Negative binomial distribution and multiplicities in $p-p(\bar{p})$ collisions

Dariusz Prorok

Institute of Theoretical Physics
University of Wrocław

Zakopane
June 12, 2011

Summary

Likelihood ratio tests are performed for the hypothesis that charged-particle multiplicity distributions measured in the limited pseudo-rapidity windows of $p - p(\bar{p})$ collisions at $\sqrt{s} = 0.9$ and 2.36 TeV are negative binomial. Results indicate that the hypothesis should be rejected in all cases of ALICE-LHC measurements, whereas should be accepted in the corresponding cases of UA5 data. Possible explanations of that and of the disagreement with the least-squares method are given.

based on:

DP, arXiv:1101.0787 [hep-ph]

Motivation

- The fitted quantity is a probability distribution function (p.d.f.), so the most natural way is to use the maximum likelihood (ML) method, where the likelihood function is constructed directly from the tested p.d.f.. Because of Wilks's Theorem one can define a statistic, the distribution of which converges to a χ^2 distribution as the number of measurements goes to infinity. Thus for the large sample the goodness-of-fit can be expressed as a p-value computed with the corresponding χ^2 distribution.

- The most commonly used method, the least-squares method (LS) (called also χ^2 minimization), has the disadvantage of providing only the qualitative measure of the significance of the fit. Only if observables are represented by Gaussian random variables with known variances, the conclusion about the goodness-of-fit equivalent to that mentioned in the first point can be derived.

Negative binomial distribution

$$P(n; p, k) = \frac{k(k+1)(k+2)...(k+n-1)}{n!}(1-p)^n p^k$$

$0 \leq p \leq 1$, k is a positive real number

$n = 0, 1, 2, ...$ - the number of charged particles in an event

$$\bar{n} = \frac{k(1-p)}{p}, \qquad V(n) = \frac{k(1-p)}{p^2}.$$

The maximum likelihood method

For N events in a sample there are N measurements of N_{ch}, say $\mathbf{X} = (X_1, X_2, ..., X_N)$.

$$L(\mathbf{X} \mid p, k) = \prod_{j=1}^{N} P(X_j; p, k)$$

The values $\hat{p}$ and $\hat{k}$ for which $L(\mathbf{X} \mid p, k)$ has its maximum are the maximum likelihood (ML) estimators of parameters p and k.

The log-likelihood function

$$\ln L(\mathbf{X} \mid p, k) = \sum_{j=1}^{N} \ln P(X_j; p, k)$$

Maximization of the log-likelihood function

$$\frac{\partial}{\partial p} \ln L(\mathbf{X} \mid p, k) = \sum_{j=1}^{N} \frac{\partial}{\partial p} \ln P(X_j; p, k) = 0$$

$$\frac{\partial}{\partial k} \ln L(\mathbf{X} \mid p, k) = \sum_{j=1}^{N} \frac{\partial}{\partial k} \ln P(X_j; p, k) = 0$$

For NBD the upper equation gives

$$\bar{n} = \langle N_{ch} \rangle \qquad \Longrightarrow \qquad \frac{1}{p} = \frac{\langle N_{ch} \rangle}{k} + 1$$

Likelihood ratio test - Wilks's theorem

X - a random variable with p.d.f $f(X, \theta)$, which depends on parameters $\theta = (\theta_1,\ \theta_2, ..., \theta_d) \in \Theta$, Θ is an open set in R^d.

$\mathbf{X} = (X_1, ..., X_N)$ - a sample of N independent observations of X

H_0 - a k-dimensional subset of Θ, $k < d$.

The maximum likelihood ratio:

$$\lambda = \frac{\max_{\theta \in H_0} L(\mathbf{X} \mid \theta)}{\max_{\theta \in \Theta} L(\mathbf{X} \mid \theta)}$$

If the hypothesis H_0 is true, *i.e.* it is true that $\theta \in H_0$, then the distribution of the statistic $-2 \ln \lambda$ converges to a χ^2 distribution with $d - k$ degrees of freedom as $N \longrightarrow \infty$.

2-in-1 χ^2 function

Let define the function:

$$\chi^2(\mathbf{X} \mid \theta)_{\theta \in H_0} = -2 \ln \frac{L(\mathbf{X} \mid \theta)}{\max_{\theta' \in \Theta} L(\mathbf{X} \mid \theta')}$$

- The minimum of χ^2 with respect to $\theta \in H_0$ is at $\hat{\theta}$ - the ML estimators.
- The test statistic $\chi^2_{min} = \chi^2(\mathbf{X} \mid \hat{\theta})$ has a χ^2 distribution in the large sample limit.

p-value of the test statistic

The probability of obtaining the value of the test statistic equal to or greater then the value just obtained for the present data set (*i.e.* χ^2_{min}), when repeating the whole experiment many times:

$$p = P(\chi^2 \geq \chi^2_{min}; n_{dof}) = \int_{\chi^2_{min}}^{\infty} f(z; n_{dof}) dz \ ,$$

$f(z; n_{dof})$ - the χ^2 p.d.f.

$n_{dof} = d - k$ - the number of degrees of freedom

2-in-1 χ^2 function for binned data

Let divide the sample $\mathbf{X} = (X_1, X_2, ..., X_N)$ into m bins defined by the number of measured charged particles $\{0, 1, 2, 3, ..., m-1\}$ and with n_i entries in the ith bin, $N = \sum_{i=1}^{m} n_i$.

$$\chi^2 = -2\ln\lambda = 2\sum_{i=1}^{m} n_i \ln\frac{n_i}{\nu_i}$$

$$\nu_i = N \cdot P(i-1; p, k)$$

Details in: G. Cowan, *Statistical data analysis*, (Oxford University Press, Oxford, 1998)

2-in-1 χ^2 function for binned data, *cont.*

$$\chi^2(p,k) = 2\sum_{i=1}^{m} n_i \ln\frac{n_i}{\nu_i} = -2\;N\sum_{i=1}^{m} P_i^{ex}\ln\frac{P(i-1;p,k)}{P_i^{ex}}$$

$P_i^{ex} = n_i/N$ - the experimental probability (frequency)

- This χ^2 function depends explicitly on the number of events in the sample!
- But does not depend on actual experimental errors!

The χ^2 function of the least-squares method

The sum of squares of normalized residuals:

$$\chi^2_{LS}(p,k) = \sum_{i=1}^{m} \frac{(P_i^{ex} - P(i-1;p,k))^2}{err_i^2}$$

err_i - the uncertainty of the ith measurement

NOT MINIMIZED HERE !!!

but

$$\chi^2_{LS} = \chi^2_{LS}(\hat{p},\hat{k})$$

$\hat{p},\hat{k}$ - ML estimators of parameters p and k

Sources of data

- UA5 Collaboration:

  R. E. Ansorge *et al.*, Z. Phys. C **43**, 357 (1989)

- ALICE Collaboration:

  K. Aamodt *et al.*, Eur. Phys. J. C **68**, 89 (2010)

Results of the LS analysis of ALICE data available in:

T.Mizoguchi, M. Biyajima, Eur. Phys. J. C **70**, 1061 (2010)

Tests of NBD for UA5 and ALICE data at $\sqrt{s} = 0.9$ TeV

Case	N_{event}	χ^2/n_{dof}	p-value [%]	χ^2_{LS}/n_{dof} with errors: quad.	sum	stat.	$\sqrt{n_i}/N_{ev}$
UA5 $\mid \eta \mid < 0.5$	8550.0	0.21	99.998	0.07	na	na	0.20
ALICE $\mid \eta \mid < 0.5$	149663.2	14.5	0	0.73	0.38	2.46	15.1
ALICE $\mid \eta \mid < 1.0$	128476.5	36.9	0	1.72	0.95	11.0	38.0
ALICE $\mid \eta \mid < 1.3$	60142.8	24.3	0	2.21	1.28	15.2	25.8
UA5 $\mid \eta \mid < 1.5$	8550.0	1.1	28.9	0.36	na	na	1.14

Tests of NBD for ALICE data at $\sqrt{s} = 2.36$ TeV

Case	N_{event}	χ^2/n_{dof}	p-value [%]	χ^2_{LS}/n_{dof} with errors: quad.	sum	stat.	$\sqrt{n_i}/N_{ev}$
ALICE $\mid \eta \mid < 0.5$	38970.79	7.0	0	0.76	0.43	3.8	7.5
ALICE $\mid \eta \mid < 1.0$	37883.99	18.5	0	2.29	1.36	18.8	20.3
ALICE $\mid \eta \mid < 1.3$	22189.40	18.2	0	4.25	2.60	39.6	20.0

ALICE data at $\sqrt{s} = 0.9$ TeV with UA5 N_{event}

Case	N_{event}	χ^2/n_{dof}	p-value [%]	χ^2_{LS}/n_{dof} with errors: quad.	sum	stat.	$\sqrt{n_i}/N_{ev}$
UA5 $\mid\eta\mid< 0.5$	8550.0	0.21	99.998	0.07	na	na	0.20
ALICE $\mid\eta\mid< 0.5$	8550.0	0.83	70.4	0.73	0.38	2.46	0.86
ALICE $\mid\eta\mid< 1.0$	8550.0	2.45	$5 \cdot 10^{-5}$	1.72	0.95	11.0	2.53
ALICE $\mid\eta\mid< 1.3$	8550.0	3.46	$7 \cdot 10^{-13}$	2.21	1.28	15.2	3.66
UA5 $\mid\eta\mid< 1.5$	8550.0	1.1	28.9	0.36	na	na	1.14

Conclusions

1. Results of the likelihood ratio tests suggest that the hypothesis about the NBD of charged-particle multiplicities measured by the ALICE Collaboration in limited pseudo-rapidity windows of proton-proton collisions at $\sqrt{s} = 0.9$ and 2.36 TeV should be rejected.
2. The significant systematic errors of ALICE data are the reasons for acceptable values of the LS test statistic for the narrowest pseudo-rapidity window cases.
3. The size of the sample is very important for the validation of a hypothesis about the p.d.f. of an observable. If the hypothesis is true, the distribution is exact for the whole population. Thus for the very large samples (as in all ALICE cases) the measured distribution should be very close to that postulated. How "close" is controlled by the size of the sample (discrepancies $\sim \sqrt{n_i}/N_{event}$), not by the size of errors, rather.