实验三：利用 R 软件单变量和多变量正态检验和置信区域等

林泽钦 3160104013 统计 1601

目录

1 实验目的和要求 **1**

1.1 实验目的 1

1.2 实验内容 1

1.3 实验环境 2

2 实验过程与结果 **2**

2.1 第一部分 2

2.2 第二部分 13

1 实验目的和要求

1.1 实验目的

通过本试验项目，能够理解并掌握如下内容：

- 单变量和多变量正态检验；
- 多变量均值向量显著性检验；
- 置信域和置信区间计算，画置信椭球等。

1.2 实验内容

1.2.1 第一部分

采用实验二 sample 样本。附表中的数据 sample.xls 进行分析。记 $X_1 = BMI, X_2 = FPG, X_3 = SBP, X_4 = DBP, X_5 = TG, X_6 = HDL - C$，并

构成一个向量。$X = (X_1, X_2, X_3, X_4, X_5, X_6)$，任选下列一项详细分析患代谢综合症的群体与没有患代谢综合症群的差异。

- 分析患代谢综合症的年龄差异
- 分析患代谢综合症的性别差异
- 分析是否吸烟对患代谢综合症的影响
- 分析是否喝酒对患代谢综合症的影响

1.2.2 第二部分

数据 ex2.1 给出了 27 名糖尿病人血清总胆固醇 (x_1)，甘油 (x_2)，空腹胰岛素 (x_3)，糖化血红蛋白 (x_4)，空腹血糖 (y) 的测量值。

- 试建立血糖 (y) 与其他指标的线性回归方程，并进行分析；
- (x_1, x_2, x_3, x_4) 是否服从多元正态？(x_1, x_2) 与 (x_3, x_4) 是否相互独立？

1.3 实验环境

- R-3.5.1
- RStudio

2 实验过程与结果

2.1 第一部分

实验中选择对患代谢综合症的年龄差异进行分析。

2.1.1 数据导入与预处理

数据的导入与预处理与实验二基本相同，主要过程为：

- 导入程序包
- 导入实验数据
- 计算 BMI
- 计算用于判断代谢综合症的各个条件
- 判断代谢综合症
- 去除异常数据以及含有 NA 的行

报告中省去了对异常值的分析，仅展示数据处理阶段的代码（对异常值的判断参见实验二的报告）。本次实验没有用到性别、吸烟、饮酒等数据，故在处理过程中删去。

```r
# import packages
library(tidyverse)
library(gridExtra)
library(devtools)
library(dplyr)
library(ggpubr)
library(MVN)

# import data
OriginData <- read_csv("Pro3Data1.csv")

# calculate BMI
DataWithBMI <-
  mutate(OriginData, BMI = 10000 * weight / (height*height))

# symptoms for diagnosing Metabolic syndrome
Overweight <- DataWithBMI$BMI >= 25
HighBloodSugar <- DataWithBMI$FPG >= 6.1
Hypertension <- DataWithBMI$sbp >= 140 | DataWithBMI$dbp >= 90
FastingBlood <- with(DataWithBMI, TG > 1.7 |
                       (gender == "男" & HDLC < 0.9) |
                       (gender == "女" & HDLC < 1.0))

# MS: Logical variable for Metabolic syndrome
CompleteData <-
  mutate(DataWithBMI, MS =
           (Overweight & HighBloodSugar & Hypertension) |
           (Overweight & HighBloodSugar & FastingBlood) |
           (Overweight & Hypertension & FastingBlood) |
           (HighBloodSugar & Hypertension & FastingBlood))
```

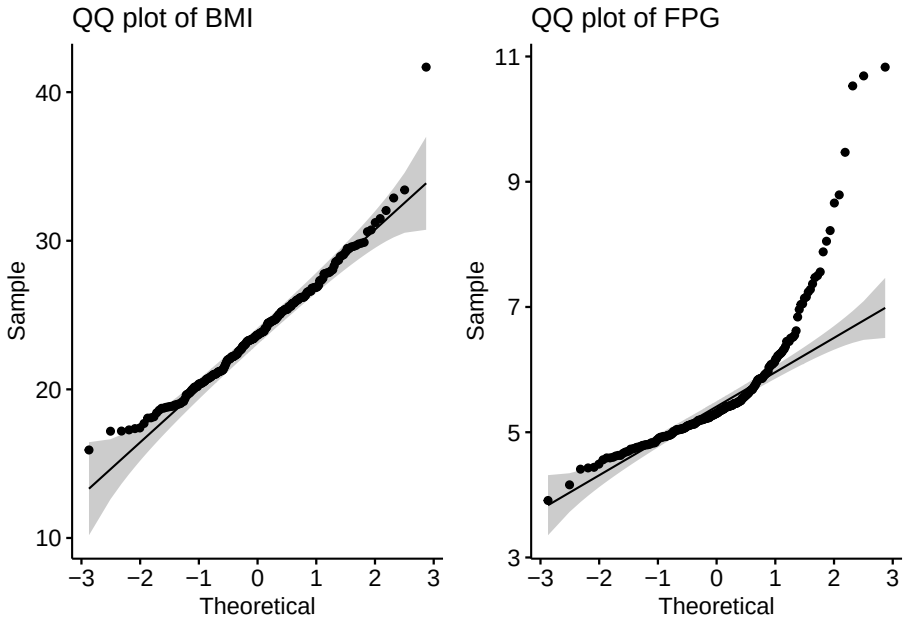
```
# remove unusual values
CompleteData <- CompleteData %>%
  filter(!((age < 10) | (BMI > 500) | (height < 60)))

# delete irrelevant columns
# and delete rows with NA
CompleteData <- CompleteData %>%
  select(age, BMI, FPG, sbp, dbp, TG, HDLC, MS) %>%
  na.omit()
```

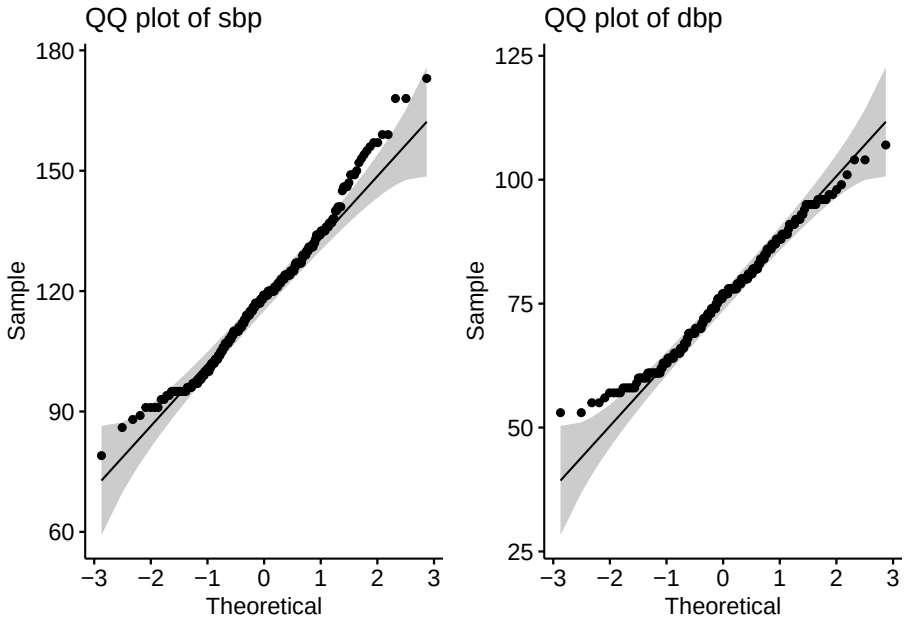
2.1.2 检验相关数据正态性

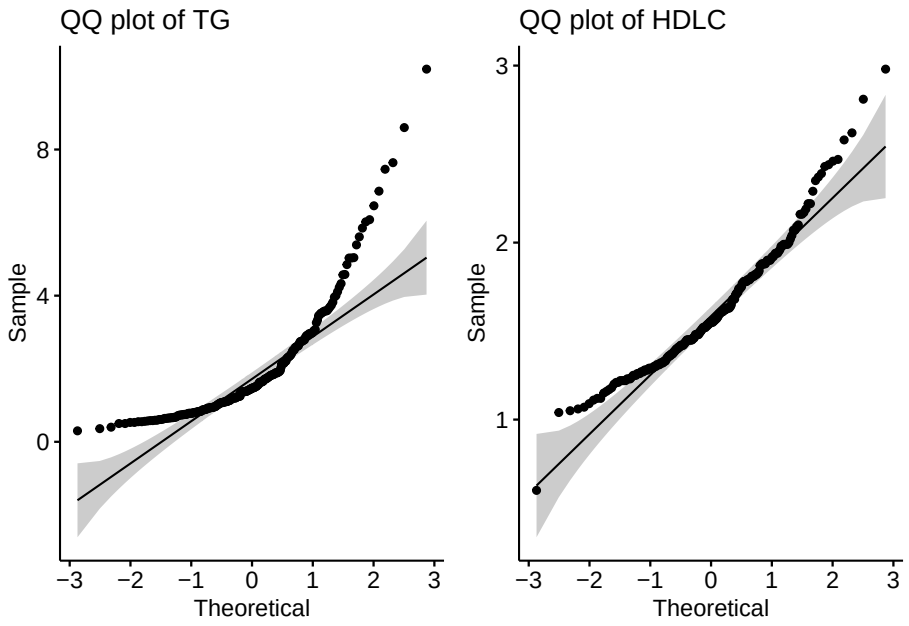
可以用 Q-Q 图对各指标数据进行观察。绘制 BMI 与 FPG 的 Q-Q 图的代码为：

```
pie1 <- ggqqplot(CompleteData$BMI,
                 main = "QQ plot of BMI")
pie2 <- ggqqplot(CompleteData$FPG,
                 main = "QQ plot of FPG")
grid.arrange(pie1, pie2, ncol=2, nrow=1)
```



类似的可以绘制出 sbp, dbp, TG, HDLC 等数据的 Q-Q 图。





初步判断，FPG 与 TG 不服从正态分布，而其它数据可以近似看作正态分布。但是具体是否可看作正态分布需要进一步的检验。下面使用假设检验来检验该数据的正态性。一元数据的常用的正态性检验方法有：

- Shapiro-Wilk test
- Cramer-von Mises test
- Lilliefors test
- Shapiro-Francia test
- Anderson-Darling test

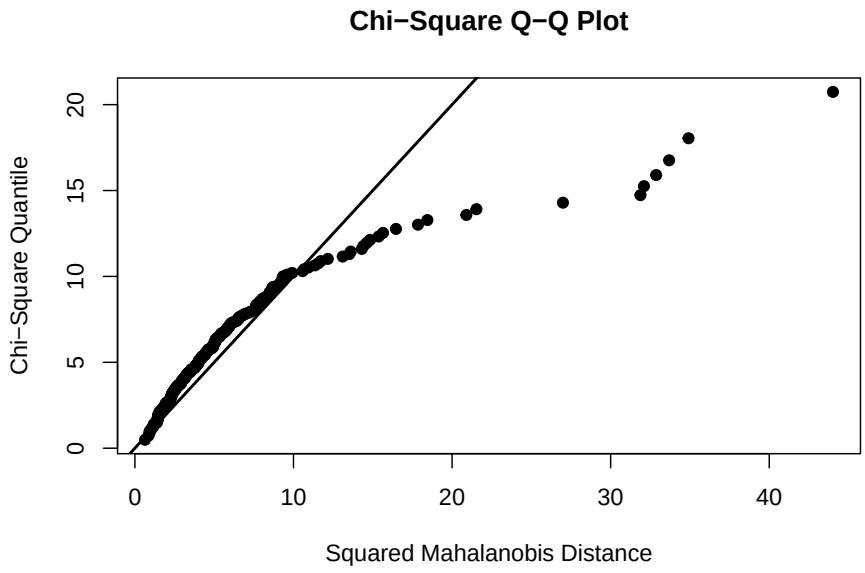
多元数据的正态性也有多种检验方法，例如：

- Mardia's test
- Henze-Zirkler test
- Royston test
- Doornik-Hansen test

实验中使用 MVN 包中的 mvn() 函数来检测数据的多元正态性。该函数可以指定使用的多元正态检验方法，同时可以对各个变量也进行正态性检验。实验中联合数据的正态检验使用了 Mardia's test 方法，单个变量正

态性的检验使用了常见的 Shapiro-Wilk test。下面的代码进行了检验，并绘制出联合数据的 χ^2 统计量的 Q-Q 图。

```
res <- mvn(CompleteData[2:7],
           mvnTest = "mardia",
           univariateTest = "SW",
           multivariatePlot = "qq")
```



首先查看各个变量正态性检验的结果：

```
res$univariateNormality
```

```
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    BMI       0.9728   1e-04      NO
## 2 Shapiro-Wilk    FPG       0.7525  <0.001      NO
## 3 Shapiro-Wilk    sbp       0.9800  0.0016      NO
## 4 Shapiro-Wilk    dbp       0.9852  0.0123      NO
## 5 Shapiro-Wilk    TG        0.7865  <0.001      NO
## 6 Shapiro-Wilk   HDLC       0.9508  <0.001      NO
```

可以看到，各个变量检验得到的 p-value 都十分小 (<0.05)，因此倾向于拒绝原假设，认为正态性不成立。接着查看联合数据的正态性检验结果：

```
res$multivariateNormality
```

```
##              Test        Statistic                p value Result
## 1 Mardia Skewness 705.950340431048 3.10020762236209e-113     NO
## 2 Mardia Kurtosis 19.8399016153122                      0     NO
## 3             MVN             <NA>                   <NA>     NO
```

同样，计算出 Mardia Skewness 与 Mardia Kurtosis 的 p-value 都十分小，因此不认为实验数据服从联合正态分布。关于实验中 mvn() 函数的使用，参考了 R 官方文档：

https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf

而 Mardia's test 的具体原理参见：Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika 36:519-530.

2.1.3 检验相关数据相关性

计算 X 的样本相关阵：

```
CompleteData %>%
  select(sbp, dbp, FPG, TG, HDLC, BMI) %>%
  cor()
```

```
##              sbp        dbp        FPG         TG        HDLC        BMI
## sbp   1.00000000  0.8058217  0.1948614  0.1943769 -0.06678692  0.3451072
## dbp   0.80582173  1.0000000  0.2679890  0.3203713 -0.12099823  0.3956302
## FPG   0.19486142  0.2679890  1.0000000  0.3879515 -0.19905304  0.2642223
## TG    0.19437686  0.3203713  0.3879515  1.0000000 -0.33294231  0.3871581
## HDLC -0.06678692 -0.1209982 -0.1990530 -0.3329423  1.00000000 -0.3401106
## BMI   0.34510723  0.3956302  0.2642223  0.3871581 -0.34011060  1.0000000
```

2.1.4 分析人群患代谢综合症的比例

首先根据年龄对数据进行分组。首先查看年龄数据的相关统计信息：

```
summary(CompleteData$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   40.00   47.00   46.72   54.00   76.00
```

将年龄分组为青年 (15-40)，中年 (41-60) 与老年 (61-76)：

```
CompleteData <- CompleteData %>%
  mutate(generation =
           ifelse(age <= 40, "Young",
           ifelse(age <= 60, "MiddleAged", "Old"))) %>%
  mutate(generation = factor(generation, ordered = TRUE,
           levels = c("Young", "MiddleAged", "Old")))
```

然后生成频率表来查看不同年龄段患代谢综合症的比例的差异：

```
MSWithGeneration <- xtabs(~ generation+MS, data = CompleteData)
prop.table(MSWithGeneration, 1)
```

```
##             MS
## generation       FALSE      TRUE
##   Young      0.9843750 0.0156250
##   MiddleAged 0.8364780 0.1635220
##   Old        0.8181818 0.1818182
```

可以看到，随着年龄段的上升，患病比率也有一定程度的上升。为了验证两者的相关性，对其进行 Fisher 精确检验：

```
fisher.test(MSWithGeneration)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  MSWithGeneration
## p-value = 0.001978
## alternative hypothesis: two.sided
```

可以看到检验的 p-value < 0.01，因此认定代谢综合症的患病率与年龄段并不独立。

2.1.5 不同年龄组是否患代谢综合症群体各类指标的均值估计

可以使用 colMeans() 函数来计算各类指标的均值。例如，对于不患病的年轻人群，各类指标的均值计算如下：

```
Young.Healthy <- CompleteData %>%
  filter(MS == FALSE, generation == "Young") %>%
  select(BMI, FPG, sbp, dbp, TG, HDLC)

colMeans(Young.Healthy)
```

```
##        BMI        FPG        sbp        dbp         TG       HDLC
##  22.510807   5.300159 111.650794  70.396825   1.344603   1.637778
```

将计算结果综合如下：

```
##                         BMI      FPG      sbp      dbp       TG     HDLC
## Young.Healthy       22.51081 5.300159 111.6508 70.39683 1.344603 1.637778
## Young.Sick          32.03510 5.190000 141.0000 85.00000 3.640000 1.390000
## MiddleAged.Healthy  23.60666 5.412632 116.5263 75.45113 1.915714 1.626316
## MiddleAged.Sick     27.46092 6.801923 131.1538 86.65385 3.726923 1.414615
## Old.Healthy         23.30660 5.591111 132.5000 79.00000 1.605556 1.667222
## Old.Sick            26.82007 6.340000 152.5000 97.50000 2.045000 1.560000
```

可以看到，在同年龄段的情况下，患病人群有着更高的 BMI, FPG, sbp, dbp 以及 TG，而 HDL-C 值就相对较低。

2.1.6 不同年龄组是否患代谢综合症群体各类指标的置信区间

可以利用 t 检验来计算中年患病人群的 BMI 指标的置信区间：

```
##
##  One Sample t-test
##
```

```
## data:  MiddleAged.Sick$BMI
## t = 41.569, df = 25, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  26.10036 28.82148
## sample estimates:
## mean of x
##  27.46092
```

如上所示，其置信程度为 95% 的执行区间为 [26.10036, 28.82148]。类似的可以计算出各组数据各项指标的置信区间。例如青年健康人群各项指标的置信区间计算如下：

```
BMI.Interval <- as.numeric(unlist(t.test(Young.Healthy$BMI)[4]))
FPG.Interval <- as.numeric(unlist(t.test(Young.Healthy$FPG)[4]))
sbp.Interval <- as.numeric(unlist(t.test(Young.Healthy$sbp)[4]))
dbp.Interval <- as.numeric(unlist(t.test(Young.Healthy$dbp)[4]))
TG.Interval <- as.numeric(unlist(t.test(Young.Healthy$TG)[4]))
HDLC.Interval <- as.numeric(unlist(t.test(Young.Healthy$HDLC)[4]))

rbind(BMI.Interval, FPG.Interval, sbp.Interval,
      dbp.Interval, TG.Interval, HDLC.Interval)
```

```
##                     [,1]       [,2]
## BMI.Interval   21.574882  23.446732
## FPG.Interval    5.127139   5.473178
## sbp.Interval  108.301364 115.000224
## dbp.Interval   67.967034  72.826617
## TG.Interval     1.107365   1.581841
## HDLC.Interval   1.545411   1.730144
```

其他的年龄组的计算结果如下，其中由于青年患病的样本只有一例，故没有对其进行计算。

青年患病：

```
## # A tibble: 1 x 6
```

```
##     BMI   FPG   sbp   dbp    TG  HDLC
##   <dbl> <dbl> <int> <int> <dbl> <dbl>
## 1  32.0  5.19   141    85  3.64  1.39
```

中年健康：

```
##                     [,1]       [,2]
## BMI.Interval   23.094161  24.119156
## FPG.Interval    5.273245   5.552018
## sbp.Interval  113.953335 119.099296
## dbp.Interval   73.643718  77.258538
## TG.Interval     1.674438   2.156991
## HDLC.Interval   1.570023   1.682608
```

中年患病：

```
##                     [,1]       [,2]
## BMI.Interval   26.100363  28.821482
## FPG.Interval    6.208887   7.394959
## sbp.Interval  125.206979 137.100713
## dbp.Interval   82.880478  90.427214
## TG.Interval     2.827404   4.626442
## HDLC.Interval   1.298401   1.530830
```

老年健康：

```
##                     [,1]       [,2]
## BMI.Interval   21.922254  24.690938
## FPG.Interval    5.164755   6.017467
## sbp.Interval  122.456355 142.543645
## dbp.Interval   73.238420  84.761580
## TG.Interval     1.249953   1.961158
## HDLC.Interval   1.478593   1.855851
```

老年患病：

```
##                     [,1]       [,2]
```

```
## BMI.Interval   22.052941  31.587191
## FPG.Interval    5.641914   7.038086
## sbp.Interval  125.480425 179.519575
## dbp.Interval   90.564019 104.435981
## TG.Interval     1.134659   2.955341
## HDLC.Interval   1.322202   1.797798
```

2.2 第二部分

2.2.1 数据导入

```
Data <- read.csv("Pro3Data2.csv")
```

2.2.2 建立线性回归方程

对数据建立线性回归方程如下：

```
lm.sol <- with(Data, lm(y ~ x1 + x2 + x3 + x4))
summary(lm.sol)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6268 -1.2004 -0.2276  1.5389  4.4467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9433     2.8286   2.101   0.0473 *
## x1            0.1424     0.3657   0.390   0.7006
## x2            0.3515     0.2042   1.721   0.0993 .
## x3           -0.2706     0.1214  -2.229   0.0363 *
## x4            0.6382     0.2433   2.623   0.0155 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01 on 22 degrees of freedom
## Multiple R-squared:  0.6008, Adjusted R-squared:  0.5282
## F-statistic: 8.278 on 4 and 22 DF,  p-value: 0.0003121
```

拟合出的回归方程为：

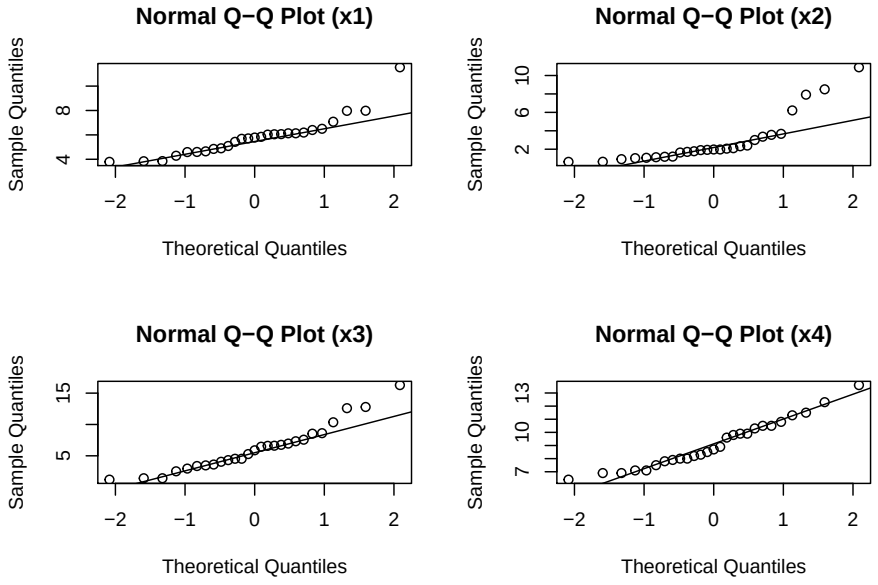
$y = 5.9433 + 0.1424\ x_1 + 0.3515\ x_2 - 0.2706\ x_3 + 0.6382\ x_4$

并且从 Coefficient 一栏的信息可以看出，x_1 的系数的拟合效果较差，其他系数拟合效果较好。

2.2.3 正态性分析

正态性的检验方法与实验的第一部分相同。

```
res <- mvn(Data[2:5],
           mvnTest = "mardia",
           univariateTest = "SW",
           univariatePlot = "qqplot")
```



各个变量的正态性：

```
res$univariateNormality
```

```
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    x1        0.8484  0.0011      NO
## 2 Shapiro-Wilk    x2        0.7318  <0.001      NO
## 3 Shapiro-Wilk    x3        0.9255  0.0537      YES
## 4 Shapiro-Wilk    x4        0.9556  0.2916      YES
```

随机向量的正态性：

```
res$multivariateNormality
```

```
##              Test        Statistic             p value Result
## 1 Mardia Skewness 44.5642943964067 0.00126381089971346     NO
## 2 Mardia Kurtosis 1.86540442491453  0.0621247376007843    YES
## 3             MVN             <NA>                <NA>     NO
```

从上述结果可以看出，x_1, x_2 各自不服从正态分布，而 x_3, x_4 各自近似服从正态分布。各个变量构成的随机向量不服从联合正态分布。假如去除变量 x_1, x_2 之后再进行检验，可以看到 x_3, x_4 构成的随机向量是近似服从联合正态分布的。

```
res <- mvn(Data[4:5],
           mvnTest = "mardia",
           univariateTest = "SW")
res$multivariateNormality
```

```
##              Test          Statistic           p value Result
## 1 Mardia Skewness   6.59114673624618 0.159137236731136    YES
## 2 Mardia Kurtosis -0.259882094800697 0.794954723139331    YES
## 3             MVN               <NA>              <NA>    YES
```

2.2.4 独立性分析

这个问题比较难办，这里只能用一些比较 naive 的方法尝试进行分析。比如可以计算样本协方差阵：

```
cor(Data[2:5])
```

```
##             x1          x2          x3         x4
## x1  1.0000000  0.63150583 -0.35479471  0.4152708
## x2  0.6315058  1.00000000 -0.03863221  0.2189743
## x3 -0.3547947 -0.03863221  1.00000000 -0.3297787
## x4  0.4152708  0.21897432 -0.32977870  1.0000000
```

可以看到，(x_1, x_2) 与 (x_3, x_4) 存在着一定的相关性。如果 (x_1, x_2) 与 (x_3, x_4) 是独立的，那么它们之间一定不相关。由此可以初步认为，它们之间不是独立的。

至于用于检验非正态分布总体变量之间独立性的较可靠方法，我目前还没找到相关资料。一种朴素的想法是，将随机向量的值域划分为若干个区间，然后将连续型变量的观测值映射为类别，然后就可以通过卡方独立性检验来检验数据的独立性。不过这种方法受划分好坏、观测数目的影响很大，实际中并不使用。在实验的例子中，观测数目只有 27，并不适合这种方法。