Capítulo 2

Sumarização dos dados

Seja X uma v.a. que denota determinada característica de interesse de uma certa população. Suponhamos que $f_X(.|\theta)$ é a sua f.d.p., onde θ designa o parâmetro. Assumimos ainda que o parâmetro é desconhecido, pelo que se pretende efectuar uma amostragem e, com base nesta, inferir sobre o valor de θ. Finalmente seja $\underline{x} = (x_1, x_2, \ldots, x_n)$ a amostra recolhida, de dimensão n, e $\underline{X} = (X_1, \ldots, X_n)$ uma amostra aleatória genérica. Designamos as n estatísticas ordinais respeitantes a $\underline{X}$ por $(X_{(1)}, \ldots, X_{(n)})$, onde $X_{(i)} \leq X_{(j)}, \forall j > i$.

Claramente que toda a informação respeitante à amostra está contida em $\underline{x}$; se n for elevado, $\underline{x}$ é uma lista de valores cuja interpretação pode ser mais ou menos difícil. Será pois desejável condensar, de algum modo, a informação contida na amostra, de forma a que a informação contida na amostra sobre o parâmetro não seja perdida e que, simultaneamente, se consiga uma menor dimensionalidade dos dados em análise. Tal fim pode ser obtido considerando *estatísticas*. Recordemos a definição de estatística.

Definição 1 *Seja $\underline{X} = (X_1, X_2, \ldots, X_n)$ uma a.a. i.i.d. à v.a. X, com f.d.p. pertencente à família $\mathcal{F} = \{f(.|\theta), \theta \in \Theta\}$, e seja $T(\underline{X})$ uma função de $\underline{X}$ tal que, dado que $X_i = x_i, i = 1, \ldots, n$, $T(\underline{x})$ toma um valor que só depende dos valores amostrais observados. Então diz-se que $T(.)$ é uma estatística.*

Assim uma estatística é qualquer função que só depende dos valores amostrais e que, em particular, a sua concretização não depende dos parâmetros eventualmente desconhecidos da f.d.p. da população.

Exemplo 2 *Seja $X \sim Binomial(10, p)$, com $p \in (0, 1)$ desconhecido. Seja ainda $T_1(\underline{X}) = \sum_{i=1}^{20} X_i$ e $T_2(\underline{X}) = \sum_{i=1}^{20}(X_i - 10p)$, onde $\underline{X} = (X_1, X_2, \ldots, X_{20}) \sim_{i.i.d.} X \sim Bin(10, p)$. $T_1(.)$ é uma estatística mas $T_2(.)$ não é.*

Qualquer estatística define uma redução na dimensionalidade dos dados, embora haja estatísticas de igual dimensionalidade à dos dados originais (pense-se, por exemplo, na estatística $T(\underline{X}) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ - estatísticas ordinais). Suponhamos que $\underline{x}$ e $\underline{y}$

são tais que $T(\underline{x}) = T(\underline{y})$. Então um utilizador que só use o valor da estatística (em vez do da amostra) tratará as duas amostras de igual forma.

Uma estatística induz uma partição no espaço de estados amostral. Seja Ω_X o espaço amostral de X, $\Omega_{\underline{X}}$ o espaço amostral de uma amostra aleatória $\underline{X}$ e $\Omega_T = \{t : t = T(\underline{x}), \underline{x} \in \Omega_X\}$ a imagem de Ω_X sob a estatística T. Então $T(.)$ induz uma partição da forma $A_t, t \in \Omega_T$, onde $A_t = \{\underline{x} \in \Omega_{\underline{X}} : T(\underline{x}) = t\}$.

Neste capítulo estudaremos formas de redução da dimensionalidade dos dados. Estaremos interessados em métodos que reduzam a dimensionalidade e que simultaneamente não percam informação sobre o parâmetro. Focaremos três formas de conseguir este objectivo: via *suficiência*, *verosimilhança* e *invariância*. Os três partilham o mesmo objectivo mas com filosofias distintas.

2.1 Princípio da Suficiência

Seja X uma v.a. e $T(.)$ uma estatística obtida como função dos valores amostrais que X pode tomar. A obtenção de uma amostra particular $\underline{x}$ pode ser conceptualmente visualizada da seguinte forma:

(a) Selecção directa de $\underline{x}$ (forma mais usual de gerar uma amostra), de acordo com a f.d.p. $f_{\underline{X}}(\underline{x}|\theta$.

(b) Obtenção de um valor particular $t \in \Omega_T$, seguido de geração de $x \in A_t$ de acordo com a f.d.p. $f_{\underline{X}}(\underline{x}|t, \theta)$.

Note-se que

$$\begin{aligned} f_{\underline{X}}(\underline{x}|\theta) &= \int_{\Omega_T} f_{\underline{X},T}(\underline{x}, t|\theta)dt \\ &= f_{\underline{X},T}(\underline{x}, t|\theta)I_{A_t}(\underline{x}) \\ &= f_T(t|\theta)f_{\underline{X}}(\underline{x}|T = t, \theta)I_{A_t}(\underline{x}). \end{aligned}$$

Uma *estatística suficiente* para o parâmetro θ é uma estatística que, num certo sentido, contém toda a informação relevante àcerca do parâmetro θ. Assim temos o **Princípio da Suficiência**, o qual se enuncia a seguir.

Definição 3 *se T é uma estatística suficiente para θ então qualquer inferência àcerca de θ depende do valor amostral $\underline{X}$ apenas através do valor $T(\underline{X})$. Assim se $\underline{x}$ e $\underline{y}$ forem dois pontos amostrais tais que $T(\underline{x}) = T(\underline{y})$ então os resultados inferenciais sobre θ são iguais quer $\underline{X} = \underline{x}$ ou $\underline{X} = \underline{y}$.*

o que equivale a dizer que

Definição 4 *T é uma estatística suficiente para a família de distribuições $\mathcal{F} = \{f(.|\theta), \theta \in \Theta\}$ (abreviadamente, T é uma estatística suficiente para θ) se*

$$f(\underline{x}|T = t, \theta) = f(\underline{x}|T = t), \quad \forall \underline{x} \in A_t.$$

Assim se T é uma estatística suficiente toda a informação contida na amostra sobre o parâmetro é retida pela estatística. Vejamos o seguinte exemplo.

Exemplo 5 *Seja* $X \sim Poisson(\lambda)$ *e* $\mathcal{F} = \{Poisson(\lambda), \lambda \in I\!R^+\}$ *e suponhamos que é retirada uma a.a. de* X *de dimensão 2. Seja* $T(X_1, X_2) = X_1 + X_2$*, pelo que*

$$f_T(t|\lambda) = \frac{e^{-2\lambda}(2\lambda)^t}{t!} I_{I\!N_0}(t)$$

Por outro lado

$$\begin{aligned} f_{X_1}(x|T=t,\lambda) &= \frac{P(X_1 = x, X_1 + X_2 = t|\lambda) I_{I\!N_0}(x)}{P(X_1 + X_2 = t|\lambda) I_{I\!N_0}(t)} \\ &= \frac{P(X_1 = x|\lambda) P(X_2 = t - x|\lambda) I_{I\!N_0}(t-x)}{P(X_1 + X_2 = t|\lambda) I_{I\!N_0}(t)} \\ &= \left(\frac{1}{2}\right)^t \binom{t}{x} I_{I\!N_0}(t) I_{I\!N_0}(t-x) \\ &= f_{X_1}(x|t) \end{aligned}$$

donde se conclui que $(X_1|T = t, \lambda) \sim Bin(t, \frac{1}{2})$*, pelo que a distribuição condicionada de* X_1 *não depende de* λ*, i.e.,* T *é uma estatística suficiente para* λ*.*

De forma a utilizar a Definição 3, é necessário verificar que $f_{\underline{X}}(\underline{x}|T = t, \theta)$ é constante para todos os valores de $\theta \in \Theta$. Note-se que se $x \notin A_t$ esta f.d.p. é nula para qualquer valor de θ, pelo que só necessitamos de indagar para $x \in A_t$. Dado que $\{\underline{X} = \underline{x}\} \subseteq \{T(\underline{X}) = T(\underline{x})\}$, então

$$\begin{aligned} f_{\underline{X}}(\underline{x}|T = t, \theta) &= \frac{f_{\underline{X},T}(\underline{x}, t|\theta)}{f_T(t|\theta)} \\ &= \frac{f_{\underline{X}}(\underline{x}|\theta)}{f_T(t|\theta)} \end{aligned}$$

concluímos que T é uma estatística suficiente se e só se o rácio da f.d.p. de $\underline{X}$ e da f.d.p. da estatística T for uma função constante em θ, para qualquer $\underline{x} \in \Omega_X$.

Claramente que um problema que se coloca na caracterização de uma estatística como suficiente prende-se com o facto de ser necessário determinar a f.d.p. da estatística, o que nem sempre é exequível. O teorema que se segue estabelece um critério que permite caracterizar qualquer estatística como sendo suficiente ou não.

Teorema 6 *(***Critério da factorização de Halmos-Savage-Bahadur***): Seja* $f_{\underline{X}}(.|\theta)$ *a f.d.p. de* $\underline{X}$*. A estatística* T *é suficiente para* θ *se e só se existirem funções* $g(.)$ *e* $h(.)$*, ambas não negativas, tais que*

$$f_{\underline{X}}(\underline{x}|\theta) = g(T(\underline{x}), \theta) h(\underline{x}), \forall \underline{x} \in \Omega_X.$$

Este teorema, para além de ultrapassar a dificuldade inerente à própria definição de estatística suficiente, permite ainda por vezes "'adivinhar"' uma estatística suficiente. Assim, dada a f.d.p. conjunta da amostra $\underline{x}$, factoriza-se em duas partes: uma dependente

do parâmetro θ e outra que não depende de θ. A parte que depende de θ, que constitui $g(T(\underline{x}),\theta)$, usualmente depende dos valores amostrais $\underline{x}$ apenas através de alguma função $T(\underline{x})$, a qual corresponderá à estatística suficiente para θ. Vejamos um exemplo.

Exemplo 7 *Considere-se a família de normais com parâmetro de escala σ fixo, $\mathcal{F} = \{f_{N(\mu,\sigma)}(.|\mu), \mu \in I\!R\}$. Factorizando a f.d.p. vem:*

$$\begin{aligned} f_{\underline{X}}(\underline{x}|\mu) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^n x_i\right)} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2} e^{-\frac{1}{2\sigma^2}\left(n\mu^2 - 2\mu\sum_{i=1}^n x_i\right)} \end{aligned}$$

pelo que se fizermos

$$g(T(\underline{x}),\mu) = e^{-\frac{1}{2\sigma^2}\left(n\mu^2 - 2\mu\sum_{i=1}^n x_i\right)}, \quad h(\underline{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2}$$

vem que $\underline{x}$ intervém em conjunção com o parâmetro μ sob a forma de $\sum_{i=1}^n x_i$, pelo que, de acordo com o teorema anterior, $T(\underline{X}) = \sum_{i=1}^n X_i$ é, neste caso, uma estatística suficiente. Claro que não é única (por exemplo $\bar{X}$ e $(X_{(1)}, \ldots, X_{(n)})$ são também estatísticas suficientes para μ).

Na caracterização de estatísticas suficientes é preciso ter em linha de conta a definição do suporte da f.d.p. O próximo exemplo ilustra como por vezes podemos cair em erros por não contemplar o suporte das f.d.p.

Exemplo 8 *Considere-se a seguinte família de f.d.p. $\mathcal{F} = \{Unif(a,b), a < b \in I\!R\}$, para a qual a f.d.p. é dada por*

$$f_X(x|a,b) = \frac{1}{b-a} I_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & c.c. \end{cases}.$$

Suponhamos que aplicamos o critério de factorização apenas na zona onde a f.d.p. é diferente de zero:

$$f_{\underline{X}}(\underline{x}|a,b) = \left(\frac{1}{b-a}\right)^n$$

pelo que seríamos levados a concluir que qualquer estatística seria suficiente para (a,b)! Mas isto porque erroneamente "'desprezámos"' a função indicatriz. Se agirmos correctamente, temos

$$f_{\underline{X}}(\underline{x}|a,b) = \left(\frac{1}{b-a}\right)^n I_{(a,b)}(x_{(1)}, x_{(n)})$$

pelo que $T(\underline{X}) = (X_{(1)}, X_{(n)})$ é uma estatística suficiente para $\mathcal{F}$.

Note-se neste exemplo que a estatística suficiente encontrada é bidimensional. De facto nem sempre as estatísticas suficientes são unidimensionais, sendo que esta situação ocorre frequentemente quando o parâmetro de interesse é multivariado. Embora não seja sempre verdade, regista-se que há tendência para que estatísticas suficientes para famílias de f.d.p. com parâmetros k-dimensionais sejam também pelo menos k-dimensionais. Há ainda casos mais bizarros, em que uma estatística suficiente para um parâmetro unidimensional tem a mesma dimensionalidade que a amostra (como é o caso da Cauchy - ver exercícios).

Para a família exponencial temos o seguinte resultado:

Teorema 9 *Sejam $X_1, X_2, \ldots, X_n$ v.a. i.i.d. a X com f.d.p. $f_X(.|\theta)$ pertencente à família exponencial. Então*

$$T(\underline{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \sum_{j=1}^{n} t_2(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right)$$

(ver definiçaõ e notação de família exponencial) é uma estatística suficiente para θ.

2.1.1 Estatísticas Suficientes Mínimas

Como vimos na secção anterior, para um dado modelo pode haver mais do que uma estatística suficiente. Por exemplo a amostra $\underline{X}$ é sempre estatística suficiente; se T for uma estatística suficiente e $r(.)$ for uma função bijectiva, com função inversa r^{-1}, então $r(T(\underline{X}))$ é ainda uma estatística suficiente (uma vez que é uma estatística equivalente a $T(\underline{\mathrm{X}})$).

A não-unicidade de estatísticas suficientes coloca a questão de saber se existe uma estatística suficiente que seja melhor que as restantes. Claramente se duas estatísticas $T_1(\underline{\mathrm{X}})$ e $T_2(\underline{\mathrm{X}})$ forem suficientes e se T_1 conseguir a maior redução dos dados, então dever-se-à optar por T_1. Esta ideia leva à definição de *estatística suficiente mínima.*

Definição 10 *Uma estatística suficiente $T(.)$ é uma estatística suficiente mínima se para qualquer outra estatística suficiente $T'(.)$, $T(.)$ for uma função de $T'(.)$.*

Esta definição não é fácil de usar. O resultado que se segue constitui uma ferramenta poderosa na classificação de estatísticas suficientes.

Teorema 11 *(***Teorema de Lehmann-Scheffé***) Seja $f_X(.|\theta)$ a f.d.p. de uma v.a. X e $f_{\underline{X}}(.|\theta)$ a f.d.p. correspondente a uma a.a. Seja $T(.)$ uma função tal que, dados $\underline{x}$ e $\underline{y}$, o ratio $\frac{f_X(\underline{x}|\theta)}{f_X(\underline{y}|\theta)}$ não depende de θ se e só se $T(\underline{x}) = T(\underline{y})$. Então $T(.)$ é uma estatística suficiente mínima.*

Se o suporte da f.d.p. depender do parâmetro θ então para que o rácio $\frac{f_X(\underline{x}|\theta)}{f_X(\underline{y}|\theta)}$ não dependa de θ o numerador e o denominador têm de ser positivos para **exactamente** os mesmos valores de θ. Geralmente esta restrição é reflectida na própria estatística suficiente mínima.

Exemplo 12 *Seja* $X \sim Unif(a,b)$ *e considere-se* $\mathcal{F} = \{Unif(a,b), a < b \in I\!R\}$. *Dado que a f.d.p. conjunta de uma amostra é dada por*

$$f_{\underline{X}}(\underline{x}|a,b) = \frac{1}{(b-a)^n} I_{(a,b)}(minx_i, maxx_i)$$

pelo que

$$\frac{f(\underline{x}|\theta)}{f(\underline{y}|\theta)} = \frac{I_{(a,b)}(minx_i, maxx_i)}{I_{(a,b)}(miny_i, maxy_i)}$$

que não depende nem de a nem de b se e só se $minx_i = miny_i$ *e* $maxx_i = maxy_i$. *Logo de acordo com o teorema de Lehmann-Scheffé, uma estatística suficiente mínima é*

$$T(\underline{X}) = (X_{(1)}, X_{(n)}).$$

Note-se por fim que uma estatística suficiente mínima para um parâmetro não é necessariamente única; na verdade qualquer função bijectiva de uma estatística suficiente mínima é ainda estatística suficiente mínima. Por exemplo, reportando ao caso anterior,

$$T_1(\underline{X}) = (\frac{X_{(n)} - X_{(1)}}{2}, X_{(n)} - X_{(1)})$$

é também uma estatística suficiente mínima. Esta propriedade tem algumas vantagens pois por vezes há estatísticas suficientes mínimas cuja interpretação é mais ou menos fácil que outras.

2.1.2 Estatísticas Ancilares

Em situação oposta às estatísticas suficientes encontram-se as *estatísticas ancilares*, as quais não contêm nenhuma informação sobre o parâmetro.

Definição 13 *Uma estatística* $S(.)$ *diz-se ser ancilar se a sua distribuição não depender do parâmetro.*

Vejamos de seguida um exemplo de uma estatística ancilar (para além das triviais).

Exemplo 14 *Considere-se* $X \sim Unif(a,b)$. *Vimos anteriormente que* $T(\underline{X}) = (X_{(1)}, X_{(n)})$ *é uma estatística suficiente mínima para* (a,b). *Tome-se agora* $S(\underline{X}) = (X_{(n)} - X_{(1)})$, *que define a amplitude de uma amostra. É possível provar que* $S(\underline{X}) \sim Beta(n-1, 2)$, *pelo que se trata realmente de uma estatística ancilar.*

Dos dois exemplos anteriores respeitantes à uniforme ressalta uma observação: enquanto que $(X_{(n)} - X_{(1)})$ é, por si só, ancilar, quando combinado em $T_1(\underline{X}) = (\frac{X_{(n)} - X_{(1)}}{2}, X_{(n)} - X_{(1)})$ dá origem a uma estatística suficiente mínima. Na verdade existem estatísticas que por si só são ancilares mas que combinadas com outras funções amostrais dão estatísticas suficientes contendo informação sobre o parâmetro.

O caso anterior da amplitude ser uma estatística ancilar não advém somente do facto de se tratar de uma uniforme mas decorre essencialmente em todas as f.d.p. pertencentes a uma família de localização. Na verdade

Teorema 15 *Sejam $X_i, i = 1, \ldots, n$ v.a. i.i.d. a X, com f.d.p. pertencente a uma família de localização. Então a estatística $S(\underline{X}) = (X_{(n)} - X_{(1)})$ é uma estatística ancilar para o parâmetro de localização.*

De forma similar, também nas famílias de escala é possível encontrar estatísticas que são sempre ancilares:

Teorema 16 *Sejam $X_i, i = 1, \ldots, n$ v.a. i.i.d. a X, com f.d.p. pertencente a uma família de escala. Então estatísticas que dependam somente das $n-1$ v.a. $\frac{X_1}{X_n}, \frac{X_2}{X_n}, \ldots, \frac{X_{n-1}}{X_n}$ são estatísticas ancilares.*

2.1.3 Estatísticas suficientes, ancilares e completas

Dado que a redução máxima dos dados sem perda de informação sobre o parâmetro é conseguida através das estatísticas suficientes mínimas, devemos ter condições que garantam a minimalidade de uma estatística suficiente. Vimos na secção anterior que uma estatística ancilar tem distribuição que não depende do parâmetro. Assim seria de esperar que qualquer estatística suficiente mínima fosse independente de estatísticas ancilares. Porém tal não é o caso. existem estatísticas que marginalmente são ancilares mas que, conjugadas com mais informação, criam estatísticas suficientes mínimas.

Já vimos, por exemplo, que no caso de uma família $Unif(a, b)$, com $a < b \in \mathbb{R}$, a estatística $X_{(n)} - X_{(1)}$ (baseada numa a.a. de dimensão n) é ancilar para (a, b) mas que $(X_{(n)} - X_{(1)}, X_{(n)} + X_{(1)})$ é uma estatística suficiente mínima para (a, b).

Porém existem casos particulares de estatísticas em que a estatística suficiente mínima é independente de qualquer estatística ancilar. Quando tal acontece estamos perante a classe de *estatísticas completas*.

Definição 17 *Seja $\mathcal{F} = \{f_T(.|\theta), \theta \in \Theta\}$ uma família de f.d.p. da estatística T. . A família $\mathcal{F}$ diz-se ser completa se para uma função $g(.)$ mensurável tal que $E[g(T(\underline{X}))|\theta] = 0$, $\forall\theta \in \Theta$ implica que $P(g(T(\underline{X})) = 0) = 1$, $\forall\theta \in \Theta$.*

Diremos alternativamente que a família de f.d.p. da estatística é completa ou que T *é uma estatística completa para o parâmetro θ.*

Esta propriedade significa que duas quaisquer funções (desde que mensuráveis) da estatística $T(.)$ que tenham a mesma média para todo o parâmetro $\theta \in \Theta$ são idênticas com probabilidade 1.

A definição de completude envolve *todos* os elementos da família de f.d.p. Assim sendo não é possivel ter estatísticas completas só para alguns elementos da família de f.d.p. Por exemplo, tome-se uma f.d.p. $N(0, 1)$ e $g(T(\underline{X}) = \sum_{i=1}^{n} X_i)$. Então $E[g(T(\underline{X}))|\theta] = 0$ mas $P(\sum_{i=1}^{n} X_i = 0) = 0$. Porém esta é uma distribuição particular, pelo que nada se pode concluir sobre completude.

Exemplo 18 *Considere-se a família de distribuições $\{Ber(p), p \in (0, 1)\}$. Para esta*

família $T(\underline{X}) = \sum_{i=1}^{n} X_i$ *é uma estatística (suficiente), com distribuição* $Bin(n, p)$*. Tome-se então uma função* g*, mensurável. Então*

$$E[g(\sum_{i=1}^{n} X_i)|\theta] = \sum_{j=0}^{n} g(j) \binom{n}{j} p^j q^{n-j}$$
$$= \frac{1}{q^n} \sum_{j=0}^{n} g(j) \binom{n}{j} r^j$$

com $r = \frac{p}{1-p}$*. Segue-se então que* $E[g(\sum_{i=1}^{n} X_i)|\theta]$ *é um polinómio de grau* n *em* r*, com coeficientes* $g(j)\binom{n}{j}$*,* $j = 0, \ldots, n$*. Como* $\binom{n}{j} \neq 0$ *então a única forma do polinómio valer* 0 *é que* $g(j) = 0, \forall j = 0, \ldots, n$*. Dado que* $T(\underline{X}) \in \{0, 1, \ldots, n\}$ *então segue-se que* $g(T(\underline{X})) = 0$ *com probabilidade* 1*, i.e.,* $\sum_{i=1}^{n} X_i$ *é uma estatística completa para* p*.*

A noção de completude é particularmente importante no âmbito das estatísticas suficientes mínimas. O próximo teorema fornece uma condição para que uma estatística suficiente mínima seja independente de uma estatística ancilar.

Teorema 19 *Se* T *é uma estatística completa e suficiente mínima então* T *é independente de qualquer estatística ancilar.*

Este teorema - também conhecido por **Teorema de Basu** - é útil para determinar a independência de duas estatísticas sem ter de determinar a f.d.p. conjunta das duas. Porém resta ainda determinar se uma estatística é completa. No geral tal pode ser bem complicado, uma vez que passa por demonstrar resultados em probabilidade e para todo o espaço paramétrico. Porém no caso da família de f.d.p. em causa pertencer à família exponencial e o **espaço paramétrico ser unidimensional** a completude pode ser determinada usando o resultado que se segue.

Teorema 20 *Seja* $\mathcal{F} = \{f(.|\theta), \theta \in \Theta\}$ *uma f.d.p. pertencente à família de f.d.p. exponencial, sendo pois que*

$$f(x|\theta) = h(x)c(\theta)e^{w(\theta)t(x)}$$

Seja ainda $X_i, i = 1, \ldots, n$ *uma a.a. de dimensão* n *proveniente de um elemento desta família. Então a estatística*

$$T(\underline{X}) = \sum_{i=1}^{n} t(X_i)$$

é uma estatística completa.

Vejamos o seguinte exemplo de aplicação deste teorema.

Exemplo 21 *Sejam* $X_1, X_2, \ldots, X_n$ n *v.a. i.i.d., com distribuição exponencial de parâmetro* θ*, e suponhamos que pretendemos calcular*

$$E\left[\frac{X_n}{X_1 + \ldots + X_n}|\theta\right]$$

Seja $g(\underline{X}) = \frac{X_n}{X_1+\ldots+X_n}$. *Note-se que*

$$g(\underline{X}) = \left(\frac{X_1 + X_2 + \ldots + X_n}{X_n}\right)^{-1} = \left(\frac{X_1}{X_n} + \ldots + \frac{X_{n-1}}{X_n} + 1\right)^{-1}$$

pelo que $g(\underline{X})$ *é função das* $n-1$ *v.a.* $\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}$. *Como uma distribuição exponencial pertence à família de escala, segue-se então que* $g(\underline{X})$ *é uma estatística ancilar. Por outro lado* $T(\underline{X}) = \sum_{i=1}^{n} X_i$ *é, pelo teorema anterior, uma estatística completa, pois a exponencial pertence à família exponencial. Então pelo teorema de Basu, vem que* $g(\underline{X})$ *e* T *são independentes entre si, pelo que*

$$\begin{aligned}\frac{1}{\theta} = E[X_n|\theta] &= E[g(\underline{X})T(\underline{X})|\theta] \\ &= E[g(\underline{X})|\theta]E[T(\underline{X})|\theta] \\ &= E[g(\underline{X})|\theta]nE[X|\theta] \\ &= E[g(\underline{X})|\theta]n\frac{1}{\theta}\end{aligned}$$

pelo que $E[g(\underline{X})|\theta] = n^{-1}$.

No caso do espaço paramétrico ser multidimensional, a completude é verificada à custa de uma condição sobre o espaço paramétrico natural (e daí a importância da parametrização natural):

Teorema 22 *A estatística suficiente para a família exponencial* k*-paramétrica é completa se o espaço paramétrico natural contém um rectângulo de* $\mathbb{R}^k$.

Por fim note-se o seguinte resultado:

Teorema 23 *Qualquer estatística suficiente completa é suficiente mínima.*

2.2 Princípio da Verosimilhança

Nesta secção estudamos uma estatística importante conhecida por *função de verosimilhança*, a qual pressupõe igualmente uma forma de redução da informação. A função de verosimilhança tem diversas utilizações (nomeadamente serve como pedra basilar na estimação paramétrica pontual), mas no contexto agora presente serve como instrumento de sumarização dos dados.

2.2.1 Função de verosimilhança

Aquamdo da abordagem de estatísticas suficientes, falou-se várias vezes na *redução dos dados sem perda de informação*. Mas o que significa exactamente a informação a que nos referimos?

Para entender o conceito de informação, teremos de recordar o conceito de *verosimilhança*, conceito chave na inferência estatística. Tome-se a seguinte definição de função de verosimilhança.

Definição 24 *Seja $f_X(.|\theta)$ uma f.d.p. e $\underline{X}$ uma a.a. proveniente de uma população com f.d.p. $f_X(.|\theta)$. Então dado que $\underline{X} = \underline{x}$, a função de θ*

$$L(\theta|\underline{x}) = f_{\underline{X}}(\underline{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

é designada por função de verosimilhança.

Segundo alguns autores, esta definição descura alguns aspectos importantes no caso das funções densidade de probabilidade serem contínuas, mas por hora ignoraremos tais argumentos.

Tome-se, por exemplo, X como sendo uma v.a. discreta, $\underline{x}$ como sendo uma concretização de uma a.a. de dimensão n, e suponhamos que comparamos a função de verosimilhança em dois pontos θ_1 e $\theta_2 \in \Theta$. Se

$$L(\theta_1|\underline{x}) = P(\underline{X} = \underline{x}|\theta_1) > P(\underline{X} = \underline{x}|\theta_2) = L(\theta_2|\underline{x})$$

então a amostra observada $\underline{x}$ é mais provável de ocorrer se $\theta = \theta_1$ que se $\theta = \theta_2$, i.e., sob a observação $\underline{x}$, θ_1 é mais plausível que θ_2.

A definição de função de verosimilhança mostra que numericamente a função de verosimilhança é igual à f.d.p. conjunta de uma amostra. Mas a interpretação é distinta: quando analisamos a f.d.p. conjunta de uma amostra encaramos o parâmetro θ como fixo e $\underline{x}$ como variável; quando analisamos a função de verosimilhança, encaramos θ como variável e $\underline{x}$ como fixo. Assim sendo não se pode dizer que a verosimilhança é uma probabilidade; por exemplo, não faz sentido adicionar verosimilhanças! Na maior parte das aplicações o que faz sentido é comparar verosimilhanças, como vimos no exemplo atrás. Daí que seja sobretudo a razão de verosimilhanças que se estuda, pelo que uma definição mais geral de verosimilhança seria considerar:

$$L(\theta) = K \prod f_X(x_i|\theta)$$

Para simplicar (uma vez que a constante K não tem expressão inferencial), toma-se $K = 1$, forma pela qual a função de verosimilhança é sempre (ou quase sempre) apresentada.

O princípio da verosimilhança traduz uma forma de usar a função de verosimilhança como um reductor e sumarizador da informação contida numa amostra.

Definição 25 *(**Princípio da verosimilhança**): Se $\underline{x}$ e $\underline{y}$ são dois pontos amostrais tais que existe uma constante $C(\underline{x}, \underline{y})$ com*

$$L(\theta|\underline{x}) = C(\underline{x}, \underline{y})L(\theta|\underline{y}), \quad \forall\theta \in \Theta$$

então as conclusões retiradas a partir das amostras $\underline{x}$ e $\underline{y}$ são iguais.

Este princípio diz que se $C(\underline{x}, \underline{y}) = 1$ então $\underline{x}$ e $\underline{y}$ contêm a mesma informação sobre θ. Mas ainda mais: se existe uma proporcionalidade entre as funções de verosimilhança, então a quantidade de informação que retêm sobre o parâmetro é idêntica. Tome-se o seguinte exemplo.

Exemplo 26 *Seja $X_i, i = 1, \ldots, n$ uma a.a. proveniente de uma população $N(\mu, \sigma^2)$, com σ^2 conhecido. Sejam $\underline{x}$ e $\underline{y}$ duas amostras aleatórias. Então*

$$C(\underline{x}, \underline{y}) = e^{-\frac{\sum (x_i - X)^2}{2\sigma^2} + \frac{\sum (y_i - \bar{y})^2}{2\sigma^2}}$$

pelo que o princípio da Verosimilhança diz que as conclusões inferenciais sobre μ para as duas amostras são iguais se $\bar{x} = \bar{y}$.

2.2.2 Informação de Fisher

Uma vez relembrado o conceito de verosimilhança, estamos agora em condições de discutir a *informação* associada a uma dada amostra e estatística..

A taxa relativa de variação da função verosimilhança designa-se por *função score* e é dada por:

$$S(\underline{x}|\theta) = \frac{d \ln L(\theta|\underline{x})}{d\theta}$$

com $S(\underline{x}|\theta) = 0$ para $\underline{x} : L(\theta|\underline{x}) = 0, \forall \theta \in \Theta$. Note-se que $S(\underline{X}|\theta)$ é uma v.a., pois depende da amostra aleatória $\underline{X}$.

A razão desta medida se interpretar como uma medida de informação prende-se com a seguinte interpretação: quanto maior for $S(\underline{x}|\theta)$ para um dado valor de $\theta = \theta_0$, mais fácil é distinguir θ_0 dos valores que lhe estão próximos e maior é, portanto, a informação que em média a observação de X fornece sobre o parâmetro.

2.2.3 Modelos regulares

Considere-se a noção de *identificabilidade* de um modelo.

Definição 27 *Um parâmetro θ de uma família de f.d.p. $\{f_X(.|\theta), \theta \in \Theta\}$ é identificável se valores distintos de θ correspondem a f.d.p. distintas, i.e.*

$$\theta \neq \theta^\star \Rightarrow \exists x \in \Omega_X : f_X(x|\theta) \neq f_X(x|\theta^\star).$$

Uma família com esta propriedade diz-se uma *família regular* ou um modelo regular.

Identificabilidade é uma propriedade de um modelo. Trata-se de uma propriedade desejável, uma vez que num modelo não identificável é muito difícil proceder a métodos inferenciais. Em particular se o modelo não for indentificável, e se $f_X(x|\theta) = f_X(x|\theta^\star), \forall x \in \Omega_X$, então regista-se que $L(\theta|\underline{x}) = L(\theta^\star|\underline{x})$, i.e., a função de verosimilhança é igual nos dois pontos.

Consideremos agora uma família de f.d.p. $\mathcal{F} = \{f_X(.|\theta), \theta \in \Theta\}$ onde Θ é um espaço unidimensional. Diremos que $\mathcal{F}$ *é uma família regular* (ou modelo regular) se as seguintes condições forem satisfeitas:

(a) A família é identificável;

(b) Θ é um intervalo aberto de $\mathbb{R}$;

(c) O suporte do modelo não depende de θ;

(d) A f.d.p. $f_X(.|\theta)$ é diferenciável em todo o $\theta \in \Theta$ e a sua derivada é integrável ao longo de Ω_X;

(e) A permutação do operador $\frac{d}{d\theta}$ com $\int dx$ é válida.

Segue-se da definição de score e da definição de modelo regular, que

$$E[S(\underline{X}|\theta)] = 0$$

Em contrapartida a variância de $S(\underline{X}|\theta)$ é designada por *medida da informação de Fisher* na amostra e é dada por:

$$I_{\underline{x}}(\theta) = E\left[\left(\frac{d \ln L(\theta|\underline{x})}{d\theta}\right)^2 |\theta\right].$$

Quanto maior for $I_{\underline{x}}(\theta)$ mais plausível será distinguir θ dos valores que lhe estão próximos das observações de $\underline{x}$; assim sendo maior será também a informação que a amostra fornece sobre o verdadeiro valor do parâmetro.

A informação de Fisher tem algumas propriedades que facilitam a sua utilização. Assim

(a) *Reparametrização*: se soubermos a informação de Fisher para um dado parâmetro θ, $I_{\underline{x}}(\theta)$, e se $\phi = g(\theta)$ for uma sua reparametrização, então a informação de Fisher para ϕ, $I_{\underline{x}}(\phi)$ relaciona-se com $I_{\underline{x}}(\theta)$ da seguinte forma:

$$I_{\underline{x}}(\phi) = I_{\underline{x}}(g(\theta))[g'(\theta)]^2$$

(b) *Expressão alternativa*: para um modelo regular com informação de Fisher finita, tem-se

$$I_{\underline{x}}(\theta) = E[-\frac{d^2 lnL(\theta|\underline{x}}{d\theta^2}|\theta]$$

(c) *Informação de Fisher para sequências de observações independentes*: se $\underline{x}_i, i = 1, 2, \ldots$ são observações provenientes de amostras independentes então

$$I_{\underline{x}_1,\ldots,\underline{x}_n}(\theta) = \sum_{i=1}^{n} I_{\underline{x}_i}(\theta)$$

Em particular a informação de Fisher de uma amostra de dimensão n é dada por

$$I_{\underline{x}}(\theta) = nI(\theta)$$

onde $I(\theta)$ é a informação de Fisher numa única informação.

(d) *Informação de Fisher para uma estatística*: Se T for uma estatística para θ, então a informação de Fisher para $T = t$, $I_t(\theta)$, está relacionada com a da amostra através de:

$$I_T(\theta) \leq I_{\underline{X}}(\theta) I_{A_t}(x)$$

verificando-se a igualdade se e só se T for uma estatística suficiente para θ.

Os resultados acima referidos podem ser estendidos ao caso de θ ser não unidimensional mas sim multiparamétrico.

2.3 Princípio da Invariância

Os dois princípios anteriores descrevem a redução da dimensionalidade associada a uma amostra de acordo com uma mesma ideia: se $\underline{x}$ e $\underline{y}$ são duas amostras com $T(\underline{x}) = T(\underline{y})$, sendo que T é uma função pré-especificada, então os resultados inferenciais sobre o parâmetro desconhecido θ são iguais quer se use uma amsotra ou outra. No caso de se usar o princípio da Suficiência, fala-se em *suficiência da estatística* T; no caso de se usar o princípio da verosimilhança $T(\underline{x})$ representa o valor comum a todas as amostras que tenham verosimilhança proporcional entre si.

O princípio da invariância afasta-se um pouco desta forma de raciocinar, embora tenha o mesmo objectivo que os anteriores (redução da dimensionalidade sem perda de informação). Assim a função T deve ser pré-especificada mas se $\underline{x}$ e $\underline{y}$ são tais que $T(\underline{x}) = T(\underline{y})$ então o princípio da invariância diz que os resultados inferenciais baseados nas duas amsotras devem ser relacionados entre si mas *não têm de ser necessariamente iguais.*

Existem dois tipos de invariância:

- *Invariância nas medições*: os resultados inferenciais não podem depender da escala em que as medições são feitas.

- *Invariância formal*: Dados dois problemas inferenciais que partilhem o mesmo espaço paramétric Θ e a mesma família de f.d.p. $\{f(.|\theta), \theta \in \Theta\}$, então os mesmos procedimentos inferenciais devem ser utilizados nos dois problemas.

O *princípio da invariância* pode então ser enunciado da seguinte forma:

Definição 28 *Se $\underline{Y} = g(\underline{X})$ é apenas uma mudança de escala no modelo tal que a estrutura formal de $\underline{Y}$ é igual à estrutura formal de $\underline{X}$ então o procedimento inferencial deve ser invariante nas medições e formalmente invariante.*

Vejamos o seguinte exemplo de aplicação do princípio da invariância.

Exemplo 29 *Seja $\{Bin(n,p), p \in (0,1)\}$ a família de f.d.p. binomiais de dimensão amostral n conhecida, e seja $T(x)$ a estimativa de p quando $X = x$, onde $X \sim Bin(n,p)$. Suponhamos agora que em vez de usar X, número de sucessos em n observações i.i.d., usamos Y como sendo o número de insucessos em n observações. Note-se que $Y \sim Bin(n, 1-p)$; seja $T^\star(y)$ uma estimativa de $1-p$ quando $Y = y$. O princípio da invariância nas medições obsriga a que*

$$T(x) = 1 - T^\star(n - x)$$

pois a diferença entre X e Y é meramente uma diferença de escala. Dado que X e Y partilham o mesmo modelo $\{Bin(n,\theta), \theta \in (0,1)\}$, então o principio da invariância formal obriga a que

$$T(z) = T^{\star}(z), \quad z = 0, 1, \ldots, n$$

pelo que os dois princípios em conjunção implicam que

$$T(x) = 1 - T^{\star}(n-x) = 1 - T(n-x)$$

Então o princípio da invariância obriga a que se trabalhe com estimadores T que obedecam a esta propriedade, o que diminui consideravelmente o número de estimadores que podemos utilizar. Por exemplo

$$T_1(x) = \frac{x}{n}, \quad T_2(x) = 0.9\frac{x}{n} + 0.05$$

são ambos estimadores invariantes para este problema. Pelo contrário

$$T_3(x) = 0.8\frac{x}{n} + 0.2$$

não é um estimador invariante para o problema.