

Working Paper n° 26

Estatísticas Ordinais

Luisa Canto e Castro Duarte

Working Paper n° 26

ISSN: 0872-895X

Depósito Legal n°: 88052/95

Luisa Canto e Castro Duarte
DEIO, Universidade de Lisboa

Março 1995

Trabalho apresentado no âmbito do Seminário de Estatística - coordenado pelo
Professor Doutor Bento Murteira

ESTATÍSTICAS ORDINAIS

Luisa Canto e Castro Duarte

Departamento de Estatística e Investigação Operacional
da Universidade de Lisboa

e

Centro de Estatística e Aplicações
da Universidade de Lisboa

Resumo: São inúmeros os problemas de análise estatística que envolvem de alguma forma as chamadas estatísticas de ordem ou estatísticas ordinais. Remetendo para os muitos títulos já existentes nesta área o leitor interessado em aprofundar este tema, apresentam-se aqui apenas alguns dos resultados fundamentais da teoria distribucional exacta (distribuições marginais e distribuições conjuntas, representação de Rényi e Teorema de Malmquist) e da teoria distribucional assintótica (estatísticas ordinais extremas e estatísticas ordinais centrais, Teorema dos Tipos Extremas e normalidade assintótica dos quantis empíricos).

Palavras Chave: Estatística Ordinal, Distribuição Limite, Valores Extremos.

Abstract: Order Statistics are of great importance in many areas of statistical theory and practice. This subject has experienced a rapid growth in recent years. Among so many relevant results, we choose to present a few concerning the basic distribution theory (marginal distributions, joint distributions, Rényi representation and Malmquist Theorem) and the asymptotic theory (extreme order statistics and central order statistics, Extremal Types Theorem and asymptotic normality of sample quantiles).

Keywords: Order Statistic, Asymptotic Distribution, Extreme Values.

1. Introdução

Dada uma amostra aleatória de dimensão n , $(X_1, X_2, \dots, X_n)$, designamos por estatística ordinal de ordem k , o elemento da amostra que ocupa a posição k na ordenação ascendente dessa amostra, e representamo-lo por $X_{k:n}$ (ou por $X_{(k)}$ quando a dimensão da amostra estiver subentendida). Tem-se pois

$$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$$

O conhecimento da teoria distribucional exacta (e assintótica) das estatísticas ordinais associadas aos mais diversos modelos tem sido de primordial importância em muitas áreas da estatística teórica e aplicada. Basta notar que, tanto a mediana amostral como outros quantis são exemplos de estatísticas ordinais. As metodologias robustas, tão em voga nos tempos recentes, consistem, essencialmente, em eliminar valores extremos de forma a aumentar a robustez e resistência dos estimadores.

Na análise de amostras censuradas (quer a censura seja de tipo I ou de tipo II) do que dispomos, na realidade, é de uma amostra de N estatísticas ordinais inferiores (associadas à amostra dos n elementos colocados inicialmente em estudo).

Entre muitas outras aplicações das estatísticas ordinais, refira-se ainda a importância crucial que os *Valores Extremos* (máximos e mínimos) têm na interpretação e previsão das mais diversas "calamidades" (temperaturas mínimas ou máximas, cheias, ciclones, tempos de vida mínimos em teoria da fiabilidade, etc.). O crescente interesse e o vasto leque de aplicabilidade deste tipo de estatísticas justificaram o aparecimento de alguns títulos específicos no tema, dos quais se podem destacar os de David (1981), Sarhan e Greenberg (1962) e Reiss (1989).

2. Teoria distribucional exacta

Seja $X_{k:n}$ a k -ésima estatística ordinal associada a uma amostra aleatória de dimensão n . Represente-se por $F_{k:n}(x)$ a sua função de distribuição (f.d.). Tem-se

$$F_{k:n}(x) = P[X_{k:n} \leq x] = P[\text{pelo menos } k \text{ dos } X_i\text{'s serem inferiores a } x].$$

No caso em que a amostra inicial $(X_1, X_2, \dots, X_n)$ é constituída por variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), a probabilidade acima é de cálculo bastante imediato. Basta notar que, se considerarmos como sucesso o facto de se ter $X_i \leq x$, então o número de elementos da amostra inferiores a x terá uma distribuição Binomial($n, F(x)$), onde $F(x)$ é a f.d. subjacente à amostra. Logo

$$F_{k:n}(x) = \sum_{j=k}^n \binom{n}{j} F^j(x) (1 - F(x))^{n-j}, 1 \leq k \leq n. \quad (1)$$

Como casos particulares temos a distribuição do *mínimo* em amostras i.i.d.,

$$\begin{aligned} F_{1:n}(x) &= \sum_{j=1}^n \binom{n}{j} F^j(x) (1 - F(x))^{n-j} \\ &= 1 - \binom{n}{0} F^0(x) (1 - F(x))^n \\ &= 1 - (1 - F(x))^n \end{aligned}$$

e a distribuição do *máximo* em amostras i.i.d.,

$$F_{n:n}(x) = F^n(x)$$

Facilmente se demonstra (basta integrar por partes) que a expressão (1) admite a seguinte representação à custa da f.d. da Beta incompleta

$$F_{k:n}(x) = \underbrace{\frac{1}{B(k, n-k+1)} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} dt}_{I_{F(x)}(k, n-k+1)} \quad (2)$$

onde $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Esta representação permite-nos obter os percentis de $X_{k:n}$ por consulta de tabelas da beta incompleta (Pearson e Hartley (1970). *Biometrika tables* - tabela 16). É interessante notar que a identificação da distribuição das estatísticas ordinais surge naturalmente através da sua f.d. (mesmo no caso discreto). Aliás, para modelos discretos, a função massa de probabilidade de $X_{k:n}$ não tem expressão explícita

$$\begin{aligned} f_{k:n}(x) &= P[X_{k:n} = x] = F_{k:n}(x) - F_{k:n}(x-1) \\ &= I_{F(x)}(k, n-k+1) - I_{F(x-1)}(k, n-k+1) \end{aligned}$$

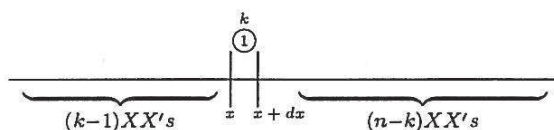
No caso contínuo, sendo $f(x)$ a função densidade de probabilidade (f.d.p.) de X e $f_{k:n}(x)$ a f.d.p. de $X_{k:n}$, obtém-se

$$f_{k:n}(x) = \frac{1}{B(k, n-k+1)} F^{k-1}(x) (1-F(x))^{n-k} f(x) \quad (3)$$

por derivação do integral que surge em (2).

Poderíamos chegar a esta mesma expressão por um cálculo probabilístico directo, o qual se irá revelar muito útil na dedução da densidade conjunta de duas ou mais estatísticas ordinais.

A ideia é começar por calcular $P[x < X_{k:n} \leq x + dx]$ e obter a partir desta a f.d.p. por passagem ao limite depois de dividida por dx .



De acordo com o diagrama acima, o acontecimento $X_{k:n} \in]x, x + dx]$ ocorre sse

- 1) $(k-1)$ dos n elementos da amostra forem não superiores a x .

Probabilidade associada: $\binom{n}{k-1} F^{k-1}(x)$.

- 2) dos $(n-k+1)$ restantes, um deles, arbitrariamente, deverá estar em $]x, x + dx]$.

Probabilidade associada: $\binom{n-k+1}{1} [F(x+dx) - F(x)]$.

- 3) Os restantes $(n-k)$ elementos deverão ser superiores a $x + dx$.

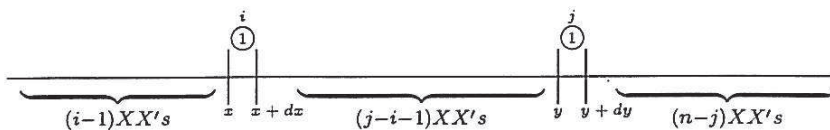
Probabilidade associada: $[1 - F(x+dx)]^{n-k}$.

Logo

$$P[x < X_{k:n} \leq x + dx] = \frac{n!}{(k-1)!(n-k)!} F^{k-1}(x) [F(x+dx) - F(x)] [1 - F(x+dx)]^{n-k}$$

resultando daqui, de imediato a expressão indicada para a densidade de $X_{k:n}$.

A partir de um raciocínio análogo para o cálculo da probabilidade do acontecimento $\{x < X_{i:n} \leq x + dx, y < X_{j:n} \leq y + dy\}$, com $1 \leq i < j \leq n$, $x < y$,



obtem-se a seguinte expressão para a f.d.p. conjunta do par $(X_{i:n}, X_{j:n})$

$$f_{i,j:n}(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F^{i-1}(x) [F(y) - F(x)]^{j-i-1} [1 - F(y)]^{n-j} f(x) f(y)$$

para $x < y$.

A generalização a qualquer número de estatísticas ordinais é agora um problema de solução quase imediata, vindo

$$f_{n_1, n_2, \dots, n_k:n}(x_1, x_2, \dots, x_k) = n! \prod_{j=0}^k \frac{[F(x_{j+1}) - F(x_j)]^{n_{j+1} - n_j - 1}}{(n_{j+1} - n_j - 1)!} \prod_{j=0}^k f(x_j) \quad (4)$$

para $x_1 < x_2 < \dots < x_k$, onde se convencionou que $x_0 = -\infty$, $x_{k+1} = +\infty$, $n_0 = 0$ e $n_{k+1} = n + 1$.

De entre estas f.d.p. conjuntas tem particular interesse a de $(X_{1:n}, \dots, X_{n:n})$, a qual vem dada por

$$f_{1,2,\dots,n:n}(x_1, x_2, \dots, x_n) = n! \prod_{j=0}^n f(x_j)$$

para $x_1 < x_2 < \dots < x_n$.

De (4) resulta ainda uma interessante propriedade para as estatísticas ordinais associadas a uma amostra i.i.d. de uma população com f.d. absolutamente contínua.

Na realidade, efectuando algumas simplificações convenientes, mostra-se facilmente que a f.d.p. condicional de $X_{j:n} | X_{1:n} = x_1, X_{2:n} = x_2, \dots, X_{i:n} = x_i$, para $j > i$, coincide com a f.d.p. condicional de $X_{j:n} | X_{i:n} = x_i$, ou seja, é válida a propriedade Markoviana para esta sequência de variáveis. Para $x \leq y$ esta última f.d.p. é dada por

$$f_{X_{j:n} | X_{i:n}}(y|x) = \frac{(n-i)!}{(j-i-1)!(n-j)!} \frac{[F(y) - F(x)]^{j-i-1} [1 - F(y)]^{n-j} f(y)}{[1 - F(y)]^{n-i}}$$

É claro que, como de costume, o conhecimento da estrutura conjunta de um vector aleatório, tem como principal vantagem a possibilidade de utilização dos métodos standard para transformações os quais nos permitem derivar a f.d.p. de qualquer função ("bem comportada") desse vector. É, pois, só uma questão de cálculos, a obtenção da f.d.p. de muitas das estatísticas elementares (mediana, amplitude amostral, trimédia, cociente extremal, etc).

Exemplo 1: Distribuição das e.o.'s e da amplitude amostral para o modelo uniforme em $[0, 1]$.

Represente-se por $U_{1:n}, U_{2:n}, \dots, U_{n:n}$ as estatísticas ordinais associadas a uma amostra de v.a.'s i.i.d. $U[0, 1]$.

Recorrendo à expressão geral indicada em (3), começamos por obter

$$f_{k:n}(u) = \frac{1}{B(k, n-k+1)} u^{k-1} (1-u)^{n-k} \quad 0 \leq u \leq 1$$

Logo $U_{k:n}$ é uma variável aleatória Beta($k, n-k+1$). Quanto à densidade conjunta de $(U_{i:n}, U_{j:n})$, tem-se

$$f_{i,j:n}(u, v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} (v-u)^{j-i-1} (1-v)^{n-j} \quad 0 \leq u \leq v \leq 1.$$

Para identificar a distribuição da amplitude amostral, $W = U_{n:n} - U_{1:n}$, considere-se a transformação

$$\begin{cases} X &= U_{1:n} \\ W &= U_{n:n} - U_{1:n} \end{cases}$$

cujo jacobiano é $|J| = 1$, vindo, por isso,

$$f_{X,W}(x, w) = f_{1,n,n}(x, x+w) \quad 0 \leq w \leq 1, \quad 0 \leq x \leq 1-w.$$

Calculando a marginal pretendida, obtem-se, para $0 \leq w \leq 1$,

$$f_W(w) = \int_0^{1-w} \frac{n!}{(n-2)!} w^{n-2} dx = \frac{1}{B(n-1, 2)} w^{n-2} (1-w)$$

donde se conclui que a amplitude de uma amostra de uniformes em $[0, 1]$ segue uma distribuição Beta($n-1, 2$).

Exemplo 2: Independência dos *spacings* em amostras de exponenciais.
Seja $(T_1, T_2, \dots, T_n)$ uma amostra aleatória de exponenciais (λ) , isto é, com f.d.p.,

$$f_T(x) = \lambda e^{-\lambda x} \quad , \quad x \geq 0.$$

Como é sabido, a “falta de memória”, traduzida através de

$$P[T < x + y | T \geq y] = P[T < x] \quad , \quad x > 0, y > 0$$

é uma das propriedades que caracteriza a exponencial.

Suponhamos agora que, para facilitar a interpretação das respectivas estatísticas ordinais, $(T_1, T_2, \dots, T_n)$ representa os tempos de vida de n máquinas postas a funcionar em simultâneo, pelo que $T_{k:n}$ representará o tempo que decorre até à k -ésima falha.

Às diferenças $T_{k:n} - T_{k-1:n}$ com $1 \leq k \leq n$, é usual chamar “spacings” entre as estatísticas ordinais, os quais, neste caso concreto, representarão os intervalos de tempo entre falhas consecutivas. (Considera-se $T_{0:n} = 0$).

Ora

$$P[T_{k:n} - T_{k-1:n} > x | T_{k-1:n} = y] = P[T_{k:n} > x + y | T_{k-1:n} = y]$$

$$= P[\text{todas as } (n - k + 1) \text{ máquinas que ainda estão em funcionamento no instante } y, \text{ durarem, pelo menos, mais } x \text{ unidades de tempo}].$$

Dada a “não memória” da exponencial, teremos, por isso,

$$P[T_{k:n} - T_{k-1:n} > x | T_{k-1:n} = y] = \{P[T > x]\}^{n-k+1} = e^{-\lambda(n-k+1)x}$$

para qualquer $x > 0$ e independentemente de y . Conclui-se assim que os spacings de exponenciais são independentes entre si e têm distribuição exponencial de parâmetro $(n - k + 1)\lambda$.

O resultado estabelecido neste exemplo serve de base a uma importante representação das estatísticas ordinais à custa de uma combinação linear de exponenciais independentes.

Representação de Rényi

Seja $T_{k:n}$ a k -ésima estatística ordinal de uma amostra i.i.d. com distribuição exponencial unitária. Vimos no exemplo anterior que as variáveis

$$V_k = (n - k + 1)(T_{k:n} - T_{k-1:n}) \quad , \quad k = 1, 2, \dots, n$$

são i.i.d. e têm, também, distribuição exponencial de parâmetro 1.

A partir daqui, facilmente se verifica que

$$T_{k:n} \stackrel{d}{=} \frac{V_1}{n} + \frac{V_2}{n-1} + \dots + \frac{V_k}{n-k+1} \quad , \quad 1 \leq k \leq n. \quad (5)$$

Para se passar ao caso geral basta lembrar que, sendo X uma v.a. com f.d. F , contínua e estritamente crescente, então $F(X)$ é uniforme em $[0, 1]$ e que, sendo U uma v.a. uniforme em $[0, 1]$, então $-\ln U$ é exponencial unitária. Seja então $(X_1, X_2, \dots, X_n)$ uma amostra i.i.d. com X . Note-se que

$$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$$

$$\Downarrow$$

$$-\ln F(X_{1:n}) \geq -\ln F(X_{2:n}) \geq \dots \geq -\ln F(X_{n:n})$$

pelo que, tendo em conta o que acima se disse,

$$-\ln F(X_{k:n}) \stackrel{d}{=} T_{n-k+1:n}.$$

De (5) seguir-se-á então finalmente a seguinte representação para $X_{k:n}$

$$X_{k:n} = F^{-1} \left(e^{-T_{n-k+1:n}} \right) = F^{-1} \left(e^{-\left(\frac{V_1}{n} + \frac{V_2}{n-1} + \dots + \frac{V_{n-k+1}}{k} \right)} \right).$$

Teorema de Malmquist

Nas condições de validade da representação de Rényi, tem-se que

$$U_{n-k+1} = \left(\frac{F(X_{k:n})}{F(X_{k+1:n})} \right)^k \quad , \quad k = 1, 2, \dots, n \quad (6)$$

são variáveis aleatórias mutuamente independentes e uniformes em $[0, 1]$. (Nota: $F(X_{n+1:n}) = 1$).

Demonstração :

Tendo em conta que

$$-\ln \frac{F(X_{k:n})}{F(X_{k+1:n})} \stackrel{d}{=} T_{n-k+1:n} - T_{n-k:n} \stackrel{d}{=} \frac{V_{n-k+1}}{k}$$

o resultado segue-se de imediato, pois poderemos escrever

$$\left(\frac{F(X_{k:n})}{F(X_{k+1:n})} \right)^k \stackrel{d}{=} e^{-V_{n-k+1}}$$

e já vimos que as v.a.'s $\{V_k\}$, $1 \leq k \leq n$ são i.i.d. exponenciais unitárias.

Vejamos agora de que modo se pode utilizar o teorema de Malmquist na construção de um algoritmo simples para simulação de estatísticas ordinais de ordem superior.

Seja $Y_{k:n} = F(X_{k:n})$ que, como se sabe, tem a distribuição da k -ésima estatística ordinal de uma amostra de uniformes em $[0, 1]$. Começando por fazer $k = n$ em (6) tem-se

$$Y_{n:n} = U_1^{\frac{1}{n}} \quad \text{com} \quad U_1 \sim U[0, 1].$$

Para $k = n - 1$, virá

$$\frac{Y_{n-1:n}}{Y_{n:n}} = U_2^{\frac{1}{n-1}} \quad \text{com} \quad U_2 \sim U[0, 1] \text{ e indep. de } U_1$$

pelo que

$$Y_{n-1:n} = U_1^{\frac{1}{n}} U_2^{\frac{1}{n-1}}.$$

Mais genericamente, ter-se-á

$$Y_{n-k:n} = U_1^{\frac{1}{n}} U_2^{\frac{1}{n-1}} \dots U_{k+1}^{\frac{1}{n-k}}$$

$$X_{n-k:n} = F^{-1}(Y_{n-k:n}).$$

Embora este resultado seja válido para qualquer k , o algoritmo só se revela eficiente na simulação de estatísticas ordinais de ordem superior (ou inferior), isto é, para k fixo, pequeno.

3. Teoria distribucional assintótica

A obtenção de distribuições limites não degeneradas para determinadas sucessões de variáveis aleatórias (de que o teorema Limite Central é exemplo típico) tem tido, desde sempre, um papel de extrema utilidade em todos os ramos da análise e modelação estatística.

Ao considerar sucessões de estatísticas ordinais, iremos ter basicamente duas situações distintas: ou a dimensão da amostra cresce para infinito mas k mantém-se fixo ou, pelo contrário, k varia com n e acompanha o seu crescimento.

Nesta secção incluem-se alguns resultados que ilustram estas duas situações, embora se deva referir que eles constituem apenas uma parcela ínfima do muito que há feito nesta área.

Tal como na teoria distribucional exacta, partindo dos modelos exponencial e uniforme é possível obter resultados assintóticos válidos para sucessões de estatísticas ordinais em modelos mais gerais.

Proposição 1

Seja $T_{k:n}$ a k -ésima estatística ordinal de uma amostra de exponenciais unitárias (k fixo). Então

$$\lim_{n \rightarrow \infty} P[nT_{k:n} \leq x] = \int_0^x \frac{t^{k-1} e^{-t}}{\Gamma(k)} dt, \quad x > 0,$$

isto é, $nT_{k:n}$ tem, assintoticamente, uma distribuição Gama($k, 1$).

Demonstração :

Recorrendo à expressão da densidade de $T_{k:n}$ (ver (3)), obtem-se

$$f_{T_{k:n}}(x) = k \binom{n}{k} (1 - e^{-x})^{k-1} e^{-(n-k)x} e^{-x}, \quad x \geq 0.$$

Logo

$$f_{nT_{k:n}}(x) = \frac{1}{n} f_{T_{k:n}}\left(\frac{x}{n}\right)$$

$$\begin{aligned}
&= \binom{n-1}{k-1} e^x (e^{x/n} - 1)^{k-1} \\
&= \frac{e^x}{(k-1)!} \prod_{j=1}^{k-1} (n-j) (e^{x/n} - 1).
\end{aligned}$$

Para cada j fixo, o factor $(n-j)(e^{x/n} - 1)$ converge para x quando n tende para infinito, e, como tal

$$f_{nT_{k:n}}(x) \xrightarrow{n \rightarrow \infty} \frac{x^{k-1}}{\Gamma(k)} e^{-x}, \quad x \geq 0.$$

Isto completa a demonstração uma vez que, pelo teorema de Scheffé, a convergência das funções de densidade implica a convergência das respectivas funções de distribuição (note-se que a recíproca não é válida...).

Este mesmo teorema poderia ser demonstrado utilizando resultados conhecidos sobre tipos de convergência de variáveis aleatórias.

Mais precisamente, se considerarmos a representação (5), iremos obter sucessivamente

$$T_{k:n} = \sum_{j=1}^k \frac{V_j}{n-j+1}$$

$\Downarrow$

$$\begin{aligned}
nT_{k:n} &= \sum_{j=1}^k \left(1 + \frac{j-1}{n-j+1}\right) V_j \\
&= \sum_{j=1}^k V_j + \sum_{j=1}^k \frac{j-1}{n-j+1} V_j
\end{aligned}$$

A segunda parcela do segundo membro da última expressão converge em probabilidade para zero, quando $n \rightarrow \infty$, visto tratar-se de uma soma finita de v.a.'s que tendem em probabilidade para zero, e a primeira parcela tem distribuição Gama($k, 1$) por se tratar de uma soma de exponenciais unitárias e independentes.

Proposição 2

Se $\{U_j\}_{j \geq 1}$ é uma sucessão de v.a.'s i.i.d. $U[0, 1]$, então, $\forall k \geq 1$

$$nU_{k:n} \xrightarrow{\mathcal{D}} \text{Gama}(k, 1)$$

e

$$n(1 - U_{n-k+1:n}) \xrightarrow{\mathcal{D}} \text{Gama}(k, 1)$$

Demonstração :

Recorde-se que $U_{n-k+1:n} \stackrel{d}{=} e^{-T_{k:n}}$. Utilizando o desenvolvimento em série da exponencial, tem-se

$$1 - U_{n-k+1:n} = T_{k:n} + \sum_{j \geq 2} \frac{(T_{k:n})^j}{j!} (-1)^{j-1}$$

ou seja

$$n(1 - U_{n-k+1:n}) = nT_{k:n} + Z_n$$

onde $Z_n = \sum_{j \geq 2} nT_{k:n} \frac{(T_{k:n})^{j-1}}{j!} (-1)^{j-1}$ converge em probabilidade para zero, quando $n \rightarrow \infty$. Logo, a distribuição assintótica de $n(1 - U_{n-k+1:n})$ é idêntica à de $nT_{k:n}$. Por outro lado, como $U_j \stackrel{d}{=} 1 - U_j$, tem-se, obviamente, $U_{k:n} \stackrel{d}{=} 1 - U_{n-k+1:n}$ o que conclui a demonstração.

Seria interessante agora, pegar nos resultados obtidos na proposição anterior e, seguindo uma estratégia análoga à utilizada na teoria distribucional exacta, obter o comportamento assintótico das estatísticas ordinais associadas a modelos mais genéricos. No entanto, fazendo as substituições convenientes, deduz-se apenas que

$$\lim_{n \rightarrow \infty} P \left[X_{k:n} \leq F^{-1} \left(\frac{x}{n} \right) \right] = \int_0^x \frac{t^{k-1} e^{-t}}{\Gamma(k)} dt$$

ou

$$\lim_{n \rightarrow \infty} P \left[X_{n-k+1:n} \geq F^{-1} \left(1 - \frac{x}{n} \right) \right] = \int_0^x \frac{t^{k-1} e^{-t}}{\Gamma(k)} dt$$

e daqui não se pode concluir que existam constantes normalizadoras $a_n \in \mathbb{R}_+$ e $b_n \in \mathbb{R}$ tais que $(X_{k:n} - b_n)/a_n$ convirja para alguma lei não degenerada. Esta questão mereceu de muitos investigadores uma atenção mais

aprofundada, podendo salientar-se, entre muitas outras, as contribuições de von Mises (1923, 1936), Fisher e Tippet (1928), Gnedenko (1943) e de Haan (1970). Uma vez que k está fixo, quando n tende para infinito, $X_{k:n}$ ($X_{n-k+1:n}$) corresponderá a uma estatística ordinal extremal mínima (máxima). É esta a razão porque é usual designar esta área de investigação em estatística como "*Teoria de Extremos*".

Tradicionalmente a teoria dos valores extremos foi desenvolvida para máximos mas os resultados são facilmente adaptáveis a mínimos tendo em conta que $\min\{X_1, X_2, \dots, X_n\} = -\max\{-X_1, -X_2, \dots, -X_n\}$. Referiremos aqui unicamente o principal resultado de teoria de extremos, o qual desempenha para a modelação de máximos um papel análogo ao que o Teorema Limite Central desempenha para as somas.

Teorema dos Tipos Extremais (Gnedenko (1943))

Seja $\{X_i\}_{i \geq 1}$ uma sucessão de v.a.'s i.i.d..

Se existirem sucessões de constantes $a_n \in \mathbb{R}_+$ e $b_n \in \mathbb{R}$ tal que

$$P \left[\frac{X_{n:n} - b_n}{a_n} \leq x \right] \xrightarrow{n \rightarrow \infty} G(x) \quad (7)$$

com $G(x)$ f.d. não degenerada, então G só poderá ser de um dos seguintes três tipos

Tipo I (Gumbel)

$$\Lambda(x) = \exp(-e^{-x}) \quad , \quad x \in \mathbb{R}$$

Tipo II (Fréchet)

$$\Phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0 \end{cases} \quad , \quad \alpha > 0.$$

Tipo III (Weibull)

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{-\alpha}) & x \leq 0 \\ 1 & x > 0 \end{cases} \quad , \quad \alpha > 0.$$

As condições necessárias para que se tenha (7) e para que G seja de um determinado tipo de entre os três acima indicados, envolvem, essencialmente,

o comportamento de cauda da função de distribuição de X .

Ao leitor interessado, aconselha-se a obra de Leadbetter, Lindgren e Rootzén (1983) por constituir um tratado bastante exaustivo e aprofundado da teoria de valores extremos em modelos estacionários.

Mencionou-se anteriormente o facto de, quando k está fixo e n cresce, as estatísticas ordinais $X_{k:n}$ e $X_{n-k+1:n}$ corresponderem a *estatísticas ordinais extremas*. Em contraponto com esta situação, admita-se agora que k cresce com n , isto é, que $k = k(n)$ e $k(n) \xrightarrow{n \rightarrow \infty} \infty$. Surgem assim as *estatísticas ordinais centrais*.

Caso particular importante é $k = [np] + 1$ com $0 < p < 1$, ou seja, o caso em que $X_{k:n}$ é o quantil empírico de ordem p .

O comportamento assintótico das estatísticas ordinais centrais (e dos quantis empíricos em particular) está caracterizado no seguinte teorema:

Teorema de Rényi

Seja $\{X_i\}_{i \geq 1}$ uma sucessão de v.a.'s i.i.d. com f.d. F absolutamente contínua. Seja $f(x)$ a correspondente f.d.p. e admita-se que $f(x)$ é contínua e positiva no intervalo $]a, b[$. Então, sendo $p \in]F(a), F(b)[$ e $k = k(n)$ tal que $|k(n) - np| = o(\sqrt{n})$, tem-se

$$\frac{X_{k:n} - F^{-1}(p)}{\sqrt{\frac{p(1-p)}{n}}} f(F^{-1}(p)) \xrightarrow{D} Z \sim \mathcal{N}(0, 1)$$

isto é

$$X_{k:n} \rightsquigarrow \mathcal{N}\left(F^{-1}(p), \sqrt{\frac{p(1-p)}{n}} \cdot \frac{1}{f(F^{-1}(p))}\right).$$

Demonstração : (esboço)

Mais uma vez se terá em conta que $X_{k:n} \stackrel{d}{=} F^{-1}(U_{k:n})$ e que $U_{k:n} \stackrel{d}{=} e^{-T_{n-k+1:n}}$ com $\{U_i\}$ sucessão de uniformes em $[0, 1]$ e $\{T_i\}$ sucessão de exponenciais unitárias. Considerando inicialmente $T_{n-k+1:n}$ que, por (5), se pode escrever como combinação linear de exponenciais independentes, bastará verificar a validade da condição de Lindberg-Feller (como é o caso) para que se possa

aplicar o teorema Limite Central a

$$\frac{T_{n-k+1:n} - M_n}{S_n}$$

onde $M_n = E(T_{n-k+1:n})$ e $S_n^2 = \text{var}(T_{n-k+1:n})$.

Por outro lado, tem-se

$$M_n = E(T_{n-k+1:n}) = \sum_{j=1}^{n-k+1} \frac{E(V_j)}{n-j+1} = \sum_{j=1}^{n-k+1} \frac{1}{n-j+1} \sim \ln \frac{n}{k} + O\left(\frac{1}{k}\right)$$

e

$$S_n^2 = \text{var}(T_{n-k+1:n}) = \sum_{j=1}^{n-k+1} \left(\frac{1}{n-j+1} \right)^2 \sim \frac{1}{k} - \frac{1}{n} + O\left(\frac{1}{k^2}\right)$$

e, consequentemente,

$$\frac{T_{n-k+1:n} - M_n}{S_n} = \frac{T_{n-k+1:n} - \ln \frac{n}{k}}{\sqrt{\frac{n-k}{nk}}} + O\left(\sqrt{\frac{n}{k(n-k)}}\right).$$

Como $k/n \rightarrow p$ implica que $\sqrt{n/[k(n-k)]} \rightarrow 0$, conclui-se então que

$$\frac{T_{n-k+1:n} + \ln p}{\sqrt{\frac{1-p}{np}}} \rightsquigarrow \mathcal{N}(0, 1).$$

Passando ao caso geral, obtém-se, numa primeira fase,

$$\begin{aligned} P \left[\frac{-\ln F(X_{k:n}) + \ln p}{\sqrt{\frac{1-p}{np}}} \leq x \right] = \\ P \left[X_{k:n} \geq F^{-1} \left(p \exp \left\{ -x \sqrt{\frac{1-p}{np}} \right\} \right) \right] \xrightarrow{n \rightarrow \infty} \Phi(x) \end{aligned}$$

seguindo-se o resultado pretendido após o desenvolvimento em série da função $F^{-1} \left(p \exp \left\{ -x \sqrt{\frac{1-p}{np}} \right\} \right)$ até à primeira ordem, em torno do ponto p .

Referências

- David, H.A. (1981). *Order Statistics*. John Wiley & Sons.
- Fisher, R.A. e Tippet, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.*, **24**, 180-190.
- Gnedenko, B. (1943). Sur la distribution limite du terme maximale d'une série aléatoire. *Ann. Math.*, **44**, 423-453.
- de Haan, L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*, Amsterdam Math. Centre Tracts, **32**, 1-124.
- Leadbetter, M.G., Lindgren, G. e Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer Verlag New York Inc.
- Malmquist, S. (1950). On a property of order statistics from a rectangular distribution. *Skand. Aktuarietidskr.*, **33**, 214-222.
- Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*, Springer-Verlag New York Inc.
- Rényi, A. (1953). On the Theory of Order Statistics. *Acta Math. Acad. Sci. Hung.*, **4**, 191-231.
- Sarhan, A.E. e Greenberg, B.G. (Eds.) (1962). *Contributions to Order Statistics*. Wiley, New York.
- von Mises, R. (1923). Über die Variationsbreite einer Beobachtungsreihe. *Sitzungsberichte der Berliner Math. Gesellschaft*, **22**, 3-8. (Reproduzido em von Mises (1964), pp. 129-134).
- von Mises, R. (1936). La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalkanique*, **1**, 141-160. (Reproduzido em von Mises (1964), pp. 271-294).
- von Mises, R. (1964). *Selected Papers of Richard von Mises*, Vol 2, American Mathematical Society, Providence.