Difference of two linearly distributed random numbers

2023-11-16

We explore the statistics of the absolute difference between two random variables, each of which is linearly distributed within a specified range. We derive the probability density function of the absolute difference, $S = |R_1 - R_2|$. We employ various methods for simulating random numbers that adhere to the derived distribution, including inverse transform sampling and geometric approaches. The results are validated through graphical comparisons of the theoretical and simulated distributions.

blog: https://tetraquark.vercel.app/posts/difference_lin_rand/

email: quarktetra@gmail.com

A couple years ago, I was reading through the lecture notes from Arpaci-Dusseau[1] on operating systems and came across some calculations on HDD seek time. I didn't quite agree with some of the hidden assumptions in the notes, and decided to take that on as a fun exercise with random variables. I wrote a blog post on it and shared it with one of the authors of the lecture notes. He liked my analysis so much so that he thought it should be extended and submitted for publication in a peer-reviewed conference. We just did that and we are waiting for the decision. Meanwhile, I want to isolate a portion of that blog post and discuss the calculation of probability density of the the absolute difference of two random numbers, which are themselves distributed linearly in a range.

Calculations

Consider a random variable R which is distributed linearly in a range $[r_\mathrm{i}, r_\mathrm{o}]$:

$$f_R(r) \equiv \frac{2r}{r_\mathrm{o}^2 - r_\mathrm{i}^2}, \text{ where } r_\mathrm{i} \leq r \leq r_\mathrm{o}. \tag{1}$$

We take two such random variables, and define their absolute difference as a new random variable:

$$S = |R_1 - R_2|. \tag{2}$$

The task is to figure out the probability density of S, i.e., $f_S(s)$. To this end, it is best to start from the cumulative distribution:

$$F_S(s) = P(|r_1 - r_2| < s) = \iint\limits_{|r_2 - r_1| < s} dr_1 dr_2 f_R(r_1) f_R(r_2). \tag{3}$$

We need to figure out the domain for which $|r_1 - r_2| < s$ is satisfied. It is the green shaded area in Figure 1.

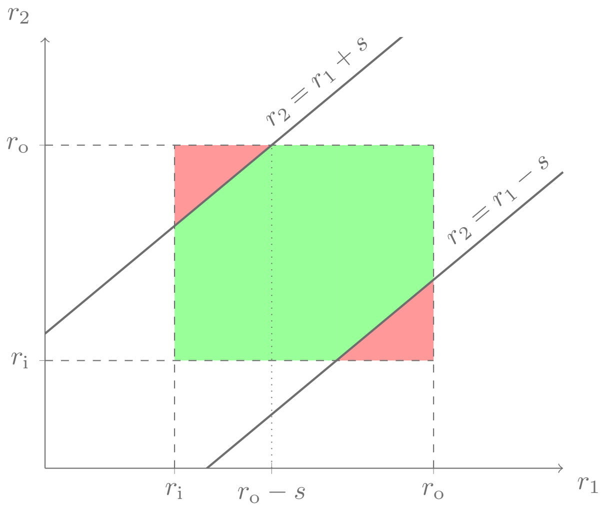

Figure 1: The domain of interest for integration. In the green shaded area $|r_2 - r_1| < s$ is satisfied.

Therefore, the cumulative probability function of the difference can be written as

$$\begin{aligned} F_S(s) &= \iint\limits_{|r_2 - r_1| < s} dr_1 dr_2 f_R(r_1) f_R(r_2) = \iint\limits_{\text{green}} dr_1 dr_2 f_R(r_1) f_R(r_2) \\ &= 1 - \iint\limits_{\text{red}} dr_1 dr_2 f_R(r_1) f_R(r_2) = 1 - 2 \int_{r_\mathrm{i}}^{r_\mathrm{o} - s} dr_1 f_R(r_1) \int_{r_1 + s}^{r_\mathrm{o}} f_R(r_2) dr_2 \\ &= 1 - 2 \int_{r_\mathrm{i}}^{r_\mathrm{o} - s} dr_1 f_R(r_1) \left[F_R(r_\mathrm{o}) - F_R(r_1 + s) \right], \end{aligned} \tag{4}$$

from which we can get the probability density by differentiating with respect to s:

$$\begin{aligned} f_S(s) &= \frac{\partial}{\partial s} F_S(s) = 2\int_{r_\mathrm{i}}^{r_\mathrm{o}-s} dr_1 f_R(r_1) f_R(r_1+s) \\ &= \frac{4}{3\left(r_\mathrm{o}^2 - r_\mathrm{i}^2\right)^2}\left[2(r_\mathrm{o}^3 - r_\mathrm{i}^3) - 3(r_\mathrm{o}^2 + r_\mathrm{i}^2)s + s^3\right]. \qquad (5) \end{aligned}$$

Simulation

We first need a way of creating random numbers with the distribution in Eq. 1. Such a linear distribution is not typically available in standard programming languages, and we have to build the distribution ourselves. There are multiple ways of doing this. For example:

1. We can take two uniform random numbers, add them up to get a triangular distribution (this simply follows from the convolution of two rectangular distributions.) We can then carve out the range we are interested by simply rejecting the instances outside.
2. Alternatively, we can take two random variables X and Y uniformly distributed in the domain defined by $r_i^2 < x^2 + y^2 < r_o^2$ with the density $\frac{1}{\pi(r_0^2 - r_i^2)}$.

We then define two new random variables $R = (X^2+Y^2)^{1/2}$ and $\Phi = \mathrm{sign}(Y)\arccos\left(\frac{X}{(X^2+Y^2)^{1/2}}\right)$. The measure of the integral will transform with the Jacobian matrix

$$\begin{aligned} \frac{dxdy}{\pi(r_0^2 - r_i^2)} &= \frac{1}{\pi(r_0^2 - r_i^2)}\left|\frac{d(x,y)}{d(r,\phi)}\right| drd\phi = \frac{1}{\pi(r_0^2 - r_i^2)}\begin{vmatrix} cos\phi & -rsin\phi \\ sin\phi & r\cos\phi \end{vmatrix} drd\phi \\ &= \frac{r}{\pi(r_0^2 - r_i^2)} drd\phi. \qquad (6) \end{aligned}$$

Upon integrating out ϕ, we pick up a factor of 2π, and get the same expression for $f(r)$ as we had in Eq. 1.

Yet another method is to use inverse transform sampling. Consider a random variable $U \sim \mathrm{Unif}(0,1)$ and define a random variable $X = F_X^{-1}(U)$. With this construction, X will have CDF as F_X. Let's give this a try. We first calculate F_R given f_R in Eq. 1:

$$F_R(r) \equiv \int_{-\infty}^{r} d\tau f_R(\tau) = \frac{r^2 - r_\mathrm{i}^2}{r_\mathrm{o}^2 - r_\mathrm{i}^2}, \text{ where } r_\mathrm{i} \le r \le r_\mathrm{o}, \qquad (7)$$

which results in

$$F_R^{-1}(U) = \sqrt{r_\mathrm{i}^2 + (r_\mathrm{o}^2 - r_\mathrm{i}^2)U}. \qquad (8)$$

In summary, if we pull numbers from a uniform random variable (U Unif(0,1)) and apply the function in Eq. 8, we will get the linear distribution. Since this is the simplest one, we will take this approach in the simulation. I will simply simulate the random numbers with $r_\mathrm{i} = 1$ and $r_\mathrm{o} = 2$, and overlay the distributions from the simulation with the expected ones from the model above. The results are shown in Figure 2 and Figure 3, which demonstrate perfect agreement with the calculated and simulated distributions.

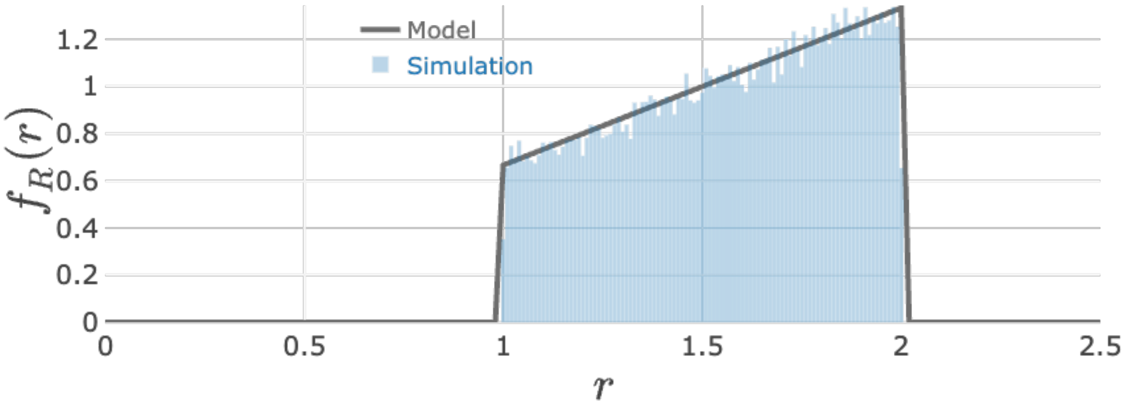

Figure 2: The linear distribution created out of the uniform random variable with $r_\mathrm{i} = 1$ and $r_\mathrm{o} = 2$.

[1] R. H. Arpaci-Dusseau and A. C. Arpaci-Dusseau, *Operating Systems: Three Easy Pieces*, 1.00 ed. https://pages.cs.wisc.edu/~remzi/OSTEP/file-disks.pdf; Arpaci-Dusseau Books, 2018.

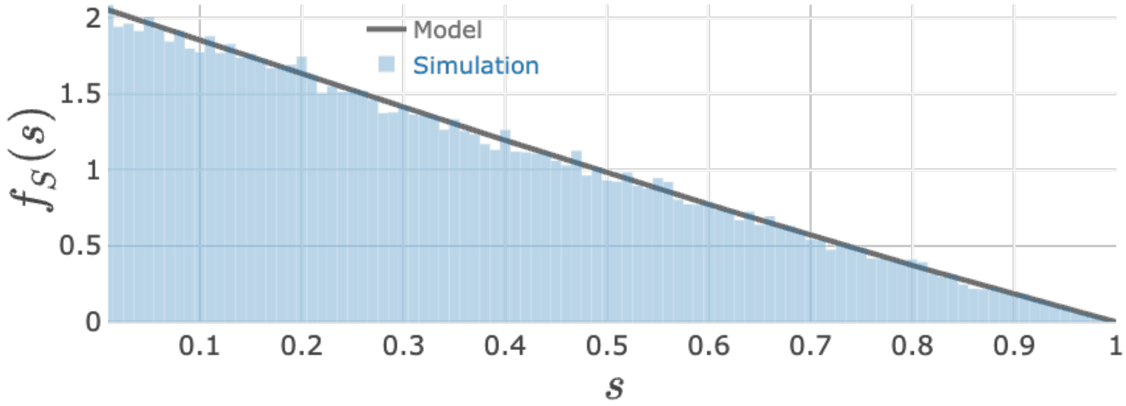

Figure 3: The distribution of the absolute difference of two linearly distributed random numbers.