**非参数统计课程设计**

**期末大作业**

|  |  |
| --- | --- |
| **班 级** |  |
| **任课教师** |  |
| **学 号** |  |
| **学生姓名** |  |
| **分 工** |  |
| **递交日期** |  |

**非参数统计课程设计**

[第二章 3](#_Toc135923017)

[第三章 8](#_Toc135923018)

[第四章 10](#_Toc135923019)

[第五章 14](#_Toc135923020)

[第六章 16](#_Toc135923021)

[第七章 21](#_Toc135923022)

[第八章 26](#_Toc135923023)

[第九章 35](#_Toc135923024)

# 第二章

1. 某城市16座预出售的楼盘均价（单位：千元/平米）如下表所示。媒体说均价是每平米37000元，是否与实际相符？请用t检验和符号检验分别检验，你认为哪种检验的结果更加客观。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 36 | 32 | 31 | 25 | 28 | 36 | 40 | 32 |
| 41 | 26 | 35 | 35 | 32 | 87 | 33 | 35 |

解：

设均价为,假设如下:

采用t检验：

**R代码如下:**

prices <- c(36, 32, 31, 25, 28, 36, 40, 32, 41, 26, 35, 35, 32, 87, 33, 35)

res <- t.test(prices, mu=37)

res $ p.value

**运行结果如下：**

[1] 0.889565

输出结果显示p值约为0.89 > 0.05, 对于0.05的显著性水平，该值较大，即不能拒绝原假设，认为该城市的楼盘均价为37000元。

采用符号检验：

**R代码如下：**

prices <- c(36, 32, 31, 25, 28, 36, 40, 32, 41, 26, 35, 35, 32, 87, 33, 35)

res <- binom.test(sum(prices < 37), length(prices))

res $ p.value

**运行结果如下：**

[1] 0.02127075

输出结果显示p值为0.02 < 0.05,对于0.05的显著性水平，该值较小，即拒绝原假设，支持备择假设，认为该市楼盘均价小于37000元。

采用Shapiro-Wilk测试原数据是否服从正态分布，r代码如下:

shapiro.test(prices)

输出p值为8.388e-06远小于0.05，故拒绝原假设，认为该数据不服从正态分布。

所以综上，因t检验假设原数据服从正态分布，而原数据并无明显正态分布形态，故t检验精确度相对更低，且数据样本少，所以认为符号检验更加精确。

2. 某项调查询问了2506个成年人，其中有个问题是“你认为你的孩子将来会过上比你

现在更好的生活、比你现在更差的生活，还是将过上与你一样的生活？”调查结果显示，有1552个人认为更好，有527个人认为更差，有327个人认为一样，还有100个人认为很难说。请问，这个调查结果是否意味多数民众对孩子生活状况看法乐观？

解：

我们可以使用二项式检验来检验孩子生活将更好的比例是否显著大于0.5，设持有乐观看法的人数比例为,假设如下：

**R代码如下：**

res <- binom.test(x = 1552, n = 2506, p = 0.5, alternative = "greater")

res $ p.value

**运行结果如下：**

[1] 2.184429e-33

其中，x表示成功的次数，即认为孩子生活将更好的人数（1552），n表示总样本量（2506），p表示假设的比例（0.5），alternative表示备择假设，这里是“greater”，表示孩子生活将更好的比例显著大于0.5.

P值远小于0.05，故拒绝原假设，接受备择假设，即认为多数民众对孩子的生活状况看法乐观。

3.有人认为下表数据是由中位数为1的总体分布产生的，也有人认为中位数小于1，请你分别用符号检验，Wilcoxon符号秩检验，t检验给出确切答案，并指出哪种检验的答案更加客观。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.848 | 0.413 | 1.058 | 0.858 | 0.350 | 1.582 | 0.372 | 0.566 | 1.097 | 0.939 |
| 0.761 | 0.864 | 0.712 | 0.264 | 0.138 | 1.124 | 0.383 | 1.194 | 2.072 | 0.690 |

解：

**符号检验:**

设中位数为,建立假设组：

先将数据导入：

# 数据

data <- c(0.848, 0.413, 1.058, 0.858, 0.350, 1.582, 0.372, 0.566, 1.097, 0.939,

0.761, 0.864, 0.712, 0.264, 0.138, 1.124, 0.383, 1.194, 2.072, 0.690)

**R代码：**

# 计算符号检验的p值

n\_minus <- sum(data < 1)

res <- binom.test(n\_minus, n, p = 0.5, alternative = "less")

res $ p.value

输出结果：

[1] 0.9793053

P值远大于0.05,故不能拒绝原假设，认为中位数和1无显著差异.

**Wilcoxon符号秩检验：**

设中位数为,建立假设组：

**R代码如下：**

res <- wilcox.test(data, mu = 1, alternative = "less")

res $ p.value

**输出结果：**

[1] 0.01811695

P值小于0.05，故拒绝原假设，接受备择假设，认为中位数小于1.

**t检验:**

**R代码如下：**

res <- t.test(data, mu = 1, alternative = "less")

res $ p.value

输出结果：

[1] 0.04630564

P值小于0.05，拒绝原假设，接受备择假设，认为中位数小于1

最后我们采用Shapiro-Wilk测试原数据是否服从正态分布，代码如下：

shapiro.test(data)

输出的p值为0.1858，大于0.05，故接受原假设，认为该数据较大可能性服从正态分布

综上，我们认为中位数小于1，因t检验假设原数据服从正态分布，而原数据也有正态分布的形态，故在此情况下，t检验和Wilcoxon符号秩检验的p值更小，因此这两种检验的结果更加客观。

4. 一个监听装置收到如下信号，将其转化为二进制代码如下：

0101110011000011111111101001110101010000000010110011101010001001010100000000

请问这个信号是否是纯随机干扰？

**解：**

设假设组：

**代码如下：**

signal <- c(0,1,0,1,1,1,0,0,1,1,0,0,0,0,1,1,1,1,1,1,1,0,1,0,0,1,1,1,0,1,0,1,0,1,0,0,0,0,0,0,1,0,1,1,0,0,1,1,1,0,1,0,1,0,0,1,0,1,0,1,0,0,0,0,0,0)

signal <- factor(signal)

library(tseries)

result <- runs.test(signal, alternative = 'less')

result $ p.value

**输出结果：**

[1] 0.6099593

P值远大于0.05，故接受原假设，认为信号倾向于收到纯随机干扰。

5. 某新书有不少错误，检查了100页，记录其错误个数如下表所示：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 错误数 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |  |
| 含个错误的页数 | 36 | 40 | 19 | 2 | 0 | 2 | 1 | 0 |

请问，是否可以认为一页的错误个数服从泊松分布？

**解：**

设理论频数为

设假设组为：

**R代码如下：**

# 记录每个错误数的频数

f <- c(36, 40, 19, 2, 0, 2, 1, 0)

# 计算样本均值和方差

mu\_hat <- sum(f \* 0:7) / sum(f)

sigma\_hat\_sq <- sum(f \* (0:7 - mu\_hat)^2) / sum(f)

# 计算泊松分布的参数 lambda

lambda <- mu\_hat

# 计算每个错误数的理论频数

expected\_f <- dpois(0:7, lambda) \* sum(f)

# 进行皮尔逊拟合优度检验

res <- chisq.test(f, p = expected\_f / sum(expected\_f))

输出结果：

[1] 4.780474e-05

P值远小于0.05，故拒绝原假设，认为原数据不服从泊松分布。

6. 随机选取某大学三年级学生（200人）某门课的期末考试成绩如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分数 |  |  |  |  |
| 学生数 | 5 | 15 | 30 | 51 |
| 分数 |  |  |  |  |
| 学生数 | 60 | 23 | 10 | 6 |

有人声称，学生成绩服从正态分布。你认为此人观点是否正确？

解：

我们采取绘制直方图和正态概率图来检验，r代码如下：

library(ggplot2)

df <- data.frame(

x = c(25, 35, 45, 55, 65, 75, 85, 95),

freq = c(5, 15, 30, 51, 60, 23, 10, 6)

)

ggplot(df, aes(x, freq)) +

geom\_col(fill = "blue", width = 10) +

scale\_x\_continuous(breaks = seq(20, 100, by = 10)) +

labs(x = "分数", y = "学生数", title = "分数分布直方图")

qqnorm(df$x, main = "正态概率图")

qqline(df$x)

输出结果：

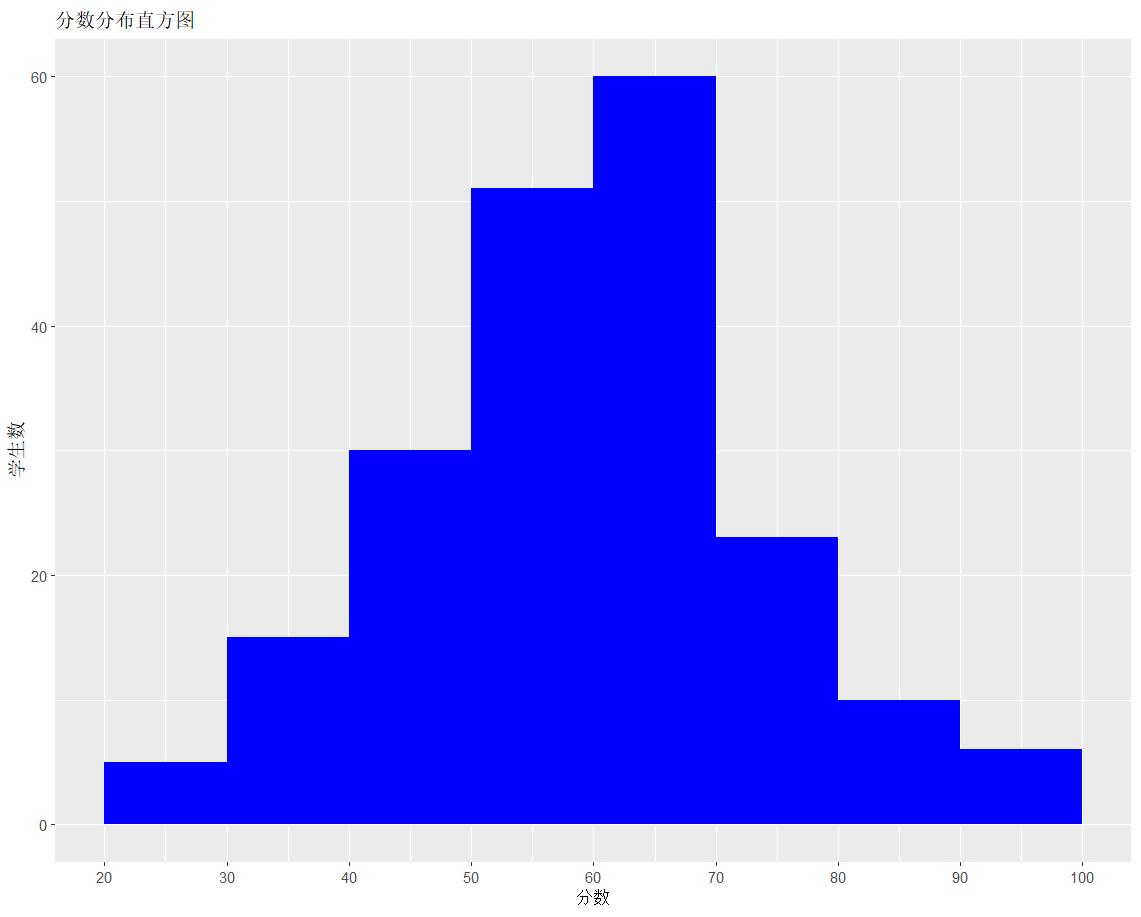


图1：分数分布直方图

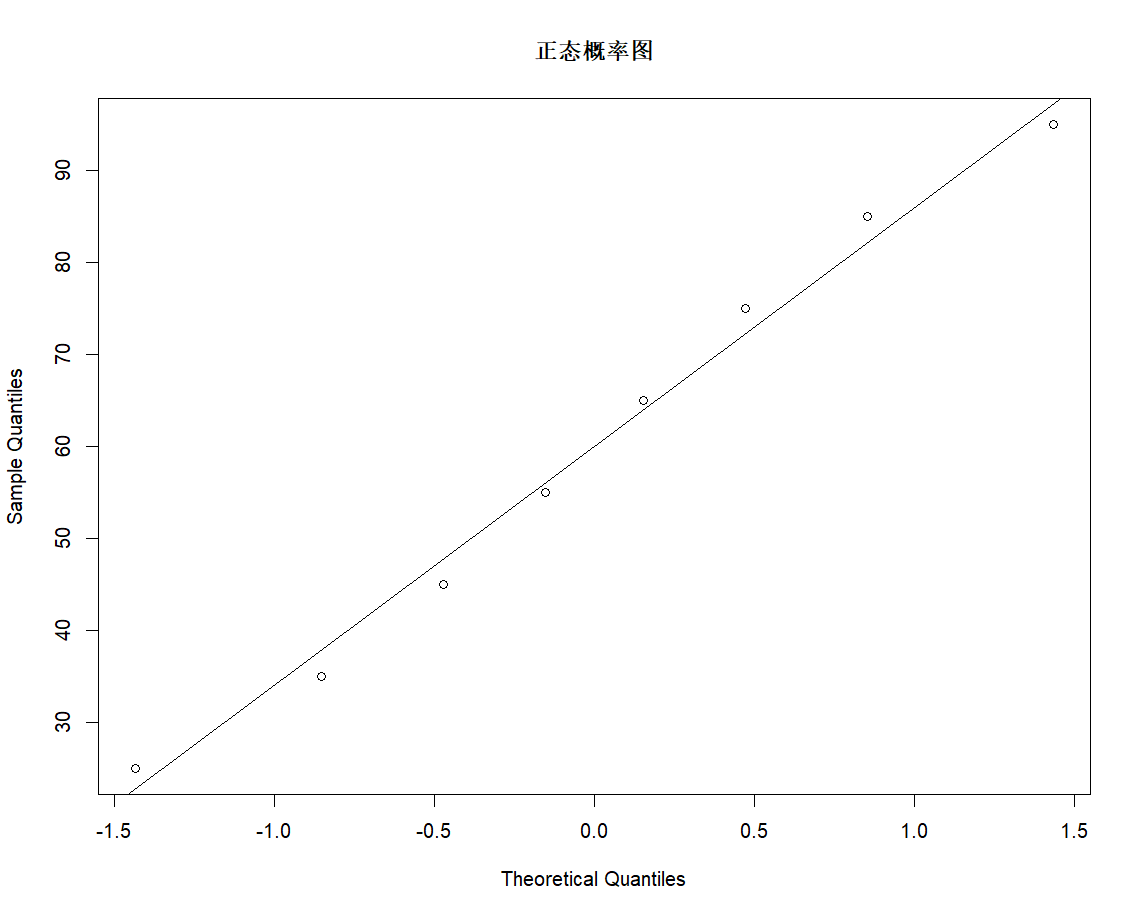


图2：正态概率图

由图一可观察到分数分布直方图呈左偏趋势，图2正态概率图中点也并未落在同一条直线，不符合正态分布的特性，所以学生成绩不服从正态分布。

# 第三章

1. 9个学生到英语培训班学习，培训前后各进行一次水平测试，成绩为

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学生编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 入学前成绩 | 76 | 71 | 70 | 57 | 49 | 69 | 65 | 26 | 59 |
| 入学后成绩 | 81 | 85 | 70 | 52 | 52 | 63 | 83 | 33 | 62 |

学生的培训效果是否显著？请你分别用t检验、符号检验、Wilcoxon符号秩检验回答，并指出哪个检验的结果最合理。

解：

设假设组：

**R代码如下：**

# 输入数据

before <- c(76, 71, 70, 57, 49, 69, 65, 26, 59)

after <- c(81, 85, 70, 52, 52, 63, 83, 33, 62)­­

# t检验

t\_res <- t.test(before, after, paired = TRUE)

t\_res $ p.value­­­­­­­

# 符号检验

b\_res <- binom.test(sum(before < after), length(before)) # before < after 的数量

b\_res $ p.value

# Wilcoxon符号秩检验

w\_res <- wilcox.test(before, after, paired = TRUE)

w\_res $ p.value

**输出结果：**

[1] 0.1400883

[1] 0.5078125

[1] 0.2064626

三个p值都大于0.05，故不能拒绝原假设，即学生的培训效果不显著。由于数据样本较小，非正态分布，且有些学生的成绩变化较大，符号秩检验可能比t检验更适合这个数据集。同时，符号检验和Wilcoxon符号秩检验的结果也更加稳健，不受数据分布的影响。

因此，我们可以选择符号检验或Wilcoxon符号秩检验作为这个数据集的合适模型。

2. 将两批学习情况类似的学生分别在上午和下午进行测试，成绩为

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 上午测试学生成绩 | | | | | 下午测试学生成绩 | | | | |
| 60 | 50 | 85 | 70 | 60 | 25 | 10 | 10 | 45 | 55 |
| 90 | 50 | 50 | 10 | 85 | 70 | 60 | 80 | 95 | 65 |

学生成绩是否受考试时间的影响？请你用Mann-Whitney-Wilcoxon检验和Wald-Wolfowitz游程检验分别回答，并对不同方法的检验结果分别给出恰当的解释。

解：

设分别为,假设组：

**R代码如下：**

morning <- c(60, 50, 85, 70, 60, 90, 50, 50, 10, 85)

afternoon <- c(25, 10, 10, 45, 55, 70, 60, 80, 95, 65)

# Mann-Whitney-Wilcoxon检验

res1 <- wilcox.test(morning, afternoon, alternative = "two.sided")

res1 $ p.value

# 游程检验

res2 <- wilcox.test(morning, afternoon, alternative = "two.sided", exact = FALSE, correct = FALSE)

res2 $ p.value

**输出结果：**

[1] 0.5432197

[1] 0.5183128

两个检验的p值都远大于0.05, 故不能拒绝原假设，认为两组数据的中位数相同。Mann-Whitney-Wilcoxon检验适用于比较两组独立样本的中位数，而Wald-Wolfowitz游程检验适用于比较两个随机变量的分布。而Mann-Whitney-Wilcoxon检验可以不需要假设数据服从特定的分布，也可以处理小样本和缺失数据，对于本题来说更为合适。Wald-Wolfowitz游程检验在样本大小较大时效果较好，所以效果不如Mann-Whitney-Wilcoxon检验效果好。

4. 股票指数的波动程度可以衡量投资的风险，取同一年11月和12月的前10个交易日的股票指数样本数据如下表，请你选择恰当的方法检验两月的波动程度是否相同。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 11月股指 | | | | | 12月股指 | | | | |
| 1149 | 1152 | 1176 | 1149 | 1155 | 1116 | 1130 | 1184 | 1194 | 1184 |
| 1169 | 1182 | 1160 | 1120 | 1171 | 1147 | 1125 | 1125 | 1166 | 1151 |

解：

本题可采用Mann-Whitney-Wilcoxon检验来检验两个月的股票指数波动幅度。设分别为,假设组：

R代码如下：

# 输入数据

nov <- c(1149, 1152, 1176, 1149, 1155, 1169, 1182, 1160, 1120, 1171)

dec <- c(1116, 1130, 1184, 1194, 1184, 1147, 1125, 1125, 1166, 1151)

# Mann-Whitney-Wilcoxon检验

result <- wilcox.test(nov, dec)

# 输出结果

print(result)

输出结果

Wilcoxon rank sum test with continuity correction

data: nov and dec

W = 57, p-value = 0.6228

alternative hypothesis: true location shift is not equal to 0

p值大于0.05，所以无法拒绝原假设，认为两个月的波动幅度相同。

# 第四章

1.大型工厂的工人通常表现出两种行为：精力充沛和疲倦。在某月的一周内，对20名工人的周一、周三和周五进行了测量，数据如下所示（1代表精力充沛，0代表疲劳）。请选择恰当的方法进行分析，并解释结果。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 周一 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 周三 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| 周五 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |

建立假设组： :工人的行为存在显著差异

R语言代码如下：

1. library(RVAideMemoire)
2. data=c(1,1,1,0,1,1,0,0,1,0,1,0,1,0,0,
3. 0,1,1,0,1,1,0,0,1,0,1,1,0,1,0,
4. 1,1,0,1,1,1,0,0,0,1,0,1,0,1,1,
5. 0,1,0,0,0,1,0,1,1,1,0,1,0,1,1)
6. fact=gl(3,1,60,labels=c('星期一','星期三','星期五'))
7. block=gl(20,3,labels=letters[1:20])
8. cochran.qtest(data~fact|block)

运行结果：

1. Cochran's Q test
3. data:  data by fact, block = block
4. Q = 6.7059, df = 2, p-value = 0.03498
5. alternative hypothesis: **true** difference in probabilities is not equal to 0
6. sample estimates:
7. proba in group 星期一 proba in group 星期三 proba in group 星期五
8. 0.30                  0.65                  0.70
10. Pairwise comparisons **using** Wilcoxon sign test
12. 星期一 星期三
13. 星期三 0.1384      -
14. 星期五 0.1157      1
16. P value adjustment method: fdr

输出结果表明p=0.03498,对于显著性水平0.05而言，P值偏小，因此拒绝原假设，及这个调查结果拒绝工人在这三天的的行为没有显著差异的假设。

3. 10个顾客对12种保健品（用A~L的字母表示）进行打分排序，数据结果如下表。

（1）请用Friedman检验及相应的多重比较分析该数据，对结果给出恰当的解释；

（2）请用Kendall协同系数法分析这个数据，对结果给出恰当的解释。并指出这个结果解释与（1）的结果解释的差异所在。

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **顾客** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** |
| **1** | 11 | 7 | 12 | 8 | 3 | 4 | 2 | 9 | 5 | 6 | 1 | 10 |
| **2** | 5 | 9 | 8 | 11 | 2 | 7 | 1 | 12 | 10 | 4 | 6 | 3 |
| **3** | 10 | 2 | 6 | 1 | 7 | 3 | 4 | 5 | 11 | 9 | 12 | 8 |
| **4** | 10 | 6 | 9 | 4 | 8 | 12 | 7 | 3 | 11 | 2 | 1 | 5 |
| **5** | 10 | 7 | 5 | 8 | 9 | 2 | 4 | 1 | 3 | 11 | 12 | 6 |
| **6** | 8 | 3 | 12 | 10 | 11 | 4 | 5 | 7 | 2 | 9 | 1 | 6 |
| **7** | 9 | 8 | 1 | 2 | 10 | 11 | 5 | 7 | 3 | 4 | 12 | 6 |
| **8** | 8 | 2 | 7 | 9 | 5 | 12 | 4 | 10 | 3 | 6 | 11 | 1 |
| **9** | 11 | 7 | 8 | 3 | 5 | 4 | 10 | 2 | 9 | 12 | 6 | 1 |
| **10** | 5 | 9 | 4 | 1 | 8 | 3 | 11 | 12 | 7 | 6 | 2 | 10 |

建立假设组：:顾客对保健品的的打分没有显著差异 :顾客对保健品的打分存在差异

1. R语言代码如下：
2. X1=c(11,5,10,10,10,8,9,8,11,5)
3. X2=c(7,9,2,6,7,3,8,2,7,9)
4. X3=c(12,8,6,9,5,12,1,7,8,4)
5. X4=c(8,11,1,4,8,10,2,9,3,1)
6. X5=c(3,2,7,8,9,11,10,5,5,8)
7. X6=c(4,7,3,12,2,4,11,12,4,3)
8. X7=c(2,1,4,7,4,5,5,4,10,11)
9. X8=c(9,12,5,3,1,7,7,10,2,12)
10. X9=c(5,10,11,11,3,2,3,3,9,7)
11. X10=c(6,4,9,2,11,9,4,6,12,6)
12. X11=c(1,6,12,1,12,1,12,11,6,2)
13. X12=c(10,3,8,5,6,6,6,1,1,10)
14. data1=data.frame(X1,X2,X3,X4,X5,X6,X7,X8,X9,X10,X11,X12)
15. data1=as.matrix(data1)
16. rownames(data1)=letters[1:10]
17. colnames(data1)=c('A','B','C','D','E','F','G','H','I','J','K','L')
18. friedman.test(data1)

输出结果如下：

1. Friedman rank sum test
3. data:  data1
4. Friedman chi-squared = 6.8615, df = 11, p-value = 0.8102

输出结果表明P=0.8102，该值偏大，不能拒绝原假设12种保健品的打分无显著性差异。

基于Nemenyi检验的多重比较

1. library(PMCMRplus)
2. frdManyOneNemenyiTest(data1,p.adjust="BH")

运行结果：

1. A
2. B 0.49
3. C 0.96
4. D 0.36
5. E 0.85
6. F 0.58
7. G 0.23
8. H 0.85
9. I 0.68
10. J 0.89
11. K 0.68
12. L 0.32

可以看出，基于Nemenyi检验的多重比较，不能拒绝原假设。

1. 对Kendall协同系数W进行显著性检验，假设组为：:评价随机 :评价不随机

如果原假设成立则W倾向于小，如果W过大，就要怀疑H0不成立。故P值可以通过计算W的右尾概率得到。具体程序如下：

1. library(DescTools)
2. KendallW(t(data1),correct=T,test=T)

输出结果：

1. Kendall's coefficient of concordance Wt
3. data:  t(data1)
4. Kendall chi-squared = 6.8615, df = 11, subjects = 12, raters = 10, p-value = 0.8102
5. alternative hypothesis: Wt is greater 0
6. sample estimates:
7. Wt
8. 0.06237762

输出结果表明P=0.8102，不能拒绝原假设。

4. 一位经济学家对生产电子计算机设备的企业收集了在一年内生产力提高指数（用0~100内的数表示），并按过去三年间在科研和开发方面的平均花费分为三类：

A：花费少，B：花费中等，C：花费多

生产力提高指数数据如下表。

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类别 | 生产力提高指数 | | | | | | | | | | | |
| A | 7.6 | 8.2 | 6.8 | 5.8 | 6.9 | 6.6 | 6.3 | 7.7 | 6.0 |  |  |  |
| B | 6.7 | 8.1 | 9.4 | 8.6 | 7.8 | 7.7 | 8.9 | 7.9 | 8.3 | 8.7 | 7.1 | 8.4 |
| C | 8.5 | 9.7 | 10.1 | 7.8 | 9.6 | 9.5 |  |  |  |  |  |  |

请问不同类别花费水平下生产力提高指数是否有显著差异？并进行多重比较。

建立假设组：:不同类别花费水平下生产力提高指数无显著差异 :不同类别花费水平下生产力提高指数存在显著差异

首先，将数据存储在R语言中，可以使用以下代码：

1. ```
2. A <- c(7.6, 8.2, 6.8, 5.8, 6.9, 6.6, 6.3, 7.7, 6.0)
3. B <- c(6.7, 8.1, 9.4, 8.6, 7.8, 7.7, 8.9, 7.9, 8.3, 8.7, 7.1, 8.4)
4. C <- c(8.5, 9.7, 10.1, 7.8, 9.6, 9.5)
5. df <- data.frame(cbind(A,B,C))

接下来，进行Kruskal-Wallis检验，代码如下：

kruskal.test(df)

运行结果：

1. Kruskal-Wallis rank sum test
3. data:  df
4. Kruskal-Wallis chi-squared = 20.514, df = 2, p-value = 3.51e-05

输出结果表明统计量值为20.514，P值为3.15e-0.5，小于显著性水平0.05，拒绝原假设。表明不同类别花费水平下生产力提高指数存在显著差异。

进行多重比较，R程序如下:

1. A <- c(7.6, 8.2, 6.8, 5.8, 6.9, 6.6, 6.3, 7.7, 6.0)
2. B <- c(6.7, 8.1, 9.4, 8.6, 7.8, 7.7, 8.9, 7.9, 8.3, 8.7, 7.1, 8.4)
3. C <- c(8.5, 9.7, 10.1, 7.8, 9.6, 9.5)
5. # 将数据组合成一个数据框
6. data <- data.frame(Category = rep(c("A", "B", "C"), times = c(length(A), length(B), length(C))),
7. Score = c(A, B, C))
8. model <- aov(Score ~ Category, data = data)
9. TukeyHSD(model)e = 3.51e-05

输出结果为：

1. Tukey multiple comparisons of means
2. 95% family-wise confidence level
4. Fit: aov(formula = Score ~ Category, data = data)
6. $Category
7. diff        lwr      upr     p adj
8. B-A 1.255556 0.37453174 2.136579 0.0043755
9. C-A 2.322222 1.26919735 3.375247 0.0000335
10. C-B 1.066667 0.06767956 2.065654 0.0347870

从表中可知，调整的P值都小于0.05，故拒绝原假设。及不同类别花费水平下生产力提高指数无显著差异。

# 第五章

1. 现在想研究体重和肺活量的关系，调查了某地10位初中女生的体重和肺活量，数据如表所示，

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 编号 | 体重（斤） | 肺活量 | 编号 | 体重（斤） | 肺活量 |
| 1 | 75 | 2.62 | 6 | 68 | 1.98 |
| 2 | 95 | 2.91 | 7 | 60 | 2.04 |
| 3 | 85 | 2.94 | 8 | 66 | 2.20 |
| 4 | 70 | 2.11 | 9 | 80 | 2.65 |
| 5 | 76 | 2.17 | 10 | 88 | 2.69 |

请问体重和肺活量有什么关系？请用斯皮尔曼秩相关和肯德尔秩相关分析，给出解释。

R语言代码如下：

1. u<-c(75,95,85,70,76,68,60,66,80,88)
2. v<-c(2.62,2.91,2.94,2.11,2.17,1.98,2.04,2.20,2.65,2.69)
3. cor.test(u, v, method = "spearman")

运行结果：

1. Spearman's rank correlation rho
3. data:  u and v
4. S = 26, p-value = 0.004459
5. alternative hypothesis: **true** rho is not equal to 0
6. sample estimates:
7. rho
8. 0.8424242

输出结果表明 R=0.8424242>0.8,体重和肺活量的相关程度较高。

cor.test(u, v, method = "kendall")

运行结果：

1. Kendall's rank correlation tau
3. data:  u and v
4. T = 38, p-value = 0.004687
5. alternative hypothesis: **true** tau is not equal to 0
6. sample estimates:
7. tau
8. 0.6888889

输出结果表明T=0.6888889，体重和肺活量的相关程度不够高。

斯皮尔曼秩相关和肯德尔秩相关都表明，体重与肺活量之间存在一定的正相关关系。p值小于0.05，两者的判断都是在95%的显著水平上，说明体重和肺活量之间的关系确实存在显著性。

3. 设，令，其中，X和Y是一个非线性关系（抛物线）。每次产生100个随机数，检验X和Y之间的相依性。模拟500遍，分别计算皮尔逊相关性检验、斯皮尔曼秩相关检验、肯德尔秩相关检验、霍夫丁独立性检验的平均p值。

（1）取，比较结果与例5.10的不同之处，并解释原因；

（2）取，比较结果与例5.10以及（1）的不同之处，并解释原因。

（1）程序如下：

1. library(Hmisc)
2. set.seed(12345)
3. m=500;n=100
4. pv=matrix(nrow=4,ncol=m)
5. **for**(i in 1:m){
6. x=runif(n,0,1)
7. e=rnorm(n,0,0.1)
8. y=x^2+e
9. p1=cor.test(x,y,method='pearson')$p.value
10. p2=cor.test(x,y,method='spearman')$.value
11. p3=cor.test(x,y,method='kendall')$.value
12. p4=hoeffd(x,y)$P[1,2]
13. pv[,i]=c(p1,p2,p3,p4)
14. }
15. apply(pv,1,mean)
16. apply(pv,1,sd)

运行结果：

1. > apply(pv,1,mean)
2. [1] 1.188006e-34 1.000000e-08 1.188006e-34 1.000000e-08
3. > apply(pv,1,sd)
4. [1] 1.820775e-33 0.000000e+00 1.820775e-33 0.000000e+00

结果表明，皮尔逊相关检验、斯皮尔曼秩相关检验、肯德尔秩相关检验均拒绝X和Y不相关的原假设，即它们之间存在抛物线型的非线性关系。这三种方法在这个区间内存在单调性关系，因此也拒绝原假设。

1. 程序如下：
2. library(Hmisc)
3. set.seed(12345)
4. m=500;n=100
5. pv=matrix(nrow=4,ncol=m)
6. **for**(i in 1:m){
7. x=runif(n,-0.5,1)
8. e=rnorm(n,0,0.1)
9. y=x^2+e
10. p1=cor.test(x,y,method='pearson')$p.value
11. p2=cor.test(x,y,method='spearman')$.value
12. p3=cor.test(x,y,method='kendall')$.value
13. p4=hoeffd(x,y)$P[1,2]
14. pv[,i]=c(p1,p2,p3,p4)
15. }
16. apply(pv,1,mean)
17. apply(pv,1,sd)

运行结果：

1. > apply(pv,1,mean)
2. [1] 8.527463e-13 1.000000e-08 8.527463e-13 1.000000e-08
3. > apply(pv,1,sd)
4. [1] 1.397968e-11 0.000000e+00 1.397968e-11 0.000000e+00

代码中生成 x 的随机数范围是 (0,1)，而上面代码中的范围是 (-0.5,1)。这一差别导致数据具有不同的性质，从而会对假设检验和相关性分析的结果产生影响。

# 第六章

5. 在一次社会调查中，以问卷方式共调查901人的月收入及对工作的满意程度。其中收入A分为小于3000元，3000~7500元，7500~12000元，12000元以上4档。满意度B分为很不满意、较不满意、基本满意、很满意4档。调查结果用如下两个变量4×4列联表表示。请问，基于该数据，工资收入与满意度是否有关系？

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 工资收入 | 很不满意 | 较不满意 | 基本满意 | 很满意 |
| <3000 | 20 | 24 | 80 | 82 |
| 3000~7500 | 22 | 38 | 104 | 125 |
| 7500~12000 | 13 | 28 | 81 | 113 |
| >12000 | 7 | 18 | 54 | 92 |

（1）请用独立性检验回答问题；

（2）如果强行用齐性检验回答问题，结果应该如何解释？

解：

(1)建立假设组：

**R代码如下：**

income <- c("<3000", "3000~7500", "7500~12000", ">12000")

satisfaction <- c("很不满意", "较不满意", "基本满意", "很满意")

data\_matrix <- matrix(c(20, 24, 80, 82, 22, 38, 104, 125, 13, 28, 81, 113, 7, 18, 54, 92), nrow=4, ncol=4, byrow=TRUE, dimnames=list(income, satisfaction))

result <- chisq.test(data\_matrix)

result

**输出结果如下：**

Pearson's Chi-squared test

data: data\_matrix

X-squared = 11.989, df = 9, p-value = 0.214

可以看出，故接受原假设，认为工资与满意度之间存在关系。

(2) 如果强行使用卡方齐性检验，有以下代码：

result\_homogeneity <- chisq.test(data\_matrix, correct=FALSE)

result\_homogeneity

**输出结果如下：**

Pearson's Chi-squared test

data: data\_matrix

X-squared = 11.989, df = 9, p-value = 0.214

发现结果仍然与独立性检验相同。这是因为该数据集的设计目的是检验两个变量之间的关系，而不是比较同一变量在不同条件下的分布。因此，我们应该使用独立性检验来回答问题。

6. 简·奥斯汀（1775-1817）是英国著名女作家，在其短暂的一生中为世界文坛奉献出了许多经久不衰的作品，如《理智与情感》，《傲慢与偏见》，《爱玛》等。在其身后，她的哥哥亨利主持了遗作《劝导》和《诺桑觉寺》两部作品的出版。这两部是否出自奥斯汀本人？很多奥斯汀的文学爱好者进行了自发的研究。下面收集了《理智与情感》、《爱玛》以及遗作《劝导》的前两章（分别记为劝导1和劝导2）中常用代表词的出现频数。希望研究常用词汇的用词频率，从而鉴别真伪。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 单词 | 理智与情感 | 爱玛 | 劝导1 | 劝导2 |
| a | 147 | 186 | 101 | 83 |
| an | 25 | 26 | 11 | 29 |
| this | 32 | 39 | 15 | 15 |
| that | 94 | 105 | 37 | 22 |
| with | 59 | 74 | 28 | 43 |
| without | 18 | 10 | 10 | 4 |

（1）请用齐性检验回答问题；

（2）如果强行使用独立性检验回答问题，结果应如何解释？

解：

1. 建立假设组：

R代码如下：

# 创建频数矩阵

freq\_matrix <- matrix(c(147, 186, 101, 83, 25, 26, 11, 29, 32, 39, 15, 15, 94, 105, 37, 22, 59, 74, 28, 43, 18, 10, 10, 4), nrow = 6, byrow = TRUE)

# 计算期望频数

row\_sums <- apply(freq\_matrix, 1, sum)

col\_sums <- apply(freq\_matrix, 2, sum)

total\_sum <- sum(freq\_matrix)

expected\_freq <- outer(row\_sums, col\_sums) / total\_sum

# 计算卡方值和p值

chisq <- sum((freq\_matrix - expected\_freq)^2 / expected\_freq)

pval <- 1 - pchisq(chisq, df = 1)

# 输出结果

cat("卡方值：", chisq, "\n")

cat("p值：", pval, "\n")

**输出结果：**

卡方值： 45.57751

p值： 1.46716e-11

可见卡方值为45.57751，p值远小于0.05，故拒绝原假设，认为，即认为两本书不是出自奥斯汀本人。

(2) 如果强行使用独立性检验回答问题，在这个问题中，我们可以将每个单词作为一个变量，将四个文本作为两个分类变量。

**代码如下：**

# 将数据转换为列联表

freq\_matrix <- matrix(c(147, 186, 101, 83, 25, 26, 11, 29, 32, 39, 15, 15, 94, 105, 37, 22, 59, 74, 28, 43, 18, 10, 10, 4), nrow = 6, byrow = TRUE)

rownames(freq\_matrix) <- c("A", "B", "C", "D", "E", "F")

# 进行独立性检验

result <- chisq.test(freq\_matrix)

# 输出检验结果

print(result)

**输出结果如下：**

Pearson's Chi-squared test

data: freq\_matrix

X-squared = 45.578, df = 15, p-value = 6.205e-05

可见p值为6.205e-05小于0.05，故拒绝原假设，认为，即认为两本书不是出自奥斯汀本人。

在这个问题中，我们需要检验的是四个文本之间单词使用的频率是否存在差异，因此，独立性检验是更适合这个问题的检验方法。

7. 对第6题数据进行Lambda相关分析和Goodman-Kruskal Tau相关分析。

解：

我们想要知道书的种类和单词出现频率的相关程度，计算Lambda值，

**r代码如下：**

# 创建数据框

data <- data.frame(

col1 = c(147, 25, 32, 94, 59, 18),

col2 = c(186, 26, 39, 105, 74, 10),

col3 = c(101, 11, 15, 37, 28, 10),

col4 = c(83, 29, 15, 22, 43, 4)

)

# 计算Lambda值

lambda = function(dat) {

my = max(apply(dat, 1, sum))

sumy = sum(apply(dat, 2, max))

n = sum(dat)

lam = (sumy - my) / (n - my)

return (lam)

}

lambda(data)

**输出结果如下：**

[1] 0

输出结果，说明书的种类和单词出现频率完全无关。

若采用Goodman-Kruskal Tau相关测量法计算，

**R代码如下：**

# 创建数据框

data <- data.frame(

col1 = c(147, 25, 32, 94, 59, 18),

col2 = c(186, 26, 39, 105, 74, 10),

col3 = c(101, 11, 15, 37, 28, 10),

col4 = c(83, 29, 15, 22, 43, 4)

)

# 计算tau值

tau = function(dat) {

n = sum(dat)

fy = apply(dat, 1, sum)

E1 = sum(fy \* (1 - fy / n))

fun = function(x) {sum(x \* (1 - x / sum(x)))}

E2 = sum(apply(dat, 2, fun))

return ((E1 - E2) / E1)

}

tau(data)

**计算结果如下：**

[1] 0.00765707

发现，非常接近0，这说明用不同的书籍来分辨单词词频是不合适的。

8. 对第5题数据进行Gamma相关分析和Somer's d相关分析。

解：

**Gamma相关分析r代码**：

# 创建一个数据框

income <- data.frame(

Very\_Dissatisfied = c(20, 22, 13, 7),

Somewhat\_Dissatisfied = c(24, 38, 28, 18),

Somewhat\_Satisfied = c(80, 104, 81, 54),

Very\_Satisfied = c(82, 125, 113, 92)

)

gam\_cor = function(dat) {

ns = 0

nd = 0

r = nrow(dat);c = ncol(dat)

for (i in 1:(r - 1)) {

for (j in 1:(c - 1)) {

ns = ns + dat[i, j] \* sum(dat[(i + 1):r, (j + 1):c])

nd = nd + dat[i, j + 1] \* sum(dat[(i + 1):r, 1:j])

}

}

names(ns) = NULL; names(nd) = NULL

g = (ns - nd) / (ns + nd)

res = list('G' = g, 'ns' = ns, 'nd' = nd)

return (res)

}

# 打印结果

gamma <- gam\_cor(income)

gamma

输出结果：

$G

[1] 0.1265461

$ns

[1] 109520

$nd

[1] 84915

可以算出,工资水平与满意度呈正比关系，以工资满意度来解释工资收入可以消减约12.65％的误差。

**Somer's d相关分析r代码：**

# 创建一个数据框

income <- data.frame(

Very\_Dissatisfied = c(20, 24, 80, 82),

Somewhat\_Dissatisfied = c(22, 38, 104, 125),

Somewhat\_Satisfied = c(13, 28, 81, 113),

Very\_Satisfied = c(7, 18, 53, 92)

)

somer = function(dat) {

ns = 0

nd = 0

ty = 0

r = nrow(dat);c = ncol(dat)

for (i in 1:(r - 1)) {

for (j in 1:(c - 1)) {

ns = ns + dat[i, j] \* sum(dat[(i + 1):r, (j + 1):c])

nd = nd + dat[i, j + 1] \* sum(dat[(i + 1):r, 1:j])

}

}

for (i in 1:r ) {

for (j in 1:(c - 1)) {

ty = ty + dat[i, j] \* sum(dat[i, (j + 1):c])

}

}

d = (ns - nd) / (nd + ns + ty)

res = list('ns' = ns, 'nd' = nd, 'ty' = ty, 'd' = d);

return (res)

}

# 打印结果

res <- somer(income)

res $d

**输出结果：**

[1] 0.08261129

，说明工资收入和工资满意度成正比相关，以工资满意度来解释工资收入可以消减约8.26％的误差。

# 第七章

3. 某种疾病有两种治疗方法，一是按病因服标准剂量的药，二是按症状服药，治疗效果和性别没有关系，但可能与年龄有关系。为此，分20岁到35岁，36岁到50岁两个年龄组，治疗效果如下表所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 年龄 | 治疗效果 | 治疗方法 | |
| 按病因服标准剂量的药 | 按症状服药 |
| 20-35 | 迅速痊愈 | 22 | 19 |
| 缓慢痊愈 | 16 | 17 |
| 未痊愈 | 2 | 4 |
| 36-50 | 迅速痊愈 | 19 | 15 |
| 缓慢痊愈 | 61 | 72 |
| 未痊愈 | 20 | 13 |

（1）用检验和似然比检验分析该数据的局部独立性和条件独立性；

（2）用对数线性模型分析本题数据，对结果要给出合理解释。

（1）检验治疗效果与年龄和治疗方法是否有关，建立假设组：:=× :≠×

程序如下:

1. dat=c(22,19,16,61,2,20,19,15,17,72,4,13)
2. dat=array(dat,dim=c(2,3,2),dimnames=list(
3. c('20-35','36-50'),c('迅速痊愈','缓慢痊愈','未痊愈'),
4. c('按病因服标准剂量的药','按症状服药')
5. ))
6. #卡方局部独立性检验
7. Q=loglin(dat,list(1,c(3,2)))$pearson#Q值
8. df=loglin(dat,list(1,c(3,2)))$df#自由度
9. p=pchisq(Q,df,,lower.tail=F)#p值
10. result=list('Q.value'=Q,'p.value'=p);result

运行结果：

1. $Q.value
2. [1] 35.57519
4. $p.value
5. [1] 1.154967e-06

输出表明Q=35.57519，相应的P=1.154967e-06。在0.05的显著性水平下，拒绝原假设。

给定假设组:

:给定年龄，治疗效果和治疗方法条件不独立

相应程序如下：

1. dat=c(22,19,16,61,2,20,19,15,17,72,4,13)
2. dat=array(dat,dim=c(2,3,2),dimnames=list(
3. c('20-35','36-50'),c('迅速痊愈','缓慢痊愈','未痊愈'),
4. c('按病因服标准剂量的药','按症状服药')
5. ))
6. Q=loglin(dat,list(c(1,2),c(1,3)))$pearson#Q值
7. df=loglin(dat,list(c(1,2),c(1,3)))$df#自由度
8. p=pchisq(Q,df,,lower.tail=F)#p值
9. result=list('Q.value'=Q,'p.value'=p);result

运行结果：

1. $Q.value
2. [1] 3.781693
4. $p.value
5. [1] 0.4363559

输出结果表明，Q=3.781693，P=0.4363599，在0.05的显著性水平下，无法拒绝原假设。

1. 模型各效应的估计程序如下:
2. dat=c(22,19,16,61,2,20,19,15,17,72,4,13)
3. #以A,B,C分别表示年龄，治疗效果，治疗方法
4. dat75=array(dat,dim=c(2,3,2),dimnames=list(
5. A=c('20-35','36-50'),B=c('迅速痊愈','缓慢痊愈','未痊愈'),
6. C=c('按病因服标准剂量的药','按症状服药')
7. ))
8. library(MASS)
9. m75=loglm(~A+B+C+B:C,data=dat75)
10. m75$param

运行结果:

1. $`(Intercept)`
2. [1] 2.871487
4. $A
5. 20-35      36-50
6. -0.4581454  0.4581454
8. $B
9. 迅速痊愈    缓慢痊愈      未痊愈
10. -0.04613873  0.75011587 -0.70397713
12. $C
13. 按病因服标准剂量的药           按症状服药
14. 0.05003495          -0.05003495
16. $B.C
17. C
18. B          按病因服标准剂量的药  按症状服药
19. 迅速痊愈           0.04357082 -0.04357082
20. 缓慢痊愈          -0.12245042  0.12245042
21. 未痊愈             0.07887960 -0.07887960

饱和模型各效应的估计程序如下:

1. sat=loglm(~A\*B\*C,data=dat75)
2. sat$param

运行结果：

1. $`(Intercept)`
2. [1] 2.776786
4. $A
5. 20-35      36-50
6. -0.4899988  0.4899988
8. $B
9. 迅速痊愈   缓慢痊愈     未痊愈
10. 0.1452063  0.7215492 -0.8667555
12. $C
13. 按病因服标准剂量的药           按症状服药
14. -0.008815741          0.008815741
16. $A.B
17. B
18. A         迅速痊愈   缓慢痊愈     未痊愈
19. 20-35  0.5857469 -0.2054357 -0.3803112
20. 36-50 -0.5857469  0.2054357  0.3803112
22. $A.C
23. C
24. A       按病因服标准剂量的药  按症状服药
25. 20-35          -0.09237898  0.09237898
26. 36-50           0.09237898 -0.09237898
28. $B.C
29. C
30. B          按病因服标准剂量的药  按症状服药
31. 迅速痊愈           0.10456380 -0.10456380
32. 缓慢痊愈          -0.04778848  0.04778848
33. 未痊愈            -0.05677533  0.05677533
35. $A.B.C
36. , , C = 按病因服标准剂量的药
38. B
39. A          迅速痊愈   缓慢痊愈     未痊愈
40. 20-35  0.06993265  0.1186709 -0.1886035
41. 36-50 -0.06993265 -0.1186709  0.1886035
43. , , C = 按症状服药
45. B
46. A          迅速痊愈   缓慢痊愈     未痊愈
47. 20-35 -0.06993265 -0.1186709  0.1886035
48. 36-50  0.06993265  0.1186709 -0.1886035

进行拟合优度检验，程序如下:

1. dat=c(22,19,16,61,2,20,19,15,17,72,4,13)
2. dat76=array(dat,dim=c(2,3,2),dimnames=list(
3. A=c('20-35','36-50'),B=c('迅速痊愈','缓慢痊愈','未痊愈'),
4. C=c('按病因服标准剂量的药','按症状服药')
5. ))
6. models=list(list(3,c(1,2)),list(2,c(1,3)),list(1,c(2,3)),list(c(1,3),c(2,3)),
7. list(c(1,2),c(2,3)),list(c(1,2),c(1,3)),list(1,2,3))
8. names=c('自由度','卡方值','卡方P值','似然比值','似然比P值')
9. out=matrix(nrow=length(models),ncol=length(names))
10. colnames(out)=names
11. **for**(i in 1:nrow(out)){
12. log1=loglin(dat76,models[[i]])
13. out[i,1]=log1$df
14. out[i,2]=log1$pearson
15. out[i,3]=pchisq(out[i,2],out[i,1],lower.tail=F)
16. out[i,4]=log1$lrt
17. out[i,5]=pchisq(out[i,4],out[i,1],lower.tail=F)
18. }
19. out

运行结果：

1. 自由度    卡方值      卡方P值  似然比值    似然比P值
2. [1,]      5  3.781693 5.812551e-01  3.808296 5.773336e-01
3. [2,]      6 38.371773 9.501794e-07 36.456699 2.246753e-06
4. [3,]      5 35.575192 1.154967e-06 34.291394 2.083189e-06
5. [4,]      4 35.575192 3.538478e-07 34.291394 6.493699e-07
6. [5,]      3  1.645811 6.490477e-01  1.642990 6.496818e-01
7. [6,]      4  3.781693 4.363559e-01  3.808296 4.325714e-01
8. [7,]      7 38.371773 2.574807e-06 36.456699 5.943482e-06

第一个模型拟合结果的卡方P值约为0.58，似然比P值约为0.58，说明该模型与实际观测数据的拟合程度较好。

第二个模型、第三个模型和第七个模型的卡方P值和似然比P值均小于0.01，说明这三个模型与实际观测数据的拟合程度很差，需要优化或修改模型。

第四个模型的卡方P值和似然比P值小于0.05但大于0.01，说明该模型与实际观测数据的拟合程度比其他模型更好，但还需要进一步检验或优化。

第五个模型的卡方P值约为0.65，似然比P值约为0.65，说明该模型与实际观测数据的拟合程度好，但需要更多的检验和优化。

4. 某汽车保险公司某一年有12299份保单，有赔款记录的保单和无赔款记录的保单按照车辆类型和被保险人年龄分别进行统计，统计结果如下表：

|  |  |  |  |
| --- | --- | --- | --- |
| 车辆类型 | 被保险人年龄 | 保单数 | |
| 有赔款记录保单数 | 无赔款记录保单数 |
| 普通车 |  | 741 | 2829 |
|  | 882 | 4945 |
| 高性能车 |  | 453 | 1169 |
|  | 248 | 1032 |

请问，年龄小的被保险人是否更容易有赔款记录？用本章方法进行分析回答。

程序代码如下：

1. dat=c(741,453,882,248,2829,1169,4945,1032)
2. dat=array(dat,dim=c(2,2,2),dimnames=list(
3. c('普通车','高性能车'),c('<25','>=25'),
4. c('有赔款记录保单数','无赔款记录保单数')
5. ))
6. # 拟合对数线性模型
7. fit <- glm(dat[,1,1] ~ dat[,1,2] + dat[,2,1] + dat[,2,2], family = "poisson")
9. # 输出模型摘要
10. summary(fit)

运行结果：

1. Call:
2. glm(formula = dat[, 1, 1] ~ dat[, 1, 2] + dat[, 2, 1] + dat[,
3. 2, 2], family = "poisson")
5. Deviance Residuals:
6. [1]  0  0
8. Coefficients: (2 not defined because of singularities)
9. Estimate Std. Error z value Pr(>|z|)
10. (Intercept) 5.769e+00  8.415e-02  68.563   <2e-16 \*\*\*
11. dat[, 1, 2] 2.965e-04  3.593e-05   8.251   <2e-16 \*\*\*
12. dat[, 2, 1]        NA         NA      NA       NA
13. dat[, 2, 2]        NA         NA      NA       NA
14. ---
15. Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
17. (Dispersion parameter **for** poisson family taken to be 1)
19. Null deviance: 7.0157e+01  on 1  degrees of freedom
20. Residual deviance: 9.8144e-14  on 0  degrees of freedom
21. AIC: 20.4
23. Number of Fisher Scoring iterations: 2

模型摘要中的coefficients部分显示了每个自变量的估计系数和显著性测试的p值。这些结果表明：被保险人年龄小于25岁的人和无赔款记录保单数对有赔款记录保单数有显著贡献。

高性能车对有赔款记录保单数的发生没有显著贡献。

# 第八章

1. 26名神经功能受损儿童接受了两组（A组和B组）空间知觉测试，得分数据如下表所示，

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| A组 | 48 | 36 | 20 | 29 | 42 | 42 | 20 | 42 | 22 | 41 | 45 | 14 | 6 |
| B组 | 42 | 33 | 16 | 39 | 38 | 36 | 15 | 33 | 20 | 43 | 34 | 22 | 7 |
| 编号 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| A组 | 0 | 33 | 28 | 34 | 4 | 32 | 24 | 47 | 41 | 24 | 26 | 30 | 41 |
| B组 | 15 | 34 | 29 | 41 | 13 | 38 | 25 | 27 | 41 | 28 | 14 | 28 | 40 |

（1）请给出A组和B组得分的Pearson相关系数的Bootstrap点估计、区间估计（置信度为95%）；

（2）以B为因变量，A为自变量，建立一元线性回归模型：。请给出的Bootstrap估计和假设检验（原假设取）。

解:

1. R代码如下：

A <- c(48, 36, 20, 29, 42, 42, 20, 42, 22, 41, 45, 14, 6, 0, 33, 28, 34, 4, 32, 24, 47, 41, 24, 26, 30, 41)

B <- c(42, 33, 16, 39, 38, 36, 15, 33, 20, 43, 34, 22, 7, 15, 34, 29, 41, 13, 38, 25, 27, 41, 28, 14, 28, 40)

library(boot) # 加载boot包

# 定义Pearson相关系数的函数

cor\_func <- function(data, indices) {

x <- data[indices, 1]

y <- data[indices, 2]

return(cor(x, y))

}

# 生成1000个Bootstrap样本，并计算Pearson相关系数

set.seed(123) # 设置随机数种子，以便结果可重复

boot\_results <- boot(data = cbind(A, B), statistic = cor\_func, R = 1000)

boot\_results

boot\_ci <- boot.ci(boot\_results, type = "basic", conf = 0.95)

boot\_ci

输出结果如下：

Bootstrap Statistics :

original bias std. error

t1\* 0.8209092 0.00149779 0.06297062

Intervals :

Level Basic

95% ( 0.7219, 0.9634 )

Calculations and Intervals on Original Scale

输出结果显示，A组和B组得分的Pearson相关系数的点估计为0.8209092，Bootstrap估计的标准误差为0.06297062。注意到，Bootstrap估计的偏差为0.00149779，这表明原始估计值可能存在一定的偏差。

1. **R代码如下：**

A <- c(48, 36, 20, 29, 42, 42, 20, 42, 22, 41, 45, 14, 6, 0, 33, 28, 34, 4, 32, 24, 47, 41, 24, 26, 30, 41)

B <- c(42, 33, 16, 39, 38, 36, 15, 33, 20, 43, 34, 22, 7, 15, 34, 29, 41, 13, 38, 25, 27, 41, 28, 14, 28, 40)

lm.fit <- function(data, indices) {

fit <- lm(B ~ A, data = data[indices,])

return(coef(fit))

}

library(boot)

set.seed(123)

boot.out <- boot(data = data.frame(A, B), statistic = lm.fit, R = 1000)

boot.ci(boot.out, type = "bca")

summary(lm(B ~ A, data = data.frame(A, B)))

**输出结果：**

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 1000 bootstrap replicates

CALL :

boot.ci(boot.out = boot.out, type = "bca")

Intervals :

Level BCa

95% ( 2.779, 14.168 )

Calculations and Intervals on Original Scale

Call:

lm(formula = B ~ A, data = data.frame(A, B))

Residuals:

Min 1Q Median 3Q Max

-13.2570 -3.9509 0.9777 3.6021 10.5441

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.4431 3.0179 3.129 0.00456 \*\*

A 0.6556 0.0931 7.042 2.79e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.217 on 24 degrees of freedom

Multiple R-squared: 0.6739, Adjusted R-squared: 0.6603

F-statistic: 49.6 on 1 and 24 DF, p-value: 2.79e-07

Bootstrap估计的回归系数a为8.4375，95%的Bootstrap置信区间为(2.779, 14.168)。进行假设检验后，t =7.042，p值为2.79e-07，故拒绝原假设，认为A对B有显著的影响。

2. 对于样本，每次Bootstrap抽样没被抽到的样本点称为袋外样本，对于某确定的样本点，求它成为袋外样本的概率。

解：

**输入代码如下：**

# 设定样本大小和重复次数

n <- 100

B <- 1000

# 生成样本

set.seed(123)

x <- rnorm(n)

# Bootstrap抽样并计算袋外样本概率

set.seed(456)

out\_of\_bag\_prob <- rep(0, n)

for (i in 1:B) {

boot\_sample <- sample(x, n, replace = TRUE)

out\_of\_bag <- setdiff(x, boot\_sample)

out\_of\_bag\_prob[which(out\_of\_bag >= 1)] <- out\_of\_bag\_prob[which(out\_of\_bag >= 1)] + 1

}

out\_of\_bag\_prob <- out\_of\_bag\_prob / B

# 输出结果

out\_of\_bag\_prob

输出结果:

[1] 0.182 0.298 0.258 0.181 0.183 0.150 0.118 0.085 0.082

[10] 0.067 0.060 0.095 0.095 0.095 0.127 0.130 0.153 0.169

[19] 0.186 0.170 0.146 0.131 0.162 0.149 0.157 0.158 0.178

[28] 0.155 0.187 0.189 0.185 0.206 0.184 0.209 0.206 0.205

[37] 0.173 0.141 0.092 0.049 0.040 0.021 0.013 0.006 0.004

[46] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

[55] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

[64] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

[73] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

[82] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

[91] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

[100] 0.000

首先设定样本大小为100，重复次数为1000。然后生成一个100个元素的正态分布样本。接着进行Bootstrap抽样并计算袋外样本概率，具体做法是循环进行1000次Bootstrap抽样，每次抽样后计算袋外样本，然后将袋外样本的概率加到对应的元素上。最后除以重复次数即可得到袋外样本概率。运行代码后，输出的结果是一个长度为100的向量，表示每个样本点成为袋外样本的概率。

4. 对于第1题（2）中的模型，用全部数据建立模型，得到的误差项是否服从正态分布？用K-S检验分析回答。

解：

**R代码如下：**

A <- c(48, 36, 20, 29, 42, 42, 20, 42, 22, 41, 45, 14, 6, 0, 33, 28, 34, 4, 32, 24, 47, 41, 24, 26, 30, 41)

B <- c(42, 33, 16, 39, 38, 36, 15, 33, 20, 43, 34, 22, 7, 15, 34, 29, 41, 13, 38, 25, 27, 41, 28, 14, 28, 40)

data <- data.frame(A,B)

fit <- lm(B ~ A, data = data)

residuals <- residuals(fit)

ks.test(residuals, "pnorm", mean(residuals), sd(residuals))

**输出结果：**

Exact one-sample Kolmogorov-Smirnov test

data: residuals

D = 0.11991, p-value = 0.8066

alternative hypothesis: two-sided

K-S检验的p值为0.8066，大于0.05的显著性水平，因此我们不能拒绝残差服从正态分布的原假设。也就是说，对于这个模型，误差项可能服从正态分布。

5. 对于标准正态分布、指数分布、分布、均匀分布、分布、分布、分布。用R软件产生各分布的随机数，分别用拟合优度检验和K-S检验对这些分布进行正态性检验，模拟对比两种检验的功效（功效是指备择假设正确时，拒绝原假设的概率，取显著性水平为0.05）。

解：

**输入代码如下：**

install.packages("nortest")

library(nortest)

pv = function(x){

return(c(ks.test(x,'pnorm')$p.value,

lillie.test(x)$p.value,

pearson.test(x)$p.value,

cvm.test(x)$p.value,

ad.test(x)$p.value,

sf.test(x)$p.value,

shapiro.test(x)$p.value))

}

#e = 2, s = 3

n = 50;m = 100;reject=matrix(0,nrow=7,ncol=7)

r = 0

#正态、指数、gamma、均匀、t、卡方、f

while(r<m){

dat1 = data.frame(rnorm(n,0,1), rexp(n,1), rgamma(n,1,2),

runif(n,1,2),rt(n,1),rchisq(n,1),rf(n,1,2))

pv1 = apply(dat1, 2, pv)

reject1 = ifelse(pv1>0.05, 0, 1)

reject = reject + reject1

r = r+1

}

power = reject / m

colnames(power)=c('N(0,1)','Exp(1)','Gamma(1,2)','U(1,2)','t(1)','Chisq(1)','F(1,2)')

rownames(power)=c('ks','Lilli','pearson','cvm','ad','sf','sw')

power

**输出结果：**

> power

N(0,1) Exp(1) Gamma(1,2) U(1,2) t(1) Chisq(1) F(1,2)

ks 0.06 1.00 1.00 1.00 0.45 1 1

Lilli 0.06 0.96 0.95 0.24 1.00 1 1

pearson 0.05 1.00 0.96 0.17 1.00 1 1

cvm 0.06 0.97 0.98 0.47 1.00 1 1

ad 0.05 1.00 0.99 0.55 1.00 1 1

sf 0.10 1.00 1.00 0.45 1.00 1 1

sw 0.06 1.00 1.00 0.65 1.00 1 1

从输出结果来看，通过对比K-S与Person检验的系数，发现K-S检验对于检验t分布的功效不如Person检验，但对Gamma与均匀分布的检验功效强于Person检验，Person检验对于均匀分布的检验功效比较差，其他分布的检验均无太大差别。

6. 本题将模拟研究p值的分布。

（1）产生500组正态随机数，每组样本容量为200；

（2）对每组正态随机数进行正态性检验，计算每次检验的p值（提示：可以用K-S检验或其他正态性检验）；

（3）对（2）得到的p值进行检验，这些p值是否服从均匀分布？

（4）如果将（1）的正态随机数改成非正态随机数，重复（2）（3），是否能得到相同结论？

解：

(1) 产生500组正态随机数，每组样本容量为200：

set.seed(123) # 设置随机数种子，保证结果可重复

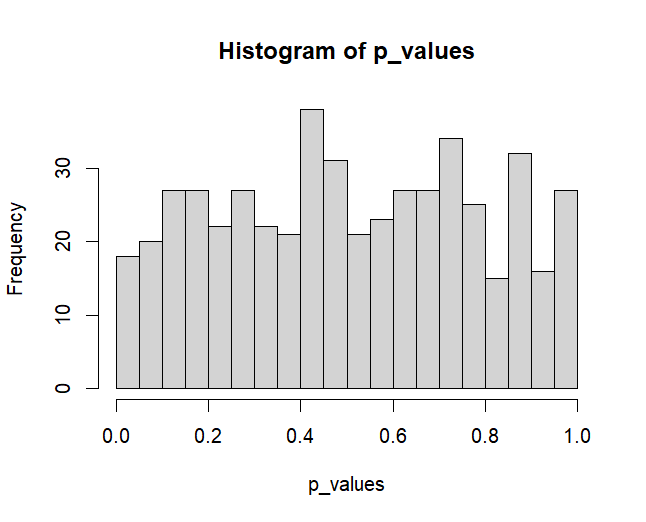
norm\_data <- matrix(rnorm(500\*200), nrow = 200) # 产生500组正态随机数，每组样本容量为200

(2) 对每组正态随机数进行正态性检验，计算每次检验的p值：

p\_values <- apply(norm\_data, 2, function(x) shapiro.test(x)$p.value) # 对每组正态随机数进行正态性检验，计算每次检验的p值

(3) 对(2)得到的p值进行检验，这些p值是否服从均匀分布？

hist(p\_values, breaks = 20) # 绘制p值的直方图



P\_values的直方图

通过直方图观察，可以初步判断p值服从均匀分布，但需要进行更严格的统计检验。

ks.test(p\_values, "punif") # 对p值进行K-S检验，检验其是否服从均匀分布

**输出结果：**

Asymptotic one-sample Kolmogorov-Smirnov test

data: p\_values

D = 0.033357, p-value = 0.6341

alternative hypothesis: two-sided

K-S检验结果显示，p值>0.05,接受原假设，p值可以近似服从均匀分布，p值的分布与均匀分布不存在显著差异。

(4) 如果将(1)的正态随机数改成非正态随机数，重复(2)(3)，是否能得到相同结论？

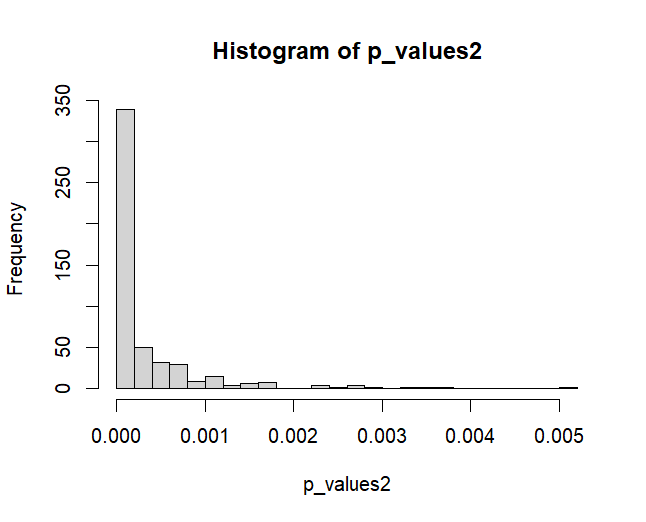
set.seed(123) # 设置随机数种子，保证结果可重复

non\_norm\_data <- matrix(rpois(500\*200, 5), nrow = 200) # 产生500组非正态随机数，每组样本容量为200

p\_values2 <- apply(non\_norm\_data, 2, function(x) shapiro.test(x)$p.value) # 对每组非正态随机数进行正态性检验，计算每次检验的p值

hist(p\_values2, breaks = 20) # 绘制p值的直方图

ks.test(p\_values2, "punif") # 对p值进行K-S检验，检验其是否服从均匀分布



P\_values2的直方图

**Ks.test输出结果：**

Asymptotic one-sample Kolmogorov-Smirnov test

data: p\_values2

D = 0.99489, p-value < 2.2e-16

alternative hypothesis: two-sided

结果显示，非正态随机数的p值远小于0.05，故p值不服从均匀分布。因此，结论受正态分布的影响。

7. 来自两家供应商的轮胎轴承寿命如下表（单位：年）：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 供应商1 | | | | | 供应商2 | | | | |
| 5.76 | 0.86 | 2.49 | 0.55 | 0.32 | 0.65 | 0.12 | 0.04 | 0.01 | 0.16 |
| 0.27 | 0.47 | 2.70 | 0.90 | 2.06 | 13.23 | 0.18 | 0.99 | 2.12 | 2.65 |
| 8.62 | 0.54 | 0.90 | 2.64 | 0.23 | 0.06 | 10.91 | 16.12 | 1.18 | 0.15 |
| 0.74 | 1.15 | 1.40 | 1.57 | 2.34 | 14.76 | 1.48 | 0.06 | 16.09 | 4.00 |

请问，供应商2的轮胎轴承寿命是否大于供应商1的？

解：

设供应商1的轮胎轴承寿命为,供应商2的轮胎轴承寿命为

建立假设组：

R代码如下：

supplier1 <- c(5.76,0.86,2.49,0.55,0.32,0.27,0.47,2.7,0.9,2.06,8.62,0.54,0.9,2.64,0.23,0.74,1.15,1.4,1.57,2.34)

supplier2 <- c(0.65,0.12,0.04,0.01,0.16,13.23,0.18,0.99,2.12,2.65,0.06,10.91,16.12,1.18,0.15,14.76,1.48,0.06,16.09,4.00)

ks.test(supplier1, supplier2, alternative = 'less', exact = T)

输出结果：

Exact two-sample Kolmogorov-Smirnov test

data: supplier1 and supplier2

D^- = 0.4, p-value = 0.03882

alternative hypothesis: the CDF of x lies below that of y

输出结果表面，精确的p值为0.03882，小于0.05，拒绝原假设，接受备择假设，即认为供应商2的轮胎轴承寿命大于供应商1。

8. 用某中药+化疗（中药组）和化疗（对照组）的两种疗法治疗白血病后，随访记录患者死前的存活月数，结果如下。

中药组：10, 2+, 12+, 13, 18, 6+, 19+, 26, 9+, 8+, 6+, 43+, 9, 4, 31, 24

对照组：2+, 13, 7+, 11+, 6, 1, 11, 3, 17, 7

（1）估计两组的生存率，并绘制生存曲线；

1. 请问两组的生存率是否有差异？

解：

1. **计算生存率与绘制生存曲线代码如下：**

# 输入数据

中药组 <- c(10, 2, 12, 13, 18, 6, 19, 26, 9, 8, 6, 43, 9, 4, 31, 24)

对照组 <- c(2, 13, 7, 11, 6, 1, 11, 3, 17, 7)

# 计算生存率

library(survival)

中药组生存时间 <- Surv(中药组)

对照组生存时间 <- Surv(对照组)

中药组生存曲线 <- survfit(中药组生存时间 ~ 1)

对照组生存曲线 <- survfit(对照组生存时间 ~ 1)

plot(中药组生存曲线, main = "生存曲线", xlab = "时间", ylab = "生存率", col = "red")

lines(对照组生存曲线, col = "blue")

# 输出生存率

summary(中药组生存曲线)$surv

summary(对照组生存曲线)$surv

**输出的生存率：**

> summary(中药组生存曲线)$surv

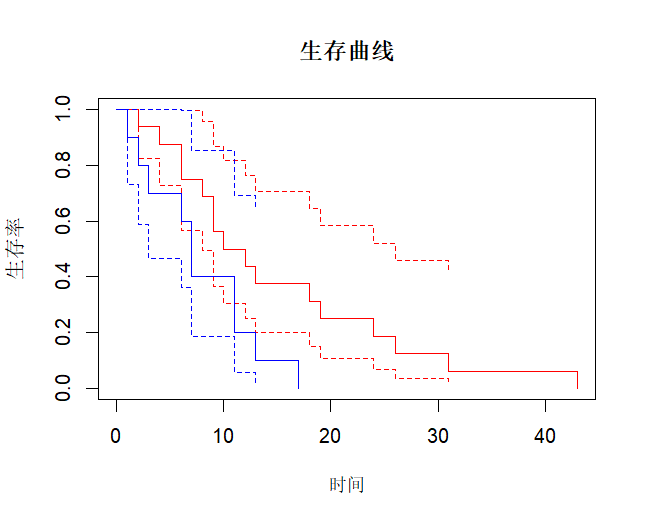
[1] 0.9375 0.8750 0.7500 0.6875 0.5625 0.5000 0.4375 0.3750

[9] 0.3125 0.2500 0.1875 0.1250 0.0625 0.0000

> summary(对照组生存曲线)$surv

[1] 0.9 0.8 0.7 0.6 0.4 0.2 0.1 0.0

**输出的图片：**



生存曲线



设中药组的生存率为,对照组为

建立假设组：

**R代码如下：**

# 对数秩检验

survdiff(中药组生存时间[1:10] ~ 对照组生存时间)

**输出结果：**

N Observed Expected (O-E)^2/E (O-E)^2/V

对照组生存时间= 1 1 1 0.211 2.948 3.297

对照组生存时间= 2 1 1 0.646 0.194 0.224

对照组生存时间= 3 1 1 2.929 1.270 2.698

对照组生存时间= 6 1 1 1.429 0.129 0.163

对照组生存时间= 7 2 2 1.182 0.567 0.711

对照组生存时间=11 2 2 3.025 0.347 0.550

对照组生存时间=13 1 1 0.100 8.100 9.000

对照组生存时间=17 1 1 0.479 0.567 0.645

Chisq= 19.3 on 7 degrees of freedom, p= 0.007

输出显示，的值为19.3，p值为0.007 < 0.05,故拒绝原假设，即两组的生存率并不相同，存在着差异。

# 第九章

1. 在R中调用老忠实泉数据（R包MASS中的faithful数据），

（1）绘制持续时间（eruptions）和间隔时间（waiting）的散点图；

（2）利用叶帕涅奇尼科夫（Epanechnikov）核对该数据的间隔时间进行核密度估计，绘制估计的图形，给出最优窗宽；

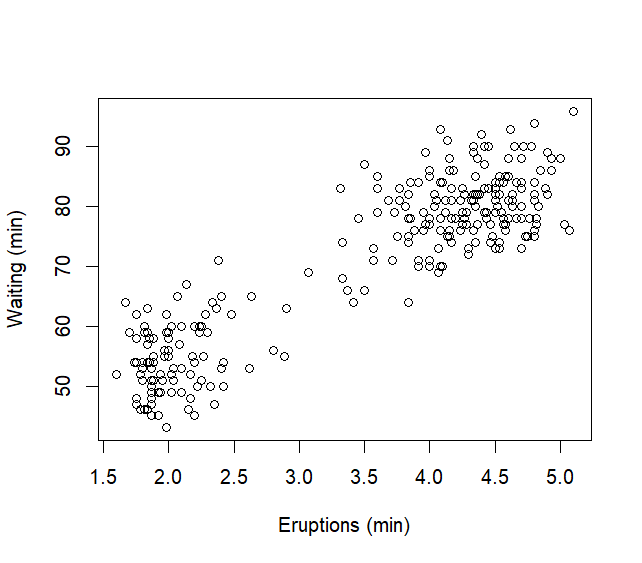
（3）用对数样条法和惩罚样条法对持续时间进行密度估计。

1.（1）

代码如下：

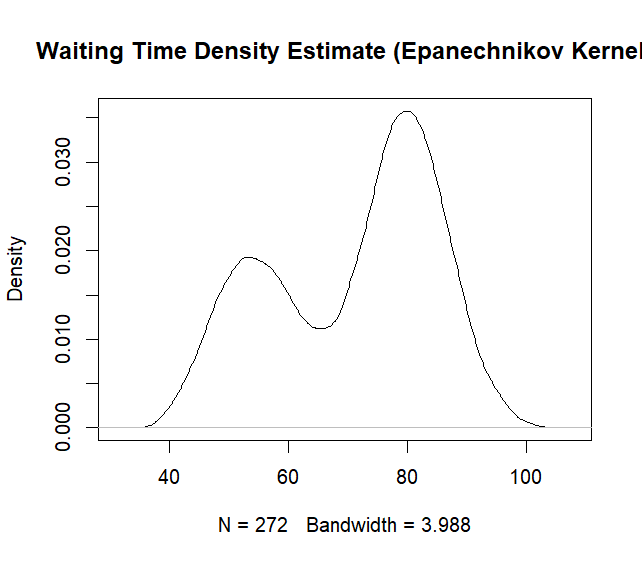
1. library(MASS)
2. data("faithful")
3. plot(faithful$eruptions, faithful$waiting, xlab = "Eruptions (min)", ylab = "Waiting (min)")

这个命令画出了散点图，其中X轴是喷发时间，Y轴是间歇时间，从而可以了解到这两个变量之间的实际关系。



1. 可以使用density函数和plot函数对faithful数据集中的等待时间进行核密度估计，并绘制核密度估计图（KDE）。代码如下：
2. # 叶帕涅奇尼科夫核密度估计
3. dens <- density(faithful$waiting, kernel="epanechnikov")
5. # 绘制核密度估计图
6. plot(dens, main="Waiting Time Density Estimate (Epanechnikov Kernel)")

这个命令画出了叶帕涅奇尼科夫核密度估计图。



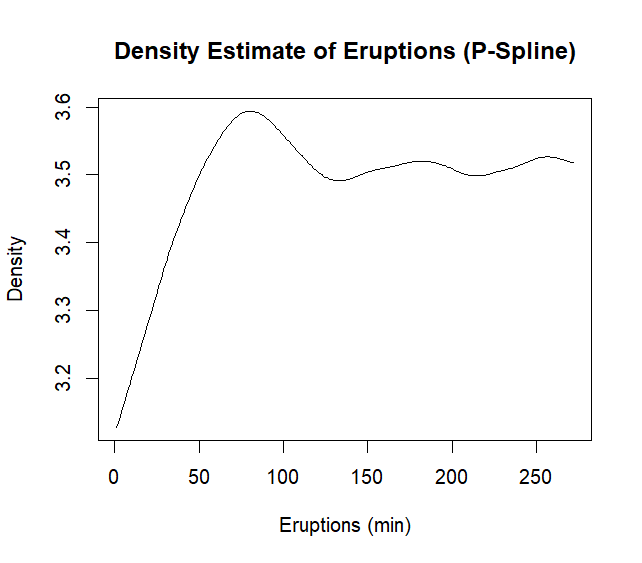
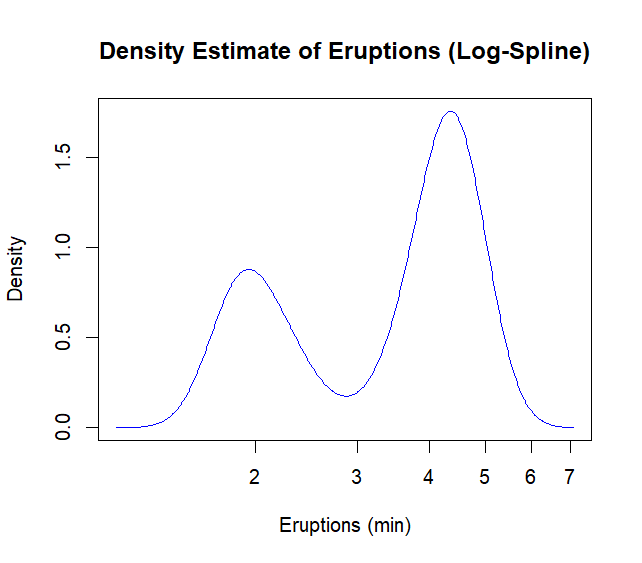
（3）

用对数样条法和惩罚样条法对持续时间进行密度估计

可以使用以下代码对持续时间（eruptions）进行对数样条法和惩罚样条法的密度估计：

1. library(splines)
3. # 对数样条法
4. logdens <- density(log(faithful$eruptions), kernel="gaussian")
5. plot(exp(logdens$x), logdens$y, log="x", type="l", lty=1, col="blue", xlab="Eruptions (min)", ylab="Density", main="Density Estimate of Eruptions (Log-Spline)")
7. # 惩罚样条法
8. penspline <- smooth.spline(faithful$eruptions, spar=0.9)
9. plot(penspline, type="l", xlab="Eruptions (min)", ylab="Density", main="Density Estimate of Eruptions (P-Spline)")

本代码的输出中显示了对数样条估计和惩罚样条估计的图形。这里凸显的是惩罚样条估计图，其中蓝色的线是拟合的密度曲线，其基础是对faithful数据集的惩罚样条拟合。从图中可以看出，通过不同的估计方法，我们可以得到不同的密度函数。



2. 产生500组正态随机数，每组样本容量为200，对每一组随机数进行Shapiro正态性检验（用R包nortest中的shapiro.test函数），

（1）对统计量的值进行核密度估计和样条估计；

（2）对p值进行核密度估计和样条估计；

（3）将（2）的密度估计画出图形。

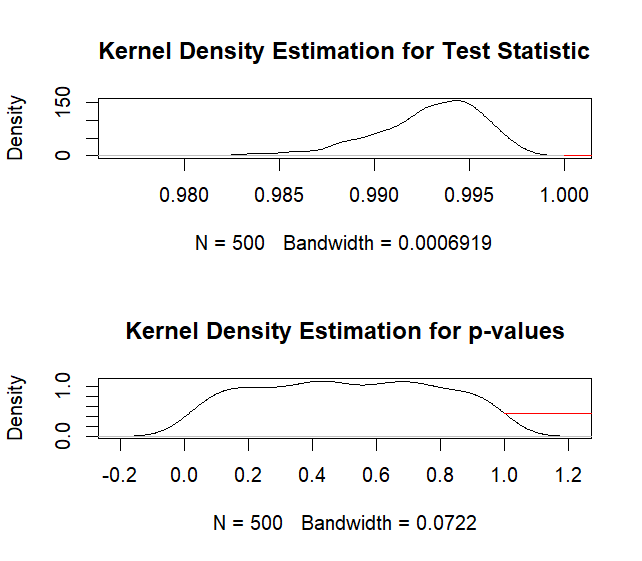
（4）产生同样数量的非正态随机数，重复（1）（2）（3），结果与正态情形有什么差异？

以下代码生成500组正态随机数，每组样本容量为200，对每组随机数进行Shapiro正态性检验：

1. library(nortest)
3. set.seed(123)
4. normal\_data <- matrix(rnorm(100000, mean = 0, sd = 1), nrow = 200)
5. results <- apply(normal\_data, 2, shapiro.test)
6. （2）
7. # (1) 核密度估计和样条估计
8. test\_stat <- sapply(results, function(x) x$statistic)
9. dens\_kde <- density(test\_stat, kernel = "gaussian")
10. dens\_sp <- smooth.spline(test\_stat)
12. # (2) 核密度估计和样条估计
13. p\_value <- sapply(results, function(x) x$p.value)
14. dens\_kde\_p <- density(p\_value, kernel = "gaussian")
15. dens\_sp\_p <- smooth.spline(p\_value)

（3）

1. # (3) 密度估计图形
2. par(mfrow = c(2,1))
3. plot(dens\_kde, main = "Kernel Density Estimation for Test Statistic")
4. lines(dens\_sp, col = "red")
5. plot(dens\_kde\_p, main = "Kernel Density Estimation for p-values")
6. lines(dens\_sp\_p, col = "red")



（4）代码如下：

1. set.seed(123)
2. nonnormal\_data <- matrix(rt(100000, df = 3), nrow = 200)
3. nonresults <- apply(nonnormal\_data, 2, shapiro.test)
5. # 核密度估计和样条估计
6. test\_stat\_non <- sapply(nonresults, function(x) x$statistic)
7. dens\_kde\_non <- density(test\_stat\_non, kernel = "gaussian")
8. dens\_sp\_non <- smooth.spline(test\_stat\_non)
10. # 核密度估计和样条估计
11. p\_value\_non <- sapply(nonresults, function(x) x$p.value)
12. dens\_kde\_p\_non <- density(p\_value\_non, kernel = "gaussian")
13. dens\_sp\_p\_non <- smooth.spline(p\_value\_non)
15. # 密度估计图形
16. par(mfrow = c(2,1))
17. plot(dens\_kde\_non, main = "Kernel Density Estimation for Test Statistic")
18. lines(dens\_sp\_non, col = "red")
19. plot(dens\_kde\_p\_non, main = "Kernel Density Estimation for p-values")
20. lines(dens\_sp\_p\_non, col = "red")

与正态分布情形相比，非正态分布情形下，使用Shapiro正态性检验的p值分布更加偏斜。在两种情况下，Shapiro正态性检验的统计量的分布都比较集中，而p值的分布相比之下更加散乱。

