一维频数表和卡方检验

一维频数表是分类数据最简单的形式，用于展示样本中某变量在不同分类值中的分布。例如，参加调查的男性和女性人数，不同种族的人数。

卡方检验

卡方检验可用于比较实际观测频数和理论频数是否相同。例如，如果 SNP1 具有“A”等位基因的频率是 90%，则理论基因型频数为：AA=0.81，AB=0.18，BB=0.01。具有 AA 基因型的期望人数是 0.81×N，AB 基因型是 0.18×N，BB 基因型是 0.01×N

X^2值的计算公式为： $X^2=\Sigma (O_i - E_i)^2/E_i$

其中：O_i = 实际观察频数；E_i = 理论期望频数 ；n = 表中格子数

X^2 = Pearson's 统计值，近似服从X^2 分布

如果X^2值的概率小于 5%，认为差异有显著性，则拒绝无效假设，说明样本频数分布与理论分布有显著差异。

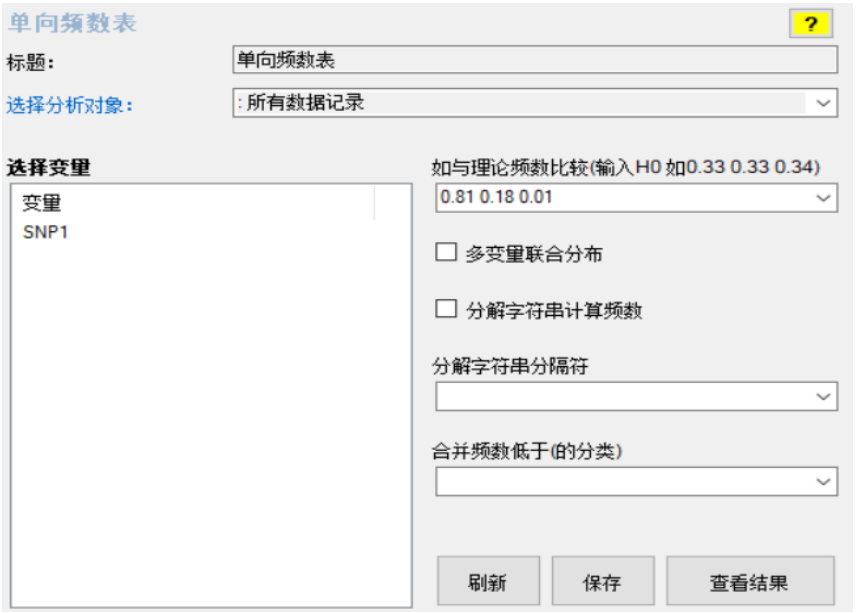
多变量联合分布的频数表：

当输入多个变量时，系统首先计算每个变量的单向频数分布，并绘出相应的构成图。如果，选择了“变量联合分布”，系统再计算多变量联合的频数分布。所谓联合分布，指几个变量取值的各种组合对应的频数，如例 2。

文字描述型变量频数统计：

本模块可以用于文字描述型记录的分析与处理。首先对原始记录进行字符分解（定义分隔符），统计子字段频率，再按子字段频率排序，频率最高者排在前面，按顺序为每个子字段生成一个（0/1）两分类变量，输出相应的数据文件与变量注解文件，如例 3。

例 1，打开练习项目 DEMO，SNP1 基因型分布与理论分布是否有显著差异，输入界面：
（或下载练习数据：www.empowerstats.com/empowerStats/exdata/demo1.xls）



当输入理论频数后，软件自动列出 SNP1 的基因型与假定的理论频数。

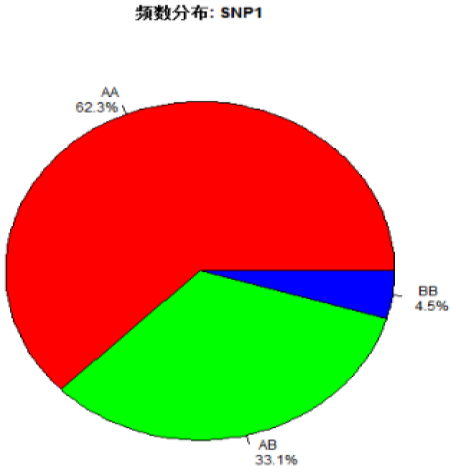
输出结果：

单向频数表：

SNP1	观察数	观察频率	理论频率	期望数
AA 62.3%	508	0.6233	0.81	660.15
AB 33.1%	270	0.3313	0.18	146.70
BB 4.5%	37	0.0454	0.01	8.15

卡方检验：

卡方值	自由度	p 值
240.8252	2	<0.000001



例 2，打开练习项目 DEMO，分析咳嗽、咳痰、气喘、气短联合分布频数，输入界面：

输出结果：

单向频数表：

	N	频率(prop)	95%CI low	95%CI upp	P.value(H0: prop=0.5)
COUGH					
no 86.4%	713	0.8642	0.8385	0.8865	<0.000001
yes 13.6%	112	0.1358	0.1135	0.1615	<0.000001
PHLEGM					
no 81.9%	676	0.8194	0.7910	0.8447	<0.000001
yes 18.1%	149	0.1806	0.1553	0.2090	<0.000001
WHEEZE					
no 87.6%	723	0.8764	0.8515	0.8977	<0.000001
yes 12.4%	102	0.1236	0.1023	0.1485	<0.000001
Shortness of breath					
no 71.0%	586	0.7103	0.6778	0.7408	<0.000001
yes 29.0%	239	0.2897	0.2592	0.3222	<0.000001

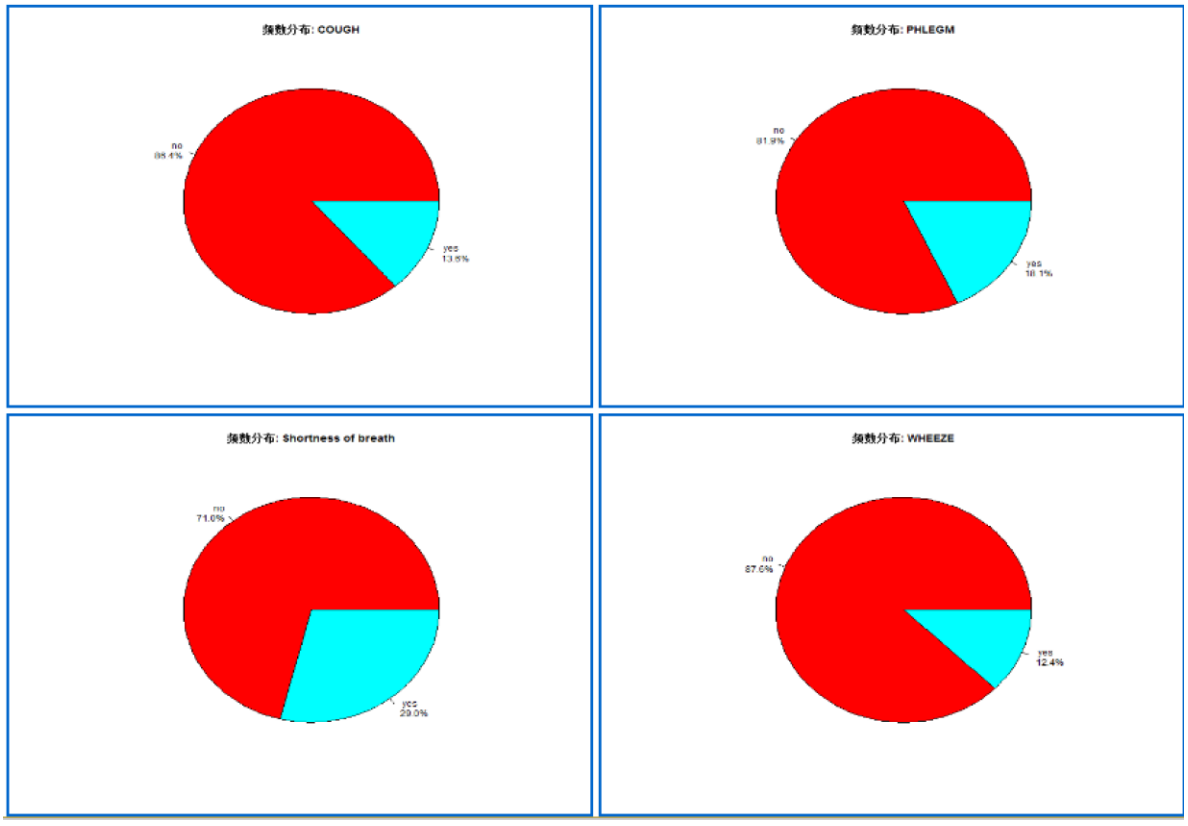
1 sample proportions test with continuity correction

多变量联合频数分布:

COUGH	PHLEGM	WHEEZE	Shortness of breath	N	频率
0	0	0	0	502	0.603365
0	0	0	1	113	0.135817
0	0	1	0	17	0.020433
0	0	1	1	25	0.030048
0	1	0	0	28	0.033654
0	1	0	1	17	0.020433
0	1	1	0	4	0.004808
0	1	1	1	7	0.008413
1	0	0	0	8	0.009615
1	0	0	1	6	0.007212
1	0	1	0	1	0.001202
1	0	1	1	4	0.004808
1	1	0	0	20	0.024038
1	1	0	1	29	0.034856
1	1	1	0	6	0.007212
1	1	1	1	38	0.045673
NA	NA	NA	NA	7	0.008413

多变量联合频数分布(除去缺失值):

COUGH	PHLEGM	WHEEZE	Shortness of breath	N	频率
0	0	0	0	502	0.608485
0	0	0	1	113	0.136970
0	0	1	0	17	0.020606
0	0	1	1	25	0.030303
0	1	0	0	28	0.033939
0	1	0	1	17	0.020606
0	1	1	0	4	0.004848
0	1	1	1	7	0.008485
1	0	0	0	8	0.009697
1	0	0	1	6	0.007273
1	0	1	0	1	0.001212
1	0	1	1	4	0.004848
1	1	0	0	20	0.024242
1	1	0	1	29	0.035152
1	1	1	0	6	0.007273
1	1	1	1	38	0.046061



例 3，下载练习数据：http://www.empowerstats.com/empowerStats/exdata/dataclean.txt
分析不孕因素（V4）频数分布，输入界面：

输出结果：

字段分解成：19 不同的子字段
生成：19 个 0/1 新变量

子字段频数分布：

Var.NAME	Substring	Frequency	%
V4.001	输卵管因素	136	48.40
V4.002	NA	37	13.17
V4.003	宫外孕术后	18	6.41
V4.004	排卵障碍	16	5.69
V4.005	多囊卵巢综合症	15	5.34
V4.006	IUI 术后	11	3.91
V4.007	其他	10	3.56
V4.008	腹腔镜术后	9	3.20
V4.009	卵巢功能低下	6	2.14
V4.010	宫腔镜术后	5	1.78
V4.011	卵巢囊肿术后	4	1.42
V4.012	子宫内膜因素	3	1.07

Var.NAME	Substring	Frequency	%
V4.013	不良孕史	2	0.71
V4.014	单角子宫	2	0.71
V4.015	慢性盆腔炎	2	0.71
V4.016	子宫内膜异位症	2	0.71
V4.017	不明原因不孕症	1	0.36
V4.018	多囊卵巢	1	0.36
V4.019	免疫性不孕症	1	0.36

新生成的变量存放在数据文件：dataclean_1_tbl1.xls

右击输出文件.htm 可看到同时输出有变量注解文件：dataclean_1_tbl1_variables.xls