Interval Spacing

Greg Kreider

January 16, 2026

Abstract

We define interval spacing as the difference in the order statistics of data over a gap of some width. We derive its density, expected value, and variance for uniform, exponential, and logistic variates. We show that interval spacing is equivalent to running a rectangular low-pass filter over the spacing, which simplifies the expressions for the expected values and introduces correlations between overlapping intervals.

The theory behind spacing, the difference between consecutive order statistics, is well-developed by now. Occasionally one sees the difference over larger gaps being used. [7] estimates the location of the mode as the midpoint of the interval with the most data points, where the size of the interval depends on the polynomial order of the data around the mode. [1] collects a test statistic over a range of gaps to determine if data has a non-zero slope. [2] and [4] study such test statistics in general for non-overlapping segments.

Let us call "interval spacing" the difference in order statistics over distances greater than one. Extending the notation used in [5] with i the upper index and w the width of the interval such that $w < i \leq n$, the interval spacing $D_{i,w}$ is

$$D_{i,w} = T_i - T_{i-w} \tag{1}$$

where T_i is the i^{th} order statistic. Alternative notations that have been used are a center point $j = i - (w/2)$ and radius $r_n = w/2$ [7, (2.1)], or start index $j = i - w + 1$ and endpoint $k = i$ of the range [1]. In this notation the spacing $D_i = D_{i,1}$.

The density of the interval spacing

$$\begin{aligned} f_{D_{i,w}}(y) = S_1 \int_{-\infty}^{\infty} &\{F_x(x)\}^{i-w-1} \{F_x(x+y) - F_x(x)\}^{w-1} \{1 - F_x(x+y)\}^{n-i} \\ &\times f_x(x)\ f_x(x+y)\ dx \end{aligned} \tag{2}$$

$$S_1 = \frac{n!}{(i-w-1)!\ (w-1)!\ (n-i)!}$$

follows from the joint density of two order statistics [8, (8) and (31) with $r = i-w$ and $r' = i$], where $F_x(x)$ and $f_x(x)$ are distribution and density functions.

Expanding [5, (2.4)] gives the same density, using $k_1 = i - w$, $t_1 = x$, $k_2 = i$, $t_2 = x + y$, $k_3 = n + 1$, $t_3 = \infty$, and $t_0 = -\infty$. (2) simplifies to the spacing's density function [5, (2.7)] when $w = 1$; notably, the factor of the distribution function raised to $w - 1$ disappears. The expected value and variance follow normally from the first two moments of this density.

$$E\Big\{D_{i,w}\Big\} = \int_0^\infty y\ f_{D_{i,w}}(y)\ dy \tag{3}$$

$$V\Big\{D_{i,w}\Big\} = E\Big\{D_{i,w}^2\Big\} - E\Big\{D_{i,w}\Big\}^2 = \int_0^\infty y^2\ f_{D_{i,w}}(y)\ dy - E\Big\{D_{i,w}\Big\}^2 \tag{4}$$

For uniform variates over the range a, b we have

$$f_{D_{i,w,unif}}(y) = \frac{n!}{(w-1)!(n-w)!}\ \left(\frac{1}{b-a}\right)^n\ y^{w-1}\ (b-y-a)^{n-w} \tag{5}$$

$$E\Big\{D_{i,w,unif}\Big\} = w\ \frac{b-a}{n+1} \tag{6}$$

$$V\Big\{D_{i,w,unif}\Big\} = w\ \frac{(n+1-w)}{n+2}\left(\frac{b-a}{n+1}\right)^2 \tag{7}$$

For exponential variates with rate parameter λ these are

$$f_{D_{i,w,exp}}(y) = w\ \binom{n-i+w}{n-i}\ \lambda\ \left\{e^{-\lambda y}\right\}^{n-i+1}\left\{1-e^{-\lambda y}\right\}^{w-1} \tag{8}$$

$$E\Big\{D_{i,w,exp}\Big\} = w\ \binom{n-i+w}{n-i}\ \frac{1}{\lambda}\ (-1)^{w-1}\sum_{k=0}^{w-1}\binom{w-1}{k}\frac{(-1)^k}{(n-i+w-k)^2} \tag{9}$$

$$E\Big\{D_{i,w,exp}^2\Big\} = w\ \binom{n-i+w}{n-i}\ \frac{2}{\lambda^2}\ \sum_{k=0}^{w-1}\binom{w-1}{k}\frac{(-1)^k}{(n-i+w-k)^3} \tag{10}$$

Use (4) and the last two equations to calculate the variance; squaring (9) and subtracting from (10) does not lead to a simpler equation. The pre-factor can also be written $(n-i+w)!/(w-1)!(n-i)!$, which will cancel a factor $(w-1)!$ from the combinatorial inside the series.

For logistic variates with mean μ and standard deviation σ we find

$$\begin{aligned} f_{D_{i,w,logis}}(y) &= \frac{1}{\sigma}\ S_2\ e^{y/\sigma}\left\{e^{y/\sigma}-1\right\}^{w-1} \\ &\qquad \times {}_2F_1\left(i, n-i+w+1; n+w+1; 1-e^{y/\sigma}\right) \end{aligned} \tag{11}$$

$$\begin{aligned} S_2 &= \frac{n!}{(n+w)!}\ \frac{(n-i+w)!}{(n-i)!}\ \frac{(i-1)!}{(i-w-1)!}\ \frac{1}{(w-1)!} \\ &= w\prod_{j=1}^{w}\frac{(i-j)(n-i+j)}{j(n+j)} \end{aligned}$$

$$E\Big\{D_{i,w,logis}\Big\} = \sum_{k=0}^{w-1} \frac{n!}{(i-w-1)!(w-1-k)!(n-i+1+k)!}\ \sigma\ (-1)^{w-1-k}$$

$$\times \left[\begin{array}{r} \dfrac{1}{i-k-1}\left[\psi(i-k)-\psi(i-k-1)\right] \\ -\displaystyle\sum_{l=1}^{n-i+w-1} \dfrac{1}{n-i+w-l}\ B(n-i+w+1-l, i-k-1) \\ -\ \dfrac{w-1-k}{n-i+w-1}\ B(n-i+w+1, i-k-2) \\ +\displaystyle\sum_{l=2}^{w-1-k} (-1)^l \sum_{j=0}^{l-1} S_3\ B(n-i+w+1+j, i-k-2-j) \end{array}\right] \tag{12}$$

$$S_3 = \frac{(w-1-k)!}{(w-1-k-l)!}\ \frac{(n-i+w-l-1)!}{(n-i+w-l+j)!}\ \frac{1}{l(l-1-j)!}$$

The second moment would require an additional integration by parts of these terms, which we do not perform. $B(a,b) = (a-1)!(b-1)!/(a+b-1)!$ is the beta function and $\psi(x)$ the dilogarithm.

The interval spacing introduces an extra factor F^{w-1} of the distribution function which requires an integration by parts to handle. This leads to more complicated expressions for the expected interval spacing than for the normal spacing, although all results reduce to the spacing versions found in [3] if $w = 1$. The uniform and exponential equations follow from known definite integrals (see Supplemental Materials for derivations). Rather than integrating the hypergeometric function in the logistic density function (11) to get the first moment, it is better to integrate first over y after combining (2) and (3), then over x. This must be done by parts, which involves a recursion down n that gives a series that must then be integrated term by term.

High-precision math libraries must be used to evaluate the series. The factorial scaling factors in (5) – (11) reach $n^w/w!$ while the spacing is of order one, so the series sum of terms of alternating sign of this size must nearly cancel.

Figure 1 plots the density of the interval spacing for draws of $n = 50$ data points. The function for uniform variates in the left graph is independent of the index i, but the others show two sets of curves. The densities with high, narrow peaks correspond to small values from the variate, found at the start of the exponential's order statistics, drawn with $i = w + 2$, or the center of the logistic, drawn with $i = n/2$. The lower, broader densities come from the tails of the distribution where the order statistics are largest. This occurs near the final indices of the expontial, drawn for $i = n - 2$, and the initial indices of the logistic, drawn for $i = w + 2$ but also valid by symmetry at the other tail. All densities skew to the right, although this is much more noticeable in the broader curves.

Figure 2 plots the expected interval spacing for exponential and logistic variates at three widths; the data size $n = 50$ has been kept small for visibility.

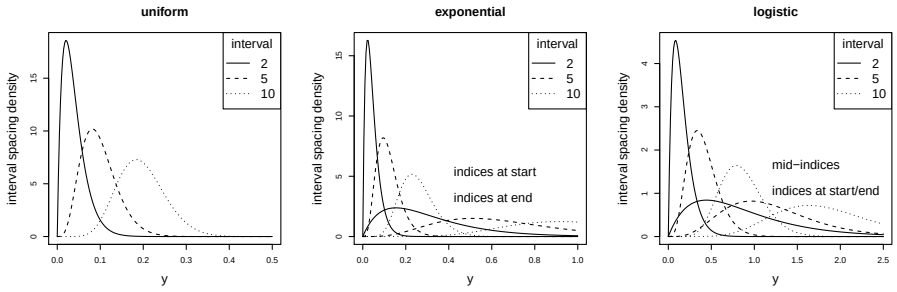

Figure 1: Density of the interval spacing at widths w of 2, 5, and 10 and indices i as noted.

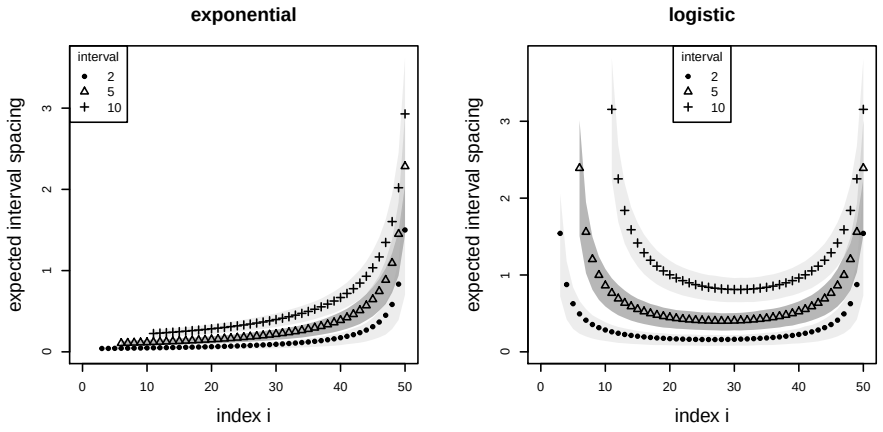

Figure 2: Expected interval spacing for exponential and logistic variates at widths w of 2, 5, and 10. Bands are inter-quartile ranges from simulations.

The grey bands are inter-quartile ranges based on ten thousand simulated draws. The expected values are not centered, lying closer to the upper quartile. This reflects the skewing of the density. Such shifting also happens for the medians, not plotted.

An interval can be broken into non-overlapping segments whose sub-spacing will add. For example, if w is even we can split the interval in half,

$$\begin{aligned} D_{i,w/2} + D_{i-w/2,w/2} &= T_i - T_{i-w/2} + T_{i-w/2} - T_{i-w} = T_i - T_{i-w} \\ &= D_{i,w} \end{aligned} \tag{13}$$

The half intervals sum to the whole. Any sub-intervals need not have the same size, which would happen in this example if w were odd, but the endpoints must match and cover the whole range. Splitting into more pieces is also possible, a process which ultimately ends by breaking the interval into w single steps. The

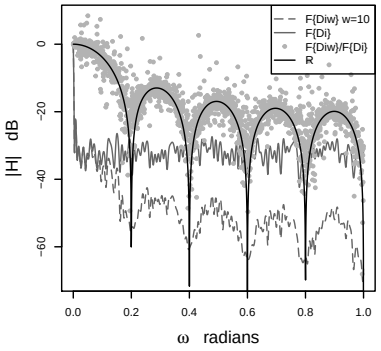

Figure 3: Interval spacing is equivalent to a rectangular low-pass filter applied to spacing.

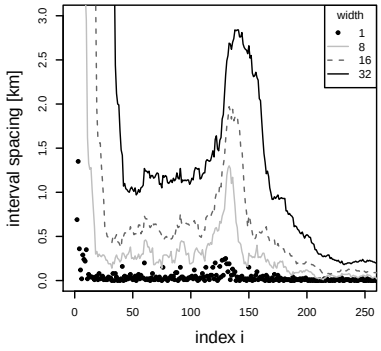

Figure 4: Low-pass filtering of spacing for increasing interval widths w.

interval spacing is the sum of these spacings,

$$D_{i,w} = \sum_{j=0}^{w-1} D_{i-j} \tag{14}$$

A sum over consecutive data points is just a rectangular low-pass filter. If the total were divided by w this would be a running mean. Because such normalization is not done — the filter kernel is a vector of w 1's rather than $1/w$'s — larger intervals amplify differences. This must be taken into account when comparing different widths. The spacing is not constant over the interval so the interval spacing is not just the spacing scaled by w, although it is close, especially where the variate's density is nearly constant, at the start of one-sided distributions or the center of two-sided.

We can demonstrate that filtering occurs by taking the ratio of the Fourier transforms of $D_{i,w}$ and D_i and comparing it to the impulse response $\mathcal{R}$ of the rectangular kernel r. That is,

$$\mathcal{F}\left\{D_{iw}\right\} = \mathcal{F}\left\{r * D_i\right\} = \mathcal{R} \cdot \mathcal{F}\left\{D_i\right\} \tag{15}$$

Figure 3 shows the ratio applied to a draw of 1200 points from a uniform distribution. The transforms of the spacing and interval spacing at $w = 10$ have been smoothed for display, but their ratio has not. Superimposed is the magnitude of $\mathcal{R}$, which has the same number of side lobes at the same height.

(14) also applies to the expected values, which allows us to simplify (9) using $E\{D_{i,exp}\}$ [3, (8)]

$$E\Big\{D_{i,w,exp}\Big\} = \sum_{j=0}^{w-1} E\Big\{D_{i,exp}\Big\} = \sum_{j=0}^{w-1} \frac{1}{\lambda\,(n-i+j+1)} \tag{16}$$

and (12) with $E\{D_{i,logis}\}$ [3, (17)]

$$E\Big\{D_{i,w,logis}\Big\} = \sum_{j=0}^{w-1} \frac{\sigma\ n}{(i-j-1)(n-i+j+1)} \tag{17}$$

The expected spacing for uniform variates, $(b-a)/(n+1)$, is repeated w times in the sum independently of j, matching (6). We could write an expression for the expected interval spacing for Gumbel variates using [3, (20)], although the result is neither simple nor illuminating. (14) can also be used for variates requiring numeric integration for their expected spacing.

Unlike (9) and (12), (16) and (17) do not require high precision libraries and can be evaluated directly.

Although a rectangular low-pass filter has a wide main band, it suppresses the sidelobes by only a moderate amount, which allows high-frequency residuals to remain, especially at sharp edges. Figure 4 plots the spacing for the depth of earthquakes under Mt. St. Helens before its eruption in 1980 [6]. Depths are considered below the surface and are negative, so the first order statistics are the deepest and the largest interval spacings are at the smallest indices and initially decrease rapidly. There are five individual large spacings between indices 40 and 100 corresponding to a depth of -7.90 – -5.62 km, and a cluster around 130, between -4.83 km and -2.87 km. The former create small coarse bumps in the $w = 8$ interval spacing, and the latter a sharp peak. This peak widens with the interval width, while the bumps merge into a single rough region without damping their range, an example of insufficient suppression of higher frequencies. Eventually at $w = 32$ one individual point falls within every interval and the bumps disappear. The signal simplifies to a nearly constant level, becoming a plateau or flat. The trailing edge of the peak beyond index 150 does become smoother with larger intervals.

If we consider overlapping intervals by lag $1 \le l < w$,

$$\begin{aligned} D_{i,w} - D_{i-l,w} &= \sum_{k=0}^{w-1} D_{i-k} - \sum_{k=0}^{w-1} D_{i-l-k} = \sum_{k=0}^{w-1} D_{i-k} - \sum_{k'=0}^{w-1+l} D_{i-k'} \\ &= \sum_{k=0}^{l-1} D_{i-k} - \sum_{k'=w}^{w-1+l} D_{i-k'} \end{aligned} \tag{18}$$

where in the second step we have shifted $k' = l + k$. The two terms in the final result do not overlap, because the common spacings $\sum_{k=l}^{w-1} D_{i-k}$ have canceled. Said differently, overlapping intervals will be correlated, sharing $w-l$ terms. The autocovariance of the interval spacing follows the convolution of the rectangular kernel (not its impulse response) with itself, which is a linear decrease with initial value equal to the variance of $D_{i,w}$ and slope $-V\{D_{i,w}\}/w$ (Figure 5).

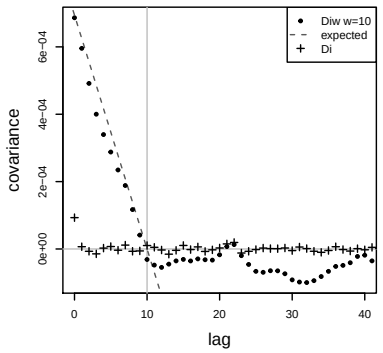

Figure 5: Auto-covariance of the interval spacing follows the self-convolution of the rectangular kernel.

Supplemental Material

The supplement for this paper includes derivations of the density, expected value, and variance of the interval spacing for uniform, exponential, and logistic variates.

References

[1] Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *The Annals of Statistics* **36**, 4, 1758–1785. doi: 10.1214/07-AOS521.

[2] Jammalamadaka, S. R., Zhou, X., and Tiwari, R. C. (1989). Asymptotic efficiencies of spacings tests for goodness of fit. *Metrika 36*, 355–377.

[3] Kreider, G. (2023). Expected spacing. *Communications in Statistics - Theory and Methods* **53**, 23, 8286–8296. doi: 10.1080/03610926.2023.2281265.

[4] Mirakhmedov, S. M. and Jammalamadaka, S. R. (2013). Higher-order expansions and efficiencies of tests based on spacings. *Journal of Nonparametric Statistics* **25**, 2, 339–359. doi: 10.1080/10485252.2012.755530.

[5] Pyke, R. (1965). Spacings. *Journal of the Royal Statistical Society, Series B* **27**, 3, 395–449.

[6] Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley and Sons, New York.

[7] Venter, J. H. (1967). On estimation of the mode. *The Annals of Mathematical Statistics 38*, 1446–1455.

[8] Wilks, S. S. (1948). Order statistics. *Bulletin of the American Mathematical Society* **54**, 1 (Jan.), 6–50.