Учреждение образования
«Брестский государственный технический университет»

Машиностроительный факультет

Кафедра «Машиностроение и эксплуатация автомобилей»

СОГЛАСОВАНО Заведующий кафедрой __________ С. В. Монтик «13» 05 ______2025 г.	СОГЛАСОВАНО Декан факультета __________ С. Р. Онысько «25» 05 ______2025 г.

ЭЛЕКТРОННЫЙ УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

«АНАЛИЗ И УПОРЯДОЧЕНИЕ ИСХОДНЫХ ДАННЫХ ПРИ СТАТИСТИЧЕСКОЙ ОБРАБОТКЕ РЕЗУЛЬТАТОВ НАУЧНЫХ ИССЛЕДОВАНИЙ»

(название дисциплины)

для специальности

7-06-0714-02 Инновационные технологии в машиностроении

Профилизация: Аддитивные и субтрактивные технологии

Составитель: Монтик Сергей Владимирович, доцент, канд. техн. наук, доцент

Рассмотрено и утверждено на заседании Научно-методического совета университета 26.06.2025 г., протокол № 4.

рег № УМК 24/25-128

Пояснительная записка

Актуальность изучения дисциплины

Владение методами подготовки, упорядочения и анализа исходных данных в процессе статистических исследований необходимо при разработке нового и модернизации существующего оборудования, оснастки и технологических процессов механосборочного производства. В связи с этим освоение дисциплины «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований», в которой изучаются данные методы является актуальным и необходимым условием подготовки высококвалифицированных магистров в области технологии машиностроения.

Цель и задачи дисциплины

Целью преподавания учебной дисциплины «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» является формирование устойчивой системы знаний, навыков и профессиональных компетенций в области анализа и упорядочения исходных данных при статистической обработке результатов научных исследований

Основными задачами изучения дисциплины являются освоение:

- математических основ анализа погрешности измерений;
- методов разработки вероятностных математических моделей;
- методов проведения дисперсионного, корреляционного, регрессионного анализа;
- методов планирование эксперимента при одно- и многофакторном регрессионном анализе, а также в задачах оптимизации;
- программного обеспечения для выполнения статистического анализа данных.

Электронный учебно-методический комплекс (ЭУМК) объединяет структурные элементы научно-методического обеспечения образования и представляет собой сборник материалов теоретического и практического характера для организации работы магистрантов специальности 7-06-0714-02 Инновационные технологии в машиностроении дневной и заочной форм получения образования по изучению дисциплины «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований».

ЭУМК разработан на основании Положения об учебно-методическом комплексе на уровне высшего образования, утвержденного Постановлением Министерства образования Республики Беларусь от 8 ноября 2022 г. № 427, и предназначен для реализации требований учебной программы по учебной дисциплине «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» для специальности 7-06-0714-02 Инновационные технологии в машиностроении.

Цели ЭУМК:

- обеспечение качественного методического сопровождения процесса обучения;

- организация эффективной самостоятельной работы студентов.

Содержание и объем ЭУМК соответствуют образовательному стандарту углубленного высшего образования ОСВО 7-06-0714-02-2023 Инновационные технологии в машиностроении и учебной программе дисциплины компонента учреждения высшего образования «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований». Материал представлен на требуемом методическом уровне и адаптирован к современным образовательным технологиям.

Структура электронного учебно-методического комплекса по дисциплине «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований»:

Теоретический раздел содержит материалы для теоретического изучения учебной дисциплины и представлен конспектом лекций.

Практический раздел содержит материалы для проведения практических учебных занятий в виде методических указаний для выполнения практических работ.

Раздел контроля знаний содержит материалы для текущей и промежуточной аттестации по учебной дисциплине - зачета, позволяющие определить соответствие результатов учебной деятельности обучающихся требованиям образовательного стандарта высшего образования и учебно-программной документации образовательных программ высшего образования.

Вспомогательный раздел включает учебную программу учреждения высшего образования по учебной дисциплине «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований».

Рекомендации по организации работы с ЭУМК:

- лекции проводятся с использованием представленных в ЭУМК конспект лекций, презентаций, персонального компьютера и мультимедийного проектора; при подготовке к зачету, практическим занятиям студенты могут использовать конспект лекций;

- практические занятия проводятся в компьютерном классе с использованием программного обеспечения MS Office.

ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

Конспект лекций по дисциплине

«АНАЛИЗ И УПОРЯДОЧЕНИЕ ИСХОДНЫХ ДАННЫХ ПРИ СТАТИСТИЧЕСКОЙ ОБРАБОТКЕ РЕЗУЛЬТАТОВ НАУЧНЫХ ИССЛЕДОВАНИЙ»

для специальности

7-06-0714-02 Инновационные технологии в машиностроении

Профилизация: Аддитивные и субтрактивные технологии

2025

1. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

1.1. Основные понятия и задачи корреляционно-регрессионного анализа (КРА)

Для технологических экспериментов характерны статистические объекты исследований, в которых имеют место стохастические или корреляционные взаимосвязи между зависимыми и независимыми переменными. Найти математическое описание этих взаимосвязей – значит получить математическую модель изучаемого процесса или объекта. В задачу корреляционно-регрессионного анализа (КРА) входит получение на основании экспериментальных данных о работе объекта (протекании процесса) математической модели объекта (процесса) и её анализ.

Ценность математической модели заключается в том, что с её помощью можно решать задачи оптимизации процесса (объекта) и предсказанию его результатов (характеристик) при изменении условий его протекания (эксплуатации). Вторую задачу называют интерполяционной.

Методы КРА могут быть применимы только для таких параметров, которые при качественном их рассмотрении, т. е. при изучении физической природы процесса (объекта), являются взаимосвязанными. На первом этапе применения этих методов обычно оценивают степень тесноты взаимосвязи значений функции отклика (y) с одним или несколькими независимыми переменными (x). Первая задача решается с помощью коэффициента парной корреляции r_{xy}, вторая – с помощью коэффициента множественной корреляции $R_{y.x_1,x_2...x_m}$

$$r_{yx} = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{nS_xS_y}, \qquad (1.1)$$

где $\overline{y}$, $\overline{x}$, S_y, S_x – средние арифметические значения и средние квадратические отклонения значений соответственно y_i и x_i в рассматриваемой выборке;

n – объём выборки (обычно n = 50–200 шт.).

Обозначим зависимую переменную цифрой 1, а независимые переменные цифрами 2, 3 … m, тогда коэффициенты парной корреляции запишутся как r_{12}, r_{13} … r_{1m}, r_{23}, r_{24} и т. д., а коэффициент множественной корреляции между y и x_1, x_2 … x_m – $R_{1.23\,\ldots\,m}$

Коэффициент множественной корреляции с использованием метода определителей находится по формуле

$$R_{1.23\,\ldots\,m} = \sqrt{1 - \frac{D}{D_{11}}}, \tag{1.2}$$

где D – определитель, составленный из всех коэффициентов парной корреляции:

$$D = \begin{vmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1m} \\ r_{21} & 1 & r_{23} & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & r_{m3} & \ldots & 1 \end{vmatrix}. \tag{1.3}$$

D_{11} – определитель, получающийся из определителя D исключением нулевого (первого слева) столбца и нулевой (верхней) строки:

$$D_{11} = \begin{vmatrix} 1 & r_{23} & \ldots & r_{2m} \\ r_{32} & 1 & \ldots & r_{3m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{m2} & r_{m3} & \ldots & 1 \end{vmatrix}. \tag{1.4}$$

Для трёх переменных

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}. \quad (1.5)$$

Значения r_{yx} и $R_{y.x_1,x_2\ldots x_m}$ находятся в пределах ±1. Если они достоверны, т. е. существенно отличаются от нуля, значит между исследуемыми факторами имеется линейная корреляционная зависимость. В противном случае эта зависимость отсутствует либо является существенно нелинейной. Если r_{yx} или $R_{y.x_1,x_2\ldots x_m}$ равны +1 или –1, что встречается крайне редко, значит между исследуемыми факторами существует функциональная взаимосвязь. Знак величин r_{yx} и $R_{y.x_1,x_2\ldots x_m}$ говорит о прямом (+) или обратном (–) характере взаимосвязей между исследуемыми факторами.

Для оценки степени тесноты нелинейных одно- и многофакторных взаимосвязей используются корреляционные отношения η.

Если корреляционный анализ подтвердил наличие взаимосвязей между исследуемыми факторами, то на следующем этапе обработки экспериментальных данных выбирают математическую модель с помощью регрессионного анализа, которая наилучшим образом описывает указанные взаимосвязи. Уравнение, с помощью которого могут быть найдены числовые значения выборочных средних функции отклика в зависимости от соответствующих значений независимых переменных, называется уравнением регрессии. В общем случае оно может быть записано в следующем виде:

$$\overline{y} = f\left(\overline{x}_1, \overline{x}_2, \overline{x}_m\right). \quad (1.6)$$

Операция замены одной функции другой, в каком-то смысле эквивалентной функцией, называется аппроксимацией. При аппроксимации неизвестных функций отклика в математической статистике наиболее часто используют полиномиальные модели. Степень полинома определяется максимальной степенью входящих в них переменных.

Выпишем полиномы различных степеней для случая двух факторов. Полином нулевой степени:

$$y = b_0. \tag{1.7}$$

Полином первой степени:

$$y = b_0 + b_1 x_1 + b_2 x_2. \tag{1.8}$$

Полином второй степени:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2. \tag{1.9}$$

Полином третьей степени:

$$\begin{gathered} y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + \\ + b_{112} x_1^2 x_2 + b_{122} x_1 x_2^2 + b_{111} x_1^3 + b_{222} x_2^3. \end{gathered} \tag{1.10}$$

Уравнения регрессии вида (1.7) и (1.8) называются линейными, а вида (1.9) и (1.10) – нелинейными. Линейные уравнения регрессии можно использовать на первом этапе исследований нелинейных корреляционных связей с тем, чтобы в дальнейшем ввести в них необходимые поправки. Расчёт коэффициентов уравнений регрессии выполняют чаще всего методом наименьших квадратов.

С позиций статистики полиномиальная модель удобна, так как позволяет увеличить степень точности аппроксимации за счет повышения порядка полинома. При этом аппроксимирующая функция остается линейной по параметрам, что облегчает все статистические операции (применение метода наименьших квадратов для оценки параметров, выбор оптимального расположения уровней факторов в области их изменения и т. д.).

При определении параметров уравнения регрессии все переменные и соотношения между ними иногда удобно выражать в стандартизованном масштабе, где за начало отсчета для каждой переменной принимается ее среднее значение, а за единицу масштаба – ее же среднее квадратическое отклонение. В стандартизованном масштабе упрощаются соотношения между переменными, что удобно при анализе многомерных связей. Значения переменных t_{x_i} в стандартизованном масштабе определяются по формуле

$$t_{x_i} = \frac{x_i - \overline{x}}{s_x}, \tag{1.11}$$

где x_i – значения переменных (зависимой или независимой) в натуральном масштабе.

Если в качестве модели процесса используется полином первой степени, то после перевода переменных в стандартизованный масштаб значительно упрощается формула для расчета коэффициента множественной корреляции:

$$R_{1.23\,\ldots\,m} = \sqrt{\beta_1^2 r_{12}^2 + \beta_2^2 r_{13}^2 + \ldots + \beta_m^2 r_{1m}^2}\,. \tag{1.12}$$

Здесь β_1, β_2, …, β_m – коэффициенты уравнения регрессии в стандартизованном масштабе.

Соответствие математической модели процесса экспериментальным данным называют адекватностью. Уравнение адекватно описывает результаты опытов, если квадратическое отклонение значений зависимой переменной y_{p_i}, рассчитанных по уравнению регрессии, от экспериментальных данных

y_i обусловлено только ошибкой воспроизведения, т. е. случайным характером этого параметра.

Применение КРА правомерно и эффективно при соблюдении ряда требований к экспериментальным данным. Содержание этих требований и методы их проверки рассмотрены ниже в п. 2 данного пособия.

1.2. Оценка достоверности результатов КРА

Поскольку результаты корреляционно-регрессионного анализа, полученные на базе ограниченного числа экспериментальных данных, являются случайными величинами, необходимо оценить их достоверность, определить доверительные интервалы, в которых находятся их истинные значения. Для этого последовательно производятся следующие операции: 1) оценка достоверности коэффициентов корреляции; 2) оценка значимости коэффициентов регрессии; 3) оценка адекватности уравнения регрессии.

1. При любом объеме выборки и нормальном многомерном распределении рассматриваемых факторов вычисляется статистика:

$$t_{\text{н}} = r\sqrt{(n-2)/(1-r^2)}, \tag{1.13}$$

имеющая распределение Стьюдента с $f = n-2$ степенями свободы. Для проверки нулевой гипотезы H_0: $\rho = 0$ (ρ – коэффициент корреляции генеральной совокупности) находят по соответствующим таблицам прил. 5 при фиксированном уровне значимости α и числе степеней свободы $f = n-2$, критическое значение $t_{\alpha/2,\, n-2}$, удовлетворяющее условию $P\left(|t| \geq t_{\alpha/2,\, n-2}\right) = \alpha$. Если наблюдаемое значение $t_{\text{н}} \geq t_{\alpha/2,\, n-2}$, то нулевую гипотезу об отсутствии линейной зависимости между переменными x и y следует отвергнуть. Этот метод часто применяют при малых объемах выборок.

Пример [4]. Пусть по выборке объема $n = 11$, извлеченной из двумерной нормальной совокупности, вычислен выборочный (эмпирический) коэффициент корреляции $r = 0,76$. Требуется проверить нулевую гипотезу H_0: $\rho = 0$ против альтернативной гипотезы H_α: $\rho \neq 0$ при уровне значимости $\alpha = 0,01$.

Решение. Вычисляем статистику

$$t_{\text{н}} = 0,76\sqrt{\frac{11-2}{1-\left(0,76\right)^2}} = 3,5.$$

По таблице распределения Стьюдента при $\alpha = 0,01$ и числе степеней свободы $f = n-2 = 9$ находим $t_{\alpha/2,\, n-2} = t_{0,005;\, 9} = 3,25.$ Так как 3,5 > 3,25, то r значимо отличается от нуля, т. е. переменные x и y являются коррелированными.

При числе наблюдений $n > 50$ надежность коэффициента корреляции можно оценить по его среднему квадратическому отклонению:

$$s_r = \left(1 - r^2\right) / \sqrt{n} \tag{1.14}$$

и нормированному отклонению:

$$t_r = \left|r\right| / s_r. \tag{1.15}$$

Достоверность коэффициента корреляции считается доказанной с вероятностью 0,997, если $t_r \geq 3$; с вероятностью 0,990 при $t_r \geq 2,58.$ Если n достаточно велико, а r близко к 0,5, границы доверительного интервала для коэффициента корреляции ρ той генеральной совокупности, из которой взята выборка, можно определить обычным способом:

$$r - t_{\text{кр}} s_r \le \rho \le r + t_{\text{кр}} s_r. \tag{1.16}$$

Значение $t_{\text{кр}}$ устанавливается по таблице функции Лапласа для выбранной вероятности. Если левая и правая части неравенства (1.16) имеют одинаковый знак, т. е. ρ не принимает нулевого значения, то r имеет достоверный знак и является значимым.

В случае выборок малых объемов доверительный интервал для r можно определить по номограмме.

Формулы (1.13–1.16) справедливы и при оценках достоверности коэффициента множественной корреляции и корреляционных отношений.

2. Проверку значимости коэффициентов регрессии можно производить двумя способами: сравнением абсолютного значения коэффициента с доверительным интервалом; с помощью t-критерия Стьюдента. В первом случае доверительный интервал для коэффициента b_i вычисляют по формуле

$$\Delta b_i = \pm t_{\text{т}} s_{b_i}, \tag{1.17}$$

где $t_{\text{т}}$ – табличное значение критерия Стьюдента при уровне значимости и числе степеней свободы, для которых определялось s_{b_i}; s_{b_i} – среднее квадратическое отклонение b_i.

Коэффициент значим, если его абсолютное значение больше доверительного интервала.

При проверке значимости коэффициентов вторым способом вычисляют

$$t_p = |b_i| / s_{b_i} \tag{1.18}$$

и сравнивают его с критическим значением этого критерия $t_{\text{кр}}$. Коэффициент значим, если $t_p > t_{\text{кр}}$ для принятого уровня значимости и числа степеней свободы, при которых определялось s_{b_i}.

Статистически незначимые коэффициенты могут быть исключены из уравнения регрессии. Причем, если уравнение получено с помощью методов планирования эксперимента, остальные коэффициенты пересчитывать не надо.

Методика определения s_{b_i} зависит от способа получения уравнения регрессии. В случае применения планирования эксперимента

$$s_{b_i}^2 = \frac{s_{(y)}^2}{nN}, \tag{1.19}$$

где $s_{(y)}^2$ – дисперсия воспроизводимости эксперимента;

n – число параллельных опытов в каждой точке матрицы при равномерном дублировании опытов (при отсутствии дублирования опытов n = 1); N – общее число опытов в матрице плана.

При равномерном дублировании опытов во всех строках матрицы плана число параллельных опытов одинаково. Для каждой строки этой матрицы вычисляют дисперсию s_j^2 результатов по данным n параллельных опытов:

$$s_j^2 = \frac{1}{n-1}\sum\nolimits_{u=1}^{n}(y_{j_u} - \bar{y}_j)^2, \tag{1.20}$$

где y_{j_u} – значение функции отклика в j-й строке для u-го опыта.

Если s_j^2 результатов опытов однородны, то дисперсия $s_{(y)}^2$ воспроизводимости эксперимента

$$s_{(y)}^2 = \frac{1}{N}\sum_{j=1}^{N} s_j^2, \tag{1.21}$$

где N – число опытов или число строк матрицы плана.

При отсутствии дублирования опытов для определения дисперсии воспроизводимости эксперимента выполняют n_0 параллельных опытов при средних уровнях всех независимых факторов (в нулевой точке плана). По результатам этих опытов вычисляют

$$s_{(y)}^2 = \frac{1}{n_0 - 1}\sum_{u=1}^{n_0} \left(y_u - \overline{y}\right)^2, \tag{1.22}$$

где y_u – значение функции отклика в u-м параллельном опыте.

При равномерном дублировании опытов число степеней свободы для расчета $s_{(y)}^2$ и, следовательно, s_{b_i} находится как $f = N\left(n-1\right)$ при отсутствии дублирования опытов $f = n_0 - 1$.

Если коэффициенты уравнения регрессии получены без планирования эксперимента (по результатам пассивного эксперимента), средние квадратические отклонения коэффициентов регрессии определяются следующим образом.

Для парной регрессии $\overline{y} = a + bx$:

$$s_b = \frac{s_y\sqrt{1 - r_{xy}^2}}{s_x\sqrt{n}}; \tag{1.23}$$

для регрессии трех переменных $\overline{y} = a + bx + cz$:

$$s_b = \frac{s_y \sqrt{1 - R^2_{y.xz}}}{s_x \sqrt{1 - r^2_{xz}} \sqrt{n}}; \tag{1.24}$$

$$s_c = \frac{s_y \sqrt{1 - R^2_{y.xz}}}{s_z \sqrt{1 - r^2_{xz}} \sqrt{n}}; \tag{1.25}$$

для регрессии многих переменных $\overline{y} = b_0 + b_1 x_1 + b_2 x_2 + {}$ $+ \ldots + b_k x_k$:

$$s_{b_1} = \frac{s_y \sqrt{1 - R^2_{y.x_1 x_2 \ldots x_k}}}{s_{x_1} \sqrt{1 - R^2_{x_1.x_2 x_3 \ldots x_k}} \cdot \sqrt{n}}. \tag{1.26}$$

Аналогичные формулы могут быть записаны для расчета s_{b_2}, s_{b_3}, …, s_{b_r}. Дальнейшая проверка значимости коэффициентов регрессии выполняется по формулам (1.17) и (1.18).

3. В зависимости от наличия сведений о дисперсии воспроизводимости эксперимента $s^2_{(y)}$ проверку адекватности уравнения регрессии можно производить по двум схемам. Первая из них применяется при отсутствии оценки дисперсии воспроизводимости, что характерно для пассивного эксперимента и состоит из следующих этапов:

а) вычисление дисперсии относительно среднего значения параметра оптимизации (остаточной дисперсии для уравнения нулевого порядка):

$$s^2_{y_0} = \frac{\sum_{i=1}^{N} \left(y_i - \overline{y} \right)^2}{N - 1}; \tag{1.27}$$

б) расчет дисперсии, характеризующей отклонение экспериментальных значений величин от найденных по уравнению регрессии. Если порядок уравнения заранее неизвестен, то в случае многофакторного пространства имеет смысл начинать с уравнения первого порядка

$$s_{y_1}^2 = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{f}, \tag{1.28}$$

где $\hat{y}_i$ – значение параметра оптимизации, вычисленное по уравнению регрессии для условий i-го опыта; $f = N - g$ – число степеней свободы; g – число коэффициентов регрессии; для линейного уравнения $g = k + 1$; для неполного квадратного уравнения, включающего члены типа $b_j x_j$ и $b_{ij} x_i x_j$, $g = k\left(k+1\right)/2 + 1$; для полного квадратного уравнения $g = k\left(k+3\right)/2 + 1$; k – число факторов;

в) вычисление опытного значения отношения дисперсий

$$F_0 = \frac{s_{y_0}^2}{s_{y_1}^2}, \tag{1.29}$$

которое затем сравнивают с критическим $F_{\text{кр}(f_1, f_2)}$. Числа степеней свободы f_1 и f_2 принимают равными соответственно знаменателям в формулах (1.27) и (1.28).

Если $F_0 \le F_{\text{кр}}$, пользоваться уравнением регрессии первого порядка не имеет смысла, так как в изученном интервале изменения уровней факторов оно описывает исследуемую систему (процесс) не лучше, чем уравнение нулевого порядка. Затем составляют уравнение второго порядка, рассчитывают

$s_{y_2}^2$ и $F_1 = s_{y_2}^2 / s_{y_1}^2$. Далее проверяют значимость этого отношения по F-критерию. Процедуру повторяют до тех пор, пока не будет выполнено условие $F_r \geq F_{\text{кр}(f_1,\, f_2)}$. Индекс r соответствует степени предпоследнего полинома.

Если известна дисперсия воспроизводимости $s_{(y)}^2$ эксперимента для оценки адекватности модели вначале рассчитывают дисперсию адекватности по формуле (1.28), а затем вычисляют опытное значение F-критерия (вторая схема):

$$F_p = s_{\text{ад}}^2 / s_{(y)}^2. \qquad (1.30)$$

Если $F_r < F_{\text{кр}(f_1,\, f_2)}$, модель считают адекватной. Значения $s_{(y)}^2$ в зависимости от характера дублирования опытов определяют по формулам (1.21) или (1.22), значения f_1 и f_2 равны соответственно знаменателям в формулах для расчета $s_{\text{ад}}^2$ и $s_{(y)}^2$. $F_{\text{кр}}(t_1,\, t_2)$ приведены в прил. 3.

2. ОЦЕНКА ПРАВОМЕРНОСТИ ПРИМЕНЕНИЯ КРА В ЗАДАННЫХ УСЛОВИЯХ

2.1. Основные положения предварительного анализа экспериментальных данных

Цель работы: изучение методики предварительного анализа экспериментальных данных при статистическом исследовании зависимостей и приобретение практических навыков.

Работа рассчитана на 5 академических часов.

Для моделирования различных случайных процессов, в том числе процессов механической обработки деталей машин, наиболее часто используется такой статистический метод как корреляционно-регрессионный анализ (КРА). При этом качество модели во многом зависит от качества экспериментальных данных, положенных в её основу. Поэтому предварительный анализ экспериментальных данных является обязательным этапом статистического исследования зависимостей [1].

При выборе КРА в качестве метода исследования необходимо выполнить анализ физической природы изучаемого процесса или объекта и убедиться в том, что между рассматриваемыми зависимыми и независимыми переменными имеются взаимосвязи. Только в этом случае для их анализа могут использоваться статистические методы.

Анализ статистических методов моделирования и распространенных методов оценки адекватности и достоверности полученных результатов показывает, что их эффективное применение возможно при соблюдении следующих требований к экспериментальным данным:

1. зависимые и независимые переменные являются случайными величинами с нормальным законом распределения*;

---

* независимые переменные могут быть и неслучайными величинами, в частности, в активных экспериментах.

2. дисперсия зависимой переменной y не зависит от абсолютных значений y – остается постоянной или однородной при различных наблюдениях y;

3. значения независимых переменных x_1, x_2 … x_m измеряются с пренебрежимо малыми ошибками по сравнению с ошибкой измерения y;

4. переменные x_1, x_2 … x_m линейно независимы;

5. процесс формирования y является стационарным и эргодическим;

6. экспериментальные данные получены из ряда независимых испытаний, наблюдений и образуют случайную выборку из данной генеральной совокупности;

7. результаты наблюдений не содержат резко выделяющихся значений, не принадлежащих к данной генеральной совокупности.

Только экспериментальные данные, отвечающие перечисленным требованиям, позволяют получить надёжные и достоверные результаты статистического исследования взаимосвязей между зависимыми (y) и независимыми (x) переменными. Рассмотрим методы проверки соблюдения указанных требований.

Соответствие распределения y и x нормальному закону распределения устанавливается либо по большим выборкам с помощью критериев Пирсона или Колмогорова [2], либо на основании анализа природы величин y или x.

Критерий Колмогорова λ даёт достаточно точные результаты даже при сравнительно небольших (около 30 шт.) объёмах выборок и прост для вычислений. Для расчёта значения λ необходимо предварительно определить эмпирические $F_n(x)$ и теоретические $F(x)$ функции предполагаемого закона распределения для каждого наблюденного значения случайной величины x (в качестве случайной величины могут рассматриваться как зависимые, так и независимые случайные величины). Методики определения $F_n(x)$ и $F(x)$ для нормального закона распределения рассмотрены в [3, 4]. Затем по максимальной разности этих функций определяется λ по следующей формуле:

$$\lambda = | F_n(x) - F(x) |_{\max} \sqrt{n} = D\sqrt{n} \tag{2.1}$$

здесь n – объём выборки, по которой производится оценка закона распределения.

По вычисленному значению λ по прил. 1 определяют $P(\lambda)$. Если $P(\lambda) > 0{,}15$, то гипотеза о соответствии эмпирического распределения нормальному распределению принимается.

Критерий χ^2 применим для любых сгруппированных совокупностей, но при достаточно большом их объёме ($n > 50$). Для вычисления χ^2 необходимо предварительно вычислить теоретические частоты для наблюденных значений эмпирического распределения, т. е. произвести сопоставление этого распределения с предполагаемым теоретическим (в нашем случае – с нормальным распределением). При этом необходимо, чтобы число эмпирических частот f_i в каждом интервале значений x было не менее 5. Если в каком-либо интервале $f_i < 5$, то такой интервал следует объединить с соседним.

Критерий χ^2 вычисляется по следующей формуле:

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(f_i - f_i'\right)^2}{f_i'}, \tag{2.2}$$

где m – число интервалов значений случайной величины x;

f_i, f_i' – эмпирическая и теоретическая частоты i-го интервала значений случайной величины x.

Методы расчёта f_i и f_i' показаны в [3, 4].

Далее необходимо определить число степеней свободы k по формуле:

$$k = m - p - 1, \tag{2.3}$$

где p – число параметров теоретического распределения.

Для нормального распределения $k = 2$ (математическое ожидание и дисперсия).

Для величины k найден закон распределения, по которому вычислены вероятности $P(\chi^2)$ для различных значений χ^2 и k. Значения $P(\chi^2)$ приведены в прил. 2. Если $P(\chi^2) > 0{,}05$, то гипотеза о соответствии рассматриваемых эмпирического и теоретического распределений принимается. Здесь принят уровень значимости $\alpha = 0{,}05$.

Необходимо отметить, что критерии согласия Пирсона и Колмогорова не позволяют однозначно установить теоретический закон распределения рассматриваемой случайной величины x. С помощью этих критериев нулевая гипотеза о виде функции распределения либо отклоняется, либо считается, что имеющаяся информация не даёт повода для отклонения выдвинутой гипотезы о виде функции распределения. Если объём выборок n невелик ($n \le 50$) или результаты измерений изменяются в узком диапазоне, то экспериментальные данные могут согласовываться с рядом законов распределения. На практике критерии согласия часто применяют для подбора закона распределения, который с наилучшей точностью описывает эмпирическое распределение из ряда рассматриваемых законов. Если критерии согласия Пирсона и Колмогорова дают различные результаты, то предпочтение следует отдать результату, полученному с помощью критерия Пирсона, который считается более строгим.

Выбор подходящего закона распределения должен базироваться прежде всего на понимании механизма изучаемого явления. Если механизм изучаемого явления неизвестен, то предварительный выбор закона распределения может быть выполнен с помощью следующих методов: по виду гистограммы частостей эмпирического распределения; с помощью графического представления эмпирической функции распределения на вероятностных бумагах; по величине эмпирического коэффициента вариации $v = s / \bar{x}$ (здесь s – среднее квадратическое отклонение случайной величины x; $\overline{x}$ – её среднее арифметическое значение); с помощью выборочных коэффициентов асимметрии и эксцесса [4, 5].

Для оценки ***однородности дисперсий y*** в условиях планирования эксперимента проводят параллельные опыты в различных точках матрицы плана, т. е. при различных значениях $x_1, x_2, \ldots x_m$. При пассивном эксперименте $s^2_{y_j}$ определяют по результатам отдельных выборок, взятых в примерно одинаковых условиях работы объекта. Если сравниваются **два** значения $s^2_{y_1}$ и $s^2_{y_2}$ при разных числах степеней их свободы f_j ($f = N - 1$, N – число параллельных опытов или объем выборки), то используется критерий Фишера, рассчитываемый как отношение большей дисперсии к меньшей:

$$F_p = \frac{s^2_{y_1}}{s^2_{y_2}}; \quad \left(s^2_{y_1} > s^2_{y_2}\right). \tag{2.4}$$

Если наблюдаемое значение F_p меньше критического $F_{\text{кр}}$ для соответствующих чисел степеней свободы и принятого уровня значимости, то опыты считаются воспроизводимыми, а дисперсии однородными [5]. В Прил. 3 приведены значения $F_{\text{кр}}$ для уровня значимости α = 0,05.

Однородность **ряда** дисперсий при одинаковом числе опытов для определения каждой из них оценивают с помощью критерия Кохрена – отношения максимальной дисперсии к сумме всех дисперсий ряда:

$$G_p = \frac{s^2_{y_{\max}}}{s^2_{y_1} + s^2_{y_2} + \ldots + s^2_{y_N}} = \frac{s^2_{y_{\max}}}{\sum_{j=1}^{N} s^2_{y_j}}, \tag{2.5}$$

где N – число параллельных опытов или различных выборок (обычно N = 2–5).

Дисперсии однородны, если расчетное значение G_p не превышает критического $G_{кр}$ [2]. В Прил. 4 приведены значения $G_{кр}$ для уровня значимости $\alpha = 0{,}05$.

Если дисперсии неоднородны, необходимо выполнить такое преобразование y, чтобы они стали однородными. Довольно часто помогает замена y на Iny.

Воспроизводимость опытов и однородность дисперсий достигается, когда выявлены и устранены источники нестабильности эксперимента, а также с помощью более точных методов и средств измерений.

Проверку ***достаточной точности измерения*** значений независимых переменных можно произвести, сопоставив ее с диапазоном изменения последних. Считается, что ошибки определения независимых переменных не должны превышать 5–7 % интервала их варьирования. Ошибки в определении значения зависимой переменной не влияют столь значительно на точность регрессионного анализа (они могут составлять до 30 % интервала варьирования).

Методы определения ошибок прямых и косвенных измерений описаны в [4, 6]. В данной работе этот вопрос не рассматривается.

Отсутствие коррелированности независимых переменных проверяется расчетом парных коэффициентов корреляции между ними [4, 5]. Способ оценки достоверности значений коэффициентов r парной корреляции зависит от объёма n выборки значений независимых переменных, по которым этот коэффициент был рассчитан.

Для выборки $n \le 30$ вычисляется критерий Стьюдента t по формуле:

$$t = \sqrt{\frac{n-2}{1-r^2}} \qquad (2.6)$$

имеющий распределение Стьюдента с $\upsilon = n - 2$ степенями свободы. Для проверки нулевой гипотезы H_0: $\rho = 0$ (здесь ρ – коэффициент корреляции генеральной совокупности) находят по таблицам распределения Стьюдента (Прил. 5) по принятому уровню значимости α и числу степеней свободы υ критическое значение $t_{\alpha/2;n-2}$, удовлетворяющее условию $P(|t| \geq t_{\alpha;n-2}) = \alpha$. Если $|t_{\text{набл}}| \geq t_{\alpha;n-2}$, то нулевую гипотезу об отсутствии линейной зависимости между рассмотренными независимыми переменными следует отвергнуть. Если же $|t_{\text{набл}}| < t_{\alpha;n-2}$, то нет оснований отвергать нулевую гипотезу о некоррелированности рассматриваемых независимых переменных.

При достаточно большом объёме выборки ($n \geq 50$) можно воспользоваться r-распределением, соответствующем $\rho = 0$ [7]. В Прил. 6 приведены квантили $r_{1-\alpha/2}$ этого распределения для некоторых уровней значимости α и числа степеней свободы $\upsilon = n - 2$. Если окажется, что найденный по выборке коэффициент корреляции r удовлетворяет неравенству

$$|r| > r_{1-\alpha/2}, \upsilon, \tag{2.7}$$

то его нужно признать значимым, т. е. нужно считать, что нулевая гипотеза неверна. А это значит, что $\rho \neq 0$ и между наблюдаемыми величинами есть корреляция. При несоблюдении условия (2.7) можно считать, что между наблюдаемыми величинами корреляция отсутствует.

Стационарным называют такой случайный процесс, основные характеристики которого $\left(M\left[x\right], \sigma^2\right)$ постоянны или однородны во времени. Поскольку при пассивном эксперименте свойства процесса определяются по одной представительной выборке, распространить полученные результаты на весь процесс в целом можно лишь при условии его стационарности.

Проверка стационарности процесса производится в следующем порядке.

По результатам измерений параметра строится случайная последовательность значений этого параметра, соответствующая порядку проведения измерений. Полученную реализацию разбивают на несколько равных отрезков (5–10), для каждого отрезка устанавливают дисперсию $s_{y_j}^2$ и с помощью критерия Кохрена по формуле (2.5) определяют, являются ли значения $s_{y_j}^2$ на каждом из отрезков оценками одной и той же генеральной дисперсии. Критическое значение критерия Кохрена можно для данных условий выбрать по прил. 4.

Затем на каждом из отрезков производится сравнение средних арифметических $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_N$, соответствующих выборочным дисперсиям $s_1^2, s_2^2, \ldots, s_N^2$, числам степеней свободы $f_1, f_2, \ldots, f_N$ (обычно $f_j = n_j - 1$, где n_j – объем соответствующей выборки, отрезка). Всем выборкам соответствует единая генеральная дисперсия σ^2. В качестве ее оценки можно взять средневзвешенную дисперсию s_y^2, которая рассчитывается по формуле (2.8)

$$\overline{s_y^2} = \frac{\sum_{j=1}^{N} f_j s_{y_j}^2}{f}, \quad f = \sum_{j}^{N} f_j. \tag{2.8}$$

Если справедлива нулевая гипотеза о равенстве всех генеральных средних, то в качестве оценки единого генерального среднего можно взять общее среднее всех элементов, отрезков, как бы объединенных в одну выборку.

Обозначим это среднее через $\overline{\overline{y}}$. Для дисперсии σ^2 можно теперь дать другую оценку:

$$\overline{s^2} = \frac{1}{N-1}\sum_{j=1}^{N} n_j \left(\overline{y_j} - \overline{\overline{y}}\right)^2, \qquad (2.9)$$

которой соответствует $N - 1$ степеней свободы. Чтобы нулевая гипотеза была справедлива, должно соблюдаться условие

$$\overline{s_y^2} / \overline{s^2} \le F_{1-p(N-1, f)}, \qquad (2.10)$$

где значение критерия Фишера $F_{1-p(N-1, f)}$ берется по Прил. 3.

Стационарная случайная функция обладает **эргодическим свойством**, если ее характеристики $M[x]$, σ^2 могут быть рассчитаны как соответствующие средние по времени для одной реализации большой продолжительности. Иными словами, эргодичность определяет способность процесса к воспроизведению своих характеристик в различных реализациях. Следовательно, условием эргодичности процесса является соблюдение описанных выше условий стационарности процесса, но рассчитанных не для различных отрезков одной реализации процесса, а для различных реализаций процесса (3–5 реализаций объёмом 50–100 шт.). Эргодичными могут быть только стационарные процессы. Если стационарный процесс обладает эргодическим свойством, его основные характеристики могут быть установлены по одной реализации достаточно большой продолжительности (объем выборки n = 50–100 шт.) Если анализ выборки не подтвердил стационарность процесса, необходимо либо учесть фактор времени в модели процесса как независимую переменную, либо найти объем выборки (период времени для ее отбора), для которой процесс будет стационарным.

Случайность и независимость результатов измерений в выборке свидетельствует об отсутствии монотонного или циклического смещения результатов измерений, вызванного влиянием неучтённого фактора, систематической погрешностью процесса. Подобный случай может иметь место при анализе размеров деталей, обрабатываемых на настроенном станке, когда вследствие изнашивания инструмента или нагрева станка, центр группирования размеров постепенно смещается при сохранении разброса размеров (значения дисперсии).

Проверку ***гипотезы о случайности выборки***, необходимую при пассивном эксперименте, можно произвести методом последовательных разностей. Этот метод заключается в следующем: по значениям x_i выборки, расположенным в последовательности их наблюдения x_1, x_2, $x_3 \ldots x_n$ образуется $n - 1$ разностей между соседними членами:

$$a_1 = x_2 - x_1,\ a_2 = x_3 - x_2,\ \ldots,\ a_{n-1} = x_n - x_{n-1}.$$

Затем определяются

$$C^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} a_i^2, \tag{2.11}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2. \tag{2.12}$$

Для оценки случайности выборки используется критерий $\tau = C^2 / s^2$. Если $\tau > \tau_q$ где τ_q – критическое значение критерия при уровне значимости q, то гипотеза «случайности» верна [2, 7]. Значения τ_q приведены в прил. 7.

Одним из важных условий правильного применения статистических оценок является **отсутствие грубых ошибок при наблюдениях**. Поэтому все грубые ошибки должны быть выявлены и исключены из рассмотрения в самом начале обработки результатов наблюдений. При решении данной задачи предполагается, что результаты наблюдений подчиняются нормальному закону распределения. Грубые погрешности возникают при случайном резком изменении условий измерения.

Для анализа резко выделяющихся значений случайных величин в выборке используют графический и аналитические методы [1]. Наиболее распространенным критерием для исключения одного экстремального значения является критерий К. Пирсона–Н. В. Смирнова–Ф. Граббса

$$\upsilon = \frac{|\bar{x} - x_i|}{s}, \tag{2.13}$$

где x_i – крайний (max или min) элемент в выборке;

$\bar{x}$, s – среднее арифметическое значение и среднее квадратическое отклонение в полной выборке (до исключения крайнего значения).

Критические значения $\upsilon_{1-\alpha}$ для данного уровня значимости α зависят только от объема выборки n. При $\upsilon > \upsilon_{1-\alpha}$ крайнее значение x отбрасывается как ошибочное. Процесс исключения подозрительных значений рекомендуется начинать с максимального значения и проводить последовательно для крайних значений вариационного ряда. После исключения ошибочного результата значения $\bar{x}$ и s необходимо пересчитать. Критерий υ справедлив для нормально распределенных случайных величин и чувствителен к отклонениям от этого закона. Для уменьшения вероятности ошибочного решения необходимо предварительно убедиться в том, что результаты получены при соблюдении требований к точности измерений в постоянных условиях реализации процесса.

3. НЕКОТОРЫЕ ВОПРОСЫ ВЫПОЛНЕНИЯ ОДНО- И МНОГОФАКТОРНЫХ ЭКСПЕРИМЕНТОВ

3.1 Методы подбора эмпирических формул по результатам однофакторного эксперимента

В результате однофакторного эксперимента получается статистический ряд измерений двух величин. В этом ряду каждому значению функции $y_1, y_2, \ldots, y_n$ соответствует определённое значение аргумента $x_1, x_2, \ldots, x_n$, где n – число измерений или объём выборки. На основе экспериментальных данных можно подобрать алгебраическое выражение функции

$$y = f(x) , \tag{3.1}$$

которое называют эмпирической формулой. Такие формулы справедливы лишь для того интервала изменения аргумента x, который имел место при проведении эксперимента. Эмпирические формулы являются приближёнными выражениями аналитических зависимостей.

Подбор эмпирических формул состоит из двух этапов.

I этап. Получение графика, показывающего изменение y при изменении x, т.е. отражающего зависимость (3.1). Выбор по этому графику примерного вида формулы (3.1).

II этап. Вычисление параметров принятой формулы, оценка её достоверности.

Для построения графиков, как правило, пользуются прямоугольной системой координат. Графики удобно строить на миллиметровой бумаге или на ЭВМ в среде *Exel*. Исходя из пределов, в которых при эксперименте изменялись значения y_i и x_i, i – номер опыта, выбирают определённые масштабы по осям так, чтобы были полностью использованы принятые длины осей. Эти масштабы выбираются произвольно, независимо друг от друга, но так, чтобы график не был слишком мелким или растянутым по одной из осей. Если интервал, в котором заключены значения измеряемых величин, лежит далеко от нуля, то целесообразно начинать деление соответствующей оси не от нуля, а от некоторого значения, лишь немногим меньшего, чем наименьшее значение, которое надо отложить на этой оси. На графике наносят точки с координатами y_i, x_i и соединяют их прямой или кривой таким образом, чтобы над и под прямой (кривой) находилось примерно одинаковое число точек. В редких случаях прямая (плавная кривая) проходят через все точки.

Иногда на одну или обе оси наносят неравномерную шкалу. Такой способ применяют в основном для приведения графика к линейному виду. Например на рис. 3.1,а изображена зависимость $y = cx^2$ с равномерными шкалами, а на графике 3.1,б – та же зависимость, когда по оси абсцисс отложены значения x^2.

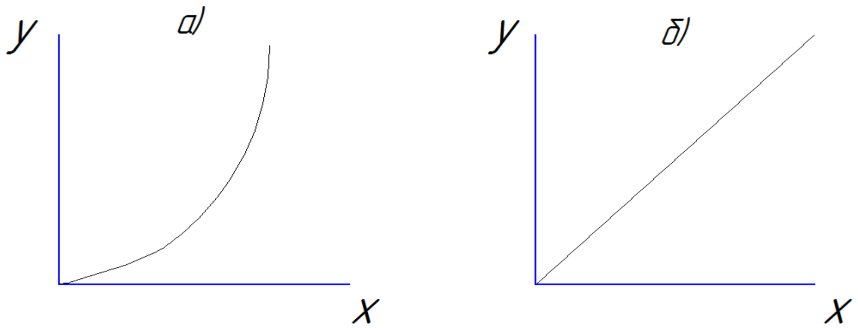


Рис. 3.1 – Графики с равномерными (а) и неравномерными (б) шкалами по осям координат

Такие масштабы называются функциональными. Из функциональных масштабов чаще всего используются логарифмический и полулогарифмический. В полулогарифмическом масштабе используются линейный масштаб по оси абсцисс и логарифмический масштаб по оси ординат. В логарифмическом масштабе используются логарифмические шкалы по обеим осям.

Для анализа вида зависимости (3.1) используют графики с равномерными шкалами по осям координат. Описанный выше метод построения графика применим для малого объёма выборки (обычно для $n \leq 10$). Для больших объёмов выборки целесообразно построить эмпирическую линию регрессии. Её пример показан на рис. 3.2.

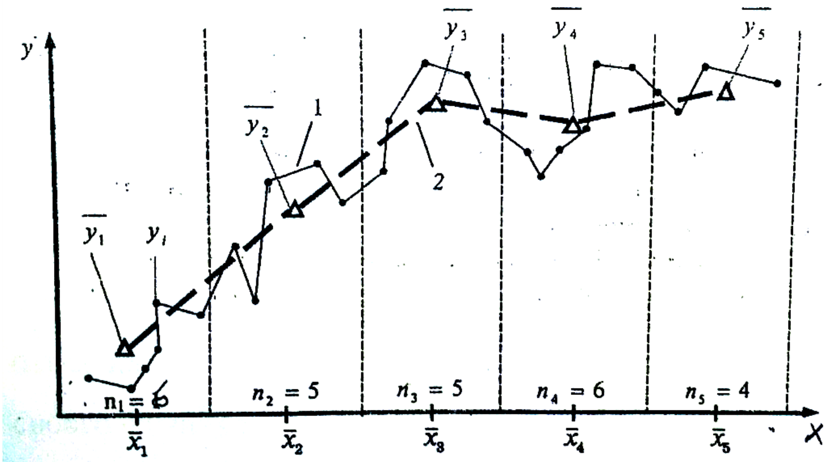


Рис. 3.2 – Пример эмпирической линии регрессии:

y_i – расчётные точки, полученные после первичной группировки экспериментальных данных на корреляционном поле; $\Delta = \overline{y_i}$ – средние значения ординат точек y_i, попавших в j-ый интервал; 1 – экспериментальная линия регрессии, полученная после первичной обработки корреляционного поля; 2 – экспериментальная линия регрессии, полученная после вторичной группировки результатов эксперимента.

Эмпирическую линию регрессии строят следующим образом:

1. Наносят на координатную сетку y-x все экспериментальные точки с координатами y_i, x_i и соединяют их ломаной линией (ломаная *1* на рис. 3.2). При $n > 30$ построение ломаной линии *1* не требуется.

2. Разбивают разброс значений x_i на 5-7 интервалов и для точек, попавших в эти интервалы определяют x_j и $\bar{y}_j$, где j – номер интервала.

3. Наносят на координатное поле точки с координатами $x_j, \bar{y}_j$ и соединяют их ломаной линией. На рис. 3.2 – это линия *2*. Она и будет эмпирической линией регрессии, с помощью которой можно оценить вид алгебраического выражения зависимости (3.1).

Если зависимость $y = f(x)$ линейна и может быть описана прямой линией

$$y_i = b_0 + b_1 x_i \ , \tag{3.2}$$

то расчёт неизвестных параметров b_0 и b_i может быть выполнен методом наименьших квадратов или с помощью корреляционно- регрессивного анализа .

При использовании МНК параметры b_0 и b_i модели (3.2) определяются по следующим уравнениям:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \ , \tag{3.3}$$

$$b_0 = \bar{y} - b_1 x_i \ , \tag{3.4}$$

где y_i, x_i – результаты эксперимента, n – число опытов, $\bar{y}$, x – средние арифметические значения аргумента x_i и функции y_i.

При использовании КРА:

$$b_1 = r\frac{S_y}{S_x}, \tag{3.5}$$

где r – коэффициент парной корреляции между значениями x и y [1] (см.ф. 3.1); S_x, S_y – средние квадратические отклонения значений x_i и y_i в выборке.

Значение b_0 в этом случае рассчитывается также по формуле 3.4.

Для выбора нелинейных эмпирических формул применяют два основных метода:

1. Линеаризация, т.е. приведение нелинейных функций к линейному виду (см. ф. 3.2) при помощи специальных преобразований.
2. Аппроксимация исследуемых зависимостей полиномами (параболами различной степени).

Рассмотрим первый метод. Он может быть реализован двумя путями.

1. Имея линейное преобразование нелинейной функции неизвестные члены уравнения типа (3.2) могут быть найдены по формулам (3.3) и (3.4).
2. Неизвестные члены линеаризованной функции могут быть определены графическим методом.

Рассмотрим пример применения первого пути для случая определения параметров степенной зависимости:

$$y = b_0 x^{b_1} \tag{3.6}$$

Прологарифмировав данное выражение, получаем:

$$lgy = lgb_0 + b_1 lgx \tag{3.7}$$

Заменяя $X = lgx$, $Y = lgy$ и $B_0 = lgb_0$, получим линейную модель:

$$Y = B_0 + b_1 X \tag{3.8}$$

Параметры B_0 и b_1 зависимости (1.8) находим с помощью формул (3.3) и (3.4), куда в качестве y_i подставим значения $Y_i = lgy_i$, в качестве x_i – значения $X_i = lgx_i$. Оценку параметра b_0 получаем потенцированием оценки B_0.

Примером применения второго пути является следующий случай.

Необходимо подобрать эмпирическую формулу для следующих результатов измерений:

Таблица 3.1 – Результаты измерений

Значения аргумента x_i	*1*	*1,5*	*2,0*	*2,5*	*3,0*	*3,5*	*4,0*	*4,5*
Значения функции y_i	*15,2*	*20,6*	*27,4*	*36,7*	*49,2*	*66,0*	*87,4*	*117,5*

На основе этих данных строится график (рис. 1.3, а), соответствующий выражению:

$$y = ae^{bx} \tag{3.9}$$

После логарифмирования выражения (1.9) $lgy = lga + bx\,lge$. Если обозначить $lg\,y = Y$, то $Y = lga + bx\,lge$, т.е. в полулогарифмических координатах выражение для Y представляет собой прямую линию (рис. 3.3,б). Подстановка в полученное уравнение координат крайних точек экспериментальных данных даёт:

$lg\,15,2 = lga + b\,lge$ и $lg\,117,5 = lga + 4,5b\,lge$.

Следовательно, $lga + b\,lge = 1,183$;

$$lga + 4{,}5b\ lge = 2{,}070\ ,$$

откуда $b = 0{,}887 / (3{,}5\ lge) = 0{,}579$; $lga = 1{,}183 - 0{,}254 = 0{,}929$; $a = 1{,}85$

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

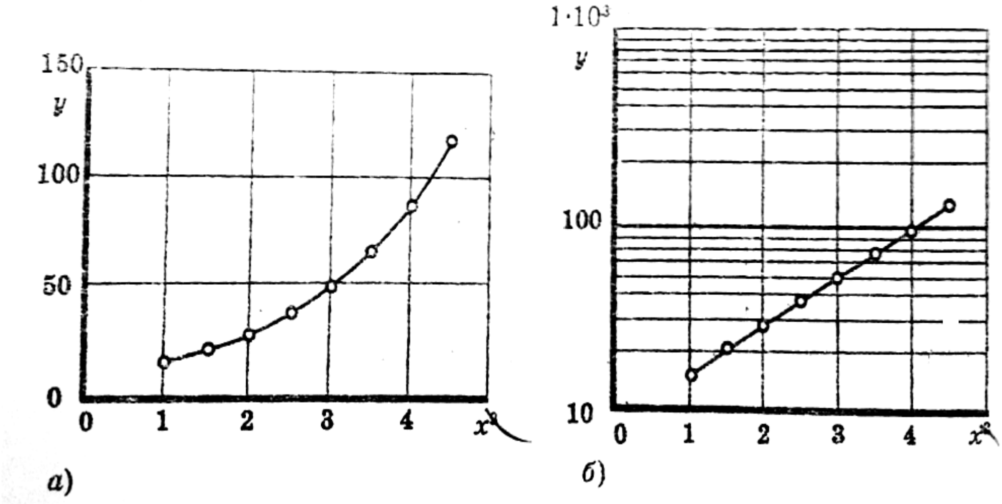


Рис. 3.3 – Подбор эмпирических характеристик графическим методом:

а – эмпирическая; *б* – спрямлённая

Графический метод позволяет визуально оценить правильность выбора вида формулы для эмпирической кривой. Если в неравномерных шкалах эмпирическая кривая превращается в прямую (рис. 3.3), это подтверждает правильность указанного выбора.

Графический метод может быть использован и непосредственно для определения параметров линеаризованной зависимости вида:

$$Y = A + BX \qquad (3.10)$$

где $Y = f_1(x,y)$ (3.11)

$X = f_2(x,y)$ (3.12)

После преобразования нелинейной исходной зависимости в линейную вида (3.10) по данным X_i, Y_i строят прямую (рис. 3.4) на координатной сетке, у которой началом является точка $Y=0$ и $X=0$. По этому графику легко определить параметры A (ордината точки пересечения прямой с осью Y) и B (тангенс угла наклона прямой с осью X): $B = tg\alpha = (Y_i - A)X_i$.

По значениям *A* и *B* с помощью зависимостей (3.11) и (3.12) можно перейти к параметрам нелинейной формулы, принятой для описания эмпирической зависимости *y=f(x)*.

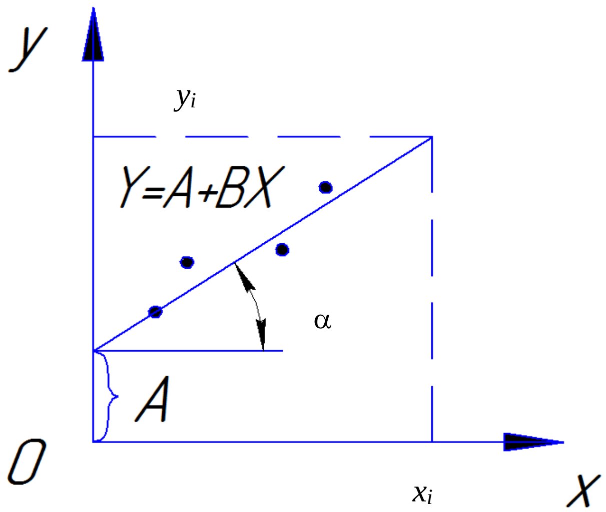
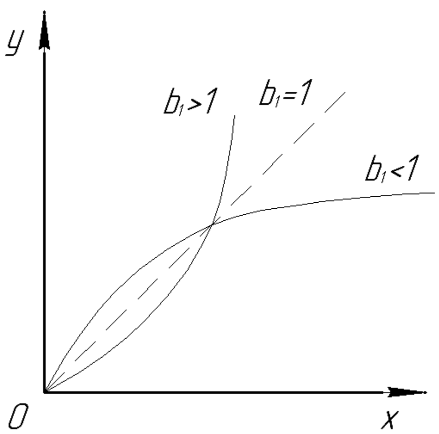
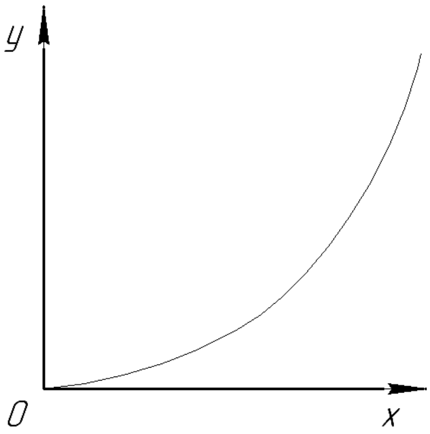
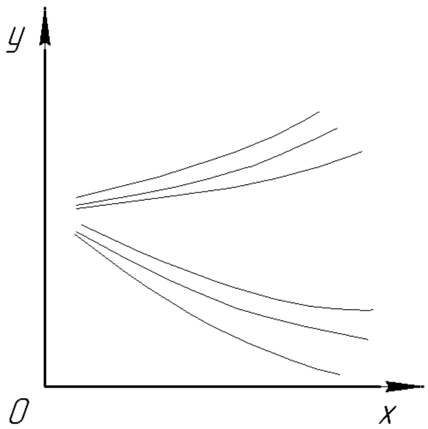
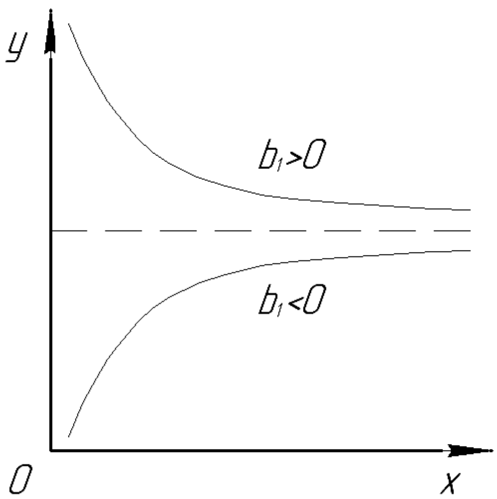


Рис. 3.4 – Графическое определение параметров *A* и *B* зависимости (1.10)

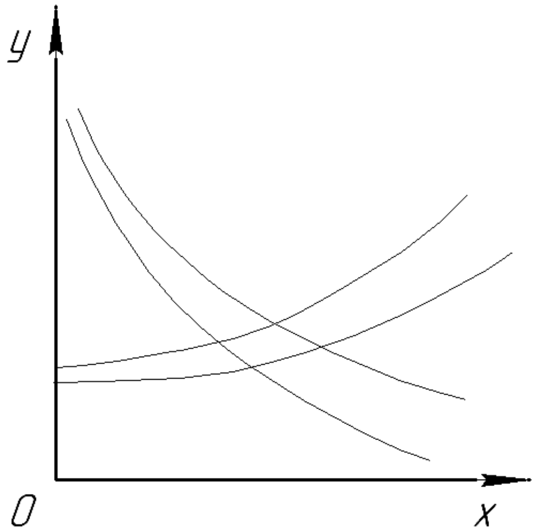
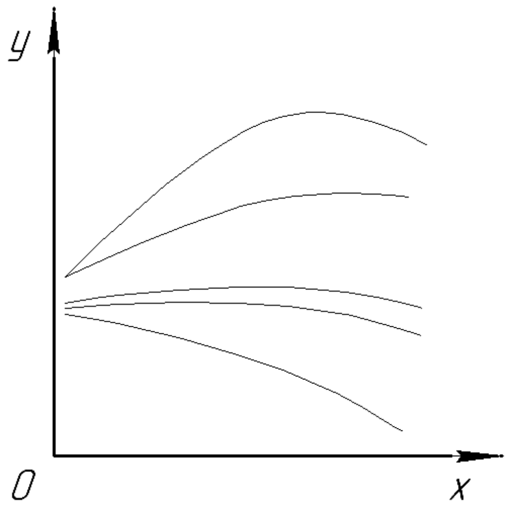
Таблица 3.2 – Виды нелинейных однофакторных моделей и методы их линеаризации

Название и формула модели	Вид графиков, характерных для данной модели	Метод линеаризации модели
1	2	3
Степенная $y = b_0 x^{b_1}$		$\lg y = \lg b_0 + b_1 \lg x$ $X = \lg x \quad Y = \lg y$ $B_0 = \lg b_0$ Линейная модель $Y = B_0 + b_1 X$ Прямая может быть построена на логарифмической сетке

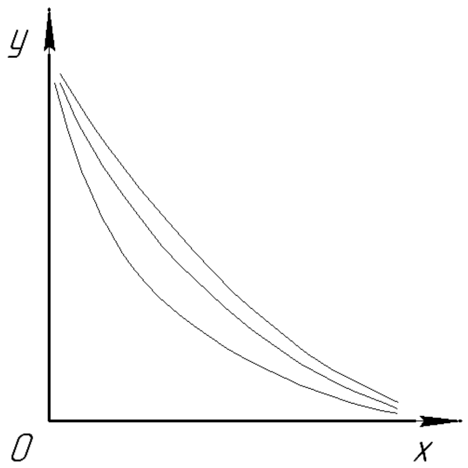
Продолжение таблицы 3.2

1	2	3
Показательная $y = b_0 b_1^x$		$\lg y = \lg b_0 + x \lg b_1$ $Y = \lg y \quad B_0 = \lg b_0$ $B_1 = \lg b_1$ Линейная модель $Y = B_0 + B_1 x$ Прямая может быть построена на полулогарифмической сетке
Экспонен-циальная $y = b_0 e^{b_1 x}$		$\lg y = \lg b_0 + b_1 x \lg e$ $Y = \lg y \quad B_0 = \lg b_0$ $B_1 = b_1 \lg e$ Линейная модель $Y = B_0 + B_1 x$ Прямая может быть построена на полулогарифмической сетке
Гиперболи-ческая $y = b_0 + \frac{b_1}{x}$		$z = 1/x$ Линейная модель $y = b_0 + b_1 z$ Прямая может быть получена на сетке прямоугольный координат

Продолжение таблицы 3.2

1	2	3
$y = a + b_0 x^{b_1}$		По экспериментальной кривой принимаем 3 произвольные точки x_1, y_1 ; x_2, y_2 ; $x_3 = \sqrt{x_1 x_2}, y_3$ и вычисляем a $a = \dfrac{y_1 y_2 - y_3^2}{y_1 + y_2 + 2y_3}$ (4.13) $\lg(y-a) = \lg b_0 + b_i \lg x$ $X = \lg x$, $Y = \lg(y-a)$ $B_0 = \lg b_0$ Линейная модель $Y = B_0 + b_1 X$ Прямая может быть построена на логарифмической сетке
$y = a + b_0 e^{b_1 x}$		Значение a определяем по методу описанному выше для модели 5, по ф. (4.13) $\lg(y-a) = \lg b_0 + x b_1 \lg e$ $Y = \lg(y-a)$ $B_0 = \lg b_0$ $B_1 = b_1 \lg e$ Линейная модель $Y = B_0 + B_1 x$ Прямая может быть построена на полулогарифмической сетке

Окончание таблицы 3.2

1	2	3
$y = \dfrac{1}{b_0 + b_1 x}$		$y = 1/z$ Линейная модель $z = b_0 + b_1 x$ Прямая может быть получена на сетке прямоугольных координат

В таблице 3.2 приведены различные виды нелинейных однофакторных моделей и методы их линеаризации. Для линеаризации функций, у которых имеются 3 независимых параметра (модели 5, 6 в табл. 3.2), наряду с логарифмированием исходной функции используют и графический метод. Для целей прогнозирования можно использовать либо полученную линеаризованную функцию, либо вернуться к первоначальному виду модели, что обычно менее удобно.

Рассмотрим второй метод подбора нелинейных эмпирических функций регрессии, т.е. аппроксимацию используемых зависимостей полиномами (параболами) вида: $y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n$. Для обоснования выбора порядка (максимальной степени) параболы необходимо исходить из следующего:

– наибольшее число экстремальных точек, которые может иметь парабола порядка n, равно $(n - 1)$, т.е. парабола второго порядка может иметь не больше одного экстремума, парабола третьего порядка – не более двух экстремальных точек и т.д.;

– согласно теореме Веерштрасса, любую непрерывную функцию (в нашем случае неизвестную истинную криволинейную функцию регрессии) можно приблизить на конечном интервале сколь угодно точно параболой порядка n;

– в большинстве случаев при обработке экспериментальных данных оказывается, что аппроксимация эмпирических зависимостей параболами выше четвёртого порядка приводит к очень незначительному увеличению точности. Поэтому считается практически нецелесообразным применять параболы порядка выше четвёртого.

Для приближённого определения порядка параболы рекомендуется по характеру расположения экспериментальных точек на корреляционном поле и по числу явно выраженных эмпирических экстремумов выбрать соответствующий порядок параболы. Например, к экспериментальным данным, нанесённым на рис. 3.5,а, целесообразно подбирать параболу второго порядка, на рис. 3.5,б – параболу третьего порядка.

Допустим, что между y и x существует криволинейная связь, аппроксимируемая параболой второго порядка:

$$y = b_0 + b_1 x + b_2 x^2 \qquad (3.14)$$

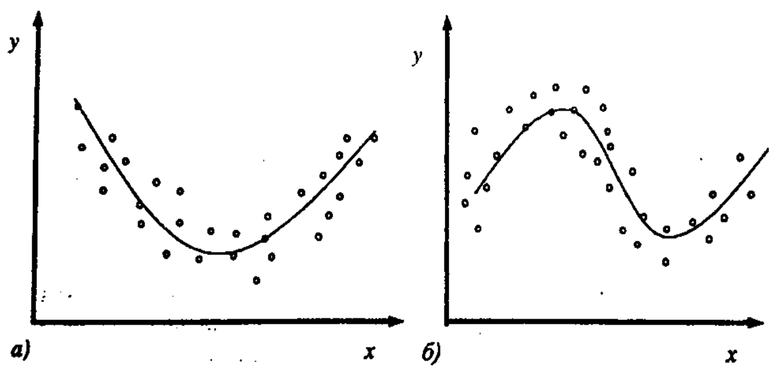


Рис. 3.5 – Характерные кривые для параболы второго порядка (а) и параболы третьего порядка (б)

Для оценки коэффициентов регрессии b_i необходимо решить следующую систему уравнений, полученную с помощью МНК:

$$\begin{cases} Nb_0 + b_1\sum_i x_i + b_2\sum_i x_i^2 = \sum_i y_i \\ b_0\sum_i x_i + b_1\sum_i x_i^2 + b_2\sum_i x_i^3 = \sum_i x_i y_i \\ b_0\sum_i x_i^2 + b_1\sum_i x_i^3 + b_2\sum_i x_i^4 = \sum_i x_i^2 y_i \end{cases} \qquad (3.15)$$

Для получения регрессионной математической модели по результатам однофакторного эксперимента необходимо предварительно установить общий вид искомой функции. Необходимо отметить, что не существует строгих математических методов, которые позволили бы априори, т.е. до проведения корреляционно-регрессионного анализа, указать общий вид функции. Обычно на практике вид функции регрессии выбирают по характеру расположения точек на корреляционном поле. При этом такой выбор обязательно необходимо дополнить логически профессиональным анализом физической сущности исследуемого процесса или явления с учётом опыта предыдущих исследований. Необходимо отметить, что выбор общего вида экспериментальной функции не является однозначным. Это значит, что одну и ту же экспериментальную зависимость можно аппроксимировать либо полиномом, либо показательной, степенной или логарифмической функцией, т.е. функциями, допускающими линеаризацию.

3.2 Оценка точности прогнозирования с помощью однофакторной регрессионной модели

Как известно, регрессионная модель $y = f(x)$ справедлива в пределах изменения фактора x в опытах, по результатам которых эта модель получена. Если использовать эту модель для значений x^*, выходящих за эти пределы, то необходимо оценить доверительный интервал для прогнозных значений y^*. Расчёт этого интервала может быть выполнен следующим образом.

Вначале рассчитывается значение несмещённой оценки дисперсии y, очищенной от влияния x, которая характеризует степень рассеивания экспериментальных точек y_i вокруг теоретической кривой y_i^T:

$$D(y)_x = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - y_i^T\right)^2. \qquad (3.16)$$

Здесь y_i^T – значения y_i, рассчитанные по уравнению регрессии для соответствующих x_i; i – номер опыта; N – число опытов.

Далее определяется значение половины величины доверительного интервала разброса среднего значения:

$$\Delta\bar{y} = t_{\frac{\alpha}{2},N}\sqrt{\frac{D(y)_x}{N}}, \qquad (3.17)$$

где $t_{\frac{\alpha}{2},N}$ – значение критерия Стьюдента.

Следующим шагом является определение величины периода упреждения (прогноза) Π:

$$\Pi = \frac{x \cdot N}{x_N} - N, \qquad (3.18)$$

где x_N – максимальное значение признака по оси абсцисс.

Например: произведена оценка какого-либо диагностического параметра y автомобиля через каждые 10 тыс. км пробега x. Всего проведено на момент анализа 12 замеров. Таким образом, $N = 12$ (число точек по оси абсцисс). Максимальное значение аргумента 120 тыс. км. Требуется спрогнозировать значение $y*$ на пробег $x^* = 155$ тыс. км. Величина Π определяется следующим образом:

$$\Pi = \frac{155 \cdot 12}{120} - 12 = 3{,}5$$

Значение половины величины доверительного интервала для прогнозируемого значения вычисляется по формуле:

$$\Delta y^* = \Delta\bar{y}\sqrt{\frac{N+1}{N} + \frac{3\left(N + 2\Pi - 1\right)^2}{N\left(N^2 - 1\right)}} \qquad (3.19)$$

Доверительный интервал для прогнозируемого значения y^* определяется соотношением:

$$y^* - \Delta y^* < M\left(y^*\right) < y^* + \Delta y^* \qquad (3.20)$$

На основании результатов расчёта рекомендуется построить график теоретической кривой y_i^T по полученной математической модели. На графике наносятся прогнозируемое значение y^* и его доверительный интервал.

Пример такого графика показан на рис. 3.6. Здесь отражена зависимость параметра потока отказов $\omega(L_i)$ карданной передачи автомобилей МАЗ-7313 от пробега L_i с начала эксплуатации.

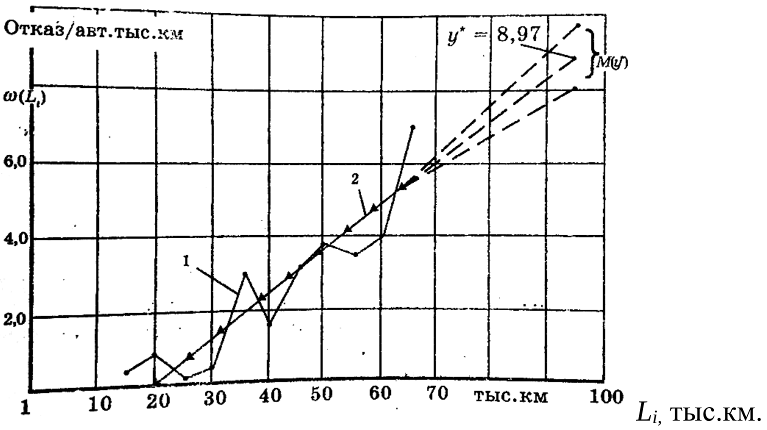


Рис. 3.6 – График зависимости параметра потока отказов от пробега:
1 – экспериментальная кривая регрессии; 2 – теоретическая кривая линейной регрессии

Испытания были проведены до $L=65$ тыс. км. При расчёте $\omega(L_i)$ для $L_i=94$ тыс.км. получили значение $\omega(L_i)=8,97$ отказ/авт.тыс.км. $\Delta y^*=0,745$. Эти значения нанесены на график $\omega=f(L_i)$ и соединены пунктирными линиями с точкой на теоретической кривой для $L=65$ тыс. км.

3.3 Выбор факторов, влияющих на изучаемый процесс или объект, сокращение их числа

Перед проведением исследований наряду с задачей выбора параметров оптимизации необходимо решить задачу выбора факторов, существенно влияющих на изучаемый процесс или объект. Вначале необходимо составить список всех факторов, которые могут влиять на изучаемые параметры оптимизации. Это может быть сделано на основании анализа природы объекта, имеющейся априорной информации, опроса специалистов и др. методами. Затем следует выделить доминирующие факторы.

Для правильного отображения объекта необходимо, чтобы его математическая модель включала все факторы, существенно влияющие на выходную величину. Отсутствие в модели хотя бы одного из существенных факторов может повлечь за собой ошибочную интерпретацию явлений, протекающих в объекте. В свою очередь это может явиться причиной многих ошибок в решениях, принимаемых с помощью модели. Например, расчетное положение оптимума окажется значительно удаленным от действительного и т.п. Кроме того, модели, в которые входят не все доминирующие факторы, как правило, являются неадекватными.

При исследовании сложных процессов, на которые воздействует большое число факторов, обычно заранее не известно, какие из этих факторов следует считать доминирующими. В сложных промышленных процессах, особенно в так называемые больших системах, число различных факторов может быть очень большим – от десятков до нескольких сотен. Степень же влияния этих факторов весьма различна; как правило, лишь небольшое число факторов оказывает существенное влияние на выходную величину, а остальные воздействуют на нее незначительно. Задача заключается в том, чтобы с помощью соответствующих методов выделить и идентифицировать существенные, доминирующие факторы на «шумовом фоне» всех остальных факторов. Если попытаться учесть в модели технологического процесса все действующие факторы, это приведет к необходимости выполнить очень большое число опытов (2^{20} – 2^{30}, для 20 – 30 факторов при варьировании факторов на двух уровнях). Это практически невозможно, да и не нужно.

Сокращение списка факторов возможно с помощью следующих методов:

1. Выбор факторов, отвечающих требованиям, предъявляемых к их идентификации.
2. Априорное ранжирование .
3. Дисперсионный анализ.
4. Отсеивающие эксперименты с помощью сверхнасыщенных и насыщенных планов (методы случайного баланса, планы Плаккета-Бермана, симплекс-планы первого прядка и др.).

При планировании эксперимента факторы должны удовлетворять следующим требованиям:

1. Управляемость – способность фактора сохранять заданное значение (уровень) в течение эксперимента.
2. Операциональность – наличие списка операций по поддержанию фактора на определенном уровне и его точному измерению.
3. Возможно высокая точность измерения.
4. Однозначность (независимость) – возможность изменения фактора независимо от других.
5. Совместимость – все комбинации факторов должны быть осуществимы и безопасны.

После исключения факторов, не удовлетворяющих указанным требованиям, следует провести ранжирование оставшихся факторов. Для этого могут быть использованы перечисленные выше методы.

Рассмотрим их основные принципы.

Методика априорного ранжирования была вкратце описана выше

Идея *дисперсионного анализа* заключается в оценке существенности доли дисперсии параметра оптимизации, вызванной влиянием какого-либо фактора А.

Если эта доля $S_{\text{А}}^2$ по сравнению с дисперсией параметра оптимизации S_y^2 является существенной (проверка осуществляется с помощью критерия Фишера) значит фактор А существенно влияет на параметр оптимизации.

Цель *отсеивающих экспериментов* – качественная принципиальная оценка степени влияния каждого из рассматриваемых факторов на параметр оптимизации при минимально возможном числе факторов. Задачи получения математической модели и оценки её адекватности при этом не ставят.

Минимальное число опытов имеет место при использовании насыщенных и сверх насыщенных планов эксперимента. Под планом эксперимента понимается схема проведения опытов, в которой указываются число опытов и уровни факторов в каждом опыте. Обычно факторы изменяются на 2 уровнях. Верхний уровень обозначают (+), нижний – (–). Степень насыщенности плана определяется числом степеней свободы φ.

$$\varphi = N - \text{k}.$$

где N – число опытов; k – число искомых параметров (коэффициентов уравнения регрессии).

Если $\varphi>0$ и $N<\text{k}$, т.е. число проведенных экспериментов превышает число оцениваемых параметров, планы называют ненасыщенными. В этом случае возможна проверка адекватности модели. При $\varphi=0$ и $N=\text{k}$ планы называют насыщенными. Для проверки адекватности здесь уже не остается степеней свободы. При φ<0, т.е. при N<k, планы называют сверх насыщенными.

Распространенным методом сокращения числа опытов является использование *дробных факторных планов* . Зависимость между параметром оптимизации и факторами здесь принимается в виде линейной модели:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i \ , \qquad (3.25)$$

где i – номер фактора, n – их число.

Насыщенные регулярные дробные факторные планы типа 2^{n-p} (здесь p – число линейных эффектов, приравненных к эффектам взаимодействий, содержат число точек эксперимента N (на уровнях +1 и -1), равное числу искомых параметров k. Для линейной модели типа (3.25)

k=n+1=N

Расчет параметров a_0, a_i зависимости (3.25) по результатам дробного факторного эксперимента производится также как и для полного факторного эксперимента . Значимость этих параметров (а, следовательно, и факторов) проверяется с помощью t-критерия. При этом проверяется справедливость неравенства:

$$|a_i| \geq t_{\text{кр}} \cdot S_i \ , \qquad (3.26)$$

где $t_{кр}$ – критическое значение распределения Стьюдента для заданного уровня значимости α и соответствующего числа степеней свободы оценки дисперсии S_i.

$$\varphi = N(\vartheta - 1)*$$

Если нулевое значение попадает в доверительный интервал для некоторого параметра, то соответствующий фактор признается несущественным, т.е. не оказывающим влияния на целевую величину.

*Здесь υ – число повторений наблюдений в каждой точке плана

4 Планирование эксперимента

Планирование эксперимента – это выбор плана эксперимента, удовлетворяющего заданным требованиям. Основным требованием при планировании эксперимента является минимизация времени и числа опытов при сохранении требуемой достоверности результатов. При *активном эксперименте* уровни факторов в каждом опыте задаются исследователем. **Уровень фактора** - фиксированное значение фактора относительно начала отсчета. Изменение уровней факторов происходит в соответствии с планом эксперимента. **План эксперимента** - совокупность данных, определяющих число, условия и порядок реализации опытов. Исследователь в каждом опыте изменяет *уровни управляемых факторов* $x_1, x_2, ..., x_n$, воздействующих на объект исследования, и регистрирует *отклик (выходной параметр)* y. В результате проведения эксперимента необходимо определить **функцию отклика** y - зависимость математического ожидания отклика от факторов $x_1, x_2, ..., x_n$:

$$y = f(x_1, x_2, ..., x_n; b_0, b_1, ...b_m).$$

Функцию отклика называют также *моделью регрессионного анализа*.

Функция отклика должна иметь числовое выражение, четкий физический или экономический смысл, быть информативной (т.е. характеризовать свойства объекта) и измеряться с требуемой точностью. Факторы должны быть управляемыми, совместимыми, независимыми, однозначными и непосредственно влиять на отклик.

Планирование эксперимента обладает рядом *преимуществ*:

- резко сокращается число испытаний;
- схема исследования объекта становится формализованной.
- проводится **последовательный эксперимент**, т.е. эксперимент, реализуемый в виде серий опытов, при этом условия проведения каждой последующей серии определяются результатами предыдущих
- процедура разработки математических моделей упрощается;

- точность математических моделей (их адекватность результатам эксперимента) повышается.

Планирование эксперимента в основном сводится к выбору числа уровней факторов и определению уровня каждого фактора в опыте. План, содержащий все возможные комбинации всех факторов на определенном числе уровней, называется **полным факторным планом,** а эксперимент, проводимый в соответствии с данным планом, – **полным факторным экспериментом (ПФЭ).** Выбранное число уровней ***p*** в сочетании с числом факторов ***k*** определяет число возможных опытов ***N***, которое равно $N = p^k$.

Для обработки результатов эксперимента факторы **нормализуют**, т.е. преобразуют натуральные значения факторов в безразмерные. Для определения параметров линейной модели достаточно каждый фактор фиксировать на одном из двух уровней: *верхнем и нижнем* (*верхний уровень* – большее значение, *нижний* – меньшее значение). Верхний уровень нормализованного фактора обозначают «+1», нижний «-1», среднее значение – «0». Нормализация факторов выполняется по следующей формуле

$$x_i = \frac{X_i - X_{CPi}}{\Delta X_i},\ \Delta X_i = \frac{X_{MAXi} - X_{MINi}}{2},\ X_{CPi} = \frac{X_{MAXi} + X_{MINi}}{2},$$

где x_i - нормализованное значение i–го фактора, X_i - натуральное значение i–го фактора; X_{CPi} - среднее значение i–го фактора, ΔX_i - интервал варьирования i–го фактора.

Рассмотрим планирование эксперимента на примере анализа зависимости отклика Y от двух факторов X_1 и X_2. При последовательном эксперименте порядок модели до опыта неизвестен. На первом этапе предполагается, что модель линейна и имеет виде

$$Y = b_0 + b_1 X_1 + b_2 X_2,$$

где X_1 и X_2 - два контролируемых фактора.

Для двух факторов количество опытов полного факторного плана равно $N = p^k = 2^2 = 4$. Составляют **матрицу плана -** стандартную форму записи условий проведения эксперимента в виде прямоугольной таблицы, строки которой отвечают опытам, столбцы – факторам (см. табл. 4.1).

Таблица 4.1 – Матрица полного факторного эксперимента для двух факторов

№ опыта u	Нормализованные значения факторов		Значение функции отклика, полученное в ходе эксперимента, y_u
	x_1	x_2	
1	*+1*	*+1*	y_1
2	*-1*	*+1*	y_2
3	*-1*	*-1*	y_3
4	*+1*	*-1*	y_4

Графические план ПФЭ для двух факторов в факторном пространстве можно представить в виде квадрата (см. рис. 4.1).

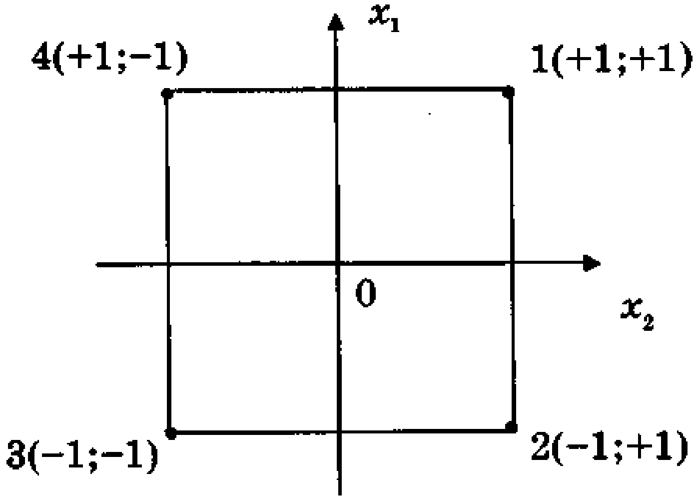


Рисунок 4.1 – Схема плана ПФЭ типа 2^2

Проводят эксперимент, определяют параметры b_0, b_1, b_2 и проверяют адекватность модели. Если модель адекватна, то заканчивают эксперимент. Адекватность модели проверяют с помощью критерия Фишера. В противном случае модель предполагается в виде

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2,$$

т. е. учитывается эффект взаимодействия факторов $X_1 X_2$, вычисляют параметры модели и проверяют ее адекватность. Коэффициент b_{12} оценивает эффект парного взаимодействия факторов X_1 и X_2, и показывает силу влияния одного фактора в зависимости от уровня другого. Матрица планирования эксперимента с учетом эффектов взаимодействия называется расширенной матрицей планирования (см. табл. 4.2).

Таблица 4.2 – Расширенная матрица полного факторного эксперимента для двух факторов

№ опыта u	Нормализованные значения факторов			Значение функции отклика, полученное в ходе эксперимента, y_u
	x_1	x_2	$x_1 \cdot x_2$	
1	+1	+1	+1	y_1
2	-1	+1	-1	y_2
3	-1	-1	+1	y_3
4	+1	-1	-1	y_4

Параметры нормализованной регрессионной модели определяются по формулам:

$$b_o = \frac{1}{N} \sum_{u=1}^{N} Y_u,$$

$$b_i = \frac{1}{N} \sum_{u=1}^{N} (x_i Y)_u,$$

$$b_{ij} = \frac{1}{N} \sum (x_i x_j Y)_u,$$

где N – число опытов; i, j – номера факторов; x_i, x_j – нормализованные значения факторов; Y_u – измеренное значение отклика в u–м опыте. Например, значение коэффициента b_1 определяется

$$b_1 = \frac{(+1) \cdot y_1 + (-1) \cdot y_2 + (-1) \cdot y_3 + (+1) \cdot y_4}{4}.$$

Если модель адекватна, то заканчивают эксперимент. Если модель не адекватна, то модель предполагается квадратичной в виде

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2 + b_{11} X_1^2 + b_{22} X_2^2.$$

Далее составляется план для определения параметров квадратичной модели, проводятся недостающие опыты, определяются параметры модели и проверяется ее адекватность.

Графическим представлением функции отклика является поверхность отклика (см. рис. 4.2).

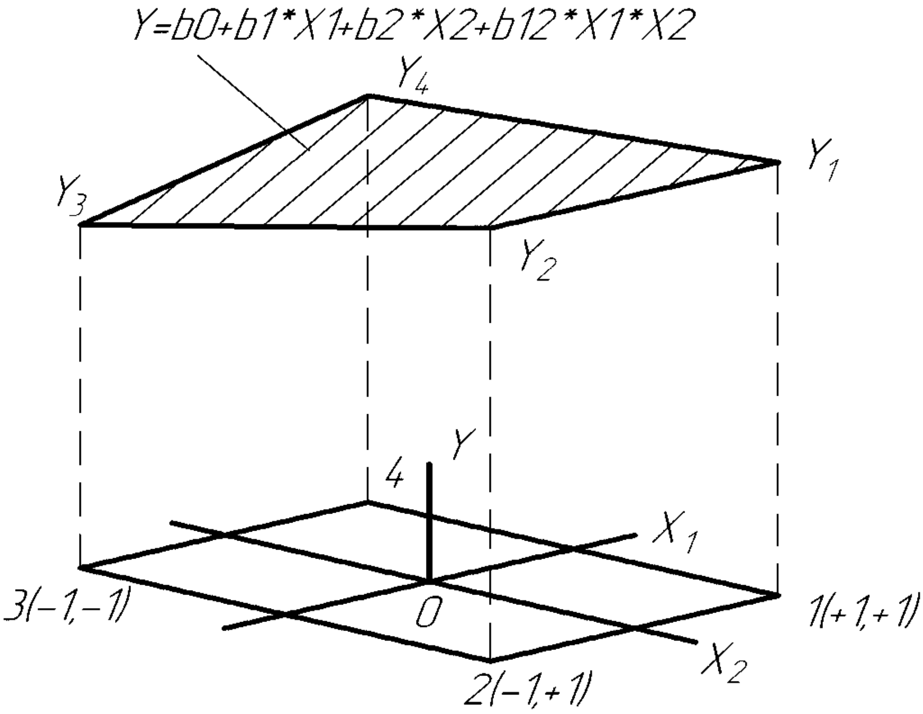


Рисунок 3.2 – Поверхность отклика $y = f(x_1, x_2)$ в факторном пространстве

После определения параметров нормализованное модели переходят к натуральной модели, с помощью которой можно определить значение отклика $\boldsymbol{y}$ по значениям факторов $\boldsymbol{x_1}$, $\boldsymbol{x_2}$.

Для трех факторов количество опытов ПФЭ равно $N = 2^3 = 8$, линейная регрессионная модель имеет вид $y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$, а матрица ПФЭ дана в табл. 4.3

Таблица 4.3 – Матрица полного факторного эксперимента для трех факторов

№ опыта, u	Нормализованные значения факторов			Значение отклика y_u
	x_1	x_2	x_3	
1	+1	+1	-1	y_1
2	-1	+1	-1	y_2
3	-1	-1	-1	y_3
4	+1	-1	-1	y_4
5	+1	+1	+1	y_5
6	-1	+1	+1	y_6
7	-1	-1	+1	y_7
8	+1	-1	+1	y_8

С увеличением количества факторов количество опытов полного факторного эксперимента резко возрастает. Например, для 5 факторов количество опытов ПФЭ равно 2^5=32.

Для сокращения количества опытов при большом числе факторов используется ***дробный факторный эксперимент (ДФЭ)*** – это эксперимент, содержащий часть комбинаций полного факторного эксперимента.

Идея построения плана дробного факторного эксперимента заключается в замене в расширенной матрице планирования ПФЭ наиболее слабого эффекта парного взаимодействия (произведения факторов) новым фактором.

Рассмотри это на примере ДФЭ для 3 факторов. Необходимо найти по результатам активного эксперимента параметры линейной модели для 3 факторов:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 .$$

Для полного факторного эксперимента необходимое количество опытов составляет $N=p^k=2^3=8$. Для сокращения количества опытов используем дробный факторный эксперимент. Для построения плана ДФЭ в расширенной матрице планирования ПФЭ типа 2^2 (см. табл. 4.3) заменяем эффект парного взаимодействия X_1* X_2 новым фактором X_3. Необходимое количество опытов ДФЭ составит 2^{3-1} =4 вместо 8 для ПФЭ (см. табл. 4.4). Параметры модели определяется как же, как и для ПФЭ.

Таблица 4.4 - Матрица планирования ДФЭ для 3 факторов (полуреплика 2^{3-1})

№ опыта, u	Нормализованные значения факторов			Значение отклика y_u
	x_1	x_2	x_3	
1	*+1*	*+1*	*+1*	y_1
2	*-1*	*+1*	*-1*	y_2
3	*-1*	*-1*	*+1*	y_3
4	*+1*	*-1*	*-1*	y_4

При составлении плана ДФЭ наиболее распространены регулярные дробные реплики, которые получают делением числа опытов соответствующего ПФЭ на число, кратное двум. Составляю дробные реплики заменой некоторых эффектов взаимодействия новыми факторами. Минимальная дробная реплика для построения линейной модели должна включать (k+1) опытов, где k – число факторов.

Список использованных источников

1. Кане, М. М. Анализ исходных данных при статистической обработке результатов научных исследований [Текст] : учебное пособие / М. М. Кане. - Минск : Вышэйшая школа, 2024. - 118, [2] с. - Библиография: с. 116 (21 назв.). - Допущено Министерством образования Республики Беларусь в качестве учебного пособия для студентов учреждений высшего образования по специальности магистратуры "Инновационные технологии в машиностроении". - 200 экз. - ISBN 978-985-06-3600-3.

2. Кане, М. М. Предварительный анализ экспериментальных данных. Оценка правомерности применения корреляционно-регрессионного анализа в заданных условиях : пособие для студентов специальностей 1-36 80 02 «Инновационные технологии в машиностроении» и 1-53 80 01 «Автоматизация» / М. М. Кане ; Белорусский национальный технический университет, Кафедра «Технология машиностроения». – Минск : БНТУ, 2020. – 49 с. – Рекомендовано УМО РБ. – ISBN 978-985-583-538-8. URL: https://rep.bntu.by/handle/data/813512.

3. Электронный учебно-методический комплекс по учебной дисциплине «Основы научных исследований, изобретательства и инновационной деятельности в машиностроении» для специальностей 1-36 01 01 «Технология машиностроения» и направления специальности 1-53 01 01 «Автоматизация технологических процессов и производств (машиностроение и приборостроение)» [Электронный ресурс] / Белорусский национальный технический университет, Кафедра «Технология машиностроения» ; сост.: М. М. Кане, В. К. Шелег, М. А. Кравчук. – Минск : БНТУ, 2023. URL: https://rep.bntu.by/handle/data/127511

ПРАКТИЧЕСКИЙ РАЗДЕЛ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«Брестский государственный технический университет»

Кафедра «Техническая эксплуатация автомобилей»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
к практическим занятиям
по дисциплине
«Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований»
для студентов специальности
7-06-0714-02 Инновационные технологии в машиностроении
Профилизация: Аддитивные и субтрактивные технологии

Брест 2025

УДК 629.331.08

Методические указания к лабораторным занятиям по дисциплине «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» для студентов специальности 7-06-0714-02 Инновационные технологии в машиностроении содержат методику проведения корреляционного и регрессионного анализа, планирования эксперимента и могут использоваться при выполнении практических работ.

Составитель: С.В. Монтик, зав. кафедрой ТЭА, доцент, к.т.н.

Практическая работа №1
КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Цель: изучить задачи и методику проведения корреляционного анализа, примеры его практического применения

Краткие сведения о корреляционном анализе

Зависимость между величинами, при которой каждому значению одной величины отвечает с соответствующей вероятностью множество возможных значений другой величины, называют **вероятностной**.

Если при наличии вероятностной зависимости между двумя величинами с изменением значения одной величины изменяется только математическое ожидания второй и наоборот, а дисперсия и тип закона распределения остаются неизменными, то для таких величин характерна **корреляционная зависимость**.

Задача корреляционного анализа – выявление наличия и тесноты (силы) связи между значениями различных случайных величин.

Примеры корреляционной связи: между величиной износа узла трения и пробегом автомобиля; между пределом прочности и пределом текучести стали; между твердостью и износостойкостью стали.

При выполнении корреляционного анализа по результатам эксперимента строят корреляционное поле, по которому по тесноте группирования точек вокруг прямой или кривой линии можно визуально судить о наличии корреляционной зависимости (см. рис.1.1).

Силу линейной статистической связи между случайным величинами X и Y оценивают **коэффициентом корреляции** r_{XY}, который принимает значения в интервале от -1 до +1 и не зависит от единиц величин X и Y. Чем больше по абсолютной величине коэффициент корреляции, тем сильнее линейная зависимость между величинами X и Y. Однако обратное не всегда верно. Силу нелинейной статистической связи между случайными величинами оценивают с помощью *корреляционного отношения*.

Задание

Используя изложенную ниже методику проведения корреляционного анализа и табличный процессор MS Excel, определите, существует ли линейная корреляционная зависимость между высотой над уровнем моря и снижением эффективной мощности двигателя автомобиля. Письменно ответить на контрольные вопросы. Исходные данные (таблица А.1), справочные данные (таблица Б.1) и шаблон отчета по лабораторной работе содержатся в файле **Лаб_раб_ОНИиИД1.xls. *Формулы нужно запрограммировать.***

Содержание отчета по лабораторной работе

Тема, цель, исходные данные, распечатка отчета по лабораторной работе (файл **Лаб_раб_ОНИиИД1.xls,** письменные ответы на контрольные вопросы.

Контрольные вопросы

1. Перечислите задачи корреляционного анализа?
2. Какая зависимость называется вероятностной?
3. Какая зависимость называется корреляционной?
4. Привести примеры корреляционной зависимости.
5. Что характеризует коэффициент корреляции, его возможные значения?

6. Изобразите примеры корреляционных полей для : а) прямой корреляционной зависимости; б) обратной корреляционной зависимости; в) нелинейной корреляционная зависимость; г) отсутствия корреляционной зависимости.

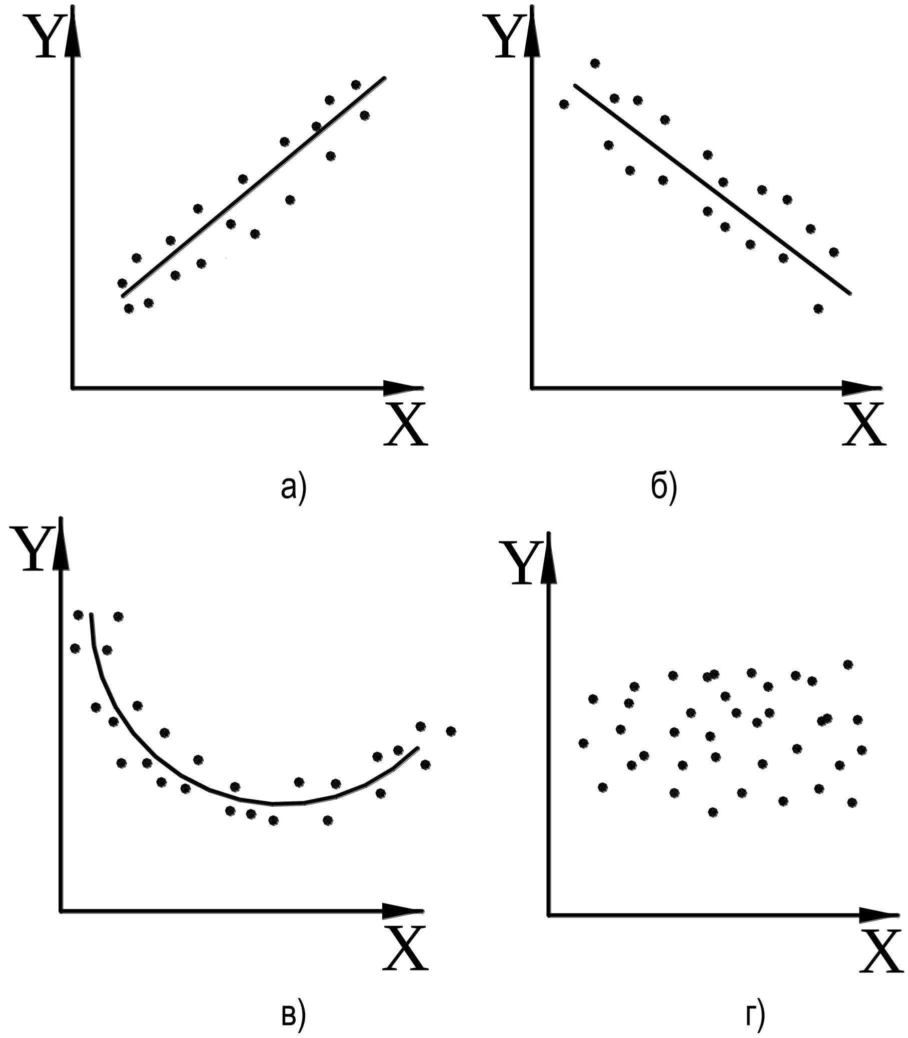


Рисунок 1.1 – Примеры корреляционных полей: а) прямая корреляционная зависимость, $r_{XY} \approx 1$; б) обратная корреляционная зависимость, $r_{XY} \approx -1$; в) нелинейная корреляционная зависимость, $r_{XY} = 0$; г) отсутствие корреляционной зависимости, $r_{XY} = 0$

Пример расчета

Исходные данные

Доверительная вероятность P=0,95

Таблица 1.1 - Результаты измерений (*исходные данные из таблицы А.1*)

Номер измерения i	Высота над уровнем моря, м **(величина X)**	Снижение эффективной мощности двигателя, % **(величина Y)**
1	*0*	*0,0*
2	*200*	*4,7*
3	*400*	*6,6*
4	*600*	*8,7*

Номер измерения i	Высота над уровнем моря, м **(величина X)**	Снижение эффективной мощности двигателя, % **(величина Y)**
5	*800*	*11,1*
6	*1000*	*7,0*
7	*1200*	*9,0*
8	*1400*	*10,5*
9	*1600*	*11,1*
10	*1800*	*14,1*

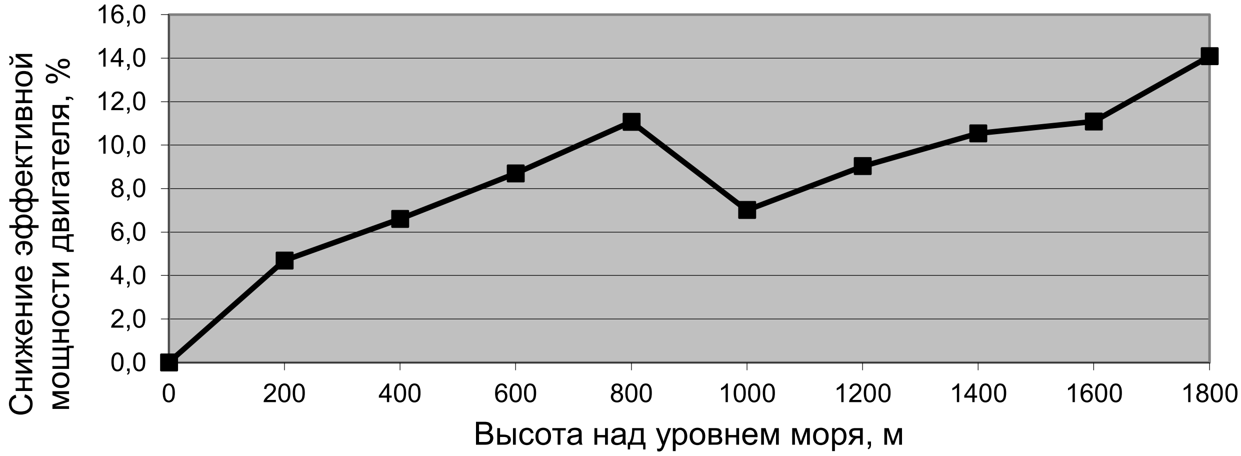


Рисунок 1.2 – Зависимость снижения эффективной мощности двигателя от высоты над уровнем моря

Порядок выполнения

1.Определите коэффициент корреляции

Эмпирический коэффициент корреляции определяется по выражению

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$

$$\sigma_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2;$$

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

где $x_i. y_i - i$- е значение случайных величин X и Y; $\bar{x}, \bar{y}, \sigma_x, \sigma_y$, – соответственно средние значения и средние квадратические отклонения случайных величин, n – количество измерений, n=10.

Для определения коэффициента корреляции в табличном процессоре MS Excel используем функцию **PEARSON**, которая возвращает **коэффициент корреляции Пирсона** — безразмерный индекс в интервале от -1,0 до 1,0 включительно, который отражает степень линейной зависимости между двумя множествами данных.

Синтаксис: **PEARSON(массив1; массив2).**

Формула имеет вид **=PEARSON(A3:A7;B3:B7).**

Для примера коэффициент корреляции r_{xy} = 0,87

2.Проверьте значимость коэффициента корреляции.

Для этого определяют наблюдаемое значение критерия Стьюдента.

$$t_H = r_{xy} \cdot \sqrt{\frac{n-2}{1-{r_{xy}}^2}},$$

где n - число измерений. Для примера $t_н$= 5,010.

Критическое значение критерия Стьюдента t_k определяется по таблице Б.1 для заданной доверительной вероятности P и числу степеней свободы f = n – 2.

При выборе t_k по таблице Б.1 следует принимать m = f + 1, т. е.

f= n – 2 = 10 – 2 = 8

m = f + 1 = 8 + 1 = 9.

Для P=0,95 и m = 9 критическое значение критерия Стьюдента t_k =2,306.

Если $|t_H| < t_k$, то коэффициент корреляции r_{xy} = 0 и связи между величинами нет. Если $|t_H| > t_k$, то коэффициент корреляции значим и между величинами X и Y существует линейная корреляционная зависимость. *Сделайте вывод о наличии зависимости между X и Y.*

Вывод (для примера): Между высотой над уровнем моря (величина X), в метрах, и снижением эффективной мощности двигателя (величина Y), в процентах, существует линейная корреляционная зависимость.

Практическая работа №2
ПРИМЕНЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА

Цель: освоить методику применения регрессионного анализа

Краткие сведения о регрессионном анализе

Задача регрессионного анализа – установление вида и параметров зависимости отклика y от уровней одного или нескольких факторов $x_1, x_2, ..., x_n$, т.е. в результате регрессионного анализа определяют вид и параметры функции

$$y = f(x_1, x_2, ..., x_n; b_0, b_1, ...b_m),$$

Данную функцию называют **регрессионной моделью**, а ее параметры $b_0, b_1, ...b_m$ – *коэффициентами регрессии*.

Коэффициенты регрессии могут вычисляться с помощью *метода наименьших квадратов,* при котором коэффициенты модели определяются из условия: сумма S квадратов отклонений экспериментальных точек y_i от точек на теоретической кривой должна быть минимальной, т.е.

$$S = \sum_{i=1}^{N} \left(y_i - f(x_1, x_2, ..., x_n; b_0, b_1, ...b_m) \right)^2 \to \min,$$

где N – количество экспериментальных точек.

Регрессионная модель позволяет определять значение отклика y по заданному значению факторов $x_1, x_2, ..., x_n$ не проводя эксперимент. Например, регрессионный анализ может использоваться для прогнозирования грузооборота или пассажирооборота автотранспортного предприятия на перспективу.

Регрессионные модели могут быть однофакторные и многофакторные, линейные и нелинейные. Примеры регрессионных моделей:

- линейная однофакторная модель $y = b_o + b_1 \cdot x$, где b_o, b_1 - параметры регрессионной модели (коэффициенты регрессии); *x* – фактор; *y* – отклик;
- нелинейная однофакторная модель $y = b_0 x^{b_1}$ (может использоваться для прогнозирования грузооборота или пассажирооборота);
- линейная многофакторная модель $y = b_o + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$;
- нелинейная многофакторная модель $y = b_o + b_1 \cdot \ln x_1 + b_2 \cdot \ln x_2 + b_3 \cdot \ln x_3$.

Задание

Более 85 % деталей тракторов и автомобилей и 95% деталей двигателей выбраковываются при износе не более 0,3 мм. Их целесообразно восстанавливать гальваническими покрытиями. Свойства гальванических покрытий определяются структурой покрытия, под которым понимают размер и форму кристаллов осажденного металла. На структуру покрытия влияет режим электролиза. Важнейшая характеристика свойств гальванических покрытий – твердость. Чем жестче режим электролиза (больше плотность тока, меньше концентрация и температура электролита), тем меньше размеры кристаллов и тем больше твердость покрытий.

Необходимо определить вид и параметры регрессионной модели, описывающей зависимость микротвердости железных покрытий H, МПа, от плотности тока Dk, А/дм^2.

В данном примере микротвердость железных покрытий является откликом Y, который зависит от уровней (значений) фактора X, которым является плотности тока. В результате регрессионного анализа необходимо установить вид и параметры зависимости отклика Y(микротвердости покрытия) от уровней фактора X (плотности тока).

Пример расчета

Исходные данные

Доверительная вероятность P=0,90

Таблица 2.1 - Результаты измерений

N п/п	Плотность тока Dk, А/дм^2	Микротвердость поверхности H, МПа
(i)	(X)	($Y_{изм}$)
1	*10*	*3000*
2	*20*	*4400*
3	*30*	*5000*
4	*40*	*5200*
5	*50*	*5400*
6	*60*	*5600*
7	*70*	*5800*
8	*80*	*6000*
9	*90*	*6200*
10	*100*	*6400*

Температура электролита составляла 80 ^{o} С, концентрация $FeCl_2$ *$4H_2O$ составляла 200 г/л.

Для определения дисперсии воспроизводимости $S_в^2$ проводились 10 дублирующих опытов (m=10) при фиксированном значении плотности тока. Дисперсия воспроизводимости $S_в^2$ определялась по формуле

$$S_в^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_{изм\,i} - \overline{Y}_{изм})^2,$$

где $\overline{Y}_{изм}$ - среднее значение измеренных величин, Yизм i – результаты ***i***-го дублирующего измерения. Дисперсия воспроизводимости $S_в^2$ = 3600 МПа^2.

Порядок выполнения

Предполагаем, что регрессионная модель имеет вид

$$Y_{расч} = b_0 + b_1 X + b_2 X^2,$$

где b_0, b_1, b_2 - коэффициенты регрессии.

Для определения коэффициентов регрессии составляем систему уравнений

$$\begin{cases} b_0 n + b_1 \sum X_i + b_2 \sum X_i^2 = \sum Y_i \\ b_0 \sum X_i + b_1 \sum X_i^2 + b_2 \sum X_i^3 = \sum (YX)_i \\ b_0 \sum X_i^2 + b_1 \sum X_i^3 + b_2 \sum X_i^4 = \sum (YX^2)_i \end{cases}$$

Для примера система имеет вид

$$\begin{cases} 10\,b_0 + 550\,b_1 + 38500\,b_2 = 5.3\cdot 10^4 \\ 550\,b_0 + 38500\,b_1 + 3.02*10^6 b_2 = 3.16\cdot 10^6 \\ 38500\,b_0 + 3.02*10^6 b_1 + 2.5*10^8\,b_2 = 2.3*10^8 \end{cases}.$$

Решая систему уравнений определяем значения коэффициентов регресии

b_0= 2760
b_1= 73,76
b_2= -0,39

Определяем дисперсию оценок коэффициентов регрессии.
Для j-го коэффициента регрессии

$$S^2(b_j) = S_в^2 C_{jj},$$

где $S_в^2$ – дисперсия воспроизводимости; C_{jj} – элемент матрицы $Ф^{-1}$, обратной информационной.

Информационная матрица имеет вид

$$Ф = \begin{pmatrix} n & \sum X_i & \sum X_i^2 \\ \sum X_i & \sum X_i^2 & \sum X_i^3 \\ \sum X_i^2 & \sum X_i^3 & \sum X_i^4 \end{pmatrix}$$

Обратная матрица определяется по формуле

$$Ф^{-1} = \frac{1}{\det Ф} \begin{pmatrix} D_{00} & D_{10} & D_{20} \\ D_{01} & D_{11} & D_{21} \\ D_{02} & D_{12} & D_{22} \end{pmatrix},$$

где det*Ф* – детерминант матрицы *Ф*, D_{jk} – алгебраическое дополнение элемента матрицы в j–ой строке и k–ом столбце.

Для примера дисперсии оценок коэффициентов регрессии

$$S^2(b_0)= 4{,}9*10^3;$$
$$S^2(b_1)= 8{,}6;$$
$$S^2(b_2)= 6{,}8*10^{-4}.$$

Проверяем значимость полученных коэффициентов регрессии. Для этого определяем наблюдаемые значения критерия Стьюдента для каждого коэффициента регрессии bj по формуле

$$t_{н j} = \frac{|b_j|}{S(b_j)}.$$

Для примера наблюдаемые значения критерия Стьюдента равны tн0= 39,11; tн1= 25,02; tн2= 15,08.

Критическое значение критерия Стьюдента определяется для P=0.90 и m=10 равно t_k=*1,833.*

Если $t_{нi}>t_k$, то коэффициент b_i значим, в противном случае – $b_i = 0$

Для примера принятые значения b_i:

b_0= 2760

b_1= 73,76

b_2= -0,39

Полученная регрессионная модель имеет вид

$$Y_{расч} = 2760 + 73.76 * X - 0.39 * X^2,$$

где Yрасч - микротвердость поверхности H, МПа; X - плотность тока Dk, А/дм^2

Число параметров модели d равно числу значимых коэффициентов регрессии. Для примера d = *3.*

Результаты расчета микротвердости поверхности по полученной модели даны в таблице 2.2 и на рисунке 2.1.

Таблица 2.2– Результаты измерений и расчета по модели

N п/п	Плотность тока Dk, А/дм^2	Микротвердость поверхности H, МПа	
		Результаты измерений	Результаты расчета по регрессионной модели
(g)	(X)	($Y_{измер}$)	($Y_{расч}$)
1	10	3000	3458,5
2	20	4400	4079
3	30	5000	4621,5
4	40	5200	5086
5	50	5400	5472,5
6	60	5600	5781
7	70	5800	6011,5
8	80	6000	6164
9	90	6200	6238,5
10	100	6400	6235

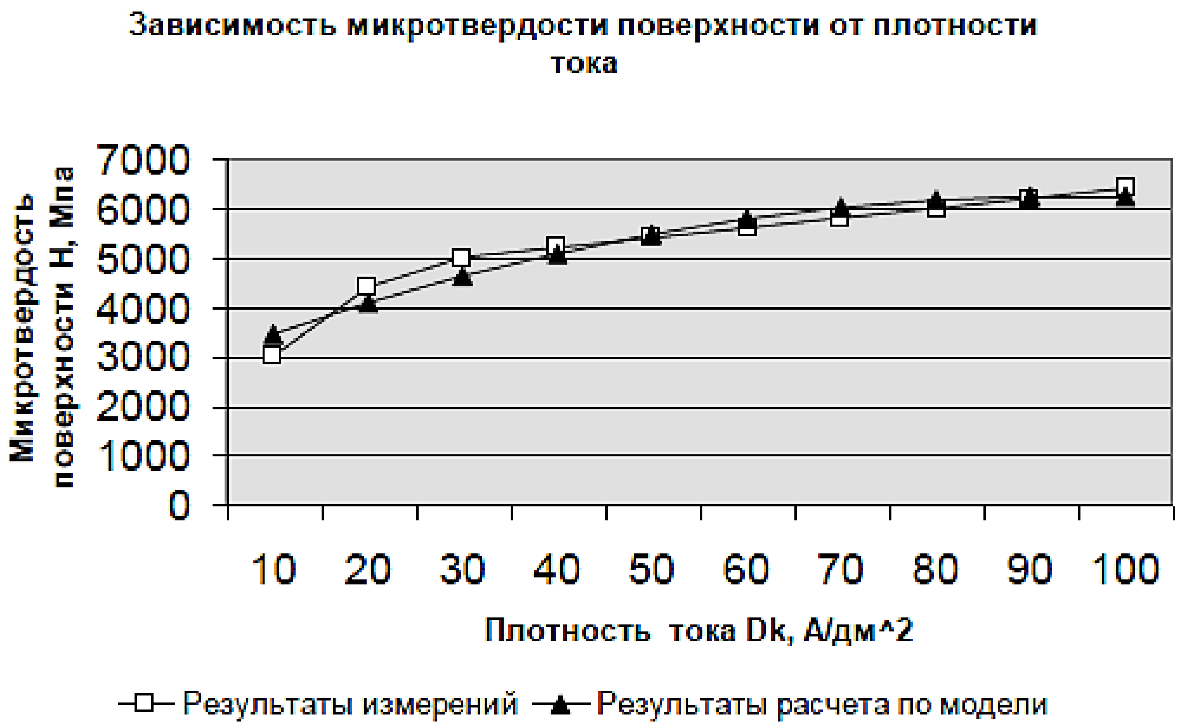


Рисунок 2.1 – Зависимость микротвердости поверхности от плотности тока (по результатам измерений и по полученной регрессионной модели)

Проверяем адекватность полученной регрессионной модели. Для этого вычисляем остаточную дисперсию $S_о^2$ и сопоставляем ее с дисперсией воспроизводимости $S_в^2$ с помощью критерия Фишера.

$$S_о^2 = \frac{1}{n-(d+1)} \sum_{g=1}^{n} (Y_{измер_g} - Y_{расч_g})^2,$$

где n – количество измерений, n = 10.

Для примера остаточная дисперсия составляет $S_о^2$= 101312 МПа^2

Наблюдаемое значение критерия Фишера определяется

$$F_н = \frac{S_о^2}{S_в^2}$$

Для примера $F_н$=28,14.

Критическое значение критерия Фишера F_k определяется для заданной доверительной вероятности P, m_1 и m_2, где m_1 и m_2 определяются в зависимости от степеней свободы f_1 и f_2, количества параметров модели ***d***, количества измерений ***n*** и количества дублирующих опытов ***m***:

f_1= n - d, f_1=10 - 3=7

$f_2= m - 1$, $f_2=10 – 1 =9$
$m_1= f_1+1$, $m_1=7+1=8$
$m_2= f_2+1$, $m_2= 9 + 1 = 10$

Для примера при P= 0,90, m_1=8, m_2=10 критическое значение критерия Фишера $F_k=2.70$.

Если $F_н<F_k$, то модель адекватна. В противном случае – не адекватна. Сделайте заключение об адекватности модели.

Вывод: **Полученная регрессионная модель адекватна**

Контрольные вопросы

1. В чем заключается задача регрессионного анализа? Что определяется в результате данного анализа?
2. Как определяются коэффициенты регрессии?
3. Для чего может использоваться регрессионная модель?
4. Запишите вид регрессионной модели при использовании полинома второй степени для описания связи между X и Y, поясните все обозначения. Запишите формулы для расчета параметров модели.
5. Как проверяют адекватность регрессионной модели?

Практическая работа №3
ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА. ОПРЕДЕЛЕНИЕ ВИДА ЛИНЕЙНОЙ ПО ПАРАМЕТРАМ МНОГОФАКТОРНОЙ МОДЕЛИ (ПОСЛЕДОВАТЕЛЬНОЕ ПЛАНИРОВАНИЕ)

Цель: освоить методику планирования полного факторного эксперимента для построения линейной двухфакторной модели

Теоретические сведения

Эксперимент - это система операций, воздействий и (или) наблюдений, направленных на получение информации об объекте при исследовательских испытаниях. Эксперимент состоит из опытов.

Опыт - это воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов. Условия опытов определяются **уровнями (значениями) факторов**.

Фактор - это переменная величина, по предположению влияющая на результаты эксперимента. **Уровень фактора** - это фиксированное значение фактора относительно начала отсчета.

Отклик - это наблюдаемая случайная переменная, по предположению, зависящая от факторов.

Пассивный эксперимент – это эксперимент, при котором уровни факторов в каждом опыте регистрируются исследователем, но не задаются. Пример пассивного эксперимента - это **подконтрольная эксплуатация автомобилей**. В этом случае в автотранспортном предприятии (АТП) выделяется специальная группа подконтрольных автомобилей (выборка), выполняющая обычную транспортную работу. На каждый автомобиль заводится специальный журнал, где фиксируется и накапливается информация о всех отказах и неисправностях, на каком пробеге они произошли или выявлены, данные о нагрузках, виде перевозимого груза, среднесуточных пробегах, пробегах до ТО и между ремонтами и т.п. Основной недостаток пассивного эксперимента: информация слишком «запаздывает», т.е. время обратной связи очень значительно. Например, от разработки какого-либо узла до момента поступления информации о его надежности из сферы эксплуатации проходит несколько лет.

Активный эксперимент - это эксперимент, в котором уровни факторов в каждом опыте задаются исследователем. Активный эксперимент проводится в соответствии с планом эксперимента.

Этапы проведения экспериментальных исследований

1. Формирование целей исследований
2. Выдвижение гипотезы об исследуемом объекте
3. Планирование эксперимента
4. Проведение эксперимента
5. Обработка и анализ результатов
6. Проверка правильности выдвинутой гипотезы
7. Выдвижение новой гипотезы, если необходимо
8. Проверка условий окончания эксперимента
9. Планирование нового эксперимента (при необходимости)

Планирование эксперимента – это выбор плана эксперимента, удовлетворяющего заданным требованиям. Основным требованием при планировании эксперимента является минимизация времени и числа опытов при сохранении требуемой достоверности результатов. Изменение уровней факторов происходит в соответствии с планом эксперимента.

План эксперимента - совокупность данных, определяющих число, условия и порядок реализации опытов. Исследователь в каждом опыте изменяет *уровни управляемых факторов* $x_1, x_2, ..., x_n$, воздействующих на объект исследования, и регистрирует *отклик (выходной параметр)* y. В результате проведения эксперимента необходимо определить **функцию отклика** y - зависимость математического ожидания отклика от факторов $x_1, x_2, ..., x_n$:

$$y = f(x_1, x_2, ..., x_n; b_0, b_1, ...b_m).$$

где $b_0, b_1, ...b_m$ – параметры модели.

Функцию отклика называют также *моделью регрессионного анализа или регрессионной моделью, а* $b_0, b_1, ...b_m$ – коэффициентами регрессии.

Функция отклика должна иметь числовое выражение, четкий физический или экономический смысл, быть информативной (т.е. характеризовать свойства объекта) и измеряться с требуемой точностью. Факторы должны быть управляемыми, совместимыми, независимыми, однозначными и непосредственно влиять на отклик [1,2].

Планирование эксперимента обладает рядом *преимуществ [2]*:

- резко сокращается число испытаний;
- схема исследования объекта становится формализованной.
- проводится **последовательный эксперимент**, т.е. эксперимент, реализуемый в

виде серий опытов, при этом условия проведения каждой последующей серии определяются результатами предыдущих
- процедура разработки математических моделей упрощается;
- точность математических моделей (их адекватность результатам эксперимента) повышается.

Планирование эксперимента в основном сводится к выбору числа уровней факторов и определению уровня каждого фактора в опыте.

План, содержащий все возможные комбинации всех факторов на определенном числе уровней, называется **полным факторным планом,** а эксперимент, проводимый в соответствии с данным планом, – **полным факторным экспериментом (ПФЭ).** Выбранное число уровней ***p*** в сочетании с числом факторов ***k*** определяет число возможных опытов ***N****,* которое равно $N = p^k$.

Для обработки результатов эксперимента факторы **нормализуют***,* т.е. преобразуют натуральные значения факторов в безразмерные. Для определения параметров линейной модели достаточно каждый фактор фиксировать на одном из двух уровней: *верхнем и нижнем* (*верхний уровень* – большее значение, *нижний* – меньшее значение). Верхний уровень нормализованного фактора обозначают «+1», нижний «-1», среднее значение – «0». Нормализация факторов выполняется по следующей формуле

$$x_i = \frac{X_i - X_{CPi}}{\Delta X_i}, \ \Delta X_i = \frac{X_{MAXi} - X_{MINi}}{2}, \ X_{CPi} = \frac{X_{MAXi} + X_{MINi}}{2}, \tag{3.1}$$

где x_i - нормализованное значение i–го фактора, X_i - натуральное значение i–го фактора; X_{CPi} - среднее значение i–го фактора, ΔX_i - интервал варьирования i–го фактора.

Рассмотрим планирование эксперимента на примере анализа зависимости отклика Y от двух факторов X_1 и X_2. При последовательном эксперименте порядок модели до опыта неизвестен. На первом этапе предполагается, что модель линейна и имеет виде

$$Y = b_0 + b_1 X_1 + b_2 X_2, \tag{3.2}$$

где X_1 и X_2 – два контролируемых фактора, b_0*,* b_1*,* b_2 – параметры модели (коэффициенты регрессии).

Для двух факторов количество опытов полного факторного плана равно $N = p^k = 2^2 = 4$. Составляют **матрицу плана –** стандартную форму записи условий проведения эксперимента в виде прямоугольной таблицы, строки которой отвечают опытам, столбцы – факторам (см. табл. 3.1).

Таблица 3.1 – Матрица полного факторного эксперимента для двух факторов

№ опыта u	Нормализованные значения факторов		Значение функции отклика, полученное в ходе эксперимента, y_u
	x_1	x_2	
1	*+1*	*+1*	y_1
2	*-1*	*+1*	y_2
3	*-1*	*-1*	y_3

№ опыта u	Нормализованные значения факторов		Значение функции отклика, полученное в ходе эксперимента, y_u
	x_1	x_2	
4	*+1*	*-1*	y_4

Графические план ПФЭ для двух факторов в факторном пространстве можно представить в виде квадрата (см. рис. 3.1).

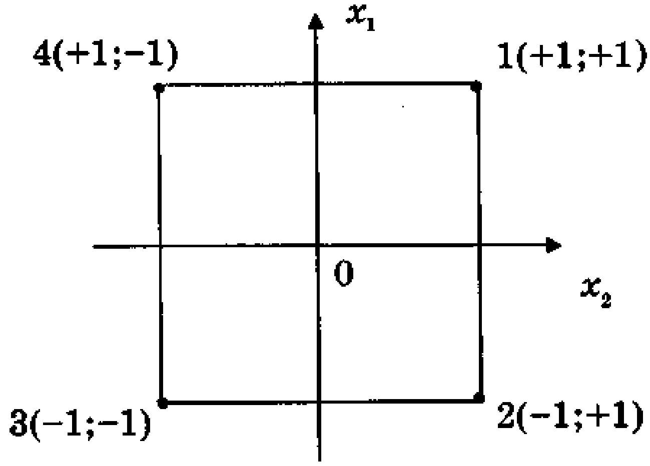


Рисунок 3.1 – Схема плана ПФЭ типа 2^2 [2]

Проводят эксперимент, определяют параметры b_0, b_1, b_2 и проверяют адекватность модели. Если модель адекватна, то заканчивают эксперимент. Адекватность модели проверяют с помощью критерия Фишера. В противном случае модель предполагается в виде

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2, \tag{3.3}$$

т. е. учитывается эффект взаимодействия факторов $X_1 X_2$, вычисляют параметры модели и проверяют ее адекватность. Коэффициент b_{12} оценивает эффект парного взаимодействия факторов X_1 и X_2, и показывает силу влияния одного фактора в зависимости от уровня другого. Матрица планирования эксперимента с учетом эффектов взаимодействия называется расширенной матрицей планирования (см. табл. 3.2).

Таблица 3.2 – Расширенная матрица полного факторного эксперимента для двух факторов

№ опыта u	Нормализованные значения факторов			Значение функции отклика, полученное в ходе эксперимента, y_u
	x_1	x_2	$x_1 \cdot x_2$	
1	*+1*	*+1*	*+1*	y_1
2	*-1*	*+1*	*-1*	y_2
3	*-1*	*-1*	*+1*	y_3
4	*+1*	*-1*	*-1*	y_4

Параметры нормализованной регрессионной модели определяются по формулам:

$$b_o = \frac{1}{N}\sum_{u=1}^{N} Y_u,$$

$$b_i = \frac{1}{N}\sum_{u=1}^{N} (x_i Y)_u,$$

$$b_{ij} = \frac{1}{N}\sum (x_i x_j Y)_u, \qquad (3.4)$$

где N – число опытов; i, j – номера факторов; x_i, x_j – нормализованные значения факторов; Y_u – измеренное значение отклика в u–м опыте. Например, значение коэффициента b_1 определяется

$$b_1 = \frac{(+1)\cdot y_1 + (-1)\cdot y_2 + (-1)\cdot y_3 + (+1)\cdot y_4}{4}.$$

Графическим представлением функции отклика является поверхность отклика (см. рис. 3.2).

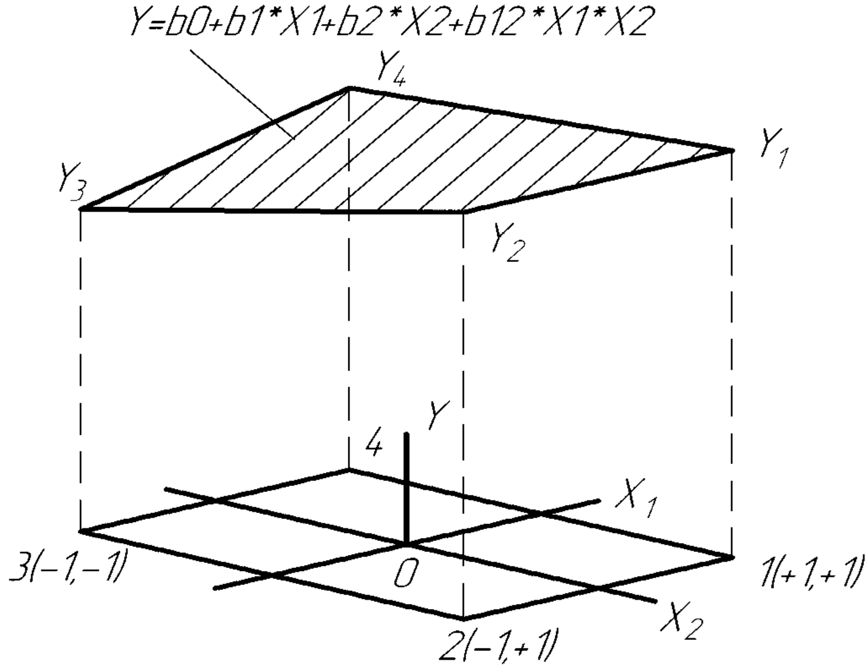


Рисунок 3.2 – Поверхность отклика $y = f(x_1, x_2)$ в факторном пространстве

Если модель адекватна, то заканчивают эксперимент. Если модель не адекватна, то модель предполагается квадратичной в виде

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2 + b_{11} X_1^2 + b_{22} X_2^2. \qquad (3.5)$$

Далее составляется план для определения параметров квадратичной модели, проводятся недостающие опыты, определяются параметры модели и проверяется ее адекватность.

После определения параметров нормализованное модели переходят к натуральной модели, с помощью которой можно определить значение отклика $\boldsymbol{Y}$ по значениям факторов $\boldsymbol{X_1}$, $\boldsymbol{X_2}$. не проводя эксперимент.

Задание

Используя изложенную ниже методику планирования полного факторного эксперимента и табличный процессор MS Excel, определите параметры модели зависимости средней толщины h (отклик Y), осажденного на катоде металла при нанесении гальванического покрытия, от плотности тока на катоде Dk (фактор X_1) и времени нанесения покрытия t (фактор X_2), а также спрогнозировать значение толщины гальванического покрытия для заданных плотности тока и времени нанесения гальванического покрытия. Модель ищем в виде (3.3). Письменно ответить на контрольные вопросы. Исходные данные (таблица А.3) и шаблон отчета по лабораторной работе содержатся в файле **Лаб_раб_ОНИиИД1.xls.** ***Формулы нужно запрограммировать.***

Пример расчета

Таблица 3.3 - Исходные данные

Наименование процесса	*Износостойкое хромирование*
Плотность металла покрытия, г/см^3	*6,9*
Электрохимический эквивалент С, г/А*ч	*0,324*
Выход металла по току, %	*13*
Пределы изменения плотности тока Dk, А/дм^2	*50 - 75*
Пределы изменения времени нанесения покрытия t, ч	*4 – 6*
Дисперсия воспроизводимости Sв, мм^2	*0,0001*
Доверительная вероятность P	*0,95*
Уровень значимости $\alpha = 1 - \mathrm{P}$	*0,05*
Определите толщину наносимого гальванического покрытия для	
Время нанесения t, ч	*6,5*
Плотности тока , А/дм^2	*65*

Порядок выполнения

1 Нормализуем натуральные значения факторов по формулам (3.1) и сводим в таблицу 3.4

Для фактора X_1 (плотности тока на катоде):

$$X_{CP1} = \frac{X_{MAX1} + X_{MIN1}}{2} = \frac{75+50}{2} = 62.5;\ \Delta X_1 = \frac{X_{MAX1} - X_{MIN1}}{2} = \frac{75-50}{2} = 12.5$$

$$x_{1MIN} = \frac{X_{1MIN} - X_{CP1}}{\Delta X_1} = \frac{50-62.5}{12.5} = -1;\ x_{1MAX} = \frac{X_{1MAX} - X_{CP1}}{\Delta X_1} = \frac{75-62.5}{12.5} = 1$$

Таблица 3.6 – Натуральные и нормализованные значения факторов

Обозначение фактора	Минимальное значение фактора		Максимальное значение фактора		Среднее значение фактора		Интервал варьирования фактора ΔX (натуральное значение)
	Натуральное	Нормализованное	Натуральное	Нормализованное	Натуральное	Нормализованное	
X_1	50	-1	75	+1	62,5	0	12,5
X_2	4	-1	6	+1	5	0	1

2 Составляем матрицу планирования полного факторного эксперимента (ПФЭ) для 2 факторов (по аналогии с табл. 3.2).

После составления матрицы планирования ПФЭ по составленному плану проводится эксперимент, при этом в каждом опыте фактор фиксируется на заданном уровне (минимальном или максимальном). В результате проведения каждого опыта определяется значение отклика Y, т.е. значение средней толщины h (мм) гальванического покрытия. Данные значения также заносятся в матрицу планирования (см. табл.3.7).

Таблица 3.7 – План полного факторного эксперимента (факторный план)

№ опыта u	X_1 (плотность тока Dk)		X_2 (временя нанесения покрытия t)		$X_1 \cdot X_2$	Значение отклика Y (Толщина покрытия h, мм), полученное в результате опыта
	натуральное значение, А/дм^2	Нормализованное значение	натуральное значение, ч	Нормализованное значение	Нормализованное значение	
1	*75*	*1*	*6*	*1*	*1*	0,2747
2	*75*	*1*	*4*	*-1*	*-1*	0,1831
3	*50*	*-1*	*6*	*1*	*-1*	0,1831
4	*50*	*-1*	*4*	*-1*	*1*	0,1221

3 Определяем коэффициенты регрессии (параметры регрессионной модели) b_0, b_1, b_2, b_3 по формулам (3.4) и их значимость.

Для коэффициента b_0:

$$b_o = \frac{1}{N}\sum_{u=1}^{N} Y_u = \frac{0.275+0.183+0.183+0.122}{4} = 0.1908$$

Для коэффициента b_1:

$$b_1 = \frac{1}{4}\sum_{u=1}^{4}(x_1 Y)_u = \frac{(+1)\cdot y_1 + (+1)\cdot y_2 + (-1)\cdot y_3 + (-1)\cdot y_4}{4} =$$

$$= \frac{(+1)\cdot 0.275 + (+1)\cdot 0.183 + (-1)\cdot 0.183 + (+1)\cdot 0.122}{4} = 0.0382.$$

Аналогично определяем значения для b_2, b_{12}: $b_2 = 0{,}0382$; $b_{12} = 0{,}0076$.

Статистическая оценка значимости коэффициентов регрессии предназначена для исключения второстепенных факторов, не влияющих на функцию отклика.

Для оценки коэффициентов регрессии сначала определяют *дисперсию воспроизводимости* эксперимента, которая характеризует разброс значений отклика при проведении параллельных (дублирующих) опытов:

$$S_y^2 = \frac{1}{N}\sum_{u=1}^{N} S_u^2,$$

где N - количество основных опытов (серий); S_u^2 - дисперсия, полученная при проведении параллельных опытов в каждой серии:

$$S_u^2 = \frac{\sum_{k=1}^{m}\left(y_{uk} - \overline{y_u}\right)^2}{m-1}, (k = 1,2,...m);$$

где y_{uk} - значение отклика в k-м опыте; $\overline{y_u}$ - среднее значение отклика в u-ой серии; количество параллельных опытов в серии (m=3).

В каждом проводимом основном опыте (серии) опыт повторялся 3 раза, а затем вычислялась дисперсия воспроизводимости. По данным эксперимента дисперсия воспроизводимости равна

$$S_y^2 = 0{,}0001.$$

Далее находим дисперсию оценок коэффициентов регрессии

$$S_b^2 = \frac{S_y^2}{N\cdot m},$$

где N- количество опытов, m – количество параллельных опытов (m=3).

Для проводимого эксперимента дисперсия оценок коэффициентов регрессии

$$S_b^2 = \frac{S_y^2}{N\cdot m} = \frac{0{,}0001}{4\cdot 3} = 0{,}0000083.$$

Доверительный интервал рассеяния коэффициентов уравнения регрессии определяют

$$\delta = S_b \cdot t(\alpha, k_2),$$

где $t(\alpha, k_2)$ - критерий Стьюдента, его принимают исходя из заданного уровня значимости α=0,05 (т.е. вероятность ошибки 5%) и числу степеней свободы $k_2 = N\cdot(m-1)$.

Для проводимого эксперимента $k_2 = 4\cdot(3-1) = 8$, а критерий Стьюдента $t(0{,}05;8) = 2{,}306$.

Доверительный интервал рассеяния коэффициентов уравнения регрессии

$$\delta = t(\alpha, k_2) \cdot S_b = t(\alpha, k_2) \cdot \sqrt{S_b^2} = 2{,}306 \cdot \sqrt{0.0000083} = 0{,}0067.$$

Если $\left|b_j\right| \geq \delta$, то коэффициент регрессии значим. Если $\left|b_j\right| < \delta$, то коэффициент регрессии не значим, т.е. $b_j = 0$.

По результатам эксперимент все коэффициенты значимы:

$$\left|0.1908\right| > 0{,}0067$$

$$\left|0.0382\right| > 0{,}0067$$

$$\left|0.0382\right| > 0{,}0067$$

$$\left|0.0076\right| > 0{,}0067.$$

Нормализованная регрессионная модель имеет вид

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_{12} \cdot x_{12} = 0.1908 - 0.0382 \cdot x_1 + 0.0382 \cdot x_2 + 0.0076 \cdot x_{12}.$$

4 Проверяем полученную регрессионную модель на адекватность для того, чтобы определить насколько расчетные значения соответствуют эмпирическим и можно ли использовать полученную регрессионную модель для расчета и прогнозирования значений отклика.

Находим опытное значение критерия Фишера $F_{ОП}$, если $F_{ОП} < F_{ТЕОР}$, то модель адекватна, и наоборот. Критерий Фишера вычисляетя [1]:

$$F_{ОП} = \frac{S_{yНА}^2}{S_y^2},$$

где $S_{yНА}^2$ - дисперсия неадекватности:

$$S_{yНА}^2 = \frac{m \sum_{u=1}^{N} \left(y_{uРАСЧ} - \overline{y_{uОПЫТ}} \right)^2}{N - n},$$

где $y_{uРАСЧ}$ - значение отклика, рассчитанное по регрессионной модели для u-й комбинации факторов, $\overline{y_{uОПЫТ}}$ - среднее значение отклика, полученное в u-м опыте, n - количество факторов.

Теоретическое значение критерия Фишера $F_{ТЕОР}$ определяется по справочным таблицам [1, 2] в зависимости от уровня значимости α и числа степеней свободы $k_1 = N - n$ и k_2.

Для расчета критерия Фишера сведем в табл. 3.8 значения отклика, полученные по регрессионной модели и полученные в ходе эксперимента. Из данной таблицы видно, что значения совпадают, поэтому считаем, что модель адекватна и расчет критерия Фишера не выполняем.

Таблица 3.8 – Сравнение значений экспериментальных и расчетных значений отклика

№ опыта, u	Значение отклика $\overline{y_{uОПЫТ}}$ (толщина покрытия h, мм), полученное в результате опыта	Значение отклика $y_{uРАСЧ}$ (толщина покрытия h, мм), рассчитанное по регрессионной модели
1	0,2747	0,2747
2	0,1831	0,1831
3	0,1831	0,1831
4	0,1221	0,1221

5 Переходим к натуральной модели и выполняем прогнозироание значения отклика для заданных значений факторов, которые лежат в области экспериментироавния.

Для перехода к натуральной модели в нормализованной модели нормализованные факторы заменяем на натуральные с помощью формул (3.1):

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_{12} \cdot x_{12} =$$

$$= b_0 + b_1 \cdot \frac{X_1 - X_{CP1}}{\Delta X_1} + b_2 \cdot \frac{X_2 - X_{CP2}}{\Delta X_2} + b_{12} \cdot \frac{X_1 - X_{CP1}}{\Delta X_1} \cdot \frac{X_2 - X_{CP2}}{\Delta X_2}$$

Для примера натуральная модель имеет вид:

$$Y = 0.1908 - 0.0382 \cdot x_1 + 0.0382 \cdot x_2 + 0.0076 \cdot x_{12} =$$

$$= 0.1908 - 0.0382 \cdot \frac{X_1 - 62.5}{12.5} + 0.0382 \cdot \frac{X_2 - 5}{1} + 0.0076 \cdot \frac{X_1 - 62.5}{12.5} \cdot \frac{X_2 - 5}{1}.$$

Выполняем прогнозирование значения отклика Y (толщина покрытия h, мм) для заданных значений факторов X_1, X_2 в соответствии с исходными данными.

Выполняем расчет с использованием натуральной модели

$$Y = 0.1908 - 0.0382 \cdot \frac{X_1 - 62.5}{12.5} + 0.0382 \cdot \frac{X_2 - 5}{1} + 0.0076 \cdot \frac{X_1 - 62.5}{12.5} \cdot \frac{X_2 - 5}{1} =$$

$$= 0.1908 - 0.0382 \cdot \frac{65 - 62.5}{12.5} + 0.0382 \cdot \frac{6.5 - 5}{1} + 0.0076 \cdot \frac{65 - 62.5}{12.5} \cdot \frac{6.5 - 5}{1} =$$

$$= 0.2579 мм$$

Содержание отчета по лабораторной работе

Тема, цель, исходные данные, распечатка отчета по лабораторной работе (файл **Лаб_раб_ОНИиИД1.xls,** письменные ответы на контрольные вопросы.

Контрольные вопросы

1. Дайте определение понятий: эксперимент, опыт, фактор, отклик, уровень фактора, план эксперимента, активный эксперимент, полный факторный эксперимент, матрица плана эксперимента.
2. Что такое подконтрольная эксплуатация автомобилей? К какому виду экспериментов она относится? Укажите его достоинства и недостатки.
3. Что определяется в ходе эксперимента?
4. Назовите основные требования при планировании эксперимента?
5. В чем заключается нормализация факторов при планировании эксперимента? С какой целью она проводится?
6. Запишите расширенную матрицу полного факторного эксперимента для двух факторов и регрессионную модель, параметры которой определяются по данной матрице.
7. Изобразите поверхность отклика для 2-х факторов.
8. Сколько и какие уровни факторов используются для определения параметров линейной модели? Назовите нормализованные значения уровней факторов?

Список используемых источников

1. Коваленко, Н. А. Научные исследования и решение инженерных задач в сфере автомобильного транспорта: учеб. пособие / Н. А. Коваленко. – Минск: Новое знание; М.: ИНФА-М, 2011 – 271 с.: ил.
2. Научные исследования и решение инженерных задач: Учебн. пособие/ С. С. Кучур, М. М. Болбас, В. К. Ярошевич. – Мн.: Адукацыя і выхаванне, 2003.
3. Кане, М. М. Предварительный анализ экспериментальных данных. Оценка правомерности применения корреляционно-регрессионного анализа в заданных условиях : пособие для студентов специальностей 1-36 80 02 «Инновационные технологии в машиностроении» и 1-53 80 01 «Автоматизация» / М. М. Кане ; Белорусский национальный технический университет, Кафедра «Технология машиностроения». – Минск : БНТУ, 2020. – 49 с. – Рекомендовано УМО РБ. – ISBN 978-985-583-538-8. URL: https://rep.bntu.by/handle/data/813512.
4. Кане, М. М. Анализ исходных данных при статистической обработке результатов научных исследований [Текст] : учебное пособие / М. М. Кане. - Минск : Вышэйшая школа, 2024. - 118, [2] с. - Библиография: с. 116 (21 назв.). - Допущено Министерством образования Республики Беларусь в качестве учебного пособия для студентов учреждений высшего образования по специальности магистратуры "Инновационные технологии в машиностроении". - 200 экз. - ISBN 978-985-06-3600-3.

РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

Вопросы к зачету

1. Основные понятия и определения теории планирования эксперимента
2. Основные принципы планирования эксперимента
3. Однофакторный эксперимент
4. Задачи выбора условий выполнения однофакторного эксперимента
5. Особенности планирования многофакторных экспериментов
6. Требования к планам экспериментов
7. Задачи выбора условий выполнения многофакторных экспериментов
8. Развитие методов планирования многофакторных экспериментов
9. Свойства и критерии планов эксперимента
10. Содержание и задачи эксперимента
11. Выбор средств измерений
12. Выбор числа наблюдений. Запись результатов
13. Выбор интервалов между экспериментальными уровнями факторов и порядка проведения опытов
14. Методы рандомизации экспериментов
15. Требования к параметру оптимизации
16. Минимизация числа параметров оптимизации с помощью корреляционного анализа
17. Расчет обобщенного параметра оптимизации
18. Экспертная оценка роли рассматриваемых факторов (параметров оптимизации) в обеспечении качества объекта (процесса)
19. Методы выбора основных факторов, влияющих на изучаемый процесс (объект)
20. Отбор наиболее существенных факторов с помощью дисперсионного анализа
21. Методы планирования отсеивающих экспериментов
22. Предварительное определение области изменения каждого фактора, внутри которой функция отклика принимает экстремальное значение
23. Оценка результатов, содержащих резко выделяющиеся опытные данные
24. Основные понятия и задачи корреляционно-регрессионного анализа
25. Предпосылки применения корреляционно-регрессионного анализа, методы проверки их соблюдения
26. Методы подбора эмпирических формул по результатам однофакторного эксперимента
27. Оценка точности прогнозирования с помощью однофакторной регрессионной модели
28. Полный факторный эксперимент типа $\mathbf{2^k}$
29. Расчет параметров математической модели по результатам полного факторного эксперимента
30. Порядок составления плана и обработки результатов полного факторного эксперимента
31. Оценка значимости коэффициентов регрессии

32.Оценка адекватности уравнения регрессии

ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

Учреждение образования
«Брестский государственный технический университет»

УТВЕРЖДАЮ
Проректор по учебной работе
_______________М.В. Нерода
. .2024

Регистрационный № /уч.

Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности

7-06-0714-02 Инновационные технологии в машиностроении

Профилизация: Аддитивные и субтрактивные технологии

2024

Учебная программа составлена на основе образовательного стандарта углубленного высшего образования ОСВО 7-06-0714-02-2023 Инновационные технологии в машиностроении и учебных планов специальности.

СОСТАВИТЕЛЬ:

Монтик С. В., заведующий кафедрой машиностроения и эксплуатации автомобилей, кандидат технических наук, доцент.

РЕЦЕНЗЕНТЫ:

Концевич П.С., заместитель директор ООО «ДжиЭсДжи Групп»;

Голуб В.М., заведующий кафедрой машиноведения учреждения образования «Брестский государственный технический университет», кандидат технических наук, доцент.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ*:*

Кафедрой машиностроения и эксплуатации автомобилей
Заведующий кафедрой С.В. Монтик
(протокол № 9 от 10.05.2024);

Методической комиссией машиностроительного факультета
Председатель методической комиссии В.П. Горбунов
(протокол №______ от _________2024);

Научно-методическим советом БрГТУ (протокол №_____ от _________2024)

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Место учебной дисциплины

Учебная дисциплина «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» относится к компоненту учреждения высшего образования и изучается на 2 курсе в 3 семестре магистрантами дневной формы получения образования и на 2 курсе в 3 семестре магистрантами заочной формы получения образования.

Целью преподавания учебной дисциплины является формирование устойчивой системы знаний, навыков и профессиональных компетенций в области анализа и упорядочения исходных данных при статистической обработке результатов научных исследований.

Основными задачами изучения дисциплины являются освоение:

- математических основ анализа погрешности измерений;
- методов разработки вероятностных математических моделей;
- методов проведения дисперсионного, корреляционного, регрессионного анализа;
- методов планирование эксперимента при одно- и многофакторном регрессионном анализе, а также в задачах оптимизации;
- программного обеспечения для выполнения статистического анализа данных.

В результате изучения учебной дисциплины «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» формируются следующие универсальные и специализированные компетенции:

УК-1: Применять методы научного познания в исследовательской деятельности, генерировать и реализовывать инновационные идеи.

СК-5: Владеть методами подготовки, упорядочения и анализа исходных данных в процессе статистических исследований при разработке нового и модернизации существующего оборудования, оснастки и технологических процессов механосборочного производства, владеть методами организации инновационной деятельности.

В результате изучения учебной дисциплины «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» магистрант должен:

знать:

- математические методы обработки погрешностей измерений, статистической проверки гипотез;
- законы распределения случайных величин, а также методы разработки вероятностных математических моделей;
- методы проведения дисперсионного, корреляционного, регрессионного анализа;

- научное планирование эксперимента при одно- и многофакторном регрессионном анализе, а также в задачах оптимизации;
- программное обеспечение для выполнения статистического анализа данных;

уметь:
- обрабатывать погрешности измерений, выполнять статистическую проверку гипотез;
- разрабатывать вероятностные математические модели;
- проводить дисперсионный, корреляционный, регрессионный анализы;
- осуществлять планирование эксперимента при одно- и многофакторном регрессионном анализе, а также в задачах оптимизации;
- применять современное программное обеспечение для статистического анализа данных.

владеть:
- методами подготовки, упорядочения и анализа исходных данных в процессе статистических исследований при разработке нового и модернизации существующего оборудования, оснастки и технологических процессов механосборочного производства.

Связь с другими учебными дисциплинами.

Дисциплина базируется на знаниях, полученных при изучении таких дисциплин, как: «Методы оптимизации технических объектов и процессов», «Научные основы технологии машиностроения», «Основы информационных технологий». Знания и умения, полученные студентами при изучении данной учебной дисциплины, необходимы для освоения дисциплин «Теоретические принципы исследований и испытаний рабочих машин», «Инновационные технологии механосборочного производства», подготовке магистерской диссертации.

План учебной дисциплины для дневной формы получения углубленного высшего образования

Код специальности (направления специальности)	Наименование специальности (направления специальности)	Курс	Семестр	Всего учебных часов	Количество зачетных единиц	Аудиторных часов (в соответствии с учебным планом УВО)					Академических часов на курсовую работу	Форма текущей аттестации
						Всего	Лекции	Лабораторные занятия	Практические занятия	Семинары		
7-06-0714-02	Инновационные технологии в машиностроении	2	3	100	3	36	30	-	6	-	-	Зачет

План учебной дисциплины для заочной формы
получения углубленного высшего образования

Код специальности (направления специальности)	Наименование специальности (направления специальности)	Курс	Семестр	Всего учебных часов	Количество зачетных единиц	Аудиторных часов (в соответствии с учебным планом УВО)					Академических часов на курсовую работу	Форма текущей аттестации
						Всего	Лекции	Лабораторные занятия	Практические занятия	Семинары		
7-06-0714-02	Инновационные технологии в машиностроении	2	3	100	3	8	4	-	4	-	-	Зачет

1 СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

1.1 ЛЕКЦИОННЫЕ ЗАНЯТИЯ, ИХ СОДЕРЖАНИЕ

Введение.

Цели, задачи и содержание учебной дисциплины, ее связь с другими дисциплинами учебного плана.

Тема 1 Погрешности измерений

Виды погрешностей измерений. Характеристики средств измерения. Уменьшение систематических погрешностей средств измерений. Расчет погрешностей средств измерений. Описание случайных погрешностей. Числовая оценка случайных погрешностей. Объединение результатов неравноточных измерений. Погрешности косвенных измерений. Обнаружение грубых погрешностей.

Тема 2 Оценка отклика, статистическая проверка гипотез

Оценка отклика. Статистическая проверка гипотез: проверка гипотезы о равенстве математического ожидания заданному значению; проверка гипотезы о равенстве средних значений; проверка гипотезы о равенстве дисперсий; проверка случайности и независимости результатов измерений в выборке.

Тема 3 Законы распределения случайных величин

Непрерывные распределения. Нормальное распределение. Равномерное распределение. Логарифмически нормальное распределение. Экспоненциальное распределение. Распределение Вейбулла. Гамма-распределение. Бета-распределение. Распределение χ^2 (распределение Пирсона). Распределение Стьюдента (t-распределение). Распределение Фишера (F-распределение). Усеченное нормальное распределение. Распределение Рэлея. Распределение Максвелла.

Дискретные распределения: биномиальное распределение (распределение Бернулли), распределение Пуассона.

Тема 4 Разработка вероятностных математических моделей

Определение вероятностной математической модели. Физические закономерности процессов формирования вероятностных распределений. Понятие о доверительной вероятности и уровне значимости. Построение интервального ряда экспериментального распределения. Расчет среднего значения и доверительного интервала. Расчет показателей вариации экспериментального распределения. Определение минимального количества измерений. Проверка экспериментальных данных на воспроизводимость результатов. Расчет эмпирических интегральной и дифференциальной функций распределения. Физический смысл интегральной и дифференциальной функций распределения. Особенности статистической обработки результатов незавершенных испытаний (цензурированных выборок). Проверка адекватности вероятностной математической модели результатам эксперимента.

Тема 5 Дисперсионный анализ

Однофакторный дисперсионный анализ. Двухфакторный дисперсионный анализ.

Тема 6 Корреляционный анализ

Парная корреляция. Многомерный корреляционный анализ. Корреляционные уравнения

Тема 7 Разработка регрессионных математических моделей

Парные и многофакторные регрессионные модели. Линейная и нелинейная зависимости. Методы подбора эмпирических формул. Вычисление коэффициентов линейной регрессионной модели, статистическая оценка значимости коэффициентов регрессионной модели. Оценка адекватности регрессионных моделей. Оценка точности прогнозирования на основании регрессионной математической модели.

Тема 8 Предпланирование эксперимента

Основные понятия и процедуры. Методы уменьшения числа параметров оптимизации: использование корреляционного анализа, функции желательности. Методы уменьшения размерности факторного пространства: априорное ранжирование факторов, критериальные математические модели, др. методы

Тема 9 Планирование эксперимента.

Основные принципы и понятия. Планирование эксперимента при оценке отклика. Планирование эксперимента при дисперсионном анализе.

Планирование при однофакторном регрессионном анализе: определение вида линейной по параметрам однофакторной модели; определение вида нелинейной по параметрам однофакторной модели; уточнение параметров однофакторной модели.

Планирование при многофакторном регрессионном анализе: определение вида линейной по параметрам многофакторной модели (последовательное планирование); определение вида нелинейной по параметрам многофакторной модели; уточнение параметров линейной по параметрам многофакторной модели.

Тема 10 Планирование эксперимента в задачах оптимизации

Основные понятия и процедуры. Планирование однофакторного поиска: метод дихотомии (половинного деления); метод Фибоначчи; метод «золотого сечения»; поиск по дискретным точкам.

Градиентные методы многофакторного поиска: планирование и анализ результатов эксперимента в исходной точке; поиск области оптимума; планирование и анализ результатов эксперимента в области оптимума.

Последовательное симплекс-планирование.

Методы случайного поиска.

Оптимизация при наличии ограничений: основные понятия; оптимизация на базе математических моделей; методы прямого поиска при оптимизации с ограничениями.

Тема 11 Выполнение анализа и статистической обработки результатов эксперимента с использованием современного программного обеспечения

Статистический анализ данных с использованием пакетов STATISTICA и EXCEL, другого программного обеспечения. Методология и технологии современного анализа данных с использованием нейронных сетей на примере STATISTICA Neural Networks.

1.2 ПРАКТИЧЕСКИЕ ЗАНЯТИЯ, ИХ СОДЕРЖАНИЕ

1. Проведение корреляционного анализа результатов эксперимента
2. Разработка регрессионной математической модели по результатам эксперимента
3. Планирование эксперимента. Определение вида линейной по параметрам многофакторной модели (последовательное планирование)

3. УЧЕБНО-МЕТОДИЧЕСКИЕ КАРТЫ УЧЕБНОЙ ДИСЦИПЛИНЫ

3.1 УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ для дневной формы получения углубленного высшего образования

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов				Количество часов самост. работы	Форма контроля знаний
		Лекции	Лабораторные занятия	Практические занятия	Семинарские занятия		
	3-й семестр						
	Введение.	0,5				2	опрос
1	Погрешности измерений	1,5				4	опрос
2	Оценка отклика, статистическая проверка гипотез	2				6	опрос
3	Законы распределения случайных величин	2				6	опрос
4	Разработка вероятностных математических моделей	4				6	опрос
5	Дисперсионный анализ	2				6	опрос
6	Корреляционный анализ	2		2		6	опрос
7	Разработка регрессионных математических моделей	4		2		6	опрос
8	Предпланирование эксперимента	2				4	опрос
9	Планирование эксперимента	4		2		6	опрос
10	Планирование эксперимента в задачах оптимизации	4				6	опрос
11	Выполнение анализа и статистической обработки результатов эксперимента с использованием современного программного обеспечения	2				6	опрос
	Всего часов по дисциплине	30		6		64	

3.2 УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ для заочной формы получения углубленного высшего образования

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов				Количество часов самост. работы	Форма контроля знаний
		Лекции	Лабораторные занятия	Практические занятия	Семинарские занятия		
	3-й семестр						
	Введение.					2	опрос

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов				Количество часов самост. работы	Форма контроля знаний
		Лекции	Лабораторные занятия	Практические занятия	Семинарские занятия		
1	Погрешности измерений					6	опрос
2	Оценка отклика, статистическая проверка гипотез					8	опрос
3	Законы распределения случайных величин					8	опрос
4	Разработка вероятностных математических моделей					8	опрос
5	Дисперсионный анализ					8	опрос
6	Корреляционный анализ					8	опрос
7	Разработка регрессионных математических моделей	2		4		8	опрос
8	Предпланирование эксперимента					8	опрос
9	Планирование эксперимента	2				8	опрос
10	Планирование эксперимента в задачах оптимизации	2				8	опрос
11	Выполнение анализа и статистической обработки результатов эксперимента с использованием современного программного обеспечения	2				8	опрос
	Всего часов по дисциплине	8		4		88	

4. ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

4.1 Перечень литературы (учебной, учебно-методической, научной, нормативной, др.)

Основная

1. Кане, М. М. Предварительный анализ экспериментальных данных. Оценка правомерности применения корреляционно-регрессионного анализа в заданных условиях : пособие для студентов специальностей 1-36 80 02 «Инновационные технологии в машиностроении» и 1-53 80 01 «Автоматизация» / М. М. Кане ; Белорусский национальный технический университет, Кафедра «Технология машиностроения». – Минск : БНТУ, 2020. – 49 с. – Рекомендовано УМО РБ. – ISBN 978-985-583-538-8. URL: https://rep.bntu.by/handle/data/813512.

2. Кане, М. М. Основы исследований, изобретательства и инновационной деятельности в машиностроении [Текст] : учебник / М. М. Кане. - Минск : Вышэйшая школа, 2018. - 366 с. : ил. - Библиогр.: с. 360–361. - ISBN 978-985-06-2829-9. URL: https://vshph.com/upload/inf/978-985-06-2829-9.pdf

Дополнительная

3. Ящерицын П. И., Махаринский Е. И. Планирование эксперимента в машиностроении: [Справ, пособие].— Мн.: Выш. шк., 1985.— 286 с.

4. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL: учебное пособие. – 2-е изд., испр. и доп. – М.: ФОРУМ, 2008. – 464 с.

5. Халяфян А.А. STATISTICA 6. Статистический анализ данных. 3-е изд. Учебник – М.: ООО «Бином-Пресс», 2007. – 512 с.

6. Нейронные сети. STATISTICA Neural Networks: Методология и технологии современного анализа данных / Под редакций В.П. Боровичкова. -2-е изд., перераб. и доп. – М.: Горячая линия – Телеком, 2008. – 392 с.

4.2. Перечень компьютерных программ, наглядных и других пособий, методических указаний и материалов, технических средств обучения, оборудования для выполнения практических работ

Электронные средства обучения

1. Набор электронных презентаций к лекциям по дисциплине «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований».

2. Пакет прикладных программ для выполнения практических работ по дисциплине «Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований» на базе табличного процессора MS Excel.

4.3. Перечень средств диагностики результатов учебной деятельности

Средства диагностики результатов учебной деятельности включают:

1. Текущая аттестация
2. Письменные отчеты по аудиторным практическим занятиям.
3. Письменный зачет.

Текущая аттестация по учебной дисциплине проводится в виде:

- контрольный опрос в устной форме.

В семестре предусмотрена одна текущая аттестация.

Обучающиеся допускаются к промежуточной аттестации по учебной дисциплине при условии успешного прохождения текущей аттестации.

Результаты текущей аттестации учитываются при проведении промежуточной аттестации по учебной дисциплине.

Промежуточная аттестация по учебной дисциплине проводится в форме письменного зачета.

4.4. Методические рекомендации по организации и выполнению самостоятельной работы обучающихся учебной дисциплине

При изучении дисциплины используются следующие формы самостоятельной работы в соответствии с п. 3 Положения о самостоятельной работе студентов учреждения образования «Брестский государственный технический университет», утвержденного ректором БрГТУ №56 от 01.06.2020:

- самостоятельное изучение тем лекционного курса по литературным источникам, имеющимся в библиотеке БрГТУ и указанным в разделе 4 настоящей программы;
- самостоятельная подготовка к выполнению практических работ по литературным источникам, имеющимся в библиотеке БрГТУ и указанным в разделе 4 настоящей программы;
- самостоятельная работа под контролем преподавателя во время практических занятий по расписанию по индивидуальным заданиям;
- самостоятельная подготовка к зачету.

Самостоятельное изучение тем лекционного курса выполняется по следующим литературным источникам:

Номе темы	Название раздела, темы	Номер литературного источника
	Введение.	[1, 2, 3]
1	Погрешности измерений	[1, 2, 3]
2	Оценка отклика, статистическая проверка гипотез	[1, 2, 3]
3	Законы распределения случайных величин	[1, 2, 3]
4	Разработка вероятностных математических моделей	[1, 2, 3]
5	Дисперсионный анализ	[1, 2, 3]
6	Корреляционный анализ	[1, 2, 3]
7	Разработка регрессионных математических моделей	[1, 2, 3]
8	Предпланирование эксперимента	[1, 2, 3]
9	Планирование эксперимента	[1, 2, 3]
10	Планирование эксперимента в задачах оптимизации	[1, 2, 3]
11	Выполнение анализа и статистической обработки результатов эксперимента с использованием современного программного обеспечения	[4, 5, 6]

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Инновационные технологии механосборочного производства	МЭА	нет	Рекомендовать к утверждению (протокол № № 9 от 10.05.2024) Зав. кафедрой МЭА С.В. Монтик ________________

Содержание учебной программы
согласовано с выпускающей кафедрой

Заведующий выпускающей кафедрой,
кандидат технических наук, доцент С.В. Монтик

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ
по учебной дисциплине
«Анализ и упорядочение исходных данных при статистической обработке результатов научных исследований»
Рег. № УД-24-1-100/уч от 28.06.2024
на 2024/2025 учебный год

№ п/п	Дополнения и изменения	Основание
1	Дополнить перечень основной литературы учебным пособием: Кане, М. М. Анализ исходных данных при статистической обработке результатов научных исследований [Текст] : учебное пособие / М. М. Кане. - Минск : Вышэйшая школа, 2024. - 118, [2] с. - Библиография: с. 116 (21 назв.). - Допущено Министерством образования Республики Беларусь в качестве учебного пособия для студентов учреждений высшего образования по специальности магистратуры "Инновационные технологии в машиностроении". - 200 экз. - ISBN 978-985-06-3600-3.	Решение кафедры «Машиностроение и эксплуатация автомобилей» (протокол № 7 от 18.02.2025)

Учебная программа пересмотрена и одобрена на заседании кафедры «Машиностроение и эксплуатация автомобилей» (протокол № 7 от 18.02.2025)

Заведующий кафедрой,
кандидат технических наук, доцент С.В.Монтик

УТВЕРЖДАЮ
Декан факультета
кандидат технических наук, доцент С.Р.Оныcько