**《概率论与数理统计》论文**



**题目：****探究大数据分析中的概率论与数理统计的应用**

**学 号： 2022211570**

**姓 名： 项 枫**

**专 业： 信息安全**

**学 院：网络空间安全学院**

**论文题目**

探究大数据分析中的概率论与数理统计的应用

**摘 要**

随着信息技术和互联网技术的不断发展，各行各业都在进行数字化转型，在这种背景下，数据信息呈现出了快速增长的态势。由于大数据具有规模大、增速快的特点，为数据分析工作增加了难度。然而将概率论与数理统计运用其中，可以提高其有效性并且可以避免一些问题的发生。对此，本文对大数据分析中的概率论与数理统计的应用进行了探究。

**关键词：**大数据分析，概率论，数理统计

**目录**

[第一章 概念阐述 3](#_Toc150976398)

[1.1 大数据分析 3](#_Toc150976399)

[1.2 概率论与数理统计 3](#_Toc150976400)

[第二章 概率论与数理统计和大数据分析的关系 4](#_Toc150976401)

[2.1 相同点 4](#_Toc150976402)

[2.1.1 研究方向 4](#_Toc150976403)

[2.1.2 发展空间 4](#_Toc150976404)

[2.2 差异点 4](#_Toc150976405)

[2.2.1 分支不同 4](#_Toc150976406)

[2.2.2 应用差异 4](#_Toc150976407)

[第三章 概率论与数理统计在大数据分析中的常用方法 5](#_Toc150976408)

[3.1 层析分析方式 5](#_Toc150976409)

[3.1.1 内容介绍 5](#_Toc150976410)

[3.1.2 基本原理 5](#_Toc150976411)

[3.1.3 计算步骤 5](#_Toc150976412)

[3.2 蒙特卡罗分析方式 7](#_Toc150976413)

[3.2.1 概念 7](#_Toc150976414)

[3.2.2 基本思路 7](#_Toc150976415)

[3.2.3 步骤 7](#_Toc150976416)

[第四章 概率论与数理统计在大数据分析中的具体应用 8](#_Toc150976417)

[4.1 在经济数据分析中的应用 8](#_Toc150976418)

[4.1.1 内容介绍 8](#_Toc150976419)

[4.1.2 常用方法 8](#_Toc150976420)

[4.1.3 意义 8](#_Toc150976421)

[4.1.4 实例 8](#_Toc150976422)

[4.2 在商业数据分析中的应用 9](#_Toc150976423)

[4.2.1 内容简介 9](#_Toc150976424)

4.2.2 [研究商业数据中的大客户流失概率方法 9](#_Toc150976425)

[4.2.3 实例 9](#_Toc150976426)

[第五章 结语 11](#_Toc150976427)

[参考文献 12](#_Toc150976428)

1. 概念阐述
   1. 大数据分析

大数据分析是最近几年社会上出现的新兴词汇。大数据分析是指利用现代化技术对规模巨大的数据进行分析。近些年来，世界逐步走进信息化时代，大数据分析技术因其具有速度快、规模大、准确度高等特点，吸引了大量用户并在各行各业广泛应用。

* 1. 概率论与数理统计

概率论与数理统计是数学的一个有特色且又十分活跃的分支，一方面，它有别开生面的研究课题，有自己独特的概念和方法，内容丰富，结果深刻；另一方面，它与其他学科又有紧密的联系，是近代数学的重要组成部分。近年来，概率论与数理统计的理论与方法已广泛应用于工业、农业、军事和科学技术中。

1. 概率论与数理统计和大数据分析的关系

随着大数据时代的不断进步与发展，人们需要在基于计算机与科学技术的基础上，在大数据分析中应用概率论与数理统计，在一定程度上拉近了概率论与数理统计和大数据分析之间的关系[1]。

* 1. 相同点
     1. 研究方向

概率论与数理统计和大数据分析的研究方向均是找出数据中的内在联系和分布规律。

* + 1. 发展空间

步入信息时代，互联网技术的迅猛发展导致大数据技术飞速发展，为二者提供了广阔的发展空间。

* 1. 差异点
     1. 分支不同

大数据分析不属于统计学的分支，而概率论与数理统计属于。

* + 1. 应用差异

二者在应用中有着较小的差异，如概率论和数理统计在DM中得到了广泛的应用和成熟的解决问题的方法和技术，在DM中占有非常重要的地位[5]。

1. 概率论与数理统计在大数据分析中的常用方法

对数据复杂性的理解与系统分析一定程度上增加了数据分析的思考方式，促进数据分析的多样化思维。不同领域间的探究方向及跨学科式的分析不断扩大现阶段科学研究的规模，进一步创新数据分析的方法。概率论与数理统计在大量数据信息进行分析及应用的过程中，可以在短时间内查找到所需数据的主要调查方向。概率论与数理统计在大数据分析中有两种应用方法分别为层析分析方式以及蒙特卡罗方式[2]。

* 1. 层析分析方式
     1. 内容介绍

层次分析法是指将一个复杂的多目标决策问题作为一个系统，将目标分解为多个目标或准则，进而分解为多指标（或准则、约束）的若干层次，通过定性指标模糊量化方法算出层次单排序（权数）和总排序，以作为目标（多指标）、多方案优化决策的系统方法。层次分析法比较适合于具有分层交错评价指标的目标系统，而且目标值又难于定量描述的决策问题[6]。

* + 1. 基本原理

层次分析法根据问题的性质和要达到的总目标，将问题分解为不同的组成因素，并按照因素间的相互关联影响以及隶属关系将因素按不同层次聚集组合，形成一个多层次的分析结构模型，从而最终使问题归结为最低层(供决策的方案、措施等)相对于最高层(总目标)的相对重要权值的确定或相对优劣次序的排定[6]。

* + 1. 计算步骤
       1. 建立层次结构模型

将决策的目标、考虑的因素（决策准则）和决策对象按它们之间的相互关系分为最高层、中间层和最低层，绘出层次结构图。 最高层是指决策的目的、要解决的问题。 最低层是指决策时的备选方案。 中间层是指考虑的因素、决策的准则。对于相邻的两层，称高层为目标层，低层为因素层。

* + - 1. 构造判断(成对比较)矩阵

在确定各层次各因素之间的权重时，如果只是定性的结果，则常常不容易被别人接受，因而Saaty等人提出一致矩阵法，即不把所有因素放在一起比较，而是两两相互比较，对此时采用相对尺度，以尽可能减少性质不同的诸因素相互比较的困难，以提高准确度。如对某一准则，对其下的各方案进行两两对比，并按其重要性程度评定等级。aij为要素i 与要素j重要性比较结果，表1列出Saaty给出的9个重要性等级及其赋值。按两两比较结果构成的矩阵称作判断矩阵。判断矩阵具有如下性质：



判断矩阵元素aij的标度方法如下：

|  |  |
| --- | --- |
| 因素i比因素j | 量化值 |
| 同等重要 | 1 |
| 稍微重要 | 3 |
| 较强重要 | 5 |
| 强烈重要 | 7 |
| 极端重要 | 9 |
| 两相邻判断的中间值 | 2，4，6，8 |

表1 比例标度表

* + - 1. 层次单排序及其一致性检验

对应于判断矩阵最大特征根λmax的特征向量，经归一化(使向量中各元素之和等于1)后记为W。W的元素为同一层次因素对于上一层次因素某因素相对重要性的排序权值，这一过程称为层次单排序。能否确认层次单排序，则需要进行一致性检验，所谓一致性检验是指对A确定不一致的允许范围。其中，n阶一致阵的唯一非零特征根为n；n 阶正互反阵A的最大特征根λ≥n，当且仅当λ=n时，A为一致矩阵。

由于λ连续的依赖于 ，则λ 比n 大的越多，A的不一致性越严重，一致性指标用CI计算，CI越小，说明一致性越大。用最大特征值对应的特征向量作为被比较因素对上层某因素影响程度的权向量，其不一致程度越大，引起的判断误差越大。因而可以用 λ-n 数值的大小来衡量A 的不一致程度。定义一致性指标为：



CI=0，有完全的一致性；CI 接近于0，有满意的一致性；CI 越大，不一致越严重。

为衡量CI 的大小，引入随机一致性指标 RI：



其中，随机一致性指标RI和判断矩阵的阶数有关，一般情况下，矩阵阶数越大，则出现一致性随机偏离的可能性也越大，其对应关系如表2：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 矩阵阶数 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| RI | 0 | 0 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 |

表2 平均随机一致性指标RI标准值(不同的标准不同，RI的值也会有微小的差异)

* + - 1. 层次总排序及其一致性检验

计算某一层次所有因素对于最高层(总目标)相对重要性的权值，称为层次总排序。这一过程是从最高层次到最低层次依次进行的[6]。

* 1. 蒙特卡罗分析方式
     1. 概念

蒙特卡罗法[7] (又称统计试验法)是描述装备运用过程中各种随机现象的基本方法，而且它特别适用于一些解析法难以求解甚至不可能求解的问题，因而在装备效能评估中具有重要地位。

* + 1. 基本思路

蒙特卡罗法[7]的基本思想是：为了求解问题，首先建立一个概率模型或随机过程，使它的参数或数字特征等于问题的解：然后通过对模型或过程的观察或抽样试验来计算这些参数或数字特征，最后给出所求解的近似值。解的精确度用估计值的标准误差来表示。蒙特卡罗法[7]的主要理论基础是概率统计理论，主要手段是随机抽样、统计试验。用蒙特卡罗法[7]求解实际问题的基本步骤为：

(1)根据实际问题的特点．构造简单而又便于实现的概率统计模型．使所求的解恰好是所求问题的概率分布或数学期望；

(2)给出模型中各种不同分布随机变量的抽样方法；

(3)统计处理模拟结果，给出问题解的统计估计值和精度估计值[8]。

* + 1. 步骤

蒙特卡罗法[7]是一种用来模拟随机现象的数学方法，这种方法在作战模拟中能直接反映作战过程中的随机性。在作战模拟中能用解析法解决的问题虽然越来越多，但有些情况下却只能采用蒙特卡罗法[7]。使用蒙特卡罗法[7]的基本步骤如下：

(1)根据作战过程的特点构造模拟模型；

(2)确定所需要的各项基础数据；

(3)使用可提高模拟精度和收敛速度的方法；

(4)估计模拟次数；

(5)编制程序并在计算机上运行；

(6)统计处理数据，给出问题的模拟结果及其精度估计。

在蒙特卡罗法[7]中，对同一个问题或现象可采用多种不同的模拟方法，它们有好有差，精度有高有低，计算量有大有小，收敛速度有快有慢，在方法的选择上有一定的技巧[9]。

1. 概率论与数理统计在大数据分析中的具体应用
   1. 在经济数据分析中的应用
      1. 内容介绍

经济数据是大数据时代的多种数据中经常见到的一种类型，经济数据在经济市场中起着至关重要的作用。通过精准地分析经济数据，有利于加快社会经济的发展，然而由于经济数据是以低密度的方式存在于互联网中，因此，在分析经济数据时具有更高的难度。然而在经济数据分析过程中运用概率论与数理统计方法，可以使经济数据分析工作更加地简单，同时其还是目前经济数据分析方法中非常简单而有效的方法[1]。

* + 1. 常用方法

在经济数据分析中应用概率论与数理统计时，经常使用的分析方法是正态概率分布，然而由于正态概率分布方法可以精准地预测与描述连续性随机变量的概率，因此，在经济金融管理领域中得到广泛普及与运用[1]。

* + 1. 意义

正态概率分布方法可以使用概率论与数理统计精准、有效地分析关于概率的全部信息，与此同时，按照分析结果可以动态掌握经济情况，这样可以保证人们充分了解市场经济规律，同时精准地分析经济信息，最终利用精准的信息更好地拟定于调整后期的决策。尽管经济市场呈现出千变万化形势，但是经济市场的改变同样体现出一定的规律与趋势，所以在分析经济数据过程中，既需要估测经济市场的变化规律与发展形势，同时需要全面考量经济市场中的风险。在经济市场中的风险与利益是一种互相存在的关系，然而通过精准地分析经济数据，既可以降低风险的发生率，同时还可以发现规避风险的手段，最终实现解决经济市场风险的效果。在经济数据分析中应用概率论与数理统计方法，同样是当前人们避免经济风险经常使用的一种方法[1]。

* + 1. 实例
       1. 在中奖概率预测的应用

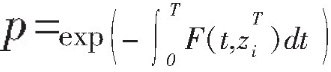
彩票的设计其实是一种数字玩法。需要通过复杂的数学计算和概率数学统计来增加中奖的机会。越来越多的人希望通过买彩票给他们带来好运和改善生活条件的机会。但是，赢得彩票的几率很小，购买彩票以提高生活质量和致富总是一种奢望。在我国的体育彩票中，超级大乐透项目中的游戏机率设定较高，而中奖的机率很低。红色球区域中的5个数字必须从35个数字中选择，蓝色球区域要从11个数字中选择2个数字作为获胜数字，35∶5的赔率是77%,11∶2的赔率是5.57%。，这种概率非常的小。但是可以说获胜原理的设置是根据一些数学定律和概率论原理设计的。根据当前的彩票游戏设置，一个人购买100次的大乐透，中奖的机会只有7次，但是由于每次的开奖都是随机抽取，并且号码不断变化，因此必须考虑其他因素。通过此计算，可以看到中奖的机率非常低。但在购买彩票之前应用概率论和数理统计知识可以使彩票购买者受益[10]。

* 1. 在商业数据分析中的应用
     1. 内容简介

在大数据发展环境下，业务数据对于企业的整体经营而言具有关键的作用。商业数据与经济数据间有着密切的关系，虽然商业数据一种经济数据，但是经济数据并非均是商业数据。企业对商业数据进行分析时会应用到费率理论及统计学的知识，可以减少出现业务数据中流失客户的情况[2]。

* + 1. 研究商业数据中的大客户流失概率方法
       1. 步骤一：构建模型

在模型建立过程中，首先，需要将大客户的基本属性相近作为基本条件，同时还需要保证流失数据符合相同的流失函数，即f0(t)。其次，流失数据的分布条件均是根据流失函数指数项exp(c,zi)T所得到的，接着寻找影响大客户流失概率的因素，划分数据的类型，同时设置指定时间段，接着归纳指定情况中大客户流失的情况，同时得到流失情况的形势图，精准计算流失形势图的标准函数，为：F(t,ziT)=f0(t)·exp (c,ziT)，从而得到某个确定客户在某个时间中流失的概率和其在流失函数中占据的位置，客户在[0,T]阶段中的流失概率是



其中p维回归参数的向量是c,p维协变量向量是ziT，同时把这个协变量界定为一种影响因素，从而可以实现研究模型的建立[1]。

* + - 1. 步骤二：回归参数的选择

研究模型构建之后，需要挑选协变量和参数，接着利用最大类似然函数计算挑选的这些回归参数。因为计算期间有多个方面的因素均对大客户流失造成直接影响，若把全部因素界定为协变量，就会增加模型维数，从而使估测参数的难度大大增加，同样不能有效确保参数的估测精准率。所以，这就需要选择性使用这些因素，为了明确协变量的数量，需要根据数理统计结果挑选协变量，如此一来才可以防止出现各种错误[1]。

* + 1. 实例
       1. 在保险问题中的应用

随着我国社会保障业务的不断深入发展，保险已成为城乡居民普遍关注的问题。在许多保险保护业务中，只有计算每个业务实际带来的收益，才能帮助人们出正确的选择。换句话说，保险办理人可以利用逆向思维来得出他们将从保险公司的经济和商业状况中受益的可能性[11]。实际上，投保人在申请保险之前可以参考保险公司的相关数据。假设每年有2 000人在一家保险公司投保，这些人发生事故的可能性约为0.005%。每个人都要为保险业务支付20元。如果发生意外事故，择期可以收到2 000元的保险费，则可以根据此数据，对保险公司所要承担的损失概率进行计算。可以通过伯努利实验和相关的计算公式来知道答案，即保险公司的损失概率会无限接近于零，该计算受到外部不确定性因素的极大影响，并且会受到投保人自身因素的影响，存在一定的发生意外的概率，但足以为保单持有人提供某些参照标准。这将有助于保单持有人做出正确的选择，并最大程度地提高自身收益[3]。

1. 结语

综上所述，概率论与数理统计在大数据分析中的应用非常广泛。信息化时代背景下，概率论和数理统计不只是大数据分析的辅助性手段，其还属于一种非常简单、有效的数据分析方法，在大数据分析中应用概率论和数理统计时，对大数据中各种数据类型的过程、趋势和影响已成为分析的主体。面对大数据的快速增长趋势，概率论和数理统计在生活中的科学合理应用，将有助于提升人们的生产生活水平。大数据分析中对概率论及数理统计的应用方法进行合理的利用，可以为人们工作及企业经营方面提供一定的便利性，满足社会经济发展的要求。在信息化不断发展的时代中，必须要掌握大数据技术、概率论以及数理统计的相关知识，对于日常的生活生产具有积极作用。

参考文献

1. 梁龙龙.概率论与数理统计在大数据分析中的应用[J].科技视界,2021,(10):34-35.
2. 赵峻婷.大数据分析中概率论与数理统计的运用探究[J].财经界,2022,(08):38-40.
3. 葛培运.概率论与数理统计在经济生活中的应用研究[J].科技视界,2021,(19):95-96.
4. 郑洽好.大数据分析中概率论与数理统计的应用探究[J].数据,2023,(02):72-73.
5. 赵冬青.概率论与数理统计在经济生活中的应用[J].商业观察,2020(05):99-100.
6. 许树柏. 实用决策方法:层次分析法原理[M]. 天津大学出版社, 1988.
7. 马天旗．高价值专利培育与评估：知识产权出版社，2018.04：217.
8. 陈磊，姚伟召，郭全魁，王秀华著,效能评估理论、方法及应用,北京邮电大学出版社,2016.01.
9. 王晖，潘高田，孙俊峰等编著,军事运筹学基础,国防工业出版社,2015.05.
10. 杨青俊.概率论与数理统计在大数据分析中的应用策略[J].今日财富(中国知识产权),2019(11):215-216.
11. 李志男.概率论与数理统计在大数据分析中的应用策略[J].数码世界，2021(3):65-66.