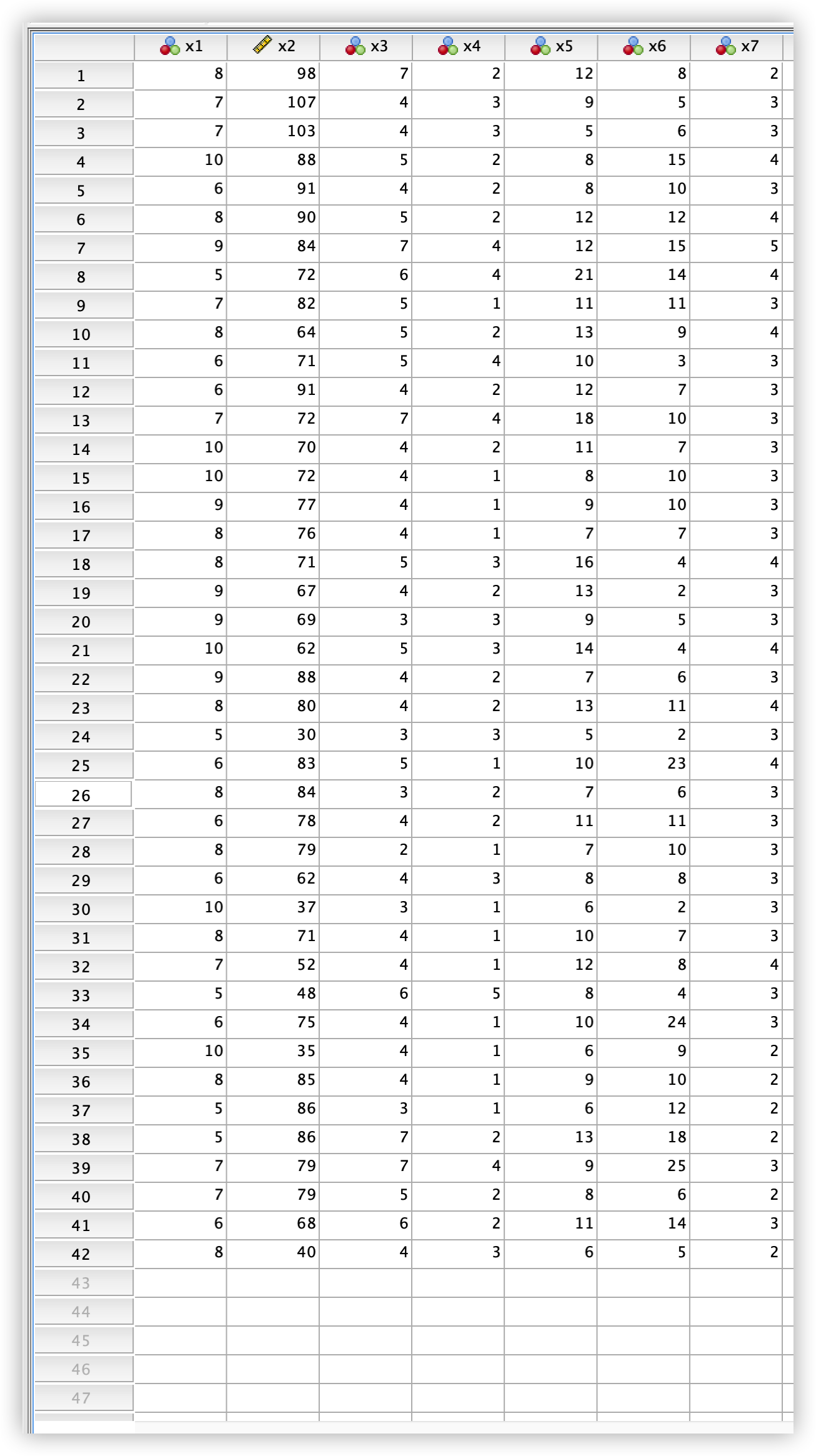
**主成分分析**

**1202110887 徐鸿飞**

**题目**（教材P154第7题）：下表是某城市再42天中午12点空气污染数据， 试分别利用样本协方差矩阵和样本相关矩阵进行主成分分析，二者的结果有何差异？原始数据的变化可否由三个或者更少的主成分反映，你能否给所选取的主成分作出解释？



主成分分析旨在利用降维的思想，把多指标转化为少数几个综合指标。在实际问题研究中，为了系统、全面地分析问题，我们必须考虑众多影响因素。因为每个因素都在不同程度上反映了所研究问题的某些信息，并且指标之间有一定的相关性，因而所得到的统计数据反映的信息在一定程度上有重叠。本文中所研究的问题变量较多，因此利用主成分分析法研究本问题，减少计算量和降低分析问题的复杂性。

针对问题一，首先将数据标准化，计算样本协方差矩阵和相关矩阵，然后分别计算样本协方差矩阵和相关矩阵的特征值和特征向量，贡献率和累计贡献率，确定选取成分个数，列出主成分方程并解释主成分意义。

针对问题二，考虑主成分的贡献率，只要主成分的累计贡献率达到80%， 就可以反映原始数据的变化，并且对所选取的主成分做出解释。

-般来说，由协方差矩阵出发求解主成分所得结果与由相关矩阵出发求解主成分所得结果有很大不同。对于度量单位不同的指标或取值范围彼此差异非常大的指标，不能直接由其协方差矩阵出发进行主成分分析，而应该考虑将数据标准化(从相关性矩阵出发)。对于同度量单位或取值范围在同量级的数据还是直接从协方差矩阵求解。

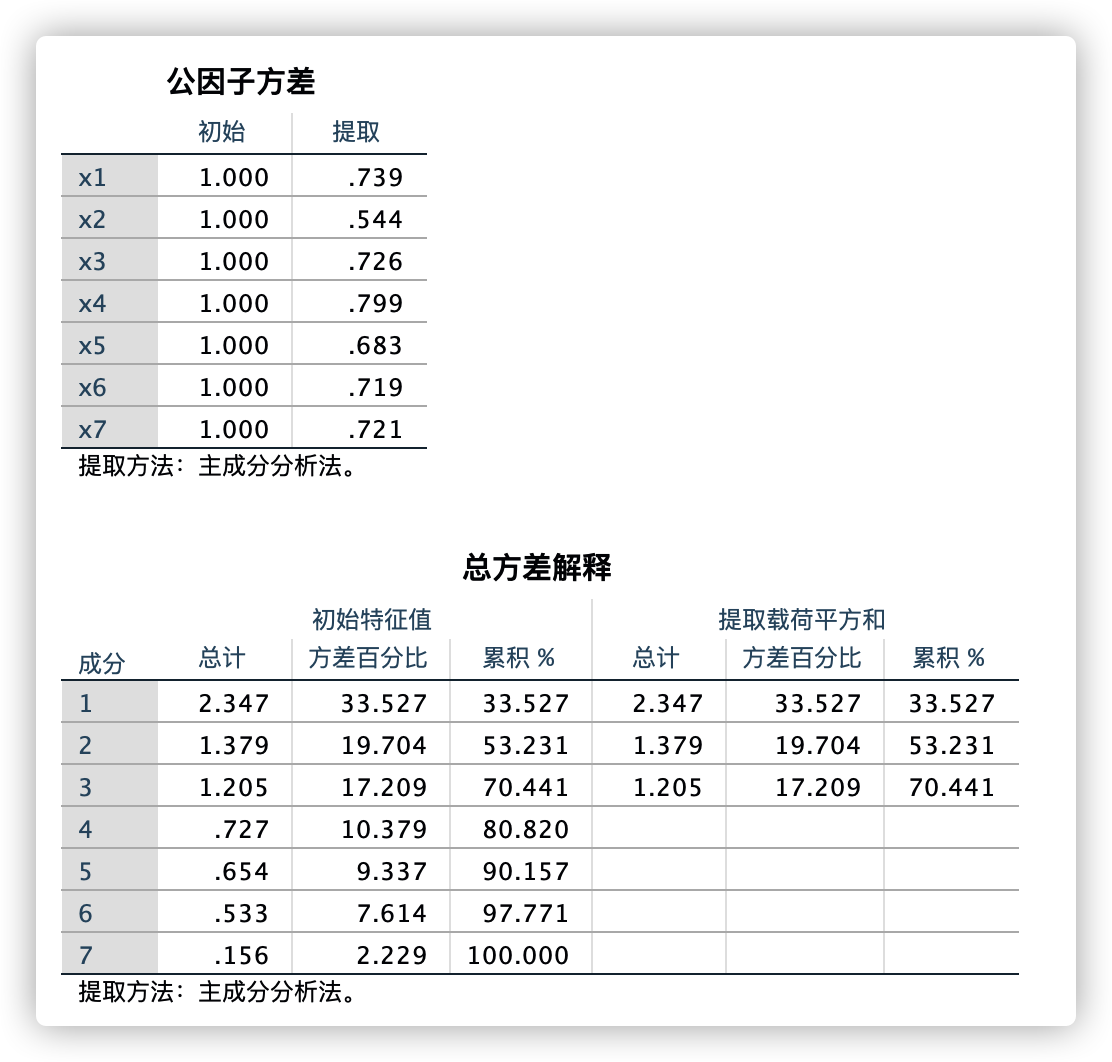
**样本相关矩阵**



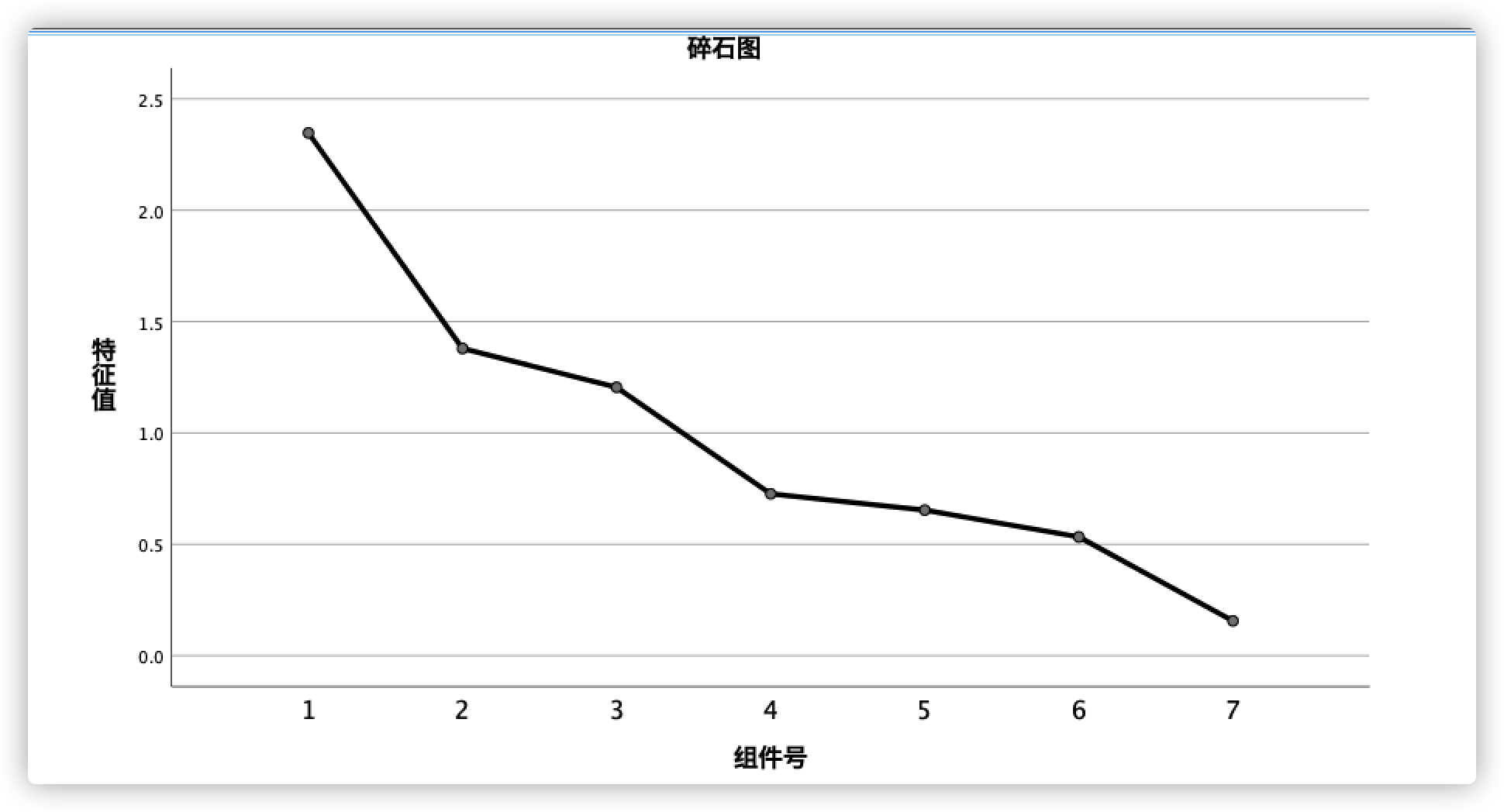
相关性矩阵是反映各个自变量之间的相关性的一个度量，相关性越大越可能分到同一组。



表明题目中的数据是否适合做主成分分析的数值，这里的值较小，说明不太适合。



对每个原始变量的提取系数，总方差解释指选取出的共同因素能够解释所有题项变量的比例。



碎石图是将特征根以图示形式展示，主要用于辅助判断因子个数。如果不确定提取多少个因子，此时可以参考碎石图结果判断因子个数。当折线由陡峭突然变得平稳时，陡峭到平稳对应的因子个数即为参考提取因子个数。比如说上图中选择4、5个因子比较合适。

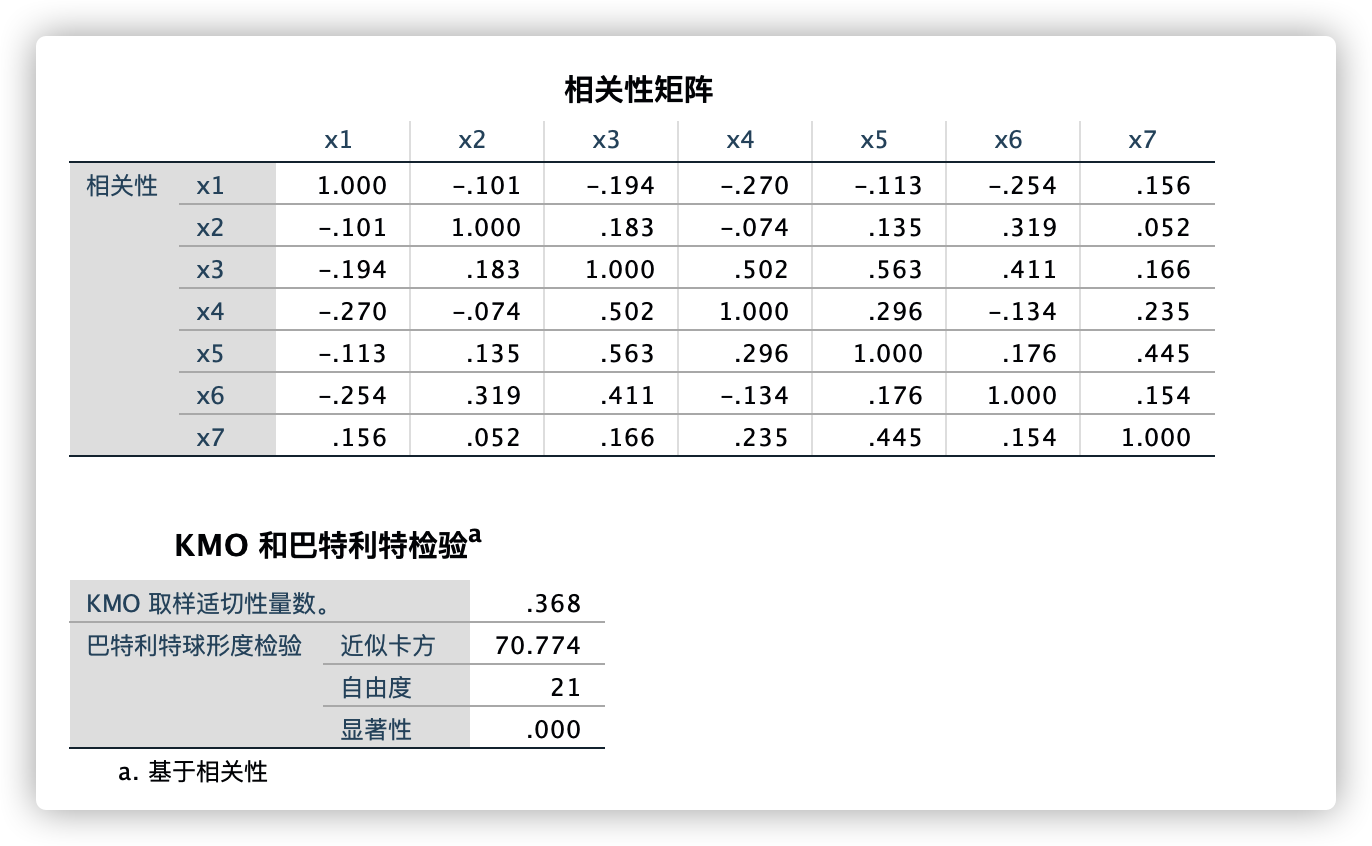


提取出的主成分对每个原始变量的提取占比，即系数。

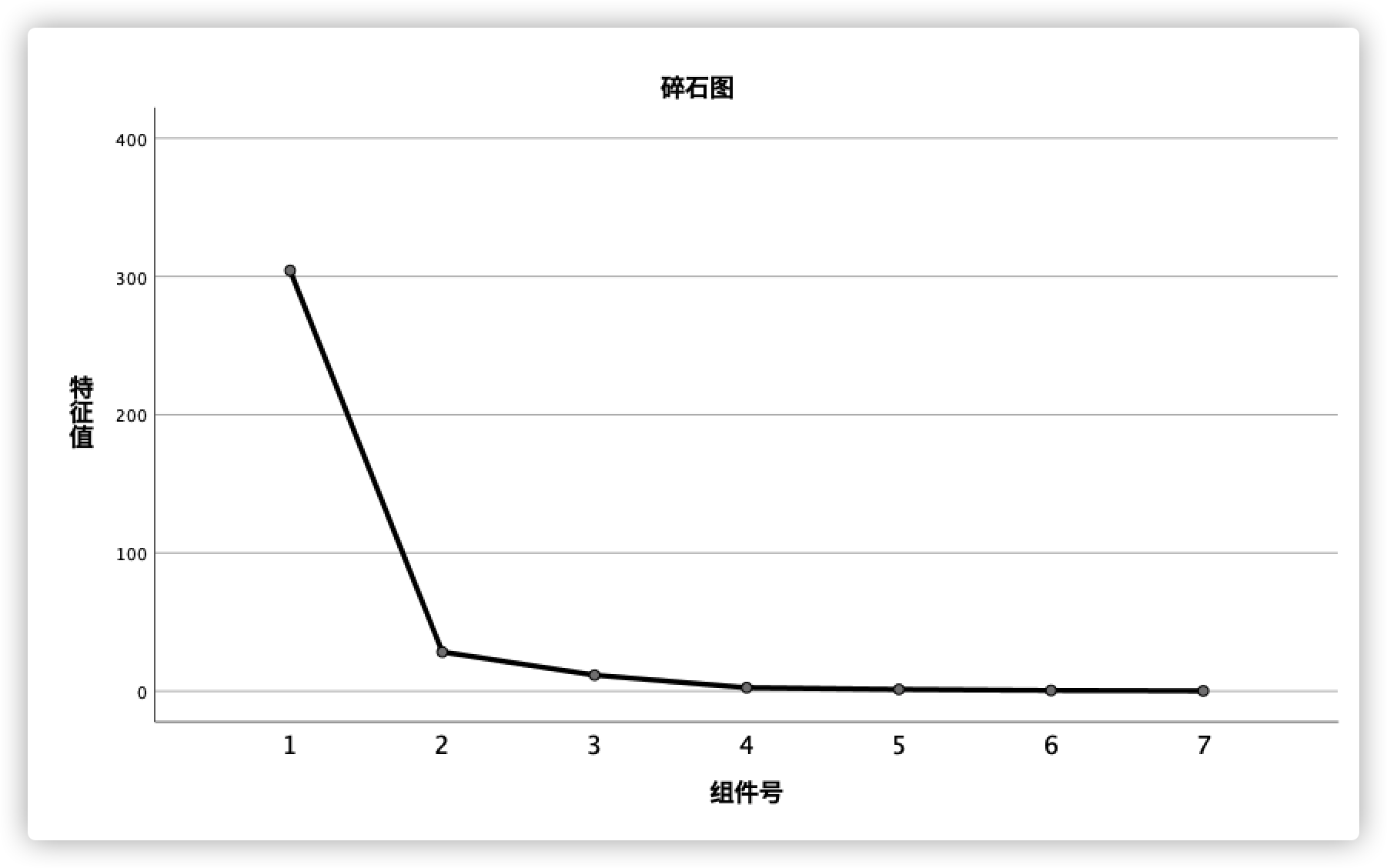
知前三个主成分分别为

可看出，Y1\*与X3\*、X5\*相关度较高，Y1\*近似是7个变量的等权重之和，反映了空气质量的综合指标，Y1\* 值越大，空气质量越差。Y2\* 与X3\*,相关度较低，由相关矩阵的主成分权重系数知，CO对空气污染指标y2的影响较小;Y3\* 与X1\*、X7\*,相关度较高，同理，由相关矩阵的主成分权重系数知，风速和HC对空气污染的影响较大。考虑前三个主成分的贡献率之和达到70.3833%，因此综合考虑来Y1\*，Y2\* 和Y3\*,来评判影响空气污染的重要指标。

**样本协方差矩阵**









这里只提取出一个主成分。

可看出，Y1与x相关度较高，而由相关矩阵的主成分权重系数知，太阳辐射对空气污染的影响最大; Y2, 与x6。相关度较高，由相关矩阵的主成分权重系数知，O3 对空气污染的影响较大; Y3与x5相关度较高，同理，由相关矩阵的主成分权重系数知，NO2, 对空气污染的影响较大。考虑前三个主成分的贡献率依次降低，得出结论:影响空气污染的最重要因素为太阳辐射。由于x2的方差较大，第一主成分主要由变量x2控制，所以所得结论与实际不符。

从协方差矩阵出发，对所有变量进行主成分分析，何从相关矩阵出发做主成分分析，两个方向得出的结果显示，原变量在第一主成分和第二主成分中的相对重要性，由于标准化而有很大变化。从协方差矩阵的角度进行主成分分析，所得第- - -主成分中，权重系数分别为-0.01、 0.9922、 0.941、 -0.0047、 0.0243、 0.1124、 0.0023， 二重相对矩阵的角度进行主成分分析，所得的第一主成分分析， 权重系数为-0.238、0.2056、0.5511、0.3776、0.4980、0.3246、0.3194。两者差距很大，并且在第二主成分中的两个系数相差更远。因此，由协方差矩阵和相关矩阵所得的主成分一-般是不同的。为了满足样本主成分累计贡献率达到80%以上，从相关矩阵出发做的主成分分析应保留4个主成分，而从协方差矩阵来看，只保留1个主成分即可。由此可知，用协方差矩阵进行主成分分析更能简化运算。本文中，由于设计的各变量的变化范围差异不大，因此应从先关矩阵出发求主成分比较合理。