# 1

# The Basic Models

## 1.1 Introduction

The central limit theorem is a key limiting result in probability and underpins a vast range of applications. Let $S_n = X_1 + \cdots + X_n$ denote the sum of independent and identically distributed scalar random variables $X_1, \ldots, X_n$ with mean $\mu$ and finite positive variance $\sigma^2$. The central limit theorem concerns the distribution of $S_n$ as $n \to \infty$. Since its mean and variance are $n\mu$ and $n\sigma^2$, the sum $S_n$ must be centered and scaled in order to have a non-trivial limiting distribution. This is achieved by considering the distribution of $Z_n = (S_n - b_n)/a_n$, where $b_n = n\mu$ and $a_n^2 = n\sigma^2$, which has a limiting standard normal, $N(0,1)$, distribution when $n \to \infty$. Formally we write

$$Z_n = \frac{S_n - b_n}{a_n} \xrightarrow{D} Z, \quad n \to \infty,$$

where $\xrightarrow{D}$ denotes convergence in distribution, a weak form of convergence under which $\Pr(Z_n \leq z)$ converges pointwise to $\Pr(Z \leq z)$ for every $z \in \mathbb{R}$ at which the latter function is continuous; since this limit is the standard normal distribution function, $\Phi(z)$, convergence occurs for all $z$. Although weak, this mode of convergence is statistically useful because it yields the approximation $\Pr(Z_n \leq z) \approx \Phi(z)$ for large enough $n$.

The central limit theorem has wide applications to data analysis because it provides finite-sample approximations for the distributions of sums and related quantities such as averages. These approximations allow one to write $S_n \dot{\sim} N(b_n, a_n^2)$ for finite—but preferably large—$n$, and often the values of $b_n$ and $a_n$ can be estimated. The validity of the underlying conditions can be assessed by suitable diagnostic plots. Moreover under certain conditions the normal limit applies also to sums of non-identically distributed or dependent variables, so the approximation applies more broadly.

An analogous result, the *extremal types theorem*, applies to sample maxima.

**Theorem 1.1 (Extremal types)** Let $M_n = \max(X_1, \ldots, X_n)$ be the maximum of a random sample $X_1, \ldots, X_n$ with distribution function $F$. If sequences $(a_n) > 0$ and $(b_n)$ can be chosen in such a way that the centred and scaled sample maximum, $(M_n - b_n)/a_n$, has a non-degenerate limiting distribution $G$, then this must be the *generalized extreme-value distribution*,

$$G(x) = \begin{cases} \exp\left[-\{1+\xi(x-\eta)/\tau\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\{-(x-\eta)/\tau\}\right], & \xi = 0, \end{cases} \quad x \in \mathbb{R}, \tag{1.1}$$

where we write $a_+ = \max(a, 0)$ for any real $a$, and where $\xi, \eta \in \mathbb{R}$ and $\tau > 0$. Put another way, $(M_n - b_n)/a_n \xrightarrow{D} Z$ as $n \to \infty$, where $Z$ has distribution function $G$. □

The 'types' are the qualitatively different distributions that arise for $\xi = 0$, $\xi > 0$ and $\xi < 0$, which are usually combined into (1.1) for statistical purposes; they are discussed in Section 1.3.

Theorem 1.1 provides a natural model for sample maxima. For example, when confronted with 30 years of annual maximum windspeeds at some location and asked to estimate the largest windspeed that might arise there over the next century, it seems natural to base the necessary extrapolation on (1.1), which often fits the data fairly well—the annual maximum is the largest of 365 daily maxima, and we might hope that $G(x)$ provides an adequate approximation to its distribution. This hope might be misplaced, however: if the daily observations show dependence and seasonality then the number of 'independent' observations from which the maximum is computed might be much smaller than 365, and then the usefulness of (1.1) might be questionable. Moreover, reducing the data to annual maxima is often undesirable, since other observations also contain information about the extremes. It turns out that a related result holds for another natural definition of rare events, namely those observations that exceed a high threshold.

**Theorem 1.2 (Exceedances)** Let $X$ be a random variable having distribution function $F$, and suppose that a function $c(u)$ can be chosen so that the limiting distribution, $H$, of $(X-u)/c(u)$, conditional on $X > u$, is non-degenerate as $u$ approaches the upper support point $x^* = \sup\{x : F(x) < 1\}$ of $X$. If such an $H$ exists, it must be the *generalized Pareto distribution*,

$$H(x) = \begin{cases} 1 - (1+\xi x/\sigma)_+^{-1/\xi} & \xi \neq 0, \\ 1 - \exp(-x/\sigma), & \xi = 0, \end{cases} \quad x > 0, \tag{1.2}$$

where $\xi \in \mathbb{R}$ and $\sigma > 0$. Put another way, as $x \to x^*$,

$$\Pr\{(X-u)/c(u) > x \mid X > u\} \to 1 - H(x).$$

□

Comparison of (1.1) and (1.2) suggests that these results must be closely related, and in fact they hold under the same conditions, with the same value of $\xi$, and with $\sigma = \tau + \xi(u - \eta)$. Because of this close connection, below we refer to the two distributions collectively as the *extremal distributions*.

In this chapter we derive these theorems and some closely related results and discuss some of their implications. Our arguments also apply to sample minima and exceedances below thresholds, since

$$\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n)$$

and $X < u$ precisely when $-X > -u$. Thus in theoretical discussion, and indeed in data analysis, we can consider whichever tail of the distribution is most convenient—provided we later remember to reverse any changes of sign!

## 1.2 Convergence of Extremes

### 1.2.1 Maxima

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function $F$ and let $x^* = \sup\{x : F(x) < 1\}$ denote their upper support point. Then, by independence, their maximum $M_n$ satisfies

$$\Pr(M_n \leq x) = \Pr(X_1 \leq x, \ldots, X_n \leq x) = \prod_{j=1}^{n} \Pr(X_j \leq x) = F(x)^n,$$

which tends to the degenerate distribution function $I(x \geq x^*)$ as $n \to \infty$, where $I(\cdot)$ is an indicator function. This is not a useful limit. Like the sum appearing in the central limit theorem, the maximum $M_n$ must be centred and scaled for a non-degenerate limit to emerge, and we therefore consider the standardized version $(M_n - b_n)/a_n$, where the real-valued sequences $b_n$ and $a_n > 0$ must be chosen so that for $x \in \mathbb{R}$,

$$\begin{aligned} \lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \lim_{n\to\infty} \Pr\left(M_n \leq b_n + a_n x\right) \\ &= \lim_{n\to\infty} F(b_n + a_n x)^n \\ &= \lim_{n\to\infty} \{1 - \Lambda_n(x)/n\}^n \\ &= \exp\{-\Lambda(x)\}, \end{aligned} \tag{1.3}$$

say, where we have set

$$\Lambda_n(x) = n\{1 - F(b_n + a_n x)\} \tag{1.4}$$

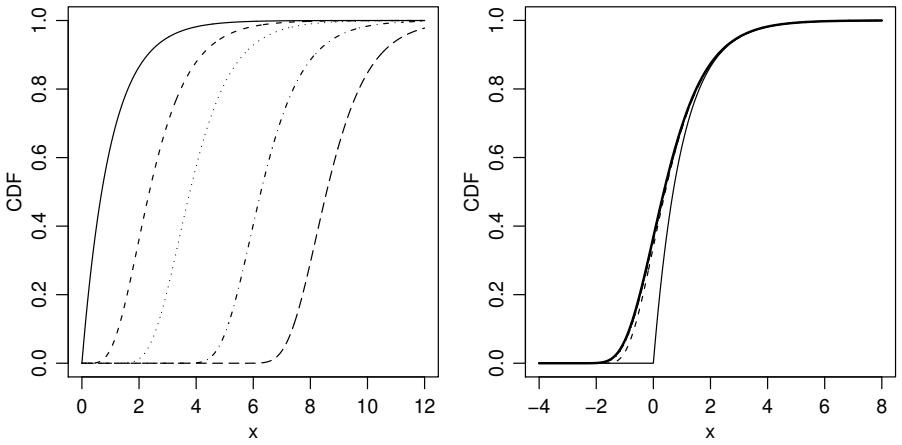


**Figure 1.1** Cumulative distribution functions of maxima (left) and renormalized maxima (right) of $m = 1$, 7, 30, 365, 3650 standard exponential variables (from left to right), with limiting Gumbel distribution (heavy).

and $\lim_{n\to\infty} \Lambda_n(x) = \Lambda(x)$. Note that the convergence of $(M_n - b_n)/a_n$ to a non-degenerate limiting random variable is equivalent to pointwise convergence of $\Lambda_n(x)$ to $\Lambda(x)$ in the set where the latter is finite.

If the limit is a distribution function $G(x)$, then $\Lambda(x)$ must be a decreasing function of $x$ satisfying

$$\lim_{x\to-\infty} \Lambda(x) = \infty, \quad \lim_{x\to\infty} \Lambda(x) = 0.$$

Moreover, if $\Lambda(x) = \infty$ for $x < x^-$ and/or $\Lambda(x) = 0$ for $x > x^+$, then the limiting distribution (1.3) places probability only in the interval $[x^-, x^+]$. If $G$ is non-degenerate, then clearly $x^-$ must be strictly less than $x^+$.

**Example 1.3 (Exponential distribution)** Let $F(x) = 1 - \exp(-x)$ for $x > 0$; then $x^* = \infty$. Provided $b_n + a_n x > 0$,

$$F(b_n + a_n x)^n = \left[1 - \exp\left\{-(b_n + a_n x)\right\}\right]^n,$$

so if we set $b_n = \log n$ and $a_n = 1$, then since any fixed $x$ will ultimately be larger than $-\log n$,

$$G(x) = \lim_{n\to\infty} F(b_n + a_n x)^n = \lim_{n\to\infty} \left(1 - \frac{e^{-x}}{n}\right)^n = \exp\left(-e^{-x}\right), \quad x \in \mathbb{R},$$

which is (1.1) with $\eta = 0$, $\tau = 1$ and $\xi = 0$.

In this case $\Lambda(x) = e^{-x}$, $x^- = -\infty$ and $x^+ = \infty$.

Figure 1.1 shows the behaviour of un-normalised and normalised maxima of exponential random samples. The normalised maxima appear to converge rapidly to the limiting distribution. □

**Example 1.4 (Uniform distribution)** Let $F(x) = x$ for $x \in [0, 1]$; then

$x^* = 1$. Provided $0 \le b_n + a_n x \le 1$, we have

$$F(b_n + a_n x)^n = (b_n + a_n x)^n,$$

so if we set $b_n = 1$, $a_n = 1/n$ and $x \le 0$, we have $(b_n + a_n x)^n \to e^x$. Since the limit must be a distribution function, we must take

$$\Lambda(x) = \begin{cases} -x, & x \le 0, \\ 0, & x > 0; \end{cases}$$

thus $\Lambda(x) = (-x)_+$. Clearly $\Lambda$ is decreasing, $x^- = -\infty$ and $x^+ = 0$, so

$$G(x) = \begin{cases} e^x, & x \le 0, \\ 1, & x > 0. \end{cases} \tag{1.5}$$

This is the distribution function of $-W$, where $W$ is a standard exponential variable. It is straightforward to check that (1.5) is (1.1) with $\eta = 1$, $\tau = 1$ and $\xi = -1$. □

**Example 1.5 (Pareto distribution)** Let $F(x) = 1 - x^{-\alpha}$ for $x > 1$ and $\alpha > 0$; then $x^* = \infty$. Provided $b_n + a_n x > 1$, we have

$$F(b_n + a_n x)^n = \left\{1 - (b_n + a_n x)^{-\alpha}\right\}^n$$

and if $b_n = 0$ and $a_n = n^{1/\alpha}$, and if $x > 0$ so that $x$ is ultimately larger than $n^{-1/\alpha}$, then

$$G(x) = \lim_{n\to\infty} F(b_n + a_n x)^n = \lim_{n\to\infty} \left(1 - \frac{x^{-\alpha}}{n}\right)^n = \exp\left(-x^{-\alpha}\right), \quad x > 0,$$

which is (1.1) with $\eta = 1$, $\tau = 1/\alpha$ and $\xi = 1/\alpha$.

In this case

$$\Lambda(x) = \begin{cases} \infty, & x \le 0, \\ x^{-\alpha}, & x > 0, \end{cases}$$

and $x^- = 0$, $x^+ = \infty$. □

An obvious question is whether these limits are unique: could other normalising sequences $\{a'_n\}$ and $\{b'_n\}$ lead to different limits? It turns out that the only other possibility is that any sequence leading to a non-degenerate limiting distribution as $n \to \infty$ must satisfy $(b_n - b'_n)/a'_n \to \eta$ and $a_n/a'_n \to \tau > 0$, corresponding to replacing $\Lambda(x)$ by $\Lambda\{(x-\eta)/\tau\}$. At the end of this section we establish the following result, from which Theorem 1.1 immediately follows.

**Lemma 1.6** The only possible non-degenerate limit for (1.4) as $n \to \infty$ is

$$\Lambda(y) = \begin{cases} \{1 + \xi(y-\eta)/\tau\}_+^{-1/\xi}, & \xi \ne 0, \\ \exp\{-(y-\eta)/\tau\}, & \xi = 0, \end{cases} \quad y \in \mathbb{R}, \tag{1.6}$$

where $\eta$ and $\xi$ are real-valued and $\tau$ is positive. □

We prove this in Section 1.2.3, but first explore some of its consequences. As the limiting distribution for a rescaled sample maximum is

$$G(y) = \exp\left\{-\Lambda(y)\right\}, \quad y \in \mathbb{R},$$

the corresponding probability density function is

$$-\dot{\Lambda}(y)\exp\left\{-\Lambda(y)\right\}, \quad y \in \mathbb{R}, \tag{1.7}$$

where

$$-\dot{\Lambda}(y) = -\frac{\mathrm{d}\Lambda(y)}{\mathrm{d}y} = \begin{cases} \tau^{-1}\left\{1+\xi(y-\eta)/\tau\right\}_+^{-1/\xi-1}, & \xi \neq 0, \\ \tau^{-1}\exp\left\{-(y-\eta)/\tau\right\}, & \xi = 0, \end{cases} \tag{1.8}$$

is non-negative because $\Lambda(y)$ is decreasing. The transformation $y \mapsto \Lambda(y)$ is strictly monotonic decreasing on the interval where $\Lambda(y)$ is finite, so if a random variable $Y$ has distribution $G$ and if $x > 0$, then

$$\begin{aligned} \Pr\{\Lambda(Y) \leq x\} &= \Pr\{Y \geq \Lambda^{-1}(x)\} \\ &= 1 - \exp\left[-\Lambda\left\{\Lambda^{-1}(x)\right\}\right] \\ &= 1 - \exp(-x), \end{aligned}$$

i.e., $W = \Lambda(Y)$ has a standard exponential distribution. Thus $Y$ can be generated as $\Lambda^{-1}(W)$.

*Largest order statistics*

It is straightforward to extend the limiting density (1.7) for the maximum to a fixed number of upper order statistics. Consider the $r$ largest order statistics $Y_r \leq \cdots \leq Y_1$ of the rescaled variables $\{(X_j - b_n)/a_n : j = 1, \ldots, n\}$, and suppose that $n \to \infty$. As $Y_1$ has the limiting distribution of the rescaled maximum $(M_n - b_n)/a_n$,

$$\Pr(Y_1 \leq y_1) = \exp\left\{-\Lambda(y_1)\right\}, \quad y_1 \in \mathbb{R},$$

and $f_{Y_1}(y_1) = \{-\dot{\Lambda}(y_1)\}\exp\{-\Lambda(y_1)\}$. The second-largest variable $Y_2$ is also the largest of an infinite number of these rescaled variables but cannot exceed $Y_1$, so below $y_1$ the distributions of $Y_2$ and $Y_1$ must be proportional. Hence

$$\Pr(Y_2 \leq y_2 \mid Y_1 = y_1) = \begin{cases} \exp\left\{\Lambda(y_1) - \Lambda(y_2)\right\}, & y_2 < y_1, \\ 1, & y_2 \geq y_1, \end{cases}$$

with corresponding conditional density

$$f_{Y_2|Y_1}(y_2 \mid y_1) = \{-\dot{\Lambda}(y_2)\}\exp\left\{\Lambda(y_1) - \Lambda(y_2)\right\}, \quad y_2 < y_1.$$

Evidently the same argument applies to $Y_3$, $Y_4$, and so forth, and after cancellations in the exponent the joint density of $Y_1, \ldots, Y_r$ is found to be

$$\begin{aligned} f_{Y_1,\ldots,Y_r}(y_1,\ldots,y_r) &= f_{Y_1}(y_1)\prod_{j=2}^{r} f_{Y_j|Y_{j-1}}(y_j \mid y_{j-1}) \\ &= \exp\{-\Lambda(y_r)\} \times \prod_{j=1}^{r} \left\{-\dot{\Lambda}(y_j)\right\}, \end{aligned} \tag{1.9}$$

where $y_r < \cdots < y_1$; this reduces to (1.7) when $r = 1$. Expression (1.9) allows inference from the $r$ highest or, with the appropriate changes lowest, values of a large sample.

The fit of this model can be checked using the fact that $\Lambda(Y_1)$, $\Lambda(Y_2) - \Lambda(Y_1)$ and so forth have independent standard exponential distributions. To see this, note that as $\Lambda(y)$ is decreasing in $y$,

$$\Pr(Y_1 \le y_1) = \Pr\{\Lambda(Y_1) \ge \Lambda(y_1)\} = \exp\{-\Lambda(y_1)\}, \quad \Lambda(y_1) > 0,$$

which implies that $\Lambda(Y_1)$ has a standard exponential distribution, and likewise

$$\begin{aligned} \Pr(Y_2 \le y_2 \mid Y_1 = y_1) &= \Pr\{\Lambda(Y_2) \ge \Lambda(y_2) \mid Y_1 = y_1\} \\ &= \exp\left[-\{\Lambda(y_2) - \Lambda(y_1)\}\right], \quad \Lambda(y_2) - \Lambda(y_1) > 0, \end{aligned}$$

yields that $\Lambda(Y_2) - \Lambda(y_1)$ is also standard exponential, conditional on $Y_1 = y_1$. But since this distribution does not depend on $y_1$, the result is also true unconditionally, i.e., $\Lambda(Y_2) - \Lambda(Y_1) \sim \exp(1)$, independent of $\Lambda(Y_1)$. By recursion we see that the differences $\Lambda(Y_j) - \Lambda(Y_{j-1})$ are independent and standard exponential for $j = 2, 3, \ldots$. Thus if an estimate $\widehat{\Lambda}$ of $\Lambda$ is available, then $\widehat{\Lambda}(y_1)$, $\widehat{\Lambda}(y_2) - \widehat{\Lambda}(y_1), \ldots$ should be close to a standard exponential sample, systematic departures from which suggest that the model is poor.

*Caveats*

Theorem 1.1 states not that maxima must follow the generalized extreme-value distribution, but rather that if a limiting distribution for maxima exists, then it must be of form (1.1). Here is an example for which no limit exists.

**Example 1.7 (Logarithmic distribution)** Consider the distribution function

$$F(x) = 1 - (\log x)^{-1}, \quad x > e.$$

In this case $\Lambda_n(x) = n/\log(b_n + a_n x)$, and if a limit $\Lambda(x)$ exists, then we must have $b_n + a_n x \to \infty$ as $n \to \infty$. Now

$$n^{-1}\log(b_n + a_n x) = n^{-1}\log a_n + n^{-1}\log(x + b_n/a_n),$$

and if the sequence $b_n/a_n$ is bounded, then any limit cannot depend on $x$. If $b_n/a_n$ is unbounded, then it must ultimately be positive, and then $\log(x +$

$b_n/a_n) \sim \log(b_n/a_n)$, and again the limit cannot depend on $x$. Thus in this case no sequences exist for which linear renormalisation of the maximum can yield a non-degenerate limiting distribution; the maxima grow too fast.

If $Y$ is Pareto with $\alpha = 1$, however, then $X = \exp(Y)$ has the distribution above, so a limiting distribution exists for $\log X$ when linearly rescaled. □

### 1.2.2 Poisson process approximation

We now outline how the convergence of maxima implies that of the rescaled sample values to a Poisson process. We first recall some useful facts about moment-generating functions and the Poisson and related distributions.

The moment-generating function of a scalar random variable $X$ is defined as $M_X(s) = \mathrm{E}\{\exp(sX)\}$, if this is finite for values of $s$ within some open set $\mathcal{S}$ containing the origin. The derivatives of $M_X(s)$ can be used to find the moments of $X$; moreover, if $M_1(s)$ and $M_2(s)$ are identical within $\mathcal{S}$, then the corresponding distributions are also identical, i.e., there is a one-one mapping between distributions and moment-generating functions. The latter are also useful in establishing convergence results: if as $n \to \infty$ the moment-generating functions $\{M_n\}$ of a sequence of random variables $\{X_n\}$ converge pointwise to $M_X$ within $\mathcal{S}$, then the random variables converge in distribution, $X_n \xrightarrow{D} X$.

These properties also hold when $X$ is vector-valued; then $s$ has the dimension of $X$ and $\mathcal{S}$ is an open set containing the origin.

Moment-generating functions have many other uses in basic statistics, some of which are explored in Problem **??.??**.

**Example 1.8 (Poisson distribution)** The moment-generating function of a Poisson random variable $X$ with mean $\lambda > 0$ is

$$\mathrm{E}(e^{sX}) = \sum_{x=0}^{\infty} e^{sx} \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=0}^{\infty} \frac{(\lambda e^s)^x}{x!} e^{-\lambda} = \exp\{\lambda(e^s - 1)\}, \quad s \in \mathbb{R}.$$

Likewise the joint moment-generating function of two independent Poisson variables with means $\lambda_1$ and $\lambda_2$ is

$$\begin{aligned} \mathrm{E}(e^{s_1 X_1 + s_2 X_2}) &= \mathrm{E}(e^{s_1 X_1})\mathrm{E}(e^{s_2 X_2}) \\ &= \exp\{\lambda_1(e^{s_1} - 1) + \lambda_2(e^{s_2} - 1)\}, \quad s_1, s_2 \in \mathbb{R}. \end{aligned} \tag{1.10}$$

The moment-generating function of $X_1 + X_2$ is obtained by setting $s_1 = s_2 = s$, in which case (1.10) reduces to $\exp\{(\lambda_1 + \lambda_2)(e^s - 1)\}$, which corresponds to a Poisson variable with mean $\lambda_1 + \lambda_2$. Clearly a finite sum of independent Poisson variables also has a Poisson distribution, and likewise for an infinite sum of independent Poisson variables for which $\sum_{j=1}^{\infty} \lambda_j$ is finite. □

**Example 1.9 (Multinomial distribution)** Let $X = (X_0, \ldots, X_D)$ be a

multinomial random variable with probability mass function

$$\Pr(X_0 = x_0, \ldots, X_D = x_D) = \frac{n!}{x_0! \cdots x_D!} p_0^{x_0} \cdots p_D^{x_D},$$

where the $p_d$ are probabilities that sum to unity and $x = (x_0, \ldots, x_D)$ lies in the set $\mathcal{X}$ of $(D+1)$-tuples of non-negative integers that sum to $n$. The corresponding moment-generating function is

$$M_X(s) = \sum_{x \in \mathcal{X}} \exp(s_0 x_0 + \cdots + s_D x_D) \frac{n!}{x_0! \cdots x_D!} p_0^{x_0} \cdots p_D^{x_D},$$

where $s = (s_0, \ldots, s_D)$, and the multinomial theorem gives

$$M_X(s) = \left\{ \sum_{d=0}^{D} p_d \exp(s_d) \right\}^n, \quad s_0, \ldots, s_D \in \mathbb{R}.$$

Suppose now that we seek the joint distribution of $X_1, \ldots, X_D$ as $n \to \infty$ and $np_d \to \lambda_d > 0$ for $d = 1, \ldots, D$. We set $s_0 = 0$ and write

$$\sum_{d=0}^{D} p_d \exp(s_d) = p_0 + \sum_{d=1}^{D} p_d e^{s_d} = 1 - n^{-1} \sum_{d=1}^{D} np_d + n^{-1} \sum_{d=1}^{D} np_d e^{s_d}$$

using the fact that $p_0 = 1 - (p_1 + \cdots + p_D)$. As $n \to \infty$ we therefore have

$$\begin{aligned} \mathrm{E}\left\{\exp(s_1 X_1 + \cdots + s_D X_D)\right\} &= \left\{ 1 + \frac{1}{n} \sum_{d=1}^{D} np_d (e^{s_d} - 1) \right\}^n \\ &\to \exp\left\{ \sum_{d=1}^{D} \lambda_d (e^{s_d} - 1) \right\}, \quad s_1, \ldots, s_D \in \mathbb{R}, \end{aligned}$$

which we see by comparison with (1.10) is the moment-generating function of $D$ independent Poisson random variables with means $\lambda_1, \ldots, \lambda_D$.

If $D = 1$ then the binomial variable $X_1$ with probability $p_1$ converges to a Poisson variable with mean $\lambda$ when $np_1 \to \lambda$. This *law of small numbers* underpins the classical Poisson approximation for binomial probabilities. □

*Limiting results for background variables*

We saw previously that the convergence of maxima is equivalent to

$$\Lambda_n(x) = n\{1 - F(b_n + a_n x)\} \to \Lambda(x), \quad n \to \infty,$$

for any $x$ at which $\Lambda(x)$ is finite. To parlay this into further results, suppose that we have an infinite series of *background variables* $X_j$ whose extremes are of interest; these might be successive rainfall amounts. We assume the $X_j$ are independent and identically distributed with distribution $F$ and define the point patterns

$$\mathcal{P}_n = \{(j/n, (X_j - b_n)/a_n) : j \in \mathbb{Z}\}, \quad n = 1, 2, \ldots,$$

in the plane $\mathcal{E} = \mathbb{R}^2$; the horizontal axis represents time, with $n$ background observations per unit of time, and the vertical axis represents their sizes. In order to obtain a limiting process that can be used as an approximation for finite samples, we suppose that $n \to \infty$, giving more and more background variables in each unit of time, and we renormalise these variables so that the maximum in each such unit has a limiting GEV distribution. We shall show that the number of points of $\mathcal{P}_n$ in any useful subset $\mathcal{A} \subset \mathcal{E}$ converges to a Poisson random variable with mean $\mu(\mathcal{A})$, and that the numbers of points in disjoint sets are independent Poisson variables. If $\mathcal{A} = (t', t] \times (x', x]$ is a finite rectangle then

$$\mu(\mathcal{A}) = (t - t')\{\Lambda(x') - \Lambda(x)\}, \quad t' < t, x' < x, \tag{1.11}$$

and the means for more complex sets are defined by addition: if $\mathcal{A}$ is a union of disjoint rectangles $\mathcal{A}_1, \ldots, \mathcal{A}_k$, then $\mu(\mathcal{A}) = \sum_{j=1}^{k} \mu(\mathcal{A}_j)$, provided this sum is finite. Equivalently we can integrate the *intensity function*

$$\dot{\mu}(t, x) = \frac{\partial^2 \mu(\mathcal{A})}{\partial t \partial x} = -\dot{\Lambda}(x), \quad t, x \in \mathbb{R},$$

over $\mathcal{A}$, giving

$$\mu(\mathcal{A}) = \int_{\mathcal{A}} \dot{\mu}(t, x)\, \mathrm{d}t\mathrm{d}x.$$

The limiting process of points on $\mathcal{E}$ is called a *Poisson point process* and the function $\mu(\cdot)$ is called its *measure*. The requirement that a Poisson variable should have a finite mean implies that it is illegitimate to consider the number of points in sets of infinite measure, such as $\mathcal{E}$.

To see where these results come from, let $\mathcal{A} = (t', t] \times (x', x]$ be a finite rectangle in $\mathcal{E}$ and let $N_n(\mathcal{A})$ denote the number of points of $\mathcal{P}_n$ in $\mathcal{A}$. Clearly $N_n(\mathcal{A})$ is a binomial variable with denominator $\lfloor nt \rfloor - \lfloor nt' \rfloor$ and probability

$$\Pr\{x' < (X_j - b_n)/a_n \leq x\} = \{\Lambda_n(x') - \Lambda_n(x)\}/n,$$

so the law of small numbers implies that $N_n(\mathcal{A})$ converges to a limiting variable $N(\mathcal{A})$ whose distribution is Poisson with mean

$$(t - t')\{\Lambda(x') - \Lambda(x)\} = \mu(\mathcal{A}).$$

This argument holds for any rectangle for which $\mu(\mathcal{A}) < \infty$, including the infinite rectangle $(t', t] \times (u, \infty)$, provided that $\Lambda(u) < \infty$.

To find the joint limiting joint distribution for disjoint finite rectangles $\mathcal{A}_1, \ldots, \mathcal{A}_k$, we first suppose that their projections onto the horizontal axis are disjoint. The binomial variables $N_n(\mathcal{A}_1), \ldots, N_n(\mathcal{A}_k)$ are independent because the background variables contributing to them are independent, so the $N_n(\mathcal{A}_j)$ converge jointly to independent Poisson variables $N(\mathcal{A}_j)$ with means $\mu(\mathcal{A}_j)$.

Now suppose that $\mathcal{A}_1$ and $\mathcal{A}_2$ are disjoint but their projections onto the

horizontal axis overlap. Without loss of generality we can take $\mathcal{A}_1 = (t', t] \times (x', x]$ and $\mathcal{A}_2 = (t', t] \times (y', y]$, where $x < y'$, and let $N_0$ count the remaining background variables in the interval $(t', t]$. Then $N_0$, $N_n(\mathcal{A}_1)$ and $N_n(\mathcal{A}_2)$ have a multinomial distribution with denominator $\lfloor nt \rfloor - \lfloor nt' \rfloor$ and the last two have probabilities

$$\{\Lambda_n(x') - \Lambda_n(x)\}/n, \quad \{\Lambda_n(y') - \Lambda_n(y)\}/n.$$

Hence the argument in Example 1.9 implies that, as $n \to \infty$, $N_n(\mathcal{A}_1)$ and $N_n(\mathcal{A}_2)$ converge to independent Poisson variables with means $\mu(\mathcal{A}_1)$ and $\mu(\mathcal{A}_2)$. Clearly the same would be true for disjoint sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$.

The above argument is illustrated in Figure 1.2, which shows realisations of the processes $\mathcal{P}_{10}$ and $\mathcal{P}_{1000}$ on the interval $[0, 10]$ for independent standard exponential background variables, and of the limiting Poisson process. The counts $N_n(\mathcal{A}_1)$, $N_n(\mathcal{A}_2)$ and $N_n(\mathcal{A}_4)$ are mutually independent for any $n$ because they are based on disjoints sets of background variables. The same is true for the triplet $N_n(\mathcal{A}_1)$, $N_n(\mathcal{A}_3)$ and $N_n(\mathcal{A}_3)$, but $N_n(\mathcal{A}_2)$ and $N_n(\mathcal{A}_3)$ are dependent for any $n$ because they are based on the same background variables. All these counts have limiting Poisson distributions, and they are all independent in the limit because the degree of dependence between $N_n(\mathcal{A}_2)$, but $N_n(\mathcal{A}_3)$ diminishes to zero. The argument extends to any finite collection of rectangles.

We see from above that if $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are disjoint rectangles, then

$$N_n(\mathcal{A}_1), \ldots, N_n(\mathcal{A}_k) \quad \xrightarrow{D} \quad N(\mathcal{A}_1), \ldots, N(\mathcal{A}_k),$$

which are independent Poisson variables with means $\mu(\mathcal{A}_1), \ldots, \mu(\mathcal{A}_k)$. Thus if $\mathcal{A} = \bigcup_{j=1}^{k} \mathcal{A}_j$, the sum $N_n(\mathcal{A}) = \sum_{j=1}^{k} N_n(\mathcal{A}_j)$ converges to a Poisson variable $N(\mathcal{A})$ with mean $\mu(\mathcal{A}) = \sum_{j=1}^{k} \mu(\mathcal{A}_j)$; this holds also for an infinite union of rectangles provided $\mu(\mathcal{A}) = \sum_{j=1}^{\infty} \mu(\mathcal{A}_j) < \infty$. Any non-pathological set can be constructed as a limit of unions of disjoint rectangles, so the count for any set likely to arise in a statistical context can be approximated by a Poisson variable, provided the corresponding mean is finite.

To reconnect this discussion with maxima, consider the largest value $Y$ of the limiting process in the set $(0, 1] \times \mathbb{R}$. This maximum is no larger than $y$ if and only if the set $\mathcal{A}_y = (0, 1] \times (y, \infty)$ is empty, so the Poisson distribution of $N(\mathcal{A}_y)$ gives

$$\Pr(Y \leq y) = \Pr\{N(\mathcal{A}_y) = 0\} = \exp\{-\mu(\mathcal{A}_y)\} = \exp\{-\Lambda(y)\},$$

which is of generalized extreme-value form. Moreover the independence of the background variables implies that maxima for disjoint time periods are independent, and that the distribution of a maximum over $T$ time units is $\exp\{-T\Lambda(y)\}$.

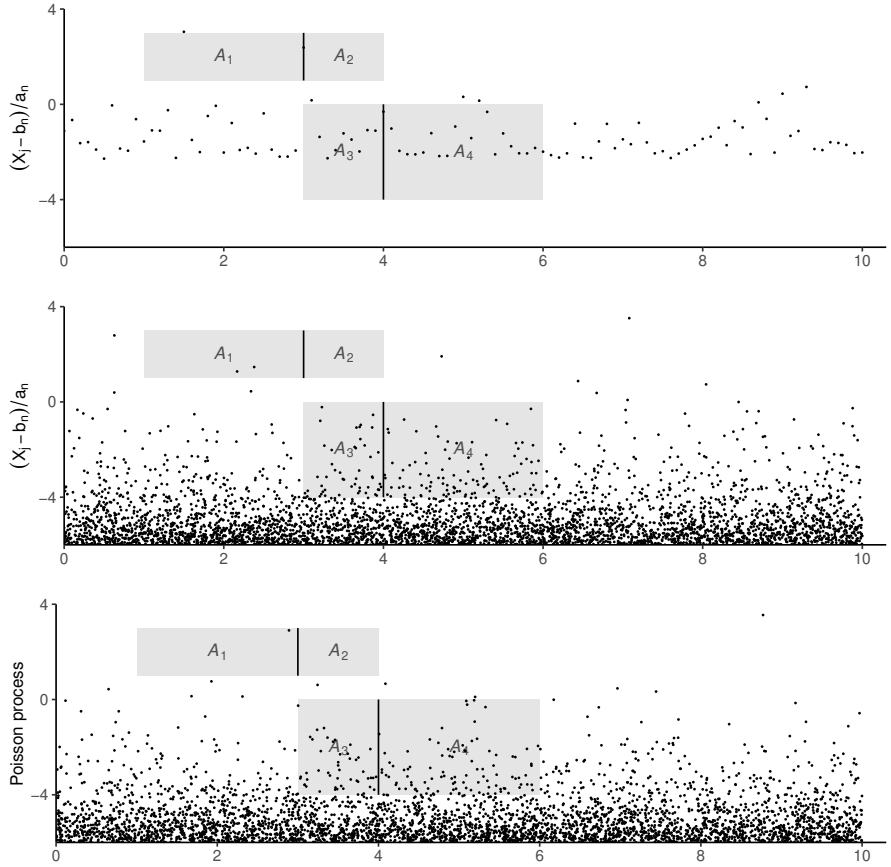


**Figure 1.2** Poisson process convergence for standard exponential variables. Top and middle panels: processes $\mathcal{P}_n$ on $[0, 10]$ for $n = 10$, 1000. Bottom panel: limiting Poisson process $\mathcal{P}$. The counts in the sets $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_4$ or $\mathcal{A}_1, \mathcal{A}_3, \mathcal{A}_4$ are independent for any $n$ because they depend on independent background variables. The counts in $\mathcal{A}_2$ and $\mathcal{A}_3$ are based on the same background variables and so are dependent for finite $n$ but become independent in the limit.

The Poisson process representation also allows us to study the limiting pattern of exceedances over a threshold $u$ for which $\Lambda(u) < \infty$. The number of such exceedances in the interval $(0, t]$ is $N\{(0, t] \times (u, \infty)\}$, which has the Poisson distribution with mean $t\Lambda(u)$, so the probability that there are no exceedances in this interval is $\exp\{-t\Lambda(u)\}$. Equivalently, the waiting time $T$ to the first exceedance satisfies

$$\Pr(T > t) = \exp\{-t\Lambda(u)\}, \quad t > 0 : \tag{1.12}$$

$T$ has the exponential distribution with parameter $\Lambda(u)$. Since events in disjoint time intervals are independent, the waiting times between successive exceedances are also mutually independent and satisfy (1.12). Thus the times of exceedances of $u$ occur according to a *homogeneous Poisson process* of rate $\Lambda(u)$ on $\mathbb{R}$.

Theorem 1.2 follows at once from this construction. When $x > 0$, the prob-

ability that the rescaled $X$ exceeds $x+u$, conditional on it exceeding $u$,

$$\Pr\{(X-b_n)/a_n > x+u \mid (X-b_n)/a_n > u\}$$

may be written as

$$\begin{aligned}\frac{\Pr\{X > b_n + a_n(x+u)\}}{\Pr(X > b_n + a_n u)} &= \frac{n\Pr\{X > b_n + a_n(x+u)\}}{n\Pr(X > b_n + a_n u)} \\ &= \frac{\Lambda_n(x+u)}{\Lambda_n(u)} \\ &\to \frac{\Lambda(x+u)}{\Lambda(u)}, \quad n \to \infty,\end{aligned}$$

and if $\sigma_u = \tau + \xi(u-\eta) > 0$, so $\Lambda(u) > 0$ and this limit is well-defined, then

$$\frac{\{1+\xi(x+u-\eta)/\tau\}_+^{-1/\xi}}{\{1+\xi(u-\eta)/\tau\}_+^{-1/\xi}} = \begin{cases}(1+\xi x/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-x/\sigma_u), & \xi = 0.\end{cases}$$

Thus the limiting probability that $(X-b_n)/a_n < x+u$, conditional on $(X-b_n)/a_n > u$, is given by (1.2).

In applications the background observations are typically both dependent and subject to trend and seasonal variation, so the Poisson process model developed above might appear unrealistic. In fact, with minor modifications, it is robust enough to furnish the basis for widely-used statistical methods.

*Poisson process*

The Poisson point process $\mathcal{P}$ constructed above is a very special instance of a powerful general model. An axiomatic approach considers a process of points, also called events, lying in a suitable space $\mathcal{E}$ for which the number $N(\mathcal{A})$ of points in any set $\mathcal{A}$ can be unambiguously defined, and for which

- the variables $N(\mathcal{A}_1), N(\mathcal{A}_2), \ldots$ for disjoint sets $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are independent; and
- $N(\mathcal{A})$ has a Poisson distribution with mean $\mu(\mathcal{A})$.

Sets $\mathcal{A}$ for which $\mu(\mathcal{A})$ is infinite must be avoided. Moreover the measure $\mu$ must be *diffuse*, i.e., $\mu(\{x\}) = 0$ for every $x \in \mathcal{E}$. For if $\mu(\{x\}) = \lambda > 0$ for some $x$, then

$$\Pr\{N(\{x\}) \geq 2\} = 1 - e^{-\lambda} - \lambda e^{-\lambda} > 0,$$

and it would be impossible to count the points in any set containing $x$. A process in which points must occur as singletons is called *simple*. If the measure has an intensity function, defined as $\dot{\mu}(x) = \partial\mu(\mathcal{A}_x)/\partial x$, where $x$ represents the top right-hand corner of a finite rectangle $\mathcal{A}_x$, and if events $x_1, \ldots, x_n$ are

observed in $\mathcal{A}$, then the corresponding density function is

$$\exp\{-\mu(\mathcal{A})\} \times \prod_{j=1}^{n} \dot{\mu}(x_j). \tag{1.13}$$

Thus when points $(t_1, y_1), \ldots, (t_n, y_n)$ have been observed within $\mathcal{A} = (0, T) \times (u, \infty)$ and the measure is defined by (1.11), the density function,

$$\exp\{-T\Lambda(u)\} \prod_{j=1}^{n} \{-\dot{\Lambda}(y_j)\},$$

provides a likelihood for inference on the parameters $\eta$, $\tau$ and $\xi$ of the extremal model. Similar expressions underpin the likelihoods for most of the Poisson processes met here.

### 1.2.3 Proof of Lemma 1.6

Let $x_* = \inf\{x : F(x) > 0\}$ and $x^* = \sup\{x : F(x) < 1\}$ denote the (possibly infinite) lower and upper support points for a random variable $X$ whose cumulative distribution function $F$ has a positive density $f$ in $\mathcal{I} = (x_*, x^*)$, let $\mathcal{H}(x) = -\log\{1 - F(x)\}$ denote the *cumulative hazard function*, and write the *hazard function* as

$$\mathcal{H}'(x) = \frac{f(x)}{1 - F(x)} = \frac{1}{r(x)}, \quad x \in \mathcal{I},$$

where $r(x) > 0$ is the *reciprocal hazard function*. If $F$ placed positive probability on $x^*$ then the maxima of a random sample from $F$ would have a degenerate limiting distribution, so $\mathcal{H}(x) \to \infty$ as $x \to x^*$.

We prove Lemma 1.6 using a version of the *von Mises conditions*: we suppose that $r(x)$ has a continuous derivative $r'(x)$ in $\mathcal{I}$ and that $\lim_{x \to x^*} r'(x) = \xi$ for some real $\xi$. These assumptions are sufficient but not necessary, but they apply broadly and allow an elementary proof. References to weaker conditions and more general proofs can be found in the Bibliographic Notes.

For given $x$, set $b_t = F^{-1}(1 - 1/t) \in \mathcal{I}$ and let $a_t$ be a positive function of $t$ for which $b_t + a_t x \in \mathcal{I}$ for all $t \geq 1$. We aim to find $\Lambda(x) = \lim_{t \to \infty} \Lambda_t(x)$, where

$$\Lambda_t(x) = t\{1 - F(b_t + a_t x)\}, \quad t \geq 1.$$

If this limit exists, then $1 - F(b_t + a_t x) \to 0$, so $b_t + a_t x \to x^*$ for every $x$ for which the limit is defined.

Suppose first that $x > 0$ and write

$$\begin{aligned} -\log \Lambda_t(x) &= -\log\{1 - F(b_t + a_t x)\} - [-\log\{1 - F(b_t)\}] \\ &= \mathcal{H}(b_t + a_t x) - \mathcal{H}(b_t) \\ &= a_t \int_0^x \mathcal{H}'(b_t + a_t u)\, \mathrm{d}u \\ &= a_t \int_0^x \frac{\mathrm{d}u}{r(b_t + a_t u)}. \end{aligned} \tag{1.14}$$

Taylor's theorem implies that for each $u \in [0, x]$ there exists $s \equiv s(u) \in (0, u)$ such that

$$r(b_t + a_t u) = r(b_t) + a_t u r'\{b_t + a_t s(u)\}. \tag{1.15}$$

The implicit function theorem implies that $s(u)$ is continuous in $u$, and it follows that $r'\{b_t + a_t s(u)\}$ is uniformly continuous for $u \in [0, x]$. Hence

$$\frac{r(b_t + a_t u)}{a_t} = \frac{r(b_t) + a_t u r'(b_t + a_t s)}{a_t} = c_t^{-1} + u r'(b_t + a_t s), \quad 0 < u < x, \tag{1.16}$$

where $c_t = a_t / r(b_t)$ is a positive function of $t$. Consequently

$$a_t \int_0^x \frac{\mathrm{d}u}{r(b_t + a_t u)} = \int_0^x \frac{c_t \mathrm{d}u}{1 + c_t r'(b_t + a_t s)u} = \int_0^x g_t(u) \frac{c_t \mathrm{d}u}{1 + \xi_t c_t u},$$

where $\xi_t = r'(b_t)$ and

$$g_t(u) = \frac{1 + c_t \xi_t u}{1 + c_t r'(b_t + a_t s)u}, \quad u \in [0, x],$$

is continuous. Moreover $g_t(u) \to 1$ for each $u$, because the fact that $b_t + a_t s \geq b_t \to x^*$ implies that both $\xi_t$ and $r'(b_t + a_t s)$ have limit $\xi$ as $t \to \infty$.

Since $r(b_t + a_t u)$ and $c_t$ are positive and $g_t(u) \to 1$, the function $1 + \xi_t c_t u$ does not change sign for $u \in [0, x]$. Hence the mean value theorem for integrals implies that there exists some $u' \in [0, x]$ such that

$$\int_0^x g_t(u) \frac{c_t \mathrm{d}u}{1 + \xi_t c_t u} = g_t(u') \int_0^x \frac{c_t \mathrm{d}u}{1 + \xi_t c_t u} = g_t(u') \times \xi_t^{-1} \log(1 + \xi_t c_t x).$$

If we now let $t \to \infty$, $g_t(u') \to 1$ and $\xi_t \to \xi$, and, since the limiting expression cannot depend on $t$, $c_t$ must converge to a positive constant $c$. Thus, the assumption that a non-degenerate limit exists implies that $a_t \sim c r(b_t)$ as $t \to \infty$, and the form of the limit is unchanged by setting $a_t = r(b_t)$ for all $t$.

With minor changes a similar argument holds for $x < 0$, and we conclude that under the stated conditions on $F$ and if the limit exists, it is of the form

$$\lim_{t \to \infty} \Lambda_t(x) = (1 + \xi x)_+^{-1/\xi},$$

where the $(\cdot)_+$ ensures that the limit is defined when $1 + \xi x$ is negative.

To see that the limit is unique up to location and scale, let $x = (y - \eta)/\tau$, so that $y \in (\eta + \tau x_*, \eta + \tau x^*)$, Then $F(x)$ is replaced by $F\{(y - \eta)/\tau\}$ and $r(x)$ is replaced by $\tau r\{(y-\eta)/\tau\}$, so $r'(x)$ is replaced by $r'\{(y-\eta)/\tau\}$. Hence $\lim_{y\to\eta+\tau x^*} r'\{(y-\eta)/\tau\} = \lim_{x\to x^*} r'(x) = \xi$, and the limit $\Lambda(x)$ is replaced by the general form $\Lambda\{(y-\eta)/\tau\}$. Replacing $x$, $b_t$ and $a_t$ by $y-\eta$, $b_t - \eta a_t/\tau$ and $a_t/\tau$ throughout the argument leads to the same result.