

Подстановка в уравнение координат крайних точек дает $\lg 15{,}2 = \lg a + в \cdot \lg e$ и $\lg 117{,}5 = \lg a + 4{,}5в \cdot \lg e$.

Следовательно, 1) $\lg a + в \cdot \lg e = 1{,}183 \cdot b = 0{,}887 \,/\, (3{,}5 \cdot \lg e)$, $b = 0{,}579$.

2) $\lg a + 4{,}5в \cdot \lg e = 2{,}07 \lg a = 1{,}183 - 0{,}254 = 0{,}929$, $a = 1{,}85$.

Окончательно эмпирическая формула имеет вид $y = 1{,}85\, e^{0{,}579x}$.

Контрольные вопросы

1. Эмпирические формулы. Что они отражают?
2. Виды математических функций, используемых при выравнивании.
3. Варианты применения графического метода выравнивания.
4. Этапы вычисления параметров формул.

5. Метод средних квадратов

При подборе эмпирических формул широко используются полиномы

$$y = A_0 + A_1 \cdot x + A_2 \cdot x^2 + A_3 \cdot x^3 + \ldots + A_n \cdot x^n, \qquad (59)$$

где $A_0, A_1, \ldots, A_n$ – постоянные коэффициенты.

Полиномами можно аппроксимировать любые результаты измерений, если они графически выражаются непрерывными функциями. Особо ценным является то, что даже при неизвестном точном выражении функции (59) можно определить значение коэффициентов A. Для определения коэффициентов A кроме графического метода, изложенного выше, применяют методы средних и наименьших квадратов.

Метод средних квадратов основан на следующем положении: по экспериментальным точкам можно построить несколько плавных кривых. Наилучшей будет та кривая, у которой разностные отклонения оказываются наименьшими, т.е. $\sum \varepsilon \approx 0$. Порядок расчета полинома сводится к следующему:

1) Определяется число членов ряда, которое обычно принимают не более 3-4.

2) В принятое выражение последовательно подставляют координаты x и y нескольких m экспериментальных точек.

3) И получают систему из (m) уравнений. Каждое уравнение приравнивается соответствующему отклонению.

$$A_0 + A_1 \cdot x_1 + A_2 \cdot x_1^2 + A_3 \cdot x_1^3 + \ldots + A_n \cdot x_1^n - y_1 = \varepsilon_1;$$

$$A_0 + A_1 \cdot x_2 + A_2 \cdot x_2^2 + A_3 \cdot x_2^3 + \ldots + A_n \cdot x_2^n - y_2 = \varepsilon_2;$$

$$\ldots$$

$$A_0 + A_1 \cdot x_m + A_2 \cdot x_m^2 + A_3 \cdot x_m^3 + \ldots + A_n \cdot x_m^n - y_m = \varepsilon_m. \qquad (60)$$

Число точек, т.е. число уравнений, должно быть не меньше числа коэффициентов A, что позволит их вычислить путем решения системы (60).

Разбивают систему начальных уравнений (60) последовательно сверху вниз на группы, число которых должно быть равно количеству коэффициентов A_0. В каждой группе складывают уравнения и получают новую систему уравнений, равную количеству групп (обычно 2-3) Решая систему, вычисляют коэффициенты A.

Метод средних квадратов обладает высокой точностью, если число точек достаточно велико (не менее 3-4). Однако степень точности можно повысить, если начальные условия сгруппировать по 2-3 варианта и вычислить для каждого варианта эмпирическую формулу. Предпочтение следует отдать той формуле, у которой $\sum\varepsilon^2 = \min$.

Для подбора эмпирической формулы можно выбрать полином типа

$$y = A_0 + A_1 \cdot x + A_2 \cdot x^2. \qquad (61)$$

Пример. Выполнено 7 измерений.

4	5	6	7	8	9	10
10,2	6,7	4,8	3,6	2,7	2,1	1,7

Путем подстановки в это уравнение значений измерений систему начальных уравнений можно разделить на три группы: 1-2, 3-4, 5-7 в виде

1. $A_0 + 4 \cdot A_1 + 16 \cdot A_2 = 10{,}2$.
2. $A_0 + 5 \cdot A_1 + 25 \cdot A_2 = 6{,}7$.

3. $A_0 + 6\cdot A_1 + 36\cdot A_2 = 4{,}8$.
4. $A_0 + 7\cdot A_1 + 49\cdot A_2 = 10{,}2$.
5. $A_0 + 8\cdot A_1 + 64\cdot A_2 = 3{,}6$.
6. $A_0 + 9\cdot A_1 + 81\cdot A_2 = 2{,}7$.
7. $A_0 + 10\cdot A_1 + 100\cdot A_2 = 1{,}7$.

Сложение уравнений в каждой группе даст

1-я гр. $2\cdot A_0 + 9\cdot A_1 + 41\cdot A_2 = 16{,}9$.
2-я гр. $2\cdot A_0 + 13\cdot A_1 + 85\cdot A_2 = 8{,}4$.
3-я гр. $3\cdot A_0 + 27\cdot A_1 + 245\cdot A_2 = 6{,}5$.

Определение из этих выражений коэффициентов A_0, A_1, A_2 приводит к эмпирической формуле

$$y = 26{,}168 - 5{,}2168\cdot x + 0{,}2811\cdot x^2. \qquad (62)$$

Метод средних квадратов может быть применен для различных кривых после их выравнивания.

Пример. Выполнено 8 измерений

3	6	9	12	15	18	21	24
57,6	41,9	31,0	22,7	16,6	12,2	8,9	6,5

Анализ кривой в системе прямоугольных координат дает возможность применить $y = ae^{-вx}$.

Произведем выравнивание путем замены переменных обозначить $Y = \lg y$, $X = \dfrac{x}{2{,}303}$. Тогда $Y = A + B\cdot X$, где $A = \lg a$, $B = в$, так как необходимо определить два параметра, то все измерения делятся на две группы по четыре измерения. Это приводит к уравнениям:

$$1{,}7604 = A + \frac{3}{2{,}303}B;\ 1{,}2201 = A + \frac{15}{2{,}303}B;\ \{2{,}303=(1 / \lg e)\};$$

$$1{,}6222 = A + \frac{6}{2{,}303}B;\ 1{,}0864 = A + \frac{18}{2{,}303}B;$$

$$1{,}4914 = A + \frac{9}{2{,}303}B;\ 0{,}9494 = A + \frac{21}{2{,}303}B;$$

$$1{,}3560 = A + \frac{12}{2{,}303}B;\ 0{,}8129 = A + \frac{24}{2{,}303}B;$$

$$6{,}2300 = 4A + \frac{30}{2{,}303}B;\ 4{,}0688 = 4A + \frac{78}{2{,}303}B.$$

После суммирования по группам можно получить систему двух уравнений с двумя неизвестными A и B, решение которых дает $A = 1{,}8952$, $a = 78{,}56$, $B = -0{,}1037$, $в = -0{,}1037$. Окончательно $y = 78{,}56 \cdot e^{-0{,}1037x}$.

Хорошие результаты при определении параметров заданного уравнения даем использование метода наименьших квадратов. Суть этого метода заключается в том, что если все измерения функции $y_1, y_2, \ldots, y_n$ произведены с одинаковой точностью и распределенные величины ошибок измерения соответствуют нормальному закону, то параметры исследуемого уравнения определяется из условия, при котором сумма квадратов отклонений измеренных значений от расчетных принимает наименьшее значение.

Для нахождения неизвестных параметров ($a_1, a_2, \ldots, a_n$) необходимо решить систему линейных уравнений

$$y_1 = a_1 \cdot x_1 + a_1 \cdot u_1 + \ldots + a_n \cdot z_1;$$

$$y_2 = a_1 \cdot x_2 + a_1 \cdot u_2 + \ldots + a_n\, z_2;$$

$$\ldots$$

$$y_n = a_1 \cdot x_m + a_1 \cdot u_m + \ldots + a_n\, z_m, \qquad (63)$$

где $y_1, y_2, \ldots, y_n$ – частные значения измеренных величин функции y, x, u, z – переменные величины.

Эту систему приводят к системе линейных уравнений путем умножения каждого уравнения соответственно на $x_1, x_2, \ldots, x_m$ и последующего их сложения, затем соответственно на $u_1, u_2, \ldots, u_m$. Это позволяет получать так называемую систему нормальных уравнений, решение которой и дает искомые коэффициенты.

$$\sum_1^m yx = a_1 \sum_1^m xx + a_2 \sum_1^m xu + \ldots + a_n \sum_1^m x\,\mathrm{z};$$

$$\sum_1^m yu = a_1 \sum_1^m ux + a_2 \sum_1^m uu + \ldots + a_n \sum_1^m u\mathrm{z};$$

$$\ldots$$

$$\sum_{1}^{m} yz = a_1 \sum_{1}^{m} z x + a_2 \sum_{1}^{m} z u + \ldots + a_n \sum_{1}^{m} z z. \qquad (64)$$

Для вычисления коэффициентов A методом наименьших квадратов необходимо пользоваться типовыми программами для ЭВМ.

Контрольные вопросы

1. Что можно аппроксимировать полиномами?

2. Положение, на котором основан метод средних квадратов.

3. Достоинство метода средних квадратов. Условие его получения.

6. Метод наименьших квадратов (МНК)

Математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов некоторых функций от искомых переменных. Он может использоваться:

– для «решения» переопределенных систем уравнений (когда количество уравнений превышает количество неизвестных), для поиска решения в случае обычных (не переопределенных) нелинейных систем уравнений;

– для аппроксимации точечных значений некоторой функцией.

МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

Суть метода заключается в том, что если все измерения функций $y_1, y_2, \ldots, y_n$ произведены с одинаковой точностью и распределенные величины ошибок измерений соответствуют нормальному закону, то параметры исследуемого уравнения определяются из условия, при котором сумма квадратов отклонений измеренных значений от расчетных принимает наименьшее значение.

Для нахождения неизвестных параметров ($a_1, a_2, \ldots, a_n$) необходимо решить систему линейных уравнений:

$$\begin{cases} y_1 = a_1 x_1 + a_2 U_1 + \ldots + a_n z_1 \\ y_1 = a_1 x_2 + a_2 U_2 + \ldots + a_n z_n \\ \ldots \\ y_n = a_1 x_m + a_2 U_m + \ldots + a_n z_m \end{cases}, \qquad (65)$$

где $y_1, \ldots, y_n$ – частные значения измеренных величин функции y, x; U, z – переменные величины.

Эту систему приводят к системе нормальных уравнений, которая получается путем умножения каждого уравнения на x_1, …, x_m, и последующего их сложения, затем – на U_1, … ,U_m и т.д.

Это позволяет получить систему нормальных уравнений:

$$\begin{cases} \sum_1^m yx = a_1 \sum_1^m xx + a_2 \sum_1^m xU + \ldots + a_n \sum_1^m xz \\ \sum_1^m yU = a_1 \sum_1^m Ux + a_2 \sum_1^m UU + \ldots + a_n \sum_1^m Uz \\ \ldots \\ \sum_1^m yz_n = a_1 \sum_1^m zx + a_2 \sum_1^m zU + \ldots + a_n \sum_1^m zz \end{cases}. \qquad (66)$$

Решение этой системы дает искомые коэффициенты.

Метод наименьших квадратов дает достаточно надежные результаты.

Контрольные вопросы

1. На чем основан метод наименьших квадратов?
2. Варианты использования метода наименьших квадратов.
3. Суть метода наименьших квадратов.

7. Корреляционный анализ

Для описания, анализа и прогнозирования явлений и процессов применяют математические модели в форме уравнений или функций. Модель процесса, отражая основные его свойства и абстрагируясь от второстепенных, позволяет судить о его поведении в определенных конкретных условиях.

Математическая модель процесса представляется чаще всего, как функция влияющих на него факторов. Некоторые из них

оказывают существенное влияние на результат, другие – весьма незначительное.

Как правило, существенных факторов немного, в то время как несущественных достаточно большое число, поэтому последними полностью пренебрегать нельзя. Как известно, источником любого богатства является труд (прошлый или настоящий). Поэтому к числу основных факторов, например, при изучении экономического процесса, относят обычно настоящий труд (или трудовые ресурсы в той или иной мере), прошлый труд (энергия, сырье, материалы, оборудование, здания, сооружения и т.д.). Вместе с тем труд прилагается при определенном состоянии внешней среды, т.е. при определенных природных условиях, поэтому соответствующие факторы также должны найти отражение в модели.

Рассмотрим двумерную модель. Предположим, что требуется установить и оценить зависимость изучаемой случайной величины от одного фактора, являющегося также случайной величиной.

Любые две случайные величины могут быть связаны либо функциональной, либо стохастической зависимостью, либо быть независимыми. Функциональная зависимость между двумя случайными величинами характеризуется тем, что каждому значению одной случайной величины соответствует определенное значение другой. Такой зависимостью связаны, например, количество купленного одноименного товара и его стоимость; количество потребленной абонентом электроэнергии и плата за нее.

Однако встречаются величины, когда каждому значению одной величины соответствует целое множество значений другой. Например, рост человека и его вес, урожайность определенной сельскохозяйственной культуры и количество внесенных удобрений и т.д.

Множество значений случайной величины Y, соответствующих фиксированному значению случайной величины X, будем рассматривать как соответствующее ему распределение случайной величины Y. Зависимость между величинами называется *стохастической*, если каждому значению одной случайной величины соответствует определенное распределение другой случайной величины, меняющееся с изменением величины X и по вариантам и по частотам.

Если стохастическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой, то эту зависимость называют *корреляционной*.

Уточним это понятие, для чего сначала введем понятие условной средней. *Условной средней yx* называют среднее арифметическое значений случайной величины *Y*, соответствующее фиксированному значению случайной

величины *X*. Если каждому значению *X* соответствует единственное значение условной средней *yx* , то очевидно, что условная средняя есть функция от *x*.

Корреляционной зависимостью Y от *X* называют функциональную зависимость условной средней y_x от *x* и обозначают $y_x = f(x)$.

Это уравнение называют уравнением регрессии *Y* по *X*. Функцию $f(x)$ называют функцией регрессии *Y* по *X*, а ее график – линией регрессии *Y* по *X*.

Аналогично определяется уравнение регрессии *X* по *Y*:

$$\bar{x}_y = \varphi(y), \tag{67}$$

где $\bar{x}_y$ – условная средняя величины *X*.

Оценка тесноты *линейной связи* между случайными величинами *Y* и *X*, то есть оценка степени рассеивания значений *X* около линии регрессии, для разных значений случайной величины *Y* (или наоборот) осуществляется с помощью коэффициента корреляции, формулу для вычисления которого получил Пирсон в начале 90-х годов XIX в.

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \tag{68}$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$; $\sigma_x = \sqrt{\bar{x}^2 - (\bar{x}^2)}$

$\sigma_y = \sqrt{\bar{y}^2 - (\bar{y}^2)}$ – элементы двумерной выборки; *n* – объем выборки.

$$\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2;\ \bar{y}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2. \tag{69}$$

Линейный коэффициент корреляции характеризует степень тесноты не всякой, а только линейной зависимости. При нели-

нейной зависимости между явлениями для измерения тесноты связи применяют так называемое корреляционное отношение, известное также под названием «индекс корреляции».

Линейная вероятностная зависимость случайных величин заключается в том, что при возрастании одной случайной величины другая имеет тенденцию возрастать или убывать по линейному закону. Эта тенденция может быть более или менее ярко выраженной, т.е. более или менее приближаться к функциональной зависимости. Рассмотрим свойства линейного коэффициента корреляции при достаточно большом объеме выборки.

1. Коэффициент корреляции принимает значения на отрезке [– 1; 1].

Знак r характеризует направление корреляционной зависимости. Если $r_{yx} > 0$, то увеличение признака X в среднем приводит к увеличению признака Y.

Если $r_{yx} < 0$, то с увеличением признака X в среднем признак Y уменьшается.

2. Если случайные величины X и Y связаны точной линейной функциональной зависимостью $y = ax + b$, то $r_{yx} \pm 1$. При этом линии регрессии Y по X и X по Y совпадают.

3. Если $r_{yx} = 0$, то линейная корреляционная связь отсутствует, а линии регрессии Y по X и X по Y параллельны осям координат.

Для качественной оценки тесноты линейной корреляционной связи величин X и Y можно воспользоваться шкалой Чеддока (табл. 5). Можно показать, что $r_{xy} = r_{yx} = r$.

На практике коэффициент корреляции находят по выборочным данным, следовательно, он будет отличаться от коэффициента корреляции генеральной совокупности. В связи с этим необходимо определить точность показателей корреляции и границы доверительных интервалов.

Выборочный коэффициент корреляции r представляет собой случайную величину, поэтому его распределение можно считать нормальным или приближенно нормальным, если:

– переменные X и Y имеют совместное нормальное или приближенно нормальное распределение;

– $r \neq \pm 1$;

– объем выборки достаточно велик.

Таблица 5

Шкала Чеддока

Теснота связи	Значение коэффициента корреляции при наличии:	
	прямой связи	обратной связи
Слабая	0,1–0,3	(– 0,3) – (– 0,1)
Умеренная	0,3–0,5	(– 0,5) – (– 0,3)
Заметная	0,5–0,7	(– 0,7) – (– 0,5)
Высокая	0,7–0,9	(– 0,9) – (– 0,7)
Весьма высокая	0,9–0,99	(– 0,99) – (– 0,9)

Пусть полученное значение выборочного коэффициента корреляции

$r \neq 0$. Закономерен вопрос: объясняется ли это действительно существующей линейной корреляционной связью между переменными X и Y в генеральной совокупности или является следствием случайного отбора элементов в выборку (т.е. при другом отборе, возможно, что $r = 0$).

Обычно в этих случаях проверяется гипотеза H_0 об отсутствии линейной корреляционной связи между переменными в генеральной совокупности, т.е. H_0: $\rho = 0$ (ρ – коэффициент корреляции генеральной совокупности) при конкурирующей гипотезе H_1: $\rho \neq 0$ и заданном уровне значимости α.

При справедливости нулевой гипотезы вычисляют статистику, которая имеет t-распределение Стьюдента с $k = n - 2$ степенями свободы.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (70)$$

Поэтому при данном уровне значимости α и k степенях свободы по табл. 2 t-критерия Стьюдента находят критическое значение случайной величины $t_{кр} = t(\alpha, k)$. Строят двустороннюю критическую область.

Тогда при $|t| \geq t_{кр}$ гипотеза H_0 должна быть отвергнута, т.е. выборочный коэффициент корреляции значимо (существенно) отличен от нуля. В этом случае можно принять, что в генеральной совокупности случайные величины X и Y с вероятностью $\gamma = 1 - \alpha$ *коррелированны*.

При $|t| < t_{кр}$ нет основания отвергать гипотезу H_0, поэтому отклонение выборочного коэффициента корреляции от нуля может носить чисто случайный характер.

Для статистически значимого линейного коэффициента корреляции целесообразно найти доверительный интервал (интервальную оценку), который с заданной надежностью $\gamma = 1 - \alpha$ содержит неизвестный генеральный коэффициент корреляции ρ. Для построения такого интервала необходимо знать выборочное распределение коэффициента корреляции, которое с ростом объема выборки очень медленно приближается к нормальному распределению.

В таких случаях прибегают к специально подобранным функциям от r, которые сходятся к хорошо изученным распределениям. Чаще всего для подбора функции применяют z-преобразование Фишера:

$$z = \frac{1}{2}\ln\frac{1+z}{1-z}. \qquad (71)$$

Распределение z уже при небольших значениях n является приближенно нормальным с математическим ожиданием

$$M(z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

и дисперсией

$$\sigma_z^2 = \frac{1}{n-3}. \qquad (72)$$

Поэтому вначале строят доверительный интервал для $M(z)$:

$$z - t_{1-\alpha}\frac{1}{\sqrt{n-3}} \le M(z) \le z + t_{1-\alpha}\frac{1}{\sqrt{n-3}}, \qquad (73)$$

где $t_{1-\alpha}$ – нормированное отклонение z, определяемое с помощью функции Лапласа $\varphi(t_{1-\alpha}) = \gamma = 1 - \alpha$.

При определении границ доверительного интервала для ρ, т.е. для перехода от z к ρ, существуют специальная таблица. При её отсутствии переход может быть осуществлен по формуле

$$r = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}. \qquad (74)$$

Контрольные вопросы

1. Как представляется математическая модель процесса?
2. Когда зависимость между величинами называется стохастической?
3. Когда зависимость между величинами называется корреляционной?
4. Когда уравнение называют уравнением регрессии Y по X?
5. Что такое оценка тесноты линейной связи между случайными величинами Y и X. С помощью чего она осуществляется?
6. Когда используется отношение «индекс корреляции»?
7. Свойства линейного коэффициента корреляции.
8. Когда величины X и Y коррелированны?

8. Регрессионный анализ

Наряду с корреляционным анализом обычно проводится и *регрессионный анализ*, который заключается в определении аналитического выражения связи зависимой случайной величины Y (называемой также *результативным признаком*) с независимыми случайными величинами X_1, X_2, … , X_m (называемыми также *факторами*).

Форма связи результативного признака Y с факторами X_1, X_2, … , X_m получила название *уравнения регрессии*. В зависимости от типа выбранного уравнения различают *линейную* и *нелинейную* регрессию (в последнем случае возможно дальнейшее уточнение: квадратичная, логарифмическая, показательная и т.д.). В зависимости от числа взаимосвязанных признаков различают *парную* и *множественную* регрессию. Если исследуется связь между двумя признаками (результативным и факторным), то регрессия называется *парной*. Если между тремя и более признаками, то регрессия называется *множественной* или *многофакторной*.

При изучении регрессии следует придерживаться определенной последовательности этапов:

1. Задание аналитической формы уравнения регрессии и определение параметров регрессии.
2. Определение в регрессии степени стохастической взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.

3. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов.

Этап 1. Выбор вида уравнения регрессии (очень важный этап анализа) производится на основании опыта предыдущих исследований, литературных источников, других соображений профессионально-теоретического характера, а также визуального наблюдения расположения «облака» точек корреляционного поля. Рассмотрим линейную модель множественной регрессии. Тогда уравнение множественной линейной регрессии может быть записано в виде:

$$\bar{y} = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + ... + a_m x_m, \qquad (75)$$

где $\bar{y}$ – теоретические значения результативного признака, полученные путем подстановки соответствующих значений факторных признаков в уравнение регрессии; x_1, x_2, … , x_m – значения факторных признаков; a_0, a_1, … , a_m – параметры уравнения (коэффициенты регрессии).

Параметры уравнения регрессии могут быть определены, например, с помощью метода наименьших квадратов. Данный метод заключается в нахождении параметров уравнения регрессии, при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выбранному уравнению регрессии.

$$S = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_1 - a_2 x_2 - a_3 x_3 - ... - a_m x_m)^2 \to \min . \qquad (76)$$

Рассматривая S в качестве функции параметров a_i и проводя математические преобразования (нахождение частных производных и равенство их нулю), получаем систему линейных уравнений с m неизвестными:

$$\begin{cases} n a_0 + a_1 \sum x_1 + + a_2 \sum x_2 + \ldots + a_m \sum x_m = \sum y \\ a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_2 x_1 + \ldots + a_m \sum x_m x_1 = \sum y x_1 \\ \ldots \\ a_0 \sum x_m + a_1 \sum x_1 x_m + a_2 \sum x_2 x_m + \ldots + a_m \sum x_m^2 = \sum y x_m \end{cases}, \qquad (77)$$

где n – число наблюдений, т.е. объем выборки; m – число факторов в уравнении регрессии.

Решив систему уравнений, определяют значения параметров a_i, являющихся коэффициентами искомого теоретического уравнения регрессии.

Этап 2. Для определения величины степени стохастической взаимосвязи результативного признака *Y* и факторов *X* необходимо знать следующие дисперсии:

• *общую дисперсию* результативного признака, отображающую влияние как основных, так и остаточных факторов

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 , \qquad (78)$$

где $\bar{y}$ – среднее значение результативного признака *Y*.

• *факторную дисперсию* результативного признака, отображающую влияние только основных факторов

$$\sigma_{\phi}^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 ; \qquad (79)$$

• *остаточную дисперсию* результативного признака, отображающую влияние только остаточных факторов

$$\sigma_0^2 = \frac{1}{n-(m+1)}\sum_{i=1}^{n}(y_i - \tilde{y})^2 . \qquad (80)$$

При корреляционной связи признака *Y* и факторов *X* выполняется соотношение $\sigma_{\text{ф}}^2 < \sigma_y^2$, при этом $\sigma_y^2 = \sigma_{\text{ф}}^2 + \sigma_0^2$.

Для анализа общего качества уравнения линейной многофакторной регрессии используют множественный коэффициент детерминации R^2, называемый также квадратом коэффициента множественной корреляции R или квадратом эмпирического корреляционного отношения *Y* по *X*.

Множественный коэффициент детерминации рассчитывается по формуле

$$R^2 = \frac{\sigma_{\text{ф}}^2}{\sigma_y^2} \qquad (81)$$

и определяет долю вариации результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную регрессионную модель.

Так как уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос об адекватности построенного уравнения генеральным данным. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F–критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}. \qquad (82)$$

Примечание. Если в уравнении регрессии свободный член $a_0 = 0$, то числитель $n - m - 1$ надо увеличить на 1, т.е. он будет равен $n - m$. В математической статистике доказывается, что если справедлива гипотеза H_0: $R^2 = 0$, то величина F имеет F-распределение с числом степеней свободы $k_1 = m$ и $k_2 = n - m - 1$.

Используя таблицу критерия Фишера – Снедекора, по заданному уровню значимости и числу степеней свободы k_1 и k_2 определяют критическую точку $F_{\text{КР}}(k_1, k_2)$.

Если $F > F_{\text{КР}}$, то гипотеза H_0 о незначимости коэффициента детерминации отвергается, в противном случае принимается.

При значениях $R^2 > 0{,}7$ считается, что вариация результативного признака Y обусловлена в основном влиянием включенных в регрессионную модель факторов X.

Для оценки адекватности уравнения регрессии часто используют показатель средней ошибки аппроксимации:

$$\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \tilde{y}|}{y_i} \cdot 100\%. \qquad (83)$$

Этап 3. Возможна ситуация, когда часть вычисленных коэффициентов регрессии не обладает необходимой степенью значимости, т.е. значения данных коэффициентов будут меньше их стандартной ошибки. В этом случае такие коэффициенты должны быть исключены из уравнения регрессии. Поэтому проверка адекватности построенного уравнения регрессии наряду с проверкой значимости коэффициента детерминации включает в себя также проверку значимости каждого коэффициента уравнения регрессии. Значимость коэффициентов регрессии проверяется с помощью t-критерия Стьюдента:

$$t = \frac{a_i}{\sigma_{a_i}}, \qquad (84)$$

где σ_{a_i} – стандартное значение ошибки для коэффициента регрессии a_i.

В математической статистике доказывается, что если гипотеза H_0: $a_i = 0$ выполняется, то величина t имеет распределение Стьюдента с числом степеней свободы $k = n - m - 1$. По табл. 2 значений t-критерия Стьюдента по заданной надежности γ и числу степеней свободы k определяют критическую точку $t_{кр}(\gamma, k)$. Гипотеза H_0: $a_i = 0$ о незначимости коэффициента регрессии отвергается, если $|t| > |t_{кр}|$. Кроме того, зная $t_{кр}$, можно найти границы доверительных интервалов для коэффициентов регрессии по формулам:

$$a_i - t_{кр}\sigma_{ai} < a_i < a_i + t_{кр}\sigma_{ai}. \qquad (85)$$

При экономической интерпретации уравнения регрессии широко используются *частные коэффициенты эластичности*, показывающие, на сколько процентов в среднем изменится значение результативного признака при изменении значения соответствующего факторного признака на 1 %, и определяемые по формуле

$$Э_{xi} = a_i \frac{\bar{x}_i}{y}, \qquad (86)$$

где $\bar{x}_i$ – среднее значение соответствующего факторного признака; y – среднее значение результативного признака Y; a_i – коэффициент регрессии при соответствующем факторном признаке.

Если имеет место парная регрессия, то выборочное уравнение линейной регрессии может быть записано в виде

$$\tilde{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}). \qquad (87)$$

Можно доказать, что это уравнение является точечной оценкой для линии регрессии генеральной совокупности $y_x = f(x)$.

В качестве оценки дисперсии σ_y^2 случайной величины Y в генеральной совокупности выступает остаточная дисперсия, определяемая выражением

$$s_y^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 . \qquad (88)$$

Деление на $(n - 2)$ приводит к тому, что s_y^2 становится несмещенной оценкой дисперсии σ_y^2. Между остаточной дисперсией $2y \cdot s$ и коэффициентом корреляции r существует зависимость, выражающая связь между линейной регрессией и линейной корреляцией $s_y^2 = \sigma_y^2(1 - r^2)$.

Контрольные вопросы

1. В чем заключается регрессионный анализ?
2. Что такое уравнение регрессии?
3. Как различают регрессию?
4. Когда регрессию называют множественной?
5. Последовательность этапов при изучении регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. Кузнецов, И. Н. Основы научных исследований : учеб. пособие. – Москва : Дашков и Ко, 2013. – 284 с.

2. Шкляр, М. Ф. Основы научных исследований : учеб. пособие. – Москва : Дашков и Ко, 2008. – 244 с.

3. Шкляр, М. Ф. Основы научных исследований : учеб. пособие. – Москва : Дашков и Ко, 2012. – 244 с.

4. Клюкин, Г. К. Основы научных исследований [Электронный ресурс] : курс лекций для студентов специальности 21.05.04.05 «Шахтное и подземное строительство» / Г. К. Клюкин; КузГТУ. – Кемерово, 2011. – Загл. с экрана. – 44 с.
http://library.kuzstu.ru/meto.php?n=90542&type=utchposob:common

5. Основы научных исследований : учебник для техн. вузов. / В. И. Крутов, И. М. Глушко, В. В. Попов [и др.]. – Москва : Высш. шк. 1989. – 400 с.

6. Капица, П. Л. Эксперимент, теория, практика. – Москва : Наука, 1977. – 351 с.

7. Румшиский, Л. З. Математическая обработка результатов эксперимента. – Москва : Наука, 1971. – 192 с.

8. Кожухар, В. М. Основы научных исследований : учеб. пособие. – Москва : Дашков и Ко, 2012. – 216 с.

9. Клюкин, Г. К. Научно-исследовательская работа [Электронный ресурс] : учеб. пособие для студентов направления подготовки 21.05.05. «Горное дело» ФГБОУ ВО «Кузбас. гос. техн. ун-т им. Т. Ф. Горбачева», Каф. стр-ва подзем. сооружений, шахт и разраб. месторождений полез. ископаемых / Г. К. Клюкин; КузГТУ. – Кемерово, 2015. – Загл. с экрана. – 43 с.

http://library.kuzstu.ru/meto.php?n=91344&type=utchposob:common