Supplement to *Expected Spacing*

Greg Kreider

DERIVATION OF EQUATIONS

SPACING FOR UNIFORM VARIATES

Density Function

The density function is

$$\begin{aligned} f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!} \int_{-\infty}^{\infty} \left(\frac{x-a}{b-a}\right)^{i-2} \left(1 - \frac{x+y-a}{b-a}\right)^{n-i} \left(\frac{1}{b-a}\right) \left(\frac{1}{b-a}\right) dx \\ &= \frac{n!}{(i-2)!(n-i)!} \left(\frac{1}{b-a}\right)^{i} \int_{-\infty}^{\infty} \left(\frac{b-x-y}{b-a}\right)^{n-i} (x-a)^{i-2}\, dx \\ &= \frac{n!}{(i-2)!(n-i)!} \left(\frac{1}{b-a}\right)^{n} \int_{a}^{b-y} (x-a)^{i-2}(b-y-x)^{n-i} dx \end{aligned}$$

where the lower bound of the integral is determined by the larger of a and $a-y$, since the two density functions are 0 below these limits, and the upper by the smaller of b and $b-y$. From [1, (3.196.3)] (please note that variables in any cited equation may be changed to avoid conflicts with our analysis)

$$\int_{\alpha}^{\beta} (x-\alpha)^{\mu-1}(\beta-x)^{\nu-1} dx = (\beta-\alpha)^{\mu+\nu-1} B(\mu,\nu)$$

With $\mu = i-1$ and $\nu = n-i+1$, both of which are positive and meet the requirements for the definite integral, and $\alpha = a$, $\beta = b-y$,

$$f_{D_i}(y) = \frac{n!}{(i-2)!(n-i)!} \left(\frac{1}{b-a}\right)^{n} (b-y-a)^{n-1} B(i-1, n-i+1)$$

Because the index and sample sizes are integers,

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \frac{(x-1)!(y-1)!}{(x+y-1)!}$$

So

$$\begin{aligned} f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!} \left(\frac{1}{b-a}\right)^{n} (b-y-a)^{n-1} \frac{(i-2)!(n-i)!}{(n-1)!} \\ &= n \left(\frac{1}{b-a}\right)^{n} (b-y-a)^{n-1} \\ &= \frac{n}{b-a} \left(\frac{b-y-a}{b-a}\right)^{n-1} \end{aligned} \tag{U.1}$$

For the standard range, $a = 0$ and $b = 1$ and this simplifies to [2, (2.3)]

Expected Spacing

The expected spacing or first moment is

$$\begin{aligned}
E\Big\{D_i\Big\} &= \int_0^\infty y f_{D_i}(y)dy \\
&= \int_0^{b-a} y\frac{n}{b-a}\left(\frac{b-a-y}{b-a}\right)^{n-1} dy \\
&= \frac{n}{(b-a)^n}\int_0^{b-a} y(b-a-y)^{n-1}dy
\end{aligned}$$

The integral's upper bound is the maximum spacing possible. Also using [1, (3.196.3)] with $\mu = 2$ and $\nu = n$,

$$\begin{aligned}
E\Big\{D_i\Big\} &= \frac{n}{(b-a)^n}(b-a)^{n+1}B(2,n) \\
&= n(b-a)\frac{1!(n-1)!}{(n+1)!} \\
&= \frac{b-a}{n+1}
\end{aligned} \tag{U.2}$$

The result again matches the formula in Pyke for the unit range.

Variance of Spacing

The variance of the spacing starts with the second moment. Once again with [1, (3.196.3)] and $\mu = 3$ and $\nu = n$,

$$\begin{aligned}
E\Big\{D_i^2\Big\} &= \int_{-\infty}^\infty y^2 f_{D_i}(y)dy \\
&= \frac{n}{(b-a)^n}\int_0^{b-a} y^2(b-a-y)^{n-1}dy \\
&= \frac{n}{(b-a)^n}(b-a)^{n+2}\frac{2!(n-1)!}{(n+2)!} \\
&= \frac{2(b-a)^2}{(n+2)(n+1)}
\end{aligned} \tag{U.3}$$

The variance is

$$\begin{aligned}
V\Big\{D_i\Big\} &= E\Big\{D_i^2\Big\} - E^2\Big\{D_i\Big\} \\
&= \frac{2(b-a)^2}{(n+2)(n+1)} - \frac{(b-a)^2}{(n+1)^2} \\
&= \frac{2(n+1)-(n+2)}{(n+2)(n+1)^2}(b-a)^2 \\
&= \frac{n}{(n+2)}\left(\frac{b-a}{n+1}\right)^2
\end{aligned} \tag{U.4}$$

This too matches Pyke. It is smaller than the square of the expected spacing by $n/(n+2)$.

SPACING FOR EXPONENTIAL VARIATES

Density Function

The density function is

$$\begin{aligned} f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!}\int_0^\infty \left(1-e^{-\lambda x}\right)^{i-2}\left\{1-\left(1-e^{-\lambda(x+y)}\right)\right\}^{n-i}\lambda e^{-\lambda x}\lambda e^{-\lambda(x+y)}dx \\ &= \frac{n!}{(i-2)!(n-i)!}\lambda^2\int_0^\infty \left(1-e^{-\lambda x}\right)^{i-2}e^{-\lambda(n-i)x}e^{-\lambda(n-i)y}e^{-\lambda 2x}e^{-\lambda y}dx \\ &= \frac{n!}{(i-2)!(n-i)!}\lambda^2 e^{-\lambda(n-i+1)y}\int_0^\infty \left(1-e^{-\lambda x}\right)^{i-2}e^{-\lambda(n-i+2)x}dx \end{aligned}$$

where the restriction $x \geq 0$ sets the lower limit. From [1, (3.312.1)]

$$\int_0^\infty \left(1-e^{-x/\beta}\right)^{\nu-1}e^{-\mu x}dx = \beta B(\beta\mu,\nu) = \beta\frac{(\beta\mu-1)!(\nu-1)!}{(\beta\mu+\nu-1)!}$$

With $\beta = 1/\lambda$, $\nu = i-1$, $\mu = \lambda(n-i+2)$, $\beta\mu = n-i+2$

$$\begin{aligned} f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!}\lambda^2 e^{-\lambda(n-i+1)}\frac{1}{\lambda}\frac{(n-i+1)!(i-2)!}{n!} \\ &= \lambda(n-i+1)e^{-\lambda(n-i+1)y} \end{aligned} \tag{E.1}$$

This is one factor of [2, (2.9)].

Expected Spacing

The expected spacing is

$$E\left\{D_i\right\} = \int_0^\infty y f_{D_i}(y)dy = \lambda(n-i+1)\int_0^\infty ye^{-\lambda(n-i+1)y}dy$$

Using [1, (2.322.1)]

$$\int xe^{ax}dx = e^{ax}\left(\frac{x}{a}-\frac{1}{a^2}\right)$$

we get

$$\begin{aligned} E\left\{D_i\right\} &= \lambda(n-i+1)\left[e^{-\lambda(n-i+1)y}\left\{\frac{y}{-\lambda(n-i+1)}-\frac{1}{\lambda^2(n-i+1)^2}\right\}\right]_0^\infty \\ &= \left[-ye^{-\lambda(n-i+1)y}-\frac{1}{\lambda(n-i+1)}e^{-\lambda(n-i+1)y}\right]_0^\infty \\ &= \frac{1}{\lambda(n-i+1)} \end{aligned} \tag{E.2}$$

where the exponential dominates at the upper bound, driving both terms to zero, and the first term drops at the lower bound.

Variance of Spacing

Following what we did for the uniform distribution, we calculate first the second moment,

$$E\left\{D_i^2\right\} = \int_0^\infty y^2 f_{D_i}(y)dy = \lambda(n-i+1)\int_0^\infty y^2e^{-\lambda(n-i+1)y}dy$$

with [1, (2.322.2)]

$$\int x^2e^{ax}dx = e^{ax}\left(\frac{x^2}{a}-\frac{2x}{a^2}+\frac{2}{a^3}\right)$$

$a = -\lambda(n-i+1)$ is also the factor before the integral, leaving

$$\begin{aligned} E\left\{D_i^2\right\} &= \left[e^{-\lambda(n-i+1)y}\left\{-y^2 - \frac{2y}{\lambda(n-i+1)} - \frac{2}{\lambda^2(n-i+1)^2}\right\}\right]_0^\infty \\ &= \frac{2}{\lambda^2(n-i+1)^2} \end{aligned} \tag{E.3}$$

Here too only the last term survives at the lower bound. The variance then follows as

$$\begin{aligned} V\left\{D_i\right\} &= E\left\{D_i^2\right\} - E^2\left\{D_i\right\} \\ &= \frac{2}{\lambda^2(n-i+1)^2} - \frac{1}{\lambda^2(n-i+1)^2} \\ &= \frac{1}{\lambda^2(n-i+1)^2} \end{aligned} \tag{E.4}$$

Unlike the uniform variance, this is the square of the expected spacing.

SPACING FOR LOGISTIC VARIATES

Density Function

Starting with the distribution's density functions and substituting $z = (x-\mu)/\sigma$, $dz = dx/\sigma$, and $w = y/\sigma$,

$$f(x) = \frac{e^{-z}}{\sigma(1+e^{-z})^2} \qquad f(x+y) = \frac{e^{-z}e^{-w}}{\sigma(1+e^{-z}e^{-w})^2}$$

$$F(x) = \frac{1}{1+e^{-z}} \qquad F(x+y) = \frac{1}{1+e^{-z}e^{-w}}$$

The spacing's density function is

$$\begin{aligned}
f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i-2}\left(1-\frac{1}{1+e^{-z}e^{-w}}\right)^{n-i}\left(\frac{e^{-z}}{\sigma(1+e^{-z})^2}\right)\left(\frac{e^{-z}e^{-w}}{\sigma(1+e^{-z}e^{-w})^2}\right)\sigma dz \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{1}{\sigma}e^{-w}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i-2}\left(\frac{e^{-z}e^{-w}}{1+e^{-z}e^{-w}}\right)^{n-i}e^{-2z}\left(\frac{1}{1+e^{-z}}\right)^{2}\left(\frac{1}{1+e^{-z}e^{-w}}\right)^{2}dz \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{1}{\sigma}e^{-w}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i}\left(\frac{1}{1+e^{-z}e^{-w}}\right)^{n-i+2}e^{-(n-i)z}e^{-(n-i)w}e^{-2z}dz \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{1}{\sigma}e^{-(n-i+1)w}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i}\left(\frac{1}{1+e^{-z}e^{-w}}\right)^{n-i+2}e^{-(n-i+2)z}dz \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{1}{\sigma}e^{-(n-i+1)w}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i}\left(\frac{e^{w}}{e^{w}+e^{-z}}\right)^{n-i+2}e^{-(n-i+2)z}dz \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{1}{\sigma}e^{w}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i}\left(\frac{1}{e^{w}+e^{-z}}\right)^{n-i+2}e^{-(n-i+2)z}dz \qquad \text{(L.1)}
\end{aligned}$$

This has the form of the definite integral [1, (3.315.1)]

$$\int_{-\infty}^{\infty}\left[\frac{1}{e^{\beta}+e^{-x}}\right]^{\nu}\left[\frac{1}{e^{\gamma}+e^{-x}}\right]^{\rho}e^{-\mu x}dx = e^{(\mu-\rho)\gamma-\beta\nu}B(\mu,\nu+\rho-\mu)\,{}_2F_1\left(\nu,\mu;\nu+\rho;1-e^{\gamma-\beta}\right)$$

We have $\beta = 0$, $\gamma = w$, $\rho = n-i+2$, $\mu = n-i+2$, and $\nu = i$. These values satisfy the conditions on the integral: $|\mathrm{Im}(\beta)| = 0 < \pi$; $|\mathrm{Im}(\gamma)| = 0 < \pi$; and $\mathrm{Re}(\nu+\rho) = n+2 > \mathrm{Re}(\mu) = n-i+2 > 0$. Now $(\mu-\rho)\gamma - \beta\nu$ is zero and the exponential falls out, leaving

$$\begin{aligned}
f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!}\frac{1}{\sigma}e^{w}B(n-i+2,i)\,{}_2F_1\left(i,n-i+2;n+2;1-e^{w}\right) \\
&= \frac{1}{\sigma}e^{w}\frac{n!}{(i-2)!(n-i)!}\frac{(n-i+1)!(i-1)!}{(n+1)!}\,{}_2F_1\left(i,n-i+2;n+2;1-e^{w}\right) \\
&= \frac{1}{\sigma}e^{y/\sigma}\frac{(n-i+1)(i-1)}{n+1}\,{}_2F_1\left(i,n-i+2;n+2;1-e^{y/\sigma}\right) \qquad \text{(L.2)}
\end{aligned}$$

Expected Spacing

To avoid integrating the hypergeometric function we can change the order of the integrals when calculating the expected spacing. Starting with (L.1) and making the same substitutions z and w, so $dw = dy/\sigma$ and the integration bounds stay the same,

$$\begin{aligned}
E\left\{D_i\right\} &= \int_0^{\infty} y f_{D_i}(y)dy \\
&= \frac{n!}{(i-2)!(n-i)!}\int_0^{\infty}\frac{y}{\sigma}e^{w}dy\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i}\left(\frac{1}{e^{w}+e^{-z}}\right)^{n-i+2}e^{-(n-i+2)z}dz \\
&= \frac{\sigma n!}{(i-2)!(n-i)!}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-z}}\right)^{i}e^{-(n-i+2)z}\int_0^{\infty}we^{w}\left(\frac{1}{e^{w}+e^{-z}}\right)^{n-i+2}dwdz
\end{aligned}$$

Doing the inner integral by parts,

$$I_{in} = \int_0^{\infty}we^{w}\left(\frac{1}{e^{w}+e^{-z}}\right)^{n-i+2}dw$$

we set $u = w$, $du = dw$, and

$$dv = e^w \left(\frac{1}{e^w + e^{-z}} \right)^{n-i+2} dw$$

Letting $\eta = n - i + 2$, $t = e^w + e^{-z}$, and $dt = e^w dw = (t - e^{-z})dw$,

$$\begin{aligned} dv &= t^{-\eta} dt \\ v &= \frac{t^{-(\eta-1)}}{-(\eta-1)} \end{aligned}$$

and the integral becomes

$$\begin{aligned} I_{in} &= \Big[uv\Big]_0^\infty - \int_0^\infty v du \\ &= \left[\frac{w}{-(\eta-1)} \left(\frac{1}{e^w + e^{-z}} \right)^{\eta-1} \right]_0^\infty - \frac{1}{-(\eta-1)} \int_0^\infty \left(\frac{1}{e^w + e^{-z}} \right)^{\eta-1} dw \end{aligned}$$

The first term goes to zero at both limits, so

$$I_{in} = \frac{1}{\eta-1} \int_0^\infty \left(\frac{1}{t} \right)^{\eta-1} \frac{1}{t - e^{-z}} dt$$

The integral has a known solution [1, (2.117.4)]

$$\int \frac{dx}{x^m(a+bx)} = \frac{(-1)^m b^{m-1}}{a^m} \ln \left(\frac{a+bx}{x} \right) + \sum_{k=1}^{m-1} \frac{(-1)^k b^{k-1}}{(m-k)a^k x^{m-k}}$$

With $m = \eta - 1$, $a = -e^{-z}$, and $b = 1$ this becomes

$$\begin{aligned} \int \left(\frac{1}{t} \right)^{\eta-1} \frac{dt}{t - e^{-z}} &= \frac{(-1)^{\eta-1}(1)^{\eta-2}}{(-e^{-z})^{\eta-1}} \ln \left(\frac{-e^{-z} + t}{t} \right) + \sum_{k=1}^{\eta-2} \frac{(-1)^k (1)^{k-1}}{(\eta-1-k)\left(-e^{-z}\right)^k} \left(\frac{1}{t} \right)^{\eta-1-k} \\ &= \frac{1}{e^{-(\eta-1)z}} \ln \left(\frac{e^w}{e^w + e^{-z}} \right) + \sum_{k=1}^{\eta-2} \frac{1}{\eta-1-k} \frac{1}{e^{-kz}} \left(\frac{1}{e^w + e^{-z}} \right)^{\eta-1-k} \\ &= e^{(\eta-1)z} \ln \left(\frac{e^w}{e^w + e^{-z}} \right) + \sum_{k=1}^{\eta-2} \frac{1}{\eta-1-k} e^{kz} \left(\frac{1}{e^w + e^{-z}} \right)^{\eta-1-k} \end{aligned}$$

The series is ignored if $n = i$. At $w = \infty$ the first term goes to $\ln(1)$, which is zero, and the exponential in the second term's denominator drives it to zero, because the exponent $\eta - 1 - k$ is positive. At $w = 0$ the exponentials go to one, leaving

$$I_{in} = \frac{-1}{\eta-1} \left\{ -e^{(\eta-1)z} \ln\left(1 + e^{-z}\right) + \sum_{k=1}^{\eta-2} \frac{1}{\eta-1-k} e^{kz} \left(\frac{1}{1 + e^{-z}} \right)^{\eta-1-k} \right\}$$

If we substitute back η we have finally

$$I_{in} = \frac{1}{n-i+1} \left\{ e^{(n-i+1)z} \ln\left(1 + e^{-z}\right) - \sum_{k=1}^{n-i} \frac{1}{n-i+1-k} e^{kz} \left(\frac{1}{1 + e^{-z}} \right)^{n-i+1-k} \right\} \tag{L.3}$$

Evaluating the outer integral,

$$\begin{aligned} E\Big\{D_i\Big\} &= \frac{\sigma n!}{(i-2)!(n-i)!} \int_{-\infty}^\infty \left(\frac{1}{1 + e^{-z}} \right)^i e^{-(n-i+2)z} I_{in} dz \\ &= \frac{\sigma n!}{(i-2)!(n-i+1)!} \int_{-\infty}^\infty e^{-z} \left(\frac{1}{1 + e^{-z}} \right)^i \ln\left(1 + e^{-z}\right) dz \\ &\quad - \frac{\sigma n!}{(i-2)!(n-i+1)!} \sum_{k=1}^{n-i} \frac{1}{n-i+1-k} \int_{-\infty}^\infty e^{-(n-i+2-k)z} \left(\frac{1}{1 + e^{-z}} \right)^{n+1-k} dz \end{aligned}$$

To evaluate the first integral, transform $t = 1 + e^{-z}$ and $dt = -e^{-z}dz$, which changes the integration limits, giving

$$\int_{-\infty}^{\infty} e^{-z}\left(\frac{1}{1+e^{-z}}\right)^i \ln\left(1+e^{-z}\right) dz = \int_{\infty}^{1} -\left(\frac{1}{t}\right)^i \ln t \, dt = \int_1^{\infty} t^{-i}\ln t\, dt$$

which we integrate by parts with $u = \ln t$, $du = dt/t$, $dv = t^{-i}dt$, and $v = t^{-i+1}/(-i+1)$. This means

$$\begin{aligned}
\int_1^{\infty} t^{-i}\ln t\, dt &= \left[\frac{1}{-i+1}t^{-i+1}\ln t\right]_1^{\infty} - \int_1^{\infty}\frac{1}{-i+1}t^{-i+1}\frac{1}{t}dt \\
&= \left[\frac{1}{-i+1}t^{-i+1}\ln t\right]_1^{\infty} - \int_1^{\infty}\frac{1}{-i+1}t^{-i}dt \\
&= \left[\frac{1}{-i+1}t^{-i+1}\ln t\right]_1^{\infty} - \left[\left(\frac{1}{-i+1}\right)^2 t^{-i+1}\right]_1^{\infty} \\
&= \left[\frac{1}{-i+1}t^{-i+1}\ln t - \frac{1}{(-i+1)^2}t^{-i+1}\right]_1^{\infty} \\
&= \left(\frac{1}{-i+1}\right)^2
\end{aligned}$$

because at the upper limit the t^{-i+1} goes to 0 (remember, $i >= 2$), and at the lower the t dependence disappears.

The second integral requires another known formula, [1, (3.314)]

$$\int_{-\infty}^{\infty} \frac{e^{-\mu x}}{\left(e^{\beta/\gamma}+e^{-x/\gamma}\right)^{\nu}}dx = \gamma e^{\beta(\mu-\frac{\nu}{\gamma})}B(\gamma\mu, \nu-\gamma\mu)$$

We have $\mu = n-i+2-k$, $\beta = 0$, $\gamma = 1$, and $\nu = n+1-k$. The integral has a number of conditions, all of which are met: $|\mathrm{Im}(\beta)| = 0 < \pi\,\mathrm{Re}(\gamma) = \pi$; $\mathrm{Re}(\nu/\gamma) = n+1-k > \mathrm{Re}(\mu) = n-i+2-k$, which holds since $i \geq 2$; and $\mathrm{Re}(\mu) = n-i+2-k > 0$, which is true at the largest $k = n-i$. With $\beta = 0$ the exponential vanishes and the integral is

$$\int_{-\infty}^{\infty} e^{-(n-i+2-k)z}\left(\frac{1}{1+e^{-z}}\right)^{n+1-k} dz = B(n-i+2-k, i-1)$$

Combining the two and expanding the beta function, then dividing out the $n-i+1-k$ factor before the second integral, gives

$$E\Big\{D_i\Big\} = \frac{\sigma n!}{(i-2)!(n-i+1)!}\left\{\left(\frac{1}{i-1}\right)^2 - \sum_{k=1}^{n-i}\frac{(n-i-k)!(i-2)!}{(n-k)!}\right\} \tag{L.4}$$

Matching to Quantile Estimator

To show this matches the derivative of the inverse c.d.f. we first want to reduce the factorials.

$$
\begin{aligned}
E\Big\{D_i\Big\} &= \frac{\sigma n!}{(i-2)!(n-i+1)!}\left\{\left(\frac{1}{i-1}\right)^2 - \sum_{k=1}^{n-i}\frac{(n-i-k)!(i-2)!}{(n-k)!}\right\} \\
&= \frac{\sigma n!}{(i-2)!(n-i+1)!}\left\{\left(\frac{1}{i-1}\right)^2 - \sum_{k=1}^{n-i}\frac{(n-i-k)!}{\prod_{j=k}^{n-i+1}(n-j)}\right\} \\
&= \frac{\sigma n!}{(i-2)!(n-i+1)!}\frac{1}{i-1}\left\{\frac{1}{i-1} - \sum_{k=1}^{n-i}\frac{(n-i-k)!}{\prod_{j=k}^{n-i}(n-j)}\right\} \\
&= \frac{\sigma n!}{(i-1)!(n-i+1)!}\frac{\prod_{j=1}^{n-i}(n-j)}{\prod_{j=1}^{n-i}(n-j)}\left\{\frac{1}{i-1} - \sum_{k=1}^{n-i}\frac{(n-i-k)!}{\prod_{j=k}^{n-i}(n-j)}\right\} \\
&= \frac{\sigma n!}{(n-1)!(n-i+1)!}\left\{\frac{\prod_{j=1}^{n-i}(n-j)}{i-1} - \sum_{k=1}^{n-i}(n-i-k)!\prod_{j=1}^{k-1}(n-j)\right\} \\
&= \frac{\sigma n}{(i-1)(n-i+1)!}\left\{\prod_{j=1}^{n-i}(n-j) - (n-i-1)!(i-1) - \sum_{k=2}^{n-i}(i-1)(n-i-k)!\prod_{j=1}^{k-1}(n-j)\right\}
\end{aligned}
$$

In the third line we have reduced the upper product limit by factoring out $(i-1)$, which combines in the fourth step with $(i-2)!$. In the fifth line we have combined the product introduced to the denominator with the $(i-1)!$ factor to get $(n-1)!$, while the one added to the numerator cancels the upper factors within the series, shifting the indices on its product. In the sixth line we have separated the $k=1$ term and factored out $i-1$ in the denominator. Combining the first product and the last in the series, $k=n-i$,

$$
\begin{aligned}
&\left\{\prod_{j=1}^{n-i}(n-j)\right\} - \left\{(i-1)(n-i-(n-i))!\prod_{j=1}^{n-i-1}(n-j)\right\} \\
&\qquad = \big[i-(i-1)\big]\prod_{j=1}^{n-i-1}(n-j) = 1\prod_{j=1}^{n-i-1}(n-j)
\end{aligned}
$$

Using this result as the new first term and matching the series at $k=n-i-1$,

$$
\begin{aligned}
&\left\{1\prod_{j=1}^{n-i-1}(n-j)\right\} - \left\{(i-1)(n-i-(n-i-1))\prod_{j=1}^{n-i-2}(n-j)\right\} \\
&\qquad = \big[(i+1)-(i-1)\big]\prod_{j=1}^{n-i-2}(n-j) = 2!\prod_{j=1}^{n-i-2}(n-j)
\end{aligned}
$$

Now the third step with $k=n-i-2$,

$$
\begin{aligned}
&\left\{2!\prod_{j=1}^{n-i-2}(n-j)\right\} - \left\{(i-1)(n-i-(n-i-2))!\prod_{j=1}^{n-i-3}(n-j)\right\} \\
&\qquad = 2!\big[(i+2)-(i-1)\big]\prod_{j=1}^{n-i-3}(n-j) = 3!\prod_{j=1}^{n-i-3}(n-j)
\end{aligned}
$$

Continue down to $k=1$. At this point we combine our new first term and the second term in the final $E\Big\{D_i\Big\}$ equation.

$$
\begin{aligned}
&\Big\{(n-1)(n-i-1)!\Big\} - \Big\{(i-1)(n-i-1)!\Big\} \\
&\qquad = (n-i)(n-i-1)! = (n-i)!
\end{aligned}
$$

Multiplying by the pre-factor we get

$$E\Big\{D_i\Big\} = \frac{\sigma n}{(i-1)(n-i+1)!}(n-i)! = \frac{\sigma n}{(i-1)(n-i+1)} \tag{L.5}$$

This equation matches $\widetilde{E}\left\{D_i\right\}$ in the main text.

Example of Matching

An example might make this clearer. Let $i = n-5$. Expanding (L.4) and multiplying terms to remove the denominators, we have

$$\begin{aligned}
E\Big\{D_i\Big\} &= \frac{\sigma n!}{6!(n-7)!}\left\{\left(\frac{1}{n-6}\right)^2 - \sum_{k=1}^{5}\frac{(5-k)!(n-7)!}{(n-k)!}\right\} \\
&= \frac{\sigma n!}{6!(n-1)!}\prod_{j=1}^{6}(n-j)\left\{\left(\frac{1}{n-6}\right)^2 - \sum_{k=1}^{5}\frac{(5-k)!(n-7)!}{(n-k)!}\right\} \\
&= \frac{\sigma n}{6!(n-6)}\prod_{j=1}^{6}(n-j)\left\{\left(\frac{1}{n-6}\right) - \sum_{k=1}^{5}\frac{(5-k)!(n-7)!(n-6)}{(n-k)!}\right\} \\
&= \frac{\sigma n}{6!(n-6)}\left\{\prod_{j=1}^{5}(n-j) - \sum_{k=1}^{5}\frac{(5-k)!(n-1)!(n-6)}{(n-k)!}\right\} \\
&= \frac{\sigma n}{6!(n-6)}\left\{\begin{array}{r}
(n-1)(n-2)(n-3)(n-4)(n-5) \\
-4!(n-6) \\
-3!(n-1)(n-6) \\
-2!(n-1)(n-2)(n-6) \\
-1!(n-1)(n-2)(n-3)(n-6) \\
-0!(n-1)(n-2)(n-3)(n-4)(n-6)
\end{array}\right\}
\end{aligned}$$

The pairs of the first and last terms, working from bottom to top, reduce one by one to

$$\begin{aligned}
\Big[(n-5) - 0!(n-6)\Big](n-1)\ldots(n-4) &= 1!(n-1)\ldots(n-4) \\
\Big[1!(n-4) - 1!(n-6)\Big](n-1)\ldots(n-3) &= 2!(n-1)\ldots(n-3) \\
\Big[2!(n-3) - 2!(n-6)\Big](n-1)(n-2) &= 3!(n-1)(n-2) \\
\Big[3!(n-2) - 3!(n-6)\Big](n-1) &= 4!(n-1) \\
4!(n-1) - 4!(n-6) &= 5!
\end{aligned}$$

So

$$E\Big\{D_i\Big\} = \frac{\sigma n}{6!(n-6)}5! = \frac{\sigma n}{6(n-6)}$$

which is the same as (L.5).

SPACING FOR GUMBEL VARIATES

Density Function

Starting with the distribution's density functions and using $z = e^{-(x+y-\mu)/\sigma}$, $dz = -(z/\sigma)dx$, and $w = e^{y/\sigma}$, so that $wz = e^{-(x-\mu)/\sigma}$,

$$f(x) = \frac{wz}{\sigma}e^{-wz} \qquad f(x+y) = \frac{z}{\sigma}e^{-z}$$
$$F(x) = e^{-wz} \qquad F(x+y) = e^{-z}$$

Including y in the z substitution simplifies $F(x+y)$, which in turn will simplify the form of the spacing's density function. The integration limits change from $x = [-\infty, +\infty]$ to $z = [+\infty, 0]$.

$$\begin{aligned}
f_{D_i}(y) &= \frac{n!}{(i-2)!(n-i)!}\int_{\infty}^{0} \left(e^{-wz}\right)^{i-2}\left(1-e^{-z}\right)^{n-i}\left(\frac{wz}{\sigma}e^{-wz}\right)\left(\frac{z}{\sigma}e^{-z}\right)\left(-\frac{\sigma}{z}dz\right) \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{w}{\sigma}\int_{0}^{\infty} e^{-(w(i-1)+1)z}z\left(1-e^{-z}\right)^{n-i}dz \\
&= \frac{n!}{(i-2)!(n-i)!}\frac{w}{\sigma}(-1)^{n-i}\int_{0}^{\infty} ze^{-(w(i-1)+1)z}\left(e^{-z}-1\right)^{n-i}dz
\end{aligned}$$

Using [1, (3.432.1)]

$$\int_0^\infty x^{\nu-1}e^{-mx}\left[e^{-x}-1\right]^p dx = \Gamma(\nu)\sum_{k=0}^{p}(-1)^k\binom{p}{k}\frac{1}{(p+m-k)^\nu}$$

with $\nu = 2$, $m = w(i-1)+1$, and $p = n-i$, we get directly

$$f_{D_i}(y) = \frac{n!}{(i-2)!(n-i!)}\frac{e^{y/\sigma}}{\sigma}(-1)^{n-i}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{\left(e^{y/\sigma}(i-1)+n-i+1-k\right)^2} \tag{G.1}$$

Expected Spacing

Letting $w = y/\sigma$ the expected spacing is

$$\begin{aligned}
E\left\{D_i\right\} &= \int_0^\infty yf_{D_i}(y)dy \\
&= \int_0^\infty \frac{n!}{(i-2)!(n-i)!}\frac{y}{\sigma}e^{y/\sigma}(-1)^{n-i}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{\left(e^{y/\sigma}(i-1)+n-i+1-k\right)^2}dy \\
&= \frac{n!}{(i-2)!(n-i)!}(-1)^{n-i}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\int_0^\infty we^w\frac{1}{\left(e^w(i-1)+n-i+1-k\right)^2}\sigma dw \\
&= \frac{n!}{(i-2)!(n-i)!}(-1)^{n-i}\frac{\sigma}{(i-1)^2}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\int_0^\infty \frac{we^w}{(e^w+\alpha)^2}dw
\end{aligned}$$

making the substitution $\alpha = (n-i+1-k)/(i-1)$ for convenience. The integral is done by parts, with $u = w$, $du = dw$, and

$$\begin{aligned}
dv &= \frac{e^w}{\left(e^w+\alpha\right)^2}dw \\
v &= -\frac{1}{e^w+\alpha}
\end{aligned}$$

so that

$$\int_0^\infty \frac{we^w}{\left(e^w+\alpha\right)^2}dw = \left[-\frac{w}{e^w+\alpha}+\int_0^\infty\frac{dw}{e^w+\alpha}\right]_0^\infty$$

The integral has the known form [1, (2.313.1)]

$$\int \frac{dx}{a+be^{mx}} = \frac{1}{am}\left[mx - \ln\left(a+be^{mx}\right)\right]$$

with $a=\alpha$, $b=1$, and $m=1$. Substituting back,

$$\begin{aligned}
\int_0^\infty \frac{we^w}{\left(e^w+\alpha\right)^2}dw &= \left[-\frac{w}{e^w+\alpha} + \frac{1}{\alpha}\Big\{w - \ln\left(e^w+\alpha\right)\Big\}\right]_0^\infty \\
&= \left[\frac{we^w}{\alpha\left(e^w+\alpha\right)} - \frac{1}{\alpha}\ln\left(e^w+\alpha\right)\right]_0^\infty \\
&= \left[\frac{w}{\alpha\left(1+\alpha e^{-w}\right)} - \frac{1}{\alpha}\ln\left(e^w+\alpha\right)\right]_0^\infty
\end{aligned}$$

As $w\to\infty$ both terms are equal and cancel, and at $w=0$ only the logarithm remains, giving

$$\int_0^\infty \frac{we^w}{\left(e^w+\alpha\right)^2}dw = \frac{1}{\alpha}\ln(\alpha+1) = \frac{i-1}{n-i+1-k}\ln\frac{n-k}{i-1}$$

Then

$$\begin{aligned}
E\Big\{D_i\Big\} &= \frac{n!}{(i-2)!(n-i)!}(-1)^{n-i}\frac{\sigma}{(i-1)^2}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{i-1}{n-i+1-k}\ln\left(\frac{n-k}{i-1}\right) \\
&= \frac{n!}{(i-2)!(n-i)!}(-1)^{n-i}\frac{\sigma}{i-1}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{n-i+1-k}\ln\left(\frac{n-k}{i-1}\right) \qquad \text{(G.2)}
\end{aligned}$$

which is the result in the main text.

To simplify this to the final form, first we can re-write the factorials.

$$\frac{n!}{(i-2)!(n-i)!}\frac{1}{i-1} = \frac{in!}{i!(n-i)!} = i\binom{n}{i}$$

Separating the logarithm and using $m=n-i$ for the second series,

$$\begin{aligned}
E\Big\{D_i\Big\} =& i\binom{n}{i}(-1)^{n-i}\sigma\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{n-i+1-k}\ln(n-k) \\
&- i\binom{n}{i}(-1)^{n-i}\sigma\sum_{k=0}^{m}(-1)^k\binom{m}{k}\frac{1}{m-k+1}\ln(i-1)
\end{aligned}$$

We want to put the second series in a standard form with a known value [1, (0.155.1)]

$$\sum_{k=1}^{n}(-1)^{k+1}\frac{1}{k+1}\binom{n}{k} = \frac{n}{n+1}$$

We do this by pulling out the first term, substituting $k'=m-k$, and adding in an extra, last term to the the sum.

$$\sum_{k=0}^{m}(-1)^k\binom{m}{k}\frac{1}{m-k+1}\ln(i-1)$$

$$
\begin{aligned}
&= \ln(i-1)\left\{\frac{1}{m+1}+\sum_{k=1}^{m}(-1)^k\frac{m!}{k!(m-k)!}\frac{1}{m-k+1}\right\}\\
&= \ln(i-1)\left\{\frac{1}{m+1}+\sum_{k'=m-1}^{0}(-1)^{m-k'}\frac{m!}{(m-k')!k'!}\frac{1}{k'+1}\right\}\\
&= \ln(i-1)\left\{\frac{1}{m+1}+(-1)^m\sum_{k'=0}^{m-1}(-1)^{k'}\binom{m}{k'}\frac{1}{k'+1}\right\}\\
&= \ln(i-1)\left\{\frac{1}{m+1}+(-1)^m+(-1)^m\sum_{k'=1}^{m-1}(-1)^{k'}\binom{m}{k'}\frac{1}{k'+1}\right\}\\
&= \ln(i-1)\left\{\frac{1}{m+1}+(-1)^m-(-1)^{2m}\binom{m}{m}\frac{1}{m+1}+(-1)^m\sum_{k'=1}^{m}(-1)^{k'}\binom{m}{k'}\frac{1}{k'+1}\right\}\\
&= \ln(i-1)\left\{(-1)^m-(-1)^m\sum_{k'=1}^{m}(-1)^{k'+1}\binom{m}{k'}\frac{1}{k'+1}\right\}\\
&= \ln(i-1)(-1)^m\left\{1-\frac{m}{m+1}\right\}\\
&= \ln(i-1)(-1)^m\frac{1}{m+1}\\
&= \ln(i-1)(-1)^{n-i}\frac{1}{n-i+1}
\end{aligned}
$$

Substituting back,

$$E\Big\{D_i\Big\} = i\binom{n}{i}(-1)^{n-i}\sigma\left\{\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{n-i+1-k}\ln(n-k)-(-1)^{n-i}\frac{1}{n-i+1}\ln(i-1)\right\}$$

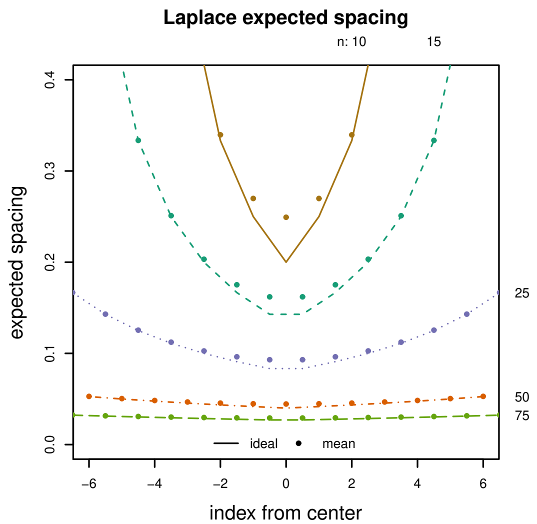
To re-write the remaining series, make the substitution $k' = n-i-k$ so that

$$\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{n-i+1-k}\ln(n-k)$$

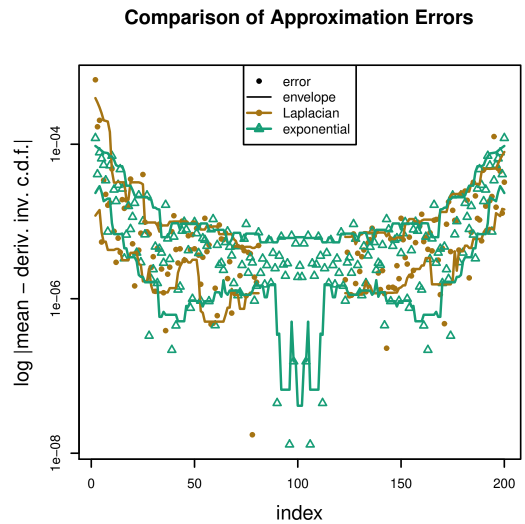
$$
\begin{aligned}
&= \sum_{k'=n-i}^{0}(-1)^{n-i-k'}\binom{n-i}{n-i-k'}\frac{1}{k'+1}\ln(i+k')\\
&= (-1)^{n-i}\sum_{k'=0}^{n-i}(-1)^{k'}\binom{n-i}{k'}\frac{1}{k'+1}\ln(i+k')
\end{aligned}
$$

The final result is

$$
\begin{aligned}
E\Big\{D_i\Big\} &= i\binom{n}{i}\sigma(-1)^{n-i}\left\{(-1)^{n-i}\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{1+k}\ln(i+k)-(-1)^{n-i}\frac{1}{n-i+1}\ln(i-1)\right\}\\
&= i\binom{n}{i}\sigma\left\{\sum_{k=0}^{n-i}(-1)^k\binom{n-i}{k}\frac{1}{1+k}\ln(i+k)-\frac{1}{n-i+1}\ln(i-1)\right\} \qquad \text{(G.3)}
\end{aligned}
$$



Supplemental Figure 1: Laplace spacing at mid-range indices for small n, showing the smooth (rounded) behavior of the mean compared to the estimator.



Supplemental Figure 2: Difference between mean spacing and quantile estimator for the exponential and Laplace distributions, showing the two are similar in the tails at small and large indices.

SUPPLEMENTAL ANALYSIS

Laplace Distribution

The discussion about the approximation error for the Laplace distribution in the main text attributes the larger error at the middle indices to rounding off the two exponentials pasted together. Supplemental Figure 1 shows the mean spacing (points) and quantile estimator (lines) for small n, where the effect can be seen. At small and large indices the Laplace behavior matches the exponential. Supplemental Figure 2 plots the approximation error, which is very noisy, and an envelope of the points at the 10% and 90% quantiles. The two envelopes are roughly the same away from the rounding in the middle. The sample size was $n = 200$ and 10 million trials generated the mean spacings.

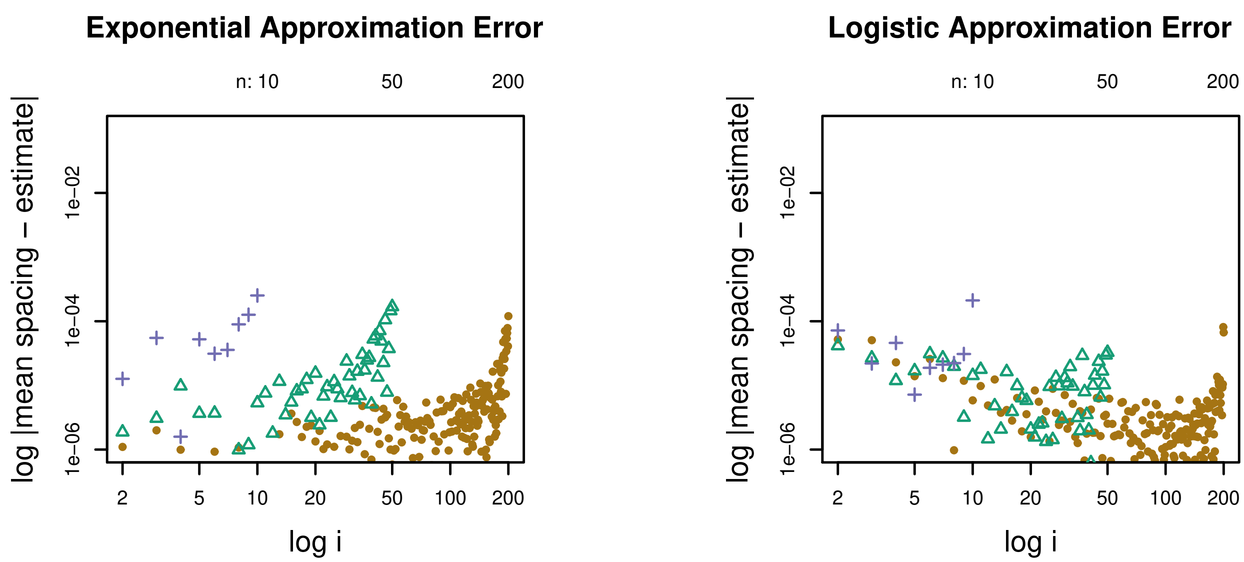
Exponential and Logistic Distributions

We have exact equations for the expected spacing of the exponential and logistic distributions. Figure 3 shows the approximation error under the same simulation conditions found in the main text. This tells us that the inherent uncertainty in the charts is 5 to 6 decades from the actual expected spacing. The tails, to the right in the exponential results and on both sides for the logistic, have a higher error, 4 decades from the actual. The simulated mean is not consistently higher or lower than the estimator, an indication that we are seeing noise.

Regression of Generic Quantile Estimator

The main text discusses corrections to the quantile estimator, either by modifying the derivative equation or changing the location p where it is evaluated. This section explores the first possibility by creating a generic form of the estimator and regressing simulated draws while varying the distribution parameters or draw size to see where the estimate fails. The next section will try the second option.

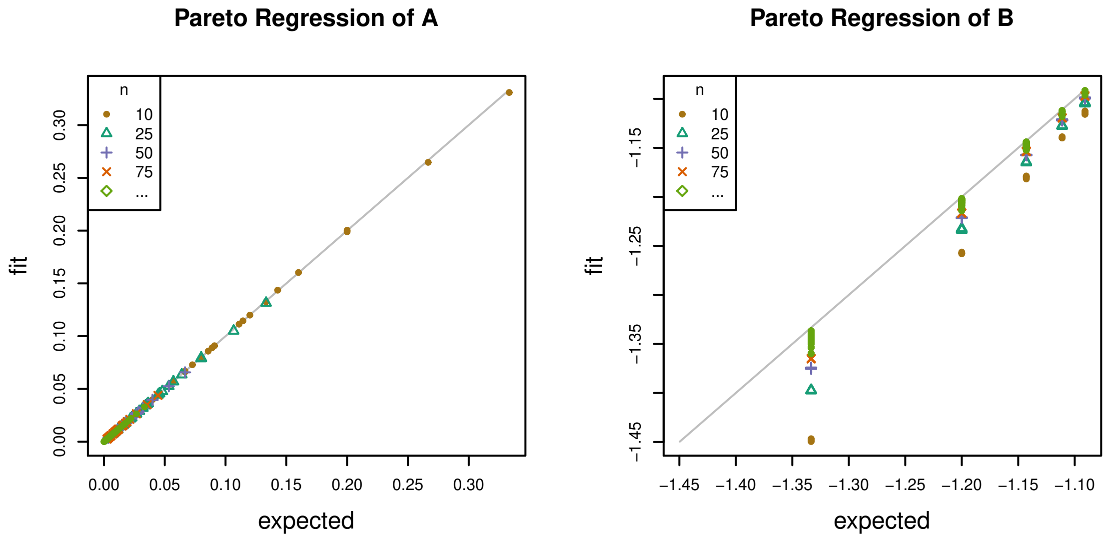
Supplemental Table 1 has the generic forms of the expected spacing and a mapping to distribution parameters, where the quantile estimator column includes the $\Delta p = 1/n$ factor but does not substitute $p = (i-1)/n$. For combinations of the sample size n and various values of each distribution parameter, 100 thousand trial draws were made and the expected spacing averaged at each index. The following graphs show the results, with the expected values of the regression variable on the x axis and the fit value on the y, for each combination. Points should lie on the 45 degree line. Points are colored for the first few n, with $n \geq 100$ lumped together; most graphs will show different behavior (color clusters) as n varies. The default set of values for n is { 10, 25, 50, 75, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900, 1000 }.



Supplemental Figure 3: Difference between mean spacing and quantile estimator for the exponential and logistic distributions.

Supplemental Table 1: Generic Quantile Estimators

	quantile estimator	regressed equation	expected values
Pareto	$\frac{b}{an}(1-p)^{-\frac{a+1}{a}}$	$A(1-p)^B$	$A = b/an; B = -(a+1)/a$
Gumbel	$\frac{\sigma}{n}\frac{1}{p}\frac{1}{-\ln(p)}$	$Ap^B \ln(p)^C$	$A = \sigma/n; B = -1; C = -1$
Rayleigh	$\frac{\sigma}{n}\frac{1}{1-p}\frac{1}{\sqrt{-2\ln(1-p)}}$	$A(1-p)^B \ln(1-p)^C$	$A = \sigma/n\sqrt{2}; B = -1; C = -1/2$
Weibull	$\frac{b}{an}\frac{1}{1-p}\left[-\ln(1-p)\right]^{\frac{1-a}{a}}$	$A(1-p)^B \ln(1-p)^C$	$A = b/an; B = -1; C = (1-a)/a$
Frechet	$\frac{\sigma}{\lambda n}\frac{1}{p}\left[-\ln(p)\right]^{-\frac{\lambda+1}{\lambda}}$	$Ap^B \ln(p)^C$	$A = \sigma/\lambda n; B = -1; C = -(\lambda+1)/\lambda$
Cauchy	$\frac{\pi\sigma}{n}\sec^2\pi(p-\frac{1}{2})$	$A\left[\sec\pi(p-\frac{1}{2})\right]^B$	$A = \pi\sigma/n; B = 2$



Supplemental Figure 4: Regression of the Pareto generic quantile estimator.

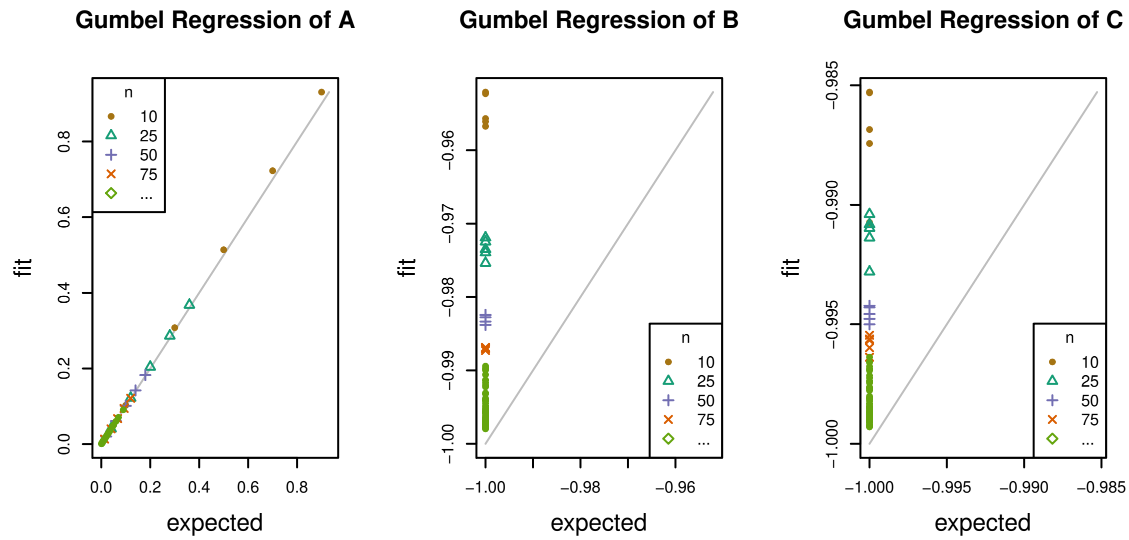
For example, there are two Pareto graphs in Supplemental Figure 4. On the left the scaling parameter A fits the ideal value b/na well; this part of the quantile estimator is correct. The power B on the right, however, does not match $-(a+1)/a$ for small n by up to 10%. It approaches the ideal as n grows, but the power also varies with a, with a larger mismatch as a grows. For example, with $a = 3$ the predicted power is -1.33, but at $n = 10$ the fit gives -1.45, at $n = 50$ -1.37, and at $n = 200$ -1.35. For $a = 9$ the predicted power is -1.111 but the fits are -1.140, -1.121, and -1.116 respectively. A correction to B would be a complicated function of a and n. The data was generated with all combinations of the default set of n plus { 1500, 2000 }, for $a = \{ 3, 5, 7, 9, 11 \}$, and for $b = \{ 2, 4, 6, 8, 10 \}$.

The Gumbel regression (Supplemental Figure 5) shows the problem is again with the power of terms, specifically the $1/p$ (B) factor. This differs from the expected value by up to 5%, and there's a clear dependence on the sample size, with the estimator drawing closer to the mean spacing as n increases. Within each cluster colored per n there is no regular dependence on the other variable, σ (not shown in these graphs). This last statement means there is no consistent ordering by σ within the cluster, the uniformity of points within the cluster changes, and the cluster spread changes erratically. The power of the $\ln p$ (C) factor is more stable, within 1.5% of the expected value of -1, although again the dependence on n seems complicated, with the cluster size changing and no regular variation with σ. The scaling parameter A is a little larger than the ideal value σ/n, up to 3.5%; the points are slightly above the ideal line. These results are based on regressions of combinations of the default set of sample sizes and σ taking the values { 1, 3, 5, 7, 9 }.

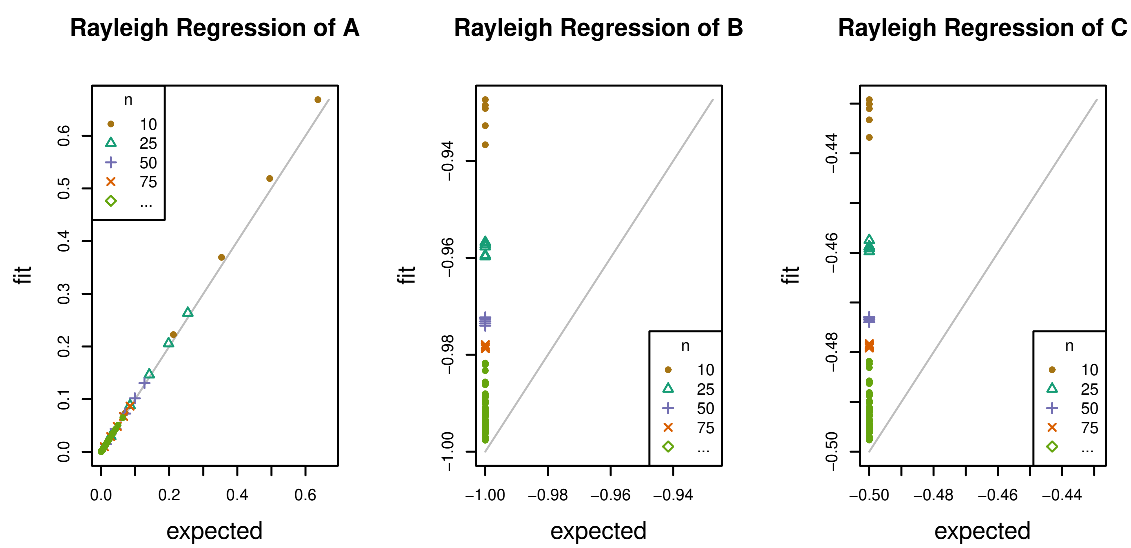
Qualitatively, the Rayleigh regressions are similar to the Gumbel (Supplemental Figure 6). The scaling parameter A leads the ideal value $\sigma/n\sqrt{2}$ but by a bit more, deviating by up to 5% instead of 3.5%, and the power parameter B of the $1/(1-p)$ factor differs from -1 by up to 7%. It depends on the sample size with the same kind of variability that is present with the Gumbel. The power parameter C of the $1/\ln(1-p)$ factor is much less stable, though, differing from -0.5 by up to 14%. That the results are similar should be expected, given the two distributions differ essentially by the square of x in the density function. The regression was done on a combination of the default set of sample sizes plus { 1500, 2000 }, and σ from { 1, 3, 5, 7, 9 }.

These same comments apply to the Weibull regressions (Supplemental Figure 7), with somewhat higher deviations, up to 10% for scaling parameter A, 8% for $1/(1-p)$ power parameter B, and 14% for $1/\ln(1-p)$ power parameter C. That last parameter depends now on the distribution parameters, separating by a and still getting closer to the expected value as the sample size increases. The trend over n is similar for each a, moving in parallel closer to the 45 degree matching line. The fits were made for combinations of the default sample sizes, a in { 1, 3, 5, 7, 9 }, and b in { 2, 3, 4, 5, 7, 9 }.

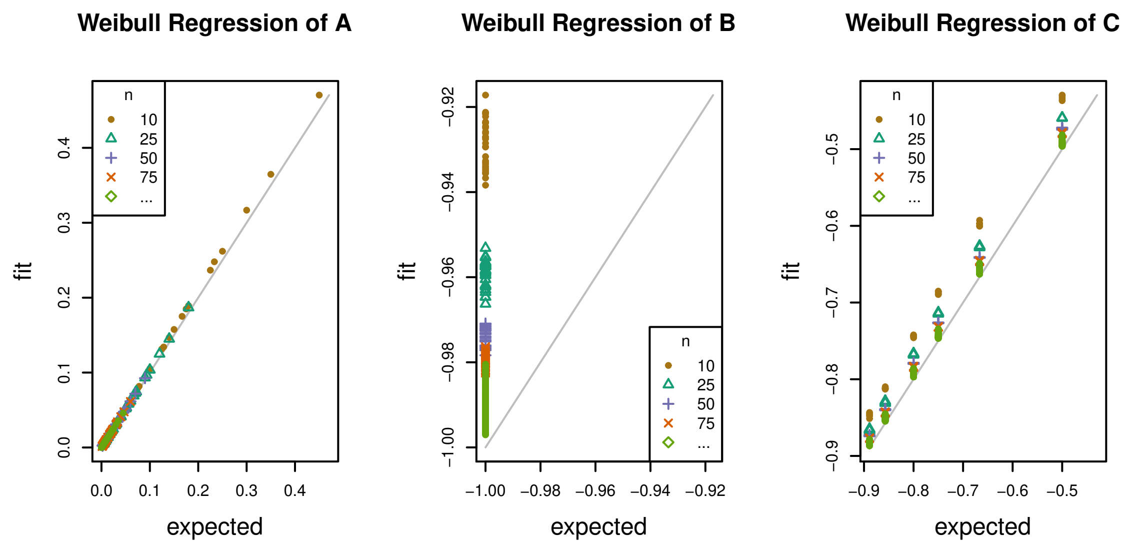
The Frechet, although similar to the other distributions in the group, has very different results (Supplemental Figure 8). The regression fits of all three parameters are worse. These are generated from combinations of the default sample sizes plus { 1500 and 2000 }, λ taking the values { 2, 3, 4, 5, 6, 7, 8, 9 }, and σ { 1/5, 1/4, 1/3, 1/2, 1, 3, 5, 7, 9 }. The scaling parameter A lags the ideal value $\sigma/\lambda n$ for small n by up to 17%. The deviation from the 45 degree line increases as λ decreases; if we had included values for $\lambda \leq 1$ A would lag even more, with



Supplemental Figure 5: Regression of the Gumbel generic quantile estimator.



Supplemental Figure 6: Regression of the Rayleigh generic quantile estimator.



Supplemental Figure 7: Regression of the Weibull generic quantile estimator.

the fit value going to 0. In the power parameter B of the $1/p$ factor, notice that there are fit values above and below the 45 degree line for a given sample size. This may reflect the asymmetry in the approximation error, where the quantile estimator is greater than the mean spacing for small indices but smaller for large. The power parameter C of the $\ln p$ factor behaves similarly to the power parameter of the Pareto distribution. Although the fit gets better (closer to the 45 degree line) as n increases, the spread shrinks as λ increases and the relationship between the distribution parameters and sample size governing this shrinkage is not clear.

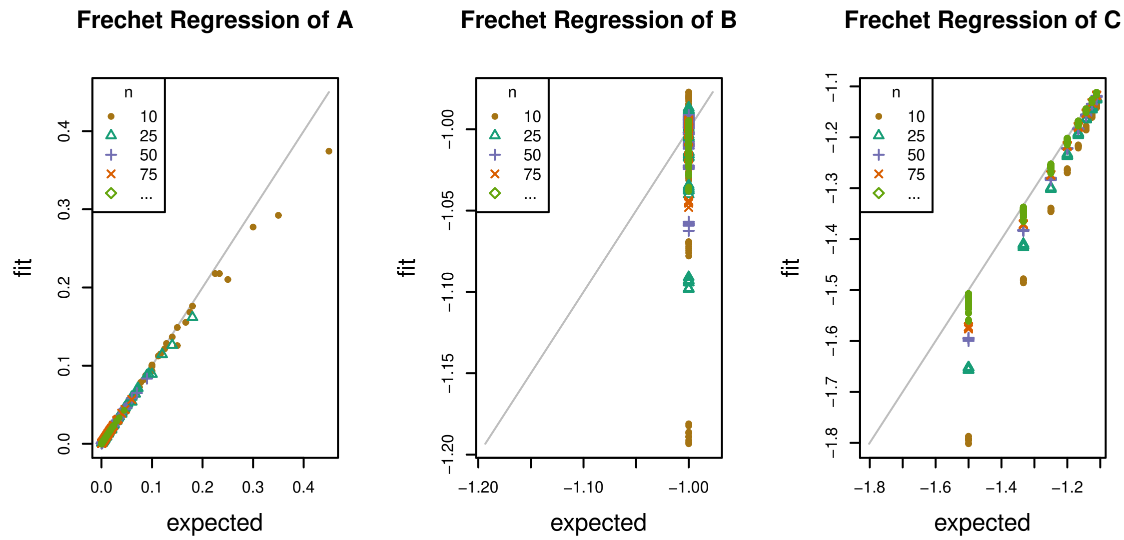
Supplemental Figure 9 shows the Cauchy fits. The fits were made for combinations of the default sample size and { 1200, 1400, 1600, 1800, 2000 }, and σ in { 1, 3, 5, 7 }. Here the scaling parameter A lags the ideal $\pi\sigma/n$ by up to 18%. The power parameter B for the secant is substantially wrong, by a factor of 1.25, but gets better as the sample size increases.

Overall, the scaling factors in the quantile estimators are close to the expected values, except for the Frechet and Cauchy distributions, or the Pareto where it seems to be exact. However, the exponents of the quantile probability, which is essentially what p is, or its logarithm, clearly do not fit the data. All distributions show a complex interaction between their parameters and the sample size, suggesting there is no simple modification to the estimators to improve the performance.

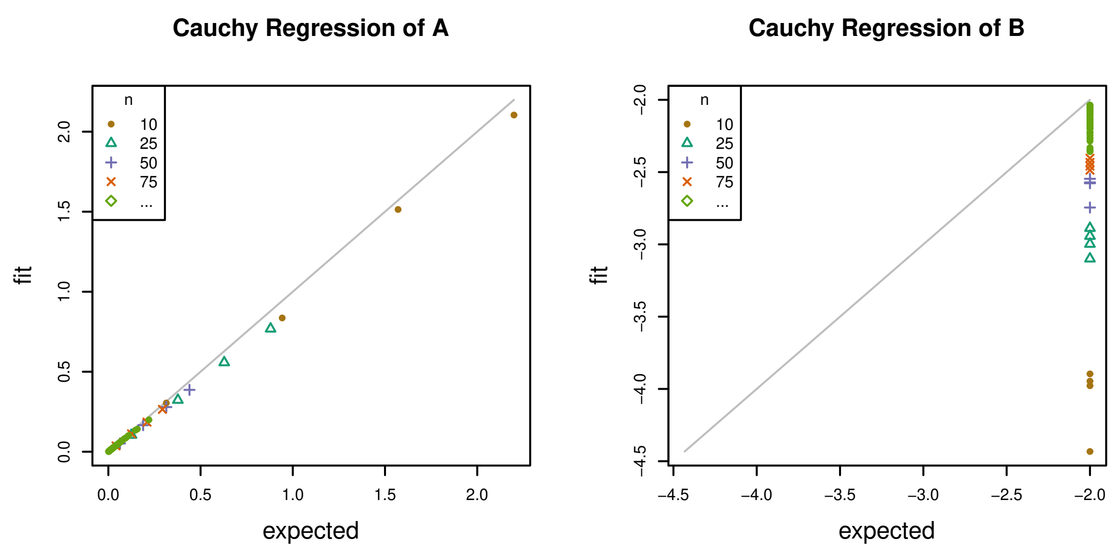
Modeling the Quantile Probability

Another check of the quantile estimators is to ask if we evaluate the derivative at the correct probabilities $p = (i-1)/n$. For the same combinations of sample sizes and distribution parameters that we used to check the generic form of the estimate, we again generate mean spacings from 100 thousand trial draws. This time we assume the equation is correct and calculate the probability for each index and combination from the mean spacing μ (Supplemental Table 2). We take the median p at each i, n and run a linear regression to get $p = mi + b$, with m the slope and b the intercept. The check is whether these values match our standard map. That is, we expect $m = 1/n$ and $b/m = -1$.

There are four difficulties. Except for the one-sided Pareto the inverse spacings are not proper functions of p, yielding indices from each of the tails. We have to pick one of the two, which requires knowing which index we are processing and to which side of the estimator's minimum it lies. This is a buried use of the probability we are trying to analyze. Second, the results near the minimum are sensitive to noise, especially as the sample size grows and the spacings become more equal. We see this in plots (not included here) of the calculated probabilities with the range of values marked, so that the curve is thick and rough at middle indices and thin at small and large indices where there is little variation. Related to this and third, if the minimum mean spacing is well above the estimator then the calculated probability will appear at a shifted index, or even remain fixed to one value. An example of this condition is found in Supplemental Figure 1 in the $n = 10$ curve, where the smallest mean value at the midpoint is close to the estimator at adjacent indices. It is impossible to reach the estimator's minimum and the calculated probability will never take this value. Finally, the distributions that use $p\ln(p)$ will require



Supplemental Figure 8: Regression of the Frechet generic quantile estimator.



Supplemental Figure 9: Regression of the Cauchy generic quantile estimator.

Supplemental Table 2: Probability Models

	estimator	probability
Pareto	$\mu = \frac{b}{a}(1-p)^{-\frac{a+1}{a}}\Delta p$	$p = 1 - \left(\frac{an\mu}{b}\right)^{-\frac{a}{a+1}}$
Cauchy	$\mu = \pi\sigma \sec^2 \pi \left(p - \frac{1}{2}\right)\Delta p$	$p = \frac{1}{2} + \frac{1}{\pi} cos^{-1}\sqrt{\frac{\pi\sigma}{n\mu}}$
Gumbel	$\mu = \sigma \frac{1}{p}\frac{-1}{\ln(p)}\Delta p$	$p \ln p = \frac{-\sigma}{n\mu}$
Rayleigh	$\mu = \sigma \frac{1}{1-p}\left[\frac{-1}{2\ln(1-p)}\right]^{\frac{1}{2}}\Delta p$	$(1-p)^2 \ln(1-p) = -\frac{1}{2}\left(\frac{\sigma}{n\mu}\right)^2$
Weibull	$\mu = \frac{b}{a}\frac{1}{1-p}\left[\frac{-1}{\ln(1-p)}\right]^{\frac{a-1}{a}}\Delta p$	$(1-p)^{\frac{a}{a-1}} \ln(1-p) = -\left(\frac{b}{an\mu}\right)^{\frac{a}{a-1}}$
Frechet	$\mu = \frac{\sigma}{\lambda}\frac{1}{p}\left[\frac{-1}{\ln(p)}\right]^{\frac{\lambda+1}{\lambda}}\Delta p$	$p^{\frac{\lambda}{1+\lambda}} \ln p = -\left(\frac{\sigma}{\lambda n\mu}\right)^{\frac{\lambda}{1+\lambda}}$

numeric root finding. The Pareto and Cauchy models can solve for p directly.

For each distribution we will present two graphs. The first plots b/m against n. It should equal -1 (grey line) and be independent of the sample size. The second plots the difference $(1/m) - n$ against n. It should be 0 and stable. Only if values are near one, or greater, do we need to check if the scaling for p is wrong. This would correspond to the difference that we see between the exponential spacing, which goes with $\Delta p = 1/n$, and the uniform, $\Delta p = 1/(n+1)$.

The Pareto model (Supplemental Figure 10) shows that the n dependence is correct, with a small 0.06 difference in the fit slope to the ideal. The offset to the index, however, is -0.86 instead of -1, and this is reasonably stable over the sample size. There is little variation in the back-calculated probability as the σ parameter changes.

The Gumbel model (Supplemental Figure 11) depends on a slightly bigger sample size than we expect, $n+0.23$, but this is not an off-by-one error. The offset to the index is again -0.85 instead of -1. Both checks are stable over the sample size. There is little variation in the calculated probability as the σ parameter changes. There is more variation for indices near the middle, as mentioned.

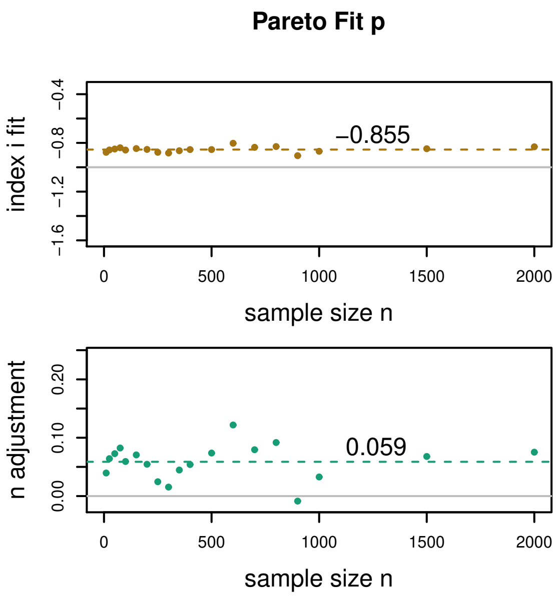
In the Rayleigh results (Supplemental Figure 12) we see a larger shift in n, 0.8, that approaches an off-by-one error. The offset for the index, -0.58, also differs more from -1 than what we have seen. This modified equation for p has the effect of moving the estimation points towards the middle of the distribution, increasing p for small i and decreasing it for large. The change is small, however, and implementing it has little effect on the approximation error. The modifications are largest for the smallest n. It is clear from analyzing results that $p \propto n$ and not $n+1$, however, even for $n = 10$. The calculated probability is stable as the σ parameter changes, and, like the Gumbel, there is more variability for middle indices.

The Weibull results (Supplemental Figure 13) fall between the Gumbel and Rayleigh. They give a model of $p = (i - 0.75)/(n + 0.48)$, with both offsets stable if $n \geq 100$. Again the greatest mismatch is at the smallest sample size. There is more variability in the calculated probability for middle indices near the minimum of the spacing than in the Rayleigh, especially for the smaller sample sizes, but these results are as both the a and b parameters vary.

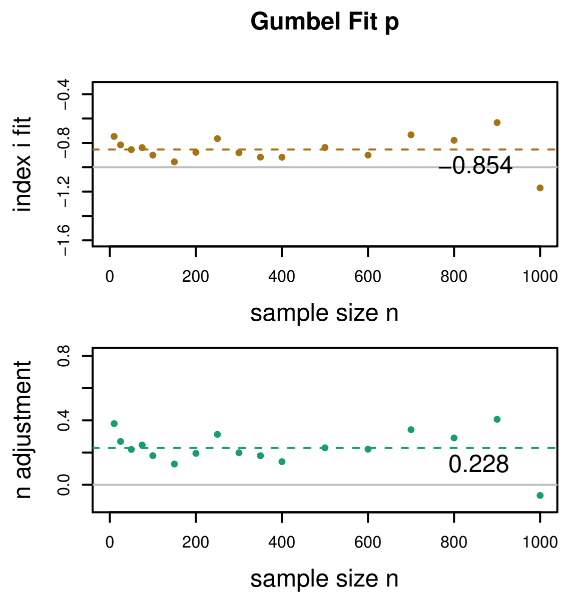
The Frechet model (Supplemental Figure 14) is not good. Both the slope and intercept change significantly as the sample size increases, stabilizing somewhere between 500 and 1000 points. We also see a high variation in the calculated probability as both distribution parameters, λ and σ, vary.

The Cauchy model (Supplemental Figure 15) is stable but differs from the base, giving $p = (i-1.82)/(n-1.7)$. The effect is to lower the probabilities at a given i. This makes the estimated spacing somewhat asymmetric, especially for small sample sizes. The change has little effect on the approximation error. The calculated probability is stable as the σ parameter varies. Because the expected spacing estimate is substantially below the mean, the calculated probability shifts away from the center, most strongly near $n/2$. This creates a gap in the middle where there are no points; they are no longer evenly spaced indices. The shift is balanced to either side so the fit correctly models the average behavior.

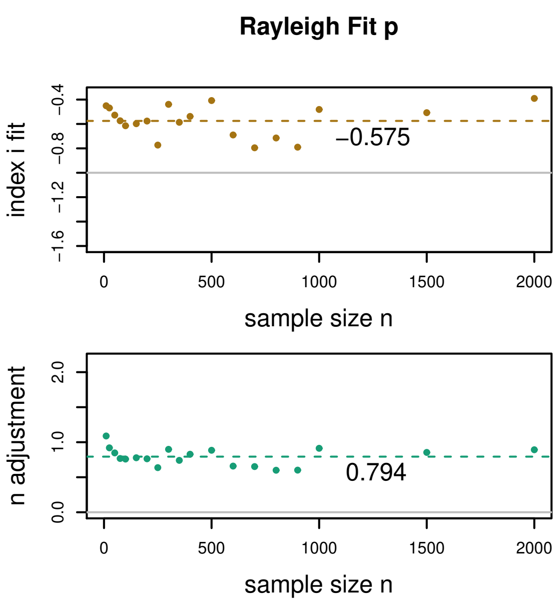
In the end, we can conclude that $p = (i - 1)/n$ is the correct placement of the quantile probabilities. This matches what we saw during development, where the dependence on n is easily checked at small sample sizes and



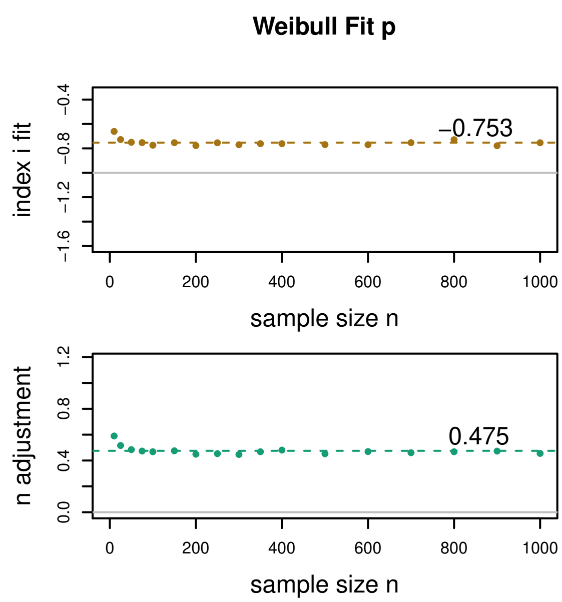
Supplemental Figure 10: Pareto check of probability model



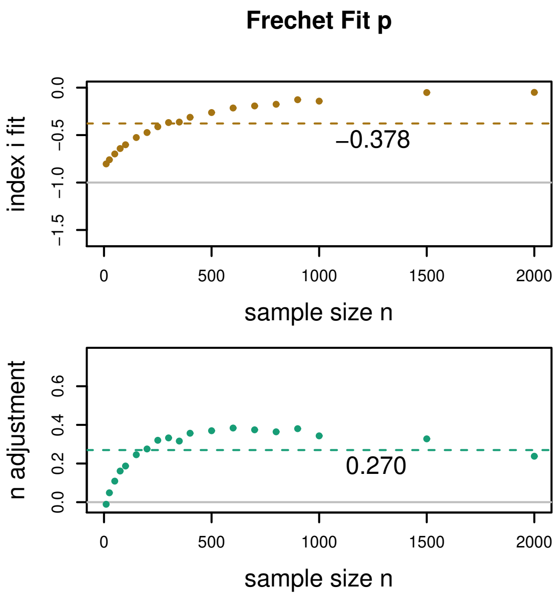
Supplemental Figure 11: Gumbel check of probability model



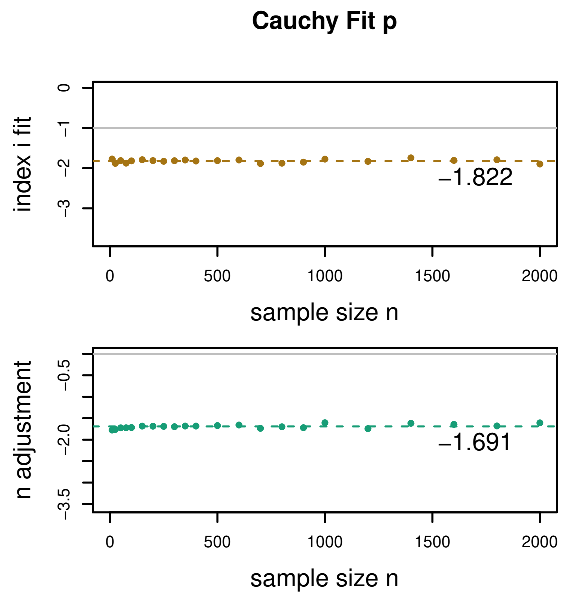
Supplemental Figure 12: Rayleigh check of probability model



Supplemental Figure 13: Weibull check of probability model



Supplemental Figure 14: Frechet check of probability model



Supplemental Figure 15: Cauchy check of probability model

any change is most noticeable. Modifying the placement according to the fits here does not substantially improve the quantile estimator, and certainly does not make it an exact estimate. The average behavior we have looked at here has been made over a large range of sample sizes, but changing the n dependence has the largest effect only at small samples. For all of these distributions, measuring the approximation errors for different denominators makes clear that $p \propto 1/n$.

Supplemental Table 3: Data File Setup

	abbrev	p1	p2
Laplace	a	σ	
Cauchy	c	σ	
exponential	a	λ	
Frechet	f	λ	σ
Gumbel	g	σ	
logistic	l	σ	
Pareto	p	a	b
Rayleigh	r	σ	
Weibull	w	b	a

SUPPLEMENTAL DATA

The files available for this paper include the data used to generate all figures. There are three sets of files.

"data_<abbrev>_<ntrial>.csv" has the mean spacing measured over a very large set of *ntrial* samples drawn for the distribution abbreviated <abbrev>, with the parameters specified in the text (normally the default location, scale, and/or rate). It also has the quantile estimator and an numeric integration of the expected spacing formula. The first column in the file is the spacing index i, then follow three columns for each sample size n, usually the set { 10, 25, 50, 75, 100, 150, 200, 250 }. "n<value>_mean" is the average spacing, "n<value>_ideal" the quantile estimator, and "n<value>_num" the numeric integration. There will be one row for each index, and values for indices that are inapplicable for a given sample size are filled with NA. There will also always be an $i = 1$ row, even though this is never valid; the row is filled with NA's.

"ufit_<abbrev>.csv" contains the regression fits of a generic version of the quantile estimator for distribution <abbrev>. The first column "n" is the sample size and "p1" and "p2" are distribution parameters whose interpretation depends on the distribution (Supplemental Table 3). "b" is the intercept of the regression fit, "m" its slope, and "rsq" the R^2 coefficient. "bexp" is the expected intercept based on the distribution parameters, and "mexp" the expected slope. The table includes a "p3" distribution parameter that is always empty (it is for a location parameter that never varies, because the spacing is independent of the distribution's location).

"pfit_<abbrv>.csv" contains the back-solved p probabilities for the third set of supplementary figures. Indices are the first columns in the table, starting at $i = 1$. As with the data files, values for invalid indices are always NA. The last columns are titled "n" for the sample size, and "p1" and "p2" for the distribution parameters. There is one row per combination of these, and the values are the probabilities corresponding to mean spacings, as per Supplemental Table 2.

References

[1] Gradshteyn, I. S. and Ryzhik, I. M. (1980) *Table of Integrals, Series, and Products.* Academic Press, corrected and enlarged edn.

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