对一次某省十校理科班5000名学生联考的成绩进行分析，内容包括如下：

1. 画出各科分数的概率分布
2. 构造各科各校成绩概览表，进行对比
3. 算出每位学生每科及总分的累积概率密度函数值，以此对学生进行重新排名。对比该排名与总分排名的变化
4. 对数学和语文成绩、数学和英语成绩做相关性检验，看数学成绩与其中哪一科关联更大
5. 说明

个别材料放在文章末尾，包括：封装好的代码，考试成绩总表。

1. 任务一，各科成绩概率分布

尽管R可以直接从xls文件读取数据，这边已将xls文件内容写入MySQL（见mysql\_client\_python.py），并且封装了RMySQL接口（见mysql\_client\_r.R），因此直接从MySQL读取。如下：

source('mysql\_client\_r.R') #加载封装好的接口

data\_frame <- get\_table('ten\_schools\_exams')

#获取整个表，get\_table()方法从mysql\_client\_r.R载入

开始前，先看一下数据结构。

打印表格头部五行内容：

print(head(data\_frame, n=5))

序号 姓名 性别 学校 数学 语文 英语 物理 化学 生物 总分

1 1 金浩穰 男 镇海中学 144 139 137 117 99 77 713

2 2 廖又菡 女 温州中学 150 131 141 118 88 80 708

3 3 李嘉胜 男 鲁迅中学 140 123 148 119 95 80 705

4 4 卢悦人 女 学军中学 143 122 147 117 93 76 698

5 5 程鹏鲲 男 杭州第二中学 148 126 135 115 91 76 691

打印表格尾部五行内容：

print(tail(data\_frame, n=5))

序号 姓名 性别 学校 数学 语文 英语 物理 化学 生物 总分

4996 4996 卢雁凡 女 温州中学 85 95 107 82 62 42 473

4997 4997 邱嘉玉 男 镇海中学 92 94 98 81 53 43 461

4998 4998 赖珂妍 女 慈溪中学 89 97 104 60 64 45 459

4999 4999 冯柔 女 杭州第二中学 72 103 95 68 62 49 449

5000 5000 宋俊才 男 慈溪中学 83 93 96 72 62 41 447

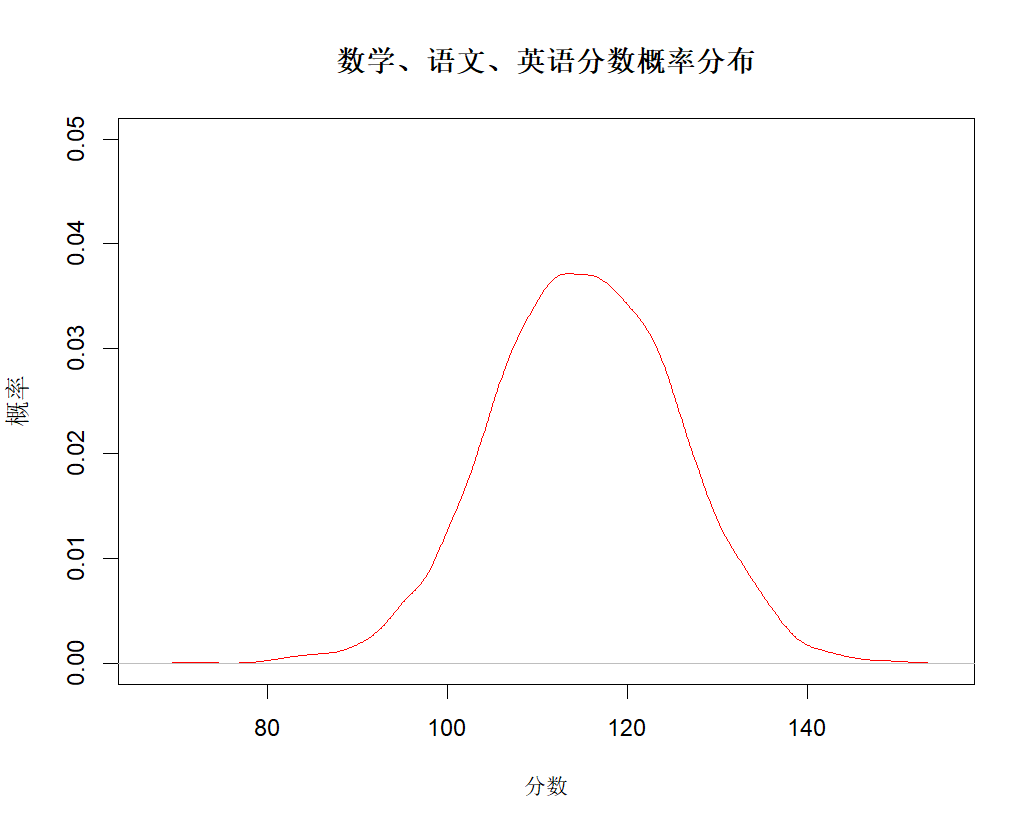
各科满分分别为，数学150，语文150，英语150，物理120，化学100，生物80。

现在，尝试画5000名学生数学、语文、英语三科成绩的概率分布。首先画数学：

math\_scores <- data\_frame[['数学']]

plot(density(math\_scores), ylim=c(0, 0.05), col='red', main='数学、语文、英语分数概率分布', xlab='分数', ylab='概率')

第一句，从表中选取“数学”字段，得到向量（是个一维数组）。第二句，density()方法获取了该数组的值和其对应的分布密度，分别作为x坐标和y坐标；ylim参数是y坐标的取值范围；col参数是分布函数的颜色；main、xlab、ylab三个参数分别是图名、x坐标轴名和y坐标轴名。输出的图如下：



接下来，在同一图上增加语文、英语两科成绩的概率分布函数。此时不再用plot()方法，而用lines()方法：

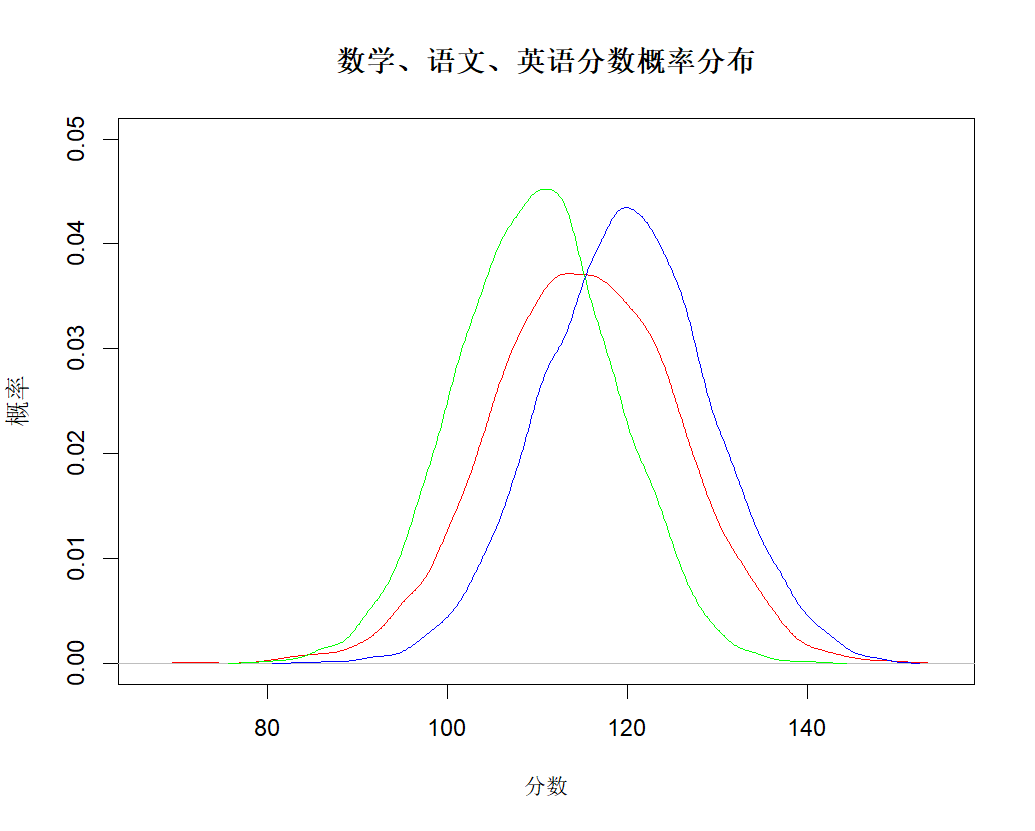
Chinese\_scores <- data\_frame[['语文']]

English\_scores <- data\_frame$英语

lines(density(Chinese\_scores), col='green')

lines(density(English\_scores), col='blue')

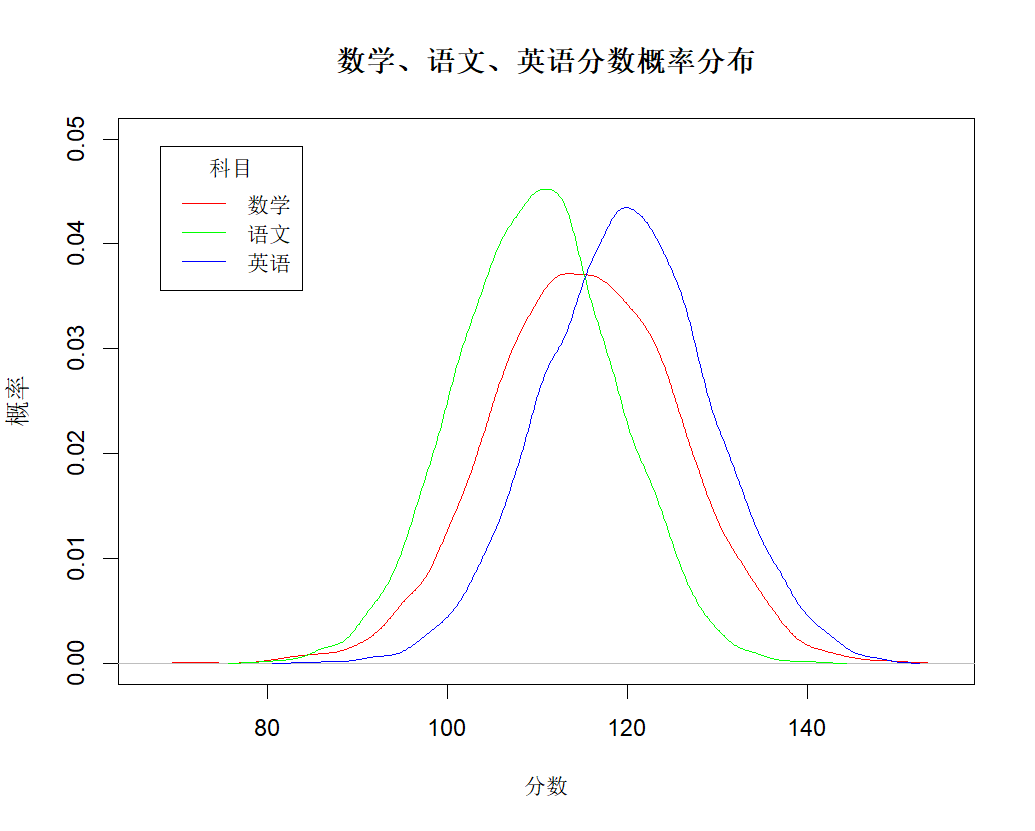
头两句，选取“语文”、“英语”两个字段，同样是两个向量。需要注意的是，data\_frame[['语文']]和data\_frame$语文，两种写法的效果是相同的。后两句，画图。此时图片变成如下：



三根线都出来了，但是不看代码的话，我们不知道那根线对应哪个科目，因此需要加个图例：

legend('topleft', inset=.05, title='科目', c('数学','语文','英语'), col=c('red','green','blue'), lty=c(1,1,1))

topleft代表图例放在左上方；title是图例的名称；c('数学','语文','英语')代表图例说明；col是三科的三种颜色；lty是line type，线的种类。请注意，后三个参数都是向量。此时图表变成如下：



可以看到，左上方图例出来了。

观察图片，5000名学生的三科成绩大致呈正态分布。试回答：

1. 三科中，哪一科更能拉分？从哪里可以看出？
2. 三科成绩的平均分和中位数大致是多少？
3. 任务二，各校各科成绩概览

需要显示各校各科成绩的关键信息，从而方便对比。这些信息包括：科目名称，学校名称，平均分，标准差，四分之一分位数（超过25%样本的分数门槛），四分之二分位数，四分之三分位数。

先来看看杭州高级中学学生的物理成绩概览。

library(sqldf)

source('mysql\_client\_r.R') #加载封装好的接口

data\_frame <- get\_table('ten\_schools\_exams') #获取整个表

首先同样加载mysql\_client\_r.R并获取整个表。但这里我们换一种方式获取“物理”字段，即使用R的sqldf包从data\_frame中获取（而非写sql从数据库读取）。R的sqldf将data\_frame当作数据库的表，传入的参数也与sql语句一致。为此，我们加载了sqldf包。接下来：

detach("package:RMySQL", unload=T)

physics\_hzgj <- sqldf('select 物理 from data\_frame where 学校="杭州高级中学"')$物理

四一分位 <- c(unname(quantile(physics\_hzgj, probs=0.25)))

四二分位 <- c(unname(quantile(physics\_hzgj, probs=0.5)))

四三分位 <- c(unname(quantile(physics\_hzgj, probs=0.75)))

平均分 <- c(mean(physics\_hzgj))

标准差 <- c(sd(physics\_hzgj))

学校 <- c('杭州高级中学')

physics\_hzgj <- data.frame(学校,四一分位,四二分位,四三分位,平均分,标准差)

print(physics\_hzgj)

第一句，释放了RMySQL包，如果不释放，R的sqldf包将与之冲突而无法使用；第二句，从data\_frame中获取了杭高学生的物理成绩。

然后，计算需要的数据。请留意quantile()方法，接收样本数据和概率作为参数，返回分位数。如其中第一句的意思是：分数样本physics\_hzgj中，达到多少分才能超过25%。Unname()方法去掉了字段名，只留下了数字。然后构造向量。

使用data.frame()方法构造数组。打印出结果如下：

学校 四一分位 四二分位 四三分位 平均分 标准差

1 杭州高级中学 90 95 101 95.206 8.686223

上述结果是杭高此次考试物理成绩的概览。可知，考到90分，就超过了该校25%的学生，95分就超过了50%，而101分就超过了75%。

然而，现在我们需要的是十所学校、6门科目以及总分的概览。太多了，我们不想把上述冗长的代码重复10乘7遍。为此，我们把上述代码修改成一个函数，接收学校名和科目为参数，返回该校该科目的成绩概览。然后，对每个科目下每所学校的成绩进行遍历。

先写这个函数：

get\_school\_subject\_summary <- function (school, subject) {

sql <- sprintf("select %s from ten\_schools\_exams where 学校= '%s'", subject, school)

school\_subject\_data <- get\_data(sql)[[subject]]

四一分位 <- c(unname(quantile(school\_subject\_data, probs=0.25)))

四二分位<- c(unname(quantile(school\_subject\_data, probs=0.5)))

四三分位 <- c(unname(quantile(school\_subject\_data, probs=0.75)))

平均分 <- c(mean(school\_subject\_data))

标准差 <- c(sd(school\_subject\_data))

学校 <- c(school)

科目 <- c(subject)

school\_subject\_summary <- data.frame(科目,学校,四一分位,四二分位,四三分位,平均分,标准差)

return (school\_subject\_summary)

}

函数主体与求杭高物理成绩概览的代码雷同，唯一需要注意的是第二行：sprintf()方法实际上用函数接收的subject和school两个参数，分别去替换了两个%s符号，返回的sql语句类似于select 物理 from ten\_schools\_exams where 学校= '杭州高级中学'。

可尝试调用该方法，运行print(get\_school\_subject\_summary(‘学军中学’, ‘化学’))，看看得到的结果格式是否与杭高物理的散装代码相同。

接下来，我们要遍历各科目，并在循环内再套一个循环，遍历各学校，然后调用上述函数。代码如下：

subjects <- c('数学','语文','英语','物理','化学','生物','总分')

schools <- unique(data\_frame$学校)

for (subject in subjects) {

subject\_ten\_schools <- data.frame(科目=character(0), 学校=character(0), 四一分位=integer(0),四二分=integer(0),四三分位=integer(0),平均分=numeric(0), 标准差=numeric(0))

for (school in schools) {

school\_subject\_summary <- get\_school\_subject\_summary(school, subject)

subject\_ten\_schools <- rbind(subject\_ten\_schools, school\_subject\_summary)

}

subject\_ten\_schools <- subject\_ten\_schools[order(-subject\_ten\_schools$平均分),]

print(subject\_ten\_schools)

}

第一句，构造了各科目名称的向量；第二句，构造了校名的向量，它是通过unique()方法，将从data\_frame中获得的“学校”字段内容去重获得的。

需要特别留意第四行，我们构造了一个名为subject\_ten\_schools的空数组，这是为了：在特定科目中，每次得到各校的成绩概览后，与该数组合并（rbind()方法接收多个数组，纵向拼接），从而最终得到该科目下十校成绩的总概览。

从第五句开始，遍历学校；第六句，我们调用了刚刚构造的函数，每次调用都获得一个名为school\_subject\_summary的成绩概览；第七句，使用rbind()方法不断将新得到的成绩概览并入老概览；第八句，按平均分对各校成绩进行排序。然后打印。

最终我们得到的结果如下：

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

9 数学 慈溪中学 109.00 116 123 115.692 9.920381

5 数学 杭州第二中学 109.00 116 122 115.440 10.154833

3 数学 鲁迅中学 108.75 115 122 115.426 10.111628

10 数学 杭州高级中学 108.00 116 122 115.384 10.319182

8 数学 诸暨中学 108.00 116 123 115.354 10.426414

7 数学 余姚中学 108.00 115 122 115.110 10.316140

4 数学 学军中学 108.00 115 122 115.106 10.265550

2 数学 温州中学 108.00 115 121 115.014 10.293966

1 数学 镇海中学 108.00 114 122 114.974 9.948632

6 数学 长兴中学 107.00 114 122 114.718 10.379496

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

10 语文 杭州高级中学 104 110.0 116 110.314 8.610208

1 语文 镇海中学 104 111.0 116 110.140 8.792441

3 语文 鲁迅中学 105 110.0 115 110.122 8.195697

7 语文 余姚中学 104 110.0 116 110.040 8.796656

2 语文 温州中学 104 109.5 116 110.026 8.851102

6 语文 长兴中学 104 110.0 116 110.006 8.813474

8 语文 诸暨中学 104 110.0 116 109.956 8.635462

4 语文 学军中学 103 109.5 116 109.706 8.965507

5 语文 杭州第二中学 104 110.0 115 109.624 8.682245

9 语文 慈溪中学 104 109.0 115 109.148 8.720447

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

2 英语 温州中学 114.75 120.5 127.00 120.348 9.496909

8 英语 诸暨中学 114.00 120.0 125.00 120.082 8.700498

10 英语 杭州高级中学 114.00 119.5 126.25 120.050 9.617510

3 英语 鲁迅中学 114.00 120.0 126.00 119.990 9.143364

5 英语 杭州第二中学 114.00 120.0 126.00 119.940 9.305137

7 英语 余姚中学 114.00 120.0 125.00 119.924 8.951563

1 英语 镇海中学 114.00 120.0 126.00 119.920 9.010659

4 英语 学军中学 113.00 120.0 127.00 119.806 9.592309

6 英语 长兴中学 113.00 120.0 126.00 119.720 9.749446

9 英语 慈溪中学 113.00 120.0 125.00 119.362 9.013403

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

8 物理 诸暨中学 90 96.0 101 95.446 8.469389

10 物理 杭州高级中学 90 95.0 101 95.206 8.686223

2 物理 温州中学 90 95.0 101 95.174 8.439798

3 物理 鲁迅中学 89 95.0 100 95.048 8.286315

6 物理 长兴中学 89 95.5 101 95.004 8.573856

1 物理 镇海中学 89 94.0 101 94.954 8.370425

9 物理 慈溪中学 89 95.0 101 94.938 8.891134

4 物理 学军中学 89 95.0 100 94.910 8.242924

5 物理 杭州第二中学 90 95.0 101 94.820 8.252503

7 物理 余姚中学 88 94.0 101 94.530 8.769430

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

3 化学 鲁迅中学 73 78 83.00 78.406 7.476695

10 化学 杭州高级中学 73 78 83.00 78.272 7.349697

4 化学 学军中学 73 78 83.00 78.240 7.495249

2 化学 温州中学 73 78 83.00 78.216 7.372248

8 化学 诸暨中学 73 78 83.00 78.188 7.423670

9 化学 慈溪中学 74 78 82.25 78.138 7.150750

6 化学 长兴中学 73 78 83.00 77.976 7.684630

1 化学 镇海中学 73 78 83.00 77.932 7.787726

7 化学 余姚中学 73 78 82.25 77.748 7.355318

5 化学 杭州第二中学 72 77 83.00 77.704 7.543620

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

3 生物 鲁迅中学 58.00 62.5 67 62.396 6.777540

10 生物 杭州高级中学 58.00 62.0 67 62.336 6.821449

4 生物 学军中学 57.00 62.0 67 62.076 6.973341

8 生物 诸暨中学 57.00 62.0 67 62.072 7.140363

7 生物 余姚中学 57.00 62.0 67 62.030 6.974697

2 生物 温州中学 57.00 62.0 66 61.996 6.951880

1 生物 镇海中学 57.00 62.0 67 61.936 7.085783

9 生物 慈溪中学 58.00 61.5 66 61.852 6.448042

6 生物 长兴中学 57.75 62.0 66 61.696 6.885812

5 生物 杭州第二中学 57.00 61.0 66 61.652 6.984588

科目 学校 四一分位 四二分位 四三分位 平均分 标准差

10 总分 杭州高级中学 559.00 579.5 603.00 581.562 34.60667

3 总分 鲁迅中学 561.00 581.0 604.00 581.388 32.55877

8 总分 诸暨中学 559.00 579.0 603.00 581.098 33.48610

2 总分 温州中学 557.75 579.0 604.00 580.774 35.28224

1 总分 镇海中学 557.00 580.0 601.00 579.860 33.81626

4 总分 学军中学 558.75 579.0 604.00 579.844 33.99516

7 总分 余姚中学 556.00 579.0 602.25 579.382 34.39841

5 总分 杭州第二中学 556.75 578.0 600.25 579.180 33.50290

9 总分 慈溪中学 557.75 579.0 602.00 579.130 33.63885

6 总分 长兴中学 554.00 582.0 601.00 579.120 35.24961

我们发现，不论是单科还是总分，十校的成绩各指标差距都不大，不符合实际情况。这是由于这份数据原本就是虚构的，且对于各校学生的分数构造得过于均匀了。

忽略上述情况，结合该结果，试分析：

1. 仅英语而言，哪所学校整体成绩最好？哪所学校内部成绩波动程度最小？
2. 就总分而言，杭高的平均分优于温州中学，而温州中学的75%分位数却更高，这说明了什么？
3. 任务三，两种排名方法的对比（总分累积概率密度排名Vs.总分排名）

我们常听到一种说法，科目A拉分，科目B不拉分，因此科目A绝不能掉队，宁可少花点时间在科目B上。这种说法有其道理。试观察图片《数学、语文、英语分数概率分布》，考虑：本次十校联考中，数学考140、语文考140、英语考140，哪个更厉害？

更直观的，执行以下代码，计算140的分数在三科中的累积概率密度：

library(sqldf)

source('mysql\_client\_r.R') #加载封装好的接口

data\_frame <- get\_table('ten\_schools\_exams') #获取整个表

print(pnorm(140, mean=mean(data\_frame$数学), sd=sd(data\_frame$数学)))

print(pnorm(140, mean=mean(data\_frame$语文), sd=sd(data\_frame$语文)))

print(pnorm(140, mean=mean(data\_frame$英语), sd=sd(data\_frame$英语)))

首先依然加载sqldf和mysql\_client\_r.R， 然后调用get\_table()获取整个表。接着，利用pnorm()，得到三个科目中140分对应的累积概率密度。

pnorm()可以理解为quantile()的反函数，前者根据分位数计算累积概率密度（考140分，能超过多少比例的人？），而后者根据累积概率密度计算分位数（超过95%的人，要考多少分？）。pnorm()主要接收三个参数：第一个是分位数（可以是一个数组）；第二个是平均数，这里通过mean()方法获得；第三个是标准差，这里通过sd()方法获得。

计算所得结果如下：

1. 0.9923879
2. 0.9997261
3. 0.9849729

这说明，在这次联考中，语文考140分超过了99.97%的人，要比数学考140分（99.24%）更厉害，而后者又比英语考140分（98.50%）更厉害。这与图表反映出的情况相符，也与我们“语文不易拉分”的认知相符。

那么，在这次联考中，以各科分数总和作为评判标准，是否合理呢？还是说以各科分数对应的累积分布概率总和作为评判标准，会更好？不妨对后者进行尝试。

首先，为了方便理解，以化学为例，在数组中加一列名为“化学\_pnorm”的向量，用来记录各位考试的化学成绩对应的累积概率密度。

data\_frame[['化学\_pnorm']] <- pnorm(data\_frame$化学, mean=mean(data\_frame$化学), sd=sd(data\_frame$化学))

print(head(data\_frame, n=5))

请注意，这里pnorm()接收的第一个参数是所有化学成绩，是一个数组，而不同于刚刚的单个成绩（即140分）。打印前五行，输出结果如下：

序号 姓名 性别 学校 数学 语文 英语 物理 化学 生物 总分 化学\_pnorm

1 1 金浩穰 男 镇海中学 144 139 137 117 99 77 713 0.9974693

2 2 廖又菡 女 温州中学 150 131 141 118 88 80 708 0.9080844

3 3 李嘉胜 男 鲁迅中学 140 123 148 119 95 80 705 0.9883073

4 4 卢悦人 女 学军中学 143 122 147 117 93 76 698 0.9771997

5 5 程鹏鲲 男 杭州第二中学 148 126 135 115 91 76 691 0.9582797

接下来，把上述代码改写成一个循环，去遍历所有科目：

for (subject in c('数学','语文','英语','物理','化学','生物')) {

subject\_pnorm <- paste(subject,'\_pnorm', sep='')

data\_frame[[subject\_pnorm]] <- pnorm(data\_frame[[subject]], mean=mean(data\_frame[[subject]]), sd=sd(data\_frame[[subject]]))

}

特别留意第二行，我们用paste()方法将学科名称和\_pnorm拼接起来，形成“化学\_pnorm”这样的格式。

此外，还需一个“总分\_pnorm”向量。这个向量不通过对“总分”执行pnorm()方法获得，而是通过将各科的pnorm相加获得：

data\_frame$总分\_pnorm <- data\_frame$数学\_pnorm + data\_frame$语文\_pnorm + data\_frame$英语\_pnorm + data\_frame$物理\_pnorm + data\_frame$化学\_pnorm + data\_frame$生物\_pnorm

可以试着再次执行print(head(data\_frame, n=5))，打印数据，会发现所有科目以及总分的pnorm都在数组内了。这里不再打印。

现在我们尝试按“总分\_pnorm”字段给考试排序，并按此排序赋予“序号\_pnorm”字段，将之与原本的排名做对比。

data\_frame\_a <- data\_frame[order(-data\_frame$总分\_pnorm),]

index <- order(-data\_frame\_a$总分\_pnorm)

data\_frame\_a$序号\_pnorm <- index

data\_frame\_b <- data\_frame\_a[c('序号','序号\_pnorm','姓名','性别','学校','数学','语文','英语','物理','化学','生物','总分','总分\_pnorm')]

print(head(data\_frame\_b, n=10))

第一行，对data\_frame按“总分\_pnorm”字段进行排序；第二行，获取按同样方法排序所得的序号；第三行，将所得的序号作为字段“序号\_pnorm”放入data\_frame，所得数组命名为data\_frame\_a；第四行，从data\_frame\_a选取部分字段（以方便展示），成为data\_frame\_b。然后打印前十行：

序号 序号\_pnorm 姓名 性别 学校 数学 语文 英语 物理 化学 生物 总分 总分\_pnorm

1 1 1 金浩穰 男 镇海中学 144 139 137 117 99 77 713 5.942406

3 3 2 李嘉胜 男 鲁迅中学 140 123 148 119 95 80 705 5.906211

7 7 3 吴炎彬 男 余姚中学 135 127 138 113 95 79 687 5.887750

2 2 4 廖又菡 女 温州中学 150 131 141 118 88 80 708 5.880695

4 4 5 卢悦人 女 学军中学 143 122 147 117 93 76 698 5.863641

6 6 6 唐明煦 女 长兴中学 129 131 139 112 100 79 690 5.852785

5 5 7 程鹏鲲 男 杭州第二中学 148 126 135 115 91 76 691 5.843092

13 13 8 方盼香 女 温州中学 137 130 137 110 94 69 677 5.729838

8 8 9 卢琮 男 镇海中学 138 117 139 111 98 80 683 5.721619

25 25 10 方惜香 女 长兴中学 129 122 134 114 93 77 669 5.714481

观察上边发现，将总分排序改为总分累积概率密度排序后，原本第7名的吴炎彬同学，升至第3名，超过了原本第2名的廖又菡同学。观察两人的分数，廖又菡同学数学、生物满分，物理也接近满分，总分较吴炎彬同学高21分。

实际上，只要打印出各科的pnorm（打印data\_frame\_a而非data\_frame\_b，或者将上边代码第4行的科目均改为“科目\_pnrom”），就发现吴炎彬同学之所以超过了廖又菡同学，就在于他物理成绩的pnorm高出约0.08，虽然按分数来说只高了7分。

那么，新的排名方式，较之按总分排名，更合理吗？这个问题较复杂，不是对比少数几位学生的成绩就能有结论的，目前我所知的分析手段也无法给出判断。

但是，我认为这其中有两个影响因素。设想一种极端的情况：对于这六门科目，5000名学生的平均分都一致，标准差也一致，那么按总分排序和按总分\_pnorm排序，应当是同等可信的。

1. 科目相关性检验

常识告诉我们，数学成绩好的，一般物理也好，但语文就不一定。接下来尝试检验数学与物理、数学与语文直接的相关性。

首先以画图的方式定性地呈现上述相关性关系。

source('mysql\_client\_r.R') #加载封装好的接口

data\_frame <- get\_table('ten\_schools\_exams') #获取整个表

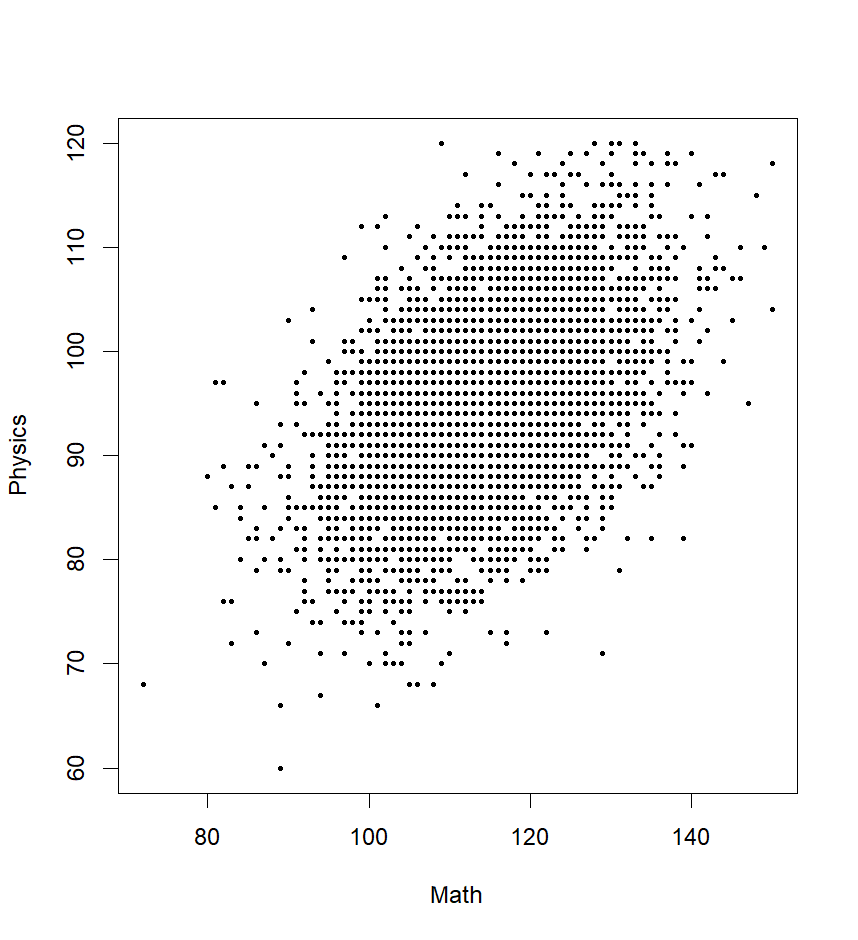
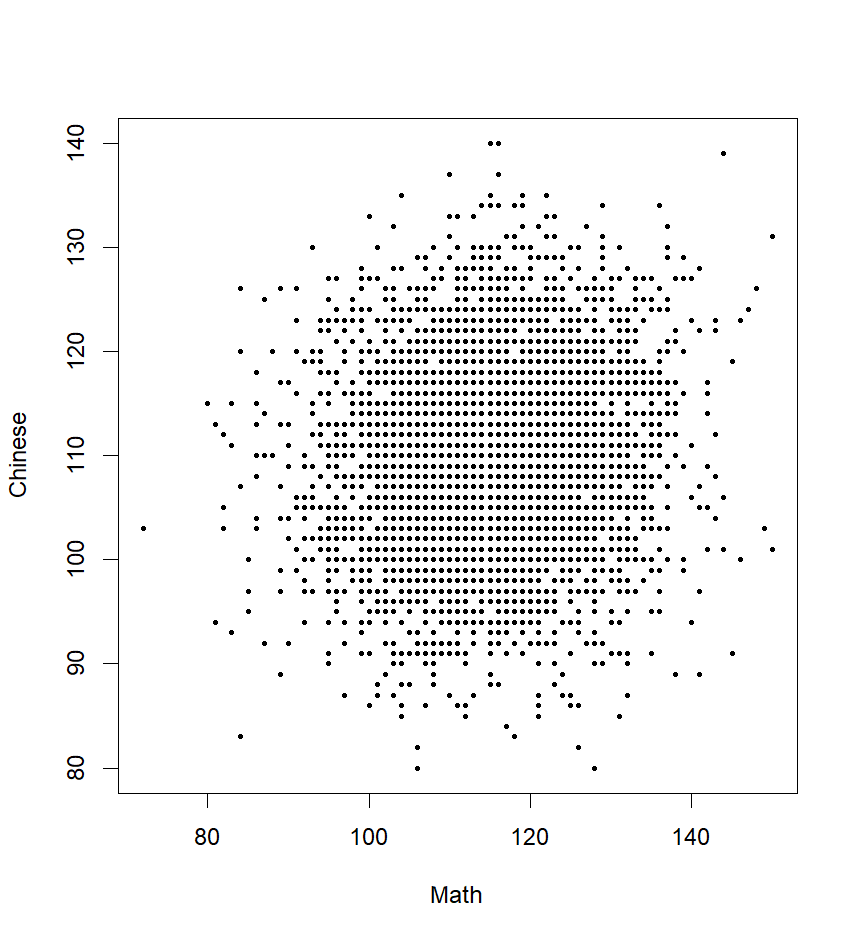
data\_frame <- data\_frame[order(data\_frame$数学),]

plot(data\_frame$数学, data\_frame$物理, type='p', xlab='Math', ylab='Physics', pch=16, cex=0.5)

首先，老样子获取表数据。然后，第三行，将数据按数学成绩由低到高排列；第四行，以数学成绩为x轴，以物理成绩为y轴，画图。pch=16表示图形类型是点，cex表示点的大小是0.5，type=’p’表示只画点（’l’只画线，’b’点线都画），pch=16。

然后将物理成绩替换为语文，再次执行：

plot(data\_frame$数学, data\_frame$语文, type='p', xlab='Math', ylab='Chinese', pch=16, cex=0.5)

可以清晰地看到，第一幅图中的图像比较扁，且向右上方倾斜，而第二幅图中的图像更圆。这初步告诉我们，相较于数学成绩与语文成绩的关系，数学成绩与物理成绩更呈现出正相关。

但是，这目前还只是初步的猜想。下面，我们求数学与语文、数学与物理的协方差和相关系数，从而较定量地判断上述关系。

print(cov(data\_frame$数学, data\_frame$物理))

print(cov(data\_frame$数学, data\_frame$语文))

cov()方法接收两个等长向量，返回其协方差。协方差代表的是两个变量的相关性。协方差为正，则两个变量同向变化，反之则反向变化。且协方差绝对值越大，则该种变化的同向或反向的程度就越高。

执行上述代码，分别返回40.11376和11.23481。据此，可以说明数学成绩越好，物理成绩也越好，且语文成绩也越好。那么是否可以说明相较于语文，物理成绩与数学成绩的正向相关性更高呢？目前还不能，因为物理成绩和语文成绩和标准差不同。

为了定量比较数学与物理、数学与语文成绩的相关性强弱，应计算两者的相关系数。执行代码：

print(cor(data\_frame$数学, data\_frame$物理))

print(cor(data\_frame$数学, data\_frame$语文))

cor()方法同样接收两个等长向量，返回的是其相关系数。相关系数等于协方差除以两个向量的标准差之积（因此上述第一行代码可以改成：

cov\_math\_physics <- cov(data\_frame$数学, data\_frame$物理)

sd\_数学 <- sd(data\_frame$数学)

sd\_物理 <- sd(data\_frame$物理)

cor\_math\_physics <- cov\_math\_physics / (sd\_数学 \* sd\_物理)

print( cor\_math\_physics )

），其值在-1和1之间，为正，则两个向量正相关，反之则负相关，为0，则不相关（相互独立），其绝对值为1，则说明两个向量呈线性关系。

上述代码得到的结果分别是0.4624419和0.1263927。这说明，在本次十校联考中，数学成绩好的学生，物理和语文成绩均偏向于更好，但物理与数学成绩的正相关性更强。