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

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

**Name:**

*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 countby2 that is a sequence of 1 to 100 in steps of 2.

b) Provide R code to create the sequence 5, 10, 10, 15, 15, 15, 20, 20, 20, 20, 25, 25, 25, 25, 25.

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

d) Generate 50 random numbers form a t- distribution with 18 degrees of freedoms. 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)** > countby2=seq(1,100,2)

> countby2

[1] 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99

**b)** > rep(seq(5,25,5),1:5)

[1] 5 10 10 15 15 15 20 20 20 20 25 25 25 25 25

**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)** 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. 3)** Nine subjects were assigned to group A and nine subjects were assigned to group B. The two groups using different training methods to improve the subjects’ read speed. The results are as following:

A: 500 230 505 404 390 200 750 700 490

B: 355 388 445 469 560 502 515 430 480

Read the data above to the R system. Perform t-test and 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 belwo**

> A<-c(500, 230, 505, 404, 390, 200, 750, 700, 490)

> B<-c(355, 388, 445, 469, 560, 502, 515, 430, 480)

> t.test(A,B)

Welch Two Sample t-test

data: A and B

t = 0.0425, df = 9.873, p-value = 0.9669

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

95 percent confidence interval:

-143.0348 148.5904

sample estimates:

mean of x mean of y

463.2222 460.4444

> wilcox.test(A,B)

Wilcoxon rank sum test

data: A and B

W = 41, p-value = 1

alternative hypothesis: true location shift is not equal to 0

Note that the p-value using t-test is 0.9669 and using Wilcoxon Rank Sum test is 1.00. 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 test if A and B come from a normal population. We can do so using the Shapiro-Wilk test as below.

> shapiro.test(A)

Shapiro-Wilk normality test

data: A

W = 0.9399, p-value = 0.581

> shapiro.test(B)

Shapiro-Wilk normality test

data: B

W = 0.9878, p-value = 0.9923

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

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

**Q.N. 4)** Grades of students taking a Statistics course and a Mathematics course simultaneously is given below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistics Grades/Math Grades | A | B | C | Other |
| A | 25 | 6 | 17 | 13 |
| B | 17 | 16 | 15 | 6 |
| C | 18 | 4 | 18 | 10 |
| Other | 10 | 8 | 11 | 20 |

Are the grades in Statistics and Mathematics course dependent? Perform the chi-square test of independence.

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

> data=c(25,6,17,13,17,16,15,6,18,4,18,10,10,8,11,20)

> m=matrix(data,nrow=4,byrow=T)

> chisq.test(m)

Pearson's Chi-squared test

data: m

X-squared = 25.5545, df = 9, p-value = 0.002415

Decision: Note that p=0.002415<0.05, so we reject the null hypothesis that the grades in Statistics and Mathematics are independent. Therefore the grades in Mathematics and Statistics are dependent.

**Q.N. 5)** An article entitled “Assessing claims made by a pizza chain” by Peter K. Dunn appeared in the Journal of Statistics Education, Volume 20, Number 1 (March 2012). The data provides the diameters of 250 pizzas, 125 each from two pizza chains (Domino’s or Eagle Boys) in Australia for a variety of crust types and toppings. The diameter of the pizzas is given in centimeters. Please visit

[http://www.amstat.org/publications/jse/jse\_data\_archive.htm](http://www.amstat.org/publications/jse/jse_data_archive.htm  and obtain pizzasize.CSV)

[and obtain pizzasize.CSV](http://www.amstat.org/publications/jse/jse_data_archive.htm  and obtain pizzasize.CSV) data file. Please note that 192 and 193 observations are not included (for unknown reason as author cited.)

1. Test whether the Eagle Boys ThiNCrust pizzas are 12 inches (30.48cm) in diameter.
2. Are Eagle Boys Thin crust pizzas bigger than Domino’s ThinNCrispy crust pizzas?

**Solution:**

**a)** Save the data pizzasize.csv and import it to R and extract the Eagle Boys ThiNCrust pizzas using R code below

**>**pizza=read.csv("C://STAT4001//pizzasize.csv", header=T)

> head(pizza,5)

ID Store CrustDescription Topping Diameter

1 1 Dominos ThinNCrispy Supreme 29.40

2 2 Dominos ThinNCrispy BBQMeatlovers 29.63

3 3 Dominos DeepPan Hawaiian 27.06

4 4 Dominos ThinNCrispy Supreme 27.45

5 5 Dominos ClassicCrust Hawaiian 26.59

>ThinPizza=subset(pizza,CrustDescription=="ThinCrust"& Store=="EagleBoys")

> head(ThinPizza,5)

ID Store CrustDescription Topping Diameter

9 9 EagleBoys ThinCrust BBQMeatlovers 30.05

11 11 EagleBoys ThinCrust SuperSupremo 29.14

14 14 EagleBoys ThinCrust Hawaiian 29.98

16 16 EagleBoys ThinCrust SuperSupremo 29.47

17 17 EagleBoys ThinCrust SuperSupremo 30.00

Now, we can perform the t-test to see whether the Eagle Boys ThiNCrust pizzas are 12 inches (30.48cm) in diameter using the R code below

> t.test(ThinPizza$Diameter,mu=30.48)

One Sample t-test

data: ThinPizza$Diameter

t = -8.8519, df = 38, p-value = 9.033e-11

alternative hypothesis: true mean is not equal to 30.48

95 percent confidence interval:

29.52225 29.87878

sample estimates:

mean of x

29.70051

Since p=9.033e-11<0.05 we reject the null hypothesis that the mean is equal to 30.48. Thus we have evidence that the thin crust pizzas are not 12 inches (30.48 cm) in diameter

**b)** In order to test whether the Eagle Boys Thin crust pizzas bigger than Domino’s ThinNCrispy crust pizzas we perform the two sample t-test by importing the required data using R code below:

> DThinPizza<-subset(pizza,CrustDescription=="ThinNCrispy" & Store=="Dominos")

> head(DThinPizza,5)

ID Store CrustDescription Topping Diameter

1 1 Dominos ThinNCrispy Supreme 29.40

2 2 Dominos ThinNCrispy BBQMeatlovers 29.63

4 4 Dominos ThinNCrispy Supreme 27.45

21 21 Dominos ThinNCrispy Supreme 26.38

22 22 Dominos ThinNCrispy Hawaiian 28.86

> t.test(ThinPizza$Diameter,DThinPizza$Diameter,alt="greater")

Welch Two Sample t-test

data: ThinPizza$Diameter and DThinPizza$Diameter

t = 5.8837, df = 74.693, p-value = 5.303e-08

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.6352648 Inf

sample estimates:

mean of x mean of y

29.70051 28.81442

Decision: Since p-value is less than 0.05 we reject the null hypothesis. Thus we have enough evidence to conclude that that Eagle Boys Thin Crust pizza has a larger diameter than Domino’s ThinNCrispy pizza.

We can also justify the application of parametric test by testing the normality of each data using Shapiro test.