1 Approximating lower bounds on S_n.

(This first paragraph below is to help keep notation straight, but probably should be removed ultimately.)

We suppose as before that X_i are iid Pareto random variables with distribution function $F(x) = 1 - x^{-\alpha}$ for $x \geq 1$, that $S_n = \sum_{i=1}^{n} X_i$, and that $M_n = \max\{X_1, X_2, ..., X_n\}$.

An approximation for the qth quantile of S_n, for small q (e.g. $q = 2.5\%$), may be constructed based on the following idea. One seeks c_n, where

$$P(S_n < c_n) \approx q. \tag{1}$$

For any value y_n,

$$\begin{aligned} P(S_n < c_n) &= P(S_n < c_n | M_n \leq y_n) P(M_n \leq y_n) \\ &\quad + P(S_n < c_n | M_n > y_n) P(M_n > y_n). \end{aligned} \tag{2}$$

The term $P(S_n < c_n | M_n \leq y_n)$ in (**??**) can be approximated using the central limit theorem, since the variables being summed are now truncated and hence have finite moments. The resulting approximation will be close for sufficiently small values of y_n and sufficiently large n. At the same time, for any reasonably large value of y_n, and for c_n satisfying (**??**) for small q, the quantity $P(S_n < c_n | M_n > y_n)$ in (**??**) will be infinitessimal; hence the entire final term in (**??**) may be considered negligeable. Thus, in approximating c_n, we suggest choosing an appropriate value of y_n, and considering the approximation

$$\begin{aligned} P(S_n < c_n) &\approx P(S_n < c_n | M_n \leq y_n) P(M_n \leq y_n) && (3) \\ &\approx \Phi\{(c_n - n\mu_{y_n})/(\sigma_{y_n}\sqrt{n})\} P(M_n \leq y_n), && (4) \end{aligned}$$

where Φ is the standard normal distribution function, and

$$\begin{aligned} \mu_y &:= E[X_1 | X_1 \leq y] = \frac{\alpha y^{1-\alpha} - \alpha}{(1-\alpha)(1-y^{-\alpha})}, && \alpha \neq 1, && (5) \\ &= \frac{ln(y)}{1-y^{-1}}, && \alpha = 1, \\ \sigma_y^2 &:= V[X_1 | X_1 \leq y] = \frac{\alpha y^{2-\alpha} - \alpha}{(2-\alpha)(1-y^{-\alpha})} - \mu_y^2, && \alpha \neq 2, && (6) \\ &= \frac{2ln(y)}{1-y^{-2}} - \mu_y^2, && \alpha = 2. \end{aligned}$$

Note that there is a tradeoff in choosing y_n in (**??**): if one selects too small a value of y_n, then the term $P(S_n < c_n | M_n > y_n)$ in (**??**) is not negligeable, so the resulting approximation may not be satisfactory. On the other hand, if y_n is too large, then the approximation of $P(S_n < c_n | M_n < y_n)$ using the central limit theorem may be unsatisfactory; this is particularly true for small n.

One option is to choose some value p^* to represent the probability $P(S_n < c_n | M_n \leq y_n)$ in (**??**). From (**??**), one then has $q/p^* = P(M_n \leq$

$y_n) = \{1 - (\frac{1}{y_n})^\alpha\}^n$, and solving this for y_n one obtains

$$y_n = \{1 - (q/p^*)^{1/n}\}^{-1/\alpha}. \tag{7}$$

One may then obtain an approximation of c_n by plugging this value of y_n from (**??**) into the first term in (**??**), yielding

$$\hat{c}_n \quad = \quad \sigma_{y_n}\sqrt{n}\Phi^{-1}(p^*) + n\mu_{y_n}, \tag{8}$$

where μ_{y_n} and σ_{y_n} are given by equations (**??**-**??**).

A naive choice for p^* is $\sqrt{q}$; this seems to balance the aforementioned tradeoff, since then $P(S_n < c_n | M_n \leq y_n) = P(M_n \leq y_n) = \sqrt{q}$.

The values reported in column 3 of Table 1 reflect the approximation $\hat{c}_n$ in (**??**), with $p^* = \sqrt{q}$ and y_n, μ_{y_n}, and σ_{y_n} given by (**??**-**??**). However, plots of ideal choices of p^* versus q, n, and α suggest that instead, the choice

$$p^* = 0.136 + 0.235q + q^2 + 0.0066\, min(n, 10) - 0.05\, max(a, 1) \tag{9}$$

provides a better approximation. Performance of the approximation $\tilde{c}_n$ resulting from the use of (**??**) with (**??**) for p^* and y_n, μ_{y_n}, and σ_{y_n} given by (**??**-**??**) is shown in column 4 of Table 1.

Although the slightly simpler $\hat{c}_n$ approximates the lower quantile quite well, in most cases there is substantial improvement from the use of the approximation $\tilde{c}_n$ which employs (**??**). This is also displayed graphically in

Figures 1 and 2 below. Figure 1 shows how the 0.02 quantile c_n and the two approximations $\hat{c}_n$ and $\tilde{c}_n$ vary with n, for $\alpha = 2/3$. One sees that the two approximations match the true quantile quite closely. Figure 2 highlights the error rates for the two approximations $\hat{c}_n$ and $\tilde{c}_n$ as a function of n, again for $\alpha = 2/3$. Table 1 and Figures 1 and 2 show the error in the approximations for $q = 0.02$, but results for other small values of q are rather similar.

Figure Captions:

Figure 1: Quantile c_n (solid curve), along with approximations $\hat{c}_n$ (dotted) and $\tilde{c}_n$ (dashed) as functions of n, for $\alpha = 2/3$ and $q = 0.02$. For each n, the values of c_n shown is the empirical 0.02 quantile from 10 million simulations of S_n.

Figure 2: Error rates for the two approximations $\hat{c}_n$ (dotted) and $\tilde{c}_n$ (dashed) as functions of n, for $\alpha = 2/3$ and $q = 0.02$. Percentage error rates are calculated as $100(\hat{c}_n - c_n)/c_n$ and $100(\tilde{c}_n - c_n)/c_n$, where each value of c_n is the empirical 0.02 quantile taken from 10 million simulations of S_n.

The table below is for $q = 0.02$. Error rates are rounded to 4 decimal places.

α	n	$(\hat{c}_n - c_n)/c_n$	$(\tilde{c}_n - c_n)/c_n$
0.5	2	0.0157	-0.0023
0.5	5	0.0120	0.0029
0.5	10	0.0079	0.0041
0.5	20	0.0052	-0.0018
0.5	50	0.0040	-0.0080
0.5	100	0.0035	-0.0021
0.67	2	0.0019	-0.0005
0.67	5	0.0012	-0.0030
0.67	10	-0.0050	0.0045
0.67	20	-0.0019	0.0027
0.67	50	-0.0010	0.0013
0.67	100	0.0005	-0.0004
1	2	-0.0110	0.0007
1	5	-0.0056	0.0000
1	10	-0.0006	0.0023
1	20	0.0019	0.0007
1	50	-0.0149	-0.0015
1	100	-0.0076	-0.0014
1.5	2	0.0003	0.0035
1.5	5	0.0045	0.0032
1.5	10	-0.0182	-0.0043
1.5	20	-0.0085	-0.0022
1.5	50	0.0017	0.0049
1.5	100	0.0059	0.0046