数据分析与挖掘复习指南

一、数据类型与描述性统计

（一）核心概念

数据按计量尺度可分为分类数据和数值数据：分类数据用于区分事物类别（如学历：本科 / 硕士 / 博士、产品类型：家电 / 服装 / 食品），无数值意义；数值数据用于衡量数量（如年龄：25 岁、每月消费：3000 元），可进一步分为连续型（如体重）和离散型（如购买商品数量）。

（二）关键知识点

1. 集中趋势指标
    1. 均值：所有数据的算术平均，反映数据 “中心位置”，但易受极端值影响（如收入数据中少数高收入者会拉高均值）。
    2. 中位数：数据排序后中间位置的数值（若样本量为偶数，取中间两个数的平均），抗极端值能力强（如收入数据常用中位数描述集中趋势）。
    3. 众数：数据中出现频率最高的数值，可用于分类数据（如某班级学生中 “近视” 人数最多，众数为 “近视”）。
2. 离散程度指标
    1. 标准差 / 方差：衡量数据偏离均值的程度，标准差越大，数据分布越分散（如两个班级数学平均分都是 80，A 班标准差 5，B 班标准差 15，说明 A 班成绩更稳定），其中标准差为σ，方差为σ^2。
    2. 应用场景：仅看集中趋势无法全面描述数据，需结合离散程度（如比较两个城市气温，不仅看平均气温，还要看气温波动）。

（三）易错点

1. 误区：所有数据都用均值描述集中趋势。

纠正：分类数据只能用众数，含极端值的数值数据建议用中位数（如房价数据）。

2.误区：标准差越小越好。

纠正：需结合业务场景，如产品质量控制中，标准差小表示质量稳定（好）；但创新项目的收益数据，标准差小可能意味着收益潜力低（需结合目标判断）。

二、概率分布

（一）核心分布及应用

1. 二项分布（$X \sim B(n,p)$）
    1. 定义：n 次独立重复试验中，每次试验仅 “成功”“失败” 两种结果，成功概率为 p，失败概率为 1-p，用于描述 “成功次数” 的分布。
    2. 关键参数：试验次数 n、成功概率 p。
    3. 应用场景：如 “投篮命中率 70%，投 6 次，计算投中 3 次的概率”“某疫苗有效率 95%，接种 100 人，计算有效人数的分布”。
2. 泊松分布（$X \sim P(\lambda)$）
    1. 定义：用于描述 “单位时间 / 空间内稀有事件发生次数” 的分布。
    2. 关键参数：λ（单位时间 / 空间内事件平均发生次数，可为非整数，如平均每小时接到 4.5 个客服电话）。
    3. 应用场景：如 “某路口每小时交通事故平均发生 0.2 次，计算 1 小时内无事故的概率”“某书店每天平均卖出 8 本某图书，计算当天卖出 10 本的概率”。
3. 正态分布（$X \sim N(\mu,\sigma^2)$）
    1. 定义：连续型数据的常见分布，概率密度曲线呈 “钟形”，对称于均值μ。
    2. 关键参数：μ（均值，决定曲线中心位置）、σ（标准差，决定曲线陡峭程度，σ越小曲线越陡）。
    3.重要性质：3σ 原则 —— 数据落在 “$\mu \pm 3\sigma$” 范围内的概率约 99.74%，可用于初步判断异常值；均值 = 中位数 = 众数。
4. 指数分布（$X \sim Exp(\lambda)$）
    1. 定义：用于描述 “两次相邻事件的时间间隔” 或 “产品失效时间” 的连续型分布。
    2. 关键参数：λ（事件发生频率，$\lambda > 0$）。
    3. 核心性质：无记忆性 ——$P(X > s+t|X > s) = P(X > t)$（如 “某灯泡已使用 1000 小时，再使用 500 小时的概率”，与 “新灯泡使用 500 小时的概率” 相等）。
    4. 应用场景：如 “某手机电池平均使用 200 小时失效，计算电池使用超过 250 小时的概率”“客服电话平均等待时间 5 分钟，计算等待超过 10 分钟的概率”。

（二）分布间关系

1. 二项分布与泊松分布：当 n 较大（如 n≥30）且 p 较小时（如 p≤0.05），二项分布可近似为泊松分布（$\lambda = np$）。
2. 二项分布与正态分布：当 n 足够大（如 n≥50）且 p 接近 0.5 时，二项分布可近似为正态分布（$\mu = np,\ \sigma^2 = np(1-p)$）。

（三）易错点

1. 误区：泊松分布的λ必须是整数。

纠正：λ是“平均发生次数”，可为小数（如每小时平均接到 2.3 个电话）。

2.误区：指数分布可用于描述“事件发生次数”。

纠正：指数分布描述“时间间隔”，而非次数（次数用泊松分布）。

三、抽样推断

（一）核心概念

抽样推断：从总体中抽取部分样本，通过样本统计量（如样本均值）估计总体参数（如总体均值μ）的统计方法，核心是“用样本推断总体”。

（二）关键知识点

1. 概率抽样方法（随机抽样，可计算抽样误差）
    1. 简单随机抽样：总体中每个单位被抽中的概率相等（如从 1000 名学生中随机抽 50 人，每人被抽中概率 0.05）。
    2. 分层抽样：将总体按某特征分组（如按年级分“大一 / 大二 / 大三”），从每组中随机抽样，抽样误差通常小于简单随机抽样（适用于总体内部差异大的场景）。
2. 样本容量影响因素
    1. 置信水平：置信水平越高（如 95%→99%），需样本量越大（需更高精度）。
    2. 总体标准差：总体波动越大（σ越大），需样本量越大（需更多数据抵消波动）。
    3. 允许误差：允许误差越小（如允许误差 ±2→±1），需样本量越大（需更高精度）。
3. 置信区间
    1. 定义：在一定置信水平下，总体参数的可能范围（如“95% 置信水平下，某地区居民平均月收入的置信区间为 [5000,6000] 元”，表示有 95% 的把握认为总体均值μ在此范围内）。
    2. 计算逻辑：样本统计量 ± 边际误差（边际误差 = 置信水平对应的临界值 × 抽样平均误差）。

（三）易错点

1. 误区：抽样误差是测量或数据处理的错误。

纠正：抽样误差是样本统计量与总体参数的天然差异（只要抽样就存在），可通过增大样本量或优化抽样方法减小，但无法消除。

2.误区：重复抽样和不重复抽样无本质区别。

纠正：重复抽样中同一单位可多次被抽中（如从 10 个球中抽 3 个，抽后放回），不重复抽样中单位仅能被抽中一次（抽后不放回）；无限总体常用重复抽样，有限总体常用不重复抽样。

四、假设检验

（一）核心逻辑

通过样本数据判断 “总体参数是否符合某假设”，本质是 “小概率反证法”—— 若原假设成立时，样本出现的概率极小（小于显著性水平α），则拒绝原假设。

（二）关键知识点

1. 假设设定
    1. 原假设（H_0）：待检验的 “无差异 / 无效应” 假设（如 “某班学生数学平均分等于 80 分”）。
    2. 备择假设（H_1）：与原假设对立的假设，分单侧（如 “平均分> 80 分” 或 “平均分 < 80 分”）和双侧（如 “平均分≠80 分”）。
2. 显著性水平与 p 值
    1. 显著性水平α：预先设定的 “小概率阈值”，常用$\alpha = 0.05$（表示 “原假设成立时，样本出现的概率≤5% 则拒绝H_0”）。
    2. p值：原假设成立时，获得当前样本及更极端样本的概率；若$p < \alpha$，拒绝H_0；若$p \geq \alpha$，接受H_0。
3. 两类错误
    1. 第一类错误（弃真错误）：H_0为真时，错误拒绝H_0，概率为α。
    2. 第二类错误（取伪错误）：H_0为假时，错误接受H_0，概率为β；α增大则β减小，需根据业务平衡（如医疗检测中，需严格控制 “漏诊”β，可适当提高α）。

（三）易错点

1. 误区：p值越小，说明效应越大。

纠正：p值反映 “证据强度”（拒绝H_0的把握），不直接代表效应大小（如样本量极大时，微小效应也可能有小p值）。

2.误区：单侧检验比双侧检验更易拒绝H_0，应优先用单侧。

纠正：单侧检验需有明确业务依据（如 “新方法效率一定高于旧方法”），无依据时必须用双侧检验，避免主观偏误。

五、方差分析

（一）核心目的

比较多个总体的均值是否相等（如比较 “3 个班级学生的数学平均分是否有显著差异”“4 种营销方案的销售额是否有显著差异”），而非比较两个样本（两个样本用 t 检验）。

（二）关键知识点（以单因素方差分析为例）

1. 前提条件
    1. 各总体服从正态分布（如每个班级的数学成绩服从$X \sim N(\mu, \sigma^2)$）。
    2. 各总体方差相等（方差齐性，如 3 个班级成绩的方差σ^2无显著差异）。
    3. 样本数据相互独立（如不同班级的学生无重叠，数据无关联）。
2. 核心计算
    1. 平方和分解：总平方和（SST）= 组间平方和（SSA，各组均值与总均值的差异）+ 组内平方和（SSE，组内数据与本组均值的差异）。
    2. F 统计量：$F = \frac{\text{组间均方（}MSA\text{）}}{\text{组内均方（}MSE\text{）}}$，其中组间均方$MSA = \frac{SSA}{k-1}$，组内均方$MSE = \frac{SSE}{n-k}$，k 为组数，n 为总样本数。
3. 结果判断
    1. 若$F > F_\alpha$（α为显著性水平，如$F_{0.05}$）或$p < \alpha$，拒绝原假设，认为至少有一个总体均值与其他不同；
    2. 若$F \leq F_\alpha$或$p \geq \alpha$，接受原假设，认为各总体均值无显著差异。

（三）易错点

1. 误区：方差分析要求各组样本容量相等。

纠正：单因素方差分析不强制要求各组样本量相等，只要满足正态、方差齐性、独立即可。

2.误区：F 统计量越小，说明组间差异越小。

纠正：F 统计量是 “组间差异” 与 “组内差异” 的比值，F越小，说明组间差异相对于组内差异越小，越难拒绝 “均值相等” 的假设。

六、回归分析

（一）核心类型与参数

1. 一元线性回归
    1. 模型形式：$y = \beta_0 + \beta_1 x + \varepsilon$，其中 y 为因变量（如销售额），x 为自变量（如广告费用），β_0为截距（x=0 时 y 的预测值），β_1为斜率（x 每增加 1 单位，y 平均变化β_1单位），ε为随机误差项（不可预测的随机波动）。
    2. 参数估计：用最小二乘法（使 “实际 y 与预测 y 的残差平方和最小”）估计β_0和β_1。
2. 多元线性回归
    1. 模型形式：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$（多个自变量，如 y = 销售额，$x_1$= 广告费用，$x_2$= 促销力度）。
    2. 关键问题：多重共线性（自变量间高度相关，如x_1= 身高，x_2= 体重），会导致参数估计不稳定（系数符号异常、标准误差大）；解决方法包括 “删除高度相关的自变量”“逐步回归”“岭回归”。

（二）拟合优度与残差

1. 决定系数R^2
    1. 定义：衡量模型对因变量变异的解释能力，取值范围 0≤R^2≤1。
    2. 意义：$R^2 = 0.7$表示 “因变量 70% 的变异可由自变量解释”，R^2越接近 1，拟合效果越好。
    3. 局限：增加自变量会使R^2增大（即使自变量无实际意义），需用调整后R^2（消除自变量数量影响），调整后R^2更适合多元回归。
2. 残差
    1. 定义：残差 = 实际y - 预测y，反映模型未解释的部分。
    2. 理想残差：服从正态分布、均值为 0、方差恒定（无趋势或波动），若残差有规律（如随 x 增大而增大），说明模型存在缺陷（需改进，如增加二次项）。

（三）易错点

1. 误区：$R^2 = 1$表示模型完美。

纠正：$R^2 = 1$可能是 “过拟合”（模型过度贴合样本数据，对新数据预测能力差），需结合测试集验证模型泛化能力。

2. 误区：相关系数显著不为 0，说明自变量与因变量有因果关系。

纠正：相关≠因果，相关系数仅反映 “线性关联程度”，因果关系需结合业务逻辑验证（如 “冰淇淋销量与溺水人数正相关，但无因果关系，共同原因是‘气温高’”）。

七、时间序列分析

（一）核心构成与平稳性

1. 时间序列构成成分
    1. 趋势（T）：长期稳定的变化方向（如逐年增长的销售额）。
    2. 季节变动（S）：周期性重复的波动（如每年春节前的销售额高峰）。
    3. 循环变动（C）：非固定周期的波动（如经济周期的繁荣与衰退，周期 3-10 年）。
    4. 随机变动（I）：无规律的随机波动（如某一天突发的促销活动影响）。
2. 平稳性
    1. 定义：时间序列的统计特征（均值μ、方差σ^2、自相关系数）不随时间推移而变化（如 “过去 5 年每月平均销售额稳定在 100 万，方差稳定在 5 万”）。
    2. 重要性：多数时间序列模型（如 AR、MA、ARIMA）依赖平稳性假设 —— 非平稳序列的均值μ/ 方差σ^2随时间变化，模型参数无法固定，预测结果不可靠；非平稳序列需通过 “差分”（如计算相邻期差值）转化为平稳序列。

（二）平滑与预测方法

1. 移动平均法
    1. 作用：平滑短期随机波动，凸显趋势；窗宽 k（计算平均的期数）为奇数时更易计算（如 k=3，用 t-1、t、t+1 期数据平均作为 t 期平滑值）。
    2. 局限：不适用于有明显季节趋势的序列（如 k=12 无法平滑年度季节波动）。
2. 指数平滑法
    1. 简单指数平滑：适用于无趋势、无季节的平稳序列（如稳定的日常客流量），仅需历史数据递归计算（权重随时间衰减，近期数据权重更大）。
    2. 二次指数平滑：适用于有线性趋势但无季节的序列（如逐年线性增长的产量），在简单平滑基础上增加趋势修正项。
3. ARIMA 模型
    1. 核心参数：p（自回归项阶数，用过去 p 期数据预测当前）、d（差分次数，将非平稳序列转化为平稳序列的次数）、q（移动平均项阶数，用过去 q 期残差预测当前）；如 $ARIMA(1,1,1)$表示“1 阶自回归、1 次差分、1 阶移动平均”。

（三）易错点

1. 误区：移动平均法的窗宽 k 越大越好。

纠正：k 过大易“过度平滑”（丢失趋势信息），k 过小易“平滑不足”（保留过多随机波动），需结合序列周期（如年度数据 k=4，季度数据 k=12）选择。

2.误区：简单指数平滑可用于有趋势的序列。

纠正：简单指数平滑无趋势修正，用于有趋势的序列会导致预测值滞后于实际趋势（如实际销售额逐年增长，预测值会低于实际值），需用二次或三次指数平滑。

八、聚类分析与数据预处理

（一）K-means 聚类

1. 核心逻辑

1.无监督学习方法：将样本按“相似度”分为 K 个簇，使簇内样本相似度高、簇间样本相似度低。

2.关键参数：聚类簇数 K（需提前设定，常用“肘部法则”—— 绘制 K 与簇内平方和的曲线，曲线拐点对应的 K 为最优值）。

3.距离度量：

1.欧氏距离：适用于连续型数据，衡量两点间直线距离（如“身高 170cm、体重 60kg”与“身高 180cm、体重 70kg”的距离，计算公式为$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$）。

2.曼哈顿距离：适用于关注“绝对差异”的场景（如“快递配送距离”，需按街道直角计算， 而非直线距离，计算公式为$|x_1-x_2|+|y_1-y_2|$）。

4. 收敛条件：簇中心（质心，簇内样本均值μ）不再变化（或变化小于阈值），此时聚类结果稳定。

（二）数据预处理

1. 缺失值处理
    1. 均值填充：适用于连续型数值数据且缺失率低（如缺失率 < 5% 的年龄数据），用非缺失值的均值填充；不适用于分类数据（如性别无法用均值填充）。
    2. 中位数填充：适用于含极端值的连续型数据（如收入数据，中位数比均值更抗极端值）。
    3. 删除样本：适用于缺失率高（如缺失率 > 30%）且样本量充足的场景，避免填充引入偏差；但样本量小时易导致数据量不足。
2. 数据标准化
    1. Z-score 标准化：将数据转化为 “均值 = 0，标准差 = 1” 的标准正态分布，公式为：标准化后值 = $\frac{\text{原始值}-\text{均值}}{\text{标准差}}$；适用于对 “数值范围敏感” 的模型（如 K-means、SVM，避免大数值变量主导聚类 / 分类结果）。
3. 异常值处理
    1. 3σ 原则：适用于正态分布数据，将 “落在$\mu \pm 3\sigma$之外” 的数据判定为异常值（概率仅 0.26%，属于小概率事件）。
    2. 处理方式：根据业务判断异常值是否为 “错误数据”（如录入错误的 “年龄 1000 岁” 需修正或删除），若为 “真实极端值”（如高收入客户），需保留并单独分析。

（三）易错点

1. 误区：K-means 的 K 值越大，聚类效果越好。

纠正：K 值过大易导致 “过聚类”（将相似样本拆分为多个簇，无业务意义），需结合业务目标（如 “将客户分为 3 类：高价值 / 中价值 / 低价值”），而非单纯追求 K 值增大。

2. 误区：缺失值都可以用均值填充。

纠正：分类数据需用 “众数填充”（如性别用 “男” 或 “女” 的众数），高缺失率数据用 “均值填充” 会严重扭曲数据分布，需用更复杂方法（如基于模型预测缺失值）。

九、数据分析报告撰写

（一）核心内容

1. 分析背景与目标：明确 “为什么做分析”（如 “为提升某产品销量，分析影响销量的关键因素”）和 “要解决什么问题”（如 “识别高潜力客户群体”）。
2. 数据来源与预处理：说明数据来源（如 “企业 CRM 系统 2023 年销售数据”）、数据范围（如 “全国 10 个城市的线下门店数据”）、预处理过程（如 “缺失值用中位数填充，异常值用 3σ 原则（落在$\mu \pm 3\sigma$之外）删除”）。

3. 分析过程与结论：用图表（如折线图展示趋势、柱状图对比差异）呈现关键发现，结论需“数据支撑”（如“广告费用每增加 1 万元，销售额平均增加 5 万元，$R^2 = 0.82$”）。
4. 决策建议：提供可落地的行动方案（如“建议将广告费用向一线城市倾斜，预计可提升销售额 15%”），而非单纯描述数据。

（二）易错点

5.误区：报告需包含详细代码实现。

纠正：数据分析报告的读者多为业务人员或决策者，无需展示代码（代码可作为附录供技术人员参考），重点是“结论 + 建议”。

6.误区：图表越多越好。

纠正：图表需“服务于结论”（如用散点图展示 x 与 y 的线性关系，用饼图展示客户群体占比），避免无关图表（如用折线图展示分类数据），防止信息冗余。