

一、问题重述

1.1 问题背景

随着科学技术的发展，无创产前检测（NIPT）已成为产前筛查中的重要工具，其准确性和安全性深受认可。然而，检测时机的选择直接影响结果的可靠性及临床决策的有效性。若检测时间过早，母体外周血中胎儿游离 DNA 浓度可能不足，易导致检测失败或假阴性结果；而过晚检测，则可能错过后续确诊性介入操作及遗传咨询的最佳窗口，影响对胎儿异常的及时判断与干预。因此，根据孕周合理选择 NIPT 检测时间，是确保结果准确、实现早期发现和科学应对胎儿异常的关键所在。

1.2 问题提出

- 问题一核心任务为探究胎儿 Y 染色体浓度与孕妇孕周数和 BMI 等指标之间的关联，需给出具体的关系模型并进行显著性的检验。
- 问题二核心任务为针对男胎孕妇，以临床证明的主要因素 BMI 为依据，进行分组以确定最佳检测时间，需分析检测误差对结果的影响。
- 问题三核心任务为在问题二的基础上，综合考虑更多因素（如身高、体重、年龄等）和达标比例，再次进行分组优化。仍需分析结果受检测误差的影响。
- 问题四核心任务是为女胎孕妇构建一个判定胎儿染色体是否异常（特指 21、18、13 号染色体非整倍体）的模型。必须综合考量多个指标，包括但不限于：相关染色体（21, 18, 13 号）的 Z 值、X 染色体的相关数据（Z 值、读段数等）、GC 含量、染色体的读段数及相关比例、孕妇的 BMI。

二、问题分析

问题一：针对探究胎儿 Y 染色体浓度与孕妇孕周数和 BMI 等指标之间的关联这一核心任务，本文先通过绘制散点图、计算其间的 Pearson 相关系数、观察线性回归的偏回归图以初步探索其间的关系，接着，计算各孕妇三项指标的组内相关系数（ICC），对初步探索阶段结果不佳的原因进行了合理的解释。在此基础上，选择建立了符合数据特性的混合效应模型，并进行似然比检验。利用模型协方差参数估算值和信息准则对比引入了随机斜率的模型 II 和只引入随机残差的模型 I，最终选择不引入随机斜率而只含有随机截距的线性混合效应模型 I。针对平均每个孕妇只有 4 个观测值的特性，采用了 Kenward-Roger 调整的 Wald 检验进行了对固定效应模型系数的显著性检验。

问题二：针对男胎孕妇，以临床证明的主要因素 BMI 为依据，进行合理分组以确定最佳检测时间这一目标，先将数据先处理为 1v1 形式，即只保留一个孕妇最早测出达标 Y 染色体浓度时的数据。对于在整个观察期结束时仍未达到阈值的孕妇，保留其最后一次检测的孕周的记录，这些数据也即右删失数据。之后，依据孕妇 BMI 的四分位数将其分为四组，本文用生存分析法以充分纳入删失数据提供的信息。先用乘积极限法绘制了各组孕妇的 KM 生存曲线，找到各组孕妇的中位和分位达标时间。在用

Schoenfeld 残差通过 HP 检验之后，建立了 Cox 回归进一步分析孕妇 BMI 对达标速度的影响。最后，利用蒙特卡罗模拟法分析检测误差对结果的影响，说明模型的稳健。

问题三：在问题二的基础上，综合考虑更多因素（如身高、体重、年龄等）和达标比例，再次进行分组。先利用 K-Means++ 算法结合肘部法则确定对孕妇分组的情况，结合主成分分析法将特征降维使得分组结果更加直观。在问题二的基础上，绘制了新分组情况下的 KM 生存曲线，之后使用 Log-Rank 检验对不同组别孕妇两两进行检验，验证分组的合理性。与问题二类似，本文建立了多因素 Cox 回归进一步分析孕妇各指标对达标速度的影响，并用森林图可视化了结果。利用蒙特卡罗模拟法分析检测误差对结果的影响，验证模型的稳健性。

问题四：针对为女胎孕妇构建一个判定胎儿染色体是否异常（特指 21、18、13 号染色体非整倍体）的模型任务，本文先对附件女胎数据特性进行了充分的分析，抛弃了传统的依据数据拟合得到分类的方法，采用无监督的簇群为 1 的聚类方法得到中心数据特性和其余数据相对于中心的距离，通过引入数据质量因子调整距离，权衡对 I、II 类错误的容错率，在此基础上重新聚类。

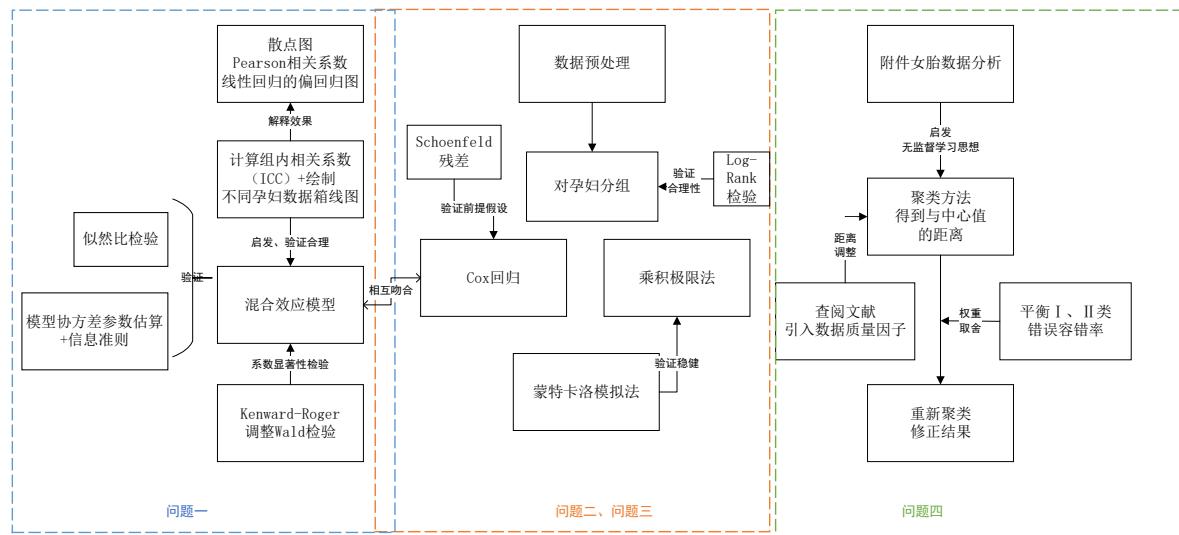


图 1 问题分析思路图

三、模型假设

1. 假设附件中 F 列末次月经时间为该孕妇月经结束时间。
2. 假设 X 染色体 Z 值的极端数据是由于检测失误引起的。
3. 假设检测过程存在 5% 的假阴性率和 5% 的假阳性率。

四、主要符号说明

符号	说明
wd_j	第 j 组孕妇的孕周数值
$S(wd_j)$	第 j 组孕妇在检测孕周 wd_j 时胎儿 Y 染色体达标的概率估计值