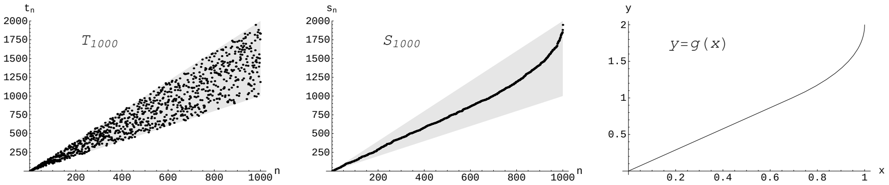
Sorting between lines

Proposed by Rick Mabry and Debbie Shepherd, LSUS, Shreveport, LA, rmabry@lsus.edu, dshepher@lsus.edu.

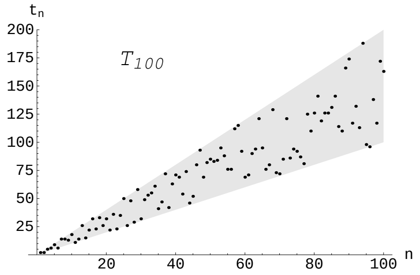
Let N be a positive integer and define the sequence $T_N = \{t_n : n = 1, 2, \ldots, N\}$, where each t_n is chosen at random (uniformly distributed) in the interval $[n, 2n]$. Let $S_N = \{s_n : n = 1, 2, \ldots, N\}$ denote the sequence of sorted values of T_N (in increasing order). For large N, when the points $\{(n, s_n) : n = 1, 2, \ldots, N\}$ are plotted in the plane, a smooth curve seems to emerge. Prove that this is indeed the case by showing that the set of normalized points

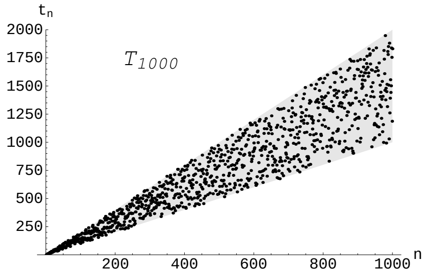
$$\left\{\left(\frac{n}{N}, \frac{s_n}{N}\right) : n = 1, 2, \ldots, N\right\}$$

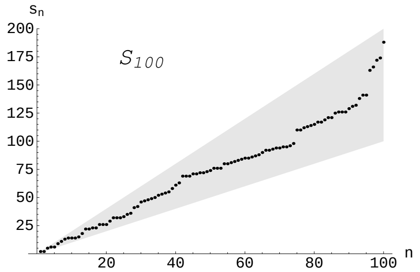
converges to points of the graph $y = g(x)$ of a smooth function $g : [0, 1] \to [0, 2]$.

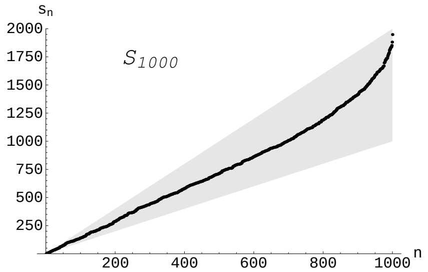


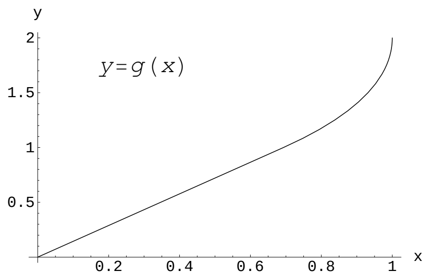
[If space were not an issue, perhaps the following pictures would be nicer.]











Dear Editor,

Should this problem be printed, we feel certain that many solvers will obtain the correct curve, but it will be interesting to see by which means they do so. This being a problem in probability, there are lurking issues of interpretation and rigor. In this problem, each of the random selections is chosen according to a *distinct* distribution function. To put it another way, each sample is of *size one*, so it is a somewhat delicate matter to see, or rather, to *prove* what happens when the number of points becomes large. We attempt a "rigorous" approach in *Solution #3.* Clearly, this is not an issue when the samples are identically distributed, for then the usual *frequentist interpretation* makes easier sense. Seeing how (or if) the solvers choose to deal with this matter should be fun. Perhaps there will be a solution involving order statistics, as well.

An *ad hoc* solution to the problem as stated is given in *Solution #1.* A fairly general approach is taken in *Solution #2.* In the case of identically distributed random variables, the solution to the analogous problem is rather well-known — it is the inverse of the common distribution function, to which our solution reduces in that situation. In the case here, of different distributions depending continuously on a parameter, we expected to find a similar result in the literature but were unable to track one down, perhaps because we do not know the terminology.

Actually, the "rigor" aspired to in *Solution #3* is more of the form of an obsessive phobia of one of us, an analyst, who is uncomfortable unless an actual limit (of the number of trials) is involved. The other, a statistician, is completely comfortable with *Solution #2*!

We hope this does not seem excessive for a *Monthly problem.* We considered submitting a short note instead but felt that would cheat the readers out of a nice exercise. (Besides, then our own rigor would have to hold up under closer scrutiny.)

R. Mabry & D. Shepherd

Solution #1. Fix a large positive integer N and consider the plot of T_N and the horizontal strip $(0, N] \times (k-1, k]$, where $k \leq 2N$ is a positive integer. The expected number of points in this strip is given by $\lambda(k) = \sum_{n=1}^{N} \int_{k-1}^{k} f_n(x)\, dx$, where $f_n(x)$ is the density function of the random variable corresponding to t_n. In our case we have

$$\lambda(k) = \begin{cases} \displaystyle\sum_{n=\lceil k/2 \rceil}^{k-1} \frac{1}{n} & \text{if} \quad 1 \leq k \leq N \\ \displaystyle\sum_{n=\lceil k/2 \rceil}^{N} \frac{1}{n} & \text{if} \quad N < k \leq 2N. \end{cases}$$

For large N we may use the asymptotic result $\sum_{n=1}^{M} 1/n = \log(M) + \gamma + O(1/M)$ and we obtain

$$\lambda(k) = \begin{cases} \log 2 + O(1/k) & \text{if} \quad 1 \leq k \leq N \\ \log(2N/k) + O(1/k) & \text{if} \quad N < k \leq 2N. \end{cases}$$

Since this is also the expected number of points of S_N in that same horizontal strip, we can see that for (n, k) to correspond (via expected values) to a point in the plot of S_N we should have $n = s(k)$, where $s(k) = \sum_{j=1}^{k} \lambda(j)$. Now using the fact that $\sum_{j=1}^{k} \log(j) = \log(k!)$, we have

$$s(k) = \begin{cases} k \log 2 + O(\log N) & \text{if} \quad 1 \leq k \leq N \\ k \log 2 + (k - N)\log N - \log(k!/N!) + O(\log N) & \text{if} \quad N < k \leq 2N. \end{cases}$$

Invoking Stirling's formula in the form $\log(m!) = (m + \frac{1}{2}) \log m - m + \frac{1}{2} \log(2\pi) + O(1/m)$ gives

$$\frac{s(k)}{N} = \begin{cases} \left(\frac{k}{N}\right) \log 2 + O\left(\frac{\log N}{N}\right) & \text{if} \quad 1 \leq k \leq N \\ \left(\frac{k}{N}\right) \left(\log 2 + 1 - \log \frac{k}{N}\right) - 1 + O\left(\frac{\log N}{N}\right) & \text{if} \quad N < k \leq 2N. \end{cases}$$

Finally, let $(u_k, v_k) = \left(\frac{s(k)}{N}, \frac{k}{N}\right)$ and let $N \to \infty$. The set of points (u_k, v_k) will converge to points on the curve $\{(h(v), v) : 0 \leq v \leq 2\}$, where

$$h(v) = \begin{cases} v \ \log 2 & \text{if} \quad 0 \leq v \leq 1 \\ v(\log 2 + 1 - \log v) - 1 & \text{if} \quad 1 < v \leq 2. \end{cases}$$

It is easy to see that h is strictly increasing. Thus it is one-to-one and its inverse is then the function g requested in the statement of the problem.

Solution #2.

Let N be large and let $\hat{t}_k = t_k/N$ and $\hat{s}_k = s_k/N$. We may assume that each $\hat{t}_k$ is obtained as a sample from a random variable uniformly distributed in the interval $[k/N, 2k/N]$ and that the $\hat{s}_k$ are obtained by sorting the $\hat{t}_k$.

We place this situation in a slightly more general setting, which is actually simpler and more convenient. (In this version of the solution we attempt to be more general. In the process we should also become more rigorous, but we save that for *Solution #3*.) Divide the interval $[0,1]$ into N equally spaced subintervals and let u_k^* denote a point in the k'th subinterval, i.e.,

$$u_k^* \in \left[\frac{k-1}{N}, \frac{k}{N}\right].$$

Assume that $\hat{t}_k$ has a probability density function $f_k = f(u_k^*, \,\cdot\,)$, where f is continuous in its first variable (which takes values in $[0,1]$). We let $F(u_k^*, \,\cdot\,)$ denote the corresponding cumulative distribution function. Our example will correspond to $u_k^* = k/N$ and the PDF given by

$$f(u,v) = \begin{cases} 1/u & \text{if} \qquad\qquad u < v < 2u \\ 0 & \text{otherwise.} \end{cases}$$

For a given v and small positive Δv, let $\mathcal{N}(v, \Delta v)$ denote the expected number of points of $\hat{t}_k$ in the interval $[v, v+\Delta v]$. For small Δv we have (rigorlessly; in the limit)

$$\mathcal{N}(v, \Delta v) = \sum_{k=1}^{N} f_k(v)\Delta v.$$

The main trick is to note that $\mathcal{N}(v, \Delta v)$ is also the number of $\hat{s}_k$ in that same interval $[v, v+\Delta v]$ and that the corresponding points of $\{(u_k^*, \hat{s}_k)\}$ lie in a strip of height Δv and width Δu given by

$$\frac{\Delta u}{\Delta v} = \frac{1}{N}\sum_{k=1}^{N} f(u_k^*, v),$$

which approaches $\int_0^1 f(u,v)\,du$ as $N \to \infty$. As $\Delta v \to 0$, so does Δu and the ratio approaches the slope of the curve $v = h(u)$. We now have

$$h'(v) = \int_0^1 f(u,v)\,du,$$

which yields

$$\begin{aligned} h(v) &= \int_{-\infty}^{v}\int_0^1 f(u,y)\,du\,dy \\ &= \int_0^1\int_{-\infty}^{v} f(u,y)\,dy\,du \\ &= \int_0^1 F(u,v)du. \end{aligned}$$

This is always a strictly increasing function of v whose inverse is the curve $v = g(u)$ sought.

In our particular example,

$$F(u,v) = \begin{cases} 1 & \text{if} \quad 0 \le u \le v/2 \\ (v-u)/u & \text{if} \quad v/2 < u \le \min(v,1) \\ 0 & \text{otherwise} \end{cases}$$

This gives

$$h(v) = \begin{cases} \int_0^{v/2} 1\ du + \int_{v/2}^{v} \frac{v-u}{u}\ du & = & v\log 2 & \text{if} \quad 0 \le v \le 1 \\ \int_0^{v/2} 1\ du + \int_{v/2}^{1} \frac{v-u}{u}\ du & = & v - 1 - v\log v + v\log 2 & \text{if} \quad 1 < v \le 2. \end{cases}$$

Solution #3. Consider the set of normalized points

$$\hat{S}_N := \left\{ \left(\frac{n}{N}, \frac{s_n}{N}\right) : n = 1, 2, \ldots, N \right\}.$$

For a subset A of the plane, let $\#_N(A)$ denote the number of points of $\hat{S}_N$ that lie in A. For a fixed $v \in \mathbb{R}$, consider the number $\hat{G}_N(v) = \#_N([0,1] \times (-\infty, v])$, which is the number of points of $\hat{S}_N$ lying at or below the horizontal line $y = v$. Since the points of $\hat{S}_N$ are each spaced horizontally $1/N$ unit apart, we need to show that the value $\hat{g}_N(v) := \frac{1}{N}\hat{G}_N(v)$ approaches a limit $g(v)$ as $N \to \infty$. In fact, we shall show that $g(v) = \int_0^1 F(x,v)dx$, where for each fixed $x \in [0,1]$, $F(x,\cdot)$ denotes the distribution function as given in *Solution #2*.

Fix v and let $\varepsilon > 0$. Since $F(\cdot, v)$ is continuous, it is Riemann-integrable and we may select a partition $x_0 = 0 < x_1 < x_2 < \cdots < x_{M-1} < x_M = 1$ of $[0,1]$ for which

$$\left| \sum_{k=1}^{M} F(x_k^*, v)\Delta x_k - \int_0^1 F(x,v)dx \right| < \varepsilon/2 \tag{1}$$

holds for any choices of $x_k^* \in I_k$, where $\Delta x_k = x_k - x_{k-1}$ and $I_k = [x_{k-1}, x_k]$.

Let $F_k^+(v)$ and $F_k^-(v)$ denote the maximum and minimum values, respectively, of $F(\cdot, v)$ on I_k. Choose N sufficiently large that

$$\Delta x_k F_k^-(v) - \frac{\varepsilon}{2M} < \frac{1}{N}\#_N(I_k \times (-\infty, v]) < \Delta x_k F_k^+(v) + \frac{\varepsilon}{2M} \tag{2}$$

for each k. [Note! At this point we feel we have attained some degree of rigor (mortis?) because we can justify the preceding step in a *frequentist* model of probability. In simpler situations in which each random variable is identically distributed, we would say that if an event has probability p, then the limit as $N \to \infty$ of the relative frequency of the event occurring in N trials approaches p. Here, we get one trial for each distinct probability, so we appeal to frequentism in this way: if the probability of an event is less than or equal to p, then the limit of the relative frequency is less than or equal to p. Such a large fuss over nothing, in all probability.] Then

$$\begin{aligned}\hat{g}_N(v) &= \sum_{k=1}^{M} \frac{1}{N} \#_N(I_k \times (-\infty, v]) \\ &< \sum_{k=1}^{M} \Big(\Delta x_k F_k^+(v) + \frac{\varepsilon}{2M} \Big) \qquad \text{(by (2))} \\ &< \int_0^1 F(x,v)dx + \varepsilon. \qquad \text{(by (1))}\end{aligned}$$

Similarly, we have

$$\hat{g}_N(v) \quad > \quad \int_0^1 F(x,v)dx - \varepsilon,$$

and together these prove that $\lim_{N\to\infty} \hat{g}_N(v) = \int_0^1 F(x,v)dx$. The rest of the solution is as in *Solution #2*.