# 

**数据特征对多元统计方法结果影响的模拟实验设计与实现**

**Design and implementation of simulation experiments for demonstrating the principles of multivariate statistical concepts**

王亚茹1  王玖1 刘海霞1 侯静1 张莉1 郭佳珂1 孙红卫1

作者单位：1 滨州医学院公共卫生与管理学院

通信作者：孙红卫，Email：hwsun2000@163.com

# 3 研究结果

## 3.1 多重共线性对多重线性回归结果的影响

### 3.1.1 实验目的

通过模拟生成数据来直观演示多重共线性对多重线性回归结果的影响，便于直观理解。

### 3.1.2 实验设计

设置1个因变量和3个自变量，自变量为正态随机变量，分别模拟自变量间相关系数是0.1，0.5，0.7，0.9，0.95，0.99，1，依据，生成自变量与因变量，其中。样本量设置为100，看对应回归分析结果的变化。并图示结果，与真实的回归系数进行对照。

### 3.1.3 实验结果

表1 模拟自变量间相关系数r的变化对多重回归系数及其P值的影响

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 自变量 | r=0.1 | r=0.5 | r=0.9 | r=0.95 | r=0.98 | r=0.99 | r=1 |
| x1 | 1.04 | 1.00 | 0.98 | 0.93 | 0.86 | 1.69 | 2.10 |
| x2 | -0.97 | -1.00 | -0.97 | -1.14 | -0.56 | -2.52 | NA# |
| x3 | 2.13 | 2.02 | 1.93 | 2.32 | 1.69 | 2.82 | NA# |
| P1\* | <0.05 | <0.05 | <0.05 | <0.05 | <0.05 | <0.05 | <0.05 |
| P2\* | <0.05 | <0.05 | <0.05 | <0.05 | 0.14 | <0.05 | NA# |
| P3\* | <0.05 | <0.05 | <0.05 | <0.05 | <0.05 | <0.05 | NA# |
| VIF1 | 1.0 | 2.1 | 9.4 | 22.1 | 59.4 | 110.5 | NA# |
| VIF2 | 1.1 | 3.2 | 8.7 | 22.4 | 57.2 | 110.6 | NA# |
| VIF3 | 1.1 | 4.2 | 16.1 | 49.1 | 104.9 | 236.8 | NA# |

\*：分别是自变量x1，x2，x3回归系数假设检验的P值。

#：自变量呈现完全共线性时，回归系数无法估计。



图1 自变量间相关系数r=0.1时的散点图



图2 自变量间相关系数r=0.5时的散点图



图3 自变量间相关系数r=0.9时的散点图

由表1可以看到，当自变量间相关程度r=0.1，0.5，0.9时，估计的回归系数真实的回归系数很接近，但当r≥0.95时，估计的回归系数与真实的回归系数差异变大，且相关越大，差异越大，当r=1时，也就是变量间完全共线性时，回归系数无法估计，对应的假设检验P值也无法估计。

容忍度（或容差，Tolerance），等于1减去该自变量为因变量，其他自变量为自变量得到的线性回归模型的决定系数。该自变量与其它变量相关越大，其决定系数越大，其容忍度越小。VIF(方差膨胀因子)，即容忍度的倒数。一般不应大于5。从表1中可以看到，随着自变量相关程度r的增大，VIF增大，特别当r≥0.95时，VIF远远大于5，表明存在多重共线性。



图4 自变量回归系数随自变量相关系数r变化曲线图

### 3.1.4 实验结论

自变量高度相关时，特别是r≥0.95时，自变量间的多重共线性会使得回归方程的回归系数估计不准确，与实际差异很大，假设检验的P值有时会出现与实际不符，此时的VIF远远大于5。特别是r=1时，自变量间完全线性，此时回归系数无法估计。当自变量出现多重共线性时，可以通过逐步回归，可以通过剔除变量，仅保留高度相关变量中的一个变量来解决多重共线性，也可以通过主成分回归、岭回归或LASSO回归等来解决。

## 3.2 变量间不同相关模式对主成分分析结果的影响

### 3.2.1 实验目的

通过模拟生成数据来直观演示变量间不同相关模式对主成分分析和因子分析结果的影响。

### 3.2.2 实验设计

设置以下两种相关模式，生成特定相关模式下的正态随机变量x1，x2与x3，x4；

（1）相关模式1：四个变量间彼此相关系数相同。看相关系数r分别是0.1，0.5，0.9时，对主成分分析特征值、贡献度、KMO的影响。

（2）相关模式2：四个变量间呈现分组相关。x1与x2相关系数a较高，x3与x4相关系数b较高，而x1，x2与x3，x4间相关c较低。比较相关模式2与相关模式1对应的主成分分析特征值、贡献度、KMO及因子分析结果。

设自变量相关矩阵为

表2 相关矩阵设计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | x1 | x2 | x3 | x4 |
| x1 | 1.0 | a | c | c |
| x2 | a | 1.0 | c | c |
| x3 | c | c | 1.0 | b |
| x4 | c | c | b | 1.0 |

相关模式1：

情景1：a=b=c=0.1

情景2：a=b=c=0.5

情景3：a=b=c=0.9

相关模式2：

情景4：a=b=0.9，c=0.1

情景5：a=b=0.5，c=0.1

### 3.2.3 实验结果

相关模式1，相关矩阵中a=b=c=r，其变量间相关阵的散点图如下。



图5 情景1(相关系数为0.1)时自变量间散点图



图6 情景1(相关系数为0.9)时自变量间散点图

表3 相关模式1下三种情景对主成分结果的影响

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 主成分 | 情景1 | | 情景2 | | 情景3 | |
| 方差 | 累积贡献 | 方差 | 累积贡献 | 方差 | 累积贡献 |
| PC1 | 0.32 | 0.31 | 0.58 | 0.58 | 0.90 | 0.90 |
| PC2 | 0.26 | 0.57 | 0.17 | 0.76 | 0.04 | 0.94 |
| PC3 | 0.22 | 0.80 | 0.15 | 0.90 | 0.03 | 0.97 |
| PC4 | 0.11 | 1.00 | 0.10 | 1.00 | 0.02 | 1.00 |
| KMO | 0.56 | | 0.76 | | 0.87 | |

由表3可以看到，随着相关系数的增大，第一主成分的方差变大，其余主成分的方差减小，当r=0.9时，第一主成分的方差占总方差的90%。KMO值也随着自变量间相关程度增大而增大，由0.56增大到0.87。

表4 相关模式1下三种情景下因子分析的结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | 情景1 | | | 情景2 | | | 情景3 | |
| F1 | F2 | h | F1 | F2 | h | F1 | h |
| x1 | 0.32 | 0.40 | 0.45 | 0.64 | 0.14 | 0.67 | 0.95 | 0.91 |
| x2 | 0.00 | 1.00 | 0.83 | 0.01 | 1.00 | 0.93 | 0.95 | 0.91 |
| x3 | 0.99 | 0.00 | 0.52 | 1.00 | 0.00 | 0.85 | 0.95 | 0.90 |
| x4 | 1.00 | 0.00 | 0.49 | 0.56 | 0.19 | 0.57 | 0.95 | 0.90 |

F1：每个变量在因子上的因子载荷；h：共同度。

图7 相关模式1下的因子载荷

由表4，图7可以看到，相关模式1下，4个自变量全相关。情景1下的因子分析结果，提取2个因子，第一因子与变量x3，x4相关较大，第二因子与x2相关较大，共同度可知，x2的共同度较高，在0.8以上，表示因子对变量x2提取较好，而对x1，x3，x4提取效果较差，此时的KMO值是0.56，提取两个因子的累积贡献度是57%，说明提取两个因子不足以代表4个变量，需要增加因子，但增加因子后降低了降维效率。情景2，提取2个因子，第一因子与变量x1，x3，x4相关较大，体现了x1，x3，x4的共性。第二因子与x2相关较大，提取两个因子的累积贡献度是76%，KMO值是0.76，提取两个因子尚可。情景3下的因子分析结果，提取1个因子，与自变量的因子载荷都很高。共同度较高，都在0.9以上。提取1个因子的累积贡献率是90%，KMO值0.87，此时因子分析的降维效率最高。综上，当变量间全相关时，相关程度较高时，因子分析效果较好，变量间相关程度较低时，因子分析效果不好。

相关模式2：以情景4和情景5都是分组相关，其中情景4组内相关为高度相关0.9，情景5组内相关系数为中度相关0.5，以情景4为例，其变量间相关阵的散点图如下。



图8 情景4下自变量间散点图

表5 自变量间呈现分组相关时对主成分结果的影响

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 主成分 | 情景3 | | 情景4 | | 情景5 | |
| 方差 | 累积贡献 | 方差 | 累积贡献 | 方差 | 累积贡献 |
| PC1 | 0.90 | 0.90 | 0.55 | 0.55 | 0.46 | 0.46 |
| PC2 | 0.04 | 0.94 | 0.41 | 0.96 | 0.28 | 0.74 |
| PC3 | 0.03 | 0.97 | 0.02 | 0.98 | 0.14 | 0.87 |
| PC4 | 0.02 | 1.00 | 0.02 | 1.00 | 0.12 | 1.00 |
| KMO | 0.87 | | 0.51 | | 0.52 | |

由表5可知，情景3的四个自变量全高度相关，情景4与情景5呈现分组相关（x1与x2相关，x3与x4相关），三者比较可知，情景3的第一主成分占比90%，而情景4的第一和第二主成分的贡献分别是55%和41%，情景5也是分组相关，但组间相关只有0.5，直到3个主成分总贡献才达到87%。KMO值在全相关时达到最大0.87，分组相关的情景4和情景5的KMO值分别为0.51和0.52。

表6 三种情景下因子分析的结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | 情景3 | | 情景4 | | | 情景5 | | |
| F1 | h | F1 | F2 | h | F1 | F2 | h |
| x1 | 0.95 | 0.91 | 0.96 | 0.09 | 0.92 | 0.79 | 0.10 | 0.63 |
| x2 | 0.95 | 0.91 | 0.95 | 0.11 | 0.92 | 0.53 | 0.25 | 0.34 |
| x3 | 0.95 | 0.90 | 0.12 | 0.95 | 0.91 | 0.09 | 0.76 | 0.58 |
| x4 | 0.95 | 0.90 | 0.15 | 0.95 | 0.91 | 0.27 | 0.53 | 0.35 |

F：每个变量在因子上的因子载荷；h：共同度。

图9 相关模式2下的因子载荷

由表6可以看到，情景3下的因子分析结果，提取1个因子，与自变量的因子载荷都很高，共同度较高，都在0.9以上。情景4，提取2个因子，第一因子与变量x1，x2相关较大，体现了x1，x2，第二因子与x3，x4相关较大，体现了x3，x4的共性，且共同度均较高，都在0.9以上。与情景3相比，情景4中如果提取1个因子，贡献率只有55%，此时一个因子只能体现x1，x2，缺少x3，x4的信息，而提取了2个因子，累积贡献率达到了96%，每个原变量的共同度也达到了0.9以上。与情景2相比，虽然情景2的KMO值是0.76，而情景4的KMO值是0.51，但情景4累积贡献度和共同度都要高于情景2，情景4的因子分析效果好于情景2。

情景5时，提取2个因子，2个因子的累积贡献率达到了74%，第一因子与变量x1，x2相关较大，第二因子与x3，x4相关较大，但共同度在0.3~0.7之间，因子分析效果不如情景4。

综上，当变量间呈分组相关时，组内相关程度较高时，因子分析效果较好，组内相关程度较低时，因子分析效果不好。而分组相关时，KMO值通常偏低。

### 3.2.4 实验结论

本次实验可知，当变量间相关较高时，主成分和因子分析的降维效果较好，少数主成分和因子能够代表原变量。当变量间相关较低时，主成分和因子分析的降维效果不好，少数主成分和因子不能够代表原变量，信息提取效果较差。当变量呈现分组相关且组内相关较大时，主成分和因子分析结果也较好，但是此时KMO值偏低。所以要综合看贡献度，共同度和因子载荷、KMO值综合来看因子分析结果。另外由于因子分析有较强的主观性在因子分析的实际应用中，还需要认真考虑公因子的实际意义，以确定实际案例中因子分析的适用性。

## 3.3 演示自变量对因变量非线性影响下，线性回归拟合和非线性回归拟合对Logsitic回归结果的影响

### 3.3.1 实验目的

通过模拟生成数据来演示在Logistic回归中，当自变量对因变量是非线性影响时，分别用线性回归拟合和非线性回归拟合估计结果的差异。

### 3.3.2 实验设计

相关研究表明定量自变量X（如产妇年龄）与二分类因变量Y（如是否低体重儿）之间存在相关关系[21]。本实验以此为背景，生成模拟数据，见图10。样本量N=500，自变量，反应变量Y服从二项分布，其中，为Y=1的概率，满足。分别将Y和X做线性Logistic回归分析和非线性Logistic回归分析，对比两种方法的拟合效果。



图10 自变量X与因变量Y的关系

### 3.3.3 实验结果

表7 模拟情景下Logistic回归线性拟合和非线性拟合的估计系数

|  |  |  |  |
| --- | --- | --- | --- |
| 自变量 | | 系数 | P |
| 真实值 | X | -3 | -- |
| X2 | 1 | -- |
| 线性拟合 | X | -2.14 | 0.004 |
| 截距 | 0.80 | <0.001 |
| 非线性拟合 | X | 0.47 | 0.003 |
| X2 | -3.63 | <0.001 |
| 截距 | 1.25 | 0.667 |

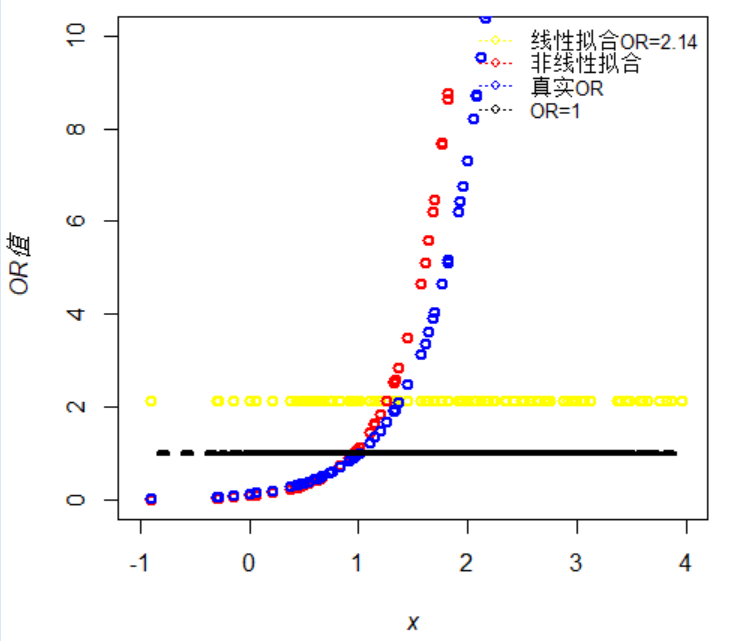


图11 线性拟合和非线性拟合的OR值比较



图12 线性模型和非线性模型的ROC曲线比较

表7为Logistic回归拟合结果，两种模型的系数均有统计学意义。进一步对线性回归和非线性回归的拟合效果加以比较。由图11知，线性拟合结果认为，OR=2.14，即自变量X是Y=1发生的危险因素，自变量越大，Y=1发生的可能性越大。非线性拟合结果认为，OR是随自变量的变化而变化的，X较小时，OR<1，随着X的增大，Y=1发生的风险逐渐降低但降低的幅度逐渐减小；X较大时，OR>1，随着X的增大，Y=1发生的风险逐渐增高且增幅也不断增高，即X过小或者过大都会造成Y=1发生的可能性较大。非线性模拟结果与本实验背景（孕妇年龄对低体重儿发生的影响）较为贴合，孕妇年龄较小时，随着年龄增长，低体重儿发生的风险逐渐降低；孕妇年龄较大时，低体重儿发生的风险随年龄的增长逐渐增高。进一步将两种拟合结果OR值与实际OR值比较可知，非线性拟合情形与实际情况确实更为吻合。由图12，通过比较ROC曲线验证，线性模型AUC=0.727，非线性模型AUC=0.881，以上结果均表明，在本实验背景下，非线性模型的拟合效果更好。

### 3.3.4 实验结论

通过模拟可以看到，如果忽视了Logistic回归中的线性性假设，会得到脱离实际的结论。至于如何通过数据来看线性性假设是否满足，可以通过散点图或者统计描述（如不同年龄组患病率）来检测是否存在非线性关系。如果存在非线性关系，在拟合Logistic回归模型时，如果自变量是定量资料，可以通过设置如模拟所示的二次项，如果是等级资料，可以通过设置哑变量（通常以患病率最低作为参考类）来看自变量与因变量之间的关联。

## 3.4 不同删失比例对Cox比例风险模型结果的影响

### 3.4.1 实验目的

通过模拟生成数据来直观演示不同删失比例对Cox比例风险模型结果的影响。

### 3.4.2 实验设计

设置包含两个自变量的生存数据，样本量N=1000，其中，x1和x2为自变量，服从多元正态分布，构建风险函数，假设指数分布生成生存时间，分别设置删失比例为5%，10%，20%，30%，40%，50%，60%，70%的情形，看Cox比例风险模型拟合结果的变化。

### 3.4.3 实验结果

表8 模拟不同删失比例变化对Cox比例风险模型系数及其P值的影响

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 自变量 | 真实值 | 5% | 10% | 20% | 30% | 40% | 50% | 60% | 70% |
| x1 | 1 | 0.931 | 0.898 | 0.816 | 0.732 | 0.730 | 0.656 | 0.501 | 0.482 |
| x2 | -1 | -0.924 | -0.924 | -0.806 | -0.751 | -0.680 | -0.676 | -0.515 | -0.593 |
| P1 | -- | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| P2 | -- | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| HR1 | 2.718 | 2.537 | 2.455 | 2.261 | 2.079 | 2.075 | 1.927 | 1.650 | 1.619 |
| HR2 | 0.368 | 0.397 | 0.397 | 0.447 | 0.472 | 0.507 | 0.509 | 0.598 | 0.618 |

由表8可知，当生存数据的删失比例为5%，10%，20%，30%，40%，50%时，估计的回归系数与真实的回归系数较为接近，而当删失比例达到60%时，估计的回归系数与真实的回归系数存在较大偏差，风险比HR1与真实值相比降幅超过50%（2.718vs1.650），风险比HR2与真实HR2相比增幅超过50%（0.368vs0.598），认为偏倚较大。随着删失比例的增大，这种差异也在增大。由以上结果认为，删失比例达到60%，结果不可信。



图13 自变量回归系数随生存数据删失比例变化曲线图

### 3.4.4 实验结论

当生存数据的删失比例较高时，特别是达到了60%，删失数据会使得Cox比例风险模型的估计不准确，与真实值相差较大，Cox模型估计结果不可靠。所以在收集预后病例资料时，要随访时间足够长，保证删失比例不要过大，从而使得分析结果更可靠。

## 3.5 变量间不同相关结构对典型相关分析结果的影响

### 3.5.1 实验目的

通过模拟生成数据来直观演示变量间不同相关结构对典型相关分析结果的影响。

### 3.5.2 实验设计

相关模式设置：两组正态随机变量（x1，x2），（y1，y2），变量间相关程度r相同，r分别是0.1，0.5，0.9，看对典型相关系数，典则变量，典型载荷的影响。

表9 相关矩阵设计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | x1 | x2 | y1 | y2 |
| x1 | 1.0 | a | a | a |
| x2 | a | 1.0 | a | a |
| y1 | a | a | 1.0 | b |
| y2 | a | a | a | 1.0 |

情景1：a=0.1

情景2：a=0.5

情景3：a=0.9

### 3.5.3 实验结果

表10 3种情景下典型相关系数

|  |  |  |  |
| --- | --- | --- | --- |
| 典型相关系数 | 情景1 | 情景2 | 情景3 |
| rC，1 | 0.237 | **0.742** | **0.933** |
| rC，2 | 0.040 | 0.043 | 0.057 |
| P1\* | 0.222 | <0.001 | <0.001 |
| P2\* | 0.693 | 0.673 | 0.575 |

\*：典型相关系数假设检验的P值。

表11 不同情景下的典型载荷

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 情景1 | | 情景2 | | 情景3 | |
|  | U1 | U2 | **U1** | U2 | **U1** | U2 |
| x1 | -0.931 | 0.364 | **-0.815** | -0.579 | **-0.971** | -0.237 |
| x2 | 0.379 | 0.925 | **-0.902** | 0.433 | **-0.966** | -0.260 |
| 方差 | 0.505 | 0.495 | **0.739** | 0.261 | **0.938** | 0.062 |
|  | V1 | V2 | **V1** | V2 | **V1** | V2 |
| y1 | -0.915 | 0.403 | **-0.937** | -0.351 | **-0.959** | -0.285 |
| y2 | 0.374 | 0.928 | **-0.826** | 0.564 | **-0.975** | -0.220 |
| 方差 | 0.489 | 0.511 | **0.780** | 0.220 | **0.935** | 0.065 |

由表11，可以看到，两组变量间全相关时，分低、中、高三种相关强度进行分析，即情景1，情景2，情景3。情景1下，典型相关系数均无统计学意义，且第一典则变量方差较小，说明提取效果差。情景2，第一对典则变量相关有统计学意义，呈现相关程度较高，且第一对典则变量方差均大于0.7，提取效果较好，典型载荷结果显示，U1与x1，x2均相关，V1与y1，y2均相关。情景3，第一对典则变量相关有统计学意义，呈高度相关，且第一对典则变量的方差均大于0.9，提取效果很好，优于情景2，其中，U1与x1，x2相关程度均较高，V1与y1，y2相关程度均较高。

计算数据在典型变量下的得分，可得到3种情形下求得的典则变量U1与V1、U2与V2为坐标的数据散点图如下。



图14 情景1下典则变量为坐标的散点图



图15 情景2下典则变量为坐标的散点图



图16 情景3下典则变量为坐标的散点图

### 3.5.4 实验结论

典型相关分析用于研究两组变量之间的线性相关情况，从每组变量中分别找出线性组合，使其相关性最大，以此实现相关的提取，用少数综合变量来反应两组变量的相关关系。本次实验从原始变量相关性的角度，探寻典型相关的适用情况。由上述结果显示，当变量间呈现低度相关时，典型相关分析无统计学意义，不适合做典型相关分析；呈现中度相关或高度相关时，典则变量的提取效果较好，降维效率高，典型相关分析效果较好。

## 3.6 类别比例不同对距离判别和Bayes判别、Logistic回归结果的影响

### 3.6.1 实验目的

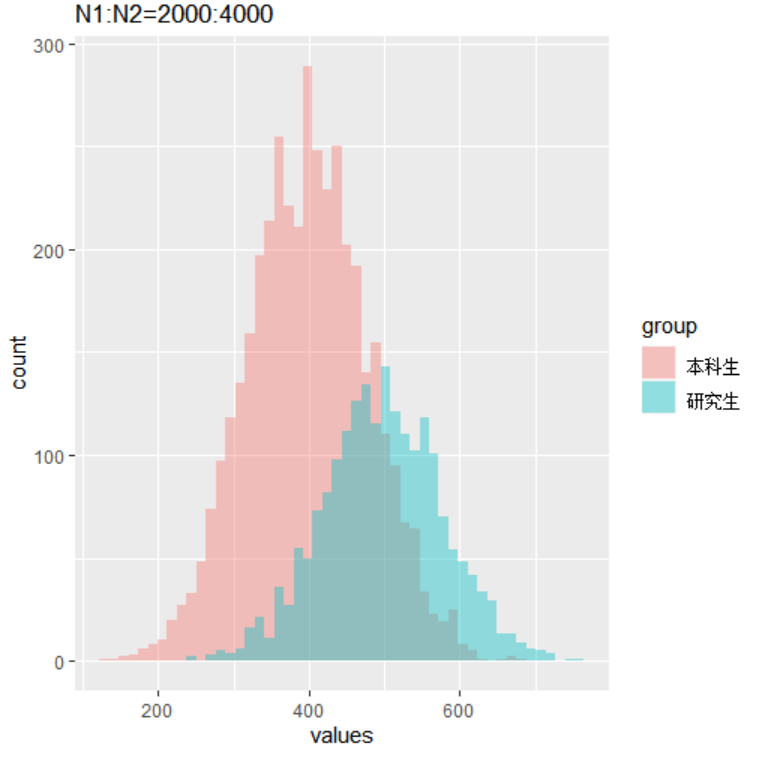
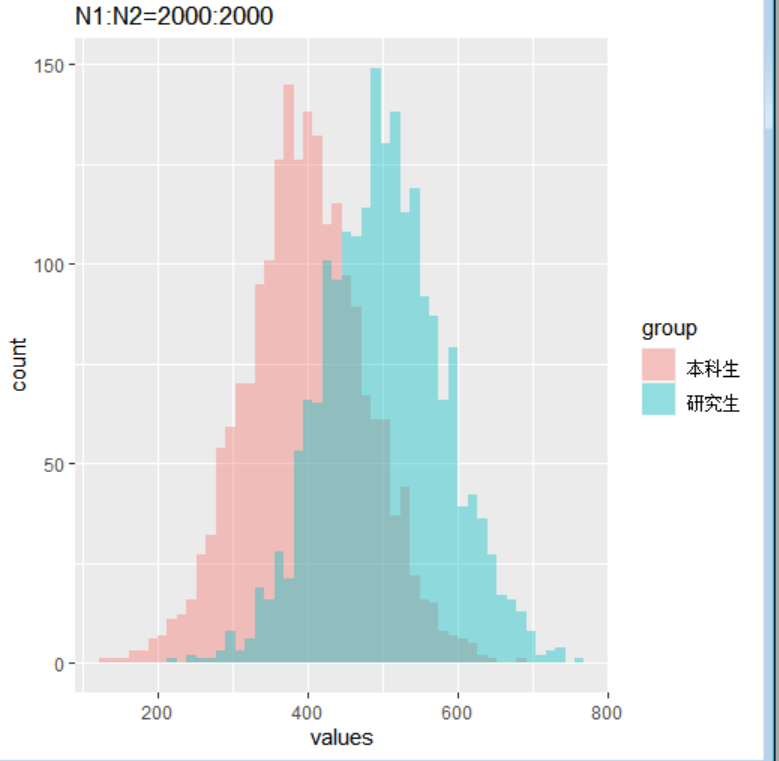
通过模拟生成数据来演示类别比例不同对距离判别和Bayes判别、Logistic回归结果的影响。

### 3.6.2 实验设计

本实验以大学英语六级成绩为背景，判别变量是英语六级考试成绩X（满分710分）。其中，假设校研究生组的成绩，校本科生组的成绩。若某该校学生六级成绩500分，试判别该生属于哪一组。

模拟不同情境，校研究生组的人数N1与校本科生组人数N2之比分别是1：1（N1=2000，N2=2000）、1：2（N1=2000，N2=4000）、1：3（N1=2000，N2=6000）、1：4（N1=2000，N2=8000），其分布见图17。

判别方法选择距离判别，Bayes判别，Logistic回归3种方法。其中，距离判别采用马氏距离得到待测样本与两组重心的距离，以此作为判别依据；Bayes判别考虑到了先验概率，即利用已知数据得到属于研究生组和属于本科生组的基本概率，以Bayes公式推出后验概率；Logistic回归进行判别预测时，以预测概率作为截断值。



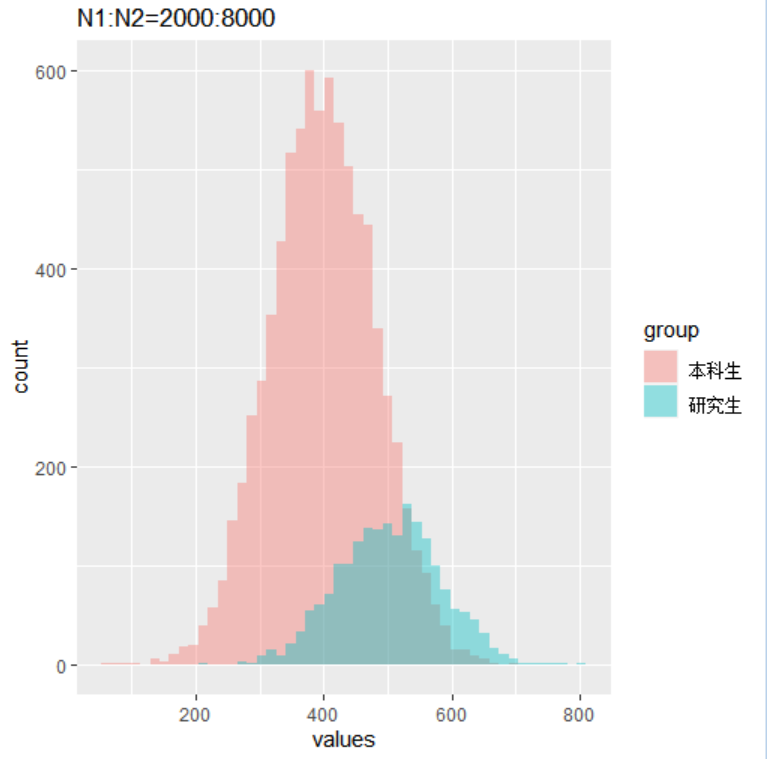
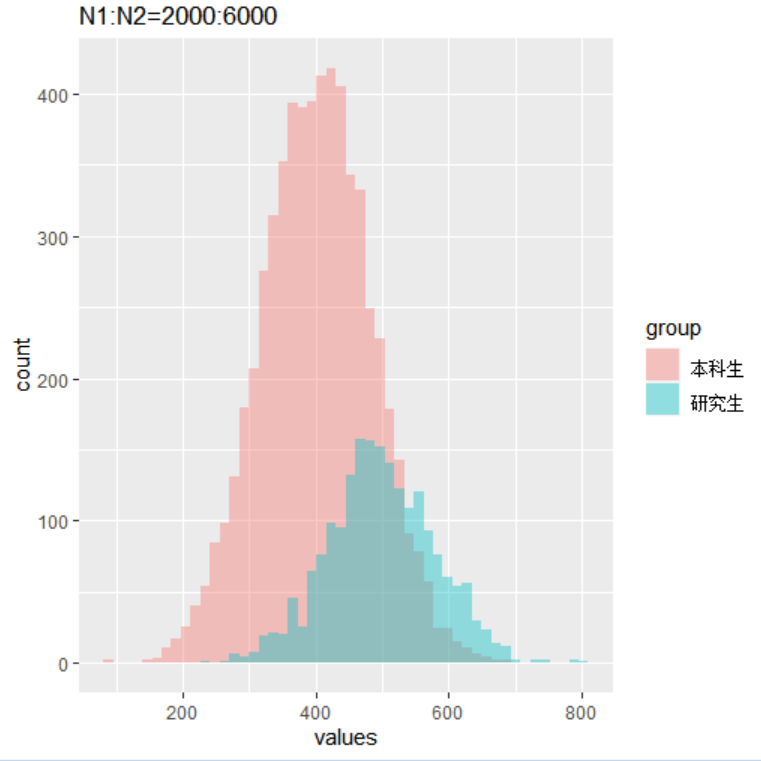


图17 不同情形下研究生与本科生六级成绩的分布直方图

### 3.6.3 实验结果

表12 模拟不同类别比例下3种判别方法对某同学成绩为500分的判别结果比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 判别方法 | N1:N2=1:1 | N1:N2=1:2 | N1:N2=1:3 | N1:N2=1:4 |
| 错判率 | 距离判别 | 0.261 | 0.270 | 0.269 | 0.261 |
| Bayes判别 | 0.261 | 0.237 | 0.201 | 0.167 |
| Logistic回归 | 0.261 | 0.237 | 0.201 | 0.167 |
| 判别结果 | 距离判别 | 研究生 | 研究生 | 研究生 | 研究生 |
| Bayes判别 | 研究生 | 研究生 | 本科生 | 本科生 |
| Logistic回归 | 研究生 | 研究生 | 本科生 | 本科生 |

由表12可以看到，当校研究生和校本科生的比例在1：1，距离判别、Bayes判别和Logistic回归的错判率相等，随着本科生比例的增加，距离判别的错判率有所增加，而Bayes判别和Logistic回归判别错判率逐渐减低。从判别结果看，N1：N2=1：1、N1：N2=1：2时，三者判别结果一致，均认为成为500的学生大概率属于校研究生组；当校研究生和校本科生的比例在1：3和1：4时，距离判别结果仍为校研究生，而Bayes判别和Logistic回归结果认为，该生属于校本科生。综上可知，随着校本科生人数占比的增大，Bayes判别和Logistic回归的错判概率更低，尤其在校本科生的数量达到校研究生数量的3倍时，判别结果发生了改变。从错判率看，此时认为Bayes判别和Logistic回归结果更好。

### 3.6.4 实验结论

距离判别的原理是比较样本与两个总体的距离，认为样本来自距离较近的总体；Bayes判别则充分利用了先验概率，得到样本的后验概率分布，以此作为判别依据；Logistic回归意在对反应变量Y的取值概率建模，根据预测变量X的不同水平计算Y=1的预测概率。该实验结果表明，当两组的类别相差较大时，距离判别不适用，适合用Bayes判别或Logistic回归。Bayes判别与Logistic回归比较，以类别比例作为先验概率时，两者等价；而在实际应用中，训练样本的类别比例可能不适合作为先验概率，比如疾病诊断，通常以该病的发病率或患病率为先验概率，此时Bayes判别由于存在先验概率（先验概率可以取各类疾病的发病率、患病率、门诊的就诊比例等），在应用上较为灵活，相较于Logistic回归更具优势。