利用SPSS软件快速整理数据的六个步骤

作者：张华 赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/2113.html

利用SPSS软件快速整理数据的六个步骤。

每年的这个时候是毕业生最忙的时候，今年也不例外。这两天办公室来了很多咨询的准毕业生，不仅将各个咨询室占满，而且会议桌也围了一圈，创单日咨询数量新高。在咨询过程中，由于拿过来的数据格式不一，清理数据的时间占了整个咨询时间的80%-90%左右，如果能提前完成数据清理，将大大提高咨询效率。下面给大家一些清理数据的SOP，希望能帮助大家快速整理好数据。

大家拿过来的数据是不是这样的：

	A	B	C	D	E
1	病例号	姓名	性别	年龄	疾病史
2	110063	刘金秀	女	72	脑血管病
3	110175	刘恩平	男	83	心脏病
4	110143	付长春	男	73	高血压
5	110047	朱克平	女	66	高血压
6	110161	程永春	男	81	脑血管病
7	110048	徐焕才	男	75	心脏病
8	110123	刘桂莲	女	74	高血压
9	110068	杨富臣	男	77	高血压
10	110062	段汝能	女	80	脑血管病
11	110100	魏柏成	男	76	高血压
12	110058	甘伟碧	女	69	脑血管病
13	110099	倪敏强	男	80	高血压
14	110250	崔巍	女	88	脑血管病
15	110198	董光申	女	84	心脏病
16	110157	毛义很	女	76	脑血管病
17	110164	杨淑娴	女	85	脑血管病
18	110125	李玉恒	男	81	高血压
19	110190	魏志云	女		心脏病
20	110194	栾忠武	男		心脏病

group1 对照组 Sheet3

第一步：合并数据

。数据在不同的sheet里是不能分析的，要把所有的数据合并到一个sheet表里，在每个表里添加一个分组变量，就可以区别是哪一组了。合并的时候要注意把相同的变量，否则数据要全错了。

病例号	姓名	group	性别	年龄	疾病史
110063	刘金秀	1	女	72	脑血管病
110175	刘恩平	1	男	83	心脏病
0143	付长春	1	男	73	高血压
10047	朱克平	1	女	66	高血压
110161	程永春	1	男	81	脑血管病
110048	徐焕才	1	男	75	心脏病
110123	刘桂莲	1	女	74	高血压
110068	杨亩臣	1	男	77	高血压

第二步：给每个一病例一个唯一编码
，便于进行溯源，进行查找和更正错误数据。方法是：插入一列，变量名设为Id，前两个编号1、2，同时选中1、2，鼠标放在右下角，出现实线的“ 十 ”字时双击。

	A	B	C
1	id	病例号	姓名
2	1	110063	刘金秀
3	2	110175	刘恩平
4		0143	付长春
5		10047	朱克平
6		110161	程永春
7		110048	徐焕才
8		110123	刘桂莲

第三步：数据导入SPSS。可以通过“ 文件-打开-数据 ”，找到文件所在的路径，将文件类型选择excel格式，打开即可。最新版的SPSS支持直接把数据拖放到SPSS上打开哦。

第四步，更改变量名

。一般的数据软件只支持变量名是“ 英文 ”或者“ 英文+数字 ”形式，虽然高级版本的SPSS可以支持中文变量名，但在多因素分析中还会出现错误，因此建议更改变量名，并在标签中进行标注。

	名称	类型	宽度	小数位数	标签
1	id	数字	8	0	编码
2	casenum	数字	12	0	病例号
3	name	字符串	9	0	姓名
4	group	数字	12	1	分组
5	sex	字符串	3	0	性别
6	age	数字	12	1	年龄
7	history	字符串	12	0	疾病史
8					

第五步：查重

。一般的统计分析方法要求各个case间是独立的，因此数据不能有重复的case，如果一个研究对象有多次随访，也应合并到一行数据里。查重方法：“ 数据-标识重复个案 ”：

每个作为主个案的最后一个匹配个案的指示符

		频率	百分比	有效百分比	累计百分比
有效	重复个案	7	2.9	2.9	2.9
	主个案	231	97.1	97.1	100.0
	总计	238	100.0	100.0	

id	casenum	name
88	110045	李四妞
104	110045	王振基
4	110047	朱克平
95	110047	赵世海
6	110048	徐焕才
55	110048	姜建平
63	110049	宋殿林
89	110049	袁桂芳
114	110051	张淑兰
115	110051	白芙蓉
41	110054	李守云
87	110054	宋玉峰
53	110056	刘珍
72	110056	白翠霞

对于重复个案，查明原因，属于完全重复者可删除，属于不同随访时，合并到一行。

第六步：数据重新编码

。统计软件一般只能对“ 数 ”进行分析，因此文本数据应转成“ 数 ”据。方法：“ 转换-自动重新编码 ”，将性别选入右框，填写一个新的变量名，点击“ 添加新名称 ”后点击确定。

在输出页面可看到编码情况，在数据页面最后一列生成“ 数据 ”。

sex into sex2 (性别)

Old Value	New Value	Value Label
男	1	男
女	2	女

连续变量转成分组变量，也可以使用重新编码功能。如将年龄分成几组，操作方法：“ 转换-重新编码为不同变量 “，将年龄选入，填写新变量名称，点击“ 变化量 ”

再点击“ 旧值和新值 ”，在弹出的界面里，左侧为旧值范围，右侧为转换成的新值，如低于或等于45岁赋为1组，则

46-60岁赋为2组，则

如此依次，点“ 继续-确定“ ，在数据最后一列即可看到新变量。

小编建议按上述步骤依次整理，经过以上几步，基本可以把数据整理完成，形成一份可分析的数据。

更多 统计方法 请访问 https://www.iikx.com/news/statistics/