CDF/ANAL/JET/CDFR/3419
Version 1.0
November 20, 1995

Assessing the Significance of a Deviation in the Tail of a Distribution

Luc Demortier

The Rockefeller University

Abstract

Several standard statistical tests can be used to quantify the significance of a deviation in the tail of a measured distribution. We study the bias introduced in these tests due to binning of the data, variation of the range over which the data are tested, and systematic uncertainties. Monte Carlo methods to compute correct significance levels are described. The results are applied to a comparison of the Run 1A inclusive jet cross section $d\sigma/dE_T$ with NLO QCD calculations.

1 Introduction

Recently a number of techniques have been proposed to quantify the significance of the discrepancy at high E_T between the measured inclusive jet cross section and the theoretical prediction [1, 2]. A simple χ^2-test is not sensitive to the fact that the high-E_T bins all deviate *in the same direction* from the prediction. On the other hand, there exist empirical distribution function tests, such as the Kolmogorov-Smirnov and the Smirnov-Cramér-von Mises tests, which do have this kind of sensitivity. Like the χ^2 test, their application involves the measurement of some deviation statistic S between data and theory, and the subsequent calculation of a significance level describing the probability to observe a deviation at least as large as S under the hypothesis that data and theory have the same parent distribution. This calculation can be done with the help of standard subroutines or tables which were originally derived under several assumptions:

1. the tests are performed on unbinned data;
2. the range of the tested distribution is not adjusted to optimize the resulting significance level;

3. there are no systematic uncertainties;

4. the theoretical model does not depend on parameters which are extracted from the data.

The relevance of the first three assumptions is to some degree overlooked in reference [2]. It is the purpose of this note to study the effect of this oversight and to propose a method to compute correct significance levels. The fourth assumption has already been discussed by Hovhannes Keutelian [3] for the Kolmogorov-Smirnov test, and will not be considered any further.

In section 2 we review the definitions of the Kolmogorov-Smirnov and Smirnov-Cramér-von Mises statistics for both one-sample and two-sample tests. We also describe the Anderson-Darling test [4, 5], which is designed to be more powerful than the first two for detecting deviations in the tail of a distribution. All three of these tests are intended to be applied to *unbinned* data. We investigate the effect of binning in section 3, and describe a method to compute correct significance levels. This method is then applied to the inclusive jet spectrum in section 4. In section 5, we study what happens to significance levels when the range of a data distribution is varied until the maximal deviation from a given model is obtained. This technique was actually employed in the inclusive jet analysis [2], but we show that, when properly applied, it offers no additional power over the standard testing method. A method to incorporate systematic uncertainties in the computation of significance levels is described and illustrated in section 6. Our conclusions are listed in section 7.

2 Definition of some Goodness-of-Fit Statistics

The goodness-of-fit statistics we are interested in are meant to be applied to the comparison of *integral* distributions, as opposed to the χ^2 statistic for example, which is applied to the comparison of distribution *densities*. Given a set of data points $\{x_i,\ i = 1,\ldots,N\}$, *sorted in ascending order*, its empirical distribution function is defined as follows:

$$S_N(x) = \begin{cases} 0 & \text{if } x < x_1 \\ i/N & \text{if } x_i \leq x < x_{i+1}, \quad i = 1,\ldots,N-1 \\ 1 & \text{if } x \geq x_N \end{cases} \tag{1}$$

2.1 One-Sample Statistics

In order to compare $S_N(x)$ to a theoretical distribution $F(x)$, we introduce three statistics: the Kolmogorov-Smirnov statistic $D_{\max}$:

$$D_{\max} \stackrel{\text{def}}{=} \sqrt{N} \sup_{-\infty < x < +\infty} |S_N(x) - F(x)| \tag{2}$$

$$= \sqrt{N} \max_{i=1,\ldots,N} \left(y_i - \frac{i-1}{N} , \frac{i}{N} - y_i \right) , \tag{3}$$

where $y_i \stackrel{\text{def}}{=} F(x_i)$, the Smirnov-Cramér-von Mises statistic W^2:

$$W^2 \stackrel{\text{def}}{=} N \int_0^1 \left(S_N(x) - F(x)\right)^2 \, dF(x) \tag{4}$$

$$= \sum_{i=1}^{N} \left(y_i - \frac{2i-1}{2N} \right)^2 + \frac{1}{12N} , \tag{5}$$

and the Anderson-Darling statistic A^2:

$$A^2 \stackrel{\text{def}}{=} N \int_0^1 \frac{\left(S_N(x) - F(x)\right)^2}{F(x)\,(1 - F(x))} \, dF(x) \tag{6}$$

$$= \sum_{i=1}^{N} \left[\left(\frac{2i-1}{N} - 2 \right) \ln(1 - y_i) - \left(\frac{2i-1}{N} \right) \ln(y_i) \right] - N . \tag{7}$$

Equalities (5) and (7) were obtained by substituting $S_N(x)$ from (1) into (4) and (6) respectively, and integrating separately over each interval over which $S_N(x)$ is constant. The statistic A^2 is identical to W^2 except for a factor $\left[F(x)\,(1 - F(x))\right]^{-1}$ in the integrand, whose purpose is to give more weight to the tails of the tested distribution. Thus one expects the Anderson-Darling statistic to be more powerful at detecting deviations in the tails.

2.2 Two-Sample Statistics

It is sometimes the case that one does not know explicitly the parent distribution function $F(x)$ of the model with which the data are to be compared. For example, the model could be a set of events obtained from a Monte Carlo simulation. One must then compare two empirical distribution functions, say $S_N(x)$ and $S_M(x)$. Let us assume that $S_N(x)$ is defined according to equation (1) from N data events $\{x_i,\ i = 1, \ldots, N\}$, and that $S_M(x)$ is similarly defined from M Monte Carlo events $\{y_i,\ i = 1, \ldots, M\}$. The two-sample Kolmogorov-Smirnov statistic is:

$$D_{\max} \stackrel{\text{def}}{=} \sqrt{\frac{N\,M}{N+M}} \sup_{-\infty < x < +\infty} \left| S_N(x) - S_M(x) \right| \tag{8}$$

The normalization factor in front of the supremum symbol ensures that this $D_{\max}$ statistic is asymptotically distributed in the same way as the one-sample statistic defined by equation (2)[1]. This way, both statistics should yield asymptotically identical tail probabilities.

[1]The Smirnov theorem guarantees that in the limit $M, N \to \infty$, with M/N constant, the two-sample and one-sample $D_{\max}$ statistics are identically distributed.

The two-sample W^2 and A^2 statistics, to be defined below, will be similarly normalized. For this note, we will not attempt to distinguish notationally between one-sample and two-sample statistics. What is intended should be clear from the context.

In order to find the two-sample statistics corresponding to W^2 and A^2, we need to find substitutes for the weight functions $dF(x)$ and $dF(x)/\left[F(x)\,(1-F(x))\right]$ in the integrals of the defining equations (4) and (6), since $F(x)$ is not known. The obvious choice is to replace $F(x)$ by $S(x)$, the empirical distribution function formed from both samples combined. Indeed, under the null-hypothesis that the x_i and y_i were drawn from the same parent population, $S(x)$ is our best estimate for $F(x)$. The two-sample Smirnov-Cramér-von Mises statistic is defined by [6]:

$$W^2 \stackrel{\text{def}}{=} \frac{N\,M}{N+M}\int_0^1 \left(S_N(x)-S_M(x)\right)^2\,dS(x) \tag{9}$$

$$= \frac{N\,M}{(N+M)^2}\left[\sum_{i=1}^{N}\left(S_N(x_i)-S_M(x_i)\right)^2+\sum_{i=1}^{M}\left(S_N(y_i)-S_M(y_i)\right)^2\right] \tag{10}$$

and the two-sample Anderson-Darling statistic by:

$$A^2 \stackrel{\text{def}}{=} \frac{N\,M}{N+M}\int_0^1 \frac{\left(S_N(x)-S_M(x)\right)^2}{S(x)\,(1-S(x))}\,dS(x) \tag{11}$$

$$= \frac{N\,M}{(N+M)^2}\left[\sum_{i=1}^{N}\frac{\left(S_N(x_i)-S_M(x_i)\right)^2}{S(x_i)\,(1-S(x_i))}+\sum_{i=1}^{M}\frac{\left(S_N(y_i)-S_M(y_i)\right)^2}{S(y_i)\,(1-S(y_i))}\right] \tag{12}$$

In this last expression, one of the sums has a term with a denominator equal to zero, since $S(x)=1$ for $x=\max\{x_i,y_i\}$. This term has a numerator equal to the square of zero however, and is therefore left out of the sum.

For computational purposes it is convenient to sort the x_i and y_i in ascending order. Then, if i is the rank of x_i in the $\{x_i\}$ sample, j the rank of y_j in the $\{y_j\}$ sample, and $R(z)$ the rank of z in the combined $\{x_i,y_j\}$ sample, we have:

$$S_N(x_i)=\frac{i}{N} \qquad S_N(y_j)=\frac{R(y_j)-j}{N} \tag{13}$$

$$S_M(x_i)=\frac{R(x_i)-i}{M} \qquad S_M(y_j)=\frac{j}{M} \tag{14}$$

$$S(x_i)=\frac{R(x_i)}{M+N} \qquad S(y_j)=\frac{R(y_j)}{M+N} \tag{15}$$

2.3 Tail Probabilities

Provided the four conditions listed in the introduction are satisfied, statistical tests based on $D_{\max}$, W^2 and A^2 are distribution-free: under the null-hypothesis, the distributions of these statistics do not depend on the form of the tested distribution $F(x)$. This can be directly seen from their definitions, since the maximum difference between $S(x)$ and $F(x)$ over the whole range of x, or the integral of this difference, is not affected by a one-to-one change in the variable x.

Once the statistics $D_{\max}$, W^2 and A^2 have been calculated, there are several ways to convert them into significance levels. We will consider two methods. The first one consists in computing the value of an analytically derived asymptotic approximation to the survivor function. The second method is to perform a Monte Carlo calculation.

Since the purpose of this note is to study how significance levels are affected by various modifications of the testing conditions, it will be useful to have analytical expressions for the survivor functions in order to be able to plot them as reference curves. The survivor function of the Kolmogorov-Smirnov statistic is given by:

$$\mathbf{S_{KS}}(\lambda) \stackrel{\text{def}}{=} \underset{N\to\infty}{\text{Prob}}\ (D_{\max} \geq \lambda) \;=\; -2\sum_{k=1}^{\infty}(-1)^k \exp(\,-2\,k^2\,\lambda^2\,) \tag{16}$$

where the notation $N \to \infty$ indicates that the formula is valid in the limit of large sample size N. In practice this means $N \geq 80$ for the Kolmogorov-Smirnov statistic. For $N < 80$, tables must be consulted [7]. $\mathbf{S_{KS}}(\lambda)$ is available in the CERN library as routine PROBKL (package code G102).

The survivor function of the Smirnov-Cramér-von Mises statistic is [4]:

$$\mathbf{S_{SCvM}}(\lambda) \stackrel{\text{def}}{=} \underset{N\to\infty}{\text{Prob}}\ (W^2 \geq \lambda) \tag{17}$$

$$= 1 - \frac{1}{\pi\sqrt{\lambda}}\sum_{k=0}^{\infty}\frac{\Gamma(k+\frac{1}{2})}{\sqrt{\pi}\,k!}\sqrt{4k+1}\,\exp\!\left(-\frac{(4k+1)^2}{16\,\lambda}\right) K_{\frac{1}{4}}\!\left(\frac{(4k+1)^2}{16\,\lambda}\right) \tag{18}$$

where $K_{\frac{1}{4}}$ is the modified Bessel function of order $1/4$. The function $\mathbf{S_{SCvM}}(\lambda)$ is not available in the CERN library, but $K_{\frac{1}{4}}$ is (routine BSIR4, package code C327). As the above series converges very rapidly, it is straightforward to translate it into computer code.

For the Anderson-Darling statistic, we have:

$$\mathbf{S_{AD}}(\lambda) \stackrel{\text{def}}{=} \underset{N\to\infty}{\text{Prob}}\ (A^2 \geq \lambda) \tag{19}$$

$$= 1 - \frac{\sqrt{8}}{\pi\sqrt{\lambda}}\sum_{k=0}^{\infty}\frac{(-1)^k\,\Gamma(k+\frac{1}{2})}{k!}\exp\!\left(-\frac{(4k+1)^2\,\pi^2}{8\,\lambda}\right) I_k(\lambda) \tag{20}$$

where:

$$I_k(\lambda) \stackrel{\text{def}}{=} \int_0^{\infty}\exp\!\left(\frac{\lambda/8}{4\,\lambda\,y^2/\,((4k+1)^2\,\pi^2)\,+\,1} - \frac{y^2}{2}\right) dy \tag{21}$$

Although $\mathbf{S_{AD}}(\lambda)$ is not available in the CERN library, it too is relatively easy to implement as a subroutine: its series converges rapidly, and the integral I_k can be performed numerically, using, for example, an open Romberg quadrature algorithm [8].

The advantage of implementing survivor functions in a computer program is that it allows one to calculate tail probabilities outside the range of available tables. However, there are many situations where the asymptotic approximation underlying these formulae is not valid, for example when the sample size is small, or when the data are binned (see section 3), or when the model depends on parameters which are extracted from the data. In all these cases, the simplest way to compute tail probabilities is via Monte Carlo algorithms. We illustrate this idea with a procedure to calculate the tail probability of a one-sample statistic:

Monte Carlo procedure 1

(1) Generate N random numbers x_i according to the theoretical distribution $F(x)$, and use them to form an "empirical" distribution function $S_N(x)$.

(2) Calculate the one-sample deviation statistic $D_{\max}$, W^2 or A^2 between $F(x)$ and $S_N(x)$.

(3) Repeat (1) and (2) a large number of times, and calculate the fraction of times that the deviation statistic is larger than the measured one. This is the desired significance level of the measurement.

Monte Carlo procedures sometimes have a disadvantage, in that large numbers of samples need to be generated in order to quantify highly significant deviations. This can become very time-consuming. It may be possible to shorten these computations however, using techniques such as importance sampling.

Figure 1 shows that calculations based on survivor functions or Monte Carlo algorithms yield identical results in their common domain of applicability. For these plots we ran 100,000 Monte Carlo experiments, for each of which we generated 1,000 random data x_i distributed according to the density:

$$\frac{dF}{dx} = \frac{125}{x^2}, \quad x \in [100, 500]. \tag{22}$$

Plots (a), (b) and (c) compare the result of Monte Carlo procedure 1 (data points) with the survivor functions (solid lines) for the one-sample Kolmogorov-Smirnov, Smirnov-Cramér-von Mises, and Anderson-Darling statistics respectively. The agreement is excellent, as expected.

For two-sample tests, the form of the distribution $F(x)$ is not necessarily known, and therefore step *(1)* in the above Monte Carlo procedure needs to be modified. As in section 2.2, let us assume that we are to compare a sample of N events with a sample of M events. The idea is to use the combined sample of $N + M$ events as a "bootstrap" estimator of $F(x)$:

Monte Carlo procedure 2

(1) Draw N events with replacement from the combined sample of $N + M$ initial events, and form their empirical distribution function $S_N(x)$.

(2) Draw M events with replacement from the combined sample of $N + M$ initial events, and form their empirical distribution function $S_M(x)$.

(3) Calculate the two-sample deviation statistic $D_{\max}$, W^2 or A^2 between $S_N(x)$ and $S_M(x)$.

(4) Repeat (1), (2) and (3) a large number of times, and calculate the fraction of times that the deviation statistic is larger than the measured one. This is the desired significance level of the measurement.

Figure 2 checks the equivalence between the above procedure (data points) and the survivor functions (solid lines), for two-sample statistics. The distributions of the two-sample $D_{\max}$, W^2 and A^2 are plotted for 100,000 trials. For plots (a), (b) and (c), the two initial samples contained 1,000 events each, drawn according to the density (22). There is good agreement for the case of W^2 and A^2, but a slight shift between the two $D_{\max}$ distributions. As expected (since the advertised equivalence is only asymptotic), this shift is reduced by increasing the size of the initial samples from 1,000 to 4,000 events each (plot d).

3 Effect of Binning the Data Sample

In the case of a complex measurement such as that of the inclusive jet cross section $d\sigma/dE_T$, one needs to unfold detector effects from the measured data, or fold in those effects in the theoretical distribution, before a meaningful comparison between data and theory can be attempted. The folding or unfolding procedure requires that the data be binned, and the goodness-of-fit tests described in the previous section must be modified in order to be applicable to binned data. Let us assume that the N data points x_i are histogrammed into B bins with contents d_j, $j = 1, \ldots, B$. The empirical distribution is now given by:

$$S_k = \frac{1}{N} \sum_{j=1}^{k} d_j \, , \quad k = 1, \ldots, B \tag{23}$$

$$N = \sum_{j=1}^{B} d_j \tag{24}$$

and is to be compared with:

$$F_k = \frac{1}{T} \sum_{j=1}^{k} t_j \, , \quad k = 1, \ldots, B \tag{25}$$

$$T = \sum_{j=1}^{B} t_j \tag{26}$$

$$t_j = F(u_j) - F(l_j) \tag{27}$$

where u_k (l_k) is the upper (lower) boundary of bin k, and $F(x)$ is the theoretical distribution function.

3.1 One-Sample Statistics for Binned Data

It is straightforward to translate the definitions (2), (4) and (6) in terms of S_k and F_k:

$$
\begin{aligned}
D_{\max(\mathrm{b})} &= \sqrt{N}\max_{k=1,\ldots,B}|S_k - F_k| && (28)\\
W^2_{(\mathrm{b})} &= N\sum_{j=1}^{B-1}(S_j - F_j)^2\,\frac{t_j}{T} && (29)\\
A^2_{(\mathrm{b})} &= N\sum_{j=1}^{B-1}\frac{(S_j - F_j)^2}{F_j\,(1-F_j)}\,\frac{t_j}{T} && (30)
\end{aligned}
$$

where the subscript (b) refers to the fact that we have binned the data before calculating these statistics. Note that the sums only need to go up to $B-1$ since by definition $S_B = F_B = 1$.

3.2 Two-Sample Statistics for Binned Data

Let $\{x_i,\ i = 1,\ldots,N\}$ and $\{y_i,\ i = 1,\ldots,M\}$ be two data samples, both histogrammed into B bins with contents d_j and d'_j respectively ($j = 1,\ldots,B$). Define the corresponding empirical distribution functions S_k and S'_k according to equation (23). Set $d''_j = d_j + d'_j$, $D'' = \sum_{j=1}^{B} d''_j$, and let S''_k be the empirical distribution function for the d''_j. The deviation statistics are then given by:

$$
\begin{aligned}
D_{\max(b)} &= \sqrt{\frac{N\,M}{N+M}}\max_{k=1,\ldots,B}|S_k - S'_k| && (31)\\
W^2_{(\mathrm{b})} &= \frac{N\,M}{N+M}\sum_{j=1}^{B-1}\left(S_j - S'_j\right)^2\,\frac{d''_j}{D''} && (32)\\
A^2_{(\mathrm{b})} &= \frac{N\,M}{N+M}\sum_{j=1}^{B-1}\frac{\left(S_j - S'_j\right)^2}{S''_j\,(1-S''_j)}\,\frac{d''_j}{D''} && (33)
\end{aligned}
$$

3.3 Tail Probabilities for Binned Data

In order to be able to convert the statistics $D_{\max(\mathrm{b})}$, $W^2_{(b)}$ and $A^2_{(b)}$ into significance levels, we need to calculate their distributions under the null-hypothesis. It can no longer be assumed that these distributions are given by the survivor functions of section 2.3, except in the limit of very fine binning. The correct distributions are most easily estimated with the Monte Carlo method, which we illustrate here for the case of one-sample statistics:

Monte Carlo procedure 3

(1) Generate N random numbers x_i according to the theoretical distribution $F(x)$ (N is the number of events in the data sample).

(2) Histogram the x_i and form the empirical distribution function S_k.

(3) Calculate the deviation statistic $D_{\max(\mathrm{b})}$, $W^2_{(\mathrm{b})}$ or $A^2_{(\mathrm{b})}$ between F_k and S_k.

(4) Repeat (1) through (3) a large number of times, and calculate the fraction of times that the deviation statistic is larger than the measured one. This is the desired significance level of the measurement.

We have applied this procedure by generating 100,000 Monte Carlo samples of $N = 1000$ events each, distributed according to equation (22). First we histogrammed each Monte Carlo sample in 10 bins from 100 to 500. The resulting distributions of $D_{\max(\mathrm{b})}$, $W^2_{(\mathrm{b})}$, and $A^2_{(\mathrm{b})}$ are plotted as data points in Figure 3. The solid lines in these figures were calculated from the survivor functions for unbinned statistics. There is a large difference between the two calculations. For the case of the Kolmogorov-Smirnov test, binning the tested distributions tends to reduce the separation between them. Therefore, the significance of a deviation calculated from binned distributions will actually be higher than what the survivor functions of section 2.3 predict. The same is true for small deviations in the Smirnov-Cramér-von Mises and Anderson-Darling tests. For higher values of W^2 and A^2 however, the data points cross over the solid curves, and deviations become actually less significant than what could be expected according to the standard survivor functions.

We also tried a finer binning, 500 bins from 100 to 500, and plotted the result in Figure 4. As expected, the disagreement between binned and unbinned is much less pronounced in this case. In conclusion, care is needed when using the standard Kolmogorov-Smirnov, Smirnov-Cramér-von Mises, or Anderson-Darling distributions to test coarsely binned data. In some sense, this can be contrasted with the χ^2 test, which requires many events per bin in order to be reliable.

Monte Carlo procedure 3 is rather impractical in the case of the inclusive jet analysis. The inclusive jet spectrum contains hundreds of thousands of events, and the generation of such a large number of random numbers for each Monte Carlo sample would require far too much processing time. In addition, it would be quite difficult to efficiently generate random numbers distributed according to the inclusive jet spectrum, because of the complexity of the analytical expression which describes this spectrum. Fortunately there is a trivially simple way to overcome these difficulties:

Monte Carlo procedure 4

(1) For each bin t_i of the theoretical distribution, generate a Poisson fluctuation $\tilde{t}_i$ with mean Nt_i/T, where N is the number of data events and $T = \sum_i t_i$. Call S_k the empirical distribution formed from the $\tilde{t}_i$.

(2) Calculate the deviation statistic $D_{\max(\mathrm{b})}$, $W^2_{(\mathrm{b})}$ or $A^2_{(\mathrm{b})}$ between S_k and the theoretical distribution F_k.

(3) Repeat (1) and (2) a large number of times, and calculate the fraction of times that the deviation statistic is larger than the measured one. This is the desired significance level of the measurement.

This procedure replaces the difficult generation of N random numbers according to $F(x)$ by the much easier generation of B Poisson random numbers. The equivalence of procedures 3 and 4 is illustrated in Figure 5, for the example described above.

4 Application to the Inclusive Jet Analysis

We are now ready to look at the effect of binning on the distributions of $D_{max(b)}$, $W^2_{(b)}$ and $A^2_{(b)}$ for the inclusive jet cross section. The theoretical prediction, smeared for detector effects, and the measured data points are listed in table 1 [9]. As the smearing procedure introduces statistical uncertainties in the theoretical distribution, two-sample statistics must be used to test consistency between data and theory.

Figure 6 shows the distribution densities of several deviation statistics, as obtained by applying Monte Carlo procedure 4 (slightly modified to handle two-sample statistics) to the theoretical prediction for the inclusive jet cross section. The densities obtained by differentiating the standard survivor functions are also shown for comparison. As more bins are included in the calculation of a statistic, the distribution of that statistic becomes more like that of the corresponding asymptotic survivor function.

A quantitative comparison between data and theory is provided in table 2. The significance levels are expressed as numbers of standard deviations for a Gaussian distribution, which are somewhat easier to comprehend than probabilities, especially when the latter are very small. The correspondence between probabilities P and numbers of standard deviations r is given by:

$$P \stackrel{\text{def}}{=} \mathrm{Prob}(|X| \geq r) \;=\; \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{r} e^{-\frac{t^2}{2}}\, dt \tag{34}$$

where X is a normal variate. The table also provides the results of a standard χ^2 test of goodness of fit. For two histograms with bin contents $\{d_i,\ i = 1, \ldots, B\}$ and $\{d'_i,\ i = 1, \ldots, B\}$, summing up to D and D' respectively, the variable:

$$X^2 \stackrel{\text{def}}{=} \sum_{i=1}^{B} \frac{\left(d_i \sqrt{D'/D} \;-\; d'_i \sqrt{D/D'}\right)^2}{d_i \;+\; d'_i} \tag{35}$$

is approximately distributed as a χ^2 variable with $B - 1$ degrees of freedom.

Bin	Jet E_T (GeV)	Number of events: Theory	Data
1	14.6	1841	1691
2	20.4	423	367
3	26.9	110969	99263
4	33.3	43435	41677
5	39.5	19176	19139
6	45.5	9677	9505
7	51.3	5165	5182
8	57.0	2933	2844
9	62.7	1757	1746
10	68.3	1088	1105
11	73.9	689	662
12	79.5	456	420
13	85.0	6350	6165
14	90.5	4405	4362
15	95.9	3076	2979
16	101.4	2216	2224
17	106.8	6276	6203
18	112.2	4584	4499
19	117.6	3379	3409
20	123.0	2525	2522
21	128.4	1899	1876

Bin	Jet E_T (GeV)	Number of events: Theory	Data
22	133.8	8738	8709
23	139.2	6759	6782
24	144.5	5227	5127
25	149.9	4071	4018
26	155.3	3156	3212
27	160.6	2472	2520
28	168.4	3534	3575
29	179.2	2245	2242
30	189.9	1457	1498
31	200.7	972	1044
32	211.5	638	695
33	224.7	586	683
34	241.0	327	344
35	257.4	186	201
36	273.8	107	129
37	292.5	77	98
38	314.4	38	52
39	336.2	19	24
40	365.5	14	21
41	418.5	4	10

Table 1: Inclusive jet event rate, smeared NLO QCD theory and raw data, as a function of jet E_T.

Deviation Statistic	Bin Range	Value of Statistic	Standard S.L.	Corrected S.L.	Smeared S.L. $\sigma_{\mathrm{stab}} = 1\%$	Smeared S.L. $\sigma_{\mathrm{stab}} = 2.5\%$
χ^2/N_{dof}	5–41	38.2/36	0.90	0.90	0.65	0.53
	10–41	35.9/31	1.15	1.15	0.95	0.79
	15–41	30.9/26	1.19	1.19	1.00	0.87
	20–41	26.8/21	1.35	1.36	1.18	1.06
	25–41	21.5/16	1.41	1.41	1.27	1.17
	30–41	10.2/11	0.65	0.64	0.60	0.58
	35–41	3.65/6	0.35	0.34	0.34	0.33
$D_{\max(b)}$	5–41	1.201	1.59	1.86	1.32	0.90
	10–41	1.395	2.05	2.38	1.67	1.27
	15–41	1.356	1.96	2.34	1.74	1.40
	20–41	1.324	1.88	2.32	1.88	1.60
	25–41	1.335	1.91	2.41	2.02	1.82
	30–41	0.923	0.91	1.52	1.39	1.31
	35–41	0.659	0.28	0.99	0.97	0.96
$W^2_{(b)}$	5–41	0.616	2.32	2.25	1.59	1.08
	10–41	1.025	3.07	3.06	1.99	1.48
	15–41	0.802	2.69	2.65	1.86	1.48
	20–41	0.770	2.63	2.50	1.96	1.66
	25–41	0.915	2.89	2.77	2.24	1.99
	30–41	0.579	2.24	2.01	1.82	1.71
	35–41	0.328	1.59	1.39	1.36	1.34
$A^2_{(b)}$	5–41	4.540	2.82	2.75	1.81	1.25
	10–41	6.808	3.54	3.52	2.18	1.61
	15–41	5.633	3.19	3.15	2.10	1.65
	20–41	5.239	3.06	2.97	2.20	1.85
	25–41	5.802	3.24	3.12	2.43	2.13
	30–41	3.060	2.23	2.09	1.86	1.73
	35–41	1.667	1.47	1.39	1.36	1.33

Table 2: Significance levels obtained from a comparison between inclusive jet data and theory. Values and significance levels are given for each of four different deviation statistics and for several E_T bin ranges. The bin numbers refer to table 1. Significance levels are expressed in equivalent numbers of standard deviations for a Gaussian distribution. Standard significance levels (column 4) were obtained from asymptotic survivor functions, whereas the corrected significance levels (column 5) were computed according to Monte Carlo procedure 4 (one million trials). The last two columns incorporate the effect of all the systematic uncertainties (cfr. section 6). The uncertainty on the E_T scale stability was taken to be 1% for column 6, and 2.5% for column 7 (see [2]).

Several points can be made about the results shown in the table. Except for the χ^2 statistic, there is a clear difference between the significance levels computed from the standard survivor functions, and those obtained from the correct Monte Carlo procedure for binned data. The difference is most pronounced in the case of $D_{\max(b)}$. It is also evident that $A^2_{(b)}$ is more powerful at detecting deviations than $W^2_{(b)}$, which is itself more powerful than $D_{\max(b)}$. The χ^2 statistic is the weakest of all four. Finally, the deviation detected by these statistics is indeed in the tail of the jet E_T distribution, since the significance levels do not change much as the first bin in the tested bin range moves from bin 5 to bin 25. Beyond bin 30, the significances start to decrease because by then the high-statistics central region of the jet E_T distribution is no longer available to "calibrate" the comparison of data with theory.

Another way to demonstrate that the observed deviation comes from high-E_T jets is to compare theory and data in the central region of the spectrum, leaving out the tail. This is done in table 3, where each of the four deviation tests is applied to the E_T region between bins 5 and 25. In all cases, data and theory are within 0.2 standard deviations of each other.

Deviation Statistic	Bin Range	Value of Statistic	Standard S.L.	Corrected S.L.	Smeared S.L. $\sigma_{\rm stab} = 1\%$	Smeared S.L. $\sigma_{\rm stab} = 2.5\%$
$\chi^2/N_{\rm dof}$	5–25	7.9/20	0.0093	0.0092	0.0058	0.0050
$D_{\max(b)}$	5–25	0.358	0.00058	0.047	0.027	0.024
$W^2_{(b)}$	5–25	0.052	0.171	0.18	0.11	0.09
$A^2_{(b)}$	5–25	0.285	0.064	0.15	0.087	0.076

Table 3: Significance levels obtained from a comparison between data and theory of the central region (bins 5 through 25) of the inclusive jet E_T spectrum. These significance levels are expressed in numbers of standard deviations for a Gaussian distribution. The columns in this table have the same meaning as in table 2.

5 Effect of Varying the Range of the Tested Distribution

Since our goal is to quantify a deviation between data and theory in the *tail* of the jet E_T spectrum, it may be tempting to try the following procedure [2]:

1. Test all bin ranges of the form i–41, where i is a bin number between 1 and 40;
2. Select the range which gives the largest deviation statistic ($D_{\max(b)}$, $W^2_{(b)}$ or $A^2_{(b)}$);

3. Convert the value of the deviation statistic into a significance level.

The third step requires some care. We are no longer doing standard Kolmogorov-Smirnov, Smirnov-Cramér-von Mises or Anderson-Darling tests as described in section 2, since we are optimizing the bin range, thereby introducing another random variable in addition to the deviation statistic itself. Let us therefore rename the deviation statistics obtained by the above procedure $D^{\max}_{\max(b)}$, $W^2_{(b)\max}$ and $A^2_{(b)\max}$. Using a simple and obvious modification of Monte Carlo procedure 4, we can plot distributions of the new statistics. This is shown in Figure 7, for the case of the inclusive jet analysis. There is a large difference between these distributions and the distributions of the standard statistics. The results of testing the deviation between data and theory are listed in table 4. They indicate that this method is actually less powerful than the simpler tests investigated in the previous section.

Deviation Statistic	Value	Bin range for which value is reached	Standard S.L.	Corrected S.L.
$D^{\max}_{\max(b)}$	1.448	28–41	2.17	1.54
$W^2_{(b)\max}$	1.236	24–41	3.39	1.99
$A^2_{(b)\max}$	7.438	24–41	3.71	2.47

Table 4: Significance levels obtained from a comparison between inclusive jet data and theory. The deviation statistics are defined in the text. The significance levels are expressed in numbers of standard deviations for a Gaussian distribution. Standard significance levels were computed from the survivor functions for $D_{\max}$, W^2 and A^2, whereas the corrected significance levels were obtained from a Monte Carlo procedure.

6 Effect of Systematic Uncertainties

So far we have not incorporated the effect of systematic uncertainties in the evaluation of significance levels. This is fairly straightforward to do when the χ^2 statistic is used; see for example [10]. For the case of the $D_{\max}$, W^2 and A^2 statistics, binned or unbinned, a different procedure must be adopted.

For binned data, the following method has been suggested in [2]. First, one adjusts the systematics to obtain the best possible agreement between data and theoretical model. This adjustment is driven by a least-squares algorithm which incorporates the effect of systematic uncertainties. It allows for some limited variation in shape of the fitted distribution. Any remaining shape difference between data and theory is then subsequently picked up by a shape-sensitive test based on $D_{\max}$, W^2 or A^2.

While this method will certainly yield a significance which is diluted by systematic effects, it is not at all clear that the least-squares algorithm will converge to the same minimum as an algorithm whose goodness-of-fit criterion is $D_{\max}$, W^2 or A^2. Therefore, it is also not clear whether the final significance properly takes into account the full range of systematic effects.

We propose here a method which avoids the above difficulty by calculating significance levels from *smeared* distributions of the deviation statistics. The smearing procedure samples the whole range of systematics with the appropriate weighting function. Let us assume we are to compare a data histogram $\{d_i\}$ with a model histogram $\{m_i\}$, and that both model and data are subject to statistical fluctuations:

Monte Carlo procedure 5

(1) For each bin of the model distribution, generate a Poisson fluctuation $\tilde{m}_i$ with mean m_i. Call S_k the empirical distribution formed from the $\tilde{m}_i$.

(2) Create a "systematic" fluctuation $\{m'_i\}$ of the model histogram. For example, if there is only one systematic uncertainty, which is Gaussian and fully correlated across bins, one would generate a single normal random number X, and shift each bin m_i by the amount $X \cdot \sigma_i$, where σ_i is the absolute systematic uncertainty on the contents m_i. If there are several systematic uncertainties, several such shifts will have to be done.

(3) For each bin of the histogram obtained from step (2), generate a Poisson fluctuation $\tilde{m}'_i$ with mean m'_i. Let S'_k be the empirical distribution function associated with the $\tilde{m}'_i$.

(4) Calculate the deviation statistic $D_{\max(\mathrm{b})}$, $W^2_{(\mathrm{b})}$ or $A^2_{(\mathrm{b})}$ between S_k and S'_k.

(5) Repeat steps (1), (2), (3) and (4) a large number of times, and calculate the fraction of times that the deviation statistic is larger than the measured one. This is the desired significance level of the measurement, smeared with systematic effects.

We have applied this procedure to the inclusive jet data of section 4. The inclusive jet analysis considers a total of eight systematic uncertainties. Since these uncertainties are independent and act in the same way on the jet E_T spectrum, it is sufficient to consider a single systematic uncertainty, equal to the bin-by-bin sum in quadrature of the original eight [1]. We will consider two cases. For the first case, the uncertainty on the stability of the absolute calibration of the calorimeter, σ_{stab}, is assumed to be 1%. This is considered to be a correct assumption [2], and leads to a combined systematic uncertainty varying from about 13% at low E_T to about 28% at high E_T. For the second case, the uncertainty on the stability of the calibration is set to its upper limit of 2.5%. Here, the combined systematic uncertainty varies from about 17% to 39%. The results of our calculations are given in the last two columns of tables 2 and 3.

Table 2 shows that the statistic $A^2_{(b)}$, calculated from bins 10 through 41, gives the highest significance: 3.52 standard deviations. When using Monte Carlo procedure 5 with the combined systematic uncertainty, we find smeared significances of 2.18 and 1.61 standard deviations, depending on the assumption about the stability of the calibration of the calorimeter. Although this result may seem disappointing, it should not come as a surprise. Figure 8 shows how the value of $A^2_{(b)}$, calculated from bins 10 through 41, varies as a function of the amount of systematic uncertainty added to the theoretical distribution. This amount is measured in numbers of standard deviations of systematic uncertainty. Horizontal lines in the plot indicate the significance levels corresponding to various values of $A^2_{(b)}$. For $\sigma_{\rm stab} = 1\%$ or 2.5%, shifting the theoretical distribution by about 2.6, respectively 1.1 standard deviations of systematic uncertainty brings it within less than one standard deviation of the data. This plot illustrates that, with the given systematic uncertainty, data and theory are not very far from each other, and that the *shape* of the systematic uncertainty can easily accomodate the high-E_T excess observed in the data.

7 Conclusions

We have reviewed the formalism of one-sample and two-sample tests with the Kolmogorov-Smirnov, Smirnov-Cramér-von Mises, and Anderson-Darling statistics. When the tested distribution is coarsely binned, as is the case for the inclusive jet E_T spectrum, we have shown that correct significance levels can not be obtained from standard published tables or subroutines, but can be calculated with Monte Carlo algorithms, of which we gave several examples. Applying this technique to the inclusive jet E_T spectrum, we found that the Anderson-Darling test is more powerful than the other two at detecting deviations in the tail of the spectrum. We then studied the effect of choosing the range of bins to include in a given test. When the criterion guiding this choice is the maximization of the discrepancy between two distributions, the corresponding significance levels are strongly biased, as one should expect. Finally, we proposed a Monte Carlo procedure to incorporate the effect of systematic uncertainties into the tests, and applied it to the inclusive jet E_T spectrum.

References

[1] Anwar Ahmad Bhatti, "A Few Minor Points in Inclusive Jet Cross Section", QCD Group Meeting of September 14, 1995;
Tom Devlin, "Jet Inclusive Cross Sections: Statistical Issues", QCD Group Meeting of September 14, 1995.

[2] Tom Devlin, "Jet Inclusive Cross Sections: A Statistical Study of the 'Excess' ", CDF note 3301, Version 3.0 (October 1, 1995).

[3] Hovhannes Keutelian, "The Kolmogorov-Smirnov Test when Parameters are Estimated from Data", CDF note 1285, Version 1.0 (April 30, 1991).

[4] T.W. Anderson and D.A. Darling, "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes", Ann. Math. Stat. **23**, 193-212 (1952).

[5] T.W. Anderson and D.A. Darling, "A Test of Goodness of Fit", J. Amer. Stat. Assoc. **49**, 765-769 (1954).

[6] Integration with respect to a discrete distribution function is discussed in many standard texts on probability; see for example section 4.2.2 in: A.F. Karr, "Probability", Springer-Verlag, 1993, 282pp.

[7] Byron P. Roe, "Probability and Statistics in Experimental Physics", Springer-Verlag, 1992, 208pp.

[8] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, "Numerical Recipes in FORTRAN, The Art of Scientific Computing", Second Edition, Cambridge University Press, 1992, 963 pp.

[9] Anwar Ahmad Bhatti, private communication. In all our calculations, we have left out the four bins at lowest E_T. These make a large contribution to a χ^2 goodness-of-fit test, and don't affect the study of the high-E_T excess (see [2]).

[10] T. Devlin, "Correlations from Systematic Corrections to Poisson-Distributed Data in Log-Likelihood Functions", CDF note 3126, Version 3.0 (August 19, 1995).

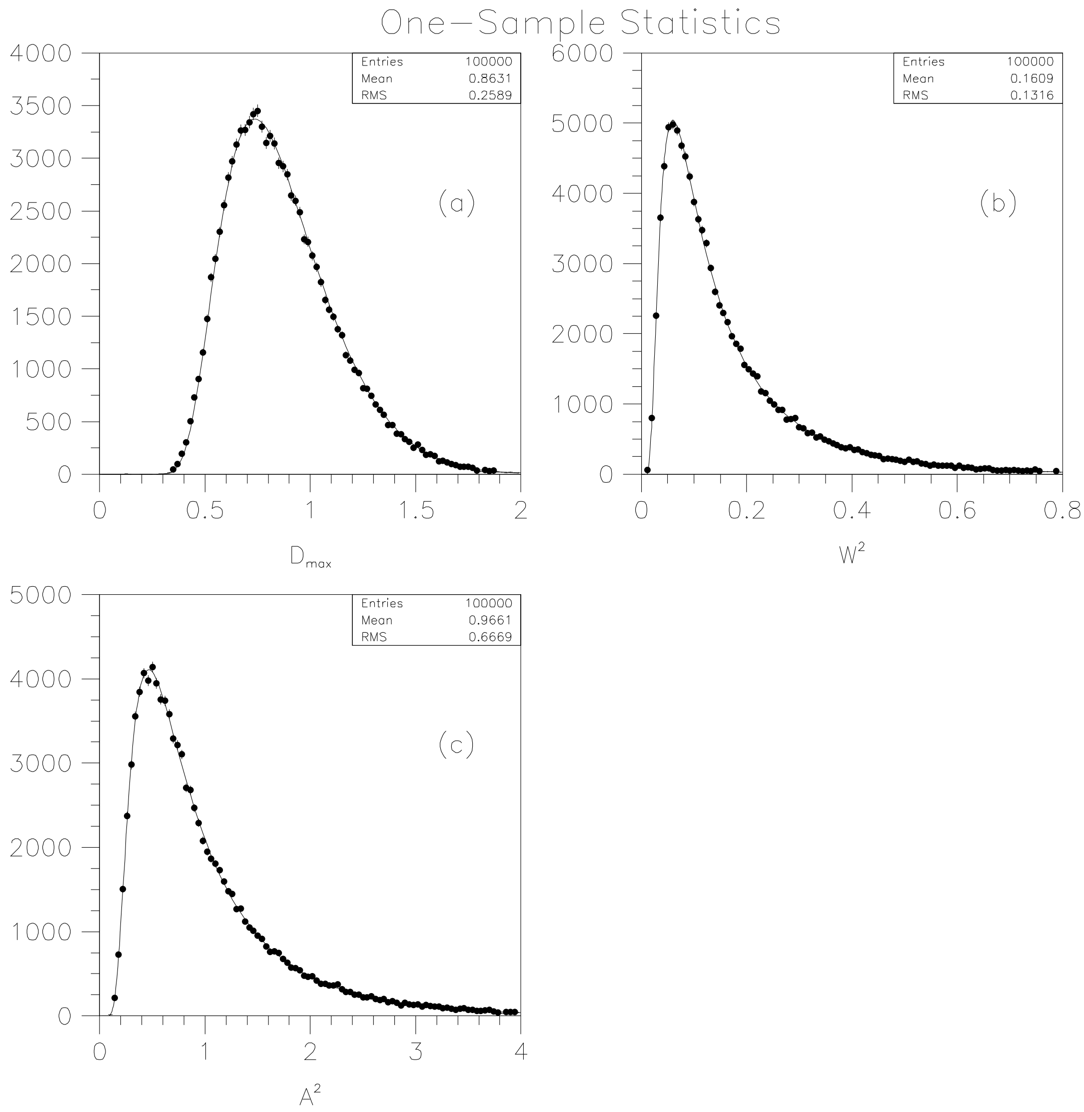

Figure 1: Distributions of (a) the Kolmogorov-Smirnov statistic, (b) the Smirnov-Cramér-von Mises statistic, and (c) the Anderson-Darling statistic for one-sample tests on unbinned data. The data points are histograms obtained by generating 100,000 trials according to Monte Carlo procedure 1. The solid lines are the result of numerically differentiating the survivor functions $\mathbf{S_{KS}}$, $\mathbf{S_{SCvM}}$ and $\mathbf{S_{AD}}$ respectively, and are normalized to the same areas as the corresponding histograms.

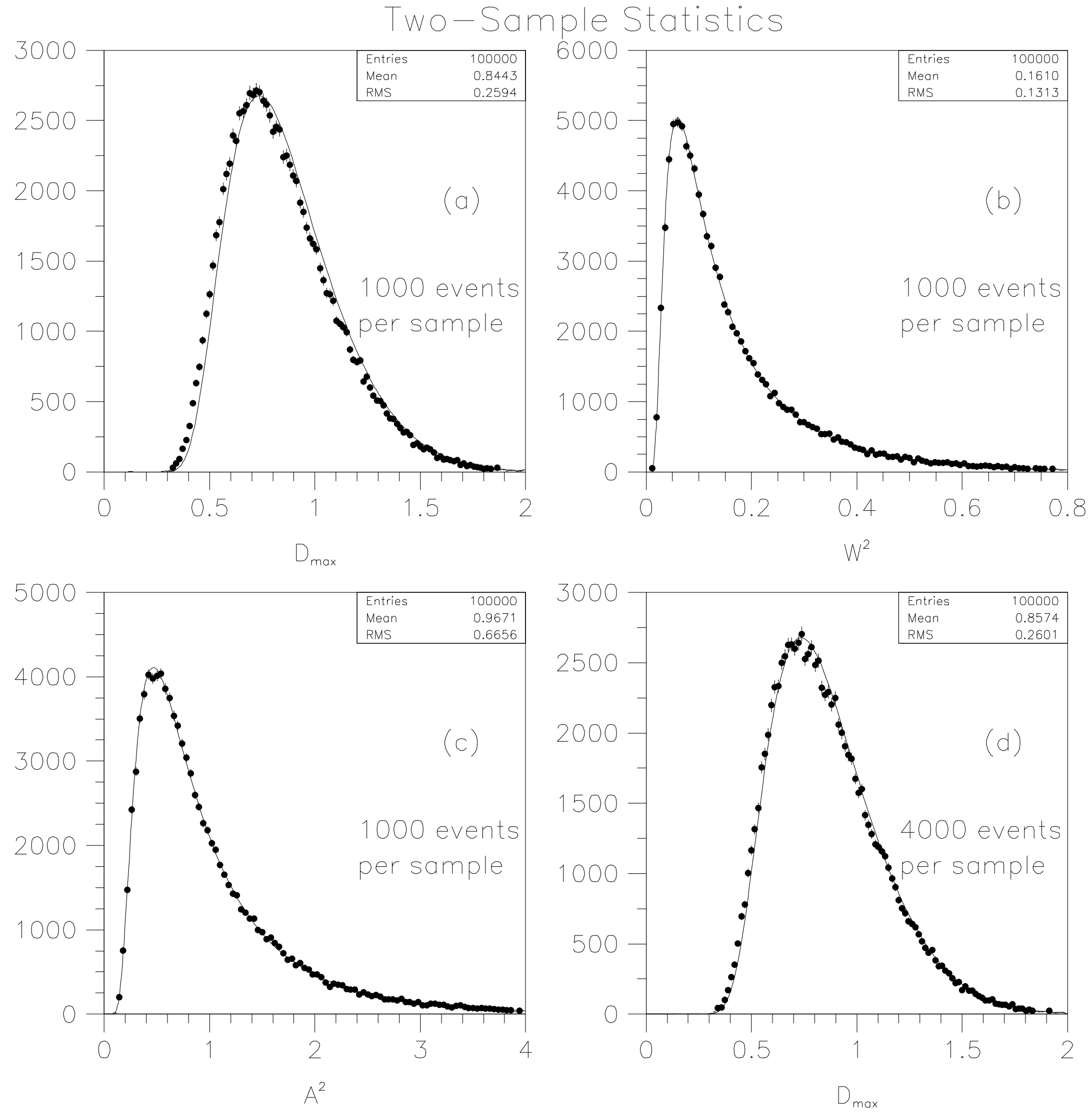

Figure 2: Distributions of the two-sample Kolmogorov-Smirnov statistic (plots a and d), Smirnov-Cramér-von Mises statistic (b), and Anderson-Darling statistic (c) for unbinned data. The data points are histograms obtained by generating 100,000 trials according to Monte Carlo procedure 2. For plots (a), (b) and (c), the initial samples contained 1,000 events each, whereas for plot (d) they contained 4,000 events each. The solid lines are the result of numerically differentiating the survivor functions $\mathbf{S_{KS}}$, $\mathbf{S_{SCvM}}$, $\mathbf{S_{AD}}$ and $\mathbf{S_{KS}}$ respectively, and are normalized to the same areas as the corresponding histograms.

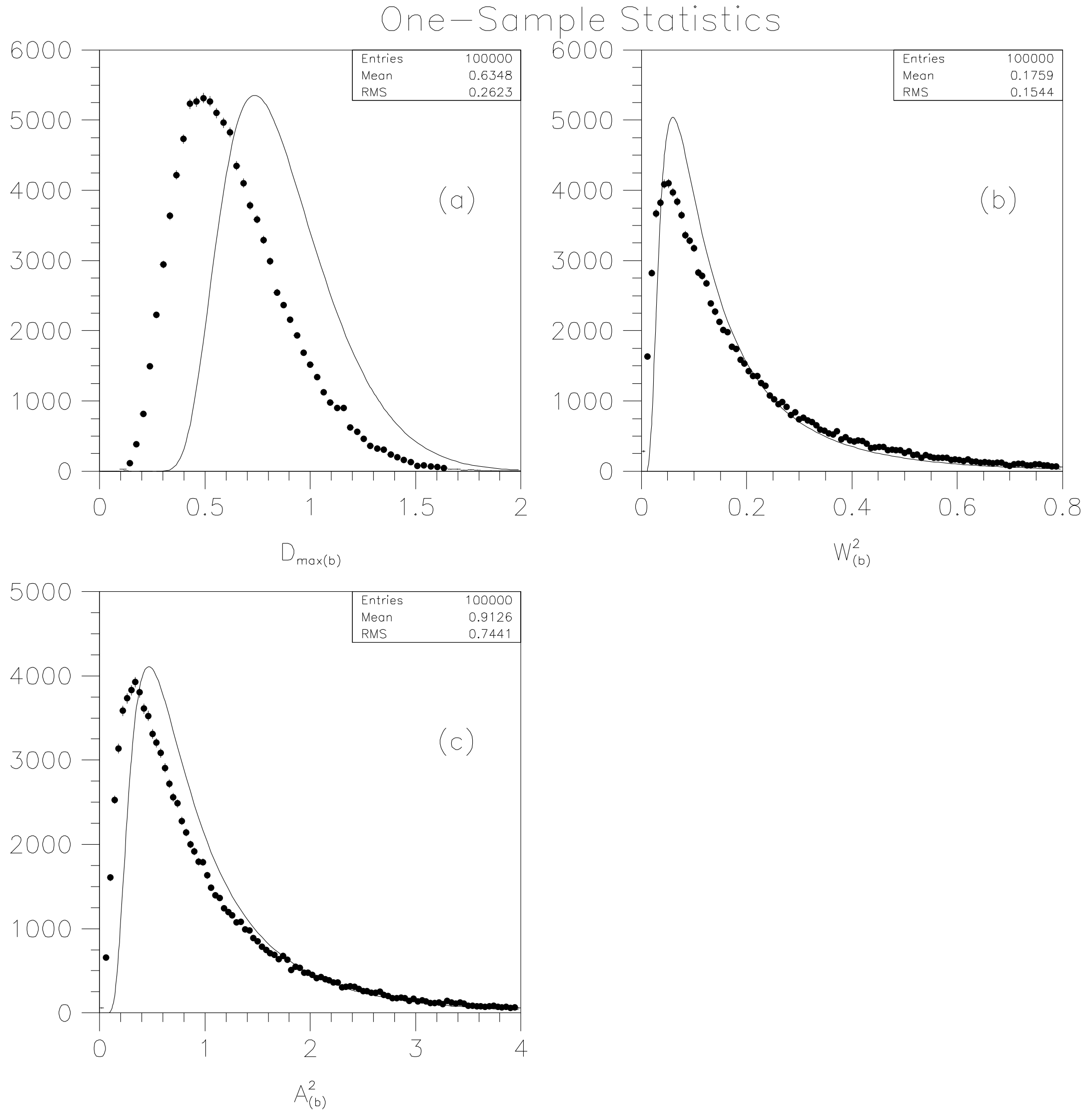

Figure 3: Distributions of (a) the Kolmogorov-Smirnov statistic, (b) the Smirnov-Cramér-von Mises statistic, and (c) the Anderson-Darling statistic for one-sample tests on coarsely binned data (see text). The data points are histograms obtained by generating 100,000 trials according to Monte Carlo procedure 3. The solid lines are the result of numerically differentiating the survivor functions $\mathbf{S_{KS}}$, $\mathbf{S_{SCvM}}$ and $\mathbf{S_{AD}}$ respectively, and are normalized to the same areas as the corresponding histograms.

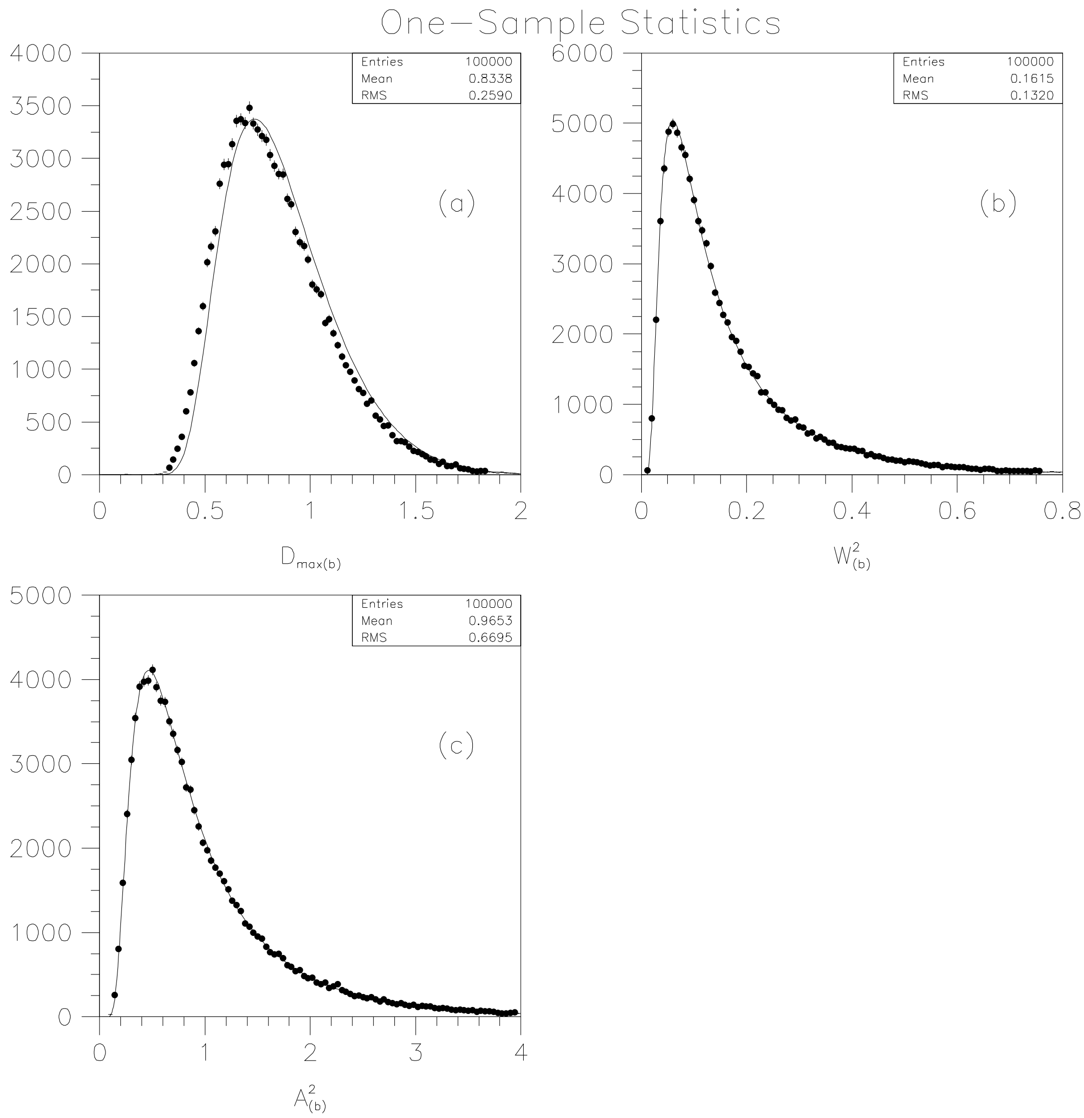

Figure 4: Distributions of (a) the Kolmogorov-Smirnov statistic, (b) the Smirnov-Cramér-von Mises statistic, and (c) the Anderson-Darling statistic for one-sample tests on finely binned data (see text). The data points are histograms obtained by generating 100,000 trials according to Monte Carlo procedure 3. The solid lines are the result of numerically differentiating the survivor functions $\mathbf{S_{KS}}$, $\mathbf{S_{SCvM}}$ and $\mathbf{S_{AD}}$ respectively, and are normalized to the same areas as the corresponding histograms.

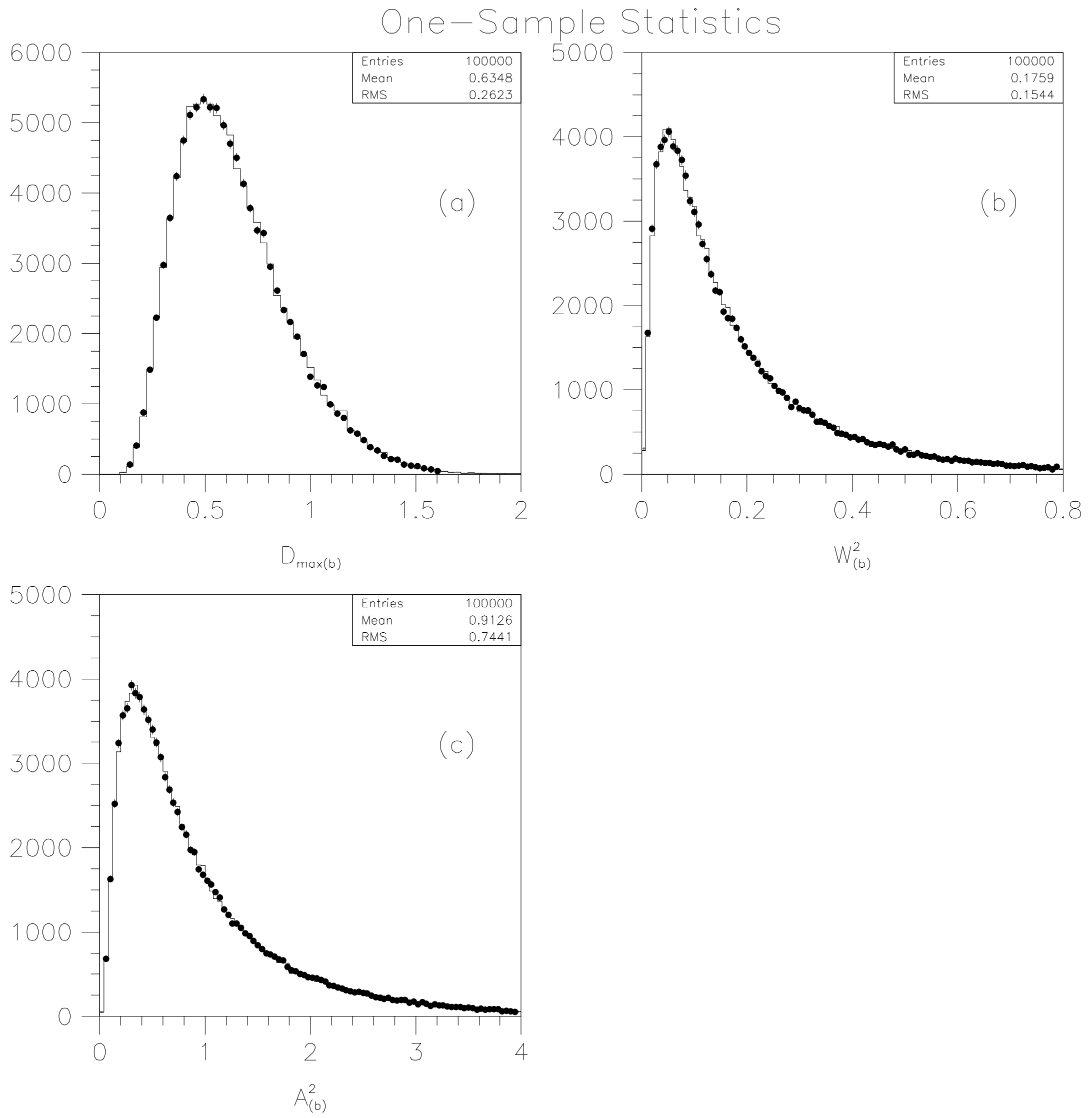

Figure 5: Distributions of the one-sample Kolmogorov-Smirnov statistic (a), Smirnov-Cramér-von Mises statistic (b), and Anderson-Darling statistic (c) for the density represented by equation (22), histogrammed in 10 bins from 100 to 500. The histograms were obtained with Monte Carlo procedure 3, the data points with procedure 4.

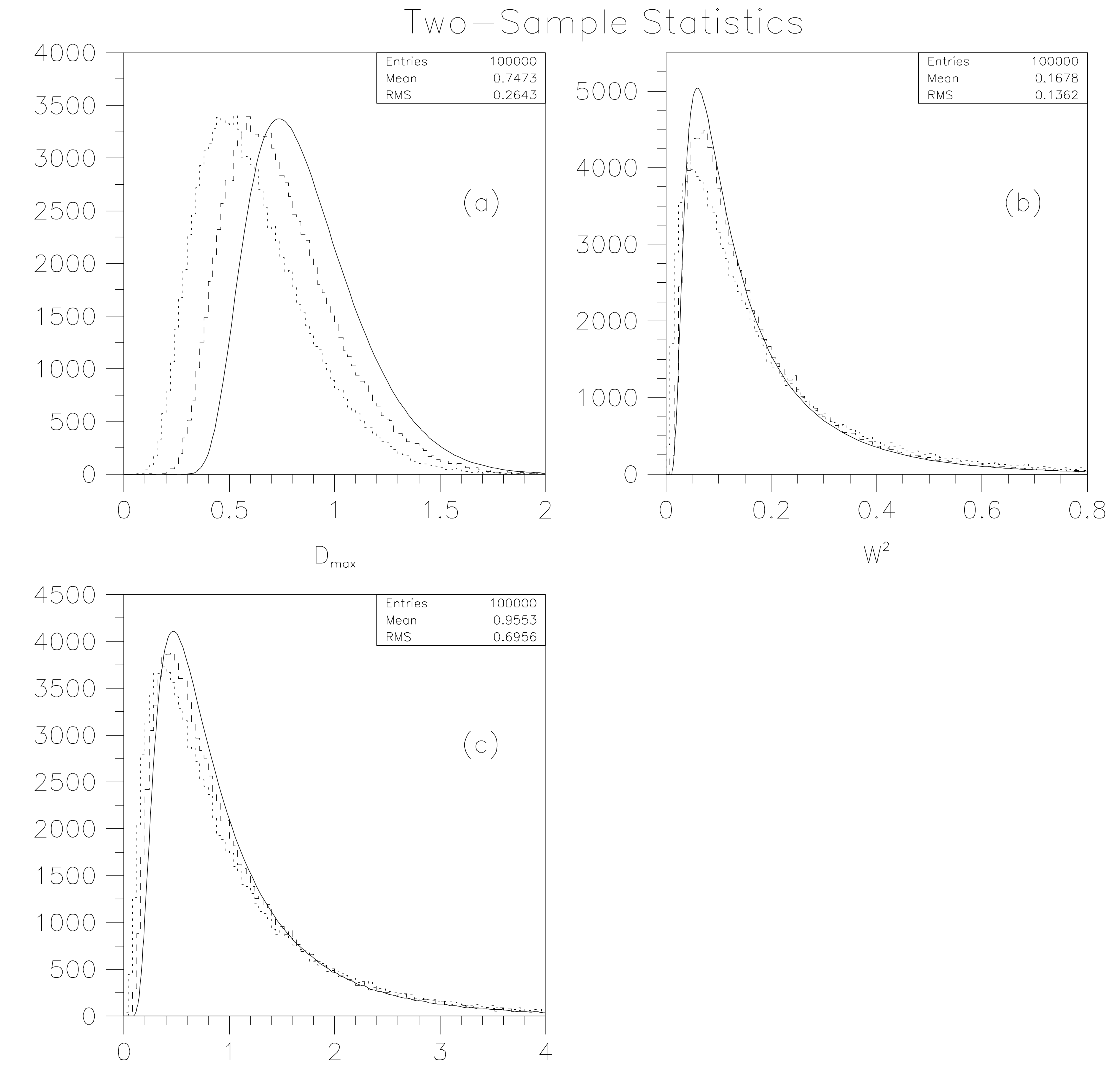

Figure 6: Distributions of the two-sample Kolmogorov-Smirnov statistic (a), Smirnov-Cramér-von Mises statistic (b), and Anderson-Darling statistic (c) for the E_T-*binned* run 1A inclusive jet sample (cfr. table 1). The dashed and dotted histograms were obtained by using bins 5 through 41, respectively 30 through 41, to calculate the statistics. The solid lines are the result of numerically differentiating the survivor functions $\mathbf{S_{KS}}$, $\mathbf{S_{SCvM}}$ and $\mathbf{S_{AD}}$ respectively, and are normalized to the same areas as the corresponding histograms.

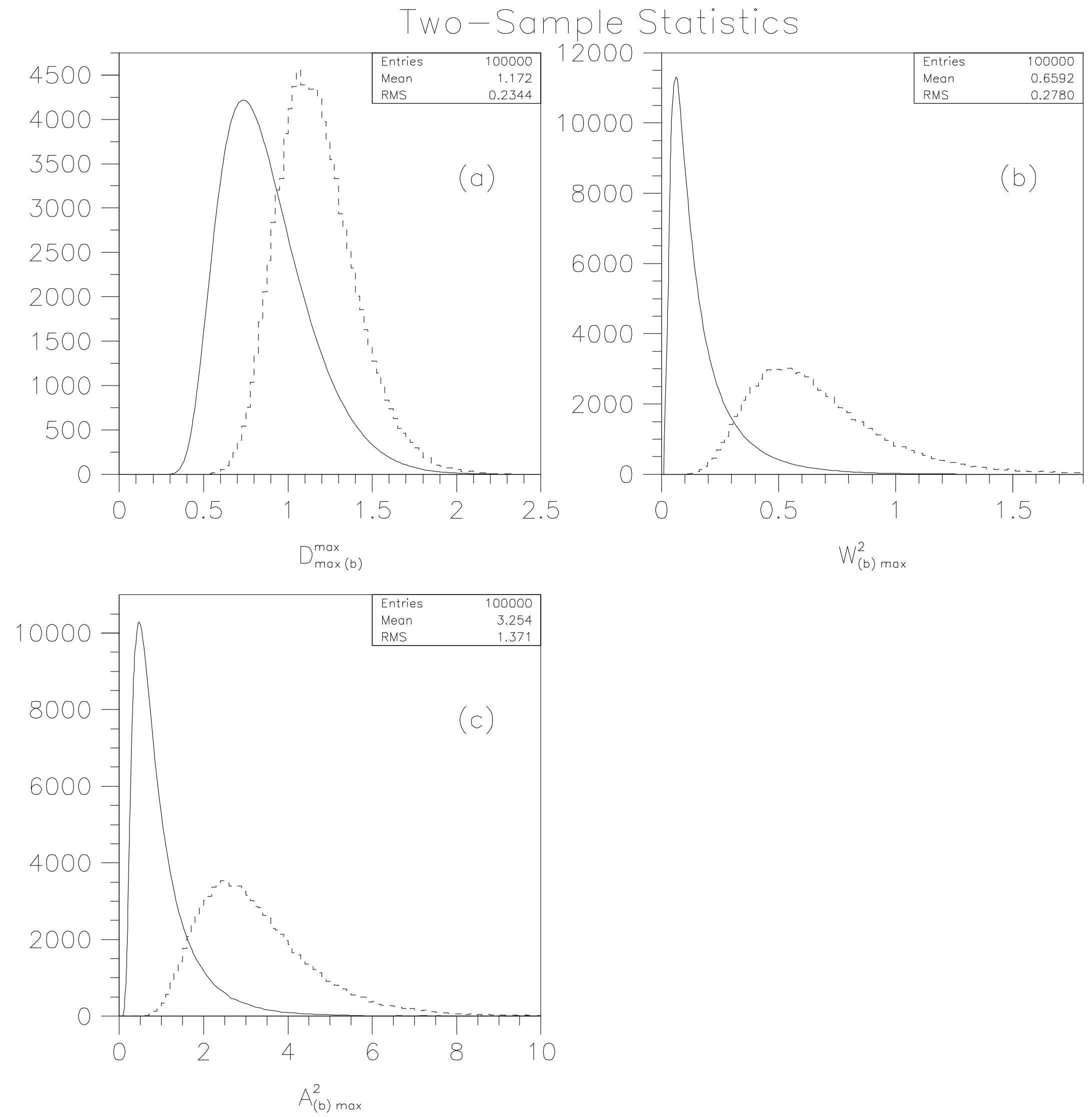

Figure 7: The dashed histograms are distributions of the two-sample deviation statistics $D^{\max}_{\max(b)}$ (plot a), $W^2_{(b)\max}$ (plot b), and $A^2_{(b)\max}$ (plot c) for the E_T-*binned* run 1A inclusive jet sample. The solid lines are the result of numerically differentiating the survivor functions $\mathbf{S_{KS}}$, $\mathbf{S_{SCvM}}$ and $\mathbf{S_{AD}}$ respectively, and are normalized to the same areas as the corresponding histograms.

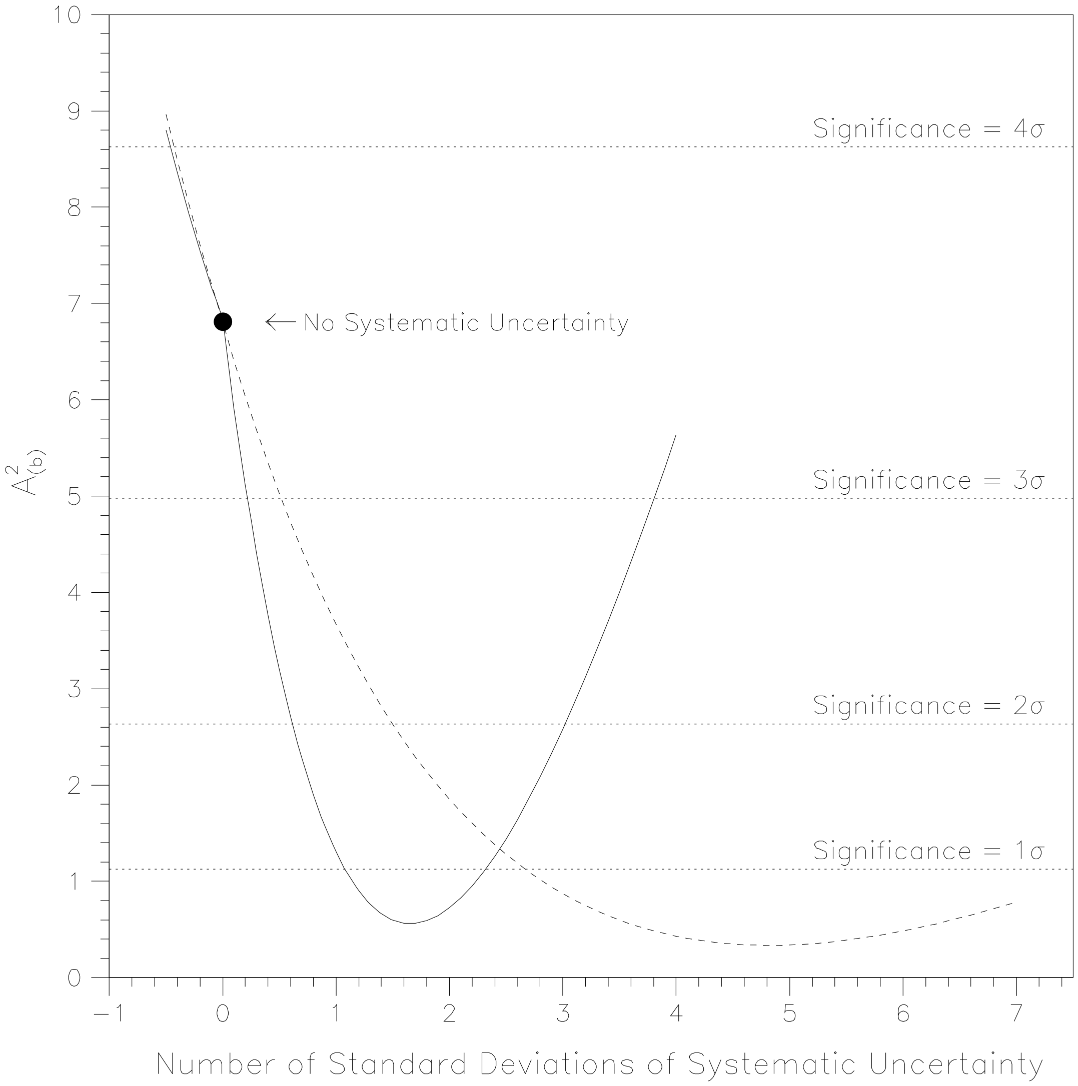

Figure 8: Comparison of inclusive jet data and theory: the two curves show the variation of the statistic $A^2_{(b)}$ as a function of the number of standard deviations of systematic uncertainty added to the theoretical distribution. The uncertainty on the stability of the absolute calibration of the calorimeter was set to 1% (dashed curve) and 2.5% (solid curve). The dotted horizontal lines show various significance levels associated with given values of $A^2_{(b)}$.