Statistical Methods in Particle Physics

Saeid Paktinat

IPM

IPM National School on the Phenomenological And Experimental aspects of the Elementary Particle Physics

16-19 Oct 2007

Outline

- Definitions
- Famous pdf's
- χ^2 and its applications

Definitions(1)

- S is a sample space, A and B are two subset of S, probability is a real value and

  $$P(A) \geq 0$$

  $$A \cap B = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$$

  $$P(S) = 1$$

- Random variable is a numerical characteristic assigned to an element of S.
- e.g, people's height, weight, …

Definitions(2)

- If x is a continuous variable f(x)dx is the probability to have a measurement which lies between x and x+dx
- f(x) is called the *probability density function* (pdf)

$$\langle u(x) \rangle = \int_{-\infty}^{+\infty} u(x)f(x)dx$$

- Different moments are defined as

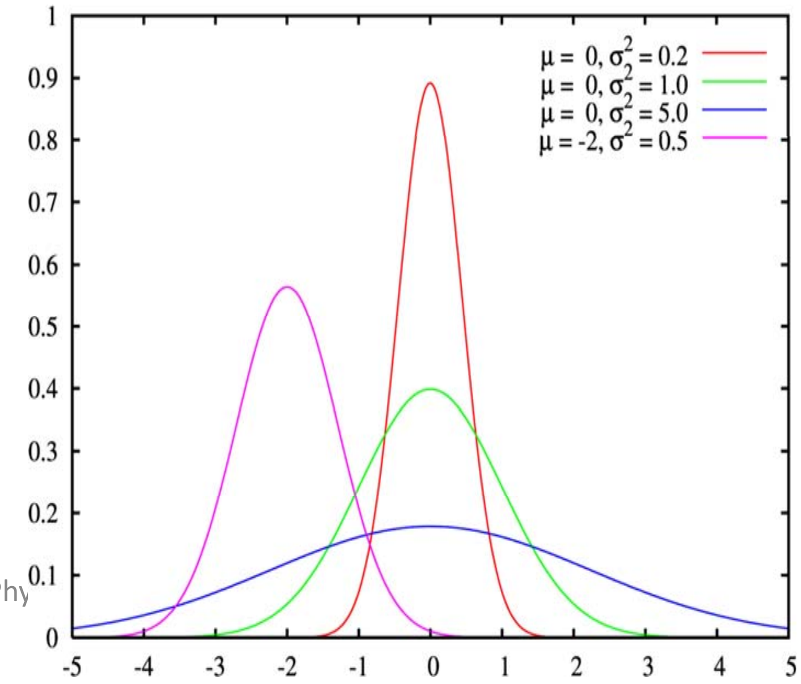
$$\alpha_n = \int_{-\infty}^{+\infty} x^n f(x)dx \quad \langle x \rangle = \alpha_1 \text{ and } \sigma^2 = \alpha_2 - \langle x \rangle^2$$

- Variance is the square of the standard deviation (Root Mean Square)

Famuos pdf's (Gaussian pdf)

When many small, independent effects are additively contributing to each observation the result follows the Gaussian (normal) distribution, e.g, people's height.

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

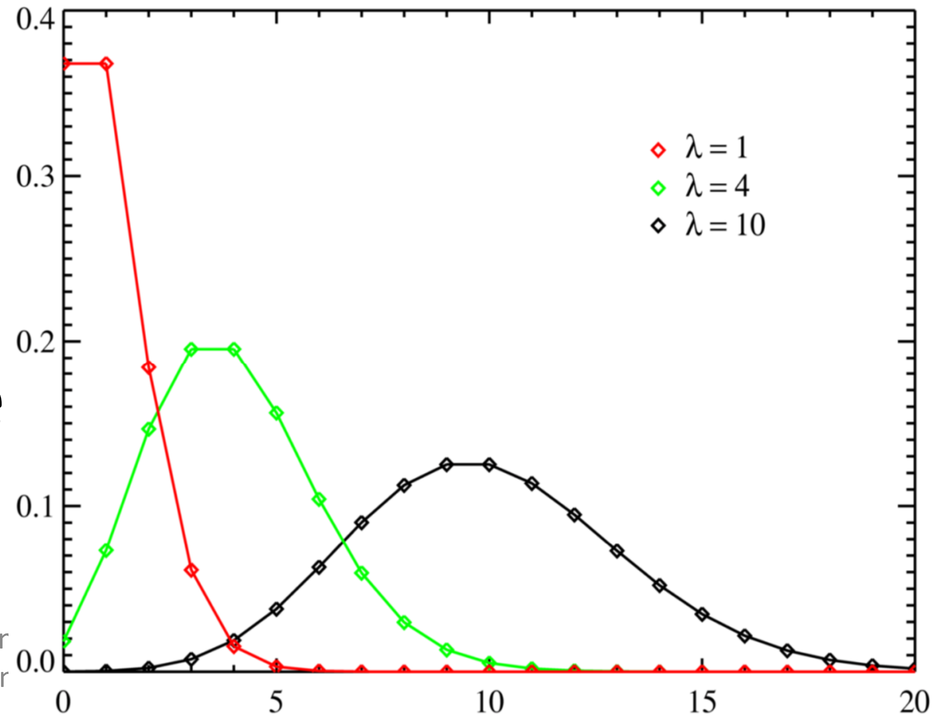


Poisson distribution

- Probability to find n events in a special range when the mean is v.
- variance is equal to v.

$$f(\mathrm{n},\mathrm{v}) = \frac{\mathrm{v}^{\mathrm{n}} e^{-\mathrm{v}}}{\mathrm{n}!}$$

- Large v approaches the Gaussian pdf.



Chi-square (χ^2) pdf

- k independent, normally distributed variables x_i:

$$z = \sum_{i=1}^{k} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

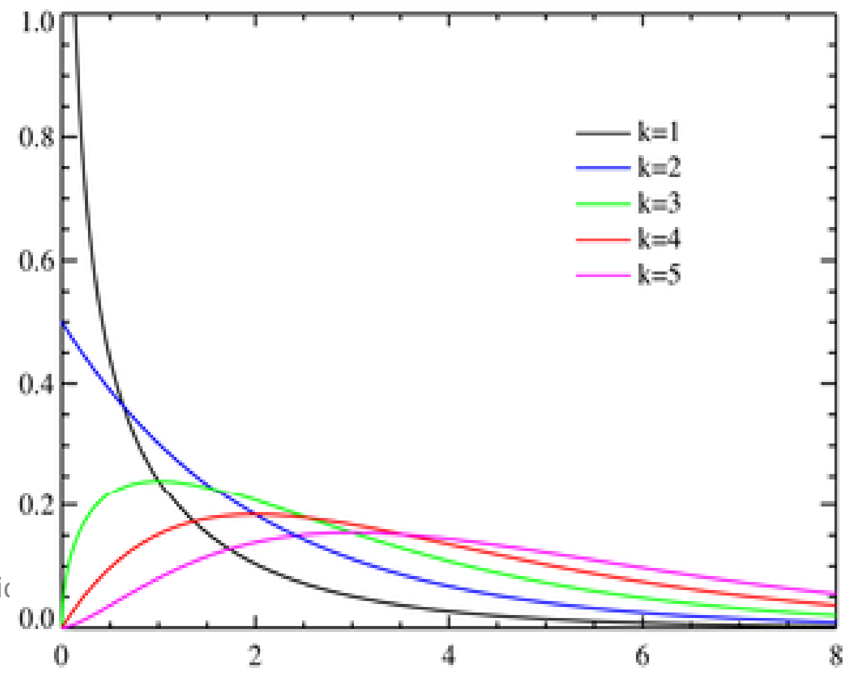
- z follows a chi-square pdf with k degrees of freedom.

$$f(x;k) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

- For the large k, it approaches

Gaussian pdf with mean

k and variance 2k.



χ^2 application

Chi-square distribution is used as a test

- To estimate the unknown parameters of a model
- To evaluate the unknown mean value of a distribution
- To quantify the goodness of fit

Parameter estimation

- Maximum likelihood
- If x_1, x_2, … x_n are the independent measurements which follow pdf $f(x,T)$ with $T(T_1, T_2, \ldots T_m)$ a vector of unknown parameters:
- Likelihood → $L = f(x_1,T) * f(x_2,T) * \ldots f(x_n,T)$
- The correct T will maximize the L.

Least squares

- (x_i, y_i) are the results of an experiment. If y_i has to follow a gaussian pdf with mean $F(x_i,T)$ and a known variance σ_i^2, the correct T will minimize

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - F(x_i, T))^2}{\sigma_i^2}$$

An example

- extract the slope and the offset of a line…

The Origins of the Chi-Square Statistic

- If the deviations from the mean follow Gaussian statistics, the probability of making any one observation is given by:

$$P_G(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2\right] \quad \text{where } x-\mu \;\rightarrow\; y_i - f(x_i)$$

- The total probability of obtaining a set of N measurements, $\{x_i, y_i\}$, is equal to the *product* of the probabilities for each data point:

$$P_{\{x,y\}} = \prod_N P_G = \left\{\prod_N \frac{1}{\sigma\sqrt{2\pi}}\right\} * \left\{\exp\left[-\tfrac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - \mathrm{f}(x_i)}{\sigma_i}\right)^2\right]\right\}$$

- Maximizing the probability is equivalent to minimizing the sum in the exponential term of $P_{\{x,y\}}$, specifically the sum of the deviations, Δy.
- The chi-square statistic is defined by this sum:

$$\chi^2 \equiv \sum_{i=1}^{N}\left(\frac{y_i - \mathrm{f}(x_i)}{\sigma_i}\right)^2$$

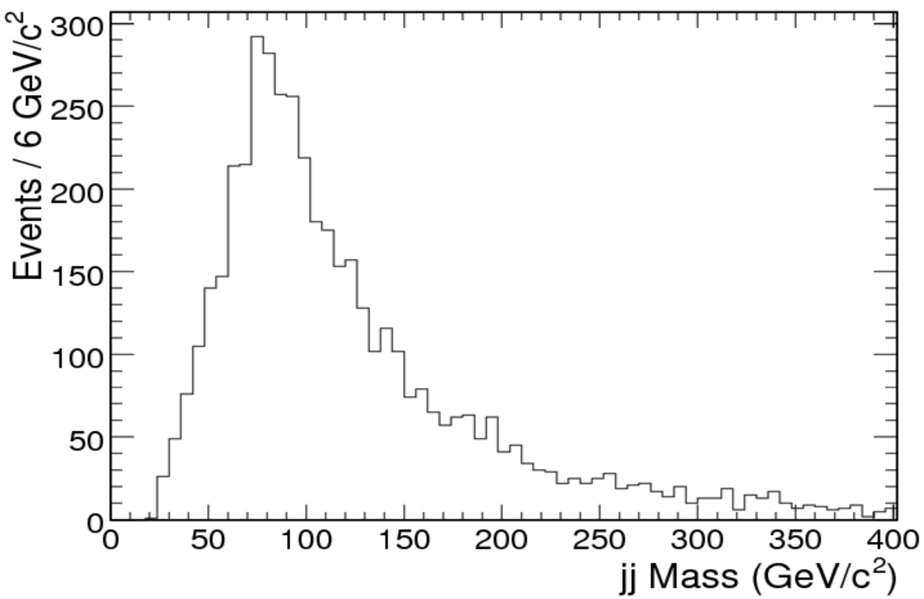
http://www.sns.gov/workshops/sns_hfir_users/posters/Laub_Chi-Square_Data_Fitting.pdf

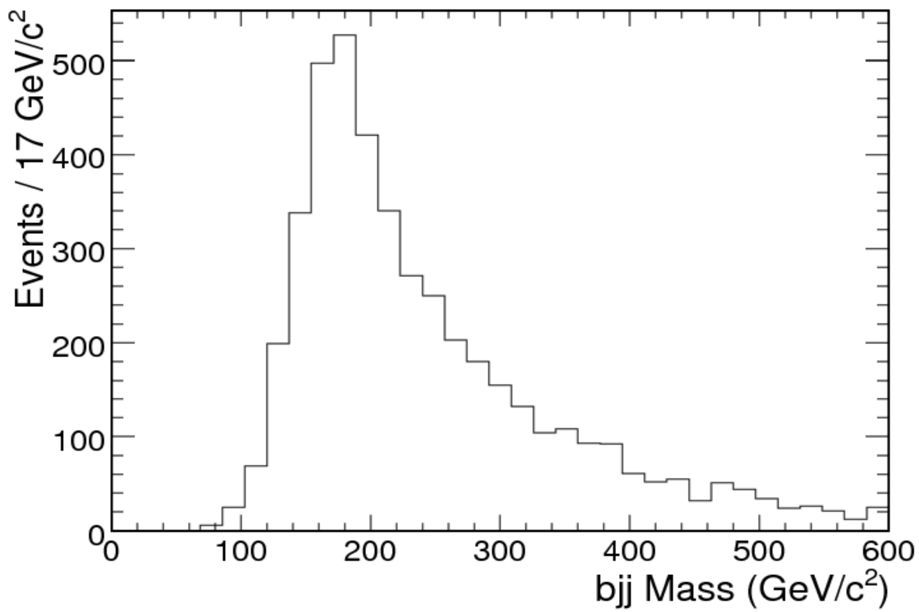
Goodness of fit

- **Goodness of fit** means how well a statistical model fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question.
- In ROOT   TMath::Prob(χ^2,ndf) gives the probability to find χ^2 with ndf.

mean value of a distribution (Kinematic Fit)

- t→ bW → bjj
- W can be made of two non b-jets
- Top can be made of the extracted W and b-jet.





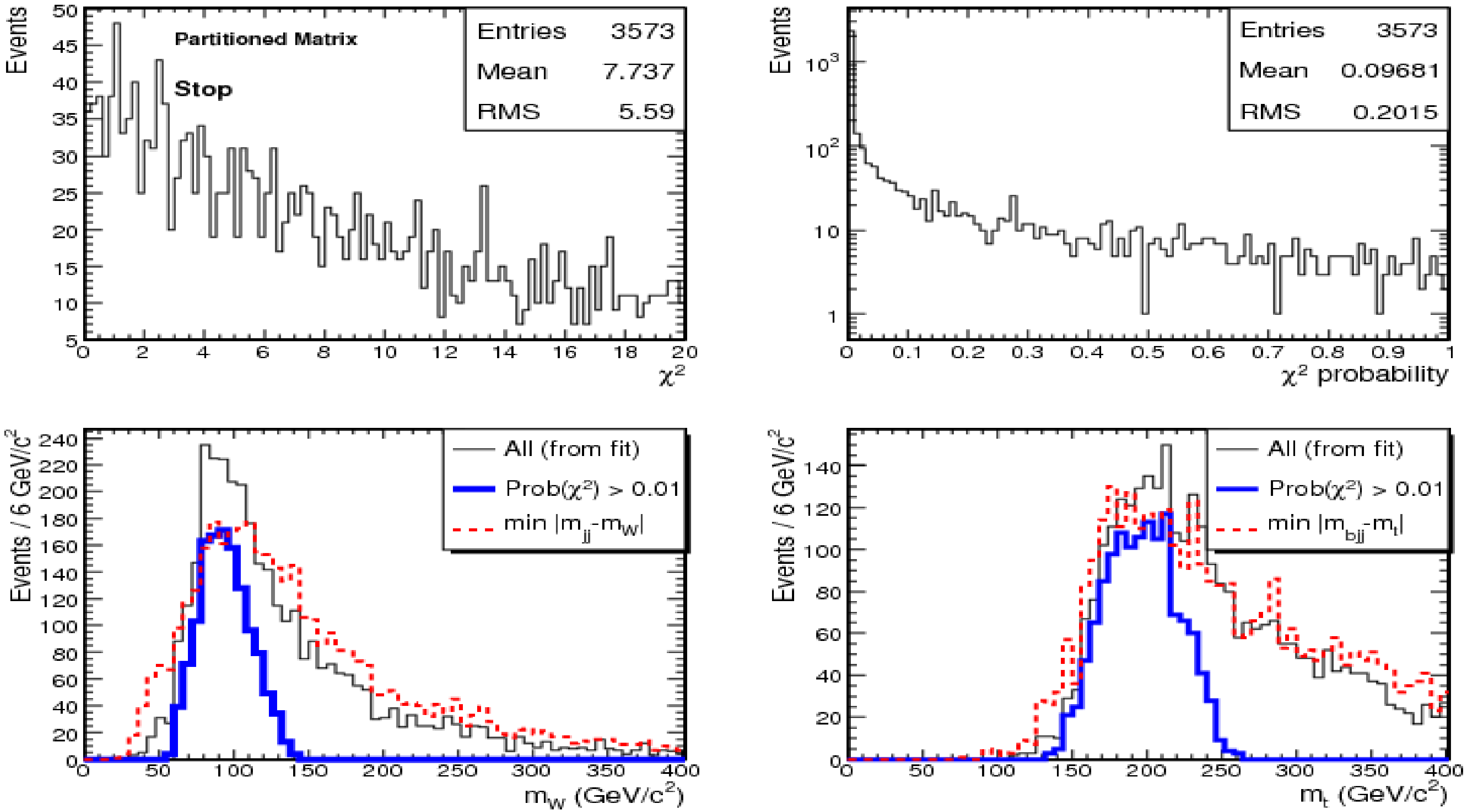
Top quark extraction

- The purpose of the analysis is not to measure the top mass→ top mass can be used with W mass as 2 constraints to find the best jet combination

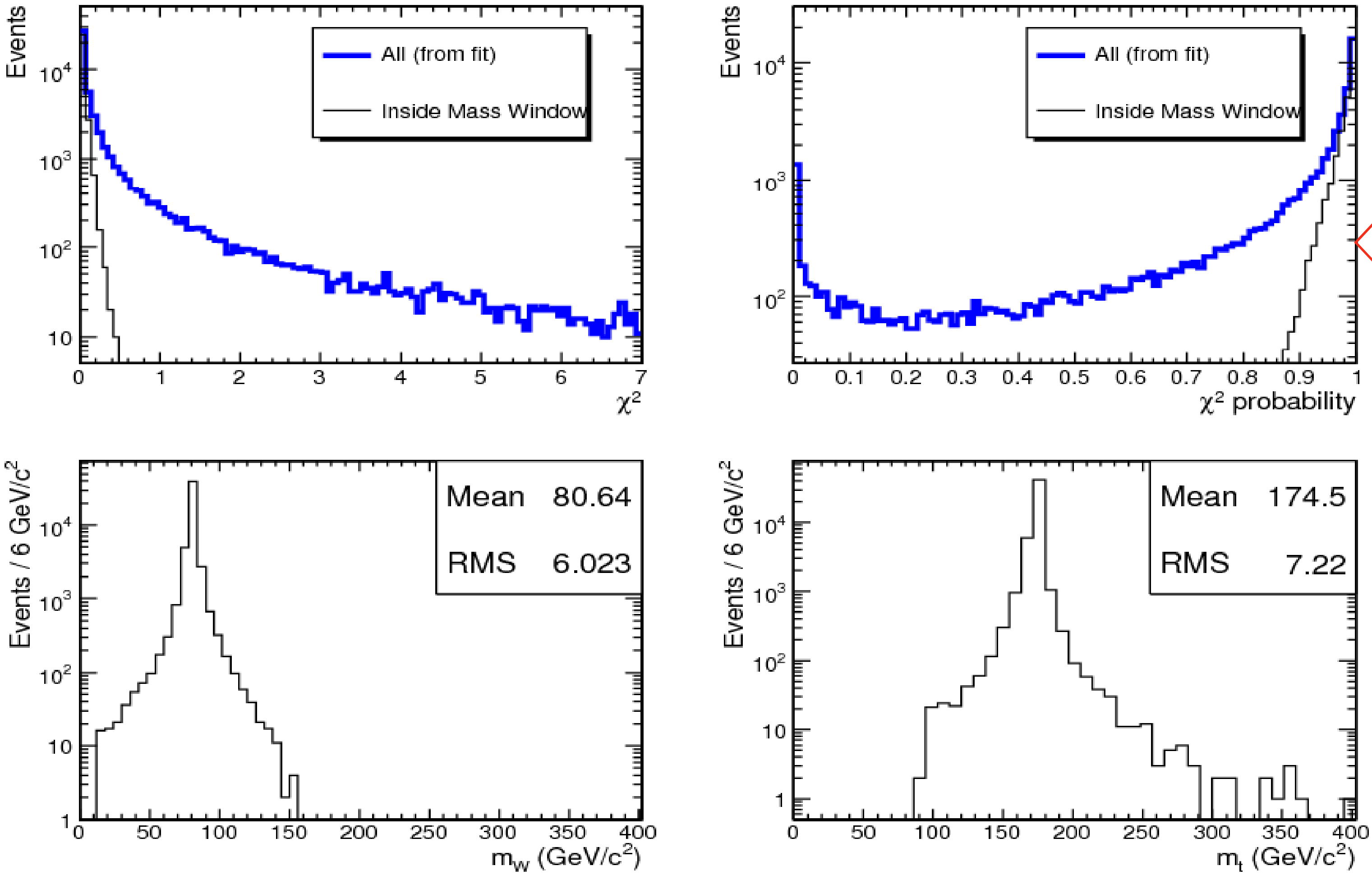
$$\chi^2 = \sum_{i=1}^{3} \frac{(E_i - E_i^m)^2}{\sigma_i^2} + \frac{(m_W - M_W)^2}{(\Gamma_W/2)^2} + \frac{(m_{Top} - M_{Top})^2}{(\Gamma_{Top}/2)^2}$$

- E_i^m is a measured energy with a gaussian distribution (E_i, σ_i).

The least χ^2 in every event



Starting from the right hypthesis Generator level info is used.



Wrongly estimated errors