Datos Extremales (2025)

MEDIA

2025-02-26

Índice general

Capítulo 1

La teoría asintótica clásica, las distribuciones extremales y sus dominios de atracción

1.1 Datos extremos

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).
- La máxima velocidad de viento en determinada dirección a lo largo de un mes (km/h).
- La temperatura ambiental mínima a lo largo de un día $(\dot{C})$.
- La temperatura ambiental mínima a lo largo de un día $(\dot{C})$
- 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

$$\max(X, Y) \tag{1.1}$$

tiene por distribución la función H definida por

$$H(t) = F(t)\, G(t) \tag{1.2}$$

Observación 2: En esta parte inicial del curso asumiremos que nuestros datos son *iid* (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 $X_1, ..., X_n$ *iid* con distribución F, entonces

$$X_n^* = \max(X_1, ..., X_n) \tag{1.3}$$

tiene distribución F_n^* dada por

$$F_n^*(t) = F(t)^n \tag{1.4}$$

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) que presentaremos en breve.

Observación 4: Si $X_1, ..., X_n$ es *iid* y definimos $Y_i = -X_i$ para todo valor de i, entonces $Y_1, ..., Y_n$ es *iid* y además

$$min(X_1, ..., X_n) = -max(Y_1, ..., Y_n) \tag{1.5}$$

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

1.2 Las distribuciones extremales

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

Se dice que una variable tiene distribución de:

-**Gumbel** si su distribución es

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

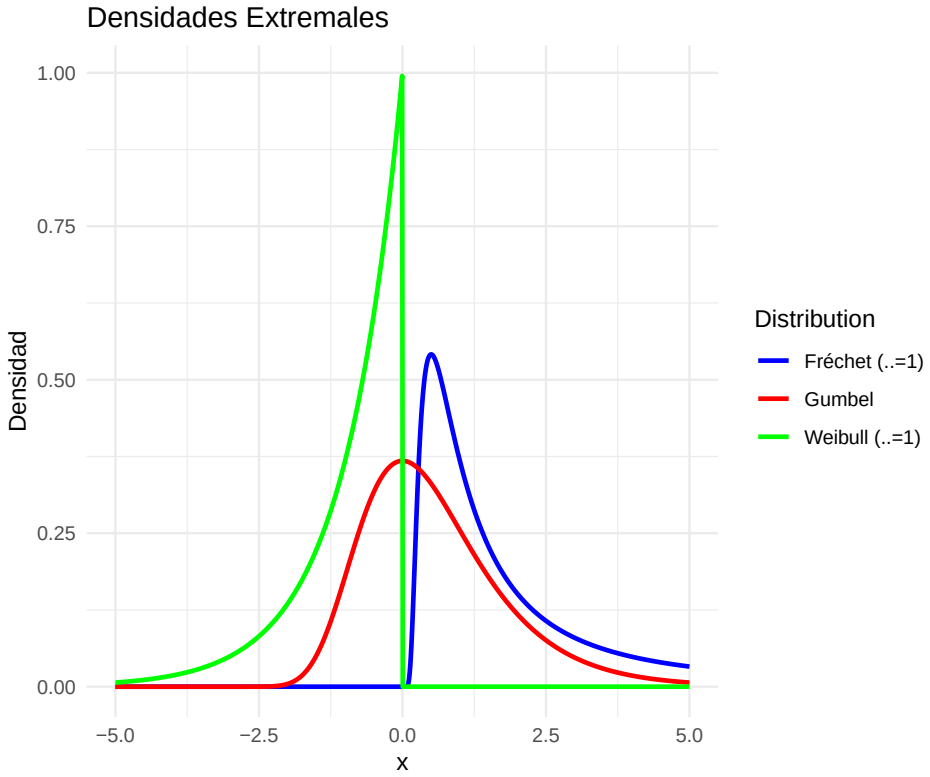
-**Weibull** de orden $\alpha > 0$ si su distribución es

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

-**Fréchet** de orden $\alpha > 0$ si su distribución es

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

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 ésta 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.



A estas versiones standard se las puede extender agregando un parámetro de recentramiento (μ) y un parámetro de escala (β).

Se dice que X tiene distribución:

- **Gumbel** : $\Lambda^{(\mu,\beta)}$ si $X = \mu + \beta Y$, donde Y tiene distribución Λ.
- **Weibull**: $\Psi^{(\mu,\beta)}$ si $X = \mu + \beta Y$, donde Y tiene distribución Ψ_α.
- **Fréchet**: $\Phi^{(\mu,\beta)}$ si $X = \mu + \beta Y$, donde Y tiene distribución Φ_α.

En general, es en este sentido que diremos que una variable es Gumbel, Weibull o Fréchet (incluyendo recentramiento y reescalamiento), pero en cálculos donde los parámetros μ y β no sean relevantes, por simplicidad, usaremos las versiones standard.

El siguiente teorema vincula las distribuciones extremales en sus formatos standard y resulta de gran utilidad práctica sobre todo al hacer tests de ajustes, etc.

Teorema 1 : ***Relaciones entre las versiones standard de las distribuciones extremales.***

X tiene distribución $\Phi_\alpha \Leftrightarrow (-1/X)$ tiene distribución $\Psi_\alpha \Leftrightarrow \log(X^\alpha)$ tiene distribución Λ.

Nota: en otros contextos de la Estadística (en particular en alguna rutinas del R), se le llama Weibull a una variable que corresponde a -X, con X Weibull como definimos nosotros.

Observación 5: Recordamos que la función Gamma (Γ), que extiende a la función factorial ($\Gamma(n) = n-1! \quad \forall n$ natural) definida por

$$\Gamma(x) = \int_0^\infty t^{u-1} e^{-t} dt \tag{1.6}$$

es una función disponible tanto en el software R como en planillas de cálculo, etc.

Teorema 2: (Tres en uno) ***Algunos datos de las distribuciones extremales.***

Parte 1

Si X tiene distribución $\Lambda^{(\mu,\beta)}$ entonces tiene:

a) **Esperanza**: $E(X) = \mu + \beta\gamma$, donde γ es la constante de Euler-Mascheroni, cuyo valor aproximado es 0.5772156649.

b) **Moda**: $\text{moda}(X) = \mu$

c) **Mediana**: $\text{med}(X) = \mu - \beta \log(\log 2) \approx \mu - 0.36651\beta$

d) **Desviación estándar**: $\sigma(X) = \frac{\beta\pi}{\sqrt{6}} \approx 1.2825\beta$

e) Si $X^+ = \max(X, 0)$, entonces $E(X + k)$ es finito para todo valor de k natural

f) Para simular computacionalmente X, se puede tomar U uniforme en $(0, 1)$ y hacer $X = \mu - \beta \log(-\log U)$.

Parte 2

Si X tiene distribución $\Psi^{(\mu,\beta)}$ entonces tiene:

a) $E(X) = \mu - \beta\Gamma(1 + 1/\alpha)$

b)
$$\text{moda}(X) = \begin{cases} \mu & \text{si } \alpha \leq 1 \\ \mu - \beta \left\{ \frac{(\alpha-1)}{\alpha} \right\}^{1/\alpha} & \text{si } \alpha > 1 \end{cases}$$

c) $\text{med}(X) = \mu - \beta(\log 2)^{\frac{1}{\alpha}}$

d) $\sigma(X) = \beta \left\{ \Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma\left(1 + \frac{1}{\alpha}\right)^2 \right\}^{1/2}$.

Parte 3

Si X tiene distribución $\Phi_\alpha^{(\mu,\beta)}$ entonces tiene:

a)
$$E(x) = \begin{cases} \mu + \beta\, \Gamma\left(1 - \frac{1}{\alpha}\right) & \text{si } \alpha > 1 \\ \infty & \text{en otro caso} \end{cases}$$

b) $\text{moda}(X) = \mu + \beta \ \left\{ \frac{\alpha}{(1+\alpha)} \right\}^{1/\alpha}$

c) $\text{med}(X) = \mu + \beta \ (\log 2)^{(-1/\alpha)}$

d)
$$\sigma(x) = \begin{cases} \mu + \left|\Gamma\left(1 - \frac{2}{\alpha}\right) - \Gamma\left(1 - \frac{1}{\alpha}\right)\right| & \text{si } \ \alpha > 2 \\ \infty & \text{si } \ 1 < \alpha \leq 2 \end{cases}$$

Observación 6: El item e) de la Parte 1 es trivialmente cierto para Weibull y tomando en cuenta el item a) de la Parte 3, es claramente falso para Fréchet.

Observación 7: El item f) de la Parte 1 en conjunto con el Teorema 1 brinda fórmulas sencillas para simular computacionalmente distribuciones Weibull o Fréchet.

Observación 8: Se generaron mil números aleatorios y aplicando el item f) de la Parte 1: se simularon mil variables Gumbel standard *iid*, calculándose su promedio, su desviación standard empírica y su mediana empírica.

```r
# Fijar semilla para reproducibilidad
set.seed(123)

# Definir parámetros
mu <- 0       # Centro
beta <- 1     # Escala
gamma <- 0.5772156649  # Constante de Euler-Mascheroni

# Número de simulaciones
n <- 1000

# Generar 1000 valores de una variable uniforme en (0,1)
U <- runif(n)

# Simular la variable Gumbel con parámetros (mu, beta)
X_gumbel <- mu - beta * log(-log(U))
```

```r
# Calcular estadísticas
esperanza <- mu + beta * gamma
moda <- mu
mediana_teorica <- mu - beta * log(log(2))
desviacion_std_teorica <- beta * pi / sqrt(6)

# Calcular estadísticas empíricas
promedio_empirico <- mean(X_gumbel)
desviacion_std_empirica <- sd(X_gumbel)
mediana_empirica <- median(X_gumbel)
```

Los resultados fueron los siguientes:

```
## ----- Resultados teóricos: -----

## Esperanza teórica: 0.5772157

## Moda teórica: 0

## Mediana teórica: 0.3665129

## Desviación estándar teórica: 1.28255

## ----- Resultados empíricos (simulación con n = 1000 ): -----

## Promedio empírico: 0.5610296

## Desviación estándar empírica: 1.261928

## Mediana empírica: 0.3376409
```

Observar que los resultados empíricos están cerca del valor esperado, desvío standard y mediana de la Gumbel standard.

A continuación presentaremos el Teorema medular de esta primera parte, expresado de la manera más llana posible. Veremos posteriormente algunos detalles con más cuidado. En particular, veremos que la continuidad de la distribución F no es una hipótesis real (ni es necesaria ni es suficiente, por eso la entrecomillamos), pero ayuda a visualizar que no vale el teorema para toda distribución F, así como veremos con cierto detalle más adelante…

Teorema 3: de Fischer-Tippet-Gnedenko (FTG)

Si $X_1, ..., X_n$ es *iid* con distribución F 'continua', llamamos F_n^* a la distribución de $max(X_1, ..., X_n)$ y n es grande, entonces existen μ real y $\beta > 0$ tales que alguna de las siguientes tres afirmaciones es correcta:

a) F_n^* se puede aproximar por la distribución de $\mu + \beta Y$, con Y variable con distribución Λ.
b) Existe $\alpha > 0$ tal que F_n^* se puede aproximar por la distribución de $\mu + \beta Y$ con Y variable con distribución Φ_α.
c) Existe $\alpha > 0$ tal que F_n^* se puede aproximar por la distribución de $\mu + \beta Y$ con Y variable con distribución Φ_α.

Lo anterior equivale a decir que la distribución del máximo de datos *continuos* e *iid*, si n es grande, puede aproximarse por una Gumbel, una Fréchet o una Weibull.

Observación 9: Como veremos con cierto detalle, cuál de las tres aproximaciones es la válida depende de cómo sea la distribución F.

Por ejemplo, veremos que:

- Si F es normal o exponencial, se aplica a F_n^* la aproximación por una Gumbel .
- Si F es uniforme, vale para F_n^* la aproximación por una Weibull.
- Si F es Cauchy, la aproximación válida para F_n^* es por una Fréchet.

Más precisamente, cuál de las tres aproximaciones es la aplicable depende de la cola de F[1].

En concreto, Weibull aparece cuando F es la distribución de una variable acotada por arriba (como la Uniforme), Gumbel para distribuciones de variables no acotadas por arriba pero con colas muy livianas (caso Exponencial y Normal) y Fréchet para colas pesadas (caso Cauchy). Finalmente, si bien aclaramos que la hipótesis de continuidad de F no es esencial, veremos que si F es la distribución Binomial o Poisson, por mencionar dos ejemplos muy conocidos y sencillos, NO se puede aplicar ninguna de las tres aproximaciones anteriores.

Observación 10. Como consecuencia del *FTG* si se tienen datos de máximos, las distribuciones extremales son "candidatas" razonables para proponer en un ajuste. Sin embargo no debe pensarse que siempre se va a lograr ajustar a una de las tres distribuciones extremales, ya que hay al menos dos causas evidentes que podrían desbaratar la aplicación del FTG:

1) Que la cantidad de registros que se consideran al calcular cada máximo no sea suficientemente grande.

2) Que los registros que se consideran al calcular cada máximo no sean *iid*[2].

[1] Los valores de $F(t)$ para valores grandes de t.

[2] Al final del capítulo 2 se verá que esto puede subsanarse con versiones más generales del FTG.

Por consiguiente el *FTG* alienta a intentar ajustar datos extremales a una de las tres distribuciones extremales, pero no siempre un tal ajuste dará un resultado afirmativo.

Ejemplo 1. Veamos un ejemplo de ajuste. Los siguientes datos corresponden a los valores, en 80 puntos geográficos distintos de la región parisina, del máximo estival del contaminante atmosférico O_3 (no perceptible sensorialmente y con impacto sanitario serio). Cada dato es el máximo registro en cada sensor a lo largo de todo un verano; el contaminante se mide diariamente, por lo cual, cada uno de nuestros 80 datos es el máximo de unas 100 lecturas diarias.

```
## [1] "Primeros 6 datos:"
```

```
##      X_i
## 1 430.30
## 2 115.70
## 3   4.48
## 4  26.95
## 5  72.27
## 6 206.40
```

Los valores se miden en unidades de referencia standarizadas que, en particular, permiten comparar las medidas de lugares diferentes, independientemente de variables relevantes como altura e incidencia solar, por trabajo previo de calibración.

El objetivo del estudio en esta etapa es conocer la distribución de estos datos y en particular estimar la probabilidad de que el máximo estival en los 80 puntos supere el valor 50 (correspondiente a existencia de riesgo moderado).

Veamos los datos que tenemos:

```
## [1] "Cálculo de estadísticos básicos"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.48   23.44   52.77  183.93  166.82 1675.00
```

Como la mayoría de tests de ajustes suponen datos *iid*, realizaremos dos tests de aleatoriedad[3]:

- Runs test (Up & Down)
- Spearman correlation of ranks

[3]En inglés es *randomness*.

Para realizar el ajuste utilizaremos el test χ^2 de ajuste[4]. Este test requiere elegir una partición más o menos arbitraria de la recta real en intervalos; sin embargo es importante que en cada intervalo caiga una cantidad suficiente de datos de la muestra; en este caso hemos tomado como extremos de los intervalos los quintiles empíricos de nuestra muestra.

Una aclaración mucho más importante es que este test requiere estimar parámetros por el método de Máxima Verosimilitud Categórica, que da resultado distintos al método de Máxima Verosimilitud a secas[5].

```
##
##  Runs Test
##
## data:  as.factor(runs_sequence)
## Standard Normal = 2.4678, p-value = 0.01359
## alternative hypothesis: two.sided
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$X_i and seq_along(data$X_i)
## S = 85949, p-value = 0.9483
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##          rho
## -0.007372289
```

Como cada dato de los 80 que disponemos es un máximo de un centenar de observaciones, intentaremos ajustarlos a una distribución extremal sabiendo que no necesariamente tendremos éxito.

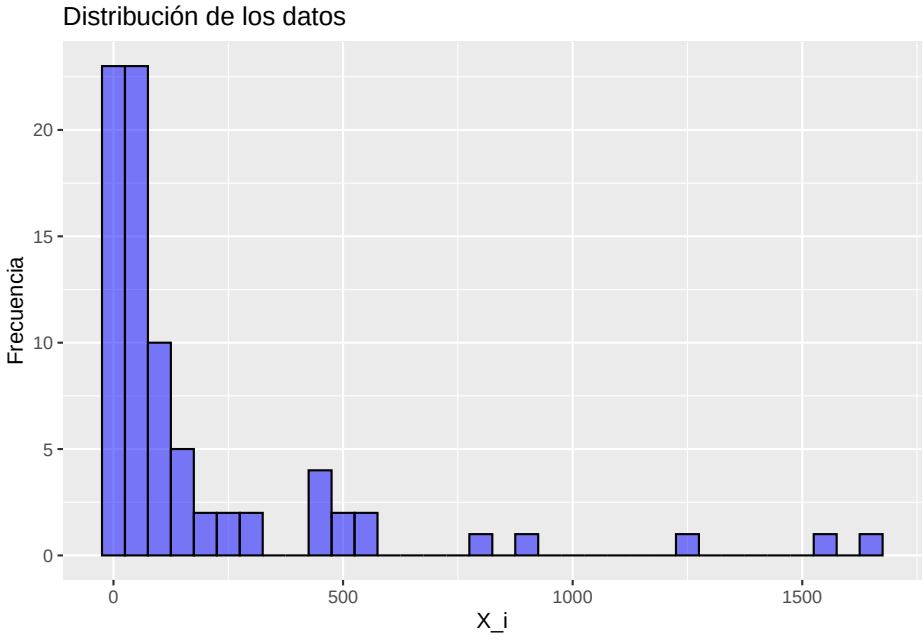
Observemos en particular que lo que pasamos por dos tests de aleatoriedad son los 80 máximos, pero no el centenar de lecturas que forman cada uno de los 80 máximos (ni siquiera tenemos esos datos originales).

Dado que visualmente se aprecian valores muy apartados, se presume una distribución de colas pesadas y por ese motivo se intenta un ajuste a una Fréchet.

---

[4]Una excelente referencia para la temática de los test χ^2 de ajuste es la introducción del trabajo Pearsonian Tests and Modifications (Jorge Graneri, CMAT, Facultad de Ciencias, 2002).

[5]Este hecho es frecuentemente ignorado y presentado erróneamente en los textos y cursos básicos de Estadística.




```r
# Parámetros del libro
loc_libro <- -6.5    #
scale_libro <- 44     #
shape_libro <- 1.04   #   (parámetro de forma positivo, Fréchet)

# Cálculo de la probabilidad de exceder el valor 50
prob_excede_50 <- 1 - pgev(50, loc = loc_libro, scale = scale_libro, shape = shape_libro)

# Mostrar la probabilidad de excedencia
print(paste("Probabilidad de excedencia del nivel 50:", round(prob_excede_50, 4)))
```

```
## [1] "Probabilidad de excedencia del nivel 50: 0.3575"
```

```r
# Proporción empírica de excedencia del nivel 50
prop_empirica <- mean(data$X_i > 50)
print(paste("Proporción empírica de excedencia del nivel 50:", round(prop_empirica, 4)))
```

```
## [1] "Proporción empírica de excedencia del nivel 50: 0.5125"
```

```r
# Intervalo de confianza para la proporción empírica
prop_ci <- prop.test(sum(data$X_i > 50), length(data$X_i))$conf.int
print(paste("Intervalo de confianza al 95%:", round(prop_ci[1], 3), "-", round(prop_ci[2], 3)))
```

```
## [1] "Intervalo de confianza al 95%: 0.399 - 0.625"
```

```r
print(paste("Probabilidad de excedencia del nivel 50:", round(prob_excede_50, 4)))
```

```
## [1] "Probabilidad de excedencia del nivel 50: 0.3575"
```

```r
# Parámetros del libro
loc_libro <- -6.5    #
scale_libro <- 44     #
shape_libro <- 1.04   #   (parámetro de forma positivo, Fréchet)

# Cálculo de la probabilidad de exceder el valor 50
prob_excede_50 <- 1 - pgev(50, loc = loc_libro, scale = scale_libro, shape = shape_libr
print(paste("Probabilidad de excedencia del nivel 50:", round(prob_excede_50, 4)))
```

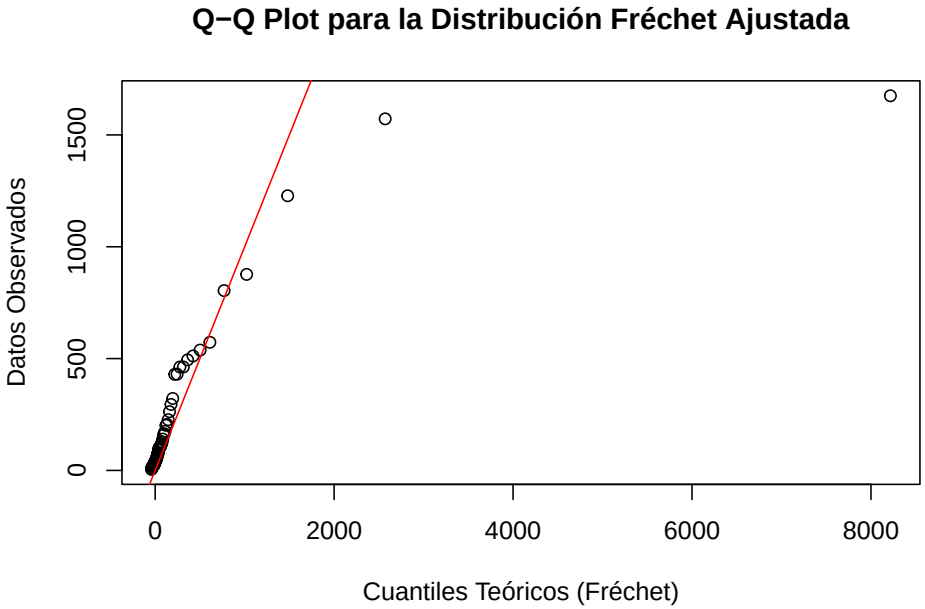
```
## [1] "Probabilidad de excedencia del nivel 50: 0.3575"
```

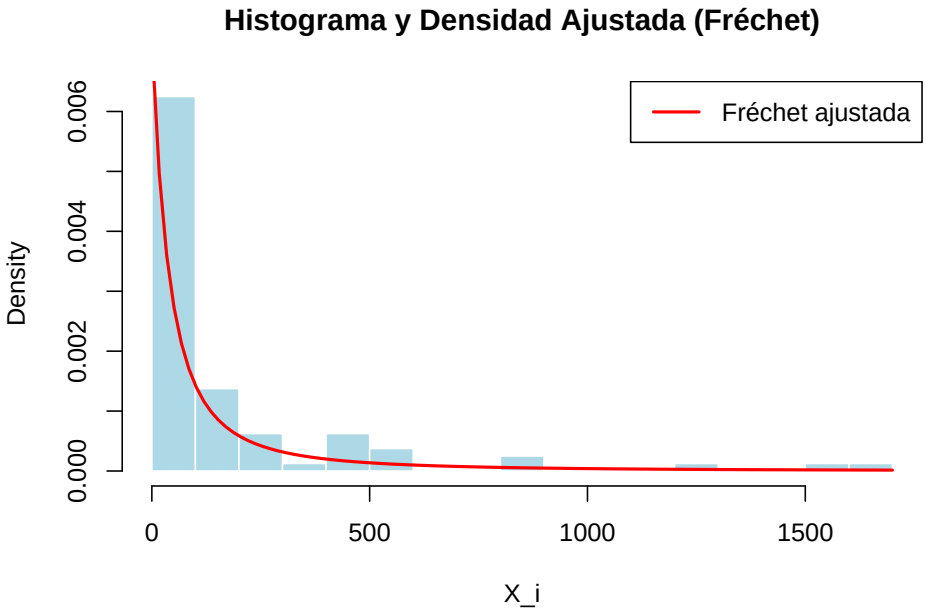
```r
# Parámetros del libro
loc_libro <- -6.5    #
scale_libro <- 44     #
shape_libro <- 1.04   #   (parámetro de forma positivo, Fréchet)

# Definir intervalos usando los quintiles empíricos
breaks <- quantile(data$X_i, probs = seq(0, 1, length.out = 6))  # 5 intervalos

# Calcular las frecuencias observadas en cada intervalo
observed_counts <- hist(data$X_i, breaks = breaks, plot = FALSE)$counts

# Calcular las probabilidades teóricas en cada intervalo usando la distribución Fréche
probs <- diff(pgev(breaks, loc = loc_libro, scale = scale_libro, shape = shape_libro))

# Convertir probabilidades en frecuencias esperadas
expected_counts <- probs * length(data$X_i)

# Realizar el test de ajuste Chi-cuadrado
chi_sq_test <- chisq.test(observed_counts, p = probs, rescale.p = TRUE)

# Mostrar los resultados del test
print("Resultados del test Chi-cuadrado con los parámetros del libro:")
```

```
## [1] "Resultados del test Chi-cuadrado con los parámetros del libro:"
```

```r
print(chi_sq_test)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed_counts
## X-squared = 4.3938, df = 4, p-value = 0.3553
```





Observación 10. Una distribución H se dice degenerada si $H(t) = 0$ ó 1 para