中 国 计 量 学 院

《 信息分析技术与应用 》

实验指导书

二级学院（部、中 心）：经济与管理学院

学 科 （ 专 业 ）：管理科学与工程

课 程 名 称：信息分析技术与应用

授 课 对 象：11 级信息管理与信息系统本科生

授 课 教 师：虎陈霞

实验一 数据整理与文件编码

一、 实验目的

熟悉 SPSS 运行界面，了解 SPSS 的功能及系统的基本组成，掌握各窗口的构成功能及使用方法，能够进行数据文件的创建和编辑。

二、 实验环境

Windows 2003 操作系统，SPSS11.5 软件。

三、 实验要求

1. 充分熟悉 SPSS 运行界面，了解 SPSS 的功能及系统的基本组成，掌握各窗口的构成功能及使用方法，熟悉帮助菜单的应用；

2. 熟练掌握对数据文件的建立、编辑等的具体操作工作。能够根据实际分析需求，定义正确的变量类型；

3. 掌握将其他数据格式的文件，调用到 SPSS 中；

四、 实验内容和步骤

1. SPSS的启动，相关菜单各功能键的熟悉；
2. 定义以下变量名：Name、 Sex 、X1 、X2、 X3 、Y， 其列方向的字串或数值是该变量的数据、输入数据到SPSS Data Editor 数据表，最后保存到指定路径。

待输入的数据

Name	Sex	X 1	X2	X3	y
王军	男	35	69	0.70	1600
宁平	女	40	74	2.50	2600
彭会	男	42	64	2.00	2100
程可	男	40	65	3.25	3200
胡兵	男	37	72	1.10	2400
玛丽	女	45	68	1.50	2200
鲁萍	女	37	66	2.00	1600
罗阳	男	44	70	3.20	2750
陈铭	男	42	65	3.00	2500
沈丹	女	41	64	2.70	2400

3. 在 excel 中自建一个包含多种数据类型的文件，调用 excel 数据文件到 SPSS，观察有何异同？如何正确调入及定义相关变量？

实验二 数据的基本统计分析

一、 实验目的

掌握运用 SPSS 软件对数据进行描述性统计分析的方法及操作步骤。

二、实验环境

Windows 2003 操作系统，SPSS 软件。

三、实验要求

1. 独立完成从建立数据文件、基本分析过程的操作；

2. 掌握对数据进行描述性统计分析的方法及操作步骤。

四、实验内容和步骤

1. 调用指定数据文件 1-1，分析不同性别（sex）、不同年龄（age）和最高受教育年限（educ）各水平的频数分布情况及描述性统计，观察其异同；用探索性分析比较不同性别其受教育年限的异同。

2. 某医生用新药甲治疗十二指肠溃疡，以药物乙作对照组，问两种方法治疗效果有无差别？

处 理	愈 合	未愈合	合计
新药甲	54	8	62
药物乙	44	20	64
合 计	98	28	126

（提示：由于此处给出的直接是频数表，因此在建立数据集时可以直接输入三个

变量——行变量、列变量和指示每个格子中频数的变量，然后用 Weight Cases 对话框指定频数变量，最后调用 Crosstabs 过程进行 X^2 检验。）

结果解析：以实验内容 2 为例，在 Statistics 子选项中，选中 Chi-square 复选框，其他按系统默认，运行后输出结果如下：

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
疗效 * 药	126	100.0%	0	.0%	126	100.0%

记录缺失值情况报告，可见126例均为有效值。

疗效 * 药 Crosstabulation

Count

		药		Total
		甲	乙	
疗效	愈合	54	44	98
	未愈合	8	20	28
Total		62	64	126

上面为列出的四格表，反映了实际试验情况，可以检查变量定义、输入是否正确。

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.133[b]	1	.013		
Continuity Correction[a]	5.118	1	.024		
Likelihood Ratio	6.304	1	.012		
Fisher's Exact Test				.018	.011
Linear-by-Linear Association	6.084	1	.014		
N of Valid Cases	126				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.78.

上表检验结果从左到右为：检验统计量值(Value)、自由度(df)、双侧近似概率(Asymp.Sig.2-sided)、双侧精确概率(Exact Sig.2-sided)、单侧精确概率(Exact Sig.1-sided)；从上到下为：Pearson卡方（Pearson Chi-Square即常用的卡方检验）、连续性校正的卡方值(Continuity Correction)、对数似然比方法计算的卡方(Likelihood Ratio)、Fisher’s确切概率法(Fisher’s Exact

Test)、线性相关的卡方值(Linear by Linear Association)、有效记录数(N of Valid Cases)。另外，Continuity Correction和Pearson卡方值处分别标注有a和b，表格下方为相应的注解：a.只为2*2表计算。b.0%个格子的期望频数小于5，最小的期望频数为13.78。因此，这里无须校正，直接采用第一行的检验结果，即χ^2=6.133，P=0.013。

实验三 相关分析与回归分析

一、实验目的

运用 SPSS 软件进行相关分析和回归分析，并对结果进行解释。

二、实验环境

Windows 2003 操作系统，SPSS 软件。

三、实验要求

1. 掌握相关分析的概念及意义，并熟练掌握对数据进行相关分析的操作步骤；
2. 了解相关分析过程中各参数的意义及选择，以及结果阅读；
3. 充分理解相关与偏相关的涵义；
4. 掌握常用回归分析原理及操作步骤；
5. 掌握回归参数的正确选择及回归结果的科学解释

四、实验内容和步骤

1. 数据文件 3-1，调查了某地 1962 年～1988 年国民收入与城乡居民储蓄存款余额，利用相关分析分析二者的关系。并尝试用城乡居民储蓄存款余额预测国民收入。

2. 数据文件 3-2，调查了某公司员工当前工资，起始工资，工作经验及受教育年限，试分析四者之间的关系。尝试建立一个以起始工资，工作经验及受教育年限等为自变量，当前工资为因变量的回归模型。

3．数据文件 3-3，调查了松柏的生长情况，分析其月生长量与月平均气温、月降雨量、月平均日照时数、月平均湿度这四个气候因素哪个因素有关。

4．调用数据文件 3-4，分析 mpg（每加仑汽油行使里程）与 weight（车重）的相互关系，制作观测量数据的散点图，根据分析结果，选择并最终确定最佳回归模型。

5．根据某医院对癌症患者的调查数据 3-5，其中的变量包括年龄（age）、患病时间（time）、肿瘤扩散等级（pathscat）、肿瘤大小（pathssize）、肿瘤史（histgrad）和癌变部位的淋巴结是否含有癌细胞（ln_yesno），试建立相关模型，对癌变部位的淋巴结是否含有癌细胞的情况进行预测。

实验四 聚类分析与因子分析

一、实验目的

运用 SPSS 软件进行聚类分析和因子分析，并对结果进行解释。

二、实验环境

Windows 2003 操作系统，SPSS 软件。

三、实验要求

1. 了解聚类分析的目的及方法的选择；

2. 掌握不同聚类方法的应用前提及实现过程；

3. 充分了解不同的聚类算法和标准化方法对聚类结果的影响；

4. 熟悉因子分析的操作，理解其与聚类分析的区别；

5. 了解因子分析的标准和解释结果；

四、实验内容和步骤

1．某市选拔了 10 名游泳运动员，分别测取了 3 组身体指标，见数据文件 4-1。欲根据不同指标将这 10 名运动员分为 4 组分别进行训练，请问如何实现？

2．数据文件 4-2 调查了一批啤酒的成分及价格，要求根据啤酒各成分含量及价格对 20 种啤酒进行分类。

3. 调用数据文件 4-3，对所调查的数据进行因子分析，指定提取两个因子。

实验内容 1 结果解析：

Initial Cluster Centers

	Cluster			
	1	2	3	4
肩宽/髋宽×100	125.00	122.00	121.00	121.00
胸厚/胸围×100	20.00	18.00	17.00	19.00
腿长/身长×100	44.00	43.00	41.00	45.00

初始类中心

Iteration History[a]

	Change in Cluster Centers			
Iteration	1	2	3	4
1	.707	.354	.707	.707
2	.000	.000	.000	.000

a. Convergence achieved due to no or small distance change. The maximum distance by which any center has changed is .000. The current iteration is 2. The minimum distance between initial centers is 2.449.

各次迭代后类中心的变化

Cluster Membership

Case Number	编号	Cluster	Distance
1	1	1	.707
2	2	2	.791
3	3	3	.707
4	4	1	.707
5	5	2	.354
6	6	4	.707
7	7	3	.707
8	8	2	1.061
9	9	2	1.275
10	10	4	.707

各观测量的分类结果

Final Cluster Centers

	Cluster			
	1	2	3	4
肩宽/髋宽×100	124.50	121.75	120.50	120.50
胸厚/胸围×100	20.00	18.00	17.00	19.00
腿长/身长×100	44.50	42.75	41.50	44.50

聚类结果形成的 4 类的类中心的 4 个变量的值。

Number of Cases in each Cluster

Cluster	1	2.000
	2	4.000
	3	2.000
	4	2.000
Valid		10.000
Missing		.000

聚类结果每类中的观测量的数目。

实验内容 3 结果解析：

骰子点数

	Observed N	Expected N	Residual
1	43	50.0	-7.0
2	49	50.0	-1.0
3	56	50.0	6.0
4	45	50.0	-5.0
5	66	50.0	16.0
6	41	50.0	-9.0
Total	300		

观测量频数分布、期望值及残差

Test Statistics

	骰子点数
Chi-Square[a]	8.960
df	5
Asymp. Sig.	.111

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.

检验结果，$X^2=8.960$，$P=0.111>0.05$，与期望分布相符。