

The Kolmogorov-Smirnov Test When Parameters Are Estimated From Data

Hovhannes Keutelian
Fermilab

Abstract

This note contains a description of a Monte Carlo procedure to prepare a distribution of Kolmogorov-Smirnov test. For a given hypothetical distribution, the result of the K-S test becomes distribution dependent if any parameter of the given distribution is estimated from data.

1 Introduction

The Kolmogorov-Smirnov test allows bin independent testing whether or not a given hypothetical distribution describes a set of observations. This test is based on comparing the cumulative distribution function $F(x)$ with the equivalent distribution of data $S_N(x)$. The cumulative function $S_N(x)$ is defined by

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

An example of $S_N(x)$ is shown in figure 1, and note that $S_N(x)$ always increases in steps of equal height, N^{-1} .

The Kolmogorov-Smirnov test is a measure of the "distance" between the experimental and hypothetical distribution functions. In this note, I use the the maximum absolute difference as the test statistic, $D_N = \max|S_N(x) - F(x)|$. If no *parameter* in $F(x)$ has been determined from the data, the variable D_N has a distribution which is independent of $F(x)$. For this case, the cumulative distribution of D_N for large N is given by[1], [2],[3],[4]

$$\lim_{N \rightarrow \infty} P(z) = 1 - 2 \sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2 z^2} \quad (2)$$

where $z = \sqrt{N}D_N$. This cumulative distribution can be used to quote the probability of finding a "distance" D equal to or greater than D_N . The difference, $1-P(z)$, gives the confidence level of the match between $F(x)$ and $S_N(x)$. The analytical function always gives higher confidence levels for finite N ; the value of $P(z)_N$ for finite N is larger compared to $P(z)$.

In the next sections, I will give the details of how to prepare a D_N distribution; and how to obtain a confidence level based on this distribution.

2 Preparing a D_N distribution

All results in this note were prepared based on an algorithm that I am about to describe. The results are obtained using the uniform probability distribution. Other probability functions can also be used to demonstrate the steps given in this algorithm. This algorithm determines the maximum distance D_N in a one dimensional distribution. The procedures are described below.

1. Ten events are selected from a uniform distribution within a range of 0 to 1. The values of these ten events are: 0.9374, 0.7629, 0.4771, 0.5111, 0.8701, 0.0684, 0.7375, 0.5615, 0.2835, 0.2508.
2. These numbers are sorted in ascending order: 0.0684, 0.2508, 0.2835, 0.4771, 0.5111, 0.5615, 0.7375, 0.7629, 0.8701, 0.9374.
3. The cumulative distribution of data is constructed, $S_N(x)$:

Range of observation	$S(x)$
[0.0000, 0.0684)	0.0
[0.0684, 0.2508)	0.1
[0.2508, 0.2835)	0.2
[0.2835, 0.4771)	0.3
[0.4771, 0.5111)	0.4
[0.5111, 0.5615)	0.5
[0.5615, 0.7375)	0.6
[0.7375, 0.7629)	0.7
[0.7629, 0.8701)	0.8
[0.8701, 0.9374)	0.9
[0.9374, 1.0000]	1.0

The plot of $S_N(x)$ is shown in figure 1.

4. The differences between $F(x)$ and $S_N()$ are calculated. These differences are listed in table 1. Note that the differences *at* and *just below* the x values *are* calculated to obtain the correct maximum absolute difference. This is an important operation. The maximum absolute difference for this sample is $D_{10} = 0.177$ and occurs at $\lim_{\epsilon \rightarrow 0} x = 0.4771 - \epsilon$,

The FORTRAN code for the above procedure is:

- $N = 10$
step = 1./float(N)

Table 1: The differences between $F(x)$ and $S_N(x)$.

Location of the event	$F(x)-S(x)$	Comment
$0.0684 - \epsilon$	0.0684	as $\lim_{\epsilon \rightarrow 0}$
0.0684	-0.0316	
$0.2508 - \epsilon$	0.1508	as $\lim_{\epsilon \rightarrow 0}$
0.2508	0.0508	
$0.2835 - \epsilon$	0.0835	as $\lim_{\epsilon \rightarrow 0}$
0.2835	-0.0165	
$0.4771 - \epsilon$	0.1771	as $\lim_{\epsilon \rightarrow 0}$
0.4771	0.0771	
$0.5111 - \epsilon$	0.1111	as $\lim_{\epsilon \rightarrow 0}$
0.5111	0.0111	
$0.5615 - \epsilon$	0.0615	as $\lim_{\epsilon \rightarrow 0}$
0.5615	-0.0385	
$0.7375 - \epsilon$	0.1375	as $\lim_{\epsilon \rightarrow 0}$
0.7375	0.0375	
$0.7629 - \epsilon$	0.0629	as $\lim_{\epsilon \rightarrow 0}$
0.7629	-0.0371	
$0.8701 - \epsilon$	0.0701	as $\lim_{\epsilon \rightarrow 0}$
0.8701	-0.0299	
$0.9374 - \epsilon$	0.0374	as $\lim_{\epsilon \rightarrow 0}$
0.9374	-0.0626	

```

do 101 i=1,N
101 x(i) = ran(iseed)
call hsort(N,x)
D = 0.0
do 111 i=1,N
S = x(i) - float(i)/float(N)
D = max(ABS(S),D)
111 D = max(ABS(S+step),D)

```

For a given x , the difference of $F(x)$ and $S_N(x)$ is stored in the variable S , then its absolute value is compared against D .

The above steps are repeated 100,000 times to obtain just as many values of D_{10} , and the distribution of 100,000 D_{10} s is shown in figure 2. This distribution can be used to set confidence levels. The confidence level for a given value of D_{10} is the area under the curve between that value and 1. In this distribution, 86130 entries have values greater than the maximum absolute difference of $D_{10} = 0.1771$. Hence, the probability of observing a distance

greater than $D_{10} = 0.1771$ is 86130/100000. Therefore, with a confidence level of 86%, the sample can be described by a uniform probability distribution.

The D_N and the $\sqrt{N}D_N$ distributions for $N=1, \dots, 9$ are shown in figures 3, 4, 5 for the purpose of pointing out the \sqrt{N} scaling. The shapes of the $\sqrt{N}D_N$ distributions converge to a common shape as N gets larger. The mean value of a D_N distribution gets smaller as N increases; this is not true for the corresponding $\sqrt{N}D_N$ distribution. The mean value of the $\sqrt{N}D_N$ distribution converges to a constant as N increases. This scaling behavior can be seen in equation 2 where $z = \sqrt{N}D_N$.

3 The K-S test for the case where parameters are estimated from data

When calculating the maximum "distance" D_N , any parameter that enters in the integral probability function can either be given or determined from data. In the previous section, the $D_N(x)$ notation is used to refer to a K-S distribution when no parameters are estimated from data. I will use $D_{Np}(x)$ notation to refer to a K-S distribution when parameters are estimated from data.

In the case where any parameter of an integral probability function is estimated from data, that data set must contain at least two events. The constraint of having a minimum number of events in a data set makes the value D_N dependent on the integral probability distribution. This constraint breaks the one-to-one relation between $F(x)$ and $S_N(x)$, hence forcing the $D_{Np}(x)$ distribution to differ from $D_N(x)$ distribution. Past discussions of the Kolmogorov-Smirnov tests with estimated parameters can be found in references [5] and [6]. The author of the first reference has used a Monte Carlo method to show that the K-S test has the same properties as in the case where no parameters are estimated from data. The author of the second reference has relied on the analytical approach to prove the the properties of K-S test. The Monte Carlo method is more efficient than the analytical approach, and is applicable for a wide variety of probability distributions.

I will use a Gaussian distribution to demonstrate the dependence of D_N on $F(x)$ when parameters are estimated from data. Except for the change of the integral probability distribution, the algorithm from the previous section is used to prepare $D_{Np}(x)$ distribution. Here is the algorithm:

- $N = 10$
step = 1./float(N)
rtmp = 0
do 101 i=1,N
x(i) = gausdv(iseed)
101 rtmp = rtmp + x(i)
mean = rtmp/float(N)

```

rtmp = 0.
do 121 i=1,N
121 rtmp = rtmp + (x(i) - mean)**2
sigma = sqrt(rtmp/float(N))

call hsort(N,x)
D = 0.
DP = 0.
do 131 i=1,N
Err = ErfLee(x(i),0.,1.)
S = Err - float(i)/float(N)
D = max(ABS(S) ,D)
D = max(ABS(S+step),D)

Err = ErfLee(x(i),mean,sigma)
S = Err - float(i)/float(N)
DP = max(ABS(S) ,DP)
DP = max(ABS(S+step),DP)
131 continue

```

The routine GAUSDV generates events from a Gaussian distribution with mean 0 and sigma 1. The results of the maximum likelihood fit are stored in the variables MEAN and SIGMA. The function ErfLee returns the integral value of a Gaussian distribution for a given x . In the above algorithm, two distances are calculated; D (D_N) and DP (D_{Np}).

With 100,000 generated samples of D_{10} and D_{10p} , the distributions of $D_{10}(x)$ and $D_{10p}(x)$ are shown in figure 6. The distribution in figure 6c is identical, within statistical uncertainties, to the distribution in figure 2. The first distribution is prepared using a Gaussian probability function and the second distribution is prepared using a uniform probability function. Therefore, for both probability functions, the variable D_N has a distribution which is independent of $F(x)$.

In figures 6a, the $D_{10}(x)$ and the $D_{10p}(x)$ distributions are superimposed on one another. The shapes of both distributions are different, where the mean and the R.M.S. values of $D_{10p}(x)$ distribution are smaller compared to $D_{10}(x)$ distribution. This information alone should cast doubt on the confidence level calculations that are based on $D_N(x)$ distribution, in the case where parameters are estimated from data. As an example, I chose a maximum distance $D = 0.150$ to calculate a confidence level. In figure 6b the area between 0.150 to 1.000 is 0.74, where the area is 0.94 in figure 6c. The difference in confidence levels is 20%. This is a significant difference that can lead to a wrong conclusion.

The plots of figure 7 show the difference in shapes of $D_{Np}(x)$ and $D_N(x)$ distributions for $N=10,15$, and 20. The displayed mean and the r.m.s. values are for $D_{Np}(x)$ distributions. Note that the \sqrt{N} also holds for the $D_{Np}(x)$ distribution. My observation is based on plots as can be seen in figure D101520d,e,f. Theoretical calculations predict the \sqrt{N} scaling only

in the case when the location (the mean value in a Gaussian distribution) and the scale (the sigma in a Gaussian distribution) parameters are estimated from data. In the case of non-linear parameters, the \sqrt{N} scaling may not hold. Hence, a $D_N(x)$ or a $D_{N_p}(x)$ distribution has to be prepared to obtain a correct confidence level.

The preparation of a D_{N_p} or a D_N distribution is straight forward for any probability function. If an algorithm that calculates D_N is already setup, all that is required is a change of the integral probability function in that algorithm. To get a reasonable estimate of a confidence level, it is enough to generate one thousand samples to prepare a D_N or a D_{N_p} distribution.

4 Comments

This note may leave an impression that a confidence level based on the K-S test is tedious and time consuming compared to a chi-square test, but this is not so. If a proper chi-square test is performed, it will take the same or a greater amount of effort and time. In the proper chi-square test a chi-square distribution has to be prepared for a specific problem. The familiar chi-square method is exact only in the limit of infinitely many observations and with linear parameter dependence; otherwise it is an approximation. Because the K-S test allows a bin independent testing of hypothesis, it is superior to the chi-square test for small samples.

5 Conclusion

Plots and examples in this note show that the standard K-S tables and equation 2 always give an overestimated confidence values in the case where parameters are estimated from data. The $D_N(x)$ and $D_{N_p}(x)$ distributions have different shapes, thus giving significantly different confidence levels. The \sqrt{N} scaling is also true for the case where parameters are estimated from data, but still the shapes of $\sqrt{N}D_N(x)$ and $\sqrt{N}D_{N_p}(x)$ distributions are different.

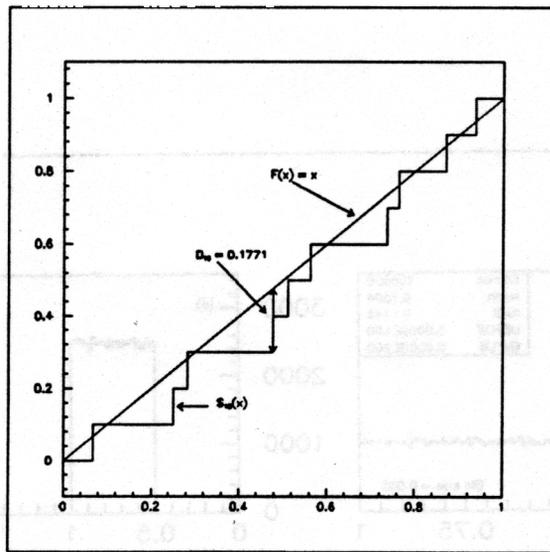


Figure 1: The cumulative distribution of 10 events belonging to a uniform distribution. Each vertical step is 0.1 unit long, and the random variable (horizontal axis) is a dimensionless quantity.

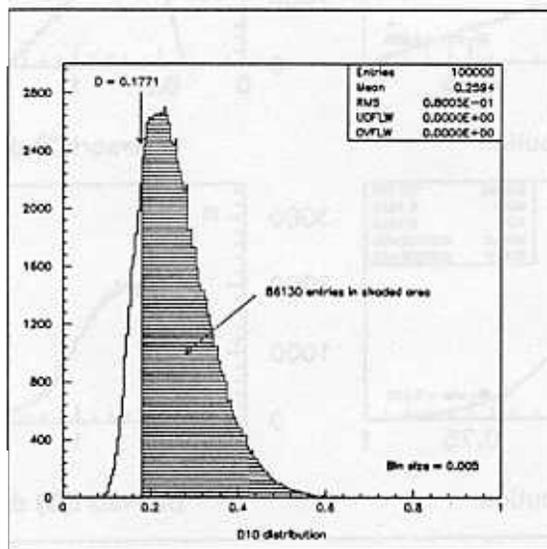


Figure 2: A D_{10} distribution based on 100,000 Monte Carlo samples. The probability of observing a maximum distance greater than 0.1771 is the area under the curve between 0.1771 and 1.0. The probability is $86130/100000 = 0.86$. The smallest value that D_{10} can have is 0.05.

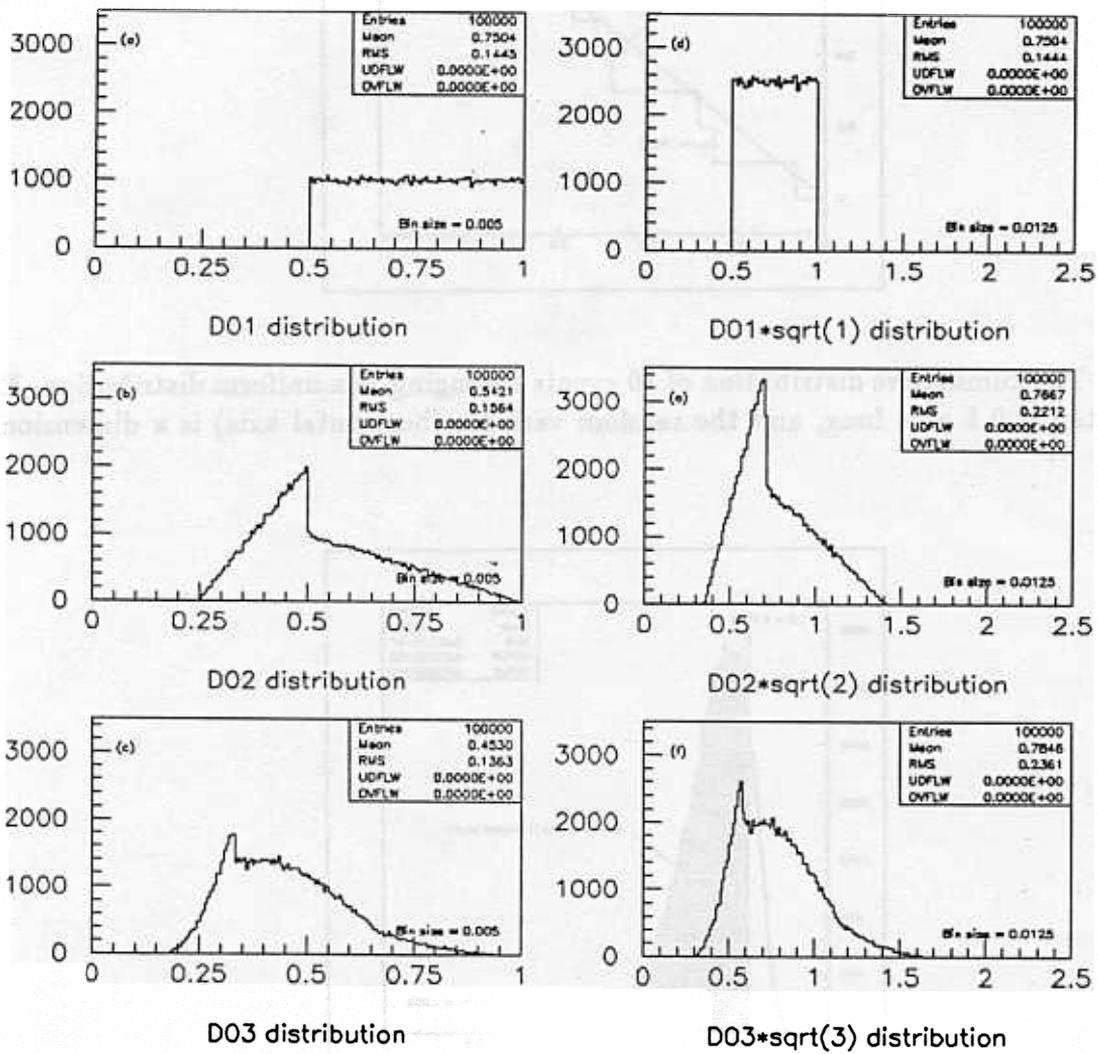


Figure 3: Plots a, b, and c show the K-S distributions for $N=1,2,3$. Plots d, e, and f show the scaled ($\sqrt{ND_N}$) K-S distributions for $N=1,2,3$.

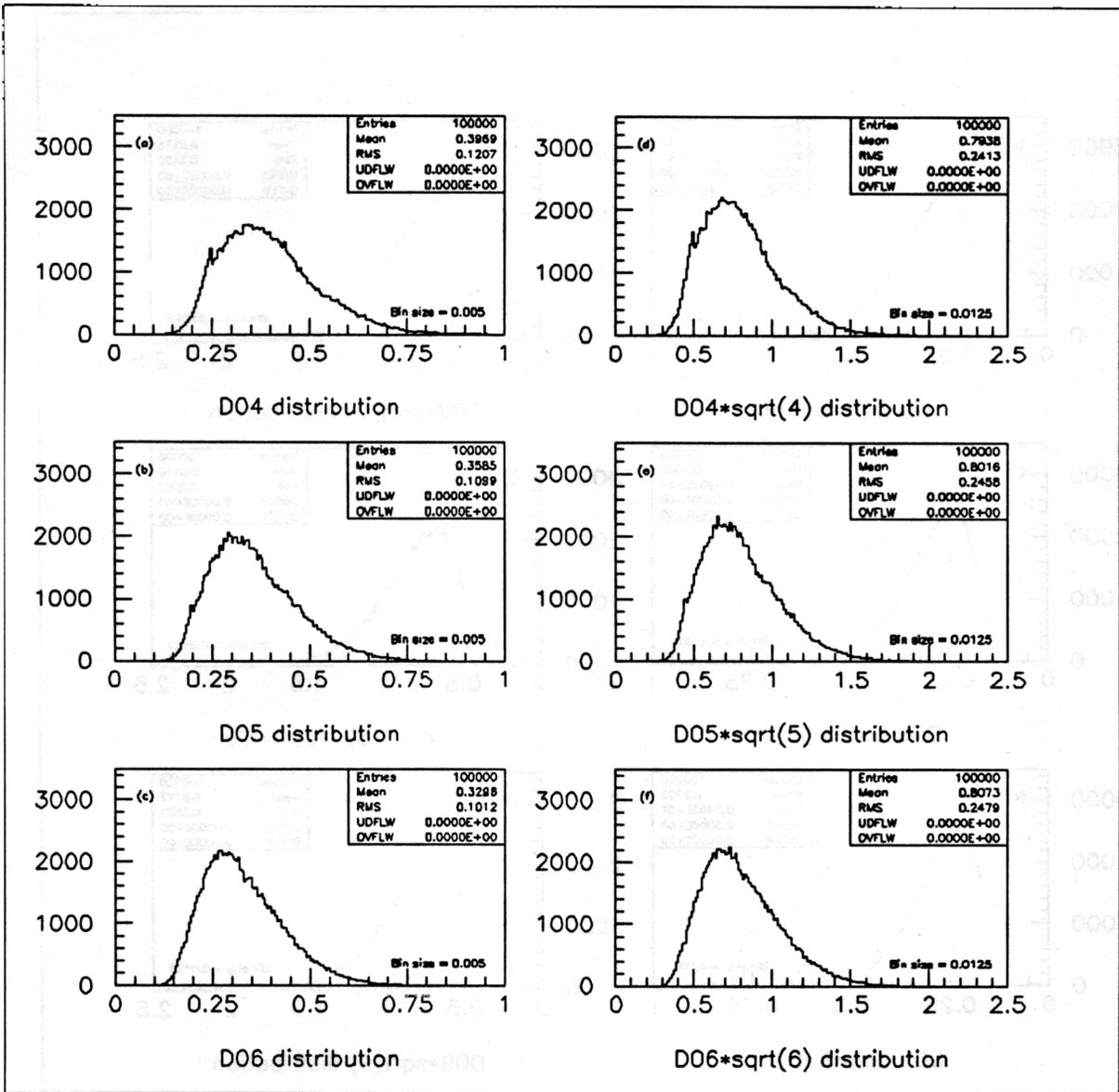
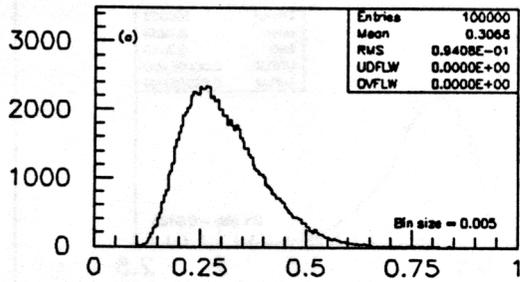
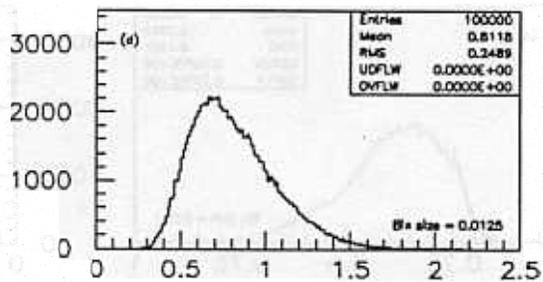


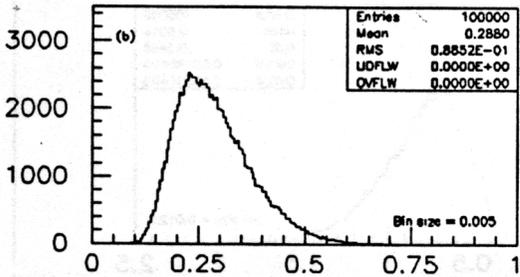
Figure 4: Plots a, b, and c show the K-S distributions for $N=4,5,6$. Plots d, e, and f show the scaled ($\sqrt{ND_N}$) K-S distributions for $N=4,5,6$.



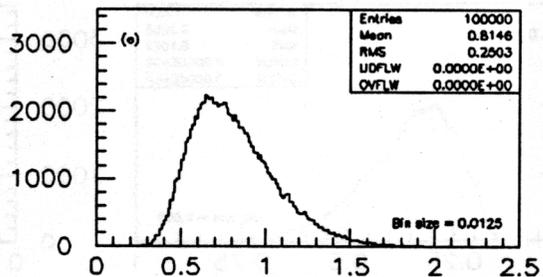
D07 distribution



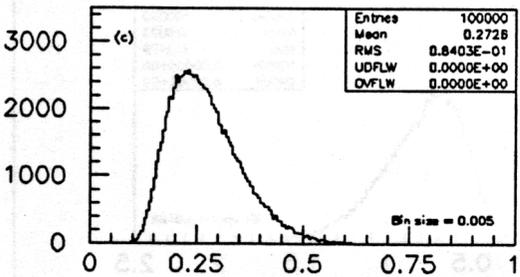
D07*sqrt(7) distribution



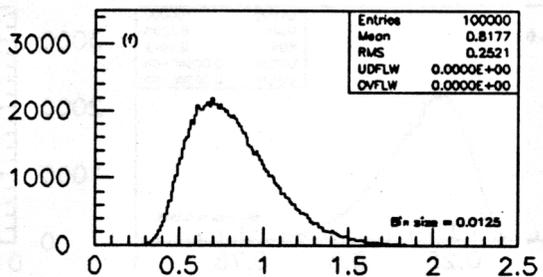
D08 distribution



D08*sqrt(8) distribution

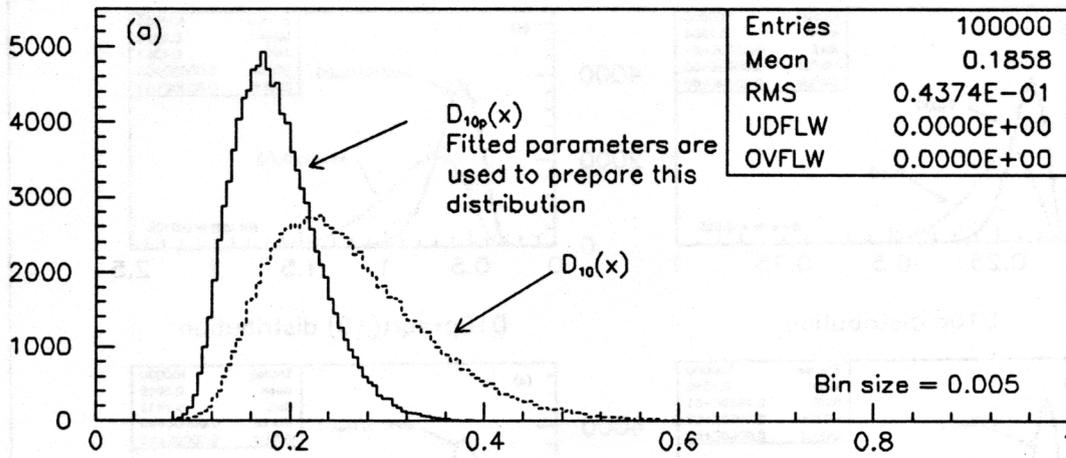


D09 distribution

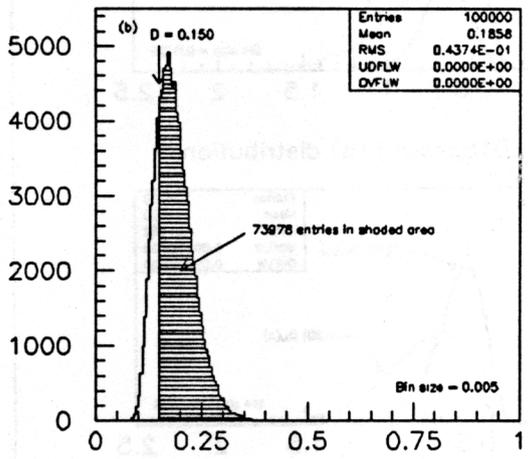


D09*sqrt(9) distribution

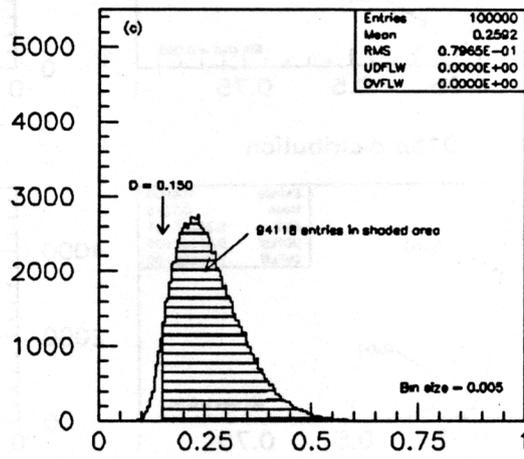
Figure 5: Plots a, b, and c show the K-S distributions for $N=7,8,9$. Plots d, e, and f show the scaled ($\sqrt{N}D_N$) K-S distributions for $N=7,8,9$.



D10p distribution

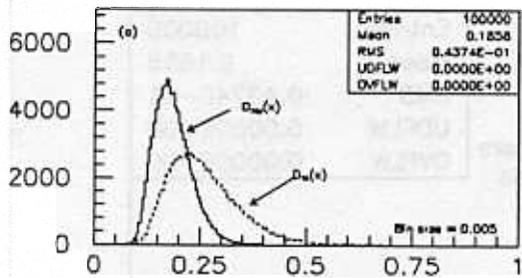


D10p distribution

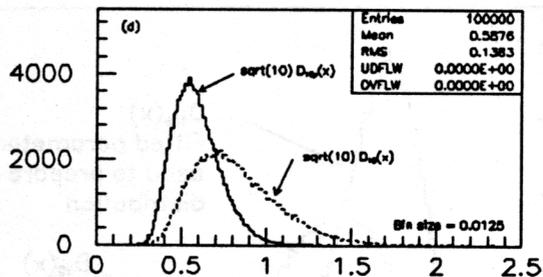


D10 distribution

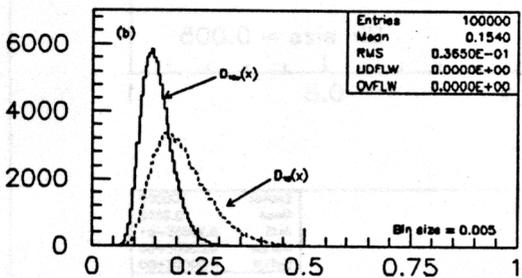
Figure 6: Plot (a) shows the superposition of the K-S distributions that are prepared with and without estimating parameters in a Gaussian distribution. The shaded areas in plots (b) and (c) give different confidence levels for the same D_{10} value.



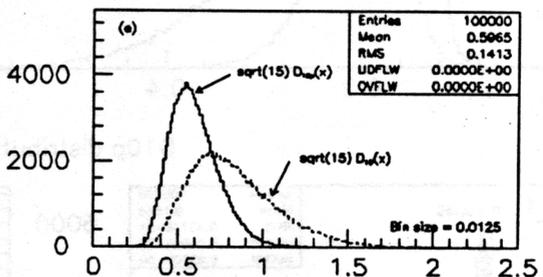
D10p distribution



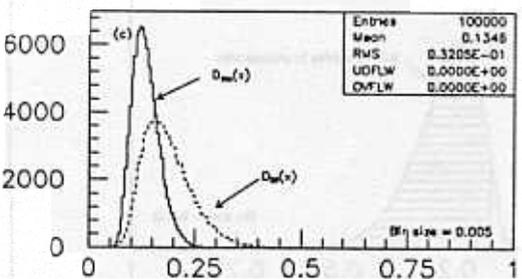
D10p*sqrt(10) distribution



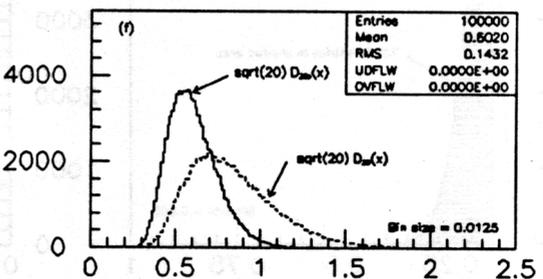
D15p distribution



D15p*sqrt(15) distribution



D20p distribution



D20p*sqrt(20) distribution

Figure 7: Plots a, b, and c show the K-S distributions for $N=10,15,20$. K-S distributions for $N=10,15,20$. The distributions with solid lines are prepared using parameters estimated from data.

References

- [1] A. G. Frodesen, O. Skjeggstad, H. Tøfte. **Probability and Statistics in Particle Physics**. UNIVERSITETSFORLAGET 1979. Pages 424-428.
- [2] W. T. Eadie, D. Drijard, F. E. James, M. Roos, B. Sadoulet. **STATISTICAL METHODS in Experimental Physics**. North-Holland Publishing Company, 1971. Pages 268-271.
- [3] W. Press, B. Flannery, S. Teukolsky, W. Vetterling. **NUMERICAL RECIPIES**. Cambridge University Press, 1986. Pages 472-475.
- [4] M. G. Kendall, A. Stuart. **The Advanced Theory of Statistics**. Charles Griffin and Co. Ltd., London, 1967, Vol. 2 The distribution of D_N for large N is derived on pages 477-480.
- [5] Hubert W Lilliefors. *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown*. American Statistical Association Journal, June 1967. Pages 399-402.
- [6] J. Durbin. *Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings*. *Biometrika*(1975),**62**,1,page 5.