1. AULA 7

Estatísticas de Ordem.

Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição F definidas em um mesmo espaço de probabilidade $(\Omega, \Im, P)$. A cada realização $w \in \Omega$, ordenamos $X_1(w), X_2(w), ..., X_n(w)$ e denotamos 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 $X_{(n;1)} < X_{(n;2)} < .... < X_{(n;n)}$.

Teorema 1.1. *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*

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

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

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

Prova:

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

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

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

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

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 1.2. 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

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

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,

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

1.1. **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 $\frac{X_{(n;n)}+X_{(n;1)}}{2}$.

Exemplo 1.3. Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição uniforme no intervalo $(0, 1)$. A função de densidade conjunta de $X_{(n;1)}, X_{(n;n)}$ é

$$f_{X_{(n;1)},X_{(n;n)}}(x, y) = n(n-1)[F(y)-F(x)]^{n-2} f(x)f(y), \ 0 < x < y < 1, \ e \ 0c.c.$$

Nosso objetivo é encontrar a fução densidade de probabilidade da amplitude R. Como variável auxiliar tomaremos a amplitude média T.

Assim $r = y - x$ e $t = \frac{x+y}{2}$,e $x = t - \frac{r}{2}$ e $y = t + \frac{r}{2}$ definem a transformação bijetora com Jacobiano

$$J = \frac{\delta x}{\delta r}\frac{\delta y}{\delta t} - \frac{\delta x}{\delta t}\frac{\delta y}{\delta r} = \frac{-1}{2} + \frac{-1}{2} = -1.$$

O valor absoluto do Jacobiano , $|J| = 1$, é 1. Se consideramos $n = 10$

$$f_{R,T}(r, t) = 10.9.[t + \frac{r}{2} - t + \frac{r}{2}]^8.1.1 = 90r^8 1_D(r, t)$$

onde D é a região de definição obtida através das regiões fechadas

$$0 < x < 1, y = 0 \Rightarrow 0 \leq t - \frac{r}{2} \leq 1, t = \frac{-r}{2} \Rightarrow -1 \leq r \leq 0;$$

$$x = 0, 0 < y < 1 \Rightarrow t - \frac{r}{2} = 0, 0 < t + \frac{r}{2} < 1 \Rightarrow 0 < r < 1;$$

$$x = 1, 0 < y < 1 \Rightarrow t - \frac{r}{2} = 1, 0 < t + \frac{r}{2} < 1 \Rightarrow -1 < r < 0;$$

$$0 < x < 1, y = 1 \Rightarrow 0 < t = \frac{-r}{2} < 1, t = 1 - \frac{-r}{2} \Rightarrow 0 < r < 1.$$

Portanto a função densidade de probabilidade da amplitude R é

$$f_R(r) = \begin{cases} \int_{-\frac{r}{2}}^{\frac{r+2}{2}} 90r^8 dt = 90r^8(r+1) & : \quad -1 < r < 0 \\ & : \\ \int_{\frac{r}{2}}^{\frac{2-r}{2}} 90r^8 dt = 90r^8(1-r) & : \quad 0 < r < 1 \end{cases}$$

1.2. **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:

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

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 $F_n(x) \to^{mq} F(x)$, $F_n(x) \to^{P} F(x)$ e $F_n(x) \to^{D} F(x)$.

Com mais rigor, Glivenko-cantelli, provou o teorema

Teorema 1.4. *Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuidas com distribuição F definidas em um espaço de probabilidade $(\Omega, \Im, P)$. Então*

$$\sup_{x \in \Re} |F_n(x) - F(x)| \to^{qc} 0.$$

Note que

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

e

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

Portanto existe uma correspondência biunívoca entre $F_n(x)$ e as 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 1.5. Se $p = \frac{1}{2}$, ζ_p é a mediana de F. Se n é impar, $n = 2m+1$ temos

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

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

Se n é par, $n = 2m$ temos

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

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

Desde que F_n é uma função escada, os p-ésimo quantis amostrais podem ser definidos como

$$\widehat{\zeta}^1_{p,n} = \sup\{x : F_n(x) \leq p\}$$
$$\widehat{\zeta}^2_{p,n} = \inf\{x : F_n(x) \geq p\}$$

definem um intervalo aberto onde podemos realizar uma interpolação linear

Teorema 1.6. *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 $\gamma^2 = \frac{p(1-p)}{f(\zeta_p)^2}$.

Exemplo 1.7. 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

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

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

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