Finite Size Corrections for EVS (i.i.d. variables)

I. INTRODUCTION

So far, we have considered only the limiting (batch size goes to infinity, $N \to \infty$) behavior. In reality, N is finite. For example, when the EVS of the brightness of galaxies are investigated, the number of galaxies in the Hercules galaxy cluster (see Fig.4. in Lecture 5-6) is $N \approx 100$. Thus we should consider corrections to the limit distributions. Up to the first correction we should seek the EVS large-N distribution for the maximum value $P(x, N)$ in the form

$$P(x, N) = P(x) + q(N)\Phi_1(x) \tag{1}$$

where $P(x)$ is the limit distribution, $\Phi_1(x)$ is the 1^{st} shape correction, and $q(N) \to 0$ for $N \to \infty$.

The convergence to the limit distribution may be fast or slow as shown in Figs.1-4 for exponential and gaussian parents.

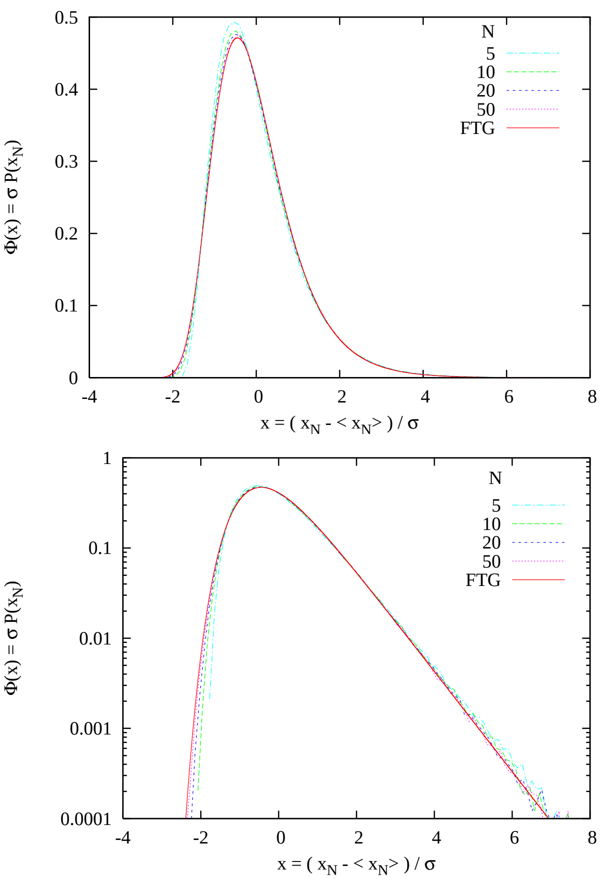

FIG. 1: Distribution of the largest for an exponential $[p(y) = e^{-x}]$ parent for various batch sizes, N. The limit distribution (Fisher-Tippett-Gumbel) is shown by red line. Note the fast convergence.

The lower panels in Figs. 2 and 4 suggests that

$$q(N) \sim \frac{1}{N} \quad \text{and} \quad q(N) \sim \frac{1}{\ln N} \tag{2}$$

for exponential and gaussian parents, respectively. One can also see on those figures that well defined shape corrections $\Phi_1(x)$ with recognizable features appear for large N.

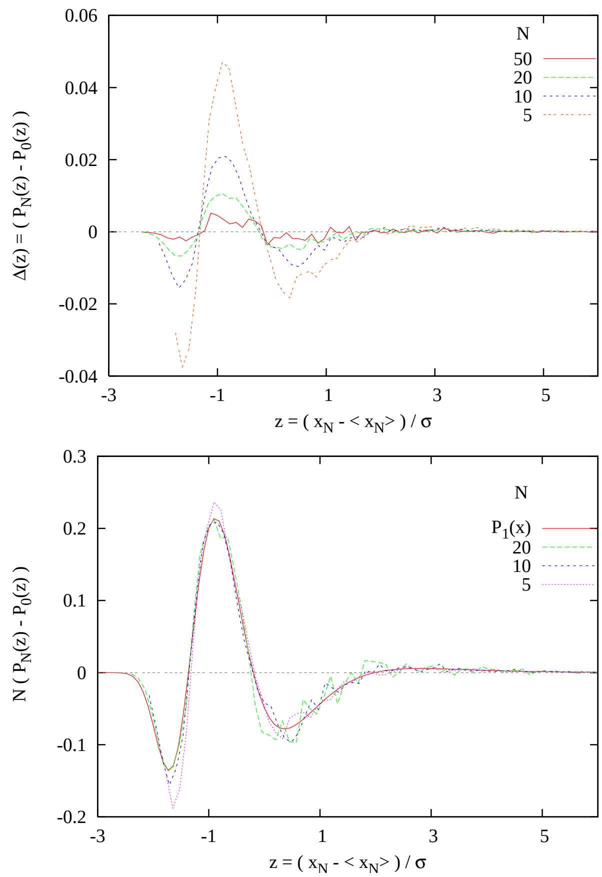

FIG. 2: Corrections to the limit distribution for an exponential $[p(y) = e^{-x}]$ parent for various batch size, N. Note the $1/N$ scaling of the shape correction.

In this lecture, we shall describe how to calculate directly $q(N)$ and $\Phi_1(x)$ for the case of exponential parent (the easiest case to calculate).

II. FINITE-SIZE CORRECTIONS IN CASE OF EXPONENTIAL PARENT

In case of an exponential parent $[p(y) = e^{-}y,\ y \geq 1]$, the integrated parent is given by

$$\mu(z) = 1 - e^{-z}\,. \tag{3}$$

Thus the integrated distribution for the largest out of a batch of size N is obtained as

$$M(z,N) = \mu^N(z) = \left(1 - e^{-z}\right)^N . \tag{4}$$

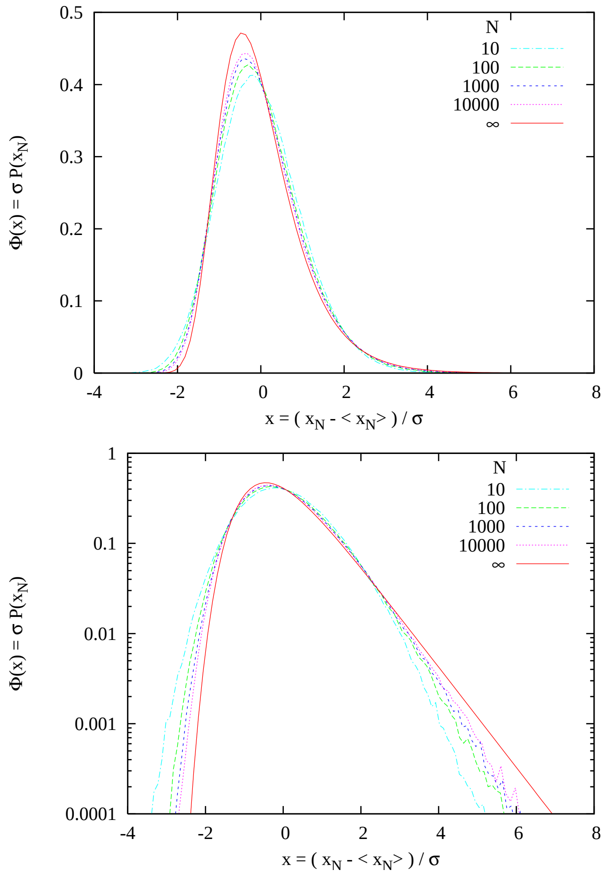

FIG. 3: Distribution of the largest for a gaussian $[p(y) \sim e^{-x^2}]$ parent for various batch sizes, N. The limit distribution (Fisher-Tippett-Gumbel) is shown by red line. Note the slow convergence.

Now, we have to make the appropriate change of variables

$$z = a_N x + b_n \tag{5}$$

and to obtain the final result in a form which is easy to compare with the experiments we shall use the standardization where

$$\langle x \rangle = 0 \quad , \quad \langle x^2 \rangle = 1 . \tag{6}$$

This standardization corresponds to plotting the experimental results against the variable $(x - \langle x \rangle)/\sqrt{\langle x^2 \rangle - \langle x \rangle^2}$.

With the above standardization, the coefficients a_N and b_N are determined as

$$\langle z \rangle = a_N \langle x \rangle + b_N = b_N \tag{7}$$

and

$$\langle z^2 \rangle - \langle z \rangle^2 = a_N^2 \langle x^2 \rangle = a_N^2 . \tag{8}$$

Thus we have to calculate a_N and b_N by evaluating $\langle z^2 \rangle$ and $\langle z \rangle$. Since $P(z,N)$ is given by

$$P(z,N) = Ne^{-z}(1 - e^{-z})^{N-1} \tag{9}$$

we have to calculate the following integrals

$$b_N = N \int_0^\infty z e^{-z}(1 - e^{-z})^{N-1} dz \tag{10}$$

$$a_N^2 = N \int_0^\infty z^2 e^{-z}(1 - e^{-z})^{N-1} dz . \tag{11}$$

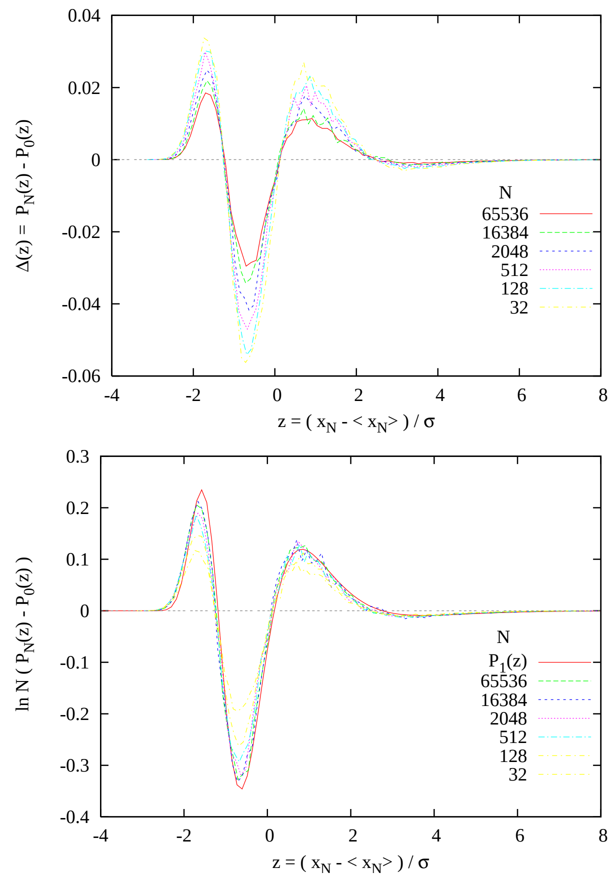

FIG. 4: Corrections to the limit distribution for a gaussian $[p(y) \sim e^{-x^2}]$ parent. Note the $1/\ln N$ scaling of the shape correction.

We shall carry out in some detail the calculation of b_N. From our previous experience with exponential parent, we know that the EVS distribution shifts with N as $\ln N$. Thus a natural change of variable to evaluate the integral is

$$z = u + \ln N \tag{12}$$

which allows to write (10) as

$$b_N = \ln N + \int_{-\ln N}^{\infty} u e^{-u} e^{(N-1)\ln(1-e^{-u}/N)} du\,. \quad (13)$$

Expanding the exponential, to order $1/N$, one finds

$$b_N = \ln N + \int_{-\ln N}^{\infty} u e^{-u-e^{-u}} du \left[1 + \frac{2e^{-u} - e^{-2u}}{2N}\right]. \quad (14)$$

As one can easily see, the lower limit of the integral can be moved to $-\infty$ since there the function to be integrated is of the order of $\exp(\ln N - \exp \ln N) \sim \exp(-N)$. Then the needed integrals can be evaluated (or looked up in integral tables):

$$\int_{-\infty}^{\infty} u e^{-u-e^{-u}} du = \gamma_E \quad (15)$$

$$\frac{1}{N}\int_{-\infty}^{\infty} u e^{-2u-e^{-u}} du = \frac{1}{N}(\gamma_E - 1) \quad (16)$$

$$-\frac{1}{2N}\int_{-\infty}^{\infty} u e^{-3u-e^{-u}} du = -\frac{1}{N}(\gamma_E - 3/2) \quad (17)$$

where $\gamma_E \approx 0.577$ is the Euler constant.

Collecting now the terms in (14), we obtain

$$b_N = \ln N + \gamma_E + \frac{1}{2N} \quad (18)$$

The calculation of a_N^2 is of similar difficulty, and all the integrals can be calculated by hand or can be found in the integral tables. The result for a_N^2 is

$$a_N^2 = \langle z^2 \rangle - \langle z \rangle^2 = \frac{\pi^2}{6} - \frac{1}{N}\,, \quad (19)$$

and so the final result to order $1/N$ for a_N is as follows

$$a_N = \frac{\pi}{\sqrt{6}} - \frac{1}{2N}\frac{\sqrt{6}}{\pi}\,, \quad (20)$$

Now we take the expression (9) for $P(z,N)$ and and make the change of variables $z = a_N x + b_N$

$$\begin{aligned} P(x,N) &= a_N P(z,N) \\ &= a_N N e^{-a_N x - b_N}(1 - e^{-a_N x - b_N})^{N-1} \quad (21) \end{aligned}$$

and expand everything to order $1/N$. This is a bit tedious calculation but, with some patience, can be done. We write out only the final result:

$$P(x,N) = P(x) + \frac{1}{N}\Phi_1(x)\,, \quad (22)$$

where $P(x)$ is the Fisher-Tippett-Gumbel distribution with the $\langle x \rangle = 0$ and $\langle x^2 \rangle = 0$ standardization

$$P(x) = \frac{\pi}{\sqrt{6}} \exp\left[-\frac{\pi}{\sqrt{6}}x - \gamma_E - e^{-\frac{\pi}{\sqrt{6}}x - \gamma_E}\right] \quad (23)$$

and the shape correction is given by

$$\begin{aligned} \Phi_{exp}(x) &= \frac{1}{2}P(x)\left[-\frac{6}{\pi^2} - 1 + \frac{\sqrt{6}}{\pi}x \right. \quad (24) \\ &\left. + \left(3 - \frac{\sqrt{6}}{\pi}x\right) e^{-\frac{\pi}{\sqrt{6}}x - \gamma_E} - e^{-\frac{2\pi}{\sqrt{6}}x - 2\gamma_E}\right]. \end{aligned}$$

As we can see, the finite-size correction for an exponential parent is small for relatively not too large N, since the convergence rate is $1/N$. Fig.2 shows this convergence as well as the shape correction (red line, denoted by $P_1(x)$).

III. FINITE-SIZE CORRECTION FOR GAUSSIAN PARENT

The calculation is somewhat more tedious as compared to the exponential case. The convergence is slow, we have

$$P(x,N) = P(x) + \frac{1}{\ln N}\Phi_G(x)\,, \quad (25)$$

where $P(x)$ is the same limit distribution as before (23), and the Gauss shape correction is given by

$$\begin{aligned} \Phi_G(x) &= \frac{1}{2}P(x)\left[\frac{\sqrt{6}}{\pi}x - \frac{6}{\pi^2}\zeta(3) + \right. \quad (26) \\ &\left. \left(-\frac{\pi^2}{12}x^2 + \frac{\sqrt{6}}{\pi}\zeta(3)x + \frac{\pi^2}{12}\right)\left(1 - e^{-\frac{\pi}{\sqrt{6}}x - \gamma_E}\right)\right]. \end{aligned}$$

where ζ is the zeta function and $\zeta(3) \approx 1.20$.

The slow convergence and the shape correction can be seen in Fig.4.

IV. CORRECTIONS IN GENERAL

The general case and the questions related to the convergence rates can be treated through renormalization group methods. In references [1, 2], one can find a detailed exposition and, in case of doubts, you can consult the authors.

Interesting and physically relevant cases we should mention are the "modified gaussian" parents

$$p(y) \sim \frac{e^{-y^\beta}}{y^\alpha} \quad , \quad y > 0\,. \quad (27)$$

Unless $\beta = 1$ and $\alpha = 0$, the correction is proportional to $\Phi_G(x)$. The convergence, however, depends on the above exponents. For example, (25) and (27) remains valid for $\beta = 2$ and $\alpha = 1$ but we have

$$P(x, N) = P(x) + \frac{1}{\ln^2 N}\Phi_G(x) \quad \text{for} \quad \alpha = 1\,, \beta = 1\,. \tag{28}$$

Fig.5 displays the results of EVS for galaxy luminosities [3]. It shows an example where taking into account the finite-size corrections was important in understanding the deviations from the expected FTG limit distribution. In this case there was another correction due to variable batch size. The variable batch size is an interesting problem of EVS but we do not have time to discuss it in the present course.

[1] G. Györgyi, N. R. Moloney, K. Ozogány, and Z. Rácz, Phys. Rev. Lett. **100**, 210601 (2008): *Finite Size Scaling in Extreme Statistics.*

[2] G. Györgyi, N. R. Moloney, K. Ozogány, Z. Rácz, and M. Droz, Phys. Rev. E **81**, 041135 (2010): *Renormalization-group theory for finite-size scaling in extreme statistics.*

[3] M. Taghizadeh-Popp, K. Ozogány, Z. Rácz, E. Regoes, A. S. Szalay, The Astrophysical Journal, 100:759 (2012): *Distribution of Maximal Luminosity of Galaxies in the Sloan Digital Sky Survey*

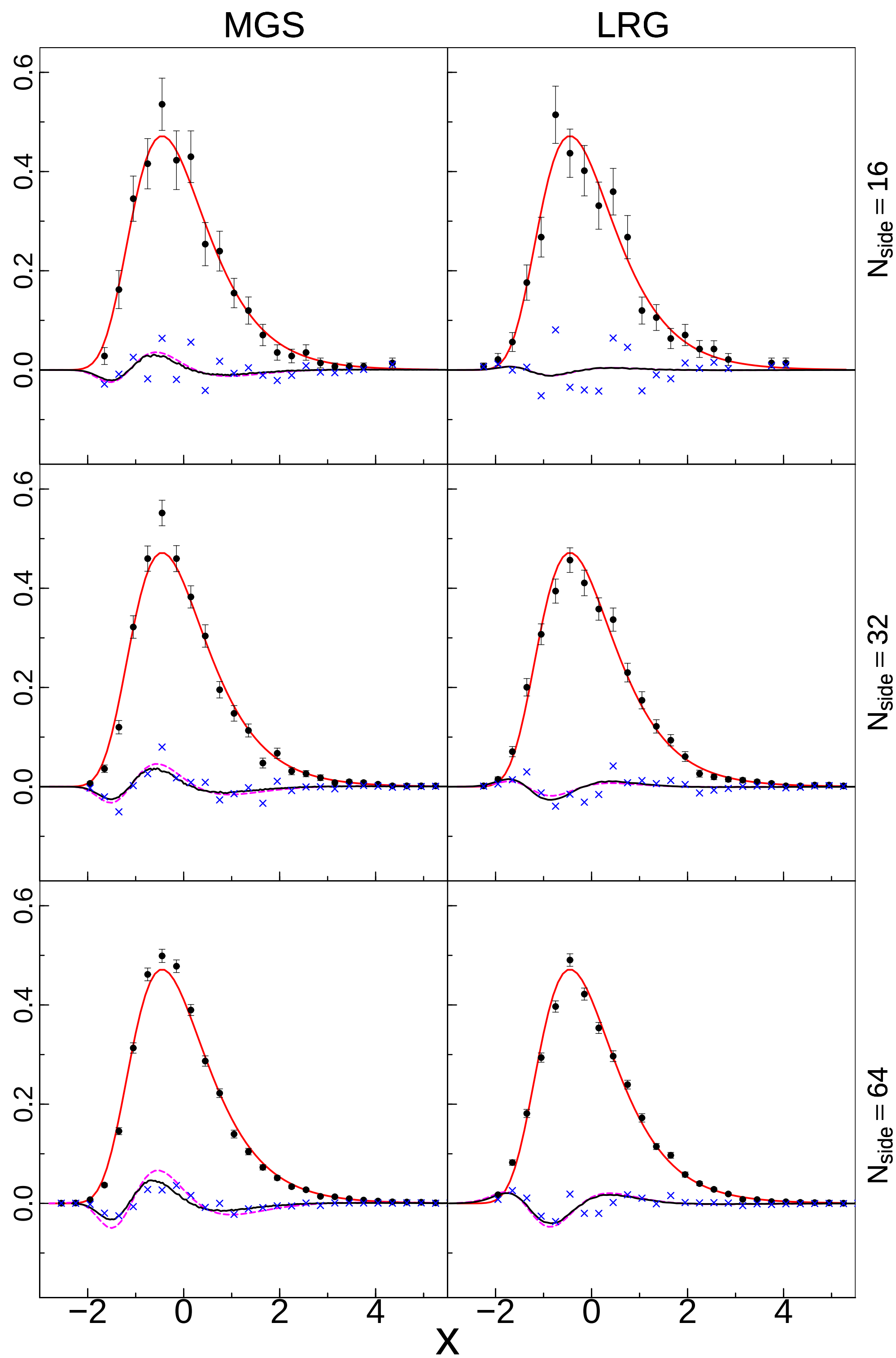

FIG. 5: The normalized maximum luminosity histograms (black circles) for N_{side} = 16, 32, 64 (from up to down) for two galaxy samples compared to the limit distribution FTG (solid red line) in scaled variables ($\langle x \rangle = 0$ and $\sigma = 1$) while blue crosses are the residuals to the FTG. For the Main Galaxy Sample (MGS), the solid magenta curves show $q(N)P_1(x) + \overline{P}_1(x)$, i.e., the first order finite size correction for the Schechter parent added to the variable batch size correction. The Large Red Galaxies (LRG) curve is different, in the sense that the parent is FTG and the finite size corrections do not appear, having corrections only due to the variable batch size ($\overline{P}_1(x)$). The black solid curves are the simulations that result from using the experimentally given luminosity distributions and sample size distributions.