****

2019级《计算机前沿创新技术（IBM校企联合课程）》课程实践报告

|  |  |
| --- | --- |
| **评分** |  |

**题目：SPSS数据分析**

学号： 19031211696

学院： 计算机科学与技术

姓名： 吕亚龙

**2019年12月19日**

SPSS数据分析

我们正处于大数据的时代，面对海量的数据，我们需要经过数据分析。数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论，并对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，数据分析可帮助人们作出判断，以便采取适当行动。数据分析的数学基础在20世纪早期就已确立，但直到计算机的出现才使得实际操作成为可能，并使得数据分析得以推广。数据分析是数学与计算机科学相结合的产物。在这次IBM讲座中，我了解到了一款很好用的软件SPSS。

SPSS是社会统计科学软件包的简称，主要运用于各个领域数据的管理和统计分析。作为世界社会科学数据分析的标准，SPSS操作界面极其友好，结果输出界面 也很美观，同时还配备了十分详细的用户手册。

SPSS的有如下种核心功能：

一、数据编辑功能：对数据进行增删改等处理，还可以根据需要对数据进行拆分、加权、排序、聚合等处理。

二、可视化功能：SPSS有很强大的绘图功能，可以根据模型自动输出描述性分析的统计图，反映不同变量间的内在关系；同时还可以由用户自定义统计图的基本属性，使数据分析报告更加美观。

三、表格编辑功能：用户可以使用SPSS绘制不同不同风格的表格，同时表格可以在查看器中编辑，也可以在专门的编辑窗口编辑。

四、联接其他软件：SPSS可以打开多种类型的数据文件，其中包括Excle、Access等等，同时用户还可以将图片保存为不同的图片格式。

五、统计功能：SPSS统计功能是进行数据分析中最重要掌握的模块，通过此功能可以完成绝地部分数理统计模型分析。

我们应该如何用SPSS进行数据分析？首先要了解一下数据分析的一般流程是什么？数据获取->数据存储->数据预处理->建模与分析->数据可视化。

常用的数据分析有如下方法。

一、描述统计：是指运用制表和分类，图形以及计筠概括性数据来描述数据的集中趋势、离散趋势、偏度、峰度。

* 缺失值填充：常用方法：剔除法、均值法、最小邻居法、比率回归法

决策树法。

* 正态性检验：很多统计方法都要求数值服从或近似服从正态分布，所

以之前需要进行正态性检验。常用方法：非参数检验的K-量检验、P-P图、Q-Q图、W检验、动差法。

二、假设检验

1、参数检验参数检验是在已知总体分布的条件下（一股要求总体服从正态分布）对一些主要的参数(如均值、百分数、方差、相关系数等）进行的检验 。

1）U验：使用条件：当样本含量n较大时，样本值符合正态分布

2）T检验 使用条件：当样本含量n较小时，样本值符合正态分布

A单样本t检验：推断该样本来自的总体均数μ与已知的某一总体均数μ0 (常为理论值或标准值)有无差别；

B配对样本t检验：当总体均数未知时，且两个样本可以配对，同对中的两者在可能会影响处理效果的各种条件方面极为相似；

C 两独立样本t检验：无法找到在各方面极为相似的两样本作配对比较时使用。

2、非参数检验非参数检验则不考虑总体分布是否已知，常常也不是针对总体参数，而是针对总体的某些一股性假设（如总体分布的位罝是否相同，总体分布是否正态）进行检验。适用情况：顺序类型的数据资料，这类数据的分布形态一般是未知的。

A 虽然是连续数据，但总体分布形态未知或者非正态；

B 总体分布虽然正态，数据也是连续类型，但样本容量极小，如下；主要方法包括：卡方检验、秩和检验、二项检验、游程检验、K-量检验等。

三、信度分析检査测量的可信度，例如调查问卷的真实性。分类：外在信度：不同时间测量时量表的一致性程度，常用方法重测信度内在信度；每个量表是否测量到单一的概念，同时组成两表的内在体项一致性如何，常用方法分半信度。

四、列联表分析用于分析离散变量或定型变量之间是否存在相关。对于二维表，可进行卡方检验，对于三维表，可作Mentel-Hanszel分层分析。列联表分析还包括配对计数资料的卡方检验、行列均为顺序变量的相关检验。

五、相关分析研究现象之间是否存在某种依存关系，对具体有依存关系的现象探讨相关方向及相关程度。

单相关：两个因素之间的相关关系叫单相关，即研究时只涉及一个自变量和一个因变量；

复相关：三个或三个以上因素的相关关系叫复相关，即研究时涉及两个或两个以上的自变量和因变量相关；

偏相关：在某一现象与多种现象相关的场合，当假定其他变量不变时，其中两个变量之间的相关关系称为偏相关。

六、方差分析使用条件：各样本须是相互独立的随机样本；各样本来自正态分布总体；各总体方差相等。

单因素方差分析：一项试验只有一个影响因素，或者存在多个影响因素时，只分析一个因素与响应变量的关系。

多因素有交互方差分析：一顼实验有多个影响因素，分析多个影响因素与响应变量的关系，同时考虑多个影响因素之间的关系。

多因素无交互方差分析：分析多个影响因素与响应变量的关系，但是影响因素之间没有影响关系或忽略影响关系。

协方差分祈：传统的方差分析存在明显的弊端，无法控制分析中存在的某些随机因素，使之影响了分祈结果的准确度。协方差分析主要是在排除了协变量的影响后再对修正后的主效应进行方差分析，是将线性回归与方差分析结合起来的一种分析方法

七、回归分析

1、一元线性回归分析：只有一个自变量X与因变量Y有关，X与Y都必须是连续型变量，因变量y或其残差必须服从正态分布。

2、多元线性回归分析：使用条件：分析多个自变量与因变量Y的关系，X与Y都必须是连续型变量，因变量y或其残差必须服从正态分布。变呈筛选方式：选择最优回归方程的变里筛选法包括全横型法（CP法）、逐步回归法，向前引入法和向后剔除法；横型诊断方法：A残差检验：观测值与估计值的差值要艰从正态分布；B强影响点判断：寻找方式一般分为标准误差法、Mahalanobis距离法；C共线性诊断：a诊断方式：容忍度、方差扩大因子法(又称膨胀系数VIF)、特征根判定法、条件指针CI、方差比例；b处理方法：增加样本容量或选取另外的回归如主成分回归、岭回归等。

3、Logistic回归分析线性回归模型要求因变量是连续的正态分布变里，且自变量和因变量呈线性关系，而Logistic回归模型对因变量的分布没有要求，一般用于因变量是离散时的情况。分类：Logistic回归模型有条件与非条件之分，条件Logistic回归模型和非条件Logistic回归模型的区别在于参数的估计是否用到了条件概率。

4、其他回归方法 非线性回归、有序回归、Probit回归、加权回归等。

八、聚类分析样本个体或指标变量按其具有的特性进行分类，寻找合理的度量事物相似性的统计量。

1、性质分类：Q型聚类分析：对样本进行分类处理，又称样本聚类分析，使用距离系数作为统计量衡量相似度，如欧式距离、极端距离、绝对距离等；R型聚类分析：对指标进行分类处理，又称指标聚类分析，使用相似系数作为统计量衡量相似度，相关系数、列联系数等

2、方法分类：系统聚类法：适用于小样本的样本聚类或指标聚类，一般用系统聚类法来聚类指标，又称分层聚类；逐步聚类法：适用于大样本的样本聚类；其他聚类法 ：两步聚类、K均值聚类等。

九、判别分析

1、判别分析：根据已掌握的一批分类明确的样品建立判别函数，使产生错判的事例最少，进而对给定的一个新样品，判断它来自哪个总体。

2、与聚类分析区别：1）聚类分析可以对样本逬行分类，也可以对指标进行分类；而判别分析只能对样本；2）聚类分析事先不知道事物的类别，也不知道分几类；而判别分析必须事先知道事物的类别，也知道分几类；3）聚类分析不需要分类的历史资料，而直接对样本进行分类；而判别分析需要分类历史资料去建立判别函数，然后才能对样本进行分类。

3、进行分类 ：1）Fisher判别分析法：以距离为判别准则来分类，即样本与哪个类的距离最短就分到哪一类，适用于两类判别；以概率为判别准则来分类，即样本属于哪一类的概率最大就分到哪一类，适用于多类判别。2）BAYES判别分析法 ：BAYES判别分析法比FISHER判别分析法更加完善和先进，它不仅能解决多类判别分析，而且分析时考虑了数据的分布状态，所以一般较多使用。

十、主成分分析将彼此梠关的一组指标变适转化为彼此独立的一组新的指标变量，并用其中较少的几个新指标变量就能综合反应原多个指标变量中所包含的主要信息 。

十一、因子分析一种旨在寻找隐藏在多变量数据中、无法直接观察到却影响或支配可测变量的潜在因子、并估计潜在因子对可测变量的影响程度以及潜在因子之间的相关性的一种多元统计分析方法。

与主成分分析比较：相同：都能够起到済理多个原始变量内在结构关系的作用；不同：主成分分析重在综合原始变适的信息。而因子分析重在解释原始变量间的关系，是比主成分分析更深入的一种多元统计方法。

用途：1）减少分析变量个数2）通过对变量间相关关系探测，将原始变量进行分类

十二、时间序列分析动态数据处理的统计方法，研究随机数据序列所遵从的统计规律，以用于解决实际问题；时间序列通常由4种要素组成：趋势、季节变动、循环波动和不规则波动。主要方法：移动平均滤波与指数平滑法、ARIMA横型、量ARIMA横型、ARIMAX模型、向呈自回归横型、ARCH族模型。

十三、生存分析用来研究生存时间的分布规律以及生存时间和相关因索之间关系的一种统计分析方法。

1、包含内容：描述生存过程，即研究生存时间的分布规律；比较生存过程，即研究两组或多组生存时间的分布规律，并进行比较；分析危险因素，即研究危险因素对生存过程的影响；建立数学模型，即将生存时间与相关危险因素的依存关系用一个数学式子表示出来。

2、方法：统计描述：包括求生存时间的分位数、中数生存期、平均数、生存函数的估计、判断生存时间的图示法，不对所分析的数据作出任何统计推断结论。非参数检验：检验分组变量各水平所对应的生存曲线是否一致，对生存时间的分布没有要求，并且检验危险因素对生存时间的影响。

3.半参数横型回归分析：在特定的假设之下，建立生存时间随多个危险因素变化的回归方程，这种方法的代表是Cox比例风险回归分析法。

4.参数模型回归分析：已知生存时间服从特定的参数横型时，拟合相应的参数模型，更准确地分析确定变量之间的变化规律。

十四、典型相关分析相关分析一般分析两个变里之间的关系，而典型相关分析是分析两组变里（如3个学术能力指标与5个在校成绩表现指标）之间相关性的一种统计分析方法。典型相关分析的基本思想和主成分分析的基本思想相似，它将一组变量与另一组变量之间单变量的多重线性相关性研究转化为对少数几对综合变量之间的简单线性相关性的研究，并且这少数几对变量所包含的线性相关性的信息几乎覆盖了原变量组所包含的全部相应信息。

十五、R0C分析R0C曲线是根据一系列不同的二分类方式(分界值或决定阈）.以真阳性率（灵敏度)为纵坐标，假阳性率（1-特异度)为横坐标绘制的曲线。

用途：R0C曲线能很容易地査出任意界限值时的对疾病的识别能力;选择最佳的诊断界限值。R0C曲线越靠近左上角，试验的准确性就越高；两种或两种以上不同诊断试验对疾病识别能力的比较，一股用R0C曲线下面积反映诊断系统的准确性。

十六、其他分析方法多重响应分析、距离分祈、项目分祈、对应分祈、决策树分析、神经网络、系统方程、蒙特卡洛模拟等。

个人感受。

通过对IBM这门课程的学习，使我很受启发和教育，开拓视野，以前很多不懂不会的知识得到了很好的补充，让我更加明确了作为一名在技术上不断学习和追求的人，不仅要具备一定的科学文化素质，同时也应该具备查找文献学习新知识以及能够寻求突破的能力，尤其要学会利用和整合资源。作为一名硕士研究生，通过对本门课程的学习，使我更加深刻的理解和认识到：学术研究不仅要闭门造车，更加要以开放的姿态来做好研究。

虽然没有接触真正的数据分析，也没有牢固的基础知识 ，但是不能否定的是这段经历让我收获了很多，这种体验带给我的冲击是很大的，让我打开了视野，看到了更大的空间，明白了企业的工作是什么模样，明白了外企的模样。也许是第一次听IBM系列讲座的原因，这两个月带给我的影响是相当深远的，是这段经历把我的视野真正带进了IT这个行业，也大大影响了我之后的职业取向。