问题 2 临床证明，男胎孕妇的 BMI 是影响胎儿 Y 染色体浓度的最早达标时间（即浓度达到或超过 4%的最早时间）的主要因素。试对男胎孕妇的 BMI 进行合理分组，给出每组的 BMI 区间和最佳 NIPT

时点，使得孕妇可能的潜在风险最小，并分析检测误差对结果的影响。

问题2：

**2.1 数据预处理**

本问题旨在探究孕妇BMI对胎儿Y染色体浓度最早达标时间的影响。考虑到同一孕妇的BMI随时间变化幅度较小（个体内变异），且个体间差异远大于时间变化量，为提高分析的准确性和可靠性，我们采用每位孕妇多次BMI测量的平均值作为其代表值进行后续统计分析。

为合理区分样本并便于计算NIPT最佳时点，我们依据Y染色体浓度达标情况对孕妇进行了科学分组：

第一组（始终达标组）：从首次检测开始Y染色体浓度即达到或超过4%的样本。该组数据保存在bmi\_Y\_always\_can\_test\_result.xlsx中，包含孕妇代码、平均BMI和最早达标天数（即首次检测时间）。

第二组（中间达标组）：初始检测未达标但后续检测达标的样本。该组数据保存在bmi\_Y\_middle\_result.xlsx中，包含孕妇代码、平均BMI和预测达标天数（通过插值法计算获得，详见附录代码Line XX）。

第三组（从不达标组）：所有检测中Y染色体浓度均未达标的样本。该组数据保存在bmi\_Y\_cannot\_test\_result.xlsx中，包含孕妇代码、平均BMI和最晚达标天数（即末次检测时间）。

为确保数据质量，若某样本曾达标但后续检测又低于4%，则视为无效样本并予以剔除。经上述处理，最终获得有效样本总量为230例。其中，第一组186例（占比80.87%），第二组37例（占比16.09%），第三组7例（占比3.04%）。这一分组策略确保了后续分析的可靠性和代表性。

分类过后总样本量为230人，第一组样本量为186人（占总比80.87%），第二组样本量为37人（占总比16.09%），第三组样本量为7人（占总比3.04%）。、bmi\_Y\_middle\_result.xlsx和bmi\_Y\_cannot\_test\_result.xlsx文件中。

为了提炼三种情况(always\_can\_test；middle；cannot\_test)孕妇的BMI特点，我们分别对三组所有样本的BMI和达标天数求了平均值，结果如下：

|  |  |  |
| --- | --- | --- |
|  | BMI | 对应天数 |
| always\_can\_test | 31.96548919 | 88.51075269 |
| middle | 33.37592436 | 112.1758142 |
| cannot\_test | 35.21942627 | 139.7142857 |

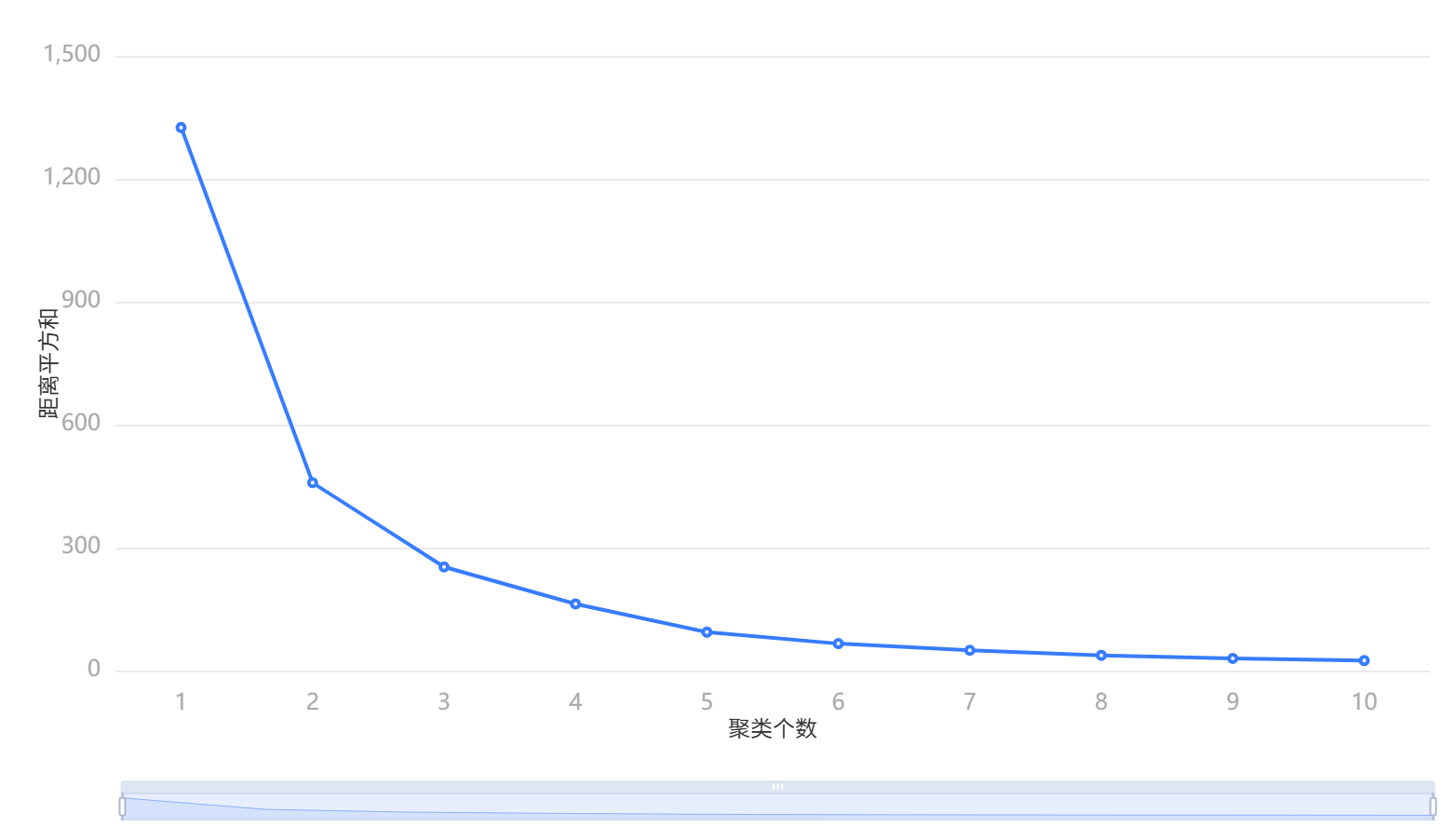
从统计数据不难看出：BMI小的个体Y染色体达标的时间有更早的趋势

以上可以省略

1. **BMI划分**

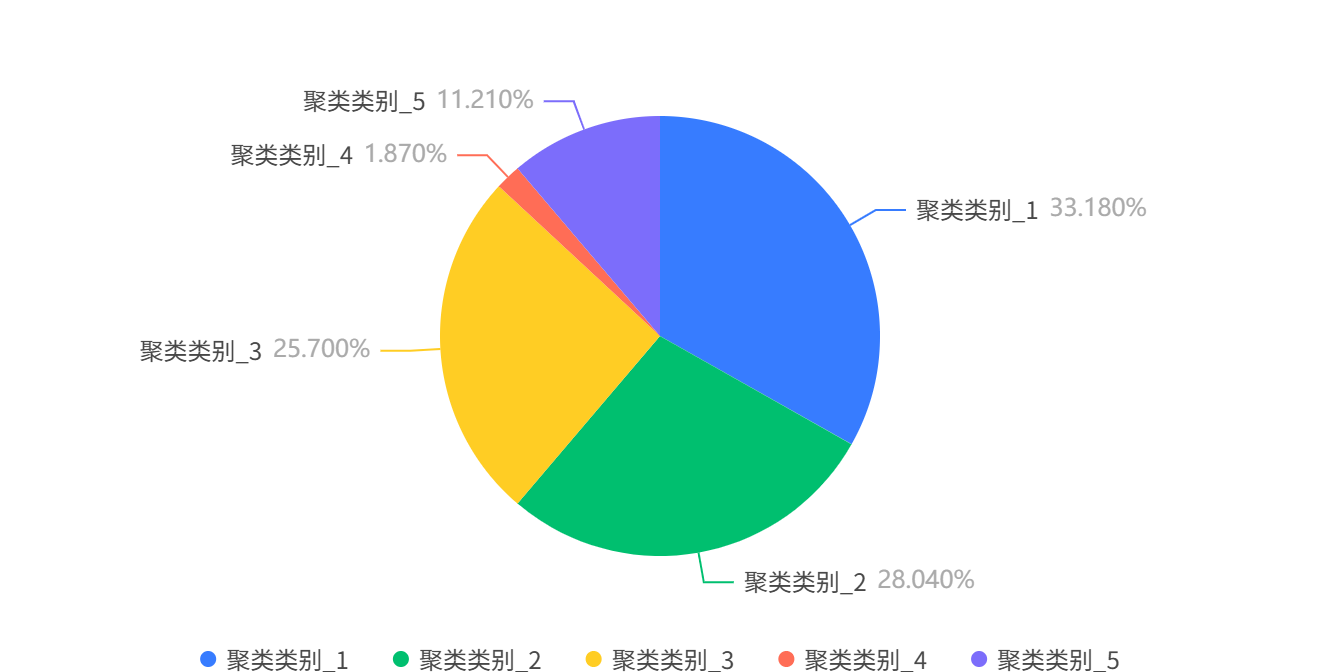
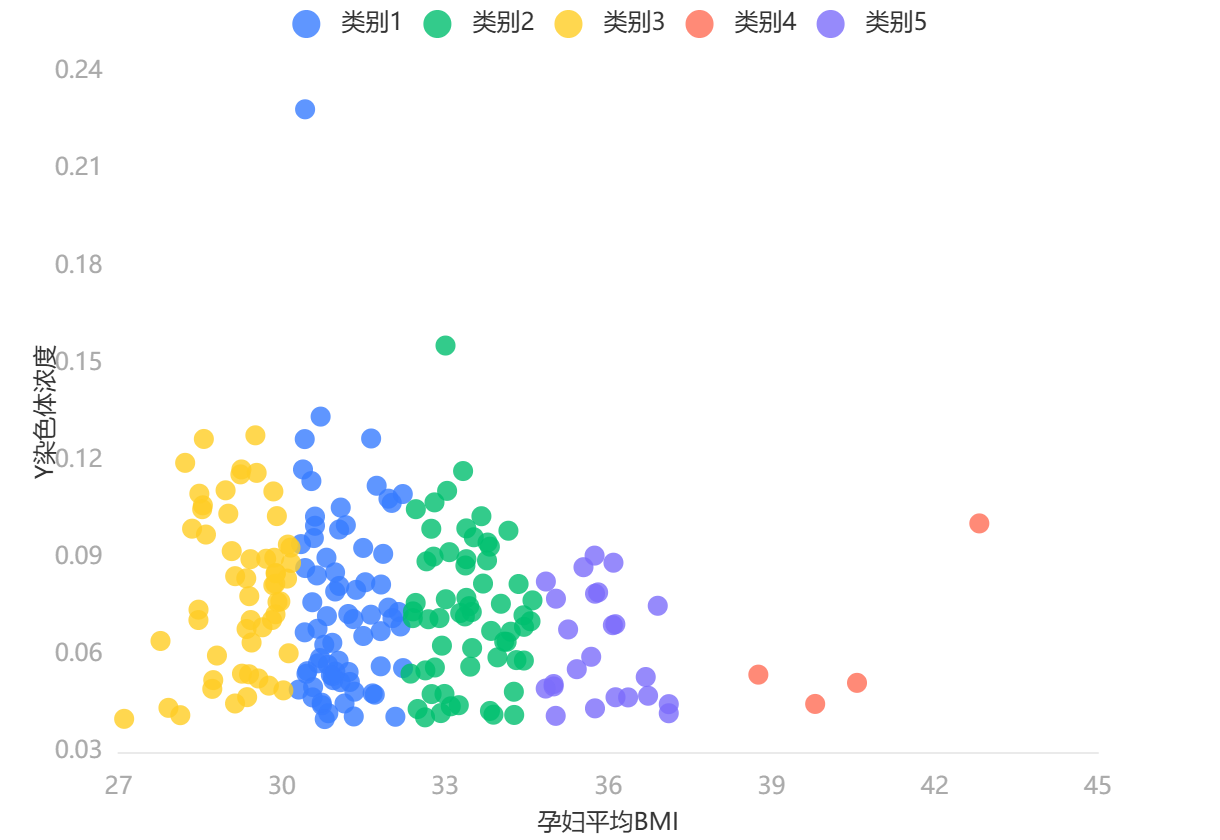
我们使用聚类算法来对BMI进行细分：

我们先用肘部法则来确定合适的划分数量，结果如下图：



在聚类个数大于5时，距离平方和下降明显变慢，故我们选择5作为聚类个数。

下面我们对Y染色体浓度/孕妇平均BMI散点图进行聚类分析，结果如下图：



**差异性分析与分类评价**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 聚类类别（平均值±标准差） | | | | | F | P |
| 类别1(n=71) | 类别2(n=60) | 类别3(n=55) | 类别5(n=24) | 类别4(n=4) |
| 孕妇平均BMI | 31.125±0.555 | 33.415±0.649 | 29.251±0.698 | 35.837±0.716 | 40.487±1.723 | 688.969 | 0.000\*\*\* |
| Y染色体浓度 | 0.076±0.03 | 0.074±0.023 | 0.081±0.024 | 0.062±0.017 | 0.063±0.026 | 2.497 | 0.044\*\* |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | |

对于变量孕妇平均BMI，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，说明变量孕妇平均BMI在聚类分析划分的类别之间存在显著性差异；  
 对于变量Y染色体浓度，显著性P值为0.044\*\*，水平上呈现显著性，拒绝原假设，说明变量Y染色体浓度在聚类分析划分的类别之间存在显著性差异；

|  |  |  |
| --- | --- | --- |
| 轮廓系数 | DBI | CH |
| 0.54 | 0.499 | 798.274 |

轮廓系数>？DBI<?，CH>?,说明聚类效果好，分类可靠。

（轮廓系数：对于一个样本集合，它的轮廓系数是所有样本轮廓系数的平均值。轮廓系数的取值范围是[-1,1]，同类别样本距离越相近不同类别样本距离越远，分数越高，聚类效果越好。  
DBI（Davies-bouldin)：该指标用来衡量任意两个簇的簇内距离之后与簇间距离之比。该指标越小表示聚类效果越好。  
CH(Calinski-Harbasz Score)：通过计算类内各点与类中心的距离平方和来度量类内的紧密度（分母），通过计算类间中心点与数据集中心点距离平方和来度量数据集的分离度（分子），CH指标由分离度与紧密度的比值得到，CH越大表示聚类效果越好。）(标成3个下标)

1. **NIPT时点计算**

**参考文献**  
[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.  
[2] Saroj,Kavita.Review:study on simple k mean and modified K mean clustering technique[J].International Journal of Computer Science Engineering and Technology,2016,6(7)：279-281.

接着我们进行每组的最佳 NIPT 时点的计算。