答案 - Statistics for Data Science Lecture 4

单项选择题: 以下哪个选项最能描述探索性数据分析（EDA）的作用？

答案: D

填空题: 在探索性数据分析（EDA）中，首先需要\_\_\_\_\_\_\_\_\_\_整个数据集，了解数据的总体情况。

答案: 浏览

判断题: 描述性统计只涉及对数据的总结，而不进行任何形式的推断或预测。

答案: True

计算题: 给定以下钻石价格数据，请计算价格的平均值。  
钻石价格（单位：美元）：[5000, 7000, 6500, 8000, 7500]

答案: 平均价格 = 7100

简答题: 简要解释DataExplorer包在R中的作用，并举例说明其在EDA中的应用场景。

答案: DataExplorer包在R中用于自动化数据分析过程，帮助数据科学家快速了解数据集的基本统计信息、缺失值情况、数据分布等。应用场景包括数据清洗、缺失值处理、变量类型识别等。

简答题: 简要描述如何使用图表和数值总结来分析钻石价格数据集，重点关注哪些变量？

答案: 可以使用直方图、箱线图等图表来分析钻石价格、重量、颜色等变量的分布情况。通过计算均值、标准差等数值总结，可以更好地理解价格与其他变量之间的关系，识别价格的影响因素。

# 答案 - Statistics for Data Science Lecture 5

单项选择题: 以下哪个选项描述了推断统计学的基本作用？

答案: A

填空题: 点估计是用于近似总体参数的\_\_\_\_\_\_\_\_\_\_。

答案: 单个数值

判断题: 无偏估计量的偏差为零。

答案: True

计算题: 已知样本的平均值为10，样本标准差为2，样本量为25，求样本均值的标准误差。

答案: 标准误差 = 2 / √25 = 0.4

简答题: 简要解释无偏估计量的定义，并举例说明。

答案: 无偏估计量是指其期望值等于总体参数的估计量。也就是说，在长期多次抽样中，无偏估计量的平均值应与总体参数相等。例如，样本均值是一个无偏估计量，它用来估计总体均值。

简答题: 简要解释如何判断一个估计量是否是好的估计量，涉及哪些统计性质？

答案: 一个好的估计量应具有无偏性、一致性和有效性。无偏性意味着估计量的期望值等于总体参数；一致性意味着随着样本量的增加，估计量趋近于真实值；有效性指估计量具有最小的方差。

# 答案 - Statistics for Data Science Lecture 6

单项选择题: 以下哪个选项正确描述了假设检验的基本作用？

答案: B

填空题: 在假设检验中，检验的假设通常是关于总体参数的\_\_\_\_\_\_\_\_\_\_。

答案: 声明

判断题: 假设检验仅通过样本数据进行推断，并不直接涉及总体数据。

答案: True

计算题: 假设一个制造公司历史上有70%的灯泡通过可靠性测试，现有新工艺生产的灯泡中，100个样本中有73个通过测试，检验该新工艺是否比旧工艺更可靠。假设显著性水平为0.05。

答案: 可以使用单样本比例检验来测试新工艺灯泡的通过率是否显著高于70%。具体的计算过程需要假设检验的Z统计量计算。

简答题: 简要解释假设检验的基本流程，并举例说明其应用。

答案: 假设检验的基本流程包括：1) 提出原假设（H0）和备择假设（H1）；2) 选择适当的检验方法并设定显著性水平；3) 计算检验统计量并求出p值；4) 比较p值与显著性水平，决定是否拒绝原假设。举例来说，制造公司可以通过假设检验判断新工艺生产的灯泡是否更可靠。

简答题: 简要描述如何使用假设检验来比较两种生产工艺的可靠性。

答案: 通过对两种工艺生产的灯泡进行抽样，分别计算每种工艺的通过率。然后，通过双样本比例检验来比较两种工艺的可靠性，检验两种工艺的通过率是否存在显著差异。

# 答案 - Statistics for Data Science Lecture 7

单项选择题: 以下哪个选项描述了统计功效的作用？

答案: B

填空题: 效应量用于衡量处理效果的\_\_\_\_\_\_\_\_\_\_。

答案: 大小

判断题: 样本大小越大，统计功效通常越高。

答案: True

计算题: 假设一个公交公司生产的刹车垫预计平均寿命为60,000英里，现有新生产工艺生产的刹车垫的样本为50个，已知通过的有73个。计算该工艺的统计功效。假设标准差为5,000英里，显著性水平为0.05。

答案: 可以使用功效分析方法计算其统计功效，需根据样本大小、效应量等参数进行估算。

简答题: 简要解释统计功效的定义，并说明提高统计功效的几种方法。

答案: 统计功效是指研究能够正确拒绝零假设的概率，避免Type II错误。提高统计功效的方法包括增加样本量、提高显著性水平、增大效应量等。

简答题: 简要描述效应量的定义和常见的效应量指标。

答案: 效应量是衡量研究中处理效果大小的指标，常见的效应量指标包括Cohen's d（用于比较两组均值的差异）和r（相关系数，衡量变量之间的关系强度）。

# 答案 - Statistics for Data Science Lecture 8

单项选择题: 以下哪个选项描述了回归分析的作用？

答案: B

填空题: 回归分析用于预测或描述\_\_\_\_\_\_\_\_\_\_之间的关系。

答案: 变量

判断题: 统计学习和机器学习的区别在于，统计学习强调模型的解释性，机器学习强调应用的规模和预测精度。

答案: True

计算题: 给定以下数据，使用线性回归预测工资（wage）与年龄（age）的关系。数据如下：  
年龄（age）：[25, 30, 35, 40, 45]  
工资（wage）：[25000, 30000, 35000, 40000, 45000]

答案: 线性回归模型为：wage = 5000 \* age - 50000  
可以计算出各个年龄对应的预测工资。

简答题: 简要解释统计学习与机器学习的区别，并说明两者的侧重点。

答案: 统计学习侧重于理解和解释数据中的模型，强调模型的解释性、精确性和不确定性；而机器学习更注重实际应用，侧重于大规模数据集的预测精度和计算效率。

简答题: 简要描述回归分析的应用，并举例说明如何使用回归模型预测工资。

答案: 回归分析用于通过已知的变量（如年龄、教育水平等）预测目标变量（如工资）。例如，可以通过回归模型建立工资与年龄、教育年限等变量之间的关系，从而预测不同年龄段的工资水平。

# 答案 - Statistics for Data Science Lecture 9

单项选择题: 以下哪个选项描述了回归分析的基本作用？

答案: B

填空题: 在回归分析中，\_\_\_\_\_\_\_\_\_\_回归是指只有一个预测变量的回归模型。

答案: 简单线性

判断题: 多元线性回归模型考虑多个预测变量，而简单线性回归仅考虑一个预测变量。

答案: True

计算题: 已知以下数据集：  
广告费（TV）：[100, 150, 200, 250, 300]  
销售额（Sales）：[30, 40, 50, 60, 70]  
请建立一个简单线性回归模型，预测广告费用（TV）与销售额（Sales）之间的关系。

答案: 可以通过最小二乘法计算回归方程，得到销售额与广告费用之间的预测关系。

简答题: 简要解释简单线性回归与多元线性回归的区别，并说明它们各自的应用场景。

答案: 简单线性回归是指只有一个预测变量的回归模型，而多元线性回归则考虑多个预测变量。简单线性回归适用于变量之间只有单一关系的情况，而多元线性回归适用于多个因素共同影响的复杂场景。

简答题: 简要描述回归模型在广告费用与销售额关系中的应用，并说明如何进行预测。

答案: 回归模型可以帮助确定广告费用（如电视、广播等）与销售额之间的关系。通过建立回归模型，研究人员可以预测给定广告费用下的销售额。例如，使用线性回归可以得到销售额与广告费用的线性关系，从而进行预测。

# 答案 - Statistics for Data Science Lecture 10

单项选择题: 以下哪个选项描述了K折交叉验证的主要优势？

答案: B

填空题: 均方误差（MSE）对\_\_\_\_\_\_\_\_\_\_比较敏感，因此在存在异常值时使用均方根误差（RMSE）更为合适。

答案: 异常值

判断题: 重复K折交叉验证通过多次重复K折交叉验证过程来减少结果的方差。

答案: True

计算题: 已知回归模型的预测值与实际值分别如下：  
预测值（ŷ）：[100, 150, 200, 250]  
实际值（y）：[120, 145, 190, 260]  
请计算均方误差（MSE）。

答案: MSE = [(120-100)² + (145-150)² + (190-200)² + (260-250)²] / 4 = 116.25

简答题: 简要解释均方误差（MSE）和均方根误差（RMSE）的区别，并说明它们的适用场景。

答案: 均方误差（MSE）是预测误差的平方平均，常用于衡量回归模型的精度，但对异常值敏感。均方根误差（RMSE）是MSE的平方根，其单位与目标变量一致，更容易解释，尤其在数据分布不正常或存在异常值时使用。

简答题: 简要描述R平方（R²）的定义及其在回归分析中的应用。

答案: R平方（R²）表示回归模型对数据的拟合程度，值范围从0到1。较高的R²值表示模型更好地解释了数据的变异。在回归分析中，R²用于评估模型的有效性，越接近1，模型越好。

# 答案 - Statistics for Data Science Lecture 10

单项选择题: 以下哪个选项描述了描述性统计的作用？

答案: B

填空题: 在统计研究中，描述数据集的关键特征时，通常需要回答\_\_\_\_\_\_\_\_\_\_、\_\_\_\_\_\_\_\_\_\_、\_\_\_\_\_\_\_\_\_\_等核心问题。

答案: 数据描述、数据集大小、变量定义

判断题: 描述性统计不涉及数据推断，只是对样本数据进行总结和展示。

答案: True

计算题: 给定以下数据集，计算其样本均值。  
样本数据：[10, 20, 30, 40, 50]

答案: 样本均值 = (10 + 20 + 30 + 40 + 50) / 5 = 30

简答题: 简要解释描述性统计的作用，并说明其在数据分析中的应用。

答案: 描述性统计用于总结和展示数据的主要特征，如样本的均值、方差、分布等。它帮助数据分析师快速了解数据的总体趋势和分布，为后续的建模和推断提供基础。

简答题: 简要描述在统计研究中如何理解和定义变量，并说明变量定义在数据分析中的重要性。

答案: 在统计研究中，变量是指数据集中可以变化的量。变量可以是定量的（如收入、年龄）或定性的（如性别、颜色）。明确变量的定义和度量单位是确保数据分析准确性和一致性的基础。