1 conceitos básicos

1.1 variável resposta e censuras

Chama-se evento de interesse, aquilo que se deseja encontrar informações sobre a ocorrência. Na análise de sobrevivência, esse evento pode ser a morte de um indivíduo, a cura, um casamento, divórcio ou funcionamento de um dispositivo ou componente de uma máquina. Em análise de sobrevivência, a variável resposta é geralmente o tempo até a ocorrência de um evento de interesse, sendo esse tempo denominado tempo de falha.

A principal característica dos dados de sobrevivência é a presença de censuras, ou seja, observações que por algum motivo não consegue-se determinar o tempo com precisão. Existem três tipos principais de censura, a mais usual é a censura à direita, esta censura acontece quando não se consegue registrar a ocorrência do evento de interesse. Em estudos médicos que analisam o tempo desde a obtenção da doença até a morte do paciente, por exemplo, este tipo de censura acontece quando o paciente é curado, morre por outra razão ou simplesmente não se pode mais observar tal paciente. Este tipo de censura tem três tipos de classificação:

- Censura do Tipo I: acontece quando a pesquisa tem um tempo pré-determinado. Ao final do estudo, as observações que não falharam são consideradas censuras. Nesse tipo de estudo, o percentual de censura é descrito como uma variável aleatória.

- Censura do Tipo II: É encontrada quando se obtém um determinado número de falhas dentro do experimento. Nesse tipo de experimento, o número de falhas deve ser determinado antes de começar o experimento, fazendo com que o número de falhas seja constante. O número de falhas, claramente deve ser menor do que o tamanho da amostra.

- Censura Aleatória: engloba os outros dois tipos de censura. Acontece quando alguns componentes não podem mais ser acompanhados ou quando o motivo da observação falhar é diferente do que interessa. Esta censura ocorre sem intervenção do pesquisador.

Além dessa censura também existem censuras importantes como a censura à esquerda e a censura intervalar. A censura à esquerda ocorre quando o evento ocorre antes do começo do experimento. Por exemplo, deseja-se observar o tempo até uma criança aprender a ler, porém no começo do estudo, algumas crianças podem já ter aprendido a ler sem saber exatamente o tempo em que ela aprendeu, isto é caracterizado como censura à esquerda.

A censura intervalar pode ser dita como um caso genérico das outras censuras. Chama-se censura intervalar quando não se sabe o tempo em que ocorreu o evento de interesse ocorreu, porém sabe-se que ele não ocorreu antes de um determinado tempo. Por exemplo, em um estudo médico é necessário que hajam visitas regulares para a detecção de certas doenças, tal como câncer. Nesse tipo de experimento, sabe-se que a doença apareceu antes do tempo de uma consulta (V), mas também sabe-se que ela apareceu depois de uma consulta (U), ou seja, a doença se manifestou no intervalo [U, V). Quando $V = \infty$ tem-se a censura a direita, e quando a o tempo $U = 0$, essa censura se torna a esquerda. Daí vem o conhecimento de caso genérico da censura.

1.2 Tempo

Como visto anteriormente, a variável resposta do experimento em questão é o tempo até o evento de interesse. Quando esse tempo pode assumir qualquer ponto real não-negativo, descreve-se essa variável como contínua. Esse é o tipo mais comum de variável na análise de sobrevivência, devido a grande diversidade de distribuições contínuas.

Em alguns casos, não faz sentido utilizar uma distribuição contínua para descrever o tempo, mas sim uma discreta. Pode-se ter como exemplo o tempo que um aluno leva para sair da universidade, pode levar 8 semestres, 9 semestres e assim por diante, ou seja, nunca vai levar um tempo real e sim um tempo pertencente aos naturais.

1.3 Funções

1.3.1 Função densidade de probabilidade

Dada uma variável aleatória contínua, não negativa, que represente o tempo de falha de uma observação. Chama-se função densidade, uma função *f*, que descreva a probabilidade de um indivíduo falhar em um intervalo de tempo, quando esse intervalo tende a zero. Sendo assim, essa função descreve a distribuição de probabilidade ao longo do intervalo de zero a infinito.

A partir desta função, é possível obter-se a função de distribuição acumulada, denominada função F. No caso contínuo, esta função é obtida a partir do cálculo da integral da função densidade sobre todo seu suporte. Ou seja, dada uma variável aleatória T:

1.3.2 função densidade de probabilidade

Segundo Meyer (1983), uma função densidade de probabilidade é uma função que que satisfaz as seguintes condições:

i $f(x) \geq 0$ para todo x,

ii $\int_{-\infty}^{+\infty} f(x)dx = 1$,

iii para quaisquer a, b com $-\infty < a < b < +\infty$, teremos $P(a \leq X \leq b) = \int_a^b f(x)dx$.

1.3.3 função de sobrevivência

A função de sobrevivência é definida como a probabilidade de um indivíduo não falhar até um determinado tempo t, ou seja, é a probabilidade de uma observação viver além do tempo t. Dada uma variável aleatória T, contínua, não negativa. Pode-se descrever a função de sobrevivência como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \int_t^\infty f(v)dv \end{aligned} \tag{1}$$

1.3.4 Função de Risco Acumulado

A função de risco acumulado é uma função que não possui uma interpretação simples, porém possui importância dentro do campo da análise de sobrevivência. Esta função, denotada como $H(t)$, pode ser definida como o logaritmo da função de sobrevivência multiplicado por menos um, ou seja:

$$H(t) = -log(S(t)) \tag{2}$$

O gráfico desta função pode assumir algumas diferentes formas. Essas formas são utilizadas para determinar possíveis modelos probabilísticos que melhor se adequam aos dados. Com relação ao comportamento da função, o gráfico pode tomar as seguintes formas:

- Reta diagonal $\Rightarrow$ Função de risco constante é adequada.

- Curva convexa ou côncava $\Rightarrow$ Função risco é monotonicamente crescente ou decrescente, respectivamente.
- Curva convexa e depois côncava $\Rightarrow$ Função risco tem forma de **U**.
- Curva côncava e depois convexa $\Rightarrow$ Função risco tem comportamento unimodal.

1.4 Modelos de Probabilidade

1.4.1 Modelo Weibull

Sendo T uma variável aleatória contínua seguindo uma distribuição Weibull, sua função de densidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp - \left\{ \left(\frac{t}{\alpha} \right)^{\gamma} \right\}, \qquad t \geq 0,$$

Onde γ e α são constantes positivas e correspondem aos parâmetros de forma e escala, respectivamente. O parâmetro α deve ter a mesma unidade de medida que t, enquanto o parâmetro γ não possui nenhuma unidade de medida.

É possível obter a função de sobrevivência para a distribuição Weibull a partir da equação 1:

$$S(t) = \exp \left\{ - \left(\frac{t}{\alpha} \right)^{\gamma} \right\}, \qquad t \geq 0 \tag{3}$$

E através da relação 2, tem-se que:

$$H(t) = - \left(\frac{t}{\alpha} \right)^{\gamma} \tag{4}$$

O parâmetro γ determina a forma da função de risco. Quando o parâmetro é menor que 1, a função de risco é monótona decrescente, caso o parâmetro seja maior que 1, a função de risco é monótona crescente e com o parâmetro assumindo valor igual a 1, a variável toma a forma de uma exponencial, que por sua vez, possui função risco constante.

1.4.2 Modelo Log-Normal

Considerando-se T uma variável aleatória contínua com distribuição Log-Normal, tem sua função densidade de probabilidade como:

$$f(t) = \frac{1}{(2\pi)^{\frac{1}{2}} \sigma t} \exp \left\{ -\frac{1}{2} \left(\frac{\log(t) - \mu}{\sigma} \right) \right\}, \qquad t > 0 \tag{5}$$

Onde μ e σ são, respectivamente, a média e o desvio padrão da variável aleatória. Ambos os parâmetros são maiores que 0.

As funções de sobrevivência e risco acumulado são descritas envolvendo a função distribuição de probabilidade da Normal Padrão devido a função distribuição de probabilidade não ser analiticamente explícita, sendo que normal padrão é descrita como:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ \frac{-u^2}{2\sigma^2} \right\} du$$

A função de sobrevivência da log-normal pode ser facilmente vista como:

$$S(t) = 1 - \Phi \left(\frac{\log(t) - \mu}{\sigma} \right) \tag{6}$$

A partir dessa equação, e da relação 2 é simples encontrar a função risco acumulado.

O parâmetro σ determina a forma da função de risco, sendo que a partir de σ¡1, a função é monótona decrescente e para valores maiores ou iguais a 1 a função é monótona crescente.

1.4.3 Modelo Log-Logístico

Para os casos onde T é uma variável aleatória contínua seguindo uma distribuição Log-Logística, sua função densidade de probabilidade é descrita como

$$f(t) = \frac{\beta \left(\dfrac{t}{\mu}\right)^{\beta-1}}{\mu \left[1 + \left(\dfrac{t}{\mu}\right)^{\beta}\right]^2}, \qquad t > 0 \tag{7}$$

Onde μ e β são constantes, ambas maiores que 0. Com base na função densidade, é possível descrever a função de sobrevivência da variável aleatória T como:

$$S(t) = \frac{1}{1 + \left(\dfrac{t}{\mu}\right)^{\beta}} \tag{8}$$

E a partir desta função pode-se encontrar a função risco acumulado:

$$H(t) = -\log \frac{1}{1 + \left(\dfrac{t}{\mu}\right)^{\beta}} \tag{9}$$

Com relação ao comportamento da função de risco, quando β é menor que 1, esta é monótona decrescente, enquanto para valores maiores que 1 a função tem um comportamento monótono crescente.

1.5 Estimadores da função de sobrevivência

1.5.1 Estimação simples

A função de sobrevivência pode ser estimada amostralmente, como a proporção dos dados que não falharam até o tempo t. Esse estimador poderia ser escrito da seguinte forma:

$$\hat{S}(t) = \frac{n^o \ de \ dados \ com \ tempo \ > \ t}{n^o \ total \ de \ individuos}, \forall \ t \ \in t \geq 0$$

Caso os dados sejam ordenados de forma crescente, pode-se representar a função de sobrevivência da seguinte forma:

$$\hat{S}(t) = \frac{n_j - d_j}{n}$$

Onde n_j é o número de indivíduos que podem falhar, d_j é o número de indivíduos que que falharam no tempo e n é o número total de indivíduos.

1.5.2 Kaplan-Meier

Os estimadores apresentados acima, não podem ser usados nesse tipo de estudo porque não existe nenhuma forma de se incluir censuras.

O estimador a ser usado nesse trabalho será o estimador não-paramétrico de Kaplan-Meier. Esse estimador é muito popular em pesquisas que usam análise de sobrevivência. O estimador é escrito da seguinte forma:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j}$$

Onde, n_j representa o número de dados em risco de falha, d_j são os dados que falharam no tempo t_j, em que, $0 \leq t_{(1)} \leq \ldots \leq t_{(n)}$, são os tempos distintos de falha. Esta técnica não utiliza covariáveis para a estimação, mas pode usar variável categóricas para verificar se as funções estimadas são diferentes.

A representação gráfica desse método se comporta em uma função da forma de escada, uma vez que a estimação entre o tempo $t_{(j)}$ e $t_{(j+1)}$ é constante.