Министерство науки и образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ ПО КУРСУ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ АНАЛИЗА ДАННЫХ»

Авторы:
Доц. Минитаева А.М.
Студент гр. ИУ6-43Б Есин А.

Москва, 2024

Оглавление

МОДУЛЬ 1 4
Семинар 1. Предварительная обработка данных 4
Выборки 4
Способы получения выборки 4
Ранжирование выборки 4
Вариационный ряд 5
Дискретный статический ряд 6
Интервальный статистический ряд 7
Эмпирическая функция распределения 8
Эмпирическая плотность распределения 10
Семинар 2. Элементы описательной статистики 12
Графическое изображение статических данных 12
МОДУЛЬ 2 17
Семинары 3-4. Проверка гипотез 17
Понятие статистической гипотезы 18
Задачи статистической проверки гипотез 18
Проверка статистических гипотез 18
Общая схема проверки статистических гипотез 19
Ошибки при проверке гипотез 20
Гипотеза о матожидании нормального распределения при известной дисперсии генеральной совокупности 21
Гипотеза о матожидании нормального распределения при неизвестной дисперсии генеральной совокупности 22
Гипотеза о сравнении генеральных дисперсий нормального распределения 24
Проверка гипотез о равенстве двух средних нормальных генеральных совокупностей, дисперсии которых известны (большие независимые выборки) 25
Проверка гипотезы о равенстве двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы (малые независимые выборки) 26
Проверка гипотезы о равенстве двух средних нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки) 29
Сравнение исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности 30
Сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события 32
Проверка гипотезы о равенстве дисперсий нескольких нормальных генеральных совокупностей по выборкам одинакового объёма (критерий Кочрена) 33

Проверка гипотезы о равенстве дисперсий нескольких нормальных совокупностей по выборкам различного объёма (критерий Бартлетта) 35
Сравнение двух вероятностей биномиальных распределений 38
Семинар 5. Статистические эксперименты 39
Понятие статистического эксперимента 39
Метод Монте-Карло 40
Разыгрывание дискретной случайной величины 43
Разыгрывание полной группы событий 44
Разыгрывание непрерывной случайной величины 45
Приближённое разыгрывание нормальной случайной величины 47
Разыгрывание дискретной двумерной случайной величины 48
Разыгрывание непрерывной двумерной случайной величины 49
Оценка надёжности простейших систем методом Монте-Карло 50
МОДУЛЬ 3 52
Семинары 6-7. Дисперсионный анализ 52
Одинаковое количество испытаний на всех уровнях 52
Неодинаковое число испытаний на различных уровнях 56
Семинар 8. Линейная регрессия. Метод наименьших квадратов 59
Математическая модель 60
Первое описание математической модели: самое общее 60
Второе описание математической модели: общее описание линейной модели 60
Третье описание математической модели: модель конкретизирована под условия примера 61
Алгоритмы проведения МНК 62
Первый алгоритм 62
Второй алгоритм 65
Третий алгоритм 67
Визуализация результатов после вычисления 68
Семинар 9. Временные ряды 69
Понятие временных рядов 69
Стационарные и нестационарные ряды 70
Анализ временных рядов 70
Моделирование временных рядов: авторегрессионные модели, скользящее среднее, ARIMA, SARIMA 72

МОДУЛЬ 1

Семинар 1. Предварительная обработка данных

Выборки

Генеральная совокупность – совокупность всех объектов, подлежащих исследованию.

Выборка (выборочная совокупность) – совокупность случайно отобранных объектов из генеральной совокупности. Она должна быть репрезентативной (представительной) – её объекты должны достаточно хорошо отражать свойства генеральной совокупности.

Выборки бывают повторные и неповторные. В повторных отобранные объекты перед отбором следующего возвращаются в генеральную совокупность, в неповторных этого не происходит.

Способы получения выборки

Виды отборов для получения выборки:

1. Простой – случайное извлечение объектов из генеральной совокупности (с возвратом или без)
2. Типический – объекты выбираются не из всей генеральной совокупности, а из её «типической» части
3. Серийный – объекты отираются сериями вместо отбора по одному
4. Механический – генеральная совокупность делится на столько частей, сколько объектов будет входить в выборку, и из каждой части выбирается по одному объекту

Объём N генеральной совокупности – количество объектов в генеральной совокупности.

Объём n выборки – количество объектов в выборке.

Предполагается, что $N >> n$.

Ранжирование выборки

Для обработки данных в выборке используется операция ранжирования – значения случайной величины сортируются в порядке возрастания.

После проведения ранжирования значения группируются так, чтобы в каждой отдельной группе значения случайной величины одинаковы. Каждое такое значение – вариант.

Ранжированный вариационный ряд распределения	
Количество полученных книг	Число студентов, получивших такое количество книг
2	7
3	9
4	9
5	5
6	6
7	3
10	1
Итого:	40

Варьирование – изменение значения варианта

Вариационный ряд

Вариационный ряд – последовательность вариантов, отсортированная по возрастанию.

Выборка:

39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 42, 41, 43, 39, 42, 41, 42, 39, 41, 37, 43, 41, 38, 43, 42, 41, 40, 41, 38, 44, 40, 39, 41, 40, 42, 40, 41, 42, 40, 43, 38, 39, 41, 41, 42.

Вариационный ряд:

xi	37	38	39	40	41	42	43	44
ni	1	3	5	8	12	9	5	2

Частота (вес) варианта – число, показывающее, сколько раз соответствующее значение вариантов встречается в ряде наблюдений. Обозначается n_i, где i – номер варианта.

Относительная частота (частость долей) варианта – отношение частоты данного варианта к общей сумме частот. Обозначается

$$p_i^* = (\frac{n_i}{n}) \text{ или } p_i^* = \frac{n_i}{\sum_{i=1}^{m} n_i}$$, где m – число вариантов

Частость – статическая вероятность появления варианта. Следовательно, естественно считать частость аналогом вероятности появления значения случайной величины X.

Дискретный статический ряд

Дискретный статический ряд – ранжированная совокупность вариантов (x_i) с соответствующими им частотами (n_i) или частостями (p_i^*).

Дискретный статический ряд удобно записывать в виде таблицы

$\sum_{i=1}^{5} n_i = 10$ $\sum_{i=1}^{5} p_i^* = 1$

$\boldsymbol{x_i}$	1	2	3	4	7
$\boldsymbol{n_i}$	2	2	3	1	2
$\frac{\mathbf{n_i}}{\mathbf{n}}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

Характеристики дискретного статистического ряда:

1) Размах варьирования $R = x_{max} - x_{min}$
2) Мода M_0^* - вариант, имеющий наибольшую частоту
3) Медиана M_e^* - значение случайной величины, находящееся в середине ряда

Пусть n – объём выборки. Тогда

$$M_e^* = \frac{x_k + x_k + 1}{2}$$, если $n = 2k$

$$M_e^* = x_{k+1}\text{, если } n = 2k + 1$$

Задача 1

Дано: Выборка задана в виде распределения частот:

$\boldsymbol{x_i}$	2	5	7
$\boldsymbol{n_i}$	1	3	6

Найти распределение относительных частот.

Решение: Найдём объём выборки: $n = 1 + 3 + 6 = 10$. Найдём относительные частоты:

$$\frac{\mathrm{n}_1}{\mathrm{n}} = \frac{1}{10} = 0.1;\ \frac{\mathrm{n}_2}{\mathrm{n}} = \frac{3}{10} = 0.3;\ \frac{\mathrm{n}_3}{\mathrm{n}} = \frac{6}{10} = 0.6$$

Напишем искомое распределение относительных частот:

$\boldsymbol{x_i}$	2	5	7
$\boldsymbol{\frac{n_i}{n}}$	0.1	0.3	0.6

Интервальный статистический ряд

Если изучаемая случайная величина X является непрерывной или число значений её велико, то составляется интервальный статистический ряд.

Сначала определяется число интервалов m в зависимости от объёмов выборки с помощью таблицы:

Объём выборки	25-40	40-60	60-100	100-200	200+
Число интервалов	5-6	6-8	7-10	8-12	10-15

Затем определяется длина частичного интервала h:

$h = \frac{x_{max} - x_{min}}{m}$, где h - шаг, m – число интервалов

Более точно шаг можно рассчитать при помощи формулы Стерджеса:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 * \lg n}$$

Если шаг оказывается дробным, то за длину интервала берётся ближайшее целое число или ближайшую простую дробь. Обычно берутся интервалы, одинаковые по длине, но интервалы разной длины допустимы.

За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{min} - \frac{h}{2}$, а конец последнего должен удовлетворять условию $x_{\text{кон}} - h \leq x_{max} \leq x_{\text{кон}}$

Промежуточный интервал получается путём прибавления шаг к концу предыдущего интервала.

Просматривая результаты наблюдений, определяется количество значений случайной величины, попавшей в каждый конкретный интервал. При этом в интервал включаются значения большие или равные нижней границе интервала и меньшие – верхней границы.

В первую строку таблицы статистического распределения вписывают частичные промежутки $[x_0, x_1), [x_1, x_2), \dots, [x_{m-1}, x_m)$

Во вторую строку статистического ряда вписывается количество наблюдений n_i, где $i = 1 \dots m$, попавших в каждый интервал (частоты соответствующих интервалов)

Эмпирическая функция распределения

Пусть получено статистическое распределение выборки, и каждому варианту из этой выборки поставлена в соответствие его частость.

Тогда эмпирическая функция (функция распределения выборки) – функция $F^*(x)$, определяющая для каждого значения x частость события

$F^*(x) = \frac{n_x}{n}$, где:

n – число выборки,

n_x – число наблюдений, меньших x

При увеличении объёма выборки частость события приближается к вероятности этого события.

Эмпирическая функция $F^*(x)$ является оценкой интегральной функции $F(x)$ в теории вероятностей.

$F^*(x)$ обладает теми же свойствами, что и $F(x)$:

1) $0 \leq F^*(x) \leq 1$
2) $F^*(x)$ - неубывающая функция
3) $F^*(-\infty) = 0$, $F^*(+\infty) = 1$

Задача 2

Дано: Найти эмпирическую функцию по данному распределению выборки:

x_i	1	4	6
n_i	10	15	25

Решение: Найдём объём выборки: $n = 10 + 15 + 25 = 50$.

Наименьшая варианта равна 1, поэтому $F^*(x) = 0$ при $x \leq 1$

Значение $X < 4$, а именно $x_1 = 1$, наблюдалось 10 раз => $F^*(x) = \frac{10}{50} = 0.2$ при $1 < x \leq 4$.

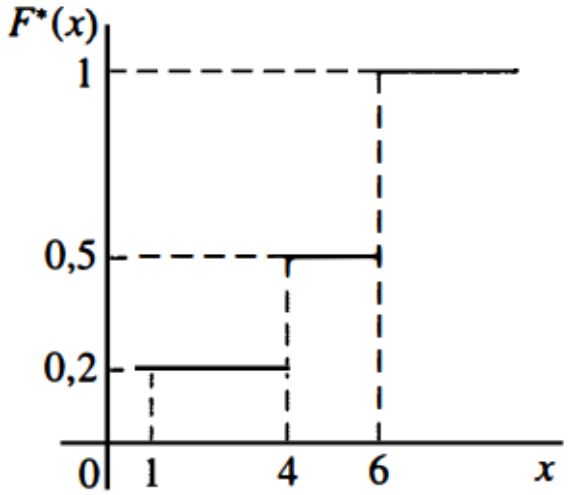
Значения $x < 6$, а именно $x_1 = 1$ и $x_2 = 4$, наблюдались $10 + 15 = 25$ раз => $F^*(x) = \frac{25}{50} = 0.5$ при $4 < x \leq 6$.

Т.к. $x = 6$ – наибольшая варианта, то $F^*(x) = 1$ при $x > 6$.

Напишем искомую эмпирическую функцию:

$$\begin{cases} 0 \text{ при } x \leq 1 \\ 0.2 \text{ при } 1 < x \leq 4 \\ 0.5 \text{ при } 4 < x \leq 6 \\ 1 \text{ при } x > 6 \end{cases}$$

График функции – на рисунке 1



Эмпирическая плотность распределения

Для интегральной функции распределения $F(x)$ справедливо приближённое равенство

$$f^*(x) = \frac{F^*(x + \Delta) - F^*(x)}{\Delta x}$$

где $F^*(x + \Delta) - F^*(x)$ – частость попадания наблюдаемых значений случайной величины X в интервал $[x; x + \Delta x)$

Таким образом, значение $f^*(x)$ характеризует плотность частоты на этом интервале.

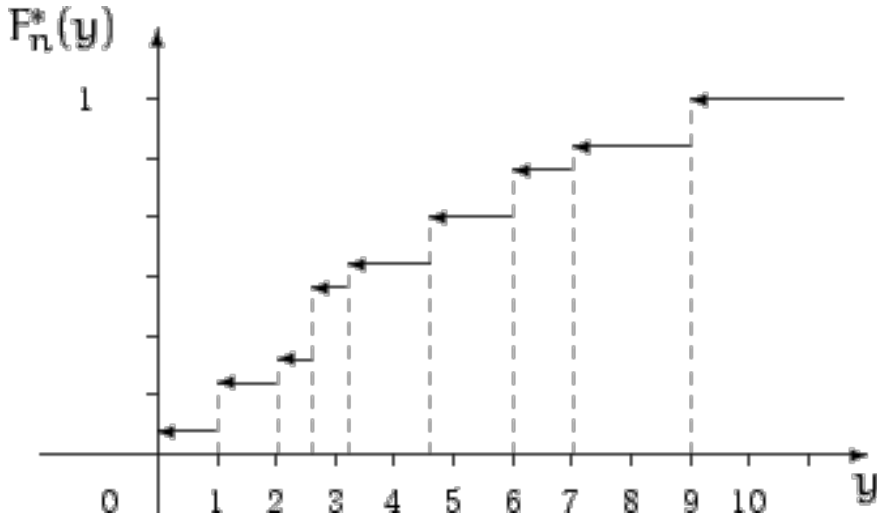
Пусть наблюдаемые значения непрерывной случайной величины представлены в виде интервального вариационного ряда.

Полагая, что p_i^* - частость попадания наблюдаемых значений в интервал $[a_i; a_i + h)$, где h - длина частичного интервала, выборочную функцию плотности $f(x)$ можно задать соотношением

$$f^*(x) = \begin{cases} 0 \text{ при } x < a_1 \\ \frac{p_i^*}{h} \text{ при } a_1 \le x \le a_{i+1}, i = 1 \dots m \\ 0 \text{ при } x > a_{m+1} \end{cases}$$

Где a_{m+1} – конец последнего m-интервала.

Т.к. функция $f^*(x)$ является аналогом распределения плотности случайной величины, площадь области под графиком этой функции равна 1.



Задача

Дано

Для выборки $2;\ 5;\ 3;\ 3;\ 6;\ 5;\ 2;\ 1;\ 5;\ 2$ построить эмпирическую функцию распределения

Решение

Статистический ряд:

x_i	1	2	3	5	6
n_i	1	3	2	3	1

По статистическому ряду запишем распределение выборки

x_i	1	2	3	5	6
p_i	0.1	0.3	0.2	0.3	0.1

ЭФР претерпевает разрыв в точках 1,2,3,5,6, а величина скачка равна соответствующей вероятности.

Таким образом, ЭФР имеет вид:

$$f^*(x) = \begin{cases} 0, x \leq 1, \\ 0 + 0.1 = 0.1, 1 < x \leq 2 \\ 0.1 + 0.3 = 0.4, 2 < x \leq 3 \\ 0.4 + 0.2 = 0.6, 3 < x \leq 5 \\ 0.6 + 0.3 = 0.9, 5 < x \leq 6 \\ 0.9 + 0.1 = 1, x > 6 \end{cases}$$

Задачи для самостоятельного решения:

1. Выборка задана в виде распределения частот:

x_i	4	7	8	12
n_i	5	2	3	10

Найти распределение относительных частот

2. Найти эмпирическую функцию по данному распределению выборки:

а)

$\boldsymbol{x_i}$	**2**	**5**	**7**	**8**
$\boldsymbol{n_i}$	**1**	**3**	**2**	**4**

б)

x_i	4	7	8
n_i	5	2	3

Семинар 2. Элементы описательной статистики

Графическое изображение статистических данных

Статистическое распределение изображается графически с помощью полигона и гистограммы.

Полигон частот – ломаная, отрезки которой соединяют точки с координатами $(x_i; n_i)$ полигоном частостей с координатами $(x_i; p_i^*)$, где $p_i^* = \frac{n_i}{n}, i = 1 \dots m$.

Полигон служит для изображения дискретного статистического ряда.

Полигон частостей – аналог многоугольника распределения дискретной случайной величины в теории вероятностей.

Гистограмма частот (частостей) – ступенчатая фигура, состоящая из прямоугольников, основания которых расположены на оси Ox и их длины равны длинам частичных интервалов (h), а высоты равны отношению:

$\frac{n_i}{h}$ - для гистограммных частот,

$\frac{n_i}{h*n}$ – для гистограммы частостей

Гистограмма – графическое изображение интервального ряда. Площадь гистограммы частот равна n, а гистограммы частостей равна 1.

Можно построить полигон для интервального ряда, если преобразовать его в дискретный ряд. В этом случае интервалы заменяют их серединными значениями и ставят в соответствие интервальные частоты (частости).

Задача

Дано

Дана выборка значений случайной величины X объёма 20:

12, 14, 19, 15, 14, 18, 13, 16, 17, 12, 18, 17, 15, 13, 17, 14, 14, 13, 14, 16

Требуется:

1) Построить дискретный вариационный ряд
2) Найти размах варьирования R, моду, медиану
3) Построить полигон частей.

Решение

Ранжируем выборку: 12, 12, 13, 13, 13, 14, 14, 14, 14, 14, 15, 15, 16, 16, 17, 17, 17, 18, 18, 19

Находим частоты вариантов и строим дискретный вариационный ряд

$$\sum_{i=1}^{8} n_i = 20$$

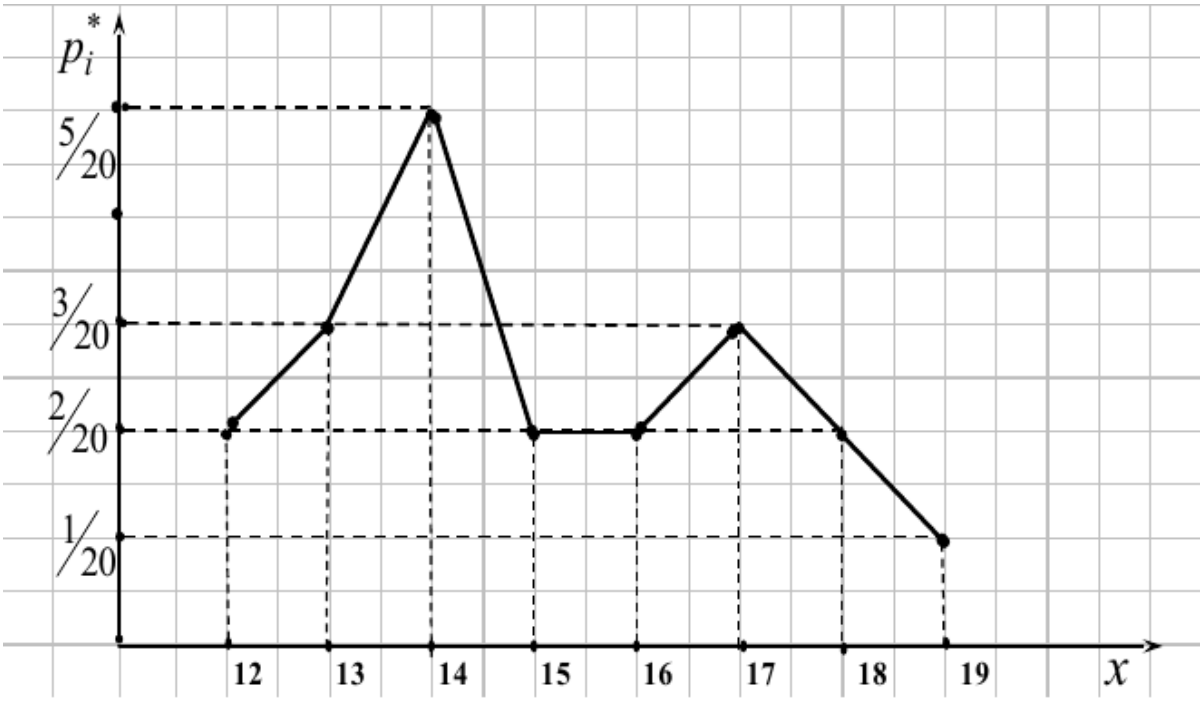
$$\sum_{i=1}^{8} p_i = 1$$

$\boldsymbol{x_i}$	12	13	14	15	16	17	18	19
$\boldsymbol{n_i}$	2	3	5	2	2	3	2	1
$\boldsymbol{p_i^* = \frac{n_i}{n}}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$

По результатам таблицы находим:

$$R = 19 - 12 = 7, M_O = 14, M_e = \frac{x_{10} + x_{11}}{2} = \frac{14 + 15}{2} = 14.5$$

Строим полигон частостей



Задача

Дано

Результаты измерений отклонений от нормы веса сердец кур-несушек дали численные значения (в мкм), приведённые в таблице

-1.760	-0.291	-0.110	-0.450	0.512
-0.158	1.701	0.634	0.720	0.490
1.531	-0.433	1.409	1.740	-0.266
-0.058	0.248	-0.095	-1.488	-0.361
0.415	-1.382	0.129	-0.361	-0.087
-0.329	0.086	0.130	-0.024	-0.882
0.318	-1.087	0.899	1.028	-1.304
0.349	-0.293	0.105	-0.056	0.757
-0.059	-0.539	-0.078	0.229	0.194
0.123	0.318	0.367	-0.992	0.529

Для данной выборки:

1) Построить интервальный ряд
2) Построить гистограммы и полигон частостей

Строим интервальный ряд

По данным таблицы определяем $x_{min} = -1.76$; $x_{max} = 1.74$

Для определения длины интервала h используем формулу Стерджеса:

$$h = \frac{x_{max} - x_{min}}{1 + 3.322 * lg50}$$

Число интервалов $m \approx 1 + 3.322 * lg50$

$h = \frac{x_{max} - x_{min}}{1+3.322*\lg n} = \frac{1.74-(-1.76)}{1+3.322*lg50} \approx \frac{3.5}{1+3.322*lg50} \approx \frac{3.5}{6.644}$

Примем $h = 0.6$, $m = 7$

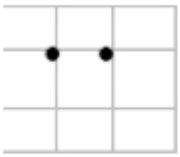
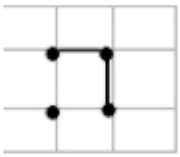
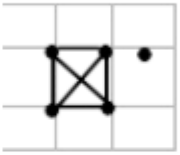
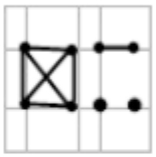
За начало первого интервала примем величину:

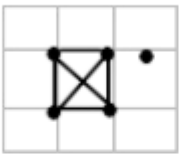
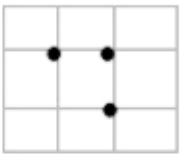
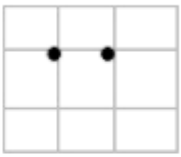
$$x_{\text{нач}} = x_{min} - \frac{h}{2} = -1.76 - 0.3 = -2.06$$

Строим интервальный ряд

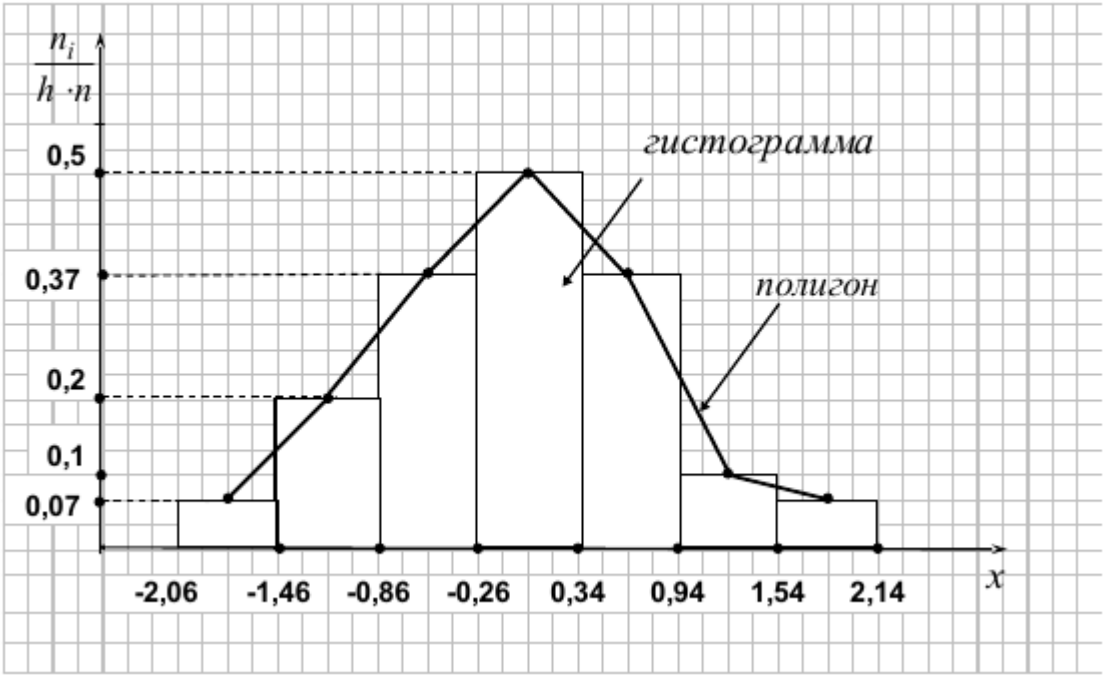
$$\sum_{i=1}^{7} n_i = 50$$

$$\sum_{i=1}^{7} p_i = 1$$

Интервалы	$[-2{,}06; -1{,}46)$	$[-1{,}46; -0{,}86)$	$[-1{,}86; -0{,}26)$	$[-0{,}26; 0{,}34)$
Подсчет частот				
Частоты n_i	2	6	11	15
Частости p_i	$\frac{2}{50}$	$\frac{6}{50}$	$\frac{11}{50}$	$\frac{15}{50}$

Интервалы	$[0{,}34; 0{,}94)$	$[0{,}94; 1{,}54)$	$[1{,}54; 2{,}14)$
Подсчет частот			
Частоты n_i	11	3	2
Частости p_i	$\frac{11}{50}$	$\frac{3}{50}$	$\frac{2}{50}$

Строим гистограмму частостей



Вершинами полигона являются середины верхних оснований прямоугольников гистограммы.

Задачи для самостоятельного решения

1. Пусть распределение частот имеет вид

x_i	2	4	6	8
n_i	3	5	7	5

   Построить полигон и гистограмму частостей.

2. Дан ряд непрерывного распределения частот

x_i	1-3	3-6	6-9	9-12
n_i	3	5	7	9

   Построить полигон и гистограмму частостей

3. Построить полигон частот и полигон частостей для данного распределения

x_i	2	7	8	15	16	17
n_i	15	35	64	55	21	10

МОДУЛЬ 2

Семинары 3-4. Проверка гипотез

Понятие статистической гипотезы

Статистическая гипотеза (гипотеза) – любое предположение о генеральной совокупности, проверяемое по выборке.

Параметрическая гипотеза – гипотеза, содержащая некоторое утверждение о параметрах распределения случайной величины (когда закон распределения считается известным). Иначе – гипотеза непараметрическая.

Нулевая гипотеза H_0 – предположение, которое выдвигается изначально, пока наблюдениями не признано обратное.

Альтернативная (конкурирующая) гипотеза H_1 – гипотеза, противоречащая нулевой, принимаемая, если основная гипотеза отвергнута.

Простая гипотеза – гипотеза, состоящая из только одного предположения.

Сложная гипотеза – гипотеза, состоящая из конечного или бесконечного количества простых гипотез.

Задачи статистической проверки гипотез

1) Относительно некой генеральной совокупности высказывается гипотеза H
2) Из генеральной совокупности извлекается выборка
3) Указывается правило, с помощью которого по выборке можно ответить на вопрос о том, следует ли отклонить гипотезу H или принять её
4) Выдвинутая гипотеза может быть правильной или неправильной => возникает необходимость её проверки

Статистическими методами нельзя доказать гипотезу, можно её только опровергнуть или не опровергнуть.

Проверка статистических гипотез

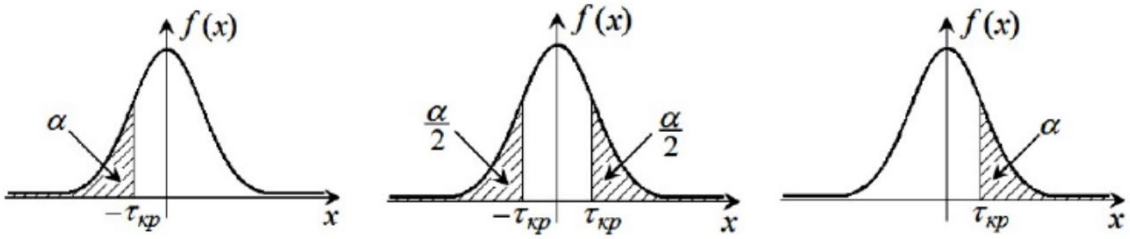
Имея гипотезы H_0 и H_1, необходимо на основе выборочных данных принять либо основную, либо альтернативную гипотезу.

Статистический критерий (критерий) проверки гипотезы – правило, по которому принимается решение, принять или отклонить гипотезу.

Статистика (тест) критерия – случайная величина τ, которая служит для проверки статистических гипотез.

Общая схема проверки статистических гипотез

1) Для основной гипотезы H_0 формулируется альтернативная гипотеза H_1.
2) Выбирается уровень значимости проверки – малое число $\alpha > 0$.
3) Рассматриваются теоретические выборки значений случайных величин, о которых сформулирована гипотеза H_0, и выбирается (формулируется) случайная величина τ. Значения и распределение τ полностью определяются по выборкам при предположении о верности H_0. Обычно τ выбирается из:
    - U – нормальное распределение
    - χ^2 – распределение Пирсона
    - T – распределение Стьюдента
    - F – Распределение Фишера-Снедекора
4) На числовой оси задают интервал D такой, что вероятность попадания случайной величины τ в интервал $P(\tau \in D) = 1 - \alpha$. Интервал D называется областью принятия гипотезы H_0, а оставшаяся область числовой оси – критической областью Величина $\tau = \tau_{\text{кр}}$ – критическое значение теста проверки. Различают 3 типа критических областей. Критическая область определяется с учётом гипотез:



5) По реализациям анализируемых выборок выбирается конкретное (наблюдаемое) значение теста τ ($\tau = \tau_{\text{набл}}$) и проверяется выполнение условия $P(\tau \in D) = 1 - \alpha$:
    - Если оно выполняется (например, $\tau_{\text{набл}} < \tau_{\text{кр}}$ для правосторонней области), то гипотеза H_0 принимается, т.е. она не противоречит опытным данным и нет оснований её опровергнуть
    - Если условие не выполняется, то полагается, что гипотеза H_0 неверна и отвергается

Для каждого критерия имеются соответствующие таблицы по которым находится критическое значение, удовлетворяющее приведённым выше соотношениям.

Принятие гипотезы H_0 не стоит расценивать как установленный абсолютно верный факт, лишь как достаточно правдоподобное, не противоречащее опыту утверждение.

Ошибки при проверке гипотез

Ошибки бывают 2 родов:

Ошибка I рода	**Ошибка II рода**
Отвергается основная гипотеза, несмотря на то, что она верна	Отвергается альтернативная гипотеза, несмотря на то, что она верна
Вероятность ошибки $P(H_1\|H_0) = \alpha$ α – уровень значимости критерия Обычно $\alpha = 0.05; 0.01; 0.005; 0.001$	Вероятность ошибки $P(H_0\|H_1) = \beta$ Величина β, как правило, заранее неизвестна
Вероятность приять верную гипотезу $P(H_0\|H_0) = 1 - \alpha$	Вероятность принять верную гипотезу $P(H_1\|H_1) = 1 - \beta$

Гипотеза о матожидании нормального распределения при известной дисперсии генеральной совокупности

Пусть генеральная совокупность X распределена по нормальному закону.

Генеральная средняя a неизвестна, но есть основания полагать, что она равна предполагаемому значению a_0.

Необходимо проверить гипотезу $H_0: a = a_0$ против альтернативной: $H_1: a \neq a_0$, или $H_1: a < a_1$, или $H_1: a > a_0$.

H_0	Статистика критерия	H_1	Область принятия H_0
$a = a_0$ $\sigma^2 = \sigma_г^2$, известно	$U = \frac{\bar{x} - a_0}{\sigma}\sqrt{n}$	$a \neq a_0$	$\lvert U\rvert < u_{кр}$ $\Phi(u_{кр}) = 0.5 - \frac{\alpha}{2}$
		$a < a_0$	$U > -u_{кр}$ $\Phi(u_{кр}) = 0.5 - \alpha$
		$a > a_0$	$U < u_{кр}$ $\Phi(u_{кр}) = 0.5 - \alpha$

Задача

Дано

Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 5$ извлечена выборка объёма $n = 100$, и по ней найдено выборочное среднее 26.5. Требуется на уровне значимости $\alpha = 0.05$ проверить гипотезу $H_0: a = a_0 = 25$

Решение

Найдём значение статистики критерия

$$U = \frac{\bar{x} - a_0}{\sigma}\sqrt{n} = \frac{26.5 - 25}{5}\sqrt{100} = 3$$

Из соотношения $\Phi(u_{кр}) = 0.5 - \frac{0.05}{2} = 0.475$ по таблице Лапласа находим $u_{кр}$ = 1.96

Т.к. $\lvert U\rvert > u_{кр}$, основная гипотеза отвергается.

Гипотеза о матожидании нормального распределения при неизвестной дисперсии генеральной совокупности

H_0	Статистика критерия	H_1	Область принятия H_0
$a = a_0$ $\sigma^2 = \sigma_г^2$, неизвестно	$T = \frac{\bar{x} - a_0}{s}\sqrt{n}$	$a \neq a_0$	$\lvert T\rvert < t_{кр}$ $t_{кр} = t_{\alpha,n-1}$ для двусторонней области
		$a < a_0$	$T > -t_{кр}$ $t_{кр} = t_{\alpha,n-1}$ для односторонней области
		$a > a_0$	$T < t_{кр}$ $t_{кр} = t_{\alpha,n-1}$ для односторонней области

s – исправленное среднее квадратичное отклонение

Значение $t_{кр}$находится по таблице Стьюдента

Задача

Дано

При выборке объёма $n = 16$, извлечённой из нормальной генеральной совокупности, найдены $x_0 = 12.4, s = 1.2$. Требуется при уровне значимости 0.05 проверить нулевую гипотезу $H_0: a = 11.8$ при альтернативной гипотезе $H_1: a = 11.8$.

Решение

Найдём наблюдаемое значение статистики критерия

$$T = \frac{\bar{x} - a_0}{s}\sqrt{n} = \frac{12.4 - 11.8}{1.2}\sqrt{16} = 2$$

Поскольку альтернативная гипотеза имеет вид

$$a \neq a_0$$

То искомая критическая область двухсторонняя. Из таблицы критических точек распределения Стьюдента найдём по уровню значимости $\alpha = 0.05$ и числу степеней свободы $k = n - 1 = 15$ критическую точку $t_{кр} = t_{кр}(0.05; 15) = 2.13$. Т.к. $|T| < t_{кр}$, то оснований отвергнуть основную гипотезу нет.

Гипотеза о сравнении генеральных дисперсий нормального распределения

Гипотезы о дисперсиях возникают довольно часто, т.к. дисперсия характеризует такие важные показатели, как точность машин, приборов, техпроцессов, степень однородности совокупностей, риск отклонения доходности активов от ожидаемого уровня и т.д.

H_0	Статистика критерия	H_1	Область принятия H_0
$\sigma_x^2 = \sigma_y^2$ a_x и a_y неизвестны	$F = \frac{s_{max}^2}{s_{min}^2}$	$\sigma_x^2 \neq \sigma_y^2$	$F < F_{\text{кр}}$ $F_{\text{кр}} = F_{\frac{\alpha}{2},n_1-1,n_2-1}$
		$\sigma_x^2 > \sigma_y^2$	$F < F_{\text{кр}}$ $F_{\text{кр}} = F_{\alpha,n_1-1,n_2-1}$

Задача

Дано

Измерения одной и той же физической величины проведены двумя методами. Получены следующие результаты:

В первом случае: $x_1 = 9.6;\ x_2 = 9.8;\ x_3 = 10; x_4 = 10.2;\ x_5 = 10.6;$

Во втором случае: $y_1 = 10.4;\ y_2 = 9.7;\ y_3 = 10; y_4 = 10.2;$

Предполагается, что результаты измерений распределены в выборке нормально и выборки независимы. Можно ли считать, что оба метода обеспечивают одинаковую точность измерений, если принять уровень значимости $\alpha = 0.1$?

Решение

Будем судить о точности методов по величине дисперсии.

Основная гипотеза $H_0: D(X) = D(Y)$

Альтернативная гипотеза $H_1: D(X) \neq D(Y)$

Найдём исправленные выборочные дисперсии:

Находим статистику: 1.48

Критическая область двусторонняя, поэтому по уровню значимости $\frac{\alpha}{2} = 0.05$ и числам степеней свободы $k_1 = 5 - 1 = 4; k_2 = 4 - 1 = 3$ находим критическую точку $F_{\text{кр}}(0.05; 4; 3) = 9.12.$

Т.к. $F_{\text{набл}} < F_{\text{кр}}$, то оснований отвергать основную гипотезу нет.

Следовательно, оба метода обеспечивают одинаковую точность измерений.

Проверка гипотез о равенстве двух средних нормальных генеральных совокупностей, дисперсии которых известны (большие независимые выборки)

Имеются две независимые выборки больших объёмов ($n_1 > 30, n_2 > 30$), по которым найдены выборочные средние. Генеральные дисперсии $D(X), D(Y)$ известны.

Необходимо проверить на уровне значимости α основную гипотезу $H_0: M(X) = M(Y)$.

H_0	Статистика критерия	H_1	Область принятия H_0
$a_x = a_y$ σ_x^2 и σ_y^2, известны	$U = \dfrac{\overline{x} - \overline{y}}{\sqrt{\dfrac{\sigma_x^2}{n_1} + \dfrac{\sigma_y^2}{n_2}}}$	$a_x \neq a_y$	$\lvert U \rvert < u_{\text{кр}}$ $\Phi(u_{\text{кр}}) = 0.5 - \dfrac{\alpha}{2}$
		$a_x < a_y$	$U > -u_{\text{кр}}$ $\Phi(u_{\text{кр}}) = 0.5 - \alpha$
		$a_x > a_y$	$U < u_{\text{кр}}$ $\Phi(u_{\text{кр}}) = 0.5 - \alpha$

Задача

Дано

Для проверки эффективности новой технологии отобраны две группы рабочих: в первой группе численностью $n_1 = 50$ человек, где применялась новая технология, выборочная средняя выработка составила $x = 85$ изделий, во второй группе численностью $n_2 = 70$ выборочная средняя - $y = 78$ изделий. Предварительно установлено, что дисперсии выработки в группах равны соответственно 100 и 74.

На уровне значимости $\alpha = 0.05$ выяснить влияние новой технологии на среднюю производительность.

Решение

Проверяемая гипотеза H_0: $a_x = a_y$, т.е. средние выработки одинаковы в обоих случаях.

В качестве альтернативной теории возьмём H_1: $a_x > a_y$

Находим фактическое значение статистики критерия

$$U = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}} = \frac{85 - 78}{\sqrt{\frac{100}{50} + \frac{74}{70}}} = 4$$

При альтернативной гипотезе H_1 по таблице Лапласа из соотношения $\Phi(u_{\text{кр}}) = 0.5 - 0.05 = 0.45$

$u_{\text{кр}} = 1.64$ – критическое значение

Т.к. $U > u_{\text{кр}}$, то гипотеза H_0 отвергается => можно сделать вывод, что новая технология повышает среднюю выработку.

Проверка гипотезы о равенстве двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы (малые независимые выборки)

Имеются две независимые выборки малых объёмов ($n_1 < 30, n_2 < 30$), по которым найдены выборочные средние и исправленные выборочные дисперсии. Генеральные дисперсии $D(X), D(Y)$ неизвестны, но предполагаются одинаковыми.

Необходимо проверить на уровне значимости α основную гипотезу $H_0: M(X) = M(Y)$.

H_0	Статистика критерия	H_1	Область принятия H_0
$a_x = a_y$ $\sigma_x^2 = \sigma_y^2$, неизвестны, но равны	$T = \frac{\bar{x} - \bar{y}}{s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s = \sqrt{\frac{s_x^2 * (n_1 - 1) + s_y^2 * (n_2 - 1)}{n_1 + n_2 - 2}}$	$a \neq a_0$	$\lvert T \rvert < t_{\text{кр}}$ $t_{\text{кр}} = t_{\alpha, n-1}$ для двусторонней области
		$a < a_0$	$T > -t_{\text{кр}}$ $t_{\text{кр}} = t_{\alpha, n-1}$ для односторонней области
		$a > a_0$	$T < t_{\text{кр}}$ $t_{\text{кр}} = t_{\alpha, n-1}$ для односторонней области

Задача

Дано

Из двух партий изделий, изготовленных на двух одинаково настроенных станках, извлечены малые выборки, объёмы которых $n = 10, m = 12$.

Получены следующие результаты:

Первый станок

Контролируемый размер изделий x_i	3.4	3.5	3.7	3.9
Частота n_i	2	3	4	1

Второй станок			
Контролируемый размер изделий y_i	3.2	3.4	3.6
Частота n_i	2	2	8

При условии значимости 0.02 проверим гипотезу $H_0: M(X) = M(Y)$ о равенстве средних размеров изделий при альтернативной гипотезе $H_1: M(X) \neq M(Y)$. Предполагается, что случайные величины X и Y распределены нормально.

Решение

Найдём выборочные средние и исправленные дисперсии для каждой выборки:

Для первой выборки: 3.6 и 0.0267 соответственно

Для второй выборки: 3.5 и 0.0255 соответственно

Для рассматриваемого критерия Стьюдента предполагается, что генеральные дисперсии одинаковы, поэтому надо сравнить дисперсии, используя критерий Фишера-Снедекора при $H_1: D(X) \neq D(Y)$.

$$F_{\text{набл}} = \frac{0.0267}{0.0255} = 1.05$$

По таблице имеем $F_{\text{кр}}(0.01; 9; 11) = 4.63$.

Т.к. $F_{\text{набл}} < F_{\text{кр}}$, дисперсии различаются незначимо.

Далее вычислим наблюдаемое значение критерия Стьюдента $T_{\text{набл}} = 1.45$.

По уровню значимости 0.02 и числу степеней свободы $k = n + m - 2 = 10 + 12 - 2 = 20$ находим по таблице Стьюдента критическую точку $t_{\text{дв.кр}}(0.02; 20) = 2.53$.

Т.к. $T_{\text{набл}} < t_{\text{дв.кр}}$, нет оснований отвергать основную гипотезу => средние размеры изделий существенно не различаются.

Проверка гипотезы о равенстве двух средних нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки)

Пусть две генеральные совокупности X и Y распределены нормально, причём их дисперсии неизвестны.

Введём обозначения:

$d_i = X_i - Y_i$ – разности вариант с одинаковыми номерами

$\overline{d} = \sum \frac{d_i}{n}$ – средняя разностей вариант с одинаковыми номерами

$s_d = \sqrt{\frac{\sum d_i^2 - \frac{[\sum d_i]}{n}}{n-1}}$ – «исправленное» среднее квадратическое отклонение.

Для того, чтобы при заданном уровне значимости α проверить основную гипотезу $H_0: M(X) = M(Y)$ о равенстве двух средних нормальных совокупностей X и Y с неизвестными дисперсиями (в случае зависимых выборок – одинакового объёма) при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$, необходимо:

1) Вычислить наблюдаемое значение критерия
2) По таблице критических точек распределения Стьюдента по заданному уровню значимости и числу степеней свободы $k = n - 1$ найти критическую точку $t_{\text{дв.к}}(\alpha, k)$:

- Если $|T| < t_{\text{дв.к}}(\alpha, k)$, оснований отвергать основную гипотезу нет
- Если $|T| < t_{\text{дв.к}}(\alpha, k)$, основная гипотеза отвергается

Задача

Дано

Двумя приборами в одном и том же порядке изменены шесть деталей и получены следующие результаты измерений (в мкм):

x_i	2	3	5	6	8	10
y_i	10	3	6	1	7	4

При уровне значимости 0.05 проверить основную гипотезу о равенстве результатов измерений в предположении, что они распределены нормально.

Решение

d_i	-8	0	-1	5	1	6

Выборочная средняя равна 0.5, «исправленное» среднее квадратическое отклонение $s_d = 5.01$.

$$T_{\text{набл}} = 0.24$$

По таблице находим критическую точку $t_{\text{дв.кр}}(0.05; 5) = 2.57$.

Сравнение исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности

Пусть n – объём выборки, по которой найдена исправленная дисперсия s^2.

При заданном уровне значимости α необходимо проверить основную гипотезу H_0: $\sigma^2 = \sigma_0^2$ о равенстве неизвестной генеральной дисперсии σ^2гипотетическому значению σ_0^2

H_0	Статистика критерия	H_1	Область принятия H_0

$\sigma^2 = \sigma_0^2$ a неизвестно	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\sigma^2 \neq \sigma_0^2$	$\chi^2_{1-\frac{\alpha}{2};n-1} < \chi^2 < \chi^2_{\frac{\alpha}{2};n-1}$
		$\sigma^2 < \sigma_0^2$	$\chi^2 > \chi^2_{1-\alpha;n-1}$
		$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha;n-1}$

Если число степеней свободы $k > 30$, то критическую точку $\chi^2_{\text{кр}}(a,k)$ можно найти из равенства Уилсона-Гильферти:

$$\chi^2_{\text{кр}}(a,k) = k\left[1 - \frac{2}{9k} + \zeta_\alpha\sqrt{\frac{2}{9k}}\right]^3$$

Где ζ_α находится, используя функцию Лапласа, из равенства

$$\Phi(\zeta_\alpha) = \frac{1-2\alpha}{2}$$

Задача

Дано

Точность работы станка-автомата проверяется по дисперсии контролируемого размера изделий, которая не должна превышать $\sigma_0^2 = 0.1$. Взята проба из 25 случайно отобранных изделий.

Получены следующие результаты:

Размер x_i	3.0	3.5	3.8	4.4	4.5
Частота n_i	2	6	9	7	1

При уровне значимости 0.05 проверить, обеспечивает ли станок требуемую точность.

Решение

$$H_0: \sigma^2 = 0.1$$

$$H_1: \sigma^2 > 0.1$$

Найдём исправленную выборочную дисперсию

$$s^2 = 0.1975$$

Найдём наблюдаемое значение критерия

$$\chi^2_{\text{набл}} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(25-1) * 0.1975}{0.1} \approx 48$$

Альтернативная гипотеза имеет вид $\sigma^2 > \sigma_0^2$ => критическая область односторонняя.

Найдём по таблице критическую точку $\chi^2_{\text{кр}}(0.05; 24) = 36.4$. Имеем $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$ => основную гипотезу отвергаем, станок не обеспечивает необходимую точность.

Сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события

Пусть по достаточно большому числу n независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, найдена относительная частота $\frac{m}{n}$, где m – число появлений события.

Требуется при заданном уровне значимости проверить основную гипотезу $H_0: p = p_0$ – неизвестная вероятность p равная гипотетической вероятности p_0.

H_0	Статистика критерия	H_1	Область принятия H_0
$p = p_0$	$U = \frac{p^* - p_0}{\sqrt{p_0 q_0}}\sqrt{n}$	$p \neq p_0$	$\lvert U\rvert < u_{\text{кр}}$ $\Phi(u_{\text{кр}}) = 0.5 - \frac{\alpha}{2}$

Достаточно большие n $np_0 > 5$ $nq_0 > 5$ $q_0 = 1 - p_0$	$p^* = \frac{m}{n}$	$p < p_0$	$U > -u_{\text{кр}}$ $\Phi(u_{\text{кр}}) = 0.5 - \alpha$
		$p > p_0$	$U < u_{\text{кр}}$ $\Phi(u_{\text{кр}}) = 0.5 - \alpha$

Задача

Дано

По 100 независимым испытаниям найдена относительная частота $\frac{m}{n} = 0.14$. При уровне системы значимости 0.05 требуется проверить основную гипотезу $H_0: p = p_0 = 0.2$при альтернативной $H_1: p \neq 0.2$

Решение

Найдём наблюдаемое значение критерия при $q_0 = 1 - p_0 = 1 - 0.2 = 0.8$

$$U_{\text{набл}} = \frac{p^* - p_0}{\sqrt{p_0 q_0}}\sqrt{n} = \frac{(0.14 - 0.2) * \sqrt{100}}{\sqrt{0.2 * 0.8}} = -1.5$$

По таблице функции Лапласа находим

$$u_{\text{кр}} = 1.96$$

Т.к. $|U_{\text{набл}}| < u_{\text{кр}}$ – нет оснований отвергать основную гипотезу (относительная частота 0.14 незначимо отличается от гипотетической вероятности 0.2).

Проверка гипотезы о равенстве дисперсий нескольких нормальных генеральных совокупностей по выборкам одинакового объёма (критерий Кочрена)

Пусть генеральные совокупности $X_1, X_2, \dots, X_m$ распределены нормально.

Из этих совокупностей извлечены m независимых выборок одинакового объёма n и по ним найдены исправленные выборочные дисперсии $s_1^2, s_2^2, \dots, s_m^2$, все с одинаковым числом степеней свободы $k = n - 1$.

Требуется на заданном уровне значимости проверить основную гипотезу об однородности дисперсий $H_0: D(X_1) = D(X_2) = \cdots = D(X_m)$ о равенстве между собой генеральных дисперсий.

В качестве критерия проверки основной гипотезы принят критерий Кочрена – отношение максимальной исправленной дисперсии к сумме всех исправленных дисперсий.

Для этого необходимо:

1) Вычислить наблюдаемое значение критерия $G_{\text{набл}} = \frac{s_{max}^2}{s_1^2 + s_2^2 + \cdots + s_l^2}$
2) По таблице критических точек распределения Кочрена найти критическую точку $G_{\text{кр}}(\alpha, k, l)$:
    - Если $G_{\text{набл}} < G_{\text{кр}}$, оснований отвергать основную гипотезу нет
    - Если $G_{\text{набл}} > G_{\text{кр}}$, основная гипотеза отвергается

При условии однородности дисперсий независимых выборок одинакового объёма в качестве оценки генеральной дисперсии принимают среднюю арифметическую исправленных дисперсий.

Задача

Дано

По четырём независимым выборкам одинакового объёма $n = 17$, извлечённым из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии: 0.21, 0.25, 0.34, 0.4.

Требуется:

1) При уровне значимости 0.05 проверить основную гипотезу об однородности дисперсий (критическая область правосторонняя)
2) Оценить генеральную дисперсию

Решение

Пункт а

Найдём наблюдаемое значение критерия Кочрена – отношение максимальной исправленной дисперсии к сумме всех дисперсий:

$$G_{\text{набл}} = \frac{0.4}{0.21 + 0.25 + 0.34 + 0.4} = \frac{1}{3}$$

Найдём по таблице критических точек распределения Кочрена по уровню значимости 0.05, числу степеней свободы $k = n - 1 = 17 - 1 = 16$ и числу выборок $l = 4$ критическую точку $G_{\text{кр}}(0.05; 16; 4) = 0.4366$.

Т.к. $G_{\text{набл}} < G_{\text{кр}}$, оснований отвергать нулевую гипотезу нет.

Пункт б

Поскольку установлена однородность дисперсий, в качестве оценки генеральной дисперсии примем среднюю арифметическую исправленных дисперсий:

$$D_{\text{г}}^{*} = \frac{0.21 + 0.25 + 0.34 + 0.4}{4} = 0.3$$

Проверка гипотезы о равенстве дисперсий нескольких нормальных совокупностей по выборкам различного объёма (критерий Бартлетта)

Пусть генеральные совокупности $X_1, X_2, \dots, X_m$ распределены нормально.

Из этих совокупностей извлечены m независимых выборок различных объёмов n_i. Некоторые объёмы могут быть равны. Если равны все объёмы, используется критерий Кочрена.

По выборкам найдены исправленные выборочные дисперсии $s_1^2, s_2^2, \dots s_m^2$. Требуется на установленном уровне значимости проверить основную гипотезу $H_0: D(X_1) = D(X_2) = \cdots = D(X_m)$ об однородности дисперсий.

Введём обозначения:

$k_i = n_i - 1$ – число степеней свободы дисперсии s_i^2

$k = \sum_{i=1}^{l} k_i$ – сумма чисел степеней свободы

$(\overline{s})^2 = \frac{1}{k}\sum_{i=1}^{l} k_i s_i^2$- средняя арифметическая исправленных дисперсий, взвешенная по числам степеней свободы

$$V = 2.303\left(k \lg(\overline{s})^2 - \sum_{i=1}^{l} k_i \, lg \, s_i^2\right)$$

$$C = 1 + \frac{1}{3(l-1)}\left(\sum_{i=`}^{l} \frac{1}{k_i} - \frac{1}{k}\right)$$

Для того, чтобы при заданном уровне значимости проверить основну. Гипотезу об однородности дисперсий нормальных совокупностей, необходимо:

1) Вычислить наблюдаемое значение критерия Бартлетта
$$B_{\text{набл}} = \frac{V}{C}$$
2) По таблице критических точек распределения по уровню значимости и числу степеней свободы найти критическую точку $\chi^2_{\text{кр}}(\alpha; l-1)$ правосторонней критической области.
   Если:
   - $B_{\text{набл}} < \chi^2_{\text{кр}}$, оснований отвергать основную гипотезу нет
   - $B_{\text{набл}} > \chi^2_{\text{кр}}$, основная гипотеза отвергается

При условии однородности дисперсии в качестве оценки генеральной дисперсии принимается среднее арифметическое исправленных дисперсий, взвешенное по числам степеней свободы

$$\overline{s^2} = \sum \frac{k_i s_i}{k}$$

Задача

Дано

По трём независимым выборкам, объёмы которых 9, 13 и 15, извлечённым из нормальных генеральных совокупностей, найдены исправленные выборки, равные 3.2, 3.8 и 6.3 соответственно.

При уровне значимости 0.05 необходимо проверить основную гипотезу об однородности дисперсий

Решение

Составим расчётную таблицу, последний столбец пока не заполняем, т.к. ещё неизвестно, понадобится ли считать C:

№ выборки	Объём выборки	Число степеней свободы	Исправленные дисперсии	$k_i s_i^2$	$\lg s_i^2$	$k_i \lg s_i^2$	$\frac{1}{k_i}$
i	n_i	k_i	s_i^2				
1	9	8	3.2	25.6	0.5051	4.408	
2	13	12	3.8	45.6	0.5798	6.9576	
3	15	14	6.3	88.2	0.7993	11.1902	
$\sum$		$k = 34$		159.4		22.1886	

Используя расчётную таблицу, найдём:

$$\overline{s^2} = 4.688$$

$$\lg \overline{s^2} = 0.6709$$

$$V = 1.43$$

По таблице по уровню значимости 0.05 и числу степеней свободы $l - 1 = 3 - 1 = 2$ находим критическую точку $\chi^2_{\text{кр}}(0.05; 2) = 6$

Т.к. $V < \chi^2_{\text{кр}}$, то $B_{\text{набл}} = \frac{V}{c} < \chi^2_{\text{кр}}$ ($C > 1$) => нет оснований отвергать основную гипотезу об однородности дисперсий (выборочные дисперсии различаются незначимо)

Сравнение двух вероятностей биномиальных распределений

Пусть в двух генеральных совокупностях проводятся независимые испытания: в результате каждого испытания событие А может появится в первой совокупности с неизвестной вероятностью p_1, а во второй – с неизвестной вероятностью p_2.

По выборкам, извлечённым из первой и второй совокупностей, найдены соответствующие частоты $w_1(A) = \frac{m_1}{n_1}$ & $w_2(A) = \frac{m_2}{n_2}$, где m_1 & m_2 – числа появлений события А, n_1, n_2 – количество испытаний.

В качестве оценок неизвестных вероятностей примем относительные частоты: $p_1 = w_1, p_2 = w_2$.

При заданной уровне значимости проверить гипотезу $H_0: p_1 = p_2 = p$ о равенстве вероятностей появления события в двух генеральных совокупностях, имеющих биномиальные распределения.

Наблюдаемое значение критерия

$$U_{\text{набл}} = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_1}}{\sqrt{\frac{m_1 + m_2}{n_2 + n_2}\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

При альтернативной гипотезе $H_1: p_1 \neq p_2$ по таблице функции Лапласа находится критическая точка $u_{\text{кр}}$ из равенства

$$\Phi\left(u_{\text{кр}}\right) = \frac{1 - \alpha}{2}$$

- Если $|U_{\text{набл}}| < u_{\text{кр}}$, то оснований отвергать основную гипотезу нет
- Если $|U_{\text{набл}}| > u_{\text{кр}}$, то основная гипотеза отвергается

При конкурирующей гипотезе $H_1: p_1 > p_2$ критическую точку правосторонней критической области находят из равенства

$$\Phi(u_{\text{кр}}) = \frac{1-2\alpha}{2}$$

Та же формула используется и для левосторонней области

Задача

Дано

За смену отказали 15 элементов устройства 1, состоящего из 800 элементов, и 25 элементов устройства 2, состоящего из 1000 элементов. При уровне значимости 0.05 проверить основную гипотезу $H_0: p_1 = p_2 = p$ о равенстве вероятностей отказа элементов обоих устройств при альтернативной гипотезе $H_1: p_1 \neq p_2$.

Решение

По условию альтернативная гипотеза имеет вид $p_1 \neq p_2$, поэтому критическая область двусторонняя. Найдём наблюдаемое значение критерия $U_{\text{набл}} = -0.89$.

Найдём критическую точку по равенству:

$$\Phi(u_{\text{кр}}) = \frac{1-\alpha}{2} = 0.475$$

По таблице функции Лапласа находим $u_{\text{кр}} = 1.96$. Т.к. $|U_{\text{набл}}| < u_{\text{кр}}$, оснований отвергать основную гипотезу нет (вероятности отказа элементов обоих устройств различаются незначительно).

Семинар 5. Статистические эксперименты

Понятие статистического эксперимента

Под экспериментами подразумевается оценка эффективности системы через имитационное моделирование, представляющее собой наблюдение

Статистический эксперимент (далее – эксперимент) – оценка эффективности системы через имитационное моделирование, представляющее собой наблюдение поведения модели системы под влиянием входных воздействий. При этом часть или все входные воздействия носят случайный характер. В результате получается набор экспериментальных данных, на основе которых могут быть оценены характеристики системы.

Для проведения экспериментов используются имитационные модели.

Имитационная модель – это формальное описание логики функционирования исследуемой системы и взаимодействия отдельных ее элементов во времени, учитывающее наиболее существенные причинно-следственные связи, присущие системе.

Метод Монте-Карло

В основе проведения экспериментов лежит метод статистических испытаний, он же метод Монте-Карло. В этом методе результат ставится в зависимость значения некоторой случайной величины, распределённой по заданному закону => результат каждого отдельного испытания несёт случайный характер.

Сущность метода Монте-Карло состоит в следующем: требуется найти значение a некоторой изучаемой величины. Для этого выбирают случайную величину X, матожидание которой равно a: $M(X) = a$.

Далее вычисляется (разыгрывается) n возможных значение x_i случайной величины X и находят их среднее арифметическое

$$\overline{x} = \frac{(\sum x_i)}{n}$$

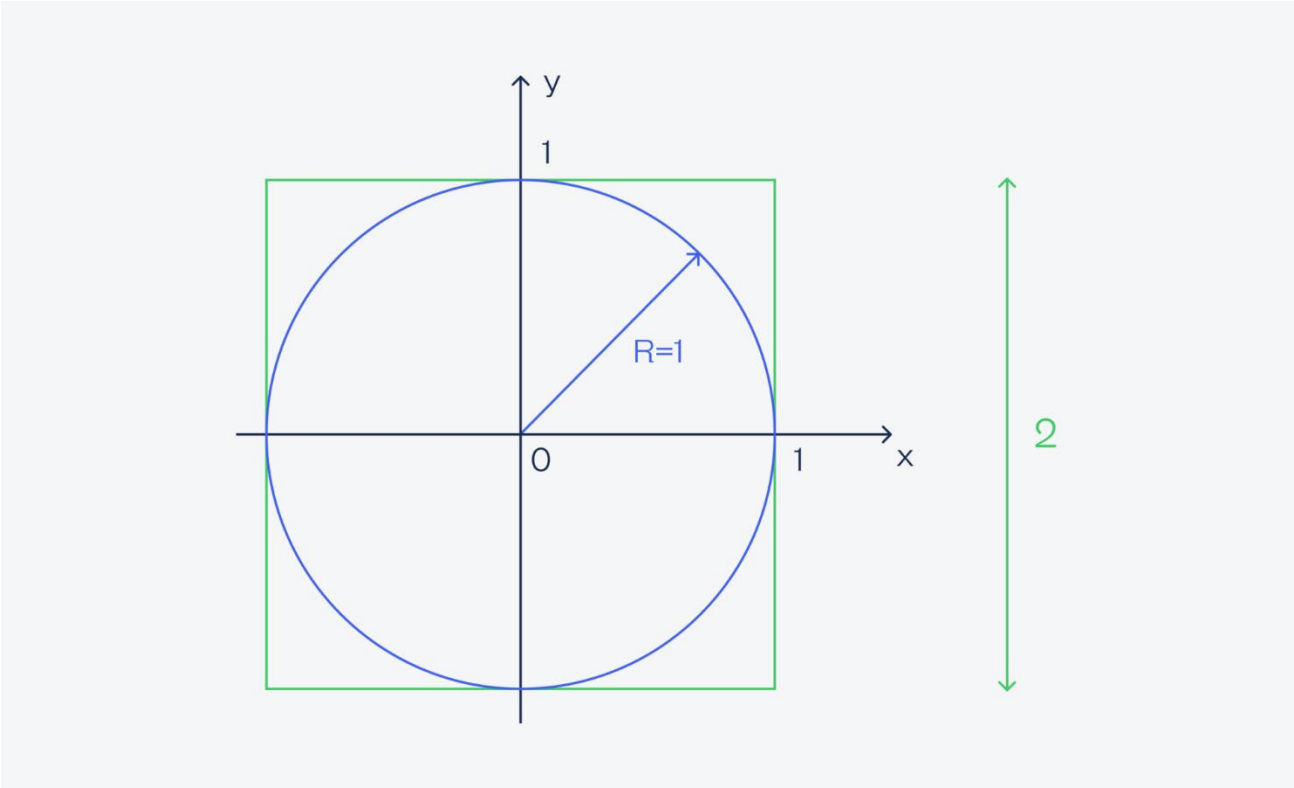
И принимается $\overline{x}$ в качестве оценки a^* искомого числа a:

$$a \cong a^* = \overline{x}$$

Таким образом, для применения метода Монте-Карло необходимо уметь разыгрывать случайную величину.

Для примера покажем классическое использование метода Монте-Карло — найдём число пи. Для этого нам понадобится круг, вписанный в квадрат,

причём у круга радиус будет равен 1. Это значит, что сторона квадрата равна 2 — это диаметр (или два радиуса) круга:



В этот квадрат мы будем случайным образом кидать песчинки и смотреть, попадут они в круг или нет (но останутся в границах квадрата). Исходя из этого набора данных мы можем посчитать отношение всех песчинок, которые попали в круг, ко всем песчинкам.

Теперь смотрим на формулы:

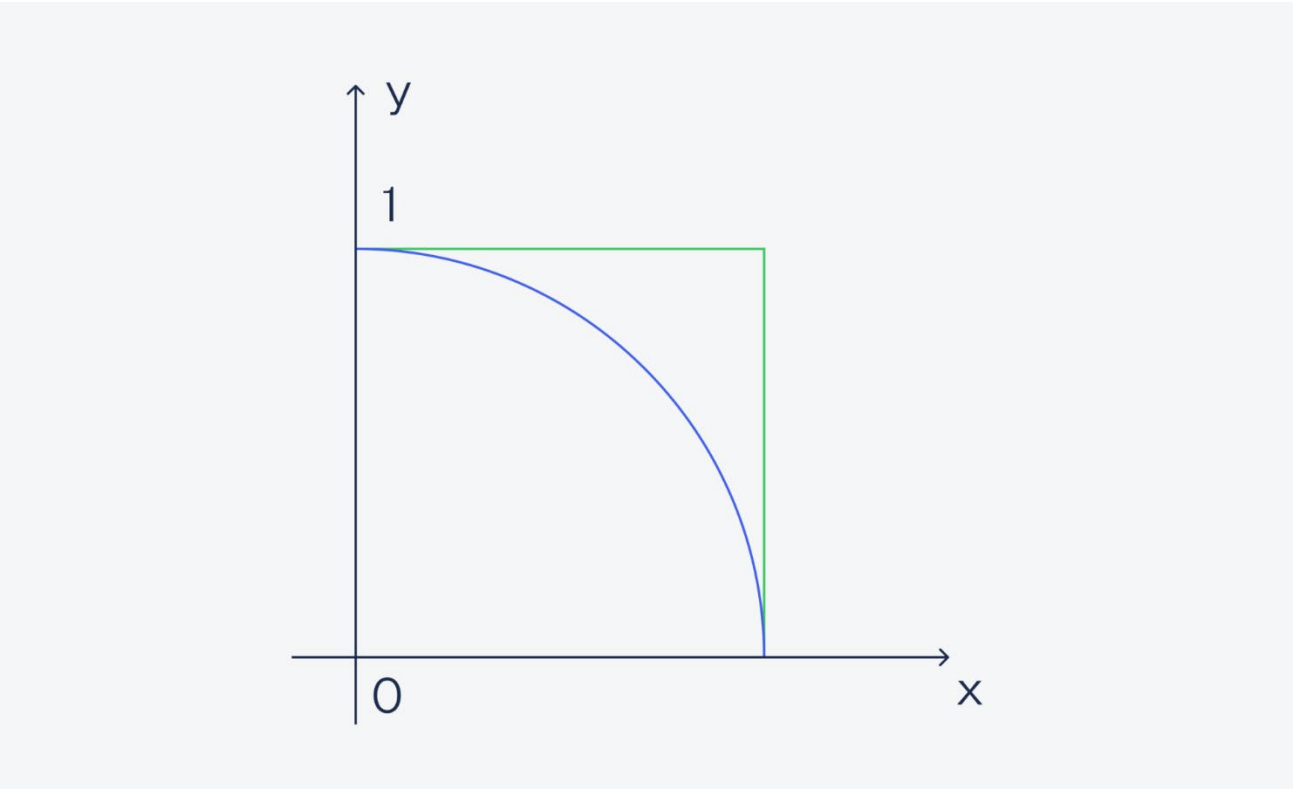
- площадь квадрата со стороной 2 равна четырём;

- площадь круга радиусом 1 равна $\pi R^2 \rightarrow \pi \times 1^2 = \pi$.

Если мы разделим площадь круга на площадь квадрата, то получим $\frac{\pi}{4}$. Но мы ещё не можем по условию посчитать площадь круга, потому что мы не знаем число π. Вместо этого мы можем разделить количество одних песчинок на другие.

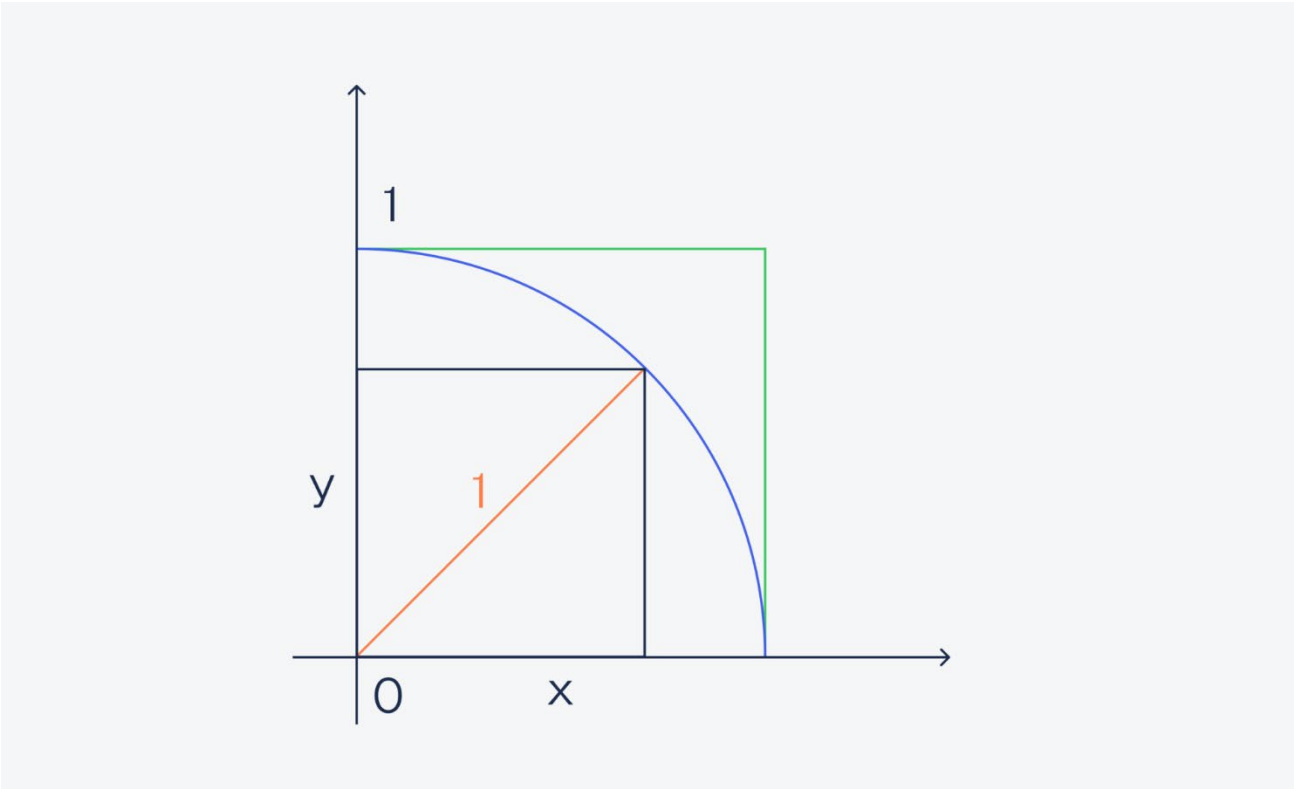
Это соотношение даст нам результат — $\frac{\pi}{4}$. Получается, что если мы умножим этот результат на 4, то получим число π, причём чем больше песчинок мы кинем, тем точнее будет результат.

Кидать песчинки будем так: в качестве координат попадания X и Y будем брать случайные числа от 0 до 1. Это значит, что все числа попадут только в один квадрант — правый верхний:



Но так как в этом квадранте ровно четверть круга и ровно четверть квадрата, то соотношение промахов и попаданий будет таким же, как если бы мы бросали песчинки в целый круг и целый квадрат.

Чтобы проверить, попадает ли песчинка в круг, используем формулу длины гипотенузы: $x^2 + y^2 = 1$ (так как гипотенуза — это радиус окружности):



Если длина гипотенузы меньше единицы — точка попадает в круг. В итоге мы посчитаем и общее количество точек, и точек, которые попали в круг. Потом мы разделим одно на другое, умножим результат на 4 и получим приближённое значение числа π.

Разыгрывание дискретной случайной величины

Введём обозначения:

R – непрерывная случайная величина, распределённая равномерно в интервале $(0; 1)$.

r_j $(j = 1, 2, \dots)$ - случайные числа (возможные значения R)

Для того, чтобы разыграть дискретную случайную величину X, заданную законом распределения

$$X\ x_1\ x_2\ \dots x_n,$$

$$p\ p_1\ p_2\ \dots p_n$$

необходимо:

1) Разбить интервал $(0, 1)$ оси Or на n частичных интервалов:
   $\Delta_1 - (0; p_1), \Delta_2 - (p_1; p_1 + p_2), \dots, \Delta_n - (p_1 + p_2 + p_3 + \dots + p_{n-1}; 1)$.

2) Выбрать случайное число r_j
   Если r_j попало в частичный интервал Δ_i, то разыгрываемая величина приняла возможное значение x_i.

Задача

Дано

Разыграть 6 возможных значений дискретной случайной величины X, закон распределения которой задан в виде таблицы:

X	2	10	18
p	0.22	0.17	0.61

Решение

Разобьём интервал $(0, 1)$ оси Or точками с координатами $0.22;\ 0.22 + 0.17 = 0.39$ на три частичных интервала $\Delta_1 - (0; 0.22), \Delta_2 - (0.22; 0.39),\ \Delta_3 - (0.39; 1)$.

Возьмём 6 случайных чисел: 0.32, 0.17, 0.9, 0.05, 0.97, 0.87. Случайное число $r_1 = 0.32$ принадлежит частичному интервалу Δ_2 => разыгрываемая дискретная случайная величина приняла значение $x_2 = 10$.

Аналогично для остальных 5 чисел. Получаем следующие разыгранные значения: 10, 2, 18, 2, 18, 18.

Разыгрывание полной группы событий

Требуется разыграть испытания, в каждом из которых наступает одно из событий полной группы, вероятности которых известны. Разыгрывание полной группы событий сводится к разыгрыванию дискретной случайной величины

Для того, чтобы разыграть испытания, в каждом из которых наступает одно из событий $A_1, A_2, \dots, A_n$ полной группы, вероятности которых $p_1, p_2, \dots, p_n$ известны, достаточно разыграть по правилу, прописанному выше, дискретную случайную величину X со следующим законом распределения:

X	1	2	…	n

p	p_1	p_2	…	p_n

Если в испытании величина X приняла значение $x_i = i$, то наступило событие A_i.

Задача

Дано

Заданы вероятности 3 событий: A_1, A_2, A_3, образующих полную группу: $p_1 = P(A_1) = 0.22; p_2 = P(A_2) = 0.31; p_3 = P(A_3) = 0.47$. Разыграть 5 испытаний, в каждом из которых появляется одной из трёх рассматриваемых событий

Решение

В соответствии с правилом настоящего параграфа надо разыграть дискретную случайную величину X с законом распределения:

X	1	2	3
p	0.22	0.31	0.47

По вышеописанному правилу разобьём интервал (0, 1) на 3 частичных интервала: $\Delta_1 = (0; 0.22);\ \Delta_2 = (0.22; 0.43);\ \Delta_3 = (0.43; 1)$.

Выберем 5 случайных чисел: $0.61, 0.19, 0.69, 0.04, 0.46$.

Случайное число $r_1 = 0{,}61$ принадлежит интервалу Δ_3, поэтому

X = 3 и, следовательно, наступило событие A_3. Аналогично найдем остальные события. В итоге получим искомую последовательность событий:
A_3, A_1, A_3, A_1, A_3.

Разыгрывание непрерывной случайной величины

Известна функция распределения $F(x)$ непрерывной случайной величины X. Требуется разыграть X (найти последовательность значений x_i).

Существует два метода разыгрывания непрерывных случайных величин.

Первый – метод обратных функций.

Суть его в том, чтобы взять случайное число r_i, приравнять его функции распределения и решить относительно x_i полученное уравнение $F(x_i) = r_i$.

Для того, чтобы разыграть значение x_i непрерывной случайной величины X, зная её плотность $f(x)$, надо выбрать случайное число r_i и решить относительно x_i уравнение

$$\int_{-\infty}^{x_i} f(x)dx = r_i$$

или уравнение

$$\int_{a}^{x_i} f(x)dx = r_i$$

Второй метод – метод суперпозиции

Для того, чтобы разыграть значение случайной величины X, функция распределения которой

$$F(x) = C_1F_1(x) + C_2F_2(x) + \cdots + C_nF_n(x)$$

Где $F_k(x)$ – функции распределения $(k = 1,2, \dots, n)$, $C_k > 0, C_1 + C_2 + \cdots + C_n = 1$, надо выбрать два независимых случайных числа r_1 & r_2 и по случайному числу r_1 разыграть возможное значение вспомогательной случайной величины Z:

Z	1	2	…	n
p	C_1	C_2	…	C_n

Если окажется, что $Z = k$, то решают относительно x уравнение $F_k(x) = r_2$.

Задача

Дано

Найти явную формулу для разыгрывания непрерывной случайной величины X, распределённой равномерно в интервале (a, b), зная её функцию распределения $F(x) = \frac{x-a}{b-a}$ $(a < x < b)$

Решение

Приравниваем заданную функцию распределения случайному числу r_i:

$$\frac{x_i - a}{b - a} = r_i$$

Решив это уравнение относительно x_i, получаем явную формулу для разыгрывания возможных значений X: $x_i = (b - a)r_i + a$

Приближённое разыгрывание нормальной случайной величины

Требуется приближённо разыграть нормальную случайную величину

Для того, чтобы приближённо разыграть возможное значение x_i нормальной случайной величины X с параметрами $a = 0$ & $\sigma = 1$, надо сложить 12 независимых случайных чисел u из полученной суммы вычесть 6:

$$x_i = \sum_{j=1}^{12} r_j - 6 = S_1 - 6$$

Если требуется приближённо разыграть нормальную случайную величину Z с математическим отклонением σ, то, разыграв x_i по приведённому выше правилу, находят искомое возможное значение по формуле

$$\zeta_i = \sigma x_i + a$$

Задача

Дано

Разыграть 4 возможных значения нормальной случайной величины с параметрами:

а) $a = 0, \sigma = 1$
б) $a = 2, \sigma = 3$

Решение

а) В соответствии с правилом разыграем возможное значение x_1 нормальной случайной величины X с параметрами $a = 0$ & $\sigma = 1$ по формуле

$$x_1 = \sum_{j=1}^{12} r_j - 6 = S_1 - 6$$

Выберем 12 случайных чисел: 0.37, 0.54, 0.2, 0.48, 0.05, 0.64, 0.89, 0.47, 0.42, 0.96, 0.24, 0.8. Сложив эти числа, получим $S_1 = 6.06$. Искомое возможное значение $x_1 = S_1 - 6 = 0.06$.
Аналогично, выбрав по 12 случайных чисел ещё три раза, получим $S_2 = 4.9, S_3 = 4.48, S_3 = 6.83$ => $x_2 = -1.1, x_2 = -1.52, x_3 = 0.83$.

б) Найдём возможные значения нормальной случайной величины Z с параметрами $a = 2, \sigma = 3$ по формуле $\zeta_i = \sigma x_i + a$.
Подставив возможные значения $x_1 = 0.06, a = 2, \sigma = 3$, получим $\zeta_1 = 3 * 0.06 + 2 = 2.18$
Аналогично найдём остальные возможные значения: $\zeta_2 = -1.3, \zeta_3 = -2.56, \zeta_4 = 4.49$

Разыгрывание дискретной двумерной случайной величины

Разыгрывание двумерной случайной величины (X, Y) сводится к разыгрыванию её составляющих.

Пусть задан закон распределения двумерной случайной величины (X, Y). Если составляющие независимы, то находятся законы их распределения и по ним разыгрывают составляющие по правилу, описанному выше.

Если составляющие зависимы, то находится закон распределения одной из них, условные законы распределения другой и разыгрывают составляющие по правилу, описанному выше.

Задача

Дано

Дискретная двумерная случайная величина (X, Y), составляющие которой независимы, задана законом распределения:

Y	X		
	x_1	x_2	x_3
y_1	0.18	0.3	0.12
y_2	0.12	0.2	0.08

Разыграть случайную величину (X, Y)

Решение

Найдём закон распределения составляющей X:

$$p_1 = P(X = x_1) = 0.18 + 0.12 = 0.3$$

$$p_2 = P(X = x_2) = 0.3 + 0.2 = 0.5$$

$$p_3 = P(X = x_3) = 0.12 + 0.08 = 0.2$$

Таким образом, искомый закон распределения имеет вид

X	x_1	x_2	x_3
p	0.3	0.5	0.2

Аналогично найдём закон распределения Y:

Y	y_1	y_2
p	0.6	0.4

Составляющие X & Y разыгрываются по правилу, описанному выше

Разыгрывание непрерывной двумерной случайной величины

Суть метода аналогична разыгрыванию дискретной двумерной случайной величины

Задача

Дано

Найти явные формулы для разыгрывания непрерывной двумерной случайной величины (X, Y), заданной плотностью вероятности $f(x, y) = \frac{3}{4}xy^2$ в области, ограниченной прямыми $x = 0, y = 0, x = 1, y = 2$.

Решение

Составляющие X и Y независимы, так как совместную плотность вероятности $f(x,\ y) = \left(\frac{3}{4}\right)xy^2$ можно представить в виде произведения двух функций, одна из которых зависит только от y, а другая - только от y.

Найдём плотность распределения составляющей X:

$$f_1(x) = \int_0^2 f(x,y)dy = \left(\frac{3}{4}\right)x \int_0^2 y^2 dy = 2x$$

Итак, $f_1(x) = 2x, 0 < x < 1$

Разыграем X по вышеописанному правилу:

$$2\int_0^{x_i} x\,dx = r_i$$

Отсюда получим явную формулу для вычисления возможных значений X:

$$x_i = \sqrt{r_i}$$

Найдём плотность распределения составляющей Y:

$$f_2(y) = \int_0^1 f(x,y)dx = \left(\frac{3}{4}\right)y^2 \int_0^1 x\,dx = \left(\frac{3}{8}\right)y^2$$

Итак, $f_2(y) = \left(\frac{3}{8}\right)y^2, 0 < y < 2$

Разыграем:

$$\left(\frac{3}{8}\right)\int_0^{y_i} y^2 dy = r_i'$$

Отсюда получим явную формулу для вычисления возможных значений y:

$$y_i = \sqrt[3]{r_i'}$$

Оценка надёжности простейших систем методом Монте-Карло

Задача

Дано

Система состоит из двух блоков, соединенных последовательно. Система отказывает при отказе хотя бы одного блока. Первый блок содержит два элемента: A, B (они соединены параллельно) и отказывает при одновременном отказе обоих элементов. Второй блок содержит один элемент C и отказывает при отказе этого элемента.

а) Найти методом Монте-Карло оценку P^* надежности (вероятности безотказной работы) системы, зная вероятности безотказной работы элементов: $P(A) = 0.8, P(B) = 0.85, P(C) = 0.6$

б) найти абсолютную погрешность $|P - P^*|$, где P - надежность системы, вычисленная аналитически. Произвести 50 испытаний.

Решение

а) Выберем три случайных числа: 0,10; 0,09; 0,73 и разыграем события A, B, C, состоящие в безотказной работе соответственно элементов A, B, C, по правилу: если случайное число меньше вероятности события, то событие наступило; если случайное число больше или равно вероятности события, то событие не наступило. Результаты испытания будем записывать в расчетную таблицу

Поскольку $P(A) = 0.8 \ \& \ 0.1 < 0.8$, то событие A наступило, т.е. элемент A в этом испытании работает безотказно. Так как $P(B) = 0.85 \ \& \ 0.09 < 0.85$, то событие B наступило, т.е. элемент B работает безотказно. Таким образом, оба элемента первого блока работают; следовательно, работает и сам первый блок. В соответствующих клетках таблицы ставим знак плюс.

Так как $P(C) = 0.6 \ \& \ 0.73 > 0.6$, то событие C не наступило, т.е. элемент C получает отказ; другими словами, второй блок, а значит, и вся система, получают отказ. В соответствующих клетках таблице ставим знак минус. Аналогично разыгрываются и остальные испытания. В таблице приведены результаты четырех испытаний.

№	Блок	След. числа, модел. эл − ты			Заключение о работе				
								блоков	системы
		A	B	C	A	B	C		
1	I	0.1	0.09		+	+		+	−
	II			0.73			−	−	
2	I	0.25	0.33		+	+		+	−
	II			0.76			−	−	
3	I	0.52	0.01		+	+		+	+
	II			0.35			+	+	
4	I	0.86	0.34		−	+		+	−
	II			0.67			−	−	

Произведя 50 испытаний, получим, что в 28 из них система работала безотказно. В качестве оценки искомой надежности P примем относительную частоту $P^* = \frac{28}{50} = 0.56$.

б) Найдем надежность системы P аналитически. Вероятности безотказной работы первого и второго блоков соответственно равны:

$$P_1 = 1 - P(A) * P(B) = 1 - 0.2 * 0.15 = 0.97; P_2 = P(C) = 0.6$$

Вероятность безотказной работы системы

$$P = P_1 * P_2 = 0.582$$

Искомая абсолютная погрешность $|P = P^*| = 0.582 = 0.56 = 0.022$.

МОДУЛЬ 3

Семинары 6-7. Дисперсионный анализ

Дисперсионный анализ (ANOVA – ANalysis Of VAriance) – метод, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. Позволяет сравнивать значения трёх и более групп.

Если:

- Одна зависимая переменная
- Одна независимая (группирующая) переменная

Тогда производится однофакторный дисперсионный анализ (One-way ANOVA)

Одинаковое количество испытаний на всех уровнях

Пусть на количественный нормально распределенный признак X воздействует фактор F, который имеет p постоянных уровней $F_1, F_2, \ldots, F_p$. На каждом уровне произведено по q испытаний. Результаты наблюдений - числа (x_{ij}, где i - номер) испытания ($i = 1, 2, \ldots, q$), j - номер уровня фактора ($j = 1, 2, \ldots, \mathrm{p}$) - записывают в виде таблицы.

№	Уровни фактора			
i	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}

… q	… x_{q1}	… x_{q2}	… …	… x_{qp}
Групповая средняя $\overline{x}_{\text{гр}}$	$\overline{x}_{\text{гр}1}$	$\overline{x}_{\text{гр}2}$	…	$\overline{x}_{\text{гр}p}$

Ставится задача: на уровне значимости α проверить нулевую гипотезу о равенстве групповых средних при допущении, что групповые генеральные дисперсии хотя и неизвестны, но одинаковы. Для решения этой задачи вводятся:

- *Общая сумма* квадратов отклонений наблюдаемых значений признака от общей средней

$$S_{\text{общ}} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij} - \overline{x}\right)^2$$

- *Факторная сумма* квадратов отклонений групповых средних от общей средней (характеризует рассеяние между группами)

$$S_{\text{факт}} = q\sum_{j=1}^{p}\left(\overline{x}_{\text{гр}j} - \overline{x}\right)^2$$

- *Остаточная сумма* квадратов отклонений наблюдаемых значений группы от своей групповой средней (характеризует рассеяние «внутри групп»)

$$S_{\square} = \sum_{i=1}^{q}\left(x_{i1} - x_{\text{гр}1}\right)^2 + \sum_{i=1}^{q}\left(x_{i2} - x_{\text{гр}2}\right)^2 + \cdots + \sum_{i=1}^{q}\left(x_{i\,p} - x_{\text{гр}\,p}\right)^2$$

Практически остаточная сумма находится по формуле:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}$$

Для вычисления общей и факторной сумм более удобны следующие формулы:

$$S_{\text{общ}} = \sum_{i=1}^{p} P_j - \frac{\left[\sum_{j=1}^{p} R_j\right]^2}{pq}$$

$$S_{\text{факт}} = \sum_{i=1}^{p} \frac{R_j^2}{q} - \frac{\left[\sum_{i=1}^{p} R_j\right]^2}{pq}$$

, где $P_j = \sum_{i=1}^{q} x_{ij}^2$ – сумма квадратов наблюдаемых значений признака на уровне F_j.

Если наблюдаемые значения признака – сравнительно большие числа, то для упрощения вычислений вычитается из каждого наблюдаемого значения одно и то же число С, примерно равное общей средней. Если уменьшенные значения $y_{ij} = x_{ij} - C$, то

$$S_{\text{общ}} = \sum_{i=1}^{p} Q_j - \frac{\left[\sum_{j=1}^{p} R_j\right]^2}{pq}$$

$$S_{\text{факт}} = \frac{\left[\sum_{j=1}^{p} T_j^2\right]}{q} - \frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq}$$

, где $Q_j = \sum_{i=1}^{q} y_{ij}^2$ – сумма квадратов уменьшенных значений признака на уровне F_j; $T_j = \sum_{i=1}^{q} y_{ij}$ – сумма уменьшенных значений признака на уровне F_j.

Разделив уже вычисленные факторную и остаточную суммы на соответствующее число степеней свободы, находим факторную и остаточную дисперсии:

$$s_{\text{факт}}^2 = \frac{S_{\text{факт}}}{p-1}; s_{\text{ост}} = \frac{S_{\text{ост}}}{p(q-1)}$$

Наконец, сравнивают факторную и остаточную дисперсии по критерию Фишера-Снедекора. Если $F_{\text{накл}} < F_{\text{кр}}$ – различие групповых средних незначимое. Если $F_{\text{накл}} > F_{\text{кр}}$ - различие групповых средних значимое.

Задача

Дано

Произведено по четыре испытания на каждом из трех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице.

№	Уровни фактора		
	F_1	F_2	F_3
1	38	20	21
2	36	24	22
3	35	26	31
4	31	30	34

$\overline{x}_{\text{гр } j}$	35	25	27

Решение

Для упрощения расчета вычтем из каждого наблюдаемого значения x_{ij} общую среднюю $\text{x} = 29$, т. е. перейдем к уменьшенным величинам: $\text{y}_{ij} = x_{ij} - 29$

Составим расчётную таблицу

№	Уровни фактора						Итог
	F_1		F_2		F_3		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	9	81	−9	81	−8	64	
2	7	49	−5	25	−7	49	
3	6	36	−3	9	2	4	
4	2	4	1	1	5	25	
$Q_j = \sum y_{ij}^2$		170		116		142	$\sum Q_j = 428$
$T_j = \sum y_{ij}$	24		−16		−8		$\sum T_j = 0$
T_j^2	576		256		64		$\sum T_j^2$

Используя итоговый столбец таблицы, найдем общую и факторную суммы квадратов отклонений, учитывая, что число уровней фактора $p = 3$, число испытаний на каждом уровне $q = 4$:

$$S_{\text{общ}} = \sum_{i=1}^{p} Q_j - \frac{\left[\sum_{j=1}^{p} R_j\right]^2}{pq} = 428 - 0 = 428$$

$$S_{\text{факт}} = \frac{\left[\sum_{j=1}^{p} T_j^2\right]}{q} - \frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq} = \frac{896}{4} - 0 = 224$$

Найдем остаточную сумму квадратов отклонений:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} = 428 - 224 = 204$$

Найдем факторную дисперсию; для этого разделим $S_{\text{факт}}$ на число степеней свободы $p - 1 = 3 - 1 = 2$:

$$s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1} = \frac{224}{2} = 112$$

Найдем остаточную дисперсию; для этого разделим S ост на

число степеней свободы $p(q-1) = 3(4-1) = 9$:

$$s^2_{\text{ост}} = \frac{S_{\text{ост}}}{p(q-1)} = \frac{204}{9} = 22.67$$

Сравним факторную и остаточную дисперсии с помощью критерия Фишера-Снедекора. Для этого сначала найдем наблюдаемое значение критерия:

$$F_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} = \frac{112}{22.67} = 4.94$$

Учитывая, что число степеней свободы числителя $k_1 = 2$, а знаменателя $k_2 = 9$ и что уровень значимости $a = 0.05$, по таблице находим критическую точку

$$F_{\text{кр}}(0.05; 2; 9) = 4.26$$

Так как $F_{\text{набл}} > F_{\text{кр}}$ - нулевую гипотезу о равенстве групповых средних отвергаем. Другими словами, групповые средние «в целом» различаются значимо.

Неодинаковое число испытаний на различных уровнях

Если число испытаний на уровне F_1 равно q_1 , на уровне $F_2 - q_2$, … , на уровне $F_p - q_p$, то общая сумма квадратов отклонений вычисляется, как и в случае одинакового числа испытаний на всех уровнях. Факторная сумма квадратов отклонений находится по формуле

$$S_{\text{факт}} = \frac{T_1^2}{q_1} + \frac{T_2^2}{q_2} + \cdots + \frac{T_p^2}{q_p} - \frac{\sum_{j=1}^{p} T_j^2}{n}$$

Где $n = q_1 + q_2 + \cdots + q_p$ – общее число испытаний.

Остальные вычисления производятся, как и в случае одинакового числа испытаний.

Задача

Дано

Произведено 13 испытаний, из них 4 - на первом уровне фактора, 4 - на втором, 3 - на третьем и 2 – на четвертом. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице

№ испытания i	Уровни фактора			
	F_1	F_2	F_3	F_4
1	1.38	1.41	1.32	1.31
2	1.38	1.42	1.33	1.33
3	1.42	1.44	1.34	-
4	1.42	1.45	-	-
$\overline{x}_{\text{гр}j}$	1.4	1.43	1.33	1.32

Решение

Перейдем к целым числам $y_{ij} = 10^2 x_{ij} - 138$. Составим расчетную таблицу

№	Уровни фактора								Итог
	F_1		F_2		F_3		F_4		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	y_{i4}	y_{i4}^2	
1	0	0	3	9	−6	36	−7	49	
2	0	0	4	16	−5	25	−5	25	

3 4	4 4	16 16	6 7	36 49	−4 –	16 –	– –	– –	
$Q_j = \sum y_{ij}^2$		32		110		77		74	$\sum Q_j = 293$
$T_j = \sum y_{ij}$	8		20		−15		−12		$\sum T_j = 1$
T_j^2	64		400		225		144		

Используя итоговый столбец и нижнюю строку таблицы, найдем общую и факторную суммы квадратов отклонений:

$$S_{\text{общ}} = \sum Q_j - \frac{\left[\sum T_j\right]^2}{n} = 293 - \frac{1^2}{13} = 292.92$$

$$S_{\text{факт}} = 262.92$$

Найдем остаточную сумму квадратов отклонений:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} = 30$$

Найдем факторную и остаточную дисперсии:

$$s^2_{\text{факт}} = 87.64$$

$$s^2_{\text{ост}} = 3.33$$

Сравним факторную и остаточную дисперсии. Для этого сначала вычислим наблюдаемое значение критерия:

$$F_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} = 26.32$$

Учитывая, что число степеней свободы числителя$k_1 = p - 1 = 4 - 1 = 3$, знаменателя $k_2 = n - p = 13 - 4 = 9$и что уровень значимости $\alpha = 0{,}05$, по таблице находим критическую точку $F_{\text{кр}}(0{,}05;\ 3;\ 9) = 3{,}86$.

Так как $F_{\text{набл}} > F_{\text{кр}}$ - нулевую гипотезу о равенстве групповых средних отвергаем. Другими словами, групповые средние различаются значимо.

Семинар 8. Линейная регрессия. Метод наименьших квадратов

Пусть в качестве моделируемого объекта будем рассматривать процесс измерения термометром температуры печи, которая линейно зависит от времени и флуктуирует в области истинного значения по нормальному закону. Необходимо по выполненным измерениям определить зависимость, по которой изменяется температура в печи. При этом ошибка найденной зависимости должна быть минимальной.

В качестве анализируемых данных будем рассматривать n значений температуры, измеренной в градусах, получаемые с использованием термометра и зависящие от времени. Эти данные представим в виде таблицы, где строки – это объекты, первый столбец – набор параметров (j=1), второй столбец – набор измеренных ответов:

Порядковый (i-ый) номер измерения x_i	Оценка температура, C°
x_1	y_1
x_2	y_2
…	…
x_n	y_n

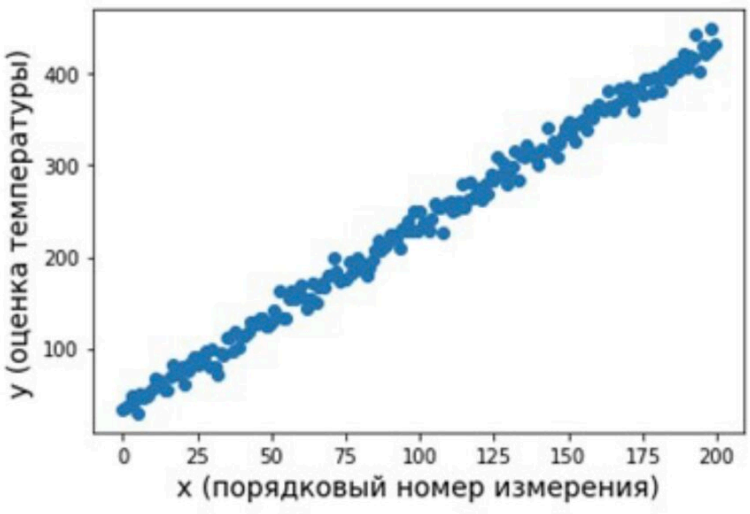
Математически регистрируемый процесс можно описать в виде

$$y_i = w_0 + w_1 x_i + \varepsilon_i,\ i = 1, \dots, n,$$

где w_0 – действительное значение свободного коэффициента модели,
w_1 – действительное значение весового коэффициента модели,
ε_i – белый гауссовский (нормальный) шум,

т.е. случайная величина, распределённая по закону Гаусса с фиксированным значением среднеквадратического отклонения (СКО), нулевым математическим ожиданием (МО) и с автокорреляционной функцией в виде дельта-функции.

Возьмём набор данных, имеющих вид:



Математическая модель

Далее математическая модель будет описана три раза. Последовательно уровень абстракции при описании модели будет понижаться и конкретизироваться.

Первое описание математической модели: самое общее

Поскольку мы рассматриваем процесс, который связывает набор наблюдаемых параметров и измеренных ответов с минимальной ошибкой этой связи, то математическую модель этого процесса в общем виде можно представить

$$\begin{cases} f\left(w^*_{j=\overline{1,u}}, x_{i=\overline{1,n},j=\overline{1,k}}\right) = \hat{y}_{i=\overline{1,n}}\,(w^*, x), \\ w^*_{j=\overline{1,m}} = \underset{w^*}{\operatorname{argmin}}\left[Q\left(\hat{y}_{i=\overline{1,n}}(w^*, x), y_{i=\overline{1,n}}\right)\right]; \end{cases} \quad (2)$$

где $x_{i=\overline{1,n},j=\overline{1,k}}$ – набор из n значений объекта, где в каждый i-ый момент измеряется k его параметров (матрица размерностью $n \times k$),

$i = \overline{1,n}$ – обозначение набора из n индексов вида $i = 1, 2, 3, \ldots, n$,

$w^*_{j=\overline{1,u}}$ – вектор параметров модели размерностью u,

$\hat{y}_{i=\overline{1,n}}(w^*, x)$ – i-ая оценка ответа,

f – функция, которая i-ый набор значений из k параметров объекта, преобразует с использованием вектора параметров w в i-ое значение ответа $\hat{y}_i$,

Q – неотрицательная функция ошибки, оценивающая расстояние между рассчитанным значениями $\hat{y}_{i=\overline{1,n}}$ и измеренным $y_{i=\overline{1,n}}$ на основе выбранной метрики.

Второе описание математической модели: общее описание линейной модели

Поскольку по условию задачи мы рассматриваем линейную зависимость, на которую влияет шум, имеющий нормальное распределение с нулевым средним значением, то используем линейную относительно параметров модель, а метрикой выберем расчёт СКО ошибки

$$\begin{cases} w_0^* + \sum_{j=1}^{m} w_j^* x_{ij} = \hat{y}_i(w^*, x), \\ Q(w^*, x) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i(w^*, x) - y_i)^2 \\ w^*_{j=\overline{1,m}} = \underset{w^*}{\operatorname{argmin}}[Q(w^*, x)] \end{cases} \quad (3)$$

где $x_{i=\overline{1,n},j=\overline{1,m}}$ – набор из n значений объекта, где в каждый i-ый момент измеряется m его параметров (матрица размерностью n×m),

w_0^* – свободный коэффициент модели,

$w^*_{j=\overline{1,m}}$ – вектор весовых коэффициентов модели размерностью m,

$\hat{y}_i(w^*, x)$ – i-ая оценка ответа, полученная с использованием свободного коэффициента w_0 и вектора параметров модели размерностью w^*,

y_i – i-ый измеренный ответ,

Q – неотрицательная функция ошибки, оценивающая расстояние между рассчитанным значениями $\hat{y}_{i=\overline{1,n}}$ и измеренным $y_{i=\overline{1,n}}$ на основе выбранной метрики.

Третье описание математической модели: модель конкретизирована под условия примера

Поскольку по условию задачи мы рассматриваем линейную зависимость одного параметра (поскольку j=1, то далее эту зависимость при описании опустим), на оценку которого влияет шум, имеющий нормальное распределение, то используем линейную модель с двумя коэффициентами

$w^* = \begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix}$, а метрикой выберем расчёт среднеквадратической ошибки

$$\begin{cases} w_0^* + w_1^* x_i = \hat{y}_i(w^*, x), \\ Q(w_0^*, w_1^*, x) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i(w^*, x) - y_i)^2 \\ [w_0^*, w_1^*] = \underset{w_0^*, w_1^*}{\operatorname{argmin}}[Q(w_0^*, w_1^*, x)] \end{cases} \quad (4)$$

где $x_{i=\overline{1,n}}$ – набор из n значений объекта, где в каждый i-ый момент измеряется один его параметр (вектор размерностью n),

w_0^* – свободный коэффициент модели,

w_1^* – весовой коэффициент модели,

$\hat{y}_i(w^*, x)$ – i-ая оценка ответа, полученная с использованием свободного коэффициента w_0^* и параметра модели w_1^*,

y_i – i-ый измеренный ответ,

Q – неотрицательная функция ошибки, оценивающая расстояние между рассчитанным значениями $\hat{y}_{i=\overline{1,n}}$ и измеренным $y_{i=\overline{1,n}}$ на основе выбранной метрики.

Было рассмотрено три модели, описывающих процесс изменения температуры в печи. Первая модель описана в наиболее общем виде, а третья приведена в виде, наиболее соответствующем начальным условиям задачи. В третьей модели учтен линейный закон изменения температуры (используется линейная модель) и нормальный закон распределения шума в измерениях (используется квадратичная функция ошибки).

Хочется отметить, что параметры математической модели процесса изменения температуры, описанной в выражении (2), в принципе, могут находиться не только методом наименьших квадратов, поскольку метрика с квадратичной функцией ошибки в явном виде не прописана. Такая общая форма записи приведена, поскольку не представляется возможным в кратком виде привести все возможные модификации МНК, у которого минимизация для различных условий может выполняться также различно.

В выражениях (3) и (4) линейные модели отличаются только количеством учитываемых параметров, причем третья модель – это редуцированная вторая модель, у которой только два коэффициента (один свободный и один весовой коэффициенты).

Далее с использованием выражения (4) и МНК, синтезируем алгоритмы, оценивающие параметры модели.

Алгоритмы проведения МНК

В общем случае задача нахождения коэффициентов модели (4) заключается в нахождении двух переменных w_0^* и w_1^* при которых функция ошибки Q принимает минимальное значение. То есть при данных w_0^* и w_1^* сумма квадратов отклонений экспериментальных данных от модельных (расчетных) значений будет наименьшей, собственно, поэтому метод наименьших квадратов и носит своё имя.

Таким образом, нахождение коэффициентов w_0^* и w_1^* сводится к нахождению экстремума функции двух переменных

$$Q(w_0^*, w_1^*, x) = \frac{1}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i)^2.$$

Для этого необходимо

– найти частные производные функции $Q(w_0^*, w_1^*, x)$ по искомым w_0^*, w_1^*;

– приравнять полученные выражения к нулю;

– решить полученную систему из двух уравнений с двумя неизвестными.

Решение полученной системы уравнений и будет являться искомыми параметрами модели.

Опишем аналитическое решение этой системы детально. Выполним расчет производной функции $Q(w_0^*, w_1^*, x)$ по параметру w_0^*.

Первый алгоритм

Начнем с применения правила дифференцирования функции представленной в виде суммы

$$\frac{\partial Q(w_0^*, w_1^*, x)}{\partial w_0^*} = \frac{\partial \left(\frac{1}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i)^2\right)}{\partial w_0^*} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial (w_0^* + w_1^* x_i - y_i)^2}{\partial w_0^*} =$$

применим правило дифференцирования сложной функции

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial (w_0^* + w_1^* x_i - y_i)^2}{\partial (w_0^* + w_1^* x_i - y_i)} \cdot \frac{\partial (w_0^* + w_1^* x_i - y_i)}{\partial w_0^*} =$$

применим правило дифференцирования функции представленной в виде суммы

$$= \frac{1}{n}\sum_{i=1}^{n} 2 \cdot (w_0^* + w_1^* x_i - y_i) \cdot \left(\frac{\partial w_0^*}{\partial w_0^*} + \frac{\partial (w_1^* x_i)}{\partial w_0^*} - \frac{\partial y_i}{\partial w_0^*}\right) =$$

$$= \frac{2}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i) \cdot (1 + 0 - 0) = \frac{2}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i).$$

Выполним расчет производной функции $Q(w_0^*, w_1^*, x)$ по параметру w_1^* и применим правило дифференцирования функции представленной в виде суммы

$$\frac{\partial Q(w_0^*, w_1^*, x)}{\partial w_1^*} = \frac{\partial \left(\frac{1}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i)^2\right)}{\partial w_1^*} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial (w_0^* + w_1^* x_i - y_i)^2}{\partial w_1^*} =$$

применим правило нахождения производной сложной функции

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial (w_0^* + w_1^* x_i - y_i)^2}{\partial (w_0^* + w_1^* x_i - y_i)} \cdot \frac{\partial (w_0^* + w_1^* x_i - y_i)}{\partial w_1^*} =$$

применим правило дифференцирования функции представленной в виде суммы

$$= \frac{1}{n}\sum_{i=1}^{n} 2 \cdot (w_0^* + w_1^* x_i - y_i) \cdot \left(\frac{\partial w_0^*}{\partial w_1^*} + \frac{\partial (w_1^* x_i)}{\partial w_1^*} - \frac{\partial y_i}{\partial w_1^*}\right) =$$

$$= \frac{2}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i) \cdot (0 + x_i - 0) = \frac{2}{n}\sum_{i=1}^{n}\left((w_0^* + w_1^* x_i - y_i) \cdot x_i\right).$$

Приравняем полученные производные к нулю и решим полученную систему уравнений

$$\begin{cases} \frac{2}{n}\sum_{i=1}^{n}(w_0^* + w_1^* x_i - y_i) = 0, \\ \frac{2}{n}\sum_{i=1}^{n}\left((w_0^* + w_1^* x_i - y_i)\cdot x_i\right) = 0. \end{cases}$$

Раскроем скобки

$$\begin{cases} \sum_{i=1}^{n} w_0^* + \sum_{i=1}^{n} w_1^* x_i - \sum_{i=1}^{n} y_i = 0, \\ \sum_{i=1}^{n} w_0^* x_i + \sum_{i=1}^{n} w_1^* x_i^2 - \sum_{i=1}^{n} y_i x_i = 0. \end{cases}$$

Вынесем постоянные множители за скобки

$$\begin{cases} n \cdot w_0^* + w_1^* \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0, \\ w_0^* \sum_{i=1}^{n} x_i + w_1^* \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i = 0. \end{cases}$$

Вынесем слагаемые с множителем «y» в правую часть уравнений

$$\begin{cases} n \cdot w_0^* + w_1^* \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ w_0^* \sum_{i=1}^{n} x_i + w_1^* \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i. \end{cases}$$

Поставим слагаемые с множителем «x» в левой части в порядке убывания степеней

$$\begin{cases} w_1^* \sum_{i=1}^{n} x_i^2 + w_0^* \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i , \\ w_1^* \sum_{i=1}^{n} x_i + n \cdot w_0^* = \sum_{i=1}^{n} y_i . \end{cases}$$

Для решения полученной системы алгебраических уравнения представим её в матричной форме

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i x_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}.$$

Выразим вектор w* с искомыми весами выполнив умножение обеих частей равенства на обратную матрицу у

$$\begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i x_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}. \qquad (5)$$

Полученное выражение (5) является решением системы уравнений и его можно уже использовать в качестве первого алгоритма оценки параметров модели.

Второй алгоритм

Выражение (5) можно упростить, выполнив аналитический расчет обратной матрицы.

Найти обратную матрицу можно с использованием, например, алгебраических дополнений. Для пояснения поиска обратной матрицы введем новую переменную А. Пусть

$$A = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix}, \text{ тогда } A^{-1} = \frac{1}{\Delta} \cdot (B)^T,$$

где $\Delta = A_{11}A_{22} - A_{12}A_{21}$ – определитель матрицы A,

$B = \begin{pmatrix} A_{22} & -A_{21} \\ -A_{12} & A_{11} \end{pmatrix}$ – матрица алгебраических дополнений, составленная из миноров $M = \begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix}$ матрицы A,

$(*)^T$ – транспонирование матрицы, указанной в скобках.

Итоговое выражение для обратной матрицы представим в виде

$$A^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{pmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{pmatrix}.$$

Перепишем выражение (5) в виде

$$\begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix} = \frac{1}{n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i} \begin{pmatrix} n & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i x_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} =$$

Упростим выражение раскрыв скобки, чтобы получить более компактную форму записи

$$= \frac{1}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} n \cdot \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i \\ -\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{n \cdot \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \frac{-\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{pmatrix}.$$

Теперь решение системы уравнений имеет вид

$$\begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix} = \begin{pmatrix} \frac{n \cdot \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \frac{-\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{pmatrix}. \qquad (6)$$

Полученное выражение (6) можно использовать в качестве второго алгоритма оценки параметров модели.

Третий алгоритм

Выполним дальнейшее упрощение полученного выражения (6), чтобы получить более компактную форму записи.

Умножим числитель и знаменатель каждого элемента матрицы на $\frac{1}{n^2}$

$$\begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix} = \begin{pmatrix} \frac{(n \cdot \sum_{i=1}^n y_i x_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i) \cdot \frac{1}{n^2}}{(n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2) \cdot \frac{1}{n^2}} \\ \frac{(-\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i x_i + \sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i) \cdot \frac{1}{n^2}}{(n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2) \cdot \frac{1}{n^2}} \end{pmatrix} =$$

раскроем скобки:

$$= \begin{pmatrix} \frac{\frac{1}{n} \cdot \sum_{i=1}^n y_i x_i - \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i\right) \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i\right)}{\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i\right)^2} \\ \frac{-\left(\frac{1}{n} \cdot \sum_{i=1}^n x_i\right) \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i x_i\right) + \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i^2\right) \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i\right)}{\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i\right)^2} \end{pmatrix}.$$

Ведем новые переменные используя термины математической статистики:

1) $\frac{1}{n} \cdot \sum_{i=1}^n x_i = \bar{x}$ – оценка МО величины x;
2) $\frac{1}{n} \cdot \sum_{i=1}^n y_i = \bar{y}$ – оценка МО величины y;
3) $\frac{1}{n} \cdot \sum_{i=1}^n y_i x_i = \overline{xy}$ – оценка МО произведения величин x и y;
4) $\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 = \overline{x^2}$ – оценка МО величины x^2;
5) $\sqrt{\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i\right)^2} = \sqrt{\overline{x^2} - \bar{x}^2} = \sigma_x$ – оценка СКО величины x;
6) $\sqrt{\frac{1}{n} \cdot \sum_{i=1}^n y_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i\right)^2} = \sigma_y$ – оценка СКО величины y;
7) $\frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} = r_{xy}$ – оценка линейного коэффициента корреляции величин x и y.

С учётом введённых переменных искомый вектор w* примет вид

$$\begin{pmatrix} w_1^* \\ w_0^* \end{pmatrix} = \begin{pmatrix} \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2} \\ \dfrac{-\bar{x} \cdot \overline{xy} + \overline{x^2} \cdot \bar{y}}{\sigma_x^2} \end{pmatrix}.$$

Упростим полученные выражения. Домножим весовой коэффициент w_1^* на $1 = \frac{\sigma_y}{\sigma_y}$.

$$w_1^* = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2} \cdot \frac{\sigma_y}{\sigma_y} = \left(\frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} \right) \cdot \frac{\sigma_y}{\sigma_x} = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$$

Прибавим к числителю w_0^* свободного коэффициента $0 = -\bar{x}^2 + \bar{x}^2$

$$w_0^* = \frac{-\bar{x} \cdot \overline{xy} + \overline{x^2} \cdot \bar{y}}{\sigma_x^2} = \frac{-\bar{x} \cdot \overline{xy} + \left(\overline{x^2} - \bar{x}^2 + \bar{x}^2 \right) \cdot \bar{y}}{\sigma_x^2} = \frac{-\bar{x} \cdot \overline{xy} + (\sigma_x^2 + \bar{x}^2) \cdot \bar{y}}{\sigma_x^2} =$$

$$= \frac{-\bar{x} \cdot \overline{xy} + \sigma_x^2 \cdot \bar{y} + \bar{x}^2 \cdot \bar{y}}{\sigma_x^2} = \frac{\sigma_x^2 \cdot \bar{y} - (\bar{x} \cdot \overline{xy} - \bar{x}^2 \cdot \bar{y})}{\sigma_x^2} = \bar{y} - \frac{\bar{x} \cdot \overline{xy} - \bar{x}^2 \cdot \bar{y}}{\sigma_x^2} =$$

$$= \bar{y} - \frac{\bar{x} \cdot (\overline{xy} - \bar{x} \cdot \bar{y})}{\sigma_x^2} \cdot \frac{\sigma_y}{\sigma_y} = \bar{y} - \frac{\bar{x} \cdot (\overline{xy} - \bar{x} \cdot \bar{y})}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \bar{y} - \bar{x} \cdot r_{xy} \cdot \frac{\sigma_y}{\sigma_x} = \bar{y} - \bar{x} \cdot w_1^*$$

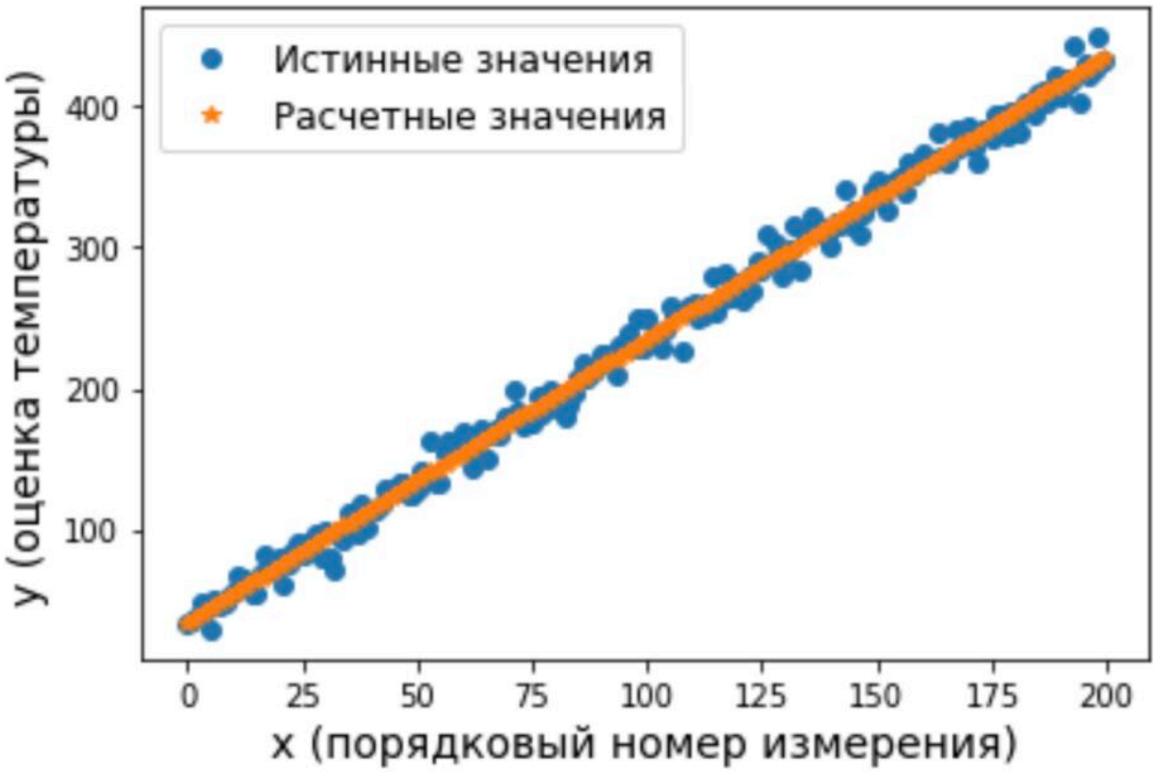
Таким образом итоговое выражение представим в виде:

$$\begin{cases} w_1^* = r_{xy} \cdot \dfrac{\sigma_y}{\sigma_x}, \\ w_0^* = \bar{y} - \bar{x} \cdot w_1^*. \end{cases} \qquad (7)$$

Полученное выражение (7) можно использовать в качестве третьего алгоритма оценки параметров модели.

Визуализация результатов после вычисления

Итак, визуализируем полученный результат, построив прямую с использованием рассчитанных коэффициентов и сопоставим их с исходными данными



Таким образом удалось выполнить оценку параметров модели, которая описывает процесс изменения температуры в печи с минимальной ошибкой. Выведенные выражения (5) – (7) и запрограммированные по ним три алгоритма, реализующие МНК при обработке одного набора измеренных температур, хотя и отличаются видом конечных выражений, дают одинаковые оценки.

Семинар 9. Временные ряды

Понятие временных рядов

Временной ряд — значения меняющихся во времени признаков, полученные в некоторые моменты времени.

Используются временные ряды для аналитики и прогнозирования, когда важно предсказать, что произойдёт с показателями в ближайшее время, будь то ближайший час, день, месяц или год.

Типы временных рядов

1) Детерминированный – ряд, в котором нет случайных аспектов/показателей, т.е. может быть выражен формулой.

Это означает наличие возможности прогнозирования поведения показателей в прошлом и предсказание их поведения в будущем

2) Недетерминированный – ряд имеет случайный аспект, что делает прогнозирование будущего сложнее. Природа таких показателей случайна и анализ производится посредством средних значений и дисперсии

Стационарные и нестационарные ряды

На наблюдение за показателями и их систематизацией влияют тенденции и сезонные эффекты. От этих условий зависит сложность моделирования системы прогнозирования. Временные ряды делятся по наличию или отсутствию тенденций и сезонных эффектов на стационарные и нестационарные.

Стационарный временной ряд – статистические свойства не зависят от времени => результат легко предсказываем

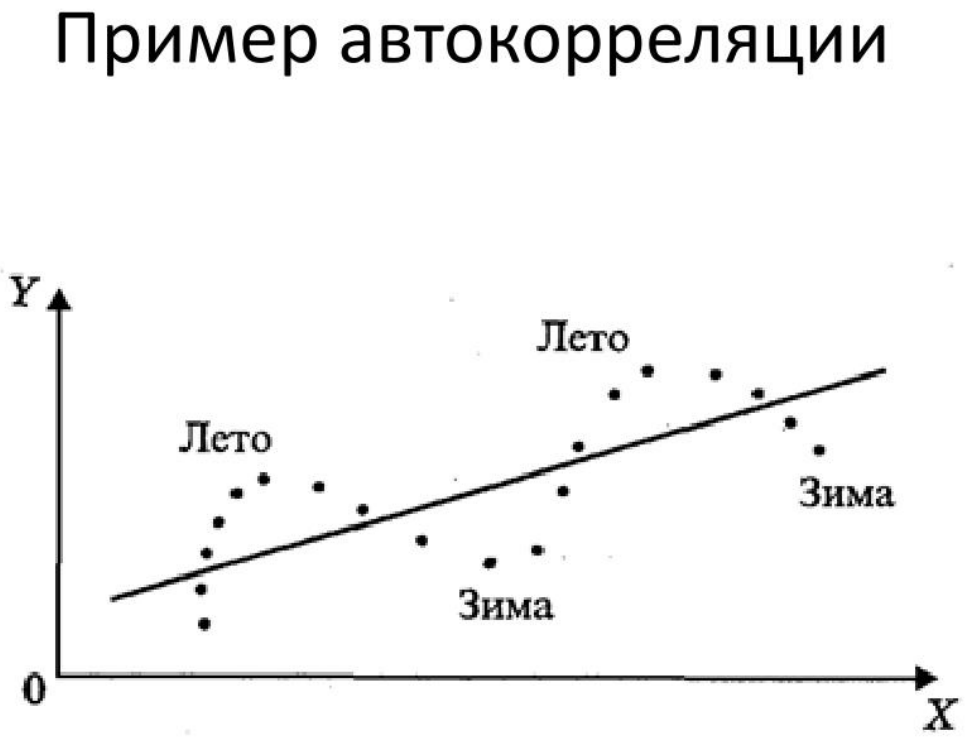
Нестационарный временной ряд – статистические свойства меняются со временем. Они показывают сезонные эффекты, тренды и другие структуры, зависящие от времени.

Для классических статистических методов удобнее создавать модели стационарных временных рядов. Если прослеживается четкая тенденция или сезонность во временных рядах, то следует смоделировать эти компоненты и удалить их из наблюдений. Из наблюдений удаляют «шум» – дополнительный компонент, который мешает усреднению данных.

Анализ временных рядов

Один из способов проверки стационарности временных рядов - это использование графических методов, таких как график временных рядов, график автокорреляции и график частной автокорреляции. График временного ряда позволяет визуализировать изменения значений ряда с течением времени, график автокорреляции показывает корреляцию между значениями ряда в разные периоды времени, а график частной автокорреляции учитывает корреляцию только между двумя значениями, пропуская все промежуточные значения.

Автокорреляция - это мера корреляции между значениями ряда с разницей во времени. Если временной ряд имеет высокую автокорреляцию, это означает, что значения ряда в разные периоды времени имеют сильную связь между собой. Автокорреляцию можно вычислить с помощью функции корреляции Пирсона, которая вычисляет корреляцию между двумя переменными. Для временных рядов это означает вычисление корреляции между значениями ряда в разные периоды времени. Если автокорреляция является значимой, то это может означать наличие тренда, цикличности или сезонности в ряде.



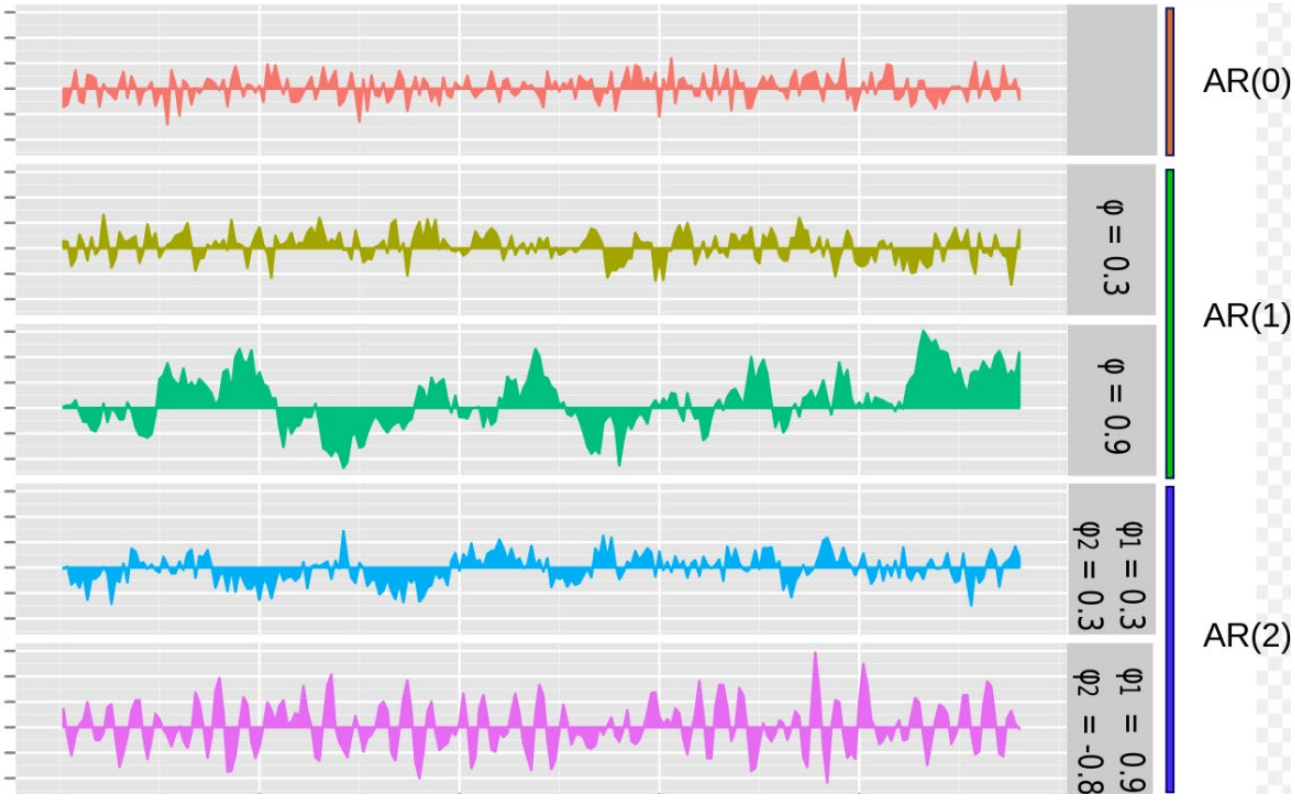
Спектральный анализ позволяет исследовать частотную составляющую ряда. Он используется для выявления скрытых цикличных паттернов во временных рядах. Для этого временной ряд разбивается на сигналы разных частот, а затем анализируется спектр этих частот. Спектральный анализ часто используется для анализа финансовых рынков, в которых цены могут изменяться в зависимости от времени и частоты.

Каждый метод имеет свои преимущества и недостатки, и их выбор зависит от конкретной области применения.

Моделирование временных рядов: авторегрессионные модели, скользящее среднее, ARIMA, SARIMA

Моделирование временных рядов является важным инструментом для прогнозирования будущих значений ряда, что позволяет принимать более обоснованные решения в различных областях.

Авторегрессионные модели (AR-модели) используют прошлые значения ряда для прогнозирования его будущих значений. Эта модель предполагает, что текущее значение ряда зависит от его предыдущих значений. Наиболее популярной AR-моделью является модель первого порядка (AR(1)), которая предполагает, что текущее значение ряда зависит только от его предыдущего значения.



Примером использования AR-модели может служить анализ финансовых рынков. Например, с помощью AR-модели можно прогнозировать будущие цены акций на основе их прошлых значений. Если взять за основу прошлые цены и применить AR-модель, можно определить вероятные цены акций в будущем.

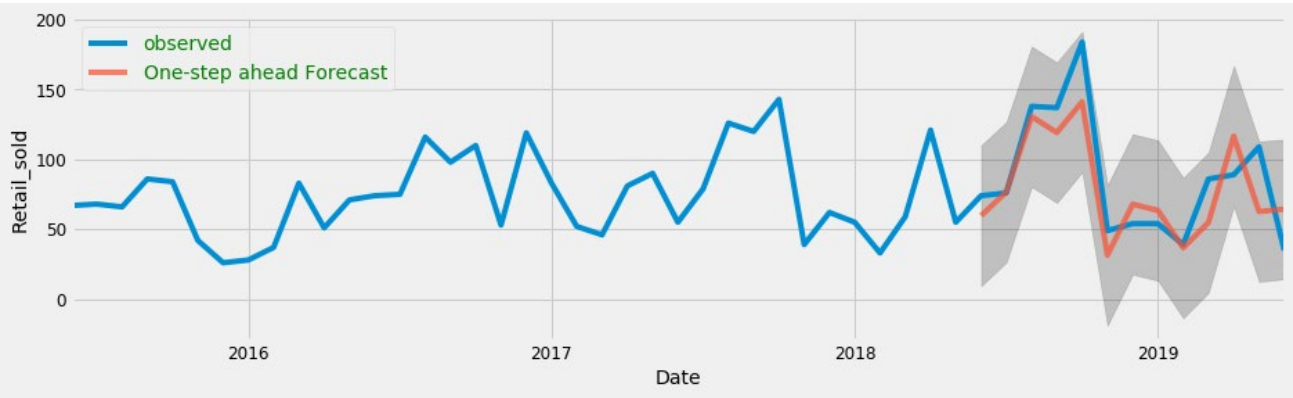
Следующий метод моделирования временных рядов - скользящее среднее (MA-модели). Эта модель использует прошлые значения ошибок - разницу между фактическими значениями ряда и его прогнозируемыми значениями, для прогнозирования будущих значений. Наиболее популярной MA-моделью является модель первого порядка (MA(1)), которая предполагает, что текущая ошибка зависит только от ее предыдущего значения.

Примером использования MA-модели может служить прогнозирование количества пользователей веб-сайта. Например, если предыдущие прогнозы ошиблись на определенный процент, можно использовать эту информацию в модели прогнозирования для улучшения точности прогноза количества пользователей.

ARIMA - это модель, комбинирующая авторегрессионные и скользящие средние модели. ARIMA позволяет моделировать данные, не являющиеся стационарными, как это не требуется для AR- и MA-моделей. ARIMA включает три параметра: параметр авторегрессии (p), параметр скользящего среднего (q) и параметр интегрирования (d).

Примером использования ARIMA может служить прогнозирование месячной выручки продукта на основе ежемесячных данных за прошлый год. Если прошлые данные имеют тренд, сезонность или циклы, можно использовать ARIMA для учета этих факторов в прогнозировании выручки.

SARIMA - это модель, комбинирующая ARIMA с сезонностью. SARIMA используется для моделирования сезонного поведения временных рядов. SARIMA - это расширение ARIMA с тремя дополнительными параметрами, определяющими сезонность: период сезонности (P), параметр авторегрессии со сезонностью (S) и параметр скользящего среднего со сезонностью (Q).



Примером использования SARIMA может служить прогнозирование продаж в интернет-магазине. Если продажи показывают сезонность, например, рост продаж в преддверии праздников, можно использовать SARIMA для прогнозирования продаж в будущих сезонах.