Statistical Data Analysis, Assignment 3

Roel Deckers

Utrecht University

September 24, 2018

1   Objective

To test the *Central Limit Theorem (CLT)* as it applies to a Uniform distribution between 0 and 1, an exponential distribution $f(x) = \exp(-x)$, and the standard "Cauchy"[1] distribution.

2   Theory

The central limit theorem states that the mean of N samples drawn from a distribution D will converge to a normal distribution D_N with mean $E[D_N] = E[D]$ and variance $var(D_N) = var(D)/N$ for large N, *if* D has well defined moments.

The uniform distribution is well defined with mean 0.5 and a standard deviation of $1/12$. The exponential distribution is well defined with both a mean an a standard deviation of 1, and the Cauchy distribution has no well defined mean or standard deviation.

3   The Distributions

Using the ROOT framework we generated our three distributions and plotted them in figure 1. We let ROOT compute the mean and standard deviations of our binned histograms, as shown in the top right of each graph, and we also computed it over our raw data. The results of these computations are given in table 1.

Note how the uniform and exponential distribution are in close agreement (with minor differences attributable to the binning procedure) but the Cauchy distribution produces wildly different results. We believe this can be attributed to a combination of the ill-defined variance of the Cauchy Distribution and the binning procedure removing all outliers.

---

[1]Technically the ratio distribution but we follow the definitions established by the assignment.

Distribution	mean
normal	0.5 ± 0.3
exponential	1 ± 1
Cauchy	$0 \pm 2 \times 10^2$

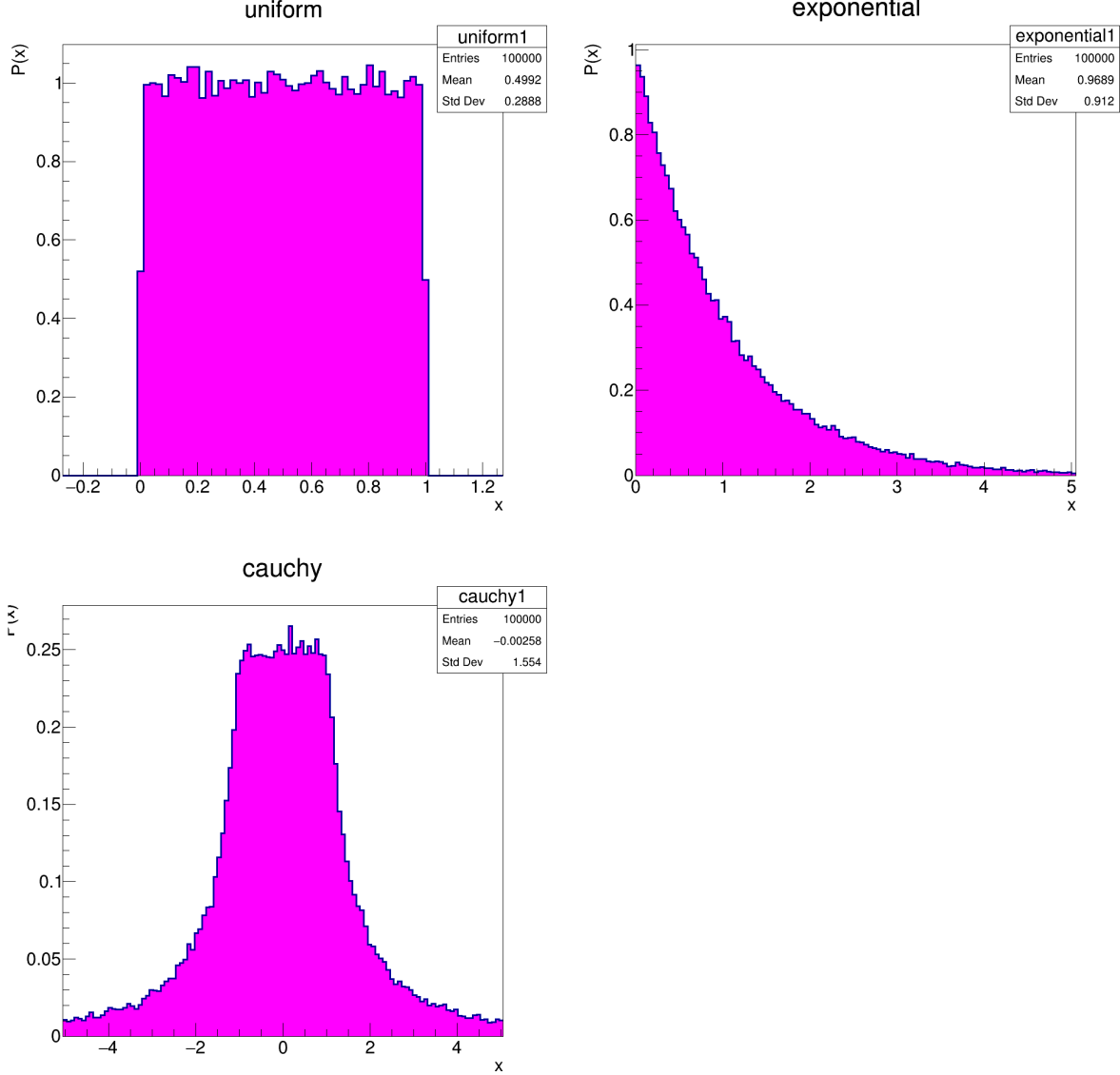Table 1: Mean and standard deviation for (unbinned) distributions with $N = 1$.

Figure 1: Our base distributions D.

4 Distributions for N samples

Subsequently we plotted the distributions of the mean of N samples from each distribution, for multiple N, in figure 2. We see that, as expected, the uniform and exponential distribution converge to a sharply peaked normal distribution as we average more samples. The Cauchy distribution however does not show much difference between $N = 10$ or $N = 100$ and if one looks closely does not resemble a normal distribution at all: the tail-ends are flared too much for a normal distribution. Out of the three, the uniform

distribution tends to a normal distribution the quickest. Probably due to it being already symmetrical.
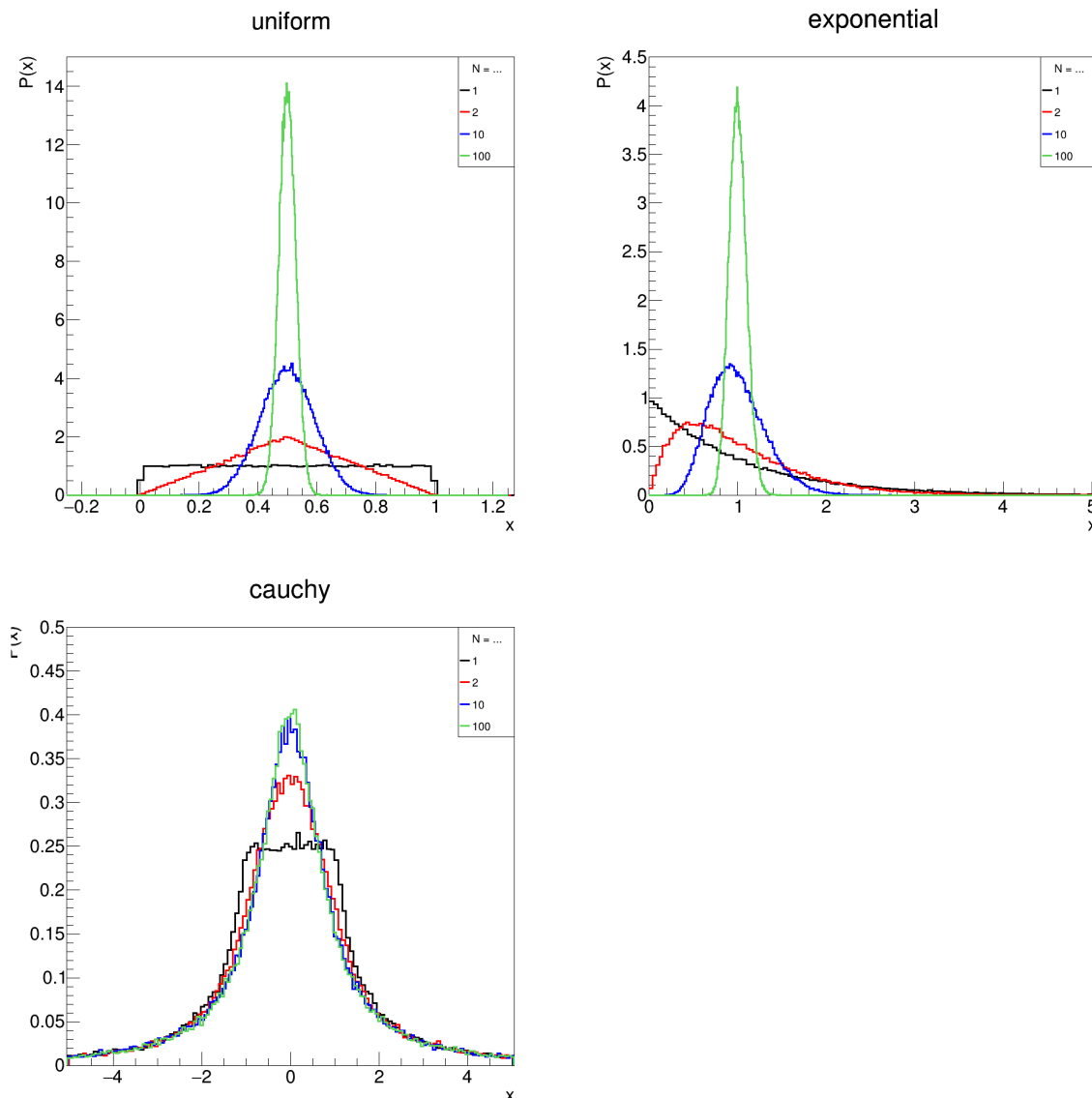

Figure 2: D_N for various D and N.

5   Mean and variance

The CLT predicts that as we increase N the mean would remain unchanged, therefore we plotted the mean of the histograms (as computed by ROOT) against N in figure 3. All of them seem to be stable, even the Cauchy distributions, this is reasonable because it is a symmetric distribution and any outliers due to ill-defined variance will be dealt with by the binning process. Furthermore, the CLT predicts that the standard deviation will go as $\sigma_{D_N} = \sigma_D/\sqrt{N}$, therefore the ratio $\sqrt{N}\sigma_{D_N}/\sigma_D$ should be equal to one if the

CLT holds. We have plotted this ratio in figure 4 using the results in table 1 for $\sigma_D{}^2$.

We see that the CLT does not hold for the Cauchy Distribution if we look at the variance, but it does hold for the uniform and exponential distribution.
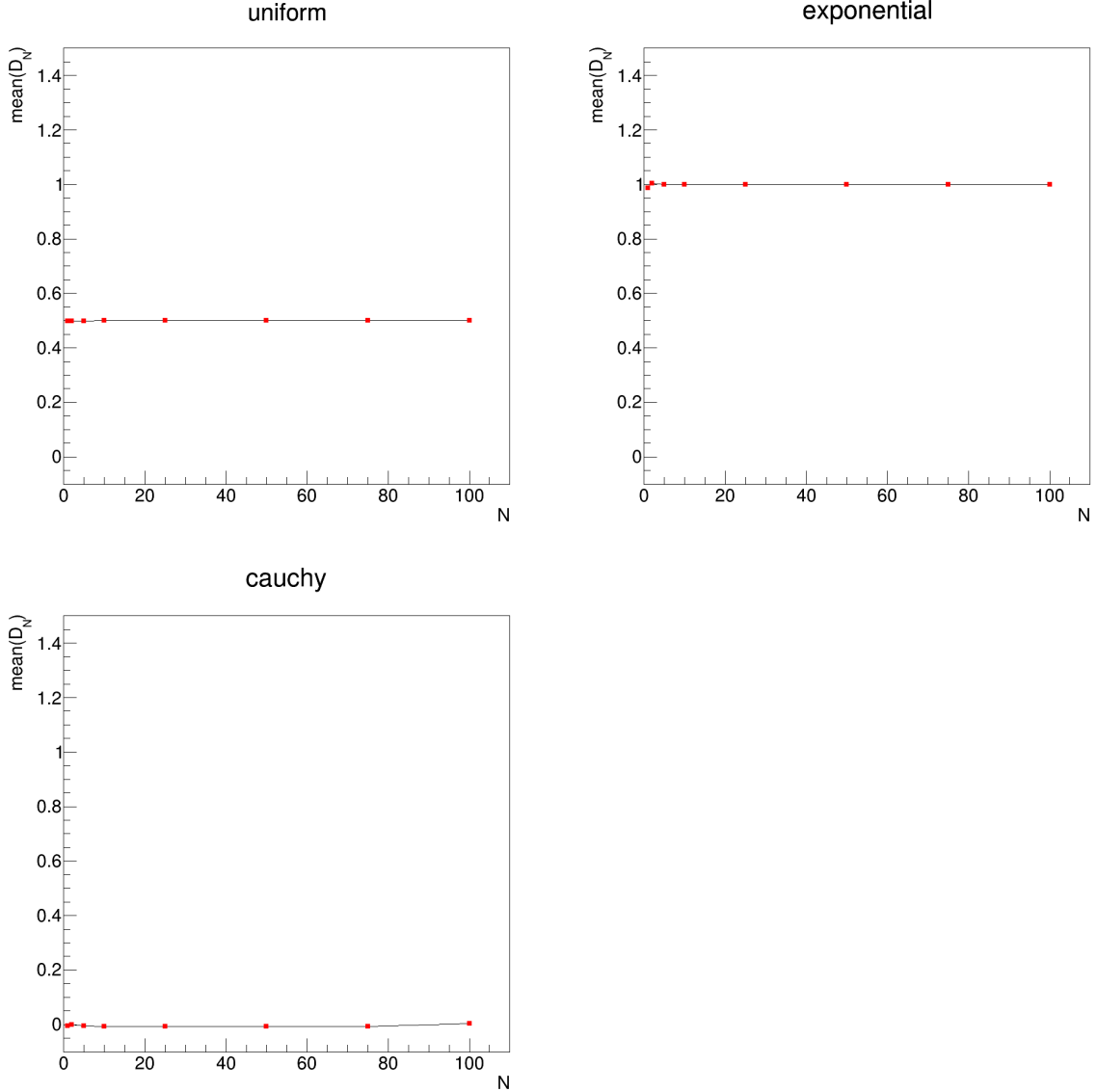

Figure 3: Means of D_N for various D as a function of N.

6 Conclusion

Unsurprisingly, the CLT holds. For uniform and exponential distributions it correctly predicts the results. For the Cauchy distribution it does not make any (correct) predictions, but this is in line with our expectations: the Cauchy distribution does not have well defined moments.

---

[2]using the histogram computed value for σ_D produces a small and fixed relative offset for uniform and exponential, but still produces wildly wrong results for Cauchy.
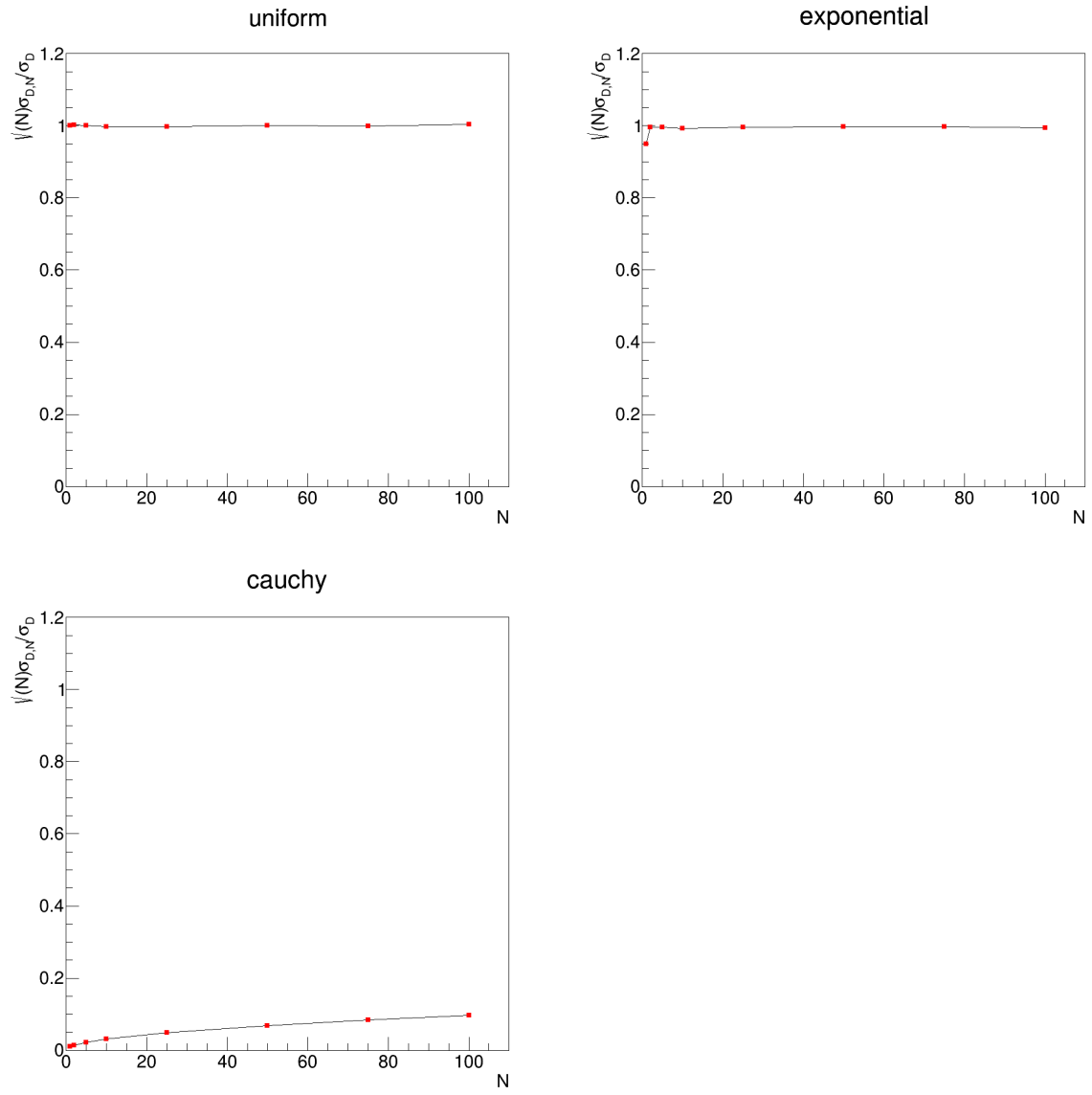
Figure 4: $\sigma_{D_N} = \sigma_D/\sqrt{N}$ for various D as function of N.