ALGUMAS TÉCNICAS DE ANÁLISE ESTATÍSTICA DE CURVAS DE SOBREVIVÊNCIA E UM EXEMPLO NO ESTUDO DE DIU

Este exemplar corresponde a redação da tese defendida pelo Sr. DJALMA ALVES AGRA e aprovada pela comissão julgadora.

Campinas, de de 1983.

PROF.DR.JOSÉ NORBERTO WALTER DAL#S

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do título de Mestre em Estatística.

Dezembro/1983.

À

Rosa Suzana, pelo apoio e incentivo ao Marcos Roberto, Djalma Junior e Dênnis Rodrigo, pela compreensão.

AGRADECIMENTOS

Sabemos que todo trabalho é fruto do interesse e esforço de um grupo de pessoas. Por esta razão quero lembrar algumas dessas pessoas e instituições que tornaram possível este trabalho:

- *Professor Doutor José Norberto Walter Dachs, orientador e amigo, pelo apoio e sugestões dadas no decorrer deste trabalho.*
- *Professor Doutor Flávio Celso Bartman, pelas sugestões dadas.*
- *Doutor Juan Dias pela cessão dos dados trabalhados.*
- *Sr. Antonio Gouvea, pelo apoio no desenvolvimento dos programas computacionais.*
- *Professor Doutor José Ferreira de Carvalho pela cessão do SAS implantado no IPEN - São Paulo.*
- *Professor Doutor Euclydes Custodio Lima Filho, pelas sugestões dadas.*
- *À Fundação Universidade do Amazonas.*
- *Instituto de Tecnologia da Amazônia - UTAM.*
- *Secretaria de Educação e Cultura do Amazonas.*
- *PICD - CAPES.*
- *Maria de Lourdes, pelo rápido e eficiente trabalho de datilografia.*

A todos os meus sinceros agradecimentos

DJALMA ALVES AGRA

Í N D I C E

INTRODUÇÃO

O estudo do tempo de sobrevivência é de muito interesse em vârias áreas de investigações científicas.

Existem na literatura, vários trabalhos que tratam por exemplo, do estudo do tempo de sobrevivência de pacientes portadores de doenças crônicas onde se aplicam as metodologias descritas neste trabalho.

O objetivo deste trabalho é de apresentar a metodologia, de forma lógica, até chegar ao modelo de regressão de Cox (1972) de forma simples, que possa ser lida por pessoas sem preparo básico em estatística, visando difundí-la, principalmente entre pesquisadores da área médica. Além disso, no presente trabalho, emprega-se o modelo de regressão de Cox em um grupo de pacientes que inseriram um DIU (Dispositivo Intra Uterino). Os dados trabalhados foram coletados pelo Ambulatório de Planejamento Familiar da UNICAMP. Nosso interesse é detetar quais são as covariáveis que contribuem para o sucesso de uma permanência do DIU inserido, por períodos de tempo mais longo.

No primeiro capítulo fazemos uma introdução da análise de sobrevivência, destacando a análise com dados censurados e as funções do tempo de sobrevivência.

No segundo capítulo apresentamos os métodos não paramétricos para estimação das funções de sobrevivência, destacando o estimador do produto limite (Kaplan-Meier, 1958).

No terceiro capítulo, apresentamos o modelo de riscos proporcionais, onde Cox (1972) introduziu uma metodologia que incorpora na análise de curvas de sobrevivência, modelos de regressão.

A análise do procedimento PHGLM (Proportional Hazard General Linear Model) do SAS (Statistical Analysis System) é feita no quarto capítulo, onde definimos as estatísticas que serão usadas na seleção de covariáveis.

A aplicação numérica da metodologia de Cox em um grupo de pacientes que inseriram um DIU (dispositivo intra uterino) está descrita no capítulo cinco. Fazemos também neste capítulo, considerações finais sobre a metodologia usada.

Dois trabalhos recentes, no Brasil, em assuntos ligados ao que se aborda nesta tese foram desenvolvidos por Barreto (1982) e Oliveira (1981) em trabalhos de dissertação de mestrado apresentadas, respectivamente ao IME-USP e ao ICMSC/USP.

CAPÍTULO I

ANÁLISE DE SOBREVIVÊNCIA

1.1.1. PRELIMINARES SOBRE A ANÁLISE DE DADOS DE SOBREVIVÊNCIA

Este capítulo tem como finalidade a análise de dados de sobrevivência, definindo-se as suas principais funções, dando um enfoque ao estudo de dados de sobrevivência em ciência médica. É conveniente afirmar que os métodos usados nos estudos clínicos, são igualmente aplicáveis na indústria, quando por exemplo, desejamos estudar os dados do tempo de sobrevivência de um dispositivo ou de um sistema de componentes eletrônicos. Abordaremos a análise de dados de sobrevivência com censuras.

É importante frisar que, os métodos estatísticos referidos acima, serão usados na análise de dados de sobrevivência derivados de estudos clínicos envolvendo seres humanos. Em razão disto, uma medição dos tempos de sobrevivência dos pacientes envolvidos no estudo, é necessária para que possamos avaliar a eficácia ou não de determinada terapia.

Assim, tempo de sobrevivência pode ser o tempo que vai da entrada de um paciente no estudo até a sua primeira reação, a duração de remissão ou outra função de reação.

Uma observação muito importante no estudo do tempo de sobrevivência, é que o ponto final, não é necessáriamente a falha (morte) do paciente. Ele pode ser caracterizado como uma recaída, o desenvolvimento de um tumor ou a primeira reação a um tratamento.

Até bem pouco tempo, o estudo de dados de sobrevivência era enfocado sob o cálculo da probabilidade de reação ou o tempo médio de vida, comparando-se em seguida, as distribuições de sobrevivência resultantes. Hoje, no entanto, sabe-se que a identificação do risco ou fatores prognósticos ligados ao desenvolvimento de uma doença é igualmente possível e importantíssimo.

É importante destacar neste capítulo uma das mais antigas técnicas estatísticas, que são as tábuas de sobrevivência ou tabelas de vida, de grande utilidade nas áreas atuarial e de saúde.

Segundo Chiang (1968), podemos classificar as tábuas de sobrevivência como: tábuas de vida de coorte e as tábuas de vida corrente. A primeira tábua de vida é também conhecida na literatura como tabela de vida de seguimento, e faz registros da experiência de mortalidade de um grupo de indivíduos durante um certo período de tempo,ou seja, desde o seu "nascimento" até a "morte" do último indivíduo do grupo. Por outro lado, uma tabela de vida corrente, registra a experiência de mortalidade da população toda em um curto período de tempo, por exemplo, um período

de um ano.

É importante observar que tanto as tabelas de vida de coorte como as tabelas de vida corrente, registram a experiência de mortalidade que podem ser decompostas em várias causas de mortes. Este estudo é conhecido como estudos de seguimento sob consideração de riscos competitivos, Chiang (1968).

Na prática, o pesquisador médico depara com o seguinte fato: há pacientes que em dado momento do estudo desaparecem sem fazer nenhuma notificação. Este problema conhecido como censura, será tratado na próxima seção.

1.1.2. ANÁLISE DE SOBREVIVÊNCIA COM DADOS CENSURADOS

Quando um pesquisador tenta construir uma técnica estatística para trabalhar com uma determinada variável aleatória, comumente ele se baseia em afirmações sobre a natureza da função de distribuição restringindo-se a uma família de distribuições indexada por um parâmetro ou um vetor de parâmetros. Acontece que, na prática, nem sempre consegue-se especificar as principais características da distribuição, isto é, não fica clara sua forma e parâmetros. Quando esta suposição é válida, diz-se que temos um modelo com a distribuição desconhecida ou não paramétrico. Na análise de dados de sobrevivência, infelizmente, alguns tempos de sobrevivência, como descritos na seção 1.1.1, não

satisfazem essas condições, ou seja, os tempos de sobrevivência de todos os pacientes no estudo não são exatos e conhecidos.

O pesquisador médico quase sempre tem que estipular o período de tempo para executar determinado tratamento em um grupo de pacientes. Assim é possível admitir pacientes em diferentes fases do estudo. O objetivo é observar os seus tempos de sobrevivência. Porém, em dados de sobrevivência, existem situações onde perde-se informações sobre determinados pacientes sendo impossível precisar o seu tempo de sobrevivência. Pacientes nestas situações são tratados como *censura*. As *censuras* podem ocorrer das seguintes formas:

i) *Perda de acompanhamento*. Neste caso o paciente sai do estudo, perdendo-se qualquer contacto com ele depois de um período de estudos, seja porque ele mudou o local de residência ou porque perdeu o interesse no estudo.

ii) *Desistência de pacientes*. A terapia usada pode ter efeitos negativos, de modo que é necessário parar o tratamento. Em outra situação, o paciente pode ainda estar em contacto com o médico, mas ele se recusa a continuar o tratamento.

iii) *Término de estudo*: Neste caso o paciente ainda está vivo no final do estudo, sendo impossível determinar o seu tempo de sobrevivência.

Em seguida, para sedimentar a idéia de censura em estudos clínicos, reproduzimos um exemplo gráfico dado por R. Miller (1980).

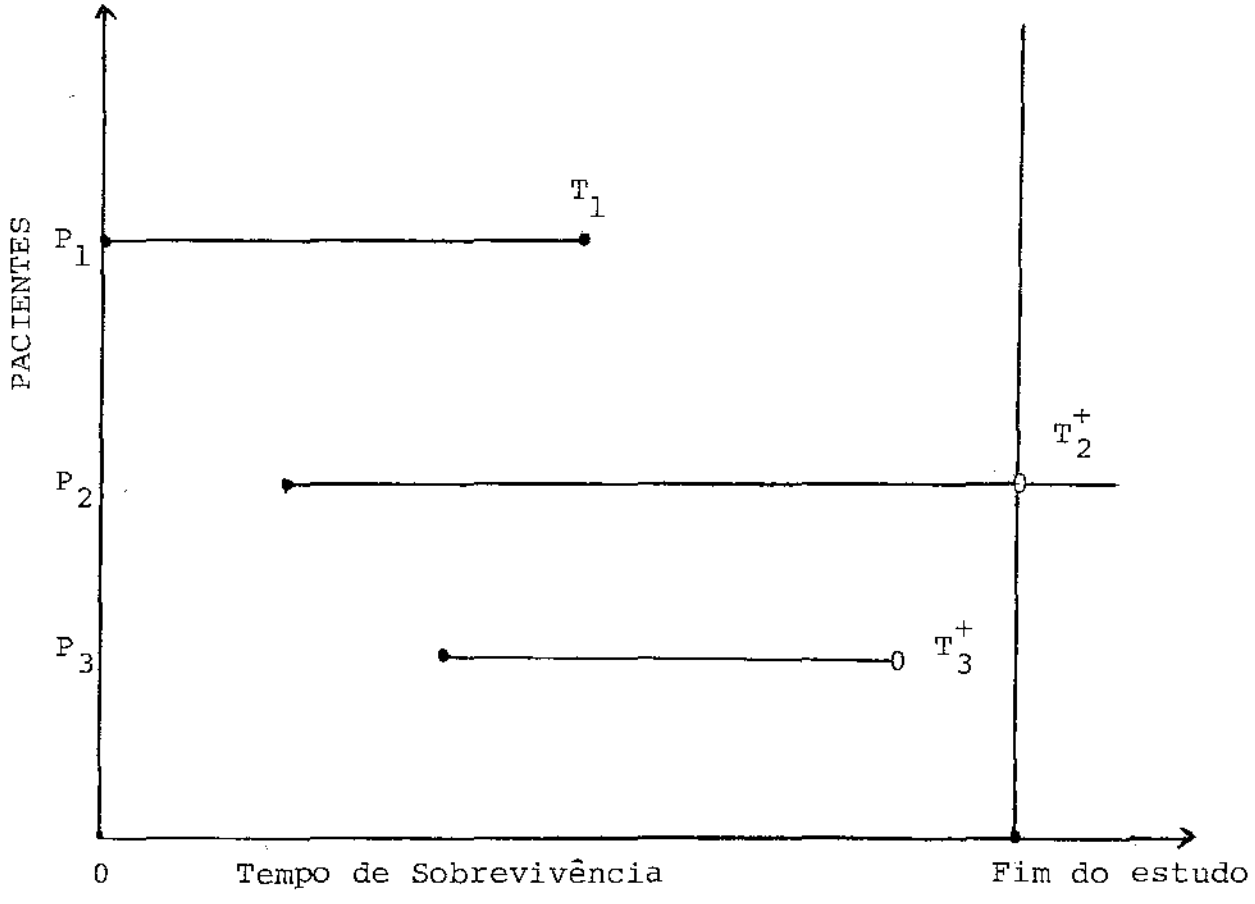


FIGURA 1

A figura 1 nos mostra P_1, paciente 1, aceito no estudo no tempo $t = 0$ e morto em T_1, caracterizando uma observação não censurada; o segundo paciente, caracterizado por P_2 entrou no estudo em andamento e vivo ao fim do estudo, aqui término é equivalente a morte, resultando em uma observação censurada T_2^+ (onde o símbolo "+" significa dado com censura), e o paciente P_3 aceito no estudo em andamento foi contactado pela última vez em T_3^+, o que caracteriza outra observação censurada.

Aqui surge uma pergunta muito natural, como é registrado o tempo de sobrevivência de um paciente censurado? Nos casos acima, por exemplo, o tempo de sobrevivência do paciente P_3 é considerado o tempo de sua entrada no estudo até o último contacto com o médico e no caso de P_2 é evidente que o tempo de sobrevivência, seja o da entrada até o término dos estudos.

Computados os tempos de sobrevivência dos pacientes no experimento clínico, o próximo passo será definirmos as principais funções de sobrevivência.

Na próxima seção definiremos as principais funções do tempo de sobrevivência que serão importantes no nosso trabalho.

1.1.3. FUNÇÕES DO TEMPO DE SOBREVIVÊNCIA

Esta seção tem como objetivo definir as principais funções de sobrevivência que no decorrer do trabalho serão uma ferramenta importantíssima.

O pesquisador tem interesse no comportamento da distribuição do tempo de sobrevivência, que pode ser caracterizada por três funções: a função sobrevivência, a função densidade de probabilidade e a função risco. Existe uma relação matemática entre essas três funções, ou seja, se uma dessas funções é dada, as outras duas podem ser obtidas a partir da primeira.

O problema básico do pesquisador na análise de dados de sobrevivência, portanto, é estimar dos dados amostrados uma das três funções citadas acima e em seguida fazer inferências sobre o modelo da população ou parâmetros.

1.2.1. DEFINIÇÕES

Considere-se uma amostra de indivíduos de uma população. Seja $T > 0$ uma variável aleatória que representa o tempo de sobrevivência de um indivíduo.

A distribuição de T pode ser caracterizada pelas seguintes funções:

1.2.2. FUNÇÃO SOBREVIVÊNCIA

Esta função, denotada por $S(t)$, é definida como a probabilidade que um indivíduo sobreviva mais do que t, isto é:

$$S(t) = P(\text{um indivíduo sobreviva mais que } t) = P(T > t). \quad (1.1)$$

Podemos ainda escrever a função sobrevivência $S(t)$, usando a definição da função distribuição $F(t)$ de t, ou seja,

$$S(t) = P(T > t) = 1 - P(T \leq t)$$

$$= 1 - P\{\text{um indivíduo falhar (morrer) antes de } t\}$$

$$= 1 - F(t) \ . \quad (1.2)$$

Em termos práticos, na ausência de censuras, estimamos a função de sobrevivência como a proporção de indivíduos que sobreviveram mais do que t, sobre o total de indivíduos no estudo, ou seja, usando a função de distribuição empírica,

$$\hat{S}(t) = \frac{\text{nº de indivíduos que sobreviveram } t}{\text{nº total de indivíduos em estudo}} \quad (1.3)$$

onde $\hat{S}(t)^*$, representa um estimador da função sobrevivência $S(t)$.

---

* leia-se S circunflexo

No exemplo 1.1 mostramos como calculamos $\hat{S}(t)$.

A função sobrevivência $S(t)$, é uma função não crescente no tempo t com as seguintes propriedades:

$$S(t) = 1 \quad \text{para} \quad t = 0$$

e

$$\lim S(t) = 0 \quad \text{se} \quad t \to \infty$$

em palavras, ela nos diz que todo indivíduo está vivo no começo do estudo e que ninguém sobrevive um tempo infinito.

De um modo geral, é muito importante esboçar graficamente a função sobrevivência $S(t)$. O gráfico da função de sobrevivência $S(t)$ é chamado de *curva de sobrevivência*.

A importância da curva de sobrevivência é que ela nos fornece a olho nu, possibilidades de compararmos por exemplo, dois ou mais tipos de tratamentos. Explicando de uma forma mais simples, uma curva de sobrevivência com declive acentuado, representa baixa taxa de sobrevivência ou curto tempo de sobrevivência, enquanto que se essa curva é menos acentuada, ela representa alta taxa de sobrevivência ou longa sobrevivência.

A curva de sobrevivência além de ser usada para comparar dados de sobrevivência entre dois ou mais grupos, é usada também para acharmos os percentis do tempo de sobrevivência, em especial, o 50º (a mediana). Neste tipo de análise usa-se a mediana porque a média é pouco resistente. Explicando, se houver no

experimento um ou mais indivíduos com tempo de vida excepcionalmente curto ou excepcionalmente longo, eles afetarão sensivelmente, o tempo médio de vida.

1.2.3. A FUNÇÃO DENSIDADE DE PROBABILIDADE

Pode-se supor em geral, que o tempo de sobrevivência T tem distribuição absolutamente contínua e existe então função densidade de probabilidade $f(t)$, ou por simplicidade de notação, a função densidade, que é definida como o limite da probabilidade que um indivíduo falhe (morra) no intervalo de t a $t+\Delta t$, dividido pelo incremento de tempo Δt, isto é,

$$f(t) = \lim_{\Delta t \to 0} \frac{P\{\text{um indivíduo morrer no intervalo } t, t+\Delta t\}}{\Delta t} \qquad (1.4)$$

Em um estudo clínico, de modo prático, a função densidade $f(t)$, é estimada como a proporção de pacientes mortos em um intervalo começando em t, dividido pelo total de pacientes em estudo, vezes o comprimento do intervalo, ou seja,

$$\hat{f}(t) = \frac{\text{nº de pacientes mortos no intervalo começando em } t}{(\text{nº total de pacientes})(\text{comprimento do intervalo})} \qquad (1.5)$$

onde $\hat{f}(t)$ é um estimador da função densidade $f(t)$.

A representação gráfica da função densidade f(t), é denominada *curva de densidade*.

A função densidade possui as seguintes propriedades:

a) f(t) é uma função não negativa, isto é,

$$f(t) \geq 0 \qquad \text{para} \qquad t \geq 0$$
$$f(t) = 0 \qquad \text{para} \qquad t < 0$$

b) a área contida entre a curva de densidade e o eixo do tempo t, é igual a um.

De modo análogo ao que foi feito com a função sobrevivência S(t), a análise gráfica da curva de densidade nos dá a razão de falha, bastando relacionar o intervalo de tempo com os picos da frequência.

A função densidade f(t), é conhecida na literatura também como razão de falha incondicional.

1.2.4. A FUNÇÃO RISCO

A função risco $\lambda(t)$, também conhecida como razão de falha instantânea do tempo de sobrevivência T, é definida como a probabilidade de falha de um indivíduo durante um intervalo de tempo, dado que ele seja sobrevivente no começo do intervalo, ou seja,

$\lambda(t)$ = P{um indivíduo falhar (morrer) no intervalo $(t, t + \Delta t)$ dado que ele seja sobrevivente no tempo t}

$$= P\{t < T < t + \Delta t \;/\; T > t\}/\Delta t \qquad (1.6)$$

Na literatura atuarial, $\lambda(t)$ é chamada de *força de mortalidade*.

A razão de falha instantânea $\lambda(t)$ é também definida em termos da função distribuição $F(t)$ e a função densidade $f(t)$, ou seja,

$$\lambda(t) = \frac{f(t)}{1 - F(t)} . \qquad (1.7)$$

Em termos práticos, a razão de falha instantânea $\lambda(t)$ é estimada como sendo a proporção de indivíduos que falham (morrem) no intervalo de tempo $(t, t + \Delta t)$, dado que eles tenham sobrevivido até o início do intervalo, isto é,

$$\hat{\lambda}(t) = \frac{m(t)}{r(t)} \qquad (1.8)$$

onde $m(t)$ denota o número de indivíduos mortos no intervalo de tempo $(t, t + \Delta t)$ e $r(t)$ denota o número de indivíduos sobreviventes restantes no começo do intervalo de tempo t.

Os atuários usam sistematicamente a razão risco médio do intervalo, onde o número de indivíduos mortos no intervalo é

dividido pelo número médio de sobreviventes no ponto médio do intervalo

$$\hat{\lambda}(t) = \frac{\text{nº de indivíduos mortos no intervalo}}{\text{(nº de indivíduos sobreviventes em t)} - \frac{1}{2}\text{(nº de mortes no intervalo)}} \qquad (1.9)$$

A função risco $\lambda(t)$ pode ser crescente, decrescente ou permanecer constante ou ser combinação destas. Para ilustrar este fato vejamos as seguintes situações.

Primeiro, pacientes portadores de leucemia aguda que não respondem a terapia indicada, tem uma razão de risco crescente. Para ilustrar a segunda afirmação vejamos um paciente ferido à bala, ele tem razão de risco decrescente, pois, o maior perigo é a operação a que se submete o paciente e este perigo diminue se a cirurgia é bem sucedida. Finalmente para exemplificar uma função de risco constante, pode-se considerar o risco de indivíduos sãos entre os 18 e 40 anos de idade, onde o principal risco de morte são os acidentes naturais.

Daremos agora um exemplo para ilustrar o cálculo das funções introduzidas. A tabela 1.1, em suas três primeiras colunas, contêm os dados de sobrevivência de 40 pacientes com mieloma (Lee,1980). Os tempos foram agrupados em intervalos de cinco meses.

A função sobrevivência estimada $\hat{S}(t)$ é calculada segundo a equação 1.3. Por exemplo, no fim do segundo intervalo, 28

dos 40 pacientes ainda estavam vivos. Assim, $\hat{S}(10)=28/40=0.700$.

Do mesmo modo a função densidade estimada $\hat{f}(t)$ é computada segundo a equação 1.5. Por exemplo, a função densidade no terceiro intervalo é $6/(40 \times 5) = 0.030$.

A função risco estimada, $\hat{\lambda}(t)$, é calculada segundo o método dado pela equação 1.9. Por exemplo a função risco estimada do primeiro intervalo é $5 / [5(40 - \frac{5}{2})] = 0.027$.

TABELA 1.1

DADOS DE SOBREVIVÊNCIA E FUNÇÕES DE SOBREVIVÊNCIA ESTIMADAS DE 40 PACIENTES COM MIELOMA

Tempo de Sobrevivência t(meses)	Nº de pacientes vivos no começo do intervalo	Nº de pacientes mortos no intervalo	$\hat{S}(t)$	$\hat{f}(t)$	$\hat{\lambda}(t)$
0 ⊢ 5	40	5	0.875	0.025	0.027
5 ⊢ 10	35	7	0.700	0.035	0.044
10 ⊢ 15	28	6	0.550	0.030	0.048
15 ⊢ 20	22	4	0.450	0.020	0.040
20 ⊢ 25	18	5	0.325	0.025	0.065
25 ⊢ 30	13	4	0.225	0.020	0.072
30 ⊢ 35	9	4	0.125	0.020	0.114
35 ⊢ 40	5	0	0.125	0.000	0.000
40 ⊢ 45	5	2	0.075	0.010	0.100
45 ⊢ 50	3	1	0.050	0.005	0.080
$\geq$ 50	2	2	0.000	-	-

FONTE: Lee, E. (1980).

Olhando para a curva de sobrevivência, figura 1.a, vê-se que o tempo de sobrevivência mediano dos pacientes com mieloma é aproximadamente 17.5 meses. Do mesmo modo, o pico de alta frequência de morte, figura 1.b, ocorre no intervalo de 5 a 10 meses. Finalmente a figura 1.c esboça o comportamento da função risco, ou seja, mostra uma tendência crescente e alcança seu pico em aproximadamente 33 meses, e então oscila. A seguir derivamos algumas relações importantes das funções de sobrevivência.

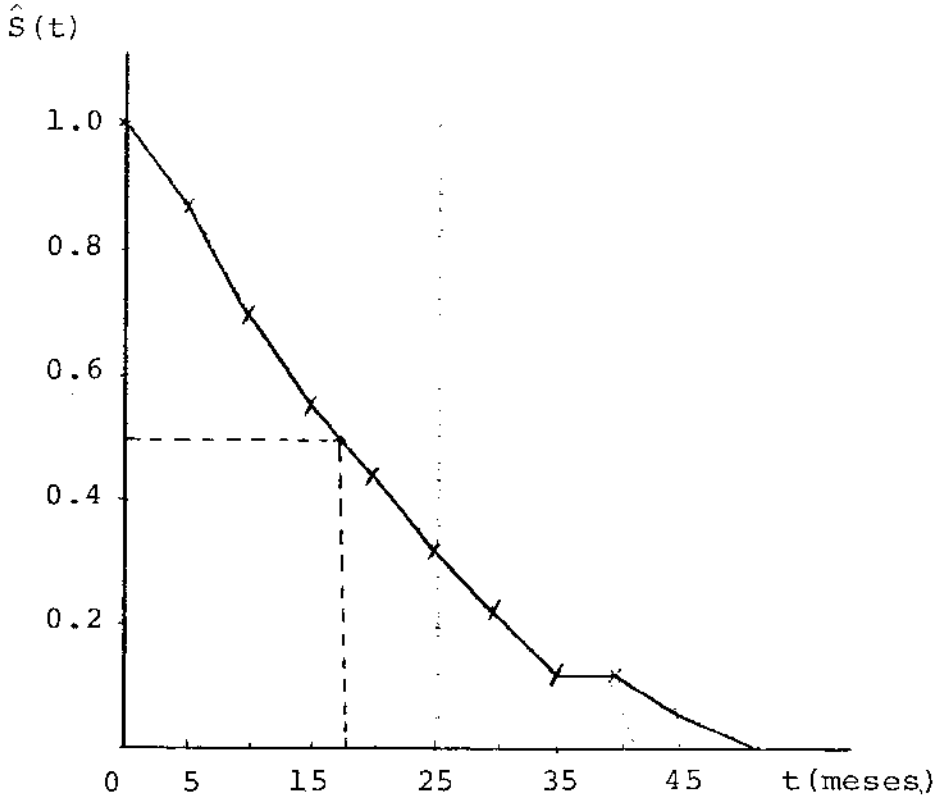


FIGURA 1.a - FUNÇÃO DE SOBREVIVÊNCIA ESTIMADA $\hat{S}(t)$ DE PACIENTES COM MIELOMA.

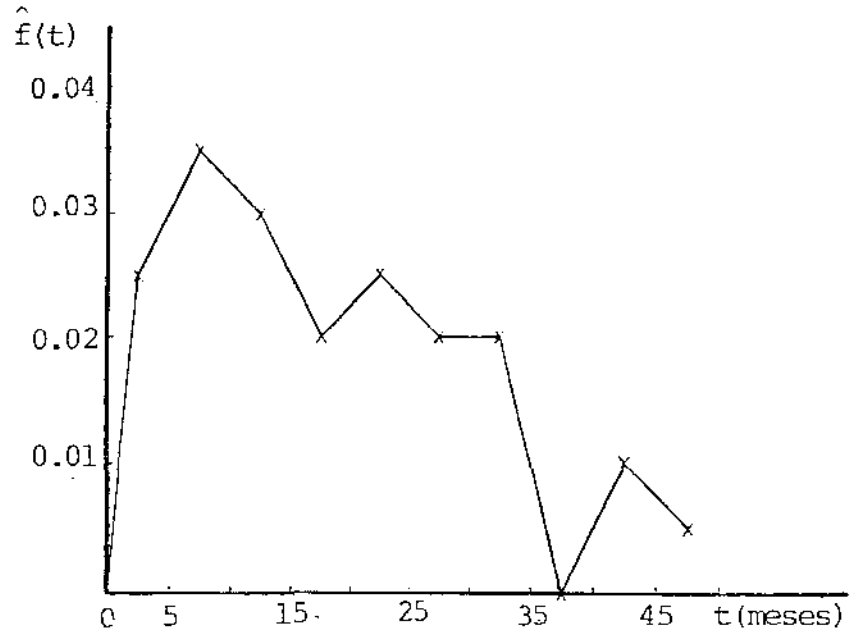


FIGURA 1.b - FUNÇÃO DENSIDADE ESTIMADA $\hat{f}(t)$ DE PACIENTES COM MIELOMA.

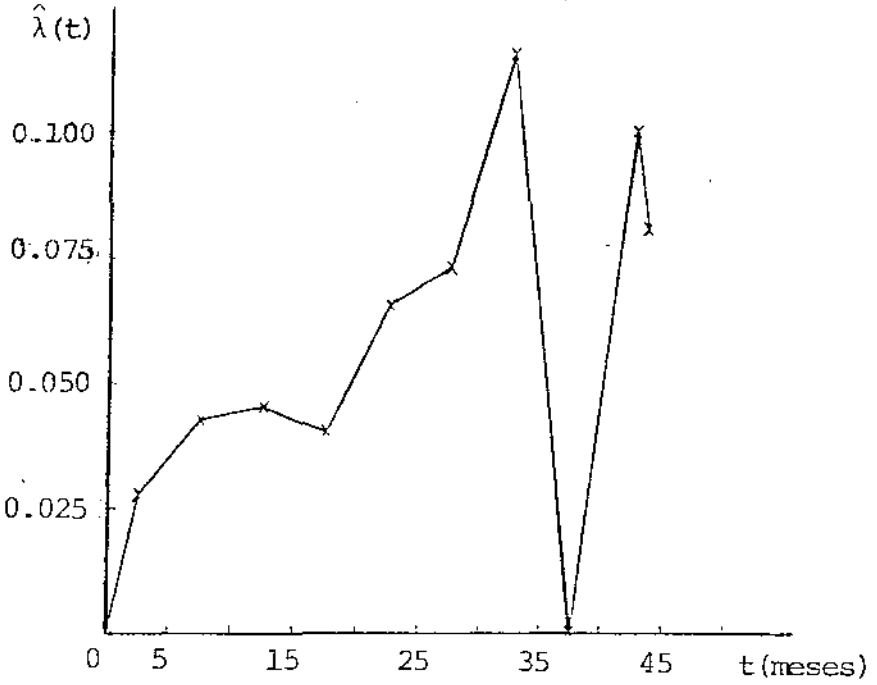


FIGURA 1.c - FUNÇÃO RISCO ESTIMADA $\hat{\lambda}(t)$ DE PACIENTES COM MIELOMA.

1.2.5. RELAÇÕES DAS FUNÇÕES DE SOBREVIVÊNCIA

Esta seção tem como objetivo mostrar que as três funções de sobrevivência definidas anteriormente, guardam entre si, uma relação matemática, ou seja, dada uma delas as outras podem ser facilmente obtidas da primeira.

A primeira relação: A função risco $\lambda(t)$ pode ser obtida das equações (1.2) e (1.7),

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \qquad (1.10)$$

Esta relação também pode ser obtida da equação (1.6).

A segunda relação: Sabe-se que a função densidade $f(t)$ pode ser obtida, derivando-se a função distribuição $F(t)$, isto é,

$$f(t) = \frac{d}{dt}[F(t)] = \frac{d}{dt}[1 - S(t)] = - S'(t) \qquad (1.11)$$

onde $S(t) = 1 - F(t)$, pela equação (1.2).

A *terceira relação*: Uma outra forma de calcularmos a função risco.

Substituindo a equação (1.11) na equação (1.10), obtem-se:

$$\lambda(t) = - \frac{S'(t)}{S(t)} = - \frac{d}{dt} \log_e S(t) \ , \qquad (1.12)$$

que em palavras diz que a derivada do logaritmo da função de sobrevivência, vezes menos um, é igual a função risco.

A *quarta relação*: Outra expressão para calcularmos a função sobrevivência. Sabemos que $\lambda(t) > 0$. Então integrando a equação (1.10) de 0 a t e usando o fato que $S(0) = 1$, obtemos:

$$\int_0^t \lambda(u)\,du = \int_0^t \frac{f(u)}{1 - F(u)}\,du =$$

$$= - \log [1 - F(u)] \Big|_0^t$$

$$= - \log [1 - F(t)]$$

$$\int_0^t \lambda(u)du = - \log S(t)$$

assim, obtemos a importantíssima relação

$$S(t) = e^{-\int_0^t \lambda(u)du} \tag{1.13}$$

A quinta relação: A função densidade pode ser determinada das equações (1.10) e (1.13), ou seja,

$$f(t) = \lambda(t)\, e^{-\int_0^t \lambda(u)du} \tag{1.14}$$

Em seguida, através do exemplo 1.1, ilustramos como podem ser obtidas algumas das relações descritas acima.

EXEMPLO 1.1. Vamos supor que o tempo de sobrevivência T de uma população tem a seguinte função densidade de probabilidade

$$f(t) = \begin{cases} \theta e^{-\theta t} & \text{se} \quad t \geq 0\,, \quad \theta > 0 \\ 0 & \text{se} \quad t < 0. \end{cases}$$

Nosso primeiro passo será encontrar a função distribuição $F(t)$ e então achar $S(t)$ usando a equação (1.2). Mas

$$F(t) = \int_0^t f(u)du$$

$$= \int_0^t \theta e^{-\theta u}\, du$$

$$= -\theta \left. \frac{e^{-\theta u}}{\theta} \right|_0^t$$

assim

$$F(t) = 1 - e^{-\theta t} \quad \text{para} \quad t \geq 0$$

logo, pela equação (1.2), a função sobrevivência é:

$$S(t) = 1 - F(t) = e^{-\theta t} \quad \text{para} \quad t \geq 0.$$

Para encontrarmos a função risco $\lambda(t)$, podemos usar a equação (1.10) ou a equação (1.12).

Usando a terceira relação tem-se:

$$\lambda(t) = -\frac{d}{dt} \log S(t) = -\frac{d}{dt} \log [e^{-\theta t}]$$

$$= -\frac{d}{dt} [-\theta t] = \theta \quad .$$

Do exemplo 1.1., função densidade de uma exponencial, vê-se que a sua função risco é o seu próprio parâmetro, ou seja, o risco, desta função independe do tempo t.

CAPÍTULO II

MÉTODOS NÃO PARAMÉTRICOS PARA ESTIMAÇÃO DAS FUNÇÕES DE SOBREVIVÊNCIA

2.1.1. PRELIMINARES

Este capítulo tem como objetivo discutir os principais métodos para estimar as funções de sobrevivência introduzidas no capítulo anterior, para dados com censuras. No exemplo da tabela 1.1, quando estimamos as funções de sobrevivência, não tínhamos a presença de censuras, ou seja, os tempos exatos de sobrevivência, eram conhecidos para todos os casos. Quando acontece o contrário, isto é, não temos os tempos exatos de vida dos pacientes em estudo, temos que usar os métodos não paramétricos, que no geral são fáceis de compreender e aplicar. Os dois métodos não paramétricos que discutiremos neste capítulo são: AS TABELAS DE VIDA e o ESTIMADOR DE KAPLAN-MEIER.

Já falamos da importância das tabelas de vida no capítulo I e na apresentação do primeiro método, seção 2.2.1, construimos uma tabela e explicamos alguns detalhes dos cálculos.

Na seção 2.2.4, apresentamos o segundo método, ilustrando a apresentação com um exemplo.

2.2.1. TABELAS DE VIDA

Uma das mais antigas técnicas usadas em análise de sobrevivência, é o clássico método de estimarmos a função de sobrevivência S(t), em estudos de epidemiologia e ciência atuarial (crescimento populacional, fertilidade, migrações, seguros), através da chamada tabela de vida. Melhor explicando, a tabela de vida é um instrumento para se poder estimar a função de sobrevivência S(t).

A construção da tabela de vida requer um número grande de observações, de modo que os tempos de sobrevivência possam ser agrupados em intervalos de classe. Esses intervalos são quase sempre, mas não necessariamente, de comprimentos iguais.

A tabela 2.1, CUTHER e EDERER 1958, ilustram os componentes de uma tabela de vida. Aproveitamos a construção da tabela e exemplificamos o cálculo da taxa de sobrevivência de S(t).

Para a construção das tabelas de vida existem basicamente dois métodos, que discutiremos nas seções 2.2.2 e 2.2.3, respectivamente.

TABELA 2.1

CÁLCULO DA FUNÇÃO DE SOBREVIVÊNCIA EM PACIENTES COM MIELOMA

Anos depois do Diagnóstico.	Vivos no início do intervalo	Mortos durante o intervalo	Perdas do seguimento no intervalo	Retirados vivos durante o intervalo	Nº efetivo exposto no risco morte	Proporção de mortos	Proporção de sobreviventes
I_i	n_i	d_i	ℓ_i	w_i	N_i	$\hat{q}_i$	$\hat{p}_i$
(0 - 1]	126	47	4	15	116.5	0.40	0.60
(1 - 2]	60	5	6	11	51.5	0.10	0.90
(2 - 3]	38	2	-	15	30.5	0.07	0.93
(3 - 4]	21	2	2	7	16.5	0.12	0.88
(4 - 5]	10	-	-	6	7.0	0.00	1.00

FONTE: CUTLER e EDERER, J. CHRONIC DIS. (1958).

2.2.2. MÉTODOS DA AMOSTRA REDUZIDA

Este método é o mais simples e consiste em estimar a função sobrevivência $S(t_k)$, usando somente aqueles indivíduos que estão em risco durante o intervalo $(0,t_k]$, e que representam a entrada no intervalo de interesse.

Seja

$$n_{(k)} = n_1 - \sum_{i=1}^{k} \ell_i - \sum_{i=1}^{k} w_i \qquad (2.1)$$

e seja

$$d_{(k)} = \sum_{i=1}^{k} d_i \qquad (2.2)$$

onde, na tabela 2.1,

I_i = o intervalo $(t_{i-1}, t_i]$;

n_i = nº de indivíduos vivos no início de I_i

d_i = nº de mortes ocorridas durante I_i

ℓ_i = perdas no seguimento durante I_i

w_i = nº de retiradas durante I_i.

Assim o estimador da função sobrevivência $S(t_k)$ é dado por:

$$S(t_k) = 1 - \frac{d_{(k)}}{n_{(k)}} \; . \qquad (2.3)$$

Para melhor compreensão da metodologia vamos usar o seguinte exemplo.

EXEMPLO 2.1. Vamos supor que desejamos achar o valor de $\hat{S}$(5 anos) ou seja, queremos estimar a probabilidade de um paciente em

estudo sobreviver 5 ou mais anos, no caso da Tabela 2.1.

Claramente da Tabela 2.1 obtemos

$$n_{(5)} = 126 - 12 - 54 = 60$$

e

$$d_{(5)} = 56$$

assim, pela equação (2.3), obtêm-se:

$$\hat{S}(5 \text{ anos}) = 1 - \frac{56}{60} = 0.06667$$

isto é, a probabilidade de um paciente sobreviver, neste caso, 5 ou mais anos, é 6,6%.

O método da amostra reduzida, é pouco usado. Uma desvantagem do método é que ele ignora as informações que estão contidas em ℓ_i e w_i.

Uma observação que deve ser feita é que este método só é válido quando não há censura nos dados, ou seja, todos os pacientes no estudo são acompanhados desde sua entrada até o final deste estudo.

2.2.3. O MÉTODO ATUARIAL

A função de sobrevivência $S(t_k)$ definida pela equação 1.2, pode ser escrita como o produto encadeado das probabilidades de sobrevivência, isto é,

$$S(t_k) = P(T > t_k)$$

$$= P(T > t_1) \cdot P(T > t_2 \mid T > t_1) \cdot P(T > t_3 \mid T > t_2) \ \ldots$$

$$\ldots \ P(T > t_{k-1} \mid T > t_{k-2}) \cdot P(T > t_k \mid T > t_{k-1})$$

$$= p_1 \cdot p_2 \cdot p_3 \cdot \ldots \cdot p_{k-1} \cdot p_k$$

onde $p_i = P(T > t_i \mid T > t_{i-1})$ é a probabilidade de sobreviver ao intervalo $(t_{i-1}, t_i]$.

Como podemos observar, o método atuarial nos dá um valor estimado para cada p_i separadamente e em seguida multiplica os valores estimados dos p_i para se estimar a função de sobrevivência $S(t_k)$.

Quando nos dados trabalhados não há perdas ou saídas, dentro de cada intervalo I_i, podemos estimar os p_i usando a fórmula (2.3).

De outro modo, com ℓ_i e w_i diferentes de zero, supomos

que, em média aqueles individuos que vieram a ser perdidos ou sairam durante o intervalo I_i, estavam em risco na metade do intervalo I_i. Por esta razão definimos o tamanho efetivo da amostra como:

$$N_i = n_i - \frac{1}{2}(\ell_i + w_i) \quad . \tag{2.4}$$

Definimos também o estimador de q_i, a proporção de mortos, como sendo

$$\hat{q}_i = \frac{d_i}{N_i} \quad . \tag{2.5}$$

De modo análogo, definimos o estimador de p_i, a proporção de sobreviventes, como sendo:

$$\hat{p}_i = 1 - \hat{q}_i \quad . \tag{2.6}$$

De posse da equação 2.6 definimos o estimador atuarial de $S(t_k)$ como:

$$\hat{S}(t_k) = \prod_{i=1}^{k} \hat{p}_i \quad . \tag{2.7}$$

Vejamos um exemplo ilustrativo.

Na Tabela 2.1, a coluna 6 contém N_i, a coluna 7 contém os $\hat{q}_i$ e a coluna 8 contém, obviamente os $\hat{p}_i$.

Assim

$$\hat{S}(5 \text{ anos}) = \prod_{i=1}^{5} \hat{p}_i = 0.44$$

em palavras, a coluna 6 é calculada pela equação 2.4 , a coluna 7 é calculada usando a equação 2.5 e a coluna 8 é calculada pela equação 2.6 .

Segundo R. Miller (1980), a estimativa da variância $S(t_k)$ é:

$$\hat{Var}(\hat{S}(t_k)) = \hat{S}^2(t_k) \sum_{i=1}^{k} \frac{d_i}{N_i(N_i - d_i)} \qquad (2.8)$$

com d_i definido na seção 2.2.1 e N_i é calculado da equação 2.4.

A necessidade do cálculo da variância surge quando precisamos construir um intervalo de confiança para a função de sobrevivência, assim como para verificar a homogeneidade dos dados trabalhados.

2.2.4. O ESTIMADOR DE KAPLAN-MEIER.

O método atuarial discutido na seção anterior, como pudemos observar, não faz nenhuma suposição paramétrica a respeito da forma da distribuição, porém considera o emprego de observações censuradas.

Um estimador simples da função de sobrevivência $S(t_k)$, mais estudado e usado atualmente, envolvendo um grupo homogêneo de indivíduos com dados censurados, foi formulado por Kaplan e Meier em 1958, herdando o nome dos autores. O estimador de Kaplan-Meier é conhecido na literatura como estimador do produto limite.

Vejamos algumas características do estimador do produto limite e semelhanças com o método atuarial. Primeiro, ele é um estimador de máxima verossimilhança. Assim, o estimador de Kaplan-Meier é um estimador suficiente. Breslow e Crowley (1974) mostraram que é um estimador não viciado. Alguns autores afirmam que o estimador do produto limite é similar ao estimador atuarial, exceto que os comprimentos dos intervalos I_i são variáveis. Outra diferença do estimador do produto limite com o estimador atuarial definido na seção 2.2 é que ele não requer preliminarmente o cálculo de $\lambda(t)$. Em seguida construimos o estimador do produto limite.

Considere $t_1, t_2, \ldots, t_n$ os tempos de falha ou censura de n indivíduos no estudo.

Seja $t_{(1)} < t_{(2)} < t_{(3)} < \ldots < t_{(n)}$, os tempos de falha ou censura ordenados dos indivíduos na amostra. No caso de não haver censura, isto é, os tempos de falhas são exatamente conhecidos, a função sobrevivência no tempo $t_{(i)}$ pode ser estimada como:

$$\hat{S}(t_{(i)}) = \frac{n-i}{n} = 1 - \frac{i}{n}$$

onde (n-i) é o número de indivíduos sobreviventes na amostra. Em outras palavras n é o número de pacientes em risco e i é o número de falhas no tempo $t_{(i)}$, respectivamente.

Quando temos falhas e censuras podemos escrever o estimador de Kaplan-Meier, usando os estimadores de p_i e q_i, respectivamente deduzidos na seção anterior, isto é, o estimador do produto limite é

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \hat{p}_i = \prod_{t_{(i)} \leq t} (1 - \hat{q}_i)$$

$$= \prod_{t_{(i)} \leq t} \left(1 - \frac{1}{n_i}\right)$$

$$= \prod_{t_{(i)} \leq t} \left(1 - \frac{1}{n-i+1}\right)$$

$$= \prod_{t_{(i)} \leq t} \frac{n-i}{n-i+1} \quad . \qquad (2.9)$$

Em termos práticos, o estimador do produto limite, tem seu cálculo facilitado, construindo-se uma tabela com as seguintes colunas:

a) COLUNA 1: esta coluna deve conter todos os tempos de sobrevivência, sejam eles censurados ou não censurados, ordenados do menor ao maior valor. Nas observações censuradas devemos afixar o símbolo "+". Se uma observação não censurada possue o mesmo valor que uma observação censurada, coloca-se em primeiro lugar a observação não censurada.

b) COLUNA 2: esta coluna, registrada por r, consiste do posto correspondente de cada observação na COLUNA 1.

c) COLUNA 3: nesta coluna, registrada por i, coloca-se apenas as observações não censuradas, ou seja, fazemos $i = r$.

d) COLUNA 4: esta coluna deverá conter o cálculo de $(n-i)/(n-i+1)$ para toda observação não censurada $t_{(i)}$. Ela nos dá a proporção de indivíduos sobreviventes de todos os $t_{(i)}$.

e) COLUNA 5: esta coluna nos dá o cálculo de $\hat{S}(t)$, que é o produto acumulado dos resultados da coluna 4.

Para melhor compreensão do cálculo do estimador do produto limite vejamos o seguinte exemplo.

EXEMPLO 2.2. (MILLER, 1981). Um experimento clínico para avaliar a eficácia da droga 6-mercaptopurina em 11 pacientes com leucemia mielogênica aguda (chamado grupo tratamento) e em outro grupo agora com 12 pacientes que não receberam a droga (chamado grupo controle) nos dão os seguintes tempos de remissão (em semanas):

Grupo tratamento

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+ .

Grupo controle

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45.

O cálculo de $\hat{S}(t)$ para o grupo tratamento, usando o estimador do produto limite, está descrito na Tabela 2.2, enquanto que o cálculo de $\hat{S}(t)$ para o grupo controle usando também o estimador do produto limite está ilustrado na Tabela 2.3.

TABELA 2.2. Cálculo da função sobrevivência para o grupo tratamento usando os dados do exemplo 2.2.

tempo $t_{(i)}$	r	i	(n-i)/(n-i+1)	$\hat{S}(t_{(i)})=\hat{S}(t_{(i-1)})\cdot(\frac{n-i}{n-i+1})$
0	-	-	-	$\hat{S}(0) = 1$
9	1	1	10/11	$\hat{S}(9) = \hat{S}(0) \times 0.91 = 0.91$
13	2	2	9/10	$\hat{S}(13) = \hat{S}(9) \times 0.9 = 0.82$
13+	3	-	-	
18	4	4	7/8	$\hat{S}(18) = \hat{S}(13) \times 0.88 = 0.72$
23	5	5	6/7	$\hat{S}(23) = \hat{S}(18) \times 0.86 = 0.62$
28+	6	-	-	
31	7	7	4/5	$\hat{S}(31) = \hat{S}(23) \times 0.8 = 0.49$
34	8	8	3/4	$\hat{S}(34) = \hat{S}(31) \times 0.75 = 0.37$
45+	9	-	-	
48	10	10	1/2	$\hat{S}(48) = \hat{S}(34) \times 0.5 = 0.18$
161+	11	-	-	

TABELA 2.3. Cálculo da função sobrevivência para o grupo controle usando os dados do exemplo 2.2.

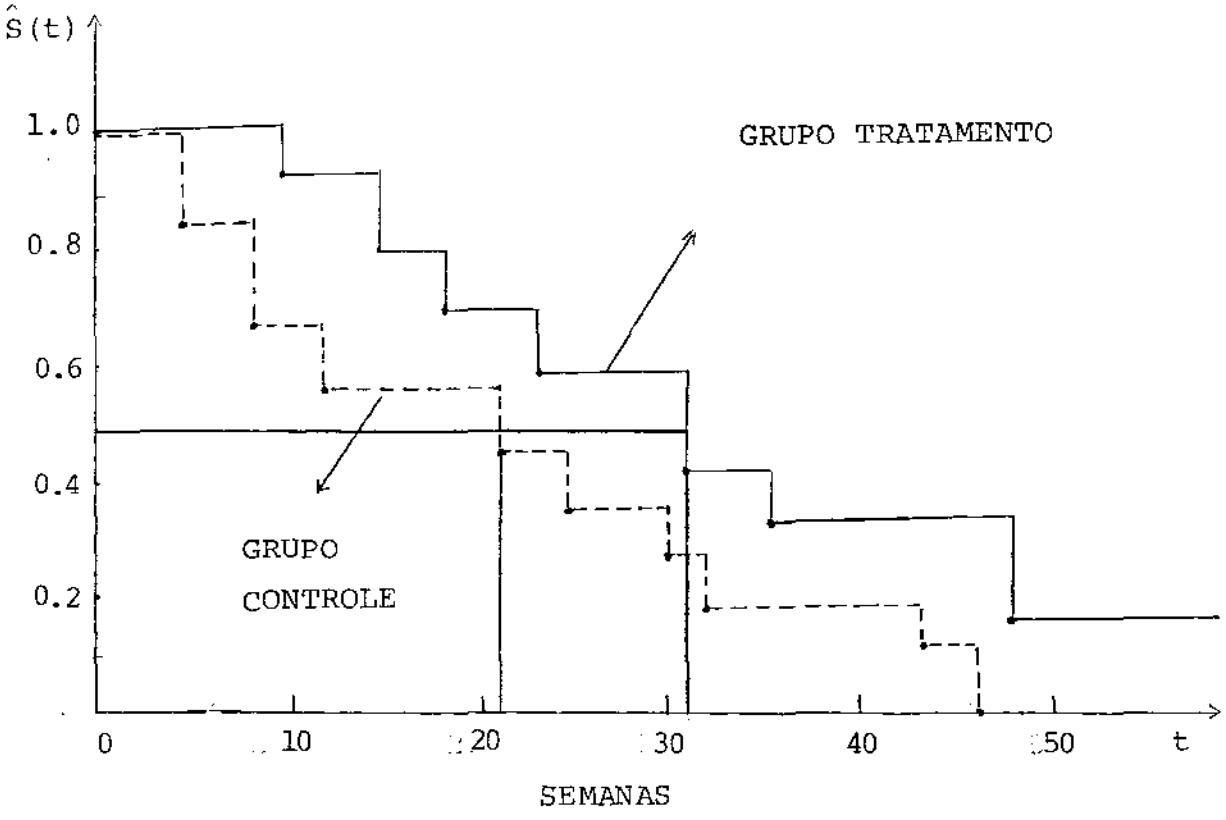
tempo $t_{(i)}$	r	i	$(n-i)/(n-i+1)$	$\hat{S}(t_{(i)}) = \hat{S}(t_{(i-1)}) \cdot \left(\frac{n-i}{n-i+1}\right)$
0	-	-		$\hat{S}(0) = 1$
5	1	1	11/12	$\hat{S}(5) = \hat{S}(0) \times 11/12 = 0.92$
5	2	2	10/11	$\hat{S}(5) = \hat{S}(5) \times 10/11 = 0.83$
8	3	3	9/10	$\hat{S}(8) = \hat{S}(5) \times 9/10 = 0.75$
8	4	4	8/9	$\hat{S}(8) = \hat{S}(8) \times 8/9 = 0.67$
12	5	5	7/8	$\hat{S}(12) = \hat{S}(8) \times 7/8 = 0.58$
16+	6	-	-	
23	7	7	5/6	$\hat{S}(23) = \hat{S}(12) \times 5/6 = 0.48$
27	8	8	4/5	$\hat{S}(27) = \hat{S}(23) \times 4/5 = 0.39$
30	9	9	3/4	$\hat{S}(30) = \hat{S}(27) \times 3/4 = 0.29$
33	10	10	2/3	$\hat{S}(33) = \hat{S}(30) \times 2/3 = 0.19$
43	11	11	1/2	$\hat{S}(43) = \hat{S}(33) \times 1/2 = 0.15$
45	12	12	0	$\hat{S}(45) = \hat{S}(43) \times 0 = 0$

Observando a figura 2.1 e recordando o que foi dito no capítulo I, podemos afirmar, de modo rústico, que o grupo tratamento tem uma curva de sobrevivência mais efetiva. No entanto , essa comparação visual nem sempre é verdadeira. Por esta razão, precisamos de outros métodos para decidirmos qual é a melhor ou melhores curvas de sobrevivência.

Esses métodos, serão estudados na seção 2.2.4.

A figura 2.1 nos mostra o gráfico do estimador do produto limite para o grupo tratamento e grupo controle

FIGURA 2.1.



Observamos na figura 2.1 que existe uma diferença de aproximadamente 10 semanas entre as medianas do grupo controle e grupo tratamento.

2.2.5. COMPARAÇÕES DE CURVAS DE SOBREVIVÊNCIA

Frequentemente o pesquisador médico vê-se diante do problema de comparar a eficácia de um determinado tratamento sobre outro. Ele precisa decidir com qual tratamento deve continuar.

No caso de comparações entre duas ou mais curvas de sobrevivência o pesquisador fica diante do mesmo tipo de problema, isto é, como decidir qual a mais representativa.

Esta seção tem como objetivo mostrar algumas técnicas de como o pesquisador pode fazer uso delas, possibilitando-lhe escolher entre várias alternativas apresentadas.

Uma das técnicas mais simples citadas por alguns autores, para verificar a existência de diferenças entre duas curvas de sobrevivência, consiste em graficar em uma mesma folha as duas curvas de sobrevivência e de modo análogo ao descrito no Capítulo I, com respeito a figura 1.a, verificar seu comportamento, fazendo sua análise a partir daí. Sabe-se no entanto, que este método nos dá somente uma idéia incompleta das diferenças existentes entre as distribuições de sobrevivência, ou seja, um simples gráfico de curvas de sobrevivência não nos revela o quanto são significativas as diferenças por ventura existentes.

Bartmann e Soares (1983) nos dão outro método simples de compararmos os tempos de sobrevivência entre dois grupos, através

da comparação do tempo total em risco. Segundo os mesmos autores, este método tem dois graves problemas, isto é, supõe que:

i) o risco de morte é independente do tempo.

ii) todos os indivíduos em cada grupo tem tempo de vida com a mesma distribuição.

Acontece que essas duas suposições nem sempre são verdadeiras.

Em razão das situações expostas, testes estatísticos são absolutamente necessários para o pesquisador que deseje decidir com mais precisão.

Existem disponíveis na literatura, para compararmos duas ou mais curvas de sobrevivência, vários testes não paramétricos que podem ser aplicados para dados com censuras assim como para dados sem censuras.

Neste trabalho, para compararmos duas curvas de sobrevivência, trataremos somente de dois destes testes, que são o teste 'log-rank' (Peto, 1972) e o teste de Cox-Mantel (Mantel, 1966, Cox, 1972).

De uma forma geral, os dois testes apresentam uma forma de calcular, relativamente simples. Em seguida descrevemos a metodologia e o uso dos testes mencionados acima.

O TESTE "LOG-RANK"

Vamos supor que aplicamos dois diferentes tratamentos a dois grupos de indivíduos e queremos aplicar o teste 'log-rank' para testar a efetividade dos tratamentos. O procedimento usual é como descrito abaixo.

O teste "log-rank" é baseado num conjunto de notas dadas (atribuidas) para as várias observações nos dois grupos estudados. De outro modo, isto significa que, se não existe diferença entre os dois tratamentos considerados, o número de falhas (mortes) num determinado período deve ser aproximadamente proporcional ao tamanho dos dois grupos naquele instante.

A comparação do que realmente acontece com o que era esperado se as distribuições do tempo de vida fosses iguais nos dois grupos, nos permite construir os testes (Bartmann e Soares, 1983).

A exemplo do que foi feito para calcularmos o estimador do produto limite, é aconselhável construirmos uma tabela para nos ajudar nos cálculos. Esta tabela será composta de 4 grupos de colunas descritas abaixo.

i) A primeira coluna registrará, em ordem crescente, os tempos de falhas e censura $t_{(i)}$.

ii) Na segunda coluna constarão os pacientes em risco, dentro de cada grupo, assim como o seu total.

iii) Na terceira coluna estarão as falhas ocorridas nos dois grupos $m_{(i)}$ e o seu total.

iv) Na última coluna temos as proporções de $r_{(i)}$ com respeito ao grupo controle $(P_{(i)})$ e ao grupo tratamento $(1 - P_{(i)})$.

Construida a tabela, o passo seguinte será calcularmos os números esperados de falhas (mortes) nos dois grupos usando as fórmulas:

$$E_1 = \sum_{i=1}^{K} m^C_{(i)} \quad P_{(i)} \tag{2.10}$$

$$E_2 = \sum_{i=1}^{K} m^T_{(i)} [1 - P_{(i)}] \tag{2.11}$$

onde K é o número de período em que ocorrem as falhas (mortes), $m_{(i)}$ e $P_{(i)}$ são como descritos acima.

Denotando-se por O_1 e O_2 respectivamente os números totais de falhas (mortes) observadas nos grupos controle e tratamento, pode-se facilmente mostrar que essa soma $(O_1 + O_2)$ é igual a soma dos valores esperados $(E_1 + E_2)$ nos dois grupos.

Para mostrarmos o fato acima citado recorremos à tabela 2.5.

Como $O_1 = 9$ e $O_2 = 21$, então $O_1 + O_2 = 30$. Do mesmo modo, observamos $E_1 = 10,8$ e $E_2 = 19,2$ logo $E_1 + E_2 = 30$, o que mostra nossa afirmação.

A discrepância entre os valores observados e esperados é medida pela estatística,

$$\chi^2 = \frac{(|O_1 - E_1| - 1/2)^2}{E_1} + \frac{(|O_2 - E_2| - 1/2)^2}{E_2} \qquad (2.12)$$

que é conhecida como estatística "log-rank".

Um teste de significância é obtido comparando-se a estatística χ^2 com uma distribuição qui-quadrado com um grau de liberdade, em razão de estarmos analisando apenas dois grupos.

Para melhor compreensão da metodologia descrita acima vejamos o seguinte exemplo (clássico na literatura).

EXEMPLO 2.3. (Anderson et all 1981). Freireich et all (1963), comparou os tempos de remissão de um grupo de pacientes com leucemia tratados com a droga 6-mercaptopurina (chamado grupo tratamento) com um grupo não tratado (chamado grupo controle).

Como resultado deste trabalho a tabela 2.4 dada a seguir, mostra os tempos de remissão (computados em semanas) dos pacientes.

TABELA 2.4

TEMPOS DE REMISSÃO (SEMANAS) DE PACIENTES COM LEUCEMIA.

TRATAMENTO:	6+, 6, 6, 6, 7, 9+, 10+, 10 , 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+
CONTROLE:	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

FONTE: Freireich et alii (1963).

OBS: O símbolo (+) denota censura.

Para calcularmos o teste "log-rank" usando os dados da tabela 2.4, vamos construir a tabela 2.5, designando por T os pacientes que estão no grupo tratamento e por C os pacientes no grupo controle.

Como podemos observar na tabela 2.4, no grupo tratamento , alguns tempos de remissão estão censurados e por esta razão, quando construimos a tabela 2.5, nas colunas correspondentes às censuras, consta uma diminuição de 0.5.

Anderson at alii (1981), justificam este fato, supondo que a censura ocorreu na metade do intervalo de tempo.

Como podemos observar é conveniente tabularmos os valores de $m_{(i)}$, $r_{(i)}$ e $P_{(i)}$. Neste caso temos $O_1 = 9$ (número de falhas registradas no grupo tratamento), $O_2 = 21$ (idem grupo controle) e $K = 17$. A primeira coluna nos dá, em ordem crescente, os 17 tempos de falhas $t_{(i)}$ das duas amostras.

De posse dos cálculos da Tabela 2.5, onde E_1 representa o número esperado de falhas no grupo controle (com 21 observações) e de modo análogo E_2 representa o número de falhas esperadas no grupo tratamento (com 21 observações), passamos ao cálculo do teste de significância.

Em primeiro lugar vamos calcular E_1 e E_2

$$E_1 = \sum_{i=1}^{17} m^{C}_{(i)} P_{(i)} = 2 \times 0.5000 + 2 \times 0.4750 + 1 \times 0.4474 + \ldots$$

$$\ldots + 2 \times 0.1429 = 10.8.$$

$$E_2 = \sum_{i=1}^{17} m^{T}_{(i)} [1 - P_{(i)}] = 2 \times 0.5000 + 2 \times 0.5250 + 1 \times 0.5526 + \ldots$$

$$\ldots + 2 \times 0.8571 = 19,2.$$

TABELA 2.5

CÁLCULO DO 'LOG-RANK' PARA OS DADOS DA TABELA 2.4.

Tempo	Em Risco (r_i)			Falhas (m_i)			Proporções		Valores Esperados	
$t_{(i)}$	T	C	TOTAL	T	C	TOTAL	$P_{(i)}$	$1-P_{(i)}$	CONTROLE	TRATAMENTO
1	21	21	42	0	2	2	0.5000	0.5000	1.0000	1.0000
2	21	19	40	0	2	2	0.4750	0.5250	0.9500	1.0500
3	21	17	38	0	1	1	0.4474	0.5526	0.4474	0.5526
4	21	16	37	0	2	2	0.4324	0.5676	0.8648	1.1352
5	21	14	35	0	2	2	0.4000	0.6000	0.8000	1.2000
6	20,5	12	32,5	3	0	3	0.3692	0.6308	0.1076	1.8924
7	17	12	29	1	0	1	0.4138	0.5862	0.4138	0.5862
8	16	12	28	0	4	4	0.4286	0.5714	1.7144	2.2856
10	14,5	8	22,5	1	0	1	0.3556	0.6444	.0.3556	0.6444
11	12,5	8	20,5	0	2	2	0.3902	0.6098	0.7704	1.2296
12	12	6	18	0	2	2	0.3333	0.6667	0.6666	1.3334
13	12	4	16	1	0	1	0.2500	0.7500	0.2500	0.7500
15	11	4	15	0	1	1	0.2667	0.7333	0.2667	0.7333
16	11	3	14	1	0	1	0.2143	0.7857	0.2143	0.7857
17	9,5	3	12,5	0	1	1	0.2400	0.7600	0.2400	0.7600
22	7	2	9	1	1	2	0.2222	0.7778	0.4444	1.5556
23	6	1	7	1	1	2	0.1429	0.8571	0.2858	1.7142
TOTAIS				9 (O_1)	21 (O_2)	30			10.7920 (E_1)	19.2082 (E_2)

Assim, substituindo os valores encontrados na equação (2.12), temos o seguinte resultado:

$$X^2 = \frac{(|9 - 19,2| - \frac{1}{2})^2}{19,2} + \frac{(|21 - 10,8| - \frac{1}{2})^2}{10,8} = 13,6 \ .$$

Deste modo, uma qui-quadrado de 13.6 é altamente significativa ($P < 0.001$), indicando a superioridade de sobrevivência do grupo de tratamento sobre o grupo controle.

Em seguida passamos a descrever o teste de Cox-Mantel.

O TESTE DE COX-MANTEL

O teste Mantel-Cox (Mantel, 1966, Cox, 1972) tem como base o mesmo princípio do teste "log-tank" descrito acima, ou seja, se não existe diferença entre os dois tratamentos, o número de falhas (mortes), num dado período $t_{(i)}$, deve ser aproximadamente proporcional ao tamanho dos grupos naquele instante.

Segundo Bartmann e Soares (1983), o teste Mantel-Cox (1972), combina a informação de tabelas de contingência 2×2 (grupo 1, grupo 2) (falhou, sobreviveu) construida para cada período i. Conforme foi dito anteriormente, o cálculo do teste Mantel - Cox (1972), é também de fácil manuseio, que passamos a descrevê-lo.

Em primeiro lugar, sejam $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ os tempos distintos de falhas (juntos) nos dois grupos. Seja $m_{(i)}$ o

número de falhas (mortes), ocorridas no período $t_{(i)}$.

Agora vamos definir as quantidades

$$U = O_2 - \sum_{i=1}^{K} m_{(i)} P_{(i)} \tag{2.13}$$

e

$$I = \sum_{i=1}^{K} \frac{m_{(i)} [r_{(i)} - m_{(i)}]}{r_{(i)} - 1} \cdot P_{(i)} \{1 - P_{(i)}\} \tag{2.14}$$

onde $r_{(i)}$ é o número total de indivíduos em risco no período $t_{(i)}$, O_2 é a soma das falhas registradas (observada) no grupo 2 (grupo controle) e as quantidades $m_{(i)}$, $P_{(i)}$ e K são respectivamente o número de falhas ocorridas no período $t_{(i)}$, a proporção de $r_{(i)}$ que pertence ao grupo controle e K é o número de períodos onde ocorrem as falhas (mortes).

Usando, novamente, os dados da Tabela 2.5, podemos facilmente encontrar os valores de U e I como definidas nas equações (2.13) e (2.14) respectivamente.

Segundo Cox (1972), um teste assintótico para duas amostras, é obtido usando a estatística.

$$T = \frac{U}{\sqrt{I}} \tag{2.14}$$

que, assintoticamente tem uma distribuição normal padrão, sob a hipótese de nulidade.

Para ilustrarmos a aplicação do teste de Mantel-Cox, vamos usar também os dados da Tabela 2.5.

Vamos supor que a hipótese nula e a hipótese alternativa sejam

$H_0 : S_T = S_C$ (a sobrevivência do grupo de pacientes que receberam o tratamento, droga 6-mercaptopurina, tem igual efeito ao grupo controle, que não recebeu a droga).

$H_1 : S_T > S_C$ (a sobrevivência do grupo de pacientes que receberam o tratamento é mais efetiva do que o grupo que não recebeu o tratamento).

Uma aplicação do teste de Cox-Mantel, com seus respectivos cálculos, é dada usando-se as equações 2.13 e 2.14 onde $K = 17$ e $O_2 = 21$.

A tabela 2.6 ilustra os cálculos das estatísticas pedidas nas equações 2.13 e 2.14 , ou seja, encontramos o valor de 10.25 para U e 6.4973 para I.

Portanto a estatística de teste é

$$T = \frac{10.25}{\sqrt{6.4973}} = 4.02.$$

TABELA 2.6

CÁLCULO DO TESTE COX-MANTEL PARA OS DADOS DA TABELA 2.4.

tempo $t_{(i)}$	$V = m_{(i)} [r_{(i)} - m_{(i)}] [P_{(i)} \; 1 - P_{(i)}]$	$\frac{V}{r_{(i)} - 1}$
1	20.00	0.4878
2	18.95	0.4859
3	17.80	0.4811
4	17.18	0.4722
5	15.84	0.4658
6	20.61	0.6543
7	6.79	0.2425
8	23.51	0.8707
10	4.93	0.2293
11	8.80	0.4513
12	7.11	0.4182
13	2.81	0.1873
15	2.74	0.1957
16	2.18	0.1677
17	2.09	0.1817
22	2.42	0.3025
23	1.22	0.2033
TOTAIS		6.4973

Como podemos observar, o valor da estatística $T = 4.02$ é bem significativo, pois na realidade, a probabilidade de significância neste caso é de $P < 0.001$.

Em face disto, evidencia-se novamente a superioridade de sobrevivência do grupo tratamento, que recebem a droga 6-mercaptopurina, ao grupo controle. Este resultado, não é uma surpresa uma vez que ao aplicarmos o teste "log-rank", já haviamos chegado ao mesmo consenso.

Embora ainda não tenhamos defrontado com problemas dessa natureza, é importante salientar aqui que, tanto no uso dos testes descritos acima, assim como na análise de curvas de sobrevivência, nos deparamos também com o problema de comparabilidade entre os grupos estudados. Explicando de uma outra forma mais simples, o tempo de sobrevivência dos indivíduos no estudo, poderá também depender de outros fatores que não seja o tipo de tratamento recebido.

A análise desse tempo de sobrevivência, será discutida com mais detalhes, no modelo de regressão de Cox, assunto do próximo capítulo.

CAPÍTULO III

MODELOS DE REGRESSÃO - O MODELO DE RISCOS PROPORCIONAIS DE COX

3.1.1. INTRODUÇÃO

Devido aos fatos descritos no final do capítulo II, Cox (1972), generalizando trabalhos anteriores (em particular o de Kaplan-Meier) introduziu uma metodologia chamada modelos de riscos proporcionais que incorpora na análise de curvas de sobrevivência descritas no capítulo 2, modelos de regressão.

Portanto, a razão de falha instantânea $\lambda(t)$, deixa de ser a mesma para todos os indivíduos no mesmo grupo, passando a incorporar a diferença existente devida a um grupo de covariáveis.

Neste trabalho sempre vamos nos referir ao caso onde o tempo de sobrevivência é continuamente distribuido o que impossibilita empates.

3.1.2. DESCRIÇÃO DO MODELO

Vamos supor que temos n indivíduos envolvidos no estudo. Com relação ao tempo de sobrevivência $t_{(i)}$, uma ou mais medidas são avaliadas, digamos as variáveis aleatórias $X_1, X_2, X_3, \ldots, X_p$. Cox (1972) chama essas variáveis, de variáveis explanatórias.

Para o i-ésimo indivíduo observado suponhamos que se tenha um vetor de covariáveis, chamado de vetor de características, ou seja,

$$\underset{\sim}{X}_i'(t) = (X_{1i}(t),\ X_{2i}(t),\ldots,X_{pi}(t)),$$

Agora o tempo de sobrevivência $t_{(i)}$, do i-ésimo paciente dependerá também dessas p variáveis independentes.

A primeira dúvida que surge (portanto deve ser bem avaliada) é se essas variáveis irão influenciar o tempo de sobrevivência de determinado indivíduo. A título de ilustração, em um estudo clínico os X_s' podem especificar, por exemplo, as características dos pacientes tais como: idade, sexo, contagem de glóbulos brancos.

Sejam $\beta_1,\beta_2,\ldots,\beta_p$ os parâmetros de regressão comuns a todos os indivíduos no estudo, cuja forma vetorial é

$$\underset{\sim}{\beta}' = (\beta_1,\beta_2,\ldots,\beta_p)\ .$$

O problema então é estabelecer uma relação entre a distribuição do tempo de sobrevivência t com o vetor de características $\underset{\sim}{X}_i'(t)$.

Cox (1972), sugere que a função risco seja usada. Seja $\lambda_i(t,\underset{\sim}{X}_i(t))$ a função risco do i-ésimo paciente no estudo.

Portanto a sugestão de Cox (1972) é feita através do seguinte modelo:

$$\lambda_i(t, \underset{\sim}{X}_i(t)) = \lambda_o(t) \cdot \exp(\beta_1 X_{1i}(t) + \beta_{2Xi}(t) + \ldots + \beta_p X_p p(t))$$

$$= \lambda_o(t) \cdot e^{\sum_{j=1}^{p} \beta_j X_{ji}(t)}$$

que pode ser escrita

$$= \lambda_o(t)\ e^{\underset{\sim}{\beta}' \underset{\sim}{X}_i(t)} \quad . \tag{3.1}$$

Explicando (3.1), vemos que o risco de um indivíduo qualquer é dado pelo produto de uma função $\lambda_o(t)$ (comum a todos os indivíduos), por um número que depende dos valores de várias variáveis explanatórias e do valor dos parâmetros de regressão. Em razão disto, fica claro o significado dado por Cox, ao chamar este modelo de riscos proporcionais.

É importante salientar que $\underset{\sim}{\beta}' \underset{\sim}{X}_i(t)$ em (3.1) pode ser substituida por qualquer função conhecida dos vetores de características X's ou dos parâmetros de regressão β's.

Como a função $\lambda_o(t)$ e o vetor de parâmetros são desconhecidos, nosso problema passa a ser a estimação dessas grandezas.

Na seção (3.3.1) descrevemos a metodologoa de estimação dos parâmetros $\beta_1, \beta_2, \ldots, \beta_p$.

Para motivar a metodologia apresentada, em seguida, mostramos um exemplo onde fazemos uso da equação 3.1.

3.2.1. UM EXEMPLO PARA INTRODUZIR A METODOLOGIA

Este exemplo (Lee, 1980) tem como objetivo mostrar ao leitor o uso da equação 3.1 dada acima.

O exemplo ilustra o problema de duas amostras (ou dois tratamentos).

Vamos supor que temos apenas uma variável, isto é, p é igual a um. Isto significa que existe somente uma variável X_1, que é definida como a seguinte variável indicadora:

$$X_{i1} = \begin{cases} 0 & \text{se o i-ésimo indivíduo é da amostra 0} \\ 1 & \text{se o i-ésimo indivíduo é da amostra 1.} \end{cases}$$

Assim, conforme a equação (3.1) as funções riscos das amostras 0 e 1 são respectivamente

$$\lambda_o(t)$$

e

$$\lambda_1(t) = \lambda_o(t)\ e^{\beta_1}\ .$$

Claramente, a função risco da amostra 1 é igual a função risco da amostra 0 multiplicada por uma constante $\exp(\beta_1)$, isto é, as duas funções riscos são proporcionais.

Vale ainda salientar que podemos escrever esta proporcionalidade em termos da função sobrevivência, ou seja,

$$S_1(t) = [S_o(t)]^K$$

onde $K = \exp(\beta_1)$.

Este fato pode ser explicado usando-se a equação 1.13, capítulo I, isto é,

$$S_1(t) = \exp\left[-\int_0^t \lambda(u,\underset{\sim}{x})du\right.$$

$$= \exp\left[-\int_0^t \lambda_o(u)\cdot e^{\beta_1}\,du\right.$$

$$= \exp\left[-e^{\beta_1}\int_0^t \lambda_o(u)du\right. = [S_o(t)]^{e^{\beta_1}}$$

O teste, para duas amostras, desenvolvido para a equação (3.1) é o teste Mantel-Cox (1972) discutido no capítulo II. Novamente mostra-se que o teste é baseado na suposição de riscos proporcionais entre as duas amostras, já mencionada quando discutimos os testes.

Ao introduzir este capítulo, afirmamos que, construido o modelo de riscos proporcionais de Cox, nosso problema era estimarmos as grandezas $\lambda_o(t)$ e o vetor de parâmetros $\hat{\underset{\sim}{\beta}}$. A seção seguinte trata da estimativa dos coeficientes de regressão do modelo.

3.3.1. ESTIMATIVA DOS COEFICIENTES DE REGRESSÃO

Conhecidas as variáveis que serão introduzidas no modelo, cujo objetivo é avaliar como essas variáveis vão influenciar o tempo de sobrevivência do i-ésimo indivíduo, faz-se necessária a estimação dos parâmetros $\beta_1, \beta_2, \ldots, \beta_p$.

Para a estimação dos β's, Cox (1972) sugere o procedimento da máxima verossimilhança, onde a função de verossimilhança é baseada sobre uma probabilidade de falha condicional. Cox a chama de verossimilhança condicional. É muito interessante reproduzirmos o argumento original de Cox sobre a verossimilhança condicional, contidos em Bartman e Soares (1983) que é o seguinte: "Suponhamos então que a função $\lambda_o(t)$ é arbitrária. Nenhuma informação sobre $\underset{\sim}{\beta}$ é dada pelos intervalos nos quais nenhuma

falha ocorre, pois, a componente $\lambda_o(t)$ pode, teoricamente, ser identicamente igual a zero em tais intervalos. Argumentamos, portanto, condicionalmente no conjunto de instantes onde as falhas ocorrem; no caso de tempos discretos vamos condicionar também nas multiplicidades observadas. Uma vez que queremos um método de análise válido para todas as $\lambda_o(t)$ possíveis, a consideração de tal distribuição condicional parece inevitável" (Cox, 1972).

Denotamos por $t_{(i)}$ o tempo de falha (morte) do indivíduo i. Suponha que esses indivíduos estejam ordenados de modo que $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$, i = 1,...,n. Seja M_{ki} o evento morte do indivíduo k no tempo $t_{(i)}$, seja ainda $t_{(i)}$ o evento de uma morte no tempo $t_{(i)}$ entre os indivíduos em risco.

Seja $R(t_{(i)})$ denotando o conjunto de indivíduos em risco no tempo $t_{(i)}$. É muito importante lembrar que $R(t_{(i)})$ consiste de todos os indivíduos cuja sobrevivência ou tempo de censura seja igual ou exceda $t_{(i)}$.

Portanto, de acordo com o modelo, a probabilidade condicional de que um indivíduo M_{ki}, no conjunto $R(t_{(i)})$, falhe (morra) no tempo $t_{(i)}$ dado que exatamente um indivíduo $t_{(i)}$ de $R(t_{(i)})$ falha (morre) no tempo $t_{(i)}$ é dado por:

$$P(M_i/t_{(i)}) = \frac{\lambda_i(t_i, \underset{\sim}{X}_i)}{\sum_{j \in R(t_{(i)})} \lambda_j(t_i, \underset{\sim}{X}_j)} = \frac{\exp(\underset{\sim}{\beta}' \underset{\sim}{X}_i(t))}{\sum_{j \in R(t_{(i)})} \exp(\underset{\sim}{\beta}' \underset{\sim}{X}_j(t))} \quad (3.2)$$

A expressão (3.2) não deve ser vista com intimidação, pois, já sabemos do capítulo I que a probabilidade de que um indivíduo morra entre os instantes $t_{(i)}$ e $t_{(i)} + \Delta t$ é dado aproximadamente por

(a) $$\lambda_o(t_{(i)}) \cdot e^{\underset{\sim}{\beta}' \underset{\sim}{X}_i(t)} \Delta t$$

enquanto que a soma dos risco de todos os indivíduos vivos é dada por

(b) $$\lambda_o(t_{(i)}) \sum_{j \in R(t_{(i)})} e^{(\underset{\sim}{\beta}' \underset{\sim}{X}_j(t))} \Delta t$$

Assim, os termos Δt e $\lambda_o(t_{(i)})$ são comuns nas duas expressões, dividindo-se (a) por (b), estes termos comuns são cancelados, resultando na expressão (3.2).

Agora fazendo-se o produto dessas probabilidades condicionais (equação 3.2) obtemos a verossimilhança condicional,

$$V_c = \prod_i \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{X}_{(i)}(t)}}{\sum_{j \in R(t_{(i)})} e^{\underset{\sim}{\beta}' \underset{\sim}{X}_j(t)}} \qquad (3.3)$$

onde o índice i nos indica que o produtório é feito sobre todas as mortes observadas.

Os estimadores para o vetor de parâmetros $\underset{\sim}{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$, podem ser obtidos, tratando-se a verossimilhança condicional, como se fosse a verossimilhança usual, ou seja, os valores estimados $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$, são obtidos maximizando o logaritmo da função de verossimilhança (3.3). Assim,

$$\mathcal{L}(\underset{\sim}{\beta}) = \sum_{i=1}^{K} \underset{\sim}{\beta}' \underset{\sim}{X}_{(i)}(t) - \sum_{i=1}^{K} \log \Big[\sum_{j \in R(t_{(i)})} \exp(\underset{\sim}{\beta}' \underset{\sim}{X}_j(t)) \Big] \quad . \qquad (3.4)$$

Claramente, os estimadores de máxima verossimilhança dos β's são os estimadores que maximizam a função de verosimilhança $\mathcal{L}(\beta)$ na equação (3.4).

Estes estimadores são obtidos resolvendo-se simultaneamente as p equações que são as derivadas de $\mathcal{L}(\beta)$ com respeito a $\beta_1, \beta_2, \ldots, \beta_p$, igualando-se a zero ou seja, dada a equação 3.4, o estimador de máxima verossimilhança $\hat{\underset{\sim}{\beta}}$, pode ser obtido como uma solução do sistema de equações (Kalbfleisch e Prentice, 1978):

$$U_g(\underset{\sim}{\beta}) = \frac{\partial \mathcal{L}(\underset{\sim}{\beta})}{\partial \beta_g} = \sum_{i=1}^{k} \{X_{gi}(t) - A_{gi}(\underset{\sim}{\beta})\} = 0 \quad (g = 1,2,\ldots,p) \tag{3.5}$$

onde $X_{gi}(t)$ é o g-ésimo elemento no vetor $\underset{\sim}{X}_{(i)}(t)$ e

$$A_{gi}(\underset{\sim}{\beta}) = \frac{\sum_{j \in R(t_{(i)})} X_{gi}(t)\, e^{\underset{\sim}{\beta}' \underset{\sim}{X}_j(t)}}{\sum_{j \in R(t_{(i)})} e^{\underset{\sim}{\beta}\underset{\sim}{X}_j(t)}} \tag{3.6}$$

De modo similar temos

$$W_{gh}(\underset{\sim}{\beta}) = -\frac{\partial \mathcal{L}(\underset{\sim}{\beta})}{\partial \beta_g \beta_h} = \sum_{i=1}^{k} C_{ghi}(\underset{\sim}{\beta}) \tag{3.7}$$

onde

$$C_{ghi}(\underset{\sim}{\beta}) = \frac{\sum_{j \in R(t_{(i)})} X_{gj}(t) X_{hj}(t)\, e^{\underset{\sim}{X}_j(t)\underset{\sim}{\beta}}}{\sum_{j \in R(t_{(i)})} e^{\underset{\sim}{X}_j(t)\underset{\sim}{\beta}}} -$$

$$- A_{gi}(\underset{\sim}{\beta})\, A_{hi}(\underset{\sim}{\beta}) \quad , \quad (g,h = 1,2,\ldots,p) \tag{3.8}$$

Existem disponíveis alguns programas de computador que encontram os estimadores de máxima verossimilhança para os

parâmetros de regressão $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$.

No nosso trabalho usamos o SAS (Statistical Analysis System) versão 79.5.

Os procedimentos do SAS que foram usados, serão detalhados no capítulo IV, que tratará exclusivamente de um exemplo.

Além da estimação do vator de parâmetros $\underset{\sim}{\beta}$, é também de interesse estimarmos a função de sobrevivência $S(t)$ para um dado do indivíduo no estudo. Assim podemos ter duas situações que são:

a) se para um determinado indivíduo tivermos os valores das variáveis explicativas todas nulas tem-se:

$$S_o(t) = e^{-\int_0^t \lambda_o(u)\,du} \tag{3.9}$$

b) na segunda hipótese se o indivíduo tiver pelo menos um valor não nulo entre as variáveis explicativas, $X_1, X_2, \ldots, X_p$, então a função sobrevivência será dada por:

$$S(t) = e^{-\int_0^t \lambda_i(u_i, \underset{\sim}{X}_i(u))\,du}$$

$$= e^{-\int_0^t \lambda_o(u) \cdot e^{\underset{\sim}{\beta}' \underset{\sim}{X}_i(u)}\,du}$$

$$= e^{-e^{\underset{\sim}{\beta}'\underset{\sim}{X}_i(u)} \int_0^t \lambda_o(u)\,du}$$

$$= S_o(t)^{\alpha} \qquad (3.10)$$

onde $\alpha = e^{\underset{\sim}{\beta}'\underset{\sim}{X}_i(u)}$.

Agora desejamos obter um estimador da função sobrevivência $S(t)$.

Uma estimativa muito usada na literatura para $S(t)$ é a sugerida por Breslow (1974), que passamos a descrevê-la.

Sejam $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$ os tempos de falhas (mortes), ordenados dos n indivíduos no estudo. Seja $m_{(i)}$ o número de falhas e seja $R(t_{(i)})$ o conjunto de indivíduos em risco no tempo $t_{(i)}$, assim

$$\hat{S}(t_{(i)}) = \prod_{k=1}^{i} \left[1 - \frac{m_{(i)}}{\sum_{j \in R(t_{(i)})} \exp(\hat{\underset{\sim}{\beta}}\,\underset{\sim}{X}_j)}\right] \qquad (3.11)$$

nos dá as estimativas da função de sobrevivência para os instantes onde as falhas ocorrem.

CAPÍTULO IV

ANÁLISE DO PROCEDIMENTO PHGLM

4.1.1. INTRODUÇÃO

O procedimento PHGLM (Proportional Hazards General Linear Model) do SAS (Statistical Analysis System), ajusta o modelo de risco proporcional de Cox - capítulo anterior - usando dados de sobrevivência, a uma única variável dependente (no nosso trabalho é a variável TTl). Um dos objetivos do modelo é detectar quais destas covariáveis influem na sobrevivência dos indivíduos em estudo. Este procedimento pode ser aplicado em dados completos (sem censuras) ou em dados com censura.

O procedimento PHGLM pode ajustar um modelo usando a eliminação "backward" assim como pode usar a técnica "stepwise", da qual fazemos uma rápida referência na seção 4.3.1.

Os dados, para serem analisados por este procedimento, devem estar ordenados, em ordem crescente, em função da variável dependente. No nosso trabalho foi usado o procedimento "SORT" do próprio SAS.

4.2.1. AS ESTATÍSTICAS USADAS NO PGHLM PARA SELEÇÃO DE VARIÁVEIS

Existem algumas técnicas que nos permitem avaliar o quanto determinada variável independente altera (influe) a sobrevivência de um grupo de indivíduos em estudo.

Neste trabalho usaremos somente as estatísticas sugeridas por Cox (1972), para testarmos os parâmetros de regressão, no modelo de risco proporcional, descrito no capítulo III.

Neste trabalho vamos usar a suposição que Cox (1972) fez, que se baseia no vetor de derivadas $U(\underset{\sim}{\beta})$. A primeira derivada do logaritmo da função de verossimilhança é dada pela equação 3.5 e menos a derivada segunda da mesma função de verossimilhança é dada pela equação 3.7, capítulo III.

O procedimento PHGLM usa a suposição descrita acima que descrevemos em seguida.

Vamos inicialmente supor que não existe nenhuma variável forçada no modelo. Deste modo, uma estatística proposta por Cox (1975) é:

$$Q_i = \frac{U_i^2(0)}{I_i(0)} \qquad (4.1)$$

onde $U_i(0)$ denota a derivada do logaritmo da função de verossimilhança calculada no 0 (zero) para o i-ésimo parâmetro (ou seja, para $\beta_i = 0$) e $I_i(0)$ é o negativo da segunda derivada do logaritmo da função de verossimilhança calculada também no

zero para o i-ésimo parâmetro. A estatística dada pela equação 4.1, tem uma distribuição assintótica qui-quadrado com um grau de liberdade.

Vale ressaltar que na equação 4.1, a verossimilhança considerada tem somente um parâmetro de regressão correspondente a i-ésima variável independente.

Agora vamos supor que, no modelo de riscos proporcionais definido pela equação 3.1, exista mais de uma variável independente, digamos p variáveis. Vamos denotar por $\underset{\sim}{B}'$ o vetor de parâmetros de regressão (vetor dos estimadores de máxima verossimilhança para as variáveis no modelo). Deste modo, as estatísticas propostas por Cox (1972,1975) são definidos por:

$$Q_p = \frac{U_p^2(\underset{\sim}{B}', 0)}{I_p(\underset{\sim}{B}', 0)} \qquad (4.2)$$

onde $U_p(\underset{\sim}{B}', 0)$ representa a derivada do logaritmo da função de verossimilhança com respeito ao p-ésimo parâmetro, calculado em $\underset{\sim}{B}'$ para os parâmetros no modelo e no zero para o p-ésimo parâmetro fora do modelo e $I_p(\underset{\sim}{B}', 0)$ é o negativo da derivada segunda, do logaritmo da função de verossimilhança nos mesmos pontos descritos acima. A estatística Q_p tem uma distribuição assintótica qui-quadrado com um grau de liberdade.

Na seção seguinte apresentamos a idéia básica do procedimento "stepwise" para a seleção de covariáveis no modelo.

4.3.1. A REGRESSÃO "STEPWISE"

A regressão "stepwise" ou regressão por passos, é um conjunto de técnicas estatísticas muito usadas para a seleção de um subconjunto de variáveis para se fazer o ajuste no modelo de regressão.

O procedimento utiliza as estatísticas definidas pelas equações 4.1 e 4.2 que em seguida passamos a descrever.

O procedimento "stepwise" pode ser apresentado em duas situações: Na primeira o procedimento acrescenta variáveis ao modelo e na segunda retira.

O procedimento citado na primeira situação, incluindo variáveis, também é abordado em duas situações. Na primeira, o procedimento incluindo variáveis, inicia de um modelo sem nenhuma variável enquanto que na segunda situação, o procedimento inicia com algumas variáveis que devem estar obrigatoriamente no modelo.

Quando o procedimento que vai incluindo variáveis começa de um modelo sem nenhuma variável, devemos incluir no modelo em primeiro lugar a variável que apresenta o maior valor da

estatística Q_i dada pela equação 4.1 entre as p-possíveis (temos p variáveis independentes). Caso ela seja significante a um nível especificado ela permanecerá no modelo. Se ela não atingir o nível especificado sairá, e o modelo não conterá nenhuma variável.

Quando o procedimento incluindo variável, começa com um modelo que já tem alguma (algumas) variável, o objetivo será saturar o modelo. Explicando, para adicionarmos outra variável no modelo, calculamos a estatística Q_p, dada pela equação 4.2 para todas as variáveis que ainda não entraram no modelo. Em seguida, aquela que tiver o maior valor e foi significativa a um nível pré-fixado, entrará no modelo. Quando calculamos a estatística Q_p e nenhuma variável atinge o nível especificado, o procedimento termina.

Como podemos observar, o que o procedimento faz na realidade é começar com um modelo reduzido saturando-o de variáveis até que atinja um número pré-fixado de variáveis ou até que a última variável a entrar no modelo não seja significante a um nível especificado.

Quando o procedimento "stepwise backward", que consiste em ir retirando variáveis no modelo, é requisitado um trabalho semelhante ao descrito acima é executado de modo inverso, ou seja, partimos de um modelo saturado até chegarmos a um modelo reduzido.

No nosso trabalho usaremos apenas o procedimento "stepwise " partindo-se de um modelo reduzido até saturá-lo.

O próximo capítulo é o exemplo onde aplicamos a metodologia descrita.

CAPÍTULO V

UMA APLICAÇÃO NUMÉRICA DO MODELO DE REGRESSÃO DE COX

5.1.1. FUNDAMENTO TEÓRICO

Neste capítulo a metodologia de Cox (1972), descrita nos capítulos anteriores, será empregada nos dados de inserção de DIUs (Dispositivo Intra Uterino), levantado pelo Ambulatório de Planejamento Familiar da UNICAMP (Universidade Estadual de Campinas).

O emprego da metodologia de Cox (1972), tem como objetivo detetar quais covariáveis são importantes na determinação do sucesso de uma permanência do dispositivo inserido por períodos de tempos mais longos.

Usando o procedimento "UNIVARIATE" do SAS pudemos verificar que algumas pacientes envolvidas no estudo tiveram dificuldades com a inserção do DIU. Essas dificuldades são de natureza médica ou pessoal. Vale salientar que algumas dificuldades sentidas por algumas pacientes, estão relacionadas com a natureza dos dispositivos.

Em seguida, destacamos alguns aspectos básicos apresentados nos dados. Deve ficar explícito que uma gestação ocorrida após a inserção do DIU, caracteriza uma falha do dispotitivo. Do mesmo

modo, uma expulsão do dispositivo, seja por causa espontânea ou provocada, ou uma remoção do dispositivo, também, por causa médica ou pessoal, caracteriza uma falha (morte). Já se uma paciente que inseriu um dispositivo e não compareceu na data marcada para retorno, é considerada uma censura, nos modelos definidos no capítulo I.

A população em estudos, é um grupo de 1883 mulheres em que se fez uma inserção de um DIU.

5.2.1. ASPECTOS SOBRE A POPULAÇÃO ESTUDADA

Como afirmamos na seção anterior, fazem parte deste estudo um grupo de 1883 mulheres que inseriram um DIU.

Para cada paciente que inseriu um dispositivo foram feitas duas fichas denominadas DIU 1 e DIU 2, onde na primeira ficha estão os dados relativos aos registros de admissão no estudo, enquanto que na segunda ficha, estão os registros relativos ao acompanhamento da paciente. Os dados trabalhados neste exemplo, correspondem à última ficha de acompanhamento de cada paciente. Em seguida descrevemos, com detalhes, o conteúdo das duas fichas de informações.

A FICHA DIU 1 - As informações contidas na ficha de admissão DIU 1, levanta em cada paciente, entre outras covariáveis, as seguintes:

i) *experiência contraceptiva* - esta covariável nos mostra qual o último método anticoncepcional usado pela paciente. Os anticoncepcionais são: DIU, PÍLULA, INJEÇÕES, OUTROS, NUNCA USOU e IGNORADO.

ii) *como terminou a última gravidez* - esta covariável, nos diz se a última gravidez da paciente terminou em um dos seguintes casos: NASCIDO VIVO, NATIMORTO, ABORTO ESPONTÂNEO, ABORTO PROVOCADO, GRAVIDEZ ECTÓPICA, NUNCA ENGRAVIDOU E IGNORADO.

iii) *número de partos e número de abortos* - estas covariáveis nos dão os respectivos números, sendo que no número de aborto constam os espontâneos e provocados.

iv) *data da última menstruação* - com esta covariável, observamos que temos algumas faltas de informações.

v) *tipo de DIU* - como a paciente aceitou a inserção do DIU - esta covariável nos mostra o tipo. Os tipos são: LIPPES, TCU - 200 e TCU - 380.

vi) *quem inseriu o dispositivo* - com esta informação é possível saber quem foi o responsável pela inserção do dispositivo da paciente. Os dispositivos foram inseridos por: MEDICO , ENFERMEIRAS, RESIDENTES e ESTAGIÁRIOS.

vii) *data da inserção do DIU e data marcada para o próximo retorno.*

viii) *mês e ano de nascimento* - os dados desta covariável, a idade em anos de cada paciente, foi calculada na época em que ocorreu a inserção na paciente.

FICHA DIU 2 - As informações contidas na ficha de acompanhamento DIU 2, a exemplo da ficha DIU 1, levanta em cada paciente, entre outras covariáveis, as seguintes:

i) *data da consulta*

ii) *como está a paciente desde a última consulta (quando houve a inserção)* - esta covariável nos mostra se desde a última consulta a paciente tem tido: DOR, HEMORRAGIA, INFECÇÃO ou OUTRA QUEIXA.

iii) *data da última mestruação* - nesta covariável consta a data ou falta de informação.

iv) *se engravidou com o DIU* - a resposta desta covariável pode ser o número 1 (SIM) e o número 2 (NÃO).

v) *se expulsou o DIU* - na resposta desta covariável consta o número 1 (SIM) e o número 2 (NÃO).

vi) *se foi retirado o DIU* - nesta covariável consta também o

número 1 (SIM) e o número 2 (NÃO).

vii) *caso tenha sido retirado* - nesta covariável consta a razão, que pode ser médica (dor, hemorragia, infecção, outra queixa, perfuração de útero, inserção com gravidez)ou pessoal (deseja ter filho, não precisa de método anticoncepcional, outra pessoal e decisão do investigador.

viii) *data de gravidez, expulsão ou retirada do DIU* - nesta covariável consta a data ou falta de informação.

ix) *se houve reinserção* - consta SIM ou NÃO.

x) *se não foi reinserido* - esta covariável diz a causa, que pode ser: NÃO INTERROMPEU O USO, marcada para reinserção, trocará por outro DIU, trocará por pílulas, fará laqueadura, trocará por outro método, não usará mais anticoncepcional, está grávida e ignorado.

Além das informações citadas, cada ficha registra a clínica que atendeu cada paciente.

De posse desses dados passamos a descrever na próxima seção, o conjunto de covariáveis usadas.

5.2.2. AS VARIÁVEIS

O conjunto de variáveis que constituem este trabalho é relativamente grande. No entanto, nem todas as variáveis listadas são de interesse da equipe de pesquisadores da CEMICAMP.

As variáveis descritas a seguir constam diretamente das fichas DIU 1 e DIU 2 ou foram obtidas transformando variáveis originais, por exemplo, em indicadores, ou são construídas a partir de mais de uma das variáveis básicas. Fazemos em seguida uma descrição mais detalhada de cada uma dessas variáveis. Foi usado, para essa descrição, o procedimento "UNIVARIATE" do SAS, versão 79.5.

VARIÁVEL PARTOS - temos sobre esta variável, 1883 informações. A tabela 5.1 nos dá um quadro demonstrativo dessa variável, através de uma distribuição de frequências.

A covariável *número de partos*, não deve ser entendida como equivalente ao número de gestações, que podem terminar em um parto ou um aborto (espontâneo ou provocado). Ela será incluída no modelo de riscos proporcionais.

TABELA 5.1 : NÚMERO DE PARTOS

Nº de partos	Frequência F	Porcentagem na cela %	Porcentagem Acumulada %
0	26	1,381	1,381
1	456	24,217	25,597
2	636	33,776	59,373
3	366	19,437	78,810
4	207	10,993	89,804
5	87	4,620	94,424
6	50	2,655	97,079
7	27	1,434	98,513
8	12	0,637	99,150
9	5	0,266	99,416
10	6	0,319	99,734
11	1	0,053	99,788
12	2	0,106	99,894
13	2	0,106	100,000
Totais	1883	100,000	

Como podemos observar na tabela 5.1, existem no conjunto de dados 26 mulheres nulíparas (nenhum parto). O outro extremo

nos mostra 2 pacientes que tiveram 12 e 13 partos respectivamente.

O número médio de partos é de 2,57 partos enquanto o desvio padrão é de 1,63 partos. O valor mediano dos partos é 2.

VARIÁVEL ABORTO - esta variável, engloba os abortos provocados e espontâneos. O número de informações tratados neste procedimento é de 1878 uma vez que existem 5 faltas de informações. A tabela 5.2 nos dá a distribuição de frequências desta variável

TABELA 5.2 : NÚMERO DE ABORTOS

Nº de abortos	Frequência F	Porcentagem na cela %	Porcentagem Acumulada %
0	1.404	74,760	74,760
1	304	16,187	90,948
2	104	5,538	96,486
3	41	2,183	98,669
4	11	0,586	99,255
5	7	0,373	99,627
6	7	0,373	100,000
Totais	1878	100,000	

Como podemos observar, existe um número significativo de mulheres que nunca abortaram, enquanto 2 mulheres no grupo tiveram 5 e 6 abortos respectivamente.

O número médio de abortos no grupo foi de 0.4 aborto, enquanto o desvio padrão da variável aborto é 0.86.

Esta covariável fará parte do procedimento PHGLM analisado na seção 5.4.1.

VARIÁVEL IDANO - a variável *idano*, a idade de cada paciente em anos, foi calculada, em cada paciente, na data da inserção do DIU. Julgamos ser esta variável, uma das mais importantes no nosso trabalho. Comprovaremos esta afirmação na análise do procedimento PHGLM na seção 5.4.1.

A tabela 5.3 nos dá a distribuição de frequência da variável *idano*.

Como podemos observar na tabela 5.3 a seguir, as idades das mulheres na época da inserção varia dos 15 (apenas uma paciente) aos 46 anos (onde temos 4 pacientes).

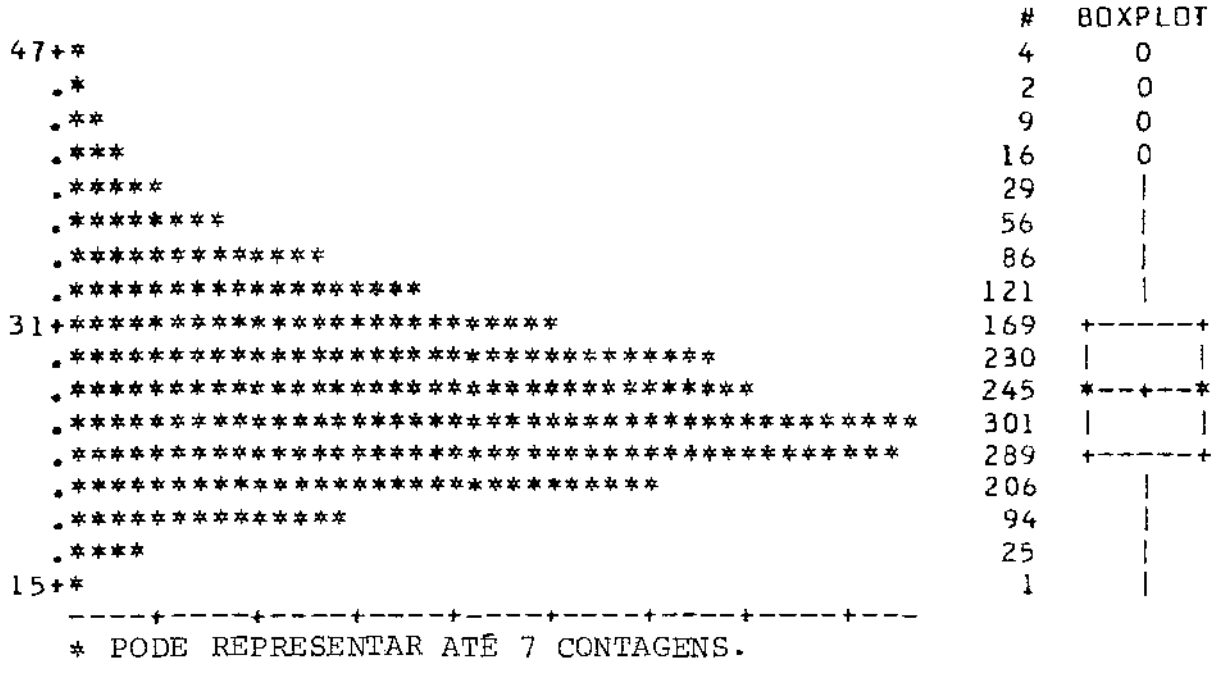
A idade média é de 26.5 anos, enquanto o desvio padrão é de 5.3 anos. A idade mediana das 1883 mulheres no estudo é de 26 anos.

Para ilustrar a variável *idano*, o gráfico 5.1, de barras, nos dá uma idéia de como essas idades se distribuem.

TABELA 5.3 - IDADE EM ANOS

Idade em anos	Frequência F	Porcentagem na cela %	Porcentagem acumulada %
15	01	0,053	0,053
16	09	0,478	0,531
17	16	0,850	1,381
18	32	1,699	3,080
19	62	3,293	6,373
20	84	4,461	10,834
21	122	6,479	17,313
22	130	6,904	24,217
23	159	8,444	32,661
24	157	8,338	40,998
25	144	7,647	48,646
26	132	7,010	55,656
27	113	6,001	61,657
28	127	6,745	68,401
29	103	5,470	73,871
30	95	5,045	78,917
31	74	3,930	82,847
32	67	3,558	86,405
33	54	2,686	89,272
34	44	2,337	91,609
35	42	2,230	93,840
36	30	1,593	95,483
37	26	1,381	96,814
38	15	0,797	97,610
39	14	0,743	98,354
40	10	0,531	98,885
41	06	0,319	99,203
42	03	0,159	99,363
43	06	0,319	99,681
44	02	0,106	99,788
46	04	0,212	100,000

GRÁFICO 5.1 - VARIÁVEL IDANO



VARIÁVEL TERMGR - As informações desta variável nos permite verificar como terminou a última gravidez de cada paciente no grupo. A tabela 5.4 nos dá a distribuição de frequência desta variável.

TABELA 5.4 : TÉRMINO DE GRAVIDEZ

Término da Gravidez	Frequência F	Porcentagem na cela %
Nascido vivo	1738	92,300
Natimorto	10	0,531
Aborto espontâneo	89	4,727
Aborto provocado	38	2,018
Nunca engravidou	8	0,424

Como podemos observar, no conjunto de 1883 pacientes, tivemos 1738 (92%) dos términos de gravidez, resultaram em nascidos vivos.

VARIÁVEL ULTANT - Esta variável nos dá o último método anticoncepcional usado pela paciente. Os resultados, registrados na tabela 5.5, nos revela o que era esperado, isto é, a predominância do uso da pílula anticoncepcional, com 1146 pacientes, o que representa, aproximadamente 61% do total de 1883 pacientes. Vê se ainda na tabela 5.5, que 65 pacientes já usavam o DIU. Outro dado a destacarmos é que 383 (20%) pacientes declararam que usavam outros métodos anticoncepcionais. De modo análogo, 216 pacientes no estudo declararam que nunca usaram nenhum tipo de anticoncepcional.

TABELA 5.5 - ÚLTIMO ANTICONCEPCIONAL USADO

Ultant	Frequência F	Porcentagem na cela %
1 - DIU	65	3,452
2 - PILULA	1146	60,860
3 - INJEÇÕES	71	3,771
4 - OUTROS	383	20,340
5 - NUNCA USOU	216	11,471
9 - FALTA DE INFORMAÇÃO	2	0,106
TOTAIS	1883	100,000

VARIÁVEL TIPO - A variável tipo, nos mostra o tipo de DIU inserido em cada paciente no estudo. A distribuição de frequências da variável tipo, com 1883 observações, estão descritas na tabela 5.6

TABELA 5.6 - TIPO DE DIU INSERIDO

Tipo	Frequência F	Porcentagem na cela %
1 - LIPPES	449	23,845
2 - TCU - 200	1110	58,948
6 - TCU - 380	324	17,207
Totais	1183	100,000

Como podemos observar, a maior quantidade de inserções de DIU foi feita com o tipo TCU - 200 (aproximadamente 59%).

Foi feito um cruzamento das variáveis TIPO e GRAVDIU (engravidou com o DIU) para verificar a eficácia do TIPO de DIU inserido.

TABELA 5.7 - CRUZAMENTO DAS VARIÁVEIS TIPO × GRAVDIU

TIPO / GRAVDIU	1 LIPPES	2 TCU - 200	6 TCU - 380	TOTAL %
SIM	16	32	1	49
%	0,85	1,70	0,05	2,60
NÃO	433	1078	323	1834
%	23,00	57,25	17,15	97,40
TOTAL	449	1110	324	1883
%	23,85	58,95	17,20	100,00

Na tabela 5.7, nas caselas correspondentes aos cruzamentos, o valor superior corresponde a frequência enquanto que o número inferior corresponde a sua porcentagem na casela.

Observando-se os percentuais obtidos, vê-se claramente a eficácia do DIU tipo 6, que é o TCU-380.

VARIÁVEL INSER - O objetivo da covariável *INSER* (quem inseriu o dispositivo na paciente) é detetar se existe diferença, na técnica de inserção dos DIUs, que altera o tempo de sobrevivência do DIU inserido em cada paciente.

Na seção 3.5.1 apresentamos o resultado do teste "log-rank" para a covariável *INSER*.

A covariável *INSER* também foi incluida no procedimento PHGLM.

A tabela 5.8 nos dá a distribuição de frequências dos profissionais que inseriram um dispositivo no grupo de 1883 pacientes.

TABELA 5.8 - QUEM INSERIU O DIU

Inserido por	Frequência	Porcentagem na cela %
MÉDICO	278	14,764
ENFERMEIRA	820	43,548
RESIDENTE	472	25,066
ESTAGIÁRIO	313	16,662
TOTAIS	1883	100,000

Como podemos observar na tabela 5.8, o maior percentual de inserções (43%) foram feitas por enfermeiras.

Fizemos novo cruzamento. Desta feita cruzamos as variáveis que inseriu com a variável EXPUL (expulsou o DIU). O objetivo deste cruzamento é verificar se a técnica de inserção do DIU por cada profissional afeta a variável EXPUL. O resultado deste cruzamento está ilustrado na tabela 5.9.

TABELA 5.9. - CRUZAMENTO DAS VARIÁVEIS INSER × EXPUL

EXPUL / INSER	NÃO 0	SIM 1	TOTAIS %
1 - MÉDICO	265	13	278
%	14,07	0,69	14,76
2 - ENFERMEIRA	781	39	820
%	41,48	2,07	43,55
3 - RESIDENTE	453	19	472
%	24,06	1,01	25,07
4 - ESTAGIÁRIO	302	11	313
%	16,04	0,58	16,62

Observando os percentuais na tabela 5.9 verifica-se que não existe diferença entre os profissionais que inseriram o dispositivo e o número de expulsões ocorridas.

VARIÁVEL RETIR - Esta variável nos dá o número de pacientes que retiram o dispositivo inserido. No nosso exemplo, tivemos 409 retiradas. As razões das retiradas são médicas e pessoais. As tabelas 5.10 e 5.11 ilustram como estão distribuidas essas razões.

TABELA 5.10 - RAZÕES MÉDICAS DAS RETIRADAS DO DIU

RAZÕES MÉDICAS	FREQUÊNCIA
1 - Dor	46
2 - Hemorragia	51
3 - Infecção	22
4 - Outra Queixa	42
5 - Perfuração do Útero	01
6 - Inserção com gravidez	01
Total	163

TABELA 5.11 - RAZÕES PESSOAIS DAS RETIRADAS DO DIU

RAZÕES PESSOAIS	FREQUÊNCIA
7 - Deseja um filho	143
8 - Não usará M.A.C.	16
9 - Outra pessoal	82
10 - Decisão do investigador	05
Total	246

Na tabela 5.11, razões pessoais das retiradas de DIU, não usará MAC significa que a paciente não usará método anticoncepcional.

VARIÁVEL INDIC 2 - A variável *INDIC 2*, que é o indicador de qualquer tipo de término, isto é, é uma variável dicotômica que assume apenas dois valores, o 1 ou 0. Designamos com o número 1 o problema de interesse. No nosso caso, o evento de interesse é a falha do dispositivo. Assim, se a paciente ficou grávida com o DIU ou se expulsou (parcialmente ou totalmente) o DIU ou se retirou o dispositivo, caracteriza o evento de interesse e é feita a contagem. Em outras palavras, quando a variável *INDIC 2* assume o valor 1 temos as observações não censuradas. Caso contrário, isto é, quando a variável *INDIC 2* é igual a zero, temos a contagem das censuras.

VARIÁVEL INDIC 1 - Esta variável tem as mesmas características da variável *INDIC 2* e será usada no procedimento SURVTEST na seção 5.3.1. A variável *INDIC 1*, quando assume o valor 2 computa as indicações não censuradas e quando assume o valor 1 computa as indicações censuradas.

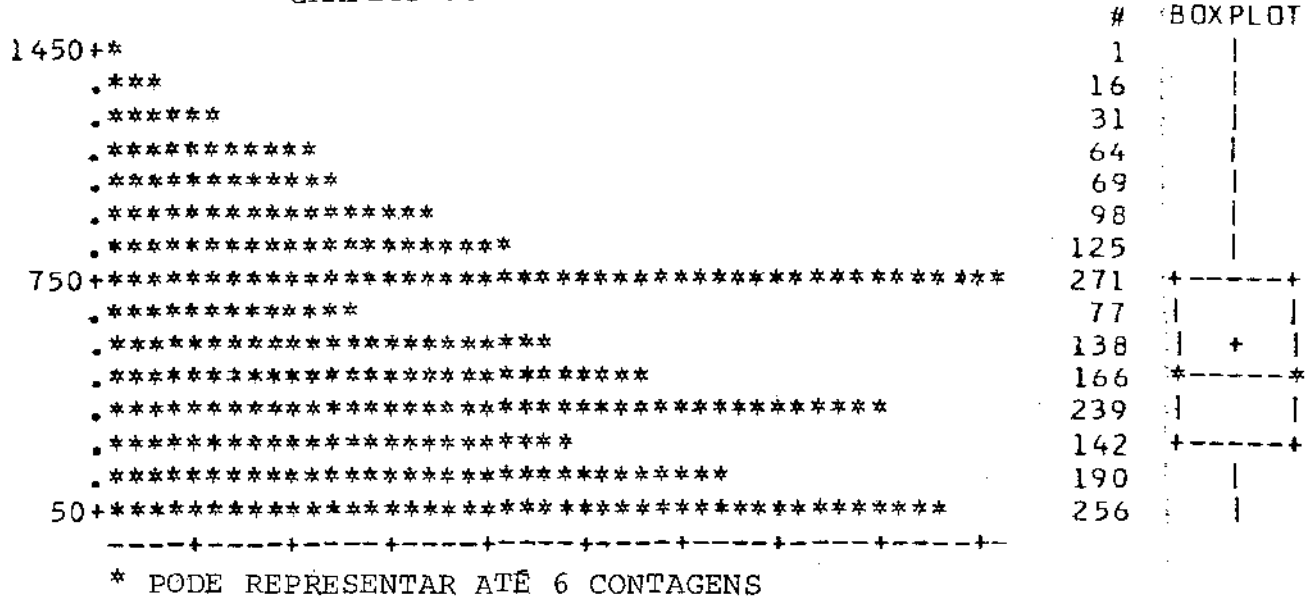
VARIÁVEL TEMPO 1 - Esta variável nos dá o tempo decorrido entre a data da consulta, que designamos por *datacon* e a data de inserção que denominamos por *datain*, ou seja, o *tempo 1 = datain datacou* . O tempo médio da variável *Tempo 1* encontrado foi de

511,6 dias (aproximadamente 17 meses) enquanto que o tempo mediano encontrado é de 459 dias (≈ 15 meses)

O gráfico 5.2 nos dá uma idéia da distribuição da variável *Tempo* 1.

A análise desta covariável, bem como as demais, foi feita usando-se o procedimento "UNIVARIATE" do SAS. Deste modo, podemos observar que os menores tempos assinalados por esta variável são 01 e 03 dias respectivamente. Observação análoga nos mostra que os maiores tempos da variável *Tempo* 1 são respectivamente 1375 e 1409 dias.

GRÁFICO 5.2 - VARIÁVEL TEMPO 1



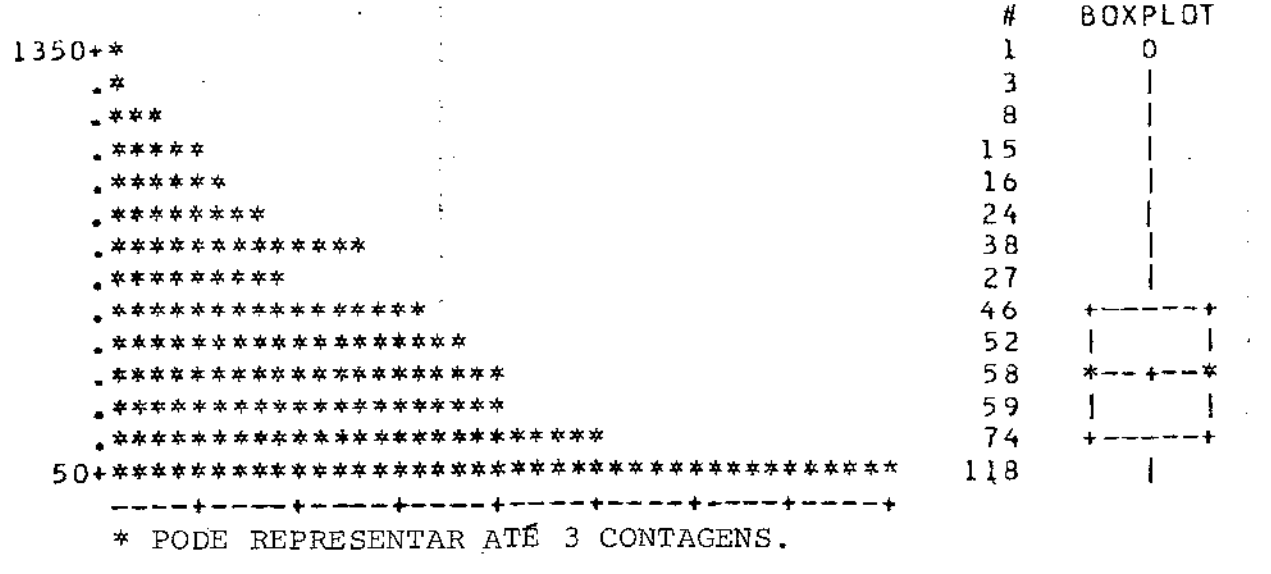
VARIÁVEL TEMPO 2 - A variável *Tempo 2* nos dá o tempo de sobrevivência do dispositivo que foi inserido na paciente. Na variável Tempo 2, o tempo é dado também, em dias.

A variável *Tempo 2* foi obtida subtraindo-se a data da retirada do dispositivo pela data da inserção. O gráfico 5.3 nos dá o comportamento desta variável.

O tempo médio da variável *Tempo 2* foi de 388,9 dias, enquanto que o tempo mediano foi de 336 dias (≃ 12 meses).

Como podemos observar, a variável *Tempo 2* tem 539 observações não censuradas (bastando somar as observações no gráfico 5.3). O total dessas observações corresponde a soma das variáveis retir, expul e gravdiu.

GRÁFICO 5.3 - VARIÁVEL TEMPO 2



De modo análogo ao que foi feito com a variável TEMPO 1, observamos que os menores tempos assumidos por esta variável são 01 e 03 dias respectivamente. Por outro lado, os maiores tempos assumidos por esta variável são 1270 e 1341 dias respectivamente.

VARIÁVEL TT 1 - No nosso trabalho a variável TT 1 desempenha um papel muito importante. Ela é a nossa variável *dependente*. Em seguida passamos a descrevê-la.

Quando a variável *INDIC 2*, definida anteriormente, assume o valor 1, então chamamos a variável *Tempo 2* de TT 1, ou seja, pelo que já foi exposto, a variável TT 1 passa agora a nos dar o tempo de sobrevivência dos dispositivos inseridos. Por outro lado, quando a variável *INDIC 2* assume o valor zero, correspondendo às censuras, a nossa variável TT 1, passa ser a variável *Tempo 1*.

Com esta variável encerramos o ciclo de descrição das variáveis. Na seção seguinte descrevemos o procedimento "SURVTEST" que calcula o teste "log-rank" definido na seção 2.2.4 do capítulo II.

5.3.1. O PROCEDIMENTO "SURVTEST"

A necessidade de usarmos o procedimento *SURVTEST* surgiu em decorrência de querermos testar o quanto é importante

determinada covariável para o nosso estudo.

Usamos o procedimento que calculou o teste "log-rank" para duas covariáveis que julgamos em princípio, serem bastante significativas que são: *INSER*, quem inseriu o dispositivo na paciente e *TIPO* de DIU usado pela paciente.

O procedimento requer que o usuário defina a sua *variável tempo*. No nosso exemplo a variável tempo, como descrita na seção anterior, é a mesma variável dependente que chamamos de variável TT1. O procedimento requer ainda que se defina a variável que computa as indicações não censuradas e censuradas. A variável assim definida é a variável INDIC 1, que quando assume o valor 2 computa as indicações não censuradas e quando assume o valor 1 computa as indicações censuradas.

Em muitos casos, segundo o que foi definido no Capítulo I, um valor para 1, computa que o indivíduo (dispositivo) está "vivo" enquanto que o número 2 computa "morte".

No primeiro exemplo, vamos testar ao nível de 5%, a hipótese nula de que não existe associação entre quem inseriu o dispositivo (INSER) e o tempo de sobrevivência do DIU(Tempo TT1). Rejeitamos H_o caso o valor da estatística χ^2 calculada seja maior que o valor crítico da distribuição $\chi^2_{n-1,\alpha}$ onde n é o número de classes da variável INSER.

A tabela 5.12 nos dá os cálculos do teste "log-rank" para a variável *INSER*, onde a variável tempo é TT1.

TABELA 5.12 - CÁLCULO DO TESTE "LOG-RANK" VARIÁVEL INSER

Quem Inseriu	N	Valor Observado	Valor Esperado	$\frac{(O-E)^2}{E}$
1 - Médico	278	80	84.66	0.26
2 - Enfermeira	820	227	239.33	0.64
3 - Residente	472	131	122.54	0.58
4 - Estagiário	312	101	92.46	0.79

A variável que indica as observações censuradas e não censuradas é INDIC 1.

O valor da estatística "log-rank", definida pela equação 2.12, Capítulo II, encontrado foi de $X^2 = 2.27$. Então, comparando-se esta estatística com uma distribuição qui-quadrado com 3(n-1, onde n=4 classes) graus de liberdade, vemos que a probabilidade de ser maior que a estatística encontrada é igual a 0,5194. Este resultado nos mostra que a técnica de inserção de cada profissional não afeta o tempo de sobrevivência do DIU.

No outro exemplo, vamos testar ao nível de 5%, a hipótese nula de que não existe associação entre o *TIPO* de DIU inserido

e o tempo que durou a inserção (variável TT 1). A variável que computa as observações censuradas e não censuradas é INDIC 1.

A tabela 5.13, nos dá os cálculos do teste "log-rank" para a variável *TIPO*.

TABELA 5.13 - CÁLCULOS DO TESTE "LOG-RANK" VARIÁVEL TIPO

TIPO DE DIU	N	VALOR OBSERVADO	VALOR ESPERADO	$\frac{(O-E)^2}{E}$
1 - LIPPES	448	146	134,97	0,99
2 - TCU - 200	1110	301	291,35	0,32
3 - TCU - 380	324	92	113,18	3,96

O valor da estatística "log-rank" encontrado é de $\chi^2 = 5.27$. Comparando-se a estatística "log-rank" calculada com uma distribuição qui-quadrada com 2(n-1, onde n=3 classes) graus de liberdade, vemos que a probabilidade de ser maior que χ^2 encontrado é igual a 0.0717. Com este resultado observamos que a variável *TIPO* não é significativa no tempo de sobrevivência do DIU ao nível de 5%.

Na seção seguinte passamos a descrever o procedimento PHGLM aplicado no grupo de pacientes que inseriram um DIU, cujas características foram descritas nas seções anteriores deste Capítulo.

5.4.1. ANÁLISE DOS DADOS DAS INSERÇÕES DE DIUs.

Os dados referentes as inserções feitas no grupo de pacientes já foram descritos em seções anteriores neste capítulo. Nesta seção, nos preocuparemos em fazer a análise numérica dos dados.

No modelo trabalhado, que passamos a descrevê-lo, usamos apenas 15 covariáveis que é o resultado de várias outras tentativas, envolvendo um grupo bem maior de covariáveis.

O procedimento de análise do modelo utilizado foi a regressão "stepwise", descrito na seção 4.3.1. do Capítulo IV, incluindo variáveis partindo do modelo sem nenhuma covariável.

Em primeiro lugar, indicamos quais variáveis entrarão no procedimento PHGLM. Essas variáveis são listadas na tabela 5.14.

Então a estatística Q_i, dada pela equação 4.1, é calculada para todas as variáveis da tabela 5.14. Aquela que tiver o maior valor entrará no modelo se o nível de significância dela for menor do que 10% (este nível é dado por P).

Em seguida temos os resultados passo a passo do procedimento "stepwise" incluindo variáveis para o grupo de pacientes que inseriram um DIU. Observe que o processo é repetido para cada modelo resultante até que nenhuma covariável seja incluida, especificada pelo nível de significância (10%) no nosso caso.

TABELA 5.14 - VARIÁVEIS QUE ENTRAM NO PHGLM

VARIÁVEL	QUI-QUADRADO	P
PARTOS	19,03	0,0000
ABORTOS	0,41	0,5206
IDADE	47,73	0,0000
INDTIP 2	0,86	0,3538
INDTIP 6	5,51	0,0189
INREST	2,37	0,1234
INDABES	0,14	0,7114
INDABPR	1,18	0,2780
INDDIU	0,55	0,4589
INDINJ	2,83	0,0923
INDOUT	10,40	0,0013
INDABOR	0,80	0,3717
INDENF	1,22	0,2702
INDEST	1,00	0,3169
INDRES	0,81	0,3668

Como podemos observar, das 1883 observações, o procedimento trabalha com 1874, pois, 9 observações foram apagadas devido a falta de informação em alguma variável.

Do total de 1874 observações, 539 são observações não cen suradas.

PASSO 1. Neste passo é incluida no modelo a variável *IDADE*, cuja estatística Qui-quadrado é 47,73 significante a menos de 0,0001. O qui-quadrado do modelo é 51.53 com 1 grau de liberdade.

A tabela 5.15 nos dá as Q estatísticas qui-quadrado ajustadas apenas para as variáveis no modelo.

TABELA 5.15 - AS Q ESTATÍSTICAS QUI-QUADRADO AJUSTADAS

VARIÁVEL	QUI-QUADRADO	P
PARTOS	1,02	0,3132
ABORTOS	0,51	0,4740
INDTIP 2	0,49	0,4830
INDTIP 6	6,01	0,0142
INREST	2,68	0,1015
INDABES	0,95	0,3301
INDABPR	1,22	0,2700
INDDIU	0,22	0,6424
INDINJ	4,34	0,0373
INDOUT	10,83	0,0010
INDABOR	2,06	0,1615
INDENF	1,25	0,2632
INDEST	1,26	0,2624
INDRES	0,83	0,3622

PASSO 2. Neste passo é adicionada a variável *INDOUT*, que possue a maior estatística qui-quadrado ao entrar no modelo 10.83(tabela

5.12) significativa a 0,001. O qui-quadrado do modelo é 63.28 com 2 graus de liberdade.

A tabela 5.16 nos dá as Q estatísticas qui-quadrado ajustadas somente para as variáveis no modelo.

TABELA 5.16 - AS Q ESTATÍSTICAS QUI-QUADRADO AJUSTADAS

VARIÁVEL	QUI-QUADRADO	P
PARTOS	1,04	0,3084
ABORTOS	1,00	0,3177
INDTIP 2	0,99	0,3209
INDTIP 6	7,38	0,0066
INDREST	3,29	0,0697
INDABES	1,23	0,2675
INDABPR	1,59	0,2067
INDDIU	0,54	0,4631
INDINJ	2,98	0,0844
INDABOR	2,67	0,1023
INDENF	1,63	0,2012
INDEST	1,27	0,2697
INDRES	1,24	0,2650

PASSO 3. Neste passo é incluida no modelo a variável *INDTIP6* com uma estatística qui-quadrado 7,38 significativa a 0,0066.

O qui-quadrado do modelo é 71.15 com 3 graus de liberdade.

A tabela 5.17 nos dá as Q estatísticas qui-quadrado ajustadas.

TABELA 5.17 - AS Q ESTATÍSTICAS QUI-QUADRADO AJUSTADAS

VARIÁVEL	QUI-QUADRADO	P
PARTOS	1,32	0,2501
ABORTO	1,30	0,2546
INDTIP 2	0,24	0,6210
INDREST	1,83	0,1765
INDABES	1,48	0,2243
INDABPR	1,64	0,2001
INDDIU	0,39	0,5335
INDINJ	3,17	0,0750
INDABOR	3,01	0,0826
INDENF	1,06	0,3042
INDEST	1,16	0,2822
INDRES	0,38	0,5389

PASSO 4. Aqui é adicionada ao modelo a variável *INDINJ* (indicador que o último anticoncepcional usado foi injeção) com uma estatística qui-quadrado igual a 3.17 significativa a 0.075. O valor qui-quadrado do modelo é 74.01 com 4 graus de liberdade.

TABELA 5.18 - AS Q ESTATÍSTICAS QUI-QUADRADO AJUSTADAS

VARIÁVEL	QUI-QUADRADO	P
PARTOS	1,62	0,2037
ABORTOS	1,18	0,2765
INDTIP 2	0,32	0,5692
INDREST	1,93	0,1647
INDABES	1,63	0,2016
INDABPR	1,71	0,1910
INDDIU	0,30	0,5835
INDABOR	3,25	0,0714
INDENF	1,06	0,3039
INDEST	1,29	0,2567
INDRES	0,37	0,5436

PASSO 5. Neste passo é adicionada a variável *INDABOR* (indicador de aborto) com uma estatística qui-quadrado igual a 3.25, significativa 0,0714. O qui-quadrado do modelo é 76.99 com 5 graus de liberdade.

TABELA 5.19 - AS Q ESTATÍSTICAS QUI-QUADRADO AJUSTADAS

VARIÁVEL	QUI-QUADRADO	P
PARTOS	1,33	0,2488
ABORTOS	0,13	0,7201
INDTIP 2	0,35	0,5515
INDREST	1,97	0,1609
INDABES	0,04	0,8415
INDABPR	0,10	0,7576
INDDIU	0,28	0,5969
INDENJ	1,13	0,2873
INDEST	1,34	0,2463
INDRES	0,36	0,5481

O procedimento pára neste passo porque nenhuma variável disponível, Tabela 5.19, atinge o nível de significância de 10%, pré-fixado, para ser incluida no modelo.

De acordo com os resultados da Tabela 5.20, observamos que, do conjunto inicial de 15 variáveis (Tabela 5.14) colocadas no procedimento PHGLM, apenas 05 delas alteram significativamente a sobrevivência dos dispositivos inseridos nas pacientes.

Essas variáveis, que Cox (1972) chama de *variáveis*

explanatórias, estão listadas pela ordem de entrada no modelo, na Tabela 5.20.

TABELA 5.20 - VARIÁVEIS QUE ALTERAM A SOBREVIVÊNCIA DOS DIUs

VARIÁVEL $X_i(t)$	BETA	ERRO PADRÃO	QUI-QUADRADO	P
IDADE	-0,00017956	0,00002512	51,11	0.0000
INDOUT	-0,42556934	0,12558018	11,48	0,0007
INDTIP 6	-0,32774692	0,11705174	7,84	0,0051
INDINJ	0,37432951	0,20261645	3,41	0,0647
INDABOR	0,30500116	0,16922852	3,25	0,0715

Interpretando os resultados obtidos, Tabela 5.20, vemos que o risco de insucesso com a inserção do DIU diminue significativamente com a idade da paciente, diminue também se a paciente usou anteriormente *outro tipo de anticoncepcional*, ou seja, se o anticoncepcional usado pela paciente não é *pílula*, nem *diu*, nem *injeção* e ainda diminue se o tipo de DIU é o TCU-380. Por outro lado, o risco aumenta se a paciente já teve algum *aborto* (provocado ou espontâneo).

CONCLUSÃO

No Capítulo I, definimos a função risco $\lambda(t)$, que, é a razão entre o número de mortes ocorridas no período t e o total de indivíduos em risco no mesmo período, ou seja, o risco estimado é dado pela equação 1.8.

$$\hat{\lambda}(t) = \frac{m(t)}{r(t)}$$

Estimados os riscos o passo seguinte é estimarmos a função sobrevivência $\hat{S}(t)$. A forma mais antiga de estimarmos a função de sobrevivência é conhecida como estimador atuarial, descrito na seção 2.2.2, que depende do cálculo prévio da função risco.

Em 1958, Kaplan-Meier desenvolveram o estimador do produto limite, seção 2.2.3, que nos permite estimarmos a função sobrevivência $\hat{S}(t)$ sem cálculo prévio da função risco.

Quando precisamos comparar duas ou mais funções de sobrevivência, para testarmos, por exemplo, a eficácia de dois ou mais tratamentos, aplicados em duas amostras, se usarmos os procedimentos citados acima para acharmos as funções sobrevivência, observamos que estes métodos tem dois graves problemas:

i) Se supõe que o risco de morte é independente do grupo.

ii) Se supõe que os grupos são homogêneos, isto é, que todos os indivíduos em cada grupo (cada amostra) tem tempos de vida com a mesma distribuição.

Acontece que, na prática, essas duas suposições raramente são verdadeiras.

Em razão disto, na seção 2.2.4, foram discutidos dois testes (o "log-rank" e Cox-Mantel) para comparar o tempo de sobrevivência em diferentes grupos, cuja validade não depende da verificação da primeira hipótese.

Sabemos que o tempo de sobrevivência pode depender também de outros fatores além do tipo de tratamento recebido. Por esta razão Cox (1972), generalizando o trabalho de Kaplan-Meier, introduziu uma metodologia (Capítulo III) que resolve este problema incorporando na análise de curvas de sobrevivência, modelos de regressão. Com isto a função risco deixa de ser a mesma para todos os indivíduos no mesmo grupo incorporando a diferença associada a uma série de covariáveis.

A grande vantagem do modelo de Cox (1972) sobre os testes desenvolvidos é que o modelo de Cox supõe que o risco instantâneo em cada indivíduo é proporcional, ou seja, o risco de um indivíduo qualquer no grupo, depende do produto de uma função $\lambda_o(t)$ (função risco comum a todos os indivíduos) por um número

$\exp(\underset{\sim}{\beta}'\underset{\sim}{X}_i(t))$ que depende de várias variáveis explicativas e dos valores dos parâmetros de regressão.

No caso do exemplo vê-se a grande vantagem do modelo de Cox que permite descrever a curva de sobrevivência como função de várias covariáveis: idade da paciente, uso de outro tipo de anticoncepcional anterior, tipo TCU 380, uso de injeções como anticoncepcional anterior e indicador de aborto. Isto é claramente superior ao que nos permitem os testes de diferenças entre tipos que também mostram que o TCU 380 comporta-se de forma diferente dos outros dois. Não só é possível detetar diferença significativa para abortos ou não como é também possível incorporar à descrição uma variável contínua como é o caso da IDADE.

Certamente, o exemplo deveria ser mais explorado de ponto de vista biomédico não cabendo, no entanto, fazê-lo aqui. O objetivo inicial de apresentar a metodologia, em suas várias formas, discutindo vantagens e desvantagens foi cumprido. O aspecto médico do problema deverá ser retomado em outro trabalho em conjunto com a equipe de pesquisadores da CEMICAMP.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ANDERSON, S., AUQUIER, A., HAUCK, W.W., OAKES,D., VANDAELE, W., WEISBERG, H.I., (1980) - *Statistical Methods for Comparative Studies*, New York, Wiley.

2. BARRETO, M.C.M. (1982) - *Análise de Curvas de Sobrevivência - O Modelo de Regressão de Cox* - Dissertação apresentada ao Instituto de Matemática, Estatística da USP para obtenção do título de Mestre em Estatística.

3. BARTMAN, F.C. e SOARES, J.F. (1983) - *Métodos Estatísticos em Medicina e Biologia* - 14º Colóquio Brasileiro de Matemática.

4. BRESLOW, N. (1974) - *Covariance Analysis of Censored Survival Data* - Biometrics 30, 89-99.

5. BRESLOW, N. e CROWLEY, J. (1974) - *A Large Sample Study of the Life Table and Product Limit Estimatives under Randon Censorship* - Annals of Statistics 2, 437-453.

6. BRESLOW, N. (1975) - *Analysis of Survival Data under a Proportional Hazards Model* - International Statistical Review, 43, 45-57.

7. CHIANG, C.L. (1968) - *Introduction to Stochastic Processes in Biostatics* - New York, Wiley.

8. COX, D.R. (1972) - *Regression Models and Life Tables* - Journal of the Royal Statistical Society, série B, 187-220 (with discussion).

9. COX, D.R. (1975) - *Partial Likelihood* - Biometrika 62, 269-276.

10. CUTLER, S. e EDERER, F. (1958) - *Maximum Utilization of the Life Table Method in Analysing Survival* - Journal of Chronical Disease 8, 699-712.

11. EFRON, B. (1977) - *The Efficiency of the Cox's Likelihood Function for Censored Data* - Journal of American Statistical Association 72, 557-565.

12. GEHAN, E.A. (1969) - *Estimating Survival Functions from the Life Table* - Journal of Chronical Disease 21, 629-644.

13. KALBFLEISCH, J.D. e PRENTICE, R.L. (1978) - *The Statistical Analysis of Faiture Time Data* - New York, Wiley.

14. KAPLAN, E.L. e MEIER, P. (1958) - *Nonparametric Estimation from Incomplete Observation* - Journal of American Statistical Association 53, 457-481.

15. KAY, R. (1977) - *Proportional Hazard Regression Models and Analysis of Censored Survival Data* - Journal of Applied Statistics 26, 227-237.

16. LAGAKOS, S.W. (1976) - *A Stochastic Model for Censored - Survival Data in the Presence of an Auxiliary Variable* - Biometrics 32, 551-559.

17. LEE, E.T. (1980) - *Statistical Methods for Survival Data Analysis* - *Lifetime Learning Publications* - Belmont, California.

18. MILLER, R.G. (1981) - *Survival Analysis* - New York, Wiley.

19. OLIVEIRA, L.A. de (1981) - *Teste do tipo Wilcoxon para Comparação de Curvas de Sobrevivência* - Dissertação apresentada ao Instituto de Ciências Matemáticas de São Carlos da USP para obtenção do Título de Mestre em Estatística.

20. PETO, R. e PETO, J. (1972) - *Asymptotically Efficient Rank Invariant Teste Procedures* - Journal of the Royal Statistical Society A 135, 185-206.

21. PRENTICE, R.L. (1973) - *Exponential Survival with Censoring and Explanatory Variables* - Biometrika 60, 279-288.

22. PRENTICE, R.L. ; KALBFLEISCH, J.D. ; PETERSON JR., A.V.; FAREWELL, V.T. e BRESLOW, N.E. (1978) - *The Analysis of Failures Times in the Presence of Competing Risks* - Biometrics 34, 541-554.

23. REINHARAT, P.S., Ed. (1980) - *S A S Supplemental Library User's Guid* - SAS Institut, Cary, N.C.