Курс: Статистические Методы Обработки Данных

Лекция 4. Идентификация формы распределений

Специальность: 1-53 01 02 – Автоматизированные системы обработки информации

УО «ГГУ им. Ф. Скорины»

Преподаватель: Бабич К.С, ст. преподаватель, 2016

Раздел 2 – Гистограммы и Операции над ними

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Раздел 2 – Гистограммы и Операции над ними

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.1 Наиболее распространенным является критерий согласия Пирсона (χ2 - критерий).

Суть χ2 - критерия состоит в вычислении величины:

$$\chi^2_{\acute{y}\hat{e}} = n\sum_{j=1}^{m}\frac{\left(\nu_j - p_j\right)^2}{p_j}$$

где m – число интервалов,

ν_j - частота попадания экспериментальных данных в j-й интервал,

p_j - вероятности в том же j-м столбце, рассчитанном по заданной модели.

Если бы наша выбранная модель во всех центрах столбцов

$$\nu_j = p_j \;\Rightarrow\; \chi^2 = 0$$

т.о. хи-квадрат – мера суммарного отклонения между модельным и экспериментальным распределением.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Величина χ2 получила своё обозначение не случайно, т.к. она подчиняется распределению χ2 с k = m – 1 – r степенями свободы.

где r – число параметров модели, определяемых по экспериментальным данным, необходимых для совмещения модели и гистограммы.

Критерий χ2 :

$$\chi^2 \le \chi^2_{p,k}$$

$p = 1 - \alpha$ -доверительная вероятность,
или вероятность <u>принятия</u> модели.
то гипотеза о распределении при данном уровне значимости не противоречит модели

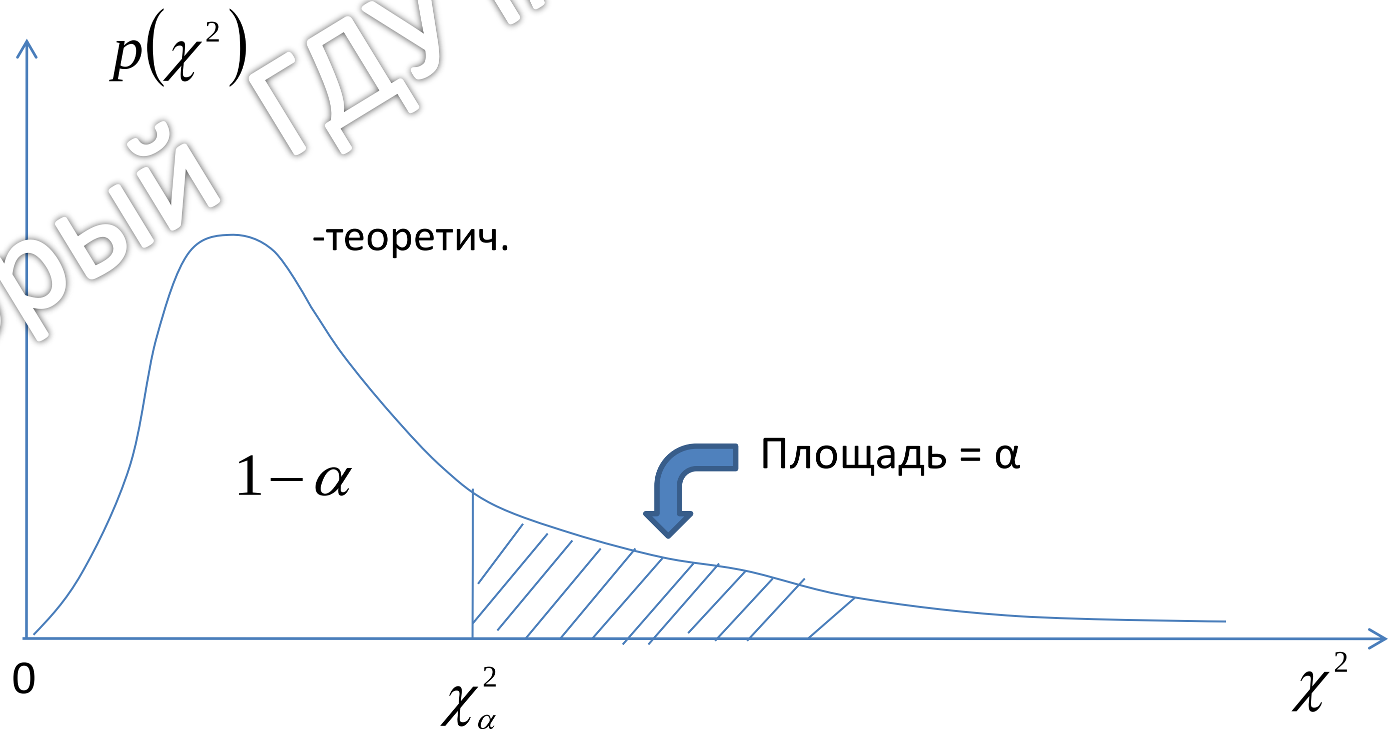
$$\chi^2 > \chi^2_{\alpha,k}$$

α - уровень значимости,
или вероятность <u>отклонения</u> модели

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Как правило, критерий хи-квадрат дает лишь отрицательный ответ однако на практике используется и для принятия положительного решения.

Распределение χ2 :



8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

В этом критерии используется максимальное значение модуля разности между экспериментальным и теоретическим распределением интегральных функций распределений.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Математически это означает, что находят:

$$D = \max |F_{экспер}(x) - F_{теор}(x)| / n$$

где F(x) – интегральная функция распределения.

и сравнивается с граничным значением вероятности

$$P \geq 2\exp\left(-2nD^2\right) = 2\exp(-2\Delta^2 / n)$$

$$\Delta = | F_{экспер} - F_{теор} |$$

Здесь P – искомая вероятность того, что искомая выборка может соответствовать выбранной модели.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Можно построить и доверительный интервал следующего типа:

$$P_{дов}[F_{экспер} - d_{\alpha} \le F_{теор} \le F_{экспер} + d_{\alpha}] = 1 - \alpha$$

где d – ширина полосы в которой лежит ***F экспериментальная***.

где критические значения

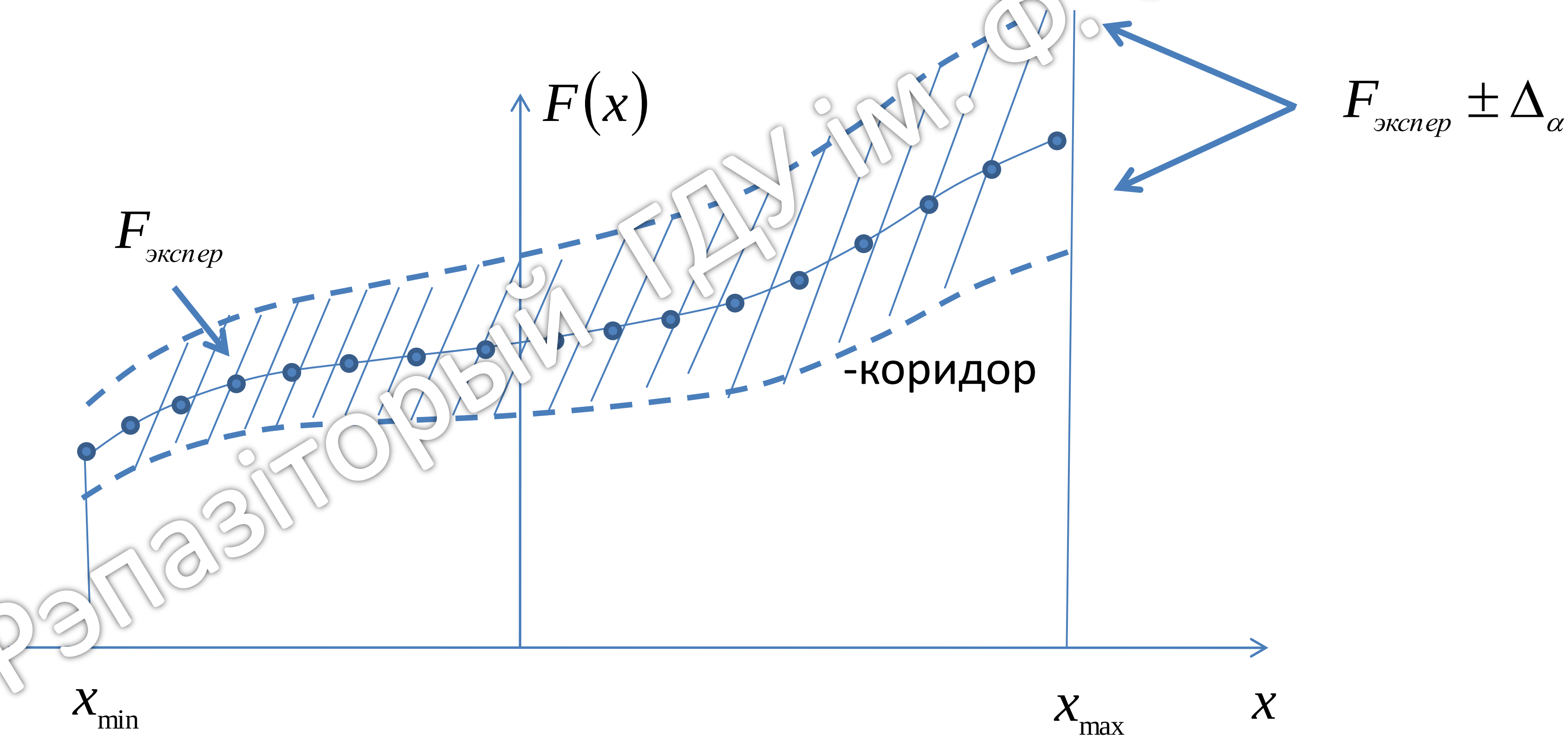
$$d_{\alpha} \approx \sqrt{-0.5\ln\left(\frac{\alpha}{2}\right)/n} \quad (n > 35)$$

Здесь P – искомая вероятность того, что искомая выборка может соответствовать полученной модели, α – уровень значимости.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Доверительный интервал представляет собой полосу с шириной $\pm d_{\alpha}$ около выбранного нами (измеренного) $F_{экспер}$ и с вероятностью 1-α истинная функция $F_{теор}(x)$ лежит внутри полосы.



8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Преимуществом перед хи-квадрат состоит в том, что не требуется группирование данных в интервалы.

Если же есть гистограммы, то они д.б. построены в одинаковых границах числе интервалов группирования.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.3 Критерий Мизеса (ω^2 – критерий)

Этот критерий также использует не сгруппированные данные.

1) Рассчитывается величина:

$$\omega^2 = \int_{-\infty}^{\infty} \left[F_{экспер} - F_{теор}\right]^2 dF_{теор}$$

$$dF_{теор} = f_{теор}(x)dx$$

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.3 Критерий Мизеса (ω^2 – критерий)

1) Рассчитывается величина:

$$\omega^2 = \int_{-\infty}^{\infty} \left[F_{экспер} - F_{теор}\right]^2 dF_{теор}$$

$$dF_{теор} = f_{теор}(x)dx$$

2) Для гистограмм $F_{экспер}$, м.б. записана в виде

$$F_{экспер}(x) = \begin{cases} 0, & x < x_1 \\ \dfrac{k}{n}, & x_k \le x \le x_{k-1} \quad (k = 1, \ldots, n-1) \\ 1, & x > x_n \end{cases}$$

k – число попаданий в k-й интервал

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.3 Критерий Мизеса (ω^2 – критерий)

3) Получаем:

$$\omega^2 = \frac{1}{n}\left\{\frac{1}{12n} + \sum_{k=1}^{n}\left[F_{теор}(x_k) - \frac{2k-1}{2n}\right]^2\right.$$

4) Затем сравнивают $n\omega^2$ с $(n\omega^2)_{крит}$ т.е. с заданным уровнем значимости α .

Таблица для n>40.

α	**$(n\omega^2)_{крит}$**	**α**	**$(n\omega^2)_{крит}$**
0,5	0,1184	0,05	0,4614
0,4	0,1467	0,03	0,5389
0,3	0,1842	0,02	0,6198
0,2	0,2412	0,01	0,7435
0,1	0,3473	0,001	1,1679

Примечание: В этом критерии более полно используется информация о выборке.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Замечания для критериев согласия

При использовании критериев согласия обычно оговаривается, что их применение корректно лишь при достаточно «больших» выборках и, как правило, указывается n>200.

Располагая соотношением для выбора m, увидим это более наглядно

$$m = \frac{\varepsilon + 1.5}{6} n^{0.4} \Rightarrow \quad n = ?$$

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Замечания для критериев согласия

При использовании критериев согласия обычно оговаривается, что их применение корректно лишь при достаточно «больших» выборках и, как правило, указывается n>200.

Располагая соотношением для выбора m, увидим это более наглядно

$$m = \frac{\varepsilon + 1.5}{6} n^{0.4} \quad \Rightarrow \quad n = \left[\frac{6m}{\varepsilon + 1.5} \right]^{2.5}$$

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Замечания для критериев согласия

Располагая соотношением для выбора m, увидим это более наглядно

$$m = \frac{\varepsilon + 1.5}{6} n^{0.4} \quad \Rightarrow \quad n = \left[\frac{6m}{\varepsilon + 1.5} \right]^{2.5}$$

Для χ2 желательно, чтобы $m = 7,\ 9,\ 11$

Тогда для нормального распределения (ε=3) имеем

$$m = 7 \div 11 \qquad \Rightarrow n = 170 \div 800 \quad \text{отсчетов}$$

Для равномерного (ε=1.8) имеем

$$m = 7 \div 11 \qquad \Rightarrow n = 600 \div 2700 \quad \text{отсчетов}$$

Поэтому выбор n>200, которое часто используется это слишком оптимистично.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Замечания для критериев согласия

Существует статистика малых отсчетов:

$$m=9 \quad (\varepsilon=3) \quad \Rightarrow \quad n=500$$

$$m=9 \quad (\varepsilon=1{,}8) \quad \Rightarrow \quad n=1080$$

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Замечания для критериев согласия

Аналогично оценим объем выборки с использованием критерия Колмогорова:

$$d_\alpha \approx \sqrt{-0.5\ln\left(\frac{\alpha}{2}\right)/n}$$

при $\alpha = 0,05$ $\quad d_\alpha = \dfrac{0,61}{\sqrt{n}}$

Если $d_{0,05} = 0,061 \Rightarrow n \sim 100$

$$d_{0,05} = 0,05 \Rightarrow n \sim 124$$

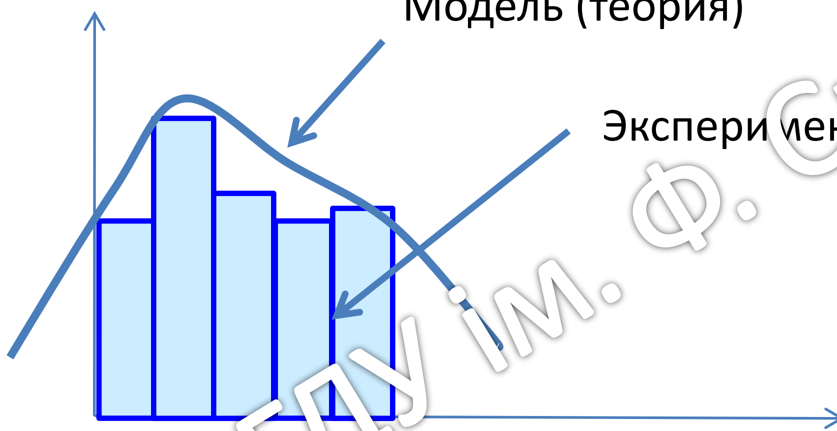
$$d_{0,05} = 0,03 \Rightarrow n \sim 400$$

$$d_{0,05} = 0,01 \Rightarrow n \sim 3600$$

Но объем 500-2500 для экспериментаторов практически редко достижим.

9. Сравнение гистограмм. Операции с гистограммами.

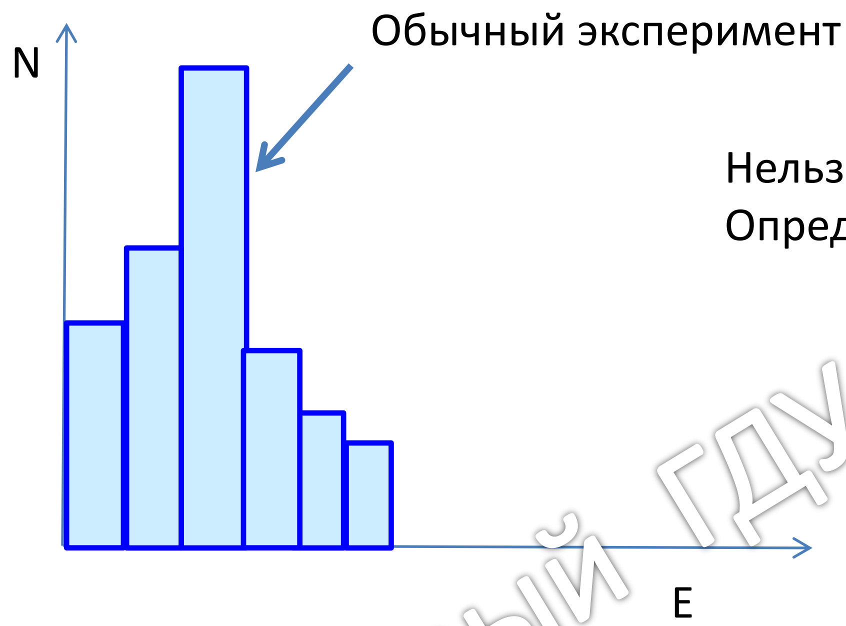
Имея выборку $x1, x2, \dots, xN$ мы стоим гистограмму - $S(i)$, (где $i=1..N$)



Но объем 500-2500 для экспериментаторов практически редко достижим.

9. Сравнение гистограмм. Операции с гистограммами.

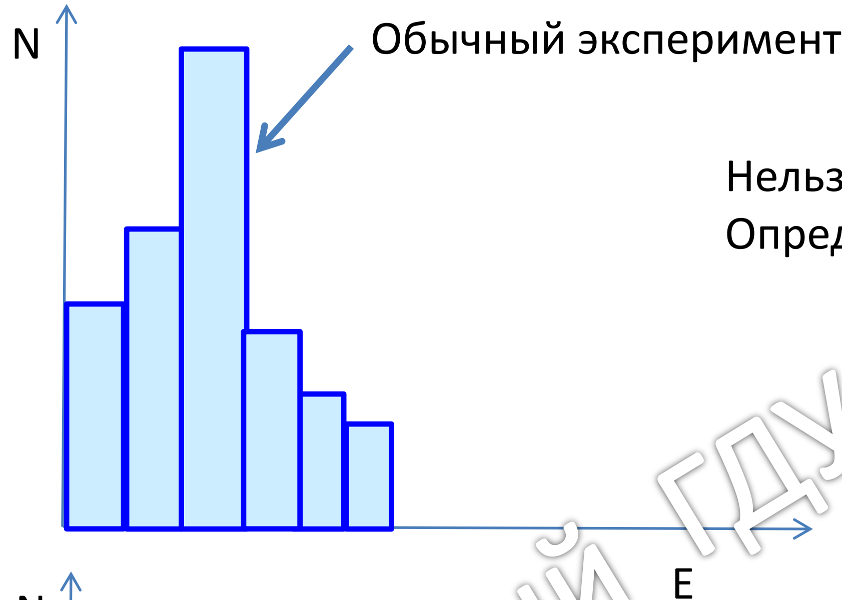
Но в физике кроме обычного эксперимента S(i)   проводят также Калибровочный эксперимент   Sк(i) .



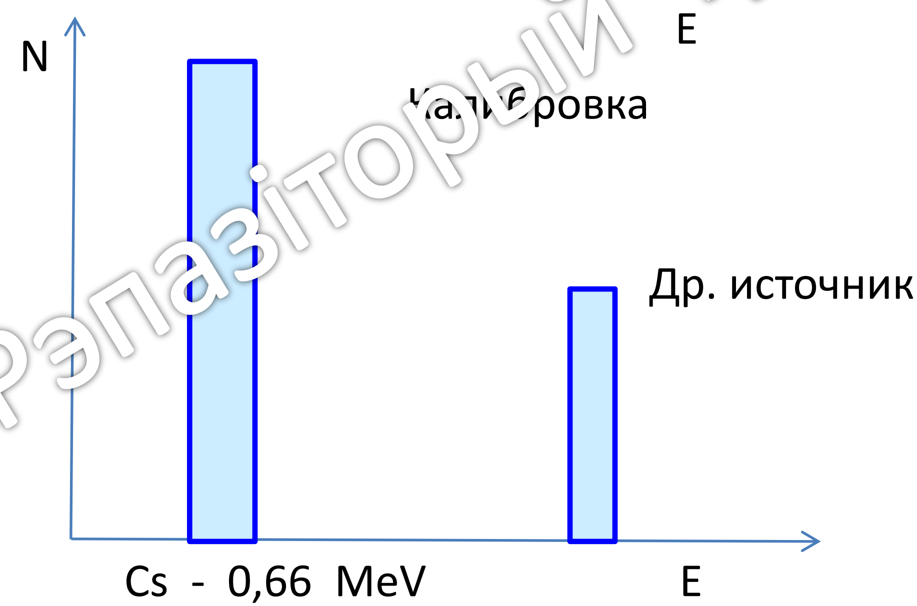
Нельзя точно сказать сколько частиц имеют Определенную энергию.

9. Сравнение гистограмм. Операции с гистограммами.

Но в физике кроме обычного эксперимента S(i) проводят также Калибровочный эксперимент Sк(i) .

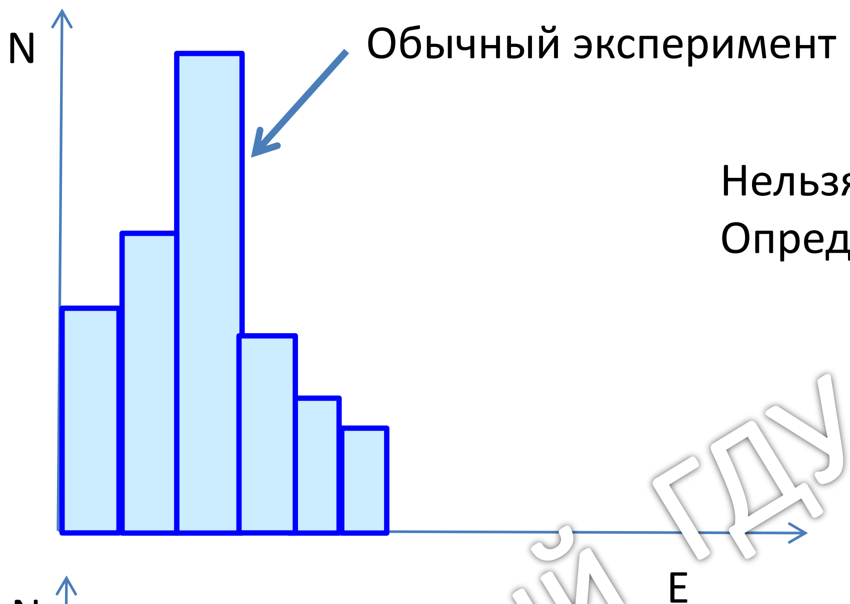


Нельзя точно сказать сколько частиц имеют Определенную энергию.

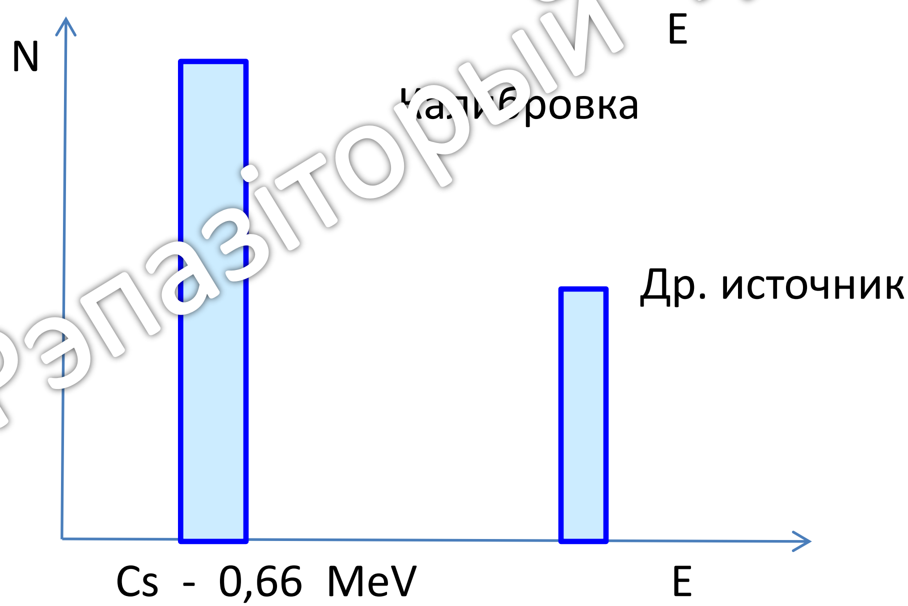


9. Сравнение гистограмм. Операции с гистограммами.

Но в физике кроме обычного эксперимента S(i) проводят также Калибровочный эксперимент Sк(i) .



Нельзя точно сказать сколько частиц имеют Определенную энергию.



1) Калибровка
2) Точки сравнения
3) Сравнение точек с гистограммами с неизвестной энергией

9. Сравнение гистограмм. Операции с гистограммами.

Калибровочный эксперимент - тоже гистограмма, которую также следует обрабатывать.

В реальности имеется:
1) Эксперимент S (i)
2) Калибровка Sк (i)

1? Зачем делать шаги 1-2-3? Как делать на прямую сравнение гистограмм?

Итак имеем гистограммы:

S1(i) - калибровочная гистограмма;
S2(i) - экспериментальная гистограмма;
i = 1 … n.

Имеются также оценки параметров гистограмм : $p_{1j} \quad p_{2j} \quad j = 1 \ldots k$

Сравнение м.б. Проведено 2-мя способами:

9. Сравнение гистограмм. Операции с гистограммами.

Сравнение м.б. Проведено 2-мя способами:

$$\chi_1^2 = \sum_{i=1}^{n} \frac{\left(S_1(i) - S_2(i)\right)^2}{D_1(i)} \qquad (9.1)$$

$D_1(i)$ - дисперсия $S_1(i) - S_2(i)$

$$\chi_2^2 = \sum_{j=1}^{k} \frac{\left(p_{1j} - p_{2j}\right)^2}{D_2(j)} \qquad (9.2)$$

$D_2(j)$ - дисперсия $p_{1j} - p_{2j}$

9. Сравнение гистограмм. Операции с гистограммами.

$$\chi_1^2 = \sum_{i=1}^{n} \frac{\left(S_1(i) - S_2(i)\right)^2}{D_1(i)} \qquad (9.1)$$

$$\chi_1^2 \Rightarrow \chi^2 \quad c \;\; n \;\; ст.свободы$$

$$\left(\chi_1^2 \leq \chi_{n,p}^2\right)$$

p – доверительная вероятность совпадения гистограмм

$$\chi_2^2 = \sum_{j=1}^{k} \frac{\left(p_{1j} - p_{2j}\right)^2}{D_2(j)} \qquad (9.2)$$

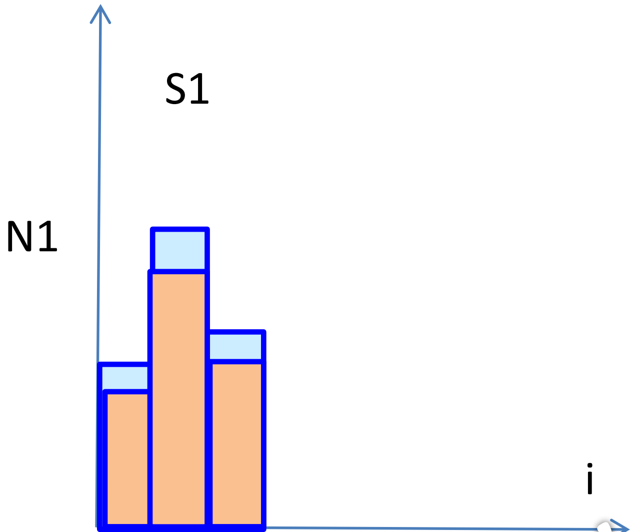
$$\chi_2^2 \Rightarrow \chi^2 \quad c \;\; k \;\; ст.свободы$$

$$\left(\chi_2^2 \leq \chi_{k,p}^2\right)$$

Этот критерий можно использовать для участков спектров.

9. Сравнение гистограмм. Операции с гистограммами.

Рассмотрим, что получается в реальности?



Пример: радиоактивный источник.
Есть активность, т.е. Испускаются частицы в ед. времени

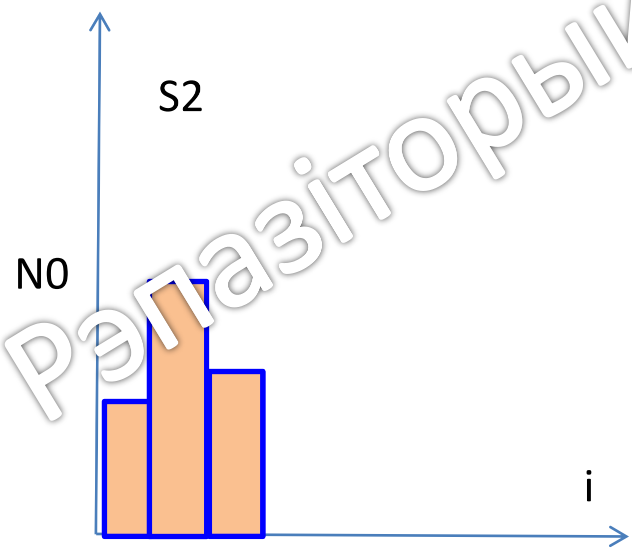
Имеем N0 – фиксированное число отсчетов.

Беря другой источник мы найдем N1

N1 и N0 совпадут!

Но если мы, к примеру, знаем, что это Cs,
а не совпадают из-за того, что кол-во Cs
Различно, то можем ввести множитель.

$$k S_1 = S_2$$



9. Сравнение гистограмм. Операции с гистограммами.

В реальности не всегда нужно «в лоб» сравнивать гистограммы, даже если они для одной и той же физической величины.

1) Одна гистограмма м.б. сдвинута относительно другой (появление фона);
2) М.б. пропорциональность (различная интенсивность одного и того же излучения;
3) Уширение сигнала (по той же причине, что и п.2

Что делать в этой ситуации?

9. Сравнение гистограмм. Операции с гистограммами.

Что делать в этой ситуации?

Запишем $S_2(i)$ в виде

$$S_2(i) = AS_1\left(\frac{i-p}{w}\right) \qquad (9.3)$$

а если уширение в каждом столбце различно, то в виде

$$S_2(i) = AS_1\left(\frac{i-p}{c\,i+w}\right) \qquad (9.4)$$

где p – сдвиг (S_2 относительно S_1)
A – пропорциональность,
w – уширение,
c i + w - коэф. уширения в каждом столбике

9. Сравнение гистограмм. Операции с гистограммами.

Тогда составим

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(S_2(i) - AS_1\left(\frac{i-p}{w}\right) \right)^2}{D_1(i)} \qquad (9.5)$$

χ^2 - с (n-k) ст. свободы, k – число параметров.

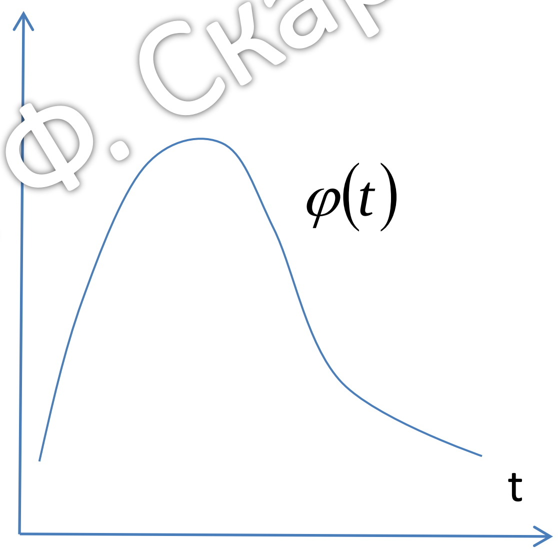
Коэффициенты A, p, w, c находятся методом наименьших квадратов, который (рассмотрим позже).

10. Преобразование Фурье.

Рассмотрим дискретное преобразование Фурье.

t – «временная» компонента
ω – «частотная» компонента

Измерения дают нам $\varphi(t)$



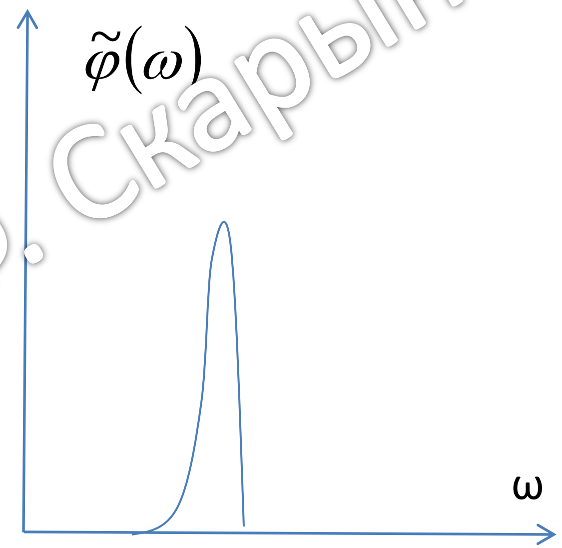
$$\varphi(t) \to N_1 \cdot \int_{-\infty}^{\infty} e^{-i\omega t} \varphi(t) dt \to \tilde{\varphi}(\omega)$$

Прямое преобразование Фурье.

10. Преобразование Фурье.

Обратное преобразование Фурье.

$$N_2 \cdot \int_{-\infty}^{\infty} e^{+i\omega t}\tilde{\varphi}(\omega)d\omega \to \varphi(t)$$



10. Преобразование Фурье.

Дискретное преобразование Фурье

Пусть M – число точек гистограммы h(i), тогда дискретное преобразование Фурье для нее равно:

$$g(k) = \frac{1}{M}\sum_{i=1}^{M} h(i)\left[\exp\left(-2\pi(i-1)j\frac{k}{M}\right)\right]$$

k=0,..M-1 *j^2=-1*

Обратное преобразование Фурье выглядит следующим образом:

$$h(i) = \sum_{k=0}^{M-1} g(k)\exp\left(j\,2\pi\frac{k}{M}i\right)$$

i=1..M

Как и в непрерывном случае преобразование Фурье показывает из каких частот «построена» гистограмма h(i).

10. Преобразование Фурье.

Как и в непрерывном случае преобразование Фурье показывает из таких частот «построена» гистограмма h(i).

Число частот для гистограмм ограничено (в силу периодичности *exp (jn)* и оно тем меньше, чем меньше число каналов (интервалов), представляющих спектр, т.е. чем шире ячейка гистограммы.

10. Преобразование Фурье.

Другой аспект использования преобразования Фурье – поиск периодичностей в гистограмме *h(i)*.

Если *h(i)* есть положение периодических функций с периодами T_l, $l=1..N$, то $g(k)$ будет иметь L пиков в точках, соответствующих частотам периодических функций.

Численное осуществление прямого преобразования Фурье и обратного преобразования Фурье проводится с помощью алгоритмов быстрого преобразования Фурье.

10. Преобразование Фурье.

Рассмотрим идею быстрого преобразования Фурье.

Стандартное преобразование Фурье можно записать в виде

$$g(k) = \sum_{i=0}^{M-1} \frac{h(i)}{M} W(k_i) \qquad k = 0..M-1 \qquad i = 1..M$$

$$W(k_i) = \exp\left[-j2\pi \frac{k}{M} i\right]$$

Для расчета этого ряда необходимо M^2 операций умножения и сложения комплексных чисел (одна операция (комплексная) эквивалентна четырем операциям сложения и умножения действительных чисел).

M^2 - долго и дорого. Что делать?

Решение: быстрое преобразование Фурье *(СУРС)*