

医学统计学实用教程 (1)

Statistical Methods in Medicine

www.MedSci.cn 张发宝 博士
2009.3



声 明

- 我们[MedSci团队](http://www.medsci.cn)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。
- 本幻灯片仅仅是第一部分，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>

统计定义

- 是一种对客观现象数量方面进行的调查研究活动；
- 是收集、整理、分析、推断、判断等认识活动的总称。
- 数据汇总仅仅是统计工作的一小部分内容。

统计三个层次：

data collection → data analysis → data mining

—— MedSci 张发宝 博士

工作生活中常见的统计学问题

- 这个药物治疗高血压有效吗？(假设检验)
- 癌症病人能活多久？(生存分析)
- 吸烟,喝酒与冠心病有关吗？(因子分析)
- 肝硬化与肝癌有关吗？(相关分析)
- 子女为什么象父母，其强度有多大？(相关与回归)
- 基因芯片的海量数据如何归类总结？(聚类分析)
- 临床不同的化疗方案，对不同的分期肿瘤病人的效果统计(方差分析)



统计学是对令人困惑费解的数字问题做出设想的艺术。

医学论文中的统计学问题

- 60年代到80年代，国外医学杂志调查结果：有统计错误的论文**20%~72%**。
- 1996年对4586篇论文统计（中华医学会系列杂志占6.9%），数据分析方法误用达**55.7%**。
- 2001年《中华预防医学杂志》：中华医学会系列杂志误用约**54%**（1995篇）。

伪造统计数据违反科学道德

➤ **1976年New Science 杂志关于科研舞弊行为的调查**

- (1) 74%的调查表反映有不正当修改数据的情况**
- (2) 17%拼凑实验结果**
- (3) 7%凭空捏造数据**
- (4) 2%故意曲解结果**

A Warning!

- Fancy statistical methods cannot rescue garbage data
- Fancy statistical methods can help you gain insight into your data, over and above what seems obvious on its face
- You should always worry about whether the sampled results are representative of the population, and whether your sample allows you to make inferences about the population.

统计学是现代医学大厦的一个重要支柱

——美国医学会杂志（JAMA）主编

统计学是挖掘数据背后的真理

—— MedSci.cn

统计学基础概念

统计资料的类型

有三种类型的资料:计量资料,计数资料,等级资料

基本概念: 变量及变量值, 研究者对每个观察单位的某项特征进行观察和测量, 这种特征称为变量, 变量的测得值叫变量值 (也叫观察值), 称为资料。按变量值的性质可将资料分为定量资料和定性资料。

1. 计量资料

定义：通过度量衡的方法，测量每一个观察单位的某项研究指标的量的大小，得到的一系列数据资料。

特点：有度量衡单位
多为连续性资料
(通过测量得到)

如患者的身高 (**cm**)、体重 (**kg**)、
血压 (**mmHg**)、脉搏 (次/分)、
红细胞计数 (**10¹² /L**)



2. 计数资料

- 定义：将全体观测单位按照某种性质或特征分组，然后再分别清点各组观察单位的个数。

- 特点：没有度量衡单位



多为间断性资料（通过枚举或记数得来）
如肤色（黑、白）、血型（**ABO**）、职业（工农兵）、性别（男女）

3. 等级资料

定义： 介于计量资料和计数资料之间的一种资料，通过半定量方法测量得到。

特点：

每一个观察单位没有确切值

各组之间有性质上的差别或程度上的不同。

①癌症分期：Ⅰ、Ⅱ、Ⅲ。

②药物疗效：治愈、好转、无效、死亡。

③尿蛋白：-,±,+,++,+++及以上

（三）资料的转化（变量类型的转化）

数值变量 —————→ 等级资料

等级资料 $\xrightarrow{\text{积分制}}$ 计量资料

- 例如：测得5人的WBC（个/ m^3 ）数如下：

3000 6000 5000 8000 12000 $\longrightarrow$ 数值变量

过低 正常 正常 正常 异常 $\longrightarrow$ 等级变量

- 若按正常3人，异常2人分组 $\rightarrow$ 二分类变量
- 若按过低1人，正常3人，过高1人分组 $\rightarrow$ 等级资料

观察单位
observations

Units; elements

变量
variables

个体 individuals

住院号	年龄	身高	体重	住院天数	职业	文化程度	分娩方式	妊娠结局
2025655	27	165	71.5	5	无	中学	顺产	足月
2025653	22	160	74.0	5	无	小学	助产	足月
2025830	25	158	68.0	6	管理员	大学	顺产	足月
2022543	23	161	69.0	5	无	中学	剖宫产	足月
2022466	25	159	62.0	11	商业	中学	剖宫产	足月
2024535	27	157	68.0	2	无	小学	顺产	早产
2025834	20	158	66.0	4	无	中学	助产	早产
2019464	24	158	70.5	3	无	中学	助产	足月
2025783	29	154	57.0	7	干部	中学	剖宫产	足月

Quantitative data Qualitative data
计量资料 计数资料

统计学中的几个基本概念

- 1、齐性与变异
- 2、总体与样本
- 3、参数与统计量
- 4、误差（标准差，标准误）
- 5、频率

1. 齐性与变异

homogeneity and variation

同质事物个体间的差异。

来源于一些未加控制或无法控制的甚至不明原因的因素。

是统计学存在的基础,从本质上说,统计学就是研究变异的科学。

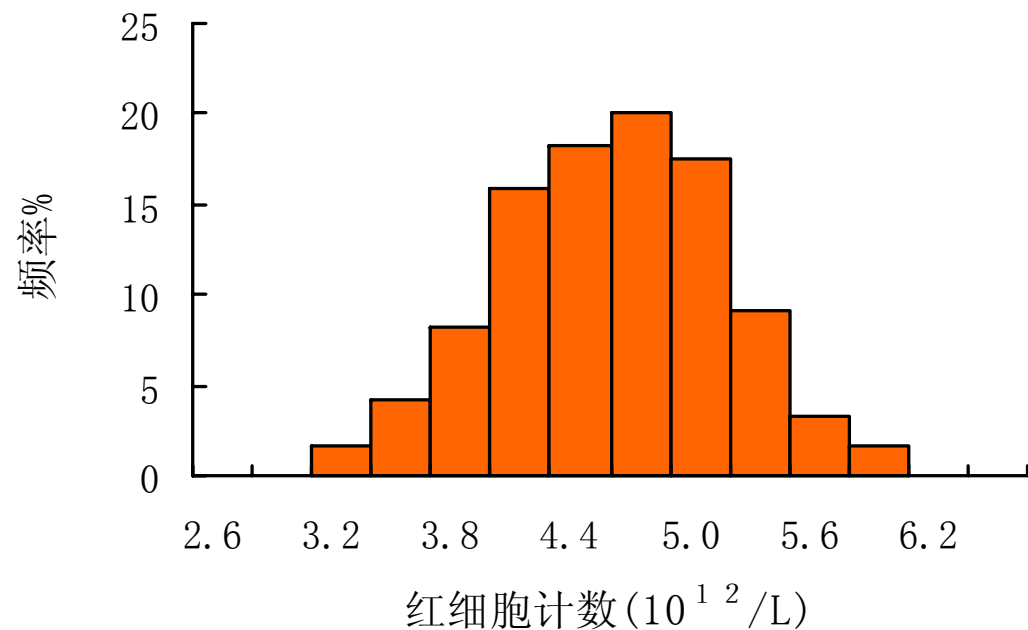


图 1-1 120名正常成年男子细胞计数直方图

表1-1 120名正常成年男子红细胞计数值($10^{12}/L$)

5.12	5.13	4.58	4.31	4.09	4.41	4.33	4.58	4.24	5.45	4.32	4.84
4.91	5.14	5.25	4.89	4.79	4.90	5.09	4.64	5.14	5.46	4.66	4.20
4.21	3.73	5.17	5.79	5.46	4.49	4.85	5.28	4.78	4.32	4.94	5.21
4.68	5.09	4.68	4.91	5.13	5.26	3.84	4.17	4.56	3.52	6.00	4.05
4.92	4.87	4.28	4.46	5.03	5.69	5.25	4.56	5.53	4.58	4.86	4.97
4.70	4.28	4.37	5.33	4.78	4.75	5.39	5.27	4.89	6.18	4.13	5.22
4.44	4.13	4.43	4.02	5.86	5.12	5.36	3.86	4.68	5.48	5.31	4.53
4.83	4.11	3.29	4.18	4.13	4.06	3.42	4.68	4.52	5.19	3.70	5.51
4.64	4.92	4.93	4.90	3.92	5.04	4.70	4.54	3.95	4.40	4.31	3.77
4.16	4.58	5.35	3.71	5.27	4.52	5.21	4.37	4.80	4.75	3.86	5.69

最大值=6.18, 最小值=3.29, 极差=2.89

算术均数=4.72, 标准差=0.57

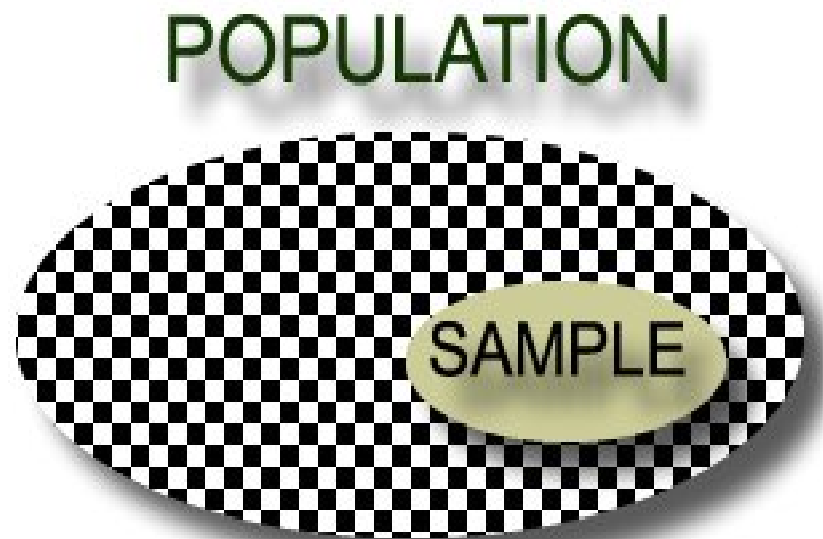
2.总体与样本 (population and sample)

总体：根据研究目的确定的同质研究对象的全体。当研究有具体而明确的指标时，总体是指该项变量值的全体。

样本：从总体中随机抽取的有**代表性**的一部分。

- 观察单位（个体）：最基本的研究单位
- 分为有限总体和无限总体。由于调查总体的不可能性、巨大性和没必要。对其中的一部分对象进行调查--
- 样本（总体与样本的关系。举例。）
- 样本选择的原则--??
- 样本量（sample size）

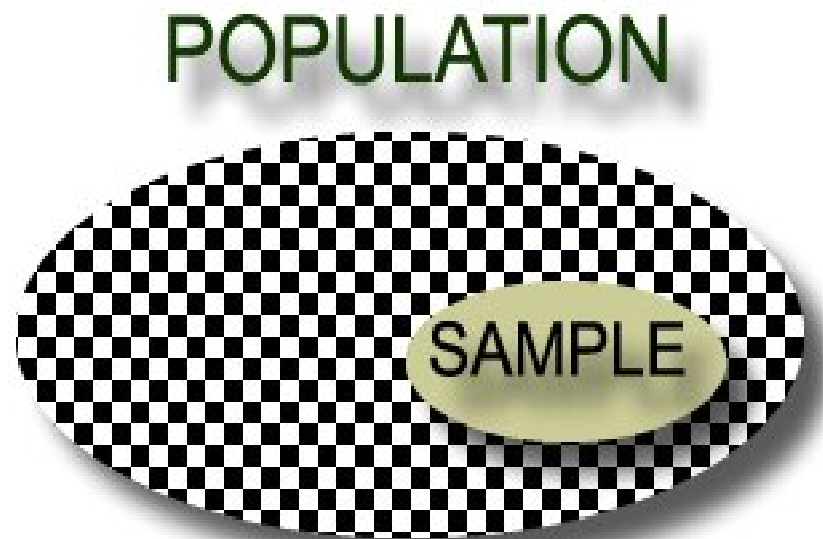
population and sample



总体：根据研究目的确定的**同质**研究对象的**全体**（集合）。分有限总体与无限总体

样本：从总体中随机抽取的部分观察单位

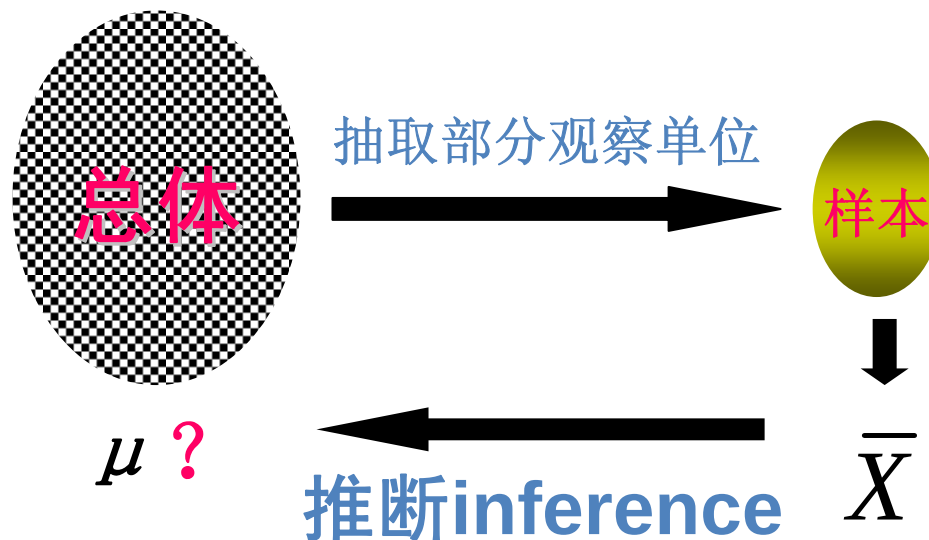
随机抽样 random sampling



为了保证样本的**可靠性**和**代表性**，需要采用随机的抽样方法（在总体中每个个体具有**相同的机会**被抽到）。

但目前几乎没有几个研究是完全按照标准的随机方法进行的！

3. 参数与统计量 parameter and statistic



参数：总体的统计指标，
如总体均数，采用希腊字母
记为 μ 。固定的常数

统计量：样本的统计指标，如样本均数，采用拉丁字母分
别记为 $\bar{X}$ 。统计量是参数附近波动的随机变量。

4. 误差

误差：统计上所说的误差泛指测量值与真值之差，样本指标与总体指标之差。主要有以下二种：系统误差和随机误差（随机测量误差,抽样误差）。

(1)系统误差：指数据搜集和测量过程中由于仪器不准确、标准不规范等原因，造成观察结果呈倾向性的偏大或偏小，这种误差称为系统误差。

特点：具有累加性

(2).随机误差：由于一些非人为的偶然因素使得结果或大或小，是不确定、不可预知的。

特点：随测量次数增加而减小。

5. 概率probability

确定性现象：在一定条件下，**一定**会发生或一定不会发生的现象。其表现结果为两种事件：肯定发生某种结果的叫**必然事件**；肯定不发生某种结果的叫**不可能事件**。

随机现象：在同样条件下**可能**会出现两种或多种结果，究竟会发生哪种结果，事先不能确定。其表现结果称为**随机事件**。随机事件的特征：①**随机性**；②**规律性**：每次发生的可能性的**大小**是确定的。

概率：描述随机事件发生的可能性大小的数值，用大写的 **P** 表示；取值 **$[0, 1]$** 。

统计学常用的方法

Terminology

statistical description	统计描述
statistical inference	统计推断
parameter estimation	参数估计
Frequency distribution	频数分布
frequency table	频数表
arithmetic Mean, average	算术平均值
standard deviation	标准差
variance	方差
range	极差, 全距, 范围
geometric mean	几何平均值
median	中位数
normal distribution	正态分布
reference range	参考值范围

一些常用看到的统计方法

率 u 检验, χ^2 , 确切概率 u , χ^2 , 确切概率似然比 χ^2 , 确切概率

构成比(分布) χ^2

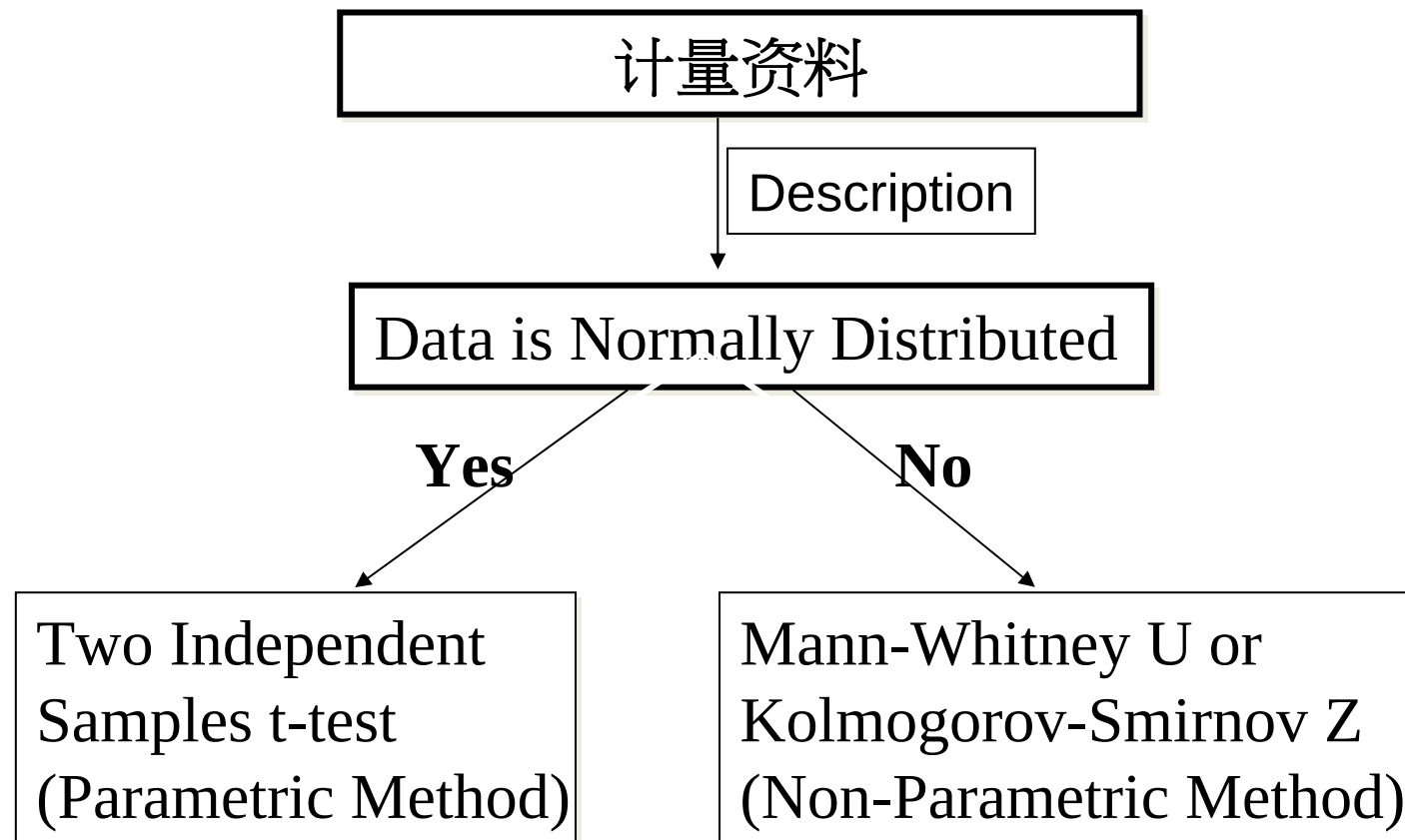
均数 u 检验, t 检验 u 检验, t 检验 方差分析, 两两比较

等级 Wilcoxon, u Wilcoxon, u Kruskal-Wallis, χ^2

方差 F 检验: F Bartlett 方差齐性

第一节 数值型变量资料的统计描述

(也称 参数统计)



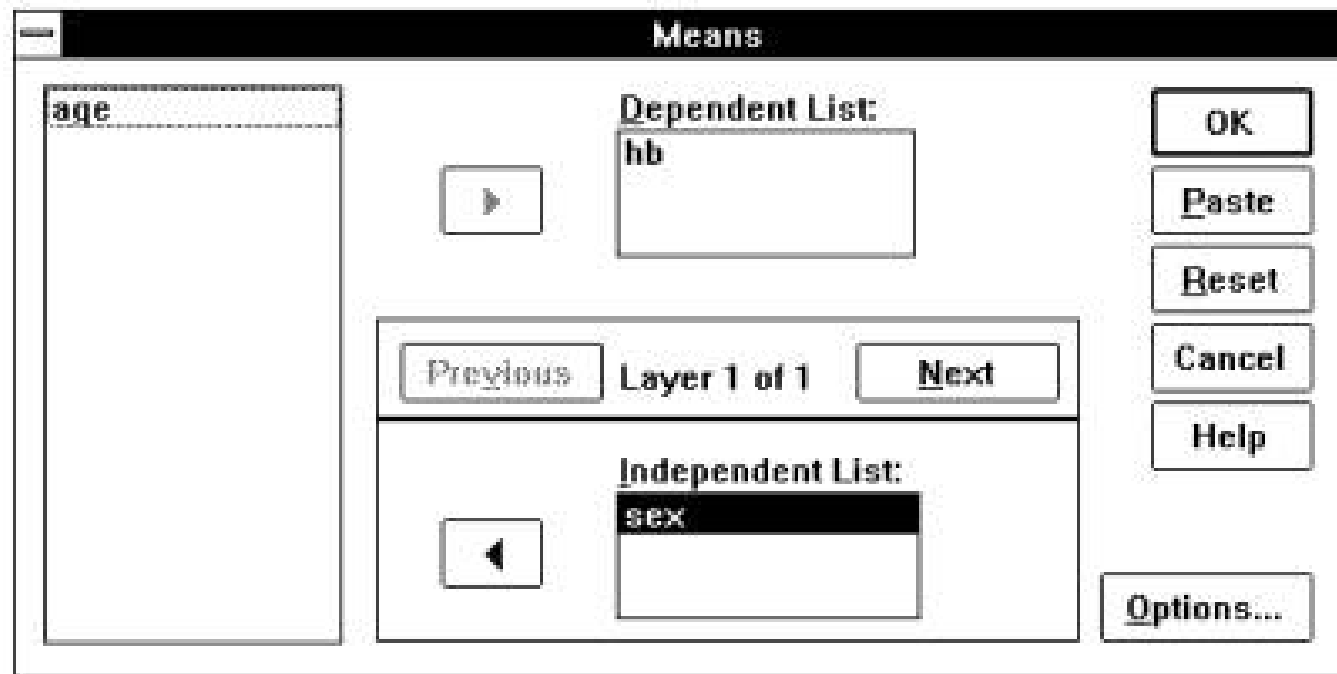
[例 5.1] 某医师测得如下血红蛋白值 (g%)，试作基本的描述性统计分析：

对象编号	性别	年龄	血红蛋白值	对象编号	性别	年龄	血红蛋白值
1	女	18	12.83	21	女	16	11.36
2	男	16	15.50	22	男	16	12.78
3	女	18	12.25	23	男	18	15.09
4	女	17	10.06	24	女	18	8.67
5	男	16	10.88	25	女	17	8.56
6	男	18	9.65	26	女	18	12.56
7	女	16	8.36	27	女	17	11.56
8	男	18	11.66	28	男	16	14.67
9	女	18	8.54	29	男	16	7.88
10	女	17	7.78	30	男	18	12.35
11	男	18	13.66	31	男	16	13.65
12	男	18	10.57	32	女	16	9.87
13	男	16	12.56	33	女	18	10.09
14	女	17	9.87	34	女	18	12.55
15	女	17	8.99	35	男	18	16.04
16	女	17	11.35	36	男	18	13.78
17	男	17	14.56	37	男	17	11.67
18	男	16	12.40	38	男	17	10.98
19	女	16	8.05	39	女	16	8.78
20	男	18	14.03	40	男	16	11.35

激活数据管理窗口，定义变量名：性别为sex，年龄为age，血红蛋白值为hb。按顺序输入数据(sex变量中，男为1，女为2)，结果见下图。

Newdata					
1:hb	12.83				
	sex	age	hb		
1	2	18	12.83		
2	1	16	15.50		
3	2	18	12.25		
4	2	17	10.06		
5	1	16	10.88		
6	1	18	9.65		
7	2	16	8.36		
8	1	18	11.66		
9	2	18	8.54		

- 激活Statistics菜单选Compare Means中的Means...项，弹出Means对话框（如图5.2示）。今欲分性别同时分年龄求血红蛋白值的均数和标准差



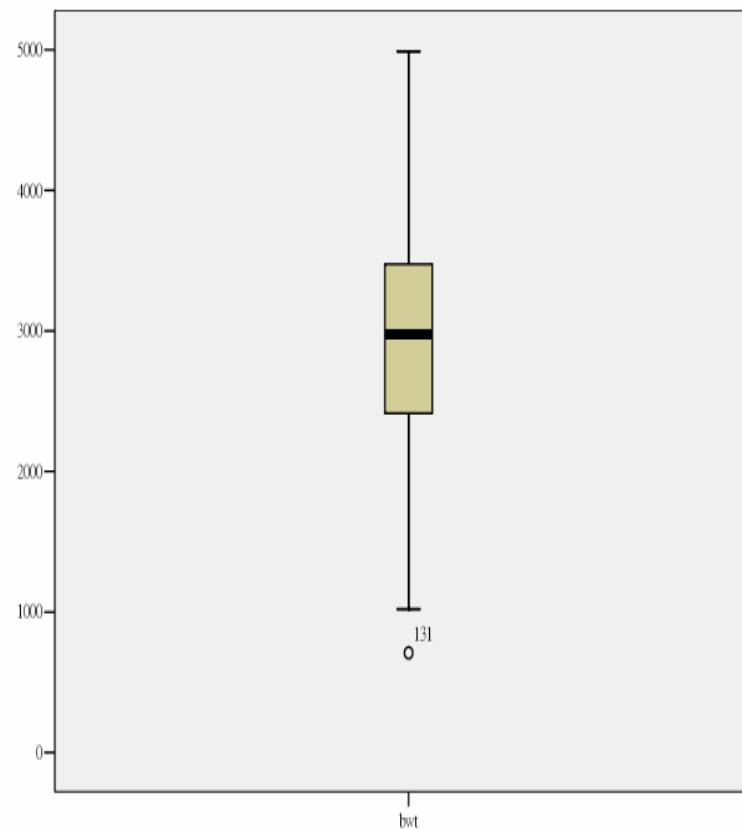
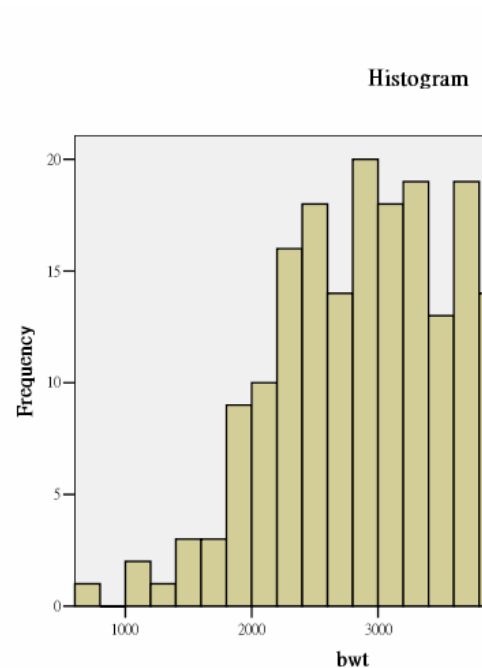


Tests of Normality

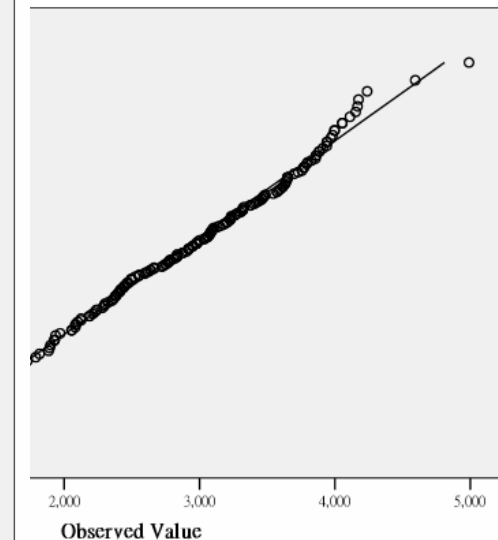
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
bwt	.043	189	.200*	.992	189	.440

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Normal Q-Q Plot of bwt



Q: Compare the BW between smoker and non-smoker

For n > 50

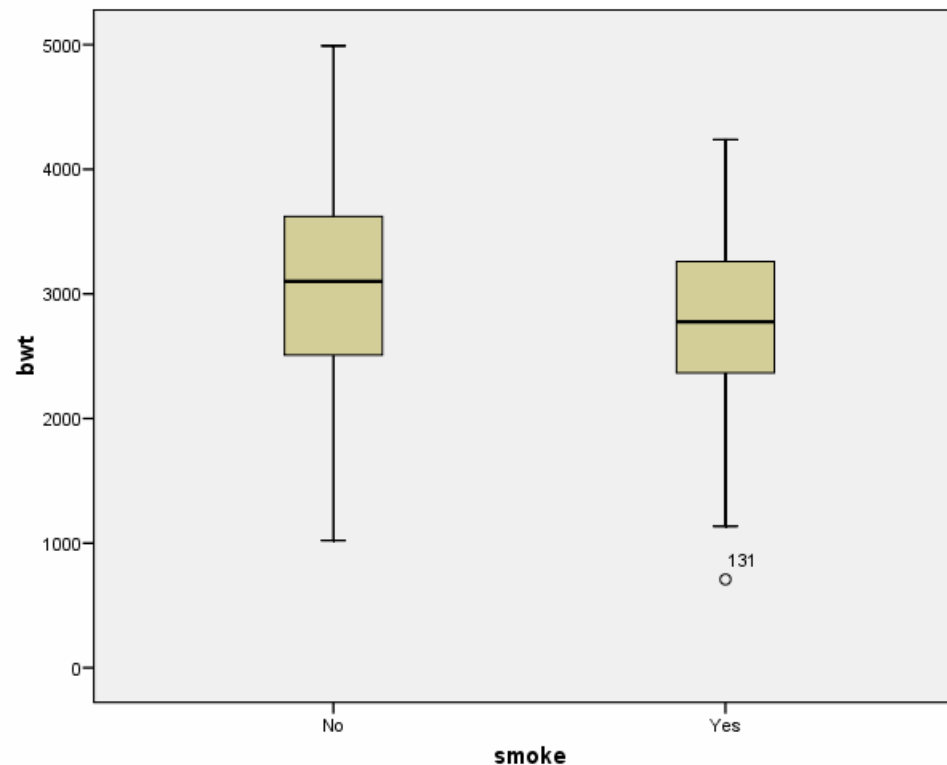
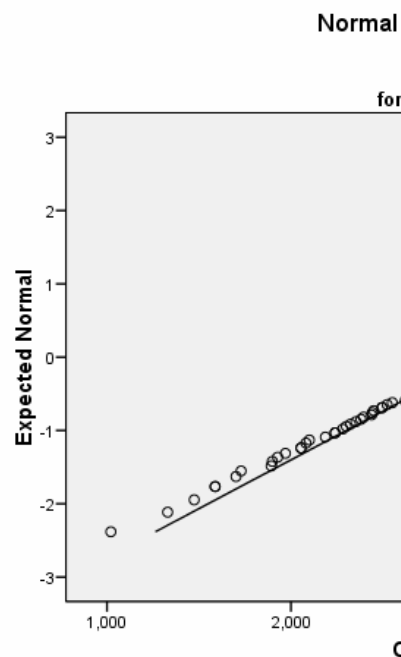
Tests of Normality

For n < 50

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
smoke		Statistic	df	Sig.	Statistic	df	Sig.
bwt	No	.060	115	.200*	.987	115	.345
	Yes	.069	74	.200*	.983	74	.410

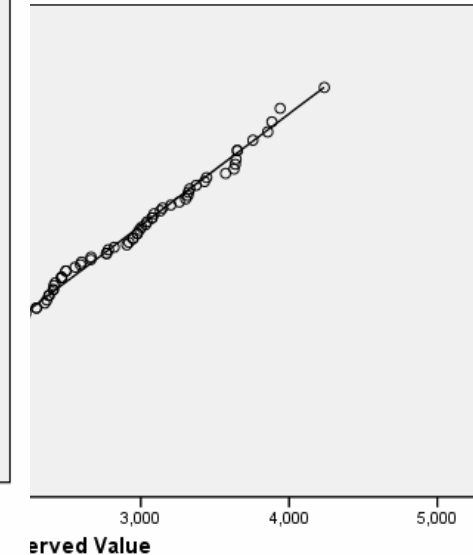
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Plot of bwt

smoke= Yes



第二节 Independent-Samples T Test过程

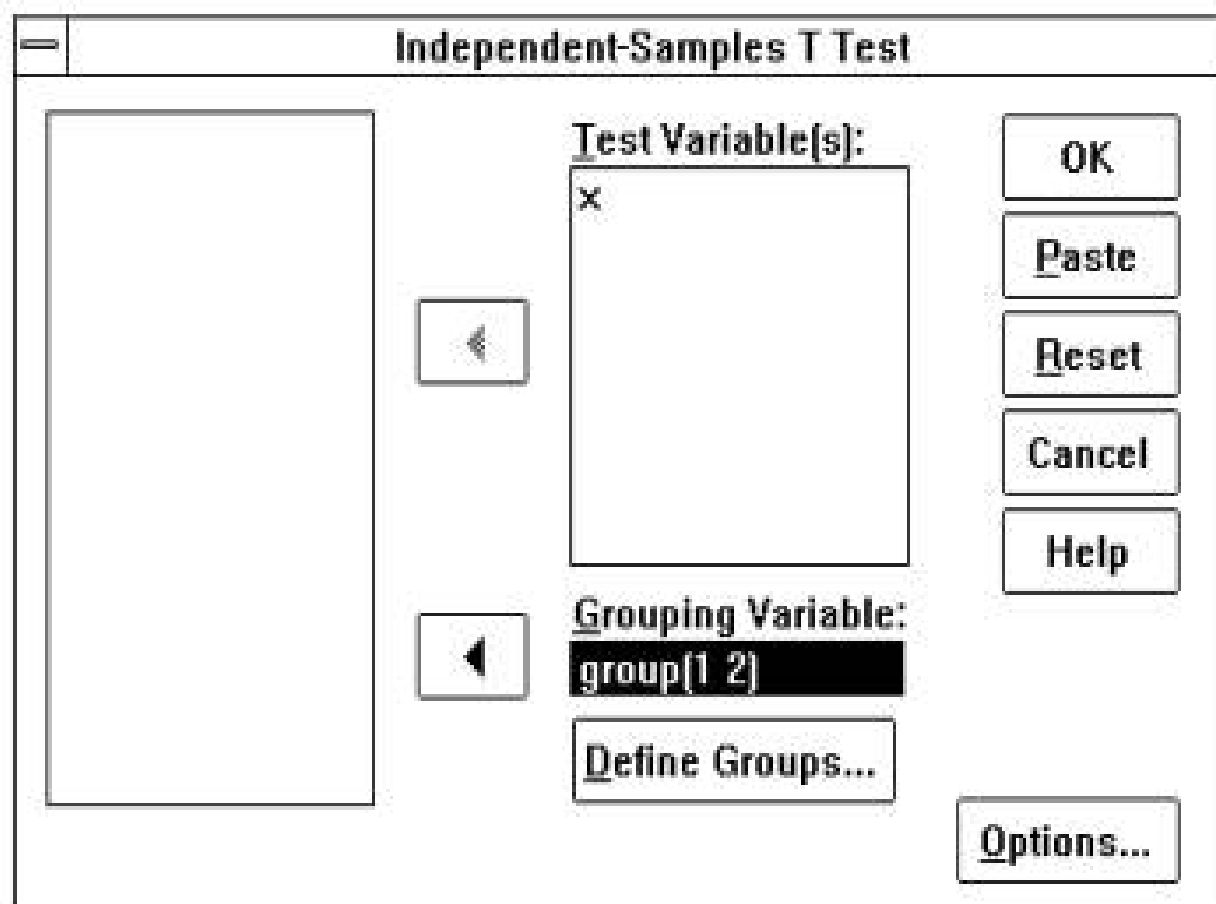
两组资料的t检验

- 分别测得14例老年性慢性支气管炎病人及11例健康人的尿中17酮类固醇排出量（mg/dl）如下，试比较两组均数有无差别。

病人	2.90	5.41	5.48	4.60	4.03	5.10	4.97
	4.24	4.36	2.72	2.37	2.09	7.10	5.92
健康人	5.18	8.79	3.14	6.46	3.72	6.64	5.60
	4.57	7.71	4.99	4.01			

- 定义变量名：把实际观察值定义为x，再定义一个变量group来区分病人与健康人。输入原始数据，在变量group中，病人输入1，健康人输入2。结果如图

Newdata					
16:group		2			
	x	group			
11	2.37	1			
12	2.09	1			
13	7.10	1			
14	5.92	1			
15	5.18	2			
16	8.79	2			
17	3.14	2			
18	6.46	2			
19	3.72	2			



SPSS中的数据输入以及统计处理的过程

The image shows the SPSS software interface with a data view in the background and two dialog boxes in the foreground.

Background Data View:

age	lwt	race
19	182	
33	155	
20	105	
21	108	
18	107	
21	124	
22	118	
17	103	
29	123	
26	113	
19	95	
19	150	
22	95	
30	107	
18	100	
18	100	
15	98	
25	118	
20	120	3
28	120	1
32	121	3

Two-Independent-Samples Tests Dialog Box:

- Test Variable List: bwt
- Grouping Variable: smoke(??)
- Test Type: ☒ Mann-Whitney U, ☐ Kolmogorov-Smirnov Z, ☐ Moses extreme reactions, ☐ Wald-Wolfowitz runs

Two Independent Samples: Defin... Sub-dialog Box:

- Group 1: 0
- Group 2: 1



Group Statistics

	smoke	N	Mean	Std. Deviation	Std. Error Mean
bwt	No	115	3054.72	752.474	70.169
	Yes	74	2773.24	660.075	76.732

Independent Samples Test

先看F, 後看 t		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
									Lower Upper
bwt	Equal variances assumed	1.519	.219	2.631	187	.009	281.478	106.975	70.446 492.511
	Equal variances not assumed			2.707	170.010	.007	281.478	103.978	76.224 486.733

Mann-Whitney Test (Nonparametric Two-samples Test)

Test Statistics^a

	bwt
Mann-Whitney U	3266.500
Wilcoxon W	6041.500
Z	-2.693
Asymp. Sig. (2-tailed)	.007

a. Grouping Variable: smoke

Group Statistics

	ANA >= 80 is Positive	N	Mean	Std. Deviation	Std. Error Mean
TBIITRAB	.00	93	48.66	25.48	2.64
	1.00	47	42.46	19.04	2.78

$$\sqrt{(25.48^2 / 93)} = 2.64$$

Independent Samples Test

Two Independent samples t-test

先看F, 後看 t		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
TBIITRAB	Equal variances assumed	6.826	.010	1.472	138	.143	6.20	4.21	-2.13	14.53
	Equal variances not assumed			1.617	118.447	.109	6.20	3.83	-1.39	13.79

Nonparametric Two Samples Test

Test Statistics^a

	TBIITRAB
Mann-Whitney U	1845.500
Wilcoxon W	2973.500
Z	-1.500
Asymp. Sig. (2-tailed)	.134

a. Grouping Variable: ANA >= 80 is Positive

Test Statistics^a

	TBIITRAB
Most Extreme Differences	.197
Absolute Positive	.087
Negative	-.197
Kolmogorov-Smirnov Z	1.098
Asymp. Sig. (2-tailed)	.179

a. Grouping Variable: ANA >= 80 is Positive

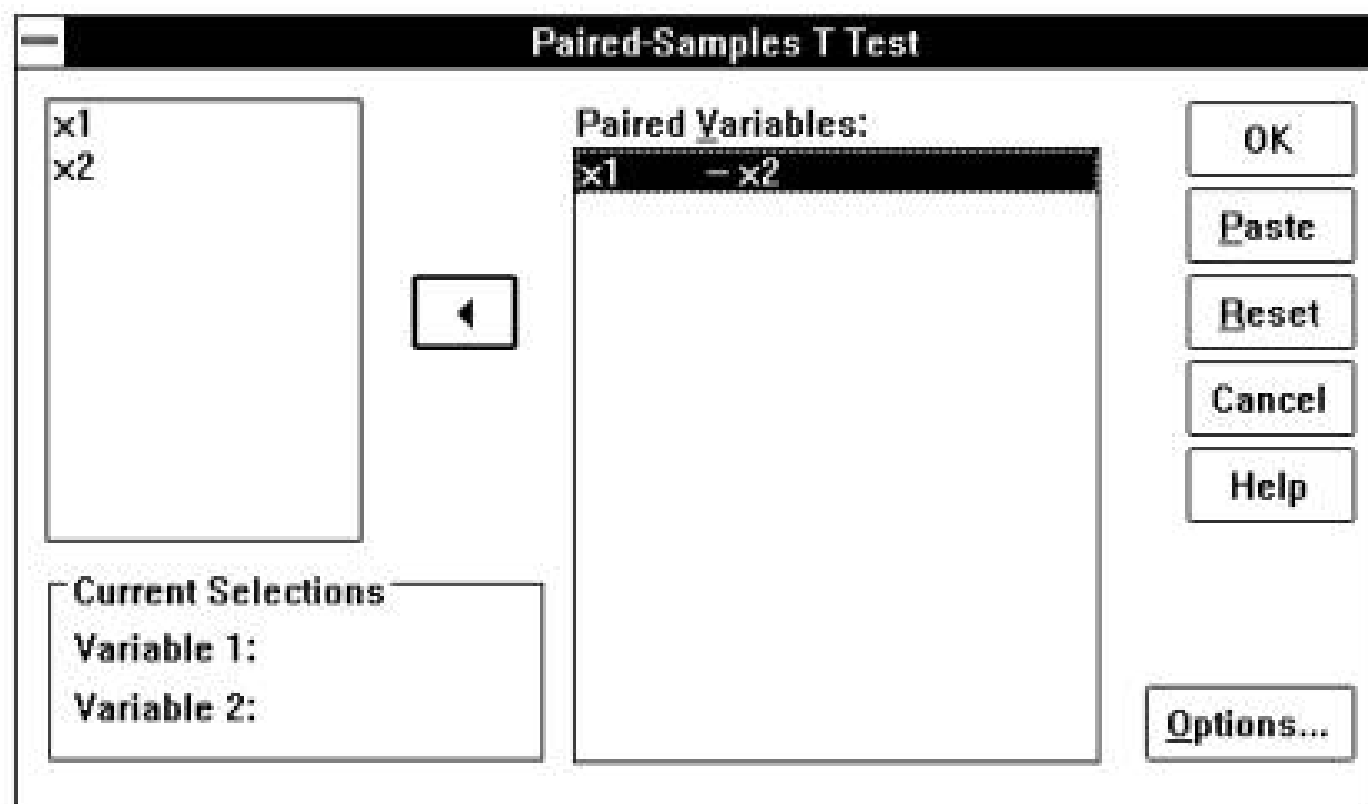
第三节 **Paired-Samples T Test**过程

- 配对资料包括：同对（年龄、性别、体重、病况等非处理因素相同或相似者）或同一研究对象分别给予两种不同处理的效果比较，以及同一研究对象处理前后的效果比较。前者推断两种效果有无差别，后者推断某种处理是否有效

- 某单位研究饲料中缺乏维生素E与肝中维生素A含量的关系，将大白鼠按性别、体重等配为8对，每对中两只大白鼠分别喂给正常饲料和维生素E缺乏饲料，一段时期后将之宰杀，测定其肝中维生素A含量 ($\mu\text{mol/L}$) 如下，问饲料中缺乏维生素E对鼠肝中维生素A含量有无影响？

大白鼠对别	肝中维生素A含量 ($\mu\text{mol/L}$)	
	正常饲料组	维生素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

Newdata					
8:x2		18.3			
	x1	x2			
1	37.20	25.70			
2	20.90	25.10			
3	31.40	18.80			
4	41.40	33.50			
5	39.80	34.00			
6	39.30	28.30			
7	36.10	26.20			
8	31.90	18.30			



第四节 **One-Way ANOVA**过程

多组间的均数比较

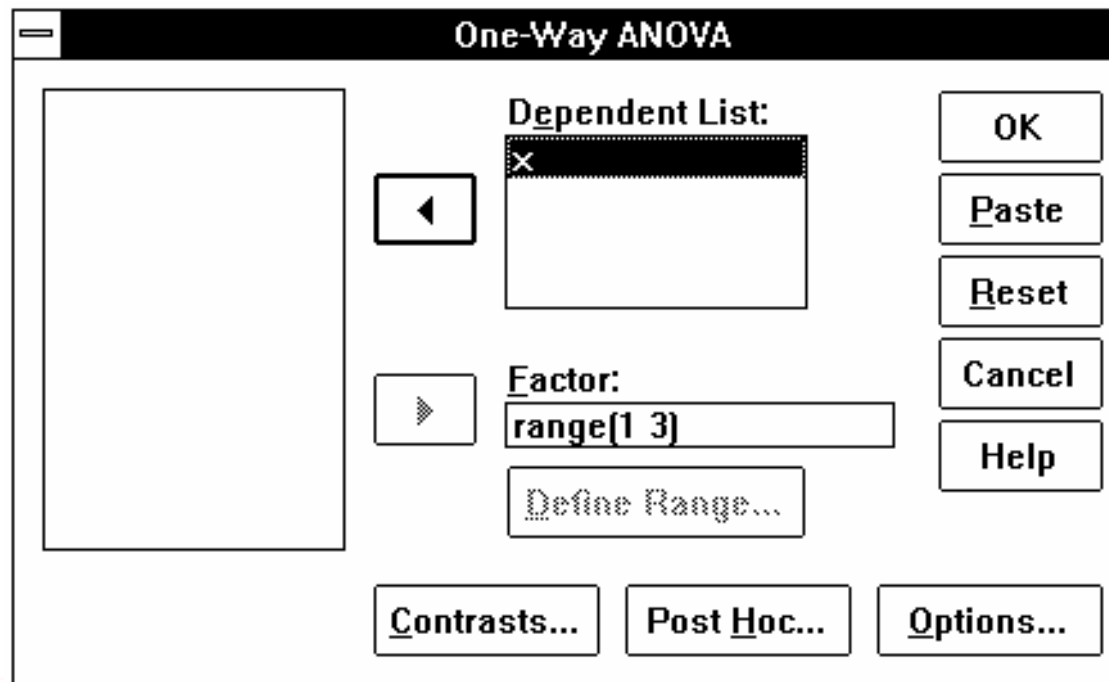
问两制剂是否有效？

对照组	甲制剂组	乙制剂组
279	129	210
334	174	285
303	110	117
338		
298		

- 实际观察值定义为x，组别用变量range表示：其中对照组的值为、甲制剂实验组的值为、乙制剂实验组的值为，输入后的结果如图

Newdata					
9:x	210				
	x	range			
4	338	1			
5	298	1			
6	129	2			
7	174	2			
8	110	2			
9	210	3			
10	285	3			
11	117	3			

- 菜单选Compare Means中的One-Way ANOVA...项，弹出One-Way ANOVA 对话框（如图5.8示）。从对话框左侧的变量列表中选x，点击➤钮使之进入Dependent List框，选range 点击➤钮使之进入Factor框，点击Define Range钮打开One-Way ANOVA: Define Range 对话框，因本例为3组比较，故在Minimum处输入1，在Maximum处输入3，点击Continue钮返回One-Way ANOVA 对话框。如果欲作多个样本均数间两两比较，可点击该对话框的Post Hoc...钮打开One-Way ANOVA: Post Hoc Multiple Comparisons对话框（如图5.9所示），这时可见在Tests框中有7种比较方法供选择：



SPSS中的数据输入以及统计处理的过程

LowBirthWeight.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : id 85

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftw	bwt
1						0	0	0	1	0	2523
2						0	0	0	0	3	2551
3						1	0	0	0	1	2557
4						1	0	0	1	2	2594
5						1	0	0	1	0	2600
6						0	0	0	0	0	2622
7						0	0	0	0	1	2637
8						0	0	0	0	1	2637
9						1	0	0	0	1	2663
10											2665
11											2722
12										1	2733
13										0	2750
14	99	0	30	107	3					1	2750
15	100	0	18	100	1					0	2769
16	101	0	18	100	1	1	0	0	0	0	2769
17	102	0	15	98	2	0	0	0	0	0	2778
18	103	0	25	118	1	1	0	0	0	3	2782
19	104	0	20	120	3	0	0	0	1	0	2807
20	105	0	28	120	1	1	0	0	0	1	2821
21	106	0	32	121	3	0	0	0	0	2	2835
22	107	0	31	100	1	0	0	0	1	3	2835
23	108	0	36	202	1	0	0	0	0	1	2836
24	109	0	28	120	3	0	0	0	0	0	2836

Tests for Several Independent Samples

Test Variable List: bwt

Grouping Variable: race(??)

Test Type

☒ Kruskal-Wallis H ☐ Median

☐ Jonckheere-Terpstra

Several Independent Samples: Define Range

Range for Grouping Variable

Minimum: 1

Maximum: 3

SPSS中Post Hoc统计的选择示意图

LowBirthWeight.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1: id 85

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt	v
1						0	0	0	1	0	2523	
2						0	0	0	0	3	2551	
3						1	0	0	0	1	2557	
4						1	0	0	1	2	2594	
5						1	0	0	1	0	2600	
6						0	0	0	0	0	2622	
7						0	0	0	0	1	2637	
8						0	0	0	0	1	2637	
9												
10												
11												
12												
13	98	0	22	95	3							
14	99	0	30	107	3							
15	100	0	18	100	1							
16	101	0	18	100	1							
17	102	0	15	98	2							
18	103	0	25	118	1							
19	104	0	20	120	3							
20	105	0	28	120	1							
21	106	0	32	121	3							
22	107	0	31	100	1							
23	108	0	36	202	1	0	0	0	0	1	2836	
24	109	0	28	120	3	0	0	0	0	0	2836	
25	111	0	25	120	3	0	0	0	1	2	2877	

One-Way ANOVA

Dependent List: bwt

Factor: race

Contrasts... Post Hoc... Options...

One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

- ☐ LSD
- ☒ Bonferroni
- ☐ Sidak
- ☐ Scheffe
- ☐ R-E-G-W F
- ☐ R-E-G-W Q
- ☐ S-N-K
- ☒ Tukey
- ☐ Tukey's-b
- ☐ Duncan
- ☐ Hochberg's GT2
- ☐ Gabriel
- ☐ Waller-Duncan
- Type I/Type II Error Ratio: 100
- ☐ Dunnett
- Control Category: Last
- Test: ☒ 2-sided ☐ < Control ☐ > Control

Equal Variances Not Assumed

- ☐ Tamhane's T2
- ☐ Dunnett's T2
- ☐ Games-Howell
- ☐ Dunnett's C

Significance level: .05

Continue Cancel Help

Post Hoc Multiple Comparisons

- Least-significant difference (LSD): 最小显著差法。 α 可指定0~1之间任何显著性水平，默认值为0.05;
- Bonferroni: Bonferroni修正差别检验法。 α 可指定0~1之间任何显著性水平，默认值为0.05;
- Duncan's multiple range test: Duncan多范围检验。只能指定 α 为0.05或0.01或0.1，默认值为0.05;
- Student-Newman-Keuls: Student-Newman-Keuls检验，简称N-K检验,亦即q检验。 α 只能为0.05;
- Tukey's honestly significant difference: Tukey显著性检验。 α 只能为0.05;
- Tukey's b: Tukey另一种显著性检验。 α 只能为0.05;
- Scheffe: Scheffe差别检验法

ANOVA

bwt

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5078209.278	2	2539104.639	4.979	.008
Within Groups	94843977.939	186	509913.860		
Total	99922187.217	188			

先F检验

先F检验, $P=0.008$
 <0.5 , 才有必要进入下一步

Multiple Comparisons

Dependent Variable: bwt
 Bonferroni

(I) race	(J) race	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	384.047*	157.872	.048	2.66	765.44
	3	300.128*	113.676	.027	25.51	574.75
2	1	-384.047*	157.872	.048	-765.44	-2.66
	3	-83.920	164.993	1.000	-482.51	314.68
3	1	-300.128*	113.676	.027	-574.75	-25.51
	2	83.920	164.993	1.000	-314.68	482.51

*, The mean difference is significant at the .05 level.

Post Hoc Tests:
 (事后检验)

Nonparametric Method
 (Kruskal-Wallis Test)

Test Statistics^{a,b}

	bwt
Chi-Square	8.598
df	2
Asymp. Sig.	.014

a. Kruskal Wallis Test

b. Grouping Variable: race

第五节 方差分析过程

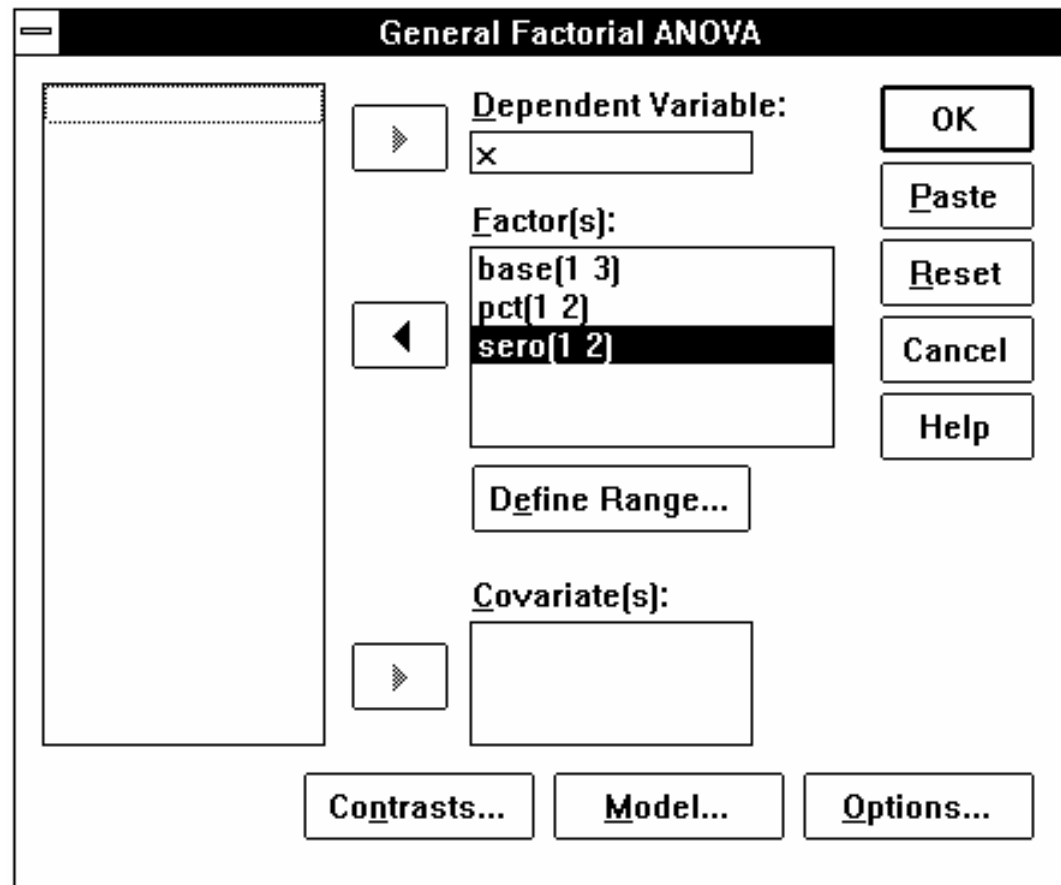
衍生:方差分析

- 两组单样本比较: T检验
- 多组单样本比较: one way ANOVA
- 那么
- 两组多样本比较呢? 方差分析
- 多组多样本比较呢? 多因素方差分析

- 为三因素析因实验的资料，请用方差分析说明不同基础液与不同血清种类对钩端螺旋体的培养计数的影响。

基础液 (A)	血清种类 (B)			
	兔血清浓度 (C)		胎盘血清浓度 (C)	
	5%	8%	5%	8%
缓冲液	648	1144	830	578
	1246	1877	853	669
	1398	1671	441	643
	909	1845	1030	1002
蒸馏水	1763	1447	920	933
	1241	1883	709	1024
	1381	1896	848	1092
	2421	1926	574	742
自来水	580	1789	1126	685
	1026	1215	1176	546
	1026	1434	1280	595
	830	1651	1212	566

- 定义变量名：基础液为base，血清种类为sero，血清浓度为pct，钩端螺旋体的培养计数为X，按顺序输入相应数值，建立数据库



选ANOVA Models中的General Factorial...项，弹出General Factorial ANOVA对话框

Multivariate过程:多元方差分析

学生 编号	甲地区			乙地区			丙地区		
	身高	体重	胸围	身高	体重	胸围	身高	体重	胸围
1	119.80	22.60	60.50	125.10	23.00	62.00	118.30	20.40	54.40
2	121.70	21.50	55.50	127.00	21.50	59.00	121.30	20.00	54.30
3	121.40	19.10	56.50	125.70	23.40	61.50	121.80	26.60	
4	124.40	21.80	60.50	114.90	17.50	52.50	124.20	22.10	
5	120.00	21.40	57.70	124.90	23.50	58.50	123.50	23.20	
6	117.00	20.10	57.00	117.60	18.90	57.00	123.00	22.90	
7	118.10	18.80	57.10	124.20	20.80	58.50	134.90	32.30	
8	118.80	22.00	61.70	117.90	20.30	61.00	123.70	22.70	
9	124.20	21.30	58.40	120.40	20.00	56.00	105.20	20.20	
10	124.90	24.00	60.80	115.00	19.70	56.50	112.20	20.80	
11	124.70	23.30	60.00	126.20	21.20	56.50	118.60	21.00	
12	123.00	22.50	60.00	125.10	22.10	58.50	112.00	23.20	
13	125.30	22.90	65.20	114.90	19.70	56.00	121.50	24.00	
14	124.20	19.50	53.80	121.50	22.00	57.00	124.50	21.50	
15	127.40	22.90	59.50	114.00	19.00	54.50	119.50	20.50	
16	128.20	22.30	60.00	118.70	19.10	54.50	122.50	23.00	
17	126.10	22.70	57.40	120.60	20.00	55.50	115.50	19.00	
18	128.70	23.50	60.40	122.90	18.50	56.00	122.50	22.50	
19	129.50	24.50	51.00	119.60	19.50	59.50	124.50	25.00	
20	126.90	25.50	61.50	112.30	20.00	58.00	125.00	25.50	

第六节 相关分析过程

相关的几种类型

★正相关 ★负相关 ★完全正相关 ★完全负相关 ★称零相关

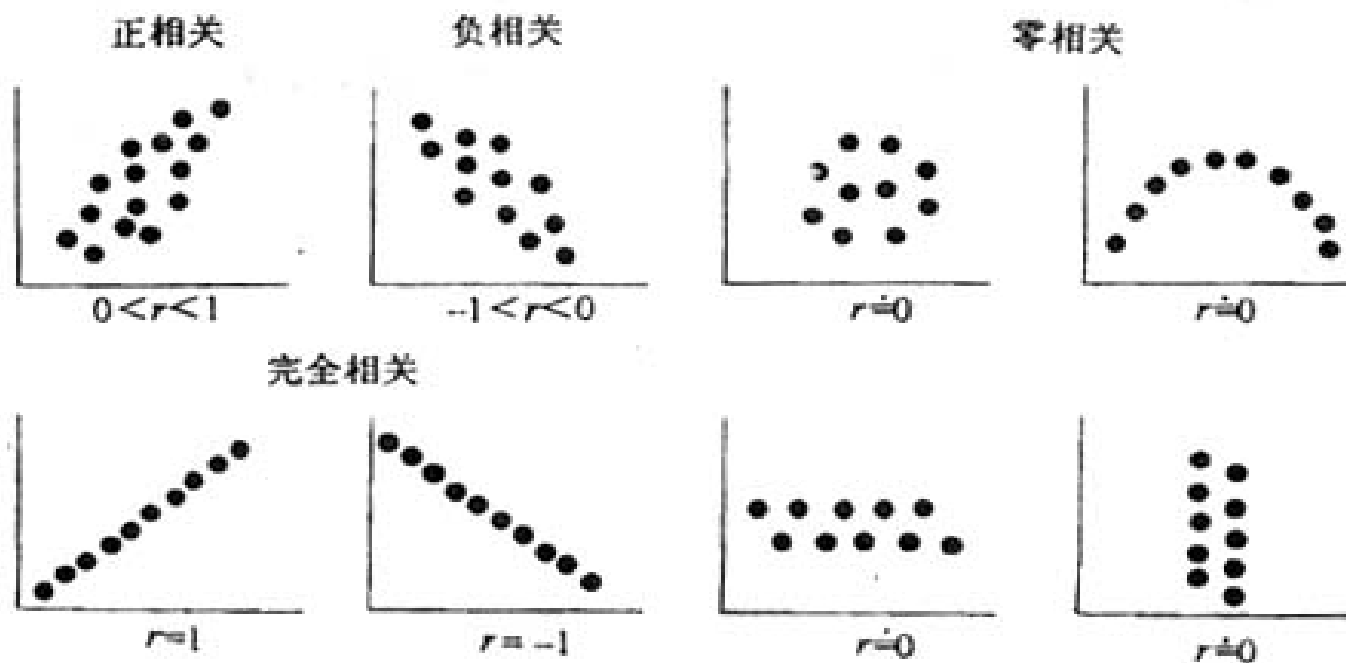


图 9.2 相关系数示意图

Bivariate过程

- 允许同时输入两变量或两个以上变量，但系统输出的是变量间两两相关的相关系数

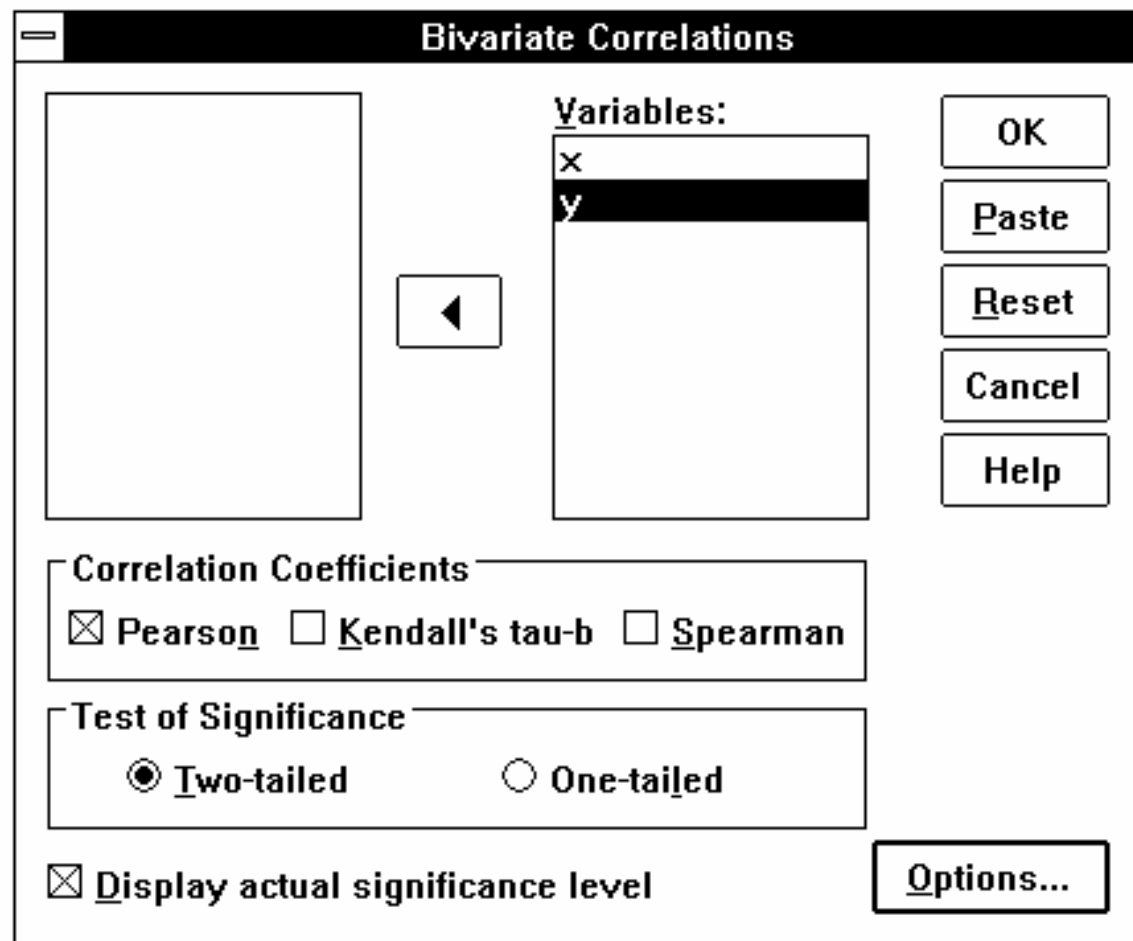
发硒与血硒的相关分析。

编号	发硒	血硒
1	74	13
2	66	10
3	88	13
4	69	11
5	91	16
6	73	9
7	66	7
8	96	14
9	58	5
10	73	10

- 定义变量名：发硒为X，血硒为Y，按顺序输入相应数值，建立数据库（图7.1）。

Newdata			
	x	y	
1	74	13	
2	66	10	
3	88	13	
4	69	11	
5	91	16	
6	73	9	
7	66	7	
8	96	14	
9	58	5	
10	73	10	

菜单选Correlate中的Bivariate...命令项，弹出Bivariate
Correlation对话框



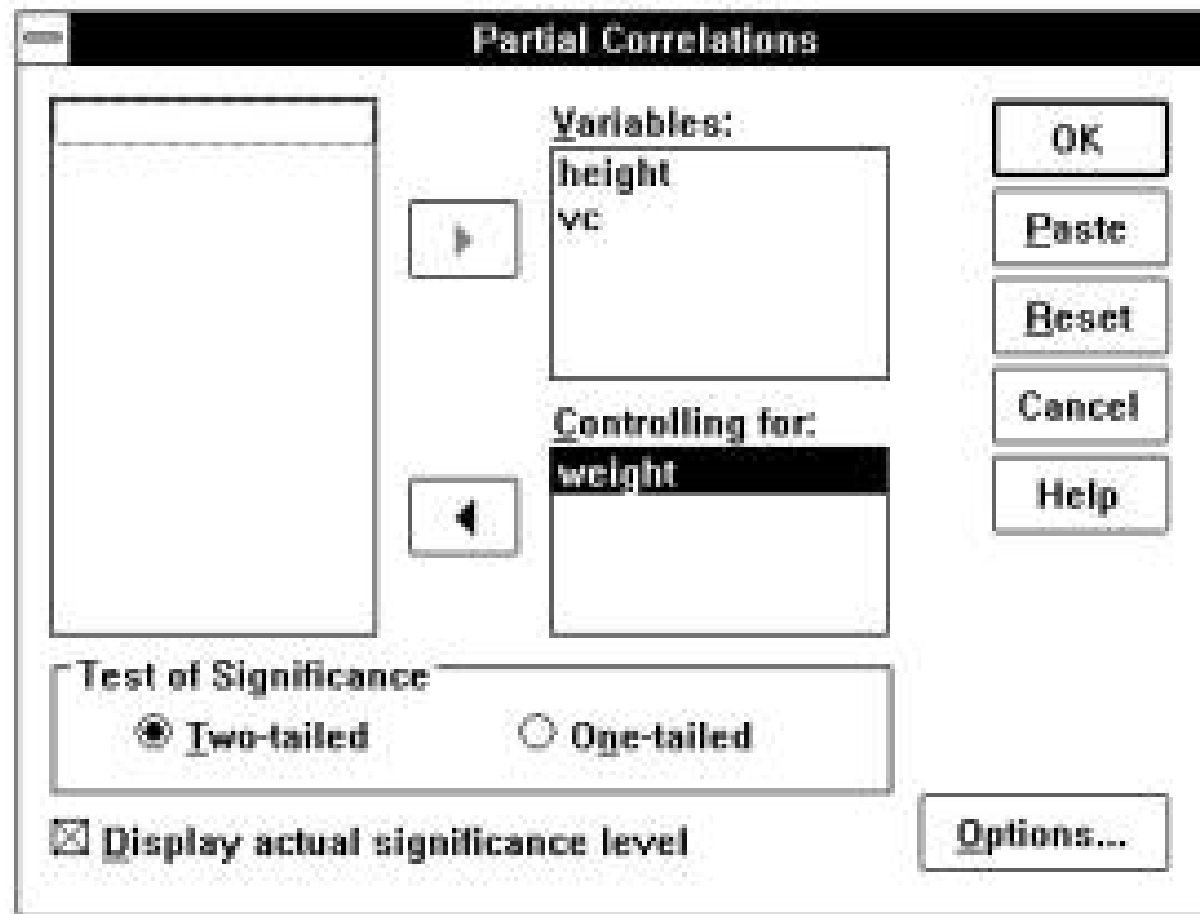
Partial过程

- 在偏相关分析中，系统可按用户的要求对两相关变量之外的某一或某些影响相关的其他变量进行控制，输出控制其他变量影响后的相关系数。

- 某地29名13岁男童身高（cm）、体重（kg）和肺活量（ml）的数据如下表, 试对该资料作控制体重影响作用的身高与肺活量相关分析

编号	身高 (cm)	体重(kg)	肺活量(ml)	编号	身高 (cm)	体重(kg)	肺活量(ml)
1	135.1	32.0	1750	16	153.0	47.2	1750
2	139.9	30.4	2000	17	147.6	40.5	2000
3	163.6	46.2	2750	18	157.5	43.3	2250
4	146.5	33.5	2500	19	155.1	44.7	2750
5	156.2	37.1	2750	20	160.5	37.5	2000
6	156.4	35.5	2000	21	143.0	31.5	1750
7	167.8	41.5	2750	22	149.4	33.9	2250
8	149.7	31.0	1500	23	160.8	40.4	2750
9	145.0	33.0	2500	24	159.0	38.5	2500
10	148.5	37.2	2250	25	158.2	37.5	2000
11	165.5	49.5	3000	26	150.0	36.0	1750
12	135.0	27.6	1250	27	144.5	34.7	2250
13	153.3	41.0	2750	28	154.6	39.5	2500
14	152.0	32.0	1750	29	156.5	32.0	1750
15	160.5	47.2	2250				

- 激活数据管理窗口，定义变量名：身高为height，体重为weight，肺活量为vc，按顺序输入相应数值，建立数据库。



- 在结果输出窗口中将看到如下统计数据：控制体重的影响后，身高与肺活量的相关系数为0.0926，经检验 $P = 0.639$ ，故身高与肺活量的线性相关不存在。（如果不控制体重的影响，则身高与肺活量的相关系数为0.5884， P 为0.001。在有控制的情况下，身高与肺活量的决定系数 $= r^2 = 0.00857$ ，而无控制的身高与肺活量决定系数 $= r^2 = 0.34621$ ，可见身高与肺活量的相关有33.764%是由体重协同作用而产生的。）

```
Controlling for..  WEIGHT
                   HEIGHT    VC
HEIGHT    1.0000    .0926
          ( 0) ( 26)
          P= .    P= .639
VC         .0926    1.0000
          ( 26) ( 0)
          P= .639  P= .
```

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed

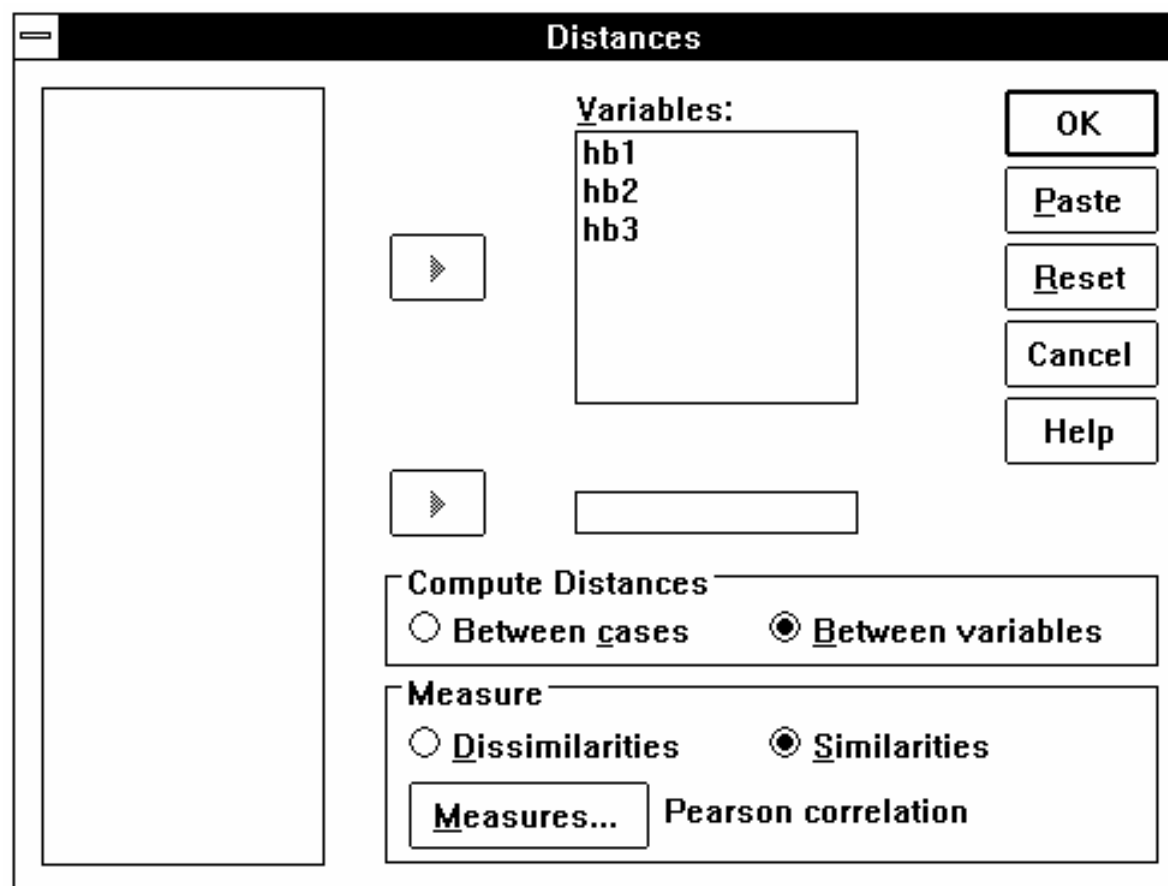
- 如果控制变量改为身高，则得如下结果：体重与肺活量的相关系数为0.5528，经检验 $P = 0.002$ ，故体重与肺活量的线性相关存在。可见，尽管肺活量与身高和体重均有关系，但如果仅仅研究其中一个变量与肺活量的相关关系时，体重的意义会更大。

距离相关分析

- 某医师对10份标准血红蛋白样品作三次平行检测，结果如下，问检测结果是否一致？

样品号	1	2	3	4	5	6	7	8	9	10
第一次	12.36	12.14	12.31	12.32	12.12	12.28	12.24	12.41	12.33	12.17
第二次	12.40	12.20	12.28	12.25	12.22	12.34	12.31	12.30	12.22	12.24
第三次	12.18	12.22	12.35	12.21	12.10	12.25	12.20	12.46	12.36	12.11

- 菜单选Correlate中的Distance...命令项



- 在Measure栏中有两种测距方式：
Dissimilarities为不相似性测距，Similarities为相似性测距。若选Dissimilarities并点击Measure...钮，弹出Distance:Dissimilarity Measure对话框（图7.6），用户可根据数据特征选用测距方法：

Distances: Dissimilarity Measures

Measure

☒ **Interval:** Euclidean distance

2 2

☐ **Counts:** Chi-square measure

☐ **Binary:** Euclidean distance

Transform Values

Standardize: None

☒ By variable

☐ By case

Transform Measures

☐ Absolute values

☐ Change sign

☐ Rescale to 0-1 range

Continue

Cancel

Help

- 1、计量资料
- Euclidean distance: 以两变量差值平方和的平方根为距离;
- Squared Euclidean distance: 以两变量差值平方和为距离;
- Chebychev: 以两变量绝对差值的最大值为距离;
- Block: 以两变量绝对差值之和为距离;
- Minkowski: 以两变量绝对差值 p 次幂之和的 p 次根为距离;
- Customized: 以两变量绝对差值 p 次幂之和的 r 次根为距离。
- 2、计数资料
- Chi-square measure: χ^2 值测距;
- Phi-square measure: ψ^2 值测距, 即将 χ^2 测距值除合计频数的平方根。
- 3、二分字符变量
- Euclidean distance: 二分差平方和的平方根, 最小为0, 最大无限;
- Squared Euclidean distance: 二分差平方和, 最小为0, 最大无限;
- Size difference: 最小距离为0, 最大无限;
- Pattern difference: 从0至1的无级测距;
- Variance: 以方差为距, 最小为0, 最大无限;
- Lance and Williams: Bray-Curtis非等距系数, 界于0至1之间。

- 若选Similarities并点击Measure...钮，弹出Distance: Similarity Measure对话框（图7.7），用户可根据数据特征选用测距方法：
-

Distances: Similarity Measures

Measure

☒ **Interval:** Pearson correlation

☐ **Binary:** Russell and Rao

Transform Values

Standardize: None

☒ **By variable**

☐ **By case**

Transform Measures

☐ **Absolute values**

☐ **Change sign**

☐ **Rescale to 0-1 range**

Continue

Cancel

Help

- **1、计量资料**
 - Pearson correlation: 以Pearson相关系数为距离;
 - Cosine: 以变量矢量的余弦值为距离, 介于-1至+1之间。
 - **2、二分字符变量**
 - Russell and Rao: 以二分点乘积为配对系数;
 - Simple matching: 以配对数与总对数的比例为配对系数;
 - Jaccard: 相似比例, 分子与分母中的配对数与非配对数给予相同的权重;
 - Dice: Dice配对系数, 分子与分母中的配对数给予加倍的权重;
 - Kulczynski 2: Kulczynski平均条件概率;
 - Sokal and Sneath 4: Sokal and Sneath 条件概率;
 - Hamann: Hamann概率;
 - Lambda: Goodman-Kruskal相似测量的 λ 值;
 - Anderberg's D: 以一个变量状态预测另一个变量状态;
 - Yule's Y: Yule综合系数, 属于 2×2 四格表的列联比例函数;
 - Yule's Q: Goodman-Kruskal γ 值, 属于 2×2 四格表的列联比例函数。
 - **3、其他型变量**
 - Ochiai: Ochiai二分余弦测量;
 - Sokal and Sneath 5: Sokal and Sneath V型相似测量;
 - Phi 4 point correlation: Pearson相关系数的平方值;
 - Dispersion: Dispersion相似测量。
-
- 本例选Similarities项, 并以Pearson correlation为测量距离。点击Continue钮返回Distance对话框, 再点击OK钮即可。

第七节 回归分析过程

回归分析是处理两个及两个以上变量间线性依存关系的统计方法。在医学领域中，此类问题很普遍，如人头发中某种金属元素的含量与血液中该元素的含量有关系，人的体表面积与身高、体重有关系。

说白了，就是N种因素，通过不同权重叠加后，形成某一固定因素间。例如：冠心病发生= $0.03 \times \text{吸烟} + 0.01 \times \text{饮酒} + 0.1 \times \text{高脂饮食} + 0.3 \times \text{高水平三油甘脂}$ 。冠心病是观察对象，而后面的因素可能从不同程度影响到冠心病，因此，与冠心病间有一个权重关系，如果将所有的因素叠加，就能够得出一个方程，并通过这个方程便能预测到任何一个人冠心病发生的可能。

非线性回归方程

(1) 整理数据

整理数据，选择合适的分析方法

(2) 画散点图

(3) 选方程

(4) 线性化

(5) 求解参数

(6) 参数带回原方程

SPSS软件可以自动完成

Linear过程

- 某医师测得10名3岁儿童的身高（cm）、体重（kg）和体表面积（cm²）资料如下。试用多元回归方法确定以身高、体重为自变量，体表面积为应变量的回归方程。

儿童编号	体表面积 (Y)	身高 (X_1)	体重 (X_2)
1	5.382	88.0	11.0
2	5.299	87.6	11.8
3	5.358	88.5	12.0
4	5.292	89.0	12.3
5	5.602	87.7	13.1
6	6.014	89.5	13.7
7	5.830	88.8	14.4
8	6.102	90.4	14.9
9	6.075	90.6	15.2
10	6.411	91.2	16.0

- 体表面积为Y，保留3位小数；身高、体重分别为X1、X2，1位小数。输入原始数据，结果如图8.1所示。

Newdata				
10:x2		16		
	y	x1	x2	
1	5.382	88.0	11.0	
2	5.299	87.6	11.8	
3	5.358	88.5	12.0	
4	5.292	89.0	12.3	
5	5.602	87.7	13.1	
6	6.014	89.5	13.7	
7	5.830	88.8	14.4	
8	6.102	90.4	14.9	
9	6.075	90.6	15.2	
10	6.411	91.2	16.0	

Linear Regression

x1
x2

Dependent:
y

Previous Block 1 of 1 Next

Independent(s):
x1
x2

Method: Enter

WLS >> Statistics... Plots... Save... Options...

OK
Paste
Reset
Cancel
Help

Curve Estimation过程

- 某地1963年调查得儿童年龄（岁） X 与锡克试验阴性率（%） Y 的资料如下，试拟合对数曲线。

年龄（岁） X	锡克试验阴性率（%） Y
1	57.1
2	76.0
3	90.9
4	93.0
5	96.7
6	95.6
7	96.2

Analyze==>Regression==>Curve estimation

SPSS中曲线拟合方程选项

Independent

☒ Variable:

☐ Time

☒ Include constant in model

Case Labels:

☐ Plot models

Models

☒ Linear ☐ Quadratic ☐ Compound ☐ Growth

☐ Logarithmic ☐ Cubic ☐ S ☐ Exponential

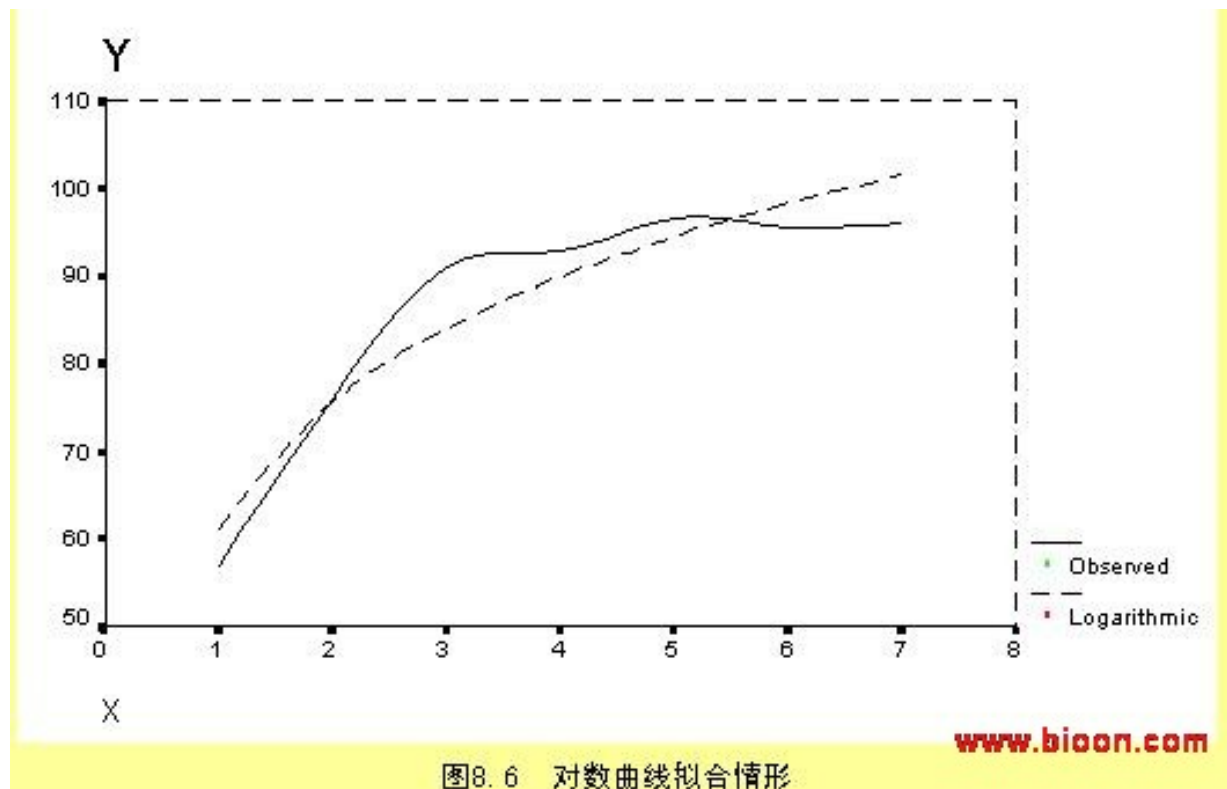
☐ Inverse ☐ Power ☐ Logistic

Upper bound:

☐ Display ANOVA table

www.bion.com

拟合曲线结果



SPSS软件可以任意选择各种曲线拟合的形式，然后可以根据拟合的R值（越接近1，表明拟合越一致），然后再选择合适的方程，这样更佳。

Binary Logistic过程

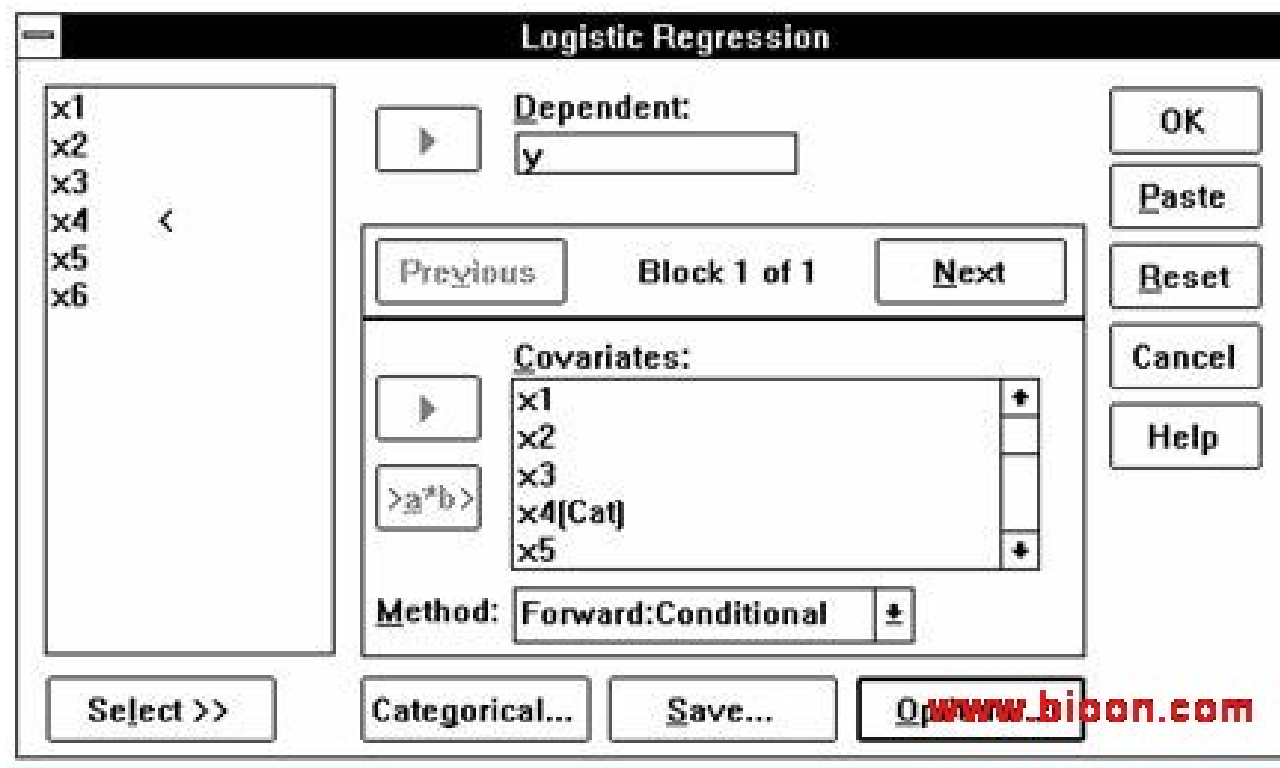
- 所谓Logistic回归，是指应变量为二级计分或二类评定的回归分析，这在医学研究中经常遇到，如：死亡与否（即生、死二类评定）的概率跟病人自身生理状况和所患疾病的严重程度有关；对某种疾病的易感性的概率（患病、不患病二类评定）与个体性别、年龄、免疫水平等有关。此类问题的解决均可借助逻辑回归来完成。

- 某医师研究男性胃癌患者发生术后院内感染的影响因素，资料如下表，请通过Logistic回归统计方法对主要影响因素进行分析。

术后感染 (有无) Y	年龄 (岁) X1	手术创伤程度 (5等级) X2	营养状态 (3等级)) X3	术前预防性抗 菌 (有无) X4	白细胞数 ($\times 10^9/L$)) X5	癌肿病理分度 (TNM得分总和)) X6
有	69	4	2	无	5.6	9
有	72	5	3	无	4.4	6
无	57	3	2	无	9.7	4
无	41	1	1	有	11.2	5
无	32	1	1	有	10.4	5
有	65	3	3	有	7.0	5
有	58	3	2	有	3.1	6
有	54	4	2	无	6.6	6
有	55	2	2	有	7.9	7
无	59	1	1	有	6.0	4
无	64	2	2	无	9.1	6
无	36	1	1	有	8.4	8
无	42	3	1	有	5.3	6
无	48	4	2	有	4.6	5
无	50	1	2	有	12.8	4

- 激活数据管理窗口，定义变量名：术后感染为Y（字符变量，有输入Y、无输入N），年龄为X1，手术创伤程度为X2，营养状态为X3，术前预防性抗菌为X4（字符变量，有输入Y、无输入N），白细胞数为X5，癌肿病理分度为X6。按要求输入原始数据。

- 菜单选Regression中的Logistic...项，弹出Logistic Regression对话框（如图8.8示）。从对话框左侧的变量列表中选y，点击➤钮使之进入Dependent框，选x1、x2、x3、x4、x5和x6，点击➤钮使之进入Covariates框；点击Method处的下拉按钮，系统提供7种方法：



Probit过程

- 完成剂量-效应关系的分析。通过概率单位使剂量-效应的S型曲线关系转化成直线，从而利用回归方程推算各效应水平的相应剂量值。

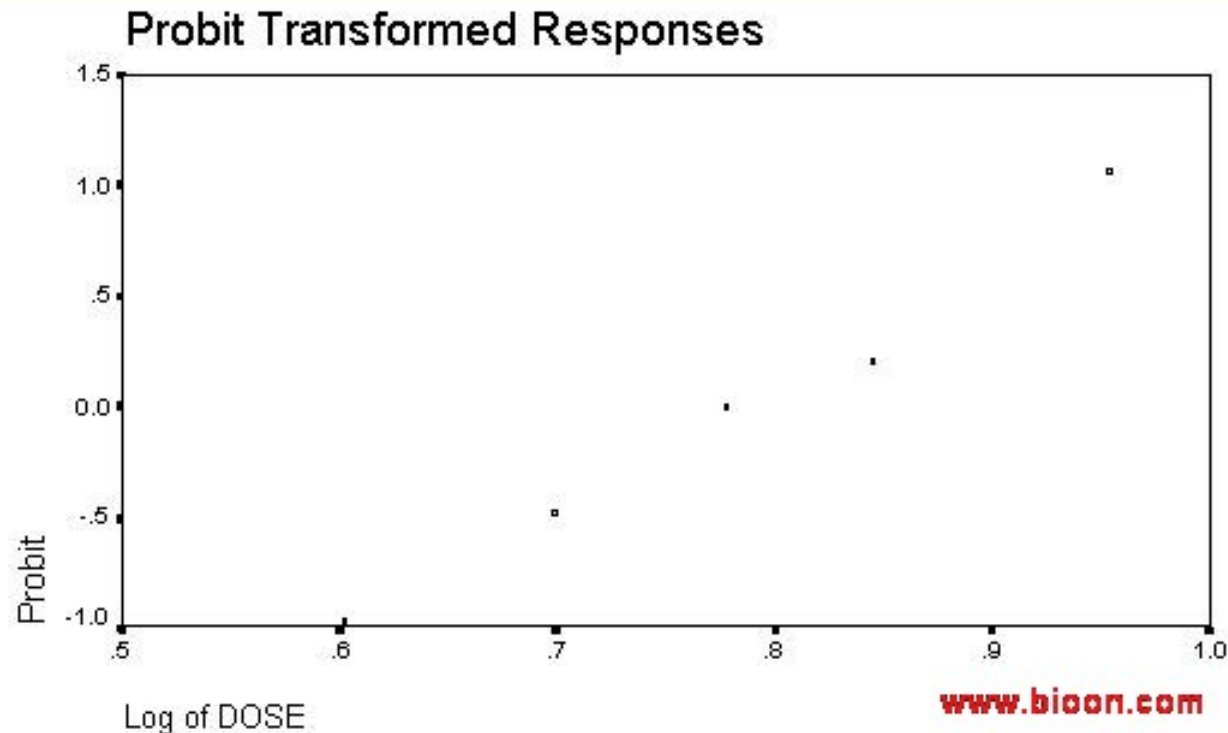
- 抗疟药环氯胍对小白鼠的毒性，试验结果如下表所示。试计算环氯胍的半数致死剂量。

剂量 (mg/kg)	动物数	死亡数
12	5	5
9	7	6
7	19	11
6	34	17
5	38	12
4	12	2
3	5	0

定义变量名：剂量为DOSE、试验动物数为OBSERVE、死亡动物数为DEATH。然后输入原始数据

- 菜单选Regression中的Probit...项，弹出Probit Analysis对话框（如图8.9示）。从对话框左侧的变量列表中选death，点击➤钮使之进入Response Frequency框；选observe，点击➤钮使之进入Total Observed框；选dose，点击➤钮使之进入Covariate(s)框，并下拉Transform菜单，选Log base 10项（即要求对剂量进行以10为底的对数转换）。
- 系统在Model栏中提供两种模型，一是概率单位模型（Probit），另一是比数比自然对数模型（Logit）。本例选用概率单位模型。

- 系统输出以剂量对数值为自变量X、以概率单位为应变量Y的回归直线散点图，从图中各点的分布状态亦可看出，**回归直线**的拟合程度是很好的。



Nonlinear过程

- 选取某地某年寿命表中40-80岁各年龄组的尚存人数资料如下表，请就该资料试拟合Gompertz曲线（ $Y = b_1 \times b_2(b_3^X)$ ）。

年龄组（岁）	年龄简化值（X）	尚存人数（Y）
40	0	81277
45	1	79258
50	2	76532
55	3	72850
60	4	67568
65	5	59911
70	6	50800
75	7	39325
80	8	28074

相关与回归的区别

- **1.意义：** 相关反映两变量的相互关系，即在两个变量中，任何一个的变化都会引起另一个的变化，是一种双向变化的关系。回归是反映两个变量的依存关系，一个变量的改变会引起另一个变量的变化，是一种单向的关系。
- **2.应用：** 研究两个变量的相互关系用相关分析。研究两个变量的依存关系用回归分析。
- **3.研究性质：** 相关是对两个变量之间的关系进行描述，看两个变量是否有关，关系是否密切，关系的性质是什么，是正相关还是负相关。回归是对两个变量做定量描述，研究两个变量的数量关系，已知一个变量值可以预测出另一个变量值，可以得到定量结果。
- **4.相关系数 r 与回归系数 b ：** r 与 b 的绝对值反映的意义不同。 r 的绝对值越大，散点图中的点越趋向于一条直线，表明两变量的关系越密切，相关程度越高。 b 的绝对值越大，回归直线越陡，说明当 x 变化一个单位时， y 的平均变化就越大。反之也是一样。

相关与回归的联系

关系：

能进行回归分析的变量之间存在相关关系。所以，对于两组新数据（两个变量）可先做散点图，求出它们的相关系数，对于确有相关关系的变量再进行回归分析，求出回归方程。

相关系数 r 与回归系数 b ：

r 与 b 的符号一致。 r 为正时， b 也为正，表示两变量是正相关，是同向变化。 r 为负时， b 也为负，表示两变量是负相关，是反向变化。 r 与 b 的假设检验结果一致，可用 r 的显著检验代替 b 的显著性检验。

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

Thank You !



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](http://www.medsci.cn)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。

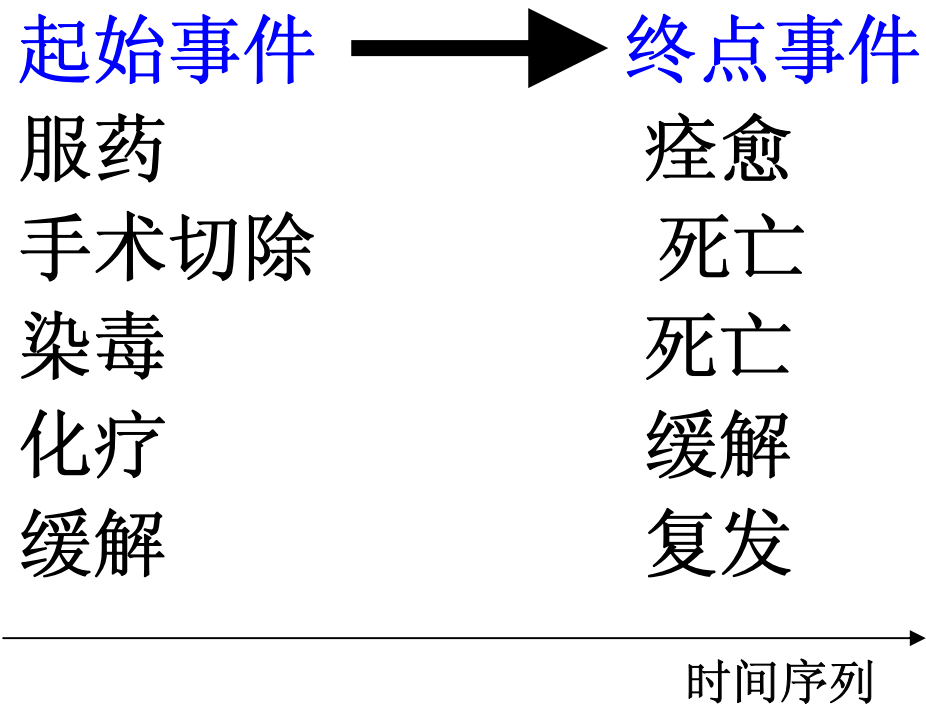
生存分析基本概念

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。

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

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

- 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。

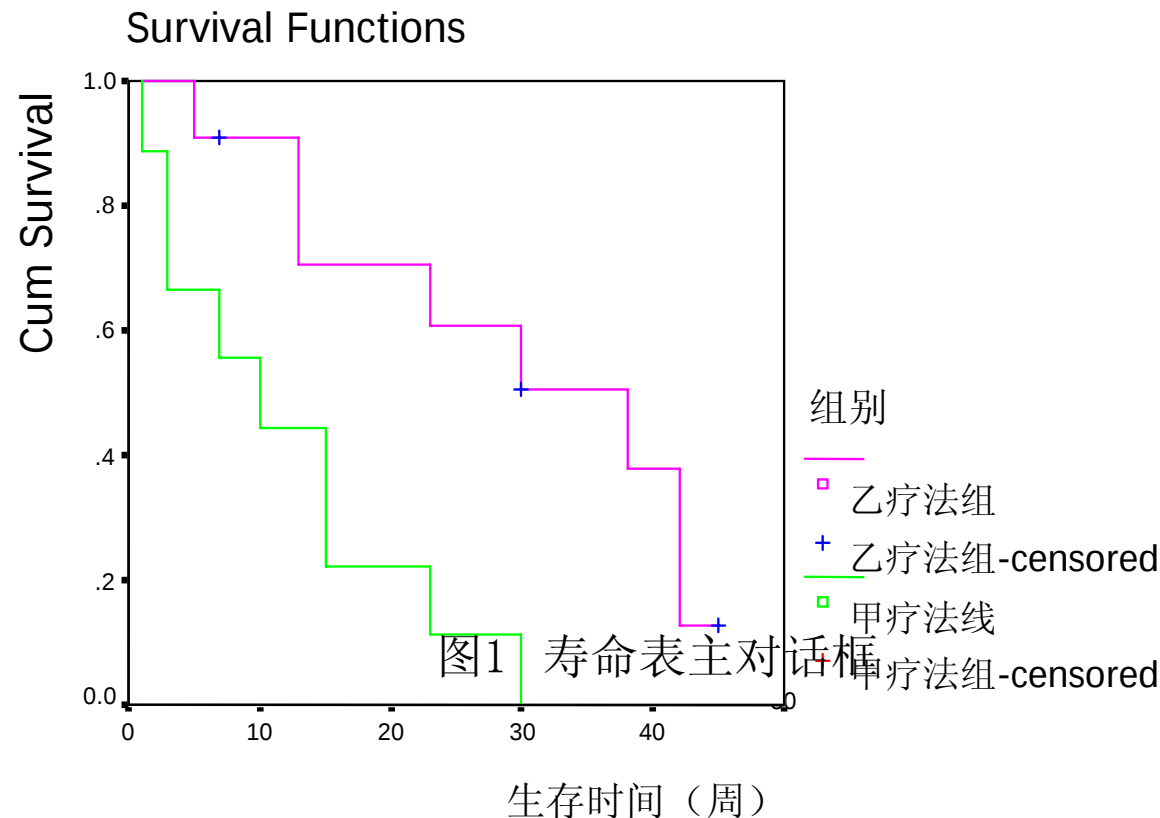
- 生存曲线是一条下降的曲线，分析时应注意曲线的高度和下降的坡度。平缓的生存曲线表示高生存率或较长生存期，陡峭的生存曲线表示低生存率或较短生存期。

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

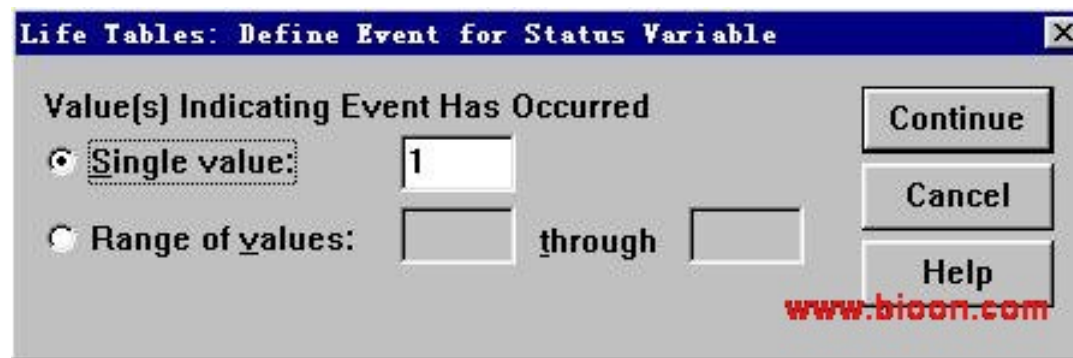
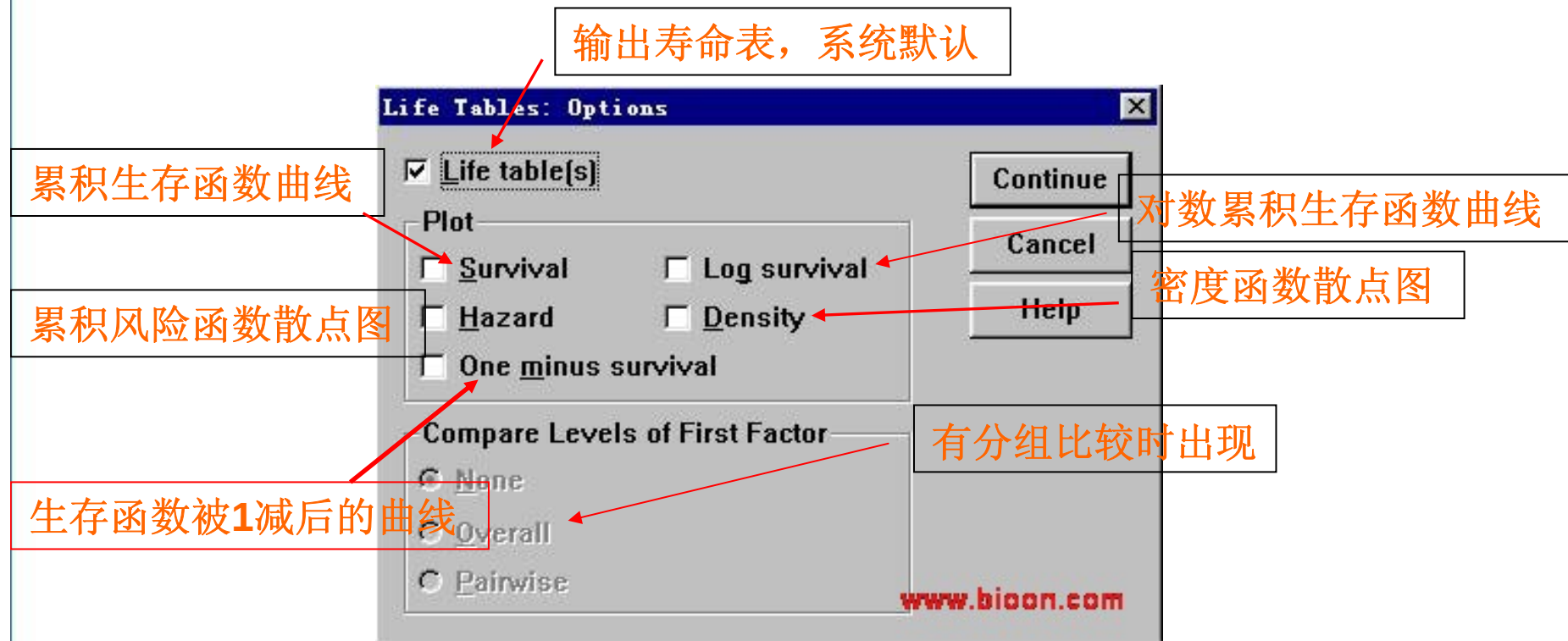


图2 定义终结事件标记值的对话框

- 击Options按钮，弹出选项对话框。



Kaplan-Meier过程

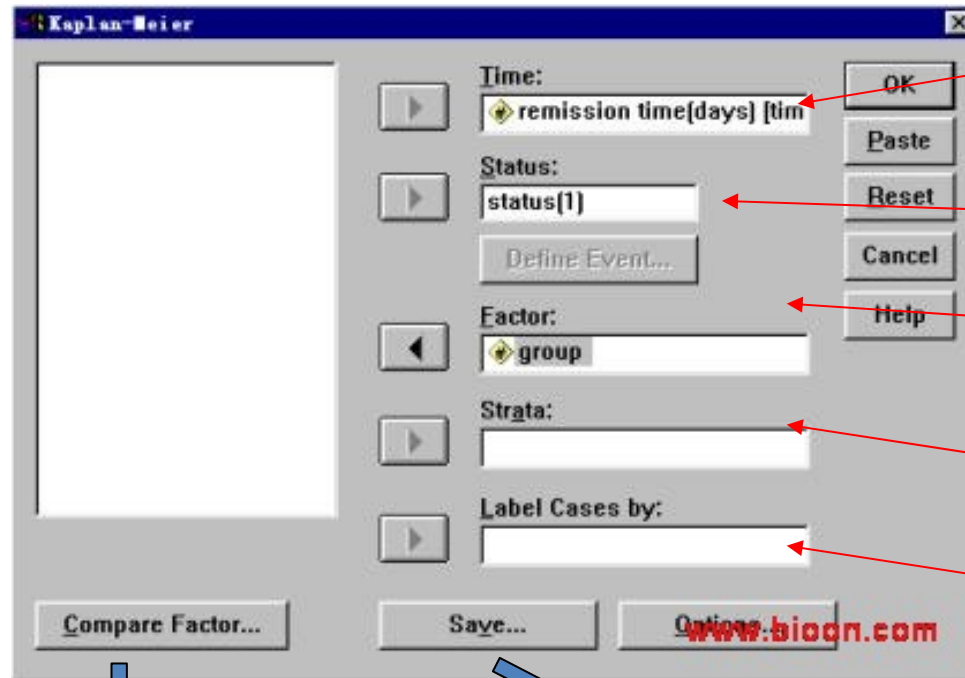
3种疗法治疗66例白血病患者的缓解时间（天）

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



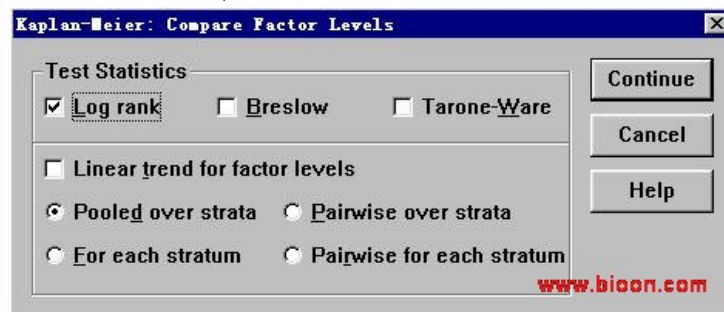
生存时间变量

生存状态变量

分组变量

分层变量

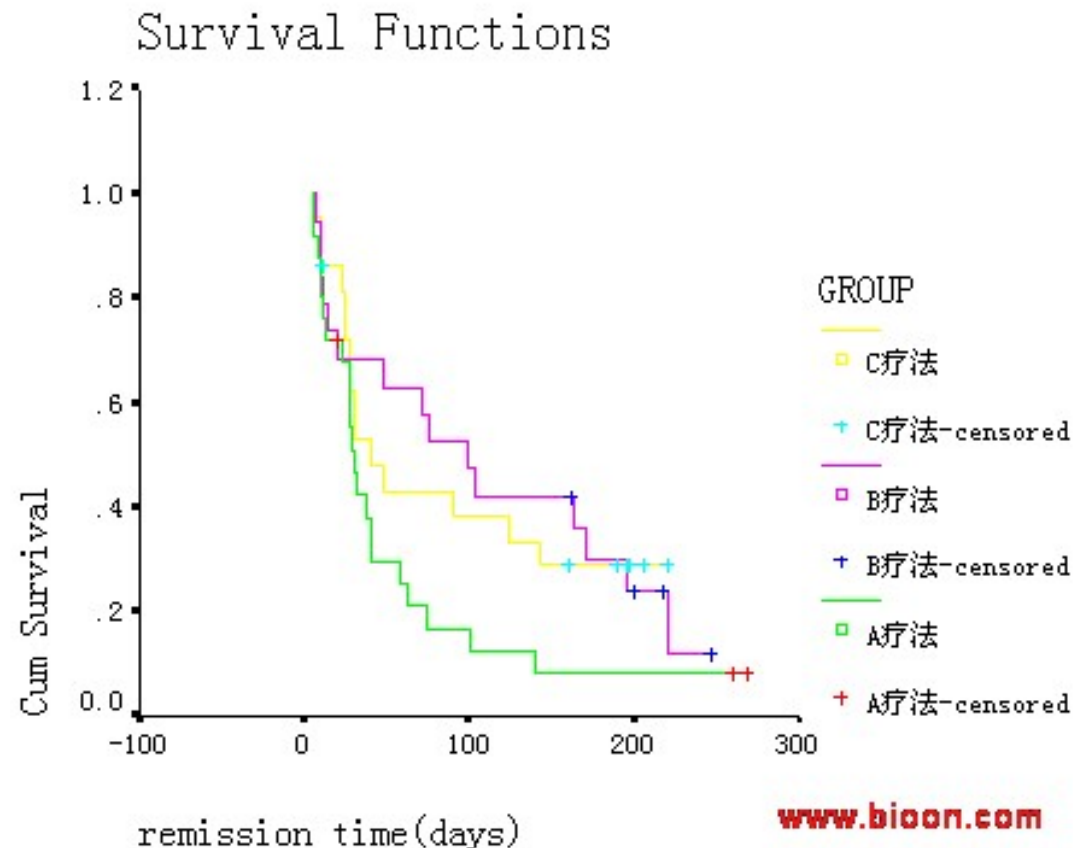
个体标记



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

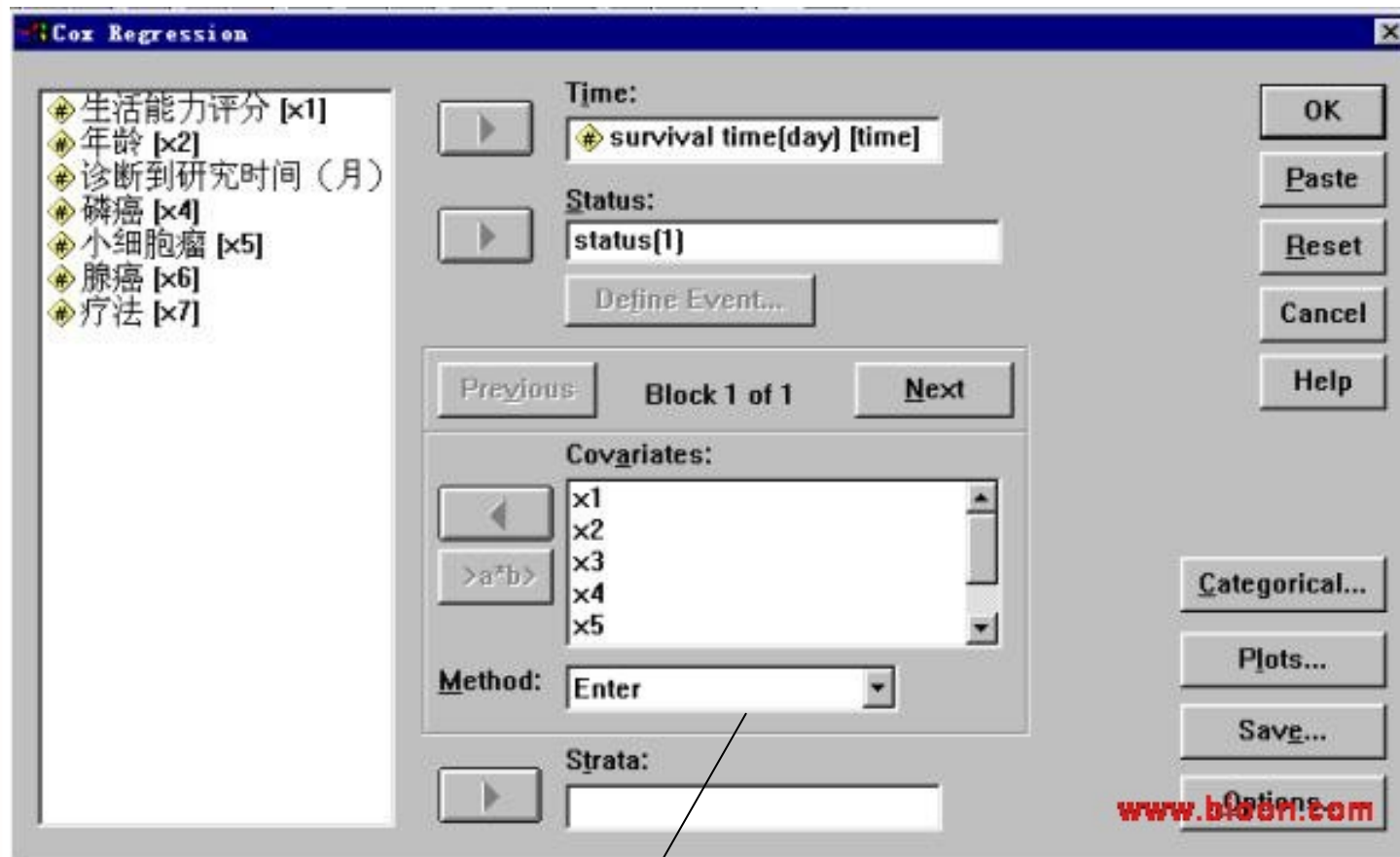
用以描述多个变量对生存时间的影响。此时可控制一个或几个因素，考察其他因素对生存时间的影响，及各因素之间的交互作用

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

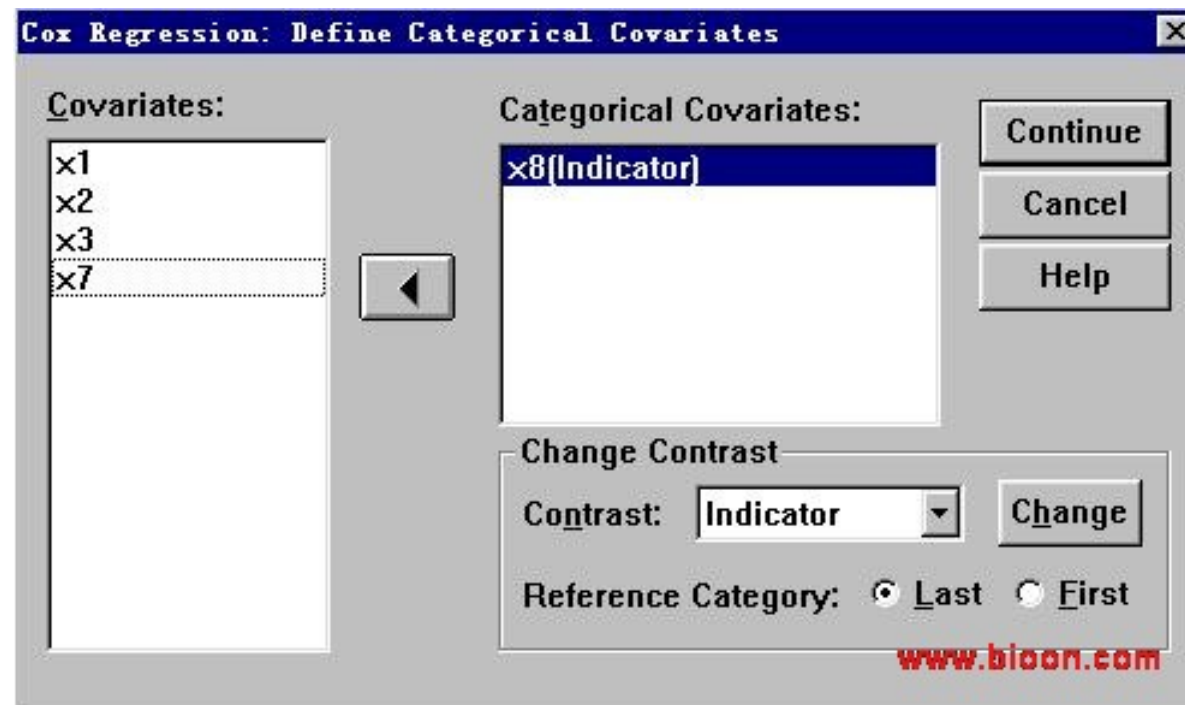
注：原数据库是用亚变量定义肺癌分类：0，0，0为其它癌；1，0，0为鳞癌；0，1，0为小细胞癌；0，0，1为腺癌。表中的最后一个变量是我加上去的癌症类别，1为鳞癌；2为小细胞癌；3为腺癌；4为其它癌。实践表明结果与用亚变量计算一样。

Cox回归主对话框

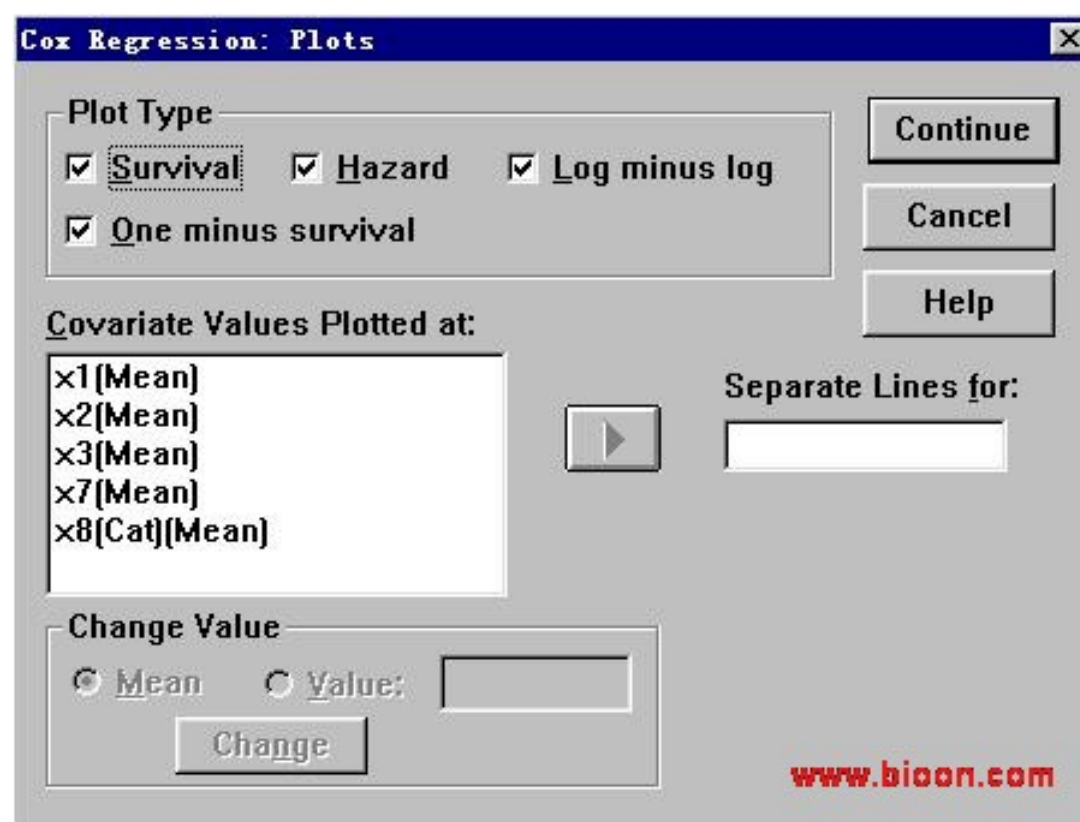


Enter: **Covariates**框内的全部变量均进入回归模型

- 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}

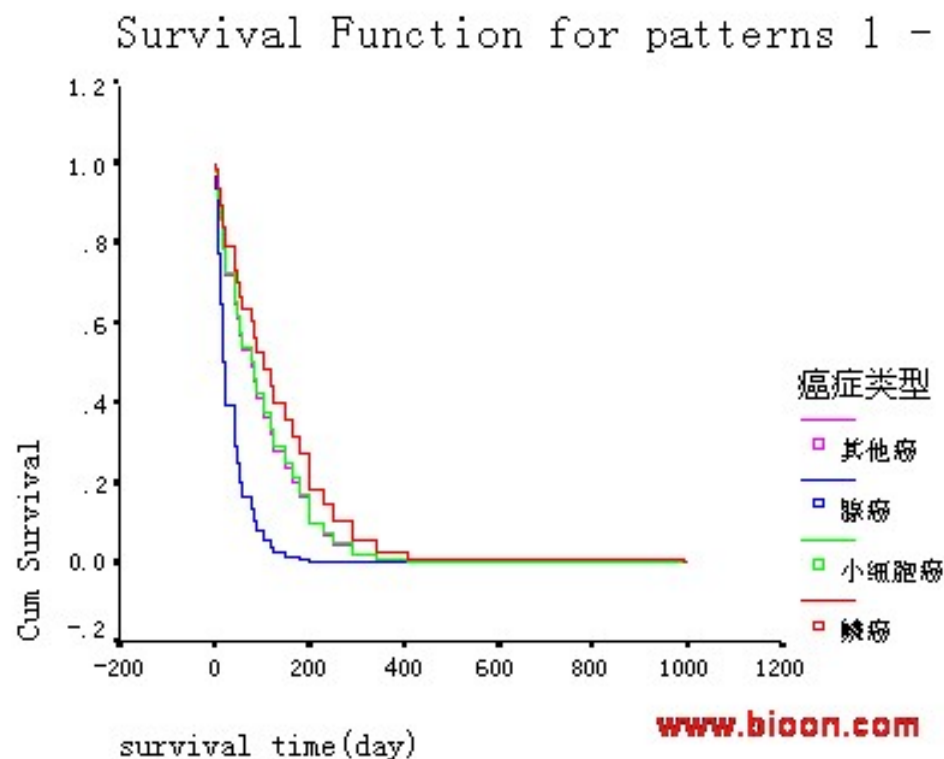
		Frequency	(1)	(2)	(3)
X8	1=鳞癌	14	1.000	.000	.000
	2=小细胞癌	11	.000	1.000	.000
	3=腺癌	5	.000	.000	1.000
	4=其他癌	10	.000	.000	.000

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

c:\spsswin\sp132.sav			
1:day	1		
	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

day

☐ Do not weight cases
 ☒ Weight cases by

◀

Frequency Variable:
death

Current Status: Do not weight cases

OK

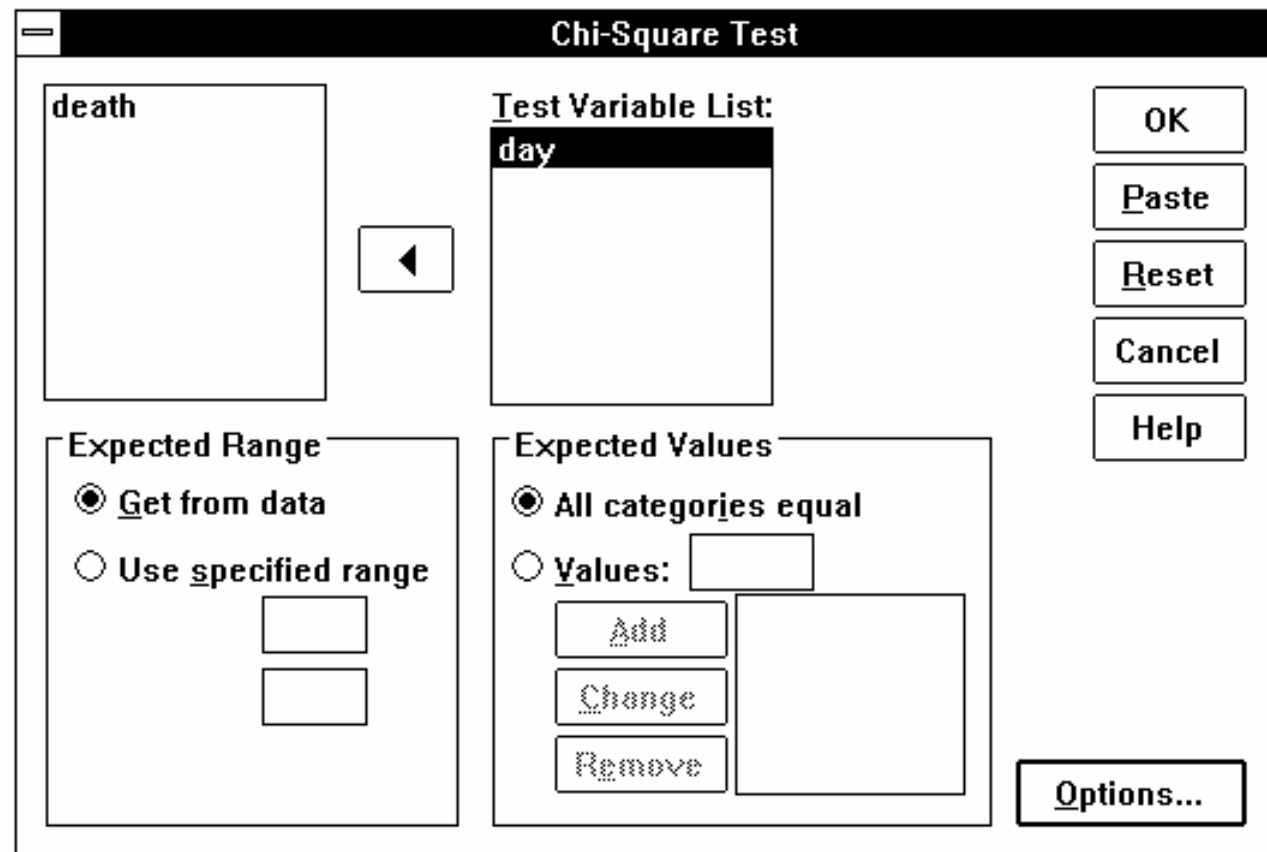
Paste

Reset

Cancel

Help

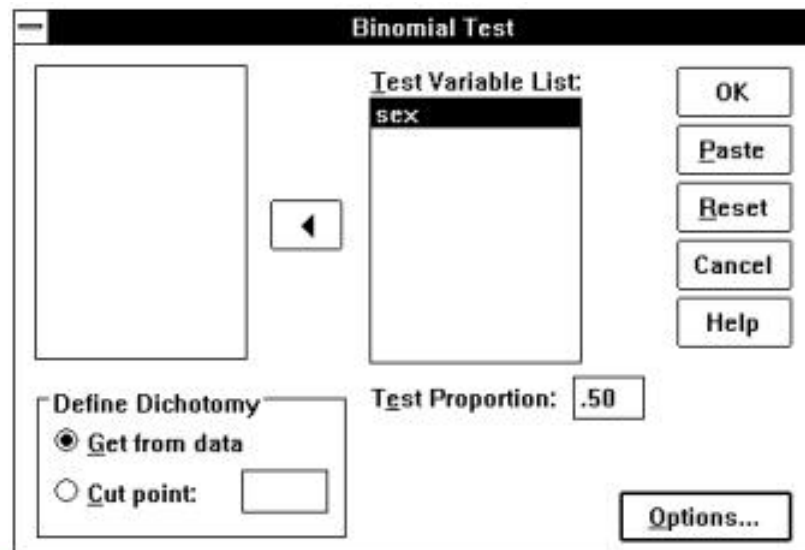
激活Statistics菜单选Nonparametric Tests中的Chi-Square...命令项，弹出Chi-Square Test对话框（图13.3）。现欲对一周内各日的死亡数进行分布分析，故在对话框左侧的变量列表中选day，点击按钮使之进入Test Variable List框，点击OK按钮即可。



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

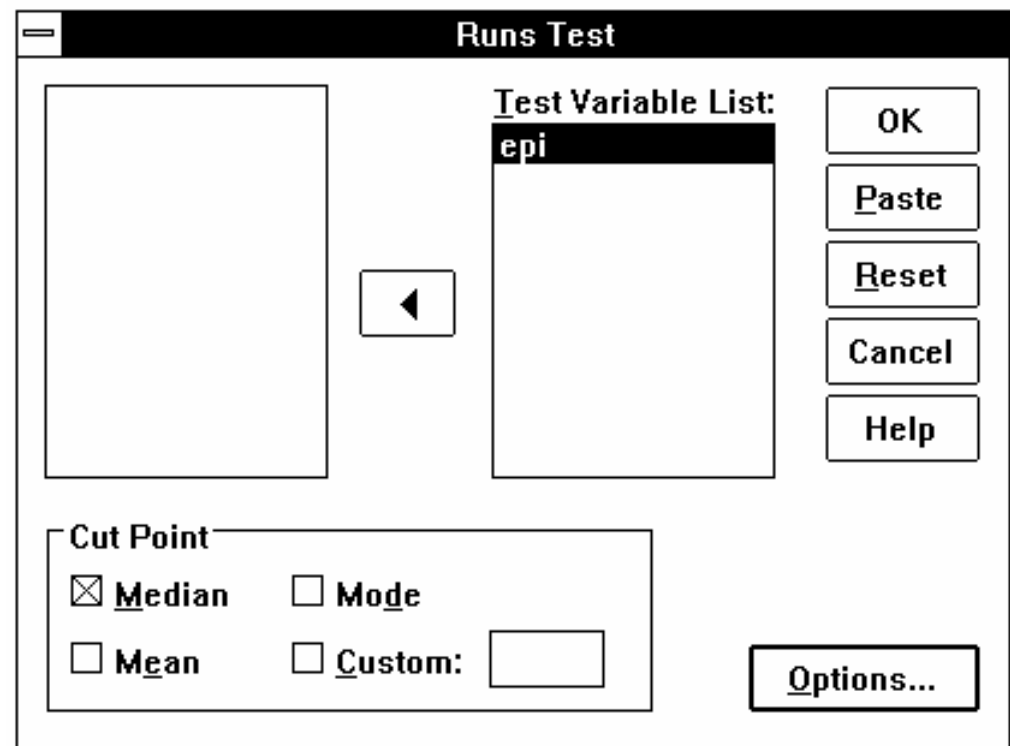


Runs（游程检验）过程

- 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			

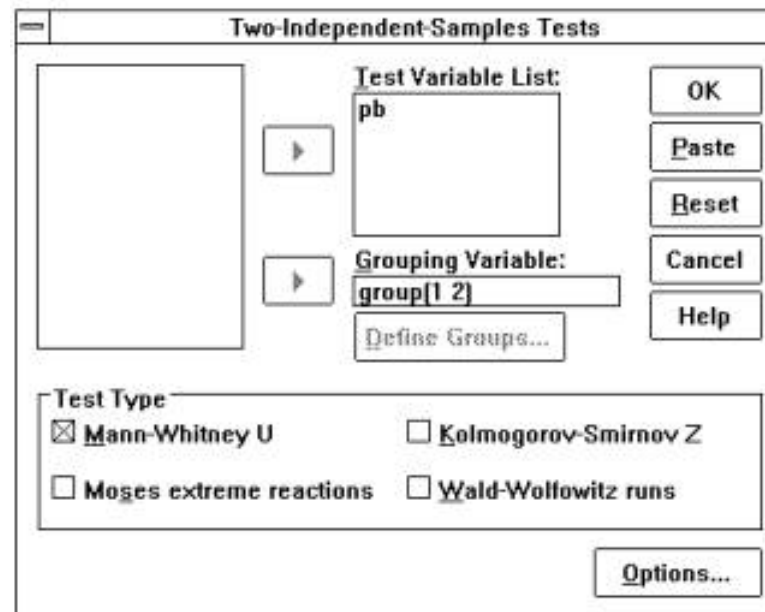


图13.7 两独立样本检验对话框

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

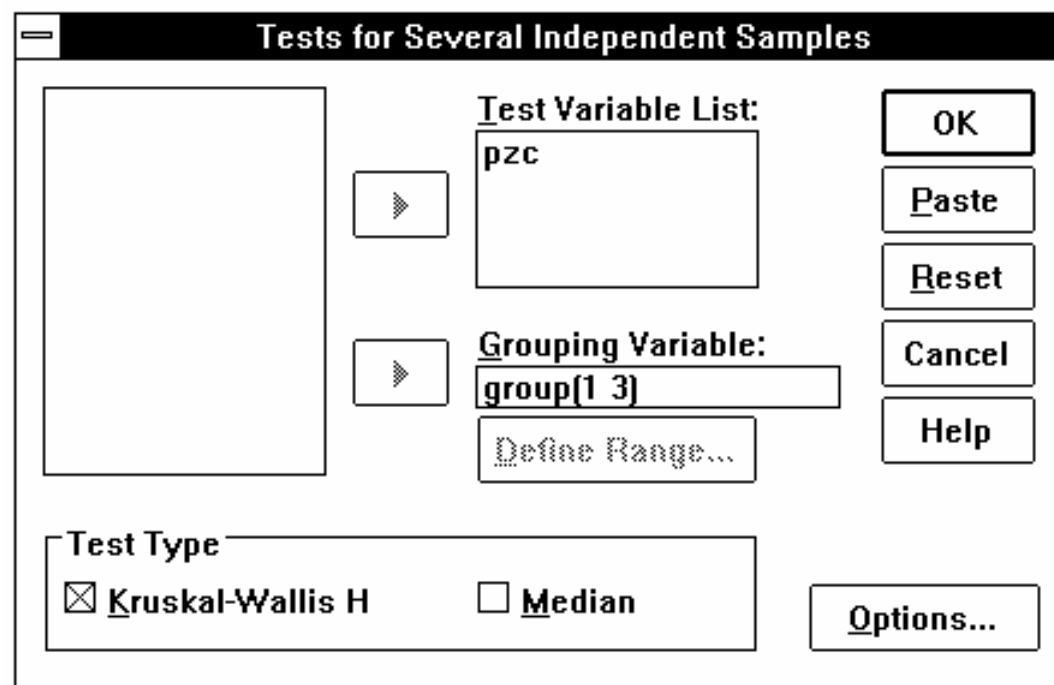


图13.8 多样本资料的秩和检验对话框

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

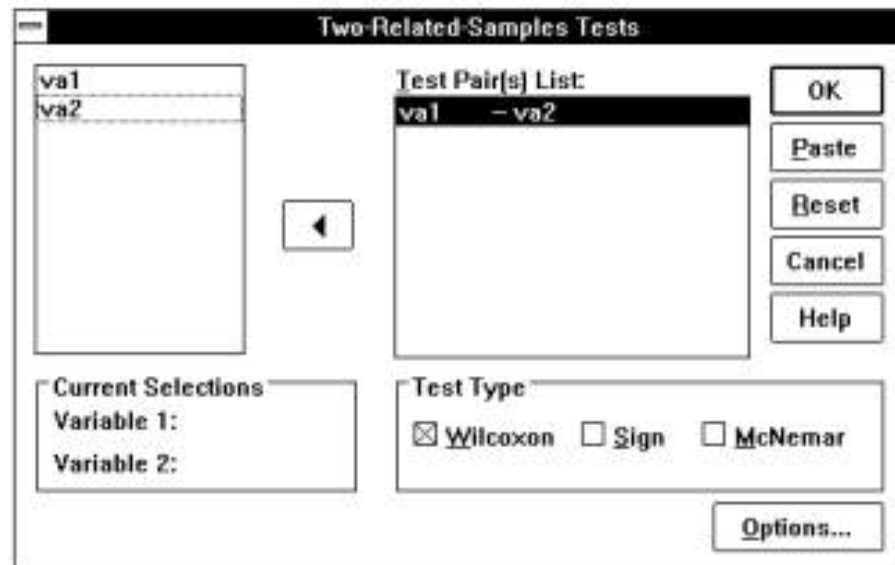


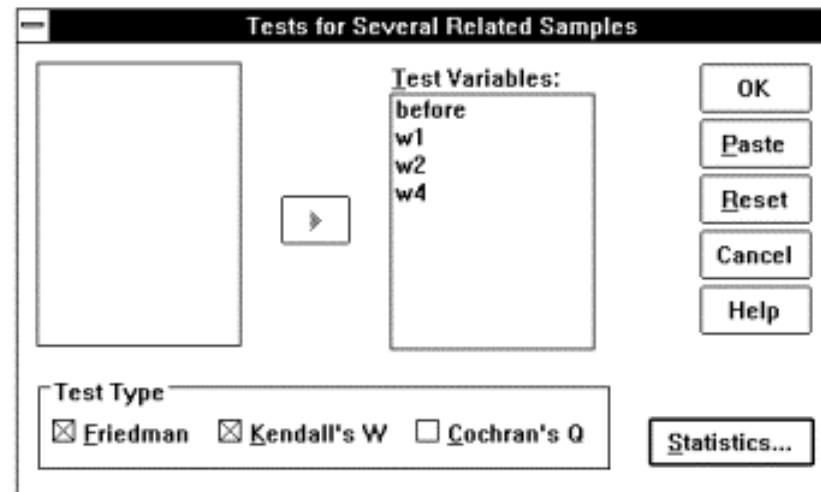
图13.9 两相关样本的秩和检验对话框

- 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

Thank You !

