实验二：利用 R 软件进行参数估计和假设检验

林泽钦 3160104013 统计 1601

目录

1 实验目的和要求

1.1 实验目的

通过本试验项目：

- 使学生理解并掌握数理统计中一些单变量参数估计和假设检验问题中在 R 软件包中的实现；
- 多维数据的均值、方差、协方差矩阵等等；

- 多维正态密度函数、分布函数，随机数等的产生；
- 二维正态密度曲线作图。

1.2 实验内容

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。$X = (X_1, X_2, X_3, X_4, X_5, X_6)$。

- 分析 X 各变量之间的相关性？
- 分析患代谢综合症的比例有没有性别差异，与吸烟或喝酒是否有关？
- 分年龄 (小于等于 30，30~50, 50~70, 70 以上)，分析 X 中的各个指标是否有年龄上的差异？
- 计算 X 样本均值、样本离差阵、样本协方差和样本相关阵。
- 分析 X_2, X_3 是否服从正态分布?

1.3 实验环境

- R-3.5.1
- RStudio

2 实验过程与结果

2.1 数据导入

为了简化数据导入与处理，实验中使用了 tidyverse 包。

```
library(tidyverse)
library(gridExtra)
library(corrgram)
```

实验中首先将提供的.xls 文件保存为.csv 文件，为了不影响读入，在导入数据之前就将中文的标题行删去，只留下英文的标题行。然后就可以直接使用 read_csv 函数将数据读入，得到一个数据类型为 tibble 的表。此数据类型的用法与数据框大致相同，不过 tidyverse 做了些许修改使得其更适合数据处理以及可视化。

```
OriginData <- read_csv("Pro2Data.csv")
```

可以看到，在生成的 tibble 中，.csv 中缺失的数据被设置为了 NA。

```
## # A tibble: 270 x 12
##         No gender   age   sbp   dbp weight height smoke drunk   FPG    TG
##      <dbl> <chr>  <int> <int> <int>  <dbl>  <dbl> <chr> <chr> <dbl> <dbl>
##  1 2.01e11 男        56    95    60   66.3   171  是    是     5.01  0.74
##  2 2.01e11 女        27   112    58   54.5   172  否    无     4.44  0.75
##  3 2.01e11 男        NA   131    92   74.9   170. <NA>  <NA>   5.95  0.66
##  4 2.01e11 女        NA    98    68   67.3   158  <NA>  <NA>   5.6   0.75
##  5 2.01e11 男        64   141    81   59.9   159  否    是     5.93  1.76
##  6 2.01e11 男        43   100    69   59.5   168. 否    无     5.68  5.03
##  7 2.01e11 女        59   107    57   47.8   160. 否    无     6.45  0.84
##  8 2.01e11 男        64   137    80   72.1   176. 否    无     5.34  0.83
##  9 2.01e11 男        NA   128    87   72.7   162. 否    是     5.41  0.97
## 10 2.01e11 女        NA   127    81   58.5   154. 否    无     5.34  0.98
## # ... with 260 more rows, and 1 more variable: HDLC <dbl>
```

首先计算 BMI，然后根据实验要求中的说明判断每个观测是否符合代谢综合症的诊断标准。BMI 的计算公式为体重 (kg) /身高 (m) 的平方。而根据代谢综合症的诊断标准，具备以下 4 项中的 3 项及以上即为代谢综合症:

- **超重**：$BMI \geq 25.0\ kg/m^2$;
- **高血糖**：$FPG \geq 6.1\ mmol/L$ 或已确诊糖尿病并治疗者;
- **高血压**：收缩压 $SBP \geq 140\ mmHg$ 或舒张压 $DBP \geq 90\ mmHg$, 或已确诊高血压并治疗者;
- **空腹血**：甘油三脂 $TG \geq 1.7\ mmol/L$ 或 $HDL - C < 0.9\ mmol/L$（男），$HDL - C < 1.0\ mmol/L$（女）.

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

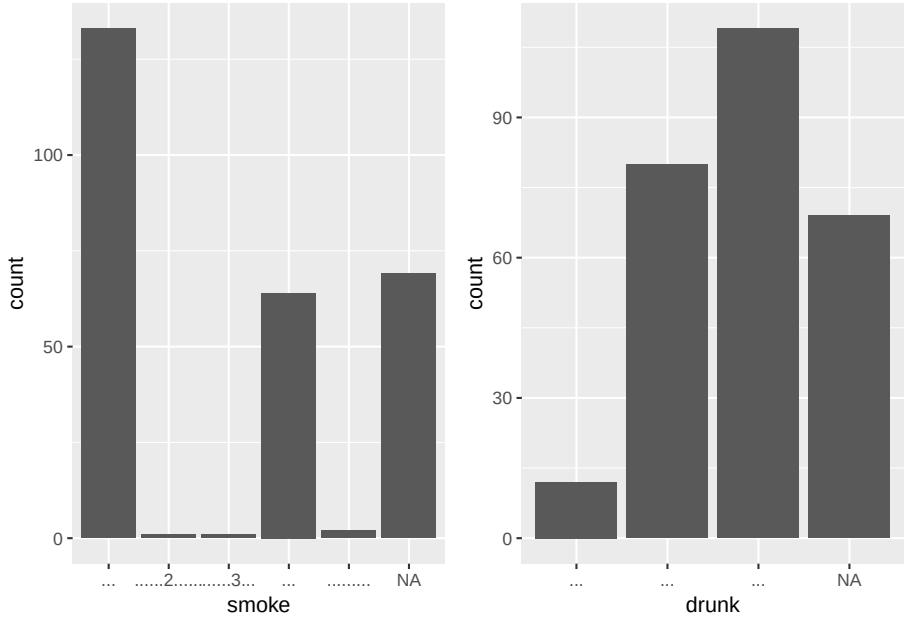
```r
# Symptoms to diagnose Metabolic syndrome
Overweight <- DataWithBMI$BMI >= 25
HighBloodSugar <- DataWithBMI$FPG >= 6.1
Hypertension <- DataWithBMI$sbp >= 140 | DataWithBMI$dbp >= 90
FastingBlood <- with(DataWithBMI, TG > 1.7 |
                       (gender == "男" & HDLC < 0.9) |
                       (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))
```

2.2 异常数据处理

2.2.1 抽烟、饮酒

首先对抽烟、饮酒的数据进行观察。



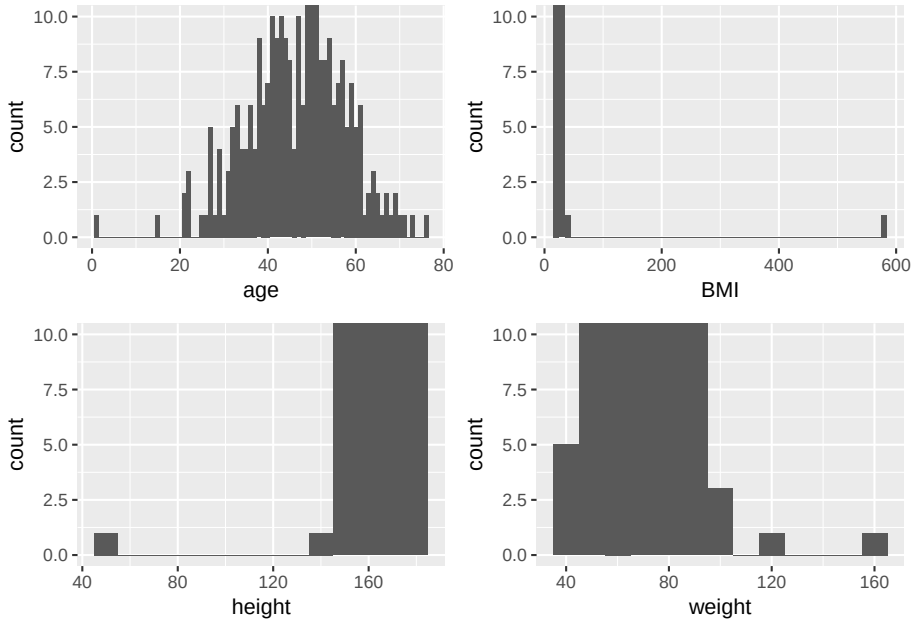
从上图可以看出，对于 smoke 变量，应把戒烟 2 个月、戒烟 3 年、已戒烟归到“是”当中；对于 drunk 变量，应把“无”改为“否”。可运行如下代码进行转换：

```
# deal with the smoke variable
CompleteData <- CompleteData %>%
  mutate(smoke = ifelse(smoke == "戒烟3年", TRUE, smoke)) %>%
  mutate(smoke = ifelse(smoke == "戒烟2个月", TRUE, smoke)) %>%
  mutate(smoke = ifelse(smoke == "已戒烟", TRUE, smoke)) %>%
  mutate(smoke = ifelse(smoke == "是", TRUE, smoke)) %>%
  mutate(smoke = ifelse(smoke == "否", FALSE, smoke))

# deal with the drunk variable
CompleteData <- CompleteData %>%
  mutate(drunk = ifelse(drunk == "是", TRUE, drunk)) %>%
  mutate(drunk = ifelse(drunk == "否", FALSE, drunk)) %>%
  mutate(drunk = ifelse(drunk == "无", FALSE, drunk))
```

2.2.2 Age, Height, Weight, BMI

同样是先对数据进行观察，为了方便观察离群点，可对 y 轴的尺度进行放大。



如图所示，需要注意的异常数据为：年龄小于 10，BMI 大于 500，身高小于 60。而对于体重，尽管出现了 160kg 的离群点，但根据常识，这仍处于可接受范围内。因此，先调出上述异常数据进行观察。

```
# Observe the unusual values
CompleteData %>%
  filter((age < 10) | (BMI > 500) | (height < 60))
```

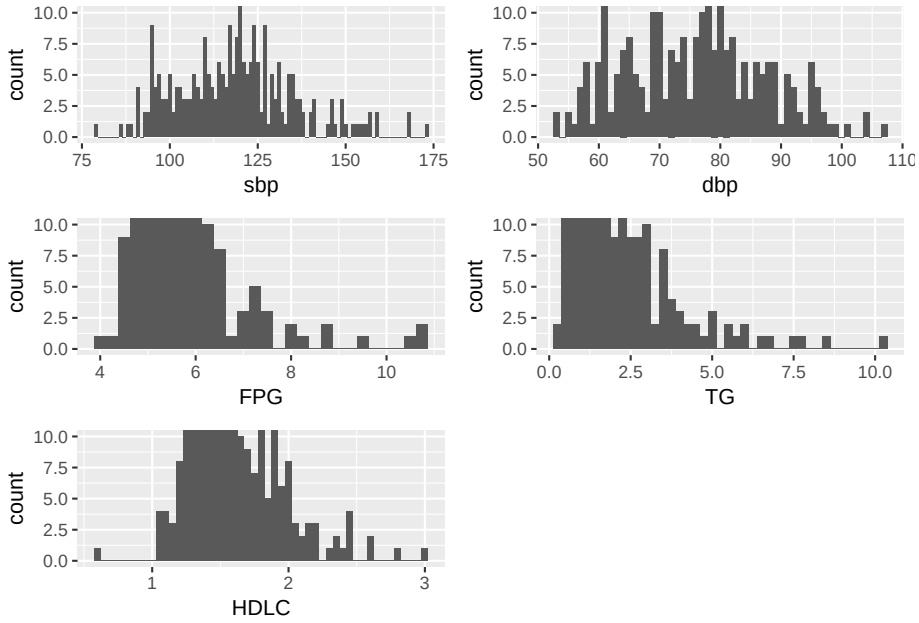
```
## # A tibble: 2 x 14
##        No gender   age   sbp   dbp weight height smoke drunk   FPG    TG
##     <dbl> <chr>  <int> <int> <int>  <dbl>  <dbl> <chr> <chr> <dbl> <dbl>
## 1 2.01e11 女         1   127    92   62.9   161   FALSE FALSE  5.73  1.84
## 2 2.01e11 女        56   108    64  164.     53.2 FALSE FALSE  5.17  1.28
## # ... with 3 more variables: HDLC <dbl>, BMI <dbl>, MS <lgl>
```

经观察，上述数据的确不符合医学规律，应从原数据中剔除。

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

2.2.3 SBP, DBP, FPG, TG, HDLC

同样是先对数据进行观察，可以根据数据的范围对坐标轴的尺度进行调整。



如上所示，少数的几个离群点均在可接受范围内，因此不需要进行额外的处理。

2.3 分析 X 各变量之间的相关性

为了对 X 进行分析，选取出变量 sbp, dbp, FPG, TG, HDL-C, BMI，然后去除缺失值。

```r
X <- CompleteData %>%
  select(sbp, dbp, FPG, TG, HDLC, BMI) %>%
  na.omit()
```

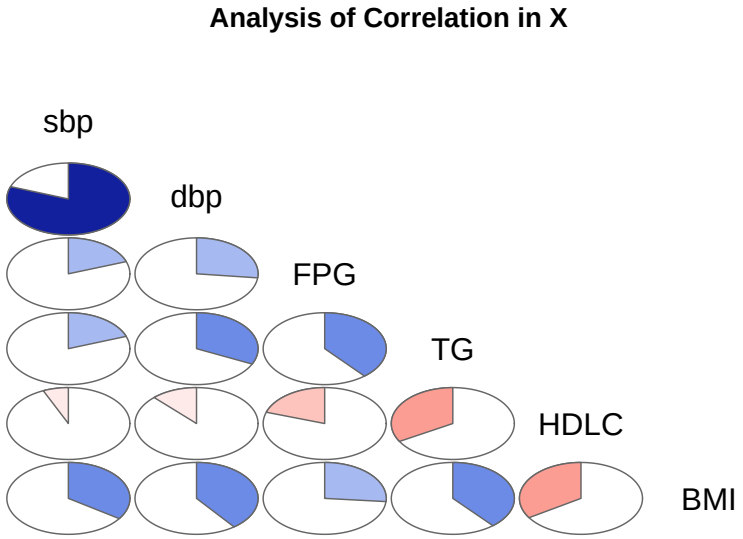
对数据 X 计算相关系数如下：

```
cor(X, method = "pearson")
```

```
##              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
```

可以用相关图对其进行可视化，如下图所示。其中蓝色、红色分别代表正、负相关，且当相关性越强时，颜色越深。

```
corrgram(X, text.panel = panel.txt,
         lower.panel = panel.pie, upper.panel = NULL,
         main = "Analysis of Correlation in X")
```



2.4 分析代谢综合症与性别、吸烟、饮酒的关系

使用二维列联表进行分析。由于用于生成列联表的 xtabs() 函数会自动忽略缺失值，所以不用先进行缺失值的处理。

2.4.1 代谢综合症与性别

用 xtabs() 函数生成列联表：

```
MSWithGender <- xtabs(~ gender + MS, CompleteData)
MSWithGender
```

```
##        MS
## gender FALSE TRUE
##      男   129   29
##      女    85    2
```

对性别与代谢综合症进行卡方独立性检验：

```
chisq.test(MSWithGender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  MSWithGender
## X-squared = 11.674, df = 1, p-value = 0.0006338
```

如上所示，$p < 0.01$，拒绝代谢综合症与性别相互独立的假设。

2.4.2 代谢综合症与吸烟

用同样的方法对吸烟与代谢综合症进行卡方独立性检验：

```
MSWithSmoke <- xtabs(~ smoke + MS, CompleteData)
chisq.test(MSWithSmoke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  MSWithSmoke
## X-squared = 4.2963, df = 1, p-value = 0.0382
```

如上所示，$p > 0.01$，没有足够理由说明代谢综合症与吸烟之间是不独立的。

2.4.3 代谢综合症与饮酒

```
MSWithDrunk <- xtabs(~ drunk + MS, CompleteData)
chisq.test(MSWithDrunk)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  MSWithDrunk
## X-squared = 9.4567, df = 1, p-value = 0.002104
```

同样，$p < 0.01$，应拒绝代谢综合症与饮酒相互独立的假设。

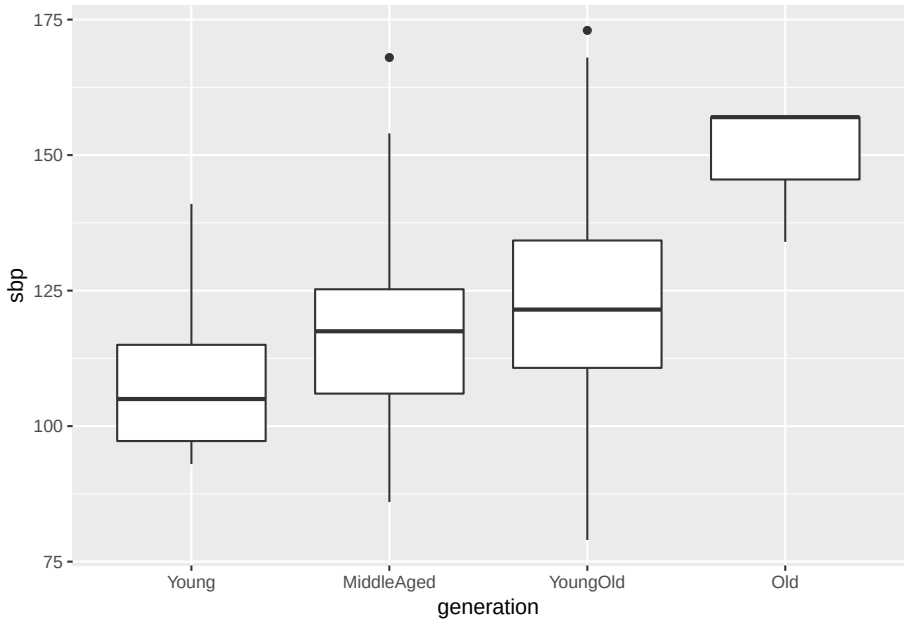
2.5 分析 X 各指标是否有年龄上的差异

首先对 X 根据年龄分组, 具体做法如下：首先从 CompleteData 提取出相关的变量，然后去除缺失值，之后根据年龄生成一个 generation 有序变量，以便根据这个变量对数据进行分组。

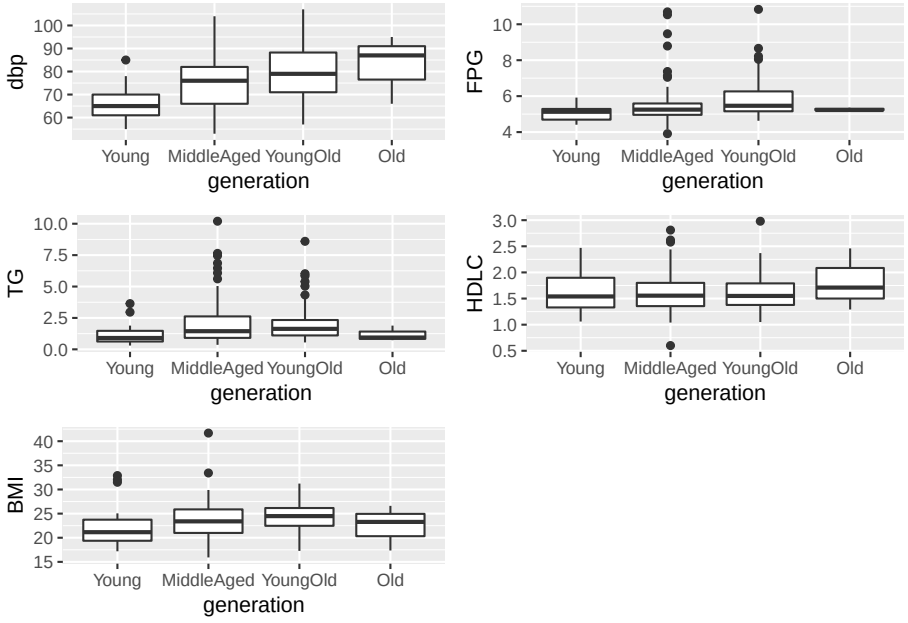
```
XWithGeneration <- CompleteData %>%
  select(age, sbp, dbp, FPG, TG, HDLC, BMI) %>%
  na.omit() %>%
  mutate(generation =
           ifelse(age <= 30, "Young",
           ifelse(age <= 50, "MiddleAged",
           ifelse(age <= 70, "YoungOld", "Old")))) %>%
  mutate(generation = factor(generation, ordered = TRUE,
           levels = c("Young", "MiddleAged", "YoungOld", "Old")))
```

以下使用箱线图来分析 X 中的各个指标是否有年龄段上的差异。以 SBP 为例，可以明显看出，随着年龄的增长，SBP 的均值有明显上升，所以可以认为 SBP 有年龄上的差异。

```
ggplot(data = XWithGeneration,
       mapping = aes(x = generation, y = sbp)) +
  geom_boxplot()
```



其他指标的箱线图如下，从图中可以看出，有明显年龄段差异的指标为 dbp。



2.6 计算 X 样本均值、样本离差阵、样本协方差和样本相关阵

2.6.1 样本均值

```
colMeans(X)
```

```
##         sbp         dbp         FPG          TG        HDLC         BMI
## 118.685714  76.000000   5.558490   1.947429   1.607755  23.798712
```

2.6.2 样本离差阵

```
cov(X) * (nrow(X) - 1)
```

```
##              sbp       dbp       FPG         TG        HDLC         BMI
## sbp  70174.80000 38192.000 788.33371 1218.77200  -94.17286 5043.79052
## dbp  38192.00000 32010.000 732.24000 1356.70000 -115.23000 3905.21282
## FPG    788.33371   732.240 233.23154  140.23565  -16.18103  222.62578
## TG    1218.77200  1356.700 140.23565  560.24068  -41.94691  505.57782
## HDLC   -94.17286  -115.230 -16.18103  -41.94691   28.33267  -99.87948
## BMI   5043.79052  3905.213 222.62578  505.57782  -99.87948 3043.86079
```

2.6.3 样本协方差阵

```
cov(X)
```

```
##               sbp         dbp        FPG         TG       HDLC        BMI
## sbp  287.6016393 156.5245902  3.2308759  4.9949672 -0.3859543 20.6712726
## dbp  156.5245902 131.1885246  3.0009836  5.5602459 -0.4722541 16.0049706
## FPG    3.2308759   3.0009836  0.9558670  0.5747363 -0.0663157  0.9124007
## TG     4.9949672   5.5602459  0.5747363  2.2960684 -0.1719136  2.0720402
## HDLC  -0.3859543  -0.4722541 -0.0663157 -0.1719136  0.1161175 -0.4093421
## BMI   20.6712726  16.0049706  0.9124007  2.0720402 -0.4093421 12.4748393
```

2.6.4 样本相关阵

```
cor(X)
```

```
##               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
```

2.7 分析 X_2, X_3 是否服从正态分布

实验中使用 W 检验来对 X_2, X_3 的正态性进行检验。W 即 Shapiro-Wilk 检验法，该方法计算 W 统计量。先分别对 X_2, X_3 进行正态性检验。

```
shapiro.test(X$FPG)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X$FPG
## W = 0.7525, p-value < 2.2e-16
```

```
shapiro.test(X$sbp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X$sbp
## W = 0.97998, p-value = 0.001561
```

如上所示，两个检验计算出的 p-value 都十分小，因此拒绝原假设，认为两个变量各自都不服从正态分布。可以预见，X_2, X_3 的联合分布也应该不服从二维正态分布。下面的检验也反映出一致的信息：

```
X %>%
  select(FPG, sbp) %>%
  t() %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.75849, p-value < 2.2e-16
```