НАЗЕМНЫЕ ТРАНСПОРТНО-ТЕХНОЛОГИЧЕСКИЕ СРЕДСТВА

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ

Кафедра прикладной математики

СТАТИСТИКА И ОБРАБОТКА ЭКСПЕРИМЕНТА

Методические указания к практическим занятиям и самостоятельной работе для обучающихся по направлению подготовки 23.05.01 Наземные транспортно-технологические средства

Составители:
Е.В. Кондрашова, А.И. Игнатов

Москва
Издательство МИСИ – МГСУ
2020

УДК 311
ББК 60.6
С78

Рецензент — доктор физико-математических наук *Б.П. Титаренко*,
профессор кафедры прикладной математики НИУ МГСУ

С78 **Статистика и обработка эксперимента** [Электронный ресурс] : методические указания к практическим занятиям и самостоятельной работе для обучающихся по направлению подготовки 23.05.01 Наземные транспортно-технологические средства / сост. : Е.В. Кондрашова, А.И. Игнатов ; Министерство науки и высшего образования Российской Федерации, Национальный исследовательский Московский государственный строительный университет, кафедра прикладной математики. — Электрон. дан. и прогр. (0,9 Мб). — Москва : Издательство МИСИ – МГСУ, 2020. — Режим доступа: http://lib.mgsu.ru. — Загл. с титул. экрана.

Методические указания составлены в соответствии с программой дисциплины «Статистика и обработка эксперимента» и знакомят студентов с основными методами статистического анализа, а также со статистическими методами планирования эксперимента. Представлены теоретические выкладки, примеры и практические задания по изучаемому курсу для закрепления обучающимися знаний, приобретенных в процессе изучения курса.

Для обучающихся по направлению 23.05.01 Наземные транспортно-технологические средства.

Учебное электронное издание

Редактор, корректор *М.Л. Манзюк*
Компьютерная верстка *А.Г. Сиволобовой*
Дизайн первого титульного экрана *Д.Л. Разумного*

Для создания электронного издания использовано:
Microsoft Word 2013, Adobe InDesign CS6, ПО Adobe Acrobat.

Подписано к использованию 13.05.2020. Объем данных 0,9 Мб.

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Национальный исследовательский
Московский государственный строительный университет»
129337, Москва, Ярославское ш., 26.

Издательство МИСИ – МГСУ.
Тел. (495) 287-49-14, вн. 13-71, (499) 188-29-75, (499) 183-97-95.
E-mail: ric@mgsu.ru, rio@mgsu.ru

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Методы математической статистики и обработки эксперимента находят широкое применение в различных областях, в том числе в производстве, в технических дисциплинах.

При изучении и сравнении различных технологических процессов, методов обработки по определенным измеряемым признакам результаты наблюдений могут быть использованы для получения оценок некоторых величин или для проверки предположений, которые называют гипотезами.

В качестве источника получения знаний на опытной основе инженерной деятельности служит эксперимент. С одной стороны, под экспериментом понимают получение сведений о характеристиках, поведении совокупности большого числа однородных объектов, с другой стороны, под этим понятием могут подразумевать информацию о свойстве некоторого объекта в течение некоторого промежутка времени (в данном случае мы говорим о многократно повторяющихся наблюдениях одного и того же объекта).

В статистике для описания любого множества объектов используется понятие совокупности. Под генеральной совокупностью понимают множество объектов, из которых проводится отбор в процессе проводимых наблюдений. Отобранные для наблюдений объекты представляют собой выборку (или выборочную совокупность) объема n, где n — число этих объектов.

Так как в результате наблюдения или эксперимента значение какой-либо характеристики определить абсолютно точно невозможно, возникает проблема появления ошибок, которые могут относиться к грубым ошибкам измерений, методическим ошибкам и случайным ошибкам. Если грубые ошибки легко выявить и устранить, а методические, возникающие в процессе отладки методики, устраняются или же учитываются, то случайные ошибки устранить невозможно. Однако можно учесть их влияние на тенденцию распределения выборочных данных.

1. ПРИКЛАДНАЯ СТАТИСТИКА

1.1. Выборочный ряд. Интервалы группировки

Предметом математической статистики является суждение о случайной величине X по результатам эксперимента, многократных однотипных наблюдений, в которых фиксируются определенные значения этой случайной величины.

Выборкой объема n называют совокупность $X_1^0, X_2^0, \ldots, X_n^0$ независимых измерений случайной величины X. При этом говорят, что выборка взята из *генеральной совокупности X*.

Значения признака, которые при переходе от одного элемента совокупности к другому изменяются, называются *вариантами* и обозначаются маленькими латинскими буквами. Ряд значений признака, расположенный в порядке возрастания или убывания, с соответствующими им весами называется *вариационным рядом.* В качестве *весов* выступают *частоты* или *относительные частоты. Частота* (m_i) показывает, сколько раз встречается тот или иной вариант в статистической совокупности. *Относительная частота* (w_i) показывает, какая часть единиц совокупности имеет тот или иной вариант к сумме всех частот ряда. Относительная частота рассчитывается по формуле

$$w_i = \frac{m_i}{\sum_{i=1}^{k} m_i}.$$

Вариационные ряды могут быть представлены в дискретной и интервальной форме. *Дискретные вариационные ряды* строят обычно в том случае, если значения изучаемого признака могут отличаться друг от друга не менее чем на некоторую конечную величину (задаются точечные значения признака) (табл. 1).

Таблица 1

Общий вид дискретного ряда

Значения признака (x_i)	x_1	x_2	...	x_k
Частоты (m_i)	m_1	m_2	...	m_k

В интервальных вариационных рядах значения признаков в них задаются в виде интервалов (табл. 2).

Таблица 2

Общий вид интервального ряда

Значения признака (x_i)	$(a_1; a_2]$	$(a_2; a_3]$	...	$(a_{i-1}; a_i]$
Частоты (m_i)	m_1	m_2	...	m_i

Разность между верхней и нижней границами интервала называется *интервальной разностью* или *длиной интервала*. В общем виде интервальную разность k_i можно представить как $k_i = x_{i\,(\max)} - x_{i\,(\min)}$. Если интервалы в вариационных рядах имеют одинаковую длину, их называют *равновеликими.*

Для определения оптимальной величины интервалов применяют *формулу Стерджесса*

$$n = 1 + [\log_2 N],$$

где n — количество интервалов; N — общее число наблюдений.

1.2. Гистограмма и полигон

Гистограмма — это ступенчатая фигура, состоящая из прямоугольников, основания которых — интервалы длины h, а высоты — w_i/h, т.е. *площадь* каждого прямоугольника равна соответствующей *относительной частоте*, а полная площадь всей гистограммы равна единице (рис. 1).

При большом числе наблюдений и увеличении количества интервалов контур гистограммы приближается к графику функции плотности вероятности, и по виду гистограммы можно предварительно сделать вывод о законе распределения изучаемой непрерывной изучаемой величины.

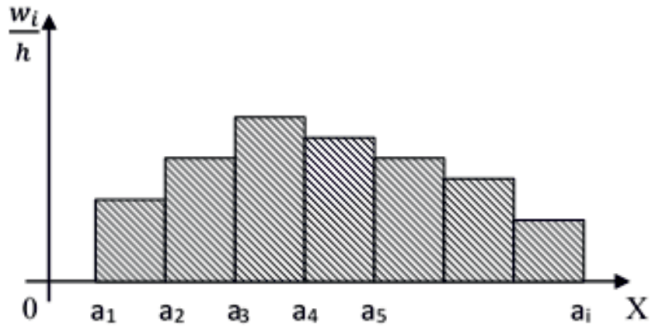

Рис. 1. Гистограмма

Полигоном относительных частот (рис. 2) называют ломаную, отрезки которой соединяют точки $(x_i; w_i)$, где x_i — середины интервалов, w_i — соответствующие им относительные частоты. По виду полигона можно выдвинуть гипотезу о законе распределения дискретной случайной величины.

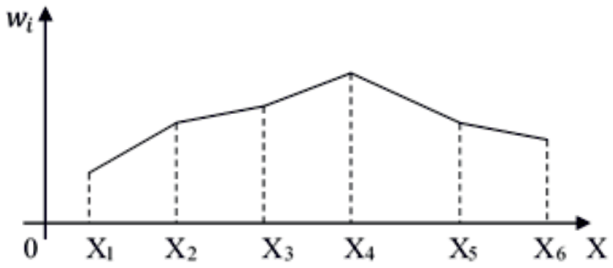

Рис. 2. Полигон частот

1.3. Эмпирическая функция распределения

Эмпирическая функция распределения (или *кумулянта*) определяется следующим равенством:

$$F^*(x) = \frac{n_x}{n},$$

где n_x — число *выборочных значений X, меньших x*.

При увеличении числа наблюдений график $F^*(x)$ приближается к графику функции распределения изучаемой случайной величины. Таким образом, график эмпирической функции распределения позволяет выдвинуть гипотезу о виде распределения.

1.4. Основные числовые выборочные характеристики

Выборочным начальным моментом k-го порядка называют величину m_k, вычисляемую по формуле

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Выборочный момент 1-го порядка обозначают $\overline{X} = m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ и называют *выборочным средним*.

Выборочным центральным начальным моментом k-го порядка называют величину v_k, вычисляемую по формуле

$$v_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

При $k = 2$ получаем *выборочную дисперсию*

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Так как эта оценка оказывается смещенной, рассматривают так называемую *исправленную выборочную дисперсию*:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Пусть имеются две выборки одинакового объема $\overline{X}_n$, $\overline{Y}_n$, тогда для них может быть найден *выборочный корреляционный момент*

$$K(\overline{X}_n, \overline{Y}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

Выборочным коэффициентом корреляции называют величину

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

Задачи для самостоятельного решения

1. По имеющимся данным, мощности предприятий за год в 2019 г. составляют:

Предприятия с годовой мощностью, тыс. т	Количество предприятий
До 500	27
500–1000	11
1000–2000	8
2000–3000	8
Свыше 3000	2

Построить гистограмму и кумуляту. Рассчитать среднюю мощность предприятий. Найти дисперсию и коэффициент вариации. Проанализировать результаты.

2. Постройте гистограмму частот, найдите среднюю заработную плату работников одного из транспортных цехов.

Заработная плата, у. е	50–70	70–100	100–150	150–175	175–200	200–225
Число работников	14	21	37	20	15	10

Найти среднее квадратическое отклонение, коэффициент вариации заработной платы.

3. Для оценки состояния деловой активности предприятий были проведены обследования и получены следующие результаты:

Показатель деловой активности	0–8	8–16	16–24	24–32
Число предприятий	10	15	8	5

Построить гистограмму распределения частот. Найти среднее значение показателя деловой активности, дисперсию, коэффициент вариации.

4. При обследовании 50 комплектов для ремонта транспортных средств установлено следующее количество запасных элементов в комплектах: 5; 4; 3; 1; 4; 5; 3; 8; 10; 1; 3; 2; 5; 6; 7; 3; 5; 2; 3; 6; 8; 3;3; 5; 5; 6; 5; 4; 8; 5; 6; 4; 8; 7; 4; 5; 7; 8; 6; 5; 7; 5; 7; 6; 7; 3; 5; 7; 3; 4. Составить вариационный ряд распределения частот. Построить кумуляту. Найти выборочное среднее и дисперсию.

5. Постройте гистограмму частот, найдите среднюю арифметическую, среднее квадратическое отклонение и коэффициент вариации для данных о выручке компании на длительной зарубежной техновыставке.

Выручка, у. е.	0–200	200–300	300–400	400–500	500–600	600–700
Число дней	3	5	9	14	8	3

6. Компанию по прокату автомобилей интересует зависимость между пробегом автомобилей (X) и стоимостью ежемесячного техобслуживания (Y). Для выяснения характера этой связи было отобрано 15 автомобилей.

X	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	13	16	15	20	19	21	26	24	30	32	30	35	34	40	39

Постройте график исходных данных и определите по нему характер зависимости. Рассчитайте выборочный коэффициент линейной корреляции Пирсона, проверьте его значимость при $\alpha = 0{,}05$. Постройте уравнение регрессии и дайте интерпретацию полученных результатов.

1.5. Оценки и методы их нахождения: метод моментов и метод максимального правдоподобия

Одна из задач математической статистики — оценить неизвестные истинные параметры генеральной совокупности.

Любая числовая функция от выборки является *статистикой*.

Статистика называется *оценкой параметра*, если при любом значении выборки считают ее значение приближенно равным неизвестному параметру. Существует ряд свойств, предъявляемых к оценкам неизвестного числового параметра, при выполнении которых оценка считается разумной.

Пусть θ — неизвестный параметр изучаемой случайной величины X.

Оценка θ^* называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру $M\theta^* = \theta$.

Оценка θ^* называется *состоятельной*, если она сходится по вероятности к оцениваемому параметру, т.е. для всякого $\varepsilon > 0,\ \lim\limits_{n\to\infty} P(|\theta^* - \theta| > \varepsilon) = 0.$

Поскольку функций от выборки можно придумать много, то нужно условиться о разумности применяемых оценок. Будем считать оценку разумной, если она несмещенная и состоятельная.

Существует несколько *методов получения оценок параметров* случайных величин.

Метод моментов заключается в приравнивании эмпирических (выборочных) моментов, вычисленных по данной выборке, и теоретических, вычисленных по предполагаемой плотности распределения, содержащей неизвестные параметры. Если число полученных моментов равно числу неизвестных параметров распределения, получают систему уравнений для вычисления этих неизвестных параметров.

Пример 1

Найти методом моментов по выборке точечную оценку $\hat{\theta}$ неизвестного параметра θ данного распределения случайной величины, заданной плотностью вероятностей $f(x) = \dfrac{2}{\theta\sqrt{\pi}} e^{-\frac{(x-1)^2}{\theta^2}}, x > 1.$

Рассмотрим математическое ожидания случайной величины, используя формулу вычисления математического ожидания для непрерывной случайной величины:

$$M(X) = \int\limits_{-\infty}^{+\infty} x f(x)\,dx = \int\limits_{1}^{+\infty} x \frac{2}{\theta\sqrt{\pi}} e^{-\frac{(x-1)^2}{\theta^2}}\,dx.$$

Вычислив данный интеграл, используя замену переменной, получаем:

$$M(X) = \frac{\theta}{\sqrt{\pi}} + 1.$$

Найденное математическое ожидание (теоретический момент первого порядка) приравниваем соответствующему эмпирическому моменту (в нашем случае — выборочному среднему):

$$\frac{\theta}{\sqrt{\pi}} + 1 = \overline{X}.$$

При решении уравнения относительно параметра θ получаем точечную оценку

$$\hat{\theta} = (\overline{X} - 1)\sqrt{\pi}.$$

Еще одним методом нахождения оценок является *метод максимального правдоподобия*, который состоит в нахождении значений неизвестного параметра распределения, при котором функция правдоподобия достигает максимума, которую мы определим ниже.

Функцией правдоподобия случайной выборки $\{X_1, \ldots, X_n\}$ из генеральной совокупности X, закон распределения которой известен с точностью до параметра $\overline{\theta}$, называется функция, определяемая как произведение вероятностей $L(X_1, \ldots, X_n, \overline{\theta}) = \prod\limits_{i=1}^{n} p(X_i, \overline{\theta})$. В случае, если генеральная совокупность имеет непрерывное распределение с плотностью вероятности, известной с точностью до параметра, функция правдоподобия имеет вид

$$L(X_1, \ldots, X_n, \overline{\theta}) = \prod_{i=1}^{n} f(X_i, \overline{\theta}).$$

Оценка $\hat{\overline{\theta}}(\overline{X}_n)$ максимального правдоподобия параметра $\overline{\theta}$ должна удовлетворять условию (необходимому условию экстремума):

$$L\left(\overline{X}_n, \hat{\overline{\theta}}\right) = \max_{\overline{\theta}} L\left(\overline{X}_n, \overline{\theta}\right).$$

Для упрощения вычислений обычно используется логарифмирование, при котором точки экстремума остаются теми же, а уравнения упрощаются. Тогда условие записывается в следующем виде

$$\frac{\partial \ln L\left(\overline{X}_n, \overline{\theta}\right)}{\partial \theta_i} = 0, i = 1, \ldots, k \text{ при } \overline{\theta} = \left(\theta_1, \ldots, \theta_k\right).$$

Полученные уравнения называются *уравнениями правдоподобия*. После того как получены их решения, можно дополнительно проверить достаточные условия максимума.

Пример 2

Используя метод максимального правдоподобия, найти точечную оценку $\hat{\theta}$ неизвестного параметра θ заданного распределения случайной величины, заданной следующей плотностью вероятностей:

$$f\left(x\right) = \frac{\theta}{2}\left(\frac{2}{x}\right)^{\theta+1}, x > 2.$$

Составим функцию правдоподобия в случае непрерывного распределения:

$$L\left(X_1, \ldots, X_n, \theta\right) = \prod_{i=1}^{n} f(X_i, \theta) = \prod_{i=1}^{n} \frac{\theta}{2}\left(\frac{2}{x}\right)^{\theta+1}.$$

Удобно рассмотреть логарифмический вид функции правдоподобия. Тогда, используя свойства логарифмирования, получим:

$$\ln L\left(X_1, \ldots, X_n, \theta\right) = \ln \prod_{i=1}^{n} \frac{\theta}{2}\left(\frac{2}{x}\right)^{\theta+1} = n \ln \theta + n\theta \ln 2 - (\theta + 1)\sum_{i=1}^{n} \ln X_i.$$

Далее решаем уравнение правдоподобия:

$$\frac{\partial \ln L\left(X_1, \ldots, X_n, \theta\right)}{\partial \theta} = \frac{n}{\theta} + n \ln 2 - \sum_{i=1}^{n} \ln X_i = 0.$$

Данное уравнение решаем относительно θ:

$$\theta = \frac{n}{\sum_{i=1}^{n} \ln X_i - n \ln 3}.$$

Задачи для самостоятельного решения

1. Найти методом моментов точечную оценку неизвестных параметра θ распределения случайной величины, заданного плотностью вероятности: $f\left(x\right) = \theta x^{-(\theta+1)}, x > 1$.

2. Найти методом моментов точечную оценку неизвестных параметров θ распределения случайной величины, заданного плотностью вероятности: $f\left(x\right) = 2\theta^2 x e^{-\theta^2 x^2}, x > 0$.

3. Найти методом максимального правдоподобия точечную оценку неизвестных параметров θ распределения случайной величины, заданного плотностью вероятности: $f\left(x\right) = 3\theta x^2 e^{-\theta x^3}, x > 0$.

4. Найти методом максимального правдоподобия точечную оценку неизвестных параметров θ распределения случайной величины, заданного плотностью вероятности: $f\left(x\right) = \theta e^{-\theta x}, x \geq 0$.

1.6. Интервальное оценивание

Доверительным интервалом для параметра θ называется интервал (θ_1,θ_2), для которого выполняется $P(\theta_1<\theta<\theta_2)=1-\alpha$. Число 1−α называют *доверительной вероятностью*, а значение α — *уровнем значимости*.

Точные доверительные интервалы строятся, как правило, в предположении нормальности данных. При том, что реальные данные могут не выглядеть нормальными, тем не менее широкое практическое применение описываемых методов дает достаточно неплохие результаты (что объясняется, в частности, асимптотической нормальностью оценок).

Для оценки математического ожидания *a* нормально распределенной совокупности при известном среднем квадратическом отклонении σ по выборке объема *n* можно использовать следующую формулу

$$\overline{X}-\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}<a<\overline{X}+\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2},$$

где $u_{1-\alpha/2}$ — квантиль уровня $1-\alpha/2$ нормального распределения (определяется по таблицам).

Для оценки математического ожидания *a* нормально распределенной совокупности при неизвестной дисперсии σ^2 генеральной совокупности

$$\overline{X}-\frac{s}{\sqrt{n}}t_{\gamma}<a<\overline{X}+\frac{s}{\sqrt{n}}t_{\gamma},$$

где t_{γ} — критическая точка распределения Стьюдента (для двусторонней области) с *n* – 1 степенями свободы и уровнем значимости $\alpha=1-\gamma$; *s* — исправленное выборочное среднее квадратическое отклонение; *n* — объем выборки.

Для неизвестной дисперсии σ²:

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}<\sigma^2<\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},$$

где χ^2_n — критические точки χ^2-распределения с *n* – 1 степенями свободы и соответствующими уровнями значимости, $\alpha=1-\gamma$ (определяется по таблицам).

Пример 3

Для компании, занимающейся поставками транспортных комплектующих, включающей 1200 офисов, составлена случайная выборка из 19 офисов. По выборке известно, что исправленное среднее квадратическое отклонение для числа работающих в офисе *s* = 25 человек. Необходимо построить 90%-й доверительный интервал для среднего квадратического отклонения числа работающих в офисе по всей компании.

Построим доверительный интервал для параметра σ по формуле

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}<\sigma^2<\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},$$

где $\chi^2_{\frac{\alpha}{2},n-1},\chi^2_{1-\frac{\alpha}{2},n-1}$ можно найти по таблице критических точек распределения хи-квадрат.

Определяем по таблице $\chi^2_{0,05;18}=28,9;\ \chi^2_{0,95;18}=9,39.$ Далее подставляем в формулу найденные табличные значения и необходимые величины и получаем искомый доверительный интервал:

$$s\sqrt{\frac{(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}}} < \sigma < s\sqrt{\frac{(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}}},\ 25\sqrt{\frac{18}{28,9}} < \sigma < 25\sqrt{\frac{18}{9,39}}.$$

Получаем $19,74 < \sigma < 34,61$.

Задачи для самостоятельного решения

1. Для оценки готовности обработки новой партии комплектов по новой технологии были проанализированы 25 модификаций станков из нескольких подразделений предприятия. Получены следующие результаты в баллах: 107, 90, 114, 88, 117, 110, 103, 120, 96, 122, 93, 100, 121, 110, 135, 85, 120, 89, 100, 126, 90, 94, 99, 116, 111. По этим данным найдите 95%-й интервал для оценки среднего балла готовности обработки всех станков предприятия.

2. По данным семи измерений некоторой величины найдены средняя результатов измерений, равная 32, и выборочная дисперсия, равная 38. Найти границы, в которых с надежностью 0,99 заключено истинное значение измеряемой величины.

3. Найти минимальный объем выборки, при котором с надежностью 0,95 точность оценки математического ожидания нормально распределенной случайной величины (по выборочному среднему) равна 0,2. Среднее квадратическое отклонение составляет 1,5.

1.7. Проверка статистических гипотез

Статистической гипотезой называется предположение о выборке, в частности любое предположение о распределении генеральной совокупности.

Правило, по которому принимается решение, какая из гипотез больше всего соответствует выборочным данным, называется *статистическим критерием* (или *критерием проверки гипотез*).

Статистические гипотезы относительно истинного значения неизвестного параметра θ распределения некоторой случайной величины называют *параметрическими гипотезами*.

Выдвинутая гипотеза называется *нулевой* (или *основной*) *гипотезой* и обозначается H_0. Гипотеза, противоречащая нулевой, называется *альтернативной* (или *конкурирующей*), обозначается H_1.

Цель статистической проверки гипотез состоит в том, чтобы на основании выборочных данных принять решение о справедливости основной гипотезы H_0.

Гипотеза называется *простой*, если сводится к утверждению о том, что значение некоторого неизвестного параметра генеральной совокупности в точности равно заданной величине, в остальных случаях — *сложной*.

При проверке гипотезы возможно появление двух ошибок.

Ошибка 1-го рода заключается в том, что в доле случаев α нулевая гипотеза может оказаться отвергнутой, в то время как она справедлива. Вероятность этой ошибки называется уровнем значимости α.

Ошибка 2-го рода заключается в том, что в доле случаев β нулевая гипотеза принимается, в то время как на самом деле она ошибочна. Вероятность ошибки второго рода β. Вероятность $1 - \beta$ называют мощностью критерия. Критерий называют более мощным, если из всех возможных критериев с заданным уровнем значимости α он обладает наибольшей мощностью.

Основным методом построения наиболее мощных статистических критериев является метод отношения правдоподобия.

Значение критерия, рассчитываемое по специальным правилам на основании выборочных данных, называется наблюдаемым значением критерия $K_{\text{набл}}$. Значения критерия, разделяющие совокупность значений критерия на область допустимых значений и критическую область, определяемые на заданном уровне значимости α по таблицам распределения случайной величины, выбранной в качестве критерия, называют критическими точками $K_{\text{кр}}$.

Областью допустимых значений называют совокупность значений критерия K, при которых нулевая гипотеза H_0 не отклоняется. *Критической областью* называют совокупность значений критерия K, при которых нулевая гипотеза H_0 отклоняется в пользу конкурирующей H_1. Различают левостороннюю, правостороннюю и двусторонние критические области.

Основной принцип проверки статистических гипотез состоит в следующем: если наблюдаемое значение критерия принадлежит критической области, то нулевая гипотеза отклоняется в пользу конкурирующей, если наблюдаемое значение критерия принадлежит области допустимых значений, то нулевую гипотезу нельзя отклонить.

Пример 4

Проверка гипотезы о неизвестном среднем a при неизвестной дисперсии σ^2.

Основная гипотеза $H_0 : a = a_0$. Альтернативная гипотеза H_1 может быть трех видов: $a \neq a_0$, $a > a_0$, $a < a_0$.

Для проверки используется статистика критерия $T = \dfrac{\overline{x} - a_0}{s}\sqrt{n}$.

Для проверки берутся критические точки $t_{\text{кр}}$ распределения Стьюдента с $n - 1$ степенью свободы и уровнем значимости. Для случая $a \neq a_0$ рассматривается двусторонняя критическая область, для остальных двух случаев — односторонняя критическая область.

В случае $a \neq a_0$, если $|T| < t_{\text{кр}}$, то нулевая гипотеза принимается, если $|T| > t_{\text{кр}}$, — отвергается. В случае $a > a_0$, если $T < t_{\text{кр}}$, то нулевая гипотеза принимается, если $T > t_{\text{кр}}$, — отвергается. В случае $a < a_0$, если $T > -t_{\text{кр}}$, то нулевая гипотеза принимается, если $T < -t_{\text{кр}}$, — отвергается.

Пример 5

Проверка гипотезы о неизвестном среднем a при известной дисперсии σ^2.

Основная гипотеза $H_0 : a = a_0$. Альтернативная гипотеза H_1 может быть трех видов: $a \neq a_0$, $a > a_0$, $a < a_0$.

Для проверки используется статистика критерия $U = \dfrac{\overline{x} - a_0}{s}\sqrt{n}$.

В случае $a \neq a_0$ критическая точка $u_{\text{кр}}$ выбирается из условия $\Phi_0\left(u_{\text{кр}}\right) = (1 - \alpha)/2$. Если $|U| < u_{\text{кр}}$, то нулевая гипотеза принимается, если $|U| > u_{\text{кр}}$ — отвергается.

В остальных случаях критическая точка $u_{\text{кр}}$ выбирается из условия $\Phi_0\left(u_{\text{кр}}\right) = 1/2 - \alpha$. В случае $a > a_0$, если $U < u_{\text{кр}}$, то нулевая гипотеза принимается, если $U > u_{\text{кр}}$, — отвергается. В случае $a < a_0$, если $U > -u_{\text{кр}}$, то нулевая гипотеза принимается, если $U < -u_{\text{кр}}$, — отвергается.

1.8. Критерии согласия

Критерии, относящиеся исключительно к виду функции распределения или плотности распределения исследуемой случайной величины, называют *критериями согласия*.

Следует понимать, что проверяют не то, что случайная величина действительно имеет определенный закон распределения, а лишь достаточно ли хорошо наблюдаемые данные согласуются с некоторым законом распределения, что помогает в дальнейшем использовать этот закон для прогнозирования поведения рассматриваемой случайной величины.

Одним из наиболее часто используемых критериев является *критерий хи-квадрат Пирсона*.

Пусть выдвинута гипотеза, что изучаемая случайная величина имеет функцию распределения $F_0(x)$ (или плотность распределения $f(x)$). Пусть далее вся область интервалов изменения величины X разбита на r непересекающихся полуинтервалов $(-\infty; C_1), [C_1, C_2), \ldots, [C_{r-1}; +\infty)$, где $C_0 = -\infty$, $C_r = +\infty$. Пусть m_i — число выборочных значений, попавших в интервал $[C_{i-1}, C_i)$.

Алгоритм проверки гипотезы по критерию следующий.

1. Из генеральной совокупности производят выборку объема n.

2. По выборке составляют сгруппированный статистический ряд.

3. Весь диапазон значений разбивается на r частичных интервалов.

4. На основании гипотетической функции распределения $F_0(x)$ находят вероятности попадания случайной величины X в частичные интервалы:

$$p_i = P\left(C_{i-1} < X < C_i\right) = F_0\left(C_i\right) - F_0(C_{i-1}),\ i = 1, 2, \ldots, r.$$

5. Вычисляется статистика $\chi^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i}$. По теореме Пирсона получаем, что если $x_1, \ldots, x_n$ — выборка из генеральной совокупности с функцией распределения $F_0(x)$, то статистика $\chi^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i}$ имеет при $n \to \infty$ распределение хи-квадрат с $r - 1$ степенями свободы, если основная гипотеза верна.

6. По таблице критических точек распределения хи-квадрат по заданному уровню значимости и числу степеней свободы $r - 1$ находятся критические точки $\chi^2_{\alpha, r-1}$.

7. Проводится сравнение значения критерия с критическим значением: если $\chi^2 > \chi^2_{\alpha, r-1}$, то нулевая гипотеза отвергается в пользу альтернативной; если $\chi^2 < \chi^2_{\alpha, r-1}$, то принимается нулевая гипотеза и считается, что гипотетическая функция распределения согласуется с опытными данными.

Если *значения параметров* гипотетической функции *неизвестны*, то рассматривается сложная гипотеза. При этом нулевая гипотеза заключается в том, что функция распределения имеет вид $F_0\left(x\right) = F(x, \theta_1, \ldots, \theta_k)$ при некоторых неизвестных значениях параметров $(\theta_1, \ldots, \theta_k)$.

Критерий проверки в данном случае имеет вид:

$$\chi^2 = \sum_{i=1}^{r} \frac{(m_i - np_i(\theta_1, \ldots, \theta_k))^2}{np_i(\theta_1, \ldots, \theta_k)}.$$

Так как истинные значения параметров $\theta_1, \ldots, \theta_k$ неизвестны, то при подстановке их оценок, найденных методом максимального правдоподобия, критерий будет получен с меньшим числом степеней свободы $r - k - 1$, где k — число параметров гипотетической функции распределения. Гипотеза принимается, если $\chi^2 < \chi^2_{\alpha, r-k-1}$.

Гипотеза об однородности предполагает, что генеральные совокупности, из которых извлечены выборки, одинаковы и им соответствуют одинаковые функции распределения.

Часто рассматривают случай двух выборок $k = 2$. Пусть есть два ряда наблюдений некоторого признака, и каждый разбит на r групп. Сгруппированный ряд принимает вид: $n_1 : m_1 \ldots m_r$, $n_2 : l_1 \ldots l_r$, где m_i, l_i — число выборочных значений в i-й группе соответственно для первого и второго наблюдений. Статистический критерий для проверки нулевой гипотезы имеет вид

$$\chi^2 = n_1 n_2 \sum_{i=1}^{r} \frac{\left(\frac{m_i}{n_1} - \frac{l_i}{n_2}\right)^2}{m_i + l_i}.$$

При $n \to \infty$, если основная гипотеза верна, критерий имеет предельное распределение хи-квадрат с $r - 1$ степенями свободы. Проверка гипотезы проводится аналогично алгоритму проверки критерия, изложенному выше: если $\chi^2 > \chi^2_{\alpha, r-1}$, то нулевая гипотеза отвергается в пользу альтернативной; если $\chi^2 < \chi^2_{\alpha, r-1}$, то принимается нулевая гипотеза и считается, что гипотетическая функция распределения согласуется с опытными данными.

Задачи для самостоятельного решения

1. Экономический анализ производительности труда предприятий позволил выдвинуть гипотезу о наличии двух типов предприятий с различной средней величиной показателя производительности труда. Выборочное обследование 42 предприятий 1-й группы дало следующие результаты: средняя производительность труда — 119 деталей. Выборочное обследование 35 предприятий 2-й группы показало, что средняя производительность труда составляет 107 деталей. Генеральные дисперсии соответственно равны 127,91 и 135,1. Считая, что выборки извлечены из нормально распределенных генеральных совокупностей X и Y, на уровне значимости 0,05 проверьте, случайно ли полученное различие средних показателей производительности труда или же имеются два типа предприятий с различной средней величиной производительности труда.

2. Партия изделий принимается в том случае, если вероятность того, что изделие окажется соответствующим стандарту, составляет не менее 0,98. Среди случайно отобранных 210 изделий проверяемой партии оказалось 195 соответствующих стандарту. Можно ли на уровне значимости $\alpha = 0{,}02$ принять партию?

3. Инженер по контролю качества проверяет среднее время горения нового вида электроламп. Для проверки в порядке случайной выборки было отобрано 100 ламп, среднее время горения которых составило 1098 часов. Среднее квадратическое отклонение времени горения составляет 112 часов. Используя уровень значимости $\alpha = 0{,}05$, проверьте гипотезу о том, что среднее время горения ламп более 1000 часов.

4. По результатам $n = 7$ независимых измерений найдено, что $\overline{x} = 82{,}48$ мм, а $S = 0{,}08$ мм. Допустив, что ошибки измерения имеют нормальное распределение, проверьте на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : \sigma^2 = 0{,}01$ мм^2 против конкурирующей гипотезы $H_1 : \sigma^2 = 0{,}005$ мм^2.

5. Используя критерий Пирсона, при уровне значимости 0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X по результатам выборки.

X	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9
N	8	10	29	26	30	25	20	24	21

6. Отдел технического контроля проверил n партий однотипных изделий и установил, что число X нестандартных изделий в одной партии имеет эмпирическое распределение, приведенное в таблице, в одной строке которой указано количество x_i нестандартных изделий в одной партии, а в другой строке — количество n_i партий, содержащих x_i нестандартных изделий. Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что случайная величина X (число нестандартных изделий в одной партии) распределена по закону Пуассона.

x_i	0	1	2	3	4	5
n_i	403	370	167	46	12	2

7. При контроле изделий в цехе были измерены диаметры 300 валиков из партии, изготовленной одним станком. В таблице приведены отклонения измеренных диаметров от номинала. На уровне значимости 0,05 проверить гипотезу, что отклонение диаметров от эталона можно описать нормальным распределением.

Границы отклонений	Середина интервала	Число валиков	Границы отклонений	Середина интервала	Число валиков
–30…–25	–27,5	2	0…5	2,5	50
–25…–20	–22,5	9	5…10	7,5	35
–20…–15	–17,5	17	10…15	12,5	20
–15…–10	–12,5	33	15…20	17,5	19
–10…–5	–7,5	40	20…25	22,5	8
–5…0	–2,5	60	25…30	27,5	7

2. ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА

2.1. Введение. Систематические и случайные погрешности

В настоящей главе пойдет речь о статистических методах обработки экспериментальных данных. Под экспериментальными данными мы будем понимать некоторую выборку, каждым элементом которой является пара значений «*фактор — зависимая переменная*». *Фактор* — это значение независимой переменной, характеризующей условия эксперимента, параметры, при которых производятся измерения. Значением фактора может быть одно число, числовой вектор. *Зависимая переменная* — значение измеряемой в эксперименте величины. Это также может быть число либо числовой вектор. Далее мы будем рассматривать только зависимые переменные, описываемые одним числом. В качестве примера рассмотрим эксперимент по измерению величины прогиба балки при различных значениях механической нагрузки и температуры. Факторами в данном эксперименте являются две величины: механическая нагрузка и температура, зависимой переменной — измеряемый в эксперименте прогиб балки, характеризуемый одним числом.

В любом реальном эксперименте на значение зависимой переменной влияет не только значение учитываемых факторов, но и значение неучитываемых факторов: неточное знание условий эксперимента, влияние случайных воздействий внешней среды, несовершенство измерительных приборов и т.д. Влияние неучитываемых факторов вносит *погрешность* в результат эксперимента, которую можно разделить на две составляющих: *систематическую погрешность* и *случайную погрешность*. Случайная погрешность меняется от эксперимента к эксперименту при сохранении неизменными условий эксперимента. Математическое ожидание такой погрешности равно нулю, и, таким образом, ее можно устранить или уменьшить, проведя большое количество однотипных измерений и усреднив по результатам этих измерений. В большинстве случаев распределение величины случайной погрешности можно считать нормальным с нулевым математическим ожиданием. Напротив, систематическую погрешность нельзя устранить усреднением по однотипным измерениям. Систематическая ошибка может быть вызвана тем, что в эксперименте не учтен какой-либо существенный фактор, который остается неизменным в данной серии измерений, но может иметь другое постоянное значение в другой серии. Другой возможный источник систематической погрешности — использование определенного измерительного прибора или методики измерений. Однако далее, если не оговорено иное, мы будем считать, что систематические погрешности отсутствуют.

Наконец, важно заметить, что выделение в эксперименте независимой переменной — фактора — и зависимой переменной зачастую носит условный характер. Математические методы позволяют лишь выявить закономерные связи между значениями условно выбранных факторов и зависимых величин, но не обосновать причинно-следственную связь между ними. Например, статистическое исследование какого-нибудь сообщества людей могло бы выявить закономерность, что люди с более темным цветом кожи в среднем имеют и более темный цвет глаз. Но было бы некорректным утверждать, что темный цвет кожи является причиной темного цвета глаз. В данном исследовании в качестве фактора можно взять цвет глаз, а в качестве зависимой переменной условиться взять цвет кожи, но можно и наоборот.

2.2. Задача регрессии

Рассмотрим экспериментальные данные, представляющие собой выборку из пар чисел (x_i, y_i), где x_i — значение фактора (который в данном эксперименте представлен одним числом), y_i — соответствующее этому значению фактора значение зависимой переменной, i — номер элемента выборки. Всю подобную выборку можно представить виде набора точек на координатной плоскости (рис. 3).

Каждое значение зависимой переменной представим в виде

$$y_i = \varphi(x_i) + \varepsilon_i,$$

где φ — неслучайная функция, ставящая в соответствие каждому числу x_i определенное значение, ε_i — случайная величина, характеризующая случайную погрешность y_i.

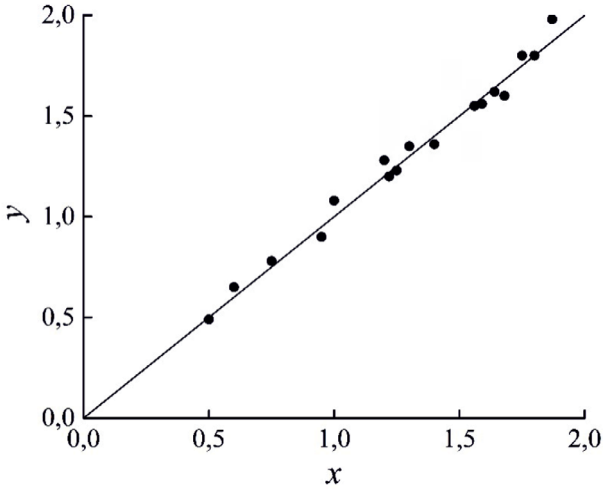

Рис. 3. Экспериментальные результаты (точки) и линия регрессии

Задача регрессии состоит в том, чтобы определить функциональную зависимость $\varphi(x)$, связывающую значения x_i и y_i на фоне случайных погрешностей ε_i. При этом функцию $\varphi(x)$ ищут в некотором классе функций, например в классе линейных функций, классе полиномов *k*-й степени и т.д. Т.е. приближают случайную зависимость СВ Y и СВ X неслучайной функцией $Y \approx y = \varphi(X)$, называемой *линией регрессии величины* Y *на величину* X. На рис. 3 показана регрессия экспериментальных данных на прямую линию.

2.3. Однофакторная линейная регрессия. Методы наибольшего правдоподобия и наименьших квадратов

В случае *однофакторной линейной регрессии* экспериментальные данные (x_i, y_i) приближаются линейной функцией $y = ax + b,$ при этом имеется только один фактор — независимая переменная *x*. Требуется определить коэффициенты a и b так, чтобы для всех (x_i, y_i) выполнялось приближенное равенство $y_i \approx ax_i + b$ с точностью до случайных погрешностей. Метод решения данной задачи может зависеть от того, каковы случайные погрешности ε_i при каждом x_i. Мы рассмотрим случай, когда величины ε_i имеют нормальное распределение с нулевым матожиданием и дисперсией σ^2, одинаковой для всех точек x_1 (т.е. дисперсия случайной погрешности не зависит от *x*).

Обратимся к *методу максимального правдоподобия*. Плотность распределения случайной погрешности имеет вид

$$f(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}},$$

функция правдоподобия для всей выборки

$$L(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n, \sigma) = \prod_{i=1}^{n} f(\varepsilon_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n} \varepsilon_i^2 \Big/ (2\sigma^2)}.$$

Для определения параметров a и b, которые входят в функцию правдоподобия через величины $\varepsilon_i = y_i - (ax_i + b)$, максимизируем функцию *L* по параметрам a и b. Это эквивалентно тому, чтобы минимизировать сумму $\sum_{i=1}^{n} \varepsilon_i^2$ по параметрам a и b. Таким образом, мы приходим к методу наименьших квадратов.

Метод наименьших квадратов (*МНК*) состоит в нахождении таких значений a и b, чтобы сумма квадратов отклонений реальных значений y_i от предсказываемых регрессией значений $\varphi(x_i)$ была минимальна: $\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)^2\xrightarrow[a,b]{}\min$. Для нахождения a и b приравнивают производные данной суммы по a и по b к нулю:

$$\begin{cases}\dfrac{\partial}{\partial a}\sum\limits_i\left(y_i-(ax_i+b)\right)^2 &=0,\\ \dfrac{\partial}{\partial b}\sum\limits_i\left(y_i-(ax_i+b)\right)^2 &=0,\end{cases}$$

откуда

$$\begin{cases}\sum\limits_i x_i\left(y_i-(ax_i+b)\right) &=0,\\ \sum\limits_i\left(y_i-(ax_i+b)\right) &=0,\end{cases}\quad\begin{cases}\sum\limits_i x_iy_i-a\sum\limits_i x_i^2-b\sum\limits_i x_i &=0\\ \sum\limits_i y_i-a\sum\limits_i x_i-bn &=0\end{cases}$$

и в итоге

$$\begin{cases}a=\dfrac{\overline{XY}-\overline{X}\,\overline{Y}}{\overline{X^2}-\left(\overline{X}\right)^2},\\ b=\overline{Y}-a\overline{X},\end{cases}$$

где $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i$, $\overline{X^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2$ и $\overline{XY}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i$ — выборочные средние.

Выражение для коэффициента a можно переписать в виде

$$a=\rho(X,Y)\sqrt{\frac{S_0^2(Y)}{S_0^2(X)}},$$

где $\rho(X,Y)$ — выборочный коэффициент корреляции выборок x_i и y_i, $S_0^2(X)$ и $S_0^2(Y)$ — выборочные дисперсии выборок x_i и y_i соответственно (см. параграф 1.4).

Пользуясь данными выборки, также можно оценить в общем случае неизвестное среднеквадратичное отклонение σ случайной погрешности. Несмещенной оценкой для σ^2 является $S^2(\varepsilon)=\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)^2$, где a и b определяются вышеуказанным способом. В знаменателе стоит число $n-2$, т.к. только $n-2$ из n слагаемых в сумме являются независимыми, поскольку все эти слагаемые связаны через два коэффициента a и b, определяемых из выборки. Если же, например, мы будем искать линию регрессии среди прямых, проходящих через начало координат, то коэффициент b по определению принимается равным нулю, и из выборки определяется только неизвестный коэффициент a. В этом случае несмещенной оценкой для σ^2 является $S^2(\varepsilon)=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-ax_i\right)^2$ (коэффициент a можно найти по МНК, но выражение для него будет отличаться от вышеприведенного (см. задачу 1 ниже).

Пример 6

Набор экспериментальных данных задан выборкой (x_i,y_i) объемом n = 4: $(x_1;y_1)=(1,0;1,2)$, $(x_2;y_2)=(2,0;2,2)$, $(x_3;y_3)=(3,0;3,7)$, $(x_4;y_4)=(4,0;4,9)$. С помощью МНК найти коэффициенты

a и b линейной регрессии Y на X, а также несмещенную оценку для дисперсии случайной погрешности.

Вычислим величины $\overline{X}$, $\overline{Y}$, $\overline{X^2}$, $\overline{XY}$. Для заданной выборки получим $\overline{X}=2,5$, $\overline{Y}=3,0$, $\overline{X^2}=7,5$, $\overline{XY}=9,075$. Тогда по формулам $a=\dfrac{\overline{XY}-\overline{X}\,\overline{Y}}{\overline{X^2}-\left(\overline{X}\right)^2}$ и $b=\overline{Y}-a\overline{X}$ найдем искомые коэффициенты $a=1,26$, $b=-0,15$. Таким образом, прямая $y=1,26x-0,15$ приближает экспериментальные данные наилучшим образом в смысле МНК. Определим теперь величины случайной погрешности $\varepsilon_i=y_i-\left(ax_i+b\right)$. $\varepsilon_1=0,09$, $\varepsilon_2=-0,17$, $\varepsilon_3=0,07$, $\varepsilon_4=0,01$. Тогда несмещенная оценка дисперсии $S^2\left(\varepsilon\right)=\dfrac{1}{2}\sum\limits_{i=1}^{4}\varepsilon_i^2=0,021$.

Задачи для самостоятельного решения

1. Найти выражение для коэффициента a прямой линии регрессии с помощью МНК в предположении, что прямая с достоверностью проходит через начало координат (т.е. что $b = 0$).

2. Набор экспериментальных данных задан выборкой $\left(x_i,y_i\right)$ объемом $n = 5$: $\left(x_1;y_1\right)=\left(-1,0;1,6\right)$, $\left(x_2;y_2\right)=\left(0,0;0,9\right)$, $\left(x_3;y_3\right)=\left(1,0;0,4\right)$, $\left(x_4;y_4\right)=\left(2,0;0,1\right)$, $\left(x_5;y_5\right)=\left(3,0;-0,6\right)$. С помощью МНК найти коэффициенты a и b линейной регрессии Y на X, а также несмещенную оценку для дисперсии случайной погрешности.

3. Набор экспериментальных данных задан выборкой $\left(x_i,y_i\right)$ объемом $n = 4$: $\left(x_1;y_1\right)=\left(1,0;1,6\right)$, $\left(x_2;y_2\right)=\left(2,0;3,1\right)$, $\left(x_3;y_3\right)=\left(3,0;4,3\right)$, $\left(x_4;y_4\right)=\left(4,0;6,0\right)$. В предположении, что прямая линия регрессии с достоверностью проходит через начало координат, найти с помощью МНК коэффициент a линейной регрессии Y на X, а также несмещенную оценку для дисперсии случайной погрешности.

4. Прочность на разрыв бумажной продукции связана с количеством твердой древесины в пульпе. На опытной установке изготавливаются десять образцов, данные для которых приведены ниже. Подберите с помощью этих данных простую линейную регрессионную модель, выражающую связь между прочностью и процентным содержанием твердой древесины.

Прочность	160	171	175	182	184	181	188	193	195	200
Процентное содержание	10	15	15	20	20	20	25	25	28	30

2.4. Полиномиальная регрессия, метод наименьших квадратов

Теперь рассмотрим случай, когда функция регрессии $\varphi\left(x\right)$ выбирается из семейства полиномов k-й степени. Задача *полиномиальной регрессии* сводится к определению неизвестных коэффициентов полинома $\varphi\left(x\right)=a_kx^k+a_{k-1}x^{k-1}+...+a_1x+a_0.$ Неизвестные коэффициенты $a_0,a_1,a_2,...,a_k$ определяются с помощью МНК исходя из условия минимизации $\sum\limits_{i=1}^{n}\varepsilon_i^2=\sum\limits_{i=1}^{n}\left(y_i-\varphi\left(x_i\right)\right)^2\xrightarrow[a]{}\min$, которое сводится к системе линейных уравнений $\sum\limits_{i=1}^{n}y_ix_i^m-\sum\limits_{i=1}^{n}x_i^m\sum\limits_{j=0}^{k}a_jx_i^j=0,\ \ m=0,...,k.$ Поменяв знаки суммирования местами, получим систему уравнений $\hat{\Xi}\vec{a}=\vec{u}$, где векторы $\vec{u}$ и $\vec{a}$ имеют длины $k+1$, матрица $\hat{\Xi}$ — размер $\left(k+1\right)\times\left(k+1\right)$, компоненты вектора $\vec{a}$ — это коэффициенты $a_0,a_1,a_2,...,a_k$, компоненты вектора $\vec{u}$ — числа $u_j=\overline{X^jY}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^jy_i$, компоненты матрицы $\hat{\Xi}$ — числа

$\theta_{jm} = \overline{X^{j+m}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{j+m}$. Решая полученную систему линейных уравнений, найдем коэффициенты $a_0, a_1, a_2, ..., a_k$.

По выборке также можно оценить σ. Несмещенной оценкой для σ^2 является

$$S^2(\varepsilon) = \frac{1}{n-k-1}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k} a_j x_j^k\right)^2,$$

где $a_0, a_1, a_2, ..., a_k$ определяются вышеуказанным способом.

Задачи для самостоятельного решения

1. Набор экспериментальных данных задан выборкой (x_i, y_i) объемом $n = 6$: $(x_1; y_1) = (0,0; 0,2)$, $(x_2; y_2) = (0,5; -1,3)$, $(x_3; y_3) = (1,0; -2,2)$, $(x_4; y_4) = (1,5; -1,4)$, $(x_5; y_5) = (2,0; -0,1)$, $(x_5; y_5) = (2,5; 2,1)$. С помощью МНК для случая $k = 2$ найти коэффициенты a_0, a_1, a_2 полиномиальной регрессии Y на X, а также несмещенную оценку для дисперсии случайной погрешности.

2.5. Многофакторная линейная регрессия

До сих пор мы рассматривали случай только одного фактора. Факторов может быть и больше. В данном разделе рассмотрим частный пример двух факторов: X и Z. Итак, имеется выборка из троек чисел (x_i, z_i, y_i). Задача *многофакторной* (*двухфакторной*) *линейной регрессии* состоит в поиске линейной функции от факторов X и Z, наилучшим образом приближающей зависимую величину Y. Т.е. необходимо найти такие коэффициенты b, a_1, a_2, чтобы $y_i \approx a_1 x_i + a_2 z_i + b$ для всех i. При этом, как и раньше, предполагаем, что при всех значениях x_i, z_i случайная ошибка имеет нормальное распределение с нулевым матожиданием и одной и той же дисперсией σ^2. Согласно МНК для определения коэффициентов b, a_1, a_2 необходимо минимизировать сумму квадратов отклонений $\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left(y_i - (a_1 x_i + a_2 z_i + b)\right)^2$ по этим коэффициентам. Дифференцируя эту сумму по коэффициентам b, a_1, a_2, получим систему из трех линейных уравнений, которую можно записать в виде матрицы $\hat{\Xi}\vec{a} = \vec{u}$, где $\vec{u} = \begin{pmatrix} \overline{Y} \\ \overline{YX} \\ \overline{YZ} \end{pmatrix}$, $\vec{a} = \begin{pmatrix} b \\ a_1 \\ a_2 \end{pmatrix}$, $\hat{\Xi} = \begin{pmatrix} 1 & \overline{X} & \overline{Z} \\ \overline{X} & \overline{X^2} & \overline{XZ} \\ \overline{Z} & \overline{XZ} & \overline{Z^2} \end{pmatrix}$.

Решая данную систему уравнений, получим

$$a_1 = \frac{S(Y)}{S(X)}\frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{1-\rho_{XZ}^2},\ a_2 = \frac{S(Y)}{S(Z)}\frac{\rho_{YZ} - \rho_{XZ}\rho_{XY}}{1-\rho_{XZ}^2},\ b = \overline{Y} - a_1\overline{X} - a_2\overline{Z},$$

где $S(X)$, $S(Y)$, $S(Z)$ соответственно — исправленные выборочные среднеквадратические отклонения выборок X, Y и Z, ρ_{XY} — выборочный коэффициент корреляции выборок X и Y, ρ_{XZ} — выборок X и Z, ρ_{YZ} — выборок Y и Z.

Задачи для самостоятельного решения

1. Выход химического процесса связан с концентрацией реагента и рабочей температурой. Подберите многофакторную линейную регрессионную модель для следующих данных.

Выход	Концентрация	Температура
81	1,0	65
89	1,0	80
83	2,0	65
91	2,0	80
79	1,0	65
87	1,0	80
84	2,0	65
90	2,0	80

2. Агент по продаже некоторого товара анализирует систему его доставки. Его интересует определение времени, необходимого для обслуживания магазина. На время доставки влияют два наиболее важных фактора: число ящиков товара и максимальное расстояние, на которое он должен доставляться. По собранным данным (см. табл. ниже) построить линейную регрессионную модель.

Число ящиков X	Расстояние Z	Время	Число ящиков X	Расстояние Z	Время
10	30	24	14	34	28
15	25	27	16	29	31
10	40	29	22	37	39
20	18	31	24	20	33
25	22	25	17	26	30
18	31	33	13	27	25
12	26	26	30	23	42
			24	33	40

2.6. Интервальные оценки в задачах регрессии

Выше мы привели точечные оценки для σ^2, для коэффициентов регрессии $a_0, a_1, a_2, ..., a_k$. Все эти оценки вычислялись исходя из выборочных данных. Поскольку от выборки к выборке экспериментальные данные могут случайно меняться, то и полученные оценки для σ^2, $a_0, a_1, a_2, ..., a_k$ также являются случайными величинам. Для определения точности точечных оценок коэффициентов регрессии возможно получить *интервальные оценки* этих коэффициентов, т.е. построить доверительные интервалы, накрывающие истинное значение коэффициентов функции регрессии с доверительной вероятностью γ.

Для случая линейной однофакторной регрессии точечная оценка σ^2 имеет вид $S^2(\varepsilon) = \dfrac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2$ (см. параграф 2.3). Каждая величина ε_i имеет нормальное распределение с нулевым матожиданием и дисперсией σ^2, а значит, сумма $\sum_{i=1}^{n}\dfrac{\varepsilon_i^2}{\sigma^2}$ есть сумма $n-2$ квадратов СВ, имеющих стандартное нормальное распределение. Такая сумма по определению имеет распределение «хи-квадрат» с $n-2$ степенями свободы χ^2_{n-2}. *Доверительный интервал* (h_1, h_2) для σ^2 построим так, что $P\left(S^2(\varepsilon) > h_2\right) = P\left(S^2(\varepsilon) < h_1\right) = (1-\gamma)/2 = \alpha/2$, т.е. на полуинтервал больше h_2 и полуинтервал меньше h_1 приходятся одинаковые вероятности $(1-\gamma)/2$. Так как СВ $\sum_{i=1}^{n}\dfrac{\varepsilon_i^2}{\sigma^2} = \dfrac{(n-2)S^2(\varepsilon)}{\sigma^2}$ имеет распределение χ^2_{n-2}, то $P\left(\dfrac{(n-2)S^2(\varepsilon)}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2},n-2}\right) = 1-\dfrac{\alpha}{2}$, где $\chi^2_{1-\frac{\alpha}{2},n-2}$ — *квантиль* уровня $1-\dfrac{\alpha}{2}$ распреде-

ления χ^2_{n-2} (находится по соответствующим таблицам). Аналогично $P\left(\frac{(n-2)S^2(\varepsilon)}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-2}\right) = \frac{\alpha}{2}$, $\chi^2_{\frac{\alpha}{2},n-2}$ — *квантиль* уровня $\frac{\alpha}{2}$ распределения χ^2_{n-2}. Т.е. событие $\chi^2_{\frac{\alpha}{2},n-2} < \frac{(n-2)S^2(\varepsilon)}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2},n-2}$ будет иметь вероятность $\left(1-\frac{\alpha}{2}\right)-\frac{\alpha}{2} = 1-\alpha = \gamma,$ т.е. оно и будет определять искомый доверительный интервал. Преобразуя неравенство $\chi^2_{\frac{\alpha}{2},n-2} < \frac{(n-2)S^2(\varepsilon)}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2},n-2},$ получим доверительный интервал $\frac{(n-2)S^2(\varepsilon)}{\chi^2_{1-\frac{\alpha}{2},n-2}} < \sigma^2 < \frac{(n-2)S^2(\varepsilon)}{\chi^2_{\frac{\alpha}{2},n-2}}$.

Построим теперь *доверительный интервал* для *коэффициента* a в *задаче однофакторной линейной регрессии*. $a = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \left(\overline{X}\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\frac{1}{n}\sum_{i=1}^{n} y_i}{\overline{X^2} - \left(\overline{X}\right)^2}$ (см. параграф 2.3). Каждое выборочное значение x_i будем рассматривать как неслучайную величину, точно задаваемую экспериментатором, а случайными будут лишь значения y_i. Тогда видно, что a является линейной функцией от y_i с неслучайными коэффициентами, зависящими от x_i. Любая линейная комбинация нормальных СВ (а все y_i являются нормальными СВ) также является нормальной СВ. Следовательно, оценка a имеет нормальное распределение. Ее матожидание равно истинному значению a (будем обозначать его $a_{\text{ист}}$) по теореме Гаусса — Маркова (оценка a несмещенная). Вычислим дисперсию оценки a.

$$Da = \frac{1}{\left(\overline{X^2} - \left(\overline{X}\right)^2\right)^2} D\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{D\left(\sum_{i=1}^{n} y_i\left(x_i - \overline{X}\right)\right)}{n^2\left(\overline{X^2} - \left(\overline{X}\right)^2\right)^2} =$$

$$= \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 Dy_i}{n^2\left(S_0^2(X)\right)^2} = \frac{\sigma^2\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n\left(S_0^2(X)\right)^2} = \frac{\sigma^2 S_0^2(X)}{n\left(S_0^2(X)\right)^2} = \frac{\sigma^2}{nS_0^2(X)} = \frac{\sigma^2}{(n-1)S^2(X)}.$$

Мы воспользовались тем, что отдельные слагаемые в сумме $\sum_{i=1}^{n} y_i\left(x_i - \overline{X}\right)$ независимы друг от друга ввиду независимости y_i и тем, что дисперсия суммы независимых СВ равна сумме дисперсий. Таким образом, СВ a имеет нормальное распределение с матожиданием $a_{\text{ист}}$ и дисперсией $\frac{\sigma^2}{nS_0^2(X)}$. Так как точное значение σ^2 в общем случае неизвестно, будем пользоваться оценкой $S^2(\varepsilon)$ для σ^2. СВ $\frac{a - a_{\text{ист}}}{\sqrt{\frac{S^2(\varepsilon)}{(n-1)S^2(X)}}}$ имеет распределение Стьюдента с n – 2 степенями свободы. Итак, должно быть выполнено неравенство

$$P\left(\frac{|a-a_{\text{ист}}|}{\sqrt{\frac{S^2(\varepsilon)}{(n-1)S^2(X)}}} < t_{\gamma,n-2}\right) = \gamma,$$

где $t_{\gamma,n-2}$ — критическая точка распределения Стьюдента с $n-2$ степенями свободы при уровне значимости $\alpha = 1-\gamma$ в случае двухсторонней критической области. Величины $t_{\gamma,n-2}$ можно найти в соответствующих таблицах. Таким образом, получим доверительный интервал для $a_{\text{ист}}$ при доверительной вероятности γ:

$$a - t_{\gamma,n-2}\sqrt{\frac{S^2(\varepsilon)}{(n-1)S^2(X)}} < a_{\text{ист}} < a + t_{\gamma,n-2}\sqrt{\frac{S^2(\varepsilon)}{(n-1)S^2(X)}}.$$

Аналогично, т.к. коэффициент $b = \overline{Y} - a\overline{X}$ линеен по выборочным значениям y_i (как и коэффициент *a*), то он также имеет нормальное распределение с матожиданием $b_{\text{ист}}$, для которого мы построим доверительный интервал. Найдем дисперсию СВ *b*. Подставляя в $b = \overline{Y} - a\overline{X}$ выражение для *a*, найдем $b = \dfrac{\overline{X^2}\,\overline{Y} - \overline{XY}\,\overline{X}}{\overline{X^2} - \left(\overline{X}\right)^2} = \dfrac{\sum_{i=1}^{n} y_i\left(\overline{X^2} - x_i\overline{X}\right)}{nS_0^2(X)}$.

$$Db = \frac{\sum_{i=1}^{n}\left(\overline{X^2} - x_i\overline{X}\right)^2 Dy_i}{\left(nS_0^2(X)\right)^2} = \frac{\sigma^2\left(n\left(\overline{X^2}\right)^2 - 2\overline{X}\,\overline{X^2}\sum_{i=1}^{n}x_i + \left(\overline{X}\right)^2\sum_{i=1}^{n}x_i^2\right)}{\left(nS_0^2(X)\right)^2} =$$

$$= \frac{\sigma^2\overline{X^2}\left(\overline{X^2} - \left(\overline{X}\right)^2\right)}{n\left(S_0^2(X)\right)^2} = \frac{\sigma^2\overline{X^2}}{nS_0^2(X)} = \frac{S_0^2(X) + \left(\overline{X}\right)^2}{nS_0^2(X)}\sigma^2 = \left(\frac{1}{n} + \frac{\left(\overline{X}\right)^2}{(n-1)S^2(X)}\right)\sigma^2.$$

Поскольку σ^2 неизвестно, то в качестве оценки дисперсии *Db* будем рассматривать величину $\left(\dfrac{1}{n} + \dfrac{\left(\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)$. Тогда СВ $\dfrac{b - b_{\text{ист}}}{\sqrt{\left(\dfrac{1}{n} + \dfrac{\left(\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)}}$ будет иметь распределение Стьюдента с $n-2$ степенями свободы. Окончательно *доверительный интервал* для $b_{\text{ист}}$ при доверительной вероятности γ имеет вид

$$b - t_{\gamma,n-2}\sqrt{\left(\frac{1}{n} + \frac{\left(\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)} < b_{\text{ист}} < b + t_{\gamma,n-2}\sqrt{\left(\frac{1}{n} + \frac{\left(\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)},$$

где $t_{\gamma,n-2}$ — квантиль уровня $\dfrac{1+\gamma}{2}$ распределения Стьюдента с $n-2$ степенями свободы.

Наконец, построим *интервальную оценку* для *оценки величины Y по функции регрессии* при произвольном фиксированном x^*, т.е. интервальную оценку для СВ $y^* = ax^* + b$, где коэффициенты *a* и *b* — коэффициенты линии регрессии, определяемые по выборочным данным. Построив доверительный интервал для $y^* = ax^* + b$ при каждом x^*, мы зададим «коридор», в котором могут лежать зна-

чения зависимой величины при любых x^* с заданной доверительной вероятностью. Итак, ax^*+b линейно зависит от выборочных значений y_i, а все коэффициенты этой линейной зависимости определяются x^* и выборочными x_i, которые мы считаем фиксированными и неслучайными. Таким образом, $y^*=ax^*+b$ представляет собой СВ с нормальным распределением с матожиданием $a_{\text{ист}}x^*+b_{\text{ист}}$. Найдем дисперсию y^*. Мы не можем просто вычислить дисперсию y^*, исходя из выражений для дисперсий *a* и *b*, т.к. $D\left(ax^*+b\right)\neq\left(x^*\right)^2 Da+Db$ ввиду зависимости СВ *a* и *b*. Поэтому, подставив выражения для *a* и *b*, сначала получим $y^*=\overline{Y}+\dfrac{\overline{XY}-\overline{X}\,\overline{Y}}{S_0^2(X)}\left(x^*-\overline{X}\right)$.

$$Dy^*=\frac{D\left(\sum\limits_{i=1}^{n}y_i\left(S_0^2(X)+\left(x_i-\overline{X}\right)\left(x^*-\overline{X}\right)\right)\right)}{\left(nS_0^2(X)\right)^2}=\sigma^2\frac{\sum\limits_{i=1}^{n}\left(S_0^2(X)+\left(x_i-\overline{X}\right)\left(x^*-\overline{X}\right)\right)^2}{\left(nS_0^2(X)\right)^2}=$$

$$=\sigma^2\frac{\sum\limits_{i=1}^{n}\left(S_0^2(X)+\left(x_i-\overline{X}\right)\left(x^*-\overline{X}\right)\right)^2}{\left(nS_0^2(X)\right)^2}=\sigma^2\left(\frac{1}{n}+\frac{\left(x^*-\overline{X}\right)^2}{nS_0^2(X)}\right)=\sigma^2\left(\frac{1}{n}+\frac{\left(x^*-\overline{X}\right)^2}{(n-1)S^2(X)}\right).$$

Поскольку σ^2 неизвестно, то в качестве оценки дисперсии y^* будем рассматривать величину $\left(\dfrac{1}{n}+\dfrac{\left(x^*-\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)$. Тогда СВ $\dfrac{y^*-\left(a_{\text{ист}}x^*+b_{\text{ист}}\right)}{\sqrt{\left(\dfrac{1}{n}+\dfrac{\left(x^*-\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)}}$ будет иметь распределение Стьюдента с $n-2$ степенями свободы. Окончательно доверительный интервал для $a_{\text{ист}}x^*+b_{\text{ист}}$ при доверительной вероятности γ имеет вид $ax^*+b-t_{\gamma,n-2}\sqrt{\left(\dfrac{1}{n}+\dfrac{\left(x^*-\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)}<a_{\text{ист}}x^*+b_{\text{ист}}$ и $a_{\text{ист}}x^*+b_{\text{ист}}<ax^*+b+t_{\gamma,n-2}\sqrt{\left(\dfrac{1}{n}+\dfrac{\left(x^*-\overline{X}\right)^2}{(n-1)S^2(X)}\right)S^2(\varepsilon)}$, где $t_{\gamma,n-2}$ — квантиль уровня $\dfrac{1+\gamma}{2}$ распределения Стьюдента с $n-2$ степенями свободы.

Теперь рассмотрим проверку гипотезы о равенстве нулю коэффициента $a_{\text{ист}}$ в однофакторной линейной регрессии $H_0: a_{\text{ист}}=0$ при альтернативной $H_1: a_{\text{ист}}\neq 0$. Основная гипотеза, по сути, означает гипотезу о том, что величина *Y* независима от *X*, если модель линейной регрессии верна. Либо возможен вариант, что модель линейной регрессии неверна, но между *Y* и *X* имеется более сложная нелинейная зависимость (этот вопрос будет рассмотрен в следующем параграфе). В предположении справедливости основной гипотезы (и справедливости самой линейной модели регрессии) рассмотрим критерий $T=\dfrac{a}{\sqrt{\dfrac{S^2(\varepsilon)}{(n-1)S^2(X)}}}$. Данная СВ *T* имеет распределение Стьюдента с $n-2$ степенями свободы (см. предыдущий параграф). Поэтому, если найденное по выборке значение *T* лежит в интервале $-t_{\gamma,n-2}<T<t_{\gamma,n-2}$, то основная гипотеза принимается. Если *T* лежит вне указан-

ного интервала, то основная гипотеза отвергается и принимается альтернативная гипотеза $a_{\text{ист}} \neq 0$, т.е. считается, что между Y и X есть статистически значимая линейная связь.

Задачи для самостоятельного решения

1. Используя данные задачи 4 из параграфа 2.3, постройте 95%-е доверительные интервалы для коэффициентов a и b однофакторной линейной регрессии.

2. Проводилось исследование влияния скорости перемешивания на количество примесей в краске, получаемой химическим способом. Оно дало следующие результаты.

Скорость перемешивания, об./мин	20	22	24	26	28	30
Примеси, %	8,4	9,5	11,8	10,4	13,3	14,8
Скорость перемешивания, об./мин	32	34	36	38	40	42
Примеси, %	13,2	14,7	16,4	16,5	18,9	18,5

Найти функцию линейной регрессии и построить 95%-е доверительные интервалы для коэффициентов a и b, а также для дисперсии случайной погрешности.

3. Закон Хаббла в астрономии гласит: скорость удаления галактики прямо пропорциональна расстоянию до нее. В таблице ниже указаны расстояния Y (в миллионах световых лет) и скорости X (в сотнях миль в секунду) для 11 галактических созвездий. Построить регрессионную модель вида $Y = aX$ и сделать интервальную оценку для коэффициента a с доверительной вероятностью 90 % (поскольку коэффициент b полагается достоверно равным 0, то при построении интервальной оценки для a используется распределение Стьюдента с $(n - 1)$-й степенью свободы, а не с $n - 2$).

Созвездие	X	Y
Дева (Virgo)	22	7,5
Пегас (Pegasus)	68	24
Персей (Perseus)	108	32
Волосы Вероники (Coma Berenices)	137	47
Большая Медведица (Ursa Major No. 1)	255	93
Большая Медведица (Ursa Major No. 2)	700	260
Лев (Leo)	315	120
Северная Корона (Corona Borealis)	390	134
Близнецы (Gemini)	405	144
Волопас (Bootes)	685	245
Гидра (Hydra)	1100	380

4. В таблице ниже представлены 50 пар наблюдений. Рассматривались переменные: X — длина «линии жизни» на левой руке в сантиметрах (с точностью до ближайших 0,15 см), Y — продолжительность жизни человека. Верно ли, что Y и X связаны линейной регрессионной зависимостью (проверить гипотезу $a_{\text{ист}} = 0$)? Для вычисления оценок коэффициентов используйте то, что $\sum_i x_i = 459{,}9$, $\sum_i x_i^2 = 4308{,}57$, $\sum_i y_i = 3333$, $\sum_i x_i y_i = 30549{,}75$.

X	Y	X	Y	X	Y	X	Y
9,75	19	7,20	61	7,80	69	6,00	76
9,00	40	7,95	62	10,05	69	8,85	77
9,60	42	8,85	62	10,50	70	9,00	80
9,75	42	8,25	65	9,15	71	9,75	82
11,25	47	8,85	65	9,45	71	10,65	82
9,45	49	9,75	65	9,45	71	13,20	82
11,25	50	8,85	66	9,45	72	7,95	83

X	Y	X	Y	X	Y	X	Y
9,00	54	9,15	66	8,10	73	7,95	86
7,95	56	10,20	66	8,85	74	9,15	88
12,00	56	9,15	67	9,60	74	9,75	88
8,10	57	7,95	68	6,45	75	9,00	94
10,20	57	8,85	68	9,75	75		
8,55	58	9,00	68	10,20	75		

2.7. Проверка адекватности линейной и полиномиальной моделей регрессии

Часто при проведении анализа экспериментальных данных вид зависимости исследуемой величины от фактора (факторов) неизвестен. Поэтому необходимо на основании экспериментальных данных предложить вид зависимости и найти ее параметры. После того, как вид зависимости (функция) предложен, необходимо показать, что предложенная модель достаточно хорошо описывает экспериментальные данные. Начнем рассмотрение с модели однофакторной линейной регрессии и *проверки адекватности описания экспериментальной зависимости однофакторной линейной моделью*.

В случае, если между X и Y имеется строгая (неслучайная) линейная зависимость $y_i = ax_i + b$, выборочный коэффициент корреляции ρ_{XY} равен ± 1. Действительно,

$$\rho_{XY} = \frac{\frac{1}{n}\sum_i \left(x_i - \overline{X}\right)\left(ax_j + b - \left(a\overline{X} + b\right)\right)}{\sqrt{\frac{1}{n}\sum_i \left(x_i - \overline{X}\right)^2 \frac{1}{n}\sum_j \left(ax_j + b - \left(a\overline{X} + b\right)\right)^2}} = \frac{a\sum_i \left(x_i - \overline{X}\right)^2}{\sqrt{a^2\left(\sum_i \left(x_i - \overline{X}\right)^2\right)^2}} = \frac{a}{|a|}.$$

Если же зависимость X и Y случайная, но все же близка к линейной, то по степени близости $|\rho_{XY}|$ к 1 можно судить о степени адекватности описания экспериментальных данных линейной регрессионной моделью.

Можно предложить и другой, более универсальный *метод проверки адекватности регрессионной модели экспериментальным данным*, который работает и в случае полиномиальной регрессионной модели, и в случае многофакторной линейной регрессии, и даже в случае, когда функция регрессии является произвольной функцией одного или нескольких факторов при условии лишь, что эта функция регрессии линейно зависит от неизвестных коэффициентов. Данный метод состоит в вычислении *коэффициента множественной корреляции R*, который равен

$$R = \sqrt{\rho_{Y\hat{Y}}} = \sqrt{\frac{\sum_i \left(y_i - \overline{Y}\right)\left(\hat{y}_i - \overline{\hat{Y}}\right)}{\sqrt{\sum_i \left(y_i - \overline{Y}\right)^2 \sum_j \left(\hat{y}_i - \overline{\hat{Y}}\right)^2}}},$$

где $\hat{y}_i = \varphi\left(x_i, z_i, \ldots\right)$ — оцененные по регрессии значения y.

Если для каждого набора значений независимых переменных $x_i, z_i, \ldots$ функция регрессии будет давать значение $\hat{y}_i$, в точности равное экспериментальному y_i (что означает полную адекватность регрессионной модели экспериментальным данным), то на координатной плоскости $(\hat{y}, y)$ точки $(\hat{y}_i, y_i)$ будут лежать вдоль прямой, соответствующей биссектрисе первого квадранта координатной плоскости (т.к. $y_i = \hat{y}_i$). И, поскольку точки $(\hat{y}_i, y_i)$ лежат вдоль прямой, выборочный коэффициент корреляции $\rho_{Y\hat{Y}}$ между ними должен в точности равняться 1. В реальном случае случайной выборки $\rho_{Y\hat{Y}}$ хотя и меньше 1, но все же близок к 1, если $\hat{y}_i$ близко к y_i. Таким образом, по степени близости

R к 1 можно судить об адекватности различных моделей регрессии и сравнивать их между собой. Например, при сравнии линейной и квадратичной полиномиальных моделей регрессии в некотором случае может получиться, что полиномиальная модель дает значительно более близкое к 1 значение R. Однако не следует использовать в качестве функции регрессии полиномы слишком большой степени (больше 4-й), лучше использовать полиномы наименьшей возможной степени, при которой R уже достаточно близок к 1. Ясно, что при наличии выборки из n элементов через них всегда можно точно провести кривую полинома $(n – 1)$-й степени, однако коэффициенты такого полинома будут при этом значительно изменяться от выборки к выборке, что неприемлемо.

Задачи для самостоятельного решения

1. Используя данные задачи 4 из параграфа 2.3, проверить на адекватность линейную регрессионную модель.

2. Используя данные задачи 1 из параграфа 2.5, проверить на адекватность модель многофакторной линейной регрессии.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Просветов Г.И. Статистика: задачи и решения. – Москва : Альфа-Пресс, 2014. – 495 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика : учеб. пособие для бакалавров. – 12-е изд. – Москва : Юрайт, 2013. – 479 с.

3. Статистические методы обработки, планирования инженерного эксперимента: учеб. пособие. – Благовещенск : Дальневосточный государственный аграрный университет, 2015. – 93 с. – URL: http://www.iprbookshop.ru/55912.html (дата обращения: 28.02.2020).

4. Горохов В.Л. Планирование и обработка экспериментов : учеб. пособие / В.Л. Горохов, В.В. Цаплин. – Санкт-Петербург : Санкт-Петербургский государственный архитектурно-строительный университет, ЭБС АСВ, 2016. – 88 с. – URL: http://www.iprbookshop.ru/63623.html (дата обращения: 28.02.2020).