EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-EP/88-174

7 December 1988

**A METHOD FOR CORRECTING OBSERVED DISTRIBUTIONS OF MULTIPLICITIES
USING THE MAXIMUM ENTROPY PRINCIPLE**

Christer Fuglesang

CERN, Geneva, Switzerland

and

Department of Physics, University of Stockholm, Stockholm, Sweden

ABSTRACT

A difficulty that can arise when one corrects an observed distribution (of multiplicities, for example) to obtain the true distribution is pointed out. It will arise if each true multiplicity can give any of a number of observed multiplicities. A method, based on the maximum entropy principle, is proposed to get a good approximation of the true distribution. The method has been developed in the work of analysing the multiplicity data of the UA5 experiment at CERN.

(Submitted to Nuclear Instruments and Methods in Physics Research)

1. INTRODUCTION

Let us consider an experiment where one wants to determine the distribution $T(n)$ (where T is for true) of some discrete variable n . For simplicity, let $n = 1, 2, \dots, N$. Owing to inefficiencies in the detector and the experimental method, one will in general not measure n , but something denoted here by m . The distribution of m will be denoted by $O(m)$ (where O is for observed). For each n , there is a probability distribution $p_n(m)$, that n true will give m observed. All N p_n -distributions are put together in a matrix called the probability matrix P_{mn} . The connection between the distributions $T(n)$ and $O(m)$ can then be written:

$$O(m) = \sum_{n=1}^N P_{mn} T(n) , \quad m = 1, 2, \dots, M . \quad (1)$$

In this article, how to get the true distribution $T(n)$, given $O(m)$ and P_{mn} , will be discussed. If one deals with limited statistics, eq. (1) cannot be solved directly to give $T(n)$ (even if P_{mn} is a square matrix). The solution will, in general, be meaningless, for example with $T(n=i) < 0$ for some i . With limited statistics it is important to distinguish between the samples [which will be denoted by primes: $O'(m)$, $T'(n)$, etc.] and the populations [$O(m)$, $T(n)$, etc.] Unless $p_n(m)$ are very narrow distributions (in principle, δ -functions), the sample $T'(n)$, which gave rise to the data $O'(m)$, can never be determined. These points will be clarified in section 2.

In section 3 a method is proposed that uses the principle of maximum entropy [1,2] to get a smooth population $T(n)$, which with a large probability is close to the one that gave the observed sample $O'(m)$. The method has been applied to charged-particle multiplicity distributions obtained by the UA5 experiment at the CERN $p\bar{p}$ Collider, at centre-of-mass energies of 200 and 900 GeV [3]. It was in this context that the method was developed, and it was motivated by the fact that an old method [4] to obtain $T(n)$ gave waves in $T(n)$ which were unphysical. The old method tried to find an approximate solution to eq. (1) in an iterative way, simultaneously applying smoothness criteria. It was found that the properties of the waves are connected with which smoothness criteria one uses [5].

How the new method works is shown in section 4, and a summary and short discussion are given in section 5.

2. DIFFICULTIES WHEN GOING FROM THE OBSERVED TO THE TRUE DISTRIBUTION

For the sake of clarity, from now on, n is taken to be the bin number for the produced particles in some reaction ($n = 1, 2, \dots, N$), and m the bin number for the observed particles ($m = 1, 2, \dots, M$)*). (At the moment trigger inefficiencies are ignored; thus all reactions are assumed to be detected.) The distribution $p_n(m)$ then denotes the probability that n produced particles give m observed particles**). There are N such distributions. When the experiment is performed, one samples not only in $T(n)$, but also in these N distributions $p_n(m)$. Thus the data sample $O'(m)$ is a product of all $N + 1$ samples $T'(n)$ and $p'_n(m)$ ($n = 1, 2, \dots, N$). Since one can never get to know the samples $p'_n(m)$ [it is only the populations $p_n(m)$, often estimated by Monte Carlo simulations, that one can

*) Usually bin number 1 corresponds to 0 particles, bin 2 to 1 particle, etc. However, in full phase space, $n = 1$ corresponds to 2 particles, $n = 4$ to 4 particles, etc., whereas $m = 1$ corresponds to 1 observed, $m = 2$ corresponds to 2 observed, etc.

**) More strictly: that the number of particles corresponding to bin n gives the number of observed particles corresponding to bin m .

know], it is clear that it is impossible to solve relation (1) to get $T'(n)$. In fact, relation (1) should be seen as two different equations, as illustrated in fig. 1: one equation connecting the populations $O(m)$, $T(n)$ and $p_n(m)$ ($n = 1, 2, \dots, N$), and the second equation relating the samples $O'(m)$, $T'(n)$ and $p'_n(m)$ ($n = 1, 2, \dots, N$). It is only when the statistical fluctuations in all the samples p'_n are small, that one can safely put $p'_n = p_n$ and solve for $T'(n)$ in eq. (1). When this is not the case, which it is likely to be if $p_n(m)$ are not narrow distributions, there is one equation where P_{mn} is known but $O(m)$ is unknown, and another equation where $O'(m)$ is known but P'_{mn} is unknown. Therefore, the aim must be to derive a population $T(n)$ that describes the data well in a statistical sense.

There is a statistical relationship between the samples and the populations. If the data bins $O'(m)$ are independent and if the number of events contributing to each bin is approximately Gaussian distributed^{*)}, the sum

$$\chi^2 = \sum_{m=1}^M \frac{\left(O'(m) - \sum_{n=1}^N P_{mn}T(n)\right)^2}{\sigma_m^2} \quad (2)$$

is χ^2 distributed (with σ_m denoting the standard deviation of the content in bin number m). The χ^2 distribution has the expectation value $E(\chi^2) = M$ and its maximum is at $M - 2$, where M is the number of data points ($M \geq 2$). Thus when the true population $T(n)$ is inserted into eq. 2, together with the data sample $O'(m)$ and P_{mn} (which is assumed to be known), it most likely gives $\chi^2 \approx M$. Note especially that it is not likely to have a very small χ^2 if M is not small. If P_{mn} is a square matrix it is even possible to find a $T(n)$ that gives $\chi^2 = 0$, but this is just the exact solution to eq. (1), which as mentioned often leads to unphysical results. Thus one does not want to minimize χ^2 [as a function of $T(n)$], but find a χ^2 -value around M .

Another, independent, statistical test of $T(n)$ is a sign test of the series

$$\left\{O'(m) - \sum_{n=1}^N P_{mn}T(n)\right\}, \quad m = 1, 2, \dots, M. \quad (3)$$

The number of sign changes ν in this series has an expectation value $E(\nu) \approx M/2$. Thus there are two criteria to use when searching for the population $T(n)$. But there is, of course, an infinite set of distributions $T(n)$ that will give $\chi^2 \approx M$ in eq. (2) and $\nu \approx M/2$ in eq. (3), and the question is which distribution should one choose?

3. METHOD FOR DERIVING $T(n)$

It has been shown [2] that the only consistent way of choosing one probability distribution, out of a set of distributions that can all equally well explain the data, is by using the maximum entropy method. Furthermore, maximum entropy always selects the simplest possible result, containing the bare minimum of structure needed to fit the data. The principle is to maximize the entropy:

$$S = - \sum_{n=1}^N T(n) \ln T(n) \quad (4)$$

under whatever constraints are imposed by the data. Ideally, then, one would like to maximize S under the constraints that χ^2 from eq. (2) is around M , and ν from eq. (3) is around $M/2$. However,

*) To satisfy this assumption, when χ^2 is to be computed the $O'(m)$ bins are grouped in the tail such that in each group there are at least five events.

this gives severe mathematical and computational problems owing to the non-linearity of these two constraints. On the other hand, under linear constraints in $T(n)$, it is fairly simple to maximize eq. (4) if the number of variables and constraints are not too large [6] (see Appendix).

Using the moments of the observed distribution, it is possible to formulate linear conditions on $T(n)$. One requires that the observed moments are exactly reproduced by the distribution $T(n)$:

$$\langle [O'(m)]^q \rangle \equiv \sum_{m=1}^M m^q O'(m) = \sum_{m=1}^M m^q \left[\sum_{n=1}^N P_{mn} T(n) \right] = \sum_{n=1}^N \left(\sum_{m=1}^M m^q P_{mn} \right) T(n) \quad (5)$$

for some q . Note that $q = 0$ gives the normalization condition. By choosing q -values such that after the maximization of S one has $\chi^2 \approx M$ and $\nu \approx M/2$, one finds a distribution for $T(n)$ which, with a large probability, is close to the true population.

The disadvantage of the conditions (5) is that they are an arbitrary way of formulating linear constraints, and the choice of q -values is also somewhat arbitrary. A motivation for using these conditions is that, for each q , the function $m^q O'(m)$ emphasizes different intervals in the variable m , as shown in fig. 2. It is also clear that the moments are less sensitive to statistical fluctuations than, for example, individual bins. Furthermore, the method has been found to work well empirically with these constraints. It does not, for example, work well if one groups bins of $O'(m)$, i.e. requires that

$$\sum_{m=m_i}^{m_{i+1}} O'(m) = \sum_{n=1}^N \left(\sum_{m=m_i}^{m_{i+1}} P_{mn} \right) T(n), \quad 1 = m_1 < m_2 < \dots < m_l = M \quad (6)$$

for a set of m_i ($i = 1, 2, \dots, l$). Actually, another advantage with the method is that no grouping of the data [or $T(n)$] is necessary (apart from when χ^2 is calculated). One could think of using, for example, central moments instead of algebraic moments, but that is basically the same thing, since the two sets of moments are functions of each other.

To get rid of the arbitrariness in the choice of q -values, an algorithm is used that searches for a set of consecutive integers [e.g. $q = (-1, 0, 1, 2)$] as few as possible and such that $\nu \approx M/2$, $\chi^2 \approx M$ and χ^2 roughly constant for close-by sets [e.g. $q = (-2, -1, 0, 1, 2)$]. One starts with $q = (0, 1)$ only, and calculates χ^2 . New constraints are added, one at a time, as long as χ^2 decreases with more than one unit. A new constraint has always the q -value one step up or down, and the direction is chosen in which χ^2 decreases most. Empirically, it has been found that at some point χ^2 becomes more or less constant, and one stops. In some cases χ^2 is significantly larger than M at this point, and then more constraints are added, until $\chi^2 \approx M$ and is again about constant. The criterion regarding ν is seldom critical, unless the original population is rather irregular. In that case, one might have to search further if ν is not around $M/2$. This was never the case when treating the UA5 data of Ref. [3], but an example is given in the next section.

This algorithm for selecting q -values is not an important feature of the method described in this article. The important features are to obtain χ^2 and ν as expected in a statistical sense. The linear constraints are introduced to get a mathematically treatable problem, while the algorithm serves, in principle, to define the q -values *a priori*, rather than *a posteriori* when one can look at the derived populations. Furthermore, the algorithm is straightforward to automatize in a computer program. An alternative way could actually be to produce a lot of distributions, using different sets of q -values, and then by inspection to choose the most 'physical' one. However, this gives a subjective bias, which usually is unwanted.

4. TEST OF THE METHOD

The method has been tested in the following way. A certain true population $T(n)$ was assumed as well as a probability matrix P_{mn} . The probabilities $p_n(m)$ were taken to be Gaussian distributed with a mean $\langle m \rangle_n$ and variance σ_n^2 given by

$$\begin{aligned}\langle m \rangle_n &= 1.0 + 0.8 n \\ \sigma_n &= -0.8 + 0.08n ,\end{aligned}\tag{7}$$

which is a rough approximation of the widest probability matrices found in the UA5 experiment [3,4] (fig. 3). Simulated experiments were performed by sampling a true multiplicity n from $T(n)$ and then sampling m from the corresponding distribution $p_n(m)$. Each ‘experiment’ was done with 7000 events (roughly the size of UA5 data samples [3]) and gave a simulated observed distribution $O'(m)$. The method described in the previous section was then applied to give a true distribution, which will be referred to sometimes as the corrected distribution.

First $T(n)$ was taken to be a negative binomial distribution

$$T(n) = P_{NB}(n; \bar{n}, k) = \binom{n+k-1}{k-1} \left(\frac{\bar{n}/k}{1 + \bar{n}/k} \right)^n \frac{1}{(1 + \bar{n}/k)^k}\tag{8}$$

with parameters $\bar{n} = 26.0$ and $k = 3.2$. The negative binomial has been shown to give good fits to many multiplicity distributions in high-energy particle reactions [3, 7], and the parameters used here are typical for UA5 data. Ten runs were made and, in table 1, the moments of the corrected distribution are compared with those of the population. The average of $\langle n \rangle$ may look a little low, but in fact $\langle n \rangle$ of all 70000 generated events was 25.95, so one can conclude that the agreement is excellent. In fig. 4 the distributions achieved are plotted together with the original negative binomial. Error bars, corresponding to a total of 7000 events, are shown at every fourth point. The distributions are mostly well within the error bars, and the overall shapes agree very well with the negative binomial.

To be in a better position to look for any systematic deviation, in fig. 5 is plotted the difference between the corrected distributions $[C(n)]$ and $T(n)$ in units of χ^2 contribution, multiplied by the sign of the difference:

$$\chi_n^2 = [T(n) - C(n)] \frac{|T(n) - C(n)|}{\sigma_n^2} ,\tag{9}$$

where σ_n is the statistical error of bin n corresponding to a total of 7000 events:

$$\sigma_n = \sqrt{7000 P_{NB}(n; 26.0, 3.2)} .\tag{10}$$

No significant systematic deviation is seen, although there may be a slight positive shift around $n = 70$. There are larger fluctuations than elsewhere at the very lowest n 's, and two runs gave very large χ^2 contributions there. This seems to be due to the relative narrow distributions $p_n(m)$ for small n (see fig. 3a) and the fact that the events here are few (of the order of 10). Thus the statistical fluctuations are large, and there is not much freedom when maximizing the entropy while fulfilling the constraints with low q -values.

The q -values found were mostly in the region $[-2, 4]$, but one run used $[-4, 5]$ and another $[-2, 7]$, which were the ‘extreme’ regions. That $q = -3$ was necessary in two runs, and even $q = -4$ in one, is due to the $p_n(m)$ distributions being narrow for low n 's and thus there is little freedom for

‘give and take’ between the smallest multiplicities. To get reasonable χ^2 in the low- m region, it is therefore sometimes necessary to put several constraints there.

Secondly, an irregular distribution was taken for $T(n)$:

$$T(n) = 0.8 P_{NB}(n; 26.0, 3.2) + 0.2 P_{Poisson}(n; 50.0) . \quad (11)$$

This gives a second peak around $n = 50$, as shown in fig. 6. It was now necessary to check not only χ^2 , but also that $\nu \approx M/2$. The resulting moments from three runs are compared with the input in table 2 and the agreement is again seen to be very good. The q -values the method produced are also given in table 2. In fig. 6 the three output curves are compared with the input population. One can see that the corrected curves make the second peak smoother, but it is still clearly seen. It is not surprising that the output is smoother than the true population, since the maximum entropy method gives the smoothest possible distribution that can explain the data and since the peak is smeared out by the probability matrix. The population for the observed distribution, i.e. $O(m) = \sum P_{mn} T(n)$, is shown by the crosses in fig. 6, and one sees how the peak structure is lost by the transformation (physically corresponding to the resolution of the detector). Indeed, it is surprising that the corrected distributions show as much of the peak structure as they do.

Finally, the result of a correction is not sensitive to which q -values one uses, as long as $\chi^2 \approx M$. In table 3, the moments for the UA5 data at 900 GeV full phase space [3], corrected with different sets of q ’s, are compared. The differences are an order of magnitude less than the statistical errors. In fig. 7, the q set given by the algorithm described in section 3 ($[-3, 5]$) is compared with the two cases $q = [-4, 5]$ and $q = [-2, 4]$. One sees that the variation is greatest around the peak, where it is of the order of the statistical error. Note that when many conditions are imposed, the corrected distribution becomes wavy. Observations of this kind have been part of the motivation for the algorithm described above to determine which q -values to use.

It should be mentioned that when correcting the real data of UA5, the trigger inefficiencies also had to be corrected for. This was done by applying the method to distributions of triggered events: $t(n) = \epsilon_n T(n)$ (where ϵ_n is the probability that an n -prong event will trigger the detector), which is normalized to one: $\sum t(n) = 1$. Then $T(n)$ is given by $T(n) = t(n)/\epsilon_n$ and by renormalizing such that $\sum T(n) = 1$.

5. SUMMARY AND DISCUSSION

A method has been described for finding a good approximation of the true distribution $T(n)$ in experiments where the true distribution is ‘smeared’ by a probability matrix P_{mn} giving an observed distribution sample $O'(m)$. The method is based on the maximum entropy principle [1,2], and the probability entropy S is maximized under the linear constraints that some moments of the observed distribution be exactly reproduced. The constraints are chosen such that a χ^2 test and a sign test of the true distribution give the statistically expected values.

The method gives very good results when the true population is fairly regular. However, if the true population is irregular, the output from the method is likely to give a smoother distribution, although the moments will be correct. This is because the true distribution has been smeared by the probability matrix, and the maximum entropy method selects the simplest distribution that can explain the observed data. If one expects an irregular distribution from physical reasons (e.g. a resonance) it would be better to try a fit of a certain functional form. In general, one should always make fits of functional forms to the data through the matrix P_{mn} (as done in ref. [3]), rather than by correcting the distribution and then making a fit to the corrected distribution.

It should again be stressed that it is impossible to correct the observed *sample* to get the true *sample* which was realized when the experiment was performed, when one has a ‘smearing’-matrix

P_{mn} . However, one can try to find the population that is the most likely in a statistical sense to have given the sample. The method presented here is not the optimal method, in the sense of mathematical statistics, but should give a good approximation and is useful practically in physical applications.

An extra difficulty in a real experiment, compared to the test in section 4, is that, in general, one has systematic errors in P_{mn} and $O'(m)$. In the case of UA5 the systematic errors are particularly large for low multiplicities [3]. When using this method it was sometimes necessary to carry out a special treatment for the 1-3 lowest bins, e.g. to group them.

Acknowledgement

I should like to thank Professor G. Ekspong for good advice, interesting discussions, and for pointing out the maximum entropy method. I wish to express my thanks also to everyone else in the UA5 Collaboration, in particular Dr. B. Åsman. Finally, I am very grateful to Drs. T. Elving and L. Eldén, of the University of Linköping, for a helpful and interesting discussion on the maximum entropy principle and related subjects.

APPENDIX

One wants to maximize the entropy

$$S = - \sum_{n=1}^N T(n) \ln T(n) \quad (A1)$$

under the linear constraints

$$\sum_{n=1}^N a_{kn} T_n = b_k, \quad k = 1, 2, \dots, K. \quad (A2)$$

[where in this paper one has

$$a_{kn} = \sum_{m=1}^M m^{q_k} P_{mn}, \quad b_k = \sum_{m=1}^M m^{q_k} O'(m).] \quad (A3)$$

Using Lagrangian multipliers λ_k ($k = 1, 2, \dots, K$) one forms

$$F(T_n, \lambda_k) = - \sum_{n=1}^N T_n \ln T_n + \sum_{k=1}^K \lambda_k \left(\sum_{n=1}^N a_{kn} T_n - b_k \right). \quad (A4)$$

The solution to eq. (A1) under the constraints (A2) is given by the (unique [6]) solution of

$$\frac{\partial F}{\partial T_n} = - \ln T_n - 1 + \sum_{k=1}^K \lambda_k a_{kn} = 0, \quad n = 1, 2, \dots, N \quad (A5)$$

$$\frac{\partial F}{\partial \lambda_k} = \sum_{n=1}^N a_{kn} T_n - b_k = 0, \quad k = 1, 2, \dots, K. \quad (A6)$$

Equation (A5) gives

$$T_n = \exp \left[\sum_{k=1}^K \lambda_k a_{kn} - 1 \right], \quad (A7)$$

which, put into eq. (A6), gives

$$\sum_{n=1}^N a_{kn} \exp \left[\sum_{k=1}^K \lambda_k a_{kn} - 1 \right] - b_k = 0. \quad (A8)$$

The last expression is a system of K non-linear equations for the K unknown λ_k 's. A generalized Newton-Raphson method (see, for example, ref. [8]) has been used to solve the system. This can always be done [6], if a start value close enough to the solution is found. We have always used $\lambda_1 = \lambda_2 = \dots = \lambda_K = 0$, which worked well. [However, it was necessary to renormalize all constraints (A2) to $b_k = 1.0$, $k = 1, 2, \dots, K$.] Finally, given λ_k , T_n is found by eq. (A7).

Table 3

Results, with different q-se', on the UA5 900 GeV full phase space data [3] (where there are 86 data bins, i.e. $M = 86$). The first row gives the 'default' result and the second row the statistical error.

r'	χ^2	ν	$\langle n \rangle$	D	C_2	C_3	C_4	C_5
$-3 \rightarrow 5$	84.5	49	35.65	19.54	1.300	2.080	3.89	8.17
-	-	-	0.25	0.25	0.007	0.036	0.13	0.52
$-2 \rightarrow 4$	94.0	43	35.65	19.54	1.301	2.080	3.89	8.16
$-2 \rightarrow 5$	95.2	46	35.66	19.53	1.300	2.078	3.88	8.16
$-2 \rightarrow 6$	89.9	48	35.67	19.53	1.300	2.077	3.88	8.15
$-2 \rightarrow 7$	77.9	50	35.68	19.52	1.299	2.076	3.88	8.15
$-3 \rightarrow 4$	85.6	47	35.65	19.54	1.300	2.080	3.89	8.17
$-3 \rightarrow 4$	84.7	49	35.65	19.54	1.300	2.080	3.89	8.17
$-3 \rightarrow 7$	79.0	52	35.71	19.50	1.298	2.073	3.87	8.12
$-4 \rightarrow 4$	92.5	45	35.73	19.49	1.298	2.071	3.86	8.11
$-4 \rightarrow 5$	84.7	47	35.74	19.49	1.297	2.070	3.86	8.09
$-4 \rightarrow 6$	80.0	48	35.74	19.49	1.297	2.069	3.86	8.09
$-4 \rightarrow 7$	79.5	48	35.74	19.49	1.297	2.069	3.86	8.09

Figure captions

Fig. 1 Illustration of how relation (1) is in fact two different equations, one connecting the populations (above) and one giving a relation between the samples that are realized when the experiment is performed (below).

Fig. 2 The distribution of $m^q O(m)$ (normalized to 1) for $q = -3, -2, \dots, 7$ (from left to right), when taking $O(m) = \text{negative binomial}(22.0, 3.0)$ [see eq. (8)].

Fig. 3 The probability distributions for observing m particles, when (a) 3, (b) 12 and (c) 60 particles are produced. The histograms show UA5 distributions (from Monte Carlo simulations) at 900 GeV for particles in the pseudorapidity region $|\eta| < 5.0$ [3], and the curves are the parametrizations used in this paper. Both sets are normalized to the number of Monte Carlo events used for the histograms.

Fig. 4 Ten corrected distributions (curves) compared with the true input population of a negative binomial distribution (diamonds). On every fourth point, error bars are shown corresponding to 7000 events.

Fig. 5 The difference between the ten corrected distributions and $T(n)$ in fig. 4, in units of χ^2 contributions per bin, χ_n^2 [eq. (9)].

Fig. 6 Three corrected distributions (curves) compared with the true input population, given by eq. (11) (diamonds). On every fourth point, error bars are shown corresponding to 7000 events. The crosses show the observed population.

Fig. 7 Corrected 900 GeV, full phase space data [3], using three sets of q -values. The diamonds show the ‘default’-set ($q = [-3, 5]$), as given by the algorithm described in section 3, and as published in ref. [3]. The curves are obtained when using $q = [-2, 4]$ and $q = [-4, 7]$.

Fig. 1

Fig. 2

Fig. 3

Fig. 4

Fig. 5

Fig. 6

Fig. 7