易侕数据清理操作指南

下载练习数据：http://www.empowerstats.com/empowerStats/exdata/dataclean_test.txt

操作步骤：

I. 第一步：将要清理的 EXCEL 或其它格式数据文件存成 Unicode Text 文件。

II. 第二步：选择要清理的数据文件，逐个点击列标题，浏览变量分布、选择清理操作。

III. 第三步：清理数据、查看与下载变量说明、浏览与下载数据。

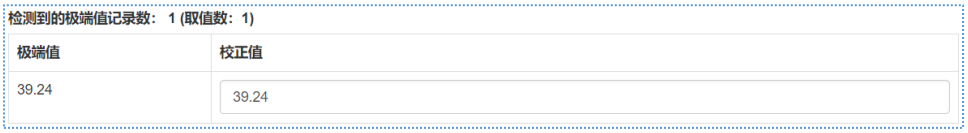
原始数据类型与清理方法：

1. **分类型字符数据**，如不孕类型。程序自动按频数排序，自动编码，必要时可修改编码。变量描述自动赋值为列标题，用户可以修改变量描述。点击“直接保存”。

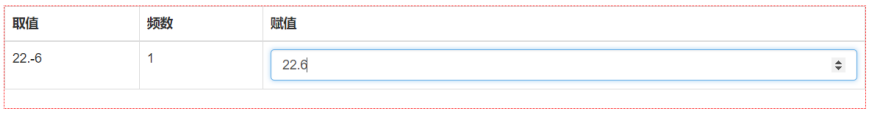
2. **数字型连续性数据**，如体重指数。程序自动统计变量分布：

N	Mean	SD	Min	P05	P10	P25	P50	P75	P90	P95	Max
205	22.70	4.51	16.61	18.70	19.12	20.20	22.04	24.61	27.20	28.25	39.24

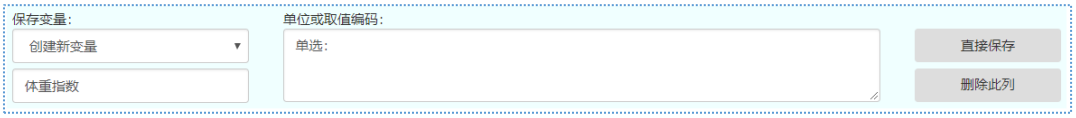
自动检测极端值并列出，用户可以对检出的极端值进行更正。如体重指数 39.24：



自动列出原始数据中错误输入的字符型数据，用户可以对之重新赋值。如体重指数 22.-6 赋值为 22.6：



对极端值与字符值（如有）进行更正与赋值后，修改变量描述（如需要），点击“直接保存”。



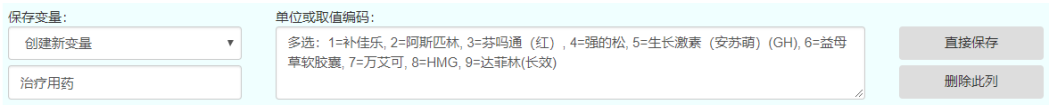
3. **文字型罗列名称**，如治疗用药。程序自动列出各种文字描述及其出现频数：

这类数据需转换成每种名称(药)是否使用的哑变量。选择分隔符，此例分隔符为“，”。如果有多种可能的分隔符，可以同时列出，中间用 OR（大小写不限）分隔，如“， or s” 表示逗号或空格分隔(S 表示空格，大小写不限)。点击“转成多选题”。

程序自动计算各名称(药)出现的频率，按频率排序，为每种名称(用药)生成一个哑变量。用户可以将相似或同种名称(药)的哑变量编号改成同一编号，以合并相似或同种名称(药)。然后点击“更新哑变量”。程序自动列出原始数据转换后各哑变量的赋值：

取值	频数	_1	_2	_3	_4	_5	_6	_7	_8	_9
补佳乐, 阿斯匹林, 芬吗通（红）	65	1	1	1	0	0	0	0	0	0
补佳乐	55	1	0	0	0	0	0	0	0	0
补佳乐, 阿斯匹林	49	1	1	0	0	0	0	0	0	0
补佳乐, 芬吗通（红）	26	1	0	1	0	0	0	0	0	0
补佳乐, 阿斯匹林, 强的松, 芬吗通（红）	3	1	1	1	1	0	0	0	0	0
生长激素（安苏萌）(GH), 补佳乐, 阿斯匹林, 芬吗通（红）	2	1	1	1	0	1	0	0	0	0
补佳乐, 阿斯匹林, 益母草软胶囊	2	1	1	0	0	0	1	0	0	0
HMG, 补佳乐	2	1	0	0	0	0	0	0	1	0
补佳乐, 万艾可, 阿斯匹林, 芬吗通（红）	2	1	1	1	0	0	0	1	0	0
补佳乐, 阿斯匹林, 强的松	1	1	1	0	1	0	0	0	0	0

修改变量描述（如需要），点击“保存哑变量”。



4. **文字描述型数据**，如移植胚胎评价。程序自动列出各种文字描述及其出现频数：

这类数据需要从文字中提取信息。通过给定起始与终止符选择要提取的信息，如要提取胚胎序号，起始符“胚胎序号：”，终止符为空格或逗号。变量描述改为胚胎序号，点击“提取信息”。

程序自动提取信息，并对提取出来的信息进行频数统计，如果提取出来的信息全为数字，自动按数字型变量处理子变量，如胚胎序号。

胚胎序号	频数	编码 更新编码
1	193	1
2	8	2
3	2	3
4	2	4

取值	频数	胚胎序号	编码 更新 保存
胚胎序号：1,2 胚胎质量：4BB×2	30	1	1
胚胎序号：1,2 胚胎质量：8C2×2	22	1	1
胚胎序号：1,2 胚胎质量：4BB×1,4BC×1	13	1	1
胚胎序号：1,2 胚胎质量：4BC×2	10	1	1
胚胎序号：1 胚胎质量：4BC×1	6	1	1
胚胎序号：1,2 胚胎质量：8C2×1,9C2×1	6	1	1
胚胎序号：1,2 胚胎质量：7C2×1,8C2×1	5	1	1

提取出胚胎序号后，点击“保存”。可以再提取胚胎质量信息（注：点击保存后，页面自动关闭该列窗口，可以再次点击列标题重新进入该窗口），设置起始与终止符为“x or ,”(注：其中的 x 符号需从数字中复制黏贴)，修改变量描述为胚胎质量，点击“提取信息”：

胚胎质量	频数	编码 更新编码
4BB	62	0
8C2	44	1
4BC	23	2
7C2	20	3
9C2	11	4
6C2	8	5
10C2	7	6

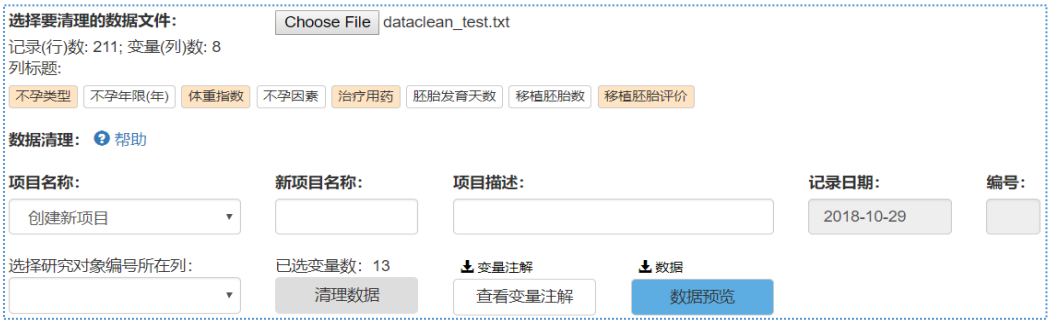
取值	频数	胚胎序号	编码 删除	胚胎质量	编码 更新 保存
胚胎序号：1,2 胚胎质量：4BB×2	30	1	1	4BB	0
胚胎序号：1,2 胚胎质量：8C2×2	22	1	1	8C2	1
胚胎序号：1,2 胚胎质量：4BB×1,4BC×1	13	1	1	4BB	0
胚胎序号：1,2 胚胎质量：4BC×2	10	1	1	4BC	2
胚胎序号：1 胚胎质量：4BC×1	6	1	1	4BC	2
胚胎序号：1,2 胚胎质量：8C2×1,9C2×1	6	1	1	8C2	1
胚胎序号：1,2 胚胎质量：7C2×1,8C2×1	5	1	1	7C2	3

如提取出来的信息含文字，程序自动对信息进行编码，用户可以对个别无法通过规定起始符与终止符的文字记录手动输入提取的信息与编码。

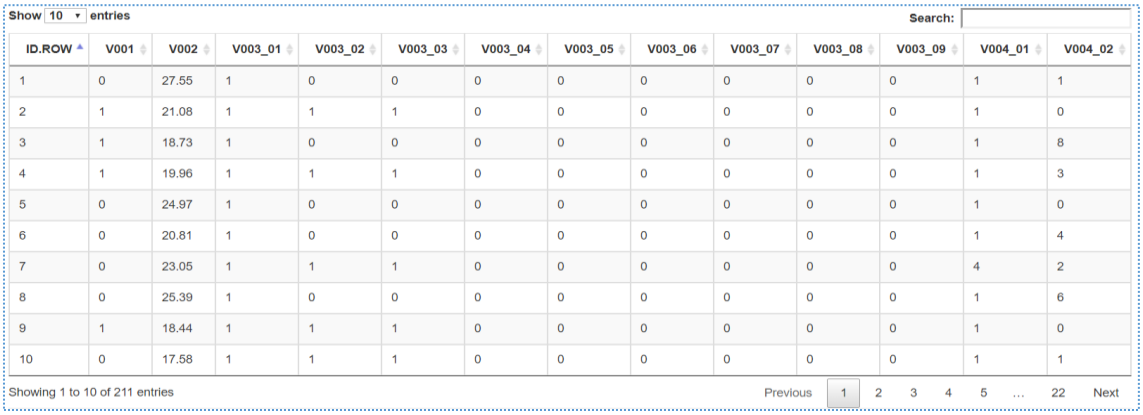
点击“保存”后保存胚胎质量信息。

其它操作：

1. 当有已清理并保存的列标题后，“清理数据”按键才被激活，已保存的列标题通过背景颜色与其它未清理的区别，输出的变量数也在“清理数据”按键上方显示。



2. 点击“清理数据”后，即可浏览数据如：



Show 10 entries Search:

ID.ROW	V001	V002	V003_01	V003_02	V003_03	V003_04	V003_05	V003_06	V003_07	V003_08	V003_09	V004_01	V004_02
1	0	27.55	1	0	0	0	0	0	0	0	0	1	1
2	1	21.08	1	1	1	0	0	0	0	0	0	1	0
3	1	18.73	1	0	0	0	0	0	0	0	0	1	8
4	1	19.96	1	1	1	0	0	0	0	0	0	1	3
5	0	24.97	1	0	0	0	0	0	0	0	0	1	0
6	0	20.81	1	0	0	0	0	0	0	0	0	1	4
7	0	23.05	1	1	1	0	0	0	0	0	0	4	2
8	0	25.39	1	0	0	0	0	0	0	0	0	1	6
9	1	18.44	1	1	1	0	0	0	0	0	0	1	0
10	0	17.58	1	1	1	0	0	0	0	0	0	1	1

Showing 1 to 10 of 211 entries Previous 1 2 3 4 5 … 22 Next

3. 点击“查看变量注解”：

变量名	取值编码	意义
ID.ROW		Row Number
V001		不孕类型
	0	原发不孕
	1	继发不孕
V002		体重指数 (kg/m2)
V003_01		治疗用药: 补佳乐
	0	No
	1	Yes
V003_02		治疗用药: 阿斯匹林
	0	No

4. 点击 变量注解 下载变量注解， 数据 下载清理后的数据。
5. 点击列标题，可以查看或关闭相应变量操作窗口。
6. 点击“删除此列”删除已保存的列。
7. 点击列变量窗口右边的下三角 图标关闭该窗口。
8. 点击列变量窗口右边的下三角 图标重置该变量（清除对该列前面所有的操作）。