多元-03-2020270026

2020270026 王姿文

3/15/2021

1.

根据$Hotelling\ T^2$统计量，可得μ的置信度为$1-\alpha$的置信域为

$$(\mu-\bar{x})^T(\frac{1}{n-1}\frac{1}{n}M_c^TM_c)^{-1}(\mu-\bar{x}) \le \frac{p(n-1)}{(n-p)}F_{1-\alpha}(p,n-p)$$

并可以此推出$a^T\mu$在多元时有置信度为$1-\alpha$的联立置信域

$$a^T\bar{x} \pm \sqrt{\frac{p(n-1)}{n-p}F_{1-\alpha}(p,n-p)}\sqrt{a^T\frac{1}{n-1}\frac{1}{n}M_c^TM_c a},\ \forall\ a \in \mathbb{R}$$

。

因此根据本题为$(\mu-\bar{x})^T\hat{\Sigma}^{-1}(\mu-\bar{x}) = \frac{1}{n}\chi_{1-\alpha}(2)$，可以推出
$\frac{1}{n-1}\frac{1}{n}M_c^TM_c = \hat{\Sigma},\ \frac{p(n-1)}{(n-p)}F_{1-\alpha}(p,n-p) = \frac{1}{n}\chi_{1-\alpha}(2)$。
故联合边界的方程为

$$\bar{x} \pm \sqrt{\hat{\Sigma}\frac{1}{n}\chi^2_{1-\alpha}(2)}$$
$$= \bar{x} \pm \sqrt{(1+\frac{1}{n})\frac{p(n-1)}{n-p}F_\alpha(p,n-p)}\hat{L}\begin{bmatrix} cos(\theta) \\ sin(\theta) \end{bmatrix},\ \forall\ \hat{L}是\hat{\Sigma}的cholesky\,factor$$

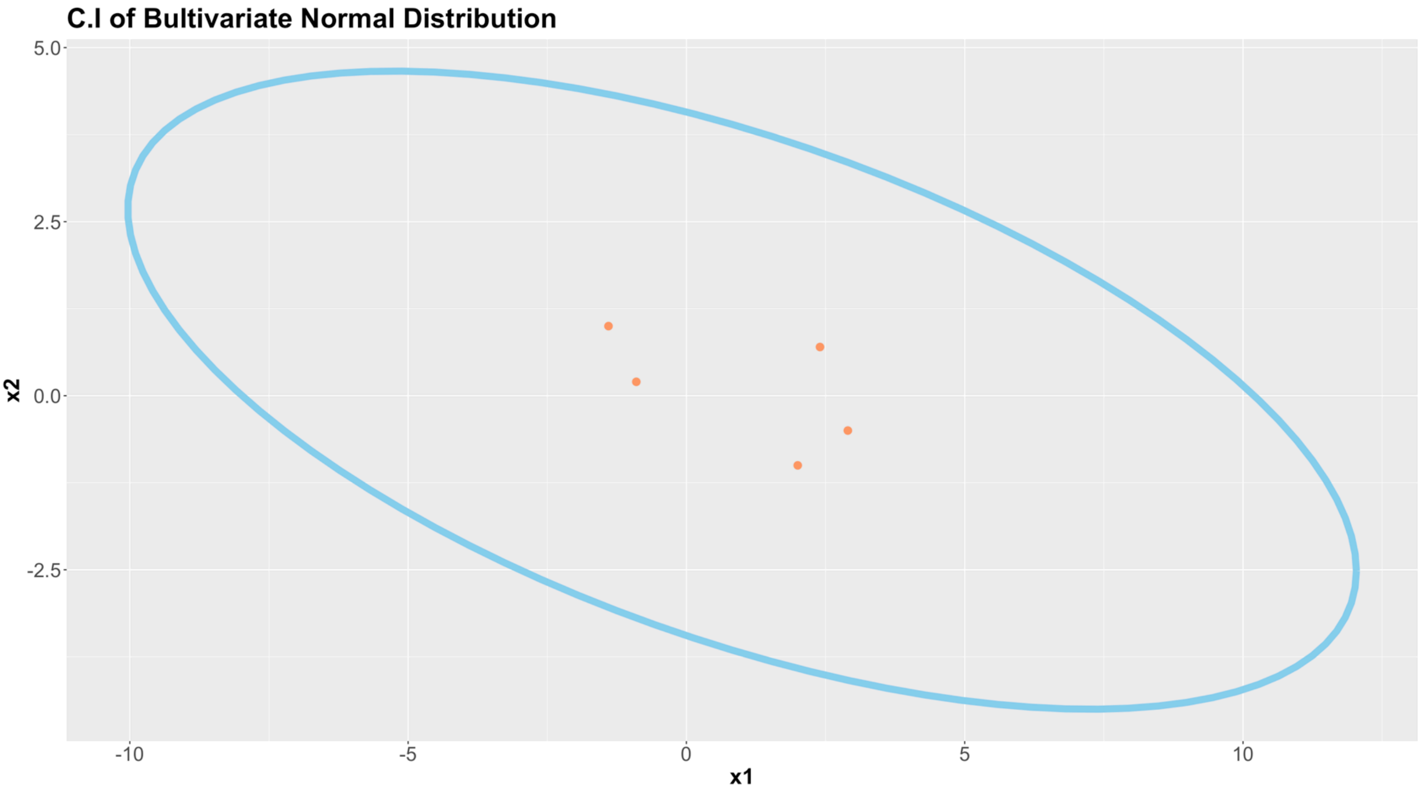
。

下为作为范例的椭圆边界图：

```r
#C.I function
pred.int.mvnorm <- function(x, alpha=.05) {
  p <- ncol(x)
  n <- nrow(x)
  Sigmahat <- var(x)
  xbar <- apply(x,2,mean)
  xbar
  theta <- seq(0, 2*pi, length=100)
  polygon <- xbar +
    sqrt(p*(n-1)/(n-p)*(1 + 1/n)*qf(alpha, p, n - p, lower.tail = FALSE))*
    t(chol(Sigmahat)) %*%
    rbind(cos(theta), sin(theta))
  t(polygon)
}

#data
x <- matrix(c(-0.9,2.4,-1.4,2.9,2.0,0.2,0.7,1.0,-0.5,-1.0),ncol=2)
dt <- as.data.frame(pred.int.mvnorm(x))
dtx <- as.data.frame(x)

#plot
ggplot(data=dt, aes(x = V1,y = V2)) +
  geom_path(color = 'skyblue',size = 3)+
  geom_point(data = dtx, aes(x = V1,y = V2),color = '#FF9966',size = 3) +
  theme(plot.title = element_text(size=25, face="bold"),
        axis.title = element_text(size=20, face="bold"),
        axis.text = element_text(size=18)) +
  labs(title ="C.I of Bultivariate Normal Distribution",
       x = 'x1', y = 'x2')
```



2.

2.1

以下令H_0：$baseline\ is\ normal\ distribution$来做正态性检验：

a. 拟合优度卡方检验

使用分段后比较观测频数和理论期望频数，下表可见分为7段而$p-value>0.05$，故接受H_0假设，推断在$\alpha=0.05$时$baseline \sim Normal\ Distribution$。

```
data(cd4,package = 'boot')
pchiTest(cd4$baseline, description='baseline')
```

```
##
## Title:
##  Pearson Chi-Square Normality Test
##
## Test Results:
##   PARAMETER:
##     Number of Classes: 7
##   STATISTIC:
##     P: 3.1
##   P VALUE:
##     Adhusted: 0.5412
##     Not adjusted: 0.7962
##
## Description:
##  baseline
```

b. Jarque-Bera偏度峰度检验:

JB = $\frac{n}{6}(偏度^2-\frac{1}{4}峰度^2)$
下表可见$p-value>0.05$，故接受H_0假设，推断在$\alpha=0.05$时$baseline \sim Normal\ Distribution$。

```
jbTest(cd4$baseline, title='baseline')
```

```
##
## Title:
##  baseline
##
## Test Results:
##   PARAMETER:
##     Sample Size: 20
##   STATISTIC:
##     LM: 0.389
##     ALM: 0.377
##   P VALUE:
##     LM p-value: 0.792
##     ALM p-value: 0.814
##     Asymptotic: 0.823
##
## Description:
##  Mon Mar 15 16:52:25 2021 by user:
```

c. Kolmogorov-Smirnov检验:

比较经验分布函数与理论分布函数的最大差，下表可见$p - value > 0.05$，故接受H_0假设，推断在$\alpha = 0.05$时$baseline \sim Normal\ Distribution$。

```
ksnormTest( (cd4$baseline - mean(cd4$baseline))/sd(cd4$baseline) )
```

```
##
## Title:
##  One-sample Kolmogorov-Smirnov test
##
## Test Results:
##   STATISTIC:
##     D: 0.0999
##   P VALUE:
##     Alternative Two-Sided: 0.9884
##     Alternative      Less: 0.7717
##     Alternative   Greater: 0.6707
##
## Description:
##  Mon Mar 15 16:52:25 2021 by user:
```

d. Shapiro-Wilk检验:

基于QQ图思想，下表可见$p - value > 0.05$，故接受H_0假设，推断在$\alpha = 0.05$时$baseline \sim Normal\ Distribution$。

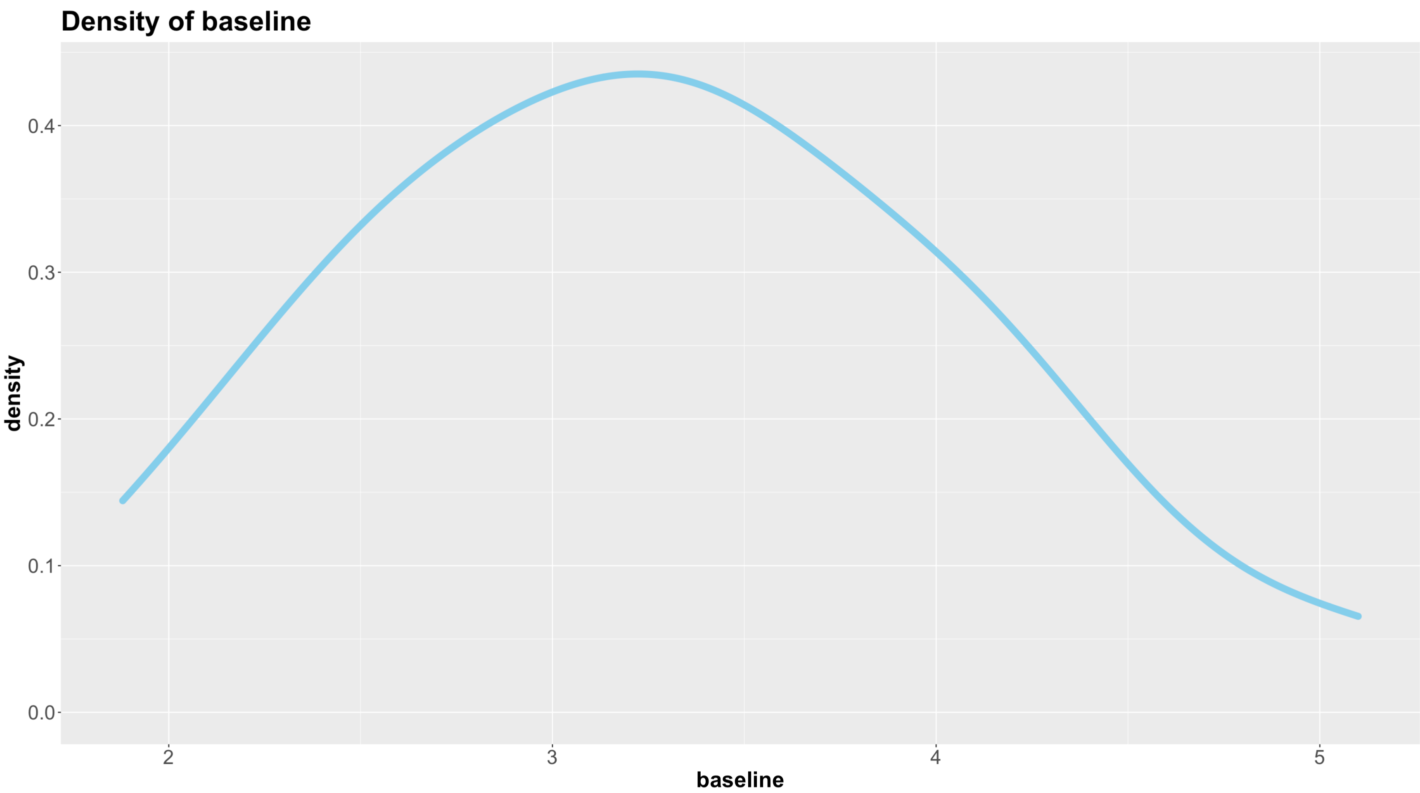
```
shapiro.test(cd4$baseline)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cd4$baseline
## W = 0.98075, p-value = 0.9434
```

e. Plot

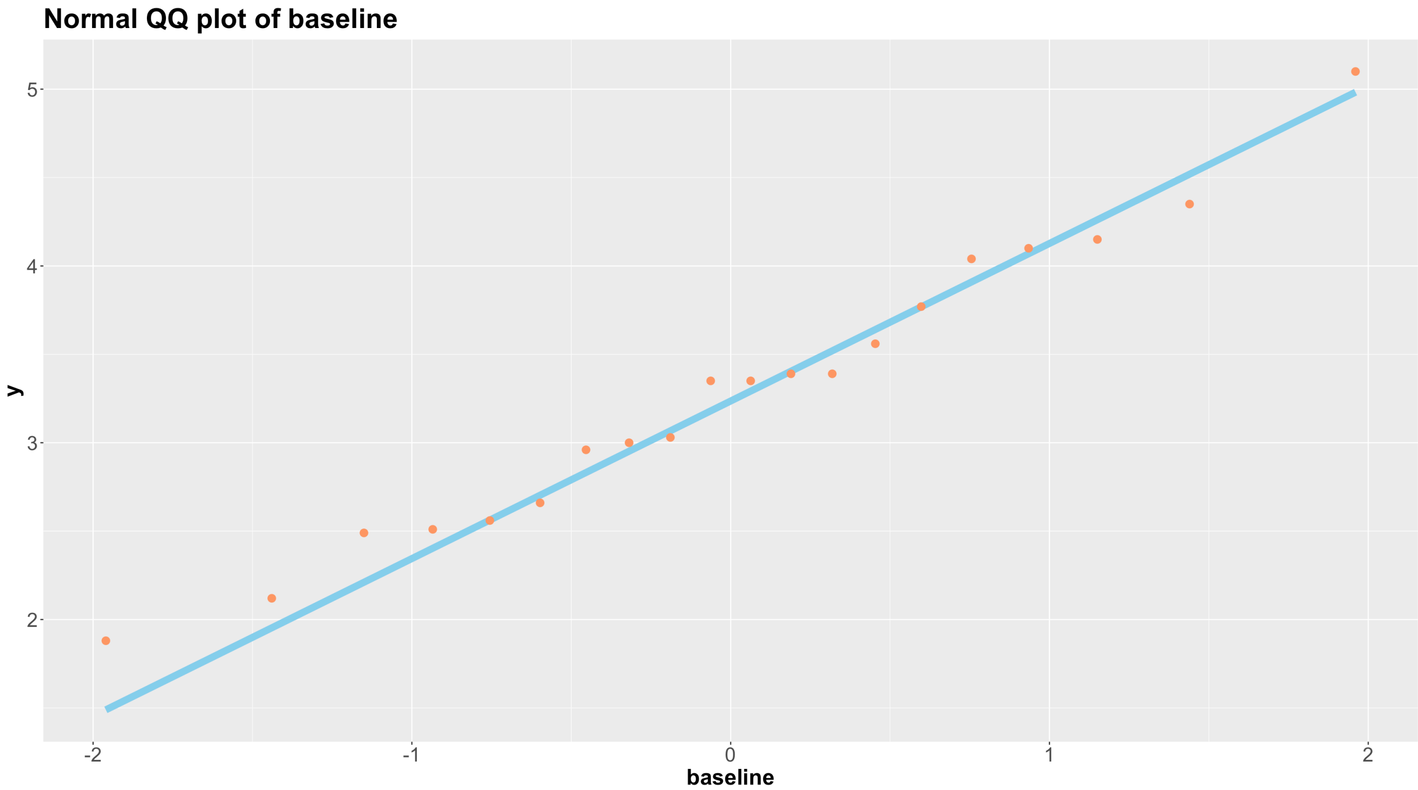
由于上述检验结果都显示 `baseline` 为正态，因此可以画图来检验下，下图确实符合正态分布但略为右偏。

```
ggplot(cd4,aes(x = baseline,y = ..density..)) +
    geom_density(color = 'skyblue',size = 3) +
    theme(plot.title = element_text(size=25, face="bold"),
                axis.title = element_text(size=20, face="bold"),
                axis.text = element_text(size=18)) +
    labs(title =paste('Density of baseline'))
```



用QQ图来画则很显然符合正态分布。

```
ggplot(cd4,aes(sample = baseline)) +
  stat_qq_line(color = 'skyblue',size = 3) +
  stat_qq(color = '#FF9966',size = 3) +
  theme(plot.title = element_text(size=25, face="bold"),
        axis.title = element_text(size=20, face="bold"),
        axis.text = element_text(size=18)) +
  labs(title =paste('Normal QQ plot of baseline'),x='baseline')
```



2.2

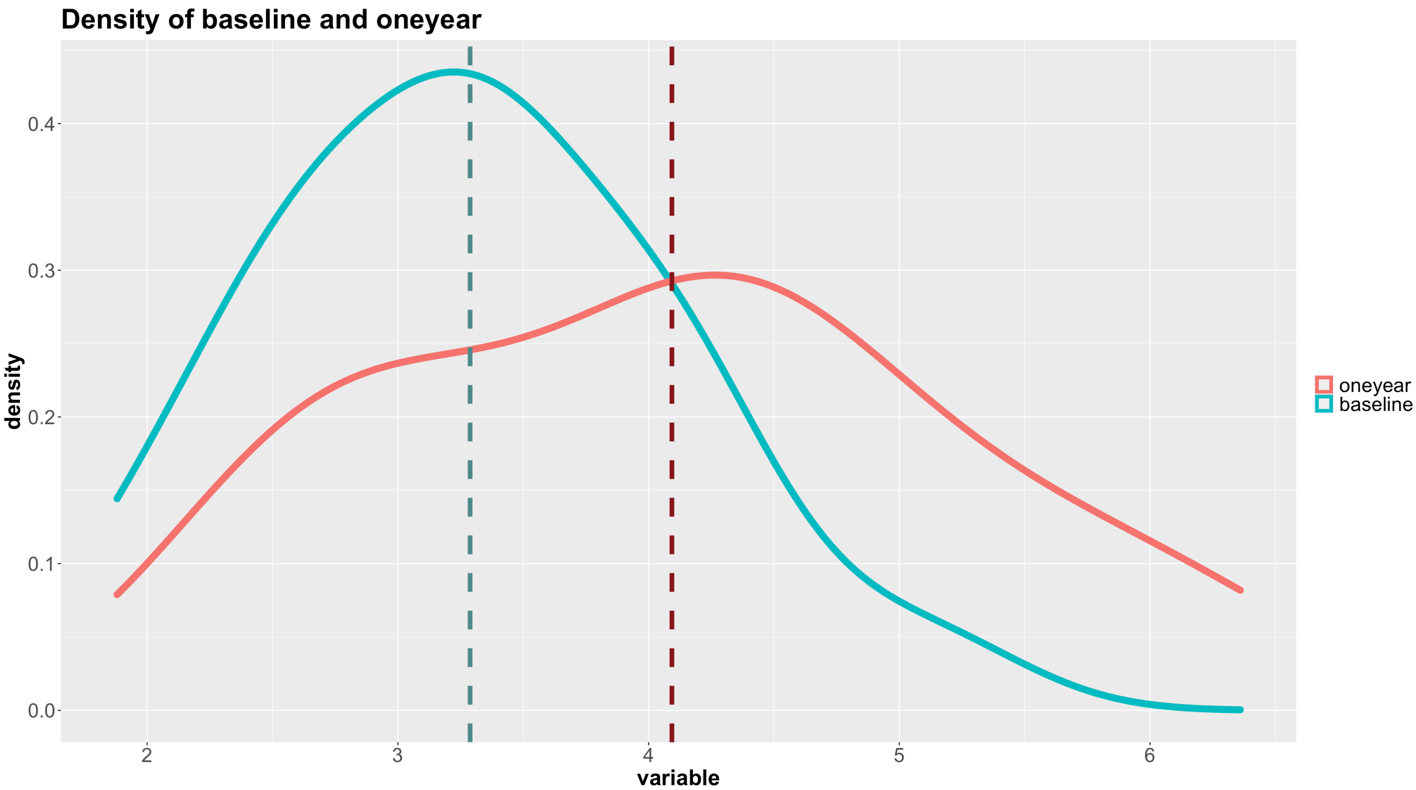
令$H_0: \mu_{baseline} = \mu_{oneyear}$，下表为成对t检验的结果，可以看出$p-value < 0.05$，拒绝$H_0$，因此两个变量的均值有显著差异，下表也能看出两个变量分别的均值。

```
t.test(cd4$baseline, cd4$oneyear)
```

```
##
##  Welch Two Sample t-test
##
## data:  cd4$baseline and cd4$oneyear
## t = -2.544, df = 33.983, p-value = 0.01568
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4480791 -0.1619209
## sample estimates:
## mean of x mean of y
##     3.288     4.093
```

下图为两个变量的分布和均值。

```
ggplot(cd4) +
  geom_density(aes(x = baseline,y = ..density..,color = 'skyblue'),size = 3,show.l
egend = TRUE) +
  geom_density(aes(x = oneyear,y = ..density..,color = 'blue'),size = 3,show.legen
d = TRUE) +
  theme(plot.title = element_text(size=25, face="bold"),
        axis.title = element_text(size=20, face="bold"),
        axis.text = element_text(size=18),
        legend.title=element_blank(),
        legend.text = element_text(size=18)) +
  labs(title =paste('Density of baseline and oneyear'),
       x = 'variable') +
  scale_colour_discrete(breaks=c("blue", "skyblue"),
                        labels=c("oneyear", "baseline")) +
  geom_vline(xintercept = 3.288,linetype = 2,color = 'darkslategray4',size = 2) +
  geom_vline(xintercept = 4.093,linetype = 2,color = 'firebrick4',size = 2)
```



Conclusion

数据的正态与否会影响后续是否需要做数据处理，cd4的数据则没有太大问题。