Entrega: curso de datos extremales

Laura Montaldo, CI: 3.512.962-7

2024-03-19

Índice

Resumen

Your abstract goes here.

Motivación y objetivo del estudio

Siguiendo a Perera, Segura, y Crisci (2021), se dice que tenemos datos extremos cuando cada dato corresponde al máximo o mínimo de varios registros. Son un caso particular de evento raro o gran desviación respecto a la media. Es por este motivo que en una gran variedad de dominios disciplinares suele ser de gran interés el trabajo con datos extremos. Además, admiten diversos enfoques. La teoría 'más' clásica de estadística de datos extremos se basa en los trabajos de Fréchet, Gumbel, Weibull, Fisher, Tippett, Gnedenko, entre otros. En este estudio, el foco va a estar puesto en esquemas que extienden a las distribuciones extremales clásicas.

Los índices de $S\&P$ son una familia de índices de renta variable[1] diseñados para medir el rendimiento del mercado de acciones en Estados Unidos que cotizan en bolsas estadounidenses. Ésta familia de índices está compuesta por una amplia variedad de índices basados en tamaño, sector y estilo. Los índices están ponderados por el criterio *float-adjusted market capitalization* (FMC). Además, se disponen de índices ponderados de manera equitativa y con límite de capitalización de mercado, como es el caso del $S\&P$ 500. Este este sentido, el $S\&P500$ entraría en el conjunto de índices ponderados por capitalización bursátil ajustada a la flotación (ver $S\&P$ Dow Jones Indices). El mismo mide el rendimiento del segmento de gran capitalización del mercado estadounidense. Es considerado como un indicador representativo del mercado de renta variable de los Estados Unidos, y está compuesto por 500 empresas constituyentes.

Se busca crear un indicador de una posible crisis bursátil. Como variable de referencia de toma la relación de precios al cierre de ayer sobre la de hoy

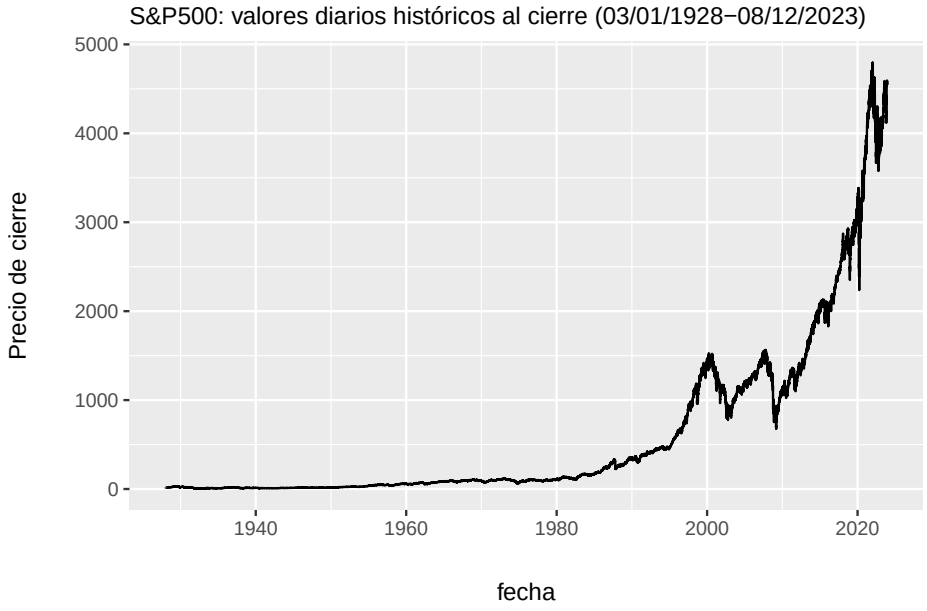
$$Indicador_t = \frac{Precio_{t-1}}{Precio_t}, \quad \text{para } t = 1, ..., T \tag{1}$$

Interpretación del Indicador:

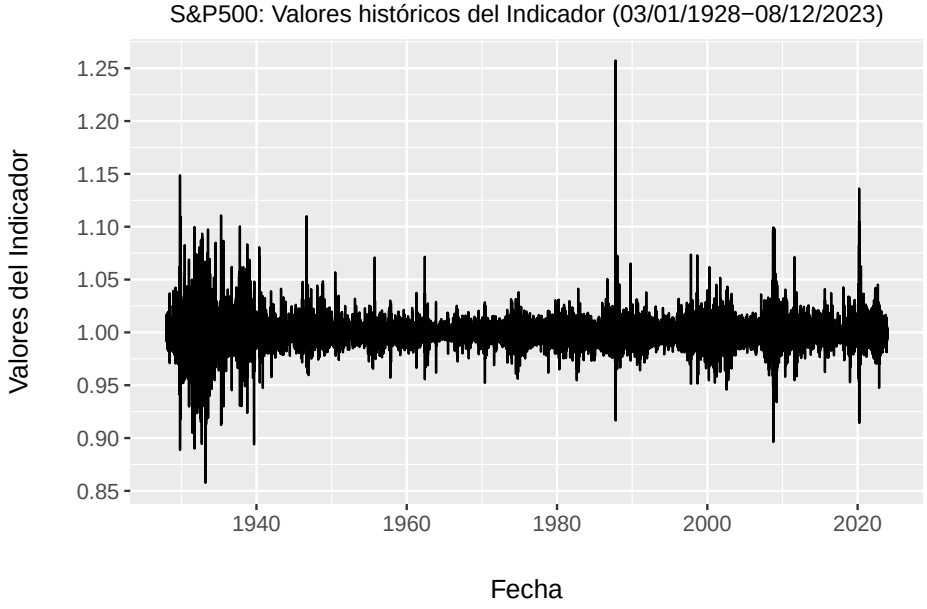
- Si el $Indicador_t \leq 1$, el precio de cierre de hoy es mayor o igual que el de ayer, lo cual podría ser considerado una señal positiva.
- Si el $Indicador_t > 1$, el precio de cierre de hoy es menor que el de ayer, lo cual podría considerarse una señal de alerta.

[1] En inglés se llaman equity indices

En las siguiente figuras se muestra la evolución histórica desde la fecha 03/01/1928 hasta 08/12/2023 del precio al cierre del día del indicar S&P 500.




```
## [1] 24100     9
```



```r
ts_relacion=df[,c('relacion')]
```

```r
result_adf <- suppressWarnings(adf.test(ts_relacion))
cat('p-valor adf:', result_adf$p.value ,'\n')
```

```
## p-valor adf: 0.01
```

```r
result_kpss <- suppressWarnings(kpss.test(ts_relacion))
cat('p-valor kpss:', result_kpss$p.value ,'\n')
```

```
## p-valor kpss: 0.1
```

```r
#install.packages('aTSA')
aTSA::adf.test(ts_relacion)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##       lag    ADF p.value
##  [1,]   0 -1.326   0.205
##  [2,]   1 -0.770   0.404
##  [3,]   2 -0.543   0.486
##  [4,]   3 -0.414   0.525
##  [5,]   4 -0.340   0.547
##  [6,]   5 -0.298   0.559
##  [7,]   6 -0.255   0.571
##  [8,]   7 -0.221   0.581
##  [9,]   8 -0.190   0.590
## [10,]   9 -0.180   0.593
## [11,]  10 -0.164   0.597
## [12,]  11 -0.152   0.601
## [13,]  12 -0.141   0.604
## [14,]  13 -0.128   0.608
## Type 2: with drift no trend
##       lag    ADF p.value
##  [1,]   0 -156.9    0.01
##  [2,]   1 -112.1    0.01
##  [3,]   2  -90.6    0.01
##  [4,]   3  -77.8    0.01
##  [5,]   4  -68.9    0.01
##  [6,]   5  -64.6    0.01
##  [7,]   6  -58.9    0.01
```

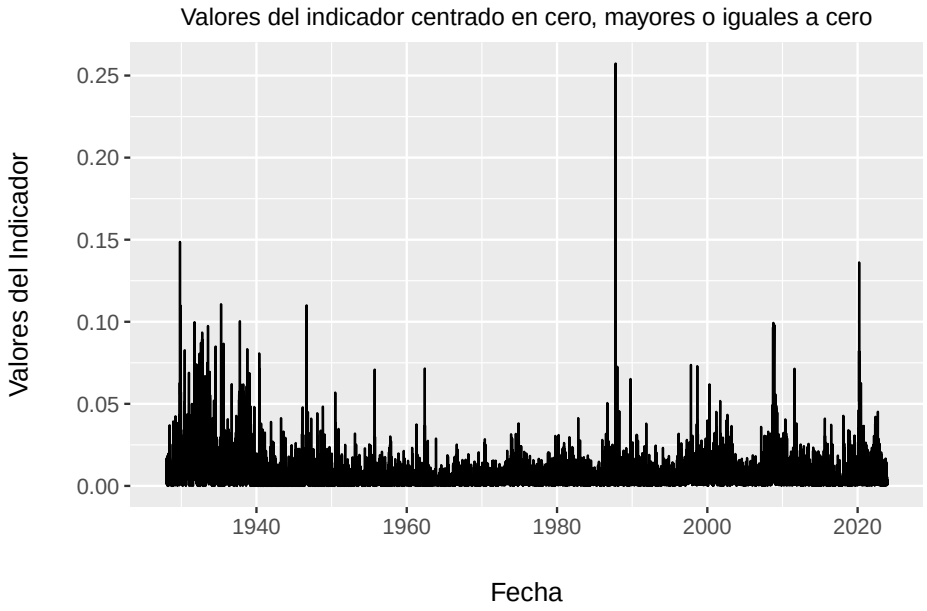
```
##  [8,]   7  -54.7    0.01
##  [9,]   8  -50.2    0.01
## [10,]   9  -47.2    0.01
## [11,]  10  -44.8    0.01
## [12,]  11  -42.6    0.01
## [13,]  12  -41.9    0.01
## [14,]  13  -40.1    0.01
## Type 3: with drift and trend
##       lag    ADF p.value
##  [1,]   0 -156.9    0.01
##  [2,]   1 -112.1    0.01
##  [3,]   2  -90.6    0.01
##  [4,]   3  -77.8    0.01
##  [5,]   4  -68.9    0.01
##  [6,]   5  -64.7    0.01
##  [7,]   6  -59.0    0.01
##  [8,]   7  -54.7    0.01
##  [9,]   8  -50.2    0.01
## [10,]   9  -47.2    0.01
## [11,]  10  -44.9    0.01
## [12,]  11  -42.6    0.01
## [13,]  12  -41.9    0.01
## [14,]  13  -40.1    0.01
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

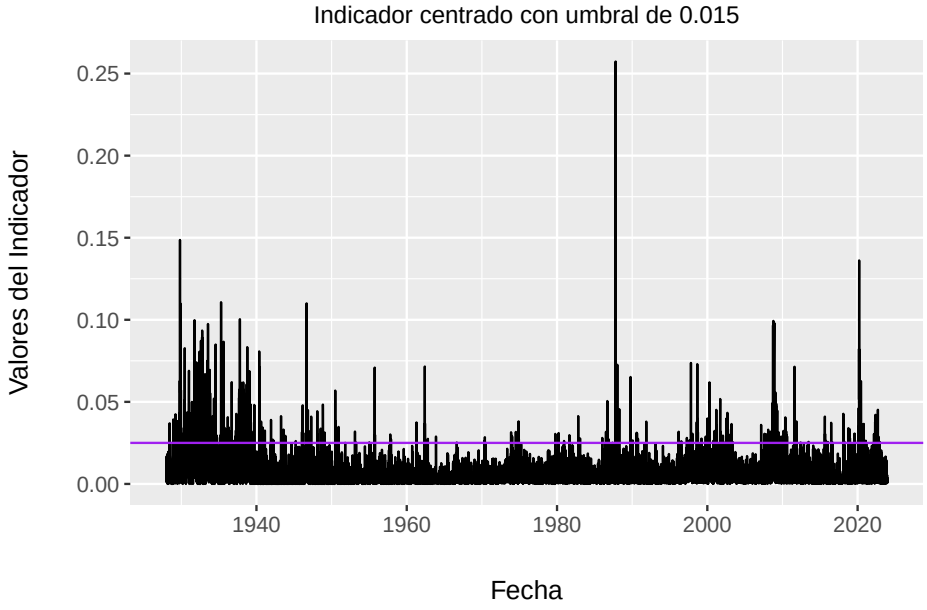
Marco teorico

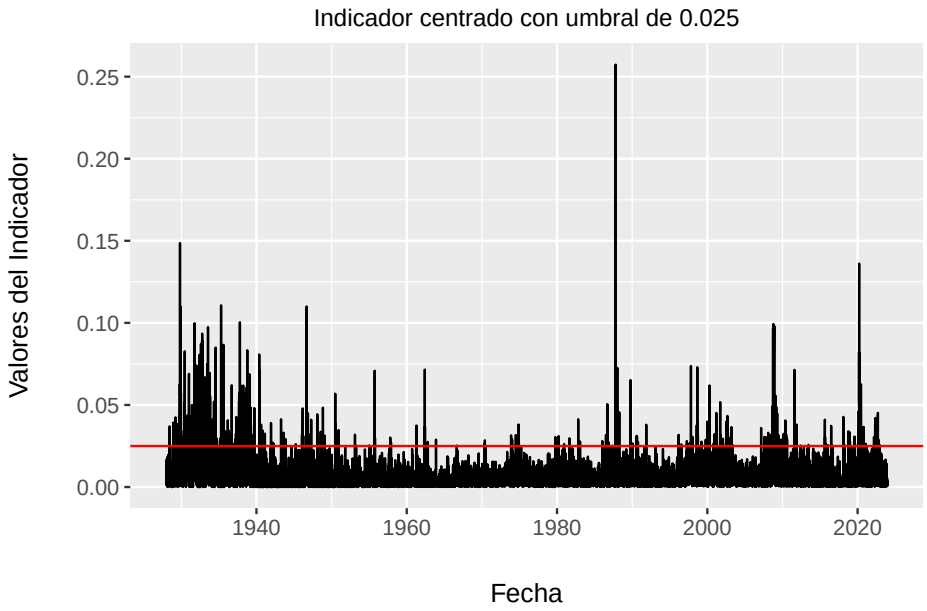
POT (Peaks Over Treshold) y variantes

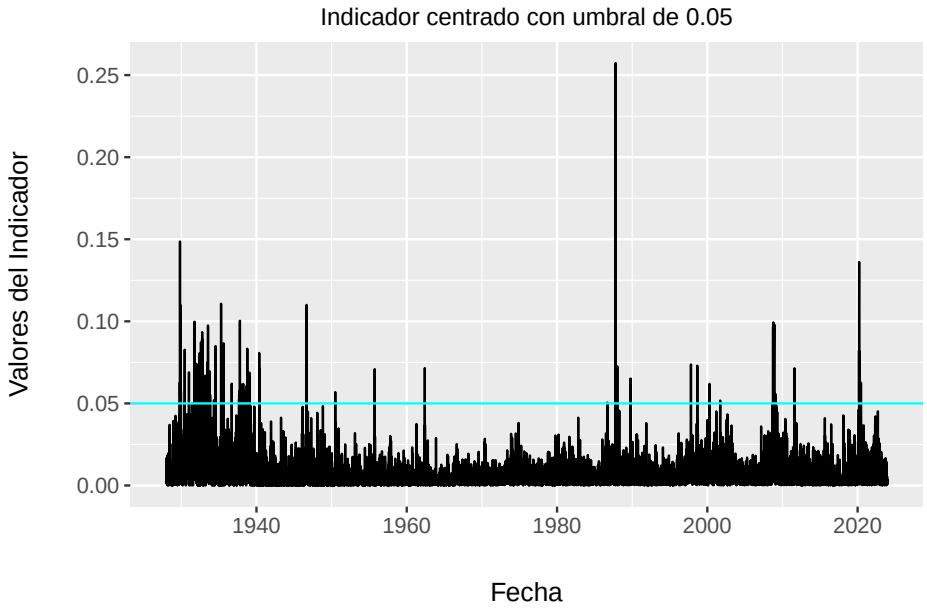
Estrategia Empírica

A la columna relativa a la relacion de precios se la resta por 1 para tener centrados los valores de la relacion de precios en cero. Y posteriormente analizar si las series, fijando distintos umbrales son estacionarias.










```
filtered_df_0_025 <- df %>%
  filter(rel_cero>= 0.025)
```

```
head(filtered_df_0_025)
```

```
##         Date  Open  High   Low Close Volume Dividends Stock.Splits relacion
```

```
## 1 1928-06-11 18.68 18.68 18.68 18.68      0         0            0 1.036938
## 2 1928-07-11 18.95 18.95 18.95 18.95      0         0            0 1.025330
## 3 1928-12-06 22.91 22.91 22.91 22.91      0         0            0 1.039284
## 4 1929-02-07 24.71 24.71 24.71 24.71      0         0            0 1.031566
## 5 1929-03-25 24.51 24.51 24.51 24.51      0         0            0 1.042432
## 6 1929-04-01 24.88 24.88 24.88 24.88      0         0            0 1.026125
##     rel_cero
## 1 0.03693793
## 2 0.02532979
## 3 0.03928414
## 4 0.03156620
## 5 0.04243162
## 6 0.02612546
```

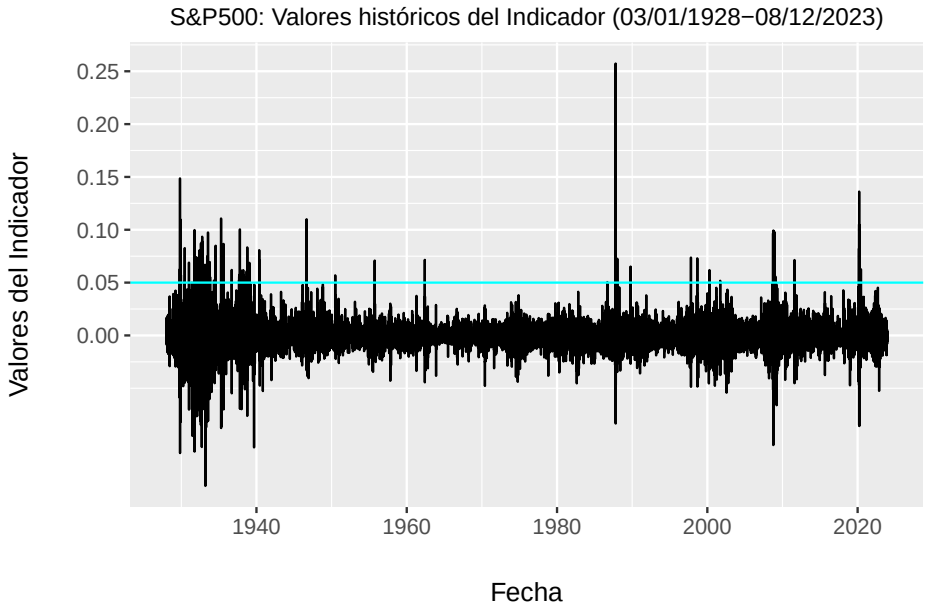
```
data=filtered_df_0_025[,c('Date', 'rel_cero')]
n=dim(data)[1]
```

```
fecha_maxima <- max(data$Date)
# Reescalar el tiempo dividiendo cada fecha por la fecha máxima
data$tiempo_reescalado <- as.numeric(data$Date - min(data$Date)) / as.numeric(fecha_maxima -
head(data)
```

```
##         Date   rel_cero tiempo_reescalado
## 1 1928-06-11 0.03693793      0.0000000000
## 2 1928-07-11 0.02532979      0.0008690614
## 3 1928-12-06 0.03928414      0.0051564311
## 4 1929-02-07 0.03156620      0.0069814600
## 5 1929-03-25 0.04243162      0.0083140209
## 6 1929-04-01 0.02612546      0.0085168019
```

```
ggplot(df, aes(x = Date, y = rel_cero)) +
  geom_line() +
  geom_hline(yintercept = 0.05, linetype = "solid", color = "cyan") +  # Add horizontal line
  ggtitle("S&P500: Valores históricos del Indicador (03/01/1928-08/12/2023)") +
  xlab("Fecha") +
  ylab("Valores del Indicador") +
  scale_x_date(limits = date_range) +
  scale_y_continuous(breaks = seq(0, ceiling(max(df$relacion)), by = 0.05)) +
  theme(
    axis.title.x = element_text(margin = margin(t = 20, b = 40)),
    axis.title.y = element_text(margin = margin(r = 20, l = 40)),
    plot.title = element_text(size = 11, hjust = 0.5),  # Center the plot title
    axis.text = element_text(size = 10),  # Adjust the size of axis text
    axis.title = element_text(size = 12),  # Adjust the size of axis titles
    legend.title = element_text(size = 10),  # Adjust the size of legend title
```

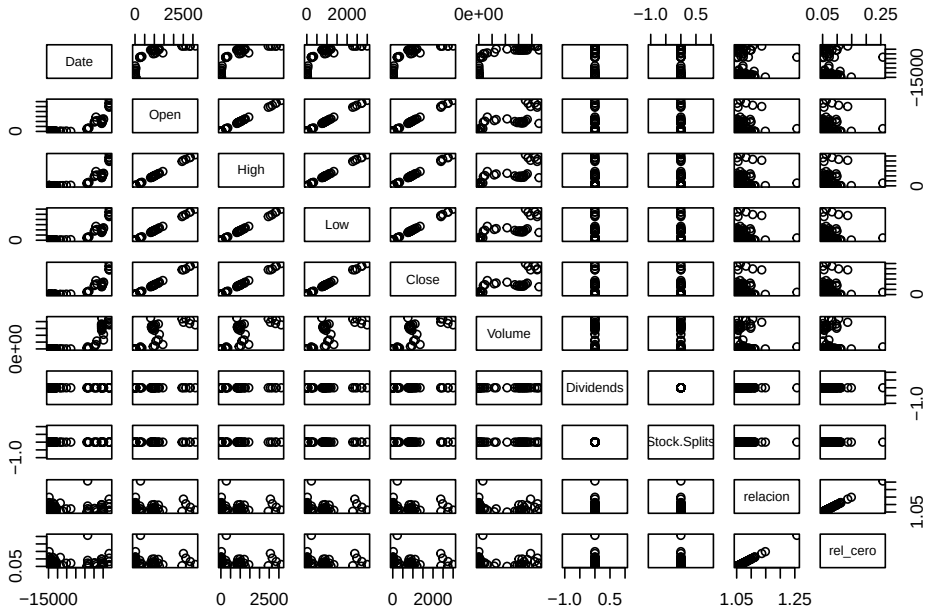
```
    legend.text = element_text(size = 8)  # Adjust the size of legend text
  )
```



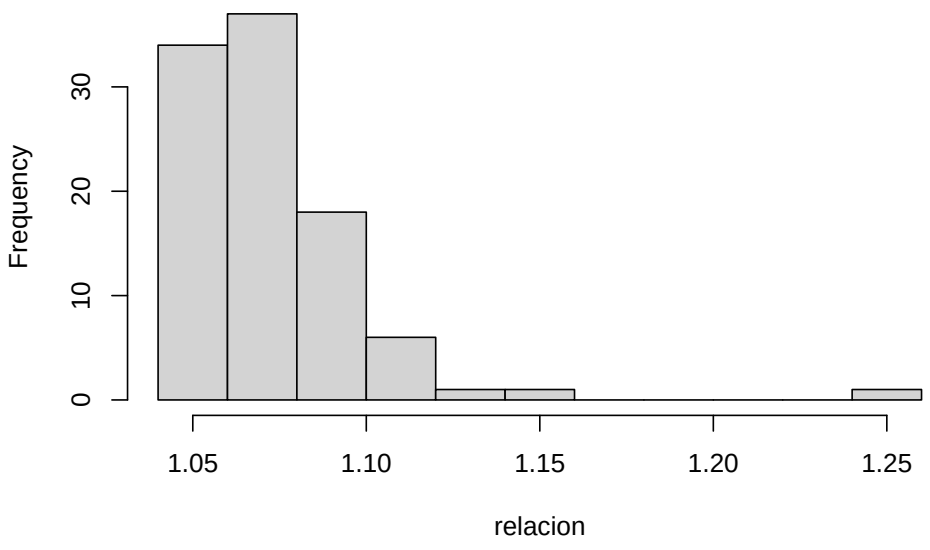

```
filtered_df_0_05 <- df %>%
  filter(rel_cero >= 0.05)
dim(filtered_df_0_05)
```

```
## [1] 98 10
```

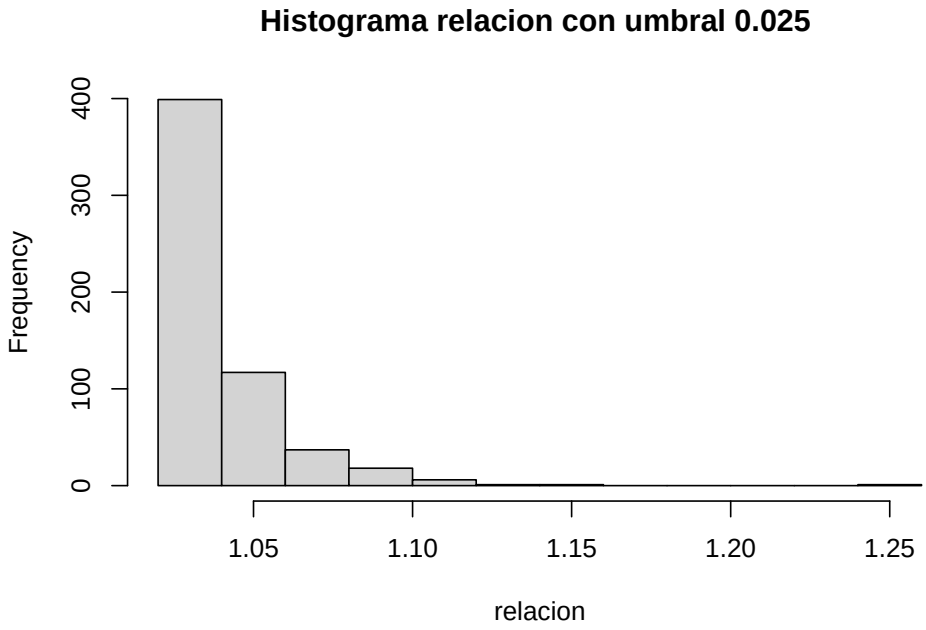
```
plot(filtered_df_0_05)
```




```
hist(filtered_df_0_05$relacion, main = "Histograma relacion con umbral 0.05 ", xlab = "relac
```




```
hist(filtered_df_0_025$relacion, main = "Histograma relacion con umbral 0.025 ", xlab = "rel
```



Apéndice

Teoría asintótica clásica y las distribuciones extremales y sus dominios de atracción

Siguiendo las Siguiendo a Perera, Segura, y Crisci (2021), se dice que tenemos *datos extremos* cuando cada dato corresponde al máximo o mínimo de varios registros.

Ejemplos de este tipo de datos son:

- La máxima altura semanal de la ola en una plataforma marina o portuaria (m)
- máxima velocidad de viento en determinada dirección a lo largo de un mes (km/h)
- temperatura ambiental mínima a lo largo de un día (C)
- La máxima velocidad de tráfico en un enlace de una red de datos de datos en una hora (Mb/s).
- El mayor registro en un conteo de Coliformes fecales sobre agua costeras al cabo de quince días.

Son un caso particular de *evento raro* o *gran desviación* respecto a la media. En resumen, en una gran variedad de dominios disciplinares suele ser de gran interés el trabajo con datos extremos, los que admiten diversos enfoques. Entre ellos, los propios al párrafo anterior (eventos raros, grandes desviaciones), que se verán en el curso. Sin embargo, el comienzo del curso se centra en la teoría más clásica de estadística de datos extremos, basada en el trabajo de Fréchet, Gumbel, Weibull, Fisher, Tippett, Gnedenko, entre otros.

Observación 1.

Se recuerda que si X e Y son variables aleatorias independientes, cuyas distribuciones son, respectivamente, F y G, entonces la variable

$$max(X, Y) \tag{2}$$

tiene por distribución la función H definida por

$$H(t) = F(t)G(t) \tag{3}$$

Observación 2.

En esta parte inicial del curso asumiremos que nuestros datos son $i.i.d$ (independientes e idénticamente distribuidos, son DOS suposiciones juntas). Esta doble suposición suele NO ser realista en aplicaciones concretas (ninguna de sus dos componentes, incluso) pero para comenzar a entender la teoría clásica, la utilizaremos por un tiempo.

Observación 3.

Resulta claramente de la Observación 1, que si tenemos datos $X_1, ..., X_n$ $i.i.d$ con distribución F, entonces

$$X_n^* = max(X_1, ..., X_n) \tag{4}$$

tiene distribución F_n^* dada por

$$F_n^*(t) = F(t)^n \tag{5}$$

Si conocemos la distribución F conoceríamos la distribución F_n^* , pero en algunos casos la lectura que queda registrada es la del dato máximo y no la de cada observación que dio lugar al mismo, por lo que a veces ni siquiera es viable estimar F.

Pero aún en los casos en que F es conocida o estimable, si n es grande, la fórmula de F_n^* puede resultar prácticamente inmanejable. En una línea de trabajo similar a la que aporta el Teorema Central del Límite en la estadística de valores medios, un teorema nos va a permitir aproximar F_n^* por distribuciones más

sencillas. Este es el Teorema de Fischer-Tippet-Gnedenko (FTG, para abreviar) que presentaremos en breve.

Observación 4.

Como $X_1, ..., X_n$ *i.i.d*, definimos $Y_i = -X_i$ para todo valor de i, entonces $Y_1, ..., Y_n$ *i.i.d* y además

$$min(X_1, ..., X_n) = -max(Y_1, ..., Y_n) \tag{6}$$

la teoría asintótica de los mínimos de datos *i.i.d* se reduce a la de los máximos, razón por la que nos concentramos aquí en estudiar el comportamiento asintótico de los máximos exclusivamente.

Definición 1: Las distribuciones extremales

Las distribuciones extremales son tres: la distribución de Gumbel; la distribución de Weibull; la distribución de Fréchet. En su versión standard o típica se definen del modo siguiente.

Distribución de Gumbel Se dice que una variable tiene distribución de Gumbel si su distribución es:

$$\Lambda(x) = exp\{-e^{-x}\} \quad \text{para todo } x \text{ real} \tag{7}$$

Cuando tomamos los máximos de variables no acotadas pero que tienen colas livianas (ej. la distribución tiene probabilidades muy bajas de tomar valores lejos de la media) los mismos convergen a una distribución asintótica extremal de Gumbel.

Para simular distribuciones de Gumbel, utilizamos el paquete **evd** de Stephenson (2002) y en particular la función **pgumbel**. Partiendo de una simulación de números aleatorios, para un secuencia de 1000 números entre $[-10, 10]$, se tienen las siguientes figuras relativas a la CDF y PDF de la distribución Gumbel.

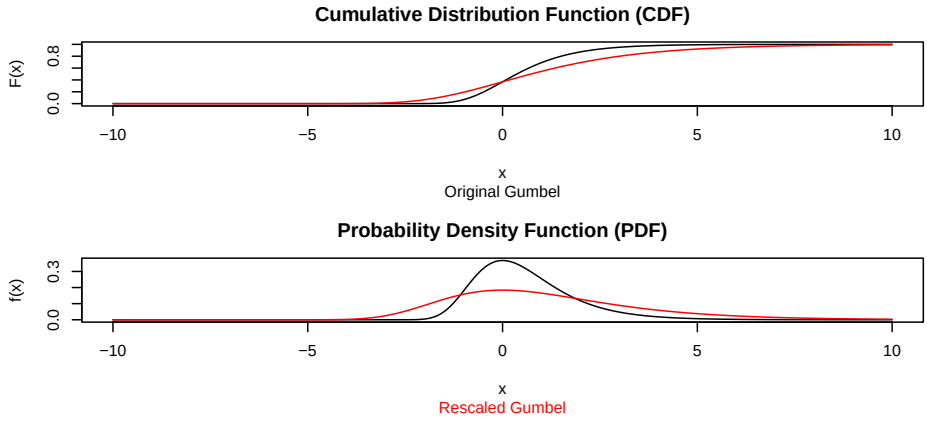
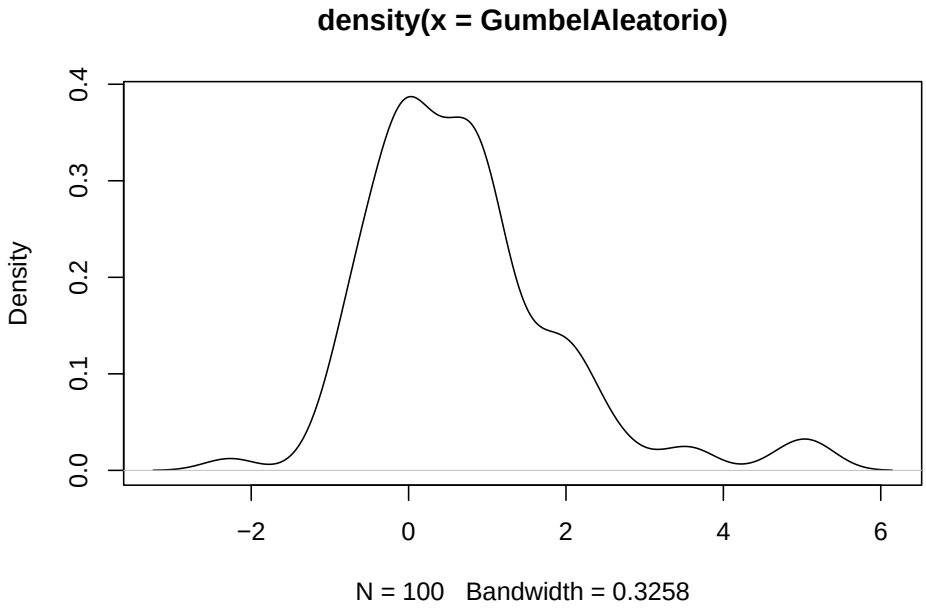


Figura 1: CDF and PDF for Gumbel distribution.

Si calculamos el valor esperado y el desvío estandard de estos valores observados y tenemos una muestra lo suficientemente grande, podremos comparar los resultados con los esperados de forma teórica.

```
# Podemos simular 100 datos aleatorios de una distribución Gumbel
GumbelAleatorio<-rgumbel(100)
plot(density(GumbelAleatorio))
```




```
-digamma(1) # Constante de Euler-Mascheroni
```

```
## [1] 0.5772157
```

```
mean(rgumbel(1000))
```

```
## [1] 0.5576658
```

```
sd(rgumbel(1000))
```

```
## [1] 1.308002
```

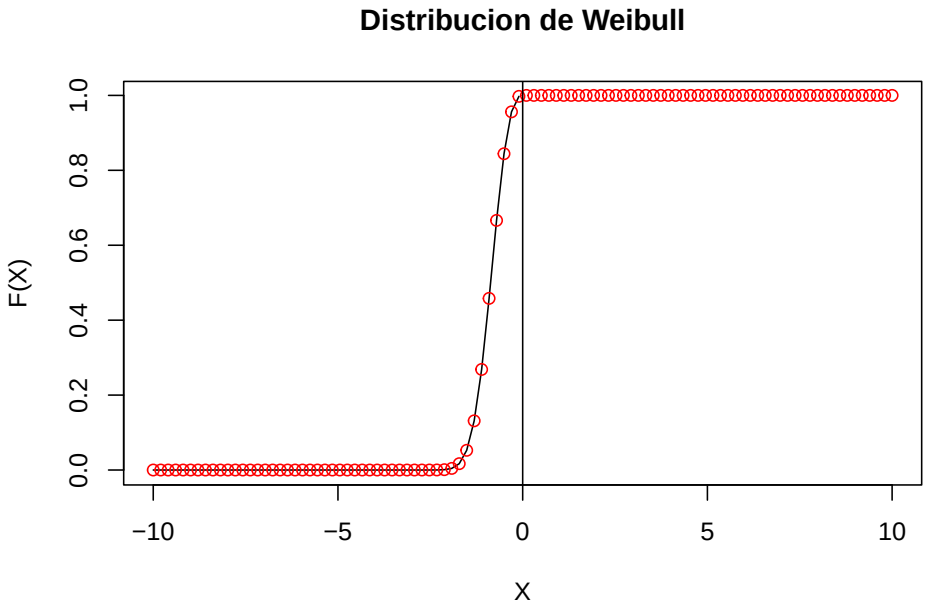
Distribución de Weibull Se dice que una variable tiene distribución de Weibull de orden $\alpha > 0$ si su distribución es:

$$\Psi_\alpha(x) = \begin{cases} exp{-(-x)^\alpha} & si\ x < 0 \\ 1 & \text{en otro caso} \end{cases} \tag{8}$$

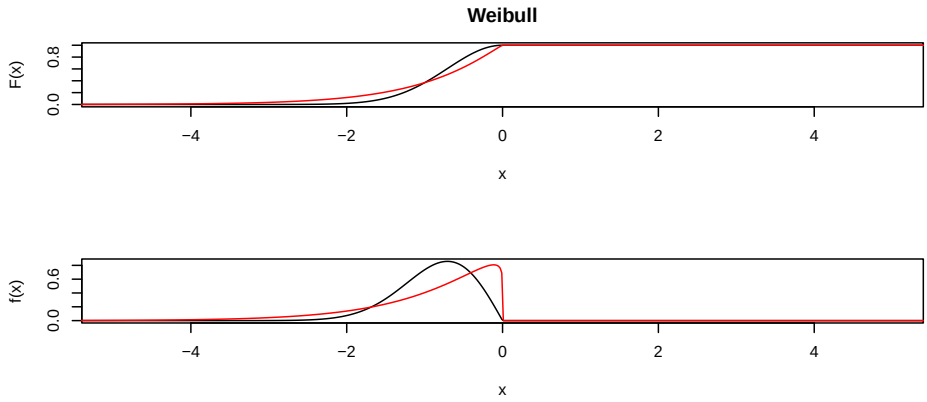
Recordemos que cuando tomamos los máximos de las variables *iid* con un rango acotado, la distribución resultante por la cual se puede aproximar es la de Weibull. En este caso, y en el resto del LAB, exp() y e son la función exponencial.

Por una única vez, calculemos la distribución de forma "manual" en el R para convencernos de la forma de la función de distribución de Weibull (Ψ). Para eso generaremos un vector auxiliar de valores x y la distribución ($F(x)$). En R la definición de la distribución es sutilmente diferente a la que vimos en el teórico (definida para positivos), pero totalmente convertible con dos cambios de signo. La función que calcula la probabilidad de una distribución Weibull es **pweibull()**. Pueden ver la definición de R utilizando help(pweibull) o ?pweibull.En R podemos saber la forma y valores de esta distribución con una función implementada en un paquete base {stats}. La función es pweibull y lleva como argumentos un vector de cuantiles (q), un argumento de forma (shape) y otro de escala (scale). Recordemos que la función plot utiliza 2 argumentos centrales (x e y) y podemos fijar los límites del gráfico (xlim e ylim), el tipo de gráfico (type) y las etiquetas de los ejes X e Y (xlab e ylab).

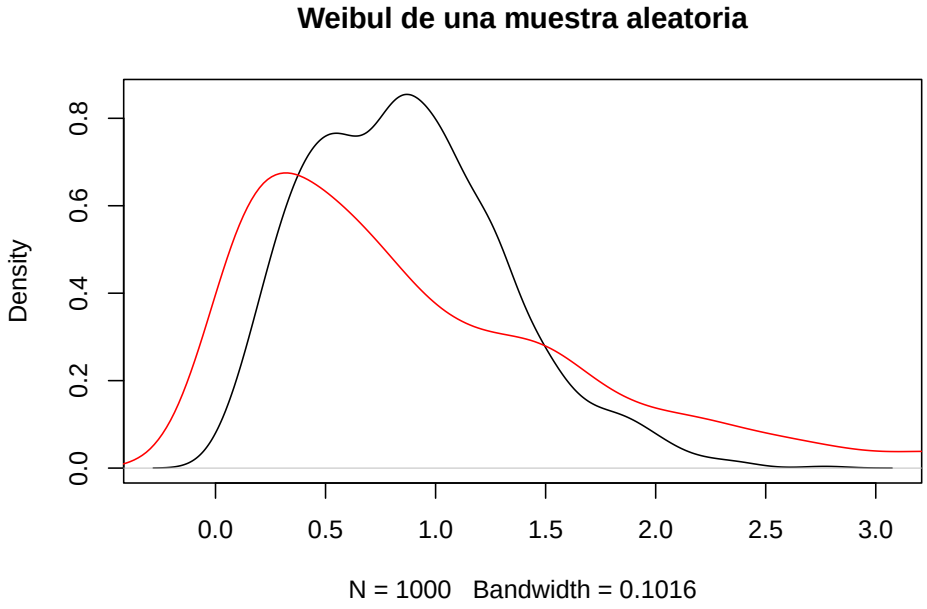
Primero generaremos un vector de numeros auxiliares equiespaciados y lo nombraremos ("x_aux"). Luego definiremos un orden (alpha=α) de la Weibull y graficaremos la función.



Veamos ahora la forma de un par de distribuciones cambiando el parámetro de orden (α), que en la función pweibull de R se nombra como shape y que define el orden de la distribución.



En R podemos también generar numeros aleatórios (técnicamente pseudo-aleatorios) de una distribución extremal. Estos simuladores de números aleatórios son útiles para comparar contra distribuciones nulas, generar modelos sintéticos para probar algorítmos, etc… Para lxs que venimos de la rama mas aplicada, muchas veces nos ayudan a entender como funcionan los modelos y a verificar si nuestra intuición es acertada respecto a la escala de ajuste de los parámetros entre otras útiles. Generaremos 2 series de 1000 números aleatórios con la función rweibull, que tiene como parámetro el número de datos que se necesitan y la forma (shape) de la distribución. Luego haremos un grafico con la densidad empírica (esto es similar a un histograma) de estos vectores.



Distribución de Fréchet Se dice que una variable tiene distribución de Fréchet de orden $\alpha > 0$ si su distribución es:

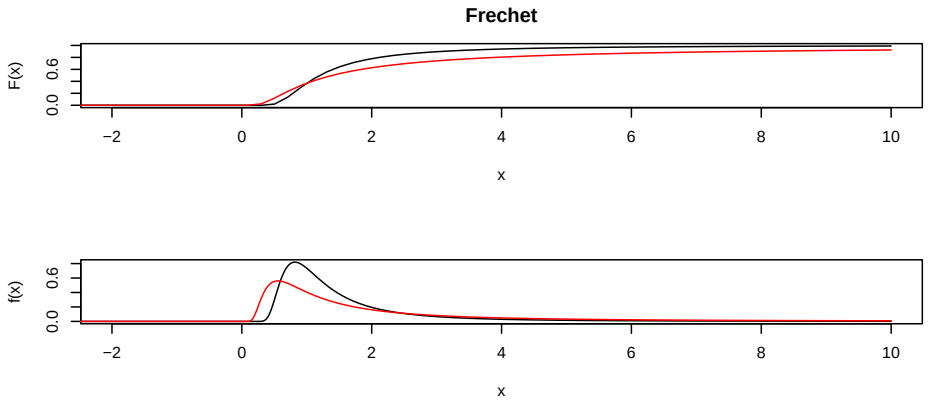
$$\Phi_\alpha(x) = \begin{cases} exp\{-x^{-\alpha}\} & si\ x > 0 \\ 0 & \text{en otro caso} \end{cases} \tag{9}$$

Esta tercera clase de variables incluyen a las distribuciones no acotadas, pero de colas pesadas. Es decir que tienen una probabilidad alta de presentar valores alejados de la media o la mediana (ej. la Cauchy). En estos casos, la distribución de sus máximos es la Frechet. Grafiquemos esta distribución para dos valores diferentes de α.

```
x_aux<- seq(-10,10, length=1000)

par(mfrow=c(3,1), mar=c(5,4,3,1))
plot(seq(-10,10,length=100), pfrechet(q=seq(-10,10,length=100),
                                       shape=2, scale=1) ,xlim=c(-2,10), type="l", ylab="F(x)
lines(seq(-10,10,length=100),
      pfrechet(q=seq(-10,10,length=100), shape=1.1, scale=1),col= "red")

plot(x_aux, dfrechet(x=x_aux,
                     shape=2, scale=1, log = FALSE) ,xlim=c(-2,10), type="l", ylab="f(x)", x
lines(x_aux, dfrechet(x=x_aux,
                      shape=1.1, scale=1, log = FALSE), col="red")
```



.

Nota:

Como los máximos en general son valores grandes, importa particularmente observar el comportamiento de estas distribuciones para x tendiendo a infinito. El límite es 1 como en toda distribución. Pero VA MAS RAPIDO a 1 la Weibull, luego la Gumbel y luego la Fréchet. Esto es indicio que la Fréchet modela datos

"más extremos", máximos de datos de colas más pesadas que la Gumbel y esta que la Weibull. Más adelante veremos esto más precisamente. En la Fréchet, la velocidad de convergencia a 1 crece al aumentar el orden. En cambio en la Weibull el orden afecta la velocidad con que va a 0 cuando x tiende a menos infinito, que crece cuanto mayor el orden. Esto quedará más claro con el Teorema 1 del curso. La visualización de las densidades de cada tipo quizás ayude a comprender mejor los pesos relativos de las colas.