

统计模拟与R语言

大作业

姓名：

学号：

专业：

班级：

得分：

**《统计模拟与R语言》大作业**

**2023-2024学年第2学期**

【**考试说明**】

1. 你将有三天的时间完成这个大作业， 5月25日上午8点下发， 5月27日晚上11：59分之前提交。
2. 将你的答案以pdf文件先提交到http://www.xzc.cn/HFSanONFd1，文件名以“班级-姓名-学号”形式，纸质版上交请看群通知后面再交，纸质版需与pdf电子版保持一致，否则成绩无效。
3. 每题的答案列在题目下方，工整排版。所提交的答案要有必要的统计步骤、公式、假设、检验、R语言分析截图、结果分析与结论等。所有R语言输出的结果必须有相应的文字分析;不必要的R语言结果输出应该被避免。
4. 你可以利用你的笔记、任何书籍或线上资源，但内容必须由你自己完成。不能与其他任何人(无论是否与课程有关)合作或讨论这个大作业。
5. 一定要在你的答案中写清楚解题思路，除此之外还应该包括R语言代码和推理，所有的结果都需要解释。
6. 在截止日期后提交，成绩将按每小时5%的比例扣分。

【**分数设置**】

|  |  |  |  |
| --- | --- | --- | --- |
| 大题 | 小题 | 分值 | 成绩 |
| 1 | （1） | 10 |  |
| （2） | 15 |  |
| （3） | 20 |  |
| 2 | （1） | 10 |  |
| （2） | 30 |  |
| 3 | （1） | 5 |  |
| （2） | 5 |  |
| （3） | 5 |  |

**【考试题目】**

1. 某校化生学院学生的实验课题为探究某种草药提取物在抑制细菌生长方面的效果是否与提取物中的四种成分 有关，现测得13组数据，见herb.csv文件，其中为草药提取物的抑菌效果，为提取物中的四种成分的含量。
2. 试用探索性数据分析方法，分别探究抑菌效果 与每种成分含量之间是否存在线性关系。
3. 建立抑菌效果 与四种成分含量 的回归方程，并对回归方程进行显著性检验与回归诊断。
4. 如果有上述检验存在问题，做相应处理后找到最优回归模型，对最优回归模型进行显著性检验和残差诊断，并解释模型的含义。
5. 某研究小组想要比较三种不同肥料在提高植物生长方面的效果是否存在显著差异。他们在相同的条件下，将一片土地分为三组，并分别施加了肥料A、肥料B和肥料C。随后，他们测量了每组植物的平均生长高度（单位：cm），得到了以下数据：

肥料A组：25, 26, 28, 24, 29, 26, 26, 28, 25, 29  
肥料B组：30, 32, 33, 31, 29, 31, 32, 30, 31, 34  
肥料C组：27, 26, 25, 24, 28, 24, 26, 26, 24, 27

1. 试用探索性数据分析方法判断不同肥料的效果是否存在显著差异。
2. 请用方差分析方法验证不同肥料的效果是否存在显著差异。（提示：做方差分析时需对其前提条件进行检验）。
3. 在课上我们探讨了大数据时代统计学研究仍然需要抽样，请对以下情况进行分析，并给出一个例子进行说明：
4. 如果通过抽样能够显著降低数据处理的复杂程度，且解决问题的效果没有太大的下降，抽样便是最优解；
5. 若随着采样率的降低，解决问题的效果也快速下降，则应寻求大数据解决方案；
6. 如果某些问题的处理效果随着数据量上升有一定的提升，但当数据大到一定规模后再增加数据量带来的效果提升并不明显，这时，应选取一个有较大规模但并非总体的数据集来处理。