

Лекция 15

- 6. Сглаживание экспериментальных зависимостей
 - 6.1. Метод наименьших квадратов
 - 6.1.1. Теоретическое обоснование метода наименьших квадратов
- 7. Проверка статистических гипотез
 - 7.1. Критерий согласия Пирсона («хи – квадрат»)
 - 7.1.1. Схема применения критерия χ^2
 - 7.1.2. Обоснование выбора весовых коэффициентов

6. Сглаживание экспериментальных зависимостей

При измерениях числовых характеристик случайных величин получают их оценки и, таким образом, наверняка имеется статистическая ошибка измерений, что связано с конечным объемом выборки.

При обработке возникает вопрос подбора кривой распределения, отражающей наиболее существенные черты экспериментального материала. Такая задача называется **задачей выравнивания или сглаживания статистических рядов**.

К этой задаче примыкает другая: **сглаживание полученной экспериментальной зависимости**. Пусть производится эксперимент с целью выявления вида функциональной зависимости:

$$y = f(x).$$

Результаты опыта, естественно, содержат ошибки измерений, которые могут иметь случайный характер. Задача сводится к подбору кривой распределения, отражающей наиболее существенные черты экспериментального материала. В решении этой задачи важное значение приобретает уровень понимания физики наблюдаемого явления, потому как "сглаживая" можно потерять очень существенный результат уникального эксперимента. Поэтому

критерии для проведения выравнивания и сглаживания, к сожалению, нельзя сформулировать для общего случая. Как правило применяют метод наименьших квадратов.

Пример : Выровнять (сгладить) результаты измерений, считая, что они имеют нормальный закон распределения.

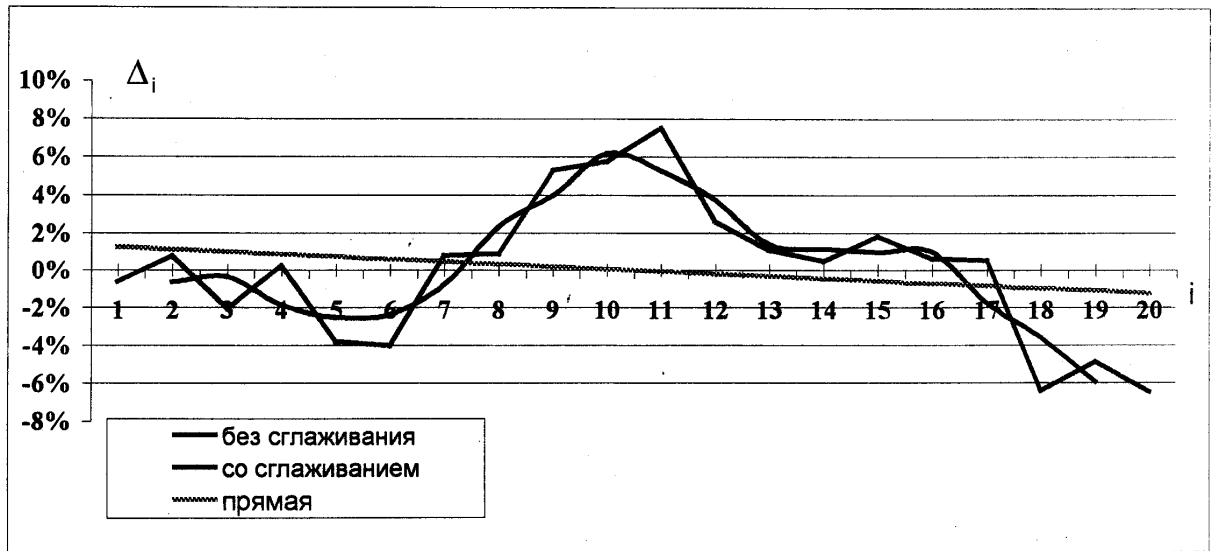
В качестве числовых параметров m_x и D_x используют их оценки m_x^* и D_x^* .

Разряд	-3,-2	-2,-1	-1,0	0,1	1,2	2,3
Частота	n_1	n_2	n_3	n_4	n_5	n_6
$W=p^*$	n_1/n	n_2/n	n_3/n	n_4/n	n_5/n	n_6/n
$P_{\text{норм}}$	$p(-2.5)=0,0175$	$p(-1.5)=0,1295$	$p(-0.5)=0,3521$	$p(0.5)=0,3521$	$p(1.5)=0,1295$	$p(2.5)=0,0175$

$$m^* = \bar{x}_1 \cdot W_1 + \bar{x}_2 \cdot W_2 + \dots; \quad \bar{x}_i = \frac{x_{i-1} + x_i}{2}; \quad D^* = \sum x_i^2 \cdot W_i - (m^*)^2$$

$$f(x_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} e^{-\frac{(x_i - m_x^*)^2}{2\sigma_x^2}} = p(\bar{x}_i).$$

В качестве примера второй задачи возьмем расчетно-графическое задание, в котором результаты эксперимента - уровень шума как функция положения на развертке дальности - представлены в виде простого статистического ряда.



Надо получить сглаженную зависимость уровня шума от положения на развертке. Если неизвестен вид зависимости, то применяют один из методов сглаживания, который выбирают, исходя из представленного экспериментального материала. Например, сглаживание по алгоритму:

$$\bar{\Delta}_{i+1} = \frac{\Delta_i + \Delta_{i+1} + \Delta_{i+2}}{3}.$$

6.1. Метод наименьших квадратов

Другой подход к сглаживанию экспериментальных данных заключается в предположении (гипотезе), что их изменение (по оси абсцисс) описывается некоторой функцией $\Delta(i)$, на которую накладывается случайная ошибка измерений, (см. "Методические указания к РГЗ по курсу ТВ и МС" п. 2.5), т.е.

$$\Delta_i = \Delta(i) + \delta_i \quad \text{или} \quad y_x = y(x) + \delta_x$$

Например, если $y(x)$ - это полином второй степени $y = ax^2 + bx + c$, то требуется найти те значения a , b и c , при которых полином **наилучшим образом** описывает экспериментальные данные. Предложение "наилучшим образом" весьма неопределенно, что вынуждает взвешенно подходить к каждому конкретному случаю. Если же каких-либо альтернативных подходов нет, в практике широко

используется **метод наименьших квадратов** (МНК). Согласно МНК требование наилучшего согласования кривой $y = f(x)$ сводится к тому, чтобы сумма квадратов отклонений экспериментальных точек от сглаживающей кривой обращалась в минимум.

МНК имеет три основных момента в применении

1. Необходимо иметь аналитическое выражение для представленного вида зависимости: **в общем случае** $y = y(x; a_1, a_2, \dots, a_n)$, **в частном случае это может быть уравнение прямой:** $y = kx + b$.
2. Должен быть обозначен метод поиска этой кривой;
3. В качестве критерия для принятия решения используется выражение:

$$L = \sum_{i=1}^n (y_i - y(x_i; a_1, a_2, \dots, a_n))^2 \rightarrow \min.$$

Например, если используется уравнение прямой, то минимум суммы соответствует нулю второй смешанной производной по параметрам k и b . Эта смешанная производная может быть представлена, в данном случае, системой двух уравнений.

В общем случае количество уравнений равно количеству параметров:

$$y = y(x; a_1, a_2, \dots, a_k);$$

$$\frac{\partial}{\partial a_j} \left\{ \sum_{i=1}^n (y_i - y(x_i; a_1, a_2, \dots, a_n))^2 \right\} = 0, \quad j = 1, 2, \dots, k;$$

$$\sum_{i=1}^n \left\{ (y_i - y(x_i; a_1, a_2, \dots, a_n)) \cdot \left(\frac{\partial y}{\partial a_j} \right)_i \right\} = 0;$$

Рассмотрим основания для применения тех формул, которые были использованы в расчетно-графическом задании

$$\Delta_i = a_1 x_i - m; \quad m = \frac{1+n}{2}; \quad n = 20; \quad x_i = 1, 2, \dots, n;$$

$$\frac{\partial}{\partial a} \left\{ \sum_{i=1}^n (y_i - a x_i - m)^2 \right\} = 0; \quad 2 \sum_{i=1}^n (y_i - a x_i - m) x_i = 0;$$

$$\sum_{i=1}^n y_i \cdot x_i - m \underbrace{\sum_{i=1}^n y_i}_0 - a \sum_{i=1}^n x_i = 0, \quad y_i = \Delta_i;$$

$$a = \frac{\Delta_1 \cdot 1 + \Delta_2 \cdot 2 + \dots + \Delta_n \cdot n}{\sum_{i=1}^n x_i} = \frac{\Delta_1 \cdot 1 + \Delta_2 \cdot 2 + \dots + \Delta_n \cdot n}{\sum_{i=1}^n x_i} = \frac{\Delta_1 \cdot 1 + \Delta_2 \cdot 2 + \dots + \Delta_n \cdot n}{\frac{1}{n} \sum_{i=1}^n x_i^2 - m^2}.$$

6.1.1. Теоретическое обоснование метода наименьших квадратов

Найденное по методу наименьших квадратов теоретическое выражение соответствует наиболее вероятному распределению ошибок, если их распределение подчиняется нормальному закону

Пусть $y_i = y_{\text{т}} + \delta_i$, где $\delta_1, \delta_2, \dots, \delta_n$ - конкретные значения, которые в результате эксперимента приняла случайная величина Δ , распределённая по нормальному закону с параметрами: $D = \sigma^2$ и $m_{\Delta} = 0$. Если считать, что значения $\delta_1, \delta_2, \dots, \delta_n$ независимы, то вероятность именно такого набора (событие А) определяется произведением:

$$P(A) = p_{\delta_1} \cdot p_{\delta_2} \cdot \dots \cdot p_{\delta_n};$$

$$p_{\delta_i} = f_{\Delta}(\delta_i) d\delta; \quad f_{\Delta} = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{\delta^2}{2\sigma^2}}; \quad P(A) = \left(\frac{1}{\sqrt{2\pi\sigma}} \right)^n e^{-\frac{1}{2\sigma^2}(\delta_1^2 + \delta_2^2 + \dots + \delta_n^2)} d\delta_n;$$

$$P_{\max}(A) \rightarrow \sum_{i=1}^n (y_i - y_{\text{т}})^2 = \min.$$

7. Проверка статистических гипотез

Статистической называют гипотезу о виде распределения исследуемой случайной величины или гипотезу о параметрах известных распределений.

Возможны и другие гипотезы, например о равенстве параметров двух или нескольких распределений, о независимости выборок

Простой называют гипотезу содержащую только одно предположение, сложная гипотеза состоит из нескольких простых. Пример простой гипотезы: "Ошибка измерения случайной величины X имеет нормальный закон" или "Ошибка измерения случайной величины X имеет равномерный закон". Это две конкурирующие гипотезы. Основную гипотезу называют нулевой (H_0), а другую – альтернативной (H_1).

Как бы хорошо при сглаживании не была подобрана теоретическая кривая всегда существуют расхождения. Возникает вопрос: чем объясняется это расхождение? Только лишь действием случайных причин или неправильно установлен (выбран) теоретический закон. Критерии, которые позволяют установить степень соответствия (согласия) выбранного закона эксперименту носят название **критериев согласия**. Идея их применения состоит в следующем.

Пусть на базе имеющегося статистического материала предстоит проверить гипотезу: **случайная величина распределена по закону $f(x)$** . Проверка гипотез состоит в доказательстве согласия между выдвинутой гипотезой и результатами эксперимента.

В МНК **сумма квадратов отклонений** теоретических значений функции $f(x)$ от результатов, полученных при экспериментах, **характеризует степень несогласия** результатов эксперимента с предложенной кривой.

Предположим, что произведено n независимых опытов (например, $n = 20$), и в результате получены оценки вероятностей попадания случайной величины X в заданные интервалы значений

	Разряд	-3; -2	-2; -1	-1; 0	0; 1	1; 2	2; 3
	Частота	$n_1=0,5$	$n_2=2,5$	$n_3=7$	$n_4=7$	$n_5=2,5$	$n_6=0,5$
	p^*	0,5/20	2,5/20	7/20	7/20	2,5/20	0,5/20
H_0	Нормальный	0,018	0,14	0,34	0,34	0,14	0,018
H_1	Равномерный.	1/6	1/6	1/6	1/6	1/6	1/6

$$x_i = \frac{R_i - \bar{R}}{\text{СКО}}$$

Если использовать только МНК, то можно получить:

$$L = \sum_{i=1}^6 p_i^* - p_i \Rightarrow \min. \quad *)$$

Однако остаётся нерешенным вопрос, является ли отличие L от нуля следствием только случайных причин, или неверно подобран закон распределения.

7.1. Критерий согласия Пирсона ("хи – квадрат")

Действительно, выражение *) не может быть использовано в качестве универсального критерия согласия и вот почему :

сумма квадратов отклонений зависит от объёма выборки n .

сумма квадратов отклонений зависит от вида закона распределения.

сумма квадратов отклонений зависит от количества разрядов, на которые мы разбили интервал возможных значений случайной величины.

$p_i^* - p_i$ тем меньше, чем больше n .

Если значение теоретической вероятности мало, то малы и разности. Следовательно, соответствующие им слагаемые в сумме имеют малый вес. Пирсон модифицировал выражение *) введением корректирующих весовых коэффициентов:

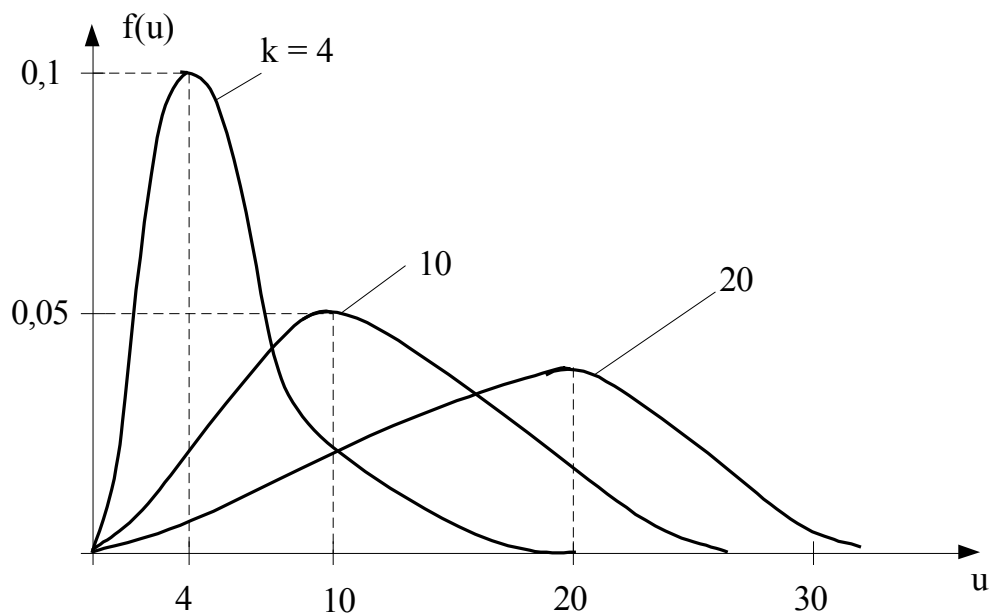
$$U = \chi^2 = \sum_{i=1}^k C_i (p_i^* - p_i)^2; \quad C_i = n \cdot \frac{1}{p_i}$$

Величина U – это случайная величина, которая является суммой квадратов случайных величин, и каждая из них имеет нормальный закон распределения. U имеет распределение

$$f(u) = \frac{1}{2^t \Gamma(t)} \cdot u^{t-1} \cdot e^{-\frac{u}{2}}; \quad t = \frac{k-s}{2},$$

где k – количество разрядов, s – это количество использованных уравнений, в которые входит вероятность. Для РГЗ $s=3$;

$$\sum_{i=1}^{k=6} p_i = 1; \quad m^* = \frac{1}{n} \sum_{i=1}^n R_i; \quad \text{СКО.}$$



$$\Gamma(t) = \int_0^{\infty} x^{t-1} \cdot e^{-x} dx; \quad t = 1, 2, 3, \dots;$$

$$\Gamma(t) = (t-1)!.$$

7.1.1. Схема применения критерия χ^2

1. Определяется мера расхождения

$$U = \chi^2 = \sum_{i=1}^k \frac{n \cdot (p_i^* - p_i)^2}{p_i}.$$

2. Определяется число степеней свободы

$$r = k - s.$$

3. По таблицам находят значение вероятности того, что в силу случайных причин U может превзойти вычисленное в пункте 1 значение.

$r \backslash p$	0,99	0,9	0,5	0,1	0,01
1	0,000	0,016	0,45	2,71	6,64
3	0,115	0,584	2,37	6,25	11,34
10	2,56	4,8	9,34	16,0	23,2
20	8,3	12,4	19,34	28,4	37,6
30	14,9	20,6	29,3	40,3	50,9

Если вероятность менее 0.1, рекомендуется проверить расчёты. Если есть возможность – повторить эксперимент. Если результат повторится, то появляется основание признать, что результаты эксперимента противоречат гипотезе.

7.1.2. Обоснование выбора весовых коэффициентов

Согласно критерию Пирсона численной мерой согласия теоретического закона распределения с экспериментальными данными служит сумма

$$U = \chi^2 = \sum_{i=1}^k \frac{n \cdot p_i^* - p_i^2}{p_i},$$

где k – количество разрядов, n – объём выборки, p_i^* – относительная частота появления величины в данном разряде, p_i – теоретическое значение вероятности появления величины в данном разряде.

Покажем, что эта сумма не зависит от n и закона распределения. Выделим i -й разряд. Назовём событие A – попадание результатов в этот разряд. Вероятность такого события описывается биномиальным распределением. И пусть случайная величина X – число попаданий в данный разряд

$$X = np^*.$$

Тогда $m_x = np$; $D_x = npq$. Найдём

$$M \left[\frac{n p_i^* - p_i^2}{p_i \cdot n} \right] = \frac{1}{np} M \left[n p_i^* - np^2 \right] = \frac{D_x}{np} = \frac{npq}{np} = 1 - p$$

$$M \chi^2 = M \left[\sum_{i=1}^k \frac{n p_i^* - p_i^2}{p_i} \right] = M \left[\sum_{i=1}^k 1 - p_i \right] = k - 1.$$

При выводе этого выражения мы считали, что эксперименты независимы. На самом деле, в РГЗ при обработке результатов мы использовали 3 уравнения, где участвовало p . С учётом этого математическое ожидание

$$M \chi^2 \cong k - s - 0,7,$$

где s – количество уравнений, равное 3.

Таким образом, можем прийти к итогу:

Если $\chi^2 < 2,3$, то результаты эксперимента скорее согласуются с теоретическим законом.

Если $\chi^2 > 2,3$, то скорее не согласуются.