Estadística climatológica con R

6. Análisis de valores extremos y periodos de retorno.

La teoría de valores extremos es una disciplina que desarrolla técnicas y modelos para describir los sucesos menos comunes. El objetivo es analizar valores extremos observados y predecirlos en el futuro (Gumbel 1958).

Los valores extremos tienen muchas aplicaciones en la práctica, como en estimaciones de ráfagas de viento, inundaciones (precipitaciones intensas), sequias, temperaturas extremas, etc.

6.1 Distribuciones usadas para el análisis de valores extremos.

6.1.1 Distribución de valores extremos generalizada

GEV en inglés (*Generalized Extreme Value distribution*)

Supongamos n variables aleatorias $X_1, X_2, \ldots, X_n$. Comúnmente, las variables aleatorias, que forman una muestra, suelen considerarse independientes e idénticamente distribuidas. Sea $M_n=Max\{ X_1, X_2, \ldots, X_n \}$ el máximo de dichas variables. Muchas veces las X_i, que tienen una función de distribución común F, corresponden a valores de un proceso medido en una escala regular de tiempo como medidas de temperatura cada hora o medias diarias, de forma que M_n representa el máximo del proceso sobre n unidades temporales de observación. Si por ejemplo, n es el número de observaciones en un mes, entonces M_n corresponde al máximo mensual.

Se llega a una función de distribución G para M_n que pertenece a alguna de las siguientes familias:

$$\text{I: } G(z) = exp\left\{-exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \ -\infty < z < \infty;$$

$$\text{II: } G(z) = \begin{cases} 0, & z \le b, \\ exp\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right], & z > b; \end{cases}$$

$$\text{III: } G(z) = \begin{cases} exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right]\right\}, & z \le b, \\ 1, & z > b; \end{cases}$$

para parámetros a>0, b y, en el caso de las familias II y III, α >0.

Estas tres clases de distribuciones son conocidas como *las distribuciones de valores extremos*, donde las de tipo **I** son la familia de ***Gumbel***; las de tipo **II** la de ***Fréchet***, y las de tipo **III** la de ***Weibull***,

cada una con su *parámetro de localización* **b** y *de escala*, **a**; y además, las familias de Fréchet y de Weibull tienen un parámetro *de forma* **α**. Más adelante veremos las características de esas distribuciones, y se profundizará más en ellas.

Todas las distribuciones anteriores se pueden condensar en una sola, que es la *Distribución de Valores Extremos generalizada* (**GEV**), cuya función de distribución es la siguiente:

$$G(z) = exp\left[-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right], \text{ para } \begin{cases} x \in \left[\mu - \frac{\sigma}{\xi}, +\infty\right[& si\ \xi > 0 \\ x \in]-\infty, +\infty[& si\ \xi = 0. \\ x \in \left]-\infty, \mu - \frac{\sigma}{\xi}\right[& si\ \xi < 0 \end{cases}$$

El modelo tiene *tres parámetros*: **μ** el parámetro de *localización*; **σ** el parámetro de *escala*; y **ξ** que es el parámetro de *forma*.

Para el caso del mínimo se llega a una función similar.

La función de distribución de valores extremos generalizada se designa por ***GEV (μ,σ,ξ)***

El caso de **ξ** *(parámetro de forma) es 0*, se corresponde con la ***distribución de Gumbel GEV(μ,σ,0).***

En R para calcular las distribuciones de valores extremos generalizadas se puede usar el paquete ***evd*** o el paquete ***extRemes***.

Con el paquete ***evd*** las características de la función de distribución de valores extremos generalizada vienen dadas como en otras funciones que vimos por:

```
> pgev(x,loc,esc,for)    # Valor de la función de distribución en x de la gev (loc,esc,for)
> dgev(x,loc,esc,for)    # Valor de la función de densidad en x de la gev (loc,esc,for)
> qgev(q,loc,esc,for)    # Cuantil de orden q de la gev (loc,esc,for)
> rgev(m,loc,esc,for)    # muestra aleatoria de tamaño m de la gev (loc,esc,for)
```

Donde ***loc*** es el factor de *localización*, ***esc*** el de *escala* y ***for*** el de *forma*.

El paquete ***evd*** tiene para la distribución de ***gumbel*** las mismas funciones cambiando simplemente *gev* por *gumbel* y quitando el parámetro de forma:
pgumbel(x,loc,esc), dgumbel(x,loc,esc), qgumbel(q,loc,esc) y rgumbel(m,loc,esc)

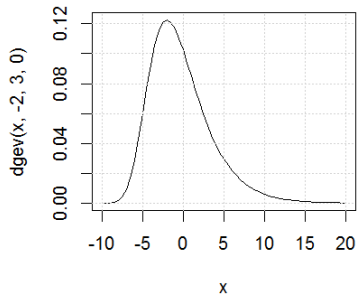
Ejemplo 6.1

Representaremos la función ***GEV(-2,3,0)*** (*Gumbel*, ya que el parámetro de forma es cero), primero su función de densidad y luego su función de distribución.

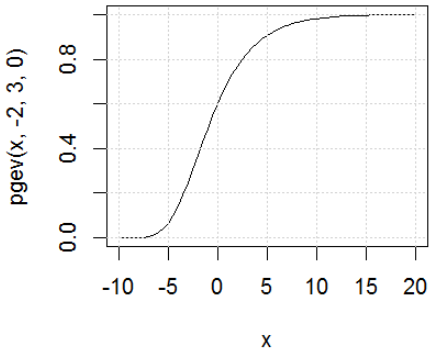
```
> library(evd)   # cargamos el paquete evd
> x <- seq(-10,20,0.5)   # Vector del eje x
> plot (x, dgev(x,-2,3,0),type='l')  # pintamos la función de densidad
```

También se podría dibujar con la función ***curve***, sin escribir el vector x

```
> curve( dgev (x,-2,3,0),-10,20)
> grid()  # dibuja las cuadriculas del fondo
```



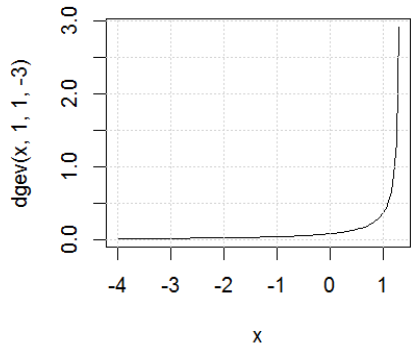
> ***plot (x, pgev(x,-2,3,0),type='l') # pintamos la función de distribución***
o mediante ***curve***:
> ***curve (pgev(x,-2,3,0),-10,20)***
> ***grid()***



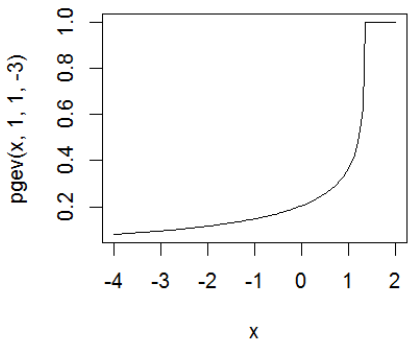
Ejemplo 6.2

Ahora una función general de valores extremos generalizada (con el parámetro de forma distinto de 0), por ejemplo la ***GEV(1,1,-3)***

> ***x <- seq(-4,1.3,0.05)***
> ***plot (x, dgev(x,1,1,-3),type='l')***
> ***grid()***



> ***x <- seq(-4,2,0.05)***
> ***plot (x, pgev(x,1,1,-3),type='l')***
> ***grid()***



Ejercicio 6.1

Representar la función de densidad y la función de distribución de la *función de Valores Extremos generalizada* ***GEV(1,1,3)*** y de ***GEV(1,3,0)*** (*Gumbel*).

Nota: la primera representarla en el intervalo de 0.5 a 10 y la segunda de -6 a 22.

6.2.1 Gumbel y Weibull

Veremos algunos casos especiales de la función general de valores extremos.

Distribución de Gumbel

En muestras de tamaño grande, donde aparecen sucesos que se presentan raramente, se utiliza con bastante éxito ***la distribución de Gumbel***

La función de distribución de Gumbel es:

$$F(x) = e^{-e^{\frac{(\mu - x)}{\beta}}}$$

La distribución de *Gumbel standard* corresponde al caso en el que μ=0 y β=1

En ese caso, la función de distribución viene dada por

$$F(x) = e^{-e^{-x}}$$

La distribución de Gumbel se ajusta muy bien para *la intensidad máxima de precipitaciones*, *precipitación máxima diaria* y *racha máxima de viento*.

La distribución de valores extremos generalizada se ajusta bien a *la temperatura máxima y mínima absoluta*, *las temperaturas mínimas de las máximas* o *máxima de las mínimas*.

Distribución de Weibull

La función de densidad de probabilidad de una variable con esta distribución es la siguiente:

$$f(x) = \frac{k}{\lambda}\left(\frac{x-\theta}{\lambda}\right)^{k-1} e^{-\left(\frac{x-\theta}{\lambda}\right)^{k}}, \text{ para } x \geq \theta.$$

donde $\boldsymbol{k}$ es el parámetro de *forma*, $\boldsymbol{\lambda} > 0$ es el parámetro de *escala* y $\boldsymbol{\theta}$ es el parámetro de localización de la distribución.

La distribución de ***Weibull*** se ajusta bien para *la velocidad del viento* (las componentes del viento por separado se ajustan a la distribución normal)

Dentro del paquete ***evd*** la distribución de ***Weibull*** viene como una ***gev*** cambiando este nombre por ***rweibull***. Ej. ***prweibull(x,loc,esc,for)*** para la función de distribución.

Niveles y periodos de retorno. Estudio gráfico de la bondad del ajuste.

Los niveles y *los periodos de retorno* tienen mucha importancia en hidrología o climatología.

Queremos responder a la pregunta de cuál es la probabilidad de que un fenómeno poco frecuente se presente una o más veces (al menos una vez) en *n* ocasiones y cuál es el tiempo medio de ocurrencia de este fenómeno.

El tiempo, intervalo de recurrencia o periodo de retorno, es el intervalo de tiempo entre dos ocurrencias de un fenómeno extremo; es decir, si un valor de un elemento meteorológico es igualado o excedido una vez cada *n* años, el tiempo de recurrencia es *n* años.

Los niveles de retorno son considerados como los cuantiles de la distribución de valores extremos. Así, si z_p (el nivel de retorno) es el cuantil de orden *p* de una variable con distribución ***GEV***, entonces *p* es la probabilidad de que z_p sea superado una vez al año; y *el período de retorno*, *1/p* , es el número de unidades de tiempo que transcurrirán en media entre dos veces en los que la variable supere el valor de z_p.

Por ejemplo, si se tiene un período de retorno de 50 años, eso equivale a una probabilidad anual de 0.02; y si el periodo de retorno es de 10 años, la probabilidad correspondiente sería de 0.10.

Si suponemos que una variable aleatoria puede ser representada por un modelo probabilístico podemos, así, estimar los periodos y niveles de retorno para estos fenómenos poco frecuentes.

Lo primero es estimar los parámetros del modelo probabilístico a partir de la muestra de datos.

Estimación de los parámetros de la distribución

Si tenemos los datos de la muestra y suponemos que se adaptan a un modelo probabilístico determinado, tenemos que obtener los parámetros de este modelo. Al ser los parámetros desconocidos, hay que estimarlos; existen muchas técnicas distintas para hacer dichas estimaciones, pero la preferible es la estimación por *máxima verosimilitud*.

Se llega a una ecuación para la que generalmente no es posible lograr una solución analítica, pero si se tienen unos datos muestrales, se pueden resolver usando algoritmos de optimización numéricos. A continuación mostraremos el método para lograr estimaciones de los parámetros por máxima verosimilitud mediante el paquete R dado un conjunto de datos concretos.

Ejemplo 6.3

Dadas las precipitaciones máximas diarias desde el año 1938 al 2022 (falta el 2006 que estuvo incompleto) de la estación 8416, Valencia, (***Valencia.Precip.Maxima.Diaria.txt***) vamos a ajustar la

distribución primero a una Distribución de Valores Extremos generalizada (***GEV***) y luego a una Gumbel y calcularemos los periodos de retorno para un número de años determinado.

> datos <- read.table('Valencia.Precip.Maxima.Diaria.txt') # lee el fichero de datos
> prec <- datos$V2 # toma solo la segunda columna, que tiene los datos de precipitacion maxima diaria

> prec

```
> prec
 [1] 169.0  27.0  57.0  96.9  35.4  36.1  70.8  25.6  80.4  53.5 102.1
[12]  47.8 152.3  77.6  57.5  43.7  39.0  51.6 262.6 125.1 123.0  40.2
[23]  36.9  65.8  92.7 111.1  44.5 147.5 148.4  30.2  59.0 108.2  40.5
[34] 140.3  57.4 101.4  70.5  81.0  36.8  71.7  32.7  49.0 101.5  51.2
[45] 105.6  76.7  62.8  34.0  98.1  51.8  56.5 134.4  38.5  78.2  65.8
[56]  59.5  55.5  32.7  41.7  19.1  61.6  24.6 165.0 127.4 128.7  52.2
[67]  93.5  37.9 178.2 112.8 119.4  50.8  47.4  49.6  72.2  43.4  47.6
[78] 130.0  76.6 136.0  33.8  92.8  47.9 172.4
```

> library(evd) # carga la librería evd de funciones de extremos

La función ***fgev*** ajusta los parámetros por máxima verosimilitud a una *distribución de valores extremos generalizada*, ***GEV(μ,σ,ξ),*** definida por ***μ*** el parámetro de localización; ***σ*** el parámetro de escala; y ***ξ*** que es el parámetro de forma.

> ajuste <- fgev(prec)

Si se quisiera ajustar a *una distribución de Gumbel* se hace de la misma forma con ***fgev***, pero forzando el parámetro de forma a 0, con ***shape***=0)

> ajuste <- fgev(prec,shape=0)

```
> ajuste

Call: fgev(x = prec)
Deviance: 845.6055

Estimates
    loc    scale    shape
54.3400  27.1624   0.2675

Standard Errors
   loc   scale   shape
3.5124  2.9405  0.1157

Optimization Information
  Convergence: successful
  Function Evaluations: 27
  Gradient Evaluations: 15
```

La pantalla que aparece nos da información sobre los parámetros estimados de *localización (loc), escala (scale)* y *forma (shape).* Los errores estándar y los resultados de la optimización por métodos numéricos.

El objeto ***ajuste*** (o el nombre que hayamos puesto) es una clase con distintas propiedades que nos dan estas informaciones:
> ajuste$estimate los parámetros estimados: *loc, scale* y *shape* y otros muchos más valores.

Se suelen representar unos *gráficos de niveles de retorno*, en los cuales se representan los niveles de retorno estimados y sus periodos de retorno asociados en una escala logarítmica. A través de estos gráficos se pueden obtener estimaciones aproximadas de niveles y periodos de retorno, aunque es mejor hacerlo numéricamente.

Otros gráficos de diagnóstico usados para comprobar cualitativamente la bondad del ajuste son los *gráficos P-P*, *Q-Q* y los *gráficos de densidad de probabilidad*.

Si ploteamos ***ajuste*** nos va a dar cuatro pantallas (hay que pinchar con el ratón dentro del grafico para ir pasando) que nos van dar una interpretación grafica de la bondad de ajuste.

> plot (ajuste)

Grafico ***P-P plot*** (*ploteo de probabilidad*) donde viene los puntos empíricos (cruces) que se tiene que aproximar a la recta de 45 %, contra más se aproximen mejor se ajustan los datos a la distribución de valores extremos. Vienen unas bandas límites entre las que tienen que estar los puntos para que el ajuste sea bueno.

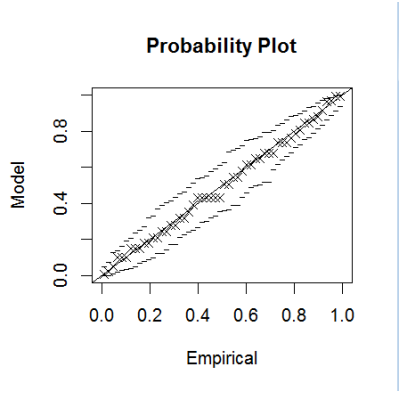


Grafico ***Q-Q plot*** (Cuantiles). De igual manera que el grafico anterior el ajuste será mejor cuanto más se aproximen los puntos (cruces) a la recta y estén entre las dos bandas límites.

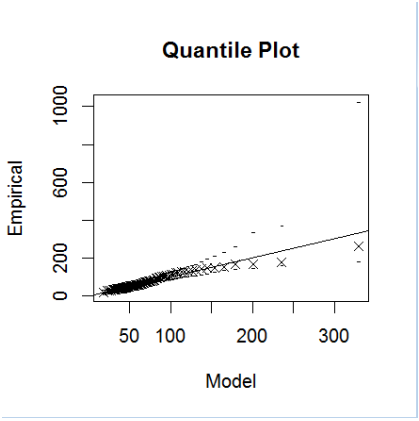


Gráfico de la función de densidad empírica y la función de valores extremos generalizada estimada. Es una comparación a simple vista del parecido.

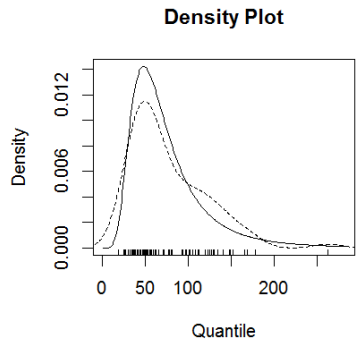
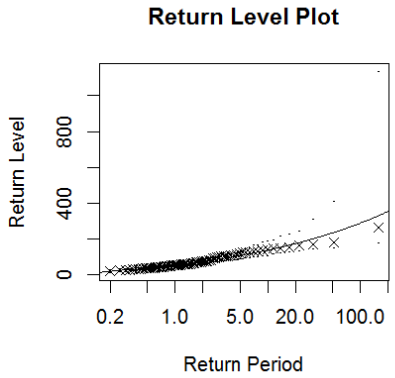


Gráfico de niveles y periodos de retorno. Se podrían obtener gráficamente los periodos de retorno, pero es mejora calcularlos numéricamente.



Se pueden mostrar los cuatro gráficos a la vez haciendo antes del ***plot***:
> par(mfrow = c(2,2))

Se vuelve otra a poner por defecto un solo gráfico con:
> par(mfrow = c(1,1))
Otra forma de comprobar la bondad del ajuste es a través del test de ***Kolmogorov-Smirnov***, que en R esta implementado con la función ***ks.test***.

El test de ***Kolmogorov-Smirnov*** sirve, en general, para ver si una distribución empírica se ajusta a una distribución teórica. Tenemos que suponer la distribución y sus parámetros, para ver si se adapta a ella o no.

El test es un *contraste de hipótesis*, suponemos *la hipótesis nula* $\boldsymbol{H_0}$: los datos proceden de una distribución normal, por ejemplo, y *la hipótesis contraria* $\boldsymbol{H_1}$: los datos no proceden de una distribución normal.

Después de aplicar el test, nos aparecen en pantalla dos números: *D* que hace referencia a la mayor diferencia absoluta observada entre la frecuencia acumulada observada y la frecuencia acumulada teórica. Si los valores observados son similares a los esperados *D* será pequeño. Por tanto el criterio para la toma de decisión entre las dos hipótesis será de la forma:

$$\text{Si } D \leq D_\alpha \Rightarrow \text{Aceptar } H_0$$
$$\text{Si } D > D_\alpha \Rightarrow \text{Rechazar } H_0$$

Donde el valor D_α se elige en función del nivel del *nivel de significación del contraste* $\boldsymbol{\alpha}$.

Otro método alternativo, que es el más usado, para la toma de decisión es el empleo de ***p-valor*** asociado con el estadístico *D* observado.

Si el ***p-valor*** es grande significa que, siendo cierta la hipótesis nula, el valor observado del estadístico *D* era esperable. Por tanto no hay razón para rechazar dicha hipótesis. Asimismo, si *el p-valor fuera pequeño*, ello indicaría que, siendo cierta la hipótesis nula, era muy difícil que se produjera el valor de *D* que efectivamente se ha observado; ello obliga a poner muy en duda, y por tanto *a rechazar, la hipótesis nula*. De esta forma, para un nivel de significación ***α***, *la regla de decisión para este contraste* es:

Si p-valor $\geq \alpha \Rightarrow$ Aceptar H_0
Si p-valor $< \alpha \Rightarrow$ Rechazar H_0

El ***ks.test*** y otros muchos test muestran el ***p-valor*** directamente, por lo que solo hay que ver si este valor es mayor o menor que ***α***, siendo ***α*** *el nivel de significación del contraste*.

Se suele tomar ***α*** como **0.05** (5 % de error). También a veces se toman α=0.01 (1 % de error) y α= 0.1 (10 % de error)

Ejemplo 6.4

Generamos *una distribución normal*

```
> x1 <- rnorm(50,1,2)  # generamos una muestra de 50 datos de distribución normal (media=1, sd= 2)
> ks.test(x1,"pnorm",1,2)   # test Kolmogorov-Smirnov (hay que poner la distribución supuesta y los parámetros de esta)
```

```
data:  x1
D = 0.0895, p-value = 0.7847
alternative hypothesis: two-sided
```

El valor *D* es pequeño, y el ***p-value*** es mayor que 0.05, luego al nivel de significación 0.05 *podemos aceptar la hipótesis de que x1 sigue una distribución normal*.

Existen otros muchos tipos de test en *R*, como el de ***Shapiro-Wilk*** para ver si una distribución se adapta bien a la normal:

```
> shapiro.test(x1)
```

```
        Shapiro-Wilk normality test

data:  x1
W = 0.9798, p-value = 0.5436
```

El *p-valor* también es mayor que *0.05* por lo que *aceptamos la hipótesis de que es normal*.

En el caso de una distribución de valores extremos generalizada o Gumbel, para ejecutar el ***ks.test*** hay que poner como argumentos los datos, el tipo de función estimada ***gev*** o ***gumbel*** con ***p*** delante (‘***pgev***’, ‘***pgumbel***’) y luego los parámetros estimados para cada distribución. Para Gumbel serán 2 solamente. Se pueden sacar de *ajuste$estimate*.

Si aplicamos el test a los resultados del ejemplo 6.5

```
> ks.test(prec,'pgev',ajuste$estimate[1],ajuste$estimate[2],ajuste$estimate[3])
```

```
        One-sample Kolmogorov-Smirnov test

data:  prec
D = 0.0772, p-value = 0.7343
alternative hypothesis: two-sided
```

El *p-valor* también es mucho mayor que *0.05* por lo que *aceptamos la hipótesis de que se adapta bien a una distribución de valores extremos generalizada*.

Para Gumbel (*shape=0*, solo tiene dos parametros):
> ks.test(prec,'pgumbel',ajuste$estimate[1],ajuste$estimate[2])

Ejercicio 6.2

Dadas las precipitaciones máximas diarias de la estación *8025, Alicante - Ciudad Jardín*. (***Alicante.Precip.Maxima.Diaria.txt***) vamos a ajustar estos datos a una distribución de Gumbel con la librería R ***evd***. Luego ver tanto gráficamente como numéricamente la bondad del ajuste y comentarla.

(Solución: parámetros de ajuste: loc=42.70 scale=22.42, p-valor=0.4323 con el ks.test)

6.2 Calculo de niveles y periodos de retorno

Como vimos arriba *los niveles de retorno* son considerados como los cuantiles de la distribución de valores extremos. Así, si z_p (el nivel de retorno) es el cuantil de orden p de una variable con distribución GEV, entonces p es la probabilidad de que z_p sea superado una vez al año; y *el período de retorno*, $1/p$, es el número de unidades de tiempo que transcurrirán en media entre dos veces en los que la variable supere el valor de z_p.

Vamos a calcular niveles y periodos de retorno por medio de *R* y el paquete ***evd***.

Ejemplo 6.5

Supongamos que una estación tiene los siguientes registros para *la precipitación máxima diaria* en toda su historia (mm):

> prec <- c(83.0, 63.0, 40.0, 35.8, 50.4, 54.4, 45.7, 68.4, 35.0, 31.3,105.9, 40.6, 31.6, 41.0, 88.5, 70.8, 52.5, 82.5, 59.2, 91.3, 44.7, 45.9, 42.3, 26.6, 44.0, 34.8, 70.9, 68.0,136.9, 45.1, 220.2, 53.1, 24.4, 75.0, 67.5, 68.3, 46.0,119.8, 25.2, 32.8, 40.4, 40.5, 22.7, 15.3, 41.5, 270.2, 59.3, 33.5, 55.4, 35.7, 28.6, 22.8, 34.6)

Para calcular ***los niveles de retorno*** definimos primero un vector con los valores de años (periodos de retorno) para los que queremos calcular estos niveles, por ejemplo para 2, 5, 10, 15, 25, 50, 100, 200 y 500 años:

> años <- c(2,5,10,15,25,50,100,200,500) # años para los que vamos a calcular los niveles de retorno

> prob <- 1/años # vector de probabilidades para esos años

> ajuste <- fgev(prec,shape=0) # ajustamos a una distribucion de Gumbel (shape=0)

> z <- ajuste$estimate # parámetros del ajuste (loc, scale y shape)

> round(qgumbel (prob, z[1], z[2], lower.tail=FALSE), 1) # calculamos los percentiles, el round (, 1) simplemente para dejar un decimal

> round(qgumbel (1-prob, z[1], z[2]), 1) # Otra forma

```
> round(qgumbel (1-prob,z[1],z[2]),1)
[1]  52.4  78.3  95.4 105.1 117.1 133.2 149.1 165.0 186.0
```

Obtenemos así la precipitaciones máximas diarias esperadas para cada uno de los periodos de retorno definidos antes: 2, 5, 10, 15, 25, 50, 100, 200 y 500 años.

Nota: en el ejemplo anterior hemos cogido una distribución de Gumbel. De igual forma se podría ajustar a otra distribución como la de valores extremos generalizada, añadiendo los parámetros de esta distribución.

Si lo que queremos ahora es calcular ***el periodo de retorno*** para una precipitación máxima concreta, procederíamos así:

> periodo.retorno <- 1 / (1 - pgumbel (preci.maxima, z[1], z[2]))

Por ejemplo, si quisiésemos obtener *el periodo de retorno* para una precipitación máxima de 100 mm en el ejemplo anterior:

> periodo.retorno <- 1 / (1 - pgumbel (100, z[1], z[2]))

```
> periodo.retorno
     loc
12.09115
```

O sea que, podríamos esperar una precipitación máxima diaria de 100 mm, por término medio, cada 12 años.

Ejercicio 6.3

Calcular los niveles de retorno para 5, 10, 25, 50, 100 y 500 años para la serie de precipitaciones máximas diarias de la estación del ejercicio 6.2 *(8025, Alicante –Ciudad Jardín).* Calcular también el periodo de retorno de una precipitación de 150 mm.

(Solución: Niveles de retorno: 76.3 93.1 114.4 130.2 145.8 182.0 El periodo de retorno es: 120 años)