

选择我们 脱颖而出

医学统计学实用教程 (2)

Statistical Methods in Medicine

www.MedSci.cn 张发宝 博士
2009.3



声 明

- 我们[MedSci团队](#)日常工作接审大量论文稿件，发现统计学问题很大，因此制作本幻灯片，希望对大家论文写作与统计有一定的帮助。
- 本幻灯片是综合目前流行的许多统计专家的讲座，并进行进一步加工而成，在此向原作者表示深深谢意！
- 幻灯片统计学软件是基于SPSS软件，因此，需要有初级统计学基础和软件基础知识。
- 本幻灯片仅仅是其中一部分，前面和后续请关注这里并进行下载：
<http://www.medsci.cn/news.asp?id=20>
- 有关SPSS教程可以参见：<http://www.bioon.com/biology/spss/Index.shtml>

第八节 生存分析 **Survival Analysis**

生存分析

- 在临床医学中,对病人治疗效果的考查.一方面可以看治疗结局的好坏,另一方面还可以通过治疗时间的长短来衡量。
- 例如某癌症病人手术后的存活时间,某种疾病治愈的时间等,把这类与时间有关的资料统称为生存资料。
- **生存分析 (survival analysis)** 是将事件的结果 (**终点事件, 纵坐标**) 和出现这一结果**所经历的时间 (横坐标)** 结合起来分析的一种统计分析方法。
- 生存分析不同于其它多因素分析的主要区别点就是生存分析考虑了每个观测出现某一结局的时间长短。

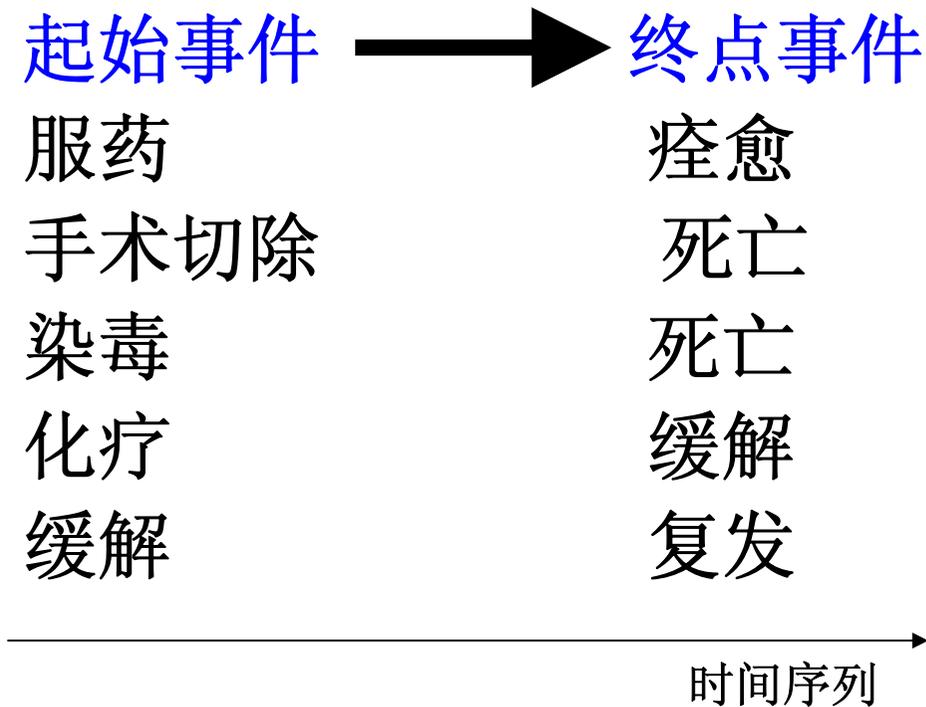
生存分析基本概念

一、生存时间

(survival time, failure time)

- 终点事件与起始事件之间的时间间隔。
- 终点事件指研究者所关心的特定结局。
- 起始事件是反映研究对象生存过程的起始特征的事件。

生存时间举例



生存时间的类型

1. 完全数据 (complete data)

从起点至死亡（死于所研究疾病）所经历的时间。

2. 截尾数据 (删失数据, censored data)

从起点至截尾点所经历的时间。

截尾原因：失访、死于其它疾病、观察结束时病人尚存活等。

例如，某肿瘤医院调查了1991-1995年间经手术治疗的大肠癌患者150例，对可能影响大肠癌术后生存时间的因素进行了调查，如性别、年龄、组织学分类、肿瘤大小、Dure's分期等。随访截止日期为2000年12月30日，随访记录见下表。

编号	性别	年龄	...	手术日期	随访终止日期	随访结局	生存时间(天)
1	男	45	...	1991.05.20	1995.06.04	死亡	1476
2	男	50	...	1992.01.12	1998.08.25	死亡	2417
3	女	36	...	1991.10.24	1994.03.18	失访	876 ⁺
4	男	52	...	1994.11.02	2000.12.30	存活	2250 ⁺
5	女	56	...	1994.06.25	1995.03.17	死亡	265
6	女	60	...	1993.12.05	1996.08.16	死于其它	985 ⁺
...							

生存时间

- 生存时间的度量单位可以是年、月、日、小时等。常用符号 t 表示，截尾数据在其右上角标记“+”。
- 生存资料的主要特点：
 - 含有截尾数据。
 - 截尾数据的特点：真实的生存时间未知，只知道比观察到的截尾生存时间要长。
 - 生存时间的分布一般不呈正态分布。

条件生存概率、生存率、生存曲线

- 条件生存概率：（**conditional probability of survival**）
表示某单位时段开始时存活的个体，到该时段结束时仍存活的可能性。
年条件生存概率表示年初尚存人口存活满**1**年的可能性。

$$p = \frac{\text{活满一年例数}}{\text{年初观察例数}}$$

- 生存率:

(**survival rate, survival function**)

指观察对象经历t个单位时段后仍存活的可能性。

$$3\text{年生存率} = \frac{\text{活满3年例数}}{\text{期初观察例数}}$$

$$5\text{年生存率} = \frac{\text{活满5年例数}}{\text{期初观察例数}}$$

- 生存曲线 (survival curve)

以观察（随访）时间为横轴，以生存率为纵轴，将各个时间点所对应的生存率连接在一起的曲线图。

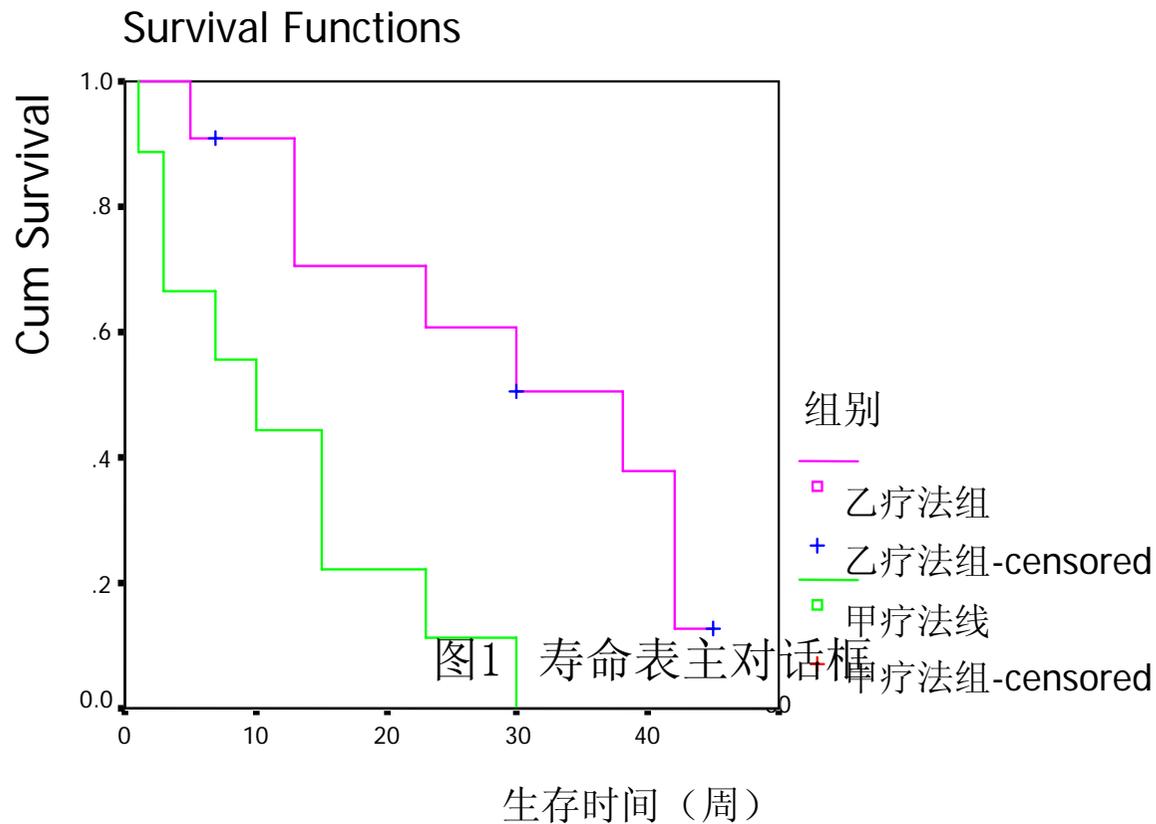
- 生存曲线是一条下降的曲线，分析时应注意曲线的高度和下降的坡度。平缓的生存曲线表示高生存率或较长生存期，陡峭的生存曲线表示低生存率或较短生存期。

生存曲线示例

某医师收集20例脑瘤患者甲、乙两疗法，治疗的生存时间（周）如下：

甲疗法组 1 3 3 7 10 15 15 23 30

乙疗法组 5 7+ 13 13 23 30 30+ 38 42 42 45+



SPSS中生存分析操作

寿命表法（Life Tables）在SPSS中操作

案例操作

某临床试验对20名第III或第IV期黑色素瘤患者进行随访研究，截至研究期结束，记录的生存资料见表1。试计算100周生存率。

12.8	15.6	24.0+	26.4	29.2	30.8+	39.2	42.0	58.4+	72.0+
77.2	82.4	87.2+	94.4+	97.2+	106.0+	114.8+	117.2+	140.0+	168.0+

注：数据后跟符号“+”表示该数据为删失数据。



- 选入变量“Status”后，【Define Event】钮被激活变黑，击该按钮，弹出定义终结事件标记值的对话框

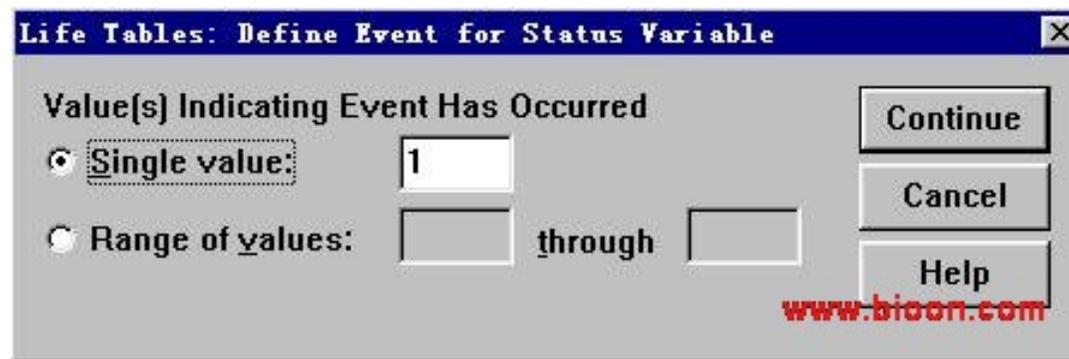


图2 定义终结事件标记值的对话框

- 击Options按钮，弹出选项对话框。



Kaplan-Meier过程

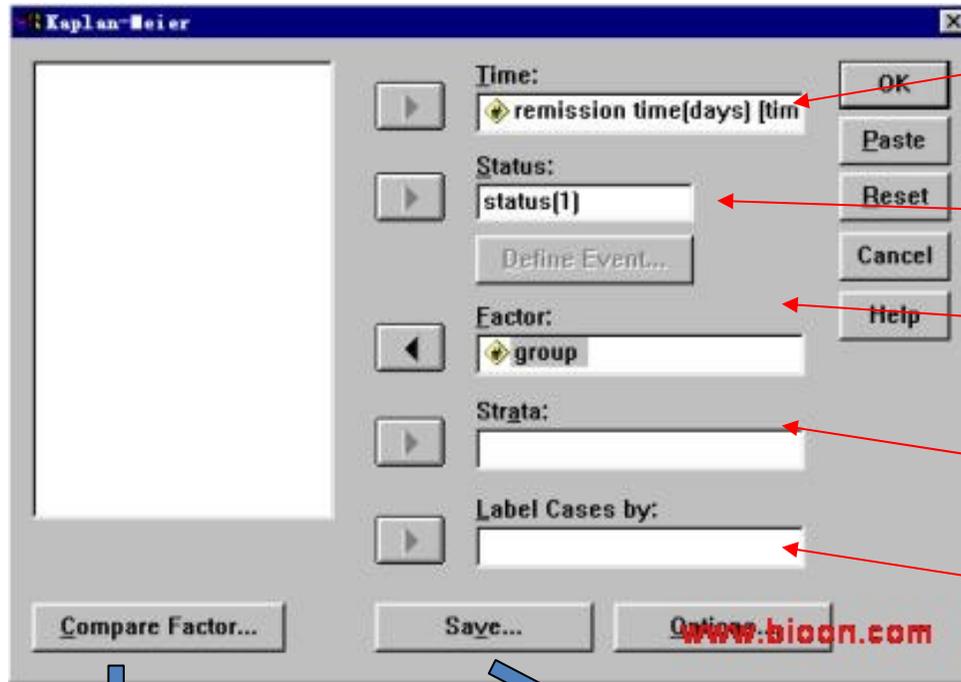
3种疗法治疗66例白血病患者的缓解时间（天）

A 疗法
4,5,9,10,11,12,13,28,28,28,29,31,32,37,41,41,57,62,74,100,139,20+,258+,269,
B疗法
8,10,10,12,14,20,48,70,75,99,103,162,169,195,220,161+,199+,217+,245+
C疗法
8,10,11,23,25,28,28,31,31,40,48,89,124,143,12+,159+,190+,196+,197+,205+,219+

注：数据后跟符号“+”表示该数据为删失数据。

Kaplan-Meier法主对话框

Analyze → Survival → Life Tables



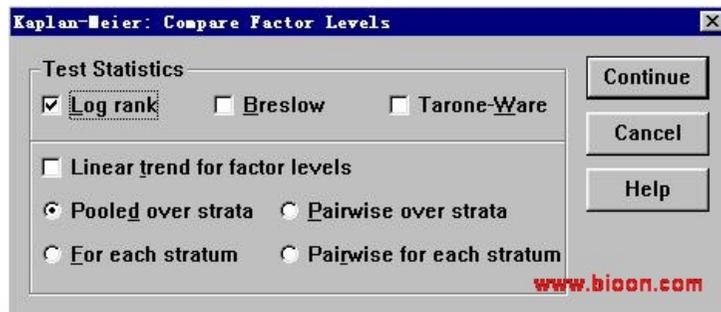
生存时间变量

生存状态变量

分组变量

分层变量

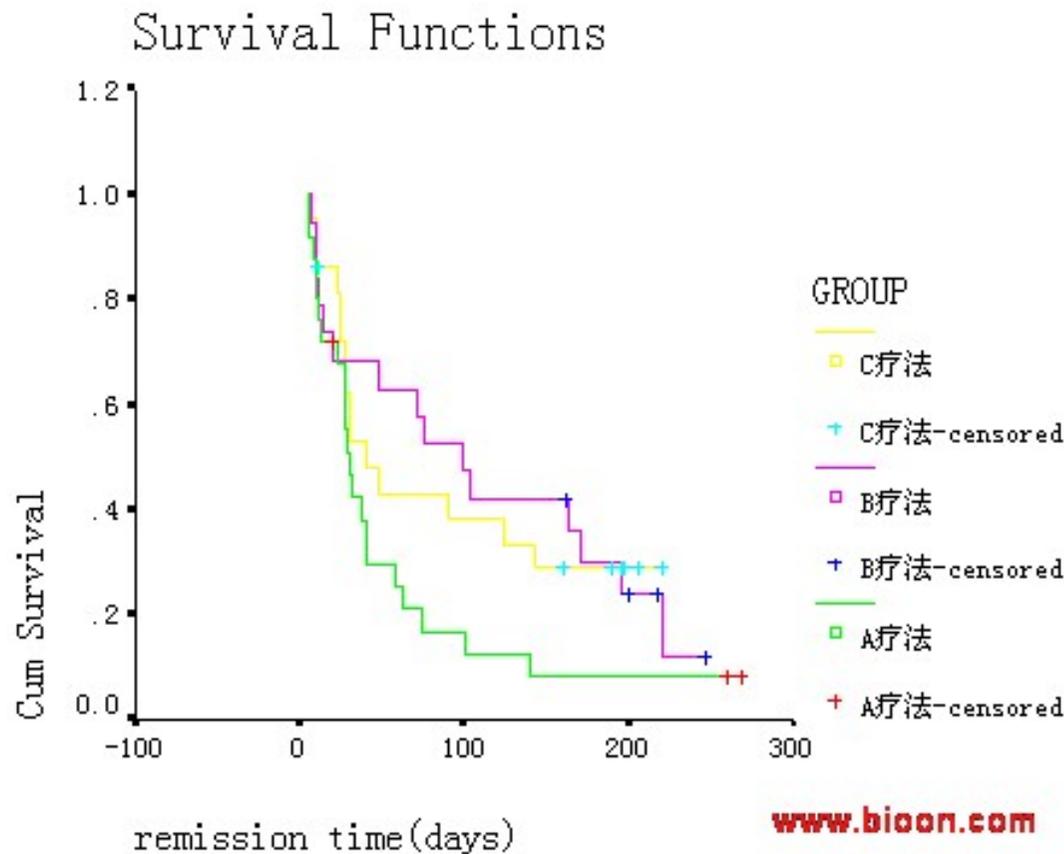
个体标记



Test Statistics for Equality of Survival Distributions for GROUP

	Statistic	df	Significance
Log Rank	4.31	2	0.1158
Breslow	3.67	2	0.1595
Tarone-Ware	4.35	2	0.1137

3种疗法的生存时间差异无显著性意义。进一步两两比较，同样没有显著性差异！



Cox Regression过程

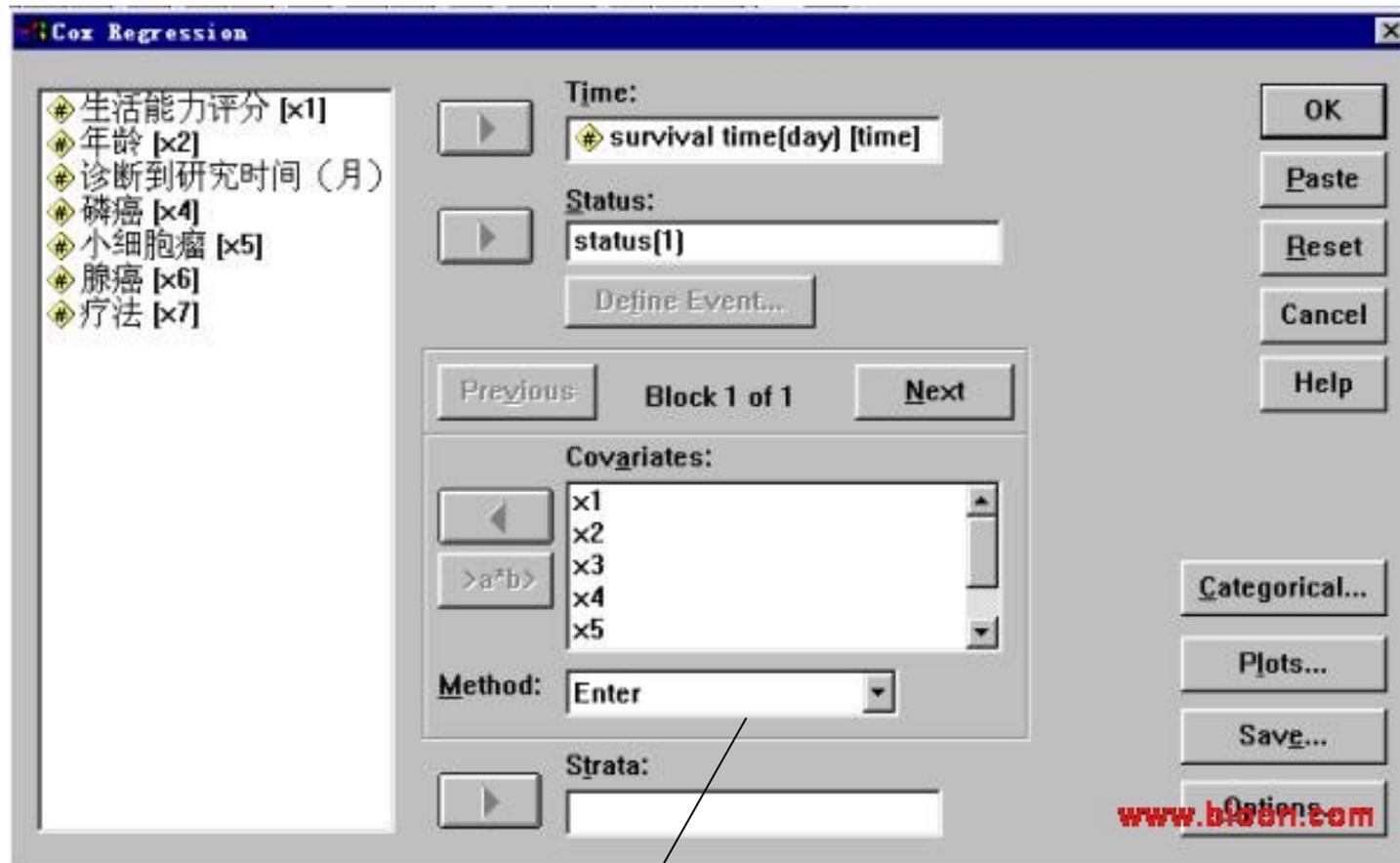
用以描述多个变量对生存时间的影响。此时可控制一个或几个因素，考察其他因素对生存时间的影响，及各因素之间的交互作用

40名肺癌患者的生存资料（详见胡克震主编的《医学随访统计方法》1993，77页）

生存时间	状态	生活能力评分	年龄	诊断到研究时间	鳞癌	小细胞癌	腺癌	疗法	癌症类别
411	1	70	64	5	1	0	0	1	1.00
126	1	60	63	9	1	0	0	1	1.00
118	1	70	65	11	1	0	0	1	1.00

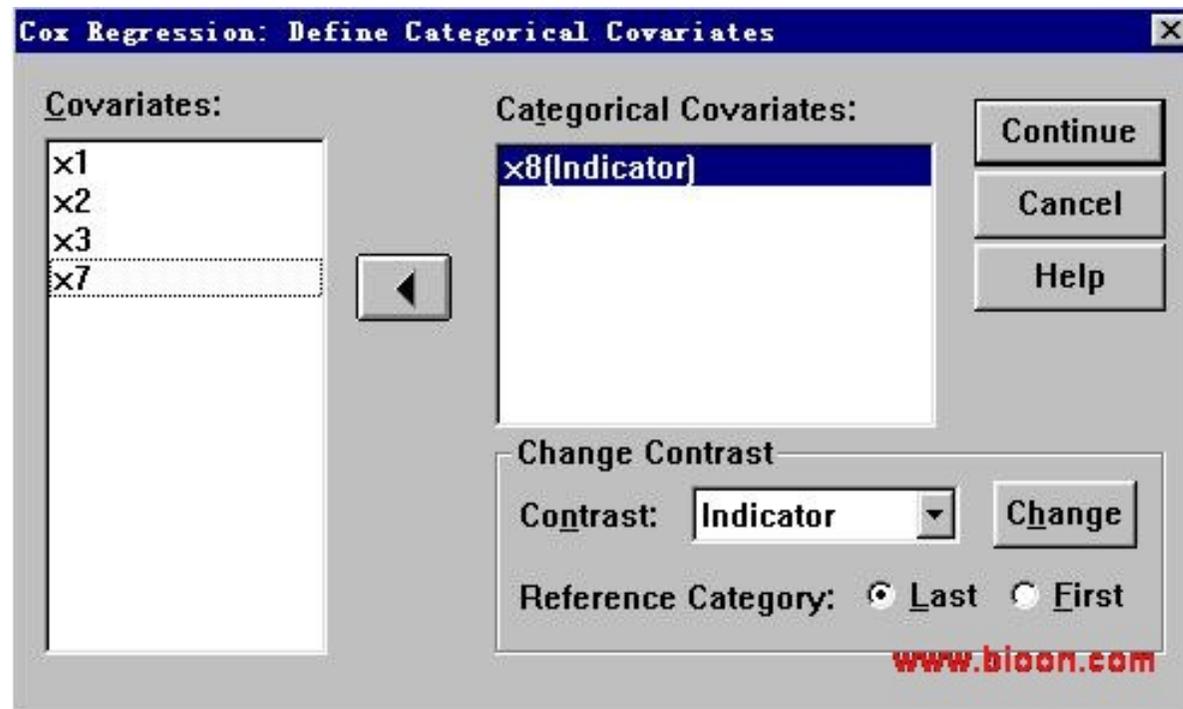
注：原数据库是用亚变量定义肺癌分类：0, 0, 0为其它癌；1, 0, 0为鳞癌；0, 1, 0为小细胞癌；0, 0, 1为腺癌。表中的最后一个变量是我加上去的癌症类别，1为鳞癌；2为小细胞癌；3为腺癌；4为其它癌。实践表明结果与用亚变量计算一样。

Cox回归主对话框

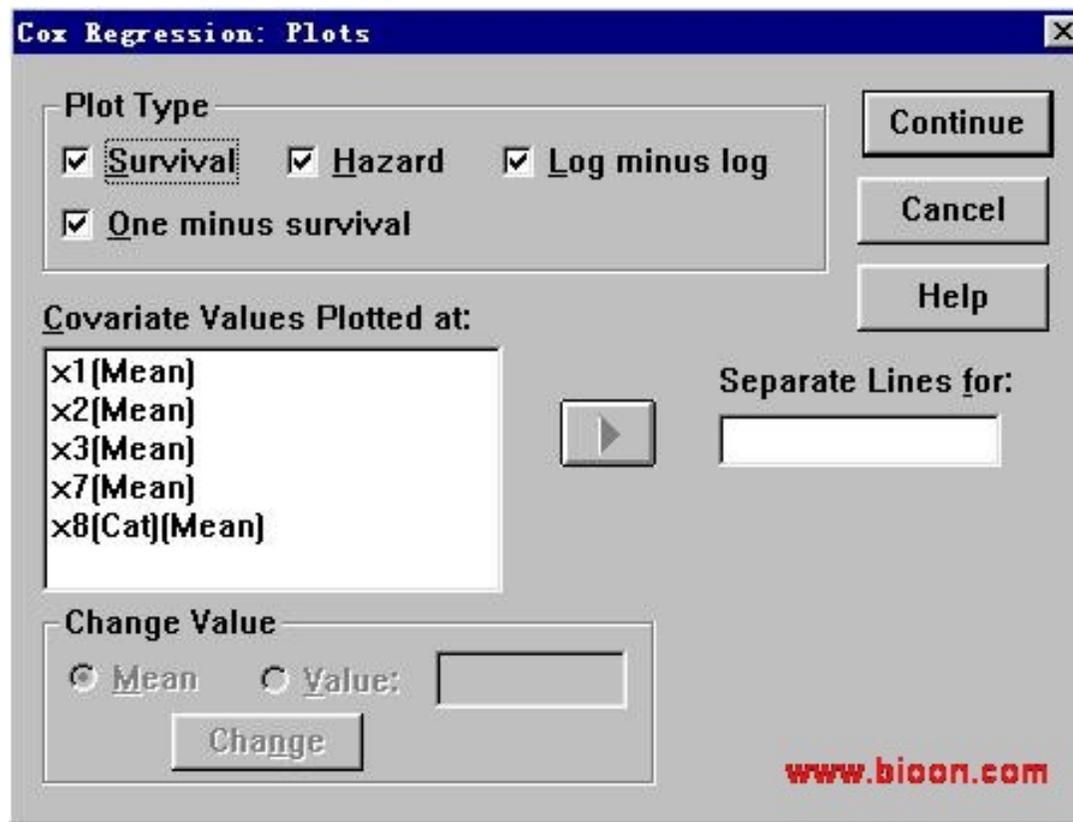


Enter: **Covariates**框内的全部变量均进入回归模型

- 选入自变量后，categorical按钮被激活。按categorical按钮，进入确定分类变量的对话框



- 【Plots】选项



Analyze==>Survival ==>Cox regression

Case Processing Summary

		N	Percent
Cases available in analysis	Event ^a	37	92.5%
	Censored	3	7.5%
	Total	40	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with non-positive time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		40	100.0%

a. Dependent Variable: survival time(day)

www.bioon.com

上表输出总例数、删失例数、失访例数。

Categorical Variable Codings^{a,b}

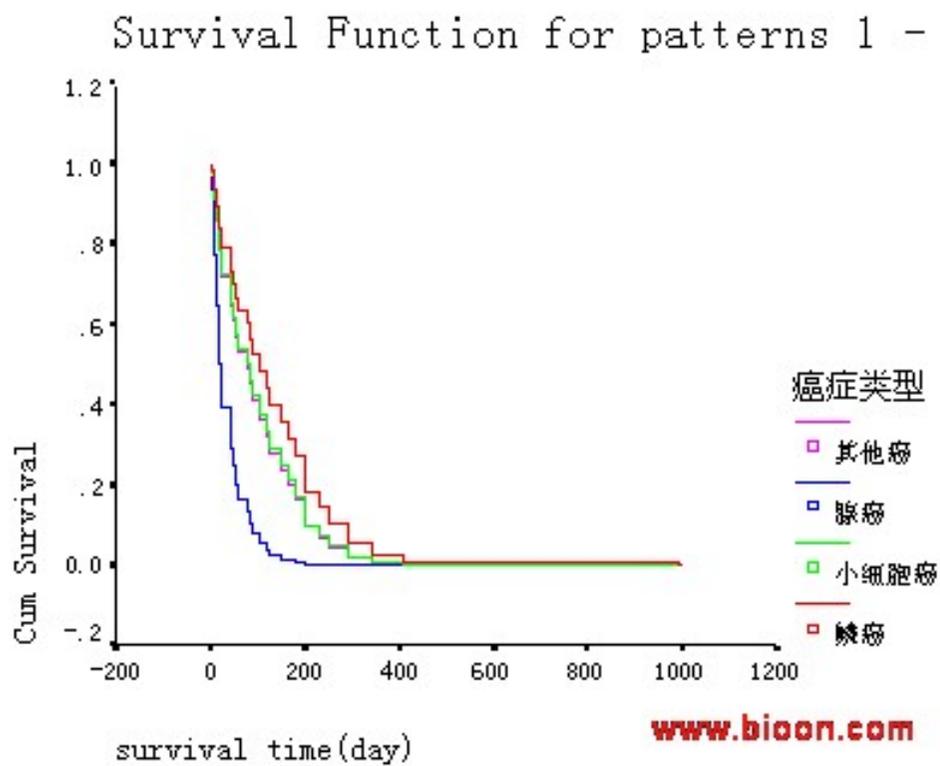
		Frequency	(1)	(2)	(3)
X8	1=鳞癌	14	1.000	.000	.000
	2=小细胞癌	11	.000	1.000	.000
	3=腺癌	5	.000	.000	1.000
	4=其他癌	10	.000	.000	.000

a. Indicator Parameter Coding

b. Category variable: X8 (癌症类型)

www.bioon.com

输出各种癌症的频数及系统所赋的亚变量x81、x82、x83值，当癌症类型是鳞癌时，x81取值为1，其它亚变量取值为0，依此类推。



输出各种癌症的累积生存函数曲线。

生存资料统计学处理方法

(一) 设计

- 目的：专业目的：据专业知识确定。

统计学目的：

- **估计**：根据样本生存资料估计总体生存率及其它有关指标（如中位生存期等），如根据脑瘤患者治疗后的生存时间资料，估计不同时间的生存率、生存曲线以及中位生存期等。
- **比较**：对不同处理组生存率进行比较，如比较不同疗法治疗脑瘤的生存率，以了解哪种治疗方案较优。

- **影响因素分析**：目的是为了探索和了解影响生存时间长短的因素，或平衡某些因素影响后，研究某个或某些因素对生存率的影响。如为改善脑瘤病人的预后，应了解影响病人预后的主要因素，包括病人的年龄、性别、病程、肿瘤分期、治疗方案等。
- **预测**：具有不同因素水平的个体生存预测，如根据脑瘤病人的年龄、性别、病程、肿瘤分期、治疗方案等预测该病人 t 年（月）生存率。

- 方法： 前瞻性队列研究
回顾性队列研究
- 确定起始事件、终点事件、随访终止日期、生存时间、截尾。
- 确定可能的影响因素、水平以及量化方法。

大肠癌影响因素量化表

变量名	因素	量化值	病例数	构成比 (%)
X1	性别	女=0	32	48.5
		男=1	34	51.5
X2	年龄 (岁)	<40=1	10	15.1
		40-60=2	37	56.1
		≥60=3	19	28.8
X3	组织学分类	乳头状腺癌=0	20	30.3
		管状腺癌=1	46	69.7
X4	肿瘤大小 (cm)	≤6=0	29	43.9
		>6=1	37	56.1

X5	细胞增殖抗原 (PCNA)			
		<55%=0	27	40.9
		≥55%=1	39	59.1
X6	淋巴管浸润	无=0	44	66.7
		有=1	22	33.3
X7	血管浸润	无=0	58	87.9
		有=1	8	12.1
Time	手术到观察结束	实际天数		
Event	结束时是否死亡	未死=0	22	33.3
		死亡=1	44	66.7

转换为 大肠癌生存资料

序号	X1	X2	X3	X4	X5	X6	X7	Time	Event
1	0	2	1	0	0	1	0	2896	0
2	0	2	1	0	0	1	0	992	1
3	1	2	1	1	0	0	0	2811	0
4	0	3	1	1	0	1	0	2052	1
5	1	2	0	1	0	0	0	2975	0
6	1	2	1	1	1	1	1	856	1
...									
65	0	1	1	1	1	0	0	584	1

分析资料

- 估计生存率： **Kaplan- Meier**法，寿命表法
- 比较（组间生存率比较）： **log-rank**检验
- 影响因素分析： **Cox**比例风险回归模型（**Cox**回归模型），是生存分析中最重要的模型之一。
- 预测（多因素分析）： **Cox**回归模型预测生存率。

生存率的估计与生存曲线

一、小样本资料生存率及其标准误的计算

- 生存率的计算：**Kaplan-Meier法**（**K-M法**、**乘积极限法**）

Kaplan-Meier法由**Kaplan**和**Meier**于**1958**年提出，直接用概率乘法定理估计生存率，故称乘积极限法（**product-limit method**），是一种非参数法，适用于小样本和大样本。

- 生存率的标准误的计算
- 生存曲线

二、大样本资料的生存分析

- 生存率的计算
寿命表法
- 生存曲线

生存曲线的log-rank检验

log-rank检验（对数秩检验、时序检验）

- 该检验属非参数检验，用于比较两组或多组生存曲线或生存时间是否相同。
- 检验统计量为卡方。
- 自由度=组数-1。
- $P \leq 0.05$ ，两组或多组生存曲线不同。
 $P > 0.05$ ，两组或多组生存曲线差别无统计学意义。

- **log-rank**检验用于整条生存曲线的比较，若比较两组某时间点处的生存率，则按下式计算：

$$u = \frac{S_1(t) - S_2(t)}{\sqrt{SE^2[S_1(t)] + SE^2[S_2(t)]}}$$

$$\alpha' = \alpha / k$$

- 如比较多个时间点处生存率，检验水准应取 **Bonferroni**校正，即 α ，其中k为比较的次数，以保证总的I型错误概率不超过 。

Cox比例风险回归模型

- **Cox比例风险回归模型**（**Cox's proportional hazards regression model**），简称**Cox回归模型**。
- 该模型由英国统计学家D.R.Cox于1972年提出，主要用于肿瘤和其它慢性病的预后分析，也可用于队列研究的病因探索。其优点：
 - 多因素分析方法
 - 不考虑生存时间分布
 - 利用截尾数据

一、Cox模型的基本形式

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

$h(t, X)$ — t 时刻风险函数、风险率或瞬时死亡率
(hazard function)。

$h_0(t)$ —基准风险函数，即所有变量都取0时 t 时刻风险函数。

X_1, X_2, \dots, X_p —协变量、影响因素、预后因素。

$\beta_1, \beta_2, \dots, \beta_p$ —回归系数。

$$RR = e^{\beta}$$

- $\beta > 0$, $RR > 1$, 说明变量X增加时, 危险率增加, 即X是危险因素。
- $\beta < 0$, $RR < 1$, 说明变量X增加时, 危险率下降, 即X是保护因素。
- $\beta = 0$, $RR = 1$, 说明变量X增加时, 危险率不变, 即X是危险无关因素。

Cox回归与多元线性回归、logistic回归的比较

	多元线性回归	logistic回归	Cox回归
数据类型	Y数值变量	Y分类变量	Y二分类变量+时间
	X数值变量、分类变量、等级变量		
模型结构			
变量筛选	前进法；后退法；逐步法		
参数估计	最小二乘法	最大似然法	最大似然法
参数检验	F-test t-test	似然比检验 Wald检验 score检验	似然比检验 Wald检验 score检验
参数解释	回归系数b	优势比OR	RR
样本含量	至少变量数的10倍	至少变量数的20倍	非截尾例数至少变量数的10倍
应用	因素分析 预测预报 Y	因素分析 预测、判别P(Y=1)	因素分析 生存预测 S(t)
方程	$\hat{y} = a + bx$	$p = p(y=1 x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$	$h(t, x) = h_0(t) \exp(\beta x)$

案例分析

某研究者观察了确诊后采取同样方案进行化疗的**26**例急性混合型白血病患者，欲了解某种不良染色体是否会影响患者病情的缓解，将治疗后**120**天内症状是否缓解作为结果变量（缓解=0；未缓解=1），有无不良染色体作为研究因素。整理资料见下表。

有无不良染色体与缓解的关系

不良染色体	缓解	未缓解	合计	缓解率 (%)
有	5	13	18	27.8
无	3	5	8	37.5
合计	8	18	26	30.8

- 考虑到例数较小，采用Fisher确切概率法，得到 $P=0.667$ ，尚不能认为不良染色体影响病情的缓解。

- 这种情况下的结果并不可靠，原因是两个比较组之间其它影响患者病情缓解的因素不一定均衡，因而需要考虑平衡其它可能的影响因素如年龄（岁）、骨髓原幼细胞分组（ $\geq 50\% = 1$ ； $< 50\% = 0$ ）、CD34表达（阳性=1；阴性=0）、性别（男=1；女=0）的作用。

多因素logistic回归分析结果

因素	回归系数	Wald卡方	P	OR
染色体	1.457	1.161	0.281	4.29
骨髓原幼 细胞分组	2.961	4.778	0.029	19.2

- 采用多因素logistic回归分析，经逐步法按0.10水准，平衡骨髓原幼细胞分组后，有无不良染色体不影响患者的缓解（ $P=0.281$ ）。

- 有一位临床医生指出，仅考虑是否缓解还不够，如果进一步利用缓解时间的长短来进行分析，信息利用得更充分。
- 费了很大辛苦，幸好查到了所有患者的缓解时间，于是采用log-rank检验比较有无不良染色体两组病人的生存曲线，得卡方**1.28**，**P=0.2579**，仍然显示患者的缓解时间与不良染色体无关。

- 生存时间的比较仍然需要考虑组间的可比性，经多因素Cox回归分析，当检验水准0.10时的结果如下。

因素	回归系数	卡方	P	RR
染色体	1.838	3.709	0.054	6.29
CD34	1.877	8.904	0.003	6.54
骨髓原幼 细胞分组	3.205	8.838	0.003	24.4

这才是最佳的统计分析手段，可以防止出现的各种假象

第十节 非参数分析

具体操作，以及适用的范围与参数分析十分类似，因此，只需要掌握何种情况采用何种合适的方法就可以了。学习时一定要注意与参数分析进行对比，自然就一通百通了。

----- www.MedSci.cn

参数统计

(parametric statistics)

↓
计量资料，呈正态分布

↓
依赖于特定分布类型，比较的是参数

非参数统计

(nonparametric statistics)

↓
对总体的分布类型不作任何要求

↓
不受总体参数的影响，比较分布或分布位置

↓
适用范围广；可用于任何类型资料(等级资料，或“>50mg”)

↓
对于符合参数统计分析条件者，采用非参数统计分析，其检验效能较低

参数统计与非参数统计原理几乎一致，适用有别

	参数统计	非参数统计
频率分析	Frequency	卡方Chi-Square
分布情况分析	正态分布统计	Runs（游程检验）
单组	平均数	1-Sample K-S
两组（独立）	Independent-Samples T Test	2 Independent Samples (如Wilcoxon-Mann-Whitney U 检验)
两组（配对）	Paired-Samples T Test	2 Related Samples (如Wilcoxon signed)
多组	One-way ANOVA	k Independent Samples（Kruskal-Wallis H检验）
多组多变量	方差分析General Factorial ANOVA	K Related Samples(如Friedman等)

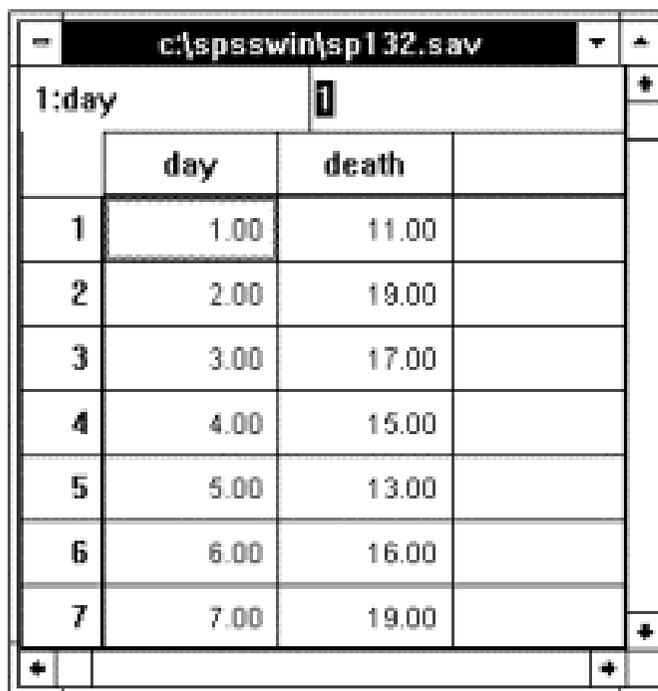
Chi-Square过程

- 卡方检验适用于配合度检验，主要用于分析实际频数与某理论频数是否相符。

某地一周内各日死亡数的分布如下表，请检验一周内各日的死亡危险性是否相同？

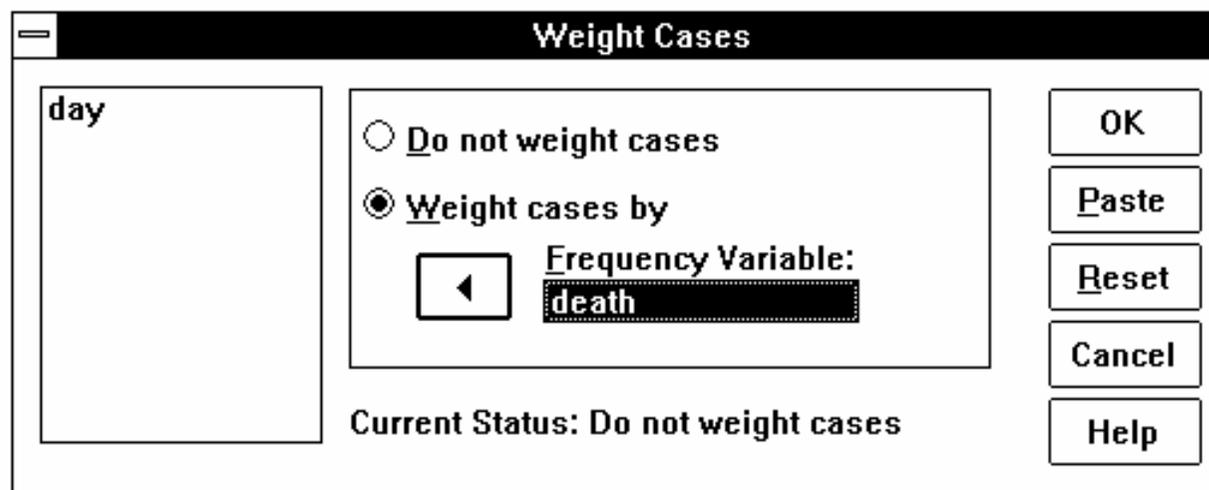
周 日	死亡数
一	11
二	19
三	17
四	15
五	15
六	16
日	19

数据录入窗口



	day	death	
1	1.00	11.00	
2	2.00	19.00	
3	3.00	17.00	
4	4.00	15.00	
5	5.00	13.00	
6	6.00	16.00	
7	7.00	19.00	

数据加权对话框



Weight Cases

day

Do not weight cases

Weight cases by

Frequency Variable: death

Current Status: Do not weight cases

OK

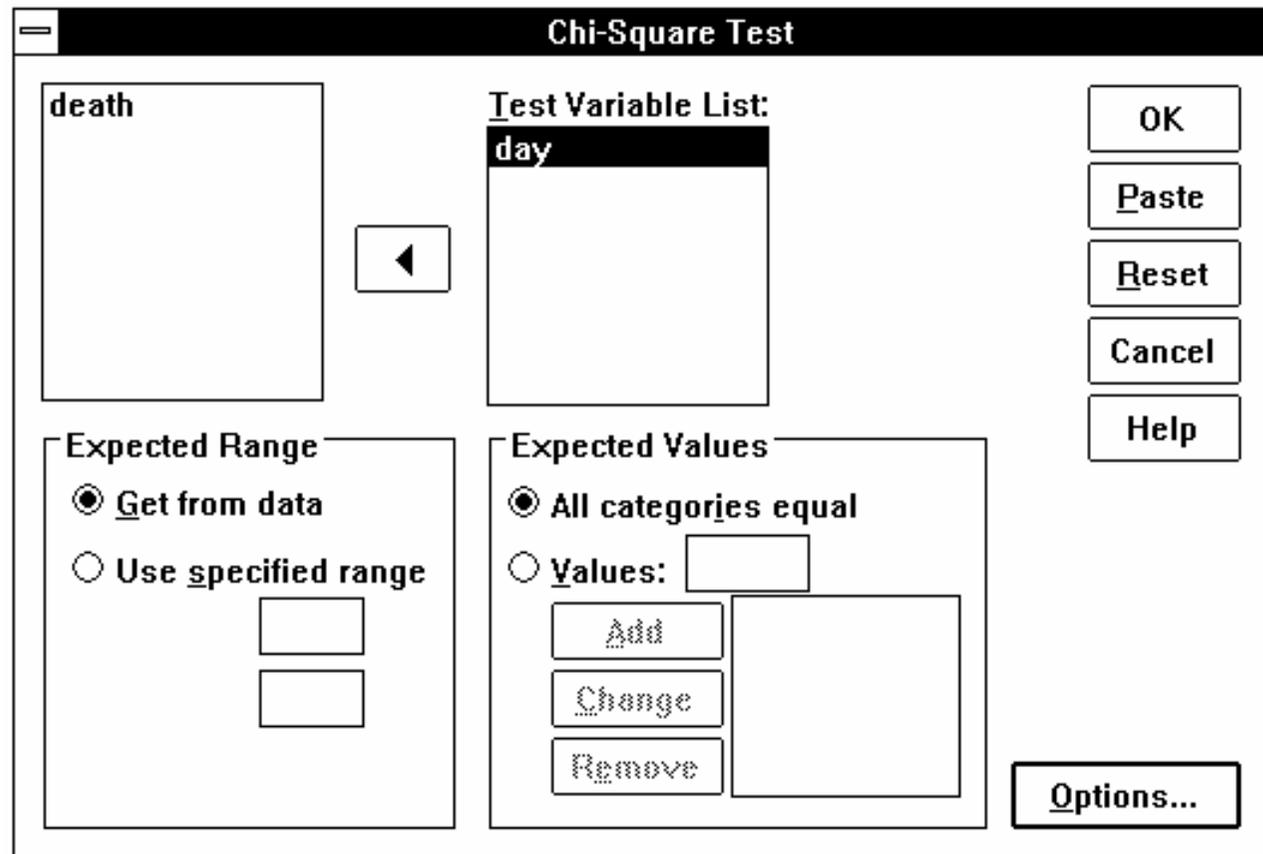
Paste

Reset

Cancel

Help

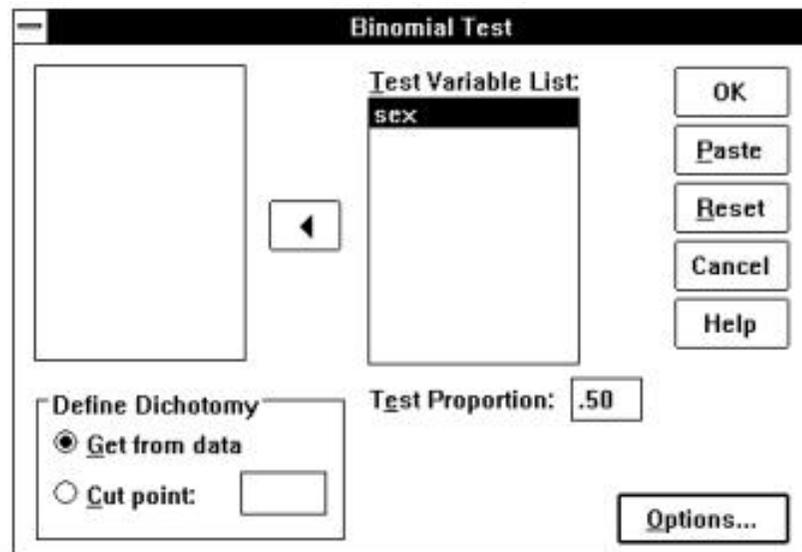
激活Statistics菜单选Nonparametric Tests中的Chi-Square...命令项，弹出Chi-Square Test对话框（图13.3）。现欲对一周内各日的死亡数进行分布分析，故在对话框左侧的变量列表中选day，点击按钮使之进入Test Variable List框，点击OK按钮即可。



Binomial过程（二项分布分析）

- 某地某一时期内出生40名婴儿，其中女性12名（定Sex=0），男性28名（定Sex=1）。问这个地方出生婴儿的性比例与通常的男女性比例（总体概率约为0.5）是否不同？

激活Statistics菜单选Nonparametric Tests中的Binomial Test...命令项，弹出Binomial Test对话框（图13.4）。在对话框左侧的变量列表中选sex，点击按钮使之进入Test Variable List框，在Test Proportion框中键入0.50，再点击OK

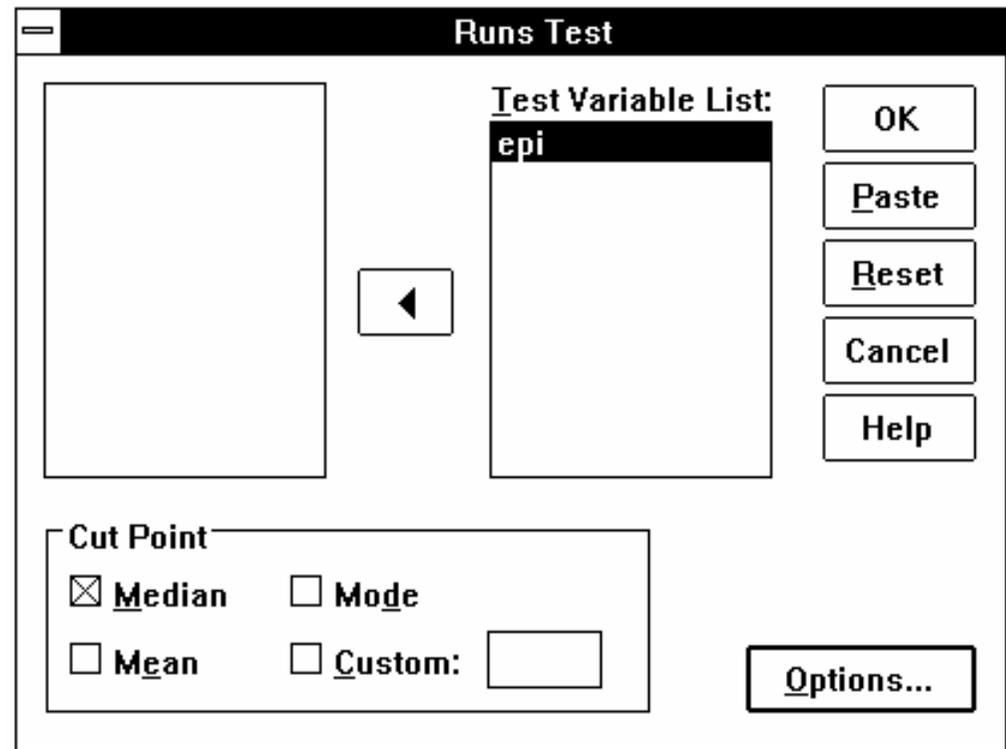


Runs（游程检验）过程

- 某村发生一种地方病，其住户沿一条河排列，调查时对发病的住户标记为“1”，对非发病的住户标记为“0”，共17户：
- 0 1 1 0 0 0 1 0 0 1 0 0 0 0 1 1 0 0 1 0 0 0 0 1 0 1
- 问病户的分布排列是呈聚集趋势，还是随机分布？

激活数据管理窗口，定义住户变量为epi。按住户顺序输入数据，发病的住户为1，非发病的住户为0。

激活Statistics菜单选Nonparametric Tests中的Runs Test...项，弹出Runs Test对话框（图13.5）。在对话框左侧的变量列表中选epi，点击按钮使之进入Test Variable List框。在临界割点Cut Point框中有四个选项：



1-Sample K-S过程

- 对单样本进行Kolmogorov-Smirnov Z检验，它将一个变量的实际频数分布与正态分布（Normal）、均匀分布（Uniform）、泊松分布（Poisson）进行比较

某地正常成年男子144人红细胞计数（万/立方毫米）的频数资料如下，问该资料的频数是否呈正态分布？

红细胞计数	人数	红细胞计数	人数
420-	2	540-	24
440-	4	560-	22
460-	7	580-	16
480-	16	600-	2
500-	20	620-	6
520-	25	640-	1

2 Independent Samples过程

调查某厂的铅作业工人7人和非铅作业工人10人的血铅值 ($\mu\text{g} / 100\text{g}$) 如下，问两组工人的血铅值有无差别？

非铅作业组	5	5	6	7	9	12	13	15	18	21
铅作业组		17	18	20	25	34	43	44		

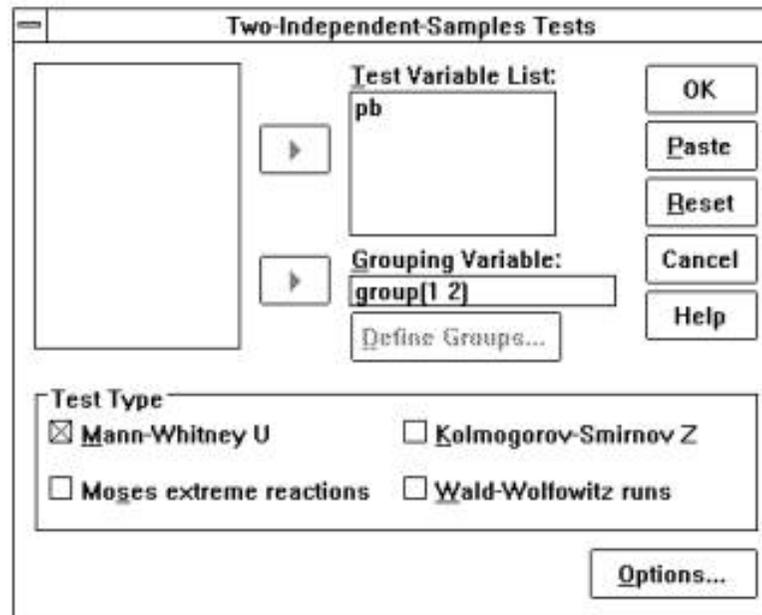


图13.7 两独立样本检验对话框

Mann-Whitney U: 主要用于判别两个独立样本所属的总体是否有相同的分布;
 Kolmogorov-Smirnov Z: 推测两个样本是否来自具有相同分布的总体;

k Independent Samples过程

(中位数检验和Kruskal-Wallis H检验)

随机抽样得以下三组人的血浆总皮质醇测定值 ($\mu\text{g} / \text{L}$)，试比较有无差异？

正常人	单纯性肥胖	皮质醇增多症
0.4	0.6	9.8
1.9	1.2	10.2
2.2	2.0	10.6
2.5	2.4	13.0
2.8	3.1	14.0
3.1	4.1	14.8
3.7	5.0	15.6
3.9	5.9	15.6
4.6	7.4	21.6
7.0	13.6	24.0

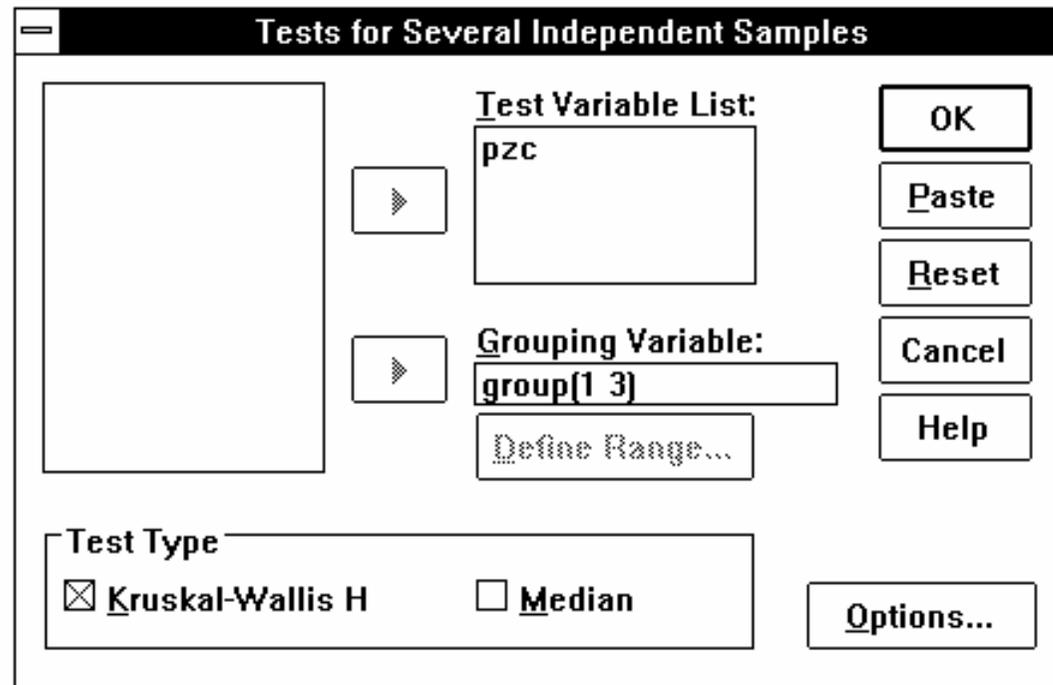


图13.8 多样本资料的秩和检验对话框

2 Related Samples过程

研究饲料中缺乏Vit E对大鼠肝中Vit A含量的关系，将大鼠按性别相同、体重相近的原则配成8对，并将每对大鼠随机分为2组（正常饲料组、Vit E缺乏饲料组），一定时间后杀死大鼠，测定肝中Vit A含量，结果如下表，问：饲料中缺乏Vit E对大鼠肝中Vit A含量有无影响？

大鼠对别	正常饲料组	Vit E 缺乏饲料组
1	37.2	25.7
2	20.9	25.1
3	31.4	18.8
4	41.4	33.5
5	39.8	34.0
6	39.3	28.3
7	36.1	26.2
8	31.9	18.3

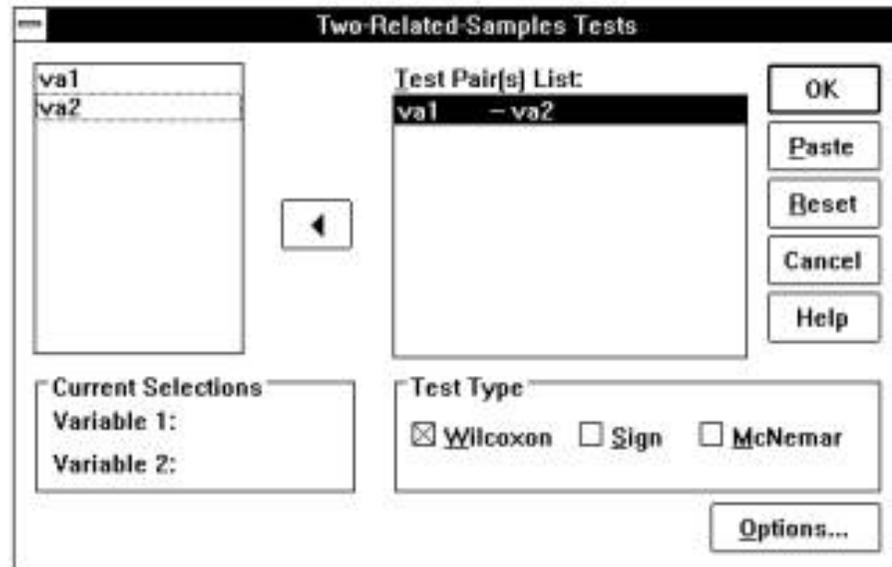


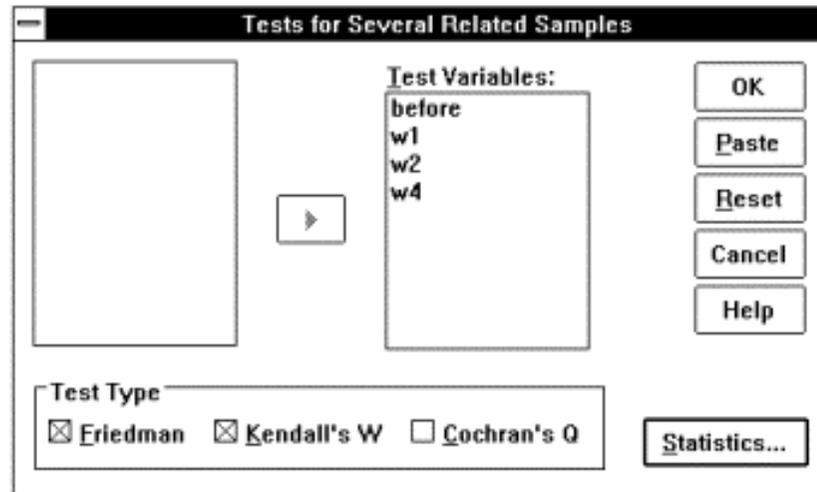
图13.9 两相关样本的秩和检验对话框

- 1、Wilcoxon: 配对符号等级秩次检验,
- 2、Sign: 符号检验;
- 3、McNemar: 以研究对象作自身对照, 检验其“前后”的变化是否显著, 该方法适用于相关的二分变量数据。

K Related Samples过程

- 用某药治疗血吸虫病患者，在治疗前和治疗后一周、二周和四周各测定7名患者血清SGPT值的变化，以观察该药对肝功能的影响，结果如下表，问：患者四个阶段的血清SGPT值有无不同

患者编号	治疗前	治 疗 后		
		一周	二周	四周
1	63	188	138	54
2	90	238	220	144
3	54	300	83	92
4	45	140	213	100
5	54	175	150	36
6	72	300	163	90
7	64	207	185	87



多个相关样本的秩和检验对话框

- 1、Friedman: 双向方差分析，考察多个相关样本是否来自同一总体；
- 2、Cochran's Q: 作为两相关样本McNemar检验的多样本推广，特别适用于定性变量和二分字符变量；
- 3、Kendall's W: Kendall和谐系数检验，通过计算Kendall和谐系数W，以检验多个相关样本是否来自同一分布的总体。

联系我们

医学生物学SCI论文编辑网 (MedSci)

- 联系人: 李欣梅博士, 张发宝 博士
- 电话: 021-64087586, 64088675
- 传真: 021-64085875
- **Email:** editing@bioon.com
- 网址: www.medsci.cn

选择我们 脱颖而出

Thank You !

