Estatísticas de Ordem e Distribuições Assintóticas de Valores Extremos

Vanderlei da Costa Bueno

Instituto de Matemática e Estatística

Universidade de São Paulo, SP. Brasil

Novembro de 2020

Estatísticas de Ordem

Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição F definidas em um espaço de probabilidade $(\Omega, \Im, P)$. A cada realização $w \in \Omega$, ordenamos $X_1(w), X_2(w), ..., X_n(w)$ e denotamos esta ordenação por por $X_{(n;1)} \leq X_{(n;2)} \leq .... \leq X_{(n;n)}$. $X_{(n;k)}$ é denominada k-ésima estatística de ordem dos $X_1, X_2, ..., X_n$. Em particular denotamos:

$$X_{(n;1)} = \min\{X_1, X_2, ..., X_n\}$$

$$X_{(n;n)} = \max\{X_1, X_2, ..., X_n\}.$$

Assumiremos que F é contínua e portanto $P(X_i = X_j) = 0, \forall i, j$ e concluímos que $X_{(n;1)} < X_{(n;2)} < .... < X_{(n;n)}$.

Estatísticas de Ordem

Teorema 1
Sob as hipóteses acima, a função densidade de probabilidade (conjunta) de $X_{(n;k)}$, ($X_{(n;i)}, X_{(n;j)}$) e de $(X_{(n;1)}, X_{(n;2)}, ..., X_{(n;n)}$) são respectivamente

$$f_{X_{(n;k)}}(x) = \frac{n!}{(k-1)!(n-k)!}(1-F(x))^{n-k}F(x)^{k-1}f(x);$$

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

$$[1-F(y)]^{n-j}f(x)f(y) \quad se \quad x < y;$$

$$f_{X_{(n;1)},X_{(n;2)},...,X_{(n;n)}}(x_1,x_2,...,x_n) = n!f(x_1)f(x_2)....f(x_n)$$

se $x_1 < x_2 < ... < x_n$.

Estatísticas de Ordem

Prova

$$f_{X_{(n;k)}}(x) = \lim_{dx \downarrow 0} \frac{F_{X_{(n;k)}}(x+dx) - F_{X_{(n;k)}}(x)}{dx} =$$

$$\lim_{dx \downarrow 0} \frac{P(x < X_{(n;k)} \leq x+dx)}{dx} =$$

$$\lim_{dx \downarrow 0} \frac{P((k-1) \;\; dos \;\; X_i's \in (-\infty, x], \quad um \;\; X_i \in (x, x+dx],}{dx}$$

$$\frac{P((n-k) \;\; dos \;\; X_i's \in (x+dx, \infty))}{dx} =$$

$$\lim_{dx \downarrow 0} \frac{n!}{(k-1)!(n-k)!} \frac{F(x)^{k-1}[F(x+dx) - F(x)][1 - F(x+dx)]^{n-k}}{dx}.$$

Prova

Como $\lim_{dx \downarrow 0} 1 - F(x + dx) = 1 - F(x)$ e
$\lim_{dx \downarrow 0} \frac{F(x+dx)-F(x)}{dx} = f(x)$ concluimos que

$$f_{X_{(n;k)}}(x) = \frac{n!}{(k-1)!(n-k)!}(1 - F(x))^{n-k} F(x)^{k-1} f(x).$$

A parte restante da prova segue com argumentos análogos.

Uma prova alternativa da demonstração acima que tem interesse em si, segue na observação abaixo.

Observação 1

Considere a função beta definida por

$$B_{n,k}(u) = \frac{n!}{(k-1)!(n-k)!} \int_0^u t^{k-1}(1-t)^{n-k} dt =$$

$$\frac{n!}{k!(n-k)!} \int_0^u (1-t)^{n-k} dt^k, 0 < u < 1.$$

Integrando por partes, temos:

$$B_{n,k}(u) = \binom{n}{k} t^k (1-t)^{n-k} |_0^u + \binom{n}{k} \int_0^u (n-k) t^k (1-t)^{n-k-1} dt =$$

$$\binom{n}{k} u^k (1-u)^{n-k} + \frac{n!}{k!(n-k-1)!} \int_0^u t^k (1-t)^{n-k-1} dt.$$

Repetindo tal processo $(n-k-1)$ vezes obtemos

$$B_{n,k}(u) = \sum_{r=k}^{n} \binom{n}{r} u^r (1-u)^{n-r}.$$

Estatísticas de Ordem

Consequentemente

$$P(X_{(n;k)} \leq x) = P(\sum_{i=1}^{n} 1_{\{X_i \leq x\}} \geq k) = \sum_{r=k}^{n} \binom{n}{r} F(x)^r (1-F(x))^{n-r} =$$

$$\frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt.$$

Se F é absolutamente contínua,

$$f_{X_{(n;k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x).$$

Estatísticas de Ordem

Funções das Estatísticas de Ordem.

A média amostral da estatísticas de ordem $\frac{\sum_{k=1}^{n} X_{(n;k)}}{n}$ é identicamente distribuida à média amostral dos $X_i's$, $\frac{\sum_{k=1}^{n} X_k}{n}$.

A mediana é definida por

$$Md = \begin{cases} X_{(n:\frac{n+1}{2})}, & n = 2k+1 \\ \frac{X_{(n:\frac{n}{2})} + X_{(n:\frac{n+1}{2})}}{2}, & n = 2k \end{cases}$$

A amplitude R é definida por $R = X_{(n;n)} - X_{(n;1)}$.

A amplitude média T é definida por $\frac{X_{(n;n)} + X_{(n;1)}}{2}$.

Estatísticas de Ordem

Função de distribuição empírica.
Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição F definidas em um espaço de probabilidade $(\Omega, \Im, P)$. A função de distribuição empírica é definida por:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

Estatísticas de Ordem

Observe que a função de distribuição empírica é um estimador não viciado e consistente da função de distribuição,

$$E[F_n(x)] = \frac{1}{n}\sum_{i=1}^{n} E[1_{\{X_i \leq x\}}] = F(x)$$

e

$$Var(F_n(x)) = E[(F_n(x) - F(x))^2] =$$

$$\frac{1}{n^2} Var(\sum_{i=1}^{n} 1_{\{X_i \leq x\}}) = \frac{F(x)(1 - F(x))}{n}$$

que converge para 0 quando n converge para o infinito. Portanto $F_n(x) \rightarrow^{mq} F(x)$, $F_n(x) \rightarrow^{P} F(x)$ e $F_n(x) \rightarrow^{D} F(x)$.

Estatísticas de Ordem

Observação 2

$$P(F_n(x) \geq \frac{k}{n}) = P(nF_n(x) \geq k) = P(X_{(n;k)} \leq x)$$

e

$$F_n(x) = \begin{cases} 0 & x < X_{(n;1)} \\ \frac{k}{n} & X_{(n;k)} \leq x < X_{(n;k+1)} \\ \\ 1 & x \geq X_{(n;n)} \end{cases}$$

Portanto existe uma correspondência biunívoca entre $F_n(x)$ e as estatísticas de ordem.

Estatísticas de Ordem

Considere um número real $p, 0 < p < 1$ e seja ζ_p o p-ésimo quantil de F, isto é, ζ_p é a única solução de $F(x) = p$, quando existir. Se, para uma estatística de ordem $X_{(n;k)}$, $\frac{k}{n}$ converge para p de maneira conveniente, $X_{(n;k)}$ é chamado o p-ésimo quantil amostral, $\widehat{\zeta}_{p,n}$. Embora existam várias maneiras de definirmos tal k, as mais adotadas são $k = k_p = [np] + 1$ e $k = k_p = [(n+1)p]$.

Exemplo 1

Se $p = \frac{1}{2}$, ζ_p é a mediana de F. Se n é impar, $n = 2m + 1$ temos

$$[np] + 1 = [\frac{(2m+1)}{2}] + 1 = [m + \frac{1}{2}] + 1 = m + 1$$

e $[(n+1)p] = [\ (2m+1+1)\ \frac{}{2]} = [m + 1] = m + 1.$

Assim o p-ésimo quantil amostral é $\widehat{\zeta}_{p,n} = X_{(n;m+1)}$.

Se n é par, $n = 2m$ temos

$$[np] + 1 = [\frac{2m}{2}] + 1 = m + 1 \quad e \quad [(n+1)\frac{1}{2}] = [m + \frac{1}{2}] = m.$$

Neste caso convencionamos definir p-ésimo quantil amostral como $\widehat{\zeta}_{p,n} = \frac{X_{(n;m+1)} + X_{(n;m)}}{2}$

Distribuições assintóticas de valores extremos

No contexto da teoria assintótica, quando $\frac{k}{n} \to_{n\to\infty} 0$ ou $\frac{k}{n} \to_{n\to\infty} 1$, estas estatísticas de ordem são denominadas de valores extremos. As populações adequadas para tais formulações são:
A) Suponha que exista $\zeta_0 > -\infty$

$$F(x) > 0 \;\; se \;\; x > \zeta_0 \;\; e \;\; F(x) = 0 \;\; se \;\; x \leq \zeta_0,$$

então ζ_0 é denominado um ponto final inferior para F. Se $\zeta_0 = -\infty$, dizemos que F tem um ponto final inferior infinito.

Estatísticas de Ordem

B) Suponha que exista $\zeta_1 < \infty$ tal que

$$F(x) < 1 \;\; se \;\; x < \zeta_1 \;\; e \;\; F(x) = 1 \;\; se \;\; x \geq \zeta_1,$$

então ζ_1 é denominado um ponto final superior para F. Se $\zeta_1 = \infty$, dizemos que F tem um ponto final superior infinito.
O comportamento das distribuições dos valores extremos dependem se a população tem pontos finais inferiores (superiores) finitos ou não.

Distribuições assintóticas de valores extremos

Teorema 2 Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição F definidas em um espaço de probabilidade $(\Omega, \Im, P)$. Se F tem um ponto final inferior finito (ζ_0) e um ponto final superior finito (ζ_1), então $X_{(n;n)} \rightarrow_P \zeta_1$ e $X_{(n;1)} \rightarrow_P \zeta_0$. Prova nas notas do professor.

Distribuições assintóticas de valores extremos

Teorema 3

Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição F definidas em um espaço de probabilidade $(\Omega, \Im, P)$. Assuma que o p-ésimo quantil $\zeta_p, 0 < p < 1$ é unicamente definido, isto é,
$\forall \varepsilon > 0, F(\zeta_p - \varepsilon) < F(\zeta_p) = p < F(\zeta_p + \varepsilon)$. Então o p-'esimo quantil amostral $X_{(n;k)}$, com $k = k_p$, é tal que $X_{(n;k)} \rightarrow^{P} \zeta_p$ e $X_{(n;k)} \rightarrow^{qc} \zeta_p$.

Prova

$$P(|X_{(n;k)} - \zeta_p| \leq \varepsilon) = P(X_{(n;k)} \leq \zeta_p + \varepsilon) - P(X_{(n;k)} \leq \zeta_p - \varepsilon).$$

Consideremos a variável aleatória Y^+ com distribuição binomial de parâmetro n e $F(\zeta_p + \varepsilon)$, de forma que

$$P(X_{(n;k)} \leq \zeta_p + \varepsilon) = P(Y^+ \geq k) = P(\frac{Y^+}{n} \geq \frac{k}{n}).$$

Contudo $\frac{Y^+}{n} \rightarrow^{qc} F(\zeta_p + \varepsilon) > p$ e $\frac{k}{n} \rightarrow p$. Concluimos que $\lim_{n\rightarrow\infty} P(X_{(n;k)} \leq \zeta_p + \varepsilon) = 1$.

Distribuições assintóticas de valores extremos

Por outro lado, consideremos a variável aleatória Y^- com distribuição binomial de parâmetro n e $F(\zeta_p - \varepsilon)$, de forma que

$$P(X_{(n;k)} \leq \zeta_p - \varepsilon) = P(Y^- \geq k) = P(\frac{Y^-}{n} \geq \frac{k}{n})$$

e $\frac{Y^-}{n} \rightarrow^{qc} F(\zeta_p - \varepsilon) < p$ e $\frac{k}{n} \rightarrow p$. Concluimos que $\lim_{n\rightarrow\infty} P(X_{(n;k)} \leq \zeta_p - \varepsilon) = 0$.
Portanto

$$\lim_{n\rightarrow\infty} P(|X_{(n;k)} - \zeta_p| \leq \varepsilon) = 1$$

e $X_{(n;k)} \rightarrow^P \zeta_p$.
A prova da convergência quase certa não será reproduzida no texto.

Distribuições assintóticas de valores extremos

Teorema 5

Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas definidas em um espaço de probabilidade $(\Omega, \Im, P)$ com distribuição F e função densidade de probabilidade $f(x)$ tal que $f(\zeta_p) > 0$. Então

$$\lim_{n\to\infty} P(\frac{\sqrt{n}(X_{(n;k)} - \zeta_p)}{\gamma} \leq x) = P(Z \leq x)$$

onde $\gamma^2 = \frac{p(1-p)}{f(\zeta_p)^2}$.

Prova

Consideremos a variável aleatória Y_n com distribuição binomial de parâmetros n e $F(\zeta_p + \frac{1}{\sqrt{n}}x)$.

$$P(\sqrt{n}(X_{(n;k)} - \zeta_p) \leq x) = P(X_{(n;k)} \leq \zeta_p + \frac{1}{\sqrt{n}}x) = P(Y_n \geq k) =$$

$$P(\frac{Y_n - nF(\zeta_p + \frac{1}{\sqrt{n}}x)}{\sqrt{nF(\zeta_p + \frac{1}{\sqrt{n}}x)(1 - F(\zeta_p + \frac{1}{\sqrt{n}}x))}} \geq$$

$$\frac{k - nF(\zeta_p + \frac{1}{\sqrt{n}}x)}{\sqrt{nF(\zeta_p + \frac{1}{\sqrt{n}}x)(1 - F(\zeta_p + \frac{1}{\sqrt{n}}x))}}).$$

Distribuições assintóticas de valores extremos

Para calcular o limite, quando $n \to \infty$ da expressão à direita usaremos o teorema do valor médio
$(\int_a^b f(x)dx = (b-a)f(a+\theta(b-a)), 0 < \theta\theta < 1)$.

$$\frac{1}{\sqrt{n}}(k - nF(\zeta_p + \frac{1}{\sqrt{n}}x)) = \frac{k}{\sqrt{n}} - \frac{n}{\sqrt{n}}[\int_0^{\zeta_p} \zeta f(z)dz + \int_{\zeta_p}^{\zeta_p + \frac{1}{\sqrt{n}}x} f(z)dz] =$$

$$\frac{k}{\sqrt{n}} - \frac{np}{\sqrt{n}} - \frac{n}{\sqrt{n}}[\zeta_p + \frac{1}{\sqrt{n}}x - \zeta_p] f(\zeta_p + \theta(\zeta_p + \frac{1}{\sqrt{n}}x) - \zeta_p), 0 < \theta < 1.$$

Portanto

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}}(k - nF(\zeta_p + \frac{1}{\sqrt{n}}x)) = -xf(\zeta_p)$$

e

$$\lim_{n\to\infty} \frac{k - nF(\zeta_p + \frac{1}{\sqrt{n}}x)}{\sqrt{nF(\zeta_p + \frac{1}{\sqrt{n}}x)(1 - F(\zeta_p + \frac{1}{\sqrt{n}}x))}}) = \frac{-xf(\zeta_p)}{\sqrt{p(1-p)}}.$$

Estatísticas de Ordem

Concluimos que, pelo teorema do limite central

$$\lim_{n\to\infty} P(\sqrt{n}(X_{(n;k)} - \zeta_p) \leq x) = P(Z \leq \frac{-xf(\zeta_p)}{\sqrt{p(1-p)}})$$

e

$$\lim_{n\to\infty} P(\frac{\sqrt{n}(X_{(n;k)} - \zeta_p)}{\gamma} \leq x) = P(Z \leq x)$$

onde $\gamma^2 = \frac{p(1-p)}{f(\zeta_p)^2}$.

Exemplo 3

Sejam $x_1, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição exponencial de par ametro λ. Desde que $P(X_1 > x) = e^{-\lambda x}$ temos que a mediana m é a solução de $e^{-\lambda m} = 0,5$, isto é $m = \frac{0,7}{\lambda}$. Portanto

$$f(m) = \lambda e^{-\frac{0,7}{\lambda}} = 0,5\lambda.$$

Estatísticas de Ordem

Pelo teorema acima

$$Md \sim N(\frac{0,7}{\lambda}, \frac{1}{4n0,25\lambda^2}).$$

Um intervalo de confiança para λ, ao nível de $0,95$ de confiança é obtido através de

$$P(-1,96 \leq (Md - \frac{0,7}{\lambda})\sqrt{n}\lambda \leq 1,96) = 0,95$$

produzindo

$$(\frac{-1,96 + 0,7\sqrt{n}}{\sqrt{n}m}, \frac{1,96 + 0,7\sqrt{n}}{\sqrt{n}m}).$$