A lo largo de la materia estudiamos distintos métodos para tratar una muestra de datos, enfocándonos en los valores centrales de la distribución de la misma, ignorando los valores extremos que esta pudiera tener y en muchos casos desechándolos. No obstante, en muchas disciplinas como la economía, la ingeniería o las ciencias ambientales, resulta importantísimo analizar estos valores extremos y los eventos que ellos producen por su alto impacto. Encontrar un análisis que lejos de ignorarlos se centre en describirlos, entonces, resulta vital.

Nos propusimos presentar la Teoría de Valores Extremos y trabajar un poco la problemática de su análisis. En particular, como en meteorología unas de las grandes divas para el estudio de eventos extremos son las olas de calor, decidimos visualizar este tema utilizando datos contenidos en un paquete de R llamado heatwaveR. Este dataset contiene las temperaturas mínimas y máximas diarias en grados Celsius registradas en Argel durante el período 01/01/1961 - 31/12/2005.

La ciudad de Argel es la capital de Argelia y se encuentra situada en el litoral mediterráneo del continente africano, en 36°46′35″N 3°03′31″E. Como esta ciudad se ubica en el hemisferio norte, particularmente en una zona subtropical y a orillas del mar Mediterráneo, su climatología es mediterránea marítima, con lo cual las temperaturas cumplen con un régimen normal de temperaturas más frías en los meses invernales de DEF y más cálidas en el verano de JJA, aunque moderadas por la influencia de este gran cuerpo de agua. Es por eso que en verano, gracias a las brisas que vienen desde el mar, las máximas no tiende a superar los 30°C; sin embargo, en esta ciudad se dan eventos de olas de calor, especialmente cuando soplan los vientos del Sur provenientes del desierto del Sahara, que son muy cálidos.

Del gráfico que muestra la serie de temperaturas máximas diarias para esta ciudad a lo largo de este periodo se observa que las mismas se encuentran entre los valores de 15°C y 30°C, con algunos casos anómalos por debajo y por arriba de estos valores. Uno se podría preguntar, entonces: ¿qué tan anómalos son estos valores, especialmente los más grandes? ¿Cómo los encuentro y analizo? ¿Qué distribución siguen los máximos? En un esfuerzo de explicar algunas de las interrogantes que nos puedan surgir después de ver esta marcha diaria de temperaturas máximas vamos, primero, a charlar un poco de algunos temas sobre teoría de valores extremos, vamos a plantear una serie de ejercicios para poner en práctica estas cuestiones y vamos, finalmente, a resolverlos juntos, explicando.

¡Empecemos! La Teoría de Valores Extremos es una rama de la estadística que se enfoca en el estudio de los eventos asociados a las colas de una distribución, tanto la correspondiente a los valores más altos como la de los más bajos de la variable aleatoria en estudio.

En la práctica existen dos aproximaciones a la Teoría de Valores Extremos: el primer método se basa en el ajuste de la distribución de los valores máximos o mínimos, mientras que en la segunda aproximación el análisis de los valores extremos se realiza a partir del análisis de los valores que exceden cierto umbral. Veamos cada uno por separado.

MODELO 1:

Según el Teorema de Fisher-Tippett-Gnedenko, hablando así entre nos, si y sólo si el máximo de una muestra de variables aleatorias independientes e idénticamente distribuidas converge, entonces lo hará hacia uno de tres tipos posibles de distribución: la distribución de Gumbel, la distribución de Fréchet o la distribución de Weibull.

Formalmente, el teorema dice:

Sea una sucesión de variables aleatorias independientes e idénticamente distribuidas provenientes de una función de distribución en común y sea donde cada representa un valor de un proceso medido en una escala temporal fija- por ejemplo temperaturas máximas o mínimas diarias- de modo que representa al máximo a lo largo de observaciones- en el ejemplo, si fuera igual a la cantidad de días que tiene un año, entonces representa la temperatura máxima anual. Si existe una sucesión de pares de números reales tal que para cada y donde es una función de distribución no degenerada[[1]](#footnote-1), entonces la distribución límite pertenece a alguna de estas tres familias que están acá:

1. Gumbel
2. Fréchet
3. Weibull

con los parámetros de posición b y de escala y en el caso de las familias 2 y 3, también el parámetro de forma.

Estas tres familias se agrupan bajo el nombre de Distribuciones para valores extremos, y tienen comportamientos distintos debidos a las diferentes conductas que toman sus colas: por ejemplo, en el caso de la distribución Weibull, tiene límite finito, mientras que para las distribuciones de Fréchet y Gumbel o mismo la densidad de F decae exponencialmente para la distribución de Gumbel, pero de forma polinómica en la de Fréchet.

Primariamente para un set de datos se podría adoptar alguna de estas tres familias y encontrar los parámetros correspondientes, pero esto no resulta útil porque habría que saber de antemano qué distribución es la apropiada para los datos y la realidad es que eso implica una suposición muy grande que a veces puede no ser acertada.

En cambio, se pueden reformular estos tres modelos, combinándolos en una sola familia de modelos llamada Generalized extreme value o GEV, como le diremos en adelante, tal que su función de distribución cumpla con esta forma

definida para estos donde los parámetros satisfacen esto y Los parámetros son de ubicación, escala y forma para y respectivamente. En particular, las distribuciones de las familias de Fréchet y Weibull se corresponden con para la primera y para la segunda, mientras que la de Gumbel se corresponde con el caso en que que se interpreta como el límite en que conduciendo a la distribución de la familia de Gumbel: , con .

La unificación de las tres familias de distribuciones de valores extremos en una sola gran familia simplifica la implementación estadística. A través de realizar inferencias sobre los mismísimos datos determinan el comportamiento más apropiado para la cola de la distribución, sin realizar juicios previos sobre cuál familia de distribuciones tomar por oportuna. Incluso la incertidumbre en el valor inferido para da una medida de la incertidumbre a la hora de elegir un tipo de familia por sobre los demás para un set de datos.

Teniendo esto en cuenta entonces el Teorema de Fisher-Tippett-Gnedenko podemos reformularlo de la siguiente manera: si existe una sucesión de pares de números reales tal que para cada y donde es una función de distribución no degenerada[[2]](#footnote-2), entonces la distribución límite pertenece a la familia GEV:

definida para donde los parámetros satisfacen que y

Para modelar extremos de una serie de observaciones independientes los datos son separados en bloques de secuencias de observaciones de longitud , con valores grandes de generando así una serie de bloques de máximos para la cual se ajusta la distribución GEV. Los bloques de datos se eligen tal que se correspondan a un período de tiempos, por ejemplo un año, en cuyo caso el es el número de observaciones en un año y los máximos se corresponden con máximos anuales.

¿Se acuerdan de nuestra serie de temperaturas máximas diarias? Teníamos 16436 observaciones. Si yo extrajera de ese set sólo el máximo de cada año de ese período me quedaría solamente con 45. Ahora, lo que dice el teorema es que a estos datos los puedo caracterizar o aproximar a una distribución que tiene nombre y apellido: GEV. Vamos a ver cómo más adelante.

MODELO 2:

Sea una sucesión de variables aleatorias independientes e idénticamente distribuidas, con distribución marginal F. Algunos valores de dicha sucesión, lógicamente, van a exceder cierto umbral- en el caso del análisis de extremos, un umbral particularmente alto. El comportamiento de los eventos extremos para dicha sucesión está dado por la siguiente probabilidad condicional:

Si la distribución F fuera conocida, entonces la distribución para los excesos (es decir, la distribución de aquellos valores que exceden el umbral ) también sería conocida; pero como no suele ser el caso, se realiza un tratamiento similar al que hicimos antes con GEV. Atentos a este teorema.

Sea una sucesión de variables aleatorias independientes e idénticamente distribuidas proveniente de una distribución , y sea . Si para grandes vale que donde

para y Entonces, para un umbral la función de distribución de condicionada a que es aproximadamente

definida para y tal que . H recibe el nombre de Distribución de Pareto generalizada.

De este modo, si el bloque de máximos tiene una distribución aproximada G, entonces la distribución de los excesos pueden ser caracterizada a través de la Distribución de Pareto generalizada. No sólo eso, los parámetros para la distribución Pareto generalizada están determinados por los de la distribución GEV de bloques máximos. Particularmente, el parámetro ξ es el mismo en ambas. Si se eligen bloques de diferentes tamaños, los valores de los parámetros para la GEV cambian, pero esto no sucede en el caso de la de Pareto: ξ permanece invariante, y no se modifica tampoco porque los cambios en μ y σ se compensan mutuamente. Resulta entonces que el parámetro de forma ξ es el dominante a la hora de determinar el comportamiento cualitativo tanto de la GEV como de la distribución Pareto generalizada: si la distribución de excesos tiene un límite superior ; si entonces no tiene límite superior. Si , la distribución tampoco tiene límite, pero debería ser reinterpretada tomando el límite llegando a

una distribución exponencial de parámetro .

Volviendo a la serie de temperaturas máximas diarias, recordemos que teníamos 16436 observaciones. Supongamos entonces que pusiera como umbral . Voy a querer quedarme, entonces, sólo con las temperaturas que excedan este valor. Veo que termino teniendo muchísimos más valores que con la metodología anterior. Esta es la principal ventaja de este método frente al otro: block-máxima, al obligar a quedarte con un solo valor máximo anual, no permite tener más de un valor extremo que merecería tener en cuenta para el análisis y que correspondan al mismo bloque. Por ejemplo, si tuviera una ola de calor con temperaturas superiores a 38 grados de 7 días de duración, sólo tendría información del día más cálido de todos (supongamos días de 38 grados, y un día de calor máximo con 40), pero es un evento largo e importante. Ahora, lo que dice el teorema es que a estos datos los puedo caracterizar o aproximar a una distribución que tiene nombre y apellido: GPD. Vamos a ver cómo más adelante.

NIVELES DE RETORNO

El nivel de retorno para series estacionarias con distribución es esencialmente un percentil, pero no cualquier percentil: el percentil de la distribución, es decir que es el valor que satisface .

Entonces veamos cómo escribir estos niveles de retorno con la metodología de block-máxima. Los estimadores de los cuantiles para extremos en una distribución anual de máximos se obtienen invirtiendo las ecuaciones

donde , siendo el nivel de retorno asociado al período de retorno tal que éste sea superado en promedio una vez cada años. En otras palabras, es excedido por el máximo anual de cualquier año particular con una probabilidad En particular, definiendo tal que

se tiene que si se realizara un gráfico vs tomando una escala logarítmica, el gráfico resulta linear cuando Si el gráfico es convexo con un límite asintótico cuando en ; en cambio, si el gráfico es cóncavo y tiende a infinito. Pero vamos a limitarnos al caso en que , por un tema de simpleza. Este gráfico es un gráfico de nivel de retorno.

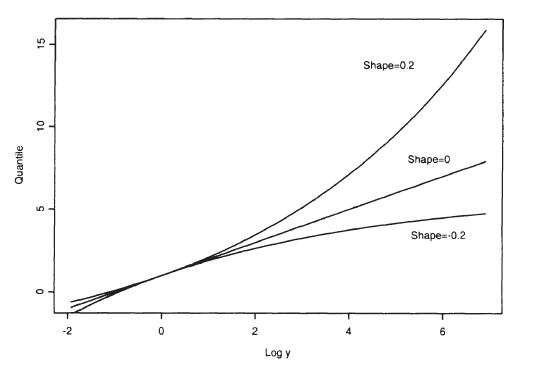


Figura 1: Gráfico de nivel de retorno para una distribución GEV con parámetros de forma ξ=-0.2, ξ=0 y ξ=0.2 respectivamente[[3]](#footnote-3).

Ahora vamos con la metodología del umbral. Suponiendo que una distribución Pareto generalizada con parámetros σ y ξ es una buena forma de modelar a los excesos a través del uso de un umbral para la variable Es decir, para

si , se obtiene

Es decir, el nivel de retorno que se excede en promedio una vez cada observaciones es la solución de la ecuación

De modo que

con un lo suficientemente grande para asegurar que Notar que esta solución para el nivel de retorno asume que . En el caso en que se tiene que

Si se realiza un gráfico de vs con una escala logarítmica se tienen resultados similares que los gráficos de niveles de retorno para el modelo GEV: si se tiene un gráfico lineal; se tiene concavidad si y convexidad si

ERROR ESTÁNDAR E INTERVALOS DE CONFIANZA

Nos enfrentamos entonces al problema de querer estimar un percentil que se encuentre en la cola de la distribución.

Entonces tengamos en cuenta por ejemplo la distribución de Gumbel. El nivel de retorno es este que está acá.

Entonces tenemos el estimador de xp a partir de estimar los parámetros de la distribución: mu y sigma, porque xp seguía una fórmula como la que vimos anteriormente. Es decir, que si conocemos a mu y sigma, también conocemos a xp. Pero conocer a xp de este modo tiene un costo. Entonces tenemos que hablar de errores de estimación y de intervalos de confianza.

Como la estimación de xp depende de las estimaciones de mu y sigma, tenemos una expresión que relaciona la varianza del estimador xp con la matriz de varianzas-covarianzas de las estimaciones de los parámetros mu y sigma, la cual es esta que tenemos acá.

Donde el tercer multiplicando es el vector de derivadas parciales de los estimadores para mu y sigma con respecto a cada una

Y V es la matriz de 2x2 de varianza covarianza de los estimadores de mu y sigma.

Ahora bien, en [estadística](https://es.wikipedia.org/wiki/Estad%C3%ADstica), se llama intervalo de confianza a un par o varios pares de números entre los cuales se estima que estará cierto valor desconocido con un determinado nivel de confianza. Esto es algo que ya vimos a lo largo de la materia pero es algo no muy explorado para niveles de retorno.

Tenemos un valor estimado, y tenemos un error en la estimación. Ya tenemos todo para construir el intervalo de confianza que nos permita no solo tener una estimación del percentil, sino, además, una medida de cuánto puede variar (para alguna probabilidad).

Un intervalo de confianza está dado de manera general, como vimos en la materia, por:

En particular, el error de un intervalo de confianza asintótico se puede escribir como

donde el primero es el percentil de una normal 0,1 y el segundo la desviación estándar del estimador, que es la raíz de la varianza del estimador, lo que obtuvimos arriba al relacional el gradiente con la matriz de varianza covarianza

para este percentil sp elegimos el alpha del nivel del intervalo:

Para un intervalo de 95% de confianza

Queremos el valor del eje x de una Normal 0,1 para el cual queda 0.025 de área a su derecha.

De forma que ya podemos armar un intervalo de confianza de nivel 0.95 , que no solo nos dice un valor puntual de la estimación del percentil z\_p de la Gumbell, sino que también nos dice qué tanto puede alejarse de este valor, si exijo una confianza en el resultado de 95%.

Con ésto logramos una interpretación más informativa del resultado, que nos da un rango de valores que podrían suceder si exigimos tal confianza.

Y ahora sí, llegamos al final del video. Esperamos que haya sido claro y nos vemos en el próximo, en que veremos cómo resolver la guía de ejercicios que preparamos para mojar un poco los pies con valores extremos.

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. Hoja 50 Coles&Tenner An introduction to statistical modeling of extreme values [↑](#footnote-ref-3)