HW1 方差分析

问题描述

研究**社会群体**及其成员的**集体行为**不仅是社会学的热门话题，也是计算机科学的热门话题。在本作业中，我们试图从集体社会和行为信息中感知社会群体的语义。给定社会群体的类别以及集体社会和行为信息的一些特征，我们的最终目标是检验这些特征能否区分这些类别。

数据

所有数据都存储在一个名为 **data.xlsx 的**文件中。

该数据集描述了从 QQ 收集到的在线群组。我们选取了 2040 个在线群组，并在 14 个列（表示为 Col[1-14]）中列出了相应的信息：

Col[1-2]：在线群组名称、群组类别。众所周知，每个 QQ 群都有一个群名来描述群的语义。为了保护隐私和直观起见，名称中的部分字符用 "*"遮盖。类别描述如下表 1 所示：

表 1.类别说明

类别	主题	不
1	在线游戏	484
2	学校校友	300
3	房屋与生活	196
4	股票市场	425
5	组织与行业	635

Col[3-14]：12 个维度特征，分别是群组规模、消息数量、友情关系密度、性别比例、平均年龄、年龄方差、地理区域、移动会话比例、会话数量、无回复会话比例、夜间会话比例、图像比例。

实验

1. **(5 分**）回忆并写出单因素方差分析所依据的假设。
2. **(5 分**）重点关注两列：类别（Col[2]）和平均年龄（Col[7]）。以 "平均年龄 "这一特征

为例，我们想测量不同类别的平均年龄是否有显著差异。请明确指出这项任务的零假设 (H0) 和备择假设 (H1)。

3. 使用您最喜欢的统计分析软件，如 Matlab、R、Excel、SPSS 或 …
   a) **(10 分**）画出 Col[7] 的经验概率密度函数，即平均年龄的经验 pdf。这个维度的数据是否服从高斯分布？检验 Col[7] 的正态性。

b) **(10 分**）在 Col[7] 中，有 5 个按类别标签划分的成分。我们将 Col[7] 中类别为 i（i = 1，...，5）的数据记为 Col[7|categoty=i]。检验各分量的正态性并检验方差的同质性。

c) **(20 分**）对 Col[7] 和 Col[2] 中的类别进行单因素方差分析检验。写下您的结论、支持性统计数据，并将您的数据可视化，以启发这一过程。

4. **(15 分**）再选择 3 列，画出各特征列的经验 pdf，并检验哪一列符合问题 1 中的假设？它们相应的对数变换如何?

5. 如何对非正态数据进行单因素方差分析?

a) **(10 分**）找出并列出可能的解决方案集。

b) **(25 分**）对您选择的 3 列进行单因素方差分析。这些特征列的差异大吗？将结果可视化。

将实验报告和必要的代码压缩成一个文件。