Combined Neyman-Pearson Chi-square: An Improved Approximation to the Poisson-likelihood Chi-square

arXiv:1903.07185

Xiangpan Ji (xji@bnl.gov)

BROOKHAVEN NATIONAL LABORATORY

July 30, 2019

DPF 2019, Northeastern University

Introduction: χ^2_{Poisson}

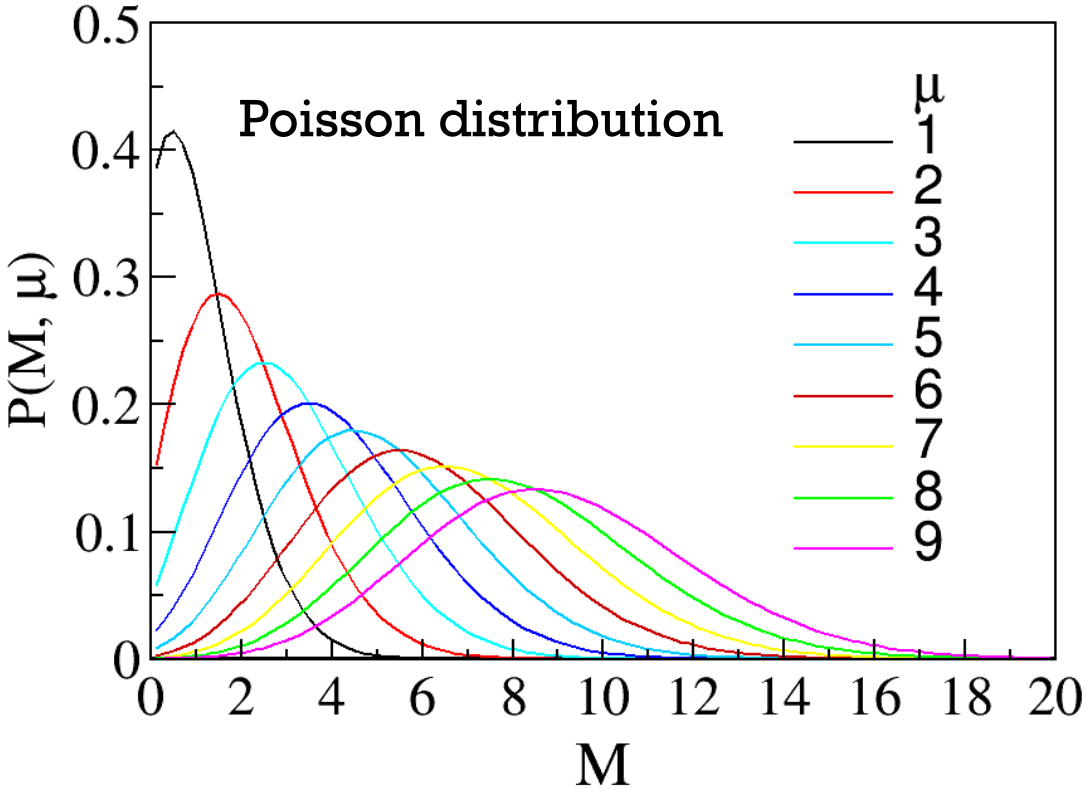
- Poisson distribution: popular for modeling the number of events occur in a measurement

- Poisson-likelihood chi-square: <u>most commonly used in HEP</u> for parameter(θ) estimation

$$\chi^2_{\text{Poisson}} = -2\ln\lambda(\boldsymbol{\theta}) = 2\sum_{i=1}^{n}\left(\mu_i(\boldsymbol{\theta}) - M_i + M_i \ln\frac{M_i}{\mu_i(\boldsymbol{\theta})}\right)$$

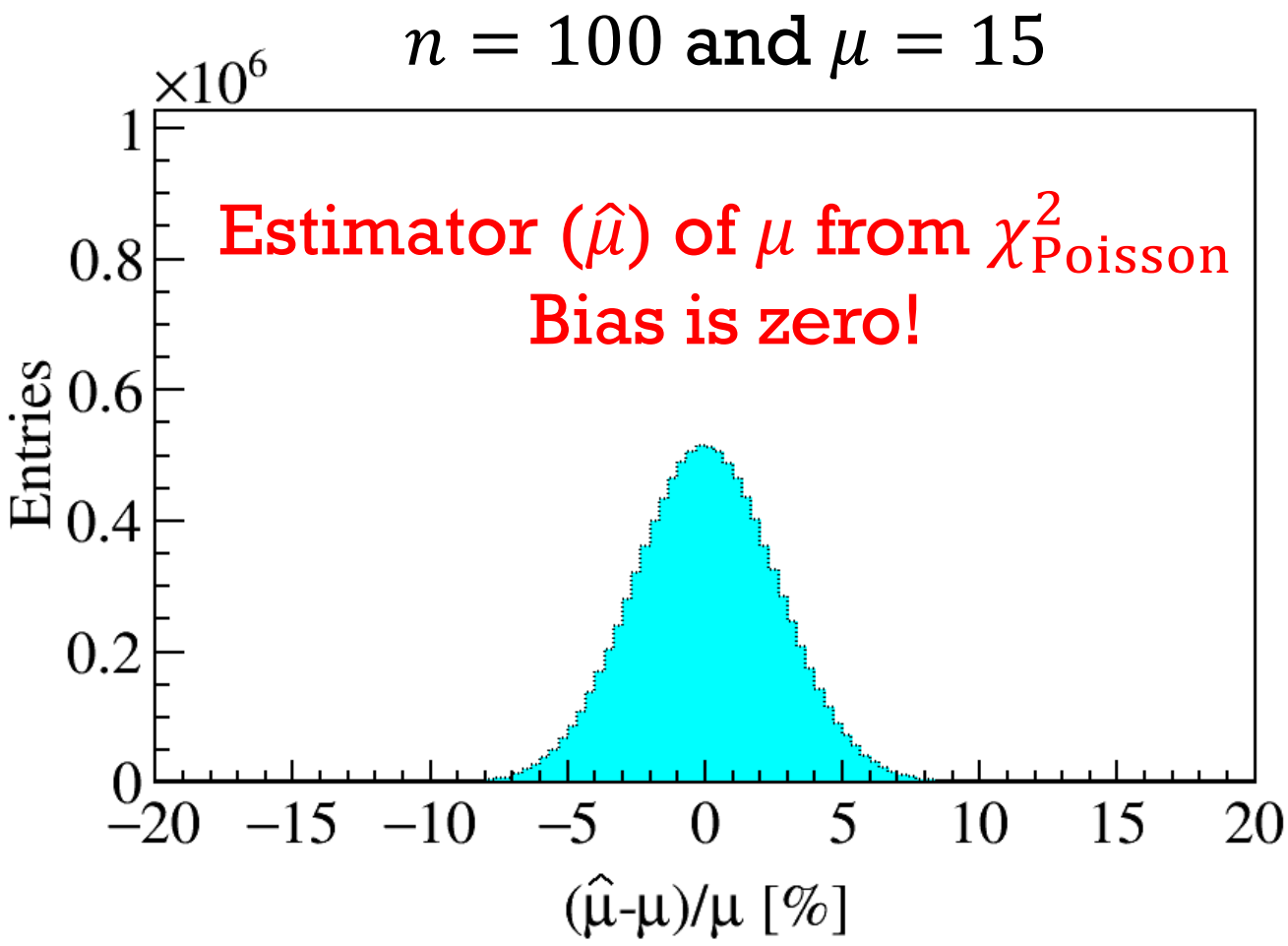
  - minimizing χ^2 → best estimation of θ
  - $\lambda(\theta)$ the likelihood ratio

M: number of events
μ: expected number of events



A Simple Example

A set of n independent counting experiments to measure a common expected value μ



Additional two common chi-squares

- At large statistics, the Poisson distribution can be approximated as Normal (Gaussian) distribution with mean μ_i and standard deviation $\sigma_i^2 = \mu_i \ (or\ M_i)$

- Then, two common chi-squares*: Pearson's and Neyman's chi-squares. <u>Popular because of their close connection to the covariance matrix format (talk later)</u>

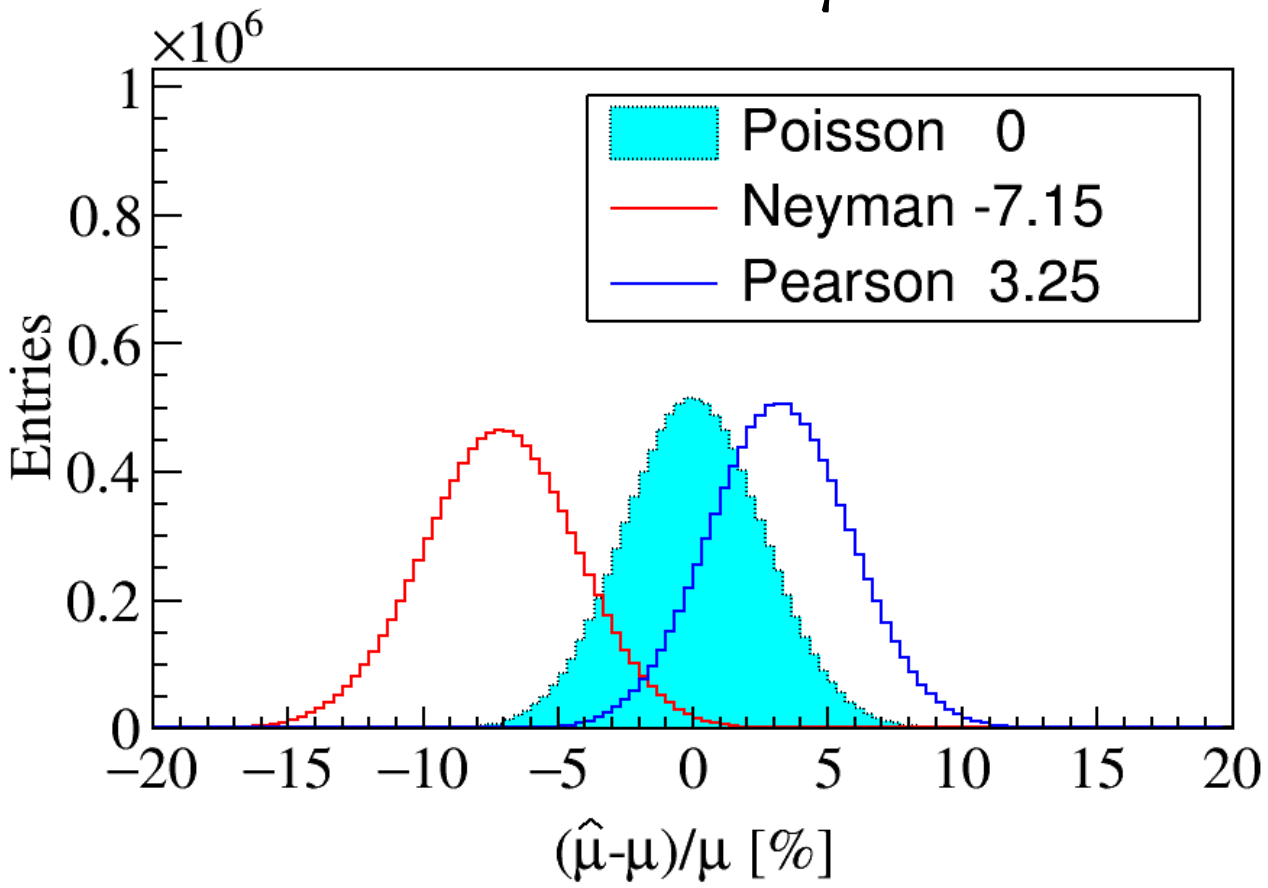
$$\chi^2_{\text{Pearson}} = \sum_i \frac{(\mu_i(\boldsymbol{\theta}) - M_i)^2}{\mu_i(\boldsymbol{\theta})} \qquad \chi^2_{\text{Neyman}} = \sum_i \frac{(\mu_i(\boldsymbol{\theta}) - M_i)^2}{M_i}$$

* *Nucl. Instrum. Meth. 221 (1984) 437*

Bias: χ^2_{Neyman} and χ^2_{Pearson}

Well known:
estimator from χ^2_{Neyman} or χ^2_{Pearson} leads to bias especially when the large statistics condition is not met

$n = 100$ and $\mu = 15$



** Nucl. Instrum. Meth. 221 (1984) 437*
** F. James, Statistical methods in experimental physics. 2006*
** Astrophys. J. 693 (2009) 822*

Biases trace back to χ^2 definition

$$\chi^2_{\text{Poisson}} = 2\sum_{i=1}^{n}\left(\mu - M_i + M_i \ln\frac{M_i}{\mu}\right)$$

$$\chi^2_{\text{Neyman}} = \sum_i \frac{(\mu - M_i)^2}{M_i}$$

$$\chi^2_{\text{Pearson}} = \sum_i \frac{(\mu - M_i)^2}{\mu}$$

Minimizing χ^2

Obtain the estimator $\hat{\mu}$

$$\hat{\mu}_{\text{Poisson}} = \frac{\sum_{i=1}^{n} M_i}{n}$$

$$\hat{\mu}_{\text{Neyman}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{M_i}} \qquad \hat{\mu}_{\text{Pearson}} = \sqrt{\frac{\sum_{i=1}^{n} M_i^2}{n}}$$

$\hat{\mu}_{\text{Neyman}} \leq \hat{\mu}_{\text{Poisson}} \leq \hat{\mu}_{\text{Pearson}}$

(equal only when all M_i are same)

$$\chi^2_{\text{Poisson}} - \chi^2_{\text{Neyman}} \approx -\sum_i^n \frac{2}{3}\frac{(\mu - M_i)^3}{M_i^2} + O\left(\frac{(\mu - M_i)^4}{M_i^3}\right)$$

$$\chi^2_{\text{Poisson}} - \chi^2_{\text{Pearson}} \approx \sum_i^n \frac{1}{3}\frac{(\mu - M_i)^3}{M_i^2} - O\left(\frac{(\mu - M_i)^4}{M_i^3}\right)$$

$\sim x$ -2

Combined Neyman-Pearson Chi-square

$$\chi^2_{\mathrm{CNP}} = \frac{1}{3}\left(\chi^2_{\mathrm{Neyman}} + 2\chi^2_{\mathrm{Pearson}}\right) = \sum_{i=1}^{n} \frac{(\mu - M_i)^2}{3/\left(\frac{1}{M_i} + \frac{2}{\mu}\right)}$$

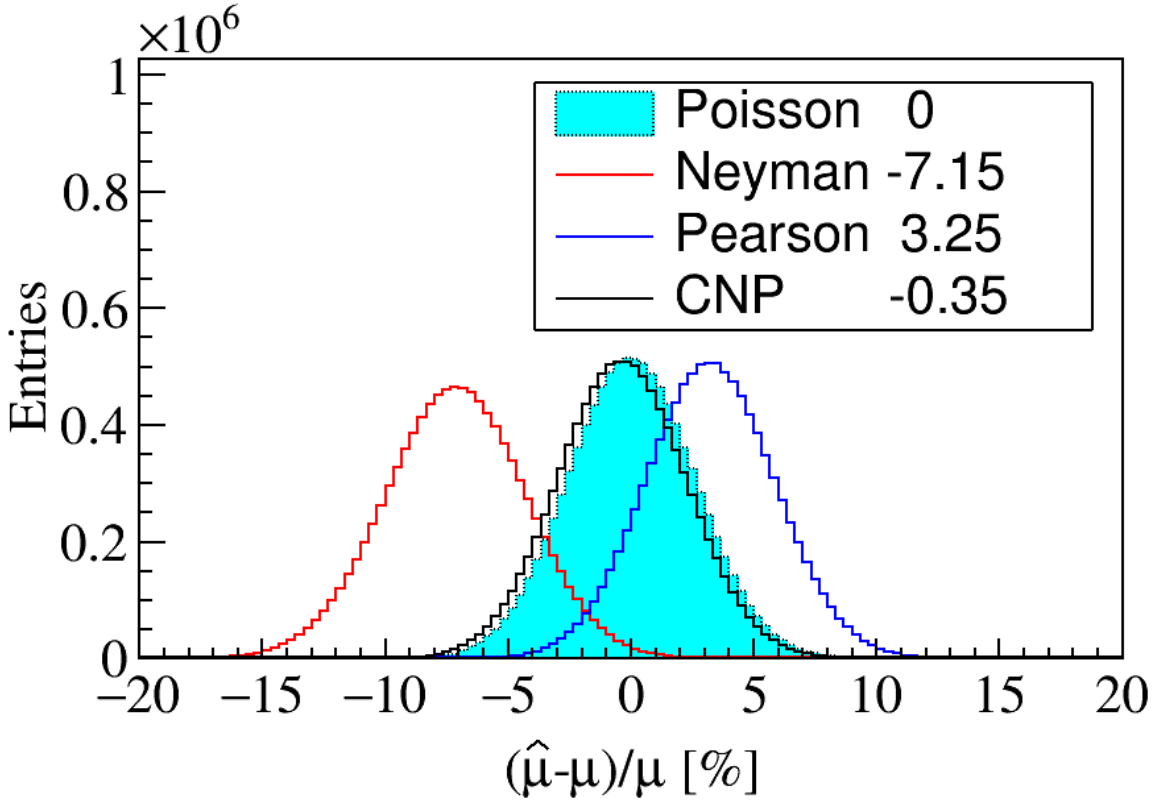
- Approximately equal to $\chi^2_{\mathrm{Poisson}}$ up to $O\left(\frac{(\mu - M_i)^4}{M_i^3}\right)$
- The estimator from χ^2_{CNP} in the previous example is

$$\hat{\mu}_{\mathrm{CNP}} = \sqrt[3]{\frac{\sum_{i=1}^{n} M_i^2}{\sum_{i=1}^{n} \frac{1}{M_i}}} = \sqrt[3]{\hat{\mu}^2_{\mathrm{Pearson}} \cdot \hat{\mu}_{\mathrm{Neyman}}}$$

<u>the geometric mean</u> of two $\hat{\mu}_{\mathrm{Pearson}}$ and one $\hat{\mu}_{\mathrm{Neyman}}$

The Simple Example

$n = 100$ and $\mu = 15$



A Realistic Example: Backgrounds + Systematics uncertainties

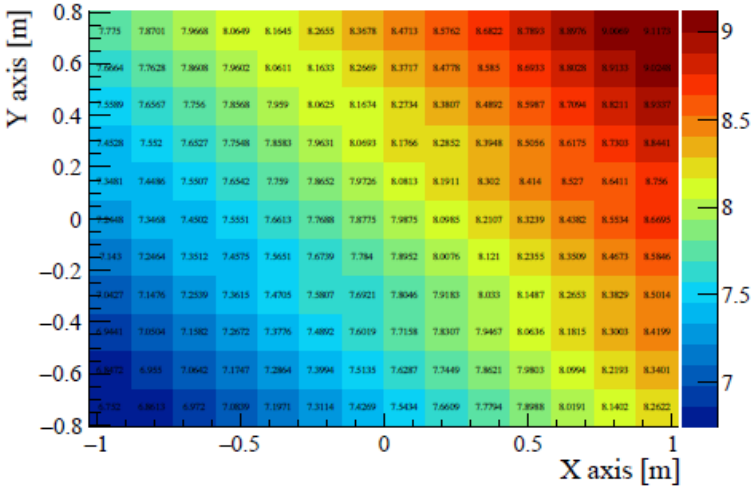
To measure the neutrino oscillation parameters, the PROSPECT measured the reactor neutrino spectrum

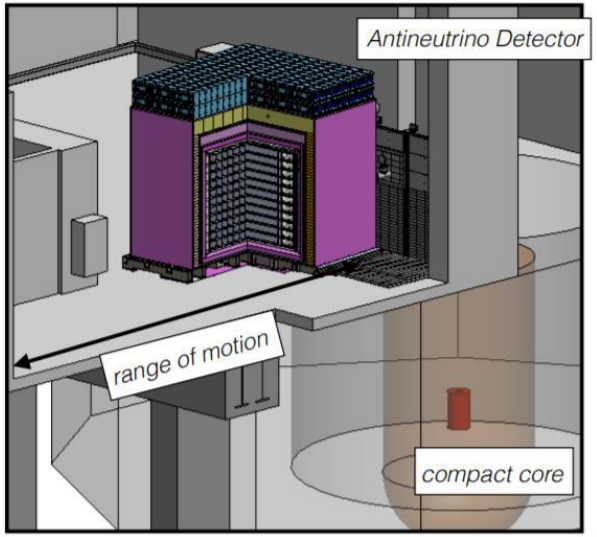
- ~100 segments, ranging from 7~9 meters to the reactor core
- number of events/segment ~200. Dividing into 16 energy bins gives ~12 events/bin
- Backgrounds and systematics uncertainties in each segment

→ An uncommon statistical regime where we need many bins (100×16) but number of events in each bin is not large

** Phys. Rev. Lett. 121 (2018) 251802*

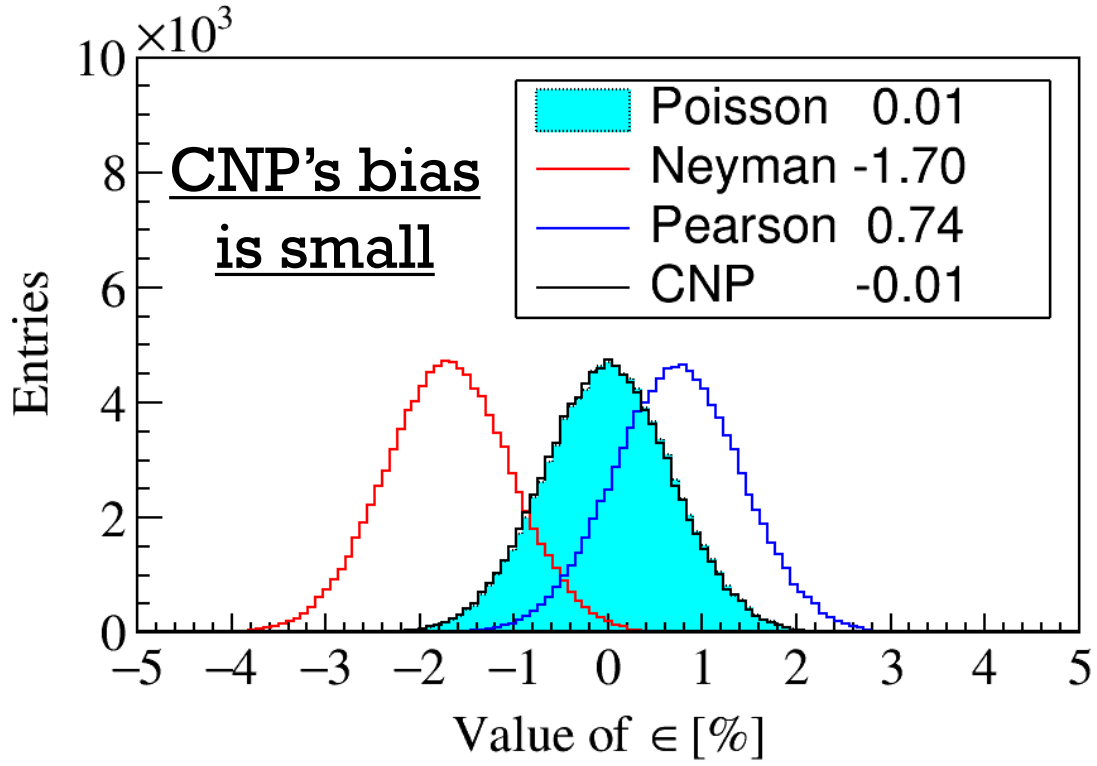
Distance to the reactor core





- Our toy experiment similar to PROSPECT: 100 detectors × 16 energy bins, true signals 30/bin, backgrounds 15/bin

- Fit the free normalization parameter ϵ (the true value is 0 in our toy experiment)



$$\chi^2_{\mathrm{CNP}} = \sum_{d=1}^{100}\sum_{i=1}^{16} \frac{\left(\mu^i_d(1+\epsilon+\epsilon_d)+b^i_d-M^i_d\right)^2}{3/(\frac{1}{M^i_d}+\frac{2}{\mu^i_d(1+\epsilon+\epsilon_d)+b^i_d})} + \sum_{d=1}^{100}\sum_{i=1}^{16} \frac{\left(b^i_d-B^i_d\right)^2}{3/(\frac{1}{B^i_d}+\frac{2}{b^i_d})} + \sum_{d=1}^{100}\left(\frac{\epsilon_d}{0.02}\right)^2$$

Denominator: CNP

Backgrounds

Uncorrelated systematics

Second Advantage
Connection to Covariance Matrix

$$\chi^2 = \sum_i^n \frac{(\mu_i(\theta) - M_i)^2}{\sigma_i^2} \begin{cases} \sigma_i^2 = M_i & \rightarrow \chi^2_{\text{Neyman}} \\ \sigma_i^2 = \mu_i(\theta) & \rightarrow \chi^2_{\text{Pearson}} \\ \sigma_i^2 = 3/\left(\frac{1}{M_i} + \frac{2}{\mu_i(\theta)}\right) & \rightarrow \chi^2_{\text{CNP}} = \frac{1}{3}\left(\chi^2_{\text{Neyman}} + 2\chi^2_{\text{Pearson}}\right) \end{cases}$$

- Both χ^2_{Neyman} and χ^2_{Pearson} are commonly used, partly due to their close connection* to the covariance-matrix formalism

$$\chi^2_{\text{cov}} = (M - \mu(\theta))^T \cdot V^{-1} \cdot (M - \mu(\theta))$$

  - Reduces the fitting number of nuisance parameters of the χ^2 function
  - Leads to a fast minimization of the χ^2 function

* *CDF Note 8661 (1999)*

Shortcoming and Advantage

At low statistics, χ^2_{Neyman} and $\chi^2_{\mathrm{Pearson}}$ have large biases

$\chi^2_{\mathrm{Poisson}}$ difficulty to use together with covariance matrix, can be computationally expensive

χ^2_{CNP} approximation to $\chi^2_{\mathrm{Poisson}}$

- Has much smaller bias
- Naturally connects to covariance matrix

Illustration
Covariance Matrix Implementation

A more complicated PROSPECT example:

There are correlations between the energy bins of the background in the same detector

$$(b_d^i)_{\text{on}} = \sum_j R_d^{ij} b_d^j$$

Backgrounds when the source on and detector on

Backgrounds when the source off and detector on (B_d^i measurement values)

R_d^{ij} is a smearing matrix

0.5 when $i = j$,

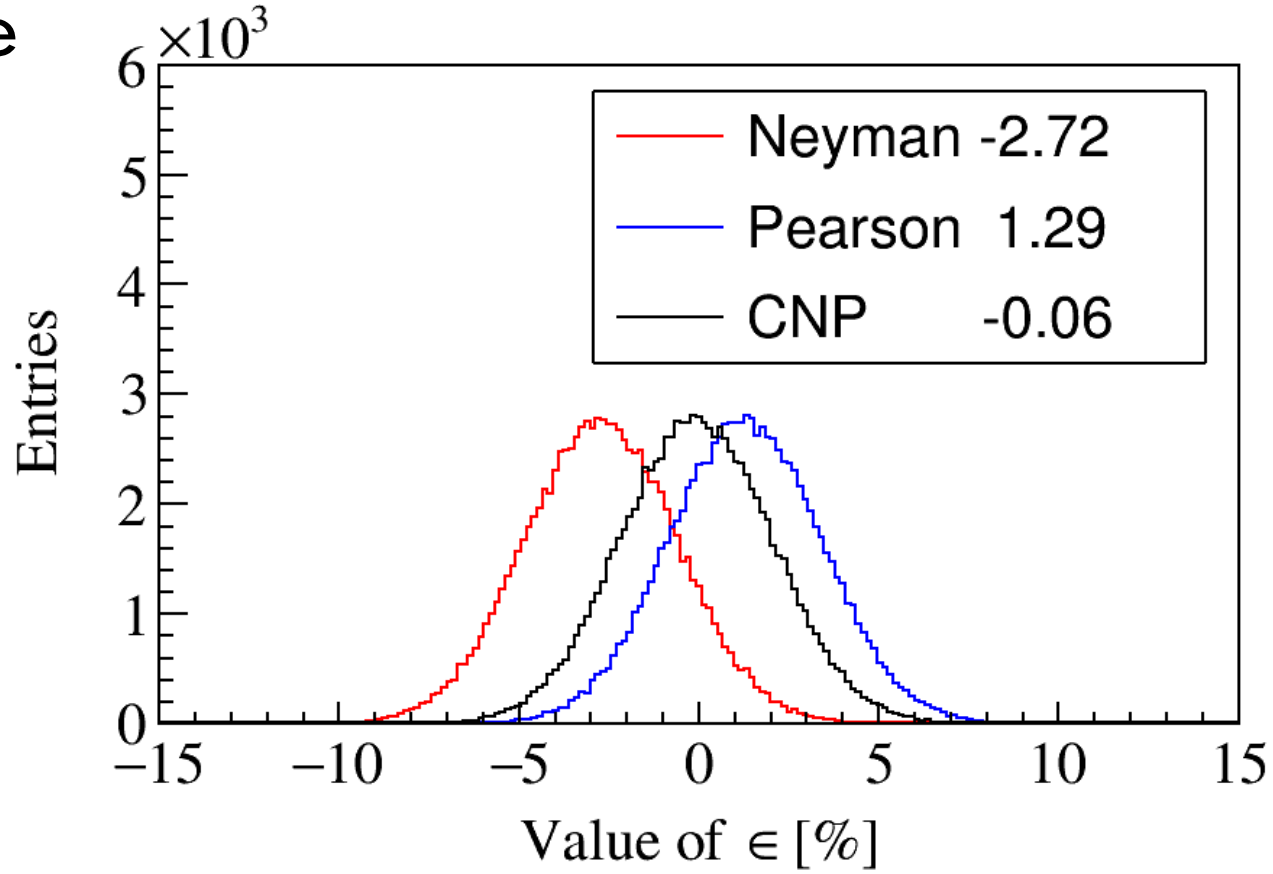
0.25 when $i = j \pm 1$

0 others

$$\left(\chi^2_{\text{CNP}}\right)_{\text{cov}} = (\boldsymbol{\mu}(1+\epsilon) + R \cdot \boldsymbol{b} - \boldsymbol{M})^T \cdot \left(V^{\text{stat}}_{\text{CNP}} + V^{\text{syst}}\right)^{-1} \cdot (\boldsymbol{\mu}(1+\epsilon) + R \cdot \boldsymbol{b} - \boldsymbol{M}) + (\boldsymbol{b} - \boldsymbol{B})^T \cdot \left(V^{\text{bkg}}_{\text{CNP}}\right)^{-1} \cdot (\boldsymbol{b} - \boldsymbol{B}),$$

All results from covariance matrix format (10 detectors × 16 energy bins)

- Poisson-likelihood chi-square doesn't have a equivalent covariance matrix format
- In the covariance matrix format, CNP chi-square also has much smaller bias than Neyman's and Pearson's chi-squares



Summary

- We proposed a new test statistics CNP chi-square: an improved approximation to the Poisson-likelihood chi-square

$$\chi^2_{\mathrm{CNP}} = \frac{1}{3}\left(\chi^2_{\mathrm{Neyman}} + 2\chi^2_{\mathrm{Pearson}}\right) = \sum_{i=1}^{n} \frac{(\mu_i(\theta) - M_i)^2}{3/\left(\frac{1}{M_i} + \frac{2}{\mu_i(\theta)}\right)}$$

- Estimator from CNP chi-square has much smaller bias than that from Neyman's or Pearson's chi-square

- CNP chi-square can use the covariance matrix format