NONPARAMETRIC STATISTICS SUMMARY

ANDREW TULLOCH

1. Basics

Theorem (The Delta Method)**.** Let Y_n be a sequence of random vectors in $\mathbb{R}^d$ such that for some $\mu \in \mathbb{R}^d$ and a random vector Z, we have $n^{\frac{1}{2}}(Y_n - \mu) \xrightarrow{d} Z$. If $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable at μ, then $n^{\frac{1}{2}}(g(Y_n) - g(\mu)) \xrightarrow{d} \nabla g(\mu)^T Z$.

Proof. For $d = 1$. Let $g'(\mu) = \nabla g(\mu)$, and let $h : \mathbb{R} \to \mathbb{R}$, by

$$h(y) = \begin{cases} \frac{g(y) - g(\mu)}{y - \mu} & y \neq \mu \\ g'(\mu) & y = \mu \end{cases} \tag{1.1}$$

Then by the continuous mapping theorem and Sltusky's theorem, $n^{\frac{1}{2}}(g(Y_n) - g(\mu)) = h(Y_n) n^{\frac{1}{2}}(Y_n - \mu) \xrightarrow{d} g'(\mu)Z$. □

Let $X_1, \dots, X_n$ be IID on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution function F. The **empirical distribution function** $\hat{F}_n$ is defined by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)\,. \tag{1.2}$$

Theorem (Glivenko-Cantelli (1933) - The Fundamental Theorem of Statistics)**.**

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{as} 0. \tag{1.3}$$

Theorem. Let $X_1, \dots, X_n \sim F$ IID. Then for every $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon \right) \leq 2e^{-2n\epsilon^2}. \tag{1.4}$$

Definition. *For $p \in (0, 1]$, the quartile function is defined by $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ and is left-continuous.*

The sample quartile function is $\hat{F}_n^{-1}(p) = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\}$.

Theorem. Let $U_1, U_2, \dots, U_n \sim U(0, 1)$ IID and $p \in (0, 1)$. Then

$$\sqrt{n}(U_{\lceil np \rceil} - p) \xrightarrow{d} N(0, p(1-p)). \tag{1.5}$$

Proof. Let $Y_1, \dots, Y_n$ IID Exp(1), let $V_n = Y_1 + \dots + Y_{\lceil np \rceil}$ and $W_n = Y_{\lceil np \rceil + 1}, \dots, Y_{n+1}$. Note that V_n, W_n are independent and

$$\frac{V_n}{V_n + W_n} =^d U_{\lceil np \rceil} \tag{1.6}$$

by previous proposition. Then

$$\sqrt{n}(\frac{V_n}{n} - p) = \frac{\sqrt{\lceil np \rceil}}{\sqrt{n}} (\frac{V_n - \lceil np \rceil}{\sqrt{\lceil np \rceil}}) + \frac{\lceil np \rceil - np}{\sqrt{n}} \xrightarrow{d} N(0, p) \tag{1.7}$$

by the CLT and Slutsky's theorem.

Similarly, $\sqrt{n}(\frac{W_n}{n} - q) \xrightarrow{d} N(0, q)$, where $q = 1 - p$, then by the delta method, with $g(x, y) = \frac{x}{x+y}$,

$$\sqrt{n}(U_{\lceil np \rceil} - p) =^d \sqrt{n}(g(\frac{V_n}{n}, \frac{W_n}{n}) - g(p, q)) \tag{1.8}$$

$$\xrightarrow{d} N(0, \nabla g(p, q)^T \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \nabla g(p, q)) \tag{1.9}$$

$$=^d n(0, p(1-p)) \tag{1.10}$$

□

Theorem. Let $p \in (0, 1)$ and let $X_1, \dots, X_n$IIDF where F is differentiable at $F^{-1}(p)$ with positive derivative $f(F^{-1}(p))$. Then

$$\sqrt{n}(X_{\lceil np \rceil} - F^{-1}(p)) \xrightarrow{d} N(0, \frac{p(1-p)}{f(F^{-1}(p))^2}) \tag{1.11}$$

Proof. Let $U_1, \dots, U_n$IID$U(0, 1)$ so that $F^{-1}(U_{\lceil np \rceil}) =^d X_{\lceil np \rceil}$. Then by the previous theorem and the delta method with $g = F^{-1}$,

$$\sqrt{n}(X_{\lceil np \rceil} - F^{-1}(p)) =^d \sqrt{n}(g(U_{\lceil np \rceil}) - g(p)) \tag{1.12}$$

$$\xrightarrow{d} N(0, \frac{p(1-p)}{f(F^{-1}(p))^2}) \tag{1.13}$$

□

Definition. *Usually, we prefer to choose h to minimize some expression measuring how well $\hat{f}_h$ estimates f as a function. We therefore define the Mean Integrated Squared Error (MSIE) as*

$$MSIE(\hat{f}_h) = \mathbb{E}\left(\int_{-\infty}^{\infty} \{\hat{f}_h(x) - f(x)\}^2 dx \right) \tag{1.14}$$

$$= \int_{-\infty}^{\infty} MSE(\hat{f}_h(x)) dx \tag{1.15}$$

$$= \int_{\infty}^{\infty} ((K_h \star f)(x) - f(x))^2 + \frac{[}{1}] h((K_n^2 \star f)(x) - (K_h \star f)^2(x)) dx \tag{1.16}$$

which is justified by Fubini's theorem as the integrand is non-negative. JAlthough exact, this expression depends on h in a complicated way. We therefore seek asymptotic approximation to calify this dependnence and faciliate an asymptotically optimal choice of h.

We need the following conditions:

(i) f is twice differentiable, $f'$is bounded, and $R(f) = \int_{-\infty}^{\infty} f''(x)^2 dx < \infty$.
(ii) $h = h_n$ is a non-random sequecne with $h \to 0$ and $nh \to \infty$ as $n \to \infty$.
(iii) K is non-negative, $\int_{-\infty}^{\infty} K(x) dx = 1$, $\int_{-\infty}^{\infty} xK(x) dx = 0$, $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x) dx < \infty$, and $R(x) < \infty$.

Theorem. Assume that the previous conditions hod. Then, for all $x \in \mathbb{R}$,

$$MSE(\hat{f}_n(x)) = \frac{R(K) f(x)}{nh} + \frac{1}{4} h^4 \mu_2^2(K) f''(x)^2 + o(\frac{1}{nh} + h^4) \tag{1.17}$$

as $n \to \infty$.

Consider now minimizing the asymptotic MISE (AMISE) $\frac{R(K)}{nh} + \frac{1}{4} h^4 \mu_2^2(K) R(f$ with respect to h, yielding the asymptotically optimal bandwidth

$$h_{AMISE} = (\frac{R(K)}{\mu_2^2(K) R(f'') n})^{\frac{1}{5}} \tag{1.18}$$

Substituting back, we obtain

$$AMISE(\hat{f}_{AMISE}) = \frac{5}{4} R(K)^{\frac{4}{5}} \mu_2(K)^{\frac{2}{5}} R(f'')^{\frac{1}{5}} n^{-\frac{4}{5}}. \tag{1.19}$$

Notice the slower rate than the typical $O(n^{-1})$ parametric rate. Notice that for the "rough" densities, with larger $R(f'')$, we should use a smaller bandwidth, and these densities are harder to estimate.

Theorem. Assume the previous assumptions (i), (ii), (iii) and that K is bounded. Then, for all $x \in \mathbb{R}$,

$$n^{\frac{2}{5}}(\hat{f}_{h_{AMISE}}(x) - f(x)) \xrightarrow{d} N(\frac{1}{2} \mu_2(K) f''(x), R(K) f(x)) \tag{1.20}$$

Theorem. If f is the $N(0, \sigma^2)$ density, then $R(f'') = \frac{3}{8\sqrt{\pi}} \sigma^{-5}$. The normal scale rate $\hat{h}_{NS}$ consists of replacing $R(f'')$ in h_{AMISE} with $\frac{3}{8\sqrt{\pi}} \hat{\sigma}^{-5}$, where $\hat{\sigma}$ is an estimate of σ. This tends to oversmooth.

The choice of kernel is coupled with the choice of bandwidth, because if we replace $K(x)$ by $\frac{1}{2} K(\frac{1}{2})$ and we halve the bandwidth, the estimate is unchanged. We therefore fix the scale by setting $\mu_2(K) = 1$. Minimizing $AMISE(\hat{f}_h)$ over K the amounts to minimizing $R(K)$ subject to

$$\int_{-\infty}^{\infty} K(x) dx = 1 \tag{1.21}$$

$$\int_{-\infty}^{\infty} xK(x) dx = 0 \tag{1.22}$$

$$\mu_2(K) = 1 \tag{1.23}$$

$$K(x) \geq 0 \tag{1.24}$$

The solution is given by the Epanechnikov kernel (1969).

$$K_E(x) = \frac{3}{4\sqrt{5}} (1 - \frac{x^2}{5}) \mathbb{I}\left(|x| \leq \sqrt{5} \right) \tag{1.25}$$

The ratio $\frac{R(K_E)}{R(K)}$ is called the **efficiency** of a kernel K, because it represents the ratio of the sample sizes needed to obtain the same $AMISE$ when using K_E compared with K.

Kernel	Efficiency
Epachnikov	1.0
Normal	0.951
Triangular	0.986
Uniform	0.930

Theorem. A natural estimator of the r-th derivative $f^{(r)}$ of f is given by

$$\hat{f}_h^{(r)}(x) = \frac{1}{nh^{r+1}} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{1.26}$$

obtained from differentiating the standard KDE for $\hat{f}$.

Under regularity conditions,

$$MSE(\hat{f}_h^{(r)}(x)) = \frac{R(K^{(r)})}{nh^{2r+1}} f(x) + \frac{1}{4} h^4 \mu_2^2 f^{(r-2)}(x)^2 + o(\frac{1}{nh} + h^4). \tag{1.27}$$

This leads to an optimal bandwidth of order $n^{-\frac{1}{2r+5}}$ and a rate of converge of $n^{-\frac{4}{2r+5}}$.

Theorem. It is possible to make the dominant integrated squared bias term of $MISE(\hat{f}_h)$ vanish by choosing $\mu_2(K) = 0$. This means we have to allow the Kernel to take negative values, so the resulting estimate need not be a density.

We can set $\hat{f}_h(x) = \max(\hat{f}_h(x), 0)$ and then renormalize, but then we lose smoothness. Nevertheless, we define K to be a k-th order kernel if writing $\mu_j(K) = \int_{-\infty}^{\infty} x^j K(x) dx$, we have

$$\mu_0(K) = 1 \tag{1.28}$$

and $\mu_j(K) = 0$ for $j = 1, \dots, k-1$, $\mu_k(K) \neq 0$, and

$$\int_{-\infty}^{\infty} |x|^k |K(x)| dx < \infty \tag{1.29}$$

If f has k continuous bounded derivatives with $R(f^{(k)}) < \infty$, then it is shown (example sheet) that $h_{AMISE} = cn^{-\frac{1}{2k+1}}$ and

$$AMISE(\hat{f}_{h_{AMISE}}) = O(n^{-\frac{2k}{2k+1}}) \tag{1.30}$$

Thus, under increasingly strong smoothness assumptions, convergence rates arbitrarily close to the parametric rate of $O(n^{-1})$ can be obtained.

The practical benefit of higher order kernels is not always apparent, and the negativity/smoothness/bandwidth selection problems mean that they are rarely used in practice.

2. Nonparameteric Regression

Assume a fixed design. The local polynomial estimator $\hat{m}_h(x; p)$ of degree p with kernel K with a bandwidth h is constructed by fitting a polynomial of degree p using weighted least squares. The weight $K_h(x_i - x)$ is associated with the weight (x_i, Y_i).

More precisely, $\hat{m}_h(x; p) = \hat{\beta}_0$ where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ whicch is minimizing

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p)^2 K_h(x_i - x) \tag{2.1}$$

where $\beta \in \mathbb{R}^{p+1}$

The theory of weighted least squares gives

$$(X^T K X)\hat{\beta} = X^T K y \tag{2.2}$$

For $p = 0$, then a simple expression (Nodorya-Watson, local constant) exists:

$$\hat{m}_h(x; 0) = \frac{\sum_{i=1}^{n} K_h(x_i - x) Y_i}{\sum_{i=1}^{n} k_h(x_i - x)} \tag{2.3}$$

For $p = 1$, we call this a local linear estimator, and we have the explicit result

$$\hat{m}_h(x; 1) = \frac{1}{n} \sum_{i=1}^{n} \frac{S_{2,h}(x) - S_{1,h}(x)(x_i - x)}{S_{2,h}(x) S_{0,h}(x) - S_{1,h}(x)^2} K_h(x_i - x) Y_i \tag{2.4}$$

with

$$S_{r,h}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^r K_h(x_i - x) \tag{2.5}$$

All local polynomial estimators of the form

$$\sum_{i=1}^{n} W(x_i, x) Y_i \tag{2.6}$$

This type of estimator is called a linear estimator. This set of weights $\{W(x_i, x)\}$ is called the **effective kernel**.

2.1. Mean Squared Error Approximations.

For convenience, let $x_i = \frac{i}{n}$. We consider the following conditions:

(i) m is twice continuosly differentiable on $[0, 1]$ and is bounded, v is continuous.
(ii) $h = h_n, h_n \to 0$, $nh \to \infty$.
(iii) K is a nonnegative probability density, symmetric, has zeros outside of $[-1, 1]$. $R(K) = \int K^2(x) dx < \infty$, and $\mu_2(K) = \int x K^2(x) < \infty$.

Theorem. Under the conditions previously, for $x \in (0, 1)$, we have

$$MSE(\hat{m}_h(x; 1)) = \frac{1}{nh} R(K) v(x) + \frac{1}{4} h^4 (m''(x))^2 \mu_2(K) + o(\frac{1}{nh} + h^4) \tag{2.7}$$

2.2. Splines.

Let $n \geq 3$, and consider for a fixed homoscedastic design

$$Y_i = m(x_i) + \sigma \epsilon_i \tag{2.8}$$

where ϵ_i are IID with $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{V}(\epsilon_i) = 1$.

Another natural idea to estimate the regression curve m is to balance the fidelity of the fit to the data and the roughness of the resulting curve. This can be done by minimizing

$$\sum_{i=1}^{n} (Y_i - \tilde{g}(x_i))^2 + \lambda \int \tilde{g}''(x)^2 dx \tag{2.9}$$

over $\tilde{g} \in S_2[a, b]$, the set of twice continuously differentiable functions on $[a, b]$. λ is a regularization parameter. As $\lambda \to \infty$, the curve is very close to the linear regression line. As $\lambda \to 0$, the resulting curve closely fits the observations.

Definition. *A cubic spline is a function $g : [a, b] \to \mathbb{R}$ satisfies*

(i) g is a cubic polynomial on $[(a, x_1), (x_1, x_2), \dots, (x_n, b)]$.
(ii) g is twice continuously differentiable on $[a, b]$.

Proposition. *For a given $\mathbf{g} = (g_1, \dots, g_n^T)$, there exists a unique natural cubic spline g with knots $x_1, \dots, x_n$ - so $g(x_i) = g_i$ for $i = 1, \dots, n$. Moreover, there exists a nonnegative definite matrix K such that*

$$\int_a^b g''(x)^2 dx = g^T K g \tag{2.10}$$

*We call g the **natural cubic spline** interploant to g at $x_1, \dots, x_n$.*

Theorem. For any $\tilde{g} \in S_2[a, b]$ satisfying $\tilde{g}(x_i) = g_i, i = 1, \dots, n$, the cubic spline interpolant to g at $\mathbf{g} = g_1, \dots, g_n$ uniquely minimizes

$$\int_a^b \tilde{g}''(x)^2 dx \tag{2.11}$$

over $\tilde{g} \in S_2[a, b]$.

Recall that $Y_i = m(x_i) + \sigma\epsilon_i$, $m \in S_2[a, b]$, $0 < x_1 < \cdots < x_n < b$. We seek to minimize

$$\mathcal{G}_\lambda(\tilde{g}) = \sum_{i=1}^{n} (Y_i - \tilde{g}(x_i))^2 + \lambda \int_a^b \tilde{g}''(x)^2 dx \tag{2.12}$$

over $\tilde{g} \in S_2[a, b]$.

Theorem. For each $\lambda > 0$, there is a unique solution $\hat{g}$ minimizing $\mathcal{G}(\tilde{g})$over $\tilde{g} \in S_2[a, b]$. This is the natural cubic spline

$$\hat{g} = (I + \lambda K)^{-1} Y \tag{2.13}$$

Cross validation method validates the estimated curve without the i-th observation by comparing the i-th value

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i,\lambda}(x_i))^2 \tag{2.14}$$

where $\hat{g}_{-i,\lambda}$ is chosen by minimizing $\mathcal{G}_\lambda$ over all data points except the i-th,

$$\sum_{j \neq i}^{n} (Y_j - \tilde{g}(x_j))^2 + \lambda \int_a^b \tilde{g}''(x)^2 dx \tag{2.15}$$

(*)

Theorem.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - \hat{g}_\lambda(x_i)}{1 - A_{ii}})^2 \tag{2.16}$$

where $A = (I + \lambda K)^{-1}$ and

$$\int_{-\infty}^{\infty} \hat{g}_\lambda''(x)^2 dx = \hat{g}_\lambda(\mathbf{x})^T K \hat{g}_\lambda(\mathbf{x}) \tag{2.17}$$

3. Nearest Neighbour Classification

Definition. *A function $g : \mathbb{R}^d \to \{0, 1\}$ is called a classifier. If the distribution of (X, Y) are known, we can minimize the risk $\mathbb{P}(g(X) \neq Y) = L(g)$ over $g : \mathbb{R}^d \to \{0, 1\}$. The minimizer $g^\star$ is called a Bayes classifier, and $L(g^d)$ is called the Bayes risk.*

Lemma. *For a classifier $\tilde{g}$ whcih has the form*

$$\tilde{g}(x) = \begin{cases} 1 & \hat{\nu}(x) > \frac{1}{2} \\ 0 & otherwise \end{cases} \tag{3.1}$$

we have

$$\mathbb{P}(\tilde{g}(X) \neq Y) - L^\star \leq 2\mathbb{E}(\|\hat{\nu}(X) - \nu(X)) \tag{3.2}$$

Definition (k-nearest neighbour classification)**.** *A k-NN classifier g_n is defined by*

$$g_n(x) = \begin{cases} 1 & \sum_{i=1}^n W_{ni}(X)\mathbb{I}(Y_i = 1) > \sum_{i=1}^n W_{ni}(X)\mathbb{I}(Y_i = 0) \\ 0 & otherwise \end{cases} \tag{3.3}$$

which is equivalent to

$$\sum_{i=1}^{n} W_{ni}(X)\mathbb{I}(Y_i = 1) > \frac{1}{2} \iff \sum_{i=1}^{n} W_{ni}(X)Y_i > \frac{1}{2} \tag{3.4}$$

where

$$W_{ni}(X) = \frac{1}{k} \tag{3.5}$$

if X_i is a k-nearest neighbour of X, and zero otherwise.

Definition. *For a certain distirbution of (X, Y), we say g_n is consistent if $\mathbb{P}(g_n(X) \neq Y) - L^\star \to 0$.*

We say g_n is strongly consistent if

$$\mathbb{P}\Big(\lim_{n\to\infty} L(g_n) = L(g^\star)\Big) = 1 \tag{3.6}$$

Theorem. If $k \to \infty$, $\frac{k}{n} \to 0$, then for all distributions of (X, Y), the k-NN estimates g_n are consistent.

4. Minimax Lower Bounds

Definition. *As a first attempt to understand a nonparametric estimation problem, we consider a minimax risk,*

$$R(\Theta) = \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta L(\tilde{j}, \theta). \tag{4.1}$$

Definition. *If we can find our $\hat{\theta}^\star$, which minimizes $\sup_{\theta\in\Theta} \mathbb{E}_\theta L(\tilde{\theta}, \theta)$ we call $\hat{\theta}^\star$ our minimax estimator. However, it is very difficult to find $\hat{\theta}^\star$. Let $c\gamma_n \leq R(\Theta) \leq C\gamma_n$, we call γ_n is minimax rate of convergence.*

Lemma (Le Cam's two points lemma)**.** *Let $\mathcal{P}$ be probability measures on $(\mathcal{X}, \mathcal{A})$, and let (Θ, d) be the pseudo-metric space, with*

$$d : \Theta \times \Theta \to [0, \infty) \tag{4.2}$$

given by

$$d(\theta_1, \theta_2) = d(\theta_2, \theta_1), d(\theta_1, \theta_2) + d(\theta_2, \theta_3) \geq Ad(\theta_1, \theta_3) \tag{4.3}$$

Let $\theta : \mathcal{P} \to \Theta$, $\theta(P)$ is the parameter of interest ($P \in \mathcal{P}$). With $\theta_0 = \theta(P_0)$, $\theta_1 = \theta(P_1)$, under two conditions,

(i) $d(\theta_0, \theta_1) \geq \delta > 0$,
(ii) $h^2(P_0, P_1) \leq C < 1$

where $h^2(P_0, P1)$ is the Hellinger distance $\int(\sqrt{dP_0} - \sqrt{dP_1})^2$, the we have for all estimators $\tilde{\theta}$,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P d(\tilde{\theta}, \theta(P)) \geq \frac{A\delta}{2}(1 - \sqrt{C}) \tag{4.4}$$

Theorem (Nonparametric regression)**.** Let $Y_i = m(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, 1)$, $x_i = \frac{i}{n}$, $m \in \Theta$ with Θ the set of all twice continuously differentiable functions on $[0, 1]$, $m''(x) < \infty$. Then for any estimator $\tilde{m}$ and any $x_0 \in [0, 1]$,

$$\sup_{m\in\Theta} \mathbb{E}\big((\tilde{m}(x) - m(x_0))^2\big) \geq Cn^{-\frac{4}{5}} \tag{4.5}$$

5. Extreme Value Theory

Let X_n be an IID sample from a distribution function F, and denote $X_{(n)} = \max\{X_1, \ldots, X_n\}$ as the maximum order statistic.

Without any normalization, $X_{(n)} \to x_\star = \inf\{x : F(x) = 1\}$.

This is not overly interesting, since the limit distribution is degenerate (we call F non-degenerate if there does not exists $a \in \mathbb{R}$ such that $F(x) = \mathbb{I}(x \geq a)$)

We may ask if there exists $\{a_n\} > 0$, $\{b_n\} > 0$, and a nondegenerate G such that

$$\mathbb{P}\Big(\frac{X_{(n)} - b_n}{a_n} \leq x\Big) \to G(x) \tag{5.1}$$

for all continuity points x of G

Classical extreme value theory starts by asking:

(i) What kind of G appears in the limit of (5.1)?
(ii) Cna we characterize F such that (5.1) holds for a specific limit distribution G?

For the first question, we have the Extremal Types theorem. For the second question, we have the "domain of attraction" problem.

Recall that $\mathbb{P}(X_{(n)} \leq x) = F(x)^n$. We say that F is in the domain of attraction of G ($F \in D(G)$) if there exists $\{a_n\} > 0$, $\{b_n\}$ and a non-degenerate G such that

$$\mathbb{P}\Big(\frac{X_{(n)} - b_n}{a_n} \leq x\Big) = [F(a_n x + b_n)^n \to G(x) \text{ for all continuity points } x \text{ of } G]. \tag{5.2}$$

and write $F(a_n x + b_n)^n \hookrightarrow G(x)$.

We say that G_1 and G_2 are of same type if $G_1(ax + b) = G_2(x)$ for some $a > 0, b$.

The next lemma shows that if $F \in D(G_1)$ and $F \in D(G_2)$, then G_1 and G_2 are of the same type.

Lemma. *Suppose X_n is an IID sample from F and there exists $\{a_n\} > 0, \{b_n\}$ and non-degenerate G such that $F(a_n x + b_n)^n \hookrightarrow G(x)$. Then there exists $\{\alpha_n\} > 0, , \{\beta_n\}$ and non-degenerate $G_\star$ such that $F(\alpha_n x_{\beta_n})^n \hookrightarrow G_\star(x)$. if and only if $\frac{\alpha_n}{a_n} \to a$ for some $a > 0$, and $\frac{\beta_n - \beta}{a_n} \to b$ for some b.*

Then we can let $G_\star(x) = G(ax + b)$.

Definition. *G is **max-stable** if for every $n \in \mathbb{N}$, there exists $\{a_n\} > 0, \{b_n\}$ such that $G^n(a_n x + b_n) = G(x)$*

Theorem. $D(g)$ is non-empty if and only if G is max-stable.

Theorem. If $F \in D(G)$, then G must belong to the following distributions (within type):

(i) Frechet - $G_{1,\alpha}(x) = \exp(-x^{-\alpha})$, $x > 0$, $\alpha > 0$
(ii) Negative Weibull - $G_{2,\alpha} = \exp(-(-x)^\alpha)$, $x < 0$, $\alpha > 0$
(iii) Gumbel - $G_3(x) = \exp(-\exp(-x))$, $x \in \mathbb{R}$.

Conversely, these distributions can appear as such limits in (5.1).

Remark. *(i) Using $X_{(1)} = -\max\{-X_1, \ldots, -x_n\}$, we have equivalent theorems in terms of normalized minima.*
(ii) Sometimes, we cannot have nondegenreate G of normalized maxima - for example $X_1, \ldots, X_n \sim Bern(\frac{1}{2})$, $X_{(n)}$.
(iii) We can combine these three types into Generalized Extreme Value Distribution (GEV) -

$$G(x; \mu, \sigma, \gamma) = \exp(-(1 + \gamma(\frac{x - \mu}{\sigma}))^{-\frac{1}{\gamma}}) \tag{5.3}$$

with $1 + \gamma(\frac{x-\mu}{\sigma}) > 0$, $\mu \in \mathbb{R}, \gamma \in \mathbb{R}, \sigma > 0$.

We have Frechet corresponds to $\gamma > 0$, $\alpha = \frac{1}{\gamma}$, NW is $\gamma < 0$, $\alpha = -\frac{1}{\gamma}$, and Gumbel corresponds to the case where $\gamma \to 0$.

5.1. Necessary and Sufficient Conditions for Convergence.

Definition. *We say a function $l : [C, \infty] \to (0, \infty)$ is "slowly varying" if $\lim_{x\to\infty} \frac{l(tx)}{l(x)} = 1$ for all $t > 0$. For example, $l(x) = \log x, \log\log x, (\log x)^\alpha$.*

We say a function $r_\alpha : [C, \infty) \to (0, \infty)$ is "regularly varying" with an index $a \in \mathbb{R}$ if $r_\alpha(x) = x^{-\alpha} l(x)$ where l is slowly varying - so $r_2(x) = x^{-2}\log x$.

*We define an **expected residual lifetime** as*

$$R(x) = \mathbb{E}(X - x|X > x) = \frac{1}{1 - F(x)}\int_x^{x_\star}(1 - F(y))dy \tag{5.4}$$

where $x_\star = \inf\{x : F(x) = 1\}$, and $\overline{F}(x) = 1 - F(x)$

Theorem. $F \in D(G_{1,\alpha})$ if and only if $x_\star = \infty$, $\overline{F}(x) = x^{-\alpha} l(x)$ where l is slowly varying. We can choose $b_n = 0$, $a_n = F^{-1}(1 - \frac{1}{n})$ for which $F^n(a_n x + b_n) \hookrightarrow G_{1,\alpha}(x)$ is satisfied.

$F \in D(G_{2,\alpha})$ if and only if $x_\star < \infty$, $\overline{F}(x_\star - \frac{1}{x}) = x^{-\alpha} l(x)$, with l slowly varying for $x > 0$. We can choose $b_n = x_\star$, $a_n = x_\star - F^{-1}(1 - \frac{1}{n})$ for convergence.

$F \in D(G_3)$ if and only if

$$\frac{\overline{F}(x + tR(x))}{\overline{F}(x)} \to e^{-t} \tag{5.5}$$

We can choose $b_n = F^{-1}(1 - \frac{1}{n})$, $a_n = R(b_n)$.

Lemma ($\star$)**.** *Suppose there exists $a_n > 0$, b_n such that $n(1 - F(a_n x + b_n)) \to u(x)$. Then*

$$F^n(a_n x + b_n) \hookrightarrow \exp(-u(x)) \tag{5.6}$$

References