举例说明探索性数据（1-2 个）； 探索性数据分析的方法（3-4 个）

探索性数据分析报告

胡成成——41724260——通信 1701

探索性数据分析（Exploratory Data Analysis，简称 EDA），是指对已有的数据（特别是调查或观察得来的原始数据）在尽量少的先验假定下进行探索，通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。特别是党我们对面对大数据时代到来的时候，各种杂乱的“脏数据”，往往不知所措，不知道从哪里开始了解目前拿到手上的数据时候，探索性数据分析就非常有效。

1. 探索性数据分析基本内容：

检查数据是否有错误：过大过小的数据均有可能是奇异值、影响点或错误数据。要找出这样的数据，并分析原因，然后决定是否从分析中删除这些数据。因为奇异值和影响点往往对分析的影响较大，不能真实反映数据的总体特征。

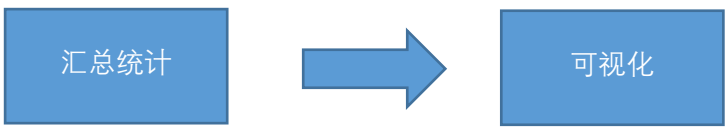
获得数据分布特征：很多分析方法对数据分布有一定的要求，例如很多检验就需要数据分布服从正态分布。因此检验数据是否正态分布，就决定了它们是否能用只对正态分布数据适用的分析方法。

对数据规律的初步观察：通过初步观察获得数据的一些内部规律，例如两个变量间是否线性相关。

2. 探索性数据分析适用场景

对数据中信息没有足够经验，不知道用何种传统方式进行分析，探索性数据分析就会非常有效。

3. 探索性分析的技术手段



3.1 汇总统计

汇总统计是量化的（如均值和方差等），用单个数和数的小集合来捕获数据集的特征，从统计学的观点看，这里所提的汇总统计过程就是对统计量的估计过

程。

3.1.1 单个属性情况

- 频率和众数
- 百分位数
- 位置度量：均值和中位数
- 散步分量：极差和方差

3.1.2 多个属性情况

- 协方差
- 相关系数

3.2 可视化

可视化技术能够让人快速吸收大量可视化信息并发现其中的模式，是十分直接且有效的数据探索性分析方法，但可视化技术具有专门性和特殊性，采用怎样的图表来描述数据及其包含的信息与具体的业务紧密相关。

4. 探索性分析的基本方法

4.1 输入参数探索分析。

将输入参数定义为离散化的变量，并参考这些参数的实际含义将其组合，构成输入参数的多种取值组合方案，多次运行模型，进行参数探索，对结果进行综合分析研究。实验结果的数量可能有几十至几十万种以上，这些结果往往需要借助计算机的强大数据表现能力来进行交互式的探索研究。

4.2 概率探索性分析。

它是对输入参数探索性分析的补充，将输入参数表示为具有特定分布函数的随机变量，运用解析方法或蒙特卡诺方法来计算结果，分析不确定性对结果的影响。但概率探索性分析也存在缺陷，它不能有效反映不确定性变量之间的因果关系，问题的某些方面可能得不到有效分析和深入理解。

4.3 混合探索分析。

最常用的探索方法是上述两者的混合，在使用不确定分布处理一些变量的不确定性后，可以将另外一些可控的关键变量恰当地用离散化的参数来表示。比如针对一些军事行动效果的分析中，可以将行动方案和威胁大小用离散化的参数表示，而将像预警时间这样的变量用概率分布表示。

5. 探索性数据分析基本步骤

5.1 明确分析目的

明确数据分析的目的，才能确保数据分析有效进行，为数据的采集、处理、分析提供清晰的指引方向。

5.2 数据收集

数据收集按照确定的数据分析的目的来收集相关数据的过程，为数据分析提供依据。一般数据来源于数据库、互联网、市场调查、公开出版物。

5.3 数据处理

数据处理包括：数据采集、数据分组、数据组织、数据计算、数据存储、数据检索、数据排序。

5.4 数据分析

数据分析分为：定性数据分析是指对词语、照片、观察结果之类的非数值型数据进行的分析。验证性数据分析是侧重于已有假设的证实或证伪。探索性数据分析是对数据进行分析从而检验假设值的形成方式，侧重于数据之中发现新的特征。

5.5 数据展示

常用柱形图、饼图、折线图等图标展示有用的信息，一目了然的发现数据的本质与作用。

5.6 报告撰写

报告撰写是整个数据分析的最后一步，是对整个数据分析过程的总结。一份优秀的报告需要一个名确的主题、清晰的目录、图文并茂描述数据、结论与建议。

6. 探索性数据分析举例

6.1 探索性数据分析在成绩分析上的应用

课程考试是高校评估学生学习成绩，检验教师教学效果的主要形式，充分发挥考试的测量、诊断、反馈、激励作用，是高校提高教学质量的重要环节。对学生成绩的分析也是提高教学质量的重要手段。

例如某学年第一学期物理成绩的频率分布直方图为例。由下图可以看出，样本分布的对称性不太好，不太符合正态分布（图中黑色曲线是拟合的正态分布的概率密度曲线），不过确定结论需要进一步检验。由此可见学生的成绩集中

靠近在接近 80 分左右，基本符合正常水平，在改革教学中可以继续保持。

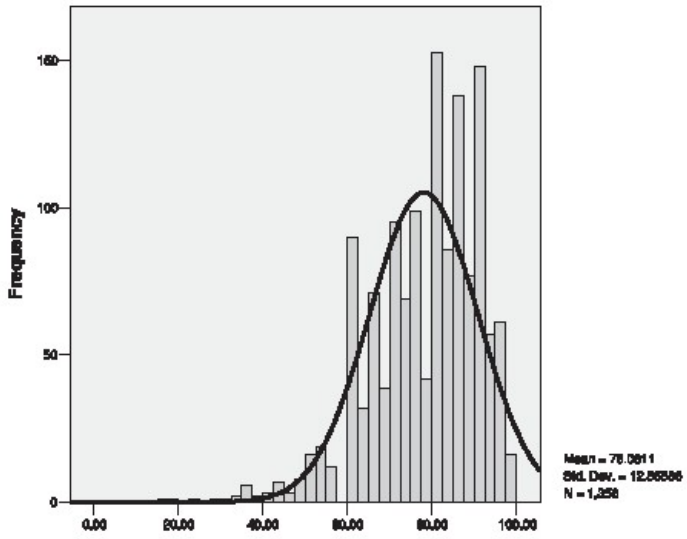


Figure 1 图片来自参考文献【2】

进一步利用非参数统计分析中单样本 K-S 检验法对历次考试成绩进行正态分布假设检验可知，该校上述科目的历次考试成绩均不呈正态分布，而这将导致基于正态分布假设的传统统计分析方法的可信度大大降低。实际上已经有许多学者注意到此种情况，他们从考试方式、学生情况、学习特点和试题质量等方面进行分析，对照统计学中应用正态分布规律的条件，得出了“考试成绩分布不服从正态规律”结论。

6.2 探索性数据分析在我国人均可支配收入的研究

人均可支配收入作为区域发展经济中衡量经济发展状况的一个重要指标，能使人们了解和把握一个国家或地区的宏观经济运行状况。

华北	收入	东北	收入	华东	收入
北京	69483	黑龙江	25450	上海	76801
天津	55244	吉林	25400	江苏	42769
河北	27199	辽宁	32172	浙江	49949
内蒙古	30405			福建	32229
山西	24134			山东	35738
				江西	20230
西北	收入	中南	收入	东南	收入
青海	20753	河南	23089	四川	19924
陕西	21030	湖南	22335	云南	19031
甘肃	17670	湖北	22972	贵州	14967
宁夏	20961	广东	44093	广西	20139
新疆	23742	海南	22045	重庆	24007
		安徽	19815	西藏	19263

Figure 2 原始数据

这里就利用 EDA 技术查探原始数据（Figure 2）的分布，作为一项非常实用的 EDA 技术，箱线图（Figure 3）可以将几个批的数据平行显示，从而对 31 个省市的人均可支配收入数据的特点进行比较分析，这样可以直观的反映出

数据的分布状况和一些较显著的特点。

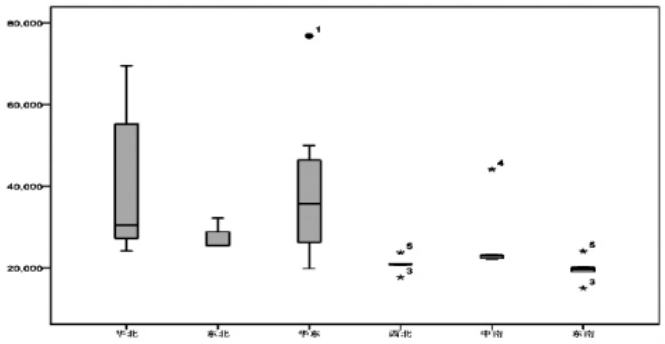


Figure 3 **数据箱线图**

得出结论：西北和东南区域人均可支配收入的总体水平低于中南 75%的城市，而东南下四分位点低于西北最小极值点则表明东南有 25%的城市人均可支配收入水平低于西北总体最低水平。

7. 参考文献：

[1]李志猛,沙基昌,谈群.探索性分析方法及其应用研究综述[J].计算机仿真,2009,26(01):32-35.

[2]张庆丰,王锋.探索性数据分析方法在成绩分析中的应用[J].安阳工学院学报,2011,10(04):102-103+122.

[3]李杨,安瑞娟.探索性数据分析在我国人均可支配收入研究中的应用[J].商业经济,2012(17):21-23.