举例一：想知道喝牛奶对感冒发病率有没有影响。

	感冒人数	未感冒人数	合计	感冒率
喝牛奶组	43	96	139	30.94%
不喝牛奶组	28	84	112	25.00%
合计	71	180	251	28.29%

喝牛奶组和不喝牛奶组的感冒率为30.94%和25.00%，两者的差别可能是抽样误差导致，，也可能是 牛奶对感冒率真的有 影响。

H

下面就就行假设了：假设喝牛奶对感冒发病率没有影响，即喝牛奶与感冒无关

所以感冒的发病率实际是（43+28）/(43+28+96+84)=28.29%

所以可以得到理论的表格

	感冒人数	未感冒人数	合计
喝牛奶组	=139*0.2829	=139*(1-0.2829)	139
不喝牛奶组	=112*0.2829	=112*(1-0.2829)	112

即下表:

	感冒人数	未感冒人数	合计
喝牛奶组	39.3231	99.6769	139
不喝牛奶组	31.6848	80.3152	112
合计	71	180	251

如果说真的没有影响的话 表格中理论值和实际值差别应该会很小。

卡方检验的计算公式

$$x^2 = \sum \frac{(A-T)^2}{T}$$

$$\frac{(43-39)^2}{39}+\frac{(96-99)^2}{99}+\cdots\cdots$$

其中:A是实际值，T为理论值

X^2值的意义:衡量理论与实际的差异程度。

经过计算可以计算得到

X^2=1.077

举例二：某个城市在某一时期内共发生交通事故600次，按不同颜色小汽车分类如下

汽车颜色	红	棕	黄	白	灰	蓝
事故次数	75	125	70	80	135	115

问：交通事故是否与汽车的颜色有关?

解：如果交通事故与汽车的颜色无关，则每种颜色的小汽车发生交通事故的可能性是一样的。即

$$P_i(\text{汽车颜色})=\frac{1}{6}，\ i=1,2,\cdots,6$$

下面计算统计量

$$\chi^2=\sum_{i=1}^{n}[\frac{(f_i-F_i)^2}{F_i}]$$

汽车颜色	f_i	P_i	$F_i=nP_i$	f_i-F_i	$\frac{(f_i-F_i)^2}{F_i}$
红	75	1/6	100	-25	6.25
棕	125	1/6	100	25	6.25
黄	70	1/6	100	-30	9
白	80	1/6	100	-20	4
灰	135	1/6	100	35	12.25
蓝	115	1/6	100	15	2.25
Σ	n=600				40

上表计算出  =40，取α=0.05，因没有估计参数，则自由度为6-0-1=5，查表得，因此我们可以认定交通事故与小汽车的颜色有关。

$$z_{0.05} = 155$$

$$\chi^2$$

$$I(X,Y) = H(X) - H(X|Y)$$

姓名	汽车颜色	是否发生事故
A	0	1
B	1	1
C	1	0

皮尔逊相关系数和余弦相似度是计算属性之间的相关性,删除相关性大的列(降维)

卡方检测是分析属性和分类目标有无相关性,把相关性大的选出(特征选择)