The method of least squares

Alexander Khanov

PHYS6260: Experimental Methods is HEP
Oklahoma State University

September 22, 2023

Formulation of the problem

- Let variable y be a function of another variable x and parameters $\mathbf{p} = p_1, \ldots, p_n$:
$$y = f(x, \mathbf{p})$$
- Suppose we have a set of N independent measurements of variable y: $\mathbf{y} = y_1, \ldots, y_N$ with known variances $\sigma_1^2, \ldots, \sigma_N^2$ taken at N values of x: $\mathbf{x} = x_1, \ldots, x_N$
- Goal: construct an estimator for $\mathbf{p}$
- Typical applications:
  - Data fitting: have several measurements taken at different times, at different positions etc.
  - Overdetermined systems: problems where the number of unknowns (parameters) is larger than the number of equations (measurements)

χ^2 sum

- Construct a function

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \mathbf{p}))^2}{\sigma_i^2} \tag{1}$$

  (this is not the estimator yet)

- Find the minimum of this function w.r.t. $\mathbf{p}$:

$$\frac{\partial \chi^2}{\partial p_i} = 0 \tag{2}$$

  (a system of n equations with n unknowns $\mathbf{p}$)

- The answer (which is a function of measurements y_i) is an estimator for parameters $\mathbf{p}$
  - the measurements y_i do not have to be Gaussian distributed, but they should be unbiased:

$$\langle y_i \rangle = f(x_i, \mathbf{p}_{\text{true}})$$

Special case: linear dependence on parameters

- In general the system of equations (2) is not easy to solve
- Special case: $f(x, \mathbf{p})$ is a linear function of parameters $\mathbf{p}$:
  - $f(x_i, \mathbf{p}) = \sum_{j=1}^{n} p_j h_j(x_i)$, or $\mathbf{f} = H\mathbf{p}$, where $H_{ij} = h_j(x_i)$
  - f doesn't have to be a linear function of x!
- In this case (2) becomes a system of linear equations w.r.t. $\mathbf{p}$

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} \frac{\left(y_i - \sum_{j=1}^{n} p_j h_j(x_i)\right)^2}{\sigma_i^2}$$

$$\frac{\partial \chi^2}{\partial p_k} = -2 \sum_{i=1}^{N} \frac{y_i - \sum_{j=1}^{n} p_j h_j(x_i)}{\sigma_i^2} h_k(x_i) = 0$$

$$\sum_{j=1}^{n} p_j \sum_{i=1}^{N} \frac{h_j(x_i) h_k(x_i)}{\sigma_i^2} = \sum_{i=1}^{N} \frac{y_i h_k(x_i)}{\sigma_i^2}$$

Correlated measurements

- If y_i are correlated with the covariance matrix $V_{ij} = \mathrm{cov}(y_i, y_j)$, then

$$\chi^2 = (\mathbf{y} - \mathbf{f})^T R (\mathbf{y} - \mathbf{f}) \tag{3}$$

  where $R = V^{-1}$

- If y_i are uncorrelated, R is diagonal:

$$R = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\sigma_N^2 \end{pmatrix}$$

  and we are back to formula (1)

Correlated measurements (2)

- Some linear algebra: if $\frac{\partial}{\partial \mathbf{x}} = \begin{pmatrix} \partial/\partial x_1 \\ \dots \\ \partial/\partial x_n \end{pmatrix}$ then for any constant vector $\mathbf{v}$ and matrix A:

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{v}^T\mathbf{x}) = \mathbf{v} \quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{v}) = \mathbf{v} \quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = A\mathbf{x} + A^T\mathbf{x}$$

- Let's apply it to (3) where $\mathbf{f} = H\mathbf{p}$:

$$\chi^2 = \mathbf{y}^T R\mathbf{y} - \mathbf{p}^T H^T R\mathbf{y} - \mathbf{y}^T RH\mathbf{p} + \mathbf{p}^T H^T RH\mathbf{p}$$

$$\frac{\partial \chi^2}{\partial \mathbf{p}} = 0 - H^T R\mathbf{y} - (\mathbf{y}^T RH)^T + (H^T RH)\mathbf{p} + (H^T RH)^T\mathbf{p} = 0$$

$R^T = R$, so $(H^T RH)\mathbf{p} = H^T R\mathbf{y}$, $\boxed{\mathbf{p} = (H^T RH)^{-1} H^T R\mathbf{y}}$

- One can show that the covariance matrix for the estimators $U_{ij} = \mathrm{cov}(p_i, p_j)$ is calculated as $U = (H^T RH)^{-1}$

Fit with a constant

- There is only one parameter p and $f(x) = p$, so $H = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$
- Eq. (1) reduces to

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - p)^2}{\sigma_i^2}$$

$$\frac{d\chi^2}{dp} = 0 \Rightarrow p = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

$$\sigma_p^2 = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1}$$

which is exactly what we had for the estimator of the mean

Fit with a straight line

- $f(x) = p_0 + p_1 x$, linear w.r.t. parameters p_0, p_1
- Minimizing χ^2, we get a system of two equations:

$$\begin{cases} p_0 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} + p_1 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \\ p_0 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + p_1 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \end{cases} \tag{4}$$

- Eq. (4) easily generalizes to an arbitrary polynomial fit

$$f(x, \mathbf{p}) = p_0 + p_1 x + \ldots + p_{n-1} x^{n-1}$$

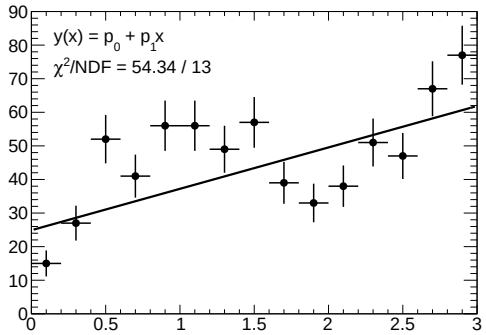
  - note that it's still linear w.r.t. parameters

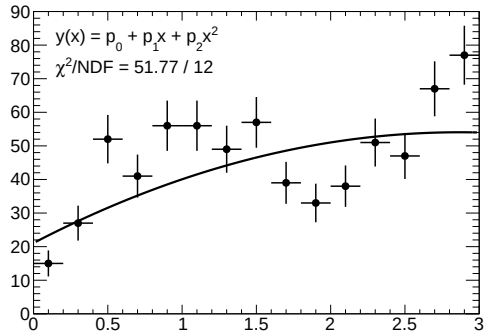
Properties of least squared method

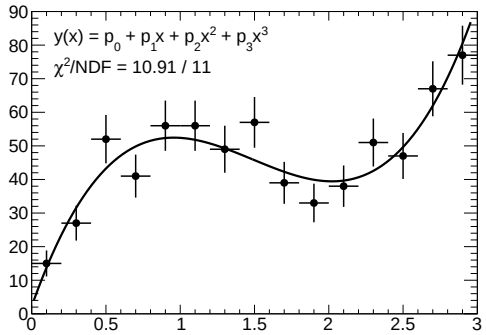
- In general, the L.S. method is neither unbiased nor efficient
- If parameter dependence is linear then estimators produced by the method are unbiased
- If measurements are Gaussian distributed then the method is asymptotically efficient (i.e. it is more and more efficient as the number of measurements increases)
  - in this case χ^2 follows the χ^2 distribution :)

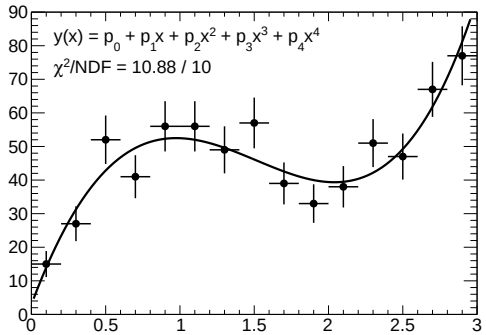
Example of least squares polynomial fit

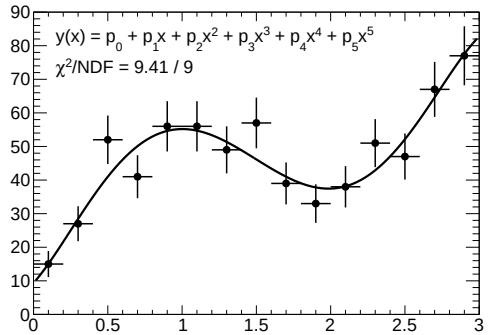
- Which fit should we use? Why?

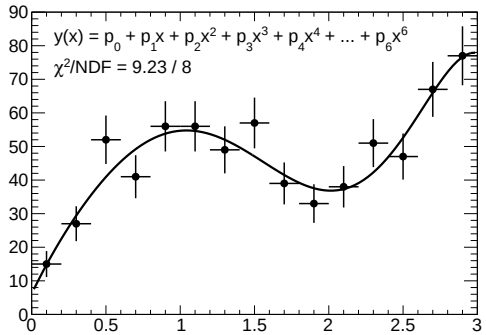












F-test

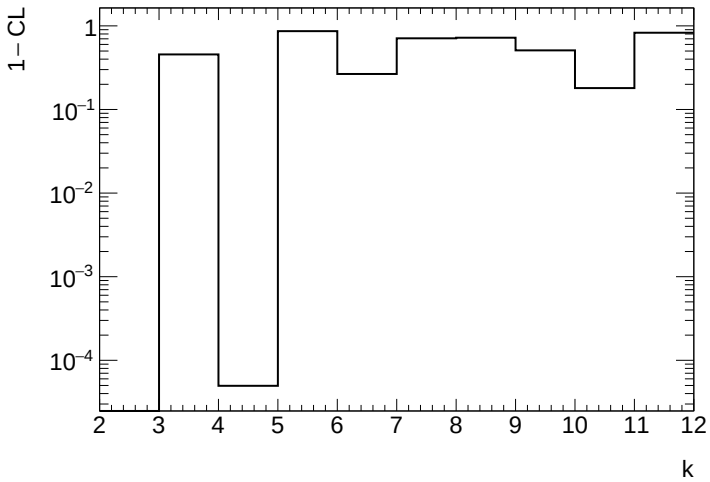
- Let n data points be fitted with two models, 1 and 2, where model 1 is "nested" within model 2
  - model 1 has k_1 parameters, and model 2 has k_2 parameters, $k_1 < k_2$
  - for any choice of parameters in model 1, the same fit can be achieved by some choice of parameters in model 2
- By construction, model 2 gives a better fit than model 1
  - the question is, does model 2 give significantly better fit than model 1
- Calucalate the F statistic:

$$F = \frac{\left(\frac{\chi_1^2 - \chi_2^2}{k_2 - k_1}\right)}{\left(\frac{\chi_2^2}{n - k_2 - 1}\right)}$$

- The null hypothesis (model 2 does not provide a significantly better fit than model 1) is rejected if the F value calculated from the data is greater than the critical value of the F-distribution (e.g. corresponding to CL=95%)

F-test results for our example

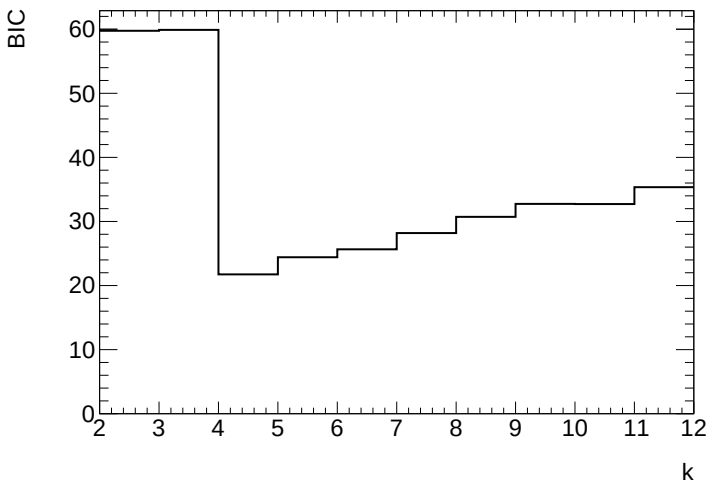
- root will calculate the F probabilities for you
- Transition from 3 to 4 appears to be significant



probability that transition $k-1 \to k$ is not significant

Bayesian Information Criterion

- In general, need to add some term to χ^2 to penalize increasing the number of fit parameters k
- Bayesian information criterion: pick the model with least $\chi^2 + k \ln n$
  - BIC is asymptotically efficient (if one of the models is correct, the probability to pick it approaches 1 as $n \to \infty$)
  - BIS does not requre the models to be nested



BIC has a minimum at $k = 4$

Effective variance

- What to do if both x and y values have errors?
  - we have a set of N independent measurements of variable y: $\mathbf{y} = y_1, \ldots, y_N$ with known variances $\sigma^2_{y1}, \ldots, \sigma^2_{yN}$ taken at N values of x: $\mathbf{x} = x_1, \ldots, x_N$ with known variances $\sigma^2_{x1}, \ldots, \sigma^2_{xN}$
- The usual approach is what is called "effective variance" method: minimize

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \mathbf{p}))^2}{\sigma^2_{yi} + (f'(x_i, \mathbf{p}))^2 \sigma^2_{xi}} \tag{5}$$

  where $f'(x_i, \mathbf{p}) = \left.\frac{\partial f}{\partial x}\right|_{x=x_i}$
- note that this ruins the linearity of minimization equations, so it's usually better to avoid it, or find the approximate minimum without x uncertainties and then improve the result by taking σ^2_{xi} into account

Combining statistical and systematic uncertainties

- Assume we have two measurements x_1 and x_2 of the same quantity x

$$x = x_1 \pm \Delta x_1(stat.) \pm \Delta x_1(syst.)$$
$$x = x_2 \pm \Delta x_2(stat.) \pm \Delta x_2(syst.)$$

Let's assume that systematic uncertainties are 100% correlated between the two measurements

- How to combine them?
  - we can assume that both measurements are constructed out of two variables: $x = r + s$
  - r is randomly distributed with variance $\sigma_r = (\Delta x(stat))^2$
  - s is randomly distributed with variance $\sigma_s = (\Delta x(syst))^2$
  - $\mathrm{cov}(r_1, r_2) = \mathrm{cov}(r_1, s_1) = \mathrm{cov}(r_1, s_2) = \mathrm{cov}(r_2, s_1) = \mathrm{cov}(r_2, s_2) = 0$
  - $\mathrm{cov}(s_1, s_2) = \sigma^2_{s1} = \sigma^2_{s2} = \sigma^2_s$

Combining statistical and systematic uncertainties (2)

- Determine covariance matrix of the measurements:

$$\sigma_{x1}^2 = \left\langle (r_1 + s_1)^2 \right\rangle - \left\langle r_1 + s_1 \right\rangle^2 = \sigma_{r1}^2 + \sigma_s^2$$
$$\sigma_{x2}^2 = \left\langle (r_2 + s_2)^2 \right\rangle - \left\langle r_2 + s_2 \right\rangle^2 = \sigma_{r2}^2 + \sigma_s^2$$
$$\mathrm{cov}(x_1, x_2) = \left\langle (r_1 + s_1)(r_2 + s_2) \right\rangle - \left\langle r_1 + s_1 \right\rangle \left\langle r_2 + s_2 \right\rangle = \sigma_s^2$$

- The covariance matrix looks like follows:

$$V = \begin{pmatrix} \sigma_{r1}^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_{r2}^2 + \sigma_s^2 \end{pmatrix}$$

The rest can be done using the formula for correlated measurements with $H = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

- This approach can be extended to any number of correlated / uncorrelated uncertainties

Binned data

- In many cases we are measuring a random quantity, and we are interested in its p.d.f.
- Suppose we want to determine the mass and the width of the Δ^{++} particle, how do we do it?
  - reminder: Δ^{++} is an unstable baryon with a mass of 1232 MeV decaying into a proton and a π^+
- Let's consider two methods: πp scattering and invariant mass measurement

Measuring the parameters of Δ^{++}: method 1

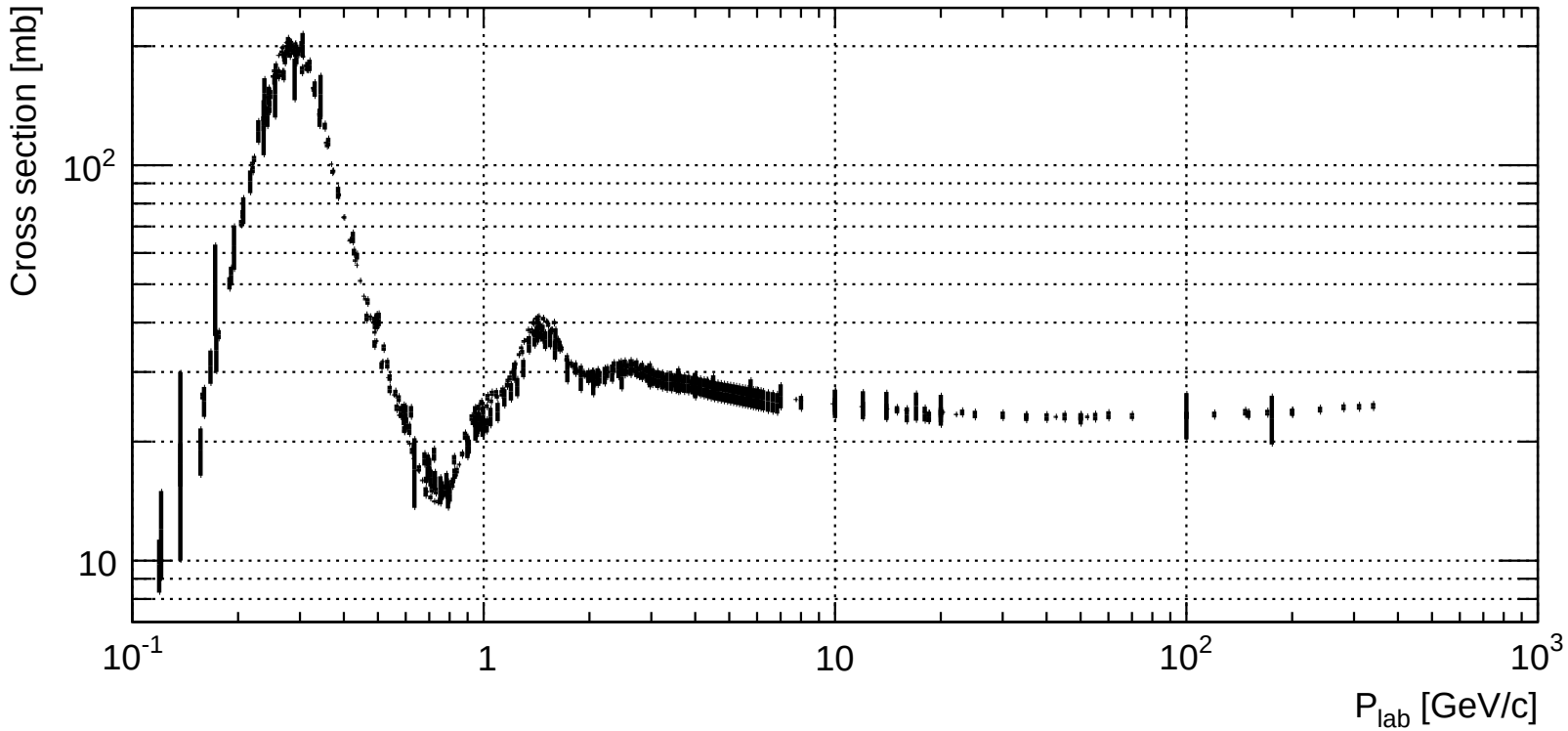
- We have a π^+ beam incident on a proton target
  - we scan a range of π energies and count the number of scattering events as a function of E (the energy of the πp system in its center of mass)
  - at each (fixed) beam energy, the number of scattering events n_i is a random (Poisson distributed) quantity with $\langle n_i \rangle = \sigma_i^2$
  - if n_i is large then Poisson can be approximated by a Gaussian with mean n_i and standard deviation $\sqrt{n_i}$
  - we assume that the points follow the Breit-Wigner formula

$$f(E) \sim \frac{(\Gamma/2)^2}{(E-M)^2 + (\Gamma/2)^2}$$

    where M and Γ are the resonance mass and width, respectively

- As the result of the experiment, we have a set of points **y** (the number of scattering events with their uncertainties) at fixed values of **x** (C.M.S. energy) which we can fit using the least squares method
  - what is the number of parameters to be determined from the fit?
  - is the parameter dependence linear?

Total $p\pi^{\pm}$ cross section (PDG summary)

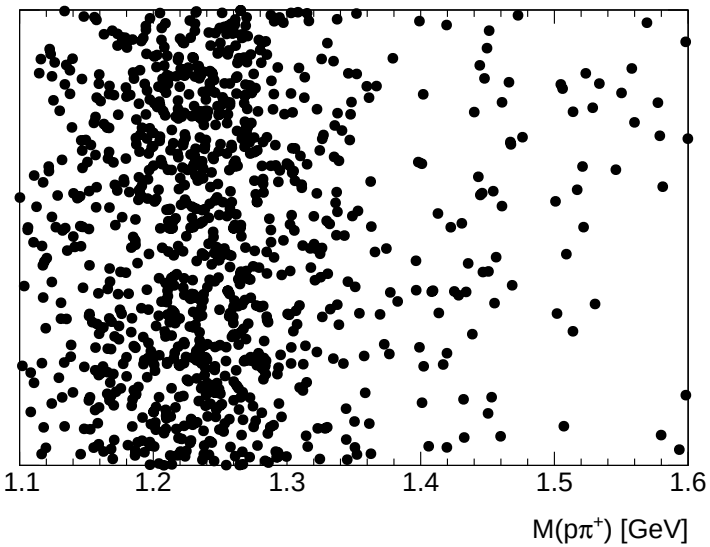


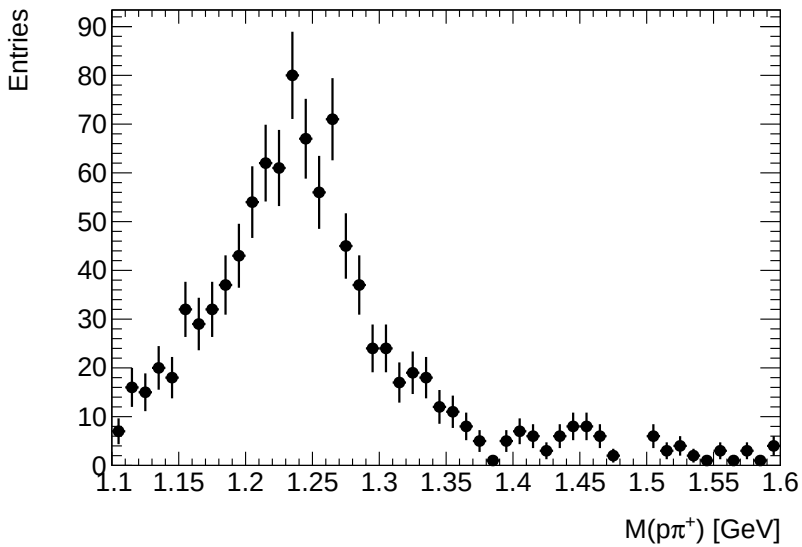
`http://pdg.lbl.gov/2013/hadronic-xsections/rpp2012-pipp_total.dat`

Measuring the parameters of Δ^{++}: method 2

- We are working in the STAR collaboration, studying the d+Au collisions
  - we are looking for proton-π^+ pairs and calculate their invariant masses
  - each proton-π^+ pair measurement results in a number with an uncertainty
- How we can use these measurements to determine the Δ^{++} mass and width?
  - we don't have "measured points" in the sense of the previous problem
  - we have "density" of the points which we need to convert to the number of events, so we can get an estimate on the density uncertainty
- What is usually used in this case is called "binning"
  - the result of the binning is a "histogram"

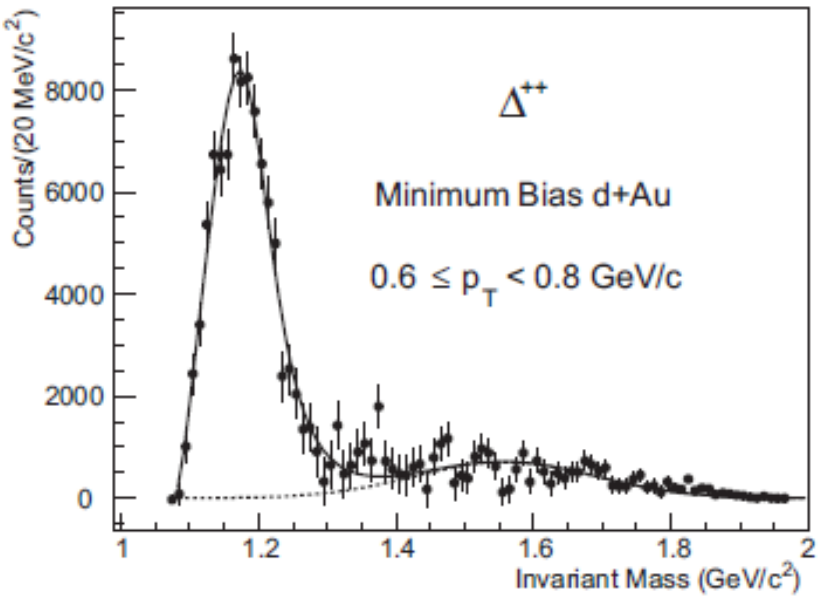
Unbinned vs binned data





is there another resonance at 1.45 GeV?

$p\pi^+$ invariant mass distribution (STAR)



arXiv:0801.0450

Optimal histogram binning

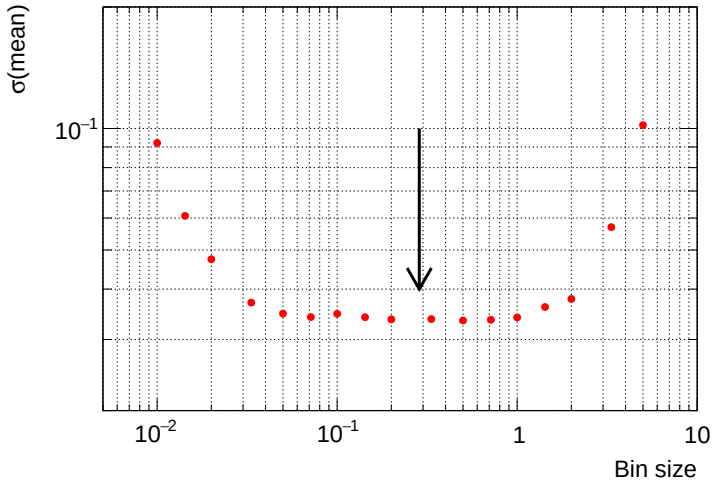
- In general, there is no such thing as universally optimal bin size, it is always problem dependent (are there any narrow peaks etc.)
- Scott: optimal bin size h can be derived from minimizing the integrated mean squared error of the histogram model

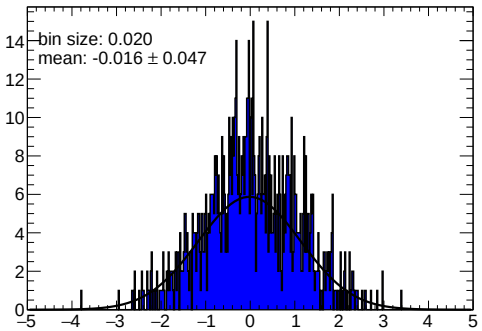
$$\mathrm{IMSE} = \int_{-\infty}^{+\infty} (f_{\mathrm{binned}}(x) - f(x))^2 \, dx$$

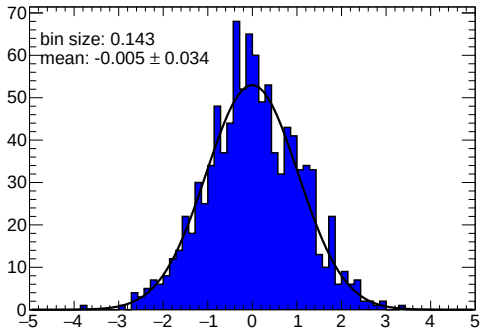
  - IMSE is asymptotically minimized by choosing $h = \left[6/N \int_{-\infty}^{+\infty} (f'(x))^2 dx\right]^{1/3}$
  - for normal distribution, $h = 3.49\sigma/N^{1/3}$
- What if the probability density is far from normal (but still fairly smooth)?

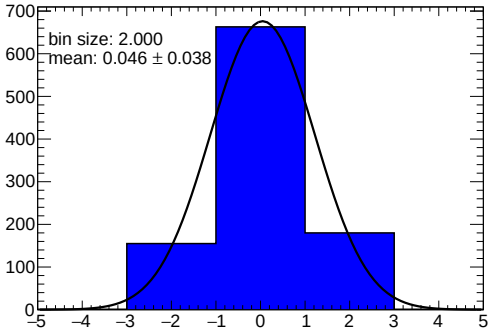
Optimal histogram binning (2)

- Freedman-Diaconis rule: $h = 2\,\mathrm{IQR}/N^{1/3}$, where IQR$=Q_3 - Q_1$ is interquartile range
  - for normal distribution, $\mathrm{IQR} = 1.35\,\sigma$
- Example: $N = 1000$ standard normal random points ($h = 0.29$)









- Bins do not have to be of equal width!
  - a popular option is to define bins such that every bin has approximately the same number of entries (≥ 5)
  - a good rule of thumb for the number of such bins: $2N^{2/5}$