Министерство образования и науки РФ

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра конструирования узлов и деталей РЭА (КУДР)

Методы планирования эксперимента

Методические указания к практическим занятиям и самостоятельной работе

Для аспирантов направления подготовки 03.06.01, 09.06.01, 10.06.01, 11.06.01, 12.06.01, 13.06.01,

Разработчик:
проф. каф. КУДР
________Еханин С.Г.

2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Цели и задачи дисциплины:

Целью при изучении дисциплины является проведение эксперимента в кратчайший срок с минимальными затратами материальных средств при высоком качестве полученных результатов.

Задачами преподавания дисциплины является освоение аспирантами современных математических методов анализа экспериментальных данных, типового программного обеспечения для обработки данных экспериментов.

В результате изучения дисциплины аспиранты должны

знать:

методы планирования эксперимента;

уметь: проводить экспериментальные исследования; выполнять анализ результатов исследования;

владеть: навыками проведения экспериментов, обработки, аппроксимации и анализа числовых результатов наблюдений, методологического анализа научного исследования.

Таким образом, чтобы обеспечить в дальнейшем высокое качество прохождения практических занятий и самостоятельной работы, необходимо уменьшить дублирование и увеличить долю основополагающего материала.

Задачами преподавания данной дисциплины является овладение аспирантами методами и приемами обработки результатов эксперимента, знаниями которые необходимы при разработке алгоритмов решения практических задач с применением средств вычислительной техники и прикладного программного обеспечения.

Основная цель большинства экспериментальных исследований состоит в нахождении такой совокупности входных управляемых переменных (факторов), при которых оптимизируемая целевая функция принимает экстремальное значение, достигается с помощью минимально возможного числа опытов при минимуме затрат времени и средств.

Знания, полученные при изучении теоретического курса, закрепляются при решении конкретных задач на практических занятиях, на которых аспиранты получают навыки по математической обработке фактического материала с применением средств вычислительной техники и компьютерных программ, планированию эксперимента, знакомятся с методологией работы.

Эти знания, несомненно, расширят кругозор аспирантов, помогут глубже понять проблемы стоящие перед современной наукой.

1. МЕТОДИКА ПРОВЕДЕНИЯ ПРАКТИЧЕСКИХ ЗАНЯТИЙ.

1.1 Обоснование тем практических занятий и методические указания к их проведению.

Математическое моделирование является одним из способов обработки информации для решения научно-технических и многих других задач [1]. Построение математических моделей обычно складывается из определения зависимостей входных и выходных переменных исследуемого объекта и определения параметров моделей (числовых значений коэффициентов), полученных на основе исследования реального объекта.

Объектом исследования называется некоторая система, преобразующая входную информацию (набор входных переменных - x) в выходную информацию (набор выходных переменных - y). Входные переменные могут изменяться под действием случайных факторов, или намеренно вручную или автоматически. В общем случае вид зависимости $y=f(x)$ неизвестен. Математическая модель, построенная на основе экспериментальных данных, должна точно описывать исследуемый объект. Основным достоинством этого метода является простота моделей и ясный физический смысл ее параметров, а также возможность составления такой модели на стадии проектирования, что дает возможность прогнозирования сложных ситуаций в дальнейшем.

I. Первая проблема, которую необходимо решить в рамках такого подхода – это **обработка экспериментальных данных**. Поэтому первые практические занятия по данной дисциплине посвящаются этой проблеме.

- **Предварительная обработка данных.** На этом этапе осуществляется устранение случайной помехи, действующей на объект. Обычно это осуществляется фильтрами (например, методом скользящего среднего и др.).

Обычно эксперименты проводятся таким образом, чтобы переменная – x менялась эквидистантно (например, через равные промежутки времени). Поэтому для увеличения информативности проведенного эксперимента необходимо проведение интерполяции и экстраполяции. *Интерполяция* – определение промежуточных значений функции между двумя полученными отсчетами. *Экстраполяция* – определение будущих значений функции с момента очередного значения до момента, находящегося за пределами полученной таблицы данных. Во многих случаях используется сплайн-интерполяция.

Сглаживанием (или фильтрацией) называется операция выделения полезного сигнала из его суммы со случайной помехой. На практике применяют несколько алгоритмов сглаживания. Наиболее часто применяют линейное сглаживание (по трем – пяти точкам), метод скользящего среднего, метод усреднения с помощью интерполяционных полиномов [2].

- **Математическая обработка данных и решение задачи идентификации** [1].

Математическая модель – это система математических соотношений, при помощи которых можно описать свойства объекта уравнением вида:

$$F(x, y, b) = 0, \qquad (1)$$

Где F – конкретный вид функции зависимости,

b - вектор параметров.

В зависимости от способа представления функции F выделяют два способа построения математических моделей **экспериментальный** (эмпирический) и **аналитический**.

Экспериментальный. На объекте проводится эксперимент с получением векторов входных и выходных параметров. Вид функции F задается по своему усмотрению, функция также как и вектор параметров b не имеет физического смысла. При задании эмпирической модели руководствуются простотой вычисления функции F. Ее задают в виде, где вектор параметров включается линейно. Этим требованиям удовлетворяют функции в виде полинома. Раздел математической статистики, занимающийся вопросами аппроксимации данных эмпирическими зависимостями, называется *регрессионным анализом*.

Аналитический способ. В этом случае вид функции F определяется физической сутью процесса. Получается система уравнений, чаще всего дифференциальных, которые сложно разрешить относительно переменных.

Задача идентификации – определение параметров математической модели по экспериментальным данным. Используя таблицу экспериментальных данных (x - вектор входных параметров, y – вектор выходных параметров), необходимо найти такие значения $b=\{b_j, j=0...p\}$ (p – порядок уравнения) модели объекта, чтобы вектор выходных параметров модели y' был максимально приближен к вектору исходных данных y. Задача идентификации заключается в минимизации некоторого критерия I, являющегося мерой близости этих двух векторов. В качестве критерия I может выступать квадратичный критерий, физический смысл которого состоит в расчете оценки дисперсии аппроксимации экспериментальных данных в выбранной модели F. Этот метод называется *методом наименьших квадратов (МНК).*

- **Оценка адекватности математической модели**. Модель считается построенной, когда определены числовые значения параметров и дан ответ на вопрос: соответствует ли модель данному объекту. Для оценки адекватности чаще всего пользуются величиной дисперсии аппроксимации.

$$S^2_{\text{àïï}} = I(b) = \frac{1}{N-p}\sum_{i=1}^{N}(y_i - y(x_i, b))^2, \qquad (2)$$

$$S_{\text{àïï}} = \sqrt{S^2_{\text{àïï}}} \text{ - среднеквадратическая погрешность.} \qquad (3)$$

Модель считается адекватной, если $S_{\text{апп}} < \Delta y$, где Δy – некоторая заданная экспериментатором точность.

II. Вторая проблема – анализ полученных зависимостей.

В любых экспериментах имеется несколько факторов, вызывающих изменчивость средних значений наблюдаемых случайных величин. Эти факторы могут принадлежать одному или нескольким источникам изменчивости.

Дисперсионный анализ дает возможность установить, существенное ли влияние оказывает тот или иной из рассматриваемых факторов на изменчивость признака, а также определить количественно «удельный вес» каждого из источников изменчивости в их общей совокупности.

Дисперсионному анализу подвергаются результаты экспериментов 3-ех видов:

1) эксперименты, в которых все факторы имеют фиксированные уровни;
2) эксперименты, в которых все факторы имеют случайные уровни;
3) эксперименты, в которых есть факторы, имеющие случайные уровни, а также, имеющие фиксированные уровни.

Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость. С этой целью производят разложение общей дисперсии σ^2 наблюдаемой совокупности, вызванной всеми источниками изменчивости, на составляющие, порожденные независимыми факторами. Каждая из составляющих дает оценку дисперсии σ^2_A, σ^2_B, …., вызванную конкретным источником изменчивости, в общей совокупности. Для проверки значимости этих составляющих оценок дисперсии их сравнивают с общей дисперсией в общей совокупности (по критерию Фишера).

Корреляционно-регрессионный анализ [2,3] дает возможность глубже понять сложный механизм причинно-следственных отношений между явлениями. Применяется для исследования интенсивности, вида и формы причинных влияний на исследуемые процессы. Можно выделить два типа связи между исследуемыми явлениями или процессами: функциональную и стохастическую.

Функциональная связь $y = f(x)$: для каждой независимой переменной x существует вполне определенное значение зависимой переменной y.

Стохастической (статистической) зависимостью называется зависимость, при которой изменение одной из величин влечет изменение распределения другой. *Статистическую* зависимость называют *корреляционной*, если при изменении значений одной величины меняется среднее значение другой.

При сравнении функциональных и статистических корреляционных зависимостей следует иметь в виду, что при функциональной зависимости, зная x можно вычислить величину y. При статистической корреляционной зависимости устанавливается лишь тенденция изменения y при изменении x. Сущность корреляционного анализа составляет изучение зависимости вариации признака от окружающих условий.

С корреляционным анализом тесно связан регрессионный. Их объединяют методы обработки данных, отличают цели и формы установления связи. В корреляционном анализе оценивается сила стохастической связи, а в регрессионном – форма.

Корреляционно-регрессионный анализ проводится в следующей последовательности.

1. Исходя из целей и задач исследования зависимости, устанавливаются результативные (y_i) и факторные (x_i) признаки.
2. По совокупности объектов определяются значения результативных и факторных признаков.
3. Обосновывается для случая парной зависимости (обычно графическим методом) модель уравнения регрессии.

4. Методом наименьших квадратов определяются параметры уравнения регрессии.
5. Определяется теснота связи между изучаемыми признаками.
6. Оценивается значимость уравнения связи, его параметров и показателей тесноты связи.

Анализ временных рядов [2,3]. Большое количество данных в технологических, экологических и др. процессах могут рассматриваться как в пространстве, так и во времени. Последовательность измерений значений переменной процесса за определенный период через одинаковые промежутки времени называется **дискретным временным рядом**:

$$Z_1, Z_2, Z_3, \ldots Z_n. \tag{4}$$

Эти последовательные наблюдения обычно зависимы. С детерминистской точки зрения ряд (4) можно представить:

$$Z_t = f(t) + \varepsilon_t, \tag{5}$$

где t = 1, 2, 3, …..n; f – гладкая (непрерывная и дифференцируемая) функция, характеризующая долгосрочное изменение в зависимости от времени – тренд; ε_t,- случайный ряд возмущений, наложенный на систематическую часть.

При наличии во временном ряду тенденции и циклических колебаний значения каждого последующего уровня зависят от предыдущих. **Корреляционная зависимость** [2-4] между последовательными уровнями временного ряда называют **автокорреляцией** уровней временного ряда.

Количественно ее можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого же ряда, сдвинутыми на один или несколько шагов во времени. Такой коэффициент называется **коэффициентом автокорреляции**. Коэффициент автокорреляции между исходным рядом и рядом, полученным путем смещения на одну единицу времени называется **коэффициентом автокорреляции первого порядка**. Аналогично можно определить коэффициенты автокорреляции более высоких порядков.

Последовательность коэффициентов автокорреляции уровней первого, второго и т.д. порядков называют **автокорреляционной функцией временного ряда**.

Анализ временных рядов преследует ряд целей. Обычно это: описание поведения ряда, построение модели для объяснения наблюдений, составление прогноза, исходя из предположения о сохранении тенденции развития в будущем. Для достижения поставленных целей используют модели, основанные на детерминистском, стохастическом, спектральном и других подходах.

В общем случае можно предположить в модели наличие следующих компонент:

1) тренд или долгосрочное колебание;
2) регулярное движение относительно тренда;
3) остаток.

Модель, в которой временной ряд представлен как сумма перечисленных выше компонент, называется **аддитивной моделью временного ряда**.

Тренд дает возможность прогнозировать основную тенденцию изменения явления во времени. Значения тренда обычно сопровождаются ошибками. Доверительный интервал для тренда определяется как

$$\hat{Z}_t \pm t_{st} \cdot S_z \,, \tag{6}$$

где Z_t – значение тренда в момент времени t, t_{st} – квантиль распределения Стьюдента для двусторонней критической области при уровне значимости α с $\nu = n\text{-}d$ степенями свободы (n – число наблюдений, d – число оцениваемых параметров, например, для уравнения прямой d = 2); S_z- среднеквадратическое отклонение членов ряда от тренда.

Стационарные временные ряды [2,3]. Временной ряд, не имеющий тренда (либо с исключенным трендом), называется стационарным.

Для выявления **циклических компонент** временного ряда применяется спектральный анализ на основе дискретного преобразования Фурье:

$$Z_t = a_0 + \sum_{j=1}^{k} (a_j \cdot cos\ jt + b_j \cdot sin\ jt) . \qquad (7)$$

Параметры a_j и b_j находятся с помощью МНК, в результате применения которого, получим:

$$a_0 = \frac{1}{n}\sum_{t=1}^{n} Z_t , \qquad a_j = \frac{2}{n}\sum_{t=1}^{n} Z_t \cdot cos\ jt , \qquad b_j = \frac{2}{n}\sum_{t=1}^{n} Z_t \cdot sin\ jt. \qquad (8)$$

После выделения тренда и циклических компонент производится **анализ остаточного ряда.** При хорошей адекватности моделей остаточный ряд будет обусловлен лишь стохастической компонентой. Поэтому анализ остаточного ряда посвящен **проверке статистических гипотез**. Процесс проверки гипотезы называется *статистическим доказательством*. Основную выдвигаемую гипотезу называют нулевой. Наряду с нулевой рассматривают альтернативную ей.

При проверке гипотезы о нормальном распределении остаточного ряда сравниваются эмпирические и теоретические (вычисленные в предположении нормальности распределения) частоты. Для этого используется статистика ХИ–квадрат Пирсона со степенями свободы $\nu = k\text{-}r\text{-}1$ (k – число интервалов, r – число оцениваемых параметров). Если ХИ-квадрат расчетный меньше или равен критической величине из распределения Пирсона, то основная гипотеза принимается. А это значит, что ранее построенные модели (тренда и циклических компонент) адекватны.

Теория планирования эксперимента [5].

Эксперимент является основным и наиболее совершенным методом познания. Основной вид эксперимента – активный, проводится в контролируемых и управляемых условиях.

Проведение активного эксперимента зачастую требует больших материальных затрат. Поэтому важной задачей является получение необходимых сведений при минимальном числе опытов.

Решением этой проблемы занимается теория планирования эксперимента (ТПА), представляющая собой раздел математической статистики.

В общем случае ТПА позволяет ответить на вопросы:

- как спланировать эксперимент, обеспечивающий при требуемой точности результатов, минимальные затраты времени и средств;

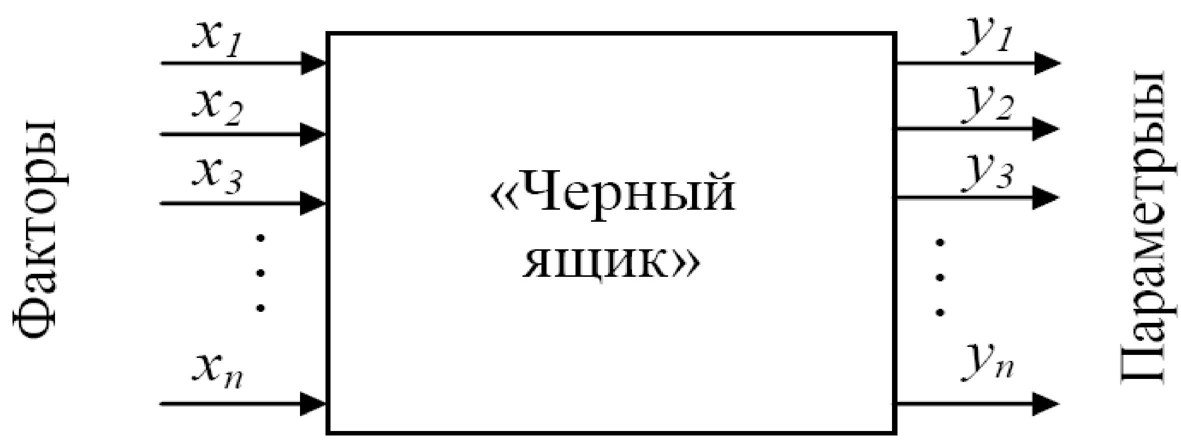
- как обработать результаты, чтобы извлечь из них максимум информации об исследуемом объекте;

- какие выводы можно сделать по результатам эксперимента и какова достоверность этих выводов.

Активный эксперимент в сочетании с методами планирования позволяет получить требуемые результаты, затратив минимальные средства и время на проведение исследования.

При планировании эксперимента исследуемый объект представляется «черным ящиком», на который воздействуют факторы xi.

Факторы должны быть совместимыми и независимыми. Совместимость предполагает допустимость любой комбинации факторов, а независимость – отсутствие между факторами корреляционной связи



Исследуемые параметры. К исследуемым параметрам также предъявляют ряд требований. Они должны быть:

- эффективными, то есть способствовать скорейшему достижению цели;

- универсальными – быть характерными не только для исследуемого объекта;

- статистически однородными, то есть определенному набору значений факторов xi с точностью до погрешности эксперимента должно соответствовать определенное значение фактора yi;

- выражаться количественно одним числом;

- легко вычисляться и иметь физический смысл;

- существовать при любом состоянии объекта.

Множество точек факторного пространства, в которых проводится эксперимент, представляется с помощью плана эксперимента.

$$x = \left\| \begin{matrix} x_1(1) & x_2(1) & \ldots & x_n(1) \\ x_1(2) & x_2(2) & \ldots & x_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(N) & x_2(N) & \cdots & x_n(N) \end{matrix} \right\| - \left\| \begin{matrix} x(1) \\ x(2) \\ \vdots \\ x(N) \end{matrix} \right\|,$$

где n – число факторов; N – число точек факторного пространства. Точка x^0 называется центром плана. Если центр плана совпадает с началом координат, то план называется центральным.

$$x^{(0)} = \frac{1}{N} \sum_{j=1}^{N} x(j),$$

Условия проведения опытов могут свободно выбираться в пределах заданных границ. Выбор соответствующего плана эксперимента позволяет обеспечить ММ разные свойства.

Наиболее распространенными являются следующие критерии.

Критерий ортогональности – когда полученные оценки коэффициентов регрессии некоррелированы (не смешаны). Замена нулем любого коэффициента в ММ в этом случае не изменяет значений остальных коэффициентов.

Критерий рототабельности – когда дисперсия выходной переменной зависит только от расстояния от центра плана.

Критерий А-оптимальности требует выбора такого плана, при котором дисперсионная матрица имеет минимальный след (минимальную сумму диагональных элементов).

Критерий D-оптимальности требует минимизации определителя дисперсионной матрицы.

Критерий G-оптимальности требует достижение наименьшей величины максимальной дисперсии зависимой переменной.

Полный факторный эксперимент.

В полном факторном эксперименте (ПФЭ) исследуется один параметр и реализуются все возможные сочетания уровней факторов. Для каждого фактора выбираются два уровня – верхний и нижний, на которых фактор варьируется.

Половина разности между верхним и нижним уровнями называется интервалом варьирования. Интервал варьирования должен быть больше погрешности измерения уровня фактора (ограничение снизу), а верхний и нижний уровни фактора не должны выходить за область его определения (ограничение сверху). На практике интервал варьирования составляет обычно 3–10% от области определения. При двух уровнях для каждого из n факторов общее число опытов составляет 2n. ПФЭ – это эксперимент типа 2n. ПФЭ позволяет получить математическую модель исследуемого объекта в виде уравнения множественной регрессии.

В зависимости от объема информации в ММ включают не все, а лишь некоторые взаимодействия первого порядка, иногда – взаимодействия второго порядка и очень редко – взаимодействия выше третьего порядка.

Связано это с тем, что учет всех взаимодействий приводит к громоздким расчетам.

Для удобства расчетов масштаб факторов выбирают так, чтобы значение верхнего уровня было равно +1, а нижнего –1.

План ПФЭ изображают в виде таблицы, столбцы которой отражают уровни факторов, а строки – номера опытов. Эти таблицы называют матрицами планирования (МП) эксперимента. Поскольку значения уровней факторов по модулю всегда равны единице, то обычно в МП записывают только знак уровня

(т. е. «+» вместо «1» и «–» вместо «–1»).

Расчет оценок коэффициентов уравнения регрессии производится по методу наименьших квадратов, при этом минимизируется сумма квадратов отклонений между экспериментальными значениями исследуемого параметра и значениями, вычисленными для тех же точек факторного пространства по уравнению регрессии.

Благодаря предварительной стандартизации масштаба факторов и ортогональности МП, расчет оценок коэффициентов регрессии в ПФЭ превращается в простую арифметическую процедуру.

Гипотезу о статистической значимости (отличии от нуля) коэффициентов регрессии проверяют по критерию Стьюдента.

Для проверки гипотезы об адекватности ММ необходимо сравнить две

дисперсии: дисперсию, рассчитанную по ММ и по экспериментальным результатам.

Для записи ММ в реальных физических величинах производят обратный переход от стандартизированного масштаба к натуральному.

1.2 Предлагаемые темы практических занятий:

1. Аппроксимация экспериментальных данных и построение гистограмм с помощью электронных таблиц Excel .

2. Расчет числовых характеристик дискретных случайных величин с использованием специальных функций в MS Excel.
3. Нормальный закон распределение. Использование функций *Excel* для вычисления значений нормального распределения и построения диаграммы нормальной функции плотности вероятности и диаграммы нормальной функции распределения.
4. Анализ вариационных рядов. Построение и графическое изображение интервальных вариационных рядов. Гистограммы распределения. Числовые характеристики интервальных вариационных рядов. Расчет показателей вариации. Моменты вариационных рядов. Коэффициенты формы распределения.
5. Выборочный метод анализа экспериментальных результатов. Виды выборок. Состоятельность выборки. Расчет числовых характеристик выборки. Построение доверительного интервала. Определение величины предельной ошибки выборочной средней и величины необходимого объема выборки.
6. Проверка статистических гипотез. Алгоритм проверки статистических гипотез. Введение выборочных данных. Формирование нулевой и конкурирующих гипотез. Задание уровня значимости. Определение выборочной статистики наблюдений. Расчет значения критерия статистики (Пирсона, Стьюдента). Проверка нулевой и конкурирующей гипотезы.
7. Корреляционно-регрессионный анализ. Графическое представление экспериментальных данных. Подбор уравнений регрессии. Расчет параметров уравнения регрессии с помощью МНК. Оценка качества уравнения с помощью средней ошибки аппроксимации. Определение коэффициента эластичности. Оценка теснота связи между переменными с помощью показателей корреляции и детерминации. Оценка значимости коэффициента корреляции и регрессии по критерию Стьюдента при уровне значимости 0,05. Определение характеристики надежности ре-

зультатов регрессионного анализа с использованием критерия Фишера при уровне значимости 0,05. Определение прогнозного значения признака.

8. Однофакторный дисперсионный анализ. Постановка задачи. Алгоритм дисперсионного анализа. Эксперимент с равным числом наблюдений по уровням. Построение вспомогательной таблицы. Вычисление вспомогательных сумм, сумм квадратов, общей суммы квадратов. Расчет сумм квадратов между группами, сумм квадратов внутри групп. Проверка нулевой гипотезы.

9. Анализ временных рядов. Графическое представление экспериментального временного ряда. Расчет коэффициента автокорреляции первого порядка. Обоснование выбора типа уравнения тренда и расчет его параметров МНК. Проверка независимости значений уровней случайной компоненты (остаточной последовательности). Проверка соответствия распределения случайной компоненты нормальному закону распределения. Определение точности трендовой модели. Выводы по прогнозу.

10. Спектральный анализ временных рядов. Графическое представление экспериментального временного ряда. Сглаживание временного ряда методом скользящей средней. Приведение начального временного ряда к стационарному виду. Спектральный анализ на основе дискретного преобразования Фурье: определение коэффициентов разложения Фурье, построение модели аппроксимации. Проверка соответствия распределения случайной компоненты нормальному закону распределения. Определение точности аппроксимационной модели. Выводы по прогнозу.

11. Планирование активных экспериментов. Метод статистических испытаний (метод Монте-Карло). Генерация случайных чисел. Построение распределения. Вычисление определенных интегралов методом Монте-

Карло. Компьютерные эксперименты на основе реализации модели пьяного прохожего.

12. Полный факторный эксперимент. В соответствии с индивидуальным заданием необходимо перейти к стандартизированному масштабу факторов, составить матрицу планирования (МП) полного факторного эксперимента (ПФЭ) и проверить ее свойства, рандомизировать опыты. Провести ПФЭ. Проверить воспроизводимость опытов. Если дисперсии неоднородны, повторить эксперимент. Рассчитать оценки коэффициентов регрессионного уравнения. Проверить статистическую значимость коэффициентов регрессии. Проверить адекватность полученной математической модели (ММ). Перейти к исходным физическим переменным. Записать полученную ММ и сделать выводы.

1.3 Методические указания к практическому занятию (пример)

В качестве примера можно привести описание методики проведения практического **занятия, посвященного выборочному методу** анализа экспериментальных результатов.

Выборка — это группа элементов, выбранная для исследования из всей совокупности имеющихся экспериментальных результатов. Конечной **целью изучения выборочной совокупности** всегда является получение информации о генеральной совокупности. Поэтому естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, то есть была бы репрезентативной или представительной. В тех случаях, когда генеральная совокупность недостаточно известна, обычно не удается предложить лучшего способа получения представительной выборки, чем случайный выбор. Поэтому случайная выборка формируется случайным отбором — из генеральной совокупности наудачу извлекается по одному объекту.

Для построения **выборочной функции распределения** весь диапазон изменения случайной величины *X* разбивают на ряд интервалов одинаковой ширины. Число интервалов обычно выбирают не менее 5 и не более 15. Затем определяют число значений случайной величины *X*, попавших в каждый интервал. Поделив эти числа на общее количество наблюдений *n*, находят относительную частоту попадания случайной величины *X* в заданные интервалы. По найденным относительным частотам строят гистограммы выборочных функций распределения.

Замена теоретической функции распределения *F(x)* на ее выборочный аналог *Fn(x)* в определении математического ожидания, дисперсии, стандартного отклонения и т. п. приводят к выборочному среднему, выборочной дисперсии, выборочному стандартному отклонению и т. д. *Выборочные характеристики являются оценками соответствующих характеристик генеральной совокупности*. Эти оценки должны удовлетворять определенным требованиям. В соответствии с важнейшими требованиями, оценки должны быть:

- несмещенными, то есть стремиться к истинному значению характеристики генеральной совокупности при неограниченном увеличении количества испытаний;
- состоятельными, то есть с ростом размера выборки оценка должна стремиться к значению соответствующего параметра генеральной совокупности с вероятностью, приближающейся к 1;
- эффективными, то есть для выборок равного объема используемая оценка должна иметь минимальную дисперсию.

Выборочные характеристики [2-4]. Среди выборочных характеристик выделяют показатели, относящиеся к центру распределения (меры положения), показатели рассеяния вариант (меры рассеяния) и меры формы распределения. К показателям, характеризующим центр распределения, относят различные виды средних (арифметическое, геометрическое и т. п.), а также моду и медиану.

Простейшим показателем, характеризующим центр выборки, является мода.

Мода — это элемент выборки с наиболее часто встречающимся значением (наиболее вероятная величина).

Средним значением выборки, или выборочным аналогом математического ожидания, называется величина

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \qquad (9)$$

Иначе говоря, среднее значение — это центр выборки, вокруг которого группируются элементы выборки. При увеличении числа наблюдений среднее приближается к математическому ожиданию.

Выборочная **медиана** — это число, которое является серединой выборки, то есть половина чисел имеет значения большие, чем медиана, а половина чисел имеет значения меньшие, чем медиана. Для нахождения медианы обычно выборку ранжируют — располагают элементы в порядке возрастания. Если количество членов ранжированного ряда нечетное, медианой является значение ряда, которое расположено посередине, то есть элемент с номером $(n + 1)/2$. Если число членов ряда четное, то медиана равна среднему членов ряда с номерами n/2 и $n/2 + 1$.

Основными **показателями рассеяния** являются *размах вариации, дисперсия выборки, стандартное отклонение и стандартная ошибка*.

Вариационный размах — это разница между максимальным и минимальным значениями элементов выборки. Эта величина является простейшей и наименее надежной мерой вариации или рассеяния элементов в выборке.

Более точно отражают рассеяние показатели, учитывающие не только крайние, но и все значения элементов выборки. **Дисперсией выборки,** или выборочным аналогом дисперсии, называется величина

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,. \qquad (10)$$

Дисперсия выборки — это параметр, характеризующий степень разброса элементов выборки относительно среднего значения. Чем больше дисперсия, тем дальше отклоняются значения элементов выборки от среднего значения.

Выборочным стандартным отклонением (среднее квадратичное отклонение) называется величина

$$s = \sqrt{s^2} . \qquad (11)$$

Это параметр, также характеризующий степень разброса элементов выборки относительно среднего значения. Чем больше среднее квадратичное отклонение, тем дальше отклоняются значения элементов выборки от среднего значения.

Стандартная ошибка или **ошибка среднего** находится из выражения

$$m = \frac{s}{\sqrt{n}} . \qquad (12)$$

Стандартная ошибка — это параметр, характеризующий степень возможного отклонения среднего значения, полученного на исследуемой ограниченной выборке, от истинного среднего значения, полученного на всей совокупности элементов. С помощью стандартной ошибки задается так называемый **доверительный интервал**. 95%-ный доверительный интервал, равный $x \pm 2m$, обозначает диапазон, в который с вероятностью 0,95 (при достаточно большом числе наблюдений $n > 30$) попадает среднее генеральной совокупности.

Выборочной квантилью называется решение уравнения

$$F_n(x) = p. \qquad (13)$$

В частности, выборочная медиана есть решение уравнения

$$F_n(x) = 0,5. \qquad (14)$$

Показателями, характеризующими **форму распределения**, являются выборочные **эксцесс** и **асимметрия**.

Эксцесс — это степень выраженности «хвостов» распределения, то есть частоты появления удаленных от среднего значений.

Асимметрия — величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает значения от -1 до 1. В случае симметричного распределения асимметрия равна 0.

Часто значения асимметрии и эксцесса используют для проверки гипотезы о том, что данные (выборка) принадлежат к определенному теоретическому распределению, в частности, нормальному распределению. Для нормального распределения асимметрия равна нулю, а эксцесс — трем.

При определении вышеперечисленных выборочных характеристик используются **специальные функции *Excel*** и **пакет анализа данных** [2].

В мастере функций Excel имеется ряд **специальных функций**, предназначенных для вычисления выборочных характеристик. Прежде всего, это *функции, характеризующие центр распределения*.

- Функция *СРЗНАЧ* вычисляет среднее арифметическое из нескольких массивов (аргументов) чисел. Аргументы *число1*, *число2*, ... — это от 1 до 30 массивов, для которых вычисляется среднее.
- Функция *МЕДИАНА* позволяет получать медиану заданной выборки.
- Функция *МОДА* вычисляет наиболее часто встречающееся значение в выборке. Например, *МОДА (10;14;5;6;10;12;13)* равняется 10.

К специальным функциям, вычисляющим выборочные характеристики, характеризующие рассеяние элементов, относятся *ДИСП*, *СТАНД0ТКЛ0Н*, *ПЕРСЕНТИЛЬ*.

- Функция *ДИСП* позволяет оценить дисперсию по выборочным данным.
- Функция *СТАНДОТКЛОН* вычисляет стандартное отклонение.
- Функция *ПЕРСЕНТИЛЬ* позволяет получить квантили заданной выборки.

Форму эмпирического распределения позволяют оценить специальные функции *ЭКСЦЕСС и СКОС*.

- Функция *ЭКСЦЕСС* вычисляет оценку эксцесса по выборочным данным. Например, *ЭКСЦЕСС (10;14;5;6;10;12;13)* равняется -1,169.
- Функция *СКОС* позволяет оценить асимметрию выборочного распределения. Например, *СКОС (10;14;5;6;10;12;13)* равняется -0,527.

В пакете Excel помимо мастера функций имеется набор более мощных инструментов для работы с несколькими выборками и углубленного анализа данных, называемый **Пакет анализа** [2], который может быть использован для решения задач статистической обработки выборочных данных. Для использования статистического *пакета анализа данных* необходимо:

- указать курсором мыши на пункт меню *Сервис* и щелкнуть левой кнопкой мыши;
- в раскрывающемся списке выбрать команду *Анализ данных*;
- выбрать необходимую строку в появившемся списке *Инструменты анализа*;
- ввести входной и выходной диапазоны и выбрать необходимые параметры.

Нахождение основных выборочных характеристик. Для определения характеристик выборки используется процедура *Описательная статистика*. Процедура позволяет получить статистический отчет, содержащий информацию о центральной тенденции и изменчивости входных данных. Для выполнения процедуры необходимо:

- выполнить команду *Сервис* ► *Анализ данных*;
- в появившемся списке Инструменты анализа выбрать строку *Описательная статистика* и нажать кнопку *ОК;*
- в появившемся диалоговом окне указать входной диапазон, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Для этого следует навести указатель мыши на левую верхнюю ячейку данных, нажать левую кнопку мыши и, не отпуская ее, протянуть указатель мыши к правой нижней

ячейке, содержащей анализируемые данные, затем отпустить левую кнопку мыши;

- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует поставить переключатель в положение *Выходной диапазон* (навести указатель мыши и щелкнуть левой клавишей), далее навести указатель мыши в поле ввода *Выходной* диапазон и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши;
- в разделе *Группировка* переключатель установить в положение по столбцам;
- установить флажок в поле *Итоговая статистика*;
- нажать кнопку *ОК*.

В результате анализа в указанном выходном диапазоне для каждого столбца данных выводятся следующие статистические характеристики: среднее, стандартная ошибка (среднего), медиана, мода, стандартное отклонение, дисперсия выборки, эксцесс, асимметричность, размах вариации, минимум, максимум, сумма, счет, наибольшее, наименьшее, уровень надежности.

2. МЕТОДИКА ПРОВЕДЕНИЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

К каждому аудиторному занятию подготовлен соответствующий раздаточный материал или компьютерный вариант раздаточного варианта. Обучающиеся имеют возможность повторить теоретические положения, которые будут использоваться на предстоящем практическом занятии. Кроме того, имеется возможность проверить качество усвоения предыдущего материала, выполнив индивидуальное задание. Варианты индивидуальных заданий подготовлены для каждого практического занятия.

Так как на каждом практическом занятии прорабатывается определенная тема, связанная с применением какого-либо раздела данного курса, проводится ознакомление с существующими методиками проведения экспериментальных исследований и приводятся попытки их анализа, то, естественно, полностью проработать данную задачу за одно практическое занятие нереально. Поэтому, доработка темы, начатой на практическом занятии, может быть перенесена на самостоятельную работу.

Поскольку каждому обучающемуся выдается индивидуальное задание по каждой теме, по результатам его выполнения можно оценивать степень усвоения предлагаемого учебного материала.

На самостоятельную проработку выносятся следующие вопросы лекционного курса:

1) Метрологическое обеспечение экспериментальных исследований. Методы и средства измерения;

2) Распределения дискретных случайных величин (Пуассона, Бернулли и др.). Числовые характеристики дискретных случайных величин;

3) Специальные законы распределения;

4) Метод наименьших квадратов (МНК);

5) Сглаживание. Подавление главных компонент;

6) Вычисление определенных интегралов методом Монте-Карло;

7) Обзор функций пакета анализа данных.

Кроме того, имеется еще один вариант самостоятельной работы – это выполнение индивидуальной работы.

Такая организация учебного процесса по данной дисциплине, на мой взгляд, является целесообразной и полезной не только для усвоения учебного материала данной дисциплины, но и для развития практических навыков. Она должна способствовать овладению аспирантами навыками самостоятельной работы и реализации индивидуального творческого мышления по основным темам курса.

Литература

1. Кистенева М.Г. Математическое моделирование процессов в биосфере: Учебное пособие. – Томск: ТУСУР, 2007. – 98 с.
2. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel: Учебное пособие для ву-зов (изд. 3-е доп. и перераб.)/ Серия «Высшее образование». Ростов н/Д: Феникс, 2005. – 480 с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. – 10,11 изд., стереотип. М.: Высшая школа, 2005. – 478 с.
4. Смирнов Г.В. Статистические методы обработки: Учебное пособие для вузов. – Томск: ТУСУР, 2007. – 105 с.
5. Решетников М.Т. Планирование эксперимента и статистическая обработка данных: Учебное пособие для вузов. – Томск: ТУСУР, 2000. – 232 с.