Введение в обработку экспериментальных ускорительных данных (практический курс)

Л.В.Кардапольцев
l.kardapoltsev@gmail.com

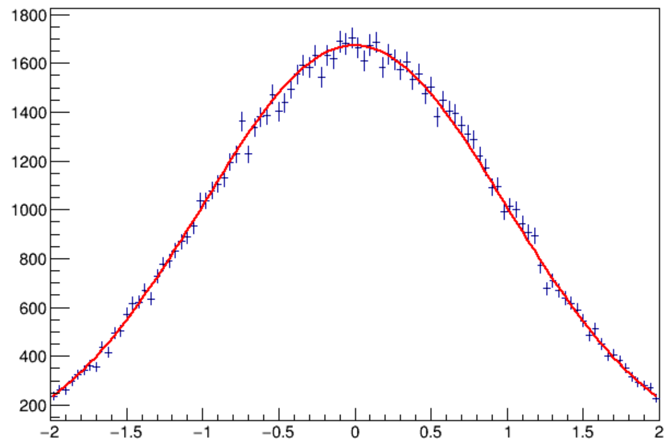
Четвертое занятие

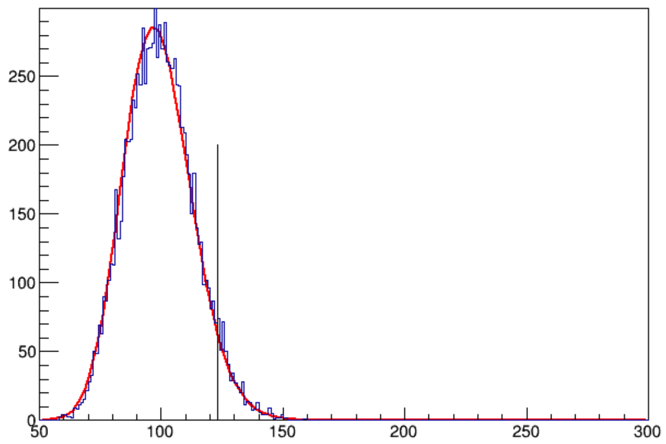
Проверка качества подгонки

Критерий χ²

- Критерий χ² является наиболее распространенным для проверки качества подгонки
- Метод заключается в том что по результатам подгонки вы вычисляете тестовую статистику

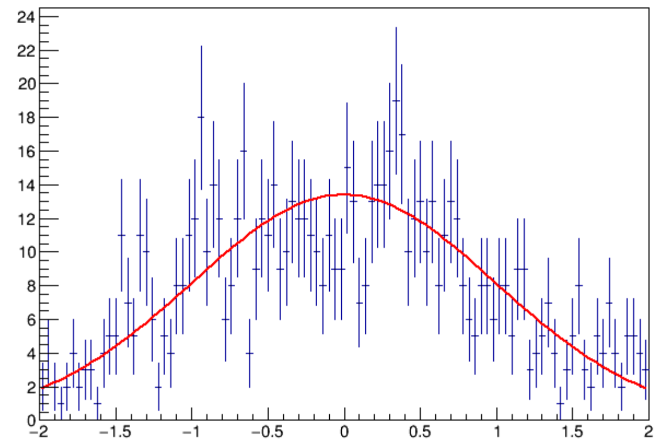
$$\chi^2 = \frac{\sum (N_{data} - N_{fit})^2}{Error^2}$$

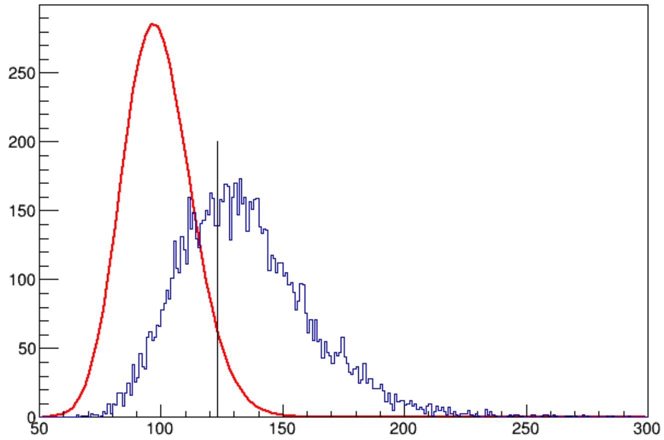




Критерий χ^2

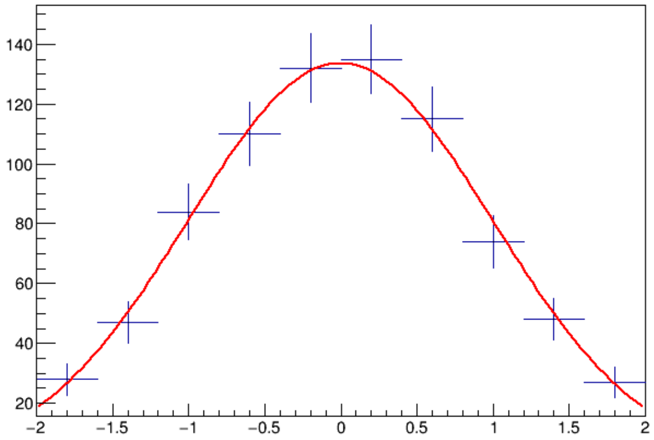
- **Если подгоночная функция действительно описывает данные и ошибки в каждом бине гауссовы, то χ^2 распределено согласно распределению χ^2 с числом степеней свободы равным числу бинов минус число параметров подгонки**
- **Если согласно предполагаемому распределению вероятность получить χ^2 больше 5%, то подгонка считается удачной**

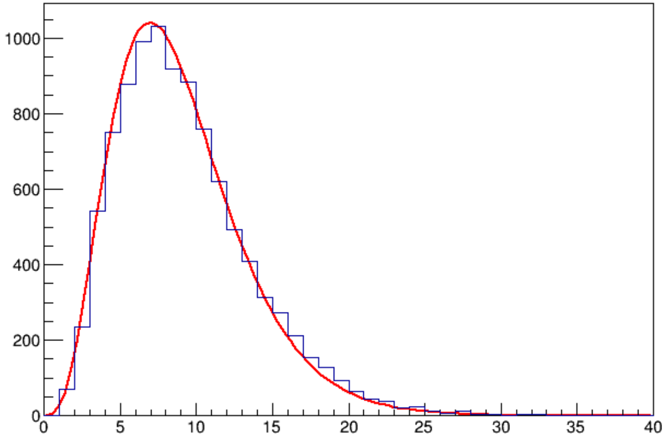




Критерий χ^2

- Этот метод подходит только для гистограмм где **в каждом бине как минимум 5 событий**
- Если событий меньше, то **может помочь укрупнение бинов**





Метод насышенной модели

- В случае если в гистограмме много пустых бинов можно использовать значение функции правдоподобия, сравненивая его со значением для так называемой «насыщенной» модели
- Для этой модели $N_{fit} == N_{data}$ для всех бинов
- Из логарифма функция правдоподобия в ROOT уже вычтено значение для насыщенной модели
- Распределение для такой функции правдоподобия нам не известно, его нужно искать используя псевдоэксперименты (ToyMC)
- Для этого нужно в цикле генерировать гистограмму с распределением согласно функции, полученной вами из подгонки
- Полученное вами значение -2 lnL нужно сравнить с распределением по -2 lnL для псевдоэкспериментов

Критерий Колмагорова-Смирнова

- **TH1::KolmogorovTest использует критерий Колмогорова-Смирнова чтобы сравнить распределения в двух гистограммах**
- **Выдаваемая вероятность должна быть распределена равномерно от 0 до 1 для совпадающих распределений**
- **Бинирование может приводить приводить к сильному искажению этого распределения**

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leqslant x},$$

где $I_{X_i \leqslant x}$ указывает, попало ли наблюдение X_i в область $(-\infty,\ x]$:

$$I_{X_i \leqslant x} = \begin{cases} 1, & X_i \leqslant x; \\ 0, & X_i > x. \end{cases}$$

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

Теорема Смирнова.

Пусть $F_{1,\ n}(x),\ F_{2,\ m}(x)$ — эмпирические функции распределения, построенные по независимым выборкам объёмом n и m случайной величины ξ.

Тогда, если $F(x) \in C^1(\mathbb{X})$, то $\forall t > 0$: $\lim\limits_{n,\ m \to \infty} P\left(\sqrt{\frac{nm}{n+m}} D_{n,\ m} \leqslant t\right) = K(t) = \sum\limits_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$, где $D_{n,\ m} = \sup\limits_x |F_{1,\ n} - F_{2,\ m}|$.

Принятие решения по критерию Смирнова.

Если статистика $\sqrt{\frac{nm}{n+m}} D_{n,\ m}$ превышает квантиль распределения Колмогорова K_α для заданного уровня значимости α, то нулевая гипотеза H_0 (об однородности выборок) отвергается. Иначе гипотеза принимается на уровне α.

Критерий Андерсона-Дарлинга

- **TH1::AndersonDarlingTest** **использует критерий** **Андерсона-Дарлигна** **чтобы сравнить распределения** **в двух гистограммах**
- **Выдаваемая вероятность должна быть распределена** **равномерно от 0 до 1 для совпадающих распределений**

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)\,(1 - F(x))}\, dF(x),$$

- **Бинирование может приводить** **приводить к искажению** **этого распределения**

Подгонка многомерных распределений

- С ростом размерности число бинов очень быстро растет, так что очень редко удается избежать бинов с малым числом событий
- Можно использовать метод псевдоэкспериментов, но у него как правило низкая чувствительность
- В ROOT для двуменрных распределений есть метод TH2::KolmogorovTest
- Строго говоря теста Колмогорова в многомерном случае не существует, так как невозможно строго определеить эмпирическую функцию распределения
- Функция распределения определена для одномерного распределения, поэтому 2D гистограмму нужно переобразовать в 1D
- Результат будет сильно зависеть от этого преобразования

Критерий смешанной выборки

$T = \frac{1}{n_k(n_a+n_b)} \sum_{i=1}^{n_a+n_b} \sum_{j=1}^{n_k} I(i,k)$ - тестовая статистика

$I(i,k) = 1$ если $i-$ое событие и $k-$е ближайшее событие принадлежат одному семплу и $I(i,k) = 0$ в противном случае.

В качестве дистанции используется

$$\sum_{\nu=1}^{D} \left(\frac{x_i^\nu - x_j^\nu}{w_\nu} \right)^2$$, в нашем случае $w_\nu = 1$.

В случае если обе выборки данных имеют одинаковое рапределение, получившееся распределение для T будет иметь распределение Гаусса со средним

$\mu_T = \frac{n_a(n_a-1)+n_b(n_b-1)}{n(n-1)}$, где $n = n_a + n_b$

и шириной

$$\lim_{n,n_k,D\to\infty} \sigma_T^2 = \frac{1}{nn_k}\left(\frac{n_a n_b}{n^2} + 4\frac{n_a^2 n_b^2}{n^4} \right)$$