EHEF NIJMEGEN

HEN-455
1 Dec. 2004

Estimating the Uncertainties of Factorial Moments

W. J. Metzger

Experimental High Energy Physics Group
Radboud University
Nijmegen, The Netherlands

1 Introduction

Factorial moments of the charged-particle multiplicity distribution are commonly used to study dynamical fluctuations. Very briefly, one wants to study the scaling properties of the probability moments of the multiplicity distribution, $C_q(M)$, defined for a division of phase space into M regions by

$$C_q(M) = \frac{1}{M} \sum_{m=1}^{M} \frac{\mu_q(m)}{\Big(\mu_1(m)\Big)^q} \tag{1}$$

where $\mu_q(m)$ is the q^{th} moment of the probability distribution that a particle occurs in phase-space interval m. Assuming that statistical fluctuations are Poissonian, the $C_q(M)$ can be estimated by the factiorial moments,[1] which for N events are given by

$$F_q(M) = \frac{1}{M} \sum_{m=1}^{M} \frac{\frac{1}{N}\sum_{i=1}^{N} f_q(n_{mi})}{\Big(\frac{1}{N}\sum_{i=1}^{N} f_1(n_{mi})\Big)^q} \tag{2}$$

where $f_q(n_{mi})$ is the unnormalized factorial moment of order q for the i^{th} event in bin m, which is given by

$$f_q(n_{mi}) = \prod_{j=0}^{q-1} (n_{mi} - j) = n_{mi}(n_{mi}-1)(n_{mi}-2)...(n_{mi}-q+1) \tag{3}$$

where n_{mi} is the number of particles in bin m of the i^{th} event. These $F_q(M)$ are also known as "vertical" factorial moments. They involve an average over bins of an average over events.

Performing the averages in the opposite order results in the "horizontal" factorial moments,

$$F_q^{\mathrm{H}}(M) = \frac{\frac{1}{N}\sum_{i=1}^{N} \frac{1}{M}\sum_{m=1}^{M} f_q(n_{mi})}{\Big(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{M}\sum_{m=1}^{M} f_1(n_{mi})\Big)^q} \tag{4}$$

which are more economically computed, but which have not been proven to be unbiased estimators of the $C_q(M)$.[2]

If the binning is chosen such that the average number of particles is the same for each bin, the denominator in (2) can be taken outside of the sum over M, and the vertical factorial moment is equal to the horizontal factorial moment.

2 Estimated statistical uncertainty —Error Propagation

In this section we derive the formulae to estimate the statistical uncertainties on the factorial moments. We follow the conventional method of "error propagation". The procedure is straightforward, but has its pitfalls.

It starts with the quantities which are actually measured, the number of particles in each bin, n_{mi}, from which moments, *i.e.*, averages over bins or events, are calculated and their variances estimated. In the case of the vertical factorial moments these moments are averages over events, for horizontal factorial moments they are averages over bins. From these moments the factorial moments themselves are calculated, and the variances of these moments are "propagated" to obtain the variances of the factorial moments.

We should keep in mind that this "propagation" involves a non-linear transformation and hence is only a first-order approximation. Also, moments estimation is non-parametric, *i.e.*, no knowledge of the true distribution of the moments is used. The estimated variance can therefore be an overestimation.

In the following $\langle\ldots\rangle$ denotes an average over events, $\langle\ldots\rangle_M$ an average over bins, and $\langle\ldots\rangle_{NM}$ an average over both bins and events.

2.1 Vertical factorial moments

Rewriting (2), we have

$$F_q(M) = \frac{1}{M}\sum_{m=1}^{M}\frac{\langle f_q(n_{mi})\rangle}{\langle f_1(n_{mi})\rangle^q} \; . \tag{5}$$

which has variance

$$V[F_q(M)] = \frac{1}{M^2}V\left[\sum_{m=1}^{M}\frac{\langle f_q(n_{mi})\rangle}{\langle f_1(n_{mi})\rangle^q}\right] \; . \tag{6}$$

As is well known,[3,4] for N events, the estimated variance of $\langle f_q(n_{mi})\rangle$ is given by

$$V[\langle f_q(n_{mi})\rangle] = \frac{\left\langle f_q^2(n_{mi})\right\rangle - \left\langle f_q(n_{mi})\right\rangle^2}{N-1} \tag{7}$$

and the estimated covariance of the averages in the numerator and denominator of (5) by

$$\operatorname{cov}\left[\langle f_q(n_{mi})\rangle, \langle f_1(n_{mi})\rangle\right] = \frac{\langle f_q(n_{mi})f_1(n_{mi})\rangle - \langle f_q(n_{mi})\rangle\langle f_1(n_{mi})\rangle}{N-1} \; . \tag{8}$$

The variance of the quotient of the two quantities, A and B^q, is, following the usual "propagation of errors"[3,4]

$$V\left[\frac{A}{B^q}\right] = \frac{V[A]}{B^{2q}} + q^2 \frac{A^2}{B^{2q+2}} V[B] - 2q \frac{A}{B^{2q+1}} \operatorname{cov}[A,B] \ . \tag{9}$$

The variance of a single term in the sum over M in (5) is then

$$\begin{aligned} V\left[\frac{\langle f_q(n_{mi})\rangle}{\langle f_1(n_{mi})\rangle^q}\right] &= \frac{V[\langle f_q(n_{mi})\rangle]}{\langle f_1(n_{mi})\rangle^{2q}} + q^2 \frac{\langle f_q(n_{mi})\rangle^2}{\langle f_1(n_{mi})\rangle^{2q+2}} V[\langle f_1(n_{mi})\rangle] \\ & -2q \frac{\langle f_q(n_{mi})\rangle}{\langle f_1(n_{mi})\rangle^{2q+1}} \operatorname{cov}\left[\langle f_q(n_{mi})\rangle, \langle f_1(n_{mi})\rangle\right] \ , \end{aligned} \tag{10}$$

which can be substituted in (6).

2.2 Horizontal factorial moments

Rewriting (4), we have

$$F_q^{\mathrm{H}}(M) = \frac{\frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} f_q(n_{mi})}{\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} f_1(n_{mi})\right)^q} = \frac{\langle f_q(n_{mi})\rangle_{NM}}{\langle f_1(n_{mi})\rangle_{NM}^q} \ . \tag{11}$$

The variance of $F_q^{\mathrm{H}}(M)$ is then

$$\begin{aligned} V\left[F_q^{\mathrm{H}}(M)\right] &= \frac{V\left[\langle f_q(n_{mi})\rangle_{NM}\right]}{\langle f_1(n_{mi})\rangle_{NM}^{2q}} + q^2 \frac{\langle f_q(n_{mi})\rangle_{NM}^2}{\langle f_1(n_{mi})\rangle_{NM}^{2q+2}} V\left[\langle f_1(n_{mi})\rangle_{NM}\right] \\ & -2q \frac{\langle f_q(n_{mi})\rangle_{NM}}{\langle f_1(n_{mi})\rangle_{NM}^{2q+1}} \operatorname{cov}\left[\langle f_q(n_{mi})\rangle_{NM}, \langle f_1(n_{mi})\rangle_{NM}\right] \end{aligned} \tag{12}$$

where

$$V\left[\langle f_q(n_{mi})\rangle_{NM}\right] = \frac{\left\langle f_q^2(n_{mi})\right\rangle_{NM} - \left\langle f_q(n_{mi})\right\rangle_{NM}^2}{NM-1} \tag{13}$$

$$\operatorname{cov}\left[\langle f_q(n_{mi})\rangle_{NM}, \langle f_1(n_{mi})\rangle_{NM}\right] = \frac{\langle f_1(n_{mi}) f_q(n_{mi})\rangle_{NM} - \langle f_1(n_{mi})\rangle_{NM}\langle f_q(n_{mi})\rangle_{NM}}{NM-1} \tag{14}$$

2.3 The *wrong* formula

It might be tempting to reason as follows when the binning has been constructed such that $\langle f_1(n_{mi})\rangle$ is the same for all m. Then $\langle f_1(n_{mi})\rangle = \langle f_1(n_{mi})\rangle_{NM} = N_{\mathrm{tr}}/NM$, where N_{tr} is the total number of tracks in the N events, and both (2) (4) can be written

$$\begin{aligned} F_q(M) = F_q^{\mathrm{H}}(M) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{m=1}^{M}\frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle_{NM}^q} = \left(\frac{NM}{N_{\mathrm{tr}}}\right)^q \langle f_q(n_{mi})\rangle_{NM} & (15) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\langle\frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle^q}\right\rangle_M = \left\langle\left\langle\frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle^q}\right\rangle_M\right\rangle \ . & (16) \end{aligned}$$

Regarding $N_{\rm tr}$ as a constant would then give, similarly to (7),

$$V\left[F_q(M)\right] = \frac{\left\langle \left\langle \frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle^q} \right\rangle_M^2 \right\rangle - \left\langle \left\langle \frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle^q} \right\rangle_M \right\rangle^2}{N-1} \ , \tag{17}$$

which is identical to that from using just the first term of (10) in (6) and to the first term of (12), as expected under the assumption of no uncertainty on $N_{\rm tr}$.

Similarly, the covariance of F_q for different numbers of bins, L and M is given by

$$\operatorname{cov}\left[F_q(L), F_q(M)\right] = \frac{\left\langle \left\langle \frac{f_q(n_{li})}{\langle f_1(n_{mi})\rangle^q} \right\rangle_L \left\langle \frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle^q} \right\rangle_M \right\rangle - \left\langle \left\langle \frac{f_q(n_{li})}{\langle f_1(n_{mi})\rangle^q} \right\rangle_L \right\rangle \left\langle \left\langle \frac{f_q(n_{mi})}{\langle f_1(n_{mi})\rangle^q} \right\rangle_M \right\rangle}{N-1} \ . \tag{18}$$

This formula is formula (4.3) of Frank Botterweck's thesis[6] and formula (9.10) of Yuan Hu's thesis,[7] both having made the assumption that the uncertainty on $\langle f_1(n_{mi})\rangle$, and thus on $N_{\rm tr}$, is negligible.

But $N_{\rm tr}$ is *not* a constant. If the experiment were repeated, giving a new sample of N events, $N_{\rm tr}$ could well be, indeed would almost certainly be, different. Moreover, the value of of $N_{\rm tr}$, and hence of $\langle f_1(n_{mi})\rangle_{NM}$, is correlated with that of $\langle f_q(n_{mi})\rangle_{NM}$. Hence, (17) and (18) are *incorrect.* Neglect of this correlation leads to an overestimation of the statistical uncertainty on F_q.

The overestimation is most dramatic for $q = 1$ where the numerator and denominator of (2) are identical, which implies $F_1 = 1 \pm 0$. However, neglect of the covariance leads to a non-zero value. The overestimation decreases as q increases.

2.4 Covariance for different M

Since the interest in the factorial moments is not so much in their values, but rather in their behavior as a function of M, in particular whether they obey a power law,

$$F_q(M) = b_q M^{\phi_q} \ , \tag{19}$$

it is necessary to have the entire covariance matrix, $\operatorname{cov}\left[F_q(L), F_q(M)\right]$. At present I only have the *wrong* formula, (18).

If we were interested in the behavior of F_q with q, we would also need the covariance matrix, $\operatorname{cov}\left[F_q(M), F_r(M)\right]$.

3 Estimated statistical uncertainty—Resampling

An alternative to error propagation, albeit a very computer intensive one, is to use resampling methods such as the bootstrap.[5] This consists of making S new samples from the original sample of N events. Each of these samples also consists of N events, which are chosen randomly with replacement from the original sample.

The uncertainty on a quantity is estimated by (essentially) the r.m.s. of the S values of the quantity obtained from these samples, the variance being given by

$$V\left[F_q(M)\right] = \frac{\sum_{s=1}^{S}\left(F_{qs}(M) - \frac{\sum_{t=1}^{S} F_{qt}(M)}{S}\right)^2}{S-1} \tag{20}$$

$$= \frac{S}{S-1}\left[\frac{\sum_{s=1}^{S} F_{qs}^2(M)}{S} - \left(\frac{\sum_{s=1}^{S} F_{qs}(M)}{S}\right)^2\right] , \tag{21}$$

where $F_{qs}(M)$ is the value of $F_q(M)$ found in sample s. Similarly, the resampling estimate of the covariance is given by

$$\mathrm{cov}\left[F_q(L), F_q(M)\right] = \frac{S}{S-1}\left[\frac{\sum_{s=1}^{S} F_{qs}(L)F_{qs}(M)}{S} - \frac{\sum_{s=1}^{S} F_{qs}(L)}{S}\frac{\sum_{s=1}^{S} F_{qs}(M)}{S}\right] . \tag{22}$$

Clearly, this method runs into difficulties if the quantity whose variance is being estimated depends heavily on the tails of distributions, since one cannot obtain events further into the tails than those present in the original sample. The method should therefore not be used for high-order factorial moments. This problem can be avoided if one has a good Monte Carlo model. Then instead of resampling the data, one can use S Monte Carlo samples of N events.

4 Comparison of propagation and resampling

Using a sample of about 9.7 million good tracks from 804125 selected Z decays from L3, variances estimated from error propagation and from resampling are compared. This data sample is not corrected for detector effects. It is very similar, if not identical, to that used by Yuan Hu[7]—some event or track selections may be slightly different. Both of the above-described resampling methods are used. In the first, 250 new samples of 804125 events are formed from the original data sample.* In the second, 250 samples are formed by randomly choosing 804125 events from the simulated JETSET sample (known as qe535) at detector level, which contains 2397175 selected events.

If the uncertainty is correctly estimated, the variance of the distribution of values of F_2 should agree with the average value of the estimated variance on F_2. They do not when the *wrong* formula, (17), is used to estimate the variance, as is shown in Figure 1 where the ratio of the uncertainty obtained by resampling to the average of that estimated by (17) is plotted as a function of M for y, p_{t}, ϕ_{major} and ϕ_{random}. The use of (17) overestimates the uncertainty by as much as a factor 5 at low M, while at large M the overestimation is less. Resampling detector-level JETSET leads to a similar conclusion (Figure 2).

On the other hand, the uncertainty estimated by (6) agrees reasonably well for all M with that found by resampling, as is shown in Figures 3 and 4.

*The number of samples is less in the case of the wrong formula.

Having seen, using resampling, that the *wrong* formula for the variance, (17), is indeed wrong, and by a large amount, it can be no surprise that the wrong formula for covariance, (18), also gives large disagreement with that obtained by resampling. This is shown in Figure 5.

Further, this disagreement is not just a consequence of the wrong formula for the covariance. The correlations themselves,

$$\rho\left[F_q(L), F_q(M)\right] = \frac{\operatorname{cov}\left[F_q(L), F_q(M)\right]}{\sqrt{V\left[F_q(L)\right] V\left[F_q(M)\right]}} \tag{23}$$

are also incorrect, as seen in Figure 6. Note that (18) overestimates the correlations, particularly at small L or M. The results of resampling the data and resampling JETSET are in good agreement, as seen in Figure 7 where the difference between the correlations estimated by the two methods is plotted.

Not having a correct formula, it is clear that we must use resampling to obtain the covariance matrix. Possibilities are to

- obtain the covariance matrix by resampling, either from data or JETSET; or
- obtain the correlation matrix by resampling and the variances by (10) and (6), from which the covariance matrix can be calculated from (23).

Best would be to do both for both data and JETSET and take any differences as systematic uncertainties.

5 Conclusions

Formulae (10) and (6) for the case of vertical and (12) for the case of horizontal factorial moments provide adequate estimates of the statistical uncertainty on the F_q, agreeing with the estimates found by resampling. Formulae (17) and (18) are based on a mistaken assumption and provide an incorrect estimate of the uncertainty.

In the absence of a correct formula for the estimated covariance of estimates of F_q for different numbers of bins, two suggestions are made, involving either incorrect estimation followed by a correction factor determined from resampling or to simply use the resampling estimate.

References

1. A. Białas and R. Peschanski, *Nucl. Phys.* B273 (1986) 703, B308 (1988) 857.

2. Chen Gang and Liu Lianshou, *Phys. Rev. D* 65 (2002) 094030.

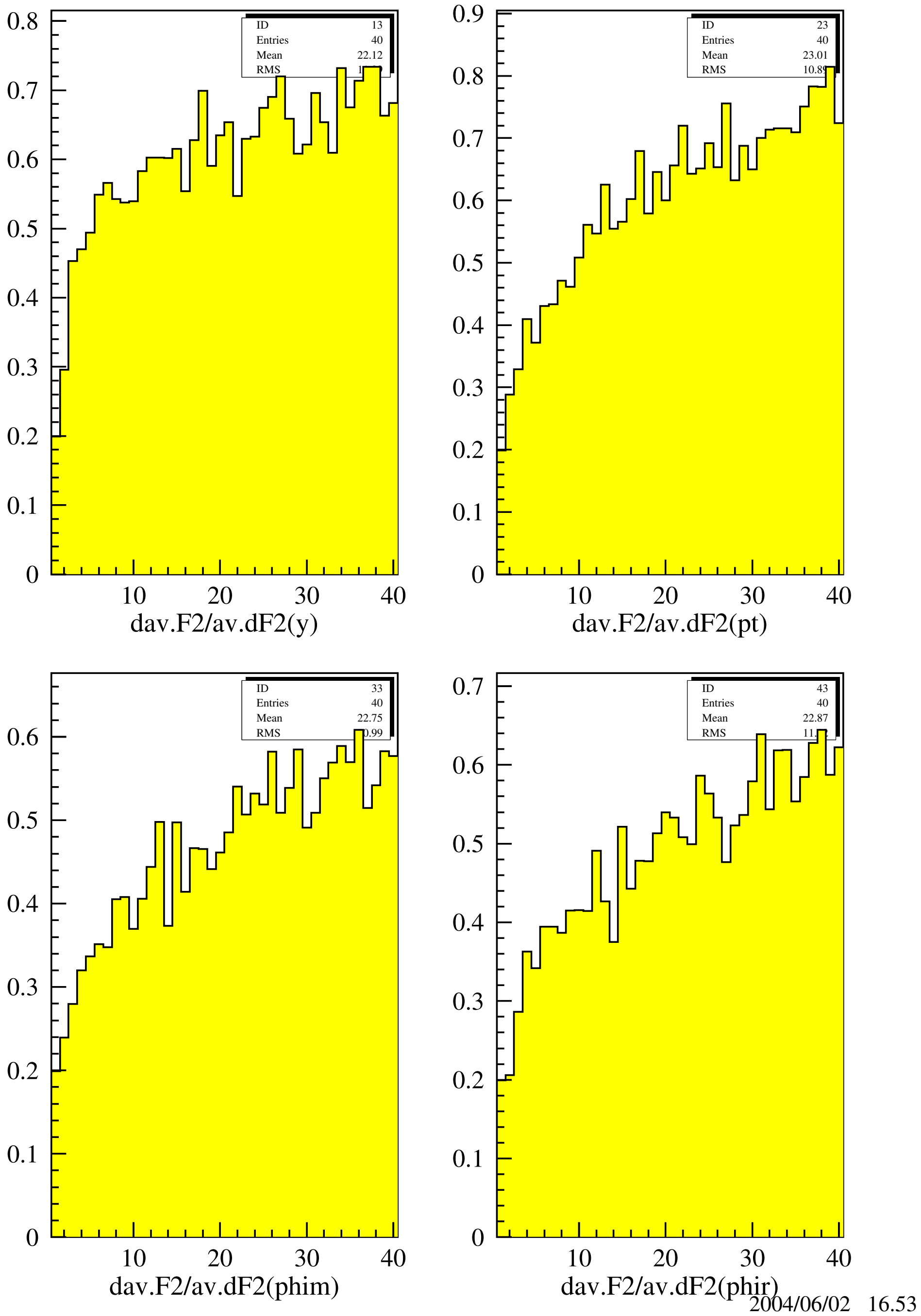
3. W. J. Metzger, *Statistical Methods in Data Analysis*, HEN-343.

4. W. T. Eadie *et al.*, *Statistical Methods in Experimental Physics* (North-Holland, 1971).

5. Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, 1993).

6. F.P.M. Botterweck, Ph.D. thesis, Nijmegen (1992).

7. Yuan Hu, Ph.D. thesis, Nijmegen (2002).



2004/06/02 16.53

Figure 1: Ratio of the uncertainty on F_2 obtained by resampling to the average estimated uncertainty on F_2 obtained using (17) as a function of M for y, p_{t}, ϕ_{major} and ϕ_{random} determined from 50 resamplings of the data.

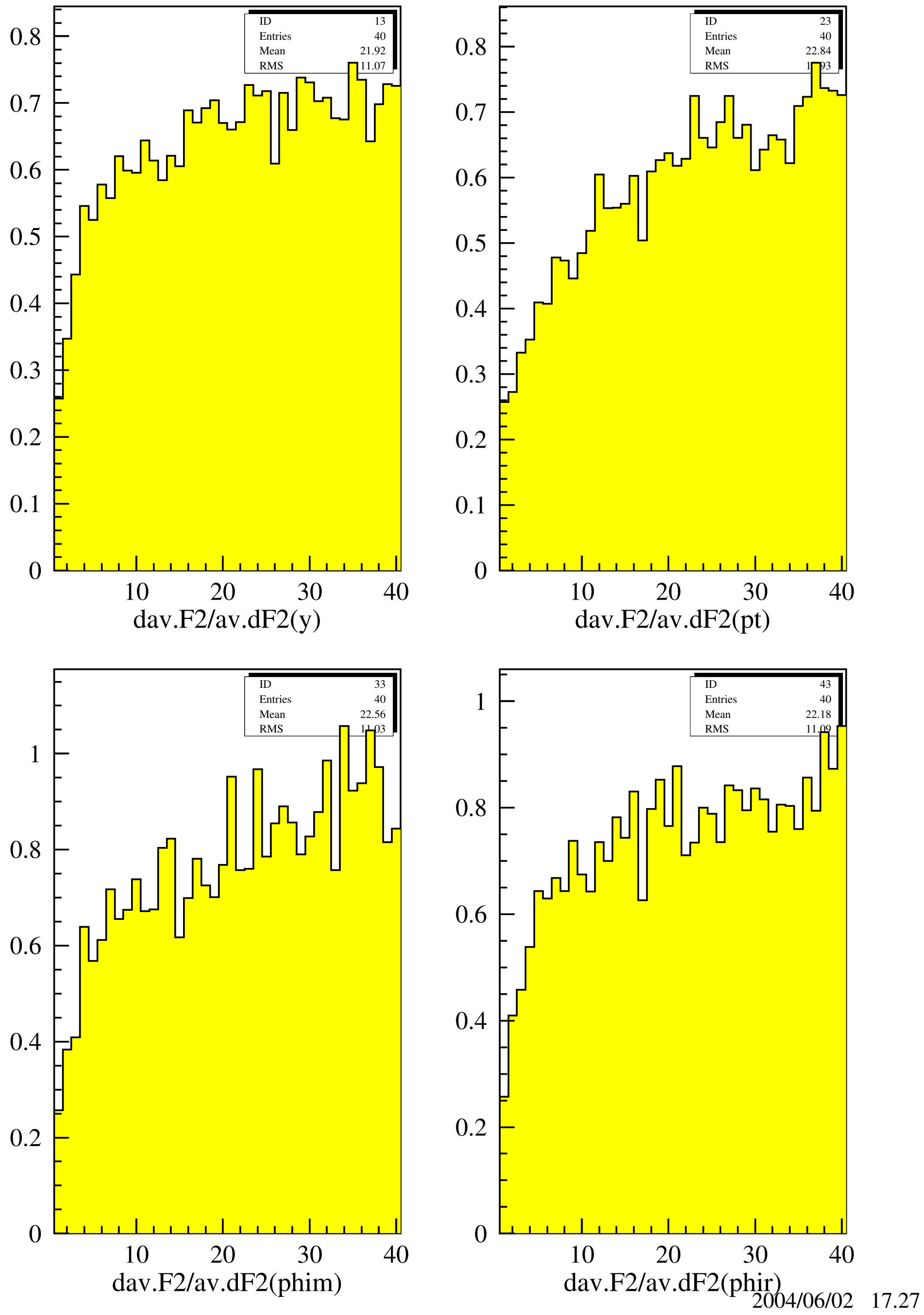


Figure 2: Ratio of the uncertainty on F_2 obtained by resampling to the average estimated uncertainty on F_2 obtained using (17) as a function of M for y, $p_{\rm t}$, $\phi_{\rm major}$ and $\phi_{\rm random}$ determined from 100 resamplings of JETSET, each sample the same size as the data sample.

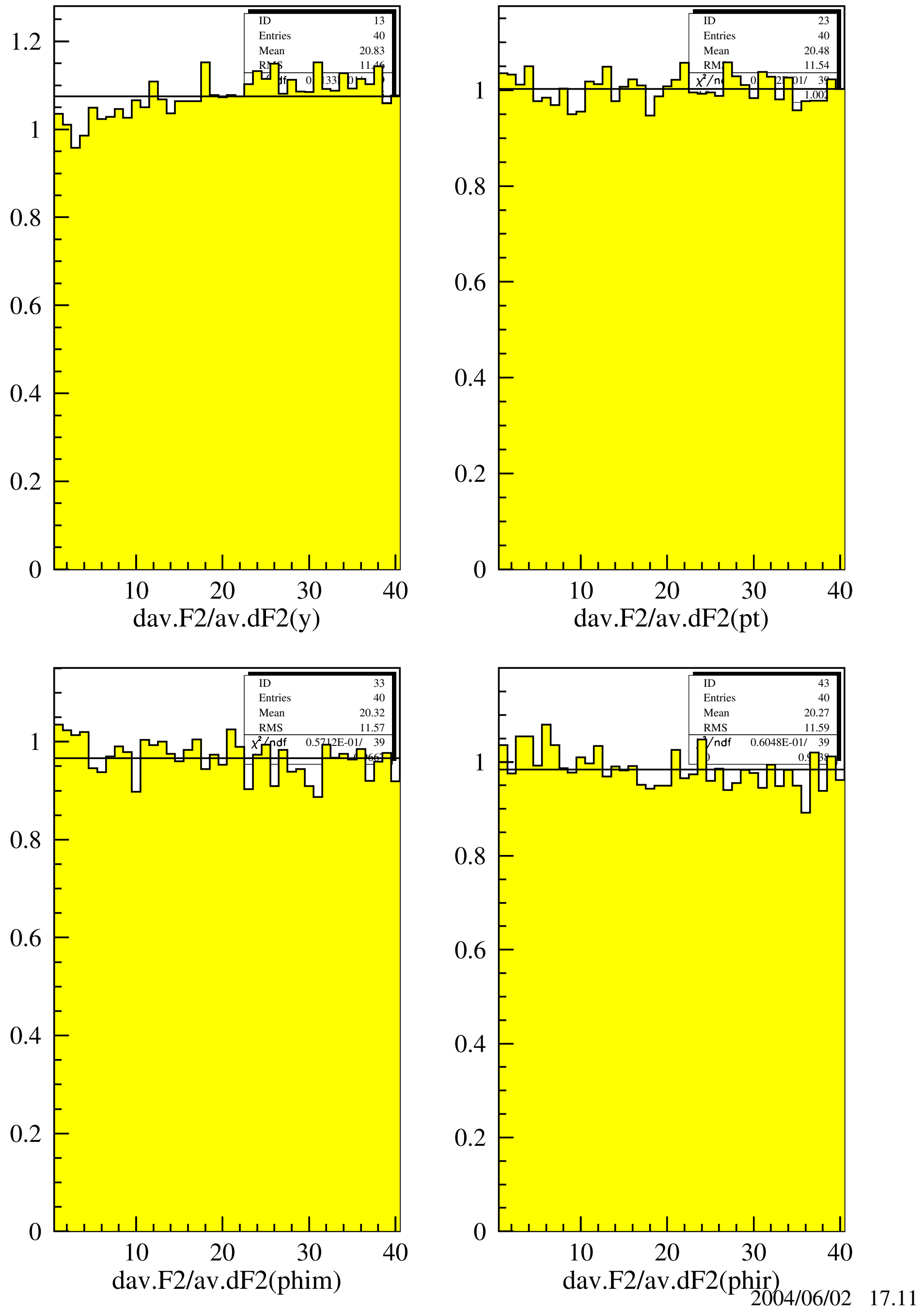


Figure 3: Ratio of the uncertainty on F_2 obtained by resampling to the average estimated uncertainty on F_2 obtained using (6) as a function of M for y, $p_{\rm t}$, $\phi_{\rm major}$ and $\phi_{\rm random}$ determined from 250 resamplings of the data.

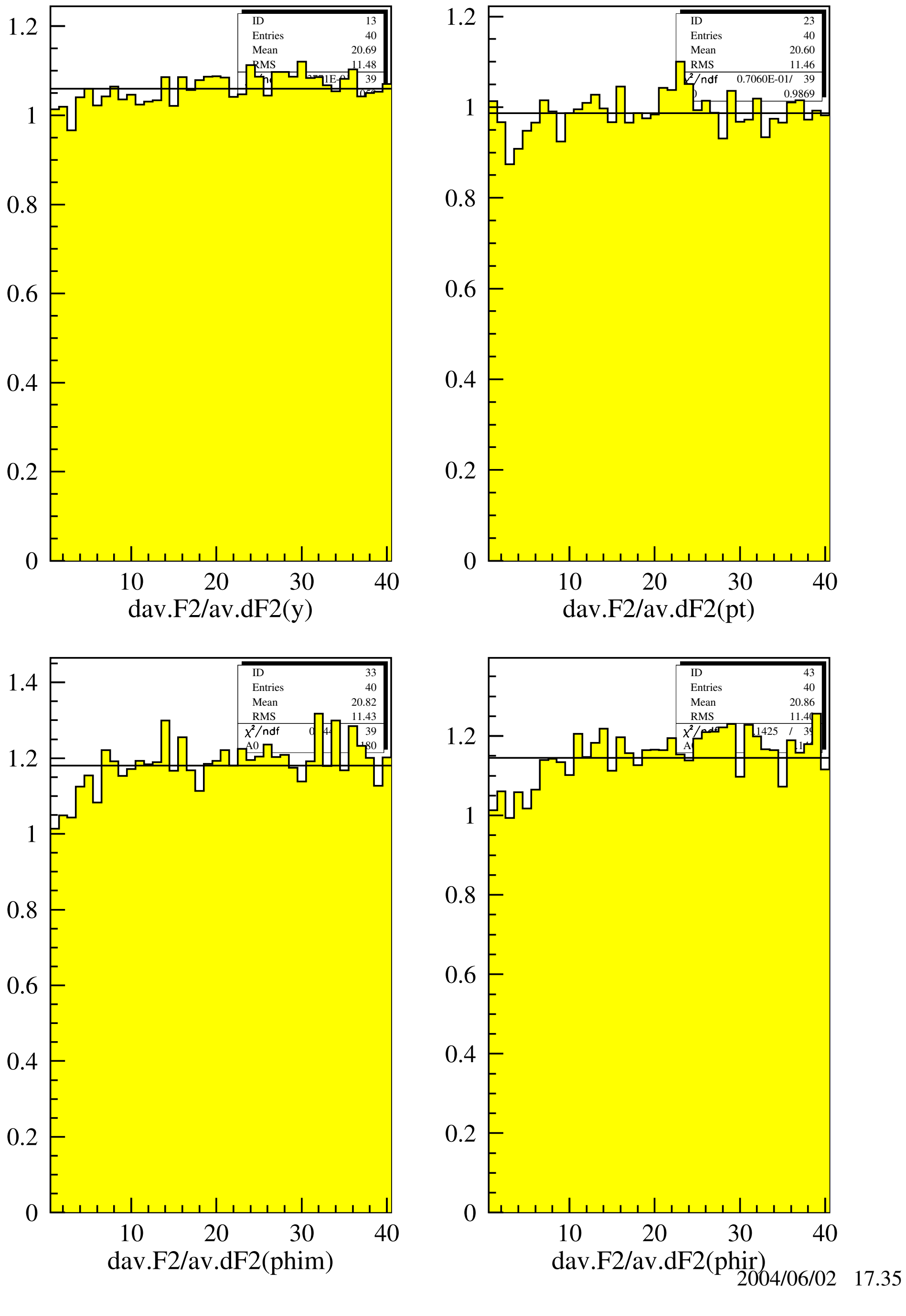


Figure 4: Ratio of the uncertainty on F_2 obtained by resampling to the average estimated uncertainty on F_2 obtained using (6) as a function of M for y, p_{t}, ϕ_{major} and ϕ_{random} determined from 250 resamplings of JETSET, each sample the same size as the data sample.

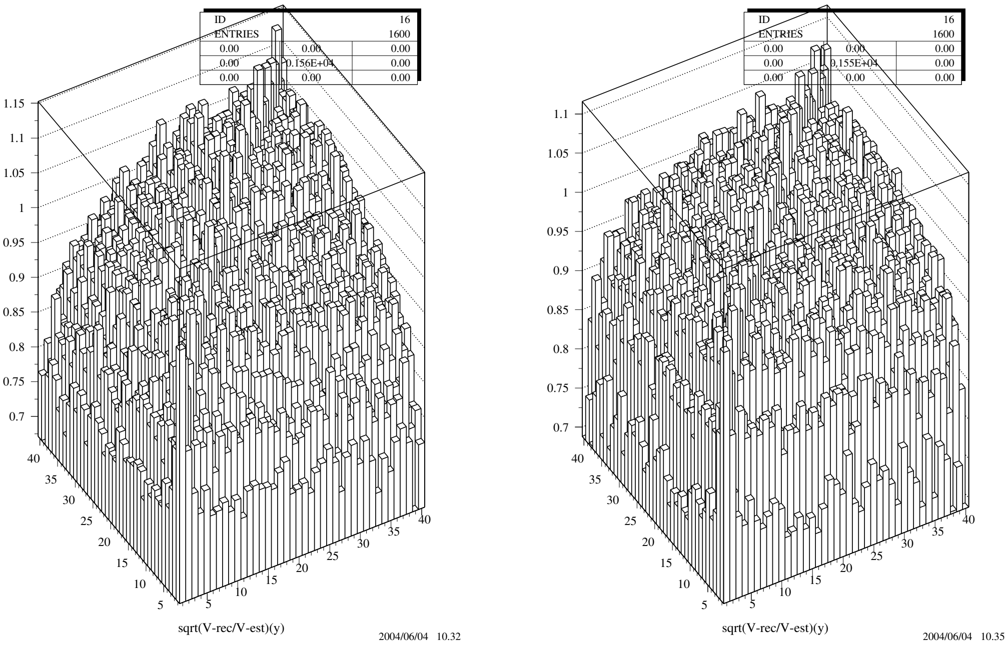


Figure 5: Ratio of the square root of the covariance of $F_2(M)$ for the rapidity, y, estimated by resampling to that estimated by (18) for (left) resampling the data and (right) resampling JETSET.

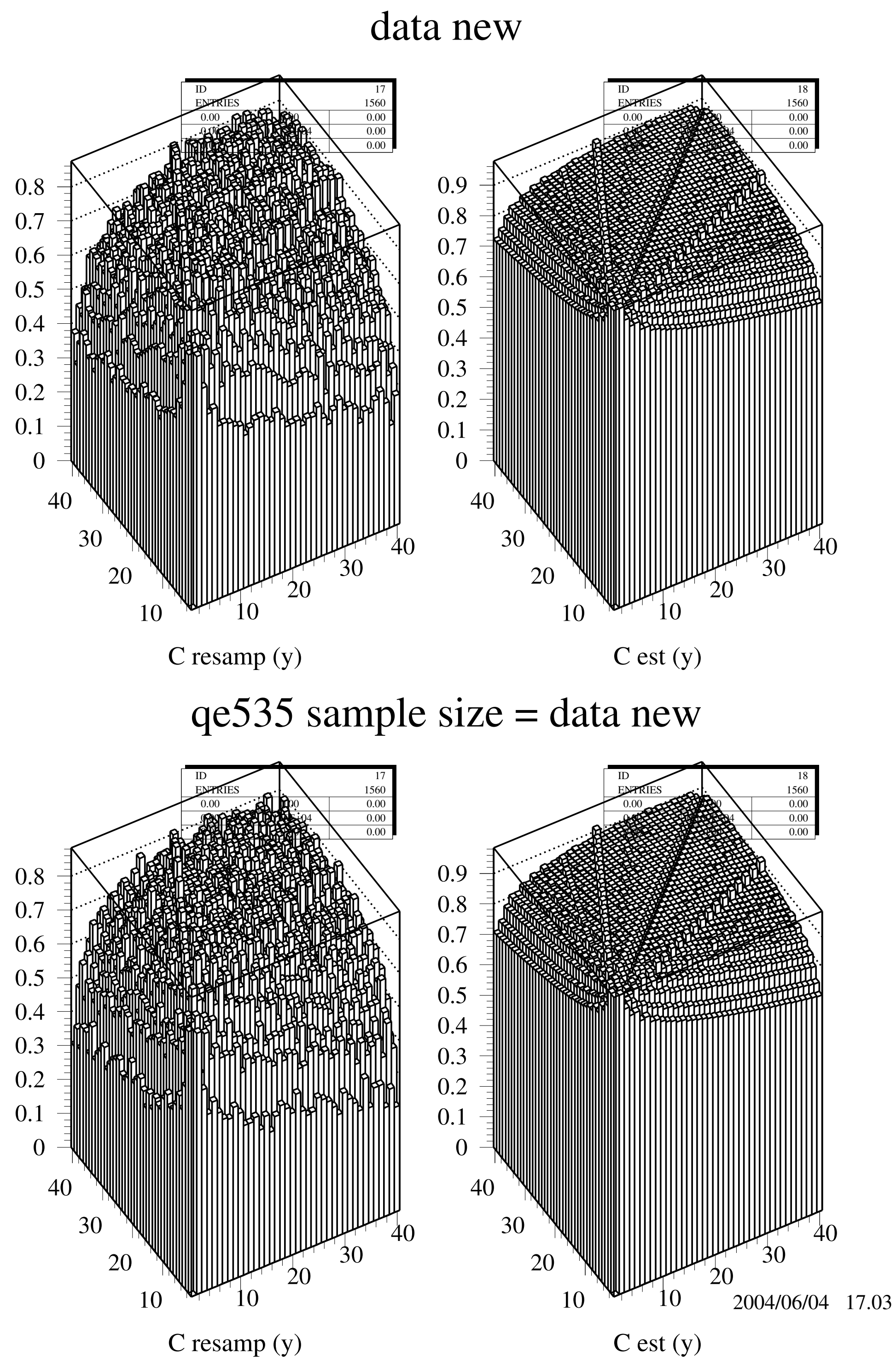


Figure 6: Left: Correlation of $F_2(M)$ for the rapidity, y, estimated (left) by resampling and (right) by (18) for (top) data and (bottom) JETSET samples of the same size as the data. The diagonal elements, which are by definition unity, are not plotted.

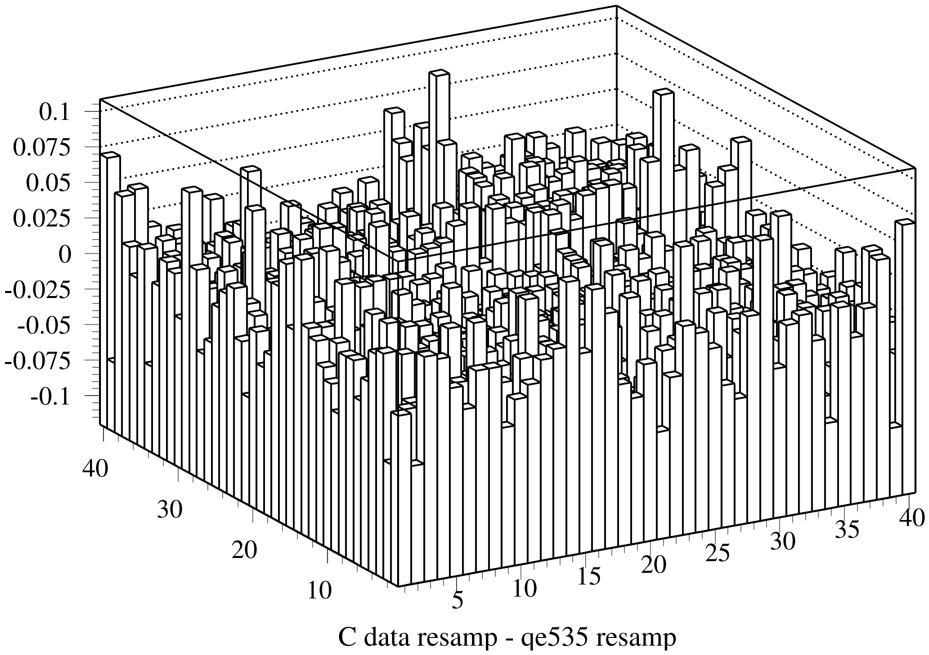


Figure 7: Difference between the correlations estimated by the resampling the data and by resampling JETSET.