**1. multiple choice R questions**

Problem 1. Which of the following R functions would you use to prepare a plot of average score for each grade in Professor Moody puzzle?

a. merge() b. histogram()  
c. sample() d. tapply()

Problem 2. (what would R say?)

u<-c(1:10)

w <-c(1,-1,3)

u[w>0]

what would R say?

a. error  
b. 1 3 4 6 7 9 10  
c. 1 3 4 5 6 7 8 9 10

Problem 3 (what would R say?

weather =data.frame(Day=c('weekday', 'weekend'), Conditions =c('sunny','rainy','cloudy')

weather

what would R say?  
  
a. error  
b. show a data frame with dim = 2,3  
c. show a data frame with dim = 3,2

Problem 4 (what would R say?)

weather =data.frame(Day=c('weekday', 'weekend'), Conditions =c('sunny','rainy', 'cloudy','snow'))  
   
what would R say?  
  
a. show a data frame with dim = 2,4  
b. show a data frame with dim = 4,2  
c. error

Problem 5   
  
weather =data.frame(Day=c('weekday', 'weekend', 'weekday', 'weekend'), Temperature =c(55,61,62,47))  
  
To select days where temperature is less than 60 you will write  
  
a) weather[weather$Temperature<60]  
b) weather[weather$Temperature<60,]  
c) weather(weather$Temperature<60)

Problem 6. Which of the following is a correct way to select 2 rows from the data frame  
traffic?

(a) traffic[1,2]  
(b) traffic[1:2,]  
(c) traffic[,c(1,2)]  
(d) traffic[c(1,2)]  
(e) traffic[2]

Problem 7. Suppose v <- c(-2,0,2,-5) is entered into the R console. What would R say  
if you enter v[v>0]?

(a) [1] TRUE FALSE FALSE TRUE  
(b) [1] 1 0 0 1  
(c) [1] 1 4

(d) [1] 2 //Partial credit given for some answers

Problem 8

weather =data.frame(Day=c('weekday', 'weekend', ‘weekday', 'weekend'), Temprature =c(55,61,62,47))  
u<-rep('warm',4)  
u[weather$Temprature<60]<-'cold'  
u

what would R say?

a) '1','0','0','1'  
b) "cold" "warm" "warm" "cold"  
c) error  
d) show data frame of dim =2,2

**2. Statistics and plotting**  
  
Problem 11. A p-value of 0.05 means:  
  
(a) There is a 95% chance that our claim is correct.  
(b) There is a 5% chance that our claim is correct.  
(c) If null hypothesis was true, then we’d expect to encounter a result equal or more extreme that we observed no more than 5% of the time.  
(d) The results are random.

Problem 12. The goal of a permutation test is to:  
  
(a) Prove that the null hypothesis is true.  
(a) Show how often the observed results could happen by random chance.  
(b) Prove that the order in which elements are added to the dataset is irrelevant.  
(c) Show that every permutation of the dataset gives the same results.

Problem 13. Central limit theorem assumes

a) Normal distribution of data  
b) Uniform distribution of data  
c) Does not make any assumptions about data distribution

Problem 14. In order to check whether indeed Holland tunnel traffic is higher than Lincoln traffic (H>L) instead of permutation test, you can use z-test.

What data is sufficient to calculate p-value for hypothesis that H>L  
  
a. Means of traffic for Lincoln and Holland tunnels respectively  
b. Standard deviations of traffic for Lincoln and Holland tunnels respectively  
c. Standard deviations of traffic for Lincoln and Holland tunnels along with the number of records for both tunnels respectively  
d. Overall standard deviation of all recorded traffic volumes for both tunnels as well as total number of records for both tunnels

//Extra credits given

For Problems 15-19 determine which type of plot would be best to use for visualizing/plotting  
the described data. Choose from the following answers

(a) Box plot  
(b) Bar plot  
(c) Scatter plot  
(d) Mosaic plot  
(e) Histogram  
(f) piechart  
(g) stacked bar plot  
(h) grouped by plot

//Credits also given if logical explanation provided for some choices other than the ones marked.

Problem 15. The distribution of scores for latecomers to the class in professor Moody data set A or E

Problem 16. The average score of "frequent" smartphone users vs average score of "frequent" question askers in professor Moody data set B

Problem 17. The price of wine vs its rating C

Problem 18. The distribution of Presidential candidates support among women, men, college educated, in NY and NJ according to a hypothetical poll. G or H

Problem 19. The height and average points-per-game for all basketball players in the NBA  
during the 2011-2012 season. C

Problem 20. When comparing happiness levels in US and Canada. 500 Americans and 100 Canadians are randomly selected from each country. The sample mean difference of happiness levels of Americans and Canadians is 0.15. The standard deviation of American sample is 1 and the standard deviation of Canadian is 0.4. Are Americans happier than Canadians? Calculate the p-value to support your hypothesis. (hint: part of the z score table is shown as follows)

|  |  |  |  |
| --- | --- | --- | --- |
| z | 0.00 | 0.01 | 0.02 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 |

Z=2.5

Problem 21. Based on the following table, let’s assume that permutation test for Volume of Traffic in Holland Tunnel being higher than in Lincoln Tunnel based on this small data set returned p=0.0003.

Would you accept the hypothesis that Holland Tunnel is busier than Lincoln Tunnel? No

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **TUNNEL** | DAY | VOLUME |
| Lincoln | Weekend | 41 |
| Holland | Weekday | 94 |
| Holland | Weekday | 87 |
| Holland | Weekday | 96 |
| Holland | weekday | 80 |
| Lincoln | weekend | 48 |
| Lincoln | weekend | 39 |
| Holland | weekend | 51 |
| Lincoln | weekend | 40 |