UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA , ESTATÍSTICA
E COMPUTAÇÃO CIENTÍFICA

Dissertação de Mestrado

TESTES DE SIGNIFICÂNCIA REPETIDOS EM ANÁLISE DE SOBREVIVÊNCIA

Autora: Estela Maris Pereira Bereta

Orientadora: Profª Drª Cicília Yuko Wada

Campinas, SP
Outubro, 2002

UNICAMP
BIBLIOTECA CENTRAL
SEÇÃO CIRCULANTE

UNICAMP

TESTES DE SIGNIFICÂNCIA REPETIDOS
EM ANÁLISE DE SOBREVIVÊNCIA

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Estela Maris Pereira Bereta e aprovada pela comissão julgadora.

Campinas, 09 de outubro de 2002.

Profª Drª Cicília Yuko Wada
Orientadora

Banca Examinadora:
1. Profª Drª Cicília Yuko Wada
2. Profª Drª Hildete Prisco Pinheiro
3. Profª Drª Maria Aparecida de Paiva Franco

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

UNIDADE BC
Nº CHAMADA UNICAMP
B452t
V EX
TOMBO BC/ 52364
PROC. 1241/03
C D
PREÇO R$ 11,00
DATA
Nº CPD

CM00179846-2

BIB ID 279876

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Bereta, Estela Maris Pereira

B452t Testes de significância repetidos em análise de sobrevivência / Estela Maris Pereira Bereta. -- Campinas, [S.P. :s.n.], 2002.

Orientador : Cicília Yuko Wada

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Testes de hipóteses estatísticas. 2. Bootstrap (Estatística). I. Wada, Cicília Yuko. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Dissertação de Mestrado defendida em 09 de outubro de 2002 e aprovada pela Banca Examinadora composta pelos Profs. Drs.

Cicília Yuko Wada

Prof (a). Dr (a). CICÍLIA YUKO WADA

Hildete Prisco Pinheiro

Prof (a). Dr (a). HILDETE PRISCO PINHEIRO

Maria Ap. Paiva Franco

Prof (a). Dr (a). MARIA APARECIDA DE PAIVA FRANCO

Sumário

Introdução 1

1 Análise de sobrevivência 6

1.1 Notação e conceitos básicos 8

1.2 Censuras 9

1.3 Métodos de estimação da função de sobrevivência 13

1.3.1 Métodos não-paramétricos 14

1.3.2 Métodos paramétricos 16

1.4 Testes não-paramétricos para a comparação de duas distribuições com observações completas 19

1.4.1 Testes de hipóteses de igualdade de duas distribuições do tipo contínuo, condicional às estatísticas de ordem observadas 20

1.4.2 A estatística "log-rank" 23

1.4.3 O teste de log-rank como um teste de escore eficiente segundo uma verossimilhança Parcial num modelo log-linear para as funções de risco em duas populações, com dados não censurados. 27

1.5 Testes não-paramétricos para comparação de duas funções de sobrevivência com observações censuradas 34

1.5.1 Extensões da estatística log-rank 35

1.5.2 A estatística "Weighted Kaplan-Meier" 36

1.5.3 Modelos de regressão 39

1.5.4 Modelo de riscos proporcionais de Cox 40

1.5.5 Estimação de máxima verossimilhança parcial para os coeficientes de regressão no modelo de riscos proporcionais 41
1.5.6 Estimação da função de sobrevivência básica 43
1.5.7 O modelo de regressão de Shen e Fleming 45
1.5.8 Estimação de $S_k(t)$ 46
1.5.9 Teste da hipótese de igualdade das curvas de sobrevivência para dois grupos 46
1.5.10 Propriedades teóricas da estatística de teste WMS 47

2 Métodos estatísticos usados em ensaios clínicos de resposta imediata 48

2.1 Ensaios Clínicos 48
2.1.1 Monitoramento de um ensaio clínico 50
2.1.2 Término antecipado de um ensaio clínico 51
2.1.3 A questão da demora na obtenção da resposta aos tratamentos 53
2.2 Método seqüencial clássico 53
2.3 Testes de significância repetidos 56
2.3.1 Testes de significância repetidos sobre a média de uma variável aleatória . 58
2.3.2 Cálculo iterativo de P[**A**], segundo Armitage, McPherson e Rowe (1969) . 60
2.3.3 A função gasto de alfa 61
2.4 Testes de significância repetidos em grupos 64
2.4.1 O movimento Browniano e as distribuições conjuntas de somas parciais de variáveis aleatórias i.i.d. 66
2.5 Construção de regiões críticas em testes de significância repetidos 68
2.5.1 Correção de Bonferroni para múltiplos testes em dados correlacionados . . 68
2.5.2 Regiões críticas segundo Pocock 70
2.5.3 Regiões críticas de O'Brien e Fleming 70
2.5.4 Construção de regiões de não rejeição de H_0 através do conceito da função gasto de alfa 72
2.6 Tabelas de Pocock e programa de Reboussin et al. para o planejamento do tamanho de amostra para um teste seqüencial em grupos sobre a diferença de duas médias 75

3 Métodos estatísticos usados em ensaios clínicos de resposta não imediata **76**

3.1 Notação 77

3.2 Resultados assintóticos sobre as distribuições conjuntas de estatísticas não-paramétricas calculadas em N análises 78

3.2.1 Comportamento ao longo do tempo do processo estocástico definido pelas estatísticas de log-rank 78

3.2.2 A distribuição conjunta de N estatísticas WKM calculadas em N análises 78

3.3 Exemplos 81

3.3.1 Exemplo 1, com entrada simultânea de indivíduos no Ensaio 81

3.3.2 Exemplo 2, com entrada de pacientes não simultânea 85

4 Estudo por simulação **90**

4.1 Estudo das distribuições de WKM e LR 90

4.2 Estudo do tamanho e poder de um teste Bootstrap de H_0 91

4.3 Resultados 94

5 Conclusões **100**

Ao Pedro

Agradecimentos

À professora Cicília Yuko Wada, pela orientação, pela compreensão e incentivo durante o desenvolvimento deste trabalho;

À professora Maria Aparecida de Paiva Franco, pelos ensinamentos, pelo apoio permanente e incansável incentivo para a conclusão deste trabalho;

Ao Pedro, pelo seu apoio, companheirismo, amor, compreensão e paciência neste período;

Aos Professores do Departamento de Estatística IMECC/UNICAMP que de forma direta ou indireta colaboraram na minha formação;

Aos meus pais, pelo apoio e incentivo;

A todos aqueles que de alguma forma contribuíram para a realização deste trabalho;

À CAPES pelo apoio financeiro.

Resumo

Este trabalho tem como objetivo principal elaborar uma exposição detalhada da teoria de testes de significância repetidos, sob o ponto de vista do nível real de teste, com particular ênfase nos testes não-paramétricos usados na área de análise de sobrevivência. Além disso, pretende-se apresentar uma estatística de teste, baseada na integração de uma diferença ponderada dos estimadores de Kaplan-Meier, para comparar funções de sobrevivência em dois grupos com a finalidade de disseminar o seu conhecimento e mostrar sua aplicação em problemas práticos. Inicialmente é feita uma revisão dos conceitos básicos da análise de sobrevivência e dos testes não-paramétricos usados neste contexto. Em seguida, são apresentados os fundamentos da teoria de testes de significância repetidos, desenvolvida a partir dos anos 60 para testes de hipóteses repetidos sobre a média de uma variável Y com distribuição normal (Armitage et al (1969), Pocock (1977), O'Brien & Fleming (1979), Lan & DeMets (1983)). Após esta revisão, são apresentadas as modificações propostas para adaptar essas técnicas para o caso de variáveis de resposta não imediatas e/ou censuradas, com distribuição assintótica conjunta normal multivariada. No final, é apresentada uma proposta para realização de um teste de significância repetido, utilizando reamostragem da estatística Weighted Kaplan-Meier (WKM) (Pepe e Fleming, 1989) e Log-rank em cada uma das análises. Um estudo por simulação desse procedimento por reamostragem foi feito.

Palavras-chave: Análise de sobrevivência; Bootstrap; Ensaios clínicos; Log-rank; Teste Kaplan-Meier ponderado; Testes de significância repetidos.

Abstract

The main objective of this work is to elaborate a detailed exposition of the theory of repeated significance tests, with particular emphasis on non parametric tests used in Survival Analysis. A second objective is to study and use in applications the somehow new Weighted Kaplan Meier statistic that was created to compare survival in two populations. This statistics is an integrated weighted diffference of Kaplan Meier estimates of survival curves in two groups. In the first chapter, a revision of basic definitions and tests used in Survival Analysis is made. After, the principles of the theory of repeated significance tests, developed around 1969 for comparison of means of a response variables Y with Normal distribution in two populations are presented (Armitage et al (1969), Pocock (1977), O'Brien & Fleming (1979), Lan & DeMets (1983)). Then, at Chapter 3, the results for non immediate or censored response variables with an asymptotic normal distributon are presented. Finally, a resampling procedure for making a repeated significance test using the Weighted Kaplan Meier (Pepe & Fleming, 1989). and log rank statistics at each interim analysis is proposed. A study by simulation of this procedure is presented.

Key-words: Bootstrap; Clinical Trials; Log Rank; Repeated significance tests; Survival Analysis; Weighted Kaplan Meier statistics.

Introdução

A teoria dos métodos estatísticos seqüenciais surgiu em aplicações no campo industrial nas áreas de controle de processo e amostragem. Tais métodos podem ser aplicados também em ensaios clínicos com seres humanos, com a finalidade de avaliar a eficiência de novos tratamentos, já que a obtenção das observações nesses estudos é feita ao longo do tempo, como resultado do planejamento do recrutamento de pacientes, ou pelo fato de que a resposta a ser observada como por exemplo, a remissão de sintomas só ocorrer após um certo período de tempo.

A análise seqüencial clássica exige a inspeção continuada dos dados. Isto embora seja desejável, é impraticável em ensaios clínicos, especialmente em grandes ensaios multicêntricos, em que as comissões do programa de consultoria do estudo podem reunir-se somente em intervalos periódicos. Além disto, por motivos práticos e éticos, tais ensaios devem ser cuidadosamente monitorados no decorrer de sua execução, de modo que uma toxicidade não esperada ou inaceitável dos tratamentos em estudo possa ser detectada o mais cedo possível para minimizar a exposição dos pacientes e não devem continuar por mais tempo que o necessário para provar a superioridade de um dos tratamentos. Por esta razão são exigidas, atualmente, nos protocolos de muitos ensaios clínicos de longa duração análises intermediárias, as quais habilitam o comitê de monitoramento dos dados a tomar a decisão de suspendê-lo assim que se evidencie a superioridade de algum dos tratamentos, além de permitirem verificar a ocorrência de efeitos colaterais adversos, a dificuldade de recrutamento de pacientes, e as violações de protocolos. Também pode ser eticamente desejável interromper o estudo tão logo os dados demonstrem uma clara equivalência dos tratamentos, uma vez que isto pode possibilitar que um novo ensaio se inicie com uma outra terapia promissora. A avaliação destes assuntos vai além de testes estatísticos rotineiros e requer o julgamento de especialistas de várias áreas tais como médicos, estatísti-

cos, e epidemiologistas, já que a decisão de terminar antecipadamente um ensaio clínico devido à evidência de toxicidade inaceitável ou de benefício da terapia, é complexa (DeMets & Lan, 1995).

Na prática, muitos ensaios clínicos não são planejados formalmente nem de maneira seqüencial, nem como procedimentos adequados a amostras de tamanho fixo; ao contrário, as observações são acumuladas e, em tempos determinados, são feitas repetidas análises estatísticas que seriam adequadas se realizadas com um tamanho de amostra pré-fixado. Além disso, não é considerado também o fato de que cada observação é utilizada em todas as análises a partir do momento do registro de sua ocorrência. Normalmente, vários testes são realizados nestes tempos com os dados acumulados, até se obter um resultado "significante" ou até que o analista decida concluir pela igualdade dos tratamentos. Esta forma de testes de significância repetidos não consiste em um procedimento válido, e uma diferença significativa ao nível de 5% no teste final, não corresponde ao nível de significância real associado ao ensaio como um todo, o verdadeiro nível de significância será maior e, assim, a evidência contra a hipótese nula de equivalência de tratamento é menor do que a do teste total. Portanto, repetidos testes da hipótese H_0 de igualdade de efeitos médios de dois tratamentos usando as estatísticas usuais e regiões críticas univariadas para cada inspeção dos dados aumentam a chance de concluir falsamente a existência de resultados benéficos de terapia, problema que foi relatado nos trabalhos de Wald (1947), Armitage et al. (1969), Armitage (1975), Robbins (1970), Meier (1975) e Canner (1977).

Armitage et al. (1969) mostra como calcular o nível de significância real α quando se faz seqüencialmente, até N, testes de uma hipótese simples sobre uma média populacional, iniciando com um teste com apenas uma observação, utilizando no j-ésimo teste ($j = 2, .., N$), todas as observações já usadas no $(j-1)$-ésimo teste e uma nova.

Muitas pesquisas estatísticas foram realizadas nos últimos anos para desenvolver, adaptar, ou modificar métodos estatísticos existentes com o propósito de obter subsídios para este processo de decisão complexo. Neste trabalho estudar-se-á aqui a abordagem freqüentista que tenta minimizar a proclamação de resultados benéficos do tratamento quando isto não for verdadeiro, controlando a probabilidade de erro tipo I. Trabalhos de Pocock (1977) e O'Brien & Fleming (1979) introduziram uma abordagem referente a análise intermediária seqüencial em grupos, que pode ser vista como uma extensão do trabalho de Armitage et al. (1969) sobre testes de

significância repetidos. A modificação daquele pesquisador baseou-se na idéia de que, quando o comitê de inspeção se encontra periodicamente, um grupo adicional de observações ou eventos é coletado e portanto, o monitoramento deve ser feito em grupos de pacientes, especificando antecipadamente o número de análises e o número de indivíduos em cada uma dessas análises. Pocock (1977), O'Brien & Fleming (1979), dentre outros, determinaram fronteiras de regiões críticas multivariadas para testes repetidos de igualdade de médias de variáveis com distribuição binomial ou normal. No entanto, nem sempre é fácil prever ou planejar quantas vezes ou exatamente quando o comitê poderá se encontrar para avaliar a segurança e benefício do tratamento, dessa forma, o número de observações analisadas entre reavaliações sucessivas pode variar.

Lan & DeMets (1983) estenderam o conceito de análise seqüencial em grupo criando uma técnica flexível denominada "gasto de alfa", que controla o nível de significância global, permitindo que não se estabeleça, a priori, o número e a época exata das análises e que permite ao pesquisador determinar uma região crítica de tamanho (α) e poder ($1-\beta$) de maior interesse para a sua investigação.

A extensão dos resultados acima mencionados para a área de sobrevivência, em que a resposta ao tratamento não é imediata e na qual podem ocorrer problemas externos ao delineamento que levem à censura de algumas observações, foi realizada, primordialmente, no campo não-paramétrico, em relação ao processo estocástico definido pelos valores do teste log-rank em uma seqüência de observações no tempo. Todo este desenvolvimento está ligado à procura de estatísticas com uma distribuição conjunta finito dimensional, igual à de uma seqüência de somas parciais de variáveis independentes com distribuição normal e faz uso de propriedades do movimento browniano, permitindo a unificação da teoria para qualquer número de análises e qualquer duração no tempo do ensaio clínico. Nesta linha de pesquisa, Selke e Siegmund (1983) determinaram que a distribuição conjunta de estatísticas log-rank $(l_1, l_2, .., l_N)$, calculadas nos tempos $0 < t_1 < t_2 < .. < t_N$, onde o número acumulado de mortes atinge um total pré-fixado de j_N mortes $(j = 1, .., N)$, ocorre da forma desejada.

A estatística *Weighted Kaplan-Meier* (WKM) de Pepe & Fleming (1989), usada para comparar funções de sobrevivência em dois grupos independentes, define um teste mais poderoso do que o determinado pela estatística log-rank, para detectar a alternativa H_1 de desigualdade de distribuições com funções de risco que se interceptam. Quando se pretende calcular *Weighted*

Kaplan-Meier (WKM) em N diversos momentos de análise, no decorrer de um ensaio clínico, com entrada de pacientes ao longo do tempo, e com ocorrência de possível censura aleatória, é preciso considerar a seqüência $(WKM)_j$, $j = 1, .., N$, cuja distribuição conjunta e matriz de covariância teórica foi estabelecida por Li (1999). Embora a matriz de covariância de $(WKM)_j$ para $j = 1, .., N$ não seja da forma exigida para que se aplique a simplificação proposta por Armitage et al. (1969), para o cálculo de fronteiras de regiões críticas para o teste de H_0, é possível proceder ao cálculo de fronteiras para determinados valores de nível de significância e poder de teste caso a caso, utilizando-se os resultados de Li (1999).

Os ensaios clínicos, muitas vezes, envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência e a forma mais eficiente de incorporar o efeito dessas covariáveis é utilizar um modelo de regressão apropriado para dados censurados. Shen & Fleming (1997) propuseram um modelo de regressão que permite riscos não proporcionais entre os grupos de tratamentos, assumindo um modelo de regressão de Cox dentro de cada grupo, quanto aos efeitos de outras covariáveis de interesse. Assim, estes autores propõem a estatística de teste *Weighted Mean Survival* (WMS), que é uma extensão da estatística WKM.

Este trabalho tem como objetivo principal elaborar uma exposição detalhada da teoria de testes de significância repetidos, sob o ponto de vista do nível real do teste, com particular relevância para os métodos não-paramétricos usados na área de análise de sobrevivência. O segundo interesse é o de apresentar uma estatística, baseada na integração de uma diferença ponderada dos estimadores de Kaplan-Meier, para comparar funções de sobrevivência em dois grupos com a finalidade de disseminar o seu conhecimento e mostrar sua aplicação em problemas práticos. Como esta estatística não está computacionalmente definida nos *softwares* estatísticos disponíveis, implementou-se programas para sua obtenção.

No capítulo 1, são abordados os temas de estimação da função de sobrevivência e testes de hipóteses para comparação de duas populações quanto à função de sobrevivência, dando ênfase ao caso não-paramétrico.

No capítulo 2, que diz respeito aos métodos estatísticos usados no monitoramento de ensaios clínicos com resposta imediata, é feita inicialmente uma introdução dos diversos tipos de ensaios clínicos e, em seguida, dos métodos seqüenciais clássicos e é apresentada a metodologia de testes de significância repetidos em dados acumulados ao longo de um ensaio clínico, quando a variável

de interesse é uma resposta imediata ao tratamento aplicado (Armitage et al., 1969; Pocock, 1977; O'Brien & Fleming, 1979). Apresentam-se as diversas regiões críticas para testes de significância repetidos sobre uma diferença de médias populacionais com observações completas, que ainda é o paradigma desta área de pesquisa.

Já no capítulo 3, que trata dos métodos estatísticos usados no monitoramento de ensaios clínicos com resposta não imediata, são expostas as deduções teóricas que permitem adaptar as técnicas apresentadas no capítulo 2 para testes de hipóteses não paramétricos repetidos dentro da área de análises de sobrevivência, como as que foram obtidas por Li (1999). Ainda nesta parte do trabalho, são apresentados os resultados obtidos sobre a distribuição conjunta das N estatísticas log-rank calculadas em N análises repetidas (Selke & Sigmund, 1983; Gail et al., 1981) e, também, sobre a distribuição conjunta assintótica das N estatísticas *Weighted Kaplan-Meier* (WKM) calculadas nestas análises (Li, 1999).

No capítulo 4, é realizado um estudo de simulação das duas estatísticas *Weighted Kaplan-Meier* (WKM) e log-rank sob o ponto de vista de testes repetidos. Este estudo compõe-se de simulações da distribuição conjunta de $(WKM)_j$ para $j = 1, .., N$ no contexto de N análises, quando as observações estão sujeitas à censura à direita, para alguns tamanhos de amostras.

Finalmente, no capítulo 5, são apresentadas as conclusões deste trabalho.

Capítulo 1

Análise de sobrevivência

Neste capítulo, é feita uma revisão dos conceitos básicos e a metodologia de análise de alguns ensaios clínicos em delineamentos aleatorizados simples para comparação de dois tratamentos com tamanho de amostra fixado, dentro da área de análise de sobrevivência .

A análise estatística de dados de sobrevivência lida com observações de tempo desde um instante inicial convencionado como $t = 0$, até que se observe um evento de interesse denominado "falha". Como exemplo, deste tipo de variável resposta, pode-se citar os tempos de vida de pacientes até à morte e os tempos decorridos desde o início do tratamento de uma doença até a remissão de seus sintomas nos diversos indivíduos. Pode ser objetivo de análise estabelecer modelos apropriados para a distribuição de probabilidade dos tempos de vida com a finalidade de fazer inferências sobre os parâmetros do modelo, comparando o efeito de diversos tratamentos e identificando os fatores que modificam a distribuição de probabilidade destes tempos de vida.

O desenvolvimento da área de análise de sobrevivência deve-se ao fato de que a metodologia estatística teve de levar em conta um aspecto peculiar referente às observações de tempos de vida T, que é a impossibilidade de observar todo o tempo T, em algumas situações. Realmente este problema, conhecido como censura nas observações, ocorre muitas vezes em estudos da área, como em ensaios clínicos de duração fixa, se a "falha" ainda não se instalou em alguns dos pacientes ao final do ensaio, ou se o acompanhamento de algum deles, foi possível por um determinado período em que ele ainda não apresentava "falha" foi depois interrompido. Normalmente, supõe-se que a observação de T pode ser impedida por uma variável C, denominada tempo de censura.

Diversos esquemas de censura têm sido considerados e serão apresentados neste capítulo. A seguir são apresentados alguns exemplos de dados de sobrevivência na área médica.

Exemplo 1: Um estudo clínico controlado foi realizado para investigar o efeito da terapia com esteróide no tratamento de hepatite viral aguda (Gregory et al., 1975). Vinte e nove pacientes com esta doença foram aleatorizados para receber um placebo ou tratamento com esteróide e cada um deles foi acompanhado por 16 semanas, ou até a morte (evento de interesse) ou até a perda do acompanhamento. Este tipo de censura chama-se censura simples do Tipo I, quando todos os pacientes são observados por tempos iguais. Os tempos de sobrevivência verificados, em semanas, para os dois grupos foram:

Tabela 1.1 Tempos de sobrevivência (em semanas) para os dois grupos de pacientes

Grupo	Tempos															
Controle	1*	2*	3	3	3*	5*	5*	16*	16*	16*	16*	16*	16*	16*	16*	
Esteróide	1	1	1	1*	4*	5	7	8	10	10*	12*	16*	16*	16*		

* indica observação censurada

Exemplo 2: Lawless (1982) apresenta um exemplo sobre um ensaio clínico em que uma droga (6MP) foi comparada com placebo a respeito da habilidade para manter a remissão em pacientes com leucemia. A Tabela 1.2 fornece os tempos de remissão de vinte e um pacientes nos grupos placebo e droga (6-MP). Neste experimento, os pacientes entraram no experimento num período de um ano e cada um foi observado por períodos diferentes de tempo. Este tipo de censura é chamada censuras múltiplas do Tipo I.

Tabela 1.2 Tempos de remissão (em semanas) para os dois grupos de pacientes

Grupo	**Tempos**													
6-MP	6	6	6	6*	7	9*	10	10*	11*	13	16	17*	19*	20*
	22	23	25*	32*	32*	34*	35*							
Placebo	1	1	2	2	3	4	4	5	5	8	8	8	8	11
	11	12	12	15	17	22	23							

* indica observação censurada

1.1 Notação e conceitos básicos

Seja T uma variável aleatória do tipo contínuo não negativa, representando o tempo de sobrevivência de um indivíduo, com função distribuição $F(t)$. A distribuição de probabilidade de T pode ainda ser caracterizada pela função de risco, pela função de risco acumulada e pela função de sobrevivência. Esta última pode ser denotada por $S(t)$ e definida como a probabilidade de um indivíduo sobreviver além de um tempo t :

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t) \text{ para } t \in R, \tag{1.1}$$

A função (1.1) é uma função contínua monótona decrescente tal que $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t)$=0.

A função de densidade de probabilidade ou taxa incondicional de falha, da variável aleatória T, é denotada por $f(t)$ e é definida como:

$$f(t) = \lim_{\Delta\to 0} \frac{P[t \leq T \leq t + \Delta t]}{\Delta t} = -\lim_{\Delta\to 0} \left(\frac{[S(t+\Delta t) - S(t)]}{\Delta t} \right) = -\frac{d}{d_t} S(t), \tag{1.2}$$

A função de risco do tempo de vida T, também conhecida como taxa de falha condicional, denotada por $h(t)$, é a taxa instantânea de falha no tempo $T = t$ condicionada à sobrevivência até o tempo t, e definida por:

$$h(t) = \lim_{\Delta\to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t}, \tag{1.3}$$

onde $P[t \leq T \leq t + \Delta t | T \geq t]$ é a probabilidade de falhar no intervalo de tempo $(t, t + \Delta t)$ pequeno, dado que o indivíduo sobreviveu ao início do intervalo. Portanto, tem-se que:

$$h(t) = \lim_{\Delta\to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t} = \lim_{\Delta\to 0} \frac{P[t \leq T \leq t + \Delta t]}{\Delta t P[T \geq t]} = \frac{f(t)}{S(t)}. \tag{1.4}$$

Uma outra função muito utilizada, é a função de risco acumulada, que é indicada por $H(t)$ e definida como:

$$H(t) = \int_0^t h(u)du. \tag{1.5}$$

Das expressões (1.2) e (1.4) obtêm-se as seguintes relações entre as funções:

$$h(t) = -\frac{\partial}{\partial t} \ln S(t), \tag{1.6}$$

$$S(t) = \exp\{-\int_0^t h(u)du\} = \exp\{-H(t)\}, \tag{1.7}$$

$$f(t) = h(t)S(t) = h(t)\exp\{-\int_0^t h(u)du\}. \tag{1.8}$$

1.2 Censuras

Uma característica dos estudos de sobrevivência é que nem sempre é possível observar o tempo de sobrevivência de um indivíduo até o final do estudo ou tempo de análise, sabendo-se apenas que ele é superior a um período de tempo observado; é usual supor que cada indivíduo tem um tempo potencial de vida T, que pode ou não ser verificado. A razão de não ser possível observar T pode ser devido ao planejamento do estudo, que estabelece uma data limite para o término de observação ou pode ser a ocorrência de um evento antes de ocorrer a falha do indivíduo. Uma observação deste tipo consiste em um tempo de vida censurado à direita e as causas associadas a ocorrência de censura, são diversas o que explica, portanto a existência de vários tipos de censura.

Censuras do tipo I

Em um ensaio clínico, os tempos potenciais de vida (ou tempos de falha) dos indivíduos são $T_1, T_2, ..., T_n$. Suponha-se que o período de coleta de observações, fixado no início, seja τ. Se o indivíduo falha ao tempo t, tal que $t < \tau$, então tem-se que t é o tempo de falha; se o indivíduo não falhou ao término do experimento, então $t = \tau$ é o tempo de censura, que é denominada censura do tipo I e pode ocorrer em dois tipos de situações:

a) Censuras simples do tipo I

Este tipo de censura ocorre quando todos os n indivíduos em estudo são observados por um igual período de tempo τ pré-fixado. Todos os n indivíduos entram no experimento juntos (ao

mesmo tempo) ou em tempos diferentes, mas todos são observados por um intervalo de tempo τ, acontecendo a falha se $t < \tau$ ou censura se $t = \tau$.

Assim, os tempos podem ser representados pelo par (t_i,δ_i), para $i = 1, 2, .., n$, onde

$$t_i = \min(T_i, \tau) \text{ e } \delta_i = \begin{cases} 1, & se\ t_i = T_i \\ 0, & se\ t_i = \tau \end{cases}$$

sendo δ_i a variável indicadora de falha.

Exemplo 3: Em um experimento que terminou após trinta semanas, seis ratos foram induzidos a desenvolver um certo tipo de câncer através de uma injeção de células tumorais. Os tempos observados até o desenvolvimento do tumor de um determinado tamanho foram:.10, 15, 30*, 25, 30*, 19. Os tempos 30* são censuras simples do tipo I. A Figura 1.1 mostra os tempos de desenvolvimentos dos tumores para os seis ratos.

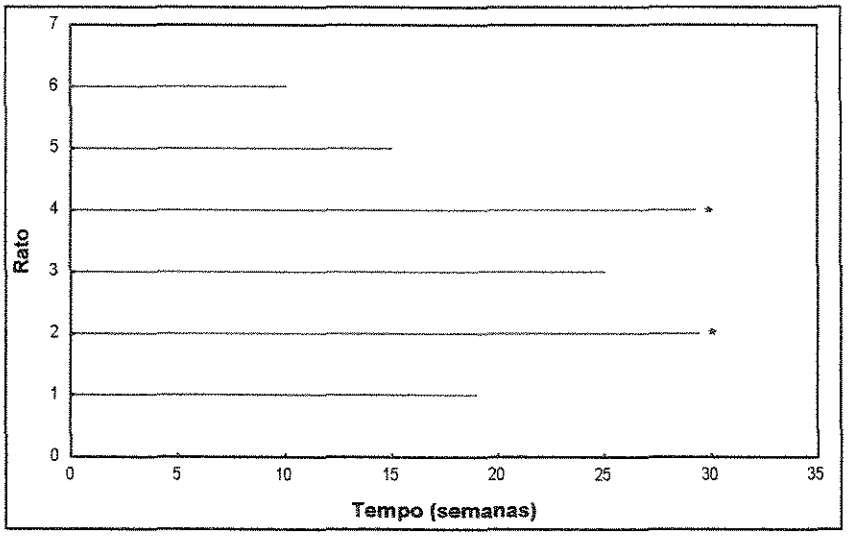

Figura 1.1: Períodos de Tempos até o aparecimento de tumor em seis ratos: (*) tempo censurado.

b) Censuras múltiplas do tipo I

Este tipo de censura ocorre quando os n indivíduos em estudo entram no experimento em tempos diferentes e são observados até uma data pré-fixada τ do calendário, na qual termina o experimento. Assim, os indivíduos ficam sujeitos a diferentes períodos de observação $L_1, L_2, ..., L_n$, isto é, cada indivíduo tem seu próprio tempo potencial de censura L_i . Se os L_i são diferentes para todos os n indivíduos, as censuras múltiplas do Tipo I ocorrem se $T_i > L_i$, onde T_i é o

tempo de falha do indivíduo i. Deste modo, os tempos observados podem ser representados pelo par (t_i, δ_i), onde:

$$t_i = min(T_i, L_i) \text{ e } \delta_i = \begin{cases} 1, & se\ T_i \leq L_i \\ 0, & se\ T_i > L_i \end{cases}$$

Note que L_i é o tempo potencial de censura e limitado por τ.

Exemplo 4: Em um estudo, seis crianças com leucemia aguda foram observadas desde a data em que entraram num período de remissão da doença e ficaram em observação até o retorno da doença ou até o fim do estudo. Os tempos de sobrevivência analisados referem-se às durações do período da remissão da doença nos pacientes medida em semanas. A Figura 1.2 ilustra, para este estudo, os períodos de remissão da doença até a recidiva ou até o final do estudo, e a Figura 1.3, a duração destes períodos de remissão da doença para o mesmo estudo.

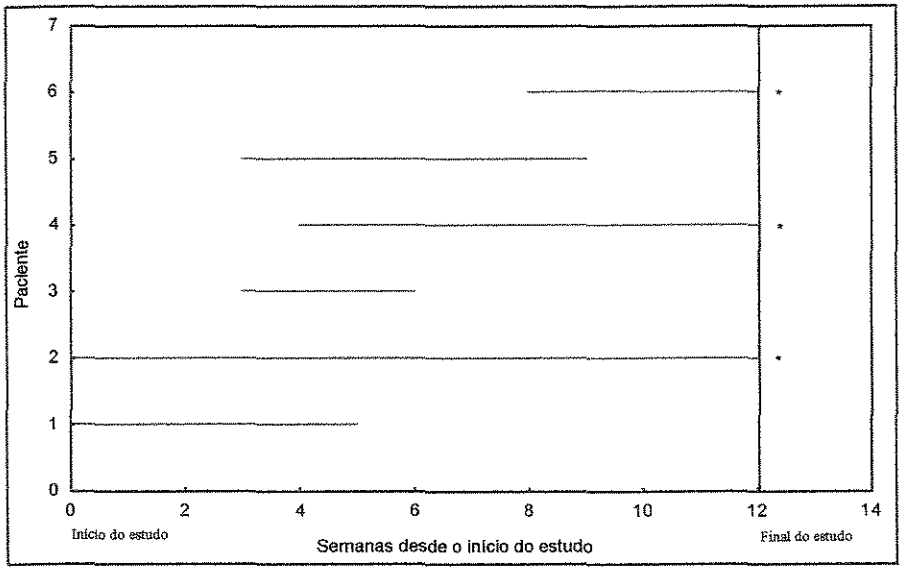

Figura 1.2: Período de remissão da doença nos pacientes até o retorno da doença ou até o final do estudo. ($\star$) indica censura múltipla do tipo I

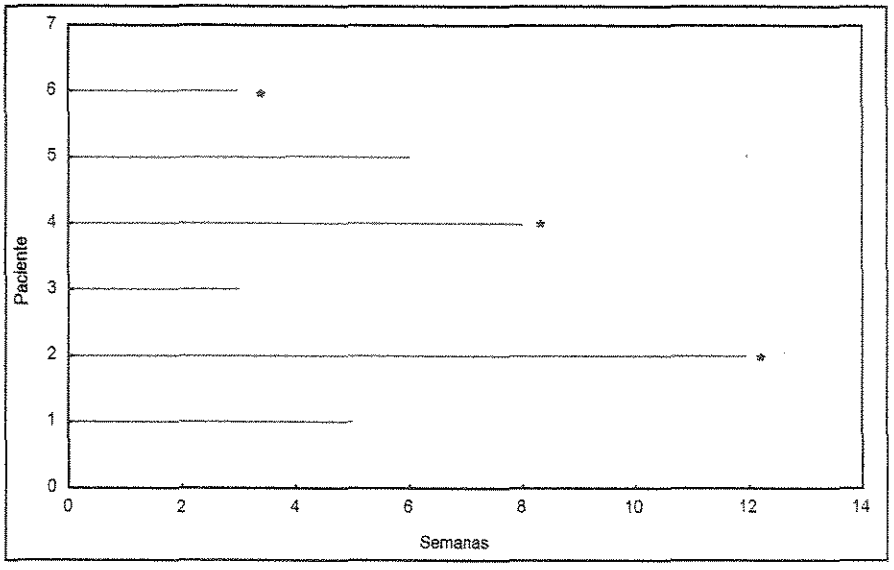

Figura 1.3: Duração do período de remissão da doença nos pacientes até o retorno da doença ou até o final do estudo. ($\star$) indica censura múltipla do tipo I.

As unidades referentes a este exemplo estão sob censura múltipla do tipo I, visto que a única informação das unidades censuradas refere-se à duração de seus períodos de remissão da doença que excedem a data do final do estudo.

Censuras do tipo II

Os experimentos industriais, geralmente, são planejados para se observar as unidades em estudo até que o evento de interesse ocorra em r delas, sendo r um número pré-fixado. Quando uma amostra de n ítens é colocada em teste até que ocorram r falhas, as $n - r$ restantes são chamadas de **censuras do Tipo II**. Neste tipo de experimento, o tempo de término é aleatório e o número r de falhas é fixado antes da realização do experimento. Se todas as unidades são colocadas em teste ao mesmo tempo, teremos as **censuras simples do Tipo II**; se essas unidades entram no experimento em tempos diferentes, as censuras são chamadas **censuras múltiplas do Tipo II**.

Em um experimento, por exemplo, em que oito componentes eletrônicos são colocados simultaneamente em teste até que cinco deles falhem, os tempos de funcionamento sem apresentar defeitos para os outros três componentes ficam censurados (censuras simples do Tipo II). A Figura 1.4 ilustra um resultado possível para este experimento.

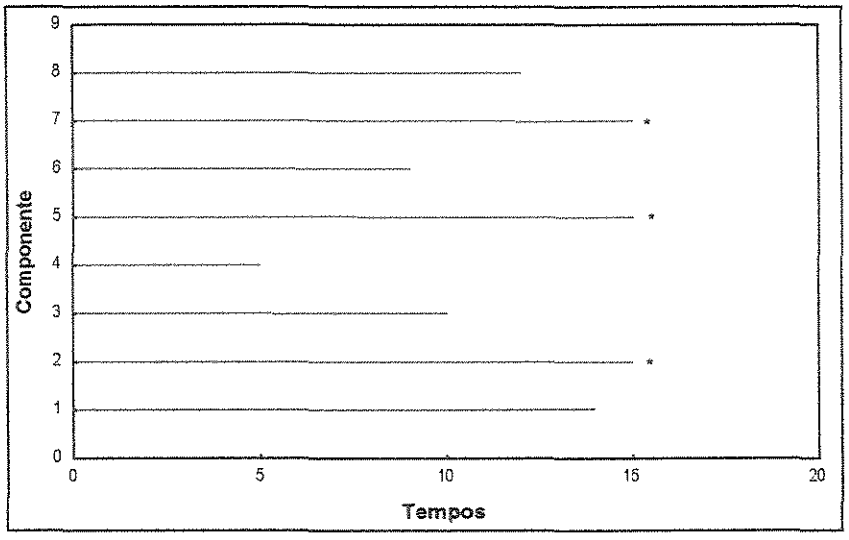

Figura 1.4: Períodos de funcionamento de oito componentes eletrônicos até a ocorrência de falha em cinco deles. (*) indica censura simples do tipo II .

Censuras aleatórias

Na censura do Tipo I, o número de falhas é aleatório, ao passo que na Censura do Tipo II este número é fixo; nos dois casos, os tempos de censuras são determinados pelo planejamento (término fixo ou número de falhas fixo). Entretanto, os tempos de censuras são freqüentemente aleatórios. Por exemplo, num ensaio clínico os pacientes entram no estudo de modo casual, de acordo com o tempo do diagnóstico. Se este estudo termina numa data pré-fixada, os tempos de censuras são aleatórios, embora se possa considerar censura do Tipo I.

Ao assumir um caráter aleatório para a censura, normalmente se considera T, o tempo potencial de vida até a falha, e C o tempo potencial até a censura como duas variáveis aleatórias independentes. Para cada indivíduo observa-se $(t_i,\ \delta_i)$, $i = 1, 2, ..., n$, onde

$$t_i = \min(T_i, C_i) \quad \text{e} \quad \delta_i = \begin{cases} 1, & \text{se } T_i \leq C_i \\ 0, & \text{se } T_i > C_i \end{cases}$$

1.3 Métodos de estimação da função de sobrevivência

A maneira de estimar a função de sobrevivência $S(t)$, está vinculada ao fato de conhecermos ou não a família de distribuições de probabilidade associada à variável aleatória T. Se não

houver informação suficiente para especificação de um modelo paramétrico para a distribuição dos tempos de sobrevivência, devemos optar por um procedimento não paramétrico para a estimação de $S(t)$.

1.3.1 Métodos não-paramétricos

Nesta seção, serão apresentadas algumas técnicas estatísticas não paramétricas usadas para analisar dados de tempo de sobrevivência na presença de censuras.

Estimador atuarial

O método atuarial estima a função de sobrevivência $S(t)$ como uma função de salto com as descontinuidades ocorrendo em instantes t_i , com $0 = t_1 < t_2 < .. < t_{k+1} = \infty$ e é usado quando não há acompanhamento contínuo do estado de vida dos pacientes, e as informações a esse respeito são coletadas segundo um acompanhamento "discreto" apenas nos instantes $t_1 < .. < t_k$. Os tempos observados de sobrevivência estão dispostos em k intervalos de classes $I_i = [t_i\ ,\ t_{i+1}),\ i = 1, 2, ..., k$. Os dados observados são as freqüências de mortes ou censuras em cada um dos intervalos, já que as observações estão agrupadas, de tal forma que temos apenas informação sobre em que intervalo cada paciente morreu ou foi censurado, mas não sabemos os tempos exatos de sobrevivência ou de censura. A função de sobrevivência é estimada por

$$\hat{S}(t_i) = \prod_{j=1}^{i-1} \hat{p}_j, \text{ de tal modo que } \Longrightarrow \begin{cases} \hat{S}(t_1) = 1, \text{ se } i = 1 \\ \hat{S}(t_i) = \hat{p}_{i-1}\hat{S}(t_{i-1}) \text{ se } i = 2, 3, ..., k \end{cases} . \tag{1.9}$$

sendo:

$\hat{p}_i = 1 - \hat{q}_i$ a proporção condicional de sobrevivência no intervalo I_i,

$\hat{q}_i = \frac{d_i}{n'_i}$ a proporção condicional de morte no intervalo I_i,

$n'_i = n_i - \frac{1}{2}(l_i + w_i),\ i = 1, ..., k$ uma aproximação para o número efetivo de indivíduos em risco no intervalo I_i,

n_i o número de indivíduos que entraram no intervalo I_i,

d_i o número de indivíduos que morreram no intervalo I_i,

l_i o número de indivíduos perdidos de observação no intervalo I_i,

w_i o número de indivíduos retirados vivos no intervalo I_i.

As censuras aqui, podem ocorrer por dois motivos: perdidos de observação ou retirados vivos devido ao término da observação no intervalo I_i ou devido a fatores éticos. Portanto, na tabela de vida, censura no intervalo I_i é definido por $(l_i + w_i)$.

Uma estimativa da variância de $\hat{S}(t)$ obtida por Greenwood (1926) é aproximada por

$$\widehat{\text{Var}}\left(\hat{S}(t_i)\right) \simeq \left[\hat{S}(t_i)\right]^2 \sum_{j=1}^{i-1} \frac{d_j}{n'_j(n_j - d_j)} \qquad para \quad i = 2, ..., k. \tag{1.10}$$

A função de risco é estimada por

$$\widehat{h(t_{mi})} = \frac{2\hat{q}_i}{b_i(1 + \hat{p}_i)} \qquad para \quad i = 2, ..., k-1. \tag{1.11}$$

onde t_{mi} é ponto médio do intervalo $[t_i\ ,\ t_{i+1})$ e b_i é o tamanho do intervalo $[t_i\ ,\ t_{i+1})$.

A variância de $\widehat{h(t_{mi})}$ é dada por

$$\widehat{\text{Var}}\left(\widehat{h(t_{mi})}\right) \simeq \frac{\left[\widehat{h(t_{mi})}\right]^2}{n'_i q_i} \left\{1 - \left[\frac{\widehat{h(t_{mi})} b_i}{2}\right]^2\right\}. \tag{1.12}$$

Estimador produto limite (Kaplan-Meier)

A estimação da função de Sobrevivência quando existe acompanhamento contínuo do estado de vida dos pacientes, deve ser feita, preferencialmente, através do estimador de Kaplan-Meier devido às suas propriedades teóricas.

Suponha que cada um de n indivíduos que pode entrar no estudo simultaneamente ou não seja acompanhado continuamente por um intervalo de tempo quanto a seu estado de vida até a ocorrência de sua falha, ou observação de censura por perda de acompanhamento, término de estudo ou outra causa. Ao final dos acompanhamentos, terão sido obtidos n tempos distintos ordenados, $t_{(1)}, t_{(2)}, ..., t_{(n)}$, até a ocorrência de falha ou censura, podendo haver mais de uma falha ao mesmo tempo. Convenciona-se que, se ocorrer empate entre uma observação de tempo de censura e uma de tempo de falha, este precede aquele. O estimador de Kaplan e Meier de $S(t)$ é definido como

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left[1 - \frac{d_i}{n_i}\right]^{\delta_i}. \tag{1.13}$$

onde,

d_i é o número de tempos de falhas ou mortes iguais a t_i;

n_i é o número de indivíduos em risco no tempo t_i, isto é, o número de indivíduos que estavam presentes no estudo, no tempo de acompanhamento t_i;

δ_i é o indicador de censura, sendo igual a 1 se a observação no tempo t_i é não censurada; caso contrário, tal indicador é igual a 0.

A variância de $\hat{S}(t)$ é estimada por

$$\widehat{\text{Var}}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{1.14}$$

O estimador de Kaplan-Meier para a função de risco acumulada $H(t)$ é dado por

$$\widehat{H(t)} = -\ln \hat{S}(t). \tag{1.15}$$

Um estimador alternativo de $H(t)$ proposto por Nelson (1969) e Altshuler (1970) é dado por

$$\widehat{H(t)} = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \qquad para \quad i = 1, 2, ..., n. \tag{1.16}$$

1.3.2 Métodos paramétricos

Uma vez escolhido o modelo paramétrico mais adequado para a distribuição dos tempos de sobrevivência, seus parâmetros podem ser estimados a partir das observações amostrais utilizando-se o método de máxima verossimilhança. Para obtenção de tais estimadores, é necessário estabelecer a função de verossimilhança considerando o tipo de censura.

Função de verossimilhança

Considere n observações independentes de (t_i, δ_i) $i = 1, .., n$, onde $t_i = min(T_i, C_i)$, δ_i é a variável indicadora de censura, supondo que T_i e C_i tenham função densidade de probabilidade $f(t, \underset{\sim}{\theta})$ e $g(t)$ respectivamente, e que $S(t, \underset{\sim}{\theta})$ seja a função de sobrevivência de T.

Função de verossimilhança considerando censuras tipo I

Neste esquema de censuras, os pacientes são observados até um tempo pré-fixado τ_0. Seja $\mathbf{d}=\{t_i : \delta_i = 1\}$ e $\mathbf{c}=\{t_i : \delta_i = 0\}$, a função de verossimilhança considerando as partições $\mathbf{d}$ e $\mathbf{c}$ é dada da seguinte forma

$$L(\theta) = \prod_{\mathbf{d}} f(t_i,\theta) * \prod_{\mathbf{c}} S(t_i,\theta). \tag{1.17}$$

O primeiro produtório é feito usando apenas os valores correspondentes a falhas e o segundo usa somente os valores de tempos de censura, o número de termos em cada produtório é aleatório.

Função de verossimilhança considerando censuras tipo II

Observa-se um grupo de n pacientes e termina-se o ensaio após a ocorrência da k-ésima morte. Sendo assim, somente as k menores observações ordenadas da amostra de tamanho n de T são observadas (k é fixo, $1 \leq k \leq n$). Para esta situação a função de verossimilhança é

$$L(\boldsymbol{\theta}) = \frac{n!}{(n-k)!} \left[\prod_{i=1}^{k} f(t_{(i)},\boldsymbol{\theta})\right] * \left[S(t_{(k)},\boldsymbol{\theta})\right]^{n-k}. \tag{1.18}$$

O número de termos, no primeiro produtório (valores completos) é fixado de antemão como k e, conseqüentemente, o número de termos no segundo é também fixado sendo $n-k$.

Função de verossimilhança no caso de censuras aleatórias quando T e C são independentes

Com censuras aleatórias, é comum a suposição de que as variáveis aleatórias contínuas T e C são independentes. As funções de sobrevivência e de densidade de C são $G_i(c)$ e $g_i(c)$ para $i = 1, 2, .., n$.

Para esta situação temos que,

$$\mathrm{P}[T_i \in (t, t+dt), \delta_i = 1; \boldsymbol{\theta}] = \mathrm{P}[T_i^0 \in (t, t+dt), C_i > t; \boldsymbol{\theta}] = G_i(t+0) f(t,\boldsymbol{\theta}) dt$$

e

$$P[T_i \in (t, t+dt), \delta_i = 0; \boldsymbol{\theta}] = g_i(t) F(t+0,\boldsymbol{\theta}) dt. \tag{1.20}$$

sendo $F(t+0,\boldsymbol{\theta}) = \lim_{u \to t+} F(u,\boldsymbol{\theta})$.

sendo $F(t+0,\theta) = \lim_{u \to t+} F(u,\theta)$.

Se F é contínua, $F(t+0,\theta) = F(t,\theta)$.

Desta forma a função de verossimilhança de (t_i, δ_i), $i = 1, 2, .., n$ pode ser escrita como:

$$\begin{aligned} L(\underset{\sim}{\theta}) &= \prod_{\mathrm{d}} f(t_i,\theta)G(t_i) * \prod_{\mathrm{c}} g(t_i)S(t_i,\theta) = \prod_{i=1}^{n} [f(t_i,\theta)G(t_i)]^{\delta_i} \prod_{i=1}^{n} [g(t_i)S(t_i,\theta)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(t_i,\theta)^{\delta_i} S(t_i,\theta)^{1-\delta_i} \prod_{i=1}^{n} G(t_i)^{\delta_i} g(t_i)^{1-\delta_i}. \end{aligned}$$

Em geral, os parâmetros de interesse são os componentes de $\boldsymbol{\theta}$ associados ao tempo de sobrevivência T. Se $G_i(t)$ e $g(t_i)$ não envolverem estes parâmetros, a função de verossimilhança pode ser escrita como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i,\theta)^{\delta_i} S(t_i,\theta)^{1-\delta_i}. \tag{1.21}$$

Estimação dos parâmetros $\boldsymbol{\theta}$ por máxima verossimilhança

Para estimar os parâmetros $\boldsymbol{\theta}$, deve-se maximizar a função de verossimilhança estabelecida de acordo com o tipo de censura ocorrido. Se $L(\boldsymbol{\theta})$ é uma função diferenciável de $\boldsymbol{\theta}$, uma condição necessária para que um valor $\widehat{\theta}$ seja um ponto de máximo da função $L(\boldsymbol{\theta})$ é que as derivadas parciais de $L(\boldsymbol{\theta})$ em relação a cada um dos parâmetros seja igual a zero quando $\boldsymbol{\theta} = \widehat{\theta}$, isto é,

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} = 0, \qquad i = 1, 2, .., k$$

No entanto, maximizar $L(\boldsymbol{\theta})$ equivale a maximizar a função log-verossimilhança definida por $l(\theta) = \ln L(\boldsymbol{\theta})$. Para que um valor $\widehat{\theta}$ de $\boldsymbol{\theta}$ seja um ponto de máximo de $L(\boldsymbol{\theta})$ é necessário que $\widehat{\theta}$ satisfaça as equações

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_i} = 0, \qquad i = 1, 2, .., k$$

Muitas vezes não é possível encontrar os estimadores analiticamente e desta forma, métodos iterativos para maximizar $L(\boldsymbol{\theta})$ são utilizados, como o de Newton-Raphson.

1.4 Testes não-paramétricos para a comparação de duas distribuições com observações completas

Nesta seção, serão revistos os conceitos de estatísticas de ordem e de postos, que são fundamentais para a teoria de testes não-paramétricos.

Uma n-upla de números reais distintos $(t_1, ..., t_n)$ é transformada de maneira única, pela ordenação crescente de seus elementos em uma n-upla $(t_{(1)}, ...t_{(n)})$ com $t_{(1)} < t_{(2)} < ... < t_{(n)}$. Define-se para i , $j = 1, ..., n$

a) P_i, posto da i-ésima componente, o número de elementos na n-upla menor que t_i tal que $P_i = j$ se $t_i = t_{(j)}$,

b) R_i, o número de elementos da n-upla $(t_1,....,t_n)$ maiores que t_i tal que $R_i = n - P_i + 1$,

c) J_i, identificação do elemento da amostra que corresponde ao posto P_i pela condição $J_i = k \Longleftrightarrow t_k = t_{(i)}$.

Por exemplo, seja uma amostra de $n = 5$ tempos de vida (2,7,4,3,8). A amostra ordenada será (2,3,4,7 e 8) e os valores de P_i e J_i estão apresentados na Tabela 1.3.

Tabela 1.3 Valores de $t_{(i)}, P_i$ e J_i para uma amostra de 5 tempos de vida

$t_{(i)}$	2	3	4	7	8
P_i	1	2	3	4	5
J_i	1	4	3	2	5

Dada, pois, uma n-upla de números reais distintos $(t_1, ..., t_n)$, ficam definidas de maneira unívoca as n-uplas

$$(t_{(1)}, ..., t_{(n)}),\ (P_1, ..., P_n) \text{ e } (J_1, ..., J_n)$$

Por outro lado,

$$(t_{(1)}, ..., t_{(n)}) \text{ e } (J_1, ..., J_n)$$

ou

$$(t_{(1)}, ..., t_{(n)}) \text{ e } (P_1, ..., P_n)$$

determinam a n-upla original $(t_1, ..., t_n)$.

1.4.1 Testes de hipóteses de igualdade de duas distribuições do tipo contínuo, condicional às estatísticas de ordem observadas

Sejam X e Y variáveis aleatórias independentes, com funções de sobrevivência do tipo contínuo indicadas por $S_1(t)$ e $S_2(t)$, respectivamente. Deseja-se testar a hipótese H_0: $S_1(t) = S_2(t)$ a partir de n_1 observações i.i.d de X e n_2 observações i.i.d de Y, a um nível de significância α.

Sob a hipótese H_0, existe uma única função densidade que será indicada por $f(x)$. Neste caso, a função densidade conjunta de uma amostra ordenada de tamanho n_1 de X e de uma amostra ordenada de tamanho n_2 de Y, é dada por

$$f(x_1)f(x_2)....f(x_{n_1})f(y_1)....f(y_{n_2}),$$

e a função densidade conjunta das estatísticas de ordem observadas $(t_{(1)} < t_{(2)} < ... < t_{n_1+n_2})$ é

$$(n_1+n_2)!f(x_1)f(x_2)....f(x_{n_1})f(y_1)....f(y_{n_2}),$$

de modo que a função densidade condicional de uma observação $\mathbf{u} = (x_1, x_2, .., x_{n_1}, y_1, .., y_{n_2})$, dado um conjunto observado de estatísticas de ordem $t_{(i)}$ $(i = 1, .., (n_1+n_2))$, é igual a

$$f(\mathbf{u}|t_{(i)}) = f(x_i, y_i|t_{(i)}) = \frac{1}{(n_1+n_2)!}$$

se as estatísticas de ordem de $\mathbf{u} = (x_1, x_2, ..., x_{n_1}, y_1, ..., y_{n_2})$ forem iguais às estatísticas dadas $(t_{(1)}, t_{(2)}, ., t_{(n_1+n_2)})$ e

$$f(\mathbf{u}/t_{(i)}) = 0$$

nos demais casos.

Uma variedade de estatísticas baseadas nas estatísticas de ordem e nos postos das observações foi proposta para criar testes condicionais da hipótese H_0: $S_1(t) = S_2(t)$, as quais são descritas a seguir:

A estatística de Wilcoxon

O teste de Wilcoxon pode ser usado para comparar funções de distribuição de duas populações independentes. Para sua construção, primeiramente, ordena-se todas as observações como se elas fossem uma única amostra e obtém-se os postos correspondentes. No caso de dados completos e sem empates, a soma dos postos correspondente à população 2 pode ser usada como estatística de teste, isto é:

$$W_2 = \sum_{i=1}^{n2} P_i$$

onde P_i é a soma dos postos ocupados pelas observações de Y na ordenação conjunta das (n_1+n_2) observações.

Sob a hipótese nula, a distribuição de probabilidade de W_2 é

$$\mathrm{P}(W_2 = w) = \frac{k(w)}{(n_1+n_2)!}. \tag{1.22}$$

onde $k(w)$ é número de permutações em que a soma de postos das n_2 observações referentes a Y é w observado e, n_1 e n_2 são os tamanhos amostrais dos dois grupos.

Os primeiros momentos centrais da distribuição de W_2, sob a hipótese nula e, condicional às estatísticas de ordem observadas, são

$$\mathrm{E}(W_2) = \frac{(n_2)(n_1+n_2+1)}{2} \qquad \mathrm{Var}(W_2) = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

Regra de Decisão:

No caso de pequenas amostras, o valor da estatística W_2 deve ser comparado com o percentil $(1-\alpha)$ de (1.22), conforme tabela para (1.22) apresentada em Hollander & Wolfe, 1973.

Para amostras grandes (pelo menos 10 em cada grupo), W_2 tem aproximadamente distribuição normal. Assim o, valor da estatística Z, abaixo definida, deve ser comparado com o percentil $(1-\alpha)$ da distribuição normal padrão.

$$Z = \frac{W_2 - \dfrac{(n_2)(n_1+n_2+1)}{2}}{[n_1 n_2(n_1+n_2+1)/12]^{1/2}} \xrightarrow{D} \mathrm{N}(0,1).$$

A estatística de Mann-Whitney

Em uma observação $\mathbf{u} = (x_1, x_2, ..., x_{n_1}; y_1, ..., y_{n_2})$, seja M_1 o número de vezes em que observações de X superam as de Y e M_2 o número de vezes em que observações de Y superam as de X:

$$M_2 = \sum I\,[Y > X],$$

prova-se que, quando não há empates, $M_2 = W_2 - \frac{n_2(n_2+1)}{2}$, onde W_2 é a estatística de Wilcoxon (Wilcoxon, 1945).

A estatística de Mann-Whitney (U) é a variável definida como a diferença entre o número de vezes em que observações de Y superam as de X e o número de vezes em que observações de X superam as de Y na amostra combinada

$$U = M_2 - M_1 = \sum I\,[Y > X] - \sum I\,[X > Y]$$

Mostra-se que

$$U = 2W_2 - (n_2)^2 - n_2 - n_2 n_1.$$

e que $E(U) = 0$ e $Var(U) = 4Var(W_2)$ (ver Hollander & Wolfe, 1973).

Sob a hipótese nula, a estatística de teste U tem, aproximadamente, distribuição normal com média zero e variância

$$Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{3},$$

Assim,

$$Z = \frac{U}{[n_1 n_2 (n_1 + n_2 + 1)/3]^{1/2}} \xrightarrow{D} N(0,1).$$

1.4.2 A estatística "log-rank"

Abordagem por meio de tabelas de contingência 2×2, construídas nos tempos de mortes

O teste estatístico de Mantel-Haenszel (1959) é utilizado comumente em situações de tabelas de classificação cruzada com variável resposta dicotômica (por exemplo, morte, sobrevida) e fatores categorizados em níveis (tratamento1, tratamento2). Num caso particular, duas tabelas de vida podem ser utilizadas, neste contexto, para comparação das taxas de sobrevida de dois grupos. Para testar a igualdade de duas funções de sobrevivência $S_1(t)$ e $S_2(t)$, Mantel considera os tempos ordenados $t_{(k)}$, de falha distintos da amostra formada pela combinação das duas amostras individuais. Este teste supõe que no tempo $t_{(k)}$ ocorrem d_k falhas e existem R_k indivíduos em risco em um tempo imediatamente inferior a $t_{(k)}$ na amostra combinada e, respectivamente, d_{1k} e R_{1k} deles no tratamento i, $i = 1, 2.$ e $j = 1, .., k$. Assim, em cada tempo de falha $t_{(k)}$, os dados podem ser dispostos em forma de uma tabela de contingência 2×2, com linhas indicando tratamentos e colunas indicando estados de sobrevivência, como está apresentado na Tabela 1.4.

Tabela 1.4 Distribuição de R indivíduos em risco segundo Tratamento e Sobrevida gerada no tempo $t_{(k)}$

Tratamento	Estado de Sobrevivência		Total
	M	S	
T_1	d_{1k}	$R_{1k} - d_{1k}$	R_{1k}
T_2	d_{2k}	$R_{1k} - d_{2k}$	R_{2k}
Total	d_k	$R_k - d_k$	R_k

A estatística do teste de Mantel compara, em cada tempo de morte, o número observado de mortes em um dos tratamentos, como por exemplo, d_{1k} com o valor esperado sob a hipótese nula H_0, condicionado ao fato de haver d_k mortes neste instante, sendo R_{1k} e R_{2k} os números de indivíduos em risco em cada tratamento.

Se os totais das marginais na Tabela 1.4 são fixos e a hipótese nula é verdadeira, então a distribuição de d_k é uma hipergeométrica. O valor esperado da casela na primeira linha e primeira coluna (Ed_{1k}), onde foi observado um valor $O_{1k} = d_{1k}$, condicional às marginais que

$$\mathrm{E}d_{1k} = d_k \frac{R_{1k}}{R_k} \text{ e sua variância } \mathrm{V}d_{1k} = \frac{dR_{2k}R_{1k}(R_k - d_k)}{R_k^2(R_k - 1)}, \tag{1.23}$$

Desta forma, se não houver diferenças entre os tratamentos no tempo $t_{(k)}$, o número esperado de mortes deve ser dividido entre os dois tratamentos de acordo com a razão entre o número de indivíduos sob risco em cada tratamento e o número total sob risco naquele instante. A estatística proposta por Mantel consiste na soma das diferenças entre o número observado e o número esperado de mortes em cada tempo $t_{(k)}$, isto é:

$$L = \sum_{k=1}^{J}(d_{1k} - \mathrm{E}d_{1k}) = \sum_{k=1}^{J}(O_{1k} - \mathrm{E}d_{1k})\,. \tag{1.24}$$

onde $O_{1k} = d_{1k}$, para $k = 1, ..., J$

Se as tabelas de contingência, para todo $t_{(k)}$, fossem independentes, a variância de L seria simplesmente a soma das variâncias condicionais $\mathrm{V}d_{1k}$, isto é,

$$\mathrm{Var}(L) = \sum_{k=1}^{J}\mathrm{V}d_{1k} = \sum_{k=1}^{J}\frac{d_k R_{2k}R_{1k}(R_k - d_k)}{R_k^2(R_k - 1)},$$

Sob a hipótese nula, L tem uma distribuição assintótica N(0,1) e, assim, um teste aproximado para a comparação das taxas de sobrevivência dos dois grupos pode ser baseado na estatística padronizada:

$$Z_L = \frac{L}{\sqrt{\mathrm{Var}(L)}} \xrightarrow{D} N(0,1), \tag{1.25}$$

ou na forma equivalente desta estatística que é dada por

$$LR = \frac{(L)^2}{\mathrm{Var}(L)}. \tag{1.26}$$

que tem, sob H_0, distribuição qui-quadrado com um grau de liberdade para grandes amostras.

Uma expressão generalizada da estatística L em (1.24) é dada pela expressão

$$\sum_{k=1}^{J} w_k(d_{1k} - \mathrm{E}d_{1k}) \tag{1.27}$$

O teste log-rank (1.24) é obtido tomando-se $w_k = 1$; quando $w_k = R_k$ (total de pessoas em risco), tem-se o teste de Wilcoxon Mann-Whitney (U).

Exemplo 4: Um estudo foi realizado para comparar dois tipos de tratamentos para um determinado tipo de câncer. Para tanto, observou-se o tempo (t) em meses até a morte de onze indivíduos com a doença que foram submetidos aos tratamentos I e II .

A Tabela 1.5 apresenta observações do tempo (t), em meses, até a morte de seis indivíduos submetidos ao tratamento I e cinco ao tratamento II, e os respectivos valores de J_i e t_i.

Tabela 1.5 Tratamento e posição (J_i) na amostra conjunta dos valores observados

Tratamento	1	1	1	1	1	1	2	2	2	2	2
J_i	1	2	3	4	5	6	7	8	9	10	11
t_i	12	58	80	90	95	100	10	14	23	56	79

Na Tabela 1.6 é feita uma representação equivalente das onze observações, através das estatísticas de ordem $t_{(i)}$ e das de postos $P_{(i)}$ e $J_{(i)}$.

Tabela 1.6 Valores de P_i e J_i referentes às estatísticas de ordem observadas

Tratamento	2	1	2	2	2	1	2	1	1	1	1
$P_{(i)}$	1	2	3	4	5	6	7	8	9	10	11
$t_{(i)}$	10	12	14	23	56	58	79	80	90	95	100
$J_{(i)}$	7	1	8	9	10	2	11	3	4	5	6

Na Tabela 1.7, para $i = 1, 2$, apresentam-se os dados já ordenados conjuntamente de forma crescente, exibindo-se os instantes em que ocorreram mortes em qualquer dos grupos, ($t_{(k)}$) e o número de mortes em cada um deles (d_{ik}). Em cada tempo $t_{(k)}$, indica-se também:

-o número de observações maiores que $t_{(k)}$ em cada um dos grupos (R_i), que é o número de indivíduos em risco de morrer no grupo i;

-o número total de observações maiores que $t_{(k)}$, (R), que corresponde ao número total de indivíduos em risco de morrer em $t_{(k)}$;

-o posto de cada observação t_k na amostra conjunta (P_k) e na amostra do tratamento i ($i = 1, 2$) P_i.

Tabela 1.7 Cálculo das Estatísticas baseadas em Postos das observações

t_k	d_{1k}	d_{2k}	R_{1k}	R_{2k}	R_k	Pd_{1k}	Pd_{2k}	Y>X	X>Y	e_{1k}	A	B
10	0	1	6	5	11	0	1	0	0	$\frac{6}{11}$	$\frac{30}{121}$	−6
12	1	0	6	4	10	2	0	0	1	$\frac{6}{10}$	$\frac{24}{100}$	4
14	0	1	5	4	9	0	3	1	0	$\frac{5}{9}$	$\frac{20}{81}$	−5
23	0	1	5	3	8	0	4	1	0	$\frac{5}{8}$	$\frac{15}{64}$	−5
56	0	1	5	2	7	0	5	1	0	$\frac{5}{7}$	$\frac{10}{49}$	−5
58	1	0	5	1	6	6	0	0	4	$\frac{5}{6}$	$\frac{5}{36}$	1
79	0	1	4	1	5	0	7	2	0	$\frac{4}{5}$	$\frac{4}{25}$	−4
80	1	0	4	0	4	8	0	0	5	$\frac{4}{4}$	0	0
90	1	0	3	0	3	9	0	0	5	$\frac{3}{3}$	0	0
95	1	0	2	0	2	10	0	0	5	$\frac{2}{2}$	0	0
100	1	0	1	0	1	11	0	0	5	1	0	0
$\sum$	n_1	n_2				W_1	W_2	M_2	M_1	$\sum E_{d_0}$	var(LR_1)	
$\sum$	6	5				46	20	5	25	8.6736	1.4722	−20

sendo:

$e_{1k} = d_k \frac{R_{1k}}{R_k}$, $\quad A = \frac{d_k R_{1k} R_{2k}(R_k - d_k)}{R_k^2(R_k - 1)}$ e $\quad B = R_k(d_{1k} - e_{1k})$.

Para testar $H_0 : S_1(t) = S_2(t)$ vs $H_1 : S_1(t) < S_2(t)$, pode-se utilizar várias estatísticas, como as que seguem abaixo.

Estatística de Wilcoxon:

O valor observado de W_2 foi 20. Desta forma, conclui-se que a hipótese nula não deve ser rejeitada ao nível $\alpha = 2,5\%$, pois $W_2 \geqslant 20 = [n_2(n_2 + n_1 + 1) - w(0,026; 6; 5)]$ conforme tabela (Hollander & Wolfe, 1973), ou seja, as funções de sobrevivência dos dois tratamentos não são diferentes.

Estatística de Mann-Whitney:

O valor obtido de M_2 foi 5. Logo $U = M_2 - M_1 = 2W_2 - (n_2)^2 - n_2 - n_1 n_2 = \sum R_k(O_k - E_k)$ = -20

Para $n_1 = 6$, $n_2 = 5$ sob H_0 P($M_2 \leq 5$)=0,041 de acordo com a tabela (Hollander & Wolfe, 1973). Portanto, não se rejeita a hipótese nula H_0, isto é, as funções de sobrevivência dos dois tratamentos não são diferentes.

Estatística de Log-rank:

$$L = \left(\sum (O_{1k} - \mathrm{E}d_{1k})\right) = -2,6736$$

$$LR = \frac{\left(\sum (O_k - \mathrm{E}d_{1k})\right)^2}{\sum_{k=1}^{J} \mathrm{V}(d_{1k})} = \frac{(-2,6736)^2}{1,4722} = 4,85541$$

O percentil $0,025$ da distribuição χ^2_1 é $5,02389$, maior que $4,85541$. Assim, as duas funções de sobrevivência dos dois tratamentos não são estatisticamente diferentes.

1.4.3 O teste de log-rank como um teste de escore eficiente segundo uma verossimilhança Parcial num modelo log-linear para as funções de risco em duas populações, com dados não censurados.

Na introdução deste trabalho, diversas estatísticas foram propostas para testar a hipótese H_0: $S_1(t) = S_2(t)$ a respeito da igualdade das funções de sobrevivência de duas variáveis aleatórias T_1 e T_2, sem conhecimento de sua forma paramétrica. No entanto, outros testes podem ser deduzidos para testar H_0 contra certas alternativas a H_0. Uma destas alternativas é

$$S_2(t) = S_1(t)^\gamma. \tag{1.28}$$

A relação acima é equivalente à seguinte relação entre as funções densidade $f_1(t)$ e $f_2(t)$:

$$f_2(t) = \gamma(S_1(t))^{\gamma-1} f_1(t)$$

conseqüentemente a relação entre as funções de risco $h_2(t) = h_1(t)\gamma$ é dada por:

$$h_2(t) = \frac{\gamma(S_1(t))^{\gamma-1} f_1(t)}{(S_2(t))} = \frac{\gamma f_1(t)}{S_1(t)} = \gamma h_1(t)$$

Por este motivo, o modelo é denominado de Riscos Proporcionais, já que para todo $t > 0$, tem-se:

$$\frac{h_2(t)}{h_1(t)} = \gamma.$$

Este modelo pode ser também expresso em um modelo linear para os logaritmos das funções de

risco:

$$\log(h_2(t)) = \log(\gamma) + \log(h_1(t))$$

A hipótese nula $H_0 : S_1(t) = S_2(t)$ passa a ser, no modelo de Riscos Proporcionais, equivalente à

$$H_0\colon\ \gamma = 1.$$

A função de verossimilhança de um conjunto de n_1 observações de T_1 e n_2 observações de T_2 é dada por (1.21) e pode ser expressa em termos das funções de risco, já que é possível estabelecer as relações:

$$h(t) = \frac{\partial \log(1 - F(t))}{\partial(t)},$$

$$f(t) = h(t)(1 - F(t)),$$

$$1 - \mathrm{F}(t) = \exp(-\int_0^t \mathrm{h}(s)ds) = \exp(-\mathrm{H}(t)).$$

A função de verossimilhança, conforme (1.21), é

$$L(t_{11}, .., t_{1n_1}; t_{21}, .., t_{2n_2}; \gamma) = \prod_1^{n_1} [S_1(t_{1i})] \prod_1^{n_1} [h_1(t_{1i})] \prod_1^{n_2} [S_2(t_{2i})] \prod_1^{n_2} [h_2(t_{2i})].$$

e, particularmente, no Modelo de Riscos Proporcionais pode ser escrita como

$$L(t_{11}, .., t_{1n_1}; t_{21}, .., t_{2n_2}; \gamma) = \prod_1^{n_1} [S_1(t_{1i})] \prod_1^{n_1} [h_1(t_{1i})] \prod_1^{n_2} [S_1(t_{2i})^\gamma] \prod_1^{n_2} [\gamma h_1(t_{2i})]. \qquad (1.29)$$

Em conseqüência, o logaritmo da função de verossimilhança é dado por:

$$\begin{aligned} l(t_{11}, .., t_{1n_1}; t_{21}, .., t_{2n_2}; \gamma) \ &= \ -(\sum_1^{n_1} H_1(t_{1i})) + \sum_1^{n_2} \gamma H_1(t_{2i}))) + \sum_1^{n_1} (\log([h_1(t_{1i})]) \\ &\quad + \sum_1^{n_2} (\log([h_1(t_{2i})]) + \sum_1^{n_2} \log(\gamma). \end{aligned}$$

A verossimilhança acima contém, além do parâmetro de interesse γ, uma função $H_1(t_{1i})$ desconhecida, que representa parâmetros de perturbação na inferência sobre γ e, deste modo não é útil para fazer inferências sobre γ. De forma alternativa em relação à verossimilhança completa acima, Cox (1972) definiu uma verossimilhança parcial, que permite a estimação de γ sem conhecer $H_1(t_{1i})$. Sua idéia foi a de transformar a observação das variáveis originais em variáveis $(u_1, u_2, ..., u_{2N})$ e de fazer uso do fato de que uma função distribuição conjunta $g(u_1, u_2, ..., u_{2N})$ sempre pode ser fatorada da seguinte maneira:

$$g(u_1, u_2, ..., u_{2N}) = g(u_1)g(u_2|u_1)g(u_3|u_1, u_2)....g(u_{2N}|u_1, u_2, ..., u_{2N-1})$$

Se for possível obter distribuições condicionais $g(u_{2k}|u_1, u_2, ..., u_{2k-1})$ não dependentes da função básica $h_1(t_1)$, mas apenas de γ ao passo que as distribuições condicionais $g(u_{2k+1}|u_1, u_2, .., u_{2k})$ dependem de γ e também de $h_1(t_i)$, define-se uma função de verossimilhança parcial LP apenas com os primeiros fatores:

$$LP(u_1, .., u_{2N}; \gamma) = \prod_{k=1}^{N} g(u_{2k}|u_1, u_2, ..., u_{2k-1}) \tag{1.30}$$

Supondo a ausência de censura, as observações completas de T_1 e de T_1 : $(t_{11}, .., t_{1n_1}; t_{21}, .., t_{2n_2})$, são representadas pelas estatísticas de ordem da amostra conjunta e pelo vetor de identificadores dos elementos da $(n_1 + n_2)$-upla, correspondente às consecutivas estatísticas de ordem:

$$(t_{(1)}, ..., t_{(n_1+n_2)}) \text{ e } (J_1, ..., J_{n_1+n2}).$$

Intercalando os valores das duas, definimos uma $2(n_1 + n_2)$-upla de valores $(u_1, u_2, .., u_{2(n_1+n_2)})$: $u_1 = t_{(1)}$; $u_2 = J_1$; $u_3 = t_{(2)}$; $u_4 = J_2$; $u_5 = t_{(3)}$ e, genericamente $u_{2k-1} = t_{(k)}$; $u_{2k} = J_k$, $k = 0, ..., N = (n_1 + n_2)$.

O cálculo das distribuições condicionais $g(u_{2k}|u_1, u_2, ..., u_{2k-1})$ é descrito abaixo e é possível observar que $g(u_2|u_1)$ é uma distribuição discreta, uma vez que u_2 pode assumir os valores $1, .., N$. Dado que todos os indivíduos estavam vivos antes de $t_{(1)}$, a probabilidade do i-ésimo

vir a morrer em um intervalo de tempo $(t_{(1)} + \gamma)$ é dada aproximadamente por:

$$h_i(t_{(1)})\gamma,$$

ao passo que a probabilidade de qualquer um dos indivíduos morrer será

$$\sum_{i=1}^{n_1+n_2} h_i(t_{(1)})\gamma.$$

Logo, a probabilidade condicional de que J_1 seja igual a i será

$$g(u_2|u_1) = P(J_1 = i|t_{(1)}) = \frac{h_i(t_{(1)})\gamma}{\sum_{j=1}^{n_1+n_2} h_j(t_{(1)})\gamma} = \frac{h_i(t_{(1)})}{\sum_{j=1}^{n_1+n_2} h_j(t_{(1)})},$$

Como no modelo em estudo

$$h_j(t_{(1)}) = (\gamma)_j h_1(t_{(1)})$$

onde $(\gamma)_j$ =1 para $j = 1, .., n_1$ e $(\gamma)_j = \gamma$ para $j = n_1 + 1, ..., n_1 + n_2$, tem-se que

$$g(u_2|u_1)=\frac{\gamma_{J_1}}{\sum_{j=1}^{n_1+n_2} \gamma_j}.$$

Da mesma forma, $g(u_4|u_1, u_2, u_3)$ depende somente do número de mortes já ocorridas e da identificação do indivíduo J_1, que é dado por:

$$g(u_4|u_1, u_2, u_3) = \frac{\gamma_{J_2}}{\sum_{j=1\ j\neq J_1}^{n_1+n_2} \gamma_j},$$

ou, de forma alternativa, indicando o conjunto de indivíduos com tempo de morte igual ou superior à k-ésima menor morte por $\mathcal{R}_k$ ={J_i para $i \geqslant k$}

$$g(u_4|u_1, u_2, u_3) = \frac{\gamma_{J_2}}{\sum_{j\in\mathcal{R}_2} \gamma_j},$$

genericamente,

$$g(u_{2k}|u_1, u_2, ..., u_{2k-1}) = \frac{\gamma_{Jk}}{\sum_{j \in \mathcal{R}_k} \gamma_j}.$$

Nota-se que esta densidade não depende funcionalmente dos valores observados de tempos de morte $(u_1, u_3, .., u_{2k-1})$, de modo que:

$$g(u_{2k}|u_1, u_2, ..., u_{2k-1}) = g(u_{2k}|u_2, u_4, ..., u_{2k-2})$$

ou

$$g(u_{2k}|u_1, u_2, ..., u_{2k-1}) = g(J_k|J_1, J_2, ..., J_{k-1}).$$

Deste modo, é possível encontrar

$$g(J_1, J_2, .., J_{N-1}, J_N) = g(J_1)g(J_2|J_1)...g(J_N|J_1, J_2, .., J_{N-1}) = \prod_{s=1}^{N} \frac{\gamma_{Js}}{\sum_{j \in \mathcal{R}_s} \gamma_j}.$$

Uma verossimilhança parcial definida pela seqüência $(u_1, u_2....u_{2(n_1+n_2)})$ será a densidade marginal de $(J_1, J_2, ..., J_{k-1}, J_N)$:

$$LP(\gamma) = \prod_{s=1}^{N} \frac{\gamma_{Js}}{\sum_{j \in \mathcal{R}_s} \gamma_j}.$$

$\mathcal{R}_s = \{J_i \text{ para } i \geqslant k\}$ é o conjunto de indivíduos em risco no instante de tempo igual à s-ésima estatística de ordem do grupo $(t_{(s)})$ formado por todos os indivíduos em estudo, J_s é o indivíduo que falha no tempo $t_{(s)}$ e γ_j é um parâmetro que vale 1, se j for um índice relativo a uma observação de T_1, e vale γ, se j for um índice relativo a uma observação de T_2.

A função de log-verossimilhança parcial relativa a γ é dada por

$$lp(\gamma) = \sum_{J_s=1}^{n_1} \log(\gamma_{Js}) + \sum_{J_s=n_1+1}^{n_1+n_2} \log(\gamma_{Js}) - \sum_{J_s=1}^{n_1+n_2} \log(\sum_{j \in \mathcal{R}_s} \gamma_j).$$

A derivada do logaritmo de LP em relação a γ, chamada função escore é indicada por $\mathrm{U}_\gamma(\gamma)$:

$$\mathrm{U}_\gamma(\gamma)=\sum_{J_s=n_1+1}^{n_1+n_2}(\frac{1}{\gamma})-\sum_{J_s=1}^{n_1+n_2}\frac{(\sum_{j\in\mathcal{R}_s}\gamma_j)}{(\sum_{j\in\mathcal{R}_s}\gamma_j)}=\sum_{J_s=n_1+1}^{n_1+n_2}(\frac{1}{\gamma})-\sum_{J_s=1}^{n_1+n_2}\frac{R_{2.s}}{(R_{2,s}\gamma+R_{1,s})}. \tag{1.31}$$

Sob a hipótese H_0: $\gamma=1$ a estatística U(1) coincide, com a estatística de "log-rank(L)" já que

$$\mathrm{U}_\gamma(1)=n_2-\sum_{J_s=1}^{n_1+n_2}\frac{R_{2.s}}{(R_{1,s}+R_{2,s})}=n_2\text{-}\sum_{J_s=1}^{n_1+n_2}[1-\frac{R_{1.s}}{(R_{1,s}+R_{2,s})}]$$

$$\mathrm{U}(1)=n_1\text{-}\sum_{J_s=1}^{n_1+n_2}\frac{R_{1.s}}{(R_{1,s}+R_{2,s})} \tag{1.32}$$

onde R_{1s} (R_{2s}) é o número de pessoas em risco no grupo 1 (2). A fórmula (1.32) pode ser ainda escrita como (1.24) isto é

$$\mathrm{U}(1)=\sum_{s=1}^{n_1+n_2}(d_{1s}-E(d_{1s}))=\sum_{s=1}^{n_1+n_2}(O_{1s}-E_{1s}) \tag{1.33}$$

No caso aqui considerado, sem observações censuradas, são calculadas n_1+n_2 tabelas de contingência e a estatística de log-rank é dada por $L=\sum_{s=1}^{n_1+n_2}(d_{1s}-E(d_{1s}))$ com $\mathrm{E}_{d_{1s}}=d_s\frac{R_{1s}}{R_s}$, onde $\sum d_{1s}=n_1$, número total de mortes no grupo 1 e $d_s=1$ é o número de mortes em qualquer um dos grupos em cada um dos tempos $t_{(s)}$.

O escore obtido através de uma verossimilhança completa, segundo uma parametrização, tem variância dada pela esperança da sua derivada em relação ao parâmetro, ou ainda do oposto da segunda derivada da função de log verossimilhança, que é denominada informação de Fisher. A segunda derivada da verossimilhança parcial Uγ é dada por:

$$\mathrm{U}'_\gamma(\gamma)=\sum_{J_s=n_1+1}^{n_1+n_2}(\frac{-1}{\gamma^2})-\sum_{J_s=1}^{n_1+n_2}\frac{R_{2.s}(-R_{2,s})}{(R_{2,s}\gamma+R_{1,s})^2}. \tag{1.34}$$

e portanto, U$'(\gamma)$ é uma soma das contribuições de cada indivíduo. Para um indivíduo, sob $H_0:\gamma=1$

$$\mathrm{U}'(1)=\sum_{J_s=n_1+1}^{n_1+n_2}(-1)+\sum_{J_s=1}^{n_1+n_2}\frac{(R_{2,s})^2}{(R_s)^2}. \tag{1.35}$$

Na parametrização acima, a estatística escore coincide com a estatística log-rank, mas o mesmo não ocorre com suas variâncias. Uma reparametrização levará a um escore que ainda coincide com a estatística log-rank e com a mesma variância que esta. No modelo acima, é comum fazer a reparametrização de γ para $\beta = \log(\gamma)$, isto é, $\gamma = e^{\beta}$, e escrever as duas funções S_1 e S_2 na forma de um modelo de regressão, introduzindo artificialmente uma covariável Z indicadora de grupo, Z valendo 0 para o grupo 1 e 1 para o grupo 2. Dessa forma, a relação (1.28) passa a ser escrita como

$$S(t, Z) = (S_1(t))^{\exp(\beta Z)} \tag{1.36}$$

O escore $\mathrm{U}_{\beta}(\beta)$, relativo à mesma amostra de n_1 valores de T_1 e n_2 valores de T_2, dados acima, está relacionado com $\mathrm{U}_{\gamma}(\gamma)$, apresentado em (1.31), da seguinte forma

$$\mathrm{U}_{\beta}(\beta) = (\mathrm{U}_{\gamma}(\gamma))e^{\beta}$$

Conseqüentemente,

$$\mathrm{U}'_{\beta}(\beta) = (\mathrm{U}'_{\gamma}(\gamma))e^{2\beta} + \mathrm{U}_{\beta}(\beta) \tag{1.37}$$

Um teste de $H_0 : \beta = 0$ versus $H_0 : \beta \neq 0$ pode ser feito utilizando-se o escore eficiente, que é $(\mathrm{U}_{\beta}(\beta))_{\beta=0}$, cuja variância é dada pelo oposto da informação de Fisher, $\mathrm{E}((\mathrm{U}'_{\beta}(\beta))_{\beta=0})$.

Substituindo (1.34) em (1.37), lembrando que se $\gamma = 1$ então $\beta = 0$, obtém-se

$$\mathrm{U}'_{\beta}(0) = (\mathrm{U}'_{\gamma}(1)) + \mathrm{U}_{\beta}(0) \tag{1.38}$$

Observando que $\mathrm{U}_{\beta}(0)$ é igual (1.32), conclui-se que a estatística escore coincide com a estatística log-rank e substituindo $\mathrm{U}_{\beta}(0)$ e também (1.35) em (1.38), obtém-se

$$\mathrm{U}'_{\beta}(0) = -(\sum_{J_s=1}^{n_1+n_2} \frac{R_{2.s}(R_{1,s})}{(R_s)^2}) \tag{1.39}$$

Em resumo, **o escore eficiente na parametrização em β do modelo de riscos proporcionais é a estatística de log-rank $\sum_{k=1}^{J} (O_{1k} - \mathrm{E}_{1k})$ cuja variância é estimada por** (1.39).

1.5 Testes não-paramétricos para comparação de duas funções de sobrevivência com observações censuradas

Muitas vezes, a observação completa de duas variáveis T_1 e T_2 é impossível. Ao invés disso, só é possível analisar (U,V,δ_X,δ_Y), onde

$$\begin{aligned} \mathrm{U} &= \min(T_1,\mathrm{C}_1),\ \mathrm{C}_1 \text{ e } T_1 \text{ independentes} \\ \mathrm{V} &= \min(T_2,\mathrm{C}_2),\ \mathrm{C}_2 \text{ e } T_2 \text{ independentes} \end{aligned} \tag{1.40}$$

$$\delta_{T_1} = \begin{cases} 1 \text{ se } \mathrm{U} = T_1 \\ 0 \text{ se } \mathrm{U} = \mathrm{C}_1 \end{cases}$$

$$\delta_{T_2} = \begin{cases} 1 \text{ se } \mathrm{V} = T_2 \\ 0 \text{ se } \mathrm{V} = \mathrm{C}_2 \end{cases}$$

Considerando-se o problema de testar a hipótese $H_0 : S_1(t) = S_2(t)$, a partir de n_1 observações i.i.d. de (U,δ_{T_1}) e n_2 observações de (V,δ_{T_2}), a função de verossimilhança passa a ser dada por

$$L(t_{11},..,t_{1n_1};t_{21},..,t_{2n_2};\gamma) \propto \prod_1^{n_1}[S_1(t_{1i})]\prod_1^{n_1}[h_1(t_{1i})]^{\delta_i}\prod_1^{n_2}[S_2(t_{2i})]\prod_1^{n_2}[h_2(t_{2i})]^{\delta_i}, \tag{1.41}$$

onde foram omitidos os fatores relativos às distribuições de C_1 e C_2, supostas não informativas sobre $S_1(t)$ e $S_2(t)$.

Para o modelo de riscos proporcionais, tem-se:

$$L(t_{11},..,t_{1n_1};t_{21},..,t_{2n_2};\gamma) \propto \prod_1^{n_1}[S_1(t_{1i})]\prod_1^{n_1}[h_1(t_{1i})]^{\delta_i}\prod_1^{n_2}[S_1(t_{2i})]^{\gamma}]\prod_1^{n_2}[\gamma h_1(t_{2i})]^{\delta_i}$$

$$l(t_{01},..,t_{0n_0};t_{11},..,t_{1n_1};\gamma) = -(\sum_{1}^{n_1} H_1(t_{1i})) + (\sum_{1}^{n_2} \gamma H_1(t_{2i})) + (\sum_{1}^{n_1} \delta_{t_{1i}} \log(h_1(t_{1i})) + \sum_{1}^{n_2} \delta_{t_{2i}} \log(h_1(t_{2i})) + \sum_{1}^{n_2} \delta_{t_{2i}} \log(\gamma)$$

A inferência sobre γ a partir da função acima, por envolver a função desconhecida $h_1(T_1)$, não é possível. Uma solução proposta é a inferência a partir de uma função de verossimilhança parcial, como foi feito para o caso sem censura.

1.5.1 Extensões da estatística log-rank

Para dados com a presença de censuras, existe uma variedade de testes de H_0, ressaltando-se, entre eles, as generalizações do teste de Wilcoxon propostas por Gehan (1965), o teste de log-rank (Mantel, 1966; Pete e Peto, 1972) e suas generalizações.

A expressão das estatísticas log-rank para testar H_0, em analogia aos testes já considerados para observações completas em 1.4.2, é dada pela fórmula:

$$\sum_{k=1}^{J} w_k \left(d_{1k} - E(d_{1k})\right). \tag{1.42}$$

onde a soma é feita apenas **para aquelas estatísticas de ordem que correspondem às observações não censuradas** da amostra conjunta, e os pesos são dados por:

w_k =1 (estatística de log-rank)

w_k =R_k (estatística de Gehan) Wilcoxon generalizado

w_k =$(R_k)^{\frac{1}{2}}$ (estatística de Tarone Ware)

w_k =$\hat{S}(t_k)$ (estatística de Peto e Peto, 1972. $\hat{S}(z_k)$ é a estimativa de Kaplan-Meier de $1-F(t_k)$ sob H_0).

$w_k = [\hat{S}(t_k)]^J$, $0 \leq J \leq 1$ (Fleming e Harrington, 1991).

Nas expressões acima, $\hat{S}(t_k)$ é a estimativa de Kaplan-Meier, a partir da amostra combinada de todas as observações calculadas no tempo de falha t_k. Estas estatísticas são chamadas estatísticas de log-rank ponderadas (Fleming, Harrington, 1991).

A escolha do peso na expressão (1.42) direciona o tipo de diferença a ser detectado nas

funções de sobrevivência. O Teste generalizado de Wilcoxon (Gehan), que utiliza o peso igual ao número de indivíduos sob risco, coloca mais peso no início do eixo do tempo. No início do estudo, todos os indivíduos estão sob risco e saindo do estado "sob risco" à medida que falham ou são censurados. O teste log-rank, por outro lado, coloca menos peso para todo o eixo do tempo, o que reforça o enfoque nos pesos maiores quando comparado ao teste de Gehan. O teste de Tarone-Ware está em uma situação intermediária e o de Gehan pondera pelo número de indivíduos sob risco que depende da experiência de sobrevivência assim como da censura. Desta forma, se o padrão de censura é nitidamente diferente nos dois grupos, o teste pode rejeitar ou aceitar não somente com base nas diferenças das sobrevivências entre os grupos, mas também devido ao padrão de censura. (Colosimo, 2001).

O teste log-rank (Mantel, 1966), já apresentado na seção 1.4.2, é o mais utilizado em análise de sobrevivência, mas deve ser usado com cautela em situações em que duas funções de risco não são proporcionais ou se cruzam. Pepe & Fleming (1989) propõem uma classe de estatísticas de teste, apresentada na próxima seção, para comparar funções de sobrevivência de dois grupos nas situações em que a suposição de riscos proporcionais não é válida.

1.5.2 A estatística "Weighted Kaplan-Meier"

No problema de comparação da distribuição dos tempos de sobrevivência sujeitos a um mecanismo de censura aleatória, em duas populações, Pepe e Feming (1989) propuseram a estatística *Weighted Kaplan-Meier* (WKM) dada por:

$$WKM = \sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{w}(t) \left[\hat{S}_{KM1}(t) - \hat{S}_{KM2}(t)\right] dt, \tag{1.43}$$

onde a diferença entre os estimadores de Kaplan-Meier para os dois grupos é ponderada em cada tempo por um peso que os autores propuseram que fosse dado por:

$$\hat{w}(t) = \frac{\hat{C}^-_{KM1}(t)\hat{C}^-_{KM2}(t)}{\hat{p}_1\hat{C}^-_{KM1}(t) + \hat{p}_2\hat{C}^-_{KM2}(t)}, \tag{1.44}$$

onde

-o sobrescrito $(-)$ indica que foi usada a versão contínua à esquerda da função,

-$T_c = \sup\{t : \hat{C}_1(t) \wedge \hat{C}_2(t)\} > 0\}$, e $\wedge$ representa o mínimo,

-$\hat{S}_i(t)$ é o estimador de Kaplan-Meier da função de sobrevivência da variável T no grupo $i = 1, 2$,

-$\hat{C}_i(t)$ é o estimador de Kaplan-Meier da função de sobrevivência da variável tempo de censura C no grupo $i = 1, 2$,

-n_i é o tamanho da amostra no grupo i e $n = n_1 + n_2$.

-Os pesos p_1 e $p_2 = 1 - p_1$ são estimados adequadamente sendo $\hat{p}_i = \frac{n_i}{n}$. Assume-se que $\lim_{n\to\infty} \frac{n_i}{n} = p_i > 0$.

-O intervalo de tempo relevante para as comparações é $[0, \tau]$ sendo $\tau = \sup\{t : \hat{S}(t) \wedge \hat{C}_1(t) \wedge \hat{C}_2(t)\} > 0\}$. Deve-se considerar somente funções $\hat{w}(.)$, que são bons estimadores de algum $w(.)$ tal que

$$\sup_{t\in(0,\tau)} |\hat{w}(t) - w(t)| \xrightarrow{P} 0. \tag{1.45}$$

$$|w(t)| \leq \Gamma\ [C_i^-(t)]^{(1/2)+\delta} \qquad \text{e} \qquad |\hat{w}(t)| \leq \Gamma\ [\hat{C}_i^-(t)]^{(1/2)+\delta} \tag{1.46}$$

para todo $t < \tau$, $i = 1, 2$ e Γ e δ constantes positivas. A condição (1.46) exige que $\hat{w}$ seja função de $\hat{C}_1^-$ e $\hat{C}_2^-$, como definida anteriormente.

Pepe & Fleming (1989) demonstram que a distribuição assintótica desta estatística, sob H_0, é normal, com média zero e variância σ^2,

$$WKM = \sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{w}(t) \left[\hat{S}_{KM1}(t) - \hat{S}_{KM2}(t)\right] dt \xrightarrow{D} N(0, \sigma^2),$$

onde

$$\sigma^2 = -\int_0^{\tau} \frac{[\int_t^{\tau} w(u)S(u)du]^2}{S^2(t)} \left(\frac{p_1 C_1^-(t) + p_2 C_2^-(t)}{C_1^-(t) * C_2^-(t)}\right) dS(t).$$

Esta expressão da variância não é intuitiva, embora no caso de dados não censurados com $w(.) = 1$, pode-se verificar que σ^2 é a variância da função distribuição $1 - S(.)$.

Os seguintes estimadores para σ^2, sendo um deles a variância não combinada e outro, a variância amostral de duas amostras combinadas, são respectivamente:

$$\hat{\sigma}^2_{nc} \equiv -\sum_{i=1}^{2} \hat{p}_{3-i} \frac{n_i}{n_i - 1} \int_0^{T_c} \frac{[\int_t^{T_c} \hat{w}(u)\hat{S}_{KMi}(u)du]^2}{\hat{S}_{KMi}(t)\hat{C}^-_{KMi}(t)\hat{S}_i^-(t)} d\hat{S}_i(t),$$

e

$$\hat{\sigma}_c^2 \equiv -\int_0^{T_c} \frac{[\int_t^{T_c} \hat{w}(u)\hat{S}_{KM}(u)du]^2}{\hat{S}_{KM}(t)\hat{S}_{KM}^{-}(t)} \left(\frac{\hat{p}_1\hat{C}_{KM1}^{-}(t) + \hat{p}_2\hat{C}_{KM2}^{-}(t)}{\hat{C}_{KM1}^{-}(t) * \hat{C}_{KM2}^{-}(t)} \right) d\hat{S}(t),$$

onde $\hat{S}(.)$ é o estimador de Kaplan-Meier da função de sobrevivência da amostra combinada.

A hipótese $H_0 : S_1 = S_2$ pode ser testada utilizando-se $(n_1 n_2)^{1/2} \frac{WKM}{\hat{\sigma}}$ e comparando-se o valor obtido com o da distribuição normal padrão.

Exemplo 5: Em um estudo clínico, observou-se 40 pacientes com leucemia após terem entrado num período de remissão da doença. Estes pacientes foram aleatorizados para receber dois tipos de tratamento, vinte receberam o tratamento A e vinte o tratamento B. Posteriormente, os pacientes ficaram em observação até o retorno da doença (recaída) ou até ao fim do estudo (indicador de censura). Os valores observados dos tempos de sobrevivência, que são as durações do período de remissão da doença nos pacientes, são apresentados na Tabela 1.8.

Tabela 1.8: Dados dos tempos de remissão em semanas de pacientes com leucemia

Tratamento A	1	3	3	6	7	7	10	12	14	15
	18	19	22	26	28*	29	34	40	48*	49*
Tratamento B	1	1	2	2	3	4	5	8	8	9
	11	12	14	16	18	21	27*	31	38*	44

Fonte: Lawless, J.F. (1982)

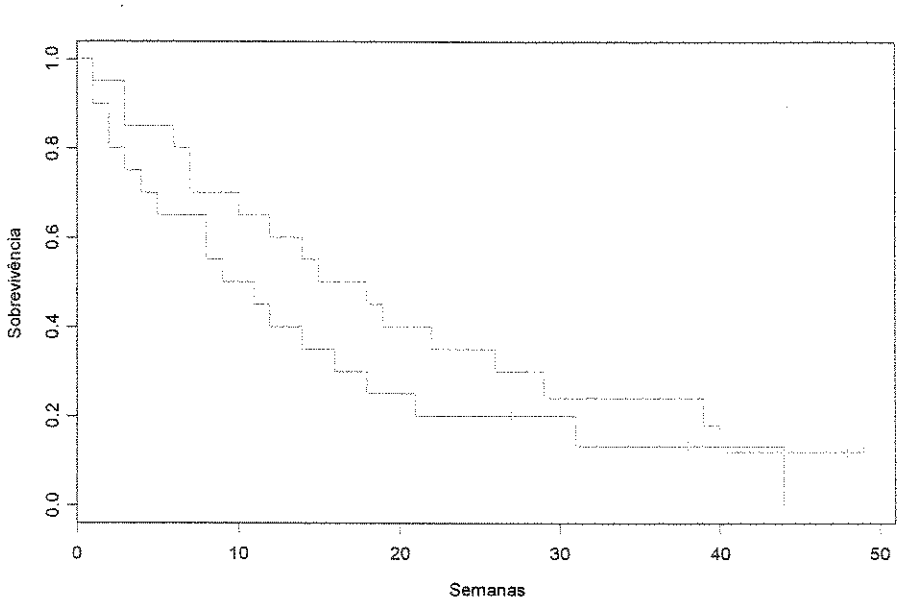

Figura 1.5: Estimativas das funções de sobrevivências dos pacientes submetidos ao tratamento A (laranja) ou tratamento B (Azul).

A hipótese a ser testada é a de que a distribuição do tempo de remissão é a mesma para os dois tratamentos. Os resultados obtidos a partir da utilização das estatísticas de teste log-rank e WKM estão dispostos na Tabela 1.9.

Tabela 1.9: Valores das estatísticas log-rank, WKM e P-valor atingido

Estatística	Nível Descritivo Atingido (P-valor)
$LR = 1,3$	0,2457$\star$
$WKM = -935,1$	0,81$\star\star$

$\star$ P-valor obtido pela distribuição assintótica da estatística.

$\star\star$ P-valor empírico obtido através de reamostragem da estatística.

Portanto, para $\alpha = 5\%$, não se rejeita H_0, isto é, as distribuições do tempo de remissão dos dois tratamentos não são estatisticamente diferentes.

1.5.3 Modelos de regressão

Os ensaios clínicos muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Por exemplo, o tempo livre da doença após o tratamento e o número de

disseminações são covariáveis que devem ser estudadas para verificar o efeito desses fatores no tempo de sobrevivência de mulheres com câncer de mama. É importante destacar, no entanto, que as técnicas não-paramétricas apresentadas nas seções anteriores não permitem a inclusão de covariáveis na análise.

A forma mais eficiente de incorporar o efeito das covariáveis envolvidas no estudo é utilizar um modelo de regressão apropriado para dados censurados. O objetivo dos modelos de regressão é determinar como T, o tempo de falha, varia com $\mathbf{z}$, vetor de covariáveis, usando um modelo apropriado para a dependência de T e $\mathbf{z}$ com dados do tipo $(t, z), i = 1, 2, ..., n$. Existem duas classes de modelos: os paramétricos e os semi-paramétricos. Neste trabalho, é considerada somente a segunda classe que tem, como caso especial, o modelo de regressão de Cox, muito utilizado em estudos clínicos. No entanto, este modelo não se ajusta a qualquer situação clínica, pois ele tem uma suposição básica que é a de riscos proporcionais, ou seja, a razão de riscos de morte relativa a dois indivíduos não muda com o tempo. A violação desta suposição pode provocar sérios vícios na estimação dos coeficientes de regressão do modelo (Struthers e Kalbfleisch, 1986).

1.5.4 Modelo de riscos proporcionais de Cox

O Modelo de riscos proporcionais de Cox supõe que o vetor de covariáveis $\mathbf{z}$ influencie multiplicativamente uma função de risco básica $h_0(t)$, de forma que a função de risco $h(t; \mathbf{z})$ de um indivíduo com vetor de covariáveis $\mathbf{z}$ seja dada por:

$$h(t; \mathbf{z}) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}). \tag{1.47}$$

A equação acima determina que a razão de funções de riscos no mesmo tempo t, entre dois indivíduos com vetores de covariáveis $\mathbf{z_1}$ e $\mathbf{z_2}$, não dependa do particular tempo t e seja dada por :

$$RR = \frac{h_1(t; \mathbf{z}_1)}{h_2(t; \mathbf{z}_2)} = \frac{h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}_1)}{h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}_2)} = \exp(\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)). \tag{1.48}$$

Devido a tal fato, este modelo recebeu o nome de "Modelo de Riscos Proporcionais", sendo $\exp(\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2))$ o fator de proporcionalidade entre funções de risco de indivíduos com vetores

de covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, respectivamente.

Da expressão (1.48) segue-se que:

$$h_1(t;\mathbf{z}_1) = h_2(t;\mathbf{z}_2)\exp(\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)).$$

Da mesma forma,

$$\frac{H_1(t;\mathbf{z}_1)}{H_2(t;\mathbf{z}_2)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_1)\int h_0(s)ds}{\exp(\boldsymbol{\beta}'\mathbf{z}_2)\int h_0(s)ds} = \exp(\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)). \tag{1.49}$$

$$H_1(t;\mathbf{z}_1) = H_2(t;\mathbf{z}_2)\exp(\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)).$$

Em termos de logaritmos, temos:

(a) $\ln h_1(t;\mathbf{z}_1) = \ln h_2(t;\mathbf{z}_2) + \boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)$ e (b) $\ln H_1(t;\mathbf{z}_1) = \ln H_2(t;\mathbf{z}_2) + \boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)$.

Segundo este modelo, todos os grupos de indivíduos que tiverem o mesmo valor K para $\exp(\boldsymbol{\beta}'\mathbf{z})$, têm uma função de risco com mesmo fator de proporcionalidade em relação à função de risco de qualquer outro grupo. Se $\mathbf{z}$ é um vetor bidimensional, a condição $\exp(\boldsymbol{\beta}'\mathbf{z}) = K$ pode ser expressa por $\beta_1 z_1 + \beta_2 z_2 = \ln(K)$.

O Modelo de riscos proporcionais pode ainda ser escrito em termos das funções de Sobrevivência da seguinte maneira:

$$S_i(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\mathbf{z}_i)} \tag{1.50}$$

1.5.5 Estimação de máxima verossimilhança parcial para os coeficientes de regressão no modelo de riscos proporcionais

O processo de estimação de máxima verossimilhança, apresentado no capítulo 3, não pode ser usado se $f(t_i, \boldsymbol{\theta})$ não for completamente determinada, como ocorre, por exemplo, no Modelo de Regressão de Riscos Proporcionais (1.47), quando $h_0(t)$ não tem sua forma especificada. Como muitas vezes o objetivo principal de uma pesquisa é estimar o efeito relativo de dois fatores, representados por duas covariáveis, quanto à sobrevivência, Cox (1972) propôs uma maneira de estimar os coeficientes de regressão, neste contexto, através do conceito de verossimilhança parcial (Cox, 1975) que passamos a expor.

Em uma amostra de n indivíduos, com d mortes, sejam $t_{(1)} \leq t_{(2)} \leq .... \leq t_{(d)}$ os d tempos

de falha ordenados e assumindo-se não empates, e os correspondentes vetores de covariáveis $\mathbf{Z}_{(1)}, \mathbf{Z}_{(2)}, ...\mathbf{Z}_{(d)}$, são representados na figura abaixo:

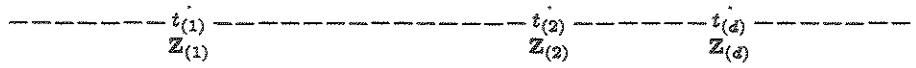

Figura 1.6: Representação dos tempos ordenados de mortes em uma amostra de tamanho n.

Seja $R(t_{(i)})$ o conjunto de indivíduos em risco antes de $t_{(i)}$(isto é, em $t_{(i)} - 0$). A função de densidade condicional de que o indivíduo (i) falhe em $t_{(i)}$, dado os pacientes em $R(t_{(i)})$ estarem em risco e que exatamente uma falha ocorra em $t_{(i)}$ é:

$$\mathrm{P}[\textit{sujeito } (i) \textit{ falha em } t_{(i)} | \textit{falha em } t_{(i)} \textit{ e } R(t_{(i)})] =$$

$$= \frac{\mathrm{P}[\text{sujeito } (i) \text{ falha em } t_{(i)} \text{ e os indivíduos em} R(t_{(i)}) \text{ sobreviveram a } t_i^-]}{\mathrm{P}[\text{falha em } t_{(i)} \text{ e dado que os indivíduos em } R(t_{(i)}) \text{ sobreviveram a } t_i^-]} \tag{1.51}$$

Por (1.4) segue que:

$$\mathrm{P}[\textit{sujeito } (i) \textit{ falha em } t_{(i)} | \textit{falha em } t_{(i)} \text{ e } R(t_{(i)})] = \frac{h_0(t_{(i)}) \exp(\boldsymbol{\beta}' z_{(i)})}{\sum_{j \in R(\, t_{(i)})} h_0(t_{(i)}) \exp(\boldsymbol{\beta}' z_j)} = \frac{\exp(\boldsymbol{\beta}' z_{(i)})}{\sum_{j \in R(\, t_{(i)})} \exp(\boldsymbol{\beta}' z_j)} \tag{1.52}$$

Assim, a verossimilhança parcial para $\boldsymbol{\beta}$ é formada pelo produto de todos os termos representados por (1.52) associados aos distintos tempos de falhas,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}' \mathbf{z}_{(i)})}{\sum_{j \in R(\, t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{z}_j)} \tag{1.53}$$

$$= \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \mathbf{z}_i)}{\sum_{j \in R(\, t_i)} \exp(\boldsymbol{\beta}' \mathbf{z}_j)} \right]^{\delta_i} \tag{1.54}$$

onde δ é o indicador de falha. Os valores de $\boldsymbol{\beta}$ que maximizam a função de verossimilhança

parcial de (1.53) são obtidos resolvendo o sistema de equações definido por $U(\beta) = 0$, onde $U(\beta)$ é o vetor escore de primeiras derivadas da função $l(\beta) = \log L(\beta)$. Isto é

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[z_i - \frac{\sum_{j \in R(t_i)} z_j \exp(\hat{\beta}' \mathbf{z}_j)}{\sum_{j \in R(t_i)} \exp(\hat{\beta}' \mathbf{z}_j)} \right] = 0 \tag{1.55}$$

A função de verossimilhança parcial (1.53) determina que os tempos de sobrevivência são contínuos e, conseqüentemente, não pressupõe a possibilidade de empates nos valores observados, mas na prática podem ocorrer empates nos tempos de falhas ou censuras devido a escala de medida. Quando ocorrem empates nas falhas ou censuras, usa-se a convenção de que a censura ocorreu após a falha, o que define as observações a serem incluídas no conjunto de risco em cada tempo de falha.

A função de verossimilhança parcial (1.53) deve ser modificada para incorporar as observações empatadas quando estas estão presentes. A verossimilhança fornecida por Breslow, 1974 é:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' \mathbf{s}_{(i)})}{\left(\sum_{j \in R(t_{(i)})} \exp(\beta' \mathbf{z}_{(j)}) \right)^{d_i}} \tag{1.56}$$

onde $\mathbf{s}_{(i)} = \sum z_i$ é a soma das covariáveis no empate e d_i é o número de empates.

1.5.6 Estimação da função de sobrevivência básica

Os coeficientes de regressão β são quantidades de maior interesse na modelagem estatística de dados mas, no modelo de Cox, também é importante a função básica de risco acumulado

$$H_0(t) = \int_0^t h_0(u) du$$

e a função de sobrevivência básica correspondente

$$S_0(t) = \exp(-H_0(t)).$$

Se $h_0(t)$ fosse especificada parametricamente, poderia ser estimada usando a função de verossimilhança.

Um estimador não paramétrico simples para $H_0(t)$, proposto por Breslow (1972), conforme Cox (1984), é obtido notando que o número total de falhas em $(0,t)$

$$D(t) = \sum_{t_j,t} d_j.$$

tem a mesma esperança que a variável aleatória $V(t)$ definida por

$$V(t) = \sum_{i=1}^{n} H_i(Y_i(t))$$

Na expressão acima, $Y_i(t) = \min(T_i, C_i, t)$, isto é, $Y_i(t)$ é o tempo de falha, nos indivíduos que já apresentaram falha antes de t, é igual ao tempo de censura para um indivíduo que saiu de observação antes de apresentar falha e é igual a t para um indivíduo ainda sob observação no tempo de calendário t.

$$V(t) = \int_0^t \sum_{j \in \Re(u)}^{n} \exp(\beta' \mathbf{z}_j) h_0(u) du,$$

e seu estimador

$$\hat{H}_0(t) = \sum_{t_j<t} \frac{d_j}{\sum_{j \in \Re(t_j)}^{n} \exp(\hat{\beta}' \mathbf{z}_j)} = \sum_{t_j<t} \frac{d_j}{\sum_{j \in \Re(t_j)} \exp(\hat{\beta}' \mathbf{z}_j)}$$

onde $\exp(\hat{\beta}' \mathbf{z}_{(j)})$ são os valores estimados de $\exp(\beta' \mathbf{z}_{(j)})$. É possível notar que para uma única amostra homogênea, este estimador reduz-se ao estimador de risco acumulado (Nelson, 1969) apresentado em (1.54).

A função de sobrevivência básica $S_0(t)$ pode ser estimada por $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$. Combinando este estimador com as estimativas $\exp(\hat{\beta}' \mathbf{z})$, $i = 1, ..., n$, obtêm-se os valores estimados para a função de risco e a de sobrevivência do indivíduo i.

$$\hat{H}_i(t) = \exp(\hat{\beta}' \mathbf{z}_i) \hat{H}_0(t),$$

$$\hat{S}_i(t) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}' \mathbf{z}_i)}.$$

A consideração de covariáveis na comparação de dois tratamentos quanto ao tempo de sobrevivência pode aumentar o poder dos testes e reduzir os vícios de estimação da diferença de sobrevivências. O modelo de Regressão de riscos proporcionais de Cox é muito utilizado para incorporar o efeito de covariáveis na comparação de dois grupos e, neste caso, uma das covariáveis é a função indicadora de grupo. O coeficiente β_0 correspondente à covariável grupo tem a interpretação de log do risco relativo de morte quando as funções de riscos são proporcionais. O teste de $\beta_0 = 0$ sob o modelo de Cox é chamado de teste de Cox para duas amostras.

Modelo estratificado de riscos proporcionais de Cox

Uma extensão do modelo de riscos proporcionais para o caso em que as funções de riscos básicas, relativos a dois grupos não são proporcionais, é dado pela expressão

$$S_k(t|\mathbf{Z}) = \exp\{-H_{0k}(t)\exp(\beta'\mathbf{Z})\}, \qquad k = 1, 2 \tag{1.57}$$

1.5.7 O modelo de regressão de Shen e Fleming

O Teste de Cox para duas amostras pode não detectar diferenças entre os tratamentos quando as funções de riscos destes tratamentos não são proporcionais ou se cruzam. Shen & Fleming (1997) propuseram um modelo de regressão que permite riscos não proporcionais entre os grupos de tratamentos, assumindo um modelo de regressão de Cox dentro de cada um deles quanto aos efeitos de outras covariáveis de interesse. O modelo é dado por:

$$S_k(t|\mathbf{Z}) = \exp\{-H_{0k}(t)\exp(\beta_k'\mathbf{Z})\}, \qquad k = 1, 2 \tag{1.58}$$

onde $S_k(t|\mathbf{Z})$ é a probabilidade de que um paciente do k-ésimo grupo sobreviva ao tempo t, sendo $H_{0k}(t)$ uma função de risco básica acumulada, $\mathbf{Z}$ um vetor $p \times 1$ de covariáveis de interesse e β_k o vetor de coeficientes de $\mathbf{Z}$ para o k -ésimo grupo. Desta forma, está implícito que o efeito das covariáveis, dado pelos coeficientes de regressão, pode depender do tratamento, de modo que o modelo não é o modelo de Cox estratificado.

1.5.8 Estimação de $S_k(t)$

Um estimador natural para a função marginal de sobrevivência $S_k(t) = E\{S_k(t|\mathbf{Z})\}$ para o k -ésimo grupo é:

$$\hat{S}_k(t) = \int_z \hat{S}_k(t|\mathbf{Z}) \mathrm{d}\hat{F}_n(z),$$

onde $\hat{F}_n$ é a distribuição empírica de $\mathbf{Z}$ e $\hat{S}_k(t|\mathbf{Z})$ pode ser estimado a partir dos dados do k -ésimo grupo,

$$\hat{S}_k(t|\mathbf{Z}) = \exp\{-\hat{H}_{0k}(t)\exp(\hat{\beta}_k'\mathbf{Z})\}$$

Mais precisamente, $\hat{S}_k(t)$ é dado por:

$$\hat{S}_k(t) = \sum_{j=1}^{2}\sum_{i=1}^{n_j} \exp\{-\hat{H}_{0k}(t)\exp(\hat{\beta}_k'\mathbf{Z}_{ij})\}, \quad k = 1, 2. \tag{1.59}$$

1.5.9 Teste da hipótese de igualdade das curvas de sobrevivência para dois grupos

A estatística WKM é baseada numa estimativa da diferença entre as funções de sobrevivência $S_1(t)$ e $S_2(t)$. Shen & Fleming (1997) propuserem uma extensão desta estatística dentro do modelo (1.58), calculando a diferença entre as funções de sobrevivência $S_1(t)$ e $S_2(t)$.

Para determinar o efeito geral de tratamento, padroniza-se a diferença da sobrevivência média calculando uma média de $\hat{S}_1(t|\mathbf{Z}) - \hat{S}_2(t|\mathbf{Z})$ sobre a distribuição das covariáveis na amostra combinada. Se τ_k representa o maior tempo de observação na amostra k, e $\tau = \min(\tau_1, \tau_2) < \infty$, onde $k = 1, 2$, sabe-se que $\hat{S}_1(t) - \hat{S}_2(t)$ é instável para t próximo a τ quando o número de pessoas em risco é pequeno; utilizando-se uma função peso $\hat{w}$, que atribui um menor peso a $\hat{S}_1(t) - \hat{S}_2(t)$ nos períodos de tempos finais, pode-se controlar esta instabilidade. Assim, Shen & Fleming (1997) propõem a estatística de teste *weighted mean survival* (WMS)

$$WMS_n(t) = \sqrt{\left(\frac{n_1 n_2}{n}\right)} \int_0^t \hat{w}(u)\{\hat{S}_1(u) - \hat{S}_2(u)\}du,$$

onde a curva de sobrevivência $\hat{S}_k(t)$, para cada grupo, é um estimador da sobrevivência média

definida na expressão (1.59) e $\hat{w}$ uma função peso dependente do dados. Os referidos autores propuseram o uso da função peso (1.44).

1.5.10 Propriedades teóricas da estatística de teste WMS

Deseja-se testar a hipótese de igualdade de duas funções de sobrevivência contra a alternativa de ordenação estocástica

$$H_1 : S_1(t) \geqslant S_2(t) \qquad \text{para todo } t,\ S_1(.) \neq S_2(.),$$

a partir dos tempos de vida de n observações sujeitos à censura à direita, das quais n_k, são provenientes da amostra $k, k = 1, 2$, e assumindo $\lim_{n\to\infty}(\frac{n_k}{n}) = 1 - p_k > 0$. Assume-se também que o tempo de falha T_{ki}, as triplas aleatórias $\{T_{ki}, U_{ki}, Z_{ki}\}(i = 1, .., n_k)$ são independentes e identicamente distribuídas, sendo U_{ki} o tempo de censura e Z_{ki} o conjunto de covariáveis para o i -ésimo indivíduo do k -ésimo grupo, considerando-se T_{ki} e U_{ki} condicionalmente independentes de Z_{ki}.

Shen & Fleming (1997) mostraram que $WMS_n(t_0)$ converge, quando o tamanho da amostra tende para o infinito, para um processo Gaussiano com média zero, e determinaram a expressão de sua variância teórica $\sigma_w^2(t_0)$.

UNICAMP
BIBLIOTECA CENTRAL

Capítulo 2

Métodos estatísticos usados em ensaios clínicos de resposta imediata

2.1 Ensaios Clínicos

Ensaios clínicos são pesquisas experimentais realizadas com o objetivo de avaliar a eficiência de uma ou mais formas de tratamento em relação a uma determinada resposta. Geralmente, estes ensaios devem ser precedidos de experimentos laboratoriais feitos com animais sobre o novo tipo de tratamento que está sendo investigado, seja ele o uso de uma nova droga ou de um novo procedimento cirúrgico, e podem ser profiláticos ou terapêuticos. Os ensaios profiláticos são planejados para avaliar a eficácia de um tratamento preventivo de uma doença, que é ministrado para uma certa moléstia a um dos grupos de indivíduos e, ao final do período de estudo, compara-se a incidência da doença nos dois grupos. Por outro lado, os ensaios terapêuticos têm por objetivo comparar um novo tratamento com o tratamento padrão de uma doença (Lee, 1992).

Foi no século XVIII que os ensaios clínicos começaram a ser desenvolvidos, destacando-se, entre eles, o de Maitland realizado em 1721, a respeito da eficácia da vacinação contra a varíola e o de Lind, realizado em 1747, quando foi descoberto o tratamento para o escorbuto. No estudo de Lind, doze tripulantes de um navio, que haviam contraído a doença, foram selecionados para receber um dos seis tratamentos, sendo submetidos a uma mesma dieta básica e divididos em seis pares de indivíduos. Cada tratamento consistia em um suplemento para a dieta básica, consistindo de duas laranjas e um limão (tratamento I), um quarto de cidra (tratamento II),

Vinagre (tratamento III), elixir de ácido sulfúrico (tratamento IV), água do mar (tratamento V) e um suplemento à escolha do paciente (tratamento VI); após seis dias, apenas os pacientes submetidos aos tratamentos I e II apresentaram melhora. Outros estudos comparativos, como os ensaios para os tratamentos contra a difteria e a febre tifóide, foram desenvolvidos entre os séculos XVIII e XIX, propiciando um maior interesse pelo uso de técnicas estatísticas na área médica. O número de pacientes envolvidos nestes estudos cresceu consideravelmente, havendo registro de um estudo de febre tifóide envolvendo mais de 11000 soldados na Índia (Lee, 1992).

O primeiro ensaio clínico que utilizou um mecanismo de atribuição aleatória de pacientes aos tratamentos em estudo, no qual aqueles não foram informados sobre qual dos tratamentos iriam receber (experimento cego), foi relatado em 1931 (Amberson et al., 1931). Neste ensaio, vinte e quatro pessoas com tuberculose pulmonar foram divididas em grupos comparáveis de 12 elementos cada, e foi lan,çada uma moeda para determinar se receberiam a injeção intravenosa de Sanocrysin (tratamento experimental) ou a de água purificada (placebo), (Friedman et al., 1985).

O experimento do Conselho de Pesquisas Médicas Britânico, conduzido por Austin Bradford Hill, em 1946, para avaliar a estreptomicina como tratamento para tuberculose pulmonar, foi o primeiro ensaio clínico aleatorizado de larga escala que utilizou números aleatórios para alocação dos pacientes a um grupo experimental e ao controle. Pode-se dizer que Hill foi o maior responsável pela introdução das técnicas modernas estatísticas na área médica. No entanto, somente nas três últimas décadas, os ensaios clínicos surgiram como um método eficaz na avaliação de intervenções médicas e, foram desenvolvidas e definidas técnicas estatísticas para planejamento e monitorização de tais ensaios.

Neste trabalho, é dada ênfase aos ensaios clínicos terapêuticos envolvendo dois tratamentos. Um ensaio clínico é geralmente realizado em três fases. A **Fase I ou Ensaio Inicial** é uma fase exploratória, em que uma nova droga, um procedimento cirúrgico, ou um tratamento é experimentado em seres humanos pela primeira vez. Geralmente, esta fase é planejada para determinar a dose máxima tolerada em seres humanos, a qual será utilizada em ensaios clínicos subseqüentes com a finalidade de verificar sua eficácia (fase II) e de compará-la com outras drogas (fase III). Nesta fase, se for possível, deve-se escolher apenas pessoas sadias para participar do ensaio, evitando assim, que pacientes muito doentes sejam expostos a alguma possível toxicidade

da droga, e a possibilidade de não se conseguir identificar um possível efeito benéfico devido a seu uso em pacientes pouco doentes ou pelo uso de medicamentos concomitantes. Na **Fase II ou Intermediária,** a dosagem da nova droga, já analisada na fase I, deve ser administrada às pessoas que apresentam a doença em estudo, com o objetivo de avaliar sua eficiência e determinar o tipo de paciente que será beneficiado. Cabe ressaltar, que este procedimento é realizado em duas etapas. Primeiramente, procura-se verificar se o benefício da droga é satisfatório ao ponto de justificar a continuidade do estudo, para posteriormente, obter uma estimativa precisa da sua eficiência. Isto pode ser feito através da proporção de pacientes na amostra que foram tratados com sucesso. Esta é uma fase exploratória, que envolve somente o novo tratamento, poucos pacientes e é realizada em um período curto de tempo, obtendo-se assim, estimativas imprecisas a respeito da segurança, eficiência e toxicidade provocadas por esta nova droga.

Nos **Ensaios Fase III ou Ensaios Clínicos Comparativos,** o objetivo é comparar a eficiência do novo tratamento estudado na fase II com a do tratamento padrão e, obter estimativas mais precisas da eficácia do primeiro para a doença em estudo. Nesta fase do ensaio, os pacientes em estudo, que devem ser semelhantes quanto aos fatores que possam afetar substancialmente o efeito do tratamento, recebem por aleatorização um de dois tratamentos. O grupo experimental é constituído pelos pacientes que são submetidos ao novo tratamento e, o de controle é composto pelos indivíduos que são submetidos ao tratamento padrão. Os dois grupos são observados durante um período de tempo, após o qual é verificado se o tratamento experimental é, de fato, o melhor.

Um ensaio clínico pode ser feito segundo dois diferentes esquemas: os de tamanho de amostra fixo e os seqüenciais. Neste último, a decisão de continuar recrutando novos pacientes após um tempo t é determinada pelas análises sobre as observações já obtidas até este tempo.

2.1.1 Monitoramento de um ensaio clínico

A responsabilidade ética do pesquisador, quanto aos indivíduos participantes de um ensaio clínico, exige que os resultados sejam monitorados durante todo tempo de sua realização. Se, durante o ensaio, os dados indicarem que um dos tratamentos é prejudicial aos indivíduos, deve-se considerar a possibilidade de interromper o estudo antes da data pré-estabelecida. Por outro lado, se estes dados demonstrarem uma clara superioridade de um dos tratamentos, o ensaio

pode também ser interrompido mais cedo, pois não seria ético dar-lhe continuidade. Se diferenças em variáveis respostas primárias ou secundárias forem inexpressivas, pode não ser justificável continuar o estudo em termos de tempo, dinheiro e esforço. Finalmente, o monitoramento de variáveis respostas pode identificar a necessidade de coletar mais dados para esclarecer questões quanto ao benefício ou toxicidade que podem surgir durante o estudo.

Para se realizar o monitoramento adequadamente, os dados devem ser coletados e processados em tempo, enquanto o ensaio avança. O monitoramento deve ser feito, preferencialmente, por uma pessoa ou grupo independente do pesquisador e, nenhuma técnica deve ser usada isoladamente para decidir se o ensaio deve prosseguir.

Monitorar dados consiste em um processo ativo, no qual tabulações e análises adicionais são sugeridas e desenvolvidas após uma revisão dos resultados obtidos do ensaio em andamento. Envolve também uma interação entre os indivíduos responsáveis pela coleta, tabulação e análise dos dados, aos quais se recomenda o não envolvimento formal com os objetivos ou com os pesquisadores. A responsabilidade de monitorar dados da variável resposta deve ser confiada a um grupo independente, constituído de especialistas de várias áreas, como médicos clínicos, pessoas com experiência em condução de ensaios clínicos, epidemiologistas e estatísticos.

A freqüência de encontros do comitê de monitoramento pode variar dependendo da fase do ensaio. As fases de recrutamento de pacientes, de seguimento e conclusão de ensaio, exigem diferentes níveis de atividade, e as reuniões não devem ser tão freqüentes a ponto de que haja poucos dados novos acumulados entre as reuniões. Em muitos ensaios clínicos longos, os comitês de monitoramento encontram-se regularmente em intervalos de quatro a seis meses, com encontros adicionais, quando necessário, e nesse período, as pessoas responsáveis pela tabulação e análise dos dados assumem a responsabilidade pelo monitoramento de situações não usuais, que devem ser comunicadas ao comitê de monitoramento.

2.1.2 Término antecipado de um ensaio clínico

A finalização antecipada de um ensaio clínico pode ser difícil, não somente porque os assuntos envolvidos são complexos, mas também devido ao fato de que a decisão final requer um consenso do comitê que monitora o referido ensaio, o qual deve basear-se, não somente nos resultados estatísticos mas, principalmente, no contexto de todos os dados disponíveis. A seguir, são

apresentadas algumas questões que devem ser consideradas no processo de decisão:

1. Possíveis diferenças nos fatores prognósticos entre os dois grupos na linha de base devem ser exploradas e os ajustes necessários devem ser feitos na análise;

2. Qualquer chance de vício na avaliação das variáveis respostas deve ser levada em consideração;

3. O possível impacto de dados faltantes deve ser avaliado;

4. Efeitos colaterais em potencial e resultados de variáveis respostas secundárias devem ser considerados;

5. A consistência interna, por exemplo, entre os subgrupos e as várias medições dos resultados deve ser examinada. Em um ensaio multicêntrico, o comitê de monitoramento deve avaliar se os resultados entre os diversos centros são consistentes; por exemplo, antes de suspender um ensaio, o referido comitê deve ter certeza que o motivo da suspensão não esteja relacionado somente a um ou a dois centros.

Um dos primeiros ensaios clínicos feitos nos Estados Unidos ilustra como a decisão de finalização antecipada pode ser controversa. O Programa de Diabetes de um Grupo de Universidades consistiu em um ensaio aleatorizado, placebo-controle, duplo-cego, realizado com a finalidade de testar a eficácia de quatro tratamentos para diabetes. A medida primária da eficácia foi o grau de lesão na retina e os quatro tratamentos utilizados foram: uma dose fixa de insulina, uma dose variável de insulina, tolbutamida e phenformina. O ensaio com o grupo tratado com tolbutamida foi terminado antecipadamente porque o comitê de monitoramento percebeu que a droga poderia ser prejudicial e não parecia ter nenhum benefício. Este grupo apresentou não só um elevado coeficiente de mortalidade devido a problemas cardiovasculares quando comparado ao grupo placebo (12.7 % vs 4.9 %), como também um coeficiente de mortalidade geral superior ao placebo (14.7% vs 10.2%). A análise da distribuição dos fatores que estão associados à mortalidade cardiovascular revelou uma desigualdade entre os grupos, pois os indivíduos tratados com tolbutamida tinham maior risco que os demais, antes do início do tratamento. Este fato, juntamente com questões relativas à correta atribuição da causa da morte, gerou consideráveis críticas e, mais tarde, o experimento com phenformina foi também terminado pelo excesso de mortalidade no grupo controle (15.2% vs 9.4%). Isto levou à revisão dos dados por um grupo independente de estatísticos que, embora tenham referendando as opiniões feitas pelo comitê de

monitoramento do referido Programa, a polêmica sobre o estudo e suas conclusões continuou por muito tempo.

2.1.3 A questão da demora na obtenção da resposta aos tratamentos

A análise estatística de dados de sobrevivência lida com observações do tempo decorrido desde um instante inicial até a ocorrência de um evento de interesse, denominado genericamente falha, embora possa ser um evento favorável à vida, como o desaparecimento dos sintomas de uma moléstia. Os ensaios clínicos que envolvem resposta temporal freqüentemente são longos, mas pode acontecer que, em alguns indivíduos, o evento de interesse não ocorra até o final do estudo de modo que seu tempo exato de falha não é conhecido. Desta forma, em um ensaio clínico comparativo da sobrevivência de pacientes segundo dois tratamentos, a informação do tempo até a falha de algum dos indivíduos é incompleta, sabendo-se apenas que este tempo é superior ao observado até o final do ensaio. Estas observações são chamadas de observações "censuradas" do tempo até a falha e precisam ser levadas em conta em qualquer análise comparativa.

Os estudos seqüenciais têm sido muito usados na área de sobrevivência e podem ser descritos dois tipos básicos conforme o modo escolhido para o encerramento do estudo. O primeiro tipo é constituído pelos **Ensaios de duração fixada** quando pacientes são observados até um tempo pré-fixado T_0; no segundo, constituído pelos **Ensaios de informação fixada,** o estudo termina assim que ocorrer um determinado número acumulado de falhas. Variações destes tipos de ensaios são obtidas, conforme é planejada a entrada de pacientes e, normalmente isto é permitido no decorrer do período de tempo até o final do estudo; ou, um mesmo grupo de pacientes fica em observação desde o início do ensaio. Nos capítulos 3 e 4, são tratadas as técnicas estatísticas desenvolvidas para analisar resultados obtidos em ensaios clínicos cuja resposta de interesse maior seja o tempo de sobrevivência.

2.2 Método seqüencial clássico

O propósito do método seqüencial clássico é minimizar o número de indivíduos que devem participar de um estudo, e a decisão de continuar alocando-os depende dos resultados daqueles que já entraram no estudo. A maioria dos métodos seqüenciais assume que o resultado da variável

resposta é conhecido em um curto período de tempo, relativo à duração do ensaio e, por essa razão, estes métodos são aplicáveis em muitos ensaios envolvendo doenças agudas. Para estudos que envolvem doenças crônicas, métodos seqüenciais clássicos não são tão úteis. O método de análise seqüencial, originalmente desenvolvido por Wald (1947) e aplicado por Armitage (1975) em ensaios clínicos envolve, testes repetidos dos dados em um experimento simples. O método assume que a única decisão a ser tomada refere-se à continuação ou não do ensaio porque um dos grupos está respondendo significativamente melhor, ou pior que o outro. Esta regra de decisão seqüencial clássica é chamada por Armitage (1975) de "plano aberto", uma vez que não existe garantia de quando uma decisão de finalização é atingida; a aderência estrita ao "plano aberto" pode significar que o estudo não tem um tamanho de amostra fixo. Muito poucos ensaios clínicos usam o "plano aberto" ou modelo seqüencial clássico, porque não há certeza de alcançar um ponto no qual o ensaio seja finalizado. O método também exige que os dados sejam arranjados em pares, uma observação de cada grupo e, em muitas situações, pares de indivíduos não são desejáveis, porque eles podem ser muito diferentes e não estar "bem combinados" em relação a variáveis prognósticas importantes.

Um exemplo de um plano seqüencial "aberto" para variáveis respostas binárias é apresentado na Figura 2.1 para um teste bilateral de uma hipótese de igualdade entre um tratamento novo (Tr_2) e um usual (Tr_1), com $\alpha = 0.05$ e poder 0.95. A hipótese nula é H_0:$P_1 = P_2$, onde P_i indica a probabilidade de sucesso do tratamento Tr_i $(i = 1,2)$. A hipótese alternativa é H_1:$P_1 \neq P_2$ e a variável escolhida para testar H_0 é D_N = diferença entre o número dentre N pares em que o tratamento Tr_2 foi superior a Tr_1 e o número de pares em que Tr_2 é inferior a Tr_1.

Usando-se um sistema de eixos cartesianos, representa-se o número N de pares de observações no eixo horizontal e, no eixo vertical, a variável D_N que é o efeito puro do tratamento Tr_2 e pode ser positiva ou negativa. O teste seqüencial é realizado verificando, para cada valor de N, a localização de (N,D_N) com relação às quatro regiões de decisão a respeito de H_0 : região em que o tratamento Tr_2 é considerado superior ao tratamento Tr_1; região em o tratamento Tr_2 é considerado inferior a Tr_1; região em que se aceita a igualdade dos tratamentos Tr_2 e Tr_1 e a região em que se deve continuar alocando novos pares até ser tomada um decisão. A Figura 2.1 mostra como são estas regiões.

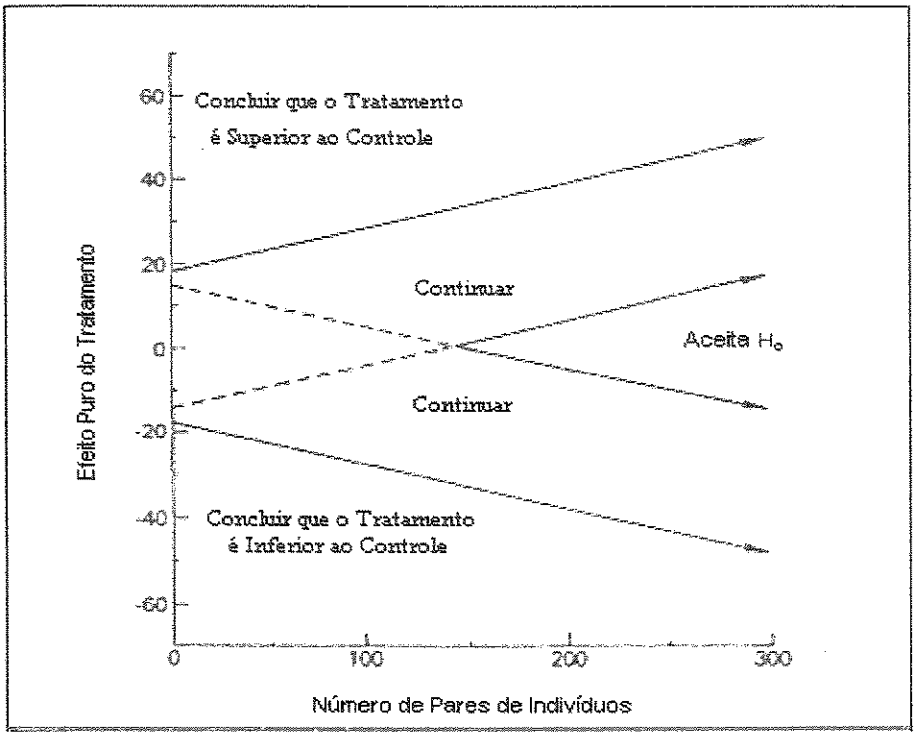

Figura 2.1: Fronteiras de um plano seqüencial aberto para dados acumulados

Fonte: Fundamentals of Clinical Trials (Friedman, Furberg, DeMets, 1985)

Armitage (1975) introduziu o modelo seqüencial "fechado" ou restrito para assegurar que um limite máximo fosse imposto ao número de indivíduos ($2N$) a ser recrutado e, assim como o "plano aberto", os dados devem ser pareados utilizando-se uma observação de cada grupo. Os limites são determinados de forma que o modelo tenha níveis específicos de α e $1-\beta$. Os limites do plano seqüencial restrito, para uma variável resposta dicotômica, são ilustrados na Figura 2.2 para a mesma hipótese bilateral usada na Figura 2.1. No eixo horizontal, representa-se o número de pares de observações e no vertical, a variável D_N de efeito puro do tratamento em N pares. A região de "aceitação" de H_0 é agora limitada pela linha vertical de abscissa igual a N. Para uma hipótese alternativa específica, Armitage (1975) dá os parâmetros das fronteiras nos casos em que a variável resposta tem distribuição binomial ou normal e o número máximo de "pares" de indivíduos necessários em termos do nível de significância e do poder do teste.

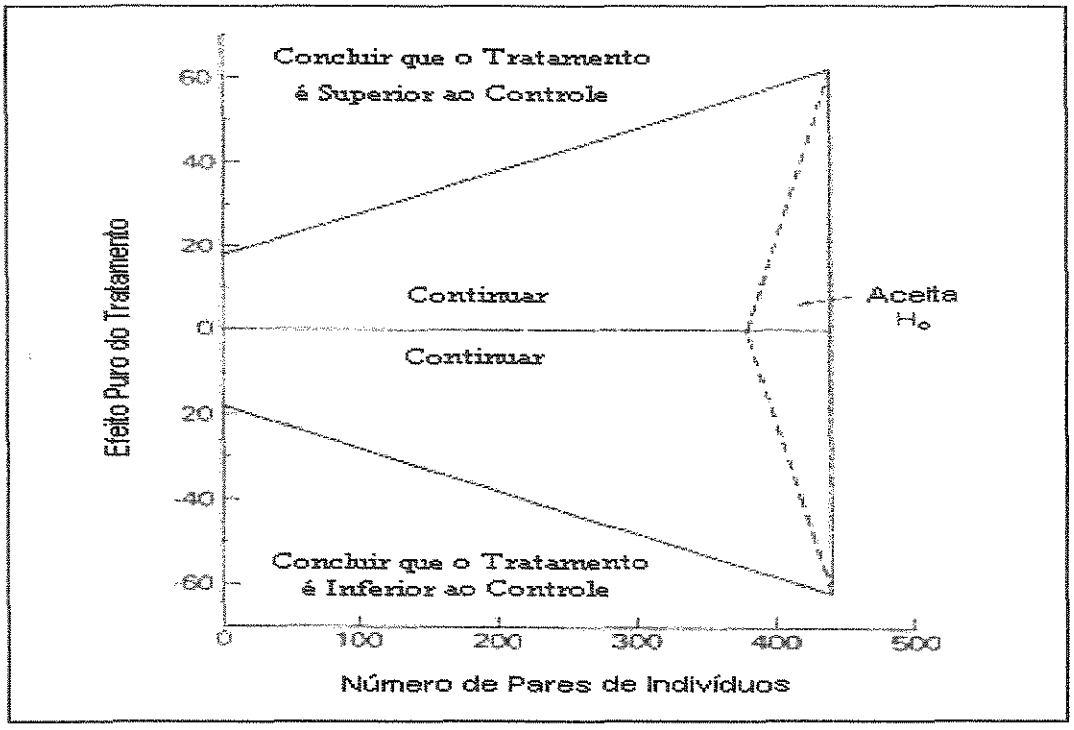

Figura 2.2: Fronteiras de um plano seqüencial fechado para dados acumulados

Fonte: Fundamentals of Clinical Trials (Friedman et al, 1985, segunda edição).

2.3 Testes de significância repetidos

A abordagem seqüencial clássica não é muito utilizada em estudos de sobrevivências ou mesmo em ensaios em que a resposta de interesse é conhecida em um curto período de tempo relativo à duração do ensaio. A maior razão de sua não aplicabilidade pode ser devido ao fato de que este método exige que os dados sejam coletados em pares de indivíduos, uma observação para cada grupo e, que seja feita a análise dos dados acumulados após cada novo par de resultados. Em muitos ensaios clínicos, isto não é necessário ou é até mesmo impraticável se os dados forem monitorados por um comitê, que tem encontros agendados espaçadamente no período de estudo. Tornou-se prática realizar vários testes de significância dos dados acumulados em diversos momentos de análise, e se não houver controle do nível de significância usado em cada análise, este não é um procedimento válido e um nível de significância de 0,05 em cada etapa, corresponderá a um nível de significância real associado ao ensaio como um todo muito maior

do que 0,05.

Nesta seção, é apresentada a metodologia desenvolvia por Armitage et al. (1969), Pocock (1977), O'Brien & Fleming (1979) e outros para controlar o nível de significância global.

Por teste de significância repetido, de uma hipótese H_0, sobre a média de uma variável aleatória X, entende-se um procedimento seqüencial com as seguintes características:

- Observações i.i.d da variável aleatória X são coletadas seqüencialmente.

- Um número total de N análises intermediárias é fixado previamente.

- No momento da primeira análise, faz-se um teste de hipótese convencional (como se o tamanho de amostra fosse fixo) da hipótese H_0 sobre a média μ_X de X a um nível de significância nominal α_1, baseado na estatística $S_1 = X_1$. Se esta hipótese H_0 for rejeitada pelo teste convencional, finaliza-se o procedimento; caso contrário, realiza-se a segunda análise, quando X_2 é observada e faz-se um teste de hipótese convencional da hipótese H_0 a um nível de significância nominal α_2, baseado na soma $S_2 = X_1 + X_2 = S_1 + X_2$. Se H_0 for rejeitada pelo teste de hipótese convencional, finaliza-se o procedimento; caso contrário, na terceira análise observa-se X_3, calcula-se $S_3 = S_2 + X_3$, e faz-se um teste de hipótese convencional da hipótese H_0 a um nível de significância nominal α_3, baseado na soma S_3.

- Procede-se, de forma semelhante, em N análises até que ocorra a rejeição de H_0 em um dos testes convencionais feitos. Se isso não ocorrer até a última análise, não rejeita-se H_0 .

Este procedimento tem um nível de significância α, denominado α global, que fica determinado pela escolha dos níveis nominais α_j ($j = 1, 2, ., N$) associados a cada uma das análises programadas. Armitage et al. (1969) demonstraram que, se testes de significância são feitos repetidamente sobre dados acumulados, sob hipótese nula, o α global será maior do que o máximo dos α_j's nominais . Por exemplo, em testes de significância repetidos sobre a média de uma variável com distribuição normal com três análises, pode-se rejeitar a hipótese nula no primeiro teste, no segundo ou no terceiro e, isto deve ser levado em conta no cálculo do α global. Se o α nominal do primeiro teste é, 0,0221, então o α global é, no mínimo, este valor, pois ainda falta incorporar a probabilidade de rejeitar a hipótese nula no segundo e terceiro testes. Da mesma forma, se forem feitos dois testes de significância a um nível α nominal de 5 %, o α global será de 8,3% e aumentará para 14,2 % se forem executados cinco testes. Assim, a probabilidade de rejeitar H_0 erroneamente cresce à medida que o número de análises aumenta. A Tabela 2.1 e a

Figura 2.3 mostram o valor de α global para diversos níveis nominais em cada análise e diversos números de análises intermediárias planejadas.

Tabela 2.1: Nível de Significância Global de um teste repetido de H_0 com N análises intermediárias, dados níveis nominais a cada etapa

Nível de Significância	Número de Análises Intermediárias (N)								
Nominal	1	2	3	4	5	10	25	50	200
0,01	0,01	0,018	0,024	0,029	0,033	0,047	0,07	0,088	0.126
0,05	0,05	0,083	0,107	0,126	0,142	0,193	0,266	0,320	0.424
0,10	0,10	0,16	0,202	0,234	0,26	0,342	0,449	0,524	0.652

Fonte: Statistical Methods For Survival Data Analysis (Lee, 1992, segunda edição)

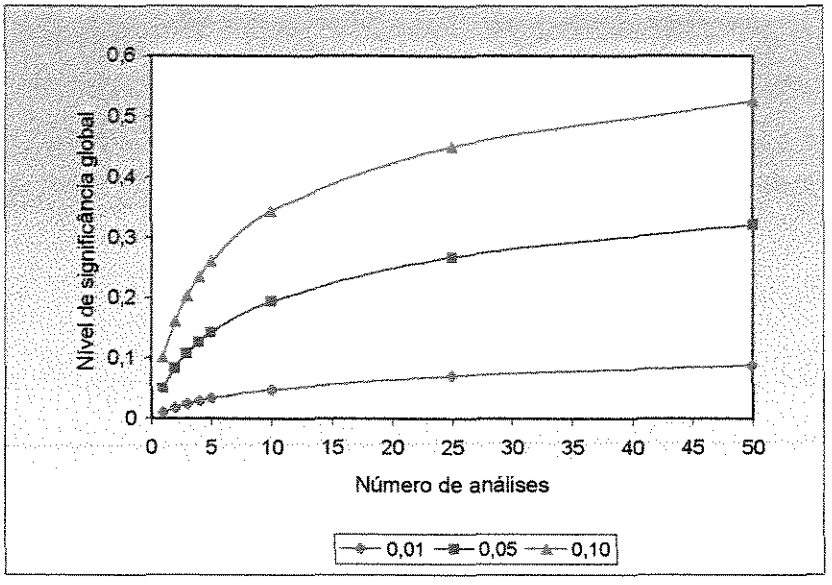

Figura 2.3: Nível de Significância Global de um procedimento de Testes de Significância Repetidos de acordo com o número de análises para os níveis nominais de 0,01; 0,05 e 0,10 a cada etapa.

2.3.1 Testes de significância repetidos sobre a média de uma variável aleatória

Caso Binomial: Em uma seqüência de N ensaios binomiais independentes, testes repetidos da hipótese $H_0 : p = p_0$ são realizados após cada um deles, usando-se a soma do número de sucessos até a etapa j, (S_j) para testar a hipótese H_0 através de um teste convencional para um tamanho de amostra fixo j. Rejeita-se H_0 na etapa j se

$$S_j \geq b_j \qquad \text{ou} \qquad S_j \leq a_j$$

onde b_i é o menor inteiro para o qual $\mathrm{P}[\mathrm{S}_j \geq \mathrm{b}_j] \leq \alpha_{/2}$ quando S_j tem distribuição binomial $(j; p_o)$ e $a_j = j - b_j$.

Se H_0 for aceito, mais um ensaio é observado e realiza-se um teste convencional baseado em S_{j+1}; o procedimento é repetido até rejeitar H_0 em uma das N etapas previstas ou, no caso de não rejeição pelo teste convencional até a N-ésima etapa, aceita-se H_0.

Caso Normal: Um ensaio consiste na observação de uma seqüência de variáveis aleatórias independentes e identicamente distribuídas $X_1, X_2, ..., X_n$ com distribuição $\mathrm{N}(\mu, 1)$. Um teste repetido em N etapas sobre o parâmetro μ é planejado da seguinte forma: após cada ensaio, utiliza-se a soma até a etapa j (S_j), que tem distribuição $\mathrm{N}(j\mu, j)$ para testar a hipótese H_0 : $\mu = \mu_0$ através de um teste convencional para um tamanho de amostra fixo j. Rejeita-se H_0 na etapa j se

$$|S_j - j\mu_0| \geq c_j,$$

onde $c_j = k\sqrt{j}$ para a constante k satisfazendo

$$\Phi(k) = 1 - \alpha_{j/2}.$$

onde $\Phi(k)$ é a função distribuição da normal com média zero e variância 1.

Se H_0 for aceito, mais um ensaio é observado e realiza-se um teste convencional baseado em S_{j+1}.

O procedimento é repetido até rejeitar H_0 em uma das N etapas previstas ou, no caso de não rejeição pelo teste convencional até a N-ésima etapa, aceita-se H_0.

Caso geral de uma variável do tipo contínuo: Um ensaio consiste na observação de uma seqüência de variáveis aleatórias independentes e identicamente distribuídas $X_1, X_2, ..., X_n$ com média μ . Indica-se por A_j a região de valores de S_j que levam à aceitação de $H_0 : \mu = \mu_0$, segundo o teste de hipótese convencional ao nível nominal de α_j, na etapa j $(j = 1, .., N)$.

Seja $\mathbf{A}_j$=$\{(s_1, s_2, ..., s_N) \in R^N;\ s_j \in \mathrm{A}_j\}$ o evento em R^N, constituído das realizações $(s_1, s_2, .., s_N) \in R^N$ que levam à aceitação de H_0, segundo o teste de hipótese convencional

ao nível nominal de α_j, na etapa j.

O procedimento seqüencial, acima definido, termina com a não rejeição de H_0 se, em todas as análises, H_0 não for rejeitada. Desta forma, a região de aceitação de H_0 é dada por:

$$\mathbf{A} = \mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N. \tag{2.1}$$

e o nível de significância global do procedimento proposto é

$$\alpha = 1\text{-}\mathrm{P}[\mathbf{A}]. \tag{2.2}$$

Como os testes intermediários são realizados em função das variáveis $S_1, S_2, ..., S_N$, para o cálculo de P[**A**] é preciso determinar sua função densidade conjunta, que pode ser escrita em termos da função densidade da variável X, f_X, como:

$$\mathrm{f}_{S_1,\cdots S_N}(s_1,s_2,...s_N)=\mathrm{f}_X(s_1)\mathrm{f}_X(s_2\text{-}s_1)..\mathrm{f}_X(s_N\text{-}s_{N-1}) \text{ para } (s_1,s_2,...s_N) \in \mathrm{R}^N \tag{2.3}$$

Logo, no caso de funções de densidade contínuas, o nível de significância global pode ser calculado através de (2.2), sabendo-se que

$$\mathrm{P}[\mathbf{A}]=\int_{A_N} .. \int_{A_1} \mathrm{f}_X(s_1)\mathrm{f}_X(s_2\text{-}s_1)...\mathrm{f}_X(s_N\text{-}s_{N-1})ds_1..ds_N \tag{2.4}$$

com uma expressão análoga em termos de somatórias para variáveis discretas.

2.3.2 Cálculo iterativo de P[A], segundo Armitage, McPherson e Rowe (1969)

A sugestão de Armitage et al. (1969) para o cálculo da integral múltipla de uma função densidade multivariada relativa às somas parciais de uma seqüência de variáveis aleatórias i.i.d. é uma opção interessante para o cálculo do α global associado a um teste de significância repetido. O caso em que esta densidade é uma normal multivariada é de maior importância, tanto no caso em que esta é a distribuição exata do conjunto de somas parciais de variáveis i.i.d., como em aplicações do teorema central do limite. Eles sugeriram que a integral acima seja calculada em N etapas sucessivas e em cada uma delas, é calculada uma integral do tipo:

$$\int_{A_j} f_j(s_j) f_X(s_{j+1}\text{-}s_j) ds_j$$

onde

$$f_1(s_j) = f_X(s_j)$$

e, para j = 2...N:

$$f_j(s_j) = \int_{A_{j-1}} f_{j-1}(s_{j-1}) f_X(s_j - s_{j-1}) ds_{j-1} \tag{2.5}$$

2.3.3 A função gasto de alfa

Lan & DeMets (1983) propuseram uma maneira de escolher os níveis nominais $\alpha(j)$ em cada análise intermediária de modo a obter um nível de significância global, α, fixado. Para entendimento de sua proposta, considere-se inicialmente o evento $\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_j$ que contém todas as realizações de testes seqüenciais repetidos que não rejeitam $\mathbf{H}_0$ até a etapa j ($j = 1, .., N$). Segundo Armitage et al.(1969), tem-se que:

$P[\mathbf{A}_1] = \int_{A_1} f_X(s_1) ds_1$

$P[\mathbf{A}_1 \cap \mathbf{A}_2] = \int_{A_2} \int_{A_1} f_X(s_1) f_X(s_2\text{-}s_1) .... ds_1 ds_2 = \int_{A_2} f_2(s_2)\, ds_2$

Da mesma forma, temos:

$P[\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_j] = \int_{A_j} f_j(s_j) ds_j$

e

$P[\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N] = \int_{A_N} f_N(s_N) ds_N$.

O complemento de $(A_1 \cap A_2 \cap .. \cap A_j)$ é o evento "**rejeitar $\mathbf{H}_0$ em uma das etapas, de 1 a j**".

Quando j cresce de 1 a N, estes eventos formam uma seqüência crescente, de modo que a seqüência de suas probabilidades $\alpha(j)$ é crescente. Assim, tem-se:

$\alpha(1) = P[(\mathbf{A}_1)^C] = 1\text{-} \int_{A_1} f_X(s_1) ds_1$

$\alpha(j) = P[(\mathbf{A}_1 \cap \mathbf{A}_2 \cap .. \cap \mathbf{A}_j)^C] = 1 - \int_{A_j} f_j(s_j) ds_j$

$\alpha(N) = P[(\mathbf{A}_1 \cap \mathbf{A}_2 \cap .. \cap \mathbf{A}_N)^C] = 1 - \int_{A_N} f_N(s_N) ds_N = \alpha$ = Nível de significância global do teste

A cada escolha de conjuntos (A_j), na determinação de um teste seqüencial, fica associada a

função crescente de j, $\alpha(j)$, denominada função de gasto de alfa (Lan & DeMets, 1983).

$$\alpha(j) = P[(\ \mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_j)^C] \tag{2.6}$$

Conforme já foi afirmado acima, a seqüência de eventos $(\mathbf{A}_1\mathbf{A}_2...\mathbf{A}_j)^C$ é não decrescente. Pode-se, então, decompor cada elemento desta seqüência em eventos disjuntos B_j, da seguinte forma:

$(\mathbf{A}_1)^C = B_1$,

$(\mathbf{A}_1\mathbf{A}_2)^C = B_1 \cup B_2$,

$(\mathbf{A}_1\mathbf{A}_2..\mathbf{A}_j)^C = B_1 \cup B_2 \cup B_3 \cup .. \cup B_j$

$(\mathbf{A}_1\mathbf{A}_2..\mathbf{A}_N)^C = B_1 \cup B_2 \cup B_3 \cup ... \cup B_N$

onde

$Bj = (\mathbf{A}_1\mathbf{A}_2..\mathbf{A}j)^C - (B_1 \cup B_2 \cup ... \cup B_{j-1})$ contém as realizações que levam à rejeição de H_0 exatamente na j-ésima etapa $(j = 1, .., N)$.

Tem-se, em conseqüência desta decomposição,

$\alpha(1) = \mathrm{P}[(\mathbf{A}_1)^C] = \mathrm{P}(\mathrm{B}_1)$

$\alpha(j) = \mathrm{P}[(\ \mathrm{A}_1 \cap \mathrm{A}_2 \cap ... \cap \mathrm{A}_j)^C] = \mathrm{P}(\mathrm{B}_1) + \mathrm{P}(\mathrm{B}_2) + ... + \mathrm{P}(\mathrm{B}_j)$

$\alpha(N) = \mathrm{P}[(\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N)^C] = \mathrm{P}(\mathrm{B}_1) + \mathrm{P}(\mathrm{B}_2) + ... + \mathrm{P}(\mathrm{B}_N) = \alpha$

Indicando por $\mathrm{P}(\mathrm{B}_j) = \pi_j$ para $j = 1, ..., N$, tem-se

$$\pi_j = \alpha(j) - \alpha(j-1) \tag{2.7}$$

π_j indica o **gasto de alfa na j-ésima etapa**. Observa-se que:

$$\alpha = \pi_1 + \pi_2 + ...\pi_N$$

Exemplo 1: Cálculo do nível global de significância α em testes de significância repetidos bicaudais da hipótese $H_0 : p = \frac{1}{2}$ sobre a média de uma variável X com distribuição binomial $(1; p)$ em dez análises, ao nível nominal de $\alpha_j = 5\%$, na j-ésima análise.

A partir da Tabela 2.2, que apresenta funções densidades binomiais com $p = \frac{1}{2}$, fazendo-se uso também de (2.4), é possível calcular o alfa global associado a uma série de testes de

significância repetidos.

Tabela 2.2: Distribuição binomial para $p = \frac{1}{2}$ e j = 1, 2, 3,...,10 $\mathrm{P}(X = s) = \binom{j}{s}\left(\frac{1}{2}\right)^j$

	1	2	3	4	5	6	7	8	9	10
0	0,5	0,25	0,125	0,0625	0,03125	0,01562	0,00781	0,0039	0,0020	0,00098
1	0,5	0,5	0,375	0,2500	0,15625	0,09375	0,05469	0,03125	0,01758	0,00976
2		0,25	0,375	0,3750	0,3125	0,23438	0,16406	0,10937	0,07031	0,04395
3			0,125	0,2500	0,3125	0,3125	0,27344	0,21875	0,16406	0,11719
4				0,0625	0,15625	0,23438	0,27344	0,27344	0,24609	0,20508
5					0,03125	0,09375	0,16406	0,21875	0,24609	0,24609
6						0,01562	0,05469	0,10937	0,16406	0,20508
7							0,00781	0,03125	0,07031	0,11719
8								0,0039	0,01758	0,04395
9									0,0020	0,00976
10										0,00098

Devido à natureza discreta da variável binomial, é impossível fazer um teste com α exatamente igual a $0,05$ em cada etapa. Desta forma, a rejeição de H_0, através de um teste de significância comum, não é possível para $j = 1, 2, 3, 4, 5$; para seis experimentos binomiais, pode-se rejeitar H_0 para $\alpha < 0,05$ se forem obtidos os valores 0 ou 6 e, com nove experimentos, se forem obtidos um dos valores 0, 1, 8 ou 9. No entanto, em testes repetidos, os valores 0 e 9 em nove experimentos nunca seriam observados, pois a hipótese já teria sido rejeitada em uma etapa anterior. Trabalhando em um nível nominal $\leq 0,05$, em dez testes repetidos bicaudais, o nível de significância global é de $0,0546875$.

Exemplo 2: Cálculo do nível global de significância α e o gasto de alfa em testes de significância repetidos bicaudais da hipótese $H_0 : \mu = 0$ sobre a média de uma variável X com distribuição normal $(\mu, 1)$ em três análises, ao nível nominal de $\alpha_j = 0,05$ na j-ésima análise.

Como os testes intermediários são realizados em termos das variáveis S_1, S_2, S_3, para o cálculo de P[**A**], é preciso integrar sua função densidade conjunta, ou seja:

$$\mathrm{P}[\mathbf{A}] = \int_{-1,96\sqrt{3}}^{1,96\sqrt{3}} \int_{-1,96\sqrt{2}}^{1,96\sqrt{2}} \int_{-1,96}^{1,96} \mathrm{f}_X(\mathrm{s}_1)\mathrm{f}_X(\mathrm{s}_2\text{-}\mathrm{s}_1)\mathrm{f}_X(\mathrm{s}_3\text{-}\mathrm{s}_2)\ \mathrm{ds}_1\mathrm{ds}_2\mathrm{ds}_3$$

De acordo com a sugestão de Armitage et al. (1969), esta integral múltipla será calculada em etapas sucessivas da seguinte forma:

$$P[A_1] = \int_{-1,96}^{1,96} f_X(s_1)ds_1 = 0,950004$$

$$f_2(s_2) = \int_{-1,96}^{1,96} f_X(s_1)f_X(s_2\text{-}s_1)\ ds_1$$

$$P[A_1 \cap A_2] = \int_{-1,96\sqrt{2}}^{1,96\sqrt{2}} f_2(s_2)ds_2 = 0,916889$$

$$f_3(s_3) = \int_{-1,96\sqrt{2}}^{1,96\sqrt{2}} f_2(s_2)f_X(s_3\text{-}s_2)ds_2$$

$$P[A] = P[A_1 \cap A_2 \cap A_3] = \int_{-1,96\sqrt{3}}^{1,96\sqrt{3}} f_3(s_3)ds_3 = 0.892752$$

Deste modo, o gasto de alfa até a primeira etapa é $\alpha(1) = 0,049996$, até a segunda etapa é $\alpha(2) = 0,083111$ e $\alpha(3) = 0,107225$.

Conforme o resultado acima, trabalhar ao nível nominal de 0,05 em três testes repetidos bicaudais corresponde, na realidade, a um nível de significância de 0,10725.

Para futuras considerações, é importante observar que a distribuição conjunta de (S_1 , S_2 , S_3), sob H_0, é uma normal multivariada de média zero e uma estrutura de covariância dada por:

$$\text{Cov}(S_i, S_j) = i \ (1 \leq i \leq j \leq 3) \tag{2.8}$$

2.4 Testes de significância repetidos em grupos

Os métodos seqüenciais são usados em ensaios nos quais os dados são examinados freqüentemente mas, em ensaios multicêntricos, as informações são inspecionadas em poucas ocasiões, por exemplo, uma vez por ano, durante cinco anos, de forma que entre duas ocasiões sucessivas, há um grupo novo de respostas ainda não examinadas. Pocock (1977) indicou as adaptações que devem ser feitas para este caso.

Denomina-se teste de significância repetido em grupos da hipótese H_0:$\mu_Y = \mu_0$, um teste em que a variável aleatória X a ser observada em cada etapa é uma soma de n outras variáveis aleatórias Y i.i.d:

$$X_1 = Y_1 + Y_2 + ... + Y_n$$

$$X_2 = Y_{n+1} + Y_{n+2} + ... + Y_{2n}$$

$$X_j = Y_{(j-1)n+1} + ... + Y_{jn} \ (j = 1, ..., N)$$

É fácil observar que as somas parciais de X e de Y estão relacionadas por:

$$S_{X_j} = S_{Y_{nj}}$$

Se as variáveis Y tiverem uma distribuição normal $(\mu, 1)$, as variáveis X são i.i.d. $\mathrm{N}(\mu n, n)$ e a distribuição conjunta de $(S_{X_1}, S_{X_2}, ..., S_{X_N})$ é uma $\mathbf{N}_N(\mu, \Sigma)$; onde

$$\Sigma = \{\sigma_{ij}\} \text{ com } \sigma_{ij} = \mathrm{Cov}(S_{Xi}, S_{Xj}) = ni; \quad 1 \leq i \leq j \leq N \tag{2.9}$$

No exemplo 2, através da integração da função densidade conjunta de (S_1, S_2, S_3), em R^3 dada por:

$$(-1,96\sqrt{j} \leq S_j \leq 1,96\sqrt{j}) \qquad j = 1,2,3$$

conclui-se que $\alpha = 0,10725$.

Para três testes de significância repetidos de H_0:$\mu_Y = 0$ em grupos de $n = 20$ observações novas, em cada intervalo entre análises, as regiões de não rejeição de H_0 serão $(-1,96\sqrt{nj}; 1,96\sqrt{nj})$ e o cálculo da função gasto de alfa até a etapa j pode ser feito utilizando-se (2.4). O resultado é que a função gasto de alfa é idêntica a já obtida para $n = 1$.

A razão disto está no fato de que, dividindo-se todos os termos da desigualdade abaixo por $\sqrt{n}$,

$$\mathrm{P}[(-1,96\sqrt{nj} \leq S_{Xj} \leq 1,96\sqrt{nj})] \ (j = 1,2,3)$$

e obtém-se

$$\mathrm{P}[((-1,96\sqrt{j}) \leq (S_{Xj})/\sqrt{n} \leq 1,96\sqrt{j})] \ (j = 1,2,3)$$

e, como $[(S_{X_1})/\sqrt{n}, (S_{X_2})/\sqrt{n}, (S_{X_3})/\sqrt{n})] \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$; onde

$$\mathbf{\Sigma}=\{\sigma_{ij}\}\text{com } \sigma_{ij}=(ni/n) = i; (1 \leq i \leq j \leq 3), \tag{2.10}$$

conclui-se que, sob H_0, tem a mesma distribuição de (S_1, S_2, S_3) e com covariância igual a dada em (2.8).

Este resultado pode ser generalizado, podendo ser enunciado da seguinte maneira: supondo que para $j = 1, ...,N$, as somas parciais S_{Xj} de uma seqüência de variáveis aleatórias indepen-

dentes e identicamente distribuídas como X com distribuição normal de variância conhecida, têm variância V(j), e Cov(S_{Xi}, S_{Xj})=V(i) ($i \leq j$) e definindo-se o tempo de informação à época da $j-$ésima análise como $t_j = \frac{V(j)}{V(N)}$, tem-se que para (j=1, .., N): $\left[\frac{S_{X_1}}{\sqrt{V(N)}}, \frac{S_{X_2}}{\sqrt{V(N)}}, .., \frac{S_{X_N}}{\sqrt{V(N)}}\right] \sim N(\mathbf{0}, \mathbf{\Sigma})$; onde

$$\Sigma = \{\sigma_{ij}\} \text{ com } \sigma_{ij} = \frac{V(i)}{V(N)} = t_i \ (1 < i \leq j \leq N). \tag{2.11}$$

Se t_j for igual a j, sua distribuição é a mesma de $(S_1, S_2, ..., S_N)$, quando a variável X tem distribuição N(0, 1). Mesmo no caso de t_j não ser igual a j, a dedução acima levará ainda a um tratamento unificado dos testes de significância repetidos, por meio da constatação de que o processo $(S_{Xj})/\sqrt{V(N)})_j$ tem propriedades características de um movimento Browniano, o que levou à elaboração de programas para planejamento de testes de significância repetidos.

2.4.1 O movimento Browniano e as distribuições conjuntas de somas parciais de variáveis aleatórias i.i.d.

Definição 1: *Um movimento Browniano (M.B.), com tendência μ e parâmetro de variância σ^2, é uma família de variáveis aleatórias W_t, $t \in R$, satisfazendo:*

(i) $W_0 = 0$;

(ii) $W_t - W_s$ tem distribuição normal de média $\mu(t-s)$ e variância $\sigma^2(t-s)$ para $0 \leq s < t < \infty$

(iii) para quaisquer $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq .. \leq s_j < t_j < \infty$, $n = 2,3.....$;

as variáveis "Acréscimos de W_t em (s_j, t_j)" $= (W_{t_j} - W_{s_j})$ j =1,2,...,n são independentes;

(iv) $W(t)$ $0 \leq t < \infty$ é uma função contínua de t;

As condições (ii) e (iii) podem ser substituídas por

(v) A distribuição conjunta de $(W_{t_1}, W_{t_2}, ..., W_{t_n})$ é uma normal multivariada com médias dadas por

$$E(W_{t_j}) = \mu t_j \quad j = 1, ..., n.$$

e estrutura de covariância dada por:

$$Cov(W_{t_i}, W_{t_j}) = \sigma^2 min(t_i, t_j).$$

para todo n e quaisquer $t_1, t_2, ..., t_n$.

Definição 2: *A partir de um M.B.* W_t, um novo movimento Browniano B_t *definido para* t *em* $[0, 1]$ *e para* $C > 0$, é obtido pela expressão abaixo:

$$B_t = \frac{W_{Ct}}{\sqrt{C}} \tag{2.12}$$

A "tendência" do movimento Browniano B_t, acima definido, é $\mu\sqrt{C}$ e o parâmetro de variância deste processo é 1.

Uma transformação para um movimento Browniano com tendência nula e variância unitária, no intervalo [0, N], é dada por:

$$\frac{W_t - n\frac{W_t}{\sigma}\mu t}{\sigma} = \frac{W_t}{\sigma} - \frac{n\mu t}{\sigma}$$

Uma transformação para um movimento Browniano com tendência nula, e intervalo de tempo [0,1], é feita através da fórmula abaixo, para $t = \frac{1}{N}, \frac{2}{N}, ..., 1$

$$B_t = \frac{W_{Nt} - Nn\mu t}{\sqrt{N}} \tag{2.13}$$

Uma transformação para o movimento Browniano padronizado em [0,1], é dada por:

$$B_t^* = \frac{B_t}{\sigma\sqrt{n}} = \frac{W_{Nt} - Nn\mu t}{\sigma\sqrt{nN}} = \frac{W_{Nt}}{\sqrt{n}\sigma\sqrt{N}} - \sqrt{N}\Delta t \tag{2.14}$$

onde $\Delta = \sqrt{n}\frac{\mu.}{\sigma}$

Logo $\sqrt{N}\Delta$ é a "tendência" do Movimento Browniano $\frac{W_{Nt}}{\sqrt{n}\sigma\sqrt{N}}$.

Numa extensão das subseções acima, se X e Y são variáveis aleatórias independentes, com distribuição normal de médias μ e $\mu + \delta$, respectivamente, e mesma variância σ_X^2, a diferença $D = X - Y$ tem distribuição $N(\delta, 2\sigma^2)$. Se as somas parciais desta variável, em grupos de n observações, forem tomadas, a distribuição conjunta destas somas é normal multivariada, com vetor de médias com j-ésima componente igual a $jn\delta$ e matriz de covariância dada por

$$Cov(S_{Di}^n, S_{Dj}^n) = n2\sigma_X^2 min(i, j)$$

$SP_j = \frac{S^n_{Di} - ni\delta}{\sigma\sqrt{2n}}$ tem distribuições conjuntas como a de um M.B. padronizado para $j = 1, .., N$ e $B^*_t = \frac{SP_{Nt}}{\sqrt{N}} = \frac{S^n_{DNt} - nNt\delta}{\sigma\sqrt{2nN}}$ como as de um M.B. padronizado em [0,1].

Portanto, no caso de diferenças de variáveis aleatórias, a tendência é

$$\frac{nN\delta}{\sigma\sqrt{2nN}} = \sqrt{\frac{Nn}{2}}\frac{\delta.}{\sigma} = \sqrt{N}\Delta$$

onde $\Delta = \sqrt{\frac{n}{2}}\frac{\delta.}{\sigma}$

Como é possível verificar posteriormente, Lan & DeMets (1983) fizeram um programa que possibilita planejar testes repetidos em grupos de maneira unificada, usando as propriedades do movimento browniano. O programa solicita que sejam dados 3 dos parâmetros: N, "tendência", α e $(1 - \beta)$, fornecendo como resposta o quarto parâmetro.

2.5 Construção de regiões críticas em testes de significância repetidos

2.5.1 Correção de Bonferroni para múltiplos testes em dados correlacionados

As múltiplas vezes que um teste é realizado precisam ser consideradas, no nível de significância dos testes estatísticos, para não ultrapassar um nível de significância global α $(0 < \alpha < 1)$ na realização de N testes repetidos conforme apresentado por Armitage et al. (1969). Uma solução bastante geral para não ultrapassar este nível de significância é usar a correção de Bonferroni, porém esta leva a um procedimento de teste altamente conservativo porque não considera o tipo especial de correlação desses testes no contexto de ensaios clínicos em que os dados são acumulados. Esta correção consiste em realizar, a cada análise, um teste com nível de significância $\frac{\alpha}{N}$ mas, pelo fato desta correção ser válida para a determinação de um limite inferior para a probabilidade da intersecção de N eventos **quaisquer**, definidos em um mesmo **espaço de probabilidade arbitrário**, é de se esperar que o nível de significância obtido seja inferior ao α desejado.

Usando a notação anteriormente apresentada, tem-se

$$\text{Nível de significância do teste} = \mathrm{P}(\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N)^C$$

Por Bonferroni,

$$\mathrm{P}(\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N) \geqq \sum_{j=1}^{N} \mathrm{P}(A_j) - (N-1)$$

Logo, tomando-se em cada análise o mesmo nível de significância nominal $\alpha_j = \frac{\alpha}{N}$, para $j = 1,.,N$ tem-se

$$\mathrm{P}(\mathbf{A}_1 = P(\mathbf{A}_2) = ... = \mathrm{P}(\mathbf{A}_N) = 1 - \frac{\alpha}{N}$$

$$\mathrm{P}(\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N) \geqq \sum_{j=1}^{N} (1 - \frac{\alpha}{N}) - (N-1) = 1 - \alpha$$

Dessa forma, o nível de significância do teste repetido satisfaz a desigualdade:

$$\mathrm{P}(\mathbf{A}_1 \cap \mathbf{A}_2 \cap ... \cap \mathbf{A}_N)^C \leq 1 - [1-\alpha] = \alpha$$

Exemplo 3: Cálculo do nível global de significância α e o gasto de alfa em testes de significância repetidos bicaudais da hipótese $H_0 : \mu = 0$ sobre a média de uma variável com distribuição $\mathrm{N}(\mu, 1)$ em três análises, usando em cada uma delas um nível nominal α^* igual a $\frac{\alpha}{3}$, conforme a Correção de Bonferroni para $\alpha = 0,05$.

Considerando-se um experimento que consiste em uma seqüência de variáveis aleatórias $X_1, X_2, ..., X_n$, independentes e normalmente distribuídas, um teste convencional de $H_0 : \mu = 0$ na etapa $j (j = 1,2,3)$ é baseado na soma parcial S_j, que tem distribuição $\mathrm{N}(\mu_j, j)$. A região de aceitação de H_0 na etapa j é

$$\mathrm{A}_j = (\text{-}2{,}394\sqrt{j}\ ;\ 2{,}394\sqrt{j})$$

Como os testes intermediários são realizados em termos das variáveis S_1, S_2, S_3, o cálculo de $\mathrm{P}[\mathbf{A}]$ consiste na integração da função densidade conjunta destas variáveis, na região apropriada, da seguinte forma:

$$\begin{aligned} \mathrm{P}[\mathbf{A}] &= \mathrm{P}[\mathbf{A}_1 \cap \mathbf{A}_2 \cap \mathbf{A}_3] = \\ &= \int_{-2{,}394\sqrt{3}}^{2{,}394\sqrt{3}} \int_{-2{,}394\sqrt{2}}^{2{,}394\sqrt{2}} \int_{-2{,}394}^{2{,}394} \mathrm{f}_X(\mathrm{s}_1)\mathrm{f}_X(\mathrm{s}_2\text{-}\mathrm{s}_1)\mathrm{f}_X(\mathrm{s}_3\text{-}\mathrm{s}_2)\ \mathrm{ds}_1\mathrm{ds}_2\mathrm{ds}_3 \end{aligned}$$

De acordo com a sugestão de Armitage et al. (1969), esta integral múltipla foi calculada em

etapas sucessivas, obtendo-se o verdadeiro nível de significância deste procedimento: $\alpha = 0,0384$.

2.5.2 Regiões críticas segundo Pocock

Valendo-se da sugestão de Armitage et al. (1969), para o caso de variáveis com distribuição N(μ, σ^2), σ^2 conhecido, e número N de análises fixado previamente, Pocock (1977) construiu tabelas para a determinação dos limites c_j das regiões A_j=$(-c_j, c_j)$, de mesmo nível de significância nominal α_j, do teste da hipótese $H_0 : \mu_X = 0$ a ser feito em cada análise, para obter um nível de significância global α sobre resultados acumulados de observações i.i.d da variável X em indivíduos ($n = 1$) ou em grupos de n indivíduos. Devido à exigência de níveis nominais intermediários iguais entre si, as tabelas de Pocock fornecem as constantes $c(N, \alpha)$ e

$$c_j = c(N, \alpha)\sqrt{j},$$

Um gráfico de c_j de $(-\ c_j)$, em termos de j, delimita uma fronteira parabólica da região de aceitação de H_0, para os valores de S_j.

As regiões críticas determinadas por Pocock para um teste bilateral da hipótese H_0, a partir da estatística padronizada z_j, na j-ésima análise são da forma:

$$|z_j| \geq c_j.$$

Os valores de c_j são calculados segundo a distribuição teórica conjunta do vetor de estatísticas z_j nas diversas análises sob H_0.

2.5.3 Regiões críticas de O'Brien e Fleming

Com uma visão bem definida de um teste multivariado, e tratando da construção de um teste de significância da hipótese

$$H_0\text{: } P_1 - P_2 = 0$$

sobre a diferença de proporções em duas amostras independentes de uma variável com distribuição binomial, O'Brien & Fleming (1979) determinaram constantes $C(N, j, \alpha)$ que satisfizessem as seguintes condições:

- o nível de significância do teste real seria α;
- um total de no máximo N análises seriam feitas;
- a hipótese H_0 seria rejeitada se o maior valor de uma estatística Qui-quadrado de Pearson obtida, em cada etapa j, de uma tabela de contingência construída com os dados observados até essa etapa, indicada por $\chi^2(j)$ superasse um valor $C(N,\alpha)$.

Os autores consideraram, para testar H_0, a estatística:

$$\mathrm{U}= \max\ (\chi^2(j))_{\leq N}$$

O evento

$$(\mathrm{U} \geqslant \mathrm{C}\ (N,\alpha))$$

pode ser decomposto nos N subconjuntos disjuntos abaixo, definidos conforme a etapa em que, pela primeira vez, foi satisfeito:

$$\chi^2(1) \geq \mathrm{U}$$

e, para $j = 2, .., N$

$$(\mathrm{U} \geqslant \mathrm{C}(N,\alpha)) \cap ((\chi^2(i \leq j-1)) < \mathrm{U}) \cap ((\chi^2(j) \geqslant \mathrm{U})$$

É possível demonstrar que

$$\mathrm{P}[(\mathrm{U} \geqslant \mathrm{C}(\mathrm{N},\alpha)) \cap (\chi^2(1) \geqslant \mathrm{U})] = \mathrm{P}[(\chi^2)_1 \geqslant \mathrm{C}(\mathrm{N},\alpha))\mathrm{N}]$$

e para $j = 2, .., N$

$$\mathrm{P}[(\mathrm{U} \geqslant \mathrm{C}(\mathrm{N},\alpha)) \cap (\chi^2(i \leq j-1) < \mathrm{U}) \cap (\chi^2(j) \geqslant \mathrm{U})] = \mathrm{P}[\chi^2(j) \geqslant \mathrm{C}(\mathrm{N},\alpha))\ \frac{N}{j}\ ]$$

Desta forma, o teste proposto por O'Brien & Fleming rejeita H_0 na etapa $j(j = 1, ., N)$ se j for o primeiro inteiro tal que

$$\chi^2(j) \geqslant \mathrm{C(N,}\alpha)\frac{N}{j}$$

2.5.4 Construção de regiões de não rejeição de H_0 através do conceito da função gasto de alfa

a) Procedimento de Slud e Wei

Slud & Wei (1982) propuseram a seguinte forma de construção de regiões de aceitação da hipótese $H_0 : \mu = 0$ a um nível de significância 2α, através de testes em grupos repetidos, bicaudais, em N etapas no máximo:

- Especificar o gasto de alfa em cada etapa j (π_j), satisfazendo $\alpha = \sum_j \pi_j$.

-Determinar as regiões $\mathrm{A}_j = (\text{-}\mathrm{c}_j\ ,\mathrm{c}_j)$, usando a sugestão de Armitage et al. (1969), através da determinação recursiva das constantes c_j $(j = 1, ., N)$ que satisfaçam

$$\mathrm{P[S_1 \notin A_1]} = 2\mathrm{P[S_1 \geqslant c_1]} = \pi_1$$

$$\mathrm{P[S_1 \in A_1;\ S_2 \notin A_2]} = 2\mathrm{P[\ |\ S_1\ | < c_1;\ S_2 \geqslant c_2]} = \pi_2 \tag{2.15}$$

$$\mathrm{P[S_1 \in A_1; ....S_{j-1} \in A_{j-1};\ S_j \notin A_j]} == 2\mathrm{P[|S_1| < c_1 .. \ |S_{j-1}| < c_{j-1};\ S_j \geqslant c_j]} = \pi_j$$

Esta idéia abandona a exigência de mesmo nível de significância nominal α_j em cada análise.

b) Procedimento de Lan e DeMets: Determinação de fronteiras de regiões de aceitação em testes de significância repetidos

Partindo da proposta de Slud & Wei (1982), de especificar uma função gasto de alfa no momento do planejamento de um teste de significância repetido, Lan & DeMets (1983) propuseram a adaptação de tal procedimento para realizar testes que não seguiam algumas das seguintes exigências até então colocadas para a realização de um teste deste tipo:

- De que um número total de N análises seja fixado previamente.
- Em cada etapa j $(j = 1, ..., N)$, um teste de H_0 é feito com base na soma dos valores das n_j observações de variáveis aleatórias i.i.d. coletadas até a etapa j.

Resumidamente, a contribuição desses autores consiste na:

- Mudança na escala de numeração das etapas, de 1 a N , para uma escala relativa aos tempos de informação, de 0 até $J(N)$;
- Definição do conceito de tempo de informação, variando de 0 até 1;
- Modificação da definição de função gasto de alfa.

Lan & DeMets (1983) generalizam o procedimento de Slud e Wei, definindo uma função gasto de alfa como qualquer função $\alpha^*(t)$ crescente definida em [0,1], satisfazendo $\alpha^*(0) = 0$ $\alpha^*(1) = \alpha$.

Para analisar de maneira unificada qualquer teste de significância repetido em N etapas, no máximo, os autores propõem que se considere uma mudança de escala de "número de ordem j da análise", em $[0, N]$ para a escala do "tempo de informação", aqui definido como $t(j) = \frac{j}{N}$ em [0,1]. Após esta mudança na escala de tempo, pode-se redefinir a função gasto de alfa, do exemplo 2 apresentado no capítulo 2 seção 2.3.3, como:

$$\alpha^*(t) = 0 \qquad 0 \leq t < 1/3$$

$$\alpha^*(t) = 0{,}05 \qquad 1/3 \leq t < 2/3$$

$$\alpha^*(t) = 0{,}0831 \qquad 2/3 \leq t < 1$$

$$\alpha^*(1) = 0{,}1072$$

Exemplos de funções $\alpha^*(t)$ para $t \in [0, 1]$,

1- $\alpha^*(t) = \alpha t$

2- $\alpha^*(t) = \alpha t^p \; (p > 0)$

3- $\alpha^*(t) = \alpha \; log\{1 + (e - 1)t\}$

4- $\alpha^*(t) = 2 - 2\Phi(z_{\frac{\alpha}{2}}/\sqrt{t})$

Escolhida uma função $\alpha^*(t)$ e o número máximo de N análises a serem feitas, o gasto de alfa no primeiro tempo de informação ($t_1 = 1/N$) deve ser $\alpha^*(t_1)$. Calcula-se o valor de c_1 satisfazendo

$$P[\mid B_{t_1} \mid \; > c_1] = \alpha(t_1)$$

e fazendo uso de que B_t tem distribuição N(0,t).

No segundo momento de análise, com tempo de informação ($t_2 = 2/N$), o gasto de alfa deve ser ($\alpha^*(t_2)$ - $\alpha^*(t_1)$). O problema, aqui, reside em determinar c_2 que satisfaça

$$P[\mid B_{t_1} \mid < c_1 ; \mid B_{t_2} \mid > c_2] = \alpha^*(t_2) - \alpha^*(t_1),$$

que pode ser calculado integrando na região apropriada a função densidade conjunta de

$$(B_{t_1}, B_{t_2}) = (B_{t_1}, B_{t_1} + (B_{t_2} - B_{t_1}))$$

A igualdade acima permite utilizar para a integração, a sugestão de Armitage et al. (1969), para o cálculo de probabilidades relativas a eventos definidos em termos de somas de variáveis aleatórias independentes, tendo em vista a propriedade (iii) de um Movimento Browniano. O primeiro componente B_{t_1} tem distribuição N$(0, t_1)$ e é independente do segundo, $(B_{t_2} - B_{t_1})$, que tem distribuição N$(0, t_2 - t_1)$, de forma que pode-se usar a idéia proposta por Armitage et al. (1969) para calcular probabilidades para dados (c_1 e c_2). Se apenas c_1 for conhecido, pode-se fazer uma busca do valor apropriado de c_2 que satisfaça a condição especificada para o gasto de alfa na segunda etapa.

É possível estender as considerações acima para fazer o monitoramento de um ensaio clínico em tempos de informação intermediários quaisquer

$$0 < t_1 < \ t_2 < ... \ < t_N = 1$$

Neste caso, especificada uma função de gasto de alfa $\alpha^*(t)$, utiliza-se o procedimento de Armitage et al. (1969) para a determinação de constantes c_j $(j = 1, ..., N)$ que satisfaçam

$$P[\mid B_{t_1} \mid < c_1 ; \mid B_{t_2} \mid < c_2 ; ... ; \mid B_{t_j} \mid < c_{j-1} \mid B_{ji} \mid > c_j] = \alpha^*(t_j) - \alpha^*(t_{j-1})$$

já que, a distribuição conjunta de

$$(B_{t_1}, .., B_{t_{j-1}}, B_{t_j}) = (B_{t_1}, B_{t_1} + (B_{t_2} - B_{t_1}), ..., B_{t_1} + (B_{t_2} - B_{t_1}) + .. + (B_{t_j} - B_{t_{j-1}}))$$

tem função densidade conjunta de somas de variáveis aleatórias independentes: B_{t_1} tem distribuição $N(0, t_1)$, $(B_{t_2} - B_{t_1})$ tem distribuição $N(0, t_2 - t_1)$,.., $(B_{tj} - B_{t_{j-1}})$ tem distribuição $N(0, t_j - t_{j-1})$.

2.6 Tabelas de Pocock e programa de Reboussin et al. para o planejamento do tamanho de amostra para um teste seqüencial em grupos sobre a diferença de duas médias

Pocock (1977) determinou regiões críticas para testes seqüenciais em grupo da hipótese $H_0 : \mu_X - \mu_Y = 0$ contra H_1: $\mu_X - \mu_Y = \delta$ de igualdade de médias de variáveis com distribuição normal a partir da estatística padronizada z_j da diferença de médias amostrais na j-ésima análise. Uma região de aceitação de H_0 pode ser construída, com um nível global de significância α, usando o conceito da função gasto de alfa. A cada etapa, um teste da hipótese H_0 será realizado com todas as observações acumuladas até então, usando a região de aceitação de H_0 determinada pela função gasto de alfa escolhida. Se H_0 for rejeitado, interrompe-se o processo; caso contrário, passa-se à etapa seguinte. Se, após as N etapas previstas, H_0 não foi rejeitado, aceita-se H_0.

O poder deste teste, $(1 - \beta)$, pode ser obtido através da integração da função densidade conjunta das estatísticas $(z_1, z_2, .., z_N)$ na região de aceitação de H_0. Pocock (1977) elaborou tabelas que, para valores dados de N, δ e para um tamanho α e um poder do teste desejado $(1 - \beta)$, fornecem o valor de $\Delta = \sqrt{\frac{n}{2\sigma^2}}\delta$. A partir de Δ obtém-se o tamanho n de cada grupo novo de pacientes para cada tratamento, $n = 2(\frac{\Delta\sigma}{\delta})^2$ que atenderá as especificações desejadas de tamanho e poder do teste para N análises. Estas tabelas podem ser utilizadas para encontrar o poder do teste para valores especificados de N, n, α, σ e, ainda, para calcular n de modo a obter poder $1 - \beta$, para um valor de α fixado, e valores especificados de N, α, σ.

Reboussin et al. (1995) escreveram um programa em que é solicitado ao usuário o número de análises (N), o nível de significância (α), o poder do teste $(1 - \beta)$ e o tipo de fronteira. O programa calcula os limites das fronteiras que delimitam a região de aceitação de H_0 e o parâmetro Δ a partir do qual obtém-se o número necessário de indivíduos sob cada tratamento, em cada análise, e o total de pacientes envolvidos no ensaio que é $2nN$.

Capítulo 3

Métodos estatísticos usados em ensaios clínicos de resposta não imediata

O uso das estatísticas log-rank e *Weighted Kaplan-Meier* (WKM) no monitoramento de ensaios clínicos, em análise de sobrevivência através de testes de significância repetidos, é abordado neste capítulo. No capítulo anterior, a hipótese nula de igualdade de duas distribuições da variável resposta reduziu-se ao caso das igualdades de duas médias, sob a hipótese de normalidade (exata ou aproximada) e as estatísticas de teste nas diversas análises tinham uma distribuição normal multivariada . Em contraste, nos ensaios na área de Sobrevivência, trabalha-se no contexto não paramétrico, existe a possibilidade de existência de valores censurados nas diversas análises, e a distribuição conjunta das estatísticas de teste tem de ser estudada caso a caso.

Diversos tipos de censura da variável resposta podem coexistir em observações colhidas em um ensaio clínico com análises intermediárias em análise de sobrevivência. Uma censura simples do tipo I pode ocorrer em ensaios clínicos com entrada simultânea de pacientes em algumas observações, em alguns instantes de análise ou em todos, devido à fixação de datas para as análises. Censuras aleatórias do tipo I poderão ocorrer para algumas observações em certas datas de análise, se a entrada de pacientes não for simultânea, mas se der, também, de forma aleatória. Finalmente, se for admitida a existência de um tempo potencial aleatório de censura C,

que impeça a observação do tempo potencial de vida até a falha, T, quando C for menor que T, pode ocorrer também censura aleatória em algumas observações. Em resumo, as observações do tempo potencial de vida até a falha, T, podem estar sujeitas simultaneamente a dois mecanismos de censura. A variável final de censura é, pois, o valor mínimo entre duas variáveis de censura, uma devida à fixação de tempos de análise e a outra, devida a um mecanismo aleatório de censura, externo ao delineamento do ensaio.

Neste capítulo, estabelece-se, inicialmente, a notação a ser usada. A seguir, é feito um resumo dos resultados teóricos acerca da distribuição conjunta das estatísticas em N momentos de análise, primeiro para a estatística log-rank e depois para a *Weighted Kaplan-Meier* (WKM). Ao final, são apresentados exemplos com dados fictícios de alguns ensaios com o cálculo das estatísticas log-rank e *Weighted Kaplan-Meier* (WKM).

3.1 Notação

Para o tratamento i, $i = 1, 2$, determina-se que n_i pacientes entrem no estudo nos tempos aleatórios $Y_{i1},...,Y_{in_i}$. Associada à entrada do indivíduo em Y_{ij}, está o par (T_{ij}, C_{ij}), onde T_{ij} e C_{ij} são, respectivamente, o tempo potencial de morte e o de censura, ambos medidos desde a entrada no estudo. Determina-se também, que T_{ij} é independente de C_{ij} e de Y_{ij}, que as triplas (Y_{ij}, T_{ij}, C_{ij}), $i = 1, 2$, $j = 1, .., n_i$, são independentes e identicamente distribuídas e que $\lim_{n\to\infty} \frac{n_i}{n} = p_i > 0$ para $i = 1, 2$, e $n = n_1 + n_2$.

Seja $X_{ij}(t) = \min(T_{ij}, C_{ij}, (t - Y_{ij})^+)$ o tempo de vida observado para o indivíduo j do grupo i até o momento de análise t. A função de sobrevivência calculada em um instante t da variável aleatória T_i é indicada por $S_i(t)$, e a de C_i por $C_i(t)$; a função de sobrevivência do tempo de entrada Y_i em um instante t, é indicada por $E_i(t)$. Desta forma, em vista das suposições feitas sobre independência destas variáveis, a função de sobrevivência da variável aleatória "tempo observado de vida até a análise no tempo t", que é $X_{ij}(t) = \min(C_{ij}, (t - Y_{ij})^+, T_i)$ é dada por $G_i(s, t) = P(\min(C_{ij}, (t - Y_{ij})^+ \geq s)) = C_i^-(s)E_i(t - s)$, para $j = 1, ..., n_i$.

Se $I[A]$ é a função indicadora de um conjunto A , seja $\Delta_{ij}(t) = I\{T_{ij} \leq \min(C_{ij}, (t - Y_{ij})^+)\}$ a variável indicadora do fato do indivíduo j $(j = 1, .., n_i)$ submetido ao tratamento $i, (i = 1$ ou $2)$ ter morrido antes do momento de análise dos dados no tempo t .

Se $R_{ij}(x,t) = I\{X_{ij}(t) \geq x\}$ para $i = 1,2$ e $j = 1,..,n_i$, $R_i(x,t) = \sum_j R_{ij}(x,t)$ é o número de pacientes submetidos ao tratamento i, com tempo de vida observado maior ou igual a x $(x < t)$ que estão ainda sob o risco de morrer, quando se faz análise dos dados no tempo t de análise. Seja $R(x,t) = R_1(x,t) + R_2(x,t)$ o número total de indivíduos em risco de morrer no tempo de análise t.

3.2 Resultados assintóticos sobre as distribuições conjuntas de estatísticas não-paramétricas calculadas em N análises

3.2.1 Comportamento ao longo do tempo do processo estocástico definido pelas estatísticas de log-rank

Na introdução aos testes de significância repetidos de uma hipótese, a variável básica X possuía distribuição normal, e a seqüência de somas parciais de variáveis independentes e com a mesma distribuição que X, constituía um processo estocástico no tempo (discreto) n, correspondente aos diversos tempos de análise. Este processo apresenta distribuições conjuntas finito-dimensionais próprias de um movimento Browniano.

Gail et al. (1981) mostraram que o valor da estatística de log-rank (1.28) e (1.32), calculada em um tempo de calendário t em que são observadas $d(t)$ mortes em um ensaio clínico com entrada não simultânea de pacientes, quando λ é próximo de $\lambda_0 = 1$, comporta-se como uma soma de variáveis aleatórias independentes W_j, $j = 1,..,d(t)$, com distribuição $\mathrm{N}(\frac{\log(\lambda)}{4}, \frac{1}{4})$. Se as N análises forem programadas para ocorrer nos tempos (aleatórios) de calendário, em que se completam $d, 2d, ..., Nd$ mortes, pode-se usar as fronteiras de regiões críticas estabelecidas para os testes de significância repetidos, como as de Pocock (1977), O'Brien & Fleming (1979), etc. Outras referências para este ponto podem ser encontradas em Tsiatis (1982), DeMets & Gail (1985), Selke & Siegmund (1983), Jenninson & Turnbull (1989).

3.2.2 A distribuição conjunta de N estatísticas WKM calculadas em N análises

Para determinar uma região crítica no teste de H_0, a partir dos valores calculados de *Weighted Kaplan-Meier* (WKM) nas N análises, no decorrer de um ensaio clínico, é preciso conhecer a distribuição conjunta dos valores desta estatística calculada em N momentos de análise. Li

(1999) mostrou que a distribuição assintótica é uma normal multivariada e determinou sua matriz de covariância teórica. Embora esta função de covariância não seja da forma exigida para que se aplique a simplificação proposta por Armitage et al. (1969) para o cálculo de fronteiras de regiões críticas para o teste de H_0, é possível proceder a este cálculo para determinados valores de nível de significância e poder de teste caso a caso, utilizando-se os resultados de Li (1999) e de Schervish(1984).

Para esta situação, Li (1999) propôs a seguinte notação para a estatística *Weighted Kaplan-Meier* (WKM) de Pepe & Fleming (1989) calculada em um instante de análise t.

$$WKM(t) = \sqrt{\frac{n_1 n_2}{n}} \int_0^{t_j} \hat{w}(s,t) \left[\hat{S}_{KM1}(s,t) - \hat{S}_{KM2}(s,t)\right] ds, \tag{3.1}$$

onde a diferença entre os estimadores de Kaplan-Meier para os dois grupos é ponderada em cada tempo por um peso dado por:

$$\hat{w}(s,t) = \frac{\hat{G}_1(u,t)\hat{G}_2(u,t)}{\hat{p}_1\hat{G}_1(t) + \hat{p}_2\hat{G}_2(t)}, \tag{3.2}$$

que é um processo aleatório com um limite determinístico $\hat{w}(s,t)$ que atribui um menor peso a $\hat{S}_{KM1}(s,t_j) - \hat{S}_{KM2}(s,t_j)$ nos períodos de tempos finais. Na presença de muita censura, é necessário determinar algumas condições de estabilidade para esta função peso que foram estabelecidas em (1.45). Com estas condições, o limite superior da integral definida em (3.1) pode ser substituído por $T^c(t) = \sup\{s : \min(\hat{G}_1(s,t), \hat{G}_2(s,t)) > 0\}$, onde $\hat{G}_i(s,t)$ é um estimador empírico de $G_i(s,t) = P(\min(C_{ij}, (t - Y_{ij})^+ \geq s) = C_i^-(s)E_i(t-s)$, sendo $C_i^-(x)$ a versão contínua à esquerda da função de sobrevivência do tempo potencial de censura aleatória C_{ij} e $E_i(x)$ a função distribuição do tempo de entrada Y_{ij}, sendo que o estimador de Kaplan-Meier de $S_{KMi}(x,t)$, é representado por $\hat{S}_{KMi}(x,t)$.

O referido autor determinou que a distribuição conjunta assintótica de $WKM(t)$, obtida nos diferentes tempo $t_1, ..., t_N$, sob a hipótese H_0, é normal multivariada com média zero. Cada covariância, relativa às estatísticas $WKM(t)$, em dois instantes de análise, $t_1 < t_2$ é dada por uma soma de integrais,

$$\sigma(t_1,t_2)=\sum_{i=1}^{2}(1-p_i)\int_0^{t_1}\frac{\left(\int_0^{t_1}w(s,t_1)S_i(x)dx\right)}{S_i(s)G_i(s,t_2)}\times\frac{\left(\int_0^{t_2}w(s,t_2)S_i(x)dx\right)}{S_i(s)G_i(s,t_2)}dH_i(s) \tag{3.3}$$

onde $H_i(s)=-\log(S_i(s))$.

A covariância $\sigma(t_1,t_2)$ pode ser estimada de duas maneiras:

a) usando, sob H_0, um estimador combinado de $S(x,t)$ indicado por $\hat{S}(s,t)$, sendo $\hat{S}(s,t)$ o estimador de Kaplan-Meier da função de sobrevivência na amostra combinada.

$$\hat{\sigma}(t_1,t_2)=\sum_{i=1}^{2}\sum_{j}^{n_i}\frac{\Delta_{ij}(t_1)\left(\int_{X_{ij}(t_1)}^{t_1}\hat{w}(x,t_1)\hat{S}(x,t_1)dx\right)\left(\int_{X_{ij}(t_1)}^{t_1}\hat{w}(x,t_2)\hat{S}(x,t_2)dx\right)}{\hat{S}(X_{ij}(t_1);t_2)\hat{G}(X_{ij}(t_1);t_2)R(X_{ij}(t_1);t_1)} \tag{3.4}$$

b) usando, sob H_0, estimadores de $S_1(x,t)$ e $S_2(x,t)$

$$\hat{\sigma}(t_1,t_2)=\sum_{i=1}^{2}\sum_{j}^{n_i}\frac{\Delta_{ij}(t_1)\left(\int_{X_{ij}(t_1)}^{t_1}\hat{w}(x,t_1)\hat{S}_i(x,t_1)dx\right)\left(\int_{X_{ij}(t_1)}^{t_1}\hat{w}(x,t_2)\hat{S}_i(x,t_2)dx\right)}{\hat{S}_i(X_{ij}(t_1);t_2)\hat{G}(X_{ij}(t_1);t_2)R(X_{ij}(t_1);t_1)} \tag{3.5}$$

A expressão do estimador da covariância $\hat{\sigma}(t_1,t_2)$, apresentada em (3.4) e (3.5), não coincide com a que foi obtida por Li (1999) pois esta última foi editada de forma incorreta no artigo. A correção foi verificada pelos autores deste trabalho e confirmada pelo autor.

Para a j-ésima análise, a estatística de teste padronizada é

$$Z(t_j)=\frac{WKM(t_j)}{\sqrt{\hat{\sigma}(t_j)}} \tag{3.6}$$

onde $\hat{\sigma}(t_j)$ é um estimador consistente da variância assintótica da estatística de teste no tempo t_j. Assim, uma estimativa consistente para a covariância entre $Z(t_1)$ e $Z(t_2)$ é

$$\frac{\hat{\sigma}(t_1,t_2)}{\{\hat{\sigma}(t_1)\hat{\sigma}(t_2)\}^{\frac{1}{2}}}$$

Seja α o nível de significância global. Para obter a região crítica em cada análise intermediária, de forma a ter um nível global α, é necessário calcular os valores de c_j , $j=1,...,N$

que satisfazem a igualdade

$$P\{|Z(t_1)| \leq c_j, ..., |Z(t_{j-1})| \leq c_{j-1}, |Z(t_j)| > c_j\} = \pi_j,$$

onde $\{\pi_j,\ j = 1, ..., N\}$ são escolhidos tal que $\sum_{j=1}^{N} \pi_j = \alpha$, sendo N o número máximo de análises intermediárias planejadas para o ensaio. Os valores de π_j podem ser especificados (Slud & Wei, 1982) ou calculados através da função gasto de alfa (Lan & DeMets, 1983). Como $\{Z(t_1), ..., Z(t_j)\}$ tem distribuição assintótica normal multivariada, e suas covariâncias podem ser estimadas consistentemente, os valores críticos podem ser obtidos por meio de integrações numéricas seqüenciais (Schervish, 1984). Feito isto, a rejeição de H_0 ocorrerá se $|Z(t_j)| > c_j$ para algum $j \leq N$.

Para implementar o procedimento gasto de alfa, é necessário especificar a função gasto de alfa, a informação e as frações de informação. A função gasto de alfa pode ser escolhida conforme se deseja gastar o α, como as de Pocock (1977) ou O'Brien & Fleming (1979). Por ser muito complicada a fórmula da variância, pode não ser possível alocar o erro tipo I de acordo com a variância da estatística. Entretanto, pode -se usar uma outra informação, como o número de eventos, para gastar o erro tipo I em cada análise. Supondo que uma das funções escolhidas seja $\alpha^*(t)$ e o estimador da fração de informação $\hat{v}_j$ para a j-ésima análise. Desta forma, a parcela do erro tipo I alocada à j-ésima análise, é $\pi_j = \alpha^*(\hat{v}_j) - \alpha^*(\hat{v}_{j-1})$.

3.3 Exemplos

3.3.1 Exemplo 1, com entrada simultânea de indivíduos no Ensaio

Se a intenção do pesquisador é realizar testes de significância repetidos não-paramétricos para avaliar se há diferença entre os tempos de sobrevivência dos tratamentos I e II, para um determinado tipo de doença, fica patente uma primeira característica destes estudos: a impossibilidade de utilizar um teste adequado, a menos que ocorra o evento em cada um dos grupos. De fato, antes da ocorrência de pelo menos um evento, os dados estão todos censurados e as estatísticas apresentadas não estão definidas. Desta forma, as análises intermediárias só podem ser programadas para ocasiões em que haja um acréscimo no número de eventos observados.

Para ilustrar esta idéia, utiliza-se o exemplo 4 da secção 1.4.1, capítulo 1, admitindo-se que o

recrutamento dos onze indivíduos foi feito em um instante inicial $t = 0$, e que estes tenham sido atribuídos aleatoriamente aos tratamentos I e II para um determinado tipo de câncer, fixando-se duas datas de análises uma no tempo 60 e a segunda, no tempo 75. Os tempos reais de sobrevida destes pacientes com as respectivas datas de análises, estão representados na Figura 3.1.

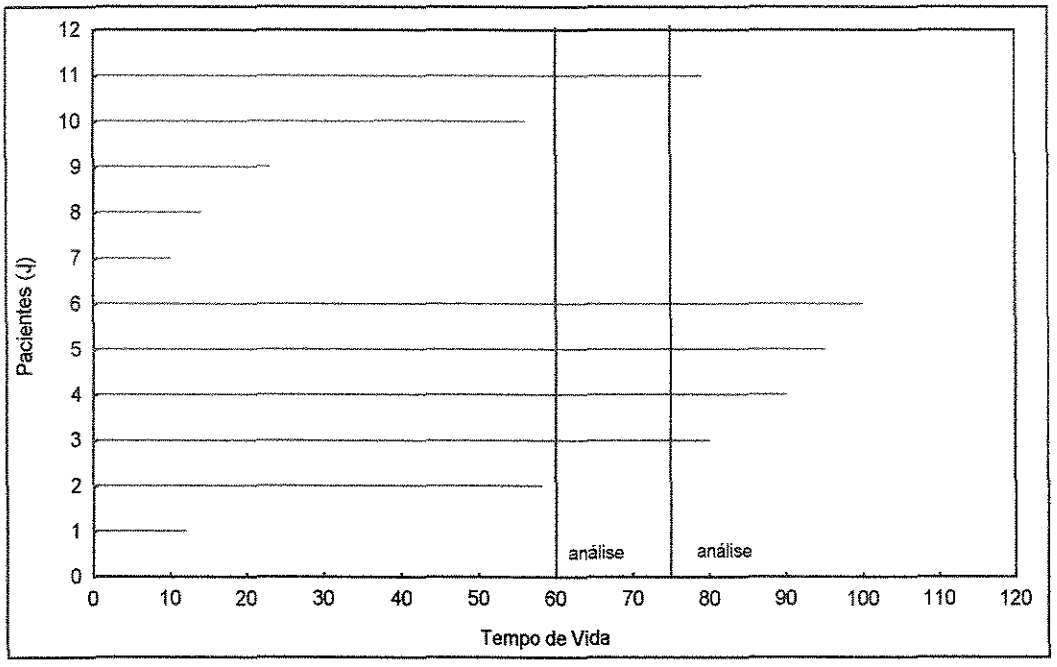

Figura 3.1: Períodos de observação de onze indivíduos desde sua entrada no estudo até sua morte. Dados referentes à Tabela 1.5. Vermelho representa tratamento 1 e azul o tratamento 2.

Analisando a Figura 3.1, percebe-se que não é possível comparar os dois grupos antes do tempo 10, pois, antes deste tempo, não ocorreu em nenhum dos grupos o evento de interesse e, as análises feitas entre os tempos 60 e 75 terão o mesmo resultado, porque entre os referidos tempos, não ocorreu um acréscimo no número de eventos observados. Desta forma, as análises só podem ser programadas para ocasiões em que haja um acréscimo no número de eventos observados.

As observações para análise nos tempos 60 e 75 estão apresentadas na Tabela 3.1 e nas Figuras 3.2 e 3.3, respectivamente, onde o símbolo (*) indica um valor censurado.

Tabela 3.1: Tratamento e posição (J_i) na amostra conjunta dos valores observados até morte ou censura (*). Dados com censura do tipo I simples τ=60.

Tratamento	1	1	1	1	1	1	2	2	2	2	2
J_i	1	2	3	4	5	6	7	8	9	10	11
t_{60}	12	58	60*	60*	60*	60*	10	14	23	56	60*
t_{70}	12	58	70*	70*	70*	70*	10	14	23	56	70*

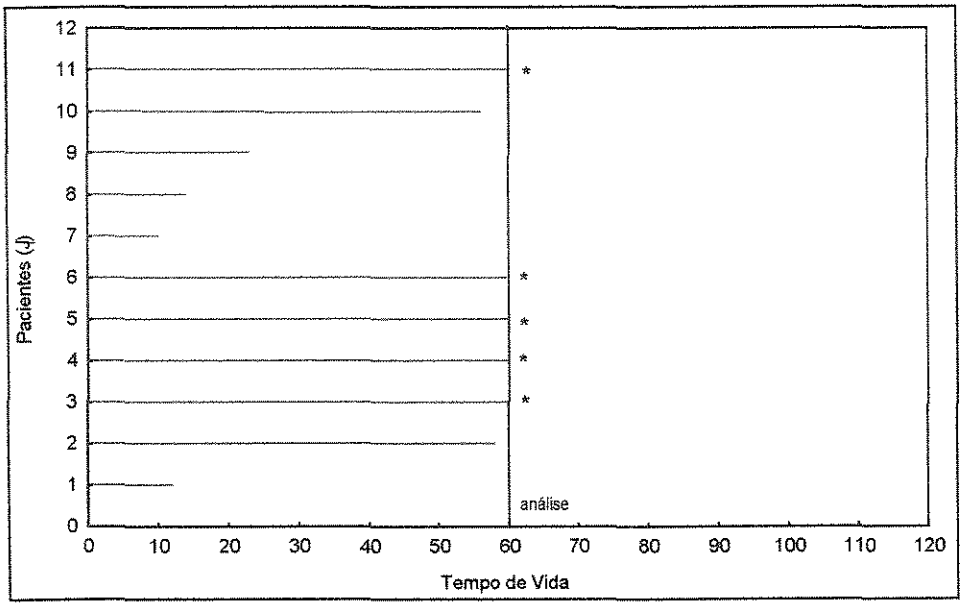

Figura 3.2: Períodos de observação de 11 indivíduos, com censura simples do tipo I (*) no tempo de análise 60. Dados referentes a Tabela 3.1. Vermelho representa o grupo 1 e azul o grupo 2.

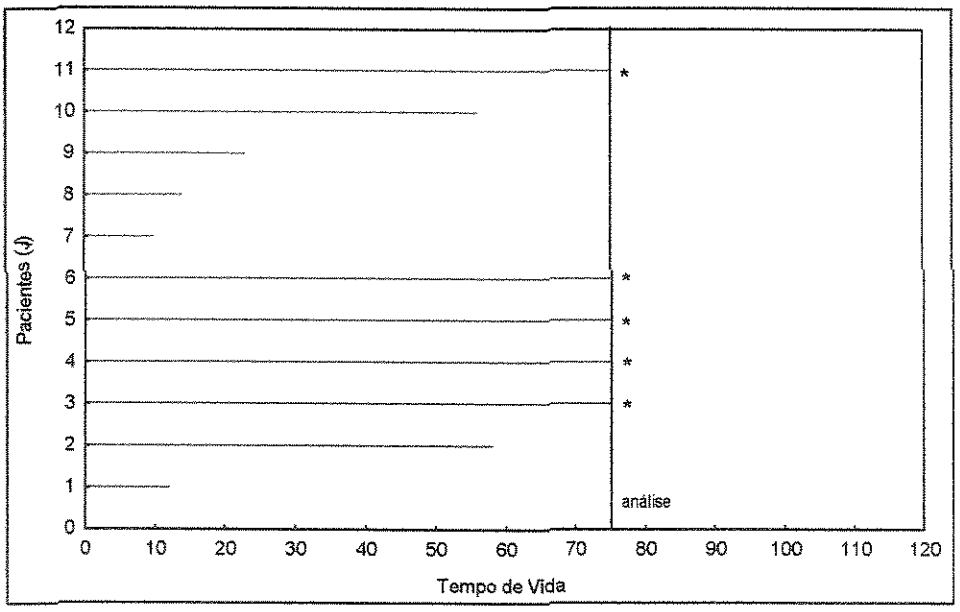

Figura 3.3: Períodos de observação de 11 indivíduos, com censura simples do tipo I (*) no tempo de análise 75. Dados referentes a Tabela 3.1. Vermelho representa o grupo 1 e azul o grupo 2.

Os resultados dos testes da hipótese H_0, nas datas de análise 60 e 75, que utilizaram as estatísticas de teste log-rank e WKM estão apresentados na Tabela 3.2.

Tabela 3.2: Valor das estatísticas log-rank, WKM e P-valor atingido em duas análises

Análise	Gasto de alfa previsto	Estatística	Nível descritivo (*P-valor*) $\star$
1	0,04396	WKM= 46,68	0,523
		LR= 2,67	0,589
2	0,00604	WKM= 46,68	0,523
		LR= 2,67	0,589

Com os resultados da Tabela 3.2, pode-se mostrar que as análises feitas nos tempos 60 e 75 apresentam o mesmo resultado e, considerando os valores das estatísticas WKM e log-rank para a primeira e segunda análise e seus respectivos P-valores empíricos para $\alpha = 0,05$, não se rejeita a hipótese nula; ou seja, as distribuições dos tempos de sobrevivência dos tratamentos I e II não são estatisticamente diferentes.

Neste exemplo, foi suposto que todos os indivíduos entraram no estudo ao mesmo tempo, e o único tipo de censura que existe em cada tempo deste estudo é de censura simples. É importante destacar que este conjunto de dados é limitado e não permite análises subseqüentes, sendo um exemplo puramente didático para mostrar que elas não podem ser planejadas em tempos quaisquer.

3.3.2 Exemplo 2, com entrada de pacientes não simultânea

Nesta seção, será apresentado o tipo de dificuldade que existe quando a entrada de indivíduos é distribuída ao longo de um intervalo de tempo, isto é, os pacientes não entram simultaneamente no estudo: o i-ésimo indivíduo entra no estudo em um tempo de calendário y_i. Após serem selecionadas, as pessoas são acompanhadas até a sua morte ou até uma data pré-determinada do calendário na qual o ensaio é finalizado. O período de tempo de calendário em que um indivíduo j é observado é denominado período de observação ou tempo de estudo e sua duração desde a sua entrada é o tempo de vida observado (X_{ij}) do indivíduo.

Neste exemplo, são considerados os dados do exemplo 4, seção 1.4.1, capítulo 1 no qual é introduzido artificialmente o **tempo de censura**. Este tempo de censura é o tempo decorrido entre a entrada do indivíduo e a ocorrência do evento que não o de interesse principal que geralmente, é a morte do indivíduo. Foram planejadas duas datas de análises: a primeira, no tempo de calendário t =18 e a segunda, no tempo $t = 22$. Então, supondo a ocorrência de censura aleatória, entrada não simultânea dos indivíduos no estudo e utilizando a notação: $a \wedge b = \min\{a, b\}$, $a^+ = \max\{a, 0\}$ no tempo $t = 18$ e $t = 22$, as observações poderiam ser descritas da forma apresentada na Tabela 3.3 e ilustradas nas Figuras 3.4 e 3.5, respectivamente.

Tabela 3.3: Dados referentes ao exemplo 4, seção 1.4.1, Capítulo 1, observados ao tempo de análise t_1 =18 e t_2 =22

J_i	1	2	3	4	5	6	7	8	9	10	11
Tratamento	1	1	1	1	1	1	2	2	2	2	2
y_{ij}	2	5	10	15	20	21	2	5	10	15	20
t_{ij}	12	58	80	90	95	100	10	14	23	56	79
c_{ij}	16	18	15	10	10	15	14	15	11	4	10
$L_j(t_1) = (18 - y_{ij})^+$	16	13	8	3	0	0	16	13	8	3	0
$t_{ij} \wedge (18 - y_{ij})^+$	12	13*	8*	3*	0	0	10	13	8*	3*	0
$c_{ij} \wedge t_{ij} \wedge (18 - y_{ij})^+$	12	13*	8*	3*	0	0	10	13	8*	3*	0
$L_j(t_2) = (22 - y_{ij})^+$	20	17	12	7	2	1	20	17	12	7	2
$t_{ij} \wedge (22 - y_{ij})^+$	12	17*	12*	7*	2*	1*	10	14	12*	7*	2*
$c_{ij} \wedge t_{ij} \wedge (22 - y_{ij})^+$	12	17*	12*	7*	2*	1*	10	14	11**	4**	2*

A oitava linha da Tabela 3.3 corresponde ao conjunto de dados observáveis ao tempo $t = 18$, onde o símbolo $*$ corresponde a um valor censurado e 0, as observações ainda não disponíveis no tempo $t = 18$. A última linha desta mesma tabela equivale ao conjunto de dados observáveis ao tempo $t = 22$; o símbolo ** indica uma observação que só é censurada pela ocorrência do evento que não o de interesse e $*$ continua representando um valor censurado pelo fato de não ter ocorrido ainda a morte, nem outro evento censurante ao tempo de análise $t = 22$.

A Figura 3.4 ilustra as observações disponíveis para a análise no tempo $t = 18$.

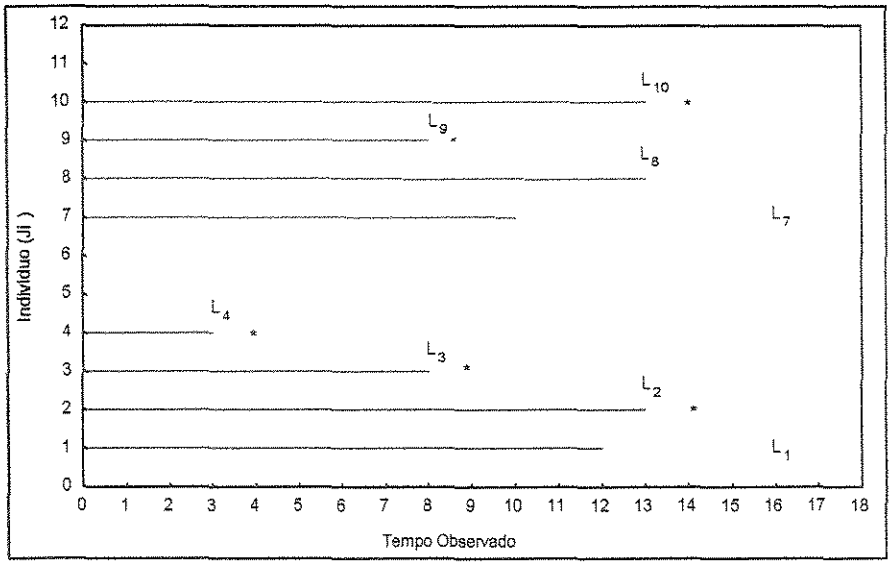

Figura 3.4: Tempos observados de vida, completos ou censurados (∗), e limites de observação L_j, para análise no tempo de calendário 18. Dados referentes a Tabela 3.3. Vermelho representa o grupo 1 e azul o grupo 2.

A Figura 3.5 ilustra as observações disponíveis para a análise no tempo $t = 22$.

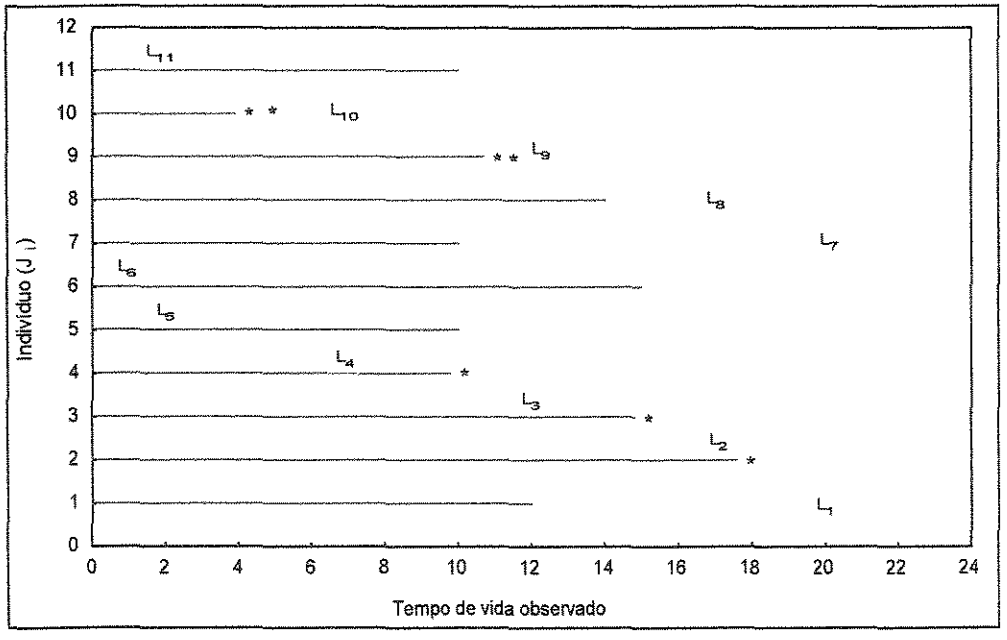

Figura 3.5: Tempos observados de vida completos ou censurados (∗ ou ∗∗) e limites de observação L_j, para análise no tempo de calendário 22. Dados referentes a Tabela 3.3. Vermelho representa o grupo 1 e azul o grupo2 .

Para avaliar se há diferença entre os tempos de sobrevivência dos tratamentos I e II para um determinado tipo de câncer, foram calculadas as estatísticas WKM e LR em duas datas de análises uma no tempo $t = 18$ outra em $t = 22$. Os resultados estão dispostos na Tabela 3.4.

Tabela 3.4: Valor das estatísticas LR e WKM em cada análise

Estatísticas	Análises	
	1	2
WKM	$0,707$	$1,211$
LR	$0,615$	$0,818$

A partir destes resultados, uma decisão quanto à não rejeição da hipótese de igualdade dos tempos de sobrevivência pode ser tomada, desde que se conheçam as fronteiras de uma região crítica adequada, de tamanho α fixado.

Como a distribuição conjunta das estatísticas não é semelhante à distribuição finito-dimensional de um movimento browniano não se pode utilizar as fronteiras de regiões críticas estabelecidas para os testes de significância repetidos propostas por Pocock (1977), O'Brien & Fleming (1979), entre outros.

No Capítulo 4 é apresentado um procedimento por reamostragem em testes repetidos que mantém o nível de significância global α que pode ser utilizado para qualquer estatística de teste. Segundo tal procedimento, foram obtidos os resultados da Tabela 3.5.

Tabela 3.5: Valor das estatísticas LR, WKM e P-valor atingido em duas análises

Análise	Gasto de alfa previsto	Estatística	Nível descritivo (*P-valor*) $\star$	Quantil empírico referente ao nível de significância global de 5%
1	0,04396	$WKM = 0,707$	0,352	1,006066
		$LR = 0,615$	0,398	3,00
2	0,00604	$WKM = 1,211$	0,292	1,868493
		$LR = 0,818$	0,246	4,898305

$\star$ *P-valor* empírico baseado em reamostragem da estatística.

Considerando, desta forma, os valores das estatísticas WKM e LR para a primeira e segunda análise e seus respectivos *P-valores* empíricos de 0,352 e 0,244 e 0,398, 0,246, respectivamente,

para $\alpha = 0,05$, não se rejeita a hipótese nula, isto é, as distribuições dos tempos de sobrevivência dos dois tratamentos não são estatisticamente diferentes.

Capítulo 4

Estudo por simulação

Nesta seção é descrito o estudo, feito por meio de simulação, da distribuição das estatísticas *Weighted Kaplan-Meier* (WKM) e log-rank (LR) sob H_0, e do tamanho e poder de um teste *bootstrap* da hipótese H_0 : $S_2(t)$=$S_1(t)$ vs H_1 : $S_2(t)$>$S_1(t)$. através destas estatísticas, no contexto de testes de significância repetidos, considerando amostras de vários tamanhos.

Como a estatística *Weighted Kaplan-Meier* (WKM) não está computacionalmente definida nos *softwares* estatísticos disponíveis, implementou-se inicialmente um programa em S-Plus para sua obtenção.

4.1 Estudo das distribuições de WKM e LR

Para o estudo das propriedades da distribuição das estatísticas WKM e (LR), simulou-se quatrocentas vezes a entrada aleatória de $2n$ pacientes em um ensaio clínico, com dois tratamentos e com planejamento de duas análises a serem feitas em dois instantes do tempo. Os tempos de entrada (Y_{ij}) e de vida (T_{ij}), para $i = 1, 2$ e $j = 1, 2, .., n_i$ destes pacientes foram gerados segundo uma distribuição exponencial $Exp(\lambda_i)$, $i = 1, 2$. Os parâmetros das distribuições de Y_{ij} e T_{ij} estão descritos nas Tabelas 4.1 e 4.2. Os tempos potenciais de censura (C_i), $i = 1, 2$ foram gerados conforme uma distribuição uniforme U$(0, c)$ para $c > 0$. O parâmetro (c) foi escolhido de forma que as amostras retiradas de cada tratamento apresentassem uma porcentagem esperada de valores censurados em torno de 25%, sob a hipótes nula. Para este estudo utilizou-se amostras de tamanhos 30, 50 e 100 por tratamento.

A Tabela 4.1 apresenta as distribuições, sob a hipótese nula H_0, das variáveis envolvidas no estudo.

Tabela 4.1: Distribuições, sob H_0, das variáveis tempo potencial de vida (T_{ij}), censura (C_i) e tempo de entrada (Y_{ij}) para uma porcentagem esperada de censura em torno de 25%.

Variável	Distribuição
Y_i	$Exp(0,50)$
T_i	$Exp(0,20)$
C_i	$U(0;20)$

Calculou-se para cada uma das 400 amostras geradas, as estatísticas WKM e LR referentes aos dois momentos de análise. Através dos valores WKM_{01l} e WKM_{02l}, $l = 1,..,400$, determinou-se a distribuição empírica de cada uma das estatísticas WKM_1 e WKM_2 que foi comparada com uma distribuição normal, através de gráficos de densidades marginais e de gráficos de quantis da distribuição normal versus quantis da distribuição empírica . A partir dos valores LR_{01l} e LR_{02l}, $l = 1,..,400$ a mesma análise foi feita para para as estatísticas LR_1 e LR_2, porém utilizando-se a distribuição qui-quadrado com 1 grau de liberdade no gráfico de quantis. Calculou-se, também, o vetor de médias, a matriz de covariâncias e a matriz de correlações das estatísticas WKM (nos dois tempos de análises) e log-rank (LR).

4.2 Estudo do tamanho e poder de um teste Bootstrap de H_0

Um teste por Bootstrap foi proposto para testar a hipótese nula. Foi feito um estudo por simulação do tamanho e poder deste teste. O procedimento geral de *Bootstrap* está explicado nos parágrafos a seguir.

O método *Bootstrap*, proposto por Efron (1979), é utilizado neste trabalho para obtenção da região crítica dos testes da hipótese H_0 definidos pelas estatísticas *Weighted Kaplan-Meier* (WKM) e log-rank (LR), no contexto de testes significância repetidos.

O método *Bootstrap* não-paramétrico consiste em uma técnica que usa a reamostragem de um conjunto de observações de uma variável aleatória X para estimar a distribuição teórica de uma função desta variável aleatória X através da distribuição empírica. A idéia é tratar os dados (amostra original) como se eles fossem a população, e retirar B amostras aleatórias

de tamanho n, com reposição dos dados como se estivesse amostrando a população e, em cada reamostra, calcular a quantidade de interesse. Este processo de gerar uma amostra original e reamostrá-la com reposição deve ser repetido um grande número de vezes (Cribari & Neto, 2000). No *Bootstrap* paramétrico, a família paramétrica da distribuição de X é conhecida e a amostra original fornece as estimativas $(\tilde{\theta})$ dos parâmetros θ desta distribuição. A seguir são geradas B amostras bootstrap da distribuição pertencentes a esta família com $\theta = \tilde{\theta}$.

O procedimento de simulação utilizado neste trabalho utiliza um tipo de *Bootstrap* não-paramétrico e segue o algoritmo abaixo:

1. Gera-se um ensaio clínico com n_i, $i = 1, 2$ pacientes para cada um dos tratamentos correspondendo a geração de vetores de tamanho n_i de tempos de entrada, tempos de censura e tempos de vida. No primeiro momento de análise (t_1) calcula-se a estatística WKM_{01} com as observações disponiveis referentes ao tempo observado de vida e a variável indicadora da censura. No instante da primeira análise o número de observações disponíveis em cada tratamento não é necessariamente o mesmo, sendo indicado por n_{i1}, $i = 1, 2$.

2. Faz-se uma reamostragem por bootstrap não-paramétrico, de tamanho B da seguinte forma:

a- Retira-se uma amostra bootstrap da amostra combinada de observações disponíveis no tempo t_1, atribuindo-se os n_{11} primeiros valores ao tratamento 1, e os n_{21} restantes ao tratamento 2. Calcula-se a estatística *(Bootstrap)* WKM, referente ao primeiro momento de análise t_1, obtendo-se WKM_{11}^*;

b- Repete-se o passo **a** independentemente B vezes, obtendo-se, ao final, as observações, $(WKM_{11}^*, ..., WKM_{B1}^*)$;

c-Determina-se a distribuição empírica marginal obtida por bootstrap de WKM_1, utilizando-se os B valores de WKM_{j1}^*, $j = 1, ..., B$ descritos nos itens **a** e **b**. Esta distribuição empírica será a distribuição de referência para realizar um teste da hipótese nula a um nível de significância α_1.

Se o valor de WKM_{01}(do passo 1) for maior que p_1, o quantil $(1 - \alpha_1)$ da distribuição empírica da estatística WKM_1, rejeita-se H_0 na primeira análise, concluindo-se o estudo.

d- Se a decisão de não rejeitar H_0 for tomada, o ensaio prossegue, com uma segunda análise no tempo t_2, utilizando as observações disponíveis da amostra gerada no passo 1. Como foi

visto no Capítulo 3, na segunda análise são utilizadas as observações disponíveis no instante t_2 referentes aos pacientes já observados no instante t_1 e de novos pacientes que entraram no ensaio no decorrer do tempo de estudo. Calcula-se a estatística WKM_{02}, referente ao segundo momento de análise. O número de observações disponíveis referentes ao tratamento i será indicado por n_{i2}

e- Retira-se uma amostra bootstrap da amostra combinada de observações disponíveis no tempo t_2, atribuindo as n_{i2} primeiros valores ao tratamento 1. Calcula-se a estatística (*Bootstrap*) de interesse obtendo-se WKM^*_{12}. Repete-se o passo a independentemente B vezes, obtendo-se, ao final, as observações, $(WKM^*_{12}, .., WKM^*_{B2})$.

f- Determina-se a distribuição empírica condicional de WKM_2 dado que WKM_1 não caiu na região crítica de tamanho α_1. Para isto é preciso retirar da amostra *Bootstrap* todas aquelas que teriam sido rejeitadas na primeira análise. Esta distribuição empírica será a distribuição de referência para um teste de H_0 na segunda análise a um nível de significância α_2. Se o valor de WKM_{02} for maior que p_2, o quantil $(1-\alpha_2)$ da distribuição empírica condicional da estatística WKM_2, rejeita-se H_0 na segunda análise, concluindo-se o estudo.

3. Retorna-se ao passo 1 repetindo-se os passos 1 e 2 R vezes, obtendo-se R simulações de um teste de H_0

Os valores de B e R devem ser grandes; no entanto, devido ao fato do algoritmo computacional requerer muito tempo de processamento, utilizou-se $B = 200$ e $R = 400$.

O tamanho global (α) do procedimento será $\alpha_1 + \alpha_2(1-\alpha_1)$.

O tamanho real empírico do teste pode ser obtido calculando-se a proporção de vezes que a hipótese nula foi rejeitada, isto é, a proporção de vezes que a estatística WKM_{01}, WKM_{02} para cada instante de análise, caiu na respectiva região crítica.

O procedimento para avaliar o poder do teste bootstrap, no contexto de testes de significância repetidos, foi feito nos mesmos moldes do esquema descrito anteriormente, com a diferença de que foram escolhidos parâmetros diferentes para a distribuição do tempo potencial de vida e de censura para os tratamentos I e II. Considerou-se amostras de tamanho 30, 50 e 100 por tratamento, e nos processos de reamostragem, duzentas reamostras. O poder real do teste foi estimado pela porcentagem de rejeições de H_0 entre as simulações.

A Tabela 4.2 apresenta os parâmetros escolhidos em cada tratamento, de forma que as

amostras retiradas de cada tratamento apresentassem uma porcentagem esperada de censura em torno de 25% para o estudo do poder do teste.

Tabela 4.2: Distribuições, sob H_1, das variáveis tempo potencial de vida (T_{ij}), censura (C_i) e tempo de entrada (Y_{ij}) para uma porcentagem esperada de censura em torno de 25%.

Variável	Tratamento	
	I	II
Y_i	$Exp(0,50)$	$Exp(0,50)$
T_i	$Exp(0,10)$	$Exp(0,20)$
C_i	$U(0;40)$	$U(0,20)$

Os mesmos passos são seguidos para um teste *Bootstrap* de significância repetido de H_0 utilizando a estatística LR.

4.3 Resultados

Nesta seção, inicialmente, é apresentado o estudo, feito por meio de simulação, da distribuição das estatísticas *Weighted Kaplan-Meier* (WKM) e log-rank (LR) sob a hipótese H_0, no contexto de testes de significância repetidos, considerando amostras de vários tamanhos. Posteriormente, são apresentados os resultados obtidos para o tamanho e poder do teste H_0 através destas estatísticas, no contexto de testes de significância repetidos para amostras de vários tamanhos.

A Tabela 4.3 apresenta as médias dos quatrocentos valores das estatísticas *Weighted Kaplan-Meier* (WKM) e log-rank (LR) em cada uma das análises, considerando-se os dados com censura para diferentes tamanhos de amostras. Os valores obtidos para WKM estão em torno de zero tanto na primeira quanto na segunda análise, e os encontrados para LR, estão próximos de um (valor esperado de uma distribuição Qui-quadrado com 1 grau de liberdade).

Tabela 4.3: Média dos 400 valores das estatísticas WKM e LR em cada instante de análise

Tamanho de amostra	WKM		LR	
	1ª Análise	2ª Análise	1ª Análise	2ª Análise
30	-0,0037	-0,0336	1,2492	1,1450
50	-0,0534	-0,0527	1,0494	0,9999
100	-0,0258	-0,0517	1,0883	1,0625

Na Tabela 4.4, encontram-se os resultados obtidos, por simulação, para a matriz de covariâncias das estatísticas WKM e LR, sob H_0, nos dois instantes de análises para os diferentes tamanhos de amostra. Os resultados encontrados para WKM revelam uma variabilidade maior na segunda análise, a qual cresce à medida que aumenta-se o tamanho da amostra e, para a estatística log-rank, isto não ocorre, pois a variabilidade é praticamente a mesma, tanto na primeira, quanto na segunda análise.

Tabela 4.4: Covariâncias das estatísticas WKM e LR nos dois instantes de análises

Tamanho de amostra	Covariâncias de WKM			Covariâncias de LR		
	S_1^2	S_2^2	S_{12}	S_1^2	S_2^2	S_{12}
30	1,8100	6,9179	2,6979	3,5425	3,1724	2,0061
50	1,6667	7,0091	2,5518	2,5026	2,1637	1,3069
100	1,7189	8,3014	2,9210	2,3101	2,4969	1,3234

A Tabelas 4.5 apresenta os resultados obtidos, por simulação, para a matriz de correlação das estatísticas WKM e LR, sob H_0, nos dois instantes de análises, para os diferentes tamanhos de amostra.

Tabela 4.5: Estimativa da correlação das estatísticas WKM e LR nos dois instantes de análises

Tamanho de amostra	Correlação de WKM	Correlação de LR
	r_{12}	r_{12}
30	0,7624	0,5984
50	0,7466	0,5616
100	0,7733	0,5511

As Figuras 4.1 e 4.2 apresentam os gráficos "q-qplots" e os histogramas da distribuição empírica da estatística WKM, considerando amostras de tamanho 30, 50 e 100, para o primeiro e segundo instantes de análise, respectivamente.

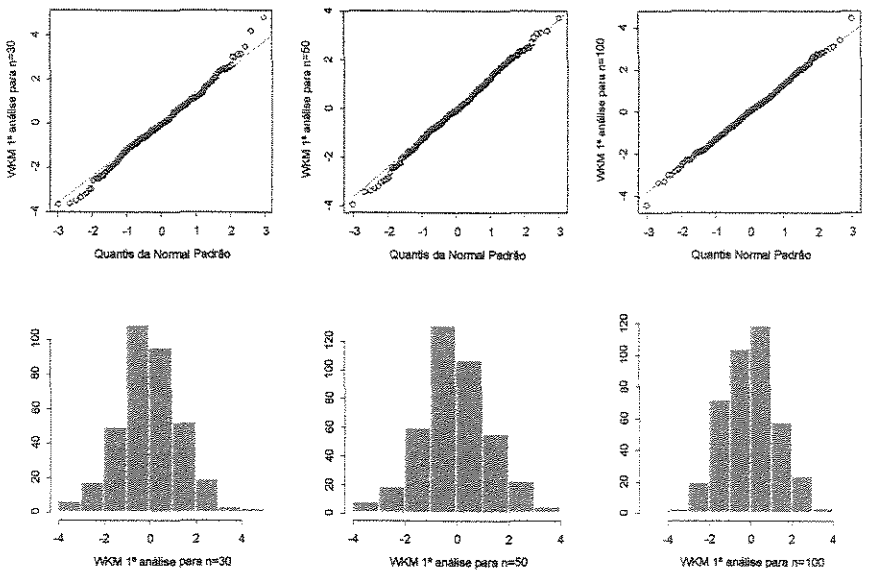

Figura 4.1: Q-qplots e Histogramas da distribuição empírica da estatística WKM no primeiro instante de análise para amostras de tamanho 30, 50 e 100, respectivamente.

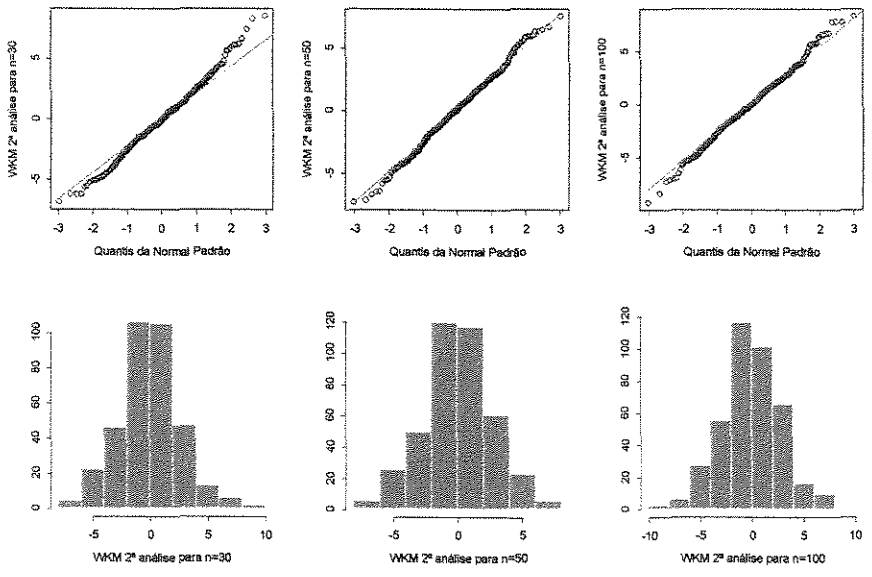

Figura 4.2: Q-qplots e Histogramas da distribuição empírica da estatística WKM no segundo instante de análise para amostras de tamanho 30, 50 e 100, respectivamente.

As Figuras 4.1 e 4.2 sugerem que a distribuição dos valores da estatística WKM, tanto na primeira, quanto na segunda análise, aproxima-se da distribuição normal, ainda que para

amostras de tamanho 30, as caudas sejam mais pesadas do que as da distribuição normal.

As figuras a seguir referem-se ao estudo realizado para a estatística log-rank (LR), sob H_0, nos dois instantes de análise, a partir de 400 amostras de tamanhos 30, 50 e 100. Nas Figuras 4.3 e 4.4, comparam-se os quantis da distribuição empírica da estatística log-rank, na primeira e na segunda análise, respectivamente, com os quantis da distribuição $\chi^2_{(1)}$, apresentam-se, também, os histogramas da distribuição empírica da estatística (LR). Nota-se, neste caso, que a distribuição empírica da estatística (LR) afasta-se da distribuição $\chi^2_{(1)}$ na cauda direita da curva, tanto na primeira como na segunda análise.

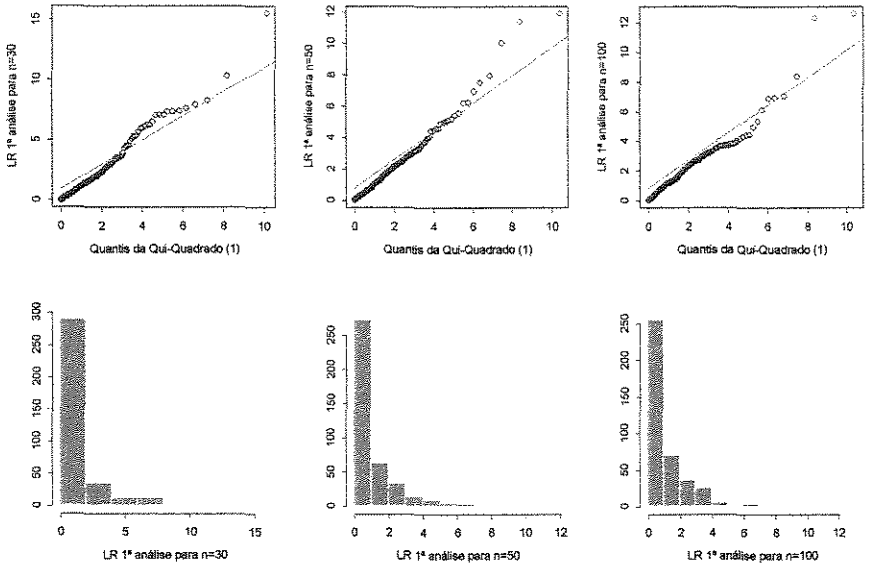

Figura 4.3: Q-qplots e Histogramas da distribuição empírica da estatística LR no primeiro instante de análise para amostras de tamanho 30, 50 e 100, respectivamente.

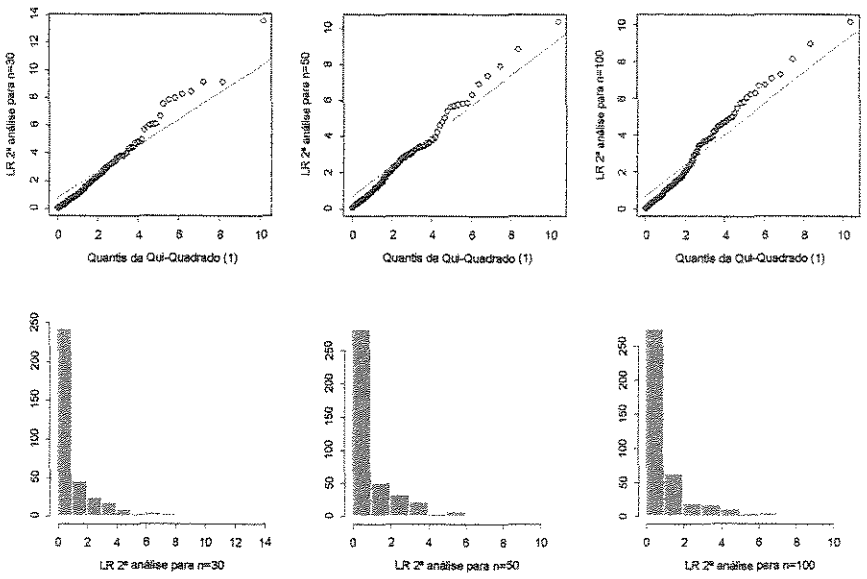

Figura 4.4: Q-qplots e Histogramas da distribuição empírica da estatística LR no segundo instante de análise para amostras de tamanho 30, 50 e 100, respectivamente.

As tabelas apresentadas, a seguir, mostram os resultados obtidos no estudo do tamanho e poder do teste da hipótese H_0, através das estatísticas WKM e LR, no contexto de testes de significância repetidos considerando duas análises, e amostras de vários tamanhos e $\alpha = 0,05$.

A Tabela 4.6 apresenta os valores obtidos do tamanho real empírico do teste bootstrap definido pelas estatísticas WKM e LR para amostras de tamanhos 30, 50 e 100. Nota-se que o teste bootstrap, usando a estatística de teste WKM, tem nível real empírico mais próximo de 0,05 para amostra de tamanho 50, embora os resultados apresentados para as de tamanho 30 e 100 não sejam ruins. Os resultados obtidos mostram que os testes bootstraps, usando a estatística de teste LR, têm nível real empírico mais próximo de 0,05 para amostras de tamanhos 50 e 100.

Tabela 4.6: Tamanho real empírico do teste da hipótese H_0 definido por WKM e LR

Tamanho da amostra	WKM	LR
30	0,0627	0,0826
50	0,0574	0,0498
100	0,0628	0,0578

Os resultados do poder real do teste bootstrap para WKM e LR, obtidos por simulação de 400 repetições, considerando $\alpha = 0,05$, são apresentados na Tabelas 4.7.

Tabela 4.7: Poder real empírico do teste da hipótese H_0 definido por WKM e LR.

Tamanho da amostra	WKM	LR
30	0,562	0,508
50	0,795	0,727
100	0,975	0,956

Os resultados acima mostram que o poder do teste de para detectar uma diferença $\delta = 0,10 - 0,20$ dos parâmetros das distribuições (exponenciais) do tempo de vida, usando a estatística de teste WKM, mostrou-se ligeiramente superior ao da estatística log-rank (LR). Para amostras de tamanho 50 e 100, o poder do teste, definido por estas estatísticas, pode ser considerado bom.

Capítulo 5

Conclusões

A proposta deste trabalho consistiu em expor a teoria de testes de significância repetidos, sob o ponto de vista do nível real de teste, com particular relevância para os testes não-paramétricos usados na área de análise de sobrevivência. As questões que foram apresentadas no capítulo 2 sustentam que a idéia de que o recrutamento de pacientes ao longo do tempo é, muitas vezes, a única maneira de se obter o tamanho da amostra planejado para fazer estudos comparativos, e que a prática de realizar várias análises no decorrer do estudo tornou-se comum entre os pesquisadores de ensaios clínicos. Este planejamento possibilita a tomada de decisão a respeito dos tratamentos antes do final do estudo, quando um destes começa a se mostrar superior ao outro. Quando se planeja fazer vários testes de significância em diversos momentos de um ensaio clínico, à medida que os dados são obtidos, é preciso cuidado para escolher o nível de significância em cada análise para se conseguir um nível α global pré-estabelecido.

As regiões críticas para um teste de igualdade de médias de uma variável com distribuição normal, a partir de estatísticas com uma distribuição conjunta finito-dimensional igual à de uma seqüência de somas parciais de variáveis independentes com distribuição normal, estão bem estabelecidas e vários autores propuseram diferentes fronteiras para a região crítica do teste; no campo não-paramétrico, procurou-se estender o uso destas fronteiras para o caso das estatísticas usadas em análise de sobrevivência. Se a estatística usada nos testes de significância repetidos é a log-rank, e se as N análises no decorrer do ensaio são programadas para ocorrerem em épocas (aleatórias) do calendário, em que se alcança números pré-fixados de mortes, pode-se usar as fronteiras de regiões críticas estabelecidas para os testes de significância repetidos, como as de

Pocock, O'Brien, etc, em que a variável resposta tem distribuição normal em cada população. Entretanto, para outras estatísticas cujas distribuições conjuntas não satisfazem estes padrões, como por exemplo, a *Weighted Kaplan-Meier* (WKM), o uso direto das fronteiras de Pocock e de outros autores não pode ser realizado, existindo a necessidade de se fazer pesquisas sobre técnicas alternativas para a realização de testes de significância repetidos.

O objetivo do presente trabalho foi estudar os métodos para calcular o nível de significância em cada análise, de modo a preservar o nível de significância global pré-estabelecido do teste. Este trabalho consistiu em uma revisão detalhada da metodologia existente, sua implementação metodológica e prática para uso em estudos clínicos, em especial quando se trata de métodos de análise de sobrevivência, tais como log-rank e *Weighted Kaplan-Meier* (WKM). No caso deste último teste, esta pesquisa apresenta um estudo de simulação não-paramétrico para o estabelecimento da regra de decisão em cada uma das análises, método este que não está descrito na literatura, mas foi implementado neste trabalho.

Através do estudo de simulação realizado, pode-se observar que, tanto para amostras pequenas quanto para amostras grandes, a aproximação pela distribuição normal para a distribuição empírica da estatística *Weighted Kaplan-Meier* (WKM), relativa ao teste de hipótese H_0 de igualdade de duas funções de sobrevivência, mostrou-se satisfatória. Com relação ao estudo sobre o tamanho real do teste, através do método de reamostragem (bootstrap) da hipótese H_0, pode-se notar que as estatísticas log-rank (LR) e *Weighted Kaplan-Meier* (WKM) apresentam praticamente o mesmo comportamento, ou seja, com tamanho real empírico próximo a $0,05$, escolhido para o teste. No que tange ao poder do teste por bootstrap, os resultados mostraram que para detectar uma diferença $\delta = 0,10 - 0,20$ dos parâmetros das distribuições (exponenciais) do tempo de vida, usando a estatística *Weighted Kaplan-Meier* (WKM) é ligeiramente superior ao da estatística log-rank (LR). Pode-se observar que, tanto o tamanho, quanto o poder do teste que utilizam esta estatística, são adequados mesmo para amostras pequenas. A desvantagem de seu uso reside no fato de que ela não está implementada, até o momento, nos *softwares* estatísticos existentes, e que o caminho aqui percorrido constitui na elaboração de um programa para sua obtenção.

Esse trabalho constitui uma abertura, inclusive na interação das áreas de sobrevivência e ensaios clínicos, propiciando a realização de futuras pesquisas nessa área.

Referências Bibliográficas

[1] Armitage, P., McPherson, C. K. and Rowe, B.C.. Repeated significance tests on accumulated data. **Journal of the Royal Statistical Society** A, v. 132, p. 235-244, 1969.

[2] Armitage, P. **Sequential Medical Trials**. 2nd edition. New York: John Wiley and Sons, 1975. 275 p.

[3] Amberson, J. B.; McMahon, B. T.; Pinner, M. A clinical trial of sanocrysin in pulmonary turberculosis. **Am Rev Turberculosis** , v. 24; p. 401-435, 1931.

[4] Breslow, N.E. Contribution to the discussion of a paper by D. R. Cox. **Journal of the Royal Statistical Society** B, v. 34, p. 216-217, *1972.*

[5] Breslow, N.E. Covariance analysis of censored survival data. **Biometrics**, v. 30, p. 89-100, 1974.

[6] Canner, P. L. Monitoring treatment differences in long-term clinical trials. **Biometrics**, v. 33, p. 603-615, 1977.

[7] Collet, D. **Modelling survival data in medical research**. London: Chapman and Hall, 1994. 347p.

[8] Colosimo, E. A. **Análise de sobrevivência aplicada**. 46ª Reunião Anual da RBRAS e 9° SEAGRO. Piracicaba, 2001. 145p.

[9] Cox, D. R. Regression models and life tables (with discussion). **Journal of the Royal Statistical Society,** B, v. 34, p. 187-220, 1972.

[10] Cox, D. R. Partial likelihood. **Biometrics**, v. 62, p. 269-276, 1975.

[11] Cox, D. R; Hinkley, D. V. **Theoretical statistics.** London: Chapman and Hall, 1974. 511p.

[12] Cox, D. R; Oakes, D. **Analysis of survival data.** London: Chapman and Hall, 1984. 201p.

[13] DeMets, D. L; Gail, M. H. Use of log-rank tests and group sequential methods at fixed calendar times. **Biometrics**, v. 41, p. 1039-1044, 1985.

[14] DeMets, D. L ;Lan, K. K. G. The alpha spending function approach to interim data analyses. **Recent Advances in Clinical Trial Design and Analysis**, p.27, 1995.

[15] Efron, B. Bootstrap methods: another look at the jackknife. **The Annals of Statistics**, v. 7, p. 1-26, 1979.

[16] Fleming, T. R.; Harrington, D. P. **Counting Processes and Survival Analysis.** New York: John Wiley, 1991.

[17] Friedman, L. M., Furberg, C. D. ;DeMets, D. L. **Fundamentals of Clinical Trials.** Second edition. PSG Publishing Company, INC. Littleton, Massachusetts, 1985. 302 p.

[18] Gail, M. H.; DeMets, D. L; Slud, E. V. Simulation studies on increments of two-sample log-rank score test for survival time data, with application to group sequential boudaries. In **Survival Analysis**, Monograph series 2. Edited by R. Johnson ;J. Crowley. Hayward, California: IMS Lecture Notes, v. 2, p. 287-301, 1982.

[19] Gehan, E. A. A generalized Wilcoxon text for comparing arbitrarily single-censored samples. **Biometrika**, v. 52, p. 203-223, 1965.

[20] Hollander, M.; Wolfe, D. A. **Nonparametric Statistical Methods.**New York: John Wiley and Sons,1973. 503 p.

[21] Jennison, C.; Turnbull, B. W. Interim analyses: repeated confidence interval approach. **J. R. Statist. Society. A**, v. 3, p. 305-361, 1989.

[22] Lan, K. K.; DeMets, D. L. Discrete sequential boundaries for clinical trials. **Biometrika**, v. 70, p. 659-663, 1983.

[23] Lan, K. K. G.; Zucker, D. Sequential monitoring of clinical trials: the role of information and brownian motion. **Statistics in Medicine, v.** 12, p. 753-765, 1993.

[24] Lawless, J. F. **Statistical models and methods for liftime data.** New York: John Wiley, 1982. 450p.

[25] Lee, E. T. **Statistical methods for survival data analysis.** Second edition. New York: John Wiley, 1992. 482 p.

[26] Li, Z. A group sequential test for survival trials: An alternative to rank-based procedures. **Biometrics,** v. 55, p. 277- 283, 1999.

[27] Mantel, N. Evaluation of Survival data and two new rank order statistics arising in its consideration. **Cancer Chemotherapy Reports**, 50, p.163-170, 1966.

[28] Mantel, N.; Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. **Journal of the National Cancer Institute**, v. 22, p. 719-748, 1959.

[29] Meier, P. Statistics and medical examination. **Biometrics**, v. 31, p. 511-529, 1975.

[30] Nelson, W. Hazard plotting for incomplete failure data. **J. Qual. Technol.**, v. 1, p. 27-52, 1969.

[31] O 'Brien, P. C.; Fleming, T. R. A multiple testing procedure for clinical trials. **Biometrics,** v. 35, p. 549-556, 1979.

[32] Pepe, M. S.; Fleming, T. R. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. **Biometrics**, v. 45, p.1468-1475, 1989.

[33] Pepe, M. S.; Fleming, T. R. Weighted Kaplan-Meier statistics: large sample and optimality considerations. **Journal of the Royal Statistical Society** B, v. 53, p. 341-352, 1991.

[34] Peto, R.; Peto, J. Asymptotically efficient rank invariant test procedures (with discussion). **Journal of the Royal Statistical Society**, A, v. 135, p. 185-206, 1972.

[35] Pocock, S. J. Group sequential methods in the design and analysis of clinical trials. **Biometrika**, v. 64, p. 191-199, 1977.

[36] Reboussin, D. M., DeMets, D. L., Kim, K.; Lan, K. K. G. Programs for computing group sequential bounds using the Lan-DeMets method. **Technical Report 95, Department of Statistics**, University of Wisconsin, 1995.

[37] Robbins, H. Statistical methods related to the law or iteraded logarithm. **Ann. Math. Stat.**, v. 41, p. 1397-1409, 1970.

[38] Schervish, M. J. Multivariate normal probabilities with error bound. **Applied Statistics**, v. 33, p. 81-94, 1984.

[39] Selke, T.; Siegmund, D. Sequential analysis of the proportional hazards model. **Biometrika**, v. 70, p. 315-326, 1983.

[40] Shen, Y.; Fleming, T. R. Weighted mean survival test statistics: a class of distance tests for censored survival data. **Journal of the Royal Statistical Society** B, v. 59, p. 269-280, 1997.

[41] Siegmund, D. Sequential Analysis: Tests and Confidence Intervals. New York: Springer, 1985.

[42] Slud, E. V.; Wei, L. J. Two-sample repeated significance tests based on the modified Wilcoxon statisitc. **Journal of the American Statistical Association**, v. 77, p. 862-868, 1982.

[43] Tsiatis, A. A. Repeated significance testing for a general class of statistics used in censored survival analysis. **Journal of the American Statistical Association,** v.77, p. 855-861, 1982.

[44] Wald, A. **Sequential Analysis**. New York: John Wiley and Sons, 1947.

[45] Whitehead, J. **The Design and Analysis of Sequential Clinical Trials.** Second edition. Chichester: Horwood, 1992. 297 p.

[46] Wilcoxon, F. Individual comparison by ranking methods. **Biometrics**, v. 1, p. 80-83, 1945.