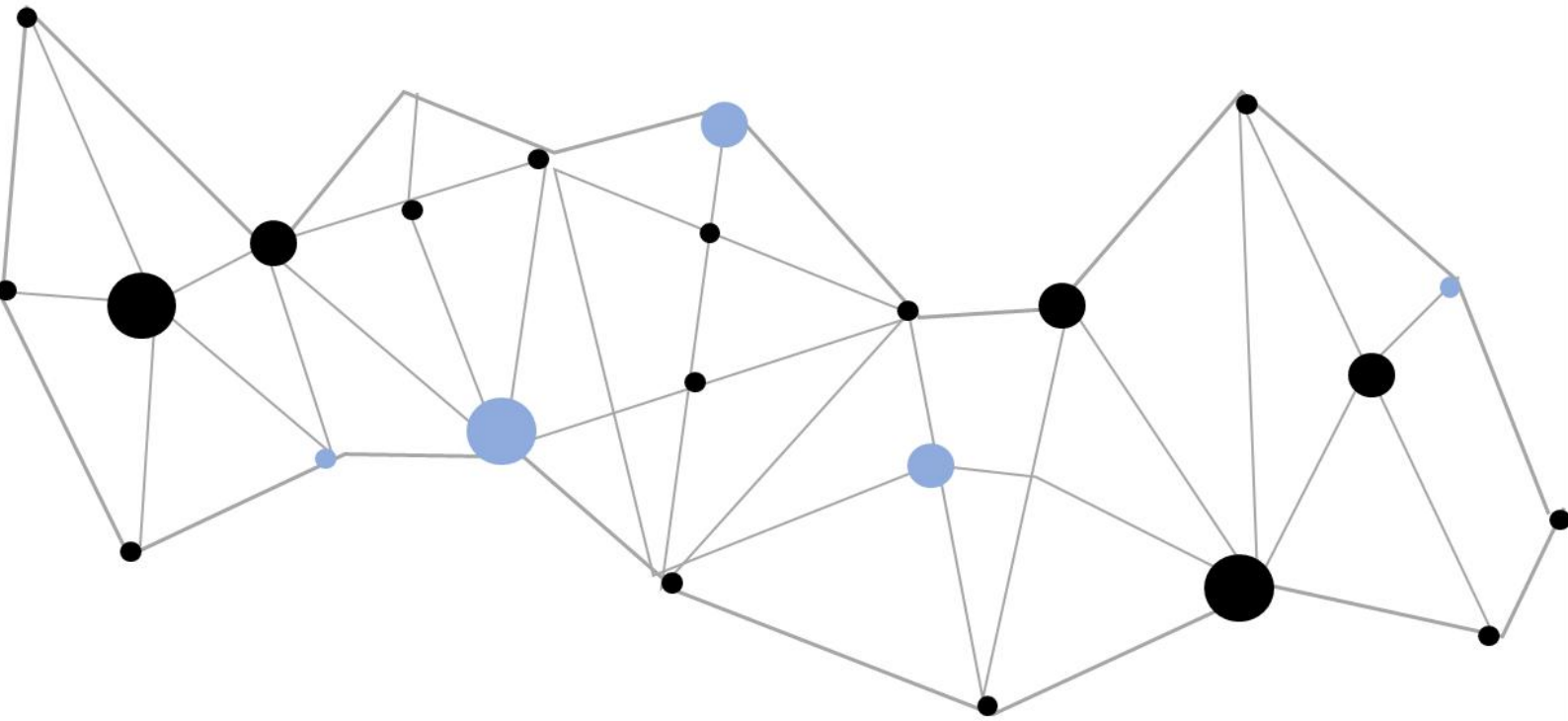
非参数统计

数 据 分 析 报 告

DATA ANALYSIS REPORT

吕明倩

19020182203614



目录

（一）每日体温监测---------- 1
1、数据集描述与简要分析---------- 1
2、 问题提出与解答---------- 2
问题一：上午和下午的体温数据是否具有相关性？---------- 2
问题二：上午和下午的体温数据分布是否相同？---------- 3
问题三：体温数据是否是随机产生的？是否存在某种趋势？---------- 5
（二）考试成绩数据分析---------- 7
问题一：各题目之间的难度有无差异？---------- 9
问题二：不同专业的同学成绩是否有差异？---------- 10
问题三：不同专业学生成绩的差异是由哪些题目造成的？---------- 12
问题四：不同专业的学生及格率是否相同？学生及格与学生专业是否具有独立性？------- 13
（三）游戏集数据分析---------- 14
问题一：探究不同手机游戏量化指标间的相互影响性---------- 17
问题二：哪一类型的游戏更受欢迎？---------- 17
问题三：哪种类型的游戏内部差异化比较明显？---------- 20
（四）马尔代夫数据分析---------- 22
问题一：各种行程之间的价格有无显著差异？---------- 24
问题二：不同店铺之间的路线报价是否存在显著差异？---------- 25
问题三：各月份前往马尔代夫旅行的价格有无显著差异？---------- 26
问题四：住宿是否含有早餐、是否是豪华住宿对旅行价格的高低有没有显著影响？------- 27
附录：R 语言处理代码---------- 28
（一）体温数据---------- 28
（二）成绩数据---------- 34
（三）游戏数据---------- 44
（四）马尔代夫数据---------- 51

（一）每日体温监测

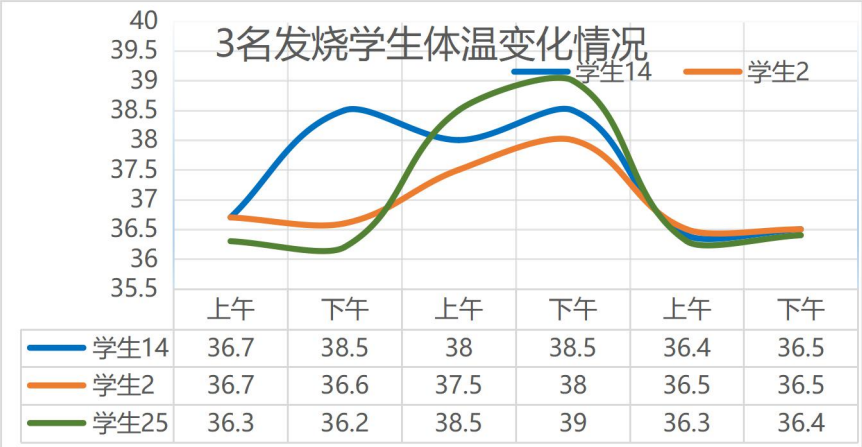
1、数据集描述与简要分析

本数据集为某班级 46 位学生的自测体温，记录了每位同学自 8 月 5 日至 10 月 31 日上午为止，每日上午和下午两次的体温数据。

由于并非所有学生都能坚持每日两次上报体温，我们得到的数据集存在部分缺失数据，但幸运地是，8 月 5 日至 9 月 16 日所有学生均完成了每日两次的体温上报，数据完整，并且有 8 位学生在本数据集截取时间段内不存在数据缺失，其学生序号分别为 4、11、20、25、28、42、44、46，其余 38 位学生的体温数据在 9 月或 10 月均存在不同程度的缺失。

此外，该数据集也存在部分不合法的数据形式，例如“36'4”、“36,3”、“3.65”、“365”、“46.4”等等，我们对数据进行合理化修正，并在承认所有学生填报数据具有真实性的前提下进行分析。

当一名学生的体温超过 37.3℃时，我们认为他发烧了。不难观测到 8 月至 10 月期间，共有 3 名同学（学生 14、学生 2、学生 25）出现发烧现象，时间分别为 8 月 8 日、8 月 27 日和 9 月 5 日，根据 3 名同学的体温变化情况（图 1-1），可以发现 3 名同学均在两天之内退烧，体温恢复到正常范围内。



	上午	下午	上午	下午	上午	下午
学生14	36.7	38.5	38	38.5	36.4	36.5
学生2	36.7	36.6	37.5	38	36.5	36.5
学生25	36.3	36.2	38.5	39	36.3	36.4

（图 1-1）

查阅资料可知：人体的正常体温范围为 36℃~37℃，因此对于体温低于 36℃的情况，我们认为此人此时体温偏低，统计发现每月体温不高于 36℃的情况如右图所示（表 1-1），总数为 242 人次。对于 8 月和 10 月，上午出现体温偏低的情况明显多于下午，我们猜测可能存在上午体温低于下午体温的现象。

人次	上午	下午
8月	60	40
9月	26	25
10月	60	31

（表 1-1）

此外，经过数据观察，我们发现对于学生 9，他出现体温偏低的情况明显多于其他同学，3 个月内共出现 59 次体温不高于 36℃的情况。同时我们观察到学生 9 的体温从未超过 36.5℃，体温均值约为 36.1℃，小于所有学生的体温均值 36.3℃，所以我们猜测学生 9 可能由于体质原因，体温数据普遍较低，从而猜测不同学生间的体温数据是存在差异的。

2、问题提出与解答

问题一：上午和下午的体温数据是否具有相关性？

我们猜测对于每个人，其上午和下午的体温可能存在一定的关系，所以我们首先将 46 位学生 3 个月的体温数据视为一个整体，由于数据量较大，我们可以剔除上午或下午存在缺失数据的学生其当天的体温数据，得到 46 位学生的 3893 对混合体温数据样本，其中每对数据分别包含该学生该日上午和下午的体温，采用 Spearman 秩相关检验和 Kendall τ 相关检验方法，分别对混合总体体温数据和按月分组的体温数据进行假设检验。

上述检验的 p 值均小于 0.05，从而，在 95%的置信水平下，我们有理由认为上午和下午的体温数据是存在相关性的，并且根据相关系数的计算（表 1-2），我们发现上午和下午体温数据存在弱正相关关系。

相关系数	8月	9月	10月	总体
Spearman	0.47	0.49	0.44	0.47
Kendall τ	0.39	0.40	0.36	0.38

（表 1-2）

同时我们可以考虑利用回归的方法说明上午和下午体温的关联程度。

①最小二乘回归（LS）

由于体温数据满足高斯-马尔可夫假设，我们可以采用最小二乘估计的方法得到回归方程：$y = 0.49x + 18.42$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.41663    0.50251   36.65   <2e-16 ***
x            0.49404    0.01382   35.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1571 on 3891 degrees of freedom
Multiple R-squared:  0.2473,    Adjusted R-squared:  0.2471
F-statistic:  1278 on 1 and 3891 DF,  p-value: < 2.2e-16
```

注意到该模型自变量的回归系数是显著的，但 R 方值仅有 0.247，效果并不好，这意味着关于下午体温的回归模型中应该还包含其他自变量。由于本数据集仅含有体温数据，所以这里我们没有引入新变量的数据支撑。

②稳健回归

由于最小二乘回归对于正态总体是更有价值的，我们在不知道总体正态性是否满足的前提下，可以选用一些稳健回归方法建立回归方程。采用最小中位数二乘回归（LMS）、最小截尾二乘回归（LTS）以及 S 估计回归，以上午体温数据为自变量、下午体温数据为因变量，建立回归方程（表 1-3）。

	LMS回归	LTS回归	S-估计回归
回归方程	y=0.5x+18.18	y=0.6x+14.56	y=0.6x+14.58

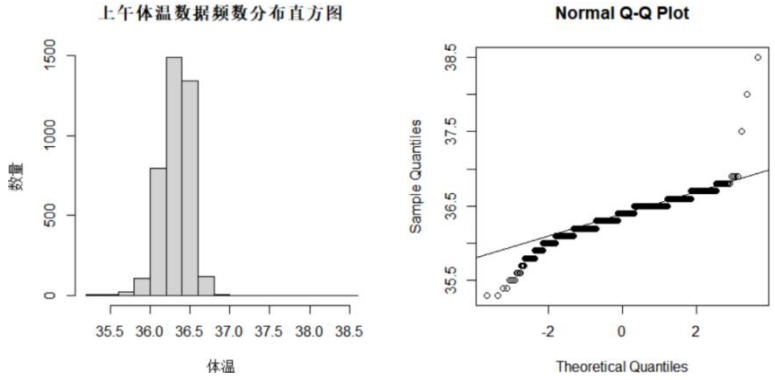
（表 1-3）

根据回归方程中自变量的回归系数，我们也不难发现上午和下午的体温数据之间存在弱正相关性。

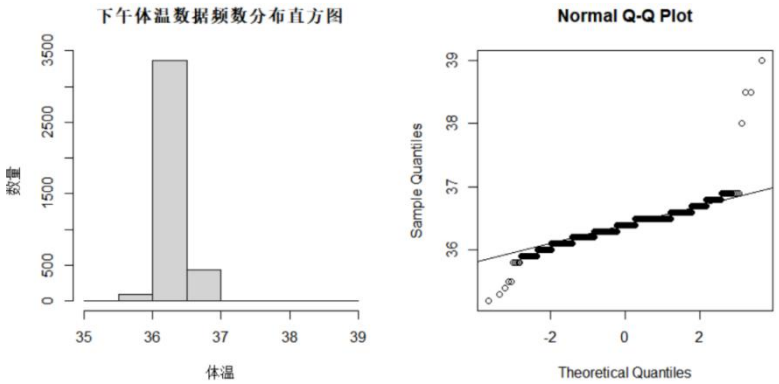
问题二：上午和下午的体温数据分布是否相同？

首先我们需要对数据的正态性作出检验。

综合采用直方图、QQ 图（图 1-2 & 图 1-3）以及 Anderson-Darling 检验、Cramer-von Mises 检验、Pearson 卡方检验、Shapiro-Francia 检验、Shapiro-Wilk 检验等多种方法进行正态检验（表 1-4）。



（图 1-2）



（图 1-3）

	检验方法	检验统计量	p值
上午	Anderson-Darling检验	A=75.547	<2.2e-16
	Cramer-von Mises检验	W=13.418	<7.37e-10
	Pearson卡方检验	P=34051	<2.2e-16
	Shapiro-Francia检验	W=0.91646	<2.2e-16
	Shapiro-Wilk检验	W=0.91776	<2.2e-16
下午	Anderson-Darling检验	A=81.514	<2.2e-16
	Cramer-von Mises检验	W=14.662	<7.37e-10
	Pearson卡方检验	P=35928	<2.2e-16
	Shapiro-Francia检验	W=0.87507	<2.2e-16
	Shapiro-Wilk检验	W=0.87677	<2.2e-16

（表 1-4）

可以看到，上述假设检验结果均拒绝正态性假设。所以我们采用两样本 Kolmogorov-Smirnov 检验（K-S 检验）来判断上午和下午的体温数据是否同分布。对于非正态总体，K-S 检验效果比较好。

给出零假设：上午和下午的体温数据具有相同的分布函数。该检验结果的 P 值为 0.01398 < 0.05，所以我们拒绝零假设，即认为上午和下午的体温分布函数存在差异。

从上述分析中我们可以看出，虽然上午和下午的体温数据并未通过 K-S 检验，但其直方图具有相似的形状，所以我们转而考虑相对更弱的条件，分析上午和下午的体温数据样本是否具有相同的位置参数和尺度参数。

位置参数检验：

①普通两样本中位数检验

首先我们仍然利用上述 46 位学生的 3893 对混合体温数据样本，由于体温数据不满足正态性假设，所以我们采用非参数方法分析其位置参数大小关系。根据经验以及数据集分析提出零假设和备择假设：

$$H_0: M_{上午} = M_{下午} \quad \text{v.s.} \quad H_1: M_{上午} < M_{下午}$$

分别采用 Brown-Mood 中位数检验、Wilcoxon (Mann-Whitney)秩和检验和正态记分检验三种方法完成位置参数检验（表 1-5）。

检验方法	Brown-Mood	Wilcoxon	正态记分
p值	0.1001	0.001352	1.29E-66
结果	接受	拒绝	拒绝

（表 1-5）

可以发现，在 95%置信水平之下，Wilcoxon (Mann-Whitney)秩和检验和正态记分检验可以拒绝原假设，认为 $M_{上午} < M_{下午}$，即我们认为下午体温普遍高于上午体温。值得注意的是上述 Brown-Mood 中位数检验并未得出相同的结论，这是由于相对于 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验而言，Brown-Mood 中位数检验只利用了符号信息，并未考虑观测值之间距离大小，所以结果的可靠性相对较低。

②成对数据检验

考虑到上午和下午的体温数据是连续的成对出现的变量，且每一对数据都是来自同一个学生，假设对于每位学生，其每天的体温观测是独立的，因此上述数据满足成对数据的定义，我们可以采用成对数据的检验来解决上述问题。

假设 D 为上午和下午的体温数据之差，给出零假设和备择假设如下：$H_0: M_D = 0$ v.s. $H_1: M_D < 0$。该检验 p 值为 2.42e-06 < 0.05，从而我们也可以得到下午体温普遍高于上午体温这一结果。

③完全区组设计

进一步考虑到不同学生的体质存在差异，所以上述检验的所有样本之间并非完全独立的，同一学生的体温数据具有相关性，但是学生之间的体质差异不是我们所关心的问题，为了消除不同学生体质差异这个因素对上午和下午体温数据差异的影响，我们考虑采用完全区组设计。

将体温数据按学生序号分类，得到 46 个区组（b=46），对于每个区组，计算每位学生三个月内上午和下午的体温均值，将其作为该区组上午体温和下午体温两个水平（k=2）的取值，得到 2×46 列联表（表 1-6）：

学生	1	2	3	4	5	6	7	8	9	10
上午体温	36.48588	36.39535	36.38816	36.51724	36.63721	36.3092	36.36437	36.3	36.12791	36.33953
下午体温	36.46588	36.38721	36.36711	36.45402	36.6186	36.34884	36.36163	36.3	36.14353	36.34535
学生	11	12	13	14	15	16	17	18	19	20
上午体温	36.47356	36.31205	36.35432	36.41047	36.30116	36.16353	36.15581	36.46747	36.3907	36.34138
下午体温	36.47241	36.38916	36.34875	36.45176	36.40235	36.20353	36.29302	36.49157	36.46163	36.33908
学生	21	22	23	24	25	26	27	28	29	30
上午体温	36.51395	36.3314	36.32209	36.32317	36.34598	36.32791	36.18372	36.37701	36.25233	36.54118
下午体温	36.51744	36.36047	36.38372	36.36707	36.37816	36.33023	36.1907	36.37701	36.21977	36.46471
学生	31	32	33	34	35	36	37	38	39	40
上午体温	36.27882	36.19651	36.37176	36.52791	36.45287	36.49765	36.44138	36.41512	36.2407	36.37857
下午体温	36.32353	36.16279	36.37647	36.5407	36.42093	36.47882	36.46491	36.37209	36.28235	36.39643
学生	41	42	43	44	45	46				
上午体温	36.35765	36.31724	36.45172	36.42299	36.36386	36.47126				
下午体温	36.35059	36.2954	36.4593	36.44253	36.3747	36.70345				

（表 1-6）

运用 Firedman 秩和检验和 Kendall 协同系数检验，这里零假设和备择假设与普通两样本中位数检验一致，我们得到上述检验的 p 值分别为 0.0002681 和 0.0003170306，因此在 95%置信水平下可以接受下午体温高于上午体温的备择假设。

综上，我们通过多种检验方法均证明了该班级学生的下午体温普遍高于上午体温，这与我们的常识性认知以及数据集描述的猜测相符合。

尺度参数检验：

为了分析上午和下午体温波动性的差异，我们需要对上午和下午体温数据进行尺度参数的假设检验。

$$H_0:\ \sigma_{上午}=\sigma_{下午}\quad \text{v.s.}\quad H_1:\ \sigma_{上午}<\sigma_{下午}$$

采用 Siegel-Tukey 检验、Mood 检验、Fligner-Killeen 检验和平方秩检验四种尺度检验方法，得到如下结果（表 1-7）。

检验方法	Siegel-Tukey	Mood	Fligner-Killeen	平方秩
p值	0.01534	0.07396	0.001748	8.05E-09
结果	拒绝	接受	拒绝	拒绝

（表 1-7）

不难发现，在 95%置信水平下，除 Mood 检验外，其余三种尺度检验均拒绝原假设，即认为下午的体温波动性显著大于上午的体温波动性。此外由于 Mood 检验的 p 值为 0.07396，所以当我们适当降低置信水平时，在 92%的置信水平下也可以得到下午体温波动性更大的结果。

问题三：体温数据是否是随机产生的？是否存在某种趋势？

由于不同学生的体温数据变化是存在差异的，所以我们不再将所有的体温数据视为一个整体，转而单独考虑某位同学的体温数据变化情况。同时考虑到时间的连续性，我们选择 3 个月内体温数据完整的 8 位同学（学生序号为 4、11、20、25、28、42、44、46）作为研究对象。

首先，我们采用游程检验探究 8 位同学的体温数据是否是随机产生的。注意到每位同学的体温数据均存在不同程度地打结现象，并且存在体温数据与样本中位数相等的情况，为了

将体温数据转化为二分数据，我们对 8 位同学的体温数据分别进行两次游程检验：

游程检验 1：样本数据大于等于样本中位数记为 1，样本数据小于样本中位数记为-1；
游程检验 2：样本数据大于样本中位数记为 1，样本数据小于等于样本中位数记为-1。

上述游程检验的零假设为数据序列是随机的，我们得到 8 位同学的 16 次游程检验结果。此外，对于体温数据不随机的学生，我们希望知道其体温变化趋势，所以我们继续进行 Cox-Stuart 趋势检验，该检验的零假设为数据序列无（增长/减少）趋势。

上述两项检验结果汇总如下：（表 1-8）

学生	游程检验	p值	结论	趋势检验p值	趋势	分析
4	1	0.7055	接受	0.4478416	无	体温数据是随机的且无趋势
	2	0.1464	接受			
11	1	6.44E-07	拒绝	0.04007166	增长	体温数据全部由36.4℃和36.5℃构成，体温波动极小，存在增长趋势
	2	无	无			
20	1	5.21E-15	拒绝	1.51E-09	减少	8月体温大部分不小于中位数，10月体温大部分不大于中位数，有明显减少趋势
	2	1.66E-08	拒绝			
25	1	0.01899	拒绝	0.5504618	无	体温数据不是随机的但无趋势，规律不明显
	2	5.86E-06	拒绝			
28	1	0.02073	拒绝	5.07E-08	增长	游程检验结果不一致，存在增长趋势
	2	0.2892	接受			
42	1	0.3087	接受	0.1409895	无	游程检验结果不一致，无趋势
	2	0.03622	拒绝			
44	1	0.001965	拒绝	0.07401609	增长	8月体温大部分不大于中位数，10月体温大部分不小于中位数，有明显增长趋势
	2	5.57E-08	拒绝			
46	1	7.93E-10	拒绝	0.1840938	无	体温数据上午大部分低于中位数，下午大部分高于中位数，无趋势
	2	1.19E-14	拒绝			

（表 1-8）

我们发现只有学生 4 的体温数据可以完全通过随机游程检验，其余同学的体温数据均存在不同类型的规律性。

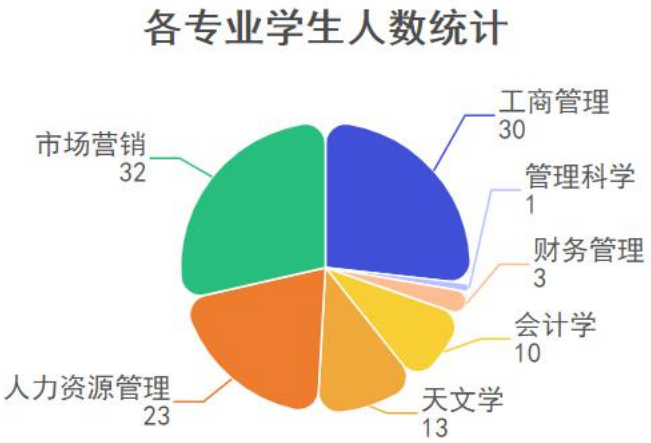
由于学生填报的体温数据可能并非真实测量所得，我们考虑基于游程检验和 Cox-Stuart 趋势检验结果，对 8 位学生填报的体温数据真实性做出合理怀疑。由于人体体温理应存在波动性（学生 11 数据波动性过小），但随着日期的增加不应该存在趋势（学生 20 有减少趋势、学生 44 有增长趋势），此外前述检验已经证明上午体温普遍低于下午体温（学生 46 的体温变化是合理的），所以我们怀疑学生 11、学生 20 和学生 44 的体温数据可能并非真实测量所得。

（二）考试成绩数据分析

1、数据集描述与简要分析

本数据集为某班级同学概率统计期中考试成绩，其中包含每位同学的专业名称、总成绩以及每道问题的得分情况。

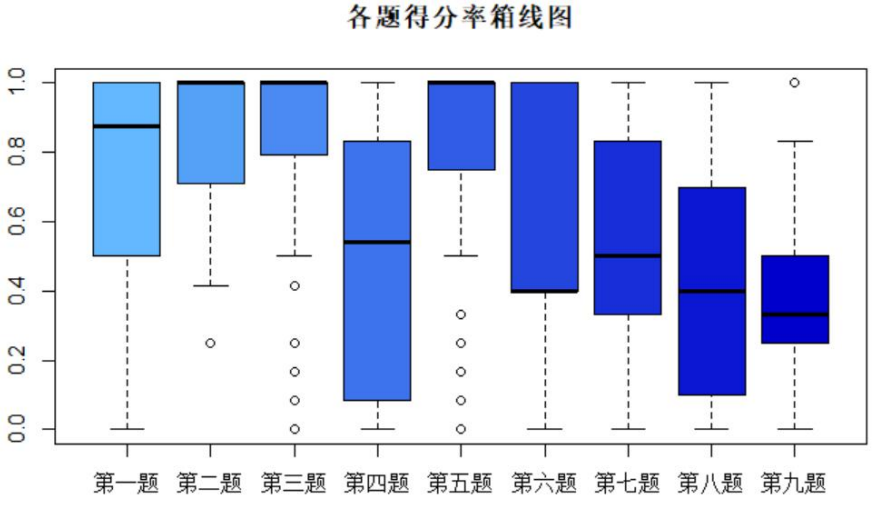
该班级由来自工商管理（30）、市场营销（32）、人力资源管理（23）、天文学（13）、会计学（10）、财务管理（3）和管理科学（1）7 个专业的 112 位同学组成（图 2-1），各专业同学的成绩均值和中位数如（表 2-1）所示，我们猜测不同专业的同学之间成绩是存在差异的。



专业	均值	中位数
工商管理	51.1	50.5
市场营销	69.44	73
人力资源管理	70.17	74
天文学	61.31	67
会计学	56.7	59.5
财务管理	67	77
管理科学	70	70

（图 2-1） **（表 2-1）**

每位同学的成绩是由 9 道题目的得分构成，由于每道题目的总分存在差异，我们考虑采用得分率来评价每道题目的作答情况，根据 9 道题得分率的箱线图（图 2-2），我们猜测 9 道题目得分率的中位数和跨度均存在差异，观察到第 2、3、5 题的得分率平均水平显著高于其他题目，猜测这三题难度不大，第 6、8、9 题的得分率平均水平较低，猜测这三题略有难度，第 4 题的得分率跨度最大，推测此题是使得学生成绩拉开差距的主要题目，经过简要分析推测 9 道题目难易程度应该存在差异，考虑采用假设检验方法验证。



（图 2-2）

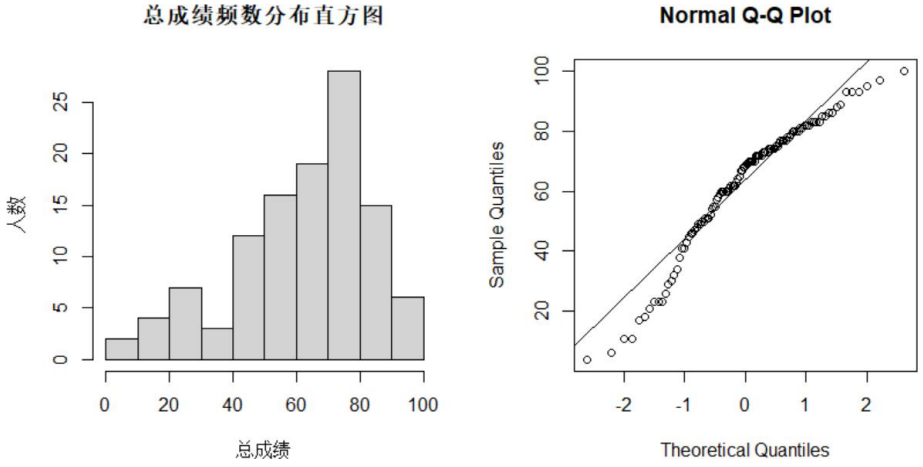
此外，根据成绩构成，假设每道问题均独立作答，我们可以将每道题目的得分视为一个变量，则总成绩受到九个变量的共同影响，由于变量数目较多，我们可以考虑采用主成分分析（PCA）方法进行降维。

```
> summary(pca,loadings=T)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.023237 1.057129 0.89449084 0.81837542 0.78574321
Proportion of Variance 0.454832 0.124169 0.08890154 0.07441537 0.06859916
Cumulative Proportion  0.454832 0.579001 0.66790257 0.74231794 0.81091710
                          Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     0.68759012 0.66689313 0.65418983 0.59687113
Proportion of Variance 0.05253113 0.04941627 0.04755159 0.03958391
Cumulative Proportion  0.86344823 0.91286450 0.96041609 1.00000000

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
第1题  0.212  0.652  0.330  0.306  0.510  0.107  0.226
第2题  0.361                -0.595        -0.380  0.516  0.234 -0.214
第3题  0.374  0.183 -0.259 -0.202 -0.104  0.589 -0.136 -0.304 -0.500
第4题  0.368 -0.135 -0.316 -0.250  0.431  0.166 -0.305  0.130  0.601
第5题  0.316  0.298  0.448 -0.205 -0.612        -0.161         0.401
第6题  0.302  0.295 -0.549  0.424 -0.277 -0.386 -0.172  0.283
第7题  0.367 -0.274         0.317        -0.193  0.352 -0.700  0.180
第8题  0.323 -0.436  0.179  0.349 -0.183  0.422  0.290  0.505
第9题  0.345 -0.287  0.428         0.232 -0.325 -0.554        -0.379
```

根据主成分分析结果，我们发现前 5 个主成分的累计贡献率可以达到 81.09%，所以可以选取前 5 个主成分作为新的独立变量，变化后的独立变量包含了原数据超过 80%的信息。

在进行假设检验之前，我们首先对所有同学的成绩数据进行正态性检验，我们以总成绩数据为例，综合采用直方图、QQ 图（图 2-3）以及 Anderson-Darling 检验、Cramer-von Mises 检验、Pearson 卡方检验、Shapiro-Francia 检验、Shapiro-Wilk 检验六种正态性检验方法（表 2-2），可以发现同学总成绩不服从正态分布，可以采用非参数方法对同学成绩进行评价。



（图 2-3）

检验方法	检验统计量	p值
Anderson-Darling检验	A=2.2958	7.48E-06
Cramer-von Mises检验	W=0.38243	3.91E-05
Pearson卡方检验	P=32	0.0007627
Shapiro-Francia检验	W=0.94053	0.0001936
Shapiro-Wilk检验	W=0.93786	5.63E-05

（表 2-2）

2、问题提出与解答

问题一：各题目之间的难度有无差异？

①位置参数检验

为了消除每题总分不同对得分的影响，我们仍然采用得分率作为评价题目作答情况的度量指标。我们假设每道题目的评分标准相同，并且每位学生每道问题的得分情况是独立的，进行多样本 Kruskal-Wallis 检验。

```
        Kruskal-Wallis rank sum test

data:  data_kw1 by group1
Kruskal-Wallis chi-squared = 245.67, df = 8, p-value < 2.2e-16
```

由于 p 值 < 0.05，因此，在 95%置信水平下，可以认为 9 道题目的得分率存在差异，从而验证了各题目之间难度存在差异性。

进一步研究不同题目得分率差异的相对大小，我们对 9 道题目中的任意两道分别进行 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验。

两种两样本位置参数检验的零假设条件均为：H_0：$M_1 = M_2$，为了表示简便，当备择假设为H_1：$M_1 < M_2$时，我们将假设检验的 p 值记录为负数，当备择假设为H_1：$M_1 > M_2$时，我们将假设检验的 p 值记录为正数。其中对于秩和检验，M_1和M_2分别代表横坐标和纵坐标对应的题目序号，而对于正态记分检验，M_1和M_2则分别代表纵坐标和横坐标对应的题目序号。上述假设检验 p 值汇总如下（表 2-3）：

	第一题	第二题	第三题	第四题	第五题	第六题	第七题	第八题	第九题	
第一题		-4.86E-05	-0.00035	2.58E-05	-0.00147	0.000935	0.000205	4.17E-09	2.76E-14	
第二题	-3.61E-15		0.416256	9.68E-16	0.354271	5.14E-12	4.91E-15	9.36E-21	1.09E-27	
第三题	-2.01E-13	-1.68E-08		7.47E-14	0.432566	6.06E-09	7.39E-12	3.52E-17	9.37E-21	
第四题	0.122907	9.19E-07	6.81E-05		-2.65E-12	-0.28176	-0.19646	0.064617	0.001198	秩和检验
第五题	-2.59E-12	-8.72E-08	-1.32E-09	-1.72E-17		2.48E-07	6.29E-10	1.39E-14	4.49E-17	
第六题	-0.22195	0.00956	0.180826	-0.00122	0.447689		-0.2577	0.002847	1.91E-05	
第七题	0.26008	1.04E-05	0.002712	-0.00129	0.021127	-0.00712		0.000732	1.33E-06	
第八题	0.001084	1.81E-11	4.83E-08	-0.31042	4.83E-06	-0.48793	0.045162		0.172909	
第九题	1.08E-07	0	8.95E-12	0.170744	5.22E-09	0.011909	0.002667	-0.41299		
正态记分检验										

（表 2-3）

可以看到，在 95%置信水平之下，绝大部分的题目之间的得分率是存在差异的，根据任意两道题目得分率的假设检验结果以及我们规定的记录方法，可以得到 9 道题目得分率平均水平的大小关系为

第 8 题&第 9 题 < 第 4 题&第 6 题&第 7 题 < 第 1 题 < 第 2 题&第 3 题&第 5 题

这与箱线图的直接观察结果相符合，第 2、3、5 题得分率较高，第 8、9 题得分率较低，所以我们有理由相信第 2、3、5 题比较容易，而第 8、9 题难度较大。

②尺度参数检验

进一步讨论每道题内部关于不同学生之间得分率的波动性是否存在差异，采用 Fligner-Killeen 检验和平方秩检验对每道题得分率进行多样本尺度参数检验。两种检验的 p 值分别为 1.557e-12 和 9.344347e-19，所以在 95%置信水平下，我们认为对于同一道题目，不同学生之间得分率的波动性存在差异。

继续进行每两道题之间的两样本平方秩检验，该检验的零假设条件均为：H_0：$\sigma_1^2 = \sigma_2^2$，当备择假设为 H_1：$\sigma_1^2 < \sigma_2^2$ 时，我们将假设检验的 p 值记录为负数，当备择假设为 H_1：$\sigma_1^2 > \sigma_2^2$ 时，我们将假设检验的 p 值记录为正数，其中 σ_1^2 和 σ_2^2 分别代表横坐标和纵坐标对应题目序号（表 2-4）。

两样本平方秩检验

	第一题	第二题	第三题	第四题	第五题	第六题	第七题	第八题	第九题
第一题		3.42E-15	9.45E-09	-0.0029	4.69E-05	-0.03206	-0.4383	-0.07269	0.000897
第二题			-1.44E-07	-9.69E-13	-1.34E-08	-5.78E-16	-1.74E-07	-3.62E-08	-0.14898
第三题				-1.16E-05	-1.28E-10	-0.06951	-0.0112	-0.00012	0.041298
第四题					0.011016	0.448366	0.000587	0.021205	9.87E-08
第五题						-0.34018	-0.12051	-0.01387	0.007274
第六题							1.69E-05	0.058827	6.96E-09
第七题								-0.11717	0.000889
第八题									4.86E-05
第九题									

（表 2-4）

其中我们可以发现一些比较明显的规律：除了第 9 题，第 2 题的得分率波动性小于其他 7 道题目；除了第 2 题，第 9 题的得分率波动性小于其他 7 道题目；除了第 6 题，第 4 题的得分率波动性明显大于其他 7 道题目。上述检验结果也与箱线图的直接观测结果一致。

问题二：不同专业的同学成绩是否有差异？

①位置参数检验

首先进行多样本 Kruskal-Wallis 检验，该检验 p 值为 0.01941 < 0.05，所以可以认为不同专业的同学成绩的平均水平存在差异。

```
        Kruskal-Wallis rank sum test

data:  data_kw2 by group2
Kruskal-Wallis chi-squared = 15.111, df = 6, p-value = 0.01941
```

其次，为了发现究竟是那些专业的学生成绩存在显著差异，进一步进行任意两个专业之间关于学生成绩的两样本 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验（表 2-5）。

上述假设检验的零假设条件均为：H_0：$M_1 = M_2$，当备择假设为 H_1：$M_1 < M_2$ 时，

我们将假设检验的 p 值记录为负数，当备择假设为 H_1：$M_1 > M_2$ 时，我们将假设检验的 p 值记录为正数，其中对于秩和检验，M_1 和 M_2 分别代表横坐标和纵坐标对应的专业名称，而对于正态记分检验，M_1 和 M_2 则分别代表纵坐标和横坐标对应的专业名称。

	工商管理	市场营销	人力资源管理	天文学	会计学	财务管理	管理科学	
工商管理		-0.002	-0.001	-0.100	-0.292	-0.136	-0.269	秩和检验
市场营销	-0.001		-0.449	0.053	0.080	0.488	0.396	
人力资源管理	-0.001	-0.492		0.010	0.073	-0.500	0.359	
天文学	-0.134	0.080	0.015		0.414	-0.209	-0.191	
会计学	-0.243	0.080	0.108	0.474		-0.249	-0.455	
财务管理	-0.116	0.420	-0.492	-0.169	-0.206		0.500	
管理科学	-0.294	0.403	0.432	-0.156	-0.396	0.342		
正态记分检验								

（表 2-5）

两种检验结果具有高度的一致性，我们发现大部分专业的同学成绩中位数不存在显著差异，仅有工商管理和市场营销、工商管理和人力资源管理、人力资源管理和天文学三组专业的同学成绩未能通过 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验。根据我们约定的记录方法，我们接受的备择假设分别为：工商管理同学成绩低于市场营销、工商管理同学成绩低于人力资源管理、人力资源管理同学成绩高于天文学。所以，我们有理由相信，工商管理和天文学专业的同学在概率统计这门课程的成绩相对不太理想。

②尺度参数检验

为了谈论同学成绩在不同专业内波动性的差异，我们首先采用 Fligner-Killeen 检验和平方秩检验对每道题得分率进行多样本尺度参数检验。

两种检验的 p 值分别为 0.04613 和 0.01245065，所以在 95%置信水平下，我们认为不同专业内同学成绩的波动性存在差异。

进一步进行任意两个专业之间的两样本平方秩检验，该检验的零假设条件均为：H_0：$\sigma_1^2 = \sigma_2^2$，当备择假设为 H_1：$\sigma_1^2 < \sigma_2^2$ 时，我们将假设检验的 p 值记录为负数，当备择假设为 H_1：$\sigma_1^2 > \sigma_2^2$ 时，我们将假设检验的 p 值记录为正数，其中 σ_1^2 和 σ_2^2 分别代表横坐标和纵坐标对应专业名称（表 2-6）。

	工商管理	市场营销	人力资源管理	天文学	会计学	财务管理	管理科学
工商管理		0.0721	0.0014	0.0135	-0.4847	0.3230	0.1281
市场营销			0.0200	0.0640	-0.0731	-0.2822	0.1283
人力资源管理				0.4710	-0.0035	-0.0225	0.1272
天文学	两样本平方秩检验结果				-0.0250	-0.0673	0.1250
会计学	人力资源管理 < 市场营销 & 财务管理					0.2296	0.1240
财务管理	人力资源管理 & 天文学 < 工商管理 & 会计学						0.1262
管理科学							

（表 2-6）

我们发现，对于同为人力资源管理专业的同学，他们之间的成绩差异比较小，对于天文学专业的同学也是这样。但对于同为市场营销专业的同学，他们之间的成绩差异比较大，对于财务管理专业、工商管理专业或会计学专业的同学也有相同的规律。

结合问题二中关于各专业中位数大小的检验结果，我们有理由怀疑工商管理专业的同学成绩平均水平较低，可能是由于存在少数分数极低的同学拉低了该专业的总体水平。

问题三：不同专业学生成绩的差异是由哪些题目造成的？

经过问题二的讨论，我们发现不同专业同学成绩平均水平存在如下差异：工商管理低于市场营销、工商管理低于人力资源管理、天文学低于人力资源管理。我们可以进一步发掘每组两个专业的成绩差异是由哪些题目引起的。

对于每组中的两个专业，关于 9 道题目得分分别进行 Brown-Mood 检验、Wilcoxon (Mann-Whitney)秩和检验和正态记分检验。上述检验的零假设条件均为：H_0：$M_1 = M_2$，当备择假设为H_1：$M_1 < M_2$时，我们将假设检验的 p 值记录为负数，当备择假设为H_1：$M_1 > M_2$时，我们将假设检验的 p 值记录为正数。假设检验结果汇总如下：（表 2-7）

专业	检验方式	第一题	第二题	第三题	第四题	第五题	第六题	第七题	第八题	第九题
工商管理&市场营销	Brown-Mood	-0.1526	NA	NA	-0.0201	NA	-0.0199	-0.1478	-0.3998	-0.0306
	Wilcoxon	-0.0224	-0.0002	-0.0429	-0.0369	-0.0182	-0.0023	-0.0948	-0.0340	-0.0110
	正态记分	-0.0001	-4.41E-08	-4.04E-06	-0.0034	-6.06E-07	-2.34E-06	-0.0079	-0.0040	-0.0009
工商管理&人力资源管理	Brown-Mood	-0.0209	NA	NA	-0.3189	NA	-0.0070	-0.0209	-0.2159	-0.2958
	Wilcoxon	-0.0157	-0.0004	-0.0266	-0.0667	-0.0568	-0.0003	-0.0381	-0.0091	-0.2137
	正态记分	-0.0001	-1.17E-07	-2.57E-06	-0.0085	-1.35E-05	-6.42E-07	-0.0089	-0.0012	-0.0227
天文学&人力资源管理	Brown-Mood	NA	NA	NA	0.0499	NA	-0.3296	-0.0864	-0.0537	-0.2611
	Wilcoxon	-0.4015	-0.1074	-0.1767	0.1117	-0.2510	-0.2128	-0.0142	-0.0019	-0.1278
	正态记分	-0.0207	-0.0001	-0.0004	0.4196	-0.0016	-0.0017	-0.0073	-0.0006	-0.0157

（表 2-7）

我们发现，由于 Brown-Mood 检验未考虑观测值距离中心的远近，所以当样本中位数达到单题最高分时，该检验失效，标记为 NA。此外，关于正态记分检验，除了天文学&人力资源管理这组中关于第四题的检验外，其余检验均拒绝原假设，认为两个专业在此题得分存在差异，因此该检验效果并不好。所以，我们根据 Wilcoxon (Mann-Whitney)秩和检验结果可以做出如下分析（表 2-8）。

专业	Wilcoxon检验结果
工商管理&市场营销	除第七题外，工商管理专业的同学在其余8道题目的得分均明显低于市场营销专业的同学
工商管理&人力资源管理	工商管理专业的同学在第一、二、三、六、七、八题得分明显低于人力资源管理的同学
天文学&人力资源管理	天文学专业的同学在第七、八题得分明显低于人力资源管理的同学

（表 2-8）

问题四：不同专业的学生及格率是否相同？学生及格与学生专业是否具有独立性？

及格率是我们比较感兴趣的一个指标，为了解决上述问题，我们首先将 7 个专业的同学及格与不及格人数汇总如下（表 2-9），采用列联表齐性和独立性的卡方检验，该检验 p 值为 0.02458，所以在 95%置信水平下，我们拒绝原假设(H_0：$p_{及格|专业i} = p_{及格|专业j}$)，认为不同专业学生的及格率不同，即学生的及格率在不同专业之间存在差异。进一步采用 Fisher 精确检验，对任意两个专业的及格率进行假设检验（表 2-10），发现工商管理与市场营销、人力资源管理、天文学三个专业的及格率均存在差异，人力资源和管理与会计学专业的及格率存在差异。

	及格	不及格	
工商管理	12	18	30
市场营销	24	8	32
人力资源管理	20	3	23
天文学	10	3	13
会计学	5	5	10
财务管理	2	1	3
管理科学	1	0	1
	74	38	112

（表 2-9）

Fisher检验	p值	两专业及格率大小
工商管理&市场营销	0.0095	不相等
工商管理&人力资源管理	0.0006	不相等
工商管理&天文学	0.0452	不相等
工商管理&会计学	0.7166	相等
市场营销&人力资源管理	0.3262	相等
市场营销&天文学	1.0000	相等
市场营销&会计学	0.2383	相等
人力资源管理&天文学	0.6454	相等
人力资源管理&会计学	0.0362	不相等
天文学&会计学	0.2213	相等

（表 2-10）

对以上四组两专业及格率差异进行比较分析，计算四组中两个及格率之差(p_hat)、相对风险(RR)、胜算比(OR)的点估计和区间估计（表 2-11）。根据估计结果可以发现，工商管理专业及格率低于市场营销、人力资源管理以及天文学，而人力资源管理专业及格率高于会计学，这与问题二中有关工商管理专业的假设检验结果相符合。

	及格率之差点估计	及格率之差区间估计	RR点估计	RR区间估计	OR点估计	OR区间估计
工商管理&市场营销	-0.35	(-0.58，-0.12)	0.53	(0.33，0.86)	0.22	(0.08，0.66)
工商管理&人力资源管理	-0.47	(-0.69，-0.25)	0.46	(0.29，0.73)	0.10	(0.02，0.41)
工商管理&天文学	-0.37	(-0.66，-0.08)	0.52	(0.31，0.88)	0.20	(0.05，0.88)
人力资源管理&会计学	0.37	(0.03，0.71)	1.74	(0.92，3.30)	6.67	(1.18，37.78)

（表 2-11）

（三）游戏集数据分析

1、数据集描述与简要分析

本数据集关于安卓手机游戏，包含 1140 个不同的手机游戏信息，对于每款手机游戏，我们记录了评分、类别、语言、热度、最后更新时间、游戏版本、资费、开发商、支持系统、评论数、喜欢数，共计 11 个数据项。

首先，将安卓手机游戏按游戏类型分类可以分为 15 个不同类别（图 3-1），其中休闲益智类型游戏最多，共有 292 个，其次动作游戏、角色扮演、射击游戏三个类型的游戏数目也均超过 100 个，游戏数目较少的类型有养成游戏、游戏工具、棋牌游戏，其游戏数目均不超过 20 个。

（图 3-1）

将安卓手机游戏按最后更新年月进行划分，可以得到每年安卓手机游戏更新数目变化情况（图 3-2），可以看到从 2010 年至 2016 年，安卓手机游戏每年更新总数目总体呈现出上升的趋势，特别是在 2016 年达到年更新 368 款游戏。在 2011 年至 2012 年，伴随着智能手机的兴起，大量新兴手机游戏涌入市场，在此期间安卓手机游戏的更新数目迅速增加，增幅均超过 200%（表 3-1），但在随后的 3 年时间内，安卓手机游戏的更新数目增幅放缓，甚至在 2014 年出现短暂的负增长，但这一现象并未持续很久，在 2016 年安卓手机游戏的更新数目增幅再次超过 64%，预计未来（2017 年以及之后）安卓手机游戏的更新数目仍会保持增长态势。

（图 3-2）

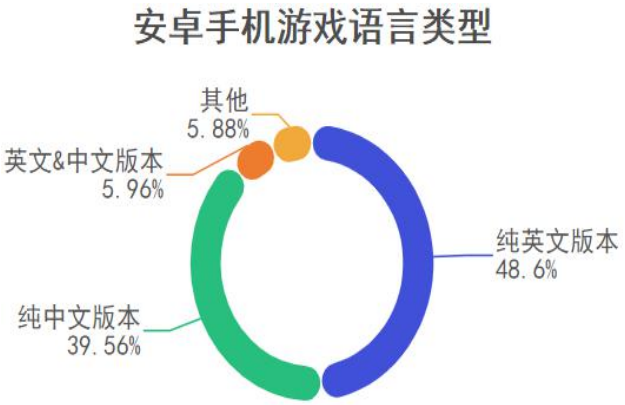
年份	增幅
2011	437.50%
2012	223.26%
2013	36.69%
2014	-11.58%
2015	33.33%
2016	64.29%

（表 3-1）

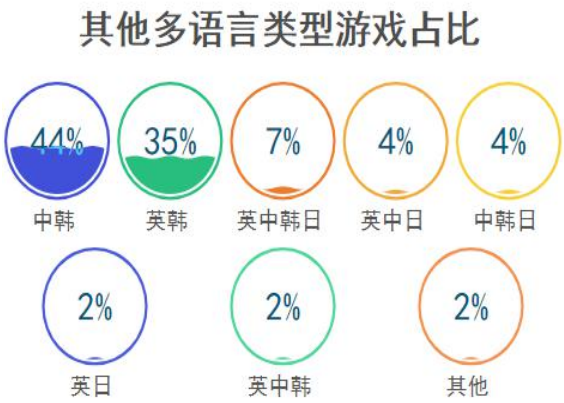
在 1140 个安卓手机游戏中，纯英文版本的游戏最多，共有 554 个，在所有游戏中占比高达 48.6%（图 3-3），数量次之的是纯中文版本的游戏，共有 451 个，占比达 39.56%，此

外还有 3 个纯韩语版本游戏和 10 个纯日语版本游戏未在图表中列出。

可以看出在安卓手机游戏市场中，英文和中文版本的游戏数目占主导地位，联合占比超过 94%，不难推测英文和中文版本的游戏受众范围更广，建议多开发英文和中文版本的游戏。



（图 3-3）

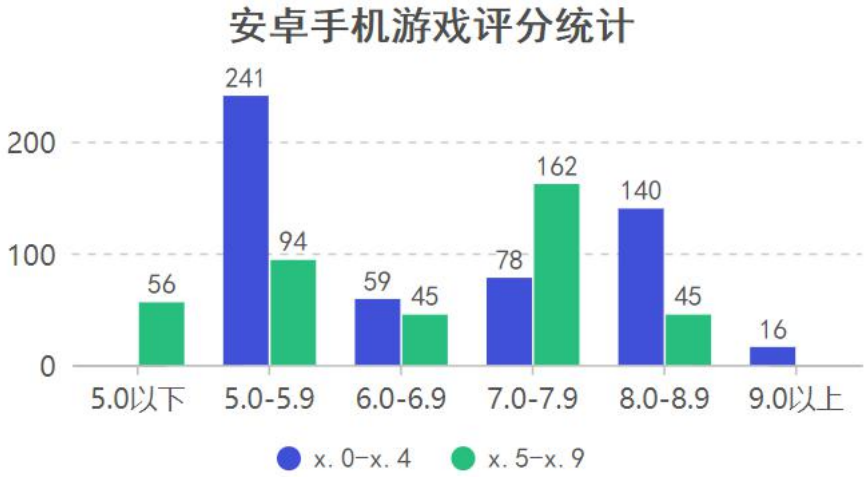


（图 3-4）

当然，在安卓手机游戏中存在多语言版本的游戏，大约有 10.9%的游戏有两种或两种以上语言版本。在多语言版本游戏中约有六成的游戏为兼有中英文版本，剩下的多语言版本游戏则是以中韩、英韩语言版本为主（图 3-4）。

关于安卓手机游戏的评分，处理 203 个缺失数据以及 2 个评分 1.0 的异常值后，安卓手机游戏的最高评分为 9.6，为 Rovio 开发的益智休闲游戏《愤怒的小鸟太空版无广告版》，最低评分为 4.5，为北京联众互动网络股份有限公司开发的棋牌游戏《单机斗地主》。

将手机游戏按评分区段分类，评分在 5.0 至 5.4 区间段的游戏最多，高达 241 个（图 3-5），评分在 7.5 至 7.9 以及 8.0 至 8.4 的游戏数目次之，分别为 162 个和 140 个，评分超过 9.0 的高分游戏仅有 16 个，说明安卓手机游戏市场仍具备很大的发展空间。



（图 3-5）

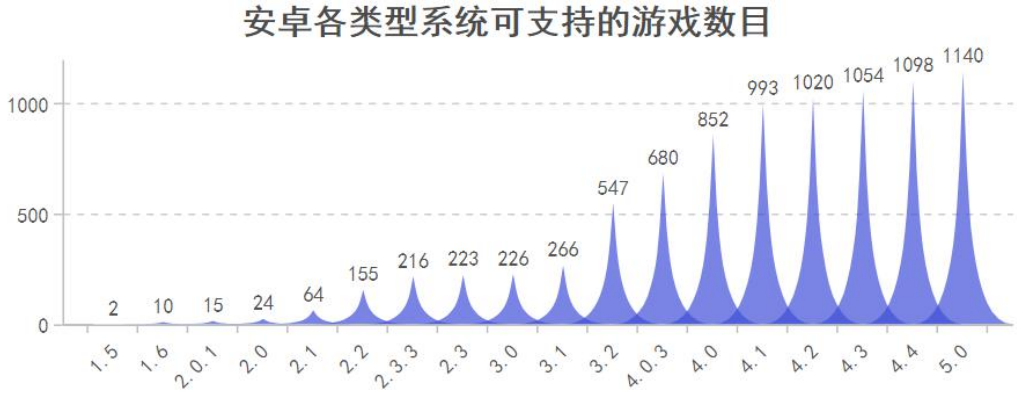
在 1140 个安卓手机游戏中，有 86 个游戏的开发商未知，除此之外，剩下的 1054 个游戏共涉及 584 家游戏开发商，其中在统计范围内仅开发一个游戏的开发商有 443 家，开发手机游戏数超过 10 款的开发商共有 9 家（表 3-2）。

开发商	Glu	Gameloft	EA	GAMEVIL	Com2us	SQEX	Rovio	SEGA世嘉	G5
数目	38	32	28	22	21	18	17	14	13

（表 3-2）

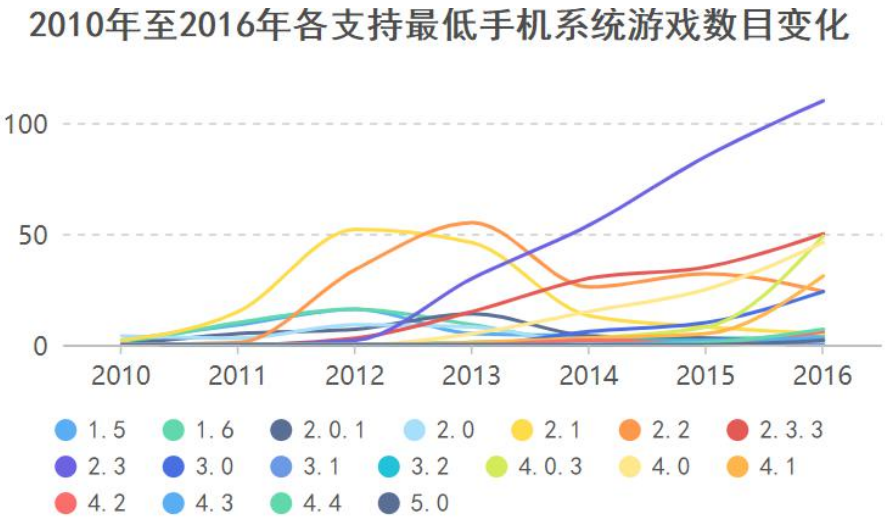
开发商 Glu 以开发 38 款手机游戏位居榜首，其主要以开发游戏类型为射击游戏（19 款），游戏的语言版本均为英文或中文版本，游戏评分均值为 6.8，累计游戏评论数超过 10 万，累计游戏喜欢数超过 2 万，是一家综合实力非常强的开发商。

由于不同游戏对手机系统的要求存在差异，算法设计复杂、占用内存更大的游戏对手机系统的要求更高，手机系统版本越高，可玩的游戏种类越多。因此，按手机系统分类，可以得到安卓各类型系统可支持的游戏数目如下（图 3-6），可以观察到，当手机系统升级到 3.2 版本时，安卓各系统可支持游戏数目开始快速增长，当手机系统升级到 4.1 时，可支持的游戏数目已经超过 87%，此时绝大多数游戏都可以正常运行。



（图 3-6）

此外，我们统计了各年份各版本游戏的研发情况（图 3-7），可以看出，在 2012 年前，安卓手机游戏支持的最低系统版本主要以 2.1 系统为主，2012 年至 2013 年实现了从 2.1 系统到 2.2 系统的转变，从 2014 年开始，安卓手机游戏支持的最低系统版本主要以 2.3 系统为主，并一直持续到 2016 年。



（图 3-7）

此外，关于该数据集还涉及到每个安卓手机游戏的热度、评论数、喜欢数等量化信息，可以作为度量不同手机游戏差异性的指标。

热度最高（95℃）的安卓手机游戏为触控科技开发的益智休闲类游戏《捕鱼达人》；最低热度为 28℃，包括《俄罗斯方块》等 74 个游戏。

评论数最高的游戏为动作游戏《GTA 侠盗猎车手:圣安地列斯》，共获得 96354 条评论，11410 个喜欢，在所有游戏中喜欢数排名第二。收获喜欢数最多的游戏为模拟经营类游戏《这是我的战争》，共获得 13323 个喜欢。

2、问题提出与解答

问题一：探究不同手机游戏量化指标间的相互影响性

一款游戏的热度、评分、评论数、喜欢数都可以作为评价手机游戏影响力和受欢迎程度的指标，但由于上述四种指标并不独立，可能存在相互影响，我们首先采用 Spearman 秩相关检验和 Kendall τ 相关检验，来评价一下上述四种指标的相关性以及协同性。

H_0：指标 X 与指标 Y 不相关（$\rho = 0$） v.s. H_1：指标 X 与指标 Y 不相关（$\rho \neq 0$）

上述假设检验的结果均为拒绝零假设，即认为上述四项指标之间是相关的，相关系数统计如下（表 3-3）：

相关系数	热度	评分	评论数	喜欢数	
热度	1.00	0.18	0.49	0.28	
评分	0.12	1.00	0.65	0.52	Spearman
评论数	0.36	0.47	1.00	0.69	秩和检验
喜欢数	0.19	0.38	0.52	1.00	
Kendall τ 检验					

（表 3-3）

我们观察到，上述四项指标之间均存在正相关关系，其中喜欢数与评论数的相关系数最大，Spearman 相关系数达到 0.69，Kendall 相关系数达到 0.52，具备较强的正相关关系，我们有理由相信，评论数较多的游戏，其喜欢数也较多。

其次评分与评论数之间、评分与喜欢数之间也存在比较强的正相关关系。所以我们认为，对于评分较高的游戏，其评论数和喜欢数普遍较多，一个游戏的评论数和喜欢数越多，其越可能获得比较高的评分。为了计算简便，后续分析我们以评分为例作为不同手机游戏间的度量指标。

此外，我们发现，游戏热度与其他三项指标之间的正相关系数相对较低，即游戏热度与其他三项指标之间的相关程度相对较低。因此在研究不同类型的游戏受欢迎程度的差异性时，除了评分之外，我们也将游戏热度这一指标纳入考量。

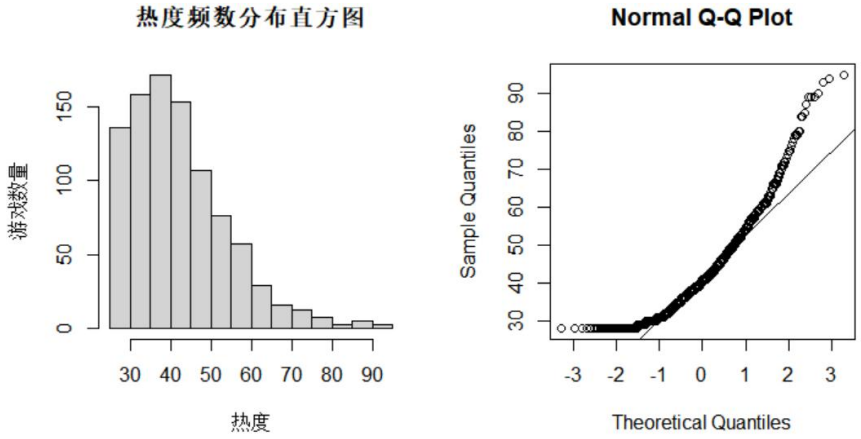
问题二：哪一类型的游戏更受欢迎？

考虑到一款游戏的热度和评分都可以直接反映出一款游戏的受欢迎程度，因此我们分别采用热度和评分作为度量指标，分析 15 类游戏的受欢迎程度。

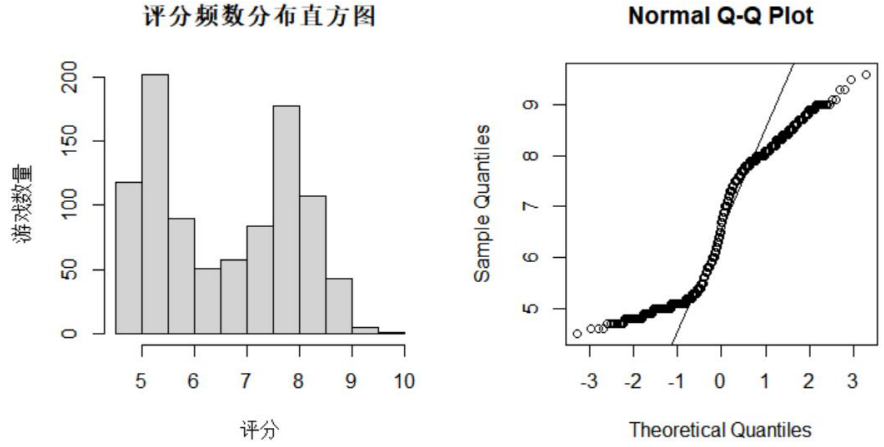
首先，分别对热度指标和评分指标进行正态性检验。通过直方图和 QQ 图（图 3-8 & 图 3-9）以及 Anderson-Darling 检验、Cramer-von Mises 检验、Pearson 卡方检验、Shapiro-Francia 检验、Shapiro-Wilk 检验等多种正态性检验方法（表 3-4），发现游戏热度和游戏评分均不服从正态分布，所以考虑采用非参数方法进行假设检验。

指标	检验方法	检验统计量	p值
热度	Anderson-Darling检验	A=17.933	<2.2e-16
	Cramer-von Mises检验	W=2.7242	<7.37e-10
	Pearson卡方检验	P=372.36	<2.2e-16
	Shapiro-Francia检验	W=0.90947	<2.2e-16
	Shapiro-Wilk检验	W=0.9091	<2.2e-16
评分	Anderson-Darling检验	A=35.41	<2.2e-16
	Cramer-von Mises检验	W=5.8954	<7.37e-10
	Pearson卡方检验	P=1065.7	<2.2e-16
	Shapiro-Francia检验	W=0.90504	<2.2e-16
	Shapiro-Wilk检验	W=0.90378	<2.2e-16

（表 3-4）



（图 3-8）



（图 3-9）

①游戏热度

由于不同游戏的热度评价是基于同一标准并且相互独立的，我们假定 15 类游戏热度具有相似的连续分布，并且所有的热度值在样本内和样本之间是独立的，形式上，假定 15 类游戏热度样本有分布函数 $F_i(x)=F(x-\theta_i)$，i=1,2,…,k，我们考虑如下假设检验问题：

$$H_0:\ \theta_1=\theta_2=\cdots=\theta_k \quad \text{v.s.} \quad H_1:\ \text{至少有一个等号不成立}$$

采用 Kruskal-Wallis 秩和检验，由于 p 值为 0.02766 < 0.05，因此，在 95%置信水平下，

可以认为 15 类游戏的热度水平存在差异。

```
        Kruskal-Wallis rank sum test

data:  data_kwl by group1
Kruskal-Wallis chi-squared = 25.773, df = 14, p-value = 0.02766
```

为了进一步研究不同类型游戏热度差异的相对大小，我们对 15 类游戏中的任意两类分别进行 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验。

上述两种假设检验的零假设条件均为：H_0：$M_1 = M_2$，当备择假设为H_1：$M_1 < M_2$时，我们将假设检验的 p 值记录为负数，当备择假设为H_1：$M_1 > M_2$时，我们将假设检验的 p 值记录为正数。其中对于秩和检验，M_1和M_2分别代表横坐标和纵坐标对应游戏类别，而对于正态记分检验，M_1和M_2则分别代表纵坐标和横坐标对应游戏类别，上述假设检验结果汇总如下（表 3-5）：

	策略塔防	动作游戏	飞行游戏	格斗游戏	角色扮演	竞速游戏	冒险解谜	模拟经营	棋牌游戏	射击游戏	体育运动	养成游戏	益智休闲	音乐游戏	游戏工具	
策略塔防		-0.258	-0.250	0.476	-0.160	-0.004	0.170	0.108	0.065	0.326	-0.095	0.483	-0.487	-0.449	-0.174	秩和检验
动作游戏	-0.093		-0.487	0.317	-0.359	-0.018	0.044	0.035	0.033	0.098	-0.193	0.356	0.226	0.380	-0.203	
飞行游戏	-0.145	-0.349		0.368	0.483	-0.034	0.077	0.057	0.039	0.172	-0.297	0.329	0.236	0.427	-0.191	
格斗游戏	-0.451	0.406	0.396		-0.231	-0.037	0.253	0.198	0.132	0.355	-0.152	0.500	-0.449	0.500	-0.153	
角色扮演	-0.070	-0.209	-0.403	-0.089		-0.036	0.021	0.014	0.013	0.067	-0.220	0.327	0.126	0.373	-0.351	
竞速游戏	-0.001	-0.009	-0.041	-0.014	-0.012		0.000	0.000	0.003	0.000	0.252	0.110	0.002	0.056	0.483	
冒险解谜	0.288	0.115	0.112	0.437	0.059	0.002		0.422	0.197	-0.248	-0.019	-0.318	-0.093	-0.217	-0.089	
模拟经营	0.185	0.082	0.067	0.302	0.021	0.002	-0.431		0.212	-0.216	-0.017	-0.285	-0.088	-0.170	-0.053	
棋牌游戏	0.103	0.067	0.063	0.237	0.024	0.007	0.340	0.376		-0.100	-0.016	-0.140	-0.052	-0.130	-0.054	
射击游戏	0.487	0.205	0.222	-0.441	0.141	0.003	-0.088	-0.078	-0.058		-0.052	-0.457	-0.257	-0.314	-0.067	
体育运动	-0.022	-0.073	-0.195	-0.055	-0.069	0.462	-0.002	-0.003	-0.007	-0.009		0.239	0.075	0.230	0.464	
养成游戏	-0.335	-0.499	0.471	-0.347	0.499	0.234	-0.192	-0.135	-0.074	-0.271	0.328		-0.497	-0.430	-0.322	
益智休闲	-0.203	-0.454	0.342	-0.201	0.366	0.023	-0.006	-0.012	-0.021	-0.041	0.118	0.496		-0.463	-0.175	
音乐游戏	-0.466	0.411	0.388	-0.420	0.367	0.082	-0.187	-0.107	-0.113	-0.238	0.220	-0.478	-0.436		-0.169	
游戏工具	-0.192	-0.222	-0.226	-0.149	-0.393	0.442	-0.069	-0.028	-0.040	-0.053	0.409	-0.393	-0.173	-0.144		
	正态记分检验															

（表 3-5）

（绿色：两种检验均拒绝原假设，蓝色：仅其中一种检验拒绝原假设）

可以看到，在 95%置信水平之下，除了体育运动、养成游戏、音乐游戏、游戏工具 4 类游戏之外，其他 10 类游戏与竞速游戏在游戏热度大小方面的假设检验，其结果均为拒绝原假设，均接受竞速游戏热度分别大于其他 10 类游戏热度的备择假设。我们发现竞速游戏的热度比绝大多数其他类型的手机游戏的热度要高，我们有理由相信竞速游戏更受欢迎，所以我们建议游戏公司可以多开发竞速游戏来获得更高的市场热度。

②游戏评分

此外，我们在承认游戏评分真实性的前提下，也可以将游戏评分作为度量游戏受欢迎程度的重要指标。在处理缺失数据后，我们利用 936 款游戏的评分数据，按照与分析游戏热度相似的方法，探究各类游戏间评分是否存在差异。

同样假设不同游戏的评分是基于同一标准的并且相互独立，满足连续性和独立性假设，我们假定 15 类游戏评分样本有分布函数$F_i(x) = F(x - \theta_i)$，i=1,2,...,k，考虑如下假设检验问题：

$$H_0:\ \theta_1=\theta_2=\cdots=\theta_k \quad \text{v.s.} \quad H_1:\ \text{至少有一个等号不成立}$$

采用 Kruskal-Wallis 秩和检验，由于 p 值为 0.00009247 < 0.05，因此，在 95%置信水平下，可以认为 15 类游戏的评分存在显著差异。

```
        Kruskal-Wallis rank sum test

data:  data_kw2 by group2
Kruskal-Wallis chi-squared = 42.792, df = 14, p-value = 9.247e-05
```

为了进一步研究不同类型游戏评分差异的相对大小，同样进行两样本 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验，其中备择假设条件以及 p 值的记录方式采用与探究不同类型游戏热度的相对大小时相同的方式，假设检验结果汇总如下（表 3-6）：

	策略塔防	动作游戏	飞行游戏	格斗游戏	角色扮演	竞速游戏	冒险解谜	模拟经营	棋牌游戏	射击游戏	体育运动	养成游戏	益智休闲	音乐游戏	游戏工具	
策略塔防		0.25852	0.00050	0.37220	-0.23855	0.01663	0.44427	0.21759	0.00715	0.09630	0.50125	0.30948	0.02862	0.18549	0.00636	秩和检验
动作游戏	0.41191		0.00070	-0.46260	-0.03982	0.04073	-0.20316	0.41163	0.01597	0.26846	-0.25795	-0.50000	0.08604	0.29383	0.00744	
飞行游戏	0.00242	0.00294		-0.00099	-0.00002	-0.02874	-0.00002	-0.00135	-0.43005	-0.00256	-0.00197	-0.00717	-0.00166	-0.01707	0.42426	
格斗游戏	-0.47866	-0.30545	-0.00043		-0.16346	0.08255	-0.37961	0.39840	0.01396	0.30074	-0.41546	0.38498	0.11422	0.19666	0.00291	
角色扮演	-0.16289	-0.01050	-0.00001	-0.14390		0.00116	0.19190	0.05774	0.00208	0.01288	0.29235	0.18719	0.00039	0.06953	0.00162	
竞速游戏	0.03350	0.11105	-0.01236	0.13479	0.00435		-0.01051	-0.09661	0.09944	-0.14309	-0.04157	-0.21463	-0.20496	-0.32205	0.06721	
冒险解谜	-0.35995	-0.05493	-0.00001	-0.27351	0.40351	-0.00226		0.21243	0.00265	0.09490	-0.49566	0.20269	0.00610	0.08890	0.00214	
模拟经营	0.24614	0.48497	-0.00085	0.44193	0.09222	-0.08932	0.19624		0.02445	0.42124	-0.26694	-0.48856	0.15669	0.25256	0.00948	
棋牌游戏	0.00584	0.01540	-0.44406	0.01373	0.00242	0.10770	0.00267	0.03657		-0.02594	-0.01459	-0.04462	-0.03751	-0.09128	0.51368	
射击游戏	0.09814	0.34460	-0.00097	0.27428	0.01805	-0.11511	0.06908	-0.40122	-0.01092		-0.12518	-0.44081	0.25260	0.37140	0.01947	
体育运动	-0.35962	-0.11647	-0.00060	-0.35284	0.45655	-0.00954	-0.46330	-0.13981	-0.00509	-0.01946		0.38472	0.08502	0.26486	0.01121	
养成游戏	0.41839	-0.38679	-0.00419	-0.49921	0.27204	-0.15716	0.25357	-0.36177	-0.02826	-0.26555	0.46559		0.28218	0.26629	0.03385	
益智休闲	0.11897	0.35606	-0.00030	0.21434	0.00549	-0.05349	0.02110	0.40559	-0.00792	-0.31702	0.18103	0.35818		-0.48770	0.02059	
音乐游戏	0.35736	-0.47745	-0.00656	0.33391	0.15633	-0.19095	0.19366	0.43025	-0.05051	-0.38163	0.44070	0.35205	-0.30469		0.04012	
游戏工具	0.02627	0.02487	-0.39933	0.00550	0.00566	0.13580	0.00507	0.01998	-0.40516	0.06355	0.03371	0.04716	0.05393	0.05397		
	正态记分检验															

（表 3-6）

（绿色：两种检验均拒绝原假设，蓝色：仅其中一种检验拒绝原假设）

在 95%置信水平之下分析上表，首先对于飞行游戏，除了棋牌游戏和游戏工具两类游戏之外，它与其他 12 类游戏在游戏评分高低上的假设检验结果均接受飞行游戏评分更小的备择假设；同样对于棋牌游戏，除了飞行游戏、竞速游戏、音乐游戏和游戏工具四类游戏外，可以接受其评分低于其他 10 类游戏的备择假设；此外，对于游戏工具也有类似结论，除了飞行游戏、竞速游戏和棋牌游戏外，可以认为其评分低于其他 11 类游戏。

因此，我们发现飞行游戏、棋牌游戏和游戏工具三种类型的游戏评分略低于其他类型的游戏，一方面，我们有理由推断，开发上述三类游戏可能难以获得高评分，受欢迎程度不高，开发商将面临一定的挑战；另一方面，上述结果也从侧面反映出市场现存的飞行游戏、棋牌游戏和游戏工具这三类游戏，其好玩程度不够，仍然存在很大的发展空间。

问题三：哪种类型的游戏内部差异化比较明显？

进一步考虑对于不同类型的游戏其游戏热度和评分的内部波动性是否存在差异，分别进行 15 类游戏热度和评分的尺度参数检验。这里我们分别以两样本和多样本平方秩检验为例说明上述问题。

首先关于游戏热度的 15 类多样本平方秩检验：

$$H_0：\ \sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2 \quad \text{v.s.} \quad H_1：\text{至少有一个等号不成立}$$

上述假设检验的 p 值为 0.00002238268，在 95%置信水平之下，我们拒绝原假设，因此我们认为 15 类手机游戏类内关于热度的波动性在不同类间存在差异，我们认为对于同一类型的不同手机游戏之间热度的差异性在不同类型之间的表现不同，这意味着可能存在热度值差异较大的游戏类别，也可能存在热度值差不多的游戏类别。

对于游戏评分的多样本平方秩检验 p 值为 0.008749428，同样拒绝原假设，我们得到了相似的结论。

进一步对 15 类游戏中的任意两类游戏进行关于热度和评分的两样本平方秩检验，零假设条件均为：$H_0：\ \sigma_1^2=\sigma_2^2$，当备择假设为 $H_1：\ \sigma_1^2<\sigma_2^2$ 时，我们将假设检验的 p 值记录为负数，当备择假设为 $H_1：\ \sigma_1^2>\sigma_2^2$ 时，我们将假设检验的 p 值记录为正数，其中对于游戏热度的平方秩检验，σ_1^2 和 σ_2^2 分别代表横坐标和纵坐标对应专业名称，而对于游戏评分的平方秩检验，σ_1^2 和 σ_2^2 分别代表纵坐标和横坐标对应专业名称（表 3-7）。

	策略塔防	动作游戏	飞行游戏	格斗游戏	角色扮演	竞速游戏	冒险解谜	模拟经营	棋牌游戏	射击游戏	体育运动	养成游戏	益智休闲	音乐游戏	游戏工具	
策略塔防		-0.2140	0.0475	0.4377	-0.1752	-0.0960	-0.4885	0.0192	0.2009	0.0368	-0.0002	-0.1750	-0.1761	0.3182	-0.2925	游戏热度尺度检验
动作游戏	-0.1348		0.0248	0.1738	0.4781	0.9040	0.1758	0.0106	0.1766	0.0042	-0.0003	-0.4164	-0.3878	0.1110	-0.4048	
飞行游戏	0.0000	0.0000		-0.1231	-0.0150	1.9040	-0.0610	-0.4549	-0.2872	-0.3191	-0.0002	-0.0477	-0.0086	-0.2522	-0.0779	
格斗游戏	-0.0025	-0.0738	0.0000		-0.2261	2.9040	-0.4814	0.1145	0.3430	0.1409	-0.0029	-0.2155	-0.1580	0.3357	-0.2763	
角色扮演	-0.1544	0.4390	0.0000	0.0446		3.9040	0.2440	0.0040	0.0870	0.0014	-0.0001	-0.3299	-0.4598	0.1582	-0.4068	
竞速游戏	0.0002	0.0000	0.0000	0.0000	0.0000		0.0799	0.0090	0.1352	0.0044	-0.0198	0.4296	0.2410	0.0613	0.4107	
冒险解谜	-0.1130	-0.3857	0.0000	0.1459	-0.3216	-0.0002		0.0235	0.1729	0.0324	-0.0002	-0.2424	-0.1005	0.3339	-0.3177	
模拟经营	0.0784	0.0089	0.0000	0.0000	0.0061	-0.0169	0.0096		-0.3661	-0.3993	0.0000	-0.0314	-0.0017	-0.3374	-0.0430	
棋牌游戏	0.0000	0.0000	0.0069	0.0000	0.0000	0.0000	0.0000	0.0000		0.3592	-0.0028	-0.0998	-0.0985	0.3748	-0.1688	
射击游戏	-0.2188	0.3314	0.0000	0.0002	0.2895	0.0000	0.2082	-0.0086	0.0000		0.0000	-0.0380	-0.0007	-0.3578	-0.0826	
体育运动	0.0000	0.0000	0.0000	0.0000	0.0000	0.0590	0.0000	0.0000	0.0000	0.0000		0.0776	0.0001	0.0022	0.0425	
养成游戏	0.0001	0.0001	0.0198	0.0000	0.0001	0.0004	0.0001	0.0001	-0.2990	0.0001	0.0001		0.3972	0.2129	0.4911	
益智休闲	0.0000	-0.0012	0.0000	-0.2840	-0.0002	0.0000	-0.0071	0.0000	0.0000	0.0000	0.0000	-0.0001		0.1153	-0.4930	
音乐游戏	0.0000	0.0000	0.0044	0.0000	0.0000	0.0000	0.0000	0.0000	0.4522	0.0000	0.0000	0.4870	0.0000		-0.1259	
游戏工具	0.0005	0.0006	0.0003	0.0003	0.0005	0.0005	0.0005	0.0005	0.0004	0.0005	0.0004	0.0004	0.0006	0.0003		
游戏评分尺度检验																

（表 3-7）

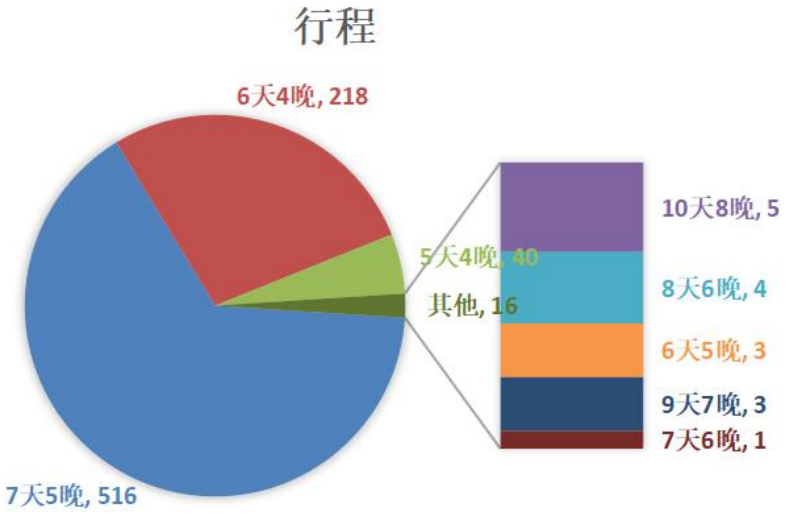
不难看出，对于不同类型的游戏，其热度的内部波动性普遍小于其评分的内部波动性。例如，对于益智休闲类型的游戏而言，此类中不同游戏间的热度差异与其他类型游戏中不同游戏的热度差异大致相同，而此类中不同游戏间的评分差异普遍大于其他类型中不同游戏的评分差异，因此我们可以发现益智休闲类游戏的评分的尺度参数较大，此类中游戏评分两极化现象更为明显，这与我们所观察到的评分最高和评分次低的游戏均为益智休闲类型游戏的现象相符。

（四）马尔代夫数据分析

1、数据集描述与简要分析

本数据集包含 2017 年 11 月至 2018 年 4 月共计 790 条马尔代夫旅游线路，涉及每条线路的名称、价格、行程、评分、评论数、店铺、出发日期、住宿信息。

其中，按行程分类，共有 7 天 5 晚、6 天 4 晚、5 天 4 晚等 8 类（图 4-1），其中行程类型主要以 7 天 5 晚为主，占比超过 65%，此外 6 天 4 晚和 5 天 4 晚两种行程也有一定的数量。按评分分类仅有 6 条线路评分 4.9、1 条线路评分 4.2，其余线路评分均为 5.0，所以此数据基本不具备选择参考性。



（图 4-1）

若按店铺分类，上述线路共来自 35 家店铺（表 4-1），其中店铺“携程自由行”在 2017 年 11 月至 2018 年 4 月的马尔代夫路线最多，高达 225 个，此外“第壹假期旅游网”和“国旅假期旗舰店”的马尔代夫路线数量紧随其后，也均超过 100 个。

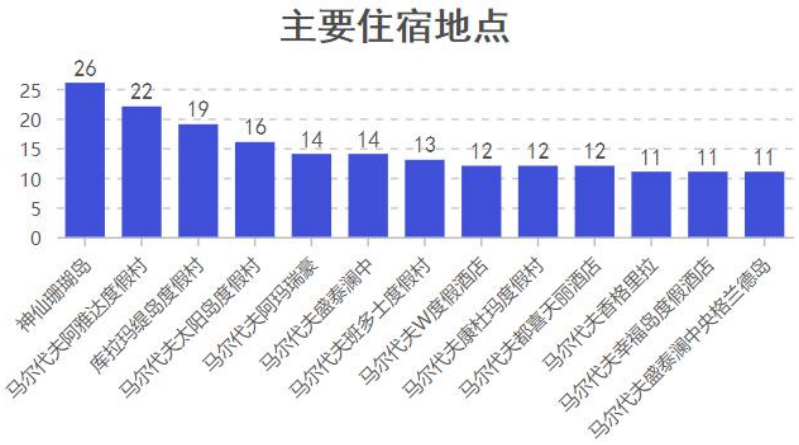
店铺	数量	店铺	数量	店铺	数量	店铺	数量
携程自由行	225	明远假期	15	心享世界	4	鲲鹏逍遥游	1
第壹假期旅游网	157	乐途国际	13	北京遇岛国旅专营店	2	青旅出境	1
国旅假期旗舰店	108	星煌旅游	10	深圳中之旅	2	青扬假期	1
国旅乐游网	47	华途旅游	9	深圳中之旅旗舰店	2	陕中旅旗舰店	1
蜗牛国旅旗舰店	41	国旅度假旗舰网	7	腾邦旅游	2	视界通旅行	1
一路行旅游	37	度假牛旅游网	6	中国旅行社专卖店	2	西安天马国旅未央...	1
四季行旅游	33	趣旅网	6	OK旅行网	1	游圣旅行	1
国旅海燕部	21	中旅世界行	5	宝中旅游飞哪儿旅行网	1	中国旅行社旗舰店	1
火星度假	21	佳天下旅游	4	金桥旅游和平店	1		

（表 4-1）

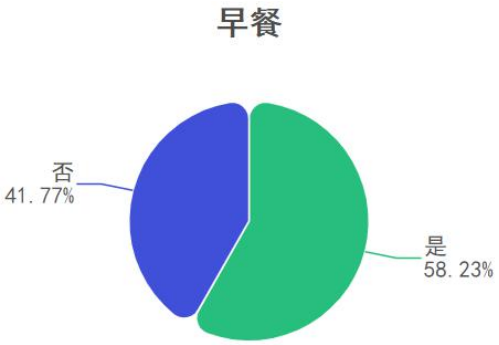
若按住宿分类，上述 790 条路线共涉及不同住宿 212 家，统计被比较多（10 条以上）线路所选择的酒店如下（图 4-2），其中共有 26 条线路选择“神仙珊瑚岛”作为住宿地，“神仙珊瑚岛”在所有住宿地中被选择的次数最多。

注意到对于不同的线路，其住宿有含有早餐与不含早餐之分（图 4-3），也有住宿是否

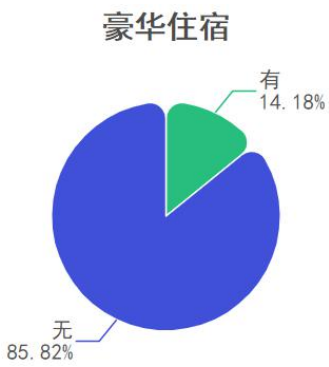
是豪华住宿之分（图 4-4）。经过统计发现超过 58%的住宿是含有早餐的，并且豪华住宿占比不足住宿总体的 15%。在这里我们猜测是否含早、是否是豪华住宿均有对该线路价格产生影响的可能性。



（图 4-2）



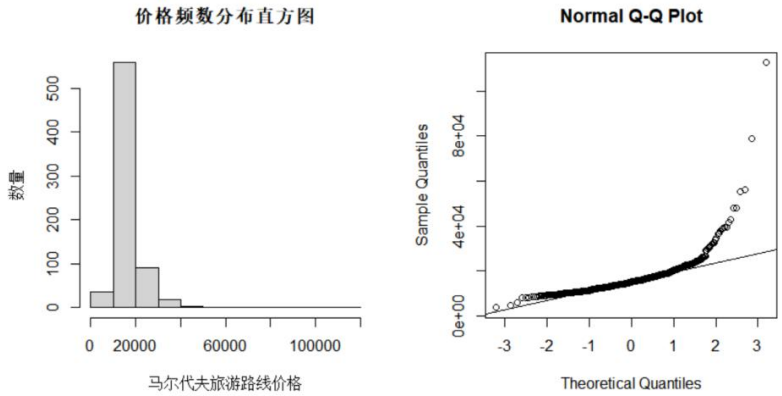
（图 4-3）



（图 4-4）

由于在 790 条线路中仅有 7 条路线评分不为 5.0，我们认为评分在各线路中不具备显著的区分性。对于评论数，观察到有超过一半的线路评论为零，所以评论数也不能直观地对一条路线的好坏作出评价。综合考量，我们选择采用最后一个量化指标——价格作为不同线路之间的选择指标。

在进行分析之前，我们首先考虑各线路价格是否服从正态分布。通过直方图和 QQ 图（图 4-5）以及 Anderson-Darling 检验、Cramer-von Mises 检验、Pearson 卡方检验、Shapiro-Francia 检验、Shapiro-Wilk 检验等多种正态性检验方法，均拒绝正态性假设，证明马尔代夫旅游路线价格不服从正态分布，我们将采用一些非参数方法对马尔代夫旅游线路价格进行评价。



（图 4-5）

2、问题提出与解答

问题一：各种行程之间的价格有无显著差异？

我们将 790 条线路按行程分为 7 类并计数，对任意两类进行 Kolmogorov-Smirnov 检验，零假设为两类行程价格具有相同的分布，经过检验发现，只有 5 天 4 晚这类行程的价格分布与其他类型不同（表 4-2），可以进一步考虑 5 天 4 晚这类行程与其他行程的位置参数的检验问题。

行程	数量
5天4晚	40
6天4晚	218
6天5晚	3
7天5晚	516
8天6晚	5
9天7晚	3
10天8晚	5

	5天4晚	6天4晚	6天5晚	7天5晚	8天6晚	9天7晚	10天8晚
5天4晚		1.10E-10	0.0357	5.15E-09	0.0008	0.0091	0.0002
6天4晚			0.6834	0.2389	0.3421	0.7015	0.1296
6天5晚				0.6655	0.2857	0.6000	0.4643
7天5晚			KS检验		0.4219	0.5616	0.1104
8天6晚	仅有5天4晚的行程价格分布与其他行程不同					0.2857	0.0794
9天7晚							0.1429
10天8晚							

（表 4-2）

我们采用两样本 Wilcoxon (Mann-Whitney)秩和检验解决上述问题。M_1 表示 5 天 4 晚行程中位数，M_2 表示其他类型行程中位数。

$$H_0:\ M_1 = M_2 \quad \text{v.s.} \quad H_1:\ M_1 < M_2$$

根据假设检验结果（表 4-3），我们接受 5 天 4 晚类型行程价格低于其他类型行程的假设，这与我们的预设结果相符合，5 天 4 晚的出行时间较短，所以费用相对较低。

Wilcoxon	6天4晚	6天5晚	7天5晚	8天6晚	9天7晚	10天8晚
5天4晚	2.56E-11	0.0083	3.30E-11	0.0015	0.0083	6.79E-05

（表 4-3）

值得注意的是，在两样本 K-S 检验中，我们发现除了 5 天 4 晚这类行程之外，其余行程的价格可以认为具有相同的分布。其中，6 天 4 晚、8 天 6 晚、9 天 7 晚、10 天 8 晚四类行程的样本量过少，所以其检验结果存在较大的偶然性，难以承认其普遍意义。

但对于 6 天 4 晚和 7 天 5 晚两类具有大量样本的行程类型，其 K-S 假设检验结果认为二者价格具有相同的分布，关于其位置参数和尺度参数的检验也认为二者的价格中位数和价格波动性没有显著差异，这点是出乎我们意料的。

所以，由于 6 天 4 晚和 7 天 5 晚两类行程的价格差异性不明显，站在消费者的角度，我们选择 7 天 5 晚的行程可以在差不多的价格下多游玩一天，对于消费者而言会更加实惠。

问题二：不同店铺之间的路线报价是否存在显著差异？

由于本数据集共涉及 35 家店铺，店铺数目较多，我们仅选择在本时间段内计划线路超过 100 条的三家大型店铺（携程自由行、第壹假期旅游网、国旅假期旗舰店）进行分析。

首先对三家店铺价格分布进行两两之间的 K-S 检验（表 4-4），我们发现“携程自由行”和“第壹假期旅游网”两家店铺的线路价格分布存在差异。

店铺	线路数量
携程自由行	210
第壹假期旅游网	157
国旅假期旗舰店	108

	携程自由行	第壹假期旅游网	国旅假期旗舰店
携程自由行		0.04258	0.2674
第壹假期旅游网			0.8936
国旅假期旗舰店			

（表 4-4）

为了进一步了解上述两家店铺关于马尔代夫旅游线路的价格差异表现，我们首先进行“携程自由行”和“第壹假期旅游网”两家店铺价格中位数检验，采用 Brown-Mood 检验、Wilcoxon (Mann-Whitney)秩和检验和正态记分检验进行双边检验（表 4-5），均接受零假设，即认为两家店铺马尔代夫旅游线路价格中位数不存在显著差异。

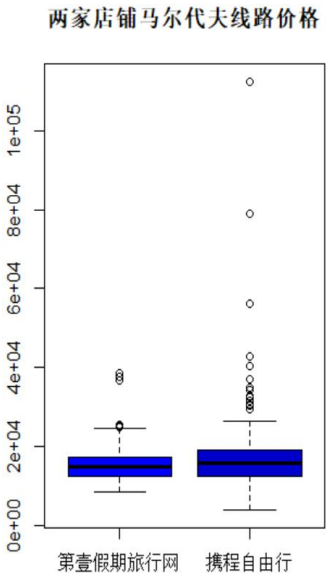
	Brown-Mood检验	Wilcoxon检验	正态记分检验
p值	0.4555	0.9741	0.5914834

（表 4-5）

所以，我们猜测上述两家店铺的价格分布差异在于其价格波动性，所以我们继续对两家店铺的线路价格进行尺度参数检验。采用 Mood 检验、Ansari-Bradley 检验、Flingner-Killeen 检验和平方秩检验进行验证（表 4-6）。我们发现上述四种尺度参数检验均拒绝零假设，接受备择假设：“携程自由行”的马尔代夫旅游线路价格波动性大于“第壹假期旅游网”，这一结论与两家店铺价格的箱线图（图 4-6）直接观察结果保持一致。

	Mood检验	AB检验	FK检验	平方秩检验
p值	0.00004083	0.0001215	0.0005027	0.000005619

（表 4-6）



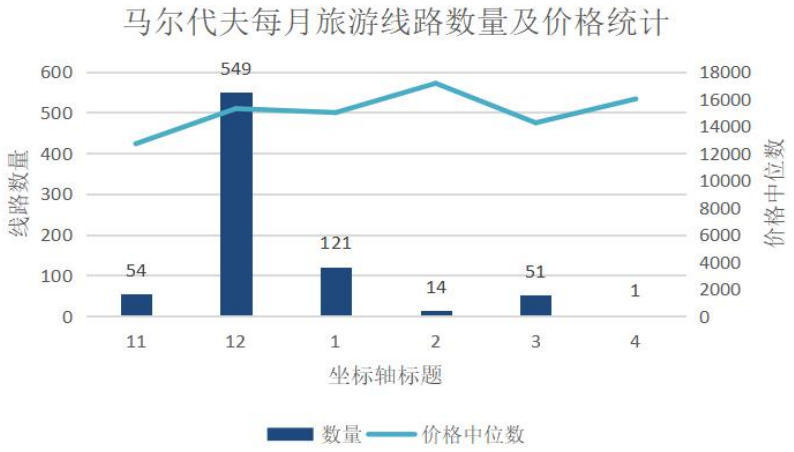
（图 4-6）

根据上述结果，我们不难猜测两家店铺的销售策略是存在差异的。第壹假期旅游网提供的线路价格差异化不大，并没有价格极低或极高的线路方案；而携程自由行提供的线路存在部分特惠线路以及豪华线路，所以为希望节省开支或者追求马尔代夫极致体验的游客提供了更多的选择机会。但注意到，两家店铺虽然采用了不同的销售策略，造成了线路价格波动性的差异，但线路价格的中位数大致相同，从而均保证了适宜的盈利。

问题三：各月份前往马尔代夫旅行的价格有无显著差异？

首先我们统计了 2017 年 11 月至 2018 年 4 月，每月前往马尔代夫的旅游线路数量以及该月旅游线路价格中位数（图 4-7）。

我们发现在 12 月出发前往马尔代夫的旅游线路数量最多，高达 549 次，远远超过其他月份的总和，所以我们合理推测 12 月是前往马尔代夫旅游的最适宜月份，并且选择在 12 月份出游马尔代夫的人数最多。



（图 4-7）

我们感兴趣的是在各个月份前往马尔代夫旅行的花费是否存在差异，由于 4 月份只有一个出发线路，样本过少，所以我们仅对 12 月至 3 月进行五样本马尔代夫路线价格中位数检验，Kruskal-Wallis 检验结果的 p 值为 0.00114，所以我们在 95%置信水平下可以认为这五个月的价格存在差异。

```
        Kruskal-Wallis rank sum test

data:  data_kw1 by group1
Kruskal-Wallis chi-squared = 18.176, df = 4, p-value = 0.00114
```

进一步进行任意两样本中位数的 Wilcoxon (Mann-Whitney)秩和检验和正态记分检验，该假设检验的零假设条件均为：H_0：$M_1 = M_2$，当备择假设为H_1：$M_1 < M_2$时，我们将假设检验的 p 值记录为负数，当备择假设为H_1：$M_1 > M_2$时，我们将假设检验的 p 值记录为正数，其中对于秩和检验，M_1和M_2分别代表横坐标和纵坐标对应的月份，而对于正态记分检验，M_1和M_2则分别代表纵坐标和横坐标对应的月份，假设检验结果记录如下（表 4-7）。

	11月	12月	1月	2月	3月	
11月		-0.0001	-0.0024	-0.0033	-0.0240	秩和检验
12月	-0.0001		0.2670	-0.0952	0.0526	
1月	-0.0159	0.0963		-0.1022	0.1766	
2月	-0.0211	-0.1738	-0.1409		0.0239	
3月	-0.0924	0.0183	0.1301	0.0201		
正态记分检验						

（表 4-7）

综合两种检验结果，我们有理由承认：在 11 月前往马尔代夫旅行的价格比其他月份都低，而在剩下的月份中，3 月前往马尔代夫会比 12 月或 2 月前往马尔代夫便宜。所以，但就旅行费用而言，我们建议避开人流高峰期 12 月，如果想在年前前往马尔代夫，可以选择在 11 月出行，如果想在年后前往马尔代夫，可以选择在 3 月出行，更容易选择到比较实惠的路线。

问题四：住宿是否含有早餐、是否是豪华住宿对旅行价格的高低有没有显著影响？

我们注意到在数据集中包含的 212 家住宿有含有早餐与不含早餐、豪华住宿与非豪华住宿之分，我们猜测这两个因素可能也会对该线路的价格产生不同程度的影响。

①早餐

首先考虑是否含有早餐这个因素，我们按住宿是否含有早餐将所有样本分为两类，我们进行两样本 K-S 分布检验以及位置参数检验，检验结果如下（表 4-8）：

	KS检验	Brown-Mood检验	Wilcoxon检验	正态记分检验
p值	0.2582	0.05896	0.1073	0.1158

（表 4-8）

在 95%置信水平下，我们无法拒绝零假设，即认为住宿是否含有早餐对其线路的价格影响不明显，所以为了方便，我们可以选择住宿包含早餐的线路。

②豪华住宿

其次，我们按住宿是否属于豪华住宿将线路分为两类，同样进行两样本分布检验，根据 K-S 检验结果（p 值=0.00005323），我们认为住宿豪华与否对该线路的价格有显著影响。

通过两样本 Brown-Mood 检验、Wilcoxon (Mann-Whitney)秩和检验和正态记分检验（表 4-9），我们可以接受豪华住宿价格高于普通住宿，所以住宿条件是否豪华是影响该线路价格的重要因素，如果该线路的住宿条件较好，会比较明显的反映在价格上。

	Brown-Mood检验	Wilcoxon检验	正态记分检验
p值	1.31E-05	4.78E-05	2.69E-04

（表 4-9）

附录：R 语言处理代码

（一）体温数据

```
library(tseries)#游程检验
library(nortest)#正态性检验包
library(fBasics)#两样本 ks 检验包

#导入 8 月数据，序号，上午，下午，月，日，学生
temp8=read.csv("E:/2021 秋季学期/非参数统计/体温/每日体温监测-上午下午 8 月.csv",header=TRUE)
head(temp8)
temp1<-as.data.frame(temp8)
attach(temp8)
summary(temp8)

#导入 9 月数据，序号，上午，下午，月，日，学生
temp9=read.csv("E:/2021 秋季学期/非参数统计/体温/每日体温监测-上午下午 9 月.csv",header=TRUE)
head(temp9)
temp1<-as.data.frame(temp9)
attach(temp9)
summary(temp9)

#导入 10 月数据，序号，上午，下午，月，日，学生
temp10=read.csv("E:/2021 秋季学期/非参数统计/体温/每日体温监测-上午下午 10 月.csv",header=TRUE)
head(temp10)
temp1<-as.data.frame(temp10)
attach(temp10)
summary(temp10)

#总数据上午下午
temp_mor=c(temp8$上午,(na.omit(temp9))$上午,(na.omit(temp10))$上午)
temp_aft=c(temp8$下午,(na.omit(temp9))$下午,(na.omit(temp10))$下午)

#原始数据横板
temph=read.csv("E:/2021 秋季学期/非参数统计/体温/体温总表横版.csv",header=TRUE)
head(temph)
temph<-as.data.frame(temph)
attach(temph)
summary(temph)
```

```
#原始数据竖板
temps=read.csv("E:/2021 秋季学期/非参数统计/体温/体温总表竖版.csv",header=TRUE)
head(temps)
temps<-as.data.frame(temps)
attach(temps)
summary(temps)

#上午和下午体温数据相关性
cor.test(temp8$上午,temp8$下午,meth="spearman")
cor.test(temp8$上午,temp8$下午,meth="kendall")
cor.test((na.omit(temp9))$上午,(na.omit(temp9))$下午,meth="spearman")
cor.test((na.omit(temp9))$上午,(na.omit(temp9))$下午,meth="kendall")
cor.test((na.omit(temp10))$上午,(na.omit(temp10))$下午,meth="spearman")
cor.test((na.omit(temp10))$上午,(na.omit(temp10))$下午,meth="kendall")
cor.test(temp_mor,temp_aft,meth="spearman")
cor.test(temp_mor,temp_aft,meth="kendall")

#正态性检验
par(mfrow=c(1,2))
hist(temp_mor,freq=T,xlab="体温",ylab="数量",main="上午体温数据频数分布直方图")
qqnorm(temp_mor);qqline(temp_mor);
ad.test(temp_mor);cvm.test(temp_mor);pearson.test(temp_mor);sf.test(temp_mor);shapiro.test(
temp_mor)
hist(temp_aft,freq=T,xlab="体温",ylab="数量",main="下午体温数据频数分布直方图")
qqnorm(temp_aft);qqline(temp_aft);
ad.test(temp_aft);cvm.test(temp_aft);pearson.test(temp_aft);sf.test(temp_aft);shapiro.test(temp
_aft)

#回归
x=temp_mor;
y=temp_aft;
summary(lm(y~x))
lqs(y~x,method="lms")
lqs(y~x,method="lts"))
lqs(y~x,method="S")

#ks 检验
ks.test(temp_mor,temp_aft)#上午和下午的温度分布有差异

#两样本位置参数检验：上午和下午  下午体温中位数高
#Brown-Mood
z=cbind(c(temp_mor,temp_aft),c(rep(1,3893),rep(2,3893)))
k=unique(z[,2]);m=median(z[,1]);m1=NULL;m2=NULL
for(i in k){m1=c(m1,sum(z[z[,2]==i,1]>m));m2=c(m2,sum(z[z[,2]==i,1]<=m))}
```

```
C=rbind(m1,m2)
fisher.test(C,alt="less")#0.1001
#Wilcoxon
wilcox.test(temp_mor,temp_aft,alt="less")#0.001352
#正态记分检验
x=temp_mor; y=temp_aft;
m=length(x);n=length(y);
w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
T=sum(w[w[,2]==1,4])
w2=sum(w[,4]^2)
S=sqrt(m*n*w2/(m+n-1)/(m+n))
Z=T/S
p=pnorm(Z)
p#1.28535e-66
#成对数据检验
wilcox.test(temp_mor,temp_aft,paired=T,alt="less")

#区组 friedman 检验
#横着是不同学生，竖着是上下午
d=matrix(0,2,46)
k=1
for(i in 3:48)
{ n1=0;n2=0
   sum1=0;sum2=0
   for(j in seq(1,174,2))
   { if(is.na(temps[,i][j])==FALSE){sum1=sum1+temps[,i][j];n1=n1+1;}
      if(is.na(temps[,i][j+1])==FALSE){sum2=sum2+temps[,i][j+1];n2=n2+1;}
   }
   d[1,k]=sum1/n1;d[2,k]=sum2/n2;k=k+1;
}
d
write(as.vector(t(d)),file="E:/2021 秋季学期/非参数统计/体温/完全区组设计矩阵.csv")
friedman.test(d)#0.0002681  上下午是有差异的
#kendall
R=apply(d,2,sum)
b=nrow(d);k=ncol(d)
S=sum((R-b*(k+1)/2)^2)
W=12*S/b^2/(k^3-k)
pchisq(b*(k-1)*W,k-1,low=F)# 0.0003170306

#两样本尺度参数检验：上午和下午  下午体温波动大
#平方秩
x=temp_mor; y=temp_aft;
```

```
m=length(x);n=length(y);
x1=abs(x-mean(x));y1=abs(y-mean(y));
xy1=c(x1,y1);xy0=c(x,y);xyi=c(rep(1,m),rep(2,n));
xy=cbind(xy1,xy0,xyi)
xy2=cbind(xy[order(xy[,1]),],1:(m+n),(1:(m+n))^2)
T1=sum(xy2[xy2[,3]==1,5]);T2=sum(xy2[xy2[,3]==2,5])
R=xy2[,5];meanR=mean(R);
S=sqrt(m*n*(sum(R^2)-(m+n)*meanR^2)/(m+n)/(m+n-1))
Zx=(T1-m*meanR)/S;Zy=(T2-n*meanR)/S
p=min(pnorm(Zx),pnorm(Zy))
p #8.049131e-09  小于

#Siegel-Tukey
z=cbind(c(temp_mor,temp_aft),c(rep(1,3893),rep(2,3893)))
x=temp_mor; y=temp_aft;
x1=x-median(outer(x,y,"-"))
xy=cbind(c(x1,y),c(rep(1,length(x)),rep(2,length(y))))
xy1=xy[order(xy[,1]),]
z=xy[,1];n=length(z);a1=2:3;b=2:3
for(i in seq(1,n,2)){ b=b+4;a1=c(a1,b) }
a2=c(1,a1+2);z=NULL
for(i in 1:n) z=c(z,(i-floor(i/2)))
b=1:2
for(i in seq(1,(n+2-2),2))
if(z[i]/2!=floor(z[i]/2)) {z[i:(i+1)]=b;b=b+2}
zz=cbind(c(0,0,z[1:(n-2)]),z[1:n])
if(n==1) R=1;
if(n==2) R=c(1,2);
if(n>2) R=c(a2[1:zz[n,1]],rev(a1[1:zz[n,2]]))
xy2=cbind(xy1,R)
Wx=sum(xy2[xy2[,2]==1,3])
Wy=sum(xy2[xy2[,2]==2,3])
nx=length(x);ny=length(y)
Wxy=Wy-0.5*ny*(ny+1);Wyx=Wx-0.5*nx*(nx+1)
pvalue=pwilcox(Wyx,nx,ny)
pvalue #0.01534668

#Mood
z=cbind(c(temp_mor,temp_aft),c(rep(1,3893),rep(2,3893)))
x=temp_mor; y=temp_aft;
m=length(x);n=length(y)
x1=x-median(outer(x,y,"-"))
xy=cbind(c(x1,y),c(rep(1,length(x)),rep(2,length(y))))
N=nrow(xy);xy1=cbind(xy[order(xy[,1]),],1:N)
```

```
R1=xy1[xy1[,2]==1,3];M=sum((R1-(N+1)/2)^2)
E1=m*(N^2-1)/12;s=sqrt(m*n*(N+1)*(N^2-4)/180)
Z=(M-E1)/s
pvalue=pnorm(Z)
pvalue # 0.07396175  小于

#ansari-bradley  不好
x=temp_mor; y=temp_aft;
ansari.test(x,y,alt="greater")
#0.4633

#fligner-killeen
fligner.test(c(temp_mor,temp_aft),c(rep(1,3893),rep(2,3893)))
#0.001748


#游程检验，零假设序列是随机的
test=sign(temps$X4-median(temps$X4))
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#0.7055
runs.test(factor(test2))#0.1464
test=sign(temps$X11-median(temps$X11))#小于
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#6.443e-07
runs.test(factor(test2))#不是二分了，只有 36.4 和 36.5
test=sign(temps$X20-median(temps$X20))#8 月大多大于等于中位数，10 月大多小于等于中位
数
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#5.213e-15
runs.test(factor(test2))#1.662e-08
test=sign(temps$X25-median(temps$X25))
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#0.01899
runs.test(factor(test2))#5.861e-06
test=sign(temps$X28-median(temps$X28))
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#0.02073
runs.test(factor(test2))#0.2892
test=sign(temps$X42-median(temps$X42))
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#0.3087
runs.test(factor(test2))#0.03622
test=sign(temps$X44-median(temps$X44))#8 月大多小于等于中位数，10 月大多大于等于中位
数
```

```
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#0.001965
runs.test(factor(test2))#5.569e-08
test=sign(temps$X46-median(temps$X46))#上午低，下午高
test1=test;test2=test;test1[test1==0]=1;test2[test2==0]=-1
runs.test(factor(test1))#7.926e-10
runs.test(factor(test2))#1.192e-14

#cox-stuart 趋势检验
D=temps$X4[1:87]-temps$X4[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#0.4478416 不拒绝 无趋势
D=temps$X11[1:87]-temps$X11[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#0.04007166 增加
D=temps$X20[1:87]-temps$X20[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#1.511817e-09 减少
D=temps$X25[1:87]-temps$X25[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#0.5504618 不拒绝 无趋势
D=temps$X28[1:87]-temps$X28[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#5.068224e-08 增加
D=temps$X42[1:87]-temps$X42[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#0.1409895 不拒绝 无趋势
D=temps$X44[1:87]-temps$X44[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#0.07401609 有趋势 增加
D=temps$X46[1:87]-temps$X46[88:174]
K=min(sum(sign(D)==1),sum(sign(D)==-1))
pbinom(K,sum(sign(D)==1)+sum(sign(D)==-1),0.5)#0.1840938 不拒绝 无趋势
```

（二）成绩数据

```
library(nortest)#正态性检验包
library(scales)#show_col 的包
library(nortest)#正态性检验包
library(RColorBrewer)#颜色数据集
colors()
brewer.pal.info
display.brewer.all()
library(psych)#碎石图
library(MASS)#回归
library(mblm)#回归

#导入数据，原始数据表
scores=read.csv("E:/2021 秋季学期/非参数统计/成绩/考试成绩.csv",header=TRUE)
head(scores)
scores<-as.data.frame(scores)
attach(scores)
summary(scores)

#各题得分率表
rates=read.csv("E:/2021 秋季学期/非参数统计/成绩/各题得分率.csv",header=TRUE)
head(rates)
rates<-as.data.frame(rates)
attach(rates)
summary(rates)

#得到各专业数据集
scores_gs<-as.data.frame(scores[6:35,]);show(scores_gs);summary(scores_gs)
scores_sc<-as.data.frame(scores[36:67,]);show(scores_sc);summary(scores_sc)
scores_rl<-as.data.frame(scores[77:99,]);show(scores_rl);summary(scores_rl)
scores_tw<-as.data.frame(scores[100:112,]);show(scores_tw);summary(scores_tw)
scores_kj<-as.data.frame(scores[c(2,3,68:75),]);show(scores_kj);summary(scores_kj)
scores_cw<-as.data.frame(scores[c(4,5,76),]);show(scores_cw);summary(scores_cw)
scores_gl<-as.data.frame(scores[1,]);show(scores_gl);summary(scores_gl)

#正态性检验
par(mfrow=c(1,2))
hist(scores$成绩,freq=T,xlab="总成绩",ylab="人数",main="总成绩频数分布直方图")
qqnorm(scores$成绩);qqline(scores$成绩);
ad.test(scores$ 成 绩 );cvm.test(scores$ 成 绩 );pearson.test(scores$ 成 绩 );sf.test(scores$ 成
绩);shapiro.test(scores$成绩)
```

```
#自定义调色板
mypalette<-colorRampPalette(c("steelblue1","blue3"))
mycolors<-mypalette(9)
show_col(mycolors)
#各题得分率的线箱图
boxplot(rates,names=c("第一题","第二题","第三题","第四题","第五题","第六题","第七题","第八题","第九题"),
        col=mycolors,main="各题得分率箱线图")

#PCA
scores_tt=scores[,4:12]
fa.parallel(scores_tt,fa="pc",n.iter=100,show.legend=F)
pca=princomp(scores_tt,cor=T)
summary(pca,loadings=T)
pca_data=predict(pca)
pca_data


#得到各题得分率向量，多样本位置参数检验
t11=rates$X1;t22=rates$X2;t33=rates$X3;t44=rates$X4;t55=rates$X5
t66=rates$X6;t77=rates$X7;t88=rates$X8;t99=rates$X9
data_kw1=c(t11,t22,t33,t44,t55,t66,t77,t88,t99)
mydata_kw1<-data.frame(data_kw1,group1=c(rep("第 1 题",112),rep("第 2 题",112),rep("第 3 题",112),
                                         rep("第 4 题",112),rep("第 5 题",112),rep("第 6 题",112),rep("第 7 题",112),
                                         rep("第 8 题",112),rep("第 9 题",112)))
kruskal.test(data_kw1~group1,data=mydata_kw1)
#任意道问题得分率差异
#wilcoxon 秩和检验
wilcox_scores=matrix(0,9,9)
for(i in 1:8)
{ for(j in (i+1):9)
   { wilcox_scores1=wilcox.test(rates[,i],rates[,j],alt="less")
     wilcox_scores2=wilcox.test(rates[,i],rates[,j],alt="greater")
     if(wilcox_scores1$p.value<wilcox_scores2$p.value)
{wilcox_scores[i,j]=wilcox_scores1$p.value*(-1)}
     else {wilcox_scores[i,j]=wilcox_scores2$p.value}
   }
}
wilcox_scores
wilcox_scores_v=as.vector(t(wilcox_scores))
write(wilcox_scores_v,file="E:/2021 秋季学期/非参数统计/成绩/任意两题得分 wilcoxon 秩和检验.csv")
```

```
#正态记分检验
normal_scores=matrix(0,9,9)
for(i in 1:8)
{ for(j in (i+1):9)
   { x=rates[,i]; y=rates[,j];
      m=112; n=112;
      w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
      w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
      T=sum(w[w[,2]==1,4])
      w2=sum(w[,4]^2)
      S=sqrt(m*n*w2/(m+n-1)/(m+n))
      Z=T/S
      p=pnorm(Z)
      if(p>0.5){p=1-p}
      else {p=p*(-1)}
      normal_scores[i,j]=p
   }
}
normal_scores
normal_scores_v=as.vector(t(normal_scores))
write(normal_scores_v,file="E:/2021 秋季学期/非参数统计/成绩/任意两题得分正态记分检
验.csv")

#尺度检验
data_fk1=cbind(c(t11,t22,t33,t44,t55,t66,t77,t88,t99),
                c(rep(1,112),rep(2,112),rep(3,112),rep(4,112),rep(5,112),
                  rep(6,112),rep(7,112),rep(8,112),rep(9,112)))
fligner.test(data_fk1[,1],data_fk1[,2])
#平方秩检验
#多样本
N=1008;k=9
d2=NULL
d=data_fk1
for(i in 1:k) d2=rbind(d2,cbind(abs(d[d[,2]==i,1]-mean(d[d[,2]==i,1])),d[d[,2]==i,1],i))
d3=cbind(d2[order(d2[,1]),],1:N,(1:N)^2)
Ti=NULL
for(i in 1:k) Ti=c(Ti,sum(d3[d3[,3]==i,5]))
ni=NULL
for(i in 1:k) ni=c(ni,nrow(d3[d3[,3]==i,]))
T=(N-1)*(sum(Ti^2/ni)-sum(Ti)^2/N)/(sum(d3[,5]^2)-sum(Ti)^2/N)
pvalue=pchisq(T,k-1,low=F)
pvalue
#两样本
square_scores=matrix(0,9,9)
```

```
for(i in 1:8)
{ for(j in (i+1):9)
   { x=rates[,i];y=rates[,j]
      m=length(x);n=length(y)
      x1=abs(x-mean(x));y1=abs(y-mean(y))
      xy1=c(x1,y1);xy0=c(x,y);xyi=c(rep(1,m),rep(2,n))
      xy=cbind(xy1,xy0,xyi)
      xy2=cbind(xy[order(xy[,1]),],1:(m+n),(1:(m+n))^2)
      T1=sum(xy2[xy2[,3]==1,5]);T2=sum(xy2[xy2[,3]==2,5])
      R=xy2[,5];meanR=mean(R)
      S=sqrt(m*n*(sum(R^2)-(m+n)*meanR^2)/(m+n)/(m+n-1))
      Zx=(T1-m*meanR)/S;Zy=(T2-n*meanR)/S
      if(pnorm(Zx)<pnorm(Zy)) {square_scores[i,j]=pnorm(Zx)*(-1)}
      else {square_scores[i,j]=pnorm(Zy)}
   }
}
square_scores
square_scores_v=as.vector(t(square_scores))
write(square_scores_v,file="E:/2021 秋季学期/非参数统计/成绩/任意两道题目得分尺度检验平方秩.csv")


#各专业同学成绩有无显著差异
#总分
data_kw2=c(scores_gs$ 成绩,scores_sc$ 成绩,scores_rl$ 成绩,scores_tw$ 成绩,scores_kj$ 成绩,scores_cw$成绩,scores_gl$成绩)
mydata_kw2<-data.frame(data_kw2,group2=c(rep("工商管理",30),rep("市场营销",32),
                                rep("人力资源管理",23),rep("天文学",13),rep("会计学",10),
                                rep("财务管理",3),rep("管理科学",1)))
kruskal.test(data_kw2~group2,data=mydata_kw2)
#任意两道问题得分率差异
#wilcoxon 秩和检验
data=c(scores_gs$ 成绩,scores_sc$ 成绩,scores_rl$ 成绩,scores_tw$ 成绩,scores_kj$ 成绩,scores_cw$成绩,scores_gl$成绩)
h=c(0,30,62,85,98,108,111,112)
wilcox_dep=matrix(0,7,7)
for(i in 1:6)
{ for(j in (i+1):7)
   { wilcox_dep1=wilcox.test(data[(h[i]+1):h[i+1]],data[(h[j]+1):h[j+1]],alt="less")
      wilcox_dep2=wilcox.test(data[(h[i]+1):h[i+1]],data[(h[j]+1):h[j+1]],alt="greater")
      if(wilcox_dep1$p.value<wilcox_dep2$p.value) {wilcox_dep[i,j]=wilcox_dep1$p.value*(-1)}
      else {wilcox_dep[i,j]=wilcox_dep2$p.value}
   }
}
```

```
wilcox_dep
wilcox_dep_v=as.vector(t(wilcox_dep))
write(wilcox_dep_v,file="E:/2021秋季学期/非参数统计/成绩/任意两专业成绩wilcoxon秩和检验.csv")
#正态记分检验
normal_dep=matrix(0,7,7)
h=c(0,30,62,85,98,108,111,112)
l=c(30,32,23,13,10,3,1)
for(i in 1:6)
{ for(j in (i+1):7)
   { x=data[(h[i]+1):h[i+1]]; y=data[(h[j]+1):h[j+1]];
      m=l[i]; n=l[j];
      w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
      w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
      T=sum(w[w[,2]==1,4])
      w2=sum(w[,4]^2)
      S=sqrt(m*n*w2/(m+n-1)/(m+n))
      Z=T/S
      p=pnorm(Z)
      if(p>0.5){p=1-p}
      else {p=p*(-1)}
      normal_dep[i,j]=p
   }
}
normal_dep
normal_dep_v=as.vector(t(normal_dep))
write(normal_dep_v,file="E:/2021秋季学期/非参数统计/成绩/任意两专业成绩正态记分检验.csv")


#尺度检验
data_fk2=list(scores_gs$成绩,scores_sc$成绩,scores_rl$成绩,scores_tw$成绩,scores_kj$成绩,scores_cw$成绩,scores_gl$成绩)
fligner.test(data_fk2)
#平方秩检验
#多样本
N=112;k=7
d2=NULL
d=cbind(data,c(rep(1,30),rep(2,32),rep(3,23),rep(4,13),rep(5,10),rep(6,3),rep(7,1)))
for(i in 1:k) d2=rbind(d2,cbind(abs(d[d[,2]==i,1]-mean(d[d[,2]==i,1])),d[d[,2]==i,1],i))
d3=cbind(d2[order(d2[,1]),],1:N,(1:N)^2)
Ti=NULL
for(i in 1:k) Ti=c(Ti,sum(d3[d3[,3]==i,5]))
ni=NULL
```

```
for(i in 1:k) ni=c(ni,nrow(d3[d3[,3]==i,]))
T=(N-1)*(sum(Ti^2/ni)-sum(Ti)^2/N)/(sum(d3[,5]^2)-sum(Ti)^2/N)
pvalue=pchisq(T,k-1,low=F)
pvalue
#两样本
square_dep=matrix(0,7,7)
for(i in 1:6)
{ for(j in (i+1):7)
   { x=data[(h[i]+1):h[i+1]]; y=data[(h[j]+1):h[j+1]];
      m=length(x);n=length(y)
      x1=abs(x-mean(x));y1=abs(y-mean(y))
      xy1=c(x1,y1);xy0=c(x,y);xyi=c(rep(1,m),rep(2,n))
      xy=cbind(xy1,xy0,xyi)
      xy2=cbind(xy[order(xy[,1]),],1:(m+n),(1:(m+n))^2)
      T1=sum(xy2[xy2[,3]==1,5]);T2=sum(xy2[xy2[,3]==2,5])
      R=xy2[,5];meanR=mean(R)
      S=sqrt(m*n*(sum(R^2)-(m+n)*meanR^2)/(m+n)/(m+n-1))
      Zx=(T1-m*meanR)/S;Zy=(T2-n*meanR)/S
      if(pnorm(Zx)<pnorm(Zy)) {square_dep[i,j]=pnorm(Zx)*(-1)}
      else {square_dep[i,j]=pnorm(Zy)}
   }
}
square_dep
square_dep_v=as.vector(t(square_dep))
write(square_dep_v,file="E:/2021 秋季学期/非参数统计/成绩/任意两个专业成绩尺度检验平
方秩.csv")

#分析总分差异主要出现在哪几道题上
#工商管理和市场营销
#两样本位置参数 Brown-Mood 检验
n1=dim(scores_gs)[1];n2=dim(scores_sc)[1];N=n1+n2
z=matrix(0,N,2)
B=c(rep(0,9))
for(i in 1:9)
{ z[,1]=c(scores_gs[,i+3],scores_sc[,i+3])
   z[,2]=c(rep(1,n1),rep(2,n2))
   m=median(z[,1]);m1=m2=NULL
   for(j in c(1,2)) {m1=c(m1,sum(z[z[,2]==j,1]>m));m2=c(m2,sum(z[z[,2]==j,1]<=m));}
   C=rbind(m1,m2)
   f1=fisher.test(C,alt="less")$p.value
   f2=fisher.test(C,alt="greater")$p.value
   if(f1<f2){B[i]=f1*(-1)}
   else {B[i]=f2}
}
```

```
B
i=2
#两样本位置参数 wilcoxon 秩和检验
w=c(rep(0,9))
for(i in 1:9)
{ w1=wilcox.test(scores_gs[,i+3],scores_sc[,i+3],alt="less")
   w2=wilcox.test(scores_gs[,i+3],scores_sc[,i+3],alt="greater")
   if(w1$p.value<w2$p.value) {w[i]=w1$p.value*(-1)}
   else {w[i]=w2$p.value}
}
w
#正态记分检验
H=c(rep(0,9))
for(i in 1:9)
{ x=scores_gs[,i+3]; y=scores_sc[,i+3];
   m=length(x);n=length(y);
   w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
   w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
   T=sum(w[w[,2]==1,4])
   w2=sum(w[,4]^2)
   S=sqrt(m*n*w2/(m+n-1)/(m+n))
   Z=T/S
   p=pnorm(Z)
   if(p>0.5){H[i]=1-p}
   else {H[i]=p*(-1)}
}
H

#工商管理和人力资源管理
#两样本位置参数 Brown-Mood 检验
n1=dim(scores_gs)[1];n2=dim(scores_rl)[1];N=n1+n2
z=matrix(0,N,2)
B=c(rep(0,9))
for(i in 1:9)
{ z[,1]=c(scores_gs[,i+3],scores_rl[,i+3])
   z[,2]=c(rep(1,n1),rep(2,n2))
   m=median(z[,1]);m1=m2=NULL
   for(j in c(1,2)) {m1=c(m1,sum(z[z[,2]==j,1]>m));m2=c(m2,sum(z[z[,2]==j,1]<=m));}
   C=rbind(m1,m2)
   f1=fisher.test(C,alt="less")$p.value
   f2=fisher.test(C,alt="greater")$p.value
   if(f1<f2){B[i]=f1*(-1)}
   else {B[i]=f2}
}
```

```
B
#两样本位置参数 wilcoxon 秩和检验
w=c(rep(0,9))
for(i in 1:9)
{ w1=wilcox.test(scores_gs[,i+3],scores_rl[,i+3],alt="less")
   w2=wilcox.test(scores_gs[,i+3],scores_rl[,i+3],alt="greater")
   if(w1$p.value<w2$p.value) {w[i]=w1$p.value*(-1)}
   else {w[i]=w2$p.value}
}
w
#正态记分检验
H=c(rep(0,9))
for(i in 1:9)
{ x=scores_gs[,i+3]; y=scores_rl[,i+3];
   m=length(x);n=length(y);
   w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
   w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
   T=sum(w[w[,2]==1,4])
   w2=sum(w[,4]^2)
   S=sqrt(m*n*w2/(m+n-1)/(m+n))
   Z=T/S
   p=pnorm(Z)
   if(p>0.5){H[i]=1-p}
   else {H[i]=p*(-1)}
}
H

#天文学和人力资源管理
#两样本位置参数 Brown-Mood 检验
n1=dim(scores_tw)[1];n2=dim(scores_rl)[1];N=n1+n2
z=matrix(0,N,2)
B=c(rep(0,9))
for(i in 1:9)
{ z[,1]=c(scores_tw[,i+3],scores_rl[,i+3])
   z[,2]=c(rep(1,n1),rep(2,n2))
   m=median(z[,1]);m1=m2=NULL
   for(j in c(1,2)) {m1=c(m1,sum(z[z[,2]==j,1]>m));m2=c(m2,sum(z[z[,2]==j,1]<=m));}
   C=rbind(m1,m2)
   f1=fisher.test(C,alt="less")$p.value
   f2=fisher.test(C,alt="greater")$p.value
   if(f1<f2){B[i]=f1*(-1)}
   else {B[i]=f2}
}
B
```

```
#两样本位置参数 wilcoxon 秩和检验
w=c(rep(0,9))
for(i in 1:9)
{ w1=wilcox.test(scores_tw[,i+3],scores_rl[,i+3],alt="less")
   w2=wilcox.test(scores_tw[,i+3],scores_rl[,i+3],alt="greater")
   if(w1$p.value<w2$p.value) {w[i]=w1$p.value*(-1)}
   else {w[i]=w2$p.value}
}
w
#正态记分检验
H=c(rep(0,9))
for(i in 1:9)
{ x=scores_tw[,i+3]; y=scores_rl[,i+3];
   m=length(x);n=length(y);
   w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
   w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
   T=sum(w[w[,2]==1,4])
   w2=sum(w[,4]^2)
   S=sqrt(m*n*w2/(m+n-1)/(m+n))
   Z=T/S
   p=pnorm(Z)
   if(p>0.5){H[i]=1-p}
   else {H[i]=p*(-1)}
}
H

#列联表，不同专业及格率是否相同
l1=matrix(c(12,18,24,8,20,3,10,3,5,5),nrow=5,ncol=2,byrow=T)
chisq.test(l1)
F=cbind(c(12,18,24,8),c(12,18,20,3),c(12,18,10,3),c(12,18,5,5),c(24,8,20,3),
          c(24,8,10,3),c(24,8,5,5),c(20,3,10,3),c(20,3,5,5),c(10,3,5,5))
L=matrix(0,10,10)
for(i in 1:10)
{ ff=matrix(F[,i],nrow=2,ncol=2,byrow=T)
   L[i,1]=fisher.test(ff)$p.value#拒绝 0.009465
   if(L[i,1]>=0.05){L[i,2:10]=NA}
   else
   { #比例差估计
      p1_hat=ff[1,1]/(ff[1,1]+ff[1,2])
      p2_hat=ff[2,1]/(ff[2,1]+ff[2,2])
      p_hat=p1_hat-p2_hat
      SE1=sqrt(p1_hat*(1-p1_hat)/(ff[1,1]+ff[1,2])+p2_hat*(1-p2_hat)/(ff[2,1]+ff[2,2]))
      CI_low1=p_hat-1.96*SE1
      CI_up1=p_hat+1.96*SE1
```

```
      L[i,2]=p_hat;L[i,3]=CI_low1;L[i,4]=CI_up1
      #相对风险 RR
      RR=p1_hat/p2_hat
      SE2=sqrt((1-p1_hat)/ff[1,1]+(1-p2_hat)/ff[2,1])
      CI_low2=RR*exp(-1.96*SE2)
      CI_up2=RR*exp(1.96*SE2)
      L[i,5]=RR;L[i,6]=CI_low2;L[i,7]=CI_up2
      #胜算比 OR
      OR=ff[1,1]*ff[2,2]/(ff[1,2]*ff[2,1])
      SE3=sqrt(1/ff[1,1]+1/ff[1,2]+1/ff[2,1]+1/ff[2,2])
      CI_low3=OR*exp(-1.96*SE3)
      CI_up3=OR*exp(1.96*SE3)
      L[i,8]=OR;L[i,9]=CI_low3;L[i,10]=CI_up3
   }
}
L
L_v=as.vector(t(L))
write(L_v,file="E:/2021 秋季学期/非参数统计/成绩/列联表.csv")
```

（三）游戏数据

```
library(nortest)#正态性检验包

#导入数据集，初始数据集已按游戏类型分类排序，第二顺序为热度
game1=read.csv("E:/2021 秋季学期/非参数统计/游戏/安卓手机游戏-类型.csv",header=TRUE)
head(game1)
game1<-as.data.frame(game1)
attach(game1)
summary(game1)

#删掉缺失评分信息以及异常评分的数据，按评分排序
game2=read.csv("E:/2021 秋季学期/非参数统计/游戏/安卓手机游戏-评分.csv",header=TRUE)
head(game2)
game2<-as.data.frame(game2)
attach(game2)
summary(game2)

#初始给定数据集，未进行任何处理
game3=read.csv("E:/2021 秋季学期/非参数统计/游戏/安卓手机游戏-热度.csv",header=TRUE)
head(game3)
game3<-as.data.frame(game3)
attach(game3)
summary(game3)

#删掉缺失评分数据后，按游戏类型分类
game4=read.csv("E:/2021 秋季学期/非参数统计/游戏/安卓手机游戏-评分-类型排序.csv",header=TRUE)
head(game4)
game4<-as.data.frame(game4)
attach(game4)
summary(game4)

#正态性检验
par(mfrow=c(1,2))
heat=game1$热度
hist(heat,freq=T,xlab="热度",ylab="游戏数量",main="热度频数分布直方图")
qqnorm(heat);qqline(heat);
ad.test(heat);cvm.test(heat);pearson.test(heat);sf.test(heat);shapiro.test(heat)
rank=game2$评分
hist(rank,freq=T,xlab="评分",ylab="游戏数量",main="评分频数分布直方图")
qqnorm(rank);qqline(rank);
ad.test(rank);cvm.test(rank);pearson.test(rank);sf.test(rank);shapiro.test(rank)
```

```
#各类型游戏热度差异
data_kw1=c(heat[1:90],heat[91:226],heat[227:263],heat[264:291],heat[292:405],
           heat[406:490],heat[491:576],heat[577:629],heat[630:647],heat[648:757],
           heat[758:803],heat[804:815],heat[816:1107],heat[1108:1127],heat[1128:1140])
mydata_kw1<-data.frame(data_kw1,group1=c(rep("策略游戏",90),rep("动作游戏",136),rep("飞行游戏",37),
                                         rep("格斗游戏",28),rep("角色扮演",114),rep("竞速游戏",85),rep("冒险解谜",86),
                                         rep("模拟经营",53),rep("棋牌游戏",18),rep("射击游戏",110),rep("体育运动",46),
                                         rep("养成游戏",12),rep("益智休闲",292),rep("音乐游戏",20),rep("游戏工具",13)))
kruskal.test(data_kw1~group1,data=mydata_kw1)
#任意两类游戏热度差异
#wilcoxon 秩和检验
h=c(0,90,226,263,291,405,490,576,629,647,757,803,815,1107,1127,1140)
wilcox_heat=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
   { wilcox_heat1=wilcox.test(heat[(h[i]+1):h[i+1]],heat[(h[j]+1):h[j+1]],alt="less")
     wilcox_heat2=wilcox.test(heat[(h[i]+1):h[i+1]],heat[(h[j]+1):h[j+1]],alt="greater")
     if(wilcox_heat1$p.value<wilcox_heat2$p.value) {wilcox_heat[i,j]=wilcox_heat1$p.value*(-1)}
     else {wilcox_heat[i,j]=wilcox_heat2$p.value}
   }
}
wilcox_heat
wilcox_heat_v=as.vector(t(wilcox_heat))
write(wilcox_heat_v,file="E:/2021 秋季学期/非参数统计/游戏/任意两类游戏热度 wilcoxon 秩和检验.csv")
#正态记分检验
h=c(0,90,226,263,291,405,490,576,629,647,757,803,815,1107,1127,1140)
l=c(90,136,37,28,114,85,86,53,18,110,46,12,292,20,13)
normal_heat=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
   { x=heat[(h[i]+1):h[i+1]]; y=heat[(h[j]+1):h[j+1]];
     m=l[i]; n=l[j];
     w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
     w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
     T=sum(w[w[,2]==1,4])
     w2=sum(w[,4]^2)
```

```
      S=sqrt(m*n*w2/(m+n-1)/(m+n))
      Z=T/S
      p=pnorm(Z)
      if(p>0.5){p=1-p}
      else {p=p*(-1)}
      normal_heat[i,j]=p
   }
}
normal_heat
normal_heat_v=as.vector(t(normal_heat))
write(normal_heat_v,file="E:/2021 秋季学期/非参数统计/游戏/任意两类游戏热度正态记分检验.csv")


#平方秩检验
#多样本
N=1140;k=15
d2=NULL
d=cbind(heat,c(rep(1,90),rep(2,136),rep(3,37),rep(4,28),rep(5,114),rep(6,85),
                rep(7,86),rep(8,53),rep(9,18),rep(10,110),rep(11,46),rep(12,12),
                rep(13,292),rep(14,20),rep(15,13)))
for(i in 1:k) d2=rbind(d2,cbind(abs(d[d[,2]==i,1]-mean(d[d[,2]==i,1])),d[d[,2]==i,1],i))
d3=cbind(d2[order(d2[,1]),],1:N,(1:N)^2)
Ti=NULL
for(i in 1:k) Ti=c(Ti,sum(d3[d3[,3]==i,5]))
ni=NULL
for(i in 1:k) ni=c(ni,nrow(d3[d3[,3]==i,]))
T=(N-1)*(sum(Ti^2/ni)-sum(Ti)^2/N)/(sum(d3[,5]^2)-sum(Ti)^2/N)
pvalue=pchisq(T,k-1,low=F)
#两样本
h=c(0,90,226,263,291,405,490,576,629,647,757,803,815,1107,1127,1140)
square_heat=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
   { x=heat[(h[i]+1):h[i+1]];y=heat[(h[j]+1):h[j+1]]
      m=length(x);n=length(y);
      x1=abs(x-mean(x));y1=abs(y-mean(y));
      xy1=c(x1,y1);xy0=c(x,y);xyi=c(rep(1,m),rep(2,n));
      xy=cbind(xy1,xy0,xyi)
      xy2=cbind(xy[order(xy[,1]),],1:(m+n),(1:(m+n))^2)
      T1=sum(xy2[xy2[,3]==1,5]);T2=sum(xy2[xy2[,3]==2,5])
      R=xy2[,5];meanR=mean(R);
      S=sqrt(m*n*(sum(R^2)-(m+n)*meanR^2)/(m+n)/(m+n-1))
      Zx=(T1-m*meanR)/S;Zy=(T2-n*meanR)/S
```

```
    if(pnorm(Zx)<pnorm(Zy)) {square_heat[i,j]=pnorm(Zx)*(-1)}
    else{square_heat[i,j]=pnorm(Zy)}
  }
}
square_heat
square_heat_v=as.vector(t(square_heat))
write(square_heat_v,file="E:/2021 秋季学期/非参数统计/游戏/任意两类游戏热度尺度检验平方秩.csv")


#各类型游戏评分差异
#wilcoxon 秩和
rankp=game4$评分
data_kw2=c(rankp[1:71],rankp[72:188],rankp[189:213],rankp[214:238],rankp[239:341],
           rankp[342:410],rankp[411:494],rankp[495:543],rankp[544:557],rankp[558:653],
           rankp[654:692],rankp[693:703],rankp[704:913],rankp[914:928],rankp[929:936])
mydata_kw2<-data.frame(data_kw2,group2=c(rep("策略游戏",71),rep("动作游戏",117),rep("飞行游戏",25),
                                         rep("格斗游戏",25),rep("角色扮演",103),rep("竞速游戏",69),rep("冒险解谜",84),
                                         rep("模拟经营",49),rep("棋牌游戏",14),rep("射击游戏",96),rep("体育运动",39),
                                         rep("养成游戏",11),rep("益智休闲",210),rep("音乐游戏",15),rep("游戏工具",8)))
kruskal.test(data_kw2~group2,data=mydata_kw2)
#任意两类游戏评分差异
#wilcoxon 秩和检验
h=c(0,71,188,213,238,341,410,494,543,557,653,692,703,913,928,936)
wilcox_rank=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
  { wilcox_rank1=wilcox.test(rankp[(h[i]+1):h[i+1]],rankp[(h[j]+1):h[j+1]],alt="less")
    wilcox_rank2=wilcox.test(rankp[(h[i]+1):h[i+1]],rankp[(h[j]+1):h[j+1]],alt="greater")
    if(wilcox_rank1$p.value<wilcox_rank2$p.value) {wilcox_rank[i,j]=wilcox_rank1$p.value*(-1)}
    else {wilcox_rank[i,j]=wilcox_rank2$p.value}
  }
}
wilcox_rank
wilcox_rank_v=as.vector(t(wilcox_rank))
write(wilcox_rank_v,file="E:/2021 秋季学期/非参数统计/游戏/任意两类游戏评分 wilcoxon 秩和检验.csv")
#正态记分检验
h=c(0,71,188,213,238,341,410,494,543,557,653,692,703,913,928,936)
l=c(71,117,25,25,103,69,84,49,14,96,39,11,210,15,8)
```

```
normal_rank=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
   { x=rankp[(h[i]+1):h[i+1]]; y=rankp[(h[j]+1):h[j+1]];
      m=l[i]; n=l[j];
      w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
      w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
      T=sum(w[w[,2]==1,4])
      w2=sum(w[,4]^2)
      S=sqrt(m*n*w2/(m+n-1)/(m+n))
      Z=T/S
      p=pnorm(Z)
      if(p>0.5){p=1-p}
      else {p=p*(-1)}
      normal_rank[i,j]=p
   }
}
normal_rank
normal_rank_v=as.vector(t(normal_rank))
write(normal_rank_v,file="E:/2021 秋季学期/非参数统计/游戏/任意两类游戏评分正态记分检
验.csv")

#平方秩检验
#多样本
N=936;k=15
d2=NULL
d=cbind(rankp,c(rep(1,71),rep(2,117),rep(3,25),rep(4,25),rep(5,103),
                rep(6,69),rep(7,84),rep(8,49),rep(9,14),rep(10,96),
                rep(11,39),rep(12,11),rep(13,210),rep(14,15),rep(15,8)))
for(i in 1:k) d2=rbind(d2,cbind(abs(d[d[,2]==i,1]-mean(d[d[,2]==i,1])),d[d[,2]==i,1],i))
d3=cbind(d2[order(d2[,1]),],1:N,(1:N)^2)
Ti=NULL
for(i in 1:k) Ti=c(Ti,sum(d3[d3[,3]==i,5]))
ni=NULL
for(i in 1:k) ni=c(ni,nrow(d3[d3[,3]==i,]))
T=(N-1)*(sum(Ti^2/ni)-sum(Ti)^2/N)/(sum(d3[,5]^2)-sum(Ti)^2/N)
pvalue=pchisq(T,k-1,low=F)
#两样本
h=c(0,71,188,213,238,341,410,494,543,557,653,692,703,913,928,936)
square_rank=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
   { x=heat[(h[i]+1):h[i+1]];y=heat[(h[j]+1):h[j+1]]
      m=length(x);n=length(y)
```

```
      x1=abs(x-mean(x));y1=abs(y-mean(y))
      xy1=c(x1,y1);xy0=c(x,y);xyi=c(rep(1,m),rep(2,n))
      xy=cbind(xy1,xy0,xyi)
      xy2=cbind(xy[order(xy[,1]),],1:(m+n),(1:(m+n))^2)
      T1=sum(xy2[xy2[,3]==1,5]);T2=sum(xy2[xy2[,3]==2,5])
      R=xy2[,5];meanR=mean(R)
      S=sqrt(m*n*(sum(R^2)-(m+n)*meanR^2)/(m+n)/(m+n-1))
      Zx=(T1-m*meanR)/S;Zy=(T2-n*meanR)/S
      if(pnorm(Zx)<pnorm(Zy)) {square_rank[i,j]=pnorm(Zx)*(-1)}
      else {square_rank[i,j]=pnorm(Zy)}
   }
}
square_rank
square_rank_v=as.vector(t(square_rank))
write(square_rank_v,file="E:/2021 秋季学期/非参数统计/游戏/任意两类游戏评分尺度检验平方秩.csv")



#各类型游戏评论数差异
data_kw2=c(comment[1:90],comment[91:226],comment[227:263],comment[264:291],comment[292:405],

comment[406:490],comment[491:576],comment[577:629],comment[630:647],comment[648:757],

comment[758:803],comment[804:815],comment[816:1107],comment[1108:1127],comment[1128:1140])
mydata_kw2<-data.frame(data_kw2,group2=c(rep("策略游戏",90),rep("动作游戏",136),rep("飞行游戏",37),
                                         rep("格斗游戏",28),rep("角色扮演",114),rep("竞速游戏",85),rep("冒险解谜",86),
                                         rep("模拟经营",53),rep("棋牌游戏",18),rep("射击游戏",110),rep("体育运动",46),
                                         rep("养成游戏",12),rep("益智休闲",292),rep("音乐游戏",20),rep("游戏工具",13)))
kruskal.test(data_kw2~group2,data=mydata_kw2)
with(data=mydata_kw2,pairwise.wilcox.test(x=data_kw2,g=group2,p.adjust.method="bonferroni"))
#尺度检验
data_fk2=list(comment[1:90],comment[91:226],comment[227:263],comment[264:291],comment[292:405],

comment[406:490],comment[491:576],comment[577:629],comment[630:647],comment[648:75
```

```
7],

comment[758:803],comment[804:815],comment[816:1107],comment[1108:1127],comment[112
8:1140])
fligner.test(data_fk2)
#单边方差检验
h=c(0,90,226,263,291,405,490,576,629,647,757,803,815,1107,1127,1140)
a=matrix(0,15,15)
for(i in 1:14)
{ for(j in (i+1):15)
   { a1=ansari.test(game1$评论数[(h[i]+1):h[i+1]],game1$评论数[(h[j]+1):h[j+1]],alt="less")
      a2=ansari.test(game1$评论数[(h[i]+1):h[i+1]],game1$评论数[(h[j]+1):h[j+1]],alt="greater")
      if(a1$p<a2$p) {a[i,j]=a1$p*(-1)}
      else {a[i,j]=a2$p}
   }
}


#相关性分析
#spearman
cor.test(game1$热度,game1$评论数,meth="spearman")
cor.test(game1$热度,game1$评论数,meth="kendall")
cor.test(game1$热度,game1$喜欢数,meth="spearman")
cor.test(game1$热度,game1$喜欢数,meth="kendall")
cor.test(game1$喜欢数,game1$评论数,meth="spearman")
cor.test(game1$喜欢数,game1$评论数,meth="kendall")
cor.test(game4$评分,game4$热度,meth="spearman")
cor.test(game4$评分,game4$热度,meth="kendall")
cor.test(game4$评分,game4$评论数,meth="spearman")
cor.test(game4$评分,game4$评论数,meth="kendall")
cor.test(game4$评分,game4$喜欢数,meth="spearman")
cor.test(game4$评分,game4$喜欢数,meth="kendall")
```

（四）马尔代夫数据

```
###4、马尔代夫
library(nortest)#正态性检验包
library(fBasics)#两样本 ks 检验包
library(RColorBrewer)#颜色数据集


#导入数据集，初始数据集
ma1=read.csv("E:/2021 秋季学期/非参数统计/马尔代夫/马尔代夫旅游.csv",header=TRUE)
head(ma1)
game1<-as.data.frame(ma1)
attach(ma1)
summary(ma1)

#导入数据集，只有三大店铺的数据集
ma2=read.csv("E:/2021 秋季学期/非参数统计/马尔代夫/马尔代夫店铺.csv",header=TRUE)
head(ma2)
game1<-as.data.frame(ma2)
attach(ma2)
summary(ma2)
#三大店铺数据
ma_dy=ma2$价格[1:157]
ma_gl=ma2$价格[158:265]
ma_xc=ma2$价格[266:490]


#导入数据集，所有行程数据集
ma3=read.csv("E:/2021 秋季学期/非参数统计/马尔代夫/马尔代夫所有行程+价格.csv",header=TRUE)
head(ma3)
game1<-as.data.frame(ma3)
attach(ma3)
summary(ma3)
#各行程数据
ma54=na.omit(ma3$价格[1:40])
ma64=na.omit(ma3$价格[41:258])
ma65=na.omit(ma3$价格[259:261])
ma75=na.omit(ma3$价格[262:777])
ma86=na.omit(ma3$价格[778:782])
ma97=na.omit(ma3$价格[783:785])
ma108=na.omit(ma3$价格[786:790])
```

```
#导入数据集，所有月份
ma4=read.csv("E:/2021 秋季学期/非参数统计/马尔代夫/马尔代夫月份.csv",header=TRUE)
head(ma4)
game1<-as.data.frame(ma4)
attach(ma4)
summary(ma4)

#正态性检验
par(mfrow=c(1,2))
#总价格，不满足正态假设
hist(na.omit(ma1$价格),freq=T,xlab="马尔代夫旅游路线价格",ylab="数量",main="价格频数分布直方图")
qqnorm(na.omit(ma1$价格));qqline(na.omit(ma1$价格));
ad.test(na.omit(ma1$价格));cvm.test(na.omit(ma1$价格));pearson.test(na.omit(ma1$价格));
sf.test(na.omit(ma1$价格));shapiro.test(na.omit(ma1$价格))

#各行程之间有无显著差异：5 天 4 晚价格低于其他行程，其他行程之间分布相同
#ks 检验
h=c(0,40,258,261,777,782,785,790)
K=matrix(0,7,7)
W=matrix(0,7,7)
for (i in 1:6)
{ for (j in (i+1):7)
   { x=na.omit(ma3$价格[(h[i]+1):h[i+1]]);y=na.omit(ma3$价格[(h[j]+1):h[j+1]])
      K[i,j]=ks.test(x,y)$p.value
      #对于 KS 检验未通过的继续进行秩和检验
      if (K[i,j]<0.05)
      { w1=wilcox.test(x,y,alt="less")$p.value
         w2=wilcox.test(x,y,alt="greater")$p.value
         if (w1<w2){ W[i,j]=w1*(-1)}
         else { W[i,j]=w2}
      }
   }
}
K
W
#除了 5 天 4 晚之外，其他的行程价格分布无显著差异


#三大店铺价格有无显著差异
ks.test(ma_dy,ma_gl)
ks.test(ma_dy,ma_xc)#有差异
ks.test(ma_gl,ma_xc)
#Brown-Mood
```

```
z=cbind(c(na.omit(ma_dy),na.omit(ma_xc)),
        c(rep(1,length(na.omit(ma_dy))),rep(2,length(na.omit(ma_xc)))))
k=unique(z[,2]);m=median(z[,1]);m1=NULL;m2=NULL
for(i in k){m1=c(m1,sum(z[z[,2]==i,1]>m));m2=c(m2,sum(z[z[,2]==i,1]<=m))}
C=rbind(m1,m2)
fisher.test(C)#0.4555
#Wilcoxon
wilcox.test(na.omit(ma_dy),na.omit(ma_xc))#0.9741
#正态记分检验
x=na.omit(ma_dy); y=na.omit(ma_xc);
m=length(x);n=length(y);
w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
T=sum(w[w[,2]==1,4])
w2=sum(w[,4]^2)
S=sqrt(m*n*w2/(m+n-1)/(m+n))
Z=T/S
p=pnorm(Z)
p#0.5914834

#尺度参数检验
#平方秩
x=na.omit(ma_dy); y=na.omit(ma_xc);
m=length(x);n=length(y);
x1=abs(x-mean(x));y1=abs(y-mean(y));
xy1=c(x1,y1);xy0=c(x,y);xyi=c(rep(1,m),rep(2,n));
xy=cbind(xy1,xy0,xyi)
xy2=cbind(xy[order(xy[,1]),],1:(m+n),(1:(m+n))^2)
T1=sum(xy2[xy2[,3]==1,5]);T2=sum(xy2[xy2[,3]==2,5])
R=xy2[,5];meanR=mean(R);
S=sqrt(m*n*(sum(R^2)-(m+n)*meanR^2)/(m+n)/(m+n-1))
Zx=(T1-m*meanR)/S;Zy=(T2-n*meanR)/S
p=min(pnorm(Zx),pnorm(Zy))
p #5.619099e-06  小于

#Mood
x=na.omit(ma_dy); y=na.omit(ma_xc);
z=cbind(c(x,y),c(rep(1,length(x)),rep(2,length(y))))
m=length(x);n=length(y)
x1=x-median(outer(x,y,"-"))
xy=cbind(c(x1,y),c(rep(1,length(x)),rep(2,length(y))))
N=nrow(xy);xy1=cbind(xy[order(xy[,1]),],1:N)
R1=xy1[xy1[,2]==1,3];M=sum((R1-(N+1)/2)^2)
E1=m*(N^2-1)/12;s=sqrt(m*n*(N+1)*(N^2-4)/180)
```

```
Z=(M-E1)/s
pvalue=pnorm(Z)
pvalue # 4.082611e-05  小于

ansari.test(x,y,alt="less")#0.0001215

#fligner-killeen
fligner.test(c(x,y),c(rep(1,length(x)),rep(2,length(y))))#0.00050275.
#第壹和携程的价格分布有差异并且差异不在于位置在于尺度，携程的价格波动大
#箱线图
data=as.data.frame(cbind(na.omit(ma_dy),na.omit(ma_xc)))
boxplot(data,names=c("第壹假期旅行网","携程自由行"),
          col=c("blue","blue3"),main="两家店铺马尔代夫线路价格")

#每月价格有无显著差异
ma_11=ma4$价格[1:54]
ma_12=ma4$价格[55:603]
ma_1=ma4$价格[604:724]
ma_2=ma4$价格[725:738]
ma_3=ma4$价格[739:789]
ma_4=ma4$价格[790:790]
median(na.omit(ma_11));median(na.omit(ma_12));median(na.omit(ma_1))
median(na.omit(ma_2));median(na.omit(ma_3));ma_4
mean(na.omit(ma_11));mean(na.omit(ma_12));mean(na.omit(ma_1))
mean(na.omit(ma_2));mean(na.omit(ma_3));ma_4

#各月价格差异
data_kw1=c(ma_11,ma_12,ma_1,ma_2,ma_3)
mydata_kw1<-data.frame(data_kw1,group1=c(rep(1,54),rep(2,549),rep(3,121),
                                rep(4,14),rep(5,51)))
kruskal.test(data_kw1~group1,data=mydata_kw1)
#任意两月价格差异
#wilcoxon 秩和检验
h=c(0,54,603,724,738,789)
wilcox_month=matrix(0,5,5)
for(i in 1:4)
{ for(j in (i+1):5)
  { wilcox_month1=wilcox.test(ma4$价格[(h[i]+1):h[i+1]],ma4$价格[(h[j]+1):h[j+1]],alt="less")
    wilcox_month2=wilcox.test(ma4$价格[(h[i]+1):h[i+1]],ma4$价格
[(h[j]+1):h[j+1]],alt="greater")
    if(wilcox_month1$p.value<wilcox_month2$p.value)
{wilcox_month[i,j]=wilcox_month1$p.value*(-1)}
    else {wilcox_month[i,j]=wilcox_month2$p.value}
  }
```

```
}
wilcox_month
wilcox_month_v=as.vector(t(wilcox_month))
write(wilcox_month_v,file="E:/2021 秋季学期/非参数统计/马尔代夫/任意两月价格 wilcoxon 秩和检验.csv")
#正态记分检验
h=c(0,54,603,724,738,789)
l=c(54,549,121,14,51)
normal_month=matrix(0,5,5)
for(i in 1:4)
{ for(j in (i+1):5)
   { x=ma4$价格[(h[i]+1):h[i+1]]; y=ma4$价格[(h[j]+1):h[j+1]];
      m=l[i]; n=l[j];
      w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
      w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
      T=sum(w[w[,2]==1,4])
      w2=sum(w[,4]^2)
      S=sqrt(m*n*w2/(m+n-1)/(m+n))
      Z=T/S
      p=pnorm(Z)
      if(p>0.5){p=1-p}
      else {p=p*(-1)}
      normal_month[i,j]=p
   }
}
normal_month
normal_month_v=as.vector(t(normal_month))
write(normal_month_v,file="E:/2021 秋季学期/非参数统计/马尔代夫/任意两月价格正态记分检验.csv")

#有无早餐对价格的影响
zaoyes=na.omit(ma4[ma4$早餐==1,]$价格)
zaono=na.omit(ma4[ma4$早餐==0,]$价格)
ks.test(zaoyes,zaono)#0.2582
#Brown-Mood
z=cbind(c(zaoyes,zaono),c(rep(1,length(zaoyes)),rep(2,length(zaono))))
k=unique(z[,2]);m=median(z[,1]);m1=NULL;m2=NULL
for(i in k){m1=c(m1,sum(z[z[,2]==i,1]>m));m2=c(m2,sum(z[z[,2]==i,1]<=m))}
C=rbind(m1,m2)
fisher.test(C,alt="less")#0.05896
#Wilcoxon
wilcox.test(zaoyes,zaono,alt="less")#0.1073
#正态记分检验
x=zaoyes; y=zaono;
```

```
m=length(x);n=length(y);
w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
T=sum(w[w[,2]==1,4])
w2=sum(w[,4]^2)
S=sqrt(m*n*w2/(m+n-1)/(m+n))
Z=T/S
p=pnorm(Z)
p#0.1157632

#豪华酒店对价格的影响
zhuyes=na.omit(ma4[ma4$豪华住宿==1,]$价格)
zhuno=na.omit(ma4[ma4$豪华住宿==0,]$价格)
ks.test(zhuyes,zhuno)#5.323e-05
#Brown-Mood
z=cbind(c(zhuyes,zhuno),c(rep(1,length(zhuyes)),rep(2,length(zhuno))))
k=unique(z[,2]);m=median(z[,1]);m1=NULL;m2=NULL
for(i in k){m1=c(m1,sum(z[z[,2]==i,1]>m));m2=c(m2,sum(z[z[,2]==i,1]<=m))}
C=rbind(m1,m2)
fisher.test(C,alt="greater")#1.311e-05
#Wilcoxon
wilcox.test(zhuyes,zhuno,alt="greater")#4.782e-05
#正态记分检验
x=zhuyes; y=zhuno;
m=length(x);n=length(y);
w=cbind(c(x,y),c(rep(1,m),rep(2,n))); w=w[order(w[,1]),]
w=cbind(w,(1:(m+n)),qnorm((1:(m+n))/(m+n+1)))
T=sum(w[w[,2]==1,4])
w2=sum(w[,4]^2)
S=sqrt(m*n*w2/(m+n-1)/(m+n))
Z=T/S
p=pnorm(Z)
1-p#0.0002689024
```