**STAT 40001/MA 59800 Statistical Computing/ Computational Statistics Fall 2013**

**Test 1- Form II (Solution)**

**Name:**

**PUID:**

*This exam consists of 5 questions of worth 100 points. Please provide the R codes that have been used to perform the calculations and graphics along with the interpretation of the output. It is important that you distinguish between the R code and the description. You may use different fonts or different colors.*

**Q.N. 1)** **Short answer questions**

a) Create a vector named countby5 that is a sequence of 5 to 100 in steps of 5.

b) Provide R code to create the sequence 5, 10, 10, 15, 15, 15, 20, 20, 20, 20.

c) Generate 100 random numbers from a normal distribution with mean 10 and variance 25. Please print first 5 observations.

d) Generate 50 random numbers form a t- distribution with 18 degrees of freedom. Please print first 5 observations.

e) The brightness dataset in the UsingR package contains the information about the brightness of stars in a sector of the sky. How many observations are included in the dataset? Please print first 5 observations.

**Solution: We can use the R code below to perform the requested task.**

**a)** > countby5

[1] 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

**b)** > rep(seq(5,20,5),1:4)

[1] 5 10 10 15 15 15 20 20 20 20

**c)** > data1=rnorm(100,10,5)

> data1[1:5]

[1] 5.159660 7.463630 4.984582 20.605701 8.343471

**d)** > data2=rt(50,df=18)

> data2[1:5]

[1] 1.1979073 -0.8357669 0.6496889 -0.0744544 1.6354301

**e)**

> library(UsingR)

> data(brightness)

> length(brightness)

[1] 966

> head(brightness,5)

[1] 9.10 9.27 6.61 8.06 8.55

**Q.N. 2)** Seven subjects were assigned to group A and eight subjects were assigned to group B. The two groups using different training methods to improve the subjects’ read speed. The results are as following:

Group A: 500 700 250 404 390 555 589

Group B: 355 388 445 469 560 502 430 480

Read the data above to the R system. Perform t-test and a nonparametric test to compare the two groups A and B. Check if it is appropriate to use t-test. Compare the results of t-test and nonparametric test.

**Solution: We entered the data and perform two sample t-test and Wilcoxon rank sum test using R code below**

> A<-c(500,700,250,404,390,555,589)

> B<-c(355,388,445,469,560,502,430,480)

> t.test(A,B)

> t.test(A,B)

Welch Two Sample t-test

data: A and B

t = 0.5002, df = 7.958, p-value = 0.6305

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

95 percent confidence interval:

-109.8024 170.5524

sample estimates:

mean of x mean of y

484.000 453.625

> wilcox.test(A,B)

Wilcoxon rank sum test

data: A and B

W = 33, p-value = 0.6126

alternative hypothesis: true location shift is not equal to 0

Note that the p-value using t-test is 0.6305 and using Wilcoxon Rank Sum test is 0.6126. In both cases p-value is higher than typical level of significance 0.05. Therefore, we have enough evidence to conclude that there no difference in the training methods.

In order to check if it was appropriate to use the t test we need to check if A and B come from a normal population. We do so using the Shapiro test as below.

> shapiro.test(A)

Shapiro-Wilk normality test

data: A

W = 0.9856, p-value = 0.982

> shapiro.test(B)

Shapiro-Wilk normality test

data: B

W = 0.9903, p-value = 0.9955

For both A and B the p-value is greater than 0.05. So we fail to reject the null hypothesis that the samples come from the normal distribution. Thus our use of the t-test is justified.

Both parametric and nonparametric test confirm that there is no difference in training methods to improve the subjects’ read speed.

**Q.N. 3)** The chickwts data are collected from an experiment to compare the effectiveness of various feed supplements on the growth rate of chickens and are available in the base package.

a) How many variables are in the database?

b) Display the information by creating side-by-side boxplot by choosing an appropriate variable. Please make sure that you have appropriately labeled the axes and use suitable title of the boxplot.

**Solution: We use the R code below to perform the requested task**

**a)**

> dim(chickwts)

[1] 71 2

> head(chickwts,5)

weight feed

1 179 horsebean

2 160 horsebean

3 136 horsebean

4 227 horsebean

5 217 horsebean

Note that there are two variables (weight and feed) and 71 observations.

**b) We can display the information graphically using R code below:**

**>** boxplot(chickwts$weight~chickwts$feed,col=c(2,3,4,5,6,7)

,xlab="Feed Type",ylab="Weight",main="Boxplot of Weight")

****

**Q.N. 4)** The National Restaurant Association is interested in determining if there is a relationship between the types of pizza pie which Americans prefer and the region of the country in which they live . The association randomly selects 285 Americans and records the category of pizza pie which best describes their preference and the region of the country they live. The data are recorded as below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type of  Pizza Pie  Preferred |  | North | South | East | West |
| Thin Crust | 40 | 30 | 35 | 45 |
| Thick Crust | 17 | 15 | 21 | 22 |
| Pan Pizza | 15 | 15 | 15 | 15 |

Can the association conclude that the types of pizza pie which Americans prefer and the region of the country they live are dependent at 0.01?

Solution: We can use the R code below to perform the chi-square test of independence.

> data=c(40,30,35,45,17,15,21,22,15,15,15,15)

> m=matrix(data,nrow=3,byrow=T)

> chisq.test(m)

Pearson's Chi-squared test

data: m

X-squared = 1.7021, df = 6, p-value = 0.945

Decision: Note that p=0.945>0.01, so we fail to reject the null hypothesis and conclude that the types of pizza pie which Americans prefer and the region of the country they live are independent.

**Q.N. 5)** Journal of Statistics Education, Volume 4, Number 2 (July 1996) include an article What's Normal? -- Temperature, Gender, and Heart Rate by A. Shoemaker. The dataset used in the article are provided in <http://www.amstat.org/publications/jse/datasets/normtemp.dat.txt>. The description of the data can be accessed in the link below.

<http://www.amstat.org/publications/jse/datasets/normtemp.txt>

1. How many variables are included in the study?
2. Print first five observations of the data.
3. Is the distribution of body temperatures normal?
4. Is the true population mean really 98.6 degrees F?
5. Is there a significant difference between males and females average temperature?

Solution: We have accessed the data using R code

>data<-read.table("http://www.amstat.org/publications/jse/ datasets /normtemp.dat.txt")

> dim(data)

[1] 130 3

There are 3 variables and 130 observations

**b)**

> head(data,5)

V1 V2 V3

1 96.3 1 70

2 96.7 1 71

3 96.9 1 74

4 97.0 1 80

5 97.1 1 73

Note that new header V1, V2 and V3 are created.

V1-Body temperature (degrees Fahrenheit)

V2- Gender,

V3- Heart rate (beats per minute)

**c)**

> shapiro.test(data$V1)

Shapiro-Wilk normality test

data: data$V1

W = 0.9866, p-value = 0.2332

Decision: p-value>0.05 so we fail to reject the null hypothesis and conclude that the body temperature are normally distributed.

d) We use the R code below to test whether the average body temperature is 98.6 F

> t.test(data$V1, mu=98.6)

One Sample t-test

data: data$V1

t = -5.4548, df = 129, p-value = 2.411e-07

alternative hypothesis: true mean is not equal to 98.6

95 percent confidence interval:

98.12200 98.37646

sample estimates:

mean of x

98.24923

Decision: p-value is less than 0.05 so we reject the null hypothesis and conclude that true mean temperature is not equal to 98.6 F.

e) We extracted the male and female temperature and perform two sample t-test using R code below.

> Mdata<-subset(data, V2=="1")

> Fdata<-subset(data, V2=="2")

> Mtemp<-Mdata$V1

> Ftemp<-Fdata$V1

> t.test(Mtemp,Ftemp)

Welch Two Sample t-test

data: Mtemp and Ftemp

t = -2.2854, df = 127.51, p-value = 0.02394

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

95 percent confidence interval:

-0.53964856 -0.03881298

sample estimates:

mean of x mean of y

98.10462 98.39385

Decision: Note that p<0.05 so reject the null hypothesis and conclude that there is a significant difference in the average body temperature between male and female.