Choosing the right bin width

Károly Banicz

Recipe

$$w_{bin} = \frac{3.49\,\sigma}{N^{\frac{1}{3}}}$$

where σ is the standard deviation, N the sample size

(Scott, D. 1979. *On optimal and data-based histograms.* Biometrika, 66:605-610.)

OR

$$w_{bin} = \frac{2\ IQR}{N^{\frac{1}{3}}}$$

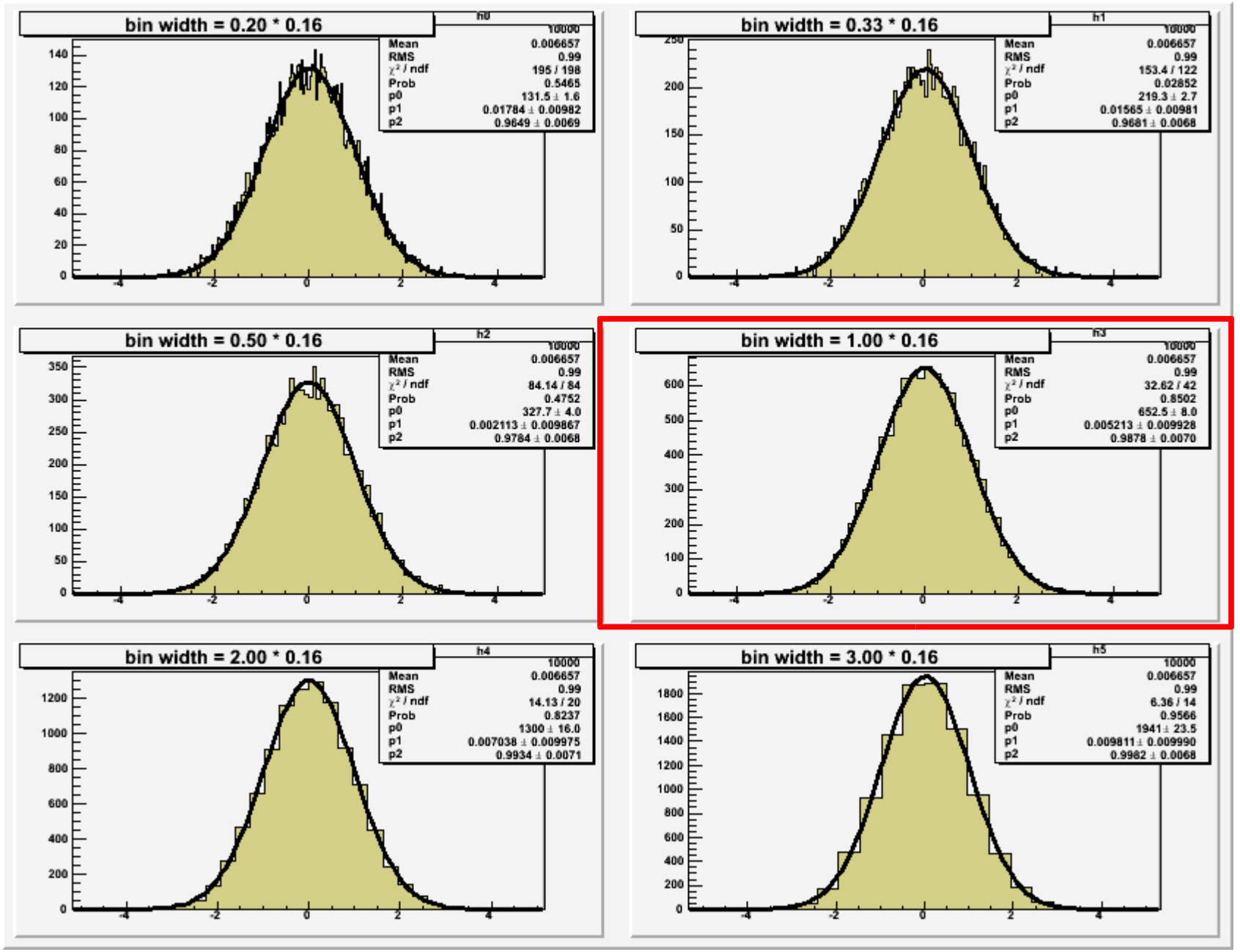
where IQR (interquartile range) is the distance between the 25th and 75th percentiles (= the middle range containing half of the sample)

(Izenman, A. J. 1991. *Recent developments in nonparametric density estimation.* Journal of the American Statistical Association, 86(413):205-224.)

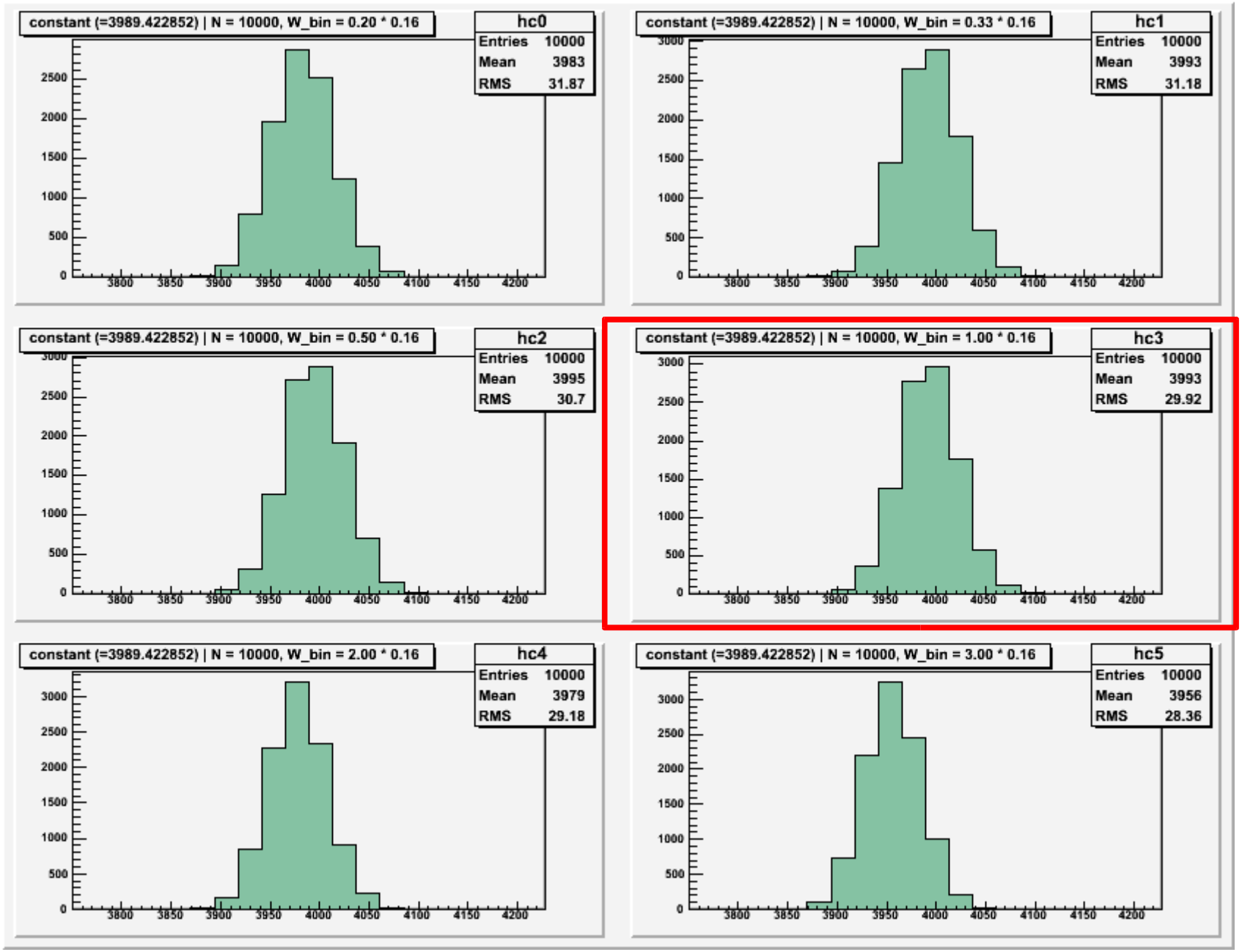
Does it work?

- generate normal p.d.f.-distributed values ($\mu = \sigma = 1$)
- fill histograms of different bin sizes with them
- fit histograms with Gaussian
- repeat thousands of times
- plot the distribution of the fitted parameters
- see which bin size leads to the most accurate fit

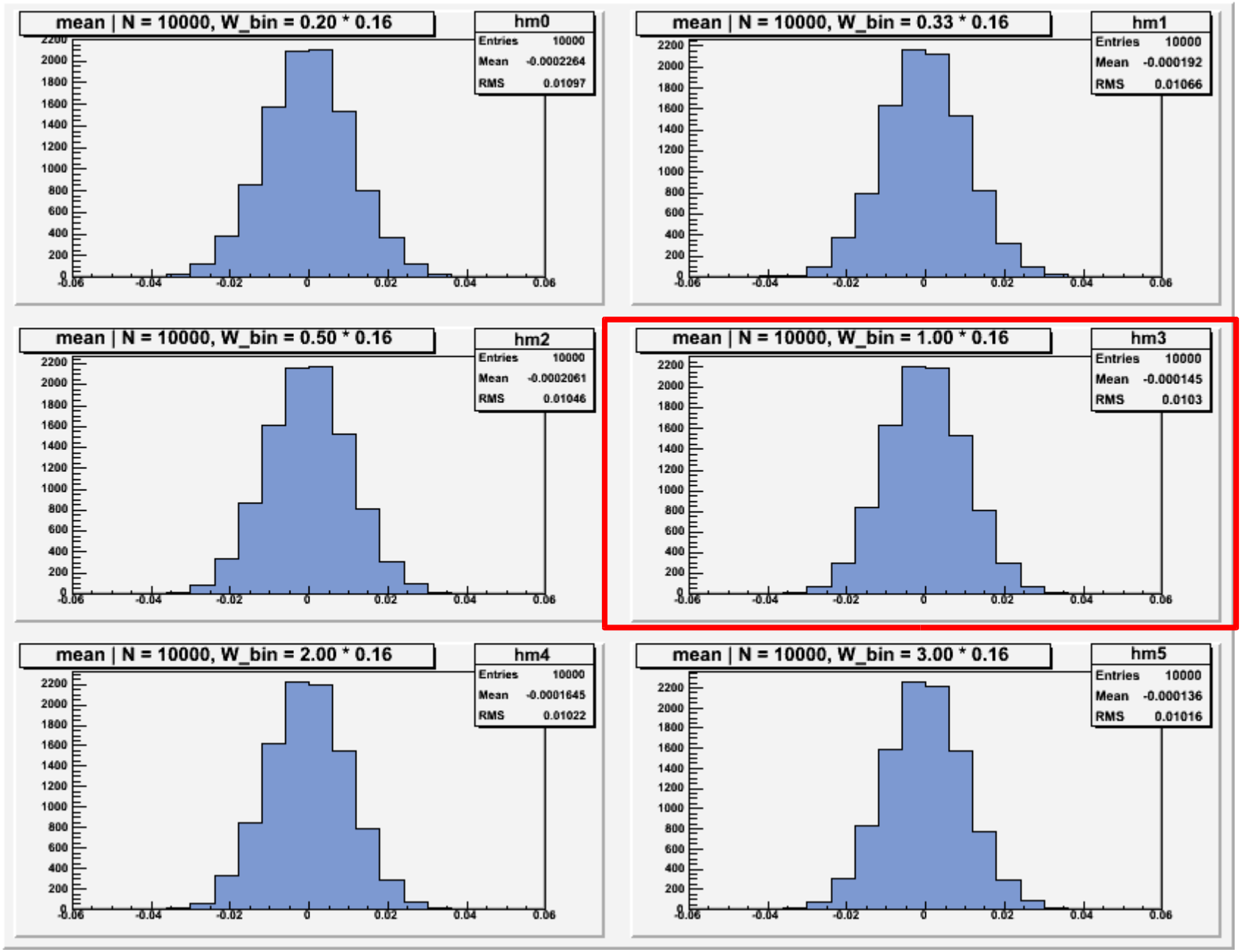
Sample size N=10000



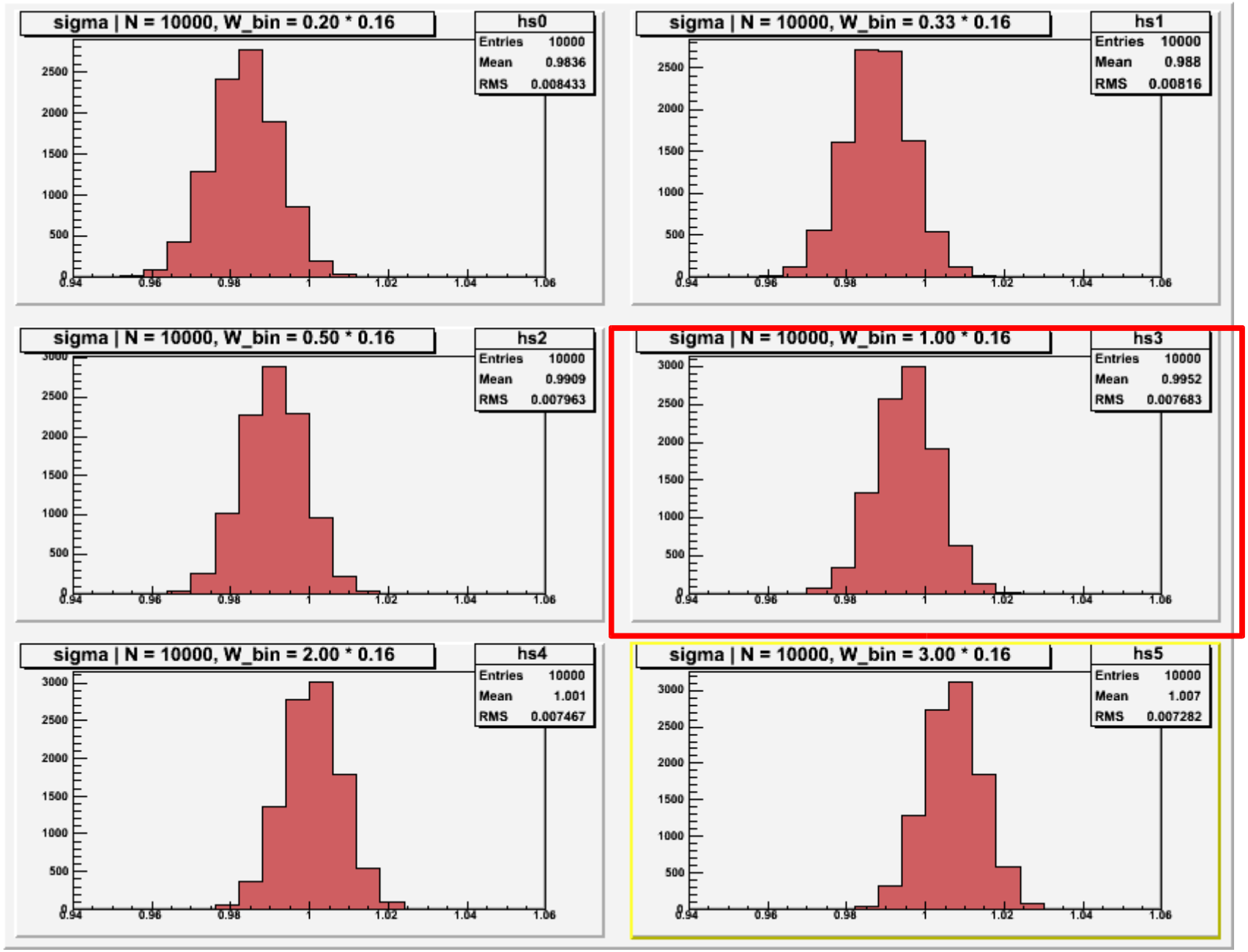
Fitted norm. const. (sample size N=10000)



Fitted mean (sample size N=10000)

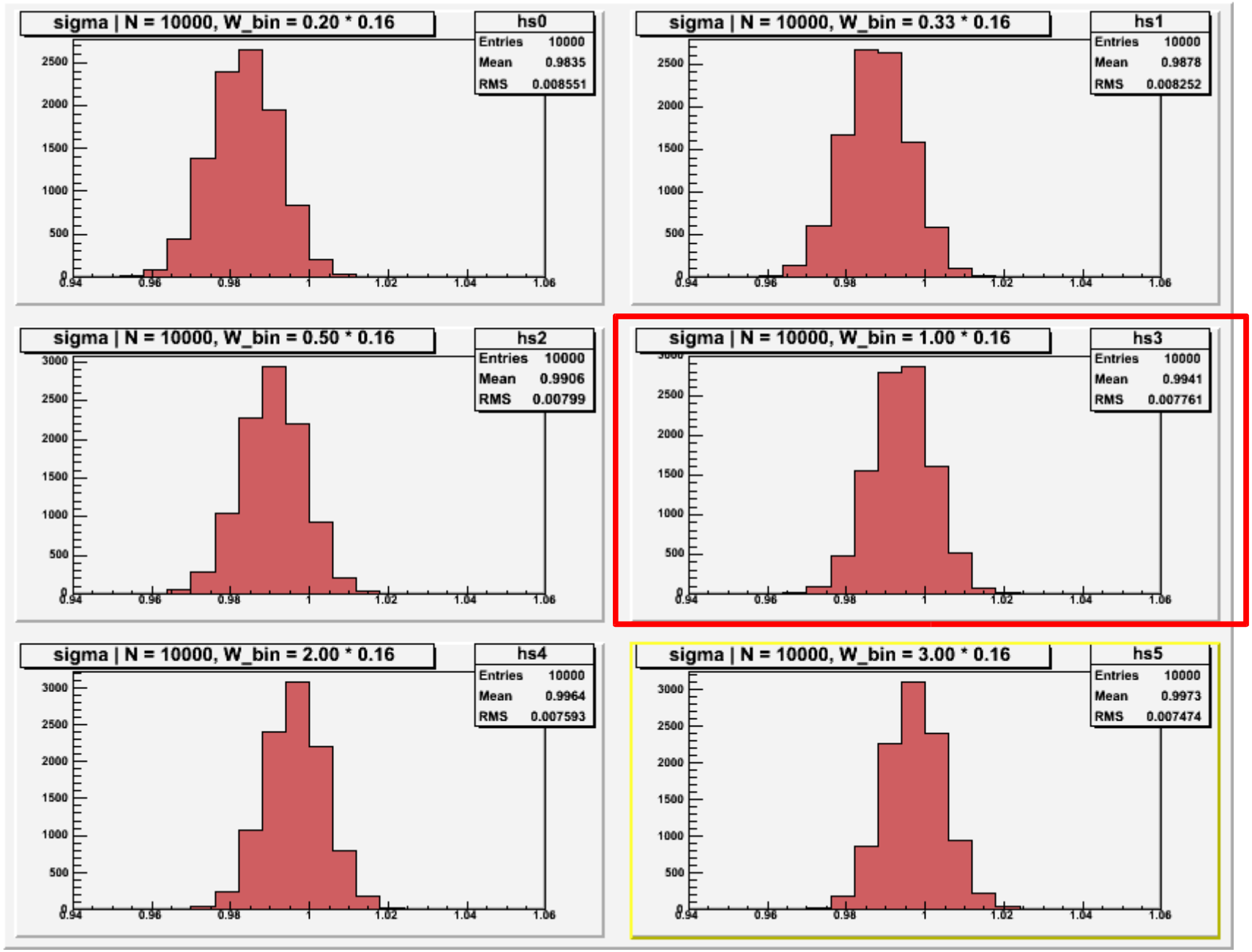


Fitted sigma (sample size N=10000)



Fitted sigma (sample size N=10000)

Fitted with option "I" (integral instead of value at bin center)



w/5

w/3

w/2

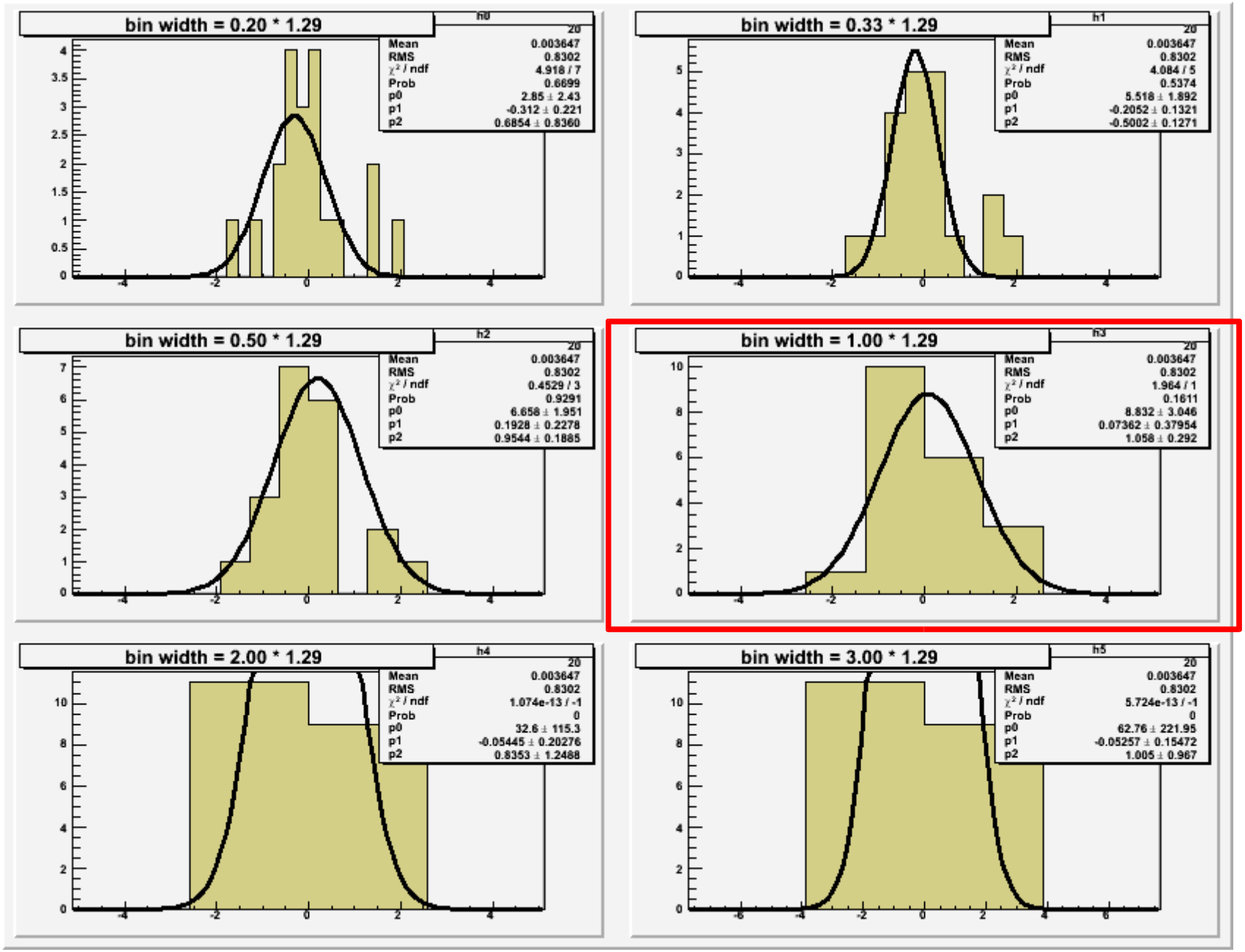
w

2w

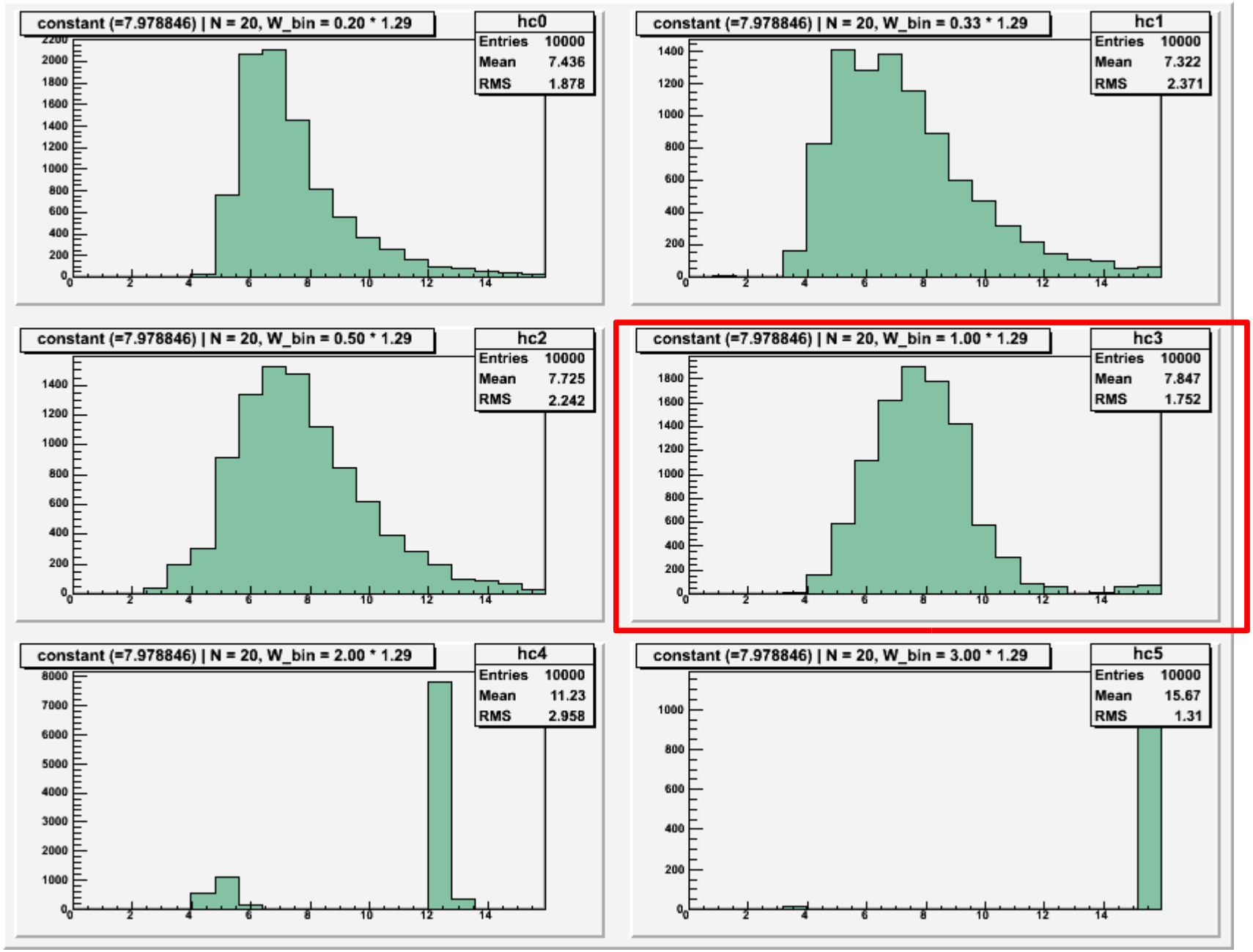
3w

prescribed bin width

Sample size N=20

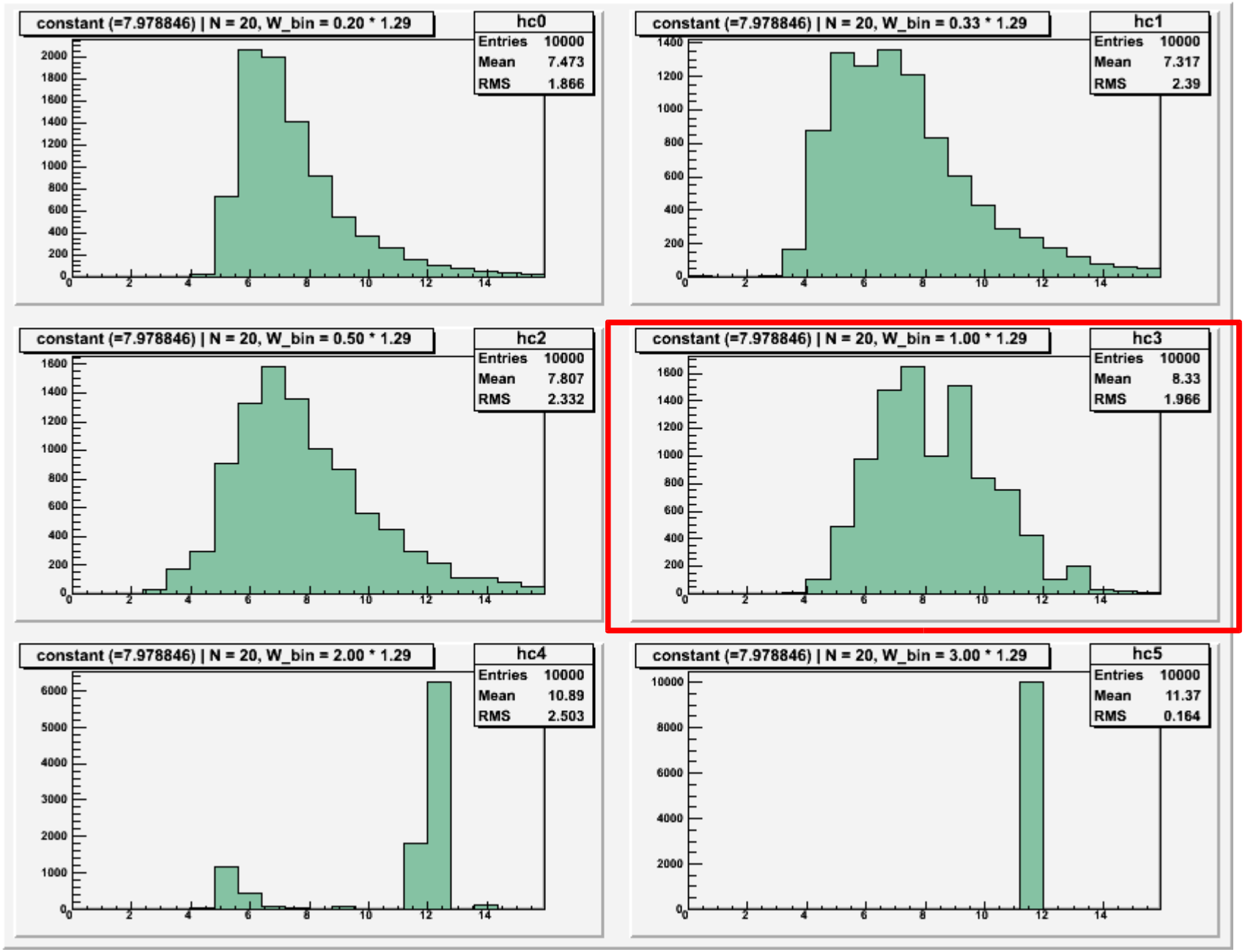


Fitted norm. const. (sample size N=20)



Fitted norm. const. (sample size N=20)

Fitted with option "I" (integral instead of value at bin center)



w/5

w/3

w/2

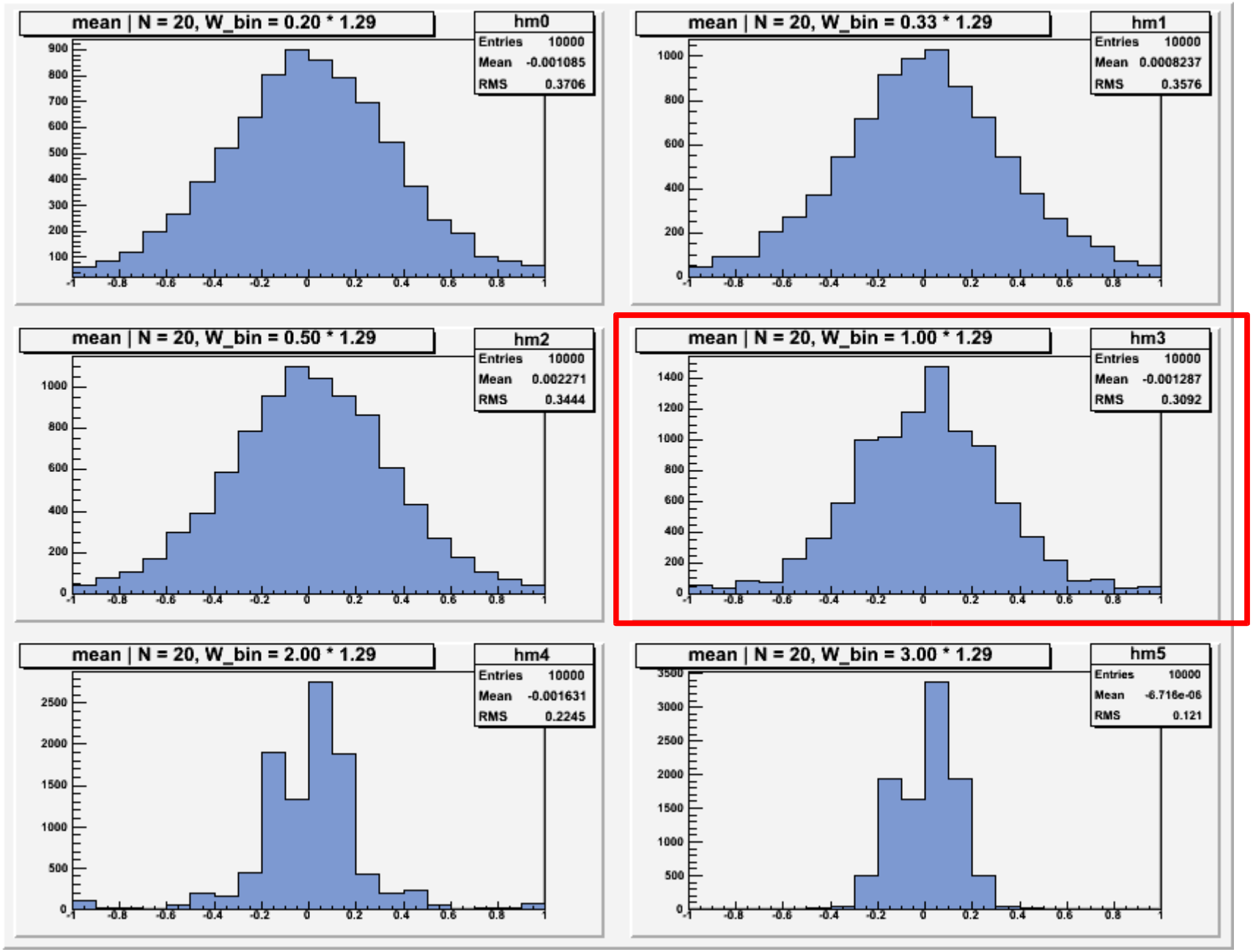
w

prescribed bin width

2w

3w

Fitted mean (sample size N=20)



w/5

w/3

w/2

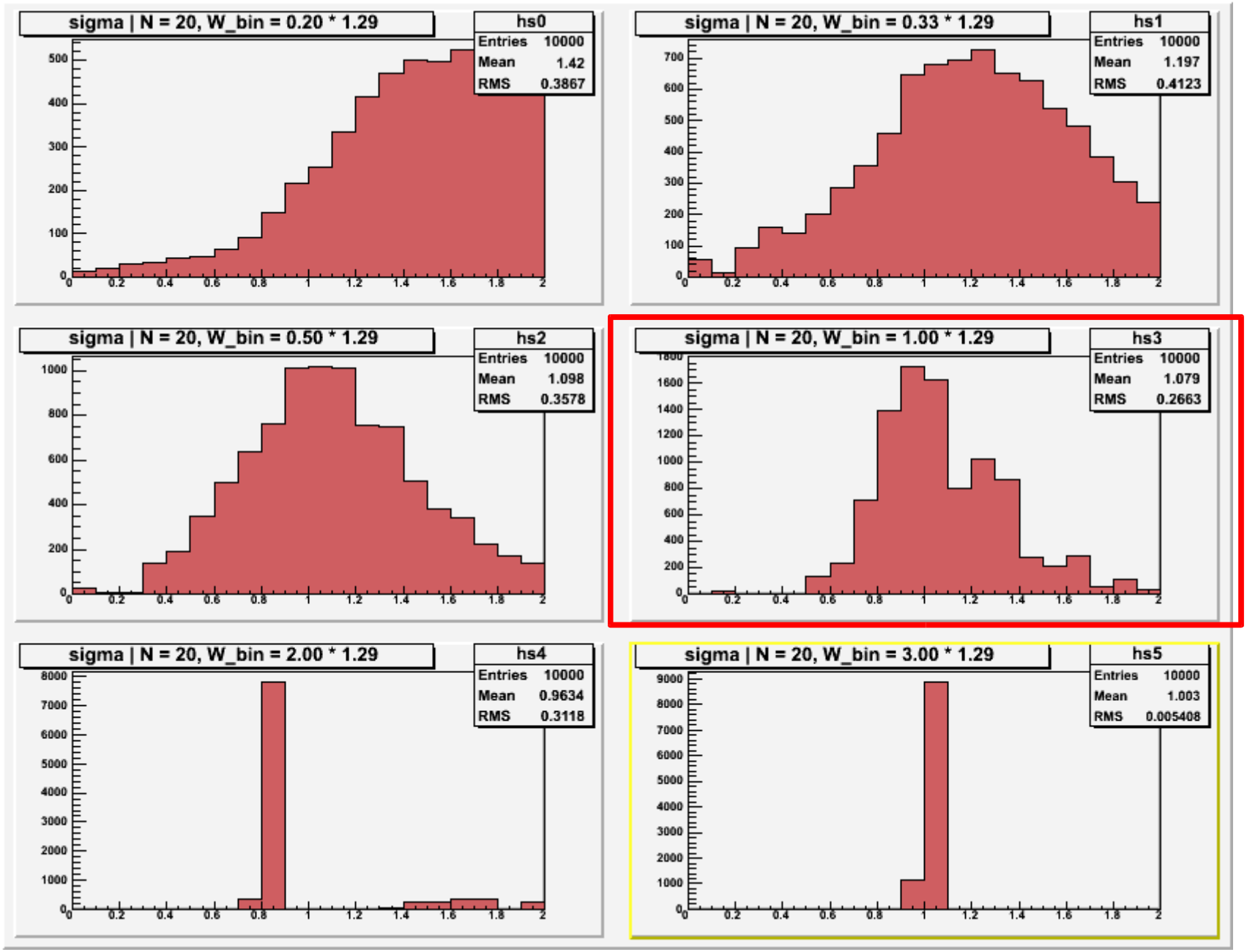
w

2w

3w

prescribed bin width

Fitted sigma (sample size N=20)



w/5

w/3

w/2

w

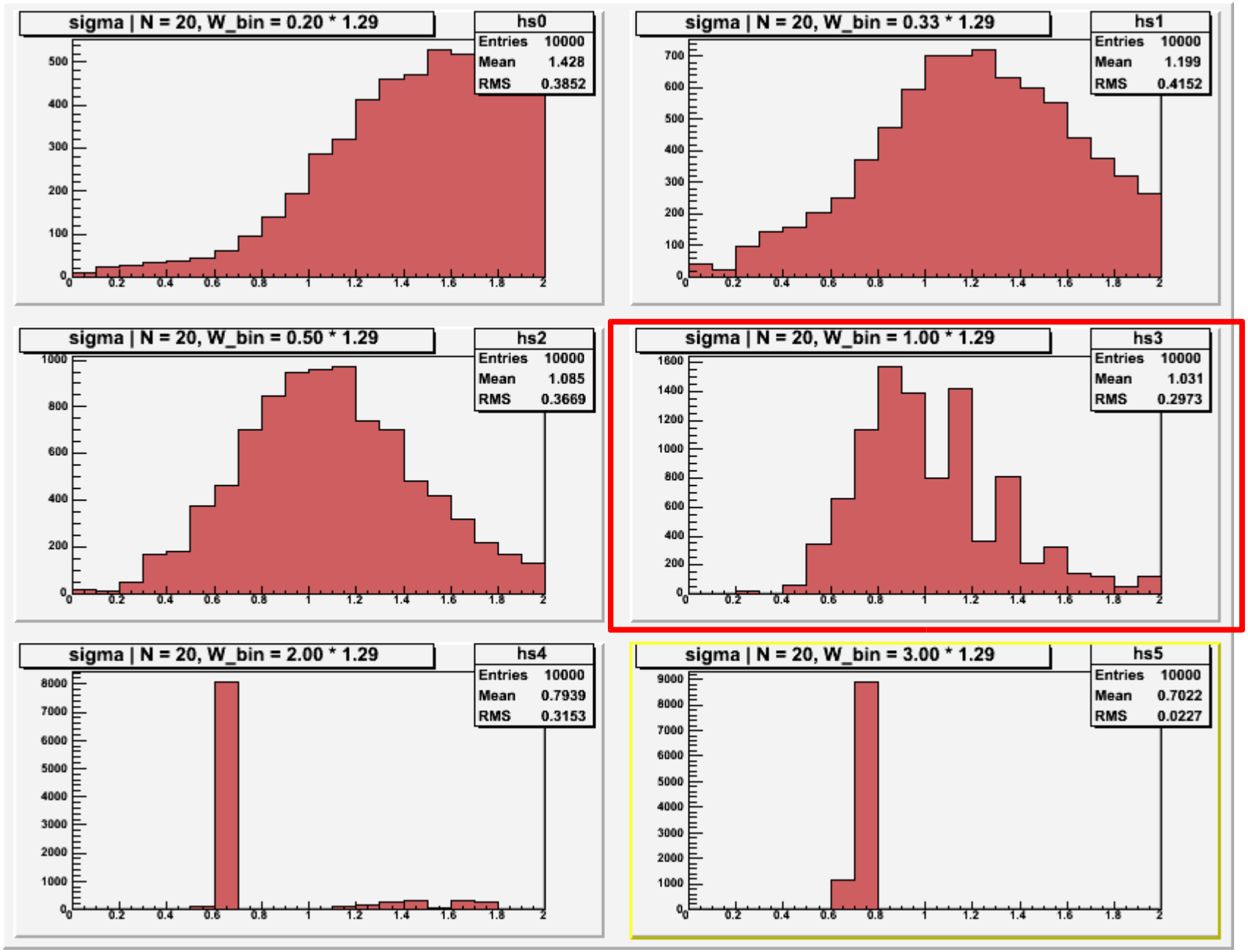
2w

3w

prescribed bin width

Fitted sigma (sample size N=20)

Fitted with option "I" (integral instead of value at bin center)



w/5

w/3

w/2

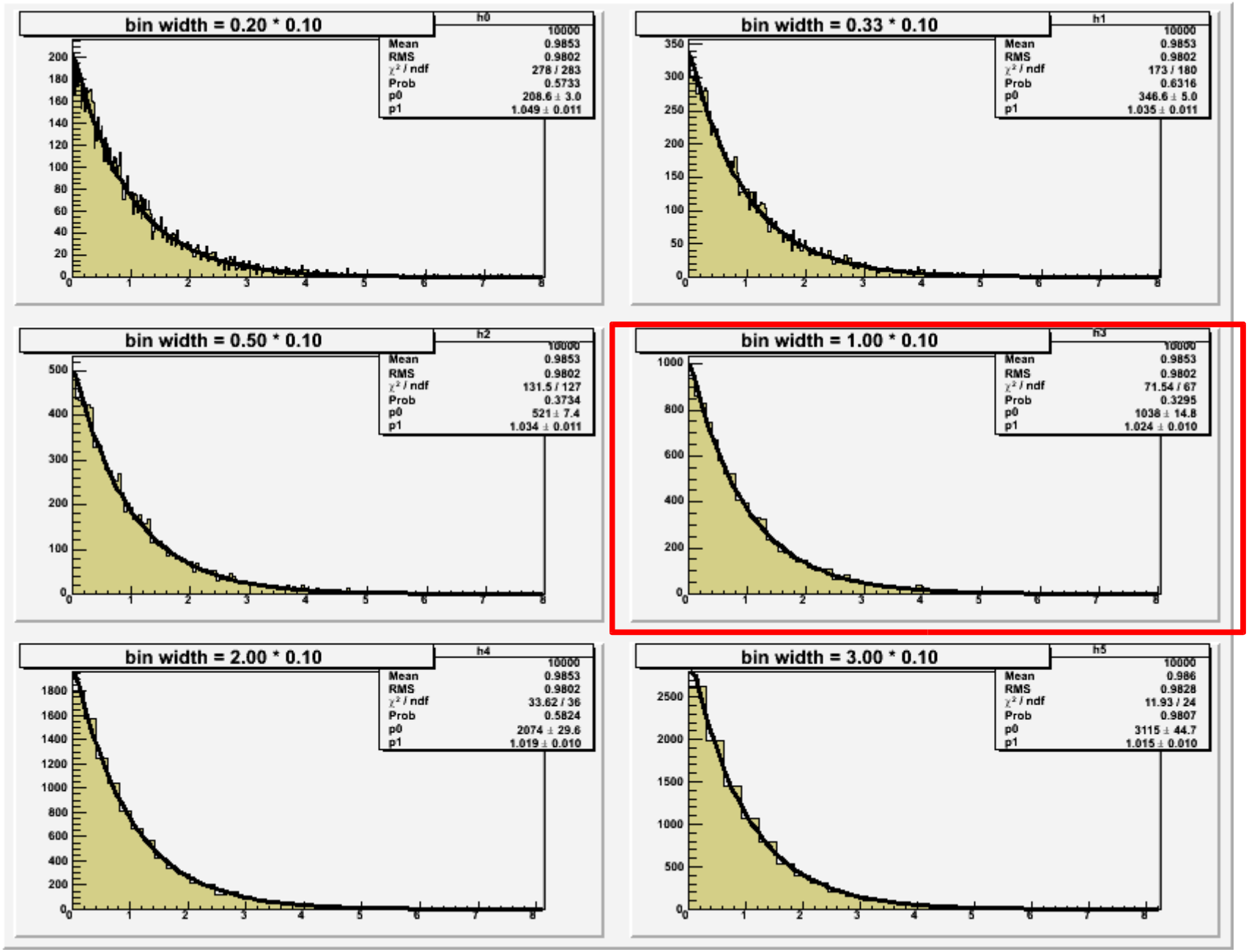
w

2w

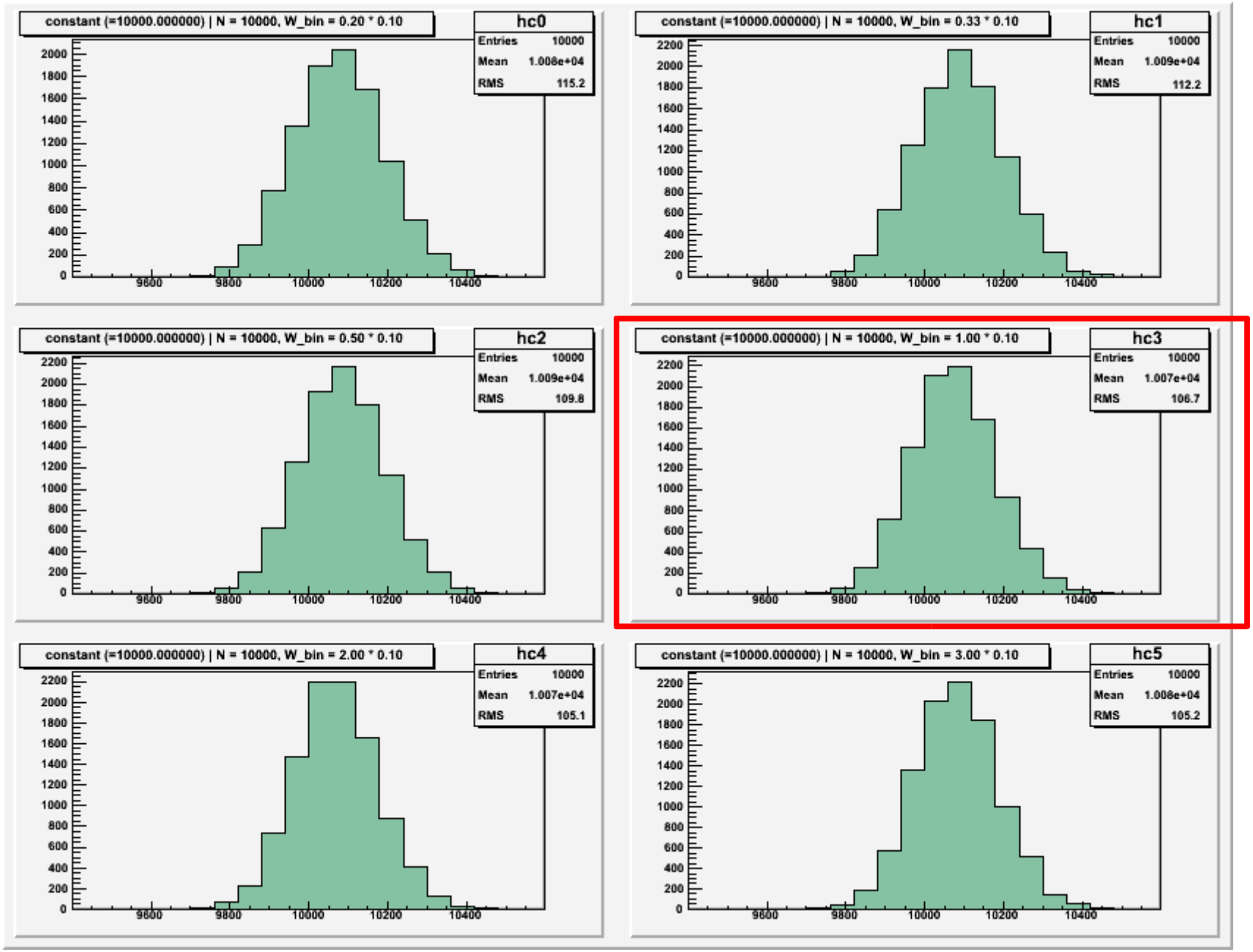
3w

prescribed bin width

Sample size N=10000



Norm. const. (sample size N=10000)



Norm. const. (sample size N=10000)

Fitted with option "I" (integral instead of value at bin center)

w/5

w/3

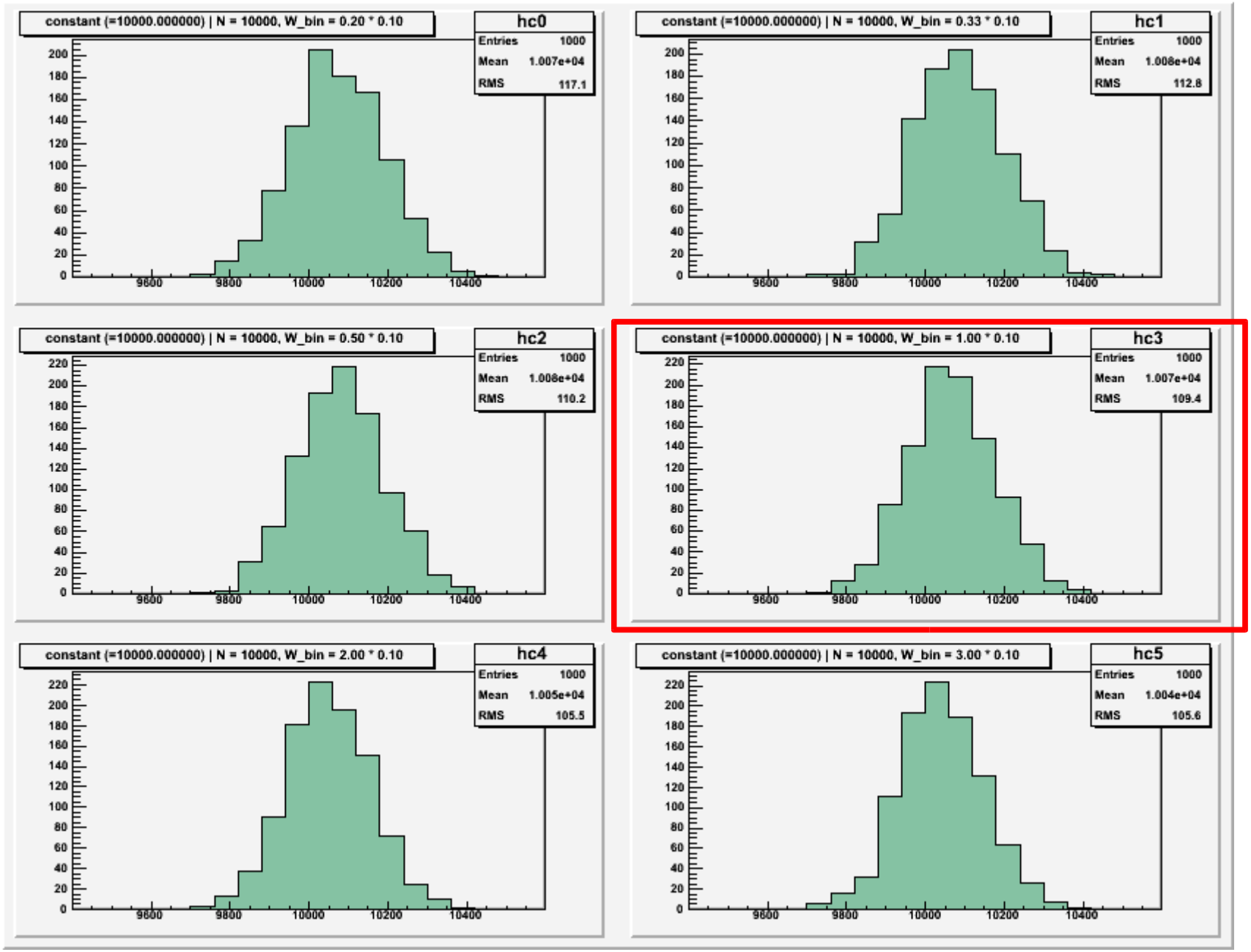
w/2

w

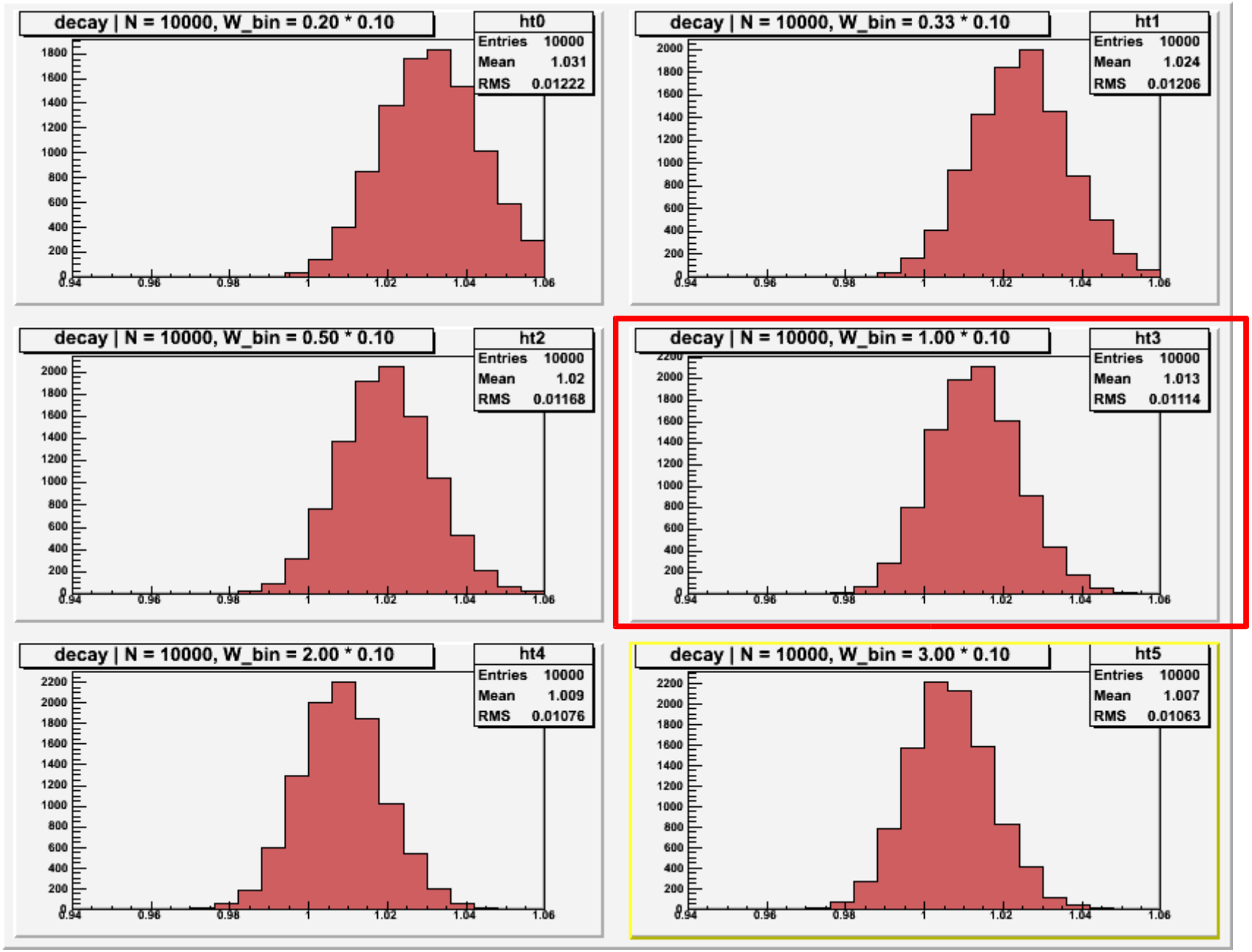
2w

3w

prescribed bin width

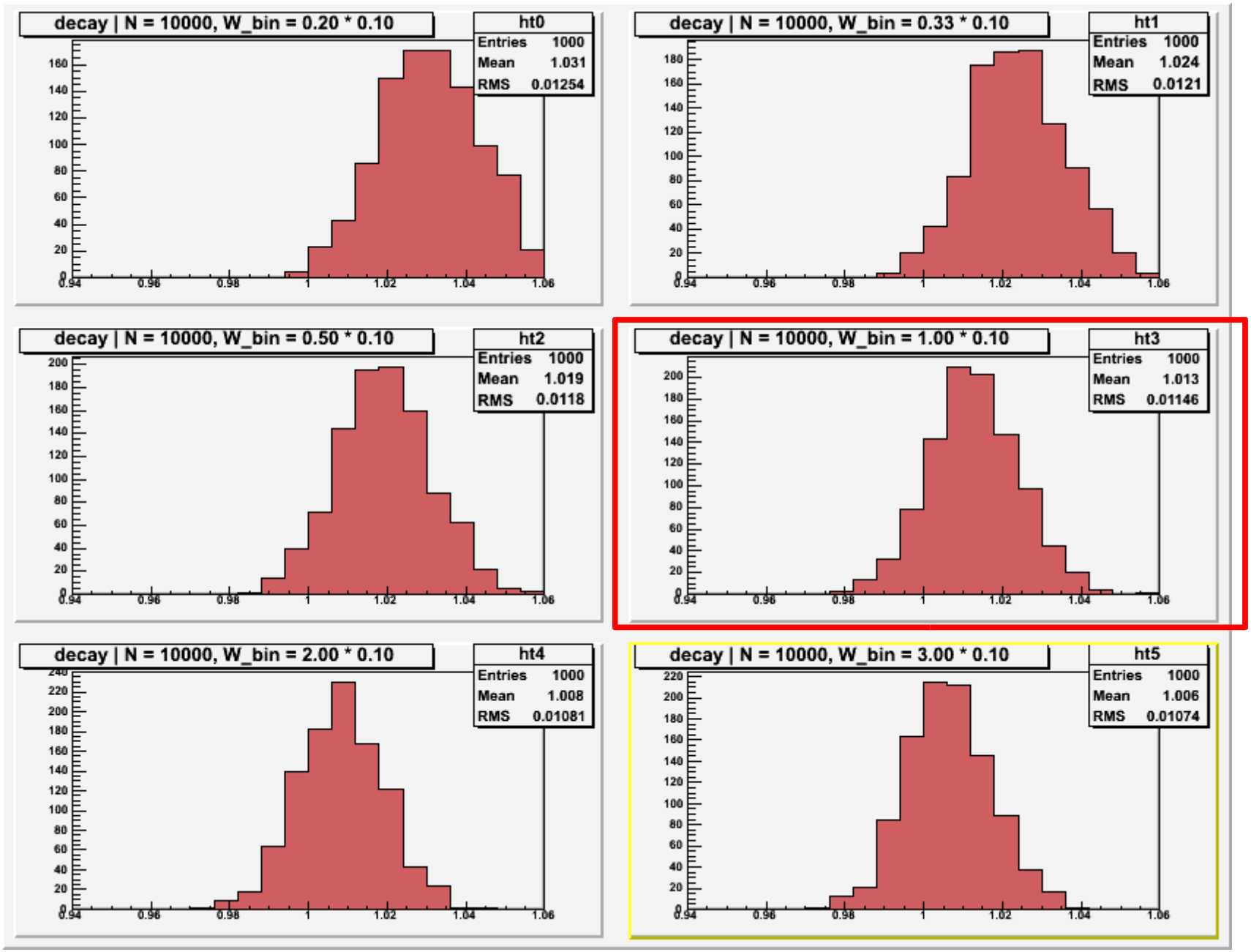


Decay const. (sample size N=10000)

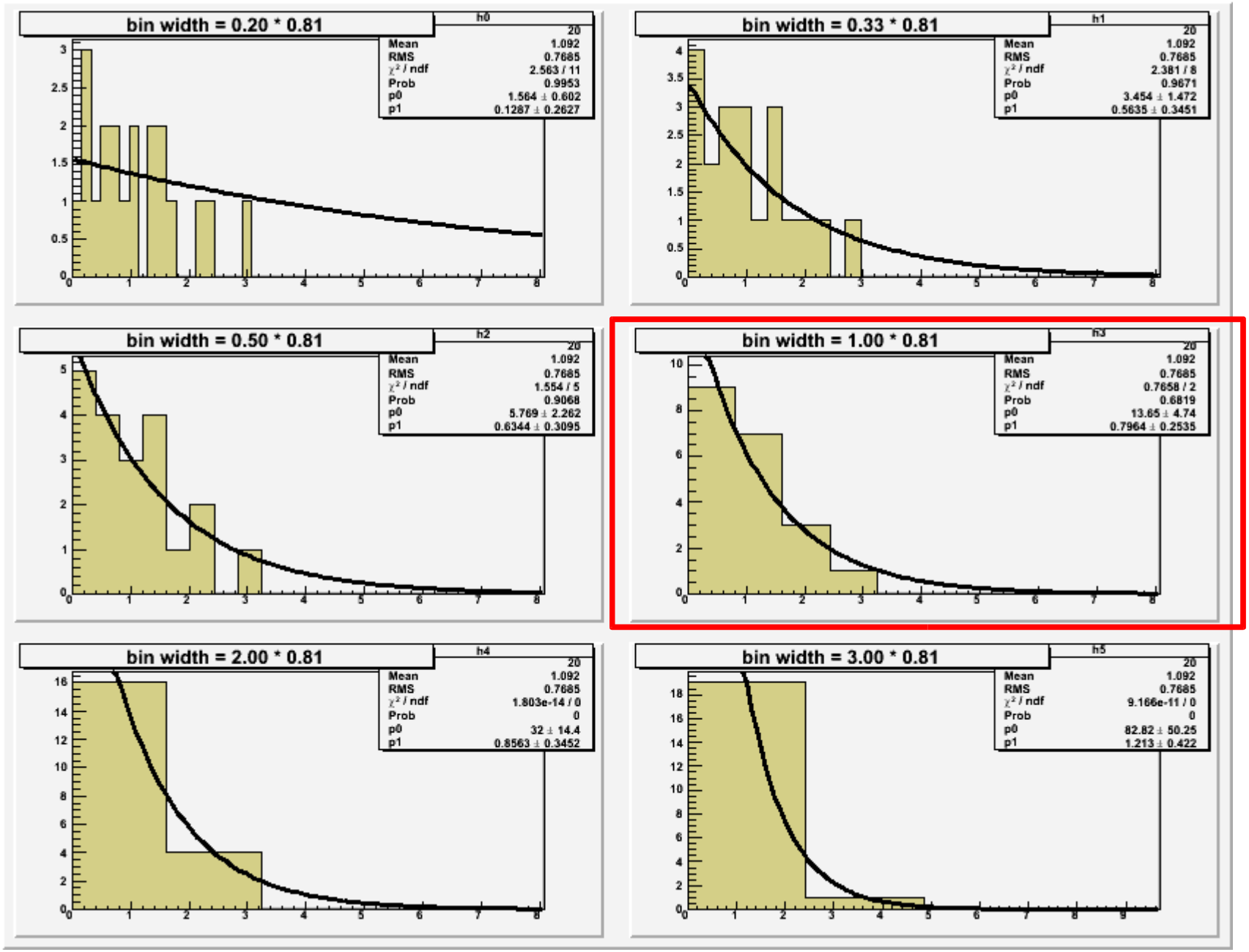


Decay const. (sample size N=10000)

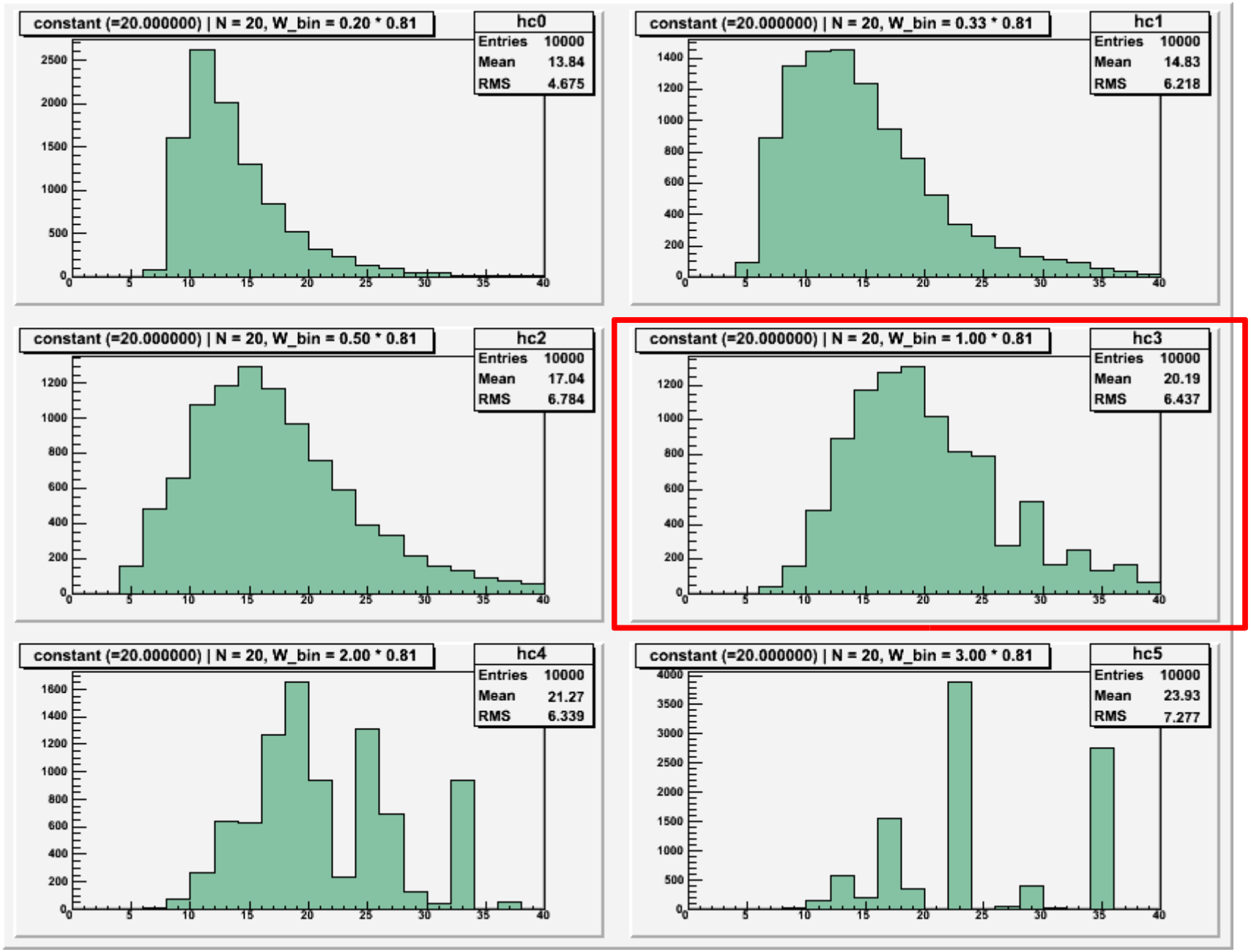
Fitted with option "I" (integral instead of value at bin center)



Sample size N=20

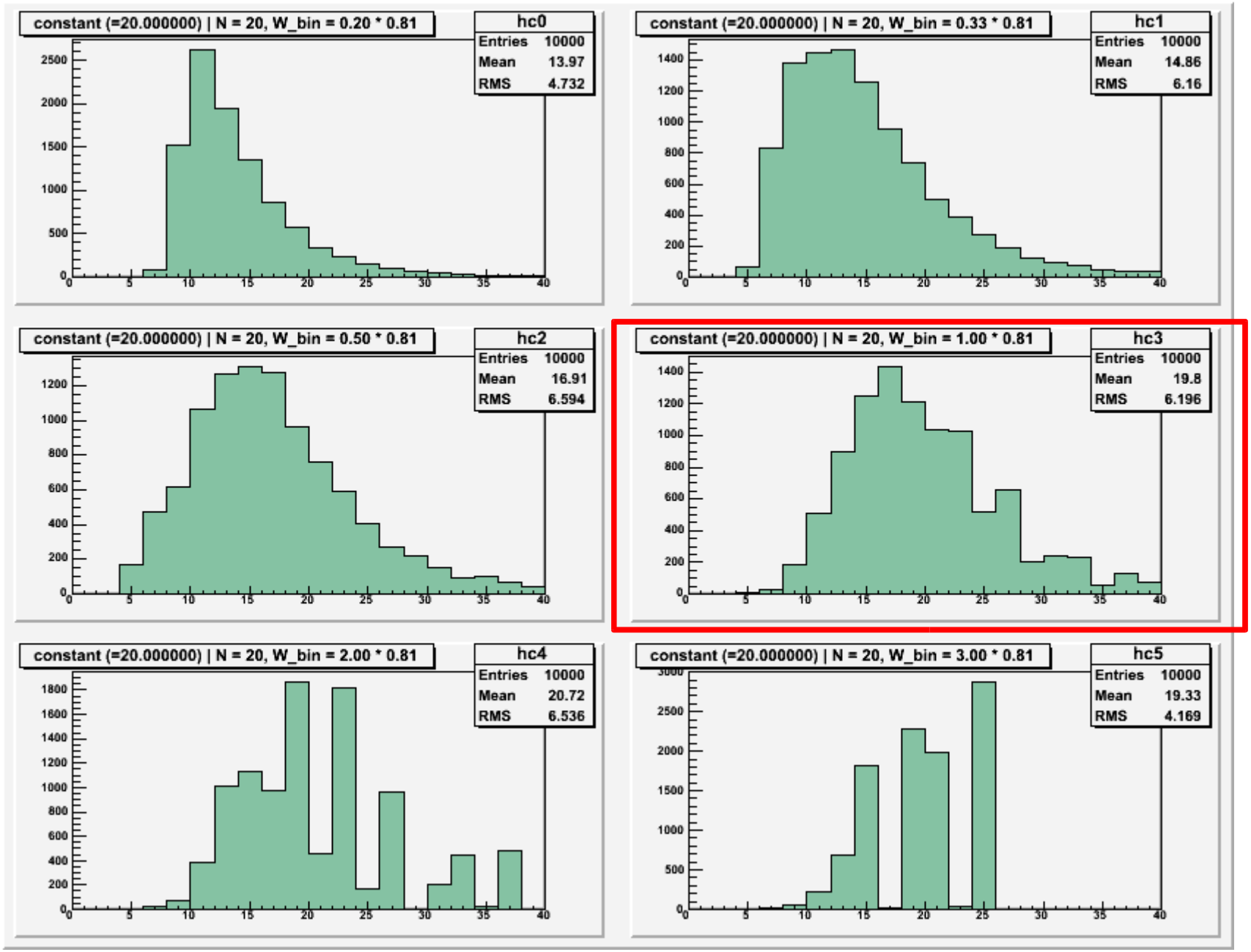


Norm. const. (sample size N=20)



Norm. const. (sample size N=20)

Fitted with option "I" (integral instead of value at bin center)



w/5

w/3

w/2

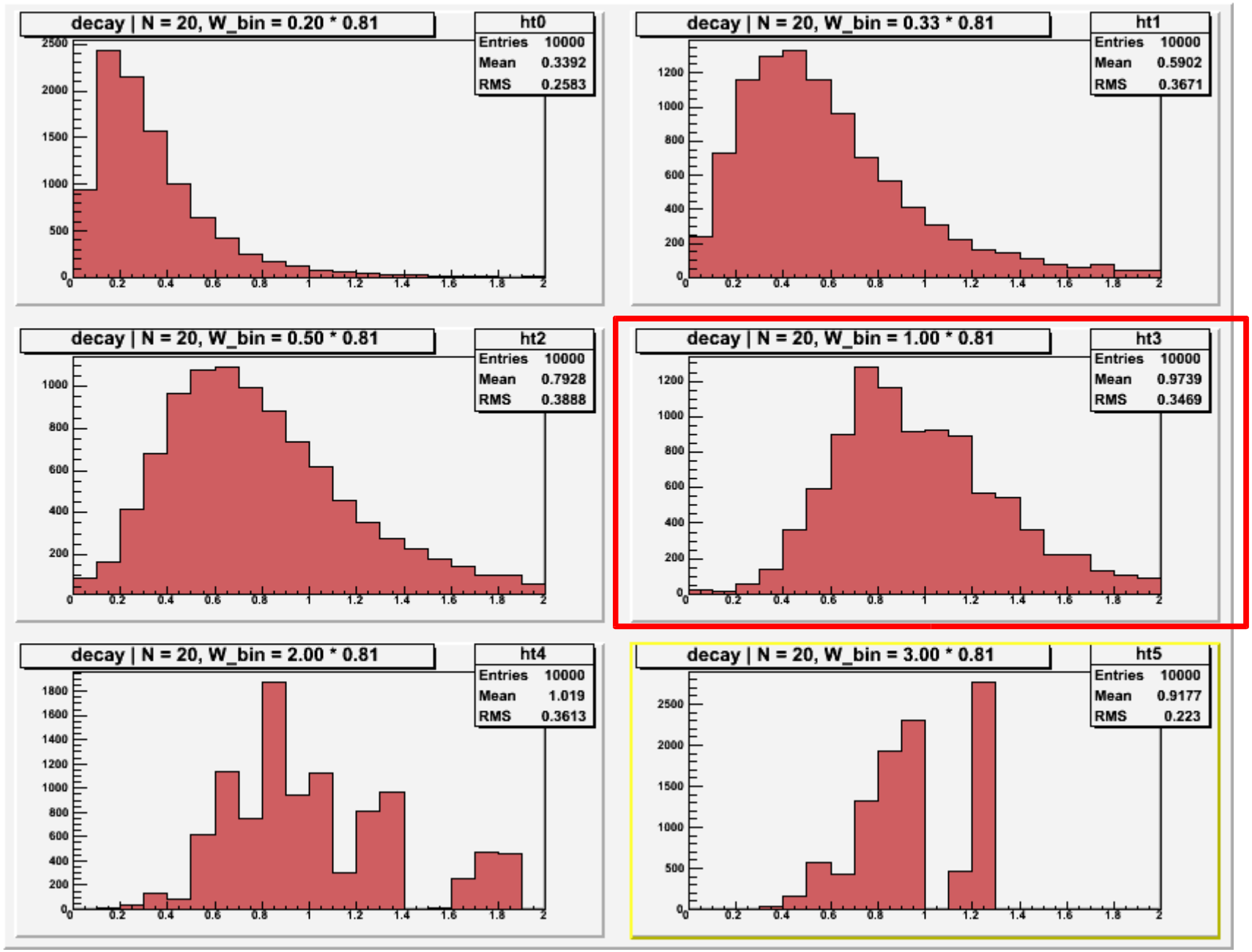
w

2w

3w

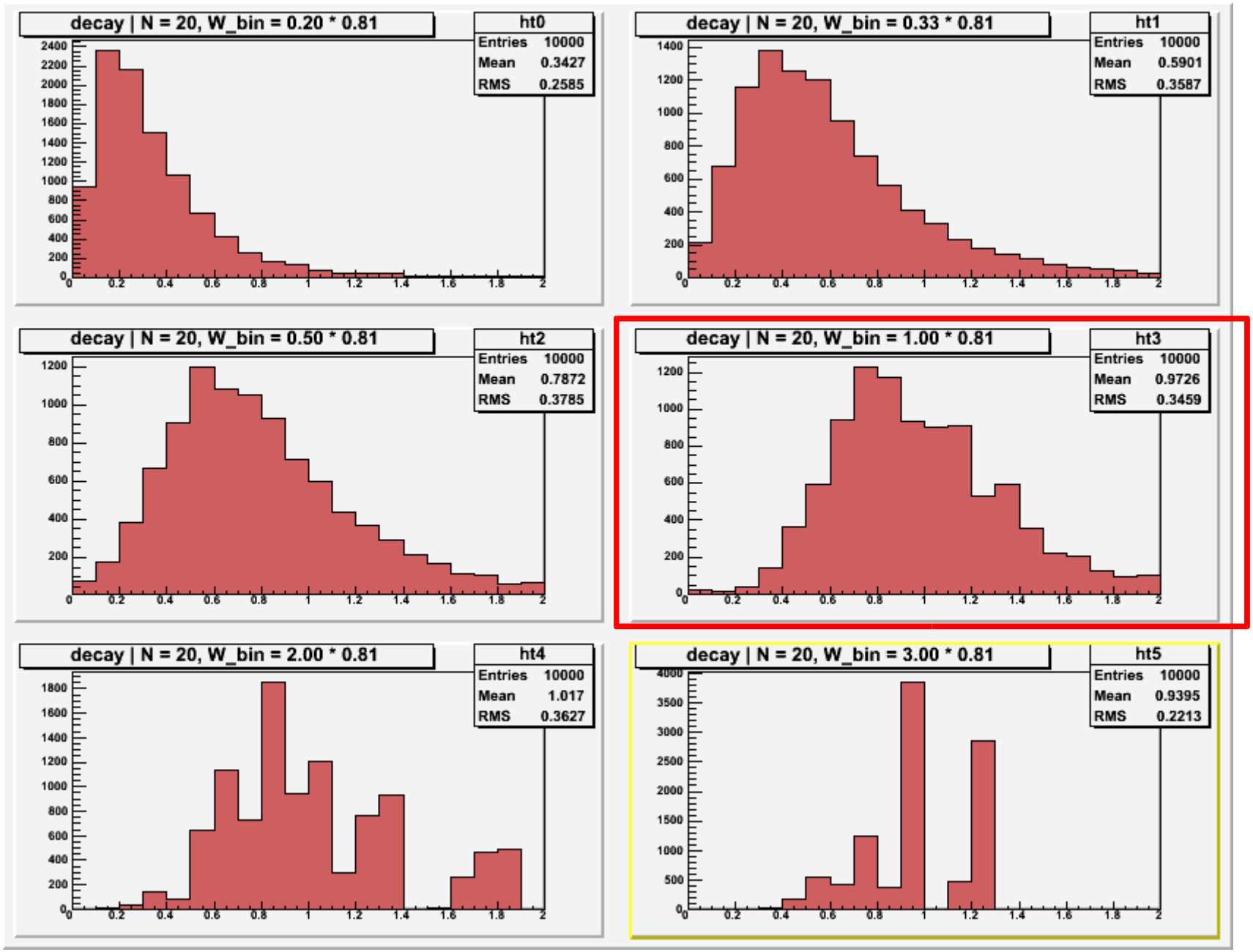
prescribed bin width

Decay const. (sample size N=20)



Decay const. (sample size N=20)

Fitted with option "I" (integral instead of value at bin center)



w/5

w/3

w/2

w

prescribed bin width

2w

3w

Conclusion

- The bin widths prescribed by the formulae do indeed seem to be the right choice.
- It matters most for low statistics.
- The option “I” of the fit doesn't seem to make much difference here.