Universidad de Buenos Aires

Facultad de Ingeniería

91.03 – ESTADÍSTICA APLICADA I

Trabajo Práctico N° 1 :

Distribuciones de Extremos

Grupo Nº 4

Docente: Montemurri, David

Alumnos: Spaventa, Maria - 98771

Rosso, Sebastian - 101633

Fecha de entrega: 20/05/2021

TP1 Estadística Aplicada I
Spaventa, María - 98771
Rosso, Sebastián - 101633

1.Objetivos

Analizar y comprender los conceptos teóricos y prácticos de las distribuciones de extremos.

2.Desarrollo

2.1. Distribuciones de los extremos

El modelo para el que se desarrolla la teoría de valores extremos está enfocado a describir el comportamiento estadıstico de:

Mn = máx{X1, ..., Xn},

donde X1, ..., Xn es una secuencia de variables aleatorias independientes con distribución común F y Mn representa el máximo del proceso sobre n unidades de tiempos de observación.

Teorema: Si existen constantes an > 0 y bn ∈ R para n ≥ 1 tales que
P $(\frac{Mn - bn}{an} < x) \rightarrow G(x)$, cuando n → ∞,
siendo G una función de distribución no degenerada, entonces G debe pertenecer a una de las siguientes familias:

- Gumbel, para las colas medias:
  $\Lambda_{\mu,\sigma}(x) = \exp(-e^{-(x-\mu)/\sigma})$ x ∈ R.
- Fréchet, para colas gruesas:
  $\Phi_{\alpha,\mu,\sigma}(x) =$ 0 si x<μ; $\exp(-((x-\mu)/\sigma)^{-\alpha})$ si x ≥ μ
- Weibull, para colas cortas o suaves:
  $\Psi_{\alpha,\mu,\sigma}(x) = \exp(-(-(x-\mu)/\sigma)^{\alpha})$ si x < μ o 1 x ≥ 0

Las anteriores distribuciones se conocen como distribuciones de valores extremos y serán las únicas a las que pueda converger la variable Mn, independientemente de cómo se distribuya. Los tres tipos de distribuciones de valores extremos (DVE) pueden ser combinados en una sola distribución con parametrización común que se conoce como la Distribución Generalizada de Valores Extremos (DGVE o GEV). La forma de esta distribución es

$$G_{\xi,\mu,\sigma}(x) = \exp\{-(1 + \xi(\frac{x-\mu}{\sigma}))^{-1/\xi}\}$$

El parámetro ξ es el parámetro de forma o índice de cola. El valor del mismo identificará la distribución y determinará el grueso de la cola. Cuanto mayor sea este índice, más gruesa será la cola. Entonces tendremos:

Para $\xi > 0 \rightarrow$ distribución de Fréchet con $\alpha = 1/\xi$.
Para $\xi < 0 \rightarrow$ distribución de Weibull con $\alpha = -1/\xi$
Para $\xi = 0 \rightarrow$ distribución de Gumbel

2.2. Criterios de convergencia

La idea de valor extremo considera una cantidad muy grande de información, es decir, un número de variables aleatorias ordenadas que tiende a infinito. Un concepto fundamental, que se utiliza de manera implícita es el de convergencia.

Dada una sucesión de variables aleatorias (X1, X2, ...) con sus respectivas funciones de distribución F1(x), F2(x),... y dadas X y F(x) otra variable aleatoria con su respectiva función de distribución, decimos que la sucesión (X1, X2, ...) converge en distribución a la variable aleatoria X si se cumple:

$$\lim_{n\to\infty} F_n(x) = F(x) \ \forall x \in R \text{ donde F es continua}$$

También es conocida como convergencia débil

2.3. Aplicaciones

Los valores extremos tienen muchas aplicaciones en la práctica. Algunas aplicaciones de la teoría de valores extremos son el estudio de la longevidad de la vida humana, la gestión de tráfico (en telecomunicaciones), la resistencia de materiales, la concentración de ozono, geología o meteorología (lluvias, vientos, etc).

A continuación se presentan algunos ejemplos concretos:
En la metalurgia, puede ser usada para determinar la calidad de algún metal pues se sabe que cuando un metal es sometido a alguna carga cíclica, en algún punto, dicho metal se romperá. Lo anterior es mayormente conocido como fatiga de materiales, y es tratando de encontrar el punto máximo en el cual un metal o material es resistente a una carga cíclica que aplicamos la TVE.

Otro ejemplo que podemos mencionar es: un neumático de un coche puede estropearse de dos formas. Por cada día que se usa el coche, el neumático se desgasta un poco más, y con el paso del tiempo y como consecuencia del deterioro acumulado, el neumático acabará rompiéndose. Pero también puede ocurrir que al conducir se pise un bache, o que el coche golpee la acera. Puede pasar que esos accidentes no tengan efectos en los neumáticos, o

que el neumático termine perforado, en cuyo caso sólo una observación sería la que causará un fallo, lo que significa que el máximo parcial supere cierto umbral.

Otro ejemplo podría ser sobre la velocidad máxima a la que circulan vehículos en una parte concreta de la autopista, ya que en función de esos datos se puede decidir el uso de coches patrulla por dicha zona; u otro ejemplo muy parecido sería el número máximo de vehículos que circulan por una intersección a una hora determinada, pues el conocer dicho máximo facilitara un mejor control del tránsito vehicular.

2.4. Simulaciones

Simulamos 50 muestras de 100 observaciones para la distribución Gamma de parámetros: 1 para la forma y 52 para la escala (promedio entre 33 y 71), en lenguaje R.
X ~ Γ(1,52)

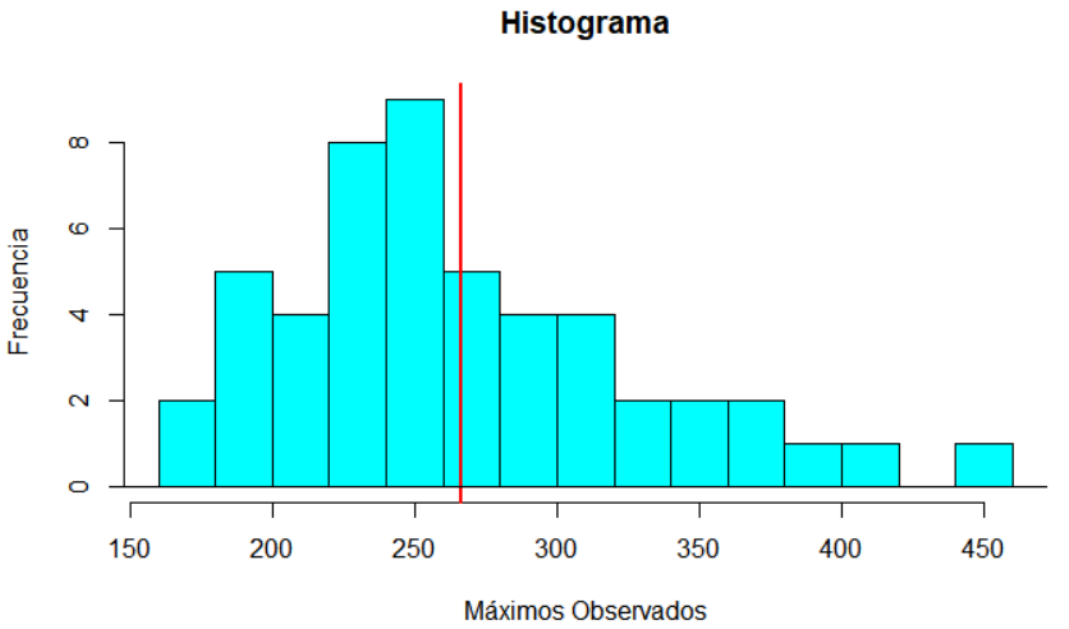
Con esto quedó generado una lista de listas que corresponden a dichas 50 simulaciones, cada una de 100 observaciones para la distribución, y podemos visualizar, por ejemplo, los resultados de la primera simulación:

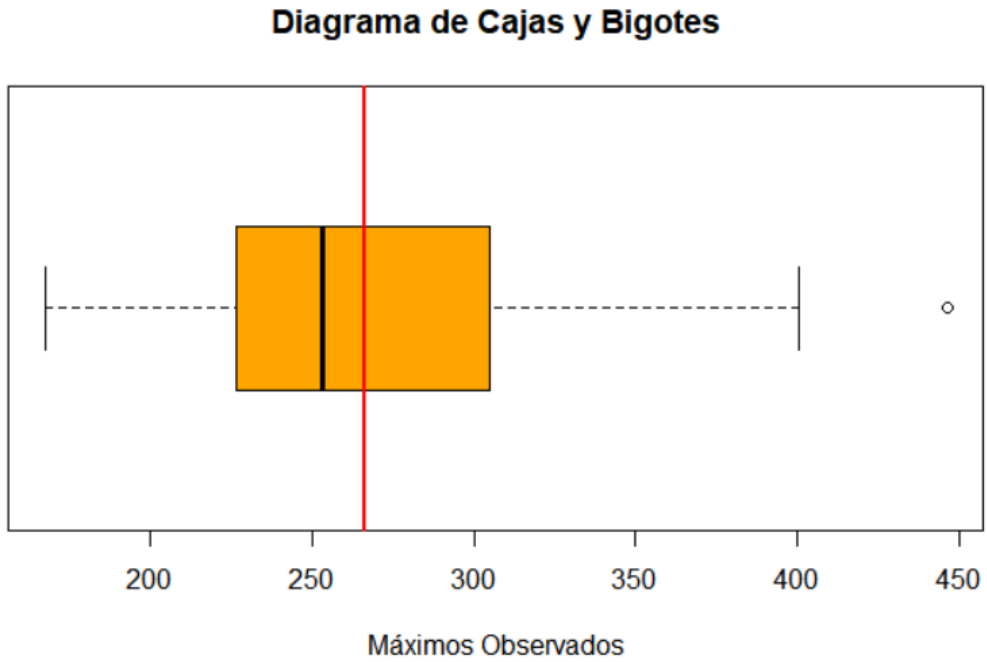
```
> X[[1]]
  [1]   4.273613   9.268978  28.643078 127.006303   8.556818  31.762089   4.362224  31.326138
  [9]  90.731410   4.875982  24.180908   1.524567  15.381370  69.653404  72.164650 107.990160
 [17]  58.480935  64.834675  16.328514 149.346156  16.068854   4.097474   2.929985 171.268946
 [25]  90.009793 108.835586  72.760350  90.599219  11.226755   4.865737 197.210155  50.781938
 [33]   3.613287  69.626326  39.084453  22.069527  33.360739  77.158223   8.791988  11.405253
 [41]  61.075766  22.024265 231.542039  19.036957  94.224343  52.442876  41.433845  48.529665
 [49]  37.480256  80.472961   9.421883  12.830403  74.241310  25.122023 159.177395 107.113036
 [57] 156.586722   1.608944  33.167329  19.720758  44.503042   1.051782  56.785742  34.230885
 [65]  23.916541  57.400201  25.210042   5.674872 446.313503 182.721392 144.658762  45.454507
 [73]  37.671424   1.507424  16.856047   4.968525  11.214824  73.120300  15.993209  45.233867
 [81]  41.407752  21.850435  13.568056   2.326720  10.456109 105.598729  88.596658  12.254287
 [89]  31.492790  37.403356  11.208306   6.489735 202.345556   9.645504   1.545020  79.016383
 [97]   5.145715  13.796752  15.116709  56.338002
```

2.5. Minimos y Maximos

Luego creamos una lista que contiene cada uno de los valores máximos observados en las distintas muestras, y a partir de ella graficamos un histograma y un diagrama de cajas y bigotes (o boxplot) que visualizamos a continuación:

```
> maximos
 [1] 446.3135 262.8784 305.1803 222.6448 385.9725 198.0463 307.0760 320.7140 267.1810 311.5197
[11] 327.6887 183.2137 363.9741 250.7067 227.2005 289.8386 193.3562 241.3617 294.2095 289.8938
[21] 274.7382 176.8358 290.2869 229.2245 400.7292 217.4226 249.9619 228.1401 274.2377 234.0201
[31] 208.0757 255.5908 226.8075 167.4307 250.5058 272.7456 188.6178 191.0907 251.9746 340.8773
[41] 201.9596 354.7878 233.4616 254.3740 312.6666 215.0805 250.1808 236.0993 254.1256 376.3286
```

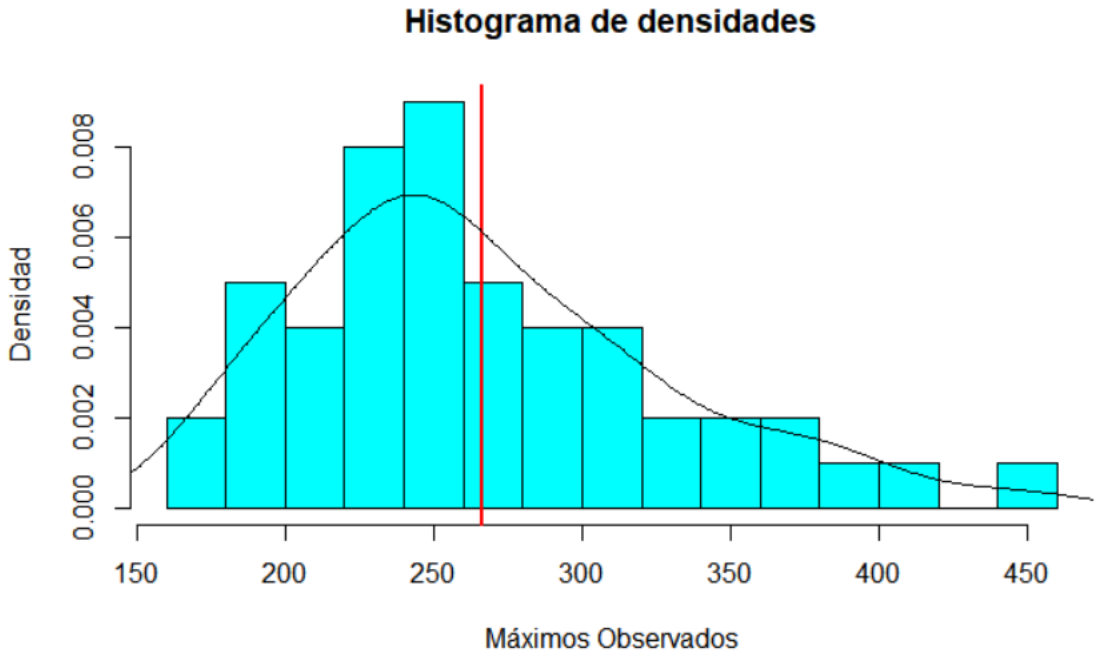




La línea roja colocada en ambos diagramas nos marca la media muestral.
En el boxplot se aprecia que el máximo de la primera muestra (446,3135) es un outlayer.

2.6. Distribución propuesta

Para proponer una distribución nos es más útil ver el histograma graficado en función de la densidad, que sería el siguiente:



Podemos notar que la curva obtenida es asimétrica positiva ya que la mayoría de los valores están concentrados a la izquierda. Además, podemos ver que la cola es media y se concentra a la derecha de la función densidad. Este tipo de curvas coincide con la Distribucion de Gumbel (Máximo).
Por otro lado, se puede decir que F(x) se encuentra en el máximo dominio de atracción de G(x) si:

Distribución Inicial F(x)	Distribución Límite para los máximos G(x)
Exponencial Gamma Normal Log-normal	Gumbel
Pareto Cauchy Log-gamma	Frechet
Uniforme Beta	Weibull

Por todo esto, nuestra distribución propuesta es Gumbel.

Calculamos en R los valores de la media μ, la varianza σ^2 y el desvío estándar σ de los máximos obtenidos:

- μ = 266.1470
- σ^2= 3848.7846
- σ = 62.0386

La estimación de los parámetros por el método de los momentos consiste en igualar los momentos poblacionales con los correspondientes momentos muestrales.

$E[x^k] = \frac{1}{n}\sum_{i=1}^{n} x_i$

Se igualan tantos momentos como parámetros haya que estimar.

En nuestro caso, tenemos una v.a con distribución Gumbel del máximo cuya función de densidad es la siguiente:

$f(x) = \frac{1}{\beta} e^{-(e^{-z}+z)}$, con $z = \frac{x-\theta}{\beta}$

Entonces, a partir del método de los momentos estimaremos los parámetros θ y β

$$E(x) = \mu = \theta + 0.5772157\,\beta$$

$$V(x) = \sigma^2 = \frac{\pi^2}{6}\beta^2$$

Nos queda entonces:

$$\overline{X} = \theta^* + 0.5772157\,\beta^*$$

$$S^2 = \frac{\pi^2}{6}\beta^{*2} \qquad con\ \beta > 0$$

Y despejamos los parámetros:

$$\theta^* = 238.226$$
$$\beta^* = 48.3713$$

2.7. Función de verosimilitud

La Estimación de Máxima Verosimilitud (EMV) es un modelo general para estimar parámetros de una distribución de probabilidad que depende de las observaciones de la muestra.

Cuando hablamos de estimación de máxima verosimilitud, debemos hablar de la función de máxima verosimilitud. Matemáticamente, dada una muestra x=(x1,…,xn) y parámetros, θ = (θ1, … , θn) entonces,

$$L(\theta/X) = \prod_{i=1}^{n} f(x_i;\theta)$$

Es decir que es la multiplicación de todas las funciones de densidad que dependen de las observaciones de la muestra (xi) y de los parámetros θ.

En este caso, haremos una transformación monótona mediante logaritmos, lo cual nos permite hacer un "cambio de escala" hacia números más pequeños. Entonces,

$$ln(L(\theta/X)) = \prod_{i=1}^{n} ln(f(x_i;\theta))$$

A partir de las fórmulas detalladas anteriormente calculamos la función densidad y posteriormente el log de la función de verosimilitud para cada distribución correspondiente a la investigación analizada anteriormente (Gumbel y Weibull) y la función original propuesta por la cátedra (Gamma). A continuación, se muestran los resultados obtenidos:

Distribución	**Ln(L)**
Gumbel	-118,7118528
Weibull	-121,4360686
Gamma	-196,9407011

Como se puede observar, la distribución más verosímil es la de Gumbel (Maximo) dado que su ln(L) es el mayor de las distribuciones propuestas.

3. Conclusiones

A partir de la muestra simulada, obtuvimos una curva de forma asimétrica positiva con una cola media concentrada a la derecha de la función densidad. Por otro lado, podemos decir que F(x) de la distribución Gamma se encuentra en el máximo dominio de atracción de G(x) si la distribución límite para los máximos es una distribución de Gumbel. Debido a todos estos motivos, anticipamos que la distribución para la muestra del valor extremo obtenido podría asemejarse con un Gumbel (Maximo).

Luego, al calcular los logaritmos de las funciones de verosimilitud, y compararlos, podemos concluir que la función de distribución Gumbel es la adecuada para predecir el comportamiento de los máximos de nuestra simulación, quedando verificada matemáticamente la hipótesis propuesta.

4. Anexo

4.1 Código

La simulación, así como los cálculos correspondientes a la manipulación de los valores de la misma fueron realizados en R, mientras que el cálculo de verosimilitudes en Excel. Ambos archivos quedan adjuntos al informe.

4.2 Bibliografía

Introducción a la Teoría de Valores Extremos - Centro de Investigación en Matemáticas, CIMAT
Trabajo de Investigación. Universidad de Granada. Análisis Estadístico de valores extremos y aplicaciones
Trabajo de Investigación. UNIVERSIDAD DE MURCIA FACULTAD DE MATEMÁTICAS. VALORES EXTREMOS Teoría y aplicaciones
Investigación. LA TEORÍA DEL VALOR EXTREMO: UNA APLICACIÓN AL SECTOR ASEGURADOR