**生统第三次作业**

一、实验室欲购进一批灯泡，打算在两个供货商之间选择一家购买。选购考虑的主要因素就是灯泡使用寿命的方差大小，为此需要对供货商提供的20个样品进行检测，得到的数据如下表所示。(20’)

|  |  |
| --- | --- |
| 供货商1 | 供货商2 |
| 6802 | 5884 |
| 5730 | 5871 |
| 5823 | 5797 |
| 5915 | 5957 |
| 5774 | 5803 |
| 5880 | 5862 |
| 5870 | 5814 |
| 5773 | 5885 |
| 5830 | 5856 |
| 5841 | 5940 |
| 5763 | 5945 |
| 5851 | 5803 |
| 5789 | 5864 |
| 5796 | 5851 |
| 5818 | 5714 |
| 5685 | 5943 |
| 5602 | 5830 |
| 5841 | 5858 |
| 5723 | 5922 |
| 5757 | 5866 |

1、检验两家供货商的灯泡使用寿命的方差有无显著差异（ɑ=0.05）（10’）

2、择最合适的检验方法检验两家供应商的灯泡使用寿命有无差别。（10’）

解：

1、检验两家供货商的灯泡使用寿命的方差有无显著差异（ɑ=0.05）

解：求两个供货商提供的样本方差是否存在显著差异，因而为双侧检验问题。建立的原假设与备择假设为：

C:\Users\hp\AppData\Local\Temp\1554026574(1).png

R code:

> a<-c(6802,5730,5823,5915,5774,5880,5870,5773,5830,5841,5763,5851,5789,5796,5818,5685,5602,5841,5723,5757)

> b<-c(5884,5871,5797,5957,5803,5862,5814,5885,5856,5940,5945,5803,5864,5851,5714,5943,5830,5858,5922,5866)

> n1<-20

> n2<-20

> var1<-var(a)

> var2<-var(b)

> F<-var1/var2

> F

[1] 15.2795

> p<-2\*(1-pf(F,df1=(n1-1),df2=(n2-1)))

> p

[1] 1.799681e-07

由于p= 1.799681e-07<0.05,因此拒绝原假设，两家供货商灯泡使用寿命的方差有显著差异。

2、选择最合适的检验方法检验两家供应商的灯泡使用寿命有无差别。

解：

使用双样本t.test

原假设：两家供应商的灯泡使用寿命无差别。

备择假设：两家供应商的灯泡使用寿命有差别。

因为两组方差不相等，使用方差不相等的两样本t.test

R code：

> t.test(a,b,var.equal=FALSE)

Welch Two Sample t-test

data: a and b

t = -0.36748, df = 21.476, p-value = 0.7169

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-133.69464 93.49464

sample estimates:

mean of x mean of y

5843.15 5863.25

因为p-value =0.7169> 0.05，因此认为原假设成立，即两家供应商的灯泡使用寿命无差别。

二、R language application (25’)

Please use R to resolve the following issues and display your R code and results.

1. For a normal random variable X with mean 4.0, and standard deviation 1.0,

1. find the probability that X is less than 2.0. (4’)
2. find the value K so that P(X>K) = 0.05. (4’)

2. When tossing a fair coin 8 times,

1. find the probability of seeing no heads (Hint: this is a binomial distribution.) (3’)
2. find the probability of seeing exactly 4 heads. (3’)
3. find the probability of seeing more than 5 heads. (3’)

3. Simulate a sample of 1000 random data points from a normal distribution with mean 100 and standard deviation 8, and store the result in a vector.

1. plot a histogram and a boxplot of the vector you just created. (4’)
2. using the data above, test the hypothesis that the mean equals 100 (using t.test). (4’)

解：

1

1. pnorm(2.0, 4,1)

0.02275013

1. qnorm(0.95, 4, 1) or qnorm(0.05, 4, 1, lower.tail=FALSE)

5.644854

2

1. dbinom(x = 0, size = 8, prob = 0.5)

0.00390625

1. dbinom(x = 4, size = 8, prob = 0.5)

0.2734375

1. 1-pbinom(q=5, size=8, prob=0.5)

0.1445312

3

1. x <- rnorm(n=1000, mean=100, sd=8)

par(mfrow=c(1,2))

hist(x)

boxplot(x)



1. t.test(x, mu=100)

One Sample t-test

data: x

t = 0.63008, df = 999, p-value = 0.5288

alternative hypothesis: true mean is not equal to 100

95 percent confidence interval:

99.66389 100.65403

sample estimates:

mean of x

100.159

三、R language application. (25’)

（Please use R to read-in and manipulate data, code and results should be displayed.）

In order to detect air quality, a city's environmental protection department conducted a random test of PM2.5 in the air every few weeks. It is known that the average value of PM2.5 per cubic meter of air in the city is 82ug/m3. In the most recent test, the value of PM2.5 per cubic meter of air(ug/m3) is shown in the homework3\_data.

1)Show your work directory (2’)

2)use R to read in the data（3’）

3)use boxplot to show the PM2.5 distribution in every month (10’)

4)get the data of month equal to March and store in data\_march (10’)

解：

## a)

getwd()

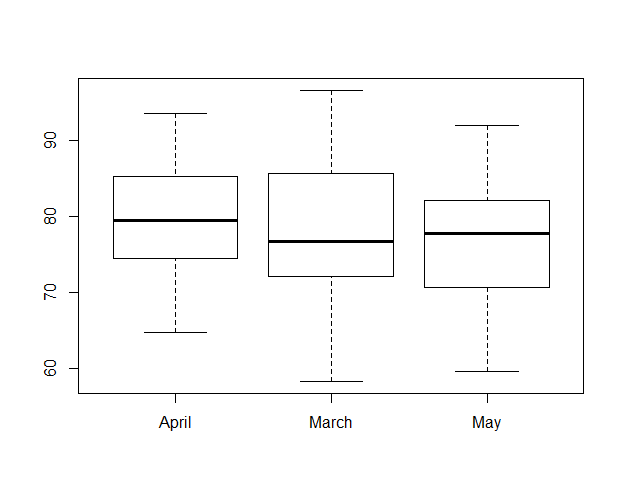
[1] "C:/Users/Chunhui Xu/Desktop"

## b)

data <- read.table("homework3\_data.txt", header = T)

## c)

boxplot(data$PM2.5\_score ~ data$Month)



## d)

data\_march <- subset(data, Month == "March")

# or

data\_march <- data[data$Month == "March",]

# or ....

2). According to recent measurements in March, when the significance level α = 0.05, whether the average PM2.5 value per cubic meter of air in the city is significantly lower than the past value (5’).

H0: u >= 82; H1: u < 82

z <- (mean(data\_march$PM2.5\_score)-82)/(sd(data\_march$PM2.5\_score)/sqrt(32))

pnorm(z)

[1] 0.008898618

So we reject H0, accept H1, the average PM2.5 value in March is lower than the past value.

3). According to recent measurements in April and May, assume that σApril2 = σMay2, Is there a significant difference in PM2.5 values for two months? (5’).

H0: uApril = uMay; H1: uApril != uMay

t.test(data[data$Month == "April",]$PM2.5\_score, data[data$Month == "May",]$PM2.5\_score, var.equal = T)

Two Sample t-test

data: data[data$Month == "April", ]$PM2.5\_score and data[data$Month == "May", ]$PM2.5\_score

t = 1.1696, df = 54, p-value = 0.2473

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.629734 6.194020

sample estimates:

mean of x mean of y

79.16429 76.88214

So we accept H0, there a significant difference in PM2.5 values for two months.

四、Suppose we draw a sample of size 20 of birthweights from a hospital, the details can be found in the homework data. The mean of national-wide birthweights is 118.(30’)

1)What is the probability that the mean birthweight of the sample falls between 100.0 and 126.0? Please list the formulas to calculate this and also the R code for it.(5’)

2)What is the 95% confidence interval of the sample mean?(5’)

3)Can we say the underlying mean birthweight from this hospital is higher than the national average?

Please list the formulas for this and also the R code for it.(5’)

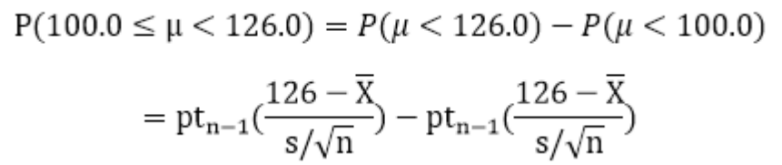
4)Test the hypothesis that the mean birthweight of sample size 20 is different from the national average (Significance level 0.05). Please list the formulas for this and also the R code for it.(5’)

5)Compute the power of the test performed in (4) with significance level 0.05.(5’)

6)To see the significance difference between the sample mean and the national mean and ensure the

type II error to be β=0.05, what is the appropriate sample size with significance level is 0.01?(5’)

解：



# 题目样本数据20且总体方差未知，故采用t分布进行计算

#（1）7分 公式2分，R代码3分，结果2分

m<‐mean(data$Birthweight)

s<‐sd(data$Birthweight)

n<‐20

p.upper<‐pt((126.0‐118)/(s/sqrt(n)),n‐1)

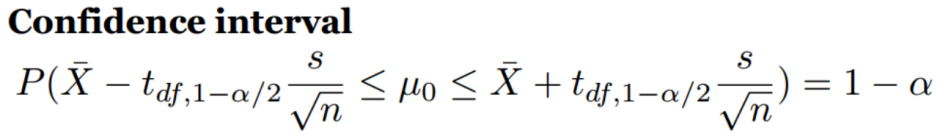
p.lower<‐pt((100.0‐118)/(s/sqrt(n)),n‐1)

p<‐p.upper‐p.lower

p

#[1] 0.9961359

So the probability that the mean birthweight of the sample falls between 100.0 and 126.0 is 0.9961359.



#(2) 6分 R代码4分，结果2分

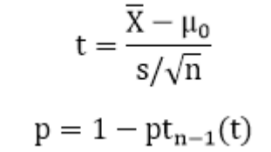
left<‐m‐qt(0.975,n‐1)\*s/sqrt(n)

right<‐m+qt(0.975,n‐1)\*s/sqrt(n)

left;right

#[1] 114.3777 125.6223

So the 95% confidence interval of the sample mean is [114.378,125.622]



#(3) 7分 公式2分，R代码3分，结果2分

#使用单尾检验

t<‐(m‐118)/s\*sqrt(n)

p<‐1‐pt(t,n‐1)

P

#[1] 0.2328299

#P=0.2328299 >0.05

##使用t.test()##

t.test(data$Birthweight, alternative = "greater", mu= 118)

# One Sample t‐test

#data: data$Birthweight

#t = 0.74454, df = 19, p‐value = 0.2328

#alternative hypothesis: true mean is greater than 118

#95 percent confidence interval:

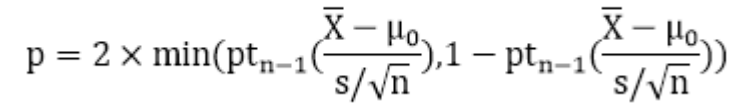
# 115.3552 Inf

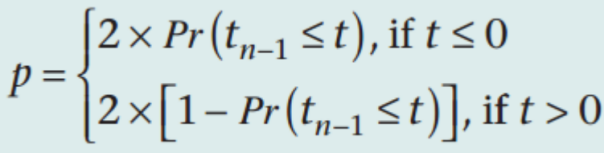
#sample estimates:

#mean of x

# 120

So the hypothesis that the underlying mean birthweight from this hospital is higher than the national average is FALSE





#（4）7分 公式2分 R代码3分 结果2分

#使用双尾检测

#又因为t>0

p<‐(1‐pt(t,n‐1))\*2

P

#[1] 0.4656597

#P=0.4656597>0.05

##使用t.test()##

t.test(data$Birthweight, mu = 118)

# One Sample t‐test

#data: data$Birthweight

#t = 0.74454, df = 19, p‐value = 0.4657

#alternative hypothesis: true mean is not equal to 118

#95 percent confidence interval:

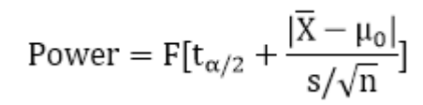
# 114.3777 125.6223

#sample estimates:

#mean of x

# 120

The hypothesis that the mean birthweight of sample size 20 is different from the national average is FALSE



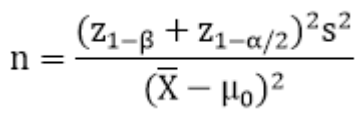
#(5) 6分 R代码4分 结果2分

power<‐pt(qt(0.025,n‐1)+abs(m‐118)/s\*sqrt(n),n‐1)

power

# [1] 0.09667934

The power of the test performed in (4) with significance level 0.05 is 0.0967.



#(6) 7分 R代码4分 结果3分

#使用t‐test计算样本量时可使用z‐test正态分布进行估算

n<‐(qnorm(1‐0.05)+qnorm(1‐0.01/2))^2\*s^2/(m‐118)^2

n

#[1] 642.7163

# n=643

#使用pwr包

pwr.t.test(d=abs(m‐118)/s,sig.level=0.01,power=1‐0.05,type="one.sample",alternative="two.sided")

# One‐sample t test power calculation

#

# n = 646.0381

# d = 0.1664842

# sig.level = 0.01

# power = 0.95

# alternative = two.sided

# n=647