Quantile Estimation by Interpolation

Martin Baxter
Electoral Calculus

17 August 2020

1. Introduction

We often want to estimate a quantile of a particular unknown distribution from empirical data. This will focus on the interpolation case, where the number of samples is large enough to cover the desired quantile, rather than extrapolating a quantile beyond the sample.

This paper will propose a formula for quantile estimation which depends on the tail shape parameter ξ of the distribution, as given by extreme value theory.

This estimate is consistent with much of the existing literature, though the limiting case can be different from existing small-sample empirical estimates.

The estimate is also speculatively extended to the case of weighted stratified sampling.

2. Estimating Quantiles

Suppose we have n independent samples $X_1, X_2, \dots X_n$ from a common real-valued distribution, with complementary distribution function $R(x) = 1 - F(x) = \mathbb{P}(X > x)$. We sort these random variables into decreasing order as $X_{1,n} \geq X_{2,n} \geq \cdots \geq X_{n,n}$.

We wish to estimate a (right-tail) quantile p of the distribution, that is $x_p \coloneqq R^{-1}(p)$, for some p fairly close to zero. We are choosing to work with the interpolation version of the problem, where n is sufficiently large to include the quantile. In other words, $np > 1$. The "extrapolation" case, where $np < 1$, will not be considered here.

The existing literature puts forward an interpolation scheme as follows.

- For each integer k in $1, \dots, n$, assign a probability p_k to the sorted value $X_{k,n}$, where, for some constant $C \in [0,1]$,

$$p_k = \frac{k - C}{n + (1 - 2C)},$$

- Then we interpolate the desired probability p between the values of p_k, to interpolate the values of $X_{k,n}$ to give the quantile estimate

$$\hat{x}_p = X_{k,n}\frac{p_{k+1} - p}{p_{k+1} - p_k} + X_{k+1,n}\frac{p - p_k}{p_{k+1} - p_k}, \text{ where } p_k \le p < p_{k+1}.$$

But there is an arbitrariness in this scheme, created by the arbitrary variable C in the formula for p_k.

We will show that $X_{k,n}$ is approximately a mean-unbiased estimator for the p_k-percentile if C is chosen to be the value = $C = \frac{1}{2}(1 + \xi)$, where ξ is the extreme-value tail shape parameter of the distribution.

The existing literature, particularly Cunnane (1978), proposes distribution-dependent values for C and equivalently ξ, such as

Distribution	**C-value**	**ξ-value**	**Paper**
Normal	0.38	-0.24	Cunnane
Gumbel/exponential	0.44	-0.12	Cunnane
Uniform	0.00	-1	Cunnane
Recommended	0.40	-0.20	Cunnane
Median	0.33	-0.33	R/SciPy/Maple

Table 2.1 : Empirical values of C from the existing literature

We will confirm that these estimates make some sense for particular percentiles of those distributions, but they are not unbiased estimators in the limiting case (other than for Uniform).

2.1 Unbiased median estimator

Suppose the objective is to get an estimate $\hat{x}_p$ of $x_p = R^{-1}(p)$ with unbiased median, so that $\mathbb{P}(\hat{x}_p < x_p) = \frac{1}{2}$. Then we should pick $C = \frac{1}{3}$. This is well known, since $R(X_{k,n}) \sim Beta(k, n + 1 - k)$, and the median of a $Beta(\alpha, \beta)$ distribution is approximately $(\alpha - \frac{1}{3})/(\alpha + \beta - \frac{2}{3})$.

2.2 Unbiased mean estimator

But suppose instead that we are looking for an unbiased mean estimator,$\hat{x}_p$, of x_p, so that $\mathbb{E}(\hat{x}_p) = x_p$. In this case, we have a new theorem, which can be stated loosely as:

Theorem 2.1. Assuming that X has a distribution which is in the domain of attraction for some non-degenerate distribution function with tail shape parameter ξ, then the quantile estimator $X_{k,n}$ is a mean-unbiased estimator for $R^{-1}(p_k)$ where

$$p_k = \frac{k - C}{n + 1 - 2C}, \text{ where } C = \frac{1}{2}(1 + \xi)$$

for sufficiently large n and k and sufficiently small k/n.

Theorem 2.2. In the special case where R is twice-differentiable and eventually either convex or concave, then an even better estimate, when calculating an estimator for x_p is given by

$$C = C_p := \frac{1}{2}\left(1 + \xi_R\left(R^{-1}(p)\right)\right), \text{ where } \xi_R(x) := \frac{\partial}{\partial x}\left(\left(-\frac{\partial}{\partial x}\log R(x)\right)^{-1}\right).$$

2.3 A note on extreme value theory

Some readers may not be fully familiar with Extreme Value Theory, which describes the shape of the tail of probability distributions. A good introduction is chapter 7 of McNeil, Frey and Embrechts (2005).

To summarise, three key results in Extreme Value Theory are as follows:

(2.3.1) Let $X_{1,n}$ be the largest of n independent samples from a specific distribution. Suppose that there exist sequences of real constants d_n and $c_n > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(c_n^{-1}(X_{1,n} - d_n) \le x\right) = H(x),$$

for some non-degenerate distribution function $H(x)$. Then there exist modified sequences of real constants b_n and $a_n > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(a_n^{-1}(X_{1,n} - b_n) \le x\right) = H_\xi(x),$$

for some unique real value ξ, where

$$H_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right), & \xi \neq 0 \\ \exp(-e^{-x}), & \xi = 0. \end{cases}$$

We call the parameter ξ the tail shape parameter of the distribution.

We also define x_F to be the maximum possible value of the distribution, which is finite if $\xi < 0$ and infinite otherwise.

(2.3.2) Let $R_u(x) = \mathbb{P}(X > u + x | X > u)$ be the conditional tail probability of the distribution. Then if *X* satisfies the conditions of 2.3.1, then there is a positive function $b(u)$ such that

$$\sup_{x\in[0,x_F-u]} \left|R_u(x) - G_{\xi,b(u)}(x)\right| \to 0 \text{ as } u \to x_F.$$

where

$$G_{\xi,b}(x) = \begin{cases} (1+\xi x/b)^{-1/\xi}, & \xi \neq 0 \\ \exp(-x/b), & \xi = 0. \end{cases}$$

The family of distributions with $\xi < 0$ is called the Weibull class, those with $\xi = 0$ are called Gumbel, and those with $\xi > 0$ are called Fréchet. Most continuous distributions satisfy the conditions above and fall into one of these three classes.

(2.3.3) In the special case where the distribution function $R(x)$ is strictly decreasing and twice-differentiable, then we can define the "local shape" parameter of the right-hand tail, $\xi_R(x)$ as

$$\xi_R(x) \coloneqq \frac{\partial}{\partial x}\left(\left(-\frac{\partial}{\partial x}\log R(x)\right)^{-1}\right) = -1 + \frac{R(x)R''(x)}{R'(x)^2}.$$

In the restricted special case where $R(x)$ is strictly decreasing, twice-differentiable and also consistently either convex or concave for large x (in other words, there exists a value u such that $R'' \geq 0$ or $R'' \leq 0$ for all $x > u$), then

$$\xi_R(x) \to \xi \text{ as } x \to x_F.$$

A version of this result in the case where $\xi = 0$ is given at (7.25) of McNeil, Frey and Embrechts (2005). The proof of the general result, if it is not already well known, follows immediately from Theorem 2.2 of Kimchi and Richter-Dyn (1978), which gives that $\frac{\partial}{\partial x}R(u + b(u)x) \to G_{\xi,1}'(x)$ and $\frac{\partial^2}{\partial x^2}R(u + b(u)x) \to G_{\xi,1}''(x)$. It is immediate that the local shape of the $G_{\xi,1}$ distribution is identically equal to ξ at all points.

There is also an equivalent local shape for the left-hand tail

$$\xi_L(x) \coloneqq \frac{\partial}{\partial x}\left(\left(-\frac{\partial}{\partial x}\log F(x)\right)^{-1}\right) = -1 + \frac{F(x)F''(x)}{F'(x)^2},$$

but ξ_R and ξ_L will in general be different for asymmetric distributions.

The table below shows some examples of a few well-known distributions, along with their (right-tail) values of shape parameter ξ and local shape parameter $\xi_R(x)$.

Distribution	**Parameters**	**ξ-Value (right tail)**	**Local $\xi_R(x)$ First-order expansion**	**Order of convergence as $p \to 0$**	**EVT Class**
Normal	$N(\mu, \sigma^2)$	0	$-\left(\frac{x-\mu}{\sigma}\right)^{-2}$	$(-\log p)^{-1}$	Gumbel
Exponential	$\exp(\lambda)$	0	0	0	Gumbel
Gamma	$\Gamma(\gamma, \lambda)$	0	$-(\gamma-1)(\lambda x)^{-2}$	$(-\log p)^{-2}$	Gumbel
Uniform		-1	-1	0	Weibull
Beta	$Beta(\alpha, \beta)$	$-\beta^{-1}$	$-\beta^{-1}\left(1 + 2\frac{\alpha-1}{\beta+1}(1-x)\right)$	$p^{1/\beta}$	Weibull
Cauchy	Centre m, Scale s	1	$1 - \frac{2}{3}\left(\frac{x-m}{s}\right)^{-2}$	p^2	Fréchet
Student-t	ν degrees of freedom	ν^{-1}	$\nu^{-1} - \frac{\nu+1}{(\nu+x^2)(\nu+2)}$	$p^{2/\nu}$	Fréchet

Table 2.3: For some well-known distributions, the shape parameter, local shape parameter (to first order), order of convergence against probability, and EVT class.

Also shown is the order of convergence which is a function of the probability p, $c(p)$ such that

$$\left|\xi_R\left(R^{-1}(p)\right) - \xi\right| < kc(p) \text{ as } p \to 0, \text{ for some } k > 0.$$

3. Proof of Theorem

We start with a lemma, which comes from a user on the Mathematics Stack Exchange website.

Lemma 3.1

If $f_n: [0,\infty] \to [0,1]$ are a family of monotone functions, and $f: [0,\infty] \to [0,1]$ is a continuous monotone function, such that $f_n(x) \to f(x)$ pointwise for any $x \in [0,\infty]$ as $n \to \infty$, then $f_n \to f$ uniformly in that

$$\sup_{x\in[0,\infty]} |f_n(x) - f(x)| \to 0 \text{ as } n \to \infty.$$

Proof: See 'Etienne' (2014).

Extension. The proof can be extended to handle the case where $f_t \to f$, for real t, as $t \to \infty$.

We will introduce some notation before we get to the main results. We are still using the complementary tail probability $R(x) \coloneqq 1 - F(x)$, the conditional tail probability $R_u(x)$ and the tail GPD (Generalised Pareto Distribution) $G_{\xi,b}(x)$ from 2.3.2.

We will also write $G_\xi(x)$ for $G_{\xi,1}(x)$. Note that G_ξ is defined over the range $(g_L, g_u]$ where $g_L = -\xi^{-1}$ for positive ξ and $-\infty$ otherwise, and $g_U = -\xi^{-1}$ for negative ξ and $+\infty$ otherwise.

Now we can state the following proposition

Proposition 3.2

There is a positive scale function $b(u)$, such that

$$R_u(b(u)x) \to G_\xi(x) \text{ uniformly on } x \in [-K_L, K_U] \text{ as } u \to x_F.$$

The lower bound K_L can be any non-negative value (as long as $K < \xi^{-1}$ in the case where ξ is positive), and the upper bound K_U is $|\xi|^{-1}$ if ξ is negative, and infinite otherwise.

Further the scale function $b(u)$ can be chosen as

$$b(u) = \begin{cases} (x_F - u)|\xi| & \text{if } \xi < 0 \\ \int_0^\infty R_u(s)ds & \text{if } \xi = 0 \\ u\xi & \text{if } \xi > 0 \end{cases}$$

Proof of Proposition 3.2

(1) In the case where $\xi < 0$, we know (McNeil Frey and Embrechts, Theorem 7.10)

$$R(x_F - x^{-1}) = x^{1/\xi} L(x),$$

where $L(x)$ is slowly varying at infinity. Thus

$$R_u(b(u)x) = \left(1 - \frac{b(u)x}{x_F - u}\right)^{-1/\xi} \frac{L((x_F - u - b(u)x)^{-1})}{L((x_F - u)^{-1})} = G_\xi(x) \frac{L((x_F - u)^{-1}(1 + \xi x)^{-1})}{L((x_F - u)^{-1})}.$$

For any fixed x, the right-hand side converges pointwise to $G_\xi(x)$ as $u \to x_F$ since L is slowly varying.

The uniform convergence follows from Lemma 3.1 on the compact interval $[K_L, K_U]$. This follows the lead of Pickands (1975). This logic will also apply to the other two cases below.

(2) In the case where $\xi = 0$, we know from Corollary 2 to Theorem 3 (Balkema and De Haan, 1974) that we have pointwise convergence

$$R_u(b(u)x) \to G_0(x) \text{ as } u \to \infty \text{ for any } x,$$

where

$$b(u) = \int_0^\infty R_u(s)ds.$$

(3) In the case where $\xi > 0$, we know (McNeil Frey and Embrechts, 2005, Theorem 7.8)

$$R(x) = x^{-1/\xi} L(x),$$

where $L(x)$ is slowly varying at infinity. Thus

$$R_u(b(u)x) = \left(1 + \frac{b(u)x}{u}\right)^{-1/\xi} \frac{L(u + b(u)x)}{L(u)} = G_\xi(x) \frac{L(u(1 + \xi x))}{L(u)}.$$

For any fixed x, the right-hand side converges pointwise to $G_\xi(x)$ as $u \to \infty$ since L is slowly varying.

□

We can now deduce a corollary which is an extension of the EVT result (2.3.2) above to extend the valid *x*-domain to include some negative values.

Corollary 3.3

There is a positive scale function $b(u)$, such that

$$\sup_{x \in [-K_L b(u),\, x_F - u]} \left| R_u(x) - G_{\xi, b(u)}(x) \right| \to 0 \text{ as } u \to x_F$$

for the same bounds K_L, K_U as in Proposition A.2. Note that $K_U b(u)$ is $(x_F - u)$ for all cases.

Proof of Corollary 3.3

We know from Proposition 3.2 that

$$\sup_{x\in[-K_L,K_U]} \left|R_u(b(u)x) - G_{\xi,1}(x)\right| \to 0 \text{ as } u \to x_F.$$

Simply set $y = b(u)x$, and note that $G_{\xi,1}(y/b(u)) = G_{\xi,b(u)}(y)$.

□

Now we make an initial proof of the theorem using the extreme value theory limit. We start by defining $U_{k,n}$ to be the *k*-th smallest of n IID $U[0,1]$ samples, so that $U_{k,n} \sim Beta(k, n+1-k)$, and $X_{k,n} \sim R^{-1}(U_{k,n})$. We know that

$$\mu_k := \mathbb{E}(U_{k,n}) = \frac{k}{n+1}, \text{ and } \sigma_k^2 := Var(U_{k,n}) = \frac{\mu_k(1-\mu_k)}{n+2}.$$

Proposition 3.4

For a fixed $y_k = R^{-1}(\mu_k)$, where $\mu_k = k/(n+1)$, define

$$H_k(z) = R(y_k) G_{\xi,b_k}(z - y_k),$$

where $b_k = b(y_k)$, and H_k is capped and floored at one and zero. Then, with $k > \xi$,

$$H_k\left(\mathbb{E}\left(H_k^{-1}(U_{k,n})\right)\right) = \frac{k - C + O(k^{-1/2})}{n+1-2C}, \text{ where } C = \tfrac{1}{2}(1+\xi).$$

Proof of Proposition 3.4

Now

$$H_k^{-1}(p) = y_k + \frac{b_k}{\xi}\left(\left(\frac{\mu_k}{p}\right)^{\xi} - 1\right),$$

so

$$\left(H_k^{-1}\right)'(\mu_k) = -b_k\mu_k^{-1}, \text{ and } \left(H_k^{-1}\right)''(\mu_k) = (1+\xi)b_k\mu_k^{-2}.$$

Thus we can use a Taylor expansion of H_k^{-1} around μ_k to see that

$$\mathbb{E}\left(H_k^{-1}(U_{k,n})\right) = y_k + C b_k \mu_k^{-2} \frac{\mu_k(1-\mu_k)}{n+2} + O\left(b_k k^{-3/2}\right).$$

Now

$$H_k'(y_k) = -\mu_k b_k^{-1}, \text{ and } H_k''(y_k) = \mu_k b_k^{-2}(1+\xi).$$

So we can use a Taylor expansion of H_k around y_k to see that

$$H_k\left(\mathbb{E}\left(H_k^{-1}(U_{k,n})\right)\right) = \mu_k - C\frac{1-\mu_k}{n+2} + O\left(n^{-1}k^{-1/2}\right).$$

The right-hand side can be re-expressed to give the stated result

$$H_k\left(\mathbb{E}\left(H_k^{-1}(U_{k,n})\right)\right) = \frac{k - C + O(k^{-1/2})}{n+1-2C}.$$

□

We can now move to the final result. This is derived by using uniform approximations of R to H, and applying the above result for H. The final result required, which says intuitively that

$$R\left(\mathbb{E}\left(R^{-1}(U_{k,n})\right)\right) \approx \frac{k-C}{n+1-2C}, \text{ where } C = \tfrac{1}{2}(1+\xi),$$

is more formally stated as follows.

Theorem 3.5

Consider a specific distribution function F, with extreme-value tail shape parameter ξ. Let us define $X_{k,n}$ as the *k*-th largest observation of n independent samples from F and $R(x) = 1 - F(x)$. For any positive ϵ, there is a critical value $\mu(\epsilon)$ such that

$$R\left(\mathbb{E}\left(X_{k,n}\right)\right) = \frac{k(1+O(\epsilon)) - C + O(k^{-1/2})}{n+1-2C}, \text{ where } C = \tfrac{1}{2}(1+\xi)$$

as long as $\mu_k = k/(n+1) < \mu(\epsilon)$ and $k > \xi$.

Proof of Theorem 3.5

By Corollary 3.3 we know there is a critical $u_0(\epsilon)$ such that the maximum distance between R_u and $G_{\xi,b(u)}$ is no more than ϵ for all $u > u_0(\epsilon)$. Define $\mu(\epsilon) = R\left(u_0(\epsilon)\right)$, then we know that

$$\left|R_u(x) - G_{\xi,b(u)}(x)\right| < \epsilon \text{ for all } u \in [R^{-1}(\mu(\epsilon)), x_F] \text{ and } x \in [-K_L b(u), x_F - u].$$

Note there is a potential problem here. The scale $b(u)$ might tend to zero if $\xi \leq 0$, so the left limit could converge to zero for non-positive ξ. But our area of interest around μ_k will have half-width of the order of $5\sigma_k \approx \mu_k(5k^{-1/2})$, so the *x*-width required is $5b(y_k)k^{-1/2}$. This is inside the left-hand bound $-K_L b(y_k)$ (for non-tiny k), which is what we need. Thus, using $u = y_k$, we have for k such that $\mu_k < \mu(\epsilon)$.

$$H_k(x) - \epsilon\mu_k \leq R(x) \leq H_k(x) + \epsilon\mu_k, \text{ for all } x \in [u - K_L b(u), x_F].$$

We can infer that

$$H_k^{-1}(p+\epsilon\mu_k) \le R^{-1}(p) \le H_k^{-1}(p-\epsilon\mu_k).$$

Following a similar logic to the proof of Proposition 3.4, we find that

$$\mathbb{E}(X_{k,n}) = \mathbb{E}\left(R^{-1}(U_{k,n})\right) \le y_k + \epsilon b_k + C b_k \mu_k^{-2} \frac{\mu_k(1-\mu_k)}{n+2} + O\left(b_k\left(\epsilon^2 + k^{-3/2}\right)\right).$$

Since H_k is decreasing, this gives a lower bound

$$H_k\left(y_k + \epsilon b_k + C b_k \mu_k^{-2} \frac{\mu_k(1-\mu_k)}{n+2} + O\left(b_k\left(\epsilon^2 + k^{-3/2}\right)\right)\right) - \epsilon\mu_k \le R\left(\mathbb{E}\left(X_{k,n}\right)\right).$$

This reduces to

$$\mu_k - 2\epsilon\mu_k - C\frac{1-\mu_k}{n+2} + O\left(\mu_k\epsilon^2 + n^{-1}k^{-1/2}\right) \le R\left(\mathbb{E}\left(X_{k,n}\right)\right).$$

That can be re-written as

$$\frac{k(1+O(\epsilon)) - C + O(k^{-1/2})}{n+1-2C} \le R\left(\mathbb{E}\left(X_{k,n}\right)\right).$$

A similar upper bound establishes the result.

□

3.6 Special Case of twice-differentiable distribution function

A more direct sketch proof is possible in the special case where *R* is strictly decreasing and twice differentiable. In this case, we can fix k and n, and define $\mu = \mathbb{E}\left(U_{k,n}\right) = k/(n+1)$, and $x = R^{-1}(\mu)$. We can take a second-order Taylor expansion of R^{-1} around μ to get

$$X_{k,n} = R^{-1}\left(U_{k,n}\right) = x + R'(x)^{-1}\left(U_{k,n} - \mu\right) - \frac{R''(x)}{2R'(x)^3}\left(U_{k,n} - \mu\right)^2.$$

Taking expectations and using the fact that $Var\left(U_{k,n}\right) \approx R(x)(1-\mu)/(n+1)$, we get

$$\mathbb{E}\left(X_{k,n}\right) = x - \frac{R(x)R''(x)}{R'(x)^2}\frac{1-\mu}{2(n+1)R'(x)}.$$

Substituting in $C_\mu = \frac{1}{2}(1 + \xi_R(R^{-1}(\mu))$, where $\xi_R(x)$ is the local shape parameter from 2.3.3, we get

$$\mathbb{E}\left(X_{k,n}\right) = x - \frac{C_\mu(1-\mu)}{2R'(x)}.$$

Using the first-order Taylor expansion of R around x gives

$$p_k \coloneqq R\left(\mathbb{E}\left(X_{k,n}\right)\right) = \mu - \frac{C_\mu(1-\mu)}{(n+1)} \approx \frac{k - C_\mu}{n+1-C_\mu}.$$

We see that the local shape parameter $\xi_R(x)$ is the effective ξ to use in the quantile estimation formula. The value of C to use in this case depends on local shape parameter as $C_\mu = \frac{1}{2}(1 + \xi_R(R^{-1}(\mu)))$.

3.6.2 Median behaviour

The formula for p_k directly above has a slightly different denominator than the main formula in 3.5. The main formula has a denominator of $(n + 1 - 2C)$, but 3.6 has $(n + 1 - C)$. For calculations in the tails, there is very little difference. But for calculations near the median, the formula in 3.6 is slightly more accurate.

As an example, consider any distribution which is asymmetric about its median. Let's take the standard exponential distribution, with median $m = \log 2$. If we define $U_{k,n} \sim Beta(k, k)$, where $k = (n + 1)/2$, then $X_{k,n} = -\log U_{k,n}$. Since $-\log x$ is a strictly convex function, we have that $\mathbb{E}(X_{k,n}) > \log 2$, and thus that $X_{k,n}$ is a biased estimator for the median. In fact, we need $X_{k,n}$ with $k \approx (2n + 3)/4$ for an unbiased estimator of the median.

We will compare the empirical differences between these formula variants in the next section.

4. Empirical Results

We can perform experiments to check the validity of our formulae. For a given distribution, we can calculate analytically (using numerical integration) the percentile for which it is the mean-unbiased estimator, that is

$$p_k = R\left(\mathbb{E}\left(X_{k,n}\right)\right).$$

We then compare this true probability with our three candidate formulas for what we will call methods 'A' (classic formula with constant *C*), 'B' (revised formula with constant *C*) and 'C' (revised formula with *p*-dependent *C*)

$$p_k^A = \frac{k - C}{n + 1 - 2C}, \text{ where } C = \frac{1}{2}(1 + \xi),$$

$$p_k^B = \frac{k - C}{n + 1 - C}, \text{ where } C = \frac{1}{2}(1 + \xi),$$

$$p_k^C = \frac{k - C_p}{n + 1 - 2C_p}, \text{ where } C_p = \frac{1}{2}\left(1 + \xi_R\left(R^{-1}(k/(n+1))\right)\right).$$

For numerical computation, we use the useful approximations to various distribution functions from Abramowitz and Stegun (1964), and the numerical integration is performed by an adaptation of the Runge-Kutta method for solving ODEs from Press et al (1988).

4.1 Example – Normal distribution

As an example, let us work with the normal distribution with zero mean and unit variance. Suppose we take $p = 99\%$ and $n = 300$, so that we are working with $(k = 3)$ the third largest observation from a sample of 300 independent observations. The mean of the corresponding Beta distribution U_k is $\mu = k/(n+1) = 0.997\%$.

But the expected value of $X_{k,n} = \Phi_R^{-1}(U_k)$ is 2.3837, which has an actual tail percentile of 0.857%, clearly different from μ. This can be compared with the three approximate p_k values:

	A 'Classic'	B 'Revised'	C '*p*-dependent'
Tail probability approximation p_k	0.833%	0.832%	0.853%
Difference to actual (x10^4)	-2.4	-2.5	-0.4

Table 4.1.1 : Three approximate probabilities for $k = 3, n = 300$, normal distribution, and comparison with true value. Comparison is given as the difference of probabilities in basis points (bps), where 1bp=10^{-4}.

We see that method 'C' is the most accurate, and the other two methods are similar for this example.

In passing we can also compare the true value with two other traditional standards for quantile estimation. Firstly, the formula $p_k = (k-1)/(n-1)$, which corresponds to the Excel function 'PERCENTILE'. This gives a value of 0.669%, which has a difference of -18.8bp to the true value. Secondly, the formula $p_k = k/(n+1)$, which corresponds to the Excel function 'PERCENTILE.EXC'. This gives a value of 0.997%, which has a difference of 14.0bp to the true value. These are both much worse than our three candidate methods.

Let's repeat this exercise using method 'A' for a range of possible values of p from 50% to 99.99%, and a range of possible values of n from 300 to one million, we get the results shown in the table below

n \ *p*	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.9%	99.99%	Worst
300	-0.0	-0.7	-1.4	-2.0	-2.4	-2.5	-2.3	-2.4	-2.8	n/a	n/a	2.8
1,000	-0.0	-0.2	-0.4	-0.6	-0.7	-0.7	-0.7	-0.6	-0.6	-0.9	n/a	0.9
3,000	-0.0	-0.1	-0.1	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	n/a	0.2
10,000	-0.0	-0.0	-0.0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.0	-0.1	0.1
30,000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0
100,000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0
300,000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0
1,000,000	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0

Table 4.1.2 : Difference between method 'A' p_k and true percentile for the normal distribution function, taken over a range of values of *p* from 50% to 99.99% (top row) and a separate range of values of the sample size *n* from 300 to one million (left-hand column). Probability differences have been multiplied by 10^4.

We see that the larger errors occur when both n and $n(1-p)$ are small. But the errors are fairly small and are equivalent to having the 99.53% percentile as an approximation for the 99.50%-percentile.

Note that we also need $n(1-p) > \xi$ for the expectation of $X_{k,n}$ to be defined.

4.2 Other distributions

We can repeat the tests from section 4.1 for other known distributions. The test distributions are described in the table below.

Distribution	**Parameters**	**Density**	**ξ value (right tail)**
Normal	Zero mean, unit variance	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$	0
Exponential	Unit scale	$\exp(-x)$	0
Gamma	$\gamma = 5, \lambda = 1$	$\Gamma(\gamma)^{-1}x^{\gamma-1}\exp(-x)$	0
Uniform	[0,1]	$I(0 \le x \le 1)$	-1
Beta	$\alpha = 4, \beta = 2$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$-\beta^{-1} = -0.5$
Cauchy	Zero centre, unit scale	$\frac{1}{\pi(1+x^2)}$	1
Student-t	$\nu = 4$	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})}\nu^{-1/2}\left(\frac{\nu}{\nu+x^2}\right)^{(\nu+1)/2}$	$\nu^{-1} = 0.25$

Table 4.2.1 : Table showing the test distributions, their parameters, probability densities, and right-tail ξ values.

For each distribution the difference between the approximate and actual values of p_k was calculated for the same ranges of *p* and *n* as in 4.1. The table below shows the worst absolute difference (over the choices of *p*) for each distribution and each tested value of *k*.

	Distribution							
Sample size	Normal	Exponential	Gamma	Uniform	Beta	Cauchy	Student-t	Worst
300	2.765	8.313	3.217	0.000	2.430	3.074	1.648	8.313
1,000	0.946	2.498	0.966	0.000	0.730	0.920	0.722	2.498
3,000	0.238	0.833	0.322	0.000	0.243	0.306	0.133	0.833
10,000	0.087	0.250	0.097	0.000	0.073	0.092	0.070	0.250
30,000	0.024	0.083	0.032	0.000	0.024	0.031	0.013	0.083
100,000	0.007	0.025	0.010	0.000	0.007	0.009	0.004	0.025
300,000	0.002	0.008	0.003	0.000	0.002	0.003	0.001	0.008
1,000,000	0.001	0.002	0.001	0.000	0.001	0.001	0.000	0.002

Table 4.2.2 : Difference between approximate method 'A' and theoretical p_k probabilities for various test distributions, taken over a range of values of *n* from 300 to one million (left-hand column) and a separate range of values of the target quantile *p* from 50% to 99.99%. The worst (over the tested *p*) absolute difference, scaled by 10^4, is shown for each tested value of *n* and each tested distribution.

We can also repeat this exercise for the other two methods. These are shown in Tables 4.2.3 and 4.2.4.

Sample size	Distribution							
	Normal	Exponential	Gamma	Uniform	Beta	Cauchy	Student-t	Worst
300	8.319	1.232	5.102	0.000	6.586	16.667	10.404	16.667
1,000	2.499	0.614	1.532	0.000	1.979	5.000	3.124	5.000
3,000	0.833	0.054	0.511	0.000	0.660	1.667	1.042	1.667
10,000	0.250	0.061	0.153	0.000	0.198	0.500	0.312	0.500
30,000	0.083	0.005	0.051	0.000	0.066	0.167	0.104	0.167
100,000	0.025	0.000	0.015	0.000	0.020	0.050	0.031	0.050
300,000	0.008	0.000	0.005	0.000	0.007	0.017	0.010	0.017
1,000,000	0.003	0.000	0.002	0.000	0.002	0.005	0.003	0.005

Table 4.2.3 : Difference between approximate method 'B' and theoretical p_k probabilities for various test distributions, taken over a range of values of *n* from 300 to one million (left-hand column) and a separate range of values of the target quantile *p* from 50% to 99.99%. The worst (over the tested *p*) absolute difference, scaled by 10^4, is shown for each tested value of *n* and each tested distribution.

Sample size	Distribution							
	Normal	Exponential	Gamma	Uniform	Beta	Cauchy	Student-t	Worst
300	1.009	1.232	1.127	0.000	0.706	0.090	1.099	1.232
1,000	0.538	0.614	0.582	0.000	0.341	0.008	0.646	0.646
3,000	0.048	0.054	0.052	0.000	0.036	0.001	0.050	0.054
10,000	0.056	0.061	0.060	0.000	0.035	0.000	0.067	0.067
30,000	0.005	0.005	0.005	0.000	0.004	0.000	0.005	0.005
100,000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
300,000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1,000,000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 4.2.4 : Difference between approximate method 'C' and theoretical p_k probabilities for various test distributions, taken over a range of values of *n* from 300 to one million (left-hand column) and a separate range of values of the target quantile *p* from 50% to 99.99%. The worst (over the tested *p*) absolute difference, scaled by 10^4, is shown for each tested value of *n* and each tested distribution.

We see that method 'C' is the best method taking over all distributions and sample sizes. Method 'B' outperforms method 'A' for some distributions, such as the exponential, which have constant local shape ξ_R. But method 'B' underperforms around the median for symmetric distributions with $\xi \neq -1$, such as Normal, Cauchy and Student-*t*.

Note that the random sampling error has standard deviation (in percentile terms) of $\sqrt{\frac{p(1-p)}{n}}$, which is generally higher than the differences between the choices of p_k. This may reduce the motivation to be particularly precise about p_k unless a high degree of accuracy is required.

4.3 Xi values against percentile

The previous two sections show that the error between the approximate and theoretical values of p_k is small.

We can also explore how the theoretical xi values change with percentile. We have an existing idea of the speed of convergence from table 2.3. The table below shows what actually happens in practice. There is a range of tested percentiles from 50% to 99.99%, and the same set of known test distributions as in 4.2.

For each percentile and each test distribution, we calculate the theoretical $\xi_R\left(R^{-1}(1-p)\right)$, where $R(x) = \mathbb{P}(X > x)$ is the tail distribution function, and ξ_R is given in 2.3.3.

These data show that there can be quite large variation between values of ξ_R at the median (where symmetric distributions will have $\xi_R = -1$) and the limit ξ.

Percentile	Normal	Exponential	Gamma	Uniform	Beta	Cauchy	Student-t
50%	-100%	0%	-61%	-100%	-129%	-100%	-100%
60%	-74%	0%	-43%	-100%	-106%	-18%	-63%
70%	-55%	0%	-31%	-100%	-89%	37%	-36%
80%	-40%	0%	-21%	-100%	-76%	73%	-15%
90%	-27%	0%	-13%	-100%	-65%	93%	2%
95%	-20%	0%	-9%	-100%	-59%	98%	11%
98%	-15%	0%	-6%	-100%	-55%	100%	17%
99%	-13%	0%	-5%	-100%	-54%	100%	20%
99.50%	-11%	0%	-4%	-100%	-52%	100%	21%
99.90%	-8%	0%	-3%	-100%	-51%	100%	23%
99.99%	-6%	0%	-2%	-100%	-50%	100%	25%
Limit ξ	0%	0%	0%	-100%	-50%	100%	25%

Table 4.3.1 : Theoretical xi value against percentiles p for a set of known test distributions.

The normal distribution, as expected, is the slowest to converge. But all distributions have ξ_R within 20% of its limiting value for all percentiles greater than or equal to 95%. This translates into a percentile error of around $10\%/n$, which is equal to 0.01% for $n = 1000$.

These data can also be seen in graphical form.

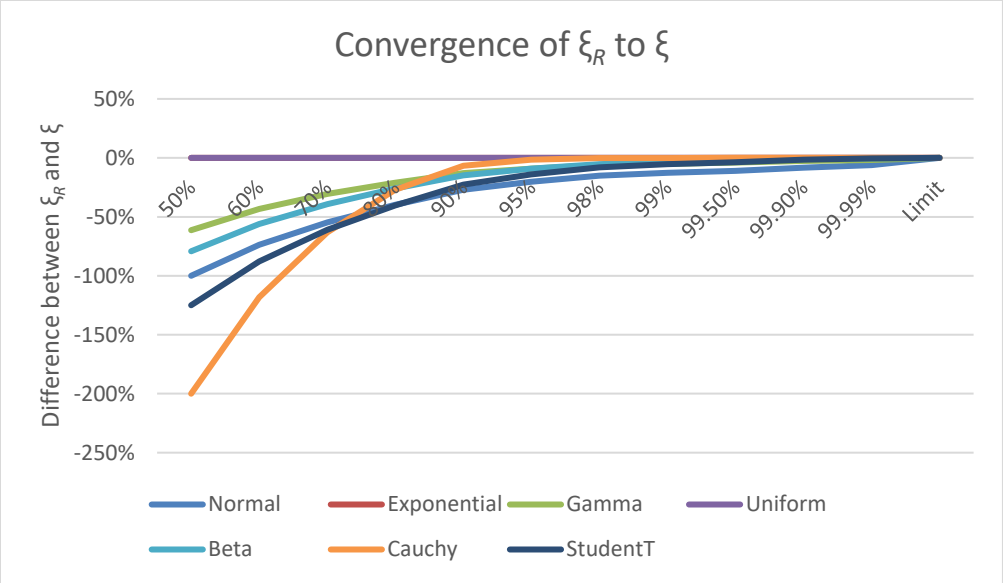

Figure 4.3.2 : Graph showing the difference between $\xi_R\left(R^{-1}(1-p)\right) - \xi$ as the percentile p tends to 100%.

The Exponential and Uniform distributions, which are of pure GPD form, have immediate convergence since $\xi_R(x) = \xi$ for all *x*.

4.4 Comparison of several approximations

We can finish our empirical section by comparing our three ξ-based approximations for p_k with three existing traditional approximations for p_k. These are called the 'Central', 'Inclusive' and 'Exclusive' cases and have these formulas for p_k:

$$p_k^{cen} = \frac{k-\frac{1}{2}}{n}, \qquad p_k^{inc} = \frac{k-1}{n-1}, \qquad p_k^{exc} = \frac{k}{n+1}.$$

We recognise these as our Method A for the cases $\xi = 0$ (Gumbel), $\xi = 1$ (eg Cauchy) and $\xi = -1$ (eg Uniform) respectively.

We fix $n = 1{,}000$ (but the results are very similar for different n), and for each approximation method and for each test distribution, we take the worst difference between the approximate and actual p_k across our range of test probabilities.

	Normal	Exponential	Gamma	Uniform	Beta	Cauchy	Student-t	Worst
Method A	0.95	2.50	0.97	0.00	0.73	0.92	0.72	2.50
Method B	2.50	0.61	1.53	0.00	1.98	5.00	3.12	5.00
Method C	0.54	0.61	0.58	0.00	0.34	0.01	0.65	0.65
Central	0.95	2.50	0.97	4.99	2.89	5.00	1.01	5.00
Inc	5.95	5.61	5.72	9.99	7.89	0.92	4.47	9.99
Exc	4.35	4.90	4.69	0.00	2.31	9.95	5.96	9.95

Table 4.4.1 : Worse case difference (in basis points) between approximate and actual p_k probabilities, taken over the range of tested percentiles, shown for each approximation method and each tested distribution.

We see that Method C is the best (has the least bad worst case), with Method A in second place. Method B has problems estimating the median of symmetric distributions with $\xi \neq -1$, such as Normal, Cauchy and Student-*t*. The three traditional methods each do well on distributions whose ξ-value is consistent with them (shown as green), but do badly for other distributions (shown as amber).

5. Quantiles of weighted stratified samples

The earlier sections have focused on the case where the samples are unbiased with independent and identically distributed observations.

But there are a number of real-world settings, including opinion polling, market research and other data-driven situations, where the sample contains some known biases. These are often corrected using a weighting scheme. This section will tentatively propose a scheme to interpolate quantiles in the presence of stratified weights.

5.1 Stratified sampling

Let's posit a stratified sampling situation. We have a total population, which is divided into subpopulations $j = 1, \ldots, m$. The subpopulations may be defined by age, gender, class, political affiliation, and so on, or by combinations of these. We know that the true fraction of subpopulation *j* in the total population is π_j.

We have a (scalar) property of interest which is defined for each individual in the population. We are interested in the distribution of *X*, which is the random value of this property for some person uniformly selected from the population.

As a model, we assume that individuals within the same subpopulation *j* have IID values of the property, drawn from a distribution with distribution function $F_j(x)$, which is unknown.

The unknown distribution function of *X* is a mixture of the subpopulation distribution functions.

$$F_X(x) = \sum_{j=1}^{m} \pi_j \, F_j(x).$$

Our known data is a sample of individuals, but the sample is not uniform. We assume that we sample uniformly within each subpopulation, but the sample is not uniform across the subpopulations. In particular the sample population fraction from subpopulation *j* is q_j, which is different from π_j.

Suppose we observe $Y_1, Y_2, \ldots, Y_n$ where Y_i is the value of the property for an individual i, who is in subpopulation J_i. What is a good estimate of a quantile of *X*?

We assume that q_j is the sample population frequency of subpopulation *j*, as

$$q_j = \frac{1}{n}\sum_{i=1}^{n} I(J_i = j).$$

Let us define the weight of individual *i* in any subpopulation as the ratio of the true subpopulation fraction to the sampled fraction:

$$w_i = \frac{\pi(J_i)}{q(J_i)}.$$

Then

$$\hat{F}_X(x) := \frac{1}{n}\sum_{i=1}^{n} w_i\, I(Y_i \le x)$$

is an unbiased estimator for $F_X(x)$. But we want a good estimator for $F_X^{-1}(p)$, which is a slightly different question.

Suppose we order the pairs (Y_i, w_i) so that Y_k is increasing. We generalise the unweighted probability attached to Y_k as p_k where

$$p_k = \frac{S_k - Cw_k}{S_n + (1-2C)w_k}, \text{ where } S_i = \sum_{r=1}^{i} w_r \text{ , and for some constant } C \in [0,1].$$

This weighted formula for p_k has the following advantageous properties

- p_k lies in $[0,1]$ and is increasing in k
- it reduces to the unweighted case (method 'A') if all the weights are equal
- it is symmetric in that

$$\frac{S_k - Cw_k}{S_n + (1-2C)w_k} + \frac{(S_n - S_{k-1}) - Cw_k}{S_n + (1-2C)w_k} = 1.$$

We also propose that the value of C should be $C = \frac{1}{2}(1+\xi)$, where ξ is the tail shape parameter of the distribution function F_X. In fact, $\xi = \max \xi_j$ if any ξ_j is non-negative (and $\xi = \min \xi_j$ otherwise), where ξ_j is the tail shape parameter of the subpopulation distribution function F_j. (Assuming that all bounded distributions have the same upper bound.)

5.2 Empirical Results

It is difficult to make analytic approximations in the weighted case. Instead, we can perform some empirical calculations to compare alternative quantile interpolation methods.

Each test has a population made up of some subpopulations, each with a specific distribution for the statistic of interest. The true frequency of each subpopulation is known, but we have a biased sample

which has a different sample frequency for the subpopulations. We use the method above to estimate various quantiles for the overall distribution, as well as the three "traditional" methods from section 4.4.

Five separate situations were tested.

5.2.1 Normal means

There are three sub-populations, all with equal frequency. These have normal distributions, all with unit variance, but means of -1, 0, and 1 respectively. There is a biased sample of 2,500 individuals, which contains 500 members of the first subgroup, 875 of the second and 1,125 of the third.

5.2.2 Normal variances

There are two sub-populations, the first of which has 25% actual frequency and the second has 75%. We have a sample of 4,750 observations, of which 250 come from the first group and 4,500 from the second. The first group has the normal distribution with zero mean and variance of 3, and the second group has the normal distribution with zero mean and variance of 1.

5.2.3 Betas

There are three sub-populations with the same actual and sample frequencies as 5.2.1. The subpopulations each have the beta distribution with (α, β) parameters respectively of $(0.5, 1), (1,1), (1,0.5)$. The overall ξ of the mixture distribution is -1.0.

5.2.4 Cauchys

There are three sub-populations with the same actual and sample frequencies as 5.2.1. The subpopulations each have the Cauchy distribution with unit scale and centres of -5, 0, and 5 respectively.

5.2.5 Various

There are three sub-populations with the same actual and sample frequencies as 5.2.1. The first is a Cauchy with centre -5 and unit scale, the second a normal with zero mean and unit variance, and the third a beta with $\alpha = \beta = 5/2$.

The quantiles estimated were: 90%, 95%, 98%, 99%, 99.9%.

The results of the tests are shown in Table 5.2.1 below. In each case the test produces a (stochastic) estimate $\hat{x}_p$ for the *p*-percentile of the distribution function F_X. The error of the estimate is expressed by the quantity $F_X\left(\mathbb{E}\left(\hat{x}_p\right)\right) - p$.

Test	Tail *Xi*	Method A	Central (*xi*=0)	Inc (*xi*=1)	Exc (*xi*=-1)
Normal Means	0	0.5	0.5	1.0	1.9
Normal Variances	0	1.7	1.7	4.6	4.9
Betas	-1	1.3	2.8	4.2	1.3
Cauchys	1	0.6	2.3	0.6	4.1
Various	1	2.1	3.2	2.1	5.3

Table 5.2.1 : The quantile estimation probability errors defined as the quantity described above in basis points (1bp=10^{-4}), with the average error absolute error taken across the five tested percentiles.

We see in every case that the weighted version of 'Method A' is the most accurate method. 'Method A' often coincides with a particular traditional method which corresponds to the tail shape parameter of that distribution.

This provides some evidence that the weighted version of Method A is a suitable method to use for interpolating quantiles in stratified samples.

But further work could be useful to establish this in more general cases.

6. Summary and Recommendations

Given these results we can now make recommendations for estimating percentiles empirically, in the case where we want an expectation-unbiased estimator of the *q*-quantile where $n(1-q) > 1$.

The general form of the estimate is as follows. We have *n* independent observations from the same distribution, $X_1, X_2, \dots, X_n$ and we sort them in decreasing order as $X_{1,n} \geq X_{2,n} \geq \cdots \geq X_{n,n}$. We want to estimate the $(1-p)$-percentile, where the tail percentile *p* satisfies $p \leq 50\%$.

We define p_k for each *k* as

$$p_k = \frac{k - C}{n + 1 - 2C},$$

for some constant *C* (see below). For a given desired tail percentile *p*, we define our estimate as

$$\hat{x}_p = X_{k,n} + \left(X_{k+1,n} - X_{k,n}\right)\frac{p - p_k}{p_{k+1} - p_k}, \text{ where } k = \lfloor (n+1)p + C(1-2p) \rfloor.$$

We will give recommendations for the choice of *C* based on the amount of information we have about the distribution

6.1 Very limited information

Suppose there is very limited information about the distribution and we are not able to estimate a value for the tail shape parameter ξ.

If the distribution has a definite hard upper bound (no tails), then assume $C = 0$.
If the distribution is known to have heavy (power-law) tails, then assume $C = 1$.
Otherwise assume $C = \frac{1}{2}$.

6.2 Limited information

Even in cases of limited information about the distribution it should be possible to estimate its tail shape parameter, ξ. This is zero for many common distributions (including the normal, exponential and gamma distributions). If the distribution has a definite hard upper bound, then ξ is likely to be negative. If the distribution is known to have very heavy tails, then ξ is likely to be positive. There are also well-known

methods for estimating ξ from the sample data, such as the Hill estimator, given in section 7.2 of McNeil, Frey and Embrechts (2005).

In this case, the best value of C to use is given by the expression $C = \frac{1}{2}(1 + \xi)$.

6.3 Richer information

In a fortunate case we have either an analytic approximation of the distribution or a good-quality estimate of the local shape parameter. That allows us either to calculate the theoretical $\xi_R(x)$ for any value of *x*, using the formula in 2.3.3, or to have an estimate for it around $x = R^{-1}(p)$. We use a modified version of the p_k formula which is slightly more accurate:

$$p_k = \frac{k - C_p}{n + 1 - C_p}, \text{ with } k = \lfloor (n+1)p + C_p(1-p) \rfloor$$

where

$$C_p = \tfrac{1}{2}(1 + \xi_R(R^{-1}(p)),$$

and $R(x)$ is the tail distribution function $R(x) = \mathbb{P}(X > x)$.

Note that it may not be necessary to calibrate (all) the parameters of the distribution since $\xi_R(R^{-1}(p))$ is insensitive to linear shifts and scales of the distribution.

It is also possible for the C_p to be outside the range [0,1] which will happen if $|\xi_R| > 1$.

6.4 Weighted stratified samples

In the case of stratified samples, where there is a weight for each observation, the following method should be followed.

Sort the observations into decreasing order as $X_1 \geq X_2 \geq \cdots \geq X_n$ with corresponding weights $w_1, w_2, \ldots, w_n$. We want to estimate the $(1-p)$-percentile, where the tail percentile *p* satisfies $p \leq 50\%$.

Make an estimate for the tail shape parameter ξ of the distribution, as in sections 6.1 or 6.2, and define $C = \frac{1}{2}(1 + \xi)$. Now define

$$p_k = \frac{S_k - Cw_k}{S_n + (1 - 2C)w_k}, \text{ where } S_r = \sum_{u=1}^{r} w_r.$$

For the tail percentile p find k such that $p_k \leq p < p_{k+1}$, and define the $(1-p)$-quantile estimate as

$$\hat{x}_p = X_k + (X_{k+1} - X_k)\frac{p - p_k}{p_{k+1} - p_k}.$$

This obvious generalisation of the unweighted case seems to work relatively well in practice.

References

Abramowitz, M. and Stegun, I.A. (editors), 1964, "Handbook of Mathematical Functions", U.S. National Bureau of Standards, reprinted Dover (1965).

Balkema, A.A. and de Haan, L., 1974, "Residual Life Time at Great Age", Annals of Probability, Vol 2 (1974), No 5, 792-804.

Cunnane, C., 1978, "Unbiased Plotting Positions – a Review", *Journal of Hydrology*, 37 (1978) 205-222.

User 'Etienne', 2014, "Sequence of monotone functions converging to a continuous limit, is the convergence uniform?", 14 June 2014, Mathematics Stack Exchange, URL https://math.stackexchange.com/questions/834126/sequence-of-monotone-functions-converging-to-a-continuous-limit-is-the-converge

Kimchi, E. and Richter-Dyn, N., 1978, "Convergence Properties of Sequences of Functions with Application to Restricted Derivative Approximation", Journal of Approximation Theory, Vol 22 (1978), 289-303.

McNeil, A.J., Frey, R. and Embrechts, P., 2005, "Quantitative Risk Management: Concepts, Techniques, Tools", Princeton

Pickands III, J., 1975, "Statistical Inference Using Extreme Order Statistics", Annals of Statistics, Vol 3 (1975), No 1, 119-131

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., 1988, "Numerical Recipes in C", Cambridge.